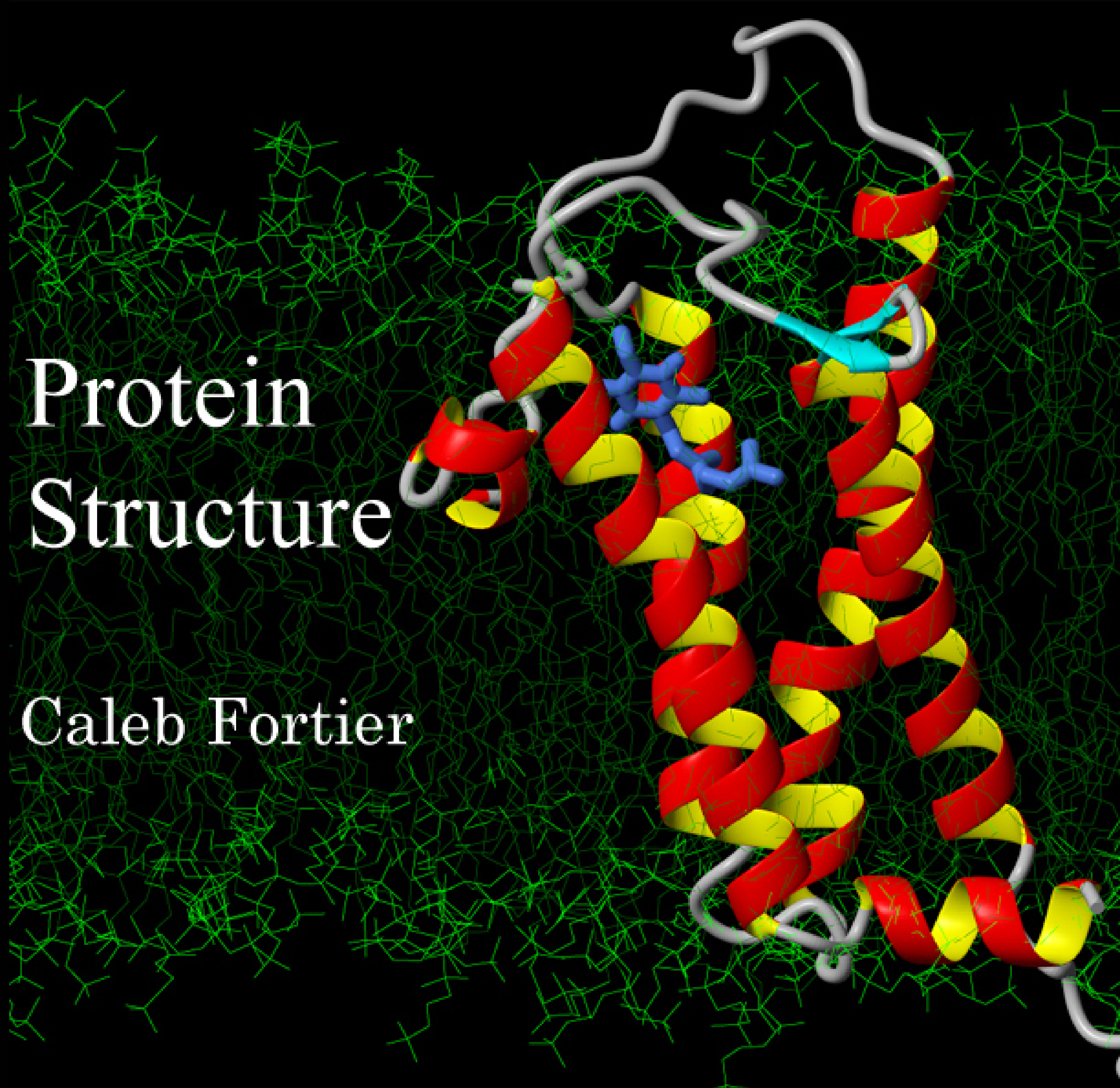


# Protein Structure

Caleb Fortier

First Edition, 2012

ISBN 978-81-323-3208-4

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

# Table of Contents

## Chapter 1

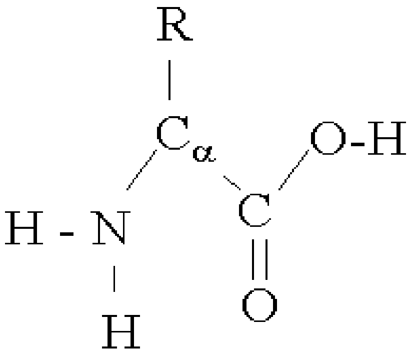
# Protein Structure

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*

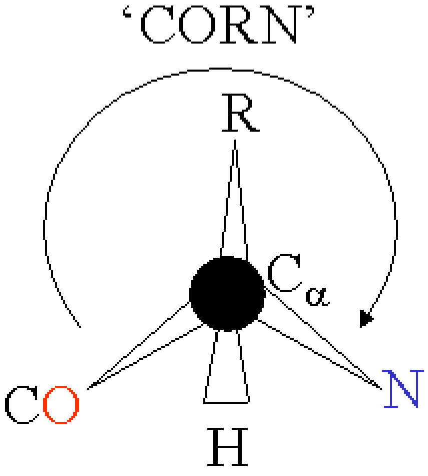


An α-amino acid. The $C_\alpha$H atom is omitted in the diagram.

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

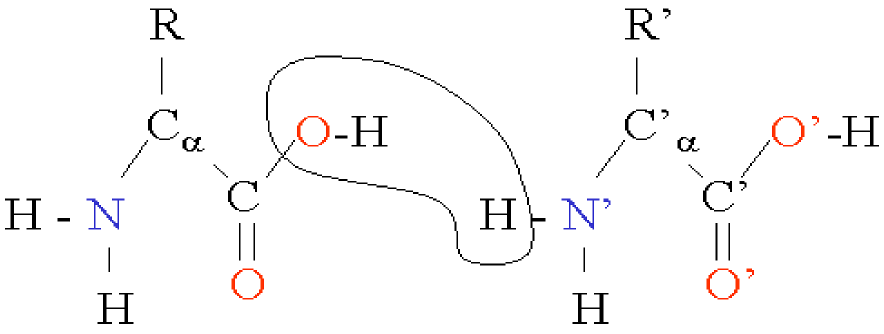
Each α-amino acid consists of a backbone part that is present in all the amino acid types, and a side chain that is unique to each type of residue. An exception from this rule is proline, where the hydrogen atom is replaced by a bond to the side chain. Because the carbon atom is bound to four different 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 $C_\alpha$ atom is viewed with the H in front, the residues read "CO-R-N" in a clockwise direction.



