Biophysics

Concepts and Applications

Nicole Hobbs

Biophysics: Concepts and Applications

Biophysics: Concepts and Applications

Nicole Hobbs

Published by The English Press,
5 Penn Plaza,
19th Floor,
New York, NY 10001, USA

ISBN: 978-1-9789-6100-5

Cataloging-in-Publication Data

Biophysics : concepts and applications / Nicole Hobbs.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-9789-6100-5
1. Biophysics. 2. Biology. 3. Medical sciences. 4. Physics. I. Hobbs, Nicole.
QH505 .B56 2021
571.4--dc23

Table of Contents

Preface

This book aims to help a broader range of students by exploring a wide variety of significant topics related to this discipline. It will help students in achieving a higher level of understanding of the subject and excel in their respective fields. This book would not have been possible without the unwavered support of my senior professors who took out the time to provide me feedback and help me with the process. I would also like to thank my family for their patience and support.

The interdisciplinary field of science that uses methods and approaches of physics to the study of biological phenomena is known as biophysics. It is concerned with the study of physical quantities such as stress, electric current, entropy and temperature in biological systems. The field uses a variety of techniques such as fluorescent imaging techniques, x-ray crystallography, atomic force microscopy, NMR spectroscopy, small angle scattering and electron microscopy. Biophysics is closely linked to other areas such as molecular biology, physiology, computational biology, developmental biology, nanotechnology, physical chemistry and systems biology. The domain finds its applications in the areas of healthcare, radiology, nanomedicine and microscopy. The book studies, analyzes and upholds the pillars of this discipline and its utmost significance in modern times. Biophysics is an upcoming field of science that has undergone rapid development over the past few decades. Coherent flow of topics, student-friendly language and extensive use of examples make this book an invaluable source of knowledge.

A brief overview of the book contents is provided below:

Chapter – Introduction

Biophysics is an interdisciplinary field of science that studies biological phenomena through the use of traditional methods and approaches of physics. A few of its types include membrane biophysics and computational biophysics. This chapter sheds light on the subject of biophysics for an easy understanding of the subject.

Chapter – Biophysical Structure and Phenomena

Biophysical structure and phenomena include the concepts of enzyme kinetics, cell division, cell migration, cell signaling, transfer RNA, myoglobin, pancreatic ribonuclease, phospholipids, etc. The topics elaborated in this chapter will help in gaining a better perspective of biophysical structure and phenomena.

Chapter – Molecular Biophysics

A biomolecule refers to the ions and molecules present in organisms essential for biological processes. Molecular biophysics deals with the study of biomolecular systems and functions such as molecular structure and dynamic behavior structure of biomolecules. This chapter discusses the subject of molecular biophysics in detail.

Chapter – Biophysical Techniques

There are various biophysical techniques that are studied under this domain. A few of them are fluorescence imaging technique, X-ray crystallography, atomic force microscopy, isothermal titration calorimetry, NMR spectroscopy, etc. This chapter has been carefully written to provide an extensive understanding of these biophysical techniques.

Chapter – Understanding Biomechanics

Biomechanics is the branch of biophysics that studies the structure, function and motion of the mechanical aspects of biological systems of living organisms and their organs and cells. Ground reaction force, Hill's muscle model, lines of non-extension, biomechatronics, force platform, etc. are some of the aspects that are studied under its domain. All the aspects related to biomechanics have been carefully analyzed in this chapter.

Nicole Hobbs

1
Introduction

Biophysics is an interdisciplinary field of science that studies biological phenomena through the use of traditional methods and approaches of physics. A few of its types include membrane biophysics and computational biophysics. This chapter sheds light on the subject of biophysics for an easy understanding of the subject.

Biophysics

Biophysics is a discipline concerned with the application of the principles and methods of physics and the other physical sciences to the solution of biological problems. The relatively recent emergence of biophysics as a scientific discipline may be attributed, in particular, to the spectacular success of biophysical tools in unravelling the molecular structure of deoxyribonucleic acid (DNA), the fundamental hereditary material, and in establishing the precisely detailed structure of proteins such as hemoglobin in order that the position of each atom may be known. Biophysics and the intimately related subject molecular biology now are firmly established as cornerstones of modern biology.

The biophysical approach is unified by a consideration of biological problems in the light of physical concepts, so that biophysics is, perforce, interdisciplinary. Biophysics may be thought of as the central circle in a two-dimensional array of overlapping circles, which include physics, chemistry, physiology, and general biology. Relations with chemistry are mediated through biochemistry and chemistry; those with physiology, through neurophysiology and sensory physiology. Biology, which may be viewed as a general subject pervading biophysical study, is evolving from a purely descriptive science into a discipline increasingly devoted to understanding the nature of the prime movers of biological events. The evolution of biology in these directions has received great impetus from the biophysical and biochemical discoveries of the 20th century. An understanding of the physical principles governing biological effects is the proper end of biophysics.

Areas of Study

The content and methods of biophysics are illustrated by examining several notable contributions to science.

Protein Structure

Within two days after the initial publication of Wilhelm Röntgen's discovery of X rays in 1895, a surgeon in Scotland used X rays to observe a needle as he extracted it from the palm of an unfortunate seamstress. Although this medical application resulted in the development of radiological diagnosis and treatment of disease by radiation, physical aspects of Röntgen's discovery also provided the means for elucidating the structure of proteins and other large molecules. The laws governing the diffraction of X rays were discovered by the two Braggs, Sir William and Sir Lawrence, who were father and son. At the Cavendish Laboratory at the University of Cambridge, where Sir Lawrence was professor, J.D. Bernal was studying the use of X-ray diffraction for the determination of the structure of large biological molecules. He had already used X rays to define the size and shape of the tobacco mosaic virus and showed it to have a regular internal structure. At the Cavendish Laboratory the group that formed around Bernal, a man of wide public and scientific interests, included the Nobel Prize winners Max Perutz and John Kendrew, who in 1937 began to use X rays to analyze two proteins fundamental to life, myoglobin and hemoglobin, both of which function in the transport of gases in the blood. Twenty-two years passed before the structures of these proteins were established; the significance of the work is that it provided the basis for an understanding of the mechanism of the action of enzymes and other proteins, an active and fruitful subject of modern investigation.

Deoxyribonucleic Acid

Interest in biophysics at the Cavendish Laboratory resulted in another important discovery, the structure of deoxyribonucleic acid (DNA), the genetic material. This achievement by a British biophysicist, Francis H.C. Crick, and by a U.S. biochemist, James Watson, was based on X-ray data obtained by Maurice Wilkins at King's College, London. When Crick first went to the Cavendish Laboratory for education in biophysics, he worked under Perutz's direction; when Watson went to the Cavendish, he and Crick began the collaboration that led to the establishment of the structure of DNA, for which Watson, Crick, and Wilkins later were awarded a Nobel Prize.

Much impetus for biophysical investigation following World War II came from the desire of physicists to move away from physics and into biology; this drive was strengthened by the publication in 1944 of Erwin Schrödinger's book What Is Life? Schrödinger, the Austrian physicist who contributed substantially to the development of wave mechanics, was anxious to determine whether biological events could be accounted for in terms of known laws of physics and chemistry, or whether a full explanation would require the

formulation of physical laws not yet known to exist. Because biological reproduction seemed to pose intractable problems, he devoted a chapter of his book to a consideration of the gene. The discussion was based on the model put forward by Max Delbrück, a physicist who had for some years been studying the genetics of viruses that infect bacteria (bacteriophages). Delbrück's summer course on bacteriophages in 1945 at Cold Spring Harbor in New York set in motion the chain of events that led to understanding the genetic code by which the sequence of the nucleotides in DNA is translated into the sequence of amino acids in a protein. The use of bacteriophage also provided an opportunity for experiments with a primitive living organism that could be studied without anatomic complexities. This aspect of biophysics has become more biochemically oriented as it has developed and is now known as molecular biology; sometimes it is considered a distinct discipline, and other times it is subsumed under the biophysical sciences.

Nerve Impulse

Important aspects of biophysics have been derived from physiology, especially in studies involving the conduction of nerve impulses. One important scientific product of World War II—the development of vastly improved electronics—largely resulted from radar devices that had been used primarily for locating aircraft. Another product, the atomic bomb, was constructed by way of nuclear reactors that could, in peace time, provide an abundant supply of radioactive isotopes, which are now of great value not only in biophysical research but also in biochemistry and medicine. These two disparate advances were important to the work of two Nobel Prize winners, Alan Hodgkin and Andrew Huxley, who showed how the flow of sodium and potassium across the membranes of nerves can be coupled to produce the action potential, a brief electrical event that initiates the action potential, which propagates the nervous signal.

A model of the nerve axon proposed by Hodgkin and Huxley grew from a 19th-century confluence of ideas. Julius Bernstein, an experimental neurophysiologist, used physical chemical theories to develop a membrane theory of nervous conduction; Hodgkin's initial experiments were designed to test specific predictions of the Bernstein hypothesis. Early in 1938 Hodgkin learned of the important results of a newly developed technique that allowed examination of the time course of nervous conduction. After World War II, Hodgkin, joined by Huxley, again took up the research. They presented their explanation of the mechanism of nervous conduction in five scientific papers between October 1951 and March 1952.

Biological Membranes

The availability of radioactive isotopes provided the technology necessary for understanding how molecules are transported across biological membranes, which are the very thin boundaries of living cells; the environment maintained by membranes in cells differs from the external environment and permits cellular function. The Danish physiologist August Krogh laid the groundwork in this subject; his pupil, Hans Ussing, developed the conceptual means by which the transport of ions (charged atoms) across

membranes can be identified. Ussing's definition of active transport made possible an understanding, at the cellular level, of the way in which ions and water are pumped into and out of living cells in order to regulate the ionic composition and water balance in cells, organs, and organisms. The molecular mechanism by which these processes occur, however, remains to be discovered.

In addition to the function of transport, membranes also are utilized as templates on which such molecules as enzymes, which must function in a sequential fashion, can be kept in the requisite order. Although great progress has been made in understanding the mechanisms by which specific atoms are assembled into large biological molecules, the principles involved in the assembly of molecules into membranes, which are organized structures of a higher degree of complexity than large molecules, are not yet very well understood. There is reason to believe that the incorporation of a molecule into a membrane endows it with properties that differ from those of a molecule in solution. A primary task of biophysics is to understand the physical character of these cooperative interactions that are essential to life.

Muscle Contraction

A.V. Hill developed exquisitely sensitive temperature sensors for measuring heat generated during muscular contraction; he initiated studies relating this heat to the thermodynamic parameters responsible for it. The electron microscope in the years following World War II made possible the description of muscular contraction at a structural level, though the mechanisms involved in the flow of heat during the process are not yet known. Simultaneously, in the 1960s, but independently, various physicists postulated the sliding-filament theory of muscular contraction, according to which muscles contract by the sliding of one filament along another and not by a springlike coiling. Remarkable advances, based on the use of techniques such as X-ray diffraction and electron microscopy, have made it possible to visualize many of the molecules involved in the process. The entire process of muscular contraction, in terms of an identification of the molecules and a description of the chemical reactions in the muscle fibre, has been almost completely explained.

Sensory Communication

The above comprise a few specific examples of the scope of biophysics. One area, difficult to discuss in specific terms, is that of sensory communication. Because stimuli, particularly those of a visible or auditory nature, can easily be specified in exact physical terms, they have excited the interest of physical scientists since before 1850. Modern electronic techniques make it relatively easy to distinguish true signals from noise; in addition, computers make possible the performance of significant experiments concerning the complex relationship between stimulus and action. Quantitative analysis of sensory response is very difficult, however, because it involves a synthesis of the action of many cells. It has been pointed out that: An adequate theory of sensory function implies an adequate theory of brain function. And an adequate theory of brain function in

its turn requires that the nervous system's behavioral repertory be predictably related to the behaviour of the elements that compose it.

Medical Physics

Medical Physics is a branch of Applied Physics, pursued by medical physicists, that uses physics principles, methods and techniques in practice and research for the prevention, diagnosis and treatment of human diseases with a specific goal of improving human health and well-being. Medical physics may further be classified into a number of sub-fields (specialties), including Radiation Oncology Physics, Medical Imaging Physics, Nuclear Medicine Physics, Medical Health Physics (Radiation Protection in Medicine), Non-ionizing Medical Radiation Physics, and Physiological Measurement. It is also closely linked to neighbouring sciences such as Biophysics, Biological Physics, and Health Physics.

Medical Physicist

Medical physicists are professionals with education and specialist training in the concepts and techniques of applying physics in medicine. Medical Physicists work in clinical, academic or research institutions.

Medical physicists working in clinical environment are health professionals, with education and specialist training in the concepts and techniques of applying physics in medicine, competent to practice independently in one or more of the subfields (specialties) of medical physics.

Role and Responsibilities

Medical physicists are involved with the application of medical physics principles and techniques for prevention, treatment and diagnosis of human disorders, illnesses and disabilities, and the protection of the patients, staff and members of the public from ionizing and non-ionizing radiation hazards. The role and responsibilities of medical physicists may be oriented toward clinical service (which includes technical and radiation safety aspects), management, education, and research and development.

Neurophysics

In recent years, physics has played a major role in the rapid advance of cognitive neuroscience. The most widely used methods of mapping brain function, using magnetic resonance imaging (fMRI), were developed in the early 1990s by USA-based physicists

working at labs in Boston, Minneapolis, Murray Hill and Washington. Professor Turner was himself one of these pioneers. Methods for interpreting measurements of magnetic fields generated by brain activity (MEG) have also seen major recent improvement in the hands of physicists. The physics of electromagnetic fields, nuclear spins, and the biophysics of blood and brain tissue have thus been combined to provide noninvasive methods that can resolve brain activity within a millimeter in space, or milliseconds in time. These methods allow the exploration of profound neuropsychological questions regarding the organization of human brain function, and even the relationship between mind and brain. With a much improved understanding of normal brain function, it is hoped that far more empirically based treatments for a wide range of neurological and psychiatric conditions can be developed.

Brain Mapping-functional and Anatomical

The most urgent challenge is to answer the question: What next can MRI methods teach us regarding the structure and function of the human brain? Because of MRI's remarkable safety record, MRI techniques can be easily explored with volunteer human subjects. High field strength allows the use of very high spatial resolution, so that details of structures less than half a millimeter in size can be distinguished. The organization of the brain's white matter, consisting of the major long nerve fibres connecting different brain regions, can also be investigated with much greater accuracy. Changes in MRI image intensity caused by local brain activity can be measured with greater sensitivity and spatial precision.

Thus the potential exists for a richer and more detailed scientific understanding of the association of brain function with particular brain structures. Until now, imaging neuroscience has relied on the quite loose association between location in space, cortical folding, and histological analyses of cell structure in cadaver brain, such as that of Brodmann, to find approximate answers to the important question: what type of neuronal territory has what type of function? But now we are discovering that many areas of grey matter have a distinctive appearance in MRI scanning, allowing us to characterize with great precision the neural substrate for particular brain processes. The goal of mapping many of these areas, to form a natural parcellation of grey matter for each individual, is perhaps within sight. Furthermore, the changes in brain structure, such as the thickness of grey matter, that are known to occur with repeated experience or practice, can be explored much more easily than at lower magnetic field strengths.

Technical Requirements

To make the best use of these powerful capabilities requires novel MRI techniques, novel hardware, and very careful attention to the safety of human subjects. At normal magnetic field strengths, MRI industry maintains routine and highly effective safety standards. Exposure to high steady magnetic fields has shown no harmful effects, on the basis of test results with hundreds of human subjects in many laboratories across

the world, and indeed there is no physical reason to anticipate such effects. But there is a safety issue at high magnetic field, for which the MRI industry has yet to develop standard procedures. To produce magnetic resonance images requires excitation of the magnetization of the spins of protons in water molecules in tissue. This is performed by delivering pulses of radio-frequency electromagnetic energy, which can in principle heat the tissue. The careful safeguards set in place by MRI scanner manufacturers must be supplemented at high field strength by electromagnetic field calculations providing detailed limits to safe scanner operation. Such calculations can also guide strategies for the most efficient scanning techniques, and the Department thus includes experts in radio-frequency modelling and hardware, to explore this possibility.

Membrane Biophysics

Biological membranes are essential components of cells and their organelles. They mainly consist of lipids and proteins. The basic structure is a double layer of lipids of about 5 nm thickness into which proteins are embedded. Membranes determine the boundary between the inside and outside of cellular compartments, and their respective ion and protein concentrations. They also control the transport of substances into the cell and are important players in the metabolism of cells. Furthermore, due to the presence of enzymatic proteins, the large overall surface of membranes gives rise to a plethora of catalytic properties.

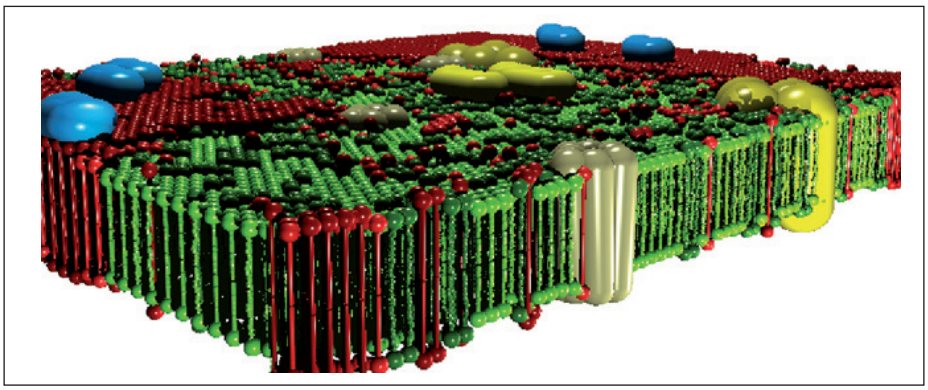
Studies on membranes are clearly a topic of biology and biophysics. However, membranes are also ensembles of many molecules and as such have features well-known in other areas, e.g., the fields of liquid crystals, polymer science, nanoscience and solid-state physics. Similar physical, physical chemistry and thermodynamic concepts to other disciplines concerned with soft matter are used, including phase behavior, diffusion, adhesion and adsorption. The field of membrane biophysics is a very important one for Soft Matter and we hereby highlight some of the more important and recent developments in this themed issue.

The basis for the understanding of biomembranes was gradually laid during the last 120 years. Overton1 proposed that the layer-separating cell organelles have the properties of a fat, Gorter and Grendel2 proposed the double-layer structure, Danielli and Davson3 suggested that the double-layer lipid membrane is covered by surface proteins, while Singer and Nicolson4 considered the membrane as a two-dimensional liquid with proteins being adsorbed to the membrane surface or spanning through the membrane interior. With time it became clear that membranes are not only homogeneous thin sheets separating two aqueous volumes, but that they possess an interesting structure with locally varying physical properties. Already in their important paper from 1972, Singer and Nicolson considered protein clusters, and Mouritsen and Bloom5 proposed that capillary forces can exist between lipids and proteins leading to domains and

aggregates. We now know that membranes are laterally inhomogeneous and display phase behavior, lipid domains, protein clusters, and compositional asymmetry between both lipid monolayers. As is true for interfaces in general, membranes have in some sense a physics of their own that can be considered separate from the 3D environment.

In an attempt to improve understanding of the function of biomembranes, much of biomembrane research in the recent decade has focused on single molecules, for example, the structure and function of individual membrane proteins, receptors or channel proteins. Seemingly, knowing more molecular detail also yields more insight into biological function. This is possibly true for molecules that act locally, e.g., enzymes. However, knowledge obtained exclusively on the molecular scale obscures important features that emerge on larger scales, for instance, elastic constants, relaxation time scales, phase behavior, domain formation, and other kinds of cooperative behavior.

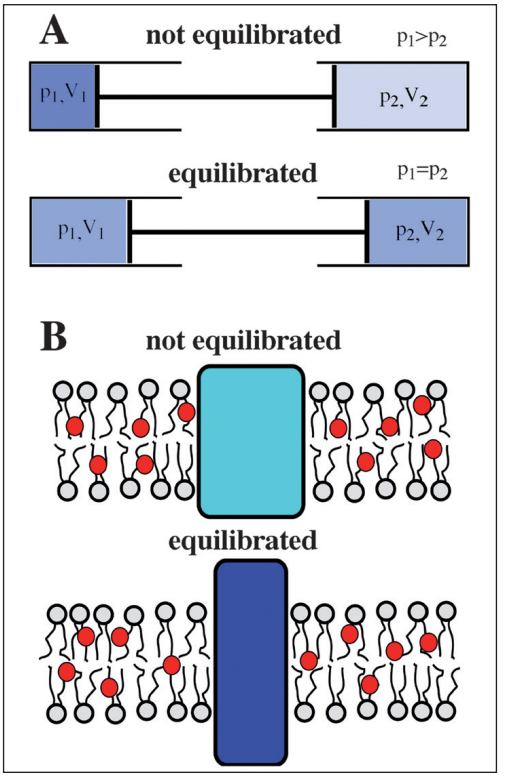
The physical theory describing such systems is thermodynamics. The second law of thermodynamics implies that the most likely state of the system occurs most often. As Einstein noted in 19106 this statement would be trivial if one did not consider the states in the proximity of the entropy maximum, i.e., the fluctuations of the system and their amplitude.A problem with focussing on molecular scales is that one may not consider the thermodynamics of the system correctly.



The present understanding of biomembrane structure is summarized in this image. The membrane consists mainly of lipids and proteins. Lipids arrange in domains and clusters and display different physical states, e.g., solid-ordered (red), liquid-ordered and liquid-disordered (green). Proteins (yellow, blue and white) are distributed differently in the various domains.

The second law is only valid for the complete system under investigation, e.g., the cell membrane as a whole, but not generally for arbitrarily selected subsystems within this membrane, e.g., a protein or an individual lipid, if one looks at them in the context of the cellular environment. One can easily construct examples that demonstrate that. Figure shows two gas containers with different pressures that are coupled by a piston. This system is not at equilibrium. At equilibrium, the pressures in both containers are

the same. In order to approach the equilibrated state, the entropy of one container increases while that of the other container decreases. This implies that in this situation the second law does not hold for the individual containers but only for the total, coupled system. One can construct similar examples from membranes interacting with proteins and drugs. The influence of the lateral pressure created by anaesthetics dissolved in membranes on protein structure and function. Even though proteins are influenced by their structure, it is still the entropy maximum of the complete membrane that determines the protein state in the presence and absence of anaesthetics. A further examplev by Weikl et al. that shows that ligand– receptor equilibria are influenced by thermal curvature fluctuations of the membrane. This effect introduces cooperativity into membrane adhesion to surfaces. Obviously, a system is not only more than, but also different from, the sum of its parts. Thus, it seems wise under many conditions not to consider singlemolecule properties, but rather system features. In thermodynamics, all the couplings between different system variables are expressed in Maxwell's relations, and there are as many of them as there are possible molecular interactions.



A: Two gas containers with different pressures are connected by a piston. When the systems equilibrates, one gas container increases its entropy while the entropy of the gas in the second container decreases. The equilibrated system represents the state of maximum entropy for the total system only. They exert a lateral pressure on the protein that may influence the protein conformation.

The susceptibilities are also functions that cannot be understood on the level of individual molecules. Some of them are the heat capacity (fluctuations in enthalpy), the volume and area compressibility (fluctuations in volume or area), bending elasticity (fluctuations in curvature), and capacitance (fluctuations in charge).

The review by Winter and Jeworrek focuses on the influence of hydrostatic pressure on the phase behavior of membranes, while the review of Sparr et al. treads gradients in chemical potentials across membranes and their influence on phase behavior, and the review of Dimova et al. considers the electrostatic potential and its influence on vesicles. The papers by Cantor et al. and Xing et al. focus on the lateral pressure profile in membranes.

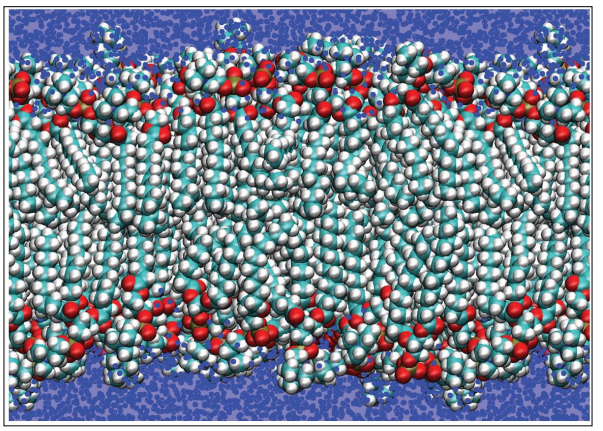
Finally, lipid vesicles are also used in drug delivery, as, for example, shown in the review of Koynova and Tenchov on the transfection properties of cationic lipids. Bunge et al. study bilayers on solid supports.

A number of excellent experts in the field of membrane biophysics contributed to this special issue of Soft Matter. We are grateful to all the authors that made this volume possible. We believe that it gives a fair insight into the present understanding of membrane biophysics and the thermodynamic couplings from a physical and physical chemistry point of view.

Consistent with the fact that about a third of the dry weight of a cell is membrane, almost half of all proteins encoded by a eukaryotic genome are membrane proteins. Thus roughly half of biological processes occur on membranes, and each of these processes will have aspects of its function that fall into the realm of physics. Just as water is the solvent for soluble proteins, the phospholipid bilayer membrane is the solvent for membrane proteins and forms the basis of the biological membrane. A semi-crystalline array that is ordered in some aspects and disordered in other aspects, a membrane has both a fluid and a solid character. It is only two molecules thick but can have an area of a millimeter squared (e.g. eggs) or a length of many meters (e.g. axons in giraffes). The phospholipid bilayer itself is stable for a range of different lipid compositions, and bilayers self-assemble upon sufficient hydration of these lipids. Membrane properties and self-interactions can therefore be extensively studied without proteins in vitro. Our understanding of the physical nature of the membrane backbone comes mostly from studies of the spectroscopic, microscopic, and electrophysiological properties of phospholipid bilayers in the absence of proteins.

Historically, physicists were attracted to aspects of biology that were intrinsically physical, such as biological electricity (e.g. the historic controversy between Galvani and Volta). Chemists and physicists differ in their approach to biology as exemplified by the issues of electron transport and the Mitchell hypothesis. Chemical pathways invoking high-energy intermediates fit the developing notion of cellular energy stored as ATP. However, Mitchell's physical hypothesis of cells storing energy in a field and in a proton gradient across a membrane was alien to biochemists focusing on pathways of intermediary metabolism: acceptance of the hypothesis took time.

Scientists who identify themselves as membrane biophysicists examine mechanisms of channel behavior, cable properties of neuronal cellular processes, membrane fusion, membrane fission, membrane organization, membrane lipid phase behavior, protein clustering, and many other topics. While cell biologists aim to identify and characterize the molecules that comprise the structures that allow the membrane to carry out these various processes, membrane biophysicists (and physical chemists) wonder how these processes work — what are the forces, energies, and pathways that explain the activity? There is a biophysical question associated with every protein and lipid in a membrane, with every one of their activities, and with all of their interactions. In voltage-dependent channel gating, it is now proposed by some that a peptide migrates across the lipid part of the membrane in response to the membrane potential. In membrane structure, it is proposed that the lipid environment controls protein clustering for activity. In apoptosis, we think that lipids are part of an apoptotic pore. In viral fusion, we think that lipids are physically stressed by proteins inducing local membrane curvature. In membrane trafficking, we think that proteins induce switch-like changes in membrane tubule diameter, rapidly cycling from wide to narrow.



A molecular dynamics simulation snapshot of a bilayer comprising the phospholipid DMPC (1,2-dimyristoyl-sn-glycero-3-phosphocholine) at 30°C. The simulation was performed on the NIH biowulf cluster using CHARMM software.

Through the work of Van't Hoff, Fick, Gibbs, Nernst, and Einstein in the 19th and 20th centuries came our current understanding of osmosis, diffusion, diffusion potentials, and Brownian motion on the physical level. These principles were then applied to biological problems of semipermeable membranes in the 50s and 60s and provided the intellectual basis of the ultimate acceptance of the Mitchell hypothesis and the Hodgkin–Huxley analysis of excitable membranes. A further advance came with electron microscopy, itself a tour de force of physics and electrical engineering that brought cell biologists the discovery that all of the intracellular organelles are bound by bilayer membranes. Electron microscopy revealed the ubiquity of membrane fusion and fission (the simultaneous coalescence or separation of two membranes) in the processes of exocytosis and endocytosis (processes by which materials are secreted and internalized, respectively) and protein synthesis through the secretory pathway. Then came the

use of reconstituted membranes to examine channels and fusion. Studies of physical principles of membranes also revealed membrane–membrane interactions, including the discovery of the hydration force, the phase behavior of lipids, and the co-existence of lipid domains.

Forces that Dominate Membrane Structure and Function

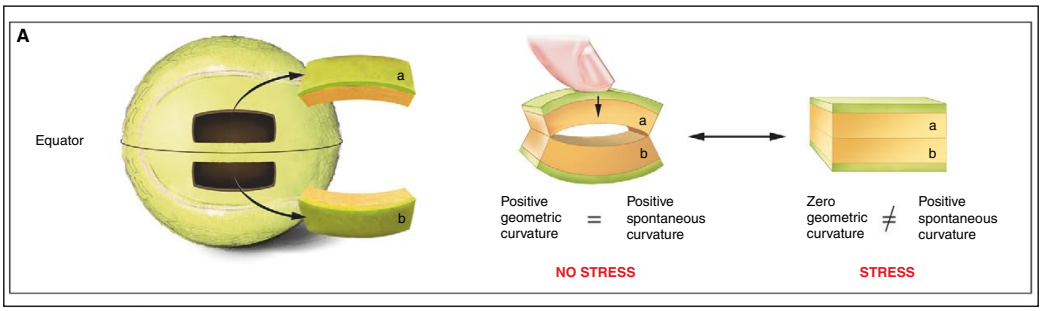
The primary force of membrane assembly is water's self-love, the hydrophobic effect that results from lipids having both polar and non-polar molecular regions. Briefly, water is so unhappy next to the acyl chains that make up the lipid tails of phospholipids and next to certain amino acid residues (hydrophobic amino acids) that acyl chains and transmembrane domains pack together to minimize contact with water. On the other hand, the head group of a lipid is very polar and happy to be near water. The bilayer structure with two planar monolayers of lipids joined together by their tails is the solution to the problem of excluding water from hydrophobic tails and transmembrane domains. To give some idea of how much energy stabilizes this structure, consider that it takes about 55 kcal/mole to remove a single acyl chain from a bilayer and put it into water. That means the energy of stabilization of the membrane is equivalent to about 16 high energy phosphate bonds (ATP, ADP) per lipid molecule. For this reason, membranes hate edges — even a 20 nm edge of a ruptured membrane would expose about a hundred lipids or cost 11000 kcal/mole. Since no other physical force comes close to generating this amount of energy and since more than 95% of the hydrophobic lateral area of a typical biological membrane is occupied by lipids, the hydrophobic effect dominates in biological membranes.

In addition to the hydrophobic effect, other intermolecular forces act to determine the net surface tension of a phospholipid membrane. The lateral distance between lipids in a bilayer membrane at equilibrium is a balance between several factors: Compression caused by the Van der Waals attraction of the acyl chains; separation caused by the entropic motions of the acyl chains to occupy all available configurations and thus push each other out of the way; close packing at the level of the carbonyl groups that form the junction between the head group and the hydrocarbon tail to avoid water incursion; and generally expansive interactions of the head groups with each other and solvent components due to their hydration, ion binding, and endogenous electrostatics. Any outside force that attempts to change this equilibrium lateral distance (about 8 Å per lipid head group) meets stiff resistance. The elastic modulus of a membrane is about 20 kT/nm. Furthermore, membranes cannot withstand any appreciable stretching and break at about 3% stretch, regardless of composition.

Because membrane shape represents the intrinsic desire of lipids to occupy structure, membrane curvature is proving to be a very important feature of a number of biological processes. The mathematical basis for the analysis of the energetics of membrane curvature and the contributions of lipids and different proteins to various intracellular structures involved in membrane trafficking has recently been reviewed. One lesson learnt was that local spontaneous curvature of the membrane

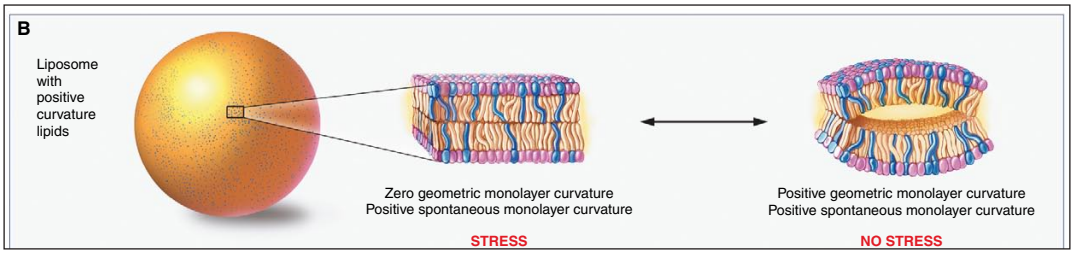
promoted by protein inserting and binding to membranes is more likely to lead to the formation of intermediates of curved organelles (e.g. during budding or endocytosis). Protein polymerization into coats can then stabilize bent membranes into vesicles whose intrinsic bilayer curvature can be tuned by the presence or absence of local curvature agents, thereby allowing variable driving forces for subsequent changes in membrane curvature.

Intrinsic to the study of membrane biophysics is the study of the composition and structure of the phospholipid bilayer. Here the X-ray diffraction studies of membranes by Luzatti and his colleagues were critical. To create a repeating structure for diffraction, they prepared multilayers of membranes. Hence, in addition to determining the electron density of material across the bilayer, they could measure the repeat distance between bilayers or the hydration volume between layers, and also determine the phase of the structure (e.g. lamellar, hexagonal, inverted hexagonal, liquid, gel, or liquid ordered). This study therefore had in its origin both the hydration forces and the phase behavior of lipids. The fact that the phases could interconvert following changes in osmotic pressure led to the proposal that lipids are polymorphic and to the idea of spontaneous curvatures of lipid monolayers. The central concept linking the spontaneous curvature of a lipid to a working concept of 'effective shape' is that, when an experimentally determined structure of a monolayer is equally partitioned into its different lipids, one can equally divide the lipidic volume into a number of selfsimilar lipids, whose shape would represent an average volumetric shape that each lipid occupies. Since lipids are fluid, each lipid rapidly interconverts between many isomeric configurations. Head group charge, methylation, motion, solvent component interactions, and other factors can shrink or expand head group hydration, changing the 'effective volume' of the phospholipids. 'Effective shape' is an idealization since no one lipid ever holds one shape for very long, but, as a practical matter, the idea of lipid shape makes curvature-driven hypotheses easier to imagine.

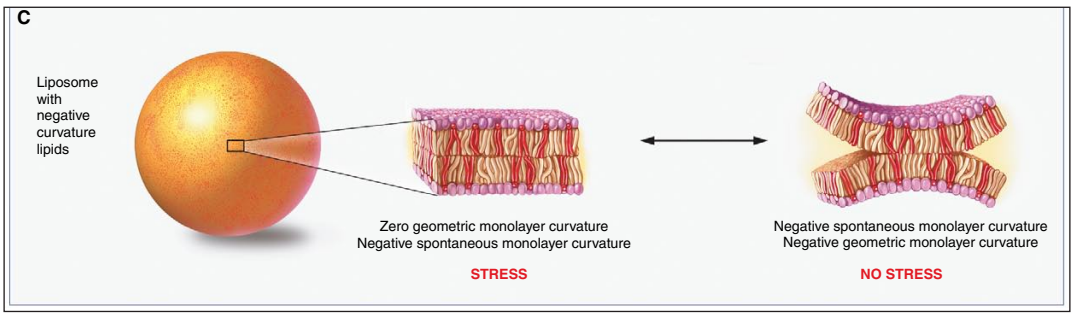


(A) An ordinary tennis ball is shown with cuts to produce two rectangular pieces of the tennis ball surface (a and b). Since the tennis ball is made of layers having different areas, there is a natural tendency of these pieces to curl. That is, in the absence of any outside forces, they spontaneously exhibit curvature, here denoted by the geometric curvature (a measure of the actual observed curvature). In the middle of (A), this spontaneous curvature of each layer is seen when the two pieces are loosely assembled together. Since these pieces of ball are elastic, it is possible to flatten these layers by

applying outside force, i.e. the curvature can go to zero (neutral) by applying pressure with fingers. However, the pieces are now stressed, and by removing fingers they will spring back to their intrinsic, spontaneous curved state. This demonstrates three features of the bending elasticity of all sheets: 1) There is a spontaneous curvature, and when geometric curvature is the same as spontaneous curvature, there is no curvature stress, 2) Curvature stress is created when the sheet is deformed to any other curvature, and 3) In the absence of external forces, the curvature stress is relieved by bending back to the spontaneous curvature.



(B) A large phospholipid vesicle is shown in water (light blue background), composed of mostly neutral lipids (yellow) but with a significant fraction of lipids having a positive spontaneous curvature (blue, e.g. lysolipids that have only one acyl tail). In this thought experiment we cut out such a small piece of membrane that it is essentially flat and we can see the individual lipids. Since membranes hate edges in water, in this cartoon we place air (yellow) around all the exposed hydrocarbon tails. The bilayer is flat, despite the fact that it is composed of two monolayers that would tend to bend if separated. Thus these flat monolayers are stressed, experiencing 'elastic frustration'. If we could allow yellow air to split the bilayer into two monolayers, we would then see them relax and exhibit their spontaneous curvature, as shown in the last figure of this panel. Since this lipid mixture has overall positive curvature, each monolayer would bend to bulge the head groups out. Essentially, the requirements of monolayer contiguity enforced by the hydrophobic effect are the 'fingers' pushing the monolayers together in aqueous solution. Similarly to the tennis ball pieces, this pushing takes work, and thus stores energy that can be used for work, e.g. for forming bent intermediates.



(C) The same as (B) except the composition is mostly neutral lipids with a significant

fraction of lipids having negative spontaneous curvature (red, e.g. phosphatidyl ethanolamine, which has a small poorly hydrated head). Now the monolayers wish to bend in the opposite direction than those in (B). Many thanks to Lydia Kibiuc for the artwork, Sol Gruner for the tennis ball idea, and Leonid Chernomordik and Michael Kozlov for their critique of the figure.

The bending energy, then, is the work required to bend the monolayer away from its intrinsic shape towards a new shape, representing a new kind of cellular energy in addition to high-energy bonds and gradients across membranes, as it is energy that can be stored in the 'bent spring' of a membrane. Since, in general, the biological membrane is constrained to a lamellar phase, and its shape can be formed by proteins acting as a scaffold, the incorporation of lipids with differing intrinsic curvatures does not change the geometric curvature of the membrane, but it does alter the work required to promote further changes, such as bending a monolayer towards the center of the bilayer membrane sufficiently to create a pore lined by headgroups, or bending a monolayer away from the center of the bilayer to create a stalk between two proximal membranes. Since a small vesicle (such as a sonicated liposome or an internal vesicle of a multivesicular body) has an outer membrane with a positive intrinsic curvature, if its lipids have slightly negative monolayer curvature (and almost all naturally occurring phospholipids are slightly negative in intrinsic monolayer curvature), then the lipids would find it very easy to bend away from the membrane and form a stalk.

Electrostatics also have a major role in membrane structure and function, but to a lesser extent than the hydrophobic effect, bending, tension, and elasticity. A significant fraction of lipids bear negative charge at physiological pH, and many proteins and glycoproteins have a multitude of charges. Surface charge in turn attracts a cloud of counterions, most easily modeled simply as a layer of negative charge at the membrane surface and a layer of positive charge a short distance (the Debye length) into the aqueous phase. Of course, actual binding of ions, such as calcium, neutralize charge (the Stern layer) and there are many instructive complexities; phospholipid membranes are a textbook example of the physics of charged surfaces in ionic solution. Specific positive charge on protein surfaces explains the binding of extrinsic proteins to membranes. Also, surface concentrations of polyphosphate inositide lipids interact with specific domains of signaling and structural proteins (complementary binding domains), and some of these lipids may in turn regulate curvature by regulating the binding of curvature-active protein domains to membranes. Studies of cells expressing mutant versions of some of these proteins show that the curvature mechanisms first detected with lipid bilayers also apply to membranes containing these proteins in vivo.

Membrane Microdomains

The challenge in the current field of membrane microheterogeneity is to discover its inner logic and categorize the types of membrane microdomains that allow for optimal organization and efficiency of the various processes that go on at the membrane surface.

It may be that there are aspects of the membrane's composition that hinder or help proteins to do the work for which they are selected, and there must be mechanisms by which the right lipids get to these proteins. Evolutionary pressures have left us with membranes having hundreds of different lipid species, representing a selective pressure at some point for each of them. These hundreds of lipid species and thousands of membrane protein species make for a large universe of potential subsets of specific lipids and proteins; each subset may cooperate to form a type of functional unit of the biological membrane. How could such hypothetical units be organized in fluid membranes?

Surprisingly, the gel phase of a membrane offers clues to good organizational principles for proteins that need to form macromolecular complexes in the plane of the membrane. These complexes are real lattices with real lattice energies and result in the packing of crystalline hydrocarbons at the right temperature. For example, hexane has a discrete intermolecular spacing of 4.1 Å when frozen, compared to an average spacing of 4.5 Å when melted. By controlling the number of lipids between them, proteins quantize the distance between themselves, perhaps nucleating a small volume where the intercalating lipids are ordered at temperatures higher than their transition temperatures.

Just as a biological membrane is a mosaic of lipids with very different curvatures that all stay flat despite their individual predilections, a biological membrane is also a mosaic of lipids with very different melting points, even those whose preferred state at physiological temperature is the gel state (solid). Thus we may imagine that proteins have a critical role in acting as well-defined solid objects in the membrane that impose constraints on lipid configurations and thus favor certain lipids for their nearest neighbor. With a huge variety of protein trans membrane domain sequences and a large variety of lipids, it is reasonable to imagine that proteins and lipids evolved together to build up biological structure in the membrane, and perhaps each membrane complex has its own specific protein and lipid compositions. Thus the growing evidence for biological membranes having significant heterogeneity may be more a reflection of this large degree of organization than the global duality of liquid-ordered vs. liquid-disordered microdomains that is seen in three-component phospholipid- and cholesterol containing bilayers.

2

Biophysical Structure and Phenomena

Biophysical structure and phenomena include the concepts of enzyme kinetics, cell division, cell migration, cell signaling, transfer RNA, myoglobin, pancreatic ribonuclease, phospholipids, etc. The topics elaborated in this chapter will help in gaining a better perspective of biophysical structure and phenomena.

Enzyme Kinetics

Enzymes are biological catalysts. They are specific to one type of reaction and to one or a small group of reactants called substrates. A catalyst speeds up the rate of a reaction without being changed itself. They are necessary as most biological reactions are very slow. Kinetics is the study of reaction rates. This will be considered in the context of enzymes where the rate of the reaction means the rate of product formation.

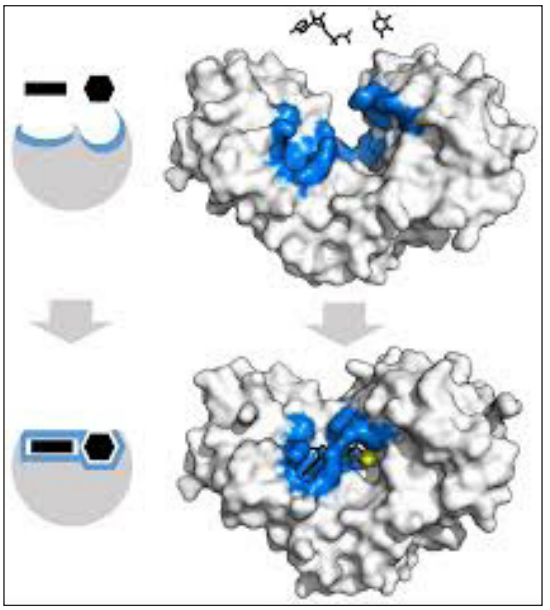
Enzyme Structure

Most enzymes are globular proteins with the exception of a few RNA enzymes (ribozymes). They have an active site made up of a few amino acids. This is where the reaction occurs. The rest of the enzyme acts as a scaffold, bringing these key amino acids together.

The active site forms a cleft or crevice that the substrate can sit in during the reaction. The cleft creates a better environment for the reaction to take place. They may do this, for example, by excluding water.

The active site is almost complementary to the substrate's shape. Therefore, when the substrate binds, the enzyme must change shape slightly to fit it. This forms the enzyme-substrate complex, also called "ES". This is the induced fit model, which is an addition to the lock and key hypothesis. Only weak bonds between the enzyme and substrate hold them in place. This is necessary to allow dissociation later on.

Enzymes have an optimum temperature and pH where they work best. Changes in pH can alter critical ionization states, while changes in temperature can disrupt important bonds. Deviations from these optimum conditions will disrupt the enzyme's structure and impact on its kinetics, eventually completely denaturing and disabling it.



Hexokinase: The Blue region corresponds to the active site and the black molecules are the substrates.

Enzyme Function

Enzymes lower the activation energy, E_a, of a particular reaction. They can do this because they have a high affinity for a transition state. The activation energy is the minimum energy needed for a reaction to occur. Enzymes assist in the reaction so that less energy is needed. This means the reaction can occur more easily. This speeds up the rate of the reaction as it allows the product to be formed faster.

An enzyme has a high affinity for the transition state (even higher than for its substrate). Therefore when the substrate binds, it is quickly forced into the transition state. This is a state that exists between the substrate and the product. The enzyme is said to facilitate the formation of the transition state.

The transition state has a high energy, making it very unstable. It can only exist transiently. The transition state spontaneously turns into the more stable product with lower energy. The enzyme will have a low affinity for the product and so the product is released.

Rate Limiting Steps

The rate limiting step in any reaction is its slowest step. It sets the pace for the entire reaction. After all, a production line can only be as productive as its slowest worker. In

enzymatic reactions, the conversion of the enzyme-substrate complex to the product is normally rate limiting. The rate of this step (and therefore the entire enzymatic reaction) is directly proportional to the concentration of ES.

The concentration of ES changes as the reaction progresses. Therefore, the rate of product formation also changes over time. When the reaction reaches equilibrium(steady state) the concentration of ES (and therefore the rate) remains relatively constant.

Reaction Kinetics

When an enzyme is added to a lot of substrate, the reaction that follows occurs in three-stages with distinct kinetics:

Phase	Concentration of ES	Rate of product formation
Pre-steady state	Burst of ES complexes form.	Slow as must first wait for ES to form. Speeds up as ES forms.
Steady state (equilibrium)	ES remains constant. It is formed as quickly as it breaks down.	Constant rate of formation. Faster than pre-steady state.
Post-steady state	Substrate depletes so fewer ES complexes can form.	Slow as there are fewer ES complexes. Slows down as substrate runs out.

The pre-steady state phase is very short as equilibrium is reached within microseconds. If you measure the rate in the first few seconds of a reaction (V_o) you are actually measuring the steady state.

Molecular Motor

Molecular motors are natural (biological) or artificial molecular machines that are the essential agents of movement in living organisms. In general terms, a motor is a device that consumes energy in one form and converts it into motion or mechanical work; for example, many protein-based molecular motors harness the chemical free energy released by the hydrolysis of ATP in order to perform mechanical work. In terms of energetic efficiency, this type of motor can be superior to currently available man-made motors. One important difference between molecular motors and macroscopic motors is that molecular motors operate in the thermal bath, an environment in which the fluctuations due to thermal noise are significant.

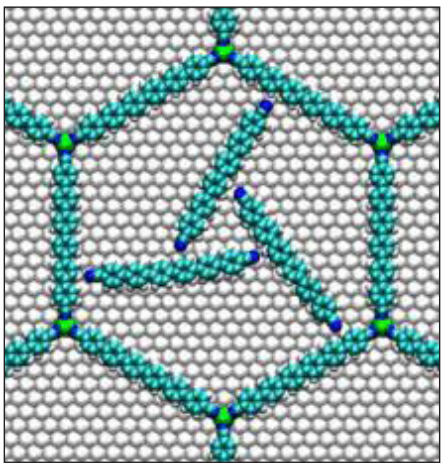
Some examples of biologically important molecular motors:

- Cytoskeletal motors:
    - Myosins are responsible for muscle contraction, intracellular cargo transport, and producing cellular tension.

- Kinesin moves cargo inside cells away from the nucleus along microtubules, in anterograde transport.
- Dynein produces the axonemal beating of cilia and flagella and also transports cargo along microtubules towards the cell nucleus, in retrograde transport.

- Polymerisation motors:
  - Actin polymerization generates forces and can be used for propulsion. ATP is used.
  - Microtubule polymerization using GTP.
  - Dynamin is responsible for the separation of clathrin buds from the plasma membrane. GTP is used.

- Rotary motors:
  - F_0F_1-ATP synthase family of proteins convert the chemical energy in ATP to the electrochemical potential energy of a proton gradient across a membrane or the other way around. The catalysis of the chemical reaction and the movement of protons are coupled to each other via the mechanical rotation of parts of the complex. This is involved in ATP synthesis in the mitochondria and chloroplasts as well as in pumping of protons across the vacuolar membrane.



Molecular dynamics simulation of a synthetic molecular motor composed of three molecules in a nanopore (outer diameter 6.7 nm) at 250 K.

  - The bacterial flagellum responsible for the swimming and tumbling of E. coli and other bacteria acts as a rigid propeller that is powered by a rotary motor. This motor is driven by the flow of protons across a membrane, possibly using a similar mechanism to that found in the F_o motor in ATP synthase.

- Nucleic acid motors:
  - RNA polymerase transcribes RNA from a DNA template.
  - DNA polymerase turns single-stranded DNA into double-stranded DNA.
  - Helicases separate double strands of nucleic acids prior to transcription or replication. ATP is used.
  - Topoisomerases reduce supercoiling of DNA in the cell. ATP is used.
  - RSC and SWI/SNF complexes remodel chromatin in eukaryotic cells. ATP is used.
  - SMC proteins responsible for chromosome condensation in eukaryotic cells.
  - Viral DNA packaging motors inject viral genomic DNA into capsids as part of their replication cycle, packing it very tightly. Several models have been put forward to explain how the protein generates the force required to drive the DNA into the capsid. An alternative proposal is that, in contrast with all other biological motors, the force is not generated directly by the protein, but by the DNA itself. In this model, ATP hydrolysis is used to drive protein conformational changes that alternatively dehydrate and rehydrate the DNA, cyclically driving it from B-DNA to A-DNA and back again. A-DNA is 23% shorter than B-DNA, and the DNA shrink/expand cycle is coupled to a protein-DNA grip/release cycle to generate the forward motion that propels DNA into the capsid.
- Enzymatic motors:
  - Catalase.
  - Urease.
  - Aldolase.
  - Hexokinase.
  - Phosphoglucose isomerase.
  - Phosphofructokinase.
  - Glucose Oxidase.
- Synthetic molecular motors have been created by chemists that yield rotation, possibly generating torque.

Theoretical Considerations

Because the motor events are stochastic, molecular motors are often modeled with the

Fokker–Planck equation or with Monte Carlo methods. These theoretical models are especially useful when treating the molecular motor as a Brownian motor.

Experimental Observation

In experimental biophysics, the activity of molecular motors is observed with many different experimental approaches, among them:

- Fluorescent methods: fluorescence resonance energy transfer (FRET), fluorescence correlation spectroscopy (FCS), total internal reflection fluorescence (TIRF).
- Magnetic tweezers can also be useful for analysis of motors that operate on long pieces of DNA.
- Neutron spin echo spectroscopy can be used to observe motion on nanosecond timescales.
- Optical tweezers (not to be confused with molecular tweezers in context) are well-suited for studying molecular motors because of their low spring constants.
- Scattering techniques: single particle tracking based on dark field microscopy or interferometric scattering microscopy (iSCAT).
- Single-molecule electrophysiology can be used to measure the dynamics of individual ion channels.

Many more techniques are also used. As new technologies and methods are developed, it is expected that knowledge of naturally occurring molecular motors will be helpful in constructing synthetic nanoscale motors.

Non-biological

Recently, chemists and those involved in nanotechnology have begun to explore the possibility of creating molecular motors *de novo*. These synthetic molecular motors currently suffer many limitations that confine their use to the research laboratory. However, many of these limitations may be overcome as our understanding of chemistry and physics at the nanoscale increases. One step toward understanding nanoscale dynamics was made with the study of catalyst diffusion in the Grubb's catalyst system. Other systems like the nanocars, while not technically motors, are also illustrative of recent efforts towards synthetic nanoscale motors.

Other non-reacting molecules can also behave as motors. This has been demonstrated by using dye molecules that move directionally in gradients of polymer solution through favorable hydrophobic interactions. Another recent study has shown that dye molecules, hard and soft colloidal particles are able to move through gradient of polymer solution through excluded volume effects.

Myoglobin

Myoglobin (symbol Mb or MB) is an iron- and oxygen-binding protein found in the muscle tissue of vertebrates in general and in almost all mammals. It is distantly related to hemoglobin which is the iron- and oxygen-binding protein in blood, specifically in the red blood cells. In humans, myoglobin is only found in the bloodstream after muscle injury. It is an abnormal finding, and can be diagnostically relevant when found in blood.

Myoglobin is the primary oxygen-carrying pigment of muscle tissues. High concentrations of myoglobin in muscle cells allow organisms to hold their breath for a longer period of time. Diving mammals such as whales and seals have muscles with particularly high abundance of myoglobin. Myoglobin is found in Type I muscle, Type II A, and Type II B, but most texts consider myoglobin not to be found in smooth muscle.

Myoglobin was the first protein to have its three-dimensional structure revealed by X-ray crystallography. This achievement was reported in 1958 by John Kendrew and associates. For this discovery, John Kendrew shared the 1962 Nobel Prize in chemistry with Max Perutz. Despite being one of the most studied proteins in biology, its physiological function is not yet conclusively established: mice genetically engineered to lack myoglobin can be viable and fertile but show many cellular and physiological adaptations to overcome the loss. Through observing these changes in myoglobin-deplete mice, it is hypothesised that myoglobin function relates to increased oxygen transport to muscle, oxygen storage and as a scavenger of reactive oxygen species.

In humans myoglobin is encoded by the *MB* gene. Myoglobin can take the forms oxymyoglobin (MbO_2), carboxymyoglobin (MbCO), and metmyoglobin (met-Mb), analogously to hemoglobin taking the forms oxyhemoglobin (HbO_2), carboxyhemoglobin (HbCO), and methemoglobin (met-Hb).

Differences from Hemoglobin

Like hemoglobin, myoglobin is a cytoplasmic protein that binds oxygen on a heme group. It harbors only one heme group, whereas hemoglobin has four. Although its heme group is identical to those in Hb, Mb has a higher affinity for oxygen than does hemoglobin. This difference is related to its different role: whereas hemoglobin transports oxygen, myoglobin's function is to store oxygen.

Meat Color

Myoglobin contains hemes, pigments responsible for the colour of red meat. The

colour that meat takes is partly determined by the degree of oxidation of the myoglobin. In fresh meat the iron atom is in the ferrous (+2) oxidation state bound to an oxygen molecule (O_2). Meat cooked well done is brown because the iron atom is now in the ferric (+3) oxidation state, having lost an electron. If meat has been exposed to nitrites, it will remain pink because the iron atom is bound to NO, nitric oxide (true of, e.g., corned beef or cured hams). Grilled meats can also take on a pink "smoke ring" that comes from the iron binding to a molecule of carbon monoxide. Raw meat packed in a carbon monoxide atmosphere also shows this same pink "smoke ring" due to the same principles. Notably, the surface of this raw meat also displays the pink color, which is usually associated in consumers' minds with fresh meat. This artificially induced pink color can persist, reportedly up to one year. Hormel and Cargill are both reported to use this meat-packing process, and meat treated this way has been in the consumer market since 2003.

Role in Disease

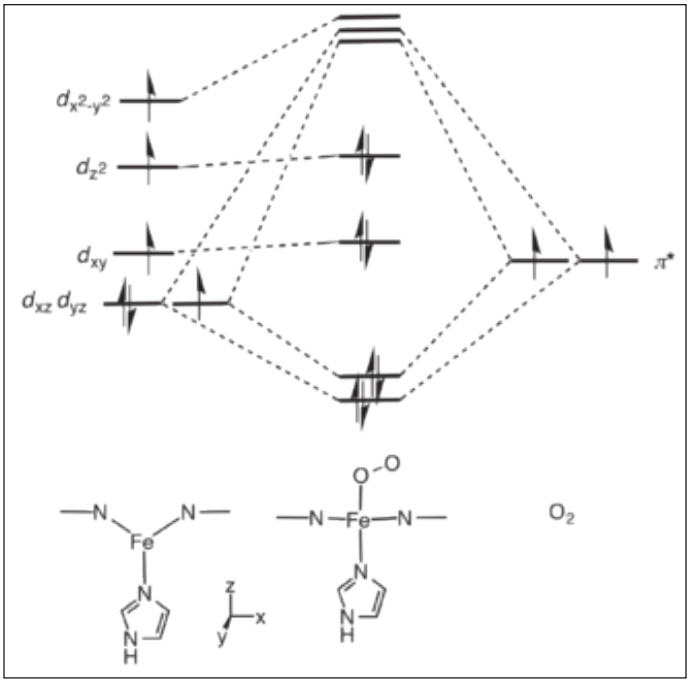
Myoglobin is released from damaged muscle tissue (rhabdomyolysis), which has very high concentrations of myoglobin. The released myoglobin is filtered by the kidneys but is toxic to the renal tubular epithelium and so may cause acute kidney injury. It is not the myoglobin itself that is toxic (it is a protoxin) but the ferrihemate portion that is dissociated from myoglobin in acidic environments (e.g., acidic urine, lysosomes).

Myoglobin is a sensitive marker for muscle injury, making it a potential marker for heart attack in patients with chest pain. However, elevated myoglobin has low specificity for acute myocardial infarction (AMI) and thus CK-MB, cardiac Troponin, ECG, and clinical signs should be taken into account to make the diagnosis.

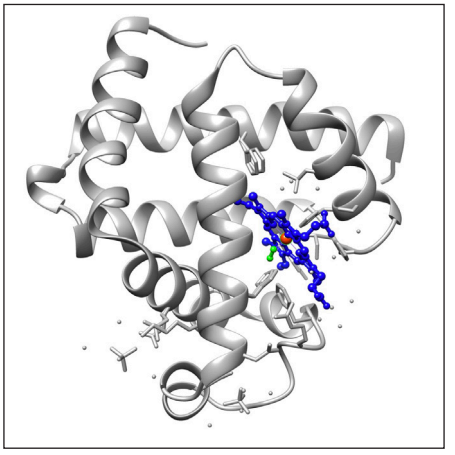
Structure and Bonding

Myoglobin belongs to the globin superfamily of proteins, and as with other globins, consists of eight alpha helices connected by loops. Myoglobin contains 154 amino acids. Myoglobin contains a porphyrin ring with an iron at its center. A *proximal* histidine group is attached directly to iron, and a *distal* histidine group hovers near the opposite face. The distal imidazole is not bonded to the iron but is available to interact with the substrate O_2. This interaction encourages the binding of O_2, but not carbon monoxide (CO), which still binds about 240× more strongly than O_2.

The binding of O_2 causes substantial structural change at the Fe center, which shrinks in radius and moves into the center of N4 pocket. O_2-binding induces "spin-pairing": The five-coordinate ferrous deoxy form is high spin and the six coordinate oxy form is low spin and diamagnetic.



Molecular orbital description of Fe-O_2 interaction in myoglobin.

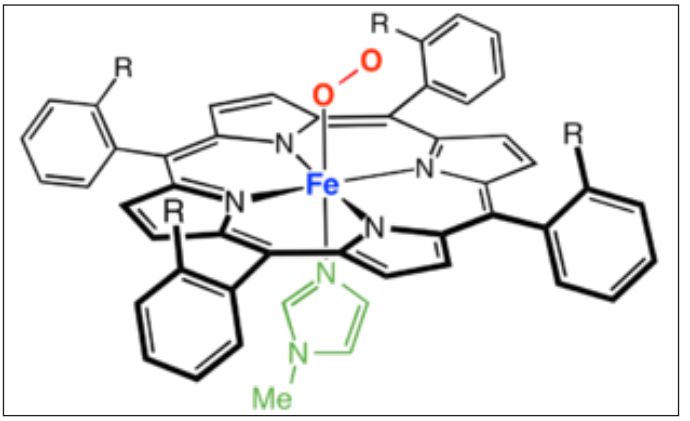


This is an image of an oxygenated myoglobin molecule. The image shows the structural change when oxygen is bound to the iron atom of the heme prosthetic group. The oxygen atoms are colored in green, the iron atom is colored in red, and the heme group is colored in blue.

Synthetic Analogues

Many models of myoglobin have been synthesized as part of a broad interest in transition metal dioxygen complexes. A well known example is the *picket fence porphyrin*, which consists of a ferrous complex of a sterically bulky derivative of tetraphenylporphyrin. In the presence of an imidazole ligand, this ferrous complex reversibly binds O_2. The O_2 substrate adopts a bent geometry, occupying the sixth position of the iron center. A key property of this model is the slow formation of the μ-oxo dimer, which is

an inactive diferric state. In nature, such deactivation pathways are suppressed by protein matrix that prevents close approach of the Fe-porphyrin assemblies.



A picket-fence porphyrin complex of Fe, with axial coordination sites occupied by methylimidazole (green) and dioxygen. The R groups flank the O_2-binding site.

Pancreatic Ribonuclease

Pancreatic ribonucleases (EC, RNase, RNase I, RNase A, pancreatic RNase, ribonuclease I, endoribonuclease I, ribonucleic phosphatase, alkaline ribonuclease, ribonuclease, gene S glycoproteins, Ceratitis capitata alkaline ribonuclease, SLSG glycoproteins, gene S locus-specific glycoproteins, S-genotype-assocd. glycoproteins, ribonucleate 3'-pyrimidino-oligonucleotidohydrolase) are pyrimidine-specific endonucleases found in high quantity in the pancreas of certain mammals and of some reptiles.

Specifically, the enzymes are involved in endonucleolytic cleavage of 3'-phosphomononucleotides and 3'-phosphooligonucleotides ending in C-P or U-P with 2',3'-cyclic phosphate intermediates. Ribonuclease can unwind the RNA helix by complexing with single-stranded RNA; the complex arises by an extended multi-site cation-anion interaction between lysine and arginine residues of the enzyme and phosphate groups of the nucleotides.

Notable Family Members

Bovine pancreatic ribonuclease is the best-studied member of the family and has served as a model system in work related to protein folding, disulfide bond formation, protein crystallography and spectroscopy, and protein dynamics. Human genome contains 8 genes that share the structure and function with bovine pancreatic ribonuclease, with 5 additional pseudo-genes. These structure and dynamics of these enzymes are related to their diverse biological functions.

Other proteins belonging to the pancreatic ribonuclease superfamily include: bovine seminal vesicle and brain ribonucleases; kidney non-secretory ribonucleases; liver-type ribonucleases; angiogenin, which induces vascularisation of normal and malignant tissues; eosinophil cationic protein, a cytotoxin and helminthotoxin with ribonuclease activity; and frog liver ribonuclease and frog sialic acid-binding lectin. The sequence of pancreatic ribonucleases contains four conserved disulfide bonds and three amino acid residues involved in the catalytic activity.

Human Genes

Human genes encoding proteins containing this domain include:

- ANG.
- RNASE1, RNASE10, RNASE12, RNASE2, RNASE3, RNASE4, RNASE6, RNASE7, and RNASE8.

Cytotoxicity

Some members of the pancreatic ribonuclease family have cytotoxic effects. Mammalian cells are protected from these effects due to their extremely high affinity for ribonuclease inhibitor (RI), which protects cellular RNA from degradation by pancreatic ribonucleases. Pancreatic ribonucleases that are not inhibited by RI are approximately as toxic as alpha-sarcin, diphtheria toxin, or ricin.

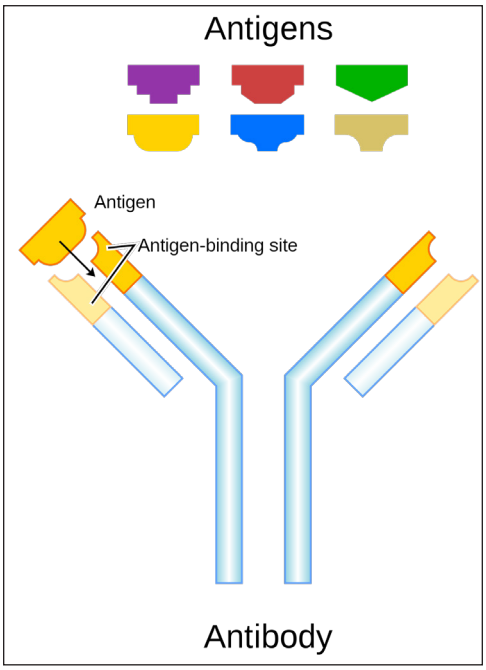
Two pancreatic ribonucleases isolated from the oocytes of the Northern leopard frog - amphinase and ranpirnase - are not inhibited by RI and show differential cytotoxicity against tumor cells. Ranpirnase was studied in a Phase III clinical trial as a treatment candidate for mesothelioma, but the trial did not demonstrate statistical significance against primary endpoints.

Antibody

An antibody (Ab), also known as an immunoglobulin (Ig), is a large, Y-shaped protein produced mainly by plasma cells that is used by the immune system to neutralize pathogens such as pathogenic bacteria and viruses. The antibody recognizes a unique molecule of the pathogen, called an antigen, via the fragment antigen-binding (Fab) variable region. Each tip of the "Y" of an antibody contains a paratope (analogous to a lock) that is specific for one particular epitope (analogous to a key) on an antigen, allowing these two structures to bind together with precision. Using this binding mechanism, an antibody can *tag* a microbe or an infected cell for attack by other parts of the immune system, or can neutralize its target directly (for example, by inhibiting a part of a microbe that is essential for its invasion and survival). Depending on the antigen, the binding

may impede the biological process causing the disease or may activate macrophages to destroy the foreign substance. The ability of an antibody to communicate with the other components of the immune system is mediated via its Fc region (located at the base of the "Y"), which contains a conserved glycosylation site involved in these interactions. The production of antibodies is the main function of the humoral immune system.

Antibodies are secreted by B cells of the adaptive immune system, mostly by differentiated B cells called plasma cells. Antibodies can occur in two physical forms, a soluble form that is secreted from the cell to be free in the blood plasma, and a membrane-bound form that is attached to the surface of a B cell and is referred to as the B-cell receptor (BCR). The BCR is found only on the surface of B cells and facilitates the activation of these cells and their subsequent differentiation into either antibody factories called plasma cells or memory B cells that will survive in the body and remember that same antigen so the B cells can respond faster upon future exposure. In most cases, interaction of the B cell with a T helper cell is necessary to produce full activation of the B cell and, therefore, antibody generation following antigen binding. Soluble antibodies are released into the blood and tissue fluids, as well as many secretions to continue to survey for invading microorganisms.



