

TEXTBOOK OF
HUMAN METABOLISM IN
HEALTH AND DISEASE

TEXTBOOK OF HUMAN METABOLISM IN HEALTH AND DISEASE

ANDREW THOMPSON



SPRINGER *Med Press*
www.springermedpress.com



SPRINGER Med Press

Published by Springer Med Press
1202 N. Orange Street Suit #600
Wilmington, DE19801 USA

© 2023 by Springer Med Press

ISBN: 979-8-88626-032-8

Textbook of Human Metabolism in Health and Disease

Andrew Thompson

This work is subject to copyright. All rights are reserved by the publisher. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without prior permission in writing from the publisher.

Notice

Practitioners and researchers must always rely on their experience and knowledge in evaluating and using information, methods, compounds, or experiments described herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made. To the fullest extent of the law, no responsibility is assumed by Springer Med Press, authors, editors, or contributors for any injury and damage to persons or property as a matter of products liability, negligence, or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from Library of Congress

For more information regarding Springer Med Press and its products, please visit the publisher's website
www.springermedpress.com

CONTENTS

Preface

ix

1.	Water, Acids, Bases and Buffers	1
	Properties of Water	1
	Physical Properties	8
	Solution	14
	Micelle	17
	Hydrophobic Effect	20
	Colligative Properties	21
	Self-ionization of Water	22
	pH Scale	24
	Buffers	26
	Henderson–Hasselbalch equation	26
	Buffer Systems of Blood	28
	Exchange of O ₂ and CO ₂	29
	Blood Buffer Calculations	30
	Measurement of pH	36
	Nuclear Magnetic Resonance	41
	Magnetic Resonance Imaging	53
	Gibbs–Donnan Effect	62
	H ⁺ Concentration and pH	64
	Respiratory Function of Hemoglobin	70
2.	Amino Acid	79
	History	80
	General Structure	80
	Isomerism	80
	Side Chains	81
	Zwitterions	81
	Isoelectric Point	81
	Occurrence and Functions in Biochemistry	82
	Non-proteinogenic Amino Acids	90
	D-amino Acid Natural Abundance	105
	Non-standard Amino Acids	105
	Non-protein Functions	107
	Uses in Industry	108
	Expanded Genetic Code	108
	Chemical Building Blocks	108
	Biodegradable Plastics	108
	Reactions	109
	Chemical Synthesis	109
	Peptide Bond Formation	129
	Biosynthesis	130
	Catabolism	130
	Physicochemical Properties of Amino Acids	130
	Table of Standard Amino Acid Abbreviations and Properties	131

3.	Protein Structure	133
	Levels of Protein Structure	133
	Protein Primary Structure	133
	Protein Secondary Structure	138
	Protein Tertiary Structure	142
	Protein Quaternary Structure	144
	Protein Structure and Stability	146
	Domains, Motifs, and Folds in Protein Structure	147
	Structural Domain	147
	Structural and Sequence Motif	147
	Supersecondary Structure	147
	Protein Fold	147
	Superdomain	147
	Protein Folding	147
	Process of Protein Folding	148
	Computational Methods for Studying Protein Folding	151
	Experimental Techniques for Studying Protein Folding	153
	Other Information	155
	Protein Structure Determination	156
	Protein Sequence Analysis: Ensembles	156
	Protein Structure Databases	157
	Structure Classification	157
	Computational Prediction of Protein Structure	158
	High Resolution	158
	X-ray Crystallography	158
	Nuclear Magnetic Resonance Spectroscopy of Proteins	172
	Electron Crystallography	178
	Electron Paramagnetic Resonance	180
4.	Thermodynamics	183
	History	184
	Etymology	184
	Branches of Thermodynamics	185
	Classical Thermodynamics	185
	Statistical Mechanics	185
	Chemical Thermodynamics	185
	Treatment of Equilibrium	185
	Laws of Thermodynamics	185
	Zeroth Law	186
	First Law of Thermodynamics	189
	Second Law of Thermodynamics	203
	Third Law of Thermodynamics	215
	System Models	217
	States and Processes	218
	Instrumentation	218
	Conjugate Variables	219
	Potentials	219
5.	Kinetics	221
	Kinetics (Physics)	221
	Biomechanics and Kinetics	221
	Rigid Body Dynamics	221
	Chemical Kinetics	223
	History	223
	Factors Affecting Reaction Rate	224
	Experimental Methods	226
	Equilibrium	226
	Free Energy	226
	Applications and Models	227
	Enzyme Kinetics	227
	Pharmacokinetics	232

Pharmacokinetic Models	233
LADME	236
Analysis	236
Population Pharmacokinetics	236
Clinical Pharmacokinetics	237
Kinesiology	237
Basics	238
Principles	238
Scope of Practice	240
Health Service	240
Technology in Kinesiology	241
Cot Analysis	241
Procedure	241
Analysis	241
Application to Genome Sequencing	242
History	242

6. Energy Metabolism 243

Metabolism	243
Types	243
ATP	244
Physical and Chemical Properties	244
Ionization in Biological Systems	245
Biosynthesis	246
Regulation of Biosynthesis	248
Functions in Cells	248
Amino Acid Activation in Protein Synthesis	250
Binding to Proteins	251
ATP Analogues	251
PC or Phosphagen System	251
Anaerobic Metabolism	251
Anaerobic Metabolism of Glucose	251
Intensity versus Duration	252
Byproducts	252
Aerobic Metabolism	252
Function	252
Significance	252
Effects	253
Misconceptions	253
Time Frame	253
Aerobic and Anaerobic System during Rest and Exercise	253
Energy Systems Review	253
Monitoring Your Energy Usage	254
Exercise Mode and Energy Usage	254
Short Duration High Intensity Exercise	255
Polyometrics	256
Variety is the Key	256
High Intensity Exercise Lasting Several Minutes	256
High Intensity and Long duration Exercise	256
Principles	257
Controversy	257
HIT and Other Training Routines	258
Rest-pause	258
High-intensity Interval Training	258
Long Duration Exercise	261

7. Simple Carbohydrates 263

Introduction	263
Terminology	263
Structure	264
Division	264

Monosaccharides	265
Structure and Nomenclature	265
Derivatives	267
Disaccharide	268
Classification	268
Formation	268
Properties	268
Assimilation	269
Common Disaccharides	269
Cellobiose	269
Isomaltose	269
Isomaltulose	270
Lactose	273
Lactulose	274
Maltose	276
Sucrose	277
Trehalose	287
Turanose	289
Nutrition	289
Classification	290
Effects of Dietary Carbohydrate Restriction	290
Metabolism	291
Catabolism	291
Carbohydrate Chemistry	291
Cyanohydrin Reaction	292
Lobry de Bruyn–van Ekenstein Transformation	292
Amadori Rearrangement	293
Nef Reaction	293
Wohl Degradation	294
Koenigs–Knorr Reaction	294
Carbohydrate Digestion	295
Trisaccharides	296
Examples	296
Maltotriose	296
Melezitose	296
Raffinose	296
Tetrasaccharides	297
Stachyose	297
<i>Bibliography</i>	299
<i>Index</i>	303

PREFACE

Medical Biochemistry is the study of chemical processes within and relating to living organisms. This book presents the fundamentals of Medical Biochemistry and related topics to students pursuing a one-or two-semester course in pre-med biochemistry or medical programs. This book is equally applicable to other health-related fields such as clinical chemistry, medical technology or pharmacology. This book focusses on the foundations and clinically relevant applications of normal human biochemistry and pathology. It presents contemporary coverage of the fundamentals of biochemistry, emphasizing relevant physiologic and pathophysiologic biochemical concepts. Users will find basic and fundamental concepts that will aid students and professionals.

The author would like to express his sincere thanks to his friends and family for their support in the completion of this book. He wants to thank Springer MedPress Corporation to recognise the intensity of the subject and help me publish this masterpiece which will educate well the newcomers. Also thanks the staff and the editorial team for their kind support and encouragement. This book has all the relevant details with the new concepts.

-Author

Water, Acids, Bases and Buffers

PROPERTIES OF WATER

Hydrogen bond

A hydrogen bond is a partially electrostatic attraction between a hydrogen (H) which is bound to a more electronegative atom such as nitrogen (N), oxygen (O), or fluorine (F), and another adjacent atom bearing a lone pair of electrons.



Fig. Properties of Water

Hydrogen bonds can occur between molecules (*intermolecular*) or within different parts of a single molecule (*intramolecular*). Depending on the nature of the donor and acceptor atoms which constitute the bond, their geometry, and environment, the energy of a hydrogen bond can vary between 1 and 40 kcal/mol. This makes them somewhat stronger than a van der Waals interaction, and weaker than fully covalent or ionic bonds. This type of bond can occur in inorganic molecules such as water and in organic molecules like DNA and proteins.

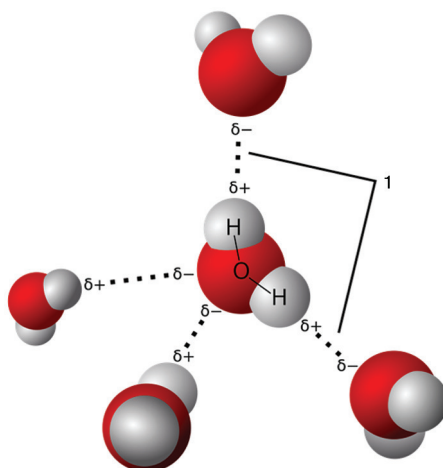


Fig. Hydrogen bond.

Intermolecular hydrogen bonding is responsible for the high boiling point of water (100 °C) compared to the other group 16 hydrides that have much weaker hydrogen bonds. Intramolecular hydrogen bonding is partly responsible for the secondary and tertiary structures of proteins and nucleic acids. It also plays an important role in the structure of polymers, both synthetic and natural.

In 2011, an IUPAC Task Group recommended a modern evidence-based definition of hydrogen bonding, which was published in the IUPAC journal *Pure and Applied Chemistry*.

This definition specifies:

- The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.

Bonding

Definitions and General Characteristics

A hydrogen atom attached to a relatively electronegative atom is the hydrogen bond *donor*. C–H bonds only participate in hydrogen bonding when the carbon atom is bound to electronegative substituents, as is the case in chloroform, CHCl_3 . In a hydrogen bond, the electronegative atom not covalently attached to the hydrogen is named proton acceptor, whereas the one covalently bound to the hydrogen is named the proton donor. In the donor molecule, the H center is protic. The donor is a Lewis base. Hydrogen bonds are represented as $\text{H}\cdots\text{Y}$ system, where the dots represent the hydrogen bond. Liquids that display hydrogen bonding (such as water) are called associated liquids.

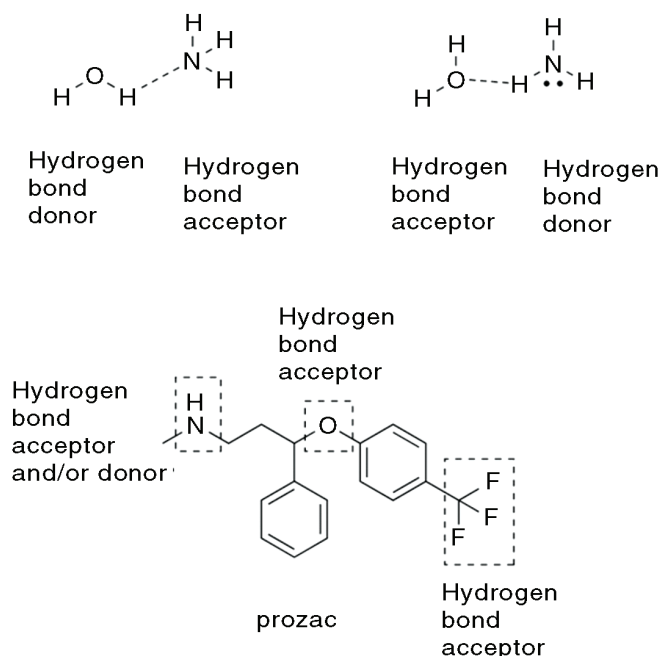


Fig. Examples of Hydrogen Bond Donating (Donors) and Hydrogen Bond Accepting Groups (Acceptors).

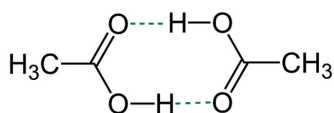


Fig. Cyclic Dimer of Acetic Acid; Dashed Green Lines Represent Hydrogen Bonds.

The hydrogen bond is often described as an electrostatic dipole-dipole interaction. However, it also has some features of covalent bonding: it is directional and strong, produces interatomic distances shorter than the sum of the van der Waals radii, and usually involves a limited number of interaction partners, which can be interpreted as a type of valence. These covalent features are more substantial when acceptors bind hydrogens from more electronegative donors.

Bond Strength

Hydrogen bonds can vary in strength from weak ($1\text{--}2\text{ kJ mol}^{-1}$) to strong (161.5 kJ mol^{-1} in the ion HF_2^-).

Typical enthalpies in vapor include:

- $\text{F-H}\cdots\text{F}$ (161.5 kJ/mol or 38.6 kcal/mol), illustrated uniquely by HF_2^- , bifluoride
- $\text{O-H}\cdots\text{N}$ (29 kJ/mol or 6.9 kcal/mol), illustrated water-ammonia
- $\text{O-H}\cdots\text{O}$ (21 kJ/mol or 5.0 kcal/mol), illustrated water-water, alcohol-alcohol
- $\text{N-H}\cdots\text{N}$ (13 kJ/mol or 3.1 kcal/mol), illustrated by ammonia-ammonia
- $\text{N-H}\cdots\text{O}$ (8 kJ/mol or 1.9 kcal/mol), illustrated water-amide
- $\text{HO-H}\cdots\text{OH}^+$ (18 kJ/mol or 4.3 kcal/mol).

Structural Details

The X-H distance is typically $\text{H-}110\text{ pm}$, whereas the $\text{H}\cdots\text{Y}$ distance is $\text{H-}160\text{ to }200\text{ pm}$. The typical length of a hydrogen bond in water is 197 pm . The ideal bond angle depends on the nature of the hydrogen bond donor. The following hydrogen bond angles between a hydrofluoric acid donor and various acceptors have been determined experimentally:

Acceptor $\cdots$ donor	VSEPR geometry	Angle ($^\circ$)
$\text{HCN}\cdots\text{HF}$	Linear	180
$\text{H}_2\text{CO}\cdots\text{HF}$	Trigonal planar	120
$\text{H}_2\text{O}\cdots\text{HF}$	Pyramidal	46
$\text{H}_2\text{S}\cdots\text{HF}$	Pyramidal	89
$\text{SO}_2\cdots\text{HF}$	Trigonal	142

Spectroscopy

Strong hydrogen bonds are revealed by downfield shifts in the ^1H NMR spectrum. For example, the acidic proton in the enol tautomer of acetylacetone appears at $\delta 15.5$, which is about 10 ppm downfield of a conventional alcohol. In the IR spectrum, hydrogen bonding shifts the X-H stretching frequency to lower energy (*i.e.*, the vibration frequency decreases). This shift reflects a weakening of the X-H bond. Certain hydrogen bonds-improper hydrogen bonds-show a blue shift of the X-H stretching frequency and a decrease in the bond length.

Theoretical Considerations

Hydrogen bonding is of continuing theoretical interest. According to a modern description O:H-O integrates both the intermolecular O:H lone pair “:” nonbond and the intramolecular H-O polar-covalent bond associated with O-O repulsive coupling.

Quantum chemical calculations of the relevant interresidue potential constants (compliance constants) revealed large differences between individual H bonds of the same type. For example, the central interresidue $\text{N-H}\cdots\text{N}$ hydrogen bond between guanine and cytosine is much stronger in comparison to the $\text{N-H}\cdots\text{N}$ bond between the adenine-thymine pair.

Theoretically, the bond strength of the hydrogen bonds can be assessed using NCI index, non-covalent interactions index, which allows a visualization of these non-covalent interactions, as its name indicates, using the electron density of the system. From interpretations of the anisotropies in the Compton profile of ordinary ice that the hydrogen bond is partly covalent. However, this interpretation was challenged. Most generally, the hydrogen bond can be viewed as a metric-dependent electrostatic scalar field between two or more intermolecular bonds. This is slightly different from the intramolecular bound states of, for example, covalent or ionic bonds; however, hydrogen bonding is generally still a bound state phenomenon, since, the interaction energy has a net negative sum. The initial theory of hydrogen bonding proposed by Linus Pauling suggested that the hydrogen bonds had a partial covalent nature. This interpretation remained controversial until NMR techniques demonstrated information transfer between hydrogen-bonded nuclei, a feat that would only be possible if the hydrogen bond contained some covalent character.

History

In the book *The Nature of the Chemical Bond*, Linus Pauling credits T. S. Moore and T. F. Winmill with the first mention of the hydrogen bond, in 1912. Moore and Winmill used the hydrogen bond to account for the fact that trimethylammonium hydroxide is a weaker base than tetramethylammonium hydroxide. The description of hydrogen bonding in its better-known setting, water, came some years later, in 1920, from Latimer and Rodebush. In that paper, Latimer and Rodebush cite work by a fellow scientist at their laboratory, Maurice Loyal Huggins, saying, “Mr. Huggins of this laboratory in some work as yet unpublished, has used the idea of a hydrogen kernel held between two atoms as a theory in regard to certain organic compounds.”

Hydrogen Bonds in Small Molecules

Water

A ubiquitous example of a hydrogen bond is found between water molecules. In a discrete water molecule, there are two hydrogen atoms and one oxygen atom. Two molecules of water can form a hydrogen bond between them; the simplest case, when only two molecules are present, is called the water dimer and is often used as a model system. When more molecules are present, as is the case with liquid water, more bonds are possible because the oxygen of one water molecule has two lone pairs of electrons, each of which can form a hydrogen bond with a hydrogen on another water molecule. This can repeat such that every water molecule is H-bonded with up to four other molecules, as shown in the figure (two through its two lone pairs, and two through its two hydrogen atoms). Hydrogen bonding strongly affects the crystal structure of ice, helping to create an open hexagonal lattice. The density of ice is less than the density of water at the same temperature; thus, the solid phase of water floats on the liquid, unlike most other substances.

Liquid water's high boiling point is due to the high number of hydrogen bonds each molecule can form, relative to its low molecular mass. Owing to the difficulty of breaking these bonds, water has a very high boiling point, melting point, and viscosity compared to otherwise similar liquids not conjoined by hydrogen bonds. Water is unique because its oxygen atom has two lone pairs and two hydrogen atoms, meaning that the total number of bonds of a water molecule is up to four.

The number of hydrogen bonds formed by a molecule of liquid water fluctuates with time and temperature. From TIP4P liquid water simulations at 25 °C, it was estimated that each water molecule participates in an average of 3.59 hydrogen bonds. At 100 °C, this number decreases to 3.24 due to the increased molecular motion and decreased density, while at 0 °C, the average number of hydrogen bonds increases to 3.69. A more recent study found a much smaller number of hydrogen bonds: 2.357 at 25 °C. The differences may be due to the use of a different method for defining and counting the hydrogen bonds.

Where the bond strengths are more equivalent, one might instead find the atoms of two interacting water molecules partitioned into two polyatomic ions of opposite charge, specifically hydroxide (OH^-) and hydronium (H_3O^+). (Hydronium ions are also known as “hydroxonium” ions.)



Indeed, in pure water under conditions of standard temperature and pressure, this latter formulation is applicable only rarely; on average about one in every 5.5×10^8 molecules gives up a proton to another water molecule, in accordance with the value of the dissociation constant for water under such conditions. It is a crucial part of the uniqueness of water.

Because water may form hydrogen bonds with solute proton donors and acceptors, it may competitively inhibit the formation of solute intermolecular or intramolecular hydrogen bonds. Consequently, hydrogen bonds between or within solute molecules dissolved in water are almost always unfavorable relative to hydrogen bonds between water and the donors and acceptors for hydrogen bonds on those solutes. Hydrogen bonds between water molecules have an average lifetime of 10^{-11} seconds, or 10 picoseconds.

Bifurcated and Over-coordinated Hydrogen Bonds in Water

A single hydrogen atom can participate in two hydrogen bonds, rather than one. This type of bonding is called “bifurcated” (split in two or “two-forked”). It can exist, for instance, in complex natural or synthetic organic molecules. It has been suggested that a bifurcated hydrogen atom is an essential step in water reorientation.

Acceptor-type hydrogen bonds (terminating on an oxygen’s lone pairs) are more likely to form bifurcation (it is called overcoordinated oxygen, OCO) than are donor-type hydrogen bonds, beginning on the same oxygen’s hydrogen.

Other Liquids

For example, hydrogen fluoride—which has three lone pairs on the F atom but only one H atom—can form only two bonds; (ammonia has the opposite problem: three hydrogen atoms but only one lone pair).



Further Manifestations of Solvent Hydrogen Bonding

- Increase in the melting point, boiling point, solubility, and viscosity of many compounds can be explained by the concept of hydrogen bonding.
- Negative azeotropy of mixtures of HF and water
- The fact that ice is less dense than liquid water is due to a crystal structure stabilized by hydrogen bonds.
- Dramatically higher boiling points of NH_3 , H_2O , and HF compared to the heavier analogues PH_3 , H_2S , and HCl, where hydrogen-bonding is absent.
- Viscosity of anhydrous phosphoric acid and of glycerol
- Dimer formation in carboxylic acids and hexamer formation in hydrogen fluoride, which occur even in the gas phase, resulting in gross deviations from the ideal gas law.
- Pentamer formation of water and alcohols in apolar solvents.

Hydrogen Bonds in Polymers

Hydrogen bonding plays an important role in determining the three-dimensional structures and the properties adopted by many synthetic and natural proteins.

DNA

In these macromolecules, bonding between parts of the same macromolecule cause it to fold into a specific shape, which helps determine the molecule's physiological or biochemical role. For example, the double helical structure of DNA is due largely to hydrogen bonding between its base pairs (as well as pi stacking interactions), which link one complementary strand to the other and enable replication.

Proteins

In the secondary structure of proteins, hydrogen bonds form between the backbone oxygens and amide hydrogens. When the spacing of the amino acid residues participating in a hydrogen bond occurs regularly between positions i and $i + 4$, an alpha helix is formed. When the spacing is less, between positions i and $i + 3$, then a 3_{10} helix is formed. When two strands are joined by hydrogen bonds involving alternating residues on each participating strand, a beta sheet is formed. Hydrogen bonds also play a part in forming the tertiary structure of protein through interaction of R-groups. The role of hydrogen bonds in protein folding has also been linked to osmolyte-induced protein stabilization. Protective osmolytes, such as trehalose and sorbitol, shift the protein folding equilibrium towards the folded state, in a concentration dependent manner. While the prevalent explanation for osmolyte action relies on excluded volume effects, that are entropic in nature, recent circular dichroism (CD) experiments have shown osmolyte to act through an enthalpic effect. The molecular mechanism for their role in protein stabilization is still not well established, though several mechanism have been proposed. Recently, computer molecular dynamics simulations suggested that osmolytes stabilize proteins by modifying the hydrogen bonds in the protein hydration layer.

Several studies have shown that hydrogen bonds play an important role for the stability between subunits in multimeric proteins. For example, a study of sorbitol dehydrogenase displayed an important hydrogen bonding network which stabilizes the tetrameric quaternary structure within the mammalian sorbitol dehydrogenase protein family. A protein backbone hydrogen bond incompletely shielded from water attack is a dehydron. Dehydrons promote the removal of water through proteins or ligand binding. The exogenous dehydration enhances the electrostatic interaction between the amide and carbonyl groups by de-shielding their partial charges. Furthermore, the dehydration stabilizes the hydrogen bond by destabilizing the nonbonded state consisting of dehydrated isolated charges. Wool, being a protein fibre, is held together by hydrogen bonds, causing wool to recoil when stretched. However, washing at high temperatures can permanently break the hydrogen bonds and a garment may permanently lose its shape.

Cellulose

Hydrogen bonds are important in the structure of cellulose and derived polymers in its many different forms in nature, such as such as cotton and flax.

Synthetic Polymers

Many polymers are strengthened by hydrogen bonds within and between the chains. Among the synthetic polymers, a well characterized example is nylon, where hydrogen bonds occur in the repeat unit and play a major role in crystallization of the material. The bonds occur between carbonyl and amine groups in the amide repeat unit. They effectively link adjacent chains, which help reinforce the material. The effect is great in aramid fibre, where hydrogen bonds stabilize the linear chains laterally. The chain axes are aligned along the fibre axis, making the fibres extremely stiff and strong. The hydrogen-bond networks make both natural and synthetic polymers sensitive to humidity levels in the atmosphere because water molecules can diffuse into the

surface and disrupt the network. Some polymers are more sensitive than others. Thus nylons are more sensitive than aramids, and nylon 6 more sensitive than nylon-11.

Symmetric Hydrogen Bond

A symmetric hydrogen bond is a special type of hydrogen bond in which the proton is spaced exactly halfway between two identical atoms. The strength of the bond to each of those atoms is equal. It is an example of a three-center four-electron bond. This type of bond is much stronger than a “normal” hydrogen bond. The effective bond order is 0.5, so its strength is comparable to a covalent bond. It is seen in ice at high pressure, and also in the solid phase of many anhydrous acids such as hydrofluoric acid and formic acid at high pressure. It is also seen in the bifluoride ion $[F-H-F]^-$. Symmetric hydrogen bonds have been observed recently spectroscopically in formic acid at high pressure (>GPa). Each hydrogen atom forms a partial covalent bond with two atoms rather than one. Symmetric hydrogen bonds have been postulated in ice at high pressure (Ice X). Low-barrier hydrogen bonds form when the distance between two heteroatoms is very small.

Dihydrogen Bond

The hydrogen bond can be compared with the closely related dihydrogen bond, which is also an intermolecular bonding interaction involving hydrogen atoms. These structures have been known for some time, and well characterized by crystallography; however, an understanding of their relationship to the conventional hydrogen bond, ionic bond, and covalent bond remains unclear. Generally, the hydrogen bond is characterized by a proton acceptor that is a lone pair of electrons in nonmetallic atoms (most notably in the nitrogen, and chalcogen groups). In some cases, these proton acceptors may be pi-bonds or metal complexes. In the dihydrogen bond, however, a metal hydride serves as a proton acceptor, thus forming a hydrogen-hydrogen interaction. Neutron diffraction has shown that the molecular geometry of these complexes is similar to hydrogen bonds, in that the bond length is very adaptable to the metal complex/hydrogen donor system.

Dynamics Probed by Spectroscopic Means

The dynamics of hydrogen bond structures in water can be probed by the IR spectrum of OH stretching vibration. In the hydrogen bonding network in protic organic ionic plastic crystals (POIPCs), which are a type of phase change material exhibiting solid-solid phase transitions prior to melting, variable-temperature infrared spectroscopy can reveal the temperature dependence of hydrogen bonds and the dynamics of both the anions and the cations. The sudden weakening of hydrogen bonds during the solid-solid phase transition seems to be coupled with the onset of orientational or rotational disorder of the ions.

Application to Drugs

Hydrogen bonding is a key to the design of drugs. According to Lipinski's rule of five the majority of orally active drug tend to have between five and ten hydrogen bonds. These interactions exist between nitrogen-hydrogen and oxygen-hydrogen centers. As with many other rules of thumb, many exceptions exist.

Hydrogen Bonding Phenomena

- Occurrence of proton tunneling during DNA replication is believed to be responsible for cell mutations.
- High water solubility of many compounds such as ammonia is explained by hydrogen bonding with water molecules.

- Deliquescence of NaOH is caused in part by reaction of OH^- with moisture to form hydrogen-bonded H_3O^- species. An analogous process happens between NaNH_2 and NH_3 , and between NaF and HF.
- The presence of hydrogen bonds can cause an anomaly in the normal succession of states of matter for certain mixtures of chemical compounds as temperature increases or decreases. These compounds can be liquid until a certain temperature, then solid even as the temperature increases, and finally liquid again as the temperature rises over the “anomaly interval”
- Smart rubber utilizes hydrogen bonding as its sole means of bonding, so that it can “heal” when torn, because hydrogen bonding can occur on the fly between two surfaces of the same polymer.

PHYSICAL PROPERTIES

Water is the chemical substance with chemical formula H_2O ; one molecule of water has two hydrogen atoms covalently bonded to a single oxygen atom. Water is a tasteless, odorless liquid at ambient temperature and pressure, and appears colorless in small quantities, although it has its own intrinsic very light blue hue. Ice also appears colorless, and water vapor is essentially invisible as a gas. Unlike other analogous hydrides of the oxygen family, water is primarily a liquid under standard conditions due to hydrogen bonding. The molecules of water are constantly moving in relation to each other, and the hydrogen bonds are continually breaking and reforming at timescales faster than 200 femtoseconds (2×10^{-13} seconds). However, these bonds are strong enough to create many of the peculiar properties of water, some of which make it integral to life.

Water, Ice, and Vapor

Within the Earth’s atmosphere and surface, the liquid phase is the most common and is the form that is generally denoted by the word “water”. The solid phase of water is known as ice and commonly takes the structure of hard, amalgamated crystals, such as ice cubes, or loosely accumulated granular crystals, like snow. There are other crystalline and amorphous phases of ice than the common hexagonal crystalline ice. The gaseous phase of water is known as water vapor (or steam). Visible steam and clouds are formed from minute droplets of water suspended in the air. Water also forms a supercritical fluid. The critical temperature is 647 K and the critical pressure is 22.064 MPa. In nature this only rarely occurs in extremely hostile conditions. A likely example of naturally occurring supercritical water is in the hottest parts of deep water hydrothermal vents, in which water is heated to the critical temperature by volcanic plumes and the critical pressure is caused by the weight of the ocean at the extreme depths where the vents are located. This pressure is reached at a depth of about 2200 meters: much less than the mean depth of the ocean (3800 meters).

Heat Capacity and Heats of Vaporization and Fusion

Water has a very high specific heat capacity of 4.1814 J/(g·K) at 25 °C—the second highest among all the heteroatomic species (after ammonia), as well as a high heat of vaporization (40.65 kJ/mol or 2257 kJ/kg at the normal boiling point), both of which are a result of the extensive hydrogen bonding between its molecules. These two unusual properties allow water to moderate Earth’s climate by buffering large fluctuations in temperature. Most of the additional energy stored in the climate system since, 1970 has accumulated in the oceans. The specific enthalpy of fusion (more commonly known as latent heat) of water is 333.55 kJ/kg at 0 °C: the same amount of energy is required to melt ice as to warm ice from –160 °C up to its melting point or to heat the same amount of water by about 80 °C. Of common substances, only that of ammonia is higher. This property confers resistance to melting on the ice of glaciers and drift ice. Before and since, the advent of mechanical refrigeration, ice was and still is in common use for retarding food spoilage. The specific heat capacity of ice at –10 °C is 2.03 J/(g·K) and the heat capacity of steam at 100 °C is 2.08 J/(g·K).

Density of Water and Ice

The density of water is about 1 gram per cubic centimetre (62 lb/cu ft): this relationship was originally used to define the gram. The density varies with temperature, but not linearly: as the temperature increases, the density rises to a peak at 3.98 °C (39.16 °F) and then decreases. This unusual negative thermal expansion below 4 °C (39 °F) is also observed in molten silica. Regular, hexagonal ice is also less dense than liquid water—upon freezing, the density of water decreases by about 9%. These effects are due to the reduction of thermal motion with cooling, which allows water molecules to form more hydrogen bonds that prevent the molecules from coming close to each other. While below 4 °C the breakage of hydrogen bonds due to heating allows water molecules to pack closer despite the increase in the thermal motion (which tends to expand a liquid), above 4 °C water expands as the temperature increases. Water near the boiling point is about 4% less dense than water at 4 °C (39 °F). Under increasing pressure, ice undergoes a number of transitions to other polymorphs with higher density than liquid water, such as ice II, ice III, high-density amorphous ice (HDA), and very-high-density amorphous ice (VHDA). The unusual density curve and lower density of ice than of water is vital to life—if water were most dense at the freezing point, then in winter the very cold water at the surface of lakes and other water bodies would sink, the lake could freeze from the bottom up, and all life in them would be killed. Furthermore, given that water is a good thermal insulator (due to its heat capacity), some frozen lakes might not completely thaw in summer. The layer of ice that floats on top insulates the water below. Water at about 4 °C (39 °F) also sinks to the bottom, thus keeping the temperature of the water at the bottom constant.

Density of Saltwater And Ice

The density of salt water depends on the dissolved salt content as well as the temperature. Ice still floats in the oceans, otherwise they would freeze from the bottom up. However, the salt content of oceans lowers the freezing point by about 1.9 °C and lowers the temperature of the density maximum of water to the former freezing point at 0 °C. This is why, in ocean water, the downward convection of colder water is *not* blocked by an expansion of water as it becomes colder near the freezing point. The oceans' cold water near the freezing point continues to sink. So creatures that live at the bottom of cold oceans like the Arctic Ocean generally live in water 4 °C colder than at the bottom of frozen-over fresh water lakes and rivers. As the surface of salt water begins to freeze (at −1.9 °C for normal salinity seawater, 3.5%) the ice that forms is essentially salt-free, with about the same density as freshwater ice. This ice floats on the surface, and the salt that is “frozen out” adds to the salinity and density of the sea water just below it, in a process known as *brine rejection*.

This denser salt water sinks by convection and the replacing seawater is subject to the same process. This produces essentially freshwater ice at −1.9 °C on the surface. The increased density of the sea water beneath the forming ice causes it to sink towards the bottom. On a large scale, the process of brine rejection and sinking cold salty water results in ocean currents forming to transport such water away from the Poles, leading to a global system of currents called the thermohaline circulation.

Miscibility and Condensation

Water is miscible with many liquids, including ethanol in all proportions. Water and most oils are immiscible usually forming layers according to increasing density from the top. This can be predicted by comparing the polarity. Water being a relatively polar compound will tend to be miscible with liquids of high polarity such as ethanol and acetone, whereas compounds with low polarity will tend to be immiscible and poorly soluble such as with hydrocarbons. As a gas, water vapor is completely miscible with air. On the other hand, the maximum water vapor pressure that is thermodynamically stable with the liquid (or solid) at a given temperature is relatively low compared with total atmospheric pressure. For example, if the vapor's partial pressure is 2% of atmospheric

pressure and the air is cooled from 25 °C, starting at about 22 °C water will start to condense, defining the dew point, and creating fog or dew. The reverse process accounts for the fog burning off in the morning. If the humidity is increased at room temperature, for example, by running a hot shower or a bath, and the temperature stays about the same, the vapor soon reaches the pressure for phase change, and then condenses out as minute water droplets, commonly referred to as steam. A saturated gas or one with 100% relative humidity is when the vapor pressure of water in the air is at equilibrium with vapor pressure due to (liquid) water; water (or ice, if cool enough) will fail to lose mass through evaporation when exposed to saturated air. Because the amount of water vapor in air is small, relative humidity, the ratio of the partial pressure due to the water vapor to the saturated partial vapor pressure, is much more useful. Vapor pressure above 100% relative humidity is called super-saturated and can occur if air is rapidly cooled, for example, by rising suddenly in an updraft.

Vapor Pressure

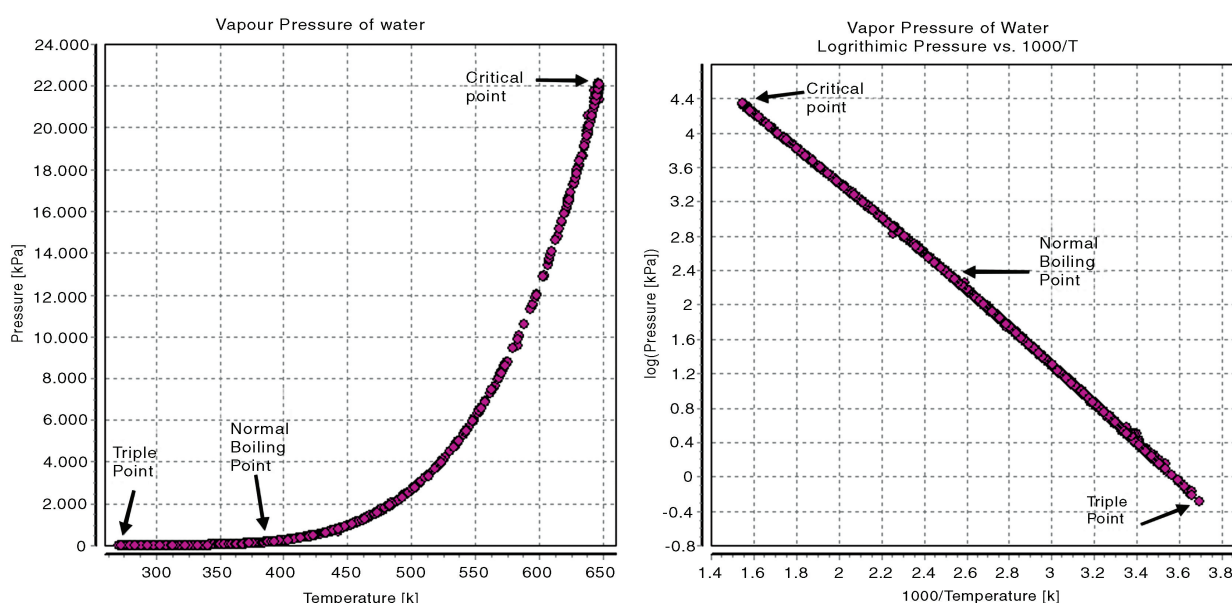


Fig. Vapor pressure diagrams of water.

Compressibility

The compressibility of water is a function of pressure and temperature. At 0 °C, at the limit of zero pressure, the compressibility is $5.1 \times 10^{-10} \text{ Pa}^{-1}$. At the zero-pressure limit, the compressibility reaches a minimum of $4.4 \times 10^{-10} \text{ Pa}^{-1}$ around 45 °C before increasing again with increasing temperature. As the pressure is increased, the compressibility decreases, being $3.9 \times 10^{-10} \text{ Pa}^{-1}$ at 0 °C and 100 megapascals (1,000 bar). The bulk modulus of water is about 2.2 GPa. The low compressibility of non-gases, and of water in particular, leads to their often being assumed as incompressible. The low compressibility of water means that even in the deep oceans at 4 km depth, where pressures are 40 MPa, there is only a 1.8% decrease in volume.

Triple Point

The temperature and pressure at which solid, liquid, and gaseous water coexist in equilibrium is called the triple point of water. This point is used to define the units of temperature (the kelvin, the SI unit of thermodynamic temperature and, indirectly, the degree Celsius and even the degree Fahrenheit). As a consequence, water's triple point temperature, as measured in these units, is a prescribed value rather than a measured quantity.

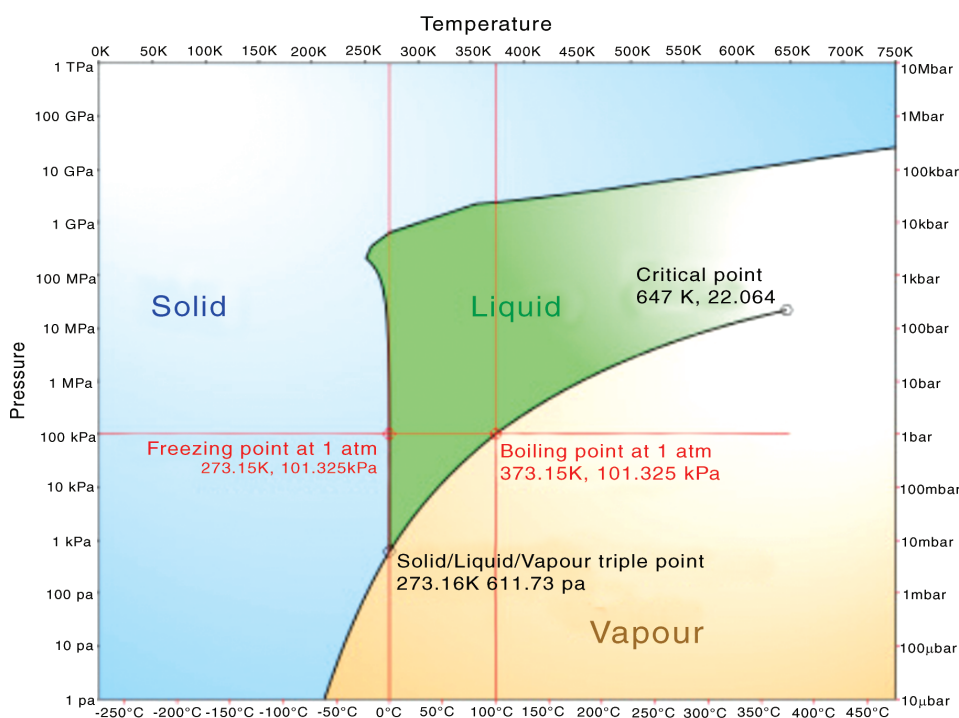


Fig. Phase diagram of water.

This pressure is quite low, about $1/166$ of the normal sea level barometric pressure of 101,325 Pa. The atmospheric surface pressure on planet Mars is 610.5 Pa, which is remarkably close to the triple point pressure. The altitude of this surface pressure was used to define zero-elevation or “sea level” on that planet.

Although it is commonly named as “*the* triple point of water”, the stable combination of liquid water, ice I, and water vapor is but one of several triple points on the phase diagram of water. Gustav Heinrich Johann Apollon Tammann in Göttingen produced data on several other triple points in the early 20th century. Kamb and others documented further triple points in the 1960s.

The various triple points of water		
Phases in stable equilibrium	Pressure	Temperature
Liquid water, ice I _h , and water vapor	611.657 Pa	273.16 K (0.01 °C)
Liquid water, ice I _h , and ice III	209.9 MPa	251 K (−22 °C)
Liquid water, ice III, and ice V	350.1 MPa	−17.0 °C
Liquid water, ice V, and ice VI	632.4 MPa	0.16 °C
Ice I _h , Ice II, and ice III	213 MPa	−35 °C
Ice II, ice III, and ice V	344 MPa	−24 °C
Ice II, ice V, and ice VI	626 MPa	−70 °C

Melting Point

The melting point of ice is 0 °C (32 °F; 273 K) at standard pressure; however, pure liquid water can be supercooled well below that temperature without freezing if the liquid is not mechanically disturbed. It can remain in a fluid state down to its homogeneous nucleation point of about 231 K (−42 °C; −44 °F). The melting point of ordinary hexagonal ice falls slightly under moderately high pressures, by 0.0073 °C (0.0131 °F)/atm or about 0.5 °C (0.90 °F)/70 atm as the stabilization energy of hydrogen bonding is exceeded by intermolecular repulsion, but as ice transforms into its allotropes above 209.9 MPa (2,072 atm), the melting point increases markedly with pressure, *i.e.*, reaching 355 K (82 °C) at 2.216 GPa (21,870 atm) (triple point of Ice VII).

Electrical Properties

Electrical Conductivity

Pure water containing no exogenous ions is an excellent insulator, but not even “deionized” water is completely free of ions. Water undergoes auto-ionization in the liquid state, when two water molecules form one hydroxide anion (OH^-) and one hydronium cation (H_3O^+). Because water is such a good solvent, it almost always has some solute dissolved in it, often a salt.

If water has even a tiny amount of such an impurity, then it can conduct electricity far more readily. It is known that the theoretical maximum electrical resistivity for water is approximately $18.2 \text{ M}\Omega\cdot\text{cm}$ ($182 \text{ k}\Omega\cdot\text{m}$) at 25°C . This figure agrees well with what is typically seen on reverse osmosis, ultra-filtered and deionized ultra-pure water systems used, for instance, in semiconductor manufacturing plants. A salt or acid contaminant level exceeding even 100 parts per trillion (ppt) in otherwise ultra-pure water begins to noticeably lower its resistivity by up to several $\text{k}\Omega\cdot\text{m}$.

In pure water, sensitive equipment can detect a very slight electrical conductivity of $0.05501 \pm 0.0001 \mu\text{S}/\text{cm}$ at 25.00°C . Water can also be electrolyzed into oxygen and hydrogen gases but in the absence of dissolved ions this is a very slow process, as very little current is conducted. In ice, the primary charge carriers are protons. Ice was previously thought to have a small but measurable conductivity of $1 \times 10^{-10} \text{ S}/\text{cm}$, but this conductivity is now thought to be almost entirely from surface defects, and without those, ice is an insulator with an immeasurably small conductivity.

Polarity and Hydrogen Bonding

An important feature of water is its polar nature. The structure has a bent molecular geometry for the two hydrogens from the oxygen vertex. The oxygen atom also has two lone pairs of electrons. One effect usually ascribed to the lone pairs is that the $\text{H}-\text{O}-\text{H}$ gas phase bend angle is 104.48° , which is smaller than the typical tetrahedral angle of 109.47° . The lone pairs are closer to the oxygen atom than the electrons sigma bonded to the hydrogens, so they require more space. The increased repulsion of the lone pairs forces the $\text{O}-\text{H}$ bonds closer to each other.

Another effect of the electronic structure is that water is a polar molecule. Due to the difference in electronegativity, there is a bond dipole moment pointing from each H to the O, making the oxygen partially negative and each hydrogen partially positive. In addition, the lone pairs of electrons on the O are in the direction opposite to the hydrogen atoms. This results in a large molecular dipole, pointing from a positive region between the two hydrogen atoms to the negative region of the oxygen atom. The charge differences cause water molecules to be attracted to each other (the relatively positive areas being attracted to the relatively negative areas) and to other polar molecules. This attraction contributes to hydrogen bonding, and explains many of the properties of water, such as solvent action. Although hydrogen bonding is a relatively weak attraction compared to the covalent bonds within the water molecule itself, it is responsible for a number of water's physical properties.

These properties include its relatively high melting and boiling point temperatures: more energy is required to break the hydrogen bonds between water molecules. In contrast, hydrogen sulfide (H_2S), has much weaker hydrogen bonding due to sulfur's lower electronegativity. H_2S is a gas at room temperature, in spite of hydrogen sulfide having nearly twice the molar mass of water. The extra bonding between water molecules also gives liquid water a large specific heat capacity. This high heat capacity makes water a good heat storage medium (coolant) and heat shield.

Cohesion and Adhesion

Water molecules stay close to each other (cohesion), due to the collective action of hydrogen bonds between water molecules. These hydrogen bonds are constantly breaking, with new bonds being formed with different water molecules; but at any given time in a sample of liquid water, a large portion of the molecules are held together by such bonds. Water also has high adhesion properties because of its polar nature. On extremely clean/smooth glass the water may form a thin film because the molecular forces between glass and water molecules (adhesive forces) are stronger than the cohesive forces. In biological cells and organelles, water is in contact with membrane and protein surfaces that are hydrophilic; that is, surfaces that have a strong attraction to water.

Irving Langmuir observed a strong repulsive force between hydrophilic surfaces. To dehydrate hydrophilic surfaces—to remove the strongly held layers of water of hydration—requires doing substantial work against these forces, called hydration forces. These forces are very large but decrease rapidly over a nanometer or less. They are important in biology, particularly when cells are dehydrated by exposure to dry atmospheres or to extracellular freezing.

Surface Tension

Water has an unusually high surface tension of 71.99 mN/m at 25 °C which is caused by the strength of the hydrogen bonding between water molecules. This allows insects to walk on water.

Capillary Action

Because water has strong cohesive and adhesive forces, it exhibits capillary action. Strong cohesion from hydrogen bonding and adhesion allows trees to transport water more than 100 m upward.

Water as a Solvent

Water is an excellent solvent due to its high dielectric constant. Substances that mix well and dissolve in water are known as hydrophilic (“water-loving”) substances, while those that do not mix well with water are known as hydrophobic (“water-fearing”) substances. The ability of a substance to dissolve in water is determined by whether or not the substance can match or better the strong attractive forces that water molecules generate between other water molecules. If a substance has properties that do not allow it to overcome these strong intermolecular forces, the molecules are precipitated out from the water. Contrary to the common misconception, water and hydrophobic substances do not “repel”, and the hydration of a hydrophobic surface is energetically, but not entropically, favorable.

When an ionic or polar compound enters water, it is surrounded by water molecules (hydration). The relatively small size of water molecules (~ 3 angstroms) allows many water molecules to surround one molecule of solute. The partially negative dipole ends of the water are attracted to positively charged components of the solute, and vice versa for the positive dipole ends. In general, ionic and polar substances such as acids, alcohols, and salts are relatively soluble in water, and non-polar substances such as fats and oils are not. Non-polar molecules stay together in water because it is energetically more favorable for the water molecules to hydrogen bond to each other than to engage in van der Waals interactions with non-polar molecules.

An example of an ionic solute is table salt; the sodium chloride, NaCl, separates into Na⁺ cations and Cl⁻ anions, each being surrounded by water molecules. The ions are then easily transported away from their crystalline lattice into solution. An example of a nonionic solute is table sugar. The water dipoles make hydrogen bonds with the polar regions of the sugar molecule (OH groups) and allow it to be carried away into solution.

Quantum Tunneling

The quantum tunneling dynamics in water was reported as early as 1992. At that time it was known that there are motions which destroy and regenerate the weak hydrogen bond by internal rotations of the substituent water monomers. On 18 March 2016, it was reported that the hydrogen bond can be broken by quantum tunneling in the water hexamer. Unlike previously reported tunneling motions in water, this involved the concerted breaking of two hydrogen bonds. Later in the same year, the discovery of the quantum tunneling of water molecules was reported.

Electromagnetic Absorption

Water is relatively transparent to visible light, near ultraviolet light, and far-red light, but it absorbs most ultraviolet light, infrared light, and microwaves. Most photoreceptors and photosynthetic pigments utilize the portion of the light spectrum that is transmitted well through water. Microwave ovens take advantage of water's opacity to microwave radiation to heat the water inside of foods. The very weak onset of absorption in the red end of the visible spectrum lends water its intrinsic blue hue.

SOLUTION

In chemistry, a solution is a special type of homogeneous mixture composed of two or more substances. The term aqueous solution is when one of the solvents is water. In such a mixture, a solute is a substance dissolved in another substance, known as a solvent. The mixing process of a solution happens at a scale where the effects of chemical polarity are involved, resulting in interactions that are specific to solvation. The solution assumes the phase of the solvent when the solvent is the larger fraction of the mixture, as is commonly the case. The concentration of a solute in a solution is the mass of that solute expressed as a percentage of the mass of the whole solution.

Characteristics

- A solution is a homogeneous mixture of two or more substances.
- The particles of solute in a solution cannot be seen by the naked eye.
- A solution does not allow beams of light to scatter.
- A solution is stable.
- The solute from a solution cannot be separated by filtration (or mechanically).
- It is composed of only one phase.

Types

Homogeneous means that the components of the mixture form a single phase. *Heterogeneous* means that the components of the mixture are of different phase. The properties of the mixture (such as concentration, temperature, and density) can be uniformly distributed through the volume but only in absence of diffusion phenomena or after their completion. Usually, the substance present in the greatest amount is considered the solvent. Solvents can be gases, liquids or solids. One or more components present in the solution other than the solvent are called solutes. The solution has the same physical state as the solvent.

Gaseous Solutions

If the solvent is a gas, only gases are dissolved under a given set of conditions. An example of a gaseous solution is air (oxygen and other gases dissolved in nitrogen). Since, interactions between molecules play

almost no role, dilute gases form rather trivial solutions. In part of the literature, they are not even classified as solutions, but addressed as mixtures.

Liquid Solutions

If the solvent is a liquid, then almost all gases, liquids, and solids can be dissolved.

Here are some examples:

- *Gas in liquid:*
 - Oxygen in water
 - Carbon dioxide in water—a less simple example, because the solution is accompanied by a chemical reaction (formation of ions). Note also that the visible bubbles in carbonated water are not the dissolved gas, but only an effervescence of carbon dioxide that has come out of solution; the dissolved gas itself is not visible since, it is dissolved on a molecular level.
- *Liquid in liquid:*
 - The mixing of two or more substances of the same chemistry but different concentrations to form a constant. (Homogenization of solutions)
 - Alcoholic beverages are basically solutions of ethanol in water.
- *Solid in liquid:*
 - Sucrose (table sugar) in water
 - Sodium chloride (NaCl) (table salt) or any other salt in water, which forms an electrolyte: When dissolving, salt dissociates into ions.
- Solutions in water are especially common, and are called aqueous solutions.
- Non-aqueous solutions are when the liquid solvent involved is not water.

Counter examples are provided by liquid mixtures that are not homogeneous: colloids, suspensions, emulsions are not considered solutions. Body fluids are examples for complex liquid solutions, containing many solutes. Many of these are electrolytes, since, they contain solute ions, such as potassium. Furthermore, they contain solute molecules like sugar and urea. Oxygen and carbon dioxide are also essential components of blood chemistry, where significant changes in their concentrations may be a sign of severe illness or injury.

Solid Solutions

If the solvent is a solid, then gases, liquids and solids can be dissolved.

- *Gas in solids:*
 - Hydrogen dissolves rather well in metals, especially in palladium; this is studied as a means of hydrogen storage.
- *Liquid in solid:*
 - Mercury in gold, forming an amalgam
 - Water in solid salt or sugar, forming moist solids
 - Hexane in paraffin wax
- *Solid in solid:*
 - Steel, basically a solution of carbon atoms in a crystalline matrix of iron atoms
 - Alloys like bronze and many others
 - Polymers containing plasticizers.

Solubility

The ability of one compound to dissolve in another compound is called solubility. When a liquid can completely dissolve in another liquid the two liquids are *miscible*. Two substances that can never mix to form a solution are

called *immiscible*. All solutions have a positive entropy of mixing. The interactions between different molecules or ions may be energetically favored or not. If interactions are unfavorable, then the free energy decreases with increasing solute concentration. At some point the energy loss outweighs the entropy gain, and no more solute particles can be dissolved; the solution is said to be saturated. However, the point at which a solution can become saturated can change significantly with different environmental factors, such as temperature, pressure, and contamination.

For some solute-solvent combinations a supersaturated solution can be prepared by raising the solubility (for example by increasing the temperature) to dissolve more solute, and then lowering it (for example by cooling). Usually, the greater the temperature of the solvent, the more of a given solid solute it can dissolve. However, most gases and some compounds exhibit solubilities that decrease with increased temperature. Such behaviour is a result of an exothermic enthalpy of solution. Some surfactants exhibit this behaviour. The solubility of liquids in liquids is generally less temperature-sensitive than that of solids or gases.

Properties

The physical properties of compounds such as melting point and boiling point change when other compounds are added. Together they are called colligative properties. There are several ways to quantify the amount of one compound dissolved in the other compounds collectively called concentration. Examples include molarity, volume fraction, and mole fraction. The properties of ideal solutions can be calculated by the linear combination of the properties of its components. If both solute and solvent exist in equal quantities (such as in a 50% ethanol, 50% water solution), the concepts of “solute” and “solvent” become less relevant, but the substance that is more often used as a solvent is normally designated as the solvent (in this example, water).

Liquid

In principle, all types of liquids can behave as solvents: liquid noble gases, molten metals, molten salts, molten covalent networks, and molecular liquids. In the practice of chemistry and biochemistry, most solvents are molecular liquids. They can be classified into polar and non-polar, according to whether their molecules possess a permanent electric dipole moment. Another distinction is whether their molecules can form hydrogen bonds (protic and aprotic solvents). Water, the most commonly used solvent, is both polar and sustains hydrogen bonds. Salts dissolve in polar solvents, forming positive and negative ions that are attracted to the negative and positive ends of the solvent molecule, respectively. If the solvent is water, hydration occurs when the charged solute ions become surrounded by water molecules. A standard example is aqueous saltwater. Such solutions are called electrolytes.

Polar solutes dissolve in polar solvents, forming polar bonds or hydrogen bonds. As an example, all alcoholic beverages are aqueous solutions of ethanol. On the other hand, non-polar solutes dissolve better in non-polar solvents. Examples are hydrocarbons such as oil and grease that easily mix with each other, while being incompatible with water. An example for the immiscibility of oil and water is a leak of petroleum from a damaged tanker, that does not dissolve in the ocean water but rather floats on the surface.

Preparation from Constituent Ingredients

It is common practice in laboratories to make a solution directly from its constituent ingredients.

There are three cases in practical calculation:

- *Case 1:* amount of solvent volume is given.
- *Case 2:* amount of solute mass is given.
- *Case 3:* amount of final solution volume is given.

In the following equations, A is solvent, B is solute, and C is concentration. Solute volume contribution is considered through ideal solution model.

- Case 1: Amount (mL) of solvent volume V_A is given. Solute mass $m_B = C V_A d_A / (100 - C/d_B)$
- Case 2: Amount of solute mass m_B is given. Solvent volume $V_A = m_B (100/C - 1/d_B)$
- Case 3: Amount (mL) of final solution volume V_t is given. Solute mass $m_B = C V_t / 100$; Solvent volume $V_A = (100/C - 1/d_B) m_B$
- Case 2: Solute mass is known, $V_A = m_B 100/C$
- Case 3: Total solution volume is known, same equation as case 1. $V_A = V_t$; $m_B = C V_A / 100$

Example: Make 2 g/100mL of NaCl solution with 1 L water Water (properties). The density of resulting solution is considered to be equal to that of water, statement holding especially for dilute solutions, so the density information is not required.

$$m_B = C V_A = (2/100) \text{ g/mL} \times 1000 \text{ mL} = 20 \text{ g.}$$

MICELLE

A micelle is an aggregate (or supramolecular assembly) of surfactant molecules dispersed in a liquid colloid. A typical micelle in aqueous solution forms an aggregate with the hydrophilic “head” regions in contact with surrounding solvent, sequestering the hydrophobic single-tail regions in the micelle centre. This phase is caused by the packing behaviour of single-tail lipids in a bilayer. The difficulty filling all the volume of the interior of a bilayer, while accommodating the area per head group forced on the molecule by the hydration of the lipid head group, leads to the formation of the micelle.

This type of micelle is known as a normal-phase micelle (oil-in-water micelle). Inverse micelles have the head groups at the centre with the tails extending out (water-in-oil micelle). Micelles are approximately spherical in shape. Other phases, including shapes such as ellipsoids, cylinders, and bilayers, are also possible. The shape and size of a micelle are a function of the molecular geometry of its surfactant molecules and solution conditions such as surfactant concentration, temperature, pH, and ionic strength. The process of forming micelles is known as micellisation and forms part of the phase behaviour of many lipids according to their polymorphism.

History

The ability of a soapy solution to act as a detergent has been recognized for centuries. However, it was only at the beginning of the twentieth century that the constitution of such solutions was scientifically studied. Pioneering work in this area was carried out by James William McBain at the University of Bristol. As early as 1913, he postulated the existence of “colloidal ions” to explain the good electrolytic conductivity of sodium palmitate solutions. These highly mobile, spontaneously formed clusters came to be called micelles, a term borrowed from biology and popularized by G.S. Hartley in his classic book *Paraffin Chain Salts: A Study in Micelle Formation*.

Solvation

Individual surfactant molecules that are in the system but are not part of a micelle are called “monomers”. Micelles represent a molecular assembly, in which the individual components are thermodynamically in equilibrium with monomers of the same species in the surrounding medium. In water, the hydrophilic “heads” of surfactant molecules are always in contact with the solvent, regardless of whether the surfactants exist as monomers or as part of a micelle. However, the lipophilic “tails” of surfactant molecules have less contact with

water when they are part of a micelle—this being the basis for the energetic drive for micelle formation. In a micelle, the hydrophobic tails of several surfactant molecules assemble into an oil-like core, the most stable form of which having no contact with water. By contrast, surfactant monomers are surrounded by water molecules that create a “cage” or solvation shell connected by hydrogen bonds. This water cage is similar to a clathrate and has an ice-like crystal structure and can be characterized according to the hydrophobic effect. The extent of lipid solubility is determined by the unfavorable entropy contribution due to the ordering of the water structure according to the hydrophobic effect.

Micelles composed of ionic surfactants have an electrostatic attraction to the ions that surround them in solution, the latter known as counterions. Although the closest counterions partially mask a charged micelle (by up to 90%), the effects of micelle charge affect the structure of the surrounding solvent at appreciable distances from the micelle. Ionic micelles influence many properties of the mixture, including its electrical conductivity. Adding salts to a colloid containing micelles can decrease the strength of electrostatic interactions and lead to the formation of larger ionic micelles. This is more accurately seen from the point of view of an effective charge in hydration of the system.

Energy of Formation

Micelles form only when the concentration of surfactant is greater than the critical micelle concentration (CMC), and the temperature of the system is greater than the critical micelle temperature, or Krafft temperature. The formation of micelles can be understood using thermodynamics: Micelles can form spontaneously because of a balance between entropy and enthalpy. In water, the hydrophobic effect is the driving force for micelle formation, despite the fact that assembling surfactant molecules is unfavorable in terms of both enthalpy and entropy of the system.

At very low concentrations of the surfactant, only monomers are present in solution. As the concentration of the surfactant is increased, a point is reached at which the unfavorable entropy contribution, from clustering the hydrophobic tails of the molecules, is overcome by a gain in entropy due to release of the solvation shells around the surfactant tails. At this point, the lipid tails of a part of the surfactants must be segregated from the water. Hence, they start to form micelles. In broad terms, above the CMC, the loss of entropy due to assembly of the surfactant molecules is less than the gain in entropy by setting free the water molecules that were “trapped” in the solvation shells of the surfactant monomers. Also important are enthalpic considerations, such as the electrostatic interactions that occur between the charged parts of surfactants.

Block Copolymer Micelles

The concept of micelles was introduced to describe the core-corona aggregates of small surfactant molecules, however it has also extended to describe aggregates of amphiphilic block copolymers in selective solvents. It is important to know the difference between these two systems. The major difference between these two types of aggregates is in the size of their building blocks. Surfactant molecules have a molecular weight which is generally of a few hundreds of grams per mole while block copolymers are generally one or two orders of magnitude larger.

Moreover, thanks to the larger hydrophilic and hydrophobic parts, block copolymers can have a much more pronounced amphiphilic nature when compared to surfactant molecules. Because of these differences in the building blocks, some block copolymer micelles behave like surfactant ones, while others don't. It is necessary therefore to make a distinction between the two situations. The former ones will belong to the *dynamic micelles* while the latter will be called *kinetically frozen micelles*.

Dynamic Micelles

Certain amphiphilic block copolymer micelles display a similar behaviour as surfactant micelles. These are generally called dynamic micelles and are characterized by the same relaxation processes assigned to surfactant exchange and micelle scission/recombination. Although the relaxation processes are the same between the two types of micelles, the kinetics of unimer exchange are very different. While in surfactant systems the unimers leave and join the micelles through a diffusion-controlled process, for copolymers the entry rate constant is slower than a diffusion controlled process. The rate of this process was found to be a decreasing power-law of the degree of polymerization of the hydrophobic block to the power $2/3$. This difference is due to the coiling of the hydrophobic block of a copolymer exiting the core of a micelle. Block copolymers which form dynamic micelles are some of the tri-block Poloxamers under the right conditions.

Kinetically Frozen Micelles

When block copolymer micelles don't display the characteristic relaxation processes of surfactant micelles, these are called *kinetically frozen micelles*. These can be achieved in two ways: when the unimers forming the micelles are not soluble in the solvent of the micelle solution, or if the core forming blocks are glassy at the temperature in which the micelles are found. Kinetically frozen micelles are formed when either of these conditions is met. A special example in which both of these conditions are valid is that of polystyrene-*b*-poly(ethylene oxide).

This block copolymer is characterized by the high hydrophobicity of the core forming block, PS, which causes the unimers to be insoluble in water. Moreover, PS has a high glass transition temperature which is, depending on the molecular weight, higher than room temperature. Thanks to these two characteristics, a water solution of PS-PEO micelles of sufficiently high molecular weight can be considered kinetically frozen. This means that none of the relaxation processes, which would drive the micelle solution towards thermodynamic equilibrium, are possible. Pioneering work on these micelles was done by Adi Eisenberg. It was also shown how the lack of relaxation processes allowed great freedom in the possible morphologies formed. Moreover, the stability against dilution and vast range of morphologies of kinetically frozen micelles make them particularly interesting, for example, for the development of long circulating drug delivery nanoparticles.

Inverse/reverse Micelles

In a non-polar solvent, it is the exposure of the hydrophilic head groups to the surrounding solvent that is energetically unfavourable, giving rise to a water-in-oil system. In this case, the hydrophilic groups are sequestered in the micelle core and the hydrophobic groups extend away from the center. These inverse micelles are proportionally less likely to form on increasing headgroup charge, since, hydrophilic sequestration would create highly unfavorable electrostatic interactions.

Supramicelles

Supramicelle is a hierarchical micelle structure (supramolecular assembly) where individual components are also micelles. Supramicelles are formed via bottom-up chemical approaches, such as self-assembly of long cylindrical micelles into radial cross-, star- or dandelion-like patterns in a specially selected solvent; solid nanoparticles may be added to the solution to act as nucleation centers and form the central core of the supermicelle. The stems of the primary cylindrical micelles are composed of various block copolymers connected by strong covalent bonds; within the supermicelle structure they are loosely held together by hydrogen bonds, electrostatic or solvophobic interactions.

Uses

When surfactants are present above the critical micelle concentration (CMC), they can act as emulsifiers that will allow a compound that is normally insoluble (in the solvent being used) to dissolve. This occurs because the insoluble species can be incorporated into the micelle core, which is itself solubilized in the bulk solvent by virtue of the head groups' favorable interactions with solvent species. The most common example of this phenomenon is detergents, which clean poorly soluble lipophilic material (such as oils and waxes) that cannot be removed by water alone. Detergents clean also by lowering the surface tension of water, making it easier to remove material from a surface. The emulsifying property of surfactants is also the basis for emulsion polymerization. Micelle formation is essential for the absorption of fat-soluble vitamins and complicated lipids within the human body. Bile salts formed in the liver and secreted by the gall bladder allow micelles of fatty acids to form. This allows the absorption of complicated lipids (*e.g.*, lecithin) and lipid-soluble vitamins (A, D, E, and K) within the micelle by the small intestine. During the process of milk-clotting, proteases act on the soluble portion of caseins, κ -casein, thus originating an unstable micellar state that results in clot formation. Micelles can also be used for targeted drug delivery as gold nanoparticles.

HYDROPHOBIC EFFECT

The hydrophobic effect is the observed tendency of nonpolar substances to aggregate in an aqueous solution and exclude water molecules. The word hydrophobic literally means “water-fearing”, and it describes the segregation of water and nonpolar substances, which maximizes hydrogen bonding between molecules of water and minimizes the area of contact between water and nonpolar molecules.

The hydrophobic effect is responsible for the separation of a mixture of oil and water into its two components. It is also responsible for effects related to biology, including: cell membranes and vesicles formation, protein folding, insertion of membrane proteins into the nonpolar lipid environment and protein-small molecule associations. Hence, the hydrophobic effect is essential to life. Substances for which this effect is observed are known as hydrophobes.

The hydrophilic groups prevent phase separation of the molecules by maintaining the hydrophobic groups in water through formation of strong hydrogen bonds with water molecules. The driving force for this self-assembly is the hydrophobic effect.

Amphiphiles

Amphiphiles are molecules that have both hydrophobic and hydrophilic domains. Detergents are composed of amphiphiles that allow hydrophobic molecules to be solubilized in water by forming micelles and bilayers (as in soap bubbles). They are also important to cell membranes composed of amphiphilic phospholipids that prevent the internal aqueous environment of a cell from mixing with external water.

Folding of Macromolecules

In the case of protein folding, the hydrophobic effect is important to understanding the structure of proteins that have hydrophobic amino acids (such as glycine, alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and methionine) clustered together within the protein. Structures of water-soluble proteins have a hydrophobic core in which side chains are buried from water, which stabilizes the folded state. Charged and polar side chains are situated on the solvent-exposed surface where they interact with surrounding water molecules. Minimizing the number of hydrophobic side chains exposed to water is the principal driving force behind the folding process, although formation of hydrogen bonds within the protein also stabilizes protein structure. The

energetics of DNA tertiary structure assembly were determined to be driven by the hydrophobic effect, in addition to Watson-Crick base pairing, which is responsible for sequence selectivity, and stacking interactions between the aromatic bases.

Protein Purification

In biochemistry, the hydrophobic effect can be used to separate mixtures of proteins based on their hydrophobicity. Column chromatography with a hydrophobic stationary phase such as phenyl-sepharose will cause more hydrophobic proteins to travel more slowly, while less hydrophobic ones elute from the column sooner. To achieve better separation, a salt may be added (higher concentrations of salt increase the hydrophobic effect) and its concentration decreased as the separation progresses.

Cause

The origin of the hydrophobic effect is not fully understood. Some argue that the hydrophobic interaction is mostly an entropic effect originating from the disruption of highly dynamic hydrogen bonds between molecules of liquid water by the nonpolar solute. A hydrocarbon chain or a similar nonpolar region of a large molecule is incapable of forming hydrogen bonds with water. Introduction of such a non-hydrogen bonding surface into water causes disruption of the hydrogen bonding network between water molecules. The hydrogen bonds are reoriented tangentially to such surface to minimize disruption of the hydrogen bonded 3D network of water molecules, and this leads to a structured water “cage” around the nonpolar surface. The water molecules that form the “cage” (or solvation shell) have restricted mobility. In the solvation shell of small nonpolar particles, the restriction amounts to some 10%.

For example, in the case of dissolved xenon at room temperature a mobility restriction of 30% has been found. In the case of larger nonpolar molecules, the reorientational and translational motion of the water molecules in the solvation shell may be restricted by a factor of two to four; thus, at 25 °C the reorientational correlation time of water increases from 2 to 4-8 picoseconds. Generally, this leads to significant losses in translational and rotational entropy of water molecules and makes the process unfavorable in terms of the free energy in the system. By aggregating together, nonpolar molecules reduce the surface area exposed to water and minimize their disruptive effect.

The hydrophobic effect can be quantified by measuring the partition coefficients of non-polar molecules between water and non-polar solvents. The partition coefficients can be transformed to free energy of transfer which includes enthalpic and entropic components, $\Delta G = \Delta H - T\Delta S$. These components are experimentally determined by calorimetry.

The hydrophobic effect was found to be entropy-driven at room temperature because of the reduced mobility of water molecules in the solvation shell of the non-polar solute; however, the enthalpic component of transfer energy was found to be favorable, meaning it strengthened water-water hydrogen bonds in the solvation shell due to the reduced mobility of water molecules. At the higher temperature, when water molecules become more mobile, this energy gain decreases along with the entropic component. The hydrophobic effect depends on the temperature, which leads to “cold denaturation” of proteins.

COLLIGATIVE PROPERTIES

In chemistry, colligative properties are properties of solutions that depend on the ratio of the number of solute particles to the number of solvent molecules in a solution, and not on the nature of the chemical species present. The number ratio can be related to the various units for concentration of solutions. The assumption that solution properties are independent of nature of solute particles is only exact for ideal solutions,

and is approximate for dilute real solutions. In other words, colligative properties are a set of solution properties that can be reasonably approximated by assuming that the solution is ideal. Here we consider only properties which result from the dissolution of nonvolatile solute in a volatile liquid solvent. They are essentially solvent properties which are changed by the presence of the solute. The solute particles displace some solvent molecules in the liquid phase and therefore reduce the concentration of solvent, so that the colligative properties are independent of the nature of the solute. The word colligative is derived from the Latin *colligatus* meaning *bound together*.

Colligative properties include:

- Relative lowering of vapor pressure
- Elevation of boiling point
- Depression of freezing point
- Osmotic pressure.

For a given solute-solvent mass ratio, all colligative properties are inversely proportional to solute molar mass. Measurement of colligative properties for a dilute solution of a non-ionized solute such as urea or glucose in water or another solvent can lead to determinations of relative molar masses, both for small molecules and for polymers which cannot be studied by other means. Alternatively, measurements for ionized solutes can lead to an estimation of the percentage of dissociation taking place. Colligative properties are mostly studied for dilute solutions, whose behaviour may often be approximated as that of an ideal solution.

Boiling Point and Freezing Point

Addition of solute to form a solution stabilizes the solvent in the liquid phase, and lowers the solvent chemical potential so that solvent molecules have less tendency to move to the gas or solid phases. As a result, liquid solutions slightly above the solvent boiling point at a given pressure become stable, which means that the boiling point increases. Similarly, liquid solutions slightly below the solvent freezing point become stable meaning that the freezing point decreases. Both the boiling point elevation and the freezing point depression are proportional to the lowering of vapor pressure in a dilute solution. These properties are colligative in systems where the solute is essentially confined to the liquid phase. Boiling point elevation (like vapor pressure lowering) is colligative for non-volatile solutes where the solute presence in the gas phase is negligible. Freezing point depression is colligative for most solutes since, very few solutes dissolve appreciably in solid solvents.

SELF-IONIZATION OF WATER

The self-ionization of water (also autoionization of water, and autodissociation of water) is an ionization reaction in pure water or in an aqueous solution, in which a water molecule, H_2O , deprotonates (loses the nucleus of one of its hydrogen atoms) to become a hydroxide ion, OH^- . The hydrogen nucleus, H^+ , immediately protonates another water molecule to form hydronium, H_3O^+ . It is an example of autoprotolysis, and exemplifies the amphoteric nature of water.

Dependence on Temperature, Pressure and Ionic Strength

The dependence of the water ionization on temperature and pressure has been investigated thoroughly. The value of $\text{p}K_{\text{w}}$ decreases as temperature increases from the melting point of ice to a minimum at c. 250 °C, after which it increases up to the critical point of water c. 374 °C. It decreases with increasing pressure.

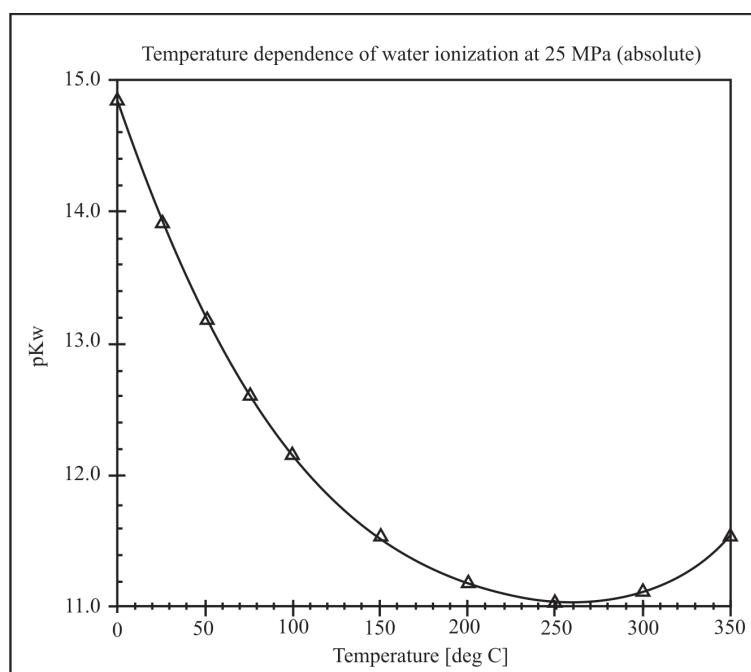


Fig. Temperature dependence of the water ionization constant at 25 MPa.

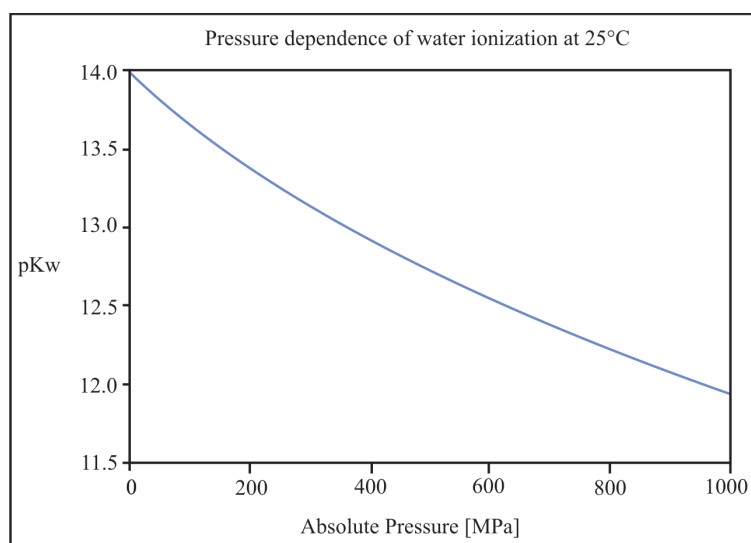


Fig. Pressure dependence of the water ionization constant at 25 °C.

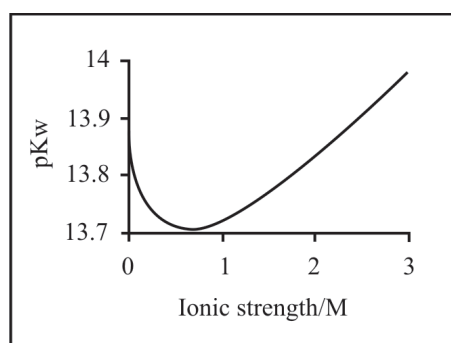


Fig. Variation of pK_w with ionic strength of NaCl solutions at 25 °C.

pK_w values for liquid water.		
Temperature	Pressure	pK_w
0 °C	0.10 MPa	14.95
25 °C	0.10 MPa	13.99
50 °C	0.10 MPa	13.26
75 °C	0.10 MPa	12.70
100 °C	0.10 MPa	12.25
150 °C	0.47 MPa	11.64
200 °C	1.5 MPa	11.31
250 °C	4.0 MPa	11.20
300 °C	8.7 MPa	11.34
350 °C	17 MPa	11.92

With electrolyte solutions, the value of pK_w is dependent on ionic strength of the electrolyte. Values for sodium chloride are typical for a 1:1 electrolyte. With 1:2 electrolytes, MX₂, pK_w decreases with increasing ionic strength. The value of K_w is usually of interest in the liquid phase. Example values for superheated steam (gas) and supercritical water fluid are given in the table.

Comparison of pK_w values for liquid water, superheated steam, and supercritical water.						
Temp. Pressure	350 °C	400 °C	450 °C	500 °C	600 °C	800 °C
0.1 MPa		47.961 ^b	47.873 ^b	47.638 ^b	46.384 ^b	40.785 ^b
17 MPa	11.920 (liquid) ^a					
25 MPa	11.551 (liquid) ^c	16.566	18.135	18.758	19.425	20.113
100 MPa	10.600 (liquid) ^c	10.744	11.005	11.381	12.296	13.544
1000 MPa	8.311 (liquid) ^c	8.178	8.084	8.019	7.952	7.957

Notes to the table. The values are for supercritical fluid except those marked:

- ^a at saturation pressure corresponding to 350 °C.
- ^b superheated steam. ^c compressed or subcooled liquid.

Relationship with the Neutral Point of Water

Water molecules dissociate into equal amounts of H₃O⁺ and OH⁻, so their concentrations are equal to 1.00×10⁻⁷ mol dm⁻³ at 25 °C. A solution in which the H₃O⁺ and OH⁻ concentrations equal each other is considered a neutral solution. In general, the pH of the neutral point is numerically equal to 1/2 pK_w. Pure water is neutral, but most water samples contain impurities. If an impurity is an acid or base, this will affect the concentrations of hydronium ion and hydroxide ion. Water samples that are exposed to air will absorb some acidic carbon dioxide to form carbonic acid (H₂CO₃) and the concentration of H₃O⁺ will increase due to the reaction H₂CO₃ + H₂O = HCO₃⁻ + H₃O⁺. The concentration of OH⁻ will decrease in such a way that the product [H₃O⁺][OH⁻] remains constant for fixed temperature and pressure. Thus these water samples will be slightly acidic. If a pH of exactly 7.0 is required, it must be maintained with an appropriate buffer solution.

PH SCALE

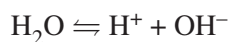
Acidic and basic are two extremes that describe a chemical property chemicals. Mixing acids and bases can cancel out or neutralize their extreme effects. A substance that is neither acidic nor basic is neutral. The pH scale measures how acidic or basic a substance is. The pH scale ranges from 0 to 14. A pH of 7 is neutral. A pH less than 7 is acidic. A pH greater than 7 is basic.

The pH scale is logarithmic and as a result, each whole pH value below 7 is ten times more acidic than the next higher value. For example, pH 4 is ten times more acidic than pH 5 and 100 times (10 times 10) more acidic than pH 6. The same holds true for pH values above 7, each of which is ten times more alkaline (another

way to say basic) than the next lower whole value. For example, pH 10 is ten times more alkaline than pH 9 and 100 times (10 times 10) more alkaline than pH 8. Pure water is neutral. But when chemicals are mixed with water, the mixture can become either acidic or basic. Examples of acidic substances are vinegar and lemon juice. Lye, milk of magnesia, and ammonia are examples of basic substances.

Ionization of Water

Water molecules exist in equilibrium with hydrogen ions and hydroxide ions.



The water equilibrium constant is written as:

$$K_w = [\text{H}^+][\text{OH}^-]$$

Experimentally, it has been found that the concentration of:

$$\text{H}^+ = \text{OH}^- = 10^{-7}$$

Therefore:

$$K_w = [10^{-7}][10^{-7}] = [10^{-14}]$$

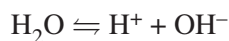
(To multiply exponential numbers—simply add the exponents.)

The values for K_w , H^+ , OH^- concentration all indicate that the equilibrium favors the reactant (water molecules). In other words, only very small amounts of H^+ and OH^- ions are present.

Effect of Acids and Bases on Water Equilibrium

If an acid (H^+) is added to the water, the equilibrium shifts to the left and the OH^- ion concentration decreases.

- *Water Equilibrium:*



If base (OH^-) is added to water, the equilibrium shifts to left and the H^+ concentration decreases.

- *Water Equilibrium Principle:* The multiplication product (addition of exponents) of H^+ and OH^- ion concentration must always be equal to 10^{-14} .

BOTH H^+ and OH^- ions are ALWAYS PRESENT in any solution. A solution is acidic if the H^+ are in excess. A solution is basic, if the OH^- ions are in excess.

Definition of PH, POH, and PKw

The concentrations of hydrogen ions and indirectly hydroxide ions are given by a pH number. pH is defined as the negative logarithm of the hydrogen ion concentration.

The equation is:

$$\text{pH} = -\log [\text{H}^+]$$

similarly,

$$\text{pOH} = -\log [\text{OH}^-] \text{ and} \\ \text{p } K_w = -\log [K_w].$$

Logarithms of numbers that are multiples of ten are merely the exponents of the number including the sign.

pH Scale

The pH scale, (0-14), is the full set of pH numbers which indicate the concentration of H^+ and OH^- ions in water. The diagram on the left gives some relationships which summarizes much of the previous discussion.

pH Scale Principle:

- H^+ ion concentration and pH relate inversely.
- OH^- ion concentration and pH relate directly.

The following statements may be made about the pH scale numbers. Complete some of them.

- Increasing pH means the H^+ ions are decreasing.
- Decreasing pH means H^+ ions are increasing.
- Increasing pH means OH^- ions are
- Decreasing pH means OH^- ions are

BUFFERS

A buffer is a solution that can resist pH change upon the addition of an acidic or basic components. It is able to neutralize small amounts of added acid or base, thus maintaining the pH of the solution relatively stable. This is important for processes and/or reactions which require specific and stable pH ranges. Buffer solutions have a working pH range and capacity which dictate how much acid/base can be neutralized before pH changes, and the amount by which it will change.

HENDERSON–HASSELBALCH EQUATION

In chemistry, the Henderson–Hasselbalch equation describes the derivation of pH as a measure of acidity (using pK_a , the negative log of the acid dissociation constant) in biological and chemical systems. The equation is also useful for estimating the pH of a buffer solution and finding the equilibrium pH in acid-base reactions (it is widely used to calculate the isoelectric point of proteins).

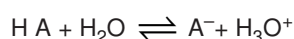
The equation is given by:

$$pH = pK_a + \log_{10} \left(\frac{[A^-]}{[HA]} \right)$$

Here, $[HA]$ is the molar concentration of the undissociated weak acid, $[A^-]$ is the molar concentration (molarity, M) of this acid's conjugate base and pK_a is $-\log_{10} K_a$ where K_a is the acid dissociation constant, that is:

$$pK_a = -\log_{10} (K_a) = -\log_{10} \left(\frac{[H_3O^+][A^-]}{[HA]} \right)$$

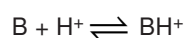
for the non-specific Brønsted acid-base reaction:



In these equations, A^- denotes the ionic form of the relevant acid. Bracketed quantities such as $[base]$ and $[acid]$ denote the molar concentration of the quantity enclosed.

For Bases

For the standard base equation:



A second form of the equation expressed in terms of K_b where K_b is the base dissociation constant:

$$pK_b = -\log_{10} (K_b) = -\log_{10} \left(\frac{[OH^-][HA]}{[A^-]} \right)$$

In analogy to the above equations, the following equation is valid:

$$\text{pOH} = \text{pK}_b + \log_{10} \left(\frac{[\text{BH}^+]}{[\text{B}]} \right)$$

where BH^+ denotes the conjugate acid of the corresponding base B. Using the properties of these terms at 25 degrees Celsius, one can synthesise an equation for pH of basic solutions in terms of pK_a and pH:

$$\text{pH} = \text{pK}_a + \log_{10} \left(\frac{[\text{B}]}{[\text{BH}^+]} \right)$$

Limitations

There are some significant approximations implicit in the Henderson–Hasselbalch equation. The most significant is the assumption that the concentration of the acid and its conjugate base at equilibrium will remain the same as the formal concentration. This neglects the dissociation of the acid and the binding of H^+ to the base. The dissociation of water and relative water concentration itself is neglected as well.

These approximations will fail when dealing with relatively strong acids or bases (pK_a more than a couple units away from 7), dilute or very concentrated solutions (less than 1 mM or greater than 1M), or heavily skewed acid/base ratios (more than 100 to 1). In high buffer dilutions, where the concentration of protons arising from water become equally or more prevalent than the buffer species themselves (at pH 7, this means buffer component concentrations of $<10^{-5}$ M formally, but practically much higher), the pK_a of the “buffer” system will tend towards neutrality.

Conventions

For bases, it is conventional to report pK_a of the conjugate acid; conversely, for acids, it is conventional to report pK_b of the conjugate base. Consequently, a base with a low pK_a has a conjugate acid that readily dissociates and therefore that the basic compound is likely to be neutral at physiological pH. This can be contrasted with a low pK_a for an acid, which means that the acid readily dissociates and the molecule is therefore likely to exist as its conjugate base at physiological pH. To interpret literature pK_a and pK_b values an inspection of the molecule to determine whether it contains acid or base functionality (or both) is important.

The Henderson–Hasselbalch equation can also be rearranged to calculate the net charge of an amino acid at a certain pH. Written in terms of the degree of dissociation α , the Henderson–Hasselbalch equation is:

$$\text{pH} = \text{pK}_a + \log_{10} \left(\frac{\alpha}{1-\alpha} \right)$$

solving for α yields,

$$\alpha = \frac{1}{10^{\text{pH} - \text{pK}_a} + 1}$$

This equation above can be used to calculate the degree of dissociation for each ionizable group at the desired pH. For ionizable groups that are able to deprotonate to a charge of -1 (e.g., carboxyl or sulfhydryl groups), multiply the calculated dissociation constant by -1. For ionizable groups that are able to deprotonate to a charge of 0, take the complement of the dissociation constant ($1-\alpha$) and multiply the constant by +1. The net charge of the amino acid will be the sum of the charges of all of the ionizable groups.

BUFFER SYSTEMS OF BLOOD

- Venous blood carries more CO_2 than arte-rial blood. Hence, the pH of venous blood is more acid than that of arterial blood by 0.01-0.03 units *i.e.*, pH 7.40 and 7.43, respectively.
- The blood buffers consists of the plasma proteins, hemoglobin, oxyhemoglobin, bicarbonates and inorganic phosphates.
- When CO_2 enters the venous blood, the small decrease in pH shifts the ratio of acid to salt in all the buffer pairs. When the ratio is shifted to form more of the acid, cations become available to form additional bicarbonates. In this respect, the plasma phosphates and bicarbonates play a minor role.
- The buffering action of the plasma proteins is important because they release sufficient cations for the carriage of about 10% of the total CO_2 .
- The phosphates in the red cells carry 25% of the total CO_2 .
- The most important buffering role is of hemoglobin and oxyhemoglobin which carry 60% of the CO_2 of the whole blood.

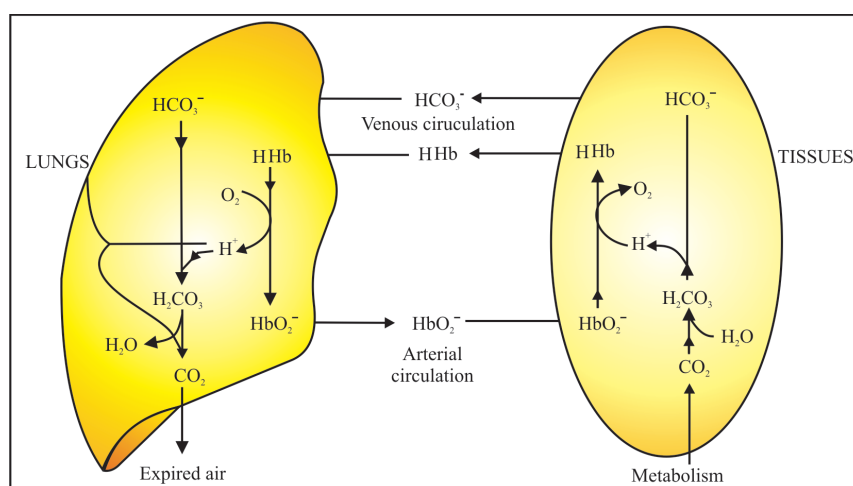


Fig. The buffering action of hemoglobin.

Hemoglobin Buffers

At the lungs the formation of oxyhemoglobin from reduced hemoglobin releases hydrogen ions which react with bicarbonate to form carbonic acid. The low CO_2 tension in the lung shifts the equilibrium towards the production of CO_2 which is continually eliminated in the expired air:



In the tissues, the oxygen tension is reduced and hence, oxy-hemoglobin dissociates delivering O_2 to the cells and reduced hemoglobin is formed. CO_2 produced by metabolism enters the blood, where it is hydrated to form H_2CO_3 which ionizes to form H^+ and HCO_3^- . Reduced hemoglobin acting as an anion accepts the H^+ ions forming acid-reduced hemoglobin (HHb). Very little change in pH occurs because the newly arrived H^+ ions are buffered by formation of very weak acid. When the blood returns to the lungs, these H^+ ions are released as a result of the formation of stronger acid (oxyhemoglobin) and the newly released H^+ ion is promptly neutralized by HCO_3^- . This reaction is necessary for the liberation of CO_2 in the lungs.

Chloride Shift

- CO_2 reacts with water to form carbonic acid (H_2CO_3), mainly inside the red cell by the enzyme carbonic anhydrase present in the red cells.

- The carbonic acid is then buffered by the intracellular buffers (Phosphate and hemoglobin) combining with potassium.
- Bicarbonate ion also returns to the plasma and exchanges with chloride which shifts into the cell when the tension of CO_2 increases in the blood.
- When the CO_2 tension is reduced, chloride leaves the cell and enters the plasma.
- Under normal conditions the red cell is impermeable to sodium or potassium. But it is permeable to hydrogen, bicarbonate and chloride ions and intracellular sources of cation (potassium) are indirectly available to the plasma by chloride (anion) exchange. This permits the carriage of additional CO_2 (as sodium bicarbonate) by plasma.
- The CO_2 entering the blood from the tissues passes into the red cells where it forms carbonic acid by carbonic anhydrase. Some of the carbonic acid returns to the plasma. The remainder reacts with hemoglobin buffers to form bicarbonate which then returns to the plasma in exchange of chloride. The chloride is neutralized by potassium in the red cells.
- All of these reactions are reversible. At the lung, when the blood becomes arterial, chloride shifts back into the plasma, thus liberating intracellular potassium to buffer the oxyhemoglobin and in the plasma, neutralizing the sodium which is liberated by the removal of CO_2 during respiration.

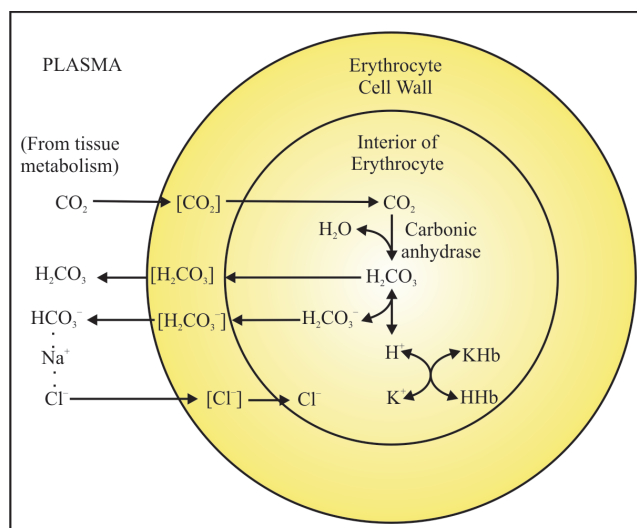


Fig. Chloride shift.

EXCHANGE OF O_2 AND CO_2

The primary function of the respiratory system is to exchange oxygen and carbon dioxide. Inhaled oxygen enters the lungs and reaches the alveoli. The layers of cells lining the alveoli and the surrounding capillaries are each only one cell thick and are in very close contact with each other. This barrier between air and blood averages about 1 micron ($\frac{1}{10,000}$ of a centimeter, or 0.000039 inch) in thickness. Oxygen passes quickly through this air-blood barrier into the blood in the capillaries. Similarly, carbon dioxide passes from the blood into the alveoli and is then exhaled. Oxygenated blood travels from the lungs through the pulmonary veins and into the left side of the heart, which pumps the blood to the rest of the body. Oxygen-deficient, carbon dioxide-rich blood returns to the right side of the heart through two large veins, the superior vena cava and the inferior vena cava. Then the blood is pumped through the pulmonary artery to the lungs, where it picks up oxygen and releases carbon dioxide.

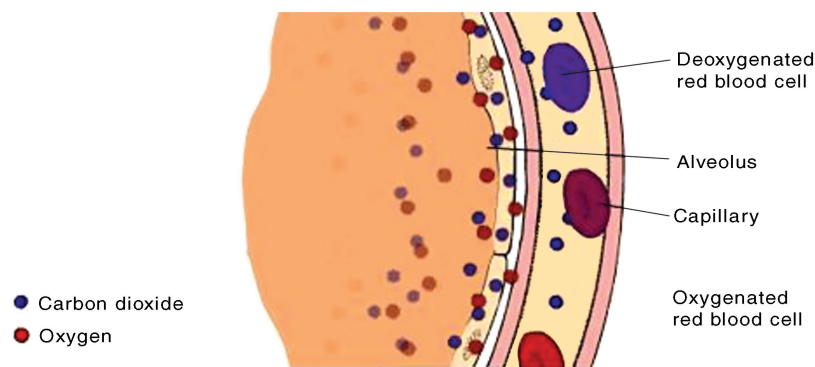


Fig. Gas Exchange Between Alveoli and Capillaries.

To support the exchange of oxygen and carbon dioxide, about 5 to 8 liters (about 1.3 to 2.1 gallons) of air per minute are brought in and out of the lungs, and about three tenths of a liter of oxygen is transferred from the alveoli to the blood each minute, even when the person is at rest. At the same time, a similar volume of carbon dioxide moves from the blood to the alveoli and is exhaled. During exercise, it is possible to breathe in and out more than 100 liters (about 26 gallons) of air per minute and extract 3 liters (a little less than 1 gallon) of oxygen from this air per minute. The rate at which oxygen is used by the body is one measure of the rate of energy expended by the body. Breathing in and out is accomplished by respiratory muscles.

Three processes are essential for the transfer of oxygen from the outside air to the blood flowing through the lungs: ventilation, diffusion, and perfusion.

- Ventilation is the process by which air moves in and out of the lungs.
- Diffusion is the spontaneous movement of gases, without the use of any energy or effort by the body, between the gas in the alveoli and the blood in the capillaries in the lungs.
- Perfusion is the process by which the cardiovascular system pumps blood throughout the lungs.

The body's circulation is an essential link between the atmosphere, which contains oxygen, and the cells of the body, which consume oxygen. For example, the delivery of oxygen to the muscle cells throughout the body depends not only on the lungs but also on the ability of the blood to carry oxygen and on the ability of the circulation to transport blood to muscle.

BLOOD BUFFER CALCULATIONS

Many people today are interested in exercise as a way of improving their health and physical abilities. But there is also concern that too much exercise, or exercise that is not appropriate for certain individuals, may actually do more harm than good. Exercise has many short-term (acute) and long-term effects that the body must be capable of handling for the exercise to be beneficial. Some of the major acute effects of exercising are shown in Figure.

When we exercise, our heart rate, systolic blood pressure, and cardiac output (the amount of blood pumped per heart beat) all increase. Blood flow to the heart, the muscles, and the skin increase. The body's metabolism becomes more active, producing CO_2 and H^+ in the muscles. We breathe faster and deeper to supply the oxygen required by this increased metabolism. Eventually, with strenuous exercise, our body's metabolism exceeds the oxygen supply and begins to use alternate biochemical processes that do not require oxygen. These processes generate lactic acid, which enters the blood stream. As we develop a long-term habit of exercise, our cardiac output and lung capacity increase, even when we are at rest, so that we can exercise longer and harder than before. Over time, the amount of muscle in the body increases, and fat is burned as its energy is needed to help fuel the body's increased metabolism.



Fig. This figure highlights some of the major acute (short-term) effects on the body during exercise.

Chemical Changes in the Blood During Exercise

During exercise, the muscles use up oxygen as they convert chemical energy in glucose to mechanical energy. This O₂ comes from hemoglobin in the blood. CO₂ and H⁺ are produced during the breakdown of glucose, and are removed from the muscle via the blood. The production and removal of CO₂ and H⁺, together with the use and transport of O₂, cause chemical changes in the blood. These chemical changes, unless offset by other physiological functions, cause the pH of the blood to drop. If the pH of the body gets too low (below 7.4), a condition known as acidosis results. This can be very serious, because many of the chemical reactions that occur in the body, especially those involving proteins, are pH-dependent. Ideally, the pH of the blood should be maintained at 7.4. If the pH drops below 6.8 or rises above 7.8, death may occur. Fortunately, we have buffers in the blood to protect against large changes in pH.

How Chemicals are Exchanged in the Body

All cells in the body continually exchange chemicals (*e.g.*, nutrients, waste products, and ions) with the external fluid surrounding them. This external fluid, in turn, exchanges chemicals with the blood being pumped throughout the body. A dominant mode of exchange between these fluids (cellular fluid, external fluid, and blood) is diffusion through membrane channels, due to a concentration gradient associated with the contents of the fluids. (Recall your experience with concentration gradients in the “Membranes, Proteins, and Dialysis” experiment.)

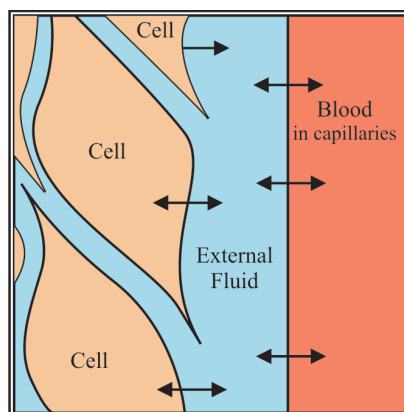


Fig. This is a schematic diagram showing the flow of species across membranes between the cells, the extracellular fluid, and the blood in the capillaries.

Hence, the chemical composition of the blood (and therefore of the external fluid) is extremely important for the cell. If, for instance, the pH of the blood and external fluid is too low (too many H^+ ions), then an excess of H^+ ions will enter the cell. As mentioned above, maintaining the proper pH is critical for the chemical reactions that occur in the body. In order to maintain the proper chemical composition inside the cells, the chemical composition of the fluids outside the cells must be kept relatively constant. This constancy is known in biology as homeostasis.

The body has a wide array of mechanisms to maintain homeostasis in the blood and extracellular fluid. The most important way that the pH of the blood is kept relatively constant is by buffers dissolved in the blood. Other organs help enhance the homeostatic function of the buffers. The kidneys help remove excess chemicals from the blood, as discussed in the Kidney Dialysis tutorial. It is the kidneys that ultimately remove (from the body) H^+ ions and other components of the pH buffers that build up in excess. Acidosis that results from failure of the kidneys to perform this excretory function is known as metabolic acidosis. However, excretion by the kidneys is a relatively slow process, and may take too long to prevent acute acidosis resulting from a sudden decrease in pH (*e.g.*, during exercise). The lungs provide a faster way to help control the pH of the blood. The increased-breathing response to exercise helps to counteract the pH-lowering effects of exercise by removing CO_2 , a component of the principal pH buffer in the blood. Acidosis that results from failure of the lungs to eliminate CO_2 as fast as it is produced is known as respiratory acidosis.

How Buffers Work: A Quantitative View

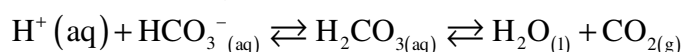
The kidneys and the lungs work together to help maintain a blood pH of 7.4 by affecting the components of the buffers in the blood. Therefore, to understand how these organs help control the pH of the blood, we must first discuss how buffers work in solution. Acid-base buffers confer resistance to a change in the pH of a solution when hydrogen ions (protons) or hydroxide ions are added or removed. An acid-base buffer typically consists of a weak acid, and its conjugate base (salt). Buffers work because the concentrations of the weak acid and its salt are large compared to the amount of protons or hydroxide ions added or removed.

When protons are added to the solution from an external source, some of the base component of the buffer is converted to the weak-acid component (thus using up most of the protons added); when hydroxide ions are added to the solution (or, equivalently, protons are removed from the solution;), protons are dissociated from some of the weak-acid molecules of the buffer, converting them to the base of the buffer (and thus replenishing most of the protons removed). However, the change in acid and base concentrations is small relative to the amounts of these species present in solution. Hence, the ratio of acid to base changes only slightly. Thus, the effect on the pH of the solution is small, within certain limitations on the amount of H^+ or OH^- added or removed.

The Carbonic-Acid-Bicarbonate Buffer in the Blood

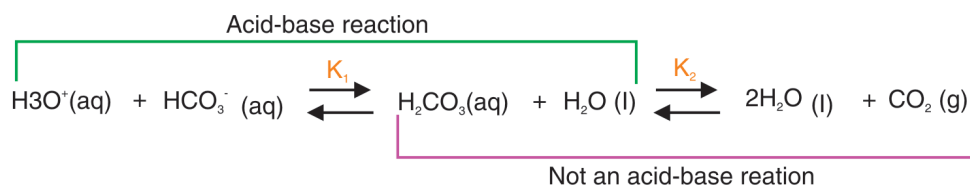
By far the most important buffer for maintaining acid-base balance in the blood is the carbonic-acid-bicarbonate buffer.

The simultaneous equilibrium reactions of interest are:



We are interested in the change in the pH of the blood; therefore, we want an expression for the concentration of H^+ in terms of an equilibrium constant and the concentrations of the other species in the reaction (HCO_3^- , H_2CO_3 , and CO_2).

To more clearly show the two equilibrium reactions in the carbonic-acid-bicarbonate buffer, Equation 1 is rewritten to show the direct involvement of water:



The equilibrium on the left is an acid-base reaction that is written in the reverse format from Equation. Carbonic acid (H_2CO_3) is the acid and water is the base. The conjugate base for H_2CO_3 is HCO_3^- (bicarbonate ion). Carbonic acid also dissociates rapidly to produce water and carbon dioxide, as shown in the equilibrium on the right of Equation. This second process is not an acid-base reaction, but it is important to the blood's buffering capacity.

$$\text{pH} = \text{pK} - \log \left(\frac{[\text{CO}_2]}{[\text{HCO}_3^-]} \right).$$

The derivation for this equation is shown in the yellow box, below. Notice that Equation 11 is in a similar form to the Henderson-Hasselbach equation presented in the introduction to the Experiment. Equation does not meet the strict definition of a Henderson-Hasselbach equation, because this equation takes into account a non-acid-base reaction (*i.e.*, the dissociation of carbonic acid to carbon dioxide and water), and the ratio in parentheses is not the concentration ratio of the acid to the conjugate base. However, the relationship shown in Equation is frequently referred to as the Henderson-Hasselbach equation for the buffer in physiological applications.

In Equation pK is equal to the negative log of the equilibrium constant, K , for the buffer.

$$\text{pK} = -\log K$$

where,

$$K = K_a/K_2.$$

This quantity provides an indication of the degree to which HCO_3^- reacts with H^+ (or with H_3O^+ as written in Equation) to form H_2CO_3 , and subsequently to form CO_2 and H_2O . In the case of the carbonic-acid-bicarbonate buffer, $\text{pK}=6.1$ at normal body temperature. As shown in Equation, the pH of the buffered solution (*i.e.*, the blood) is dependent only on the ratio of the amount of CO_2 present in the blood to the amount of HCO_3^- (bicarbonate ion) present in the blood (at a given temperature, so that pK remains constant). This ratio remains relatively constant, because the concentrations of both buffer components (HCO_3^- and CO_2) are very large, compared to the amount of H^+ added to the blood during normal activities and moderate exercise. When H^+ is added to the blood as a result of metabolic processes, the amount of HCO_3^- (relative to the amount of CO_2) decreases; however, the amount of the change is tiny compared to the amount of HCO_3^- present in the blood.

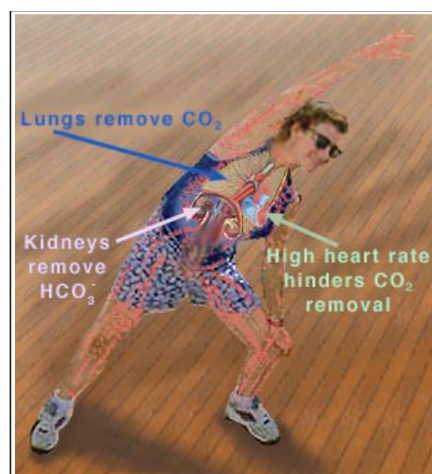


Fig. This figure shows the major organs that help control the blood concentrations of CO_2 and HCO_3^- , and thus help control the pH of the blood. Removing CO_2 from the blood helps increase the pH. Removing HCO_3^- from the blood helps lower the pH.

This optimal buffering occurs when the pH is within approximately 1 pH unit from the pK value for the buffering system, *i.e.*, when the pH is between 5.1 and 7.1. However, the normal blood pH of 7.4 is outside the optimal buffering range; therefore, the addition of protons to the blood due to strenuous exercise may be too great for the buffer alone to effectively control the pH of the blood. When this happens, other organs must help control the amounts of CO_2 and HCO_3^- in the blood. The lungs remove excess CO_2 from the blood, and the kidneys remove excess HCO_3^- from the body (helping to lower the pH). The lungs' removal of CO_2 from the blood is somewhat impeded during exercise when the heart rate is very rapid; the blood is pumped through the capillaries very quickly, and so there is little time in the lungs for carbon dioxide to be exchanged for oxygen. The ways in which these three organs help to control the blood pH through the bicarbonate buffer system are highlighted in Figure, above.

Why the Optimal Buffering Capacity is at $\text{pH}=\text{pK}$

Why is the buffering capacity of the carbonic-acid-bicarbonate buffer highest when the pH is close to the pK value, but lower at normal blood pH? The answer to this question lies in the shape of the titration curve for the buffer, which is shown in Figure, below.

Titration Curve for the Bicarbonate-Buffer System

It is possible to plot a titration curve for this buffer system, just as you did for your solution in the acid-base-equilibria experiment. In this plot, the vertical axis shows the pH of the buffered solution (in this case, the blood). The horizontal axis shows the composition of the buffer: on the left-hand side of the plot, most of the buffer is in the form of carbonic acid or carbon dioxide, and on the right-hand side of the plot, most of the buffer is in the form of bicarbonate ion. Note that as acid is added, the pH decreases and the buffer shifts towards greater H_2CO_3 and CO_2 concentration. Conversely, as base is added, the pH increases and the buffer shifts towards greater HCO_3^- concentration.

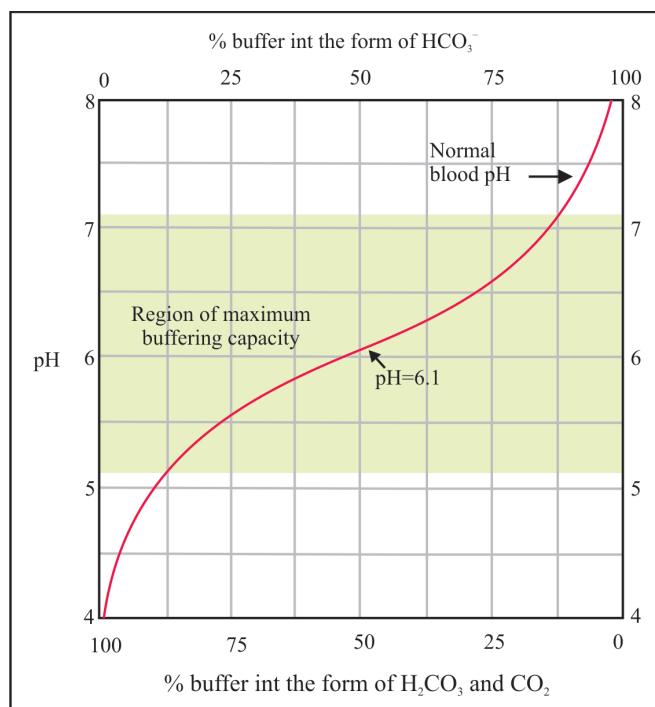


Fig. This is the titration curve for the carbonic-acid-bicarbonate buffer. Note that the pH of the blood (7.4) lies outside the region of greatest buffering capacity (green).

Note: The percent buffer in the form of HCO_3^- is given by the formula:

$$\% \text{ buffer in the form of } \text{HCO}_3^- = \frac{[\text{HCO}_3^-]}{[\text{HCO}_3^-] + [\text{H}_2\text{CO}_3] + [\text{CO}_3]} \times 100$$

The slope of the curve is flattest where the pH is equal to the pK value (6.1) for the buffer. Here, the buffering capacity is greatest because a shift in the relative concentrations of bicarbonate and carbon dioxide produces only a small change in the pH of the solution. However, at pH values higher than 7.1, the slope of the curve is much higher. Here, a shift in the relative concentrations of bicarbonate and carbon dioxide produces a large change in the pH of the solution. Hence, at the physiological blood pH of 7.4, other organs must help to control the amounts of HCO_3^- and CO_2 in the blood to keep the pH relatively constant, as described above.

Other PH-Buffer Systems in the Blood

Other buffers perform a more minor role than the carbonic-acid-bicarbonate buffer in regulating the pH of the blood. The phosphate buffer consists of phosphoric acid (H_3PO_4) in equilibrium with dihydrogen phosphate ion (H_2PO_4^-) and H^+ . The pK for the phosphate buffer is 6.8, which allows this buffer to function within its optimal buffering range at physiological pH.

The phosphate buffer only plays a minor role in the blood, however, because H_3PO_4 and H_2PO_4^- are found in very low concentration in the blood. Hemoglobin also acts as a pH buffer in the blood. Recall from the “Hemoglobin” tutorial from Chem 151 that hemoglobin protein can reversibly bind either H^+ (to the protein) or O_2 (to the Fe of the heme group), but that when one of these substances is bound, the other is released (as explained by the Bohr effect). During exercise, hemoglobin helps to control the pH of the blood by binding some of the excess protons that are generated in the muscles. At the same time, molecular oxygen is released for use by the muscles.

Equilibrium Shifts: A Qualitative View

The above discussion treats changes in pH quantitatively. These changes in pH can be described qualitatively, also. A qualitative view is very useful for predicting how the pH will change in response to external conditions (such as exercise). The principle used for this qualitative view is known as Le Châtelier’s Principle.

Le Chatelier’s Principle

When a reactant or product of an equilibrium reaction is added to a solution that is at equilibrium, the added species will react to change the concentrations of the reactants and products in the solution until a new equilibrium is established (but the ratio of the concentrations given in the mass action expression is the same, because the equilibrium constant, K , is a constant at a given temperature). This process is known as a shift in the equilibrium. In 1884, Henri Le Châtelier developed a rule to predict how a system in equilibrium will shift when the conditions of the system are changed.

Although this rule can oversimplify the changes that occur in certain situations, it is a powerful and useful tool for predicting the *direction* of an equilibrium shift. Le Châtelier’s Principle states that “if a change in conditions ([an external] ‘stress’) is imposed on a system at equilibrium, the equilibrium position will shift in a direction that tends to reduce that change in conditions” (Zumdahl, 208). For example, if the concentration of one of the products of an equilibrium reaction is increased in a solution that was at equilibrium (prior to the concentration increase), the equilibrium will shift so as to reduce the concentration of the product, *i.e.*, more

reactant will be generated. Of course, the reverse equilibrium shift would occur when the concentration of a product is decreased.

External Stress	Predicted Equilibrium Shift
Change in pressure	Affects gases only.
Concentration of reactant(s) increased	Equilibrium shifts toward products (to the right).
Concentration of product(s) increased	Equilibrium shifts toward reactants (to the left).
Temperature increased (exothermic reaction)	Equilibrium shifts toward reactants (to the left).
Temperature increased (endothermic reaction)	Equilibrium shifts toward products (to the right).

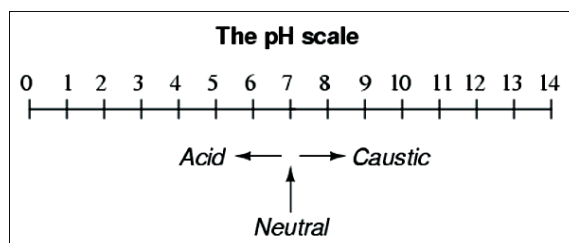
The effect of a temperature change can also be predicted by Le Châtelier's Principle. (In the case of a change in temperature, the equilibrium constant actually changes.) If a reaction is exothermic, "heat" is treated as a "product" (e.g., $A + B \rightarrow C + D + \text{"heat"}$). If a reaction is endothermic, "heat" is treated as a "reactant" (e.g., $\text{"heat"} + A + B \rightarrow C + D$). Increasing the temperature can be thought of as increasing the amount of "heat" in the reaction. Examples of external stresses and the equilibrium shifts predicted by Le Châtelier's Principle are shown in the purple table below.

Example: Removal of Bicarbonate by the Kidneys

Le Châtelier's Principle can be used to explain how the kidneys help prevent excessively high pH (a condition known as alkalosis). When the pH of the blood is too high, the kidneys remove bicarbonate ion (HCO_3^-) from the blood. As the kidneys decrease the blood's concentration of HCO_3^- , the equilibrium reaction in Equation is shifted towards the left to compensate for the loss in HCO_3^- , according to Le Châtelier's Principle. When the equilibrium shifts to the left, more H^+ ions are generated together with HCO_3^- ions. As a result, the pH decreases.

MEASUREMENT OF PH

A very important measurement in many liquid chemical processes (industrial, pharmaceutical, manufacturing, food production, *etc.*) is that of pH: the measurement of hydrogen ion concentration in a liquid solution. A solution with a low pH value is called an "acid," while one with a high pH is called a "caustic." The common pH scale extends from 0 (strong acid) to 14 (strong caustic), with 7 in the middle representing pure water (neutral):



pH is defined as follows: The lower-case letter "p" in pH stands for the negative common (base ten) logarithm, while the upper-case letter "H" stands for the element hydrogen. Thus, pH is a logarithmic measurement of the number of moles of hydrogen ions (H^+) per liter of solution. Incidentally, the "p" prefix is also used with other types of chemical measurements where a logarithmic scale is desired, pCO_2 (Carbon Dioxide) and pO_2 (Oxygen) being two such examples.

