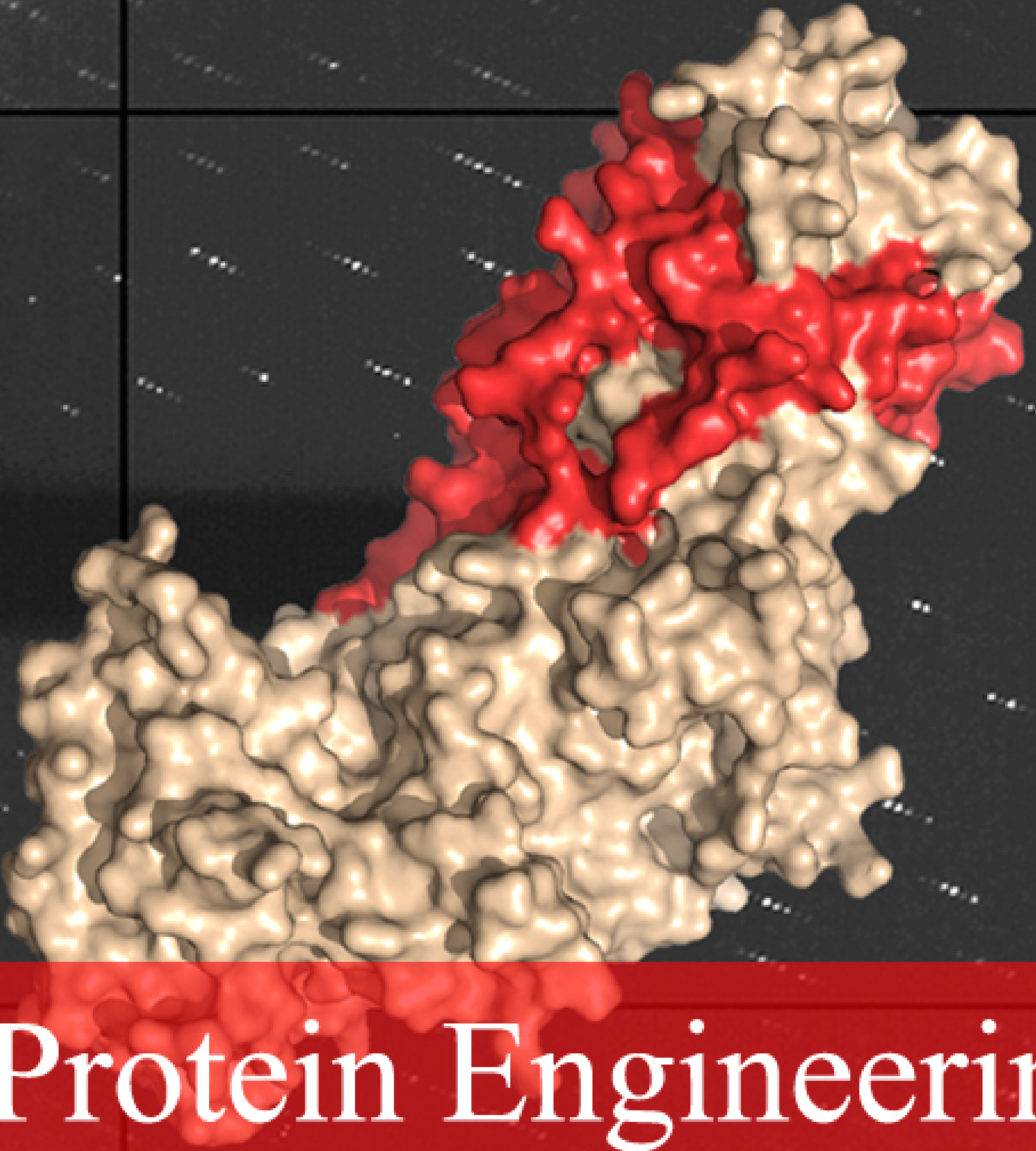


Protein Engineering

Sharlene Boswell

First Edition, 2012

ISBN 978-81-323-3015-8

© All rights reserved.

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Introduction

Protein engineering is the process of developing useful or valuable proteins. It is a young discipline, with much research taking place into the understanding of protein folding and recognition for protein design principles.

There are two general strategies for protein engineering, *rational design* and *directed evolution*. These techniques are not mutually exclusive; researchers will often apply both. In the future, more detailed knowledge of protein structure and function, as well as advancements in high-throughput technology, may greatly expand the capabilities of protein engineering. Eventually, even unnatural amino acids may be incorporated thanks to a new method that allows the inclusion of novel amino acids in the genetic code.

Rational design of proteins

In rational protein design, the scientist uses detailed knowledge of the structure and function of the protein to make desired changes. This generally has the advantage of being inexpensive and technically easy, since site-directed mutagenesis techniques are well-developed. However, its major drawback is that detailed structural knowledge of a protein is often unavailable, and even when it is available, it can be extremely difficult to predict the effects of various mutations.

Computational protein design algorithms seek to identify novel amino acid sequences that are low in energy when folded to the pre-specified target structure. While the sequence-conformation space that needs to be searched is large, the most challenging requirement for computational protein design is a fast, yet accurate, energy function that can distinguish optimal sequences from similar suboptimal ones.

Directed evolution

In directed evolution, random mutagenesis is applied to a protein, and a selection regime is used to pick out variants that have the desired qualities. Further rounds of mutation and selection are then applied. This method mimics natural evolution and generally produces superior results to rational design. An additional technique known as DNA shuffling mixes and matches pieces of successful variants in order to produce better results. This process mimics the recombination that occurs naturally during sexual reproduction. The great advantage of directed evolution is that it requires no prior structural knowledge of a protein, nor is it necessary to be able to predict what effect a given mutation will have. Indeed, the results of directed evolution experiments are often surprising in that desired changes are often caused by mutations that were not expected to have that effect. The drawback is that they require high-throughput, which is not feasible for all proteins. Large amounts of recombinant DNA must be mutated and the products screened for desired qualities. The sheer number of variants often requires expensive robotic equipment to automate the process. Furthermore, not all desired activities can be easily screened for.

Examples of engineered proteins

Using computational methods, a protein with a novel fold has been designed, known as Top7, as well as sensors for unnatural molecules. The engineering of fusion proteins has yielded rilonacept, a pharmaceutical which has secured FDA approval for the treatment of cryopyrin-associated periodic syndrome.

Another computational method, IPRO, successfully engineered the switching of cofactor specificity of Candida boidinii xylose reductase. Iterative Protein Redesign and Optimization (IPRO) redesigns proteins to increase or give specificity to native or novel substrates and cofactors. This is done by repeatedly randomly perturbing the backbones of the proteins around specified design positions, identifying the lowest energy combination of rotamers, and determining if the new design has a lower binding energy than previous ones. The iterative nature of this process allows IPRO to make additive mutations to the protein sequence that collectively improve the specificity towards the desired substrates and/or cofactors. Details on how to download the software implemented in Python and experimental testing of predictions are outlined in the following paper.

Chapter 1

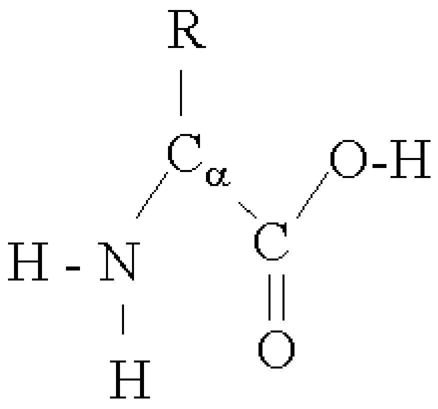
Protein Structure

Proteins are an important class of biological macromolecules present in all organisms. All proteins are polymers of amino acids. Classified by their physical size, proteins are nanoparticles (definition: 1–100 nm). Each protein polymer – also known as a polypeptide – consists of a sequence of 20 different L-α-amino acids, also referred to as residues. For chains under 40 residues the term peptide is frequently used instead of protein. To be able to perform their biological function, proteins fold into one or more specific spatial conformations, driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions, Van Der Waals forces, and hydrophobic packing. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure. This is the topic of the scientific field of structural biology, which employs techniques such as X-ray crystallography, NMR spectroscopy, and dual polarisation interferometry to determine the structure of proteins.

Protein structures range in size from tens to several thousand residues Very large aggregates can be formed from protein subunits: for example, many thousand actin molecules assemble into a microfilament.

A protein may undergo reversible structural changes in performing its biological function. The alternative structures of the same protein are referred to as different conformations, and transitions between them are called conformational changes.

Protein covalent structure and stereochemistry

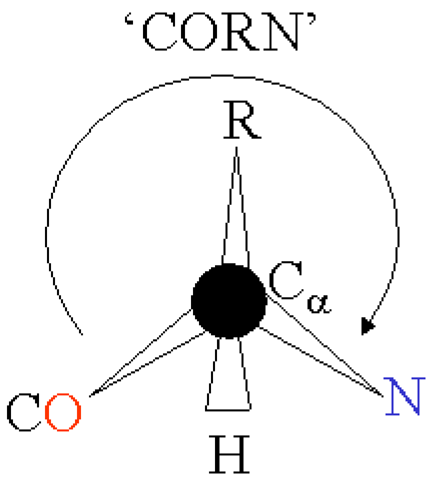


An α-amino acid. The C_αH atom is omitted in the diagram.

Protein amino acids are combined into a single polypeptide chain in a condensation reaction. This reaction is catalysed by the ribosome in a process known as translation.

Amino acid residues

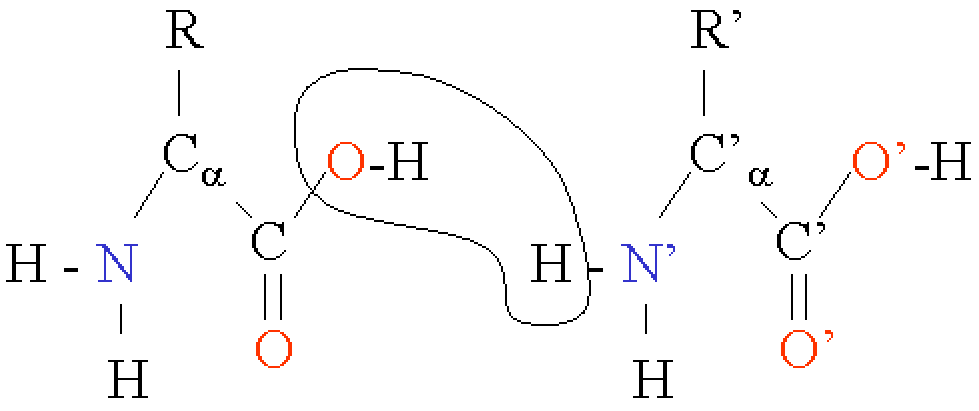
Each α-amino acid consists of a backbone part that is present in all the amino acid types, and a side chain that is unique to each type of residue. An exception from this rule is proline, where the hydrogen atom is replaced by a bond to the side chain. Because the carbon atom is bound to four different groups it is chiral, however only one of the isomers occur in biological proteins. Glycine however, is not chiral since its side chain is a hydrogen atom. A simple mnemonic for correct L-form is "CORN": when the C_α atom is viewed with the H in front, the residues read "CO-R-N" in a clockwise direction.



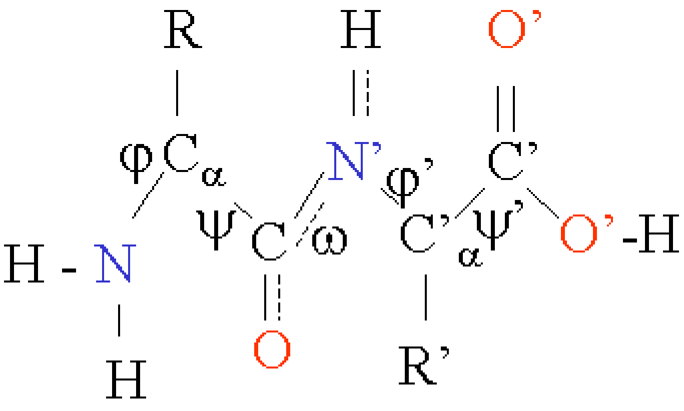
CO-R-N rule

The 20 naturally occurring amino acids have different physical and chemical properties, including their electrostatic charge, pKa, hydrophobicity, size and specific functional groups. These properties play a major role in molding protein structure.

The peptide bond



Two amino acids



Bond angles for ψ and ω

The peptide bond tend to be planar due to the delocalization of the electrons from the double bond. The rigid peptide dihedral angle, ω (the bond between C_1 and N) is always close to 180 degrees. The dihedral angles phi φ (the bond between N and Cα) and psi ψ (the bond between Cα and C_1) can have a certain range of possible values. These angles are the internal degrees of freedom of a protein, they control the protein's conformation. They are restrained by geometry to allowed ranges typical for particular secondary structure elements, and represented in a Ramachandran plot. A few important bond lengths are given in the table below.

Peptide bond	Average length	Single bond	Average length	Hydrogen bond	Average (±30)
Cα – C	153 pm	C - C	154 pm	O-H --- O-H	280 pm
C - N	133 pm	C - N	148 pm	N-H --- O=C	290 pm

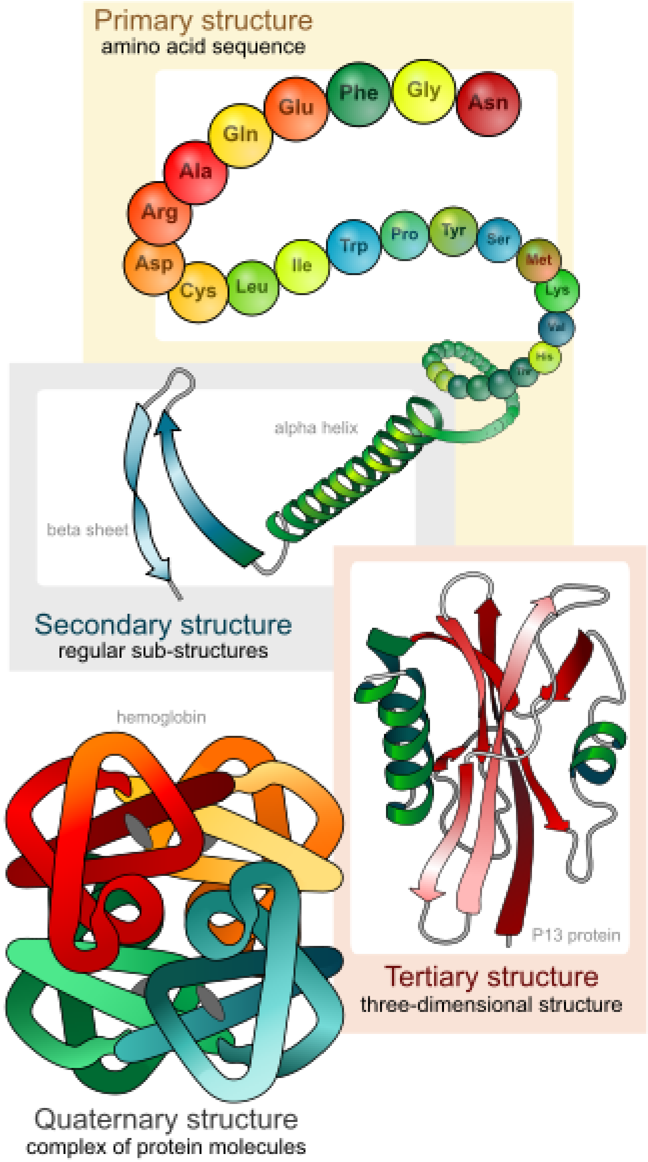
N - Cα	146 pm	C - O	143 pm	O-H --- O=C	280 pm

Side-chain conformation

The atoms along the side chain are named with Greek letters in Greek alphabetical order: α, β, γ, δ, ϵ, and so on. C_{α} refers to the carbon atom of the backbone closest to the carbonyl group of that amino acid, C_{β} the second closest and so on. The dihedral angles around the bonds between these atoms are named χ1, χ2, χ3, etc. The dihedral angle of the first movable atom of the side chain, γ, defined as N-Cα-Cβ-*X*γ, is named χ1. Side chains tend to adopt different staggered conformations called *gauche(-)*, *trans*, and *gauche(+)*, which corresponds to rotation angles of 60°, 180°, and -60°, respectively, around the sp3-sp3 bonds.

The diversity of side-chain conformations is often expressed in rotamer libraries. A rotamer library is a collection of rotamers for each residue type. Side-chain dihedral angles are not evenly distributed, but for most side chain types, the χ angles occur in tight clusters around certain values. Rotamer libraries therefore are usually derived from statistical analysis of side-chain conformations in known structures of proteins by clustering observed conformations or by dividing dihedral angle space into bins, and determining an average conformation in each bin.

Levels of protein structure



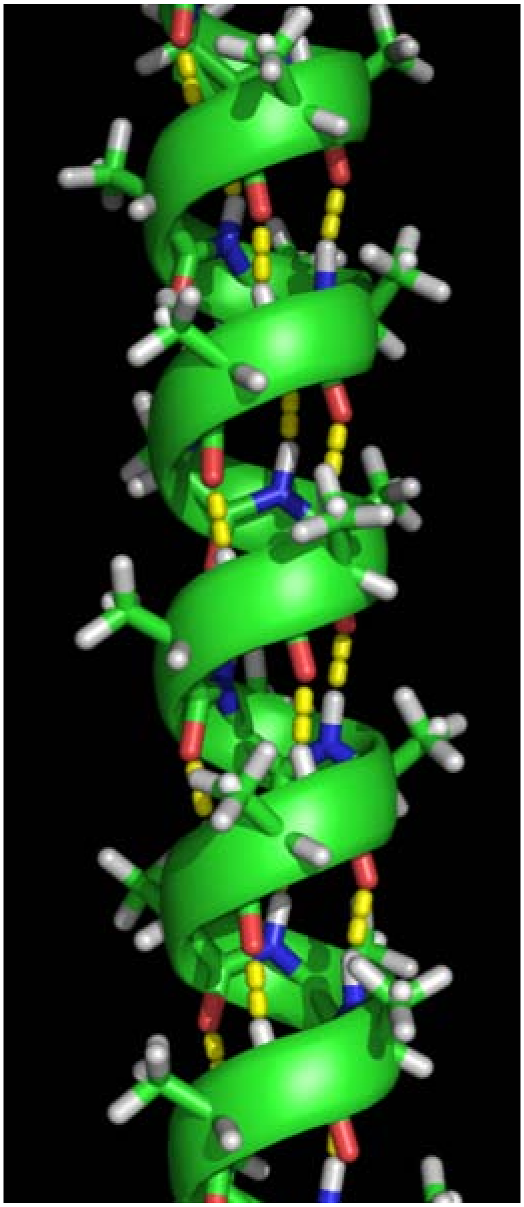
Protein structure, from primary to quaternary structure.

There are four distinct levels of protein structure.

Primary structure

The primary structure refers to the sequence of the different amino acids of the peptide or protein. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code. Post-translational modifications such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene.

Secondary structure



An alpha-helix with hydrogen bonds (yellow dots)

Secondary structure refers to highly regular local sub-structures. Two main types of secondary structure, the alpha helix and the beta strand, were suggested in 1951 by Linus

Pauling and coworkers.. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. Some parts of the protein are ordered but do not form any regular structures. They should not be confused with random coil, an unfolded polypeptide chain lacking any fixed three-dimensional structure. Several sequential secondary structures may form a "supersecondary unit".

Tertiary structure

Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by the *non-specific* hydrophobic interactions (the burial of hydrophobic residues from water), but the structure is stable only when the parts of a protein domain are locked into place by *specific* tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment.

Quaternary structure

Quaternary structure is a larger assembly of several protein molecules or polypeptide chains, usually called subunits in this context. The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, and a tetramer if it contains four subunits. The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer. Multimers made up of identical subunits are referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits are referred to with a prefix of "hetero-" (e.g. a heterotetramer, such as the two alpha and two beta chains of hemoglobin). Many proteins do not have the quaternary structure and function as monomers.

Domains, motifs, and folds in protein structure

Protein are frequently described as consisting from several structural units.

- A **structural domain** is an element of the protein's overall structure that is self-stabilizing and often folds independently of the rest of the protein chain. Many domains are not unique to the protein products of one gene or one gene family but instead appear in a variety of proteins. Domains often are named and singled out because they figure prominently in the biological function of the protein they belong to; for example, the "calcium-binding domain of calmodulin". Because they are independently stable, domains can be "swapped" by genetic engineering between one protein and another to make chimeras.

- The **structural and sequence motifs** refer to short segments of protein three-dimensional structure or amino acid sequence that were found in a large number of different proteins.

- The **supersecondary structure** refers to a specific combination of secondary structure elements, such as beta-alpha-beta units or helix-turn-helix motif. Some of them may be also referred to as structural motifs.

- **Protein fold** refers to the general protein architecture, like helix bundle, beta-barrel, Rossman fold or different "folds" provided in the Structural Classification of Proteins database.

Despite the fact that there are about 100,000 different proteins expressed in eukaryotic systems, there are many fewer different domains, structural motifs and folds. This is partly a consequence of evolution, since genes or parts of genes can be doubled or moved around within the genome. This means that, for example, a protein domain might be moved from one protein to another thus giving the protein a new function. Because of these mechanisms, pathways and mechanisms tend to be reused in several different proteins.

Protein folding

An unfolded polypeptide folds into its characteristic three-dimensional structure from random coil.

Protein structure determination

Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows one to measure the 3D density distribution of electrons in the protein (in the crystallized state) and thereby infer the 3D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by Nuclear Magnetic Resonance techniques. The secondary structure composition can be determined via circular dichroism or dual polarisation interferometry. Cryo-electron microscopy has recently become a means of determining protein structures to high resolution (less than 5 angstroms or 0.5 nanometer) and is anticipated to increase in power as a tool for high resolution work in the next decade. This technique is still a valuable resource for researchers working with very large protein complexes such as virus coat proteins and amyloid fibers.

Structure classification

Protein structures can be classified based on their similarity or a common evolutionary origin. SCOP and CATH databases provide two different structural classifications of proteins.

Computational prediction of protein structure

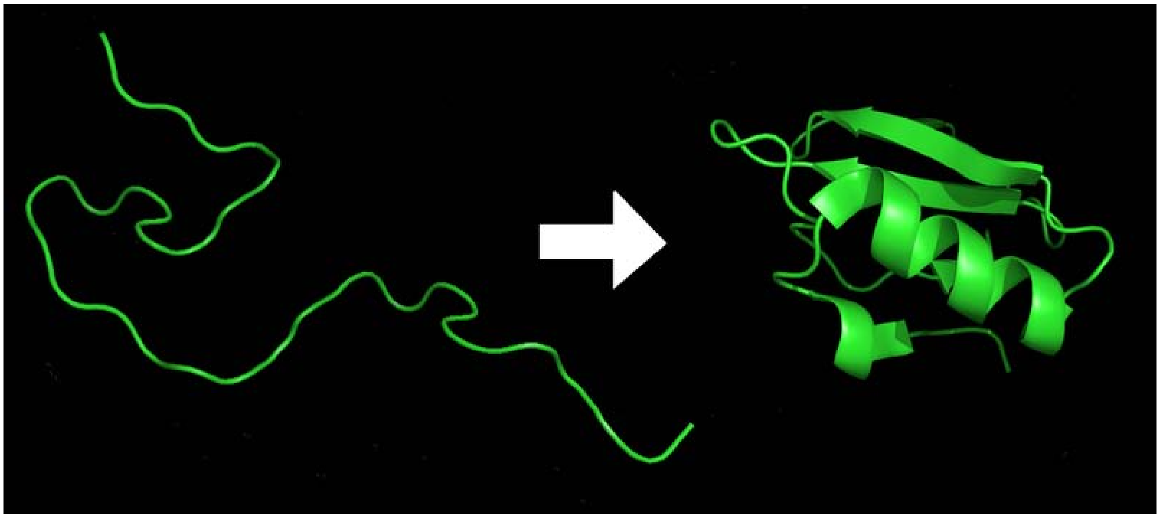
The generation of a protein sequence is much easier than the determination of a protein structure. However, the structure of a protein gives much more insight in the function of the protein than its sequence. Therefore, a number of methods for the computational prediction of protein structure from its sequence have been developed. *Ab initio* prediction methods use just the sequence of the protein. Threading and Homology Modeling methods can build a 3D model for a protein of unknown structure from experimental structures of evolutionary related proteins.

Protein structure related software

There are software to aid researchers working on, often overlapping, different aspects of protein structure. The most basic functionality is providing structure visualization. Analysis of protein structure can be facilitated by software that aligns structures. In the absence of existing structures for a given protein sequence, there are methods to predict or to model the structure of such sequences based on known protein structures. And given models of known or predicted structures, one can use software to verify them for errors, predict protein conformational changes, or predict substrate binding sites.

Chapter 2

Protein Folding



Protein before and after folding.

Protein folding 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 developed three-dimensional structure (the left hand side of the neighboring figure). Amino acids interact with each other to produce a well-defined three-dimensional structure, the folded protein (the right hand side of the figure), known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence (Anfinsen's dogma).

The correct three-dimensional structure is essential to function, although some parts of functional proteins may remain unfolded Failure to fold into the intended shape usually produces inactive proteins with different properties including toxic prions. Several neurodegenerative and other diseases are believed to result from the accumulation of *misfolded* (incorrectly folded) proteins. Many allergies are caused by the folding of the proteins, for the immune system does not produce antibodies for certain protein structures.

Known facts

Relationship between folding and amino acid sequence

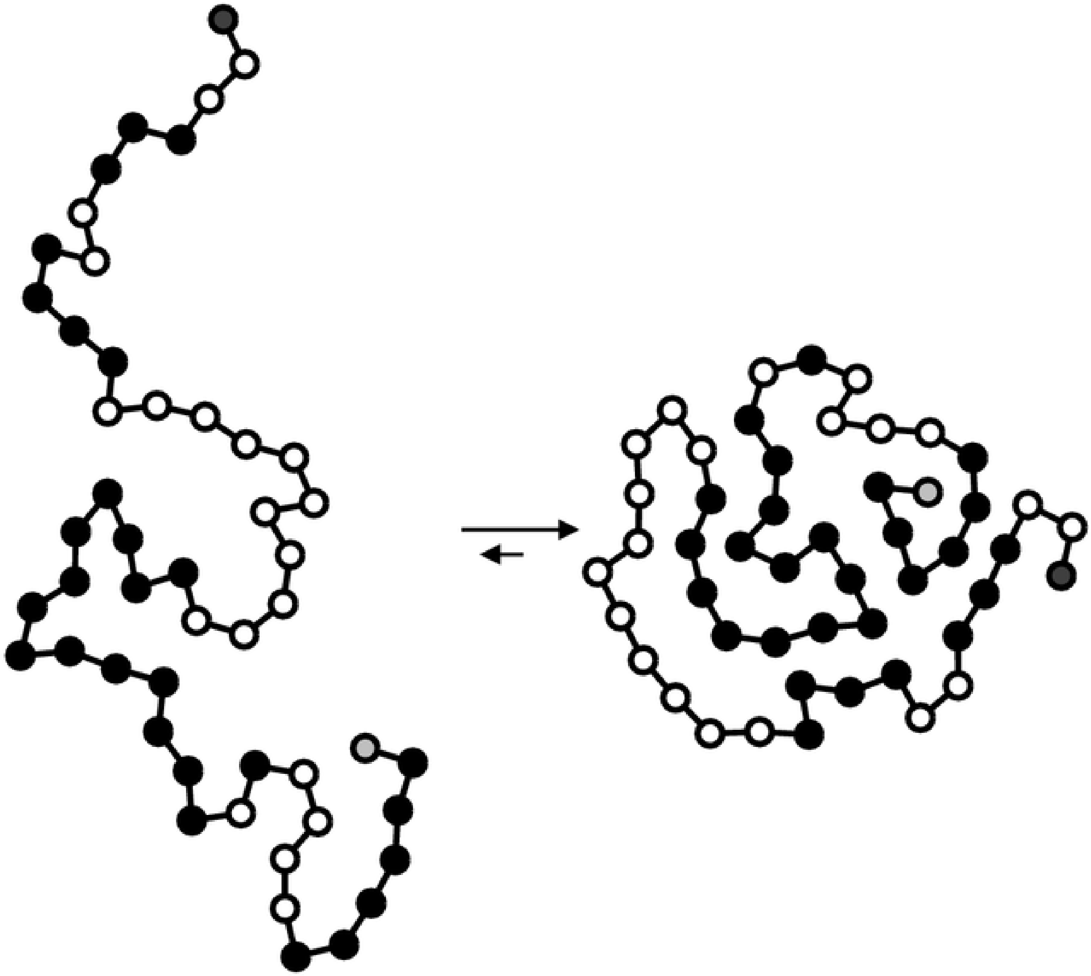


Illustration of the main driving force behind protein structure formation. In the compact fold (to the right), the hydrophobic amino acids (shown as black spheres) are in general shielded from the solvent.

The amino-acid sequence (or primary structure) of a protein determines its native conformation. A protein molecule folds 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 temperature, and the presence of molecular chaperones.

Folded proteins usually have a hydrophobic core in which side chain packing stabilizes the folded state, and charged or polar side chains occupy the solvent-exposed surface where they interact with surrounding water. Minimizing the number of hydrophobic side-chains exposed to water is an important driving force behind the folding process. Formation of intramolecular hydrogen bonds provides another important contribution to protein stability. 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.

The process of folding *in vivo* 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. Specialized proteins called chaperones assist in the folding of other proteins. A well studied example is the bacterial GroEL system, which assists in the folding of globular proteins. In eukaryotic organisms chaperones are known as heat shock proteins. Although most globular proteins are able to assume their native state unassisted, chaperone-assisted folding is often necessary in the crowded intracellular environment to prevent aggregation; chaperones are also used to prevent misfolding and aggregation which may occur as a consequence of exposure to heat or other changes in the cellular environment.

There are two models of protein folding that are currently being confirmed. The first is the diffusion collision model in which a nucleus is formed, then the secondary structure, and finally these secondary structures are collided together and pack tightly together. The next model is the nucleation-condensation model, in which the secondary and tertiary structure of the protein is made at the same time. Finally, recent studies have shown that some proteins show characteristics of both of these folding models.

For the most part, scientists have been able to study many identical molecules folding together *en masse*. At the coarsest level, it appears that in transitioning to the native state, a given amino acid sequence takes on roughly the same route and proceeds through roughly the same intermediates and transition states. Often folding involves first the establishment of regular secondary and supersecondary structures, particularly alpha helices and beta sheets, and afterwards tertiary structure. Formation of quaternary structure usually involves the "assembly" or "coassembly" of subunits that have already folded. The regular alpha helix and beta sheet structures fold rapidly because they are stabilized by intramolecular hydrogen bonds, as was first characterized by Linus Pauling. Protein folding may involve covalent bonding in the form of disulfide bridges formed between two cysteine residues or the formation of metal clusters. Shortly before settling into their more energetically favourable native conformation, molecules may pass through an intermediate "molten globule" state.

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. Folding is a spontaneous process independent of energy inputs from nucleoside triphosphates. The passage of the folded state is mainly guided by hydrophobic interactions, formation of intramolecular hydrogen bonds, and van der Waals forces, and it is opposed by conformational entropy.

Disruption of the native state

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). High concentrations of solutes, extremes of pH, mechanical forces, and the presence of chemical denaturants can do the same. 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.

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 chaperones or heat shock proteins, which assist other proteins both in folding and in remaining folded. Some proteins never fold in cells at all except with the assistance of chaperone molecules, which either isolate individual proteins so that their folding is not interrupted by interactions with other proteins or help to unfold misfolded proteins, giving them a second chance to refold properly. This function is crucial to prevent the risk of precipitation into insoluble amorphous aggregates.

Incorrect protein folding and neurodegenerative disease

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 intracytoplasmic aggregation diseases such as Huntington's and Parkinson's disease. These age onset degenerative diseases are associated with the multimerization of misfolded proteins into insoluble, extracellular aggregates and/or intracellular inclusions including cross-beta sheet amyloid fibrils; it is not 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. 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.

Effect of external factors on the folding of Proteins

Several external factors such as temperature, external fields (electric, magnetic), , molecular crowding , limitation of space could have a big influence on the folding of proteins . Modification of the local minima by external factors can also induce modifications of the folding trajectory.

The Levinthal paradox and kinetics

The Levinthal paradox observes that if a protein were to fold by sequentially sampling all possible conformations, it would take an astronomical amount of time to do so, even if the conformations were sampled at a rapid rate (on the nanosecond or picosecond scale). Based upon the observation that proteins fold much faster than this, Levinthal then proposed that a random conformational search does not occur, and the protein must, therefore, fold through a series of meta-stable intermediate states.

The duration of the folding process varies dramatically depending on the protein of interest. When studied outside the cell, the slowest folding proteins require many minutes or hours to fold primarily due to proline isomerization, and must pass through a number of intermediate states, like checkpoints, before the process is complete. On the other hand, very small single-domain proteins with lengths of up to a hundred amino acids typically fold in a single step. Time scales of milliseconds are the norm and the very fastest known protein folding reactions are complete within a few microseconds.

Energy landscape theory of protein folding

The protein folding phenomenon was largely an experimental endeavor until the formulation of an energy landscape theory of proteins by Joseph Bryngelson and Peter Wolynes in the late 1980s and early 1990s. This approach introduced the *principle of minimal frustration*,. This principle says that nature has chosen amino acid sequences so that the folded state of the protein is very stable. Additionally, the undesired interactions between amino acids along the folding pathway are reduced making the acquisition of the folded state a very fast process. Even though nature has reduced the level of *frustration* in proteins, some degree of it remains up to now as can be observed in the presence of local minima in the energy landscape of proteins. A consequence of these evolutionarily selected sequences is that proteins are generally thought to have globally "funneled energy landscapes" (coined by José Onuchic) that are largely directed towards the native state. This "folding funnel" landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather than being restricted to a single mechanism. The theory is supported by both computational simulations of model proteins and experimental studies, and it has been used to improve methods for protein structure prediction and design. The description of protein folding by the leveling free-energy landscape is also consistent with the 2nd law of thermodynamics. Physically, thinking of landscapes in terms of visualizable potential or total energy surfaces simply with maxima, saddle points, minima and funnels, rather like geographic landscapes, is perhaps a little misleading. The relevant description is really a highly dimensional phase space in which manifolds might take a variety of more complicated topological forms: see for example .