Each antibody binds to a specific antigen; an interaction similar to a lock and key.

Antibodies are glycoproteins belonging to the immunoglobulin superfamily. They constitute most of the gamma globulin fraction of the blood proteins. They are typically made of basic structural units—each with two large heavy chains and two small light chains. There are several different types of antibody heavy chains that define the five different types of crystallisable fragments (Fc) that may be attached to the antigen-binding fragments. The five different types of Fc regions allow antibodies to be grouped into

five *isotypes*. Each Fc region of a particular antibody isotype is able to bind to its specific Fc Receptor (FcR), except for IgD, which is essentially the BCR, thus allowing the antigen-antibody complex to mediate different roles depending on which FcR it binds. The ability of an antibody to bind to its corresponding FcR is further modulated by the structure of the glycans present at conserved sites within its Fc region. The ability of antibodies to bind to FcRs helps to direct the appropriate immune response for each different type of foreign object they encounter. For example, IgE is responsible for an allergic response consisting of mast cell degranulation and histamine release. IgE's Fab paratope binds to allergic antigen, for example house dust mite particles, while its Fc region binds to Fc receptor ε. The allergen-IgE-FcRε interaction mediates allergic signal transduction to induce conditions such as asthma.

Though the general structure of all antibodies is very similar, a small region at the tip of the protein is extremely variable, allowing millions of antibodies with slightly different tip structures, or antigen-binding sites, to exist. This region is known as the *hypervariable region*. Each of these variants can bind to a different antigen. This enormous diversity of antibody paratopes on the antigen-binding fragments allows the immune system to recognize an equally wide variety of antigens. The large and diverse population of antibody paratope is generated by random recombination events of a set of gene segments that encode different antigen-binding sites (or *paratopes*), followed by random mutations in this area of the antibody gene, which create further diversity. This recombinational process that produces clonal antibody paratope diversity is called V(D)J or VJ recombination. Basically, the antibody paratope is polygenic, made up of three genes, V, D, and J. Each paratope locus is also polymorphic, such that during antibody production, one allele of V, one of D, and one of J is chosen. These gene segments are then joined together using random genetic recombination to produce the paratope. The regions where the genes are randomly recombined together is the hyper variable region used to recognise different antigens on a clonal basis.

Antibody genes also re-organize in a process called class switching that changes the one type of heavy chain Fc fragment to another, creating a different isotype of the antibody that retains the antigen-specific variable region. This allows a single antibody to be used by different types of Fc receptors, expressed on different parts of the immune system.

Forms

The membrane-bound form of an antibody may be called a *surface immunoglobulin* (sIg) or a *membrane immunoglobulin* (mIg). It is part of the *B cell receptor* (BCR), which allows a B cell to detect when a specific antigen is present in the body and triggers B cell activation. The BCR is composed of surface-bound IgD or IgM antibodies and associated Ig-α and Ig-β heterodimers, which are capable of signal transduction. A typical human B cell will have 50,000 to 100,000 antibodies bound to its surface. Upon antigen binding, they cluster in large patches, which can exceed 1 micrometer in diameter, on lipid rafts that isolate the BCRs from most other cell signaling receptors. These patches may improve the efficiency of the cellular immune response. In humans, the

cell surface is bare around the B cell receptors for several hundred nanometers, which further isolates the BCRs from competing influences.

Antibody–antigen Interactions

The antibody's paratope interacts with the antigen's epitope. An antigen usually contains different epitopes along its surface arranged discontinuously, and dominant epitopes on a given antigen are called determinants.

Antibody and antigen interact by spatial complementarity (lock and key). The molecular forces involved in the Fab-epitope interaction are weak and non-specific – for example electrostatic forces, hydrogen bonds, hydrophobic interactions, and van der Waals forces. This means binding between antibody and antigen is reversible, and the antibody's affinity towards an antigen is relative rather than absolute. Relatively weak binding also means it is possible for an antibody to cross-react with different antigens of different relative affinities.

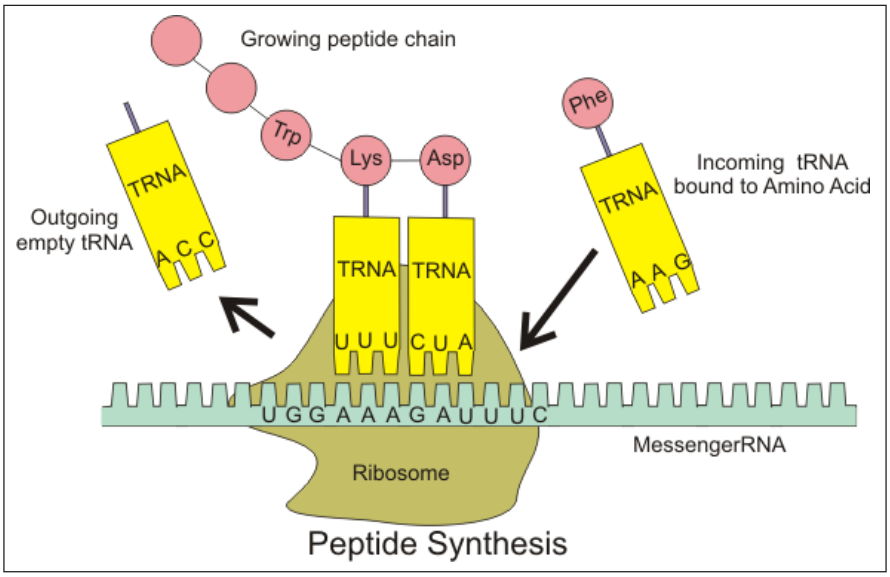
Often, once an antibody and antigen bind, they become an immune complex, which functions as a unitary object and can act as an antigen in its own right, being countered by other antibodies. Similarly, haptens are small molecules that provoke no immune response by themselves, but once they bind to proteins, the resulting complex or hapten-carrier adduct is antigenic.

Transfer RNA

A transfer RNA molecule is used in translation and consists of a single RNA strand that is only about 80 nucleotides long, containing an anticodon on the other end; the anticodon base-pairs with a complementary codon on mRNA and transfer RNA (abbreviated tRNA and formerly referred to as sRNA, for soluble RNA) is an adaptor molecule composed of RNA, typically 76 to 90 nucleotides in length, that serves as the physical link between the mRNA and the amino acid sequence of proteins. tRNA does this by carrying an amino acid to the protein synthetic machinery of a cell (ribosome) as directed by a 3-nucleotide sequence (codon) in a messenger RNA (mRNA). As such, tRNAs are a necessary component of translation, the biological synthesis of new proteins in accordance with the genetic code.

While the specific nucleotide sequence of an mRNA specifies which amino acids are incorporated into the protein product of the gene from which the mRNA is transcribed, the role of *tRNA* is to specify which sequence from the genetic code corresponds to which amino acid. The mRNA encodes a protein as a series of contiguous codons, each of which is recognized by a particular tRNA. One end of the tRNA matches the genetic code in a three-nucleotide sequence called the anticodon. The anticodon forms three

complementary base pairs with a codon in mRNA during protein biosynthesis. On the other end of the tRNA is a covalent attachment to the amino acid that corresponds to the anticodon sequence. Each type of tRNA molecule can be attached to only one type of amino acid, so each organism has many types of tRNA. Because the genetic code contains multiple codons that specify the same amino acid, there are several tRNA molecules bearing different anticodons which carry the same amino acid.



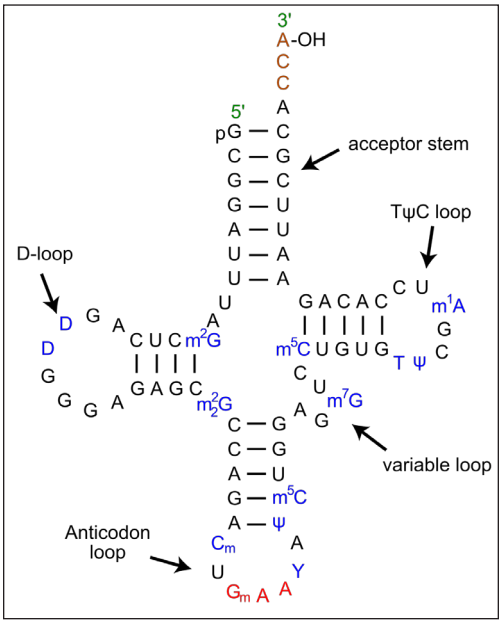
The interaction of tRNA and mRNA in protein synthesis.

The covalent attachment to the tRNA 3' end is catalyzed by enzymes called aminoacyl tRNA synthetases. During protein synthesis, tRNAs with attached amino acids are delivered to the ribosome by proteins called elongation factors, which aid in association of the tRNA with the ribosome, synthesis of the new polypeptide, and translocation (movement) of the ribosome along the mRNA. If the tRNA's anticodon matches the mRNA, another tRNA already bound to the ribosome transfers the growing polypeptide chain from its 3' end to the amino acid attached to the 3' end of the newly delivered tRNA, a reaction catalyzed by the ribosome. A large number of the individual nucleotides in a tRNA molecule may be chemically modified, often by methylation or deamidation. These unusual bases sometimes affect the tRNA's interaction with ribosomes and sometimes occur in the anticodon to alter base-pairing properties.

Structure

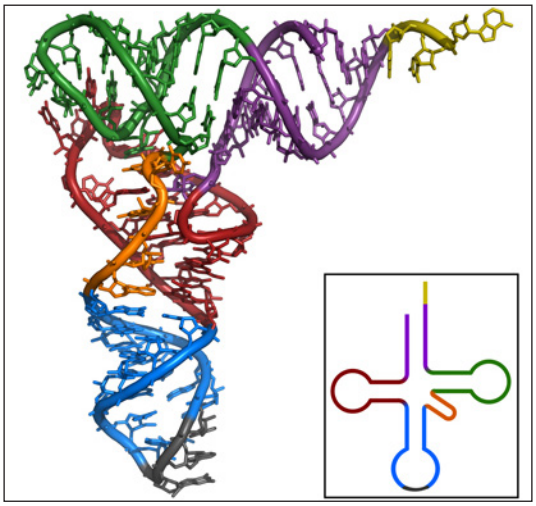
The structure of tRNA can be decomposed into its primary structure, its secondary structure (usually visualized as the *cloverleaf structure*), and its tertiary structure (all tRNAs have a similar L-shaped 3D structure that allows them to fit into the P and A sites of the ribosome). The cloverleaf structure becomes the 3D L-shaped structure through coaxial stacking of the helices, which is a common RNA tertiary structure

motif. The lengths of each arm, as well as the loop 'diameter', in a tRNA molecule vary from species to species. The tRNA structure consists of the following:



Secondary *cloverleaf structure* of $tRNA^{Phe}$ from yeast.

- A 5'-terminal phosphate group.
- The acceptor stem is a 7- to 9-base pair (bp) stem made by the base pairing of the 5'-terminal nucleotide with the 3'-terminal nucleotide (which contains the CCA 3'-terminal group used to attach the amino acid). In general, such 3'-terminal tRNA-like structures are referred to as 'genomic tags'. The acceptor stem may contain non-Watson-Crick base pairs.



Tertiary structure of tRNA. *CCA tail* in yellow, *Acceptor stem* in purple, *Variable loop* in orange, *D arm* in red, *Anticodon arm* in blue with *Anticodon* in black, *T arm* in green.

- The CCA tail is a cytosine-cytosine-adenine sequence at the 3' end of the tRNA molecule. The amino acid loaded onto the tRNA by aminoacyl tRNA synthetases, to form aminoacyl-tRNA, is covalently bonded to the 3'-hydroxyl group on the CCA tail. This sequence is important for the recognition of tRNA by enzymes and critical in translation. In prokaryotes, the CCA sequence is transcribed in some tRNA sequences. In most prokaryotic tRNAs and eukaryotic tRNAs, the CCA sequence is added during processing and therefore does not appear in the tRNA gene.
- The D arm is a 4- to 6-bp stem ending in a loop that often contains dihydrouridine.
- The anticodon arm is a 5-bp stem whose loop contains the anticodon. The tRNA 5'-to-3' primary structure contains the anticodon but in reverse order, since 3'-to-5' directionality is required to read the mRNA from 5'-to-3'.
- The T arm is a 4- to 5- bp stem containing the sequence TΨC where Ψ is pseudouridine, a modified uridine.
- Bases that have been modified, especially by methylation (e.g. tRNA (guanine-N7-)-methyltransferase), occur in several positions throughout the tRNA. The first anticodon base, or wobble-position, is sometimes modified to inosine (derived from adenine), queuosine (derived from guanine), uridine-5-oxyacetic acid (derived from uracil), 5-methylaminomethyl-2-thiouridine (derived from uracil), or lysidine (derived from cytosine).

3D animated GIF showing the structure of phenylalanine-tRNA from yeast (PDB ID 1ehz). White lines indicate base pairing by hydrogen bonds. In the orientation shown, the acceptor stem is on top and the anticodon on the bottom.

Anticodon

An anticodon is a unit made up of three nucleotides that correspond to the three bases of the codon on the mRNA. Each tRNA contains a distinct anticodon triplet sequence

that can form 3 complementary base pairs to one or more codons for an amino acid. Some anticodons can pair with more than one codon due to a phenomenon known as wobble base pairing. Frequently, the first nucleotide of the anticodon is one not found on mRNA: inosine, which can hydrogen bond to more than one base in the corresponding codon position. In the genetic code, it is common for a single amino acid to be specified by all four third-position possibilities, or at least by both pyrimidines and purines; for example, the amino acid glycine is coded for by the codon sequences GGU, GGC, GGA, and GGG. Other modified nucleotides may also appear at the first anticodon position—sometimes known as the "wobble position"—resulting in subtle changes to the genetic code, as for example in mitochondria. Per cell, 61 types of tRNA would be required to provide a one-to-one correspondence between tRNA molecules and codons that specify amino acids, as there are 61 sense codons of the standard genetic code. However, many cells contain fewer than 61 types of tRNAs because the wobble base is capable of binding to several, though not necessarily all, of the codons that specify a particular amino acid. A minimum of 31 tRNAs are required to translate, unambiguously, all 61 sense codons; the maximum observed is 41.

Aminoacylation

Aminoacylation is the process of adding an aminoacyl group to a compound. It covalently links an amino acid to the CCA 3' end of a tRNA molecule. Each tRNA is aminoacylated (or *charged*) with a specific amino acid by an aminoacyl tRNA synthetase. There is normally a single aminoacyl tRNA synthetase for each amino acid, despite the fact that there can be more than one tRNA, and more than one anticodon for an amino acid. Recognition of the appropriate tRNA by the synthetases is not mediated solely by the anticodon, and the acceptor stem often plays a prominent role. Reaction:

- Amino acid + ATP → aminoacyl-AMP + PPi.
- Aminoacyl-AMP + tRNA → aminoacyl-tRNA + AMP.

Certain organisms can have one or more aminoacyl tRNA synthetases missing. This leads to charging of the tRNA by a chemically related amino acid, and by use of an enzyme or enzymes, the tRNA is modified to be correctly charged. For example, *Helicobacter pylori* has glutaminyl tRNA synthetase missing. Thus, glutamate tRNA synthetase charges tRNA-glutamine(tRNA-Gln) with glutamate. An amidotransferase then converts the acid side chain of the glutamate to the amide, forming the correctly charged gln-tRNA-Gln.

Binding to Ribosome

The ribosome has three binding sites for tRNA molecules that span the space between the two ribosomal subunits: The A (aminoacyl), P (peptidyl), and E (exit) sites. In addition, the ribosome has two other sites for tRNA binding that are used during mRNA

decoding or during the initiation of protein synthesis. These are the T site (named elongation factor Tu) and I site (initiation). By convention, the tRNA binding sites are denoted with the site on the small ribosomal subunit listed first and the site on the large ribosomal subunit listed second. For example, the A site is often written A/A, the P site, P/P, and the E site, E/E. The binding proteins like L27, L2, L14, L15, L16 at the A- and P- sites have been determined by affinity labeling by A.P. Czernilofsky et al.

Once translation initiation is complete, the first aminoacyl tRNA is located in the P/P site, ready for the elongation cycle described below. During translation elongation, tRNA first binds to the ribosome as part of a complex with elongation factor Tu (EF-Tu) or its eukaryotic (eEF-1) or archaeal counterpart. This initial tRNA binding site is called the A/T site. In the A/T site, the A-site half resides in the small ribosomal subunit where the mRNA decoding site is located. The mRNA decoding site is where the mRNA codon is read out during translation. The T-site half resides mainly on the large ribosomal subunit where EF-Tu or eEF-1 interacts with the ribosome. Once mRNA decoding is complete, the aminoacyl-tRNA is bound in the A/A site and is ready for the next peptide bond to be formed to its attached amino acid. The peptidyl-tRNA, which transfers the growing polypeptide to the aminoacyl-tRNA bound in the A/A site, is bound in the P/P site. Once the peptide bond is formed, the tRNA in the P/P site is deacylated, or has a free 3' end, and the tRNA in the A/A site carries the growing polypeptide chain. To allow for the next elongation cycle, the tRNAs then move through hybrid A/P and P/E binding sites, before completing the cycle and residing in the P/P and E/E sites. Once the A/A and P/P tRNAs have moved to the P/P and E/E sites, the mRNA has also moved over by one codon and the A/T site is vacant, ready for the next round of mRNA decoding. The tRNA bound in the E/E site then leaves the ribosome.

The P/I site is actually the first to bind to aminoacyl tRNA, which is delivered by an initiation factor called IF2 in bacteria. However, the existence of the P/I site in eukaryotic or archaeal ribosomes has not yet been confirmed. The P-site protein L27 has been determined by affinity labeling by E. Collatz and A.P. Czernilofsky.

tRNA Genes

Organisms vary in the number of tRNA genes in their genome. For example, the nematode worm *C. elegans*, a commonly used model organism in genetics studies, has 29,647 genes in its nuclear genome, of which 620 code for tRNA. The budding yeast *Saccharomyces cerevisiae* has 275 tRNA genes in its genome.

In the human genome, which, according to January 2013 estimates, has about 20,848 protein coding genes in total, there are 497 nuclear genes encoding cytoplasmic tRNA molecules, and 324 tRNA-derived pseudogenes—tRNA genes thought to be no longer functional (although pseudo tRNAs have been shown to be involved in antibiotic resistance in bacteria). Regions in nuclear chromosomes, very similar in sequence to mitochondrial tRNA genes, have also been identified (tRNA-lookalikes). These tRNA-lookalikes

are also considered part of the nuclear mitochondrial DNA (genes transferred from the mitochondria to the nucleus).

As with all eukaryotes, there are 22 mitochondrial tRNA genes in humans. Mutations in some of these genes have been associated with severe diseases like the MELAS syndrome. Cytoplasmic tRNA genes can be grouped into 49 families according to their anticodon features. These genes are found on all chromosomes, except the 22 and Y chromosome. High clustering on 6p is observed (140 tRNA genes), as well on 1 chromosome. The HGNC, in collaboration with the Genomic tRNA Database (GtRNAdb) and experts in the field, has approved unique names for human genes that encode tRNAs.

Evolution

The top half of tRNA (consisting of the T arm and the acceptor stem with 5'-terminal phosphate group and 3'-terminal CCA group) and the bottom half (consisting of the D arm and the anticodon arm) are independent units in structure as well as in function. The top half may have evolved first including the 3'-terminal genomic tag which originally may have marked tRNA-like molecules for replication in early RNA world. The bottom half may have evolved later as an expansion, e.g. as protein synthesis started in RNA world and turned it into a ribonucleoprotein world (RNP world). This proposed scenario is called genomic tag hypothesis. In fact, tRNA and tRNA-like aggregates have an important catalytic influence (i. e. as ribozymes) on replication still today. These roles may be regarded as 'molecular (or chemical) fossils' of RNA world.

Genomic tRNA content is a differentiating feature of genomes among biological domains of life: Archaea present the simplest situation in terms of genomic tRNA content with a uniform number of gene copies, Bacteria have an intermediate situation and Eukarya present the most complex situation. Eukarya present not only more tRNA gene content than the other two kingdoms but also a high variation in gene copy number among different isoacceptors, and this complexity seem to be due to duplications of tRNA genes and changes in anticodon specificity.

Evolution of the tRNA gene copy number across different species has been linked to the appearance of specific tRNA modification enzymes (uridine methyltransferases in Bacteria, and adenosine deaminases in Eukarya), which increase the decoding capacity of a given tRNA. As an example, tRNAAla encodes four different tRNA isoacceptors (AGC, UGC, GGC and CGC). In Eukarya, AGC isoacceptors are extremely enriched in gene copy number in comparison to the rest of isoacceptors, and this has been correlated with its A-to-I modification of its wobble base. This same trend has been shown for most amino acids of eukaryal species. Indeed, the effect of these two tRNA modifications is also seen in codon usage bias. Highly expressed genes seem to be enriched in codons that are exclusively using codons that will be decoded by these modified tRNAs, which suggests a possible role of these codons—and consequently of these tRNA modifications—in translation efficiency.

tRNA-derived Fragments

tRNA-derived fragments (or tRFs) are short molecules that emerge after cleavage of the mature tRNAs or the precursor transcript. Both cytoplasmic and mitochondrial tRNAs can produce fragments. There are at least four structural types of tRFs believed to originate from mature tRNAs, including the relatively long tRNA halves and short 5'-tRFs, 3'-tRFs and i-tRFs. The precursor tRNA can be cleaved to produce molecules from the 5' leader or 3' trail sequences. Cleavage enzymes include Angiogenin, Dicer, RNase Z and RNase P. Especially in the case of Angiogenin, the tRFs have a characteristically unusual cyclic phosphate at their 3' end and a hydroxyl group at the 5' end. tRFs appear to play a role in RNA interference, specifically in the suppression of retroviruses and retrotransposons that use tRNA as a primer for replication. Half-tRNAs cleaved by angiogenin are also known as tiRNAs. The biogensis of smaller fragments, including those that function as piRNAs, are less understood.

tRFs have multiple dependencies and roles; such as exhibiting significant changes between sexes, among races and disease status. Functionally, they can be loaded on Ago and act through RNAi pathways, participate in the formation of stress granules, displace mRNAs from RNA-binding proteins or inhibit translation. At the system or the organismal level, the four types of tRFs have a diverse spectrum of activities. Functionally, tRFs are associated with viral infection, cancer, cell proliferation and also with epigenetic transgenerational regulation of metabolism. tRFs are not restricted to humans and have been shown to exist in multiple organisms.

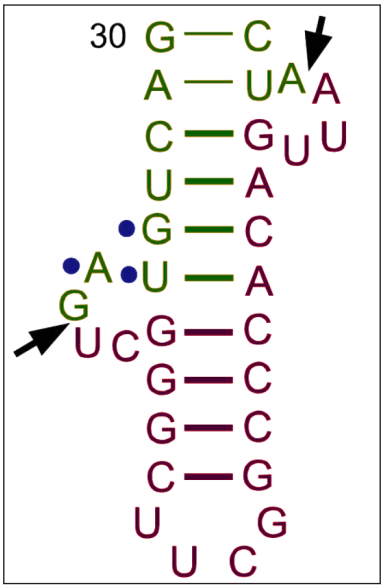
Engineered tRNAs

Artificial suppressor elongator tRNAs are used to incorporate unnatural amino acids at nonsense codons placed in the coding sequence of a gene. Engineered initiator tRNAs ($tRNA^{fMet2}$ with CUA anticodon encoded by metY gene) have been used to initiate translation at the amber stop codon UAG. This type of engineered tRNA is called a nonsense suppressor tRNA because it suppresses the translation stop signal that normally occurs at UAG codons. The amber initiator tRNA inserts methionine and glutamine at UAG codons preceded by a strong Shine-Dalgarno sequence. An investigation of the amber initiator tRNA showed that it was orthogonal to the regular AUG start codon showing no detectable off-target translation initiation events in a genomically recoded *E. coli* strain.

tRNA Biogenesis

In eukaryotic cells, tRNAs are transcribed by RNA polymerase III as pre-tRNAs in the nucleus. RNA polymerase III recognizes two highly conserved downstream promoter sequences: The 5' intragenic control region (5'-ICR, D-control region, or A box), and the 3'-ICR (T-control region or B box) inside tRNA genes. The first promoter begins at +8 of mature tRNAs and the second promoter is located 30–60 nucleotides

downstream of the first promoter. The transcription terminates after a stretch of four or more thymidines.



Bulge-helix-bulge motive of tRNA intron.

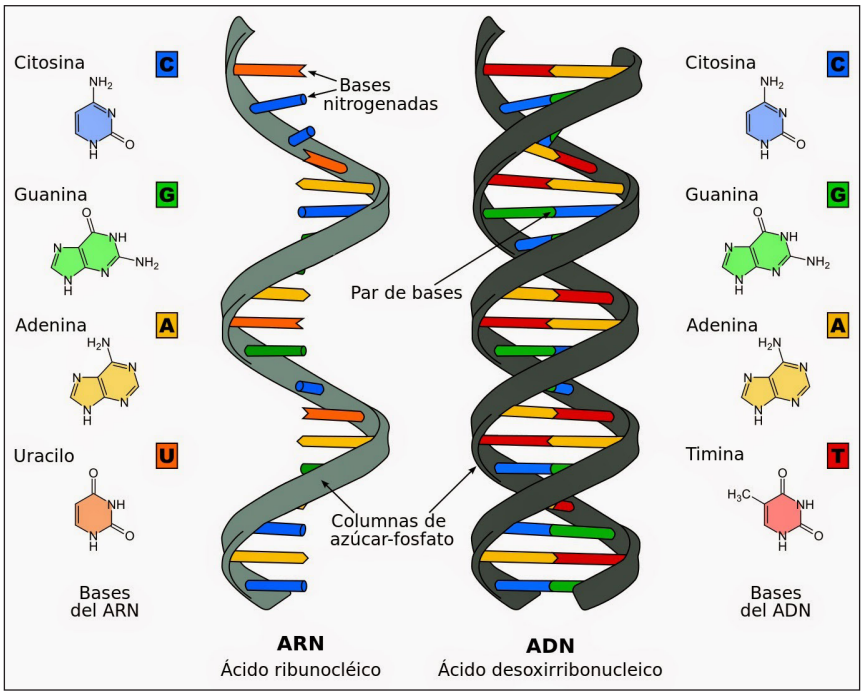
Pre-tRNAs undergo extensive modifications inside the nucleus. Some pre-tRNAs contain introns that are spliced, or cut, to form the functional tRNA molecule; in bacteria these self-splice, whereas in eukaryotes and archaea they are removed by tRNA-splicing endonucleases. Eukaryotic pre-tRNA contains bulge-helix-bulge (BHB) structure motif that is important for recognition and precise splicing of tRNA intron by endonucleases. This motif position and structure are evolutionarily conserved. However, some organisms, such as unicellular algae have a non-canonical position of BHB-motif as well as 5'- and 3'-ends of the spliced intron sequence. The 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme. A notable exception is in the archaeon *Nanoarchaeum equitans,* which does not possess an RNase P enzyme and has a promoter placed such that transcription starts at the 5' end of the mature tRNA. The non-templated 3' CCA tail is added by a nucleotidyl transferase. Before tRNAs are exported into the cytoplasm by Los1/Xpo-t, tRNAs are aminoacylated. The order of the processing events is not conserved. For example, in yeast, the splicing is not carried out in the nucleus but at the cytoplasmic side of mitochondrial membranes.

Nucleic Acids

Nucleic acids are the biopolymers, or small biomolecules, essential to all known forms of life. The term *nucleic acid* is the overall name for DNA and RNA. They are composed of nucleotides, which are the monomers made of three components: a 5-carbon sugar,

a phosphate group and a nitrogenous base. If the sugar is a compound ribose, the polymer is RNA (ribonucleic acid); if the sugar is derived from ribose as deoxyribose, the polymer is DNA (deoxyribonucleic acid).

Nucleic acids are the most important of all biomolecules. These are found in abundance in all living things, where they function to create and encode and then store information of every living cell of every life-form organism on Earth. In turn, they function to transmit and express that information inside and outside the cell nucleus—to the interior operations of the cell and ultimately to the next generation of each living organism. The encoded information is contained and conveyed via the nucleic acid sequence, which provides the 'ladder-step' ordering of nucleotides within the molecules of RNA and DNA.



Nucleic acids RNA (left) and DNA (right).

Strings of nucleotides are bonded to form helical backbones—typically, one for RNA, two for DNA—and assembled into chains of base-pairs selected from the five primary, or canonical, nucleobases, which are: adenine, cytosine, guanine, thymine, and uracil. Thymine occurs only in DNA and uracil only in RNA. Using amino acids and the process known as protein synthesis, the specific sequencing in DNA of these nucleobase-pairs enables storing and transmitting coded instructions as genes. In RNA, base-pair sequencing provides for manufacturing new proteins that determine the frames and parts and most chemical processes of all life forms.

Occurrence and Nomenclature

The term *nucleic acid* is the overall name for DNA and RNA, members of a family

of biopolymers, and is synonymous with *polynucleotide*. Nucleic acids were named for their initial discovery within the nucleus, and for the presence of phosphate groups (related to phosphoric acid). Although first discovered within the nucleus of eukaryotic cells, nucleic acids are now known to be found in all life forms including within bacteria, archaea, mitochondria, chloroplasts, viruses, and viroids. (There is debate as to whether viruses are living or non-living). All living cells contain both DNA and RNA (except some cells such as mature red blood cells), while viruses contain either DNA or RNA, but usually not both. The basic component of biological nucleic acids is the nucleotide, each of which contains a pentose sugar (ribose or deoxyribose), a phosphate group, and a nucleobase. Nucleic acids are also generated within the laboratory, through the use of enzymes (DNA and RNA polymerases) and by solid-phase chemical synthesis. The chemical methods also enable the generation of altered nucleic acids that are not found in nature, for example peptide nucleic acids.

Molecular Composition and Size

Nucleic acids are generally very large molecules. Indeed, DNA molecules are probably the largest individual molecules known. Well-studied biological nucleic acid molecules range in size from 21 nucleotides (small interfering RNA) to large chromosomes (human chromosome 1 is a single molecule that contains 247 million base pairs).

In most cases, naturally occurring DNA molecules are double-stranded and RNA molecules are single-stranded. There are numerous exceptions, however—some viruses have genomes made of double-stranded RNA and other viruses have single-stranded DNA genomes, and, in some circumstances, nucleic acid structures with three or four strands can form.

Nucleic acids are linear polymers (chains) of nucleotides. Each nucleotide consists of three components: a purine or pyrimidine nucleobase (sometimes termed *nitrogenous base* or simply *base*), a pentose sugar, and a phosphate group. The substructure consisting of a nucleobase plus sugar is termed a nucleoside. Nucleic acid types differ in the structure of the sugar in their nucleotides–DNA contains 2'-deoxyribose while RNA contains ribose (where the only difference is the presence of a hydroxyl group). Also, the nucleobases found in the two nucleic acid types are different: adenine, cytosine, and guanine are found in both RNA and DNA, while thymine occurs in DNA and uracil occurs in RNA.

The sugars and phosphates in nucleic acids are connected to each other in an alternating chain (sugar-phosphate backbone) through phosphodiester linkages. In conventional nomenclature, the carbons to which the phosphate groups attach are the 3'-end and the 5'-end carbons of the sugar. This gives nucleic acids directionality, and the ends of nucleic acid molecules are referred to as 5'-end and 3'-end. The nucleobases are

joined to the sugars via an N-glycosidic linkage involving a nucleobase ring nitrogen (N-1 for pyrimidines and N-9 for purines) and the 1' carbon of the pentose sugar ring.

Non-standard nucleosides are also found in both RNA and DNA and usually arise from modification of the standard nucleosides within the DNA molecule or the primary (initial) RNA transcript. Transfer RNA (tRNA) molecules contain a particularly large number of modified nucleosides.

Topology

Double-stranded nucleic acids are made up of complementary sequences, in which extensive Watson-Crick base pairing results in a highly repeated and quite uniform double-helical three-dimensional structure. In contrast, single-stranded RNA and DNA molecules are not constrained to a regular double helix, and can adopt highly complex three-dimensional structures that are based on short stretches of intramolecular base-paired sequences including both Watson-Crick and noncanonical base pairs, and a wide range of complex tertiary interactions.

Nucleic acid molecules are usually unbranched, and may occur as linear and circular molecules. For example, bacterial chromosomes, plasmids, mitochondrial DNA, and chloroplast DNA are usually circular double-stranded DNA molecules, while chromosomes of the eukaryotic nucleus are usually linear double-stranded DNA molecules. Most RNA molecules are linear, single-stranded molecules, but both circular and branched molecules can result from RNA splicing reactions. The total amount of pyrimidines is equal to the total amount of purines. The diameter of the helix is about 20Å.

Sequences

One DNA or RNA molecule differs from another primarily in the sequence of nucleotides. Nucleotide sequences are of great importance in biology since they carry the ultimate instructions that encode all biological molecules, molecular assemblies, subcellular and cellular structures, organs, and organisms, and directly enable cognition, memory, and behavior. Enormous efforts have gone into the development of experimental methods to determine the nucleotide sequence of biological DNA and RNA molecules, and today hundreds of millions of nucleotides are sequenced daily at genome centers and smaller laboratories worldwide.

Types

Deoxyribonucleic Acid

Deoxyribonucleic acid (DNA) is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms. The DNA segments carrying this genetic information are called genes. Likewise, other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all known forms of life. DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are, therefore, anti-parallel. Attached to each sugar is one of four types of molecules called nucleobases (informally, bases). It is the sequence of these four nucleobases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA in a process called transcription. Within cells DNA is organized into long structures called chromosomes. During cell division these chromosomes are duplicated in the process of DNA replication, providing each cell its own complete set of chromosomes. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the chromosomes, chromatin proteins such as histones compact and organize DNA. These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

Ribonucleic Acid

Ribonucleic acid (RNA) functions in converting genetic information from genes into the amino acid sequences of proteins. The three universal types of RNA include transfer RNA (tRNA), messenger RNA (mRNA), and ribosomal RNA (rRNA). Messenger RNA acts to carry genetic sequence information between DNA and ribosomes, directing protein synthesis. Ribosomal RNA is a major component of the ribosome, and catalyzes peptide bond formation. Transfer RNA serves as the carrier molecule for amino acids to be used in protein synthesis, and is responsible for decoding the mRNA. In addition, many other classes of RNA are now known.

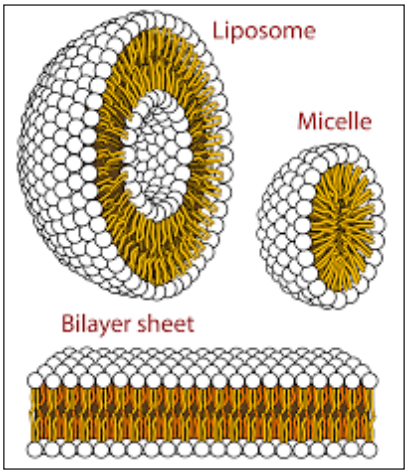
Artificial Nucleic Acid

Artificial nucleic acid analogues have been designed and synthesized by chemists, and include peptide nucleic acid, morpholino- and locked nucleic acid, glycol nucleic acid, and threose nucleic acid. Each of these is distinguished from naturally occurring DNA or RNA by changes to the backbone of the molecules.

Biological Membranes

A biological membrane or biomembrane is an enclosing or separating membrane that acts as a selectively permeable barrier within living things. Biological membranes, in

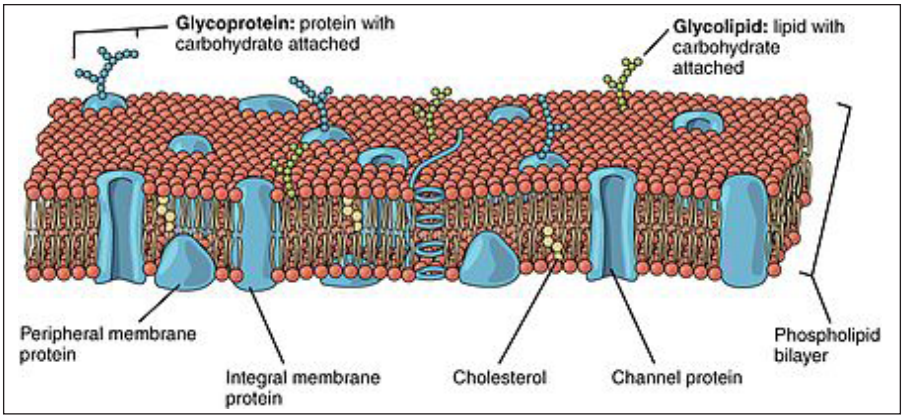
the form of eukaryotic cell membranes, consist of a phospholipid bilayer with embedded, integral and peripheral proteins used in communication and transportation of chemicals and ions. The bulk of lipid in a cell membrane provides a fluid matrix for proteins to rotate and laterally diffuse for physiological functioning. Proteins are adapted to high membrane fluidity environment of lipid bilayer with the presence of an annular lipid shell, consisting of lipid molecules bound tightly to surface of integral membrane proteins. The cell membranes are different from the isolating tissues formed by layers of cells, such as mucous membranes, basement membranes, and serous membranes.



Cross-sectional view of the structures that can be formed by phospholipids in an aqueous solution.

Composition

Asymmetry



A fluid membrane model of the phospholipid bilayer.

The lipid bilayer consists of two layers- an outer leaflet and an inner leaflet. The components of bilayers are distributed unequally between the two surfaces to create asymmetry between the outer and inner surfaces. This asymmetric organization is important for cell functions such as cell signaling. The asymmetry of the biological membrane

reflects the different functions of the two leaflets of the membrane. As seen in the fluid membrane model of the phospholipid bilayer, the outer leaflet and inner leaflet of the membrane are asymmetrical in their composition. Certain proteins and lipids rest only on one surface of the membrane and not the other.

Both the plasma membrane and internal membranes have cytosolic and exoplasmic faces. This orientation is maintained during membrane trafficking – proteins, lipids, glycoconjugates facing the lumen of the ER and Golgi get expressed on the extracellular side of the plasma membrane. In eucaryotic cells, new phospholipids are manufactured by enzymes bound to the part of the endoplasmic reticulum membrane that faces the cytosol. These enzymes, which use free fatty acids as substrates, deposit all newly made phospholipids into the cytosolic half of the bilayer. To enable the membrane as a whole to grow evenly, half of the new phospholipid molecules then have to be transferred to the opposite monolayer. This transfer is catalyzed by enzymes called flippases. In the plasma membrane, flippases transfer specific phospholipids selectively, so that different types become concentrated in each monolayer.

Using selective flippases is not the only way to produce asymmetry in lipid bilayers, however. In particular, a different mechanism operates for glycolipids—the lipids that show the most striking and consistent asymmetric distribution in animal cells.

Lipids

The biological membrane is made up of lipids with hydrophobic tails and hydrophilic heads. The hydrophobic tails are hydrocarbon tails whose length and saturation is important in characterizing the cell. Lipid rafts occur when lipid species and proteins aggregate in domains in the membrane. These help organize membrane components into localized areas that are involved in specific processes, such as signal transduction.

Red blood cells, or erythrocytes, have a unique lipid composition. The bilayer of red blood cells is composed of cholesterol and phospholipids in equal proportions by weight. Erythrocyte membrane plays a crucial role in blood clotting. In the bilayer of red blood cells is phosphatidylserine. This is usually in the cytoplasmic side of the membrane. However, it is flipped to the outer membrane to be used during blood clotting.

Proteins

Phospholipid bilayers contain different proteins. These membrane proteins have various functions and characteristics and catalyze different chemical reactions. Integral proteins span the membranes with different domains on either side. Integral proteins hold strong association with the lipid bilayer and cannot easily become detached. They will dissociate only with chemical treatment that breaks the membrane. Peripheral proteins are unlike integral proteins in that they hold weak interactions with the surface of the bilayer and can easily become dissociated from the membrane.

Peripheral proteins are located on only one face of a membrane and create membrane asymmetry.

Some examples of plasma membrane proteins and their functions		
Functional class	Protein example	Specific function
Transporters	Na+ Pump	actively pumps Na+ out of cells and K+ in.
Anchors	integrins	link intracellular actin filaments to extracellular matrix proteins.
Receptors	platelet-derived growth factor receptor.	binds extracellular PDGF and, as a consequence, generates intracellular signals that cause the cell to grow and divide.
Enzymes	adenylyl cyclase.	catalyzes the production of intracellular signaling molecule cyclic AMP in response to extracellular signals.

Oligosaccharides

Oligosaccharides are sugar containing polymers. In the membrane, they can be covalently bound to lipids to form glycolipids or covalently bound to proteins to form glycoproteins. Membranes contain sugar-containing lipid molecules known as glycolipids. In the bilayer, the sugar groups of glycolipids are exposed at the cell surface, where they can form hydrogen bonds. Glycolipids provide the most extreme example of asymmetry in the lipid bilayer. Glycolipids perform a vast number of functions in the biological membrane that are mainly communicative, including cell recognition and cell-cell adhesion. Glycoproteins are integral proteins. They play an important role in the immune response and protection.

Formation

The phospholipid bilayer is formed due to the aggregation of membrane lipids in aqueous solutions. Aggregation is caused by the hydrophobic effect, where hydrophobic ends come into contact with each other and are sequestered away from water. This arrangement maximises hydrogen bonding between hydrophilic heads and water while minimising unfavorable contact between hydrophobic tails and water. The increase in available hydrogen bonding increases the entropy of the system, creating a spontaneous process.

Function

Biological molecules are amphiphilic or amphipathic, i.e. are simultaneously hydrophobic and hydrophilic. The phospholipid bilayer contains charged hydrophilic headgroups, which interact with polar water. The layers also contain hydrophobic tails, which meet with the hydrophobic tails of the complementary layer. The hydrophobic tails are usually fatty acids that differ in lengths. The interactions of lipids, especially the hydrophobic tails, determine the lipid bilayer physical properties such as fluidity.

Membranes in cells typically define enclosed spaces or compartments in which cells may maintain a chemical or biochemical environment that differs from the outside. For example, the membrane around peroxisomes shields the rest of the cell from peroxides, chemicals that can be toxic to the cell, and the cell membrane separates a cell from its surrounding medium. Peroxisomes are one form of vacuole found in the cell that contain by-products of chemical reactions within the cell. Most organelles are defined by such membranes, and are called "membrane-bound" organelles.

Selective Permeability

Probably the most important feature of a biomembrane is that it is a selectively permeable structure. This means that the size, charge, and other chemical properties of the atoms and molecules attempting to cross it will determine whether they succeed in doing so. Selective permeability is essential for effective separation of a cell or organelle from its surroundings. Biological membranes also have certain mechanical or elastic properties that allow them to change shape and move as required. Generally, small hydrophobic molecules can readily cross phospholipid bilayers by simple diffusion.

Particles that are required for cellular function but are unable to diffuse freely across a membrane enter through a membrane transport protein or are taken in by means of endocytosis, where the membrane allows for a vacuole to join onto it and push its contents into the cell. Many types of specialized plasma membranes can separate cell from external environment: apical, basolateral, presynaptic and postsynaptic ones, membranes of flagella, cilia, microvillus, filopodia and lamellipodia, the sarcolemma of muscle cells, as well as specialized myelin and dendritic spine membranes of neurons. Plasma membranes can also form different types of "supramembrane" structures such as caveolae, postsynaptic density, podosome, invadopodium, desmosome, hemidesmosome, focal adhesion, and cell junctions. These types of membranes differ in lipid and protein composition.

Distinct types of membranes also create intracellular organelles: endosome; smooth and rough endoplasmic reticulum; sarcoplasmic reticulum; Golgi apparatus; lysosome; mitochondrion (inner and outer membranes); nucleus (inner and outer membranes); peroxisome; vacuole; cytoplasmic granules; cell vesicles (phagosome, autophagosome, clathrin-coated vesicles, COPI-coated and COPII-coated vesicles) and secretory vesicles (including synaptosome, acrosomes, melanosomes, and chromaffin granules). Different types of biological membranes have diverse lipid and protein compositions. The content of membranes defines their physical and biological properties. Some components of membranes play a key role in medicine, such as the efflux pumps that pump drugs out of a cell.

Fluidity

The hydrophobic core of the phospholipid bilayer is constantly in motion because of rotations around the bonds of lipid tails. Hydrophobic tails of a bilayer bend and lock

together. However, because of hydrogen bonding with water, the hydrophilic head groups exhibit less movement as their rotation and mobility are constrained. This results in increasing viscosity of the lipid bilayer closer to the hydrophilic heads.

Below a transition temperature, a lipid bilayer loses fluidity when the highly mobile lipids exhibits less movement becoming a gel-like solid. The transition temperature depends on such components of the lipid bilayer as the hydrocarbon chain length and the saturation of its fatty acids. Temperature-dependence fluidity constitutes an important physiological attribute for bacteria and cold-blooded organisms. These organisms maintain a constant fluidity by modifying membrane lipid fatty acid composition in accordance with differing temperatures.

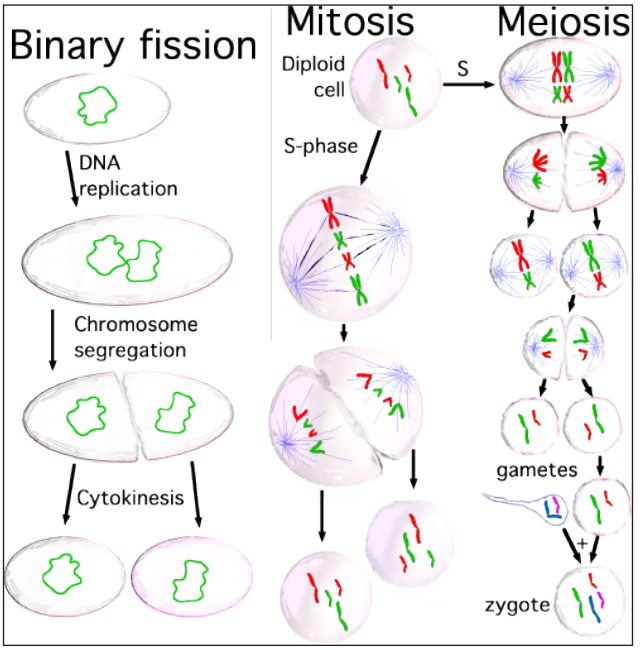
In animal cells, membrane fluidity is modulated by the inclusion of the sterol cholesterol. This molecule is present in especially large amounts in the plasma membrane, where it constitutes approximately 20% of the lipids in the membrane by weight. Because cholesterol molecules are short and rigid, they fill the spaces between neighboring phospholipid molecules left by the kinks in their unsaturated hydrocarbon tails. In this way, cholesterol tends to stiffen the bilayer, making it more rigid and less permeable.

For all cells, membrane fluidity is important for many reasons. It enables membrane proteins to diffuse rapidly in the plane of the bilayer and to interact with one another, as is crucial, for example, in cell signaling. It permits membrane lipids and proteins to diffuse from sites where they are inserted into the bilayer after their synthesis to other regions of the cell. It allows membranes to fuse with one another and mix their molecules, and it ensures that membrane molecules are distributed evenly between daughter cells when a cell divides. If biological membranes were not fluid, it is hard to imagine how cells could live, grow, and reproduce.

Cell Division

Cell division is the process by which a parent cell divides into two or more daughter cells. Cell division usually occurs as part of a larger cell cycle. In eukaryotes, there are two distinct types of cell division: a vegetative division, whereby each daughter cell is genetically identical to the parent cell (mitosis), and a reproductive cell division, whereby the number of chromosomes in the daughter cells is reduced by half to produce haploid gametes (meiosis). Meiosis results in four haploid daughter cells by undergoing one round of DNA replication followed by two divisions. Homologous chromosomes are separated in the first division, and sister chromatids are separated in the second division. Both of these cell division cycles are used in the process of sexual reproduction at some point in their life cycle. Both are believed to be present in the last eukaryotic common ancestor.

Prokaryotes (bacteria) undergo a vegetative cell division known as binary fission, where their genetic material is segregated equally into two daughter cells. While binary fission may be the means of division by most prokaryotes, there are alternative manners of division, such as budding, that have been observed. All cell divisions, regardless of organism, are preceded by a single round of DNA replication.



Three types of cell division.

For simple unicellular microorganisms such as the amoeba, one cell division is equivalent to reproduction – an entire new organism is created. On a larger scale, mitotic cell division can create progeny from multicellular organisms, such as plants that grow from cuttings. Mitotic cell division enables sexually reproducing organisms to develop from the one-celled zygote, which itself was produced by meiotic cell division from gametes. After growth, cell division by mitosis allows for continual construction and repair of the organism. The human body experiences about 10 quadrillion cell divisions in a lifetime.

The primary concern of cell division is the maintenance of the original cell's genome. Before division can occur, the genomic information that is stored in chromosomes must be replicated, and the duplicated genome must be separated cleanly between cells. A great deal of cellular infrastructure is involved in keeping genomic information consistent between generations.

Phases of Eukaryotic Cell Division

Interphase

Interphase is the process a cell must go through before mitosis, meiosis, and cytokinesis.

Interphase consists of three main phases: G_1, S, and G_2. G_1 is a time of growth for the cell where specialized cellular functions occur in order to prepare the cell for DNA Replication. There are checkpoints during interphase that allow the cell to be either advance or halt further development. In S phase, the chromosomes are replicated in order for the genetic content to be maintained. During G_2, the cell undergoes the final stages of growth before it enters the M phase, where spindles are synthesized. The M phase, can be either mitosis or meiosis depending on the type of cell. Germ cells, or gametes, undergo meiosis, while somatic cells will undergo mitosis. After the cell proceeds successfully through the M phase, it may then undergo cell division through cytokinesis. The control of each checkpoint is controlled by cyclin and cyclin-dependent kinases. The progression of interphase is the result of the increased amount of cyclin. As the amount of cyclin increases, more and more cyclin dependent kinases attach to cyclin signaling the cell further into interphase. At the peak of the cyclin, attached to the cyclin dependent kinases this system pushes the cell out of interphase and into the M phase, where mitosis, meiosis, and cytokinesis occur. There are three transition checkpoints the cell has to go through before entering the M phase. The most important being the G_1-S transition checkpoint. If the cell does not pass this checkpoint, it results in the cell exiting the cell cycle.

Prophase

Prophase is the first stage of division. The nuclear envelope is broken down, long strands of chromatin condense to form shorter more visible strands called chromosomes, the nucleolus disappears, and microtubules attach to the chromosomes at the kinetochores present in the centromere. Microtubules associated with the alignment and separation of chromosomes are referred to as the spindle and spindle fibers. Chromosomes will also be visible under a microscope and will be connected at the centromere. During this condensation and alignment period in meiosis, the homologous chromosomes undergo a break in their double-stranded DNA at the same locations, followed by a recombination of the now fragmented parental DNA strands into non-parental combinations, known as crossing over. This process is evidenced to be caused in a large part by the highly conserved Spo11 protein through a mechanism similar to that seen with topoisomerase in DNA replication and transcription.

Metaphase

In metaphase, the centromeres of the chromosomes convene themselves on the *metaphase plate* (or *equatorial plate*), an imaginary line that is equidistant from the two centrosome poles and held together by complex complexes known as cohesins. Chromosomes line up in the middle of the cell by microtubule organizing centers (MTOCs) pushing and pulling on centromeres of both chromatids thereby causing the chromosome to move to the center. At this point the chromosomes are still condensing and are currently one step away from being the most coiled and condensed they will be,

and the spindle fibers have already connected to the kinetochores. During this phase all the microtubules, with the exception of the kinetochores, are in a state of instability promoting their progression towards anaphase. At this point, the chromosomes are ready to split into opposite poles of the cell towards the spindle to which they are connected.

Anaphase

Anaphase is a very short stage of the cell cycle and occurs after the chromosomes align at the mitotic plate. Kinetochores emit anaphase-inhibition signals until their attachment to the mitotic spindle. Once the final chromosome is properly aligned and attached the final signal dissipates and triggers the abrupt shift to anaphase. This abrupt shift is caused by the activation of the anaphase-promoting complex and its function of tagging degradation of proteins important towards the metaphase-anaphase transition. One of these proteins that is broken down is securin which through its breakdown releases the enzyme separase that cleaves the cohesin rings holding together the sister chromatids thereby leading to the chromosomes separating. After the chromosomes line up in the middle of the cell, the spindle fibers will pull them apart. The chromosomes are split apart while the sister chromatids move to opposite sides of the cell. As the sister chromatids are being pulled apart, the cell and plasma are elongated by non-kinetochore microtubules.

Telophase

Telophase is the last stage of the cell cycle in which a cleavage furrow splits the cells cytoplasm (cytokinesis) and chromatin. This occurs through the synthesis of a new nuclear envelopes that forms around the chromatin gathered at each pole and the reformation of the nucleolus as the chromosomes decondense their chromatin back to the loose state it possessed during interphase. The division of the cellular contents is not always equal and can vary by cell type as seen with oocyte formation where one of the four daughter cells possess the majority of the cytoplasm.

Cytokinesis

The last stage of the cell division process is cytokinesis. In this stage there is a cytoplasmic division that occurs at the end or either mitosis or meiosis. At this stage there is a resulting irreversible separation leading to two daughter cells. Cell division plays a important role in determining the fate of the cell. This is due to there being the possibility of an asymmetric division. This as a result leads to cytokinesis producing unequal daughter cells containing completely different amounts or concentrations of fate-determining molecules.

Variants

Cells are broadly classified into two main categories: simple, non-nucleated prokaryotic cells, and complex, nucleated eukaryotic cells. Due to their structural differences,

eukaryotic and prokaryotic cells do not divide in the same way. Also, the pattern of cell division that transforms eukaryotic stem cells into gametes (sperm cells in males or egg cells in females), termed meiosis, is different from that of the division of somatic cells in the body. Image of the mitotic spindle in a human cell showing microtubules in green, chromosomes (DNA) in blue, and kinetochores in red.

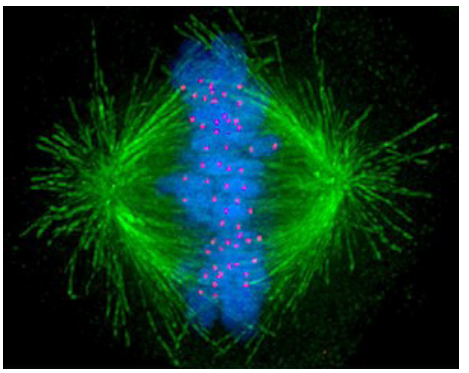


Image of the mitotic spindle in a human cell showing microtubules in green, chromosomes (DNA) in blue, and kinetochores in red.

Degradation

Multicellular organisms replace worn-out cells through cell division. In some animals, however, cell division eventually halts. In humans this occurs, on average, after 52 divisions, known as the Hayflick limit. The cell is then referred to as senescent. With each division the cells telomeres, protective sequences of DNA on the end of a chromosome that prevent degradation of the chromosomal DNA, shorten. This shortening has been correlated to negative effects such as age related diseases and shortened lifespans in humans. Cancer cells, on the other hand, are not thought to degrade in this way, if at all. An enzyme complex called telomerase, present in large quantities in cancerous cells, rebuilds the telomeres through synthesis of telomeric DNA repeats, allowing division to continue indefinitely.

Cell Migration

Cell migration is a central process in the development and maintenance of multicellular organisms. Tissue formation during embryonic development, wound healing and immune responses all require the orchestrated movement of cells in particular directions to specific locations. Cells often migrate in response to specific external signals, including chemical signals and mechanical signals. Errors during this process have serious consequences, including intellectual disability, vascular disease, tumor formation and metastasis. An understanding of the mechanism by which cells migrate may lead to the development of novel therapeutic strategies for controlling, for example, invasive tumour cells. The cells are complex and have organelles.

Due to the highly viscous environment (low Reynolds number), cells need to permanently produce forces in order to move. Cells achieve active movement by very different mechanisms. Many less complex prokaryotic organisms (and sperm cells) use flagella or cilia to propel themselves. Eukaryotic cell migration typically is far more complex and can consist of combinations of different migration mechanisms. It generally involves drastic changes in cell shape which are driven by the cytoskeleton. Two very distinct migration scenarios are crawling motion (most commonly studied) and blebbing motility. A paradigmatic example of crawling motion is the case of fish epidermal keratocytes, which have been extensively used in research and teaching.

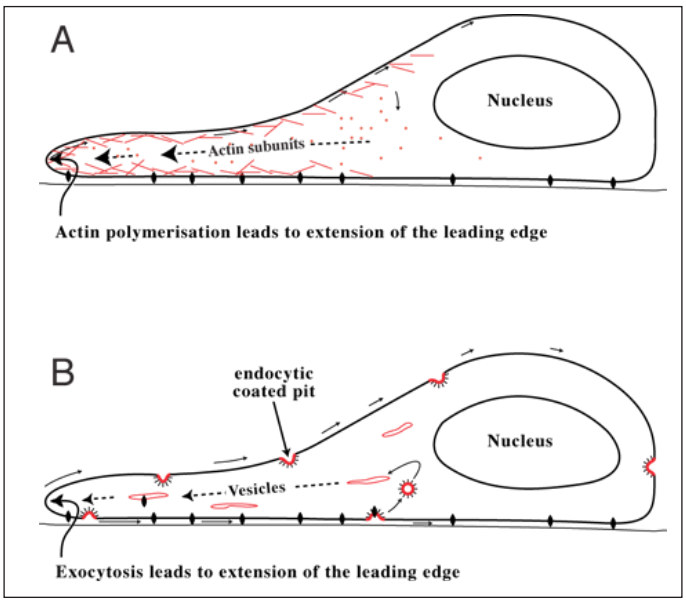
Cell Migration Studies

The migration of cultured cells attached to a surface is commonly studied using microscopy. As cell movement is very slow, a few μm/minute, time-lapse microscopy videos are recorded of the migrating cells to speed up the movement. Such videos reveal that the leading cell front is very active, with a characteristic behavior of successive contractions and expansions. It is generally accepted that the leading front is the main motor that pulls the cell forward.

Common Features

The processes underlying mammalian cell migration are believed to be consistent with those of (non-spermatozooic) locomotion. Observations in common include:

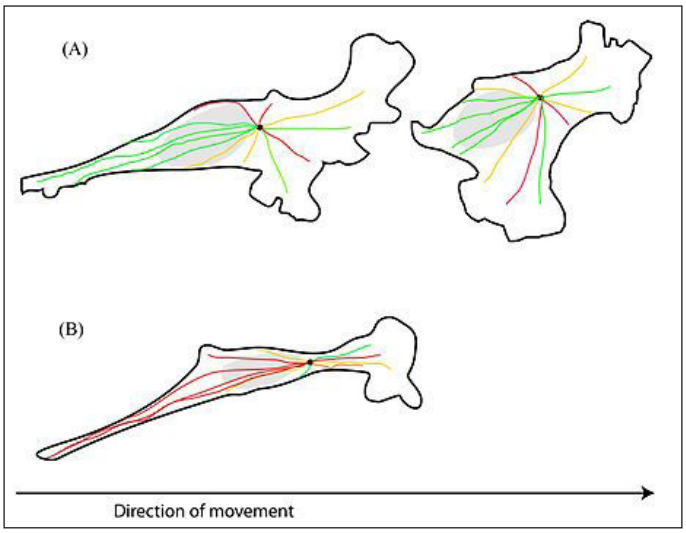
- Cytoplasmic displacement at leading edge (front).
- Laminar removal of dorsally-accumulated debris toward trailing edge (back).



Two different models for how cells move. A) Cytoskeletal model. B) Membrane Flow Model.

The latter feature is most easily observed when aggregates of a surface molecule are cross-linked with a fluorescent antibody or when small beads become artificially bound to the front of the cell.

Other eukaryotic cells are observed to migrate similarly. The amoeba Dictyostelium discoideum is useful to researchers because they consistently exhibit chemotaxis in response to cyclic AMP; they move more quickly than cultured mammalian cells; and they have a haploid genome that simplifies the process of connecting a particular gene product with its effect on cellular behaviour.



(A) Dynamic microtubules are necessary for tail retraction and are distributed at the rear end in a migrating cell. Green, highly dynamic microtubules; yellow, moderately dynamic microtubules and red, stable microtubules. (B) Stable microtubules act as struts and prevent tail retraction and thereby inhibit cell migration.

Molecular Processes of Migration

There are two main theories for how the cell advances its front edge: the cytoskeletal model and membrane flow model. It is possible that both underlying processes contribute to cell extension.

Cytoskeletal Model

Leading Edge

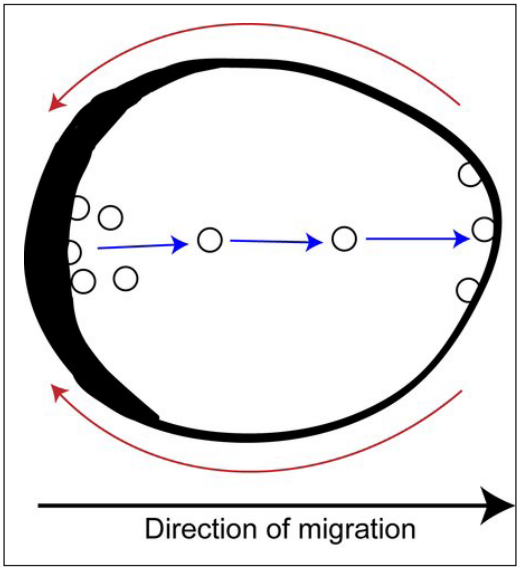
Experimentation has shown that there is rapid actin polymerisation at the cell's front edge. This observation has led to the hypothesis that formation of actin filaments "push" the leading edge forward and is the main motile force for advancing the cell's front edge. In addition, cytoskeletal elements are able to interact extensively and intimately with a cell's plasma membrane.

Trailing Edge

Other cytoskeletal components (like microtubules) have important functions in cell migration. It has been found that microtubules act as "struts" that counteract the contractile forces that are needed for trailing edge retraction during cell movement. When microtubules in the trailing edge of cell are dynamic, they are able to remodel to allow retraction. When dynamics are suppressed, microtubules cannot remodel and, therefore, oppose the contractile forces. The morphology of cells with suppressed microtubule dynamics indicate that cells can extend the front edge (polarized in the direction of movement), but have difficulty retracting their trailing edge. On the other hand, high drug concentrations, or microtubule mutations that depolymerize the microtubules, can restore cell migration but there is a loss of directionality. It can be concluded that microtubules act both to restrain cell movement and to establish directionality.

Membrane Flow Model

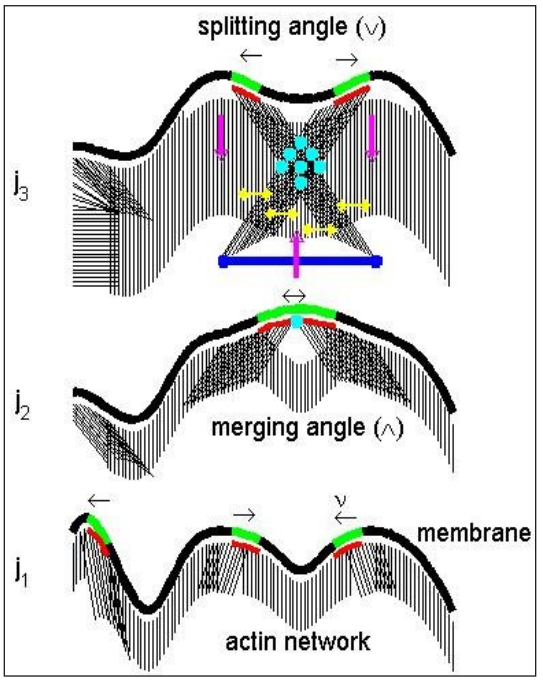
Studies have also shown that the front of the migration is the site at which the membrane is returned to the cell surface from internal membrane pools at the end of the endocytic cycle. This has led to the hypothesis that extension of the leading edge occurs primarily by addition of membrane at the front of the cell. If so, the actin filaments that form at the front might stabilize the added membrane so that a structured extension, or lamella, is formed rather than a bubble-like structure (or bleb) at its front. For a cell to move, it is necessary to bring a fresh supply of "feet" (proteins called integrins, which attach a cell to the surface on which it is crawling) to the front. It is likely that these feet are endocytosed toward the rear of the cell and brought to the cell's front by exocytosis, to be reused to form new attachments to the substrate.



Rearward membrane flow (red arrows) and vesicle trafficking from back to front (blue arrows) drive adhesion-independent migration.

Mechanistic basis of Amoeboid Migration

Adhesive crawling is not the only migration mode exhibited by eukaryotic cells. Importantly, metastatic cancer cells and immune cells like macrophages and neutrophils have been found to be capable of adhesion-independent migration. The mechanistic basis of this migration mode is less understood than either eukaryotic cell crawling or flagella-based swimming by microorganisms. The physicist E. M. Purcell theorized that under conditions of low Reynolds number fluid dynamics, which apply at the cellular scale, rearward surface flow could provide a mechanism for microscopic objects to swim forward. After some decades, experimental support for this model was provided using optogenetics. It was shown that cells migrating in an amoeboid fashion without adhesions exhibit plasma membrane flow towards the cell rear that can propel cells by exerting tangential forces on the surrounding fluid. Polarized trafficking of membrane-containing vesicles from the rear to the front of the cell helps maintain cell size. Rearward membrane flow was also observed in *Dictyostelium discoideum* cells. Interestingly, the migration of supracellular clusters has also been found to be supported by a similar mechanism of rearward surface flow.



Schematic representation of the collective biomechanical and molecular mechanism of cell motion.

Collective Biomechanical and Molecular Mechanism of Cell Motion

Based on some mathematical models, recent studies hypothesize a novel biological model for collective biomechanical and molecular mechanism of cell motion. It is proposed that microdomains weave the texture of cytoskeleton and their interactions mark

the location for formation of new adhesion sites. According to this model, microdomain signaling dynamics organizes cytoskeleton and its interaction with substratum. As microdomains trigger and maintain active polymerization of actin filaments, their propagation and zigzagging motion on the membrane generate a highly interlinked network of curved or linear filaments oriented at a wide spectrum of angles to the cell boundary. It is also proposed that microdomain interaction marks the formation of new focal adhesion sites at the cell periphery. Myosin interaction with the actin network then generate membrane retraction/ruffling, retrograde flow, and contractile forces for forward motion. Finally, continuous application of stress on the old focal adhesion sites could result in the calcium-induced calpain activation, and consequently the detachment of focal adhesions which completes the cycle.

Polarity in Migrating Cells

Migrating cells have a polarity—a front and a back. Without it, they would move in all directions at once, i.e. spread. How this polarity is formulated at a molecular level inside a cell is unknown. In a cell that is meandering in a random way, the front can easily give way to become passive as some other region, or regions, of the cell forms a new front. In chemotaxing cells, the stability of the front appears enhanced as the cell advances toward a higher concentration of the stimulating chemical. This polarity is reflected at a molecular level by a restriction of certain molecules to particular regions of the inner cell surface. Thus, the phospholipid PIP3 and activated Rac and CDC42 are found at the front of the cell, whereas Rho GTPase and PTEN are found toward the rear.