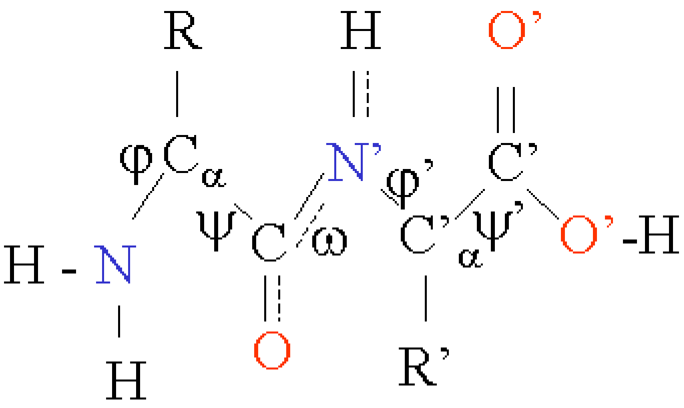
CO-R-N rule

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 |

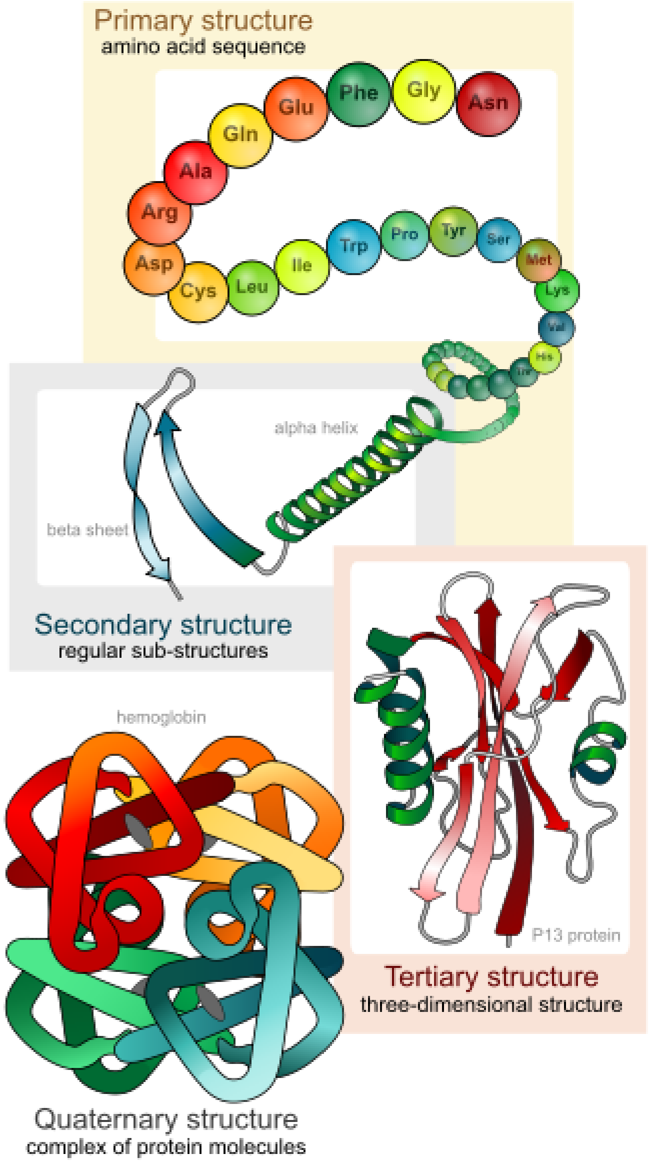
| | | | | | |
|---|---|---|---|---|---|
| N - Cα | 146 pm | C - O | 143 pm | O-H --- O=C | 280 pm |

## 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. $C_\alpha$ refers to the carbon atom of the backbone closest to the carbonyl group of that amino acid, $C_\beta$ 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*



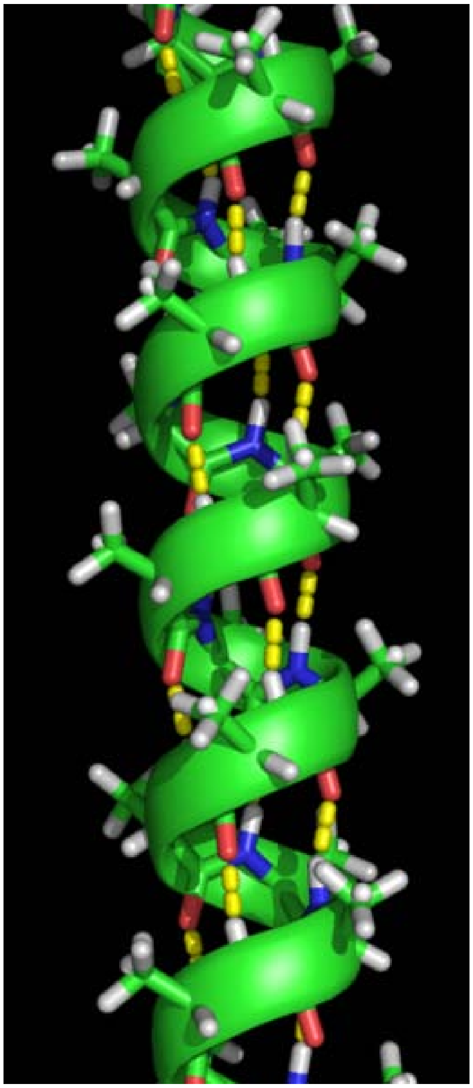
**Protein structure**, from primary to quaternary structure.

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

# 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).

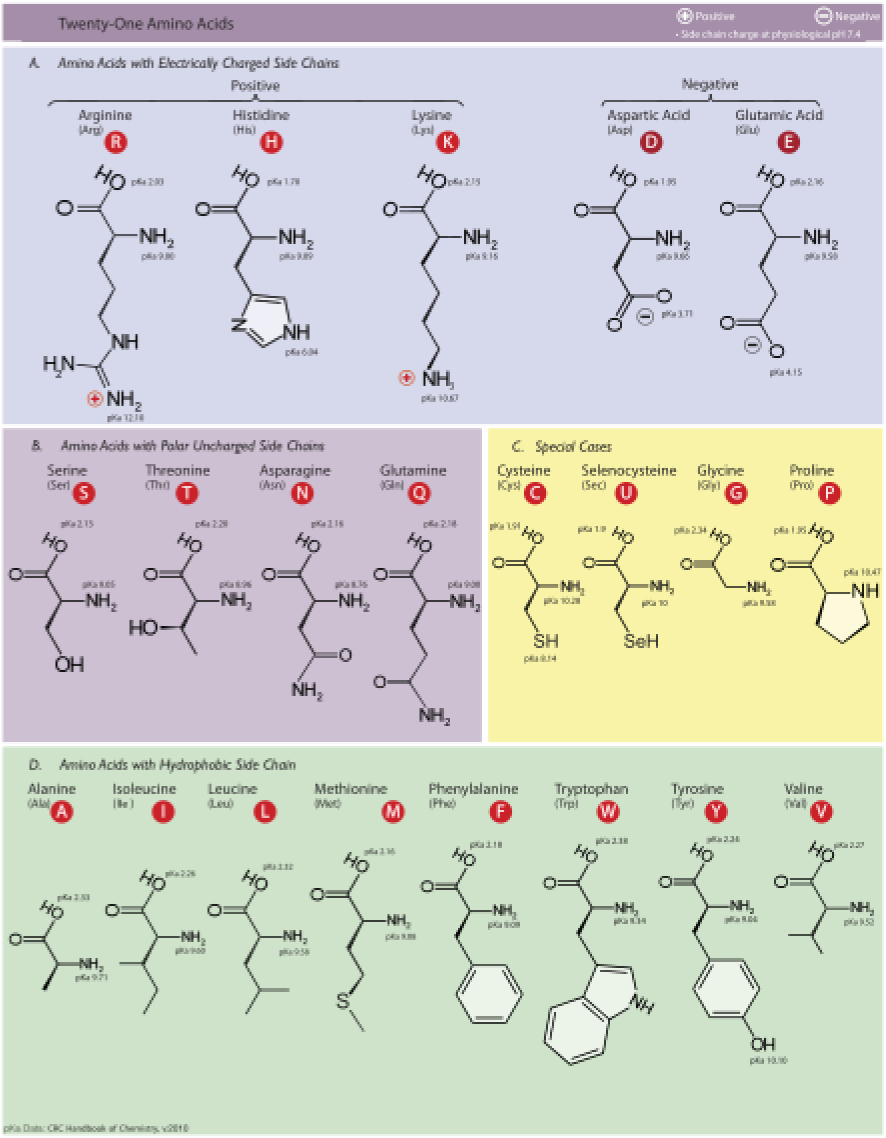
**Non-proteinogenic** amino acids are either not found in proteins (like carnitine, GABA, or L-DOPA), or are not produced directly and in isolation by standard cellular machinery (like hydroxyproline and selenomethionine). The latter often results from posttranslational modification of proteins.