Techniques for studying protein folding

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 helicies 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 can be used to measure equilibrium unfolding of the protein by measuring the change in this absorption as a function of denaturant concentration or temperature. A denaturant melt measures the free energy of unfolding as well as the protein's m value, or denaturant dependence. A temperature melt measures the melting temperature (T_m) of the protein. This type of spectroscopy can also be combined with fast-mixing devices, such as stopped flow, to measure protein folding kinetics and to generate chevron plots.

Dual Polarisation Interferometry

Dual polarisation interferometry is a surface based technique for measuring the optical properties of molecular layers. When used to characterise 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.

Vibrational circular dichroism of proteins

The more recent developments of vibrational circular dichroism (VCD) techniques for proteins, currently involving Fourier transform (FFT) instruments, provide powerful means for determining protein conformations in solution even for very large protein molecules. Such VCD studies of proteins are often combined with X-ray diffraction of protein crystals, FT-IR data for protein solutions in heavy water (D_2O), or *ab initio* quantum computations to provide unambiguous structural assignments that are unobtainable from CD.

Modern 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. These are experimental methods for rapidly triggering the folding of a sample of unfolded protein, and then observing the resulting dynamics. Fast techniques in widespread 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.

Computational prediction of protein tertiary structure

De novo or *ab initio* techniques for computational protein structure prediction are related to, but strictly distinct from, studies involving protein folding. Molecular Dynamics (MD) is an important tool for studying protein folding and dynamics in silico. 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. In order to simulate long time folding processes (beyond about 1 microsecond), like folding of small-size proteins (about 50 residues) or larger, some approximations or simplifications in protein models need to be introduced. An approach using reduced protein representation (pseudo-atoms representing groups of atoms are defined) and statistical potential is not only useful in protein structure prediction, but is also capable of reproducing the folding pathways.

There are distributed computing projects which use idle CPU or GPU time of personal computers to solve problems such as protein folding or prediction of protein structure. One such prominent example being the Folding@Home project. People can run these programs on their computer or PlayStation 3 to support them.

Experimental techniques of protein structure determination

Folded structures of proteins are routinely determined by X-ray crystallography and NMR. Dynamic methods to characterise protein folding such as dual polarisation interferomety and CD provide a measurement of conformation and conformational change rather than structure.

Chapter 3

Protein Design & Fusion Protein

Protein Design

Protein design is the design of new protein molecules, either from scratch or by making calculated variations on a known structure. The use of rational protein design techniques is a major aspect of protein engineering.

The design of minimalist computer models of proteins (lattice proteins), and the secondary structural modification of real proteins, began in the mid-1990s. The *de novo* design of real proteins became possible shortly afterwards, and the 21st century has seen the creation of small proteins with real biological functions including chiroselective catalysis, ion detection, and antiviral behaviour. There is great hope that the design of these and larger proteins will have applications in medicine and bioengineering. Recent computational redesign was capable of experimentally switching the cofactor specificity of Candida boidinii xylose reductase from NADPH to NADH.

Overview

The number of possible amino acid sequences is enormous, but only a subset of them will fold reliably and quickly to a single native state. Protein design involves identifying novel sequences within this subset, in particular those with a physiologically active native state. Physically, the native state of a protein is the conformational free energy minimum for the chain. Therefore protein design is the search for sequences which have the chosen structure as a free energy minimum. In a sense it is the reverse of structure prediction: a tertiary structure is specified, and a sequence is identified which will fold to it. Hence it is also referred to as *inverse folding*. Prion diseases like Mad Cow Disease are helpful examples of how important it is that designer proteins possess only one possible stable conformation. In Mad Cow Disease, there exists a healthy protein with a fatal weakness: there is another conformation this protein can "comfortably" take; the abnormally folded shape has very little free energy and is therefore very stable. For reasons that are not yet fully understood, this mis-folded prion protein has the ability to catalyze other proteins of its type to also adopt the mis-folded prion shape, which results in a disease-generating cascade of previously functional proteins quickly becoming misfolded. They lose the ability to perform their intended function in the new conformation, and have a tendency to form aggregates called plaques. The buildup of these aggregates in the brain leads to progressive neuronal death, and eventually death of the entire organism.

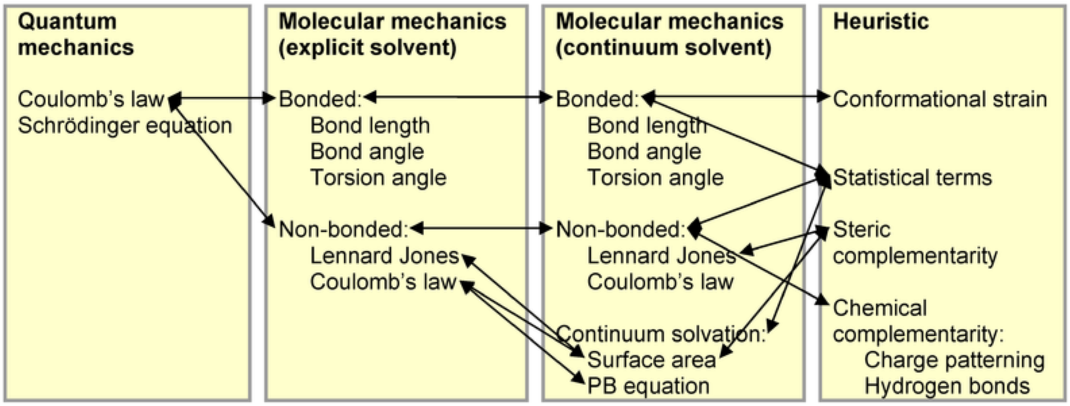
It is therefore easy to see both how important it is that a designer protein have only one possible stable tertiary structure, and that researchers exercise extreme diligence to ensure that this remains the case in all environments - especially *in vivo*.

Protein design requires an understanding of the molecular interactions that stabilize proteins in specific folded configurations; experience has shown, however, that it does not require an understanding of the dynamical process by which proteins fold,

Protein design can be accomplished using computer models, which, while simplifying the problem, are able to generate sequences that fold to the desired structure. Computational protein design algorithms search the sequence-conformation space for sequences that are low in energy when folded to the target structure. This search space is large; currently the most challenging requirement for computational protein design is a fast, yet accurate, energy function that can distinguish optimal sequences from similar suboptimal ones. Using computational methods, a protein with a novel fold has been designed, as well as sensors for unnatural molecules.

On the other hand, it is widely believed that not all possible protein structures are *designable*, which means that there are compact configurations of the chain which no sequences can fold to. In particular, conformations which are poor in secondary structures are unlikely to be designable. The designability of given structures is an issue that is still poorly understood.

Models of protein structure and function used in protein design



Comparison of various potential energy functions

Computational protein design algorithms use models of protein energetics to evaluate how mutations would affect a protein's structure and function. These energy functions typically include a combination of molecular mechanics, statistical (i.e. knowledge-based), and other empirical terms. However, the trend has been towards using more physically based potential energy functions.

Ancestral sequence reconstruction

Ancestral reconstruction techniques have been used to design proteins with putative ancient functions.

Software

Iterative Protein Redesign and Optimization. IPRO redesigns proteins to increase or give specificity to native or novel substrates and cofactors. This is done by repeatedly randomly perturbing the backbones of the proteins around specified design positions, identifying the lowest energy combination of rotamers, and determining if the new design has a lower binding energy than previous ones. The iterative nature of this process allows IPRO to make additive mutations to the protein sequence that collectively improve the specificity towards the desired substrates and/or cofactors. Experimental testing of predictions by IPRO successfully switched the cofactor preference of Candida boidinii xylose reductase from NADPH to NADH. Details on how to download the software implemented in Python and experimental testing of predictions are outlined in the following paper.

EGAD: A Genetic Algorithm for protein Design. A free, open-source software package for protein design and prediction of mutation effects on protein folding stabilities and binding affinities. EGAD can also consider multiple structures simultaneously for designing specific binding proteins or locking proteins into specific conformational states. In addition to natural protein residues, EGAD can also consider free-moving ligands with or without rotatable bonds. EGAD can be used with single or multiple processors.

RosettaDesign. A software package, under active development and free for academic use, that has seen extensive successful use. RosettaDesign is accessible via a web server.

SHARPEN. A permissive open-source library for protein design and structure prediction. SHARPEN offers a variety of combinatorial optimization methods (e.g. Monte Carlo, Simulated Annealing, FASTER) and can score proteins using the successful Rosetta all-atom force field or molecular mechanics force fields (OPLSaa). In addition to the protein modeling library, SHARPEN includes tools for scalable distributed computing.

WHAT IF software for protein modelling, design, validation, and visualisation.

Abalone software for protein modelling and visualisation.

Fusion Protein

Fusion proteins or **chimeric proteins** are proteins created through the joining of two or more genes which originally coded for separate proteins. Translation of this *fusion gene* results in a single polypeptide with functional properties derived from each of the original proteins. *Recombinant fusion proteins* are created artificially by recombinant DNA technology for use in biological research or therapeutics. *Chimeric mutant proteins* occur naturally when a complex mutation, such as a chromosomal translocation, tandem duplication, or retrotransposition creates a novel coding sequence containing parts of the coding sequences from two different genes. Naturally occurring fusion proteins are commonly found in cancer cells, where they may function as oncoproteins. The bcr-abl fusion protein is a well-known example of an oncogenic fusion protein, and is considered to be the primary oncogenic driver of chronic myelogenous leukemia.

Functions

Some fusion proteins combine whole peptides and therefore contains all functional domains of the original proteins. However, other fusion proteins, especially those that are naturally occurring, combine only portions of coding sequences and therefore do not maintain the original functions of the parental genes that formed them.

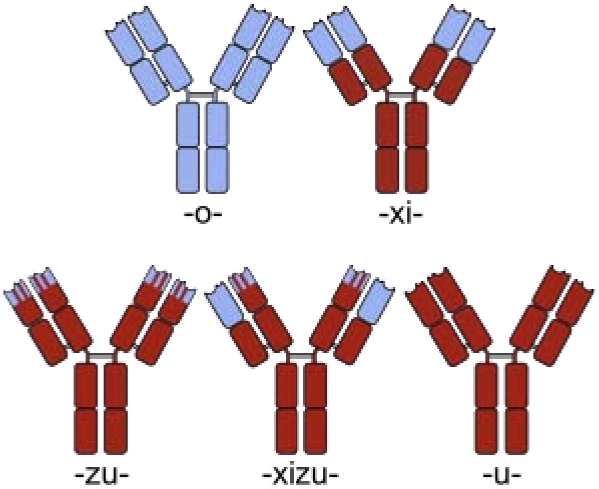
Many whole gene fusions are fully functional, and can still act to replace the original peptides. Some, however, experience interactions between the two proteins that can modify their functions. Beyond these effects, some gene fusions may cause regulatory changes that alter when and where these genes act. For partial gene fusions, the shuffling of different active sites and binding domains can potentially result in new proteins with novel functions.

Recombinant technology

A **recombinant fusion protein** is a protein created through genetic engineering of a fusion gene. This typically involves removing the stop codon from a cDNA sequence coding for the first protein, then appending the cDNA sequence of the second protein in frame through ligation or overlap extension PCR. That DNA sequence will then be expressed by a cell as a single protein. The protein can be engineered to include the full sequence of both original proteins, or only a portion of either.

If the two entities are proteins, often linker (or "spacer") peptides are also added which make it more likely that the proteins fold independently and behave as expected. Especially in the case where the linkers enable protein purification, linkers in protein or peptide fusions are sometimes engineered with cleavage sites for proteases or chemical agents which enable the liberation of the two separate proteins. This technique is often used for identification and purification of proteins, by fusing a GST protein, FLAG peptide, or a hexa-his peptide (6xHis-tag) which can be isolated using affinity chromatography with nickel or cobalt resins. Fusion proteins can also be manufactured with toxins or antibodies attached to them in order to study disease development.

Chimeric protein drugs



Sketches of mouse (top left), chimeric (top right) and humanized (bottom left) monoclonal antibodies. Human parts are shown in brown, non-human parts in blue.

The purpose of creating fusion proteins in drug development is to impart properties from each of the "parent" proteins to the resulting chimeric protein. Several chimeric protein drugs are currently available for medical use.

Many chimeric protein drugs are monoclonal antibodies whose specificity for a target molecule was developed using mice and hence were initially "mouse" antibodies. As non-human proteins, mouse antibodies tend to evoke an immune reaction if administered to humans. The chimerization process involves engineering the replacement of segments of the antibody molecule that distinguish it from a human antibody. For example, human constant domains can be introduced, thereby eliminating most of the potentially immunogenic portions of the drug without altering its specificity for the intended therapeutic target. Antibody nomenclature indicates this type of modification by inserting *-xi-* into the non-proprietary name (e.g. abci-*xi*-mab). If parts of the variable domains are also replaced by human portions, *humanized* antibodies are obtained. Although not conceptually distinct from chimeras, this type is indicated using *-zu-* such as in dacli-*zu*-mab.

In addition to chimeric and humanized antibodies, there are other pharmaceutical purposes for the creation of chimeric constructs. Etanercept, for example, is a TNFα blocker created through the combination of a tumor necrosis factor receptor (TNFR) with the immunoglobulin G1 Fc segment. TNFR provides specificity for the drug target and the antibody Fc segment is believed to add stability and deliverability of the drug.

Natural occurrence

Naturally occurring fusion genes are most commonly created when a chromosomal translocation replaces the terminal exons of one gene with intact exons from a second gene. This creates a single gene which can be transcribed, spliced, and translated to produce a functional fusion protein. Many important cancer-promoting oncogenes are fusion genes produced in this way.

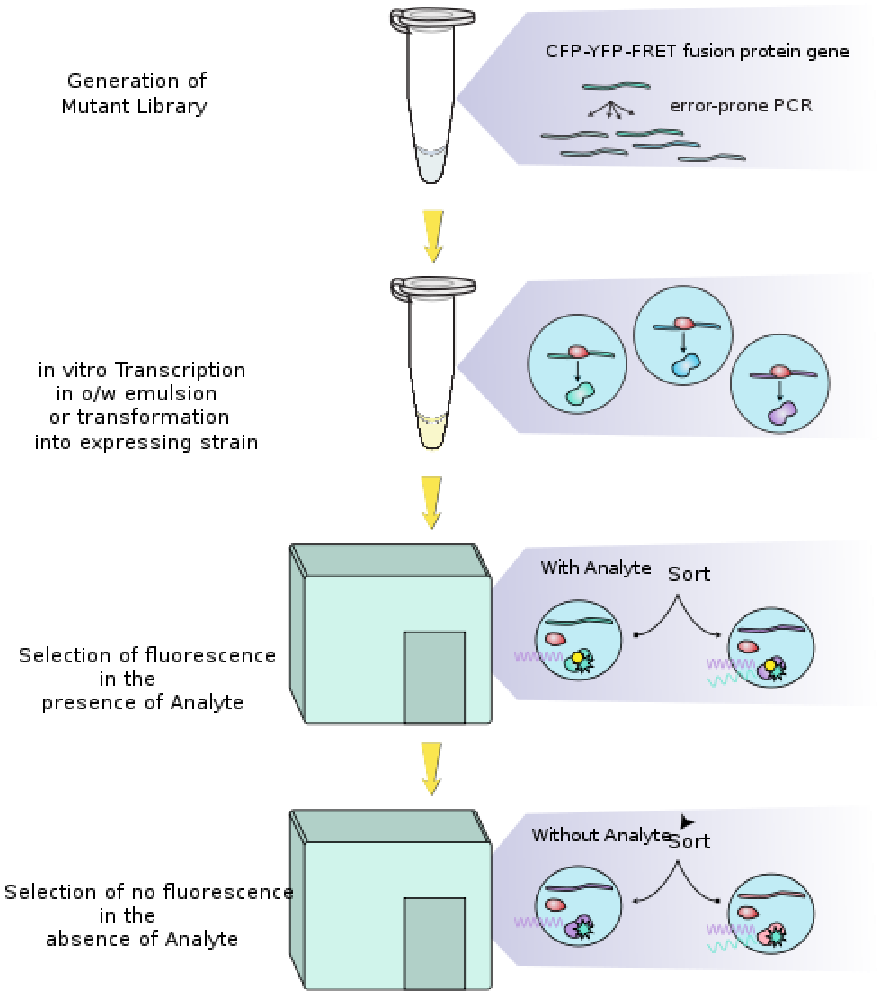
Examples include:

- Gag-onc fusion protein
- Bcr-abl fusion protein
- Tpr-met fusion protein

Antibodies are fusion proteins produced by VDJ recombination.

Chapter 4

Directed Evolution


Generation of Mutant Library
CFP-YFP-FRET fusion protein gene
error-prone PCR
in vitro Transcription in o/w emulsion or transformation into expressing strain
Selection of fluorescence in the presence of Analyte
With Analyte
Sort
Selection of no fluorescence in the absence of Analyte
Without Analyte
Sort

An example of a possible round to evolve a protein based fluorescent sensor for a specific analyte using two consecutive FACS sortings

Directed evolution is a method used in protein engineering to harness the power of natural selection to evolve proteins or RNA with desirable properties not found in nature.

A typical experiment

A typical directed evolution experiment involves three steps:

1. *Diversification:* The gene encoding the protein of interest is mutated and/or recombined at random to create a large library of gene variants. Techniques commonly used in this step are error-prone PCR and DNA shuffling.
2. *Selection:* The library is tested for the presence of mutants (variants) possessing the desired property using a screen or selection. Screens enable the researcher to identify and isolate high-performing mutants by hand, while selections automatically eliminate all nonfunctional mutants.
3. *Amplification:* The variants identified in the selection or screen are replicated manyfold, enabling researchers to sequence their DNA in order to understand what mutations have occurred.

Together, these three steps are termed a "round" of directed evolution. Most experiments will perform more than one round. In these experiments, the "winners" of the previous round are diversified in the next round to create a new library. At the end of the experiment, all evolved protein or RNA mutants are characterized using biochemical methods.

Likelihood of success

The likelihood of success in a directed evolution experiment is directly related to the total library size, as evaluating more mutants increases the chances of finding one with the desired properties. Performing multiple rounds of evolution is useful not only because a new library of mutants is created in each round, but because each new library uses better mutants as templates. The experiment is analogous to climbing a hill on a landscape where elevation is a function of the desired property. The goal is to reach the summit, which represents the best mutant. Each round of selection samples mutants on all sides of the starting template and selects the mutant with the highest elevation, thereby climbing the hill. A new round samples mutants on all sides of this new template and picks the highest of these, and so on until the summit is reached.

In vivo *and* in vitro

Directed evolution can be performed in living cells (*in vivo* evolution) or may not involve cells at all (*in vitro* evolution). *In vivo* evolution has the advantage of selecting for properties in a cellular environment, which is useful when the evolved protein or RNA is to be used in living organisms, but *in vitro* evolution is often more versatile in the types of selections that can be performed. Furthermore, *in vitro* evolution experiments can generate larger libraries because the library DNA need not be inserted into cells, the currently limiting step.

Advantages

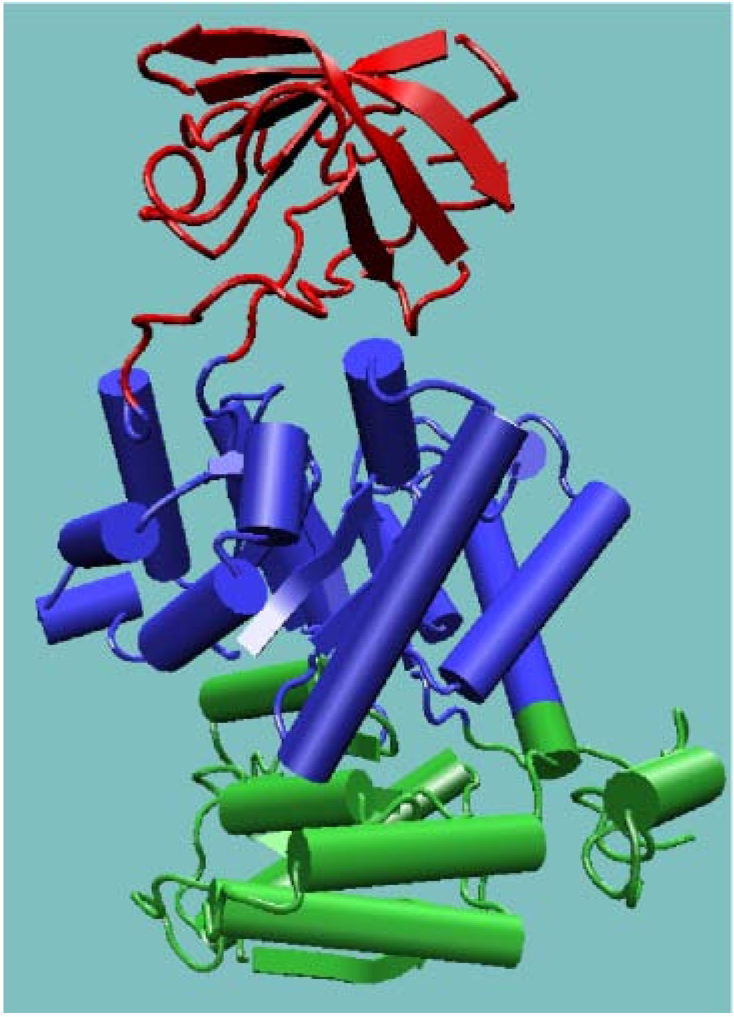
The advantage of the directed evolution approach is that the researcher need not understand the mechanism of the desired activity in order to improve it. An alternative method is rational design of site-directed mutagenesis based on X-ray crystallography data.

Uses

Most directed evolution projects seek to evolve properties that are useful to humans in an agricultural, medical or industrial context (biocatalysis). It is thus possible to use this method to optimize properties that were not selected for in the original organism. This may include catalytic activity, catalytic specificity, thermostability and many others.

Chapter 5

Protein Domain



Pyruvate kinase, a protein from three domains (PDB 1pkn)

A **protein domain** is a part of protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist of several structural domains. One domain may appear in a variety of different proteins. Molecular evolution uses domains as building blocks and these may be recombined in different arrangements to create proteins with different functions. Domains vary in length from between about 25 amino acids up to 500 amino acids in length. The shortest domains such as zinc fingers are stabilized by metal ions or disulfide bridges. Domains often form functional units, such as the calcium-binding EF hand domain of calmodulin. Because they are independently stable, domains can be "swapped" by genetic engineering between one protein and another to make chimeric proteins.

Background

The concept of the **domain** was first proposed in 1973 by Wetlaufer after X-ray crystallographic studies of hen lysozyme and papain and by limited proteolysis studies of immunoglobulins . Wetlaufer defined domains as stable units of protein structure that could fold autonomously. In the past domains have been described as units of:

- compact structure
- function and evolution
- folding .

Each definition is valid and will often overlap, i.e. a compact structural domain that is found amongst diverse proteins is likely to fold independently within its structural environment. Nature often brings several domains together to form multidomain and multifunctional proteins with a vast number of possibilities . In a multidomain protein, each domain may fulfil its own function independently, or in a concerted manner with its neighbours. Domains can either serve as modules for building up large assemblies such as virus particles or muscle fibres, or can provide specific catalytic or binding sites as found in enzymes or regulatory proteins.

An appropriate example is pyruvate kinase, a glycolytic enzyme that plays an important role in regulating the flux from fructose-1,6-biphosphate to pyruvate. It contains an all-β regulatory domain, an α/β-substrate binding domain and an α/β-nucleotide binding domain, connected by several polypeptide linkers. Each domain in this protein occurs in diverse sets of protein families.

The central α/β-barrel substrate binding domain is one of the most common enzyme folds. It is seen in many different enzyme families catalysing completely unrelated reactions. The α/β-barrel is commonly called the TIM barrel named after triose phosphate isomerase, which was the first such structure to be solved. It is currently classified into 26 homologous families in the CATH domain database . The TIM barrel is formed from a sequence of β-α-β motifs closed by the first and last strand hydrogen bonding together, forming an eight stranded barrel. There is debate about the evolutionary origin of this domain. One study has suggested that a single ancestral enzyme could have diverged into several families, while another suggests that a stable TIM-barrel structure has evolved through convergent evolution .

The TIM-barrel in pyruvate kinase is 'discontinuous', meaning that more than one segment of the polypeptide is required to form the domain. This is likely to be the result of the insertion of one domain into another during the protein's evolution. It has been shown from known structures that about a quarter of structural domains are discontinuous. The inserted β-barrel regulatory domain is 'continuous', made up of a single stretch of polypeptide.

Covalent association of two domains represents a functional and structural advantage since there is an increase in stability when compared with the same structures non-covalently associated . Other, advantages are the protection of intermediates within inter-domain enzymatic clefts that may otherwise be unstable in aqueous environments, and a fixed stoichiometric ratio of the enzymatic activity necessary for a sequential set of reactions .

Domains are units of protein structure

The primary structure (string of amino acids) of a protein ultimately encodes its uniquely folded 3D conformation. The most important factor governing the folding of a protein into 3D structure is the distribution of polar and non-polar side chains. Folding is driven by the burial of hydrophobic side chains into the interior of the molecule so to avoid contact with the aqueous environment. Generally proteins have a core of hydrophobic residues surrounded by a shell of hydrophilic residues. Since the peptide bonds themselves are polar they are neutralised by hydrogen bonding with each other when in the hydrophobic environment. This gives rise to regions of the polypeptide that form regular 3D structural patterns called secondary structure. There are two main types of secondary structure: α-helices and β-sheets.

Some simple combinations of secondary structure elements have been found to frequently occur in protein structure and are referred to as supersecondary structure or motifs. For example, the β-hairpin motif consists of two adjacent antiparallel β-strands joined by a small loop. It is present in most antiparallel β structures both as an isolated ribbon and as part of more complex β-sheets. Another common super-secondary structure is the β-α-β motif, which is frequently used to connect two parallel β-strands. The central α-helix connects the C-termini of the first strand to the N-termini of the second strand, packing its side chains against the β-sheet and therefore shielding the hydrophobic residues of the β-strands from the surface.

Structural alignment is an important tool for determining domains.

Tertiary structure of domains

Several motifs pack together to form compact, local, semi-independent units called domains. The overall 3D structure of the polypeptide chain is referred to as the protein's tertiary structure. Domains are the fundamental units of tertiary structure, each domain containing an individual hydrophobic core built from secondary structural units connected by loop regions. The packing of the polypeptide is usually much tighter in the interior than the exterior of the domain producing a solid-like core and a fluid-like surface. In fact, core residues are often conserved in a protein family, whereas the residues in loops are less conserved, unless they are involved in the protein's function. Protein tertiary structure can be divided into four main classes based on the secondary structural content of the domain.

- All-α domains have a domain core built exclusively from α-helices. This class is dominated by small folds, many of which form a simple bundle with helices running up and down.
- All-β domains have a core comprising of antiparallel β-sheets, usually two sheets packed against each other. Various patterns can be identified in the arrangement of the strands, often giving rise to the identification of recurring motifs, for example the Greek key motif.
- α+β domains are a mixture of all-α and all-β motifs. Classification of proteins into this class is difficult because of overlaps to the other three classes and therefore is not used in the CATH domain database.
- α/β domains are made from a combination of β-α-β motifs that predominantly form a parallel β-sheet surrounded by amphipathic α-helices. The secondary structures are arranged in layers or barrels.

Domains have limits on size

Domains have limits on size. The size of individual structural domains varies from 36 residues in E-selectin to 692 residues in lipoxygenase-1, but the majority, 90%, have less than 200 residues with an average of approximately 100 residues. Very short domains, less than 40 residues, are often stabilised by metal ions or disulfide bonds. Larger domains, greater than 300 residues, are likely to consist of multiple hydrophobic cores.

Domains and quaternary structure

Many proteins have a quaternary structure, which consists of several polypeptide chains that associate into an oligomeric molecule. Each polypeptide chain in such a protein is called a subunit. Hemoglobin, for example, consists of two α and two β subunits. Each of the four chains has an all-α globin fold with a heme pocket.

Domain swapping is a mechanism for forming oligomeric assemblies. In domain swapping, a secondary or tertiary element of a monomeric protein is replaced by the same element of another protein. Domain swapping can range from secondary structure elements to whole structural domains. It also represents a model of evolution for functional adaptation by oligomerisation, e.g. oligomeric enzymes that have their active site at subunit interfaces.

Domains as evolutionary modules

Nature is a tinkerer and not an inventor, new sequences are adapted from pre-existing sequences rather than invented. Domains are the common material used by nature to generate new sequences, they can be thought of as genetically mobile units, referred to as 'modules'. Often, the C and N termini of domains are close together in space, allowing them to easily be "slotted into" parent structures during the process of evolution. Many domain families are found in all three forms of life, Archaea, Bacteria and Eukarya. Domains that are repeatedly found in diverse proteins are often referred to as modules, examples can be found among extracellular proteins associated with clotting, fibrinolysis, complement, the extracellular matrix, cell surface adhesion molecules and cytokine receptors.

Molecular evolution gives rise to families of related proteins with similar sequence and structure. However, sequence similarities can be extremely low between proteins that share the same structure. Protein structures may be similar because proteins have diverged from a common ancestor. Alternatively, some folds may be more favored than others as they represent stable arrangements of secondary structures and some proteins may converge towards these folds over the course of evolution . There are currently about 45,000 experimentally determined protein 3D structures deposited within the Protein Data Bank (PDB). However this set contains a lot of identical or very similar structures. All proteins should be classified to structural families to understand their evolutionary relationships. Structural comparisons are best achieved at the domain level. For this reason many algorithms have been developed to automatically assign domains in proteins with known 3D structure.

The CATH domain database classifies domains into approximately 800 fold families, ten of these folds are highly populated and are referred to as 'super-folds'. Super-folds are defined as folds for which there are at least three structures without significant sequence similarity. The most populated is the α/β-barrel super-fold as described previously.

Multidomain proteins

The majority of genomic proteins, two-thirds in unicellular organisms and more than 80% in metazoa, are multidomain proteins created as a result of gene duplication events. Many domains in multidomain structures could have once existed as independent proteins. More and more domains in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes. For example, vertebrates have a multi-enzyme polypeptide containing the GAR synthetase, AIR synthetase and GAR transformylase modules (GARs-AIRs-GARt; GAR: glycinamide ribonucleotide synthetase/transferase; AIR: aminoimidazole ribonucleotide synthetase). In insects, the polypeptide appears as GARs-(AIRs)2-GARt, in yeast GARs-AIRs is encoded separately from GARt, and in bacteria each domain is encoded separately.

Origin

Multidomain proteins are likely to have emerged from a selective pressure during evolution to create new functions. Various proteins have diverged from common ancestors by different combinations and associations of domains. Modular units frequently move about, within and between biological systems through mechanisms of genetic shuffling:

- transposition of mobile elements including horizontal transfers (between species);

- gross rearrangements such as inversions, translocations, deletions and duplications;
- homologous recombination;
- slippage of DNA polymerase during replication.

Types of organisation

The simplest multidomain organisation seen in proteins is that of a single domain repeated in tandem. The domains may interact with each other or remain isolated, like beads on string. The giant 30,000 residue muscle protein titin comprises about 120 fibronectin-III-type and Ig-type domains. In the serine proteases, a gene duplication event has led to the formation of a two β-barrel domain enzyme. The repeats have diverged so widely that there is no obvious sequence similarity between them. The active site is located at a cleft between the two β-barrel domains, in which functionally important residues are contributed from each domain. Genetically engineered mutants of the chymotrypsin serine protease were shown to have some proteinase activity even though their active site residues were abolished and it has therefore been postulated that the duplication event enhanced the enzyme's activity.

Modules frequently display different connectivity relationships, as illustrated by the kinesins and ABC transporters. The kinesin motor domain can be at either end of a polypeptide chain that includes a coiled-coil region and a cargo domain. ABC transporters are built with up to four domains consisting of two unrelated modules, ATP-binding cassette and an integral membrane module, arranged in various combinations.

Not only do domains recombine, but there are many examples of a domain having been inserted into another. Sequence or structural similarities to other domains demonstrate that homologues of inserted and parent domains can exist independently. An example is that of the 'fingers' inserted into the 'palm' domain within the polymerases of the Pol I family. Since a domain can be inserted into another, there should always be at least one continuous domain in a multidomain protein. This is the main difference between definitions of structural domains and evolutionary/functional domains. An evolutionary domain will be limited to one or two connections between domains, whereas structural domains can have unlimited connections, within a given criterion of the existence of a common core. Several structural domains could be assigned to an evolutionary domain.

Domains are autonomous folding units

Folding

Protein folding - the unsolved problem : Since the seminal work of Anfinsen over forty years ago, the goal to completely understand the mechanism by which a polypeptide rapidly folds into its stable native conformation remains elusive. Many experimental folding studies have contributed much to our understanding, but the principles that govern protein folding are still based on those discovered in the very first studies of folding. Anfinsen showed that the native state of a protein is thermodynamically stable, the conformation being at a global minimum of its free energy.

Folding is a directed search of conformational space allowing the protein to fold on a biologically feasible time scale. The Levinthal paradox states that if an averaged sized protein would sample all possible conformations before finding the one with the lowest energy, the whole process would take billions of years. Proteins typically fold within 0.1 and 1000 seconds, therefore the protein folding process must be directed some way through a specific folding pathway. The forces that direct this search are likely to be a combination of local and global influences whose effects are felt at various stages of the reaction.