It is believed that filamentous actins and microtubules are important for establishing and maintaining a cell's polarity. Drugs that destroy actin filaments have multiple and complex effects, reflecting the wide role that these filaments play in many cell processes. It may be that, as part of the locomotory process, membrane vesicles are transported along these filaments to the cell's front. In chemotaxing cells, the increased persistence of migration toward the target may result from an increased stability of the arrangement of the filamentous structures inside the cell and determine its polarity. In turn, these filamentous structures may be arranged inside the cell according to how molecules like PIP3 and PTEN are arranged on the inner cell membrane. And where these are located appears in turn to be determined by the chemoattractant signals as these impinge on specific receptors on the cell's outer surface.

Although microtubules have been known to influence cell migration for many years, the mechanism by which they do so has remained controversial. On a planar surface, microtubules are not needed for the movement, but they are required to provide directionality to cell movement and efficient protrusion of the leading edge. When present, microtubules retard cell movement when their dynamics are suppressed by drug treatment or by tubulin mutations.

Inverse Problems in the Context of Cell Motility

An area of research called inverse problems in cell motility has been established. This approach is based on the idea that behavioral or shape changes of a cell bear information about the underlying mechanisms that generate these changes. Reading cell motion, namely, understanding the underlying biophysical and mechanochemical processes, is of paramount importance. The mathematical models developed in these works determine some physical features and material properties of the cells locally through analysis of live cell image sequences and uses this information to make further inferences about the molecular structures, dynamics, and processes within the cells, such as the actin network, microdomains, chemotaxis, adhesion, and retrograde flow.

Cell Signalling

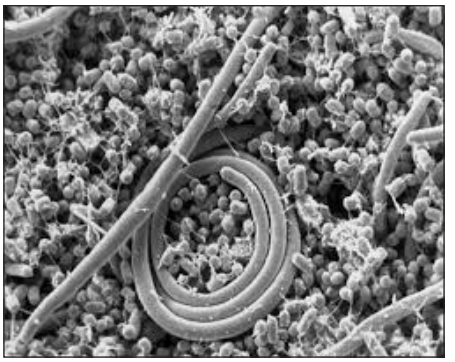
In biology, cell signaling is part of any communication process that governs basic activities of cells and coordinates multiple-cell actions. The ability of cells to perceive and correctly respond to their microenvironment is the basis of development, tissue repair, and immunity, as well as normal tissue homeostasis. Errors in signaling interactions and cellular information processing may cause diseases such as cancer, autoimmunity, and diabetes. By understanding cell signaling, clinicians may treat diseases more effectively and, theoretically, researchers may develop artificial tissues.

Systems biology studies the underlying structure of cell-signaling networks and how changes in these networks may affect the transmission and flow of information (signal transduction). Such networks are complex systems in their organization and may exhibit a number of emergent properties, including bistability and ultrasensitivity. Analysis of cell-signaling networks requires a combination of experimental and theoretical approaches, including the development and analysis of simulations and modeling. Long-range allostery is often a significant component of cell-signaling events.

All cells receive and respond to signals from their surroundings. This is accomplished by a variety of signal molecules that are secreted or expressed on the surface of one cell and bind to a receptor expressed by the other cells, thereby integrating and coordinating the function of the many individual cells that make up organisms. Each cell is programmed to respond to specific extracellular signal molecules. Extracellular signaling usually entails the following steps:

- Synthesis and release of the signaling molecule by the signaling cell.
- Transport of the signal to the target cell.
- Binding of the signal by a specific receptor leading to its activation.
- Initiation of signal-transduction pathways.

Between Organisms



Example of signaling between bacteria. *Salmonella enteritidis* uses N-Acyl homoserine lactone for Quorum sensing.

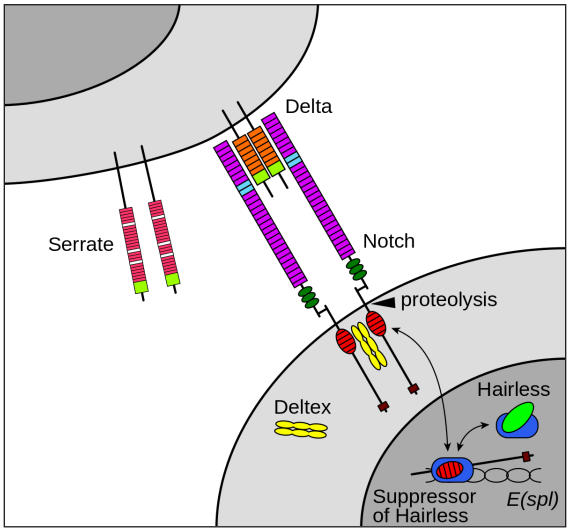
Cell signaling has been most extensively studied in the context of human diseases and signaling between cells of a single organism. However, cell signaling may also occur between the cells of two different organisms. In many mammals, early embryo cells exchange signals with cells of the uterus. In the human gastrointestinal tract, bacteria exchange signals with each other and with human epithelial and immune system cells. For the yeast *Saccharomyces cerevisiae* during mating, some cells send a peptide signal (mating factor pheromones) into their environment. The mating factor peptide may bind to a cell surface receptor on other yeast cells and induce them to prepare for mating.

Classification

Cell signaling can be classified as either mechanical or biochemical based on the type of the signal. Mechanical signals are the forces exerted on the cell and the forces produced by the cell. These forces can both be sensed and responded to by the cells. Biochemical signals are biochemical molecules such as proteins, lipids, ions, and gases. These signals can be categorized based on the distance between signaling and responder cells. Signaling within, between, and amongst cells is subdivided into the following classifications:

- Intracrine signals are produced by the target cell that stay within the target cell.
- Autocrine signals are produced by the target cell, are secreted, and affect the target cell itself via receptors. Sometimes autocrine cells can target cells close by if they are the same type of cell as the emitting cell. An example of this are immune cells.
- Juxtacrine signals target adjacent (touching) cells. These signals are transmitted along cell membranes via protein or lipid components integral to the membrane and are capable of affecting either the emitting cell or cells immediately adjacent.

- Paracrine signals target cells in the vicinity of the emitting cell. Neurotransmitters represent an example.
- Endocrine signals target distant cells. Endocrine cells produce hormones that travel through the blood to reach all parts of the body.



Notch-mediated juxtacrine signal between adjacent cells.

Cells communicate with each other via direct contact (juxtacrine signaling), over short distances (paracrine signaling), or over large distances and/or scales (endocrine signaling).

Some cell–cell communication requires direct cell–cell contact. Some cells can form gap junctions that connect their cytoplasm to the cytoplasm of adjacent cells. In cardiac muscle, gap junctions between adjacent cells allow for action potential propagation from the cardiac pacemaker region of the heart to spread and coordinate the contraction of the heart.

The notch signaling mechanism is an example of juxtacrine signaling (also known as contact-dependent signaling) in which two adjacent cells must make physical contact in order to communicate. This requirement for direct contact allows for very precise control of cell differentiation during embryonic development. In the worm *Caenorhabditis elegans*, two cells of the developing gonad each have an equal chance of terminally differentiating or becoming a uterine precursor cell that continues to divide. The choice of which cell continues to divide is controlled by competition of cell surface signals. One cell will happen to produce more of a cell surface protein that activates the Notch receptor on the adjacent cell. This activates a feedback loop or system that reduces Notch expression in the cell that will differentiate and that increases Notch on the surface of the cell that continues as a stem cell.

Many cell signals are carried by molecules that are released by one cell and move to make contact with another cell. *Endocrine* signals are called hormones. Hormones are

produced by endocrine cells and they travel through the blood to reach all parts of the body. Specificity of signaling can be controlled if only some cells can respond to a particular hormone. *Paracrine* signals such as retinoic acid target only cells in the vicinity of the emitting cell. Neurotransmitters represent another example of a paracrine signal. Some signaling molecules can function as both a hormone and a neurotransmitter. For example, epinephrine and norepinephrine can function as hormones when released from the adrenal gland and are transported to the heart by way of the blood stream. Norepinephrine can also be produced by neurons to function as a neurotransmitter within the brain. Estrogen can be released by the ovary and function as a hormone or act locally via paracrine or autocrine signaling. Active species of oxygen and nitric oxide can also act as cellular messengers. This process is dubbed redox signaling.

Multicellular Organisms

In a multicellular organism, signaling between cells occurs either through release into the extracellular space, divided in paracrine signaling (over short distances) and endocrine signaling (over long distances), or by direct contact, known as juxtacrine signaling. Autocrine signaling is a special case of paracrine signaling where the secreting cell has the ability to respond to the secreted signaling molecule. Synaptic signaling is a special case of paracrine signaling (for chemical synapses) or juxtacrine signaling (for electrical synapses) between neurons and target cells. Signaling molecules interact with a target cell as a ligand to cell surface receptors, and/or by entering into the cell through its membrane or endocytosis for intracrine signaling. This generally results in the activation of second messengers, leading to various physiological effects.

A particular molecule is generally used in diverse modes of signaling, and therefore a classification by mode of signaling is not possible. At least three important classes of signaling molecules are widely recognized, although non-exhaustive and with imprecise boundaries, as such membership is non-exclusive and depends on the context:

- Hormones are the major signaling molecules of the endocrine system, though they often regulate each other's secretion via local signaling (e.g. islet of Langerhans cells), and most are also expressed in tissues for local purposes (e.g. angiotensin) or failing that, structurally related molecules are (e.g. PTHrP).
- Neurotransmitters are signaling molecules of the nervous system, also including neuropeptides and neuromodulators. Neurotransmitters like the catecholamines are also secreted by the endocrine system into the systemic circulation.
- Cytokines are signaling molecules of the immune system, with a primary paracrine or juxtacrine role, though they can during significant immune responses have a strong presence in the circulation, with systemic effect (altering iron metabolism or body temperature). Growth factors can be considered as cytokines or a different class.

Signaling molecules can belong to several chemical classes: lipids, phospholipids, amino acids, monoamines, proteins, glycoproteins, or gases. Signaling molecules binding surface receptors are generally large and hydrophilic (e.g. TRH, Vasopressin, Acetylcholine), while those entering the cell are generally small and hydrophobic (e.g. glucocorticoids, thyroid hormones, cholecalciferol, retinoic acid), but important exceptions to both are numerous, and a same molecule can act both via surface receptor or in an intracrine manner to different effects. In intracrine signaling, once inside the cell, a signaling molecule can bind to intracellular receptors, other elements, or stimulate enzyme activity (e.g. gasses). The intracrine action of peptide hormones remains a subject of debate.

Hydrogen sulfide is produced in small amounts by some cells of the human body and has a number of biological signaling functions. Only two other such gases are currently known to act as signaling molecules in the human body: nitric oxide and carbon monoxide.

Signaling Receptors

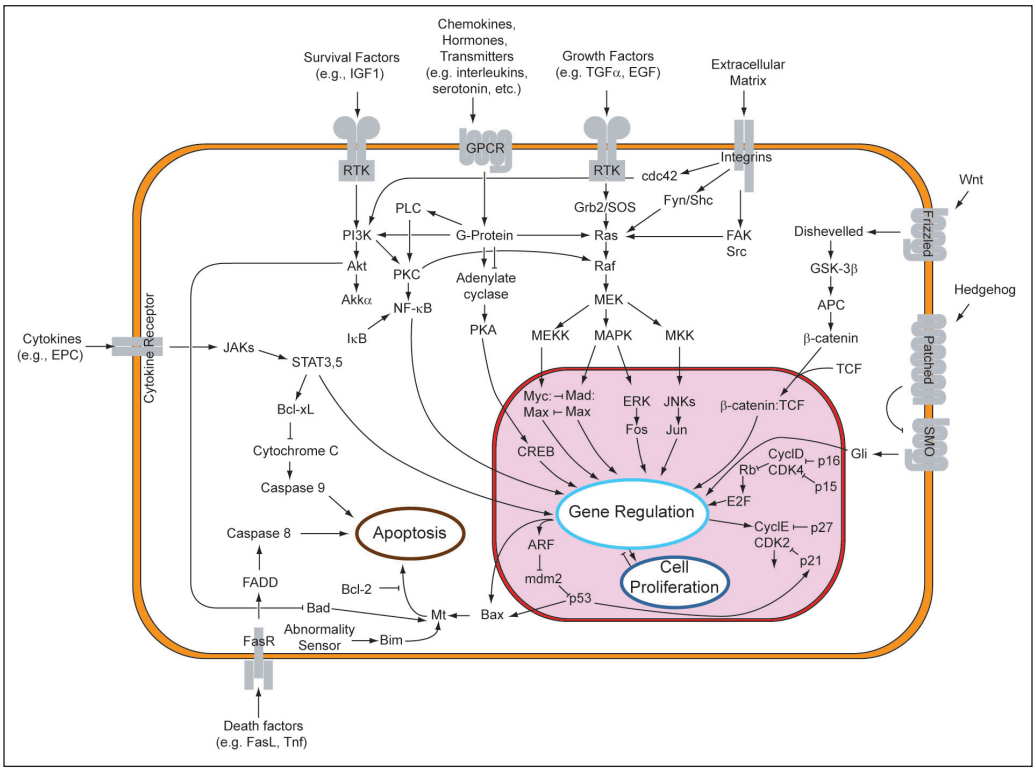
Cells receive information from their neighbors through a class of proteins known as receptors. Notch is a cell surface protein that functions as a receptor. Animals have a small set of genes that code for signaling proteins that interact specifically with Notch receptors and stimulate a response in cells that express Notch on their surface. Molecules that activate (or, in some cases, inhibit) receptors can be classified as hormones, neurotransmitters, cytokines, and growth factors, in general called receptor ligands. Ligand receptor interactions such as that of the Notch receptor interaction, are known to be the main interactions responsible for cell signaling mechanisms and communication.

Notch acts as a receptor for ligands that are expressed on adjacent cells. While some receptors are cell-surface proteins, others are found inside cells. For example, estrogen is a hydrophobic molecule that can pass through the lipid bilayer of the membranes. As part of the endocrine system, intracellular estrogen receptors from a variety of cell types can be activated by estrogen produced in the ovaries.

A number of transmembrane receptors for small molecules and peptide hormones, as well as intracellular receptors for steroid hormones exist, giving cells the ability to respond to a great number of hormonal and pharmacological stimuli. In diseases, often, proteins that interact with receptors are aberrantly activated, resulting in constitutively activated downstream signals.

For several types of intercellular signaling molecules that are unable to permeate the hydrophobic cell membrane due to their hydrophilic nature, the target receptor is expressed on the membrane. When such a signaling molecule activates its receptor, the signal is carried into the cell usually by means of a second messenger such as cAMP.

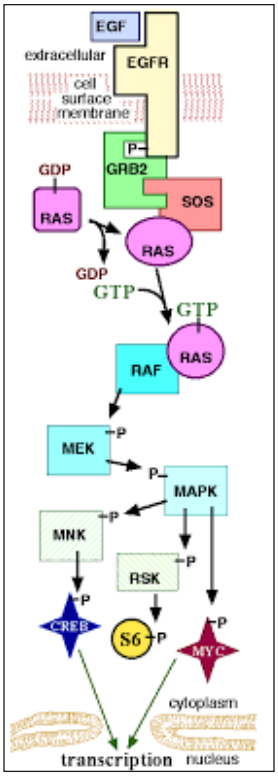
Signaling Pathways



Overview of signal transduction pathways.

In some cases, receptor activation caused by ligand binding to a receptor is directly coupled to the cell's response to the ligand. For example, the neurotransmitter GABA can activate a cell surface receptor that is part of an ion channel. GABA binding to a $GABA_A$ receptor on a neuron opens a chloride-selective ion channel that is part of the receptor. $GABA_A$ receptor activation allows negatively charged chloride ions to move into the neuron, which inhibits the ability of the neuron to produce action potentials. However, for many cell surface receptors, ligand-receptor interactions are not directly linked to the cell's response. The activated receptor must first interact with other proteins inside the cell before the ultimate physiological effect of the ligand on the cell's behavior is produced. Often, the behavior of a chain of several interacting cell proteins is altered following receptor activation. The entire set of cell changes induced by receptor activation is called a signal transduction mechanism or pathway.

In the case of Notch-mediated signaling, the signal transduction mechanism can be relatively simple. As shown in figure, the activation of Notch can cause the Notch protein to be altered by a protease. Part of the Notch protein is released from the cell surface membrane and takes part in gene regulation. Cell signaling research involves studying the spatial and temporal dynamics of both receptors and the components of signaling pathways that are activated by receptors in various cell types.



Key components of a signal transduction pathway.

A more complex signal transduction pathway is shown in figure. This pathway involves changes of protein–protein interactions inside the cell, induced by an external signal. Many growth factors bind to receptors at the cell surface and stimulate cells to progress through the cell cycle and divide. Several of these receptors are kinases that start to phosphorylate themselves and other proteins when binding to a ligand. This phosphorylation can generate a binding site for a different protein and thus induce protein–protein interaction. In figure, the ligand (called epidermal growth factor (EGF)) binds to the receptor (called EGFR). This activates the receptor to phosphorylate itself. The phosphorylated receptor binds to an adaptor protein (GRB2), which couples the signal to further downstream signaling processes. For example, one of the signal transduction pathways that are activated is called the mitogen-activated protein kinase (MAPK) pathway. The signal transduction component labeled as “MAPK” in the pathway was originally called “ERK,” so the pathway is called the MAPK/ERK pathway. The MAPK protein is an enzyme, a protein kinase that can attach phosphate to target proteins such as the transcription factor MYC and, thus, alter gene transcription and, ultimately, cell cycle progression. Many cellular proteins are activated downstream of the growth factor receptors (such as EGFR) that initiate this signal transduction pathway.

Some signaling transduction pathways respond differently, depending on the amount of signaling received by the cell. For instance, the hedgehog protein activates different

genes, depending on the amount of hedgehog protein present. Complex multi-component signal transduction pathways provide opportunities for feedback, signal amplification, and interactions inside one cell between multiple signals and signaling pathways.

Intra- and Inter-species Signaling

Molecular signaling can occur between different organisms, whether unicellular or multicellular. The emitting organism produces the signaling molecule, secretes it into the environment, where it diffuses, and it is sensed or internalized by the receiving organism. In some cases of interspecies signaling, the emitting organism can actually be a host of the receiving organism, or vice versa.

Intraspecies signaling occurs especially in bacteria, yeast, social insects, but also many vertebrates. The signaling molecules used by multicellular organisms are often called pheromones. They can have such purposes as alerting against danger, indicating food supply, or assisting in reproduction. In unicellular organisms such as bacteria, signaling can be used to 'activate' peers from a dormant state, enhance virulence, defend against bacteriophages, etc. In quorum sensing, which is also found in social insects, the multiplicity of individual signals has the potentiality to create a positive feedback loop, generating coordinated response. In this context, the signaling molecules are called autoinducers. This signaling mechanism may have been involved in evolution from unicellular to multicellular organisms. Bacteria also use contact-dependent signaling, notably to limit their growth.

Molecular signaling can also occur between individuals of different species. This has been particularly studied in bacteria. Different bacterial species can coordinate to colonize a host and participate in common quorum sensing. Therapeutic strategies to disrupt this phenomenon are being investigated. Interactions mediated through signaling molecules are also thought to occur between the gut flora and their host, as part of their commensal or symbiotic relationship. Gram negative microbes deploy bacterial outer membrane vesicles for intra- and inter-species signaling in natural environments and at the host-pathogen interface. Additionally, interspecies signaling occurs between multicellular organisms. In *Vespa mandarinia*, individuals release a scent that directs the colony to a food source.

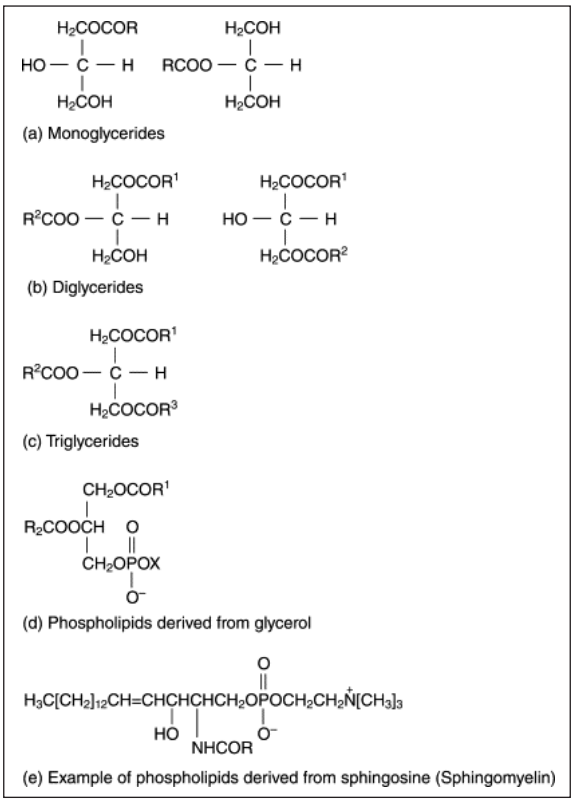
Computational Models

Recent approaches to better understand elements of pathway crosstalk, complex ligand-receptor binding, and signaling network dynamics have been aided by the use of systems biology approaches. Computational models often take aim at compiling information from published literature to generate a coherent set of signaling components and their associated interactions. The development of computational models allows for a more in-depth probing of cell signaling pathways at a global level by manipulating different variables and systemically evaluating the resulting response. The use of analytical models for the study of signal transduction has been heavily applied in the

fields of pharmacology and drug discovery to assess receptor-ligand interactions and pharmacokinetics as well as the flow of metabolites in large networks. A commonly applied strategy to model cell signaling mechanisms is through the use of ordinary differential equation (ODE) models by expressing the time-dependent concentration of a signaling molecule as a function of other molecules downstream and/or upstream within the pathway. ODE models have already been applied for dynamic analysis of the Mitogen-activated protein kinase, Estrogen receptor alpha, and MTOR signaling pathways among numerous others.

Phospholipids

Phospholipids are a major component of the lipid bilayer portion of biological membranes and 31P NMR has provided much detail regarding membrane dynamics and morphology.

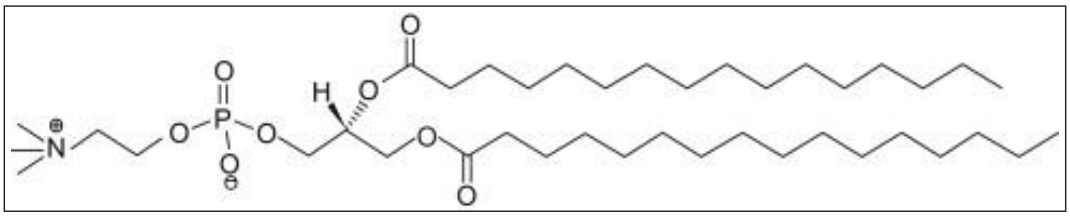


The representative structure of common lipids: (a) Monoglycerides (b) diglycerides, (c) triglycerides, (d) phospholipids derived from glycerol, (e) example of phospholipids derived from sphingosine (Sphingomyelin).

Phospholipids are mostly made from glycerides by substituting one of the three fatty acids by a phosphate group with some other molecule attached to its end. The

other form of phospholipids is sphingomyelin, which is derived from sphingosine instead of glycerol. Phospholipids are soluble in both water and oil (amphiphilic) because the hydrocarbon tails of two fatty acids are still hydrophobic, but the phosphate group end is hydrophilic. Phospholipids are the major component of cell membrane to form lipid bilayers. Figure shows the representative structure of common lipids.

Phospholipids are amphiphilic (amphipathic) molecules having polar (hydrophilic) headgroups and nonpolar (hydrophobic) hydrocarbon chains. Phospholipids are diverse molecules with different headgroups, and different lengths and degrees of saturation of hydrocarbon chains. Phospholipids can be classified as synthetic, such as dimyristoylphosphatidylcholine and dipalmitoylphosphatidylcholine (DPPC), and natural, such as egg (EPC) or soya (SPC) phosphatidylcholine. Natural phospholipids comprise a mixture of phosphatidylcholines having chains of different lengths and degrees of unsaturation.



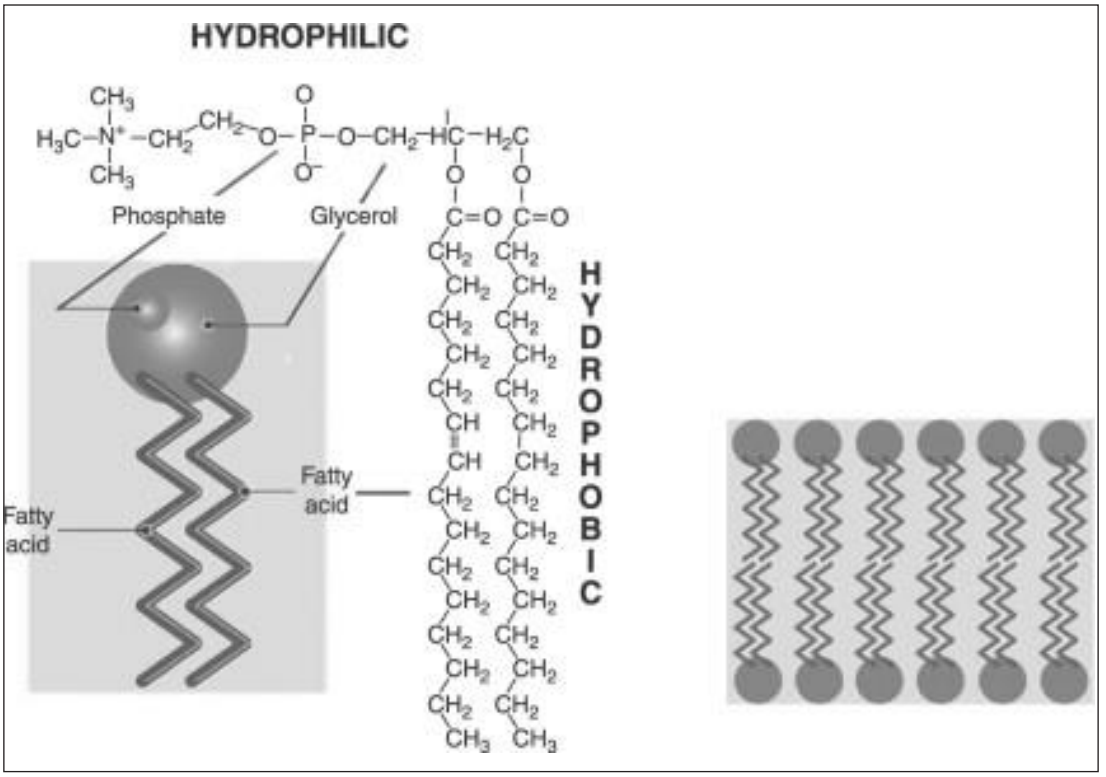
Chemical structure of the synthetic phospholipid DPPC.

Amphiphilic molecules undergo a geometry-dependent rearrangement when aqueous phase is added, in order to mask the hydrophobic moieties from the polar aqueous environment. If an amphiphilic molecule has a cylindrical geometry (i.e., the cross-sectional area of polar and nonpolar moieties are approximately equal), adjacent molecules align parallel to each other to form a monolayer, which in turn arranges symmetrically with another monolayer to form a bilayer sheet called the lamella or lamellar (cubic) phase.

Lasic has suggested a mechanism which describes the formation of liposomes from phospholipid molecules. Briefly, when aqueous phase is added to a thin film of phospholipid, the hydration of the outer monolayer predominates compared to the inner layers, resulting in the expansion of the polar headgroups of phospholipid molecules and the formation of “blisters”. Aqueous phase penetration through these “blisters” results in the formation of phospholipid bilayers, which develop into tubular fibrils that in turn increase the surface area in contact with the aqueous phase. The bilayer sheets will have hydrophobic moieties of the amphiphileexposed to the aqueous phase. This is thermodynamically unfavorable, which causes the bilayers to round off and form multilamellar liposomes. However, it is essential to bear in mind that for liposomes to form, the hydration procedure must be undertaken at a temperature significantly higher than that of the phase transition (T_m) temperature of the phospholipid(s) selected. This is the temperature at which the

phospholipid membrane passes from a tightly ordered gel ($L_{\beta'}$) phase to the less ordered liquid-crystalline (L_{α}) phase, where the flexibility of phospholipid molecules is higher.

Phospholipids play a major role in cell membranes. They generally consist of diglycerides, a phosphate group (a molecule of phosphoric acid) and an organic molecule (choline). A diglyceride is a glyceride composed of two fatty acid chains that are covalently bound to a single glycerol molecule via an ester linkage. Glycerol ($C_3H_8O_3$), containing three hydroxyl groups (–OH), is responsible for the solubility of phospholipid molecules in water. Glycerol acts as a backbone by its attachment to both fatty acids (hydrocarbon) chains and a phosphate group. Fatty acid molecules are either saturated or unsaturated and are insoluble in water. Thus, the phospholipid molecule is composed of a hydrophobic moiety comprising two fatty acid chains, and a hydrophilic head made of glycerol and phosphate. The bilayer structure is formed when the fatty acid moiety of one layer faces the fatty acid moiety of another layer and the head groups face the water.



Structure of phospholipid and bilayer.

Phospholipids are either natural or synthetic. Naturally occurring phospholipids include soya phosphatidylcholine (SPC) and egg phosphatidylcholine (EPC), while synthetic phospholipids include dipalmitoyl phosphatidylcholine (DPPC) and dimyrestoyl phosphatidylcholine (DMPC).

References

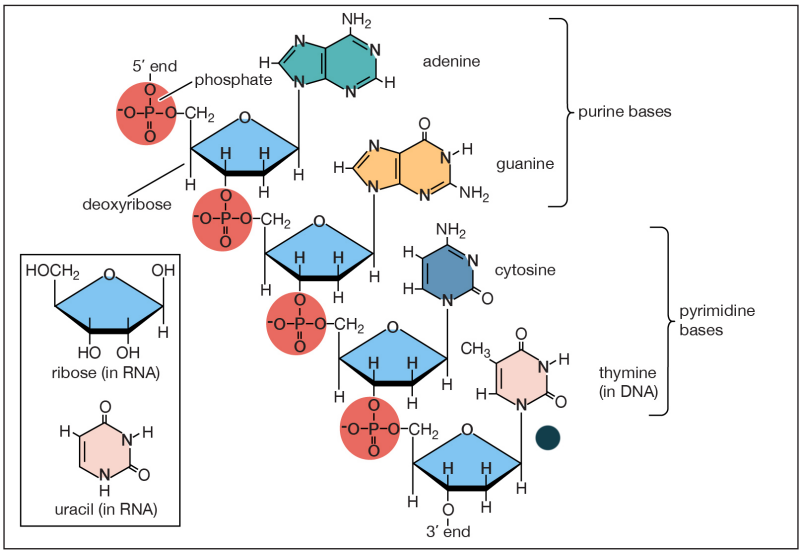
- Maxwell Myer W (2004). Greer JG, Foerster J, Lukens JN, Rodgers GM, Paraskevas F (eds.). Wintrobe's clinical hematology (11 ed.). Hagerstown, MD: Lippincott Williams & Wilkins. pp. 453–456. ISBN 978-0-7817-3650-3
- Enzyme-activity, enzyme-kinetics, basics: teachmephysiology.com, Retrieved 18 July, 2020
- Cleland WW (January 2005). "The use of isotope effects to determine enzyme mechanisms". Archives of Biochemistry and Biophysics. 433 (1): 2–12. doi:10.1016/j.abb.2004.08.027. PMID 15581561
- Phospholipid, engineering: sciencedirect.com, Retrieved 20 July, 2020
- Clark BF (October 2006). "The crystal structure of tRNA" (PDF). Journal of Biosciences. 31 (4): 453–7. doi:10.1007/BF02705184. PMID 17206065

3
Molecular Biophysics

A biomolecule refers to the ions and molecules present in organisms essential for biological processes. Molecular biophysics deals with the study of biomolecular systems and functions such as molecular structure and dynamic behavior structure of biomolecules. This chapter discusses the subject of molecular biophysics in detail.

Biomolecules

Biomolecule, also called biological molecules are any of numerous substances that are produced by cells and living organisms. Biomolecules have a wide range of sizes and structures and perform a vast array of functions. The four major types of biomolecules are carbohydrates, lipids, nucleic acids, and proteins.

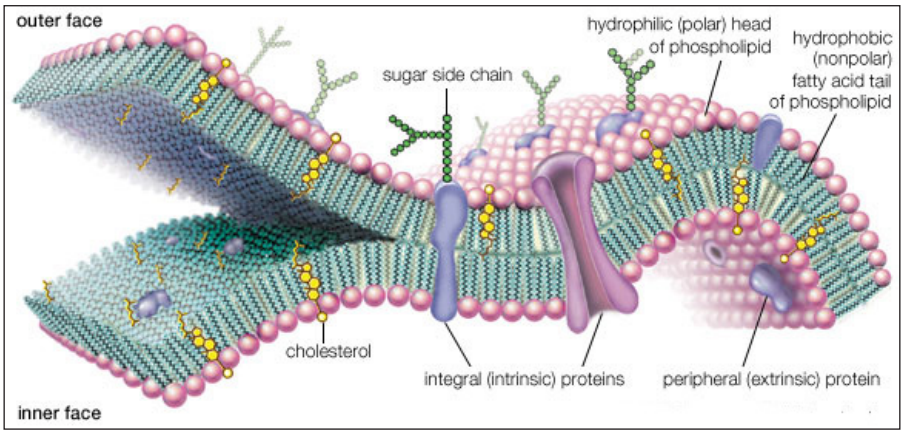


Portion of polynucleotide chain of deoxyribonucleic acid (DNA). The inset shows the corresponding pentose sugar and pyrimidine base in ribonucleic acid (RNA).

Among biomolecules, nucleic acids, namely DNA and RNA, have the unique function of storing an organism's genetic code—the sequence of nucleotides that determines

the amino acid sequence of proteins, which are of critical importance to life on Earth. There are 20 different amino acids that can occur within a protein; the order in which they occur plays a fundamental role in determining protein structure and function. Proteins themselves are major structural elements of cells. They also serve as transporters, moving nutrients and other molecules in and out of cells, and as enzymes and catalysts for the vast majority of chemical reactions that take place in living organisms. Proteins also form antibodies and hormones, and they influence gene activity.

Likewise, carbohydrates, which are made up primarily of molecules containing atoms of carbon, hydrogen, and oxygen, are essential energy sources and structural components of all life, and they are among the most abundant biomolecules on Earth. They are built from four types of sugar units—monosaccharides, disaccharides, oligosaccharides, and polysaccharides. Lipids, another key biomolecule of living organisms, fulfill a variety of roles, including serving as a source of stored energy and acting as chemical messengers. They also form membranes, which separate cells from their environments and compartmentalize the cell interior, giving rise to organelles, such as the nucleus and the mitochondrion, in higher (more complex) organisms.

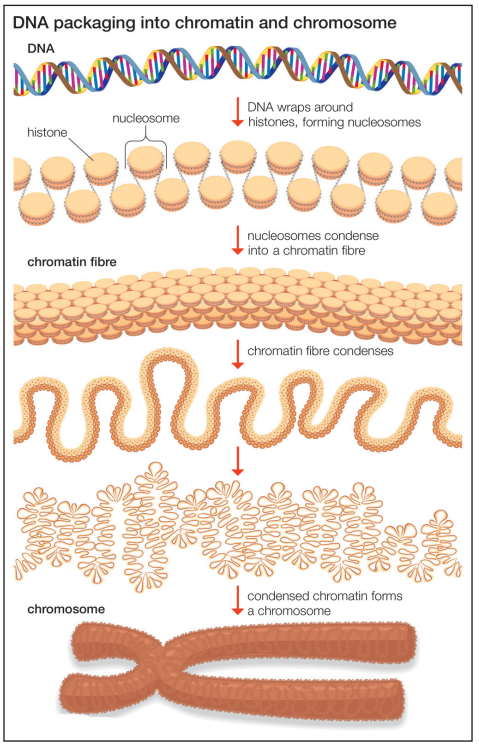


Molecular view of the cell membrane.

Intrinsic proteins penetrate and bind tightly to the lipid bilayer, which is made up largely of phospholipids and cholesterol and which typically is between 4 and 10 nanometers (nm; 1 nm = 10^{-9} metre) in thickness. Extrinsic proteins are loosely bound to the hydrophilic (polar) surfaces, which face the watery medium both inside and outside the cell. Some intrinsic proteins present sugar side chains on the cell's outer surface.

All biomolecules share in common a fundamental relationship between structure and function, which is influenced by factors such as the environment in which a given biomolecule occurs. Lipids, for example, are hydrophobic ("water-fearing"); in water, many spontaneously arrange themselves in such a way that the hydrophobic ends of the molecules are protected from the water, while the hydrophilic ends are exposed to the water. This arrangement gives rise to lipid bilayers, or two layers of phospholipid molecules, which form the membranes of cells and organelles. In another example, DNA,

which is a very long molecule—in humans, the combined length of all the DNA molecules in a single cell stretched end to end would be about 1.8 metres (6 feet), whereas the cell nucleus is about 6 μm ($6*10^{-6}$ metre) in diameter—has a highly flexible helical structure that allows the molecule to become tightly coiled and looped. This structural feature plays a key role in enabling DNA to fit in the cell nucleus, where it carries out its function in coding genetic traits.



DNA wraps around proteins called histones to form units known as nucleosomes.
These units condense into a chromatin fibre, which condenses further to form a chromosome.

Molecular Biophysics

Molecular biophysics is a rapidly evolving interdisciplinary area of research that combines concepts in physics, chemistry, engineering, mathematics and biology. It seeks to understand biomolecular systems and explain biological function in terms of molecular structure, structural organization, and dynamic behaviour at various levels of complexity (from single molecules to supramolecular structures, viruses and small living systems). This discipline covers topics such as the measurement of molecular forces, molecular associations, allosteric interactions, Brownian motion, and cable theory. The technical challenges of molecular biophysics are formidable, and the discipline has required development of specialized equipment and procedures capable of imaging and manipulating minute living structures, as well as novel experimental approaches.

Molecular biophysics typically addresses biological questions similar to those in biochemistry and molecular biology, seeking to find the physical underpinnings of biomolecular phenomena. Scientists in this field conduct research concerned with understanding the interactions between the various systems of a cell, including the interactions between DNA, RNA and protein biosynthesis, as well as how these interactions are regulated. A great variety of techniques are used to answer these questions.

Fluorescent imaging techniques, as well as electron microscopy, X-ray crystallography, NMR spectroscopy, atomic force microscopy (AFM) and small-angle scattering (SAS) both with X-rays and neutrons (SAXS/SANS) are often used to visualize structures of biological significance. Protein dynamics can be observed by neutron spin echo spectroscopy. Conformational change in structure can be measured using techniques such as dual polarisation interferometry, circular dichroism, SAXS and SANS. Direct manipulation of molecules using optical tweezers or AFM, can also be used to monitor biological events where forces and distances are at the nanoscale. Molecular biophysicists often consider complex biological events as systems of interacting entities which can be understood e.g. through statistical mechanics, thermodynamics and chemical kinetics. By drawing knowledge and experimental techniques from a wide variety of disciplines, biophysicists are often able to directly observe, model or even manipulate the structures and interactions of individual molecules or complexes of molecules.

Areas of Research

Computational Biology

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, ecological, behavioral, and social systems. The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science and evolution. Computational biology has become an important part of developing emerging technologies for the field of biology. Molecular modelling encompasses all methods, theoretical and computational, used to model or mimic the behaviour of molecules. The methods are used in the fields of computational chemistry, drug design, computational biology and materials science to study molecular systems ranging from small chemical systems to large biological molecules and material assemblies.

Membrane Biophysics

Membrane biophysics is the study of biological membrane structure and function using physical, computational, mathematical, and biophysical methods. A combination of these methods can be used to create phase diagrams of different types of membranes, which yields information on thermodynamic behavior of a membrane and its components. As opposed to membrane biology, membrane biophysics focuses on quantitative

information and modeling of various membrane phenomena, such as lipid raft formation, rates of lipid and cholesterol flip-flop, protein-lipid coupling, and the effect of bending and elasticity functions of membranes on inter-cell connections.

Motor Proteins

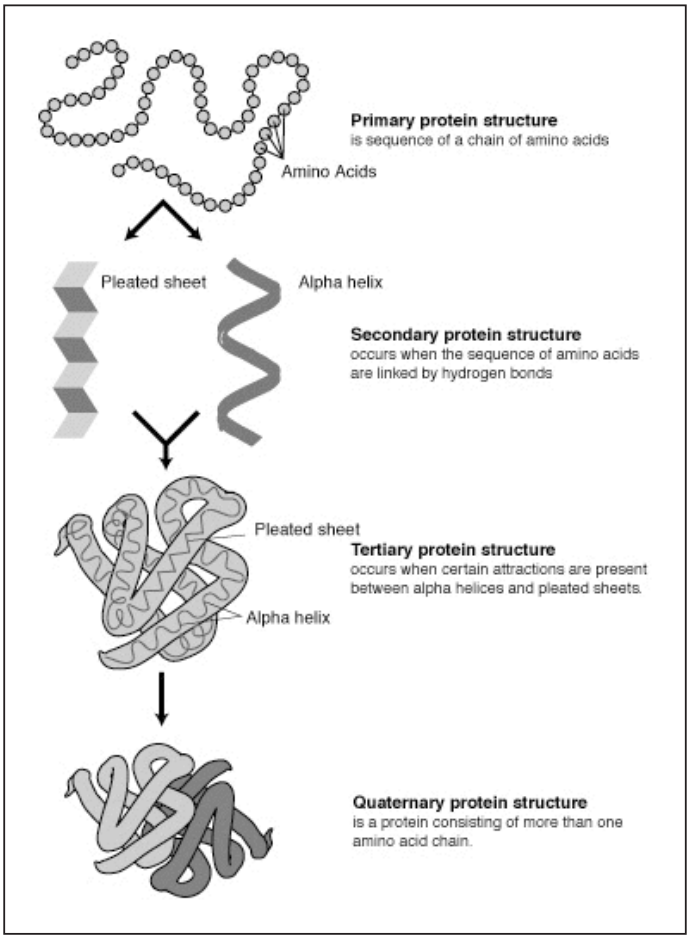
Motor proteins are a class of molecular motors that can move along the cytoplasm of animal cells. They convert chemical energy into mechanical work by the hydrolysis of ATP. A good example is the muscle protein myosin which "motors" the contraction of muscle fibers in animals. Motor proteins are the driving force behind most active transport of proteins and vesicles in the cytoplasm. Kinesins and cytoplasmic dyneins play essential roles in intracellular transport such as axonal transport and in the formation of the spindle apparatus and the separation of the chromosomes during mitosis and meiosis. Axonemal dynein, found in cilia and flagella, is crucial to cell motility, for example in spermatozoa, and fluid transport, for example in trachea. Some biological machines are motor proteins, such as myosin, which is responsible for muscle contraction, kinesin, which moves cargo inside cells away from the nucleus along microtubules, and dynein, which moves cargo inside cells towards the nucleus and produces the axonemal beating of motile cilia and flagella. "In effect, the [motile cilium] is a nanomachine composed of perhaps over 600 proteins in molecular complexes, many of which also function independently as nanomachines...Flexible linkers allow the mobile protein domains connected by them to recruit their binding partners and induce long-range allostery via protein domain dynamics. Other biological machines are responsible for energy production, for example ATP synthase which harnesses energy from proton gradients across membranes to drive a turbine-like motion used to synthesise ATP, the energy currency of a cell. Still other machines are responsible for gene expression, including DNA polymerases for replicating DNA, RNA polymerases for producing mRNA, the spliceosome for removing introns, and the ribosome for synthesising proteins. These machines and their nanoscale dynamics are far more complex than any molecular machines that have yet been artificially constructed.

These molecular motors are the essential agents of movement in living organisms. In general terms, a motor is a device that consumes energy in one form and converts it into motion or mechanical work; for example, many protein-based molecular motors harness the chemical free energy released by the hydrolysis of ATP in order to perform mechanical work. In terms of energetic efficiency, this type of motor can be superior to currently available man-made motors. Richard Feynman theorized about the future of nanomedicine. He wrote about the idea of a medical use for biological machines. Feynman and Albert Hibbs suggested that certain repair machines might one day be reduced in size to the point that it would be possible to "swallow the doctor".

These biological machines might have applications in nanomedicine. For example, they could be used to identify and destroy cancer cells. Molecular nanotechnology is a speculative subfield of nanotechnology regarding the possibility of engineering molecular

assemblers, biological machines which could re-order matter at a molecular or atomic scale. Nanomedicine would make use of these nanorobots, introduced into the body, to repair or detect damages and infections. Molecular nanotechnology is highly theoretical, seeking to anticipate what inventions nanotechnology might yield and to propose an agenda for future inquiry. The proposed elements of molecular nanotechnology, such as molecular assemblers and nanorobots are far beyond current capabilities.

Protein Folding



Constituent amino-acids can be analyzed to predict secondary, tertiary and quaternary protein structure.

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. As the polypeptide chain is being

synthesized by a 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, known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence or primary structure.

Protein Structure Prediction

Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine, in drug design, biotechnology and in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

Spectroscopy

Spectroscopic techniques like NMR, spin label electron spin resonance, Raman spectroscopy, infrared spectroscopy, circular dichroism, and so on have been widely used to understand structural dynamics of important biomolecules and intermolecular interactions.

Motor Proteins

Motor proteins are a class of molecular motors that can move along the cytoplasm of animal cells. They convert chemical energy into mechanical work by the hydrolysis of ATP. Flagellar rotation, however, is powered by a proton pump.

Cellular Functions

The best prominent example of a motor protein is the muscle protein myosin which “motors” the contraction of muscle fibers in animals. Motor proteins are the driving force behind most active transport of proteins and vesicles in the cytoplasm. Kinesins and cytoplasmic dyneins play essential roles in intracellular transport such as axonal transport and in the formation of the spindle apparatus and the separation of the chromosomes during mitosis and meiosis. Axonemal dynein, found in cilia and flagella, is crucial to cell motility, for example in spermatozoa, and fluid transport, for example in trachea.

Diseases Associated with Motor Protein Defects

The importance of motor proteins in cells becomes evident when they fail to fulfill their function. For example, kinesin deficiencies have been identified as the cause for Charcot-Marie-Tooth disease and some kidney diseases. Dynein deficiencies can lead to chronic infections of the respiratory tract as cilia fail to function without dynein. Numerous myosin deficiencies are related to disease states and genetic syndromes. Because myosin II is essential for muscle contraction, defects in muscular myosin predictably cause myopathies. Myosin is necessary in the process of hearing because of its role in the growth of stereocilia so defects in myosin protein structure can lead to Usher syndrome and non-syndromic deafness.

Cytoskeletal Motor Proteins

Motor proteins utilizing the cytoskeleton for movement fall into two categories based on their substrates: Microfilaments or microtubules. Actin motors such as myosin move along microfilaments through interaction with actin, and microtubule motors such as dynein and kinesin move along microtubules through interaction with tubulin. There are two basic types of microtubule motors: plus-end motors and minus-end motors, depending on the direction in which they "walk" along the microtubule cables within the cell.

Actin Motors

Myosin

Myosins are a superfamily of actin motor proteins that convert chemical energy in the form of ATP to mechanical energy, thus generating force and movement. The first identified myosin, myosin II, is responsible for generating muscle contraction. Myosin II is an elongated protein that is formed from two heavy chains with motor heads and two light chains. Each myosin head contains actin and ATP binding site. The myosin heads bind and hydrolyze ATP, which provides the energy to walk toward the plus end of an actin filament. Myosin II are also vital in the process of cell division. For example, non-muscle myosin II bipolar thick filaments provide the force of contraction needed to divide the cell into two daughter cells during cytokinesis. In addition to myosin II, many other myosin types are responsible for variety of movement of non-muscle cells. For example, myosin is involved in intracellular organization and the protrusion of actin-rich structures at the cell surface. Myosin V is involved in vesicle and organelle transport. Myosin XI is involved in cytoplasmic streaming, wherein movement along microfilament networks in the cell allows organelles and cytoplasm to stream in a particular direction. Eighteen different classes of myosins are known.

Genomic representation of myosin motors:

- Fungi (yeast): 5.
- Plants (Arabidopsis): 17.

- Insects (Drosophila): 13.
- Mammals (human): 40.
- Chromadorea (nematode C. elegans): 15.

Microtubule Motors

Kinesin

Kinesins are a group of related motor proteins that use a microtubule track in anterograde movement. They are vital to spindle formation in mitotic and meiotic chromosome separation during cell division and are also responsible for shuttling mitochondria, Golgi bodies, and vesicles within eukaryotic cells. Kinesins have two heavy chains and two light chains per active motor. The two globular head motor domains in heavy chains can convert the chemical energy of ATP hydrolysis into mechanical work to move along microtubules. The direction in which cargo is transported can be towards the plus-end or the minus-end, depending on the type of kinesin. In general, kinesins with N-terminal motor domains move their cargo towards the plus ends of microtubules located at the cell periphery, while kinesins with C-terminal motor domains move cargo towards the minus ends of microtubules located at the nucleus. Fourteen distinct kinesin families are known, with some additional kinesin-like proteins that cannot be classified into these families.

Genomic representation of kinesin motors:

- Fungi (yeast): 6.
- Plants (Arabidopsis thaliana): 61.
- Insects (Drosophila melanogaster): 25.
- Mammals (human): 45.

Dynein

Dyneins are microtubule motors capable of a retrograde sliding movement. Dynein complexes are much larger and more complex than kinesin and myosin motors. Dyneins are composed of two or three heavy chains and a large and variable number of associated light chains. Dyneins drive intracellular transport toward the minus end of microtubules which lies in the microtubule organizing center near the nucleus. The dynein family has two major branches. Axonemal dyneins facilitate the beating of cilia and flagella by rapid and efficient sliding movements of microtubules. Another branch is cytoplasmic dyneins which facilitate the transport of intracellular cargos. Compared to 15 types of axonemal dynein, only two cytoplasmic forms are known.

Genomic representation of dynein motors:

- Fungi (yeast): 1.

- Plants (Arabidopsis thaliana): 0.
- Insects (Drosophila melanogaster): 13.
- Mammals (human): 14-15.

Plant-specific Motors

In contrast to animals, fungi and non-vascular plants, the cells of flowering plants lack dynein motors. However, they contain a larger number of different kinesins. Many of these plant-specific kinesin groups are specialized for functions during plant cell mitosis. Plant cells differ from animal cells in that they have a cell wall. During mitosis, the new cell wall is built by the formation of a cell plate starting in the center of the cell. This process is facilitated by a phragmoplast, a microtubule array unique to plant cell mitosis. The building of cell plate and ultimately the new cell wall requires kinesin-like motor proteins.

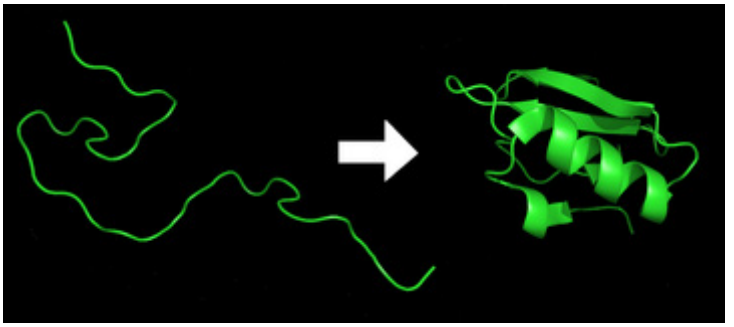
Another motor protein essential for plant cell division is kinesin-like calmodulin-binding protein (KCBP), which is unique to plants and part kinesin and part myosin.

Other Molecular Motors

Besides the motor proteins above, there are many more types of proteins capable of generating forces and torque in the cell. Many of these molecular motors are ubiquitous in both prokaryotic and eukaryotic cells, although some, such as those involved with cytoskeletal elements or chromatin, are unique to eukaryotes. The motor protein prestin, expressed in mammalian cochlear outer hair cells, produces mechanical amplification in the cochlea. It is a direct voltage-to-force converter, which operates at the microsecond rate and possesses piezoelectric properties.

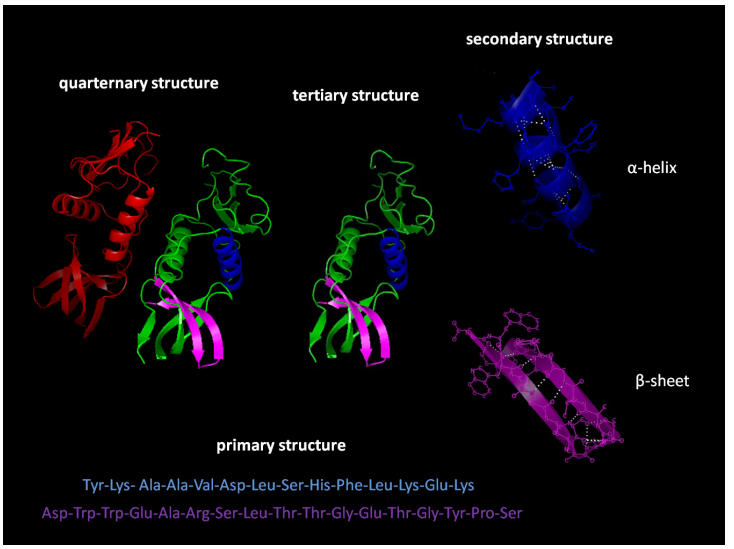
Protein Folding

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 a 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. As the polypeptide chain is being synthesized by a 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, known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence or primary structure.



Protein before and after 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. It happens in cooking, in burns, in proteinopathies, and in other contexts.



Results of protein folding.

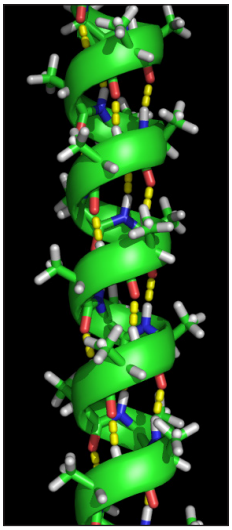
The duration of the folding process varies dramatically depending on the protein of interest. When studied outside the cell, the slowest folding proteins require many minutes or hours to fold primarily due to proline isomerization, and must pass through a number of intermediate states, like checkpoints, 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.

Process of Protein Folding

Primary Structure

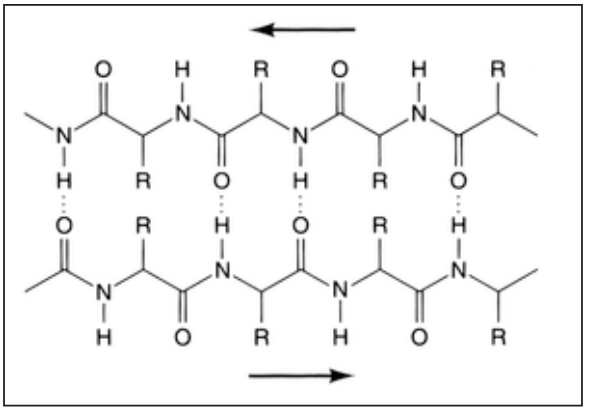
The primary structure of a protein, its linear amino-acid sequence, determines its native conformation. The specific amino acid residues and their position in the polypeptide chain are the determining factors for which portions of the protein fold closely together and form its three-dimensional conformation. 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.

Secondary Structure



The alpha helix spiral formation.

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. The β pleated sheet is a structure that forms with the backbone bending over itself to form the hydrogen bonds. 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.



An anti-parallel beta pleated sheet displaying hydrogen bonding within the backbone.

Tertiary Structure

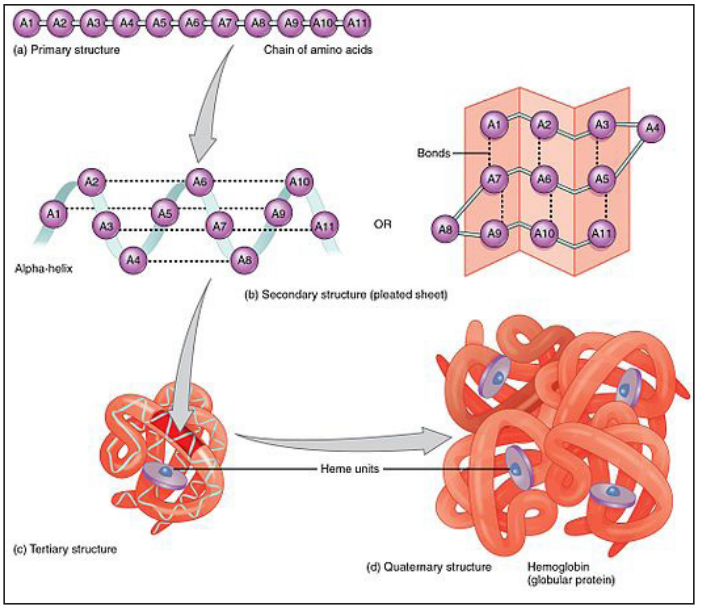
The alpha helices and beta pleated sheets can be amphipathic in nature, or contain a hydrophilic portion and a hydrophobic portion. This property of secondary structures aids in the tertiary structure of a protein in which the folding occurs so that the hydrophilic sides are facing the aqueous environment 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

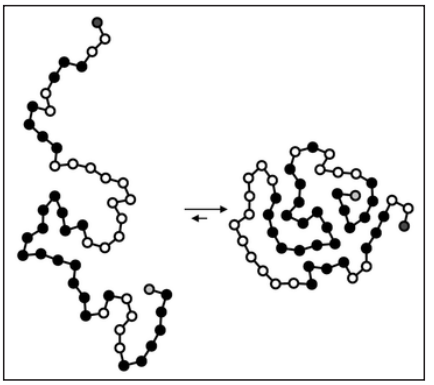
Folding is a spontaneous process that is mainly guided by hydrophobic interactions, formation of intramolecular hydrogen bonds, van der Waals forces, and it is opposed by conformational entropy. The process of folding often begins co-translationally, so that the N-terminus of the protein begins to fold 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.



All forms of protein structure.

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



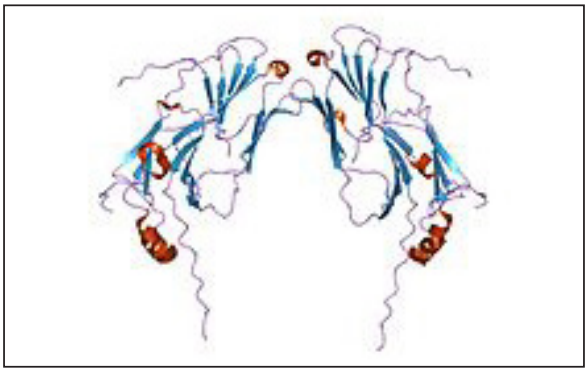
Hydrophobic collapse. In the compact fold (to the right), the hydrophobic amino acids (shown as black spheres) collapse toward the center to become shielded from aqueous environment.

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 delta G 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.

In proteins with globular folds, hydrophobic amino acids tend to be interspersed along the primary sequence, rather than randomly distributed or clustered together. However, proteins that have recently been born de novo, which tend to be intrinsically disordered, show the opposite pattern of hydrophobic amino acid clustering along the primary sequence.

Chaperones



Example of a small eukaryotic heat shock protein.

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 toward 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.

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 structures 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.

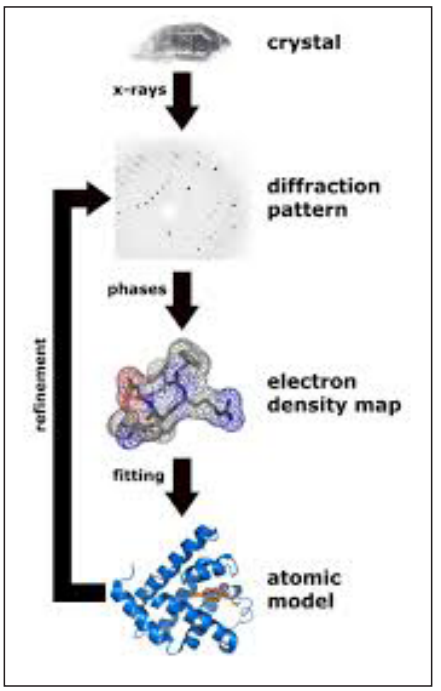
Aggregated proteins are associated with prion-related illnesses such as Creutzfeldt–Jakob disease, bovine spongiform encephalopathy (mad cow disease), amyloid-related illnesses such as Alzheimer's disease and familial amyloid cardiomyopathy or polyneuropathy, as well as intracellular aggregation diseases 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.

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



Steps of 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 denaturation temperature (Tm) 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 (FT) 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.

Biotin Painting

Biotin painting enables condition-specific cellular snapshots of unfolded proteins. Biotin 'painting' shows a bias towards predicted Intrinsically disordered proteins.

Computational Studies of Protein Folding

Computational studies 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.

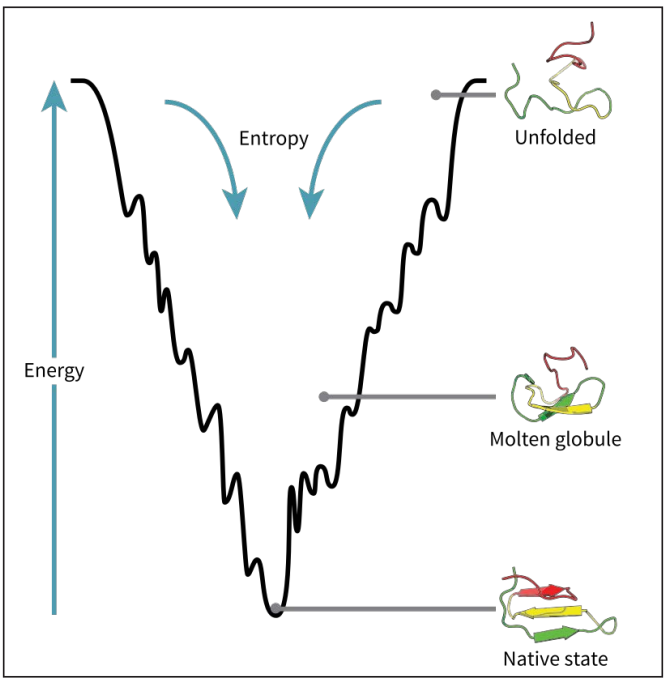
Levinthal's Paradox

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.

Energy Landscape of Protein Folding

The configuration space of a protein during folding can be visualized as energy landscape. According to Joseph Bryngelson and Peter Wolynes, proteins follow the principle of minimal frustration meaning that naturally evolved proteins have optimized their folding energy landscapes., and that nature has chosen amino acid sequences so that the folded state of the protein is sufficiently stable. In addition, the acquisition of the folded state had to become a sufficiently 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.



The energy funnel by which an unfolded polypeptide chain assumes its native structure.

A consequence of these evolutionarily selected sequences is that proteins are generally thought to have globally "funneled energy landscapes" that are largely directed toward 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.

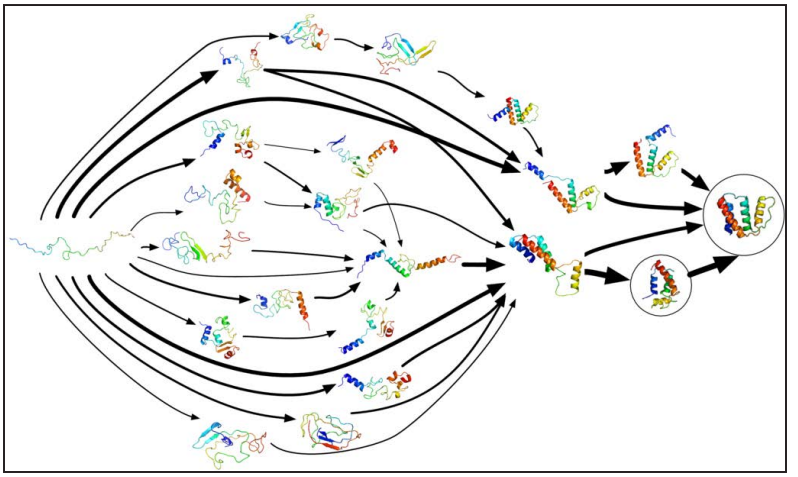
The unfolded polypeptide chain begins at the top of the funnel where it may assume the largest number of unfolded variations and is in its highest energy state. Energy landscapes such as these indicate that there are a large number of initial possibilities, but only a single native state is possible; 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.

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.

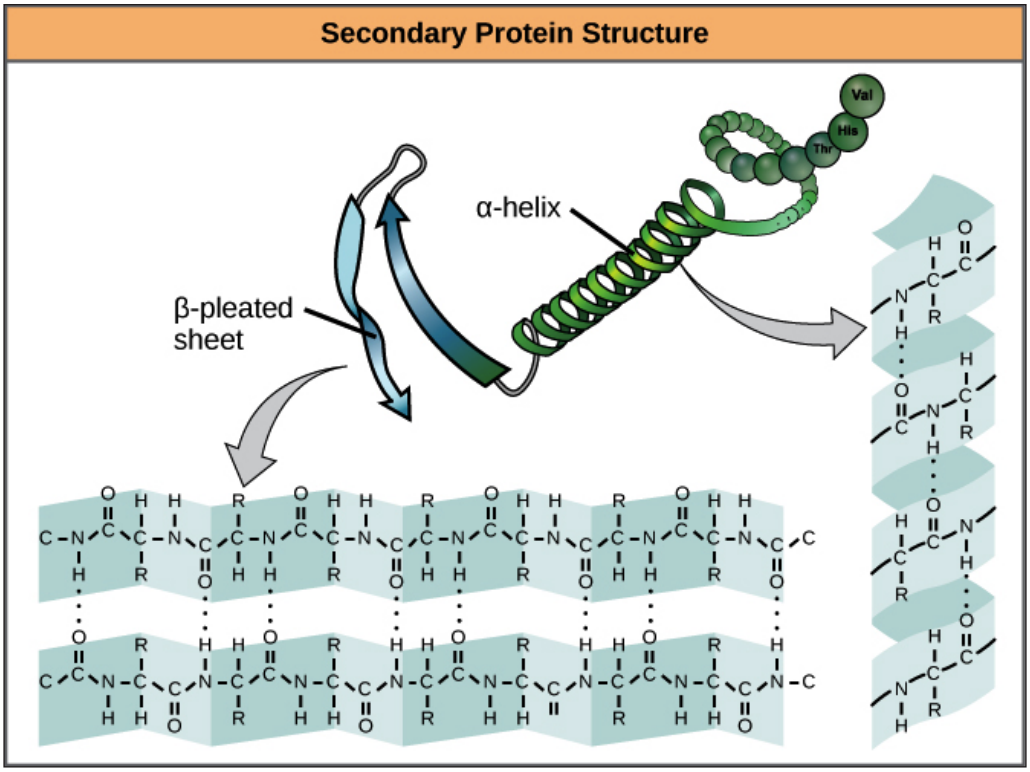


The possible shapes and folding pathways a protein can take as it condenses from its initial randomly coiled state (left) into its native 3D structure (right).

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.