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*

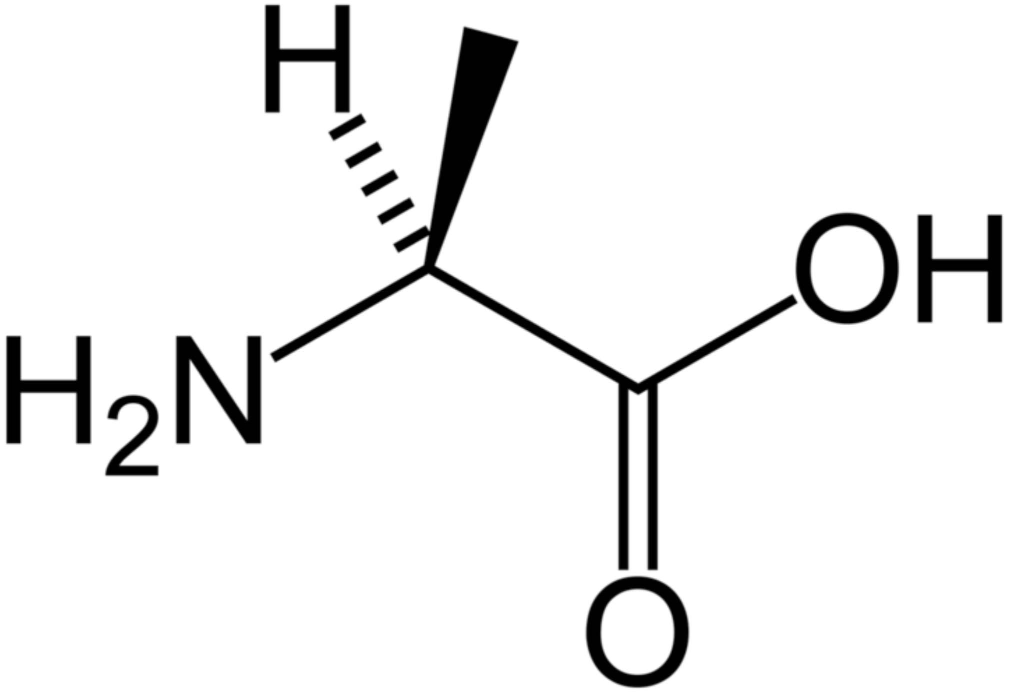
The following illustrates the structures and abbreviations of the 21 amino acids that are directly encoded for protein synthesis by the genetic code of eukaryotes. The structures given below are standard chemical structures, not the typical zwitterion forms that exist in aqueous solutions.



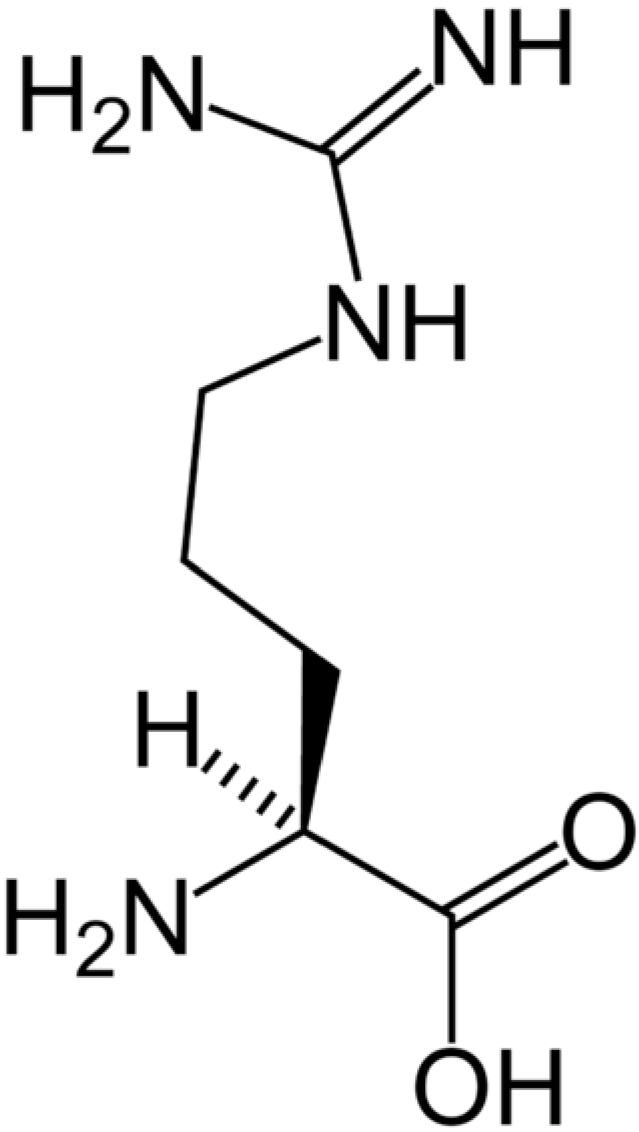
pKa Data: CRC Handbook of Chemistry, v2010

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

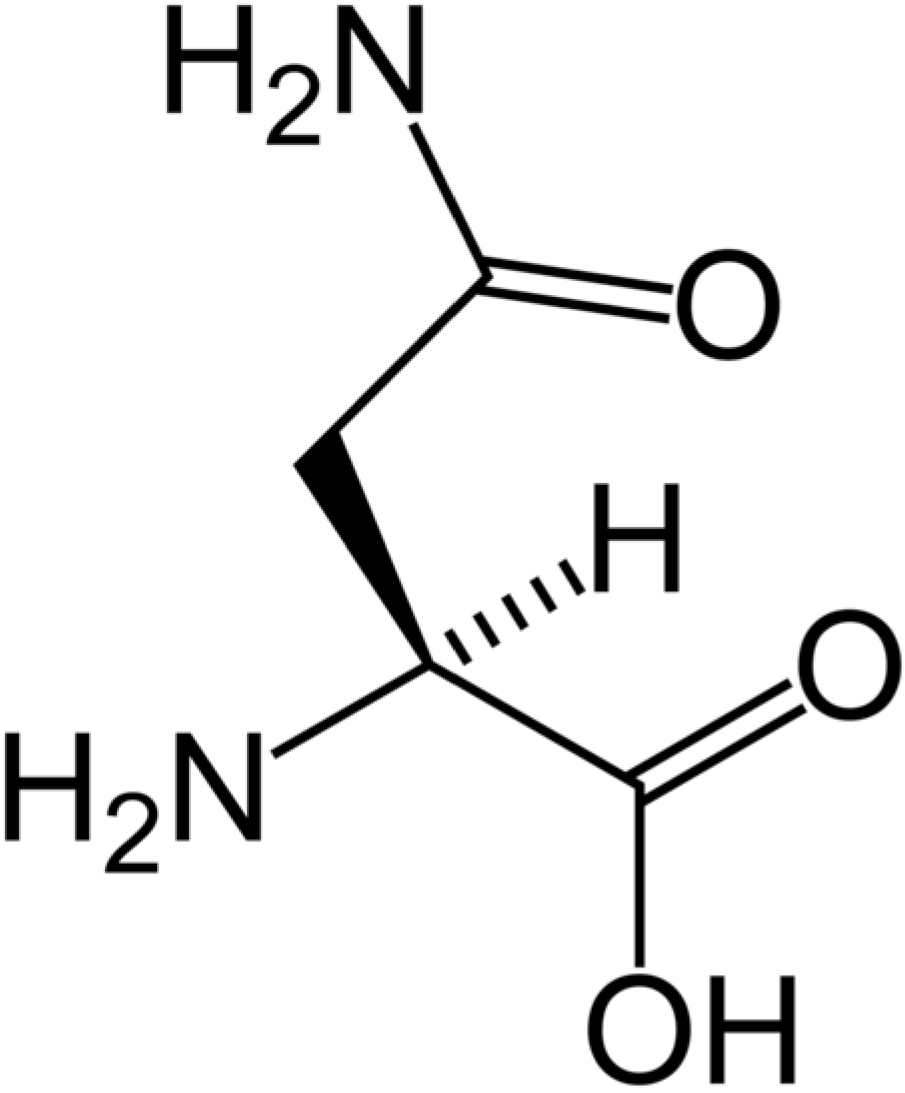
Grouped table of 21 amino acids' structures, nomenclature, and their side groups' pKa's.