Advances in experimental and theoretical studies have shown that folding can be viewed in terms of energy landscapes, where folding kinetics is considered as a progressive organisation of an ensemble of partially folded structures through which a protein passes on its way to the folded structure. This has been described in terms of a folding funnel, in which an unfolded protein has a large number of conformational states available and there are fewer states available to the folded protein. A funnel implies that for protein folding there is a decrease in energy and loss of entropy with increasing tertiary structure formation. The local roughness of the funnel reflects kinetic traps, corresponding to the accumulation of misfolded intermediates. A folding chain progresses toward lower intra-chain free-energies by increasing its compactness. The chains conformational options become increasingly narrowed ultimately toward one native structure.

Advantage of domains in protein folding

The organisation of large proteins by structural domains represents an advantage for protein folding, with each domain being able to individually fold, accelerating the folding process and reducing a potentially large combination of residue interactions. Furthermore, given the observed random distribution of hydrophobic residues in proteins, domain formation appears to be the optimal solution for a large protein to bury its hydrophobic residues while keeping the hydrophilic residues at the surface.

However, the role of inter-domain interactions in protein folding and in energetics of stabilisation of the native structure, probably differs for each protein. In T4 lysozyme, the influence of one domain on the other is so strong that the entire molecule is resistant to proteolytic cleavage. In this case, folding is a sequential process where the C-terminal domain is required to fold independently in an early step, and the other domain requires the presence of the folded C-terminal domain for folding and stabilisation.

It has been found that the folding of an isolated domain can take place at the same rate or sometimes faster than that of the integrated domain. Suggesting that unfavourable interactions with the rest of the protein can occur during folding. Several arguments suggest that the slowest step in the folding of large proteins is the pairing of the folded domains. This is either because the domains are not folded entirely correctly or because the small adjustments required for their interaction are energetically unfavourable, such as the removal of water from the domain interface.

Domains and protein flexibility

The presence of multiple domains in proteins gives rise to a great deal of flexibility and mobility, leading to **protein domain dynamics**. Domain motions can be inferred by comparing different structures of a protein, or they can be directly observed using spectra measured by neutron spin echo spectroscopy. They can also be suggested by sampling in extensive molecular dynamics trajectories. Domain motions are important for:

- catalysis;
- regulatory activity;
- transport of metabolites;
- formation of protein assemblies; and
- cellular locomotion.

One of the largest observed domain motions is the `swivelling' mechanism in pyruvate phosphate dikinase. The phosphoinositide domain swivels between two states in order to bring a phosphate group from the active site of the nucleotide binding domain to that of the phosphoenolpyruvate/pyruvate domain. The phosphate group is moved over a distance of 45A involving a domain motion of about 100 degrees around a single residue.In enzymes, the closure of one domain onto another captures a substrate by an induced fit, allowing the reaction to take place in a controlled way. A detailed analysis by Gerstein led to the classification of two basic types of domain motion; hinge and shear. Only a relatively small portion of the chain, namely the inter-domain linker and side chains undergo significant conformational changes upon domain rearrangement.

Hinges by secondary structures

A study by Hayward found that the termini of α-helices and β-sheets form hinges in a large number of cases. Many hinges were found to involve two secondary structure elements acting like hinges of a door, allowing an opening and closing motion to occur. This can arise when two neighbouring strands within a β-sheet situated in one domain, diverge apart as they join the other domain. The two resulting termini then form the bending regions between the two domains. α-helices that preserve their hydrogen bonding network when bent are found to behave as mechanical hinges, storing `elastic energy' that drives the closure of domains for rapid capture of a substrate.

Helical to extended conformation

The interconversion of helical and extended conformations at the site of a domain boundary is not uncommon. In calmodulin, torsion angles change for five residues in the middle of a domain linking α-helix. The helix is split into two, almost perpendicular, smaller helices separated by four residues of an extended strand.

Shear motions

Shear motions involve a small sliding movement of domain interfaces, controlled by the amino acid side chains within the interface. Proteins displaying shear motions often have a layered architecture: stacking of secondary structures. The interdomain linker has merely the role of keeping the domains in close proximity.

Domain motion and functional dynamics in enzymes

The analysis of the internal dynamics of structurally different, but functionally similar enzymes has highlighted a common relationship between the positioning of the active site and the two principal protein sub-domains. In fact, for several members of the hydrolase superfamily, the catalytic site is located close to the interface separating the two principal quasi-rigid domains. Such positioning appears instrumental for maintaining the precise geometry of the active site, while allowing for an appreciable functionally-oriented modulation of the flanking regions resulting from the relative motion of the two sub-domains.

Domain definition from structural co-ordinates

The importance of domains as structural building blocks and elements of evolution has brought about many automated methods for their identification and classification in proteins of known structure. Automatic procedures for reliable domain assignment is essential for the generation of the domain databases, especially as the number of protein structures is increasing. Although the boundaries of a domain can be determined by visual inspection, construction of an automated method is not straightforward. Problems occur when faced with domains that are discontinuous or highly associated. The fact that there is no standard definition of what a domain really is has meant that domain assignments have varied enormously, with each researcher using a unique set of criteria.

A structural domain is a compact, globular sub-structure with more interactions within it than with the rest of the protein. Therefore, a structural domain can be determined by two visual characteristics; its compactness and its extent of isolation. Measures of local compactness in proteins have been used in many of the early methods of domain assignment and in several of the more recent methods.

Methods

One of the first algorithms used a Cα-Cα distance map together with a hierarchical clustering routine that considered proteins as several small segments, 10 residues in length. The initial segments were clustered one after another based on inter-segment distances; segments with the shortest distances were clustered and considered as single segments thereafter. The stepwise clustering finally included the full protein. Go also exploited the fact that inter-domain distances are normally larger than intra-domain distances; all possible Cα-Cα distances were represented as diagonal plots in which there were distinct patterns for helices, extended strands and combinations of secondary structures.

The method by Sowdhamini and Blundell clusters secondary structures in a protein based on their Cα-Cα distances and identifies domains from the pattern in their dendrograms. As the procedure does not consider the protein as a continuous chain of amino acids there are no problems in treating discontinuous domains. Specific nodes in these dendrograms are identified as tertiary structural clusters of the protein, these include both super-secondary structures and domains. The DOMAK algorithm is used to create the 3Dee domain database. It calculates a 'split value' from the number of each type of contact when the protein is divided arbitrarily into two parts. This split value is large when the two parts of the structure are distinct.

The method of Wodak and Janin was based on the calculated interface areas between two chain segments repeatedly cleaved at various residue positions. Interface areas were calculated by comparing surface areas of the cleaved segments with that of the native structure. Potential domain boundaries can be identified at a site where the interface area was at a minimum. Other methods have used measures of solvent accessibility to calculate compactness.

The PUU algorithm incorporates a harmonic model used to approximate inter-domain dynamics. The underlying physical concept is that many rigid interactions will occur within each domain and loose interactions will occur between domains. This algorithm is used to define domains in the FSSP domain database.

Swindells (1995) developed a method, DETECTIVE, for identification of domains in protein structures based on the idea that domains have a hydrophobic interior. Deficiencies were found to occur when hydrophobic cores from different domains continue through the interface region.

RigidFinder is a novel method for identification of protein rigid blocks (domains and loops) from two different conformations. Rigid blocks are defined as blocks where all inter residue distances are conserved across conformations.

A general method to identify *dynamical domains*, that is protein regions that behave approximately as rigid units in the course of structural fluctuations, has been introduced by Potestio et al. and, among other applications was also used to compare the consistency of the dynamics-based domain subdivisions with standard structure-based ones. The method, termed PiSQRD, is publicly available in the form of a webserver. The latter allows users to optimally subdivide single-chain or multimeric proteins into quasi-rigid domains based on the collective modes of fluctuation of the system. By default the latter are calculated through an elastic network model; alternatively pre-calculated essential dynamical spaces can be uploaded by the user.

Chapter 6

Proteomics

Robotic preparation of MALDI mass spectrometry samples on a sample carrier.

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of the genes. The word "proteome" is a blend of "**prote**in" and "gen**ome**", and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student. The proteome is the entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or system. This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes.

Complexity of the problem

After genomics, proteomics is considered the next step in the study of biological systems. It is much more complicated than genomics mostly because while an organism's genome is more or less constant, the proteome differs from cell to cell and from time to time. This is because distinct genes are expressed in distinct cell types. This means that even the basic set of proteins which are produced in a cell needs to be determined.

In the past this was done by mRNA analysis, but this was found not to correlate with protein content. It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present.

Post-translational modifications

Not only does the translation from mRNA cause differences, many proteins are also subjected to a wide variety of chemical modifications after translation. Many of these post-translational modifications are critical to the protein's function.

Phosphorylation

One such modification is phosphorylation, which happens to many enzymes and structural proteins in the process of cell signaling. The addition of a phosphate to particular amino acids—most commonly serine and threonine mediated by serine/threonine kinases, or more rarely tyrosine mediated by tyrosine kinases—causes a protein to become a target for binding or interacting with a distinct set of other proteins that recognize the phosphorylated domain.

Because protein phosphorylation is one of the most-studied protein modifications, many "proteomic" efforts are geared to determining the set of phosphorylated proteins in a particular cell or tissue-type under particular circumstances. This alerts the scientist to the signaling pathways that may be active in that instance.

Ubiquitination

Ubiquitin is a small protein that can be affixed to certain protein substrates by enzymes called E3 ubiquitin ligases. Determining which proteins are poly-ubiquitinated can be helpful in understanding how protein pathways are regulated. This is therefore an additional legitimate "proteomic" study. Similarly, once it is determined what substrates are ubiquitinated by each ligase, determining the set of ligases expressed in a particular cell type will be helpful.

Additional modifications

Listing all the protein modifications that might be studied in a "Proteomics" project would require a discussion of most of biochemistry; therefore, a short list will serve here to illustrate the complexity of the problem. In addition to phosphorylation and ubiquitination, proteins can be subjected to (among others) methylation, acetylation, glycosylation, oxidation and nitrosylation. Some proteins undergo ALL of these modifications, often in time-dependent combinations, aptly illustrating the potential complexity one has to deal with when studying protein structure and function.

Distinct proteins are made under distinct settings

Even if one is studying a particular cell type, that cell may make different sets of proteins at different times, or under different conditions. Furthermore, as mentioned, any one protein can undergo a wide range of post-translational modifications.

Therefore a "proteomics" study can become quite complex very quickly, even if the object of the study is very restricted. In more ambitious settings, such as when a biomarker for a tumor is sought - when the proteomics scientist is obliged to study sera samples from multiple cancer patients - the amount of complexity that must be dealt with is as great as in any modern biological project.

Limitations to genomic study

Scientists are very interested in proteomics because it gives a much better understanding of an organism than genomics. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Second, as mentioned above many proteins experience post-translational modifications that profoundly affect their activities; for example some proteins are not active until they become phosphorylated. Methods such as phosphoproteomics and glycoproteomics are used to study post-translational modifications. Third, many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications. Fourth, many proteins form complexes with other proteins or RNA molecules, and only function in the presence of these other molecules. Finally, protein degradation rate plays an important role in protein content.

Methods of studying proteins

Determining proteins which are post-translationally modified

One way in which a particular protein can be studied is to develop an antibody which is specific to that modification. For example, there are antibodies which only recognize certain proteins when they are tyrosine-phosphorylated, known as phospho-specific antibodies; also, there are antibodies specific to other modifications. These can be used to determine the set of proteins that have undergone the modification of interest.

For sugar modifications, such as glycosylation of proteins, certain lectins have been discovered which bind sugars. These too can be used.

A more common way to determine post-translational modification of interest is to subject a complex mixture of proteins to electrophoresis in "two-dimensions", which simply means that the proteins are electrophoresed first in one direction, and then in another... this allows small differences in a protein to be visualized by separating a modified protein from its unmodified form. This methodology is known as "two-dimensional gel electrophoresis".

Recently, another approach has been developed called PROTOMAP which combines SDS-PAGE with shotgun proteomics to enable detection of changes in gel-migration such as those caused by proteolysis or post translational modification.

Determining the existence of proteins in complex mixtures

Classically, antibodies to particular proteins or to their modified forms have been used in biochemistry and cell biology studies. These are among the most common tools used by practicing biologists today.

For more quantitative determinations of protein amounts, techniques such as ELISAs can be used.

For proteomic study, more recent techniques such as matrix-assisted laser desorption/ionization (MALDI) have been employed for rapid determination of proteins in particular mixtures and increasingly electrospray ionization (ESI).

Computational methods in studying protein biomarkers

Computational predictive models have shown that extensive and diverse feto-maternal protein trafficking occurs during pregnancy and can be readily detected non-invasively in maternal whole blood. This computational approach circumvented a major limitation, the abundance of maternal proteins interfering with the detection of fetal proteins, to fetal proteomic analysis of maternal blood. Computational models can use fetal gene transcripts previously identified in maternal whole blood to create a comprehensive proteomic network of the term neonate. Such work shows that the fetal proteins detected in pregnant woman’s blood originate from a diverse group of tissues and organs from the developing fetus. The proteomic networks contain many biomarkers that are proxies for development and illustrate the potential clinical application of this technology as a way to monitor normal and abnormal fetal development.

An information theoretic framework has also been introduced for biomarker discovery, integrating biofluid and tissue information. This new approach takes advantage of functional synergy between certain biofluids and tissues with the potential for clinically significant findings not possible if tissues and biofluids were considered individually. By conceptualizing tissue-biofluid as information channels, significant biofluid proxies can be identified and then used for guided development of clinical diagnostics. Candidate biomarkers are then predicted based on information transfer criteria across the tissue-biofluid channels. Significant biofluid-tissue relationships can be used to prioritize clinical validation of biomarkers.

Establishing protein-protein interactions

Most proteins function in collaboration with other proteins, and one goal of proteomics is to identify which proteins interact. This is especially useful in determining potential partners in cell signaling cascades.

Several methods are available to probe protein-protein interactions. The traditional method is yeast two-hybrid analysis. New methods include protein microarrays, immunoaffinity chromatography followed by mass spectrometry, dual polarisation interferometry, Microscale Thermophoresis and experimental methods such as phage display and computational methods

Practical applications of proteomics

One of the most promising developments to come from the study of human genes and proteins has been the identification of potential new drugs for the treatment of disease. This relies on genome and proteome information to identify proteins associated with a disease, which computer software can then use as targets for new drugs. For example, if a certain protein is implicated in a disease, its 3D structure provides the information to design drugs to interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, will inactivate the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease. As genetic differences among individuals are found, researchers expect to use these techniques to develop personalized drugs that are more effective for the individual.

Biomarkers

The FDA defines a biomarker as, “A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention”.

Understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked immunosorbent assay (ELISA) or mass spectrometry.

Current research methodologies

There are many approaches to attempting to characterize the human proteome, which is estimated to exceed 100,000 unique forms, 25,000 genes plus post-translational modifications.

In addition, first promising attempts to decipher the proteom of animal tumors have recently been reported.

Chapter 7

Proteinogenic Amino Acid

Proteinogenic amino acids are those amino acids that can be found in proteins and require cellular machinery coded for in the genetic code of any organism for their isolated production. There are 22 standard amino acids, but only 21 are found in eukaryotes. Of the 22, 20 are directly encoded by the universal genetic code. Humans can synthesize 11 of these 20 from each other or from other molecules of intermediary metabolism. The other 9 must be consumed in the diet, and so are called *essential amino acids*; those are histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. The remaining two, selenocysteine and pyrrolysine, are incorporated into proteins by unique synthetic mechanisms.

The word *proteinogenic* means "protein building". Proteinogenic amino acids can be assembled into a polypeptide (the subunit of a protein) through a process called translation (the second stage of protein biosynthesis, part of the overall process of gene expression).

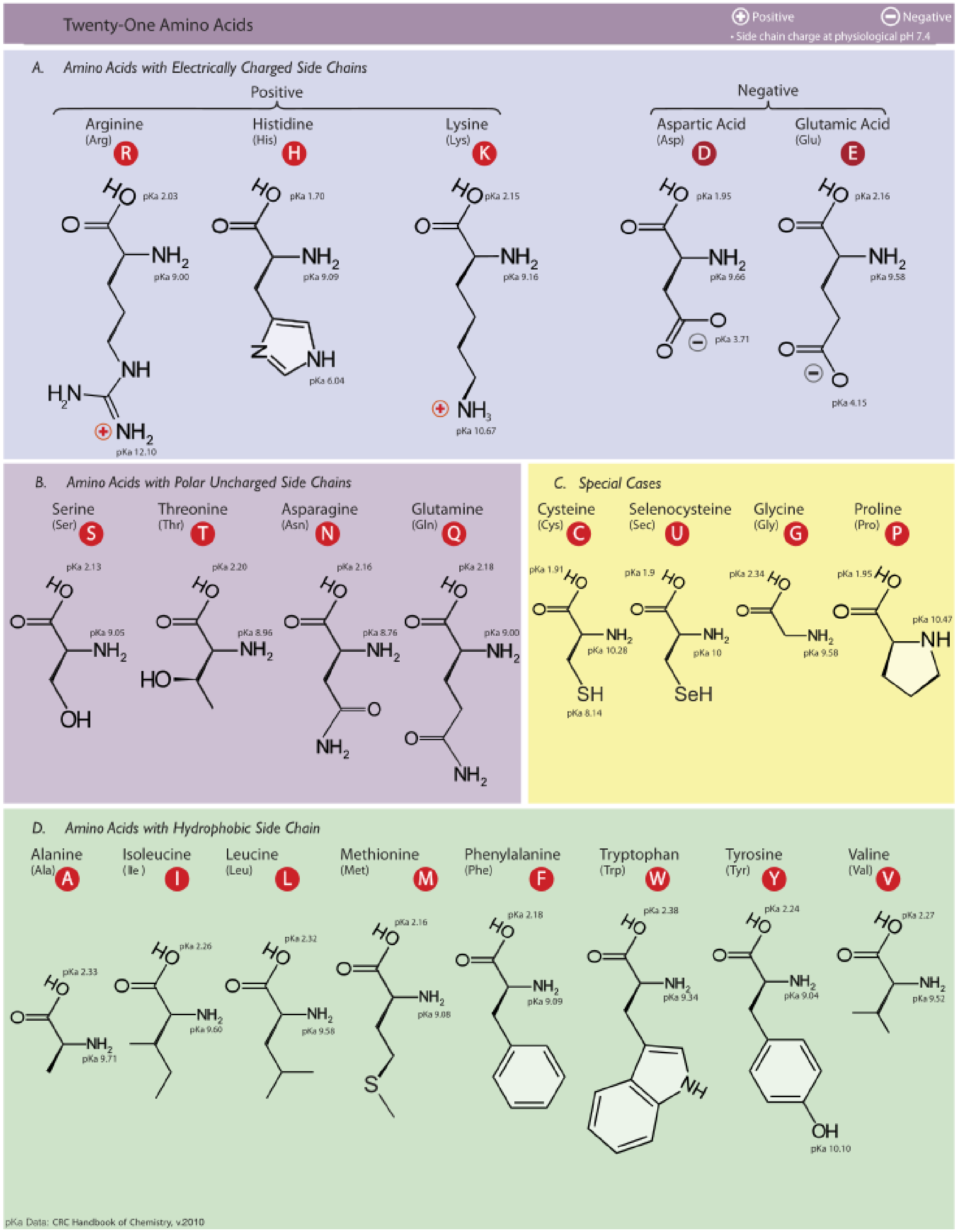
Non-proteinogenic amino acids are either not found in proteins (like carnitine, GABA, or L-DOPA), or are not produced directly and in isolation by standard cellular machinery (like hydroxyproline and selenomethionine). The latter often results from posttranslational modification of proteins.

There are clear reasons why organisms have not evolved to incorporate certain non-proteinogenic amino acids into proteins: for example, ornithine and homoserine cyclize against the peptide backbone and fragment the protein with relatively short half-lives, while others are toxic because they can be mistakenly incorporated into proteins, such as the arginine analog canavanine.

Non-proteinogenic amino acids are found in nonribosomal peptides, which are not produced by the ribosome during translation.

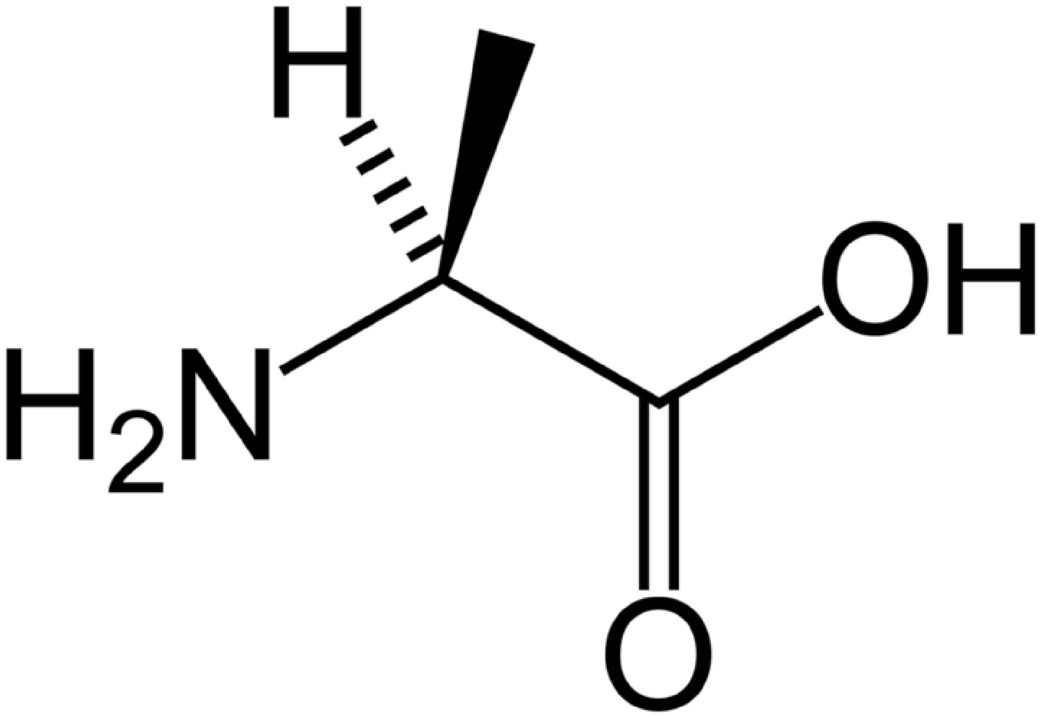
Structures

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

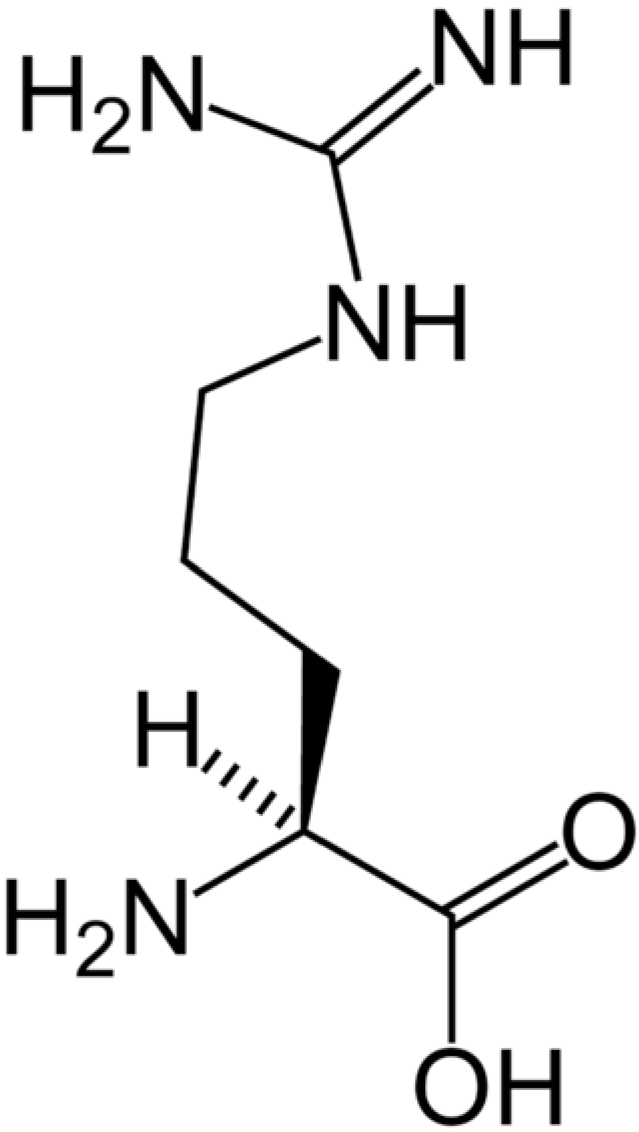


Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010

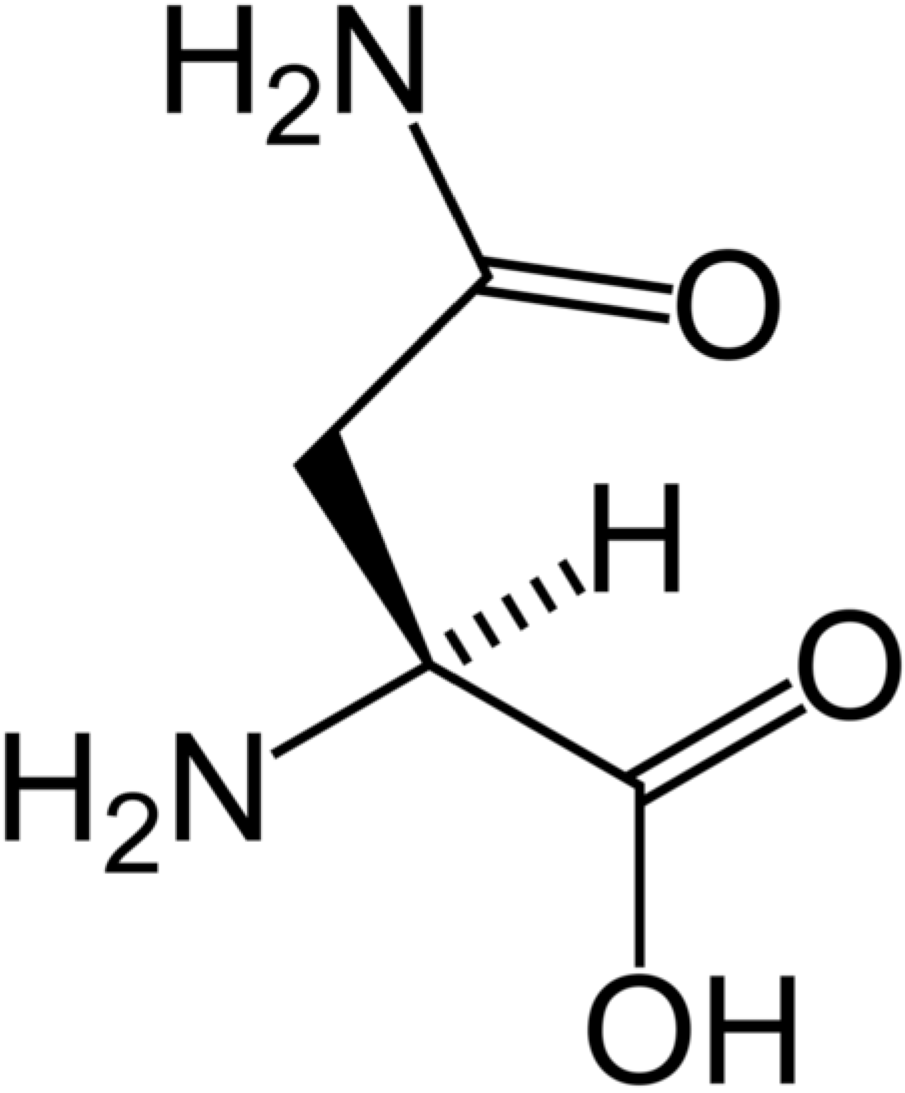
Grouped table of 21 amino acids' structures, nomenclature, and their side groups' pKa's.



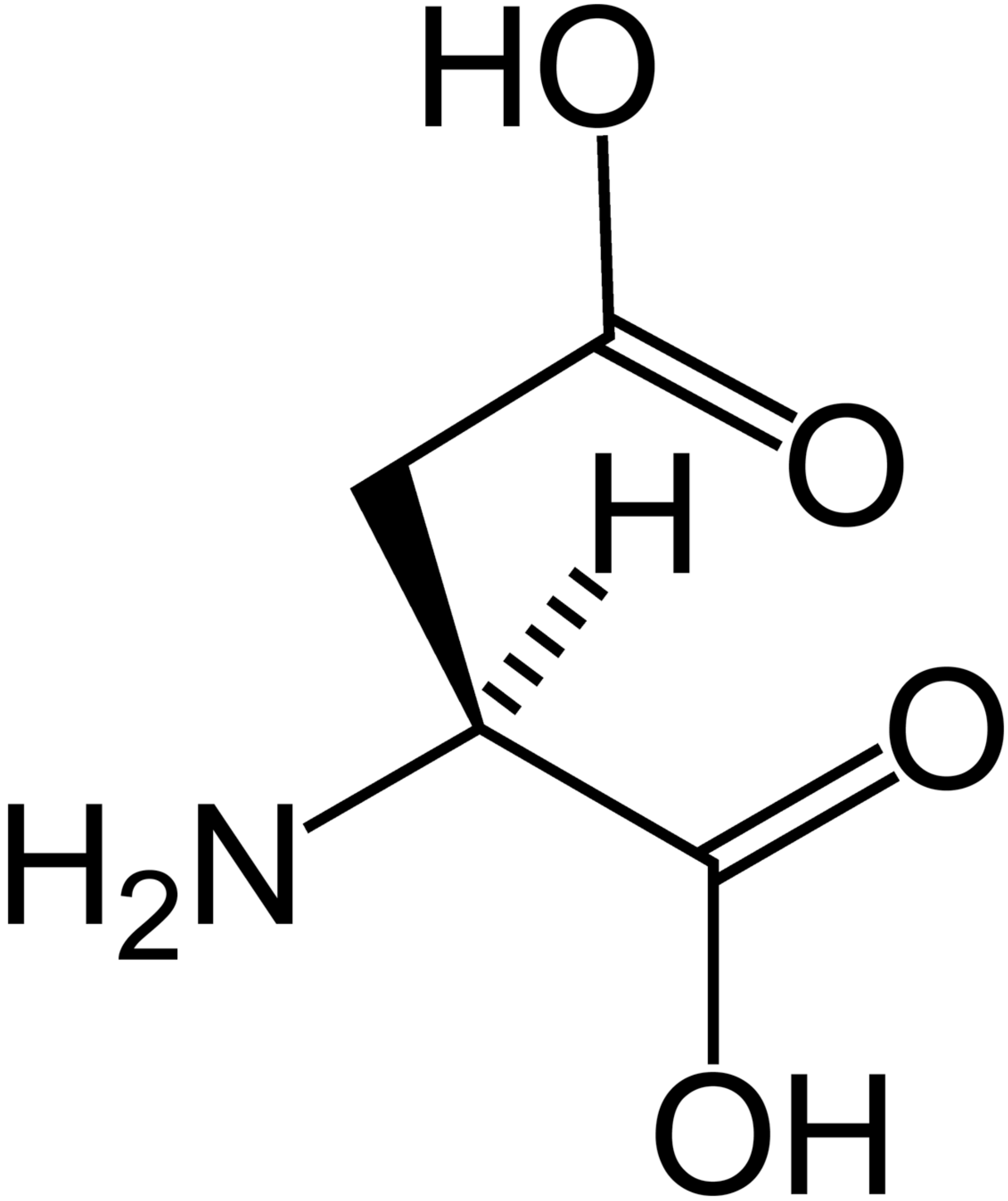
L-Alanine
(Ala / A)



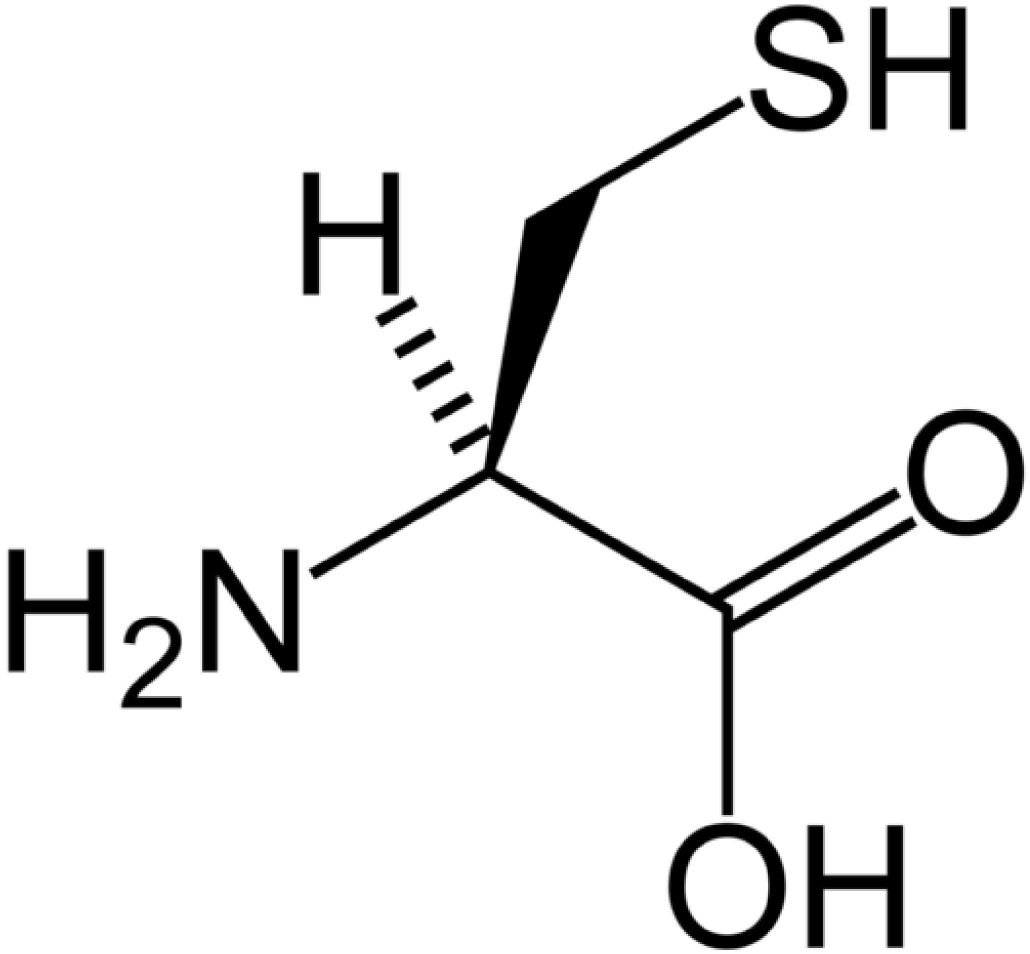
L-Arginine
(Arg / R)