Protein Structure Prediction

Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

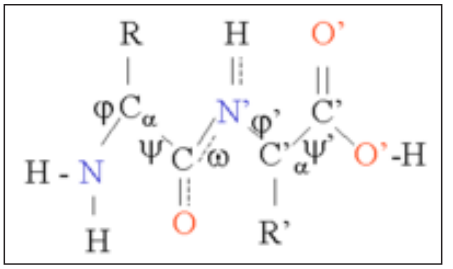


Constituent amino-acids can be analyzed to predict secondary, tertiary and quaternary protein structure.

Protein Structure and Terminology

Proteins are chains of amino acids joined together by peptide bonds. Many conformations of this chain are possible due to the rotation of the chain about each Cα atom. It is these conformational changes that are responsible for differences in the three dimensional structure of proteins. Each amino acid in the chain is polar, i.e. it has separated positive and negative charged regions with a free carbonyl group, which can act as hydrogen bond acceptor and an NH group, which can act as hydrogen bond donor. These groups can therefore interact in the protein structure. The 20 amino acids can be classified according to the chemistry of the side chain which also plays an important structural role. Glycine takes on a special position, as it has the smallest side chain, only one hydrogen atom, and therefore can increase the local flexibility in the protein structure. Cysteine on the other hand can react with another cysteine residue and thereby form a cross link stabilizing the whole structure.

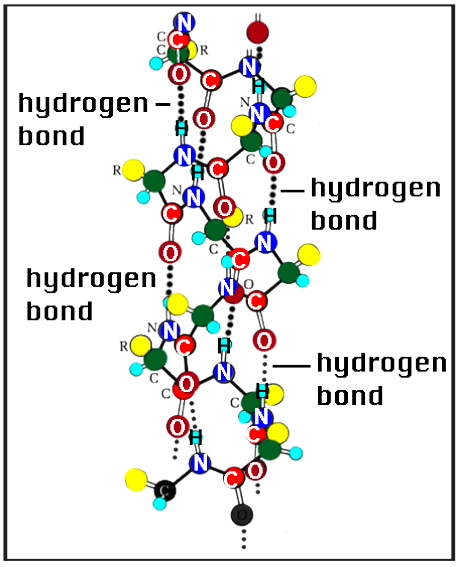
The protein structure can be considered as a sequence of secondary structure elements, such as α helices and β sheets, which together constitute the overall three-dimensional configuration of the protein chain. In these secondary structures regular patterns of H bonds are formed between neighboring amino acids, and the amino acids have similar Φ and Ψ angles.



Bond angles for ψ and ω.

The formation of these structures neutralizes the polar groups on each amino acid. The secondary structures are tightly packed in the protein core in a hydrophobic environment. Each amino acid side group has a limited volume to occupy and a limited number of possible interactions with other nearby side chains, a situation that must be taken into account in molecular modeling and alignments.

α Helix



An alpha-helix with hydrogen bonds (yellow dots).

The α helix is the most abundant type of secondary structure in proteins. The α helix has 3.6 amino acids per turn with an H bond formed between every fourth residue; the average length is 10 amino acids (3 turns) or 10 Å but varies from 5 to 40 (1.5 to 11 turns). The alignment of the H bonds creates a dipole moment for the helix with a resulting partial positive charge at the amino end of the helix. Because this region has free NH2 groups, it will interact with negatively charged groups such as phosphates. The most common location of α helices is at the surface of protein cores, where they provide an interface with the aqueous environment. The inner-facing side of the helix tends to have hydrophobic amino acids and the outer-facing side hydrophilic amino acids. Thus, every third of four amino acids along the chain will tend to be hydrophobic,

a pattern that can be quite readily detected. In the leucine zipper motif, a repeating pattern of leucines on the facing sides of two adjacent helices is highly predictive of the motif. A helical-wheel plot can be used to show this repeated pattern. Other α helices buried in the protein core or in cellular membranes have a higher and more regular distribution of hydrophobic amino acids, and are highly predictive of such structures. Helices exposed on the surface have a lower proportion of hydrophobic amino acids. Amino acid content can be predictive of an α -helical region. Regions richer in alanine (A), glutamic acid (E), leucine (L), and methionine (M) and poorer in proline (P), glycine (G), tyrosine (Y), and serine (S) tend to form an α helix. Proline destabilizes or breaks an α helix but can be present in longer helices, forming a bend.

β Sheet

β sheets are formed by H bonds between an average of 5–10 consecutive amino acids in one portion of the chain with another 5–10 farther down the chain. The interacting regions may be adjacent, with a short loop in between, or far apart, with other structures in between. Every chain may run in the same direction to form a parallel sheet, every other chain may run in the reverse chemical direction to form an anti parallel sheet, or the chains may be parallel and anti parallel to form a mixed sheet. The pattern of H bonding is different in the parallel and anti parallel configurations. Each amino acid in the interior strands of the sheet forms two H bonds with neighboring amino acids, whereas each amino acid on the outside strands forms only one bond with an interior strand. Looking across the sheet at right angles to the strands, more distant strands are rotated slightly counterclockwise to form a left-handed twist. The Cα atoms alternate above and below the sheet in a pleated structure, and the R side groups of the amino acids alternate above and below the pleats. The Φ and Ψ angles of the amino acids in sheets vary considerably in one region of the Ramachandran plot. It is more difficult to predict the location of β sheets than of α helices. The situation improves somewhat when the amino acid variation in multiple sequence alignments is taken into account.

Loop

Loops are regions of a protein chain that are: 1) Between α helices and β sheets, 2) Of various lengths and three-dimensional configurations, and 3) On the surface of the structure. Hairpin loops that represent a complete turn in the polypeptide chain joining two antiparallel β strands may be as short as two amino acids in length. Loops interact with the surrounding aqueous environment and other proteins. Because amino acids in loops are not constrained by space and environment as are amino acids in the core region, and do not have an effect on the arrangement of secondary structures in the core, more substitutions, insertions, and deletions may occur. Thus, in a sequence alignment, the presence of these features may be an indication of a loop. The positions of introns in genomic DNA sometimes correspond to the locations of loops in the encoded protein. Loops also tend to have charged and polar amino acids and are frequently a

component of active sites. A detailed examination of loop structures has shown that they fall into distinct families.

Coils

A region of secondary structure that is not an α helix, a β sheet, or a recognizable turn is commonly referred to as a coil.

Protein Classification

Proteins may be classified according to both structural and sequence similarity. For structural classification, the sizes and spatial arrangements of secondary structures described in the above paragraph are compared in known three-dimensional structures. Classification based on sequence similarity was historically the first to be used. Initially, similarity based on alignments of whole sequences was performed. Later, proteins were classified on the basis of the occurrence of conserved amino acid patterns. Databases that classify proteins by one or more of these schemes are available. In considering protein classification schemes, it is important to keep several observations in mind. First, two entirely different protein sequences from different evolutionary origins may fold into a similar structure. Conversely, the sequence of an ancient gene for a given structure may have diverged considerably in different species while at the same time maintaining the same basic structural features. Recognizing any remaining sequence similarity in such cases may be a very difficult task. Second, two proteins that share a significant degree of sequence similarity either with each other or with a third sequence also share an evolutionary origin and should share some structural features also. However, gene duplication and genetic rearrangements during evolution may give rise to new gene copies, which can then evolve into proteins with new function and structure.

The more commonly used terms for evolutionary and structural relationships among proteins are listed below. Many additional terms are used for various kinds of structural features found in proteins. Descriptions of such terms may be found at the CATH Web site, the Structural Classification of Proteins (SCOP) Web site, and a Glaxo-Wellcome tutorial on the Swiss bioinformatics Expasy Web site.

- Active site: A localized combination of amino acid side groups within the tertiary (three-dimensional) or quaternary (protein subunit) structure that can interact with a chemically specific substrate and that provides the protein with biological activity. Proteins of very different amino acid sequences may fold into a structure that produces the same active site.
- Architecture: Is the relative orientations of secondary structures in a three-dimensional structure without regard to whether or not they share a similar loop structure.
- Fold (Topology): A type of architecture that also has a conserved loop structure.

- Blocks: Is a conserved amino acid sequence pattern in a family of proteins. The pattern includes a series of possible matches at each position in the represented sequences, but there are not any inserted or deleted positions in the pattern or in the sequences. By way of contrast, sequence profiles are a type of scoring matrix that represents a similar set of patterns that includes insertions and deletions.

- Class: A term used to classify protein domains according to their secondary structural content and organization. Four classes were originally recognized by Levitt and Chothia, and several others have been added in the SCOP database. Three classes are given in the CATH database: mainly-α, mainly-β, and α–β, with the α–β class including both alternating α/β and α+β structures.

- Core: The portion of a folded protein molecule that comprises the hydrophobic interior of α-helices and β-sheets. The compact structure brings together side groups of amino acids into close enough proximity so that they can interact. When comparing protein structures, as in the SCOP database, core is the region common to most of the structures that share a common fold or that are in the same superfamily. In structure prediction, core is sometimes defined as the arrangement of secondary structures that is likely to be conserved during evolutionary change.

- Domain (Sequence Context): A segment of a polypeptide chain that can fold into a three-dimensional structure irrespective of the presence of other segments of the chain. The separate domains of a given protein may interact extensively or may be joined only by a length of polypeptide chain. A protein with several domains may use these domains for functional interactions with different molecules.

- Family (Sequence Context): A group of proteins of similar biochemical function that are more than 50% identical when aligned. This same cutoff is still used by the Protein Information Resource (PIR). A protein family comprises proteins with the same function in different organisms (orthologous sequences) but may also include proteins in the same organism (paralogous sequences) derived from gene duplication and rearrangements. If a multiple sequence alignment of a protein family reveals a common level of similarity throughout the lengths of the proteins, PIR refers to the family as a homeomorphic family. The aligned region is referred to as a homeomorphic domain, and this region may comprise several smaller homology domains that are shared with other families. Families may be further subdivided into subfamilies or grouped into superfamilies based on respective higher or lower levels of sequence similarity. The SCOP database reports 1296 families and the CATH database, reports 1846 families.

  When the sequences of proteins with the same function are examined in greater detail, some are found to share high sequence similarity. They are obviously

members of the same family by the above criteria. However, others are found that have very little, or even insignificant, sequence similarity with other family members. In such cases, the family relationship between two distant family members A and C can often be demonstrated by finding an additional family member B that shares significant similarity with both A and C. Thus, B provides a connecting link between A and C. Another approach is to examine distant alignments for highly conserved matches.

At a level of identity of 50%, proteins are likely to have the same three-dimensional structure, and the identical atoms in the sequence alignment will also superimpose within approximately 1 Å in the structural model. Thus, if the structure of one member of a family is known, a reliable prediction may be made for a second member of the family, and the higher the identity level, the more reliable the prediction. Protein structural modeling can be performed by examining how well the amino acid substitutions fit into the core of the three-dimensional structure.

- Family (Structural Context): As used in the FSSP database (Families of structurally similar proteins) and the DALI/FSSP Web site, two structures that have a significant level of structural similarity but not necessarily significant sequence similarity.

- Fold: Similar to structural motif, includes a larger combination of secondary structural units in the same configuration. Thus, proteins sharing the same fold have the same combination of secondary structures that are connected by similar loops. An example is the Rossman fold comprising several alternating α helices and parallel β strands. In the SCOP, CATH, and FSSP databases, the known protein structures have been classified into hierarchical levels of structural complexity with the fold as a basic level of classification.

- Homologous domain (Sequence Context): An extended sequence pattern, generally found by sequence alignment methods, that indicates a common evolutionary origin among the aligned sequences. A homology domain is generally longer than motifs. The domain may include all of a given protein sequence or only a portion of the sequence. Some domains are complex and made up of several smaller homology domains that became joined to form a larger one during evolution. A domain that covers an entire sequence is called the homeomorphic domain by PIR (Protein Information Resource).

- Module: A region of conserved amino acid patterns comprising one or more motifs and considered to be a fundamental unit of structure or function. The presence of a module has also been used to classify proteins into families.

- Motif (Sequence Context): A conserved pattern of amino acids that is found in

two or more proteins. In the Prosite catalog, a motif is an amino acid pattern that is found in a group of proteins that have a similar biochemical activity, and that often is near the active site of the protein. Examples of sequence motif databases are the Prosite catalog and the Stanford Motifs Database.

- Motif (Structural Context): A combination of several secondary structural elements produced by the folding of adjacent sections of the polypeptide chain into a specific three-dimensional configuration. An example is the helix-loop-helix motif. Structural motifs are also referred to as supersecondary structures and folds.

- Position-specific scoring matrix (Sequence Context, also known as weight or Scoring Matrix): Represents a conserved region in a multiple sequence alignment with no gaps. Each matrix column represents the variation found in one column of the multiple sequence alignment.

- Position-specific scoring matrix—3D (Structural Context): Represents the amino acid variation found in an alignment of proteins that fall into the same structural class. Matrix columns represent the amino acid variation found at one amino acid position in the aligned structures.

- Primary structure: The linear amino acid sequence of a protein, which chemically is a polypeptide chain composed of amino acids joined by peptide bonds.

- Profile (Sequence Context): A scoring matrix that represents a multiple sequence alignment of a protein family. The profile is usually obtained from a well-conserved region in a multiple sequence alignment. The profile is in the form of a matrix with each column representing a position in the alignment and each row one of the amino acids. Matrix values give the likelihood of each amino acid at the corresponding position in the alignment. The profile is moved along the target sequence to locate the best scoring regions by a dynamic programming algorithm. Gaps are allowed during matching and a gap penalty is included in this case as a negative score when no amino acid is matched. A sequence profile may also be represented by a hidden Markov model, referred to as a profile HMM.

- Profile (Structural Context): A scoring matrix that represents which amino acids should fit well and which should fit poorly at sequential positions in a known protein structure. Profile columns represent sequential positions in the structure, and profile rows represent the 20 amino acids. As with a sequence profile, the structural profile is moved along a target sequence to find the highest possible alignment score by a dynamic programming algorithm. Gaps may be included and receive a penalty. The resulting score provides an indication as to whether or not the target protein might adopt such a structure.

- Quaternary Structure: The three-dimensional configuration of a protein molecule comprising several independent polypeptide chains.
- Secondary Structure: The interactions that occur between the C, O, and NH groups on amino acids in a polypeptide chain to form α-helices, β-sheets, turns, loops, and other forms, and that facilitate the folding into a three-dimensional structure.
- Superfamily: A group of protein families of the same or different lengths that are related by distant yet detectable sequence similarity. Members of a given superfamily thus have a common evolutionary origin. Originally, Dayhoff defined the cutoff for superfamily status as being the chance that the sequences are not related of 10 6, on the basis of an alignment score. Proteins with few identities in an alignment of the sequences but with a convincingly common number of structural and functional features are placed in the same superfamily. At the level of three-dimensional structure, superfamily proteins will share common structural features such as a common fold, but there may also be differences in the number and arrangement of secondary structures. The PIR resource uses the term homeomorphic superfamilies to refer to superfamilies that are composed of sequences that can be aligned from end to end, representing a sharing of single sequence homology domain, a region of similarity that extends throughout the alignment. This domain may also comprise smaller homology domains that are shared with other protein families and superfamilies. Although a given protein sequence may contain domains found in several superfamilies, thus indicating a complex evolutionary history, sequences will be assigned to only one homeomorphic superfamily based on the presence of similarity throughout a multiple sequence alignment. The superfamily alignment may also include regions that do not align either within or at the ends of the alignment. In contrast, sequences in the same family align well throughout the alignment.
- Supersecondary Structure: A term with similar meaning to a structural motif. Tertiary structure is the three-dimensional or globular structure formed by the packing together or folding of secondary structures of a polypeptide chain.

Secondary Structure

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins based only on knowledge of their amino acid sequence. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as “extended” conformations), or turns. The success of a prediction is determined by comparing it to the results of the DSSP algorithm (or similar e.g. STRIDE) applied to the crystal structure of the protein. Specialized algorithms have been developed for the detection of specific well-defined patterns such as transmembrane helices and coiled coils in proteins.

The best modern methods of secondary structure prediction in proteins reach about 80% accuracy; this high accuracy allows the use of the predictions as feature improving fold recognition and ab initio protein structure prediction, classification of structural motifs, and refinement of sequence alignments. The accuracy of current protein secondary structure prediction methods is assessed in weekly benchmarks such as LiveBench and EVA.

Early methods of secondary structure prediction, introduced in the 1960s and early 1970s, focused on identifying likely alpha helices and were based mainly on helix-coil transition models. Significantly more accurate predictions that included beta sheets were introduced in the 1970s and relied on statistical assessments based on probability parameters derived from known solved structures. These methods, applied to a single sequence, are typically at most about 60-65% accurate, and often underpredict beta sheets. The evolutionary conservation of secondary structures can be exploited by simultaneously assessing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of an aligned column of amino acids. In concert with larger databases of known protein structures and modern machine learning methods such as neural nets and support vector machines, these methods can achieve up to 80% overall accuracy in globular proteins. The theoretical upper limit of accuracy is around 90%, partly due to idiosyncrasies in DSSP assignment near the ends of secondary structures, where local conformations vary under native conditions but may be forced to assume a single conformation in crystals due to packing constraints. Limitations are also imposed by secondary structure prediction's inability to account for tertiary structure; for example, a sequence predicted as a likely helix may still be able to adopt a beta-strand conformation if it is located within a beta-sheet region of the protein and its side chains pack well with their neighbors. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

Other Improvements

It is reported that in addition to the protein sequence, secondary structure formation depends on other factors. For example, it is reported that secondary structure tendencies depend also on local environment, solvent accessibility of residues, protein structural class, and even the organism from which the proteins are obtained. Based on such observations, some studies have shown that secondary structure prediction can be improved by addition of information about protein structural class, residue accessible surface area and also contact number information.

Tertiary Structure

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein

structures—typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy—is lagging far behind the output of protein sequences.

The protein structure prediction remains an extremely difficult and unresolved undertaking. The two main problems are calculation of protein free energy and finding the global minimum of this energy. A protein structure prediction method must explore the space of possible protein structures which is astronomically large. These problems can be partially bypassed in "comparative" or homology modeling and fold recognition methods, in which the search space is pruned by the assumption that the protein in question adopts a structure that is close to the experimentally determined structure of another homologous protein. On the other hand, the de novo or ab initio protein structure prediction methods must explicitly resolve these problems. The progress and challenges in protein structure prediction has been reviewed in Zhang 2008.

Before Modeling

Most tertiary structure modelling methods, such as Rosetta, are optimized for modelling the tertiary structure of single protein domains. A step called domain parsing, or domain boundary prediction, is usually done first to split a protein into potential structural domains. As with the rest of tertiary structure prediction, this can be done comparatively from known structures or ab initio with the sequence only (usually by machine learning, assisted by covariation). The structures for individual domains are docked together in a process called domain assembly to form the final tertiary structure.

Ab Initio Protein Modeling

Energy- and Fragment-based Methods

Ab initio- or de novo- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e., global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. To predict protein structure de novo for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers (such as Blue Gene or MDGRAPE-3) or distributed computing (such as Folding@home, the Human Proteome Folding Project and Rosetta@Home). Although these computational barriers are vast, the potential benefits of structural genomics (by predicted or experimental methods) make ab initio structure prediction an active research field.

As of 2009, a 50-residue protein could be simulated atom-by-atom on a supercomputer for 1 millisecond. As of 2012, comparable stable-state sampling could be done

on a standard desktop with a new graphics card and more sophisticated algorithms. A much larger simulation timescales can be achieved using coarse-grained modeling.

Evolutionary Covariation to Predict 3D Contacts

As sequencing became more commonplace in the 1990s several groups used protein sequence alignments to predict correlated mutations and it was hoped that these co-evolved residues could be used to predict tertiary structure (using the analogy to distance constraints from experimental procedures such as NMR). The assumption is when single residue mutations are slightly deleterious, compensatory mutations may occur to restabilize residue-residue interactions. This early work used what are known as local methods to calculate correlated mutations from protein sequences, but suffered from indirect false correlations which result from treating each pair of residues as independent of all other pairs.

In 2011, a different, and this time global statistical approach, demonstrated that predicted coevolved residues were sufficient to predict the 3D fold of a protein, providing there are enough sequences available (>1,000 homologous sequences are needed). The method, EVfold, uses no homology modeling, threading or 3D structure fragments and can be run on a standard personal computer even for proteins with hundreds of residues. The accuracy of the contacts predicted using this and related approaches has now been demonstrated on many known structures and contact maps, including the prediction of experimentally unsolved transmembrane proteins.

Comparative Protein Modeling

Comparative protein modelling uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has been suggested that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins.

These methods may also be split into two groups:

- Homology modeling is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment. Unsurprisingly, homology modelling is most accurate when the target and template have similar sequences.

- Protein threading scans the amino acid sequence of an unknown structure against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. This type of method is also known as 3D-1D fold recognition due to its compatibility analysis between three-dimensional structures and linear protein sequences. This method has also given rise to methods performing an inverse folding search by evaluating the compatibility of a given structure with a large database of sequences, thus predicting which sequences have the potential to produce a given fold.

Side-chain Geometry Prediction

Accurate packing of the amino acid side chains represents a separate problem in protein structure prediction. Methods that specifically address the problem of predicting side-chain geometry include dead-end elimination and the self-consistent mean field methods. The side chain conformations with low energy are usually determined on the rigid polypeptide backbone and using a set of discrete side chain conformations known as "rotamers." The methods attempt to identify the set of rotamers that minimize the model's overall energy.

These methods use rotamer libraries, which are collections of favorable conformations for each residue type in proteins. Rotamer libraries may contain information about the conformation, its frequency, and the standard deviations about mean dihedral angles, which can be used in sampling. Rotamer libraries are derived from structural bioinformatics or other statistical analysis of side-chain conformations in known experimental structures of proteins, such as by clustering the observed conformations for tetrahedral carbons near the staggered (60°, 180°, -60°) values.

Rotamer libraries can be backbone-independent, secondary-structure-dependent, or backbone-dependent. Backbone-independent rotamer libraries make no reference to backbone conformation, and are calculated from all available side chains of a certain type. Secondary-structure-dependent libraries present different dihedral angles and/or rotamer frequencies for α-helix, β-sheet, or coil secondary structures. Backbone-dependent rotamer libraries present conformations and/or frequencies dependent on the local backbone conformation as defined by the backbone dihedral angles ϕ and Ψ, regardless of secondary structure.

The modern versions of these libraries as used in most software are presented as multidimensional distributions of probability or frequency, where the peaks correspond to the dihedral-angle conformations considered as individual rotamers in the lists. Some versions are based on very carefully curated data and are used primarily for structure validation, while others emphasize relative frequencies in much larger data sets and are the form used primarily for structure prediction, such as the Dunbrack rotamer libraries.

Side-chain packing methods are most useful for analyzing the protein's hydrophobic core, where side chains are more closely packed; they have more difficulty addressing the looser constraints and higher flexibility of surface residues, which often occupy multiple rotamer conformations rather than just one.

Prediction of Structural Classes

Statistical methods have been developed for predicting structural classes of proteins based on their amino acid composition, pseudo amino acid composition and functional domain composition. Secondary structure predicion also implicitly generates such a prediction for singular domains.

Quaternary Structure

In the case of complexes of two or more proteins, where the structures of the proteins are known or can be predicted with high accuracy, protein–protein docking methods can be used to predict the structure of the complex. Information of the effect of mutations at specific sites on the affinity of the complex helps to understand the complex structure and to guide docking methods.

Computational Biology

Computational biology is a branch of biology involving the application of computers and computer science to the understanding and modeling of the structures and processes of life. It entails the use of computational methods (e.g., algorithms) for the representation and simulation of biological systems, as well as for the interpretation of experimental data, often on a very large scale.

Underpinnings of Computational Biology

The beginnings of computational biology essentially date to the origins of computer science. British mathematician and logician Alan Turing, often called the father of computing, used early computers to implement a model of biological morphogenesis (the development of pattern and form in living organisms) in the early 1950s, shortly before his death. At about the same time, a computer called MANIAC, built at the Los Alamos National Laboratory in New Mexico for weapons research, was applied to such purposes as modeling hypothesized genetic codes. (Pioneering computers had been used even earlier in the 1950s for numeric calculations in population genetics, but the first instances of authentic computational modeling in biology were the work by Turing and by the group at Los Alamos).

By the 1960s, computers had been applied to deal with much more-varied sets of analyses, namely those examining protein structure. These developments marked the rise

of computational biology as a field, and they originated from studies centred on protein crystallography, in which scientists found computers indispensable for carrying out laborious Fourier analyses to determine the three-dimensional structure of proteins.

Starting in the 1950s, taxonomists began to incorporate computers into their work, using the machines to assist in the classification of organisms by clustering them based on similarities of sets of traits. Such taxonomies have been useful particularly for phylogenetics (the study of evolutionary relationships). In the 1960s, when existing techniques were extended to the level of DNA sequences and amino acid sequences of proteins and combined with a burgeoning knowledge of cellular processes and protein structures, a whole new set of computational methods was developed in support of molecular phylogenetics. These computational methods entailed the creation of increasingly sophisticated techniques for the comparison of strings of symbols that benefited from the formal study of algorithms and the study of dynamic programming in particular. Indeed, efficient algorithms always have been of primary concern in computational biology, given the scale of data available, and biology has in turn provided examples that have driven much advanced research in computer science. Examples include graph algorithms for genome mapping (the process of locating fragments of DNA on chromosomes) and for certain types of DNA and peptide sequencing methods, clustering algorithms for gene expression analysis and phylogenetic reconstruction, and pattern matching for various sequence search problems.

Beginning in the 1980s, computational biology drew on further developments in computer science, including a number of aspects of artificial intelligence (AI). Among these were knowledge representation, which contributed to the development of ontologies (the representation of concepts and their relationships) that codify biological knowledge in “computer-readable” form, and natural-language processing, which provided a technological means for mining information from text Perhaps most significantly, the subfield of machine learning found wide use in biology, from modeling sequences for purposes of pattern recognition to the analysis of high-dimensional (complex) data from large-scale gene-expression studies.

Applications of Computational Biology

Initially, computational biology focused on the study of the sequence and structure of biological molecules, often in an evolutionary context. Beginning in the 1990s, however, it extended increasingly to the analysis of function. Functional prediction involves assessing the sequence and structural similarity between an unknown and a known protein and analyzing the proteins’ interactions with other molecules. Such analyses may be extensive, and thus computational biology has become closely aligned with systems biology, which attempts to analyze the workings of large interacting networks of biological components, especially biological pathways.

Biochemical, regulatory, and genetic pathways are highly branched and interleaved, as well as dynamic, calling for sophisticated computational tools for their modeling and analysis. Moreover, modern technology platforms for the rapid, automated (high-throughput) generation of biological data have allowed for an extension from traditional hypothesis-driven experimentation to data-driven analysis, by which computational experiments can be performed on genome-wide databases of unprecedented scale. As a result, many aspects of the study of biology have become unthinkable without the power of computers and the methodologies of computer science.

Distinctions among Related Fields

How best to distinguish computational biology from the related field of bioinformatics, and to a lesser extent from the fields of mathematical and theoretical biology, has long been a matter of debate. The terms bioinformatics and computational biology are often used interchangeably, even by experts, and many feel that the distinctions are not useful. Both fields fundamentally are computational approaches to biology. However, whereas bioinformatics tends to refer to data management and analysis using tools that are aids to biological experimentation and to the interpretation of laboratory results, computational biology typically is thought of as a branch of biology, in the same sense that computational physics is a branch of physics. In particular, computational biology is a branch of biology that is uniquely enabled by computation. In other words, its formation was not defined by a need to deal with scale; rather, it was defined by virtue of the techniques that computer science brought to the formulation and solving of challenging problems, to the representation and examination of domain knowledge, and ultimately to the generation and testing of scientific hypotheses.

Computational biology is more easily distinguished from mathematical biology, though there are overlaps. The older discipline of mathematical biology was concerned primarily with applications of numerical analysis, especially differential equations, to topics such as population dynamics and enzyme kinetics. It later expanded to include the application of advanced mathematical approaches in genetics, evolution, and spatial modeling. Such mathematical analyses inevitably benefited from computers, especially in instances involving systems of differential equations that required simulation for their solution. The use of automated calculation does not in itself qualify such activities as computational biology. However, mathematical modeling of biological systems does overlap with computational biology, particularly where simulation for purposes of prediction or hypothesis generation is a key element of the model. A useful distinction in this regard is that between numerical analysis and discrete mathematics; the latter, which is concerned with symbolic rather than numeric manipulations, is considered foundational to computer science, and in general its applications to biology may be considered aspects of computational biology.

Computational biology can also be distinguished from theoretical biology (which itself is sometimes grouped with mathematical biology), though again there are

significant relationships. Theoretical biology often focuses on mathematical abstractions and speculative interpretations of biological systems that may or may not be of practical use in analysis or amenable to computational implementation. Computational biology generally is associated with practical application, and indeed journals and annual meetings in the field often actively encourage the presentation of biological analyses using real data along with theory. On the other hand, important contributions to computational biology have arisen through aspects of theoretical biology derived from information theory, network theory, and nonlinear dynamical systems (among other areas). As an example, advances in the mathematical study of complex networks have increased scientists' understanding of naturally occurring interactions among genes and gene products, providing insight into how characteristic network architectures may have arisen in the course of evolution and why they tend to be robust in the face of perturbations such as mutations.

References

- Pirovano W, Heringa J (2010). "Protein secondary structure prediction". Methods in Molecular Biology. 609: 327–48. doi:10.1007/978-1-60327-241-4_19. ISBN 978-1-60327-240-7. PMID 20221928
- Biomolecule, science: britannica.com, Retrieved 09 May, 2020
- Mathews DH (June 2006). "Revolutions in RNA secondary structure prediction". Journal of Molecular Biology. 359 (3): 526–32. doi:10.1016/j.jmb.2006.01.067. PMID 16500677
- Computational-biology, science: britannica.com, Retrieved 16 April, 2020
- Lee S, Tsai FT (2005). "Molecular chaperones in protein quality control". Journal of Biochemistry and Molecular Biology. 38 (3): 259–65. doi:10.5483/BMBRep.2005.38.3.259. PMID 15943899

4
Biophysical Techniques

There are various biophysical techniques that are studied under this domain. A few of them are fluorescence imaging technique, X-ray crystallography, atomic force microscopy, isothermal titration calorimetry, NMR spectroscopy, etc. This chapter has been carefully written to provide an extensive understanding of these biophysical techniques.

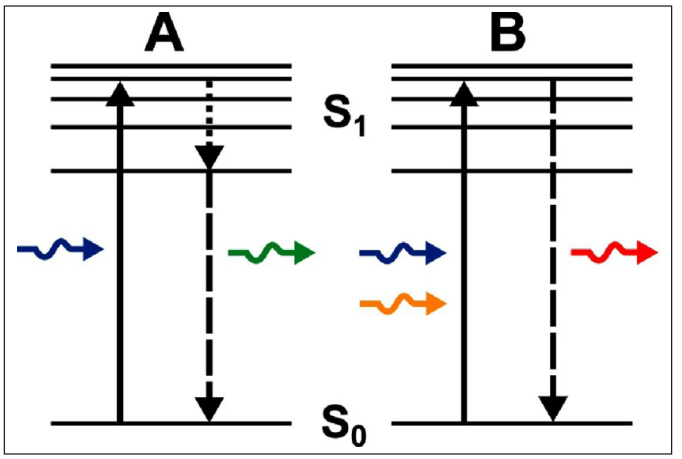
Fluorescence Imaging Technique

The advent of optical microscopy has had a significant impact on the life sciences specially in visualizing the molecular architecture of the cell. The development of fluorescent probes has allowed observations at the single molecule level whilst the recent introduction of super resolution microscopy has allowed imaging of an array of nanoscopic biological complexes with unprecedented resolution.

Fluorescence

A quantum mechanical treatment of an atom reveals the probability of an associated electron, possessing an amount of available energy, of occupying any energy level of the atom. At scales larger than molecular scales, levels group into bands, sometimes referred to as *states*. Often, electrons occupy the lowest energy state (referred to as the *ground state*). Electrons could also occupy higher energy states if the ground state is fully occupied or if they are provided with the energy differential sufficient to *excite* them from the ground state to any unoccupied upper energy state. The energy differential can be provided in the form of light. Light is composed of quantized elementary particles known as *photons*. It is now established, following years of experimentation, that photons possess two natures, a wave nature and a particle nature. Besides mass and speed, which reflect their particle nature, photons are also characterized with wavelength and frequency, which reflect their wave nature. Wavelength describes how often a wave repeats itself on the length scale, and frequency, describes how often a wave repeats itself but on the time scale. The energy of a photon is directly proportional to its frequency and inversely proportional to its wavelength.

A coherent light source (e.g. laser) emitting photons with a known wavelength can be used to excite an electron in an atom, molecule or bulk, to an energy state higher than the ground state provided that the energy difference is less than the energy of the excitation photon. Following excitation, electrons could spontaneously, or be stimulated to, relax to the ground state. Either result in the emission of a photon of different energy, and thus different wavelength, to the excitation photon. This process is known as *fluorescence*. The wavelength of a spontaneously emitted photon is higher than that emitted stimulatingly. These processes can be figuratively depicted using a simplified Jablonski diagram. Emission at different wavelengths can be split using a *dichroic* mirror; a mirror that has different transmission properties at different wavelengths. This implies that the excitation and emission can be split; the latter to be collected and the former to be blocked.



Schematic diagram representing energy levels in an atom.

(A) Spontaneous emission: An electron excited (solid black arrow) by an incident photon (blue arrow) from the ground state (S_0) to the first state (S_1) spontaneously relaxes by internal conversion (dotted line) to the lowest energy level of S_1 and then to S_0 by emitting another photon (green arrow) of higher wavelength. (B) Stimulated emission: An electron excited (solid black arrow) by an incident photon (blue arrow) from the ground state (S_0) to the first state (S_1) can be stimulated to relax to S_0 by another photon (orange arrow) of higher wavelength and emitting a photon (red arrow) of higher wavelength. The wavelength of a spontaneously emitted photon is higher than that produced by stimulated emission.

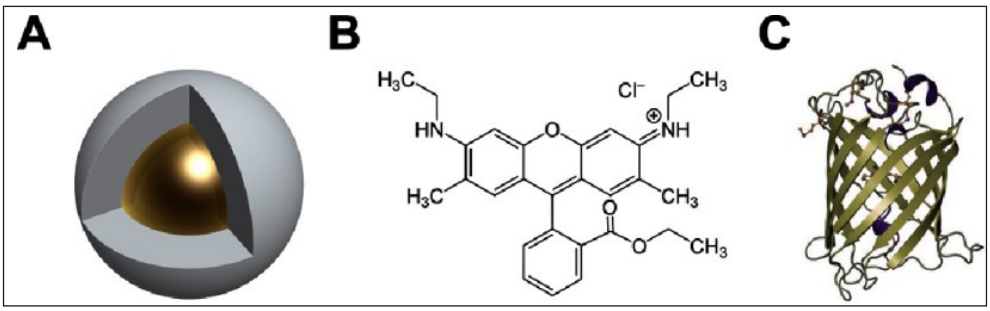
Fluorescent Labels

Observation of single biological molecules is far beyond the realm of the naked eye and white light microscopy. To facilitate observation at the single molecule level, a fluorescent molecule, referred to as a *label* or a *fluorophore*, is attached to the biological molecule of interest, excited using a coherent light source and the emitted fluorescence is, consequently, collected. Improved camera technologies and fluorescent labels have rendered this possible.

The decision to choose an appropriate fluorescent label relies on three important parameters: Photo stability, brightness and size. Following excitation, an electron could relax to the ground state, and emit a photon, or relax to a non-emissive state and not emit. The latter process is known as *photoblinking*. Relaxation to the non-emissive state is followed by relaxation to the ground state and, subsequently, resumed ability to fluoresce after excitation. It is evident that this process is temporary and thus the term blinking. *Photobleaching*, or permanent loss of fluorescence, can occur due to an alteration to the chemical structure of a fluorescent label following long periods of laser exposure. Typically, photo blinking occurs on the scale of a few seconds, whilst photobleaching occurs after 0.1 –1000 seconds of continuous exposure depending on the label.

Although often implicated as a limiting factor in single particle tracking experiments, photobleaching has been useful in determining the stoichiometry of protein complexes. It was shown that by exciting singly-labeled monomers in an oligomeric complex and recording the overall decrease in intensity caused by photobleaching, the stoichiometry of the complex under study can be deduced by counting the photobleaching steps.

In general, photostability (i.e. photo - blinking and bleaching) determine the fate of a fluorescent label and can sometimes limit its application in biophysical investigations. Whilst photo stability determines the traceability of a fluorescent label, brightness determines its detectability. *Quantum yield* or the ratio of the number of photons emitted to the number of photons received by a fluorescent label. Higher quantum yields denote higher probabilities of fluorescence detection. There is a limit, however, on how bright a label could be. A large label is typically brighter than a smaller label. Yet, for many biological processes, where the molecule of interest is smaller than the label or its dynamics can be affected by the size of the label, a large label cannot be used. Other parameters such as labeling schemes, coupling efficiency, and compatibility with live-cell imaging are also considered when choosing an appropriate label.



Schematic diagram representing different fluorescent labels.

(A) Quantum dots are inorganic semi-conducting nanocrystals composed of a core (e.g. CdTe, CdSe, InP, InGaP etc.) and a shell (e.g. ZnS). The shell can be functionalized (i.e. grafted) with a chemical moiety to enable targeted labeling of a protein. Furthermore, the size and composition of the shell can be tuned to improve the photo stability of a

quantum dot for single molecule imaging purposes. Though possessing high quantum yields, quantum dots suffer from poor intercellular internalization and uncharacterized size-dependent effects on protein dynamics. (B) Organic dyes are small fluorescent moieties that can be linked to the protein under study via a chemical tag. The small size and high quantum yield of organic dyes render them suitable labels for single molecule imaging. However, their potential phototoxic effects make a large subset, yet, unsuitable for imaging in living cells. Shown in the figure is the chemical structure of the organic dye Rhodamine 6G. (C) Fluorescent proteins are genetically expressible labels which are suitable for imaging in living systems. However, their utility for single molecule imaging is dependent on their proper folding and the size of the protein under study. Shown in the figure is the structure of the Green Fluorescent Protein (GFP).

Table: Comparison between different fluorescent labels.

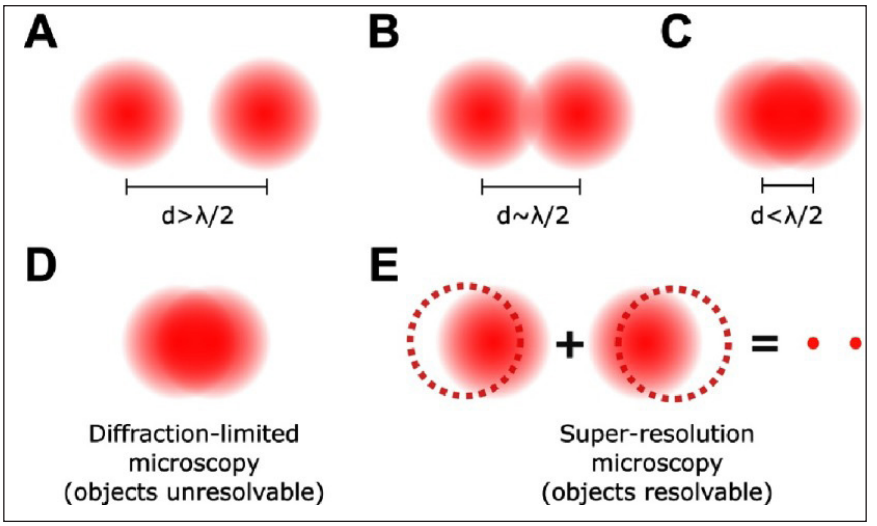
Fluorescent label	Size / structure	Photo bleaching / blinking times	Brightness	Labeling methodologies	Reported applications	Available labels
Quantum Dot (QD)	6 –60 nm / Inorganic nanocrystal	>1000 s / <1 ms	200 – 2000 $mM.cm^{-1}$	Bio conjugation, or electrostatic interaction.	Tracking of single membrane and motor proteins *in vitro and in vivo*.	CdTe, CdSe, InP, InGaP, PbS, PbSe, and ZnS-coated CdSe
Organic dye	0.5 nm / organic molecule	1 – 10 s / 0.5 s	45 – 250 $mM.cm^{-1}$	Targeted labeling via a peptide tag (e.g. FLAG-tag, HA-tag or His-tag), or, a protein tag (e.g. Halo-tag).	Super resolution imaging of histone proteins *in vivo*. Tracking of membrane and motor proteins *in vitro*.	For a list of organic dyes refer to.
Fluorescent protein (FP)	3 nm / protein	13 – 174 s / 1 -2 s	4 – 95 $mM. cm^{-1}$	Protein expression or targeted labeling via a peptide or protein tag.	Imaging the dynamics of filopodia, microtubules, transcription factors, chromosomes, myosin, and, cadherin *in vivo*.	For a list of fluorescent proteins refer to.

Fluorescence Microscopy Techniques

Diffraction-limited Microscopy Techniques

The wave nature of light imposes a fundamental limit on the dimensions of the smallest object that can be observed using light microscopy. The limit was derived by Ernest Abbe who found that any two objects could be resolved if the distance in between is approximately larger than half of the wavelength of the excitation source. This implies

that any two fluorescently-labeled entities, observed using visible light (450 –650 nm), cannot be discerned if they are closer than 200 –300 nm. Fluorescence microscopy techniques which cannot overcome that resolution (diffraction) limit are known as diffraction-limited.



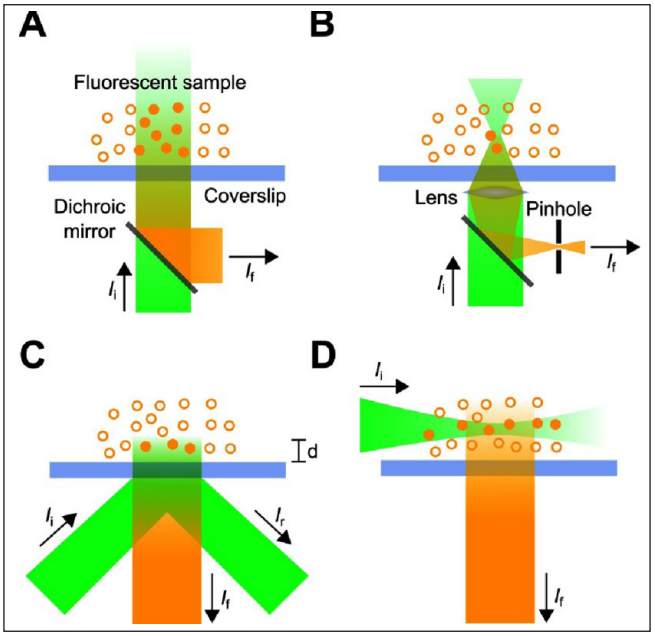
Schematic diagrams representing the diffraction limit of light.

Due to the wave nature of light, a point object observed in the far field is seen as an airy disk. If the distance between two point objects is (A) greater than, or, (B) equal to half of wavelength of the excitation source, the objects become resolvable, and can be precisely located. If the distance between the two point objects is (C) less than half of the wavelength of the excitation source, the objects become unresolvable. This resolution limit is known as the Abbe diffraction limit. Closely located point objects which are not resolvable using diffraction-limited microscopy techniques (D) can be resolved using super-resolution microscopy techniques (E). Unlike diffraction-limited microscopy techniques which operate by detecting the fluorescence from all the molecules in a single image, super-resolution microscopy detects the fluorescence from each molecule, or a subset of molecules, at different times; overlapping the images obtained for distinct fluorophores results in a super-resolved image.

Epi-fluorescence Illumination Microscopy

The use of epi-illumination had originally referred to the arrangement of optical components that permits illumination from above the specimen. The term now refers to any orthogonal scheme of illumination whereby the excitation source is directed perpendicular to the plane of the specimen. Epi-fluorescence illumination has, and still is, implemented using xenon arc & mercury vapor lamps, Tungsten-Halogen, and Light Emitting Diodes (LEDs). However, the strong divergence and incoherence of these excitation sources meant that single-molecules could hardly be observed in thin samples even at substantially-high light levels. To observe single molecules using epi-fluorescence

illumination, a laser beam is directed through an objective lens onto a sample and the fluorescence emission is collected using the same objective. The emission and excitation are separated, where then, the former can be imaged. Although this scheme serves to effectively image thin (<1 μm) samples, it fails to image dense thick samples where the fluorescence signals from the different imaging planes destructively contribute to the acquired image. The simple illumination scheme has served to uncover the reversible assembly of the 'Tat A' component of the twin-arginine protein transport system.



Schematic diagrams representing various diffraction-limited fluorescence microscopy techniques.

In Epi-illumination (A), an upwardly-directed laser beam, under filling a lens objective, is directed perpendicular to the plane of the fluorescent sample and the fluorescence, from a large portion of the sample, is collected using the same objective. For large samples, the acquired image is blurred as a result of the overlapping of the fluorescence signal obtained from the different imaging planes. In confocal microscopy (B), an upwardly-directed laser beam is focused at specific planes of a sample to excite fluorophores inside the focusing spot. By scanning the beam laterally across the sample, a full image of a specific plane can be acquired. Acquisition rates in confocal microscopy are limited by the scanning speed of the beam rendering image acquisition slow. To resolve that problem, light is directed at an angle, above the critical angle, to undergo total internal reflection at the interface (C). To satisfy the boundary conditions at the interface, an evanescent (decaying) wave propagates for a few hundred nanometers above the boundary exciting fluorophores at the vicinity of the interface. The acquired image is not blurred as a result of the emission being limited to a thin imaging plane. Furthermore, acquisition rates in TIRF microscopy is only limited by the camera frame rate rendering image acquisition fast. In light sheet microscopy (D) a laser beam is directed though a cylindrical objective, or a two-dimensional photonic lattice, creating a

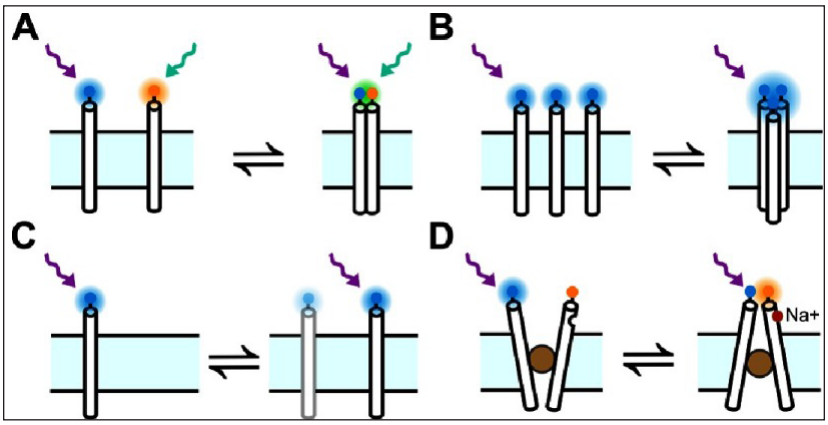
light sheet that excites a section through a fluorescent sample. The fluorescence from the excited plane is orthogonally collected. LSFM combines fine optical-sectioning with fast acquisition rates. I_i is the incident beam, I_r is the reflected beam, and, I_f is the fluorescence emission.

Confocal Microscopy

To ameliorate the aforementioned problem, an upwardly-directed laser beam is focused inside the sample and a micron-sized pinhole is placed in the emission path to restrict the emission to that from a micron-sized spot and eliminate any out-of-focus signal. To image a single plane, the spot is laterally scanned across the sample. Scanning is typically performed by an electrically-expandable material, a set of electronically-controlled mirrors or a multiple-hole spinning disk which deflect the focused beam laterally across the sample. In all cases, acquisition is not set by the detector's frame rate, as in epi-illumination microscopy, but by the scanning speed. Although the fine sectioning of biological samples in confocal microscopy produces high-resolution images, the limitation on the acquisition speed renders confocal microscopy an unsuitable technique for imaging highly dynamic processes.

Total Internal Reflection Fluorescence Microscopy (TIRF-M)

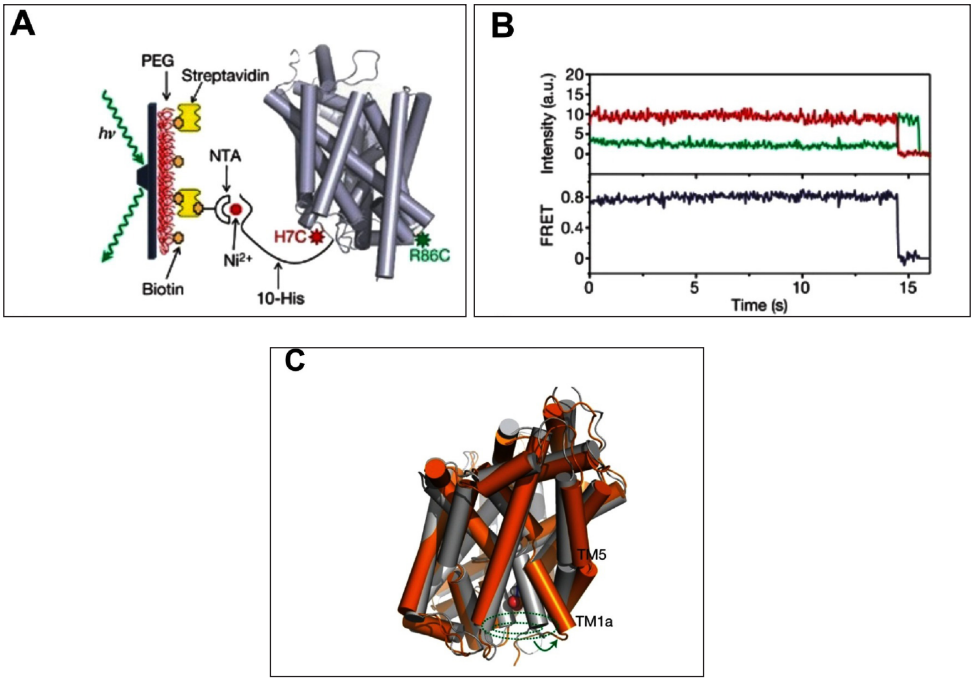
Light incident at a boundary between two different media changes its direction of propagation; this process is known as *refraction*. If the angle of incidence exceeds the *critical angle* dictated by the ratio of the refractive indices of the two mediums, light undergoes *total internal reflection*. To satisfy boundary conditions, an evanescent (decaying) field propagates parallel to the plane of the boundary. The depth of penetration of the evanescent field is dictated by the angle of incidence and the ratio of the medium's refractive indices; however, it can be adjusted to a few hundred nanometers. The evanescent field exclusively excites fluorophores at the vicinity of the boundary. This illumination scheme is known as Total Internal Reflection Fluorescence Microscopy (TIRF-M).



Schematic diagrams representing various single molecule techniques used in studying the function and dynamics of membrane proteins.

(A) The binding of two different fluorescently-labeled subunits can be detected by the presence of transient co-localized fluorescence bursts in two emissions channels. (B) The clustering of multiple subunits can be detected observing an increase in fluorescence. The number of subunits in a single cluster can be deduced by dividing the fluorescence signal of the cluster by the fluorescence signal of a single subunit. Alternatively, a time sequence of the cluster is acquired as it photobleaches and the number of subunits can be deduced by counting the photobleaching steps. (C) The random movement (diffusion) of a protein can be imaged and the hydrodynamic properties of the membrane, and surrounding, can be obtained from the calculated diffusion coefficient. (D) The gating and alternating action of an ion channel, or transporter, can also be detected by single-molecule FRET. In FRET, energy is transferred between a donor (blue) and acceptor (orange) when they are less than 10 nm apart. The ratio of the emission of the acceptor to the emission of the donor, known as the FRET ratio, is a direct measure of how far both fluorophores are. By labeling ion channels, and other biomolecules, at closely located sites, using suitable FRET pairs, and monitoring the FRET ratio, it is possible to observe proteins dynamics in real time. Reproduced with permission from.

Due to its simple implementation, TIRF-M found numerous applications in biology, especially where it relates to imaging lipid membranes and its constituents. The superior signal-to-noise ratio of TIRF-M has been configured for real time measurements on numerous transporters and channels at the single molecule level. Although lipid membranes and their constituents can be easily imaged in TIRF-M, the richness of intra-cellular biology is not accessible to this technique.

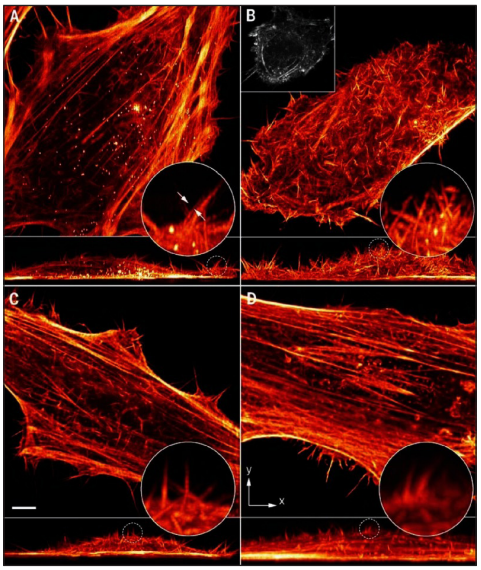


Reported example for the use of single molecule FRET in imaging the gating dynamics of the neurotransmitter transporter homologue (LeuT).

(A) The diffusion of doubly-labeled LeuT is restricted through binding to a glass substrate. A laser beam undergoes total internal reflection at the glass / water interface and the evanescent field excites the fluorescently labeled protein. (B) The emission from the donor and acceptor dyes is recorded and the FRET ratio, defined as the ratio of the acceptor-to-donor emission, is calculated under different physiological conditions. (C) The observed dynamic changes can be translated into their structural context with the aid of molecular dynamics simulations.

Light Sheet Fluorescence Microscopy (LSFM)

Light sheet microscopy was developed to overcome the limitations imposed by confocal and TIRF microscopies. Light sheet microscopy relies on an orthogonal illumination scheme whereby a couple of microns' thick sheet of light, produced by directing a Gaussian-shaped beam though a cylindrical lens objective, illuminates a section of a fluorescently labeled sample. The fluorescence signal from the section is collected by another objective placed perpendicular to the excitation beam.



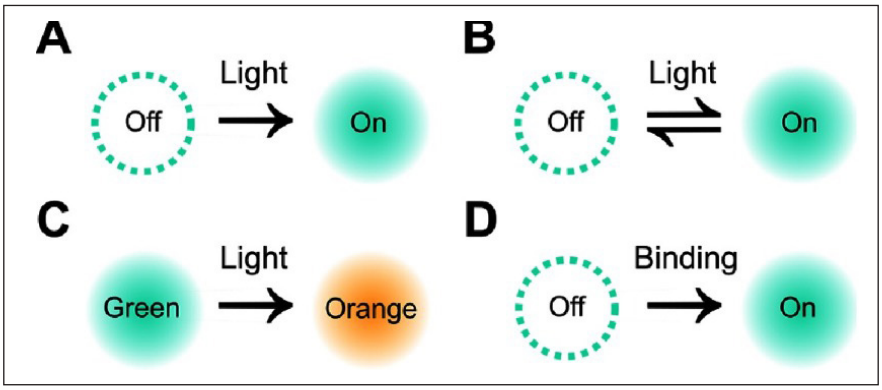
Hela cells labeled with the fluorescent protein mEmerald imaged using lattice light sheet microscopy in different configurations.

Although this scheme enables video rate imaging and effectively reduces photo bleaching, the thickness of the light sheet does not permit fine sectioning, and, therefore images suffer from reduced contrast. Lattice light sheet microscopy was recently reported by Chen and coworkers where the cylindrical-lens system was replaced by a two-dimensional optical lattice. This configuration produces a hundred nanometers' ultra-thin light sheets that can be used to illuminate thin sections of a sample. This improvement permits 3D intra-cellular imaging at sub second temporal resolutions combined with unprecedented axial and lateral spatial resolutions. Light sheet microscopy has been instrumental in uncovering the organization of RNA polymerase II *in vivo*, tracking

single molecules on the apical membrane of living cells and imaging mouse embryos during the first development stages.

Super Resolution Microscopy Techniques

Since many biological molecules perform their functions cooperatively, in nanoscopic molecular complexes (5 nm –100 nm), diffraction-limited fluorescence microscopy techniques fail to uncover the architecture, and eventually dynamics, of those biological machineries. To overcome the diffraction limit, a set of fluorescence microscopy techniques, known as super resolution microscopies, were developed since 1994. Super resolution microscopies rely on localizing individual fluorophores by sequentially modulating their fluorescence and constructing a super-resolved image by super position of the exclusively localized fluorophores.



Schematic representation of some fluorescence modulation schemes.

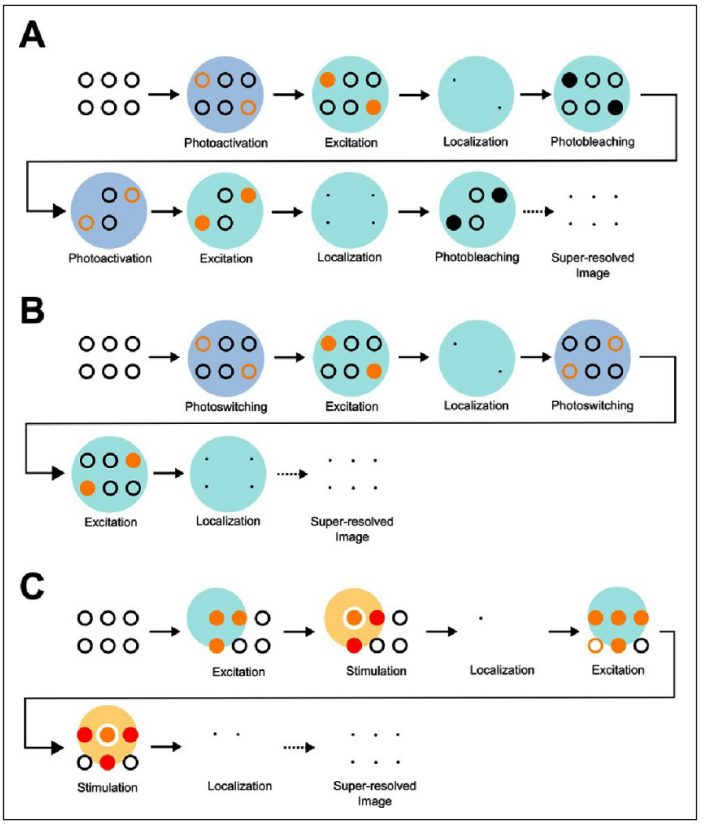
(A) Photoactivation: A deactivated (switched off) protein is irreversibly activated (switched on) by the application of light. (B) Photoswitching: A deactivated protein is reversibly activated by the application of light. (C) Stimulated emission: An active protein can spontaneously emit (green) or be stimulated to emit (orange). (D) Binding-induced activation: A deactivated (switched off) protein is irreversibly activated (switched on) upon binding to a substrate. In a fluorescent sample, the fluorescence of some fluorophores is sequentially modulated discriminating them from other fluorophores whose fluorescence has not been modulated. Successive modulations result in obtaining a super resolved image where the fluorescence from single molecules is well resolved.

Photoactivated Localization Microscopy (PALM)

The engineering of photoactivatable fluorescent proteins has prompted the development of PALM in 2006. In PALM, a weak short-wavelength laser beam randomly activates a subset of inactive fluorophores. The activation beam has a low intensity enough to ensure that fluorophores in the same diffraction limited region are unlikely to be all activated at the same time. Another beam, of longer wavelength and higher intensity,

is used to excite the activated fluorophores and the emission is subsequently collected. Since the distance between the majorities of the active fluorophore pairs is larger than the diffraction limit, the centers of the associated airy disks can be determined and the position of the fluorophores can be precisely located.

Following localization, the active fluorophores are photobleached and another subset of inactive fluorophores is activated, excited, localized and photobleached. To ensure the localization of the majority of the fluorophores residing in the desired optical section, the aforementioned steps are repeated and the images of the localized fluorophores are overlapped until a satisfactory complete super-resolved image is produced.



Schematic diagrams demonstrating the mechanism of operation of PALM, STORM and STED.

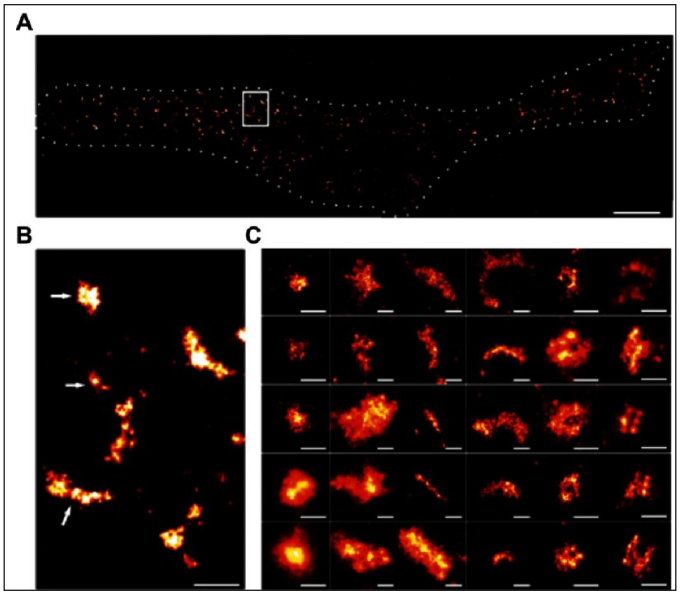
In PALM (A), a randomly selected subset of a large number of closely-located inactive fluorophores (empty black circles) is activated using a weak short wavelength laser beam. The activated fluorophores (empty orange circles) are excited using a longer wavelength laser beam and the fluorescence (solid orange circles) is collected. Since the distance between the exclusively excited fluorophores is larger than the diffraction limit, they can be localized (black dots). The fluorophores are photo bleached under continuous excitation and another set of fluorophores is activated, excited, localized and photo bleached. This process is repeated until a complete super resolved image is obtained. In STORM (B), a randomly selected subset of a large number of closely

located inactive switchable fluorophores (empty black circles) is switched using a weak short wavelength laser beam. The activated fluorophores (empty orange circles) are excited using a longer wavelength laser beam and the fluorescence (solid orange circles) is collected and localized. Instead of undergoing photo bleaching, the activated fluorophores are deactivated, or switched off, to a dark state by the application of a weak short wavelength laser beam, which in turn randomly activates (switches on) another subset of inactivated fluorophores. This is repeated until a complete super resolved image is obtained. In STED (C) a laser beam (large blue circle) excites a large number of active fluorophores (orange solid circles) in a diffraction limited region. A doughnut shaped beam (large orange circle), centered on the center of the excitation beam, de-excites all the excited fluorophores (red solid circles) except for the very few which are located at the center (orange solid circles). The stimulation of the de-excited fluorophores results in the emission of photons having a wavelength longer than that emitted spontaneously from the center of the doughnut shaped beam. Both emissions are filtered and the fluorophores at the center of the beam are localized (black dots) accordingly. By scanning the beam and repeating the excitation and stimulation processes, a complete super-resolved image is obtained.

Since multiple fluorophores can be localized at the same time, PALM is considered to be a wide-field imaging technique. Although acquisition rates in wide-field imaging techniques (e.g. LSFM and TIRF) are typically considered to be limited by the detector's frame rate, acquisition in PALM is relatively slow; limited by the total number of frames needed to acquire a complete super-resolved image. The affordable engineering of photoactivatable fluorophores and the applicability of PALM in TIRF and confocal (fPALM) configurations have encouraged its use in visualizing the dynamic architecture of the endocytic machinery[1], and, the nanoscopic oligomerization of hemagglutinin in biological membranes.

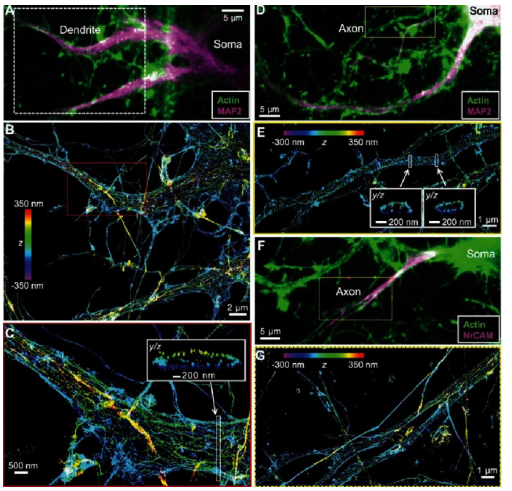
Stochastic Optical Reconstruction Microscopy (STORM)

The photobleaching of active fluorophores, in PALM, has been traditionally seen as an additional time consuming step, which if eliminated, would enable faster imaging. PALM and STORM share common operation principles, in particular, the random activation and localization of a subset of fluorophores. Nonetheless, the use of STORM is distinguished by the use of photoswitchable fluorophores which can be activated (turned on) and deactivated (turned off) at will. This improvement discards the requirement for successively photobleaching and activating fluorophores, thereby shortening the total acquisition time. STORM was first demonstrated using single fluorophores, however, the use of fluorophore-pairs was recently reported where, the emission of an activator fluorophore, following excitation, was shown to activate a photoswitchable reporter fluorophore; this process, when iterated, yields super-resolved images. STORM has been recently used in elucidating the mechanism of actions of the Nucleoid-associated proteins in chromosome organization, visualizing the architecture of apoptotic pores, and, the periodic arrangement of actin in axons.



Apoptosis, a form of programmed cell death, is fundamental to the development of multi-cellular organisms.

BAX, a pore-forming protein and key player in the cell suicide program, promotes apoptosis by irreversibly permeabilizing the mitochondrial outer membrane. The organization of BAX pores is unresolvable using diffraction-limited microscopy techniques. (A) AF647-anti-GFP nanobodies-labeled BAX in the outer membrane of apoptotic mitochondria (dashed white line) of HeLa cells imaged using dSTORM; a derivative of STORM that eliminates the requirement for using an activation laser by embedding the specimen in 'blinking buffer'. Scale bar = 5 µm. (B) enlarged region of (A) showing non-random structures. Scale bar = 500 nm. (C) Gallery of apoptotic pores forming varied structures: Dots, aggregates, lines, arcs, rings and double lines. Scale bar = 100 nm.

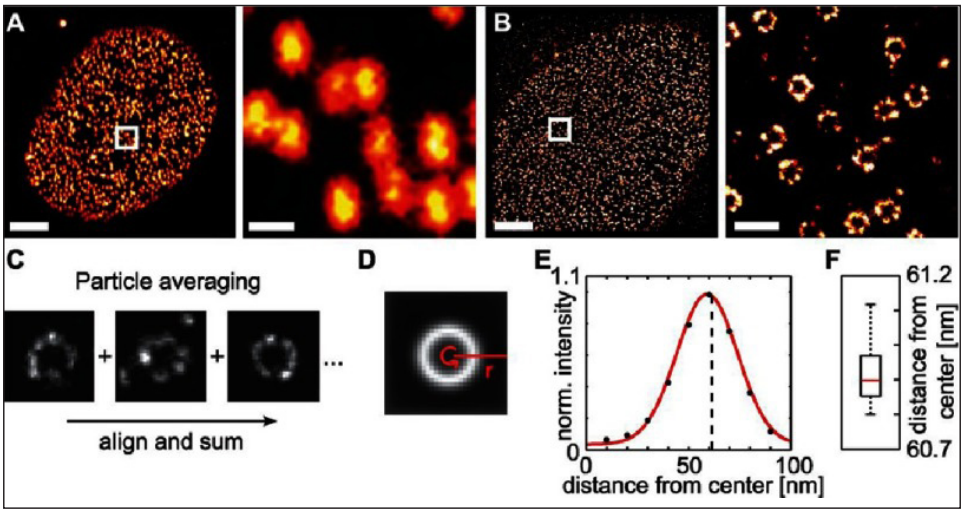


Spatial organization of actin in neuronal axons and dendrites imaged using STORM.