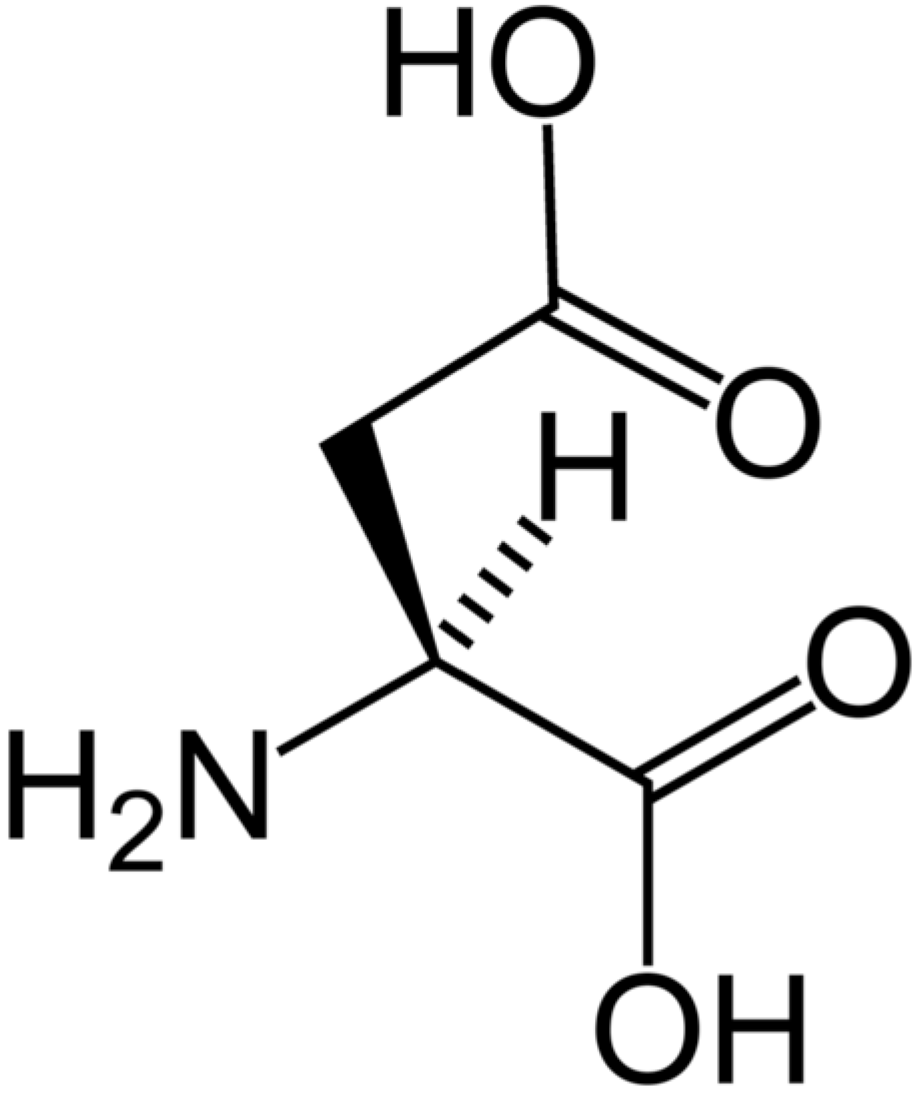
L-Alanine
(Ala / A)



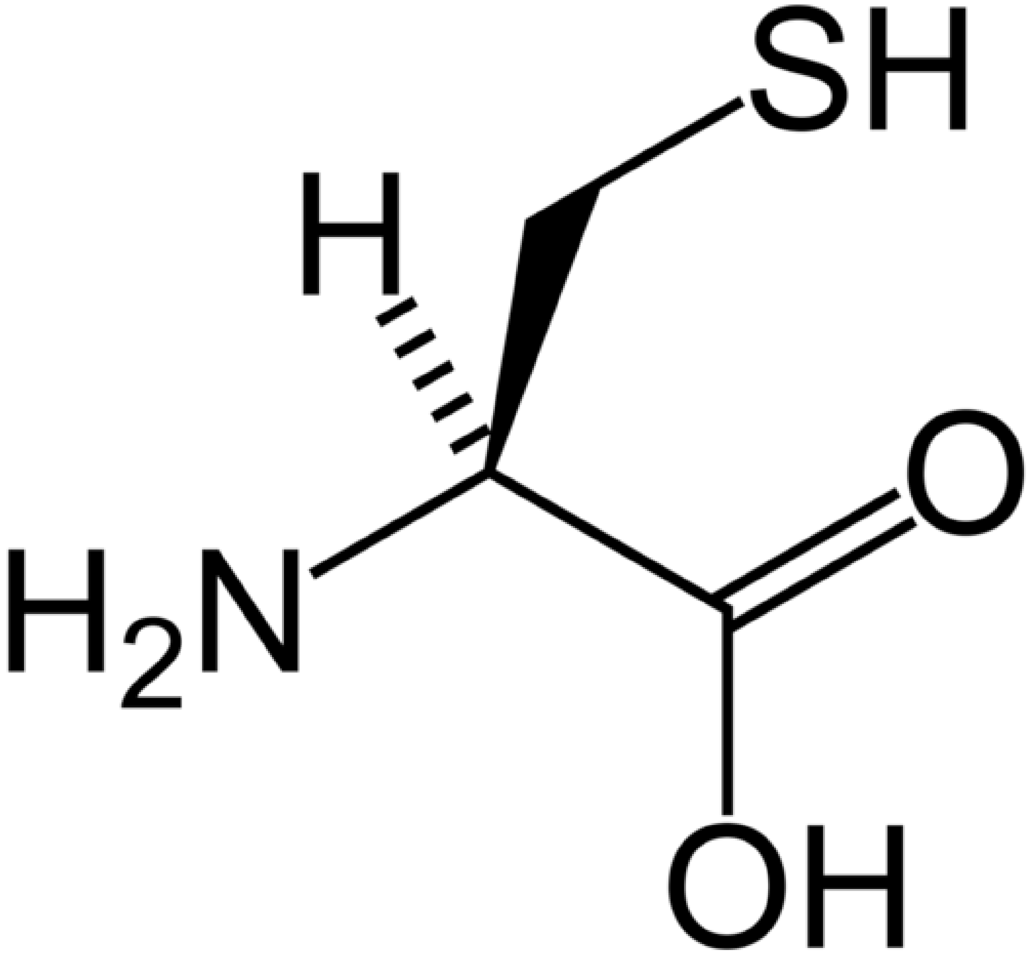
L-Arginine
(Arg / R)