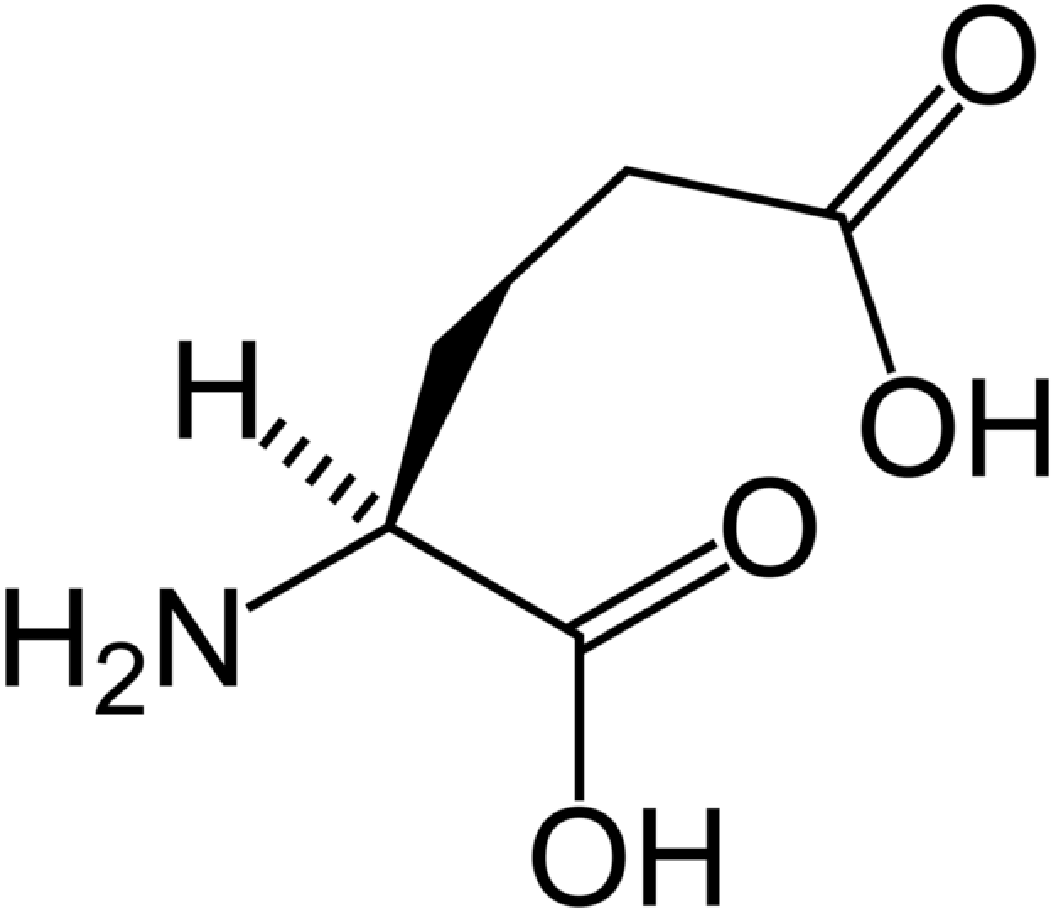
L-Asparagine
(Asn / N)



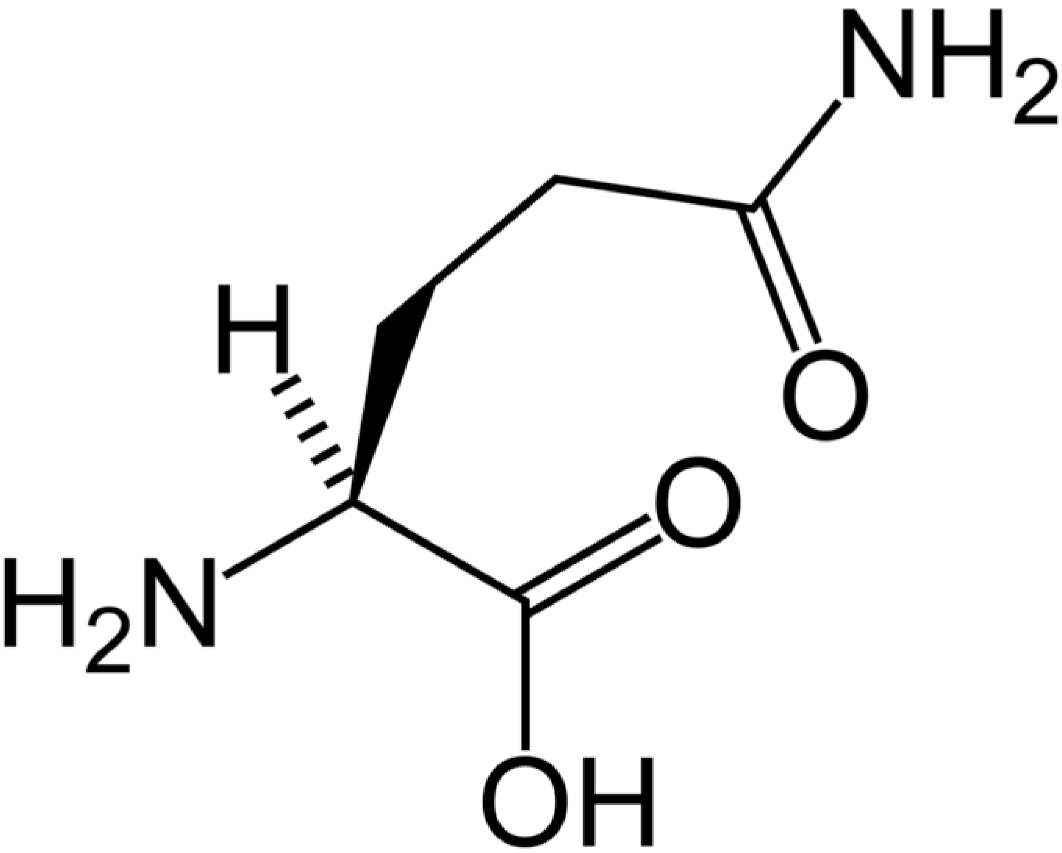
L-Aspartic acid
(Asp / D)



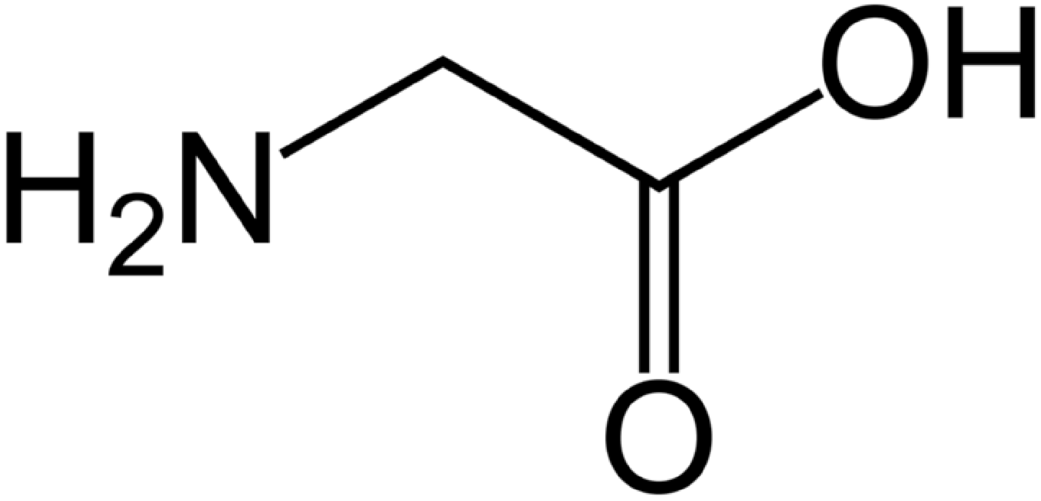
L-Cysteine
(Cys / C)



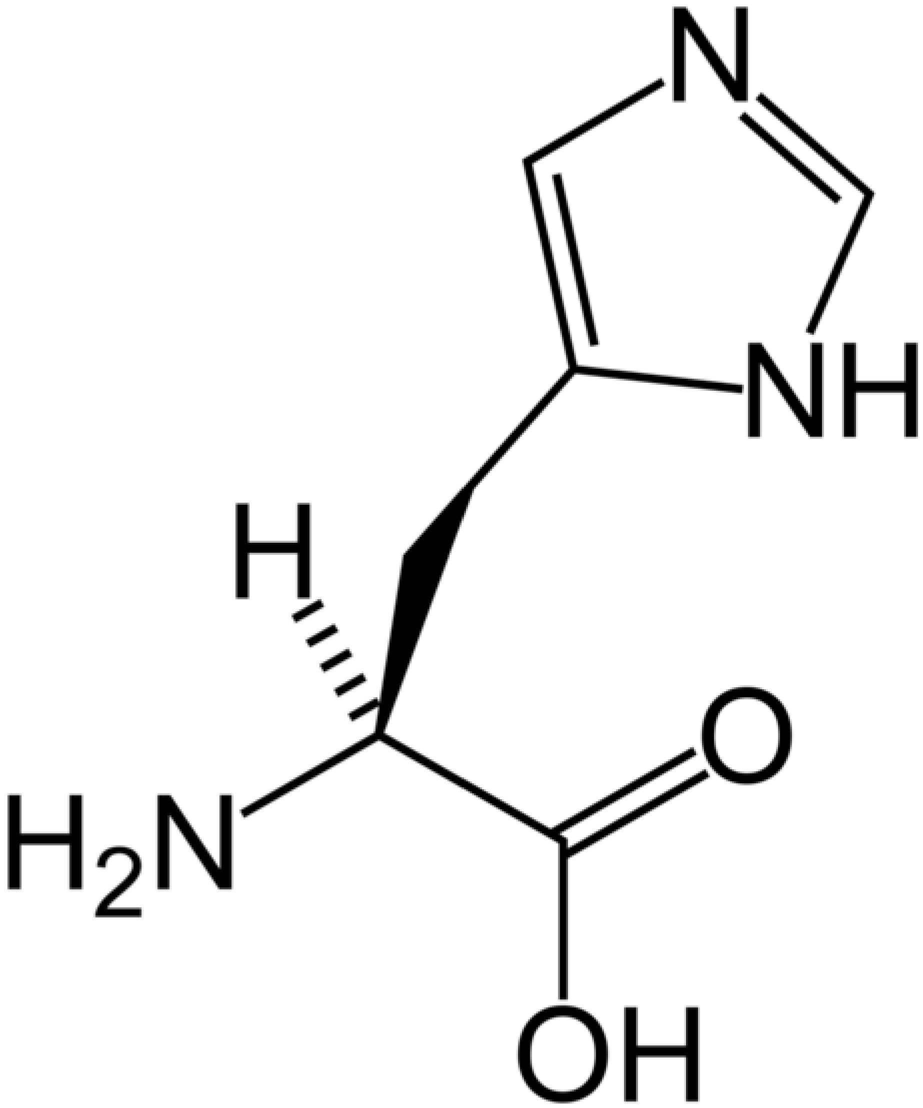
L-Glutamic acid
(Glu / E)



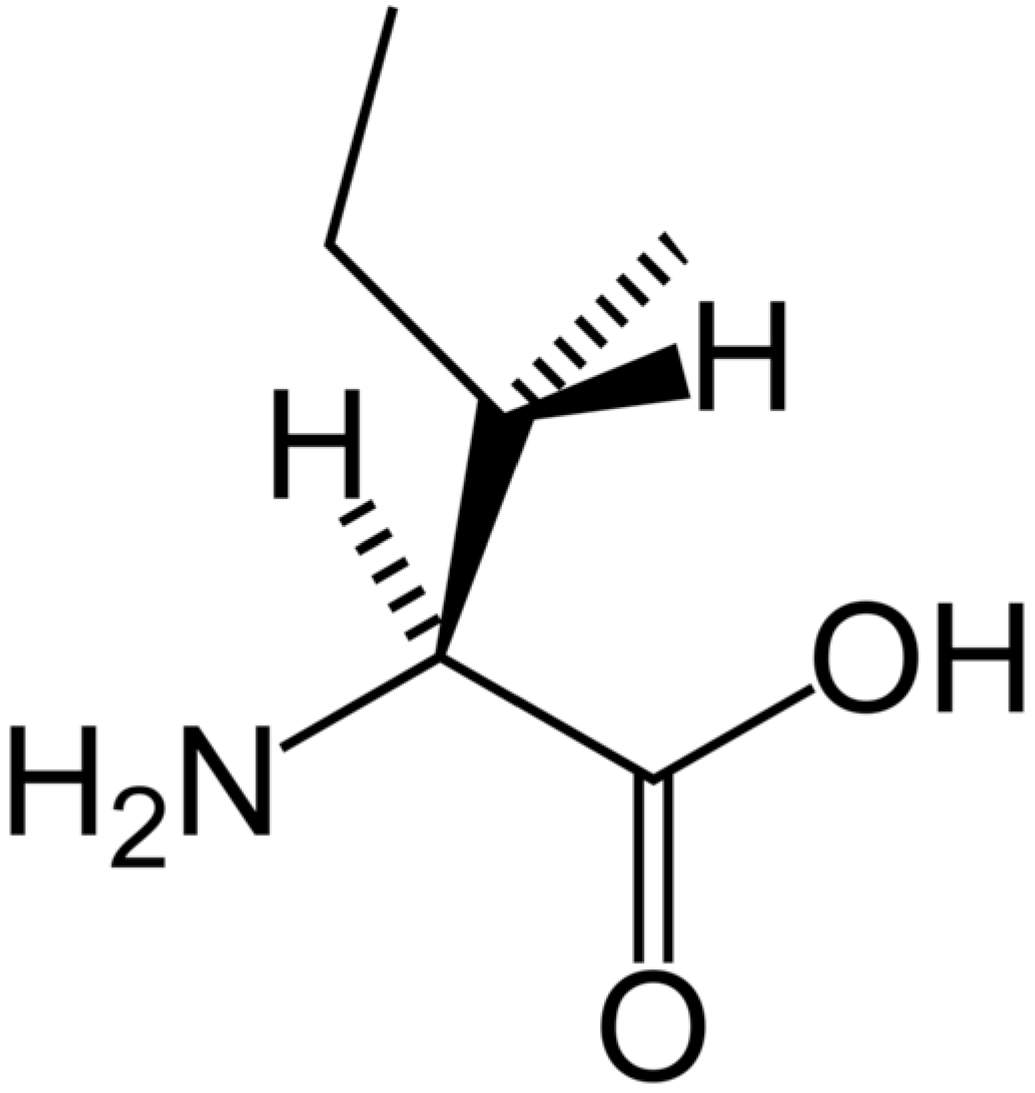
L-Glutamine
(Gln / Q)



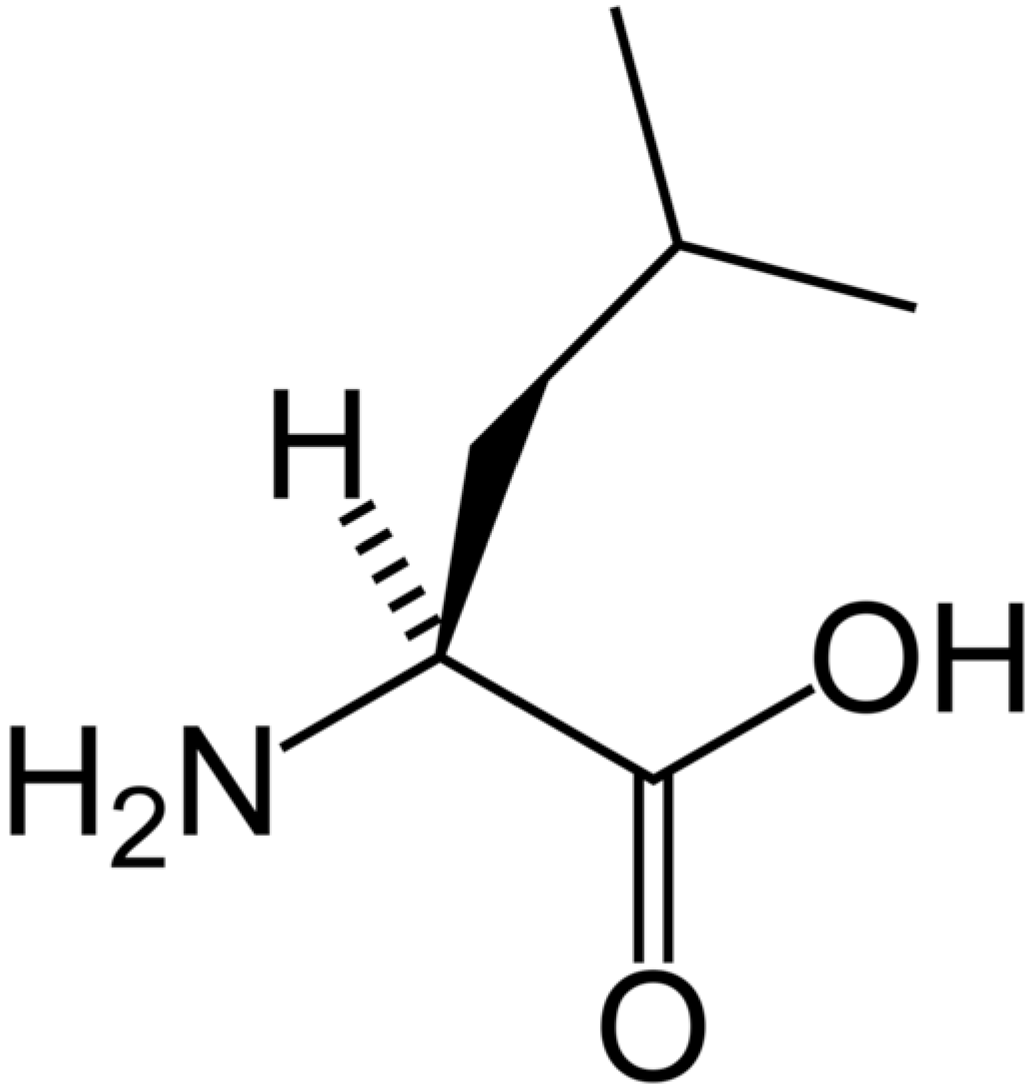
Glycine
(Gly / G)



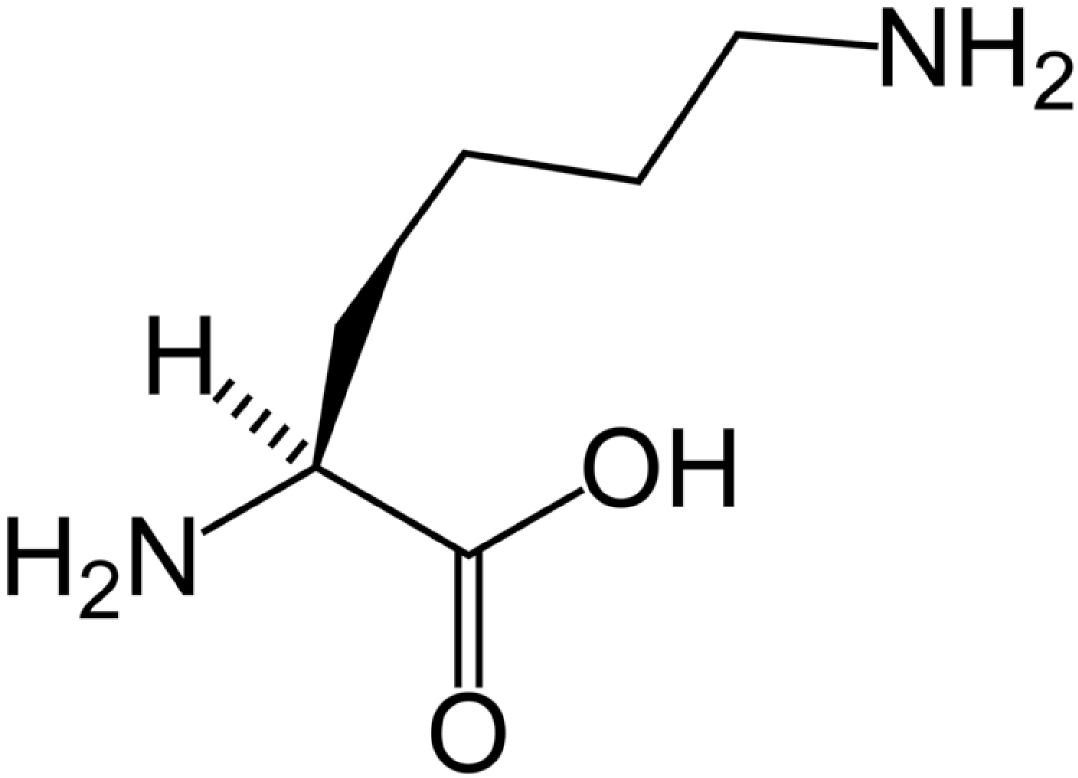
L-Histidine
(His / H)



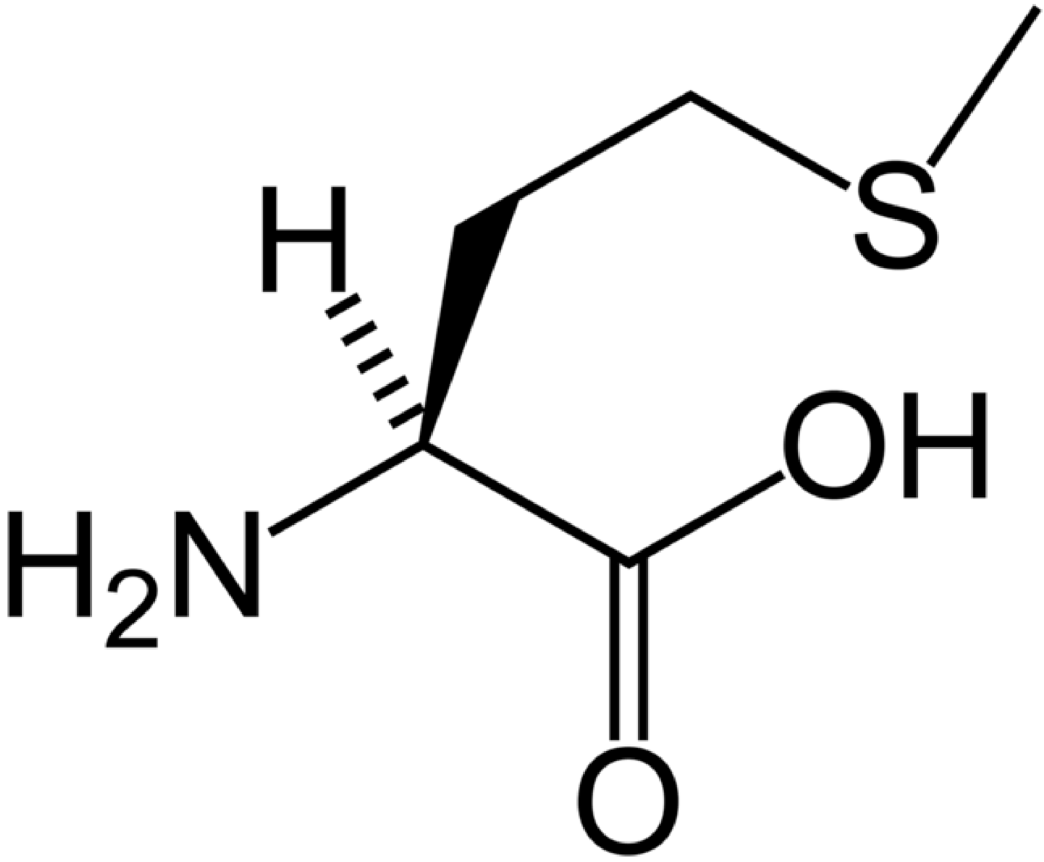
L-Isoleucine
(Ile / I)



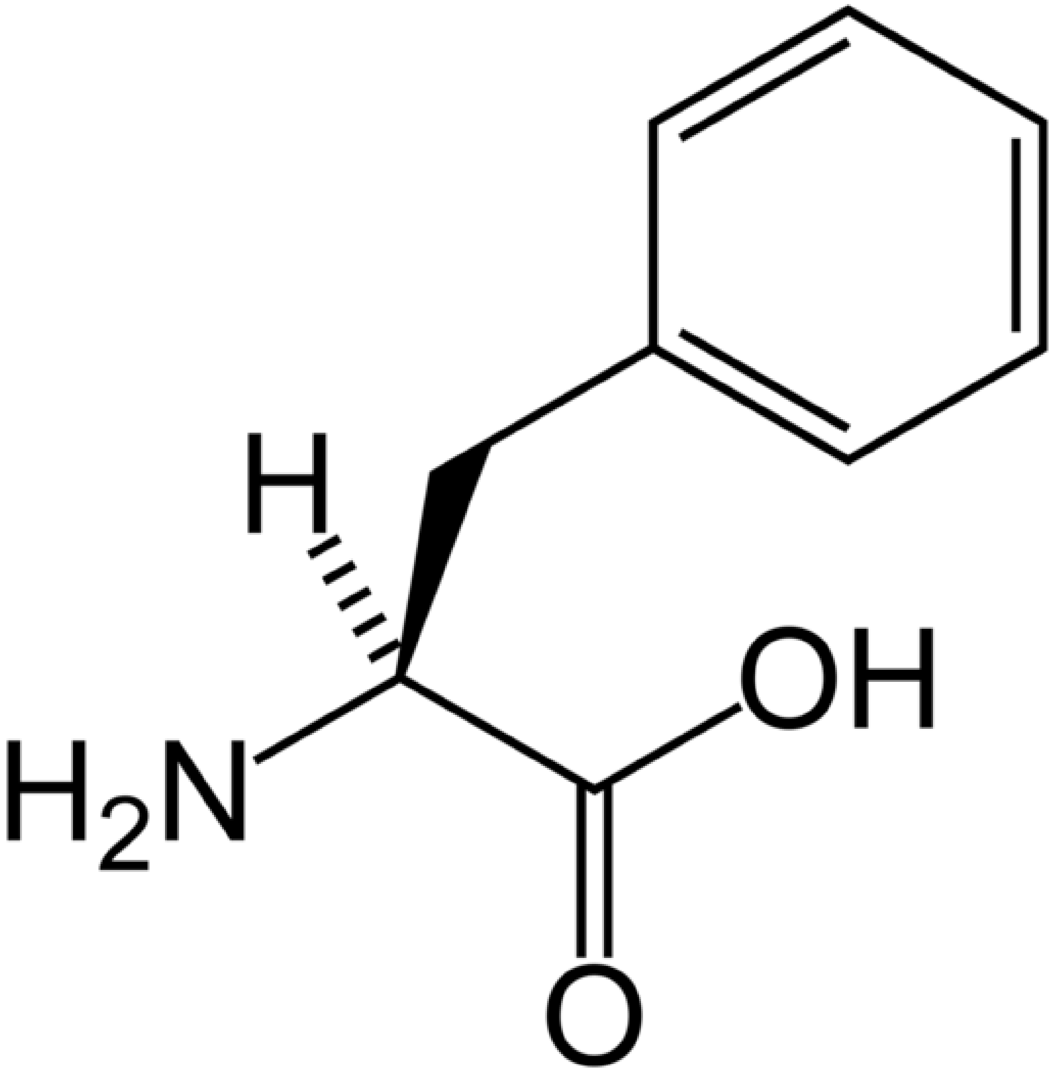
L-Leucine
(Leu / L)



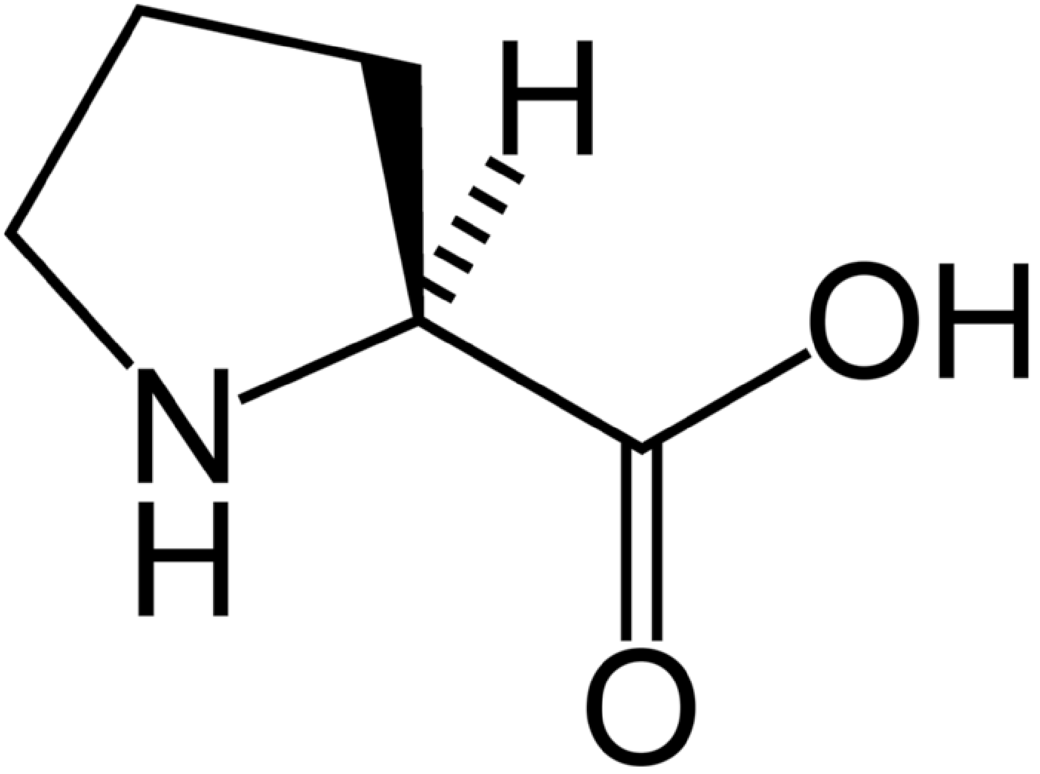
L-Lysine
(Lys / K)



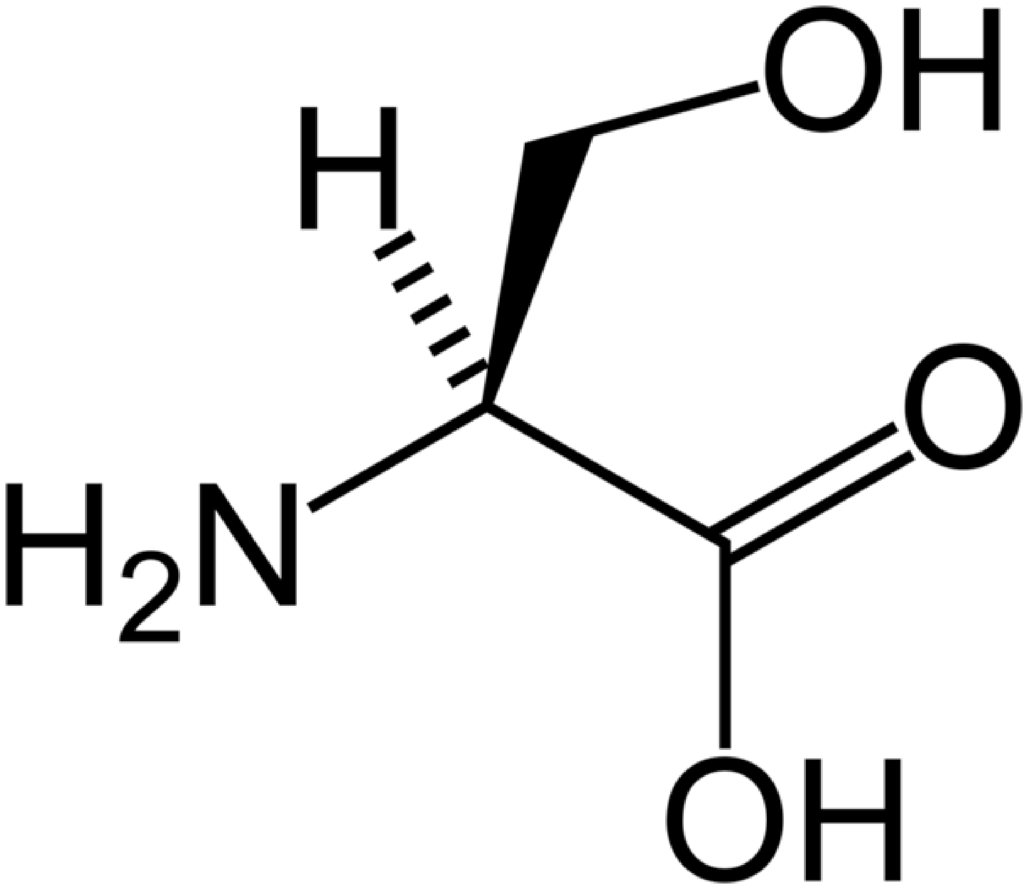
L-Methionine
(Met / M)