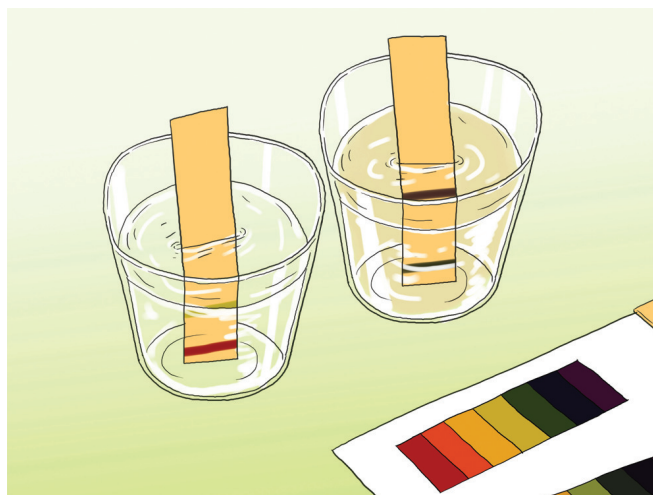
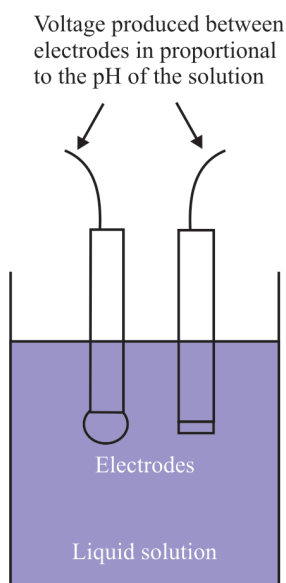


Fig. Ways to measure the pH of water.

The logarithmic pH scale works like this: a solution with 10^{-12} moles of H^+ ions per liter has a pH of 12; a solution with 10^{-3} moles of H^+ ions per liter has a pH of 3. While very uncommon, there is such a thing as an acid with a pH measurement below 0 and a caustic with a pH above 14. Such solutions, understandably, are quite concentrated and *extremely* reactive.

While pH can be measured by color changes in certain chemical powders (the “litmus strip” being a familiar example from high school chemistry classes), continuous process monitoring and control of pH requires a more sophisticated approach. The most common approach is the use of a specially-prepared electrode designed to allow hydrogen ions in the solution to migrate through a selective barrier, producing a measurable potential (voltage) difference proportional to the solution’s pH: The design and operational theory of pH electrodes is a very complex subject, explored only briefly here. What is important to understand is that these two electrodes generate a voltage directly proportional to the pH of the solution. At a pH of 7 (neutral), the electrodes will produce 0 volts between them. At a low pH (acid) a voltage will be developed of one polarity, and at a high pH (caustic) a voltage will be developed of the opposite polarity.

An unfortunate design constraint of pH electrodes is that one of them (called the *measurement* electrode) must be constructed of special glass to create the ion-selective barrier needed to screen out hydrogen ions from all the other ions floating around in the solution.

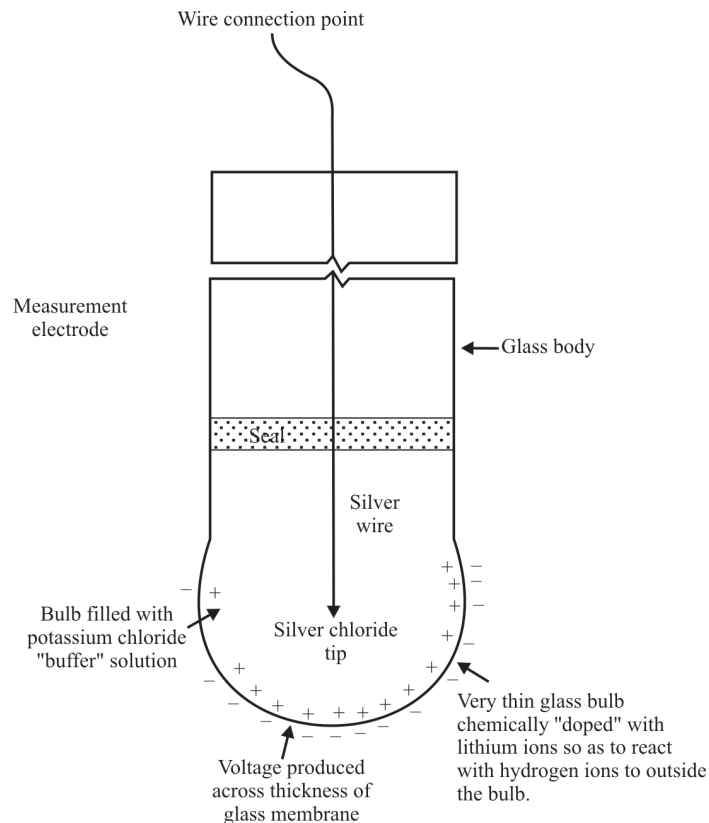


This glass is chemically doped with lithium ions, which is what makes it react electrochemically to hydrogen ions. Of course, glass is not exactly what you would call a “conductor;” rather, it is an extremely good insulator. This presents a major problem if our intent is to measure voltage between the two electrodes. The circuit path from one electrode contact, through the glass barrier, through the solution, to the other electrode, and back through the other electrode’s contact, is one of *extremely* high resistance.

The other electrode (called the *reference* electrode) is made from a chemical solution of neutral (7) pH buffer solution (usually potassium chloride) allowed to exchange ions with the process solution through a porous separator, forming a relatively low resistance connection to the test liquid. At first, one might be inclined to ask: why not just dip a metal wire into the solution to get an electrical connection to the liquid? The reason this will not work is because metals tend to be highly reactive in ionic solutions and can produce a significant voltage across the interface of metal-to-liquid contact. The use of a wet chemical interface with the measured solution is necessary to avoid creating such a voltage, which of course would be falsely interpreted by any measuring device as being indicative of pH.

Here is an illustration of the measurement electrode’s construction.

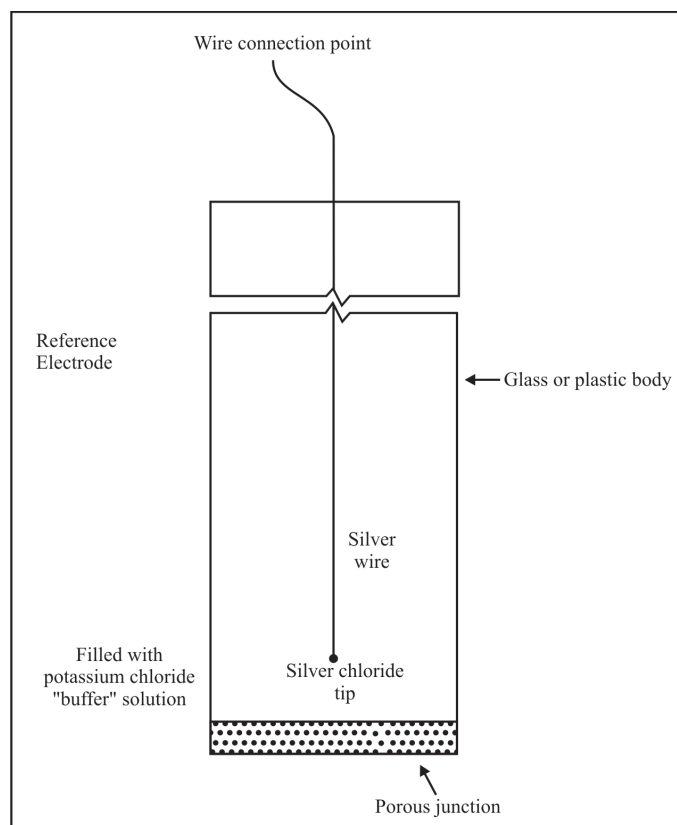
Note the thin, lithium-doped glass membrane across which the pH voltage is generated:



Here is an illustration of the reference electrode’s construction.

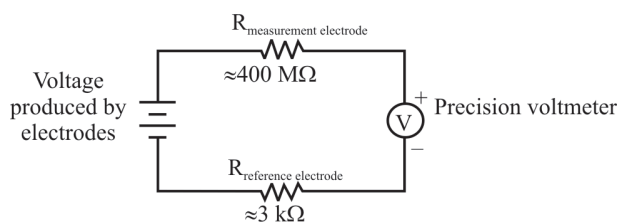
The porous junction shown at the bottom of the electrode is where the potassium chloride buffer and process liquid interface with each other:

The measurement electrode’s purpose is to generate the voltage used to measure the solution’s pH. This voltage appears across the thickness of the glass, placing the silver wire on one side of the voltage and the liquid solution on the other. The reference electrode’s purpose is to provide the stable, zero-voltage connection to the liquid solution so that a complete circuit can be made to measure the glass electrode’s voltage. While the reference electrode’s connection to the test liquid may only be a few kilo-ohms, the glass electrode’s resistance may range from ten to nine hundred mega-ohms, depending on electrode design!



Being that any current in this circuit must travel through *both* electrodes' resistances (and the resistance presented by the test liquid itself), these resistances are in series with each other and therefore add to make an even greater total. An ordinary analog or even digital voltmeter has much too low of an internal resistance to measure voltage in such a high-resistance circuit.

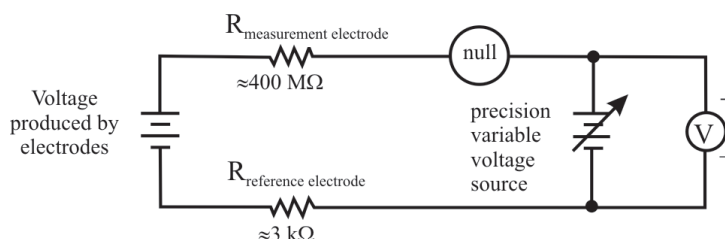
The equivalent circuit diagram of a typical pH probe circuit illustrates the problem:



Even a very small circuit current traveling through the high resistances of each component in the circuit (especially the measurement electrode's glass membrane), will produce relatively substantial voltage drops across those resistances, seriously reducing the voltage seen by the meter. Making matters worse is the fact that the voltage differential generated by the measurement electrode is very small, in the millivolt range (ideally 59.16 millivolts per pH unit at room temperature). The meter used for this task must be very sensitive and have an extremely high input resistance.

The most common solution to this measurement problem is to use an amplified meter with an extremely high internal resistance to measure the electrode voltage, so as to draw as little current through the circuit as possible. With modern semiconductor components, a voltmeter with an input resistance of up to $10^{17} \Omega$ can be built with little difficulty. Another approach, seldom seen in contemporary use, is to use a potentiometric "null-balance" voltage measurement setup to measure this voltage without drawing *any* current from the circuit under test.

If a technician desired to check the voltage output between a pair of pH electrodes, this would probably be the most practical means of doing so using only standard benchtop metering equipment:



As usual, the precision voltage supply would be adjusted by the technician until the null detector registered zero, then the voltmeter connected in parallel with the supply would be viewed to obtain a voltage reading. With the detector “nulled” (registering exactly zero), there should be zero current in the pH electrode circuit, and therefore no voltage dropped across the resistances of either electrode, giving the real electrode voltage at the voltmeter terminals. Wiring requirements for pH electrodes tend to be even more severe than thermocouple wiring, demanding very clean connections and short distances of wire (10 yards or less, even with gold-plated contacts and shielded cable) for accurate and reliable measurement. As with thermocouples, however, the disadvantages of electrode pH measurement are offset by the advantages: good accuracy and relative technical simplicity. Few instrumentation technologies inspire the awe and mystique commanded by pH measurement, because it is so widely misunderstood and difficult to troubleshoot.

Without elaborating on the exact chemistry of pH measurement, a few words of wisdom can be given here about pH measurement systems:

- All pH electrodes have a finite life, and that lifespan depends greatly on the type and severity of service. In some applications, a pH electrode life of one month may be considered long, and in other applications the same electrode(s) may be expected to last for over a year.
- Because the glass (measurement) electrode is responsible for generating the pH-proportional voltage, it is the one to be considered suspect if the measurement system fails to generate sufficient voltage change for a given change in pH (approximately 59 millivolts per pH unit), or fails to respond quickly enough to a fast change in test liquid pH.
- If a pH measurement system “drifts,” creating offset errors, the problem likely lies with the reference electrode, which is supposed to provide a zero-voltage connection with the measured solution.
- Because pH measurement is a logarithmic representation of ion concentration, there is an incredible range of process conditions represented in the seemingly simple 0-14 pH scale. Also, due to the nonlinear nature of the logarithmic scale, a change of 1 pH at the top end (say, from 12 to 13 pH) does not represent the same quantity of chemical activity change as a change of 1 pH at the bottom end (say, from 2 to 3 pH). Control system engineers and technicians must be aware of this dynamic if there is to be any hope of *controlling* process pH at a stable value.
- The following conditions are hazardous to measurement (glass) electrodes: high temperatures, extreme pH levels (either acidic or alkaline), high ionic concentration in the liquid, abrasion, hydrofluoric acid in the liquid (HF acid dissolves glass!), and any kind of material coating on the surface of the glass.
- Temperature changes in the measured liquid affect both the response of the measurement electrode to a given pH level (ideally at 59 mV per pH unit), and the actual pH of the liquid. Temperature measurement devices can be inserted into the liquid, and the signals from those devices used to compensate for the effect of temperature on pH measurement, but this will only compensate for the measurement electrode’s mV/pH response, not the actual pH change of the process liquid!

Advances are still being made in the field of pH measurement, some of which hold great promise for overcoming traditional limitations of pH electrodes. One such technology uses a device called a *field-effect transistor* to electrostatically measure the voltage produced by an ion-permeable membrane rather than measure the voltage with an actual voltmeter circuit. While this technology harbors limitations of its own, it is at least a pioneering concept, and may prove more practical at a later date.

NUCLEAR MAGNETIC RESONANCE

Nuclear magnetic resonance (NMR) is a physical phenomenon in which nuclei in a magnetic field absorb and re-emit electromagnetic radiation. This energy is at a specific resonance frequency which depends on the strength of the magnetic field and the magnetic properties of the isotope of the atoms; in practical applications, the frequency is similar to VHF and UHF television broadcasts (60–1000 MHz). NMR allows the observation of specific quantum mechanical magnetic properties of the atomic nucleus. Many scientific techniques exploit NMR phenomena to study molecular physics, crystals, and non-crystalline materials through nuclear magnetic resonance spectroscopy. NMR is also routinely used in advanced medical imaging techniques, such as in magnetic resonance imaging (MRI). All isotopes that contain an odd number of protons and/or neutrons have an intrinsic magnetic moment and angular momentum, in other words a nonzero spin, while all nuclides with even numbers of both have a total spin of zero. The most commonly studied nuclei are ^1H and ^{13}C , although nuclei from isotopes of many other elements (*e.g.*, ^2H , ^6Li , ^{10}B , ^{11}B , ^{14}N , ^{15}N , ^{17}O , ^{19}F , ^{23}Na , ^{29}Si , ^{31}P , ^{35}Cl , ^{113}Cd , ^{129}Xe , ^{195}Pt) have been studied by high-field NMR spectroscopy as well.

A key feature of NMR is that the resonance frequency of a particular substance is directly proportional to the strength of the applied magnetic field. It is this feature that is exploited in imaging techniques; if a sample is placed in a non-uniform magnetic field then the resonance frequencies of the sample's nuclei depend on where in the field they are located. Since, the resolution of the imaging technique depends on the magnitude of magnetic field gradient, many efforts are made to develop increased field strength, often using superconductors. The effectiveness of NMR can also be improved using hyperpolarization, and/or using two-dimensional, three-dimensional and higher-dimensional multi-frequency techniques.

The principle of NMR usually involves two sequential steps:

- The alignment (polarization) of the magnetic nuclear spins in an applied, constant magnetic field B_0 .
- The perturbation of this alignment of the nuclear spins by employing an electro-magnetic, usually radio frequency (RF) pulse. The required perturbing frequency is dependent upon the static magnetic field (H_0) and the nuclei of observation.

The two fields are usually chosen to be perpendicular to each other as this maximizes the NMR signal strength. The resulting response by the total magnetization (M) of the nuclear spins is the phenomenon that is exploited in NMR spectroscopy and magnetic resonance imaging. Both use intense applied magnetic fields (H_0) in order to achieve dispersion and very high stability to deliver spectral resolution, the details of which are described by chemical shifts, the Zeeman effect, and Knight shifts (in metals). NMR phenomena are also utilized in low-field NMR, NMR spectroscopy and MRI in the Earth's magnetic field (referred to as Earth's field NMR), and in several types of magnetometers.

History

Nuclear magnetic resonance was first described and measured in molecular beams by Isidor Rabi in 1938, by extending the Stern–Gerlach experiment, and in 1944, Rabi was awarded the Nobel Prize in Physics for this

work. In 1946, Felix Bloch and Edward Mills Purcell expanded the technique for use on liquids and solids, for which they shared the Nobel Prize in Physics in 1952.

Yevgeny Zavoisky likely observed nuclear magnetic resonance in 1941, well before Felix Bloch and Edward Mills Purcell, but dismissed the results as not reproducible. Russell H. Varian filed the “Method and means for correlating nuclear properties of atoms and magnetic fields”, U.S., Patent 2,561,490 on July 24, 1951. Varian Associates developed the first NMR unit called NMR HR-30 in 1952. Purcell had worked on the development of radar during World War II at the Massachusetts Institute of Technology’s Radiation Laboratory. His work during that project on the production and detection of radio frequency power and on the absorption of such RF power by matter laid the foundation for Rabi’s discovery of NMR. Rabi, Bloch, and Purcell observed that magnetic nuclei, like ^1H and ^{31}P , could absorb RF energy when placed in a magnetic field and when the RF was of a frequency specific to the identity of the nuclei. When this absorption occurs, the nucleus is described as being *in resonance*. Different atomic nuclei within a molecule resonate at different (radio) frequencies for the same magnetic field strength. The observation of such magnetic resonance frequencies of the nuclei present in a molecule allows any trained user to discover essential chemical and structural information about the molecule. The development of NMR as a technique in analytical chemistry and biochemistry parallels the development of electromagnetic technology and advanced electronics and their introduction into civilian use.

Theory of Nuclear Magnetic Resonance

Nuclear Spin and Magnets

All nucleons, that is neutrons and protons, composing any atomic nucleus, have the intrinsic quantum property of spin. The overall spin of the nucleus is determined by the spin quantum number S . If the numbers of both the protons and neutrons in a given nuclide are even then $S = 0$, *i.e.*, there is no overall spin. Then, just as electrons pair up in atomic orbitals, so do even numbers of protons or even numbers of neutrons, (both of which are also spin $\frac{1}{2}$ particles and hence, fermions) pair up giving zero overall spin.

However, a proton and neutron will have lower energy when their spins are parallel, not anti-parallel. Parallel spin alignment does not infringe upon the Pauli Exclusion Principle. The lowering of energy for parallel spins has to do with the quark structure of these two nucleons. Therefore, the spin ground state for the deuteron (the deuterium nucleus, or the ^2H isotope of hydrogen)—that has only a proton and a neutron—corresponds to a spin value of 1, *not of zero*.

The single, isolated deuteron therefore exhibits an NMR absorption spectrum characteristic of a quadrupolar nucleus of spin 1, which in the “rigid” state at very low temperatures is a characteristic (‘Pake’) *doublet*, (not a singlet as for a single, isolated ^1H , or any other isolated fermion or dipolar nucleus of spin $1/2$). On the other hand, because of the Pauli Exclusion Principle, the tritium isotope of hydrogen must have a pair of anti-parallel spin neutrons (of total spin zero for the neutron-spin pair), plus a proton of spin $1/2$. Therefore, the character of the tritium nucleus is again magnetic dipolar, *not quadrupolar*—like its non-radioactive deuteron cousin—and the tritium nucleus total spin value is again $1/2$, just like for the simpler, abundant hydrogen isotope, ^1H nucleus (the *proton*). The NMR absorption (radio) frequency for tritium is however slightly higher than that of ^1H because the tritium nucleus has a slightly higher gyromagnetic ratio than ^1H . In many other cases of *non-radioactive* nuclei, the overall spin is also non-zero. For example, the ^{27}Al nucleus has an overall spin value $S = \frac{5}{2}$. A non-zero spin is thus always associated with a non-zero magnetic moment (μ) via the relation $\mu = \gamma S$, where γ is the gyromagnetic ratio. It is this magnetic moment that allows the observation of NMR absorption spectra caused by transitions between nuclear spin levels. Most nuclides (with some rare exceptions) that have both even numbers of protons and even numbers of neutrons, also have zero nuclear magnetic moments, and they also

have zero magnetic dipole and quadrupole moments. Hence, such nuclides do not exhibit any NMR absorption spectra. Thus, ^{18}O is an example of a nuclide that has no NMR absorption, whereas ^{13}C , ^{31}P , ^{35}Cl and ^{37}Cl are nuclides that do exhibit NMR absorption spectra. The last two nuclei are quadrupolar nuclei whereas the preceding two nuclei (^{13}C and ^{31}P) are dipolar ones.

Electron spin resonance (ESR) is a related technique in which transitions between electronic spin levels are detected rather than nuclear ones. The basic principles are similar but the instrumentation, data analysis, and detailed theory are significantly different. Moreover, there is a much smaller number of molecules and materials with unpaired electron spins that exhibit ESR (or electron paramagnetic resonance (EPR)) absorption than those that have NMR absorption spectra. ESR has much higher sensitivity than NMR does.

Values of Spin Angular Momentum

The angular momentum associated with nuclear spin is quantized. This means both that the magnitude of angular momentum is quantized (*i.e.*, S can only take on a restricted range of values), and also that the orientation of the associated angular momentum is quantized. The associated quantum number is known as the magnetic quantum number, m , and can take values from $+S$ to $-S$, in integer steps. Hence, for any given nucleus, there are a total of $2S + 1$ angular momentum states. The z -component of the angular momentum vector (S) is therefore $S_z = m\hbar$, where $\hbar$ is the reduced Planck constant.

The z -component of the magnetic moment is simply:

$$\mu_z = \gamma S_z = \gamma m\hbar.$$

Spin Behaviour in a Magnetic Field

Consider nuclei which have a spin of one-half, like ^1H , ^{13}C or ^{19}F . The nucleus has two possible spin states: $m = 1/2$ or $m = -1/2$ (also referred to as spin-up and spin-down, or sometimes α and β spin states, respectively). These states are degenerate; that is, they have the same energy. Hence, the number of atoms in these two states will be approximately equal at thermal equilibrium.

If a nucleus is placed in a magnetic field, however, the interaction between the nuclear magnetic moment and the external magnetic field mean the two states no longer have the same energy. The energy of a magnetic moment μ when in a magnetic field B_0 is given by:

$$E = -\mu \cdot B_0 = -\mu_x B_{0x} - \mu_y B_{0y} - \mu_z B_{0z}.$$

Usually the z -axis is chosen to be along B_0 , and the above expression reduces to:

$$E = -\mu_z B_0,$$

or alternatively:

$$E = -\gamma m \hbar B_0.$$

As a result, the different nuclear spin states have different energies in a non-zero magnetic field. In less formal language, we can talk about the two spin states of a spin $1/2$ as being *aligned* either with or against the magnetic field. If γ is positive (true for most isotopes) then $m = 1/2$ is the lower energy state.

The energy difference between the two states is:

$$\Delta E = \gamma \hbar B_0,$$

and this difference results in a small population bias towards the lower energy state.

Magnetic Resonance by Nuclei

Resonant absorption by nuclear spins will occur only when electromagnetic radiation of the correct frequency (*e.g.*, equaling the Larmor precession rate) is being applied to match the energy difference between the nuclear

spin levels in a constant magnetic field of the appropriate strength. The energy of an absorbed photon is then $E = h\nu_0$, where ν_0 is the resonance radiofrequency that has to match (that is, it has to be equal to the Larmor precession frequency ν_L of the nuclear magnetization in the constant magnetic field B_0). Hence, a magnetic resonance absorption will only occur when $\Delta E = h\nu_0$, which is when $\nu_0 = \gamma B_0/2\pi$. Such magnetic resonance frequencies typically correspond to the radio frequency (or RF) range of the electromagnetic spectrum for magnetic fields up to roughly 20 T. It is this magnetic resonant absorption which is detected in NMR.

As a result, the energy gap is reduced, and the frequency required to achieve resonance is also reduced. This shift in the NMR frequency due to the electronic molecular orbital coupling to the external magnetic field is called chemical shift, and it explains why NMR is able to probe the chemical structure of molecules, which depends on the electron density distribution in the corresponding molecular orbitals. If a nucleus in a specific chemical group is shielded to a higher degree by a higher electron density of its surrounding molecular orbital, then its NMR frequency will be shifted “upfield” (that is, a lower chemical shift), whereas if it is less shielded by such surrounding electron density, then its NMR frequency will be shifted “downfield” (that is, a higher chemical shift).

Unless the local symmetry of such molecular orbitals is very high (leading to “isotropic” shift), the shielding effect will depend on the orientation of the molecule with respect to the external field (B_0). In solid-state NMR spectroscopy, magic angle spinning is required to average out this orientation dependence in order to obtain values close to the average chemical shifts. This is unnecessary in conventional NMR investigations of molecules, since, rapid “molecular tumbling” averages out the chemical shift anisotropy (CSA). In this case, the term “average” chemical shift (ACS) is used.

Relaxation

The process called population relaxation refers to nuclei that return to the thermodynamic state in the magnet. This process is also called T_1 , “spin-lattice” or “longitudinal magnetic” relaxation, where T_1 refers to the mean time for an individual nucleus to return to its thermal equilibrium state of the spins. Once the nuclear spin population is relaxed, it can be probed again, since, it is in the initial, equilibrium (mixed) state. The precessing nuclei can also fall out of alignment with each other (returning the net magnetization vector to a non-precessing field) and stop producing a signal. This is called T_2 or *transverse relaxation*. Because of the difference in the actual relaxation mechanisms involved (for example, intermolecular versus intramolecular magnetic dipole-dipole interactions), T_1 is usually (except in rare cases) longer than T_2 (that is, slower spin-lattice relaxation, for example because of smaller dipole-dipole interaction effects).

In practice, the value of T_2^* which is the actually observed decay time of the observed NMR signal, or free induction decay (to $1/e$ of the initial amplitude immediately after the resonant RF pulse), also depends on the static magnetic field inhomogeneity, which is quite significant. (There is also a smaller but significant contribution to the observed FID shortening from the RF inhomogeneity of the resonant pulse). In the corresponding FT-NMR spectrum—meaning the Fourier transform of the free induction decay—the T_2^* time is inversely related to the width of the NMR signal in frequency units. Thus, a nucleus with a long T_2 relaxation time gives rise to a very sharp NMR peak in the FT-NMR spectrum for a very homogeneous (“well-shimmed”) static magnetic field, whereas nuclei with shorter T_2 values give rise to broad FT-NMR peaks even when the magnet is shimmed well. Both T_1 and T_2 depend on the rate of molecular motions as well as the gyromagnetic ratios of both the resonating and their strongly interacting, next-neighbour nuclei that are not at resonance.

A Hahn echo decay experiment can be used to measure the dephasing time, as shown in the animation below. The size of the echo is recorded for different spacings of the two pulses. This reveals the decoherence which is not refocused by the 180° pulse. In simple cases, an exponential decay is measured which is described by the T_2 time.

NMR Spectroscopy

NMR spectroscopy is one of the principal techniques used to obtain physical, chemical, electronic and structural information about molecules due to either the chemical shift, Zeeman effect, or the Knight shift effect, or a combination of both, on the resonant frequencies of the nuclei present in the sample. It is a powerful technique that can provide detailed information on the topology, dynamics and three-dimensional structure of molecules in solution and the solid state. Thus, structural and dynamic information is obtainable (with or without “magic angle” spinning (MAS)) from NMR studies of quadrupolar nuclei (that is, those nuclei with spin $S > 1/2$) even in the presence of magnetic “dipole-dipole” interaction broadening (or simply, dipolar broadening) which is always much smaller than the quadrupolar interaction strength because it is a magnetic vs. an electric interaction effect. Additional structural and chemical information may be obtained by performing double-quantum NMR experiments for quadrupolar nuclei such as ^2H . Also, nuclear magnetic resonance is one of the techniques that has been used to design quantum automata, and also build elementary quantum computers.

Continuous-wave (CW) Spectroscopy

In its first few decades, nuclear magnetic resonance spectrometers used a technique known as continuous-wave spectroscopy (CW spectroscopy). Although NMR spectra could be, and have been, obtained using a fixed magnetic field and sweeping the frequency of the electromagnetic radiation, this more typically involved using a fixed frequency source and varying the current (and hence, magnetic field) in an electromagnet to observe the resonant absorption signals. This is the origin of the counterintuitive, but still common, “high field” and “low field” terminology for low frequency and high frequency regions respectively of the NMR spectrum.

As of 1996, CW instruments were still used for routine work because the older instruments were cheaper to maintain and operate, often operating at 60 MHz with correspondingly weaker (non-superconducting) electromagnets cooled with water rather than liquid helium. One radio coil operated continuously, sweeping through a range of frequencies, while another orthogonal coil, designed not to receive radiation from the transmitter, received signals from nuclei that reoriented in solution. As of 2014, low-end refurbished 60 MHz and 90 MHz systems were sold as FT-NMR instruments, and in 2010 the “average workhorse” NMR instrument was configured for 300 MHz.

CW spectroscopy is inefficient in comparison with Fourier analysis techniques since, it probes the NMR response at individual frequencies in succession. Since, the NMR signal is intrinsically weak, the observed spectrum suffers from a poor signal-to-noise ratio. This can be mitigated by signal averaging *i.e.*, adding the spectra from repeated measurements. While the NMR signal is constant between scans and so adds linearly, the random noise adds more slowly—proportional to the square root of the number of spectra. Hence, the overall signal-to-noise ratio increases as the square-root of the number of spectra measured.

Fourier-transform Spectroscopy

Most applications of NMR involve full NMR spectra, that is, the intensity of the NMR signal as a function of frequency. Early attempts to acquire the NMR spectrum more efficiently than simple CW methods involved illuminating the target simultaneously with more than one frequency. A revolution in NMR occurred when short pulses of radio-frequency radiation began to be used—centered at the middle of the NMR spectrum. In simple terms, a short pulse of a given “carrier” frequency “contains” a range of frequencies centered about the carrier frequency, with the range of excitation (bandwidth) being inversely proportional to the pulse duration, *i.e.*, the Fourier transform of a short pulse contains contributions from all the frequencies in the neighborhood of the principal frequency. The restricted range of the NMR frequencies made it relatively easy to use short (millisecond to microsecond) radio frequency pulses to excite the entire NMR spectrum.

Applying such a pulse to a set of nuclear spins simultaneously excites all the single-quantum NMR transitions. In terms of the net magnetization vector, this corresponds to tilting the magnetization vector away from its equilibrium position (aligned along the external magnetic field). The out-of-equilibrium magnetization vector precesses about the external magnetic field vector at the NMR frequency of the spins. This oscillating magnetization vector induces a current in a nearby pickup coil, creating an electrical signal oscillating at the NMR frequency. This signal is known as the free induction decay (FID), and it contains the vector sum of the NMR responses from all the excited spins. In order to obtain the frequency-domain NMR spectrum (NMR absorption intensity vs. NMR frequency) this time-domain signal (intensity vs. time) must be Fourier transformed.

Fortunately the development of Fourier Transform NMR coincided with the development of digital computers and the digital Fast Fourier Transform. Fourier methods can be applied to many types of spectroscopy. Richard R. Ernst was one of the pioneers of pulse NMR, and he won a Nobel Prize in chemistry in 1991 for his work on Fourier Transform NMR and his development of multi-dimensional NMR.

Multi-dimensional NMR Spectroscopy

The use of pulses of different shapes, frequencies and durations in specifically designed patterns or *pulse sequences* allows the spectroscopist to extract many different types of information about the molecule. Multi-dimensional nuclear magnetic resonance spectroscopy is a kind of FT NMR in which there are at least two pulses and, as the experiment is repeated, the pulse sequence is systematically varied. In *multidimensional nuclear magnetic resonance* there will be a sequence of pulses and, at least, one variable time period. In three dimensions, two time sequences will be varied. In four dimensions, three will be varied.

There are many such experiments. In one, these time intervals allow (amongst other things) magnetization transfer between nuclei and, therefore, the detection of the kinds of nuclear-nuclear interactions that allowed for the magnetization transfer. Interactions that can be detected are usually classified into two kinds. There are *through-bond* interactions and *through-space* interactions, the latter usually being a consequence of the nuclear Overhauser effect. Experiments of the nuclear Overhauser variety may be employed to establish distances between atoms, as for example by 2D-FT NMR of molecules in solution.

Although the fundamental concept of 2D-FT NMR was proposed by Jean Jeener from the Free University of Brussels at an International Conference, this idea was largely developed by Richard Ernst who won the 1991 Nobel prize in Chemistry for his work in FT NMR, including multi-dimensional FT NMR, and especially 2D-FT NMR of small molecules. Multi-dimensional FT NMR experiments were then further developed into powerful methodologies for studying biomolecules in solution, in particular for the determination of the structure of biopolymers such as proteins or even small nucleic acids. In 2002 Kurt Wüthrich shared the Nobel Prize in Chemistry (with John Bennett Fenn and Koichi Tanaka) for his work with protein FT NMR in solution.

Solid-state NMR Spectroscopy

This technique complements X-ray crystallography in that it is frequently applicable to molecules in a liquid or liquid crystal phase, whereas crystallography, as the name implies, is performed on molecules in a solid phase. Though nuclear magnetic resonance is used to study solids, extensive atomic-level molecular structural detail is especially challenging to obtain in the solid state. There is little signal averaging by thermal motion in the solid state, where most molecules can only undergo restricted vibrations and rotations at room temperature, each in a slightly different electronic environment, therefore exhibiting a different NMR absorption peak. Such a variation in the electronic environment of the resonating nuclei results in a blurring of the observed spectra—which is often only a broad Gaussian band for non-quadrupolar spins in a solid—thus making the interpretation of such “dipolar” and “chemical shift anisotropy” (CSA) broadened spectra either very difficult or impossible.

Professor Raymond Andrew at the University of Nottingham in the UK pioneered the development of high-resolution solid-state nuclear magnetic resonance. He was the first to report the introduction of the MAS (magic angle sample spinning; MASS) technique that allowed him to achieve spectral resolution in solids sufficient to distinguish between chemical groups with either different chemical shifts or distinct Knight shifts. In MASS, the sample is spun at several kilohertz around an axis that makes the so-called magic angle θ_m (which is $\sim 54.74^\circ$, where $\cos^2 \theta_m = 1/3$) with respect to the direction of the static magnetic field B_0 ; as a result of such magic angle sample spinning, the chemical shift anisotropy bands are averaged to their corresponding average (isotropic) chemical shift values. The above expression involving $\cos^2 \theta_m$ has its origin in a calculation that predicts the magnetic dipolar interaction effects to cancel out for the specific value of θ_m called the magic angle. One notes that correct alignment of the sample rotation axis as close as possible to θ_m is essential for cancelling out the dipolar interactions whose strength for angles sufficiently far from θ_m is usually greater than ~ 10 kHz for C–H bonds in solids, for example, and it is thus greater than their CSA values.

There are different angles for the sample spinning relative to the applied field for the averaging of quadrupole interactions and paramagnetic interactions, correspondingly $\sim 30.6^\circ$ and $\sim 70.1^\circ$.

A concept developed by Sven Hartmann and Erwin Hahn was utilized in transferring magnetization from protons to less sensitive nuclei (popularly known as cross-polarization) by M.G. Gibby, Alex Pines and John S. Waugh. Then, Jake Schaefer and Ed Stejskal demonstrated also the powerful use of cross-polarization under MASS conditions (CP-MAS) which is now routinely employed to measure high resolution spectra of low-abundance and low-sensitivity nuclei, namely carbon-13, in solids.

Sensitivity

Because the intensity of nuclear magnetic resonance signals and, hence, the sensitivity of the technique depends on the strength of the magnetic field the technique has also advanced over the decades with the development of more powerful magnets. Advances made in audio-visual technology have also improved the signal-generation and processing capabilities of newer instruments.

As noted above, the sensitivity of nuclear magnetic resonance signals is also dependent on the presence of a magnetically susceptible nuclide and, therefore, either on the natural abundance of such nuclides or on the ability of the experimentalist to artificially enrich the molecules, under study, with such nuclides. The most abundant naturally occurring isotopes of hydrogen and phosphorus (for example) are both magnetically susceptible and readily useful for nuclear magnetic resonance spectroscopy. In contrast, carbon and nitrogen have useful isotopes but which occur only in very low natural abundance. Other limitations on sensitivity arise from the quantum-mechanical nature of the phenomenon. For quantum states separated by energy equivalent to radio frequencies, thermal energy from the environment causes the populations of the states to be close to equal. Since, incoming radiation is equally likely to cause stimulated emission (a transition from the upper to the lower state) as absorption, the NMR effect depends on an excess of nuclei in the lower states.

Several factors can reduce sensitivity, including:

- Increasing temperature, which evens out the population of states. Conversely, low temperature NMR can sometimes yield better results than room-temperature NMR, providing the sample remains liquid.
- Saturation of the sample with energy applied at the resonant radiofrequency. This manifests in both CW and pulsed NMR; in the first case (CW) this happens by using too much continuous power that keeps the upper spin levels completely populated; in the second case (pulsed), each pulse (that is at least a 90° pulse) leaves the sample saturated, and four to five times the (longitudinal) relaxation time ($5T_1$) must pass before the next pulse or pulse sequence can be applied. For single pulse experiments, shorter RF pulses that tip the magnetization by less than 90° can be used, which loses some intensity of the signal, but allows for shorter *recycle delays*. The optimum there is called an

Ernst angle, after the Nobel laureate. Especially in solid state NMR, or in samples with very few nuclei with spins greater than zero, (diamond with the natural 1% of carbon-13 is especially troublesome here) the longitudinal relaxation times can be on the range of hours, while for proton-NMR they are more on the range of one second.

- Non-magnetic effects, such as electric-quadrupole coupling of spin-1 and spin-3/2 nuclei with their local environment, which broaden and weaken absorption peaks N, an abundant spin-1 nucleus, is difficult to study for this reason. High resolution NMR instead probes molecules using the rarer N isotope, which has spin-1/2.

Isotopes

Many isotopes of chemical elements can be used for NMR analysis.

Commonly used nuclei:

- ^1H , the most commonly used spin-1/2 nucleus in NMR investigation, has been studied using many forms of NMR. Hydrogen is highly abundant, especially in biological systems. It is the nucleus most sensitive to NMR signal (apart from ^3H which is not commonly used due to its instability and radioactivity). Proton NMR produces narrow chemical shift with sharp signals. Fast acquisition of quantitative results (peak integrals in stoichiometric ratio) is possible due to short relaxation time. The ^1H signal has been the sole diagnostic nucleus used for clinical magnetic resonance imaging.
- ^2H , a spin 1 nucleus commonly utilized as signal-free medium in the form of deuterated solvents during proton NMR, to avoid signal interference from hydrogen-containing solvents in measurement of ^1H solutes. Also used in determining the behaviour of lipids in lipid membranes and other solids or liquid crystals as it is a relatively non-perturbing label which can selectively replace ^1H . Alternatively, ^2H can be detected in media specially labeled with ^2H . Deuterium resonance is commonly used in high-resolution NMR spectroscopy to monitor drifts in the magnetic field strength (lock) and to improve the homogeneity of the external magnetic field.
- ^3He , is very sensitive to NMR. There is a very low percentage in natural helium, and subsequently has to be purified from ^4He . It is used mainly in studies of endohedral fullerenes, where its chemical inertness is beneficial to ascertaining the structure of the entrapping fullerene.
- ^{11}B , more sensitive than ^{10}B , yields sharper signals. Quartz tubes must be used as borosilicate glass interferes with measurement.
- ^{13}C spin-1/2, is widely used, despite its relative paucity in naturally occurring carbon (approximately 1%). It is stable to nuclear decay. Since, there is a low percentage in natural carbon, spectrum acquisition on samples which have not been experimentally enriched in ^{13}C takes a long time. Frequently used for labeling of compounds in synthetic and metabolic studies. Has low sensitivity and wide chemical shift, yields sharp signals. Low percentage makes it useful by preventing spin-spin couplings and makes the spectrum appear less crowded. Slow relaxation means that spectra are not integrable unless long acquisition times are used.
- ^{14}N , spin-1, medium sensitivity nucleus with wide chemical shift. Its large quadrupole moment interferes in acquisition of high resolution spectra, limiting usefulness to smaller molecules and functional groups with a high degree of symmetry such as the headgroups of lipids.
- ^{15}N , spin-1/2, relatively commonly used. Can be used for labeling compounds. Nucleus very insensitive but yields sharp signals. Low percentage in natural nitrogen together with low sensitivity requires high concentrations or expensive isotope enrichment.
- ^{17}O , spin-5/2, low sensitivity and very low natural abundance (0.037%), wide chemical shifts range

(up to 2000 ppm). Quadrupole moment causing a line broadening. Used in metabolic and biochemical studies in studies of chemical equilibria.

- ^{19}F , spin-1/2, relatively commonly measured. Sensitive, yields sharp signals, has wide chemical shift.
- ^{31}P , spin-1/2, 100% of natural phosphorus. Medium sensitivity, wide chemical shifts range, yields sharp lines. Spectra tend to have a moderate amount of noise. Used in biochemical studies and in coordination chemistry where phosphorus containing ligands are involved.
- ^{35}Cl and ^{37}Cl , broad signal. ^{35}Cl significantly more sensitive, preferred over ^{37}Cl despite its slightly broader signal. Organic chlorides yield very broad signals, its use is limited to inorganic and ionic chlorides and very small organic molecules.
- ^{43}Ca , used in biochemistry to study calcium binding to DNA, proteins, etc., Moderately sensitive, very low natural abundance.
- ^{195}Pt , used in studies of catalysts and complexes.

Other nuclei (usually used in the studies of their complexes and chemical binding, or to detect presence of the element):

- | | | |
|---------------------------------------|---|---|
| • ^6Li , ^7Li | • ^9Be | • ^{19}F |
| • ^{21}Ne | • ^{23}Na | • ^{25}Mg |
| • ^{27}Al | • ^{29}Si | • ^{31}P |
| • ^{33}S | • ^{39}K , ^{40}K , ^{41}K | • ^{45}Sc |
| • ^{47}Ti , ^{49}Ti | • ^{50}V , ^{51}V | • ^{53}Cr |
| • ^{55}Mn | • ^{57}Fe | • ^{59}Co |
| • ^{61}Ni | • ^{63}Cu , ^{65}Cu | • ^{67}Zn |
| • ^{69}Ga , ^{71}Ga | • ^{73}Ge | • ^{75}As |
| • ^{77}Se | • ^{81}Br | • ^{87}Rb |
| • ^{87}Sr | • ^{95}Mo | • ^{109}Ag |
| • ^{113}Cd | • ^{119}Sn | • ^{125}Te |
| • ^{127}I | • ^{133}Cs | • ^{135}Ba , ^{137}Ba |
| • ^{139}La | • ^{183}W | • ^{199}Hg . |

Applications

NMR is extensively used in medicine in the form of magnetic resonance imaging. NMR is used industrially mainly for routine analysis of chemicals. The technique is also used, for example, to measure the ratio between water and fat in foods, monitoring of flow of corrosive fluids in pipes, or to study molecular structures such as catalysts.

Medicine

The application of nuclear magnetic resonance best known to the general public is magnetic resonance imaging for medical diagnosis and magnetic resonance microscopy in research settings, however, it is also widely used in chemical studies, notably in NMR spectroscopy such as proton NMR, carbon-13 NMR, deuterium NMR and phosphorus-31 NMR. Biochemical information can also be obtained from living tissue (*e.g.*, human brain tumors) with the technique known as *in vivo* magnetic resonance spectroscopy or chemical shift NMR Microscopy. These studies are possible because nuclei are surrounded by orbiting electrons, which are charged particles that generate small, local magnetic fields that add to or subtract from the external magnetic field, and so will partially shield the nuclei. The amount of shielding depends on the exact local environment. For example, a hydrogen bonded to an oxygen will be shielded differently from a hydrogen bonded to a carbon atom. In addition, two hydrogen nuclei can interact via a process known as spin-spin coupling, if they are on the same

molecule, which will split the lines of the spectra in a recognizable way. As one of the two major spectroscopic techniques used in metabolomics, NMR is used to generate metabolic fingerprints from biological fluids to obtain information about disease states or toxic insults.

Chemistry

By studying the peaks of nuclear magnetic resonance spectra, chemists can determine the structure of many compounds. It can be a very selective technique, distinguishing among many atoms within a molecule or collection of molecules of the same type but which differ only in terms of their local chemical environment. NMR spectroscopy is used to unambiguously identify known and novel compounds, and as such, is usually required by scientific journals for identity confirmation of synthesized new compounds. By studying T_2 information, a chemist can determine the identity of a compound by comparing the observed nuclear precession frequencies to known frequencies. Further structural data can be elucidated by observing *spin-spin coupling*, a process by which the precession frequency of a nucleus can be influenced by the magnetization transfer from nearby chemically bound nuclei. Spin-spin coupling is observed in NMR of hydrogen-1 (^1H NMR), since, its natural abundance is nearly 100%; isotope enrichment is required for most other elements.

Because the nuclear magnetic resonance *timescale* is rather slow, compared to other spectroscopic methods, changing the temperature of a T_2^* experiment can also give information about fast reactions, such as the Cope rearrangement or about structural dynamics, such as ring-flipping in cyclohexane. At low enough temperatures, a distinction can be made between the axial and equatorial hydrogens in cyclohexane. An example of nuclear magnetic resonance being used in the determination of a structure is that of buckminsterfullerene (often called “buckyballs”, composition C_{60}).

This now famous form of carbon has 60 carbon atoms forming a sphere. The carbon atoms are all in identical environments and so should see the same internal H field. Unfortunately, buckminsterfullerene contains no hydrogen and so ^{13}C nuclear magnetic resonance has to be used. ^{13}C spectra require longer acquisition times since, carbon-13 is not the common isotope of carbon (unlike hydrogen, where ^1H is the common isotope). However, in 1990 the spectrum was obtained by R. Taylor and co-workers at the University of Sussex and was found to contain a single peak, confirming the unusual structure of buckminsterfullerene.

Purity Determination (w/w NMR)

NMR is primarily used for structural determination, however it can also be used for purity determination, providing that the structure and molecular weight of the compound is known. This technique requires the use of an internal standard of a known purity. Typically this standard will have a high molecular weight to facilitate accurate weighing, but relatively few protons so as to give a clear peak for later integration *e.g.*, 1,2,4,5-tetrachloro-3-nitrobenzene. Accurately weighed portions of the standard and sample are combined and analysed by NMR. Suitable peaks are selected for both compounds and the purity of the sample determined via the following equation.

$$\text{Purity} = \frac{w_{\text{std}} \times n[\text{H}]_{\text{std}} \times \text{MW}_{\text{spl}}}{w_{\text{spl}} \times \text{MW}_{\text{std}} \times n[\text{H}]_{\text{spl}}} \times P$$

Where:

- w_{std} : Weight of internal standard
- w_{spl} : Weight of sample

- n_{std} : The integrated area of the peak selected for comparison in the standard, corrected for the number of protons in that functional group
- n_{spl} : The integrated area of the peak selected for comparison in the sample, corrected for the number of protons in that functional group
- MW_{std} : Molecular weight of standard
- MW_{spl} : Molecular weight of sample
- P : Purity of internal standard.

Non-destructive Testing

Nuclear magnetic resonance is extremely useful for analyzing samples non-destructively. Radio waves and static magnetic fields easily penetrate many types of matter and anything that is not inherently ferromagnetic. For example, various expensive biological samples, such as nucleic acids, including RNA and DNA, or proteins, can be studied using nuclear magnetic resonance for weeks or months before using destructive biochemical experiments. This also makes nuclear magnetic resonance a good choice for analyzing dangerous samples.

Acquisition of Dynamic Information

In addition to providing static information on molecules by determining their 3D structures in solution, one of the remarkable advantages of NMR over X-ray crystallography is that it can be used to obtain important dynamic information.

Data Acquisition in the Petroleum Industry

Another use for nuclear magnetic resonance is data acquisition in the petroleum industry for petroleum and natural gas exploration and recovery. Initial research in this domain began in the 1950s, however, the first commercial instruments were not released until the early 1990s. A borehole is drilled into rock and sedimentary strata into which nuclear magnetic resonance logging equipment is lowered. Nuclear magnetic resonance analysis of these boreholes is used to measure rock porosity, estimate permeability from pore size distribution and identify pore fluids (water, oil and gas). These instruments are typically low field NMR spectrometers.

NMR logging, a subcategory of electromagnetic logging, measures the induced magnet moment of hydrogen nuclei (protons) contained within the fluid-filled pore space of porous media (reservoir rocks). Unlike conventional logging measurements (*e.g.*, acoustic, density, neutron, and resistivity), which respond to both the rock matrix and fluid properties and are strongly dependent on mineralogy, NMR-logging measurements respond to the presence of hydrogen protons. Because these protons primarily occur in pore fluids, NMR effectively responds to the volume, composition, viscosity, and distribution of these fluids, for example oil, gas or water. NMR logs provide information about the quantities of fluids present, the properties of these fluids, and the sizes of the pores containing these fluids.

From this information, it is possible to infer or estimate:

- The volume (porosity) and distribution (permeability) of the rock pore space
- Rock composition
- Type and quantity of fluid hydrocarbons
- Hydrocarbon producibility

The basic core and log measurement is the T_2 decay, presented as a distribution of T_2 amplitudes versus time at each sample depth, typically from 0.3 ms to 3 s. The T_2 decay is further processed to give the total pore volume (the total porosity) and pore volumes within different ranges of T_2 . The most common volumes are the

bound fluid and free fluid. A permeability estimate is made using a transform such as the Timur-Coates or SDR permeability transforms. By running the log with different acquisition parameters, direct hydrocarbon typing and enhanced diffusion are possible.

Flow Probes for NMR Spectroscopy

Recently, real-time applications of NMR in liquid media have been developed using specifically designed flow probes (flow cell assemblies) which can replace standard tube probes. This has enabled techniques that can incorporate the use of high performance liquid chromatography (HPLC) or other continuous flow sample introduction devices.

Process Control

NMR has now entered the arena of real-time process control and process optimization in oil refineries and petrochemical plants. Two different types of NMR analysis are utilized to provide real time analysis of feeds and products in order to control and optimize unit operations. Time-domain NMR (TD-NMR) spectrometers operating at low field (2–20 MHz for ^1H) yield free induction decay data that can be used to determine absolute hydrogen content values, rheological information, and component composition. These spectrometers are used in mining, polymer production, cosmetics and food manufacturing as well as coal analysis. High resolution FT-NMR spectrometers operating in the 60 MHz range with shielded permanent magnet systems yield high resolution ^1H NMR spectra of refinery and petrochemical streams. The variation observed in these spectra with changing physical and chemical properties is modeled using chemometrics to yield predictions on unknown samples. The prediction results are provided to control systems via analogue or digital outputs from the spectrometer.

Earth's Field NMR

In the Earth's magnetic field, NMR frequencies are in the audio frequency range, or the very low frequency and ultra low frequency bands of the radio frequency spectrum. Earth's field NMR (EFNMR) is typically stimulated by applying a relatively strong dc magnetic field pulse to the sample and, after the end of the pulse, analyzing the resulting low frequency alternating magnetic field that occurs in the Earth's magnetic field due to free induction decay (FID). These effects are exploited in some types of magnetometers, EFNMR spectrometers, and MRI imagers. Their inexpensive portable nature makes these instruments valuable for field use and for teaching the principles of NMR and MRI. An important feature of EFNMR spectrometry compared with high-field NMR is that some aspects of molecular structure can be observed more clearly at low fields and low frequencies, whereas other aspects observable at high fields are not observable at low fields.

This is because:

- Electron-mediated heteronuclear J -couplings (spin-spin couplings) are field independent, producing clusters of two or more frequencies separated by several Hz, which are more easily observed in a fundamental resonance of about 2 kHz. "Indeed it appears that enhanced resolution is possible due to the long spin relaxation times and high field homogeneity which prevail in EFNMR."
- Chemical shifts of several ppm are clearly separated in high field NMR spectra, but have separations of only a few millihertz at proton EFNMR frequencies, so are usually lost in noise *etc.*

Zero Field NMR

In zero field NMR all magnetic fields are shielded such that magnetic fields below 1 nT (nanotesla) are achieved and the nuclear precession frequencies of all nuclei are close to zero and indistinguishable. Under

those circumstances the observed spectra are no-longer dictated by chemical shifts but primarily by J -coupling interactions which are independent of the external magnetic field. Since, inductive detection schemes are not sensitive at very low frequencies, on the order of the J -couplings (typically between 0 and 1000 Hz), alternative detection schemes are used. Specifically, sensitive magnetometers turn out to be good detectors for zero field NMR. A zero magnetic field environment does not provide any polarization hence, it is the combination of zero field NMR with hyperpolarization schemes that makes zero field NMR attractive.

Quantum Computing

NMR quantum computing uses the spin states of nuclei within molecules as qubits. NMR differs from other implementations of quantum computers in that it uses an ensemble of systems, in this case molecules.

Magnetometers

Various magnetometers use NMR effects to measure magnetic fields, including proton precession magnetometers (PPM) (also known as proton magnetometers), and Overhauser magnetometers.

SNMR

It is based on the principle of Nuclear magnetic resonance (NMR) and measurements can be used to indirectly estimate the water content of saturated and unsaturated zones in the earth's subsurface. SNMR is used to estimate aquifer properties, including quantity of water contained in the aquifer, Porosity, and Hydraulic conductivity.

Makers of NMR Equipment

Major NMR instrument makers include Thermo Fisher Scientific, Magritek, Oxford Instruments, Bruker, Spinlock SRL, General Electric, JEOL, Kimble Chase, Philips, Siemens AG, and Agilent Technologies, Inc. (who own Varian, Inc.).

MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body in both health and disease. MRI scanners use strong magnetic fields, electric field gradients, and radio waves to generate images of the organs in the body. MRI does not involve X-rays or the use of ionizing radiation, which distinguishes it from CT or CAT scans. Magnetic resonance imaging is a medical application of nuclear magnetic resonance (NMR). NMR can also be used for *imaging* in other NMR applications such as NMR spectroscopy. While the hazards of X-rays are now well-controlled in most medical contexts, MRI may still be seen as a better choice than a CT scan. MRI is widely used in hospitals and clinics for medical diagnosis, staging of disease and follow-up without exposing the body to radiation. However, MRI may often yield different diagnostic information compared with CT. There may be risks and discomfort associated with MRI scans. Compared with CT scans, MRI scans typically take longer and are louder, and they usually need the subject to enter a narrow, confining tube.

In addition, people with some medical implants or other non-removable metal inside the body may be unable to undergo an MRI examination safely. MRI was originally called 'NMRI' (nuclear magnetic resonance imaging) and is a form of NMR, though the use of 'nuclear' in the acronym was dropped to avoid negative associations with the word. Certain atomic nuclei are able to absorb and emit radio frequency energy when placed in an external magnetic field. In clinical and research MRI, hydrogen atoms are most often used to generate a detectable radio-

frequency signal that is received by antennas in close proximity to the anatomy being examined. Hydrogen atoms are naturally abundant in people and other biological organisms, particularly in water and fat. For this reason, most MRI scans essentially map the location of water and fat in the body. Pulses of radio waves excite the nuclear spin energy transition, and magnetic field gradients localize the signal in space. By varying the parameters of the pulse sequence, different contrasts may be generated between tissues based on the relaxation properties of the hydrogen atoms therein.

Since its development in the 1970s and 1980s, MRI has proven to be a highly versatile imaging technique. While MRI is most prominently used in diagnostic medicine and biomedical research, it also may be used to form images of non-living objects. MRI scans are capable of producing a variety of chemical and physical data, in addition to detailed spatial images. The sustained increase in demand for MRI within health systems has led to concerns about cost effectiveness and overdiagnosis.

Mechanism

Construction and Physics

To perform a study, the person is positioned within an MRI scanner that forms a strong magnetic field around the area to be imaged. In most medical applications, protons (hydrogen atoms) in tissues containing water molecules create a signal that is processed to form an image of the body. First, energy from an oscillating magnetic field temporarily is applied to the patient at the appropriate resonance frequency. The excited hydrogen atoms emit a radio frequency signal, which is measured by a receiving coil. The radio signal may be made to encode position information by varying the main magnetic field using gradient coils. As these coils are rapidly switched on and off they create the characteristic repetitive noise of an MRI scan. The contrast between different tissues is determined by the rate at which excited atoms return to the equilibrium state. Exogenous contrast agents may be given to the person to make the image clearer.

The major components of an MRI scanner are the main magnet, which polarizes the sample, the shim coils for correcting shifts in the homogeneity of the main magnetic field, the gradient system which is used to localize the MR signal and the RF system, which excites the sample and detects the resulting NMR signal. The whole system is controlled by one or more computers. MRI requires a magnetic field that is both strong and uniform. The field strength of the magnet is measured in teslas—and while the majority of systems operate at 1.5 T, commercial systems are available between 0.2 and 7 T. Most clinical magnets are superconducting magnets, which require liquid helium. Lower field strengths can be achieved with permanent magnets, which are often used in “open” MRI scanners for claustrophobic patients. Recently, MRI has been demonstrated also at ultra-low fields, *i.e.*, in the microtesla-to-millitesla range, where sufficient signal quality is made possible by prepolarization (on the order of 10–100 mT) and by measuring the Larmor precession fields at about 100 microtesla with highly sensitive superconducting quantum interference devices (SQUIDS).

T1 and T2

Each tissue returns to its equilibrium state after excitation by the independent relaxation processes of T1 (spin-lattice; that is, magnetization in the same direction as the static magnetic field) and T2 (spin-spin; transverse to the static magnetic field). To create a T1-weighted image, magnetization is allowed to recover before measuring the MR signal by changing the repetition time (TR). This image weighting is useful for assessing the cerebral cortex, identifying fatty tissue, characterizing focal liver lesions and in general for obtaining morphological information, as well as for post-contrast imaging. To create a T2-weighted image, magnetization is allowed to

decay before measuring the MR signal by changing the echo time (TE). This image weighting is useful for detecting edema and inflammation, revealing white matter lesions and assessing zonal anatomy in the prostate and uterus.

The standard display of MRI images is to represent fluid characteristics in black and white images, where different tissues turn out as follows:

Signal	T1-weighted	T2-weighted
High	• Fat • Subacute hemorrhage • Melanin • Protein-rich fluid • Slowly flowing blood • Paramagnetic substances, such as gadolinium, manganese, copper • Cortical pseudolaminar necrosis	• More water content, as in edema, tumor, infarction, inflammation and infection • Extracellularly located methemoglobin in subacute hemorrhage
Inter-mediate	Gray matter darker than white matter	White matter darker than grey matter
Low	• Bone • Urine • CSF • Air • More water content, as in edema, tumor, infarction, inflammation, infection, hyperacute or chronic hemorrhage • Low proton density as in calcification	• Bone • Air • Fat • Low proton density, as in calcification and fibrosis • Paramagnetic material, such as deoxyhemoglobin, intracellular methemoglobin, iron, ferritin, hemosiderin, melanin • Protein-rich fluid

Diagnostics

Usage by Organ or System

MRI has a wide range of applications in medical diagnosis and more than 25,000 scanners are estimated to be in use worldwide. MRI affects diagnosis and treatment in many specialties although the effect on improved health outcomes is uncertain. MRI is the investigation of choice in the preoperative staging of rectal and prostate cancer and, has a role in the diagnosis, staging, and follow-up of other tumors.

Neuroimaging

MRI is the investigative tool of choice for neurological cancers, as it has better resolution than CT and offers better visualization of the posterior fossa. The contrast provided between grey and white matter makes MRI the best choice for many conditions of the central nervous system, including demyelinating diseases, dementia, cerebrovascular disease, infectious diseases, and epilepsy. Since, many images are taken milliseconds apart, it shows how the brain responds to different stimuli, enabling researchers to study both the functional and structural brain abnormalities in psychological disorders. MRI also is used in guided stereotactic surgery and radiosurgery for treatment of intracranial tumors, arteriovenous malformations, and other surgically treatable conditions using a device known as the N-localizer.

Cardiovascular

Cardiac MRI is complementary to other imaging techniques, such as echocardiography, cardiac CT, and nuclear medicine. Its applications include assessment of myocardial ischemia and viability, cardiomyopathies, myocarditis, iron overload, vascular diseases, and congenital heart disease.

Musculoskeletal

Applications in the musculoskeletal system include spinal imaging, assessment of joint disease, and soft tissue tumors.

Liver and Gastrointestinal

Hepatobiliary MR is used to detect and characterize lesions of the liver, pancreas, and bile ducts. Focal or diffuse disorders of the liver may be evaluated using diffusion-weighted, opposed-phase imaging, and dynamic contrast enhancement sequences.

Extracellular contrast agents are used widely in liver MRI and newer hepatobiliary contrast agents also provide the opportunity to perform functional biliary imaging. Anatomical imaging of the bile ducts is achieved by using a heavily T2-weighted sequence in magnetic resonance cholangiopancreatography (MRCP). Functional imaging of the pancreas is performed following administration of secretin. MR enterography provides non-invasive assessment of inflammatory bowel disease and small bowel tumors. MR-colonography may play a role in the detection of large polyps in patients at increased risk of colorectal cancer.

Angiography

Magnetic resonance angiography (MRA) generates pictures of the arteries to evaluate them for stenosis (abnormal narrowing) or aneurysms (vessel wall dilatations, at risk of rupture). MRA is often used to evaluate the arteries of the neck and brain, the thoracic and abdominal aorta, the renal arteries, and the legs (called a “run-off”).

A variety of techniques can be used to generate the pictures, such as administration of a paramagnetic contrast agent (gadolinium) or using a technique known as “flow-related enhancement” (*e.g.*, 2D and 3D time-of-flight sequences), where most of the signal on an image is due to blood that recently moved into that plane. Techniques involving phase accumulation (known as phase contrast angiography) can also be used to generate flow velocity maps easily and accurately. Magnetic resonance venography (MRV) is a similar procedure that is used to image veins. In this method, the tissue is now excited inferiorly, while the signal is gathered in the plane immediately superior to the excitation plane—thus imaging the venous blood that recently moved from the excited plane.

Contrast Agents

MRI for imaging anatomical structures or blood flow do not require contrast agents as the varying properties of the tissues or blood provide natural contrasts. However, for more specific types of imaging, exogenous contrast agents may be given intravenously, orally, or intra-articularly. The most commonly used intravenous contrast agents are based on chelates of gadolinium. In general, these agents have proved safer than the iodinated contrast agents used in X-ray radiography or CT. Anaphylactoid reactions are rare, occurring in approx. 0.03–0.1%. Of particular interest is the lower incidence of nephrotoxicity, compared with iodinated agents, when given at usual doses—this has made contrast-enhanced MRI scanning an option for patients with renal impairment, who would otherwise not be able to undergo contrast-enhanced CT.

Although gadolinium agents have proved useful for patients with renal impairment, in patients with severe renal failure requiring dialysis there is a risk of a rare but serious illness, nephrogenic systemic fibrosis, which may be linked to the use of certain gadolinium-containing agents. The most frequently linked is gadodiamide, but other agents have been linked too.

Although a causal link has not been definitively established, current guidelines in the United States are that dialysis patients should only receive gadolinium agents where essential, and that dialysis should be performed

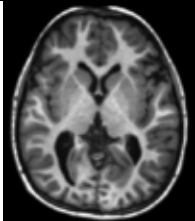
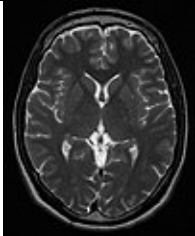
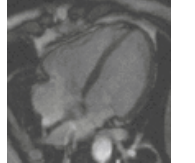
as soon as possible after the scan to remove the agent from the body promptly. In Europe, where more gadolinium-containing agents are available, a classification of agents according to potential risks has been released. Recently, a new contrast agent named gadoxetate, brand name Eovist (US) or Primovist (EU), was approved for diagnostic use: this has the theoretical benefit of a dual excretion path.

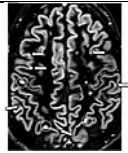
Sequences

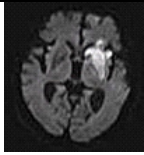
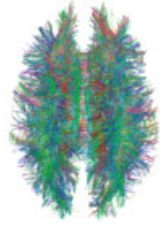
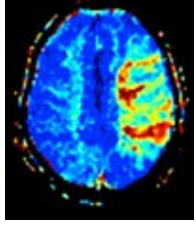
An MRI sequence is a particular setting of radiofrequency pulses and gradients, resulting in a particular image appearance. The T1 and T2 weighting can also be described as MRI sequences.

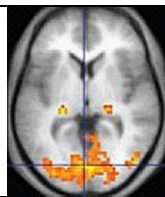
Overview Table

This table does not include uncommon and experimental sequences.

Group	Sequence	Abbr.	Physics	Main clinical distinctions	Example
Spin echo	T1 weighted	T1	Measuring spin–lattice relaxation by using a short repetition time (TR) and echo time (TE)	Lower signal for more water content, as in edema, tumor, infarction, inflammation, infection, hyperacute or chronic hemorrhage High signal for fat High signal for paramagnetic substances, such as MRI contrast agents Standard foundation and comparison for other sequences	
	T2 weighted	T2	Measuring spin–spin relaxation by using long TR and TE times	Higher signal for more water content Low signal for fat Low signal for paramagnetic substances Standard foundation and comparison for other sequences	
	Proton density weighted	PD	Long TR (to reduce T1) and short TE (to minimize T2)	Joint disease and injury. High signal from meniscus tears (pictured)	
Gradient echo	Steady-state free precession	SSFP	Maintenance of a steady, residual transverse magnetisation over successive cycles.	Creation of cardiac MRI videos (pictured).	
Inversion recovery	Short tau inversion recovery	STIR	Fat suppression by setting an inversion time where the signal of fat is zero	High signal in edema, such as in more severe stress fracture Shin splints pictured:	

	Fluid-attenuated inversion recovery	FLAIR	Fluid suppression by setting an inversion time that nulls fluids	High signal in lacunar infarction, multiple sclerosis (MS) plaques, subarachnoid haemorrhage and meningitis (pictured).	
	Double inversion recovery	DIR	Simultaneous suppression of cerebrospinal fluid and white matter by two inversion times	High signal of multiple sclerosis plaques (pictured)	

Diffusion weighted (DWI)	Conventional	DWI	Measure of Brownian motion of water molecules	High signal within minutes of cerebral infarction (pictured).	
	Apparent diffusion coefficient	ADC	Reduced T2 weighting by taking multiple conventional DWI images with different DWI weighting, and the change corresponds to diffusion	Low signal minutes after cerebral infarction (pictured)	
	Diffusion tensor	DTI	Mainly tractography (pictured) by an overall greater Brownian motion of water molecules in the directions of nerve fibers	Evaluating white matter deformation by tumors Reduced fractional anisotropy may indicate dementia	
Perfusion weighted (PWI)	Dynamic susceptibility contrast	DSC	Gadolinium contrast is injected, and rapid repeated imaging (generally gradient-echo echo-planar T2 weighted) quantifies susceptibility-induced signal loss	In cerebral infarction, the infarcted core and the penumbra have decreased perfusion (pictured).	
	Dynamic contrast enhanced	DCE	Measuring shortening of the spin-lattice relaxation (T1) induced by a gadolinium contrast bolus		
	Arterial spin labelling	ASL	Magnetic labeling of arterial blood below the imaging slab, which subsequently enters the region of interest. It does not need gadolinium contrast.		

Functional MRI (fMRI)	Blood-oxygen-level dependent imaging	BOLD	Changes in oxygen saturation-dependent magnetism of hemoglobin reflects tissue activity.	Localizing highly active brain areas before surgery	
Magnetic resonance angiography (MRA) and venography	Time-of-flight	TOF	Blood entering the imaged area is not yet magnetically saturated, giving it a much higher signal when using short echo time and flow compensation.	Detection of aneurysm, stenosis, or dissection	
	Phase-contrast magnetic resonance imaging	PC-MRA	Two gradients with equal magnitude, but opposite direction, are used to encode a phase shift, which is proportional to the velocity of spins.	Detection of aneurysm, stenosis, or dissection (pictured)	(VIPR)
Susceptibility-weighted		SWI	Sensitive for blood and calcium, by a fully flow compensated, long echo, gradient recalled echo (GRE) pulse sequence to exploit magnetic susceptibility differences between tissues	Detecting small amounts of hemorrhage (diffuse axonal injury pictured) or calcium	

Other Specialized Configurations

Magnetic Resonance Spectroscopy

Magnetic resonance spectroscopy (MRS) is used to measure the levels of different metabolites in body tissues. The MR signal produces a spectrum of resonances that corresponds to different molecular arrangements of the isotope being “excited”. This signature is used to diagnose certain metabolic disorders, especially those affecting the brain, and to provide information on tumor metabolism. Magnetic resonance spectroscopic imaging (MRSI) combines both spectroscopic and imaging methods to produce spatially localized spectra from within the sample or patient. The spatial resolution is much lower (limited by the available SNR), but the spectra in each voxel contains information about many metabolites. Because the available signal is used to encode spatial and spectral information, MRSI requires high SNR achievable only at higher field strengths (3 T and above).

Real-time MRI

Real-time MRI refers to the continuous monitoring (“filming”) of moving objects in real time. While many different strategies have been developed since, the early 2000s, a recent development reported a real-time MRI technique based on radial FLASH and iterative reconstruction that yields a temporal resolution of 20 to 30 milliseconds for images with an in-plane resolution of 1.5 to 2.0 mm. The new method promises to add important information about diseases of the joints and the heart. In many cases MRI examinations may become easier and more comfortable for patients.

Interventional MRI

The lack of harmful effects on the patient and the operator make MRI well-suited for interventional radiology, where the images produced by an MRI scanner guide minimally invasive procedures. Such procedures must be done with no ferromagnetic instruments. A specialized growing subset of interventional MRI is intraoperative MRI, in which doctors use an MRI in surgery. Some specialized MRI systems allow imaging concurrent with the surgical procedure. More typical, however, is that the surgical procedure is temporarily interrupted so that MRI can verify the success of the procedure or guide subsequent surgical work.

Magnetic Resonance Guided Focused Ultrasound

In MRgFUS therapy, ultrasound beams are focused on a tissue—guided and controlled using MR thermal imaging—and due to the significant energy deposition at the focus, temperature within the tissue rises to more than 65 °C (150 °F), completely destroying it. This technology can achieve precise ablation of diseased tissue. MR imaging provides a three-dimensional view of the target tissue, allowing for precise focusing of ultrasound energy. The MR imaging provides quantitative, real-time, thermal images of the treated area. This allows the physician to ensure that the temperature generated during each cycle of ultrasound energy is sufficient to cause thermal ablation within the desired tissue and if not, to adapt the parameters to ensure effective treatment.

Multinuclear Imaging

Hydrogen is the most frequently imaged nucleus in MRI because it is present in biological tissues in great abundance, and because its high gyromagnetic ratio gives a strong signal. However, any nucleus with a net nuclear spin could potentially be imaged with MRI. Such nuclei include helium-3, lithium-7, carbon-13, fluorine-19, oxygen-17, sodium-23, phosphorus-31 and xenon-129. ^{23}Na and ^{31}P are naturally abundant in the body, so can be imaged directly. Gaseous isotopes such as ^3He or ^{129}Xe must be hyperpolarized and then inhaled as their nuclear density is too low to yield a useful signal under normal conditions. ^{17}O and ^{19}F can be administered in sufficient quantities in liquid form (*e.g.*, ^{17}O -water) that hyperpolarization is not a necessity. Using helium or xenon has the advantage of reduced background noise, and therefore increased contrast for the image itself, because these elements are not normally present in biological tissues.

Moreover, the nucleus of any atom that has a net nuclear spin and that is bonded to a hydrogen atom could potentially be imaged via heteronuclear magnetization transfer MRI that would image the high-gyromagnetic-ratio hydrogen nucleus instead of the low-gyromagnetic-ratio nucleus that is bonded to the hydrogen atom. In principle, heteronuclear magnetization transfer MRI could be used to detect the presence or absence of specific chemical bonds.

Multinuclear imaging is primarily a research technique at present. However, potential applications include functional imaging and imaging of organs poorly seen on ^1H MRI (*e.g.*, lungs and bones) or as alternative contrast agents. Inhaled hyperpolarized ^3He can be used to image the distribution of air spaces within the lungs. Injectable solutions containing ^{13}C or stabilized bubbles of hyperpolarized ^{129}Xe have been studied as contrast agents for angiography and perfusion imaging. ^{31}P can potentially provide information on bone density and structure, as well as functional imaging of the brain. Multinuclear imaging holds the potential to chart the distribution of lithium in the human brain, this element finding use as an important drug for those with conditions such as bipolar disorder.

Molecular Imaging by MRI

MRI has the advantages of having very high spatial resolution and is very adept at morphological imaging and functional imaging. MRI does have several disadvantages though. First, MRI has a sensitivity of around 10^{-3} mol/L to 10^{-5} mol/L, which, compared to other types of imaging, can be very limiting. This problem stems

from the fact that the population difference between the nuclear spin states is very small at room temperature. For example, at 1.5 teslas, a typical field strength for clinical MRI, the difference between high and low energy states is approximately 9 molecules per 2 million. Improvements to increase MR sensitivity include increasing magnetic field strength, and hyperpolarization via optical pumping or dynamic nuclear polarization. There are also a variety of signal amplification schemes based on chemical exchange that increase sensitivity.

To achieve molecular imaging of disease biomarkers using MRI, targeted MRI contrast agents with high specificity and high relaxivity (sensitivity) are required. To date, many studies have been devoted to developing targeted-MRI contrast agents to achieve molecular imaging by MRI. Commonly, peptides, antibodies, or small ligands, and small protein domains, such as HER-2 affibodies, have been applied to achieve targeting. To enhance the sensitivity of the contrast agents, these targeting moieties are usually linked to high payload MRI contrast agents or MRI contrast agents with high relaxivities. A new class of gene targeting MR contrast agents (CA) has been introduced to show gene action of unique mRNA and gene transcription factor proteins. This new CA can trace cells with unique mRNA, microRNA and virus; tissue response to inflammation in living brains. The MR reports change in gene expression with positive correlation to TaqMan analysis, optical and electron microscopy.

Economics

In the UK, the price of a clinical 1.5-tesla MRI scanner is around £920,000/US\$1.4 million, with the lifetime maintenance cost broadly similar to the purchase cost. In the Netherlands, the average MRI scanner costs around €1 million, with a 7-T MRI having been taken in use by the UMC Utrecht in December 2007, costing €7 million. Construction of MRI suites could cost up to US\$500,000/€370,000 or more, depending on project scope. Pre-polarizing MRI (PMRI) systems using resistive electromagnets have shown promise as a low-cost alternative and have specific advantages for joint imaging near metal implants, however they are likely unsuitable for routine whole-body or neuroimaging applications.

MRI scanners have become significant sources of revenue for health care providers in the US. This is because of favorable reimbursement rates from insurers and federal government programs. Insurance reimbursement is provided in two components, an equipment charge for the actual performance and operation of the MRI scan and a professional charge for the radiologist's review of the images and/or data. In the US Northeast, an equipment charge might be \$3,500/€2,600 and a professional charge might be \$350/€260, although the actual fees received by the equipment owner and interpreting physician are often significantly less and depend on the rates negotiated with insurance companies or determined by the Medicare fee schedule. For example, an orthopedic surgery group in Illinois billed a charge of \$1,116/€825 for a knee MRI in 2007, but the Medicare reimbursement in 2007 was only \$470.91/€350. Many insurance companies require advance approval of an MRI procedure as a condition for coverage.

In the US, the Deficit Reduction Act of 2005 significantly reduced reimbursement rates paid by federal insurance programs for the equipment component of many scans, shifting the economic landscape. Many private insurers have followed suit. In the United States, an MRI of the brain with and without contrast billed to Medicare Part B entails, on average, a technical payment of US\$403/€300 and a separate payment to the radiologist of US\$93/€70. In France, the cost of an MRI exam is approximately €150/US\$205. This covers three basic scans including one with an intravenous contrast agent as well as a consultation with the technician and a written report to the patient's physician. In Japan, the cost of an MRI examination (excluding the cost of contrast material and films) ranges from US\$155/€115 to US\$180/€133, with an additional radiologist professional fee of US\$17/€12.50. In India, the cost of an MRI examination including the fee for the radiologist's opinion comes to around ₹ 3000–4000 (€37–49/US\$50–60), excluding the cost of contrast material. In the UK the retail price for an MRI scan privately ranges between £350 and £700 (€405–810).

Safety

MRI is in general a safe technique, although injuries may occur as a result of failed safety procedures or human error. Contraindications to MRI include most cochlear implants and cardiac pacemakers, shrapnel, and metallic foreign bodies in the eyes. The safety of MRI during the first trimester of pregnancy is uncertain, but it may be preferable to other options. Since, MRI does not use any ionizing radiation, its use is generally favored in preference to CT when either modality could yield the same information. In certain cases, MRI is not preferred as it may be more expensive, time-consuming, and claustrophobia-exacerbating. MRI uses powerful magnets and can therefore cause magnetic materials to move at great speeds posing risk. Deaths have occurred.

Overuse

Medical societies issue guidelines for when physicians should use MRI on patients and recommend against overuse. MRI can detect health problems or confirm a diagnosis, but medical societies often recommend that MRI not be the first procedure for creating a plan to diagnose or manage a patient's complaint. A common case is to use MRI to seek a cause of low back pain; the American College of Physicians, for example, recommends against this procedure as unlikely to result in a positive outcome for the patient.

Artifacts

An MRI artifact is a visual artifact, that is, an anomaly during visual representation. Many different artifacts can occur during magnetic resonance imaging (MRI), some affecting the diagnostic quality, while others may be confused with pathology. Artifacts can be classified as patient-related, signal processing-dependent and hardware (machine)-related. Artifacts remain a problematic in magnetic resonance imaging (MRI). Some affect the quality of the examination, while others may be confused with pathology.

Non-medical Use

MRI is used industrially mainly for routine analysis of chemicals. The nuclear magnetic resonance technique is also used, for example, to measure the ratio between water and fat in foods, monitoring of flow of corrosive fluids in pipes, or to study molecular structures such as catalysts.

History

In the late 1970s, physicists Peter Mansfield and Paul Lauterbur, developed MRI-related techniques, like the echo-planar imaging (EPI) technique. Mansfield and Lauterbur were awarded the 2003 Nobel Prize in Physiology or Medicine for their “discoveries concerning magnetic resonance imaging”.

GIBBS–DONNAN EFFECT

The Gibbs–Donnan effect (also known as the Donnan's effect, Donnan law, Donnan equilibrium, or Gibbs–Donnan equilibrium) is a name for the behaviour of charged particles near a semi-permeable membrane that sometimes fail to distribute evenly across the two sides of the membrane. The usual cause is the presence of a different charged substance that is unable to pass through the membrane and thus creates an uneven electrical charge. For example, the large anionic proteins in blood plasma are not permeable to capillary walls. Because small cations are attracted, but are not bound to the proteins, small anions will cross capillary walls away from the anionic proteins more readily than small cations.

Some ionic species can pass through the barrier while others cannot. The solutions may be gels or colloids as well as solutions of electrolytes, and as such the phase boundary between gels, or a gel and a liquid, can also act as a selective barrier. The electric potential arising between two such solutions is called the Donnan potential. The effect is named after the American physicist Josiah Willard Gibbs and the British chemist Frederick G. Donnan. The Donnan equilibrium is prominent in the triphasic model for articular cartilage proposed by Mow and Lai, as well as in electrochemical fuel cells and dialysis. The Donnan effect is extra osmotic pressure attributable to cations (Na^+ and K^+) attached to dissolved plasma proteins.

Example

The presence of a charged impermeant ion (for example, a protein) on one side of a membrane will result in an asymmetric distribution of permeant charged ions.

The Gibbs–Donnan equation at equilibrium states (assuming permeant ions are Na^+ and Cl^-):

$$[\text{Na}_{\text{Side 1}}] \times [\text{Cl}_{\text{Side 1}}] = [\text{Na}_{\text{Side 2}}] \times [\text{Cl}_{\text{Side 2}}]$$

Example:

Start	Equilibrium	Osmolarity
Side 1: 9 Na, 9 Cl	Side 1: 6 Na, 6 Cl	Side 1: 12
Side 2: 9 Na, 9 Protein	Side 2: 12 Na, 3 Cl, 9 Protein	Side 2: 24

Double Donnan

Note that Sides 1 and 2 are no longer in osmotic equilibrium (*i.e.*, the total osmolytes on each side are not the same). *In vivo*, ion balance does not equilibrate at the proportions that would be predicted by the Gibbs–Donnan model, because the cell cannot tolerate the attendant large influx of water. This is balanced by instating a functionally impermeant cation, Na^+ , extracellularly to counter the anionic protein. Na^+ does cross the membrane via leak channels (the permeability is approximately 1/10 that of K^+ , the most permeant ion) but, as per the pump-leak model, it is extruded by the Na^+/K^+ -ATPase.

pH Change

Because there is a difference in concentration of ions on either side of the membrane, the pH may also differ when protons are involved. In many instances, from ultrafiltration of proteins to ion exchange chromatography, the pH of the buffer adjacent to the charged groups of the membrane is different from the pH of the rest of the buffer solution. When the charged groups are negative (basic), then they will attract protons so that the pH will be lower than the surrounding buffer. When the charged groups are positive (acidic), then they will repel protons so that the pH will be higher than the surrounding buffer.

Physiological Applications

Red Blood Cells

When tissue cells are in a protein-containing fluid, the Donnan effect of the cytoplasmic proteins is equal and opposite to the Donnan effect of the extracellular proteins. The opposing Donnan effects cause chloride ions to migrate inside the cell, increasing the intracellular chloride concentration. The Donnan effect may explain why some red blood cells do not have active sodium pumps; the effect relieves the osmotic pressure of plasma proteins, which is why sodium pumping is less important for maintaining the cell volume.

Neurology

Brain tissue swelling, known as cerebral oedema, results from brain injury and other traumatic head injuries that can increase intracranial pressure (ICP). Negatively charged molecules within cells create a fixed charge density, which increases intracranial pressure through the Donnan effect. ATP pumps maintain a negative membrane potential even though negative charges leak across the membrane; this action establishes a chemical and electrical gradient.

The negative charge in the cell and ions outside the cell creates a thermodynamic potential; if damage occurs to the brain and cells lose their membrane integrity, ions will rush into the cell to balance chemical and electrical gradients that were previously established. The membrane voltage will become zero, but the chemical gradient will still exist. To neutralize the negative charges within the cell, cations flow in, which increases the osmotic pressure inside relative to the outside of the cell. The increased osmotic pressure forces water to flow into the cell and tissue swelling occurs.

H⁺ CONCENTRATION AND PH

Acid–base Reaction

An acid–base reaction is a chemical reaction that occurs between an acid and a base, which can be used to determine pH. Several theoretical frameworks provide alternative conceptions of the reaction mechanisms and their application in solving related problems; these are called the acid–base theories, for example, Brønsted–Lowry acid–base theory.

Their importance becomes apparent in analyzing acid–base reactions for gaseous or liquid species, or when acid or base character may be somewhat less apparent. The first of these concepts was provided by the French chemist Antoine Lavoisier, around 1776.

It is important to think of the acid–base reaction models as theories that compliment each other. For example the current Lewis model has the broadest definition of what an acid and base are, with the Bronsted–Lowry theory being a subset of what acids and bases are, and the Arrhenius theory being the most restrictive.

Acid–base Definitions

Historic Development

Lavoisier's Oxygen Theory of Acids

The first scientific concept of acids and bases was provided by Lavoisier in around 1776.

Since, Lavoisier's knowledge of strong acids was mainly restricted to oxoacids, such as HNO₃ (nitric acid) and H₂SO₄ (sulfuric acid), which tend to contain central atoms in high oxidation states surrounded by oxygen, and since, he was not aware of the true composition of the hydrohalic acids (HF, HCl, HBr, and HI), he defined acids in terms of their containing *oxygen*, which in fact he named from Greek words meaning “acid-former” (from the Greek (*oxys*) meaning “acid” or “sharp” and (*geinomai*) meaning “engender”). The Lavoisier definition held for over 30 years, until the 1810 article and subsequent lectures by Sir Humphry Davy in which he proved the lack of oxygen in H₂S, H₂Te, and the hydrohalic acids. However, Davy failed to develop a new theory, concluding that “acidity does not depend upon any particular elementary substance, but upon peculiar arrangement of various substances”. One notable modification of oxygen theory was provided by Berzelius, who stated that acids are oxides of nonmetals while bases are oxides of metals.

Liebig's Hydrogen Theory of Acids

In 1838, Justus von Liebig proposed that an acid is a hydrogen-containing substance in which the hydrogen could be replaced by a metal. This redefinition was based on his extensive work on the chemical composition of organic acids, finishing the doctrinal shift from oxygen-based acids to hydrogen-based acids started by Davy. Liebig's definition, while completely empirical, remained in use for almost 50 years until the adoption of the Arrhenius definition.

Arrhenius Definition

The first modern definition of acids and bases in molecular terms was devised by Svante Arrhenius. A hydrogen theory of acids, it followed from his 1884 work with Friedrich Wilhelm Ostwald in establishing the presence of ions in aqueous solution and led to Arrhenius receiving the Nobel Prize in Chemistry in 1903.

As defined by Arrhenius:

- An Arrhenius acid is a substance that dissociates in water to form hydrogen ions (H^+); that is, an acid increases the concentration of H^+ ions in an aqueous solution.

This causes the protonation of water, or the creation of the hydronium (H_3O^+) ion. Thus, in modern times, the symbol H^+ is interpreted as a shorthand for H_3O^+ , because it is now known that a bare proton does not exist as a free species in aqueous solution.

- An Arrhenius base is a substance that dissociates in water to form hydroxide (OH^-) ions; that is, a base increases the concentration of OH^- ions in an aqueous solution.

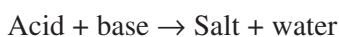
The Arrhenius definitions of acidity and alkalinity are restricted to aqueous solutions, and refer to the concentration of the solvent ions. Under this definition, pure H_2SO_4 and HCl dissolved in toluene are not acidic, and molten NaOH and solutions of calcium amide in liquid ammonia are not alkaline. This led to the development of the Brønsted-Lowry theory and subsequent Lewis theory to account for these non-aqueous exceptions. *Overall, to qualify as an Arrhenius acid, upon the introduction to water, the chemical must either cause, directly or otherwise:*

- An increase in the aqueous hydronium concentration, or
- A decrease in the aqueous hydroxide concentration.

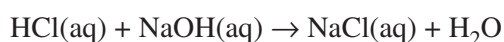
Conversely, to qualify as an Arrhenius base, upon the introduction to water, the chemical must either cause, directly or otherwise:

- A decrease in the aqueous hydronium concentration, or
- An increase in the aqueous hydroxide concentration.

The reaction of an acid with a base is called a neutralization reaction. The products of this reaction are a salt and water.



In this traditional representation an acid-base neutralization reaction is formulated as a double-replacement reaction. For example, the reaction of hydrochloric acid, HCl , with sodium hydroxide, NaOH , solutions produces a solution of sodium chloride, NaCl , and some additional water molecules.



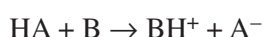
The modifier (aq) in this equation was implied by Arrhenius, rather than included explicitly. It indicates that the substances are dissolved in water. Though all three substances, HCl , NaOH and NaCl are capable of existing as pure compounds, in aqueous solutions they are fully dissociated into the aquated ions H^+ , Cl^- , Na^+ and OH^- .

Brønsted-Lowry Definition

The Brønsted-Lowry definition, formulated in 1923, independently by Johannes Nicolaus Brønsted in Denmark and Martin Lowry in England, is based upon the idea of protonation of bases through the de-protonation

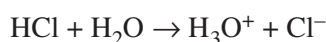
of acids—that is, the ability of acids to “donate” hydrogen ions (H^+)—otherwise known as protons—to bases, which “accept” them. An acid–base reaction is, thus, the removal of a hydrogen ion from the acid and its addition to the base. The removal of a hydrogen ion from an acid produces its *conjugate base*, which is the acid with a hydrogen ion removed. The reception of a proton by a base produces its *conjugate acid*, which is the base with a hydrogen ion added. Unlike the previous definitions, the Brønsted–Lowry definition does not refer to the formation of salt and solvent, but instead to the formation of *conjugate acids* and *conjugate bases*, produced by the transfer of a proton from the acid to the base. In this approach, acids and bases are fundamentally different in behaviour from salts, which are seen as electrolytes, subject to the theories of Debye, Onsager, and others. An acid and a base react not to produce a salt and a solvent, but to form a new acid and a new base. The concept of neutralization is thus absent. Brønsted–Lowry acid–base behaviour is formally independent of any solvent, making it more all-encompassing than the Arrhenius model. The calculation of pH under the Arrhenius model depended on alkalis (bases) dissolving in water (aqueous solution). The Brønsted–Lowry model expanded what could be pH tested using insoluble and soluble solutions (gas, liquid, solid).

The general formula for acid–base reactions according to the Brønsted–Lowry definition is:



where HA represents the acid, B represents the base, BH^+ represents the conjugate acid of B, and A^- represents the conjugate base of HA.

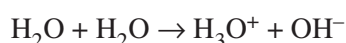
For example, a Brønsted–Lowry model for the dissociation of hydrochloric acid (HCl) in aqueous solution would be the following:



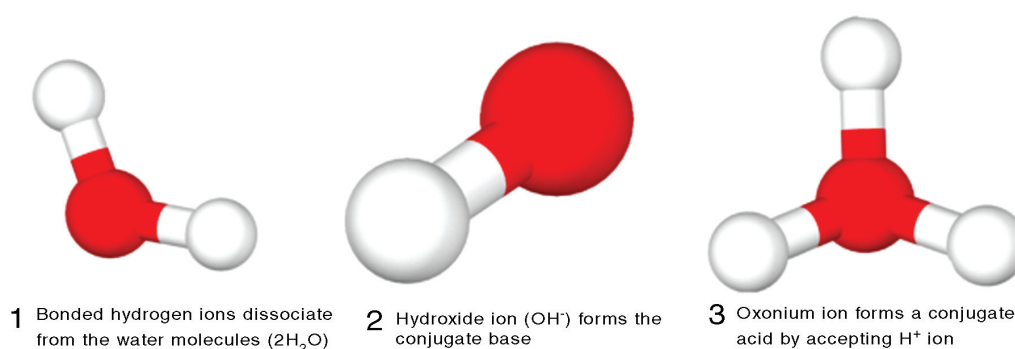
The removal of H^+ from the HCl produces the chloride ion, Cl^- , the conjugate base of the acid. The addition of H^+ to the H_2O (acting as a base) forms the hydronium ion, H_3O^+ , the conjugate acid of the base.

Water is amphoteric—that is, it can act as both an acid and a base.

The Brønsted–Lowry model explains this, showing the dissociation of water into low concentrations of hydronium and hydroxide ions:

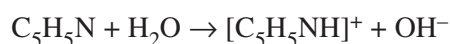


This equation is demonstrated in the image below:

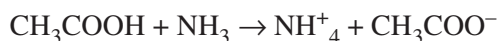


Here, one molecule of water acts as an acid, donating an H^+ and forming the conjugate base, OH^- , and a second molecule of water acts as a base, accepting the H^+ ion and forming the conjugate acid, H_3O^+ .

As an example of water acting as an acid, consider an aqueous solution of pyridine, $\text{C}_5\text{H}_5\text{N}$.



In this example, a water molecule is split into a hydrogen ion, which is donated to a pyridine molecule, and an hydroxide ion. In the Brønsted–Lowry model, the solvent does not necessarily have to be water that is required by the Arrhenius Acid–Base model. For example, consider what happens when acetic acid, CH_3COOH , dissolves in liquid ammonia.



An H^+ ion is removed from acetic acid, forming its conjugate base, the acetate ion, CH_3COO^- . The addition of an H^+ ion to an ammonia molecule of the solvent creates its conjugate acid, the ammonium ion, NH_4^+ .

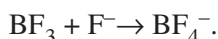
The Brønsted–Lowry model calls hydrogen-containing substances (like HCl) acids. Thus, some substances, which many chemists considered to be acids, such as SO_3 or BCl_3 , are excluded from this classification due to lack of hydrogen. Gilbert N. Lewis wrote in 1938, “To restrict the group of acids to those substances that contain hydrogen interferes as seriously with the systematic understanding of chemistry as would the restriction of the term oxidizing agent to substances containing oxygen.” Furthermore, KOH and KNH_2 are not considered Brønsted bases, but rather salts containing the bases OH^- and NH_2^- .

Lewis Definition

The hydrogen requirement of Arrhenius and Brønsted–Lowry was removed by the Lewis definition of acid–base reactions, devised by Gilbert N. Lewis in 1923, in the same year as Brønsted–Lowry, but it was not elaborated by him until 1938. Instead of defining acid–base reactions in terms of protons or other bonded substances, the Lewis definition defines a base (referred to as a *Lewis base*) to be a compound that can donate an *electron pair*, and an acid (a *Lewis acid*) to be a compound that can receive this electron pair.

For example, boron trifluoride, BF_3 is a typical Lewis acid. It can accept a pair of electrons as it has a vacancy in its octet. The fluoride ion has a full octet and can donate a pair of electrons.

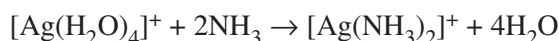
Thus,



is a typical Lewis acid, Lewis base reaction. All compounds of group 13 elements with a formula AX_3 can behave as Lewis acids. Similarly, compounds of group 15 elements with a formula DY_3 , such as amines, NR_3 , and phosphines, PR_3 , can behave as Lewis bases. Adducts between them have the formula $\text{X}_3\text{A} \leftarrow \text{DY}_3$ with a dative covalent bond, shown symbolically as $\leftarrow$, between the atoms A (acceptor) and D (donor). Compounds of group 16 with a formula DX_2 may also act as Lewis bases; in this way, a compound like an ether, R_2O , or a thioether, R_2S , can act as a Lewis base. The Lewis definition is not limited to these examples. For instance, carbon monoxide acts as a Lewis base when it forms an adduct with boron trifluoride, of formula $\text{F}_3\text{B} \leftarrow \text{CO}$.

Adducts involving metal ions are referred to as co-ordination compounds; each ligand donates a pair of electrons to the metal ion.

The reaction,



can be seen as an acid–base reaction in which a stronger base (ammonia) replaces a weaker one (water).

The Lewis and Brønsted–Lowry definitions are consistent with each other since, the reaction:



is an acid–base reaction in both theories.

Solvent System Definition

One of the limitations of the Arrhenius definition is its reliance on water solutions. Edward Curtis Franklin studied the acid–base reactions in liquid ammonia in 1905 and pointed out the similarities to the water-based Arrhenius theory. Albert F. O. Germann, working with liquid phosgene, COCl_2 , formulated the solvent-based theory in 1925, thereby generalizing the Arrhenius definition to cover aprotic solvents.

Germann pointed out that in many solutions, there are ions in equilibrium with the neutral solvent molecules:

- *Solvonium:* A generic name for a positive ion.

- *Solvate*: A generic name for a negative ion.

For example, water and ammonia undergo such dissociation into hydronium and hydroxide, and ammonium and amide, respectively:

- $2 \text{H}_2\text{O} \rightarrow \text{H}_3\text{O}^+ + \text{OH}^-$
- $2 \text{NH}_3 \rightleftharpoons \text{NH}_4^+ + \text{NH}_2^-$

Some aprotic systems also undergo such dissociation, such as dinitrogen tetroxide into nitrosonium and nitrate, antimony trichloride into dichloroantimonium and tetrachloroantimonate, and phosgene into chlorocarboxonium and chloride:

- $\text{N}_2\text{O}_4 \rightarrow \text{NO}^+ + \text{NO}_3^-$
- $2 \text{SbCl}_3 \rightarrow \text{SbCl}_2^+ + \text{SbCl}_4^-$
- $\text{COCl}_2 \rightarrow \text{COCl}^+ + \text{Cl}^-$

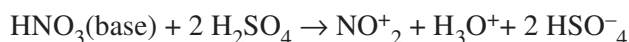
A solute that causes an increase in the concentration of the solvonium ions and a decrease in the concentration of solvate ions is defined as an *acid*. A solute that causes an increase in the concentration of the solvate ions and a decrease in the concentration of the solvonium ions is defined as a *base*.

Thus, in liquid ammonia, KNH_2 (supplying NH_2^-) is a strong base, and NH_4NO_3 (supplying NH_4^+) is a strong acid. In liquid sulfur dioxide (SO_2), thionyl compounds (supplying SO^{2+}) behave as acids, and sulfites (supplying SO_3^{2-}) behave as bases.

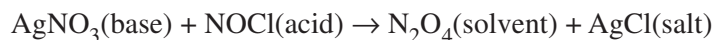
The non-aqueous acid–base reactions in liquid ammonia are similar to the reactions in water:

- $2 \text{NaNH}_2(\text{base}) + \text{Zn}(\text{NH}_2)_2(\text{amphiphilic amide}) \rightarrow \text{Na}_2[\text{Zn}(\text{NH}_2)_4]$
- $2 \text{NH}_4\text{I}(\text{acid}) + \text{Zn}(\text{NH}_2)_2(\text{amphiphilic amide}) \rightarrow [\text{Zn}(\text{NH}_2)_4]\text{I}_2$

Nitric acid can be a base in liquid sulfuric acid:



The unique strength of this definition shows in describing the reactions in aprotic solvents; for example, in liquid N_2O_4 :



Because the solvent system definition depends on the solute as well as on the solvent itself, a particular solute can be either an acid or a base depending on the choice of the solvent: HClO_4 is a strong acid in water, a weak acid in acetic acid, and a weak base in fluorosulfonic acid; this characteristic of the theory has been seen as both a strength and a weakness, because some substances (such as SO_3 and NH_3) have been seen to be acidic or basic on their own right. On the other hand, solvent system theory has been criticized as being too general to be useful. Also, it has been thought that there is something intrinsically acidic about hydrogen compounds, a property not shared by non-hydrogenic solvonium salts.

Lux–Flood Definition

This acid–base theory was a revival of oxygen theory of acids and bases, proposed by German chemist Hermann Lux in 1939, further improved by Håkon Flood circa 1947 and is still used in modern geochemistry and electrochemistry of molten salts. This definition describes an acid as an oxide ion (O^{2-}) acceptor and a base as an oxide ion donor.

For example:

- $\text{MgO}(\text{base}) + \text{CO}_2(\text{acid}) \rightarrow \text{MgCO}_3$
- $\text{CaO}(\text{base}) + \text{SiO}_2(\text{acid}) \rightarrow \text{CaSiO}_3$
- $\text{NO}_3^-(\text{base}) + \text{S}_2\text{O}_7^{2-}(\text{acid}) \rightarrow \text{NO}_2^+ + 2 \text{SO}_4^{2-}$

This theory is also useful in the systematisation of the reactions of noble gas compounds, especially the xenon oxides, fluorides, and oxofluorides.

Usanovich Definition

Mikhail Usanovich developed a general theory that does not restrict acidity to hydrogen-containing compounds, but his approach, published in 1938, was even more general than Lewis theory. Usanovich's theory can be summarized as defining an acid as anything that accepts negative species or donates positive ones, and a base as the reverse. This defined the concept of redox (oxidation-reduction) as a special case of acid–base reactions

Some examples of Usanovich acid–base reactions include:

- $\text{Na}_2\text{O}(\text{base}) + \text{SO}_3(\text{acid}) \rightarrow 2 \text{Na}^+ + \text{SO}_4^{2-}$ (species exchanged: O^{2-} anion)
- $3 (\text{NH}_4)_2\text{S}(\text{base}) + \text{Sb}_2\text{S}_5(\text{acid}) \rightarrow 6 \text{NH}_4^+ + 2 \text{SbS}_4^{3-}$ (species exchanged: 3 S^{2-} anions)
- $2\text{Na}(\text{base}) + \text{Cl}_2(\text{acid}) \rightarrow 2\text{Na}^+ + 2\text{Cl}^-$ (species exchanged: 2 electrons)

Rationalizing the Strength of Lewis Acid–base Interactions

HSAB Theory

In 1963, [Ralph Person] proposed a qualitative concept known as the Hard and Soft Acids and Bases principle. later made quantitative with help of Robert Parr in 1984. ‘Hard’ applies to species that are small, have high charge states, and are weakly polarizable. ‘Soft’ applies to species that are large, have low charge states and are strongly polarizable. Acids and bases interact, and the most stable interactions are hard–hard and soft–soft. This theory has found use in organic and inorganic chemistry.

ECW model:

The ECW Model is quantitative model that describes and predicts the strength of Lewis acid base interactions, $-\Delta H$. The model assigned E and C parameters to many Lewis acids and bases. Each acid is characterized by an E_A and a C_A . Each base is likewise characterized by its own E_B and C_B . The E and C parameters refer, respectively, to the electrostatic and covalent contributions to the strength of the bonds that the acid and base will form.

The equation is,

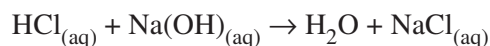
$$-\Delta H = E_A E_B + C_A C_B + W$$

The W term represents a constant energy contribution for acid–base reaction such as the cleavage of a dimeric acid or base. The equation predicts reversal of acids and base strengths. The graphical presentations of the equation show that there is no single order of Lewis base strengths or Lewis acid strengths.

Acid–base Equilibrium

The reaction of a strong acid with a strong base is essentially a quantitative reaction.

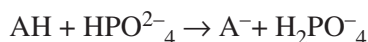
For example,



In this reaction both the sodium and chloride ions are spectators as the neutralization reaction,



does not involve them. With weak bases addition of acid is not quantitative because a solution of a weak base is a buffer solution. A solution of a weak acid is also a buffer solution. When a weak acid reacts with a weak base an equilibrium mixture is produced. For example, adenine, written as AH can react with a hydrogen phosphate ion, HPO_4^{2-} .



The equilibrium constant for this reaction can be derived from the acid dissociation constants of adenine and of the dihydrogen phosphate ion.

$$[A^-][H^+] = K_{a1}[AH][HPO_4^{2-}][H^+] = K_{a2}[H_2PO_4^-]$$

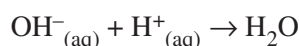
The notation $[X]$ signifies “concentration of X”. When these two equations are combined by eliminating the hydrogen ion concentration, an expression for the equilibrium constant, K is obtained.

$$[A^-][H_2PO_4^-] = K[AH][HPO_4^{2-}]; K = K_{a1}/K_{a2}$$

Acid–alkali Reaction

An acid–alkali reaction is a special case of an acid–base reaction, where the base used is also an alkali. When an acid reacts with an alkali salt (a metal hydroxide), the product is a metal salt and water. Acid–alkali reactions are also neutralization reactions.

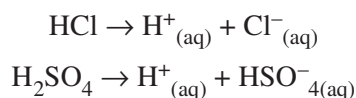
In general, acid–alkali reactions can be simplified to,



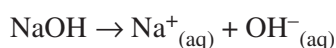
by omitting spectator ions.

Acids are in general pure substances that contain hydrogen cations (H^+) or cause them to be produced in solutions.

Hydrochloric acid (HCl) and sulfuric acid (H_2SO_4) are common examples. In water, these break apart into ions:



The alkali breaks apart in water, yielding dissolved hydroxide ions:



RESPIRATORY FUNCTION OF HEMOGLOBIN

Hemoglobin (American) or haemoglobin (British) abbreviated Hb or Hgb, is the iron-containing oxygen-transport metalloprotein in the red blood cells of all vertebrates (with the exception of the fish family Channichthyidae) as well as the tissues of some invertebrates. Hemoglobin in the blood carries oxygen from the lungs or gills to the rest of the body (*i.e.*, the tissues). There it releases the oxygen to permit aerobic respiration to provide energy to power the functions of the organism in the process called metabolism. A healthy individual has “12 to 16” grams of haemoglobin in every 100 ml of blood.

In mammals, the protein makes up about 96% of the red blood cells’ dry content (by weight), and around 35% of the total content (including water). Hemoglobin has an oxygen-binding capacity of 1.34 mL O_2 per gram, which increases the total blood oxygen capacity seventy-fold compared to dissolved oxygen in blood. The mammalian hemoglobin molecule can bind (carry) up to four oxygen molecules. Hemoglobin is involved in the transport of other gases: It carries some of the body’s respiratory carbon dioxide (about 20–25% of the total) as carbaminohemoglobin, in which CO_2 is bound to the globin protein. The molecule also carries the important regulatory molecule nitric oxide bound to a globin protein thiol group, releasing it at the same time as oxygen.

Hemoglobin is also found outside red blood cells and their progenitor lines. Other cells that contain hemoglobin include the A9 dopaminergic neurons in the substantia nigra, macrophages, alveolar cells, lungs, retinal pigment epithelium, hepatocytes, mesangial cells in the kidney, endometrial cells, cervical cells and vaginal epithelial cells. In these tissues, hemoglobin has a non-oxygen-carrying function as an antioxidant and a regulator of iron metabolism.

Hemoglobin and hemoglobin-like molecules are also found in many invertebrates, fungi, and plants. In these organisms, hemoglobins may carry oxygen, or they may act to transport and regulate other small molecules and ions such as carbon dioxide, nitric oxide, hydrogen sulfide and sulfide. A variant of the molecule, called leghemoglobin, is used to scavenge oxygen away from anaerobic systems, such as the nitrogen-fixing nodules of leguminous plants, before the oxygen can poison (deactivate) the system.

Genetics

Hemoglobin consists of protein subunits (the “globin” molecules), and these proteins, in turn, are folded chains of a large number of different amino acids called polypeptides. The amino acid sequence of any polypeptide created by a cell is in turn determined by the stretches of DNA called genes. In all proteins, it is the amino acid sequence that determines the protein’s chemical properties and function.

There is more than one hemoglobin gene: in humans, hemoglobin A (the main form of hemoglobin present) is coded for by the genes, *HBA1*, *HBA2*, and *HBB*. The amino acid sequences of the globin proteins in hemoglobins usually differ between species. These differences grow with evolutionary distance between species. For example, the most common hemoglobin sequences in humans and chimpanzees are nearly identical, differing by only one amino acid in both the alpha and the beta globin protein chains. These differences grow larger between less closely related species. Even within a species, different variants of hemoglobin always exist, although one sequence is usually a “most common” one in each species. Mutations in the genes for the hemoglobin protein in a species result in hemoglobin variants. Many of these mutant forms of hemoglobin cause no disease. Some of these mutant forms of hemoglobin, however, cause a group of hereditary diseases termed the *hemoglobinopathies*. The best known hemoglobinopathy is sickle-cell disease, which was the first human disease whose mechanism was understood at the molecular level. A (mostly) separate set of diseases called thalassemias involves underproduction of normal and sometimes abnormal hemoglobins, through problems and mutations in globin gene regulation. All these diseases produce anemia.

Variations in hemoglobin amino acid sequences, as with other proteins, may be adaptive. For example, hemoglobin has been found to adapt in different ways to high altitudes. Organisms living at high elevations experience lower partial pressures of oxygen compared to those at sea level. This presents a challenge to the organisms that inhabit such environments because hemoglobin, which normally binds oxygen at high partial pressures of oxygen, must be able to bind oxygen when it is present at a lower pressure. Different organisms have adapted to such a challenge. For example, recent studies have suggested genetic variants in deer mice that help explain how deer mice that live in the mountains are able to survive in the thin air that accompanies high altitudes. A researcher from the University of Nebraska-Lincoln found mutations in four different genes that can account for differences between deer mice that live in lowland prairies versus the mountains. After examining wild mice captured from both highlands and lowlands, it was found that: the genes of the two breeds are “virtually identical—except for those that govern the oxygen-carrying capacity of their hemoglobin”. “The genetic difference enables highland mice to make more efficient use of their oxygen”, since, less is available at higher altitudes, such as those in the mountains.

Mammoth hemoglobin featured mutations that allowed for oxygen delivery at lower temperatures, thus enabling mammoths to migrate to higher latitudes during the Pleistocene. This was also found in hummingbirds that inhabit the Andes. Hummingbirds already expend a lot of energy and thus have high oxygen demands and yet Andean hummingbirds have been found to thrive in high altitudes. Non-synonymous mutations in the hemoglobin gene of multiple species living at high elevations (*Oreotrochilus*, *A. castelnaudii*, *C. violifer*, *P. gigas*, and *A. viridicuada*) have caused the protein to have less of an affinity for inositol hexaphosphate (IHP), a molecule found in birds that has a similar role as 2,3-BPG in humans; this results in the ability to bind oxygen in lower partial pressures.

Birds' unique circulatory lungs also promote efficient use of oxygen at low partial pressures of O_2 . These two adaptations reinforce each other and account for birds' remarkable high-altitude performance. Hemoglobin adaptation extends to humans, as well. Studies have found that a small number of native Tibetan women have a genotype which codes for hemoglobin to be more highly saturated with oxygen. Natural selection seems to be the main force working on this gene because the mortality rate of offspring is significantly lower for women with higher hemoglobin-oxygen affinity when compared to the mortality rate of offspring from women with low hemoglobin-oxygen affinity. While the exact genotype and mechanism by which this occurs is not yet clear, selection is acting on these women's ability to bind oxygen in low partial pressures, which overall allows them to better sustain crucial metabolic processes.

Synthesis

Hemoglobin (Hb) is synthesized in a complex series of steps. The heme part is synthesized in a series of steps in the mitochondria and the cytosol of immature red blood cells, while the globin protein parts are synthesized by ribosomes in the cytosol. Production of Hb continues in the cell throughout its early development from the proerythroblast to the reticulocyte in the bone marrow. At this point, the nucleus is lost in mammalian red blood cells, but not in birds and many other species. Even after the loss of the nucleus in mammals, residual ribosomal RNA allows further synthesis of Hb until the reticulocyte loses its RNA soon after entering the vasculature (this hemoglobin-synthetic RNA in fact gives the reticulocyte its reticulated appearance and name).

Structure

Hemoglobin has a quaternary structure characteristic of many multi-subunit globular proteins. Most of the amino acids in hemoglobin form alpha helices, and these helices are connected by short non-helical segments. Hydrogen bonds stabilize the helical sections inside this protein, causing attractions within the molecule, which then causes each polypeptide chain to fold into a specific shape. Hemoglobin's quaternary structure comes from its four subunits in roughly a tetrahedral arrangement.

In most vertebrates, the hemoglobin molecule is an assembly of four globular protein subunits. Each subunit is composed of a protein chain tightly associated with a non-protein prosthetic heme group. Each protein chain arranges into a set of alpha-helix structural segments connected together in a globin fold arrangement. Such a name is given because this arrangement is the same folding motif used in other heme/globin proteins such as myoglobin. This folding pattern contains a pocket that strongly binds the heme group.

A heme group consists of an iron (Fe) ion (charged atom) held in a heterocyclic ring, known as a porphyrin. This porphyrin ring consists of four pyrrole molecules cyclically linked together (by methine bridges) with the iron ion bound in the center. The iron ion, which is the site of oxygen binding, coordinates with the four nitrogen atoms in the center of the ring, which all lie in one plane. The iron is bound strongly (covalently) to the globular protein via the N atoms of the imidazole ring of F8 histidine residue (also known as the proximal histidine) below the porphyrin ring. A sixth position can reversibly bind oxygen by a coordinate covalent bond, completing the octahedral group of six ligands. Oxygen binds in an "end-on bent" geometry where one oxygen atom binds to Fe and the other protrudes at an angle. When oxygen is not bound, a very weakly bonded water molecule fills the site, forming a distorted octahedron.

Even though carbon dioxide is carried by hemoglobin, it does not compete with oxygen for the iron-binding positions but is bound to the protein chains of the structure. The iron ion may be either in the Fe^{2+} or in the Fe^{3+} state, but ferrihemoglobin (methemoglobin) (Fe^{3+}) cannot bind oxygen. In binding, oxygen temporarily and reversibly oxidizes (Fe^{2+}) to (Fe^{3+}) while oxygen temporarily turns into the superoxide ion, thus iron must exist in the +2 oxidation state to bind oxygen. If superoxide ion associated to Fe^{3+} is protonated, the hemoglobin iron

will remain oxidized and incapable of binding oxygen. In such cases, the enzyme methemoglobin reductase will be able to eventually reactivate methemoglobin by reducing the iron center. In adult humans, the most common hemoglobin type is a tetramer (which contains four subunit proteins) called *hemoglobin A*, consisting of two α and two β subunits non-covalently bound, each made of 141 and 146 amino acid residues, respectively. This is denoted as $\alpha_2\beta_2$. The subunits are structurally similar and about the same size. Each subunit has a molecular weight of about 16,000 daltons, for a total molecular weight of the tetramer of about 64,000 daltons (64,458 g/mol). Thus, 1 g/dL = 0.1551 mmol/L. Hemoglobin A is the most intensively studied of the hemoglobin molecules. In human infants, the hemoglobin molecule is made up of 2 α chains and 2 γ chains. The gamma chains are gradually replaced by β chains as the infant grows. The four polypeptide chains are bound to each other by salt bridges, hydrogen bonds, and the hydrophobic effect.

Oxygen Saturation

In general, hemoglobin can be saturated with oxygen molecules (oxyhemoglobin), or desaturated with oxygen molecules (deoxyhemoglobin).

Oxyhemoglobin

Oxyhemoglobin is formed during physiological respiration when oxygen binds to the heme component of the protein hemoglobin in red blood cells. This process occurs in the pulmonary capillaries adjacent to the alveoli of the lungs. The oxygen then travels through the blood stream to be dropped off at cells where it is utilized as a terminal electron acceptor in the production of ATP by the process of oxidative phosphorylation. It does not, however, help to counteract a decrease in blood pH. Ventilation, or breathing, may reverse this condition by removal of carbon dioxide, thus causing a shift up in pH.

Hemoglobin exists in two forms, a *taut (tense) form* (T) and a *relaxed form* (R). Various factors such as low pH, high CO_2 and high 2,3 BPG at the level of the tissues favor the taut form, which has low oxygen affinity and releases oxygen in the tissues. Conversely, a high pH, low CO_2 , or low 2,3 BPG favors the relaxed form, which can better bind oxygen. The partial pressure of the system also affects O_2 affinity where, at high partial pressures of oxygen (such as those present in the alveoli), the relaxed (high affinity, R) state is favoured. Inversely, at low partial pressures (such as those present in respiring tissues), the (low affinity, T) tense state is favoured. Additionally, the binding of oxygen to the iron(II) heme pulls the iron into the plane of the porphyrin ring, causing a slight conformational shift. The shift encourages oxygen to bind to the three remaining heme units within hemoglobin (thus, oxygen binding is cooperative).

Deoxygenated Hemoglobin

Deoxygenated hemoglobin is the form of hemoglobin without the bound oxygen. The absorption spectra of oxyhemoglobin and deoxyhemoglobin differ. The oxyhemoglobin has significantly lower absorption of the 660 nm wavelength than deoxyhemoglobin, while at 940 nm its absorption is slightly higher. This difference is used for the measurement of the amount of oxygen in a patient's blood by an instrument called a pulse oximeter. This difference also accounts for the presentation of cyanosis, the blue to purplish color that tissues develop during hypoxia. Deoxygenated hemoglobin is paramagnetic; it is weakly attracted to magnetic fields. In contrast, oxygenated hemoglobin exhibits diamagnetism, a weak repulsion from a magnetic field.

Evolution of Vertebrate Hemoglobin

Scientists agree that the event that separated myoglobin from hemoglobin occurred after lampreys diverged from jawed vertebrates. This separation of myoglobin and hemoglobin allowed for the different functions of

the two molecules to arise and develop: myoglobin has more to do with oxygen storage while hemoglobin is tasked with oxygen transport. The α - and β -like globin genes encode the individual subunits of the protein. The predecessors of these genes arose through another duplication event also after the gnathostome common ancestor derived from jawless fish, approximately 450–500 million years ago.