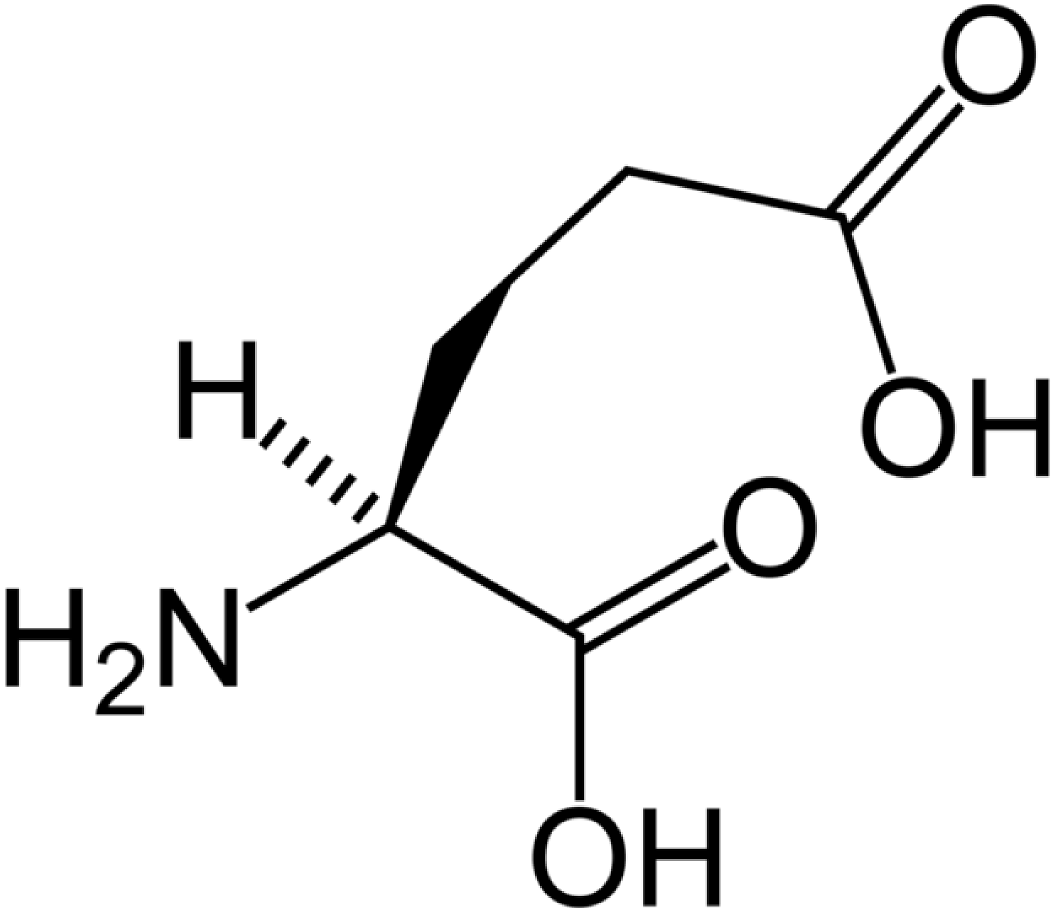
L-Asparagine
(Asn / N)



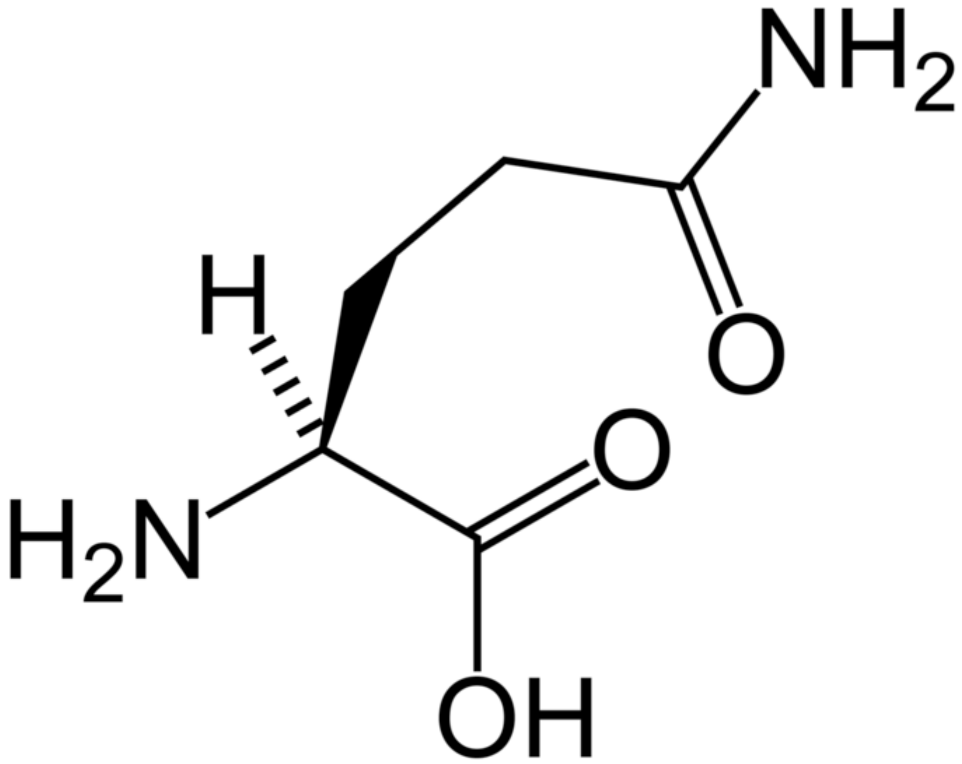
L-Aspartic acid
(Asp / D)



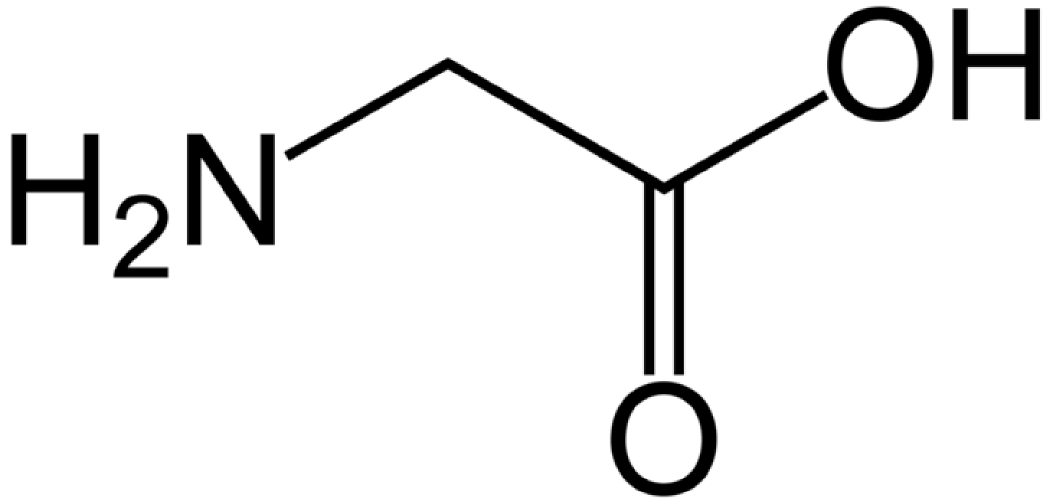
L-Cysteine
(Cys / C)