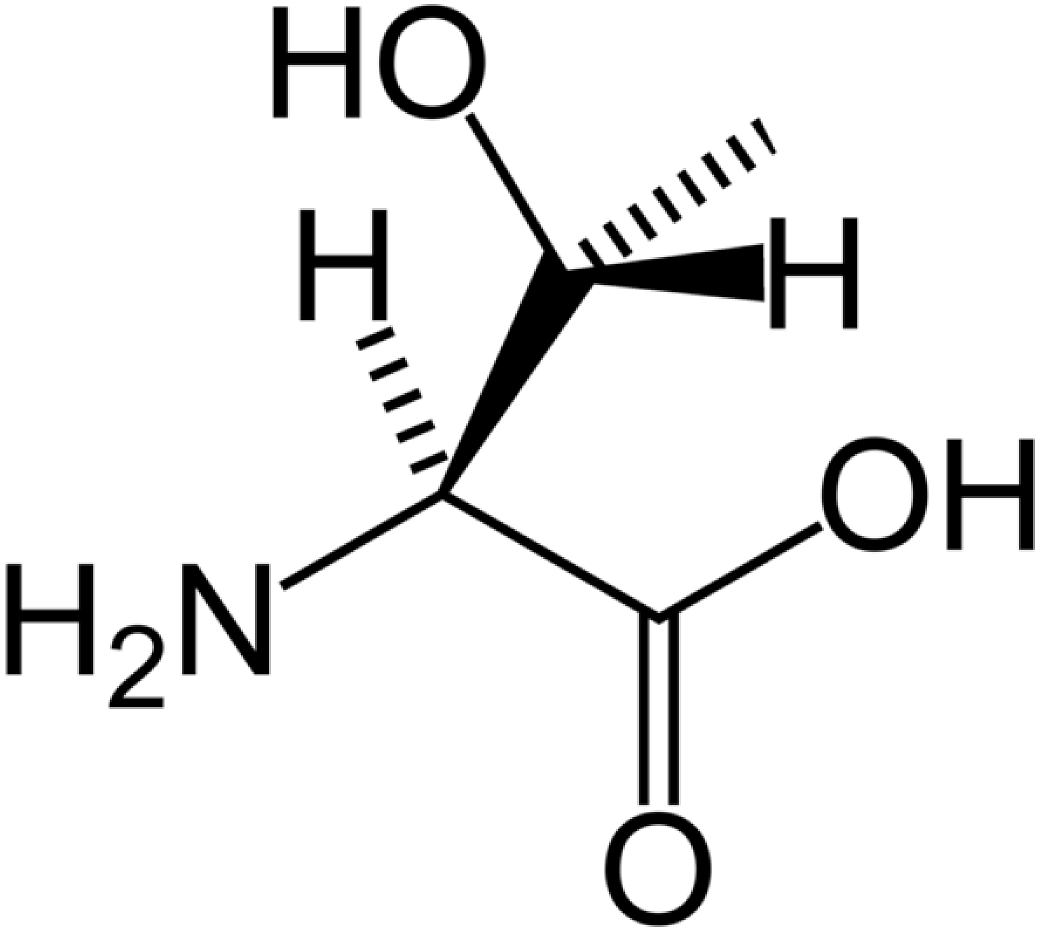
L-Phenylalanine
(Phe / F)



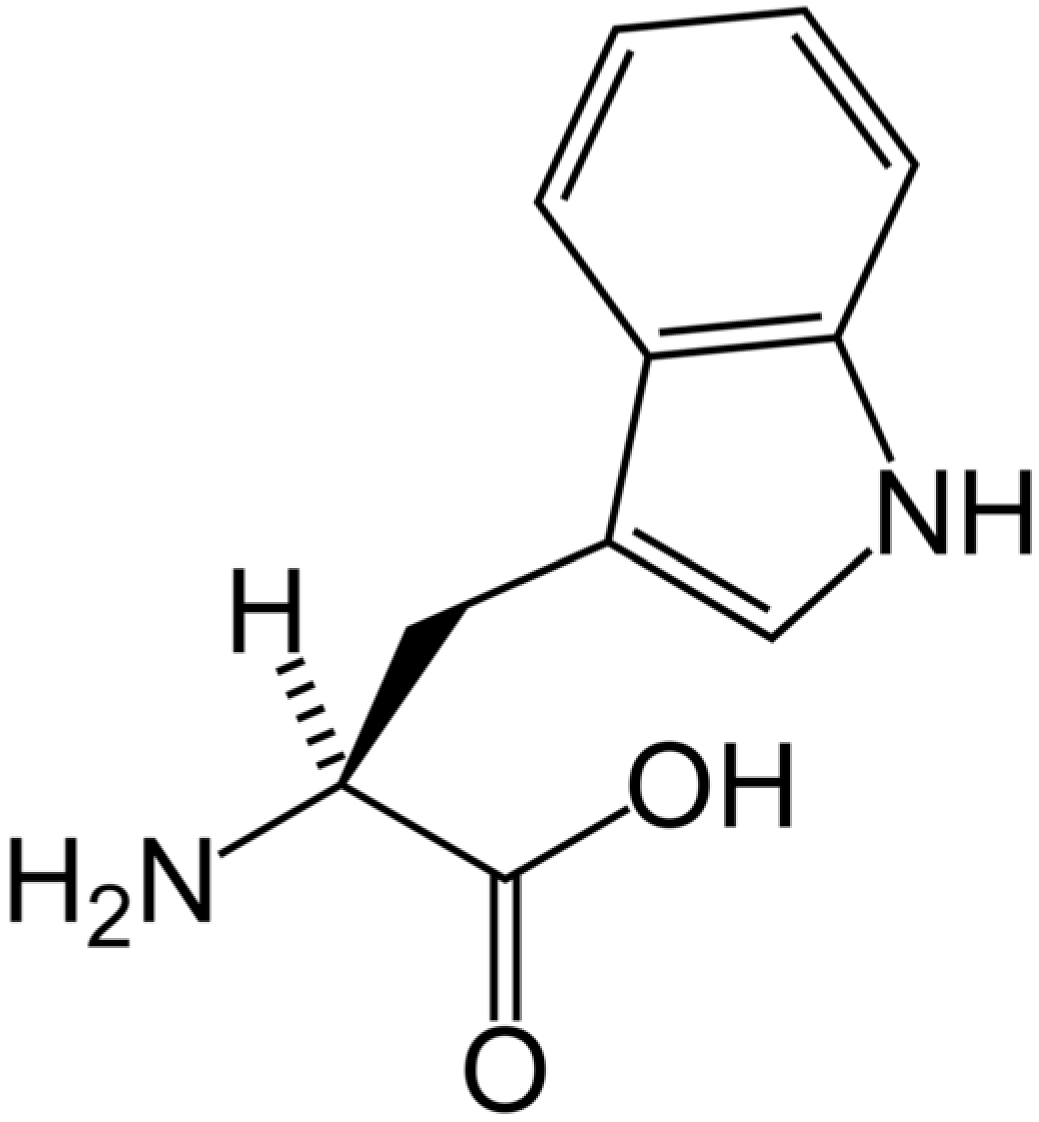
L-Proline
(Pro / P)



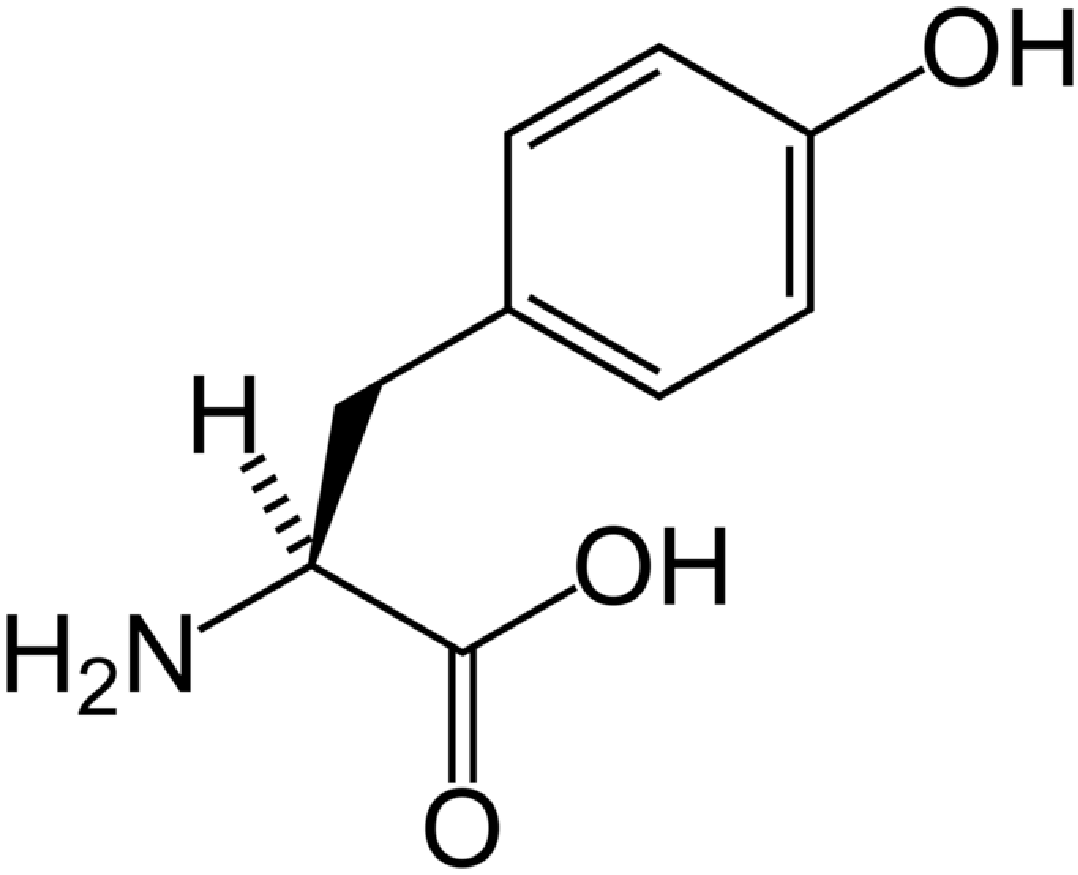
L-Serine
(Ser / S)



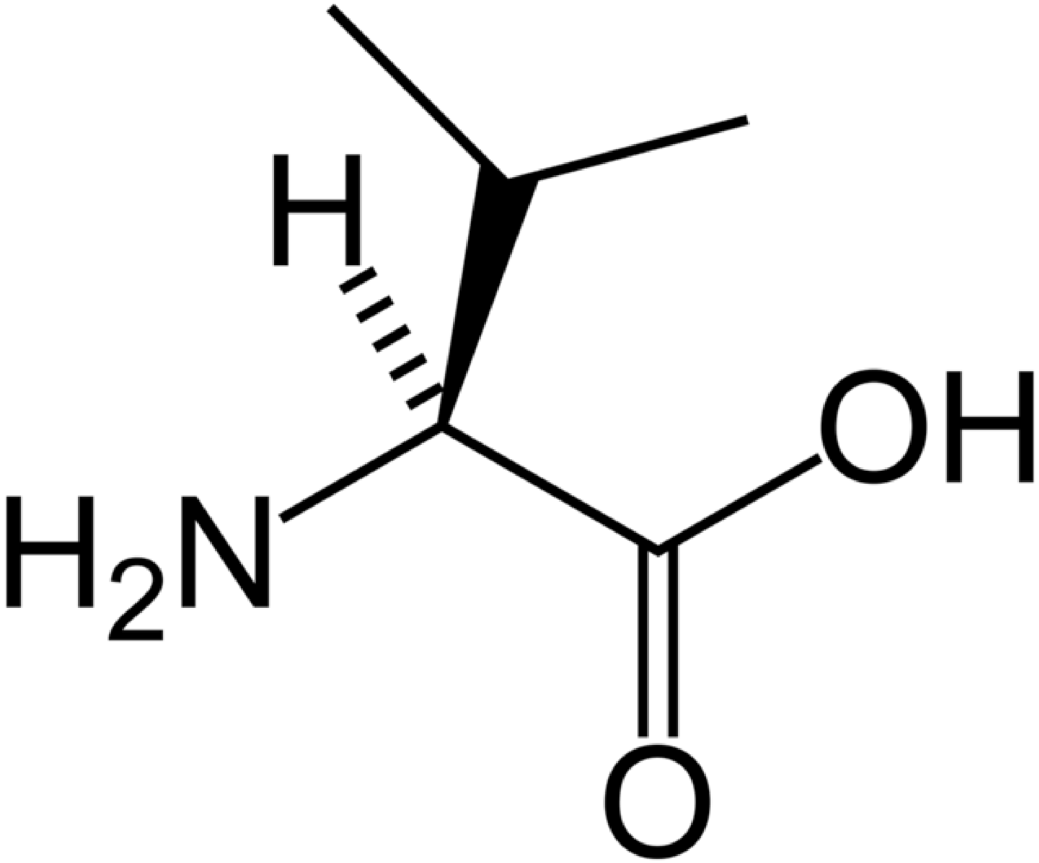
L-Threonine
(Thr / T)



L-Tryptophan
(Trp / W)

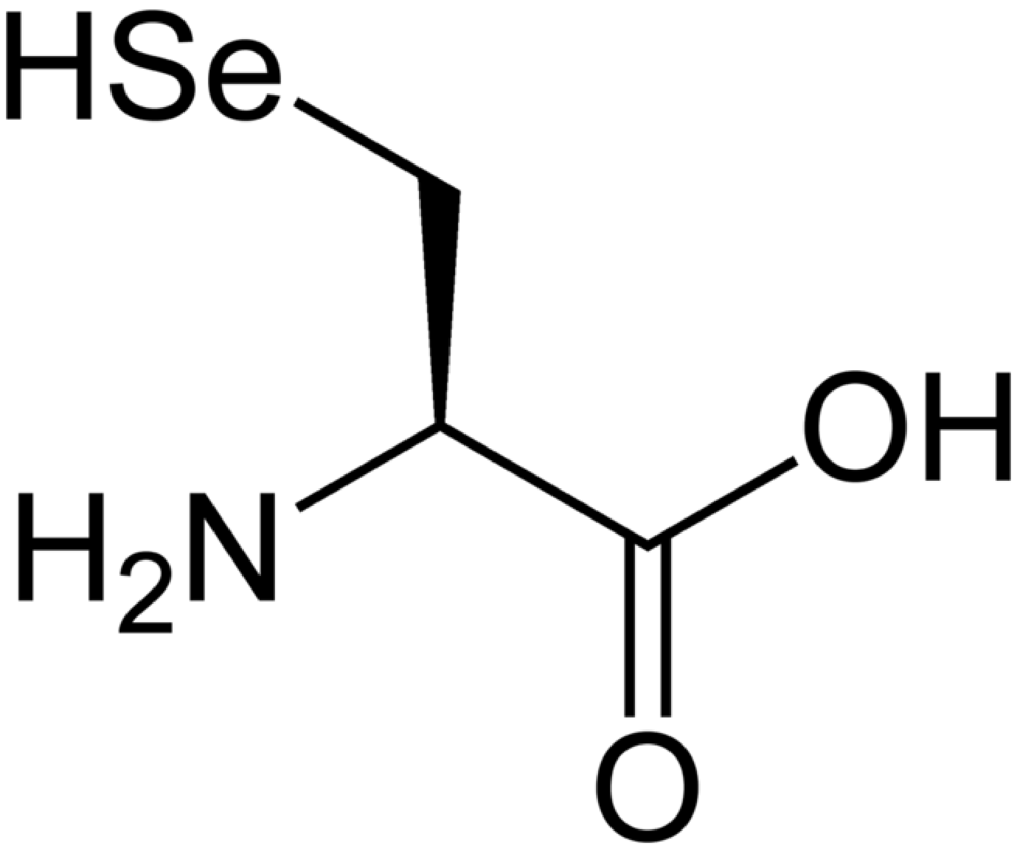


L-Tyrosine
(Tyr / Y)

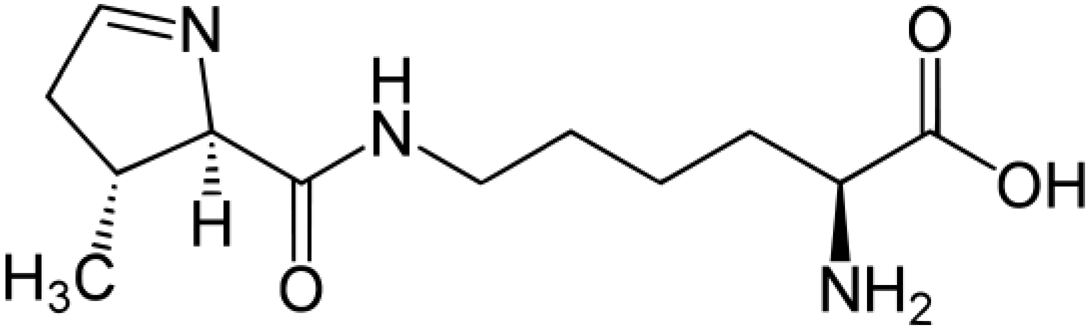


L-Valine
(Val / V)

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



L-Selenocysteine



L-Pyrrolysine
(Pyl / O)

Non-specific abbreviations

Sometimes the specific identity of an amino acid cannot be determined unambiguously. Certain protein sequencing techniques do not distinguish among certain pairs. Thus, the following codes are used:

- *Asx* (*B*) is "asparagine or aspartic acid"
- *Glx* (*Z*) is "glutamic acid or glutamine"
- *Xle* (*J*) is "leucine or isoleucine"

In addition, the symbol *X* is used to indicate an amino acid that is completely unidentified.

Chemical properties

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

General chemical properties

Amino Acid	**Short**	**Abbrev.**	**Avg. Mass (Da)**	**pI**	**pK$_1$ (α-COOH)**	**pK$_2$ (α-$^+$NH$_3$)**
Alanine	A	Ala	89.09404	6.01	2.35	9.87
Cysteine	C	Cys	121.15404	5.05	1.92	10.70
Aspartic acid	D	Asp	133.10384	2.85	1.99	9.90
Glutamic acid	E	Glu	147.13074	3.15	2.10	9.47
Phenylalanine	F	Phe	165.19184	5.49	2.20	9.31
Glycine	G	Gly	75.06714	6.06	2.35	9.78
Histidine	H	His	155.15634	7.60	1.80	9.33
Isoleucine	I	Ile	131.17464	6.05	2.32	9.76
Lysine	K	Lys	146.18934	9.60	2.16	9.06
Leucine	L	Leu	131.17464	6.01	2.33	9.74
Methionine	M	Met	149.20784	5.74	2.13	9.28
Asparagine	N	Asn	132.11904	5.41	2.14	8.72
Pyrrolysine	O	Pyl				
Proline	P	Pro	115.13194	6.30	1.95	10.64
Glutamine	Q	Gln	146.14594	5.65	2.17	9.13
Arginine	R	Arg	174.20274	10.76	1.82	8.99
Serine	S	Ser	105.09344	5.68	2.19	9.21
Threonine	T	Thr	119.12034	5.60	2.09	9.10

Selenocysteine	U	Sec	168.053			
Valine	V	Val	117.14784	6.00	2.39	9.74
Tryptophan	W	Trp	204.22844	5.89	2.46	9.41
Tyrosine	Y	Tyr	181.19124	5.64	2.20	9.21

Side chain properties

Amino Acid	Short	Abbrev.	Side chain	Hydro-phobic	pKa	Polar	pH	Small	Tiny	Aromatic or Aliphatic	van der Waals volume
Alanine	A	Ala	$-CH_3$	X	-	-	-	X	X	-	67
Cysteine	C	Cys	$-CH_2SH$	X	8.18	-	acidic	X	-	-	86
Aspartic acid	D	Asp	$-CH_2COOH$	-	3.90	X	acidic	X	-	-	91
Glutamic acid	E	Glu	$-CH_2CH_2COOH$	-	4.07	X	acidic	-	-	-	109
Phenylalanine	F	Phe	$-CH_2C_6H_5$	X	-	-	-	-	-	Aromatic	135
Glycine	G	Gly	-H	X	-	-	-	X	X	-	48
Histidine	H	His	$-CH_2-C_3H_3N_2$	-	6.04	X	weak basic	-	-	Aromatic	118
Isoleucine	I	Ile	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	-	Aliphatic	124
Lysine	K	Lys	$-(CH_2)_4NH_2$	-	10.54	X	basic	-	-	-	135
Leucine	L	Leu	$-CH_2CH(CH_3)_2$	X	-	-	-	-	-	Aliphatic	124
Methionine	M	Met	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	-	124
Asparagine	N	Asn	$-CH_2CONH_2$	-	-	X	-	X	-	-	96
Pyrrolysine	O	Pyl									
Proline	P	Pro	$-CH_2CH_2CH_2-$	X	-	-	-	X	-	-	90
Glutamine	Q	Gln	$-CH_2CH_2CONH_2$	-	-	X	-	-	-	-	114
Arginine	R	Arg	$-(CH_2)_3NH-C(NH)NH_2$	-	12.48	X	strongly basic	-	-	-	148
Serine	S	Ser	$-CH_2OH$	-	-	X	-	X	X	-	73
Threonine	T	Thr	$-CH(OH)CH_3$	-	-	X	weak acidic	X	-	-	93
Selenocysteine	U	Sec	$-CH_2SeH$	X	5.73	-	-	X	-	-	
Valine	V	Val	$-CH(CH_3)_2$	X	-	-	-	X	-	Aliphatic	105
Tryptophan	W	Trp	$-CH_2C_8H_6N$	X	-	-	-	-	-	Aromatic	163
Tyrosine	Y	Tyr	$-CH_2-C_6H_4OH$	-	10.46	X	-	-	-	Aromatic	141

Note: The pKa values of amino acids are typically slightly different when the amino acid is inside a protein. Protein pKa calculations are sometimes used to calculate the change in the pKa value of an amino acid in this situation.

Gene expression and biochemistry

Amino Acid	Short	Abbrev.	Codon(s)	Occurrence in human proteins (%)	Essential‡ in humans
Alanine	A	Ala	GCU, GCC, GCA, GCG	7.8	-
Cysteine	C	Cys	UGU, UGC	1.9	Conditionally

Aspartic acid	D	Asp	GAU, GAC	5.3	-
Glutamic acid	E	Glu	GAA, GAG	6.3	Conditionally
Phenylalanine	F	Phe	UUU, UUC	3.9	Yes
Glycine	G	Gly	GGU, GGC, GGA, GGG	7.2	Conditionally
Histidine	H	His	CAU, CAC	2.3	Yes
Isoleucine	I	Ile	AUU, AUC, AUA	5.3	Yes
Lysine	K	Lys	AAA, AAG	5.9	Yes
Leucine	L	Leu	UUA, UUG, CUU, CUC, CUA, CUG	9.1	Yes
Methionine	M	Met	AUG	2.3	Yes
Asparagine	N	Asn	AAU, AAC	4.3	-
Pyrrolysine	O	Pyl	UAG*		-
Proline	P	Pro	CCU, CCC, CCA, CCG	5.2	-
Glutamine	Q	Gln	CAA, CAG	4.2	-
Arginine	R	Arg	CGU, CGC, CGA, CGG, AGA, AGG	5.1	Conditionally
Serine	S	Ser	UCU, UCC, UCA, UCG, AGU, AGC	6.8	-
Threonine	T	Thr	ACU, ACC, ACA, ACG	5.9	Yes
Selenocysteine	U	Sec	UGA**		-
Valine	V	Val	GUU, GUC, GUA, GUG	6.6	Yes
Tryptophan	W	Trp	UGG	1.4	Yes
Tyrosine	Y	Tyr	UAU, UAC	3.2	Conditionally
Stop codon†	-	Term	UAA, UAG, UGA	-	-

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

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

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

‡ An essential amino acid cannot be synthesized in humans and must, therefore, be supplied in the diet. Conditionally essential amino acids are not normally required in the diet, but must be supplied exogenously to specific populations that do not synthesize it in adequate amounts.

Mass spectrometry

In mass spectrometry of peptides and proteins, it is useful to know the masses of the residues. The mass of the peptide or protein is the sum of the residue masses plus the mass of water.

Amino Acid	Short	Abbrev.	Formula	Mon. Mass§ (Da)	Avg. Mass (Da)
Alanine	A	Ala	C_3H_5NO	71.03711	71.0788
Cysteine	C	Cys	C_3H_5NOS	103.00919	103.1388
Aspartic acid	D	Asp	$C_4H_5NO_3$	115.02694	115.0886
Glutamic acid	E	Glu	$C_5H_7NO_3$	129.04259	129.1155
Phenylalanine	F	Phe	C_9H_9NO	147.06841	147.1766
Glycine	G	Gly	C_2H_3NO	57.02146	57.0519
Histidine	H	His	$C_6H_7N_3O$	137.05891	137.1411
Isoleucine	I	Ile	$C_6H_{11}NO$	113.08406	113.1594
Lysine	K	Lys	$C_6H_{12}N_2O$	128.09496	128.1741
Leucine	L	Leu	$C_6H_{11}NO$	113.08406	113.1594
Methionine	M	Met	C_5H_9NOS	131.04049	131.1986
Asparagine	N	Asn	$C_4H_6N_2O_2$	114.04293	114.1039
Pyrrolysine	O	Pyl	$C_{12}H_{21}N_3O_3$	255.15829	255.3172
Proline	P	Pro	C_5H_7NO	97.05276	97.1167
Glutamine	Q	Gln	$C_5H_8N_2O_2$	128.05858	128.1307
Arginine	R	Arg	$C_6H_{12}N_4O$	156.10111	156.1875
Serine	S	Ser	$C_3H_5NO_2$	87.03203	87.0782
Threonine	T	Thr	$C_4H_7NO_2$	101.04768	101.1051
Selenocysteine	U	Sec	C_3H_5NOSe	150.95364	150.0388
Valine	V	Val	C_5H_9NO	99.06841	99.1326
Tryptophan	W	Trp	$C_{11}H_{10}N_2O$	186.07931	186.2132
Tyrosine	Y	Tyr	$C_9H_9NO_2$	163.06333	163.1760

§ Monoisotopic mass

Stoichiometry and metabolic cost in cell

Following table lists the abundance of amino acids in E.coli cell and the metabolic cost (ATP) for synthesis the amino acids. Negative numbers indicate the metabolic processes are energy favorable and do not cost net ATP of the cell. Note that the abundance of amino acids include amino acids in free-form and in polymerization form (proteins).

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

Remarks

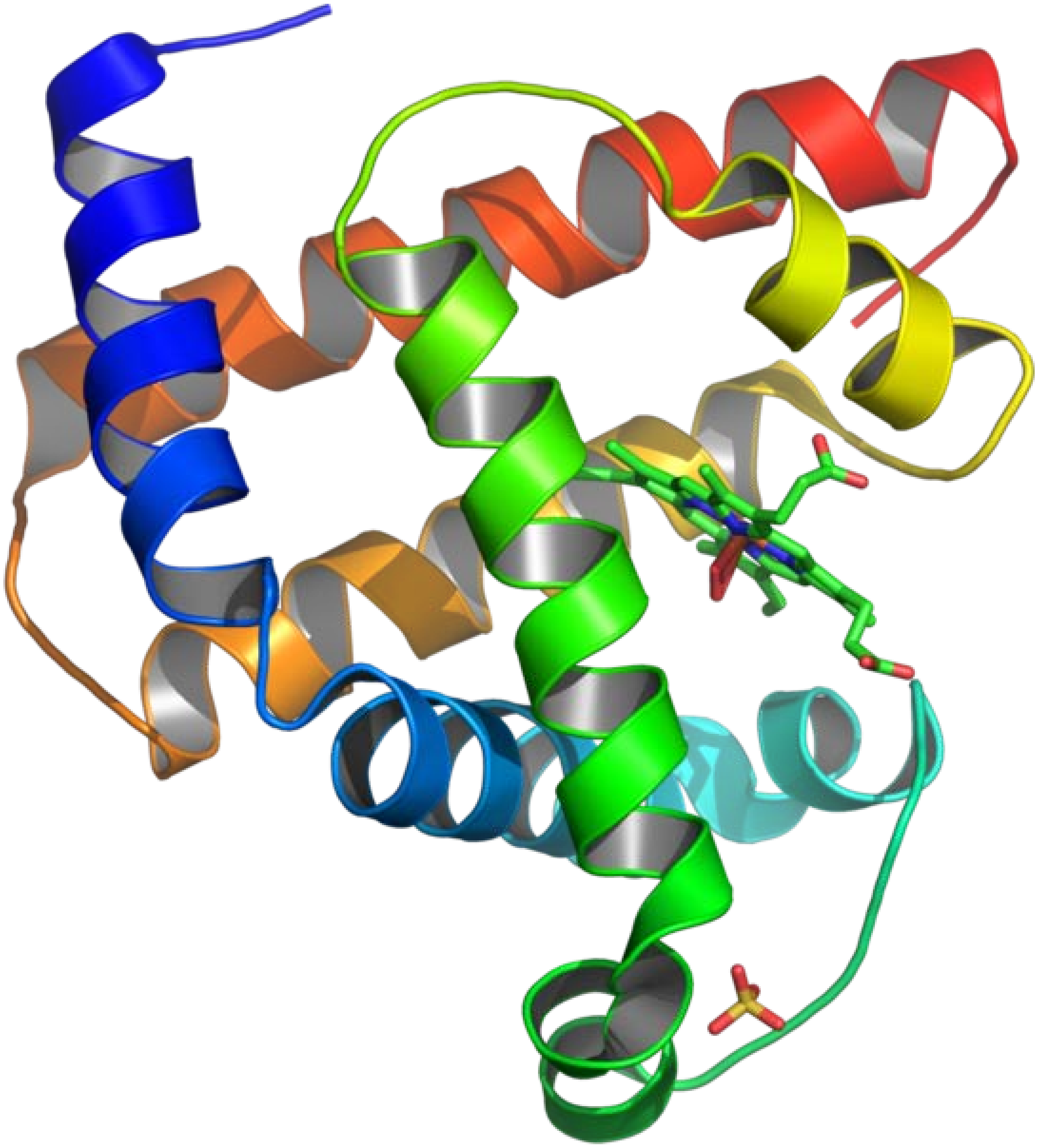
Amino Acid	Abbrev.		Remarks
Alanine	A	Ala	Very abundant, very versatile. More stiff than glycine, but small enough to pose only small steric limits for the protein conformation. It behaves fairly neutrally, and can be located in both hydrophilic regions on the protein outside and the hydrophobic areas inside.
Asparagine or aspartic acid	B	Asx	A placeholder when either amino acid may occupy a position.
Cysteine	C	Cys	The sulfur atom bonds readily to heavy metal ions. Under oxidizing conditions, two cysteines can join together in a disulfide bond to form the amino acid cystine. When cystines are part of a protein, insulin for example, the tertiary structure is stabilized, which makes the protein more resistant to denaturation; therefore, disulfide bonds are common in proteins that have to function in harsh environments including digestive enzymes (e.g., pepsin and chymotrypsin) and structural proteins (e.g., keratin). Disulfides are also found in

			peptides too small to hold a stable shape on their own (eg. insulin).
Aspartic acid	D	Asp	Behaves similarly to glutamic acid. Carries a hydrophilic acidic group with strong negative charge. Usually is located on the outer surface of the protein, making it water-soluble. Binds to positively-charged molecules and ions, often used in enzymes to fix the metal ion. When located inside of the protein, aspartate and glutamate are usually paired with arginine and lysine.
Glutamic acid	E	Glu	Behaves similar to aspartic acid. Has longer, slightly more flexible side chain.
Phenylalanine	F	Phe	Essential for humans. Phenylalanine, tyrosine, and tryptophan contain large rigid aromatic group on the side-chain. These are the biggest amino acids. Like isoleucine, leucine and valine, these are hydrophobic and tend to orient towards the interior of the folded protein molecule. Phenylalanine can be converted into Tyrosine.
Glycine	G	Gly	Because of the two hydrogen atoms at the α carbon, glycine is not optically active. It is the smallest amino acid, rotates easily, adds flexibility to the protein chain. It is able to fit into the tightest spaces, e.g., the triple helix of collagen. As too much flexibility is usually not desired, as a structural component it is less common than alanine.
Histidine	H	His	In even slightly acidic conditions protonation of the nitrogen occurs, changing the properties of histidine and the polypeptide as a whole. It is used by many proteins as a regulatory mechanism, changing the conformation and behavior of the polypeptide in acidic regions such as the late endosome or lysosome, enforcing conformation change in enzymes. However only a few histidines are needed for this, so it is comparatively scarce.
Isoleucine	I	Ile	Essential for humans. Isoleucine, leucine and valine have large aliphatic hydrophobic side chains. Their molecules are rigid, and their mutual hydrophobic interactions are important for the correct folding of proteins, as these chains tend to be located inside of the protein molecule.
Leucine or isoleucine	J	Xle	A placeholder when either amino acid may occupy a position
Lysine	K	Lys	Essential for humans. Behaves similarly to arginine. Contains a long flexible side-chain with a positively-charged end. The flexibility of the chain makes lysine and arginine suitable for binding to molecules with many negative charges on their surfaces. E.g., DNA-binding proteins have their active regions rich with arginine and lysine. The strong charge makes these two amino acids prone to be located on the outer hydrophilic surfaces of the proteins; when they are found inside, they are

			usually paired with a corresponding negatively-charged amino acid, e.g., aspartate or glutamate.
Leucine	L	Leu	Essential for humans. Behaves similar to isoleucine and valine.
Methionine	M	Met	Essential for humans. Always the first amino acid to be incorporated into a protein; sometimes removed after translation. Like cysteine, contains sulfur, but with a methyl group instead of hydrogen. This methyl group can be activated, and is used in many reactions where a new carbon atom is being added to another molecule.
Asparagine	N	Asn	Similar to aspartic acid. Asn contains an amide group where Asp has a carboxyl.
Pyrrolysine	O	Pyl	Similar to lysine, with a pyrroline ring attached.
Proline	P	Pro	Contains an unusual ring to the N-end amine group, which forces the CO-NH amide sequence into a fixed conformation. Can disrupt protein folding structures like α helix or β sheet, forcing the desired kink in the protein chain. Common in collagen, where it often undergoes a posttranslational modification to hydroxyproline.
Glutamine	Q	Gln	Similar to glutamic acid. Gln contains an amide group where Glu has a carboxyl. Used in proteins and as a storage for ammonia. The most abundant Amino Acid in the body.
Arginine	R	Arg	Functionally similar to lysine.
Serine	S	Ser	Serine and threonine have a short group ended with a hydroxyl group. Its hydrogen is easy to remove, so serine and threonine often act as hydrogen donors in enzymes. Both are very hydrophilic, therefore the outer regions of soluble proteins tend to be rich with them.
Threonine	T	Thr	Essential for humans. Behaves similarly to serine.
Selenocysteine	U	Sec	Selenated form of cysteine, which replaces sulfur.
Valine	V	Val	Essential for humans. Behaves similarly to isoleucine and leucine.
Tryptophan	W	Trp	Essential for humans. Behaves similarly to phenylalanine and tyrosine. Precursor of serotonin. Naturally fluorescent.
Unknown	X	Xaa	Placeholder when the amino acid is unknown or unimportant.
Tyrosine	Y	Tyr	Behaves similarly to phenylalanine (precursor to Tyrosine) and tryptophan. Precursor of melanin, epinephrine, and thyroid hormones. Naturally fluorescent, although fluorescence is usually quenched by energy transfer to tryptophans.
Glutamic acid or glutamine	Z	Glx	A placeholder when either amino acid may occupy a position.