The development of α and β genes created the potential for hemoglobin to be composed of multiple subunits, a physical composition central to hemoglobin's ability to transport oxygen. Having multiple subunits contributes to hemoglobin's ability to bind oxygen cooperatively as well as be regulated allosterically. Subsequently, the α gene also underwent a duplication event to form the *HBA1* and *HBA2* genes. These further duplications and divergences have created a diverse range of α - and β -like globin genes that are regulated so that certain forms occur at different stages of development. Most ice fish of the family Channichthyidae have lost their hemoglobin genes as an adaptation to cold water.

Iron's Oxidation State in Oxyhemoglobin

Assigning oxygenated hemoglobin's oxidation state is difficult because oxyhemoglobin (Hb-O_2), by experimental measurement, is diamagnetic (no net unpaired electrons), yet the lowest-energy (ground-state) electron configurations in both oxygen and iron are paramagnetic (suggesting at least one unpaired electron in the complex).

The lowest-energy form of oxygen, and the lowest energy forms of the relevant oxidation states of iron, are these:

- Triplet oxygen, the lowest-energy molecular oxygen species, has two unpaired electrons in antibonding π^* molecular orbitals.
- Iron(II) tends to exist in a high-spin $3d^6$ configuration with four unpaired electrons.
- Iron(III) ($3d^5$) has an odd number of electrons, and thus must have one or more unpaired electrons, in any energy state.

All of these structures are paramagnetic (have unpaired electrons), not diamagnetic. Thus, a non-intuitive (e.g., a higher-energy for at least one species) distribution of electrons in the combination of iron and oxygen must exist, in order to explain the observed diamagnetism and no unpaired electrons.

The two logical possibilities to produce diamagnetic (no net spin) Hb-O_2 are:

- Low-spin Fe^{2+} binds to singlet oxygen. Both low-spin iron and singlet oxygen are diamagnetic. However, the singlet form of oxygen is the higher-energy form of the molecule.
- Low-spin Fe^{3+} binds to O_2^- (the superoxide ion) and the two unpaired electrons couple antiferromagnetically, giving observed diamagnetic properties. Here, the iron has been oxidized (has lost one electron), and the oxygen has been reduced (has gained one electron).

Another possible model in which low-spin Fe^{4+} binds to peroxide, O_2^{2-} , can be ruled out by itself, because the iron is paramagnetic (although the peroxide ion is diamagnetic). Here, the iron has been oxidized by two electrons, and the oxygen reduced by two electrons.

Direct experimental data:

- X-ray photoelectron spectroscopy suggests iron has an oxidation state of approximately 3.2.
- Infrared vibrational frequencies of the O-O bond suggests a bond length fitting with superoxide (a bond order of about 1.6, with superoxide being 1.5).
- X-ray Absorption Near Edge Structures at the iron K-edge. The energy shift of 5 eV between deoxyhemoglobin and oxyhemoglobin, as for all the methemoglobin species, strongly suggests an actual local charge closer to Fe^{3+} than Fe^{2+} .

Thus, the nearest formal oxidation state of iron in Hb-O_2 is the +3 state, with oxygen in the “1 state (as superoxide, O_2^-). The diamagnetism in this configuration arises from the single unpaired electron on superoxide

aligning antiferromagnetically with the single unpaired electron on iron (in a low-spin d^5 state), to give no net spin to the entire configuration, in accordance with diamagnetic oxyhemoglobin from experiment. The second choice of the logical possibilities above for diamagnetic oxyhemoglobin being found correct by experiment, is not surprising: singlet oxygen (possibility #1) is an unrealistically high energy state. Model 3 leads to unfavorable separation of charge (and does not agree with the magnetic data), although it could make a minor contribution as a resonance form.

Iron's shift to a higher oxidation state in Hb-O_2 decreases the atom's size, and allows it into the plane of the porphyrin ring, pulling on the coordinated histidine residue and initiating the allosteric changes seen in the globulins. Early postulates by bio-inorganic chemists claimed that possibility #1 (above) was correct and that iron should exist in oxidation state II. This conclusion seemed likely, since, the iron oxidation state III as methemoglobin, when *not* accompanied by superoxide. O_2^- to "hold" the oxidation electron, was known to render hemoglobin incapable of binding normal triplet O_2 as it occurs in the air. It was thus assumed that iron remained as Fe(II) when oxygen gas was bound in the lungs. The iron chemistry in this previous classical model was elegant, but the required presence of the diamagnetic, high-energy, singlet oxygen molecule was never explained. It was classically argued that the binding of an oxygen molecule placed high-spin iron(II) in an octahedral field of strong-field ligands; this change in field would increase the crystal field splitting energy, causing iron's electrons to pair into the low-spin configuration, which would be diamagnetic in Fe(II). This forced low-spin pairing is indeed thought to happen in iron when oxygen binds, but is not enough to explain iron's change in size. Extraction of an additional electron from iron by oxygen is required to explain both iron's smaller size and observed increased oxidation state, and oxygen's weaker bond.

The assignment of a whole-number oxidation state is a formalism, as the covalent bonds are not required to have perfect bond orders involving whole electron transfer. Thus, all three models for paramagnetic Hb-O_2 may contribute to some small degree (by resonance) to the actual electronic configuration of Hb-O_2 . However, the model of iron in Hb-O_2 being Fe(III) is more correct than the classical idea that it remains Fe(II).

Cooperativity

When oxygen binds to the iron complex, it causes the iron atom to move back towards the center of the plane of the porphyrin ring. At the same time, the imidazole side-chain of the histidine residue interacting at the other pole of the iron is pulled towards the porphyrin ring. This interaction forces the plane of the ring sideways towards the outside of the tetramer, and also induces a strain in the protein helix containing the histidine as it moves nearer to the iron atom. This strain is transmitted to the remaining three monomers in the tetramer, where it induces a similar conformational change in the other heme sites such that binding of oxygen to these sites becomes easier. As oxygen binds to one monomer of hemoglobin, the tetramer's conformation shifts from the T (tense) state to the R (relaxed) state. This shift promotes the binding of oxygen to the remaining three monomer's heme groups, thus saturating the hemoglobin molecule with oxygen. In the tetrameric form of normal adult hemoglobin, the binding of oxygen is, thus, a cooperative process. The binding affinity of hemoglobin for oxygen is increased by the oxygen saturation of the molecule, with the first molecules of oxygen bound influencing the shape of the binding sites for the next ones, in a way favorable for binding. This positive cooperative binding is achieved through steric conformational changes of the hemoglobin protein complex as discussed above; *i.e.*, when one subunit protein in hemoglobin becomes oxygenated, a conformational or structural change in the whole complex is initiated, causing the other subunits to gain an increased affinity for oxygen. As a consequence, the oxygen binding curve of hemoglobin is sigmoidal, or *S*-shaped, as opposed to the normal hyperbolic curve associated with noncooperative binding. The dynamic mechanism of the cooperativity in hemoglobin and its relation with the low-frequency resonance has been discussed.

Binding for Ligands Other than Oxygen

Besides the oxygen ligand, which binds to hemoglobin in a cooperative manner, hemoglobin ligands also include competitive inhibitors such as carbon monoxide (CO) and allosteric ligands such as carbon dioxide (CO₂) and nitric oxide (NO). The carbon dioxide is bound to amino groups of the globin proteins to form carbaminohemoglobin; this mechanism is thought to account for about 10% of carbon dioxide transport in mammals. Nitric oxide can also be transported by hemoglobin; it is bound to specific thiol groups in the globin protein to form an S-nitrosothiol, which dissociates into free nitric oxide and thiol again, as the hemoglobin releases oxygen from its heme site. This nitric oxide transport to peripheral tissues is hypothesized to assist oxygen transport in tissues, by releasing vasodilatory nitric oxide to tissues in which oxygen levels are low.

Competitive

The binding of oxygen is affected by molecules such as carbon monoxide (for example, from tobacco smoking, exhaust gas, and incomplete combustion in furnaces). CO competes with oxygen at the heme binding site. Hemoglobin's binding affinity for CO is 250 times greater than its affinity for oxygen, meaning that small amounts of CO dramatically reduce hemoglobin's ability to transport oxygen. Since, carbon monoxide is a colorless, odorless and tasteless gas, and poses a potentially fatal threat, carbon monoxide detectors have become commercially available to warn of dangerous levels in residences. When hemoglobin combines with CO, it forms a very bright red compound called carboxyhemoglobin, which may cause the skin of CO poisoning victims to appear pink in death, instead of white or blue. When inspired air contains CO levels as low as 0.02%, headache and nausea occur; if the CO concentration is increased to 0.1%, unconsciousness will follow. In heavy smokers, up to 20% of the oxygen-active sites can be blocked by CO.

In similar fashion, hemoglobin also has competitive binding affinity for cyanide (CN⁻), sulfur monoxide (SO), and sulfide (S²⁻), including hydrogen sulfide (H₂S). All of these bind to iron in heme without changing its oxidation state, but they nevertheless inhibit oxygen-binding, causing grave toxicity.

The iron atom in the heme group must initially be in the ferrous (Fe²⁺) oxidation state to support oxygen and other gases' binding and transport (it temporarily switches to ferric during the time oxygen is bound, as explained above). Initial oxidation to the ferric (Fe³⁺) state without oxygen converts hemoglobin into "hemoglobin" or methemoglobin, which cannot bind oxygen. Hemoglobin in normal red blood cells is protected by a reduction system to keep this from happening. Nitric oxide is capable of converting a small fraction of hemoglobin to methemoglobin in red blood cells. The latter reaction is a remnant activity of the more ancient nitric oxide dioxygenase function of globins.

Allosteric

Carbon dioxide occupies a different binding site on the hemoglobin. Carbon dioxide is more readily dissolved in deoxygenated blood, facilitating its removal from the body after the oxygen has been released to tissues undergoing metabolism. This increased affinity for carbon dioxide by the venous blood is known as the Haldane effect.

Through the enzyme carbonic anhydrase, carbon dioxide reacts with water to give carbonic acid, which decomposes into bicarbonate and protons:



Hence, blood with high carbon dioxide levels is also lower in pH (more acidic). Hemoglobin can bind protons and carbon dioxide, which causes a conformational change in the protein and facilitates the release of

oxygen. Protons bind at various places on the protein, while carbon dioxide binds at the α -amino group. Carbon dioxide binds to hemoglobin and forms carbaminohemoglobin. This decrease in hemoglobin's affinity for oxygen by the binding of carbon dioxide and acid is known as the Bohr effect. The Bohr effect favors the T state rather than the R state. (shifts the O_2 -saturation curve to the *right*). Conversely, when the carbon dioxide levels in the blood decrease (*i.e.*, in the lung capillaries), carbon dioxide and protons are released from hemoglobin, increasing the oxygen affinity of the protein.

A reduction in the total binding capacity of hemoglobin to oxygen (*i.e.*, shifting the curve down, not just to the right) due to reduced pH is called the root effect. This is seen in bony fish. It is necessary for hemoglobin to release the oxygen that it binds; if not, there is no point in binding it. The sigmoidal curve of hemoglobin makes it efficient in binding (taking up O_2 in lungs), and efficient in unloading (unloading O_2 in tissues). In people acclimated to high altitudes, the concentration of 2,3-Bisphosphoglycerate (2,3-BPG) in the blood is increased, which allows these individuals to deliver a larger amount of oxygen to tissues under conditions of lower oxygen tension. This phenomenon, where molecule Y affects the binding of molecule X to a transport molecule Z, is called a *heterotropic* allosteric effect. Hemoglobin in organisms at high altitudes has also adapted such that it has less of an affinity for 2,3-BPG and so the protein will be shifted more towards its R state. In its R state, hemoglobin will bind oxygen more readily, thus allowing organisms to perform the necessary metabolic processes when oxygen is present at low partial pressures.

Animals other than humans use different molecules to bind to hemoglobin and change its O_2 affinity under unfavorable conditions. Fish use both ATP and GTP. These bind to a phosphate "pocket" on the fish hemoglobin molecule, which stabilizes the tense state and therefore decreases oxygen affinity. GTP reduces hemoglobin oxygen affinity much more than ATP, which is thought to be due to an extra hydrogen bond formed that further stabilizes the tense state. Under hypoxic conditions, the concentration of both ATP and GTP is reduced in fish red blood cells to increase oxygen affinity.

A variant hemoglobin, called fetal hemoglobin (HbF , $\alpha_2\gamma_2$), is found in the developing fetus, and binds oxygen with greater affinity than adult hemoglobin. This means that the oxygen binding curve for fetal hemoglobin is left-shifted (*i.e.*, a higher percentage of hemoglobin has oxygen bound to it at lower oxygen tension), in comparison to that of adult hemoglobin. As a result, fetal blood in the placenta is able to take oxygen from maternal blood. Hemoglobin also carries nitric oxide (NO) in the globin part of the molecule. This improves oxygen delivery in the periphery and contributes to the control of respiration. NO binds reversibly to a specific cysteine residue in globin; the binding depends on the state (R or T) of the hemoglobin. The resulting S-nitrosylated hemoglobin influences various NO-related activities such as the control of vascular resistance, blood pressure and respiration. NO is not released in the cytoplasm of erythrocytes but transported by an anion exchanger called AE1 out of them.

Types in Humans

Hemoglobin variants are a part of the normal embryonic and fetal development. They may also be pathologic mutant forms of hemoglobin in a population, caused by variations in genetics. Some well-known hemoglobin variants, such as sickle-cell anemia, are responsible for diseases and are considered hemoglobinopathies. Other variants cause no detectable pathology, and are thus considered non-pathological variants.

In the embryo:

- Gower 1 ($\zeta_2\varepsilon_2$)
- Gower 2 ($\alpha_2\varepsilon_2$) (PDB: 1A9W)
- Hemoglobin Portland I ($\zeta_2\gamma_2$)
- Hemoglobin Portland II ($\zeta_2\beta_2$).

In the fetus:

- Hemoglobin F ($\alpha_2\gamma_2$) (PDB: 1FDH).

After birth:

- Hemoglobin A ($\alpha_2\beta_2$) (PDB: 1BZ0)–The most common with a normal amount over 95%
- Hemoglobin A₂ ($\alpha_2\delta_2$)– δ chain synthesis begins late in the third trimester and, in adults, it has a normal range of 1.5–3.5%
- Hemoglobin F ($\alpha_2\gamma_2$)–In adults Hemoglobin F is restricted to a limited population of red cells called F-cells. However, the level of Hb F can be elevated in persons with sickle-cell disease and beta-thalassemia.

Variant forms that cause disease:

- Hemoglobin D-Punjab–($\alpha_2\beta^D_2$)–A variant form of hemoglobin.
- Hemoglobin H (β_4)–A variant form of hemoglobin, formed by a tetramer of β chains, which may be present in variants of α thalassemia.
- Hemoglobin Barts (γ_4)–A variant form of hemoglobin, formed by a tetramer of γ chains, which may be present in variants of α thalassemia.
- Hemoglobin S ($\alpha_2\beta^S_2$)–A variant form of hemoglobin found in people with sickle cell disease. There is a variation in the β -chain gene, causing a change in the properties of hemoglobin, which results in sickling of red blood cells.
- Hemoglobin S ($\alpha_2\beta^S_2$)–A variant form of hemoglobin found in people with sickle cell disease. There is a variation in the β -chain gene, causing a change in the properties of hemoglobin, which results in sickling of red blood cells.
- Hemoglobin C ($\alpha_2\beta^C_2$)–Another variant due to a variation in the β -chain gene. This variant causes a mild chronic hemolytic anemia.
- Hemoglobin E ($\alpha_2\beta^E_2$)–Another variant due to a variation in the β -chain gene. This variant causes a mild chronic hemolytic anemia.
- Hemoglobin AS–A heterozygous form causing sickle cell trait with one adult gene and one sickle cell disease gene
- Hemoglobin SC disease–A compound heterozygous form with one sickle gene and another encoding Hemoglobin C.
- Hemoglobin Hopkins-2–A variant form of hemoglobin that is sometimes viewed in combination with Hemoglobin S to produce sickle cell disease.

Amino Acid

Amino acids are organic compounds containing amine ($-\text{NH}_2$) and carboxyl ($-\text{COOH}$) functional groups, along with a side chain (R group) specific to each amino acid. The key elements of an amino acid are carbon (C), hydrogen (H), oxygen (O), and nitrogen (N), although other elements are found in the side chains of certain amino acids. About 500 naturally occurring amino acids are known (though only 20 appear in the genetic code) and can be classified in many ways. They can be classified according to the core structural functional groups' locations as alpha-(α -), beta-(β -), gamma-(γ -) or delta-(δ -) amino acids; other categories relate to polarity, pH level, and side chain group type (aliphatic, acyclic, aromatic, containing hydroxyl or sulfur, *etc.*). In the form of proteins, amino acid residues form the second-largest component (water is the largest) of human muscles and other tissues. Beyond their role as residues in proteins, amino acids participate in a number of processes such as neurotransmitter transport and biosynthesis.

In biochemistry, amino acids having both the amine and the carboxylic acid groups attached to the first (alpha-) carbon atom have particular importance. They are known as 2-, alpha-, or α -amino acids (generic formula $\text{H}_2\text{NCHRCOOH}$ in most cases, where R is an organic substituent known as a “side chain”); often the term “amino acid” is used to refer specifically to these. They include the 22 proteinogenic (“protein-building”) amino acids, which combine into peptide chains (“polypeptides”) to form the building-blocks of a vast array of proteins. These are all L-stereoisomers (“left-handed” isomers), although a few D-amino acids (“right-handed”) occur in bacterial envelopes, as a neuromodulator (D-serine), and in some antibiotics. Twenty of the proteinogenic amino acids are encoded directly by triplet codons in the genetic code and are known as “standard” amino acids. The other two (“non-standard” or “non-canonical”) are selenocysteine (present in many prokaryotes as well as most eukaryotes, but not coded directly by DNA), and pyrrolysine (found only in some archaea and one bacterium). Pyrrolysine and selenocysteine are encoded via variant codons; for example, selenocysteine is encoded by stop codon and SECIS element. *N*-formylmethionine (which is often the initial amino acid of proteins in bacteria, mitochondria, and chloroplasts) is generally considered as a form of methionine rather than as a separate proteinogenic amino acid. Codon–tRNA combinations not found in nature can also be used to “expand” the genetic code and form novel proteins known as alloproteins incorporating non-proteinogenic amino acids.

Many important proteinogenic and non-proteinogenic amino acids have biological functions. For example, in the human brain, glutamate (standard glutamic acid) and gamma-amino-butyric acid (“GABA”, non-standard gamma-amino acid) are, respectively, the main excitatory and inhibitory neurotransmitters. Hydroxyproline, a major component of the connective tissue collagen, is synthesised from proline. Glycine is a biosynthetic precursor to porphyrins used in red blood cells. Carnitine is used in lipid transport. Nine proteinogenic amino acids are called “essential” for humans because they cannot be produced from other compounds by the human body and so must be taken in as food. Others may be conditionally essential for certain ages or medical conditions. Essential amino acids may also differ between species. Because of their biological significance, amino acids

are important in nutrition and are commonly used in nutritional supplements, fertilizers, and food technology. Industrial uses include the production of drugs, biodegradable plastics, and chiral catalysts.

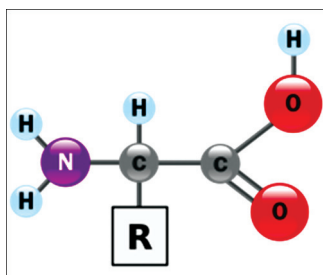


Fig. The structure of an alpha amino acid in its unionized form.

HISTORY

The first few amino acids were discovered in the early 19th century. In 1806, French chemists Louis-Nicolas Vauquelin and Pierre Jean Robiquet isolated a compound in asparagus that was subsequently named asparagine, the first amino acid to be discovered. Cystine was discovered in 1810, although its monomer, cysteine, remained undiscovered until 1884. Glycine and leucine were discovered in 1820. The last of the 20 common amino acids to be discovered was threonine in 1935 by William Cumming Rose, who also determined the essential amino acids and established the minimum daily requirements of all amino acids for optimal growth. The unity of the chemical category was recognized by Wurtz in 1865, but he gave no particular name to it. Usage of the term “amino acid” in the English language is from 1898, while the German term, *Aminosäure*, was used earlier. Proteins were found to yield amino acids after enzymatic digestion or acid hydrolysis. In 1902, Emil Fischer and Franz Hofmeister independently proposed that proteins are formed from many amino acids, whereby bonds are formed between the amino group of one amino acid with the carboxyl group of another, resulting in a linear structure that Fischer termed “peptide”.

GENERAL STRUCTURE

In the structure shown at the top of the page, R represents a side chain specific to each amino acid. The carbon atom next to the carboxyl group (which is therefore numbered 2 in the carbon chain starting from that functional group) is called the α -carbon. Amino acids containing an amino group bonded directly to the alpha carbon are referred to as *alpha amino acids*. These include amino acids such as proline which contain secondary amines, which used to be often referred to as “imino acids”.

ISOMERISM

The alpha amino acids are the most common form found in nature, but only when occurring in the L-isomer. The alpha carbon is a chiral carbon atom, with the exception of glycine which has two indistinguishable hydrogen atoms on the alpha carbon. Therefore, all alpha amino acids but glycine can exist in either of two enantiomers, called L or D amino acids, which are mirror images of each other. While L-amino acids represent all of the amino acids found in proteins during translation in the ribosome, D-amino acids are found in some proteins produced by enzyme posttranslational modifications after translation and translocation to the endoplasmic reticulum, as in exotic sea-dwelling organisms such as cone snails. They are also abundant components of the peptidoglycan cell walls of bacteria, and D-serine may act as a neurotransmitter in the brain. D-amino acids are used in racemic crystallography to create centrosymmetric crystals, which (depending on the protein) may allow for easier and more robust protein structure determination. The L and D convention for amino acid configuration refers not to the optical activity of the amino acid itself but rather to the optical activity of the isomer of glyceraldehyde from which that amino acid can, in theory, be synthesized (D-glyceraldehyde is dextrorotatory; L-glyceraldehyde is levorotatory).

In alternative fashion, the (*S*) and (*R*) designators are used to indicate the absolute stereochemistry. Almost all of the amino acids in proteins are (*S*) at the α carbon, with cysteine being (*R*) and glycine non-chiral. Cysteine has its side chain in the same geometric position as the other amino acids, but the *R/S* terminology is reversed because of the higher atomic number of sulfur compared to the carboxyl oxygen gives the side chain a higher priority, whereas the atoms in most other side chains give them lower priority.

SIDE CHAINS

In amino acids that have a carbon chain attached to the α -carbon (such as lysine, shown to the right) the carbons are labeled in order as α , β , γ , δ , and so on. In some amino acids, the amine group is attached to the β or γ -carbon, and these are therefore referred to as *beta* or *gamma amino acids*. Amino acids are usually classified by the properties of their side chain into four groups. The side chain can make an amino acid a weak acid or a weak base, and a hydrophile if the side chain is polar or a hydrophobe if it is nonpolar. The chemical structures of the 22 standard amino acids, along with their chemical properties, are described more fully in the article on these proteinogenic amino acids. The phrase “branched-chain amino acids” or BCAA refers to the amino acids having aliphatic side chains that are non-linear; these are leucine, isoleucine, and valine. Proline is the only proteinogenic amino acid whose side-group links to the α -amino group and, thus, is also the only proteinogenic amino acid containing a secondary amine at this position. In chemical terms, proline is, therefore, an imino acid, since, it lacks a primary amino group, although it is still classed as an amino acid in the current biochemical nomenclature, and may also be called an “N-alkylated alpha-amino acid”.

ZWITTERIONS

The α -carboxylic acid group of amino acids is a weak acid, meaning that it releases a hydron (such as a proton) at moderate pH values. In other words, carboxylic acid groups ($-\text{CO}_2\text{H}$) can be deprotonated to become negative carboxylates ($-\text{CO}_2^-$). The negatively charged carboxylate ion predominates at pH values greater than the pK_a of the carboxylic acid group (mean for the 20 common amino acids is about 2.2.). In a complementary fashion, the α -amine of amino acids is a weak base, meaning that it accepts a proton at moderate pH values. In other words, α -amino groups (NH_2) can be protonated to become positive α -ammonium groups ($^+\text{NH}_3$). The positively charged α -ammonium group predominates at pH values less than the pK_a of the α -ammonium group (mean for the 20 common α -amino acids is about 9.4). Because all amino acids contain amine and carboxylic acid functional groups, they share amphiprotic properties. Below pH 2.2, the predominant form will have a neutral carboxylic acid group and a positive α -ammonium ion (net charge +1), and above pH 9.4, a negative carboxylate and neutral α -amino group (net charge -1). But at pH between 2.2 and 9.4, an amino acid usually contains both a negative carboxylate and a positive α -ammonium group, as shown in structure (2) on the right, so has net zero charge. This molecular state is known as a zwitterion, from the German Zwitter meaning *hermaphrodite* or *hybrid*. The fully neutral form (structure (1) on the left) is a very minor species in aqueous solution throughout the pH range (less than 1 part in 10^7). Amino acids exist as zwitterions also in the solid phase, and crystallize with salt-like properties unlike typical organic acids or amines.

ISOELECTRIC POINT

The variation in titration curves when the amino acids can be grouped by category. With the exception of tyrosine, using titration to distinguish among hydrophobic amino acids is problematic. At pH values between the two pK_a values, the zwitterion predominates, but coexists in dynamic equilibrium with small amounts of net negative and net positive ions. At the exact midpoint between the two pK_a values, the trace amount of net negative and trace of net positive ions exactly balance, so that average net charge of all forms present is zero. This pH is

known as the isoelectric point pI , so $pI = \frac{1}{2}(pK_{a1} + pK_{a2})$. The individual amino acids all have slightly different pK_a values, so have different isoelectric points. For amino acids with charged side chains, the pK_a of the side chain is involved. Thus for Asp, Glu with negative side chains, $pI = \frac{1}{2}(pK_{a1} + pK_{aR})$, where pK_{aR} is the side chain pK_a . Cysteine also has potentially negative side chain with $pK_{aR} = 8.14$, so pI should be calculated as for Asp and Glu, even though the side chain is not significantly charged at neutral pH. For His, Lys, and Arg with positive side chains, $pI = \frac{1}{2}(pK_{aR} + pK_{a2})$. Amino acids have zero mobility in electrophoresis at their isoelectric point, although this behaviour is more usually exploited for peptides and proteins than single amino acids. Zwitterions have minimum solubility at their isoelectric point and some amino acids (in particular, with non-polar side chains) can be isolated by precipitation from water by adjusting the pH to the required isoelectric point.

OCCURRENCE AND FUNCTIONS IN BIOCHEMISTRY

Proteinogenic Amino Acids

Proteinogenic amino acids are amino acids that are incorporated biosynthetically into proteins during translation. The word “proteinogenic” means “protein creating”. Throughout known life, there are 22 genetically encoded (proteinogenic) amino acids, 20 in the standard genetic code and an additional 2 that can be incorporated by special translation mechanisms. In contrast, non-proteinogenic amino acids are amino acids that are either not incorporated into proteins (like GABA, L-DOPA, or triiodothyronine), misincorporated in place of a genetically encoded amino acid, or not produced directly and in isolation by standard cellular machinery (like hydroxyproline). The latter often results from post-translational modification of proteins. Some non-proteinogenic amino acids are incorporated into nonribosomal peptides which are synthesized by non-ribosomal peptide synthetases. Both eukaryotes and prokaryotes can incorporate selenocysteine into their proteins via a nucleotide sequence known as a SECIS element, which directs the cell to translate a nearby UGA codon as selenocysteine (UGA is normally a stop codon). In some methanogenic prokaryotes, the UAG codon (normally a stop codon) can also be translated to pyrrolysine.

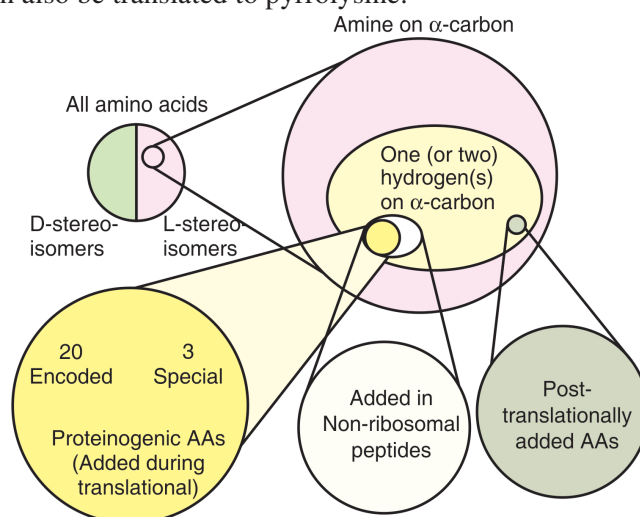


Fig. Proteinogenic amino acids are a small fraction of all amino acids.

In eukaryotes, there are only 21 proteinogenic amino acids, the 20 of the standard genetic code, plus selenocysteine. Humans can synthesize 12 of these from each other or from other molecules of intermediary metabolism. The other nine must be consumed (usually as their protein derivatives), and so they are called essential amino acids. The essential amino acids are histidine, isoleucine, leucine, lysine, methionine,

phenylalanine, threonine, tryptophan, and valine (*i.e.*, H, I, L, K, M, F, T, W, V). The proteinogenic amino acids have been found to be related to the set of amino acids that can be recognized by ribozyme autoaminoacylation systems. Thus, non-proteinogenic amino acids would have been excluded by the contingent evolutionary success of nucleotide-based life forms. Other reasons have been offered to explain why certain specific non-proteinogenic amino acids are not generally incorporated into proteins; for example, ornithine and homoserine cyclize against the peptide backbone and fragment the protein with relatively short half-lives, while others are toxic because they can be mistakenly incorporated into proteins, such as the arginine analog canavanine.

Structures

The following illustrates the structures and abbreviations of the 21 amino acids that are directly encoded for protein synthesis by the genetic code of eukaryotes. The structures given below are standard chemical structures, not the typical zwitterion forms that exist in aqueous solutions.

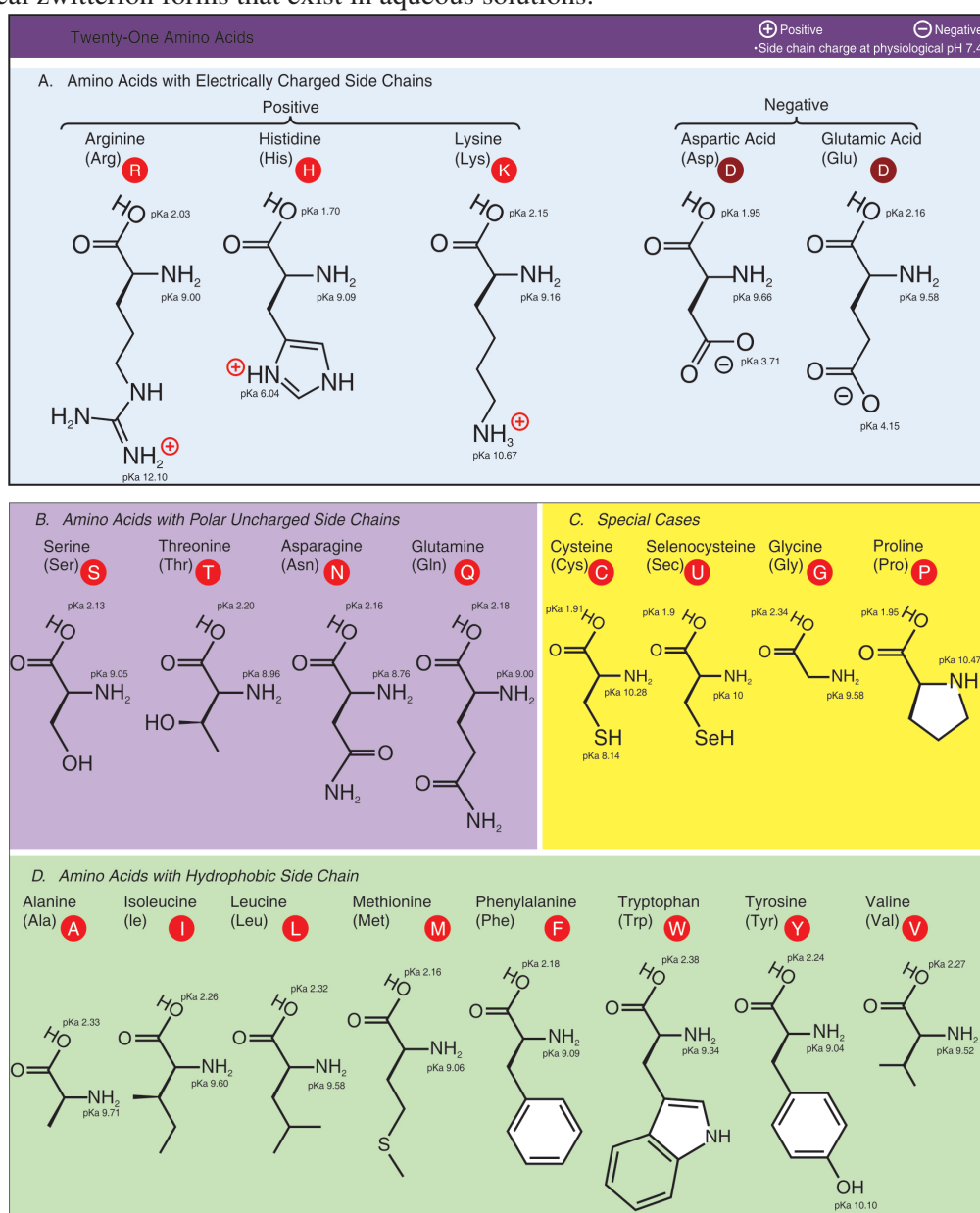
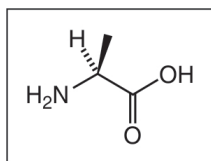
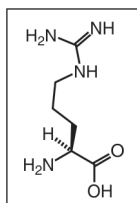
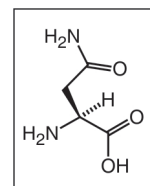
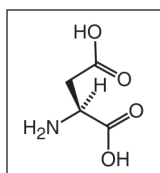
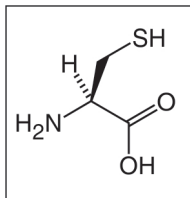
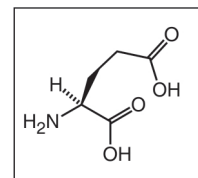
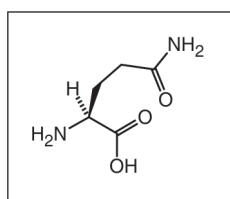
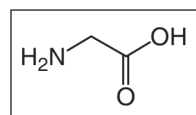
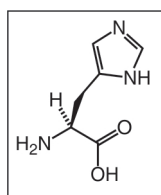
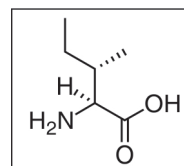
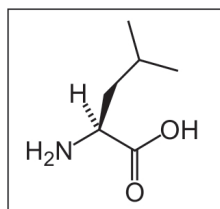
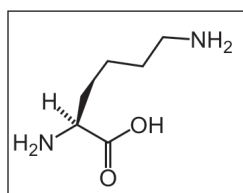
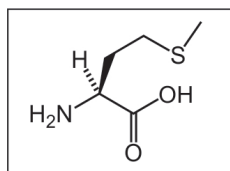
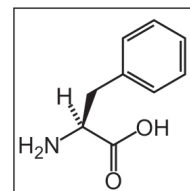
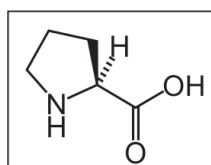
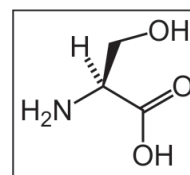


Fig. Grouped table of 21 amino acids' structures, nomenclature, and their side groups' pKa values.

**Fig.** L-Alanine (Ala / A).**Fig.** L-Arginine (Arg/R).**Fig.** L-Asparagine: (Asn/N).**Fig.** L-Aspartic acid (Asp/D).**Fig.** L-Cysteine (Cys/C).**Fig.** L-Glutamic acid (Glu/E).**Fig.** L-Glutamine (Gln/Q).**Fig.** Glycine (Gly/G).**Fig.** L-Histidine (His/H).**Fig.** L-Isoleucine (Ile/I).**Fig.** L-Leucine (Leu/L).**Fig.** L-Lysine (Lys/K).**Fig.** L-Methionine (Met/M).**Fig.** L-Phenylalanine (Phe/F).**Fig.** L-Proline (Pro/P).**Fig.** L-Serine (Ser/S).

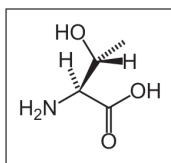


Fig. L-Threonine (Thr/T).

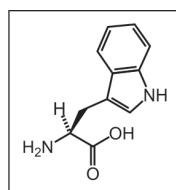


Fig. L-Tryptophan (Trp/W).

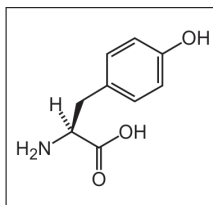


Fig. L-Tyrosine (Tyr/Y).

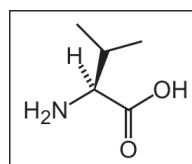


Fig. L-Valine (Val/V).

IUPAC/IUBMB now also recommends standard abbreviations for the following two amino acids:

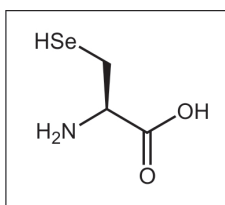


Fig. L-Selenocysteine (Sec/U).

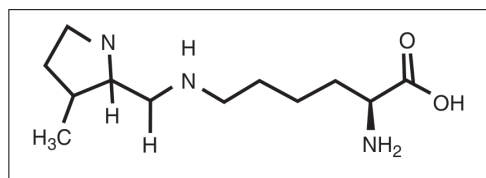


Fig. L-Pyrrolysine (Pyl/O).

Chemical Properties

Following is a table listing the one-letter symbols, the three-letter symbols, and the chemical properties of the side chains of the standard amino acids. The masses listed are based on weighted averages of the elemental isotopes at their natural abundances. Forming a peptide bond results in elimination of a molecule of water, so the mass of an amino acid unit within a protein chain is reduced by 18.01524 Da.

General chemical properties:

Amino acid	Short	Abbrev.	Avg. mass (Da)	pI	pK ₁ (α -COOH)	pK ₂ (ϵ -NH ₃ ⁺)
Alanine	A	Ala	89.09404	6.01	2.35	9.87
Cysteine	C	Cys	121.15404	5.05	1.92	10.70
Aspartic acid	D	Asp	133.10384	2.85	1.99	9.90
Glutamic acid	E	Glu	147.13074	3.15	2.10	9.47
Phenylalanine	F	Phe	165.19184	5.49	2.20	9.31
Glycine	G	Gly	75.06714	6.06	2.35	9.78
Histidine	H	His	155.15634	7.60	1.80	9.33
Isoleucine	I	Ile	131.17464	6.05	2.32	9.76
Lysine	K	Lys	146.18934	9.60	2.16	9.06
Leucine	L	Leu	131.17464	6.01	2.33	9.74
Methionine	M	Met	149.20784	5.74	2.13	9.28
Asparagine	N	Asn	132.11904	5.41	2.14	8.72
Pyrrolysine	O	Pyl	255.31			
Proline	P	Pro	115.13194	6.30	1.95	10.64
Glutamine	Q	Gln	146.14594	5.65	2.17	9.13
Arginine	R	Arg	174.20274	10.76	1.82	8.99
Serine	S	Ser	105.09344	5.68	2.19	9.21
Threonine	T	Thr	119.12034	5.60	2.09	9.10
Selenocysteine	U	Sec	168.053	5.47	1.91	10
Valine	V	Val	117.14784	6.00	2.39	9.74
Tryptophan	W	Trp	204.22844	5.89	2.46	9.41
Tyrosine	Y	Tyr	181.19124	5.64	2.20	9.21

Side Chain Properties

Amino acid	Short	Abbrev.	Side chain	Hydro-phobic	pKa	Polar	pH	Small	Tiny	Aromatic or Aliphatic	Van der Waals volume (Å ³)
Alanine	A	Ala	-CH ₃	X	-	-	-	X	X	Aliphatic	67
Cysteine	C	Cys	-CH ₂ SH	X	8.55	-	Acidic	X	X	-	86
Aspartic acid	D	Asp	-CH ₂ COOH	-	3.67	X	Acidic	X	-	-	91
Glutamic acid	E	Glu	-CH ₂ CH ₂ COOH	-	4.25	X	Acidic	-	-	-	109
Phenylalanine	F	Phe	-CH ₂ C ₆ H ₅	X	-	-	-	-	-	Aromatic	135
Glycine	G	Gly	-H	X	-	-	-	X	X	-	48
Histidine	H	His	-CH ₂ -C ₃ H ₃ N ₂	-	6.54	X	Weak basic	-	-	Aromatic	118
Isoleucine	I	Ile	-CH(CH ₃)CH ₂ CH ₃	X	-	-	-	-	-	Aliphatic	124
Lysine	K	Lys	-(CH ₂) ₄ NH ₂	-	10.40	X	Basic	-	-	-	135
Leucine	L	Leu	-CH ₂ CH(CH ₃) ₂	X	-	-	-	-	-	Aliphatic	124
Methionine	M	Met	-CH ₂ CH ₂ SCH ₃	X	-	-	-	-	-	Aliphatic	124
Asparagine	N	Asn	-CH ₂ CONH ₂	-	-	X	-	X	-	-	96
Pyrrolysine	O	Pyl	-(CH ₂) ₄ NHCOC ₄ H ₉ NCH ₃	-	N.D.	X	Weak basic	-	-	-	
Proline	P	Pro	-CH ₂ CH ₂ CH ₂ -	X	-	-	-	X	-	-	90
Glutamine	Q	Gln	-CH ₂ CH ₂ CONH ₂	-	-	X	-	-	-	-	114
Arginine	R	Arg	-(CH ₂) ₃ NH-C(NH)NH ₂	-	12.3	X	Strongly basic	-	-	-	148
Serine	S	Ser	-CH ₂ OH	-	-	X	-	X	X	-	73
Threonine	T	Thr	-CH(OH)CH ₃	-	-	X	-	X	-	-	93
Selenocysteine	U	Sec	-CH ₂ SeH	-	5.43	-	Acidic	X	X	-	
Valine	V	Val	-CH(CH ₃) ₂	X	-	-	-	X	-	Aliphatic	105
Tryptophan	W	Trp	-CH ₂ C ₈ H ₆ N	-	-	X	-	-	-	Aromatic	163

§: Values for Asp, Cys, Glu, His, Lys and Tyr were determined using the amino acid residue placed centrally in an alanine pentapeptide. The value for Arg is from Pace *et al.* (2009). The value for Sec is from Byun and Kang (2011). N.D.: The pKa value of Pyrrolysine has not been reported. *Note:* The pKa value of an amino-acid residue in a small peptide is typically slightly different when it is inside a protein. Protein pKa calculations are sometimes used to calculate the change in the pKa value of an amino-acid residue in this situation.

Gene Expression and Biochemistry

Amino acid	Short	Abbrev.	Codon(s)	Occurrence in Archaeal proteins (%)&	Occurrence in Bacteria proteins (%)&	Occurrence in Eukaryote proteins (%)&	Occurrence in human proteins (%)&	Essential‡ in humans
Alanine	A	Ala	GCU, GCC, GCA, GCG	8.2	10.06	7.63	7.01	No
Cysteine	C	Cys	UGU, UGC	0.98	0.94	1.76	2.3	Conditionally
Aspartic acid	D	Asp	GAU, GAC	6.21	5.59	5.4	4.73	No
Glutamic acid	E	Glu	GAA, GAG	7.69	6.15	6.42	7.09	Conditionally
Phenylalanine	F	Phe	UUU, UUC	3.86	3.89	3.87	3.65	Yes
Glycine	G	Gly	GGU, GGC, GGA, GGG	7.58	7.76	6.33	6.58	Conditionally
Histidine	H	His	CAU, CAC	1.77	2.06	2.44	2.63	Yes
Isoleucine	I	Ile	AUU, AUC, AUA	7.03	5.89	5.1	4.33	Yes
Lysine	K	Lys	AAA, AAG	5.27	4.68	5.64	5.72	Yes
Leucine	L	Leu	UUA, UUG, CUU, CUC, CUA, CUG	9.31	10.09	9.29	9.97	Yes
Methionine	M	Met	AUG	2.35	2.38	2.25	2.13	Yes

Asparagine	N	Asn	AAU, AAC	3.68	3.58	4.28	3.58	No
Pyrrolysine	O	Pyl	UAG*	0	0	0	0	No
Proline	P	Pro	CCU, CCC, CCA, CCG	4.26	4.61	5.41	6.31	No
Glutamine	Q	Gln	CAA, CAG	2.38	3.58	4.21	4.77	No
Arginine	R	Arg	CGU, CGC, CGA, CGG, AGA, AGG	5.51	5.88	5.71	5.64	Conditionally
Serine	S	Ser	UCU, UCC, UCA, UCG, AGU, AGC	6.17	5.85	8.34	8.33	No
Threonine	T	Thr	ACU, ACC, ACA, ACG	5.44	5.52	5.56	5.36	Yes
Selenocysteine	U	Sec	UGA**	0	0	0	>0	No
Valine	V	Val	GUU, GUC, GUA, GUG	7.8	7.27	6.2	5.96	Yes
Tryptophan	W	Trp	UGG	1.03	1.27	1.24	1.22	Yes
Tyrosine	Y	Tyr	UAU, UAC	3.35	2.94	2.87	2.66	Conditionally
Stop codon†	-	Term	UAA, UAG, UGA††				-	-

Note:

* UAG is normally the amber stop codon, but encodes pyrrolysine if a PYLIS element is present.

** UGA is normally the opal (or umber) stop codon, but encodes selenocysteine if a SECIS element is present.

† The stop codon is not an amino acid, but is included for completeness.

†† UAG and UGA do not always act as stop codons.

‡ An essential amino acid cannot be synthesized in humans and must, therefore, be supplied in the diet. Conditionally essential amino acids are not normally required in the diet, but must be supplied exogenously to specific populations that do not synthesize it in adequate amounts. and Occurrence of amino acids is based on 135 Archaea, 3775 Bacteria, 614 Eukaryota proteomes and human proteome (21,006 proteins) respectively.

Mass Spectrometry

In mass spectrometry of peptides and proteins, knowledge of the masses of the residues is useful. The mass of the peptide or protein is the sum of the residue masses plus the mass of water (Monoisotopic mass = 18.01056 Da; average mass = 18.0153 Da). The residue masses are calculated from the tabulated chemical formulas and atomic weights. In mass spectrometry, ions may also include one or more protons (Monoisotopic mass = 1.00728 Da; average mass = 1.0074 Da).

Amino Acid	Short	Abbrev.	Formula	Mon. Mass§ (Da)	Avg. Mass (Da)
Alanine	A	Ala	C ₃ H ₅ NO	71.03711	71.0779
Cysteine	C	Cys	C ₃ H ₅ NOS	103.00919	103.1429
Aspartic acid	D	Asp	C ₄ H ₅ NO ₃	115.02694	115.0874
Glutamic acid	E	Glu	C ₅ H ₇ NO ₃	129.04259	129.1140
Phenylalanine	F	Phe	C ₉ H ₉ NO	147.06841	147.1739
Glycine	G	Gly	C ₂ H ₃ NO	57.02146	57.0513
Histidine	H	His	C ₆ H ₇ N ₃ O	137.05891	137.1393
Isoleucine	I	Ile	C ₆ H ₁₁ NO	113.08406	113.1576
Lysine	K	Lys	C ₆ H ₁₂ N ₂ O	128.09496	128.1723
Leucine	L	Leu	C ₆ H ₁₁ NO	113.08406	113.1576
Methionine	M	Met	C ₅ H ₉ NOS	131.04049	131.1961
Asparagine	N	Asn	C ₄ H ₆ N ₂ O ₂	114.04293	114.1026
Pyrrolysine	O	Pyl	C ₁₂ H ₁₉ N ₃ O ₂	237.14773	237.2982
Proline	P	Pro	C ₅ H ₇ NO	97.05276	97.1152
Glutamine	Q	Gln	C ₅ H ₈ N ₂ O ₂	128.05858	128.1292
Arginine	R	Arg	C ₆ H ₁₂ N ₄ O	156.10111	156.1857
Serine	S	Ser	C ₃ H ₅ NO ₂	87.03203	87.0773
Threonine	T	Thr	C ₄ H ₇ NO ₂	101.04768	101.1039
Selenocysteine	U	Sec	C ₃ H ₅ NOSe	150.95364	150.0489
Valine	V	Val	C ₅ H ₉ NO	99.06841	99.1311
Tryptophan	W	Trp	C ₁₁ H ₁₀ N ₂ O	186.07931	186.2099
Tyrosine	Y	Tyr	C ₉ H ₉ NO ₂	163.06333	163.1733

- § Monoisotopic mass

Stoichiometry and Metabolic Cost in Cell

The table below lists the abundance of amino acids in *E. coli* cells and the metabolic cost (ATP) for synthesis the amino acids. Negative numbers indicate the metabolic processes are energy favorable and do not cost net ATP of the cell. The abundance of amino acids includes amino acids in free form and in polymerization form (proteins).

Amino acid	Abundance (# of molecules ($\times 10^8$) per <i>E. coli</i> cell)	ATP cost in synthesis under aerobic condition	ATP cost in synthesis under anaerobic condition
Alanine	2.9	-1	1
Cysteine	0.52	11	15
Aspartic acid	1.4	0	2
Glutamic acid	1.5	-7	-1
Phenylalanine	1.1	-6	2
Glycine	3.5	-2	2
Histidine	0.54	1	7
Isoleucine	1.7	7	11
Lysine	2.0	5	9
Leucine	2.6	-9	1
Methionine	0.88	21	23
Asparagine	1.4	3	5
Proline	1.3	-2	4
Glutamine	1.5	-6	0
Arginine	1.7	5	13
Serine	1.2	-2	2
Threonine	1.5	6	8
Tryptophan	0.33	-7	7
Tyrosine	0.79	-8	2
Valine	2.4	-2	2

Catabolism

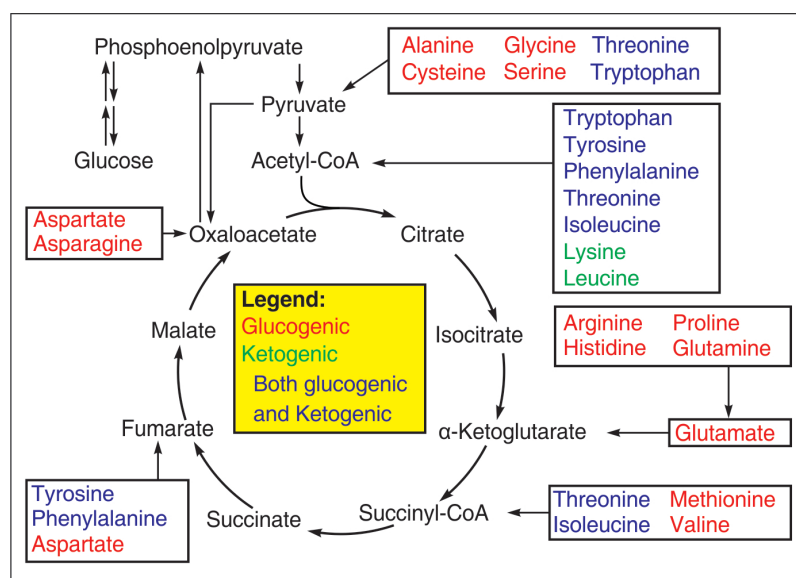


Fig. Amino acids can be classified according to the properties of their main products as either of:

- Glucogenic, with the products having the ability to form glucose by gluconeogenesis
- Ketogenic, with the products not having the ability to form glucose: These products may still be used for ketogenesis or lipid synthesis.
- Amino acids catabolized into both glucogenic and ketogenic products.

Remarks

Amino Acid	Abbrev.		Remarks
Alanine	A	Ala	Very abundant and very versatile, it is more stiff than glycine, but small enough to pose only small steric limits for the protein conformation. It behaves fairly neutrally, and can be located in both hydrophilic regions on the protein outside and the hydrophobic areas inside.
Asparagine or aspartic acid	B	Asx	A placeholder when either amino acid may occupy a position
Cysteine	C	Cys	The sulfur atom bonds readily to heavy metal ions. Under oxidizing conditions, two cysteines can join together in a disulfide bond to form the amino acid cystine. When cystines are part of a protein, insulin for example, the tertiary structure is stabilized, which makes the protein more resistant to denaturation; therefore, disulfide bonds are common in proteins that have to function in harsh environments including digestive enzymes (e.g., pepsin and chymotrypsin) and structural proteins (e.g., keratin). Disulfides are also found in peptides too small to hold a stable shape on their own (e.g. insulin).
Aspartic acid	D	Asp	Asp behaves similarly to glutamic acid, and carries a hydrophilic acidic group with strong negative charge. Usually, it is located on the outer surface of the protein, making it water-soluble. It binds to positively charged molecules and ions, and is often used in enzymes to fix the metal ion. When located inside of the protein, aspartate and glutamate are usually paired with arginine and lysine.
Glutamic acid	E	Glu	Glu behaves similarly to aspartic acid, and has a longer, slightly more flexible side chain.
Phenylalanine	F	Phe	Essential for humans, phenylalanine, tyrosine, and tryptophan contain a large, rigid aromatic group on the side chain. These are the biggest amino acids. Like isoleucine, leucine, and valine, these are hydrophobic and tend to orient towards the interior of the folded protein molecule. Phenylalanine can be converted into tyrosine.
Phenylalanine	F	Phe	Essential for humans, phenylalanine, tyrosine, and tryptophan contain a large, rigid aromatic group on the side chain. These are the biggest amino acids. Like isoleucine, leucine, and valine, these are hydrophobic and tend to orient towards the interior of the folded protein molecule. Phenylalanine can be converted into tyrosine.
Glycine	G	Gly	Because of the two hydrogen atoms at the α carbon, glycine is not optically active. It is the smallest amino acid, rotates easily, and adds flexibility to the protein chain. It is able to fit into the tightest spaces, e.g., the triple helix of collagen. As too much flexibility is usually not desired, as a structural component, it is less common than alanine.
Histidine	H	His	His is essential for humans. In even slightly acidic conditions, protonation of the nitrogen occurs, changing the properties of histidine and the polypeptide as a whole. It is used by many proteins as a regulatory mechanism, changing the conformation and behavior of the polypeptide in acidic regions such as the late endosome or lysosome, enforcing conformation change in enzymes. However, only a few histidines are needed for this, so it is comparatively scarce.
Isoleucine	I	Ile	Ile is essential for humans. Isoleucine, leucine, and valine have large aliphatic hydrophobic side chains. Their molecules are rigid, and their mutual hydrophobic interactions are important for the correct folding of proteins, as these chains tend to be located inside of the protein molecule.
Leucine or isoleucine	J	Xle	A placeholder when either amino acid may occupy a position
Lysine	K	Lys	Lys is essential for humans, and behaves similarly to arginine. It contains a long, flexible side chain with a positively charged end. The flexibility of the chain makes lysine and arginine suitable for binding to molecules with many negative charges on their surfaces. E.g., DNA-binding proteins have their active regions rich with arginine and lysine. The strong charge makes these two amino acids prone to be located on the outer hydrophilic surfaces of the proteins; when they are found inside, they are usually paired with a corresponding negatively charged amino acid, e.g., aspartate or glutamate.
Leucine	L	Leu	Leu is essential for humans, and behaves similarly to isoleucine and valine.

Life Based on Alternative Proteinogenic Sets

The proteinogenic set used by known life on Earth appears to be arbitrarily selected by evolution, according to current knowledge, from many hundreds of possible alpha-type amino acids. Xenobiology studies hypothetical life forms that could be constructed using alternative sets using expanded genetic codes. Miller-type experiments on artificial abiogenesis show that alpha-type amino acids predominate in water-based ‘primordial soups’, but beta-type amino acids dominate when less water is present. Both alpha- and beta-based sets could form the basis for alternative protein constructions and life forms.

NON-PROTEINOGENIC AMINO ACIDS

In biochemistry, non-coded or non-proteinogenic amino acids are those not naturally encoded or found in the genetic code of any organism. Despite the use of only 22 amino acids (21 in eukaryotes) by the translational machinery to assemble proteins (the proteinogenic amino acids), over 140 amino acids are known to occur naturally in proteins and thousands more may occur in nature or be synthesized in the laboratory.

Many non-proteinogenic amino acids are noteworthy because they are:

- Intermediates in biosynthesis,
- Post-translationally formed in proteins,
- Possess a physiological role (*e.g.*, Components of bacterial cell walls, neurotransmitters and toxins),
- Natural or man-made pharmacological compounds,
- Present in meteorites and in prebiotic experiments (*e.g.*, Miller–Urey experiment).

Definition by Negation

Technically, any organic compound with an amine ($-\text{NH}_2$) and a carboxylic acid ($-\text{COOH}$) functional group is an amino acid. The proteinogenic amino acids are a small subset of this group that possess a central carbon atom (α - or 2-) bearing an amino group, a carboxyl group, a side chain and an α -hydrogen in a *levo* conformation, with the exception of glycine, which is achiral, and proline, whose amine group is a secondary amine and is consequently frequently referred to as an imino acid for traditional reasons, albeit not an imino.

The genetic code encodes 20 standard amino acids for incorporation into proteins during translation. However, there are two extra proteinogenic amino acids: selenocysteine and pyrrolysine. These non-standard amino acids do not have a dedicated codon, but are added in place of a stop codon when a specific sequence is present, UGA codon and SECIS element for selenocysteine, UAG PYLIS downstream sequence for pyrrolysine. All other amino acids are termed “non-proteinogenic”.

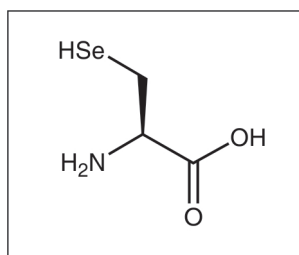


Fig. Selenocysteine. This amino acid contains a selenol group on its β -carbon.

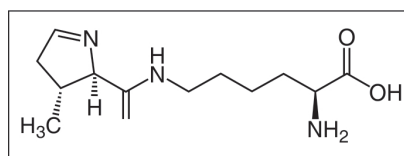


Fig. Pyrrolysine. This amino acid is formed by joining to the ϵ -amino group of lysine a carboxylated pyrroline ring.

There are various groups of amino acids:

- 20 standard amino acids
- 22 proteinogenic amino acids
- Over 80 amino acids created abiotically in high concentrations
- About 900 are produced by natural pathways
- Over 118 engineered amino acids have been placed into protein

These groups overlap, but are not identical. All 22 proteinogenic amino acids are biosynthesised by organisms and some, but not all, of them also are abiotic (found in prebiotic experiments and meteorites). Some natural amino acids, such as norleucine, are misincorporated translationally into proteins due to infidelity of the protein-synthesis process. Many amino acids, such as ornithine, are metabolic intermediates produced biosynthetically, but not incorporated translationally into proteins. Post-translational modification of amino-acid residues in proteins leads to the formation of many proteinaceous, but non-proteinogenic, amino acids. Other amino acids are solely found in abiotic mixes (*e.g.*, α -methylnorvaline). Over 30 unnatural amino acids have been inserted translationally into protein in engineered systems, yet are not biosynthetic.

Nomenclature

In addition to the IUPAC numbering system to differentiate the various carbons in an organic molecule, by sequentially assigning a number to each carbon, including those forming a carboxylic group, the carbons along the side-chain of amino acids can also be labelled with Greek letters, where the α -carbon is the central chiral carbon possessing a carboxyl group, a side chain and, in α -amino acids, an amino group—the carbon in carboxylic groups is not counted. (Consequently, the IUPAC names of many non-proteinogenic α -amino acids start with 2-amino-and end in-ic acid.)

Natural, but Non- L- α -amino Acids

Most natural amino acids are α -amino acids in the L conformation, but some exceptions exist.

Non-alpha

Some non- α amino acids exist in organisms. In these structures, the amine group displaced further from the carboxylic acid end of the amino acid molecule. Thus a β amino acid has the amine group bonded to the second carbon away, and a γ amino acid has it on the third. Examples include β -alanine, GABA, and δ -aminolevulinic acid.

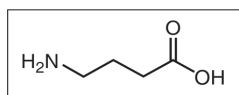


Fig. β -alanine: an amino acid produced by aspartate 1-decarboxylase and a precursor to coenzyme A and the peptides carnosine and anserine.

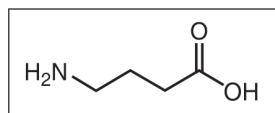


Fig. γ -Aminobutyric acid (GABA): A neurotransmitter in animals.

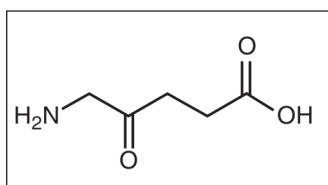


Fig. δ-Aminolevulinic acid: an intermediate in tetrapyrrole biosynthesis (haem, chlorophyll, cobalamin *etc.*).

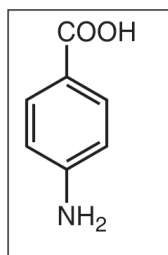


Fig. 4-Aminobenzoic acid (PABA): An intermediate in folate biosynthesis

The reason why α-amino acids are used in proteins has been linked to their frequency in meteorites and prebiotic experiments. An initial speculation on the deleterious properties of β-amino acids in terms of secondary structure, turned out to be incorrect.

D-amino Acids

Some amino acids contain the opposite absolute chirality, chemicals that are not available from normal ribosomal translation/transcription machinery. Most bacterial cells walls are formed by peptidoglycan, a polymer composed of amino sugars crosslinked with short oligopeptides bridged between each other. The oligopeptide is non-ribosomally synthesised and contains several peculiarities including D-amino acids, generally D-alanine and D-glutamate.

A further peculiarity is that the former is racemised by a PLP-binding enzymes (encoded by *alr* or the homologue *dadX*), whereas the latter is racemised by a cofactor independent enzyme (*murI*). Some variants are present, in *Thermotoga* spp. D-lysine is present and in certain vancomycin-resistant bacteria D-serine is present (*vanT* gene). In animals, some D-amino acids are neurotransmitters.

Without a Hydrogen on the α-carbon

All proteinogenic amino acids have at least one hydrogen on the α-carbon. Glycine has two hydrogens, and all others have one hydrogen and one side-chain. Replacement of the remaining hydrogen with a larger substituent, such as a methyl group, distorts the protein backbone. In some fungi α-amino isobutyric acid is produced as a precursor to peptides, some of which exhibit antibiotic properties. This compound is similar to alanine, but possesses an additional methyl group on the α-carbon instead of a hydrogen. It is therefore achiral. Another compound similar to alanine without an α-hydrogen is dehydroalanine, which possess a methylene sidechain. It is one of several naturally occurring dehydroamino acids.

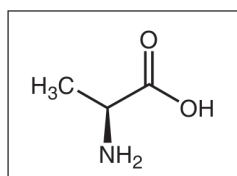


Fig. Alanine.

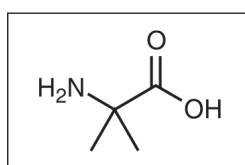


Fig. Aminoisobutyric Acid.

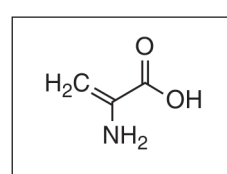


Fig. Dehydroalanine.

Twin Amino Acid Stereocentres

A subset of L- α -amino acids are ambiguous as to which of two ends is the α -carbon. In proteins a cysteine residue can form a disulfide bond with another cysteine residue, thus crosslinking the protein. Two crosslinked cysteines form a cystine molecule. Cysteine and methionine are generally produced by direct sulfurylation, but in some species they can be produced by transsulfuration, where the activated homoserine or serine is fused to a cysteine or homocysteine forming cystathionine. A similar compound is lanthionine, which can be seen as two alanine molecules joined via a thioether bond and is found in various organisms. Similarly, djenkolic acid, a plant toxin from jengkol beans, is composed of two cysteines connected by a methylene group. Diaminopimelic acid is both used as a bridge in peptidoglycan and is used a precursor to lysine (via its decarboxylation).

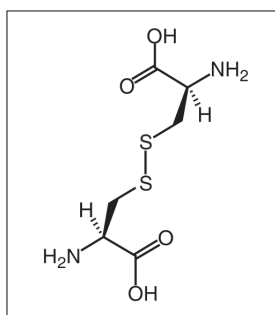


Fig. Cystine.

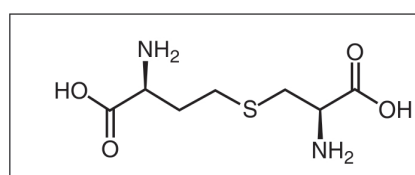


Fig. Cystathionine.

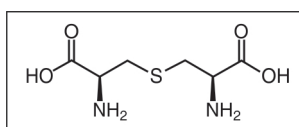


Fig. Lanthionine.

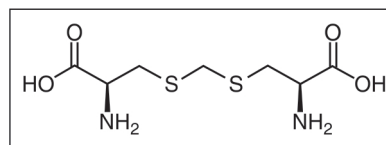


Fig. Djenkolic acid.

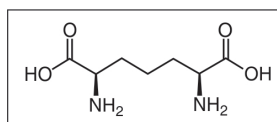


Fig. Diaminopimelic acid.

Prebiotic Amino Acids and Alternative Biochemistries

In meteorites and in prebiotic experiments (*e.g.*, Miller–Urey experiment) many more amino acids than the twenty standard amino acids are found, several of which at higher concentrations than the standard ones: It has been conjectured that if amino acid based life were to arise in parallel elsewhere in the universe, no more than 75% of the amino acids would be in common. The most notable anomaly is the lack of aminobutyric acid.

Proportion of amino acids relative to glycine (%)		
Molecule	Electric discharge	Murchinson meteorite
Glycine	100	100
Alanine	180	36
α -Amino-n-butyric acid	61	19
Norvaline	14	14
Valine	4.4	
Norleucine	1.4	
Leucine	2.6	
Isoleucine	1.1	
Alloisoleucine	1.2	
t-leucine	< 0.005	

α -Amino-n-heptanoic acid	0.3	
Proline	0.3	22
Pipecolic acid	0.01	11
α,β -diaminopropionic acid	1.5	
α,γ -diaminobutyric acid	7.6	
Ornithine	< 0.01	
lysine	< 0.01	
Aspartic acid	7.7	13
Glutamic acid	1.7	20
Serine	1.1	
Threonine	0.2	
Allothreonine	0.2	
Methionine	0.1	
Homocysteine	0.5	
Homoserine	0.5	
β -Alanine	4.3	10
β -Amino-n-butyric acid	0.1	5
β -Aminoisobutyric acid	0.5	7
γ -Aminobutyric acid	0.5	7
α -Aminoisobutyric acid	7	33
isovaline	1	11
Sarcosine	12.5	7
N-ethyl glycine	6.8	6
N-propyl glycine	0.5	
N-isopropyl glycine	0.5	
N-methyl alanine	3.4	3
N-ethyl alanine	< 0.05	
N-methyl β -alanine	1.0	
N-ethyl β -alanine	< 0.05	
isoserine	1.2	
α -hydroxy- γ -aminobutyric acid	17	

Straight Side Chain

The genetic code has been described as a frozen accident and the reasons why there is only one standard amino acid with a straight chain (alanine) could simply be redundancy with valine, leucine and isoleucine. However, straight chained amino acids are reported to form much more stable alpha helices.

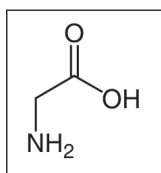


Fig. Glycine (Hydrogen side-chain).

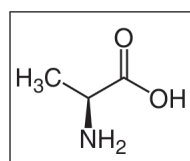


Fig. Alanine (Methyl side-chain).

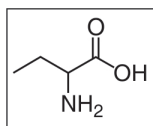


Fig. α -aminobutyric acid (Ethyl side-chain).

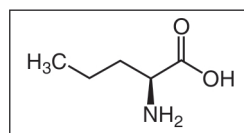


Fig. Norvaline (*n*-Propyl side-chain).

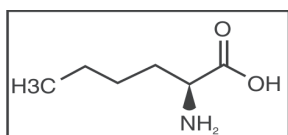


Fig. Norleucine (*n*-Butyl side-chain).

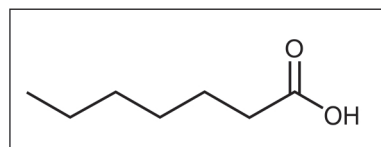


Fig. Homonorleucine (*n*-Pentyl side-chain).

Chalcogen

Serine, homoserine, O-methyl-homoserine and O-ethyl-homoserine possess an hydroxymethyl, hydroxyethyl, O-methyl-hydroxymethyl and O-methyl-hydroxyethyl side chain. Whereas cysteine, homocysteine, methionine and ethionine possess the thiol equivalents. The selenol equivalents are selenocysteine, selenohomocysteine, selenomethionine and selenoethionine. Amino acids with the next chalcogen down are also found in nature: several species such as *Aspergillus fumigatus*, *Aspergillus terreus*, and *Penicillium chrysogenum* in the absence of sulfur are able to produce and incorporate into protein tellurocysteine and telluromethionine. Hydroxyglycine, an amino acid with a hydroxyl side-chain, is highly unstable.

Expanded Genetic Code

An expanded genetic code is an artificially modified genetic code in which one or more specific codons have been re-allocated to encode an amino acid that is not among the 20 common naturally-encoded proteinogenic amino acids.

The key prerequisites to expand the genetic code are:

- The non-standard amino acid to encode,
- An unused codon to adopt,
- A tRNA that recognises this codon, and
- A tRNA synthetase that recognises only that tRNA and only the non-standard amino acid.

Expanding the genetic code is an area of research of synthetic biology, an applied biological discipline whose goal is to engineer living systems for useful purposes. The genetic code expansion enriches the repertoire of useful tools available to science.

It is noteworthy that the genetic code for all organisms is basically the same, so that all living beings use the same 'genetic language'. In general, the introduction of new functional unnatural amino acids into proteins of living cells breaks the universality of the genetic language, which ideally leads to alternative life forms. Proteins are produced thanks to the translational system molecules, which decode the RNA messages into a string of amino acids. The translation of genetic information contained in messenger RNA (mRNA) into a protein is catalysed by ribosomes. Transfer RNAs (tRNA) are used as keys to decode the mRNA into its encoded polypeptide. The tRNA recognizes a specific three nucleotide codon in the mRNA with a complementary sequence called the anticodon on one of its loops. Each three nucleotide codon is translated into one of twenty naturally occurring amino acids. There is at least one tRNA for any codon, and sometimes multiple codons code for the same amino acid.

Many tRNAs are compatible with several codons. An enzyme called an aminoacyl tRNA synthetase covalently attaches the amino acid to the appropriate tRNA. Most cells have a different synthetase for each amino acid (20 or more synthetases). On the other hand, some bacteria have fewer than 20 aminoacyl tRNA synthetases, and introduce the "missing" amino acid(s) by modification of a structurally related amino acid by an aminotransferase enzyme. A feature exploited in the expansion of the genetic code is the fact that the aminoacyl tRNA synthetase often does not recognize the anticodon, but another part of the tRNA, meaning that if the anticodon were to be mutated the encoding of that amino acid would change to a new codon. In the ribosome, the information in mRNA is translated into a specific amino acid when the mRNA codon matches with the complementary anticodon of a tRNA, and the attached amino acid is added onto a growing polypeptide chain. When it is released from the ribosome, the polypeptide chain folds into a functioning protein.

In order to incorporate a novel amino acid into the genetic code several changes are required. First, for successful translation of a novel amino acid, the codon to which the novel amino acid is assigned cannot already code for one of the 20 natural amino acids. Usually a nonsense codon (stop codon) or a four-base codon

are used. Second, a novel pair of tRNA and aminoacyl tRNA synthetase are required, these are called the orthogonal set. The orthogonal set must not crosstalk with the endogenous tRNA and synthetase sets, while still being functionally compatible with the ribosome and other components of the translation apparatus. The active site of the synthetase is modified to accept only the novel amino acid. Most often, a library of mutant synthetases is screened for one which charges the tRNA with the desired amino acid. The synthetase is also modified to recognize only the orthogonal tRNA. The tRNA synthetase pair is often engineered in other bacteria or eukaryotic cells. In this area of research, the 20 encoded proteinogenic amino acids are referred to as standard amino acids, or alternatively as natural or canonical amino acids, while the added amino acids are called non-standard amino acids (NSAAs), or unnatural amino acids (uAAs; term not used in papers dealing with natural non-proteinogenic amino acids, such as phosphoserine), or non-canonical amino acids.

Non-standard Amino Acids

The first element of the system is the amino acid that is added to the genetic code of a certain strain of organism. Over 71 different NSAAs have been added to different strains of *E. coli*, yeast or mammalian cells. Due to technical details (easier chemical synthesis of NSAAs, less crosstalk and easier evolution of the aminoacyl-tRNA synthase), the NSAAs are generally larger than standard amino acids and most often have a phenylalanine core but with a large variety of different substituents. These allow a large repertoire of new functions, such as labelling as a fluorescent reporter (*e.g.*, dansylalanine) or to produce translationally protein in *E. coli* with Eukaryotic post-translational modifications (*e.g.*, phosphoserine, phosphothreonine, and phosphotyrosine).

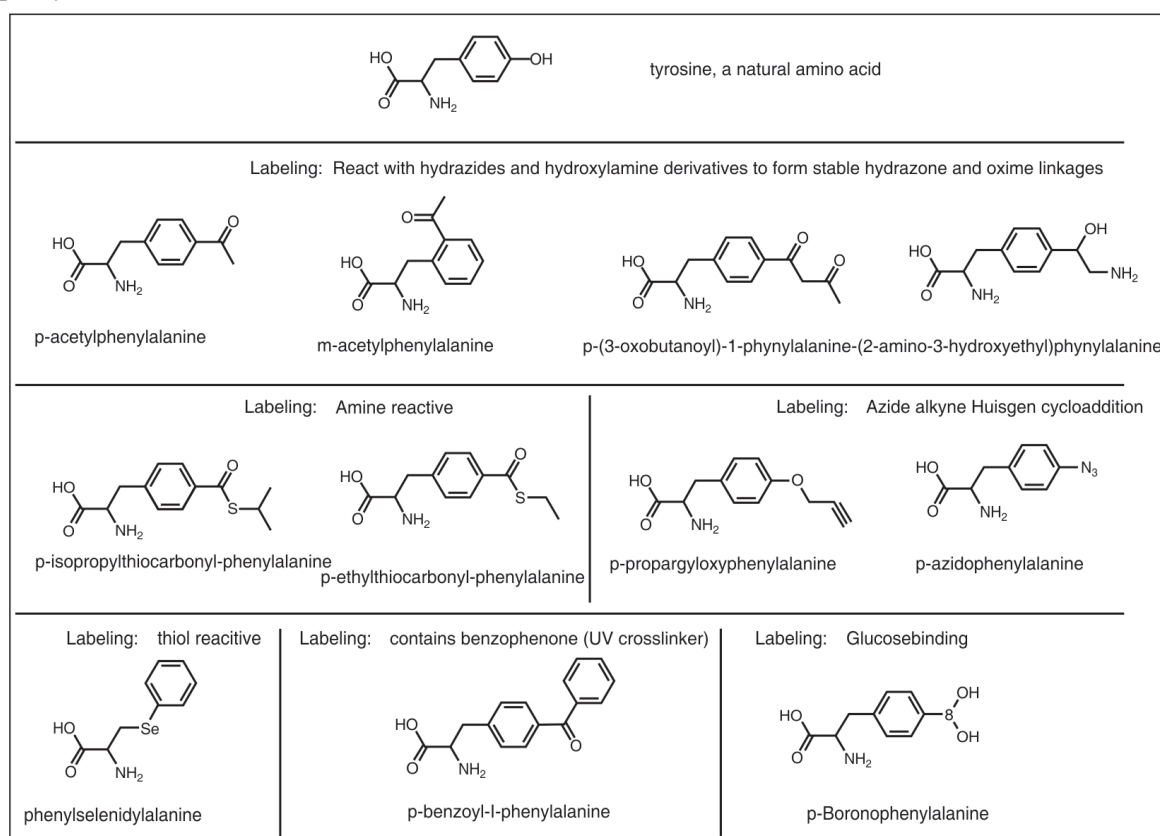


Fig. Tyrosine and some synthetic tyrosine variants used for protein labelling.

Different variants of tyrosine have been synthesized and can be incorporated into proteins using an expanded genetic code. The variants depicted here are all used for chemical or photochemical linking.

This means that the incorporated AA specifically reacts with either a particular chemical group (such as hydrazides, amines, azides, or thiols) or can be UV-activated to crosslink with other AAs.

Unnatural amino acids incorporated into proteins include heavy atom containing amino acids to facilitate certain x-ray crystallographic studies; amino acids with novel steric/packing and electronic properties; photocrosslinking amino acids which can be used to probe protein-protein interactions *in vitro* or *in vivo*; keto, acetylene, azide, and boronate containing amino acids which can be used to selectively introduce a large number of biophysical probes, tags, and novel chemical functional groups into proteins *in vitro* or *in vivo*; redox active amino acids to probe and modulate electron transfer; photocaged and photoisomerizable amino acids to photoregulate biological processes; metal binding amino acids for catalysis and metal ion sensing; amino acids that contain fluorescent or infra-red active side chains to probe protein structure and dynamics; α -hydroxy acids and D-amino acids as probes of backbone conformation and hydrogen bonding interactions; and sulfated amino acids and mimetics of phosphorylated amino acids as probes of posttranslational modifications.

Availability of the non-standard amino acid requires that the organism either import it from the medium or biosynthesised it. In the first case, the unnatural amino acid is first synthesised chemically in optically pure L-form. It is then added to the growth medium of the cell. Generally a library of compounds is tested to see which can be imported and incorporated, but often the various transport systems can handle unnatural amino acids with apolar side-chains. In the second case, a biosynthetic paths need to be engineered. One example is an *E. coli* strain that biosynthesizes a novel, previously unnatural amino acid (p-aminophenylalanine) from basic carbon sources and includes this amino acid in its genetic code. Another example is the production of phosphoserine, which is a natural metabolite and as a consequence the pathway flux had to be altered to increase its production.

Codon Assignment

Another element of the system is a codon to allocate to the new amino acid. A major problem for the genetic code expansion is that there are no free codons. The genetic code has a nonrandom layout that shows tell-tale signs of various phases of primordial evolution, however, it has since, frozen into place and is near-universally conserved. Nevertheless, some codons are rarer than others. In fact, in *E. coli* (and all organisms) the codon usage is not equal, but presents several rare codons, the rarest being the amber stop codon (UAG).

Amber Codon Suppression

The possibility of reassigning codons was realized by Normanly *et al.* in 1990, when a viable mutant strain of *E. coli* read through the UAG (“amber”) stop codon. This was possible thanks to the rarity of this codon and the fact that release factor 1 alone makes the amber codon terminate translation. Later, in the Schultz lab, the tRNA^{Tyr}/tyrosyl-tRNA synthetase (TyrRS) from *Methanococcus jannaschii*, an archaeobacterium, was used to introduce a tyrosine instead of STOP, the default value of the amber codon. This was possible because of the differences between the endogenous bacterial synthases and the orthologous archaeal synthase, which do not recognize each other. Subsequently, the group evolved the orthogonal tRNA/synthase pair to utilise the non-standard amino acid *O*-methyltyrosine. This was followed by the larger naphthylalanine and the photocrosslinking benzoylphenylalanine, which proved the potential utility of the system.

The amber codon is the least used codon in *Escherichia coli*, but hijacking it results in a substantial loss of fitness. One study in fact found that there were at least 83 peptides majorly affected by the readthrough. Additionally, the labelling was incomplete. As a consequence, several strains have been made to reduce the fitness cost, including the removal of all amber codons from the genome. In most *E. coli* K-12 strains (*viz.*, *Escherichia coli* (molecular biology) for strain pedigrees) there are 314 UAG stop codons. Consequently, a gargantuan amount of work has gone into the replacement of these. One approach pioneered by the group of Prof. George Church from Harvard, was dubbed MAGE in CAGE: this relied on a multiplex transformation and subsequent strain recombination to remove all UAG codons—the latter part presented a halting point in a first

paper, but was overcome. This resulted in the *E. coli* strain C321. DA, which lacks all UAG codons and RF1. This allowed an experiment to be done with this strain to make it “addicted” to the amino acid biphenylalanine by evolving several key enzymes to require it structurally, therefore putting its expanded genetic code under positive selection.

Rare Sense Codon Reassignment

In addition to the amber codon, rare sense codons have also been considered for use. The AGG codon codes for arginine, but a strain has been successfully modified to make it code for 6-*N*-allyloxycarbonyl-lysine. Another candidate is the AUA codon, which is unusual in that its respective tRNA has to differentiate against AUG that codes for methionine (primordially, isoleucine, hence, its location). In order to do this, the AUA tRNA has a special base, lysidine. The deletion of the synthase (*tilS*) was possible thanks to the replacement of the native tRNA with that of *Mycoplasma mobile* (no lysidine). The reduced fitness is a first step towards pressuring the strain to lose all instances of AUA, allowing it to be used for genetic code expansion.

Four Base Codons

Other approaches include the addition of extra base pairing or the use of orthologous ribosomes that accept in addition to the regular triplet genetic code, tRNAs with quadruple code. This allowed the simultaneous usage of two unnatural amino acids, *p*-azidophenylalanine (AzPhe) and N6-[(2-propynyloxy)carbonyl]lysine (CAK), which cross-link with each other by Huisgen cycloaddition.

tRNA/synthase Pair

Another key element is the tRNA/synthase pair. The orthologous set of synthetase and tRNA can be mutated and screened through directed evolution to charge the tRNA with a different, even novel, amino acid. Mutations to the plasmid containing the pair can be introduced by error-prone PCR or through degenerate primers for the synthetase’s active site. Selection involves multiple rounds of a two-step process, where the plasmid is transferred into cells expressing chloramphenicol acetyl transferase with a premature amber codon.

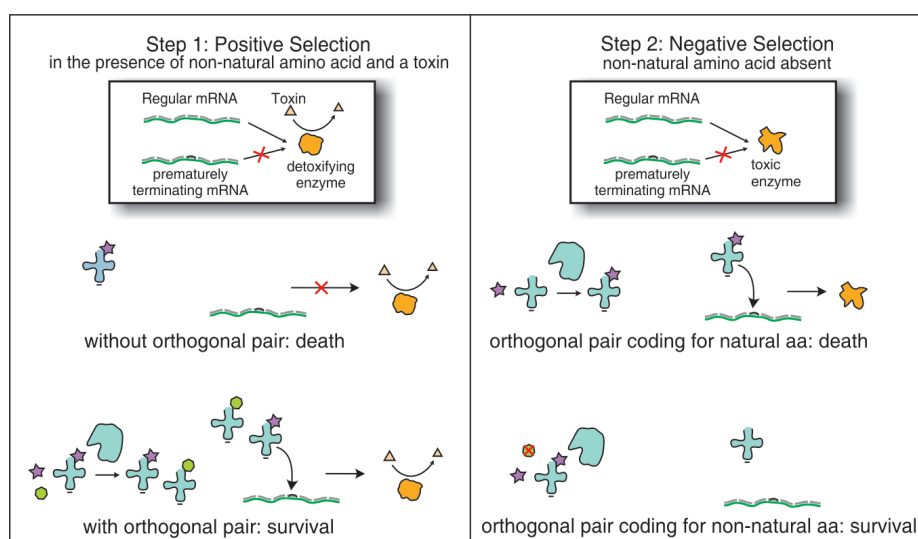


Fig. Several methods for selecting the synthetase that accepts only the non-natural amino acid have been developed. One of which is by using a combination of positive and negative selection.

In the presence of toxic chloramphenicol and the non-natural amino acid, the surviving cells will have overridden the amber codon using the orthogonal tRNA aminoacylated with either the standard amino acids or the non-natural one. To remove the former, the plasmid is inserted into cells with a barnase gene (toxic) with a premature amber

codon but without the non-natural amino acid, removing all the orthogonal synthetases that do not specifically recognize the non-natural amino acid. In addition to the recoding of the tRNA to a different codon, they can be mutated to recognize a four-base codon, allowing additional free coding options. The non-natural amino acid, as a result, introduces diverse physicochemical and biological properties in order to be used as a tool to explore protein structure and function or to create novel or enhanced protein for practical purposes.

Orthogonal Sets in *E.coli*

The orthogonal pairs of synthetase and tRNA that work for one organism may not work for another, as the synthetase may mis-aminoacylate endogenous tRNAs or the tRNA be mis-aminoacylated itself by an endogenous synthetase. As a result, the sets created to date differ between organisms.

- tRNA^{Tyr}-TyrRS pair from the archaeon *Methanococcus jannaschii*
- tRNA^{Lys}-LysRS pair from the archaeon *Pyrococcus horikoshii*
- tRNA^{Glu}-GluRS pair from *Methanosarcina mazei*
- leucyl-tRNA synthetase from *Methanobacterium thermoautotrophicum* and a mutant leucyl tRNA derived from *Halobacterium* sp
- tRNA^{Amber}-PylRS pair from the archaeon *Methanosarcina barkeri* and *Methanosarcina mazei*
- tRNA^{Amber}-3-Iodotyrosyl-tRNA synthetase (variant of aminoacyl-tRNA synthetase from *Methanocaldococcus jannaschii*).

Orthogonal Sets in Yeast

- tRNA^{Tyr}-TyrRS pair from *Escherichia coli*
- tRNA^{Leu}-LeuRS pair from *Escherichia coli*
- tRNA^{iMet} from human and GlnRS from *Escherichia coli*
- tRNA^{Amber}-PylRS pair from the archaeon *Methanosarcina barkeri* and *Methanosarcina mazei*
- tRNA^{Amber}-4,5-dimethoxy-2-nitrobenzyl-cysteinyl-tRNA synthetase.

Orthogonal Sets in Mammalian Cells

- tRNA^{Tyr}-TyrRS pair from *Bacillus stearothermophilus*
- modified tRNA^{Trp}-TrpRS pair from *Bacillus subtilis* trp
- tRNA^{Leu}-LeuRS pair from *Escherichia coli*
- tRNA^{Amber}-PylRS pair from the archaeon *Methanosarcina barkeri* and *Methanosarcina mazei*
- In 2017 a mouse engineered with an extended genetic code that can produce proteins with unnatural amino acids was reported.

Orthogonal Ribosomes

Similarly to orthogonal tRNAs and aminoacyl tRNA synthetases (aaRSs), orthogonal ribosomes have been engineered to work in parallel to the natural ribosomes. Orthogonal ribosomes ideally use different mRNA transcripts than their natural counterparts and ultimately should draw on a separate pool of tRNA as well. This should alleviate some of the loss of fitness which currently still arises from techniques such as Amber codon suppression. Additionally, orthogonal ribosomes can be mutated and optimized for particular tasks, like the recognition of quadruplet codons. Such an optimization is not possible, or highly disadvantageous for natural ribosomes.

o-Ribosome

In 2005 three sets of ribosomes were published, which did not recognize natural mRNA, but instead translated a separate pool of orthogonal mRNA (o-mRNA). This was achieved by changing the recognition sequence of the mRNA, the Shine-Dalgarno sequence, and the corresponding recognition sequence in the 16S rRNA of

ribosomes, the so-called Anti-Shine-Dalgarno-Sequence. This way the base pairing, which is usually lost if either sequence is mutated, stays available. However the mutations in the 16S rRNA were not limited to the obviously base-pairing nucleotides of the classical Anti-Shine-Dalgarno sequence.

Ribo-X

In 2007 the group of Jason W. Chin presented an orthogonal ribosome, which was optimized for Amber codon suppression. The 16S rRNA was mutated in such a way that it bound the release factor RF1 less strongly than the natural ribosome does. This ribosome did not eliminate the problem of lowered cell fitness caused by suppressed stop codons in natural proteins. However through the improved specificity it raised the yields of correctly synthesized target protein significantly (from ~20% to >60% percent for one amber codon to be suppressed and from <1% to >20% for two amber codons).

Ribo-Q

In 2010 the group of Jason W. Chin presented a further optimized version of the orthogonal ribosome. The Ribo-Q is a 16S rRNA optimized to recognize tRNAs, which have quadruplet anti-codons to recognize quadruplet codons, instead of the natural triplet codons. With this approach the number of possible codons rises from 64 to 256. Even accounting for a variety of stop codons, more than 200 different amino acids could potentially be encoded this way.

Ribosome Stapppling

The orthogonal ribosomes described above all focus on optimizing the 16S rRNA. Thus far, this optimized 16S rRNA was combined with natural large-subunits to form orthogonal ribosomes. If the 23S rRNA, the main RNA-component of the large ribosomal subunit, is to be optimized as well, it had to be assured, that there was no crosstalk in the assembly of orthogonal and natural ribosomes. To ensure that optimized 23S rRNA would only form into ribosomes with the optimized 16S rRNA, the two rRNAs were combined into one transcript. By inserting the sequence for the 23S rRNA into a loop-region of the 16S rRNA sequence, both subunits still adopt functioning folds. Since, the two rRNAs are linked and thus in constant proximity, they preferably bind each other, not other free floating ribosomal subunits.

Engineered Peptidyl Transferase Center

In 2014 it was shown that by altering the peptidyl transferase center of the 23S rRNA, ribosomes could be created which draw on orthogonal pools of tRNA. The 3' end of tRNAs is universally conserved to be CCA. The two cytidines base pair with two guanines the 23S rRNA to bind the tRNA to the ribosome. This interaction is required for translational fidelity. However, by co-mutating the binding nucleotides in such a way, that they can still base pair, the translational fidelity can be conserved. The 3'-end of the tRNA is mutated from CCA to CGA, while two cytidine nucleotides in the ribosomes A- and P-sites are mutated to guanine. This leads to ribosomes which do not accept naturally occurring tRNAs as substrates and to tRNAs, which cannot be used as substrate by natural ribosomes. To use such tRNAs effectively, they would have to be aminoacylated by specific, orthogonal aaRSs. Most naturally occurring aaRSs recognize the 3'-end of their corresponding tRNA. aaRSs for these 3'-mutated tRNAs are not available yet. Thus far, this system has only been shown to work in an in-vitro translation setting where the aminoacylation of the orthogonal tRNA was achieved using so called "flexizymes". Flexizymes are ribozymes with tRNA-amino-acylation activity.

Applications

With an expanded genetic code, the unnatural amino acid can be genetically directed to any chosen site in the protein of interest. The high efficiency and fidelity of this process allows a better control of the placement of the

modification compared to modifying the protein post-translationally, which, in general, will target all amino acids of the same type, such as the thiol group of cysteine and the amino group of lysine. Also, an expanded genetic code allows modifications to be carried out *in vivo*.

The ability to site-specifically direct lab-synthesized chemical moieties into proteins allows many types of studies that would otherwise be extremely difficult, such as:

- Probing protein structure and function: By using amino acids with slightly different size such as *O*-methyltyrosine or dansylalanine instead of tyrosine, and by inserting genetically coded reporter moieties (color-changing and/or spin-active) into selected protein sites, chemical information about the protein's structure and function can be measured.
- Probing the role of post-translational modifications in protein structure and function: By using amino acids that mimic post-translational modifications such as phosphoserine, biologically active protein can be obtained, and the site-specific nature of the amino acid incorporation can lead to information on how the position, density, and distribution of protein phosphorylation effect protein function.
- Identifying and regulating protein activity: By using photocaged amino acids, protein function can be "switched" on or off by illuminating the organism.
- Changing the mode of action of a protein: One can start with the gene for a protein that binds a certain sequence of DNA and, by inserting a chemically active amino acid into the binding site, convert it to a protein that cuts the DNA rather than binding it.
- Improving immunogenicity and overcoming self-tolerance: By replacing strategically chosen tyrosines with *p*-nitro phenylalanine, a tolerated self-protein can be made immunogenic.
- Selective destruction of selected cellular components: using an expanded genetic code, unnatural, destructive chemical moieties (sometimes called "chemical warheads") can be incorporated into proteins that target specific cellular components.
- Producing better protein: the evolution of T7 bacteriophages on a non-evolving *E. coli* strain that encoded 3-iodotyrosine on the amber codon, resulted in a population fitter than wild-type thanks to the presence of iodotyrosine in its proteome

Future

The expansion of the genetic code is still in its infancy. Current methodology uses only one non-standard amino acid at the time, whereas ideally multiple could be used.

Recoded Synthetic Genome

One way to achieve the encoding of multiple unnatural amino acids is by rewriting genome synthetically. In 2010, at the cost of \$40 million an organism, *Mycoplasma laboratorium*, was constructed that was controlled by a synthetic genome. Due to the larger genome size this is not possible with *E. coli*, however several methods are being developed to overcome this, such as the fragmentation of the genome into separate linear chromosomes. In addition to the elimination of the usage of rare codons, the specificity of the system needs to be increased as many tRNA recognise several codons

Expanded Genetic Alphabet

Another approach is to expand the number of nucleobases to increase the coding capacity. An unnatural base pair (UBP) is a designed subunit (or nucleobase) of DNA which is created in a laboratory and does not occur in nature. A demonstration of UBPs were achieved *in vitro* by Ichiro Hirao's group at RIKEN institute in Japan. In 2002, they developed an unnatural base pair between 2-amino-8-(2-thienyl)purine (s) and pyridine-2-one (y) that functions *in vitro* in transcription and translation for the site-specific incorporation of non-standard amino

acids into proteins. In 2006, they created 7-(2-thienyl)imidazo[4,5-b]pyridine (Ds) and pyrrole-2-carbaldehyde (Pa) as a third base pair for replication and transcription. Afterwards, Ds and 4-[3-(6-aminohexanamido)-1-propynyl]-2-nitropyrrole (Px) was discovered as a high fidelity pair in PCR amplification. In 2013, they applied the Ds-Px pair to DNA aptamer generation by *in vitro* selection (SELEX) and demonstrated the genetic alphabet expansion significantly augment DNA aptamer affinities to target proteins.

In 2012, a group of American scientists led by Floyd Romesberg, a chemical biologist at the Scripps Research Institute in San Diego, California, published that his team designed an unnatural base pair (UBP). The two new artificial nucleotides or *Unnatural Base Pair* (UBP) were named d5SICS and dNaM. More technically, these artificial nucleotides bearing hydrophobic nucleobases, feature two fused aromatic rings that form a (d5SICS–dNaM) complex or base pair in DNA. In 2014 the same team from the Scripps Research Institute reported that they synthesized a stretch of circular DNA known as a plasmid containing natural T-A and C-G base pairs along with the best-performing UBPs Romesberg's laboratory had designed, and inserted it into cells of the common bacterium *E. coli* that successfully replicated the unnatural base pairs through multiple generations. This is the first known example of a living organism passing along an expanded genetic code to subsequent generations.

This was in part achieved by the addition of a supportive algal gene that expresses a nucleotide triphosphate transporter which efficiently imports the triphosphates of both d5SICSTP and dNaMTP into *E. coli* bacteria. Then, the natural bacterial replication pathways use them to accurately replicate the plasmid containing d5SICS–dNaM. The successful incorporation of a third base pair into a living micro-organism is a significant breakthrough towards the goal of greatly expanding the number of amino acids which can be encoded by DNA, thereby expanding the potential for living organisms to produce novel proteins. The artificial strings of DNA do not encode for anything yet, but scientists speculate they could be designed to manufacture new proteins which could have industrial or pharmaceutical uses. In May 2014, researchers announced that they had successfully introduced two new artificial nucleotides into bacterial DNA, and by including individual artificial nucleotides in the culture media, were able to passage the bacteria 24 times; they did not create mRNA or proteins able to use the artificial nucleotides.

Related Methods

Selective Pressure Incorporation (SPI) Method for Production of Alloproteins

There have been many studies that have produced protein with non-standard amino acids, but they do not alter the genetic code. These protein, called alloprotein, are made by incubating cells with an unnatural amino acid in the absence of a similar coded amino acid in order for the former to be incorporated into protein in place of the latter, for example L-2-aminohexanoic acid (Ahx) for methionine (Met).

These studies rely on the natural promiscuous activity of the aminoacyl tRNA synthetase to add to its target tRNA an unnatural amino acid (*i.e.*, analog) similar to the natural substrate, for example methionyl-tRNA synthetase's mistaking isoleucine for methionine. In protein crystallography, for example, the addition of selenomethionine to the media of a culture of a methionine-auxotrophic strain results in proteins containing selenomethionine as opposed to methionine (*viz.*, Multi-wavelength anomalous dispersion for reason).

Another example is that photoleucine and photomethionine are added instead of leucine and methionine to cross-label protein. Similarly, some tellurium-tolerant fungi can incorporate tellurocysteine and telluromethionine into their protein instead of cysteine and methionine. The objective of expanding the genetic code is more radical as it does not replace an amino acid, but it adds one or more to the code. On the other hand, proteome-wide replacements are most efficiently performed by global amino acid substitutions. For example, global proteome-wide substitutions of natural amino acids with fluorinated analogs have been attempted in *E. coli* and *B. subtilis*. A complete tryptophan substitution with thienopyrrole-alanine in response

to 20899 UGG codons in *E. coli* was reported in 2015 by Budisa and Söll. Moreover, many biological phenomena, such as protein folding and stability, are based on synergistic effects at many positions in the protein sequence.

In this context, the SPI method generates recombinant protein variants or alloproteins directly by substitution of natural amino acids with unnatural counterparts. An amino acid auxotrophic expression host is supplemented with an amino acid analog during target protein expression. This approach avoids the pitfalls of suppression-based methods and it is superior to it in terms of efficiency, reproducibility and an extremely simple experimental setup. Numerous studies demonstrated how global substitution of canonical amino acids with various isosteric analogs caused minimal structural perturbations but dramatic changes in thermodynamic, folding, aggregation spectral properties and enzymatic activity.

In vitro synthesis:

- The genetic code expansion described above is *in vivo*. An alternative is the change of coding *in vitro* translation experiments. This requires the depletion of all tRNAs and the selective reintroduction of certain aminoacylated-tRNAs, some chemically aminoacylated.

Chemical Synthesis

There are several techniques to produce peptides chemically, generally it is by solid-phase protection chemistry. This means that any (protected) amino acid can be added into the nascent sequence. In November 2017, a team from the Scripps Research Institute reported having constructed a semi-synthetic *E. coli* bacteria genome using six different nucleic acids (versus four found in nature).

The two extra ‘letters’ form a third, unnatural base pair. The resulting organisms were able to thrive and synthesize proteins using “unnatural amino acids”. The unnatural base pair used is dNaM–dTPT3. This unnatural base pair has been demonstrated previously, but this is the first report of transcription and translation of proteins using an unnatural base pair.

Roles

In cells, especially autotrophs, several non-proteinogenic amino acids are found as metabolic intermediates. However, despite the catalytic flexibility of PLP-binding enzymes, many amino acids are synthesised as keto-acids (*e.g.*, 4-methyl-2-oxopentanoate to leucine) and aminated in the last step, thus keeping the number of non-proteinogenic amino acid intermediates fairly low. Ornithine and citrulline occur in the urea cycle, part of amino acid catabolism. In addition to primary metabolism, several non-proteinogenic amino acids are precursors or the final production in secondary metabolism to make small compounds or non-ribosomal peptides (such as some toxins).

Post-translationally Incorporated into Protein

Despite not being encoded by the genetic code as proteinogenic amino acids, some non-standard amino acids are nevertheless found in proteins. These are formed by post-translational modification of the side chains of standard amino acids present in the target protein.

These modifications are often essential for the function or regulation of a protein; for example, in Gamma-carboxyglutamate the carboxylation of glutamate allows for better binding of calcium cations, and in hydroxyproline the hydroxylation of proline is critical for maintaining connective tissues. Another example is the formation of hypusine in the translation initiation factor EIF5A, through modification of a lysine residue. Such modifications can also determine the localization of the protein, *e.g.*, the addition of long hydrophobic groups can cause a protein to bind to a phospholipid membrane.

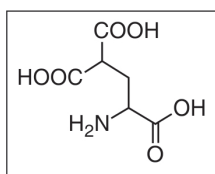


Fig. Carboxyglutamic acid. Whereas glutamic acid possess one γ -carboxyl group, Carboxyglutamic acid possess two.

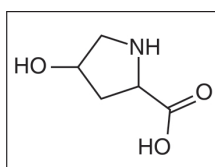


Fig. Hydroxyproline. This imino acid differs from proline due to a hydroxyl group on carbon 4.

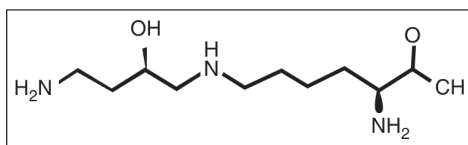


Fig. Hypusine. This amino acid is obtained by adding to the ϵ -amino group of a lysine a 4-aminobutyl moiety (obtained from spermidine).

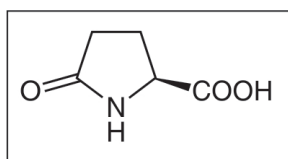


Fig. Pyroglutamic acid.

There is some preliminary evidence that aminomalonic acid may be present, possibly by misincorporation, in protein.

Toxic Analogues

Several non-proteinogenic amino acids are toxic due to their ability to mimic certain properties of proteinogenic amino acids, such as thialysine. Some non-proteinogenic amino acids are neurotoxic by mimicking amino acids used as neurotransmitters (*i.e.*, not for protein biosynthesis), *e.g.*, Quisqualic acid, canavanine or azetidine-2-carboxylic acid. Cephalosporin C has an α -amino adipic acid (homoglutamate) backbone that is amidated with a cephalosporin moiety. Penicillamine is a therapeutic amino acid, whose mode of action is unknown.

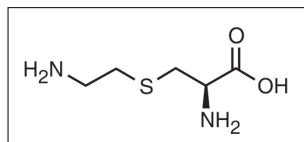


Fig. Thialysine.

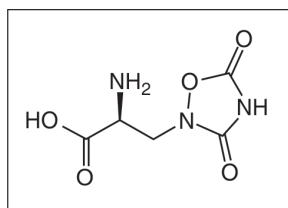


Fig. Quisqualic acid.

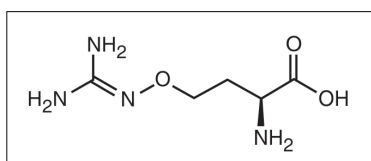


Fig. Canavanine.

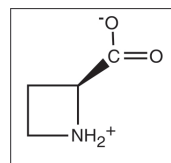


Fig. azetidine-2-carboxylic acid.

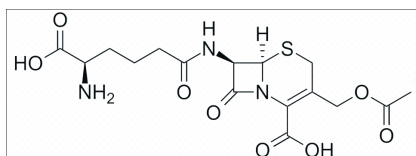


Fig. Cephalosporin C.

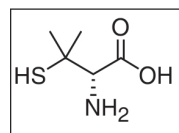


Fig. Penicillamine.

Naturally-occurring cyanotoxins can also include non-proteinogenic amino acids. Microcystin and nodularin, for example, are both derived from ADDA, a β -amino acid.

Not Amino Acids

Taurine is an amino sulfonic acid and not an amino acid, however it is occasionally considered as such as the amounts required to suppress the auxotroph in certain organisms (*e.g.*, cats) are closer to those of “essential amino acids” (amino acid auxotrophy) than of vitamins (cofactor auxotrophy). The osmolytes, sarcosine and glycine betaine are derived from amino acids, but have a secondary and quaternary amine respectively.

D-AMINO ACID NATURAL ABUNDANCE

D-isomers are uncommon in live organisms. For instance, gramicidin is a polypeptide made up from mixture of D- and L-amino acids. Other compounds containing D-amino acid are tyrocidine and valinomycin. These compounds disrupt bacterial cell walls, particularly in Gram-positive bacteria. Only 837 D-amino acids were found in Swiss-Prot database (187 million amino acids analysed).

NON-STANDARD AMINO ACIDS

The 20 amino acids that are encoded directly by the codons of the universal genetic code are called *standard* or *canonical* amino acids. A modified form of methionine (*N*-formylmethionine) is often incorporated in place of methionine as the initial amino acid of proteins in bacteria, mitochondria and chloroplasts. Other amino acids are called *non-standard* or *non-canonical*. Most of the non-standard amino acids are also non-proteinogenic (*i.e.*, they cannot be incorporated into proteins during translation), but two of them are proteinogenic, as they can be incorporated translationally into proteins by exploiting information not encoded in the universal genetic code. The two non-standard proteinogenic amino acids are selenocysteine (present in many non-eukaryotes as well as most eukaryotes, but not coded directly by DNA) and pyrrolysine (found only in some archaea and one bacterium). The incorporation of these non-standard amino acids is rare. For example, 25 human proteins include selenocysteine (Sec) in their primary structure, and the structurally characterized enzymes (selenoenzymes) employ Sec as the catalytic moiety in their active sites. Pyrrolysine and selenocysteine are encoded via variant codons. For example, selenocysteine is encoded by stop codon and SECIS element.

IN HUMAN NUTRITION

When taken up into the human body from the diet, the 20 standard amino acids either are used to synthesize proteins and other biomolecules or are oxidized to urea and carbon dioxide as a source of energy. The oxidation

pathway starts with the removal of the amino group by a transaminase; the amino group is then fed into the urea cycle. The other product of transamination is a keto acid that enters the citric acid cycle. Glucogenic amino acids can also be converted into glucose, through gluconeogenesis.

Of the 20 standard amino acids, nine (His, Ile, Leu, Lys, Met, Phe, Thr, Trp and Val) are called essential amino acids because the human body cannot synthesize them from other compounds at the level needed for normal growth, so they must be obtained from food. In addition, cysteine, taurine, tyrosine, and arginine are considered semiessential amino-acids in children (though taurine is not technically an amino acid), because the metabolic pathways that synthesize these amino acids are not fully developed. The amounts required also depend on the age and health of the individual, so it is hard to make general statements about the dietary requirement for some amino acids. Dietary exposure to the non-standard amino acid BMAA has been linked to human neurodegenerative diseases, including ALS.

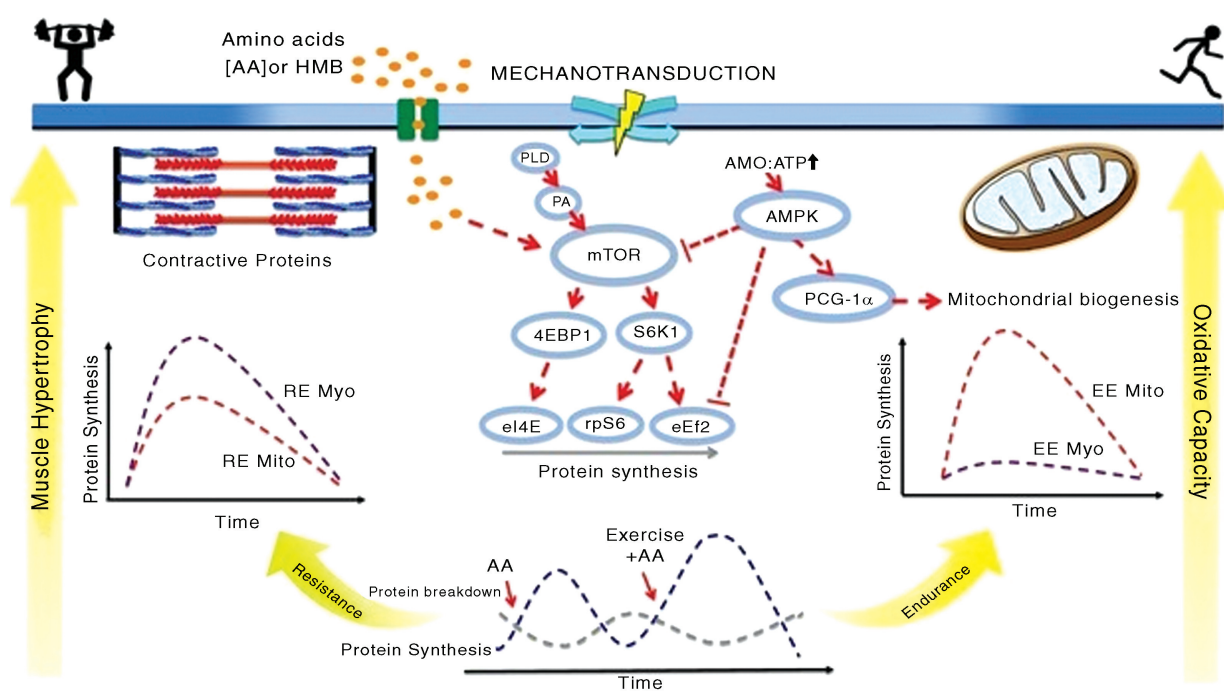


Fig. Diagram of the molecular signaling cascades that are involved in myofibrillar muscle protein synthesis and mitochondrial biogenesis in response to physical exercise and specific amino acids or their derivatives (primarily L-leucine and HMB). Many amino acids derived from food protein promote the activation of mTORC1 and increase protein synthesis by signaling through Rag GTPases.

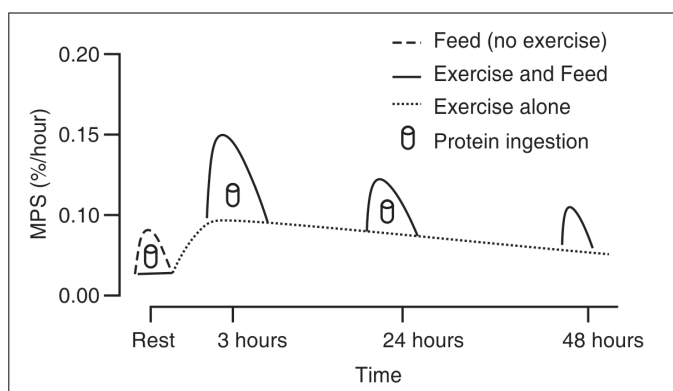


Fig. Resistance training stimulates muscle protein synthesis (MPS) for a period of up to 48 hours following exercise (shown by lighter dotted line). Ingestion of a protein-rich meal at any point during this period will augment the exercise-induced increase in muscle protein synthesis (shown by solid lines).

USES IN INDUSTRY

Amino acids are used for a variety of applications in industry, but their main use is as additives to animal feed. This is necessary, since, many of the bulk components of these feeds, such as soybeans, either have low levels or lack some of the essential amino acids: Lysine, methionine, threonine, and tryptophan are most important in the production of these feeds. In this industry, amino acids are also used to chelate metal cations in order to improve the absorption of minerals from supplements, which may be required to improve the health or production of these animals. The food industry is also a major consumer of amino acids, in particular, glutamic acid, which is used as a flavour enhancer, and aspartame (aspartyl-phenylalanine-1-methyl ester) as a low-calorie artificial sweetener. Similar technology to that used for animal nutrition is employed in the human nutrition industry to alleviate symptoms of mineral deficiencies, such as anemia, by improving mineral absorption and reducing negative side effects from inorganic mineral supplementation.

The chelating ability of amino acids has been used in fertilizers for agriculture to facilitate the delivery of minerals to plants in order to correct mineral deficiencies, such as iron chlorosis. These fertilizers are also used to prevent deficiencies from occurring and improving the overall health of the plants. The remaining production of amino acids is used in the synthesis of drugs and cosmetics. Similarly, some amino acids derivatives are used in pharmaceutical industry. They include 5-HTP (5-hydroxytryptophan) used for experimental treatment of depression, L-DOPA (L-dihydroxyphenylalanine) for Parkinson's treatment, and eflornithine drug that inhibits ornithine decarboxylase and used in the treatment of sleeping sickness.

EXPANDED GENETIC CODE

Since, 2001, 40 non-natural amino acids have been added into protein by creating a unique codon (recoding) and a corresponding transfer-RNA:aminoacyl-tRNA-synthetase pair to encode it with diverse physicochemical and biological properties in order to be used as a tool to exploring protein structure and function or to create novel or enhanced proteins.

Nullomers

Nullomers are codons that in theory code for an amino acid, however in nature there is a selective bias against using this codon in favor of another, for example bacteria prefer to use CGA instead of AGA to code for arginine. This creates some sequences that do not appear in the genome. This characteristic can be taken advantage of and used to create new selective cancer-fighting drugs and to prevent cross-contamination of DNA samples from crime-scene investigations.

CHEMICAL BUILDING BLOCKS

Amino acids are important as low-cost feedstocks. These compounds are used in chiral pool synthesis as enantiomerically pure building-blocks. Amino acids have been investigated as precursors chiral catalysts, *e.g.*, for asymmetric hydrogenation reactions, although no commercial applications exist.

BIODEGRADABLE PLASTICS

Amino acids are under development as components of a range of biodegradable polymers. These materials have applications as environmentally friendly packaging and in medicine in drug delivery and the construction of prosthetic implants. These polymers include polypeptides, polyamides, polyesters, polysulfides, and polyurethanes with amino acids either forming part of their main chains or bonded as side chains. These modifications alter the physical properties and reactivities of the polymers. An interesting example of such materials is polyaspartate, a

water-soluble biodegradable polymer that may have applications in disposable diapers and agriculture. Due to its solubility and ability to chelate metal ions, polyaspartate is also being used as a biodegradable anti-scaling agent and a corrosion inhibitor. In addition, the aromatic amino acid tyrosine is being developed as a possible replacement for toxic phenols such as bisphenol A in the manufacture of polycarbonates.

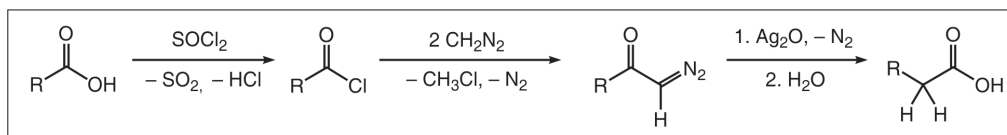
REACTIONS

As amino acids have both a primary amine group and a primary carboxyl group, these chemicals can undergo most of the reactions associated with these functional groups. These include nucleophilic addition, amide bond formation, and imine formation for the amine group, and esterification, amide bond formation, and decarboxylation for the carboxylic acid group. The combination of these functional groups allow amino acids to be effective polydentate ligands for metal-amino acid chelates. The multiple side chains of amino acids can also undergo chemical reactions. The types of these reactions are determined by the groups on these side chains and are, therefore, different between the various types of amino acid.

CHEMICAL SYNTHESIS

Arndt–Eistert Reaction

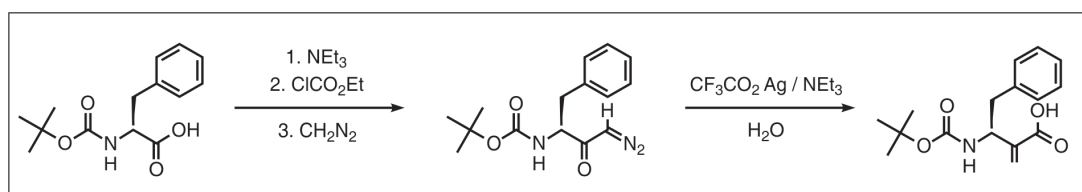
The Arndt–Eistert synthesis is a series of chemical reactions designed to convert a carboxylic acid to a higher carboxylic acid homologue (*i.e.*, contains one additional carbon atom) and is considered a homologation process. Named for the German chemists Fritz Arndt (1885–1969) and Bernd Eistert (1902–1978), Arndt–Eistert synthesis is a popular method of producing β -amino acids from α -amino acids. Acid chlorides react with diazomethane to give diazoketones. In the presence of a nucleophile (water) and a metal catalyst (Ag_2O), diazoketones will form the desired acid homologue.



While the classic Arndt–Eistert synthesis uses thionyl chloride to convert the starting acid to an acid chloride, any procedure can be used that will generate an acid chloride. Diazoketones are typically generated as described here, but other methods such as diazo-group transfer can also apply. Since, diazomethane is toxic and violently explosive, many safer alternatives have been developed, such as the usage of ynolates (Kowalski ester homologation) or diazo(trimethylsilyl)methane.