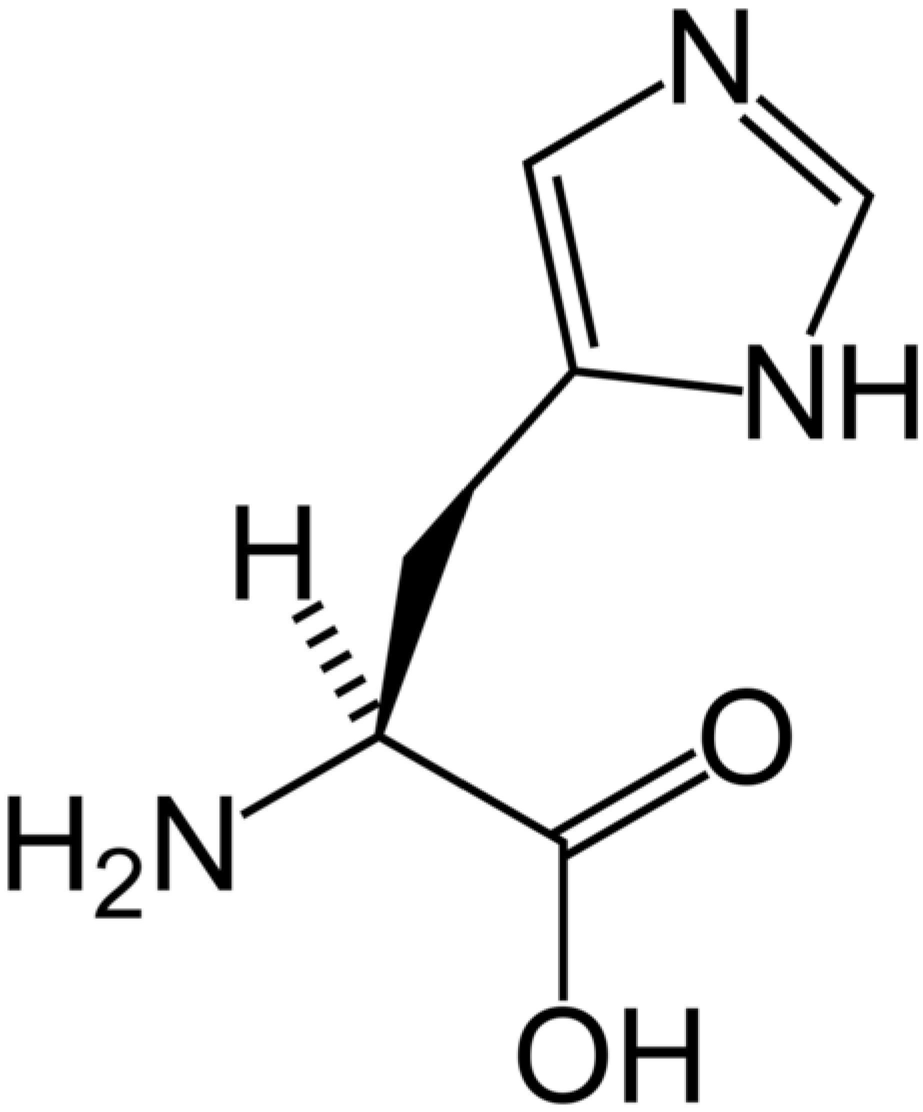
L-Glutamic acid
(Glu / E)



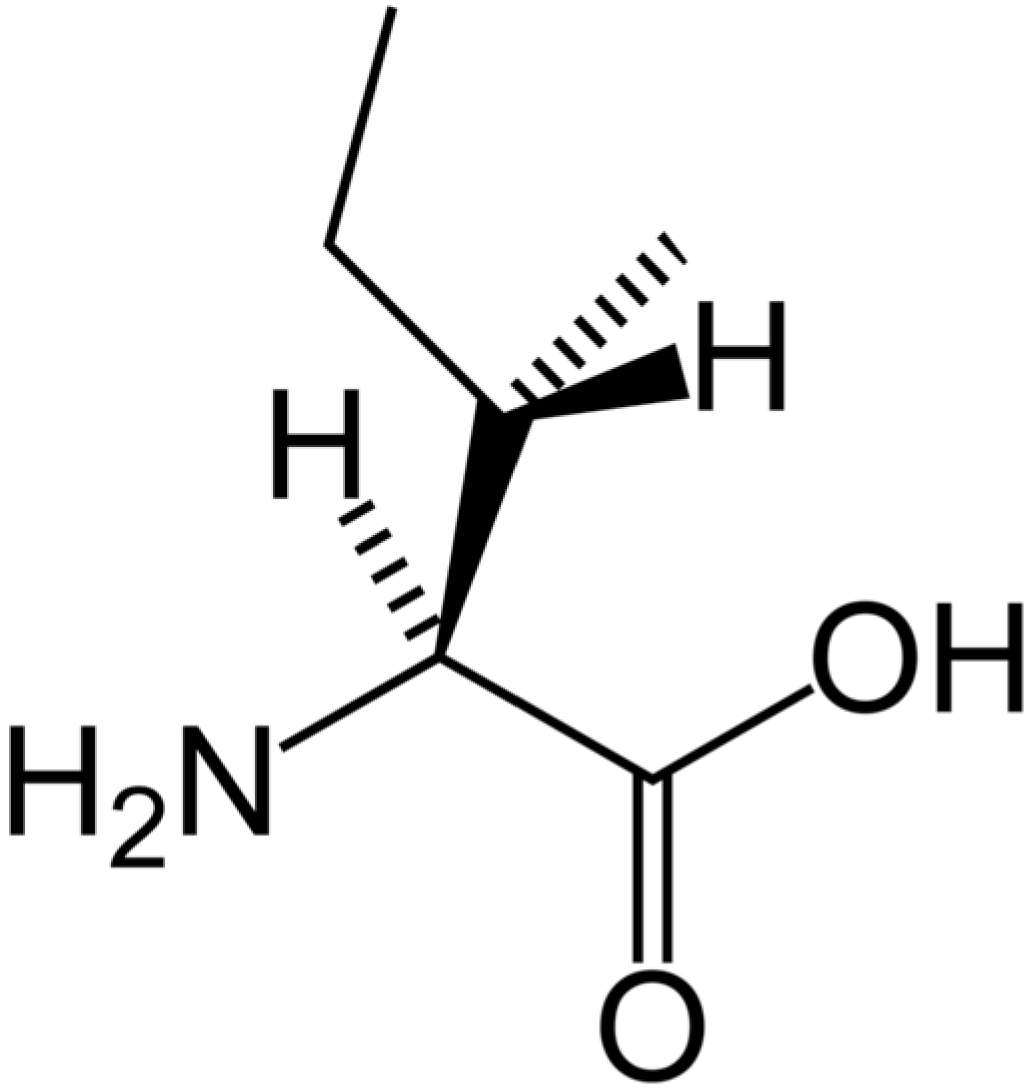
L-Glutamine
(Gln / Q)