Chapter 8

Protein



A representation of the 3D structure of the protein myoglobin showing colored alpha helices. This protein was the first to have its structure solved by X-ray crystallography.

Towards the right-center among the coils, a prosthetic group called a heme group is shown colored largely in green.

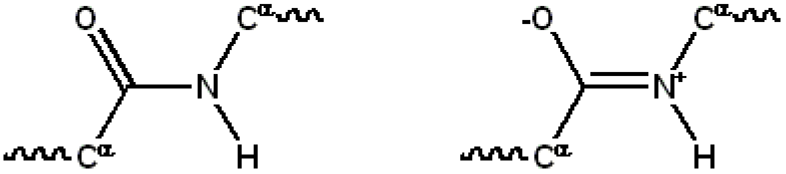
Proteins are biochemical compounds consisting of one or more polypeptides typically folded into a globular or fibrous form in a biologically functional way. A polypeptide is a single linear polymer chain of amino acids bonded together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. In general, the genetic code specifies 20 standard amino acids; however, in certain organisms the genetic code can include selenocysteine—and in certain archaea—pyrrolysine. Shortly after or even during synthesis, the residues in a protein are often chemically modified by post-translational modification, which alters the physical and chemical properties, folding, stability, activity, and ultimately, the function of the proteins. Sometimes proteins have non-peptide groups attached, which can be called prosthetic groups or cofactors. Proteins can also work together to achieve a particular function, and they often associate to form stable complexes.

One of the most distinguishing features of polypeptides is their ability to fold into a globular state, or "structure". The extent to which proteins fold into a defined structure varies widely. Some proteins fold into a highly rigid structure with small fluctuations and are therefore considered to be single structure. Other proteins undergo large rearrangements from one conformation to another. This conformational change is often associated with a signaling event. Thus, the structure of a protein serves as a medium through which to regulate either the function of a protein or activity of an enzyme. Not all proteins requiring a folding process in order to function, as some function in an unfolded state.

Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism.

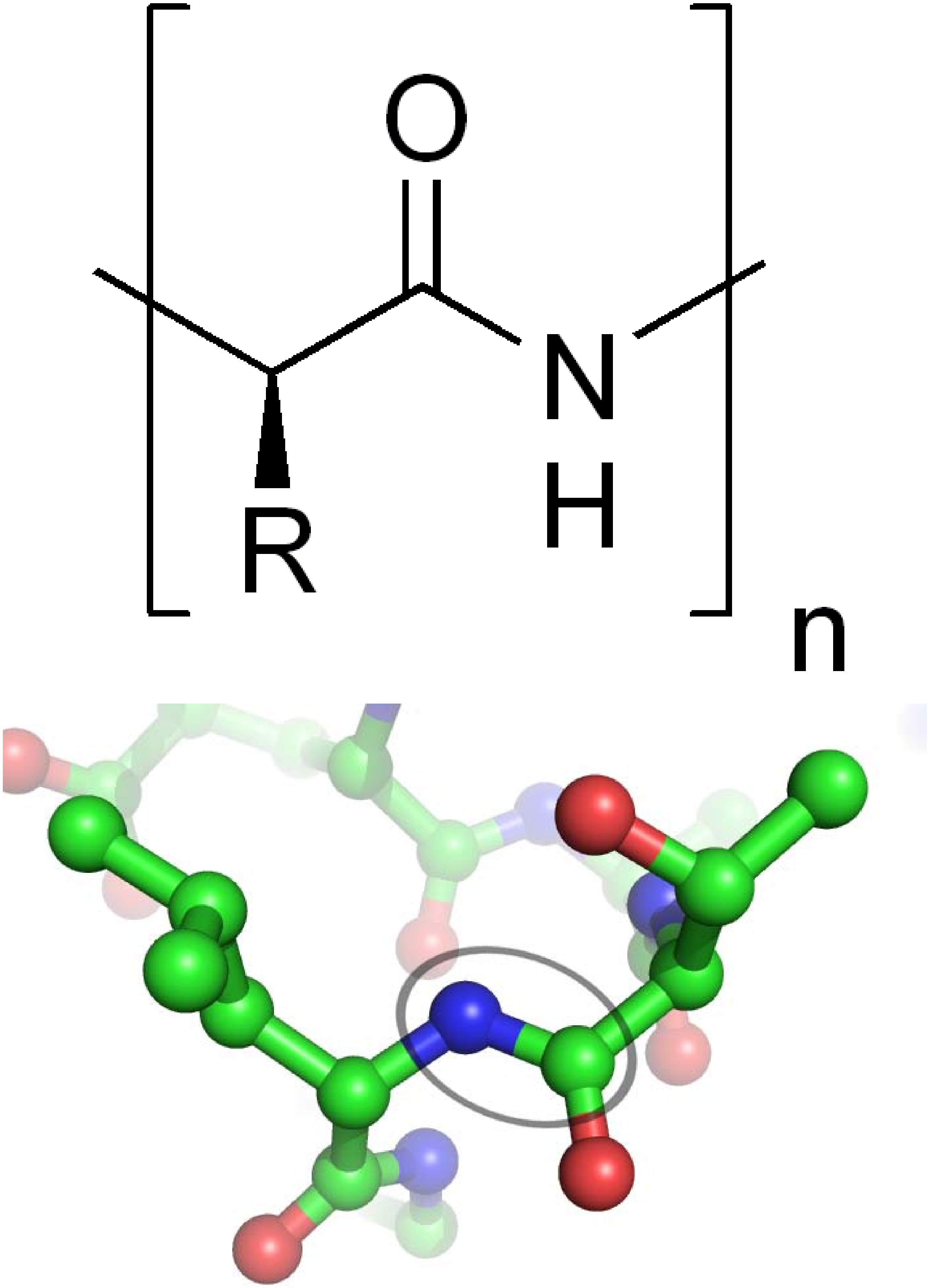
Proteins were first described by the Dutch chemist Gerhardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838. Early nutritional scientists such as the German Carl von Voit believed that protein was the most important nutrient for maintaining the structure of the body, because it was generally believed that "flesh makes flesh." The central role of proteins as enzymes in living organisms was however not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was in fact a protein. The first protein to be sequenced was insulin, by Frederick Sanger, who won the Nobel Prize for this achievement in 1958. The first protein structures to be solved were hemoglobin and myoglobin, by Max Perutz and Sir John Cowdery Kendrew, respectively, in 1958. The three-dimensional structures of both proteins were first determined by X-ray diffraction analysis; Perutz and Kendrew shared the 1962 Nobel Prize in Chemistry for these discoveries. Proteins may be purified from other cellular components using a variety of techniques such as ultracentrifugation, precipitation, electrophoresis, and chromatography; the advent of genetic engineering has made possible a number of methods to facilitate purification. Methods commonly used to study protein structure and function include immunohistochemistry, site-directed mutagenesis, nuclear magnetic resonance and mass spectrometry. Distributed computing is a relatively new tool researchers are using to examine the infamously complex interactions that govern protein folding; the statistical analysis techniques employed to calculate a protein's probable tertiary structure from its amino acid sequence (primary structure) are well-suited for the distributed computing environment, which has made this otherwise prohibitively expensive and time consuming problem significantly more manageable.

Biochemistry



Resonance structures of the peptide bond that links individual amino acids to form a protein polymer

Most proteins consist of linear polymers built from series of up to 20 different L-α-amino acids. All proteinogenic amino acids possess common structural features, including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. Only proline differs from this basic structure as it contains an unusual ring to the N-end amine group, which forces the CO–NH amide moiety into a fixed conformation. The side chains of the standard amino acids, detailed in the list of standard amino acids, have a great variety of chemical structures and properties; it is the combined effect of all of the amino acid side chains in a protein that ultimately determines its three-dimensional structure and its chemical reactivity.

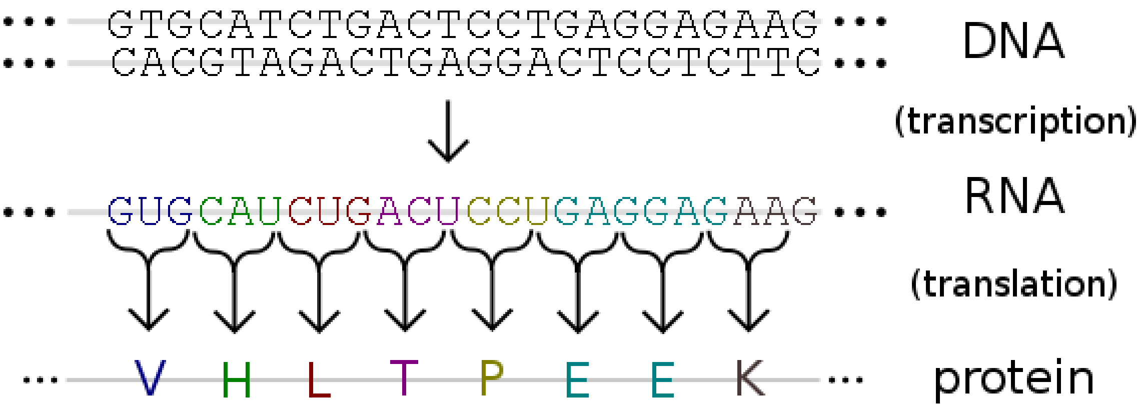


Chemical structure of the peptide bond (left) and a peptide bond between leucine and threonine (right)

The amino acids in a polypeptide chain are linked by peptide bonds. Once linked in the protein chain, an individual amino acid is called a *residue,* and the linked series of carbon, nitrogen, and oxygen atoms are known as the *main chain* or *protein backbone.* The peptide bond has two resonance forms that contribute some double-bond character and inhibit rotation around its axis, so that the alpha carbons are roughly coplanar. The other two dihedral angles in the peptide bond determine the local shape assumed by the protein backbone. The end of the protein with a free carboxyl group is known as the C-terminus or carboxy terminus, whereas the end with a free amino group is known as the N-terminus or amino terminus.

The words *protein*, *polypeptide,* and *peptide* are a little ambiguous and can overlap in meaning. *Protein* is generally used to refer to the complete biological molecule in a stable conformation, whereas *peptide* is generally reserved for a short amino acid oligomers often lacking a stable three-dimensional structure. However, the boundary between the two is not well defined and usually lies near 20–30 residues. *Polypeptide* can refer to any single linear chain of amino acids, usually regardless of length, but often implies an absence of a defined conformation.

Synthesis



••• GTGCATCTGACTCCTGAGGAGAAG •••
••• CACGTAGACTGAGGACTCCTCTTC •••
DNA
(transcription)
••• GUGCAUCUGACUCCUGAGGAGAAG •••
RNA
(translation)
··· V H L T P E E K ···
protein

The DNA sequence of a gene encodes the amino acid sequence of a protein.

Proteins are assembled from amino acids using information encoded in genes. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination designates an amino acid, for example AUG (adenine-uracil-guanine) is the code for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code, with some amino acids specified by more than one codon. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a *primary transcript*) using various forms of post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. In prokaryotes the mRNA may either be used as soon as it is produced, or be bound by a ribosome after having moved away from the nucleoid. In contrast, eukaryotes make mRNA in the cell nucleus and then translocate it across the nuclear membrane into the cytoplasm, where protein synthesis then takes place. The rate of protein synthesis is higher in prokaryotes than eukaryotes and can reach up to 20 amino acids per second.

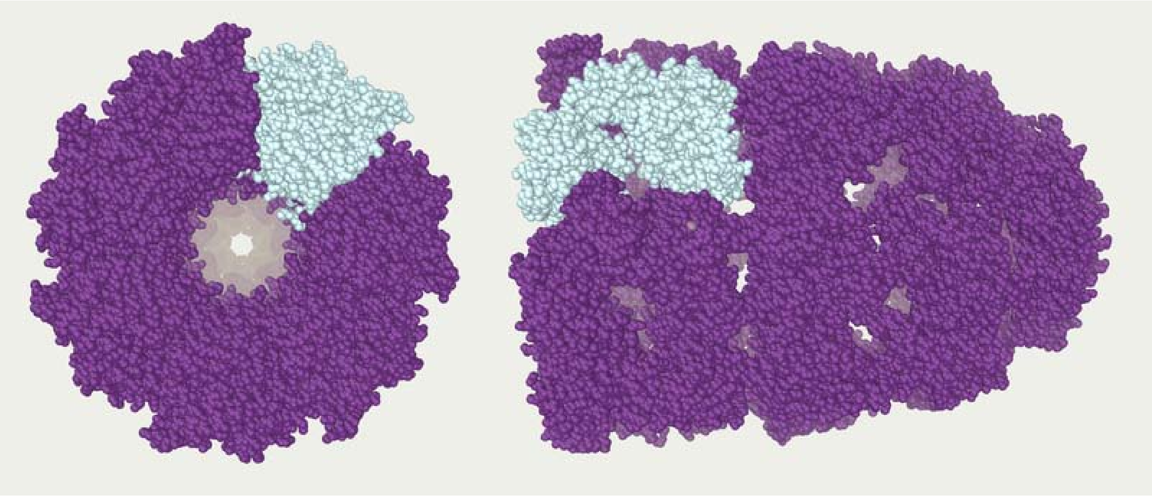
The process of synthesizing a protein from an mRNA template is known as translation. The mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon located on a transfer RNA molecule, which carries the amino acid corresponding to the codon it recognizes. The enzyme aminoacyl tRNA synthetase "charges" the tRNA molecules with the correct amino acids. The growing polypeptide is often termed the *nascent chain*. Proteins are always biosynthesized from N-terminus to C-terminus.

The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of *daltons* (synonymous with atomic mass units), or the derivative unit kilodalton (kDa). Yeast proteins are on average 466 amino acids long and 53 kDa in mass. The largest known proteins are the titins, a component of the muscle sarcomere, with a molecular mass of almost 3,000 kDa and a total length of almost 27,000 amino acids.

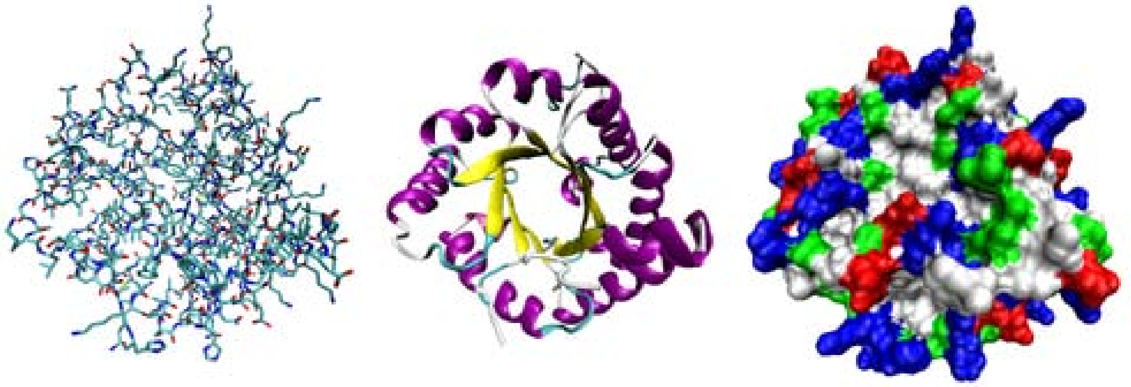
Chemical synthesis

Short proteins can also be synthesized chemically by a family of methods known as peptide synthesis, which rely on organic synthesis techniques such as chemical ligation to produce peptides in high yield. Chemical synthesis allows for the introduction of non-natural amino acids into polypeptide chains, such as attachment of fluorescent probes to amino acid side chains. These methods are useful in laboratory biochemistry and cell biology, though generally not for commercial applications. Chemical synthesis is inefficient for polypeptides longer than about 300 amino acids, and the synthesized proteins may not readily assume their native tertiary structure. Most chemical synthesis methods proceed from C-terminus to N-terminus, opposite the biological reaction.

Structure



The crystal structure of the chaperonin. Chaperonins assist protein folding.

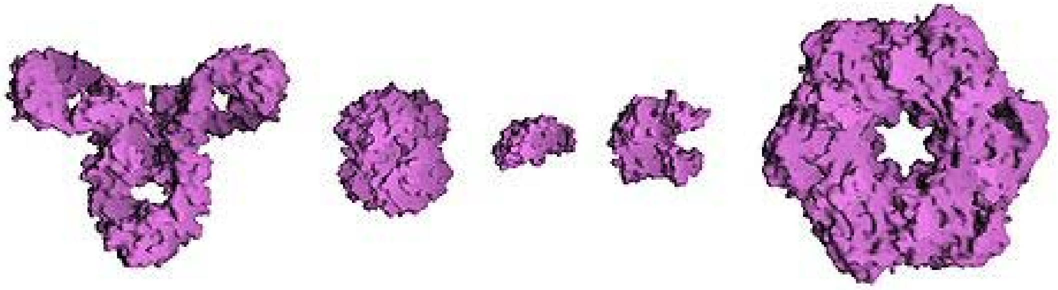


Three possible representations of the three-dimensional structure of the protein triose phosphate isomerase. Left: all-atom representation colored by atom type. Middle: Simplified representation illustrating the backbone conformation, colored by secondary structure. Right: Solvent-accessible surface representation colored by residue type (acidic residues red, basic residues blue, polar residues green, nonpolar residues white)

Most proteins fold into unique 3-dimensional structures. The shape into which a protein naturally folds is known as its native conformation. Although many proteins can fold unassisted, simply through the chemical properties of their amino acids, others require the aid of molecular chaperones to fold into their native states. Biochemists often refer to four distinct aspects of a protein's structure:

- *Primary structure*: the amino acid sequence.
- *Secondary structure*: regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the alpha helix, beta sheet and turns. Because secondary structures are local, many regions of different secondary structure can be present in the same protein molecule.
- *Tertiary structure*: the overall shape of a single protein molecule; the spatial relationship of the secondary structures to one another. Tertiary structure is generally stabilized by nonlocal interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, disulfide bonds, and even post-translational modifications. The term "tertiary structure" is often used as synonymous with the term *fold*. The tertiary structure is what controls the basic function of the protein.
- *Quaternary structure*: the structure formed by several protein molecules (polypeptide chains), usually called *protein subunits* in this context, which function as a single protein complex.

Proteins are not entirely rigid molecules. In addition to these levels of structure, proteins may shift between several related structures while they perform their functions. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as "conformations", and transitions between them are called *conformational changes.* Such changes are often induced by the binding of a substrate molecule to an enzyme's active site, or the physical region of the protein that participates in chemical catalysis. In solution proteins also undergo variation in structure through thermal vibration and the collision with other molecules.



Molecular surface of several proteins showing their comparative sizes. From left to right are: immunoglobulin G (IgG, an antibody), hemoglobin, insulin (a hormone), adenylate kinase (an enzyme), and glutamine synthetase (an enzyme).

Proteins can be informally divided into three main classes, which correlate with typical tertiary structures: globular proteins, fibrous proteins, and membrane proteins. Almost all globular proteins are soluble and many are enzymes. Fibrous proteins are often structural, such as collagen, the major component of connective tissue, or keratin, the protein component of hair and nails. Membrane proteins often serve as receptors or provide channels for polar or charged molecules to pass through the cell membrane.

A special case of intramolecular hydrogen bonds within proteins, poorly shielded from water attack and hence promoting their own dehydration, are called dehydrons.

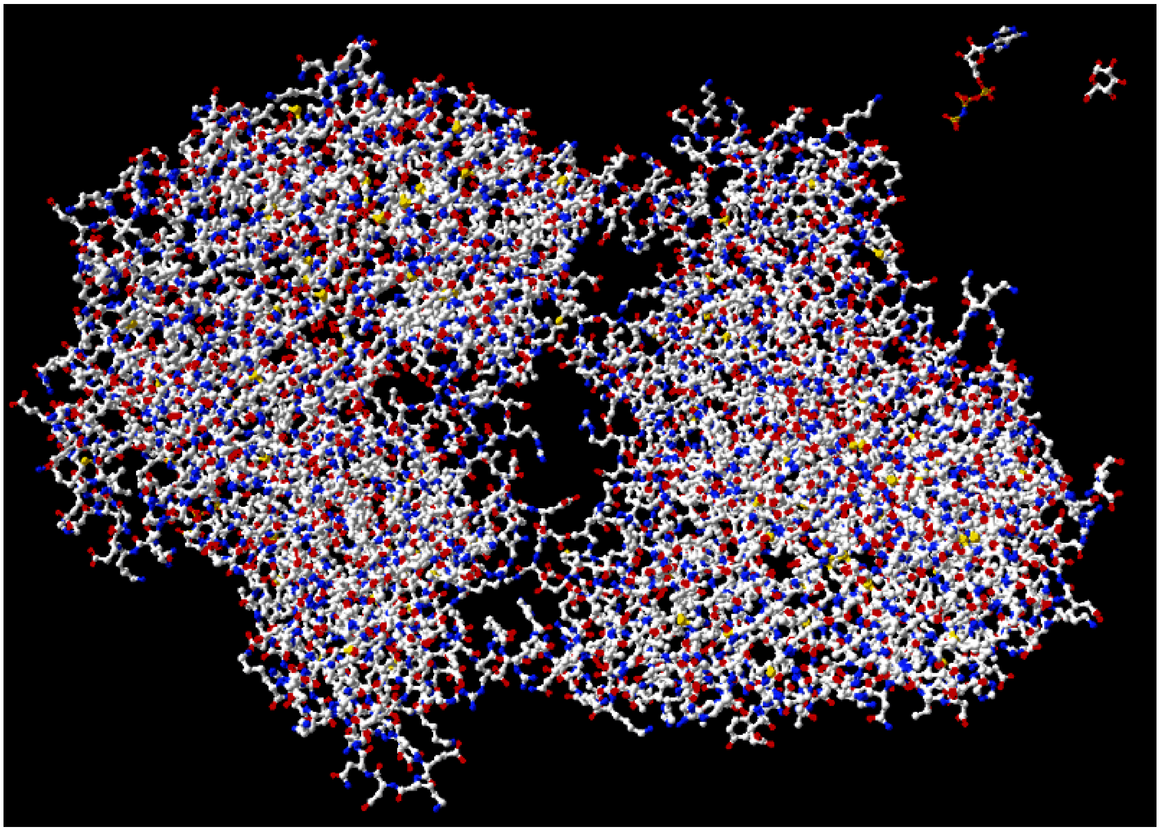
Structure determination

Discovering the tertiary structure of a protein, or the quaternary structure of its complexes, can provide important clues about how the protein performs its function. Common experimental methods of structure determination include X-ray crystallography and NMR spectroscopy, both of which can produce information at atomic resolution. However, NMR experiments are able to provide information from which a subset of distances between pairs of atoms can be estimated, and the final possible conformations for a protein are determined by solving a distance geometry problem. Dual polarisation interferometry is a quantitative analytical method for measuring the overall protein conformation and conformational changes due to interactions or other stimulus. Circular dichroism is another laboratory technique for determining internal beta sheet/ helical composition of proteins. Cryoelectron microscopy is used to produce lower-resolution structural information about very large protein complexes, including assembled viruses; a variant known as electron crystallography can also produce high-resolution information in some cases , especially for two-dimensional crystals of membrane proteins. Solved structures are usually deposited in the Protein Data Bank (PDB), a freely available resource from which structural data about thousands of proteins can be obtained in the form of Cartesian coordinates for each atom in the protein.

Many more gene sequences are known than protein structures. Further, the set of solved structures is biased toward proteins that can be easily subjected to the conditions required in X-ray crystallography, one of the major structure determination methods. In particular, globular proteins are comparatively easy to crystallize in preparation for X-ray crystallography. Membrane proteins, by contrast, are difficult to crystallize and are underrepresented in the PDB. Structural genomics initiatives have attempted to remedy these deficiencies by systematically solving representative structures of major fold classes. Protein structure prediction methods attempt to provide a means of generating a plausible structure for proteins whose structures have not been experimentally determined.

Cellular functions

Proteins are the chief actors within the cell, said to be carrying out the duties specified by the information encoded in genes. With the exception of certain types of RNA, most other biological molecules are relatively inert elements upon which proteins act. Proteins make up half the dry weight of an *Escherichia coli* cell, whereas other macromolecules such as DNA and RNA make up only 3% and 20%, respectively. The set of proteins expressed in a particular cell or cell type is known as its proteome.



The enzyme hexokinase is shown as a simple ball-and-stick molecular model. To scale in the top right-hand corner are two of its substrates, ATP and glucose.

The chief characteristic of proteins that also allows their diverse set of functions is their ability to bind other molecules specifically and tightly. The region of the protein responsible for binding another molecule is known as the binding site and is often a depression or "pocket" on the molecular surface. This binding ability is mediated by the tertiary structure of the protein, which defines the binding site pocket, and by the chemical properties of the surrounding amino acids' side chains. Protein binding can be extraordinarily tight and specific; for example, the ribonuclease inhibitor protein binds to human angiogenin with a sub-femtomolar dissociation constant ($<10^{-15}$ M) but does not bind at all to its amphibian homolog onconase (>1 M). Extremely minor chemical changes such as the addition of a single methyl group to a binding partner can sometimes suffice to nearly eliminate binding; for example, the aminoacyl tRNA synthetase specific to the amino acid valine discriminates against the very similar side chain of the amino acid isoleucine.

Proteins can bind to other proteins as well as to small-molecule substrates. When proteins bind specifically to other copies of the same molecule, they can oligomerize to form fibrils; this process occurs often in structural proteins that consist of globular monomers that self-associate to form rigid fibers. Protein–protein interactions also regulate enzymatic activity, control progression through the cell cycle, and allow the assembly of large protein complexes that carry out many closely related reactions with a common biological function. Proteins can also bind to, or even be integrated into, cell membranes. The ability of binding partners to induce conformational changes in proteins allows the construction of enormously complex signaling networks. Importantly, as interactions between proteins are reversible, and depend heavily on the availability of different groups of partner proteins to form aggregates that are capable to carry out discrete sets of function, study of the interactions between specific proteins is a key to understand important aspects of cellular function, and ultimately the properties that distinguish particular cell types.

Enzymes

The best-known role of proteins in the cell is as enzymes, which catalyze chemical reactions. Enzymes are usually highly specific and accelerate only one or a few chemical reactions. Enzymes carry out most of the reactions involved in metabolism, as well as manipulating DNA in processes such as DNA replication, DNA repair, and transcription. Some enzymes act on other proteins to add or remove chemical groups in a process known as post-translational modification. About 4,000 reactions are known to be catalyzed by enzymes. The rate acceleration conferred by enzymatic catalysis is often enormous—as much as 10^{17}-fold increase in rate over the uncatalyzed reaction in the case of orotate decarboxylase (78 million years without the enzyme, 18 milliseconds with the enzyme).

The molecules bound and acted upon by enzymes are called substrates. Although enzymes can consist of hundreds of amino acids, it is usually only a small fraction of the residues that come in contact with the substrate, and an even smaller fraction—three to four residues on average—that are directly involved in catalysis. The region of the enzyme that binds the substrate and contains the catalytic residues is known as the active site.

Cell signaling and ligand binding

Ribbon diagram of a mouse antibody against cholera that binds a carbohydrate antigen

Many proteins are involved in the process of cell signaling and signal transduction. Some proteins, such as insulin, are extracellular proteins that transmit a signal from the cell in which they were synthesized to other cells in distant tissues. Others are membrane proteins that act as receptors whose main function is to bind a signaling molecule and induce a biochemical response in the cell. Many receptors have a binding site exposed on the cell surface and an effector domain within the cell, which may have enzymatic activity or may undergo a conformational change detected by other proteins within the cell.

Antibodies are protein components of adaptive immune system whose main function is to bind antigens, or foreign substances in the body, and target them for destruction.

Antibodies can be secreted into the extracellular environment or anchored in the membranes of specialized B cells known as plasma cells. Whereas enzymes are limited in their binding affinity for their substrates by the necessity of conducting their reaction, antibodies have no such constraints. An antibody's binding affinity to its target is extraordinarily high.

Many ligand transport proteins bind particular small biomolecules and transport them to other locations in the body of a multicellular organism. These proteins must have a high binding affinity when their ligand is present in high concentrations, but must also release the ligand when it is present at low concentrations in the target tissues. The canonical example of a ligand-binding protein is haemoglobin, which transports oxygen from the lungs to other organs and tissues in all vertebrates and has close homologs in every biological kingdom. Lectins are sugar-binding proteins which are highly specific for their sugar moieties. Lectins typically play a role in biological recognition phenomena involving cells and proteins. Receptors and hormones are highly specific binding proteins.

Transmembrane proteins can also serve as ligand transport proteins that alter the permeability of the cell membrane to small molecules and ions. The membrane alone has a hydrophobic core through which polar or charged molecules cannot diffuse. Membrane proteins contain internal channels that allow such molecules to enter and exit the cell. Many ion channel proteins are specialized to select for only a particular ion; for example, potassium and sodium channels often discriminate for only one of the two ions.

Structural proteins

Structural proteins confer stiffness and rigidity to otherwise-fluid biological components. Most structural proteins are fibrous proteins; for example, actin and tubulin are globular and soluble as monomers, but polymerize to form long, stiff fibers that comprise the cytoskeleton, which allows the cell to maintain its shape and size. Collagen and elastin are critical components of connective tissue such as cartilage, and keratin is found in hard or filamentous structures such as hair, nails, feathers, hooves, and some animal shells.

Other proteins that serve structural functions are motor proteins such as myosin, kinesin, and dynein, which are capable of generating mechanical forces. These proteins are crucial for cellular motility of single celled organisms and the sperm of many multicellular organisms which reproduce sexually. They also generate the forces exerted by contracting muscles.

Methods of study

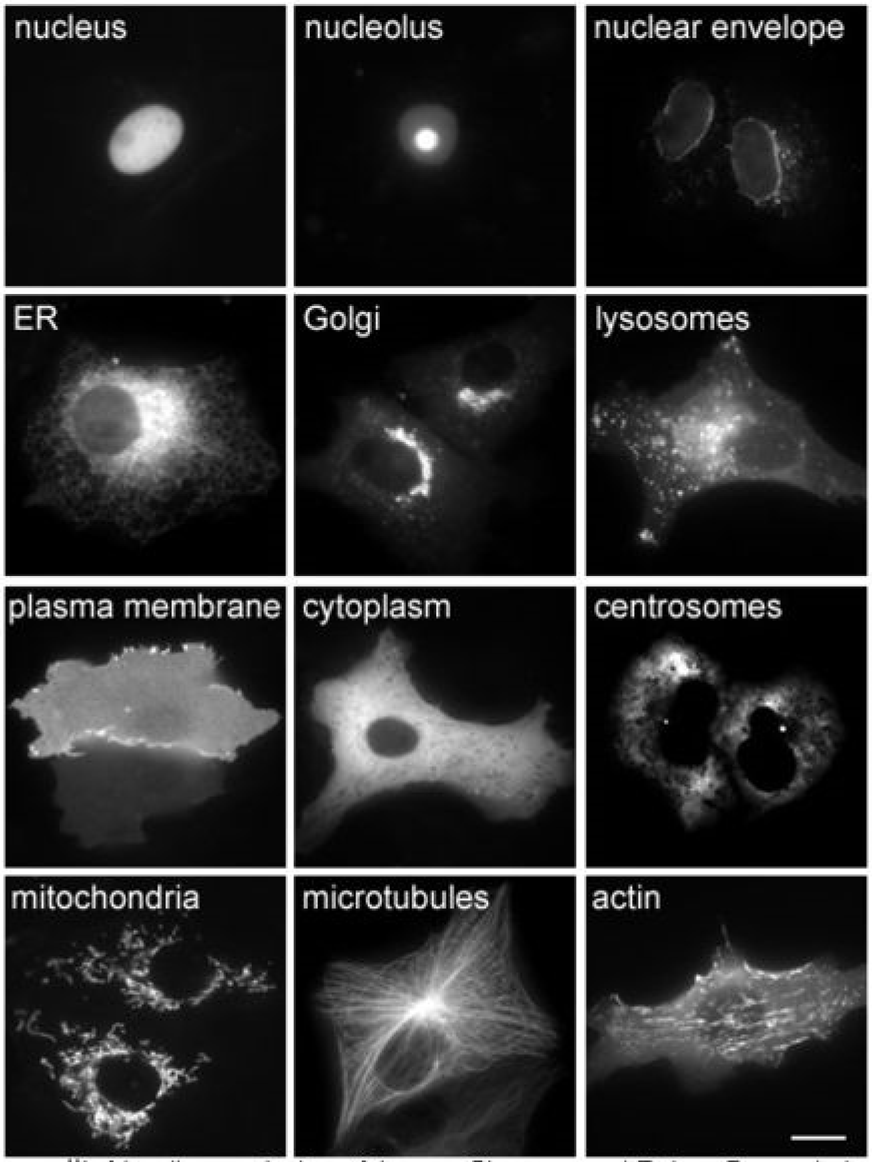
As some of the most commonly studied biological molecules, the activities and structures of proteins are examined both *in vitro* and *in vivo*. *In vitro* studies of purified proteins in controlled environments are useful for learning how a protein carries out its function: for example, enzyme kinetics studies explore the chemical mechanism of an enzyme's catalytic activity and its relative affinity for various possible substrate molecules. By contrast, *in vivo* experiments on proteins' activities within cells or even within whole organisms can provide complementary information about where a protein functions and how it is regulated.