(A, F) Diffraction-limited images of Alexa-Fluor647-labeled actin filaments in hippocampal neurons. (B, C, E, G) Super-resolved enlargements of the areas marked in (A, B, D, F) revealing the periodic distribution of actin in neurons with a uniform spacing of 180 –190 nm.

Ground State Depletion followed by Individual Return Microscopy (GSDIM)

GSDIM relies in its operation on depleting the ground state by exciting all fluorophores in a field of view to an upper state. The excited electrons relax to a long-lived dark "off" state by irradiative emission. Individual molecules return back, stochastically, from the dark state to the ground state by emitting a photon. The stochastically-emitted fluorescence is localized and a complete super-resolved image of a sample is constructed following the return, and localization, of all the excited fluorophores. Similar to other super resolution microscopy techniques, GSDIM has been instrumental in facilitating numerous discoveries as, for instance, visualizing the structural organization of the nuclear pore complex.



Structure of the Nuclear Pore Complex (NPC) analyzed by super resolution microscopy and particle averaging.

GFP-tagged nucleoporins (NUPs) in the nuclear apical membrane of U2OS cells imaged using (A) confocal microscopy, and, (B) GSDIM. Scale bars = 1 μm (left), 300 nm (right). (C) Alignment and superposition of thousands of super-resolved NUPs results in a (D) circular form whose (E, F) average radius can be easily deduced. By labeling individual NUPs at different sites and measuring the radius of the computed circle, the alignment of the NUPs in individual NPCs can be deduced.

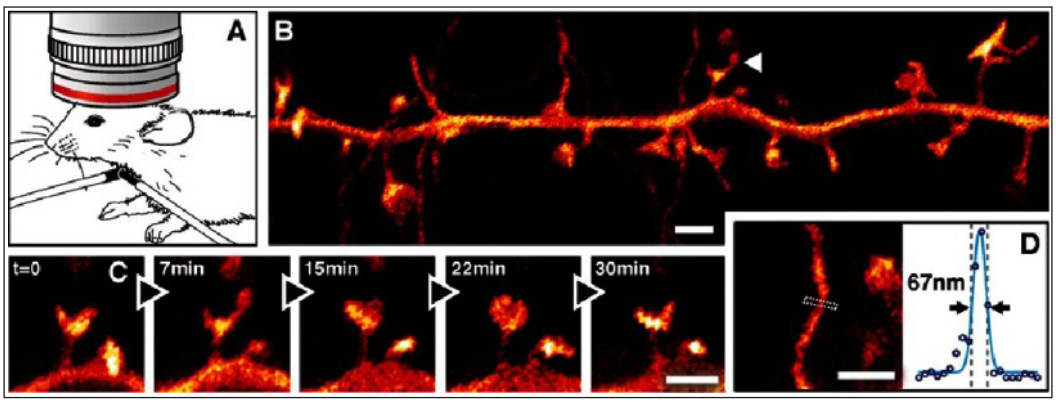
Stimulated Emission by Depletion (STED) Microscopy

As opposed to stochastic functional techniques like STORM and PALM, where a random subset of fluorophores is imaged at any one time, STED is a spatially-targeted

technique denoting that a few fluorophores at a certain position are chosen to be imaged at any one time. STED exploits the difference in the wavelength between spontaneous and stimulated emission. A collimated beam excites a large number of fluorophores in an optical section and another, doughnut shaped beam, of larger wavelength, centered at the excitation beam forcefully depresses the majority of the fluorophores to the ground state. The size of the annulus of the doughnut shaped beam can be adjusted down to molecular scales. The fluorophores in the middle of the doughnut-shaped beam, emit spontaneously whilst the forcefully depleted fluorophores emit stimulatingly.

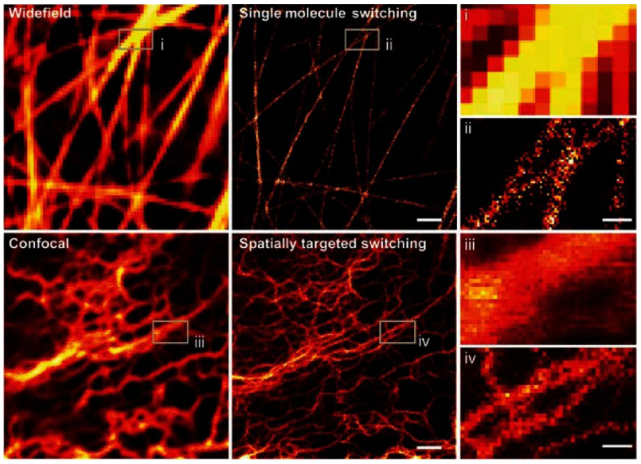
Since both emissions possess different wavelengths, they can be separated and the spontaneously emitted fluorescence signal is localized. Scanning the excitation, and stimulation, beams laterally across the sample yields a complete super-resolved image of the optical section. It can be shown that the obtained resolution using STED is inversely proportional to the intensity of the doughnut shaped beam. Therefore, the use of STED requires the engineering of special fluorophores that can withstand high laser powers and that the entire sample is not damaged at such high laser intensities.

STED has been used in the study of a number of biological systems including the clustering of the envelope protein during the budding of HIV-1 particles, the periodic organization of neural cytoskeleton[6]and the transient formation of cholesterol-mediated nanoscopic complexes in biological membranes. STED is the first super resolution microscopy technique to be successfully tested on living animals. Images obtained using diffraction-limited and super-resolution microscopies are shown in figure.



STED nanoscopy of the somatosensory cortex in the brain of a living mouse.

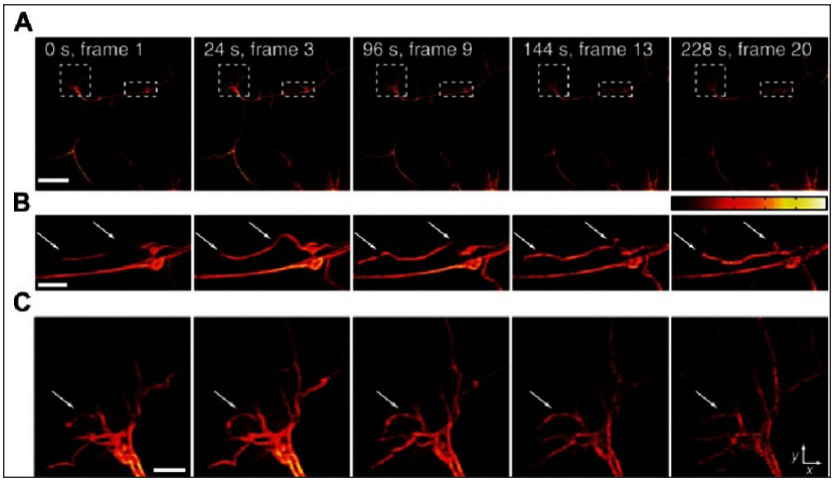
(A) Optical access to the cortex of anesthetized YFP-expressing heterozygous mice is facilitated by the presence of a cover glass inserted through a sealed hole in the skull of the mouse. (B) Super-resolved projection of neuronal structures. (C) Temporal dynamics of spine morphology. (D) Measurement of the width of a spine cross-section reveals a four-fold resolution improvement compared to diffraction-limited microscopy. Scale bar = 1 μm.



Living ptk2 cells expressing the fluorescent dye Dreiklang-Map2 imaged using:
(i) Epi-fluorescence microscopy, (ii) STORM, (iii) Confocal microscopy, and, (iv) STED microscopy.

Structured Illumination Microscopy (SIM)

As opposed to STED, where the true likeness of a sample is obtained after imaging, SIM is another super resolution technique that exploits the ability of periodic illumination to produce negative imprints of a sample. SIM employs a grid-like light pattern to illuminate a sample. The illumination pattern is periodic with high intensity regions a wavelength-separated from low intensity regions. The high intensity regions excite the underlying fluorophores leaving the fluorophores underlying the low intensity regions in the ground state. By rotating the illumination pattern, an image of dark regions on a bright background (negative imprint) can be obtained.



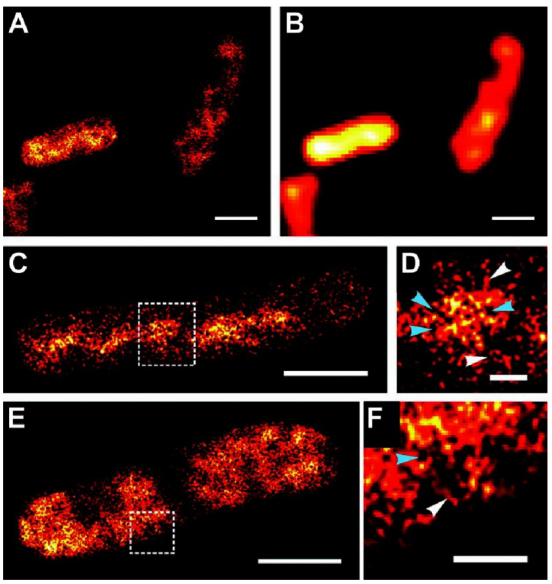
(A) Time series of filopodia-like structures expressing the fluorescent protein Lifeact-Dronpa imaged using parallelized STED nanoscopy.

Scale bar = 10 μm. (B, C) Enlargements of the regions marked in (A). Scale bar = 2 μm.

Binding-activation Localization Microscopy (BALM).

To obtain a true super-resolved image, the negative imprints undergo advanced image processing, therefore, SIM requires that the sample be immobilized during pattern rotation and on other hand, non-trivial computation expertise to deconvolve the obtained images. The speed of acquisition is fast but is generally limited by the speed of pattern rotation and the total number of images acquired during this process. Recently, however, SIM's pattern illumination scheme has inspired the development of a 100,000 folds faster variant of STED microscopy. The reported scheme relies on orthogonal illumination obtained by passing two STED beams across two perpendicularly aligned diffraction gratings that split the beams onto periodic positions across the sample plane. The recently developed technique has been demonstrated on the imaging of 100-squared microns of GFP-labeled Keratin at less than a second temporal resolution and the dynamics of filopodia-like structures at relevant time scales.

Dissimilar to PALM, STORM, STED and SSIM which can only provide positional information, BALM can also provide binding kinetics. BALM was developed with the aim of visualizing the organization of the bacterial chromosomes in E coli. The DNA binding dyes, YOYO-1 and PicoGreen, were used to visualize the organization of DNA strands at a resolution of 10 nm. The fluorescence of the DNA-binding dyes increases up to 800-fold upon binding. A complete super-resolved image of the sample under study is obtained by the repeated localization and photobleaching of the sequentially binding fluorophores. BALM was also used to visualize the structural organization of amyloid fibrils.



PicoGreen-labeled nucleoid structures in *E.Coli* imaged using (A) BALM and (B) Diffraction-limited microscopy.

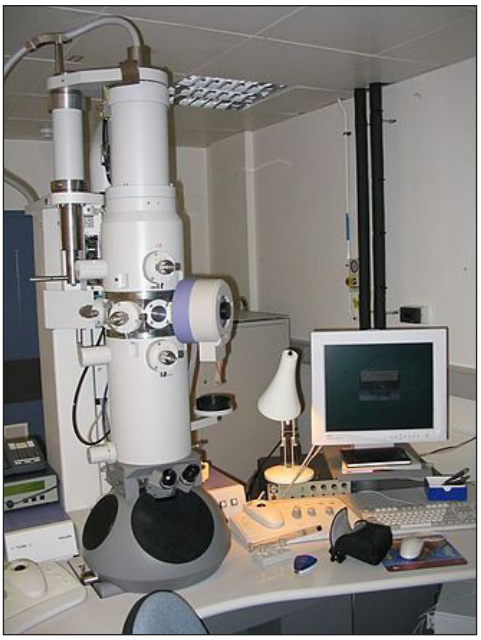
(C, E) Spatiotemporal organization of nucleoids during bacterial division. (D, F) are enlarged regions of (C, E) showing void regions. Scale bars = 1 μm (A - C, E), 200 nm (D, F).

Scanning Near Field Optical Microscopy (SNOM)

SNOM was developed to break the diffraction limit of light by abolishing the far field, lens-based, imaging system using a tip brought close to a sample, in the near field, to allow imaging at sub-diffraction limited resolutions. SNOM is commonly implemented in a TIRF configuration, whereby the evanescent field resulting from the total internal reflection of an incident beam can be detected and imaged using a fiber probe placed at a distance less than its penetration depth. The use of a scanning probe to construct a super resolved image of a sample implies that the acquisition rates using SNOM are limited by the probe scanning speed and that the sample, and probe, needs to be stabilized for good quality imaging. These requirements limit the use of SNOM in biological investigations.

Electron Microscope

An electron microscope is a microscope that uses a beam of accelerated electrons as a source of illumination. As the wavelength of an electron can be up to 100,000 times shorter than that of visible light photons, electron microscopes have a higher resolving power than light microscopes and can reveal the structure of smaller objects. A scanning transmission electron microscope has achieved better than 50 pm resolution in annular dark-field imaging mode and magnifications of up to about 10,000,000× whereas most light microscopes are limited by diffraction to about 200 nm resolution and useful magnifications below 2000×.



A modern transmission electron microscope.

Electron microscope constructed by Ernst Ruska.

Electron microscopes use shaped magnetic fields to form electron optical lens systems that are analogous to the glass lenses of an optical light microscope. Electron microscopes are used to investigate the ultrastructure of a wide range of biological and inorganic specimens including microorganisms, cells, large molecules, biopsy samples, metals, and crystals. Industrially, electron microscopes are often used for quality control and failure analysis. Modern electron microscopes produce electron micrographs using specialized digital cameras and frame grabbers to capture the images.

Types

Transmission Electron Microscope (TEM)

The original form of the electron microscope, the transmission electron microscope (TEM), uses a high voltage electron beam to illuminate the specimen and create an image. The electron beam is produced by an electron gun, commonly fitted with a tungsten filament cathode as the electron source. The electron beam is accelerated by an anode typically at +100 keV (40 to 400 keV) with respect to the cathode, focused by electrostatic and electromagnetic lenses, and transmitted through the specimen that is in part transparent to electrons and in part scatters them out of the beam. When it emerges from the specimen, the electron beam carries information about the structure of the specimen that is magnified by the objective lens system of the microscope. The spatial variation in this information (the “image”) may be viewed by projecting the magnified electron image onto a fluorescent viewing screen coated with a phosphor or scintillator material such as zinc sulfide. Alternatively, the image can be photographically

recorded by exposing a photographic film or plate directly to the electron beam, or a high-resolution phosphor may be coupled by means of a lens optical system or a fibre optic light-guide to the sensor of a digital camera. The image detected by the digital camera may be displayed on a monitor or computer.

The resolution of TEMs is limited primarily by spherical aberration, but a new generation of hardware correctors can reduce spherical aberration to increase the resolution in high-resolution transmission electron microscopy (HRTEM) to below 0.5 angstrom (50 picometres), enabling magnifications above 50 million times. The ability of HRTEM to determine the positions of atoms within materials is useful for nano-technologies research and development.

Transmission electron microscopes are often used in electron diffraction mode. The advantages of electron diffraction over X-ray crystallography are that the specimen need not be a single crystal or even a polycrystalline powder, and also that the Fourier transform reconstruction of the object's magnified structure occurs physically and thus avoids the need for solving the phase problem faced by the X-ray crystallographers after obtaining their X-ray diffraction patterns.

One major disadvantage of the transmission electron microscope is the need for extremely thin sections of the specimens, typically about 100 nanometers. Creating these thin sections for biological and materials specimens is technically very challenging. Semiconductor thin sections can be made using a focused ion beam. Biological tissue specimens are chemically fixed, dehydrated and embedded in a polymer resin to stabilize them sufficiently to allow ultrathin sectioning. Sections of biological specimens, organic polymers, and similar materials may require staining with heavy atom labels in order to achieve the required image contrast.

Serial-section Electron Microscopy (ssEM)

One application of TEM is serial-section electron microscopy (ssEM), for example in analyzing the connectivity in volumetric samples of brain tissue by imaging many thin sections in sequence.

Scanning Electron Microscope (SEM)

The SEM produces images by probing the specimen with a focused electron beam that is scanned across a rectangular area of the specimen (raster scanning). When the electron beam interacts with the specimen, it loses energy by a variety of mechanisms. The lost energy is converted into alternative forms such as heat, emission of low-energy secondary electrons and high-energy backscattered electrons, light emission (cathodoluminescence) or X-ray emission, all of which provide signals carrying information about the properties of the specimen surface, such as its topography and composition. The image displayed by an SEM maps the varying intensity of any of these signals into the

image in a position corresponding to the position of the beam on the specimen when the signal was generated. In the SEM image of an ant shown below and to the right, the image was constructed from signals produced by a secondary electron detector, the normal or conventional imaging mode in most SEMs.

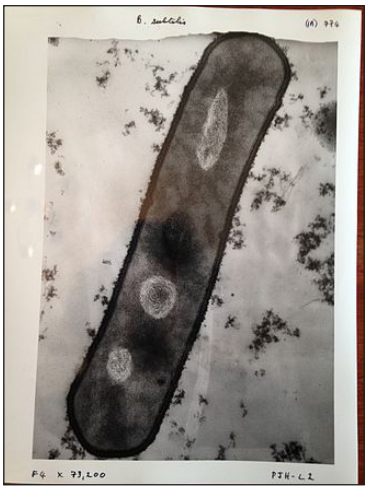
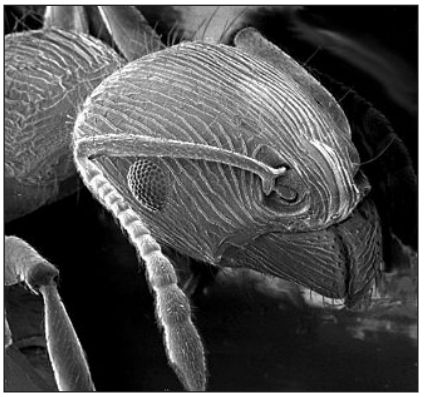


Image of *Bacillus subtilis* taken with a 1960s electron microscope.

Generally, the image resolution of an SEM is lower than that of a TEM. However, because the SEM images the surface of a sample rather than its interior, the electrons do not have to travel through the sample. This reduces the need for extensive sample preparation to thin the specimen to electron transparency. The SEM is able to image bulk samples that can fit on its stage and still be maneuvered, including a height less than the working distance being used, often 4 millimeters for high-resolution images. The SEM also has a great depth of field, and so can produce images that are good representations of the three-dimensional surface shape of the sample. Another advantage of SEMs comes with environmental scanning electron microscopes (ESEM) that can produce images of good quality and resolution with hydrated samples or in low, rather than high, vacuum or under chamber gases. This facilitates imaging unfixed biological samples that are unstable in the high vacuum of conventional electron microscopes.



An image of an ant in a scanning electron microscope.

Reflection Electron Microscope (REM)

In the reflection electron microscope (REM) as in the TEM, an electron beam is incident on a surface but instead of using the transmission (TEM) or secondary electrons (SEM), the reflected beam of elastically scattered electrons is detected. This technique is typically coupled with reflection high energy electron diffraction (RHEED) and *reflection high-energy loss spectroscopy (RHELS)*. Another variation is spin-polarized low-energy electron microscopy (SPLEEM), which is used for looking at the microstructure of magnetic domains.

Scanning Transmission Electron Microscope (STEM)

The STEM rasters a focused incident probe across a specimen that (as with the TEM) has been thinned to facilitate detection of electrons scattered *through* the specimen. The high resolution of the TEM is thus possible in STEM. The focusing action (and aberrations) occur before the electrons hit the specimen in the STEM, but afterward in the TEM. The STEMs use of SEM-like beam rastering simplifies annular dark-field imaging, and other analytical techniques, but also means that image data is acquired in serial rather than in parallel fashion. Often TEM can be equipped with the scanning option and then it can function both as TEM and STEM.

Scanning Tunneling Microscopy (STM)

In STM, a conductive tip held at a voltage is brought near a surface, and a profile can be obtained based on the tunneling probability of an electron from the tip to the sample since it is a function of distance.

Color

In their most common configurations, electron microscopes produce images with a single brightness value per pixel, with the results usually rendered in grayscale. However, often these images are then colorized through the use of feature-detection software, or simply by hand-editing using a graphics editor. This may be done to clarify structure or for aesthetic effect and generally does not add new information about the specimen.

In some configurations information about several specimen properties is gathered per pixel, usually by the use of multiple detectors. In SEM, the attributes of topography and material contrast can be obtained by a pair of backscattered electron detectors and such attributes can be superimposed in a single color image by assigning a different primary color to each attribute. Similarly, a combination of backscattered and secondary electron signals can be assigned to different colors and superimposed on a single color micrograph displaying simultaneously the properties of the specimen.

Some types of detectors used in SEM have analytical capabilities, and can provide several items of data at each pixel. Examples are the Energy-dispersive X-ray spectroscopy

(EDS) detectors used in elemental analysis and Cathodoluminescence microscope (CL) systems that analyse the intensity and spectrum of electron-induced luminescence in (for example) geological specimens. In SEM systems using these detectors, it is common to color code the signals and superimpose them in a single color image, so that differences in the distribution of the various components of the specimen can be seen clearly and compared. Optionally, the standard secondary electron image can be merged with the one or more compositional channels, so that the specimen's structure and composition can be compared. Such images can be made while maintaining the full integrity of the original signal, which is not modified in any way.

Sample Preparation

An insect coated in gold for viewing with a scanning electron microscope.

Materials to be viewed under an electron microscope may require processing to produce a suitable sample. The technique required varies depending on the specimen and the analysis required:

- Chemical fixation: For biological specimens aims to stabilize the specimen's mobile macromolecular structure by chemical crosslinking of proteins with aldehydes such as formaldehyde and glutaraldehyde, and lipids with osmium tetroxide.
- Negative stain: Suspensions containing nanoparticles or fine biological material (such as viruses and bacteria) are briefly mixed with a dilute solution of an electron-opaque solution such as ammonium molybdate, uranyl acetate (or formate), or phosphotungstic acid. This mixture is applied to a suitably coated EM grid, blotted, then allowed to dry. Viewing of this preparation in the TEM should be carried out without delay for best results. The method is important in microbiology for fast but crude morphological identification, but can also be used as the basis for high-resolution 3D reconstruction using EM tomography methodology when carbon films are used for support. Negative staining is also used for observation of nanoparticles.

- Cryofixation: Freezing a specimen so rapidly, in liquid ethane that the water forms vitreous (non-crystalline) ice. This preserves the specimen in a snapshot of its solution state. An entire field called cryo-electron microscopy has branched from this technique. With the development of cryo-electron microscopy of vitreous sections (CEMOVIS), it is now possible to observe samples from virtually any biological specimen close to its native state.
- Dehydration or replacement of water with organic solvents such as ethanol or acetone, followed by critical point drying or infiltration with embedding resins. Also freeze drying.
- Embedding, Biological specimens: After dehydration, tissue for observation in the transmission electron microscope is embedded so it can be sectioned ready for viewing. To do this the tissue is passed through a 'transition solvent' such as propylene oxide (epoxypropane) or acetone and then infiltrated with an epoxy resin such as Araldite, Epon, or Durcupan; tissues may also be embedded directly in water-miscible acrylic resin. After the resin has been polymerized (hardened) the sample is thin sectioned (ultrathin sections) and stained – it is then ready for viewing.
- Embedding materials: After embedding in resin, the specimen is usually ground and polished to a mirror-like finish using ultra-fine abrasives. The polishing process must be performed carefully to minimize scratches and other polishing artifacts that reduce image quality.
- Metal shadowing: Metal (e.g. platinum) is evaporated from an overhead electrode and applied to the surface of a biological sample at an angle. The surface topography results in variations in the thickness of the metal that are seen as variations in brightness and contrast in the electron microscope image.
- Replication: A surface shadowed with metal (e.g. platinum, or a mixture of carbon and platinum) at an angle is coated with pure carbon evaporated from carbon electrodes at right angles to the surface. This is followed by removal of the specimen material (e.g. in an acid bath, using enzymes or by mechanical separation) to produce a surface replica that records the surface ultrastructure and can be examined using transmission electron microscopy.
- Sectioning: Produces thin slices of the specimen, semitransparent to electrons. These can be cut on an ultramicrotome with a glass or diamond knife to produce ultra-thin sections about 60–90 nm thick. Disposable glass knives are also used because they can be made in the lab and are much cheaper.
- Staining: Uses heavy metals such as lead, uranium or tungsten to scatter imaging electrons and thus give contrast between different structures, since many (especially biological) materials are nearly "transparent" to electrons (weak

phase objects). In biology, specimens can be stained "en bloc" before embedding and also later after sectioning. Typically thin sections are stained for several minutes with an aqueous or alcoholic solution of uranyl acetate followed by aqueous lead citrate.

- Freeze-fracture or Freeze-etch: A preparation method particularly useful for examining lipid membranes and their incorporated proteins in "face on" view. The fresh tissue or cell suspension is frozen rapidly (cryofixation), then fractured by breaking or by using a microtome while maintained at liquid nitrogen temperature. The cold fractured surface (sometimes "etched" by increasing the temperature to about −100 °C for several minutes to let some ice sublime) is then shadowed with evaporated platinum or gold at an average angle of 45° in a high vacuum evaporator. The second coat of carbon, evaporated perpendicular to the average surface plane is often performed to improve the stability of the replica coating. The specimen is returned to room temperature and pressure, then the extremely fragile "pre-shadowed" metal replica of the fracture surface is released from the underlying biological material by careful chemical digestion with acids, hypochlorite solution or SDS detergent. The still-floating replica is thoroughly washed free from residual chemicals, carefully fished up on fine grids, dried then viewed in the TEM.
- Freeze-fracture replica immunogold labeling (FRIL): The freeze-fracture method has been modified to allow the identification of the components of the fracture face by immunogold labeling. Instead of removing all the underlying tissue of the thawed replica as the final step before viewing in the microscope the tissue thickness is minimized during or after the fracture process. The thin layer of tissue remains bound to the metal replica so it can be immunogold labeled with antibodies to the structures of choice. The thin layer of the original specimen on the replica with gold attached allows the identification of structures in the fracture plane. There are also related methods which label the surface of etched cells and other replica labeling variations.
- Ion beam milling: Thins samples until they are transparent to electrons by firing ions (typically argon) at the surface from an angle and sputtering material from the surface. A subclass of this is focused ion beam milling, where gallium ions are used to produce an electron transparent membrane in a specific region of the sample, for example through a device within a microprocessor. Ion beam milling may also be used for cross-section polishing prior to SEM analysis of materials that are difficult to prepare using mechanical polishing.
- Conductive coating: An ultrathin coating of electrically conducting material, deposited either by high vacuum evaporation or by low vacuum sputter coating of the sample. This is done to prevent the accumulation of static electric fields at the specimen due to the electron irradiation required during imaging. The coating materials include gold, gold/palladium, platinum, tungsten, graphite, etc.

- Earthing: To avoid electrical charge accumulation on a conductively coated sample, it is usually electrically connected to the metal sample holder. Often an electrically conductive adhesive is used for this purpose.

Disadvantages

Jeol Transmission and scanning electron microscope.

Electron microscopes are expensive to build and maintain, but the capital and running costs of confocal light microscope systems now overlaps with those of basic electron microscopes. Microscopes designed to achieve high resolutions must be housed in stable buildings (sometimes underground) with special services such as magnetic field canceling systems.

The samples largely have to be viewed in vacuum, as the molecules that make up air would scatter the electrons. An exception is liquid-phase electron microscopy using either a closed liquid cell or an environmental chamber, for example, in the environmental scanning electron microscope, which allows hydrated samples to be viewed in a low-pressure (up to 20 Torr or 2.7 kPa) wet environment. Various techniques for in situ electron microscopy of gaseous samples have been developed as well.

Scanning electron microscopes operating in conventional high-vacuum mode usually image conductive specimens; therefore non-conductive materials require conductive coating (gold/palladium alloy, carbon, osmium, etc.). The low-voltage mode of modern microscopes makes possible the observation of non-conductive specimens without coating. Non-conductive materials can be imaged also by a variable pressure (or environmental) scanning electron microscope.

Small, stable specimens such as carbon nanotubes, diatom frustules and small mineral crystals (asbestos fibres, for example) require no special treatment before being examined in the electron microscope. Samples of hydrated materials, including almost all biological specimens have to be prepared in various ways to stabilize them, reduce their thickness (ultrathin sectioning) and increase their electron optical contrast (staining). These processes may result in *artifacts*, but these can usually be identified by comparing the results obtained by using radically different specimen preparation methods. Since the 1980s, analysis of cryofixed, vitrified specimens has also become increasingly used by scientists, further confirming the validity of this technique.

X-ray Crystallography

X-ray crystallography is a technique used for determining the high-resolution, three-dimensional crystal structures of atom and molecules and has been fundamental in the development of many scientific fields. In its first decades of application, it is mainly used for determining the size of atoms, the lengths and types of chemical bonds, the atomic-scale differences among various materials, as well as the crystalline integrity, grain orientation, grain size, film thickness and interface roughness of the related materials, especially minerals and alloys. This method could also reveal the structure and function of many biological molecules like vitamins, drugs, proteins and nucleic acids. Up to date, it 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 explain the unusual electronic or elastic properties of a material, shed light on chemical interactions and processes, or function as the basis for designing pharmaceuticals against diseases. In particular, protein have been extensively put into structure determination by X-ray crystallography, which is also employed routinely in determining how a pharmaceutical drug interacts with its protein target and what changes might improve it. X-ray crystallography method has advantages of no damage to samples, free of pollution, low environmental requirements, high performance and precision over other measuring tools. These advantages make X-ray crystallography the most convenient and important manner to investigate the microstructure of materials.

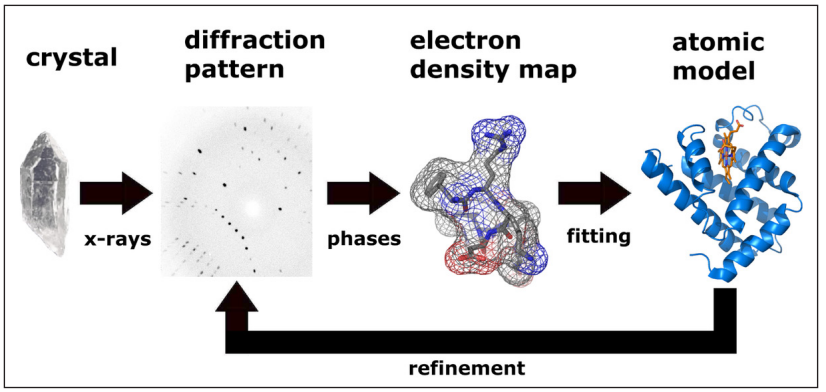
Principle

In a single-crystal X-ray diffraction measurement, a crystal is mounted on a goniometer, which is used to position the crystal at selected orientations. The crystal is illuminated with a finely focused monochromatic beam of X-rays, leading to a diffraction pattern of regularly spaced spots known as reflections. X-ray crystallography works in a manner of elastic scattering with the outgoing X-rays having the same energy and wavelength as the incoming X-rays, which get an altered direction after diffraction.

A crystallographer can then produce a three-dimensional picture of the density of electrons within the crystal by measuring the angles and intensities of these diffracted beams under the assistance of the mathematical method Fourier transforms. From this electron density, the mean positions of the atoms, chemical bonds, crystallographic disorder, and some other information in the crystal can be determined. Poor resolution or even errors may occur if the crystals are too small, or not uniform enough in their internal makeup.

Workflow

The technique of single crystal X-ray crystallography has three basic steps. The first and usually most difficult step is to produce an adequate crystal of the studied material. The crystal should be sufficiently large with all dimensions larger than 0.1 mm, pure in composition and regular in structure, and have no significant internal imperfections such as cracks or twinning. The crystal is subsequently placed in an intense beam of X-rays, usually of a single wavelength, to produce regular reflection pattern. The angles and intensities of diffracted X-rays are measured with each compound having a unique diffraction pattern. Previous reflections disappear and new ones appear along with the gradual rotation of the crystal, and the intensity of every spot is recorded at every orientation of the crystal. Multiple data sets may have to be collected since each set covers slightly more than half a full rotation of the crystal and typically contains tens of thousands of reflections. Ultimately, these collected data are combined computationally with complementary chemical information to obtain and refine a model from the arrangement of atoms within the crystal. The final refined model of the atomic arrangement is called a crystal structure and usually stored in a public database.



Workflow for solving the molecular structure by X-ray crystallography.

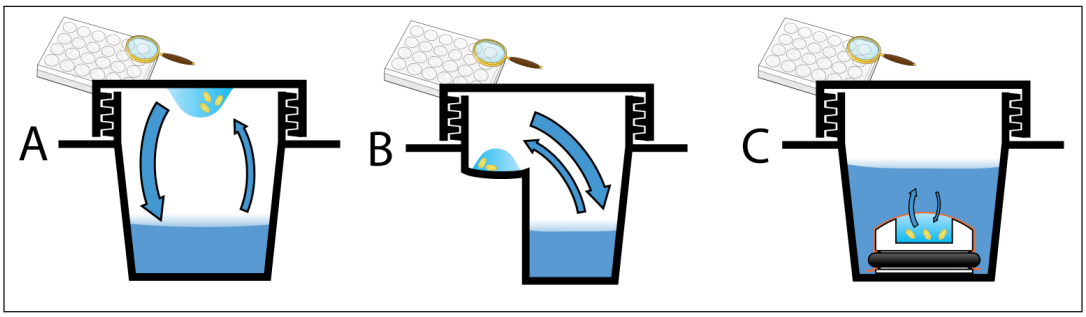
Procedure

Crystallization

In most cases, the generation of a diffraction-quality crystal is a primary barrier to solve its atomic-resolution structure. A pure crystal of high regularity is a general requirement

in crystallography to solve the structure of a complicated arrangement of atoms. There are many methods to cultivate crystal, such as gas diffusion, liquid phase diffusion, temperature gradient, vacuum sublimation, convection and so on, and the most widely adopted methodology is gas phase diffusion, which can be further divided into hanging drop, sitting drop, oil drop and microdialysis. The crystallography of small molecules and macromolecular differs in the range of possible techniques applied to produce diffraction-quality crystals. Small molecules have few degrees of conformational freedom, and can be crystallized by a wide range of methods. On the contrary, macromolecules have too many degrees of freedom to achieve a perfect crystallization so as to maintain a stable structure.

The range of crystallization conditions is also restricted to solution conditions where biomacromolecules remain folded configuration. There are several factors known to inhibit or ruin crystallization. Crystals generally grow at a constant temperature and are protected from shocks or vibrations that possibly disturb the crystallization. Impurities in the molecules or crystallization solutions are also inimical to crystallization. Molecules with high conformational flexibility or high tendcency to self-assemble into regular helices are often unwilling to assemble into crystals. A slight change in molecular properties can even lead to large differences in crystallization behavior. After acquiring the initial conditions of crystal growth, it is often necessary to optimize the crystallization conditions by adjusting precipitant concentration, pH value, sample concentration, temperature and ionic strength.



Three methods of preparing crystals: A Hanging drop; B Sitting drop; C Microdialysis.

Data Collection

Diffraction experiments are needed after obtaining single crystal. The X-ray irradiating to the crystal is diffracted, and the diffraction data are recorded. X-ray has are two main sources, one of which applied in the common X-ray instrument produces X-rays with multiple characteristic wavelengths by bombarding copper targets or molybdenum targets with high energy electron flow. Another one is the X-ray with variable wavelength generated through synchrotron radiation. X-rays from synchrotron radiation can be grouped into angular dispersion synchrotron radiation (ADXD) and energy dispersive synchrotron radiation (EDXRD). The experimental principle of ADXD is the same as that of the normal X-ray diffractometer, while the wavelength is lower

and the energy is higher. The incident light of EDXRD is white light with a continuous wavelength, and the diffraction signal is collected at a fixed angle. In comparison with ADXD, EDXRD also has a lower resolution ratio and technical requirement. Diffraction data, including location and intensity of diffraction points, are often by image plates or CCD detectors.

Data Analysis

The analysis of the recorded diffraction data could indicate the corresponding crystal system and Bravais lattice of crystal, and reveal the miller index and intensity of each diffraction point in the inverted space. Frist of all, indexation, strength integration, consolidation, and amplitude reduction are executed in data analysis procedure. Some problems in the determination of crystallographic parameters are then discussed, which is followed by common data collection and program processing. The intensity data of crystal diffraction finally undergoes a quality assessment.

Crystal Symmetry, Unit Cell and Image Scaling

Each recorded series of two-dimensional diffraction patterns corresponding to a different crystal orientation, is converted into a three-dimensional model of the electron density, which is completed by the mathematical technique of Fourier transforms. Each spot has a corresponding type of variation in the electron density and which variation corresponds to which spot (indexing) must be determined. 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) are also necessary to be figured out.

Data processing commences with the reflections indexation, which means identifying the dimensions of the unit cells and which image peak stands for which position in reciprocal space. A byproduct of indexing is to determine the crystal symmetry. The data is then integrated after having assigned symmetry. The hundreds of images containing the thousands of reflections are converted into a single file that consists of records of the miller index of each reflection and intensity for each reflection. These various images taken at different orientations of the crystal are merged and scaled firstly 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.

The optimization of intensity scale is critical for the peaks intensity since they are the key information from which the structure is determined. The repetitive technique of crystallographic data collection and the high symmetry of crystalline materials lead the diffractometer to repeatedly record many symmetry-equivalent reflections, allowing the calculation of symmetry-related *R*-factor, which is a reliable index based upon how similar are the measured intensities of symmetry-equivalent reflections, thus evaluating the quality of the data.

Initial Phasing

The data collected from a diffraction experiment represents a reciprocal space of the crystal lattice. The size and shape of the unit cell, and the inherent symmetry within the crystal govern the position of each diffraction 'spot', whose intensity is recorded and proportional to the square of the structure factor amplitude. The structure factor contains information involving both amplitude and phase of a wave, both of which must be known to obtain an interpretable electron density map that enables a crystallographer to build a starting model of the molecule. During a diffraction experiment, the phase cannot be directly recorded, which is known as the phase problem. The estimates of initial phase can be finished in a variety of ways such as *Ab initio* phasing, direct methods, molecular replacement, anomalous X-ray scattering and heavy atom methods.

Model Building and Phase Refinement

An initial model can be established after obtaining initial phases. This model can be applied to refine the phases, atomic positions and respective Debye-Waller factors with the aim of fitting the observed diffraction data, thus getting an improved model and ideally yielding a better set of phases. A new model can then be fit to a novel electron density map and a further round of refinement is performed, which continuously proceeds until the correlation between the diffraction data and the model is maximized. The agreement is measured by an *R*-factor defined as:

$$R = \frac{\sum \text{all reflections} \left| F_O - F_c \right|}{\sum \text{all reflection} \left| F_O \right|},$$

where *F* is the structural factor. A similar quality criterion R_{free} is calculated from a subset of reflections that are not included in the structure refinement. Model qualities including chemical bonding features of stereochemistry, hydrogen bonding and the distribution of bond lengths and angles are complementarily measured. Phase bias is a serious problem in such iterative model building, but can be checked by a common technique of omit maps. In many cases, crystallographic disorder smears the electron density map and weakly scattering atoms are routinely invisible. It is also likely that a single atom appears multiple times in an electron density map. In still other cases, the covalent structure deduced for the molecule is detected to be incorrect or changed.

Deposition of the Structure

Once the model of a molecular structure is finalized, it would be often deposited in crystallographic databases such as the Cambridge Structural Database for small molecules, the Inorganic Crystal Structure Database for inorganic compounds or the Protein Data Bank for protein structures.

NMR Spectroscopy

Nuclear magnetic resonance spectroscopy, most commonly known as NMR spectroscopy or magnetic resonance spectroscopy (MRS), is a spectroscopic technique to observe local magnetic fields around atomic nuclei. The sample is placed in a magnetic field and the NMR signal is produced by excitation of the nuclei sample with radio waves into nuclear magnetic resonance, which is detected with sensitive radio receivers. The intramolecular magnetic field around an atom in a molecule changes the resonance frequency, thus giving access to details of the electronic structure of a molecule and its individual functional groups. As the fields are unique or highly characteristic to individual compounds, in modern organic chemistry practice, NMR spectroscopy is the definitive method to identify monomolecular organic compounds. Similarly, biochemists use NMR to identify proteins and other complex molecules. Besides identification, NMR spectroscopy provides detailed information about the structure, dynamics, reaction state, and chemical environment of molecules. The most common types of NMR are proton and carbon-13 NMR spectroscopy, but it is applicable to any kind of sample that contains nuclei possessing spin.

A 900 MHz NMR instrument with a 21.1 T magnet at HWB-NMR.

NMR spectra are unique, well-resolved, analytically tractable and often highly predictable for small molecules. Different functional groups are obviously distinguishable, and identical functional groups with differing neighboring substituents still give distinguishable signals. NMR has largely replaced traditional wet chemistry tests such as color reagents or typical chromatography for identification. A disadvantage is that a relatively large amount, 2–50 mg, of a purified substance is required, although it may be recovered through a workup. Preferably, the sample should be dissolved in a solvent, because NMR analysis of solids requires a dedicated magic angle spinning machine and may not give equally well-resolved spectra. The timescale of NMR is relatively long,

and thus it is not suitable for observing fast phenomena, producing only an averaged spectrum. Although large amounts of impurities do show on an NMR spectrum, better methods exist for detecting impurities, as NMR is inherently not very sensitive - though at higher frequencies, sensitivity is higher.

Correlation spectroscopy is a development of ordinary NMR. In two-dimensional NMR, the emission is centered around a single frequency, and correlated resonances are observed. This allows identifying the neighboring substituents of the observed functional group, allowing unambiguous identification of the resonances. There are also more complex 3D and 4D methods and a variety of methods designed to suppress or amplify particular types of resonances. In nuclear Overhauser effect (NOE) spectroscopy, the relaxation of the resonances is observed. As NOE depends on the proximity of the nuclei, quantifying the NOE for each nucleus allows for construction of a three-dimensional model of the molecule.

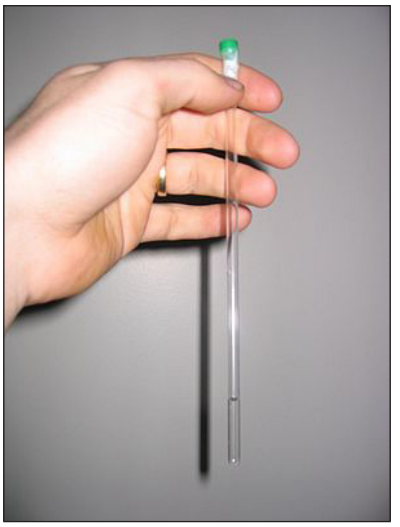
Cutaway of an NMR magnet that shows its structure: radiation shield, vacuum chamber, liquid nitrogen vessel, liquid helium vessel, and cryogenic shims.

NMR spectrometers are relatively expensive; universities usually have them, but they are less common in private companies. Between 2000 and 2015, an NMR spectrometer cost around 500,000 - 5 million USD. Modern NMR spectrometers have a very strong, large and expensive liquid helium-cooled superconducting magnet, because resolution directly depends on magnetic field strength. Less expensive machines using permanent magnets and lower resolution are also available, which still give sufficient performance for certain applications such as reaction monitoring and quick checking of samples. There are even benchtop nuclear magnetic resonance spectrometers. NMR can be observed in magnetic fields less than a millitesla. Low-resolution NMR produces broader peaks which can easily overlap one another causing issues in resolving complex structures. The use of higher strength magnetic fields result in clear resolution of the peaks and is the standard in industry.

NMR Techniques

Resonant Frequency

When placed in a magnetic field, NMR active nuclei (such as ^{1}H or ^{13}C) absorb electromagnetic radiation at a frequency characteristic of the isotope. The resonant frequency, energy of the radiation absorbed, and the intensity of the signal are proportional to the strength of the magnetic field. For example, in a 21 Tesla magnetic field, hydrogen atoms (commonly referred to as protons) resonate at 900 MHz. It is common to refer to a 21 T magnet as a 900 MHz magnet since hydrogen is the most common nucleus detected, however different nuclei will resonate at different frequencies at this field strength in proportion to their nuclear magnetic moments.



The NMR sample is prepared in a thin-walled glass tube - an NMR tube.

Sample Handling

An NMR spectrometer typically consists of a spinning sample-holder inside a very strong magnet, a radio-frequency emitter and a receiver with a probe (an antenna assembly) that goes inside the magnet to surround the sample, optionally gradient coils for diffusion measurements, and electronics to control the system. Spinning the sample is usually necessary to average out diffusional motion, however some experiments call for a stationary sample when solution movement is an important variable. For instance, measurements of diffusion constants (*diffusion ordered spectroscopy* or DOSY) are done using a stationary sample with spinning off, and flow cells can be used for online analysis of process flows.

Deuterated Solvents

The vast majority of molecules in a solution are solvent molecules, and most regular solvents are hydrocarbons and so contain NMR-active protons. In order to avoid

detecting only signals from solvent hydrogen atoms, deuterated solvents are used where 99+% of the protons are replaced with deuterium (hydrogen-2). The most widely used deuterated solvent is deuterochloroform ($CDCl_3$), although other solvents may be used for various reasons, such as solubility of a sample, desire to control hydrogen bonding, or melting or boiling points. The chemical shifts of a molecule will change slightly between solvents, and the solvent used will almost always be reported with chemical shifts. NMR spectra are often calibrated against the known solvent residual proton peak instead of added tetramethylsilane.

Shim and Lock

To detect the very small frequency shifts due to nuclear magnetic resonance, the applied magnetic field must be constant throughout the sample volume. High resolution NMR spectrometers use shims to adjust the homogeneity of the magnetic field to parts per billion (ppb) in a volume of a few cubic centimeters. In order to detect and compensate for inhomogeneity and drift in the magnetic field, the spectrometer maintains a "lock" on the solvent deuterium frequency with a separate lock unit, which is essentially an additional transmitter and RF processor tuned to the lock nucleus (deuterium) rather than the nuclei of the sample of interest. In modern NMR spectrometers shimming is adjusted automatically, though in some cases the operator has to optimize the shim parameters manually to obtain the best possible resolution.

Acquisition of Spectra

Upon excitation of the sample with a radio frequency (60–1000 MHz) pulse, a nuclear magnetic resonance response - a free induction decay (FID) - is obtained. It is a very weak signal, and requires sensitive radio receivers to pick up. A Fourier transform is carried out to extract the frequency-domain spectrum from the raw time-domain FID. A spectrum from a single FID has a low signal-to-noise ratio, but it improves readily with averaging of repeated acquisitions. Good ^{1}H NMR spectra can be acquired with 16 repeats, which takes only minutes. However, for elements heavier than hydrogen, the relaxation time is rather long, e.g. around 8 seconds for ^{13}C. Thus, acquisition of quantitative heavy-element spectra can be time-consuming, taking tens of minutes to hours.

Following the pulse, the nuclei are, on average, excited to a certain angle vs. the spectrometer magnetic field. The extent of excitation can be controlled with the pulse width, typically ca. 3-8 μs for the optimal 90° pulse. The pulse width can be determined by plotting the (signed) intensity as a function of pulse width. It follows a sine curve, and accordingly, changes sign at pulse widths corresponding to 180° and 360° pulses.

Decay times of the excitation, typically measured in seconds, depend on the effectiveness of relaxation, which is faster for lighter nuclei and in solids, and slower for heavier nuclei and in solutions, and they can be very long in gases. If the second excitation

pulse is sent prematurely before the relaxation is complete, the average magnetization vector has not decayed to ground state, which affects the strength of the signal in an unpredictable manner. In practice, the peak areas are then not proportional to the stoichiometry; only the presence, but not the amount of functional groups is possible to discern. An inversion recovery experiment can be done to determine the relaxation time and thus the required delay between pulses. A 180° pulse, an adjustable delay, and a 90° pulse is transmitted. When the 90° pulse exactly cancels out the signal, the delay corresponds to the time needed for 90° of relaxation. Inversion recovery is worthwhile for quantitive ^{13}C, 2D and other time-consuming experiments.

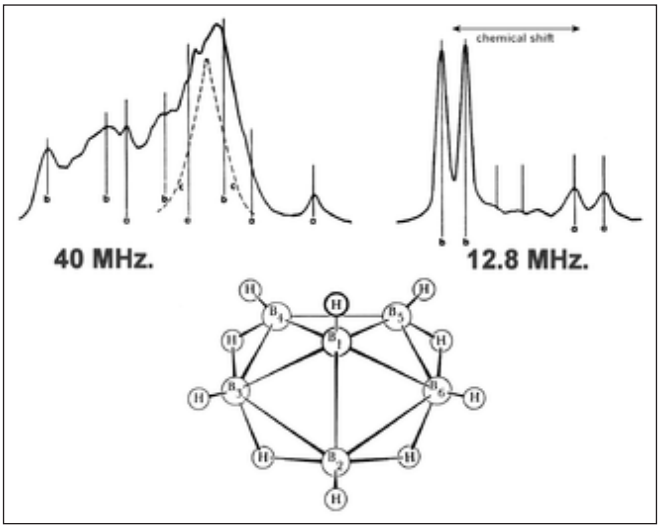
Chemical Shift

A spinning charge generates a magnetic field that results in a magnetic moment proportional to the spin. In the presence of an external magnetic field, two spin states exist (for a spin 1/2 nucleus): One spin up and one spin down, where one aligns with the magnetic field and the other opposes it. The difference in energy (ΔE) between the two spin states increases as the strength of the field increases, but this difference is usually very small, leading to the requirement for strong NMR magnets (1-20 T for modern NMR instruments). Irradiation of the sample with energy corresponding to the exact spin state separation of a specific set of nuclei will cause excitation of those set of nuclei in the lower energy state to the higher energy state.

For spin 1/2 nuclei, the energy difference between the two spin states at a given magnetic field strength is proportional to their magnetic moment. However, even if all protons have the same magnetic moments, they do not give resonant signals at the same frequency values. This difference arises from the differing electronic environments of the nucleus of interest. Upon application of an external magnetic field, these electrons move in response to the field and generate local magnetic fields that oppose the much stronger applied field. This local field thus "shields" the proton from the applied magnetic field, which must therefore be increased in order to achieve resonance (absorption of rf energy). Such increments are very small, usually in parts per million (ppm). For instance, the proton peak from an aldehyde is shifted ca. 10 ppm compared to a hydrocarbon peak, since as an electron-withdrawing group, the carbonyl deshields the proton by reducing the local electron density. The difference between 2.3487 T and 2.3488 T is therefore about 42 ppm. However a frequency scale is commonly used to designate the NMR signals, even though the spectrometer may operate by sweeping the magnetic field, and thus the 42 ppm is 4200 Hz for a 100 MHz reference frequency (rf).

However, given that the location of different NMR signals is dependent on the external magnetic field strength and the reference frequency, the signals are usually reported relative to a reference signal, usually that of TMS (tetramethylsilane). Additionally, since the distribution of NMR signals is field dependent, these frequencies are divided by the spectrometer frequency. However, since we are dividing Hz by MHz, the resulting number would be too small, and thus it is multiplied by a million. This operation

therefore gives a locator number called the "chemical shift" with units of parts per million. In general, chemical shifts for protons are highly predictable since the shifts are primarily determined by simpler shielding effects (electron density), but the chemical shifts for many heavier nuclei are more strongly influenced by other factors including excited states ("paramagnetic" contribution to shielding tensor).



Example of the chemical shift: NMR spectrum of hexaborane B_6H_{10} showing peaks shifted in frequency, which give clues as to the molecular structure.

The chemical shift provides information about the structure of the molecule. The conversion of the raw data to this information is called *assigning* the spectrum. For example, for the ^{1}H-NMR spectrum for ethanol (CH_3CH_2OH), one would expect signals at each of three specific chemical shifts: one for the C*H*$_3$ group, one for the C*H*$_2$ group and one for the O*H* group. A typical CH_3 group has a shift around 1 ppm, a CH_2 attached to an OH has a shift of around 4 ppm and an OH has a shift anywhere from 2–6 ppm depending on the solvent used and the amount of hydrogen bonding. While the O atom does draw electron density away from the attached H through their mutual sigma bond, the electron lone pairs on the O bathe the H in their shielding effect.

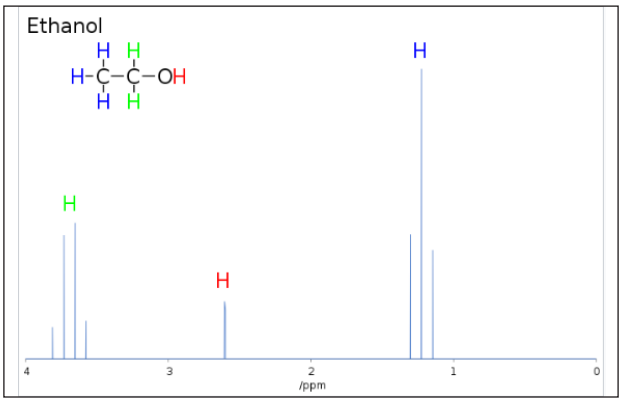
In paramagnetic NMR spectroscopy, measurements are conducted on paramagnetic samples. The paramagnetism gives rise to very diverse chemical shifts. In ^{1}H NMR spectroscopy, the chemical shift range can span up to thousands of ppm.

Because of molecular motion at room temperature, the three methyl protons *average out* during the NMR experiment (which typically requires a few ms). These protons become degenerate and form a peak at the same chemical shift.

The shape and area of peaks are indicators of chemical structure too. In the example above—the proton spectrum of ethanol—the CH_3 peak has three times the area of the OH peak. Similarly the CH_2 peak would be twice the area of the OH peak but only 2/3 the area of the CH_3 peak.

Software allows analysis of signal intensity of peaks, which under conditions of optimal relaxation, correlate with the number of protons of that type. Analysis of signal intensity is done by integration—the mathematical process that calculates the area under a curve. The analyst must integrate the peak and not measure its height because the peaks also have *width*—and thus its size is dependent on its area not its height. However, it should be mentioned that the number of protons, or any other observed nucleus, is only proportional to the intensity, or the integral, of the NMR signal in the very simplest one-dimensional NMR experiments. In more elaborate experiments, for instance, experiments typically used to obtain carbon-13 NMR spectra, the integral of the signals depends on the relaxation rate of the nucleus, and its scalar and dipolar coupling constants. Very often these factors are poorly known - therefore, the integral of the NMR signal is very difficult to interpret in more complicated NMR experiments.

J-coupling



Example ^{1}H NMR spectrum (1-dimensional) of ethanol plotted as signal intensity vs. chemical shift. There are three different types of H atoms in ethanol regarding NMR. The hydrogen (H) on the -OH group is not coupling with the other H atoms and appears as a singlet, but the CH_3- and the -CH_2- hydrogens are coupling with each other, resulting in a triplet and quartet respectively.

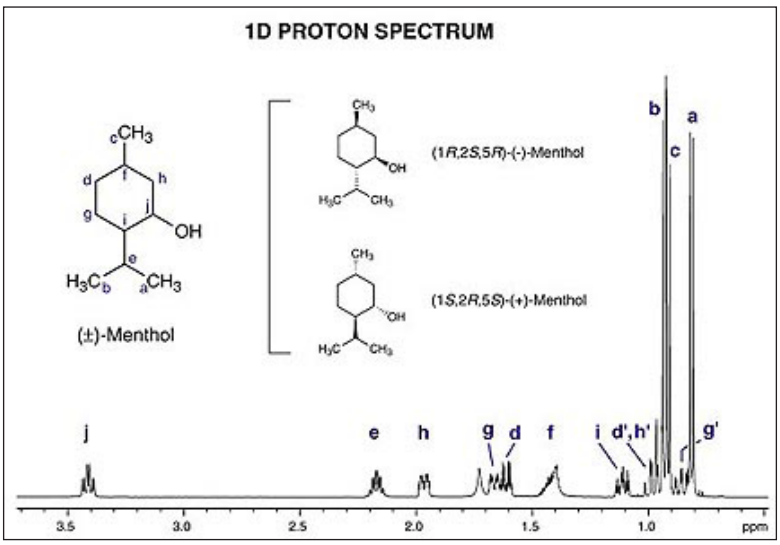
Multiplicity	Intensity Ratio
Singlet (s)	1
Doublet (d)	1:1
Triplet (t)	1:2:1
Quartet (q)	1:3:3:1
Quintet	1:4:6:4:1
Sextet	1:5:10:10:5:1
Septet	1:6:15:20:15:6:1

Some of the most useful information for structure determination in a one-dimensional NMR spectrum comes from J-coupling or scalar coupling (a special case of spin-spin

coupling) between NMR active nuclei. This coupling arises from the interaction of different spin states through the chemical bonds of a molecule and results in the splitting of NMR signals. For a proton, the local magnetic field is slightly different depending on whether an adjacent nucleus points towards or against the spectrometer magnetic field, which gives rise to two signals per proton instead of one. These splitting patterns can be complex or simple and, likewise, can be straightforwardly interpretable or deceptive. This coupling provides detailed insight into the connectivity of atoms in a molecule.

Coupling to n equivalent (spin ½) nuclei splits the signal into a n+1 multiplet with intensity ratios following Pascal's triangle. Coupling to additional spins will lead to further splittings of each component of the multiplet e.g. coupling to two different spin ½ nuclei with significantly different coupling constants will lead to a *doublet of doublets* (abbreviation: dd). Note that coupling between nuclei that are chemically equivalent (that is, have the same chemical shift) has no effect on the NMR spectra and couplings between nuclei that are distant (usually more than 3 bonds apart for protons in flexible molecules) are usually too small to cause observable splittings. *Long-range* couplings over more than three bonds can often be observed in cyclic and aromatic compounds, leading to more complex splitting patterns.

For example, in the proton spectrum for ethanol, the CH_3 group is split into a *triplet* with an intensity ratio of 1:2:1 by the two neighboring CH_2 protons. Similarly, the CH_2 is split into a *quartet* with an intensity ratio of 1:3:3:1 by the three neighboring CH_3 protons. In principle, the two CH_2 protons would also be split again into a *doublet* to form a *doublet of quartets* by the hydroxyl proton, but intermolecular exchange of the acidic hydroxyl proton often results in a loss of coupling information.



Coupling to any spin ½ nuclei such as phosphorus-31 or fluorine-19 works in this fashion (although the magnitudes of the coupling constants may be very different). But the splitting patterns differ from for nuclei with spin greater than ½ because the

spin quantum number has more than two possible values. For instance, coupling to deuterium (a spin 1 nucleus) splits the signal into a *1:1:1 triplet* because the spin 1 has three spin states. Similarly, a spin 3/2 nucleus splits a signal into a *1:1:1:1 quartet* and so on.

Coupling combined with the chemical shift (and the integration for protons) tells us not only about the chemical environment of the nuclei, but also the number of *neighboring* NMR active nuclei within the molecule. In more complex spectra with multiple peaks at similar chemical shifts or in spectra of nuclei other than hydrogen, coupling is often the only way to distinguish different nuclei.

^{1}H NMR spectrum of menthol with chemical shift in ppm on the horizontal axis. Each magnetically inequivalent proton has a characteristic shift, and couplings to other protons appear as splitting of the peaks into multiplets: e.g. peak *a*, because of the three magnetically equivalent protons in methyl group *a*, couple to one adjacent proton (*e*) and thus appears as a doublet.

Second-order (or Strong) Coupling

The coupling constant is small in comparison with the difference in NMR frequencies between the inequivalent spins. If the shift separation decreases (or the coupling strength increases), the multiplet intensity patterns are first distorted, and then become more complex and less easily analyzed (especially if more than two spins are involved). Intensification of some peaks in a multiplet is achieved at the expense of the remainder, which sometimes almost disappear in the background noise, although the integrated area under the peaks remains constant. In most high-field NMR, however, the distortions are usually modest and the characteristic distortions (*roofing*) can in fact help to identify related peaks.

Some of these patterns can be analyzed with the method published by John Pople, though it has limited scope. Second-order effects decrease as the frequency difference between multiplets increases, so that high-field (i.e. high-frequency) NMR spectra display less distortion than lower frequency spectra. Early spectra at 60 MHz were more prone to distortion than spectra from later machines typically operating at frequencies at 200 MHz or above.

Magnetic Inequivalence

More subtle effects can occur if chemically equivalent spins (i.e., nuclei related by symmetry and so having the same NMR frequency) have different coupling relationships to external spins. Spins that are chemically equivalent but are not indistinguishable (based on their coupling relationships) are termed magnetically inequivalent. For example, the 4 H sites of 1,2-dichlorobenzene divide into two chemically equivalent pairs by symmetry, but an individual member of one of the pairs has different couplings to the spins making up the other pair. Magnetic inequivalence can lead to highly complex

spectra which can only be analyzed by computational modeling. Such effects are more common in NMR spectra of aromatic and other non-flexible systems, while conformational averaging about C-C bonds in flexible molecules tends to equalize the couplings between protons on adjacent carbons, reducing problems with magnetic inequivalence.

Correlation Spectroscopy

Correlation spectroscopy is one of several types of two-dimensional nuclear magnetic resonance (NMR) spectroscopy or 2D-NMR. This type of NMR experiment is best known by its acronym, COSY. Other types of two-dimensional NMR include J-spectroscopy, exchange spectroscopy (EXSY), Nuclear Overhauser effect spectroscopy (NOESY), total correlation spectroscopy (TOCSY) and heteronuclear correlation experiments, such as HSQC, HMQC, and HMBC. In correlation spectroscopy, emission is centered on the peak of an individual nucleus; if its magnetic field is correlated with another nucleus by through-bond (COSY, HSQC, etc.) or through-space (NOE) coupling, a response can also be detected on the frequency of the correlated nucleus. Two-dimensional NMR spectra provide more information about a molecule than one-dimensional NMR spectra and are especially useful in determining the structure of a molecule, particularly for molecules that are too complicated to work with using one-dimensional NMR. The first two-dimensional experiment, COSY, was proposed by Jean Jeener, a professor at Université Libre de Bruxelles, in 1971. This experiment was later implemented by Walter P. Aue, Enrico Bartholdi and Richard R. Ernst, who published their work in 1976.

Solid-state Nuclear Magnetic Resonance

Solid-state 900 MHz (21.1 T) NMR spectrometer at
the Canadian National Ultrahigh-field NMR Facility for Solids.

A variety of physical circumstances do not allow molecules to be studied in solution, and at the same time not by other spectroscopic techniques to an atomic level, either. In solid-phase media, such as crystals, microcrystalline powders, gels, anisotropic solutions, etc., it is in particular the dipolar coupling and chemical shift anisotropy that become dominant to the behaviour of the nuclear spin systems. In conventional

solution-state NMR spectroscopy, these additional interactions would lead to a significant broadening of spectral lines. A variety of techniques allows establishing high-resolution conditions, that can, at least for ^{13}C spectra, be comparable to solution-state NMR spectra.

Two important concepts for high-resolution solid-state NMR spectroscopy are the limitation of possible molecular orientation by sample orientation, and the reduction of anisotropic nuclear magnetic interactions by sample spinning. Of the latter approach, fast spinning around the magic angle is a very prominent method, when the system comprises spin 1/2 nuclei. Spinning rates of ca. 20 kHz are used, which demands special equipment. A number of intermediate techniques, with samples of partial alignment or reduced mobility, is currently being used in NMR spectroscopy.

Applications in which solid-state NMR effects occur are often related to structure investigations on membrane proteins, protein fibrils or all kinds of polymers, and chemical analysis in inorganic chemistry, but also include "exotic" applications like the plant leaves and fuel cells. For example, Rahmani et al. studied the effect of pressure and temperature on the bicellar structures' self-assembly using deuterium NMR spectroscopy.

Biomolecular NMR Spectroscopy

Proteins

Much of the innovation within NMR spectroscopy has been within the field of protein NMR spectroscopy, an important technique in structural biology. A common goal of these investigations is to obtain high resolution 3-dimensional structures of the protein, similar to what can be achieved by X-ray crystallography. In contrast to X-ray crystallography, NMR spectroscopy is usually limited to proteins smaller than 35 kDa, although larger structures have been solved. NMR spectroscopy is often the only way to obtain high resolution information on partially or wholly intrinsically unstructured proteins. It is now a common tool for the determination of Conformation Activity Relationships where the structure before and after interaction with, for example, a drug candidate is compared to its known biochemical activity. Proteins are orders of magnitude larger than the small organic molecules but the basic NMR techniques and some NMR theory also applies. Because of the much higher number of atoms present in a protein molecule in comparison with a small organic compound, the basic 1D spectra become crowded with overlapping signals to an extent where direct spectral analysis becomes untenable. Therefore, multidimensional (2, 3 or 4D) experiments have been devised to deal with this problem. To facilitate these experiments, it is desirable to isotopically label the protein with ^{13}C and ^{15}N because the predominant naturally occurring isotope ^{12}C is not NMR-active and the nuclear quadrupole moment of the predominant naturally occurring ^{14}N isotope prevents high resolution information from being obtained from this nitrogen isotope. The most important method used for structure determination of proteins utilizes NOE experiments to measure distances

between atoms within the molecule. Subsequently, the distances obtained are used to generate a 3D structure of the molecule by solving a distance geometry problem. NMR can also be used to obtain information on the dynamics and conformational flexibility of different regions of a protein.

Nucleic Acids

"Nucleic acid NMR" is the use of NMR spectroscopy to obtain information about the structure and dynamics of polynucleic acids, such as DNA or RNA. As of 2003, nearly half of all known RNA structures had been determined by NMR spectroscopy.

Nucleic acid and protein NMR spectroscopy are similar but differences exist. Nucleic acids have a smaller percentage of hydrogen atoms, which are the atoms usually observed in NMR spectroscopy, and because nucleic acid double helices are stiff and roughly linear, they do not fold back on themselves to give "long-range" correlations. The types of NMR usually done with nucleic acids are ^{1}H or proton NMR, ^{13}C NMR, ^{15}N NMR, and ^{31}P NMR. Two-dimensional NMR methods are almost always used, such as correlation spectroscopy (COSY) and total coherence transfer spectroscopy (TOCSY) to detect through-bond nuclear couplings, and nuclear Overhauser effect spectroscopy (NOESY) to detect couplings between nuclei that are close to each other in space.