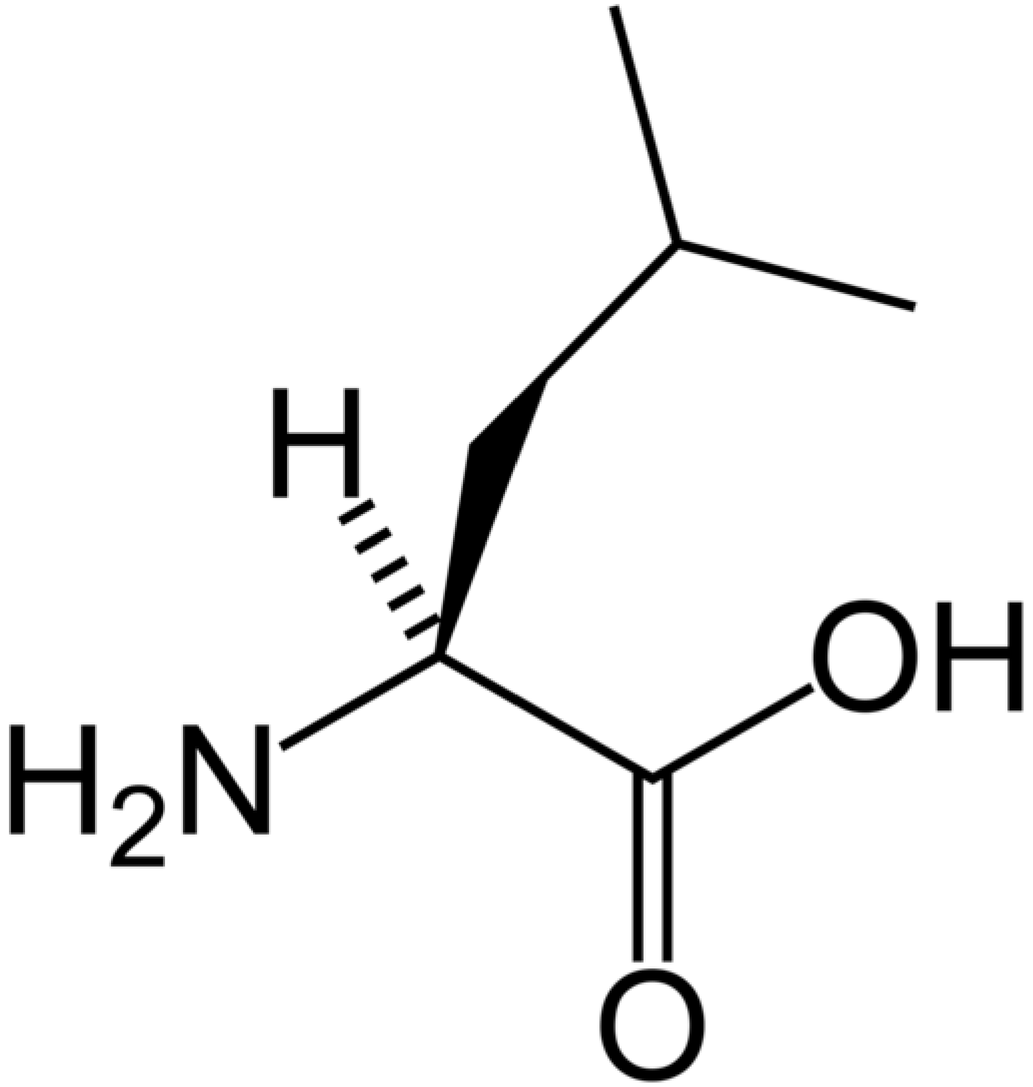
Glycine
(Gly / G)



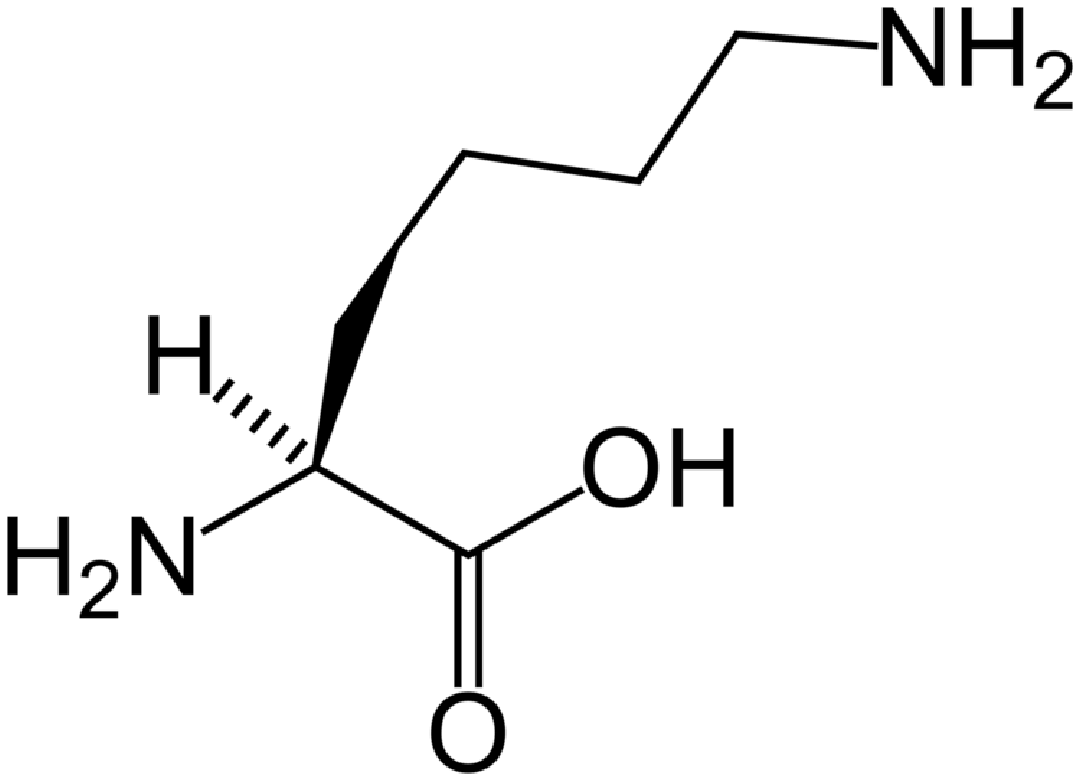
L-Histidine
(His / H)



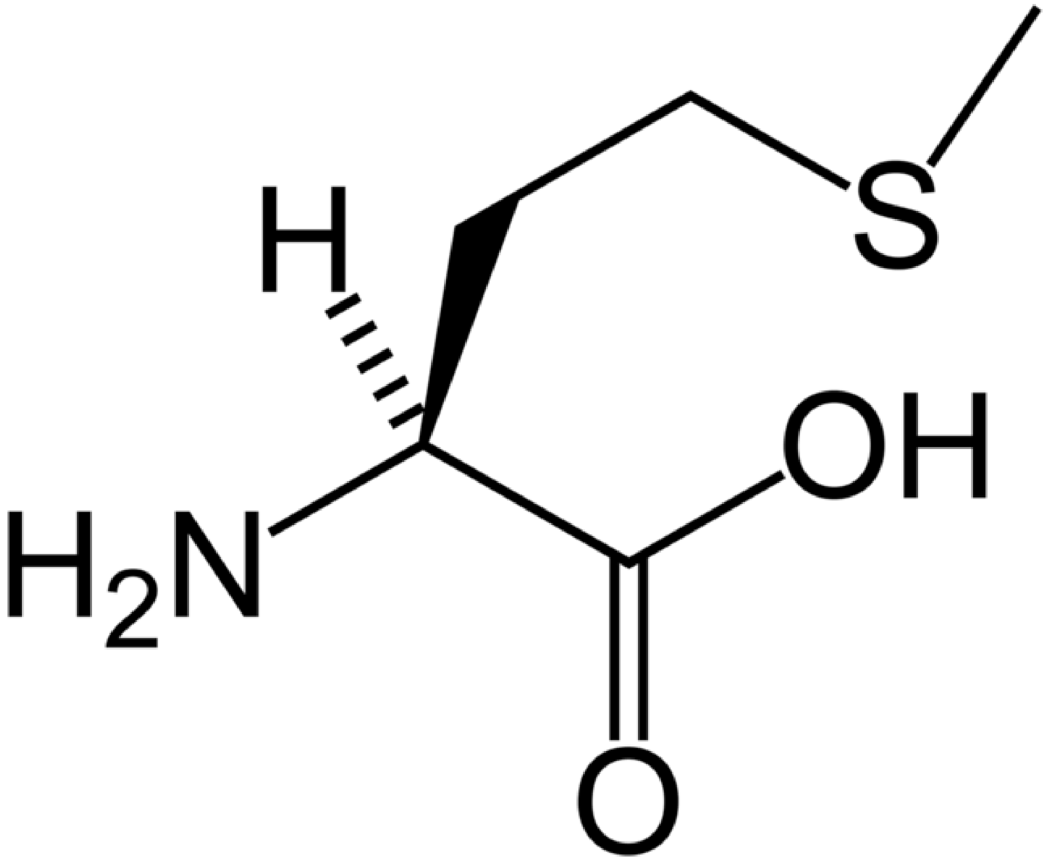
L-Isoleucine
(Ile / I)