Protein purification

In order to perform *in vitro* analysis, a protein must be purified away from other cellular components. This process usually begins with cell lysis, in which a cell's membrane is disrupted and its internal contents released into a solution known as a crude lysate. The resulting mixture can be purified using ultracentrifugation, which fractionates the various cellular components into fractions containing soluble proteins; membrane lipids and proteins; cellular organelles, and nucleic acids. Precipitation by a method known as salting out can concentrate the proteins from this lysate. Various types of chromatography are then used to isolate the protein or proteins of interest based on properties such as molecular weight, net charge and binding affinity. The level of purification can be monitored using various types of gel electrophoresis if the desired protein's molecular weight and isoelectric point are known, by spectroscopy if the protein has distinguishable spectroscopic features, or by enzyme assays if the protein has enzymatic activity. Additionally, proteins can be isolated according their charge using electrofocusing.

For natural proteins, a series of purification steps may be necessary to obtain protein sufficiently pure for laboratory applications. To simplify this process, genetic engineering is often used to add chemical features to proteins that make them easier to purify without affecting their structure or activity. Here, a "tag" consisting of a specific amino acid sequence, often a series of histidine residues (a "His-tag"), is attached to one terminus of the protein. As a result, when the lysate is passed over a chromatography column containing nickel, the histidine residues ligate the nickel and attach to the column while the untagged components of the lysate pass unimpeded. A number of different tags have been developed to help researchers purify specific proteins from complex mixtures.

Cellular localization



with friendly permission of Jeremy Simpson and Rainer Pepperkok

Proteins in different cellular compartments and structures tagged with green fluorescent protein (here, white)

The study of proteins *in vivo* is often concerned with the synthesis and localization of the protein within the cell. Although many intracellular proteins are synthesized in the cytoplasm and membrane-bound or secreted proteins in the endoplasmic reticulum, the specifics of how proteins are targeted to specific organelles or cellular structures is often unclear. A useful technique for assessing cellular localization uses genetic engineering to express in a cell a fusion protein or chimera consisting of the natural protein of interest linked to a "reporter" such as green fluorescent protein (GFP). The fused protein's position within the cell can be cleanly and efficiently visualized using microscopy, as shown in the figure opposite.

Other methods for elucidating the cellular location of proteins requires the use of known compartmental markers for regions such as the ER, the Golgi, lysosomes/vacuoles, mitochondria, chloroplasts, plasma membrane, etc. With the use of fluorescently tagged versions of these markers or of antibodies to known markers, it becomes much simpler to identify the localization of a protein of interest. For example, indirect immunofluorescence will allow for fluorescence colocalization and demonstration of location. Fluorescent dyes are used to label cellular compartments for a similar purpose.

Other possibilities exist, as well. For example, immunohistochemistry usually utilizes an antibody to one or more proteins of interest that are conjugated to enzymes yielding either luminescent or chromogenic signals that can be compared between samples, allowing for localization information. Another applicable technique is cofractionation in sucrose (or other material) gradients using isopycnic centrifugation. While this technique does not prove colocalization of a compartment of known density and the protein of interest, it does increase the likelihood, and is more amenable to large-scale studies.

Finally, the gold-standard method of cellular localization is immunoelectron microscopy. This technique also uses an antibody to the protein of interest, along with classical electron microscopy techniques. The sample is prepared for normal electron microscopic examination, and then treated with an antibody to the protein of interest that is conjugated to an extremely electro-dense material, usually gold. This allows for the localization of both ultrastructural details as well as the protein of interest.

Through another genetic engineering application known as site-directed mutagenesis, researchers can alter the protein sequence and hence its structure, cellular localization, and susceptibility to regulation. This technique even allows the incorporation of unnatural amino acids into proteins, using modified tRNAs, and may allow the rational design of new proteins with novel properties.

Proteomics and bioinformatics

The total complement of proteins present at a time in a cell or cell type is known as its proteome, and the study of such large-scale data sets defines the field of proteomics, named by analogy to the related field of genomics. Key experimental techniques in proteomics include 2D electrophoresis, which allows the separation of a large number of proteins, mass spectrometry, which allows rapid high-throughput identification of proteins and sequencing of peptides (most often after in-gel digestion), protein microarrays, which allow the detection of the relative levels of a large number of proteins present in a cell, and two-hybrid screening, which allows the systematic exploration of protein–protein interactions. The total complement of biologically possible such interactions is known as the interactome. A systematic attempt to determine the structures of proteins representing every possible fold is known as structural genomics.

The large amount of genomic and proteomic data available for a variety of organisms, including the human genome, allows researchers to efficiently identify homologous proteins in distantly related organisms by sequence alignment. Sequence profiling tools can perform more specific sequence manipulations such as restriction enzyme maps, open reading frame analyses for nucleotide sequences, and secondary structure prediction.

From this data phylogenetic trees can be constructed and evolutionary hypotheses developed using special software like ClustalW regarding the ancestry of modern organisms and the genes they express. The field of bioinformatics seeks to assemble, annotate, and analyze genomic and proteomic data, applying computational techniques to biological problems such as gene finding and cladistics.

Structure prediction and simulation

Complementary to the field of structural genomics, protein structure prediction seeks to develop efficient ways to provide plausible models for proteins whose structures have not yet been determined experimentally. The most successful type of structure prediction, known as homology modeling, relies on the existence of a "template" structure with sequence similarity to the protein being modeled; structural genomics' goal is to provide sufficient representation in solved structures to model most of those that remain. Although producing accurate models remains a challenge when only distantly related template structures are available, it has been suggested that sequence alignment is the bottleneck in this process, as quite accurate models can be produced if a "perfect" sequence alignment is known. Many structure prediction methods have served to inform the emerging field of protein engineering, in which novel protein folds have already been designed. A more complex computational problem is the prediction of intermolecular interactions, such as in molecular docking and protein–protein interaction prediction.

The processes of protein folding and binding can be simulated using such technique as molecular mechanics, in particular, molecular dynamics and Monte Carlo, which increasingly take advantage of parallel and distributed computing (Folding@Home project; molecular modeling on GPU). The folding of small alpha-helical protein domains such as the villin headpiece and the HIV accessory protein have been successfully simulated *in silico*, and hybrid methods that combine standard molecular dynamics with quantum mechanics calculations have allowed exploration of the electronic states of rhodopsins.

Nutrition

Most microorganisms and plants can biosynthesize all 20 standard amino acids, while animals (including humans) must obtain some of the amino acids from the diet. The amino acids that an organism cannot synthesize on its own are referred to as essential amino acids. Key enzymes that synthesize certain amino acids are not present in animals — such as aspartokinase, which catalyzes the first step in the synthesis of lysine, methionine, and threonine from aspartate. If amino acids are present in the environment, microorganisms can conserve energy by taking up the amino acids from their surroundings and downregulating their biosynthetic pathways.

In animals, amino acids are obtained through the consumption of foods containing protein. Ingested proteins are then broken down into amino acids through digestion, which typically involves denaturation of the protein through exposure to acid and hydrolysis by enzymes called proteases. Some ingested amino acids are used for protein biosynthesis, while others are converted to glucose through gluconeogenesis, or fed into the citric acid cycle. This use of protein as a fuel is particularly important under starvation conditions as it allows the body's own proteins to be used to support life, particularly those found in muscle. Amino acids are also an important dietary source of nitrogen.

History and etymology

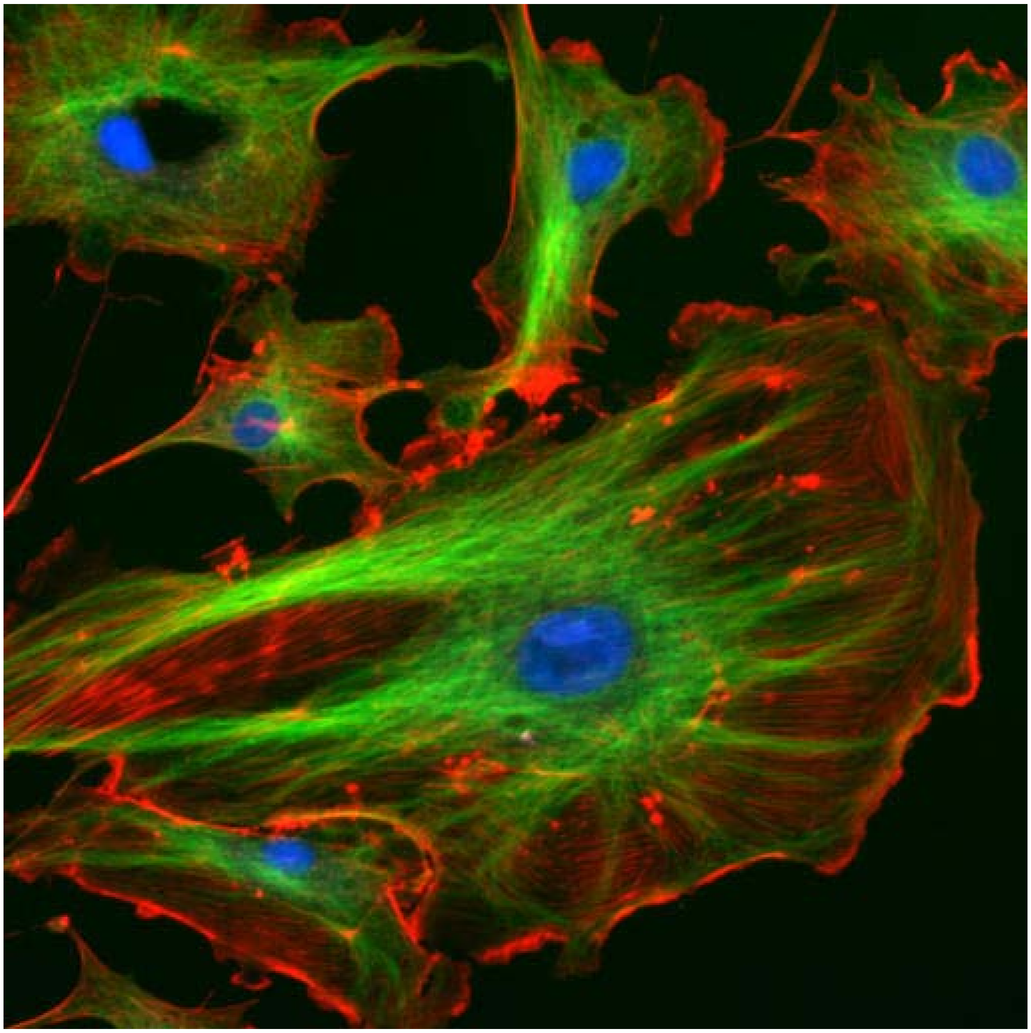
Proteins were recognized as a distinct class of biological molecules in the eighteenth century by Antoine Fourcroy and others, distinguished by the molecules' ability to coagulate or flocculate under treatments with heat or acid. Noted examples at the time included albumin from egg whites, blood serum albumin, fibrin, and wheat gluten. Dutch chemist Gerhardus Johannes Mulder carried out elemental analysis of common proteins and found that nearly all proteins had the same empirical formula, $C_{400}H_{620}N_{100}O_{120}P_1S_1$. He came to the erroneous conclusion that they might be composed of a single type of (very large) molecule. The term "protein" to describe these molecules was proposed in 1838 by Mulder's associate Jöns Jakob Berzelius; protein is derived from the Greek word πρωτεῖος (*proteios*), meaning "primary", "in the lead", or "standing in front". Mulder went on to identify the products of protein degradation such as the amino acid leucine for which he found a (nearly correct) molecular weight of 131 Da.

The difficulty in purifying proteins in large quantities made them very difficult for early protein biochemists to study. Hence, early studies focused on proteins that could be purified in large quantities, e.g., those of blood, egg white, various toxins, and digestive/metabolic enzymes obtained from slaughterhouses. In the 1950s, the Armour Hot Dog Co. purified 1 kg of pure bovine pancreatic ribonuclease A and made it freely available to scientists; this gesture helped ribonuclease A become a major target for biochemical study for the following decades.

Linus Pauling is credited with the successful prediction of regular protein secondary structures based on hydrogen bonding, an idea first put forth by William Astbury in 1933. Later work by Walter Kauzmann on denaturation, based partly on previous studies by Kaj Linderstrøm-Lang, contributed an understanding of protein folding and structure mediated by hydrophobic interactions. In 1949 Fred Sanger correctly determined the amino acid sequence of insulin, thus conclusively demonstrating that proteins consisted of linear polymers of amino acids rather than branched chains, colloids, or cyclols. The first atomic-resolution structures of proteins were solved by X-ray crystallography in the 1960s and by NMR in the 1980s. As of 2009, the Protein Data Bank has over 55,000 atomic-resolution structures of proteins. In more recent times, cryo-electron microscopy of large macromolecular assemblies and computational protein structure prediction of small protein domains are two methods approaching atomic resolution.

Chapter 9

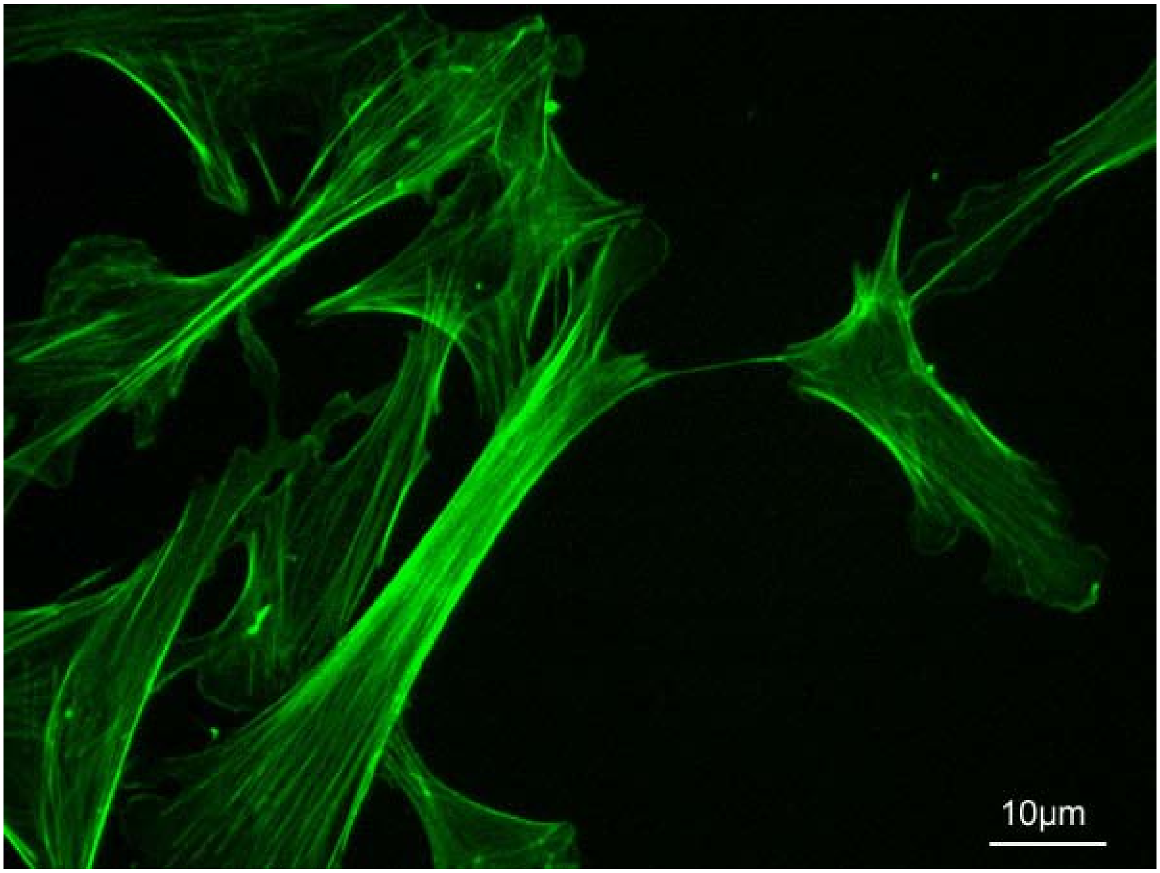
Cytoskeleton



The eukaryotic cytoskeleton. Actin filaments are shown in red, microtubules in green, and the nuclei are in blue.

The **cytoskeleton** (also CSK) is a cellular "scaffolding" or "skeleton" contained within the cytoplasm and is made out of protein. The cytoskeleton is present in all cells; it was once thought to be unique to eukaryotes, but recent research has identified the prokaryotic cytoskeleton. It has structures such as flagella, cilia and lamellipodia and plays important roles in both intracellular transport (the movement of vesicles and organelles, for example) and cellular division. The concept of a protein mosaic that dynamically coordinated cytoplasmic biochemistry was proposed by Rudolph Peters in 1929 while the term (*cytosquelette*, in French) was first introduced by French embryologist Paul Wintrebert in 1931.

The eukaryotic cytoskeleton



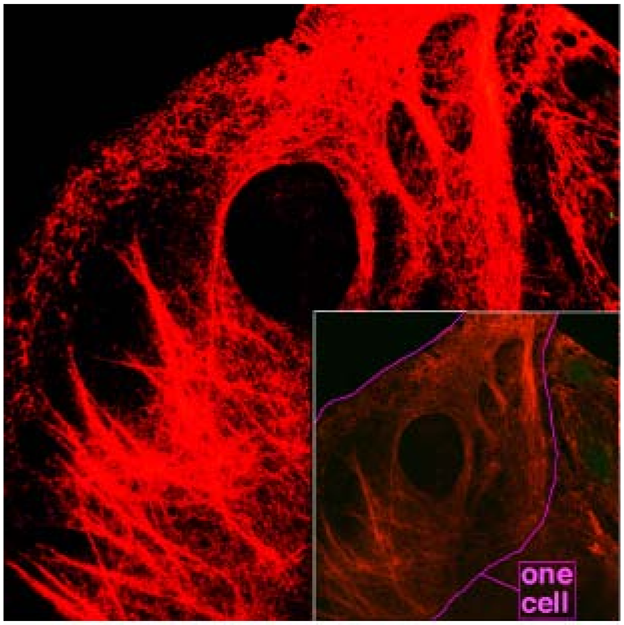
Actin cytoskeleton of mouse embryo fibroblasts, stained with phalloidin.

Eukaryotic cells contain three main kinds of cytoskeletal filaments, which are microfilaments, intermediate filaments, and microtubules. The cytoskeleton provides the cell with structure and shape, and by excluding macromolecules from some of the cytosol it adds to the level of macromolecular crowding in this compartment. Cytoskeletal elements interact extensively and intimately with cellular membranes.

Microfilaments

These are the thinnest filaments of the cytoskeleton. They are composed of linear polymers of actin subunits, and generate force by elongation at one end of the filament coupled with shrinkage at the other, causing net movement of the intervening strand. They also act as tracks for the movement of myosin molecules that attach to the microfilament and "walk" along them.

Intermediate filaments



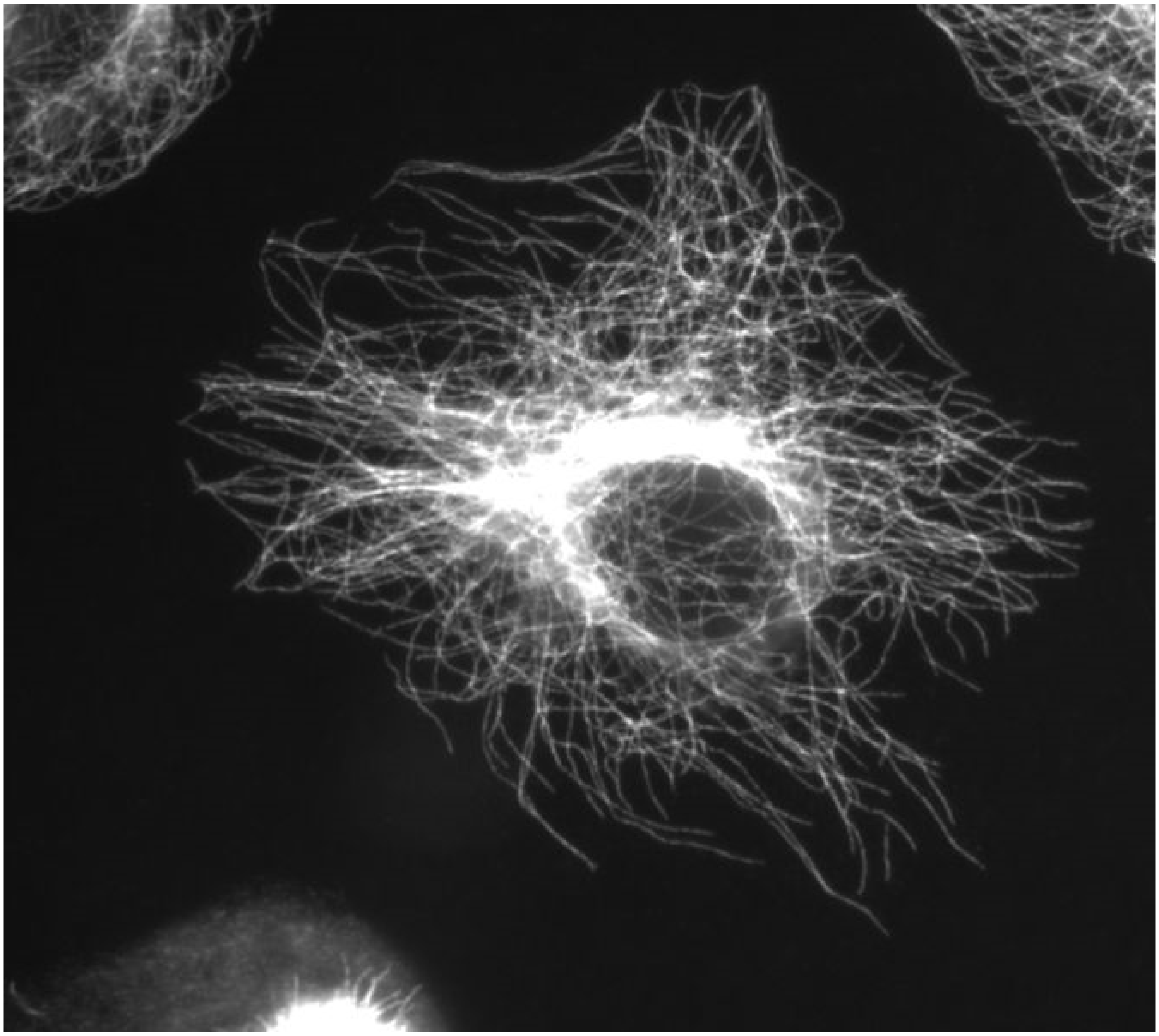
Microscopy of keratin filaments inside cells.

These filaments, around 10 nanometers in diameter, are more stable (strongly bound) than actin filaments, and heterogeneous constituents of the cytoskeleton. Although little work has been done on intermediate filaments in plants, there is some evidence that cytosolic intermediate filaments might be present, and plant nuclear filaments have been detected. Like actin filaments, they function in the maintenance of cell-shape by bearing tension (microtubules, by contrast, resist compression. It may be useful to think of micro- and intermediate filaments as cables, and of microtubules as cellular support beams). Intermediate filaments organize the internal tridimensional structure of the cell, anchoring organelles and serving as structural components of the nuclear lamina and sarcomeres. They also participate in some cell-cell and cell-matrix junctions.

Different intermediate filaments are:

- made of vimentins, being the common structural support of many cells.
- made of keratin, found in skin cells, hair and nails.
- neurofilaments of neural cells.
- made of lamin, giving structural support to the nuclear envelope.

Microtubules



Microtubules in a gel fixated cell.

Microtubules are hollow cylinders about 23 nm in diameter (lumen = approximately 15nm in diameter), most commonly comprising 13 protofilaments which, in turn, are polymers of alpha and beta tubulin. They have a very dynamic behaviour, binding GTP for polymerization. They are commonly organized by the centrosome.

In nine triplet sets (star-shaped), they form the centrioles, and in nine doublets oriented about two additional microtubules (wheel-shaped) they form cilia and flagella. The latter formation is commonly referred to as a "9+2" arrangement, wherein each doublet is connected to another by the protein dynein. As both flagella and cilia are structural components of the cell, and are maintained by microtubules, they can be considered part of the cytoskeleton.

They play key roles in:

- intracellular transport (associated with dyneins and kinesins, they transport organelles like mitochondria or vesicles).
- the axoneme of cilia and flagella.
- the mitotic spindle.
- synthesis of the cell wall in plants.

Comparison

Cytoskeleton type	Diameter (nm)	Structure	Subunit examples
Microfilaments	6	double helix	actin
Intermediate filaments	10	two anti-parallel helices/dimers, forming tetramers	• vimentin (mesenchyme) • glial fibrillary acidic protein (glial cells) • neurofilament proteins (neuronal processes) • keratins (epithelial cells) • nuclear lamins
Microtubules	23	protofilaments, in turn consisting of tubulin subunits	α- and β-tubulin

The prokaryotic cytoskeleton

The cytoskeleton was previously thought to be a feature only of eukaryotic cells, but homologues to all the major proteins of the eukaryotic cytoskeleton have recently been found in prokaryotes. Although the evolutionary relationships are so distant that they are not obvious from protein sequence comparisons alone, the similarity of their three-dimensional structures and similar functions in maintaining cell shape and polarity provides strong evidence that the eukaryotic and prokaryotic cytoskeletons are truly homologous. However, some structures in the bacterial cytoskeleton may have yet to be identified.

FtsZ

FtsZ was the first protein of the prokaryotic cytoskeleton to be identified. Like tubulin, FtsZ forms filaments in the presence of GTP, but these filaments do not group into tubules. During cell division, FtsZ is the first protein to move to the division site, and is essential for recruiting other proteins that synthesize the new cell wall between the dividing cells.

MreB and ParM

Prokaryotic actin-like proteins, such as MreB, are involved in the maintenance of cell shape. All non-spherical bacteria have genes encoding actin-like proteins, and these proteins form a helical network beneath the cell membrane that guides the proteins involved in cell wall biosynthesis.

Some plasmids encode a partitioning system that involves an actin-like protein ParM. Filaments of ParM exhibit dynamic instability, and may partition plasmid DNA into the dividing daughter cells by a mechanism analogous to that used by microtubules during eukaryotic mitosis.

Crescentin

The bacterium *Caulobacter crescentus* contains a third 3rd protein, crescentin, that is related to the intermediate filaments of eukaryotic cells. Crescentin is also involved in maintaining cell shape, such as helical and vibrioid forms of bacteria, but the mechanism by which it does this is currently unclear.

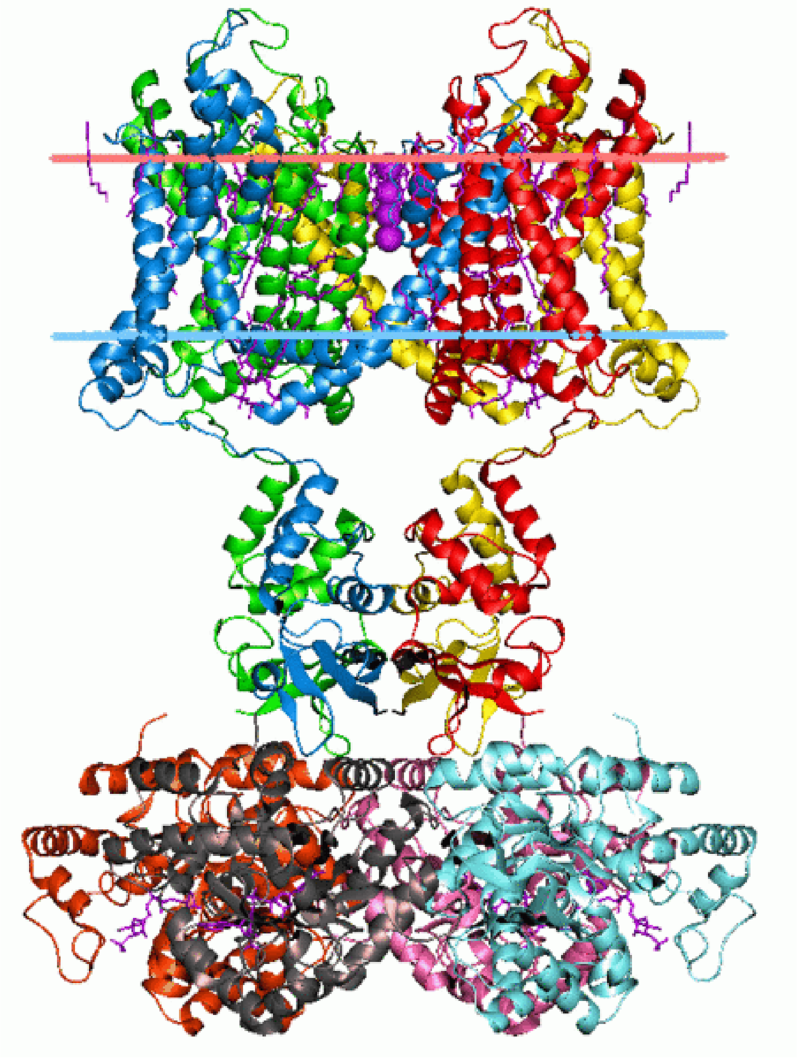
History

Microtrabeculae

A fourth eukaryotic cytoskeletal element, *microtrabeculae*, was proposed by Keith Porter based on images obtained from high-voltage electron microscopy of whole cells in the 1970s. The images showed short, filamentous structures of unknown molecular composition associated with known cytoplasmic structures. Porter proposed that this microtrabecular structure represented a novel filamentous network distinct from microtubules, filamentous actin, or intermediate filaments. It is now generally accepted that microtrabeculae are nothing more than an artifact of certain types of fixation treatment, although we have yet to fully understand the complexity of the cell's cytoskeleton.

Chapter 10

Membrane Protein



Crystal structure of Potassium channel KvAP. Calculated hydrocarbon boundaries of the lipid bilayer are indicated by red and blue dots.

A **membrane protein** is a protein molecule that is attached to, or associated with the membrane of a cell or an organelle. More than half of all proteins interact with membranes.

Function

Biological membranes consist of a phospholipid bilayer and a variety of proteins that accomplish vital biological functions.

- Structural proteins are attached to microfilaments in the cytoskeleton which ensures stability of the cell.
- Cell adhesion molecules allow cells to identify each other and interact. Such proteins are involved in immune response, for example.
- Membrane enzymes produce a variety of substances essential for cell function.
- Membrane receptor proteins serve as connection between the cell's internal and external environments.
- Transport proteins play an important role in the maintenance of concentrations of ions. These transport proteins come in two forms: carrier proteins and channel proteins.

Main categories

Membrane proteins can be divided into three categories:

- Integral membrane proteins, penetrating the lipid bilayer
- Peripheral membrane proteins, external but bound with noncovalent bonds
- Lipid-anchored protein, external but bound with covalent bonds

An alternative classification is to divide all membrane proteins to integral and *amphitropic*.

- The *amphitropic* are proteins that can exist in two alternative states: a water-soluble and a lipid bilayer-bound. The amphitropic protein category includes water-soluble channel-forming polypeptide toxins, which associate irreversibly with membranes, but excludes peripheral proteins that interact with other membrane proteins rather than with lipid bilayer.
- The *integral* proteins can be found only in the membrane-bound state.

Integral membrane proteins

Integral membrane proteins are permanently attached to the membrane. They can be defined as those proteins which require a detergent (such as SDS or Triton X-100) or some other apolar solvent to be displaced. They can be classified according to their relationship with the bilayer:

- Integral polytopic proteins, also known as "transmembrane proteins," are proteins that are permanently attached to the lipid membrane and span across the membrane (at least once). The transmembrane regions of the proteins are either beta-barrels or alpha-helical. The alpha-helical domains are present in all types of biological membranes including outer membranes. The beta-barrels were found only in outer membranes of Gram-negative bacteria, lipid-rich cell walls of a few Gram-positive bacteria, and outer membranes of mitochondria and chloroplasts.

- Integral monotopic proteins are proteins that are permanently attached to the lipid membrane from only one side and do not span across the membrane.

Peripheral membrane proteins

Peripheral membrane proteins are temporarily attached either to the lipid bilayer or to integral proteins by a combination of hydrophobic, electrostatic, and other non-covalent interactions. Peripheral proteins dissociate following treatment with a polar reagent, such as a solution with an elevated pH or high salt concentrations.

Integral and peripheral proteins may be post-translationally modified, with added fatty acid or prenyl chains, or GPI (glycosylphosphatidylinositol), which may be anchored in the lipid bilayer.

Polypeptide toxins

Classification of membrane proteins to integral and peripheral does not include some polypeptide toxins, such as colicin A or alpha-hemolysin, and certain proteins involved in apoptosis. These proteins are water-soluble but can aggregate and associate irreversibly with the lipid bilayer and form alpha-helical or beta-barrel transmembrane channels.