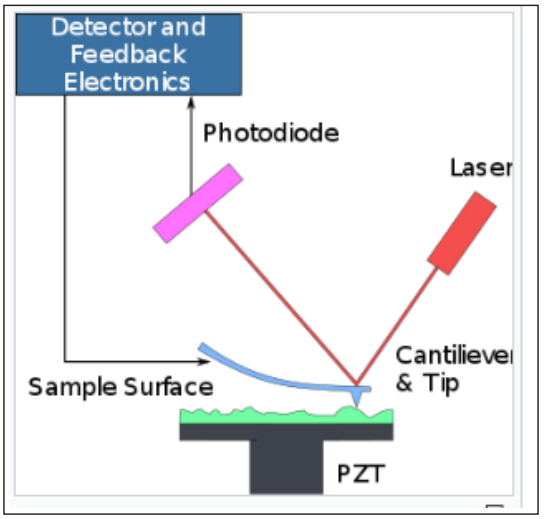
Parameters taken from the spectrum, mainly NOESY cross-peaks and coupling constants, can be used to determine local structural features such as glycosidic bond angles, dihedral angles (using the Karplus equation), and sugar pucker conformations. For large-scale structure, these local parameters must be supplemented with other structural assumptions or models, because errors add up as the double helix is traversed, and unlike with proteins, the double helix does not have a compact interior and does not fold back upon itself. NMR is also useful for investigating nonstandard geometries such as bent helices, non-Watson–Crick basepairing, and coaxial stacking. It has been especially useful in probing the structure of natural RNA oligonucleotides, which tend to adopt complex conformations such as stem-loops and pseudoknots. NMR is also useful for probing the binding of nucleic acid molecules to other molecules, such as proteins or drugs, by seeing which resonances are shifted upon binding of the other molecule.

Carbohydrates

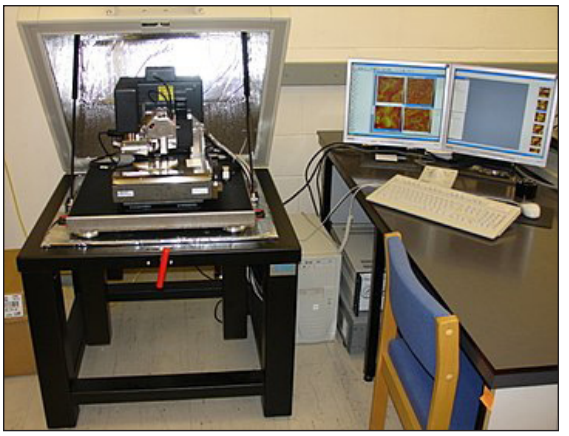
Carbohydrate NMR spectroscopy addresses questions on the structure and conformation of carbohydrates. The analysis of carbohydrates by 1H NMR is challenging due to the limited variation in functional groups, which leads to 1H resonances concentrated in narrow bands of the NMR spectrum. In other words, there is poor spectral dispersion. The anomeric proton resonances are segregated from the others due to fact that the anomeric carbons bear two oxygen atoms. For smaller carbohydrates, the dispersion of the anomeric proton resonances facilitates the use of 1D TOCSY experiments to investigate the entire spin systems of individual carbohydrate residues.

Atomic Force Microscopy

Atomic force microscopy (AFM) or scanning force microscopy (SFM) is a very-high-resolution type of scanning probe microscopy (SPM), with demonstrated resolution on the order of fractions of a nanometer, more than 1000 times better than the optical diffraction limit.



The probe on the end of the cantilever is scanned over the surface changing the amount of laser light reflected into the photodiode. The height of the cantilever is then adjusted to restore the response signal resulting in the measured cantilever displacement tracing the surface.



An atomic force microscope on the left with controlling computer on the right.

Atomic force microscopy (AFM) is a type of scanning probe microscopy (SPM), with demonstrated resolution on the order of fractions of a nanometer, more than 1000 times better than the optical diffraction limit. The information is gathered by “feeling” or

"touching" the surface with a mechanical probe. Piezoelectric elements that facilitate tiny but accurate and precise movements on (electronic) command enable precise scanning.

Abilities

The AFM has three major abilities: Force measurement, topographic imaging, and manipulation. In force measurement, AFMs can be used to measure the forces between the probe and the sample as a function of their mutual separation. This can be applied to perform force spectroscopy, to measure the mechanical properties of the sample, such as the sample's Young's modulus, a measure of stiffness.

For imaging, the reaction of the probe to the forces that the sample imposes on it can be used to form an image of the three-dimensional shape (topography) of a sample surface at a high resolution. This is achieved by raster scanning the position of the sample with respect to the tip and recording the height of the probe that corresponds to a constant probe-sample interaction. The surface topography is commonly displayed as a pseudocolor plot. Although the initial publication about the atomic force microscopy by Binnig, Quate and Gerber in 1986 speculated about the possibility of achieving atomic resolution, profound experimental challenges needed to be overcome before atomic resolution of defects and step edges in ambient (liquid) conditions was demonstrated in 1993 by Ohnesorge and Binnig. True atomic resolution of the silicon 7x7 surface - the atomic images of this surface obtained by STM had convinced the scientific community of the spectacular spatial resolution of scanning tunneling microscopy - had to wait a little longer before it was shown by Giessibl.

In manipulation, the forces between tip and sample can also be used to change the properties of the sample in a controlled way. Examples of this include atomic manipulation, scanning probe lithography and local stimulation of cells. Simultaneous with the acquisition of topographical images, other properties of the sample can be measured locally and displayed as an image, often with similarly high resolution. Examples of such properties are mechanical properties like stiffness or adhesion strength and electrical properties such as conductivity or surface potential. In fact, the majority of SPM techniques are extensions of AFM that use this modality.

Other Microscopy Technologies

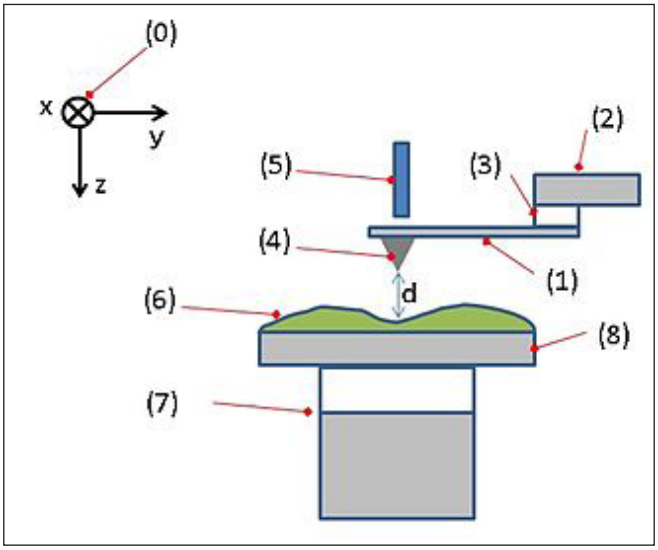
The major difference between atomic force microscopy and competing technologies such as optical microscopy and electron microscopy is that AFM does not use lenses or beam irradiation. Therefore, it does not suffer from a limitation in spatial resolution due to diffraction and aberration, and preparing a space for guiding the beam (by creating a vacuum) and staining the sample are not necessary.

There are several types of scanning microscopy including scanning probe microscopy (which includes AFM, scanning tunneling microscopy (STM) and near-field scanning

optical microscope (SNOM/NSOM), STED microscopy (STED), and scanning electron microscopy and Electrochemical AFM, EC-AFM). Although SNOM and STED use visible, infrared or even terahertz light to illuminate the sample, their resolution is not constrained by the diffraction limit.

Configuration

Figure shows an AFM, which typically consists of the following features. Numbers in parentheses correspond to numbered features in figure. Coordinate directions are defined by the coordinate system (0).



Typical configuration of an AFM. (1) Cantilever, (2) upport for cantilever, (3) Piezoelectric element(to oscillate cantilever at its eigen frequency.), (4) Tip (Fixed to open end of a cantilever, acts as the probe), (5) Detector of deflection and motion of the cantilever, (6) Sample to be measured by AFM, (7) xyz drive, (moves sample (6) and stage (8) in x, y, and z directions with respect to a tip apex (4)), and (8) Stage.

The small spring-like cantilever: (1) Carried by the support (2). Optionally, a piezoelectric element (typically made of a ceramic material) (3) Oscillates the cantilever (1). The sharp tip (4) Fixed to the free end of the cantilever (1). The detector (5) Records the deflection and motion of the cantilever (1). The sample (6) Mounted on the sample stage (8). An xyz drive (7) Permits to displace the sample (6) and the sample stage (8) in x, y, and z directions with respect to the tip apex (4). Although figure shows the drive attached to the sample, the drive can also be attached to the tip, or independent drives can be attached to both, since it is the relative displacement of the sample and tip that needs to be controlled. Controllers and plotter are not shown in figure.

According to the configuration, the interaction between tip and sample, which can be an atomic scale phenomenon, is transduced into changes of the motion of cantilever

which is a macro scale phenomenon. Several different aspects of the cantilever motion can be used to quantify the interaction between the tip and sample, most commonly the value of the deflection, the amplitude of an imposed oscillation of the cantilever, or the shift in resonance frequency of the cantilever.

Detector

The detector of AFM measures the deflection (displacement with respect to the equilibrium position) of the cantilever and converts it into an electrical signal. The intensity of this signal will be proportional to the displacement of the cantilever.

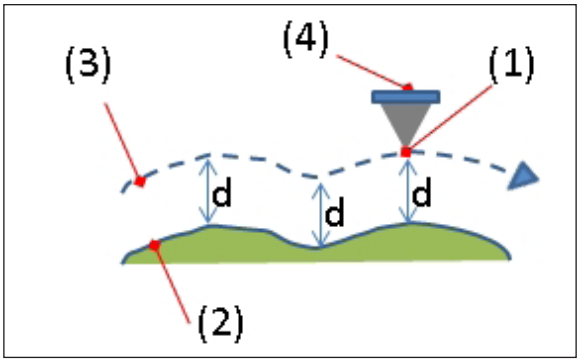
Various methods of detection can be used, e.g. interferometry, optical levers, the piezoresistive method, the piezoelectric method, and STM-based detectors.

Image Formation

When using the AFM to image a sample, the tip is brought into contact with the sample, and the sample is raster scanned along an x-y grid. Most commonly, an electronic feedback loop is employed to keep the probe-sample force constant during scanning. This feedback loop has the cantilever deflection as input, and its output controls the distance along the z axis between the probe support and the sample support. As long as the tip remains in contact with the sample, and the sample is scanned in the x-y plane, height variations in the sample will change the deflection of the cantilever. The feedback then adjusts the height of the probe support so that the deflection is restored to a user-defined value (the setpoint). A properly adjusted feedback loop adjusts the support-sample separation continuously during the scanning motion, such that the deflection remains approximately constant. In this situation, the feedback output equals the sample surface topography to within a small error.

Historically, a different operation method has been used, in which the sample-probe support distance is kept constant and not controlled by a feedback (servo mechanism). In this mode, usually referred to as ‘constant height mode’, the deflection of the cantilever is recorded as a function of the sample x-y position. As long as the tip is in contact with the sample, the deflection then corresponds to surface topography. The main reason this method is not very popular anymore, is that the forces between tip and sample are not controlled, which can lead to forces high enough to damage the tip or the sample. It is however common practice to record the deflection even when scanning in ‘constant force mode’, with feedback. This reveals the small tracking error of the feedback, and can sometimes reveal features that the feedback was not able to adjust for.

The AFM signals, such as sample height or cantilever deflection, are recorded on a computer during the x-y scan. They are plotted in a pseudocolor image, in which each pixel represents an x-y position on the sample, and the color represents the recorded signal.



Topographic image forming by AFM. (1): Tip apex, (2): Sample surface, (3): Z-orbit of Tip apex, (4): Cantilever.

The AFM was invented by IBM scientists in 1982. The precursor to the AFM, the scanning tunneling microscope (STM), was developed by Gerd Binnig and Heinrich Rohrer in the early 1980s at IBM Research - Zurich, a development that earned them the 1986 Nobel Prize for Physics. Binnig invented the atomic force microscope and the first experimental implementation was made by Binnig, Quate and Gerber in 1986.

The first commercially available atomic force microscope was introduced in 1989. The AFM is one of the foremost tools for imaging, measuring, and manipulating matter at the nanoscale.

Applications

The AFM has been applied to problems in a wide range of disciplines of the natural sciences, including solid-state physics, semiconductor science and technology, molecular engineering, polymer chemistry and physics, surface chemistry, molecular biology, cell biology, and medicine.

Applications in the field of solid state physics include: (a) The identification of atoms at a surface, (b) the evaluation of interactions between a specific atom and its neighboring atoms, and (c) the study of changes in physical properties arising from changes in an atomic arrangement through atomic manipulation.

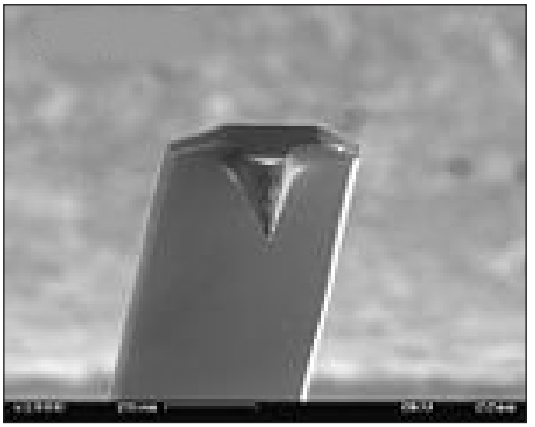
In molecular biology, AFM can be used to study the structure and mechanical properties of protein complexes and assemblies. For example, AFM has been used to image microtubules and measure their stiffness.

In cellular biology, AFM can be used to attempt to distinguish cancer cells and normal cells based on a hardness of cells, and to evaluate interactions between a specific cell and its neighboring cells in a competitive culture system. AFM can also be used to indent cells, to study how they regulate the stiffness or shape of the cell membrane or wall.

In some variations, electric potentials can also be scanned using conducting cantilevers. In more advanced versions, currents can be passed through the tip to probe the

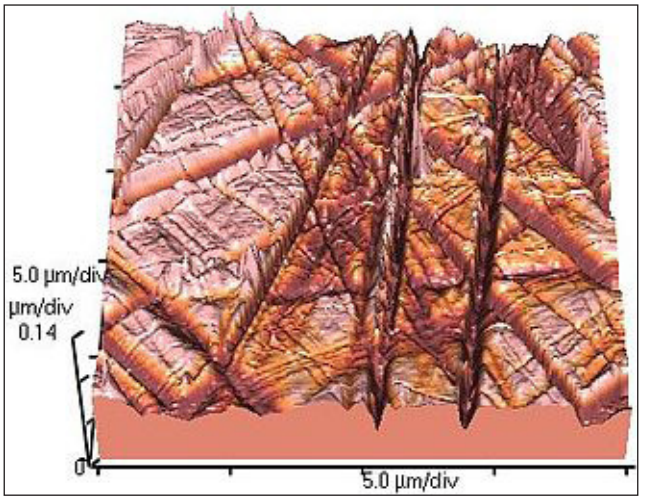
electrical conductivity or transport of the underlying surface, but this is a challenging task with few research groups reporting consistent data.

Principles



Electron micrograph of a used AFM cantilever.

The AFM consists of a cantilever with a sharp tip (probe) at its end that is used to scan the specimen surface. The cantilever is typically silicon or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought into proximity of a sample surface, forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law. Depending on the situation, forces that are measured in AFM include mechanical contact force, van der Waals forces, capillary forces, chemical bonding, electrostatic forces, magnetic forces (see magnetic force microscope, MFM), Casimir forces, solvation forces, etc. Along with force, additional quantities may simultaneously be measured through the use of specialized types of probes.



Atomic force microscope topographical scan of a glass surface. The micro and nano-scale features of the glass can be observed, portraying the roughness of the material. The image space is (x,y,z) = (20 µm × 20 µm × 420 nm).

The AFM can be operated in a number of modes, depending on the application. In general, possible imaging modes are divided into static (also called *contact*) modes and a variety of dynamic (non-contact or "tapping") modes where the cantilever is vibrated or oscillated at a given frequency.

Imaging Modes

AFM operation is usually described as one of three modes, according to the nature of the tip motion: contact mode, also called static mode (as opposed to the other two modes, which are called dynamic modes); tapping mode, also called intermittent contact, AC mode, or vibrating mode, or, after the detection mechanism, amplitude modulation AFM; non-contact mode, or, again after the detection mechanism, frequency modulation AFM.

Despite the nomenclature, repulsive contact can occur or be avoided both in amplitude modulation AFM and frequency modulation AFM, depending on the settings.

Contact Mode

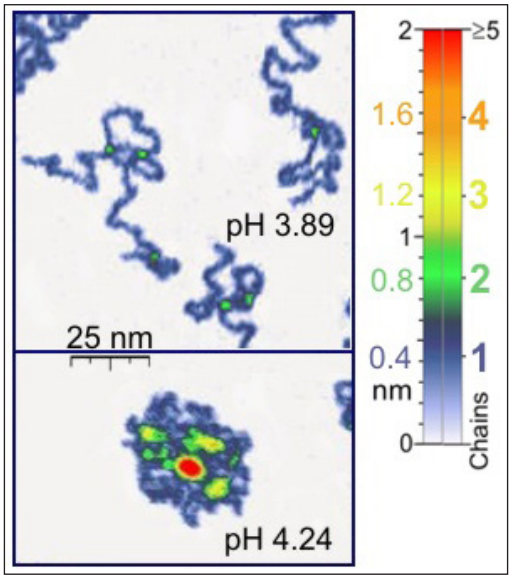
In contact mode, the tip is "dragged" across the surface of the sample and the contours of the surface are measured either using the deflection of the cantilever directly or, more commonly, using the feedback signal required to keep the cantilever at a constant position. Because the measurement of a static signal is prone to noise and drift, low stiffness cantilevers (i.e. cantilevers with a low spring constant, k) are used to achieve a large enough deflection signal while keeping the interaction force low. Close to the surface of the sample, attractive forces can be quite strong, causing the tip to "snap-in" to the surface. Thus, contact mode AFM is almost always done at a depth where the overall force is repulsive, that is, in firm "contact" with the solid surface.

Tapping Mode

In ambient conditions, most samples develop a liquid meniscus layer. Because of this, keeping the probe tip close enough to the sample for short-range forces to become detectable while preventing the tip from sticking to the surface presents a major problem for contact mode in ambient conditions. Dynamic contact mode (also called intermittent contact, AC mode or tapping mode) was developed to bypass this problem. Nowadays, tapping mode is the most frequently used AFM mode when operating in ambient conditions or in liquids.

In tapping mode, the cantilever is driven to oscillate up and down at or near its resonance frequency. This oscillation is commonly achieved with a small piezo element in the cantilever holder, but other possibilities include an AC magnetic field (with magnetic cantilevers), piezoelectric cantilevers, or periodic heating with a modulated laser beam. The amplitude of this oscillation usually varies from several nm to 200 nm. In tapping mode, the frequency and amplitude of the driving signal are kept constant, leading to a constant amplitude of the cantilever oscillation as long as there is no drift or interaction with the surface. The interaction of forces acting on the cantilever when the tip comes

close to the surface, Van der Waals forces, dipole-dipole interactions, electrostatic forces, etc. cause the amplitude of the cantilever's oscillation to change (usually decrease) as the tip gets closer to the sample. This amplitude is used as the parameter that goes into the electronic servo that controls the height of the cantilever above the sample. The servo adjusts the height to maintain a set cantilever oscillation amplitude as the cantilever is scanned over the sample. A *tapping AFM* image is therefore produced by imaging the force of the intermittent contacts of the tip with the sample surface.



Single polymer chains (0.4 nm thick) recorded in a tapping mode under aqueous media with different pH.

Although the peak forces applied during the contacting part of the oscillation can be much higher than typically used in contact mode, tapping mode generally lessens the damage done to the surface and the tip compared to the amount done in contact mode. This can be explained by the short duration of the applied force, and because the lateral forces between tip and sample are significantly lower in tapping mode over contact mode. Tapping mode imaging is gentle enough even for the visualization of supported lipid bilayers or adsorbed single polymer molecules (for instance, 0.4 nm thick chains of synthetic polyelectrolytes) under liquid medium. With proper scanning parameters, the conformation of single molecules can remain unchanged for hours, and even single molecular motors can be imaged while moving.

When operating in tapping mode, the phase of the cantilever's oscillation with respect to the driving signal can be recorded as well. This signal channel contains information about the energy dissipated by the cantilever in each oscillation cycle. Samples that contain regions of varying stiffness or with different adhesion properties can give a contrast in this channel that is not visible in the topographic image. Extracting the sample's material properties in a quantitative manner from phase images, however, is often not feasible.

Non-contact Mode

In non-contact atomic force microscopy mode, the tip of the cantilever does not contact the sample surface. The cantilever is instead oscillated at either its resonant frequency (frequency modulation) or just above (amplitude modulation) where the amplitude of oscillation is typically a few nanometers (<10 nm) down to a few picometers. The van der Waals forces, which are strongest from 1 nm to 10 nm above the surface, or any other long-range force that extends above the surface acts to decrease the resonance frequency of the cantilever. This decrease in resonant frequency combined with the feedback loop system maintains a constant oscillation amplitude or frequency by adjusting the average tip-to-sample distance. Measuring the tip-to-sample distance at each (x,y) data point allows the scanning software to construct a topographic image of the sample surface.

Non-contact mode AFM does not suffer from tip or sample degradation effects that are sometimes observed after taking numerous scans with contact AFM. This makes non-contact AFM preferable to contact AFM for measuring soft samples, e.g. biological samples and organic thin film. In the case of rigid samples, contact and non-contact images may look the same. However, if a few monolayers of adsorbed fluid are lying on the surface of a rigid sample, the images may look quite different. An AFM operating in contact mode will penetrate the liquid layer to image the underlying surface, whereas in non-contact mode an AFM will oscillate above the adsorbed fluid layer to image both the liquid and surface.

Schemes for dynamic mode operation include frequency modulation where a phase-locked loop is used to track the cantilever's resonance frequency and the more common amplitude modulation with a servo loop in place to keep the cantilever excitation to a defined amplitude. In frequency modulation, changes in the oscillation frequency provide information about tip-sample interactions. Frequency can be measured with very high sensitivity and thus the frequency modulation mode allows for the use of very stiff cantilevers. Stiff cantilevers provide stability very close to the surface and, as a result, this technique was the first AFM technique to provide true atomic resolution in ultra-high vacuum conditions.

In amplitude modulation, changes in the oscillation amplitude or phase provide the feedback signal for imaging. In amplitude modulation, changes in the phase of oscillation can be used to discriminate between different types of materials on the surface. Amplitude modulation can be operated either in the non-contact or in the intermittent contact regime. In dynamic contact mode, the cantilever is oscillated such that the separation distance between the cantilever tip and the sample surface is modulated.

Amplitude modulation has also been used in the non-contact regime to image with atomic resolution by using very stiff cantilevers and small amplitudes in an ultra-high vacuum environment.

Topographic Image

Image formation is a plotting method that produces a color mapping through changing the x-y position of the tip while scanning and recording the measured variable, i.e. the intensity of control signal, to each x-y coordinate. The color mapping shows the measured value corresponding to each coordinate. The image expresses the intensity of a value as a hue. Usually, the correspondence between the intensity of a value and a hue is shown as a color scale in the explanatory notes accompanying the image.

What is the Topographic Image of Atomic Force Microscope?

Operation mode of image forming of the AFM are generally classified into two groups from the viewpoint whether it uses z-Feedback loop (not shown) to maintain the tip-sample distance to keep signal intensity exported by the detector. The first one (using z-Feedback loop), said to be "constant **XX** mode" (**XX** is something which kept by z-Feedback loop).

Topographic image formation mode is based on abovementioned "constant **XX** mode", z-Feedback loop controls the relative distance between the probe and the sample through outputting control signals to keep constant one of frequency, vibration and phase which typically corresponds to the motion of cantilever (for instance, voltage is applied to the Z-piezoelectric element and it moves the sample up and down towards the Z direction.

Topographic Image of FM-AFM

When the distance between the probe and the sample is brought to the range where atomic force may be detected, while a cantilever is excited in its natural eigen frequency (f_0), a phenomenon occurs that the resonance frequency (f) of the cantilever shifts from its original resonance frequency (natural eigen frequency). In other words, in the range where atomic force may be detected, the frequency shift ($df=f-f_0$) will be observed. So, when the distance between the probe and the sample is in the non-contact region, the frequency shift increases in negative direction as the distance between the probe and the sample gets smaller.

When the sample has concavity and convexity, the distance between the tip-apex and the sample varies in accordance with the concavity and convexity accompanied with a scan of the sample along x-y direction (without height regulation in z-direction). As a result, the frequency shift arises. The image in which the values of the frequency obtained by a raster scan along the x-y direction of the sample surface are plotted against the x-y coordination of each measurement point is called a constant-height image.

On the other hand, the df may be kept constant by moving the probe upward and downward in z-direction using a negative feedback (by using z-feedback loop) while the

raster scan of the sample surface along the x-y direction. The image in which the amounts of the negative feedback (the moving distance of the probe upward and downward in z-direction) are plotted against the x-y coordination of each measurement point is a topographic image. In other words, the topographic image is a trace of the tip of the probe regulated so that the df is constant and it may also be considered to be a plot of a constant-height surface of the df.

Therefore, the topographic image of the AFM is not the exact surface morphology itself, but actually the image influenced by the bond-order between the probe and the sample, however, the topographic image of the AFM is considered to reflect the geographical shape of the surface more than the topographic image of a scanning tunnel microscope.

Force Spectroscopy

Another major application of AFM (besides imaging) is force spectroscopy, the direct measurement of tip-sample interaction forces as a function of the gap between the tip and sample (the result of this measurement is called a force-distance curve). For this method, the AFM tip is extended towards and retracted from the surface as the deflection of the cantilever is monitored as a function of piezoelectric displacement. These measurements have been used to measure nanoscale contacts, atomic bonding, Van der Waals forces, and Casimir forces, dissolution forces in liquids and single molecule stretching and rupture forces. Furthermore, AFM was used to measure, in an aqueous environment, the dispersion force due to polymer adsorbed on the substrate. Forces of the order of a few piconewtons can now be routinely measured with a vertical distance resolution of better than 0.1 nanometers. Force spectroscopy can be performed with either static or dynamic modes. In dynamic modes, information about the cantilever vibration is monitored in addition to the static deflection.

Problems with the technique include no direct measurement of the tip-sample separation and the common need for low-stiffness cantilevers, which tend to ‘snap’ to the surface. These problems are not insurmountable. An AFM that directly measures the tip-sample separation has been developed. The snap-in can be reduced by measuring in liquids or by using stiffer cantilevers, but in the latter case a more sensitive deflection sensor is needed. By applying a small dither to the tip, the stiffness (force gradient) of the bond can be measured as well.

Biological Applications and Other

Force spectroscopy is used in biophysics to measure the mechanical properties of living material (such as tissue or cells) or detect structures of different stiffness buried into the bulk of the sample using the stiffness tomography. Another application was to measure the interaction forces between from one hand a material stuck on the tip of

the cantilever, and from another hand the surface of particles either free or occupied by the same material. From the adhesion force distribution curve, a mean value of the forces has been derived. It allowed to make a cartography of the surface of the particles, covered or not by the material.

Identification of Individual Surface Atoms

The AFM can be used to image and manipulate atoms and structures on a variety of surfaces. The atom at the apex of the tip "senses" individual atoms on the underlying surface when it forms incipient chemical bonds with each atom. Because these chemical interactions subtly alter the tip's vibration frequency, they can be detected and mapped. This principle was used to distinguish between atoms of silicon, tin and lead on an alloy surface, by comparing these 'atomic fingerprints' to values obtained from large-scale density functional theory (DFT) simulations.

The trick is to first measure these forces precisely for each type of atom expected in the sample, and then to compare with forces given by DFT simulations. The team found that the tip interacted most strongly with silicon atoms, and interacted 24% and 41% less strongly with tin and lead atoms, respectively. Thus, each different type of atom can be identified in the matrix as the tip is moved across the surface.

Probe

An AFM probe has a sharp tip on the free-swinging end of a cantilever that is protruding from a holder. The dimensions of the cantilever are in the scale of micrometers. The radius of the tip is usually on the scale of a few nanometers to a few tens of nanometers. (Specialized probes exist with much larger end radii, for example probes for indentation of soft materials). The cantilever holder, also called holder chip – often 1.6 mm by 3.4 mm in size – allows the operator to hold the AFM cantilever/probe assembly with tweezers and fit it into the corresponding holder clips on the scanning head of the atomic force microscope.

This device is most commonly called an "AFM probe", but other names include "AFM tip" and "cantilever" (employing the name of a single part as the name of the whole device). An AFM probe is a particular type of SPM (scanning probe microscopy) probe.

AFM probes are manufactured with MEMS technology. Most AFM probes used are made from silicon (Si), but borosilicate glass and silicon nitride are also in use. AFM probes are considered consumables as they are often replaced when the tip apex becomes dull or contaminated or when the cantilever is broken. They can cost from a couple of tens of dollars up to hundreds of dollars per cantilever for the most specialized cantilever/probe combinations.

Just the tip is brought very close to the surface of the object under investigation, the cantilever is deflected by the interaction between the tip and the surface, which is what

the AFM is designed to measure. A spatial map of the interaction can be made by measuring the deflection at many points on a 2D surface.

Several types of interaction can be detected. Depending on the interaction under investigation, the surface of the tip of the AFM probe needs to be modified with a coating. Among the coatings used are gold – for covalent bonding of biological molecules and the detection of their interaction with a surface, diamond for increased wear resistance and magnetic coatings for detecting the magnetic properties of the investigated surface. Another solution exists to achieve high resolution magnetic imaging : having the probe equip with a microSQUID. The AFM tips is fabricated using silicon micro machining and the precise positioning of the microSQUID loop is done by electron beam lithography.

The surface of the cantilevers can also be modified. These coatings are mostly applied in order to increase the reflectance of the cantilever and to improve the deflection signal.

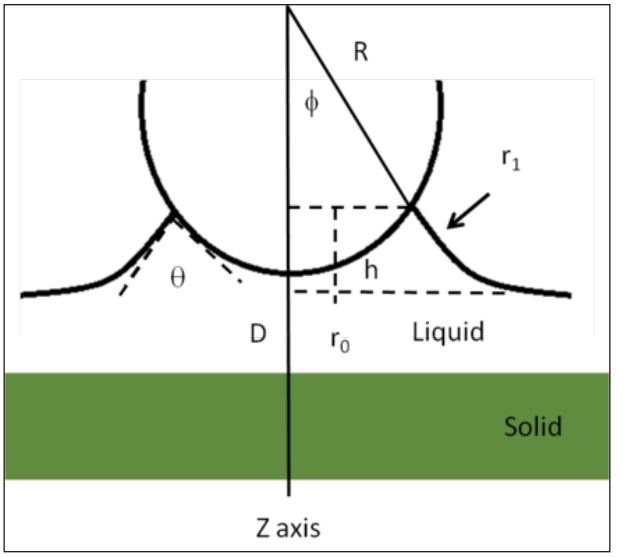
Forces vs. Tip Geometry

The forces between the tip and the sample strongly depend on the geometry of the tip. Various studies were exploited in the past years to write the forces as a function of the tip parameters.

Among the different forces between the tip and the sample, the water meniscus forces are highly interesting, both in air and liquid environment. Other forces must be considered, like the Coulomb force, van der Waals forces, double layer interactions, solvation forces, hydration and hydrophobic forces.

Water Meniscus

Water meniscus forces are highly interesting for AFM measurements in air. Due to the ambient humidity, a thin layer of water is formed between the tip and the sample during air measumements. The resulting capillary force gives rise to a strong attractive force that pulls the tip onto the surface. In fact, the adhesion force measured between tip and sample in ambient air of finite humidity is usually dominated by capillary forces. As a consequence, it is difficult to pull the tip away from the surface. For soft samples including many polymers and in particular biological materials, the strong adhesive capillary force gives rise to sample degradation and destruction upon imaging in contact mode. Historically, these problems were an important motivation for the development of dynamic imaging in air (e.g. ‘tapping mode’). During tapping mode imaging in air, capillary bridges still form. Yet, for suitable imaging conditions, the capillary bridges are formed and broken in every oscillation cycle of the cantilever normal to the surface, as can be inferred from an analysis of cantilever amplitude and phase vs. distance curves. As a consequence, destructive shear forces are largely reduced and soft samples can be investigated.



Model for AFM water meniscus.

In order to quantify the equilibrium capillary force, it is necessary to start from the Laplace equation for pressure:

$$P = \gamma_L \left(\frac{1}{r_1} + \frac{1}{r_o}\right) \simeq \frac{\gamma_L}{r_{eff}},$$

where γ_L is the surface energy and r_o and r_1 are defined in the figure.

The pressure is applied on an area of:

$$A \simeq 2\pi R \simeq [r_{eff}(1+\cos\theta)+h],$$

where d, θ, and h are defined in the figure.

The force which pulles together the two surfaces is:

$$F = 2\pi R\gamma_L\left(1+\cos\theta+\frac{h}{r_{eff}}\right).$$

The same formula could also be calculated as a function of relative humidity.

Gao calculated formulas for different tip geometries. As an example, the forse decreases by 20% for a conical tip with respect to a spherical tip.

When these forces are calculated, a difference must be made between the wet on dry situation and the wet on wet situation.

For a spherical tip, the force is:

$$f_m = -2\pi R\gamma_L(\cos\theta+\cos\phi)\left(1-\frac{dh}{dD}\right)f \quad \text{for dry on wet,}$$

$f_m = -2\pi R\gamma_L \frac{dr_o}{dD}$ for wet on wet,

where θ is the contact angle of the dry sphere and φ is the immersed angle, as shown in the figure Also R,h and D are illustrated in the same figure.

For a conical tip, the formula becomes:

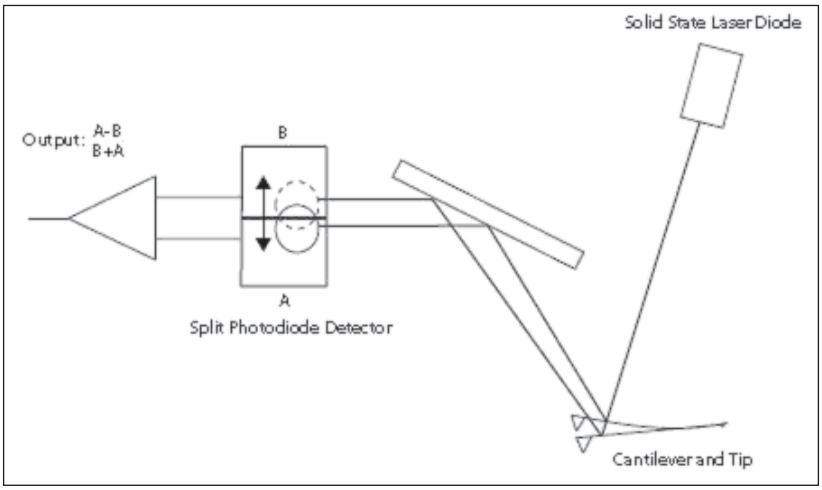
$f_m = -2\pi R\gamma_L \frac{\tan\delta}{\cos\delta}(\cos\theta + \sin\delta)(hD)(1 - \frac{dh}{dD})$ for dry on wet,

$f_m = -2\pi R\gamma_L (\frac{1}{\cos\delta} + \sin\delta)(r_o)(\frac{dr_o}{dD})$ for wet on wet,

where δ is the half cone angle and r_o and h are parameters of the meniscus profile.

AFM Cantilever-deflection Measurement

Beam-deflection Measurement



AFM beam-deflection detection.

The most common method for cantilever-deflection measurements is the beam-deflection method. In this method, laser light from a solid-state diode is reflected off the back of the cantilever and collected by a position-sensitive detector (PSD) consisting of two closely spaced photodiodes, whose output signal is collected by a differential amplifier. Angular displacement of the cantilever results in one photodiode collecting more light than the other photodiode, producing an output signal (the difference between the photodiode signals normalized by their sum), which is proportional to the deflection of the cantilever. The sensitivity of the beam-deflection method is very high, a noise floor on the order of 10 fm $Hz^{-1/2}$ can be obtained routinely in a well-designed system. Although this method is sometimes called the 'optical lever' method, the signal is not amplified if the beam path is made longer. A longer beam path increases the motion of the reflected spot on the photodiodes, but also widens the spot by the same amount due to diffraction, so that the same amount of optical power is moved from one photodiode to

the other. The 'optical leverage' (output signal of the detector divided by deflection of the cantilever) is inversely proportional to the numerical aperture of the beam focusing optics, as long as the focused laser spot is small enough to fall completely on the cantilever. It is also inversely proportional to the length of the cantilever.

The relative popularity of the beam-deflection method can be explained by its high sensitivity and simple operation, and by the fact that cantilevers do not require electrical contacts or other special treatments, and can therefore be fabricated relatively cheaply with sharp integrated tips.

Other Deflection-measurement Methods

Many other methods for beam-deflection measurements exist.

- Piezoelectric detection: Cantilevers made from quartz (such as the qPlus configuration), or other piezoelectric materials can directly detect deflection as an electrical signal. Cantilever oscillations down to 10pm have been detected with this method.
- Laser Doppler vibrometry: A laser Doppler vibrometer can be used to produce very accurate deflection measurements for an oscillating cantilever (thus is only used in non-contact mode). This method is expensive and is only used by relatively few groups.
- Scanning tunneling microscope (STM): The first atomic microscope used an STM complete with its own feedback mechanism to measure deflection. This method is very difficult to implement, and is slow to react to deflection changes compared to modern methods.
- Optical interferometry: Optical interferometry can be used to measure cantilever deflection. Due to the nanometre scale deflections measured in AFM, the interferometer is running in the sub-fringe regime, thus, any drift in laser power or wavelength has strong effects on the measurement. For these reasons optical interferometer measurements must be done with great care (for example using index matching fluids between optical fibre junctions), with very stable lasers. For these reasons optical interferometry is rarely used.
- Capacitive detection: Metal coated cantilevers can form a capacitor with another contact located behind the cantilever. Deflection changes the distance between the contacts and can be measured as a change in capacitance.
- Piezoresistive detection: Cantilevers can be fabricated with piezoresistive elements that act as a strain gauge. Using a Wheatstone bridge, strain in the AFM cantilever due to deflection can be measured. This is not commonly used in vacuum applications, as the piezoresistive detection dissipates energy from the system affecting Q of the resonance.

Piezoelectric Scanners

AFM scanners are made from piezoelectric material, which expands and contracts proportionally to an applied voltage. Whether they elongate or contract depends upon the polarity of the voltage applied. Traditionally the tip or sample is mounted on a 'tripod' of three piezo crystals, with each responsible for scanning in the *x*,*y* and *z* directions. In 1986, the same year as the AFM was invented, a new piezoelectric scanner, the tube scanner, was developed for use in STM. Later tube scanners were incorporated into AFMs. The tube scanner can move the sample in the *x*, *y*, and *z* directions using a single tube piezo with a single interior contact and four external contacts. An advantage of the tube scanner compared to the original tripod design, is better vibrational isolation, resulting from the higher resonant frequency of the single element construction, in combination with a low resonant frequency isolation stage. A disadvantage is that the *x*-*y* motion can cause unwanted *z* motion resulting in distortion. Another popular design for AFM scanners is the flexure stage, which uses separate piezos for each axis, and couples them through a flexure mechanism.

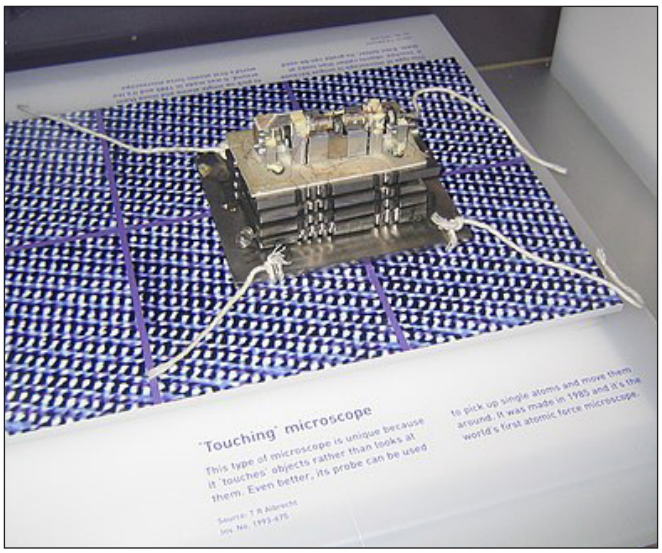
Scanners are characterized by their sensitivity, which is the ratio of piezo movement to piezo voltage, i.e., by how much the piezo material extends or contracts per applied volt. Because of differences in material or size, the sensitivity varies from scanner to scanner. Sensitivity varies non-linearly with respect to scan size. Piezo scanners exhibit more sensitivity at the end than at the beginning of a scan. This causes the forward and reverse scans to behave differently and display hysteresis between the two scan directions. This can be corrected by applying a non-linear voltage to the piezo electrodes to cause linear scanner movement and calibrating the scanner accordingly. One disadvantage of this approach is that it requires re-calibration because the precise non-linear voltage needed to correct non-linear movement will change as the piezo ages. This problem can be circumvented by adding a linear sensor to the sample stage or piezo stage to detect the true movement of the piezo. Deviations from ideal movement can be detected by the sensor and corrections applied to the piezo drive signal to correct for non-linear piezo movement. This design is known as a 'closed loop' AFM. Non-sensored piezo AFMs are referred to as 'open loop' AFMs.

The sensitivity of piezoelectric materials decreases exponentially with time. This causes most of the change in sensitivity to occur in the initial stages of the scanner's life. Piezoelectric scanners are run for approximately 48 hours before they are shipped from the factory so that they are past the point where they may have large changes in sensitivity. As the scanner ages, the sensitivity will change less with time and the scanner would seldom require recalibration, though various manufacturer manuals recommend monthly to semi-monthly calibration of open loop AFMs.

Advantages of Piezoelectric Scanners

AFM has several advantages over the scanning electron microscope (SEM). Unlike the

electron microscope, which provides a two-dimensional projection or a two-dimensional image of a sample, the AFM provides a three-dimensional surface profile. In addition, samples viewed by AFM do not require any special treatments (such as metal/carbon coatings) that would irreversibly change or damage the sample, and does not typically suffer from charging artifacts in the final image. While an electron microscope needs an expensive vacuum environment for proper operation, most AFM modes can work perfectly well in ambient air or even a liquid environment. This makes it possible to study biological macromolecules and even living organisms. In principle, AFM can provide higher resolution than SEM. It has been shown to give true atomic resolution in ultra-high vacuum (UHV) and, more recently, in liquid environments. High resolution AFM is comparable in resolution to scanning tunneling microscopy and transmission electron microscopy. AFM can also be combined with a variety of optical microscopy and spectroscopy techniques such as fluorescent microscopy of infrared spectroscopy, giving rise to scanning near-field optical microscopy, nano-FTIR and further expanding its applicability. Combined AFM-optical instruments have been applied primarily in the biological sciences but have recently attracted strong interest in photovoltaics and energy-storage research, polymer sciences, nanotechnology and even medical research.

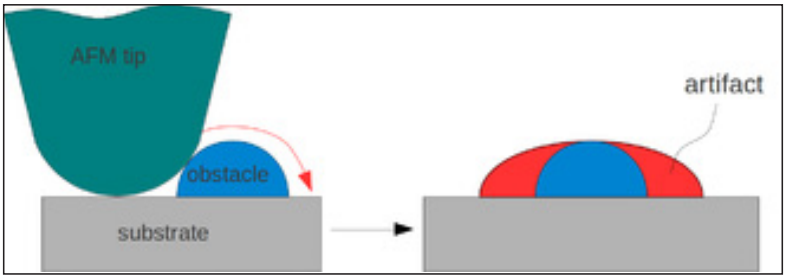


The first atomic force microscope.

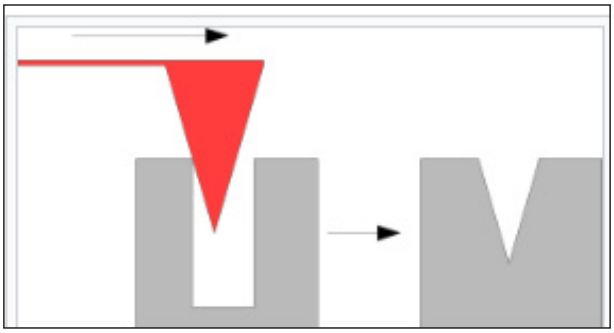
Disadvantages of Piezoelectric Scanners

A disadvantage of AFM compared with the scanning electron microscope (SEM) is the single scan image size. In one pass, the SEM can image an area on the order of square millimeters with a depth of field on the order of millimeters, whereas the AFM can only image a maximum scanning area of about 150×150 micrometers and a maximum height on the order of 10-20 micrometers. One method of improving the scanned area size for AFM is by using parallel probes in a fashion similar to that of millipede data storage.

The scanning speed of an AFM is also a limitation. Traditionally, an AFM cannot scan images as fast as an SEM, requiring several minutes for a typical scan, while an SEM is capable of scanning at near real-time, although at relatively low quality. The relatively slow rate of scanning during AFM imaging often leads to thermal drift in the image making the AFM less suited for measuring accurate distances between topographical features on the image. However, several fast-acting designs were suggested to increase microscope scanning productivity including what is being termed videoAFM (reasonable quality images are being obtained with videoAFM at video rate: faster than the average SEM). To eliminate image distortions induced by thermal drift, several methods have been introduced.



Showing an AFM artifact arising from a tip with a high radius of curvature with respect to the feature that is to be visualized.



AFM artifact, steep sample topography.

AFM images can also be affected by nonlinearity, hysteresis, and creep of the piezoelectric material and cross-talk between the x, y, z axes that may require software enhancement and filtering. Such filtering could "flatten" out real topographical features. However, newer AFMs utilize real-time correction software (for example, feature-oriented scanning) or closed-loop scanners, which practically eliminate these problems. Some AFMs also use separated orthogonal scanners (as opposed to a single tube), which also serve to eliminate part of the cross-talk problems.

As with any other imaging technique, there is the possibility of image artifacts, which could be induced by an unsuitable tip, a poor operating environment, or even by the sample itself, as depicted on the right. These image artifacts are unavoidable; however, their occurrence and effect on results can be reduced through various methods. Artifacts resulting from a too-coarse tip can be caused for example by inappropriate

handling or de facto collisions with the sample by either scanning too fast or having an unreasonably rough surface, causing actual wearing of the tip.

Due to the nature of AFM probes, they cannot normally measure steep walls or overhangs. Specially made cantilevers and AFMs can be used to modulate the probe sideways as well as up and down (as with dynamic contact and non-contact modes) to measure sidewalls, at the cost of more expensive cantilevers, lower lateral resolution and additional artifacts.

Small-angle Scattering

Small-angle scattering (SAS) is a scattering technique based on deflection of collimated radiation away from the straight trajectory after it interacts with structures that are much larger than the wavelength of the radiation. The deflection is small (0.1-10°) hence the name *small-angle*. SAS techniques can give information about the size, shape and orientation of structures in a sample.

SAS is a powerful technique for investigating large-scale structures from 10 Å up to thousands and even several tens of thousands of angstroms. The most important feature of the SAS method is its potential for analyzing the inner structure of disordered systems, and frequently the application of this method is a unique way to obtain direct structural information on systems with random arrangement of density inhomogeneities in such large-scales.

Currently, the SAS technique, with its well-developed experimental and theoretical procedures and wide range of studied objects, is a self-contained branch of the structural analysis of matter. SAS can refer to small angle neutron scattering (SANS) or small angle X-ray scattering (SAXS).

Applications

Small-angle scattering is particularly useful because of the dramatic increase in forward scattering that occurs at phase transitions, known as critical opalescence, and because many materials, substances and biological systems possess interesting and complex features in their structure, which match the useful length scale ranges that these techniques probe. The technique provides valuable information over a wide variety of scientific and technological applications including chemical aggregation, defects in materials, surfactants, colloids, ferromagnetic correlations in magnetism, alloy segregation, polymers, proteins, biological membranes, viruses, ribosome and macromolecules. While analysis of the data can give information on size, shape, etc., without making any model assumptions a preliminary analysis of the data can only give information on the radius of gyration for a particle using Guinier's equation.

Continuum Description

SAS patterns are typically represented as scattered intensity as a function of the magnitude of the scattering vector $q = 4\pi \sin(\theta)/\lambda$. Here 2θ is the angle between the incident beam and the detector measuring the scattered intensity, and λ is the wavelength of the radiation. One interpretation of the scattering vector is that it is the resolution or yardstick with which the sample is observed. In the case of a two-phase sample, e.g. small particles in liquid suspension, the only contrast leading to scattering in the typical range of resolution of the SAS is simply $\Delta\rho$, the difference in average scattering length density between the particle and the surrounding liquid, because variations in ρ due to the atomic structure only become visible at higher angles. This means that the total integrated intensity of the SAS pattern (in 3D) is an invariant quantity proportional to the square $\Delta\rho^2$. In 1-dimensional projection, as usually recorded for an isotropic pattern this invariant quantity becomes:

$$\int I(q)q^2\,dx,$$

where the integral runs from q=0 to wherever the SAS pattern is assumed to end and the diffraction pattern starts. It is also assumed that the density does not vary in the liquid or inside the particles, i.e. there is binary contrast. SAXS is described in terms of the electronic density where SANS is described in terms of a neutron scattering length density.

Porod's Law

At wave numbers that are relatively large on the scale of SAS, but still small when compared to wide-angle Bragg diffraction, local interface intercorrelations are probed, whereas correlations between opposite interface segments are averaged out. For smooth interfaces, one obtains Porod's law:

$$I(q) \sim Sq^{-4}.$$

This allows the surface area *S* of the particles to be determined with SAS. This needs to be modified if the interface is rough on the scale q^{-1}. If the roughness can be described by a fractal dimension *d* between 2-3 then Porod's law becomes:

$$I(q) \sim S'q^{-(6-d)}.$$

Scattering from Particles

Small-angle scattering from particles can be used to determine the particle shape or their size distribution. A small-angle scattering pattern can be fitted with intensities calculated from different model shapes when the size distribution is known. If the shape is known, a size distribution may be fitted to the intensity. Typically one assumes the particles to be spherical in the latter case.

If the particles are in solution and known to have uniform size dispersity, then a typical strategy is to measure different concentrations of particles in the solution. From the obtained SAXS patterns one can extrapolate to the intensity pattern one would get for a single particle. This is a necessary procedure that eliminates the *concentration effect*, which is a small shoulder that appears in the intensity patterns due to the proximity of neighbouring particles. The average distance between particles is then roughly the distance $2\pi/q^*$, where q^* is the position of the shoulder on the scattering vector range q. The shoulder thus comes from the structure of the solution and this contribution is called *the structure factor*. One can write for the small-angle X-ray scattering intensity:

$$I(q) = P(q)S(q),$$

where:

- $I(q)$ is the intensity as a function of the magnitude q of the scattering vector.
- $P(q)$ is the form factor.
- $S(q)$ is the structure factor.

When the intensities from low concentrations of particles are extrapolated to infinite dilution, the structure factor is equal to 1 and no longer disturbs the determination of the particle shape from the form factor $p(q)$. One can then easily apply the Guinier approximation (also called Guinier law, after André Guinier), which applies only at the very beginning of the scattering curve, at small q-values. According to the Guinier approximation the intensity at small q depends on the radius of gyration of the particle.

An important part of the particle shape determination is usually the distance distribution function $p(r)$, which may be calculated from the intensity using a Fourier transform:

$$p(r) = \frac{r^2}{2\pi^2}\int_0^\infty I(q)\frac{\sin qr}{qr}q^2 dq.$$

The distance distribution function $p(r)$, is related to the frequency of certain distances within the particle. Therefore it goes to zero at the largest diameter of the particle. It starts from zero at $r = 0$ due to the multiplication by r^2. The shape of the $p(r)$, -function already tells something about the shape of the particle. If the function is very symmetric, the particle is also highly symmetric, like a sphere. The distance distribution function should not be confused with the size distribution.

The particle shape analysis is especially popular in biological small-angle X-ray scattering, where one determines the shapes of proteins and other natural colloidal polymers.

Biophotonics

The term biophotonics denotes a combination of biology and photonics, with photonics being the science and technology of generation, manipulation, and detection of photons, quantum units of light. Photonics is related to electronics and photons. Photons play a central role in information technologies, such as fiber optics, the way electrons do in electronics.

Biophotonics can also be described as the "development and application of optical techniques, particularly imaging, to the study of biological molecules, cells and tissue". One of the main benefits of using the optical techniques which make up biophotonics is that they preserve the integrity of the biological cells being examined.

Biophotonics has therefore become the established general term for all techniques that deal with the interaction between biological items and photons. This refers to emission, detection, absorption, reflection, modification, and creation of radiation from biomolecular, cells, tissues, organisms, and biomaterials. Areas of application are life science, medicine, agriculture, and environmental science. Similar to the differentiation between "electric" and "electronics," a difference can be made between applications such as Therapy and surgery, which use light mainly to transfer energy, and applications such as diagnostics, which use light to excite matter and to transfer information back to the operator. In most cases, the term biophotonics refers to the latter type of application.

Applications

Biophotonics is an interdisciplinary field involving the interaction between electromagnetic radiation and biological materials including: tissues, cells, sub-cellular structures, and molecules in living organisms.

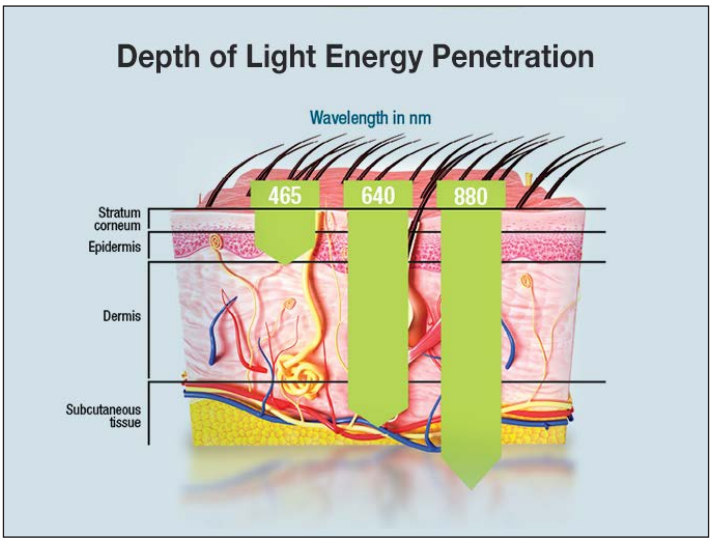
Recent biophotonics research has created new applications for clinical diagnostics and therapies involving fluids, cells, and tissues. These advances are allowing scientists and physicians opportunities for superior, non-invasive diagnostics for vascular and blood flow, as well as tools for better examination of skin lesions. In addition to new diagnostic tools, the advancements in biophotonics research have provided new photothermal, photodynamic, and tissue therapies.

- *Dermatology*: By observing the numerous and complex interactions between light and biological materials, the field of biophotonics presents a unique set of diagnostic techniques that medical practitioners can utilize. Biophotonic imaging provides the field of dermatology with the only non-invasive technique available for diagnosing skin cancers. Traditional diagnostic procedures for skin cancers involve visual assessment and biopsy, but a new Laser Induced Fluorescence spectroscopy technique allow dermatologists to compare spectrographs of

a patient's skin with spectrographs known to correspond with malignant tissue. This provides doctors with earlier diagnosis and treatment options.

"Among optical techniques, an emerging imaging technology based on laser scanning, the optical coherence tomography or OCT imaging is considered to be a useful tool to differentiate healthy from malignant skin tissue." The information is immediately accessible and eliminates the need for skin excision. This also eliminates the need for the skin samples to be processed in a lab which reduces labor costs and processing time.

Furthermore, these optical imaging technologies can be used during traditional surgical procedures to determine the boundaries of lesions to ensure that the entirety of the diseased tissue is removed. This is accomplished by exposing nanoparticles that have been dyed with a fluorescing substance to the acceptable light photons. Nanoparticles that are functionalized with fluorescent dyes and marker proteins will congregate in a chosen tissue type. When the particles are exposed to wavelengths of light that correspond to the fluorescent dye, the unhealthy tissue glows. This allows for the attending surgeon to quickly visually identify boundaries between healthy and unhealthy tissue, resulting in less time on the operating table and higher patient recovery. "Using dielectrophoretic microarray devices, nanoparticles and DNA biomarkers were rapidly isolated and concentrated onto specific microscopic locations where they were easily detected by epifluorescent microscopy."



Shows the depth of penetration of light through human skin.

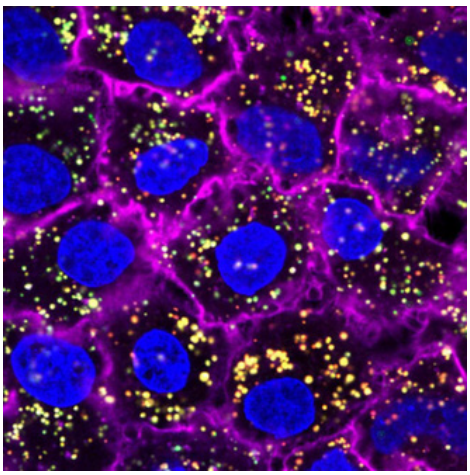
- Optical Tweezers: Optical tweezers (or traps) are scientific tools employed to maneuver microscopic particles such as atoms, DNA, bacteria, viruses, and other types of nanoparticles. It uses the lights momentum to exert small forces on a sample. This technique allows for the organizing and sorting of cells, the tracking of the movement of bacteria, and the changing of cell structure.

- Laser Micro-scalpels are a combination of fluorescence microscopy and a femtosecond laser "can penetrate up to 250 micrometers into tissue and target single cells in 3-D space." The technology, which was patented by researchers at the University of Texas at Austin, means that surgeons can excise diseased or damaged cells without disturbing or damaging healthy surrounding cells in delicate surgeries involving areas such as the eyes and vocal chords.

- Photoacoustic microscopy (PAM) is an imaging technology that utilizes both laser technology and ultrasound technology. This dual imaging modality is far superior at imaging deep tissue and vascular tissues than previous imaging technologies. The improvement in resolution provides higher quality images of deep tissues and vascular systems, allowing non-invasive differentiation of cancerous tissues vs healthy tissue by observing such things as "water content, oxygen saturation level, and hemoglobin concentration." Researchers have also been able to use PAM to diagnose endometriosis in rats.

- Low Level Laser Therapy (LLLT), although somewhat controversial as to it efficacy, can be used to treat wounds by repairing tissue and preventing tissue death. However, more recent studies indicate that LLLT is more useful for reducing inflammation and assuaging chronic joint pain. In addition, it is believed that LLLT could possibly prove to be useful in the treatment of severe brain injury or trauma, stroke, and degenerative neurological diseases.

- Photodynamic Therapy (PT) uses photosynthesizing chemicals and oxygen to induce a cellular reaction to light. It can be used to kill cancer cells, treat acne, and reduce scarring. PT can also kill bacteria, viruses, and fungi. The technology provides treatment with little to no long-term side effects, is less invasive than surgery and can be repeated more often than radiation. Treatment is limited, however, to surfaces and organs that can be exposed to light, which eliminates deep tissue cancer treatments.



Nano particles injected into a tumor to use photothermal therapy.

- Photothermal therapy most commonly uses nanoparticles made of a noble metal to convert light into heat. The nanoparticles are engineered to absorb light in the 700-1000 nm range, where the human body is optically transparent. When the particles are hit by light they heat up, disrupting or destroying the surrounding cells via hyperthermia. Because the light used does not interact with tissue directly, photothermal therapy has few long term side effects and it can be used to treat cancers deep within the body.

FRET

Fluorescence Resonance Energy Transfer, also known as Foerster Resonance Energy Transfer (FRET in both cases) is the term given to the process where two excited "fluorophores" pass energy one to the other non-radiatively (i.e., without exchanging a photon). By carefully selecting the excitation of these flurophores and detecting the emission, FRET has become one of the most widely used techniques in the field of Biophotonics, giving scientists the chance to investigate sub-cellular environments.

Biofluorescence

Biofluorescence describes the absorption of ultraviolet or visible light and the sub sequential emission of photons at a lower energy level (S_1 excited state relaxes to S_0 ground state) by intrinsically fluorescent proteins or by synthetic fluorescent molecules covalently attached to a biomarker of interest. Biomarkers are molecules indicative or disease or distress and are a typically monitored systemically in a living organism, or by using an ex vivo tissue sample for microscopy, or in vitro: in the blood, urine, sweat, saliva, interstitial fluid, aqueous humor, or sputum. Stimulating light excites an electron, raising energy to an unstable level. This instability is unfavorable, so the energized electron is returned to a stable state almost as immediately as it becomes unstable.The time delay between excitation and re-emission that occurs when returning to the stable ground state causes the photon that is re-emitted to be a different color (i.e. it relaxes to a lower energy and thus the photon emitted is at a shorter wavelength, as governed by the Plank-Einstein relation $\mathrm{E} = \frac{\mathrm{hc}}{\lambda}$) than the excitation light that was absorbed. This return to stability corresponds with the release of excess energy in the form of fluorescent light. This emission of light is only observable whilst the exciation light is still providing photons to the fluorescent molecule and is typically excited by blue or green light and emits purple, yellow, orange, green, cyan, or red. Biofluorescence is often confused with the following forms of biotic light: Bioluminescence and biophosphorescence.

Bioluminescence

Bioluminescence differs from biofluorescence in that it is the natural production of light by chemical reactions within an organism, whereas biofluorescence and biophosphorescence are the absorption and reemission of light from the natural environment.

Biophosphorescence

Biophosphorescence is similar to biofluorescence in its requirement of light at specified wavelengths as a provider of excitation energy. The difference here lies in the relative stability of the energized electron. Unlike with biofluorescence, here the electron retains stability in the forbidden triplet state (unpaired spins), with a longer delay in emitting light resulting in the effect that it continues to "glow-in-the-dark" even long after the stimulating light source has been removed.

Biolasing

A biolaser is when laser light is generated by or from within a living cell. Imaging in biophotonics often relies on laser light, and integration with biological systems is seen as a promising route to enhancing sensing and imaging techniques. Biolasers, like any laser system, requires three components gain medium, optical feedback structure and pump source. For the gain medium, a variety of naturally produced fluorescent proteins can be used in different laser structure. Enclosing an optical feedback structure in a cell has been demonstrated using cell vacuoles, as well as using fully enclosed laser systems such as dye doped polymer microspheres, or semiconductor nanodisks lasers.

Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a physical technique used to determine the thermodynamic parameters of interactions in solution. It is most often used to study the binding of small molecules (such as medicinal compounds) to larger macromolecules (proteins, DNA etc.). It consists of two cells which are enclosed in an adiabatic jacket. The compounds to be studied are placed in the sample cell, while the other cell, the reference cell, is used as a control and contains the buffer in which the sample is dissolved.

Thermodynamic Measurements

ITC is a quantitative technique that can determine the binding affinity (K_a), enthalpy changes (ΔH), and binding stoichiometry (n) of the interaction between two or more molecules in solution. From these initial measurements, Gibbs free energy changes (ΔG) and entropy changes (ΔS) can be determined using the relationship:

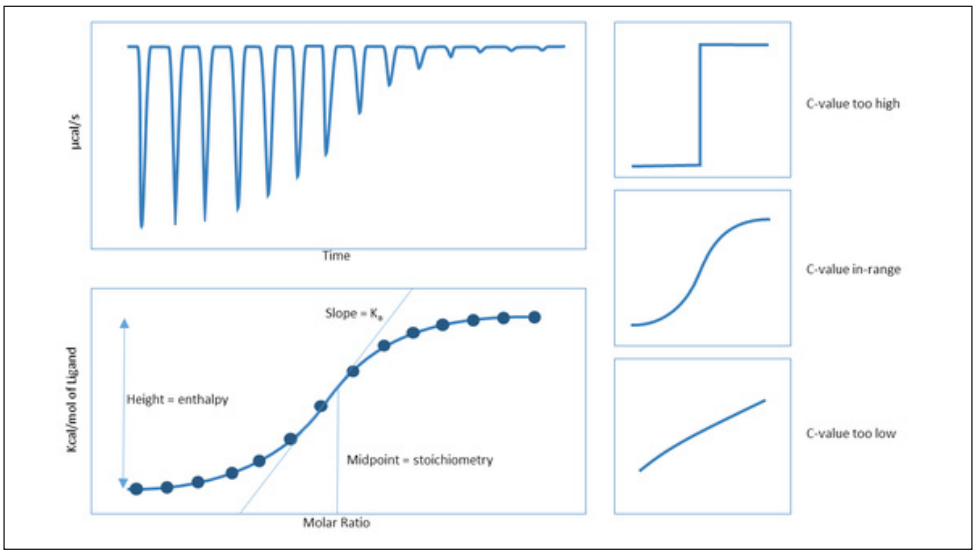
$$\Delta G = -RT \ln K_a = \Delta H - T\Delta S,$$

where R is the gas constant and T is the absolute temperature.

For accurate measurements of binding affinity, the curve of the thermogram must be sigmoidal. The profile of the curve is determined by the c-value, which is calculated using the equation:

$$c = n * Ka * M$$

where n is the stoichiometry of the binding, K_a is the association constant and M is the concentration of the molecule in the cell.

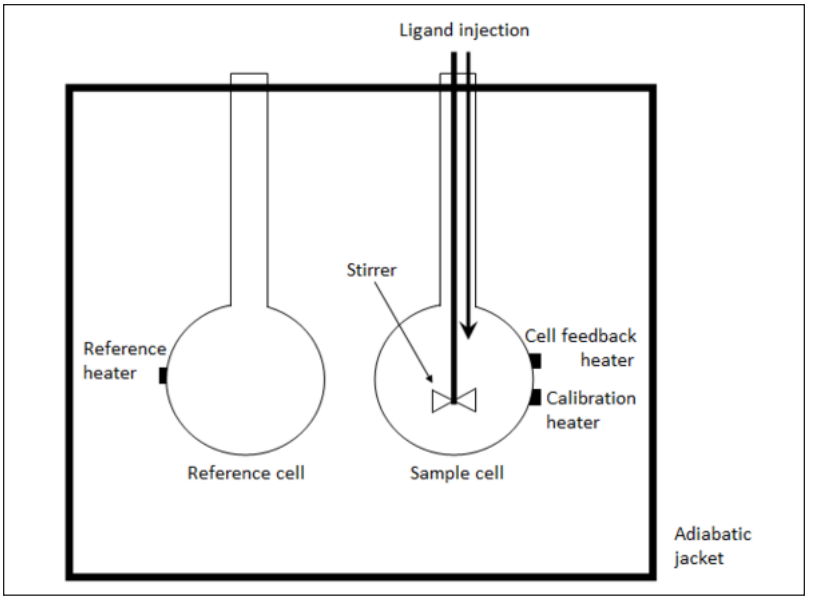


ITC thermogram.