Reaction Mechanism

The key step in the Arndt–Eistert synthesis is the metal-catalyzed Wolff rearrangement of the diazoketone to form a ketene. In the insertion homologation of *t*-BOC protected (*S*)-phenylalanine (2-amino-3-phenylpropanoic acid), *t*-BOC protected (*S*)-3-amino-4-phenylbutanoic acid is formed.



Wolff rearrangement of the α -diazoketone intermediate forms a ketene via a 1,2-rearrangement, which is subsequently hydrolysed to form the carboxylic acid. The consequence of the 1,2-rearrangement is that the

methylene group alpha to the carboxyl group in the product is the methylene group from the diazomethane reagent. It has been demonstrated that the rearrangement preserves the stereochemistry of the chiral centre as the product formed from *t*-BOC protected (*S*)-phenylalanine retains the (*S*) stereochemistry with a reported enantiomeric excess of at least 99%. Heat, light, platinum, silver, and copper salts will also catalyze the Wolff rearrangement to produce the desired acid homologue.

Variations

In the Newman–Beal modification, addition of triethylamine to the diazomethane solution will avoid the formation of α -chloromethylketone side-products.

Corey–Link Reaction

In organic chemistry, the Corey–Link reaction is a name reaction that converts a 1,1,1-trichloro-2-keto structure into a 2-aminocarboxylic acid (an alpha amino acid) or other acyl functional group with control of the chirality at the alpha position. The reaction is named for E.J. Corey and John Link, who first reported the reaction sequence.

Process

The first stage of the process is the reduction of the carbonyl to give a 1,1,1-trichloro-2-hydroxy structure. The original protocol used catecholborane with a small amount of one enantiomer of CBS catalyst (a Corey–Itsuno reduction). The choice of chirality of the catalyst thus gives selectivity for one or the other stereochemistry of the alcohol in the product. This 2-hydroxy structure is then reacted with azide and a nucleophilic base. The multistep reaction mechanism begins with deprotonation of the alcohol, followed by the oxygen-anion attacking the adjacent trichloromethyl position to form an epoxide. The azide then opens this ring by an S_N2 reaction to give a 2-azido structure whose stereochemistry is inverted compared to the original 2-hydroxy.

The oxygen, having become attached to the first carbon of the chain during the epoxide formation, simultaneously displaces a second chlorine atom there to form an acyl chloride. An additional nucleophilic reactant, such as hydroxide or an alkoxide, then triggers an acyl substitution there to produce a carboxylic acid or ester. Various other nucleophiles can be used to generate other acyl functional groups. This sequence of steps gives a 2-azido compound, which is then reduced to the 2-amino compound in a separate reaction, typically a Staudinger reaction.

Bargellini Reaction

The Bargellini reaction involves the same type of dichloroepoxy intermediate, formed by a different method, that reacts with a single structure containing two nucleophilic groups. It thus gives products such as morpholinones or piperazinones, alpha-amino esters or amides in which the amine is tethered to the acyl substituent group.

Miller–Urey Experiment

The Miller–Urey experiment (or Miller experiment) was a chemical experiment that simulated the conditions thought at the time to be present on the early Earth, and tested the chemical origin of life under those conditions. The experiment supported Alexander Oparin's and J. B. S. Haldane's hypothesis that putative conditions on the primitive Earth favoured chemical reactions that synthesized more complex organic compounds from simpler inorganic precursors. Considered to be the classic experiment investigating abiogenesis, it was conducted in 1952 by Stanley Miller, with assistance from Harold Urey, at the University of Chicago and later the University

of California, San Diego and published the following year. After Miller's death in 2007, scientists examining sealed vials preserved from the original experiments were able to show that there were actually well over 20 different amino acids produced in Miller's original experiments. That is considerably more than what Miller originally reported, and more than the 20 that naturally occur in life. More recent evidence suggests that Earth's original atmosphere might have had a composition different from the gas used in the Miller experiment. But prebiotic experiments continue to produce racemic mixtures of simple to complex compounds under varying conditions.

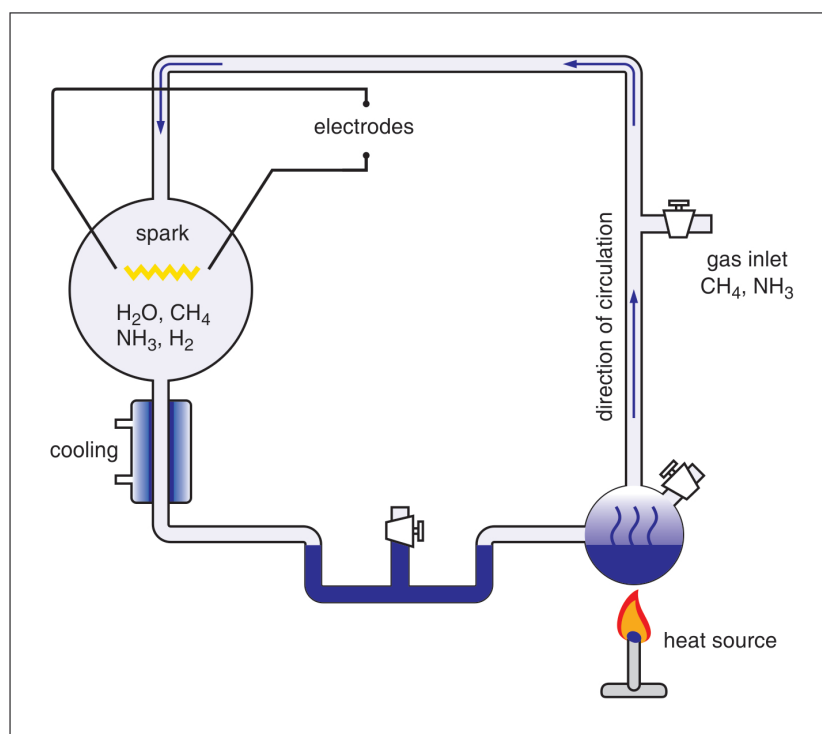


Fig. The experiment.

Experiment

The experiment used water (H_2O), methane (CH_4), ammonia (NH_3), and hydrogen (H_2). The chemicals were all sealed inside a sterile 5-liter glass flask connected to a 500 ml flask half-full of liquid water. The liquid water in the smaller flask was heated to induce evaporation, and the water vapour was allowed to enter the larger flask. Continuous electrical sparks were fired between the electrodes to simulate lightning in the water vapour and gaseous mixture, and then the simulated atmosphere was cooled again so that the water condensed and trickled into a U-shaped trap at the bottom of the apparatus.

After a day, the solution collected at the trap had turned pink in colour. At the end of one week of continuous operation, the boiling flask was removed, and mercuric chloride was added to prevent microbial contamination. The reaction was stopped by adding barium hydroxide and sulfuric acid, and evaporated to remove impurities. Using paper chromatography, Miller identified five amino acids present in the solution: glycine, α -alanine and β -alanine were positively identified, while aspartic acid and α -aminobutyric acid (AABA) were less certain, due to the spots being faint. In a 1996 interview, Stanley Miller recollected his lifelong experiments following his original work and stated: "Just turning on the spark in a basic pre-biotic experiment will yield 11 out of 20 amino acids." As observed in all subsequent experiments, both left-handed (L) and right-handed (D) optical isomers were created in a racemic mixture. In biological systems, almost all of the compounds are non-racemic, or homochiral. The original experiment remains today under the care of Miller and Urey's former

student Jeffrey Bada, a professor at the UCSD, Scripps Institution of Oceanography. The apparatus used to conduct the experiment is on display at the Denver Museum of Nature and Science.

Chemistry of Experiment

One-step reactions among the mixture components can produce hydrogen cyanide (HCN), formaldehyde (CH_2O), and other active intermediate compounds (acetylene, cyanoacetylene, *etc.*):

- $\text{CO}_2 \rightarrow \text{CO} + [\text{O}]$ (atomic oxygen)
- $\text{CH}_4 + 2[\text{O}] \rightarrow \text{CH}_2\text{O} + \text{H}_2\text{O}$
- $\text{CO} + \text{NH}_3 \rightarrow \text{HCN} + \text{H}_2\text{O}$
- $\text{CH}_4 + \text{NH}_3 \rightarrow \text{HCN} + 3\text{H}_2$ (BMA process)

The formaldehyde, ammonia, and HCN then react by Strecker synthesis to form amino acids and other biomolecules:

- $\text{CH}_2\text{O} + \text{HCN} + \text{NH}_3 \rightarrow \text{NH}_2\text{-CH}_2\text{-CN} + \text{H}_2\text{O}$
- $\text{NH}_2\text{-CH}_2\text{-CN} + 2\text{H}_2\text{O} \rightarrow \text{NH}_3 + \text{NH}_2\text{-CH}_2\text{-COOH}$ (glycine)

Furthermore, water and formaldehyde can react, via Butlerov's reaction to produce various sugars like ribose.

The experiments showed that simple organic compounds of building blocks of proteins and other macromolecules can be formed from gases with the addition of energy.

Other Experiments

This experiment inspired many others. In 1961, Joan Oró found that the nucleotide base adenine could be made from hydrogen cyanide (HCN) and ammonia in a water solution. His experiment produced a large amount of adenine, the molecules of which were formed from 5 molecules of HCN. Also, many amino acids are formed from HCN and ammonia under these conditions. Experiments conducted later showed that the other RNA and DNA nucleobases could be obtained through simulated prebiotic chemistry with a reducing atmosphere. There also had been similar electric discharge experiments related to the origin of life contemporaneous with Miller–Urey. An article in *The New York Times* (March 8, 1953:E9), titled “Looking Back Two Billion Years” describes the work of Wollman (William) M. MacNevin at The Ohio State University, before the Miller *Science* paper was published in May 1953. MacNevin was passing 100,000 volt sparks through methane and water vapor and produced “resinous solids” that were “too complex for analysis.” The article describes other early earth experiments being done by MacNevin. It is not clear if he ever published any of these results in the primary scientific literature.

K. A. Wilde submitted a paper to *Science* on December 15, 1952, before Miller submitted his paper to the same journal on February 10, 1953. Wilde's paper was published on July 10, 1953. Wilde used voltages up to only 600 V on a binary mixture of carbon dioxide (CO_2) and water in a flow system. He observed only small amounts of carbon dioxide reduction to carbon monoxide, and no other significant reduction products or newly formed carbon compounds. Other researchers were studying UV-photolysis of water vapor with carbon monoxide. They have found that various alcohols, aldehydes and organic acids were synthesized in reaction mixture. More recent experiments by chemists Jeffrey Bada, one of Miller's graduate students, and Jim Cleaves at Scripps Institution of Oceanography of the University of California, San Diego were similar to those performed by Miller. However, Bada noted that in current models of early Earth conditions, carbon dioxide and nitrogen (N_2) create nitrites, which destroy amino acids as fast as they form. When Bada performed the Miller-type experiment with the addition of iron and carbonate minerals, the products were rich in amino acids. This suggests the origin of significant amounts of amino acids may have occurred on Earth even with an atmosphere containing carbon dioxide and nitrogen.

Earth's Early Atmosphere

Some evidence suggests that Earth's original atmosphere might have contained fewer of the reducing molecules than was thought at the time of the Miller–Urey experiment. There is abundant evidence of major volcanic eruptions 4 billion years ago, which would have released carbon dioxide, nitrogen, hydrogen sulfide (H_2S), and sulfur dioxide (SO_2) into the atmosphere.

Experiments using these gases in addition to the ones in the original Miller–Urey experiment have produced more diverse molecules. The experiment created a mixture that was racemic (containing both L and D enantiomers) and experiments since, have shown that “in the lab the two versions are equally likely to appear”; however, in nature, L amino acids dominate. Later experiments have confirmed disproportionate amounts of L or D oriented enantiomers are possible. Originally it was thought that the primitive secondary atmosphere contained mostly ammonia and methane. However, it is likely that most of the atmospheric carbon was CO_2 with perhaps some CO and the nitrogen mostly N_2 . In practice gas mixtures containing CO, CO_2 , N_2 , etc., give much the same products as those containing CH_4 and NH_3 so long as there is no O_2 . The hydrogen atoms come mostly from water vapor. In fact, in order to generate aromatic amino acids under primitive earth conditions it is necessary to use less hydrogen-rich gaseous mixtures. Most of the natural amino acids, hydroxyacids, purines, pyrimidines, and sugars have been made in variants of the Miller experiment. More recent results may question these conclusions.

The University of Waterloo and University of Colorado conducted simulations in 2005 that indicated that the early atmosphere of Earth could have contained up to 40 percent hydrogen—implying a much more hospitable environment for the formation of prebiotic organic molecules. The escape of hydrogen from Earth's atmosphere into space may have occurred at only one percent of the rate previously believed based on revised estimates of the upper atmosphere's temperature. One of the authors, Owen Toon notes: “In this new scenario, organics can be produced efficiently in the early atmosphere, leading us back to the organic-rich soup-in-the-ocean concept. I think this study makes the experiments by Miller and others relevant again.” Outgassing calculations using a chondritic model for the early earth complement the Waterloo/Colorado results in re-establishing the importance of the Miller–Urey experiment. In contrast to the general notion of early earth's reducing atmosphere, researchers at the Rensselaer Polytechnic Institute in New York reported the possibility of oxygen available around 4.3 billion years ago. Their study reported in 2011 on the assessment of Hadean zircons from the earth's interior (magma) indicated the presence of oxygen traces similar to modern-day lavas. This study suggests that oxygen could have been released in the earth's atmosphere earlier than generally believed.

Extraterrestrial Sources

Conditions similar to those of the Miller–Urey experiments are present in other regions of the solar system, often substituting ultraviolet light for lightning as the energy source for chemical reactions. The Murchison meteorite that fell near Murchison, Victoria, Australia in 1969 was found to contain over 90 different amino acids, nineteen of which are found in Earth life. Comets and other icy outer-solar-system bodies are thought to contain large amounts of complex carbon compounds (such as tholins) formed by these processes, darkening surfaces of these bodies. The early Earth was bombarded heavily by comets, possibly providing a large supply of complex organic molecules along with the water and other volatiles they contributed. This has been used to infer an origin of life outside of Earth: the panspermia hypothesis.

Recent Related Studies

In recent years, studies have been made of the amino acid composition of the products of “old” areas in “old” genes, defined as those that are found to be common to organisms from several widely separated species,

assumed to share only the last universal ancestor (LUA) of all extant species. These studies found that the products of these areas are enriched in those amino acids that are also most readily produced in the Miller–Urey experiment. This suggests that the original genetic code was based on a smaller number of amino acids—only those available in prebiotic nature—than the current one.

Jeffrey Bada, himself Miller’s student, inherited the original equipment from the experiment when Miller died in 2007. Based on sealed vials from the original experiment, scientists have been able to show that although successful, Miller was never able to find out, with the equipment available to him, the full extent of the experiment’s success. Later researchers have been able to isolate even more different amino acids, 25 altogether. Bada has estimated that more accurate measurements could easily bring out 30 or 40 more amino acids in very low concentrations, but the researchers have since, discontinued the testing. Miller’s experiment was therefore a remarkable success at synthesizing complex organic molecules from simpler chemicals, considering that all life uses just 20 different amino acids.

In 2008, a group of scientists examined 11 vials left over from Miller’s experiments of the early 1950s. In addition to the classic experiment, reminiscent of Charles Darwin’s envisioned “warm little pond”, Miller had also performed more experiments, including one with conditions similar to those of volcanic eruptions. This experiment had a nozzle spraying a jet of steam at the spark discharge.

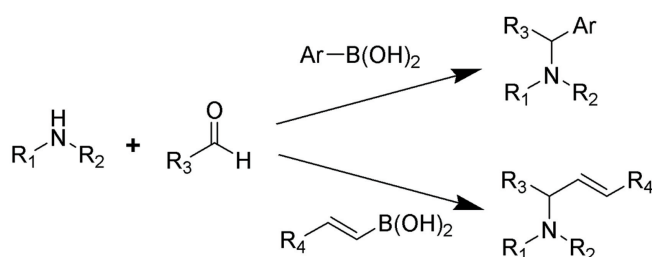
By using high-performance liquid chromatography and mass spectrometry, the group found more organic molecules than Miller had. Interestingly, they found that the volcano-like experiment had produced the most organic molecules, 22 amino acids, 5 amines and many hydroxylated molecules, which could have been formed by hydroxyl radicals produced by the electrified steam. The group suggested that volcanic island systems became rich in organic molecules in this way, and that the presence of carbonyl sulfide there could have helped these molecules form peptides.

Amino Acids Identified

Below is a table of amino acids produced and identified in the “classic” 1952 experiment, as published by Miller in 1953, the 2008 re-analysis of vials from the volcanic spark discharge experiment, and the 2010 re-analysis of vials from the H₂S-rich spark discharge experiment.

Petasis Reaction

The Petasis reaction (alternatively called the Petasis borono–Mannich (PBM) reaction) is the chemical reaction of an amine, aldehyde, and vinyl- or aryl-boronic acid to form substituted amines.



Reported in 1993 by Nicos Petasis as a practical method towards the synthesis of a geometrically pure antifungal agent, naftifine, the Petasis reaction can be described as a variation of the Mannich reaction. Rather than generating an enolate to form the substituted amine product, in the Petasis reaction, the vinyl group of the organoboronic acid serves as the nucleophile. In comparison to other methods of generating allyl amines, the Petasis reaction tolerates a multifunctional scaffold, with a variety of amines and organoboronic acids as potential starting materials.

Amino acid	Produced in experiment			Proteinogenic
	Miller–Urey (1952)	Volcanic spark discharge (2008)	H ₂ S-rich spark discharge (2010)	
Glycine	✓	✓	✓	Yes
α -Alanine	✓	✓	✓	Yes
β -Alanine	✓	✓	✓	No
Aspartic acid	✓	✓	✓	Yes
α -Aminobutyric acid	✓	✓	✓	No
Serine	✗	✓	✓	Yes
Isoserine	✗	✓	✓	No
α -Aminoisobutyric acid	✗	✓	✓	No
β -Aminoisobutyric acid	✗	✓	✓	No
β -Aminobutyric acid	✗	✓	✓	No
γ -Aminobutyric acid	✗	✓	✓	No
Valine	✗	✓	✓	Yes
Isovaline	✗	✓	✓	No
Glutamic acid	✗	✓	✓	Yes
Norvaline	✗	✓	✗	No
α -Aminoadipic acid	✗	✓	✗	No
Homoserine	✗	✓	✗	No
2-Methylserine	✗	✓	✗	No
β -Hydroxyaspartic acid	✗	✓	✗	No
Ornithine	✗	✓	✗	No
2-Methylglutamic acid	✗	✓	✗	No
Phenylalanine	✗	✓	✗	Yes
Homocysteic acid	✗	✗	✓	No
S-methylcysteine	✗	✗	✓	No
Methionine	✗	✗	✓	Yes
Methionine sulfoxide	✗	✗	✓	No
Methionine sulfone	✗	✗	✓	No
Isoleucine	✗	✗	✓	Yes
Leucine	✗	✗	✓	Yes
Ethionine	✗	✗	✓	No

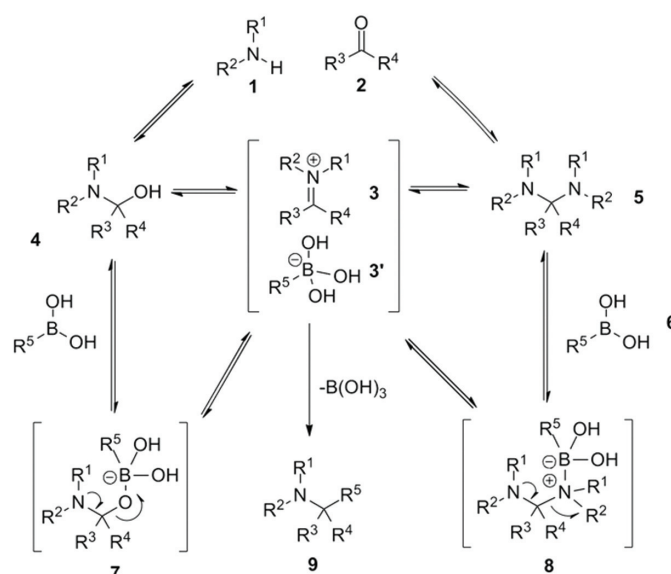
Additionally, the reaction does not require anhydrous or inert conditions. As a mild, selective synthesis, the Petasis reaction is useful in generating α -amino acids, and is utilized in combinatorial chemistry and drug discovery.

Reaction Mechanism

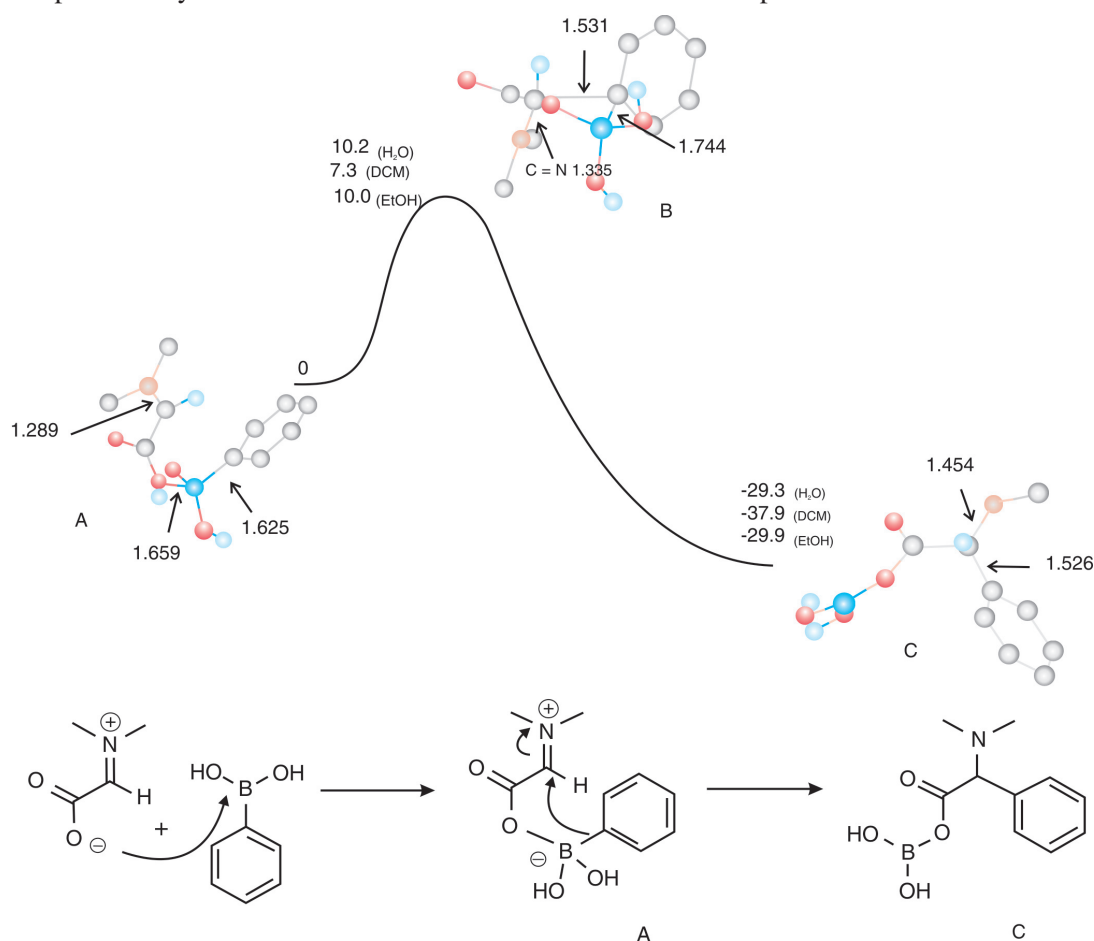
The mechanism of the Petasis reaction is not fully understood; however, it is similar to the Mannich reaction at its early stage. In the Mannich reaction, an imine or iminium salt serves as the electrophile to which the nucleophile adds; however, in the Petasis reaction it is not clear which intermediate serves as the electrophile. Petasis proposes that the reaction is characterized by a complex equilibrium among the three starting materials and various intermediates, and the final product is formed via a rate-determining and irreversible C-C bond formation step.

The condensation between amine 1 and carbonyl 2 forms hemiaminal 4, which is in a complex equilibrium with iminium ion 3 and aminal 5. Boronic acid 6 react with hemiaminal 4 and aminal 5 in a reversible fashion via intermediate 7 and 8 respectively, forming again electrophilic iminium ion 3, this time accompanied by nucleophilic boronate 3'.

Note that there are no evidences suggesting that boronic acid alone can directly react with iminium ions: In addition to needing acid for an appreciable amount of iminium salt to be generated, it has been shown that vinyl boronic acids do not react efficiently with preformed iminium salts (like Eschenmoser's salt). The irreversible C-C bond migration between 3 and 3' then follows, furnishing desired product 9 with loss of boric acid. All intermediates will ultimately lead to the final product, as the reaction between 3 and 3' is irreversible, pulling the equilibrium of the entire system towards the final product.

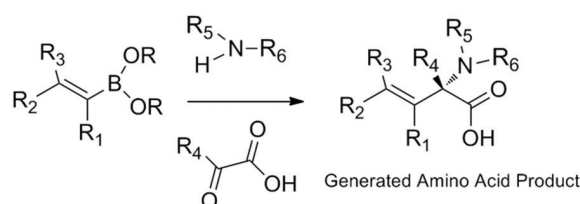


Density functional theory (DFT) studies have been performed to study the mechanism of Petasis reaction. Starting with the Petasis reaction between glyoxylic acid, dimethylamine and phenylboronic acid, Gois *et al.* reported that the migration of the boronic acid substituent (phenyl group) of the "ate complex" A incurs an energy barrier of 10kcal/mol and a five-membered transition state B. Formation of quaternary boron salts during the reaction has been experimentally confirmed by Hansen and coworkers. They reported that, in the absence of amine, an upfield ^{11}B shift is observed after the addition of glyoxylic acid to a solution of phenylboronic acid. This is presumably due to the formation of electron rich boronate species.

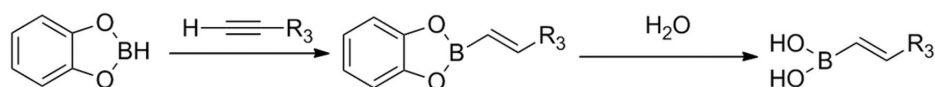


Preparation

The Petasis reaction proceeds under mild conditions, without the use of strong acids, bases, or metals. Unlike the Strecker and Ugi syntheses, the Petasis reaction avoids the use of cyanide and isocyanide reagents. The amine is mixed with the carbonyl substrate using either dioxane or toluene as a solvent at 90 °C for 10 minutes. Subsequently, the boronic acid is added to the mixture and product is generated, either after 30 minutes at 90 °C, or after several hours at 25 °C. In α -amino acid synthesis, α -keto acids, such as glyoxylic and pyruvic acid, are stirred in ethanol, toluene, or dichloromethane with amines and vinyl boronic acids at 25–50 °C for 12–48 h to give the corresponding β , γ -unsaturated compounds.



One of the most attractive features of the Petasis reaction is its use of boronic acids as a nucleophilic source. Unlike most vinyl substrates, vinyl boronic acids are stable to air and water and can be removed during workup with a simple extraction. Many boronic acid derivatives are easy to prepare and with the advent of the Suzuki coupling, a larger number of them are now commercially available. In the seminal report of the reaction, the organoboronic acids were prepared by hydroboration of terminal alkynes with catecholborane.



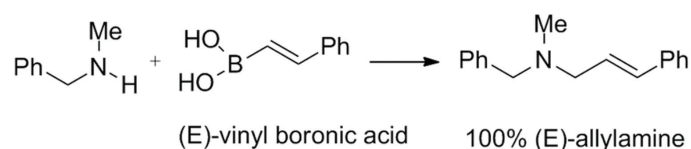
Other methods of generating boronic acids were also reported.

Reaction Scope and Synthetic Applications

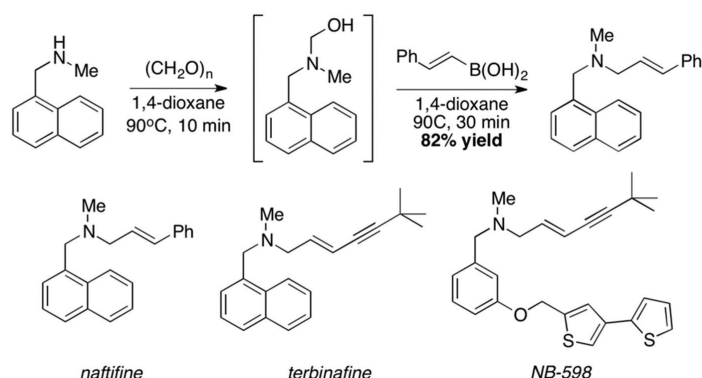
A wide variety of functional groups including alcohols, carboxylic acids, and amines are tolerated in the Petasis Reaction. Known substrates that are compatible with reaction conditions include vinylboronate esters, arylboronate esters, and potassium organotrifluoroborates. Additionally, a variety of substituted amines can be used other than secondary amines. Tertiary aromatic amines, hydrazines, hydroxylamines, sulfonamides, and indoles have all been reported.

Synthesis of Allyl Amines

Petasis and coworkers proposed, in their seminal study, that vinyl boronic acids can react with the adducts of secondary amines and paraformaldehyde to give tertiary allyl amines. The geometry of the double bond of the starting vinyl boronic acid is completely retained in the final product. Yield is typically in the good to excellent range. *The following reaction is particularly effective, hitting a yield of 96%:*

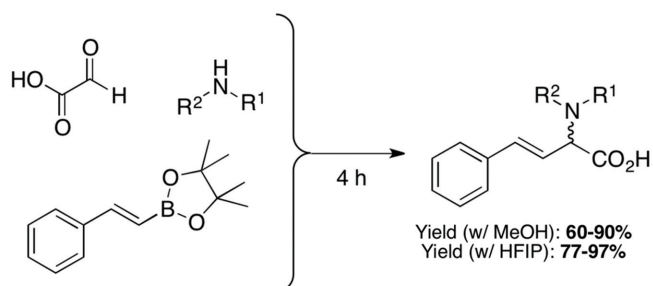


Petasis and coworkers used this reaction to synthesize naftifine, a very potent topical antimycotic, in one step in 82% yield. Other compounds with related structure include terbinafine and NB598.

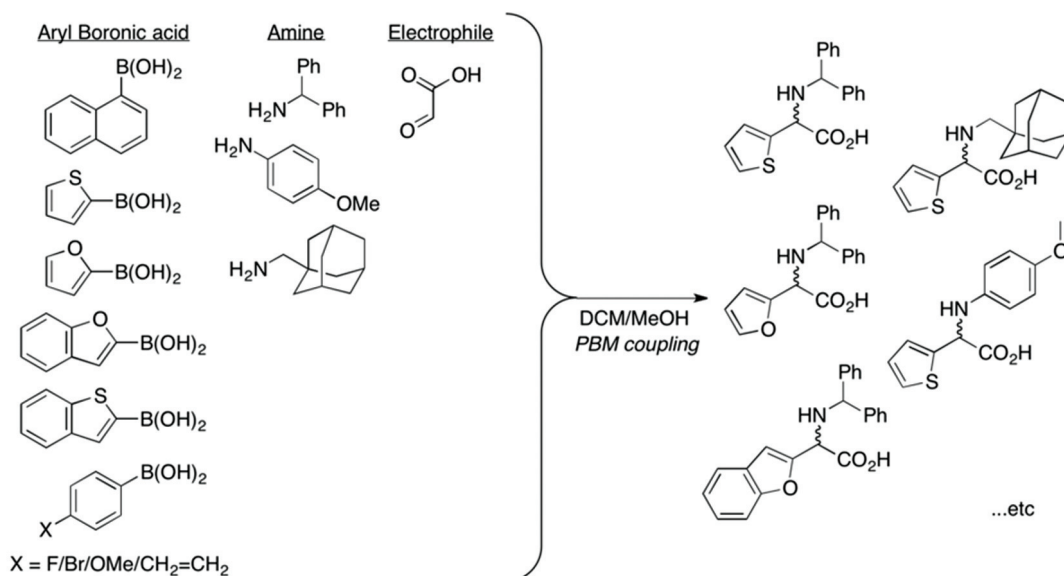


Synthesis of Amino Acids

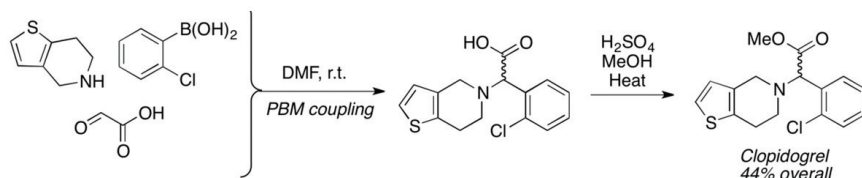
β,γ -unsaturated, N-substituted amino acids are conveniently prepared through the condensation of organoboronic acids, boronates, or boronic esters with amines and glyoxylic acids. Yields are typically in the range of 60–80%, and a wide variety of polar or non-polar solvents can be employed (although DCM and MeOH is typically used). Free amino acids that do not have N-substitutions can be prepared by using trityl amine or bis-(4-methoxyphenyl)methyl amine, followed by deprotection under acidic conditions. Piettre and coworkers found out that the usage of highly polar protic solvents like Hexafluoroisopropanol (HFIP) can shorten reaction time and improve yield. Microwave irradiation was also used to promote the reaction in methanol.



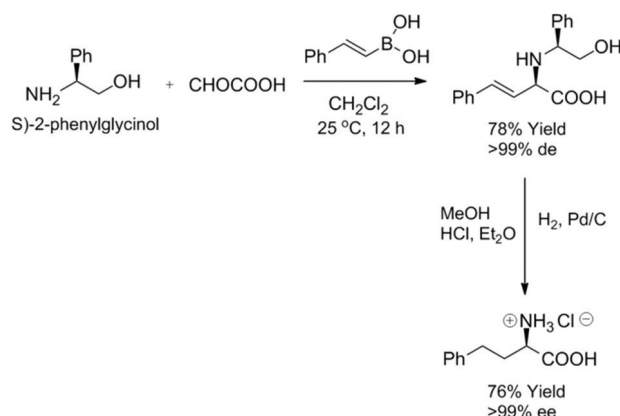
Apart from vinyl boronic acids, aryl boronic acids and other heterocyclic derivatives can also be used in Petasis multicomponent coupling. Possible substrate scope includes thienyl, pyridyl, furyl, and benzofuranyl, 1-naphthyl, and aryl groups with either electron-donating or electron-withdrawing substituent.



Clopidogrel, an antiplatelet agent, was racemically synthesized by Kalinski and coworkers in two steps, using Petasis reaction as the key strategy. Acid-catalyzed esterification immediately following the multicomponent coupling steps to afford Clopidogrel in 44% overall yield.

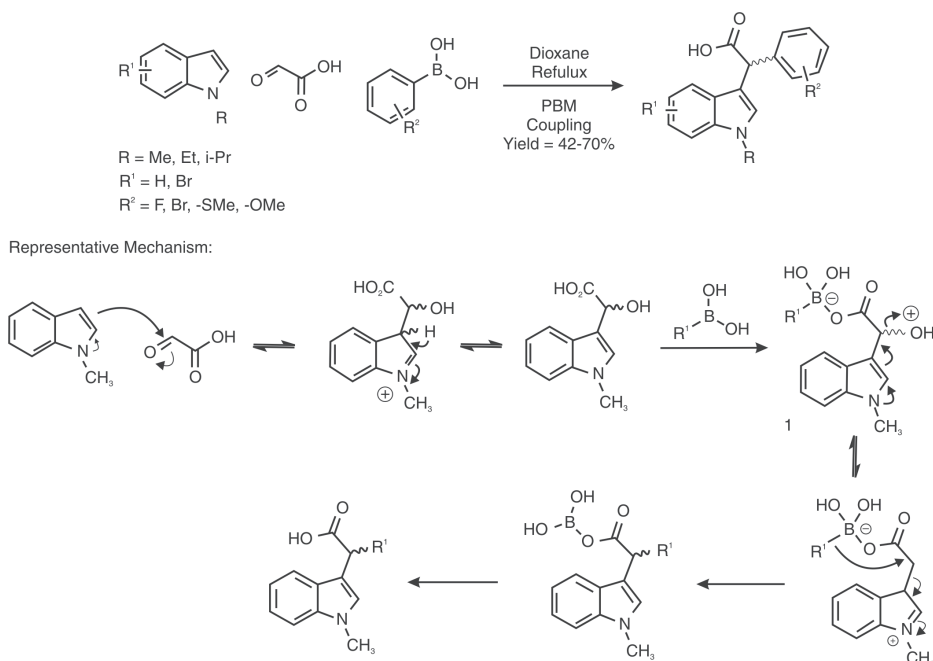


The Petasis reaction exhibits high degrees of stereocontrol when a chiral amine or aldehyde is used as a substrate. When certain chiral amines, such as (S)-2-phenylglycinol, are mixed with an α -keto acid and vinyl boronic acid at room temperature, the corresponding allylamine is formed as a single diastereomer. Furthermore, enantiomeric purity can be achieved by hydrogenation of the diastereoselective product. In the reaction with (S)-2-phenylglycinol, (R)-2-phenylglycinol is generated in 76% yield.

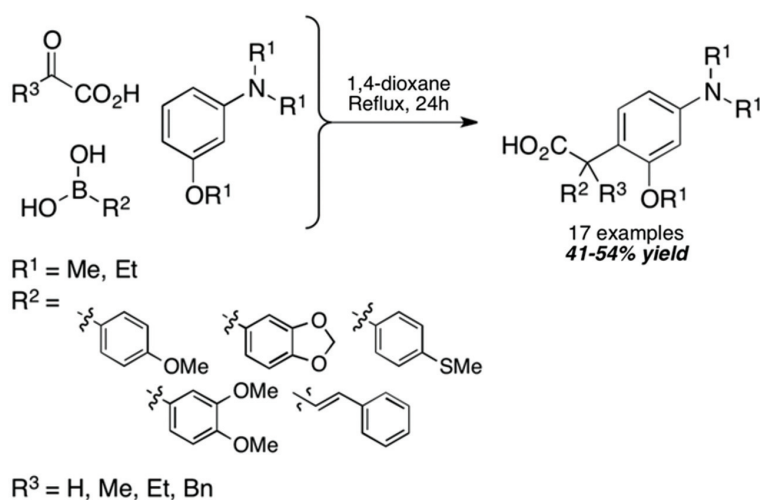


Unconventional Synthesis of Carboxylic Acids

Apart from amino-acids, Petasis borono-Mannich reaction can also be used to prepare carboxylic acids, albeit with unconventional mechanisms. Naskar *et al.* reported the use of *N*-substituted indoles as amine equivalent.

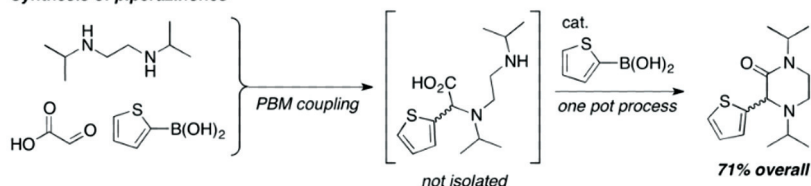
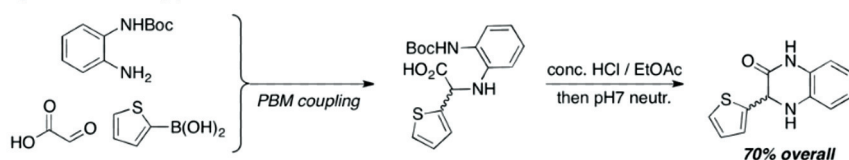
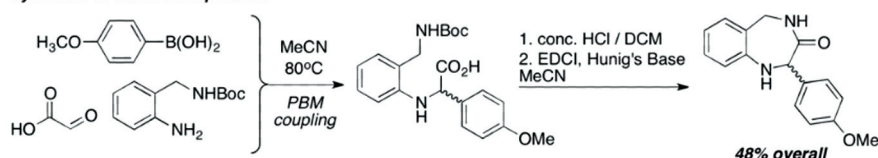


Naskar *et al.* also proposed the usage of tertiary aromatic amines in Petasis reaction as another equivalent of amine nucleophile. The mechanism is similar to the N-substituted indole case. The reaction is carried out under harsh conditions (24-hr reflux in 1,4-dioxane), but the resultant carboxylic acid is obtained in reasonable yield (41–54% yield). Note that the usage of α -ketoacids instead of glyoxylic acid does not diminish yields. 1,3,5-trioxygenated benzene derivatives can also be used in lieu of tertiary aromatic amines.

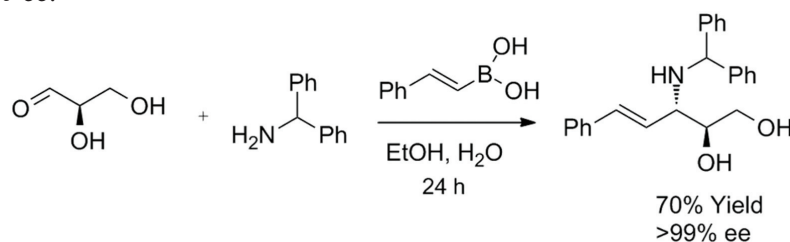


When used as nitrogen nucleophiles, amino acids can furnish various iminodicarboxylic acid derivatives. High diastereoselectivity is usually observed, and the newly formed stereocenter usually share the same configuration with the starting amino acid.

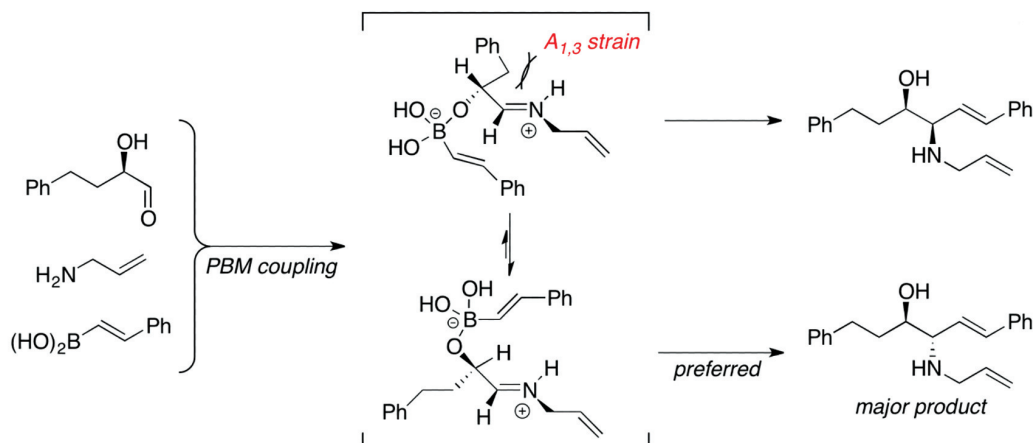
When diamines are used in PBM reactions, heterocycles of various structures, such as piperazinones, benzopiperazinones, and benzodiazepinones, are efficiently prepared. Lactamization reactions are commonly employed to form the heterocycles, usually under strongly acidic conditions.

Synthesis of piperazinones**Synthesis of benzopiperazinones****Synthesis of benzodiazepinones****Synthesis of Amino Alcohols**

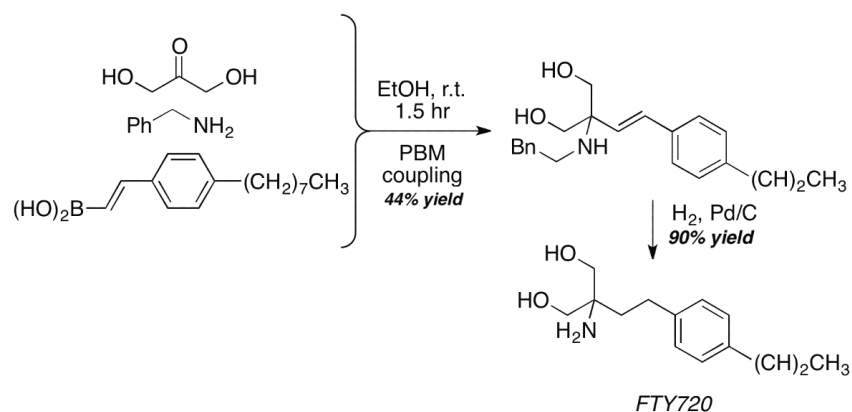
When a α -hydroxy aldehyde is used as a substrate in the synthesis of β -amino alcohols, a single diastereomer is generated. This reaction forms exclusively anti-product, confirmed by ^1H NMR spectroscopy. The product does not undergo racemization, and when enantiomerically pure α -hydroxy aldehydes are used, enantiomeric excess can be achieved. It is believed that the boronic acid first reacted with the chiral hydroxyl group, furnishing a nucleophilic alkenyl boronate, followed by face selective, intramolecular migration of the alkenyl group into the electrophilic iminium carbon, forming the desired C-C bond irreversibly. In the reaction of enantiomerically pure glyceraldehydes, the corresponding 3-amino 1,2-diol product is formed in 70% yield and greater than 99% ee.



Pyne and coworkers suggested that diastereoselectivity arises from the reaction of the more stable (and, in this case, more reactive) conformation of the ate complex, where 1,3 allylic strain is minimized.

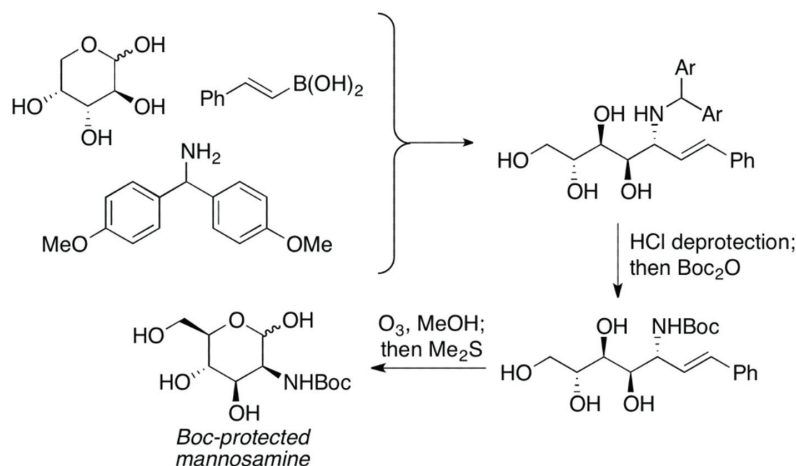


Using dihydroxyacetone, a somewhat unconventional aldehyde equivalent, Sugiyama *et al.* is able to use Petasis reaction to assemble the core structure of FTY720 (a potent immunosuppressive agent) in 40% yield. A straightforward hydrogenation then follows to afford the product via a one-step benzyl-group removal and C-C double bond hydrogenation.



Synthesis of Amino Polyols and Amino Sugars

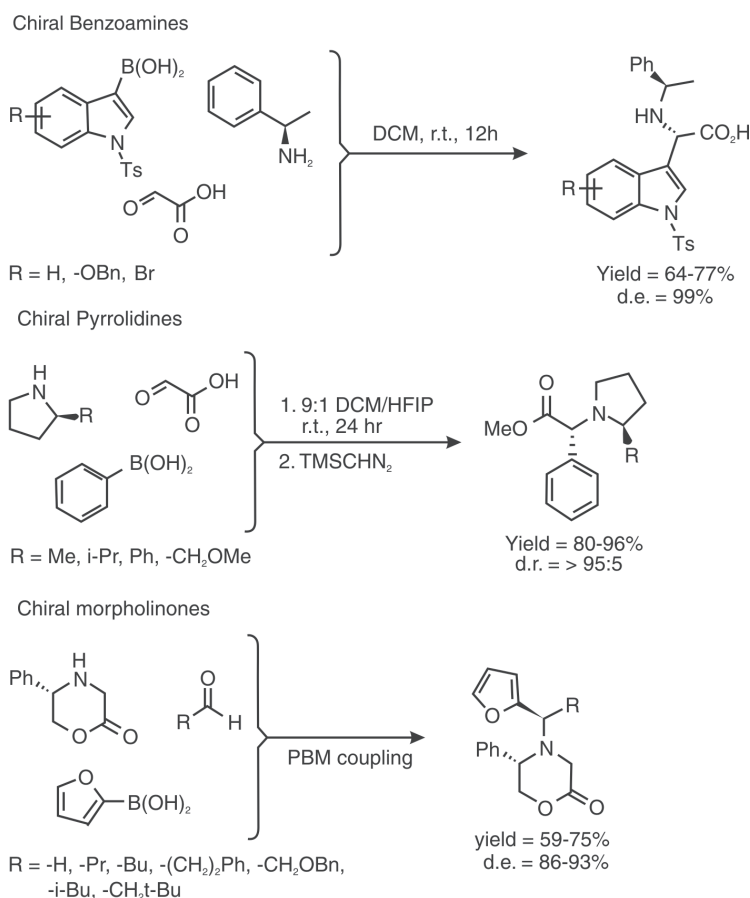
Petasis and coworkers reported the usage of unprotected carbohydrates as the carbonyl component in PBM reactions. It is used as the equivalent of α -hydroxyl aldehydes with pre-existing chirality, and the aminopolyol product is usually furnished with moderate to good yield, with excellent selectivity. A wide variety of carbohydrates, as well as nitrogen nucleophiles (ex. amino acids), can be used to furnish highly stereochemically-enriched products. The aminopolyol products can then undergo further reactions to prepare aminosugars. Petasis used this reaction to prepare Boc-protected mannosamine from D-arabinose.



Applications in Enantioselective Synthesis

With Chiral Amine Nucleophile

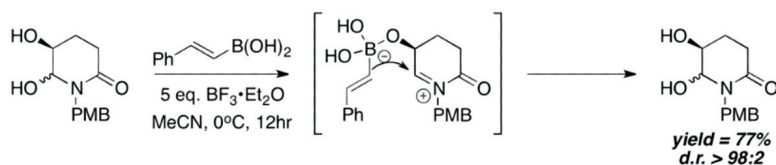
Generally speaking, when chiral amine is used in Petasis coupling, the stereochemical outcome of Petasis reaction is strongly correlated to the chirality of the amine, and high to excellent diastereoselectivity is observed even without the usage of bulky chiral inducing groups. Chiral benzyl amines, 2-substituted pyrrolidines, and 5-substituted 2-morpholinones have been shown to induce good to excellent diastereomeric excess under different Petasis reaction conditions.



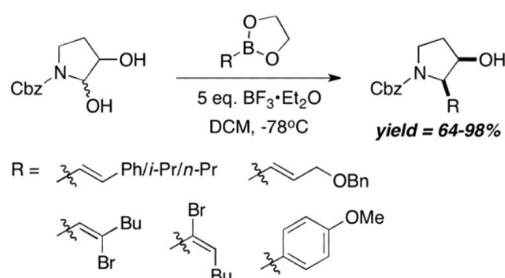
With Chiral N-acyliminium Ions

Chiral N-acyliminium ion “starting materials” are generally prepared by in-situ dehydration of cyclic hemiaminal. They also carry a chiral hydroxyl group that is in proximity with the iminium carbon; boronic acids react with such chiral hydroxyl groups to form a chiral and electron-rich boronate species, followed by side-selective and intramolecular boronate vinyl/aryl transfer into the iminium carbon. Hence, the reaction is highly diastereoselective, with *cis*-boronate aryl/vinyl transfer being the predominant pathway. Hydroxypyrrolidines and Hydroxy- γ - and δ -lactams have been shown to react very diastereoselectively, with good to excellent yield. However, such procedures are limited to the usage of vinyl- or electron-rich aryl-boronic acids only.

Hydroxy δ -lactam as substrates:



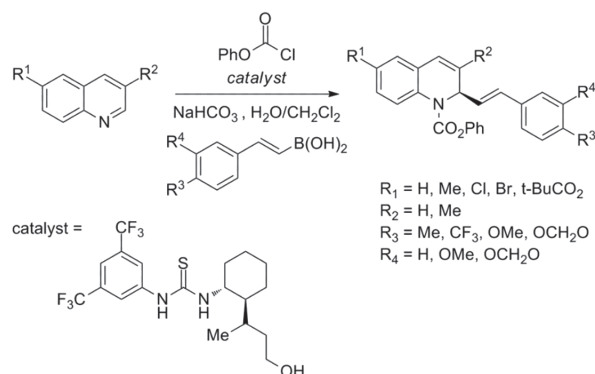
Hydroxy pyrrolidines as substrates:



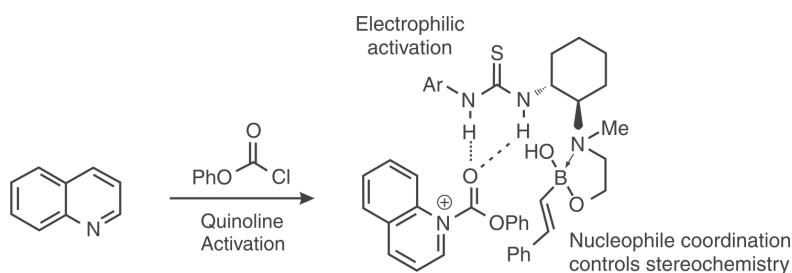
Batey and coworkers take advantage of the high diastereoselectivity of this reaction to prepare ($\pm$)-6-deoxycastanospermine in 7 steps, with an impressive overall yield of 32% (from the vinyl boronic ester). The key acyclic precursor to deoxycastanospermine (A) is formed first by condensing vinyl boronic ester 1 with Cbz-protected hydroxy-pyrrolidine 2 with a PBM coupling, followed by dihydroxylation and TBS protection. A then undergo intramolecular cyclization via a one-pot imine formation and reduction sequel, followed by TBS deprotection, to afford ($\pm$)-6-deoxycastanospermine.

With Thiourea Catalyst

Takemoto and coworkers of Kyoto University recently reported an enantioselective Petasis-type reaction to transform quinolines into respective chiral 1,2-dihydroquinolines (product) using alkenyl boronic acids and chiral thiourea catalyst. Good yields (59–78%) and excellent enantioselectivities (82–96%) are reported.



Takemoto and coworkers observed that addition of chloroformates are required as electrophilic activating agents, and the reaction does not proceed without them. Also, a 1,2-amino alcohol functionality is required on the catalyst for the reaction to proceed stereoselectively. They rationalize these findings by suggesting that the chloroformate reagent reacted with the quinoline nitrogen to make an N-acylated quinolinium intermediate B, which is further activated by electrophilic chiral thiourea. They also suggest that the 1,2-amino alcohol functionality of the catalyst is chelating to the alkenyl boronic acids and that such chelation directed the stereochemical outcome.

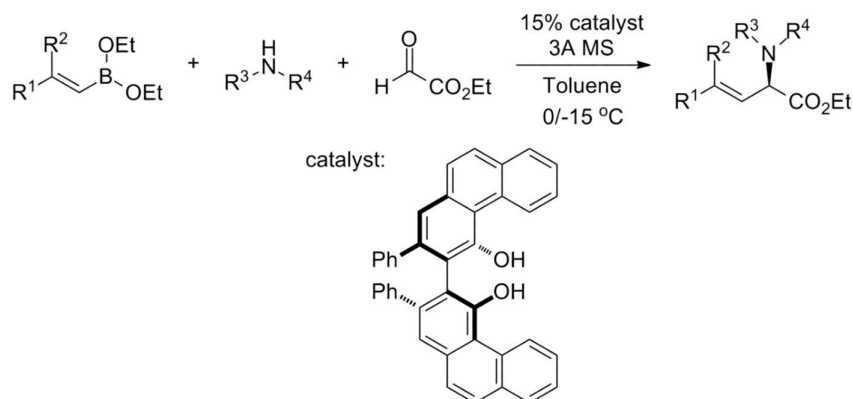


With Chiral Biphenols

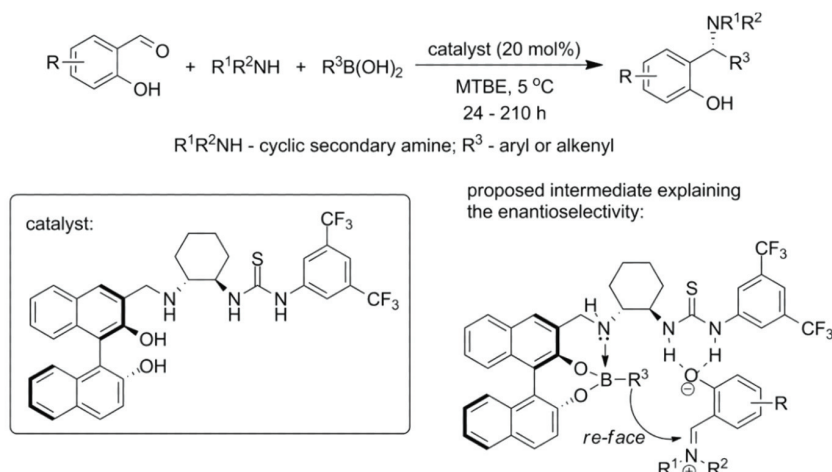
Schaus and Lou of Boston University reported the following reaction, in which chiral α -amino acids with various functionalities are conveniently furnished by mixing alkenyl diethyl boronates, secondary amines, glyoxylates and chiral biphenol catalyst in toluene in one-pot:

This reaction tolerates a wide range of functionalities, both on the sides of alkenyl boronates and the secondary amine: the electron-richness of the substrates does not affect the yield and enantioselectivity, and sterically demanding substrates (dialkylsubstituted alkenyl boronates and amines with α -stereocenter) only compromise enantioselectivity slightly. Reaction rates do vary on a case-by-case basis. Interestingly, under the reported condition, boronic acids substrates failed to give any enantioselectivity. Also, 3Å molecular sieve is used in the reaction system. While the authors did not provide the reason for such usage in the paper, it was speculated that

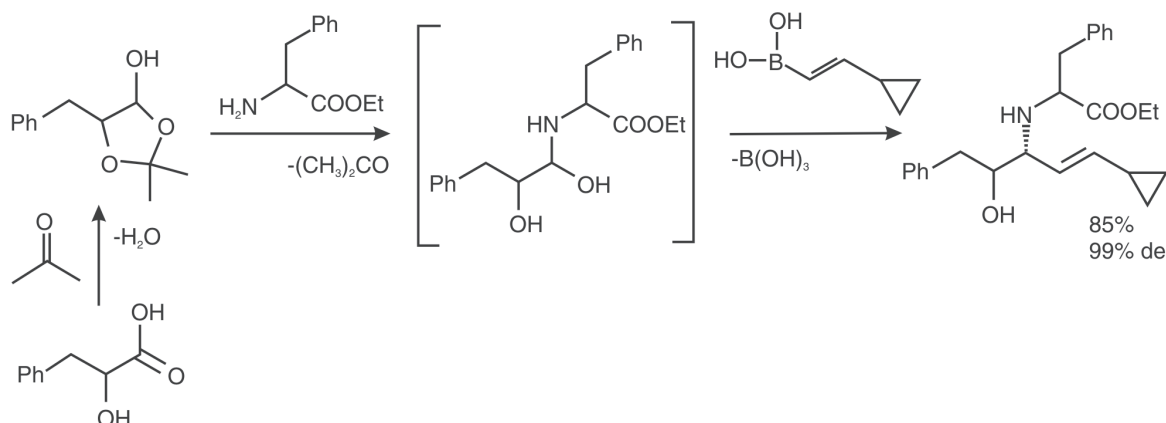
3 Å molecular sieves act as water scavenger and prevent the decomposition of alkenyl diethyl boronates into their respective boronic acids. The catalyst could be recycled from the reaction and reused without compromising yield or enantioselectivity.



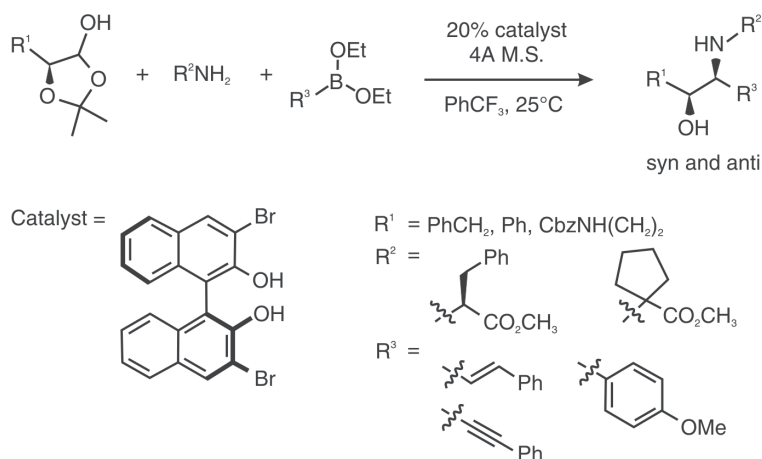
More recently, Yuan with coworkers from Chengdu Institute of Organic Chemistry, Chinese Academy of Science combined both approaches (chiral thiourea catalyst and chiral biphenol) in a single catalyst, reporting for the first time the catalytic system that is capable of performing enantioselective Petasis reaction between salicylaldehydes, cyclic secondary amines and aryl- or alkenylboronic acids:



In one application the Petasis reaction is used for quick access to a multifunctional scaffold for divergent synthesis. The reactants are the lactol derived from L-phenyl-lactic acid and acetone, L-phenylalanine methyl ester and a boronic acid. The reaction takes place in ethanol at room temperature to give the product, an anti-1,2-amino alcohol with a diastereomeric excess of 99% 002E.



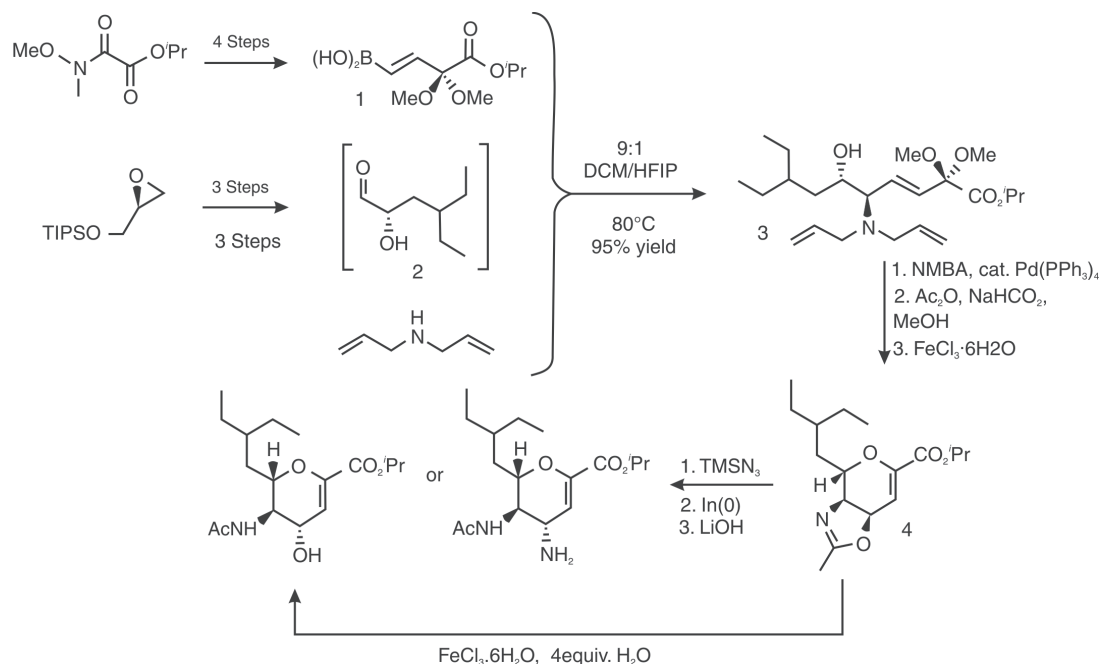
Notice that the authors cannot assess syn-1,2-amino alcohol with this method due to intrinsic mechanistic selectivity, and the authors argue that such intrinsic selectivity hampers their ability to access the full matrix of stereoisomeric products for the usage of small molecule screening. In a recent report, Schaus and co-workers reported that syn amino alcohol can be obtained with the following reaction condition, using a chiral dibromo-biphenol catalyst their group developed:



Although the syn vs. anti diastereomeric ratio ranges from mediocre to good (1.5:1 to 7.5:1), the substrate scope for such reactions remain rather limited, and the diastereoselectivity is found to be dependent on the stereogenic center on the amine starting material.

Petasis Reaction and Total Synthesis

Beau and coworkers assembled the core dihydropyran framework of zanamivir congeners via a combination of PBM reaction and Iron(III)-promoted deprotection-cyclization sequence. A stereochemically-defined α -hydroxyaldehyde 2, diallylamine and a dimethylketal-protected boronic acid 1 is coupled to form the acyclic, stereochemically-defined amino-alcohol 3, which then undergoes an Iron(III)-promoted cyclization to form a bicyclic dihydropyran 4. Selective opening of the oxazoline portion of the dihydropyran intermediate 4 with water or trimethylsilyl azide then furnish downstream products that have structures resembling the Zanamivir family members.



Wong and coworkers prepared N-acetylneuraminic acid with a PBM coupling, followed by nitron-[3+2] cycloaddition. Vinylboronic acid is first coupled with L-arabinose 1 and Bis(4-methoxyphenyl) methanamine 2 to form an stereochemically-defined allyl amine 3. Afterwards, the sequence of dipolar cycloaddition, base-mediated N-O bond breakage and hydrolysis then complete the synthesis of N-acetylneuraminic acid.

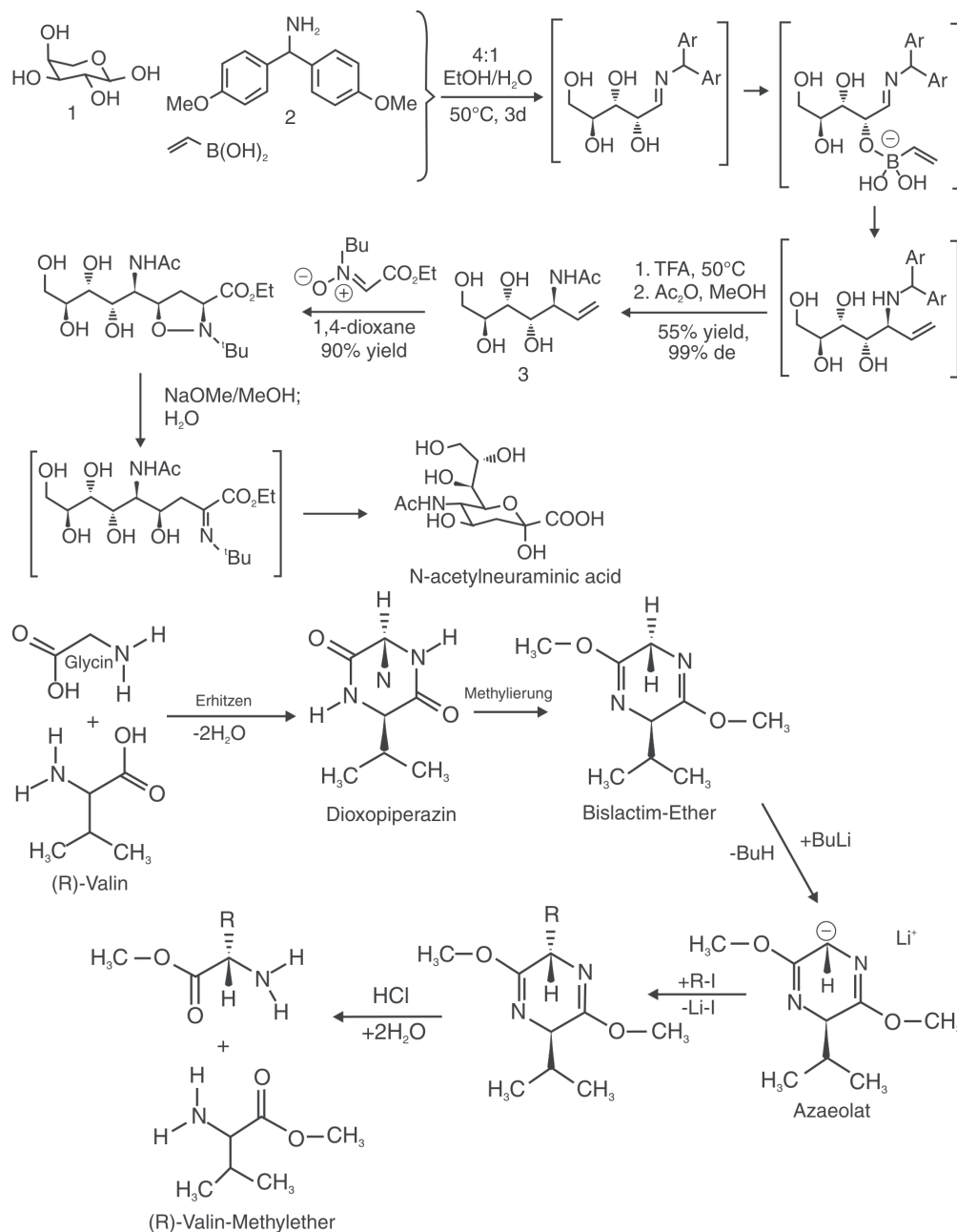
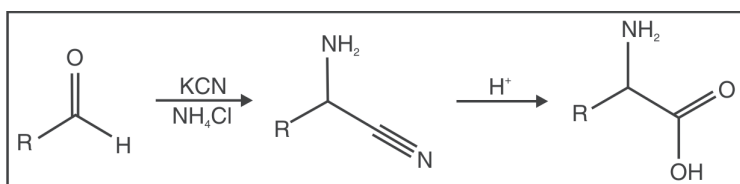


Fig. Reaction mechanism for the Schöllkopf method.

Strecker Amino Acid Synthesis

The Strecker amino acid synthesis, also known simply as the Strecker synthesis, was discovered by German chemist Adolph Strecker, and is a term used for a series of chemical reactions that synthesize an amino acid from an aldehyde or ketone. The aldehyde is condensed with ammonium chloride in the presence of potassium cyanide to form an α -aminonitrile, which is subsequently hydrolyzed to give the desired amino acid. In the

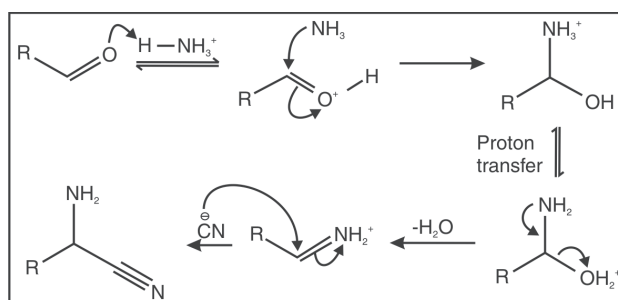
original Strecker reaction acetaldehyde, ammonia, and hydrogen cyanide combined to form after hydrolysis alanine.



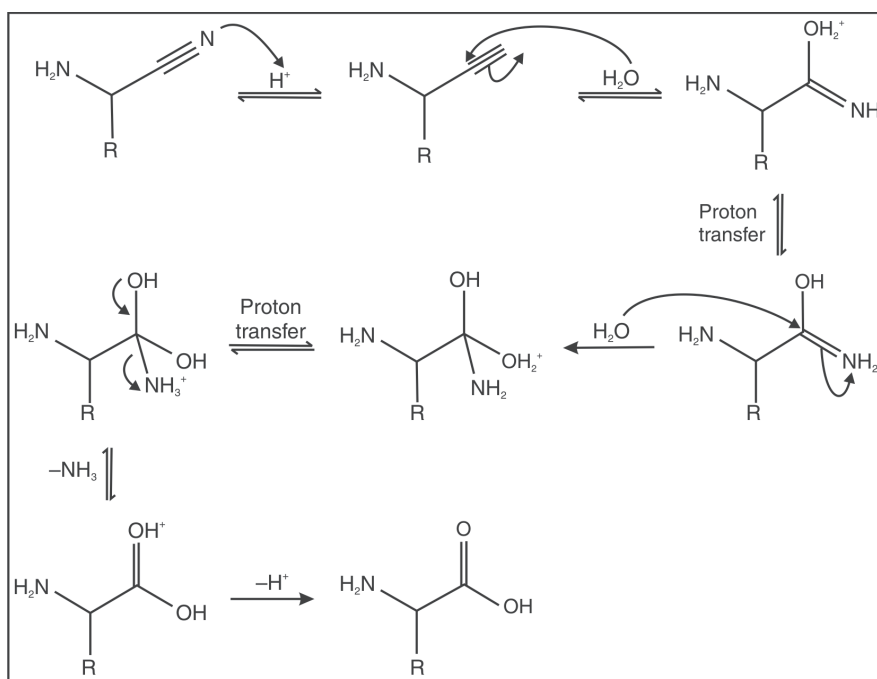
While usage of ammonium salts gives unsubstituted amino acids, primary and secondary amines also successfully give substituted amino acids. Likewise, the usage of ketones, instead of aldehydes, gives α,α -disubstituted amino acids. The traditional synthesis of Adolph Strecker from 1850 gives racemic α -amino nitriles, but several procedures utilizing asymmetric auxiliaries or asymmetric catalysts have been developed.

Reaction Mechanism

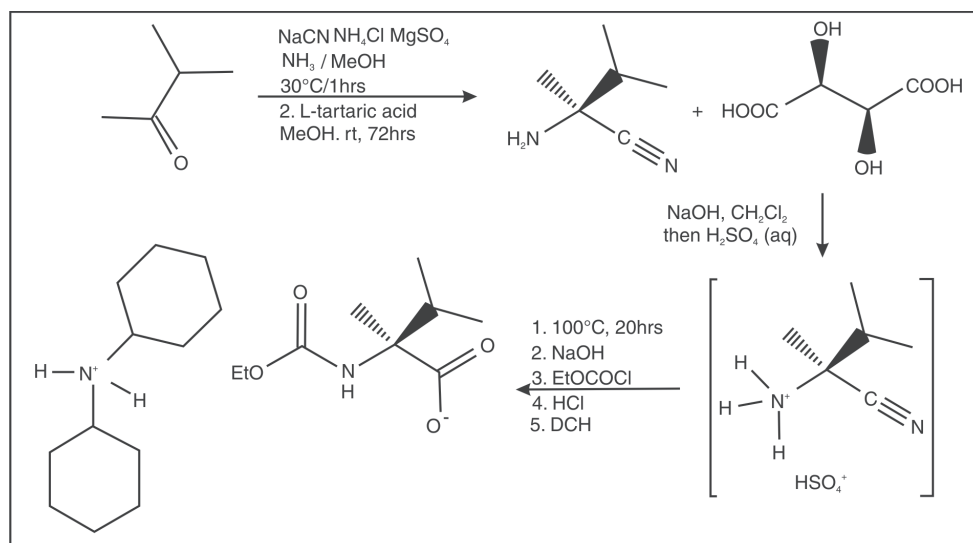
In the first part of the reaction, the carbonyl oxygen of an aldehyde is protonated, followed by a nucleophilic attack of ammonia to the carbonyl carbon. After subsequent proton exchange, water is cleaved from the iminium ion intermediate. A cyanide ion then attacks the iminium carbon yielding an aminonitrile.



In the second part of the Strecker Synthesis the nitrile nitrogen of the aminonitrile is protonated, and the nitrile carbon is attacked by a water molecule. A 1,2-diamino-diol is then formed after proton exchange and a nucleophilic attack of water to the former nitrile carbon. Ammonia is subsequently eliminated after the protonation of the amino group, and finally the deprotonation of a hydroxyl group produces an amino acid.



One example of the Strecker synthesis is a multikilogram scale synthesis of an L-valine derivative starting from Methyl isopropyl ketone:



Asymmetric Strecker Reactions

The asymmetric Strecker reaction was pioneered by Kaoru Harada in 1963. By replacing ammonia with (S)-α-phenylethylamine as chiral auxiliary the ultimate reaction product was chiral alanine. The first asymmetric synthesis via a chiral catalyst was reported in 1996.

Catalytic Asymmetric Strecker Reactions

Catalytic asymmetric Strecker reaction can be effected using thiourea-derived catalyst. In 2012, a BINOL-derived catalyst was employed to generate chiral cyanide anion.

PEPTIDE BOND FORMATION

As both the amine and carboxylic acid groups of amino acids can react to form amide bonds, one amino acid molecule can react with another and become joined through an amide linkage. This polymerization of amino acids is what creates proteins. This condensation reaction yields the newly formed peptide bond and a molecule of water. In cells, this reaction does not occur directly; instead, the amino acid is first activated by attachment to a transfer RNA molecule through an ester bond. This aminoacyl-tRNA is produced in an ATP-dependent reaction carried out by an aminoacyl tRNA synthetase. This aminoacyl-tRNA is then a substrate for the ribosome, which catalyzes the attack of the amino group of the elongating protein chain on the ester bond. As a result of this mechanism, all proteins made by ribosomes are synthesized starting at their N-terminus and moving towards their C-terminus.

However, not all peptide bonds are formed in this way. In a few cases, peptides are synthesized by specific enzymes. For example, the tripeptide glutathione is an essential part of the defenses of cells against oxidative stress. This peptide is synthesized in two steps from free amino acids. In the first step, gamma-glutamylcysteine synthetase condenses cysteine and glutamic acid through a peptide bond formed between the side chain carboxyl of the glutamate (the gamma carbon of this side chain) and the amino group of the cysteine. This dipeptide is then condensed with glycine by glutathione synthetase to form glutathione. In chemistry, peptides are synthesized by a variety of reactions. One of the most-used in solid-phase peptide synthesis uses the aromatic oxime derivatives of amino acids as activated units. These are added in sequence onto the growing peptide chain, which is attached

to a solid resin support. The ability to easily synthesize vast numbers of different peptides by varying the types and order of amino acids (using combinatorial chemistry) has made peptide synthesis particularly important in creating libraries of peptides for use in drug discovery through high-throughput screening.

BIOSYNTHESIS

In plants, nitrogen is first assimilated into organic compounds in the form of glutamate, formed from alpha-ketoglutarate and ammonia in the mitochondrion. In order to form other amino acids, the plant uses transaminases to move the amino group to another alpha-keto carboxylic acid. For example, aspartate aminotransferase converts glutamate and oxaloacetate to alpha-ketoglutarate and aspartate. Other organisms use transaminases for amino acid synthesis, too. Nonstandard amino acids are usually formed through modifications to standard amino acids. For example, homocysteine is formed through the transsulfuration pathway or by the demethylation of methionine via the intermediate metabolite S-adenosyl methionine, while hydroxyproline is made by a posttranslational modification of proline. Microorganisms and plants can synthesize many uncommon amino acids. For example, some microbes make 2-aminoisobutyric acid and lanthionine, which is a sulfide-bridged derivative of alanine. Both of these amino acids are found in peptidic antibiotics such as alamethicin. However, in plants, 1-aminocyclopropane-1-carboxylic acid is a small disubstituted cyclic amino acid that is a key intermediate in the production of the plant hormone ethylene.

CATABOLISM

Amino acids must first pass out of organelles and cells into blood circulation via amino acid transporters, since, the amine and carboxylic acid groups are typically ionized. Degradation of an amino acid, occurring in the liver and kidneys, often involves deamination by moving its amino group to alpha-ketoglutarate, forming glutamate. This process involves transaminases, often the same as those used in amination during synthesis. In many vertebrates, the amino group is then removed through the urea cycle and is excreted in the form of urea. However, amino acid degradation can produce uric acid or ammonia instead. For example, serine dehydratase converts serine to pyruvate and ammonia. After removal of one or more amino groups, the remainder of the molecule can sometimes be used to synthesize new amino acids, or it can be used for energy by entering glycolysis or the citric acid cycle, as detailed in image at right.

PHYSICOCHEMICAL PROPERTIES OF AMINO ACIDS

The 20 amino acids encoded directly by the genetic code can be divided into several groups based on their properties. Important factors are charge, hydrophilicity or hydrophobicity, size, and functional groups. These properties are important for protein structure and protein–protein interactions. The water-soluble proteins tend to have their hydrophobic residues (Leu, Ile, Val, Phe, and Trp) buried in the middle of the protein, whereas hydrophilic side chains are exposed to the aqueous solvent. (Note that in biochemistry, a residue refers to a specific monomer within the polymeric chain of a polysaccharide, protein or nucleic acid.) The integral membrane proteins tend to have outer rings of exposed hydrophobic amino acids that anchor them into the lipid bilayer. In the case part-way between these two extremes, some peripheral membrane proteins have a patch of hydrophobic amino acids on their surface that locks onto the membrane. In similar fashion, proteins that have to bind to positively charged molecules have surfaces rich with negatively charged amino acids like glutamate and aspartate, while proteins binding to negatively charged molecules have surfaces rich with positively charged chains like lysine and arginine. There are different hydrophobicity scales of amino acid residues. Some amino acids have special properties such as cysteine, that can form covalent disulfide bonds to other cysteine residues, proline that forms a cycle to the polypeptide backbone, and glycine that is more flexible than other amino acids. Many

proteins undergo a range of posttranslational modifications, when additional chemical groups are attached to the amino acids in proteins. Some modifications can produce hydrophobic lipoproteins, or hydrophilic glycoproteins. These type of modification allow the reversible targeting of a protein to a membrane. For example, the addition and removal of the fatty acid palmitic acid to cysteine residues in some signaling proteins causes the proteins to attach and then detach from cell membranes.

TABLE OF STANDARD AMINO ACID ABBREVIATIONS AND PROPERTIES

Amino acid	3-letter	1-letter	Side chain class	Side chain polarity	Side chain charge (pH 7.4)	Hydropathy index	Absorbance ϵ_{max} (nm)	$\epsilon_{\text{cat}}/\epsilon_{\text{max}}$ (mM ⁻¹ cm ⁻¹)	MW (weight)	Occurrence in proteins (%)	Coding in the Standard Genetic Code
Alanine	Ala	A	Aliphatic	Nonpolar	Neutral	1.8			89.094	8.76	GCN
Arginine	Arg	R	Basic	Basic polar	Positive	-4.5			174.203	5.78	MGN, CGY
Asparagine	Asn	N	Amide	Polar	Neutral	-3.5			132.119	3.93	AAY
Aspartic acid	Asp	D	Acid	Acidic polar	Negative	-3.5			133.104	5.49	GAY
Cysteine	Cys	C	Sulfur-containing	Nonpolar	Neutral	2.5	250	0.3	121.154	1.38	UGY
Glutamic acid	Glu	E	Acid	Acidic polar	Negative	-3.5			147.131	6.32	GAR
Glutamine	Gln	Q	Amide	Polar	Neutral	-3.5			146.146	3.9	CAR
Glycine	Gly	G	Aliphatic	Nonpolar	Neutral	-0.4			75.067	7.03	GGN
Histidine	His	H	Basic aromatic	Basic polar	Positive (10%) neutral (90%)	-3.2	211	5.9	155.156	2.26	CAY
Isoleucine	Ile	I	Aliphatic	Nonpolar	Neutral	4.5			131.175	5.49	AUH
Leucine	Leu	L	Aliphatic	Nonpolar	Neutral	3.8			131.175	9.68	YUR, CUY
Lysine	Lys	K	Basic	Basic polar	Positive	-3.9			146.189	5.19	AAR
Methionine	Met	M	Sulfur-containing	Nonpolar	Neutral	1.9			149.208	2.32	AUG
Phenylalanine	Phe	F	Aromatic	Nonpolar	Neutral	2.8	257, 206, 188	0.2, 9.3, 60.0	165.192	3.87	UUY
Proline	Pro	P	Cyclic	Nonpolar	Neutral	-1.6			115.132	5.02	CCN
Serine	Ser	S	Hydroxyl-containing	Polar	Neutral	-0.8			105.093	7.14	UCN, AGY
Threonine	Thr	T	Hydroxyl-containing	Polar	Neutral	-0.7			119.119	5.53	ACN
Tryptophan	Trp	W	Aromatic	Nonpolar	Neutral	-0.9	280, 219	5.6, 47.0	204.228	1.25	UGG
Tyrosine	Tyr	Y	Aromatic	Polar	Neutral	-1.3	274, 222, 193	1.4, 8.0, 48.0	181.191	2.91	UAY
Valine	Val	V	Aliphatic	Nonpolar	Neutral	4.2			117.148	6.73	GUN

Two additional amino acids are in some species coded for by codons that are usually interpreted as stop codons:

21st and 22nd amino acids	3-letter	1-letter	MW (weight)
Selenocysteine	Sec	U	168.064
Pyrrolysine	Pyl	O	255.313

Ambiguous amino acids	3-letter	1-letter	Amino Acids Included	Codons
Any / unknown	Xaa	X	All	NNN
Asparagine or aspartic acid	Asx	B	D, N	RAY
Glutamine or glutamic acid	Glx	Z	E, Q	SAR
Leucine or Isoleucine	Xle	J	I, L	
Hydrophobic		Φ	V, I, L, F, W, Y, M	
Aromatic		Ω	F, W, Y, H	
Aliphatic (non-aromatic)		Ψ	V, I, L, M	VUN, UUR
Small		π	P, G, A, S	BCN, RGY, GGR
Hydrophilic		ζ	S, T, H, N, Q, E, D, K, R	
Positively charged		+	K, R, H	ARR, CRY, CGR
Negatively charged		-	D, E	GAN

In addition to the specific amino acid codes, placeholders are used in cases where chemical or crystallographic analysis of a peptide or protein cannot conclusively determine the identity of a residue. They are also used to summarise conserved protein sequence motifs. The use of single letters to indicate sets of similar residues is similar to the use of abbreviation codes for degenerate bases. Unk is sometimes used instead of Xaa, but is less standard. In addition, many non-standard amino acids have a specific code. For example, several peptide drugs, such as Bortezomib and MG132, are artificially synthesized and retain their protecting groups, which have specific codes. Bortezomib is Pyz-Phe-boroLeu, and MG132 is Z-Leu-Leu-Leu-al. To aid in the analysis of protein structure, photo-reactive amino acid analogs are available. These include photoleucine (pLeu) and photomethionine (pMet).

Protein Structure

Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. Proteins are polymers—specifically polypeptides—formed from sequences of amino acids, the monomers of the polymer. A single amino acid monomer may also be called a residue indicating a repeating unit of a polymer. Proteins form by amino acids undergoing condensation reactions, in which the amino acids lose one water molecule per reaction in order to attach to one another with a peptide bond. By convention, a chain under 30 amino acids is often identified as a peptide, rather than a protein. To be able to perform their biological function, proteins fold into one or more specific spatial conformations driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions, Van der Waals forces, and hydrophobic packing. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure. This is the topic of the scientific field of structural biology, which employs techniques such as X-ray crystallography, NMR spectroscopy, and dual polarisation interferometry to determine the structure of proteins.

Protein structures range in size from tens to several thousand amino acids. By physical size, proteins are classified as nanoparticles, between 1–100 nm. Very large aggregates can be formed from protein subunits. For example, many thousands of actin molecules assemble into a microfilament. A protein generally undergo reversible structural changes in performing its biological function. The alternative structures of the same protein are referred to as different conformational isomers, or simply, conformations, and transitions between them are called conformational changes.

LEVELS OF PROTEIN STRUCTURE

There are four distinct levels of protein structure.

PROTEIN PRIMARY STRUCTURE

Protein primary structure is the linear sequence of amino acids in a peptide or protein. By convention, the primary structure of a protein is reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end. Protein biosynthesis is most commonly performed by ribosomes in cells. Peptides can also be synthesized in the laboratory. Protein primary structures can be directly sequenced, or inferred from DNA sequences.

Formation

Biological

Amino acids are polymerised via peptide bonds to form a long backbone, with the different amino acid side chains protruding along it. In biological systems, proteins are produced during translation by a cell's ribosomes.

Some organisms can also make short peptides by non-ribosomal peptide synthesis, which often use amino acids other than the standard 20, and may be cyclised, modified and cross-linked.

Chemical

Peptides can be synthesised chemically via a range of laboratory methods. Chemical methods typically synthesise peptides in the opposite order to biological protein synthesis (starting at the C-terminus).

Notation

Protein sequence is typically notated as a string of letters, listing the amino acids starting at the amino-terminal end through to the carboxyl-terminal end. Either a three letter code or single letter code can be used to represent the 20 naturally occurring amino acids, as well as mixtures or ambiguous amino acids (similar to nucleic acid notation). Peptides can be directly sequenced, or inferred from DNA sequences. Large sequence databases now exist that collate known protein sequences.

20 natural amino acid notation		
Amino Acid	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Ambiguous amino acid notation		
Symbol	Description	Residues represented
X	Any amino acid, or unknown	All
B	Aspartate or Asparagine	D, N
Z	Glutamate or Glutamine	E, Q
J	Leucine or Isoleucine	I, L
Φ	Hydrophobic	V, I, L, F, W, Y, M
Ω	Aromatic	F, W, Y, H
Ψ	Aliphatic	V, I, L, M
π	Small	P, G, A, S
ζ	Hydrophilic	S, T, H, N, Q, E, D, K, R
+	Positively charged	K, R, H
-	Negatively charged	D, E

Modification

In general, polypeptides are unbranched polymers, so their primary structure can often be specified by the sequence of amino acids along their backbone. However, proteins can become cross-linked, most commonly by disulfide bonds, and the primary structure also requires specifying the cross-linking atoms, *e.g.*, specifying the cysteines involved in the protein's disulfide bonds. Other crosslinks include desmosine.

Isomerisation

The chiral centers of a polypeptide chain can undergo racemization. Although it does not change the sequence, it does affect the chemical properties of the sequence. In particular, the L-amino acids normally found in proteins can spontaneously isomerize at the C^α atom to form D-amino acids, which cannot be cleaved by most proteases. Additionally, proline can form stable trans-isomers at the peptide bond.

Posttranslational Modification

Finally, the protein can undergo a variety of posttranslational modifications, which are briefly summarized here. The N-terminal amino group of a polypeptide can be modified covalently, *e.g.*,

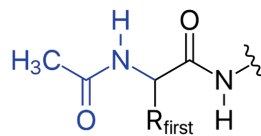


Fig. N-terminal acetylation.

- **Acetylation**— $C(=O)-CH_3$: The positive charge on the N-terminal amino group may be eliminated by changing it to an acetyl group (N-terminal blocking).
- **Formylation**— $C(=O)H$: The N-terminal methionine usually found after translation has an N-terminus blocked with a formyl group. This formyl group (and sometimes the methionine residue itself, if followed by Gly or Ser) is removed by the enzyme deformylase.
- **Pyroglutamate**

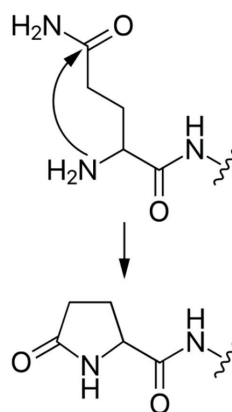


Fig. Formation of pyroglutamate from an N-terminal glutamine.

An N-terminal glutamine can attack itself, forming a cyclic pyroglutamate group.

- **Myristoylation**— $C(=O)-(CH_2)_{12}-CH_3$: Similar to acetylation. Instead of a simple methyl group, the myristoyl group has a tail of 14 hydrophobic carbons, which make it ideal for anchoring proteins to cellular membranes.

The C-terminal carboxylate group of a polypeptide can also be modified, *e.g.*,

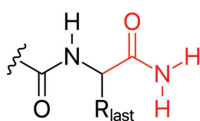


Fig. C-terminal amidation.

- *Amidation:* The C-terminus can also be blocked (thus, neutralizing its negative charge) by amidation.
- *Glycosyl phosphatidylinositol (GPI) attachment:* Glycosyl phosphatidylinositol is a large, hydrophobic phospholipid prosthetic group that anchors proteins to cellular membranes. It is attached to the polypeptide C-terminus through an amide linkage that then connects to ethanolamine, thence to sundry sugars and finally to the phosphatidylinositol lipid moiety.

Finally, the peptide side chains can also be modified covalently, *e.g.*,

- *Phosphorylation:* Aside from cleavage, phosphorylation is perhaps the most important chemical modification of proteins. A phosphate group can be attached to the sidechain hydroxyl group of serine, threonine and tyrosine residues, adding a negative charge at that site and producing an unnatural amino acid. Such reactions are catalyzed by kinases and the reverse reaction is catalyzed by phosphatases. The phosphorylated tyrosines are often used as “handles” by which proteins can bind to one another, whereas phosphorylation of Ser/Thr often induces conformational changes, presumably because of the introduced negative charge. The effects of phosphorylating Ser/Thr can sometimes be simulated by mutating the Ser/Thr residue to glutamate.
- *Glycosylation:* A catch-all name for a set of very common and very heterogeneous chemical modifications. Sugar moieties can be attached to the sidechain hydroxyl groups of Ser/Thr or to the sidechain amide groups of Asn. Such attachments can serve many functions, ranging from increasing solubility to complex recognition. All glycosylation can be blocked with certain inhibitors, such as tunicamycin.
- *Deamidation (succinimide formation):* In this modification, an asparagine or aspartate side chain attacks the following peptide bond, forming a symmetrical succinimide intermediate. Hydrolysis of the intermediate produces either aspartate or the β -amino acid, iso(Asp). For asparagine, either product results in the loss of the amide group, hence, “deamidation”.
- *Hydroxylation:* Proline residues may be hydroxylated at either of two atoms, as can lysine (at one atom). Hydroxyproline is a critical component of collagen, which becomes unstable upon its loss. The hydroxylation reaction is catalyzed by an enzyme that requires ascorbic acid (vitamin C), deficiencies in which lead to many connective-tissue diseases such as scurvy.
- *Methylation:* Several protein residues can be methylated, most notably the positive groups of lysine and arginine. Arginine residues interact with the nucleic acid phosphate backbone and commonly form hydrogen bonds with the base residues, particularly guanine, in protein–DNA complexes. Lysine residues can be singly, doubly and even triply methylated. Methylation does *not* alter the positive charge on the side chain, however.
- *Acetylation:* Acetylation of the lysine amino groups is chemically analogous to the acetylation of the N-terminus. Functionally, however, the acetylation of lysine residues is used to regulate the binding of proteins to nucleic acids. The cancellation of the positive charge on the lysine weakens the electrostatic attraction for the (negatively charged) nucleic acids.
- *Sulfation:* Tyrosines may become sulfated on their O^{η} atom. Somewhat unusually, this modification occurs in the Golgi apparatus, not in the endoplasmic reticulum. Similar to phosphorylated tyrosines, sulfated tyrosines are used for specific recognition, *e.g.*, in chemokine receptors on the cell surface. As with phosphorylation, sulfation adds a negative charge to a previously neutral site.
- *Prenylation and palmitoylation*— $C(=O)-(CH_2)_{14}-CH_3$: The hydrophobic isoprene (*e.g.*, farnesyl, geranyl, and geranylgeranyl groups) and palmitoyl groups may be added to the S^{γ} atom of cysteine

residues to anchor proteins to cellular membranes. Unlike the GPI and myristoyl anchors, these groups are not necessarily added at the termini.

- *Carboxylation*: A relatively rare modification that adds an extra carboxylate group (and, hence, a double negative charge) to a glutamate side chain, producing a Gla residue. This is used to strengthen the binding to “hard” metal ions such as calcium.
- *ADP-ribosylation*: The large ADP-ribosyl group can be transferred to several types of side chains within proteins, with heterogeneous effects. This modification is a target for the powerful toxins of disparate bacteria, *e.g.*, *Vibrio cholerae*, *Corynebacterium diphtheriae* and *Bordetella pertussis*.
- *Ubiquitination and SUMOylation*: Various full-length, folded proteins can be attached at their C-termini to the sidechain ammonium groups of lysines of other proteins. Ubiquitin is the most common of these, and usually signals that the ubiquitin-tagged protein should be degraded.

Most of the polypeptide modifications listed above occur *post-translationally*, *i.e.*, after the protein has been synthesized on the ribosome, typically occurring in the endoplasmic reticulum, a subcellular organelle of the eukaryotic cell. Many other chemical reactions (*e.g.*, cyanylation) have been applied to proteins by chemists, although they are not found in biological systems.

Cleavage and Ligation

In addition to those listed above, the most important modification of primary structure is peptide cleavage (by chemical hydrolysis or by proteases). Proteins are often synthesized in an inactive precursor form; typically, an N-terminal or C-terminal segment blocks the active site of the protein, inhibiting its function. The protein is activated by cleaving off the inhibitory peptide. Some proteins even have the power to cleave themselves. Typically, the hydroxyl group of a serine (rarely, threonine) or the thiol group of a cysteine residue will attack the carbonyl carbon of the preceding peptide bond, forming a tetrahedrally bonded intermediate [classified as a hydroxyoxazolidine (Ser/Thr) or hydroxythiazolidine (Cys) intermediate]. This intermediate tends to revert to the amide form, expelling the attacking group, since, the amide form is usually favored by free energy, (presumably due to the strong resonance stabilization of the peptide group).

However, additional molecular interactions may render the amide form less stable; the amino group is expelled instead, resulting in an ester (Ser/Thr) or thioester (Cys) bond in place of the peptide bond. This chemical reaction is called an N-O acyl shift.

The ester/thioester bond can be resolved in several ways:

- Simple hydrolysis will split the polypeptide chain, where the displaced amino group becomes the new N-terminus. This is seen in the maturation of glycosylasparaginase.
- A β -elimination reaction also splits the chain, but results in a pyruvoyl group at the new N-terminus. This pyruvoyl group may be used as a covalently attached catalytic cofactor in some enzymes, especially decarboxylases such as S-adenosylmethionine decarboxylase (SAMDC) that exploit the electron-withdrawing power of the pyruvoyl group.
- Intramolecular transesterification, resulting in a *branched* polypeptide. In inteins, the new ester bond is broken by an intramolecular attack by the soon-to-be C-terminal asparagine.
- Intermolecular transesterification can transfer a whole segment from one polypeptide to another, as is seen in the Hedgehog protein autoprocessing.

History

The proposal that proteins were linear chains of α -amino acids was made nearly simultaneously by two scientists at the same conference in 1902, the 74th meeting of the Society of German Scientists and Physicians,

held in Karlsbad. Franz Hofmeister made the proposal in the morning, based on his observations of the biuret reaction in proteins. Hofmeister was followed a few hours later by Emil Fischer, who had amassed a wealth of chemical details supporting the peptide-bond model. For completeness, the proposal that proteins contained amide linkages was made as early as 1882 by the French chemist E. Grimaux.

Despite these data and later evidence that proteolytically digested proteins yielded only oligopeptides, the idea that proteins were linear, unbranched polymers of amino acids was not accepted immediately. Some well-respected scientists such as William Astbury doubted that covalent bonds were strong enough to hold such long molecules together; they feared that thermal agitations would shake such long molecules asunder. Hermann Staudinger faced similar prejudices in the 1920s when he argued that rubber was composed of macromolecules.

Thus, several alternative hypotheses arose. The colloidal protein hypothesis stated that proteins were colloidal assemblies of smaller molecules. This hypothesis was disproved in the 1920s by ultracentrifugation measurements by Theodor Svedberg that showed that proteins had a well-defined, reproducible molecular weight and by electrophoretic measurements by Arne Tiselius that indicated that proteins were single molecules. A second hypothesis, the cyclol hypothesis advanced by Dorothy Wrinch, proposed that the linear polypeptide underwent a chemical cyclol rearrangement $C=O + HN \rightarrow C(OH)-N$ that crosslinked its backbone amide groups, forming a two-dimensional *fabric*. Other primary structures of proteins were proposed by various researchers, such as the diketopiperazine model of Emil Abderhalden and the pyrrol/piperidine model of Troensegaard in 1942. Although never given much credence, these alternative models were finally disproved when Frederick Sanger successfully sequenced insulin and by the crystallographic determination of myoglobin and hemoglobin by Max Perutz and John Kendrew.

Primary Structure in Other Molecules

Any linear-chain heteropolymer can be said to have a “primary structure” by analogy to the usage of the term for proteins, but this usage is rare compared to the extremely common usage in reference to proteins. In RNA, which also has extensive secondary structure, the linear chain of bases is generally just referred to as the “sequence” as it is in DNA (which usually forms a linear double helix with little secondary structure). Other biological polymers such as polysaccharides can also be considered to have a primary structure, although the usage is not standard.

Relation to Secondary and Tertiary Structure

The primary structure of a biological polymer to a large extent determines the three-dimensional shape (tertiary structure). Protein sequence can be used to predict local features, such as segments of secondary structure, or trans-membrane regions. However, the complexity of protein folding currently prohibits predicting the tertiary structure of a protein from its sequence alone. Knowing the structure of a similar homologous sequence (for example a member of the same protein family) allows highly accurate prediction of the tertiary structure by homology modeling. If the full-length protein sequence is available, it is possible to estimate its general biophysical properties, such as its isoelectric point. Sequence families are often determined by sequence clustering, and structural genomics projects aim to produce a set of representative structures to cover the sequence space of possible non-redundant sequences.

PROTEIN SECONDARY STRUCTURE

Protein secondary structure is the three dimensional form of *local segments* of proteins. The two most common secondary structural elements are alpha helices and beta sheets, though beta turns and omega loops

occur as well. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its three dimensional tertiary structure. Secondary structure is formally defined by the pattern of hydrogen bonds between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone. Secondary structure may alternatively be defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot regardless of whether it has the correct hydrogen bonds. The concept of secondary structure was first introduced by Kaj Ulrik Linderstrøm-Lang at Stanford in 1952. Other types of biopolymers such as nucleic acids also possess characteristic secondary structures.

Types

The most common secondary structures are alpha helices and beta sheets. Other helices, such as the 3_{10} helix and π helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix. Other extended structures such as the polyproline helix and alpha sheet are rare in native state proteins but are often hypothesized as important protein folding intermediates. Tight turns and loose, flexible loops link the more “regular” secondary structure elements. The random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure.

Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as “helix breakers” because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine (“MALEK” in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine and phenylalanine) and C^β -branched amino acids (isoleucine, valine, and threonine) prefer to adopt β -strand conformations. However, these preferences are not strong enough to produce a reliable method of predicting secondary structure from sequence alone. Low frequency collective vibrations are thought to be sensitive to local rigidity within proteins, revealing beta structures to be generically more rigid than alpha or disordered proteins. Neutron scattering measurements have directly connected the spectral feature at ~ 1 THz to collective motions of the secondary structure of beta-barrel protein GFP. Hydrogen bonding patterns in secondary structures may be significantly distorted, which makes automatic determination of secondary structure difficult. There are several methods for formally defining protein secondary structure (*e.g.*, DSSP, DEFINE, STRIDE, ScrewFit, SST).

Structural features of the three major forms of protein helices			
Geometry attribute	α -helix	3_{10} helix	π -helix
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å (0.15 nm)	2.0 Å (0.20 nm)	1.1 Å (0.11 nm)
Radius of helix	2.3 Å (0.23 nm)	1.9 Å (0.19 nm)	2.8 Å (0.28 nm)
Pitch	5.4 Å (0.54 nm)	6.0 Å (0.60 nm)	4.8 Å (0.48 nm)

DSSP Classification

The Dictionary of Protein Secondary Structure, in short DSSP, is commonly used to describe the protein secondary structure with single letter codes. The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling *et al.* in 1951 (before any protein structure had ever been experimentally determined).

There are eight types of secondary structure that DSSP defines:

- G = 3-turn helix (3_{10} helix). Min length 3 residues.
- H = 4-turn helix (α helix). Minimum length 4 residues.
- I = 5-turn helix (π helix). Minimum length 5 residues.

- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand in parallel and/or anti-parallel β -sheet conformation. Min length 2 residues.
- B = residue in isolated β -bridge (single pair β -sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment).
- C = coil (residues which are not in any of the above conformations).

‘Coil’ is often codified as ‘ ’ (space), C (coil) or ‘–’ (dash). The helices (G, H and I) and sheet conformations are all required to have a reasonable length. This means that 2 adjacent residues in the primary structure must form the same hydrogen bonding pattern. If the helix or sheet hydrogen bonding pattern is too short they are designated as T or B, respectively. Other protein secondary structure assignment categories exist (sharp turns, Omega loops, *etc.*), but they are less frequently used.

Secondary structure is defined by hydrogen bonding, so the exact definition of a hydrogen bond is critical. The standard hydrogen-bond definition for secondary structure is that of DSSP, which is a purely electrostatic model. It assigns charges of $\pm q_1 \approx H-0.42e$ to the carbonyl carbon and oxygen, respectively, and charges of $\pm q_2 H-0.20e$ to the amide hydrogen and nitrogen, respectively.

The electrostatic energy is,

$$E = q_1 q_2 \left(\frac{1}{r_{\text{ON}}} + \frac{1}{r_{\text{CH}}} - \frac{1}{r_{\text{OH}}} - \frac{1}{r_{\text{CN}}} \right) \cdot 332 \text{ kcal / mol.}$$

According to DSSP, a hydrogen-bond exists if and only if E is less than -0.5 kcal/mol (-2.1 kJ/mol). Although the DSSP formula is a relatively crude approximation of the *physical* hydrogen-bond energy, it is generally accepted as a tool for defining secondary structure.

SST Classification

SST is a Bayesian method to assign secondary structure to protein coordinate data using the Shannon information criterion of Minimum Message Length (MML) inference. SST treats any assignment of secondary structure as a potential hypothesis that attempts to explain (compress) given protein coordinate data. The core idea is that the *best* secondary structural assignment is the one that can explain (compress) the coordinates of a given protein coordinates in the most economical way, thus linking the inference of secondary structure to lossless data compression.

SST accurately delineates any protein chain into regions associated with the following assignment types:

- E = (Extended) strand of a β -pleated sheet
- G = Right-handed 3_{10} helix
- H = Right-handed α -helix
- I = Right-handed π -helix
- g = Left-handed 3_{10} helix
- h = Left-handed α -helix
- i = Left-handed π -helix
- 3 = 3_{10} -like Turn
- 4 = α -like Turn
- 5 = π -like Turn
- T = Unspecified Turn
- C = Coil
- – = Unassigned residue

SST detects π and 3_{10} helical caps to standard α -helices, and automatically assembles the various extended strands into consistent β -pleated sheets. It provides a readable output of dissected secondary structural elements, and a corresponding PyMol-loadable script to visualize the assigned secondary structural elements individually.

Experimental Determination

The rough secondary-structure content of a biopolymer (*e.g.*, “this protein is 40% α -helix and 20% β -sheet.”) can be estimated spectroscopically. For proteins, a common method is far-ultraviolet (far-UV, 170–250 nm) circular dichroism. A pronounced double minimum at 208 and 222 nm indicate α -helical structure, whereas a single minimum at 204 nm or 217 nm reflects random-coil or β -sheet structure, respectively. A less common method is infrared spectroscopy, which detects differences in the bond oscillations of amide groups due to hydrogen-bonding. Finally, secondary-structure contents may be estimated accurately using the chemical shifts of an initially unassigned NMR spectrum.

Prediction

Predicting protein tertiary structure from only its amino acid sequence is a very challenging problem, but using the simpler secondary structure definitions is more tractable. Early methods of secondary-structure prediction were restricted to predicting the three predominate states: helix, sheet, or random coil. These methods were based on the helix-or-sheet-forming propensities of individual amino acids, sometimes coupled with rules for estimating the free energy of forming secondary structure elements. The first widely used techniques to predict protein secondary structure from the amino acid sequence were the Chou–Fasman method and the GOR method. Although such methods claimed to achieve ~60% accurate in predicting which of the three states (helix/sheet/coil) a residue adopts, blind computing assessments later showed that the actual accuracy was much lower. A significant increase in accuracy (to nearly ~80%) was made by exploiting multiple sequence alignment; knowing the full distribution of amino acids that occur at a position (and in its vicinity, typically ~7 residues on either side) throughout evolution provides a much better picture of the structural tendencies near that position. For illustration, a given protein might have a glycine at a given position, which by itself might suggest a random coil there. However, multiple sequence alignment might reveal that helix-favoring amino acids occur at that position (and nearby positions) in 95% of homologous proteins spanning nearly a billion years of evolution. Moreover, by examining the average hydrophobicity at that and nearby positions, the same alignment might also suggest a pattern of residue solvent accessibility consistent with an α -helix. Taken together, these factors would suggest that the glycine of the original protein adopts α -helical structure, rather than random coil. Several types of methods are used to combine all the available data to form a 3-state prediction, including neural networks, hidden Markov models and support vector machines. Modern prediction methods also provide a confidence score for their predictions at every position.

Secondary-structure prediction methods were evaluated by the Critical Assessment of protein Structure Prediction (CASP) experiments and continuously benchmarked, *e.g.*, by EVA (benchmark). Based on these tests, the most accurate methods were Psipred, SAM, PORTER, PROF, and SABLE. The chief area for improvement appears to be the prediction of β -strands; residues confidently predicted as β -strand are likely to be so, but the methods are apt to overlook some β -strand segments (false negatives). There is likely an upper limit of ~90% prediction accuracy overall, due to the idiosyncrasies of the standard method (DSSP) for assigning secondary-structure classes (helix/strand/coil) to PDB structures, against which the predictions are benchmarked. Accurate secondary-structure prediction is a key element in the prediction of tertiary structure, in all but the simplest (homology modeling) cases. For example, a confidently predicted pattern of six secondary structure elements $\beta\alpha\beta\beta\alpha\beta$ is the signature of a ferredoxin fold.

Applications

Both protein and nucleic acid secondary structures can be used to aid in multiple sequence alignment. These alignments can be made more accurate by the inclusion of secondary structure information in addition to simple sequence information. This is sometimes less useful in RNA because base pairing is much more highly conserved than sequence. Distant relationships between proteins whose primary structures are unalignable can sometimes be found by secondary structure. It has been shown that α -helices are more stable, robust to mutations and designable than β -strands in natural proteins, thus designing functional all- α proteins is likely to be easier than designing proteins with both helices and strands; this has been recently confirmed experimentally.

PROTEIN TERTIARY STRUCTURE

Protein tertiary structure is the three dimensional shape of a protein. The tertiary structure will have a single polypeptide chain “backbone” with one or more protein secondary structures, the protein domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular protein determine its tertiary structure. The protein tertiary structure is defined by its atomic coordinates. These coordinates may refer either to a protein domain or to the entire tertiary structure. A number of tertiary structures may fold into a quaternary structure.

History

The science of the tertiary structure of proteins has progressed from one of hypothesis to one of detailed definition. Although Emil Fischer had suggested proteins were made of polypeptide chains and amino acid side chains, it was Dorothy Maud Wrinch who incorporated geometry into the prediction of protein structures. Wrinch demonstrated this with the *Cyclol* model, the first prediction of the structure of a globular protein. Contemporary methods are able to determine, without prediction, tertiary structures to within 5 Å (0.5 nm) for small proteins (<120 residues) and, under favorable conditions, confident secondary structure predictions.

Determinants

Globular proteins have a core of hydrophobic amino acid residues and a surface region of water-exposed, charged, hydrophilic residues. This arrangement may stabilise interactions within the tertiary structure. For example, in secreted proteins, which are not bathed in cytoplasm, disulfide bonds between cysteine residues help to maintain the tertiary structure. There is a commonality of stable tertiary structures seen in proteins of diverse function and diverse evolution. For example, the TIM barrel, named for the enzyme triosephosphateisomerase, is a common tertiary structure as is the highly stable, dimeric, coiled coil structure. Hence, proteins may be classified by the structures they hold. Databases of proteins which use such a classification include *SCOP* and *CATH*.

Stability of Native States

The native state or native conformation of a protein is its most typical conformation in a cellular environment.

Chaperone Proteins

It is commonly assumed that the native state of a protein is also the most thermodynamically stable and that a protein will reach its native state, given its chemical kinetics, before it is translated. Protein chaperones within the cytoplasm of a cell assist a newly synthesised polypeptide to attain its native state. Some chaperone proteins

are highly specific in their function, for example, protein disulfide isomerase; others are general in their function and may assist most globular proteins, for example, the prokaryotic GroEL/GroES system of proteins and the homologous eukaryotic heat shock proteins (the Hsp60/Hsp10 system).

Kinetic Traps

Folding kinetics may trap a protein in a high-energy conformation. The high-energy conformation may contribute to the function of the protein. For example, the influenza hemagglutinin protein is a single polypeptide chain which when activated, is proteolytically cleaved to form two polypeptide chains. The two chains are held in a high-energy conformation. When the local pH drops, the protein undergoes an energetically favorable conformational rearrangement that enables it to penetrate the host cell membrane.

Metastability

Some tertiary protein structures may exist in long-lived states that are not the expected most stable state. For example, many serpinous (serine protease inhibitors) show this metastability. They undergo a conformational change when a loop of the protein is cut by a protease.

Cytoplasmic Environment

Prediction of protein tertiary structure relies on knowing the protein's primary structure and comparing the possible predicted tertiary structure with known tertiary structures in protein data banks. This only takes into account the cytoplasmic environment present at the time of protein synthesis to the extent that a similar cytoplasmic environment may also have influenced the structure of the proteins recorded in the protein data bank.

Ligand Binding

The structure of a protein, for example an enzyme, may change upon binding of its natural ligands, for example a cofactor. In this case, the structure of the protein bound to the ligand is known as holo structure, of the unbound protein as apo structure.

Determination

The knowledge of the tertiary structure of soluble globular proteins is more advanced than that of membrane proteins because the former are easier to study with available technology.

X-ray Crystallography

X-ray crystallography is the most common tool used to determine protein structure. It provides high resolution of the structure but it does not give information about protein's conformational flexibility.

NMR

Protein NMR gives comparatively lower resolution of protein structure. It is limited to smaller proteins. However, it can provide information about conformational changes of a protein in solution.

Dual Polarisation Interferometry

Dual polarisation interferometry provides complimentary information about surface captured proteins. It assists in determining structure and conformation changes over time.

Projects

Prediction Algorithm

The Folding@home project at Stanford University is a distributed computing research effort which uses approximately 5 petaFLOPS ($H=10 \times 86$ petaFLOPS) of available computing. It aims to find an algorithm which will consistently predict protein tertiary and quaternary structures given the protein's amino acid sequence and its cellular conditions.

Protein Aggregation Diseases

Protein aggregation diseases such as Alzheimer's disease and Huntington's disease and prion diseases such as bovine spongiform encephalopathy can be better understood by constructing (and reconstructing) disease models. This is done by causing the disease in laboratory animals, for example, by administering a toxin, such as MPTP to cause Parkinson's disease, or through genetic manipulation. Protein structure prediction is a new way to create disease models, which may avoid the use of animals.

Protein Tertiary Structure Retrieval Project (CoMOGrad)

Matching patterns in tertiary structure of a given protein to huge number of known protein tertiary structures and retrieve most similar ones in ranked order is in the heart of many research areas like function prediction of novel proteins, study of evolution, disease diagnosis, drug discovery, antibody design etc., The CoMOGrad project at BUET is a research effort to devise an extremely fast and much precise method for protein tertiary structure retrieval and develop online tool based on research outcome.

PROTEIN QUATERNARY STRUCTURE

Protein quaternary structure is the number and arrangement of multiple folded protein subunits in a multi-subunit complex. It includes organisations from simple dimers to large homooligomers and complexes with defined or variable numbers of subunits. It can also refer to biomolecular complexes of proteins with nucleic acids and other cofactors.

Description and Examples

Many proteins are actually assemblies of multiple polypeptide chains. The quaternary structure refers to the number and arrangement of the protein subunits with respect to one another. Examples of proteins with quaternary structure include hemoglobin, DNA polymerase, and ion channels. Enzymes composed of subunits with diverse functions are sometimes called holoenzymes, in which some parts may be known as regulatory subunits and the functional core is known as the catalytic subunit. Other assemblies referred to instead as multiprotein complexes also possess quaternary structure. Examples include nucleosomes and microtubules. Changes in quaternary structure can occur through conformational changes within individual subunits or through reorientation of the subunits relative to each other. It is through such changes, which underlie cooperativity and allostery in "multimeric" enzymes, that many proteins undergo regulation and perform their physiological function. The above definition follows a classical approach to biochemistry, established at times when the distinction between a protein and a functional, proteinaceous unit was difficult to elucidate. More recently, people refer to protein-protein interaction when discussing quaternary structure of proteins and consider all assemblies of proteins as protein complexes.