Intracellular localization

Proteins are specifically targeted to many different types of biological membranes

Membrane Protein Complexes

Membrane Proteins commonly function as complexes. These complexes are vital to cellular function. Understanding how these complexes are assembled, degraded, and their composition are crucial to understanding their function and regulation. Reoccurring in recent literature are the ideas that: membrane protein complexes assemble in an orderly fashion, chaperones aide assembly by preventing unfavorable interactions, and membrane proteins can be interchanged in existing complexes. Membrane protein complexes assemble through the orderly assembly of intermediates. For example, the simple membrane-embedded four subunit complex, cytochrome bo3 of Escherichia coli, is assembled via two intermediate complexes. This suggests a linearly organized assembly pathway. Although interactions between other subunits could lead to the formation of many intermediates, they do not occur. Ordered assembly could be the cell's protection against harmful intermediates. Chaperones interact with membrane proteins guiding their assembly. They aide in preventing the assembly of dead-end and toxic intermediates, as well as unwanted aggregations. Via chaperones assembly can occur through inactive intermediates potentially preventing damaging interactions they could cause. Membrane protein complexes are not fixed entities. Though a process called dynamic exchange, membrane proteins are exchanged in and out of exsitisting protein complexes. This has its implications as a repair mechanism and in regulation.

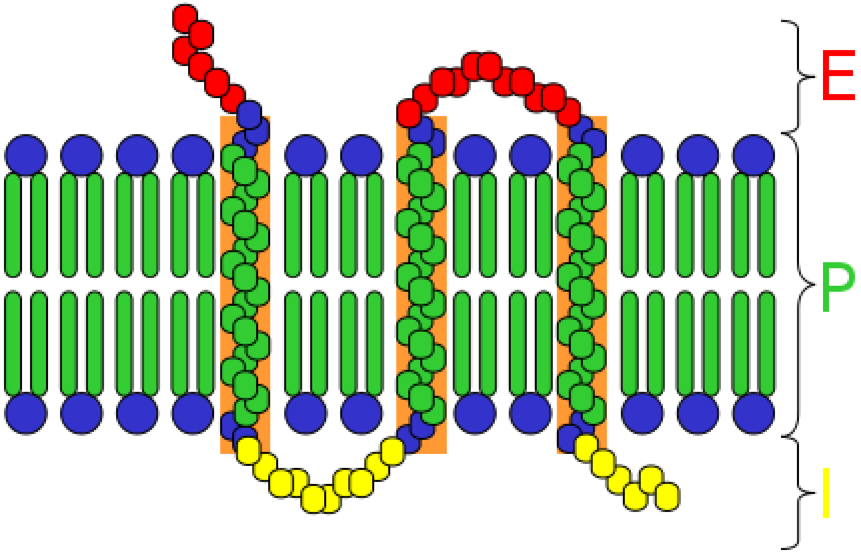
Membrane Protein Structures

The structures of membrane proteins are stabilized by weak interactions and influenced by additional interactions with the solubilizing environment. The influence of the environment on membrane protein structures is especially significant. Despite the significant functional importance of membrane proteins, the structural biology has been particularly challenging as shown by the low number of membrane protein structures determined. Integral membrane proteins are present in a heterogeneous environment that poses major obstacles for existing structural methodologies.

Many of the successful membrane protein structures are characterized by X-ray crystallography and are very large structures in which the interactions with the membrane mimetic environments can be anticipated to be small in comparison to those within the protein structures. The small domains are particularly sensitive to the influence of membrane mimetic environments, potentially leading to non-native structures. Fortunately, there are many sample preparation conditions that can be chosen for crystallization and for solution NMR. All membrane protein structural biology should be subjected to careful scrutiny; through a combination of structural methodologies it should be possible to achieve an understanding of the native functional state for membrane protein structures.

Chapter 11

Integral Membrane Protein



E=extracellular space; P=plasma membrane; I=intracellular space

An **integral membrane protein** (**IMP**) is a protein molecule (or assembly of proteins) that is permanently attached to the biological membrane. Proteins that cross the membrane are surrounded by "annular" lipids, which are defined as lipids that are in direct contact with a membrane protein. Such proteins can be separated from the biological membranes only using detergents, nonpolar solvents, or sometimes denaturing agents.

IMPs comprise a very significant fraction of the proteins encoded in an organism's genome.

All transmembrane proteins are IMPs, but not all IMPs are transmembrane proteins.

Structure

Three-dimensional structures of only ~160 different integral membrane proteins are currently determined at atomic resolution by X-ray crystallography or nuclear magnetic resonance spectroscopy due to the difficulties with extraction and crystallization. In addition, structures of many water-soluble domains of IMPs are available in the Protein Data Bank. Their membrane-anchoring α-helices have been removed to facilitate the extraction and crystallization.

IMPs can be divided into two groups:

1. Integral polytopic proteins (Transmembrane proteins)
2. Integral monotopic proteins

Integral Polytopic Protein

The most common type of IMP is the transmembrane protein (TM), which spans the entire biological membrane. **Single-pass** membrane proteins cross the membrane only once, while **multi-pass** membrane proteins weave in and out, crossing several times. Single pass TM proteins can be categorized as Type I, which are positioned such that their amino-terminus is outside of the membrane, or Type II, which have their carboxy-terminus outside of the membrane.

Integral Monotopic Proteins

Integral monotopic proteins, are permanently attached to the membrane from one side. Such domains require detergents for extraction or crystallization, even after removal of their transmembrane helices. Therefore, they are often classified as integral monotopic *proteins*

Determination of Protein Structure

The use of hydropathy plots helps determine integral protein structures based on the hydrophobic and hydrophilic characteristics of alpha helical integral proteins.

Function

IMPs include transporters, channels, receptors, enzymes, structural membrane-anchoring domains, proteins involved in accumulation and transduction of energy, and proteins responsible for cell adhesion. Classification of transporters can be found in Transporter Classification database.

Examples

Examples of integral membrane proteins:

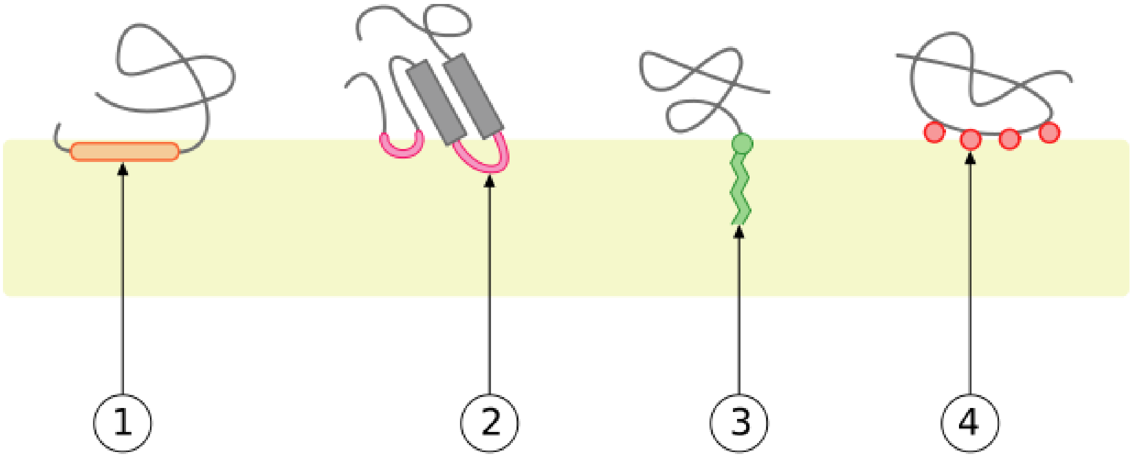
- Insulin receptor
- Some types of cell adhesion proteins or cell adhesion molecules (CAMs) such as Integrins, Cadherins, NCAMs, or Selectins.
- Some types of receptor proteins
- Glycophorin
- Rhodopsin
- Band 3
- CD36
- GPR30

Chapter 12

Peripheral Membrane Protein

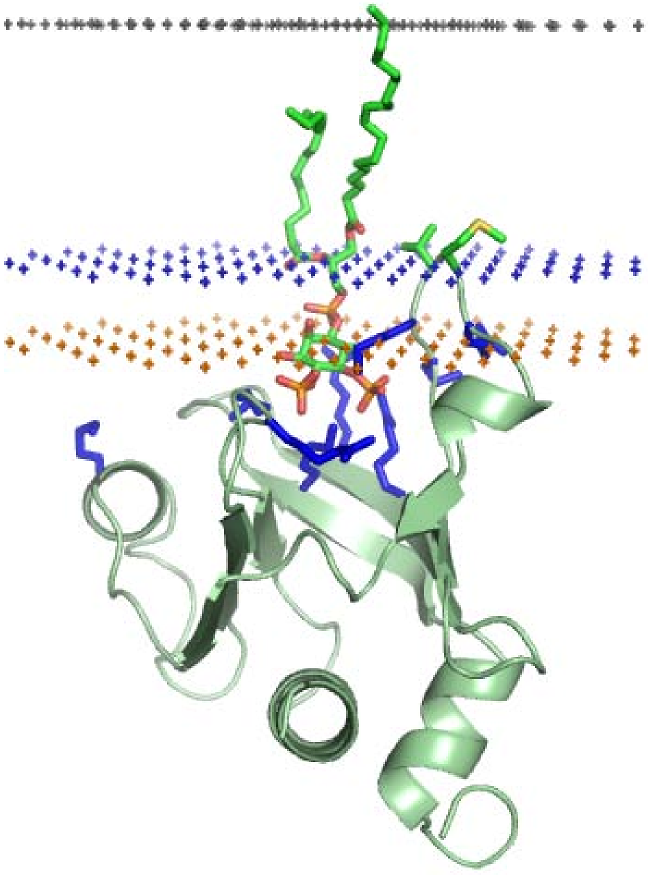
Peripheral membrane proteins are proteins that adhere only temporarily to the biological membrane with which they are associated. These molecules attach to integral membrane proteins, or penetrate the peripheral regions of the lipid bilayer. The regulatory protein subunits of many ion channels and transmembrane receptors, for example, may be defined as peripheral membrane proteins. In contrast to integral membrane proteins, peripheral membrane proteins tend to collect in the water-soluble component, or fraction, of all the proteins extracted during a protein purification procedure. Proteins with GPI anchors are an exception to this rule and can have purification properties similar to those of integral membrane proteins.

The reversible attachment of proteins to biological membranes has shown to regulate cell signaling and many other important cellular events, through a variety of mechanisms. For example, the close association between many enzymes and biological membranes may bring them into close proximity with their lipid substrate(s). Membrane binding may also promote rearrangement, dissociation, or conformational changes within many protein structural domains, resulting in an activation of their biological activity. Additionally, the positioning of many proteins are localized to either the inner or outer surfaces or leaflets of their resident membrane. This facilitates the assembly of multi-protein complexes by increasing the probability of any appropriate protein-protein interactions.



Schematic representation of the different types of interaction between monotopic membrane proteins and the cell membrane: 1. interaction by an amphipathic α-helix parallel to the membrane plane (in-plane membrane helix) 2. interaction by a hydrophobic loop 3. interaction by a covalently bound membrane lipid (*lipidation*) 4. electrostatic or ionic interactions with membrane lipids (*e.g.* through a calcium ion)

Binding of peripheral proteins to the lipid bilayer



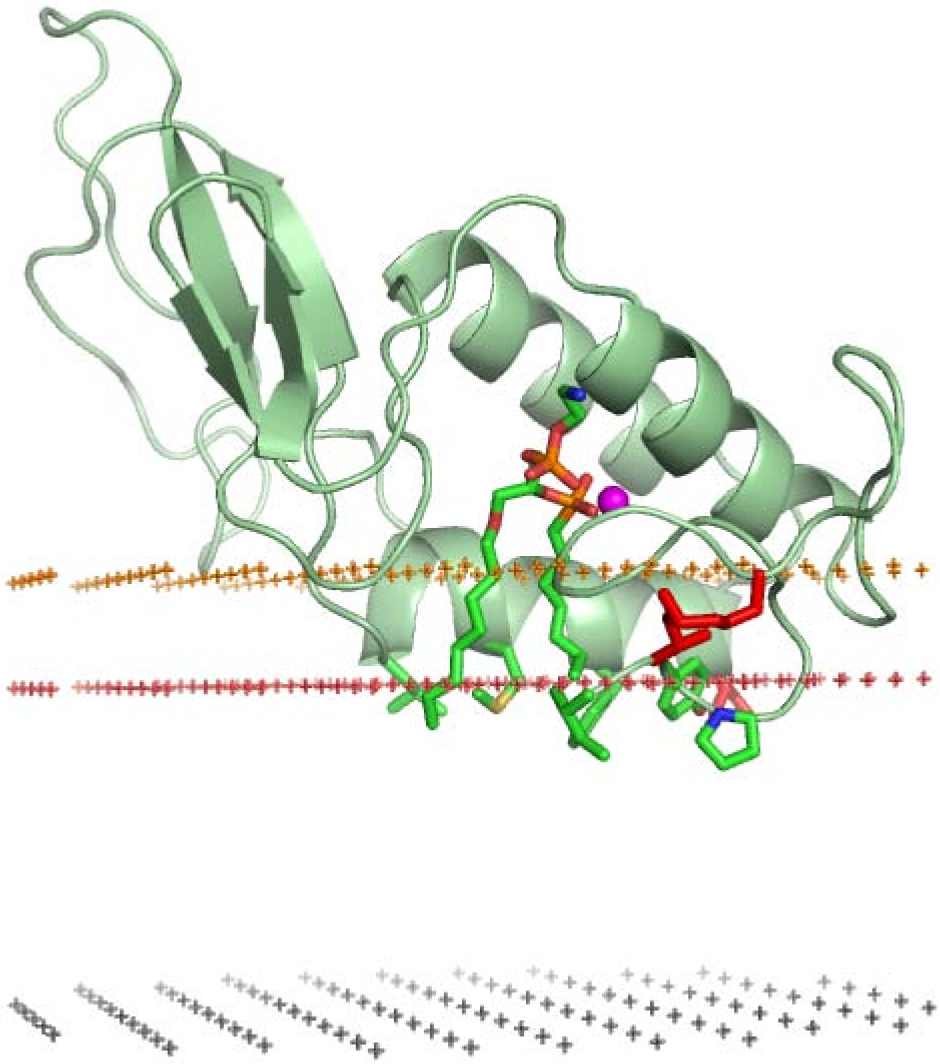
PH domain of phospholipase C delta 1. Middle plane of the lipid bilayer - black dots. Boundary of the hydrocarbon core region - blue dots (intracellular side). Layer of lipid phosphates - yellow dots.

Peripheral membrane proteins may interact with other proteins or directly with the lipid bilayer. In the latter case, they are then known as *amphitropic* proteins. Some proteins, such as G-proteins and certain protein kinases, interact with transmembrane proteins and the lipid bilayer simultaneously. Some polypeptide hormones, antimicrobial peptides, and neurotoxins accumulate at the membrane surface prior to locating and interacting with their cell surface receptor targets, which may themselves be peripheral membrane proteins.

The Phospholipid bilayer that forms the cell surface membrane consists of a hydrophobic inner core region sandwiched between two regions of hydrophilicity, one at the inner surface and one at the outer surface of the cell membrane. The inner and outer surfaces, or interfacial regions, of model phospholipid bilayers have been shown to have a thickness of around 8 to 10 Å, although this may be wider in biological membranes that include large amounts of gangliosides or lipopolysaccharides. The hydrophobic inner core region of typical biological membranes may have a thickness of around 27 to 32 Å, as estimated by Small angle X-ray scattering (SAXS). The boundary region between the hydrophobic inner core and the hydrophilic interfacial regions is very narrow, at around 3Å. Moving outwards away from the hydrophobic core region and into the interfacial hydrophilic region, the effective concentration of water rapidly changes across this boundary layer, from nearly zero to a concentration of around 2M. The phosphate groups within phospholipid bilayers are fully hydrated or saturated with water and are situated around 5 Å outside the boundary of the hydrophobic core region.

Some water-soluble proteins associate with lipid bilayers *irreversibly* and can form transmembrane alpha-helical or beta-barrel channels. Such transformations occur in pore forming toxins such as colicin A, alpha-hemolysin, and others. They may also occur in [[Bcl-2-associated X protein|BcL-2 like protein , in some amphiphilic antimicrobial peptides , and in certain annexins . These proteins are usually described as peripheral as one of their conformational states is water-soluble or only loosely associated with a membrane.

Membrane binding mechanisms



Bee venom phospholipase A2 (1poc). Middle plane of the lipid bilayer - black dots. Boundary of the hydrocarbon core region - red dots (extracellular side). Layer of lipid phosphates - yellow dots.

The association of a protein with a lipid bilayer may involve significant changes within tertiary structure of a protein. These may include the folding of regions of protein structure that were previously unfolded or a re-arrangement in the folding or a refolding of the membrane-associated part of the proteins . It also may involve the formation or dissociation of protein quaternary structures or oligomeric complexes, and specific binding of ions, ligands, or regulatory lipids.

Typical amphitropic proteins must interact strongly with the lipid bilayer in order to perform their biological functions. These include the enzymatic processing of lipids and other hydrophobic substances, membrane anchoring, and the binding and transfer of small nonpolar compounds between different cellular membranes. These proteins may be anchored to the bilayer as a result of hydrophobic interactions between the bilayer and exposed nonpolar residues at the surface of a protein, by specific non-covalent binding interactions with regulatory lipids , or through their attachment to covalently-bound lipid anchors.

It has been shown that the membrane binding affinities of many peripheral proteins depend on the specific lipid composition of the membrane with which they are associated.

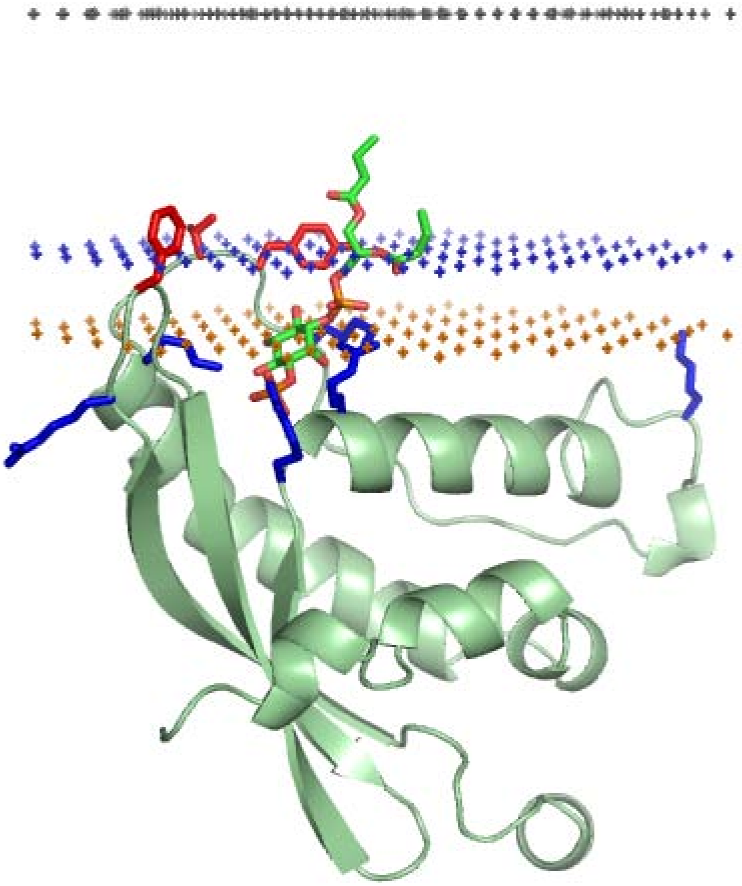
Non-specific hydrophobic association

Amphitropic proteins associate with lipid bilayers via various hydrophobic anchor structures. Such as amphiphilic α-helixes, exposed nonpolar loops, post-translationally acylated or lipidated amino acid residues, or acyl chains of specifically bound regulatory lipids such as phosphatidylinositol phosphates. Hydrophobic interactions have been shown to be important even for highly cationic peptides and proteins, such as the polybasic domain of the MARCKS protein or histactophilin, when their natural hydrophobic anchors are present.

Covalently bound lipid anchors

Lipid anchored proteins are covalently attached to different fatty acid acyl chains on the cytoplasmic side of the cell membrane via palmitoylation, myristoylation, or prenylation. At the cell surface, on the opposite side of the cell membrane lipid anchored proteins are covalently attached to the lipids glycosylphosphatidylinositol (GPI) and cholesterol. Protein association with membranes through the use of acylated residues is a reversible process, as the acyl chain can be buried in a protein's hydrophobic binding pocket after dissociation from the membrane. This process occurs within the beta-subunits of G-proteins . Perhaps because of this additional need for structural flexibility, lipid anchors are usually bound to the highly flexible segments of proteins tertiary structure that are not well resolved by protein crystallographic studies.

Specific protein-lipid binding



P40phox PX domain of NADPH oxidase Middle plane of the lipid bilayer - black dots. Boundary of the hydrocarbon core region - blue dots (intracellular side). Layer of lipid phosphates - yellow dots.

Some cytosolic proteins are recruited to different cellular membranes by recognizing certain types of lipid found within a given membrane. Binding of a protein to a specific lipid occurs via specific membrane-targeting structural domains that occur within the protein and have specific binding pockets for the lipid head groups of the lipids to which they bind. This is a typical biochemical protein-ligand interaction, and is stabilized by the formation of intermolecular hydrogen bonds, van der Waals interactions, and hydrophobic interactions between the protein and lipid ligand. Such complexes are also stabilized by the formation of ionic bridges between the aspartate or glutamate residues of the protein and lipid phosphates via interveening calcium ions (Ca^{2+}). Such ionic bridges can occur and are stable when ions (such as Ca^{2+}) are already bound to a protein in solution, prior to lipid binding. The formation of ionic bridges is seen in the protein-lipid interaction between both protein C2 type domains and annexins..

Protein-lipid electrostatic interactions

Any positively charged protein will be attracted to a negatively charged membrane by nonspecific electrostatic interactions. However, not all peripheral peptides and proteins are cationic, and only certain sides of membrane are negatively charged. These include the cytoplasmic side of plasma membranes, the outer leaflet of outer bacterial membranes and mitochondrial membranes. Therefore, electrostatic interactions play an important role in membrane targeting of electron carriers such as cytochrome c, cationic toxins such as charybdotoxin, and specific membrane-targeting domains such as some PH domains, C1 domains, and C2 domains.

Electrostatic interactions are strongly dependent on the ionic strength of the solution. These interactions are relatively weak at the physiological ionic strength (0.14M NaCl): ~3 to 4 kcal/mol for small cationic proteins, such as cytochrome c, charybdotoxin or hisactophilin.

Spatial position in membrane

Orientations and penetration depths of many amphitropic proteins and peptides in membranes are studied using site-directed spin labeling, chemical labeling, measurement of membrane binding affinities of protein mutants, fluorescence spectroscopy, solution or solid-state NMR spectroscopy, ATR FTIR spectroscopy, X-ray or neutron diffraction, and computational methods.

Two distinct membrane-association modes of proteins have been identified. Typical water-soluble proteins have no exposed nonpolar residues or any other hydrophobic anchors. Therefore, they remain completely in aqueous solution and do not penetrate into the lipid bilayer, which would be energetically costly. Such proteins interact with bilayers only electrostatically, for example, ribonuclease and poly-lysine interact with membranes in this mode. However, typical amphitropic proteins have various hydrophobic anchors that penetrate the interfacial region and reach the hydrocarbon interior of the membrane. Such proteins "deform" the lipid bilayer, decreasing the temperature of lipid fluid-gel transition. The binding is usually a strongly exothermic reaction. Association of amphiphilic α-helices with membranes occurs similarly. Intrinsically unstructured or unfolded peptides with nonpolar residues or lipid anchors can also penetrate the interfacial region of the membrane and reach the hydrocarbon core, especially when such peptides are cationic and interact with negatively charged membranes.

Categories of peripheral proteins

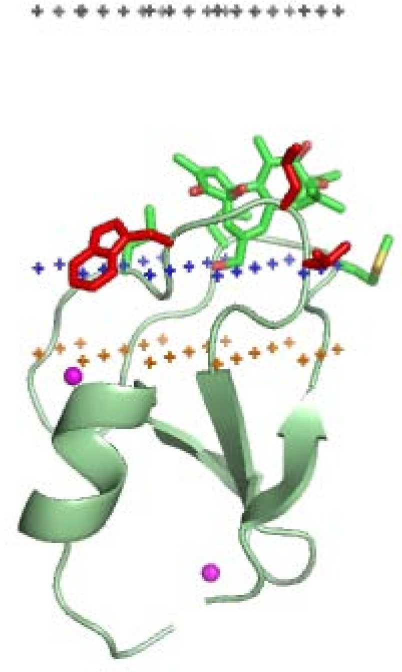
Enzymes

Peripheral enzymes participate in metabolism of different membrane components, such as lipids (phospholipases and cholesterol oxidases), cell wall oligosaccharides (glycosyltransferase and transglycosidases), or proteins (signal peptidase and palmitoyl protein thioesterases). Lipases can also digest lipids that form micelles or nonpolar droplets in water.

Class	Function	Physiology	Structure
Alpha/beta hydrolase fold	Catalyzes the hydrolysis of chemical bonds.	Includes bacterial, fungal, gastric and pancreatic lipases, palmitoyl protein thioesterases, cutinase, and cholinesterases	
Phospholipase A2 (secretory and cytosolic)	Hydrolysis of sn-2 fatty acid bond of phospholipids.	Lipid digestion, membrane disruption, and lipid signaling.	
Phospholipase C	Hydrolyzes PIP2, a phosphatidylinositol, into two second messagers, inositol triphosphate and diacylglycerol.	Lipid signaling	
Cholesterol oxidases	Oxidizes and isomerizes cholesterol to cholest-4-en-3-one.	Depletes cellular membranes of cholesterol, used in bacterial pathogenesis.	
Carotenoid oxygenase	Cleaves carotenoids.	Carotenoids function in both plants and animals as hormones (includes vitamin A in humans), pigments, flavors, floral scents and defense compounds.	
Lipoxygenases	Iron-containing enzymes that catalyze the dioxygenation of polyunsaturated fatty acids.	In animals lipoxygenases are involved in the synthesis of inflammatory mediators known as leukotrienes.	
Alpha toxins	Cleave phospholipids in the cell membrane, similar	Bacterial pathogenesis, particularly by	

	to Phospholipase C.	*Clostridium perfringens*.
Sphingomyelinase C	A phosphodiesterase, cleaves phosphodiester bonds.	Processing of lipids such as sphingomyelin.
Glycosyltransferases: MurG and Transglycosidases	Catalyzes the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds.	Biosynthesis of disaccharides, oligosaccharides and polysaccharides (glycoconjugates), MurG is involved in bacterial peptidoglycan biosynthesis.
Ferrochelatase	Converts protoporphyrin IX into heme.	Involved in porphyrin metabolism, protoporphyrins are used to strengthen egg shells.
Myotubularin-related protein family	Lipid phosphatase that dephosphorylates PtdIns3P and PtdIns(3,5)P2.	Required for muscle cell differentiation.
Dihydroorotate dehydrogenases	Oxidation of dihydroorotate (DHO) to orotate.	Biosynthesis of pyrimidine nucleotides in prokaryotic and eukaryotic cells.
Glycolate oxidase	Catalyses the oxidation of α-hydroxy acids to the corresponding α-ketoacids.	In green plants, the enzyme participates in photorespiration. In animals, the enzyme participates in production of oxalate.

Membrane-targeting domains (“lipid clamps”)



C1 domain of PKC-delta (1ptr) Middle plane of the lipid bilayer - black dots. Boundary of the hydrocarbon core region - blue dots (cytoplasmic side). Layer of lipid phosphates - yellow dots.

Membrane-targeting domains associate specifically with head groups of their lipid ligands embedded into the membrane. These lipid ligands are present in different concentrations in distinct types of biological membranes (for example, PtdIns3P can be found mostly in membranes of early endosomes, PtdIns(3,5)P2 in late endosomes, and PtdIns4P in the Golgi). Hence, each domain is targeted to a specific membrane.

- C1 domains bind diacylglycerol and phorbol esters.
- C2 domains bind phosphatidylserine or phosphatidylcholine
- Pleckstrin homology domains , PX domains , and Tubby domains bind different phosphoinositides
- FYVE domains are more specific for PtdIns3P.
- ENTH domains bind PtdIns(3,4)P2 or PtdIns(4,5)P2.
- ANTH domain binds PtdIns(4,5)P2.
- Proteins from ERM (ezrin/radixin/moesin) family bind PtdIns(4,5)P2.

- Other phosphoinositide-binding proteins include phosphotyrosine-binding domain and certain PDZ domains. They bind PtdIns(4,5)P2.
- Discoidin domains of blood coagulation factors
- ENTH, VHS and ANTH domains

Structural domains

Structural domains mediate attachment of other proteins to membranes. Their binding to membranes can be mediated by calcium ions (Ca^{2+}) that form bridges between the acidic protein residues and phosphate groups of lipids, as in annexins or GLA domains.

Class	Function	Physiology	Structure
Annexins	Calcium-dependent intracellular membrane/ phospholipid binding.	Functions include vesicle trafficking, membrane fusion and ion channel formation.	
Synapsin I	Coats synaptic vesicles and binds to several cytoskeletal elements.	Functions in the regulation of neurotransmitter release.	
Synuclein	Unknown cellular function.	Thought to play a role in regulating the stability and/or turnover of the plasma membrane. Associated with both Parkinson's disease and Alzheimer's disease.	
GLA-domains of the coagulation system	Gamma-carboxyglutamate (GLA) domains are responsible for the high-affinity binding of calcium ions.	Involved in function of clotting factors in the blood coagulation cascade.	
Spectrin and α-actinin-2	Found in several cytoskeletal and microfilament proteins.	Maintenance of plasma membrane integrity and cytoskeletal structure.	

Transporters of small hydrophobic molecules

These peripheral proteins function as carriers of non-polar compounds between different types of cell membranes or between membranes and cytosolic protein complexes. The transported substances are phosphatidylinositol, tocopherol, gangliosides, glycolipids, sterol derivatives, retinol, or fatty acids.

- Glycolipid transfer proteins

- Lipocalins including retinol binding proteins and fatty acid-binding proteins
- Polyisoprenoid-binding protein
- Ganglioside GM2 activator proteins
- CRAL-TRIO domain (α-Tocopherol and phosphatidylinositol sec14p transfer proteins)
- Sterol carrier proteins
- Phosphatidylinositol transfer proteins and STAR domains
- Oxysterol-binding protein

Electron carriers

These proteins are involved in electron transport chains. They include cytochrome c, cupredoxins, high potential iron protein, adrenodoxin reductase, some flavoproteins, and others.

Polypeptide hormones, toxins, and antimicrobial peptides

Many hormones, toxins, inhibitors, or antimicrobial peptides interact specifically with transmembrane protein complexes. They can also accumulate at the lipid bilayer surface, prior to binding their protein targets. Such polypeptide ligands are often positively charged and interact electrostatically with anionic membranes.

Some water-soluble proteins and peptides can also form transmembrane channels. They usually undergo oligomerization, significant conformational changes, and associate with membranes irreversibly. 3D structure of one such transmembrane channel, α-hemolysin, has been determined. In other cases, the experimental structure represents a water-soluble conformation that interacts with the lipid bilayer peripherally, although some of the channel-forming peptides are rather hydrophobic and therefore were studied by NMR spectroscopy in organic solvents or in the presence of micelles.

Class	Proteins	Physiology
Venom toxins	• Scorpion venom • Snake venom • Conotoxins • Poneratoxin (insect)	Well known types of biotoxins include neurotoxins, cytotoxins, hemotoxins and necrotoxins. Biotoxins have two primary functions: predation (snake, scorpion and cone snail toxins) and defense (honeybee and ant toxins).
Sea anemone toxins	• Sea anemone sodium channel inhibitory toxin • Neurotoxin III • Cytolysins	Inhibition of sodium and potassium channels and membrane pore formation are the primary actions of over 40 known Sea anemone peptide toxins. Sea anemone are carnivorous animals and use toxins in predation and defense; anemone toxin is of similar toxicity as the most toxic organophosphate chemical warfare agents.

Bacterial toxins	• Perfringolysin O • Botulinum toxin B • Heat-stable enterotoxin B • δ-Endotoxins • Bacteriocins • Lantibiotic peptides • Gramicidin S	Microbial toxins are the primary virulence factors for a variety of pathogenic bacteria. Some toxins, are Pore forming toxins that lyse cellular membranes. Other toxins inhibit protein synthesis or activate second messenger pathways causing dramatic alterations to signal transduction pathways critical in maintaining a variety of cellular functions. Several bacterial toxins can act directly on the immune system, by acting as superantigens and causing massive T cell proliferation, which overextends the immune system. Botulinum toxin is a neurotoxin that prevents neuro-secretory vesicles from docking/fusing with the nerve synapse plasma membrane, inhibiting neurotransmitter release.
Fungal Toxins	• Cyclic lipopeptide antibiotics Surfactin and daptomycin • Peptaibols	These peptides are characterized by the presence of an unusual amino acid, α-aminoisobutyric acid, and exhibit antibiotic and antifungal properties due to their membrane channel-forming activities.
Antimicrobial peptides	• HP peptide • Saposin B and NK-lysin • Lactoferricin B • Magainin , Moricins , and Pleurocidin	The modes of action by which antimicrobial peptides kill bacteria is varied and includes disrupting membranes, interfering with metabolism, and targeting cytoplasmic components. In contrast to many conventional antibiotics these peptides appear to be bacteriocidal instead of bacteriostatic.
Defensins	• Insect defensins • Plant defensins : Cyclotides and thionins	Defensins are a type of antimicrobial peptide; and are an important component of virtually all innate host defenses against microbial invasion. Defensins penetrate microbial cell membranes by way of electrical attraction, and form a pore in the membrane allowing efflux, which ultimately leads to the lysis of microorganisms.
Neuronal peptides	• Tachykinin peptides	These proteins excite neurons, evoke behavioral responses, are potent vasodilatators, and are responsible for contraction in many types of smooth muscle.
Apoptosis regulators	• Bcl-2	Members of the Bcl-2 family govern mitochondrial outer membrane permeability. Bcl-2 itself suppresses apoptosis in a variety

of cell types including lymphocytes and neuronal cells.