Instrument

An isothermal titration calorimeter is composed of two identical cells made of a highly efficient thermally conducting and chemically inert material such as Hastelloy alloy or gold, surrounded by an adiabatic jacket. Sensitive thermopile/thermocouple circuits are used to detect temperature differences between the reference cell (filled with buffer or water) and the sample cell containing the macromolecule. Prior to addition of ligand, a constant power (<1 mW) is applied to the reference cell. This directs a feedback circuit, activating a heater located on the sample cell. During the experiment, ligand is titrated into the sample cell in precisely known aliquots, causing heat to be either taken up or evolved (depending on the nature of the reaction). Measurements consist of the time-dependent input of power required to maintain equal temperatures between the sample and reference cells.

In an exothermic reaction, the temperature in the sample cell increases upon addition of ligand. This causes the feedback power to the sample cell to be decreased (remember: A reference power is applied to the reference cell) in order to maintain an equal temperature between the two cells. In an endothermic reaction, the opposite occurs; the feedback circuit increases the power in order to maintain a constant temperature (isothermic/isothermal operation).



Schematic of an ITC instrument.

Observations are plotted as the power needed to maintain the reference and the sample cell at an identical temperature against time. As a result, the experimental raw data consists of a series of spikes of heat flow (power), with every spike corresponding to one ligand injection. These heat flow spikes/pulses are integrated with respect to time, giving the total heat exchanged per injection. The pattern of these heat effects as a function of the molar ratio [ligand]/[macromolecule] can then be analysed to give the thermodynamic parameters of the interaction under study. Degassing samples is often necessary in order to obtain good measurements as the presence of gas bubbles within the sample cell will lead to abnormal data plots in the recorded results. The entire experiment takes place under computer control.

Application in Drug Discovery

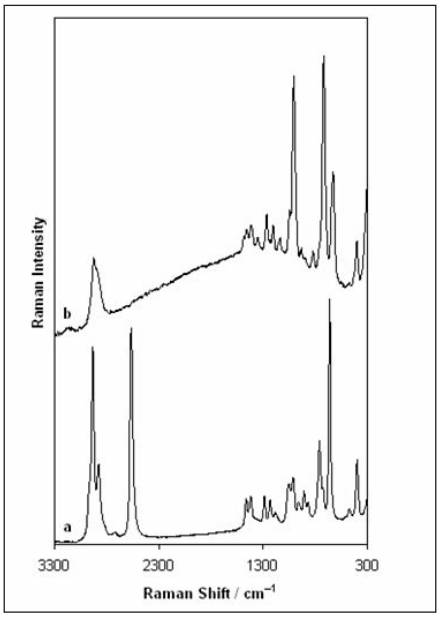
ITC is one of the latest techniques to be used in characterizing binding affinity of ligands for proteins. It is typically used as a secondary screening technique in high throughput screening. ITC is particularly useful as it gives not only the binding affinity, but also the thermodynamics of the binding. This thermodynamic characterization allows for further optimization of compounds.

Surface Enhanced Raman Spectroscopy

Surface-enhanced Raman spectroscopy or surface-enhanced Raman scattering (SERS) is a surface-sensitive technique that enhances Raman scattering by molecules adsorbed on rough metal surfaces or by nanostructures such as plasmonic-magnetic silica

nanotubes. The enhancement factor can be as much as 10^{10} to 10^{11}, which means the technique may detect single molecules.

Raman spectrum of liquid 2-mercaptoethanol (below) and SERS spectrum of 2-mercaptoethanol monolayer formed on roughened silver (above). Spectra are scaled and shifted for clarity. A difference in selection rules is visible: Some bands appear only in the bulk-phase Raman spectrum or only in the SERS spectrum.



Mechanisms

The exact mechanism of the enhancement effect of SERS is still a matter of debate in the literature. There are two primary theories and while their mechanisms differ substantially, distinguishing them experimentally has not been straightforward. The electromagnetic theory proposes the excitation of localized surface plasmons, while the chemical theory proposes the formation of charge-transfer complexes. The chemical theory is based on resonance Raman spectroscopy, in which the frequency coincidence (or resonance) of the incident photon energy and electron transition greatly enhances Raman scattering intensity. Research in 2015 on a more powerful extension of the SERS technique called SLIPSERS (Slippery Liquid-Infused Porous SERS) has further supported the EM theory.

Electromagnetic Theory

The increase in intensity of the Raman signal for adsorbates on particular surfaces occurs because of an enhancement in the electric field provided by the surface. When the incident light in the experiment strikes the surface, localized surface plasmons are excited. The field enhancement is greatest when the plasmon frequency, ω_p is in resonance with the radiation ($\omega = \omega_p / \sqrt{3}$ for spherical particles). In order for scattering

to occur, the plasmon oscillations must be perpendicular to the surface; if they are in-plane with the surface, no scattering will occur. It is because of this requirement that roughened surfaces or arrangements of nanoparticles are typically employed in SERS experiments as these surfaces provide an area on which these localized collective oscillations can occur. SERS enhancement can occur even when an excited molecule is relatively far apart from the surface which hosts metallic nanoparticles enabling surface plasmon phenomena.

The light incident on the surface can excite a variety of phenomena in the surface, yet the complexity of this situation can be minimized by surfaces with features much smaller than the wavelength of the light, as only the dipolar contribution will be recognized by the system. The dipolar term contributes to the plasmon oscillations, which leads to the enhancement. The SERS effect is so pronounced because the field enhancement occurs twice. First, the field enhancement magnifies the intensity of incident light, which will excite the Raman modes of the molecule being studied, therefore increasing the signal of the Raman scattering. The Raman signal is then further magnified by the surface due to the same mechanism that excited the incident light, resulting in a greater increase in the total output. At each stage the electric field is enhanced as E^2, for a total enhancement of E^4.

The enhancement is not equal for all frequencies. For those frequencies for which the Raman signal is only slightly shifted from the incident light, both the incident laser light and the Raman signal can be near resonance with the plasmon frequency, leading to the E^4 enhancement. When the frequency shift is large, the incident light and the Raman signal cannot both be on resonance with ω_p, thus the enhancement at both stages cannot be maximal.

The choice of surface metal is also dictated by the plasmon resonance frequency. Visible and near-infrared radiation (NIR) are used to excite Raman modes. Silver and gold are typical metals for SERS experiments because their plasmon resonance frequencies fall within these wavelength ranges, providing maximal enhancement for visible and NIR light. Copper's absorption spectrum also falls within the range acceptable for SERS experiments. Platinum and palladium nanostructures also display plasmon resonance within visible and NIR frequencies.

Chemical Theory

Resonance Raman spectroscopy explains the huge enhancement of Raman scattering intensity. Intermolecular and intramolecular charge transfers significantly enhance Raman spectrum peaks. In particular, the enhancement is huge for species adsorbing the metal surface due to the high-intensity charge transfers from the metal surface with wide band to the adsorbing species. This resonance Raman enhancement is dominant in SERS for species on small nanoclusters with considerable band gaps, because surface plasmon appears only in metal surface with near-zero band gaps. This chemical mechanism probably occurs in concert with the electromagnetic mechanism for metal surface.

Surfaces

While SERS can be performed in colloidal solutions, today the most common method for performing SERS measurements is by depositing a liquid sample onto a silicon or glass surface with a nanostructured noble metal surface. While the first experiments were performed on electrochemically roughened silver, now surfaces are often prepared using a distribution of metal nanoparticles on the surface as well as using lithography or porous silicon as a support. Two dimensional silicon nanopillars decorated with silver have also been used to create SERS active substrates. The most common metals used for plasmonic surfaces are silver and gold; however, aluminium has recently been explored as an alternative plasmonic material, because its plasmon band is in the UV region, contrary to silver and gold. Hence, there is great interest in using aluminium for UV SERS. It has, however, surprisingly also been shown to have a large enhancement in the infrared, which is not fully understood. In the current decade, it has been recognized that the cost of SERS substrates must be reduced in order to become a commonly used analytical chemistry measurement technique. To meet this need, plasmonic paper has experienced widespread attention in the field, with highly sensitive SERS substrates being formed through approaches such as soaking, in-situ synthesis, screen printing and inkjet printing.

The shape and size of the metal nanoparticles strongly affect the strength of the enhancement because these factors influence the ratio of absorption and scattering events. There is an ideal size for these particles, and an ideal surface thickness for each experiment. Particles that are too large allow the excitation of multipoles, which are nonradiative. As only the dipole transition leads to Raman scattering, the higher-order transitions will cause a decrease in the overall efficiency of the enhancement. Particles that are too small lose their electrical conductance and cannot enhance the field. When the particle size approaches a few atoms, the definition of a plasmon does not hold, as there must be a large collection of electrons to oscillate together. An ideal SERS substrate must possess high uniformity and high field enhancement. Such substrates can be fabricated on a wafer scale and label-free superresolution microscopy has also been demonstrated using the fluctuations of surface enhanced Raman scattering signal on such highly uniform, high-performance plasmonic metasurfaces.

Applications

SERS substrates are used to detect the presence of low-abundance biomolecules, and can therefore detect proteins in bodily fluids. Early detection of pancreatic cancer biomarkers was accomplished using SERS-based immunoassay approach. A SERS-base multiplex protein biomarker detection platform in a microfluidic chip to detect is used to detect several protein biomarkers to predict the type of disease and critical biomarkers and increase the chance of diagnosis between diseases with similar biomarkers (PC, OVC, and pancreatitis). This technology has been utilized to detect urea and blood plasma label free in human serum and may become the next generation in cancer detection and screening.

The ability to analyze the composition of a mixture on the nano scale makes the use of SERS substrates beneficial for environmental analysis, pharmaceuticals, material sciences, art and archeological research, forensic science, drug and explosives detection, food quality analysis, and single algal cell detection. SERS combined with plasmonic sensing can be used for high-sensitivity and quantitative detection of biomolecular interaction, and to study redox processes at the single molecule level.

Immunoassays

SERS-based immunoassays can be used for detection of low-abundance biomarkers. For example, antibodies and gold particles can be used to quantify proteins in serum with high sensitivity and specificity.

Oligonucleotide Targeting

SERS can be used to target specific DNA and RNA sequences using a combination of gold and silver nanoparticles and Raman-active dyes, such as Cy3. Specific single nucleotide polymorphisms (SNP) can be identified using this technique. The gold nanoparticles facilitate the formation of a silver coating on the dye-labeled regions of DNA or RNA, allowing SERS to be performed. This has several potential applications: For example, Cao et al. report that gene sequences for HIV, Ebola, Hepatitis, and Bacillus Anthracis can be uniquely identified using this technique. Each spectrum was specific, which is advantageous over fluorescence detection; some fluorescent markers overlap and interfere with other gene markers. The advantage of this technique to identify gene sequences is that several Raman dyes are commercially available, which could lead to the development of non-overlapping probes for gene detection.

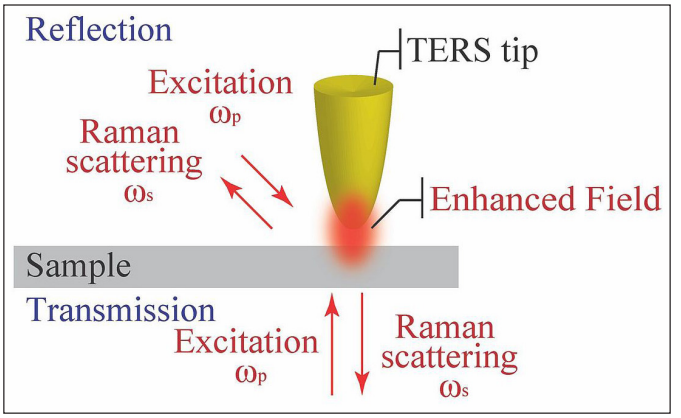
Tip Enhanced Raman Spectroscopy

TERS (Tip Enhanced Raman Spectroscopy) brings Raman spectroscopy into nanoscale resolution imaging. TERS is a super-resolution chemical technique. Better yet, it is a label-free super-resolution imaging technique which has been extended by our novel technology into an important new imaging technology.

TERS imaging is performed with an AFM/Raman system, where a Scanning Probe microscope (SPM that can be used in atomic force, scanning tunneling, or normal/shear force mode) is integrated with a confocal Raman spectrometer through an opto-mechanical coupling. The scanning probe microscope allows for nanoscale imaging, the optical coupling brings the excitation laser to the functionalized tip (or probe), and the spectrometer analyzes the Raman (or otherwise scattered) light providing a hyperspectral image with nanometer scale chemical contrast.

A TERS system is based on a metallic tip (generally made of gold or silver) employed to concentrate the incident light field at the apex. The tip acts as a nano-source of light

and local field enhancer, greatly improving the Raman sensitivity (by a factor of 10^3 _10^7) and reducing the probed volume to the "nano" region immediately below the tip. The optical coupling that combines the two instruments uses a confocal scheme. Two different configurations exist for this coupling: one in transmission and one in reflection, having their own advantages and drawbacks.



Reflection and transmission TERS configurations.

The transmission configuration allows the use of the highest numerical aperture (NA) objectives (including immersion objectives) giving high power density at the focus point and enabling the collection of high signal level, but it can only be used for transparent samples. The reflection configuration can be used for any kind of samples (opaque and transparent) but is limited to lower NA objectives. By combining point-by-point scanning with simultaneous spectrum acquisition, near-field Raman mappings can be performed with lateral resolution down to ten nanometers or less.

References

- Rudenberg, H. Gunther; Rudenberg, Paul G. (2010). "Origin and Background of the Invention of the Electron Microscope". Advances in Imaging and Electron Physics. 160. pp. 207–286. doi:10.1016/S1076-5670(10)60006-7. ISBN 978-0-12-381017-5
- What-is-tip-enhanced-raman-spectroscopy, nanoraman, microscopy-and-imaging, measurement-and-control-technique, technology: horiba.com, Retrieved 22 April, 2020
- Ravelli RB, Garman EF (October 2006). "Radiation damage in macromolecular cryocrystallography". Current Opinion in Structural Biology. 16 (5): 624–9. doi:10.1016/j.sbi.2006.08.001. PMID 16938450
- O'Brien, R., Ladbury, J.E. and Chowdry B.Z. (2000) Isothermal titration calorimetry of biomolecules. Chapter 10 in Protein-Ligand interactions: hydrodynamics and calorimetry Ed. Harding, S.E. and Chowdry, B.Z, Oxford University Press. ISBN 0-19-963746-6
- Svergun DI; Koch MHJ (2003). "Small-angle scattering studies of biological macromolecules in solution". Rep. Prog. Phys. 66 (10): 1735–1782. Bibcode:2003RPPh...66.1735S. doi:10.1088/0034-4885/66/10/R05

5

Understanding Biomechanics

Biomechanics is the branch of biophysics that studies the structure, function and motion of the mechanical aspects of biological systems of living organisms and their organs and cells. Ground reaction force, Hill's muscle model, lines of non-extension, biomechatronics, force platform, etc. are some of the aspects that are studied under its domain. All the aspects related to biomechanics have been carefully analyzed in this chapter.

Biomechanics is the study of how the systems and structures of biological organisms, from the smallest plants to the largest animals, react to various forces and external stimuli. In humans, biomechanics often refers to the study of how the skeletal and musculature systems work under different conditions. In biomechanics more generally, scientists often try to apply physics and other mathematically based forms of analysis to discover the limits and capabilities of biological systems.

In a way, biomechanics has been around since the inquiring ancient Greek and Roman minds began dissecting animals and vivisecting humans to discover the inner systems of our bodies. Many of the great philosophers and scientists of our past tried their hand at some form of biomechanics, from Aristotle, who wrote *On the Motion of Animals* in the 4th century BC, to Leonardo da Vinci, who studied human muscle and joint function in 15th century Italy. In the 19th century, scores of Europeans were incredibly fascinated, for some reason, with the gait of horses and extensively studied the biomechanics of a horse's galloping motion. Today, rather than a field that scientists and philosophers dabble in, biomechanics is its own branch of human and biological science, with entire departments in hospitals and universities devoted to the subject's study.

Biomechanics combines the field of engineering mechanics with the fields of biology and physiology. Biomechanics applies mechanical principles to the human body in order to understand the mechanical influences on bone and joint health. Forces that load the joints are generated by muscles and transmitted by tendons. Bones must withstand these forces. Developments in the field of biomechanics have improved our

understanding of normal and pathologic gait, mechanics of neuromuscular control, and mechanics of growth and form. This knowledge has contributed to the development of medical diagnostic and treatment procedures. It has provided the basis for the design and manufacture of medical implants and orthotic devicesand has enhanced rehabilitation therapy practices. Biomechanics has also been used to improve human performance in the workplace and in athletic competition.

In the human body these are the effects of the application of forces to the human body. In the case of the upper limb this could include internal or external forces. Knowing and understanding the range of forces that are unique in the upper limb will help the health care team to appropriately create the intended design and treatment plan. In the case of the upper limb, muscles are typically attached close to joints, favoring efficiency, speed, and fine motor movements. For the most part, internal muscles of the upper limb such as the lumbricals, triceps, biceps, and flexor digitorum are optimized for fine quick movements over a larger ROM. Correctly designed orthoses should consider the internal forces of the limb segments being treated and the application of external forces that will be applied to the limbs as required by the patient's environment, chosen profession, or activities of daily living.

Biomechanics is considered one of the core disciplines to understand principles of physiology and pathophysiology in the human. Decades of biomechanics research revealed many primary principles regulating growth, tissue formation, organism morphogenesis, and tissue regeneration. Not surprisingly, biomechanics is one of the core sciences used to understand how dynamic loading in the central nervous system (CNS) causes injury. This key transition from physiology to pathophysiology is especially important, as early pathophysiologic changes to the CNS can strongly influence cellular decisions to survive, reintegrate, and repair the injured CNS. This long-term response, initiated when forces are transferred to the cellular/molecular scale at the time of injury, is a key contributor in the outcome of head-injured and spinal cord-injured patients.

Compared to nearly every other application of biomechanics and human body, biomechanics of traumatic loading has three unique components. First, the mechanical event is nearly always considered as a single event, rather than a series of cyclical loadings that are applied to organ/tissue/cellular preparations. In other areas of biomechanics, cyclical loading of tissue/cellular components is critical in defining the homeostatic response and the adaptation of this response during disease. Examples include the periodic distension of the vascular wall during the cardiovascular cycle, or the frequent loading and unloading of orthopedic soft tissues during gait. In contrast, the brain and spinal cord are considered "mechanically protected" organs and do not have a clear constant level of mechanical stimulation. Second, traumatic loading is probably the fastest event studied in biomechanics, especially considering the very recent work on blast injury biomechanics. The very brief nature of the event – the acceleration/deceleration event in a blunt impact event is typically delivered in less than 50 milliseconds, while the mechanical loading in a blast event lasts only a few milliseconds – present

enormous technical challenges because traditional biomechanical tools to study the forces, displacements, deformations, and stresses in tissue/cellular structures are often lacking. Third, the functional and physical wiring of the brain and spinal cord remain largely a mystery and are the focus of many efforts across the neuroscienceand neurologic sciences. Any biomechanically based study of events at the cellular and molecular scale must consider the physical wiring of the structure and our ability to mechanically interrogate individual components of these circuits to lay the groundwork for a more integrative view of how the brain and spinal cord will respond to the mechanical forces occurring during injury.

Given this unique set of circumstances, it is not surprising that there has been a large increase in the use and development of simplified laboratory models to deconstruct some of these experimental complexities. Foremost among these challenges is developing new simplified laboratory models that develop repeatable, reproducible, and *informative models* of the clinical condition. Certainly, as we know more about how mechanical loading affects the function of the brain and spinal cord at many length scales – at the organ level, at the tissue level, and at the cellular/molecular level – we will become much more facile with developing models that accurately reproduce these loadings *in vitro*.

The conversion of the mechanical input to the resulting biochemical cascades – termed mechanotransmission – triggers the evolving injury patterns in the brain and spinal cord after traumatic injury. The concept of mechanotransmission has transformed biomechanics from a study of structure-property relationships into a more integrated structure-function-property triad across lengths scales in the CNS. At the most reductionist level, this means that we are now beginning to explore how some cells respond directly to the mechanical load with a set of biochemical signatures, even now extending these into genomic signatures. Not surprisingly, as the mechanical load is adjusted to either individual cells or clusters of cells, one may find that some cell types respond while others do not.

Sports Biomechanics

Sports biomechanics is a quantitative based study and analysis of professional athletes and sports activities in general. It can simply be described as the physics of sports. In this subfield of biomechanics the laws of mechanics are applied in order to gain a greater understanding of athletic performance through mathematical modeling, computer simulation and measurement. Biomechanics is the study of the structure and function of biological systems by means of the methods of mechanics (the branch of physics involving analysis of the actions of forces). Within mechanics there are two sub-fields of study: statics, which is the study of systems that are in a state of constant motion either at rest (with no motion) or moving with a constant velocity; and dynamics, which is the

study of systems in motion in which acceleration is present, which may involve kinematics (the study of the motion of bodies with respect to time, displacement, velocity, and speed of movement either in a straight line or in a rotary direction) and kinetics (the study of the forces associated with motion, including forces causing motion and forces resulting from motion). Sports biomechanists help people obtain optimal muscle recruitment and performance. A biomechanist also uses their knowledge to apply proper load barring techniques to preserve the body.

Human Movement Biomechanics

All life forms on earth are subjected to gravity. How the form, function and motion of these biological systems are affected directly or indirectly by gravity has been a subject of scientific study over centuries. It started with the development of mechanics that were concerned with the behavior of physical bodies when subjected to forces or displacements, and the subsequent effect of the bodies on their environment. More than 3000 years ago Babylonian astronomers studied the problem of planetary positions. The Renaissance astronomer Nicolaus Copernicus published his heliocentric theory in 1543, which was in contrast to the widely accepted geocentric system model popular during the 13th–17th centuries that had been proposed by Aristotle and Claudius Ptolemy. During the early modern period, renowned scientists, including Galileo, Kepler and Newton, laid the foundation of classical mechanics. Galileo Galilei worked theoretically and experimentally on the motions of bodies, particularly of falling bodies. Johannes Kepler empirically discovered his laws of planetary motion which gave an approximate description of the motion of planets around the sun and provided one of the foundations for Isaac Newton's theory of universal gravitation.

In 1687 Isaac Newton used the newly developed mathematics of calculus to give a detailed mathematical explanation of mechanics and published his classic *Philosophiae Naturalis Principia Mathematica*. In this three-volume publication he formulated the law of universal gravitation and the three laws of motion which can be applied to the motion of planets in the heavens and all forms of movements on earth. Since all life forms on earth are under the influence of universal gravitation, there is no doubt that mechanics not only governs the motions of lifeless objects, but also affects the form, motion and function of the biological systems on earth. The mechanical interactions within the biological systems and with the environment received attention from early scholars such as Aristotle, Leonardo da Vinci, Galileo Galilei, Johannes Kepler, Rene Descartes, and Isaac Newton, among others, as well as from scientists of the present day. Their efforts on the discovery of these mechanical interactions have come to form a discipline of research called *biomechanics*.

Biomechanics is the study of continuum mechanics (that is, the study of loads, motion, stress, and strain of solids and fluids) of biological systems and the mechanical

effects on the body's movement, size, shape and structure. Mechanical influence on biological systems can be found at multiple levels, from molecular and cellular, all the way up to the tissue, organ and system level. Therefore, the study of biomechanics of humans ranges from the inner workings of a cell, the mechanical properties of soft and hard tissues, to the development and movement of the neuromusculoskeletal system of the body. *Molecular biomechanics* refers to the study of how mechanical forces and deformation affect the conformation, binding/reaction, function and transport of biomolecules, such as DNA, RNA and proteins, and how mechanobiochemistry couples in biomolecular motors and ion channel flows, etc. *Cellular biomechanics* is concerned with the study of how cells sense mechanical forces or deformations, and transduce them into biological responses, especially for the study of how mechanical forces alter cell growth, differentiation, movement, signal transduction, protein secretion and transport, gene expression and regulation.

The properties of living tissues are affected by applied loads and deformations, and *tissue biomechanics* is mainly concerned with the growth and remodeling of tissues as a response to applied mechanical stimuli. For example, the effects of elevated blood pressure on the mechanics of the arterial wall, and the behavior of cardiomyocytes within a heart with a cardiac infarct, have been widely regarded as instances in which living tissue is remodeled as a direct consequence of applied loads. Another example is Wolff's law of bone remodeling, developed by Julius Wolff in the 19th century. Wolff's laws states that the internal architecture of the trabecular bone and the external cortical bone in a healthy person or animal will adapt to the loads placed on the bone and it will remodel itself over time to become stronger to resist that type of loading. The converse is also true. If the loading on a bone decreases, the bone will become weaker owing to turnover and a lack of stimulus for continued remodeling that is required to maintain bone mass.

At the system level, mechanical factors also affect the form, performance and function of the musculoskeletal system. Human movement is achieved by a complex and highly coordinated mechanical interaction between bones, muscles, ligaments and joints within the musculoskeletal system under the control of the nervous system. Muscles generate tensile forces and apply moments at joints with short lever arms in order to provide static and dynamic stability of the body under gravitational and other loads while regularly performing precise limb control. Any injury or lesion of any of the individual elements of the musculoskeletal system will change the mechanical interaction and cause degradation, instability or disability of movement. On the other hand, proper modification, manipulation and control of the mechanical environment can help prevent injury, correct abnormality, and speed healing and rehabilitation.

One of the most succes Therefore, understanding the biomechanics and loading of each element is helpful for studying disease etiology, making treatment decisions and evaluating the effects of treatment. However, because of ethical considerations and technological limitations, direct measurement of the forces transmitted in the human body is

possible only in exceptional circumstances, such as through instrumented implants. A further challenge is the redundant nature of the musculoskeletal system. In the human body there are more joints and muscles than are necessary for performing our daily motor tasks. Therefore, a certain task can be achieved by more than one musculoskeletal strategy. However, this compensatory mechanism is essential for coping with the consequences of injuries or diseases to the musculoskeletal system, but it makes it difficult to determine the internal forces noninvasively.

Currently, combining noninvasive measurements of the movement, such as the position of segments and the strain on force-measuring instruments, with computer graphics-based anatomical modeling is a useful approach to estimating these loadings. In this approach it is essential to integrate the techniques of motion analysis and medical imaging. They include measuring human motion and external loads, developing three-dimensional (3D) computer graphics-based biomechanical models based on medical imaging, calculating internal forces and validating the results. A validated 3D computer biomechanical model can then be applied to the simulation of various movements and surgical procedures. A recording of an electromyogram (EMG) of the active muscles can further be used to understand the muscle activity during human movement. However, developing a precise and noninvasive method for measuring the internal force within the human body for clinical and other purposes still remains a great challenge in the field of human biomechanics and motion analysis.

Human Motion Analysis

All movements and changes in movements arise from the action of forces, both internal and external. A change in the force acting on an object is necessary for moving an object from a stationary position or for changing its velocity. The amount of change in the velocity of an object depends on the magnitude and direction of the applied force. Newton's laws of motion give a clear relationship between the changing force and the resultant change in movement, and this is applicable to all forms of movement, including human locomotion. Human motion analysis is the systematic study of human motion by careful observation, augmented by instrumentation for measuring body movements, body mechanics and the activity of the muscles. It aims to gather quantitative information about the mechanics of the musculoskeletal system during the execution of a motor task. A special branch of human motion analysis is gait analysis which is specific to the study of human walking, and is used to assess, plan and treat individuals with conditions affecting their ability to walk. The following is a brief account of the history of human motion analysis/gait analysis.

In pursuing mental and physical excellence, the ancient Greeks found that harmony of mind and body required athletic activity to complement the pursuit of knowledge. Their interest in sport and human movement can be seen in the predominance of kinematic representations of Greek athletics in the artistic media. With the mechanical, mathematical and anatomical paradigms developed during Greek antiquity, the great

philosopher Aristotle wrote the first book on human movement, which is the first scientific analysis of human and animal movement in terms of observing and describing muscular action and movement.

The great figure from the Renaissance, Leonardo da Vinci, was the first to study human anatomy through dissections of at least 30 cadaveric bodies. He was particularly interested in the structure of the human body as it relates to performance, center of gravity, balance and center of resistance. He identified muscles and nerves in the human body and described the mechanics of the body during standing, walking up- and downhill, rising from a sitting position, jumping, and human gait. He also suggested that cords be attached to a skeleton at the points of origin and insertion of the muscles to demonstrate the progressive action and interaction of various muscles during movement.

Even though Leonardo da Vinci had produced very detailed descriptions of the human body, it was not until the mid-16th century when Andreas Vesalius published the first anatomy book, *De Humani Corporis Fabrica*, which gave him the credit of being "The Father of Modern Anatomy". Another two chief figures of the Renaissance were Galileo Galilei and Giovani Alfonso Borelli. While Galilei applied mechanical theory to study animal movement, and published a treatise *De Animaliam Motibus (The movement of Animals)*, Borelli published *De Motu Animalum (On the Motion of Animals)* in 1680 which successfully clarified muscular movement and body dynamics. Borelli also estimated the center of mass of the entire body by stretching the body out on a rigid platform that was supported on a knife edge and then repositioned until it was balanced. Borelli is often considered the "Father of Biomechanics".

During the age of enlightenment, Wilhelm Eduard Weber and his younger brother, Eduard Friedrich Weber published the results of their collaborative study on the mechanism of walking in mankind in 1836. Since then, the study of human motion has greatly progressed from an observatory/descriptive science to one based on quantitative measurements, for which Étienne-Jules Marey made the most important breakthrough. He determined a series of actions of human locomotion in various forms according to measurements of the effort exerted at each moment using graphical methods, and glass-plate and celluloid film chronophotography. Carlet added a heel and separate forefoot chambers to Marey's pressure-recording shoes, obtaining more measurements of the onset and duration of weight-bearing, and the vertical reaction force.

Among the noted scientists, Eadweard Muybridge began what was probably the first assessment of gait and deserves to be considered the "Father of Modern Gait Analysis". Since Muybridge found that it was not possible to capture the rapid limb movements of horses in motion by eye, he improved photography by creating a camera with a shutterspeed of up to 1/100 of a second and recorded the motion in men, women, children, animals and birds. With the aid of computer vision and the techniques of pattern recognition and artificial intelligence, photogrammetry using photographs, radiographs, and video images continued to develop after Muybridge's invention. Stereophotogrammetry, the technique of measuring 3D landmark coordinates, was then

developed by the "Father of Stereophotogrammetry", Carl Pulfrich. Christian Wilhelm Braune and Otto Fischer then used analytical close-range photogrammetry, combined with the geometrical properties of central projection from multi-camera observations, to estimate the 3D position data from digitized and noisy image data. In the 1890s, they used stereophotogrammetry and ground reaction force (GRF) measurements to study the biomechanics of human gait under loaded and unloaded conditions using their pioneering 3D mathematical technique based on Newtonian mechanics. They also carefully studied the mass, volume center of mass, and body segments of three adult male cadavers and introduced the use of regression equations for estimating body segment parameters, based on the length and mass of body segments. To date, the mathematical methodology of gait analysis developed by Braune and Fischer has essentially remained unchanged in modern gait analysis. It is noted that during the age of Enlightenment, computers had not yet been developed. Therefore, manual involvement was necessary for determining the specific markers on the human body in each image, which was not only laborious and time-consuming, but also one of the main sources of errors. This inevitably limited the clinical application of the mainly two-dimensional (2D) measurement and analysis of the human motion during that period.

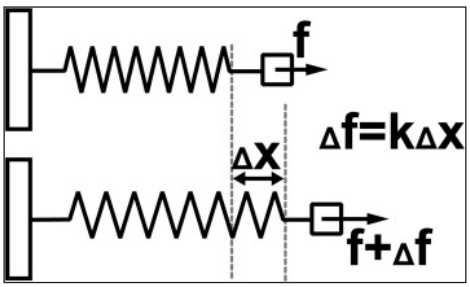
In the modern era, human motion analysis developed rapidly as the knowledge of anatomy and mechanics, and measurement technology was progressively established. In the 1940s, Harold Eugene Edgerton pioneered high-speed stroboscopic photography that was used to photograph objects in motion at a frequency of several million exposures per second. During the 1970s, video camera systems, such as infrared high-speed cameras, began their widespread application in the analysis of pathological gait, producing detailed motion analysis results within realistic cost and time constraints. With the collocation of high-speed computers and video camera systems, 3D analysis of human motion became feasible. However, it had to wait until after World War II to make its debut in clinical 3D applications.

After the war there were many retired soldiers who had sustained limb injuries and who needed orthopedic treatment, prostheses, orthoses and subsequent rehabilitation for recovery of functional activities, especially for level walking. In order to provide better medical services and achieve treatment goals, numerous investigators devoted themselves to the study of gait analysis for clinical applications. Among them was Verne Thompson Inman who began by applying the theory of mechanical engineering to clinical problems, such as designing prostheses for amputees. He studied the biomechanics of locomotion and proved the assumption that the most efficient gait pattern is achieved by minimizing vertical and lateral excursions of the body's center of gravity (COG). He also identified the so-called gait determinants for normal walking, i.e., features of the movement pattern that minimizes these COG excursions. He suggested that these features be used to determine whether a movement pattern is normal or pathological. Following Inman's work, Jacquelin Perry divided the gait cycle into five stance phase periods and three swing phase periods. David Sutherland refined the

definition of the gait cycle to have three periods of stance, namely initial double support, single limb stance, and second double support. He also carried out a comprehensive investigation on the walking patterns in a total of nearly 300 normal children aged from 1 to 7 years for the study of the development of mature gait.

Because of the tedious nature of processing and analyzing cine film, the need for a more scientific approach to automate the process led to the development of the 3D Vicon motion capture system by Professor John P. Paul and his PhD students, Mick Jarrett and Brian Andrews. This system was made to capture human movement data in numerical digits instead of analog data, and which is now widely used in the study of human motion.

Apart from kinematic measurements using video cameras, Dr J. Robert Close used a 16-mm movie camera with a sound track for studying the phasic action of muscles in subjects after muscle transfers for poliomyelitis. Doctor Close was the first to record synchronously the kinesiological electromyography (KEMG) of one muscle and kinematic data on cine film. Jacquelin Perry pioneered using fine wire electrodes to record the gait electromyogram (EMG) and used it as a primary clinical tool in determining the appropriateness of surgical procedures to correct gait deformities. Because muscles are the engines for producing active movements, EMG has been a useful assessment tool for detecting the electrical activity of specific muscles and assessing their contribution to movement or gait. Between 1944 and 1947, Vern Inman and colleagues added KEMG to other measurements, i.e., 3D force and energy, in the study of walking in normal subjects and amputees, and thus significantly moved the science of gait analysis forward.



Determining the nonmeasurable force (Δf) in the spring from the measured deformation of the spring (Δx) using Hooke's law.

The essential aim of human motion analysis is to understand the mechanical function of the musculoskeletal system during the execution of a motor task. Since the forces for generating movement in the musculoskeletal system are too difficult to measure noninvasively, combined experimental and mathematical modeling approaches have been used. The power of mathematical modeling is that it enables the values of parameters which are difficult or impossible to measure to be calculated from the values of quantities which can be measured. An example of this approach is determining the force in a spring which cannot be measured directly. With the measureable deformation of

the spring, the force in the spring can be calculated using Hooke's law that relates the deformation with the force in the spring. Therefore, numerous studies have used mathematical modeling in conjunction with noninvasive experimental measurements to calculate nonmeasurable internal forces in the musculoskeletal system through inverse dynamics analysis.

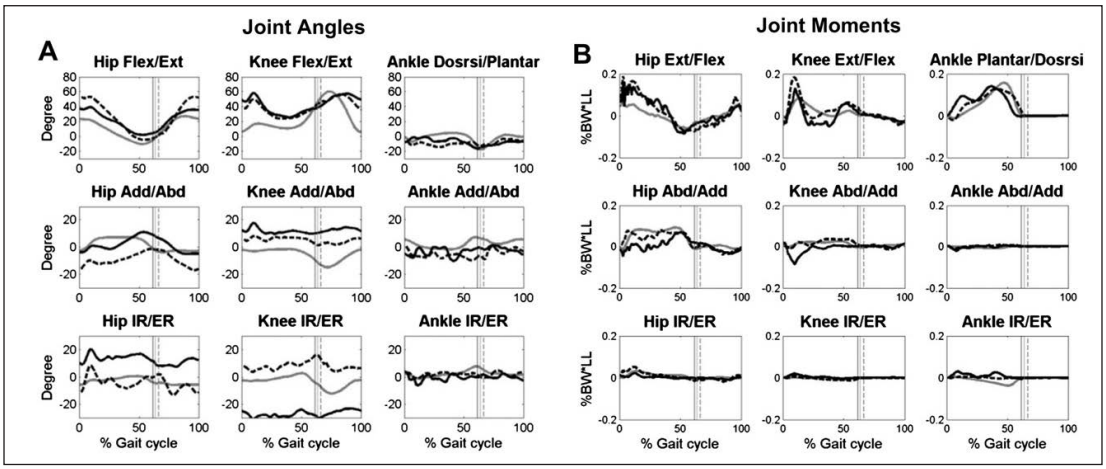
The measurable values of quantities in human motion analysis are usually the motion of the musculoskeletal system defined by skin markers and measured by the motion capture systems, and the external forces applied to this musculoskeletal system measured by force plates. With the 3D trajectories of skin markers obtained using stereophotogrammetry, and the GRFs and center of pressure (COP) measured using force plates, intersegmental forces and internal moments at the joints of the lower limbs are then calculated from the solution of equations based on Newton's laws of motion. This approach is called inverse dynamics analysis.

Advances in Techniques of Human Motion Analysis

Stereophotogrammetry-based human motion analysis has been widely used in the diagnosis and subsequent planning and evaluation of treatment of neuromusculoskeletal pathology. However, there is still room for technological improvement, including mathematical modeling of the musculoskeletal system, quantitative validation of the models for internal force estimation, and techniques to quantify and minimize measurement errors, i.e., soft tissue artifacts (STA). New technology for accurately determining the skeletal motion of a joint is also needed to supplement the measurement of the gross motion by the stereophotogrammetry systems.

One of the most successful applications of clinical gait analysis is the surgical planning in cerebral palsy (CP). A previous study of 70 CP patients showed that after clinical gait analysis 89% of the original treatment plans were altered and 39% of the recommended procedures were not done. However, this relies on extensive team work in the interpretation of a huge bulk of data derived from 3D motion analysis, and which has been a huge obstacle for applying gait analysis in a clinical setting. Computer graphics-based models may be useful in providing easy access to and visualization of gait analysis results by creating a "look and feel" environment for the user to examine the experimental data and interpret the analytical results with relatively less effort. Traditionally, the musculoskeletal system was modeled mathematically as a multi-link system with individualized model parameters, such as the length of each segment, the joint centre positions, and the lines of action of the muscles and tendons. The joints were often modeled as ball-and-socket joints. This simplification made it difficult to describe the precise motion of the joint, which directly affected the accuracy of the lines of action and lever arms of the surrounding muscles and tendons, and thus the estimated internal forces. To address this problem, a mathematical model of the human pelvis-leg system in the sagittal plane, with an anatomical model of the knee, was developed to calculate forces transmitted by the structural elements of the system.

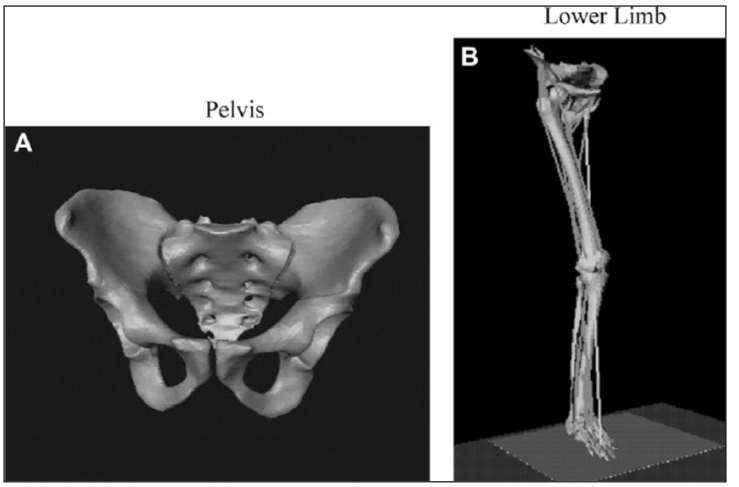
The sagittal plane model underestimates the value of the maximum axial force in the femur during walking by about 30% but suggests that 70% is due to the action of the extensors or flexors. With advances in computer graphics technology, a complete 3D computer graphics-based geometrical model of the locomotor system was developed with the anatomical structures reconstructed from images produced by computed tomography (CT) and magnetic resonance imaging (MRI). While information in the anatomical plane of interest is helpful for clinical reference, a complete 3D data display provides more realistic support for clinical decision-making because the locomotor system itself and its responses to the internal and external forces are 3D in nature. In addition, it would not be an easy task for clinicians to reconstruct 3D knowledge of a complex movement by themselves from individual components without any adequate technical support. In this regard, 3D solid modeling and image display could be a good tool for presenting results and surgical simulation.



Traditional gait output of (A) joint angles and (B) joint moments at the hip, knee and ankle of the right limb (black, solid line) and left limb (black, dotted line) in a typical patient with spastic diplegia cerebral palsy and in that of the healthy controls (gray, solid line) during level walking. BW: Body weight; LL: leg length.

Success of any theoretical model relies heavily on the validation of its assumptions. The geometrical model of the locomotor apparatus developed by the author was validated by EMG and telemetered force data from two instrumented patients. The model was used to study the influence of activity of hip flexors and extensors on the forces in the femur during isometric exercises and during level walking. Kinematic and kinetic data, together with simultaneous electromyography (EMG) and *in vivo* axial forces transmitted along the prostheses from two patients implanted with instrumented massive proximal femoral prostheses, were obtained. A comparison of the levels of the calculated axial forces in the femur model agreed well with the simultaneous telemetered forces for isometric tests. Interaction between the muscles and the bones during isometric tests was examined and biarticular muscles were shown to

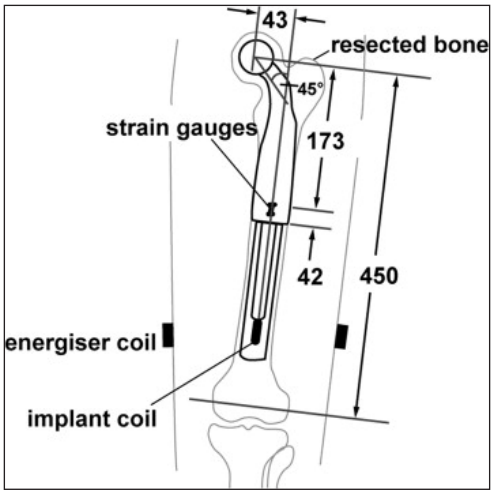
play a major role in modulating forces in bones. The study supports the hypothesis that muscles balance the external limb moments, not only at joints, but also along the limbs, decreasing the bending moments but increasing the axial compressive forces in bones. It is thus suggested that appropriate simulation of muscle force is necessary in *in vitro* laboratory experiments and in theoretical studies of load transmission in bones. Knowledge of the forces and moments transmitted by the bones is essential for the design and fixation of implants and their preclinical testing. There is a similar need to understand the effects of the mechanical environment on fracture repair and limb lengthening procedures. Quantitative validation of internal forces provided in this study is of great help for the design of prostheses and related rehabilitation applications.



Reconstruction of a 3D geometric computer model of (A) the pelvis and (B) the locomotor system from computed tomography.

Success of any theoretical model relies heavily on the validation of its assumptions. The geometrical model of the locomotor apparatus developed by the author was validated by EMG and telemetered force data from two instrumented patients. The model was used to study the influence of activity of hip flexors and extensors on the forces in the femur during isometric exercises and during level walking. Kinematic and kinetic data, together with simultaneous electromyography (EMG) and *in vivo* axial forces transmitted along the prostheses from two patients implanted with instrumented massive proximal femoral prostheses, were obtained. A comparison of the levels of the calculated axial forces in the femur model agreed well with the simultaneous telemetered forces for isometric tests. Interaction between the muscles and the bones during isometric tests was examined and biarticular muscles were shown to play a major role in modulating forces in bones. The study supports the hypothesis that muscles balance the external limb moments, not only at joints, but also along the limbs, decreasing the bending moments but increasing the axial compressive forces in bones. It is thus suggested that appropriate simulation of muscle force is necessary in *in vitro* laboratory experiments and in theoretical studies of

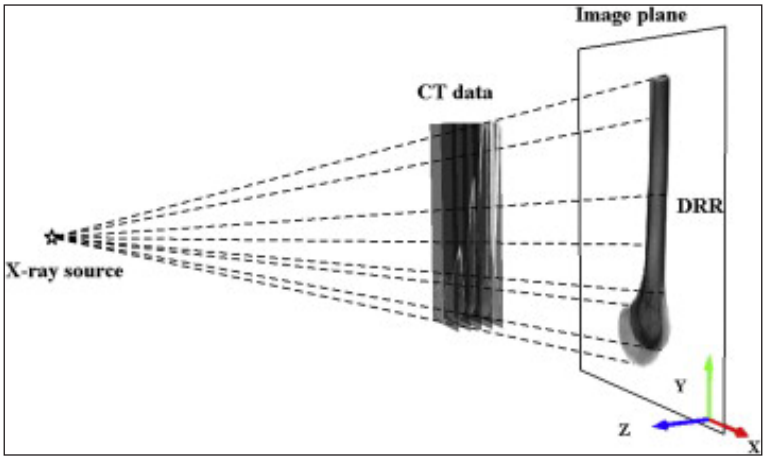
load transmission in bones. Knowledge of the forces and moments transmitted by the bones is essential for the design and fixation of implants and their preclinical testing. There is a similar need to understand the effects of the mechanical environment on fracture repair and limb lengthening procedures. Quantitative validation of internal forces provided in this study is of great help for the design of prostheses and related rehabilitation applications.



Instrumented massive proximal femoral prosthesis that enables the measurement of femoral axial forces *in vivo*.

Knowledge of accurate kinematics of natural human joints, including 3D rigid body and surface kinematics, is essential for the understanding of their function and for many clinical applications. For this purpose, an accurate measurement method for the kinematics of skeletal motion is needed. Several techniques are available for measuring 3D joint kinematics, but few allow noninvasive measurement with submillimeter accuracy. Imaging methods, such as MRI and CT may be used to provide 3D geometry and poses of bones, but they are limited to static and nonweight-bearing conditions. Skin marker-based methods, including stereophotogrammetry and electromagnetic tracking systems have been widely used in clinical settings for measuring the 3D kinematics of the human body during functional activities. However, STAs are difficult to prevent without the use of invasive bone pins. STAs are characterized by marker movement in relation to the underlying bone caused by skin deformation and displacement, and have been regarded as the most critical source of error in human movement analysis because of their great influence on the estimation of skeletal system kinematics. In order to reduce the effects of STA, several techniques have been used for measuring 3D skeletal kinematics, including the use of external fixators, intracortical pins and percutaneous tracking devices, medical imaging methods, as well as model-based registration methods. While the use of invasive approaches enables direct measurement of the movement of bones and thus the STA of the overlying skin markers, they affect the motion of the subject and alter the soft tissue motion. The risk of infection when applying these approaches is also of great concern. Medical imaging methods, such as MRI

and X-ray fluoroscopy, allow noninvasive measurement of unrestricted joint motion *in vivo* but they are limited to 2D or static measurements.

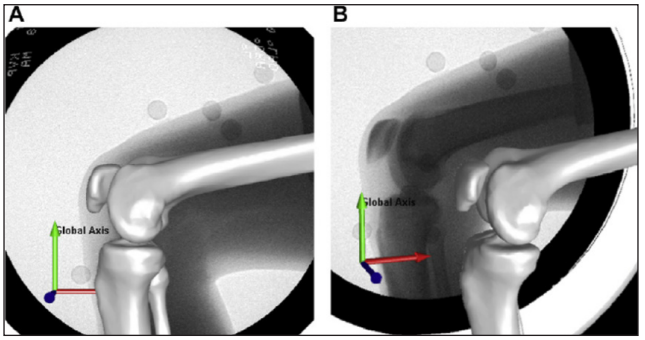


Perspective projection model of the fluoroscopy system. X- and Y-axes define the image plane. Rays of the point source X-ray passing through the CT volume are projected onto the image plane to form a DRR.

Surface model-based registration methods using dynamic fluoroscopy systems have been proposed for the accurate estimation of 3D poses (positions and orientations) of knee prosthesis components. All these approaches work by registering the known 3D CAD surface models of the prosthesis component to the dynamic fluoroscopic images. The surface model of a prosthesis component is projected onto the fluoroscopic image plane, and the pose of the component is then determined as the 3D pose of the component model that gives the best correspondence between the fluoroscopic image and the projected contours and areas of the model component. These methods have been shown to have high accuracy because the metallic components have precisely known geometric features and produce sharp edges in fluoroscopic images. The method has also been assumed to be applicable to natural knee joints but bones differ fundamentally from total knee replacement (TKR) components in their form and internal structure, resulting in completely different fluoroscopic images. Bone edges are less well-defined and it has been suggested that bone edge attenuation is the primary factor limiting the theoretical accuracy of this type of method. Under real life conditions, the accuracy would likely be worse. While methods using bi-planar fluoroscopy with better accuracy than single-plane fluoroscopy have been proposed, biplanar fluoroscopy methods inevitably increase radiation doses and unacceptably constrain the motion of the patient.

In order to bridge the gap in the *in vivo* measurement of knee kinematics, the author and colleagues have developed a volumetric model-based 2D to 3D registration method with a new similarity measure, called the weighted edge-matching score (WEMS), for measuring natural knee kinematics with single-plane fluoroscopy. The 3D poses of the bones were obtained by registering their volumetric CT models to the corresponding fluoroscopy images using digitally reconstructed radiographs (DRR). At each image

frame, an optimization procedure was used to find the pose of the bone using the DRR which best matched the fluoroscopic image according to the WEMS similarity measure. The performance of this registration method in the measurement of the knee poses was previously evaluated using a cadaveric knee. The means and standard deviations of the discrepancies between the measurements and the gold standard were 0.24 ± 0.77 mm, 0.41 ± 3.06 mm and 0.59 ± 1.13 mm for in-plane translation, out-of-plane translation, and all rotations, respectively.



A volumetric model-based 2D to 3D registration method with a new similarity measure, called the weighted edge-matching score (WEMS), for measuring natural knee kinematics with single-plane fluoroscopy. (A) Lateral view and (B) oblique view.

Since STA is critical in the measurement accuracy of skin marker-based stereophotogrammetry, it is necessary to quantify the STA during functional activities for establishing guidelines for selecting marker locations and for developing compensation methods in human motion analysis. Three-dimensional movements of skin markers relative to the underlying bones in normal subjects during functional activities were measured for the first time using a noninvasive method based on integrating 3D fluoroscopy (WEMS) and stereophotogrammetry. Generally, thigh markers had greater STA than shank markers, and the STA of markers close to the knee joint were greater than those away from the joint. The STA of the markers showed nonlinear relationships with knee flexion angles and some of their patterns were different between activities. The STA of a marker also appeared to vary among subjects and was affected by more than one joint. These results suggest that correction of STA of individual markers during a functional activity using a linear error model based on STA measured from static isolated joint positions may not produce satisfactory results. In addition to improving the experimental measurements by judiciously determining marker placement, compensation of STA in human motion analysis may have to consider the multijoint nature of functional activities by using a global compensation approach with individual anthropometric data.

Techniques for reducing the effects of STA can be divided up into those which model the skin surface and those which include joint motion constraints. Traditional segment-based methods treat each body segment separately without imposing joint

constraints, resulting in apparent dislocations at joints predominantly because of STA. Including joint constraints in a global STA minimization approach, called a global optimization method (GOM), is regarded as an effective solution for reducing the effects of STA on the skeletal system kinematics. The GOM, first proposed by the author and colleagues, is based on minimizing the weighted sum of the squared distances between measured and model-determined marker positions with joint constraints for simultaneously determining the spatial pose of all segments of a multi-link model of the locomotor musculoskelatal system. Numerical experiments were used to show that the GOM was capable of eliminating joint dislocations and giving more accurate model position and orientation estimates. It was suggested that, with joint constraints and a global error compensation scheme, the effects of measurement errors on the reconstruction of the musculoskeletal system and subsequent mechanical analyses could be reduced globally. The GOM minimizes errors not only in flexion/extension at joints, but also in axial rotations and ab/adductions, and this improvement contributes to extending the applicability of gait analysis to clinical problems.

Applications of Human Motion Analysis

With the development of theoretical and experimental methods to improve accuracy and reliability, human motion analysis has become a useful investigative and diagnostic tool in many research and clinical areas, such as medicine, ergonomics and sports, to name but a few. Through human motion analysis, the deviations from normal movement in terms of the altered kinematic, kinetic or EMG patterns can be identified and then used to evaluate the neuromusculoskeletal conditions, to help with subsequent treatment planning and to assess the efficacy of treatment in various patient groups, such as those with CP, stroke, knee osteoarthritis (OA), diabetes mellitus (DM) and spinal cord injury (SCI). Human motion analysis has also seen many applications in assistive technology, such as prosthetics and orthotics, in which accurate evaluation of critical joint motion characteristics can be obtained from human motion measurements. In sports science and medicine, human motion analysis is also widely used to help optimize athletic performance and to identify mechanisms of common sports injuries and the accompanied posture-related or movement-related problems.

Evaluation of Musculoskeletal Pathology and Assessment of Treatment Efficacy

Pathological motion is a result of the response and compensations of the neuromusculoskeletal system for the underlying pathologies. Analysis of the deviations of motion should thus permit an evaluation of the pathology to be made, and should be helpful for subsequent planning and assessment of treatment. Many leading orthopedic hospitals worldwide now have gait laboratories which are routinely used in a large number of cases, both to design treatment plans, and for follow-up monitoring.

Based on gait analysis results, the development of treatment regimes advanced significantly in the 1980s, including in orthopedic surgery. For example, gait analysis has been used in assisting with evaluating the pathological condition and with surgical planning for children with CP. CP is characterized by impaired motor control affecting movement and posture with bony deformities and changes in the joint and muscle properties. Common treatment options for the paralysis of spastic muscles in patients with CP include Botox injection or the lengthening, re-attachment or detachment of particular tendons. Corrections of distorted bony anatomy are also commonly undertaken in these patients. It has been shown that motion analysis data can alter recommendations for surgical intervention and ultimately reduce the amount of surgery in children with CP. From 2005, in cooperation with the Pediatric Division, Department of Orthopedic Surgery, National Taiwan University Hospital, clinical gait analysis services have been offered for patients with neuromusculoskeletal pathology at the Motion Analysis Laboratory at the Institute of Biomedical Engineering, National Taiwan University. A team of orthopedic surgeons, rehabilitation physicians, physical therapists, occupational therapists and biomechanical engineers work together to support the service.

Motion analysis techniques have been used to study the effects of the severity of knee OA on the compensatory gait patterns and whether these gait changes would help unloading the diseased knee. Knee OA, resulting in biomechanical changes of the lower limb during level walking, affects an increasing proportion of the population. Compared to the normal group, patients with mild and severe bilateral medial knee OA all had obviously increased pelvic anterior tilt, swing-pelvis list, decreased standing knee abduction, as well as decreased standing hip flexor and knee extensor moments. The severe group also demonstrated increased hip abduction, knee extension and ankle plantarflexion. For the mild group, listing and anterior tilting of the pelvis were responsible for successfully reducing the extensor moment and maintaining a normal abductor moment at the diseased knee. At the expense of a greater hip abductor moment and with extra compensatory changes at other joints, the severe group successfully diminished the knee extensor moment but failed to decrease the abductor moment. These results suggest that, in addition to training of the knee muscles, training of the hip muscles and pelvic control are essential for rehabilitating patients with knee OA, especially for patients with more severe OA.

Gait analysis was also useful for investigating the effects of treatment on gait patterns in patients with knee OA. While TKR has been the main choice of treatment for advanced knee OA, acupuncture has been a popular alternative in managing patients with mild to moderate knee OA. Acupuncture has been shown to be effective in pain relief and anesthesia, and has been suggested in traditional Chinese medicine for treating various types of functional disabilities, including knee OA. After acupuncture stimulation, patients with bilateral medial compartment knee OA reported significantly reduced pain and walked with higher speed and greater step length, as well as better body weight transfer through increased dynamic joint ranges of motion and increased joint moments. The differences between the two groups after treatment suggest that the

significantly improved gait performance in the experimental group may be associated with the pain relief after treatment, but the relatively small decrease in pain in the sham group, even without subjective factors involved, was not enough to induce significant improvements in gait patterns. Gait analysis combined with the visual analog scales can be useful for evaluating the true effect of acupuncture treatment for patients with neuromusculoskeletal diseases and movement disorders. The study not only demonstrated the short-term effects of acupuncture treatment on knee OA, but also justified the need for study of its long-term effects.

Another example is the study of patients with anterior cruciate ligament deficiency (ACLD) during obstacle-crossing. The ACL has both structural and proprioceptive functions so ACLD is known to lead to instability, a decrease of the muscular strength, and impaired somatosensory feedback of the knee. Obstacle-crossing presents an ideal motor task in which both the structural and proprioceptive functions of the ACL can be evaluated because a safe and successful obstacle-crossing requires stability of the body provided mainly by the stance limb and sufficient foot clearance of the swing limb. The former depends on the stability of the joints while the latter emphasizes the sensory function of the joints. These demands may not be met in ACLD patients who have impairments in both structural stability and sensory feedback of the affected joint, or patients who have restored structural stability but have residual impaired sensory feedback of the affected joint after anterior cruciate ligament reconstruction (ACLR).

Therefore, detailed analysis and study of the ACLD subjects during obstacle-crossing would be helpful for establishing a more complete picture of the integration and interaction of the structural and sensory functions of the ACL during functional activity. In the study by Lu et al., patients with ACLD were found to show statistically the same toe-obstacle clearance, crossing speeds, heel-obstacle distances and toe-obstacle distances when crossing obstacles with the unaffected limb leading compared to healthy controls. However, they showed significantly increased peak hip extensor and ankle plantarflexor moments and decreased peak knee extensor moments in the affected stance limb, with significant increases in the anterior tilt of the pelvis and flexion at both hips when the unaffected leading toe was above the obstacle. These results suggest that patients with ACLD adopted a strategy of decreasing the peak knee extensor moments (quadriceps net effort) in order to avoid anterior displacement of the tibia of the affected trailing stance limb during obstacle-crossing. In order to prevent quadriceps contraction, patients with ACLD may have to shift the center of body mass forward and thus cause greater pelvic anterior tilt and hip flexion, both in the swing and stance limb, with normal leading toe-clearance for safe obstacle-crossing.

These examples demonstrate the use of motion analysis in evaluating pathology of the neuromusculoskeletal system and assessing treatment. The analysis results are also helpful for improving the management of relevant patient populations.

Prevention of Injury, Design of Prostheses and Orthoses in Rehabilitation

Since Dr. Vern Inman initially applied gait analysis to lower limb prostheses research, it has been clear that study of gait not only allows a better understanding of the neuromusculoskeletal compensations for underlying pathologies, but also helps identify the mechanisms of abnormal gait. Study of pathological and assisted gait enable the evaluation of the efficacy of orthoses, such as functional knee braces and lateral-wedged insoles, and the design and development of fall-prevention strategies during obstacle-crossing.

Use of functional knee braces has been suggested to provide protection and to improve kinetic performance of the knee in ACL-injured patients. However, the efficacy of knee bracing in achieving these goals is still controversial. Lu et al. compared the 3D kinetics of the knee between bracing conditions and between limbs in order to understand the immediate effects of functional bracing on walking performance in individuals with ACL injuries, including ACLD and ACLR subjects. They found that functional knee bracing did not significantly affect the kinetics of the unaffected knees for either group. For the ACLD group, bracing significantly increased the peak abductor moments in the affected knees and reduced the bilateral kinetic asymmetry in the frontal plane. With functional knee braces, significantly greater peak moments and impulses of the abductors and extensors, and reduced bilateral kinetic asymmetry in the sagittal and frontal planes, were found in the ACLR group. For ACLR patients, functional bracing can be recommended to help achieve better bilateral kinetic symmetry during gait. For ACLD patients, in addition to bracing, supplementary emphasis on rehabilitative exercise for better kinetic knee performance in the sagittal plane is needed.

Hemiparetic subjects after a cerebral vascular accident or stroke may show reduced functional walking ability owing to impaired ankle and knee control. In order to increase weight bearing over the paretic side during stance, and to correct abnormal gait patterns during walking, an ankle-foot orthosis on the paretic limb is frequently prescribed for these patients. Motion analysis was used to assess the immediate effects of a 5° lateral-wedged insole, applied to either the nonparetic or the paretic side, on the weight-bearing symmetry and the moment changes of lower-limb joints in stroke patients with hemiparesis. The use of a lateral-wedged insole over the nonparetic side was shown to improve the stance symmetry and tended to reduce the paretic knee abductor loads. Applying the wedged-insole on the paretic side during ambulation would significantly decrease the abductor moments at the ipsilateral hip and knee when compared to the contralateral side (nonparetic side).

Falls as a result of unsuccessfully negotiating obstacles often lead to physical injuries which may result in great costs including medical and social expenses. Knowledge of the mechanics of the locomotor system and control strategies adopted during this activity is helpful for understanding and identifying the risk factors for tripping which are important for preventing falls, for the design of fall-prevention devices and for planning

programs aimed at preventing trip-related falls. Because the elderly and older patients with knee OA, highly functioning patients subsequent to stroke and patients with type II DM are shown to be highly prone to falling, the author studied the biomechanical differences between the elderly or these patient groups and healthy young subjects in this functional activity. Apart from using traditional kinematic and kinetic analyses, a novel approach, namely quantification of the interjoint coordination using the continuous relative phase (CRP) method, was used to study obstacle-crossing. Interjoint coordination provides information on how the neuromusculoskeletal system organizes the redundant degrees of freedom (DOF) of the joints in order to achieve a smooth, efficient and accurate functional movement. Therefore, analyses of the patterns and variability of the interjoint coordination were performed to gain more insight into the control of the locomotor system in normal adults, and to identify the control deficits in the elderly and patients with knee OA during obstacle negotiation.

Sports Medicine

Sports biomechanics is a very active area within the field of biomechanical research. The specific goals of sports biomechanics research include performance enhancement, injury prevention and safety for many elite, leisure and rehabilitation sports. Motion analysis plays a key role in professional sports training, aiming to optimize and improve athletic performance. Examples include determining joint loadings in the lower extremities during elliptical exercises and their changes with different pedal rates, step lengths and step heights and study of the joint kinematics of the lower limbs and the COP movements in wrestlers during tackle defense.

Elliptical exercise (EE) has been developed as a low-impact aerobic exercise modality and has been shown to be beneficial for developing and maintaining cardiorespiratory fitness. Over the last decade it has increased in popularity in fitness training and clinical applications. During EE, the feet are constrained by pedals to follow an elliptical trajectory which could lead to the possibility of producing disadvantageous joint loads. In addition, the forces transmitted in the lower limbs were found to be closely related to potential musculoskeletal overuse injuries. Therefore, complete knowledge of the loading in the lower limbs during EE is helpful for improving the design of elliptical trainers (ET) and is essential for ensuring an efficient and safe exercise environment, especially for patients. However, little is known regarding the loadings applied to the lower limb during EE. The author used a detailed 3D dynamic analysis of the lower extremities to determine the differences in lower-limb kinematics and kinetics between EE and level walking. The results showed that EE had smaller loading rates around heelstrike when compared to level walking. Reduced vertical pedal reaction force (PRF) and loading rates during EE were achieved by greater compensatory hip flexor and knee extensor moments. Therefore, users' joint function and muscle strength, especially at the knee, must be considered in order to avoid injuries when using ET for athletic and rehabilitative training.

The effects of pedal rates, step height and step length on the biomechanics of lower limbs during EE were also studied. With increasing pedal rates, the medial, anterior and posterior PRF, as well as the medial and vertical loading rates, all increased during EE. As step length increases, harmful joint loadings, especially at the hip joint, could increase during EE. When step height increased, a more flexed posture achieved by greater peak hip and knee flexion and hip abduction during swing phase were used to compensate for the change in pedal trajectory and to keep the body stable. Increased hip extensor and decreased knee extensor moments in late swing then occurred in the more flexed posture in conjunction with reduced posterior shear forces that could shift the line of action of the PRF more anterior to the hip joint center and less posterior to the knee joint center. Therefore, the harmful joint loadings, in particular at the knee joints, and the corresponding risks of injuries may be reduced during EE with increasing step heights.

Wrestling is one of the oldest and most popular competitive sports in the world. However, knowledge of the biomechanics of wrestling is limited and the biomechanical risk factors of injuries are still unclear. The author used 3D motion analysis to investigate the joint kinematics of the lower limbs and the COP movements in Greco-Roman style (GR) and free style (FS) wrestlers during defense tackle from three different directions. The results showed that the wrestlers who majored in GR had the tendency to resist tackle attacks longer than the FS group. The FS wrestlers tended to have greater anterior-posterior (A/P) motion of the COP with significant increased knee flexion when compared to the GR group. This flexed knee strategy may be related to the training the FS wrestlers received and the rules of the sport. Significantly increased transverse- and frontal-plane joint angles at the knee and ankle may be the risk factors related to the knee and ankle injuries commonly found in wrestlers. Therefore, it was suggested that strengthening of the lower-limb muscles may be helpful for reducing these injuries during wrestling competitions. These examples show that motion analysis is an efficient technique for identifying beneficial and damaging factors in performing sports and exercises.

Biomechatronics

Biomechatronics is an applied interdisciplinary science that aims to integrate biology, mechanics, and electronics. It also encompasses the fields of robotics and neuroscience. Biomechatronic devices encompass a wide range of applications from the development of prosthetic limbs to engineering solutions concerning respiration, vision, and the cardiovascular system.

How it Works?

Biomechatronics mimics how the human body works. For example, four different steps must occur to be able to lift the foot to walk. First, impulses from the motor center of the

brain are sent to the foot and leg muscles. Next the nerve cells in the feet send information, providing feedback to the brain, enabling it to adjust the muscle groups or amount of force required to walk across the ground. Different amounts of force are applied depending on the type of surface being walked across. The leg's muscle spindle nerve cells then sense and send the position of the floor back up to the brain. Finally, when the foot is raised to step, signals are sent to muscles in the leg and foot to set it down.

Biosensors

Biosensors are used to detect what the user wants to do or their intentions and motions. In some devices the information can be relayed by the user's nervous system or muscle system. This information is related by the biosensor to a controller which can be located inside or outside the biomechatronic device. In addition biosensors receive information about the limb position and force from the limb and actuator. Biosensors come in a variety of forms. They can be wires which detect electrical activity, needle electrodes implanted in muscles, and electrode arrays with nerves growing through them.

Mechanical Sensors

The purpose of the mechanical sensors is to measure information about the biomechatronic device and relate that information to the biosensor or controller.

Controller

The controller in a biomechatronic device relays the user's intentions to the actuators. It also interprets feedback information to the user that comes from the biosensors and mechanical sensors. The other function of the controller is to control the biomechatronic device's movements.

Actuator

The actuator is an artificial muscle. Its job is to produce force and movement. Depending on whether the device is orthotic or prosthetic the actuator can be a motor that assists or replaces the user's original muscle.