Nomenclature

The number of subunits in an oligomeric complex is described using names that end in-mer (Greek for “part, subunit”). Formal and Greco-Latinate names are generally used for the first ten types and can be used for up to twenty subunits, whereas higher order complexes are usually described by the number of subunits, followed by-meric.

1 = monomer/subunit	2 = dimer	3 = trimer
4 = tetramer	5 = pentamer	6 = hexamer
7 = heptamer	8 = octamer	9 = nonamer
10 = decamer	11 = undecamer	12 = dodecamer
13 = tridecamer	14 = tetradecamer	15 = pentadecamer*
16 = hexadecamer	17 = heptadecamer*	18 = octadecamer
19 = nonadecamer	20 = eicosamer	21-mer
22-mer	23-mer* <i>etc.</i> ,	

*No known examples.

Although complexes higher than octamers are rarely observed for most proteins, there are some important exceptions. Viral capsids are often composed of multiples of 60 proteins. Several molecular machines are also found in the cell, such as the proteasome (four heptameric rings = 28 subunits), the transcription complex and the spliceosome. The ribosome is probably the largest molecular machine, and is composed of many RNA and protein molecules. In some cases, proteins form complexes that then assemble into even larger complexes. In such cases, one uses the nomenclature, *e.g.*, “dimer of dimers” or “trimer of dimers”, to suggest that the complex might dissociate into smaller sub-complexes before dissociating into monomers.

Determination

Protein quaternary structure can be determined using a variety of experimental techniques that require a sample of protein in a variety of experimental conditions. The experiments often provide an estimate of the mass of the native protein and, together with knowledge of the masses and/or stoichiometry of the subunits, allow the quaternary structure to be predicted with a given accuracy. It is not always possible to obtain a precise determination of the subunit composition for a variety of reasons. The number of subunits in a protein complex can often be determined by measuring the hydrodynamic molecular volume or mass of the intact complex, which requires native solution conditions.

For *folded* proteins, the mass can be inferred from its volume using the partial specific volume of 0.73 ml/g. However, volume measurements are less certain than mass measurements, since, *unfolded* proteins appear to have a much larger volume than folded proteins; additional experiments are required to determine whether a protein is unfolded or has formed an oligomer.

Prediction

Some bioinformatics methods were developed for predicting the quaternary structural attributes of proteins based on their sequence information by using various modes of pseudo amino acid composition.

Direct Mass Measurement of Intact Complexes

- Sedimentation-equilibrium analytical ultracentrifugation
- Electrospray mass spectrometry
- Mass Spectrometric Immunoassay MSIA

Direct Size Measurement of Intact Complexes

- Static light scattering
- Size exclusion chromatography (requires calibration)
- Dual polarisation interferometry.

Indirect Size Measurement of Intact Complexes

- Sedimentation-velocity analytical ultracentrifugation (measures the translational diffusion constant)
- Dynamic light scattering (measures the translational diffusion constant)
- Pulsed-gradient protein nuclear magnetic resonance (measures the translational diffusion constant)
- Fluorescence polarization (measures the rotational diffusion constant)
- Dielectric relaxation (measures the rotational diffusion constant)
- Dual polarisation interferometry (measures the size and the density of the complex)

Methods that measure the mass or volume under unfolding conditions (such as MALDI-TOF mass spectrometry and SDS-PAGE) are generally not useful, since, non-native conditions usually cause the complex to dissociate into monomers. However, these may sometimes be applicable; for example, the experimenter may apply SDS-PAGE after first treating the intact complex with chemical cross-link reagents.

Protein–protein Interactions

Proteins are capable of forming very tight complexes. For example, ribonuclease inhibitor binds to ribonuclease A with a roughly 20 fM dissociation constant. Other proteins have evolved to bind specifically to unusual moieties on another protein, *e.g.*, biotin groups (avidin), phosphorylated tyrosines (SH2 domains) or proline-rich segments (SH3 domains).

PROTEIN STRUCTURE AND STABILITY

Protein stability is depended upon a few factors such as 1) Non-covalent electrostatic interactions 2) Hydrophobic interactions These interaction energies are from the order of 20-40 kJ/mol. Proteins are very sensitive to changing temperatures and a change in temperature may result in unfolding or denaturation. Protein denaturation may result in loss of function, and loss of native state. X-ray Crystallography, and calorimetry indicates that there is no general mechanism that describes the effect of temperature change on the functions and structure of proteins. This is due to the fact that proteins do not represent a uniform class of chemical entities from an energetic point of view. The structure and stability of an individual protein depends on the ratio of its polar and non-polar residues. They contribute to the conformational and the net enthalpies of local and non-local interactions. Taking the weak intermolecular interactions responsible for structural integrity into consideration, it is hard to predict the effects of temperature because there are too many unknown factors contributing to the hypothetical free energy balance and its temperature dependence. Internal salt linkages produce thermal stability, and whether cold temperature results in the destabilization of these linkages is unknown. In principle, the free energy of stabilization of soluble globular proteins does not exceed 50-100 kJ/mol. The stabilization is based on the equivalent of few hydrogen bonds, ion pairs, or hydrophobic interactions, even though numerous intramolecular interactions results in stabilization. Taking into consideration the large number of hydrogen bonds that take place for the stabilization of secondary structures, and the stabilization of the inner core through hydrophobic interactions, the free energy of stabilization emerges as small difference between large numbers. Therefore, the structure of a native protein is not optimized for the maximum stability.

DOMAINS, MOTIFS, AND FOLDS IN PROTEIN STRUCTURE

Proteins are frequently described as consisting of several structural units. These units include domains, motifs, and folds. Despite the fact that there are about 100,000 different proteins expressed in eukaryotic systems, there are many fewer different domains, structural motifs and folds.

STRUCTURAL DOMAIN

A structural domain is an element of the protein's overall structure that is self-stabilizing and often folds independently of the rest of the protein chain. Many domains are not unique to the protein products of one gene or one gene family but instead appear in a variety of proteins. Domains often are named and singled out because they figure prominently in the biological function of the protein they belong to; for example, the "calcium-binding domain of calmodulin". Because they are independently stable, domains can be "swapped" by genetic engineering between one protein and another to make chimera proteins.

STRUCTURAL AND SEQUENCE MOTIF

The structural and sequence motifs refer to short segments of protein three-dimensional structure or amino acid sequence that were found in a large number of different proteins.

SUPERSECONDARY STRUCTURE

The supersecondary structure refers to a specific combination of secondary structure elements, such as β - α - β units or a helix-turn-helix motif. Some of them may be also referred to as structural motifs.

PROTEIN FOLD

A protein fold refers to the general protein architecture, like a helix bundle, β -barrel, Rossman fold or different "folds" provided in the Structural Classification of Proteins database. A related concept is protein topology that refers to the arrangement of contacts within the protein.

SUPERDOMAIN

A superdomain consists of two or more nominally unrelated structural domains that are inherited as a single unit and occur in different proteins. An example is provided by the protein tyrosine phosphatase domain and C2 domain pair in PTEN, several tensin proteins, auxilin and proteins in plants and fungi. The PTP-C2 superdomain evidently came into existence prior to the divergence of fungi, plants and animals is therefore likely to be about 1.5 billion years old.

PROTEIN FOLDING

Protein folding is the physical process by which a protein chain acquires its native 3-dimensional structure, a conformation that is usually biologically functional, in an expeditious and reproducible manner. It is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil. Each protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of amino acids. This polypeptide lacks any stable (long-lasting) three-dimensional structure (the left hand side of the first figure). As the polypeptide chain is being synthesized by the ribosome, the linear chain begins to fold into its three dimensional structure. Folding begins to occur even during translation of the polypeptide chain. Amino acids interact with each other to produce a well-defined three-dimensional

structure, the folded protein (the right hand side of the figure), known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence or primary structure (Anfinsen's dogma). The energy landscape describes the folding pathways in which the unfolded protein is able to assume its native state.

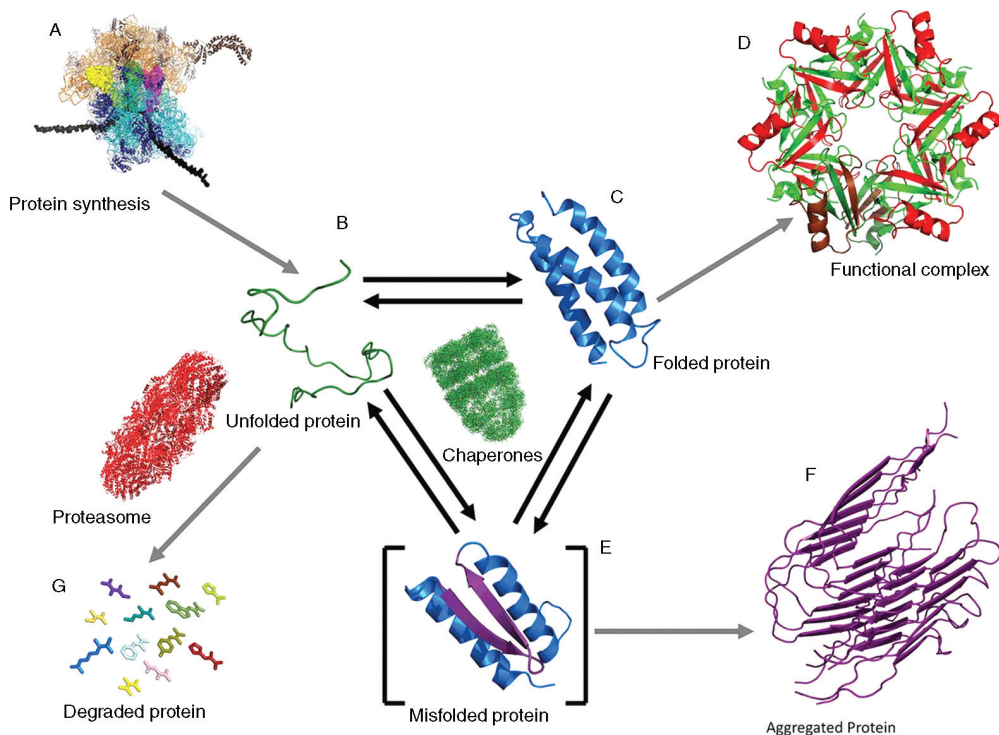


Fig. Protein folding.

The correct three-dimensional structure is essential to function, although some parts of functional proteins may remain unfolded, so that protein dynamics is important. Failure to fold into native structure generally produces inactive proteins, but in some instances misfolded proteins have modified or toxic functionality. Several neurodegenerative and other diseases are believed to result from the accumulation of amyloid fibrils formed by misfolded proteins. Many allergies are caused by incorrect folding of some proteins, because the immune system does not produce antibodies for certain protein structures. Denaturation of proteins is a process of transition from the folded to the unfolded state.

PROCESS OF PROTEIN FOLDING

Primary Structure

The primary structure of a protein, its linear amino-acid sequence, determines its native conformation. The specific amino acid residues and their position in the polypeptide chain are the determining factors for which portions of the protein fold closely together and form its three dimensional conformation. The amino acid composition is not as important as the sequence. The essential fact of folding, however, remains that the amino acid sequence of each protein contains the information that specifies both the native structure and the pathway to attain that state. This is not to say that nearly identical amino acid sequences always fold similarly. Conformations differ based on environmental factors as well; similar proteins fold differently based on where they are found.

Formation of a secondary structure is the first step in the folding process that a protein takes to assume its native structure. Characteristic of secondary structure are the structures known as alpha helices and beta sheets that fold rapidly because they are stabilized by intramolecular hydrogen bonds, as was first characterized by Linus

Pauling. Formation of intramolecular hydrogen bonds provides another important contribution to protein stability. α -helices are formed by hydrogen bonding of the backbone to form a spiral shape (refer to figure on the right).

The β pleated sheet is a structure that forms with the backbone bending over itself to form the hydrogen bonds (as displayed in the figure to the left). The hydrogen bonds are between the amide hydrogen and carbonyl oxygen of the peptide bond. There exists anti-parallel β pleated sheets and parallel β pleated sheets where the stability of the hydrogen bonds is stronger in the anti-parallel β sheet as it hydrogen bonds with the ideal 180 degree angle compared to the slanted hydrogen bonds formed by parallel sheets.

Tertiary Structure

The alpha helices and beta pleated sheets can be amphipathic in nature, or contain a hydrophilic portion and a hydrophobic portion. This property of secondary structures aids in the tertiary structure of a protein in which the folding occurs so that the hydrophilic sides are facing the aqueous environment surrounding the protein and the hydrophobic sides are facing the hydrophobic core of the protein. Secondary structure hierarchically gives way to tertiary structure formation. Once the protein's tertiary structure is formed and stabilized by the hydrophobic interactions, there may also be covalent bonding in the form of disulfide bridges formed between two cysteine residues. Tertiary structure of a protein involves a single polypeptide chain; however, additional interactions of folded polypeptide chains give rise to quaternary structure formation.

Quaternary Structure

Tertiary structure may give way to the formation of quaternary structure in some proteins, which usually involves the “assembly” or “coassembly” of subunits that have already folded; in other words, multiple polypeptide chains could interact to form a fully functional quaternary protein.

Driving Forces of Protein Folding

Folding is a spontaneous process that is mainly guided by hydrophobic interactions, formation of intramolecular hydrogen bonds, van der Waals forces, and it is opposed by conformational entropy. The process of folding often begins co-translationally, so that the N-terminus of the protein begins to fold while the C-terminal portion of the protein is still being synthesized by the ribosome; however, a protein molecule may fold spontaneously during or after biosynthesis. While these macromolecules may be regarded as “folding themselves”, the process also depends on the solvent (water or lipid bilayer), the concentration of salts, the pH, the temperature, the possible presence of cofactors and of molecular chaperones. Proteins will have limitations on their folding abilities by the restricted bending angles or conformations that are possible. These allowable angles of protein folding are described with a two-dimensional plot known as the Ramachandran plot, depicted with psi and phi angles of allowable rotation.

Hydrophobic Effect

Protein folding must be thermodynamically favorable within a cell in order for it to be a spontaneous reaction. Since, it is known that protein folding is a spontaneous reaction, then it must assume a negative Gibbs free energy value. Gibbs free energy in protein folding is directly related to enthalpy and entropy. For a negative ΔG to arise and for protein folding to become thermodynamically favorable, then either enthalpy, entropy, or both terms must be favorable. Minimizing the number of hydrophobic side-chains exposed to water is an important driving force behind the folding process. The hydrophobic effect is the phenomenon in which the hydrophobic chains of a protein collapse into the core of the protein (away from the hydrophilic environment). In an aqueous environment, the water molecules tend to aggregate around the hydrophobic regions or side

chains of the protein, creating water shells of ordered water molecules. An ordering of water molecules around a hydrophobic region increases order in a system and therefore contributes a negative change in entropy (less entropy in the system).

The water molecules are fixed in these water cages which drives the hydrophobic collapse, or the inward folding of the hydrophobic groups. The hydrophobic collapse introduces entropy back to the system via the breaking of the water cages which frees the ordered water molecules. The multitude of hydrophobic groups interacting within the core of the globular folded protein contributes a significant amount to protein stability after folding, because of the vastly accumulated van der Waals forces (specifically London Dispersion forces). The hydrophobic effect exists as a driving force in thermodynamics only if there is the presence of an aqueous medium with an amphiphilic molecule containing a large hydrophobic region. The strength of hydrogen bonds depends on their environment; thus, H-bonds enveloped in a hydrophobic core contribute more than H-bonds exposed to the aqueous environment to the stability of the native state.

Chaperones

Molecular chaperones are a class of proteins that aid in the correct folding of other proteins *in vivo*. Chaperones exist in all cellular compartments and interact with the polypeptide chain in order to allow the native three-dimensional conformation of the protein to form; however, chaperones themselves are not included in the final structure of the protein they are assisting in.

Chaperones may assist in folding even when the nascent polypeptide is being synthesized by the ribosome. Molecular chaperones operate by binding to stabilize an otherwise unstable structure of a protein in its folding pathway, but chaperones do not contain the necessary information to know the correct native structure of the protein they are aiding; rather, chaperones work by preventing incorrect folding conformations. In this way, chaperones do not actually increase the rate of individual steps involved in the folding pathway towards the native structure; instead, they work by reducing possible unwanted aggregations of the polypeptide chain that might otherwise slow down the search for the proper intermediate and they provide a more efficient pathway for the polypeptide chain to assume the correct conformations.

Chaperones are not to be confused with folding catalysts, which actually do catalyze the otherwise slow steps in the folding pathway. Examples of folding catalysts are protein disulfide isomerases and peptidyl-prolyl isomerases that may be involved in formation of disulfide bonds or interconversion between *cis* and *trans* stereoisomers, respectively. Chaperones are shown to be critical in the process of protein folding *in vivo* because they provide the protein with the aid needed to assume its proper alignments and conformations efficiently enough to become “biologically relevant”. This means that the polypeptide chain could theoretically fold into its native structure without the aid of chaperones, as demonstrated by protein folding experiments conducted *in vitro*; however, this process proves to be too inefficient or too slow to exist in biological systems; therefore, chaperones are necessary for protein folding *in vivo*. Along with its role in aiding native structure formation, chaperones are shown to be involved in various roles such as protein transport, degradation, and even allow denatured proteins exposed to certain external denaturant factors an opportunity to refold into their correct native structures.

A fully denatured protein lacks both tertiary and secondary structure, and exists as a so-called random coil. Under certain conditions some proteins can refold; however, in many cases, denaturation is irreversible. Cells sometimes protect their proteins against the denaturing influence of heat with enzymes known as heat shock proteins (a type of chaperone), which assist other proteins both in folding and in remaining folded. Some proteins never fold in cells at all except with the assistance of chaperones which either isolate individual proteins so that their folding is not interrupted by interactions with other proteins or help to unfold misfolded proteins, allowing them to refold into the correct native structure. This function is crucial to prevent the risk of

precipitation into insoluble amorphous aggregates. The external factors involved in protein denaturation or disruption of the native state include temperature, external fields (electric, magnetic), molecular crowding, and even the limitation of space, which can have a big influence on the folding of proteins. High concentrations of solutes, extremes of pH, mechanical forces, and the presence of chemical denaturants can contribute to protein denaturation, as well. These individual factors are categorized together as stresses. Chaperones are shown to exist in increasing concentrations during times of cellular stress and help the proper folding of emerging proteins as well as denatured or misfolded ones.

Under some conditions proteins will not fold into their biochemically functional forms. Temperatures above or below the range that cells tend to live in will cause thermally unstable proteins to unfold or denature (this is why boiling makes an egg white turn opaque). Protein thermal stability is far from constant, however; for example, hyperthermophilic bacteria have been found that grow at temperatures as high as 122 °C, which of course requires that their full complement of vital proteins and protein assemblies be stable at that temperature or above.

COMPUTATIONAL METHODS FOR STUDYING PROTEIN FOLDING

The study of protein folding includes three main aspects related to the prediction of protein stability, kinetics, and structure. A recent review summarizes the available computational methods for protein folding.

Energy Landscape of Protein Folding

The protein folding phenomenon was largely an experimental endeavor until the formulation of an energy landscape theory of proteins by Joseph Bryngelson and Peter Wolynes in the late 1980s and early 1990s. This approach introduced the *principle of minimal frustration*. This principle says that nature has chosen amino acid sequences so that the folded state of the protein is very stable. In addition, the undesired interactions between amino acids along the folding pathway are reduced, making the acquisition of the folded state a very fast process. Even though nature has reduced the level of *frustration* in proteins, some degree of it remains up to now as can be observed in the presence of local minima in the energy landscape of proteins. A consequence of these evolutionarily selected sequences is that proteins are generally thought to have globally “funneled energy landscapes” (coined by José Onuchic) that are largely directed towards the native state. This “folding funnel” landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather than being restricted to a single mechanism.

The theory is supported by both computational simulations of model proteins and experimental studies, and it has been used to improve methods for protein structure prediction and design. The description of protein folding by the leveling free-energy landscape is also consistent with the 2nd law of thermodynamics. Physically, thinking of landscapes in terms of visualizable potential or total energy surfaces simply with maxima, saddle points, minima, and funnels, rather like geographic landscapes, is perhaps a little misleading. The relevant description is really a high-dimensional phase space in which manifolds might take a variety of more complicated topological forms.

The unfolded polypeptide chain begins at the top of the funnel where it may assume the largest number of unfolded variations and is in its highest energy state. Energy landscapes such as these indicate that there are a large number of initial possibilities, but only a single native state is possible; however, it does not reveal the numerous folding pathways that are possible. A different molecule of the same exact protein may be able to follow marginally different folding pathways, seeking different lower energy intermediates, as long as the same

native structure is reached. Different pathways may have different frequencies of utilization depending on the thermodynamic favorability of each pathway. This means that if one pathway is found to be more thermodynamically favorable than another, it is likely to be used more frequently in the pursuit of the native structure. As the protein begins to fold and assume its various conformations, it always seeks a more thermodynamically favorable structure than before and thus continues through the energy funnel. Formation of secondary structures is a strong indication of increased stability within the protein, and only one combination of secondary structures assumed by the polypeptide backbone will have the lowest energy and therefore be present in the native state of the protein. Among the first structures to form once the polypeptide begins to fold are alpha helices and beta turns, where alpha helices can form in as little as 100 nanoseconds and beta turns in 1 microsecond.

There exists a saddle point in the energy funnel landscape where the transition state for a particular protein is found. The transition state in the energy funnel diagram is the conformation that must be assumed by every molecule of that protein if the protein wishes to finally assume the native structure. No protein may assume the native structure without first passing through the transition state. The transition state can be referred to as a variant or premature form of the native state rather than just another intermediary step. The folding of the transition state is shown to be rate-determining, and even though it exists in a higher energy state than the native fold, it greatly resembles the native structure. Within the transition state, there exists a nucleus around which the protein is able to fold, formed by a process referred to as “nucleation condensation” where the structure begins to collapse onto the nucleus.

Two models of protein folding are currently being confirmed:

- The diffusion collision model, in which first a nucleus forms, then the secondary structure, and finally these secondary structures collide and pack tightly together.
- The nucleation-condensation model, in which the secondary and tertiary structures of the protein are made at the same time.

Recent studies have shown that some proteins show characteristics of both of these folding models. For the most part, scientists have been able to study many identical molecules folding together *en masse*. At the coarsest level, it appears that in transitioning to the native state, a given amino acid sequence takes roughly the same route and proceeds through roughly the same intermediates and transition states. Often folding involves first the establishment of regular secondary and supersecondary structures, in particular alpha helices and beta sheets, and afterwards tertiary structure.

Modeling of Protein Folding

De novo or *ab initio* techniques for computational protein structure prediction are related to, but strictly distinct from, experimental studies of protein folding. Molecular Dynamics (MD) is an important tool for studying protein folding and dynamics in silico. First equilibrium folding simulations were done using implicit solvent model and umbrella sampling. Because of computational cost, *ab initio* MD folding simulations with explicit water are limited to peptides and very small proteins. MD simulations of larger proteins remain restricted to dynamics of the experimental structure or its high-temperature unfolding. Long-time folding processes (beyond about 1 millisecond), like folding of small-size proteins (about 50 residues) or larger, can be accessed using coarse-grained models. The 100-petaFLOP distributed computing project Folding@home created by Vijay Pande's group at Stanford University simulates protein folding using the idle processing time of CPUs and

GPUs of personal computers from volunteers. The project aims to understand protein misfolding and accelerate drug design for disease research. Long continuous-trajectory simulations have been performed on Anton, a massively parallel supercomputer designed and built around custom ASICs and interconnects by D. E. Shaw Research. The longest published result of a simulation performed using Anton is a 2.936 millisecond simulation of NTL9 at 355 K.

EXPERIMENTAL TECHNIQUES FOR STUDYING PROTEIN FOLDING

While inferences about protein folding can be made through mutation studies, typically, experimental techniques for studying protein folding rely on the gradual unfolding or folding of proteins and observing conformational changes using standard non-crystallographic techniques.

X-ray Crystallography

X-ray crystallography is one of the more efficient and important methods for attempting to decipher the three dimensional configuration of a folded protein. To be able to conduct X-ray crystallography, the protein under investigation must be located inside a crystal lattice. To place a protein inside a crystal lattice, one must have a suitable solvent for crystallization, obtain a pure protein at supersaturated levels in solution, and precipitate the crystals in solution. Once a protein is crystallized, x-ray beams can be concentrated through the crystal lattice which would diffract the beams or shoot them outwards in various directions. These exiting beams are correlated to the specific three-dimensional configuration of the protein enclosed within. The x-rays specifically interact with the electron clouds surrounding the individual atoms within the protein crystal lattice and produce a discernible diffraction pattern. Only by relating the electron density clouds with the amplitude of the x-rays can this pattern be read and lead to assumptions of the phases or phase angles involved that complicate this method. Without the relation established through a mathematical basis known as Fourier transform, the “phase problem” would render predicting the diffraction patterns very difficult. Emerging methods like multiple isomorphous replacement use the presence of a heavy metal ion to diffract the x-rays into a more predictable manner, reducing the number of variables involved and resolving the phase problem.

Fluorescence Spectroscopy

Fluorescence spectroscopy is a highly sensitive method for studying the folding state of proteins. Three amino acids, phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp), have intrinsic fluorescence properties, but only Tyr and Trp are used experimentally because their quantum yields are high enough to give good fluorescence signals. Both Trp and Tyr are excited by a wavelength of 280 nm, whereas only Trp is excited by a wavelength of 295 nm. Because of their aromatic character, Trp and Tyr residues are often found fully or partially buried in the hydrophobic core of proteins, at the interface between two protein domains, or at the interface between subunits of oligomeric proteins. In this apolar environment, they have high quantum yields and therefore high fluorescence intensities. Upon disruption of the protein's tertiary or quaternary structure, these side chains become more exposed to the hydrophilic environment of the solvent, and their quantum yields decrease, leading to low fluorescence intensities. For Trp residues, the wavelength of their maximal fluorescence emission also depend on their environment.

Fluorescence spectroscopy can be used to characterize the equilibrium unfolding of proteins by measuring the variation in the intensity of fluorescence emission or in the wavelength of maximal emission as functions of a denaturant value. The denaturant can be a chemical molecule (urea, guanidinium hydrochloride), temperature,

pH, pressure, *etc.*, The equilibrium between the different but discrete protein states, *i.e.*, native state, intermediate states, unfolded state, depends on the denaturant value; therefore, the global fluorescence signal of their equilibrium mixture also depends on this value. One thus obtains a profile relating the global protein signal to the denaturant value. The profile of equilibrium unfolding may enable one to detect and identify intermediates of unfolding. General equations have been developed by Hugues Bedouelle to obtain the thermodynamic parameters that characterize the unfolding equilibria for homomeric or heteromeric proteins, up to trimers and potentially tetramers, from such profiles. Fluorescence spectroscopy can be combined with fast-mixing devices such as stopped flow, to measure protein folding kinetics, generate a chevron plot and derive a Phi value analysis.

Circular Dichroism

Circular dichroism is one of the most general and basic tools to study protein folding. Circular dichroism spectroscopy measures the absorption of circularly polarized light. In proteins, structures such as alpha helices and beta sheets are chiral, and thus absorb such light. The absorption of this light acts as a marker of the degree of foldedness of the protein ensemble. This technique has been used to measure equilibrium unfolding of the protein by measuring the change in this absorption as a function of denaturant concentration or temperature. A denaturant melt measures the free energy of unfolding as well as the protein's *m* value, or denaturant dependence. A temperature melt measures the melting temperature (T_m) of the protein. As for fluorescence spectroscopy, circular-dichroism spectroscopy can be combined with fast-mixing devices such as stopped flow to measure protein folding kinetics and to generate chevron plots.

Vibrational Circular Dichroism of Proteins

The more recent developments of vibrational circular dichroism (VCD) techniques for proteins, currently involving Fourier transform (FFT) instruments, provide powerful means for determining protein conformations in solution even for very large protein molecules. Such VCD studies of proteins are often combined with X-ray diffraction of protein crystals, FT-IR data for protein solutions in heavy water (D_2O), or *ab initio* quantum computations to provide unambiguous structural assignments that are unobtainable from CD.

Protein Nuclear Magnetic Resonance Spectroscopy

Protein folding is routinely studied using NMR spectroscopy, for example by monitoring hydrogen-deuterium exchange of backbone amide protons of proteins in their native state, which provides both the residue-specific stability and overall stability of proteins.

Dual Polarisation Interferometry

Dual polarisation interferometry is a surface-based technique for measuring the optical properties of molecular layers. When used to characterize protein folding, it measures the conformation by determining the overall size of a monolayer of the protein and its density in real time at sub-Angstrom resolution, although real-time measurement of the kinetics of protein folding are limited to processes that occur slower than ~10 Hz. Similar to circular dichroism, the stimulus for folding can be a denaturant or temperature.

Studies of Folding with High Time Resolution

The study of protein folding has been greatly advanced in recent years by the development of fast, time-resolved techniques. Experimenters rapidly trigger the folding of a sample of unfolded protein and observe the

resulting dynamics. Fast techniques in use include neutron scattering, ultrafast mixing of solutions, photochemical methods, and laser temperature jump spectroscopy. Among the many scientists who have contributed to the development of these techniques are Jeremy Cook, Heinrich Roder, Harry Gray, Martin Gruebele, Brian Dyer, William Eaton, Sheena Radford, Chris Dobson, Alan Fersht, Bengt Nölting and Lars Konermann.

Proteolysis

Proteolysis is routinely used to probe the fraction unfolded under a wide range of solution conditions (*e.g.*, Fast parallel proteolysis (FASTpp)).

Optical Tweezers

Single molecule techniques such as optical tweezers and AFM have been used to understand protein folding mechanisms of isolated proteins as well as proteins with chaperones. Optical tweezers have been used to stretch single protein molecules from their C- and N-termini and unfold them to allow study of the subsequent refolding. The technique allows one to measure folding rates at single-molecule level; for example, optical tweezers have been recently applied to study folding and unfolding of proteins involved in blood coagulation. von Willebrand factor (vWF) is a protein with an essential role in blood clot formation process. It discovered—using single molecule optical tweezers measurement—that calcium-bound vWF acts as a shear force sensor in the blood. Shear force leads to unfolding of the A2 domain of vWF, whose refolding rate is dramatically enhanced in the presence of calcium. Recently, it was also shown that the simple src SH3 domain accesses multiple unfolding pathways under force.

OTHER INFORMATION

Incorrect Protein Folding and Neurodegenerative Disease

A protein is considered to be misfolded if it cannot achieve its normal native state. This can be due to mutations in the amino acid sequence or a disruption of the normal folding process by external factors. The misfolded protein typically contains β -sheets that are organized in a supramolecular arrangement known as a cross- β structure. These β -sheet-rich assemblies are very stable, very insoluble, and generally resistant to proteolysis. The structural stability of these fibrillar assemblies is caused by extensive interactions between the protein monomers, formed by backbone hydrogen bonds between their β -strands. The misfolding of proteins can trigger the further misfolding and accumulation of other proteins into aggregates or oligomers. The increased levels of aggregated proteins in the cell leads to formation of amyloid-like structures which can cause degenerative disorders and cell death. The amyloids are fibrillary structure that contain intermolecular hydrogen bonds, which are highly insoluble, and made from converted protein aggregates. Therefore, the proteasome pathway may not be efficient enough to degrade the misfolded proteins prior to aggregation. Misfolded proteins can interact with one another and form structured aggregates and gain toxicity through intermolecular interactions.

Aggregated proteins are associated with prion-related illnesses such as Creutzfeldt–Jakob disease, bovine spongiform encephalopathy (mad cow disease), amyloid-related illnesses such as Alzheimer's disease and familial amyloid cardiomyopathy or polyneuropathy, as well as intracellular aggregation diseases such as Huntington's and Parkinson's disease. These age onset degenerative diseases are associated with the aggregation of misfolded proteins into insoluble, extracellular aggregates and/or intracellular inclusions including cross- β amyloid fibrils. It is not completely clear whether the aggregates are the cause or merely a reflection of the loss of protein homeostasis, the balance between synthesis, folding, aggregation and protein turnover. Recently the European Medicines Agency

approved the use of Tafamidis or Vyndaqel (a kinetic stabilizer of tetrameric transthyretin) for the treatment of transthyretin amyloid diseases. This suggests that the process of amyloid fibril formation (and not the fibrils themselves) causes the degeneration of post-mitotic tissue in human amyloid diseases. Misfolding and excessive degradation instead of folding and function leads to a number of proteopathy diseases such as antitrypsin-associated emphysema, cystic fibrosis and the lysosomal storage diseases, where loss of function is the origin of the disorder. While protein replacement therapy has historically been used to correct the latter disorders, an emerging approach is to use pharmaceutical chaperones to fold mutated proteins to render them functional.

Levinthal's Paradox and Kinetics

In 1969, Cyrus Levinthal noted that, because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an astronomical number of possible conformations. An estimate of 3^{300} or 10^{143} was made in one of his papers. Levinthal's paradox is a thought experiment based on the observation that if a protein were folded by sequentially sampling of all possible conformations, it would take an astronomical amount of time to do so, even if the conformations were sampled at a rapid rate (on the nanosecond or picosecond scale). Based upon the observation that proteins fold much faster than this, Levinthal then proposed that a random conformational search does not occur, and the protein must, therefore, fold through a series of meta-stable intermediate states. The duration of the folding process varies dramatically depending on the protein of interest. When studied outside the cell, the slowest folding proteins require many minutes or hours to fold primarily due to proline isomerization, and must pass through a number of intermediate states, like checkpoints, before the process is complete. On the other hand, very small single-domain proteins with lengths of up to a hundred amino acids typically fold in a single step. Time scales of milliseconds are the norm and the very fastest known protein folding reactions are complete within a few microseconds.

PROTEIN STRUCTURE DETERMINATION

Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows one to measure the three-dimensional (3-D) density distribution of electrons in the protein, in the crystallized state, and thereby infer the 3-D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by nuclear magnetic resonance (NMR) techniques. For larger protein complexes, cryo-electron microscopy can determine protein structures. The resolution is typically lower than that of X-ray crystallography, or NMR, but the maximum resolution is steadily increasing. This technique is still a particularly valuable for very large protein complexes such as virus coat proteins and amyloid fibers.

General secondary structure composition can be determined via circular dichroism. Vibrational spectroscopy can also be used to characterize the conformation of peptides, polypeptides, and proteins. Two-dimensional infrared spectroscopy has become a valuable method to investigate the structures of flexible peptides and proteins that cannot be studied with other methods. A more qualitative picture of protein structure is often obtained by proteolysis, which is also useful to screen for more crystallizable protein samples. Novel implementations of this approach, including fast parallel proteolysis (FASTpp), can probe the structured fraction and its stability without the need for purification. Once a protein's structure has been experimentally determined, further detailed studies can be done computationally, using molecular dynamic simulations of that structure.

PROTEIN SEQUENCE ANALYSIS: ENSEMBLES

Proteins are often thought of as relatively stable structures that have a set tertiary structure and experience conformational changes as a result of being modified by other proteins or as part of enzymatic activity. However

proteins have varying degrees of stability and some of the less stable variants are intrinsically disordered proteins. These proteins exist and function in a relatively ‘disordered’ state lacking a stable tertiary structure. As a result, they are difficult to describe in a standard protein structure model that was designed for proteins with a fixed tertiary structure.

Conformational ensembles have been devised as a way to provide a more accurate and ‘dynamic’ representation of the conformational state of intrinsically disordered proteins. Conformational ensembles function by attempting to represent the various conformations of intrinsically disordered proteins within an ensemble file (the type found at the Protein Ensemble Database). Protein ensemble files are a representation of a protein that can be considered to have a flexible structure. Creating these files requires determining which of the various theoretically possible protein conformations actually exist. One approach is to apply computational algorithms to the protein data in order to try to determine the most likely set of conformations for an ensemble file.

There are multiple methods for preparing data for the Protein Ensemble Database that fall into two general methodologies—pool and molecular dynamics (MD) approaches (diagrammed in the figure). The pool based approach uses the protein’s amino acid sequence to create a massive pool of random conformations. This pool is then subjected to more computational processing that creates a set of theoretical parameters for each conformation based on the structure. Conformational subsets from this pool whose average theoretical parameters closely match known experimental data for this protein are selected. The molecular dynamics approach takes multiple random conformations at a time and subjects all of them to experimental data. Here the experimental data is serving as limitations to be placed on the conformations (*e.g.*, known distances between atoms). Only conformations that manage to remain within the limits set by the experimental data are accepted. This approach often applies large amounts of experimental data to the conformations which is a very computationally demanding task.

Protein	Data Type	Protocol	PED ID	References
Sic1/Cdc4	NMR and SAXS	Pool-based	PED9AAA	
p15 PAF	NMR and SAXS	Pool-based	PED6AAA	
MKK7	NMR	Pool-based	PED5AAB	
Beta-synuclein	NMR	MD-based	PED1AAD	
P27 KID	NMR	MD-based	PED2AAA	

PROTEIN STRUCTURE DATABASES

Protein structure database is a database that is modeled around the various experimentally determined protein structures. The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way. Data included in protein structure databases often includes 3D coordinates as well as experimental information, such as unit cell dimensions and angles for x-ray crystallography determined structures.

Though most instances, in this case either proteins or a specific structure determinations of a protein, also contain sequence information and some databases even provide means for performing sequence based queries, the primary attribute of a structure database is structural information, whereas sequence databases focus on sequence information, and contain no structural information for the majority of entries. Protein structure databases are critical for many efforts in computational biology such as structure based drug design, both in developing the computational methods used and in providing a large experimental dataset used by some methods to provide insights about the function of a protein.

STRUCTURE CLASSIFICATION

Protein structures can be grouped based on their similarity or a common evolutionary origin. The Structural Classification of Proteins database and CATH database provide two different structural classifications of proteins. Shared structure between proteins is considered evidence of evolutionary relatedness between proteins and is used group proteins together into protein superfamilies.

COMPUTATIONAL PREDICTION OF PROTEIN STRUCTURE

The generation of a protein sequence is much easier than the determination of a protein structure. However, the structure of a protein gives much more insight in the function of the protein than its sequence. Therefore, a number of methods for the computational prediction of protein structure from its sequence have been developed. *Ab initio* prediction methods use just the sequence of the protein. Threading and homology modeling methods can build a 3-D model for a protein of unknown structure from experimental structures of evolutionarily-related proteins, called a protein family.

HIGH RESOLUTION

Cryogenic Electron Microscopy

Electron cryomicroscopy (CryoEM) is an electron microscopy (EM) technique where the sample is cooled to cryogenic temperatures. There are various types of electron microscopy, some of which has been performed at cryogenic temperatures:

Transmission Electron Cryomicroscopy

Transmission electron cryomicroscopy (CryoTEM) is a transmission electron microscopy technique is used in structural biology.

- Electron crystallography, method to determine the arrangement of atoms in solids using a TEM
- Electron cryotomography (CryoET), a specialized application of where samples are imaged as they are tilted

Scanning Electron Cryomicroscopy

Scanning electron cryomicroscopy (CryoSEM), is scanning electron microscopy technique with a scanning electron microscope's cold stage in a cryogenic chamber.

X-RAY CRYSTALLOGRAPHY

X-ray crystallography is a technique used for determining the atomic and molecular structure of a crystal, in which the crystalline atoms cause a beam of incident X-rays to diffract into many specific directions. By measuring the angles and intensities of these diffracted beams, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal. From this electron density, the mean positions of the atoms in the crystal can be determined, as well as their chemical bonds, their crystallographic disorder, and various other information. Since, many materials can form crystals—such as salts, metals, minerals, semiconductors, as well as various inorganic, organic, and biological molecules—X-ray crystallography has been fundamental in the development of many scientific fields. In its first decades of use, this method determined the size of atoms, the lengths and types of chemical bonds, and the atomic-scale differences among various materials, especially minerals and alloys.

The method also revealed the structure and function of many biological molecules, including vitamins, drugs, proteins and nucleic acids such as DNA. X-ray crystallography is still the chief method for characterizing the atomic structure of new materials and in discerning materials that appear similar by other experiments. X-ray crystal structures can also account for unusual electronic or elastic properties of a material, shed light on chemical interactions and processes, or serve as the basis for designing pharmaceuticals against diseases.

In a single-crystal X-ray diffraction measurement, a crystal is mounted on a goniometer. The goniometer is used to position the crystal at selected orientations. The crystal is illuminated with a finely focused monochromatic beam of X-rays, producing a diffraction pattern of regularly spaced spots known as *reflections*. The two-dimensional images taken at different orientations are converted into a three-dimensional model of the density of electrons within the crystal using the mathematical method of Fourier transforms, combined with chemical data known for the sample. Poor resolution (fuzziness) or even errors may result if the crystals are too small, or not uniform enough in their internal makeup.

X-ray crystallography is related to several other methods for determining atomic structures. Similar diffraction patterns can be produced by scattering electrons or neutrons, which are likewise interpreted by Fourier transformation. If single crystals of sufficient size cannot be obtained, various other X-ray methods can be applied to obtain less detailed information; such methods include fibre diffraction, powder diffraction and (if the sample is not crystallized) small-angle X-ray scattering (SAXS). If the material under investigation is only available in the form of nanocrystalline powders or suffers from poor crystallinity, the methods of electron crystallography can be applied for determining the atomic structure. For all above mentioned X-ray diffraction methods, the scattering is elastic; the scattered X-rays have the same wavelength as the incoming X-ray. By contrast, *inelastic* X-ray scattering methods are useful in studying excitations of the sample such as plasmons, crystal-field and orbital excitations, magnons, and phonons, rather than the distribution of its atoms.

History

Early Scientific History of Crystals and X-rays

Crystals, though long admired for their regularity and symmetry, were not investigated scientifically until the 17th century. Johannes Kepler hypothesized in his work *Strena seu de Nive Sexangula* (A New Year's Gift of Hexagonal Snow) (1611) that the hexagonal symmetry of snowflake crystals was due to a regular packing of spherical water particles. The Danish scientist Nicolas Steno (1669) pioneered experimental investigations of crystal symmetry. Steno showed that the angles between the faces are the same in every exemplar of a particular type of crystal, and René Just Haüy (1784) discovered that every face of a crystal can be described by simple stacking patterns of blocks of the same shape and size. Hence, William Hallows Miller in 1839 was able to give each face a unique label of three small integers, the Miller indices which remain in use today for identifying crystal faces. Haüy's study led to the correct idea that crystals are a regular three-dimensional array (a Bravais lattice) of atoms and molecules; a single unit cell is repeated indefinitely along three principal directions that are not necessarily perpendicular. In the 19th century, a complete catalog of the possible symmetries of a crystal was worked out by Johan Hessel, Auguste Bravais, Evgraf Fedorov, Arthur Schönflies and (belatedly) William Barlow (1894). From the available data and physical reasoning, Barlow proposed several crystal structures in the 1880s that were validated later by X-ray crystallography; however, the available data were too scarce in the 1880s to accept his models as conclusive.

Wilhelm Röntgen discovered X-rays in 1895, just as the studies of crystal symmetry were being concluded. Physicists were initially uncertain of the nature of X-rays, but soon suspected (correctly) that they were waves of electromagnetic radiation, in other words, another form of light. At that time, the wave model of light—specifically, the Maxwell theory of electromagnetic radiation—was well accepted among scientists, and experiments by Charles Glover Barkla showed that X-rays exhibited phenomena associated with electromagnetic waves, including transverse polarization and spectral lines akin to those observed in the visible wavelengths. Single-slit experiments in the laboratory of Arnold Sommerfeld suggested that X-rays had a wavelength of about 1 angstrom. However, X-rays are composed of photons, and thus are not only waves of electromagnetic

radiation but also exhibit particle-like properties. Albert Einstein introduced the photon concept in 1905, but it was not broadly accepted until 1922, when Arthur Compton confirmed it by the scattering of X-rays from electrons. Therefore, these particle-like properties of X-rays, such as their ionization of gases, caused William Henry Bragg to argue in 1907 that X-rays were *not* electromagnetic radiation. Nevertheless, Bragg's view was not broadly accepted and the observation of X-ray diffraction by Max von Laue in 1912 confirmed for most scientists that X-rays were a form of electromagnetic radiation.

X-ray Diffraction

Crystals are regular arrays of atoms, and X-rays can be considered waves of electromagnetic radiation. Atoms scatter X-ray waves, primarily through the atoms' electrons. Just as an ocean wave striking a lighthouse produces secondary circular waves emanating from the lighthouse, so an X-ray striking an electron produces secondary spherical waves emanating from the electron. This phenomenon is known as elastic scattering, and the electron (or lighthouse) is known as the *scatterer*. A regular array of scatterers produces a regular array of spherical waves. Although these waves cancel one another out in most directions through destructive interference, they add constructively in a few specific directions, determined by Bragg's law:

$$2d \sin\theta = n\lambda$$

Here d is the spacing between diffracting planes, θ is the incident angle, n is any integer, and λ is the wavelength of the beam. These specific directions appear as spots on the diffraction pattern called *reflections*. Thus, X-ray diffraction results from an electromagnetic wave (the X-ray) impinging on a regular array of scatterers (the repeating arrangement of atoms within the crystal).

X-rays are used to produce the diffraction pattern because their wavelength λ is typically the same order of magnitude (1–100 angstroms) as the spacing d between planes in the crystal. In principle, any wave impinging on a regular array of scatterers produces diffraction, as predicted first by Francesco Maria Grimaldi in 1665. To produce significant diffraction, the spacing between the scatterers and the wavelength of the impinging wave should be similar in size. For illustration, the diffraction of sunlight through a bird's feather was first reported by James Gregory in the later 17th century. The first artificial diffraction gratings for visible light were constructed by David Rittenhouse in 1787, and Joseph von Fraunhofer in 1821. However, visible light has too long a wavelength (typically, 5500 angstroms) to observe diffraction from crystals. Prior to the first X-ray diffraction experiments, the spacings between lattice planes in a crystal were not known with certainty.

The idea that crystals could be used as a diffraction grating for X-rays arose in 1912 in a conversation between Paul Peter Ewald and Max von Laue in the English Garden in Munich. Ewald had proposed a resonator model of crystals for his thesis, but this model could not be validated using visible light, since, the wavelength was much larger than the spacing between the resonators. Von Laue realized that electromagnetic radiation of a shorter wavelength was needed to observe such small spacings, and suggested that X-rays might have a wavelength comparable to the unit-cell spacing in crystals. Von Laue worked with two technicians, Walter Friedrich and his assistant Paul Knipping, to shine a beam of X-rays through a copper sulfate crystal and record its diffraction on a photographic plate. After being developed, the plate showed a large number of well-defined spots arranged in a pattern of intersecting circles around the spot produced by the central beam. Von Laue developed a law that connects the scattering angles and the size and orientation of the unit-cell spacings in the crystal, for which he was awarded the Nobel Prize in Physics in 1914.

Scattering

As described in the mathematical derivation below, the X-ray scattering is determined by the density of electrons within the crystal. Since, the energy of an X-ray is much greater than that of a valence electron, the

scattering may be modeled as Thomson scattering, the interaction of an electromagnetic ray with a free electron. This model is generally adopted to describe the polarization of the scattered radiation.

The intensity of Thomson scattering for one particle with mass m and elementary charge q is:

$$I_o = I_c \left(\frac{q^4}{m^2 c^4} \right) \frac{1 + \cos^2 2\theta}{2} = I_e 7.94 \cdot 10^{-26} \frac{1 + \cos^2 2\theta}{2} = I_e f$$

Hence, the atomic nuclei, which are much heavier than an electron, contribute negligibly to the scattered X-rays.

Development from 1912 to 1920

After Von Laue's pioneering research, the field developed rapidly, most notably by physicists William Lawrence Bragg and his father William Henry Bragg. In 1912–1913, the younger Bragg developed Bragg's law, which connects the observed scattering with reflections from evenly spaced planes within the crystal. The Braggs, father and son, shared the 1915 Nobel Prize in Physics for their work in crystallography. The earliest structures were generally simple and marked by one-dimensional symmetry. However, as computational and experimental methods improved over the next decades, it became feasible to deduce reliable atomic positions for more complicated two- and three-dimensional arrangements of atoms in the unit-cell.

The potential of X-ray crystallography for determining the structure of molecules and minerals—then only known vaguely from chemical and hydrodynamic experiments—was realized immediately. The earliest structures were simple inorganic crystals and minerals, but even these revealed fundamental laws of physics and chemistry. The first atomic-resolution structure to be “solved” (*i.e.*, determined) in 1914 was that of table salt.

The distribution of electrons in the table-salt structure showed that crystals are not necessarily composed of covalently bonded molecules, and proved the existence of ionic compounds. The structure of diamond was solved in the same year, proving the tetrahedral arrangement of its chemical bonds and showing that the length of C–C single bond was 1.52 angstroms. Other early structures included copper, calcium fluoride (CaF_2 , also known as *fluorite*), calcite (CaCO_3) and pyrite (FeS_2) in 1914; spinel (MgAl_2O_4) in 1915; the rutile and anatase forms of titanium dioxide (TiO_2) in 1916; pyrochroite $\text{Mn}(\text{OH})_2$ and, by extension, brucite $\text{Mg}(\text{OH})_2$ in 1919;. Also in 1919 sodium nitrate (NaNO_3) and caesium dichloriodide (CsICl_2) were determined by Ralph Walter Graystone Wyckoff, and the wurtzite (hexagonal ZnS) structure became known in 1920.

The structure of graphite was solved in 1916 by the related method of powder diffraction, which was developed by Peter Debye and Paul Scherrer and, independently, by Albert Hull in 1917. The structure of graphite was determined from single-crystal diffraction in 1924 by two groups independently. Hull also used the powder method to determine the structures of various metals, such as iron and magnesium.

Cultural and Aesthetic Importance

In what has been called his scientific autobiography, *The Development of X-ray Analysis*, Sir William Lawrence Bragg mentioned that he believed the field of crystallography was particularly welcoming to women because the techno-aesthetics of the molecular structures resembled textiles and household objects. Bragg was known to compare crystal formation to “curtains, wallpapers, mosaics, and roses”. In 1951, the Festival Pattern Group at the Festival of Britain hosted a collaborative group of textile manufacturers and experienced crystallographers to design lace and prints based on the X-ray crystallography of insulin, china clay, and hemoglobin. One of the leading scientists of the project was Dr. Helen Megaw (1907–2002), the Assistant Director of Research at the Cavendish Laboratory in Cambridge at the time. Megaw is credited as one of the central figures who took inspiration from crystal diagrams and saw their potential in design. In 2008, the Wellcome Collection in London curated an exhibition on the Festival Pattern Group called “From Atom to Patterns”.

Contributions to Chemistry and Material Science

X-ray crystallography has led to a better understanding of chemical bonds and non-covalent interactions. The initial studies revealed the typical radii of atoms, and confirmed many theoretical models of chemical bonding, such as the tetrahedral bonding of carbon in the diamond structure, the octahedral bonding of metals observed in ammonium hexachloroplatinate (IV), and the resonance observed in the planar carbonate group and in aromatic molecules. Kathleen Lonsdale's 1928 structure of hexamethylbenzene established the hexagonal symmetry of benzene and showed a clear difference in bond length between the aliphatic C–C bonds and aromatic C–C bonds; this finding led to the idea of resonance between chemical bonds, which had profound consequences for the development of chemistry. Her conclusions were anticipated by William Henry Bragg, who published models of naphthalene and anthracene in 1921 based on other molecules, an early form of molecular replacement.

Also in the 1920s, Victor Moritz Goldschmidt and later Linus Pauling developed rules for eliminating chemically unlikely structures and for determining the relative sizes of atoms. These rules led to the structure of brookite (1928) and an understanding of the relative stability of the rutile, brookite and anatase forms of titanium dioxide.

The distance between two bonded atoms is a sensitive measure of the bond strength and its bond order; thus, X-ray crystallographic studies have led to the discovery of even more exotic types of bonding in inorganic chemistry, such as metal-metal double bonds, metal-metal quadruple bonds, and three-center, two-electron bonds. X-ray crystallography—or, strictly speaking, an inelastic Compton scattering experiment—has also provided evidence for the partly covalent character of hydrogen bonds. In the field of organometallic chemistry, the X-ray structure of ferrocene initiated scientific studies of sandwich compounds, while that of Zeise's salt stimulated research into “back bonding” and metal- π complexes.

Finally, X-ray crystallography had a pioneering role in the development of supramolecular chemistry, particularly in clarifying the structures of the crown ethers and the principles of host-guest chemistry. In material sciences, many complicated inorganic and organometallic systems have been analyzed using single-crystal methods, such as fullerenes, metalloporphyrins, and other complicated compounds. Single-crystal diffraction is also used in the pharmaceutical industry, due to recent problems with polymorphs. The major factors affecting the quality of single-crystal structures are the crystal's size and regularity; recrystallization is a commonly used technique to improve these factors in small-molecule crystals. The Cambridge Structural Database contains over 800,000 structures as of September 2016; over 99% of these structures were determined by X-ray diffraction.

Mineralogy and Metallurgy

Since the 1920s, X-ray diffraction has been the principal method for determining the arrangement of atoms in minerals and metals. The application of X-ray crystallography to mineralogy began with the structure of garnet, which was determined in 1924 by Menzer. A systematic X-ray crystallographic study of the silicates was undertaken in the 1920s. This study showed that, as the Si/O ratio is altered, the silicate crystals exhibit significant changes in their atomic arrangements. Machatschki extended these insights to minerals in which aluminium substitutes for the silicon atoms of the silicates. The first application of X-ray crystallography to metallurgy likewise occurred in the mid-1920s.

Most notably, Linus Pauling's structure of the alloy Mg_2Sn led to his theory of the stability and structure of complex ionic crystals. On October 17, 2012, the Curiosity rover on the planet Mars at “Rocknest” performed the first X-ray diffraction analysis of Martian soil. The results from the rover's CheMin analyzer revealed the presence of several minerals, including feldspar, pyroxenes and olivine, and suggested that the Martian soil in the sample was similar to the “weathered basaltic soils” of Hawaiian volcanoes.

Early Organic and Small Biological Molecules

The first structure of an organic compound, hexamethylenetetramine, was solved in 1923. This was followed by several studies of long-chain fatty acids, which are an important component of biological membranes. In the 1930s, the structures of much larger molecules with two-dimensional complexity began to be solved. A significant advance was the structure of phthalocyanine, a large planar molecule that is closely related to porphyrin molecules important in biology, such as heme, corrin and chlorophyll. X-ray crystallography of biological molecules took off with Dorothy Crowfoot Hodgkin, who solved the structures of cholesterol (1937), penicillin (1946) and vitamin B₁₂ (1956), for which she was awarded the Nobel Prize in Chemistry in 1964. In 1969, she succeeded in solving the structure of insulin, on which she worked for over thirty years.

Biological Macromolecular Crystallography

Crystal structures of proteins (which are irregular and hundreds of times larger than cholesterol) began to be solved in the late 1950s, beginning with the structure of sperm whale myoglobin by Sir John Cowdery Kendrew, for which he shared the Nobel Prize in Chemistry with Max Perutz in 1962. Since, that success, 132055 X-ray crystal structures of proteins, nucleic acids and other biological molecules have been determined. For comparison, the nearest competing method in terms of structures analyzed is nuclear magnetic resonance (NMR) spectroscopy, which has resolved 11904 chemical structures. Moreover, crystallography can solve structures of arbitrarily large molecules, whereas solution-state NMR is restricted to relatively small ones (less than 70 kDa). X-ray crystallography is now used routinely by scientists to determine how a pharmaceutical drug interacts with its protein target and what changes might improve it. However, intrinsic membrane proteins remain challenging to crystallize because they require detergents or other means to solubilize them in isolation, and such detergents often interfere with crystallization. Such membrane proteins are a large component of the genome, and include many proteins of great physiological importance, such as ion channels and receptors. Helium cryogenics are used to prevent radiation damage in protein crystals.

On the other end of the size scale, even relatively small molecules may pose challenges for the resolving power of X-ray crystallography. The structure assigned in 1991 to the antibiotic isolated from a marine organism, diazonamide A (C₄₀H₃₄Cl₂N₆O₆, molar mass 765.65 g/mol), proved to be incorrect by the classical proof of structure: a synthetic sample was not identical to the natural product.

The mistake was attributed to the inability of X-ray crystallography to distinguish between the correct-OH/>NH and the interchanged-NH₂/-O-groups in the incorrect structure. With advances in instrumentation, however, it is now routinely possible to distinguish between such similar groups using modern single-crystal X-ray diffractometers.

Relationship to Other Scattering Techniques

Elastic vs. Inelastic Scattering

X-ray crystallography is a form of elastic scattering; the outgoing X-rays have the same energy, and thus same wavelength, as the incoming X-rays, only with altered direction. By contrast, *inelastic scattering* occurs when energy is transferred from the incoming X-ray to the crystal, *e.g.*, by exciting an inner-shell electron to a higher energy level. Such inelastic scattering reduces the energy (or increases the wavelength) of the outgoing beam. Inelastic scattering is useful for probing such excitations of matter, but not in determining the distribution of scatterers within the matter, which is the goal of X-ray crystallography. X-rays range in wavelength from 10 to 0.01 nanometers; a typical wavelength used for crystallography is 1 Å (0.1 nm), which is on the scale of

covalent chemical bonds and the radius of a single atom. Longer-wavelength photons (such as ultraviolet radiation) would not have sufficient resolution to determine the atomic positions. At the other extreme, shorter-wavelength photons such as gamma rays are difficult to produce in large numbers, difficult to focus, and interact too strongly with matter, producing particle-antiparticle pairs. Therefore, X-rays are the “sweet spot” for wavelength when determining atomic-resolution structures from the scattering of electromagnetic radiation.

Other X-ray Techniques

Other forms of elastic X-ray scattering include powder diffraction, Small-Angle X-ray Scattering (SAXS) and several types of X-ray fibre diffraction, which was used by Rosalind Franklin in determining the double-helix structure of DNA. In general, single-crystal X-ray diffraction offers more structural information than these other techniques; however, it requires a sufficiently large and regular crystal, which is not always available. These scattering methods generally use *monochromatic* X-rays, which are restricted to a single wavelength with minor deviations. A broad spectrum of X-rays (that is, a blend of X-rays with different wavelengths) can also be used to carry out X-ray diffraction, a technique known as the Laue method. This is the method used in the original discovery of X-ray diffraction. Laue scattering provides much structural information with only a short exposure to the X-ray beam, and is therefore used in structural studies of very rapid events (Time resolved crystallography). However, it is not as well-suited as monochromatic scattering for determining the full atomic structure of a crystal and therefore works better with crystals with relatively simple atomic arrangements.

The Laue back reflection mode records X-rays scattered backwards from a broad spectrum source. This is useful if the sample is too thick for X-rays to transmit through it. The diffracting planes in the crystal are determined by knowing that the normal to the diffracting plane bisects the angle between the incident beam and the diffracted beam. A Grevinger chart can be used to interpret the back reflection Laue photograph.

Electron and Neutron Diffraction

Other particles, such as electrons and neutrons, may be used to produce a diffraction pattern. Although electron, neutron, and X-ray scattering are based on different physical processes, the resulting diffraction patterns are analyzed using the same coherent diffraction imaging techniques. As derived below, the electron density within the crystal and the diffraction patterns are related by a simple mathematical method, the Fourier transform, which allows the density to be calculated relatively easily from the patterns. However, this works only if the scattering is *weak*, *i.e.*, if the scattered beams are much less intense than the incoming beam. Weakly scattered beams pass through the remainder of the crystal without undergoing a second scattering event. Such re-scattered waves are called “secondary scattering” and hinder the analysis. Any sufficiently thick crystal will produce secondary scattering, but since, X-rays interact relatively weakly with the electrons, this is generally not a significant concern. By contrast, electron beams may produce strong secondary scattering even for relatively thin crystals (>100 nm). Since, this thickness corresponds to the diameter of many viruses, a promising direction is the electron diffraction of isolated macromolecular assemblies, such as viral capsids and molecular machines, which may be carried out with a cryo-electron microscope. Moreover, the strong interaction of electrons with matter (about 1000 times stronger than for X-rays) allows determination of the atomic structure of extremely small volumes. The field of applications for electron crystallography ranges from bio molecules like membrane proteins over organic thin films to the complex structures of (nanocrystalline) intermetallic compounds and zeolites.

Neutron diffraction is an excellent method for structure determination, although it has been difficult to obtain intense, monochromatic beams of neutrons in sufficient quantities. Traditionally, nuclear reactors have been used, although sources producing neutrons by spallation are becoming increasingly available. Being

uncharged, neutrons scatter much more readily from the atomic nuclei rather than from the electrons. Therefore, neutron scattering is very useful for observing the positions of light atoms with few electrons, especially hydrogen, which is essentially invisible in the X-ray diffraction. Neutron scattering also has the remarkable property that the solvent can be made invisible by adjusting the ratio of normal water, H_2O , and heavy water, D_2O .

Methods

Overview of Single-crystal X-ray Diffraction

The oldest and most precise method of X-ray crystallography is *single-crystal X-ray diffraction*, in which a beam of X-rays strikes a single crystal, producing scattered beams. When they land on a piece of film or other detector, these beams make a *diffraction pattern* of spots; the strengths and angles of these beams are recorded as the crystal is gradually rotated. Each spot is called a *reflection*, since, it corresponds to the reflection of the X-rays from one set of evenly spaced planes within the crystal. For single crystals of sufficient purity and regularity, X-ray diffraction data can determine the mean chemical bond lengths and angles to within a few thousandths of an angstrom and to within a few tenths of a degree, respectively. The atoms in a crystal are not static, but oscillate about their mean positions, usually by less than a few tenths of an angstrom. X-ray crystallography allows measuring the size of these oscillations.

Procedure

The technique of single-crystal X-ray crystallography has three basic steps. The first—and often most difficult—step is to obtain an adequate crystal of the material under study. The crystal should be sufficiently large (typically larger than 0.1 mm in all dimensions), pure in composition and regular in structure, with no significant internal imperfections such as cracks or twinning.

In the second step, the crystal is placed in an intense beam of X-rays, usually of a single wavelength (*monochromatic X-rays*), producing the regular pattern of reflections. The angles and intensities of diffracted X-rays are measured, with each compound having a unique diffraction pattern. As the crystal is gradually rotated, previous reflections disappear and new ones appear; the intensity of every spot is recorded at every orientation of the crystal. Multiple data sets may have to be collected, with each set covering slightly more than half a full rotation of the crystal and typically containing tens of thousands of reflections. In the third step, these data are combined computationally with complementary chemical information to produce and refine a model of the arrangement of atoms within the crystal. The final, refined model of the atomic arrangement—now called a *crystal structure*—is usually stored in a public database.

Limitations

As the crystal's repeating unit, its unit cell, becomes larger and more complex, the atomic-level picture provided by X-ray crystallography becomes less well-resolved (more “fuzzy”) for a given number of observed reflections. Two limiting cases of X-ray crystallography—“small-molecule” (which includes continuous inorganic solids) and “macromolecular” crystallography—are often discerned. *Small-molecule crystallography* typically involves crystals with fewer than 100 atoms in their asymmetric unit; such crystal structures are usually so well resolved that the atoms can be discerned as isolated “blobs” of electron density. By contrast, *macromolecular crystallography* often involves tens of thousands of atoms in the unit cell. Such crystal structures are generally less well-resolved (more “smeared out”); the atoms and chemical bonds appear as tubes of electron density, rather than as isolated atoms. In general, small molecules are also easier to crystallize than macromolecules; however, X-ray crystallography has proven possible even for viruses and proteins with hundreds of thousands

of atoms, through improved crystallographic imaging and technology. Though normally X-ray crystallography can only be performed if the sample is in crystal form, new research has been done into sampling non-crystalline forms of samples.

Crystallization

Although crystallography can be used to characterize the disorder in an impure or irregular crystal, crystallography generally requires a pure crystal of high regularity to solve the structure of a complicated arrangement of atoms. Pure, regular crystals can sometimes be obtained from natural or synthetic materials, such as samples of metals, minerals or other macroscopic materials. The regularity of such crystals can sometimes be improved with macromolecular crystal annealing and other methods. However, in many cases, obtaining a diffraction-quality crystal is the chief barrier to solving its atomic-resolution structure.

Small-molecule and macromolecular crystallography differ in the range of possible techniques used to produce diffraction-quality crystals. Small molecules generally have few degrees of conformational freedom, and may be crystallized by a wide range of methods, such as chemical vapor deposition and recrystallization. By contrast, macromolecules generally have many degrees of freedom and their crystallization must be carried out so as to maintain a stable structure. For example, proteins and larger RNA molecules cannot be crystallized if their tertiary structure has been unfolded; therefore, the range of crystallization conditions is restricted to solution conditions in which such molecules remain folded.

Protein crystals are almost always grown in solution. The most common approach is to lower the solubility of its component molecules very gradually; if this is done too quickly, the molecules will precipitate from solution, forming a useless dust or amorphous gel on the bottom of the container. Crystal growth in solution is characterized by two steps: *nucleation* of a microscopic crystallite (possibly having only 100 molecules), followed by *growth* of that crystallite, ideally to a diffraction-quality crystal. The solution conditions that favor the first step (nucleation) are not always the same conditions that favor the second step (subsequent growth). The crystallographer's goal is to identify solution conditions that favor the development of a single, large crystal, since, larger crystals offer improved resolution of the molecule. Consequently, the solution conditions should *disfavor* the first step (nucleation) but *favor* the second (growth), so that only one large crystal forms per droplet. If nucleation is favored too much, a shower of small crystallites will form in the droplet, rather than one large crystal; if favored too little, no crystal will form whatsoever. Other approaches involve, crystallizing proteins under oil, where aqueous protein solutions are dispensed under liquid oil, and water evaporates through the layer of oil.

Different oils have different evaporation permeabilities, therefore yielding changes in concentration rates from different precipitant/protein mixture. The technique relies on bringing the protein directly into the nucleation zone by mixing protein with the appropriate amount of precipitant to prevent the diffusion of water out of the drop. It is extremely difficult to predict good conditions for nucleation or growth of well-ordered crystals. In practice, favorable conditions are identified by *screening*; a very large batch of the molecules is prepared, and a wide variety of crystallization solutions are tested. Hundreds, even thousands, of solution conditions are generally tried before finding the successful one.

The various conditions can use one or more physical mechanisms to lower the solubility of the molecule; for example, some may change the pH, some contain salts of the Hofmeister series or chemicals that lower the dielectric constant of the solution, and still others contain large polymers such as polyethylene glycol that drive the molecule out of solution by entropic effects. It is also common to try several temperatures for encouraging crystallization, or to gradually lower the temperature so that the solution becomes supersaturated. These methods require large amounts of the target molecule, as they use high concentration of the molecule(s) to be crystallized.

Due to the difficulty in obtaining such large quantities (milligrams) of crystallization-grade protein, robots have been developed that are capable of accurately dispensing crystallization trial drops that are in the order of 100 nanoliters in volume. This means that 10-fold less protein is used per experiment when compared to crystallization trials set up by hand (in the order of 1 microliter).

Several factors are known to inhibit or mar crystallization. The growing crystals are generally held at a constant temperature and protected from shocks or vibrations that might disturb their crystallization. Impurities in the molecules or in the crystallization solutions are often inimical to crystallization. Conformational flexibility in the molecule also tends to make crystallization less likely, due to entropy.

Ironically, molecules that tend to self-assemble into regular helices are often unwilling to assemble into crystals. Crystals can be marred by twinning, which can occur when a unit cell can pack equally favorably in multiple orientations; although recent advances in computational methods may allow solving the structure of some twinned crystals. Having failed to crystallize a target molecule, a crystallographer may try again with a slightly modified version of the molecule; even small changes in molecular properties can lead to large differences in crystallization behaviour.

Data Collection

Mounting the Crystal

The crystal is mounted for measurements so that it may be held in the X-ray beam and rotated. There are several methods of mounting. In the past, crystals were loaded into glass capillaries with the crystallization solution (the mother liquor). Nowadays, crystals of small molecules are typically attached with oil or glue to a glass fibre or a loop, which is made of nylon or plastic and attached to a solid rod. Protein crystals are scooped up by a loop, then flash-frozen with liquid nitrogen. This freezing reduces the radiation damage of the X-rays, as well as the noise in the Bragg peaks due to thermal motion (the Debye-Waller effect). However, untreated protein crystals often crack if flash-frozen; therefore, they are generally pre-soaked in a cryoprotectant solution before freezing. Unfortunately, this pre-soak may itself cause the crystal to crack, ruining it for crystallography. Generally, successful cryo-conditions are identified by trial and error.

The capillary or loop is mounted on a goniometer, which allows it to be positioned accurately within the X-ray beam and rotated. Since, both the crystal and the beam are often very small, the crystal must be centered within the beam to within ~25 micrometers accuracy, which is aided by a camera focused on the crystal. The most common type of goniometer is the “kappa goniometer”, which offers three angles of rotation: the ω angle, which rotates about an axis perpendicular to the beam; the κ angle, about an axis at ~50° to the ω axis; and, finally, the τ angle about the loop/capillary axis. When the κ angle is zero, the ω and τ axes are aligned. The κ rotation allows for convenient mounting of the crystal, since, the arm in which the crystal is mounted may be swung out towards the crystallographer. The oscillations carried out during data collection (mentioned below) involve the ω axis only. An older type of goniometer is the four-circle goniometer, and its relatives such as the six-circle goniometer.

X-ray Sources

Rotating anode:

- Small scale can be done on a local X-ray tube source, typically coupled with an image plate detector. These have the advantage of being (relatively) inexpensive and easy to maintain, and allow for quick screening and collection of samples. However, the wavelength light produced is limited by anode material, typically copper. Further, intensity is limited by the power applied and cooling capacity available to

avoid melting the anode. In such systems, electrons are boiled off of a cathode and accelerated through a strong electric potential of ~50 kV; having reached a high speed, the electrons collide with a metal plate, emitting *bremsstrahlung* and some strong spectral lines corresponding to the excitation of inner-shell electrons of the metal. The most common metal used is copper, which can be kept cool easily, due to its high thermal conductivity, and which produces strong K_{α} and K_{β} lines. The K_{β} line is sometimes suppressed with a thin (~10 μm) nickel foil. The simplest and cheapest variety of sealed X-ray tube has a stationary anode (the Crookes tube) and run with ~2 kW of electron beam power. The more expensive variety has a rotating-anode type source that run with ~14 kW of e-beam power.

- X-rays are generally filtered (by use of X-ray filters) to a single wavelength (made monochromatic) and collimated to a single direction before they are allowed to strike the crystal. The filtering not only simplifies the data analysis, but also removes radiation that degrades the crystal without contributing useful information. Collimation is done either with a collimator (basically, a long tube) or with a clever arrangement of gently curved mirrors. Mirror systems are preferred for small crystals (under 0.3 mm) or with large unit cells (over 150 $\text{\AA}$).
- Rotating anodes were used by Joanna (Joka) Maria Vandenberg in the first experiments that demonstrated the power of X rays for quick (in real time production) screening of large InGaAsP thin film wafers for quality control of quantum well lasers.

Synchrotron radiation:

- Synchrotron radiation are some of the brightest lights on earth. It is the single most powerful tool available to X-ray crystallographers. It is made of X-ray beams generated in large machines called synchrotrons. These machines accelerate electrically charged particles, often electrons, to nearly the speed of light and confine them in a (roughly) circular loop using magnetic fields.
- Synchrotrons are generally national facilities, each with several dedicated beamlines where data is collected without interruption. Synchrotrons were originally designed for use by high-energy physicists studying subatomic particles and cosmic phenomena. The largest component of each synchrotron is its electron storage ring. This ring is actually not a perfect circle, but a many-sided polygon. At each corner of the polygon, or sector, precisely aligned magnets bend the electron stream. As the electrons' path is bent, they emit bursts of energy in the form of X-rays.
- Using synchrotron radiation frequently has specific requirements for X-ray crystallography. The intense ionizing radiation can cause radiation damage to samples, particularly macromolecular crystals. Cryo crystallography protects the sample from radiation damage, by freezing the crystal at liquid nitrogen temperatures (~100 K). However, synchrotron radiation frequently has the advantage of user selectable wavelengths, allowing for anomalous scattering experiments which maximizes anomalous signal. This is critical in experiments such as SAD and MAD.

Free electron laser:

- Recently, free-electron lasers have been developed for use in X-ray crystallography. These are the brightest X-ray sources currently available; with the X-rays coming in femtosecond bursts. The intensity of the source is such that atomic resolution diffraction patterns can be resolved for crystals otherwise too small for collection. However, the intense light source also destroys the sample, requiring multiple crystals to be shot. As each crystal is randomly oriented in the beam, hundreds of thousands of individual diffraction images must be collected in order to get a complete data-set. This method, serial femtosecond crystallography, has been used in solving the structure of a number of protein crystal structures, sometimes noting differences with equivalent structures collected from synchrotron sources.

Recording the Reflections

When a crystal is mounted and exposed to an intense beam of X-rays, it scatters the X-rays into a pattern of spots or *reflections* that can be observed on a screen behind the crystal. A similar pattern may be seen by shining a laser pointer at a compact disc. The relative intensities of these spots provide the information to determine the arrangement of molecules within the crystal in atomic detail. The intensities of these reflections may be recorded with photographic film, an area detector or with a charge-coupled device (CCD) image sensor. The peaks at small angles correspond to low-resolution data, whereas those at high angles represent high-resolution data; thus, an upper limit on the eventual resolution of the structure can be determined from the first few images. Some measures of diffraction quality can be determined at this point, such as the mosaicity of the crystal and its overall disorder, as observed in the peak widths. Some pathologies of the crystal that would render it unfit for solving the structure can also be diagnosed quickly at this point.

One image of spots is insufficient to reconstruct the whole crystal; it represents only a small slice of the full Fourier transform. To collect all the necessary information, the crystal must be rotated step-by-step through 180° , with an image recorded at every step; actually, slightly more than 180° is required to cover reciprocal space, due to the curvature of the Ewald sphere. However, if the crystal has a higher symmetry, a smaller angular range such as 90° or 45° may be recorded. The rotation axis should be changed at least once, to avoid developing a “blind spot” in reciprocal space close to the rotation axis. It is customary to rock the crystal slightly (by $0.5\text{--}2^\circ$) to catch a broader region of reciprocal space. Multiple data sets may be necessary for certain phasing methods. For example, MAD phasing requires that the scattering be recorded at least three (and usually four, for redundancy) wavelengths of the incoming X-ray radiation. A single crystal may degrade too much during the collection of one data set, owing to radiation damage; in such cases, data sets on multiple crystals must be taken.

Data Analysis

Crystal Symmetry, Unit Cell, and Image Scaling

The recorded series of two-dimensional diffraction patterns, each corresponding to a different crystal orientation, is converted into a three-dimensional model of the electron density; the conversion uses the mathematical technique of Fourier transforms, which is explained below. Each spot corresponds to a different type of variation in the electron density; the crystallographer must determine *which* variation corresponds to *which* spot (*indexing*), the relative strengths of the spots in different images (*merging and scaling*) and how the variations should be combined to yield the total electron density (*phasing*).

Data processing begins with *indexing* the reflections. This means identifying the dimensions of the unit cell and which image peak corresponds to which position in reciprocal space. A byproduct of indexing is to determine the symmetry of the crystal, *i.e.*, its *space group*. Some space groups can be eliminated from the beginning. For example, reflection symmetries cannot be observed in chiral molecules; thus, only 65 space groups of 230 possible are allowed for protein molecules which are almost always chiral. Indexing is generally accomplished using an *autoindexing* routine. Having assigned symmetry, the data is then *integrated*. This converts the hundreds of images containing the thousands of reflections into a single file, consisting of (at the very least) records of the Miller index of each reflection, and an intensity for each reflection (at this state the file often also includes error estimates and measures of partiality (what part of a given reflection was recorded on that image)).

A full data set may consist of hundreds of separate images taken at different orientations of the crystal. The first step is to merge and scale these various images, that is, to identify which peaks appear in two or more images (*merging*) and to scale the relative images so that they have a consistent intensity scale. Optimizing the

intensity scale is critical because the relative intensity of the peaks is the key information from which the structure is determined. The repetitive technique of crystallographic data collection and the often high symmetry of crystalline materials cause the diffractometer to record many symmetry-equivalent reflections multiple times. This allows calculating the symmetry-related R-factor, a reliability index based upon how similar are the measured intensities of symmetry-equivalent reflections, thus assessing the quality of the data.

Initial Phasing

The data collected from a diffraction experiment is a reciprocal space representation of the crystal lattice. The position of each diffraction 'spot' is governed by the size and shape of the unit cell, and the inherent symmetry within the crystal. The intensity of each diffraction 'spot' is recorded, and this intensity is proportional to the square of the *structure factor* amplitude. The structure factor is a complex number containing information relating to both the amplitude and phase of a wave. In order to obtain an interpretable *electron density map*, both amplitude and phase must be known (an electron density map allows a crystallographer to build a starting model of the molecule). The phase cannot be directly recorded during a diffraction experiment: this is known as the phase problem. Initial phase estimates can be obtained in a variety of ways:

- *Ab initio phasing or direct methods*: This is usually the method of choice for small molecules (<1000 non-hydrogen atoms), and has been used successfully to solve the phase problems for small proteins. If the resolution of the data is better than 1.4 Å (140 pm), direct methods can be used to obtain phase information, by exploiting known phase relationships between certain groups of reflections.
- *Molecular replacement*: If a related structure is known, it can be used as a search model in molecular replacement to determine the orientation and position of the molecules within the unit cell. The phases obtained this way can be used to generate *electron density maps*.
- *Anomalous X-ray scattering (MAD or SAD phasing)*: The X-ray wavelength may be scanned past an absorption edge of an atom, which changes the scattering in a known way. By recording full sets of reflections at three different wavelengths (far below, far above and in the middle of the absorption edge) one can solve for the substructure of the anomalously diffracting atoms and hence, the structure of the whole molecule. The most popular method of incorporating anomalous scattering atoms into proteins is to express the protein in a methionine auxotroph (a host incapable of synthesizing methionine) in a media rich in seleno-methionine, which contains selenium atoms. A MAD experiment can then be conducted around the absorption edge, which should then yield the position of any methionine residues within the protein, providing initial phases.
- *Heavy atom methods (multiple isomorphous replacement)*: If electron-dense metal atoms can be introduced into the crystal, direct methods or Patterson-space methods can be used to determine their location and to obtain initial phases. Such heavy atoms can be introduced either by soaking the crystal in a heavy atom-containing solution, or by co-crystallization (growing the crystals in the presence of a heavy atom). As in MAD phasing, the changes in the scattering amplitudes can be interpreted to yield the phases. Although this is the original method by which protein crystal structures were solved, it has largely been superseded by MAD phasing with selenomethionine.

Model Building and Phase Refinement

Having obtained initial phases, an initial model can be built. This model can be used to refine the phases, leading to an improved model, and so on. Given a model of some atomic positions, these positions and their respective Debye-Waller factors (or B-factors, accounting for the thermal motion of the atom) can be refined to fit the observed diffraction data, ideally yielding a better set of phases. A new model can then be fit to the new electron density map and a further round of refinement is carried out. This continues until the correlation between the diffraction data and the model is maximized.

The agreement is measured by an R -factor defined as,

$$R = \frac{\sum_{\text{all reflections}} |F_o - F_c|}{\sum_{\text{all reflections}} |F_o|}$$

where F is the structure factor. A similar quality criterion is R_{free} , which is calculated from a subset (~10%) of reflections that were not included in the structure refinement. Both R factors depend on the resolution of the data. As a rule of thumb, R_{free} should be approximately the resolution in angstroms divided by 10; thus, a data-set with 2 Å resolution should yield a final $R_{\text{free}} \sim 0.2$. Chemical bonding features such as stereochemistry, hydrogen bonding and distribution of bond lengths and angles are complementary measures of the model quality. Phase bias is a serious problem in such iterative model building. *Omit maps* are a common technique used to check for this. It may not be possible to observe every atom in the asymmetric unit. In many cases, Crystallographic disorder smears the electron density map. Weakly scattering atoms such as hydrogen are routinely invisible. It is also possible for a single atom to appear multiple times in an electron density map, *e.g.*, if a protein sidechain has multiple (<4) allowed conformations. In still other cases, the crystallographer may detect that the covalent structure deduced for the molecule was incorrect, or changed. For example, proteins may be cleaved or undergo post-translational modifications that were not detected prior to the crystallization.

Disorder

A common challenge in refinement of crystal structures results from crystallographic disorder. Disorder can take many forms but in general involves the coexistence of two or more species or conformations. Failure to recognize disorder results in flawed interpretation. Pitfalls from improper modeling of disorder are illustrated by the discounted hypothesis of bond stretch isomerism. Disorder is modelled with respect to the relative population of the components, often only two, and their identity. In structures of large molecules and ions, solvent and counterions are often disordered.

Deposition of the Structure

Once the model of a molecule's structure has been finalized, it is often deposited in a crystallographic database such as the Cambridge Structural Database (for small molecules), the Inorganic Crystal Structure Database (ICSD) (for inorganic compounds) or the Protein Data Bank (for protein structures). Many structures obtained in private commercial ventures to crystallize medicinally relevant proteins are not deposited in public crystallographic databases.

Applications of X-ray Diffraction

X-ray diffraction has wide and various applications in the chemical, biochemical, physical, material and mineralogical sciences. Laue claimed in 1937 that the technique “has extended the power of observing minute structure ten thousand times beyond that given us by the microscope”. X-ray diffraction is analogous to a microscope with atomic-level resolution which shows the atoms and their electron distribution. X-ray diffraction, electron diffraction, and neutron diffraction give information about the structure of matter, crystalline and non-crystalline, at the atomic and molecular level. In addition, these methods may be applied in the study of properties of all materials, inorganic, organic or biological. Due to the importance and variety of applications of diffraction studies of crystals, many Nobel Prizes have been awarded for such studies.

X-ray Method for Investigation of Drugs

X-ray diffraction has been used for the identification of antibiotic drugs such as: eight β -lactam (ampicillin sodium, penicillin G procaine, cefalexin, ampicillin trihydrate, benzathine penicillin, benzylpenicillin sodium,

cefotaxime sodium, Ceftriaxone sodium), three tetracycline (doxycycline hydrochloride, oxytetracycline dehydrate, tetracycline hydrochloride) and two macrolide (azithromycin, erythromycin estolate) antibiotic drugs. Each of these drugs has a unique XRD pattern that makes their identification possible.

X-ray Method for Investigation of Textile Fibers and Polymers

Forensic examination of any trace evidence is based upon Locard's exchange principle. This states that "every contact leaves a trace". In practice, even though a transfer of material has taken place, it may be impossible to detect, because the amount transferred is very small. Textile fibers are a mixture of crystalline and amorphous substances. Therefore, the measurement of the degree of crystalline gives useful data in the characterization of fibers using X-ray diffractometry. It has been reported that X-ray diffraction was used to identify of a "crystalline" deposit which was found on a chair. The deposit was found to be amorphous, but the diffraction pattern present matched that of polymethylmethacrylate. Pyrolysis mass spectrometry later identified the deposit as polymethylcyanoacrylate of Boin crystal parameters.

X-ray method for investigation of bones

Hiller investigated the effects of heating and burning on bone mineral using X-ray diffraction (XRD) techniques. The bone samples were heated in temperature of 500, 700, and 900 C° for 15 and 45 min. The results show bone crystals began to change during the first 15 min of heating at 500 C° and above. At higher temperatures, thickness and shape of crystals of bones appear stabilized, but when the samples were heated at lower temperature or for shorter period, XRD traces showed extreme changes in crystal parameters.

Integrated Circuits

X-ray diffraction has been demonstrated as a method for investigating the complex structure of integrated circuits.

NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY OF PROTEINS

Nuclear magnetic resonance spectroscopy of proteins (usually abbreviated protein NMR) is a field of structural biology in which NMR spectroscopy is used to obtain information about the structure and dynamics of proteins, and also nucleic acids, and their complexes. The field was pioneered by Richard R. Ernst and Kurt Wüthrich at the ETH, and by Ad Bax, Marius Clore, and Angela Gronenborn at the NIH, among others. Structure determination by NMR spectroscopy usually consists of several phases, each using a separate set of highly specialized techniques. The sample is prepared, measurements are made, interpretive approaches are applied, and a structure is calculated and validated.

NMR involves the quantum mechanical properties of the central core ("nucleus") of the atom. These properties depend on the local molecular environment, and their measurement provides a map of how the atoms are linked chemically, how close they are in space, and how rapidly they move with respect to each other. These properties are fundamentally the same as those used in the more familiar magnetic resonance imaging (MRI), but the molecular applications use a somewhat different approach, appropriate to the change of scale from millimeters (of interest to radiologists) to nanometers (bonded atoms are typically a fraction of a nanometer apart), a factor of a million. This change of scale requires much higher sensitivity of detection and stability for long term measurement. In contrast to MRI, structural biology studies do not directly generate an image, but rely on complex computer calculations to generate three-dimensional molecular models. Currently most samples are examined in a solution in water, but methods are being developed to also work with solid samples. Data collection

relies on placing the sample inside a powerful magnet, sending radio frequency signals through the sample, and measuring the absorption of those signals. Depending on the environment of atoms within the protein, the nuclei of individual atoms will absorb different frequencies of radio signals. Furthermore, the absorption signals of different nuclei may be perturbed by adjacent nuclei. This information can be used to determine the distance between nuclei. These distances in turn can be used to determine the overall structure of the protein.

A typical study might involve how two proteins interact with each other, possibly with a view to developing small molecules that can be used to probe the normal biology of the interaction (“chemical biology”) or to provide possible leads for pharmaceutical use (drug development). Frequently, the interacting pair of proteins may have been identified by studies of human genetics, indicating the interaction can be disrupted by unfavorable mutations, or they may play a key role in the normal biology of a “model” organism like the fruit fly, yeast, the worm *C. elegans*, or mice. To prepare a sample, methods of molecular biology are typically used to make quantities by bacterial fermentation. This also permits changing the isotopic composition of the molecule, which is desirable because the isotopes behave differently and provide methods for identifying overlapping NMR signals.

Sample Preparation

Protein nuclear magnetic resonance is performed on aqueous samples of highly purified protein. Usually, the sample consists of between 300 and 600 microlitres with a protein concentration in the range 0.1–3 millimolar. The source of the protein can be either natural or produced in a production system using recombinant DNA techniques through genetic engineering. Recombinantly expressed proteins are usually easier to produce in sufficient quantity, and this method makes isotopic labeling possible. The purified protein is usually dissolved in a buffer solution and adjusted to the desired solvent conditions. The NMR sample is prepared in a thin-walled glass tube.

Data Collection

Protein NMR utilizes multidimensional nuclear magnetic resonance experiments to obtain information about the protein. Ideally, each distinct nucleus in the molecule experiences a distinct electronic environment and thus has a distinct chemical shift by which it can be recognized. However, in large molecules such as proteins the number of resonances can typically be several thousand and a one-dimensional spectrum inevitably has incidental overlaps. Therefore, multidimensional experiments that correlate the frequencies of distinct nuclei are performed. The additional dimensions decrease the chance of overlap and have a larger information content, since, they correlate signals from nuclei within a specific part of the molecule. Magnetization is transferred into the sample using pulses of electromagnetic (radiofrequency) energy and between nuclei using delays; the process is described with so-called pulse sequences. Pulse sequences allow the experimenter to investigate and select specific types of connections between nuclei. The array of nuclear magnetic resonance experiments used on proteins fall in two main categories—one where magnetization is transferred through the chemical bonds, and one where the transfer is through space, irrespective of the bonding structure. The first category is used to assign the different chemical shifts to a specific nucleus, and the second is primarily used to generate the distance restraints used in the structure calculation, and in the assignment with unlabelled protein.

Depending on the concentration of the sample, on the magnetic field of the spectrometer, and on the type of experiment, a single multidimensional nuclear magnetic resonance experiment on a protein sample may take hours or even several days to obtain suitable signal-to-noise ratio through signal averaging, and to allow for sufficient evolution of magnetization transfer through the various dimensions of the experiment. Other things being equal, higher-dimensional experiments will take longer than lower-dimensional experiments. Typically, the first experiment to be measured with an isotope-labelled protein is a 2D heteronuclear single quantum

correlation (HSQC) spectrum, where “heteronuclear” refers to nuclei other than ^1H . In theory, the heteronuclear single quantum correlation has one peak for each H bound to a heteronucleus. Thus, in the ^{15}N -HSQC one signal is expected for each amino acid residue with the exception of proline, which has no amide-hydrogen due to the cyclic nature of its backbone. Tryptophan and certain other residues with N-containing sidechains also give rise to additional signals. The ^{15}N -HSQC is often referred to as the fingerprint of a protein because each protein has a unique pattern of signal positions. Analysis of the ^{15}N -HSQC allows researchers to evaluate whether the expected number of peaks is present and thus to identify possible problems due to multiple conformations or sample heterogeneity. The relatively quick heteronuclear single quantum correlation experiment helps determine the feasibility of doing subsequent longer, more expensive, and more elaborate experiments. It is not possible to assign peaks to specific atoms from the heteronuclear single quantum correlation alone.

Resonance Assignment

In order to analyze the nuclear magnetic resonance data, it is important to get a resonance assignment for the protein, that is to find out which chemical shift corresponds to which atom. This is typically achieved by sequential walking using information derived from several different types of NMR experiment. The exact procedure depends on whether the protein is isotopically labelled or not, since, a lot of the assignment experiments depend on carbon-13 and nitrogen-15.

Homonuclear Nuclear Magnetic Resonance

With unlabelled protein the usual procedure is to record a set of two dimensional homonuclear nuclear magnetic resonance experiments through correlation spectroscopy (COSY), of which several types include conventional correlation spectroscopy, *total correlation* spectroscopy (TOCSY) and nuclear Overhauser effect spectroscopy (NOESY). A two-dimensional nuclear magnetic resonance experiment produces a two-dimensional spectrum. The units of both axes are chemical shifts. The COSY and TOCSY transfer magnetization through the chemical bonds between adjacent protons. The conventional correlation spectroscopy experiment is only able to transfer magnetization between protons on adjacent atoms, whereas in the total correlation spectroscopy experiment the protons are able to relay the magnetization, so it is transferred among all the protons that are connected by adjacent atoms. Thus in a conventional correlation spectroscopy, an alpha proton transfers magnetization to the beta protons, the beta protons transfers to the alpha and gamma protons, if any are present, then the gamma proton transfers to the beta and the delta protons, and the process continues. In total correlation spectroscopy, the alpha and all the other protons are able to transfer magnetization to the beta, gamma, delta, epsilon if they are connected by a continuous chain of protons.

The continuous chain of protons are the sidechain of the individual amino acids. Thus these two experiments are used to build so called spin systems, that is build a list of resonances of the chemical shift of the peptide proton, the alpha protons and all the protons from each residue's sidechain. Which chemical shifts corresponds to which nuclei in the spin system is determined by the conventional correlation spectroscopy connectivities and the fact that different types of protons have characteristic chemical shifts. To connect the different spinsystems in a sequential order, the nuclear Overhauser effect spectroscopy experiment has to be used. Because this experiment transfers magnetization through space, it will show crosspeaks for all protons that are close in space regardless of whether they are in the same spin system or not. The neighbouring residues are inherently close in space, so the assignments can be made by the peaks in the NOESY with other spin systems. One important problem using homonuclear nuclear magnetic resonance is overlap between peaks. This occurs when different protons have the same or very similar chemical shifts. This problem becomes greater as the protein becomes larger, so homonuclear nuclear magnetic resonance is usually restricted to small proteins or peptides.

Nitrogen-15 Nuclear Magnetic Resonance

The most commonly performed ^{15}N experiment is the ^1H - ^{15}N HSQC. The experiment is highly sensitive and therefore can be performed relatively quickly. It is often used to check the suitability of a protein for structure determination using NMR, as well as for the optimization of the sample conditions. It is one of the standard suite of experiments used for the determination of the solution structure of protein. The HSQC can be further expanded into three- and four dimensional NMR experiments, such as ^{15}N -TOCSY-HSQC and ^{15}N -NOESY-HSQC.

Carbon-13 and Nitrogen-15 Nuclear Magnetic Resonance

When the protein is labelled with carbon-13 and nitrogen-15 it is possible to record triple resonance experiments that transfer magnetisation over the peptide bond, and thus connect different spin systems through bonds. This is usually done using some of the following experiments, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB and CBCA(CO)NH. All six experiments consist of a ^1H - ^{15}N plane (similar to a HSQC spectrum) expanded with a carbon dimension. In the HN(CA)CO, each H_N plane contains the peaks from the carbonyl carbon from its residue as well the preceding one in the sequence.

The HNCO contains the carbonyl carbon chemical shift from only the preceding residue, but is much more sensitive than HN(CA)CO. These experiments allow each ^1H - ^{15}N peak to be linked to the preceding carbonyl carbon, and sequential assignment can then be undertaken by matching the shifts of each spin system's own and previous carbons. The HNCA and HN(CO)CA works similarly, just with the alpha carbons (C_α) rather than the carbonyls, and the HNCACB and the CBCA(CO)NH contains both the alpha carbon and the beta carbon (C_β). Usually several of these experiments are required to resolve overlap in the carbon dimension. This procedure is usually less ambiguous than the NOESY-based method, since, it is based on through bond transfer. In the NOESY-based methods additional peaks corresponding to atoms that are close in space but that do not belong to sequential residues will appear, confusing the assignment process. Following the initial sequential resonance assignment it is usually possible to extend the assignment from the C_α and C_β to the rest of the sidechain using experiments such as HCCH-TOCSY, which is basically a TOCSY experiment resolved in an additional carbon dimension.

Restraint Generation

In order to make structure calculations a number of experimentally determined restraints have to be generated. These fall into different categories; the most widely used are distance restraints and angle restraints.

Distance Restraints

A crosspeak in a NOESY experiment signifies spatial proximity between the two nuclei in question. Thus each peak can be converted into a maximum distance between the nuclei, usually between 1.8 and 6 angstroms. The intensity of a NOESY peak is proportional to the distance to the minus 6th power, so the distance is determined according to intensity of the peak. The intensity-distance relationship is not exact, so usually a distance range is used. It is of great importance to assign the NOESY peaks to the correct nuclei based on the chemical shifts. If this task is performed manually it is usually very labor-intensive, since, proteins usually have thousands of NOESY peaks. Some computer programs such as PASD/XPLOR-NIH, UNIO, CYANA, ARIA/CNS, and AUDANA/PONDEROSA-C/S in the Integrative NMR platform perform this task automatically on manually pre-processed listings of peak positions and peak volumes, coupled to a structure calculation. Direct access to the raw NOESY data without the cumbersome need of iteratively refined peak lists is so far only granted by the PASD algorithm

implemented in XPLOR-NIH, the ATNOS/CANDID approach implemented in the UNIO software package, and the PONDEROSA-C/S and thus indeed guarantees objective and efficient NOESY spectral analysis. To obtain as accurate assignments as possible it is a great advantage to have access to carbon-13 and nitrogen-15 NOESY experiments, since, they help to resolve overlap in the proton dimension. This leads to faster and more reliable assignments, and in turn to better structures.

Angle Restraints

In addition to distance restraints, restraints on the torsion angles of the chemical bonds, typically the psi and phi angles, can be generated. One approach is to use the Karplus equation, to generate angle restraints from coupling constants. Another approach uses the chemical shifts to generate angle restraints. Both methods use the fact that the geometry around the alpha carbon affects the coupling constants and chemical shifts, so given the coupling constants or the chemical shifts, a qualified guess can be made about the torsion angles.

Orientation Restraints

The analyte molecules in a sample can be partially ordered with respect to the external magnetic field of the spectrometer by manipulating the sample conditions. Common techniques include addition of bacteriophages or bicelles to the sample, or preparation of the sample in a stretched polyacrylamide gel. This creates a local environment that favours certain orientations of nonspherical molecules. Normally in solution NMR the dipolar couplings between nuclei are averaged out because of the fast tumbling of the molecule. The slight overpopulation of one orientation means that a residual dipolar coupling remains to be observed. The dipolar coupling is commonly used in solid state NMR and provides information about the relative orientation of the bond vectors relative to a single global reference frame. Typically the orientation of the N-H vector is probed in a HSQC like experiment. Initially, residual dipolar couplings were used for refinement of previously determined structures, but attempts at de novo structure determination have also been made.

Hydrogen–deuterium Exchange

NMR spectroscopy is nucleus specific. Thus it can distinguish between hydrogen and deuterium. The amide protons in the protein exchange readily with the solvent, and, if the solvent contains a different isotope, typically deuterium, the reaction can be monitored by NMR spectroscopy. How rapidly a given amide exchanges reflects its solvent accessibility. Thus amide exchange rates can give information on which parts of the protein are buried, hydrogen bonded etc.,. A common application is to compare the exchange of a free form versus a complex. The amides that become protected in the complex, are assumed to be in the interaction interface.

Structure Calculation

The experimentally determined restraints can be used as input for the structure calculation process. Researchers, using computer programs such as XPLOR-NIH, CYANA or GeNMR attempt to satisfy as many of the restraints as possible, in addition to general properties of proteins such as bond lengths and angles. The algorithms convert the restraints and the general protein properties into energy terms, and then try to minimize this energy. The process results in an ensemble of structures that, if the data were sufficient to dictate a certain fold, will converge.

Structure Validation

Is important to note that the ensemble of structures obtained is an “experimental model”, *i.e.*, a representation of certain kind of experimental data. To acknowledge this fact is really important because it means that the

model could be a good or bad representation of that experimental data. In general, the quality of a model will depend on both the quantity and quality of experimental data used to generate it and the correct interpretation of such data. It is important to remember that every experiment has associated errors. Random errors will affect the reproducibility and precision of the resulting structures. If the errors are systematic, the accuracy of the model will be affected. The precision indicates the degree of reproducibility of the measurement and is often expressed as the variance of the measured data set under the same conditions. The accuracy, however, indicates the degree to which a measurement approaches its “true” value.

Ideally, a model of a protein will be more accurate the more fit the actual molecule that represents and will be more precise as there is less uncertainty about the positions of their atoms. In practice there is no “standard molecule” against which to compare models of proteins, so the accuracy of a model is given by the degree of agreement between the model and a set of experimental data. Historically, the structures determined by NMR have been, in general, of lower quality than those determined by X-ray diffraction.

This is due, in part, to the lower amount of information contained in data obtained by NMR. Because of this fact, it has become common practice to establish the quality of NMR ensembles, by comparing it against the unique conformation determined by X-ray diffraction, for the same protein. However, the X-ray diffraction structure may not exist, and, since, the proteins in solution are flexible molecules, a protein represented by a single structure may lead to underestimate the intrinsic variation of the atomic positions of a protein. A set of conformations, determined by NMR or X-ray crystallography may be a better representation of the experimental data of a protein than a unique conformation.

The utility of a model will be given, at least in part, by the degree of accuracy and precision of the model. An accurate model with relatively poor precision could be useful to study the evolutionary relationships between the structures of a set of proteins, whereas the rational drug design requires both precise and accurate models. A model that is not accurate, regardless of the degree of precision with which it was obtained will not be very useful. Since, protein structures are experimental models that can contain errors, it is very important to be able to detect these errors. The process aimed at the detection of errors is known as validation. There are several methods to validate structures, some are statistical like PROCHECK and WHAT IF while others are based on physical principles as CheShift, or a mixture of statistical and physics principles PSVS.

Dynamics

In addition to structures, nuclear magnetic resonance can yield information on the dynamics of various parts of the protein. This usually involves measuring relaxation times such as T_1 and T_2 to determine order parameters, correlation times, and chemical exchange rates. NMR relaxation is a consequence of local fluctuating magnetic fields within a molecule. Local fluctuating magnetic fields are generated by molecular motions. In this way, measurements of relaxation times can provide information of motions within a molecule on the atomic level. In NMR studies of protein dynamics the nitrogen-15 isotope is the preferred nucleus to study because its relaxation times are relatively simple to relate to molecular motions. This, however, requires isotope labeling of the protein. The T_1 and T_2 relaxation times can be measured using various types of HSQC-based experiments. The types of motions that can be detected are motions that occur on a time-scale ranging from about 10 picoseconds to about 10 nanoseconds. In addition slower motions, which take place on a time-scale ranging from about 10 microseconds to 100 milliseconds, can also be studied.

However, since, nitrogen atoms are found mainly in the backbone of a protein, the results mainly reflect the motions of the backbone, which is the most rigid part of a protein molecule. Thus, the results obtained from nitrogen-15 relaxation measurements may not be representative for the whole protein. Therefore, techniques utilising relaxation measurements of carbon-13 and deuterium have recently been developed, which enables systematic studies of motions of the amino acid side-chains in proteins. A challenging and special case of study

regarding dynamics and flexibility of peptides and full-length proteins is represented by disordered structures. Nowadays, it is an accepted concept that proteins can exhibit a more flexible behaviour known as disorder or lack of structure; however, it is possible to describe an ensemble of structures instead of a static picture representing a fully functional state of the protein. Many advances are represented in this field in particular in terms of new pulse sequences, technological improvement, and rigorous training of researchers in the field.

NMR Spectroscopy on Large Proteins

Traditionally, nuclear magnetic resonance spectroscopy has been limited to relatively small proteins or protein domains. This is in part caused by problems resolving overlapping peaks in larger proteins, but this has been alleviated by the introduction of isotope labelling and multidimensional experiments. Another more serious problem is the fact that in large proteins the magnetization relaxes faster, which means there is less time to detect the signal. This in turn causes the peaks to become broader and weaker, and eventually disappear. Two techniques have been introduced to attenuate the relaxation: transverse relaxation optimized spectroscopy (TROSY) and deuteration of proteins. By using these techniques it has been possible to study proteins in complex with the 900 kDa chaperone GroES-GroEL.

Automation of the Process

Structure determination by NMR has traditionally been a time consuming process, requiring interactive analysis of the data by a highly trained scientist. There has been a considerable interest in automating the process to increase the throughput of structure determination and to make protein NMR accessible to non-experts. The two most time consuming processes involved are the sequence-specific resonance assignment (backbone and side-chain assignment) and the NOE assignment tasks. Several different computer programs have been published that target individual parts of the overall NMR structure determination process in an automated fashion. Most progress has been achieved for the task of automated NOE assignment. So far, only the FLYA and the UNIO approach were proposed to perform the entire protein NMR structure determination process in an automated manner without any human intervention. Alternatively, modules in the NMRFAM-SPARKY such as APES (two-letter-code: ae), PINE/PINE-SPARKY (two-letter-code: ep) and PONDEROSA (two-letter-code: c3, up) can be used to operate automation and visual verification in step-wise manner. Efforts have also been made to standardize the structure calculation protocol to make it quicker and more amenable to automation.

ELECTRON CRYSTALLOGRAPHY

Electron crystallography is a method to determine the arrangement of atoms in solids using a transmission electron microscope (TEM).

Comparison with X-ray Crystallography

It can complement X-ray crystallography for studies of very small crystals (<0.1 micrometers), both inorganic, organic, and proteins, such as membrane proteins, that cannot easily form the large 3-dimensional crystals required for that process. Protein structures are usually determined from either 2-dimensional crystals (sheets or helices), polyhedrons such as viral capsids, or dispersed individual proteins. Electrons can be used in these situations, whereas X-rays cannot, because electrons interact more strongly with atoms than X-rays do. Thus, X-rays will travel through a thin 2-dimensional crystal without diffracting significantly, whereas electrons can be used to form an image. Conversely, the strong interaction between electrons and protons makes thick (*e.g.*, 3-dimensional > 1 micrometer) crystals impervious to electrons, which only penetrate

short distances. One of the main difficulties in X-ray crystallography is determining phases in the diffraction pattern. Because of the complexity of X-ray lenses, it is difficult to form an image of the crystal being diffracted, and hence, phase information is lost. Fortunately, electron microscopes can resolve atomic structure in real space and the crystallographic structure factor phase information can be experimentally determined from an images Fourier transform. The Fourier transform of an atomic resolution image is similar, but different, to a diffraction pattern—with reciprocal lattice spots reflecting the symmetry and spacing of a crystal. Aaron Klug was the first to realize that the phase information could be read out directly from the Fourier transform of an electron microscopy image that had been scanned into a computer, already in 1968. For this, and his studies on virus structures and transfer-RNA, Klug received the Nobel Prize for chemistry in 1982.

Radiation Damage

A common problem to X-ray crystallography and electron crystallography is radiation damage, by which especially organic molecules and proteins are damaged as they are being imaged, limiting the resolution that can be obtained. This is especially troublesome in the setting of electron crystallography, where that radiation damage is focused on far fewer atoms. One technique used to limit radiation damage is electron cryomicroscopy, in which the samples undergo cryofixation and imaging takes place at liquid nitrogen or even liquid helium temperatures. Because of this problem, X-ray crystallography has been much more successful in determining the structure of proteins that are especially vulnerable to radiation damage.

Protein Structures Determined by Electron Crystallography

The first electron crystallographic protein structure to achieve atomic resolution was bacteriorhodopsin, determined by Richard Henderson and coworkers at the Medical Research Council Laboratory of Molecular Biology in 1990. However, already in 1975 Unwin and Henderson had determined the first membrane protein structure at intermediate resolution (7 Ångström), showing for the first time the internal structure of a membrane protein, with its alpha-helices standing perpendicular to the plane of the membrane. Since, then, several other high-resolution structures have been determined by electron crystallography, including the light-harvesting complex, the nicotinic acetylcholine receptor, and the bacterial flagellum.

Application to Inorganic Materials

Electron crystallographic studies on inorganic crystals using high-resolution electron microscopy (HREM) images were first performed by Aaron Klug in 1978 and by Sven Hovmöller and coworkers in 1984. HREM images were used because they allow to select (by computer software) only the very thin regions close to the edge of the crystal for structure analysis. This is of crucial importance since, in the thicker parts of the crystal the exit-wave function (which carries the information about the intensity and position of the projected atom columns) is no longer linearly related to the projected crystal structure. Moreover, not only do the HREM images change their appearance with increasing crystal thickness, they are also very sensitive to the chosen setting of the defocus Δf of the objective lens. To cope with this complexity Michael O'Keefe started in the early 1970s to develop image simulation software which allowed to understand and interpret the observed contrast changes in HREM images.

There was a serious disagreement in the field of electron microscopy of inorganic compounds; while some have claimed that “the phase information is present in EM images” others have the opposite view that “the phase information is lost in EM images”. The reason for these opposite views is that the word “phase” has been used with different meanings in the two communities of physicists and crystallographers. The physicists are more concerned about the “electron wave phase”—the phase of a wave moving through the sample during exposure

by the electrons. This wave has a wavelength of about 0.02–0.03 Ångström (depending on the accelerating voltage of the electron microscope). Its phase is related to the phase of the undiffracted direct electron beam. The crystallographers, on the other hand, mean the “crystallographic structure factor phase” when they simply say “phase”. This phase is the phase of standing waves of potential in the crystal (very similar to the electron density measured in X-ray crystallography).

Each of these waves have their specific wavelength, called d-value for distance between so-called Bragg planes of low/high potential. These d-values range from the unit cell dimensions to the resolution limit of the electron microscope, *i.e.*, typically from 10 or 20 Ångströms down to 1 or 2 Ångströms. Their phases are related to a fixed point in the crystal, defined in relation to the symmetry elements of that crystal. The crystallographic phases are a property of the crystal, so they exist also outside the electron microscope. The electron waves vanish if the microscope is switched off. In order to determine a crystal structure, it is necessary to know the crystallographic structure factors, but not to know the electron wave phases. A more detailed discussion how (crystallographic structure factor) phases link with the phases of the electron wave can be found in.

Just as with proteins, it has been possible to determine the atomic structures of inorganic crystals by electron crystallography. For simpler structure it is sufficient to use three perpendicular views, but for more complicated structures, also projections down ten or more different diagonals may be needed.

In addition to electron microscopy images, it is also possible to use electron diffraction (ED) patterns for crystal structure determination. The utmost care must be taken to record such ED patterns from the thinnest areas in order to keep most of the structure related intensity differences between the reflections (quasi-kinematical diffraction conditions). Just as with X-ray diffraction patterns, the important crystallographic structure factor phases are lost in electron diffraction patterns and must be uncovered by special crystallographic methods such as direct methods, maximum likelihood or (more recently) by the charge-flipping method. On the other hand, ED patterns of inorganic crystals have often a high resolution (= interplanar spacings with high Miller indices) much below 1 Ångström. This is comparable to the point resolution of the best electron microscopes. Under favourable conditions it is possible to use ED patterns from a single orientation to determine the complete crystal structure. Alternatively a hybrid approach can be used which uses HRTEM images for solving and intensities from ED for refining the crystal structure. Recent progress for structure analysis by ED was made by introducing the Vincent-Midgley precession technique for recording electron diffraction patterns. The thereby obtained intensities are usually much closer to the kinematical intensities, so that even structures can be determined that are out of range when processing conventional (selected area) electron diffraction data.

Crystal structures determined via electron crystallography can be checked for their quality by using first-principles calculations within density functional theory (DFT). This approach was for the first time applied for the validation of several metal-rich structures which were only accessible by HRTEM and ED, respectively. Recently, two very complicated zeolite structures have been determined by electron crystallography combined with X-ray powder diffraction. These are more complex than the most complex zeolite structures determined by X-ray crystallography.

ELECTRON PARAMAGNETIC RESONANCE

Electron paramagnetic resonance (EPR) or electron spin resonance (ESR) spectroscopy is a method for studying materials with unpaired electrons. The basic concepts of EPR are analogous to those of nuclear magnetic resonance (NMR), but it is electron spins that are excited instead of the spins of atomic nuclei. EPR spectroscopy is particularly useful for studying metal complexes or organic radicals. EPR was first observed in Kazan State University by Soviet physicist Yevgeny Zavoisky in 1944, and was developed independently at the same time by Brebis Bleaney at the University of Oxford.

Applications

EPR/ESR spectroscopy is used in various branches of science, such as biology, chemistry and physics, for the detection and identification of free radicals and paramagnetic centers such as F-centers. EPR is a sensitive, specific method for studying both radicals formed in chemical reactions and the reactions themselves. For example, when ice (solid H_2O) is decomposed by exposure to high-energy radiation, radicals such as H, OH, and HO_2 are produced. Such radicals can be identified and studied by EPR. Organic and inorganic radicals can be detected in electrochemical systems and in materials exposed to UV light. In many cases, the reactions to make the radicals and the subsequent reactions of the radicals are of interest, while in other cases EPR is used to provide information on a radical's geometry and the orbital of the unpaired electron. EPR/ESR spectroscopy is also used in geology and archaeology as a dating tool. It can be applied to a wide range of materials such as carbonates, sulfates, phosphates, silica or other silicates.

Medical and biological applications of EPR also exist. Although radicals are very reactive, and so do not normally occur in high concentrations in biology, special reagents have been developed to spin-label molecules of interest. These reagents are particularly useful in biological systems. Specially-designed nonreactive radical molecules can attach to specific sites in a biological cell, and EPR spectra can then give information on the environment of these so-called spin labels or spin probes. Spin-labeled fatty acids have been extensively used to study dynamic organisation of lipids in biological membranes, lipid-protein interactions and temperature of transition of gel to liquid crystalline phases. A type of dosimetry system has been designed for reference standards and routine use in medicine, based on EPR signals of radicals from irradiated polycrystalline α -alanine (the alanine deamination radical, the hydrogen abstraction radical, and the $(\text{CO}^\bullet(\text{OH}))=\text{C}(\text{CH}_3)\text{NH}_2^+$ radical). This method is suitable for measuring gamma and x-rays, electrons, protons, and high-linear energy transfer (LET) radiation of doses in the 1 Gy to 100 kGy range.

EPR/ESR spectroscopy can be applied only to systems in which the balance between radical decay and radical formation keeps the free radicals concentration above the detection limit of the spectrometer used. This can be a particularly severe problem in studying reactions in liquids. An alternative approach is to slow down reactions by studying samples held at cryogenic temperatures, such as 77 K (liquid nitrogen) or 4.2 K (liquid helium). An example of this work is the study of radical reactions in single crystals of amino acids exposed to x-rays, work that sometimes leads to activation energies and rate constants for radical reactions.

The study of radiation-induced free radicals in biological substances (for cancer research) poses the additional problem that tissue contains water, and water (due to its electric dipole moment) has a strong absorption band in the microwave region used in EPR spectrometers. EPR/ESR also has been used by archaeologists for the dating of teeth. Radiation damage over long periods of time creates free radicals in tooth enamel, which can then be examined by EPR and, after proper calibration, dated. Alternatively, material extracted from the teeth of people during dental procedures can be used to quantify their cumulative exposure to ionizing radiation. People exposed to radiation from the Chernobyl disaster have been examined by this method.

Radiation-sterilized foods have been examined with EPR spectroscopy, the aim being to develop methods to determine whether a particular food sample has been irradiated and to what dose. EPR/ESR spectroscopy has been used to measure properties of crude oil, in particular asphaltene and vanadium content. EPR measurement of asphaltene content is a function of spin density and solvent polarity. Prior work dating to the 1960s has demonstrated the ability to measure vanadium content to sub-ppm levels. In the field of quantum computing, pulsed EPR is used to control the state of electron spin qubits in materials such as diamond, silicon and gallium arsenide.

Thermodynamics

Thermodynamics is a branch of physics concerned with heat and temperature and their relation to other forms of energy and work. The behaviour of these quantities is governed by the four laws of thermodynamics, irrespective of the composition or specific properties of the material or system in question. The laws of thermodynamics are explained in terms of microscopic constituents by statistical mechanics. Thermodynamics applies to a wide variety of topics in science and engineering, especially physical chemistry, chemical engineering and mechanical engineering. Historically, thermodynamics developed out of a desire to increase the efficiency of early steam engines, particularly through the work of French physicist Nicolas Léonard Sadi Carnot (1824) who believed that engine efficiency was the key that could help France win the Napoleonic Wars. Scottish physicist Lord Kelvin was the first to formulate a concise definition of thermodynamics in 1854 which stated, “*Thermo-dynamics is the subject of the relation of heat to forces acting between contiguous parts of bodies, and the relation of heat to electrical agency.*”

The initial application of thermodynamics to mechanical heat engines was extended early on to the study of chemical compounds and chemical reactions. Chemical thermodynamics studies the nature of the role of entropy in the process of chemical reactions and has provided the bulk of expansion and knowledge of the field. Other formulations of thermodynamics emerged in the following decades. Statistical thermodynamics, or statistical mechanics, concerned itself with statistical predictions of the collective motion of particles from their microscopic behaviour. In 1909, Constantin Carathéodory presented a purely mathematical approach to the field in his axiomatic formulation of thermodynamics, a description often referred to as *geometrical thermodynamics*. A description of any thermodynamic system employs the four laws of thermodynamics that form an axiomatic basis. The first law specifies that energy can be exchanged between physical systems as heat and work. The second law defines the existence of a quantity called entropy, that describes the direction, thermodynamically, that a system can evolve and quantifies the state of order of a system and that can be used to quantify the useful work that can be extracted from the system.

In thermodynamics, interactions between large ensembles of objects are studied and categorized. Central to this are the concepts of the thermodynamic *system* and its *surroundings*. A system is composed of particles, whose average motions define its properties, and those properties are in turn related to one another through equations of state. Properties can be combined to express internal energy and thermodynamic potentials, which are useful for determining conditions for equilibrium and spontaneous processes. With these tools, thermodynamics can be used to describe how systems respond to changes in their environment. This can be applied to a wide variety of topics in science and engineering, such as engines, phase transitions, chemical reactions, transport phenomena, and even black holes. The results of thermodynamics are essential for other fields of physics and for chemistry, chemical engineering, aerospace engineering, mechanical engineering, cell biology, biomedical engineering, materials science, and economics, to name a few. This article is focused mainly on classical thermodynamics which primarily studies systems in thermodynamic equilibrium. Non-

equilibrium thermodynamics is often treated as an extension of the classical treatment, but statistical mechanics has brought many advances to that field.

HISTORY

The history of thermodynamics as a scientific discipline generally begins with Otto von Guericke who, in 1650, built and designed the world's first vacuum pump and demonstrated a vacuum using his Magdeburg hemispheres. Guericke was driven to make a vacuum in order to disprove Aristotle's long-held supposition that 'nature abhors a vacuum'. Shortly after Guericke, the English physicist and chemist Robert Boyle had learned of Guericke's designs and, in 1656, in coordination with English scientist Robert Hooke, built an air pump.

Using this pump, Boyle and Hooke noticed a correlation between pressure, temperature, and volume. In time, Boyle's Law was formulated, which states that pressure and volume are inversely proportional. Then, in 1679, based on these concepts, an associate of Boyle's named Denis Papin built a steam digester, which was a closed vessel with a tightly fitting lid that confined steam until a high pressure was generated. Later designs implemented a steam release valve that kept the machine from exploding. By watching the valve rhythmically move up and down, Papin conceived of the idea of a piston and a cylinder engine. He did not, however, follow through with his design. Nevertheless, in 1697, based on Papin's designs, engineer Thomas Savery built the first engine, followed by Thomas Newcomen in 1712. Although these early engines were crude and inefficient, they attracted the attention of the leading scientists of the time.

The fundamental concepts of heat capacity and latent heat, which were necessary for the development of thermodynamics, were developed by Professor Joseph Black at the University of Glasgow, where James Watt was employed as an instrument maker. Black and Watt performed experiments together, but it was Watt who conceived the idea of the external condenser which resulted in a large increase in steam engine efficiency. Drawing on all the previous work led Sadi Carnot, the "father of thermodynamics", to publish *Reflections on the Motive Power of Fire* (1824), a discourse on heat, power, energy and engine efficiency. The book outlined the basic energetic relations between the Carnot engine, the Carnot cycle, and motive power. It marked the start of thermodynamics as a modern science. The first thermodynamic textbook was written in 1859 by William Rankine, originally trained as a physicist and a civil and mechanical engineering professor at the University of Glasgow. The first and second laws of thermodynamics emerged simultaneously in the 1850s, primarily out of the works of William Rankine, Rudolf Clausius, and William Thomson (Lord Kelvin). The foundations of statistical thermodynamics were set out by physicists such as James Clerk Maxwell, Ludwig Boltzmann, Max Planck, Rudolf Clausius and J. Willard Gibbs. During the years 1873-76 the American mathematical physicist Josiah Willard Gibbs published a series of three papers, the most famous being *On the Equilibrium of Heterogeneous Substances*, in which he showed how thermodynamic processes, including chemical reactions, could be graphically analyzed, by studying the energy, entropy, volume, temperature and pressure of the thermodynamic system in such a manner, one can determine if a process would occur spontaneously. Also Pierre Duhem in the 19th century wrote about chemical thermodynamics. During the early 20th century, chemists such as Gilbert N. Lewis, Merle Randall, and E. A. Guggenheim applied the mathematical methods of Gibbs to the analysis of chemical processes.

ETYMOLOGY

The etymology of *thermodynamics* has an intricate history. It was first spelled in a hyphenated form as an adjective (*thermo-dynamic*) and from 1854 to 1868 as the noun *thermo-dynamics* to represent the science of generalized heat engines. American biophysicist Donald Haynie claims that *thermodynamics* was coined in 1840 from the Greek *therme*, meaning heat and *dynamis*, meaning power. However, this etymology has been cited as unlikely. Pierre Perrot claims that the term *thermodynamics* was coined by James Joule in 1858 to designate the science of relations between heat and power, however, Joule never used that term, but used instead the term *perfect*

thermo-dynamic engine in reference to Thomson's 1849 phraseology. By 1858, *thermo-dynamics*, as a functional term, was used in William Thomson's paper "An Account of Carnot's Theory of the Motive Power of Heat."

BRANCHES OF THERMODYNAMICS

The study of thermodynamical systems has developed into several related branches, each using a different fundamental model as a theoretical or experimental basis, or applying the principles to varying types of systems.

CLASSICAL THERMODYNAMICS

Classical thermodynamics is the description of the states of thermodynamic systems at near-equilibrium, that uses macroscopic, measurable properties. It is used to model exchanges of energy, work and heat based on the laws of thermodynamics. The qualifier *classical* reflects the fact that it represents the first level of understanding of the subject as it developed in the 19th century and describes the changes of a system in terms of macroscopic empirical (large scale, and measurable) parameters. A microscopic interpretation of these concepts was later provided by the development of *statistical mechanics*.

STATISTICAL MECHANICS

Statistical mechanics, also called statistical thermodynamics, emerged with the development of atomic and molecular theories in the late 19th century and early 20th century, and supplemented classical thermodynamics with an interpretation of the microscopic interactions between individual particles or quantum-mechanical states. This field relates the microscopic properties of individual atoms and molecules to the macroscopic, bulk properties of materials that can be observed on the human scale, thereby explaining classical thermodynamics as a natural result of statistics, classical mechanics, and quantum theory at the microscopic level.

CHEMICAL THERMODYNAMICS

Chemical thermodynamics is the study of the interrelation of energy with chemical reactions or with a physical change of state within the confines of the laws of thermodynamics.

TREATMENT OF EQUILIBRIUM

Equilibrium thermodynamics is the systematic study of transformations of matter and energy in systems as they approach equilibrium. The word equilibrium implies a state of balance. In an equilibrium state there are no unbalanced potentials, or driving forces, within the system. A central aim in equilibrium thermodynamics is: given a system in a well-defined initial state, subject to accurately specified constraints, to calculate what the state of the system will be once it has reached equilibrium. Non-equilibrium thermodynamics is a branch of thermodynamics that deals with systems that are not in thermodynamic equilibrium. Most systems found in nature are not in thermodynamic equilibrium because they are not in stationary states, and are continuously and discontinuously subject to flux of matter and energy to and from other systems. The thermodynamic study of non-equilibrium systems requires more general concepts than are dealt with by equilibrium thermodynamics. Many natural systems still today remain beyond the scope of currently known macroscopic thermodynamic methods.

LAWS OF THERMODYNAMICS

The four laws of thermodynamics define fundamental physical quantities (temperature, energy, and entropy) that characterize thermodynamic systems at thermal equilibrium. The laws describe how these quantities behave under various circumstances, and forbid certain phenomena (such as perpetual motion).

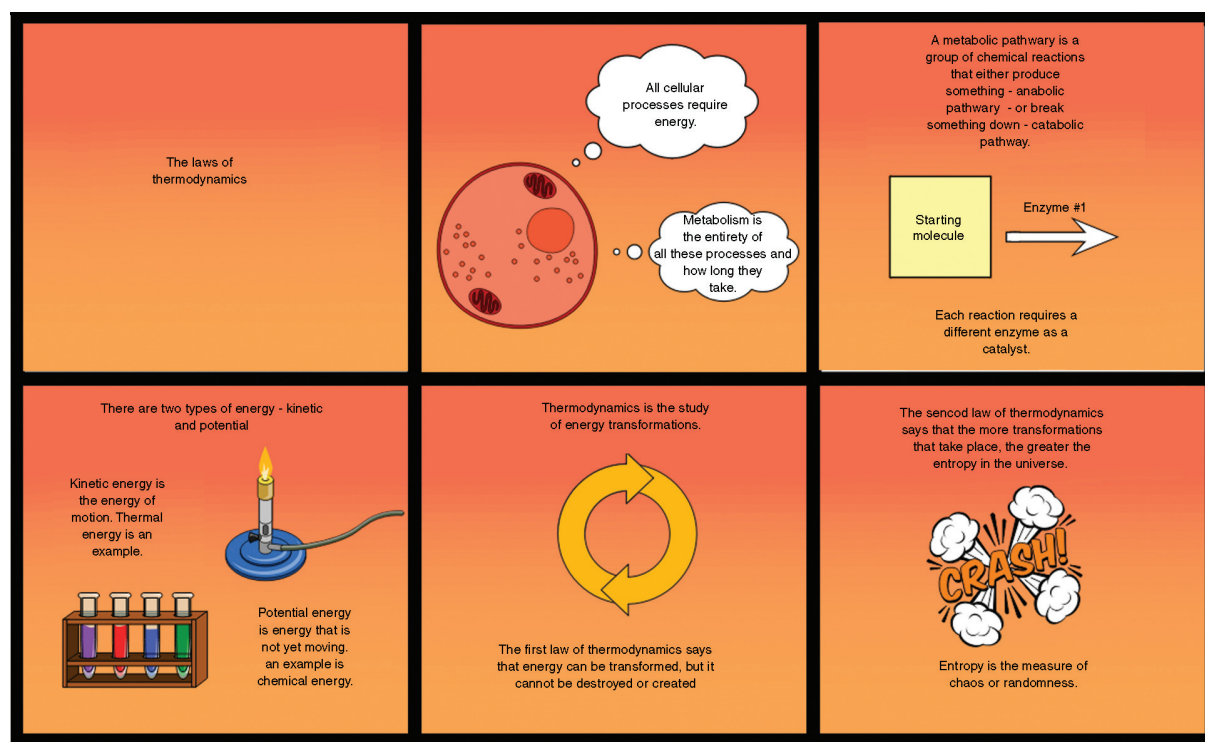


Fig. Laws of thermodynamics.

The four laws of thermodynamics are:

- **Zeroth law of thermodynamics:** If two systems are in thermal equilibrium with a third system, they are in thermal equilibrium with each other. This law helps define the concept of temperature.
- **First law of thermodynamics:** When energy passes, as work, as heat, or with matter, into or out from a system, the system's internal energy changes in accord with the law of conservation of energy. Equivalently, perpetual motion machines of the first kind (machines that produce work with no energy input) are impossible.
- **Second law of thermodynamics:** In a natural thermodynamic process, the sum of the entropies of the interacting thermodynamic systems increases. Equivalently, perpetual motion machines of the second kind (machines that spontaneously convert thermal energy into mechanical work) are impossible.
- **Third law of thermodynamics:** The entropy of a system approaches a constant value as the temperature approaches absolute zero. With the exception of non-crystalline solids (glasses) the entropy of a system at absolute zero is typically close to zero, and is equal to the natural logarithm of the product of the quantum ground states.

There have been suggestions of additional laws, but none of them achieves the generality of the four accepted laws, and they are not mentioned in standard textbooks. The laws of thermodynamics are important fundamental laws in physics and they are applicable in other natural sciences.

ZEROth LAW

The zeroth law of thermodynamics states that if two thermodynamic systems are each in thermal equilibrium with a third, then they are in thermal equilibrium with each other. Accordingly, thermal equilibrium between systems is a transitive relation. Two systems are said to be in the relation of thermal equilibrium if they are linked by a wall permeable only to heat and they do not change over time. As a convenience of language, systems are sometimes also said to be in a relation of thermal equilibrium if they are not linked so as to be able to transfer heat to each other, but would not do so if they were connected by a wall permeable only to heat. The

physical meaning is expressed by Maxwell in the words: “All heat is of the same kind”. Another statement of the law is “All diathermal walls are equivalent”. The law is important for the mathematical formulation of thermodynamics, which needs the assertion that the relation of thermal equilibrium is an equivalence relation. This information is needed for a mathematical definition of temperature that will agree with the physical existence of valid thermometers.

Zeroth Law as Equivalence Relation

A thermodynamic system is by definition in its own state of internal thermodynamic equilibrium, that is to say, there is no change in its observable state (*i.e.*, macrostate) over time and no flows occur in it. One precise statement of the zeroth law is that the relation of thermal equilibrium is an equivalence relation on pairs of thermodynamic systems. In other words, the set of all systems each in its own state of internal thermodynamic equilibrium may be divided into subsets in which every system belongs to one and only one subset, and is in thermal equilibrium with every other member of that subset, and is not in thermal equilibrium with a member of any other subset. This means that a unique “tag” can be assigned to every system, and if the “tags” of two systems are the same, they are in thermal equilibrium with each other, and if different, they are not. This property is used to justify the use of empirical temperature as a tagging system. Empirical temperature provides further relations of thermally equilibrated systems, such as order and continuity with regard to “hotness” or “coldness”, but these are not implied by the standard statement of the zeroth law.

*If it is defined that a thermodynamic system is in thermal equilibrium with itself (*i.e.*, thermal equilibrium is reflexive), then the zeroth law may be stated as follows:*

- If a body *C*, be in thermal equilibrium with two other bodies, *A* and *B*, then *A* and *B* are in thermal equilibrium with one another.

This statement asserts that thermal equilibrium is a left-Euclidean relation between thermodynamic systems. If we also define that every thermodynamic system is in thermal equilibrium with itself, then thermal equilibrium is also a reflexive relation. Binary relations that are both reflexive and Euclidean are equivalence relations.

Thus, again implicitly assuming reflexivity, the zeroth law is therefore often expressed as a right-Euclidean statement:

- If two systems are in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.

One consequence of an equivalence relationship is that the equilibrium relationship is symmetric: If *A* is in thermal equilibrium with *B*, then *B* is in thermal equilibrium with *A*. Thus we may say that two systems are in thermal equilibrium with each other, or that they are in mutual equilibrium.

Another consequence of equivalence is that thermal equilibrium is a transitive relationship and is occasionally expressed as such:

- If *A* is in thermal equilibrium with *B* and if *B* is in thermal equilibrium with *C*, then *A* is in thermal equilibrium with *C*.

A reflexive, transitive relationship does not guarantee an equivalence relationship. In order for the above statement to be true, *both* reflexivity *and* symmetry must be implicitly assumed.

It is the Euclidean relationships which apply directly to thermometry. An ideal thermometer is a thermometer which does not measurably change the state of the system it is measuring. Assuming that the unchanging reading of an ideal thermometer is a valid “tagging” system for the equivalence classes of a set of equilibrated thermodynamic systems, then if a thermometer gives the same reading for two systems, those two systems are in thermal equilibrium, and if we thermally connect the two systems, there will be no subsequent change in the state of either one. If the readings are different, then thermally connecting the two systems will cause a change in the states of both systems and when the change is complete, they will both yield the same thermometer reading. The zeroth law provides no information regarding this final reading.

Foundation of Temperature

The zeroth law establishes thermal equilibrium as an equivalence relationship. An equivalence relationship on a set (such as the set of all systems each in its own state of internal thermodynamic equilibrium) divides that set into a collection of distinct subsets (“disjoint subsets”) where any member of the set is a member of one and only one such subset. In the case of the zeroth law, these subsets consist of systems which are in mutual equilibrium. This partitioning allows any member of the subset to be uniquely “tagged” with a label identifying the subset to which it belongs. Although the labeling may be quite arbitrary, temperature is just such a labeling process which uses the real number system for tagging. The zeroth law justifies the use of suitable thermodynamic systems as thermometers to provide such a labeling, which yield any number of possible empirical temperature scales, and justifies the use of the second law of thermodynamics to provide an absolute, or thermodynamic temperature scale. Such temperature scales bring additional continuity and ordering (*i.e.*, “hot” and “cold”) properties to the concept of temperature.

In the space of thermodynamic parameters, zones of constant temperature form a surface, that provides a natural order of nearby surfaces. One may therefore construct a global temperature function that provides a continuous ordering of states. The dimensionality of a surface of constant temperature is one less than the number of thermodynamic parameters, thus, for an ideal gas described with three thermodynamic parameters P , V and N , it is a two-dimensional surface. For example, if two systems of ideal gases are in equilibrium, then $P_1 V_1 / N_1 = P_2 V_2 / N_2$ where P_i is the pressure in the i th system, V_i is the volume, and N_i is the amount (in moles, or simply the number of atoms) of gas. The surface $PV/N = \text{constant}$ defines surfaces of equal thermodynamic temperature, and one may label defining T so that $PV/N = RT$, where R is some constant. These systems can now be used as a thermometer to calibrate other systems. Such systems are known as “ideal gas thermometers”.

In a sense, focused on in the zeroth law, there is only one kind of diathermal wall or one kind of heat, as expressed by Maxwell’s dictum that “All heat is of the same kind”. But in another sense, heat is transferred in different ranks, as expressed by Sommerfeld’s dictum “Thermodynamics investigates the conditions that govern the transformation of heat into work. It teaches us to recognize temperature as the measure of the work-value of heat. Heat of higher temperature is richer, is capable of doing more work. Work may be regarded as heat of an infinitely high temperature, as unconditionally available heat.” This is why temperature is the particular variable indicated by the zeroth law’s statement of equivalence.

Physical Meaning of the Usual Statement of the Zeroth Law

The present article states the zeroth law as it is often summarized in textbooks. Nevertheless, this usual statement perhaps does not explicitly convey the full physical meaning that underlies it. The underlying physical meaning was perhaps first clarified by Maxwell in his 1871 textbook.

In Carathéodory’s (1909) theory, it is postulated that there exist walls “permeable only to heat”, though heat is not explicitly defined in that paper. This postulate is a physical postulate of existence. It does not, however, as worded just previously, say that there is only one kind of heat. This paper of Carathéodory states as proviso 4 of its account of such walls: “Whenever each of the systems S_1 and S_2 is made to reach equilibrium with a third system S_3 under identical conditions, systems S_1 and S_2 are in mutual equilibrium”. It is the function of this statement in the paper, not there labeled as the zeroth law, to provide not only for the existence of transfer of energy other than by work or transfer of matter, but further to provide that such transfer is unique in the sense that there is only one kind of such wall, and one kind of such transfer. This is signaled in the postulate of this paper of Carathéodory that precisely one non-deformation variable is needed to complete the specification of a thermodynamic state, beyond the necessary deformation variables, which are not restricted in number. It is therefore not exactly clear what Carathéodory means when in the introduction of this paper he writes “*It is*

possible to develop the whole theory without assuming the existence of heat, that is of a quantity that is of a different nature from the normal mechanical quantities.”

Maxwell (1871) discusses at some length ideas which he summarizes by the words “All heat is of the same kind”. Modern theorists sometimes express this idea by postulating the existence of a unique one-dimensional *hotness manifold*, into which every proper temperature scale has a monotonic mapping. This may be expressed by the statement that there is only one kind of temperature, regardless of the variety of scales in which it is expressed. Another modern expression of this idea is that “All diathermal walls are equivalent”. This might also be expressed by saying that there is precisely one kind of non-mechanical, non-matter-transferring contact equilibrium between thermodynamic systems.

These ideas may be regarded as helping to clarify the physical meaning of the usual statement of the zeroth law of thermodynamics. It is the opinion of Lieb and Yngvason (1999) that the derivation from statistical mechanics of the law of entropy increase is a goal that has so far eluded the deepest thinkers. Thus the idea remains open to consideration that the existence of heat and temperature are needed as coherent primitive concepts for thermodynamics, as expressed, for example, by Maxwell and Planck. On the other hand, Planck in 1926 clarified how the second law can be stated without reference to heat or temperature, by referring to the irreversible and universal nature of friction in natural thermodynamic processes.

History

According to Arnold Sommerfeld, Ralph H. Fowler invented the title ‘the zeroth law of thermodynamics’ when he was discussing the 1935 text of Saha and Srivastava. They write on page 1 that “every physical quantity must be measurable in numerical terms”. They presume that temperature is a physical quantity and then deduce the statement “If a body A is in temperature equilibrium with two bodies B and C, then B and C themselves will be in temperature equilibrium with each other”. They then in a self-standing paragraph italicize as if to state their basic postulate: “*Any of the physical properties of A which change with the application of heat may be observed and utilised for the measurement of temperature.*” They do not themselves here use the term ‘zeroth law of thermodynamics’. There are very many statements of these physical ideas in the physics literature long before this text, in very similar language. What was new here was just the label ‘zeroth law of thermodynamics’. Fowler, with co-author Edward A.

Guggenheim, wrote of the zeroth law as follows:

- ...we introduce the postulate: If two assemblies are each in thermal equilibrium with a third assembly, they are in thermal equilibrium with each other.

They then proposed that “it may be shown to follow that the condition for thermal equilibrium between several assemblies is the equality of a certain single-valued function of the thermodynamic states of the assemblies, which may be called the temperature t , any one of the assemblies being used as a “thermometer” reading the temperature t on a suitable scale. This postulate of the “*Existence of temperature*” could with advantage be known as *the zeroth law of thermodynamics*”. The first sentence of this present article is a version of this statement. It is not explicitly evident in the existence statement of Fowler and Guggenheim that temperature refers to a unique attribute of a state of a system, such as is expressed in the idea of the hotness manifold. Also their statement refers explicitly to statistical mechanical assemblies, not explicitly to macroscopic thermodynamically defined systems.

FIRST LAW OF THERMODYNAMICS

The first law of thermodynamics is a version of the law of conservation of energy, adapted for thermodynamic systems. The law of conservation of energy states that the total energy of an isolated system is constant; energy can be transformed from one form to another, but can be neither created nor destroyed.

The first law is often formulated,

$$\Delta U = Q - W.$$

It states that the change in the internal energy ΔU of a closed system is equal to the amount of heat Q supplied to the system, minus the amount of work W done by the system on its surroundings. An equivalent statement is that perpetual motion machines of the first kind are impossible.

History

Investigations into the nature of heat and work and their relationship began with the invention of the first engines used to extract water from mines. Improvements to such engines so as to increase their efficiency and power output came first from mechanics that worked with such machines but only slowly advanced the art. Deeper investigations that placed those on a mathematical and physics basis came later. The first law of thermodynamics was developed empirically over about half a century. The first full statements of the law came in 1850 from Rudolf Clausius and from William Rankine; Rankine's statement is considered less distinct relative to Clausius'. A main aspect of the struggle was to deal with the previously proposed caloric theory of heat.

In 1840, Germain Hess stated a conservation law for the so-called 'heat of reaction' for chemical reactions. His law was later recognized as a consequence of the first law of thermodynamics, but Hess's statement was not explicitly concerned with the relation between energy exchanges by heat and work. According to Truesdell (1980), Julius Robert von Mayer in 1841 made a statement that meant that "in a process at constant pressure, the heat used to produce expansion is universally interconvertible with work", but this is not a general statement of the first law.

Original Statements: The "thermodynamic Approach"

The original nineteenth century statements of the first law of thermodynamics appeared in a conceptual framework in which transfer of energy as heat was taken as a primitive notion, not defined or constructed by the theoretical development of the framework, but rather presupposed as prior to it and already accepted. The primitive notion of heat was taken as empirically established, especially through calorimetry regarded as a subject in its own right, prior to thermodynamics. Jointly primitive with this notion of heat were the notions of empirical temperature and thermal equilibrium. This framework also took as primitive the notion of transfer of energy as work. This framework did not presume a concept of energy in general, but regarded it as derived or synthesized from the prior notions of heat and work. By one author, this framework has been called the "thermodynamic" approach.

The first explicit statement of the first law of thermodynamics, by Rudolf Clausius in 1850, referred to cyclic thermodynamic processes.

- In all cases in which work is produced by the agency of heat, a quantity of heat is consumed which is proportional to the work done; and conversely, by the expenditure of an equal quantity of work an equal quantity of heat is produced.

Clausius also stated the law in another form, referring to the existence of a function of state of the system, the internal energy, and expressed it in terms of a differential equation for the increments of a thermodynamic process.

This equation may be described as follows:

- In a thermodynamic process involving a closed system, the increment in the internal energy is equal to the difference between the heat accumulated by the system and the work done by it.

Because of its definition in terms of increments, the value of the internal energy of a system is not uniquely defined. It is defined only up to an arbitrary additive constant of integration, which can be adjusted to give

arbitrary reference zero levels. This non-uniqueness is in keeping with the abstract mathematical nature of the internal energy. The internal energy is customarily stated relative to a conventionally chosen standard reference state of the system. The concept of internal energy is considered by Bailyn to be of “enormous interest”. Its quantity cannot be immediately measured, but can only be inferred, by differencing actual immediate measurements. Bailyn likens it to the energy states of an atom, that were revealed by Bohr’s energy relation $h\nu = E_{n'} - E_n$. In each case, an unmeasurable quantity (the internal energy, the atomic energy level) is revealed by considering the difference of measured quantities (increments of internal energy, quantities of emitted or absorbed radiative energy).

Conceptual Revision: The “mechanical Approach”

In 1907, George H. Bryan wrote about systems between which there is no transfer of matter (closed systems): “Definition. When energy flows from one system or part of a system to another otherwise than by the performance of mechanical work, the energy so transferred is called *heat*.” This definition may be regarded as expressing a conceptual revision, as follows. This was systematically expounded in 1909 by Constantin Carathéodory, whose attention had been drawn to it by Max Born. Largely through Born’s influence, this revised conceptual approach to the definition of heat came to be preferred by many twentieth-century writers. It might be called the “mechanical approach”.

Energy can also be transferred from one thermodynamic system to another in association with transfer of matter. Born points out that in general such energy transfer is not resolvable uniquely into work and heat moieties. In general, when there is transfer of energy associated with matter transfer, work and heat transfers can be distinguished only when they pass through walls physically separate from those for matter transfer.

The “mechanical” approach postulates the law of conservation of energy. It also postulates that energy can be transferred from one thermodynamic system to another adiabatically as work, and that energy can be held as the internal energy of a thermodynamic system. It also postulates that energy can be transferred from one thermodynamic system to another by a path that is non-adiabatic, and is unaccompanied by matter transfer. Initially, it “cleverly” (according to Bailyn) refrains from labelling as ‘heat’ such non-adiabatic, unaccompanied transfer of energy. It rests on the primitive notion of *walls*, especially adiabatic walls and non-adiabatic walls, defined as follows. Temporarily, only for purpose of this definition, one can prohibit transfer of energy as work across a wall of interest.

Then walls of interest fall into two classes:

- Those such that arbitrary systems separated by them remain independently in their own previously established respective states of internal thermodynamic equilibrium; they are defined as adiabatic; and
- Those without such independence; they are defined as non-adiabatic.

This approach derives the notions of transfer of energy as heat, and of temperature, as theoretical developments, not taking them as primitives. It regards calorimetry as a derived theory. It has an early origin in the nineteenth century, for example in the work of Helmholtz, but also in the work of many others.

Conceptually Revised Statement, According to the Mechanical Approach

The revised statement of the first law postulates that a change in the internal energy of a system due to any arbitrary process, that takes the system from a given initial thermodynamic state to a given final equilibrium thermodynamic state, can be determined through the physical existence, for those given states, of a reference process that occurs purely through stages of adiabatic work.

The revised statement is then,

- For a closed system, in any arbitrary process of interest that takes it from an initial to a final state of internal thermodynamic equilibrium, the change of internal energy is the same as that for a reference adiabatic work process that links those two states. This is so regardless of the path of the process of interest, and regardless of whether it is an adiabatic or a non-adiabatic process. The reference adiabatic work process may be chosen arbitrarily from amongst the class of all such processes.

This statement is much less close to the empirical basis than are the original statements, but is often regarded as conceptually parsimonious in that it rests only on the concepts of adiabatic work and of non-adiabatic processes, not on the concepts of transfer of energy as heat and of empirical temperature that are presupposed by the original statements. Largely through the influence of Max Born, it is often regarded as theoretically preferable because of this conceptual parsimony. Born particularly observes that the revised approach avoids thinking in terms of what he calls the “imported engineering” concept of heat engines.

Basing his thinking on the mechanical approach, Born in 1921, and again in 1949, proposed to revise the definition of heat. In particular, he referred to the work of Constantin Carathéodory, who had in 1909 stated the first law without defining quantity of heat. Born’s definition was specifically for transfers of energy without transfer of matter, and it has been widely followed in textbooks (examples:). Born observes that a transfer of matter between two systems is accompanied by a transfer of internal energy that cannot be resolved into heat and work components. There can be pathways to other systems, spatially separate from that of the matter transfer, that allow heat and work transfer independent of and simultaneous with the matter transfer. Energy is conserved in such transfers.

Description

The first law of thermodynamics for a closed system was expressed in two ways by Clausius. One way referred to cyclic processes and the inputs and outputs of the system, but did not refer to increments in the internal state of the system. The other way referred to an incremental change in the internal state of the system, and did not expect the process to be cyclic. A cyclic process is one that can be repeated indefinitely often, returning the system to its initial state. Of particular interest for single cycle of a cyclic process are the net work done, and the net heat taken in (or ‘consumed’, in Clausius’ statement), by the system.

In a cyclic process in which the system does net work on its surroundings, it is observed to be physically necessary not only that heat be taken into the system, but also, importantly, that some heat leave the system. The difference is the heat converted by the cycle into work. In each repetition of a cyclic process, the net work done by the system, measured in mechanical units, is proportional to the heat consumed, measured in calorimetric units. The constant of proportionality is universal and independent of the system and in 1845 and 1847 was measured by James Joule, who described it as the *mechanical equivalent of heat*.

In a non-cyclic process, the change in the internal energy of a system is equal to net energy added as heat to the system minus the net work done by the system, both being measured in mechanical units. Taking ΔU as a change in internal energy, one writes,

$$\Delta U = Q - W \text{ (sign, convention, of, Clausius, and, generally, in, this, article)}$$

where Q denotes the net quantity of heat supplied to the system by its surroundings and W denotes the net work done by the system. This sign convention is implicit in Clausius’ statement of the law given above. It originated with the study of heat engines that produce useful work by consumption of heat.

Often nowadays, however, writers use the IUPAC convention by which the first law is formulated with work done on the system by its surroundings having a positive sign.

With this now often used sign convention for work, the first law for a closed system may be written:

$$\Delta U = Q + W \text{ (sign, convention, of, IUPAC)}$$

This convention follows physicists such as Max Planck, and considers all net energy transfers to the system as positive and all net energy transfers from the system as negative, irrespective of any use for the system as an engine or other device. When a system expands in a fictive quasistatic process, the work done by the system on the environment is the product, $P dV$, of pressure, P , and volume change, dV , whereas the work done *on* the system is $-P dV$.

Using either sign convention for work, the change in internal energy of the system is:

$$dU = \delta Q - PdV \text{ (quasi-static process),}$$

where δQ denotes the infinitesimal amount of heat supplied to the system from its surroundings.

Work and heat are expressions of actual physical processes of supply or removal of energy, while the internal energy U is a mathematical abstraction that keeps account of the exchanges of energy that befall the system.

Thus the term heat for Q means “that amount of energy added or removed by conduction of heat or by thermal radiation”, rather than referring to a form of energy within the system. Likewise, the term work energy for W means “that amount of energy gained or lost as the result of work”. Internal energy is a property of the system whereas work done and heat supplied are not. A significant result of this distinction is that a given internal energy change ΔU can be achieved by, in principle, many combinations of heat and work.

Various Statements of the Law for Closed Systems

The law is of great importance and generality and is consequently thought of from several points of view. Most careful textbook statements of the law express it for closed systems. It is stated in several ways, sometimes even by the same author. For the thermodynamics of closed systems, the distinction between transfers of energy as work and as heat is central and is within the scope of the present article. For the thermodynamics of open systems, such a distinction is beyond the scope of the present article, but some limited comments are made on it in the section below headed ‘First law of thermodynamics for open systems’.

There are two main ways of stating a law of thermodynamics, physically or mathematically. They should be logically coherent and consistent with one another.

An example of a physical statement is that of Planck (1897/1903):

- It is in no way possible, either by mechanical, thermal, chemical, or other devices, to obtain perpetual motion, *i.e.*, it is impossible to construct an engine which will work in a cycle and produce continuous work, or kinetic energy, from nothing.

This physical statement is restricted neither to closed systems nor to systems with states that are strictly defined only for thermodynamic equilibrium; it has meaning also for open systems and for systems with states that are not in thermodynamic equilibrium.

An example of a mathematical statement is that of Crawford (1963):

For a given system we let ΔE^{kin} = large-scale mechanical energy, ΔE^{pot} = large-scale potential energy, and ΔE^{tot} = total energy. The first two quantities are specifiable in terms of appropriate mechanical variables, and by definition,

$$E^{\text{tot}} = E^{\text{kin}} + E^{\text{pot}} + U.$$

For any finite process, whether reversible or irreversible,

$$\Delta E^{\text{tot}} = \Delta E^{\text{kin}} + \Delta E^{\text{pot}} + \Delta U.$$

The first law in a form that involves the principle of conservation of energy more generally is,

$$\Delta E^{\text{tot}} = Q + W.$$

Here Q and W are heat and work added, with no restrictions as to whether the process is reversible, quasistatic, or irreversible.

This statement by Crawford, for W , uses the sign convention of IUPAC, not that of Clausius. Though it does not explicitly say so, this statement refers to closed systems, and to internal energy U defined for bodies in states of thermodynamic equilibrium, which possess well-defined temperatures. The history of statements of the law for closed systems has two main periods, before and after the work of Bryan (1907), of Carathéodory (1909), and the approval of Carathéodory's work given by Born (1921). The earlier traditional versions of the law for closed systems are nowadays often considered to be out of date.

Carathéodory's celebrated presentation of equilibrium thermodynamics refers to closed systems, which are allowed to contain several phases connected by internal walls of various kinds of impermeability and permeability (explicitly including walls that are permeable only to heat). Carathéodory's 1909 version of the first law of thermodynamics was stated in an axiom which refrained from defining or mentioning temperature or quantity of heat transferred. That axiom stated that the internal energy of a phase in equilibrium is a function of state, that the sum of the internal energies of the phases is the total internal energy of the system, and that the value of the total internal energy of the system is changed by the amount of work done adiabatically on it, considering work as a form of energy. That article considered this statement to be an expression of the law of conservation of energy for such systems. This version is nowadays widely accepted as authoritative, but is stated in slightly varied ways by different authors. Such statements of the first law for closed systems assert the existence of internal energy as a function of state defined in terms of adiabatic work. Thus heat is not defined calorimetrically or as due to temperature difference. It is defined as a residual difference between change of internal energy and work done on the system, when that work does not account for the whole of the change of internal energy and the system is not adiabatically isolated.

The 1909 Carathéodory statement of the law in axiomatic form does not mention heat or temperature, but the equilibrium states to which it refers are explicitly defined by variable sets that necessarily include "non-deformation variables", such as pressures, which, within reasonable restrictions, can be rightly interpreted as empirical temperatures, and the walls connecting the phases of the system are explicitly defined as possibly impermeable to heat or permeable only to heat.

According to Münster (1970), "A somewhat unsatisfactory aspect of Carathéodory's theory is that a consequence of the Second Law must be considered at this point [in the statement of the first law], *i.e.*, that it is not always possible to reach any state 2 from any other state 1 by means of an adiabatic process." Münster instances that no adiabatic process can reduce the internal energy of a system at constant volume. Carathéodory's paper asserts that its statement of the first law corresponds exactly to Joule's experimental arrangement, regarded as an instance of adiabatic work. It does not point out that Joule's experimental arrangement performed essentially irreversible work, through friction of paddles in a liquid, or passage of electric current through a resistance inside the system, driven by motion of a coil and inductive heating, or by an external current source, which can access the system only by the passage of electrons, and so is not strictly adiabatic, because electrons are a form of matter, which cannot penetrate adiabatic walls. The paper goes on to base its main argument on the possibility of quasi-static adiabatic work, which is essentially reversible. The paper asserts that it will avoid reference to Carnot cycles, and then proceeds to base its argument on cycles of forward and backward quasi-static adiabatic stages, with isothermal stages of zero magnitude.

Sometimes the concept of internal energy is not made explicit in the statement.

Sometimes the existence of the internal energy is made explicit but work is not explicitly mentioned in the statement of the first postulate of thermodynamics. Heat supplied is then defined as the residual change in internal energy after work has been taken into account, in a non-adiabatic process. A respected modern author states the first law of thermodynamics as "Heat is a form of energy", which explicitly mentions neither internal energy nor adiabatic work. Heat is defined as energy transferred by thermal contact with a reservoir, which has a temperature, and is generally so large that addition and removal of heat do not alter its temperature.

A current student text on chemistry defines heat thus: “*heat* is the exchange of thermal energy between a system and its surroundings caused by a temperature difference.” The author then explains how heat is defined or measured by calorimetry, in terms of heat capacity, specific heat capacity, molar heat capacity, and temperature. A respected text disregards the Carathéodory’s exclusion of mention of heat from the statement of the first law for closed systems, and admits heat calorimetrically defined along with work and internal energy. Another respected text defines heat exchange as determined by temperature difference, but also mentions that the Born (1921) version is “completely rigorous”. These versions follow the traditional approach that is now considered out of date, exemplified by that of Planck (1897/1903).

Evidence for the First Law of Thermodynamics for Closed Systems

The first law of thermodynamics for closed systems was originally induced from empirically observed evidence, including calorimetric evidence. It is nowadays, however, taken to provide the definition of heat via the law of conservation of energy and the definition of work in terms of changes in the external parameters of a system. The original discovery of the law was gradual over a period of perhaps half a century or more, and some early studies were in terms of cyclic processes. The following is an account in terms of changes of state of a closed system through compound processes that are not necessarily cyclic. This account first considers processes for which the first law is easily verified because of their simplicity, namely adiabatic processes (in which there is no transfer as heat) and adynamic processes (in which there is no transfer as work).

Adynamic Processes

A complementary observable aspect of the first law is about heat transfer. Adynamic transfer of energy as heat can be measured empirically by changes in the surroundings of the system of interest by calorimetry. This again requires the existence of adiabatic enclosure of the entire process, system and surroundings, though the separating wall between the surroundings and the system is thermally conductive or radiatively permeable, not adiabatic. A calorimeter can rely on measurement of sensible heat, which requires the existence of thermometers and measurement of temperature change in bodies of known sensible heat capacity under specified conditions; or it can rely on the measurement of latent heat, through measurement of masses of material that change phase, at temperatures fixed by the occurrence of phase changes under specified conditions in bodies of known latent heat of phase change. The calorimeter can be calibrated by adiabatically doing externally determined work on it. The most accurate method is by passing an electric current from outside through a resistance inside the calorimeter. The calibration allows comparison of calorimetric measurement of quantity of heat transferred with quantity of energy transferred as work. According to one textbook, “The most common device for measuring ΔU is an adiabatic bomb calorimeter.” According to another textbook, “Calorimetry is widely used in present day laboratories.” According to one opinion, “Most thermodynamic data come from calorimetry...” According to another opinion, “The most common method of measuring “heat” is with a calorimeter.”

When the system evolves with transfer of energy as heat, without energy being transferred as work, in an adynamic process, the heat transferred to the system is equal to the increase in its internal energy:

$$Q_{A \rightarrow B}^{\text{adynamic}} = \Delta V.$$

- The internal energy of an isolated system remains constant.

General Case for Irreversible Processes

If, in a process of change of state of a closed system, the energy transfer is not under a practically zero temperature gradient and practically frictionless, then the process is irreversible. Then the heat and work transfers

may be difficult to calculate, and irreversible thermodynamics is called for. Nevertheless, the first law still holds and provides a check on the measurements and calculations of the work done irreversibly on the system $W_{A \rightarrow B}^{\text{path } P_0, \text{reversible}}$, and the heat transferred irreversibly to the system $Q_{A \rightarrow B}^{\text{path } P_0, \text{reversible}}$, which belong to the same particular process defined by its particular irreversible path, P_1 , through the space of thermodynamic states.

$$W_{A \rightarrow B}^{\text{path } P_1, \text{irreversible}} + W_{A \rightarrow B}^{\text{path } P_1, \text{irreversible}} = \Delta V.$$

This means that the internal energy U is a function of state and that the internal energy change ΔU between two states is a function only of the two states.

Overview of the Weight of Evidence for the Law

The first law of thermodynamics is so general that its predictions cannot all be directly tested. In many properly conducted experiments it has been precisely supported, and never violated. Indeed, within its scope of applicability, the law is so reliably established, that, nowadays, rather than experiment being considered as testing the accuracy of the law, it is more practical and realistic to think of the law as testing the accuracy of experiment. An experimental result that seems to violate the law may be assumed to be inaccurate or wrongly conceived, for example due to failure to account for an important physical factor. Thus, some may regard it as a principle more abstract than a law.

State Functional Formulation for Infinitesimal Processes

When the heat and work transfers in the equations above are infinitesimal in magnitude, they are often denoted by δ , rather than exact differentials denoted by d , as a reminder that heat and work do not describe the *state* of any system.

The integral of an inexact differential depends upon the particular path taken through the space of thermodynamic parameters while the integral of an exact differential depends only upon the initial and final states. If the initial and final states are the same, then the integral of an inexact differential may or may not be zero, but the integral of an exact differential is always zero. The path taken by a thermodynamic system through a chemical or physical change is known as a thermodynamic process.

The first law for a closed homogeneous system may be stated in terms that include concepts that are established in the second law. The internal energy U may then be expressed as a function of the system's defining state variables S , entropy, and V , volume: $U = U(S, V)$. In these terms, T , the system's temperature, and P , its pressure, are partial derivatives of U with respect to S and V . These variables are important throughout thermodynamics, though not necessary for the statement of the first law. Rigorously, they are defined only when the system is in its own state of internal thermodynamic equilibrium. For some purposes, the concepts provide good approximations for scenarios sufficiently near to the system's internal thermodynamic equilibrium.

The first law requires that:

$$dU = \delta Q - \delta W$$

(closed system, general process, quasi-static or irreversible).

Then, for the fictive case of a reversible process, dU can be written in terms of exact differentials. One may imagine reversible changes, such that there is at each instant negligible departure from thermodynamic equilibrium within the system. This excludes isochoric work. Then, mechanical work is given by $\delta W = -P dV$ and the quantity of heat added can be expressed as $\delta Q = T dS$.

For these conditions,

$$dU = T dS - P dV \text{ (closed system, reversible process).}$$

While this has been shown here for reversible changes, it is valid in general, as U can be considered as a thermodynamic state function of the defining state variables S and V :

$$dU = T dS - P dV \text{ (closed system, general process, quasi-static or irreversible).}$$

Equation is known as the fundamental thermodynamic relation for a closed system in the energy representation, for which the defining state variables are S and V , with respect to which T and P are partial derivatives of U . It is only in the fictive reversible case, when isochoric work is excluded, that the work done and heat transferred are given by $-P dV$ and $T dS$.

In the case of a closed system in which the particles of the system are of different types and, because chemical reactions may occur, their respective numbers are not necessarily constant, the fundamental thermodynamic relation for dU becomes:

$$dU = TdS - PdV + \sum_i \mu_i dN_i.$$

where dN_i is the (small) increase in amount of type- i particles in the reaction, and μ_i is known as the chemical potential of the type- i particles in the system. If dN_i is expressed in mol then μ_i is expressed in J/mol. If the system has more external mechanical variables than just the volume that can change, the fundamental thermodynamic relation further generalizes to:

$$dU = TdS - \sum_i X_i dx_i + \sum_j \mu_j dN_j.$$

Here the X_i are the generalized forces corresponding to the external variables x_i . The parameters X_i are independent of the size of the system and are called intensive parameters and the x_i are proportional to the size and called extensive parameters.

For an open system, there can be transfers of particles as well as energy into or out of the system during a process. For this case, the first law of thermodynamics still holds, in the form that the internal energy is a function of state and the change of internal energy in a process is a function only of its initial and final states, as noted in the section below headed First law of thermodynamics for open systems.

A useful idea from mechanics is that the energy gained by a particle is equal to the force applied to the particle multiplied by the displacement of the particle while that force is applied. Now consider the first law without the heating term: $dU = -PdV$. The pressure P can be viewed as a force (and in fact has units of force per unit area) while dV is the displacement (with units of distance times area). We may say, with respect to this work term, that a pressure difference forces a transfer of volume, and that the product of the two (work) is the amount of energy transferred out of the system as a result of the process. If one were to make this term negative then this would be the work done on the system.

It is useful to view the TdS term in the same light: here the temperature is known as a “generalized” force (rather than an actual mechanical force) and the entropy is a generalized displacement. Similarly, a difference in chemical potential between groups of particles in the system drives a chemical reaction that changes the numbers of particles, and the corresponding product is the amount of chemical potential energy transformed in process.

For example, consider a system consisting of two phases: liquid water and water vapor. There is a generalized “force” of evaporation that drives water molecules out of the liquid. There is a generalized “force” of condensation that drives vapor molecules out of the vapor. Only when these two “forces” (or chemical potentials) are equal is there equilibrium, and the net rate of transfer zero. The two thermodynamic parameters that form a generalized force-displacement pair are called “conjugate variables”. The two most familiar pairs are, of course, pressure-volume, and temperature-entropy.

Spatially Inhomogeneous Systems

Classical thermodynamics is initially focused on closed homogeneous systems (*e.g.*, Planck 1897/1903), which might be regarded as ‘zero-dimensional’ in the sense that they have no spatial variation. But it is desired to study also systems with distinct internal motion and spatial inhomogeneity. For such systems, the principle of conservation of energy is expressed in terms not only of internal energy as defined for homogeneous systems, but also in terms of kinetic energy and potential energies of parts of the inhomogeneous system with respect to each other and with respect to long-range external forces.

How the total energy of a system is allocated between these three more specific kinds of energy varies according to the purposes of different writers; this is because these components of energy are to some extent mathematical artefacts rather than actually measured physical quantities. For any closed homogeneous component of an inhomogeneous closed system, if E denotes the total energy of that component system, one may write;

$$E = E^{\text{kin}} + E^{\text{pot}} + U$$

Where E^{kin} and E^{pot} denote respectively the total kinetic energy and the total potential energy of the component closed homogeneous system, and U denotes its internal energy.

Potential energy can be exchanged with the surroundings of the system when the surroundings impose a force field, such as gravitational or electromagnetic, on the system. A compound system consisting of two interacting closed homogeneous component subsystems has a potential energy of interaction E_{12}^{pot} between the subsystems.

Thus, in an obvious notation, one may write,

$$E = E_1^{\text{kin}} + E_1^{\text{pot}} + U_1 = E_2^{\text{kin}} + E_2^{\text{pot}} + U_2 + E_{12}^{\text{pot}}$$

The quantity E_{12}^{pot} in general lacks an assignment to either subsystem in a way that is not arbitrary, and this stands in the way of a general non-arbitrary definition of transfer of energy as work. On occasions, authors make their various respective arbitrary assignments.

The distinction between internal and kinetic energy is hard to make in the presence of turbulent motion within the system, as friction gradually dissipates macroscopic kinetic energy of localised bulk flow into molecular random motion of molecules that is classified as internal energy. The rate of dissipation by friction of kinetic energy of localised bulk flow into internal energy, whether in turbulent or in streamlined flow, is an important quantity in non-equilibrium thermodynamics. This is a serious difficulty for attempts to define entropy for time-varying spatially inhomogeneous systems.

First Law of Thermodynamics for Open Systems

For the first law of thermodynamics, there is no trivial passage of physical conception from the closed system view to an open system view. For closed systems, the concepts of an adiabatic enclosure and of an adiabatic wall are fundamental. Matter and internal energy cannot permeate or penetrate such a wall. For an open system, there is a wall that allows penetration by matter. In general, matter in diffusive motion carries with it some internal energy, and some microscopic potential energy changes accompany the motion. An open system is not adiabatically enclosed.

There are some cases in which a process for an open system can, for particular purposes, be considered as if it were for a closed system. In an open system, by definition hypothetically or potentially, matter can pass between the system and its surroundings. But when, in a particular case, the process of interest involves only hypothetical or potential but no actual passage of matter, the process can be considered as if it were for a closed system.

Internal Energy for an Open System

Since, the revised and more rigorous definition of the internal energy of a closed system rests upon the possibility of processes by which adiabatic work takes the system from one state to another, this leaves a problem for the definition of internal energy for an open system, for which adiabatic work is not in general possible. According to Max Born, the transfer of matter and energy across an open connection “cannot be reduced to mechanics”. In contrast to the case of closed systems, for open systems, in the presence of diffusion, there is no unconstrained and unconditional physical distinction between convective transfer of internal energy by bulk flow of matter, the transfer of internal energy without transfer of matter (usually called heat conduction and work transfer), and change of various potential energies.

The older traditional way and the conceptually revised (Carathéodory) way agree that there is no physically unique definition of heat and work transfer processes between open systems. In particular, between two otherwise isolated open systems an adiabatic wall is by definition impossible. This problem is solved by recourse to the principle of conservation of energy. This principle allows a composite isolated system to be derived from two other component non-interacting isolated systems, in such a way that the total energy of the composite isolated system is equal to the sum of the total energies of the two component isolated systems. Two previously isolated systems can be subjected to the thermodynamic operation of placement between them of a wall permeable to matter and energy, followed by a time for establishment of a new thermodynamic state of internal equilibrium in the new single unpartitioned system. The internal energies of the initial two systems and of the final new system, considered respectively as closed systems as above, can be measured. Then the law of conservation of energy requires that,

$$\Delta U_s + \Delta U_o = 0,$$

where ΔU_s and ΔU_o denote the changes in internal energy of the system and of its surroundings respectively. This is a statement of the first law of thermodynamics for a transfer between two otherwise isolated open systems, that fits well with the conceptually revised and rigorous statement of the law stated above.

For the thermodynamic operation of adding two systems with internal energies U_1 and U_2 , to produce a new system with internal energy U , one may write $U = U_1 + U_2$; the reference states for U , U_1 and U_2 should be specified accordingly, maintaining also that the internal energy of a system be proportional to its mass, so that the internal energies are extensive variables. There is a sense in which this kind of additivity expresses a fundamental postulate that goes beyond the simplest ideas of classical closed system thermodynamics; the extensivity of some variables is not obvious, and needs explicit expression; indeed one author goes so far as to say that it could be recognized as a fourth law of thermodynamics, though this is not repeated by other authors.

Also of course,

$$\Delta N_s + \Delta N_o = 0,$$

where ΔN_s and ΔN_o denote the changes in mole number of a component substance of the system and of its surroundings respectively. This is a statement of the law of conservation of mass.

Process of Transfer of Matter between an Open System and its Surroundings

A system connected to its surroundings only through contact by a single permeable wall, but otherwise isolated, is an open system. If it is initially in a state of contact equilibrium with a surrounding subsystem, a thermodynamic process of transfer of matter can be made to occur between them if the surrounding subsystem is subjected to some thermodynamic operation, for example, removal of a partition between it and some further surrounding subsystem. The removal of the partition in the surroundings initiates a process of exchange between the system and its contiguous surrounding subsystem.

An example is evaporation. One may consider an open system consisting of a collection of liquid, enclosed except where it is allowed to evaporate into or to receive condensate from its vapor above it, which may be considered as its contiguous surrounding subsystem, and subject to control of its volume and temperature.

A thermodynamic process might be initiated by a thermodynamic operation in the surroundings, that mechanically increases in the controlled volume of the vapor. Some mechanical work will be done within the surroundings by the vapor, but also some of the parent liquid will evaporate and enter the vapor collection which is the contiguous surrounding subsystem. Some internal energy will accompany the vapor that leaves the system, but it will not make sense to try to uniquely identify part of that internal energy as heat and part of it as work. Consequently, the energy transfer that accompanies the transfer of matter between the system and its surrounding subsystem cannot be uniquely split into heat and work transfers to or from the open system. The component of total energy transfer that accompanies the transfer of vapor into the surrounding subsystem is customarily called 'latent heat of evaporation', but this use of the word heat is a quirk of customary historical language, not in strict compliance with the thermodynamic definition of transfer of energy as heat. In this example, kinetic energy of bulk flow and potential energy with respect to long-range external forces such as gravity are both considered to be zero. The first law of thermodynamics refers to the change of internal energy of the open system, between its initial and final states of internal equilibrium.

Open System with Multiple Contacts

An open system can be in contact equilibrium with several other systems at once. This includes cases in which there is contact equilibrium between the system, and several subsystems in its surroundings, including separate connections with subsystems through walls that are permeable to the transfer of matter and internal energy as heat and allowing friction of passage of the transferred matter, but immovable, and separate connections through adiabatic walls with others, and separate connections through diathermic walls impermeable to matter with yet others. Because there are physically separate connections that are permeable to energy but impermeable to matter, between the system and its surroundings, energy transfers between them can occur with definite heat and work characters. Conceptually essential here is that the internal energy transferred with the transfer of matter is measured by a variable that is mathematically independent of the variables that measure heat and work.

With such independence of variables, the total increase of internal energy in the process is then determined as the sum of the internal energy transferred from the surroundings with the transfer of matter through the walls that are permeable to it, and of the internal energy transferred to the system as heat through the diathermic walls, and of the energy transferred to the system as work through the adiabatic walls, including the energy transferred to the system by long-range forces. These simultaneously transferred quantities of energy are defined by events in the surroundings of the system. Because the internal energy transferred with matter is not in general uniquely resolvable into heat and work components, the total energy transfer cannot in general be uniquely resolved into heat and work components. Under these conditions, the following formula can describe the process in terms of externally defined thermodynamic variables, as a statement of the first law of thermodynamics:

$$(3)DU_0 = Q - W - \sum_{i=1}^m \Delta U_i$$

(suitably defined surrounding subsystems, general process, quasi – static or irreversible)

where ΔU_0 denotes the change of internal energy of the system, and ΔU_i denotes the change of internal energy of the i th of the m surrounding subsystems that are in open contact with the system, due to transfer between the

system and that i th surrounding subsystem, and Q denotes the internal energy transferred as heat from the heat reservoir of the surroundings to the system, and W denotes the energy transferred from the system to the surrounding subsystems that are in adiabatic connection with it. The case of a wall that is permeable to matter and can move so as to allow transfer of energy as work is not considered here.

Combination of First and Second Laws

If the system is described by the energetic fundamental equation, $U_0 = U_0(S, V, N_j)$, and if the process can be described in the quasi-static formalism, in terms of the internal state variables of the system, then the process can also be described by a combination of the first and second laws of thermodynamics, by the formula,

$$(4)dU_0 = TdS - PdV + \sum_{j=1}^n \mu_j dN_j \triangleright$$

(suitably defined surrounding subsystems, general process, quasi – static or irreversible)

where there are n chemical constituents of the system and permeably connected surrounding subsystems, and where T, S, P, V, N_j , and μ_j , are defined as above.

For a general natural process, there is no immediate term-wise correspondence between equations (3) and (4), because they describe the process in different conceptual frames.

Nevertheless, a conditional correspondence exists. There are three relevant kinds of wall here: purely diathermal, adiabatic, and permeable to matter. If two of those kinds of wall are sealed off, leaving only one that permits transfers of energy, as work, as heat, or with matter, then the remaining permitted terms correspond precisely. If two of the kinds of wall are left unsealed, then energy transfer can be shared between them, so that the two remaining permitted terms do not correspond precisely.

For the special fictive case of quasi-static transfers, there is a simple correspondence. For this, it is supposed that the system has multiple areas of contact with its surroundings. There are pistons that allow adiabatic work, purely diathermal walls, and open connections with surrounding subsystems of completely controllable chemical potential (or equivalent controls for charged species). Then, for a suitable fictive quasi-static transfer, one can write,

$$\delta Q = TdS \text{ and } \delta W = PdV$$

(suitably defined surrounding subsystems, quasi – static transfers of energy)

For fictive quasi-static transfers for which the chemical potentials in the connected surrounding subsystems are suitably controlled, these can be put into equation (4) to yield,

$$(5)dU_0 = \delta Q - \delta W + \sum_{j=1}^n \mu_j dN_j$$

(suitably defined surrounding subsystems, quasi – static transfers)

The reference does not actually write equation (5), but what it does write is fully compatible with it. Another helpful account is given by Tschoegl.

There are several other accounts of this, in apparent mutual conflict.

Non-equilibrium Transfers

The transfer of energy between an open system and a single contiguous subsystem of its surroundings is considered also in non-equilibrium thermodynamics. The problem of definition arises also in this case. It may be allowed that the wall between the system and the subsystem is not only permeable to matter and to internal

energy, but also may be movable so as to allow work to be done when the two systems have different pressures. In this case, the transfer of energy as heat is not defined.

Methods for study of non-equilibrium processes mostly deal with spatially continuous flow systems. In this case, the open connection between system and surroundings is usually taken to fully surround the system, so that there are no separate connections impermeable to matter but permeable to heat. Except for the special case mentioned above when there is no actual transfer of matter, which can be treated as if for a closed system, in strictly defined thermodynamic terms, it follows that transfer of energy as heat is not defined. In this sense, there is no such thing as ‘heat flow’ for a continuous-flow open system. Properly, for closed systems, one speaks of transfer of internal energy as heat, but in general, for open systems, one can speak safely only of transfer of internal energy. A factor here is that there are often cross-effects between distinct transfers, for example that transfer of one substance may cause transfer of another even when the latter has zero chemical potential gradient.

Usually transfer between a system and its surroundings applies to transfer of a state variable, and obeys a balance law, that the amount lost by the donor system is equal to the amount gained by the receptor system. Heat is not a state variable. For his 1947 definition of “heat transfer” for discrete open systems, the author Prigogine carefully explains at some length that his definition of it does not obey a balance law. He describes this as paradoxical.

The situation is clarified by Gyarmati, who shows that his definition of “heat transfer”, for continuous-flow systems, really refers not specifically to heat, but rather to transfer of internal energy, as follows. He considers a conceptual small cell in a situation of continuous-flow as a system defined in the so-called Lagrangian way, moving with the local center of mass. The flow of matter across the boundary is zero when considered as a flow of total mass. Nevertheless, if the material constitution is of several chemically distinct components that can diffuse with respect to one another, the system is considered to be open, the diffusive flows of the components being defined with respect to the center of mass of the system, and balancing one another as to mass transfer. Still there can be a distinction between bulk flow of internal energy and diffusive flow of internal energy in this case, because the internal energy density does not have to be constant per unit mass of material, and allowing for non-conservation of internal energy because of local conversion of kinetic energy of bulk flow to internal energy by viscosity.

Gyarmati shows that his definition of “the heat flow vector” is strictly speaking a definition of flow of internal energy, not specifically of heat, and so it turns out that his use here of the word heat is contrary to the strict thermodynamic definition of heat, though it is more or less compatible with historical custom, that often enough did not clearly distinguish between heat and internal energy; he writes “that this relation must be considered to be the exact definition of the concept of heat flow, fairly loosely used in experimental physics and heat technics.” Apparently in a different frame of thinking from that of the above-mentioned paradoxical usage in the earlier sections of the historic 1947 work by Prigogine, about discrete systems, this usage of Gyarmati is consistent with the later sections of the same 1947 work by Prigogine, about continuous-flow systems, which use the term “heat flux” in just this way. This usage is also followed by Glansdorff and Prigogine in their 1971 text about continuous-flow systems.

They write: “Again the flow of internal energy may be split into a convection flow ρuv and a conduction flow. This conduction flow is by definition the heat flow W .

Therefore: $j[U] = \rho uv + W$ where u denotes the [internal] energy per unit mass. [These authors actually use the symbols E and e to denote internal energy but their notation has been changed here to accord with the notation of the present article. These authors actually use the symbol U to refer to total energy, including kinetic energy of bulk flow.]” This usage is followed also by other writers on non-equilibrium thermodynamics such as

Lebon, Jou, and Casas-Vásquez, and de Groot and Mazur. This usage is described by Bailyn as stating the non-convective flow of internal energy, and is listed as his definition number 1, according to the first law of thermodynamics. This usage is also followed by workers in the kinetic theory of gases. This is not the *ad hoc* definition of “reduced heat flux” of Haase. In the case of a flowing system of only one chemical constituent, in the Lagrangian representation, there is no distinction between bulk flow and diffusion of matter. Moreover, the flow of matter is zero into or out of the cell that moves with the local center of mass. In effect, in this description, one is dealing with a system effectively closed to the transfer of matter. But still one can validly talk of a distinction between bulk flow and diffusive flow of internal energy, the latter driven by a temperature gradient within the flowing material, and being defined with respect to the local center of mass of the bulk flow. In this case of a virtually closed system, because of the zero matter transfer, as noted above, one can safely distinguish between transfer of energy as work, and transfer of internal energy as heat.

SECOND LAW OF THERMODYNAMICS

The second law of thermodynamics states that the total entropy of an isolated system can never decrease over time. The total entropy can remain constant in ideal cases where the system is in a steady state (equilibrium), or is undergoing a reversible process. In all spontaneous processes, the total entropy always increases and the process is irreversible. The increase in entropy accounts for the irreversibility of natural processes, and the asymmetry between future and past. Historically, the second law was an empirical finding that was accepted as an axiom of thermodynamic theory. Statistical thermodynamics, classical or quantum, explains the microscopic origin of the law. The second law has been expressed in many ways. Its first formulation is credited to the French scientist Sadi Carnot, who in 1824 showed that there is an upper limit to the efficiency of conversion of heat to work, in a heat engine.

The first law of thermodynamics provides the basic definition of internal energy, associated with all thermodynamic systems, and states the rule of conservation of energy. The second law is concerned with the direction of natural processes. It asserts that a natural process runs only in one sense, and is not reversible. For example, heat always flows spontaneously from hotter to colder bodies, and never the reverse, unless external work is performed on the system. The explanation of the phenomena was given in terms of entropy. Total entropy (S) can never decrease over time for an isolated system because the entropy of an isolated system spontaneously evolves towards thermodynamic equilibrium: the entropy should stay the same or increase.

In a fictive reversible process, an infinitesimal increment in the entropy (dS) of a system is defined to result from an infinitesimal transfer of heat (δQ) to a closed system (which allows the entry or exit of energy-but not matter) divided by the common temperature (T) of the system and the surroundings which supply the heat:

$$dS = \frac{\delta Q}{T} (\text{closed system, idealized fictive reversible process})$$

Different notations are used for infinitesimal amounts of heat (δ) and infinitesimal amounts of entropy (d) because entropy is a function of state, while heat, like work, is not. For an actually possible infinitesimal process without exchange of matter with the surroundings, the second law requires that the increment in system entropy be greater than that:

$$dS > \frac{\delta Q}{T} (\text{closed system, actually possible, irreversible process})$$

This is because a general process for this case may include work being done on the system by its surroundings, which must have frictional or viscous effects inside the system, and because heat transfer actually occurs only irreversibly, driven by a finite temperature difference. Introducing a set of internal variables ξ to describe the deviation of a thermodynamic system from the equilibrium state, one can record the equality,

$$dS = \frac{\delta Q}{T} = \frac{1}{T} \sum_j \Xi_j \delta \xi_j \quad (\text{Closed system, actually possible, irreversible process}).$$

The second term represents work of internal variables that can be perturbed by external influences, but the system cannot perform any positive work via internal variables. This statement introduces the impossibility of the reversion of evolution of the thermodynamic system in time and can be considered as a formulation of *the second principle of thermodynamics*—the formulation, which is, of course, equivalent to the formulation of the principle in terms of entropy. The zeroth law of thermodynamics in its usual short statement allows recognition that two bodies in a relation of thermal equilibrium have the same temperature, especially that a test body has the same temperature as a reference thermometric body. For a body in thermal equilibrium with another, there are indefinitely many empirical temperature scales, in general respectively depending on the properties of a particular reference thermometric body. The second law allows a distinguished temperature scale, which defines an absolute, thermodynamic temperature, independent of the properties of any particular reference thermometric body.

Various Statements of the Law

The second law of thermodynamics may be expressed in many specific ways, the most prominent classical statements being the statement by Rudolf Clausius (1854), the statement by Lord Kelvin (1851), and the statement in axiomatic thermodynamics by Constantin Carathéodory (1909). These statements cast the law in general physical terms citing the impossibility of certain processes. The Clausius and the Kelvin statements have been shown to be equivalent.

Carnot's Principle

The historical origin of the second law of thermodynamics was in Carnot's principle. It refers to a cycle of a Carnot heat engine, fictively operated in the limiting mode of extreme slowness known as quasi-static, so that the heat and work transfers are between subsystems that are always in their own internal states of thermodynamic equilibrium. The Carnot engine is an idealized device of special interest to engineers who are concerned with the efficiency of heat engines. Carnot's principle was recognized by Carnot at a time when the caloric theory of heat was seriously considered, before the recognition of the first law of thermodynamics, and before the mathematical expression of the concept of entropy. Interpreted in the light of the first law, it is physically equivalent to the second law of thermodynamics, and remains valid today.

It states:

- The efficiency of a quasi-static or reversible Carnot cycle depends only on the temperatures of the two heat reservoirs, and is the same, whatever the working substance. A Carnot engine operated in this way is the most efficient possible heat engine using those two temperatures.

Clausius Statement

The German scientist Rudolf Clausius laid the foundation for the second law of thermodynamics in 1850 by examining the relation between heat transfer and work. His formulation of the second law, which was published in German in 1854, is known as the *Clausius statement*:

- Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time.

The statement by Clausius uses the concept of 'passage of heat'. As is usual in thermodynamic discussions, this means 'net transfer of energy as heat', and does not refer to contributory transfers one way and the other. Heat cannot spontaneously flow from cold regions to hot regions without external work being performed on the system, which is evident from ordinary experience of refrigeration, for example. In a refrigerator, heat flows from cold to hot, but only when forced by an external agent, the refrigeration system.

Kelvin Statement

Lord Kelvin expressed the second law as:

- It is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects.

Equivalence of the Clausius and the Kelvin Statements

Suppose there is an engine violating the Kelvin statement: *i.e.*, one that drains heat and converts it completely into work in a cyclic fashion without any other result. Now pair it with a reversed Carnot engine as shown by the figure. The net and sole effect of this newly created engine consisting of the two engines mentioned is transferring heat;

$$\Delta Q = Q \left(\frac{1}{\eta} - 1 \right)$$

from the cooler reservoir to the hotter one, which violates the Clausius statement. Thus a violation of the Kelvin statement implies a violation of the Clausius statement, *i.e.*, the Clausius statement implies the Kelvin statement. We can prove in a similar manner that the Kelvin statement implies the Clausius statement, and hence, the two are equivalent.

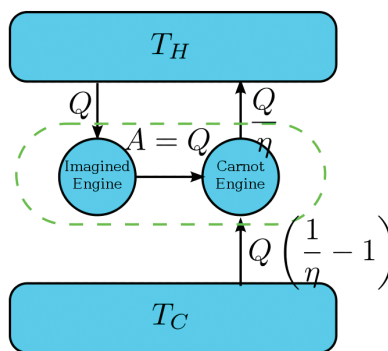


Fig. Derive Kelvin Statement from Clausius Statement.

Planck's Proposition

Planck offered the following proposition as derived directly from experience. This is sometimes regarded as his statement of the second law, but he regarded it as a starting point for the derivation of the second law.

- It is impossible to construct an engine which will work in a complete cycle, and produce no effect except the raising of a weight and cooling of a heat reservoir.

Relation between Kelvin's Statement and Planck's Proposition

It is almost customary in textbooks to speak of the “Kelvin-Planck statement” of the law, as for example in the text by ter Haar and Wergeland. One text gives a statement very like Planck’s proposition, but attributes it to Kelvin without mention of Planck.

One monograph quotes Planck’s proposition as the “Kelvin-Planck” formulation, the text naming Kelvin as its author, though it correctly cites Planck in its references. The reader may compare the two statements quoted just above here. *The Kelvin–Planck statement (or the heat engine statement) of the second law of thermodynamics states that:*

- It is impossible to devise a cyclically operating device, the sole effect of which is to absorb energy in the form of heat from a single thermal reservoir and to deliver an equivalent amount of work.

Planck's Statement

Planck stated the second law as follows.

- Every process occurring in nature proceeds in the sense in which the sum of the entropies of all bodies taking part in the process is increased. In the limit, *i.e.*, for reversible processes, the sum of the entropies remains unchanged.

Rather like Planck's statement is that of Uhlenbeck and Ford for *irreversible phenomena*.

- ... in an irreversible or spontaneous change from one equilibrium state to another (as for example the equalization of temperature of two bodies A and B, when brought in contact) the entropy always increases.

Principle of Carathéodory

Constantin Carathéodory formulated thermodynamics on a purely mathematical axiomatic foundation.

His statement of the second law is known as the Principle of Carathéodory, which may be formulated as follows:

- In every neighborhood of any state S of an adiabatically enclosed system there are states inaccessible from S.

With this formulation, he described the concept of adiabatic accessibility for the first time and provided the foundation for a new subfield of classical thermodynamics, often called geometrical thermodynamics. It follows from Carathéodory's principle that quantity of energy quasi-statically transferred as heat is a holonomic process function, in other words,

$$\delta Q = TdS$$

Though it is almost customary in textbooks to say that Carathéodory's principle expresses the second law and to treat it as equivalent to the Clausius or to the Kelvin-Planck statements, such is not the case. To get all the content of the second law, Carathéodory's principle needs to be supplemented by Planck's principle, that isochoric work always increases the internal energy of a closed system that was initially in its own internal thermodynamic equilibrium.

Planck's Principle

In 1926, Max Planck wrote an important paper on the basics of thermodynamics.

He indicated the principle:

- The internal energy of a closed system is increased by an adiabatic process, throughout the duration of which, the volume of the system remains constant.

This formulation does not mention heat and does not mention temperature, nor even entropy, and does not necessarily implicitly rely on those concepts, but it implies the content of the second law. A closely related statement is that "Frictional pressure never does positive work." Using a now-obsolete form of words, Planck himself wrote: "The production of heat by friction is irreversible."

Not mentioning entropy, this principle of Planck is stated in physical terms. It is very closely related to the Kelvin statement given just above. It is relevant that for a system at constant volume and mole numbers, the entropy is a monotonic function of the internal energy. Nevertheless, this principle of Planck is not actually

Planck's preferred statement of the second law, which is quoted above, in a previous sub-section of the present section of this present article, and relies on the concept of entropy.

A statement that in a sense is complementary to Planck's principle is made by Borgnakke and Sonntag.

They do not offer it as a full statement of the second law:

- ... there is only one way in which the entropy of a [closed] system can be decreased, and that is to transfer heat from the system.

Differing from Planck's just foregoing principle, this one is explicitly in terms of entropy change. Removal of matter from a system can also decrease its entropy.

Statement for a System that Has a known Expression of its Internal Energy as a Function of its Extensive State Variables

The second law has been shown to be equivalent to the internal energy U being a weakly convex function, when written as a function of extensive properties (mass, volume, entropy...).

Corollaries

Perpetual Motion of the Second Kind

Before the establishment of the second law, many people who were interested in inventing a perpetual motion machine had tried to circumvent the restrictions of first law of thermodynamics by extracting the massive internal energy of the environment as the power of the machine. Such a machine is called a "perpetual motion machine of the second kind". The second law declared the impossibility of such machines.

Carnot Theorem

Carnot's theorem (1824) is a principle that limits the maximum efficiency for any possible engine. The efficiency solely depends on the temperature difference between the hot and cold thermal reservoirs.

Carnot's theorem states:

- All irreversible heat engines between two heat reservoirs are less efficient than a Carnot engine operating between the same reservoirs.
- All reversible heat engines between two heat reservoirs are equally efficient with a Carnot engine operating between the same reservoirs.

In his ideal model, the heat of caloric converted into work could be reinstated by reversing the motion of the cycle, a concept subsequently known as thermodynamic reversibility. Carnot, however, further postulated that some caloric is lost, not being converted to mechanical work. Hence, no real heat engine could realise the Carnot cycle's reversibility and was condemned to be less efficient.

Though formulated in terms of caloric, rather than entropy, this was an early insight into the second law.

Clausius Inequality

The Clausius theorem (1854) states that in a cyclic process:

$$\oint \frac{\delta Q}{T} \leq 0.$$

The equality holds in the reversible case and the ' $<$ ' is in the irreversible case. The reversible case is used to introduce the state function entropy. This is because in cyclic processes the variation of a state function is zero from state functionality.

Thermodynamic Temperature

For an arbitrary heat engine, the efficiency is:

$$\eta = \frac{W_n}{q_H} = \frac{q_H - q_C}{q_H} = 1 - \frac{q_C}{q_H} \quad (1)$$

where W_n is for the net work done per cycle. Thus the efficiency depends only on q_C/q_H .

Carnot's theorem states that all reversible engines operating between the same heat reservoirs are equally efficient. Thus, any reversible heat engine operating between temperatures T_1 and T_2 must have the same efficiency, that is to say, the efficiency is the function of temperatures only:

$$\frac{q_C}{q_H} = f(T_H, T_C) \quad (2)$$

In addition, a reversible heat engine operating between temperatures T_1 and T_3 must have the same efficiency as one consisting of two cycles, one between T_1 and another (intermediate) temperature T_2 , and the second between T_2 and T_3 .

This can only be the case if,

$$f(T_1, T_3) = \frac{q_3}{q_1} = \frac{q_2 q_3}{q_1 q_2} = f(T_1, T_2) f(T_2, T_3).$$

Now consider the case where T_1 is a fixed reference temperature: the temperature of the triple point of water. Then for any T_2 and T_3 ,

$$f(T_2, T_3) = \frac{f(T_1, T_3)}{f(T_1, T_2)} = \frac{278.16 \cdot f(T_1, T_3)}{278.16 f(T_1, T_2)}.$$

Therefore, if thermodynamic temperature is defined by,

$$T = 273.16 \times f(T_1, T)$$

then the function f , viewed as a function of thermodynamic temperature, is simply,

$$f(T_2, T_3) = \frac{T_3}{T_2},$$

and the reference temperature T_1 will have the value 273.16. (Any reference temperature and any positive numerical value could be used—the choice here corresponds to the Kelvin scale.)

Entropy

According to the Clausius equality, for a reversible process,

$$\oint \frac{\delta Q}{T} = 0$$

That means the line integral $\int_L \frac{\delta Q}{T}$ path independent.

So we can define a state function S called entropy, which satisfies,

$$dS = \frac{\delta Q}{T}$$

With this we can only obtain the difference of entropy by integrating the above formula. To obtain the absolute value, we need the third law of thermodynamics, which states that $S=0$ at absolute zero for perfect crystals.

For any irreversible process, since, entropy is a state function, we can always connect the initial and terminal states with an imaginary reversible process and integrating on that path to calculate the difference in entropy.

Now reverse the reversible process and combine it with the said irreversible process. Applying Clausius inequality on this loop,

$$-\Delta S + \int \frac{\delta Q}{T} = \oint \frac{\delta Q}{T} < 0$$

Thus,

$$-\Delta S \geq + \int \frac{\delta Q}{T}$$

where the equality holds if the transformation is reversible.

Notice that if the process is an adiabatic process, then,

$$\delta Q = 0, \quad \text{so} \quad \Delta S \geq 0.$$

Energy, Available Useful Work

An important and revealing idealized special case is to consider applying the Second Law to the scenario of an isolated system (called the total system or universe), made up of two parts: a sub-system of interest, and the sub-system's surroundings. These surroundings are imagined to be so large that they can be considered as an *unlimited* heat reservoir at temperature T_R and pressure P_R —so that no matter how much heat is transferred to (or from) the sub-system, the temperature of the surroundings will remain T_R ; and no matter how much the volume of the sub-system expands (or contracts), the pressure of the surroundings will remain P_R .

Whatever changes to dS and dS_R occur in the entropies of the sub-system and the surroundings individually, according to the Second Law the entropy S_{tot} of the isolated total system must not decrease:

$$dS_{tot} = dS + dS_R \geq 0$$

According to the first law of thermodynamics, the change dU in the internal energy of the sub-system is the sum of the heat δq added to the sub-system, *less* any work δw done *by* the sub-system, *plus* any net chemical energy entering the sub-system $\sum \mu_{iR} N_i$, so that:

$$dU = \delta q - \delta w + d(\sum \mu_i R N_i)$$

where μ_{iR} are the chemical potentials of chemical species in the external surroundings.

Now the heat leaving the reservoir and entering the sub-system is,

$$\delta q = T_R (-dS_R) \leq T_R dS$$

where we have first used the definition of entropy in classical thermodynamics (alternatively, in statistical thermodynamics, the relation between entropy change, temperature and absorbed heat can be derived); and then the Second Law inequality from above.

It therefore follows that any net work δw done by the sub-system must obey,

$$\delta w \leq -dU + T_R dS + \sum \mu_i R dN_i$$

It is useful to separate the work δw done by the subsystem into the *useful* work δw_u that can be done *by* the sub-system, over and beyond the work $p_R dV$ done merely by the sub-system expanding against the surrounding external pressure, giving the following relation for the useful work (exergy) that can be done:

$$\delta w_u \leq -d(U - T_R S + p_R V - \sum \mu_i R N_i)$$

It is convenient to define the right-hand-side as the exact derivative of a thermodynamic potential, called the *availability* or *exergy* E of the subsystem,

$$E = U - T_R S + p_R V - \sum \mu_i R N_i$$

The Second Law therefore implies that for any process which can be considered as divided simply into a subsystem, and an unlimited temperature and pressure reservoir with which it is in contact,

$$dE + \delta w_u \leq 0$$

i.e., the change in the subsystem's exergy plus the useful work done *by* the subsystem (or, the change in the subsystem's exergy less any work, additional to that done by the pressure reservoir, done *on* the system) must be less than or equal to zero. In sum, if a proper *infinite-reservoir-like* reference state is chosen as the system surroundings in the real world, then the Second Law predicts a decrease in E for an irreversible process and no change for a reversible process.

$$dS_{\text{tot}} \leq 0 \text{ is equivalent to } dE + \delta w_u \leq 0$$

This expression together with the associated reference state permits a design engineer working at the macroscopic scale (above the thermodynamic limit) to utilize the Second Law without directly measuring or considering entropy change in a total isolated system.. Those changes have already been considered by the assumption that the system under consideration can reach equilibrium with the reference state without altering the reference state. An efficiency for a process or collection of processes that compares it to the reversible ideal may also be found This approach to the Second Law is widely utilized in engineering practice, environmental accounting, systems ecology, and other disciplines.

History

The first theory of the conversion of heat into mechanical work is due to Nicolas Léonard Sadi Carnot in 1824. He was the first to realize correctly that the efficiency of this conversion depends on the difference of temperature between an engine and its environment. Recognizing the significance of James Prescott Joule's work on the conservation of energy, Rudolf Clausius was the first to formulate the second law during 1850, in this form: heat does not flow *spontaneously* from cold to hot bodies. While common knowledge now, this was contrary to the caloric theory of heat popular at the time, which considered heat as a fluid. From there he was able to infer the principle of Sadi Carnot and the definition of entropy (1865).

Established during the 19th century, the Kelvin-Planck statement of the Second Law says, "It is impossible for any device that operates on a cycle to receive heat from a single reservoir and produce a net amount of work.—This was shown to be equivalent to the statement of Clausius.

The ergodic hypothesis is also important for the Boltzmann approach. It says that, over long periods of time, the time spent in some region of the phase space of microstates with the same energy is proportional to the volume of this region, *i.e.*, that all accessible microstates are equally probable over a long period of time. Equivalently, it says that time average and average over the statistical ensemble are the same. There is a traditional doctrine, starting with Clausius, that entropy can be understood in terms of molecular 'disorder' within a macroscopic system. This doctrine is obsolescent.

Account Given by Clausius

In 1856, the German physicist Rudolf Clausius stated what he called the "second fundamental theorem in the mechanical theory of heat—in the following form:

$$\int \frac{\delta Q}{T} = -N$$

where Q is heat, T is temperature and N is the “equivalence-value” of all uncompensated transformations involved in a cyclical process. Later, in 1865, Clausius would come to define “equivalence-value” as entropy. On the heels of this definition, that same year, the most famous version of the second law was read in a presentation at the Philosophical Society of Zurich on April 24, in which, in the end of his presentation, Clausius concludes:

The entropy of the universe tends to a maximum.

This statement is the best-known phrasing of the second law. Because of the looseness of its language, *e.g.*, universe, as well as lack of specific conditions, *e.g.*, open, closed, or isolated, many people take this simple statement to mean that the second law of thermodynamics applies virtually to every subject imaginable. This is not true; this statement is only a simplified version of a more extended and precise description.

In terms of time variation, the mathematical statement of the second law for an isolated system undergoing an arbitrary transformation is:

$$\frac{dS}{dt} \geq 0$$

where,

S is the entropy of the system and t is time.

The equality sign applies after equilibration.

An alternative way of formulating of the second law for isolated systems is:

$$\frac{dS}{dt} = \dot{S}_i \text{ with } \dot{S}_i \geq 0$$

with $\dot{S}_i$ the sum of the rate of entropy production by all processes inside the system. The advantage of this formulation is that it shows the effect of the entropy production. The rate of entropy production is a very important concept since, it determines (limits) the efficiency of thermal machines. Multiplied with ambient temperature T_a it gives the so-called dissipated energy $P_{\text{diss}} = T_a \dot{S}_i$. *The expression of the second law for closed systems (so, allowing heat exchange and moving boundaries, but not exchange of matter) is:*

$$\frac{dS}{dt} = \frac{\dot{Q}}{T} + \dot{S}_i \text{ with } \dot{S}_i \geq 0$$

Here,

$\dot{Q}$ is the heat flow into the system

T is the temperature at the point where the heat enters the system.

The equality sign holds in the case that only reversible processes take place inside the system. If irreversible processes take place (which is the case in real systems in operation) the $>$ -sign holds. If heat is supplied to the system at several places we have to take the algebraic sum of the corresponding terms. *For open systems (also allowing exchange of matter):*

$$\frac{dS}{dt} = \frac{\dot{Q}}{T} + \dot{S} + \dot{S}_i \text{ with } \dot{S}_i \geq 0$$

Here $\dot{S}$ is the flow of entropy into the system associated with the flow of matter entering the system. It should not be confused with the time derivative of the entropy. If matter is supplied at several places we have to take the algebraic sum of these contributions.

Statistical Mechanics

Statistical mechanics gives an explanation for the second law by postulating that a material is composed of atoms and molecules which are in constant motion. A particular set of positions and velocities for each particle in

the system is called a microstate of the system and because of the constant motion, the system is constantly changing its microstate. Statistical mechanics postulates that, in equilibrium, each microstate that the system might be in is equally likely to occur, and when this assumption is made, it leads directly to the conclusion that the second law must hold in a statistical sense.

That is, the second law will hold on average, with a statistical variation on the order of $1/N$ where N is the number of particles in the system. For everyday (macroscopic) situations, the probability that the second law will be violated is practically zero. However, for systems with a small number of particles, thermodynamic parameters, including the entropy, may show significant statistical deviations from that predicted by the second law. Classical thermodynamic theory does not deal with these statistical variations.

Living Organisms

There are two principal ways of formulating thermodynamics:

1. Through passages from one state of thermodynamic equilibrium to another, and
2. Through cyclic processes, by which the system is left unchanged, while the total entropy of the surroundings is increased.

These two ways help to understand the processes of life. This topic is mostly beyond the scope of this present article, but has been considered by several authors, such as Erwin Schrödinger, Léon Brillouin and Isaac Asimov. It is also the topic of current research.

To a fair approximation, living organisms may be considered as examples of (b). Approximately, an animal's physical state cycles by the day, leaving the animal nearly unchanged. Animals take in food, water, and oxygen, and, as a result of metabolism, give out breakdown products and heat. Plants take in radiative energy from the sun, which may be regarded as heat, and carbon dioxide and water. They give out oxygen. In this way they grow. Eventually they die, and their remains rot. This can be regarded as a cyclic process. Overall, the sunlight is from a high temperature source, the sun, and its energy is passed to a lower temperature sink, the soil. This is an increase of entropy of the surroundings of the plant. Thus animals and plants obey the second law of thermodynamics, considered in terms of cyclic processes. Simple concepts of efficiency of heat engines are hardly applicable to this problem because they assume closed systems.

From the thermodynamic viewpoint that considers (a), passages from one equilibrium state to another, only a roughly approximate picture appears, because living organisms are never in states of thermodynamic equilibrium. Living organisms must often be considered as open systems, because they take in nutrients and give out waste products.

Thermodynamics of open systems is currently often considered in terms of passages from one state of thermodynamic equilibrium to another, or in terms of flows in the approximation of local thermodynamic equilibrium. The problem for living organisms may be further simplified by the approximation of assuming a steady state with unchanging flows. General principles of entropy production for such approximations are subject to unsettled current debate or research. Nevertheless, ideas derived from this viewpoint on the second law of thermodynamics are enlightening about living creatures.

Gravitational Systems

In systems that do not require for their descriptions the general theory of relativity, bodies always have positive heat capacity, meaning that the temperature rises with energy. Therefore, when energy flows from a high-temperature object to a low-temperature object, the source temperature is decreased while the sink temperature is increased; hence, temperature differences tend to diminish over time. This is not always the case for systems in which the gravitational force is important and the general theory of relativity is required. Such systems can spontaneously

change towards uneven spread of mass and energy. This applies to the universe in large scale, and consequently it may be difficult or impossible to apply the second law to it. Beyond this, the thermodynamics of systems described by the general theory of relativity is beyond the scope of the present article.

Non-equilibrium States

The theory of classical or equilibrium thermodynamics is idealized. A main postulate or assumption, often not even explicitly stated, is the existence of systems in their own internal states of thermodynamic equilibrium. In general, a region of space containing a physical system at a given time, that may be found in nature, is not in thermodynamic equilibrium, read in the most stringent terms. In looser terms, nothing in the entire universe is or has ever been truly in exact thermodynamic equilibrium. For purposes of physical analysis, it is often enough convenient to make an assumption of thermodynamic equilibrium. Such an assumption may rely on trial and error for its justification. If the assumption is justified, it can often be very valuable and useful because it makes available the theory of thermodynamics. Elements of the equilibrium assumption are that a system is observed to be unchanging over an indefinitely long time, and that there are so many particles in a system, that its particulate nature can be entirely ignored. Under such an equilibrium assumption, in general, there are no macroscopically detectable fluctuations. There is an exception, the case of critical states, which exhibit to the naked eye the phenomenon of critical opalescence. For laboratory studies of critical states, exceptionally long observation times are needed. In all cases, the assumption of thermodynamic equilibrium, once made, implies as a consequence that no putative candidate “fluctuation” alters the entropy of the system.

It can easily happen that a physical system exhibits internal macroscopic changes that are fast enough to invalidate the assumption of the constancy of the entropy. Or that a physical system has so few particles that the particulate nature is manifest in observable fluctuations. Then the assumption of thermodynamic equilibrium is to be abandoned. There is no unqualified general definition of entropy for non-equilibrium states.

There are intermediate cases, in which the assumption of local thermodynamic equilibrium is a very good approximation, but strictly speaking it is still an approximation, not theoretically ideal. For non-equilibrium situations in general, it may be useful to consider statistical mechanical definitions of other quantities that may be conveniently called ‘entropy’, but they should not be confused or conflated with thermodynamic entropy properly defined for the second law. These other quantities indeed belong to statistical mechanics, not to thermodynamics, the primary realm of the second law. Even though the applicability of the second law of thermodynamics is limited for non-equilibrium systems, the laws governing such systems are still being discussed. One of the guiding principles for systems which are far from equilibrium is the maximum entropy production principle. It states that a system away from equilibrium evolves in such a way as to maximize entropy production, given present constraints. The physics of macroscopically observable fluctuations is beyond the scope of this article.

Arrow of Time

Thermodynamic Asymmetry in Time. The second law of thermodynamics is a physical law that is not symmetric to reversal of the time direction. This does not conflict with notions that have been observed of the fundamental laws of physics, namely CPT symmetry, since, the second law applies statistically, it is hypothesized, on time-asymmetric *boundary conditions*. The second law has been proposed to supply a partial explanation of the difference between moving forward and backwards in time, such as why the cause precedes the effect (the causal arrow of time).

Irreversibility

Irreversibility in thermodynamic processes is a consequence of the asymmetric character of thermodynamic operations, and not of any internally irreversible microscopic properties of the bodies. Thermodynamic operations

are macroscopic external interventions imposed on the participating bodies, not derived from their internal properties. There are reputed “paradoxes” that arise from failure to recognize this.

Loschmidt's Paradox

Loschmidt's paradox, also known as the reversibility paradox, is the objection that it should not be possible to deduce an irreversible process from the time-symmetric dynamics that describe the microscopic evolution of a macroscopic system. In the opinion of Schrödinger, “It is now quite obvious in what manner you have to reformulate the law of entropy—or for that matter, all other irreversible statements—so that they be capable of being derived from reversible models. You must not speak of one isolated system but at least of two, which you may for the moment consider isolated from the rest of the world, but not always from each other.” The two systems are isolated from each other by the wall, until it is removed by the thermodynamic operation, as envisaged by the law. The thermodynamic operation is externally imposed, not subject to the reversible microscopic dynamical laws that govern the constituents of the systems. It is the cause of the irreversibility. The statement of the law in this present article complies with Schrödinger's advice. The cause–effect relation is logically prior to the second law, not derived from it.

Poincaré Recurrence Theorem

The Poincaré recurrence theorem considers a theoretical microscopic description of an isolated physical system. This may be considered as a model of a thermodynamic system after a thermodynamic operation has removed an internal wall. The system will, after a sufficiently long time, return to a microscopically defined state very close to the initial one. The Poincaré recurrence time is the length of time elapsed until the return. It is exceedingly long, likely longer than the life of the universe, and depends sensitively on the geometry of the wall that was removed by the thermodynamic operation. The recurrence theorem may be perceived as apparently contradicting the second law of thermodynamics. More obviously, however, it is simply a microscopic model of thermodynamic equilibrium in an isolated system formed by removal of a wall between two systems.

For a typical thermodynamical system, the recurrence time is so large (many many times longer than the lifetime of the universe) that, for all practical purposes, one cannot observe the recurrence. One might wish, nevertheless, to imagine that one could wait for the Poincaré recurrence, and then re-insert the wall that was removed by the thermodynamic operation. It is then evident that the appearance of irreversibility is due to the utter unpredictability of the Poincaré recurrence given only that the initial state was one of thermodynamic equilibrium, as is the case in macroscopic thermodynamics. Even if one could wait for it, one has no practical possibility of picking the right instant at which to re-insert the wall. The Poincaré recurrence theorem provides a solution to Loschmidt's paradox. If an isolated thermodynamic system could be monitored over increasingly many multiples of the average Poincaré recurrence time, the thermodynamic behaviour of the system would become invariant under time reversal.

Maxwell's Demon

James Clerk Maxwell imagined one container divided into two parts, *A* and *B*. Both parts are filled with the same gas at equal temperatures and placed next to each other, separated by a wall. Observing the molecules on both sides, an imaginary demon guards a microscopic trapdoor in the wall. When a faster-than-average molecule from *A* flies towards the trapdoor, the demon opens it, and the molecule will fly from *A* to *B*. The average speed of the molecules in *B* will have increased while in *A* they will have slowed down on average. Since, average molecular speed corresponds to temperature, the temperature decreases in *A* and increases in *B*, contrary to the

second law of thermodynamics. One response to this question was suggested in 1929 by Leó Szilárd and later by Léon Brillouin. Szilárd pointed out that a real-life Maxwell's demon would need to have some means of measuring molecular speed, and that the act of acquiring information would require an expenditure of energy. Maxwell's demon repeatedly alters the permeability of the wall between *A* and *B*. It is therefore performing thermodynamic operations on a microscopic scale, not just observing ordinary spontaneous or natural macroscopic thermodynamic processes.

Quotations

- The law that entropy always increases holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations—then so much the worse for Maxwell's equations. If it is found to be contradicted by observation—well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.—*Sir Arthur Stanley Eddington, The Nature of the Physical World (1927)*
- There have been nearly as many formulations of the second law as there have been discussions of it.—*Philosopher/Physicist P.W. Bridgman, (1941)*
- Clausius is the author of the sibyllic utterance, “The energy of the universe is constant; the entropy of the universe tends to a maximum.” The objectives of continuum thermomechanics stop far short of explaining the “universe”, but within that theory we may easily derive an explicit statement in some ways reminiscent of Clausius, but referring only to a modest object: an isolated body of finite size.—*Truesdell, C., Muncaster, R.G. (1980). Fundamentals of Maxwell's Kinetic Theory of a Simple Monatomic Gas, Treated as a Branch of Rational Mechanics, Academic Press, New York, ISBN 0-12-701350-4, p.17.*

THIRD LAW OF THERMODYNAMICS

The third law of thermodynamics is sometimes stated as follows, regarding the properties of systems in equilibrium at absolute zero temperature:

- The entropy of a perfect crystal at absolute zero is exactly equal to zero.

At absolute zero (zero kelvin), the system must be in a state with the minimum possible energy, and the above statement of the third law holds true provided that the perfect crystal has only one minimum energy state.

Entropy is related to the number of accessible microstates, and for a system consisting of many particles, quantum mechanics indicates that there is only one unique state (called the ground state) with minimum energy. If the system does not have a well-defined order (if its order is glassy, for example), then in practice there will remain some finite entropy as the system is brought to very low temperatures as the system becomes locked into a configuration with non-minimal energy. The constant value is called the residual entropy of the system.

The Nernst–Simon statement of the third law of thermodynamics concerns thermodynamic processes at a fixed, low temperature:

- The entropy change associated with any condensed system undergoing a reversible isothermal process approaches zero as the temperature at which it is performed approaches 0 K.

Here a condensed system refers to liquids and solids.

A classical formulation by Nernst (actually a consequence of the Third Law) is:

- It is impossible for any process, no matter how idealized, to reduce the entropy of a system to its absolute-zero value in a finite number of operations.

It was proven in 2017 by Masanes and Oppenheim.

There also exists a formulation of the Third Law which approaches the subject by postulating a specific energy behaviour:

- If the composite of two thermodynamic systems constitutes an isolated system, then any energy exchange in any form between those two systems is bounded.

History

The 3rd law was developed by the chemist Walther Nernst during the years 1906–12, and is therefore often referred to as Nernst's theorem or Nernst's postulate. The third law of thermodynamics states that the entropy of a system at absolute zero is a well-defined constant. This is because a system at zero temperature exists in its ground state, so that its entropy is determined only by the degeneracy of the ground state. In 1912 Nernst stated the law thus: "It is impossible for any procedure to lead to the isotherm $T = 0$ in a finite number of steps."

An alternative version of the third law of thermodynamics as stated by Gilbert N. Lewis and Merle Randall in 1923:

- If the entropy of each element in some (perfect) crystalline state be taken as zero at the absolute zero of temperature, every substance has a finite positive entropy; but at the absolute zero of temperature the entropy may become zero, and does so become in the case of perfect crystalline substances.

This version states not only ΔS will reach zero at 0 K, but S itself will also reach zero as long as the crystal has a ground state with only one configuration. Some crystals form defects which cause a residual entropy. This residual entropy disappears when the kinetic barriers to transitioning to one ground state are overcome.

With the development of statistical mechanics, the third law of thermodynamics (like the other laws) changed from a *fundamental* law (justified by experiments) to a *derived* law (derived from even more basic laws). The basic law from which it is primarily derived is the statistical-mechanics definition of entropy for a large system:

$$S - S_0 = k_B \ln \Omega$$

Where S is entropy, k_B is the Boltzmann constant, and Ω is the number of microstates consistent with the macroscopic configuration. The counting of states is from the reference state of absolute zero, which corresponds to the entropy of S_0 .

Explanation

In simple terms, is used for convenience the third law states that the entropy of a perfect crystal of a pure substance approaches zero as the temperature approaches zero. The alignment of a perfect crystal leaves no ambiguity as to the location and orientation of each part of the crystal. As the energy of the crystal is reduced, the vibrations of the individual atoms are reduced to nothing, and the crystal becomes the same everywhere.

The third law provides an absolute reference point for the determination of entropy at any other temperature. The entropy of a system, determined relative to this zero point, is then the *absolute* entropy of that system. Mathematically, the absolute entropy of any system at zero temperature is the natural log of the number of ground states times Boltzmann's constant $k_B = 1.38 \times 10^{-23} \text{ J K}^{-1}$.

The entropy of a *perfect* crystal lattice as defined by Nernst's theorem is zero provided that its ground state is unique, because $\ln(1) = 0$. If the system is composed of one-billion atoms, all alike, and lie within the matrix of a perfect crystal, the number of permutations of one-billion identical things taken one-billion at a time is $\Omega = 1$.

Hence,;

$$S - S_0 = k_B \ln \Omega = k_B \ln 1 = 0$$

The difference is zero, hence, the initial entropy S_0 can be any selected value so long as all other such calculations include that as the initial entropy. As a result, the initial entropy value of zero is selected $S_0 = 0$ is used for convenience.

$$S - S_0 = S - 0 = 0$$

$$S = 0.$$

Systems with Non-zero Entropy at Absolute Zero

An example of a system which does not have a unique ground state is one whose net spin is a half-integer, for which time-reversal symmetry gives two degenerate ground states. For such systems, the entropy at zero temperature is at least $k_B \ln(2)$ (which is negligible on a macroscopic scale). Some crystalline systems exhibit geometrical frustration, where the structure of the crystal lattice prevents the emergence of a unique ground state. Ground-state helium (unless under pressure) remains liquid. In addition, glasses and solid solutions retain large entropy at 0 K, because they are large collections of nearly degenerate states, in which they become trapped out of equilibrium. Another example of a solid with many nearly-degenerate ground states, trapped out of equilibrium, is ice Ih, which has “proton disorder”. For the entropy at absolute zero to be zero, the magnetic moments of a perfectly ordered crystal must themselves be perfectly ordered; from an entropic perspective, this can be considered to be part of the definition of a “perfect crystal”. Only ferromagnetic, antiferromagnetic, and diamagnetic materials can satisfy this condition. However, ferromagnetic materials do not, in fact, have zero entropy at zero temperature, because the spins of the unpaired electrons are all aligned and this gives a ground-state spin degeneracy. Materials that remain paramagnetic at 0 K, by contrast, may have many nearly-degenerate ground states (for example, in a spin glass), or may retain dynamic disorder (a quantum spin liquid).

SYSTEM MODELS

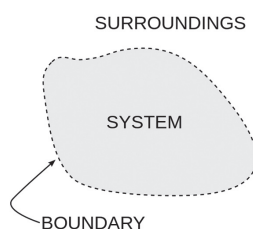


Fig. A diagram of a generic thermodynamic system.

An important concept in thermodynamics is the thermodynamic system, which is a precisely defined region of the universe under study. Everything in the universe except the system is called the *surroundings*. A system is separated from the remainder of the universe by a *boundary* which may be a physical boundary or notional, but which by convention defines a finite volume. Exchanges of work, heat, or matter between the system and the surroundings take place across this boundary. In practice, the boundary of a system is simply an imaginary dotted line drawn around a volume within which is going to be a change in the internal energy of that volume. Anything that passes across the boundary that effects a change in the internal energy of the system needs to be accounted for in the energy balance equation.

The volume can be the region surrounding a single atom resonating energy, such as Max Planck defined in 1900; it can be a body of steam or air in a steam engine, such as Sadi Carnot defined in 1824; it can be the body of a tropical cyclone, such as Kerry Emanuel theorized in 1986 in the field of atmospheric thermodynamics; it could also be just one nuclide (*i.e.*, a system of quarks) as hypothesized in quantum thermodynamics, or the event horizon of a black hole. Boundaries are of four types: fixed, movable, real, and imaginary. For example, in an engine, a fixed boundary means the piston is locked at its position, within which a constant volume process might occur. If the piston is allowed to move that boundary is movable while the cylinder and cylinder head boundaries are fixed. For closed systems, boundaries are real while for open systems boundaries are often imaginary. In the case of a jet engine, a fixed imaginary boundary might be assumed at the intake of the engine,

fixed boundaries along the surface of the case and a second fixed imaginary boundary across the exhaust nozzle. Generally, thermodynamics distinguishes three classes of systems, defined in terms of what is allowed to cross their boundaries:

Table. Interactions of Thermodynamic Systems.

Type of system	Mass flow	Work	Heat
Open	✓	✓	✓
Closed	✗	✓	✓
Thermally isolated	✗	✓	✗
Mechanically isolated	✗	✗	✓
Isolated	✗	✗	✗

As time passes in an isolated system, internal differences of pressures, densities, and temperatures tend to even out. A system in which all equalizing processes have gone to completion is said to be in a state of thermodynamic equilibrium. Once in thermodynamic equilibrium, a system's properties are, by definition, unchanging in time. Systems in equilibrium are much simpler and easier to understand than are systems which are not in equilibrium. Often, when analysing a dynamic thermodynamic process, the simplifying assumption is made that each intermediate state in the process is at equilibrium, producing thermodynamic processes which develop so slowly as to allow each intermediate step to be an equilibrium state and are said to be reversible processes.

STATES AND PROCESSES

When a system is at equilibrium under a given set of conditions, it is said to be in a definite thermodynamic state. The state of the system can be described by a number of state quantities that do not depend on the process by which the system arrived at its state. They are called intensive variables or extensive variables according to how they change when the size of the system changes. The properties of the system can be described by an equation of state which specifies the relationship between these variables. State may be thought of as the instantaneous quantitative description of a system with a set number of variables held constant.

A thermodynamic process may be defined as the energetic evolution of a thermodynamic system proceeding from an initial state to a final state. It can be described by process quantities. Typically, each thermodynamic process is distinguished from other processes in energetic character according to what parameters, such as temperature, pressure, or volume, etc., are held fixed; Furthermore, it is useful to group these processes into pairs, in which each variable held constant is one member of a conjugate pair.

Several commonly studied thermodynamic processes are:

- Adiabatic process: occurs without loss or gain of energy by heat
- Isenthalpic process: occurs at a constant enthalpy
- Isentropic process: a reversible adiabatic process, occurs at a constant entropy
- Isobaric process: occurs at constant pressure
- Isochoric process: occurs at constant volume (also called isometric/isovolumetric)
- Isothermal process: occurs at a constant temperature
- Steady state process: occurs without a change in the internal energy.

INSTRUMENTATION

There are two types of thermodynamic instruments, the meter and the reservoir. A thermodynamic meter is any device which measures any parameter of a thermodynamic system. In some cases, the thermodynamic parameter is actually defined in terms of an idealized measuring instrument. For example, the zeroth law states that if two bodies are in thermal equilibrium with a third body, they are also in thermal equilibrium with each other. This principle, as noted by James Maxwell in 1872, asserts that it is possible to measure temperature. An

idealized thermometer is a sample of an ideal gas at constant pressure. From the ideal gas law $pV=nRT$, the volume of such a sample can be used as an indicator of temperature; in this manner it defines temperature. Although pressure is defined mechanically, a pressure-measuring device, called a barometer may also be constructed from a sample of an ideal gas held at a constant temperature. A calorimeter is a device which is used to measure and define the internal energy of a system.

A thermodynamic reservoir is a system which is so large that its state parameters are not appreciably altered when it is brought into contact with the system of interest. When the reservoir is brought into contact with the system, the system is brought into equilibrium with the reservoir. For example, a pressure reservoir is a system at a particular pressure, which imposes that pressure upon the system to which it is mechanically connected. The Earth's atmosphere is often used as a pressure reservoir. If ocean water is used to cool a power plant, the ocean is often a temperature reservoir in the analysis of the power plant cycle.

CONJUGATE VARIABLES

The central concept of thermodynamics is that of energy, the ability to do work. By the First Law, the total energy of a system and its surroundings is conserved. Energy may be transferred into a system by heating, compression, or addition of matter, and extracted from a system by cooling, expansion, or extraction of matter. In mechanics, for example, energy transfer equals the product of the force applied to a body and the resulting displacement. Conjugate variables are pairs of thermodynamic concepts, with the first being akin to a "force" applied to some thermodynamic system, the second being akin to the resulting "displacement," and the product of the two equalling the amount of energy transferred.

The common conjugate variables are:

- Pressure-volume (the mechanical parameters);
- Temperature-entropy (thermal parameters);
- Chemical potential-particle number (material parameters).

POTENTIALS

Thermodynamic potentials are different quantitative measures of the stored energy in a system. Potentials are used to measure the energy changes in systems as they evolve from an initial state to a final state. The potential used depends on the constraints of the system, such as constant temperature or pressure. For example, the Helmholtz and Gibbs energies are the energies available in a system to do useful work when the temperature and volume or the pressure and temperature are fixed, respectively.

The five most well known potentials are:

Name	Symbol	Formula	Natural variables
Internal energy	U	$\int \left(TdS - pdV + \sum_i \mu_i dN_i \right)$	S, V, {N _i }
Helmholtz free energy	F	$U - TS$	T, V, {N _i }
Enthalpy	H	$U + pdV$	S, p, {N _i }
Gibbs free energy	G	$U + pV - TS$	T, p, {N _i }
Landau Potential (Grand potential)	Ω, Φ_G	$U - TS - \sum_i \mu_i N_i$	T, V, {μ _i }

Kinetics

KINETICS (PHYSICS)

In physics and engineering, kinetics is the branch of classical mechanics that is concerned with the relationship between motion and its causes, specifically, forces and torques. Since, the mid-20th century, the term “dynamics” (or “analytical dynamics”) has largely superseded “kinetics” in physics textbooks, though the term is still used in engineering. In plasma physics, kinetics refers to the study of continua in velocity space. This is usually in the context of non-thermal (non-Maxwellian) velocity distributions, or processes that perturb thermal distributions. These “kinetic plasmas” cannot be adequately described with fluid equations.

BIOMECHANICS AND KINETICS

One of the major components of kinetics is analyzing one's center of pressure. Center of pressure (COP) is often tested using a force plate in a biomechanics lab. Many scientists don't always have access to these expensive and unique equipment. Force plates can be any from thousands to tens of thousands of dollars. Scientists and students today have turned to an alternative way, by using Nintendo Wii Balance Board system as a force plate. One 2014 San Diego State University study has shown a correlation of $r = 0.99$ for the center of pressure measured using the Balance Board compared to a gold standard measure. The students expect their results to find that ankle bracing will reduce sway and decrease ankle instability and falls. Postural sway is the measurement of the time and distance a subject spends away from an ideal center of pressure.

Measurement of a subject's center of pressure has been used as a method of quantifying postural stability. Investigators have theorized that if ankle injuries cause differentiation and functional instability, a subject's postural sway should be altered. Bracing has been known to provide cutaneous afferent feedback in maintaining postural control. The feedback follows the same neurological pathways in which the external ankle brace enhances the replication of the joint position. Although some may argue that braces have a negative impact on sport performance, the effectiveness of their mechanical stability makes them of high priority and necessity. Due to the high incidence of ankle injury, the control of stability and body sway remain of most importance in reduction of injury and better quality of life. The term *kinetics* is also used to refer to chemical kinetics, particularly in chemical physics and physical chemistry. In such uses, a qualifier is often used or implied, for example: “physical kinetics”, “crystal growth kinetics”, and so on.

RIGID BODY DYNAMICS

Rigid-body dynamics studies the movement of systems of interconnected bodies under the action of external forces. The assumption that the bodies are rigid, which means that they do not deform under the action of applied forces, simplifies the analysis by reducing the parameters that describe the configuration of the system

to the translation and rotation of reference frames attached to each body. This excludes bodies that display fluid, highly elastic, and plastic behaviour. The dynamics of a rigid body system is described by the laws of kinematics and by the application of Newton's second law (kinetics) or their derivative form Lagrangian mechanics. The solution of these equations of motion provides a description of the position, the motion and the acceleration of the individual components of the system and overall the system itself, as a function of time. The formulation and solution of rigid body dynamics is an important tool in the computer simulation of mechanical systems.

Rigid Body in Three Dimensions

Orientation or Attitude Descriptions

Several methods to describe orientations of a rigid body in three dimensions have been developed. They are summarized in the following sections.

Euler Angles

The first attempt to represent an orientation is attributed to Leonhard Euler. He imagined three reference frames that could rotate one around the other, and realized that by starting with a fixed reference frame and performing three rotations, he could get any other reference frame in the space (using two rotations to fix the vertical axis and other to fix the other two axes). The values of these three rotations are called Euler angles.

Tait–Bryan Angles

These are three angles, also known as yaw, pitch and roll, Navigation angles and Cardan angles. Mathematically they constitute a set of six possibilities inside the twelve possible sets of Euler angles, the ordering being the one best used for describing the orientation of a vehicle such as an airplane. In aerospace engineering they are usually referred to as Euler angles.

Orientation Vector

Euler also realized that the composition of two rotations is equivalent to a single rotation about a different fixed axis (Euler's rotation theorem). Therefore, the composition of the former three angles has to be equal to only one rotation, whose axis was complicated to calculate until matrices were developed. Based on this fact he introduced a vectorial way to describe any rotation, with a vector on the rotation axis and module equal to the value of the angle. Therefore, any orientation can be represented by a rotation vector (also called Euler vector) that leads to it from the reference frame. When used to represent an orientation, the rotation vector is commonly called orientation vector, or attitude vector. A similar method, called axis-angle representation, describes a rotation or orientation using a unit vector aligned with the rotation axis, and a separate value to indicate the angle.

Orientation Matrix

With the introduction of matrices the Euler theorems were rewritten. The rotations were described by orthogonal matrices referred to as rotation matrices or direction cosine matrices. When used to represent an orientation, a rotation matrix is commonly called orientation matrix, or attitude matrix. The above-mentioned Euler vector is the eigenvector of a rotation matrix (a rotation matrix has a unique real eigenvalue). The product of two rotation matrices is the composition of rotations. Therefore, as before, the orientation can be given as the rotation from the initial frame to achieve the frame that we want to describe. The configuration space of a non-symmetrical object in n -dimensional space is $SO(n) \times \mathbb{R}^n$. Orientation may be visualized by attaching a basis of tangent vectors to an object. The direction in which each vector points determines its orientation.

Orientation Quaternion

Another way to describe rotations is using rotation quaternions, also called versors. They are equivalent to rotation matrices and rotation vectors. With respect to rotation vectors, they can be more easily converted to and from matrices. When used to represent orientations, rotation quaternions are typically called orientation quaternions or attitude quaternions.

Newton's Second Law in Three Dimensions

To consider rigid body dynamics in three-dimensional space, Newton's second law must be extended to define the relationship between the movement of a rigid body and the system of forces and torques that act on it. Newton formulated his second law for a particle as, "The change of motion of an object is proportional to the force impressed and is made in the direction of the straight line in which the force is impressed." Because Newton generally referred to mass times velocity as the "motion" of a particle, the phrase "change of motion" refers to the mass times acceleration of the particle, and so this law is usually written as

$$F = ma,$$

where F is understood to be the only external force acting on the particle, m is the mass of the particle, and a is its acceleration vector. The extension of Newton's second law to rigid bodies is achieved by considering a rigid system of particles.

Applications

- For the analysis of robotic systems
- For the biomechanical analysis of animals, humans or humanoid systems
- For the analysis of space objects
- For the understanding of strange motions of rigid bodies.
- For the design and development of dynamics-based sensors like gyroscopic sensors etc.,
- For the design and development of various stability enhancement applications in automobiles etc.,
- For improving the graphics of video games which involves rigid bodies.

CHEMICAL KINETICS

Chemical kinetics, also known as reaction kinetics, is the study of rates of chemical processes. Chemical kinetics includes investigations of how different experimental conditions can influence the speed of a chemical reaction and yield information about the reaction's mechanism and transition states, as well as the construction of mathematical models that can describe the characteristics of a chemical reaction.

HISTORY

In 1864, Peter Waage and Cato Guldberg pioneered the development of chemical kinetics by formulating the law of mass action, which states that the speed of a chemical reaction is proportional to the quantity of the reacting substances. Van't Hoff studied chemical dynamics and published in 1884 his famous "Etudes de dynamique chimique". In 1901 he was awarded by the first Nobel Prize in Chemistry "in recognition of the extraordinary services he has rendered by the discovery of the laws of chemical dynamics and osmotic pressure in solutions". After van 't Hoff, chemical kinetics deals with the experimental determination of reaction rates from which rate laws and rate constants are derived. Relatively simple rate laws exist for zero order reactions (for which reaction rates are independent of concentration), first order reactions, and second order reactions, and can be derived for others.

Elementary reactions follow the law of mass action, but the rate law of stepwise reactions has to be derived by combining the rate laws of the various elementary steps, and can become rather complex. In consecutive reactions, the rate-determining step often determines the kinetics. In consecutive first order reactions, a steady state approximation can simplify the rate law. The activation energy for a reaction is experimentally determined through the Arrhenius equation and the Eyring equation. The main factors that influence the reaction rate include: the physical state of the reactants, the concentrations of the reactants, the temperature at which the reaction occurs, and whether or not any catalysts are present in the reaction.

Gorban and Yablonsky have suggested that the history of chemical dynamics can be divided into three eras. The first is the van 't Hoff wave searching for the general laws of chemical reactions and relating kinetics to thermodynamics. The second may be called the Semenov—Hinshelwood wave with emphasis on reaction mechanisms, especially for chain reactions. The third is associated with Aris and the detailed mathematical description of chemical reaction networks.

FACTORS AFFECTING REACTION RATE

Nature of the Reactants

Depending upon what substances are reacting, the reaction rate varies. Acid/base reactions, the formation of salts, and ion exchange are fast reactions. When covalent bond formation takes place between the molecules and when large molecules are formed, the reactions tend to be very slow. Nature and strength of bonds in reactant molecules greatly influence the rate of its transformation into products.

Physical State

The physical state (solid, liquid, or gas) of a reactant is also an important factor of the rate of change. When reactants are in the same phase, as in aqueous solution, thermal motion brings them into contact. However, when they are in different phases, the reaction is limited to the interface between the reactants. Reaction can occur only at their area of contact; in the case of a liquid and a gas, at the surface of the liquid. Vigorous shaking and stirring may be needed to bring the reaction to completion. This means that the more finely divided a solid or liquid reactant the greater its surface area per unit volume and the more contact it with the other reactant, thus the faster the reaction. To make an analogy, for example, when one starts a fire, one uses wood chips and small branches—one does not start with large logs right away. In organic chemistry, on water reactions are the exception to the rule that homogeneous reactions take place faster than heterogeneous reactions.

Surface Area of Solids

In a solid, only those particles that are at the surface can be involved in a reaction. Crushing a solid into smaller parts means that more particles are present at the surface, and the frequency of collisions between these and reactant particles increases, and so reaction occurs more rapidly. For example, Sherbet (powder) is a mixture of very fine powder of malic acid (a weak organic acid) and sodium hydrogen carbonate. On contact with the saliva in the mouth, these chemicals quickly dissolve and react, releasing carbon dioxide and providing for the fizzy sensation.

Also, fireworks manufacturers modify the surface area of solid reactants to control the rate at which the fuels in fireworks are oxidised, using this to create different effects. For example, finely divided aluminium confined in a shell explodes violently. If larger pieces of aluminium are used, the reaction is slower and sparks are seen as pieces of burning metal are ejected.

Concentration

The reactions are due to collisions of reactant species. The frequency with which the molecules or ions collide depends upon their concentrations. The more crowded the molecules are, the more likely they are to collide and react with one another. Thus, an increase in the concentrations of the reactants will usually result in the corresponding increase in the reaction rate, while a decrease in the concentrations will usually have a reverse effect. For example, combustion will occur more rapidly in pure oxygen than in air (21% oxygen). The rate equation shows the detailed dependence of the reaction rate on the concentrations of reactants and other species present. Different mathematical forms are possible depending on the reaction mechanism. The actual rate equation for a given reaction is determined experimentally and provides information about the reaction mechanism.

Temperature

Temperature usually has a major effect on the rate of a chemical reaction. Molecules at a higher temperature have more thermal energy. Although collision frequency is greater at higher temperatures, this alone contributes only a very small proportion to the increase in rate of reaction. Much more important is the fact that the proportion of reactant molecules with sufficient energy to react (energy greater than activation energy: $E > E_a$) is significantly higher and is explained in detail by the Maxwell–Boltzmann distribution of molecular energies. The ‘rule of thumb’ that the rate of chemical reactions doubles for every 10 °C temperature rise is a common misconception. This may have been generalized from the special case of biological systems, where the α (temperature coefficient) is often between 1.5 and 2.5. A reaction’s kinetics can also be studied with a temperature jump approach. This involves using a sharp rise in temperature and observing the relaxation time of the return to equilibrium. A particularly useful form of temperature jump apparatus is a shock tube, which can rapidly jump a gas’s temperature by more than 1000 degrees.

Catalysts

A catalyst is a substance that alters the rate of a chemical reaction but remains chemically unchanged afterwards. The catalyst increases the rate of the reaction by providing a different reaction mechanism to occur with a lower activation energy. In autocatalysis a reaction product is itself a catalyst for that reaction leading to positive feedback. Proteins that act as catalysts in biochemical reactions are called enzymes. Michaelis–Menten kinetics describe the rate of enzyme mediated reactions. A catalyst does not affect the position of the equilibrium, as the catalyst speeds up the backward and forward reactions equally. In certain organic molecules, specific substituents can have an influence on reaction rate in neighbouring group participation.