Balance

In biomechanics, balance is an ability to maintain the line of gravity (vertical line from centre of mass) of a body within the base of support with minimal postural sway. Sway is the horizontal movement of the centre of gravity even when a person is standing still. A certain amount of sway is essential and inevitable due to small perturbations within the body (e.g., breathing, shifting body weight from one foot to the other or from forefoot to rearfoot) or from external triggers (e.g., visual distortions, floor translations).

An increase in sway is not necessarily an indicator of dysfunctional balance so much as it is an indicator of decreased sensorimotor control.

A woman demonstrating the ability to balance.

Maintaining balance requires coordination of input from multiple sensory systems including the vestibular, somatosensory, and visual systems.

- Vestibular system: Sense organs that regulate equilibrium (equilibrioception); directional information as it relates to head position (internal gravitational, linear, and angular acceleration).
- Somatosensory system: Senses of proprioception and kinesthesia of joints; information from skin and joints (pressure and vibratory senses); spatial position and movement relative to the support surface; movement and position of different body parts relative to each other.
- Visual system: Reference to verticality of body and head motion; spatial location relative to objects.

A waiter balancing wine glasses.

The senses must detect changes of spatial orientation with respect to the base of support, regardless of whether the body moves or the base is altered. There are environmental factors that can affect balance such as light conditions, floor surface changes, alcohol, drugs, and ear infection.

There is balance impairments associated with aging. Age-related decline in the ability of the above systems to receive and integrate sensory information contributes to poor balance in older adults. As a result, the elderly are at an increased risk of falls. In fact, one in three adults aged 65 and over will fall each year. In the case of an individual standing quietly upright, the limit of stability is defined as the amount of postural sway at which balance is lost and corrective action is required.

Body sway can occur in all planes of motion, which make it an increasingly difficult ability to rehabilitate. There is strong evidence in research showing that deficits in postural balance is related to the control of medial-lateral stability and an increased risk of falling. To remain balanced, a person standing must be able to keep the vertical projection of their center of mass within their base of support, resulting in little medial-lateral or anterior-posterior sway. Ankle sprains are one of the most frequently occurring injuries among athletes and physically active people. The most common residual disability post ankle sprain is instability along with body sway. Mechanical instability includes insufficient stabilizing structures and mobility that exceed physiological limits. Functional instability involves recurrent sprains or a feeling of giving way of the ankle. Nearly 40% of patients with ankle sprains suffer from instability and an increase in body sway. Injury to the ankle causes a proprioceptive deficit and impaired postural control. Individuals with muscular weakness, occult instability, and decreased postural control are more susceptible to ankle injury than those with better postural control.

Balance can be severely affected in individuals with neurological conditions. People who suffer a stroke or spinal cord injury for example, can struggle with this ability. Impaired balance is strongly associated with future function and recovery after a stroke, and is the strongest predictor of falls.

Another population where balance is severely affected is Parkinson's disease patients. A study done by Nardone and Schieppati showed that individuals with Parkinson's disease problems in balance have been related to a reduced limit of stability and an impaired production of anticipatory motor strategies and abnormal calibration.

Balance can also be negatively affected in a normal population through fatigue in the musculature surrounding the ankles, knees, and hips. Studies have found, however, that muscle fatigue around the hips (gluteals and lumbar extensors) and knees have a greater effect on postural stability (sway). It is thought that muscle fatigue leads to a decreased ability to contract with the correct amount of force or accuracy. As a result, proprioception and kinesthetic feedback from joints are altered so that conscious joint awareness may be negatively effected.

Balance Training

Since balance is a key predictor of recovery and is required in so many of our activities

of daily living, it is often introduced into treatment plans by physiotherapists and occupational therapists when dealing with geriatrics, patients with neurological conditions, or others whom they have determined it to be beneficial.

Balance.

Balance training in stroke patients has been supported. Methods commonly used and proven to be effective for this population include sitting or standing balance practice with various progressions including reaching, variations in base of support, use of tilt boards, gait training varying speed, and stair climbing exercises. Another method to improve balance is perturbation training, which is an external force applied to a person's center of mass in an attempt to move it from the base of support. The type of training should be determined by a physiotherapist and will depend on the nature and severity of the stroke, stage of recovery, and the patient's abilities and impairments after the stroke.

Balance training.

Populations such as the elderly, children with neuromuscular diseases, and those with motor deficits such as chronic ankle instability have all been studied and balance training has been shown to result in improvements in postural sway and improved "one-legged stance balance" in these groups. The effects of balance training can be measured by more varied means, but typical quantitative outcomes are centre of pressure (CoP), postural sway, and static/dynamic balance, which are measured

by the subject's ability to maintain a set body position while undergoing some type of instability.

Some types of exercise (gait, balance, co-ordination and functional tasks; strengthening exercise; 3D exercise and multiple exercise types) improve clinical balance outcomes in older people, and are seemingly safe. There is still insufficient evidence supporting general physical activity, computerized balance programs or vibration plates.

Functional Balance Assessments

Functional tests of balance focus on maintenance of both static and dynamic balance, whether it involves a type of perturbation/change of CoM or during quiet stance. Standardized tests of balance are available to allow allied health care professionals to assess an individual's postural control. Some functional balance tests that are available are:

- Romberg Test: Used to determine proprioceptive contributions to upright balance. Subject remains in quiet standing while eyes are open. If this test is not difficult enough, there is a Sharpened Romberg's test. Subjects would have to have their arms crossed, feet together and eyes closed. This decreases the base of support, raises the subject's center of mass, and prevents them from using their arms to help balance.
- Functional Reach Test: Measures the maximal distance one can reach forward beyond arm's length while maintaining feet planted in a standing position.
- Berg Balance Scale: Measures static and dynamic balance abilities using functional tasks commonly performed in everyday life. One study reports that the Berg Balance Scale is the most commonly used assessment tool throughout stroke rehabilitation, and found it to be a sound measure of balance impairment in patients following a stroke.
- Performance-Oriented Mobility Assessment (POMA): Measures both static and dynamic balance using tasks testing balance and gait.
- Timed Up and Go Test: Measures dynamic balance and mobility.
- Balance Efficacy Scale: Self-report measure that examines an individual's confidence while performing daily tasks with or without assistance.
- Star Excursion Test: A dynamic balance test that measures single stance maximal reach in multiple directions.
- Balance Evaluation Systems Test (BESTest): Tests for 6 unique balance control methods to create a specialized rehabilitation protocol by identifying specific balance deficits.

- The Mini-Balance Evaluation Systems Test (Mini-BESTest): Is a short form of the Balance Evaluation System Test that is used widely in both clinical practice and research. The test is used to assess balance impairments and includes 14 items of dynamic balance task, divided in to four subcomponents: anticipatory postural adjustments, reactive postural control, sensory orientation and dynamic gait. Mini-BESTest has been tested for mainly neurological diseases, but also other diseases.
- BESS: The BESS (Balance Error Scoring System) is a commonly used way to assess balance. It is known as a simple and affordable way to get an accurate assessment of balance, although the validity of the BESS protocol has been questioned. The BESS is often used in sports settings to assess the effects of mild to moderate head injury on one's postural stability. The BESS tests three separate stances (double leg, single leg, tandem) on two different surfaces (firm surface and medium density foam) for a total of six tests. Each test is 20 seconds long, with the entire time of the assessment approximately 5–7 minutes. The first stance is the double leg stance. The participant is instructed to stand on a firm surface with feet side by side with hands on hips and eyes closed. The second stance is the single leg stance. In this stance the participant is instructed to stand on their non-dominant foot on a firm surface with hands on hips and eyes closed. The third stance is the tandem stance. The participant stands heel to toe on a firm surface with hands on hips and eyes closed. The fourth, fifth, and sixth stances repeat in order stances one, two, and three except the participant performs these stances on a medium density foam surface. The BESS is scored by an examiner who looks for deviations from the proper stances. A deviation is noted when any of the following occurs in the participant during testing: opening the eyes, removing hands from the hips, stumbling forward or falling, lifting the forefoot or heel off the testing surface, abduction or flexion of the hip beyond 30 degrees, or remaining out of the proper testing position for more than 5 seconds.

Concussion (or mild traumatic brain injury) have been associated with imbalance among sports participants and military personnel. Some of the standard balance tests may be too easy or time-consuming for application to these high-functioning groups, s. Expert recommendations have been gathered concerning balance assessments appropriate to military service-members.

Quantitative (Computerized) Assessments

Due to recent technological advances, a growing trend in balance assessments has become the monitoring of center of pressure (terrestrial locomotion) (CoP), the reaction vector of center of mass on the ground, path length for a specified duration. With quantitative assessments, minimal CoP path length is suggestive of good balance. Laboratory-grade force plates are considered the "gold-standard" of measuring

CoP. The NeuroCom Balance Manager (NeuroCom, Clackamas, OR, United States) is a commercially available dynamic posturography system that uses computerized software to track CoP during different tasks. These different assessments range from the sensory organization test looking at the different systems that contribute through sensory receptor input to the limits of stability test observing a participant's ankle range of motion, velocity, and reaction time. While the NeuroCom is considered the industry standard for balance assessments, it does come at a steep price (about $250,000).

Within the past 5 years research has headed toward inexpensive and portable devices capable of measuring CoP accurately. Recently, Nintendo's Wii balance board (Nintendo, Kyoto, Japan) has been validated against a force plate and found to be an accurate tool to measure CoP. This is very exciting as the price difference in technology ($25 vs $10,000) makes the Wii balance board a suitable alternative for clinicians to use quantitative balance assessments. Other inexpensive, custom-built force plates are being integrated into this new dynamic to create a growing field of research and clinical assessment that will benefit many populations.

Fatigue's effect on balance.

Fatigue's Effect on Balance

The complexity of balance allows for many confounding variables to affect a person's ability to stay upright. Fatigue (medical), causing central nervous system (CNS) dysfunction, can indirectly result in the inability to remain upright. This is seen repeatedly in clinical populations (e.g. Parkinson's disease, multiple sclerosis). Another major concern regarding fatigue's effect on balance is in the athletic population. Balance testing has become a standard measure to help diagnose concussions in athletes, but due to the fact that athletes can be extremely fatigued has made it hard for clinicians to accurately determine how long the athletes need to rest before fatigue is gone, and they can measure balance to determine if the athlete is concussed. This can have devastating effects when looking at college and professional games where the athlete is depended upon by a community. So far, researchers have only been able to estimate that athletes

need anywhere from 8–20 minutes of rest before testing balance That can be a huge difference depending on the circumstances.

Other Factors Influencing Balance

Age, gender, and height have all been shown to impact an individual's ability to balance and the assessment of that balance. Typically, older adults have more body sway with all testing conditions. Tests have shown that older adults demonstrate shorter functional reach and larger body sway path lengths. Height also influences body sway in that as height increases, functional reach typically decreases. However, this test is only a measure of anterior and posterior sway. This is done to create a repeatable and reliable clinical balance assessment tool. A 2011 Cochrane Review found that specific types of exercise (such as gait, balance, co-ordination and functional tasks; strengthening exercises; 3D exercises [e.g. Tai Chi] and combinations of these) can help improve balance in older adults. However, there was no or limited evidence on the effectiveness of general physical activities, such as walking and cycling, computer-based balance games and vibration plates.

Voluntary Control of Balance

While balance is mostly an automatic process, voluntary control is common. Active control usually takes place when a person is in a situation where balance is compromised. This can have the counter-intuitive effect of increasing postural sway during basic activities such as standing. One explanation for this effect is that conscious control results in over-correcting an instability and "may inadvertently disrupt relatively automatic control processes." While concentration on an external task "promotes the utilization of more automatic control processes."

Balance and Dual-tasking

Supra-postural tasks are those activities that rely on postural control while completing another behavioral goal, such as walking or creating a text message while standing upright. Research has demonstrated that postural stability operates to permit the achievement of other activities. In other words, standing in a stable upright position is not at all beneficial if one falls as soon as any task is attempted. In a healthy individual, it is believed that postural control acts to minimize the amount of effort required (not necessarily to minimize sway), while successfully accomplishing the supra-postural task. Research has shown that spontaneous reductions in postural sway occur in response to the addition of a secondary goal.

McNevin and Wulf found an increase in postural performance when directing an individual's attention externally compared to directing attention internally That is, focusing attention on the effects of one's movements rather than on the movement itself will boost

performance. This results from the use of more automatic and reflexive control processes. When one is focused on their movements (internal focus), they may inadvertently interfere with these automatic processes, decreasing their performance. Externally focusing attention improves postural stability, despite increasing postural sway at times. It is believed that utilizing automatic control processes by focusing attention externally enhances both performance and learning. Adopting an external focus of attention sub-sequently improves the performance of suprapostural tasks, while increasing postural stability.

Center of Pressure

The center of pressure is the point where the total sum of a pressure field acts on a body, causing a force to act through that point. The total force vector acting at the center of pressure is the value of the integrated vectorial pressure field. The resultant force and center of pressure location produce equivalent force and moment on the body as the original pressure field. Pressure fields occur in both static and dynamic fluid mechanics. Specification of the center of pressure, the reference point from which the center of pressure is referenced, and the associated force vector allows the moment generated about any point to be computed by a translation from the reference point to the desired new point. It is common for the center of pressure to be located on the body, but in fluid flows it is possible for the pressure field to exert a moment on the body of such magnitude that the center of pressure is located outside the body.

Hydrostatic Example (Dam)

Since the forces of water on a dam are hydrostatic forces, they vary linearly with depth. The total force on the dam is then the integral of the pressure multiplied by the width of the dam as a function of the depth. The center of pressure is located at the centroid of the triangular shaped pressure field $\frac{2}{3}$ from the top of the water line. The hydrostatic force and tipping moment on the dam about some point can be computed from the total force and center of pressure location relative to the point of interest.

Usage for Sailboats

Center of pressure is used in sailboat design to represent the position on a sail where the aerodynamic force is concentrated. The relationship of the aerodynamic center of pressure on the sails to the hydrodynamic center of pressure (referred to as the center of lateral resistance) on the hull determines the behavior of the boat in the wind. This behavior is known as the “helm” and is either a weather helm or lee helm. A slight amount of weather helm is thought by some sailors to be a desirable situation, both from the standpoint of the “feel” of the helm, and the tendency of the boat to head

slightly to windward in stronger gusts, to some extent self-feathering the sails. Other sailors disagree and prefer a neutral helm.

The fundamental cause of "helm", be it weather or lee, is the relationship of the center of pressure of the sail plan to the center of lateral resistance of the hull. If the center of pressure is astern of the center of lateral resistance, a weather helm, the tendency of the vessel is to want to turn into the wind. If the situation is reversed, with the center of pressure forward of the center of lateral resistance of the hull, a "lee" helm will result, which is generally considered undesirable, if not dangerous. Too much of either helm is not good, since it forces the helmsman to hold the rudder deflected to counter it, thus inducing extra drag beyond what a vessel with neutral or minimal helm would experience.

Aircraft Aerodynamics

A stable configuration is desirable not only in sailing, but in aircraft design as well. Aircraft design therefore borrowed the term center of pressure. And like a sail, a rigid non-symmetrical airfoil not only produces lift, but a moment. The center of pressure of an aircraft is the point where all of the aerodynamic pressure field may be represented by a single force vector with no moment. A similar idea is the aerodynamic center which is the point on an airfoil where the pitching moment produced by the aerodynamic forces is constant with angle of attack.

The aerodynamic center plays an important role in analysis of the longitudinal static stability of all flying machines. It is desirable that when the pitch angle and angle of attack of an aircraft are disturbed (by, for example wind shear/vertical gust) that the aircraft returns to its original trimmed pitch angle and angle of attack without a pilot or autopilot changing the control surface deflection. For an aircraft to return towards its trimmed attitude, without input from a pilot or autopilot, it must have positive longitudinal static stability.

Missile Aerodynamics

Missiles typically do not have a preferred plane or direction of maneuver and thus have symmetric airfoils. Since the center of pressure for symmetric airfoils is relatively constant for small angle of attack, missile engineers typically speak of the complete center of pressure of the entire vehicle for stability and control analysis. In missile analysis, the center of pressure is typically defined as the center of the additional pressure field due to a change in the angle of attack off of the trim angle of attack.

For unguided rockets the trim position is typically zero angle of attack and the center of pressure is defined to be the center of pressure of the resultant flow field on the entire vehicle resulting from a very small angle of attack (that is, the center of pressure in the limit as angle of attack goes to zero). For positive stability in missiles, the total vehicle

center of pressure defined as given above must be further from the nose of the vehicle than the center of gravity. In missiles at lower angles of attack, the contributions to the center of pressure are dominated by the nose, wings, and fins. The normalized normal force coefficient derivative with respect to the angle of attack of each component multiplied by the location of the center of pressure can be used to compute a centroid representing the total center of pressure. The center of pressure of the added flow field is behind the center of gravity and the additional force "points" in the direction of the added angle of attack; this produces a moment that pushes the vehicle back to the trim position.

In guided missiles where the fins can be moved to trim the vehicles in different angles of attack, the center of pressure is the center of pressure of the flow field at that angle of attack for the undeflected fin position. This is the center of pressure of any small change in the angle of attack (as defined above). Once again for positive static stability, this definition of center of pressure requires that the center of pressure be further from the nose than the center of gravity. This ensures that any increased forces resulting from increased angle of attack results in increased restoring moment to drive the missile back to the trimmed position. In missile analysis, positive static margin implies that the complete vehicle makes a restoring moment for any angle of attack from the trim position.

Movement of Center of Pressure for Aerodynamic Fields

The center of pressure on a symmetric airfoil typically lies close to 25% of the chord length behind the leading edge of the airfoil. (This is called the "quarter-chord point"). For a symmetric airfoil, as angle of attack and lift coefficient change, the center of pressure does not move. It remains around the quarter-chord point for angles of attack below the stalling angle of attack. The role of center of pressure in the control characterization of aircraft takes a different form than in missiles.

On a cambered airfoil the center of pressure does not occupy a fixed location. For a conventionally cambered airfoil, the center of pressure lies a little behind the quarter-chord point at maximum lift coefficient (large angle of attack), but as lift coefficient reduces (angle of attack reduces) the center of pressure moves toward the rear. When the lift coefficient is zero an airfoil is generating no lift but a conventionally cambered airfoil generates a nose-down pitching moment, so the location of the center of pressure is an infinite distance behind the airfoil.

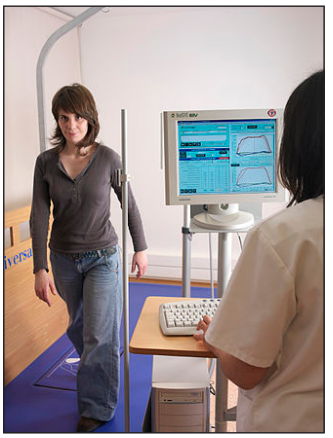
For a reflex-cambered airfoil, the center of pressure lies a little ahead of the quarter-chord point at maximum lift coefficient (large angle of attack), but as lift coefficient reduces (angle of attack reduces) the center of pressure moves forward. When the lift coefficient is zero an airfoil is generating no lift but a reflex-cambered airfoil generates a nose-up pitching moment, so the location of the center of pressure is an infinite distance ahead of the airfoil. This direction of movement of the center of pressure on a reflex-cambered airfoil has a stabilising effect.

The way the center of pressure moves as lift coefficient changes makes it difficult to use the center of pressure in the mathematical analysis of longitudinal static stability of an aircraft. For this reason, it is much simpler to use the aerodynamic center when carrying out a mathematical analysis. The aerodynamic center occupies a fixed location on an airfoil, typically close to the quarter-chord point.

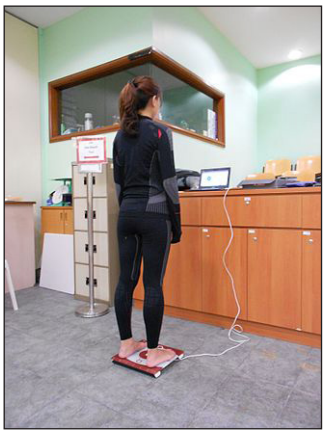
The aerodynamic center is the conceptual starting point for longitudinal stability. The horizontal stabilizer contributes extra stability and this allows the center of gravity to be a small distance aft of the aerodynamic center without the aircraft reaching neutral stability. The position of the center of gravity at which the aircraft has neutral stability is called the neutral point.

Force Platform

Force platforms or force plates are measuring instruments that measure the ground reaction forces generated by a body standing on or moving across them, to quantify balance, gait and other parameters of biomechanics. Most common areas of application are medicine and sports.



Woman walking on a force platform.



Stabilometric force platform in sport.

Operation

The simplest force platform is a plate with a single pedestal, instrumented as a load cell. Better designs have a pair of rectangular plates, although triangular can also work, one over another with load cells or triaxial force transducers between them at the corners. Like single-force platforms, dual-force platforms can be used to assess performance in double leg tests and strength and power asymmetries in unilateral jump and isometric tests. However, they also provide an additional level of intelligence on neuromuscular

status by evaluating the force distribution between limbs during double-limb tests, revealing critical information on strength asymmetries and compensatory strategies.

The simplest force plates measure only the vertical component of the force in the geometric center of the platform. More advanced models measure the three-dimensional components of the single equivalent force applied to the surface and its point of application, usually called the centre of pressure (CoP), as well as the vertical moment of force. Cylindrical force plates have also been constructed for studying arboreal locomotion, including brachiation.

Force platforms may be classified as single-pedestal or multi-pedestal and by the transducer (force and moment transducer) type: Strain gauge, piezoelectric sensors, capacitance gauge, piezoresistive, etc., each with its advantages and drawbacks. Single pedestal models, sometimes called load cells, are suitable for forces that are applied over a small area. For studies of movements, such as gait analysis, force platforms with at least three pedestals and usually four are used to permit forces that migrate across the plate. For example, during walking ground reaction forces start at the heel and finish near the big toe.

Force platforms should be distinguished from pressure measuring systems that, although they too quantify centre of pressure, do not directly measure the applied force vector. Pressure measuring plates are useful for quantifying the pressure patterns under a foot over time but cannot quantify horizontal or shear components of the applied forces.

The measurements from a force platform can be either studied in isolation, or combined with other data, such as limb kinematics to understand the principles of locomotion. If an organism makes a standing jump from a force plate, the data from the plate alone is sufficient to calculate acceleration, work, power output, jump angle, and jump distance using basic physics. Simultaneous video measurements of leg joint angles and force plate output can allow the determination of torque, work and power at each joint using a method called inverse dynamics.

Recent Developments in Technology

Advancements in technology have allowed force platforms to take on a new role within the kinetics field. Traditional laboratory-grade force plates cost (usually in the thousands) have made them very impractical for the everyday clinician. However, Nintendo introduced the Wii Balance Board (WBB) in 2007 and changed the structure of what a force plate can be. By 2010, it was found that the WBB is a valid and reliable instrument to measure the weight distribution, when directly compared to the “gold-standard” laboratory-grade force plate, while costing less than $100. More so, this has been verified in both healthy and clinical populations. This is possible due to the four force transducers found in the corners of the WBB. These studies are conducted using customized software, such as LabVIEW that can be integrated with the board to be able to measure

the amount of body sway or the CoP path length during trials for time. The other benefit to having a posturography system, such as the WBB, is that it is portable so clinicians around the world are able to measure body sway quantitatively, instead of relying on the subjective, clinical balance assessments currently in use.

According to Digital Trends, Nintendo's Wii and the WiiU successor product have both been discontinued as of March 2016. This exemplifies one of the issues arising from the adoption of inexpensive off-the-shelf consumer products re-purposed for medical measurements. Further issues with such adoption arise from the regulatory and standards bodies around the world. Force platforms used for measuring a patients balance and mobility performance are classified by the U.S. FDA (United States Food and Drug Administration) as Class I Medical Devices. As such they must be manufactured to certain quality standards as established by ISO (International Standards Organization)ISO 9001 Quality Management Principles or ISO 13485 Medical Device Quality Management Systems. The European Union's MDD (Medical Device Directive) also classifies force platforms used for medical measurements as Class I medical devices and require medical CE certification for importation and use in the European Union for such medical applications. A notable recent standard, ASTM F3109-16 Standard Test Method for Verification of Multi-Axis Force Measuring Platforms presents a framework for manufactures and users to verify the performance of Force platforms across the extents of their working surface. Standards such as these are used by manufactures of medical grade force platforms to ensure that measurements made on a patient population are accurate, repeatable and reliable. In short, inexpensive consumer grade entertainment components may be a poor choice for medical measurements given the lack of continuity of such products and their legal, regulatory and perhaps quality unsuitability for such applications.

Ground Reaction Force

In physics, and in particular in biomechanics, the ground reaction force (GRF) is the force exerted by the ground on a body in contact with it. For example, a person standing motionless on the ground exerts a contact force on it (equal to the person's weight) and at the same time an equal and opposite ground reaction force is exerted by the ground on the person.

In the above example, the ground reaction force coincides with the notion of a normal force. However, in a more general case, the GRF will also have a component parallel to the ground, for example when the person is walking – a motion that requires the exchange of horizontal (frictional) forces with the ground.

The use of the word reaction derives from Newton's third law, which essentially states that if a force, called *action*, acts upon a body, then an equal and opposite force, called *reaction*, must act upon another body. The force exerted by the ground is conventionally referred to as the reaction, although, since the distinction between action and

reaction is completely arbitrary, the expression *ground action* would be, in principle, equally acceptable.

The component of the GRF parallel to the surface is the frictional force. When slippage occurs the ratio of the magnitude of the frictional force to the normal force yields the coefficient of static friction.

GRF is often observed to evaluate force production in various groups within the community. One of these groups studied often are athletes to help evaluate a subject's ability to exert force and power. This can help create baseline parameters when creating strength and conditioning regimens from a rehabilitation and coaching standpoint. Plyometric jumps such as a drop-jump is an activity often used to build greater power and force which can lead to overall better ability on the playing field. When landing from a safe height in a bilateral comparisons on GRF in relation to landing with the dominant foot first followed by the non-dominant limb, literature has shown there were no significances in bilateral components with landing with the dominant foot first faster than the non-dominant foot on the GRF of the drop-jump or landing on vertical GRF output.

Hill's Muscle Model

In biomechanics, Hill's muscle model refers to either Hill's equations for tetanized muscle contraction or to the 3-element model. They were derived by the famous physiologist Archibald Vivian Hill.

Equation to Tetanized Muscle

This is a popular state equation applicable to skeletal muscle that has been stimulated to show Tetanic contraction. It relates tension to velocity with regard to the internal thermodynamics. The equation is:

$$(v+b)(F+a)=b(F_o+a),$$

where,

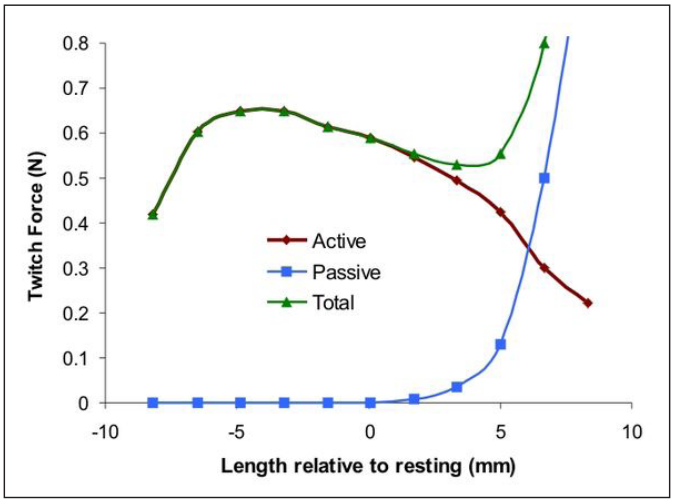
- F is the tension (or load) in the muscle.
- v is the velocity of contraction.
- F_o is the maximum isometric tension (or load) generated in the muscle.
- a coefficient of shortening heat.
- $b = a \cdot v_o / F_o$.
- v_o is the maximum velocity, when $F = o$.

Although Hill's equation looks very much like the van der Waals equation, the former has units of energy dissipation, while the latter has units of energy. Hill's equation demonstrates that the relationship between F and v is hyperbolic. Therefore, the higher the load applied to the muscle, the lower the contraction velocity. Similarly, the higher the contraction velocity, the lower the tension in the muscle. This hyperbolic form has been found to fit the empirical constant only during isotonic contractions near resting length.

The muscle tension decreases as the shortening velocity increases. This feature has been attributed to two main causes. The major appears to be the loss in tension as the cross bridges in the contractile element and then reform in a shortened condition. The second cause appears to be the fluid viscosity in both the contractile element and the connective tissue. Whichever the cause of loss of tension, it is a viscous friction and can therefore be modeled as a fluid damper.

Three-element Model

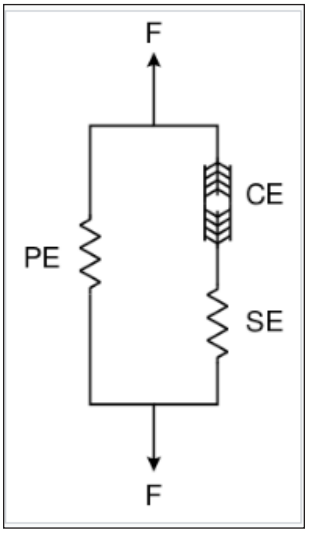
The three-element Hill muscle model is a representation of the muscle mechanical response. The model is constituted by a contractile element (CE) and two non-linear spring elements, one in series (SE) and another in parallel (PE). The active force of the contractile element comes from the force generated by the actin and myosin cross-bridges at the sarcomere level. It is fully extensible when inactive but capable of shortening when activated. The connective tissues (fascia, epimysium, perimysium and endomysium) that surround the contractile element influences the muscle's force-length curve. The parallel element represents the passive force of these connective tissues and has a soft tissue mechanical behavior. The parallel element is responsible for the muscle passive behavior when it is stretched, even when the contractile element is not activated. The series element represents the tendon and the intrinsic elasticity of the myofilaments. It also has a soft tissue response and provides energy storing mechanism.



Muscle length vs Force. In Hill's muscle model the active and passive forces are respectively F_{CE} and F_{PE}.

The net force-length characteristics of a muscle is a combination of the force-length characteristics of both active and passive elements. The forces in the contractile element, in the series element and in the parallel element, F_{CE}, F_{SE} and F_{PE}, respectively, satisfy:

$$F = F_{PE} + F_{SE} \quad \text{and} \quad F_{CE} = F_{SE} .$$



Hill's elastic muscle model. F: Force; CE: Contractile Element; SE: Series Element; PE: Parallel Element.

On the other hand, the muscle length L and the lengths L_{CE}, L_{SE} and L_{PE}, of those elements satisfy:

$$L = L_{PE} \quad \text{and} \quad L = L_{CE} + L_{SE} .$$

During isometric contractions the series elastic component is under tension and therefore is stretched a finite amount. Because the overall length of the muscle is kept constant, the stretching of the series element can only occur if there is an equal shortening of the contractile element itself.

Viscoelasticity

Muscles present viscoelasticity, therefore a viscous damper may be included in the model, when the dynamics of the second-order critically damped twitch is regarded. One common model for muscular viscosity is an exponential form damper, where:

$$F_D = k(\dot{L}_D)^a ,$$

is added to the model's global equation, whose k and a are constants.

Lines of Non-extension

In the field of biomechanics, the lines of non-extension are notional lines running across the human body along which body movement causes neither stretching or contraction. Discovered by Arthur Iberall in work beginning in the 1940s, as part of research into space suit design, they have been further developed by Dava Newman in the development of the Space Activity Suit.

They were originally mapped by Iberall by drawing a series of circles over a portion of the body and then watching their deformations as the wearer walked around or performed various tasks. The circles deform into ellipses as the skin stretches over the moving musculature, and these deformations were recorded. After a huge number of such measurements the data is then examined to find all of the possible deformations of the circles, and more importantly, the non-moving points on them where the original circle and the deformed ellipse intersect (at four points per circle). By mapping these points over the entire body, a series of lines are produced. These lines may then be used to direct the placement of tension elements in a spacesuit to enable constant suit pressure regardless of the motion of the body.

Plant Biomechanics

The mechanical design of a plant is fundamental to its growth and reproductive performance and so to its fitness. Leaves must be positioned to efficiently capture sunlight, commonly requiring support by stems and branches, and roots must be able to push through soil to acquire water and mineral nutrients and to ensure anchorage. The transport system must be able to sustain pressures in conducting fluids, often to tens of meters above ground. Stems and branches need to withstand static loads, such as the mass of the crown, and dynamic loading as a result of wind storms and arboreal animals. Plants must be able to display reproductive organs to biotic pollinators and dispersers (and withstand their mass and movement) or ensure that pollen and diaspores reach effective air currents. In addition, plant organs may require protection from damage by herbivores. However, structural reinforcement can be costly, achieved at the expense of fundamental functions such as carbon gain and reproduction. Furthermore, fitness is more than just survival in a particular abiotic environment; plant design must be considered in the plant's competitive context. Hence plant design must be "smart" design or growth and competitiveness will be severely compromised. Consequently, mechanical design from the scale of cellular anatomy to whole-plant architecture is fundamental to survival and likely to be generally under strong selective pressure.

Plant design at any level is likely to be constrained by conflicting influences on different functions, the effects of which will depend on the local environment. Furthermore,

mechanical properties can co-vary, with trade-offs among some properties. Therefore, there may be compromises. In addition, as the number of functions of an organism or organ increases, the number of equally efficient designs may increase, but potentially with a decline in efficiency of performance of any particular function. Hence, even though there may be mechanical trends along environmental gradients, a variety of mechanical solutions may occur within any environment. Some designs may be far from optimal due to constraints imposed by phylogenetic history. Furthermore, if plant parts are exposed to different environments during development there may be a heterogeneous profile of mechanics among and within plants of the same species. Plant tissues are more complex than most engineering structures, making their mechanical properties often difficult to measure, interpret, and predict. All these factors contribute to the complexity of understanding the functional and ecological significance of biomechanical design.

An increasing amount is known about the importance of mechanical properties of individual organs of herbaceous and woody plants, i.e., stems, branches, roots, leaves, and reproductive parts, and how these properties relate to or are affected by the local environment. However, our understanding of how these properties interact with each other over time and space is much less evident, especially in terms of plant fitness and distribution. Some plant responses to external mechanical stresses have positive secondary effects on, for example, susceptibility to insect attack, hardiness to drought and temperature stress, and the ability to capture sunlight. Nevertheless, plant investment in withstanding mechanical perturbations can have detrimental effects on plant size and fecundity; thus, fitness may be reduced unless the increased chance of survival compensates for reduced fecundity. Even less is known about downstream implications of plant mechanics on ecosystem structure and function, for example, on habitat quality, species diversity, trophic interactions, and nutrient cycling.

How Environmental Conditions Influence Plant Mechanical Design

Fundamental design traits, i.e. structure and architecture, at both the cellular and whole plant level, are directly influenced by the immediate environment in terms of resource supply and biotic/abiotic stress. Plants can respond to stress over evolutionary time (adaptation), but can also respond to temporal and spatial fluctuations in external stresses through adjustment in their shape and structure (phenotypic plasticity), allowing them to counterbalance the effects of the stress over their life span. Self-supporting terrestrial plants can be exposed to a range of changing stresses, such as wind, avalanches, and other soil mass movements. Vertical uprooting forces or trampling from grazing animals commonly have to be resisted by herbaceous species. In addition, plants may be exposed to changes in resource availability (both temporal and spatial), which may lead to mechanical responses. Significant examples of how plant mechanical properties change depending on the local environment can be found in self-supporting and lianoid forms of the same species, e.g., western poison oak. In such species, seedlings are often self-supporting until certain environmental cues are

received, e.g., light availability or the proximity of a host support. The self-supporting structure can then change to a lianoid form, hence altering significantly the ecological function of that individual over a period of time (a subject extensively reviewed by Rowe and Speck.

Abiotic stress: Stem and root mechanical design—in woody plants, wind loading may be the most commonly experienced mechanical abiotic stress. Many species also need to grow upwards to the canopy, so investment in mechanical support, anchorage, and light capture needs to be finely balanced. Trees in particular may reach extreme heights, e.g., Sequoiadendron giganteum Lindl. And Eucalyptus regnans F.Muell, and thus need rigid stems with a high biomass investment to prevent trunk failure by buckling or bending. A tree may achieve such rigidity by, for example, increased stem cross-section or a wood (xylem) structure capable of resisting the imposed static and dynamic mechanical stresses experienced throughout its life. Wood structure in trees is highly diverse both spatially (geographically and between populations and species) and temporally (throughout the life of an individual). This diversity reflects evolutionary history and particular ecological conditions, as well as the various tradeoffs among mechanical support, transport, and economy of construction. Wood density and conductivity are basic traits that determine growth and mechanical support and also influence photosynthesis, so that variation in density may be more related to mechanical support of the water column, which is under negative pressure, than to support of the tree. In addition, wood density in tropical savannah trees and Mediterranean broadleaved trees depends significantly on soil moisture content. Air temperature also affects wood density, by influencing the viscosity of sap as it ascends the stem, which in turn results in changes in cell lumen size. Therefore, trends in wood density must be strongly related to the geographic location and available water supply of a species, as well as the mechanical constraints imposed on individuals.

Roots are also particularly sensitive to fluctuations in water, nutrients, and mechanical stress. When external dynamic loading on the shoot system occurs due to, for example, wind, the adaptive growth responses in the root system are often more striking than those of the shoots. The vigor of a root system is fundamental to survival of a self-supporting plant, not only with regard to resource uptake, but also to anchorage strength. In many situations, even if the shoot is grazed by animals or damaged through mechanical loading, a well-anchored root system will ensure survival. Root systems of mature trees need to be large and deep enough to withstand overturning due to mechanical loading on the stem. The structural proximal roots tend to anchor the tree to the soil, whereas the finer distal roots serve a more absorptive function. Tree root systems with a higher degree of branching and vertical growth are usually better anchored than more shallow root systems, although soil type has a strong influence on the mode of failure. In young trees and woody shrubs, overturning is less problematic. Saplings and shrubs often display more fibrous root systems, with less secondary thickening and with a relatively large overall surface area. Increased root area and branching not only improve

nutrient absorption, but augment soil–root friction and resistance of vertical uprooting. Herbs and young woody plants are particularly vulnerable to vertical uprooting by grazing animals, and a well-anchored root system may improve plant survival.

In plants subjected to recurring mechanical stresses such as wind loading or other mechanical perturbation, several responses are observed. Leaf number and area may diminish, which reduces windinduced drag on the crown, but also reduces photosynthesis. Plants develop a stunted appearance, as stem elongation is reduced but radial growth is increased; hence, trees growing on windy sites tend to have a highly tapered shape. There is increased allocation of nitrogen and carbon to structural components, usually in the areas where mechanical stresses are highest, i.e., the base of the trunk and branches and along certain lateral roots and tap roots. Windward lateral roots held in tension during loading have been shown to develop more branches per unit area of soil, thus increasing surface area and root–soil friction. Plant internal structure also undergoes chemical and morphological changes, with the formation of thick-walled tracheid cells in trees.

This flexure wood is more resistant than normal wood in compression, but due to the structure of cellulose in the cell walls, is also more flexible, with a lower modulus of elasticity (E), hence allowing the stem and branches to flex more easily during wind sway. Combined, these responses to mechanical loading result in a reduced bending, lodging, or overturning moment, when the stem or crown is subjected to an external force. However, plant responses may differ depending on the abiotic stress applied and the species. Smith and Ennos pointed out that species plasticity will reflect the evolutionary history and current ecological conditions specific to each species. These authors showed that plants exposed to wind loading and stem flexure react in different and opposite ways. In experiments on Helianthus annuus L., wind-loaded plants were taller with a higher stem hydraulic conductivity whilst flexed plants were shorter with a lower stem hydraulic conductivity. Hence, the plant reaction to wind was the sum of two separate responses: a physiological response to air flow around the plant and the response of the stem to mechanical bending. How plants react to wind loading will therefore depend on organ size and mechanical properties, e.g., leaf stiffness or stem flexibility, and may explain why the responses of different species to wind loading are so variable.

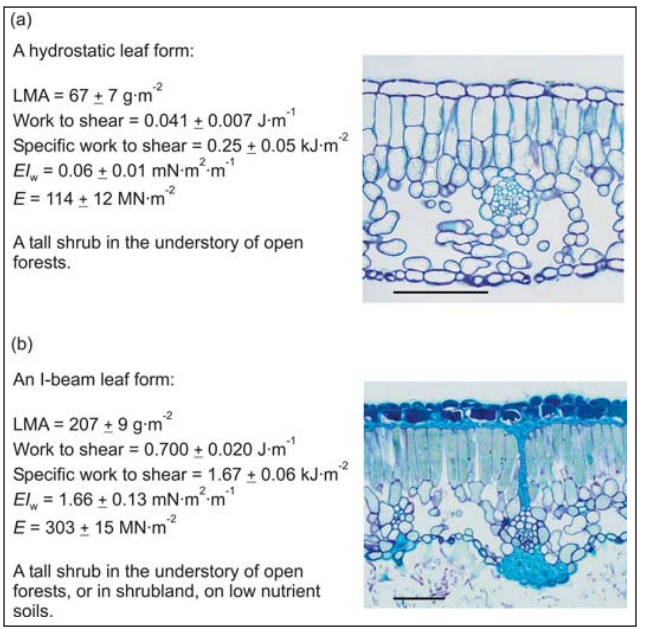
Certain trade-offs must therefore exist between different plant functions, including survival and reproduction. The mechanical stability of self-supporting plants is clearly important for survival, but can plant mechanical properties shift the emphasis from one factor to another? Examples of shifts from reproduction to stress resistance have been recorded in several species. The fitness of both Capsella bursa-pastoris L. and Brassica napus L. decreased significantly when subjected to mechanical perturbation. Delays in anthesis and reductions in flower number or size occurred, thus this trade-off is costly to fitness, unless benefits gained through the development of a mechanically hardened phenotype counterbalance costs to flower production. Similarly, Pigliucci showed that wind loading significantly increased branching in certain populations of Arabidopsis thaliana L. from different geographic locations, which in turn negatively

affected fecundity. Plants subjected to 16 h of daily wind loading flowered later and with more basal fruiting branches than those exposed to 0 or 6 h of daily wind. A high degree of genetic differentiation existed among the different populations, suggesting that many genes exist that underlie quantitative traits, thus offering greater phenotypic flexibility.

Although wind loading may decrease fecundity, the response of plants to this ubiquitous physical stress has been shown to improve other traits important to survival. Berthier and Stokes demonstrated that phototropic movements were significantly augmented in seedlings of Pinus pinaster Ait. Subjected to daily intermittent wind loading, thus improving light capture in exposed plants and hence their competitive strategy. Similarly, in inclined saplings of the same species, Berthier and Stokes found that leaning plants exposed to daily wind loading straightened significantly faster than leaning control plants not exposed to wind. Hence, wind stress may actually confer an ecological advantage to young trees in terms of competitive strategy and survival by stimulating tropic processes. A comparable example can also be found within the forest understory. Tree seedlings and herbs in forests are often damaged by falling debris such as branches and leaves. When a gap appears, e.g., after tree fall, the high levels of light stimulate compensatory growth in damaged understory plants. The mechanisms underlying these physiological responses are not yet fully understood, but Berthier and Stokes suggest that crosstalk exists between different sensory pathways in plant cells and that by stimulating mechanosensors through mechanical loading, several other reactions may be induced, including tropic movements.

Support and protection of leaves: Leaves generally need to support their own mass to align at an optimal angle to the sun, usually more or less horizontally but depending partly on the particular light environment. Horizontal alignment maximizes bending forces, so the optimum design for light interception may maximize the risk of bending failure. In addition, leaves can be repeatedly exposed to dynamic bending and twisting loads from wind, snow, hail, falling debris, and animals, and abrasion with other foliage, often with considerable force, yet are typically relatively flimsy structures compared with stems and roots. Hence, leaf design must be a compromise between carbon gain (including issues of thermoregulation and water loss) and the costs of sustaining static and dynamic loads. The leaf structure of many vascular plants is such that forces perpendicular to the surface are spread over a larger area, with small lateral deflections absorbed by the stringers (veins and the cells that connect them to both surfaces) that hold the leaf surfaces together, giving them the property of stressskin panels. Such structures are relatively flexible in bending and torsion and can deform elastically to some extent without damage. Hence, even relatively soft leaves can survive a moderate degree of mechanical stress. Even so, there is considerable variation in the structure and mechanics of leaves that is assumed to have adaptive significance in support and protection, in conjunction with other functions. However, the mechanical design of leaf laminae is poorly understood compared with that of petioles and stems.

Some leaves use predominantly hydrostatic support around a skeleton of vascular tissue, an effective form of support if leaf water status can be maintained. The mechanical properties and arrangements of vascular tissue reflect demand for both supply and support. In large and thin "hydrostatic" leaves, the cost of veins supporting the periphery may be substantial since the vein mass must increase at a higher power than the length, so above a critical leaf size it may be more efficient to produce multiple leaves than a single leaf of similar area. However, leaf size and shape (via effects on boundary layers) also influence rates of photosynthesis and water loss and costs of branch support, so in some environments it may still be more efficient to invest in large leaves, even if the cost of self-support is relatively high. Plants may have specialized structures that stiffen the leaf, such as a sub-epidermal layer of sclerenchyma, or enhanced I-beams formed by bundle sheath extensions of sclerenchyma meeting stiffened sub-epidermal layers, shifting reliance from hydrostatic toward structural support. Because these leaves are less dependent on water for support they may be more common where water is seasonally limiting. The most effective position for stiffening tissues such as collenchyma and sclerenchyma is close to the outer surfaces (distant from the neutral axis of bending). Since this design may compromise carbon gain by both diluting photosynthetic tissue and attenuating light reaching tissues below, it may be common only in species adapted to sunny environments. Silica can provide stiffening at potentially low cost, although its density may impose other costs. Some devices with major support and strengthening functions, such as silica and sclerenchyma bundle sheath extensions, may have other significant roles, such as in water transport to the parenchyma and epidermis for the latter and protection against herbivores.



Contrasting leaf anatomy, mechanics and ecology of two species from woodland in south-eastern Australia. (a) Prostanthera lasianthos Labill. (Lamiaceae); (b) Banksia

marginata Cav. (Proteaceae). The light micrographs are cross-sections across a secondary vein (sections provided by P. Peeters). Bar ¼ 100 lm. Mechanical properties were measured following Read et al.

The gross design of the lamina also influences support. Leaf support is best explained by the structural property EI (flexural stiffness or resistance to bending), the product of E and the shape of the transverse section of the structure (I, the second moment of area). I can increase by a change in size (e.g., leaf thickness), but can also vary without changing size (and therefore investment) since it is affected by shape relative to the direction of the force. For example, a long, narrow leaf blade will bend more than an ovate blade of the same area. Hence monocot leaves, with their typical linear shape that maximizes leaf area while minimizing self-shading and stem costs, have a variety of design features that enhance support. These can include a high proportion of longitudinal main veins, deposition of silica, revolute margins, longitudinal folding and curling, and terete blades, all of which assist in stiffening the blade. The longitudinal V-fold seen in some monocot blades allows stiffness for support plus twisting, which reduces drag. Similarly, petioles must be able to support leaf blades, and the maximum size of a leaf may be constrained by the load on the petiole. Both the bulk tissue stiffness and the dimensions of the petiole, including length and transverse geometry, influence leaf support.

Some capacity of the petiole and lamina to bend and twist allows reorientation in response to environmental change (e.g., changed position of light source) or rapid repositioning in response to strong loads such as wind to reduce drag and damage, the latter being particularly important for large leaves. When exposed to strong loads leaves must either resist bending and torsion or deform without fracture to escape damage. Leaf shape is not necessarily constant; local forces exerted by wind on a leaf depend partly on leaf shape, but shape itself is influenced by wind (Vogel, 2003). Thin leaves, leaflets, and leaf clusters readily fold and reconfigure in strong wind, reducing drag and consequent damage. Leaf movement may have the extra advantage of increasing air turbulence close to the leaf surface, thinning the boundary layer and increasing the capacity for heat and gas exchange and photosynthesis, although leaf flutter under very low wind speeds does not seem to be sufficient to generate this effect. Strong winds may have adverse effects, reducing carbon gain in a thin leaf by cooling and drying the leaves or curling the leaves and reducing their effective surface area. The petiole also contributes substantially to leaf bending and twisting by properties of its material (tissue types and arrangements) and shape. In particular, its cross-sectional shape can have a marked effect on the ratio of flexural to torsional rigidity.

A groove or flattening of the upper petiole surface (a departure from a circular cross-section) can increase the ratio of flexural to torsional rigidity, increasing the capacity to twist into the wind, thereby reducing drag. Tissues also contrast in their mechanical properties. Parenchyma and collenchyma act hydrostatically; under water stress a petiole can lose stiffness so the leaf wilts, reducing drag as well as exposure to high

irradiance and further water loss. In contrast, sclerenchyma is a non-hydrostatic tissue with high stiffness in bending and twisting. There may be a trend towards flexible petioles with collenchyma in the hypodermis in trees exposed to strong winds but petioles with subepidermal sclerenchyma in understory species. Plasticity in lamina and petiole form occurs both between and within plants in response to contrasting exposure to wind and light. For example, Woodward showed that some upland species of Phleum have greater sclerification than lowland species, but that sclerification could increase in lowland species in response to wind speed; Niklas showed that Acer saccharum L. trees on windy sites produce fewer and smaller leaves on shorter but more flexible petioles than on sheltered sites, reducing drag and damage. Wind can also lead to abrasive damage; abrasion from windborne particles or foliar contact during wind experiments resulted in increased epidermal conductance in needles of Picea sitchensis (Bong.) Carr. and Pinus sylvestris L.

The economics of support and supply is complex, not only at the level of the leaf but at the level of the whole plant, since it is influenced by efficiencies of leaf arrangement in both space and time. Furthermore, the product of branching plus leaf orientation may be critical in competition, not only because of its primary influence on carbon gain, but figure Contrasting leaf anatomy, mechanics and ecology of two species from woodland in south-eastern Australia. (a) Prostanthera lasianthos Labill. (Lamiaceae); (b) Banksia marginata Cav. (Proteaceae). The light micrographs are cross-sections across a secondary vein. Bar ¼ 100 lm. Mechanical properties were measured following Read et al. by shading competing neighbors below. For example, a forest dominated by trees that have vertically oriented leavescan support a richer understory of light-demanding plants than forests of trees with horizontally oriented leaves. There may be other downstream effects on community composition and dynamics. Leaves that use hydrostatic support should have less dilution of photosynthetic tissue by structural tissue, so may be preferred by herbivores for their higher concentration of nutrients that are easier to access. If so, and if some communities in particular environments (e.g., mesic) have a greater proportion of species with hydrostatic leaves, there may be a greater density and diversity of herbivores and their predators. Hence, the ecology of leaf support and protection should influence not just competitive success, but higher trophic levels.

Biotic stress: Leaf mechanics in protection from herbivores—Animals can cause significant damage to plants. Resistance of species to trampling has been associated with higher tensile strength of organs and stem flexibility. The tensile strength of some organs may increase in plants exposed to trampling. However, the most widespread source of biotic damage is herbivory. Here we focus on leaves, but mechanical properties are likely to contribute to defense of other organs. There is strong evidence that leaf mechanical properties, including strength (commonly measured as the force to fracture per unit area over which the force is applied) and toughness (usually regarded as the energy or work to fracture), deter herbivores. Studies of herbivory have used various mechanical measures, including work to fracture, force to fracture, and strength from

both tensile tests and compressive punching and shearing tests. Foliar strength and work to fracture are often, but not always, correlated but are rarely both measured, so it is difficult to know which is of most significance in ecological studies. For these reasons, and since terms are not used consistently among studies, we here use "tough" to refer to this range of properties. Variation in these properties among plants results from differences in the amount and composition (e.g., lignification) of cell wall, thickness of the cuticle, leaf thickness and density (influencing leaf mass per area, LMA), and amounts and arrangements of particular tissues, including specialized tissues such as sclerenchyma, each contributing differently to each mechanical property. Silica is well known in monocots, in external hooked structures that sharpen blade margins and in cell walls and cytoplasm. It has suggested roles in support and antiherbivore defense, both in spinescent structures that must be rigid to be effective deterrants and in leaves (including dicots), and may also protect against microorganisms.

There is diverse evidence for the defensive roles of mechanical properties. First, there is often a strong negative association between leaf toughness and herbivore damage across species, with low levels of herbivory in sclerophyllous vegetation. Second, tough leaves may be eaten less than softer leaves of the same species that have developed under different conditions. Third, mature toughened leaves or stems may be eaten less or more slowly than young expanding tissue of the same species, even when young tissue has higher levels of chemical defense. There are also many examples of specific effects of leaf mechanics on herbivores. For example, tough leaves can have components that wear insect mandibles, and substantial energy demands are required by leaf-cutting ants in cutting leaves. Ants may avoid mature leaves because of their toughness, even when better for growth of symbiotic fungi than young leaves. Toughness can limit feeding by early instars of insects and may increase in leaves produced after herbivore attack. Some caterpillars roll leaves around expanding buds to feed on the shaded expanding leaves that are less tough than unshaded leaves. Feeding time and rate of development can be longer on tougher leaves. In some cases herbivores avoid tough leaves or grow more slowly than on a diet of soft leaves, but feed successfully when the leaves are ground to a powder, implying a mechanical constraint to feeding or assimilation. Mechanical constraints can act in a complex way, such that delayed access to nutrients by herbivores can alter the ratio of assimilated nutrients, thereby constraining growth. Larger head mass of insects is reported in grass specialists than other foliage feeders of similar size, and feeding on tough grasses can induce development of larger heads in caterpillars. Mechanical defense might also have indirect effects on herbivore survival, if slowed feeding increases vulnerability to abiotic stress and natural enemies, and may have complex effects on herbivore behavior across a range of scales.

Protection from herbivory may be conferred directly by mechanical constraints to feeding or indirectly by dilution of nutrients by the structural features that confer the mechanical properties, and nutrients such as proteins may be less available due to binding with lignin, i.e., the properties that make the leaf tough are associated with those that

reduce digestibility, both of which may affect diet selection. In addition, leaf chemistry can vary along gradients in mechanical properties so that soft leaves and even parts of leaves are more nutritious, sometimes with lower levels of chemical defenses. Hence it can be difficult to discern the exact role of mechanical properties. Mechanical defenses can potentially be overcome by herbivores with appropriate musculature and mouthparts, but there may still be significant energy costs (direct or indirect) incurred during acquisition through to assimilation. Hence mechanical defenses may be best considered not only in absolute terms (e.g., whether a herbivore has the bite force to acquire and process the food), but in terms of the cost of acquisition per nutrition gained.

Relatively little is known about the specific mechanical properties that deter herbivores. We might expect that mechanical constraints to feeding relate to the way the herbivore uses its mouthparts, but there is still much to learn about the various ways in which teeth and invertebrate mouthparts work to acquire and mechanically process their plant diet. Leaf "structural" properties, but not "material" properties (per unit leaf thickness), correlated with feeding by moth larvae across 20 plant species, and shearing properties correlated with feeding more than tensile properties, the latter consistent with small animals having to cut through individual veins. Densities of chewing insects, but not sucking insects, correlated with leaf strength of forest understory species. Leaf structure can influence invertebrate herbivores at a fine scale since they are small enough to feed selectively across a leaf, often avoiding the toughest parts. A dense wall of hairs can be a significant mechanical barrier to feeding insects, and lignified bundle sheath extensions may provide physical barriers to movement of leaf miners. Some vertebrates even seem to discriminate among and within leaves in relation to mechanical gradients, although others may use toughness to select among feeding sites rather than among leaves. Hence plants must defend themselves across a range of scales, from insects influenced by cell traits to large mammals that ingest whole leaves or even branches.

Plants have a range of anti-herbivore defenses, including chemistry, spinescence, and attraction of natural enemies, but there is relatively little evidence for the manner in which mechanical properties are balanced against these other defenses. Limited data for phenolics vs. toughness across species suggests there is no simple association (J. Read, unpublished data). This is perhaps not surprising since phenolics, although useful for comparisons because of their widespread occurrence, have multiple effects, including photoprotection. However, there is some evidence of "trade-offs" between defenses, for example, in declining levels of chemical defenses as expanding leaves toughen, and between different forms of physical defense.

Mechanical properties can provide defense indirectly. For example, slippery wax barriers on stems of Macaranga myrmecophytes limit access by generalist ants that might compete with or prey on the resident "wax-running" ant partners that protect the host from herbivores. Macaranga species are often pioneers, and the deposition of slippery

waxes on the stems may incur direct costs, but not the additional indirect costs to photosynthesis associated with leaf toughening, which could be critical to fastgrowing pioneers.

If mechanical properties influence plant selection by herbivores, they may significantly influence plant growth, survival, and patterns of species' abundance, but prediction of the latter can be difficult because of the variety of ways in which plants resist herbivory and the complexity of scaling up to community-level properties. Leaf mechanical and underlying structural traits may also influence local densities and feeding guilds of herbivores and consequently the community abundance and composition of herbivores and potentially their predators. In addition, for herbivores limited by leaf mechanics (or in concert with nutrition and chemical defense) and restricted to feeding on immature leaves, the window of opportunity for feeding may be small, particularly in seasonal climates where leaf expansion is synchronous and restricted in duration or where environmental stress limits production of new leaves. If temporal variation in the mechanical profile of vegetation influences temporal patterns of herbivore abundance, higher trophic levels may be affected. Hence leaf mechanics should have a significant role in the organization and functioning (e.g., diversity, trophic relationships, and productivity) of aboveground ecosystems, a role that has been little addressed to date.

Biomechanics of Reproduction and Regeneration

Reproductive structures: Reproductive structures vary substantially in their mechanics, from often delicate flowers to robust woody fruits, but relatively little is known about their mechanical traits. Animal-pollinated flowers may need to be large if supplying the appropriate food reward for a bird or mammal or clustered to concentrate the reward in space. Larger inflorescences may require co-evolved stem properties for support and to minimize drag to reduce damage, particularly if the perianth is delicate. Flowers on long stems may be more apparent to pollinators or better able to deposit pollen in wind currents, but are more exposed to wind damage. For example, laterally facing daffodil flowers are supported on stems that are elliptical in cross-section, allowing twisting, thereby reducing drag. Ennos noted that inflorescences of the lowland sedge Carex acutiformis Ehrh. were borne on tall stems that were triangular in section with low torsional rigidity, allowing twisting in wind, reducing drag and possibly self-pollination; however, mountain species that must resist stronger winds had shorter stems that were more circular in section and better able to resist buckling. Some flowers or their parts (e.g., the labellum of orchids and the style supporting the pollen presenter in Proteaceae species) may need to be stiffened to support the mass and activity of pollinators, but few data are available on this subject.

Fruits have evolved for protection and dispersal of seeds, often with specialized mechanical design that enhances one or both functions. Possibly the greatest degree of protection is afforded against seed predators and fire. Species in regions where wildfires are common are often serotinous, with mature seeds stored in the canopy for an extended period in robust fruits that range from leathery to massive woody structures until

the advent of more suitable post-fire conditions for seedling recruitment. Protection of seeds from the heat of a fire is related to wall thickness and fruit size, but woody fruits may also confer protection against insect larvae and seed-eating birds. In Hakea (Proteaceae), protection from seed eaters may provide a better explanation than fire for the particularly dense and thick-walled follicles of some species, and reproductive organs are sometimes also protected by a "nest" of tough, spinescent leaves. In the same family, but from rainforest, Macadamia ternifolia F.Muell. has a follicle with a wall as hard as commercial-grade aluminum, but with twice the strength (stronger than concrete or brick) and half the density. In the Australian tropical rainforest a higher content of N has been recorded in seeds (adjusted for mass) that have marked protection, including by woody fruits. Seeds of some plants have hard seed coats, often associated with seed dormancy and potentially affording protection from both herbivores and pathogens, i.e., enabling a long life span, as well as contributing to dormancy. For animal-dispersed seeds, a hard seed coat or stony endocarp may provide mechanical protection as the seed is handled or passes through the disperser's gut. In a different vein, the changes in fruit texture accompanying ripening of fleshy fruits may be used as a selection cue by primates, in conjunction with visual and olfactory cues.

Reproductive parts are often motile to a greater or lesser degree, with some wind-dispersed pollen, fruits, and seeds mechanically designed for passive "flight." Added mass increases wing-loading (mass : area), thereby decreasing dispersal capacity; hence there is a potential trade-off between dispersal distance (by wind) and the nutritional reserves and protective seed coverings (including fruit tissue) that might confer an advantage during dispersal, any dormancy period, and establishment. Wind-dispersed seeds and fruits may have devices such as hairs, fibers, and wings that increase drag, or in larger diaspores lift-to-drag, thereby maximizing distance of dispersal. Winged fruits of some species glide downwards in a helical manner, in which a lift-generating airfoil, for example, reduces the descent rate. Dispersal of spores, pollen, and seeds is sometimes achieved by dynamic release, following cellular dehydration in reproductive structures such as fern sporangia, anthers at dehiscence, and fruits (and similarly, the mechanical design of vegetative parts allows sun-tracking by leaves, leaf-folding as in some grasses and legumes, and trap-closing in fly-traps, not discussed in this review), often involving specialized cells and patterns of wallthickening. Dynamic loading by wind can assist in spore and pollen release, and ovulate organs may be designed to capture pollen aerodynamically. Some plants have a highly specialized mechanical system of pollen transfer. These include buzz pollination (e.g., in Solanaceae), whereby a stamen is vibrated by rapid contractions of the flight muscles of pollen-collecting insects, releasing small quantities of pollen (avoiding over-exploitation) through poricidal anthers, and piston (Fabaceae) and lever systems (e.g., Salvia), in which the pollinator lands on, or pushes against, a staminal lever that moves to deposit pollen onto the visitor. Hence mechanical design has often been finely tuned in plants to achieve motility of diaspores and pollen.

Vegetative structures in regeneration: The mechanical properties of stem and root systems have in some species evolved to become either highly resistant to breaking or to

the contrary easily snapped or pulled, which allows for a greater dispersal of seeds and vegetative parts. For example, the stem base of the fast-growing herbaceous annual Galium aparine L. is highly extensible and thus able to withstand high breaking strains. Uprooting may therefore be avoided as the mericarps hook on to passing mammals. In many species, however, the opposite strategy may improve dispersal. Bobich and Nobel studied the morphology, mechanics, and rooting ability of four species of cholla (Opuntia spp.) cactus in the Mojave and Sonoran deserts. Only one species, O. bigelovii Engelm., showed evidence of vegetative reproduction in the field and was also the only species that had both relatively stiff, but weak terminal joints, as well as joints that demonstrated a high rooting ability. This species also had the highest frequency of spines per area of tubercule, thus increasing the chance that as a vertebrate brushed past the cactus, terminal segments would be broken off at the weak joint and also catch onto the animal's pelt, hence allowing for greater dispersal. In a study of the brittleness (a measure of ease of fracture) of twig bases of willows (Salix spp.), Beismann et al. found that S. fragilis L. and S. rubens Schrank were the most brittle, thus enabling these riparian species to propagate vegetatively downstream via broken twigs. However, in other species, e.g., S. triandra L. and S. viminalis L., the loss of twigs may serve more as a mechanical safety mechanism, as these riverside species are often submerged during flooding and the loss of twigs will reduce drag forces during inundation, thus preventing stem breakage. In the one subalpine species tested, S. appendiculata Vill., low stresses at fracture and yield were noted. Survival throughout the winter is a priority for this species, which is often buried under heavy snow or subjected to avalanches and so should be able to withstand high strains from either wind or flooding.

Root anchorage quality of saplings and vegetative propagules can also determine the success of reproduction in riparian species. In uprooting experiments, Salix elaeagnos Scop. saplings growing in disturbed sites close to the river channel were found to have a relatively higher resistance to uprooting than Populus nigra L. saplings that occupied more stabilized sites. However, cuttings of P. nigra demonstrated superior growth and anchorage ability under severe flood conditions, thus leading to greater success in vegetative reproduction. Differences in root anchorage were attributed to differences in morphology and tensile strength between saplings and cuttings. Root tensile strength increases with decreasing diameter and is highly related to cellulose content. Herbs with highly fibrous root systems will thus better resist vertical uprooting than species with fewer or thicker roots. Fibrous-rooted grasses are also more deeply rooted than the average herbaceous plant, thus improving uprooting resistance. Ennos and Fitter therefore proposed that such species would better resist uprooting by grazing animals. If the root system remains intact in the ground, regrowth of shoots will also be more vigorous after damage.

Aquatic macrophytes also have a large number of thin and branched roots. This increases surface area along with tensile strength and ensures a high root–soil friction. Aquatic plants are subject not only to pulling forces arising from waves, but also to

grazing forces from birds. Although stem breakage will lead to a loss of fitness, uprooting is more damaging. Schutten et al. demonstrated that aquatic species had either relatively weak stems or a substantial investment in anchorage. In plants that break, e.g. Myriophyllum spicatum L., regrowth of the shoot is possible. Certain aquatic plants may also have evolved a weaker stem section above the sediment surface where breakage occurs, thus sacrificing the shoot, but protecting the perennating root system. A large body of literature is available on the mechanics of wave-swept aquatic plants and algae, which is reviewed elsewhere.

The mode of failure that a tree can undergo during a wind storm can have similar consequences for regeneration. Many broadleaved species are able to resprout when branches are broken or the stem is broken after a storm. By investing in a well-anchored root system, many broadleaved species break in the trunk (when soil conditions are not limiting to root system development). Niklas suggested that broadleaved species invest a large proportion of their biomass in the main stems and establish a class of stems (branches and twigs) prone to mechanical failure during storms, thus providing a margin of safety against catastrophic wind damage. By ensuring good anchorage and a strong stem base, broadleaved trees can resprout after storm damage, thus prolonging their life and maintaining their position in the canopy. However, this hypothesis does not extend to most conifer species, which usually cannot resprout and often die after severe wounding. Therefore, conifer seedlings need to grow quickly to benefit from gaps caused by storm damage, and in a mixed forest these young conifers will be in competition with resprouting broadleaved species if windstorms cause gaps in the canopy. A long-term dynamic and spatial analysis of regeneration and juvenile tree growth in a forest damaged by a windstorm, along with mechanical properties of each species studied, would be useful in understanding the ecological implications and how to manage such forests.

References

- What-is-biomechanics-definition-applications, academy: study.com, Retrieved 10 August, 2020
- Simon P.R. Jenkins (2005) Sports Science Handbook The Essential Guide to Kinesiology, Sport and Exercise Science Volulme 1: A-H, Page 294 ISBN 0906522 36 6
- Biomechanics, neuroscience: sciencedirect.com, Retrieved 26 February, 2020
- Boone, Tommy. “Basic Concepts in Sports Biomechanics”. Archived from the original on 28 October 2011. Retrieved 27 October 2011
- Plant-biomechanics-in-ecological-context-51192388: researchgate.net, Retrieved 06 March, 2020
- Davidson; Madigan, Nussbaum (2004). “Effects of lumbar extensor fatigue and fatigue rate on postural sway”. European Journal of Applied Physiology. 93 (92): 183–189. doi:10.1007/s00421-004-1195-1. PMID 15549370

Permissions

Index