Pressure

Increasing the pressure in a gaseous reaction will increase the number of collisions between reactants, increasing the rate of reaction. This is because the activity of a gas is directly proportional to the partial pressure of the gas. This is similar to the effect of increasing the concentration of a solution. In addition to this straightforward mass-action effect, the rate coefficients themselves can change due to pressure. The rate coefficients and products of many high-temperature gas-phase reactions change if an inert gas is added to the mixture; variations on this effect are called fall-off and chemical activation. These phenomena are due to exothermic or endothermic reactions occurring faster than heat transfer, causing the reacting molecules to have non-thermal energy distributions (non-Boltzmann distribution). Increasing the pressure increases the heat transfer rate between the reacting molecules and the rest of the system, reducing this effect.

Condensed-phase rate coefficients can also be affected by (very high) pressure; this is a completely different effect than fall-off or chemical-activation. It is often studied using diamond anvils. A reaction's kinetics can also be studied with a pressure jump approach. This involves making fast changes in pressure and observing the relaxation time of the return to equilibrium.

EXPERIMENTAL METHODS

The experimental determination of reaction rates involves measuring how the concentrations of reactants or products change over time. For example, the concentration of a reactant can be measured by spectrophotometry at a wavelength where no other reactant or product in the system absorbs light. For reactions which take at least several minutes, it is possible to start the observations after the reactants have been mixed at the temperature of interest.

Fast Reactions

For faster reactions, the time required to mix the reactants and bring them to a specified temperature may be comparable or longer than the half-life of the reaction.

Special methods to start fast reactions without slow mixing step include:

- Stopped flow methods, which can reduce the mixing time to the order of a millisecond
- Chemical relaxation methods such as temperature jump and pressure jump, in which a pre-mixed system initially at equilibrium is perturbed by rapid heating or depressurization so that it is no longer at equilibrium, and the relaxation back to equilibrium is observed. For example, this method has been used to study the neutralization $\text{H}_3\text{O}^+ + \text{OH}^-$ with a half-life of 1 μs or less under ordinary conditions.
- Flash photolysis, in which a laser pulse produces highly excited species such as free radicals, whose reactions are then studied.

EQUILIBRIUM

While chemical kinetics is concerned with the rate of a chemical reaction, thermodynamics determines the extent to which reactions occur. In a reversible reaction, chemical equilibrium is reached when the rates of the forward and reverse reactions are equal (the principle of detailed balance) and the concentrations of the reactants and products no longer change. This is demonstrated by, for example, the Haber–Bosch process for combining nitrogen and hydrogen to produce ammonia. Chemical clock reactions such as the Belousov–Zhabotinsky reaction demonstrate that component concentrations can oscillate for a long time before finally attaining the equilibrium.

FREE ENERGY

In general terms, the free energy change (ΔG) of a reaction determines whether a chemical change will take place, but kinetics describes how fast the reaction is. A reaction can be very exothermic and have a very positive entropy change but will not happen in practice if the reaction is too slow. If a reactant can produce two different products, the thermodynamically most stable one will in general form, except in special circumstances when the reaction is said to be under kinetic reaction control. The Curtin–Hammett principle applies when determining the product ratio for two reactants interconverting rapidly, each going to a different product. It is possible to make predictions about reaction rate constants for a reaction from free-energy relationships. The kinetic isotope effect is the difference in the rate of a chemical reaction when an atom in one of the reactants is replaced by one

of its isotopes. Chemical kinetics provides information on residence time and heat transfer in a chemical reactor in chemical engineering and the molar mass distribution in polymer chemistry.

APPLICATIONS AND MODELS

The mathematical models that describe chemical reaction kinetics provide chemists and chemical engineers with tools to better understand and describe chemical processes such as food decomposition, microorganism growth, stratospheric ozone decomposition, and the chemistry of biological systems. These models can also be used in the design or modification of chemical reactors to optimize product yield, more efficiently separate products, and eliminate environmentally harmful by-products. When performing catalytic cracking of heavy hydrocarbons into gasoline and light gas, for example, kinetic models can be used to find the temperature and pressure at which the highest yield of heavy hydrocarbons into gasoline will occur. Chemical Kinetics is frequently validated and explored through modeling in specialized packages as a function of ordinary differential equation-solving (ODE-solving) and curve-fitting.

ENZYME KINETICS

Enzyme kinetics is the study of the chemical reactions that are catalysed by enzymes. In enzyme kinetics, the reaction rate is measured and the effects of varying the conditions of the reaction are investigated. Studying an enzyme's kinetics in this way can reveal the catalytic mechanism of this enzyme, its role in metabolism, how its activity is controlled, and how a drug or an agonist might inhibit the enzyme.

Enzymes are usually protein molecules that manipulate other molecules—the enzymes' substrates. These target molecules bind to an enzyme's active site and are transformed into products through a series of steps known as the enzymatic mechanism,



These mechanisms can be divided into single-substrate and multiple-substrate mechanisms. Kinetic studies on enzymes that only bind one substrate, such as triosephosphate isomerase, aim to measure the affinity with which the enzyme binds this substrate and the turnover rate. Some other examples of enzymes are phosphofructokinase and hexokinase, both of which are important for cellular respiration (glycolysis).

When enzymes bind multiple substrates, such as dihydrofolate reductase (shown right), enzyme kinetics can also show the sequence in which these substrates bind and the sequence in which products are released. An example of enzymes that bind a single substrate and release multiple products are proteases, which cleave one protein substrate into two polypeptide products. Others join two substrates together, such as DNA polymerase linking a nucleotide to DNA. Although these mechanisms are often a complex series of steps, there is typically one *rate-determining step* that determines the overall kinetics. This rate-determining step may be a chemical reaction or a conformational change of the enzyme or substrates, such as those involved in the release of product(s) from the enzyme.

Knowledge of the enzyme's structure is helpful in interpreting kinetic data. For example, the structure can suggest how substrates and products bind during catalysis; what changes occur during the reaction; and even the role of particular amino acid residues in the mechanism. Some enzymes change shape significantly during the mechanism; in such cases, it is helpful to determine the enzyme structure with and without bound substrate analogues that do not undergo the enzymatic reaction. Not all biological catalysts are protein enzymes; RNA-based catalysts such as ribozymes and ribosomes are essential to many cellular functions, such as RNA splicing and translation. The main difference between ribozymes and enzymes is that RNA catalysts are composed of nucleotides, whereas enzymes are composed of amino acids. Ribozymes also perform a more limited set of reactions, although their reaction mechanisms and kinetics can be analysed and classified by the same methods.

General Principles

The reaction catalysed by an enzyme uses exactly the same reactants and produces exactly the same products as the uncatalysed reaction. Like other catalysts, enzymes do not alter the position of equilibrium between substrates and products. However, unlike uncatalysed chemical reactions, enzyme-catalysed reactions display saturation kinetics. For a given enzyme concentration and for relatively low substrate concentrations, the reaction rate increases linearly with substrate concentration; the enzyme molecules are largely free to catalyse the reaction, and increasing substrate concentration means an increasing rate at which the enzyme and substrate molecules encounter one another. However, at relatively high substrate concentrations, the reaction rate asymptotically approaches the theoretical maximum; the enzyme active sites are almost all occupied by substrates resulting in saturation, and the reaction rate is determined by the intrinsic turnover rate of the enzyme. The substrate concentration midway between these two limiting cases is denoted by K_M . Thus, K_M is the substrate concentration value in which the substrate concentration is reaching halfway of the maximum reaction velocity. The two most important kinetic properties of an enzyme are how easily the enzyme becomes saturated with a particular substrate, and the maximum rate it can achieve. Knowing these properties suggests what an enzyme might do in the cell and can show how the enzyme will respond to changes in these conditions.

Enzyme Assays

Enzyme assays are laboratory procedures that measure the rate of enzyme reactions. Since, enzymes are not consumed by the reactions they catalyse, enzyme assays usually follow changes in the concentration of either substrates or products to measure the rate of reaction. There are many methods of measurement. Spectrophotometric assays observe change in the absorbance of light between products and reactants; radiometric assays involve the incorporation or release of radioactivity to measure the amount of product made over time. Spectrophotometric assays are most convenient since, they allow the rate of the reaction to be measured continuously. Although radiometric assays require the removal and counting of samples (*i.e.*, they are discontinuous assays) they are usually extremely sensitive and can measure very low levels of enzyme activity. An analogous approach is to use mass spectrometry to monitor the incorporation or release of stable isotopes as substrate is converted into product. The most sensitive enzyme assays use lasers focused through a microscope to observe changes in single enzyme molecules as they catalyse their reactions. These measurements either use changes in the fluorescence of cofactors during an enzyme's reaction mechanism, or of fluorescent dyes added onto specific sites of the protein to report movements that occur during catalysis. These studies are providing a new view of the kinetics and dynamics of single enzymes, as opposed to traditional enzyme kinetics, which observes the average behaviour of populations of millions of enzyme molecules.

An example progress curve for an enzyme assay is shown above. The enzyme produces product at an initial rate that is approximately linear for a short period after the start of the reaction. As the reaction proceeds and substrate is consumed, the rate continuously slows (so long as substrate is not still at saturating levels). To measure the initial (and maximal) rate, enzyme assays are typically carried out while the reaction has progressed only a few percent towards total completion. The length of the initial rate period depends on the assay conditions and can range from milliseconds to hours. However, equipment for rapidly mixing liquids allows fast kinetic measurements on initial rates of less than one second. These very rapid assays are essential for measuring pre-steady-state kinetics, which are discussed below. Most enzyme kinetics studies concentrate on this initial, approximately linear part of enzyme reactions. However, it is also possible to measure the complete reaction curve and fit this data to a non-linear rate equation. This way of measuring enzyme reactions is called progress-curve analysis. This approach is useful as an alternative to rapid kinetics when the initial rate is too fast to measure accurately.

Single-substrate Reactions

Enzymes with single-substrate mechanisms include isomerases such as triosephosphate isomerase or bisphosphoglycerate mutase, intramolecular lyases such as adenylate cyclase and the hammerhead ribozyme, an RNA lyase. However, some enzymes that only have a single substrate do not fall into this category of mechanisms. Catalase is an example of this, as the enzyme reacts with a first molecule of hydrogen peroxide substrate, becomes oxidised and is then reduced by a second molecule of substrate. Although a single substrate is involved, the existence of a modified enzyme intermediate means that the mechanism of catalase is actually a ping-pong mechanism, a type of mechanism that is discussed in the *Multi-substrate reactions* section below.

Multi-substrate Reactions

Multi-substrate reactions follow complex rate equations that describe how the substrates bind and in what sequence. The analysis of these reactions is much simpler if the concentration of substrate A is kept constant and substrate B varied. Under these conditions, the enzyme behaves just like a single-substrate enzyme and a plot of v by $[S]$ gives apparent K_M and V_{max} constants for substrate B. If a set of these measurements is performed at different fixed concentrations of A, these data can be used to work out what the mechanism of the reaction is. For an enzyme that takes two substrates A and B and turns them into two products P and Q, there are two types of mechanism: ternary complex and ping-pong.

Ternary-complex Mechanisms

In these enzymes, both substrates bind to the enzyme at the same time to produce an EAB ternary complex. The order of binding can either be random (in a random mechanism) or substrates have to bind in a particular sequence (in an ordered mechanism). When a set of v by $[S]$ curves (fixed A, varying B) from an enzyme with a ternary-complex mechanism are plotted in a Lineweaver–Burk plot, the set of lines produced will intersect. Enzymes with ternary-complex mechanisms include glutathione *S*-transferase, dihydrofolate reductase and DNA polymerase. The following links show short animations of the ternary-complex mechanisms of the enzymes dihydrofolate reductase and DNA polymerase.

Ping-pong Mechanisms

As shown on the right, enzymes with a ping-pong mechanism can exist in two states, E and a chemically modified form of the enzyme E^* ; this modified enzyme is known as an intermediate. In such mechanisms, substrate A binds, changes the enzyme to E^* by, for example, transferring a chemical group to the active site, and is then released. Only after the first substrate is released can substrate B bind and react with the modified enzyme, regenerating the unmodified E form. When a set of v by $[S]$ curves (fixed A, varying B) from an enzyme with a ping-pong mechanism are plotted in a Lineweaver–Burk plot, a set of parallel lines will be produced.

This is called a secondary plot. Enzymes with ping-pong mechanisms include some oxidoreductases such as thioredoxin peroxidase, transferases such as acylneuraminidase and cytidyltransferase and serine proteases such as trypsin and chymotrypsin. Serine proteases are a very common and diverse family of enzymes, including digestive enzymes (trypsin, chymotrypsin, and elastase), several enzymes of the blood clotting cascade and many others. In these serine proteases, the E^* intermediate is an acyl-enzyme species formed by the attack of an active site serine residue on a peptide bond in a protein substrate. A short animation showing the mechanism of chymotrypsin is linked [here](#).

Non-Michaelis–Menten Kinetics

Some enzymes produce a sigmoid v by $[S]$ plot, which often indicates cooperative binding of substrate to the active site. This means that the binding of one substrate molecule affects the binding of subsequent substrate molecules. This behaviour is most common in multimeric enzymes with several interacting active sites. Here, the mechanism of cooperation is similar to that of hemoglobin, with binding of substrate to one active site altering the affinity of the other active sites for substrate molecules. Positive cooperativity occurs when binding of the first substrate molecule *increases* the affinity of the other active sites for substrate. Negative cooperativity occurs when binding of the first substrate *decreases* the affinity of the enzyme for other substrate molecules.

Allosteric enzymes include mammalian tyrosyl tRNA-synthetase, which shows negative cooperativity, and bacterial aspartate transcarbamoylase and phosphofructokinase, which show positive cooperativity. Cooperativity is surprisingly common and can help regulate the responses of enzymes to changes in the concentrations of their substrates. Positive cooperativity makes enzymes much more sensitive to $[S]$ and their activities can show large changes over a narrow range of substrate concentration. Conversely, negative cooperativity makes enzymes insensitive to small changes in $[S]$. The Hill equation (biochemistry) is often used to describe the degree of cooperativity quantitatively in non-Michaelis–Menten kinetics. The derived Hill coefficient n measures how much the binding of substrate to one active site affects the binding of substrate to the other active sites. A Hill coefficient of <1 indicates negative cooperativity and a coefficient of >1 indicates positive cooperativity.

Pre-steady-state Kinetics

In the first moment after an enzyme is mixed with substrate, no product has been formed and no intermediates exist. The study of the next few milliseconds of the reaction is called pre-steady-state kinetics. Pre-steady-state kinetics is therefore concerned with the formation and consumption of enzyme–substrate intermediates (such as ES or E^*) until their steady-state concentrations are reached. This approach was first applied to the hydrolysis reaction catalysed by chymotrypsin. Often, the detection of an intermediate is a vital piece of evidence in investigations of what mechanism an enzyme follows. For example, in the ping–pong mechanisms that are shown above, rapid kinetic measurements can follow the release of product P and measure the formation of the modified enzyme intermediate E^* . In the case of chymotrypsin, this intermediate is formed by an attack on the substrate by the nucleophilic serine in the active site and the formation of the acyl-enzyme intermediate.

In the figure to the right, the enzyme produces E^* rapidly in the first few seconds of the reaction. The rate then slows as steady state is reached. This rapid burst phase of the reaction measures a single turnover of the enzyme. Consequently, the amount of product released in this burst, shown as the intercept on the y-axis of the graph, also gives the amount of functional enzyme which is present in the assay.

Chemical Mechanism

An important goal of measuring enzyme kinetics is to determine the chemical mechanism of an enzyme reaction, *i.e.*, the sequence of chemical steps that transform substrate into product. The kinetic approaches discussed above will show at what rates intermediates are formed and inter-converted, but they cannot identify exactly what these intermediates are. Kinetic measurements taken under various solution conditions or on slightly modified enzymes or substrates often shed light on this chemical mechanism, as they reveal the rate-determining step or intermediates in the reaction. For example, the breaking of a covalent bond to a hydrogen atom is a common rate-determining step. Which of the possible hydrogen transfers is rate determining can be shown by measuring the kinetic effects of substituting each hydrogen by deuterium, its stable isotope. The rate

will change when the critical hydrogen is replaced, due to a primary kinetic isotope effect, which occurs because bonds to deuterium are harder to break than bonds to hydrogen. It is also possible to measure similar effects with other isotope substitutions, such as $^{13}\text{C}/^{12}\text{C}$ and $^{18}\text{O}/^{16}\text{O}$, but these effects are more subtle.

Isotopes can also be used to reveal the fate of various parts of the substrate molecules in the final products. For example, it is sometimes difficult to discern the origin of an oxygen atom in the final product; since, it may have come from water or from part of the substrate. This may be determined by systematically substituting oxygen's stable isotope ^{18}O into the various molecules that participate in the reaction and checking for the isotope in the product. The chemical mechanism can also be elucidated by examining the kinetics and isotope effects under different pH conditions, by altering the metal ions or other bound cofactors, by site-directed mutagenesis of conserved amino acid residues, or by studying the behaviour of the enzyme in the presence of analogues of the substrate(s).

Mechanisms of Catalysis

The favoured model for the enzyme–substrate interaction is the induced fit model. This model proposes that the initial interaction between enzyme and substrate is relatively weak, but that these weak interactions rapidly induce conformational changes in the enzyme that strengthen binding. These conformational changes also bring catalytic residues in the active site close to the chemical bonds in the substrate that will be altered in the reaction. Conformational changes can be measured using circular dichroism or dual polarisation interferometry. After binding takes place, one or more mechanisms of catalysis lower the energy of the reaction's transition state by providing an alternative chemical pathway for the reaction. Mechanisms of catalysis include catalysis by bond strain; by proximity and orientation; by active-site proton donors or acceptors; covalent catalysis and quantum tunnelling. Enzyme kinetics cannot prove which modes of catalysis are used by an enzyme. However, some kinetic data can suggest possibilities to be examined by other techniques. For example, a ping-pong mechanism with burst-phase pre-steady-state kinetics would suggest covalent catalysis might be important in this enzyme's mechanism. Alternatively, the observation of a strong pH effect on V_{max} but not K_{m} might indicate that a residue in the active site needs to be in a particular ionisation state for catalysis to occur.

History

In 1902 Victor Henri proposed a quantitative theory of enzyme kinetics, but at the time the experimental significance of the hydrogen ion concentration was not yet recognized. After Peter Lauritz Sørensen had defined the logarithmic pH-scale and introduced the concept of buffering in 1909 the German chemist Leonor Michaelis and Dr. Maud Leonora Menten (a postdoctoral researcher in Michaelis's lab at the time) repeated Henri's experiments and confirmed his equation, which is now generally referred to as Michaelis-Menten kinetics (sometimes also *Henri-Michaelis-Menten kinetics*). Their work was further developed by G. E. Briggs and J. B. S. Haldane, who derived kinetic equations that are still widely considered today a starting point in modeling enzymatic activity.

The major contribution of the Henri-Michaelis-Menten approach was to think of enzyme reactions in two stages. In the first, the substrate binds reversibly to the enzyme, forming the enzyme-substrate complex. This is sometimes called the Michaelis complex. The enzyme then catalyzes the chemical step in the reaction and releases the product. The kinetics of many enzymes is adequately described by the simple Michaelis-Menten model, but all enzymes have internal motions that are not accounted for in the model and can have significant contributions to the overall reaction kinetics. This can be modeled by introducing several Michaelis-Menten pathways that are connected with fluctuating rates, which is a mathematical extension of the basic Michaelis-Menten mechanism.

PHARMACOKINETICS

Pharmacokinetics sometimes abbreviated as PK, is a branch of pharmacology dedicated to determining the fate of substances administered to a living organism. The substances of interest include any chemical xenobiotic such as: pharmaceutical drugs, pesticides, food additives, cosmetics, etc., It attempts to analyze chemical metabolism and to discover the fate of a chemical from the moment that it is administered up to the point at which it is completely eliminated from the body. Pharmacokinetics is the study of how an organism affects a drug, whereas pharmacodynamics (PD) is the study of how the drug affects the organism. Both together influence dosing, benefit, and adverse effects.

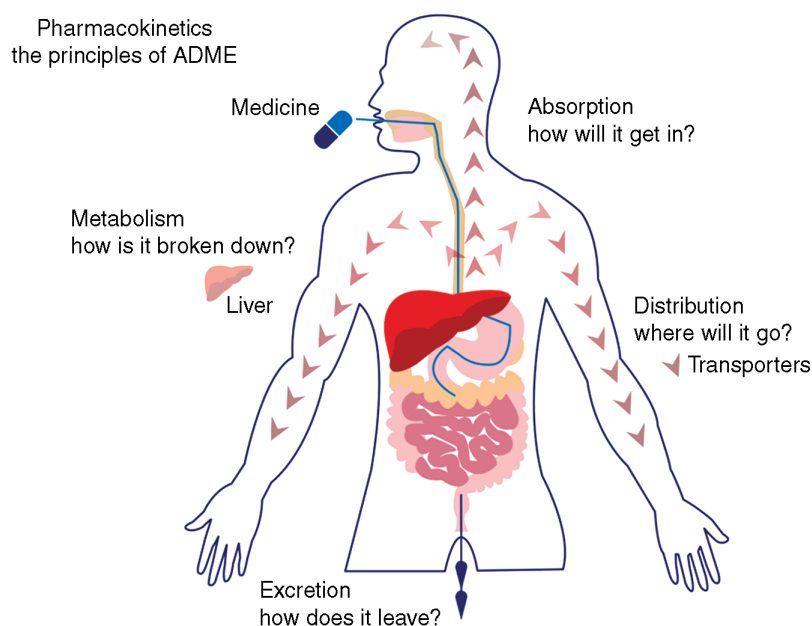


Fig. Pharmacokinetics.

Pharmacokinetics describes how the body affects a specific xenobiotic/chemical after administration through the mechanisms of absorption and distribution, as well as the metabolic changes of the substance in the body (e.g., by metabolic enzymes such as cytochrome P450 or glucuronosyltransferase enzymes), and the effects and routes of excretion of the metabolites of the drug. Pharmacokinetic properties of chemicals are affected by the route of administration and the dose of administered drug. These may affect the absorption rate.

Models have been developed to simplify conceptualization of the many processes that take place in the interaction between an organism and a chemical substance. One of these, the multi-compartmental model, is the most commonly used approximations to reality; however, the complexity involved in adding parameters with that modelling approach means that *monocompartmental models* and above all *two compartmental models* are the most-frequently used. *The various compartments that the model is divided into are commonly referred to as the ADME scheme (also referred to as LADME if liberation is included as a separate step from absorption):*

- Liberation—the process of release of a drug from the pharmaceutical formulation.
- Absorption—the process of a substance entering the blood circulation.
- Distribution—the dispersion or dissemination of substances throughout the fluids and tissues of the body.
- Metabolism (or biotransformation, or inactivation)—the recognition by the organism that a foreign substance is present and the irreversible transformation of parent compounds into daughter metabolites.
- Excretion—the removal of the substances from the body. In rare cases, some drugs irreversibly accumulate in body tissue.

The two phases of metabolism and excretion can also be grouped together under the title elimination. The study of these distinct phases involves the use and manipulation of basic concepts in order to understand the process dynamics. For this reason in order to fully comprehend the *kinetics* of a drug it is necessary to have detailed knowledge of a number of factors such as: the properties of the substances that act as excipients, the characteristics of the appropriate biological membranes and the way that substances can cross them, or the characteristics of the enzyme reactions that inactivate the drug.

All these concepts can be represented through mathematical formulas that have a corresponding graphical representation. The use of these models allows an understanding of the characteristics of a molecule, as well as how a particular drug will behave given information regarding some of its basic characteristics such as its acid dissociation constant (pKa), bioavailability and solubility, absorption capacity and distribution in the organism. The model outputs for a drug can be used in industry (for example, in calculating bioequivalence when designing generic drugs) or in the clinical application of pharmacokinetic concepts. Clinical pharmacokinetics provides many performance guidelines for effective and efficient use of drugs for human-health professionals and in veterinary medicine.

PHARMACOKINETIC MODELS

Pharmacokinetic modelling is performed by noncompartmental or compartmental methods. Noncompartmental methods estimate the exposure to a drug by estimating the area under the curve of a concentration-time graph. Compartmental methods estimate the concentration-time graph using kinetic models. Noncompartmental methods are often more versatile in that they do not assume any specific compartmental model and produce accurate results also acceptable for bioequivalence studies. The final outcome of the transformations that a drug undergoes in an organism and the rules that determine this fate depend on a number of interrelated factors. A number of functional models have been developed in order to simplify the study of pharmacokinetics. These models are based on a consideration of an organism as a number of related compartments. The simplest idea is to think of an organism as only one homogenous compartment. This *monocompartmental model* presupposes that blood plasma concentrations of the drug are a true reflection of the drug's concentration in other fluids or tissues and that the elimination of the drug is directly proportional to the drug's concentration in the organism (first order kinetics).

However, these models do not always truly reflect the real situation within an organism. For example, not all body tissues have the same blood supply, so the distribution of the drug will be slower in these tissues than in others with a better blood supply. In addition, there are some tissues (such as the brain tissue) that present a real barrier to the distribution of drugs, that can be breached with greater or lesser ease depending on the drug's characteristics. If these relative conditions for the different tissue types are considered along with the rate of elimination, the organism can be considered to be acting like two compartments: one that we can call the *central compartment* that has a more rapid distribution, comprising organs and systems with a well-developed blood supply; and a *peripheral compartment* made up of organs with a lower blood flow. Other tissues, such as the brain, can occupy a variable position depending on a drug's ability to cross the barrier that separates the organ from the blood supply.

This *two compartment model* will vary depending on which compartment elimination occurs in. The most common situation is that elimination occurs in the central compartment as the liver and kidneys are organs with a good blood supply. However, in some situations it may be that elimination occurs in the peripheral compartment or even in both. This can mean that there are three possible variations in the two compartment model, which still do not cover all possibilities. This model may not be applicable in situations where some of the enzymes responsible for metabolizing the drug become saturated, or where an active elimination mechanism is present that is independent of the drug's plasma concentration. In the real world each tissue will have its own distribution

characteristics and none of them will be strictly linear. If we label the drug's volume of distribution within the organism Vd_F and its volume of distribution in a tissue Vd_T the former will be described by an equation that takes into account all the tissues that act in different ways, that is:

$$Vd_F = Vd_{T1} + Vd_{T2} + Vd_{T3} + \dots + Vd_{Tn}$$

This represents the *multi-compartment model* with a number of curves that express complicated equations in order to obtain an overall curve. A number of computer programs have been developed to plot these equations. However complicated and precise this model may be, it still does not truly represent reality despite the effort involved in obtaining various distribution values for a drug. This is because the concept of distribution volume is a relative concept that is not a true reflection of reality. The choice of model therefore comes down to deciding which one offers the lowest margin of error for the drug involved.

Noncompartmental Analysis

Noncompartmental PK analysis is highly dependent on estimation of total drug exposure. Total drug exposure is most often estimated by area under the curve (AUC) methods, with the trapezoidal rule (numerical integration) the most common method. Due to the dependence on the length of x in the trapezoidal rule, the area estimation is highly dependent on the blood/plasma sampling schedule. That is, the closer time points are, the closer the trapezoids reflect the actual shape of the concentration-time curve.

Compartmental Analysis

Compartmental PK analysis uses kinetic models to describe and predict the concentration-time curve. PK compartmental models are often similar to kinetic models used in other scientific disciplines such as chemical kinetics and thermodynamics. The advantage of compartmental over some noncompartmental analyses is the ability to predict the concentration at any time. The disadvantage is the difficulty in developing and validating the proper model. Compartment-free modelling based on curve stripping does not suffer this limitation. The simplest PK compartmental model is the one-compartmental PK model with IV bolus administration and first-order elimination. The most complex PK models (called PBPK models) rely on the use of physiological information to ease development and validation.

Single-compartment Model

Linear pharmacokinetics is so-called because the graph of the relationship between the various factors involved (dose, blood plasma concentrations, elimination, etc.) gives a straight line or an approximation to one. For drugs to be effective they need to be able to move rapidly from blood plasma to other body fluids and tissues.

The change in concentration over time can be expressed as,

$$C = C_{\text{initial}} \times e^{-k_{\text{el}} \times t}$$

Multi-compartmental Models

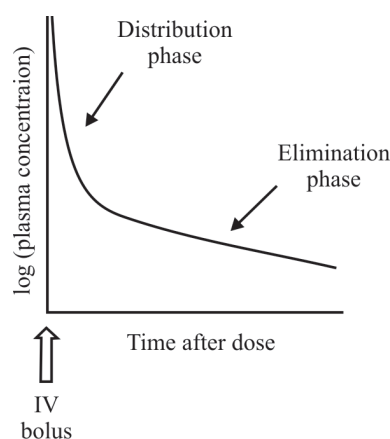
The graph for the non-linear relationship between the various factors is represented by a curve; the relationships between the factors can then be found by calculating the dimensions of different areas under the curve. The models used in *non-linear pharmacokinetics* are largely based on Michaelis–Menten kinetics.

A reaction's factors of non-linearity include the following:

- *Multiphasic absorption: Drugs injected intravenously are removed from the plasma through two primary mechanisms:*

- Distribution to body tissues and
- Metabolism + excretion of the drugs.

The resulting decrease of the drug's plasma concentration follows a biphasic pattern.



- Plasma drug concentration vs time after an IV dose
 - *Alpha phase*: An initial phase of rapid decrease in plasma concentration. The decrease is primarily attributed to drug distribution from the central compartment (circulation) into the peripheral compartments (body tissues). This phase ends when a pseudo-equilibrium of drug concentration is established between the central and peripheral compartments.
 - *Beta phase*: A phase of gradual decrease in plasma concentration after the alpha phase. The decrease is primarily attributed to drug metabolism and excretion.
 - Additional phases (gamma, delta, etc.) are sometimes seen.
- A drug's characteristics make a clear distinction between tissues with high and low blood flow.
- *Enzymatic saturation*: When the dose of a drug whose elimination depends on biotransformation is increased above a certain threshold the enzymes responsible for its metabolism become saturated. The drug's plasma concentration will then increase disproportionately and its elimination will no longer be constant.
- *Induction or enzymatic inhibition*: Some drugs have the capacity to inhibit or stimulate their own metabolism, in negative or positive feedback reactions. As occurs with fluvoxamine, fluoxetine and phenytoin. As larger doses of these pharmaceuticals are administered the plasma concentrations of the unmetabolized drug increases and the elimination half-life increases. It is therefore necessary to adjust the dose or other treatment parameters when a high dosage is required.
- The kidneys can also establish active elimination mechanisms for some drugs, independent of plasma concentrations.

It can therefore be seen that non-linearity can occur because of reasons that affect the entire pharmacokinetic sequence: Absorption, distribution, metabolism and elimination.

Variable Volume in Time Models

Variable volume pharmacokinetic models can be drug centered models that imply a volume of drug distribution to be that volume in which the drug is distributed at that elapsed time following drug administration. Another possibility occurs when the body volume is changing in time, which would occur, for example, during dialysis when the volume in which a drug can be distributed is itself changing in time.

LADME

A number of phases occur once the drug enters into contact with the organism, these are described using the acronym LADME:

- Liberation of the active substance from the delivery system,
- Absorption of the active substance by the organism,
- Distribution through the blood plasma and different body tissues,
- Metabolism that is inactivation of the xenobiotic substance, and finally
- Excretion or elimination of the substance or the products of its metabolism.

Some textbooks combine the first two phases as the drug is often administered in an active form, which means that there is no liberation phase. Others include a phase that combines distribution, metabolism and excretion into a disposition phase. Other authors include the drug's toxicological aspect in what is known as *ADME-Tox* or *ADMET*. Each of the phases is subject to physico-chemical interactions between a drug and an organism, which can be expressed mathematically. Pharmacokinetics is therefore based on mathematical equations that allow the prediction of a drug's behaviour and which place great emphasis on the relationships between drug plasma concentrations and the time elapsed since, the drug's administration.

ANALYSIS

Bioanalytical Methods

Bioanalytical methods are necessary to construct a concentration-time profile. Chemical techniques are employed to measure the concentration of drugs in biological matrix, most often plasma. Proper bioanalytical methods should be selective and sensitive. For example, microscale thermophoresis can be used to quantify how the biological matrix/liquid affects the affinity of a drug to its target.

Mass Spectrometry

Pharmacokinetics is often studied using mass spectrometry because of the complex nature of the matrix (often plasma or urine) and the need for high sensitivity to observe concentrations after a low dose and a long time period. The most common instrumentation used in this application is LC-MS with a triple quadrupole mass spectrometer. Tandem mass spectrometry is usually employed for added specificity. Standard curves and internal standards are used for quantitation of usually a single pharmaceutical in the samples. The samples represent different time points as a pharmaceutical is administered and then metabolized or cleared from the body. Blank samples taken before administration are important in determining background and ensuring data integrity with such complex sample matrices. Much attention is paid to the linearity of the standard curve; however it is common to use curve fitting with more complex functions such as quadratics since, the response of most mass spectrometers is not linear across large concentration ranges. There is currently considerable interest in the use of very high sensitivity mass spectrometry for microdosing studies, which are seen as a promising alternative to animal experimentation.

POPULATION PHARMACOKINETICS

Population pharmacokinetics is the study of the sources and correlates of variability in drug concentrations among individuals who are the target patient population receiving clinically relevant doses of a drug of interest.

Certain patient demographic, pathophysiological, and therapeutical features, such as body weight, excretory and metabolic functions, and the presence of other therapies, can regularly alter dose-concentration relationships. For example, steady-state concentrations of drugs eliminated mostly by the kidney are usually greater in patients suffering from renal failure than they are in patients with normal renal function receiving the same drug dosage. Population pharmacokinetics seeks to identify the measurable pathophysiologic factors that cause changes in the dose-concentration relationship and the extent of these changes so that, if such changes are associated with clinically significant shifts in the therapeutic index, dosage can be appropriately modified. An advantage of population pharmacokinetic modelling is its ability to analyse sparse data sets (sometimes only one concentration measurement per patient is available).

CLINICAL PHARMACOKINETICS

Clinical pharmacokinetics (arising from the clinical use of population pharmacokinetics) is the direct application to a therapeutic situation of knowledge regarding a drug's pharmacokinetics and the characteristics of a population that a patient belongs to (or can be ascribed to). An example is the relaunch of the use of ciclosporin as an immunosuppressor to facilitate organ transplant. The drug's therapeutic properties were initially demonstrated, but it was almost never used after it was found to cause nephrotoxicity in a number of patients. However, it was then realized that it was possible to individualize a patient's dose of ciclosporin by analysing the patients plasmatic concentrations (pharmacokinetic monitoring). This practice has allowed this drug to be used again and has facilitated a great number of organ transplants. Clinical monitoring is usually carried out by determination of plasma concentrations as this data is usually the easiest to obtain and the most reliable.

The main reasons for determining a drug's plasma concentration include:

- Narrow therapeutic range (difference between toxic and therapeutic concentrations)
- High toxicity
- High risk to life.

Drugs where pharmacokinetic monitoring is recommended include:

Medications for which monitoring is recommended			
Antiepileptic medication + Phenytoin + Carbamazepine + Valproic acid + Lamotrigine + Ethosuximide + Phenobarbital + Primidone	Cardioactive medication + Digoxin + Lidocaine	Immunosuppressor medication + Ciclosporin + Tacrolimus + Sirolimus + Everolimus + Mycophenolate	Antibiotic medication + Gentamicin + Tobramycin + Amikacin + Vancomycin
Bronchodilator medication + Theophylline	Cytostatic medication + Methotrexate + 5-Fluorouracil + Irinotecan	* Antiviral (HIV) medication + Efavirenz + Tenofovir + Ritonavir	Coagulation factors + Factor VIII, + Factor IX, + Factor VIIa, + Factor XI

KINESIOLOGY

Kinesiology is the scientific study of human or non-human body movement. Kinesiology addresses physiological, biomechanical, and psychological dynamic principles and mechanisms of movement. Applications of kinesiology to human health (*i.e.*, human kinesiology) include biomechanics and orthopedics; strength and conditioning; sport psychology; methods of rehabilitation, such as physical and occupational therapy; and sport and exercise. Studies of human and animal motion include measures from motion tracking systems,

electrophysiology of muscle and brain activity, various methods for monitoring physiological function, and other behavioral and cognitive research techniques.

BASICS

Kinesiology is the study of human and nonhuman animal-body movements, performance, and function by applying the sciences of biomechanics, anatomy, physiology, psychology, and neuroscience. Applications of kinesiology in human-health include physical education teacher, rehabilitation, health and safety, health promotion, workplaces, sport and exercise industries. A bachelor's degree in kinesiology can provide strong preparation for graduate study in biomedical research, as well as in professional programs, such as medicine.

Whereas the term “kinesiologist” is neither a licensed nor professional designation in the United States nor most countries (with the exception of Canada), individuals with training in this area can teach physical education, provide consulting services, conduct research and develop policies related to rehabilitation, human motor performance, ergonomics, and occupational health and safety. In North America, kinesiologists may study to earn a Bachelor of Science, Master of Science, or Doctorate of Philosophy degree in Kinesiology or a Bachelor of Kinesiology degree, while in Australia or New Zealand, they are often conferred an Applied Science (Human Movement) degree (or higher). Many doctoral level faculty in North American kinesiology programs received their doctoral training in related disciplines, such as neuroscience, mechanical engineering, psychology, and physiology. The world's first kinesiology department was launched in 1967 at the University of Waterloo, Canada.

PRINCIPLES

Adaptation through Exercise

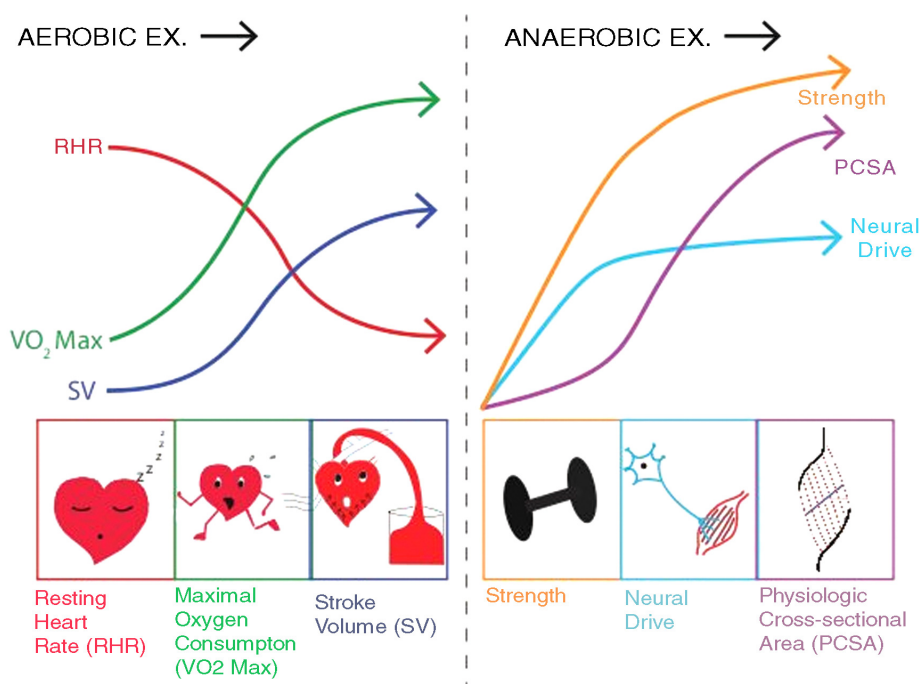


Fig. Summary of long-term adaptations to regular aerobic and anaerobic exercise.

Aerobic exercise can cause several central cardiovascular adaptations, including an increase in stroke volume (SV) and maximal aerobic capacity ($\text{VO}_2 \text{ max}$), as well as a decrease in resting heart rate (RHR). Long-term adaptations to resistance training, the most common form of anaerobic exercise, include muscular hypertrophy, an increase in the physiological cross-sectional area (PCSA) of muscle(s), and an increase in neural drive, both of which lead to increased muscular strength. Neural adaptations begin more quickly and plateau prior to the hypertrophic response.

Adaptation through exercise is a key principle of kinesiology that relates to improved fitness in athletes as well as health and wellness in clinical populations. Exercise is a simple and established intervention for many movement disorders and musculoskeletal conditions due to the neuroplasticity of the brain and the adaptability of the musculoskeletal system. Therapeutic exercise has been shown to improve neuromotor control and motor capabilities in both normal and pathological populations. There are many different types of exercise interventions that can be applied in kinesiology to athletic, normal, and clinical populations. Aerobic exercise interventions help to improve cardiovascular endurance. Anaerobic strength training programs can increase muscular strength, power, and lean body mass. Decreased risk of falls and increased neuromuscular control can be attributed to balance intervention programs. Flexibility programs can increase functional range of motion and reduce the risk of injury.

As a whole, exercise programs can reduce symptoms of depression and risk of cardiovascular and metabolic diseases. Additionally, they can help to improve quality of life, sleeping habits, immune system function, and body composition. The study of the physiological responses to physical exercise and their therapeutic applications is known as exercise physiology, which is an important area of research within kinesiology.

Neuroplasticity

Neuroplasticity is also a key scientific principle used in kinesiology to describe how movement and changes in the brain are related. The human brain adapts and acquires new motor skills based on this principle, which includes both adaptive and maladaptive brain changes.

Adaptive Plasticity

Recent empirical evidence indicates the significant impact of physical activity on brain function; for example, greater amounts of physical activity are associated with enhanced cognitive function in older adults. The effects of physical activity can be distributed throughout the whole brain, such as higher gray matter density and white matter integrity after exercise training, and/or on specific brain areas, such as greater activation in prefrontal cortex and hippocampus. Neuroplasticity is also the underlying mechanism of skill acquisition. For example, after long-term training, pianists showed greater gray matter density in sensorimotor cortex and white matter integrity in the internal capsule compared to non-musicians.

Maladaptive Plasticity

Maladaptive plasticity is defined as neuroplasticity with negative effects or detrimental consequences in behaviour. Movement abnormalities may occur among individuals with and without brain injuries due to abnormal remodeling in central nervous system. Learned non-use is an example commonly seen among patients with brain damage, such as stroke. Patients with stroke learned to suppress paretic limb movement after unsuccessful experience in paretic hand use; this may cause decreased neuronal activation at adjacent areas of the infarcted motor cortex. There are many types of therapies that are designed to overcome maladaptive plasticity in clinic and research, such as constraint-induced movement therapy (CIMT), body weight support treadmill training (BWSTT) and virtual reality therapy. These interventions are shown to enhance motor function in paretic limbs and stimulate cortical reorganization in patients with brain damage.

Motor Redundancy

Motor redundancy is a widely used concept in kinesiology and motor control which states that, for any task the human body can perform, there are effectively an unlimited number of ways the nervous system could achieve that task.

This redundancy appears at multiple levels in the chain of motor execution:

- Kinematic redundancy means that for a desired location of the endpoint (*e.g.*, the hand or finger), there are many configurations of the joints that would produce the same endpoint location in space.
- Muscle redundancy means that the same net joint torque could be generated by many different relative contributions of individual muscles.
- Motor unit redundancy means that for the same net muscle force could be generated by many different relative contributions of motor units within that muscle.

The concept of motor redundancy is explored in numerous studies, usually with the goal of describing the relative contribution of a set of motor elements (*e.g.*, muscles) in various human movements, and how these contributions can be predicted from a comprehensive theory. Two distinct (but not incompatible) theories have emerged for how the nervous system coordinates redundant elements: simplification and optimization. In the simplification theory, complex movements and muscle actions are constructed from simpler ones, often known as primitives or synergies, resulting in a simpler system for the brain to control. In the optimization theory, motor actions arise from the minimization of a control parameter, such as the energetic cost of movement or errors in movement performance.

SCOPE OF PRACTICE

In Canada, kinesiology is a professional designation as well as an area of study. In the province of Ontario the scope has been officially defined as, “the assessment of human movement and performance and its rehabilitation and management to maintain, rehabilitate or enhance movement and performance” Kinesiologists work in a variety of roles as health professionals. They work as rehabilitation providers in hospitals, clinics and private settings working with populations needing care for musculoskeletal, cardiac and neurological conditions. They provide rehabilitation to persons injured at work and in vehicular accidents. Kinesiologists also work as functional assessment specialists, exercise therapists, ergonomists, return to work specialists, case managers and medical legal evaluators. They can be found in hospital, long term care, clinic, work, and community settings. Additionally, kinesiology is applied in areas of health and fitness for all levels of athletes, but more often found with training of elite athletes.

HEALTH SERVICE

- *Health Promotion:* Kinesiologists working in the health promotion industry work with individuals to enhance the health, fitness, and well-being of the individual. Kinesiologists can be found working in fitness facilities, personal training/corporate wellness facilities, and industry.
- *Clinical/Rehabilitation:* Kinesiologists work with individuals with disabling conditions to assist in regaining their optimal physical function. They work with individuals in their home, fitness facilities, rehabilitation clinics, and at the worksite. They also work alongside physiotherapists and occupational therapists.
- *Ergonomics:* Kinesiologists work in industry to assess suitability of design of workstations and provide suggestions for modifications and assistive devices.
- *Health and Safety:* Kinesiologists are involved in consulting with industry to identify hazards and provide recommendations and solutions to optimize the health and safety of workers.

- *Disability Management/Case Coordination:* Kinesiologists recommend and provide a plan of action to return an injured individual to their optimal function in all aspects of life.
- *Management/Research/Administration/Health and Safety:* Kinesiologists frequently fulfill roles in all above areas, perform research, and manage businesses.
- *Health Education:* Kinesiologists working in health education teach people about behaviors that promote wellness. They develop and implement strategies to improve the health of individuals and communities. Community health workers collect data and discuss health concerns with members of specific populations or communities.
- *Athletic Training:* Kinesiologists working in athletic training work in cooperation with physicians. Athletic trainers strive to prevent athletes from suffering injuries, diagnose them if they have suffered an injury and apply the appropriate treatment.
- *Athletic Coaches and Scouts:* Kinesiologists who pursue a career as an athletic coach develop new talent and guide an athlete's progress in a specific sport. They teach amateur or professional athletes the skills they need to succeed at their sport. Many coaches are also involved in scouting. Scouts look for new players and evaluate their skills and likelihood for success at the college, amateur, or professional level.
- *Physical Education Teacher:* Kinesiologists working as physical education teachers are responsible for teaching fitness, sports and health. They help students stay both mentally and physically fit by teaching them to make healthy choices.

TECHNOLOGY IN KINESIOLOGY

Motion capture technology has application in measuring human movement, and thus kinesiology. Historically, motion capture labs have recorded high fidelity data. While accurate and credible, these systems can come at high capital and operational costs. Modern-day systems have increased accessibility to mocap technology.

COT ANALYSIS

C₀t analysis, a technique based on the principles of DNA reassociation kinetics, is a biochemical technique that measures how much repetitive DNA is in a DNA sample such as a genome. It is used to study genome structure and organization and has also been used to simplify the sequencing of genomes that contain large amounts of repetitive sequence.

PROCEDURE

The procedure involves heating a sample of genomic DNA until it denatures into the single stranded-form, and then slowly cooling it, so the strands can pair back together. While the sample is cooling, measurements are taken of how much of the DNA is base paired at each temperature. The amount of single and double-stranded DNA is measured by rapidly diluting the sample, which slows reassociation, and then binding the DNA to a hydroxylapatite column. The column is first washed with a low concentration of sodium phosphate buffer, which elutes the single-stranded DNA, and then with high concentrations of phosphate, which elutes the double stranded DNA. The amount of DNA in these two solutions is then measured using a spectrophotometer.

ANALYSIS

Since, a sequence of single-stranded DNA needs to find its complementary strand to reform a double helix, common sequences renature more rapidly than rare sequences. Indeed, the rate at which a sequence will

reassociate is proportional to the number of copies of that sequence in the DNA sample. A sample with a highly-repetitive sequence will renature rapidly, while complex sequences will renature slowly. However, instead of simply measuring the percentage of double-stranded DNA versus time, the amount of renaturation is measured relative to a C_0t value. The C_0t value is the product of C_0 (the initial concentration of DNA), t (time in seconds), and a constant that depends on the concentration of cations in the buffer. Repetitive DNA will renature at low C_0t values, while complex and unique DNA sequences will renature at high C_0t values. The fast renaturation of the repetitive DNA is because of the availability of numerous complementary sequences.

APPLICATION TO GENOME SEQUENCING

C_0t filtration is a technique that uses the principles of DNA renaturation kinetics to separate the repetitive DNA sequences that dominate many eukaryotic genomes from “gene-rich” single/low-copy sequences. This allows DNA sequencing to concentrate on the parts of the genome that are most informative and interesting, which will speed up the discovery of new genes and make the process more efficient.

HISTORY

It was first developed and utilized by Roy Britten and his colleagues at the Carnegie Institution of Washington in the 1960s. Of particular note, it was through C_0t analysis that the redundant (repetitive) nature of eukaryotic genomes was first discovered. However, it wasn't until the breakthrough DNA reassociation kinetics experiments of Britten and his colleagues that it was shown that not all DNA coded for genes. In fact, their experiments demonstrated that the majority of eukaryotic genomic DNA is composed of repetitive, non-coding elements.

Energy Metabolism

METABOLISM

Definition

Metabolism is the set of life-sustaining chemical transformations within the cells of living organisms. The three main purposes of metabolism are the conversion of food/fuel to energy to run cellular processes, the conversion of food/fuel to building blocks for proteins, lipids, nucleic acids, and some carbohydrates, and the elimination of nitrogenous wastes. These enzyme-catalyzed reactions allow organisms to grow and reproduce, maintain their structures, and respond to their environments. The word metabolism can also refer to the sum of all chemical reactions that occur in living organisms, including digestion and the transport of substances into and between different cells, in which case the set of reactions within the cells is called intermediary metabolism or intermediate metabolism. Metabolism is usually divided into two categories: catabolism, the *breaking down* of organic matter for example, the breaking down of glucose to pyruvate, by cellular respiration, and anabolism, the *building up* of components of cells such as proteins and nucleic acids. Usually, breaking down releases energy and building up consumes energy. The chemical reactions of metabolism are organized into metabolic pathways, in which one chemical is transformed through a series of steps into another chemical, by a sequence of enzymes. Enzymes are crucial to metabolism because they allow organisms to drive desirable reactions that require energy that will not occur by themselves, by coupling them to spontaneous reactions that release energy. Enzymes act as catalysts that allow the reactions to proceed more rapidly. Enzymes also allow the regulation of metabolic pathways in response to changes in the cell's environment or to signals from other cells.

The metabolic system of a particular organism determines which substances it will find nutritious and which poisonous. For example, some prokaryotes use hydrogen sulfide as a nutrient, yet this gas is poisonous to animals. The speed of metabolism, the metabolic rate, influences how much food an organism will require, and also affects how it is able to obtain that food. A striking feature of metabolism is the similarity of the basic metabolic pathways and components between even vastly different species. For example, the set of carboxylic acids that are best known as the intermediates in the citric acid cycle are present in all known organisms, being found in species as diverse as the unicellular bacterium *Escherichia coli* and huge multicellular organisms like elephants. These striking similarities in metabolic pathways are likely due to their early appearance in evolutionary history, and their retention because of their efficacy.

TYPES

Metabolism refers to the dynamic changes of the molecules within a cell, especially those small molecules used as sources of energy and as precursors for the synthesis of proteins, lipids, and nucleic acids. These reactions

occur in the steady state rather than all at once. *Steady state* refers to dynamic equilibrium, or homeostasis, where the individual molecules change but the rate at which they are made equals the rate at which they are destroyed. Concentrations of individual molecules in metabolic reactions are therefore kept relatively constant, while any individual molecules are present only for a brief time. Metabolism therefore is said to be an open chemical system. Metabolic reactions can be catabolic (directed towards the breakdown of larger molecules to produce energy), or anabolic (directed towards the energy consuming synthesis of cellular components from smaller molecules).

ATP

Adenosine triphosphate (ATP) is a nucleotide (also called a nucleoside triphosphate) and is a small molecule used in cells as a coenzyme. It is often referred to as the “molecular unit of currency” of intracellular energy transfer. ATP transports chemical energy within cells for metabolism. Most cellular functions need energy in order to be carried out: synthesis of proteins, synthesis of membranes, movement of the cell, cellular division, transport of various solutes etc., ATP is the molecule that carries energy to the place where the energy is needed. When ATP breaks into adenosine diphosphate (ADP) and phosphate (P_i), the breakdown of the last covalent link of phosphate (a simple- PO_4) liberates energy that is used in reactions where it is needed.

It is one of the end products of photophosphorylation, aerobic respiration, and fermentation, and is used by enzymes and structural proteins in many cellular processes, including biosynthetic reactions, motility, and cell division. One molecule of ATP contains adenine, ribose, and three phosphate groups, and it is produced by a wide variety of enzymes, including ATP synthase, from adenosine diphosphate (ADP) or adenosine monophosphate (AMP) and various phosphate group donors. Substrate-level phosphorylation, oxidative phosphorylation in cellular respiration, and photophosphorylation in photosynthesis are three major mechanisms of ATP biosynthesis. Metabolic processes that use ATP as an energy source convert it back into its precursors. ATP is therefore continuously recycled in organisms: the human body, which on average contains only 250 grams (8.8 oz) of ATP, turns over its own body weight equivalent in ATP each day.

ATP is used as a substrate in signal transduction pathways by kinases that phosphorylate proteins and lipids. It is also used by adenylate cyclase, which uses ATP to produce the second messenger molecule cyclic AMP. The ratio between ATP and AMP is used as a way for a cell to sense how much energy is available and control the metabolic pathways that produce and consume ATP. Apart from its roles in signaling and energy metabolism, ATP is also incorporated into nucleic acids by polymerases in the process of transcription. ATP is the neurotransmitter believed to signal the sense of taste.

The structure of this molecule consists of a purine base (adenine) attached by the 9 nitrogen atom to the 12 carbon atom of a pentose sugar (ribose). Three phosphate groups are attached at the 5 carbon atom of the pentose sugar. It is the addition and removal of these phosphate groups that inter-convert ATP, ADP and AMP. When ATP is used in DNA synthesis, the ribose sugar is first converted to deoxyribose by ribonucleotide reductase. ATP was discovered in 1929 by Karl Lohmann, and independently by Cyrus Fiske and Yellapragada Subbarow of Harvard Medical School, but its correct structure was not determined until some years later. It was proposed to be the intermediary molecule between energy-yielding and energy-requiring reactions in cells by Fritz Albert Lipmann in 1941. It was first artificially synthesized by Alexander Todd in 1948.

PHYSICAL AND CHEMICAL PROPERTIES

ATP consists of adenosine—composed of an adenine ring and a ribose sugar—and three phosphate groups (triphosphate). The phosphoryl groups, starting with the group closest to the ribose, are referred to as the alpha

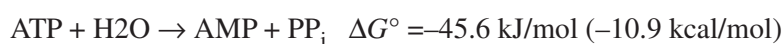
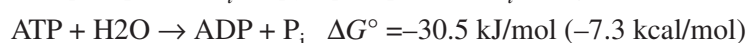
(α), beta (β), and gamma (γ) phosphates. Consequently, it is closely related to the adenosine nucleotide, a monomer of RNA. ATP is highly soluble in water and is quite stable in solutions between pH 6.8 and 7.4, but is rapidly hydrolysed at extreme pH. Consequently, ATP is best stored as an anhydrous salt.

ATP is an unstable molecule in unbuffered water, in which it hydrolyses to ADP and phosphate. This is because the strength of the bonds between the phosphate groups in ATP is less than the strength of the hydrogen bonds (hydration bonds), between its products (ADP and phosphate), and water. Thus, if ATP and ADP are in chemical equilibrium in water, almost all of the ATP will eventually be converted to ADP. A system that is far from equilibrium contains Gibbs free energy, and is capable of doing work. Living cells maintain the ratio of ATP to ADP at a point ten orders of magnitude from equilibrium, with ATP concentrations fivefold higher than the concentration of ADP. This displacement from equilibrium means that the hydrolysis of ATP in the cell releases a large amount of free energy.

Two phosphoanhydride bonds (those that connect adjacent phosphates) in an ATP molecule are responsible for the high energy content of this molecule. In the context of biochemical reactions, these anhydride bonds are frequently—and sometimes controversially—referred to as *high-energy bonds* (despite the fact it takes energy to break bonds). Energy stored in ATP may be released upon hydrolysis of the anhydride bonds. The primary phosphate group on the ATP molecule that is hydrolyzed when energy is needed to drive anabolic reactions is the γ -phosphate group. Located the farthest from the ribose sugar, it has a higher energy of hydrolysis than either the α - or β -phosphate. The bonds formed after hydrolysis—or the phosphorylation of a residue by ATP—are lower in energy than the phosphoanhydride bonds of ATP. During enzyme-catalyzed hydrolysis of ATP or phosphorylation by ATP, the available free energy can be harnessed by a living system to do work.

Any unstable system of potentially reactive molecules could potentially serve as a way of storing free energy, if the cell maintained their concentration far from the equilibrium point of the reaction. However, as is the case with most polymeric biomolecules, the breakdown of RNA, DNA, and ATP into simpler monomers is driven by both energy-release and entropy-increase considerations, in both standard concentrations, and also those concentrations encountered within the cell.

The standard amount of energy released from hydrolysis of ATP can be calculated from the changes in energy under non-natural (standard) conditions, then correcting to biological concentrations. The net change in heat energy (enthalpy) at standard temperature and pressure of the decomposition of ATP into hydrated ADP and hydrated inorganic phosphate is -30.5 kJ/mol, with a change in free energy of 3.4 kJ/mol. *The energy released by cleaving either a phosphate (P_i) or pyrophosphate (PP_i) unit from ATP at standard state of 1 M are:*



These values can be used to calculate the change in energy under physiological conditions and the cellular ATP/ADP ratio. However, a more representative value (which takes AMP into consideration) called the Energy charge is increasingly being employed. The values given for the Gibbs free energy for this reaction are dependent on a number of factors, including overall ionic strength and the presence of alkaline earth metal ions such as Mg^{2+} and Ca^{2+} . Under typical cellular conditions, ΔG is approximately -57 kJ/mol (-14 kcal/mol).

IONIZATION IN BIOLOGICAL SYSTEMS

ATP (adenosine triphosphate) has multiple groups with different acid dissociation constants. In neutral solution, ionized ATP exists mostly as ATP^{4-} , with a small proportion of ATP^{3-} . As ATP has several negatively charged groups in neutral solution, it can chelate metals with very high affinity. The binding constant for various metal ions are (given as per mole) as Mg^{2+} (9554), Na^+ (13), Ca^{2+} (3722), K^+ (8), Sr^{2+} (1381) and Li^+ (25). Due to the strength of these interactions, ATP exists in the cell mostly in a complex with Mg^{2+} .

BIOSYNTHESIS

The ATP concentration inside the cell is typically 1–10 mM. ATP can be produced by redox reactions using simple and complex sugars (carbohydrates) or lipids as an energy source. For complex fuels to be synthesized into ATP, they first need to be broken down into smaller, more simple molecules. Carbohydrates are hydrolysed into simple sugars, such as glucose and fructose. Fats (triglycerides) are metabolised to give fatty acids and glycerol. The overall process of oxidizing glucose to carbon dioxide is known as cellular respiration and can produce about 30 molecules of ATP from a single molecule of glucose. ATP can be produced by a number of distinct cellular processes; the three main pathways used to generate energy in eukaryotic organisms are glycolysis and the citric acid cycle/oxidative phosphorylation, both components of cellular respiration; and beta-oxidation. The majority of this ATP production by a non-photosynthetic aerobic eukaryote takes place in the mitochondria, which can make up nearly 25% of the total volume of a typical cell.

Glycolysis

In glycolysis, glucose and glycerol are metabolized to pyruvate via the glycolytic pathway. In most organisms, this process occurs in the cytosol, but, in some protozoa such as the kinetoplastids, this is carried out in a specialized organelle called the glycosome. Glycolysis generates a net two molecules of ATP through substrate phosphorylation catalyzed by two enzymes: PGK and pyruvate kinase. Two molecules of NADH are also produced, which can be oxidized via the electron transport chain and result in the generation of additional ATP by ATP synthase. The pyruvate generated as an end-product of glycolysis is a substrate for the Krebs Cycle.

Glucose

In the mitochondrion, pyruvate is oxidized by the pyruvate dehydrogenase complex to the acetyl group, which is fully oxidized to carbon dioxide by the citric acid cycle (also known as the Krebs cycle). Every “turn” of the citric acid cycle produces two molecules of carbon dioxide, one molecule of the ATP equivalent guanosine triphosphate (GTP) through substrate-level phosphorylation catalyzed by succinyl-CoA synthetase, three molecules of the reduced coenzyme NADH, and one molecule of the reduced coenzyme FADH₂. Both of these latter molecules are recycled to their oxidized states (NAD⁺ and FAD, respectively) via the electron transport chain, which generates additional ATP by oxidative phosphorylation. The oxidation of an NADH molecule results in the synthesis of 2–3 ATP molecules, and the oxidation of one FADH₂ yields between 1–2 ATP molecules. The majority of cellular ATP is generated by this process. Although the citric acid cycle itself does not involve molecular oxygen, it is an obligately aerobic process because O₂ is needed to recycle the reduced NADH and FADH₂ to their oxidized states. In the absence of oxygen the citric acid cycle will cease to function due to the lack of available NAD⁺ and FAD.

The generation of ATP by the mitochondrion from cytosolic NADH relies on the malate-aspartate shuttle (and to a lesser extent, the glycerol-phosphate shuttle) because the inner mitochondrial membrane is impermeable to NADH and NAD⁺. Instead of transferring the generated NADH, a malate dehydrogenase enzyme converts oxaloacetate to malate, which is translocated to the mitochondrial matrix. Another malate dehydrogenase-catalyzed reaction occurs in the opposite direction, producing oxaloacetate and NADH from the newly transported malate and the mitochondrion's interior store of NAD⁺. A transaminase converts the oxaloacetate to aspartate for transport back across the membrane and into the intermembrane space.

In oxidative phosphorylation, the passage of electrons from NADH and FADH₂ through the electron transport chain powers the pumping of protons out of the mitochondrial matrix and into the intermembrane space. This creates a proton motive force that is the net effect of a pH gradient and an electric potential gradient across the inner mitochondrial membrane. Flow of protons down this potential gradient—that is, from the intermembrane

space to the matrix—provides the driving force for ATP synthesis by ATP synthase. This enzyme contains a rotor subunit that physically rotates relative to the static portions of the protein during ATP synthesis.

Most of the ATP synthesized in the mitochondria will be used for cellular processes in the cytosol; thus it must be exported from its site of synthesis in the mitochondrial matrix. The inner membrane contains an antiporter, the ADP/ATP translocase, which is an integral membrane protein used to exchange newly synthesized ATP in the matrix for ADP in the intermembrane space. This translocase is driven by the membrane potential, as it results in the movement of about 4 negative charges out of the mitochondrial membrane in exchange for 3 negative charges moved inside. However, it is also necessary to transport phosphate into the mitochondrion; the phosphate carrier moves a proton in with each phosphate, partially dissipating the proton gradient.

Beta Oxidation

Fatty acids can also be broken down to acetyl-CoA by beta-oxidation. Each round of this cycle reduces the length of the acyl chain by two carbon atoms and produces one NADH and one FADH₂ molecule, which are used to generate ATP by oxidative phosphorylation. Because NADH and FADH₂ are energy-rich molecules, dozens of ATP molecules can be generated by the beta-oxidation of a single long acyl chain. The high energy yield of this process and the compact storage of fat explain why it is the most dense source of dietary calories.

Fermentation

Fermentation entails the generation of energy via the process of substrate-level phosphorylation in the absence of a respiratory electron transport chain. In most eukaryotes, glucose is used as both an energy store and an electron donor.

The equation for the oxidation of glucose to lactic acid is:



Anaerobic Respiration

Anaerobic respiration is the process of respiration using an electron acceptor other than O₂. In prokaryotes, multiple electron acceptors can be used in anaerobic respiration. These include nitrate, sulfate or carbon dioxide. These processes lead to the ecologically important processes of denitrification, sulfate reduction and acetogenesis, respectively.

ATP Replenishment by Nucleoside Diphosphate Kinases

ATP can also be synthesized through several so-called “replenishment” reactions catalyzed by the enzyme families of nucleoside diphosphate kinases (NDKs), which use other nucleoside triphosphates as a high-energy phosphate donor, and the ATP:guanido-phosphotransferase family.

ATP Production During Photosynthesis

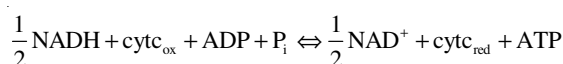
In plants, ATP is synthesized in thylakoid membrane of the chloroplast during the light-dependent reactions of photosynthesis in a process called photophosphorylation. Here, light energy is used to pump protons across the chloroplast membrane. This produces a proton-motive force and this drives the ATP synthase, exactly as in oxidative phosphorylation. Some of the ATP produced in the chloroplasts is consumed in the Calvin cycle, which produces triose sugars. The total quantity of ATP in the human body is about 0.2 moles. The majority of ATP is not usually synthesised *de novo*, but is generated from ADP by the aforementioned

processes. Thus, at any given time, the total amount of ATP + ADP remains fairly constant. The energy used by human cells requires the hydrolysis of 100 to 150 moles of ATP daily, which is around 50 to 75 kg. A human will typically use up his or her body weight of ATP over the course of the day. This means that each ATP molecule is recycled 500 to 750 times during a single day ($100/0.2 = 500$). ATP cannot be stored, hence, its consumption closely follows its synthesis. However a total of around 5 g of ATP is used by cell processes at any time in the body.

REGULATION OF BIOSYNTHESIS

ATP production in an aerobic eukaryotic cell is tightly regulated by allosteric mechanisms, by feedback effects, and by the substrate concentration dependence of individual enzymes within the glycolysis and oxidative phosphorylation pathways. Key control points occur in enzymatic reactions that are so energetically favorable that they are effectively irreversible under physiological conditions. In glycolysis, hexokinase is directly inhibited by its product, glucose-6-phosphate, and pyruvate kinase is inhibited by ATP itself. The main control point for the glycolytic pathway is phosphofructokinase (PFK), which is allosterically inhibited by high concentrations of ATP and activated by high concentrations of AMP. The inhibition of PFK by ATP is unusual, since, ATP is also a substrate in the reaction catalyzed by PFK; the biologically active form of the enzyme is a tetramer that exists in two possible conformations, only one of which binds the second substrate fructose-6-phosphate (F6P). The protein has two binding sites for ATP—the active site is accessible in either protein conformation, but ATP binding to the inhibitor site stabilizes the conformation that binds F6P poorly.

A number of other small molecules can compensate for the ATP-induced shift in equilibrium conformation and reactivate PFK, including cyclic AMP, ammonium ions, inorganic phosphate, and fructose-1,6-and-2,6-biphosphate. The citric acid cycle is regulated mainly by the availability of key substrates, particularly the ratio of NAD^+ to NADH and the concentrations of calcium, inorganic phosphate, ATP, ADP, and AMP. Citrate—the molecule that gives its name to the cycle—is a feedback inhibitor of citrate synthase and also inhibits PFK, providing a direct link between the regulation of the citric acid cycle and glycolysis. In oxidative phosphorylation, the key control point is the reaction catalyzed by cytochrome c oxidase, which is regulated by the availability of its substrate—the reduced form of cytochrome c. *The amount of reduced cytochrome c available is directly related to the amounts of other substrates:*



which directly implies this equation:

$$\frac{[\text{cytc}_{\text{red}}]}{[\text{cytc}_{\text{ox}}]} = \left(\frac{[\text{NADH}]}{[\text{NAD}^+]} \right)^{\frac{1}{2}} \left(\frac{[\text{ADP}][\text{P}_i]}{[\text{ATP}]} \right) K_{\text{eq}}$$

Thus, a high ratio of $[\text{NADH}]$ to $[\text{NAD}^+]$ or a high ratio of $[\text{ADP}][\text{P}_i]$ to $[\text{ATP}]$ imply a high amount of reduced cytochrome c and a high level of cytochrome c oxidase activity. An additional level of regulation is introduced by the transport rates of ATP and NADH between the mitochondrial matrix and the cytoplasm.

FUNCTIONS IN CELLS

Metabolism, Synthesis, and Active Transport

ATP is consumed in the cell by energy-requiring (endergonic) processes and can be generated by energy-releasing (exergonic) processes. In this way ATP transfers energy between spatially separate metabolic reactions. ATP is the main energy source for the majority of cellular functions. This includes the synthesis of

macromolecules, including DNA and RNA, and proteins. ATP also plays a critical role in the transport of macromolecules across cell membranes, *e.g.*, exocytosis and endocytosis.

Roles in Cell Structure and Locomotion

ATP is critically involved in maintaining cell structure by facilitating assembly and disassembly of elements of the cytoskeleton. In a related process, ATP is required for the shortening of actin and myosin filament crossbridges required for muscle contraction. This latter process is one of the main energy requirements of animals and is essential for locomotion and respiration.

Cell Signaling

Extracellular Signaling

Extracellular ATP (eATP) is also a signalling molecule. ATP, ADP, or adenosine are recognised by purinergic receptors, or purinoreceptors, which might be the most abundant receptors in mammalian tissues. In humans, this signalling role is important in both the central and peripheral nervous system. Activity-dependent release of ATP from synapses, axons and glia activates purinergic membrane receptors known as P2. The P2Y receptors are G protein-coupled receptors, which are *metabotropic*, and primarily modulate intracellular calcium and, to a lesser extent, cyclic AMP levels. Though named between P2Y₁ and P2Y₁₅, only nine members of the P2Y family have been cloned, and some are only related through weak homology and several (P2Y₅, P2Y₇, P2Y₉, P2Y₁₀) do not function as receptors that raise cytosolic calcium.

The P2X *ionotropic* receptor subgroup comprises seven members (P2X₁–P2X₇), which are ligand-gated Ca²⁺-permeable ion channels that open when bound to an extracellular purine nucleotide, like ATP. In contrast to P2 receptors (agonist rank order of potency: ATP > ADP > AMP > ADO), purinergic nucleoside triphosphates like ATP are not strong agonists of P1 receptors, which are strongly activated by adenosine and other nucleosides (ADO > AMP > ADP > ATP).

P1 receptors have A1, A2a, A2b, and A3 subtypes (the “A” is standard nomenclature for indicating an *adenosine receptor* subtype), all of which are G protein-coupled receptors, A1 and A3 being coupled to G_i, and A2a and A2b being coupled to G_s. All adenosine receptors were shown to activate at least one subfamily of mitogen-activated protein kinases. The actions of adenosine are often antagonistic or synergistic to the actions of ATP. In the CNS, adenosine has multiple functions, such as modulation of neural development, neuron and glial signalling and the control of innate and adaptive immune systems.

Intracellular Signaling

ATP is critical in signal transduction processes. It is used by kinases as the source of phosphate groups in their phosphate transfer reactions. Kinase activity on substrates such as proteins or membrane lipids are a common form of signal transduction. Phosphorylation of a protein by a kinase can activate a cascade such as the mitogen-activated protein kinase cascade.

ATP is also used by adenylate cyclase most commonly in G protein-coupled receptor signal transduction pathways and is transformed to the second messenger molecule cyclic AMP, which is involved in triggering calcium signals by the release of calcium from intracellular stores. This form of signal transduction is particularly important in brain function, although it is involved in the regulation of a multitude of other cellular processes.

DNA and RNA Synthesis

In all known organisms, the Deoxyribonucleotides that make up DNA are synthesized by the action of ribonucleotide reductase (RNR) enzymes on their corresponding ribonucleotides. These enzymes reduce the sugar residue from ribose to deoxyribose by removing oxygen from the 2' hydroxyl group; the substrates are ribonucleoside diphosphates and the products deoxyribonucleoside diphosphates (the latter are denoted dADP, dCDP, dGDP, and dUDP respectively.)

All ribonucleotide reductase enzymes use a common sulphhydryl radical mechanism reliant on reactive cysteine residues that oxidize to form disulfide bonds in the course of the reaction. RNR enzymes are recycled by reaction with thioredoxin or glutaredoxin. The regulation of RNR and related enzymes maintains a balance of dNTPs relative to each other and relative to NTPs in the cell. Very low dNTP concentration inhibits DNA synthesis and DNA repair and is lethal to the cell, while an abnormal ratio of dNTPs is mutagenic due to the increased likelihood of the DNA polymerase incorporating the wrong dNTP during DNA synthesis. Regulation of or differential specificity of RNR has been proposed as a mechanism for alterations in the relative sizes of intracellular dNTP pools under cellular stress such as hypoxia.

In the synthesis of the nucleic acid RNA, adenosine derived from ATP is one of the four nucleotides incorporated directly into RNA molecules by RNA polymerases. The energy driving this polymerization comes from cleaving off a pyrophosphate (two phosphate groups). The process is similar in DNA biosynthesis, except that ATP is reduced to the deoxyribonucleotide dATP, before incorporation into DNA.

AMINO ACID ACTIVATION IN PROTEIN SYNTHESIS

Aminoacyl-tRNA synthetase enzymes utilize ATP as an energy source to attach a tRNA molecule to its specific amino acid, forming an aminoacyl-tRNA complex, ready for translation at ribosomes. The energy is made available by ATP hydrolysis to adenosine monophosphate (AMP) as two phosphate groups are removed. Amino acid activation refers to the attachment of an amino acid to its Transfer RNA (tRNA). Aminoacyl transferase binds Adenosine triphosphate (ATP) to amino acid, PP_i is released. Aminoacyl transferase binds AMP-amino acid to tRNA. The AMP is used in this step.

Amino Acid Activation

During amino acid activation the amino acids (aa) are attached to their corresponding tRNA. The coupling reactions are catalysed by a group of enzymes called aminoacyl-tRNA synthetases (named after the reaction product aminoacyl-tRNA or aa-tRNA). The coupling reaction proceeds in two steps:

1. $\text{aa} + \text{ATP} \longrightarrow \text{aa-AMP} + \text{PP}_i$
2. $\text{aa-AMP} + \text{tRNA} \longrightarrow \text{aa-tRNA} + \text{AMP}$

The amino acid is coupled to the penultimate nucleotide at the 3'-end of the tRNA (the A in the sequence CCA) via an ester bond (roll over in illustration). The formation of the ester bond conserves a considerable part of the energy from the activation reaction. This stored energy provides the majority of the energy needed for peptide bond formation during translation.

Each of the 20 amino acids are recognized by its specific aminoacyl-tRNA synthetase. The synthetases are usually composed of one to four protein subunits. The enzymes vary considerably in structure although they all perform the same type of reaction by binding ATP, one specific amino acid and its corresponding tRNA.

The specificity of the amino acid activation is as critical for the translational accuracy as the correct matching of the codon with the anticodon. The reason is that the ribosome only sees the anticodon of the tRNA during translation. Thus, the ribosome will not be able to discriminate between tRNAs with the same anticodon but linked to different amino acids. The error frequency of the amino acid activation reaction is approximately 1 in 10000 despite the small structural differences between some of the amino acids.

BINDING TO PROTEINS

Some proteins that bind ATP do so in a characteristic protein fold known as the Rossmann fold, which is a general nucleotide-binding structural domain that can also bind the coenzyme NAD. The most common ATP-binding proteins, known as kinases, share a small number of common folds; the protein kinases, the largest kinase superfamily, all share common structural features specialized for ATP binding and phosphate transfer. ATP in complexes with proteins, in general, requires the presence of a divalent cation, almost always magnesium, which binds to the ATP phosphate groups. The presence of magnesium greatly decreases the dissociation constant of ATP from its protein binding partner without affecting the ability of the enzyme to catalyze its reaction once the ATP has bound. The presence of magnesium ions can serve as a mechanism for kinase regulation.

ATP ANALOGUES

Biochemistry laboratories often use *in vitro* studies to explore ATP-dependent molecular processes. Enzyme inhibitors of ATP-dependent enzymes such as kinases are needed to examine the binding sites and transition states involved in ATP-dependent reactions. ATP analogs are also used in X-ray crystallography to determine a protein structure in complex with ATP, often together with other substrates. Most useful ATP analogs cannot be hydrolyzed as ATP would be; instead they trap the enzyme in a structure closely related to the ATP-bound state. Adenosine 5'(γ -thiotriphosphate) is an extremely common ATP analog in which one of the gamma-phosphate oxygens is replaced by a sulfur atom; this molecule is hydrolyzed at a dramatically slower rate than ATP itself and functions as an inhibitor of ATP-dependent processes. In crystallographic studies, hydrolysis transition states are modeled by the bound vanadate ion. However, caution is warranted in interpreting the results of experiments using ATP analogs, since, some enzymes can hydrolyze them at appreciable rates at high concentration.

PC OR PHOSPHAGEN SYSTEM

A muscle cell has some amount of ATP floating around that it can use immediately, but not very much—only enough to last for about three seconds. To replenish the ATP levels quickly, muscle cells contain a high-energy phosphate compound called creatine phosphate. The phosphate group is removed from creatine phosphate by an enzyme called creatine kinase, and is transferred to ADP to form ATP. The cell turns ATP into ADP, and the phosphagen rapidly turns the ADP back into ATP. As the muscle continues to work, the creatine phosphate levels begin to decrease. Together, the ATP levels and creatine phosphate levels are called the phosphagen system. The phosphagen system can supply the energy needs of working muscle at a high rate, but only for 8 to 10 seconds.

ANAEROBIC METABOLISM

Glucose, a simple sugar, is the body's primary source of energy. Cells in the body can quickly and efficiently convert glucose to adenosine triphosphate, or ATP—the energy currency that enables all bodily functions to occur. Glucose breaks down to ATP by 2 pathways: aerobic metabolism, which requires oxygen, and anaerobic metabolism, which occurs without oxygen. Exercise can help increase the efficiency of anaerobic metabolism of glucose.

ANAEROBIC METABOLISM OF GLUCOSE

Anaerobic metabolism of glucose is a step-wise biochemical process called glycolysis or fermentation and can be performed by most cells in humans, animals and plants. Glycolysis occurs in the cytosol — the liquid

portion of the cell — and produces energy quickly but not that efficiently. The anaerobic breakdown of 1 glucose molecule results in the production of 6 ATP molecules, 2 pyruvic acid molecules and 4 hydrogen ions, according to “Guyton and Hall Textbook of Medical Physiology.” By comparison, aerobic metabolism of 1 glucose molecule results in 38 ATP molecules.

INTENSITY VERSUS DURATION

Muscle cells use anaerobic glycolysis to produce ATP very quickly, about 2.5 times faster than aerobic metabolism, enabling powerful bursts of activity. Historically this could fuel “fight or flight” survival activities, such as sprinting from predators or lifting heavy objects. Unfortunately, the intensity, or high power output, of these activities can only be maintained for short periods because of the quick use of the fuel source — glucose — and the production of acidic byproducts. On average, glycolysis can only produce ATP for 30 seconds to 1 minute, although with training this can be increased, according to a 2009 study published in “Sports Medicine.” Activities that are sustained for more than 45 minutes increasingly use aerobic metabolism to produce ATP.

BYPRODUCTS

The pyruvic acid and hydrogen ions that are produced by anaerobic glycolysis must be managed. Pyruvic acid can undergo 2 different processes: It can enter the mitochondria, where it is further metabolized aerobically, or it can be converted to lactic acid. Conveniently, lactic acid can bind to hydrogen, creating lactate, which can exit the cell and eliminate buildup inside the muscle. Unfortunately, this process does not occur fast enough to balance the production of hydrogen ions by glycolysis, and the muscle cell becomes acidic, creating the burning sensation felt during intense exercise.

TRAINING THE GLYCOLYTIC SYSTEM

Anaerobic training is an effort to increase the ability to produce ATP using the glycolytic pathway and to manage its byproducts. According to the American College of Sports Medicine, power training should consist of 30 to 45 seconds of intense exercise, such as weightlifting or sprinting, with 3 to 5 minutes of rest between sets. Individual physical capacity, training status and target goals should be incorporated into any training regime. Speak to a doctor or personal trainer if you have any concerns about a new exercise regime.

AEROBIC METABOLISM

Aerobic metabolism is a way for your cells to convert fat, carbohydrate and sometimes protein into energy, but only in the presence of oxygen. Aerobic metabolism is slow, so it is useful for sustained activities like jogging or dancing rather than short bursts of effort like sprinting or weightlifting.

FUNCTION

Aerobic metabolism converts fat and carbohydrate into units of cellular energy called ATP. Aerobic metabolism is very efficient, producing 34 molecules of ATP from every molecule of glucose, compared to the two ATP molecules anaerobic metabolism generates. Aerobic metabolism is also the only means by which your cells can use fat for fuel, which is part of the reason why aerobic activities like swimming or cycling are such an effective means to lose weight.

SIGNIFICANCE

Your body performs aerobic metabolism all day long to provide energy for activities of daily living. Because aerobic metabolism requires water and oxygen, it is the reason that humans breathe constantly and part of the

reason why you need to drink water. Carbohydrates are a necessary ingredient for aerobic metabolism—even to burn fat. Because the muscles and liver can only store a limited amount of carbohydrate between meals, aerobic metabolism is also the reason why you need to eat carbohydrates every day. Although your body can metabolize protein for energy in the absence of carbohydrates, it is a slow and inefficient process. Aerobic metabolism in the absence of carbohydrate can only keep up with activities with a very low energy cost, leaving you feeling sluggish.

EFFECTS

Aerobic metabolism produces carbon dioxide and water as byproducts. The extra carbon dioxide production from keeping up with the demands of exercise is what makes you breathe hard when you exercise. According to the American College of Sports Medicine, at their maximum exercise capacity, humans only use 75 percent of the oxygen they inhale, meaning that there is no shortage of air coming in. Instead, it is the extra carbon dioxide buildup in the blood and the need to expel it that makes you short of breath.

MISCONCEPTIONS

Activities aren't solely aerobic or anaerobic. In fact, anaerobic metabolism is the first step in aerobic respiration. Both types of metabolism carry on simultaneously in different proportions depending on your effort level. Even as you sit reading, your cells are still performing anaerobic metabolism. The reason that you aren't out of breath at rest is that your cells can easily clear anaerobic metabolism's byproducts as quickly as you produce them.

TIME FRAME

Aerobic metabolism can only sustain a certain level of activity before it can't process oxygen fast enough to keep up with the cells' demands. Anaerobic metabolism supports any additional activity beyond this point, producing lactic acid as a byproduct. As the Sports Fitness Advisor explains, eventually your body reaches a point where it is producing lactic acid faster than it can clear it, forcing you to slow down. The amount of time it takes for a person to hit that "wall" depends on the intensity of exercise and her individual fitness level.

AEROBIC AND ANAEROBIC SYSTEM DURING REST AND EXERCISE

While most people know that aerobic exercise is good for the heart and that resistance training helps build lean body mass, most people don't fully understand how these different types of exercise elicit very different responses within our bodies. A basic understanding of how our body uses energy during different forms of exercise is critical for designing an effective exercise program. We will focus on energy systems—*i.e.*, how the body utilizes fat, carbohydrate, and protein to produce energy—and how these energy systems are relied upon during different forms of exercise. This article will give you a better understanding of how your body converts the food you eat into usable energy and how targeting specific energy systems will help you achieve your personal health and fitness goals.

ENERGY SYSTEMS REVIEW

In general, there are three basic energy systems:

- The phosphagen system (also referred to as the immediate energy system),
- The glycolytic energy system (also referred to as the nonoxidative or anaerobic system), and
- Mitochondrial respiration (also referred to as the oxidative or aerobic system).

Regardless of what energy system is used, the end result is the production of adenosine triphosphate (or ATP). ATP is extracted from the food we eat (fat, carbohydrate, and protein) and is required for the biochemical reactions involved in any muscle contraction. The intensity and duration of the activity dictates which foodstuffs are broken down as well as which energy system predominates. It is important to keep in mind, however, that no energy system acts alone. The relative contribution from each system depends on the intensity and duration of the activity. The Phosphagen or Immediate Energy System The phosphagen system is active during all-out exercise that lasts about 5 to 10 seconds such as a 100-meter dash, diving, jumping, lifting a heavy weight, dashing up a flight of stairs, or any other activity that involves a maximal, short burst of power. This system relies on stored ATP and to a larger extent, creatine phosphate to provide immediate energy. For any maximal intensity exercise lasting longer than 10 seconds, assistance from other sources of energy is required.

Glycolytic Energy or Anaerobic System The glycolytic energy system (also called glycolysis) involves the partial break down of glucose to a molecule called pyruvate. During this process, a relatively small amount of energy is produced. When oxygen demands exceed the oxygen supply, pyruvate is converted to lactate. Under these circumstances, glycolysis is often referred to as “fast” or “anaerobic” glycolysis. Anaerobic glycolysis is a key contributor to the total energy requirements for moderate to high intensity exercise lasting about one to two minutes. Although this system can provide a rapid source of energy, it is only about half as fast as the phosphagen system. When there is enough oxygen to meet the oxygen demands of the activity, such as during prolonged light to moderate intensity exercise, glycolysis proceeds much slower and the pyruvate that is formed participates in the formation of additional energy via aerobic processes. In this case, glycolysis is sometimes referred to as “aerobic” or “slow” glycolysis. We often think of low to moderate intensity aerobic exercise as a good way to burn a significant amount of fat. While this is true, aerobic energy can be derived from carbohydrates and to a much smaller extent, protein. In fact, most people don’t realize that even during light to moderate exercise, carbohydrates can provide up to 40 to 60 percent of the total energy requirements. In contrast, protein is not a preferred source of energy during any form of exercise (assuming an adequate diet) and generally contributes less than 10 percent of the total energy requirements.

MONITORING YOUR ENERGY USAGE

One of the most effective methods of determining the predominant energy system during a specific form of exercise is by monitoring your heart rate. Heart rate monitoring can help you determine the intensity of your workout as well as estimate the heart rate at which you transition from aerobic to anaerobic exercise (*i.e.*, from carbohydrate and fat usage to predominantly carbohydrate). While the transition point differs from person to person, you can get a general idea of where you transition from aerobic to anaerobic exercise by watching for substantial increases in heart rate, muscle fatigue, or in breathing depth and frequency. If you are truly engaging in anaerobic exercise, you will not be able to sustain the intensity of the exercise for longer than about one to two minutes. If you notice your intensity dropping off, you were probably performing anaerobic exercise. In contrast, if you are able to sustain your exercise intensity longer than about two minutes, you are probably exercising aerobically. As your fitness improves, you will be able to perform higher intensity exercise for longer periods of time.

EXERCISE MODE AND ENERGY USAGE

Keep in mind that although resistance training doesn’t necessarily burn a significant number of calories, it can provide significant health and fitness benefits. Not only does resistance training increase lean body mass (*i.e.*, muscle), which burns more calories than fat even while at rest, engaging in a regular resistance training can have positive effects on elements such as cholesterol, glucose metabolism, and bone density, to name a few.

Circuit Training Circuit training is sometimes considered a type of resistance training, but it is actually a compromise between resistance training and cardiovascular training. Essentially, circuit training can improve muscle endurance as well as provide modest gains in aerobic capacity. Because it is generally a low to moderate intensity workout that is sustained for an hour or more, circuit training is primarily an aerobic activity.

Aerobic Exercise

Walking, Jogging, Traditional Hi-Lo Aerobics, and Step Aerobics “Aerobic” exercise is typically touted as a great way to burn a lot of fat. While this is not necessarily incorrect, it can be misleading. For example, at about 25 percent of aerobic capacity (*i.e.*, low intensity exercise), fat is the primary source of fuel, but you are not burning a significant number of calories. If your goal is to lose weight, the key consideration is the net deficit in calories, not where the calories come from. As exercise intensity increases, the number of calories burned also increases. Therefore, while it is true that fat contributes a greater percentage of the total energy during lower intensity exercise, at higher intensity exercise, the total quantity of fat utilized may be greater for exercise performed for an equivalent period of time.

SO HOW DOES ENERGY USAGE AFFECT YOUR WORKOUT?

If you don’t have a specific goal in mind, but simply want to improve your overall health, the American College of Sports Medicine recommends moderate intensity physical activity performed for at least 20 to 30 minutes, excluding time spent warming up and cooling down, 3 to 5 times a week. If, on the other hand, you are training for some type of competitive event, make sure that your training program emphasizes the type of activity involved in that event. For example, if you are training for a triathlon, engaging in a power lifting training program three days a week will not make the best use of your time. You need to actively engage in running, biking, and swimming. Finally, if your goal is to lose weight, caloric deficit is key. You should aim for a caloric deficit of about 500 calories a day through decreased energy intake, increased energy expenditure, or a combination of the two. Although there are numerous types of exercise that are effective for weight loss, a combination of regular aerobic exercise and resistance training is a good place to start.

Table. Nutrient usage at different exercise intensities.

Nutrient	At Rest	Light to moderate intensity exercise	High-intensity endurance exercise	High-intensity sprint-type exercise
Protein	2-5%	2-5%	5-8%	2%
Carbohydrate	35%	40%	70%	95%
Fat	60%	55%	15%	3%

SHORT DURATION HIGH INTENSITY EXERCISE

After setting your alarm for the fifth time this, and still not getting out of bed early to work out, you realize once again you’re not going to workout. Don’t give up just yet. With a few minutes of your time, it’s still possible to increase your heart rate and get your blood pumping.

LET’S GET IT STARTED

Before starting your workout, begin with at least a five minute warm up. Your warm up should include a low intensity cardiovascular activity or activities such as jogging in place, jumping jacks or a light jog. Create a warm up that is specific to your program by using exercises that target specific muscles. An example would be using high knees, butt kicks, or lunges. There are many benefits to warming up including increased blood flow, increased heart rate, and warming up your muscles.

NOTHING BUT THE BASICS

When you're short on time but still need to work out, grab a jump rope. Try various techniques when jumping rope such as high knees, side to side, and squat jumps. If it has been a while since, you have jumped rope, start slowly aiming for 30 seconds. Try and perform as many 30 second sets as you can but, be sure to rest in between. If you don't have access to a jump rope, that's okay jumping jacks will do just fine. For more intense exercises try adding mountain climbers or bear crawls, these full body movements not only get your heart pumping but also work your entire body. Slowly increase your time while decreasing your rest time. A few sets and you'll be sweating in no time.

GO OUTSIDE

There are many activities you can do outside, including sprints and running. If you normally run outside, try changing your program. Instead of running long distances all week, use one day to run short distances for time. For example, instead of running a few miles, run one mile as fast as you can run. Sprints can also be added to your program including shuttle runs or running for a specified distance such as 40 yards 10 times. To increase your intensity add pushups in between your sprints.

PLYOMETRICS

Plyometrics have been shown to improve power, strength and endurance. Squat jumps, burpee's, single leg hops, scissors, and box jumps are movements that you can do just about anywhere. For box jumps, use a stair if you don't have access to a box. Before beginning plyometrics be sure to try the movements first in order to practice the technique and form. You may find your heart rate increasing rapidly with these full body movements. Start slowly aiming for five to 10 repetitions, increasing the number once you're able to perform 10 repetitions with resting.

VARIETY IS THE KEY

Don't be afraid to change your program often. If you're lifting weights, try using these exercises in between sets to increase your heart rate. On a strictly cardio day, aim for at least 25 to 30 minutes, if you don't have enough time, exercise for long as long as you can. Any amount is better than nothing.

HIGH INTENSITY EXERCISE LASTING SEVERAL MINUTES

The oxygen dissociation curve shifted less to the right in venous blood draining from muscle in eight insulin-deficient diabetics working at a constant submaximal workload than in seven normal controls (28.7 mm. Hg vs. 30.8 mm Hg; P less than 0.05). This diminution of the in-vivo Bohr effect at the muscle tissue level during exercise in diabetics was due to a significantly smaller decrease of venous blood pH (down to 7.33 vs. 7.27 in normals; P less than 0.05), probably a consequence of an altered muscle metabolism in insulin deficiency. Although no glucose was taken up, even during exercise, and less lactate was produced by insulin-deficient muscle (P less than 0.05), the differences in venous blood pH appeared to be brought about mainly by a different CO_2 production of the exercising muscle in the two groups. The response of Krebs cycle activity to exercise in insulin-deficient muscle might have been inadequate, as suggested by the increased 3-hydroxybutyrate/acetoacetate ratio in the venous blood observed in the normal controls but not in the diabetics. Furthermore, proportionally less of the arterial ketone body concentration was utilized by the working muscle in the insulin-deficient diabetics. Changes in erythrocyte 2,3-diphosphoglycerate did not contribute to the differences in the in-vivo Bohr effect.

HIGH INTENSITY AND LONG DURATION EXERCISE

High Intensity Training (HIT) is a form of strength training popularized in the 1970s by Arthur Jones, the founder of Nautilus. The training focuses on performing quality weight training repetitions to the point of momentary

muscular failure. The training takes into account the number of repetitions, the amount of weight, and the amount of time the muscle is exposed to tension in order to maximize the amount of muscle fibre recruitment.

PRINCIPLES

The fundamental principles of High Intensity Training (HIT) are that exercise should be brief, infrequent, and intense. Exercises are performed with a high level of effort, or intensity, where it is thought that it will stimulate the body to produce an increase in muscular strength and size. Advocates of HIT believe that this method is superior for strength and size building to most other methods which, for example, may stress lower weights with larger volume (sets x reps). As strength increases, HIT techniques will have the weight/resistance increased progressively where it is thought that it will provide the muscles with adequate overload to stimulate further improvements. There is an inverse relationship between how intensely and how long one can exercise. As a result, high intensity workouts are generally kept brief. After a High Intensity workout, as with any workout, the body requires time to recover and produce the responses stimulated during the workout, so there is more emphasis on rest and recovery in the HIT philosophy than in most other weight training methods. In any workout, not just HIT, training schedules should allow adequate time between workouts for recovery (and adaptation).

While many typical HIT programs comprise a single-set per exercise, tri-weekly, full-body workout, many variations exist in specific recommendations of set and exercise number, workout routines, volume and frequency of training. The common thread is an emphasis on a high level of effort, relatively brief and infrequent (*i.e.*, not daily) training, and the cadence of a lift, which will be very slow compared to a non-HIT weight training routine. Most HIT advocates stress the use of controlled lifting speeds and strict form, with special attention paid to avoiding any bouncing, jerking, or yanking of the weight or machine movement arm during exercise. Technical HIT advice varies from lifting the weights smoothly and at a natural pace, to timing the lifts, peaking at hold and descent. In extreme cases, it may take up to 30 seconds to complete a single repetition. Also emphasized when near exhaustion in order to further exhaust the muscle or muscles exercised: doing static holds for periods of time, and negative reps (lowering the weight). This will stimulate further growth and strength because muscles are weakest in positive/contracting movements (sometimes referred to as first stage failure of a muscle). Although you may not be able to lift a weight for another rep you will almost certainly be able to hold it statically for a further period (second stage of failure) and finally lower a weight at a slow controlled speed (third stage of failure). Until all three (lifting, holding and lowering) parts of an exercise can no longer be completed in a controlled manner a muscle cannot be considered thoroughly exhausted/exercised.

CONTROVERSY

A large number of skeptics dispute the methods and results claimed by HIT advocates. Some of the criticism asserts that HIT violates much conventional “wisdom” in weight training. By always using a weight that one can lift 8-12 times, using 4 second negatives, and so on, it has flown in the face of the exercise establishment. There exists also a controversy related to the development of HIT and its originality. Near the close of the 19th century, a medical doctor by the name of Gustav Zander developed a complete set of machines and a workout method remarkably close to that promoted by inventor and HIT enthusiast Arthur Jones in the early 1970s. Jones stated:

- So, in attempts to improve my exercise results, I designed and built a total of about twenty very sophisticated exercise machines, then believing that these were the first exercise machines ever built by anybody. But many years later, I learned that a doctor named Gustav Zander had designed and built a number of exercise machines in Europe nearly a hundred years before I built my first one; I did not copy Zander’s work and learned nothing from him, was not even aware of his work until long after I had made the same discoveries that he had made. But if I had known about, and understood,

Zander's work, it would have saved me a lot of time and a rather large fortune in money, because the man was a genius; his only problem was that he lived about a century ahead of his time, at a time when very few people cared about exercise and even fewer knew anything about it.

Regardless of who originally developed the systems (and machines) it is clear that through Arthur Jones and his company and a crew of HIT advocates, the principles and concepts of HIT became popularized.

HIT AND OTHER TRAINING ROUTINES

HIT will target a single body part with one or two exercises, and generally a single set of 6-10 reps for upper body exercises and either 8-15 or more commonly 12-20 reps for lower body exercises, done to momentary muscular failure. Deadlifts usually have a rep range of 1-5 reps. Cadence for a HIT workout is supposed to be smooth, but not always Super Slow. A standard HIT cadence is usually 3-1-4-1. For clarity, here are two examples of how the cadence would be for an exercise.

- *On the Lat Pulldown exercise the cadence is as follows:* 3 seconds pulling down (Positive movement), followed by a 1-second pause and squeeze (at full contraction), followed by a 4-second return (Negative movement), followed by a 1-second rest. This completes 1 rep.
- *On the Barbell Squat the cadence is as follows:* 4 seconds lowering the bar (the Negative movement), followed by a 1-second pause (at the bottom), followed by 3 seconds raising up the bar (the Positive movement), followed by a 1-second rest at the top. This completes 1 rep.

HIT stresses intensity over repetition. Many weightlifters will use a HIT routine to help break a 'plateau'- meaning they will use HIT temporarily when another routine stops giving desired results. Some HIT trainees will use HIT exclusively as well-Arthur Jones believed HIT was all that was required. Different strength training authors from Ellington Darden and Mike Mentzer to Dorian Yates and Gordon LaVelle have called their system HIT, with each individual having credited Arthur Jones for the formulation of its basic tenet principles. However, there has never been a clear and consistent guideline on how to utilize HIT. Darden advocated full body routines, while Yates recommended to split the workouts into four different sessions a week. Mentzer believed that no more than one set to muscular failure per body part was all that was required, yet Yates and LaVelle believed that more than one exercise per body part is necessary to get complete development as a bodybuilder.

REST-PAUSE

A former Mr. Universe, the late Mike Mentzer achieved his lifetime best condition from performing rest-pause, an old system of lifting involving single-rep maxima interspersed with brief (10 second) rest periods. Rest-pause has the advantages of old-school power training while also allowing for enough overall reps to be performed for hypertrophy and cardiovascular exercise purposes.

HIGH-INTENSITY INTERVAL TRAINING

High-intensity interval training (HIIT), also called high-intensity intermittent exercise (HIIE) or sprint interval training (SIT), is a form of interval training, a cardiovascular exercise strategy alternating short periods of intense anaerobic exercise with less intense recovery periods. HIIT is the concept where one performs a short burst of high-intensity (or max-intensity) exercise followed by a brief low-intensity activity, repeatedly, until too exhausted to continue. Though there is no universal HIIT session duration, these intense workouts typically last under 30 minutes, with times varying based on a participant's current fitness level. HIIT workouts provide improved athletic capacity and condition as well as improved glucose metabolism. Compared with other regimens, HIIT may not be as effective for treating hyperlipidemia and obesity, or improving muscle and bone mass. However, research has shown that HIIT regimens successfully produced significant reductions in the fat mass of the whole-body. Some researchers also note that HIIT requires "an extremely high level of subject motivation" and question whether the general population could safely or practically tolerate the extreme nature of the exercise regimen.

Procedure

High-intensity interval training can be described as an exercise session composed entirely of HIIT techniques, or as a component of an exercise plan. HIIT exercise sessions generally consist of a warm up period, then several repetitions of high-intensity exercise separated by medium intensity exercise for recovery, then a cool down period. The high-intensity exercise should be done at near maximum intensity. The medium exercise should be about 50% intensity. The number of repetitions and length of each depends on the exercise, but may be as little as three repetitions with just 20 seconds of intense exercise. The specific exercises performed during the high-intensity portions vary. There is no specific formula to HIIT. Depending on one's level of cardiovascular development, the moderate-level intensity can be as slow as walking. A common formula involves a 2:1 ratio of work to recovery periods, for example, 30–40 seconds of hard sprinting alternated with 15–20 seconds of jogging or walking. The entire HIIT session may last between four and thirty minutes, meaning that it is considered to be an excellent way to maximize a workout that is limited on time. Use of a clock or timer is recommended to keep accurate times, the number of rounds, and intensity.

Branch

Peter Coe Regimen

A type of high-intensity interval training with short recovery periods was used in the 1970s by the athletics coach Peter Coe when setting sessions for his son Sebastian Coe. Inspired by the principles propounded by the German coach and university professor Woldemar Gerschler and the Swedish physiologist Per-Olof Åstrand, Coe set sessions involving repeated fast 200 metre runs with only 30 seconds recovery between each fast run.

Tabata Regimen

A version of HIIT was based on a 1996 study by Professor Izumi Tabata (田畑泉) *et al.* initially involving Olympic speedskaters. The study used 20 seconds of ultra-intense exercise (at an intensity of about 170% of $\text{VO}_{2\text{max}}$) followed by 10 seconds of rest, repeated continuously for 4 minutes (8 cycles). The exercise was performed on a mechanically braked cycle ergometer. Tabata called this the IE1 protocol. In the original study, athletes using this method trained 4 times per week, plus another day of steady-state training, and obtained gains similar to a group of athletes who did steady state training (70% $\text{VO}_{2\text{max}}$) 5 times per week. The steady state group had a higher $\text{VO}_{2\text{max}}$ at the end (from 52 to 57 $\text{mL}/(\text{kg}\cdot\text{min})$), but the Tabata group had started lower and gained more overall (from 48 to 55 $\text{mL}/(\text{kg}\cdot\text{min})$). Also, only the Tabata group had gained anaerobic capacity benefits. It is important to note that in the original study from 1996, participants were disqualified if they could not keep a steady cycling pace of 85RPM for the full 20 seconds of work. In popular culture, “Tabata training” has now come to refer to a wide variety of HIIT protocols and exercise regimens that may or may not have similar benefits to those found in Tabata's original study.

Gibala Regimen

Professor Martin Gibala and his team at McMaster University in Canada have been researching high-intensity exercise for several years. Their 2010 study on students uses 3 minutes for warming up, then 60 seconds of intense exercise (at 95% of $\text{VO}_{2\text{max}}$) followed by 75 seconds of rest, repeated for 8–12 cycles (sometimes referred to as “The Little Method”). Subjects using this method training 3 times per week obtained gains similar to what would be expected from subjects who did steady state (50–70% $\text{VO}_{2\text{max}}$) training five times per week. While still a demanding form of training, this exercise protocol could be used by the general public with nothing more than an average exercise bike.

Gibala's group published a less intense version of their regimen in a 2011 paper in *Medicine and Science in Sports and Exercise*. This was intended as a gentler option for sedentary people who had done no exercise for over a year. It included 3 minutes of warm-up, 10 repetitions of 60-second bursts at 60% peak power (80–95% of heart rate reserve) each followed by 60 seconds of recovery, and then a 5-minute cool-down.

Zuniga Regimen

Jorge Zuniga, assistant professor of exercise science at Creighton University, set out to determine how to fit the highest volume of work and oxygen consumption into the smallest amount of time. He found that intervals of 30 seconds at 90% of maximum power output followed by 30 seconds of rest allowed for the highest VO_2 consumption and the longest workout duration at specified intensity. Alternative protocols considered included 100% of maximum power output on the same interval schedule, similar to the Coe regimen, and 90% of maximum power output for three minutes, similar to traditional interval training. Zuniga's protocol has been implemented to great success by his students participating in Creighton's Army ROTC program. Cadets completing the protocol twice a week saw greater improvements in APFT scores than in years past.

Vol्लाard Regimen

Dr Niels Vol्लाard at the University of Stirling proposed that when high-intensity intervals are done at 'all-out' intensities, associated health benefits plateau after performing 2 or 3 sprint repetitions. This led to the development of a 10-minute exercise routine consisting of easy pedalling interspersed with two 20-second 'all-out' cycling sprints. In a 2017 meta-analysis, Vol्लाard indeed showed that common protocols with as many as 6 to 10 repetitions of 30-second 'all-out' sprints do not improve aerobic fitness more than the '2x20-s' protocol. It is claimed that this short protocol may remove many of the drawbacks that make other high-intensity interval training protocols unsuitable for the general population. In a BBC *Horizon* programme in February 2012, Jamie Timmons, professor of systems biology at the University of Loughborough, put Michael Mosley through this exercise bike regimen, but with three sprints instead of two. This was done three times a week for a total of 30 minutes of exercise per week (3 minutes of intense exercise), plus warm-up and recovery time.

Regimen Comparison

Wood *et al.* compared High-intensity interval training of eight 1-minute bouts at 85% W_{max} interspersed with a 1-minute active recovery at 25% W_{max} v Sprint interval training of eight 30-second bouts at 130% W_{max} interspersed with 90-second active recovery at 25% W_{max} . (Total time matched at 24 mins including warm up and cool down). Their conclusion was "HIIT is the recommended routine" but "the magnitude of differences in various parameters between regimens was small; therefore, preference for either modality may be up to the individual".

Health Effects

Cardiovascular Fitness

A 2015 systematic review and meta-analysis of randomized controlled trials found that HIIT training and traditional endurance training both lead to significantly improved cardiovascular fitness in healthy adults ages 18–45 but greater improvements in VO_2 max were seen in those participating in the HIIT exercise regimen. Another analysis also found that HIIT regimens of one month or longer effectively improve cardiovascular fitness in adolescents and lead to moderate improvements in body composition. Furthermore, a separate systematic

review and meta-analysis of seven small randomized controlled trials found that HIIT (defined as four intervals of four minutes at 85–95% of max heart rate with three-minute intervals at 60–70% of max heart rate) was more effective than moderate-intensity continuous training at improving blood vessel function and markers of blood vessel health.

- Cardiovascular disease: A 2015 meta-analysis comparing HIIT to moderate intensity continuous training (MICT) in people with coronary artery disease found that HIIT leads to greater improvements in VO_2 max but that MICT leads to greater reductions in body weight and heart rate. A 2014 meta-analysis found that the cardiorespiratory fitness, as measured by VO_2 max, of individuals with lifestyle-induced chronic cardiovascular or metabolic diseases (including high blood pressure, obesity, heart failure, coronary artery disease, or metabolic syndrome) who completed a HIIT exercise program was nearly double that of individuals who completed a MICT exercise program.

Metabolic Effects

HIIT significantly lowers insulin resistance compared to continuous training or control conditions and leads to modestly decreased fasting blood glucose levels and increased weight loss compared to those who do not undergo a physical activity intervention.

LONG DURATION EXERCISE

Concordant with the functional roles of the brain structures that exhibit increased gray matter volumes, regular exercise over a period of several months has been shown to persistently improve numerous executive functions and several forms of memory. In particular, consistent aerobic exercise has been shown to improve attentional control, information processing speed, cognitive flexibility (*e.g.*, task switching), inhibitory control, working memory updating and capacity, declarative memory, and spatial memory. In healthy young and middle-aged adults, the effect sizes of improvements in cognitive function are largest for indices of executive functions and small to moderate for aspects of memory and information processing speed. It may be that in older adults, individuals benefit cognitively by taking part in both aerobic and resistance type exercise of at least moderate intensity. Individuals who have a sedentary lifestyle tend to have impaired executive functions relative to other more physically active non-exercisers. A reciprocal relationship between exercise and executive functions has also been noted: improvements in executive control processes, such as attentional control and inhibitory control, increase an individual's tendency to exercise.

Simple Carbohydrates

INTRODUCTION

A carbohydrate is a biomolecule consisting of carbon (C), hydrogen (H) and oxygen (O) atoms, usually with a hydrogen–oxygen atom ratio of 2:1 (as in water); in other words, with the empirical formula $C_m(H_2O)_n$ (where m may be different from n). This formula holds true for monosaccharides. Some exceptions exist; for example, deoxyribose, a sugar component of DNA, has the empirical formula $C_5H_{10}O_4$. The carbohydrates are technically hydrates of carbon; structurally it is more accurate to view them as aldoses and ketoses.

The term is most common in biochemistry, where it is a synonym of ‘saccharide’, a group that includes sugars, starch, and cellulose. The saccharides are divided into four chemical groups: monosaccharides, disaccharides, oligosaccharides, and polysaccharides. Monosaccharides and disaccharides, the smallest (lower molecular weight) carbohydrates, are commonly referred to as sugars. The word *saccharide* comes from the Greek *σάκχαρον* (*sákkharon*), meaning “sugar”. While the scientific nomenclature of carbohydrates is complex, the names of the monosaccharides and disaccharides very often end in the suffix-ose. For example, grape sugar is the monosaccharide glucose, cane sugar is the disaccharide sucrose, and milk sugar is the disaccharide lactose. Carbohydrates perform numerous roles in living organisms. Polysaccharides serve for the storage of energy (*e.g.*, starch and glycogen) and as structural components (*e.g.*, cellulose in plants and chitin in arthropods). The 5-carbon monosaccharide ribose is an important component of coenzymes (*e.g.*, ATP, FAD and NAD) and the backbone of the genetic molecule known as RNA. The related deoxyribose is a component of DNA. Saccharides and their derivatives include many other important biomolecules that play key roles in the immune system, fertilization, preventing pathogenesis, blood clotting, and development.

Carbohydrates are found in a wide variety of foods. The important sources are cereals (wheat, maize, rice), potatoes, sugarcane, fruits, table sugar (sucrose), bread, milk, etc., Starch and sugar are the important carbohydrates in our diet. Starch is abundant in potatoes, maize, rice and other cereals. Sugar appears in our diet mainly as sucrose (table sugar), which is added to drinks and many prepared foods such as jam, biscuits and cakes, and glucose and fructose which occur naturally in many fruits and some vegetables. Glycogen is a carbohydrate found in the liver and muscles (as energy source). Cellulose in the cell wall of all plant tissue is a carbohydrate. It is important in our diet as fibre which helps to maintain a healthy digestive system.

TERMINOLOGY

In scientific literature, the term “carbohydrate” has many synonyms, like “sugar” (in the broad sense), “saccharide”, “ose”, “glucide”, “hydrate of carbon” or “polyhydroxy compounds with aldehyde or ketone”. Some of these terms, specially “carbohydrate” and “sugar”, are also used with other meanings. In food science and in many informal contexts, the term “carbohydrate” often means any food that is particularly rich in the complex carbohydrate starch

(such as cereals, bread and pasta) or simple carbohydrates, such as sugar (found in candy, jams, and desserts). Often in lists of nutritional information, such as the USDA National Nutrient Database, the term “carbohydrate” (or “carbohydrate by difference”) is used for everything other than water, protein, fat, ash, and ethanol. This will include chemical compounds such as acetic or lactic acid, which are not normally considered carbohydrates. It also includes dietary fibre which is a carbohydrate but which does not contribute much in the way of food energy (kilocalories), even though it is often included in the calculation of total food energy just as though it were a sugar. In the strict sense, “sugar” is applied for sweet, soluble carbohydrates, many of which are used in food.

STRUCTURE

Formerly the name “carbohydrate” was used in chemistry for any compound with the formula $C_m(H_2O)_n$. Following this definition, some chemists considered formaldehyde (CH_2O) to be the simplest carbohydrate, while others claimed that title for glycolaldehyde. Today, the term is generally understood in the biochemistry sense, which excludes compounds with only one or two carbons and includes many biological carbohydrates which deviate from this formula. For example, while the above representative formulas would seem to capture the commonly known carbohydrates, ubiquitous and abundant carbohydrates often deviate from this. For example, carbohydrates often display chemical groups such as: *N*-acetyl (*e.g.*, chitin), sulphate (*e.g.*, glycosaminoglycans), carboxylic acid (*e.g.*, sialic acid) and deoxy modifications (*e.g.*, fucose and sialic acid). Natural saccharides are generally built of simple carbohydrates called monosaccharides with general formula $(CH_2O)_n$ where *n* is three or more. A typical monosaccharide has the structure $H-(CHOH)_x(C=O)-(CHOH)_y-H$, that is, an aldehyde or ketone with many hydroxyl groups added, usually one on each carbon atom that is not part of the aldehyde or ketone functional group.

Examples of monosaccharides are glucose, fructose, and glyceraldehydes. However, some biological substances commonly called “monosaccharides” do not conform to this formula (*e.g.*, uronic acids and deoxy-sugars such as fucose) and there are many chemicals that do conform to this formula but are not considered to be monosaccharides (*e.g.*, formaldehyde CH_2O and inositol $(CH_2O)_6$). The open-chain form of a monosaccharide often coexists with a closed ring form where the aldehyde/ketone carbonyl group carbon ($C=O$) and hydroxyl group ($-OH$) react forming a hemiacetal with a new $C-O-C$ bridge. Monosaccharides can be linked together into what are called polysaccharides (or oligosaccharides) in a large variety of ways. Many carbohydrates contain one or more modified monosaccharide units that have had one or more groups replaced or removed. For example, deoxyribose, a component of DNA, is a modified version of ribose; chitin is composed of repeating units of *N*-acetyl glucosamine, a nitrogen-containing form of glucose.

DIVISION

Carbohydrates are polyhydroxy aldehydes, ketones, alcohols, acids, their simple derivatives and their polymers having linkages of the acetal type. They may be classified according to their degree of polymerization, and may be divided initially into three principal groups, namely sugars, oligosaccharides and polysaccharides

The major dietary carbohydrates		
Class (DP*)	Subgroup	Components
Sugars (1–2)	Monosaccharides	Glucose, galactose, fructose, xylose
	Disaccharides	Sucrose, lactose, maltose, trehalose
	Polyols	Sorbitol, mannitol
Oligosaccharides (3–9)	Malto-oligosaccharides	Maltodextrins
	Other oligosaccharides	Raffinose, stachyose, fructo-oligosaccharides
Polysaccharides (>9)	Starch	Amylose, amylopectin, modified starches
	Non-starch polysaccharides	Cellulose, hemicellulose, pectins, hydrocolloids

DP * = Degree of polymerization

MONOSACCHARIDES

Monosaccharides also called simple sugars, are the most basic units of carbohydrates. They are fundamental units of carbohydrates and cannot be further hydrolyzed to simpler compounds. The general formula is $C_nH_{2n}O_n$. They are the simplest form of sugar and are usually colorless, water-soluble, and crystalline solids. Some monosaccharides have a sweet taste. Examples of monosaccharides include glucose (dextrose), fructose (levulose), and galactose. Monosaccharides are the building blocks of disaccharides (such as sucrose and lactose) and polysaccharides (such as cellulose and starch). Further, each carbon atom that supports a hydroxyl group (so, all of the carbons except for the primary and terminal carbon) is chiral, giving rise to a number of isomeric forms, all with the same chemical formula. For instance, galactose and glucose are both aldohexoses, but have different physical structures and chemical properties.

STRUCTURE AND NOMENCLATURE

With few exceptions (*e.g.*, deoxyribose), monosaccharides have this chemical formula: $C_x(H_2O)_y$, where conventionally $x \geq 3$. Monosaccharides can be classified by the number x of carbon atoms they contain: Triose (3) tetrose (4), pentose (5), hexose (6), heptose (7), and so on.

The most important monosaccharide, glucose, is a hexose. Examples of heptoses include the ketoses mannoheptulose and sedoheptulose. Monosaccharides with eight or more carbons are rarely observed as they are quite unstable. In aqueous solutions monosaccharides exist as rings if they have more than four carbons.

Linear-chain Monosaccharides

Simple monosaccharides have a linear and unbranched carbon skeleton with one carbonyl ($C=O$) functional group, and one hydroxyl (OH) group on each of the remaining carbon atoms. Therefore, the molecular structure of a simple monosaccharide can be written as $H(CHOH)_n(C=O)(CHOH)_mH$, where $n + 1 + m = x$; so that its elemental formula is $C_xH_{2x}O_x$.

By convention, the carbon atoms are numbered from 1 to x along the backbone, starting from the end that is closest to the $C=O$ group. Monosaccharides are the simplest units of carbohydrates and the simplest form of sugar. If the carbonyl is at position 1 (that is, n or m is zero), the molecule begins with a formyl group $H(C=O)-$ and is technically an aldehyde. In that case, the compound is termed an aldose. Otherwise, the molecule has a keto group, a carbonyl ($C=O$) between two carbons; then it is formally a ketone, and is termed a ketose. Ketoses of biological interest usually have the carbonyl at position 2.

The various classifications above can be combined, resulting in names such as “aldohexose” and “ketotriose”.

A more general nomenclature for open-chain monosaccharides combines a Greek prefix to indicate the number of carbons (tri-, tetra-, pent-, hex-, etc.) with the suffixes “-ose” for aldoses and “-ulose” for ketoses. In the latter case, if the carbonyl is not at position 2, its position is then indicated by a numeric infix. So, for example, $H(C=O)(CHOH)_4H$ is pentose, $H(CHOH)(C=O)(CHOH)_3H$ is pentulose, and $H(CHOH)_2(C=O)(CHOH)_2H$ is pent-3-ulose.

Open-chain Stereoisomers

Two monosaccharides with equivalent molecular graphs (same chain length and same carbonyl position) may still be distinct stereoisomers, whose molecules differ in the three-dimensional arrangement of the bonds of certain atoms. This happens only if the molecule contains a stereogenic center, specifically a carbon atom that is chiral (connected to four distinct molecular sub-structures). Those four bonds can have any of two configurations in space distinguished by their handedness. In a simple open-chain monosaccharide, every carbon

is chiral except the first and the last atoms of the chain, and (in ketoses) the carbon with the keto group. For example, the triketose $\text{H}(\text{CHOH})(\text{C}=\text{O})(\text{CHOH})\text{H}$ (glycerone, dihydroxyacetone) has no stereogenic center, and therefore exists as a single stereoisomer. The other triose, the aldose $\text{H}(\text{C}=\text{O})(\text{CHOH})_2\text{H}$ (glyceraldehyde), has one chiral carbon—the central one, number 2—which is bonded to groups H , OH , $\text{C}(\text{OH})\text{H}_2$, and $\text{C}(\text{O})\text{H}$. Therefore, it exists as two stereoisomers whose molecules are mirror images of each other (like a left and a right glove). Monosaccharides with four or more carbons may contain multiple chiral carbons, so they typically have more than two stereoisomers. The number of distinct stereoisomers with the same diagram is bounded by 2^c , where c is the total number of chiral carbons.

The Fischer projection is a systematic way of drawing the skeletal formula of an acyclic monosaccharide so that the handedness of each chiral carbon is well specified. Each stereoisomer of a simple open-chain monosaccharide can be identified by the positions (right or left) in the Fischer diagram of the chiral hydroxyls (the hydroxyls attached to the chiral carbons). Most stereoisomers are themselves chiral (distinct from their mirror images). In the Fischer projection, two mirror-image isomers differ by having the positions of all chiral hydroxyls reversed right-to-left. Mirror-image isomers are chemically identical in non-chiral environments, but usually have very different biochemical properties and occurrences in nature.

While most stereoisomers can be arranged in pairs of mirror-image forms, there are some non-chiral stereoisomers that are identical to their mirror images, in spite of having chiral centers. This happens whenever the molecular graph is symmetrical, as in the 3-ketopentoses $\text{H}(\text{CHOH})_2(\text{CO})(\text{CHOH})_2\text{H}$, and the two halves are mirror images of each other. In that case, mirroring is equivalent to a half-turn rotation. For this reason, there are only three distinct 3-ketopentose stereoisomers, even though the molecule has two chiral carbons.

Distinct stereoisomers that are not mirror-images of each other usually have different chemical properties, even in non-chiral environments. Therefore, each mirror pair and each non-chiral stereoisomer may be given a specific monosaccharide name. For example, there are 16 distinct aldohexose stereoisomers, but the name “glucose” means a specific pair of mirror-image aldohexoses. In the Fischer projection, one of the two glucose isomers has the hydroxyl at left on C3, and at right on C4 and C5; while the other isomer has the reversed pattern. These specific monosaccharide names have conventional three-letter abbreviations, like “Glu” for glucose and “Thr” for threose.

Generally, a monosaccharide with n asymmetrical carbons has 2^n stereoisomers. The number of open chain stereoisomers for an aldose monosaccharide is larger by one than that of a ketose monosaccharide of the same length. Every ketose will have $2^{(n-3)}$ stereoisomers where $n > 2$ is the number of carbons. Every aldose will have $2^{(n-2)}$ stereoisomers where $n > 2$ is the number of carbons. These are also referred to as epimers which have the different arrangement of OH and H groups at the asymmetric or chiral carbon atoms (this does not apply to those carbons having the carbonyl functional group).

Configuration of Monosaccharides

Like many chiral molecules, the two stereoisomers of glyceraldehyde will gradually rotate the polarization direction of linearly polarized light as it passes through it, even in solution. The two stereoisomers are identified with the prefixes D- and L-, according to the sense of rotation: D-glyceraldehyde is dextrorotatory (rotates the polarization axis clockwise), while L-glyceraldehyde is levorotatory (rotates it counterclockwise).

The D- and L- prefixes are also used with other monosaccharides, to distinguish two particular stereoisomers that are mirror-images of each other. For this purpose, one considers the chiral carbon that is furthest removed from the $\text{C}=\text{O}$ group. Its four bonds must connect to H , OH , $\text{C}(\text{OH})\text{H}$, and the rest of the molecule. If the molecule can be rotated in space so that the directions of those four groups match those of the analog groups in

D-glyceraldehyde's C2, then the isomer receives the D-prefix. Otherwise, it receives the L-prefix. In the Fischer projection, the D-and L-prefixes specifies the configuration at the carbon atom that is second from bottom: D-if the hydroxyl is on the right side, and L-if it is on the left side. Note that the D-and L-prefixes do not indicate the direction of rotation of polarized light, which is a combined effect of the arrangement at all chiral centers. However, the two enantiomers will always rotate the light in opposite directions, by the same amount.

Cyclisation of Monosaccharides

A monosaccharide often switches from the acyclic (open-chain) form to a cyclic form, through a nucleophilic addition reaction between the carbonyl group and one of the hydroxyls of the same molecule. The reaction creates a ring of carbon atoms closed by one bridging oxygen atom. The resulting molecule has an hemiacetal or hemiketal group, depending on whether the linear form was an aldose or a ketose. The reaction is easily reversed, yielding the original open-chain form. In these cyclic forms, the ring usually has 5 or 6 atoms. These forms are called furanoses and pyranoses, respectively—by analogy with furan and pyran, the simplest compounds with the same carbon-oxygen ring (although they lack the double bonds of these two molecules). For example, the aldohexose glucose may form a hemiacetal linkage between the hydroxyl on carbon 1 and the oxygen on carbon 4, yielding a molecule with a 5-membered ring, called glucofuranose.

The same reaction can take place between carbons 1 and 5 to form a molecule with a 6-membered ring, called glucopyranose. Cyclic forms with a 7-atom ring (the same of oxepane), rarely encountered, are called heptoses. For many monosaccharides (including glucose), the cyclic forms predominate, in the solid state and in solutions, and therefore the same name commonly is used for the open-and closed-chain isomers. Thus, for example, the term “glucose” may signify glucofuranose, glucopyranose, the open-chain form, or a mixture of the three. Cyclization creates a new stereogenic center at the carbonyl-bearing carbon. The -OH group that replaces the carbonyl's oxygen may end up in two distinct positions relative to the ring's midplane. Thus each open-chain monosaccharide yields two cyclic isomers (anomers), denoted by the prefixes α -and β -. The molecule can change between these two forms by a process called mutarotation, that consists in a reversal of the ring-forming reaction followed by another ring formation.

Haworth Projection

The stereochemical structure of a cyclic monosaccharide can be represented in a Haworth projection. In this diagram, the α -isomer for the pyranose form of a D-aldohexose has the -OH of the anomeric carbon below the plane of the carbon atoms, while the β -isomer has the -OH of the anomeric carbon above the plane. Pyranoses typically adopt a chair conformation, similar to that of cyclohexane. In this conformation, the α -isomer has the -OH of the anomeric carbon in an axial position, whereas the β -isomer has the -OH of the anomeric carbon in equatorial position (considering D-aldohexose sugars).

DERIVATIVES

A large number of biologically important modified monosaccharides exist:

- *Amino sugars such as:*
 - Galactosamine
 - Glucosamine
 - Sialic acid
 - N-Acetylglucosamine
- *Sulfosugars such as:*
 - Sulfoquinovose

DISACCHARIDE

A disaccharide (also called a double sugar or biose) is the sugar formed when two monosaccharides (simple sugars) are joined by glycosidic linkage. Like monosaccharides, disaccharides are soluble in water. Three common examples are sucrose, lactose, and maltose. Disaccharides are one of the four chemical groupings of carbohydrates (monosaccharides, disaccharides, oligosaccharides, and polysaccharides). The most common types of disaccharides—sucrose, lactose, and maltose—have twelve carbon atoms, with the general formula $C_{12}H_{22}O_{11}$. The differences in these disaccharides are due to atomic arrangements within the molecule.

The joining of simple sugars into a double sugar happens by a condensation reaction, which involves the elimination of a water molecule from the functional groups only. Breaking apart a double sugar into its two simple sugars is accomplished by hydrolysis with the help of a type of enzyme called a disaccharidase. As building the larger sugar ejects a water molecule, breaking it down consumes a water molecule. These reactions are vital in metabolism. Each disaccharide is broken down with the help of a corresponding disaccharidase (sucrase, lactase, and maltase).

CLASSIFICATION

There are two functionally different classes of disaccharides:

1. Reducing disaccharides, in which one monosaccharide, the reducing sugar of the pair, still has a free hemiacetal unit that can perform as a reducing aldehyde group; cellobiose and maltose are examples of reducing disaccharides, each with one hemiacetal unit, the other occupied by the glycosidic bond, which prevents it from acting as a reducing agent.
2. Non-reducing disaccharides, in which the component monosaccharides bond through an acetal linkage between their anomeric centers. This results in neither monosaccharide being left with a hemiacetal unit that is free to act as a reducing agent. Sucrose and trehalose are examples of non-reducing disaccharides because their glycosidic bond is between their respective hemiacetal carbon atoms. The reduced chemical reactivity of the non-reducing sugars in comparison to reducing sugars, may be an advantage where stability in storage is important.

FORMATION

The formation of a disaccharide molecule from two monosaccharide molecules proceeds by displacing a hydroxyl radical from one molecule and a hydrogen nucleus (a proton) from the other, so that the now vacant bonds on the monosaccharides join the two monomers together. The vacant bonds on the hydroxyl radical and the proton unite in their turn, forming a molecule of water, that then goes free. Because of the removal of the water molecule from the product, the term of convenience for such a process is “dehydration reaction” (also “condensation reaction” or “dehydration synthesis”). For example, milk sugar (lactose) is a disaccharide made by condensation of one molecule of each of the monosaccharides glucose and galactose, whereas the disaccharide sucrose in sugar cane and sugar beet, is a condensation product of glucose and fructose. Maltose, another common disaccharide, is condensed from two glucose molecules. The dehydration reaction that bonds monosaccharides into disaccharides (and also bonds monosaccharides into more complex polysaccharides) forms what are called glycosidic bonds.

PROPERTIES

The glycosidic bond can be formed between any hydroxyl group on the component monosaccharide. So, even if both component sugars are the same (*e.g.*, glucose), different bond combinations (regiochemistry) and stereochemistry (*alpha*- or *beta*-) result in disaccharides that are diastereoisomers with different chemical and

physical properties. Depending on the monosaccharide constituents, disaccharides are sometimes crystalline, sometimes water-soluble, and sometimes sweet-tasting and sticky-feeling.

ASSIMILATION

Digestion involves breakdown to the monosaccharides.

COMMON DISACCHARIDES

Disaccharide	Unit 1	Unit 2	Bond
Sucrose (<i>table sugar, cane sugar, beet sugar, or saccharose</i>)	Glucose	Fructose	$\alpha(1\rightarrow2)\beta$
Lactulose	Galactose	Fructose	$\beta(1\rightarrow4)$
Lactose (<i>milk sugar</i>)	Galactose	Glucose	$\beta(1\rightarrow4)$
Maltose (<i>malt sugar</i>)	Glucose	Glucose	$\alpha(1\rightarrow4)$
Trehalose	Glucose	Glucose	$\alpha(1\rightarrow1)\alpha$
Cellobiose	Glucose	Glucose	$\beta(1\rightarrow4)$
Chitobiose	Glucosamine	Glucosamine	$\beta(1\rightarrow4)$

Less common disaccharides include:

Disaccharide	Units	Bond
Kojibiose	Two glucose monomers	$\alpha(1\rightarrow2)$
Nigerose	Two glucose monomers	$\alpha(1\rightarrow3)$
Isomaltose	Two glucose monomers	$\alpha(1\rightarrow6)$
β,β -Trehalose	Two glucose monomers	$\beta(1\rightarrow1)\beta$
α,β -Trehalose	Two glucose monomers	$\alpha(1\rightarrow1)\beta$
Sophorose	Two glucose monomers	$\beta(1\rightarrow2)$
Laminaribiose	Two glucose monomers	$\beta(1\rightarrow3)$
Gentiobiose	Two glucose monomers	$\beta(1\rightarrow6)$
Turanose	A glucose monomer and a fructose monomer	$\alpha(1\rightarrow3)$
Maltulose	A glucose monomer and a fructose monomer	$\alpha(1\rightarrow4)$
Palatinose	A glucose monomer and a fructose monomer	$\alpha(1\rightarrow6)$
Gentiobiulose	A glucose monomer and a fructose monomer	$\beta(1\rightarrow6)$
Mannobiose	Two mannose monomers	either $\alpha(1\rightarrow2)$, $\alpha(1\rightarrow3)$, $\alpha(1\rightarrow4)$, or $\alpha(1\rightarrow6)$
Melibiose	A galactose monomer and a glucose monomer	$\alpha(1\rightarrow6)$
Melibiulose	A galactose monomer and a fructose monomer	$\alpha(1\rightarrow6)$
Rutinose	A rhamnose monomer and a glucose monomer	$\alpha(1\rightarrow6)$
Rutinulose	A rhamnose monomer and a fructose monomer	$\beta(1\rightarrow6)$
Xylobiose	Two xylopyranose monomers	$\beta(1\rightarrow4)$

CELLOBIOSE

Cellobiose is a disaccharide with the formula $C_{12}H_{22}O_{11}$. Cellobiose, a reducing sugar, consists of two β -glucose molecules linked by a $\beta(1\rightarrow4)$ bond. It can be hydrolyzed to glucose enzymatically or with acid. Cellobiose has eight free alcohol (OH) groups, one acetal linkage and one hemiacetal linkage, which give rise to strong inter- and intramolecular hydrogen bonds. It can be obtained by enzymatic or acidic hydrolysis of cellulose and cellulose rich materials such as cotton, jute, or paper. Treatment of cellulose with acetic anhydride and sulfuric acid, gives cellobiose octoacetate, which is no longer a hydrogen bond donor (though it is still a hydrogen bond acceptor) and is soluble in nonpolar organic solvents.

ISOMALTOSE

Isomaltose is a disaccharide similar to maltose, but with a $\alpha(1\rightarrow6)$ -linkage instead of the $\alpha(1\rightarrow4)$ -linkage. Both of the sugars are glucose, which is a pyranose sugar. Isomaltose is a reducing sugar. Isomaltose is produced

when high maltose syrup is treated with the enzyme transglucosidase (TG) and is one of the major components in the mixture isomaltooligosaccharide. It is a product of the caramelization of glucose.

ISOMALTULOSE

Isomaltulose is a disaccharide carbohydrate composed of glucose and fructose. The glucose and fructose are linked by an α -1,6-glycosidic bond (chemical name: 6-*O*- α -D-glucopyranosyl-D-fructose). Isomaltulose is present in honey and sugarcane extracts. It tastes similar to sucrose (table sugar) with half the sweetness. Isomaltulose is also known by the trade name Palatinose, which is manufactured by enzymatic rearrangement (isomerization) of sucrose from beet sugar. The enzyme and its source were discovered in Germany in 1950, and since, then its physiological role and physical properties have been studied extensively. Isomaltulose has been used as an alternative to sugar in foods in Japan since, 1985, in the EU since, 2005, in the US since, 2006, and in Australia and New Zealand since, 2007, besides other countries worldwide.

Analytical methods for characterization and assay of commercial isomaltulose are laid down, for example, in the Food Chemicals Codex. Its physical properties closely resemble those of sucrose, making it easy to use in existing recipes and processes. Isomaltulose is hydrogenated to produce isomalt, a minimally digestible carbohydrate that like dietary fibre is fermented in the large intestine or colon to short-chain fatty acids. Isomalt is used as a sugar replacer, for example in sugar-free candies and confectionery. Like sucrose, isomaltulose can be digested to glucose and fructose. However, while in sucrose the glucose is linked to the anomeric carbon of the fructose (an α -1,2 linkage), in isomaltulose the linkage is to the 6 carbon (α -1,6), making isomaltulose a reducing sugar, unlike sucrose. The fructose in isomaltulose exists in a ring structure that readily opens to exhibit a carbonyl group as in ketones and aldehydes, which explains why isomaltulose is a reducing sugar. In comparison with sucrose and most other carbohydrates, isomaltulose is not a significant substrate for oral bacteria. Consequently acid production from isomaltulose in the mouth is too slow to promote tooth decay).

Function

In nutrition, isomaltulose is a source of food energy. It provides the same amount of energy as does sucrose. The energy from both isomaltulose and sucrose is used to fuel physical, mental, physiological and metabolic activities. By way of the heat release during fueling these activities both sucrose and isomaltulose help to keep the body warm. In foods, both isomaltulose and sucrose are sweeteners. Isomalt is less sweet than sucrose.. In food preparation and processing, both isomaltulose and sucrose have similar characteristics allowing recipes that use sucrose able to use isomaltulose instead or together.

Available Carbohydrate

Isomaltulose is an available carbohydrate like sucrose and most others sugars or maltodextrins. Available or “nutritive” carbohydrates serve the body as a source of energy for its physiological functions.

Whether a carbohydrate becomes available to the body largely depends on its digestion, absorption, and metabolism,

- A carbohydrate that is absorbed (with or without prior digestion) and is fully metabolized in humans (*i.e.*, is not excreted into urine) is known to nutritionists as ‘available carbohydrate’.

When present in foods and beverages that are ingested by humans, isomaltulose is digested completely and absorbed. Its intestinal digestion involves the enzyme isomaltase, which is located at the surface of the brush border lining the inner wall of the small intestine. This enzyme is otherwise involved in the digestion of α -1,6 linkages present in starch. The products of isomaltulose digestion are glucose and fructose, which are absorbed and enter blood. Once absorbed, the glucose and fructose follow the same metabolic pathways through the body

as if they were derived from sucrose. While fructose is mostly converted to glucose or glycogen stores in the liver, glucose from the small intestine and liver is distributed via blood to different parts of the body where it serves cellular metabolism as an energy source directly or indirectly after storage as glycogen in the tissues of the body especially in skeletal muscle.

Source of Energy

Based on the heat released on complete burning of isomaltulose in oxygen to carbon dioxide and water, and its categorization as availability of carbohydrate, the food energy value of isomaltulose is identical to that of sucrose. For both, it is 4 kcal/g (17 kJ/g), a value that is used in food labelling or dietary planning.

Slow and Sustained Release of Carbohydrate and Energy

Isomaltulose is slow to be digested, absorbed and so is released as glucose and fructose into the bloodstream gradually. After ingestion, the enzymatic digestion of sucrose and isomaltulose occur on the same sucrase-isomaltase enzyme complex, which is located in the small intestine. Several studies show that this complex breaks down isomaltulose slowly compared to sucrose. The maximum rate at which an enzyme can use a substrate is known as the $V_{\max}$. The $V_{\max}$ of isomaltase for isomaltulose is 4.5 times slower than that of sucrase for sucrose. As a result of its slow digestion, isomaltulose reaches more distally in the human small intestine than does sucrose, as evidenced by their different incretin hormones responses. The incretin hormone glucose-dependent insulintropic polypeptide (GIP) is secreted from the earlier (proximal) part of the small intestine in lower amounts after isomaltulose than for sucrose ingestion, whereas the incretin hormone glucagon-like peptide-1 (GLP-1) is secreted from a later (distal) part of the small intestine in higher amounts with isomaltulose compared with sucrose. Compared with sucrose, the absorption of energy as carbohydrate from isomaltulose is prolonged, as was illustrated for example by Ang and Linn in 2014. The resulting sustained energy supply to the body from isomaltulose is reflected in the prolonged shape of the blood glucose concentration response curve.

Lower Blood Glucose and Insulin Response

The lower blood glucose and insulin concentrations after isomaltulose than after sucrose (or glucose) ingestion is evident for several population subgroups. Consequently, isomaltulose has a low glycemic index (GI). Using standardized methodology, isomaltulose has a GI of 32 as reported the University of Sydney Glycaemic Index Research Service, which list isomaltulose in an online searchable database GI database. The GI value of 32 compared to 67 for sucrose, and 100 for glucose, makes isomaltulose a particularly low-GI carbohydrate ($GI < 55$). Confirmation of a low glycaemic response to isomaltulose is provided in numerous studies for different population groups including healthy people, overweight or obese persons, prediabetic persons, and type 1 or type 2 diabetes patients (*e.g.*). Among these studies, all show the lower blood glucose response of isomaltulose and where tested also show the associated reduction in the blood insulin response. A significant role for the incretin hormone GLP-1 has been established, which occurs in response to distal carbohydrate absorption and limits the rise in blood glucose concentration after a meal. A claim corresponding to the low glycemic response of isomaltulose and its potential to lower the blood glucose response to foods when replacing other sugars has been approved as laid down in EU legislation following the publication of a positive opinion from the European Food Safety Authority.

Improvements to Blood Glucose Control

In the long-term when eating a diet including carbohydrate, reducing undesirably high concentrations of glucose in blood, and the associated demand for insulin, is supportive of the prevention and management of

diabetes mellitus, cardiovascular disease, and possibly overweight and obesity—as indicated by the International Carbohydrate Quality Consortium consensus of expert nutrition scientists. Continuous monitoring of 24-h blood glucose concentration following diets including isomaltulose instead of sucrose lowers the blood glucose profile over the day, as a result of a lower blood glucose response to individual meals. A lower glycemic diet can be achieved by choosing foods with low or reduced glycemic properties instead of many commonly consumed foods.

More specifically by choosing lower GI foods from within each food group (fruit, vegetable, whole grains, *etc.*). Isomaltulose in place of sucrose and many other carbohydrates has enabled the glycemic profile of foods to be reduced. Several studies provide evidence of improvements in both blood glucose control and lipid metabolism in both diabetic and non-diabetic persons upon regular consumption of isomaltulose when compared with other carbohydrates such as sucrose, maltodextrin, or glucose.

Effect on Fat Oxidation

When compared with other ingested carbohydrates, isomaltulose ingestion associates with higher rates of fat oxidation and lower rates of fat storage. Mechanistically this involves a lower blood glucose concentration, which then provides a lower stimulus to insulin secretion, which in turn allows more fatty acids to be released from adipose tissue for oxidation as an energy source. The lower insulin concentration also decreases carbohydrate oxidation so allowing more fatty acids to be oxidized. A lower insulin concentration also lowers the rate of liver free-fatty acid recycling via plasma triglycerides and reduces the storage of triglycerides in adipose tissue.

Practical implications seem to be several, higher rates of fat oxidation occur after isomaltulose than after higher glycaemic carbohydrates ingestion in many studies, yet the studies have different foci:

- Isomaltulose in weight management and body composition: Studies have looked at the effects on fat oxidation and other metabolic responses when replacing sugars with isomaltulose in meals (or drinks) taken by healthy or overweight to obese adults, with or without impaired glucose tolerance, and while largely sedentary. These studies have shown isomaltulose to have a role in reducing body fatness, at least central obesity. Abdominal fat decreases when consuming isomaltulose instead of sucrose (sugar replacement) or instead of breakfast calories (largely carbohydrate replacement).
- Isomaltulose in physical activity and sports nutrition: Others studies have looked at the potential benefits of slow and sustained release of carbohydrate during physical activity. Using isomaltulose in place of other ingested carbohydrates, higher rates of fat oxidation also occur during endurance activities, when preserving glycogen is important. In addition, trials using a recovery protein drink have shown that incorporating isomaltulose and a nutritional supplement (β -hydroxy- β -methylbutyrate) may help recovery from resistance exercise—so reducing of muscle damage and improving athletic performance.
- Isomaltulose in type 1 diabetes patients engaging in physical activity: In people having type 1 diabetes, the ingestion of isomaltulose instead of glucose during moderate carbohydrate loading prior to exercise improves blood glucose control, protects against hypoglycemia while maintaining running performance. The reduced risk of exercise-induced hypoglycemia arises in part from need of less insulin by injection (50% lower) when using isomaltulose and in part from the higher contribution of fat oxidation to energy metabolism, which preserves glycogen stores to counter the risk of hypoglycemia.

Cognitive Performance (Mood and Memory)

Carbohydrates by way of the glucose they supply can improve cognitive performance (mood and memory). Several studies comparing isomaltulose with higher glycaemic carbohydrates taken at breakfast show improvements in mood and memory in healthy children, middle-aged adults, and aged adult.

Oral Health

Isomaltulose is ‘kind to teeth’. Fermentation of carbohydrates by bacteria in the mouth (especially on the teeth) is responsible for the formation of dental plaque and oral acids. The acid initiates tooth demineralisation and tooth decay (dental caries). Isomaltulose largely resists fermentation by oral bacteria and is the first carbohydrate of its kind to show by pH telemetry a negligible acid production on teeth. The evidence is strong and provides the basis for ‘toothkind’ claims approved by both the Food and Drug Administration in the USA and European authorities following a positive opinion from the European Food Safety Authority.

Use

Isomaltulose is used in foods, drinks and health products owing to several of its properties:

- *Taste profile:* Isomaltulose is used as a food ingredient and alternative to other sugars and maltodextrins in foods and beverages. In these it provides a natural sucrose-like sweetness profile with a sweetening power about half that of sucrose, and no aftertaste.
- *Moisture:* Isomaltulose has very low moisture absorption (hygroscopy) therefore it has free-flowing properties in instant powders, which because of their low risk of lumping can easily be made up into drink or other type of product.
- *Stability:* Isomaltulose is highly stable during processing, including acidic conditions and environments where bacteria might grow. In sports beverages, for instance, isotonicity (osmotic pressure equal to that of fluids in the body) can be maintained during storage over the beverage’s shelf-life.
- *Applications:* Examples of products in which isomaltulose finds application include: baked goods, pastry glazings and icings, breakfast cereals, cereal bars, dairy produce, sugar confectionery (*e.g.*, chocolates, jellies, chewy confections and chewing or bubble gum), frozen desserts, fruit-juice beverages, malt beverages, sports beverages, energy drinks, instant drinks, and special and clinical nutrition feeds.
- *Recognition:* The use of isomaltulose in foods and drinks is recognised in many regions worldwide. For example, it is generally recognized as safe (GRAS) by the U.S., Food and Drug Administration, is approved as a novel food by the European Commission, and in Japan has the status FOSHU (food for specific health use).

LACTOSE

Lactose is a disaccharide. It is a sugar composed of galactose and glucose. Lactose makes up around 2–8% of milk (by weight). The name comes from *lac* (gen. *lactis*), the Latin word for milk, plus the suffix *-ose* used to name sugars. The compound is a white, water-soluble, non-hygroscopic solid with a mildly sweet taste. It is used in the food industry.

Structure and Reactions

Lactose is a disaccharide derived from the condensation of galactose and glucose, which form a β -1 $\rightarrow$ 4 glycosidic linkage. Its systematic name is β -D-galactopyranosyl-(1 $\rightarrow$ 4)-D-glucose. The glucose can be in either the α -pyranose form or the β -pyranose form, whereas the galactose can only have the β -pyranose form: hence, α -lactose and β -lactose refer to the anomeric form of the glucopyranose ring alone.

Lactose is hydrolysed to glucose and galactose, isomerised in alkaline solution to lactulose, and catalytically hydrogenated to the corresponding polyhydric alcohol, lactitol. Lactulose is a commercial product, used for treatment of constipation.

Isolation

Several million tons are produced annually as a by-product of the dairy industry. Whey is made up of 6.5% solids of which 4.8% is lactose, which is purified by crystallisation. Whey or milk plasma is the liquid remaining after milk is curdled and strained, for example in the production of cheese. Lactose comprises about 2–8% of milk by weight. Industrially, lactose is produced from whey permeate—that is whey filtrated for all major proteins. The protein fraction is used in infant nutrition and sport nutrition while the permeate can be evaporated to 60–65% solids and crystallized while cooling. Lactose can also be isolated by dilution of whey with ethanol.

Metabolism

Infant mammals nurse on their mothers to drink milk, which is rich in lactose. The intestinal villi secrete the enzyme lactase (β -D-galactosidase) to digest it. This enzyme cleaves the lactose molecule into its two subunits, the simple sugars glucose and galactose, which can be absorbed. Since, lactose occurs mostly in milk, in most mammals, the production of lactase gradually decreases with maturity due to a lack of continuing consumption.

Many people with ancestry in Europe, West Asia, South Asia, the Sahel belt in West Africa, East Africa and a few other parts of Central Africa maintain lactase production into adulthood. In many of these areas, milk from mammals such as cattle, goats, and sheep is used as a large source of food. Hence, it was in these regions that genes for lifelong lactase production first evolved. The genes of adult lactose tolerance have evolved independently in various ethnic groups. By descent, more than 70% of western Europeans can drink milk as adults, compared with less than 30% of people from areas of Africa, eastern and south-eastern Asia and Oceania. In people who are lactose intolerant, lactose is not broken down and provides food for gas-producing gut flora, which can lead to diarrhea, bloating, flatulence, and other gastrointestinal symptoms.

Lactation

Lactose's synthesis is essentially one-way and is unable to cross the Golgi membrane. The resulting accumulation of lactose is the primary source of osmotic pressure, drawing in the bulk of the water content of milk.

Applications

Its mild flavour and easy handling properties has led to its use as a carrier and stabiliser of aromas and pharmaceutical products. Lactose is not added directly to many foods, because its solubility is less than that of other sugars commonly used in food. Infant formula is a notable exception, where the addition of lactose is necessary to match the composition of human milk. Lactose is not fermented by most yeast during brewing, which may be used to advantage. For example, lactose may be used to sweeten stout beer; the resulting beer is usually called a milk stout or a cream stout. Yeast belonging to the genus *Kluyveromyces* have a unique industrial application as they are capable of fermenting lactose for ethanol production. Surplus lactose from the whey by-product of dairy operations is a potential source of alternative energy. Another significant lactose use is in the pharmaceutical industry. Lactose is added to tablet and capsule drug products as an ingredient because of its physical and functional properties, *i.e.*, compressibility and cost effective use. For similar reasons it can be used to cut (dilute) illicit drugs.

LACTULOSE

Lactulose is a non-absorbable sugar used in the treatment of constipation and hepatic encephalopathy. It is used by mouth for constipation and either by mouth or in the rectum for hepatic encephalopathy. It generally

begins working after eight to twelve hours but may take up to two days to improve constipation. Common side effects include abdominal bloating and cramps. There is the potential for electrolyte problems to occur as a result of diarrhea it produces. No evidence of harm to the baby has been found when used during pregnancy. It is generally regarded as safe during breastfeeding. It is classified as an osmotic laxative. Lactulose was first made in 1929 and has been used medically since the 1950s. It is on the World Health Organization's List of Essential Medicines, the most effective and safe medicines needed in a health system. It is available as a generic and brand-name product. In various countries it could be purchased for about US\$0.16 per 15 ml of syrup (10 g of lactulose) in 2015. In the United States the cost of this amount is about US\$0.63. Lactulose is made from the milk sugar lactose, which is composed of two simple sugars, galactose and glucose.

Medical Uses

Constipation

Lactulose is used in the treatment of chronic constipation in patients of all ages as a long-term treatment. Lactulose is used for chronic idiopathic constipation, *i.e.*, chronic constipation occurring without any identifiable cause. Lactulose may be used to counter the constipating effects of opioids, and in the symptomatic treatment of hemorrhoids as a stool softener. The dosage of lactulose for chronic idiopathic constipation is adjusted depending on the constipation severity and desired effect, from a mild stool softener to causing diarrhea. Dosage is reduced in case of galactosemia as most preparations contain the monosaccharide galactose due to its synthesis process.

Hyperammonemia

Lactulose is useful in treating hyperammonemia (high blood ammonia), which can lead to hepatic encephalopathy. Lactulose helps trap the ammonia (NH_3) in the colon and bind to it. It does this by using gut flora to acidify the colon, transforming the freely diffusible ammonia into ammonium (NH_4^+) which can no longer diffuse back into the blood. It is also useful for preventing hyperammonemia caused as a side effect of administration of valproic acid. Lactulose for hepatic encephalopathy generally requires relatively large oral dosages three or four times a day with episodic diarrhea and constant flatulence almost a certain side effect. People who take lactulose at this level of dosage generally end up wearing an adult diaper and plastic pants for any activities away from home or at night (with a chux pad for the bed) because the diarrhea can occur swiftly and without much warning.

Small Intestine Bacterial Overgrowth

Lactulose is used as a test of small intestine bacterial overgrowth (SIBO). Recently the reliability of it for diagnosing SIBO has been seriously questioned. A large amount of it is given with subsequent testing of molecular hydrogen gas in the breath. The test is positive if an increase in exhaled hydrogen occurs before that which would be expected by normal colonocyte digestion. An earlier result has been hypothesized to indicate digestion occurring within the small intestine. An alternate explanation for differences in results is the variance in small bowel transit time among tested subjects.

Special Populations

No evidence of harm to the baby has been found when used during pregnancy. It is generally regarded as safe during breastfeeding.

Side Effects

Common side effects of lactulose are abdominal cramping, borborygmus and flatulence. In normal individuals, overdose is considered uncomfortable, but not life-threatening. Uncommon side effects are nausea and vomiting. In sensitive individuals, such as the elderly or people with reduced kidney function, excess lactulose dosage can result in dehydration and electrolyte disturbances such as low magnesium levels. Ingestion of lactulose does not cause a weight gain because it is nondigestible with no nutritional value. Although lactulose is less likely to cause dental caries than sucrose, as a sugar there is potential for this. This should be taken into consideration when taken by people with a high susceptibility to this condition.

Mechanism of Action

It is a disaccharide (double-sugar) formed from one molecule each of the simple sugars (monosaccharides) fructose and galactose. Lactulose is not normally present in raw milk but is a product of heat-processes: the greater the heat, the greater amount of this substance (from 3.5 mg/L in low temperature pasteurized milk to 744 mg/L in in-container sterilized milk). It is produced commercially by isomerization of lactose. Lactulose is not absorbed in the small intestine nor broken down by human enzymes, thus stays in the digestive bolus through most of its course, causing retention of water through osmosis leading to softer, easier to pass stool.

It has a secondary laxative effect in the colon, where it is fermented by the gut flora, producing metabolites which have osmotic powers and peristalsis-stimulating effects (such as acetate), but also methane associated with flatulence. Lactulose is metabolized in the colon by bacterial flora to short-chain fatty acids, including lactic acid and acetic acid. These partially dissociate, acidifying the colonic contents (increasing the H^+ concentration in the gut). This favors the formation of the nonabsorbable NH_4^+ from NH_3 , trapping NH_3 in the colon and effectively reducing plasma NH_3 concentrations. Lactulose is therefore effective in treating hepatic encephalopathy. Specifically it is effective as secondary prevention of hepatic encephalopathy in people with cirrhosis. Moreover, recent studies showed improved cognitive functions of in people with cirrhosis with minimal hepatic encephalopathy treated with lactulose.

MALTOSE

Maltose also known as maltobiose or malt sugar, is a disaccharide formed from two units of glucose joined with an $\alpha(1\rightarrow4)$ bond. In the isomer isomaltose, the two glucose molecules are joined with an $\alpha(1\rightarrow6)$ bond. Maltose is the two-unit member of the amylose homologous series, the key structural motif of starch. When beta-amylase breaks down starch, it removes two glucose units at a time, producing maltose. An example of this reaction is found in germinating seeds, which is why it was named after malt. Unlike sucrose, it is a reducing sugar.

History

Maltose was 'discovered' by Augustin-Pierre Dubrunfaut, although this discovery was not widely accepted until it was confirmed in 1872 by Irish chemist and brewer Cornelius O'Sullivan. Its name comes from malt, combined with the suffix '-ose' which is used in names of sugars and other carbohydrates. Maltose making and use in China goes as far back as the Shang dynasty. In Japan, history of maltose has been recorded since, the period of Emperor Jimmu.

Structure and Nomenclature

Maltose is a disaccharide: The carbohydrates are generally divided into monosaccharides, oligosaccharides, and polysaccharides depending on the number of sugar subunits. Maltose, with two sugar units, is an oligosaccharide, specifically a disaccharide, because it consists of two glucose molecules. Glucose is a hexose: a monosaccharide containing six carbon atoms.

The two glucose units are in the pyranose form and are joined by an O-glycosidic bond, with the first carbon (C_1) of the first glucose linked to the fourth carbon (C_4) of the second glucose, indicated as $(1 \rightarrow 4)$. The link is characterized as α because the glycosidic bond to the anomeric carbon (C_1) is in the opposite plane from the CH_2OH substituent in the same ring (C_6 of the first glucose). If the glycosidic bond to the anomeric carbon (C_1) were in the same plane as the CH_2OH substituent, it would be classified as a $\beta(1 \rightarrow 4)$ bond, and the resulting molecule would be cellobiose. The anomeric carbon (C_1) of the second glucose molecule, which is not involved in a glycosidic bond, could be either an α - or β -anomer depending on the bond direction of the attached hydroxyl group relative to the CH_2OH substituent of the same ring, resulting in either α -maltose or β -maltose. An isomer of maltose is isomaltose. This is similar to maltose but instead of a bond in the $\alpha(1 \rightarrow 4)$ position, it is in the $\alpha(1 \rightarrow 6)$ position, the same bond that is found at the branch points of glycogen and amylopectin.

Properties

Like glucose, maltose is a reducing sugar, because the ring of one of the two glucose units can open to present a free aldehyde group; the other one cannot because of the nature of the glycosidic bond. Maltose can be broken down to glucose by the maltase enzyme, which catalyses the hydrolysis of the glycosidic bond. Maltose in aqueous solution exhibits mutarotation, because the α and β isomers that are formed by the different conformations of the anomeric carbon have different specific rotations, and in aqueous solutions, these two forms are in equilibrium. It has a sweet taste, but is only about 30-60% as sweet as sugar, depending on the concentration. A 10% solution of maltose is 35% as sweet as sucrose.

Sources and Absorption

Maltose is a component of malt, a substance which is obtained in the process of allowing grain to soften in water and germinate. It is also present in highly variable quantities in partially hydrolysed starch products like maltodextrin, corn syrup and acid-thinned starch. In humans, maltose is broken down by various maltase enzymes, providing two glucose molecules which can be further processed: either broken down to provide energy, or stored as glycogen. The lack of the sucrase-isomaltase enzyme in humans causes sucrose intolerance, but because there are four different maltase enzymes, complete maltose intolerance is extremely rare.

SUCROSE

Sucrose is common table sugar. It is a disaccharide, a molecule composed of two monosaccharides: glucose and fructose. Sucrose is produced naturally in plants, from which table sugar is refined. It has the formula $C_{12}H_{22}O_{11}$. For human consumption, sucrose is extracted, and refined, from either sugar cane or sugar beet. Sugar mills are located where sugar cane is grown to crush the cane and produce raw sugar which is shipped around the world for refining into pure sucrose. Some sugar mills also process the raw sugar into pure sucrose. Sugar beet factories are located in colder climates where the beet is grown and process the beets directly into

refined sugar. The sugar refining process involves washing the raw sugar crystals before dissolving them into a sugar syrup which is filtered and then passed over carbon to remove any residual colour. The by-now clear sugar syrup is then concentrated by boiling under a vacuum and crystallized as the final purification process to produce crystals of pure sucrose. These crystals are clear, odourless, and have a sweet taste. *En masse*, the crystals appear white. Sugar is often an added ingredient in food production and food recipes. About 175 million tonnes of sugar were produced worldwide in 2013.

Physical and Chemical Properties

Structural *O*- α -D-glucopyranosyl-(1 $\rightarrow$)- β -D-fructofuranoside

In sucrose, the components glucose and fructose are linked via an ether bond between C1 on the glucosyl subunit and C2 on the fructosyl unit. The bond is called a glycosidic linkage. Glucose exists predominantly as two isomeric “pyranoses” (α and β), but only one of these forms links to the fructose. Fructose itself exists as a mixture of “furanoses”, each of which having α and β isomers, but only one particular isomer links to the glucosyl unit. What is notable about sucrose is that, unlike most disaccharides, the glycosidic bond is formed between the reducing ends of both glucose and fructose, and not between the reducing end of one and the nonreducing end of the other. This linkage inhibits further bonding to other saccharide units. Since, it contains no anomeric hydroxyl groups, it is classified as a non-reducing sugar.

Sucrose crystallizes in the monoclinic space group $P2_1$ with room-temperature lattice parameters $a = 1.08631$ nm, $b = 0.87044$ nm, $c = 0.77624$ nm, $\beta = 102.938^\circ$. The purity of sucrose is measured by polarimetry, through the rotation of plane-polarized light by a solution of sugar. The specific rotation at 20°C using yellow “sodium-D” light (589 nm) is $+66.47^\circ$. Commercial samples of sugar are assayed using this parameter. Sucrose does not deteriorate at ambient conditions.

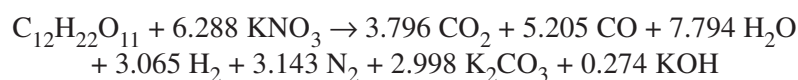
Thermal and Oxidative Degradation

Combusts to carbon dioxide and water. Mixing sucrose with the oxidizer potassium nitrate produces the fuel known as rocket candy that is used to propel amateur rocket motors.

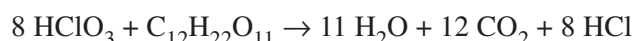


This reaction is somewhat simplified though. Some of the carbon does get fully oxidized to carbon dioxide, and other reactions, such as the water-gas shift reaction also take place.

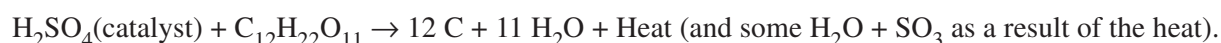
A more accurate theoretical equation is:



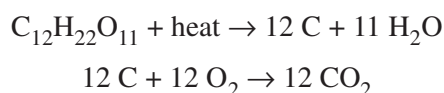
Sucrose burns with chloric acid, formed by the reaction of hydrochloric acid and potassium chlorate:



Sucrose can be dehydrated with sulfuric acid to form a black, carbon-rich solid, as indicated in the following idealized equation:



The formula for sucrose’s decomposition can be represented as 2 step reaction, first simplified reaction is dehydration of sucrose to pure Carbon and water then carbon oxidises to CO_2 with O_2 from air.



Hydrolysis

Hydrolysis breaks the glycosidic bond converting sucrose into glucose and fructose. Hydrolysis is, however, so slow that solutions of sucrose can sit for years with negligible change. If the enzyme sucrase is added, however, the reaction will proceed rapidly. Hydrolysis can also be accelerated with acids, such as cream of tartar or lemon juice, both weak acids. Likewise, gastric acidity converts sucrose to glucose and fructose during digestion, the bond between them being an acetal bond which can be broken by an acid.

Synthesis and Biosynthesis of Sucrose

The biosynthesis of sucrose proceeds via the precursors UDP-glucose and fructose 6-phosphate, catalyzed by the enzyme sucrose-6-phosphate synthase. The energy for the reaction is gained by the cleavage of Uridine diphosphate (UDP). Sucrose is formed by plants and cyanobacteria but not by other organisms. Sucrose is found naturally in many food plants along with the monosaccharide fructose. In many fruits, such as pineapple and apricot, sucrose is the main sugar. In others, such as grapes and pears, fructose is the main sugar.

Chemical Synthesis

Although sucrose is almost invariably isolated from natural sources, its chemical synthesis was first achieved in 1953 by Raymond Lemieux.

Sources

In nature, sucrose is present in many plants, and in particular their roots, fruits and nectars, because it serves as a way to store energy, primarily from photosynthesis. Many mammals, birds, insects and bacteria accumulate and feed on the sucrose in plants and for some it is their main food source. Seen from a human consumption perspective, honeybees are especially important because they accumulate sucrose and produce honey, an important foodstuff all over the world. The carbohydrates in honey itself primarily consists of fructose and glucose with trace amounts of sucrose only. As fruits ripen, their sucrose content usually rise sharply, but some fruits contain almost no sucrose at all. This includes grapes, cherries, blueberries, blackberries, figs, pomegranates, tomatoes, avocados, lemons and limes. Sucrose is a naturally occurring sugar, but with the advent of industrialization, it has been increasingly refined and consumed in all kinds of processed foods.

Production

History of Sucrose Refinement

The production of table sugar has a long history. Some scholars claim Indians discovered how to crystallize sugar during the Gupta dynasty, around AD 350. Other scholars point to the ancient manuscripts of China, dated to the 8th century BC, where one of the earliest historical mentions of sugar cane is included along with the fact that their knowledge of sugar cane was derived from India. Further, it appears that by about 500 BC, residents of present-day India began making sugar syrup and cooling it in large flat bowls to make raw table sugar crystals that were easier to store and transport. In the local Indian language, these crystals were called *khanda*, which is the source of the word *candy*.

The army of Alexander the Great was halted on the banks of river Indus by the refusal of his troops to go further east. They saw people in the Indian subcontinent growing sugarcane and making *granulated*, *salt-like sweet powder*, locally called *sakhar*, pronounced as *sakcharon* in Greek. On their return journey, the Greek

soldiers carried back some of the “honey-bearing reeds”. Sugarcane remained a limited crop for over a millennium. Sugar was a rare commodity and traders of sugar became wealthy. Venice, at the height of its financial power, was the chief sugar-distributing center of Europe. Arabs started producing it in Sicily and Spain. Only after the Crusades did it begin to rival honey as a sweetener in Europe. The Spanish began cultivating sugarcane in the West Indies in 1506 (Cuba in 1523). The Portuguese first cultivated sugarcane in Brazil in 1532.

Sugar remained a luxury in much of the world until the 18th century. Only the wealthy could afford it. In the 18th century, the demand for table sugar boomed in Europe and by the 19th century it had become regarded as a human necessity. The use of sugar grew from use in tea, to cakes, confectionery and chocolates. Suppliers marketed sugar in novel forms, such as solid cones, which required consumers to use a sugar nip, a pliers-like tool, in order to break off pieces.

The demand for cheaper table sugar drove, in part, colonization of tropical islands and nations where labour-intensive sugarcane plantations and table sugar manufacturing could thrive. Growing sugar cane crop in hot humid climates, and producing table sugar in high temperature sugar mills was harsh, inhumane work. The demand for cheap and docile labour for this work, in part, first drove slave trade from Africa (in particular West Africa), followed by indentured labour trade from South Asia (in particular India). Millions of slaves, followed by millions of indentured laborers were brought into the Caribbean, Indian Ocean, Pacific Islands, East Africa, Natal, north and eastern parts of South America, and southeast Asia. The modern ethnic mix of many nations, settled in the last two centuries, has been influenced by table sugar.

Beginning in the late 18th century, the production of sugar became increasingly mechanized. The steam engine first powered a sugar mill in Jamaica in 1768, and, soon after, steam replaced direct firing as the source of process heat. During the same century, Europeans began experimenting with sugar production from other crops. Andreas Marggraf identified sucrose in beet root and his student Franz Achard built a sugar beet processing factory in Silesia (Prussia). However, the beet-sugar industry really took off during the Napoleonic Wars, when France and the continent were cut off from Caribbean sugar. In 2010, about 20 percent of the world’s sugar was produced from beets. Today, a large beet refinery producing around 1,500 tonnes of sugar a day needs a permanent workforce of about 150 for 24-hour production.

Current Trends

Table sugar (sucrose) comes from plant sources. Two important sugar crops predominate: sugarcane (*Saccharum spp.*) and sugar beets (*Beta vulgaris*), in which sugar can account for 12% to 20% of the plant’s dry weight. Minor commercial sugar crops include the date palm (*Phoenix dactylifera*), sorghum (*Sorghum vulgare*), and the sugar maple (*Acer saccharum*). Sucrose is obtained by extraction of these crops with hot water, concentration of the extract gives syrups, from which solid sucrose can be crystallized. In 2013, worldwide production of table sugar amounted to 175 million tonnes.

Most cane sugar comes from countries with warm climates, because sugarcane does not tolerate frost. Sugar beets, on the other hand, grow only in cooler temperate regions and do not tolerate extreme heat. About 80 percent of sucrose is derived from sugarcane, the rest almost all from sugar beets.

In 2010, Brazil, India, European Union, China, Thailand, and United States were the major sugar-producing countries in the world. Brazil produced about 40 million tonnes of table sugar in 2013, while India produced 25 million, EU-27 countries 16 million, China 14 million, Thailand about 10 million, and United States over 7 million. The country rankings for table sugar production change with each year’s sugarcane crop harvest and as new sugar production plants are commissioned worldwide.

Viewed by region, Asia predominates in cane sugar production, with large contributions from India, China, Thailand, and other countries combining to account for 40% of global production in 2006. South America comes in second place with 32% of global production; Africa and Central America each produce 8% and Australia 5%. The United States, the Caribbean, and Europe make up the remainder, with roughly 3% each.

Beet sugar comes from regions with cooler climates: northwest and eastern Europe, northern Japan, plus some areas in the United States (including California). In the northern hemisphere, the beet-growing season ends with the start of harvesting around September. Harvesting and processing continues until March in some cases. The availability of processing plant capacity and the weather both influence the duration of harvesting and processing—the industry can store harvested beets until processed, but a frost-damaged beet becomes effectively unprocessable. Brazil is the world's largest sugar exporter at 29 million tonnes in the year 2013. The European Union (EU) has become the world's second-largest sugar exporter. The Common Agricultural Policy of the EU sets maximum quotas for members' production to match supply and demand, and a price. Europe exports excess production quota (approximately 5 million tonnes in 2003). Part of this, “quota” sugar, gets subsidised from industry levies, the remainder (approximately half) sells as “C quota” sugar at market prices without subsidy. These subsidies and a high import tariff make it difficult for other countries to export to the EU states, or to compete with the Europeans on world markets.

The United States sets high sugar prices to support its producers, with the effect that many former purchasers of sugar have switched to corn syrup (beverage manufacturers) or moved out of the country (candy manufacturers).

India consumes the most sugar at 26 million tonnes of table sugar in 2013. EU-27 is in second place at 18 million and China is third at above 16 million. Low prices of sugar are expected to stimulate global consumption and trade, with exports forecast 4 percent higher at 59 million tonnes. The low prices of glucose syrups produced from wheat and corn (maize) threaten the traditional sugar market. Used in combination with artificial sweeteners, they can allow drink manufacturers to produce very low-cost goods.

High-fructose Corn Syrup

In the United States, there are tariffs on the importation of sugar, and subsidies for the production of maize (corn). High-fructose corn syrup (HFCS) is significantly cheaper than refined sucrose as a sweetener. This has led to sucrose being partially displaced in U.S., industrial food production by HFCS and other non-sucrose natural sweeteners. Some people regard HFCS as unhealthy. However, clinical nutritionists, medical authorities, and the United States Food and Drug Administration have dismissed such concerns because “Sucrose, HFCS, invert sugar, honey, and many fruits and juices deliver the same sugars in the same ratios to the same tissues within the same time frame to the same metabolic pathways”. While scientific authorities agree that dietary sugars are a source of empty calories associated with certain health problems, the belief that glucose-fructose syrups such as HFCS are especially unhealthy is not supported by scientific evidence. The FDA does endorse limiting the consumption of all added sugars, including HFCS.

Types

Cane

Since, the 6th century BC, cane sugar producers have crushed the harvested vegetable material from sugarcane in order to collect and filter the juice. They then treat the liquid (often with lime (calcium oxide)) to remove impurities and then neutralize it. Boiling the juice then allows the sediment to settle to the bottom for dredging out, while the scum rises to the surface for skimming off. In cooling, the liquid crystallizes, usually in the process of stirring, to produce sugar crystals. Centrifuges usually remove the uncrystallized syrup. The producers can then either sell the sugar product for use as is, or process it further to produce lighter grades. The later

processing may take place in another factory in another country. Sugarcane is a major component of Brazilian agriculture; the country is the world's largest producer of sugarcane and its derivative products, such as crystallized sugar and ethanol (ethanol fuel).

Beet

Beet sugar producers slice the washed beets, then extract the sugar with hot water in a “diffuser”. An alkaline solution (“milk of lime” and carbon dioxide from the lime kiln) then serves to precipitate impurities. After filtration, evaporation concentrates the juice to a content of about 70% solids, and controlled crystallisation extracts the sugar. A centrifuge removes the sugar crystals from the liquid, which gets recycled in the crystalliser stages. When economic constraints prevent the removal of more sugar, the manufacturer discards the remaining liquid, now known as molasses, or sells it on to producers of animal feed. Sieving the resultant white sugar produces different grades for selling.

Cane versus beet:

- It is difficult to distinguish between fully refined sugar produced from beet and cane. One way is by isotope analysis of carbon. Cane uses C4 carbon fixation, and beet uses C3 carbon fixation, resulting in a different ratio of ^{13}C and ^{12}C isotopes in the sucrose. Tests are used to detect fraudulent abuse of European Union subsidies or to aid in the detection of adulterated fruit juice.
- Sugar cane tolerates hot climates better, but the production of sugar cane needs approximately four times as much water as the production of sugar beet. As a result, some countries that traditionally produced cane sugar (such as Egypt) have built new beet sugar factories since, about 2008. Some sugar factories process both sugar cane and sugar beets and extend their processing period in that way.
- The production of sugar leaves residues that differ substantially depending on the raw materials used and on the place of production. While cane molasses is often used in food preparation, humans find molasses from sugar beets unpalatable, and it consequently ends up mostly as industrial fermentation feedstock (for example in alcohol distilleries), or as animal feed. Once dried, either type of molasses can serve as fuel for burning.
- Pure beet sugar is difficult to find, so labelled, in the marketplace. Although some brands label their product clearly as “pure cane sugar”, beet sugar is almost always labeled simply as sugar or pure sugar. Interviews with the 5 major beet sugar-producing companies revealed that many store brands or “private label” sugar products are pure beet sugar. The lot code can be used to identify the company and the plant from which the sugar came, enabling beet sugar to be identified if the codes are known.

Culinary Sugars

Mill white:

- Mill white, also called plantation white, crystal sugar or superior sugar is produced from raw sugar. It is exposed to sulfur dioxide during the production to reduce the concentration of colour compounds and helps prevent further colour development during the crystallization process. Although common to sugarcane-growing areas, this product does not store or ship well. After a few weeks, its impurities tend to promote discoloration and clumping; therefore this type of sugar is generally limited to local consumption.

Blanco directo:

- Blanco directo, a white sugar common in India and other south Asian countries, is produced by precipitating many impurities out of cane juice using phosphoric acid and calcium hydroxide, similar to the carbonatation technique used in beet sugar refining. Blanco directo is more pure than mill white sugar, but less pure than white refined.

White refined:

- White refined is the most common form of sugar in North America and Europe. Refined sugar is made by dissolving and purifying raw sugar using phosphoric acid similar to the method used for blanco directo, a carbonation process involving calcium hydroxide and carbon dioxide, or by various filtration strategies. It is then further purified by filtration through a bed of activated carbon or bone char. Beet sugar refineries produce refined white sugar directly without an intermediate raw stage.

White refined sugar is typically sold as granulated sugar, which has been dried to prevent clumping and comes in various crystal sizes for home and industrial use:

- Coarse-grain, such as *sanding sugar* (also called “pearl sugar”, “decorating sugar”, *nibbed sugar* or *sugar nibs*) is a coarse grain sugar used to add sparkle and flavour atop baked goods and candies. Its large reflective crystals will not dissolve when subjected to heat.
- Granulated, familiar as table sugar, with a grain size about 0.5 mm across. “Sugar cubes” are lumps for convenient consumption produced by mixing granulated sugar with sugar syrup.
- Caster (or *castor*) (0.35 mm), a very fine sugar in Britain and other Commonwealth countries, so-named because the grains are small enough to fit through a castor which is small vessel with a perforated top, from which to sprinkle sugar at table. Commonly used in baking and mixed drinks, it is sold as “superfine” sugar in the United States. Because of its fineness, it dissolves more quickly than regular white sugar and is thus especially useful in meringues and cold liquids. Castor sugar can be prepared at home by grinding granulated sugar for a couple of minutes in a mortar or food processor.
- Powdered, *10X sugar*, *confectioner’s sugar* (0.060 mm), or *icing sugar* (0.024 mm), produced by grinding sugar to a fine powder. The manufacturer may add a small amount of anticaking agent to prevent clumping—either cornstarch (1% to 3%) or tri-calcium phosphate.

Brown sugar comes either from the late stages of cane sugar refining, when sugar forms fine crystals with significant molasses content, or from coating white refined sugar with a cane molasses syrup (blackstrap molasses). Brown sugar’s colour and taste becomes stronger with increasing molasses content, as do its moisture-retaining properties. Brown sugars also tend to harden if exposed to the atmosphere, although proper handling can reverse this.

Measurement

Dissolved Sugar Content

Scientists and the sugar industry use degrees Brix (symbol °Bx), introduced by Adolf Brix, as units of measurement of the mass ratio of dissolved substance to water in a liquid. A 25 °Bx sucrose solution has 25 grams of sucrose per 100 grams of liquid; or, to put it another way, 25 grams of sucrose sugar and 75 grams of water exist in the 100 grams of solution. The Brix degrees are measured using an infrared sensor. This measurement does not equate to Brix degrees from a density or refractive index measurement, because it will specifically measure dissolved sugar concentration instead of all dissolved solids. When using a refractometer, one should report the result as “refractometric dried substance” (RDS). One might speak of a liquid as having 20 °Bx RDS. This refers to a measure of percent by weight of *total* dried solids and, although not technically the same as Brix degrees determined through an infrared method, renders an accurate measurement of sucrose content, since, sucrose in fact forms the majority of dried solids. The advent of in-line infrared Brix measurement sensors has made measuring the amount of dissolved sugar in products economical using a direct measurement.

Consumption

Refined sugar was a luxury before the 18th century. It became widely popular in the 18th century, then graduated to becoming a necessary food in the 19th century. This evolution of taste and demand for sugar as an essential food ingredient unleashed major economic and social changes. Eventually, table sugar became sufficiently cheap and common enough to influence standard cuisine and flavored drinks. Sucrose forms a major element in confectionery and desserts. Cooks use it for sweetening—its fructose component, which has almost double the sweetness of glucose, makes sucrose distinctively sweet in comparison to other carbohydrates. It can also act as a food preservative when used in sufficient concentrations. Sucrose is important to the structure of many foods, including biscuits and cookies, cakes and pies, candy, and ice cream and sorbets. It is a common ingredient in many processed and so-called “junk foods”.

Nutritional Information

Fully refined sugar is 99.9% sucrose, thus providing only carbohydrate as dietary nutrient and 390 kilocalories per 100 g serving (USDA data, right table). There are no micronutrients of significance in fully refined sugar (right table).

Metabolism of Sucrose

In humans and other mammals, sucrose is broken down into its constituent monosaccharides, glucose and fructose, by sucrase or isomaltase glycoside hydrolases, which are located in the membrane of the microvilli lining the duodenum. The resulting glucose and fructose molecules are then rapidly absorbed into the bloodstream. In bacteria and some animals, sucrose is digested by the enzyme invertase.

Sucrose is an easily assimilated macronutrient that provides a quick source of energy, provoking a rapid rise in blood glucose upon ingestion. Sucrose, as a pure carbohydrate, has an energy content of 3.94 kilocalories per gram (or 17 kilojoules per gram). When large amounts of refined food that contain high percentages of sucrose are consumed, beneficial nutrients can be displaced from the diet, which can contribute to an increased risk for chronic disease. The rapidity with which sucrose raises blood glucose can cause problems for people suffering from defective glucose metabolism, such as persons with hypoglycemia or diabetes mellitus. Sucrose can contribute to the development of metabolic syndrome. In an experiment with rats that were fed a diet one-third of which was sucrose, the sucrose first elevated blood levels of triglycerides, which induced visceral fat and ultimately resulted in insulin resistance. Another study found that rats fed sucrose-rich diets developed high triglycerides, hyperglycemia, and insulin resistance. A 2004 study recommended that the consumption of sucrose-containing drinks should be limited due to the growing number of people with obesity and insulin resistance.

Human Health

Human beings have long sought sugars, but aside from wild honey and fruits, have not had access to the large quantities that characterize the modern diet after the industrial age. Studies have indicated potential links between consumption of free sugars, including sucrose which is particularly prevalent in processed foods, and health hazards, including obesity and tooth decay. It is also considered as causing endogenous glycation processes since, it metabolises into glucose and fructose in the body.

Tooth Decay

Tooth decay (dental caries) has become a pronounced health hazard associated with the consumption of sugars, especially sucrose. Oral bacteria such as *Streptococcus mutans* live in dental plaque and metabolize any sugars (not just sucrose, but also glucose, lactose, fructose, and cooked starches) into lactic acid. The resultant lactic acid lowers the pH of the tooth's surface, stripping it of minerals in the process known as tooth decay.

All 6-carbon sugars and disaccharides based on 6-carbon sugars can be converted by dental plaque bacteria into acid that demineralizes teeth, but sucrose may be uniquely useful to *Streptococcus sanguinis* (formerly *Streptococcus sanguis*) and *Streptococcus mutans*. Sucrose is the only dietary sugar that can be converted to sticky glucans (dextran-like polysaccharides) by extracellular enzymes. These glucans allow the bacteria to adhere to the tooth surface and to build up thick layers of plaque. The anaerobic conditions deep in the plaque encourage the formation of acids, which leads to carious lesions. Thus, sucrose could enable *S. mutans*, *S. sanguinis* and many other species of bacteria to adhere strongly and resist natural removal, *e.g.*, by flow of saliva, although they are easily removed by brushing. The glucans and levans (fructose polysaccharides) produced by the plaque bacteria also act as a reserve food supply for the bacteria. Such a special role of sucrose in the formation of tooth decay is much more significant in light of the almost universal use of sucrose as the most desirable sweetening agent. Widespread replacement of sucrose by high-fructose corn syrup (HFCS) has not diminished the danger from sucrose. If smaller amounts of sucrose are present in the diet, they will still be sufficient for the development of thick, anaerobic plaque and plaque bacteria will metabolise other sugars in the diet, such as the glucose and fructose in HFCS.

Glycemic Index

Sucrose is a disaccharide made up of 50% glucose and 50% fructose and has a glycemic index of 65. Sucrose is digested rapidly, but has a relatively low glycemic index due to its content of fructose, which has a minimal effect on blood glucose. As with other sugars, sucrose is digested into its components via the enzyme sucrase to glucose (blood sugar) and fructose. The glucose component is transported into the blood (90%) and excess glucose is converted to temporary storage in the liver—named glycogen. The fructose is either bonded to cellulose and transported out the GI tract or processed by the liver into citrates, aldehydes, and, for the most part, lipid droplets (fat). As the glycemic index measures the speed at which glucose is released into the bloodstream a refined sugar containing glucose is considered high-glycemic. As with other sugars, over-consumption may cause an increase in blood sugar levels from a normal 90 mg/dL to up over 150 mg/dL (5 mmol/l to over 8.3 mmol/l).

Diabetes Mellitus

Diabetes mellitus, a disease that causes the body to metabolize sugar poorly, occurs when either:

- The body attacks the cells producing insulin, the hormone that allows the metabolizing of sugar (Type 1 diabetes)
- The body's cells exhibit impaired responses to insulin (Type 2 diabetes).

When glucose builds up in the bloodstream, it can cause two problems:

1. In the short term, cells become starved for energy because they do not have access to the glucose
2. In the long term, frequent glucose build-up increases the acidity of the blood, damaging many of the body's organs, including the eyes, kidneys, nerves, and/or heart.

Authorities advise diabetics to avoid sugar-rich foods to prevent adverse reactions.

Obesity

The National Health and Nutrition Examination Survey I and their follow-on studies as part of a series indicate that the population in the United States has increased its proportion of energy consumption from carbohydrates and decreased its proportion from total fat while obesity has increased. This implies, along with the United Nations report cited below, that obesity may correlate better with sugar consumption than with fat consumption, and that reducing fat consumption while increasing sugar consumption may increase the level of obesity. The following table summarizes this study (based on the proportion of energy intake from different food sources for US Adults 20–74 years old, as carried out by the U.S., Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD):

Year	Sex	Carbohydrate	Fat	Protein	Obesity
1971	Male	42.4%	36.9%	16.5%	12.1%
1971	Female	45.4%	36.1%	16.9%	16.6%
2000	Male	49.0%	32.8%	15.5%	27.7%
2000	Female	51.6%	32.8%	15.1%	34.0%

Added sugar is not always evident in food products. While expected in desserts, candies, and soft drinks, sugar is also added to a wide range of non-sweet items such as bread, crackers, potato chips, peanut butter, soup, salad dressing, ketchup, mayonnaise, and many other common sauces. Forms of added sugar include technically accurate, but misleading, terms such as cane juice, evaporated cane juice, corn syrup and corn syrup solids, malt syrup, rice syrup, dextrose, maltose, maltodextrin, molasses, treacle, and xylose. A 2002 study conducted by the U.S., National Academy of Sciences concluded that, due to discrepancies in data from different studies, it could not set a tolerable upper intake level, since, “there is no clear and consistent association between increased intakes of added sugars and BMI.” However, it explains that this may be due to the underreporting of the consumption of added sugars. (BMI, or “body mass index,” is a measure of weight and height used to estimate body fat.)

Gout

The occurrence of gout is connected with an excess production of uric acid. A diet rich in sucrose may lead to gout as it raises the level of insulin, which prevents excretion of uric acid from the body. As the concentration of uric acid in the body increases, so does the concentration of uric acid in the joint liquid and beyond a critical concentration, the uric acid begins to precipitate into crystals. Researchers have implicated sugary drinks high in fructose in a surge in cases of gout.

UN Dietary Recommendation

In 2015, the World Health Organization (WHO) published a new guideline on sugars intake for adults and children, as a result of an extensive review of the available scientific evidence by a multidisciplinary group of experts. The guideline recommends that both adults and children reduce the intake of free sugars (monosaccharides and disaccharides added to foods and beverages by the manufacturer, cook or consumer, and sugars naturally present in honey, syrups, fruit juices and fruit juice concentrates) to less than 10% of total energy intake. A reduction to below 5% of total energy intake brings additional health benefits, especially in what regards dental caries.

Religious Concerns

The sugar refining industry often uses bone char (calcinated animal bones) for decolorizing. About 25% of sugar produced in the U.S., is processed using bone char as a filter, the remainder being processed with activated carbon. As bone char does not seem to remain in finished sugar, Jewish religious leaders consider sugar filtered through it to be pareve, meaning that it is neither meat nor dairy and may be used with either type of food. However, the bone char must source to a kosher animal (*e.g.*, cow, sheep) for the sugar to be kosher.

Trade and Economics

One of the most widely-traded commodities in the world throughout history, sugar accounts for around 2% of the global dry cargo market. International sugar prices show great volatility, ranging from around 3 to over 60 cents per pound in the past 50 years. About 100 of the world’s 180 countries produce sugar from beet or cane, a few more refine raw sugar to produce white sugar, and all countries consume sugar. Consumption of sugar ranges from around 3 kilograms per person per annum in Ethiopia to around 40 kg/person/yr in Belgium. Consumption per capita rises with income per capita until it reaches a plateau of around 35 kg per person per

year in middle income countries. Many countries subsidize sugar production heavily. The European Union, the United States, Japan, and many developing countries subsidize domestic production and maintain high tariffs on imports. Sugar prices in these countries have often exceeded prices on the international market by up to three times; today, with world market sugar futures prices currently strong, such prices typically exceed world prices by two times. Within international trade bodies, especially in the World Trade Organization, the “G20” countries led by Brazil have long argued that, because these sugar markets in essence exclude cane sugar imports, the G20 sugar producers receive lower prices than they would under free trade. While both the European Union and United States maintain trade agreements whereby certain developing and less developed countries (LDCs) can sell certain quantities of sugar into their markets, free of the usual import tariffs, countries outside these preferred trade régimes have complained that these arrangements violate the “most favoured nation” principle of international trade. This has led to numerous tariffs and levies in the past.

In 2004, the WTO sided with a group of cane sugar exporting nations (led by Brazil and Australia) and ruled the EU sugar-régime and the accompanying ACP-EU Sugar Protocol (whereby a group of African, Caribbean, and Pacific countries receive preferential access to the European sugar market) illegal. In response to this and to other rulings of the WTO, and owing to internal pressures on the EU sugar-régime, the European Commission proposed on 22 June 2005 a radical reform of the EU sugar-régime, cutting prices by 39% and eliminating all EU sugar exports. The African, Caribbean, Pacific and least developed country sugar exporters reacted with dismay to the EU sugar proposals. On 25 November 2005, the Council of the EU agreed to cut EU sugar prices by 36% as from 2009. In 2007, it seemed that the U.S., Sugar programme could become the next target for reform. However, some commentators expected heavy lobbying from the U.S., sugar industry, which donated \$2.7 million to US House and US Senate incumbents in the 2006 US election, more than any other group of US food-growers. Especially prominent lobbyists include The Fanjul Brothers, so-called “sugar barons” who made the single largest individual contributions of soft money to both the Democratic and Republican parties in the political system of the United States of America.

Small quantities of sugar, especially specialty grades of sugar, reach the market as ‘fair trade’ commodities; the fair trade system produces and sells these products with the understanding that a larger-than-usual fraction of the revenue will support small farmers in the developing world. However, whilst the Fairtrade Foundation offers a premium of \$60.00 per tonne to small farmers for sugar branded as “Fairtrade”, government schemes such as the U.S., Sugar programme and the ACP Sugar Protocol offer premiums of around \$400.00 per tonne above world market prices. However, the EU announced on 14 September 2007 that it had offered “to eliminate all duties and quotas on the import of sugar into the EU”. The US Sugar Association has launched a campaign to promote sugar over artificial substitutes. The Association now aggressively challenges many common beliefs regarding negative side-effects of sugar consumption. The campaign aired a high-profile television commercial during the 2007 Primetime Emmy Awards on FOX Television. The Sugar Association uses the trademark tagline “Sugar: sweet by nature”.

TREHALOSE

Trehalose is a sugar consisting of two molecules of glucose. It is also known as mycose or tremalose. Some bacteria, fungi, plants and invertebrate animals synthesize it as a source of energy, and to survive freezing and lack of water. Extracting trehalose was once a difficult and costly process, but around 2000, the Hayashibara company (Okayama, Japan) confirmed an inexpensive extraction technology from starch for mass production. Trehalose has high water retention capabilities, and is used in food, cosmetics, and as a drug.

Structure

Trehalose is a disaccharide formed by a 1,1-glucoside bond between two α -glucose units. There are two other isomers, but they are not found in nature. It is found in nature as a disaccharide and also as a monomer in some polymers.

Synthesis

There are at least three biological pathways for trehalose biosynthesis. There is an industrial process where trehalose is derived from corn starch.

Chemical Properties

Trehalose is a nonreducing sugar formed from two glucose units joined by a 1–1 alpha bond, giving it the name of α -D-glucopyranosyl-(1→1)- α -D-glucopyranoside. The bonding makes trehalose very resistant to acid hydrolysis, and therefore is stable in solution at high temperatures, even under acidic conditions. The bonding also keeps nonreducing sugars in closed-ring form, such that the aldehyde or ketone end groups do not bind to the lysine or arginine residues of proteins (a process called glycation). Trehalose is less soluble than sucrose, except at high temperatures (>80 °C). Trehalose forms a rhomboid crystal as the dihydrate, and has 90% of the calorific content of sucrose in that form. Anhydrous forms of trehalose readily regain moisture to form the dihydrate. Anhydrous forms of trehalose can show interesting physical properties when heat-treated.

Trehalose aqueous solutions show a concentration-dependent clustering tendency. Owing to their ability to form hydrogen bonds between one another, they self-associate in water to form clusters of various sizes. All-atom molecular dynamics simulations have shown that when reaching a concentration of 1.5–2.2 molar, the trehalose molecular clusters percolate and form large, continuous aggregates within the system. Trehalose directly interacts with nucleic acids, facilitates melting of double stranded DNA and stabilizes single-stranded nucleic acids.

Biological Properties

Organisms ranging from bacteria, yeast, fungi, insects, invertebrates, and lower and higher plants have enzymes that can make trehalose. In nature, trehalose can be found in plants, and microorganisms. In animals, trehalose is prevalent in shrimp, and also in insects, including grasshoppers, locusts, butterflies, and bees, in which trehalose serves as blood-sugar. Trehalose is then broken down into glucose by the catabolic enzyme trehalase for use. Trehalose is also present in the nutrition exchange liquid of hornets and their larvae. Trehalose is the major carbohydrate energy storage molecule used by insects for flight. One possible reason for this is that the glycosidic linkage of trehalose, when acted upon by an insect trehalase, releases two molecules of glucose, which is required for the rapid energy requirements of flight. This is double the efficiency of glucose release from the storage polymer starch, for which cleavage of one glycosidic linkage releases only one glucose molecule. In plants, trehalose is seen in sunflower seeds, moonwort, *Selaginella* plants, and sea algae. Within the fungi, it is prevalent in some mushrooms, such as shiitake (*Lentinula edodes*), oyster, king oyster, and golden needle. Even within the plant kingdom, *Selaginella* (sometimes called the resurrection plant), which grows in desert and mountainous areas, may be cracked and dried out, but will turn green again and revive after rain because of the function of trehalose. The two prevalent theories as to how trehalose works within the organism in the state of cryptobiosis are the vitrification theory, a state that prevents ice formation, or the water displacement theory, whereby water is replaced by trehalose.

Nutritional and Dietary Properties

Trehalose is rapidly broken down into glucose by the enzyme trehalase, which is present in the brush border of the intestinal mucosa of omnivores (including humans) and herbivores. It causes less of a spike in blood sugar than glucose. Trehalose has about 45% the sweetness of sucrose at concentrations above 22%, but when the concentration is reduced, its sweetness decreases more quickly than that of sucrose, so that a 2.3% solution

tastes 6.5 times less sweet as the equivalent sugar solution. It is commonly used in prepared frozen foods, like ice cream, because it lowers the freezing point of foods. Deficiency of trehalase enzyme is unusual in humans, except in the Greenlandic Inuit, where it occurs in 10–15% of the population.

Medical Use

Trehalose is an ingredient, along with hyaluronic acid, in an artificial tears product used to treat dry eye.

Research

In 1832, H.A.L. Wiggers discovered trehalose in an ergot of rye, and in 1859 Marcellin Berthelot isolated it from trehala manna, a substance made by weevils, and named it trehalose. In January 2018 research was published showing that some strains of *Clostridium difficile* found in the human gut microbiome are especially responsive to trehalose; the study found that in the lab and in mice, these strains grew better and produced more toxins when fed trehalose. The authors pointed out a correlation between the timing of the introduction of trehalose into the food supply when the cost of production came down around 2000, and a rise in cases of hospital-acquired *C. difficile* infections, and suggested that trehalose might be causing the rise in infections. It was not known at that time whether trehalose actually caused the rise in *C. difficile* infections or affects *C. difficile* in the human gut.

TURANOSE

Turanose is a reducing disaccharide. The D-isomer is naturally occurring. Its systematic name is α -D-glucopyranosyl-(1 $\rightarrow$ 3)- α -D-fructofuranose. It is an analog of sucrose not metabolized by higher plants, but rather acquired through the action of sucrose transporters for intracellular carbohydrate signaling. In addition to its involvement in signal transduction, D-(+)-turanoose can also be used as a carbon source by many organisms including numerous species of bacteria and fungi.

NUTRITION

Carbohydrate consumed in food yields 3.87 kilocalories of energy per gram for simple sugars, and 3.57 to 4.12 kilocalories per gram for complex carbohydrate in most other foods. Relatively high levels of carbohydrate are associated with processed foods or refined foods made from plants, including sweets, cookies and candy, table sugar, honey, soft drinks, breads and crackers, jams and fruit products, pastas and breakfast cereals. Lower amounts of carbohydrate are usually associated with unrefined foods, including beans, tubers, rice, and unrefined fruit. Animal-based foods generally have the lowest carbohydrate levels, although milk does contain a high proportion of lactose.

Organisms typically cannot metabolize all types of carbohydrate to yield energy. Glucose is a nearly universal and accessible source of energy. Many organisms also have the ability to metabolize other monosaccharides and disaccharides but glucose is often metabolized first. In *Escherichia coli*, for example, the lac operon will express enzymes for the digestion of lactose when it is present, but if both lactose and glucose are present the lac operon is repressed, resulting in the glucose being used first. Polysaccharides are also common sources of energy. Many organisms can easily break down starches into glucose; most organisms, however, cannot metabolize cellulose or other polysaccharides like chitin and arabinoxylans. These carbohydrate types can be metabolized by some bacteria and protists. Ruminants and termites, for example, use microorganisms to process cellulose. Even though these complex carbohydrates are not very digestible, they represent an important dietary element for humans, called dietary fibre. Fibre enhances digestion, among other benefits. Based on the effects on risk of heart disease and obesity in otherwise healthy middle-aged adults, the Institute of Medicine recommends that American and Canadian adults get between 45–65% of dietary energy from whole-grain carbohydrates. The

Food and Agriculture Organization and World Health Organization jointly recommend that national dietary guidelines set a goal of 55–75% of total energy from carbohydrates, but only 10% directly from sugars (their term for simple carbohydrates).

CLASSIFICATION

Nutritionists often refer to carbohydrates as either simple or complex. However, the exact distinction between these groups can be ambiguous. The term *complex carbohydrate* was first used in the U.S., Senate Select Committee on Nutrition and Human Needs publication *Dietary Goals for the United States* (1977) where it was intended to distinguish sugars from other carbohydrates (which were perceived to be nutritionally superior). However, the report put “fruit, vegetables and whole-grains” in the complex carbohydrate column, despite the fact that these may contain sugars as well as polysaccharides. This confusion persists as today some nutritionists use the term complex carbohydrate to refer to any sort of digestible saccharide present in a whole food, where fibre, vitamins and minerals are also found (as opposed to processed carbohydrates, which provide energy but few other nutrients). The standard usage, however, is to classify carbohydrates chemically: simple if they are sugars (monosaccharides and disaccharides) and complex if they are polysaccharides (or oligosaccharides).

In any case, the simple vs. complex chemical distinction has little value for determining the nutritional quality of carbohydrates. Some simple carbohydrates (*e.g.*, fructose) raise blood glucose slowly, while some complex carbohydrates (starches), especially if processed, raise blood sugar rapidly. The speed of digestion is determined by a variety of factors including which other nutrients are consumed with the carbohydrate, how the food is prepared, individual differences in metabolism, and the chemistry of the carbohydrate.

The USDA's *Dietary Guidelines for Americans 2010* call for moderate-to high-carbohydrate consumption from a balanced diet that includes six one-ounce servings of grain foods each day, at least half from whole grain sources and the rest from enriched. The glycemic index (GI) and glycemic load concepts have been developed to characterize food behaviour during human digestion. They rank carbohydrate-rich foods based on the rapidity and magnitude of their effect on blood glucose levels. Glycemic index is a measure of how quickly food glucose is absorbed, while glycemic load is a measure of the total absorbable glucose in foods. The insulin index is a similar, more recent classification method that ranks foods based on their effects on blood insulin levels, which are caused by glucose (or starch) and some amino acids in food.

EFFECTS OF DIETARY CARBOHYDRATE RESTRICTION

Carbohydrates are a common source of energy in living organisms; however, no single carbohydrate is an essential nutrient in humans. Humans are able to obtain all of their energy requirement from protein and fats, though the potential for some negative health effects of extreme carbohydrate restriction remains, as the issue has not been studied extensively yet. However, in the case of dietary fibre–indigestible carbohydrates which are not a source of energy–inadequate intake can lead to significant increases in mortality.

Following a diet consisting of very low amounts of daily carbohydrate for several days will usually result in higher levels of blood ketone bodies than an isocaloric diet with similar protein content. This relatively high level of ketone bodies is commonly known as ketosis and is very often confused with the potentially fatal condition often seen in type 1 diabetics known as diabetic ketoacidosis. Somebody suffering ketoacidosis will have much higher levels of blood ketone bodies along with high blood sugar, dehydration and electrolyte imbalance. Long-chain fatty acids cannot cross the blood–brain barrier, but the liver can break these down to produce ketones. However, the medium-chain fatty acids octanoic and heptanoic acids can cross the barrier and be used by the brain, which normally relies upon glucose for its energy. Gluconeogenesis allows humans to synthesize some glucose from specific amino acids: from the glycerol backbone in triglycerides and in some cases from fatty acids.

METABOLISM

Carbohydrate metabolism denotes the various biochemical processes responsible for the formation, breakdown and interconversion of carbohydrates in living organisms.

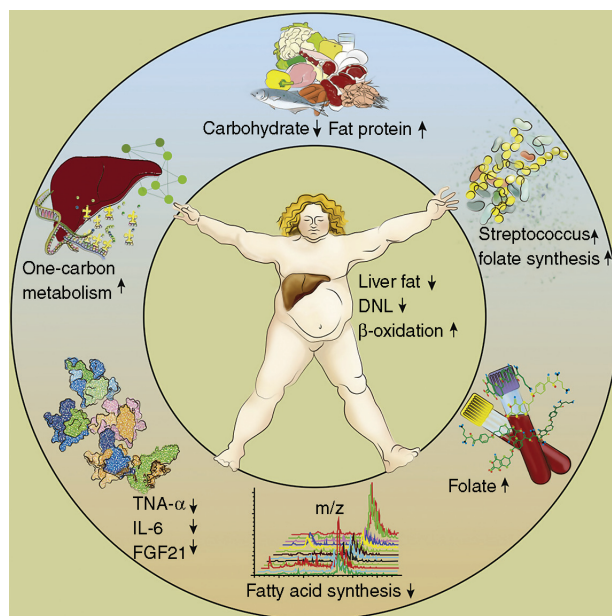


Fig. Metabolism.

The most important carbohydrate is glucose, a simple sugar (monosaccharide) that is metabolized by nearly all known organisms. Glucose and other carbohydrates are part of a wide variety of metabolic pathways across species: plants synthesize carbohydrates from carbon dioxide and water by photosynthesis storing the absorbed energy internally, often in the form of starch or lipids. Plant components are consumed by animals and fungi, and used as fuel for cellular respiration. Oxidation of one gram of carbohydrate yields approximately 4 kcal of energy, while the oxidation of one gram of lipids yields about 9 kcal. The human body stores between 300 to 500 g of carbohydrates depending on body weight, with the skeletal muscle contributing to a large portion of the storage. Energy obtained from metabolism (*e.g.*, oxidation of glucose) is usually stored temporarily within cells in the form of ATP. Organisms capable of aerobic respiration metabolize glucose and oxygen to release energy with carbon dioxide and water as byproducts.

CATABOLISM

Catabolism is the metabolic reaction which cells undergo to extract energy. There are two major metabolic pathways of monosaccharide catabolism: glycolysis and the citric acid cycle. In glycolysis, oligo/polysaccharides are cleaved first to smaller monosaccharides by enzymes called glycoside hydrolases. The monosaccharide units can then enter into monosaccharide catabolism. In some cases, as with humans, not all carbohydrate types are usable as the digestive and metabolic enzymes necessary are not present.

CARBOHYDRATE CHEMISTRY

Carbohydrate Acetalisation

In carbohydrate chemistry carbohydrate acetalisation is an organic reaction and a very effective means of providing a protecting group.

The example below depicts the acetalisation reaction of D-ribose:

1. With acetone or 2,2-dimethoxypropane as the acetalisation reagent the reaction is under thermodynamic reaction control and results in the pentose
2. The latter reagent in itself is an acetal and therefore the reaction is actually a cross-acetalisation. Kinetic reaction control results from 2-methoxypropene as the reagent. D-ribose in itself is a hemiacetal and in equilibrium with the pyranose
3. In aqueous solution ribose is 75% pyranose and 25% furanose and a different acetal 4 is formed.

Selective acetalization of carbohydrate and formation of acetals possessing atypical properties is achieved by using arylsulfonyl acetals. An example of arylsulfonyl acetals as carbohydrate-protective groups are phenylsulfonylethylidene acetals. These acetals are resistant to the acid hydrolysis and can be deprotected easily by classical reductive conditions.

CYANOHYDRIN REACTION

A cyanohydrin reaction is an organic chemical reaction by an aldehyde or ketone with a cyanide anion or a nitrile to form a cyanohydrin. This nucleophilic addition is a reversible reaction but with aliphatic carbonyl compounds equilibrium is in favor of the reaction products. The cyanide source can be potassium cyanide, sodium cyanide or trimethylsilyl cyanide. With aromatic aldehydes such as benzaldehyde, the benzoin condensation is a competing reaction. The reaction is used in carbohydrate chemistry as a chain extension method for example that of D-xylose.

LOBRY DE BRUYN–VAN EKENSTEIN TRANSFORMATION

In carbohydrate chemistry, the Lobry de Bruyn–van Ekenstein transformation also known as the Lobry de Bruyn–Alberda–van Ekenstein transformation is the base or acid catalyzed transformation of an aldose into the ketose isomer or vice versa, with a tautomeric enediol as reaction intermediate. Ketoses may be transformed into 3-ketoses, etcetera. The enediol is also an intermediate for the epimerization of an aldose or ketose. The reactions are usually base catalyzed, but can also take place under acid or neutral conditions. A typical rearrangement reaction is that between the aldose glyceraldehyde and the ketose dihydroxyacetone in a chemical equilibrium. The Lobry de Bruyn–van Ekenstein transformation is relevant for the industrial production of certain ketoses and was discovered in 1885 by Cornelis Adriaan Lobry van Troostenburg de Bruyn and Willem Alberda van Ekenstein.

Aldose-ketose Transformation

The following scheme describes the interconversion between an aldose and a ketose, where R is any organic residue. The equilibrium or the *reactant to product ratio* depends on concentration, solvent, pH and temperature. At equilibrium the aldose and ketose form a mixture which in the case of the glyceraldehyde and dihydroxyacetone is also called glycerose. A related reaction is the alpha-ketol rearrangement.

Epimerization

The carbon atom at which the initial deprotonation takes place is a stereocenter. If, for example, D-glucose (an Aldose) rearranges to D-fructose, the ketose, the stereochemical configuration is lost in the enol form. In the chemical reaction the enol can be protonated from two faces, resulting in the backformation of glucose or the formation of the epimer D-mannose. The final product is a mix of D-glucose, D-fructose and D-mannose.

AMADORI REARRANGEMENT

The Amadori rearrangement is an organic reaction describing the acid or base catalyzed isomerization or rearrangement reaction of the *N*-glycoside of an aldose or the glycosylamine to the corresponding 1-amino-1-deoxy-ketose. The Amadori rearrangement was discovered by the organic chemist Mario Amadori (1886–1941), who in 1925 reported this reaction while studying the Maillard reaction. The reaction is important in carbohydrate chemistry.

The reaction mechanism is demonstrated starting from the reaction of D-mannose in its closed: (1) and open-form (2) with ammonia to produce the 1,1-amino-alcohol (3), which is unstable and loses water to the glycosylamine (again the open imine (5) and the closed form hemiaminal (4)), which is the starting point for the actual Amadori rearrangement. By treatment of the glycosylamine with pyridine and acetic anhydride, the imine group rearranges and the intermediate enol, in turn, rearranges to the ketone. In this particular reaction, all the alcohol and amino groups are acylated as well. The reaction is associated with the Maillard reaction in which the reagents are naturally occurring sugars and amino acids. A recent study has revealed the possibility of Amadori rearrangement during interaction between oxidized dextran and gelatine.

Amadori Product

An Amadori product is an intermediate in the production of an advanced glycation end-product (AGE) as a result of glycation.

The formation of an advanced glycation end-product involves the following steps:

- *Formation of a Schiff base:* for example the aldehyde group of a glucose molecule will combine with the amino group of a lysine molecule (in a protein) to form an imine or Schiff base, which is a double bond between the carbon atom of the glucose and the nitrogen atom of the lysine.
- *Formation of an Amadori product:* the Amadori product is a re-arrangement from the Schiff base, wherein the hydrogen atom from the hydroxyl group adjacent to the carbon-nitrogen double bond moves to bond to the nitrogen, leaving a ketone.
- *Formation of an advanced glycation end-product (AGE):* the Amadori product is oxidized, most often by transition metal catalysis.

The first two steps in this reaction are both reversible, but the last step is irreversible.

NEF REACTION

The Nef reaction is an organic reaction describing the acid hydrolysis of a salt of a primary or secondary nitroalkane (1) to an aldehyde or a ketone (3) and nitrous oxide (4). The reaction has been the subject of several literature reviews. The reaction was reported in 1894 by the chemist John Ulric Nef, who treated the sodium salt of nitroethane with sulfuric acid resulting in an 85–89% yield of nitrous oxide and at least 70% yield of acetaldehyde. However, the reaction was pioneered a year earlier in 1893 by Konovalov, who converted the potassium salt of 1-phenylnitroethane with sulfuric acid to acetophenone.

The *Nef reaction* should not be confused with the Nef synthesis.

Reaction Mechanism

The reaction mechanism starting from the nitronate salt as the resonance structures 1a and 1b is depicted below: The salt is protonated forming the nitronic acid 2 (in some cases these nitronates have been isolated) and once more to the iminium ion 3. This intermediate is attacked by water in a nucleophilic addition forming 4 which loses a proton and then water to the 1-nitroso-alkanol 5 which is believed to be responsible for the deep-blue colour of the reaction mixture in many Nef reactions. This intermediate rearranges to hyponitrous acid 6

(forming nitrous oxide 6c through 6b) and the oxonium ion 7 which loses a proton to form the carbonyl compound. Note that formation of the nitronate salt from the nitro compound requires an alpha hydrogen atom and therefore the reaction fails with tertiary nitro compounds.

Scope

The Nef reaction is frequently encountered in organic synthesis. It has been applied in carbohydrate chemistry as a chain-extension method for aldoses for example in the isotope labeling of C¹⁴-D-mannose and C¹⁴-D-glucose from D-arabinose and C14-nitromethane (the first step here is a Henry reaction): The opposite reaction is the Wohl degradation. The reaction is also used in combination with the Michael reaction in the synthesis of γ -keto-carbonyls such as: or 2,5-heptanedione Hydrolysis of nitro compounds with strong acid without the intermediate salt stage results in the formation of carboxylic acids and hydroxylamine salts. The hydrolysis step of the Nef reaction can also be performed with Lewis acids such as tin(IV) chloride and iron(III) chloride or oxidizing agents, such as oxone.

WOHL DEGRADATION

The Wohl degradation in carbohydrate chemistry is a chain contraction method for aldoses. The classic example is the conversion of glucose to arabinose as shown below. The reaction is named after the German chemist Alfred Wohl (1863–1939). In one modification, d-glucose is converted to the glucose oxime by reaction with hydroxylamine and sodium methoxide. In the second step the pentaacetyl glycononitrile is formed by reaction with acetic anhydride in acetic acid with sodium acetate. In this reaction step the oxime is converted into the nitrile with simultaneous conversion of all the alcohol groups to acetate groups. In the final step sodium methoxide in methanol is added, leading to removal of all the acetate groups and ejection of the nitrile group and collapse of the second carbon from a tetrahedral structure to an aldehyde.

Ruff–Fenton Degradation

In a variation, the Ruff–Fenton degradation (Otto Ruff 1898, H.J.H. Fenton 1893) converts the aldose first to the alpha-hydroxy-carboxylic acid with bromine and calcium hydroxide and then to the shortened aldose by reaction with Iron(III) sulfate and hydrogen peroxide.

KOENIGS–KNORR REACTION

The Koenigs–Knorr reaction in organic chemistry is the substitution reaction of a glycosyl halide with an alcohol to give a glycoside. It is one of the oldest and simplest glycosylation reactions. It is named after Wilhelm Koenigs (1851–1906), student of von Bayer and fellow student with Hermann Emil Fischer, and Edward Knorr, student of Koenigs. In its original form, Koenigs and Knorr treated *acetobromoglucose* with alcohols in the presence of silver carbonate. Shortly afterwards Fischer and Armstrong reported very similar findings. In the above example, the stereochemical outcome is determined by the presence of the neighboring group at C2 that lends anchimeric assistance, resulting in the formation of a 1,2-trans stereochemical arrangement. Esters (*e.g.*, acetyl, benzoyl, pivalyl) generally provide good anchimeric assistance, whereas ethers (*e.g.*, benzyl, methyl *etc.*) do not, leading to mixtures of stereoisomers.

Mechanism

In the first step of the mechanism, the glycosyl bromide reacts with silver carbonate upon elimination of silver bromide and the silver carbonate anion to the oxocarbenium ion. From this structure a dioxolanium is formed,

which is attacked by methanol via an SN_2 mechanism at the carbonyl carbon atom. This attack leads to the inversion. After cleavage of a hydrogen atom, the glycoside is formed. The reaction can also be applied to carbohydrates with other protecting groups. In the oligosaccharide synthesis in place of the methanol other carbohydrates are used, which have been modified with protective groups in such a way that only one hydroxyl group is accessible.

History

The method was later transferred by Emil Fischer and Burckhardt Helferich to other chloro-substituted purines and produced thus for the first time synthetic nucleosides. It was later improved and modified by numerous chemists.

Alternative Reactions

Generally, the Koenigs–Knorr reaction refers to the use of glycosyl chlorides, bromides and more recently iodides as glycosyl donors. The Koenigs–Knorr reaction can be performed with alternative promoters such as various heavy metal salts including mercuric bromide/mercuric oxide, mercuric cyanide and silver triflate. When mercury salts are used, the reaction is normally called the Helferich method. Other glycosidation methods are Fischer glycosidation, use of glycosyl acetates, thioglycosides, glycosyl trichloroacetimidates, glycosyl fluorides or n-pentenyl glycosides as glycosyl donors, or intramolecular aglycon delivery.

CARBOHYDRATE DIGESTION

All carbohydrates absorbed in the small intestine must be hydrolyzed to monosaccharides prior to absorption. Hydrolysis precedes transport of monosaccharides in hamster intestine. From sucrose, glucose is taken up much faster than fructose. Monosaccharide transport saturates with D-glucose at 30 mM. Digestion of starch begins with the action of salivary alpha-amylase/ptyalin, although its activity is slight in comparison with that of pancreatic amylase in the small intestine. Amylase hydrolyzes starch to alpha-dextrin, which are then digested by gluco-amylase (alpha-dextrinases) to maltose and maltotriose. The products of digestion of alpha-amylase and alpha-dextrinase, along with dietary disaccharides are hydrolyzed to their corresponding monosaccharides by enzymes (maltase, isomaltase, sucrase and lactase) present in the brush border of the small intestine. In the typical Western diet, digestion and absorption of carbohydrates is fast and takes place usually in the upper small intestine. However, when the diet contains carbohydrates not easily digestible, digestion and absorption take place mainly in the ileal portion of the intestine.

Digestion of food continues while simplest elements are absorbed. The absorption of most digested food occurs in the small intestine through the brush border of the epithelium covering the villi (small hair-like structure). It is not a simple diffusion of substances, but is active and requires energy use by the epithelial cells.

During the phase of carbohydrate absorption, fructose is transported into the intestinal cell's cytosol, glucose and galactose competes with other [lNal] transporter required for operation. From the cytosol, monosaccharides pass into the capillaries by simple or facilitated diffusion. Carbohydrates not digested in the small intestine, including resistant starch foods such as potato, bean, oat, wheat flour, and several monosaccharide oligosaccharides and starch, are digested in a variable when they reach the large intestine. The bacterial flora metabolize these compounds anaerobically in the absence of oxygen. This produces gases (hydrogen, carbon dioxide and methane) and short-chain fatty acids (acetate, propionate, butyrate). The gases are absorbed and excreted by breathing or through the anus (flatulence). Fatty acids are rapidly metabolized. Butyrate is used mainly by cells in the colon and acetate is absorbed into the blood and taken up by the liver, muscle and other tissue. Propionate is an important precursor of glucose in some animals, but not humans.

TRISACCHARIDES

Trisaccharides are oligosaccharides composed of three monosaccharides with two glycosidic bonds connecting them. Similar to the disaccharides, each glycosidic bond can be formed between any hydroxyl group on the component monosaccharides. Even if all three component sugars are the same (*e.g.*, glucose), different bond combinations (regiochemistry) and stereochemistry (alpha-or beta-) result in trisaccharides that are diastereoisomers with different chemical and physical properties

EXAMPLES

Trisaccharide	Unit 1	Bond	Unit 2	Bond	Unit 3
Isomaltotriose	Glucose	$\alpha(1\rightarrow6)$	Glucose	$\alpha(1\rightarrow6)$	Glucose
Nigerotriose	Glucose	$\alpha(1\rightarrow3)$	Glucose	$\alpha(1\rightarrow3)$	Glucose
Maltotriose	Glucose	$\alpha(1\rightarrow4)$	Glucose	$\alpha(1\rightarrow4)$	Glucose
Melezitose	Glucose	$\alpha(1\rightarrow2)$	Fructose	$\alpha(1\rightarrow3)$	Glucose
Maltotriulose	Glucose	$\alpha(1\rightarrow4)$	Glucose	$\alpha(1\rightarrow4)$	Fructose
Raffinose	Galactose	$\alpha(1\rightarrow6)$	Glucose	$\beta(1\rightarrow2)$	Fructose
Kestose	Glucose	$\alpha(1\leftrightarrow2)$	Fructose	$\beta(1\leftarrow2)$	Fructose

MALTOTRIOSE

Maltotriose is a trisaccharide (three-part sugar) consisting of three glucose molecules linked with α -1,4 glycosidic bonds. It is most commonly produced by the digestive enzyme alpha-amylase (a common enzyme in human saliva) on amylose in starch. The creation of both maltotriose and maltose during this process is due to the random manner in which alpha amylase hydrolyses α -1,4glycosidic bonds. It is the shortest chain oligosaccharide that can be classified as maltodextrin.

MELEZITOSE

Melezitose, also spelled melicitose, is a nonreducing trisaccharidesugar that is produced by many plant sap eating insects, including aphids such as *Cinara pilicornis* by an enzyme reaction. This is beneficial to the insects, as it reduces the stress of osmosis by reducing their own water potential. The melezitose is part of the honeydew which acts as an attractant for ants and also as a food for bees. This is useful to the lice as they have a symbiotic relationship with ants. Melezitose can be partially hydrolyzed to glucose and turanose the latter of which is an isomer of sucrose.

RAFFINOSE

Raffinose is a trisaccharide composed of galactose, glucose, and fructose. It can be found in beans, cabbage, brussels sprouts, broccoli, asparagus, other vegetables, and whole grains. Raffinose can be hydrolyzed to D-galactose and sucrose by the enzyme α -galactosidase (α -GAL), an enzyme not found in the human digestive tract. α -GAL also hydrolyzes other α -galactosides such as stachyose, verbascose, and galactinol, if present. The enzyme does not cleave β -linked galactose, as in lactose.

The raffinose family of oligosaccharides (RFOs) are alpha-galactosyl derivatives of sucrose, and the most common are the trisaccharide raffinose, the tetrasaccharidestachyose, and the pentasaccharide verbascose. RFOs are almost ubiquitous in the plant kingdom, being found in a large variety of seeds from many different families, and they rank second only to sucrose in abundance as soluble carbohydrates.

Humans and other monogastric animals (pigs and poultry) do not possess the α -GAL enzyme to break down RFOs and these oligosaccharides pass undigested through the stomach and upper intestine. In the lower intestine, they are fermented by gas-producing bacteria that do possess the α -GAL enzyme and make carbon dioxide,

methane or hydrogen—leading to the flatulence commonly associated with eating beans and other vegetables. α -GAL is present in digestive aids such as the product Beano. Procedures concerning cryopreservation have used raffinose to provide hypertonicity for cell desiccation prior to freezing. Either raffinose or sucrose is used as a base substance for sucralose.

TETRASACCHARIDES

A tetrasaccharide is a carbohydrate which gives upon hydrolysis four molecules of the same or different monosaccharides. For example, stachyose upon hydrolysis gives one molecule each of glucose and fructose and two molecules of galactose. The general formula of a tetrasaccharide is typically $C_{24}H_{42}O_{21}$.

Structure and occurrence of trisaccharides		
Name	chemical compound	function/occur
Lychnose (1- α -Galactosyl-raffinose)	O- α -D-Galp-(1 $\rightarrow$ 6)-O- α -D-Glup-(1 $\rightarrow$ 2)-O- β -D-Fruf-(1 $\rightarrow$ 1)-O- α -D-Galp	
Maltotetraose	O- α -D-Glcp-(1 $\rightarrow$ 4)-O- α -D-Glcp-(1 $\rightarrow$ 4)-O- α -D-Glcp-(1 $\rightarrow$ 4)-D-Glcp	In starch syrup
Nigerotetraose	O- α -D-Glcp-(1 $\rightarrow$ 3)-O- α -D-Glcp-(1 $\rightarrow$ 3)-O- α -D-Glcp-(1 $\rightarrow$ 3)-D-Glcp	
Nystose (β -D-Fructosyl-1-kestose)	O- α -D-Glcp-(1 $\rightarrow$ 2)- β -D-Fruf-(1 $\rightarrow$ 2)- β -D-Fruf-(1 $\rightarrow$ 2)- β -D-Fruf	
Sesamose	O- α -D-Galp-(1 $\rightarrow$ 6)-O- α -D-Galp-(1 $\rightarrow$ 6)-O- β -D-Fruf-(2 $\rightarrow$ 1)-O- α -D-Glcp	
Stachyose	O- α -D-Galp-(1 $\rightarrow$ 6)-O- α -D-Galp-(1 $\rightarrow$ 6)-O- α -D-Glcp-(1 $\rightarrow$ 2)- β -D-Fruf	Widespread in plants (artichoke, soybean)

STACHYOSE

Stachyose is a tetrasaccharide consisting of two α -D-galactose units, one α -D-glucose unit, and one β -D-fructose unit sequentially linked as gal(α 1 $\rightarrow$ 6)gal(α 1 $\rightarrow$ 6)glc(α 1 $\leftrightarrow$ 2 β)fru. Together with related oligosaccharides such as raffinose, Stachyose occurs naturally in numerous vegetables (*e.g.*, green beans, soybeans and other beans) and other plants. Stachyose is less sweet than sucrose, at about 28% on a weight basis. It is mainly used as a bulk sweetener or for its functional oligosaccharide properties. Stachyose is not completely digestible by humans and delivers 1.5 to 2.4 kcal/g (6 to 10 kJ/g).

Bibliography

- Allen, M. P., and Tildesley, D. J.: *Computer Simulation of Liquids*. Oxford University Press, New York; 2001.
- Berkert, U., and Allinger, N. L.: *Molecular Mechanics*. American Chemical Society, Washington, D.C; 200.
- Bohne, A., Lang, E., and von der Lieth, C.-W. Mol. Model.; 2001.
- Bohne, A., Lang, E., and von der Lieth, C.-W. Bioinformatics. 2001.
- Bonneau, R., and Baker, D. Annu. Rev. Biophys. Biomol. Struct; 2002.
- Brañde, C.-I. Q. Rev. Biophys. 2002.
- Brooks C. I., III, Karplus, M., and Pettitt, B. M.: *Proteins: A Theoretical Prospective of Dynamics, Structure and Thermodynamics*. John Wiley & Sons, New York; 2005.
- Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M.: J. Comput. Chem.; 2003.
- Bryant S. H., and Altschul S. F.: Curr. Opin. Struct. Biol.; 2006.
- Bystroff, C., and Kraut, J.: *Biochemistry*; 2005.
- Chang, G., Guida, W. C., and Still, W. C.: J. Am. Chem. Soc.; 2005.
- China, G., Padron, G., Hooft, R. W. W., Sander, C., and Vriend, G. (1995) Proteins; 2005.
- Chothia, C., and Lesk, A. M. EMBO J; 2001.
- Clark, M., Cramer R. D., III, and van Opensch, N. J. Comput. Chem.; 2001.
- Crichton, T. E. J. Mol. Biol.; 2005.
- Dandekar, T., and Argos, P. J. Mol. Biol.; 2003.
- Eisen, M. B., Wiley, D. C., Karplus, M., and Hubbard, R. E. Proteins; 2005.
- Ewing, T., and Kuntz, I. D. J. Comput. Chem.; 2007.
- Feng, Z. K., and Sippl, M. J. Fold Des; 2004.
- Gibrat, J.-F., Madej, T., and Bryant, S. H. Curr. Opin. Struct. Biol; 2011.
- Gogonea, V., Urez, D. S., van der Vaart, A., and Merz, K. M. Jr. Curr. Opin. Struct. Biol.; 2001.

- Hehre, J. W., Radom, L., Schleyer, P., and Pople, J. *Ab Initio Molecular Orbital Theory*. John Wiley & Sons, New York; 2001.
- Herman, J., Berendsen, H. J. C., van Gunsteren, W., and Postma, J. P. M. *Biopolymers*; 2001.
- Hilbert, M., Bohm, G., and Jaenicke, R. *Proteins*; 2000.
- Hoijtje, H.-D., and Folkeis, G. (1997) *Molecular Modeling: Basic Principles and Applications*. VCH, New York. Howard, A. E., and Kollman, P. A. (1988) *J. Med. Chem.*; 2001.
- Holm, L., and Sander, C. *J. Mol. Biol.*; 20003.
- Jalaie, M., and Lipkowitz, K. B. *Rev. Comput. Chem.*; 2000.
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. *J. Mol. Biol.*, 2005.
- Jorgensen, W. L., and Tirado-Rives, J. *J. Amer. Chem. Soc.*; 2001.
- Jorgensen, W. L., Chandrasekhas, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). *J. Chem. Phys.*; 2005.
- Koshland, D. E. Jr. (1958) *Proc. Natl. Acad. Sci. U.S.A.*; 2001.
- Koza, J.: *Genetic Programming*. MIT Press, Cambridge; 2001.
- Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) *J. Mol. Biol.* 161:269—288. Lest, A. M., and Chothia, C. (1984) *J. Mol. Biol.*; 2001.
- Leach, A. R.: *Molecular Modeling Principles and Applications*. Longman, Essex; 2001.
- Levine, M., Stuart, D., and Williams J.: *Acta Crystallogr*; 2004.
- Levitt, M., and Chothia, C. *Nature*; 2003.
- Lu, G. G.: *J. Appl. Crystallogr*; 2000.
- Mathew, D. H., Sabina, J., Zuker, M., and Turner, D. H. *J. Mol. Biol.*; 2000.
- McCammon, J. A., and Harvey, S. C.: *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, New York; 2004.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) *J. Chem. Phys.* 2002.
- Morris, G. M., Goodsell, D. S., Huey, R., and Olson, A. J. *J. Comput. Aided Mol. Des*; 2001.
- Needleman, S. B., and Wunsch, C. D. *J. Mol. Biol.*; 2005.
- Nemathy, G., Porttle, M. S., and Scheraga, H. J. *Phys. Chem.*; 2005.
- Osawa, E., and Lipkowitz, K. B. *Rev. Comput. Chem.*; 2005.
- Perutz, M. F.: *Mechanisms of Cooperativity and Allosteric Regulation in Proteins*. Cambridge University Press, Cambridge; 2009.
- Phillips, D. C. *Proc. Natl Acad. Sci., U.S.A.*; 2005.
- Rarey, M., Kramer, B., Lenguer, T., and Klebe, G. *J. Mol. Biol.*; 2006.
- Richardson, D. C., and Richardson, J. S. *Protein Sci.*; 2002.
- Richardson, D. C., and Richardson, J. S. *Trends Biochem. Sci.*; 2004.
- Schlecht, M. F.: *Molecular Modeling on the PC*. Wiley-VCH, New York; 2004.
- Scrutton, N. S., Berry, A., and Perham, R. N. *Nature*; 2001.
- Shindyalov, I. N., and Bourne, P. E.: *Protein Eng.*; 2002.
- Shindyalov, I. N., and Bourne, P. E.: *Nucleic Acids Res*; 2001.
- Stewart, D. E., Weiner, P. K., and Wampller, J. E. *J. Mol. Graph*; 2002.
- Szustakowski, J. D., and Weng, Z. P. *Proteins*; 2000.

- Tsai, C. S. *Int. J. Biochem. Cell Biol.*; 2005.
- Tulinsky, A., Park, C. H., and Skzypezak-Jankun, E. J. *Mol. Biol.*; 2002.
- Warshel, A: *Computer Modeling of Chemical Reactions in Enzymes and Solutions*. John Wiley & Sons, New York; 2002.
- Weiner, S. J., Kollman, P. A., Nguyen, D. T., and Case, D. A. *J. Comput. Chem.*; 2002.
- Weiner, S. J., Kollmann, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr., and Weiner, P. J. *Am. Chem. Soc.*; 2004.
- Wiberg, K. B. *J. Am. Chem. Soc.*; 2002.
- Wilks, H. M., Halsall, D. J., Atkinson, T., Chia, W. N., Clarke, A. R., and Holbrook, J. J.: *Biochemistry*; 2002.
- Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R., and Holbrook, J. J. *Science* ; 2001.
- Xavier, D., Mark, A. E. and van Gunsteren, W. F. *J. Comput. Chem.*; 2008.

Index

A

Abbreviations 83, 85, 131, 132, 266
Abundant 42, 47, 48, 54, 60, 80, 107, 113,
249, 263, 264
Accessible 178, 180, 210, 215, 248, 289, 295
Acetalisation 291, 292
Adaptation 72, 74, 238, 239, 257
Adaptive 71, 239, 249
Aerobic 70, 238, 239, 244, 246, 248, 251,
252, 253, 254, 255,
260, 261, 291
Amadori 292, 293
Anaerobic 71, 238, 239, 247, 251, 252, 253,
254, 257, 258, 259,
260, 266, 285, 290
Aspartame 108

B

Biodegradable 80, 108, 109
Biomechanics 221, 237, 238, 254
Biosynthesis 79, 90, 92, 104, 107, 130, 133,
149, 256, 257, 259,
263, 290, 297
Bruyn 292

C

Catabolism 88, 90, 103, 130, 243, 291
Catalytic 103, 105, 125, 129, 137, 144, 227,
231
Cathode 168

Cellobiose 268, 269, 277
Colligative 16, 21, 22, 24
Cryogenic 158, 163, 181
Crystallography 7, 46, 48, 51, 53, 80, 82, 102,
105, 133, 136, 143,
145, 146, 148, 153,
154, 156, 157, 158,
159, 160, 161, 162,
163, 164, 165, 166,
167, 168, 169, 170,
177, 178, 179, 180,
181, 183, 251, 253
Cyanohydrin 292
Cytosol 72, 246, 247, 251, 295

D

Diagnosis 49, 53, 55, 62, 144
Disaccharide 263, 265, 268, 269, 270, 273,
276, 277, 278, 285,
286, 287, 289, 290,
295, 296

E

Ekenstein 292
Electrophoresis 82
Elimination 85, 101, 137, 233, 234, 235, 236,
243, 268, 294
Endogenous 96, 97, 99, 284
Ethanol 9, 15, 16, 117, 120, 125, 264, 274, 282
Etymology 184
Eukaryotic 96, 137, 143, 147, 242, 246, 248

Expenditure 190, 215, 255

F

Fermentation 173, 244, 247, 251, 273, 282

G

Gene 61, 71, 72, 74, 78, 86, 92, 98, 101, 102, 147, 242

Genomic 241, 242

Genotype 72

H

Hasselbalch 26, 27

Henderson 26, 27, 33, 179

Heterogeneity 174

Homogenization 15

Homology 138, 141, 158, 249

Hydrophobic 13, 17, 18, 19, 20, 21, 73, 81, 102, 103, 130, 131, 133, 135, 136, 142, 146, 149, 150, 153

I

Inflammation 55, 61

Integration 50, 190, 234

Ionization 12, 22, 23, 25, 160, 245

Isoelectric 26, 81, 82, 138

Isolation 82, 163, 274

Isomaltose 269, 276, 277

Isomaltulose 270, 271, 272, 273

Isomerism 80, 171

K

Kinesiology 237, 238, 239, 240, 241

Kinetics 19, 142, 143, 151, 154, 156, 221, 222, 223, 224, 225, 226, 227, 228, 230, 231, 233, 234, 241, 242

L

Lactation 274

Lactose 263, 265, 268, 273, 274, 275, 276, 284, 289, 296

Lactulose 273, 274, 275, 276

Legislation 271

M

Macrophages 70

Maltose 268, 269, 270, 276, 277, 286, 295

Maltotriose 295, 296

Manipulation 144, 233

Melezitose 296

Metabolism 28, 30, 59, 70, 76, 82, 103, 107, 212, 227, 232, 233, 235, 236, 243, 244, 248, 251, 252, 253, 254, 256, 258, 268, 270, 271, 272, 274, 284, 290, 291

Microscopy 49, 61, 156, 158, 179, 180

Microtubules 144

Monosaccharide 263, 264, 265, 266, 267, 268, 269, 275, 277, 279, 291, 295

Morphological 54, 60

Motif 72, 147, 276

N

Nomenclature 81, 83, 91, 145, 249, 263, 265, 277

O

Occurrence 7, 82, 87, 195, 286

Optical 61, 80, 111, 154, 155

Optimal 34, 35, 80, 240, 241

P

Paramagnetic 43, 47, 56, 73, 74, 75, 180, 181, 217

Pharmacokinetics 232, 233, 234, 236, 237

Phosphagen 251, 253, 254

Phosphate 29, 35, 69, 70, 77, 136, 241, 244, 245, 246, 247, 248, 249, 250, 251, 254, 279, 283

Physicochemical 99, 108, 130

Plyometrics 256

Polymerization 19, 20, 88, 129, 250, 264

Polypeptide 71, 72, 73, 95, 105, 130, 135, 136, 137, 138, 142, 143, 144, 147, 148, 149, 150, 151, 152, 156, 227, 271

Polysaccharides 138, 263, 264, 265, 268, 277, 285, 289, 290, 291

Potentials 183, 185, 197, 213, 229, 240

Proteinogenic 79, 81, 82, 83, 90, 91, 92, 95, 96, 103, 104, 105

Protocol 110, 178, 259, 260, 287

Q

Quaternary 6, 72, 105, 116, 142, 144, 145, 149, 153

R

Radiation 14, 41, 42, 43, 45, 47, 53, 62, 159, 160, 161, 163, 164, 167, 168, 169, 179, 181, 193

Radioactive 42

Raffinose 296, 297

Recombination 19, 97

Replication 6, 7, 102, 221

Resonance 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 59, 60, 62, 75, 137, 146, 154, 156, 162, 163, 172, 173, 174, 175, 177, 178, 180, 293

Respiration 29, 70, 73, 77, 227, 243, 244, 246, 247, 249, 253, 291

Respiratory 29, 30, 32, 70, 75

S

Simulation 153, 179, 222, 299

Sophisticated 37, 257

Spectroscopy 3, 7, 41, 44, 45, 46, 47, 48, 49, 50, 52, 53, 59, 74, 121, 133, 141, 153, 154, 155, 156, 163, 172, 174, 176, 178, 180, 181

Stachyose 296, 297

Sucrose 15, 263, 265, 268, 270, 271, 272, 273,

276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 288, 289, 295, 296, 297

T

Tertiary 2, 6, 21, 117, 120, 138, 139, 141, 142, 143, 144, 149, 150, 152, 153, 156, 159, 168, 296

Tetrasaccharides 297

Therapeutic 104, 237, 239

Therefore 25

Thermodynamics 18, 150, 151, 183, 184, 185, 186, 187, 188, 189, 190, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 209, 211, 212, 213, 214, 215, 216, 217, 218, 219, 224, 226, 234

Transformation 97, 159, 188, 209, 211, 224, 232, 292

Trehalose 6, 268, 287, 297

Trisaccharides 295, 296

Turanose 289, 296

V

Vectors 176, 222, 223

Z

Zwitterions 81, 82