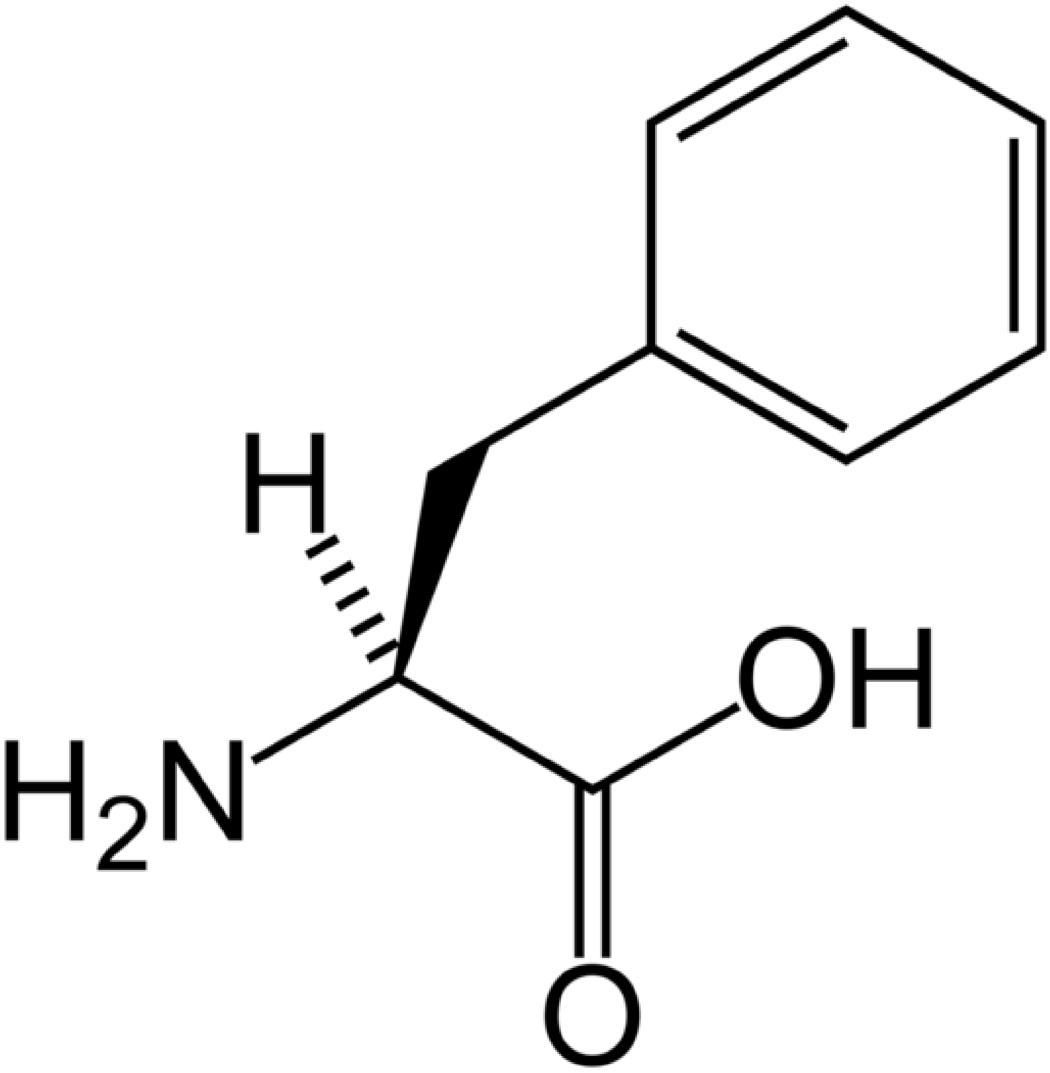
L-Leucine
(Leu / L)



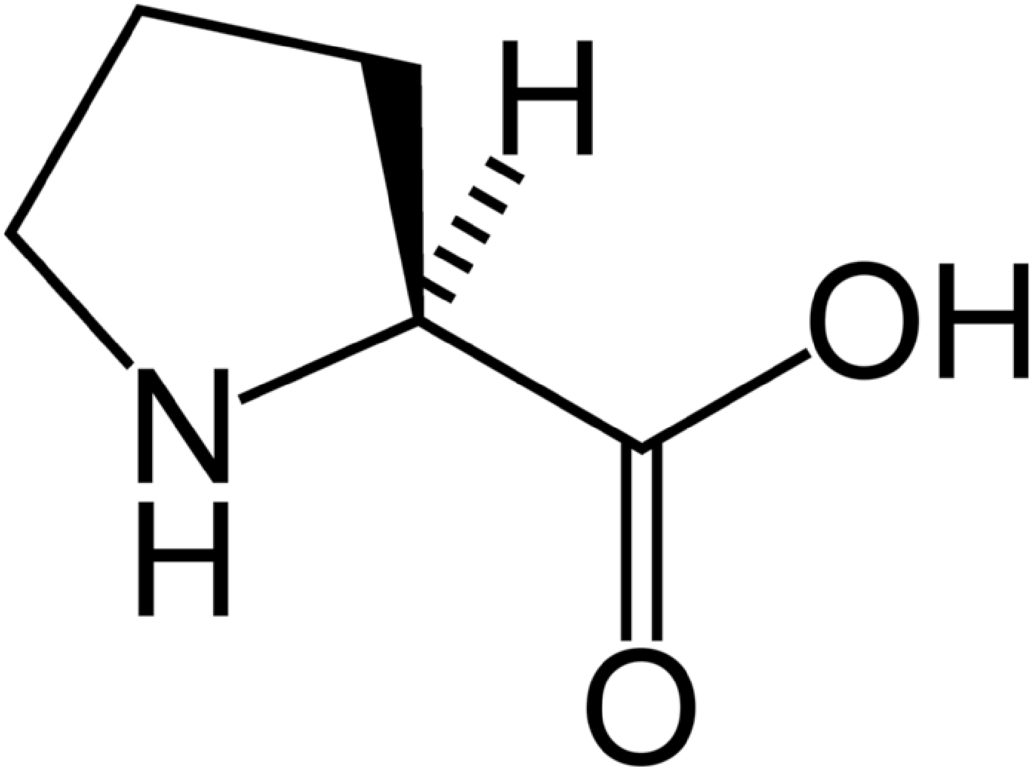
L-Lysine
(Lys / K)



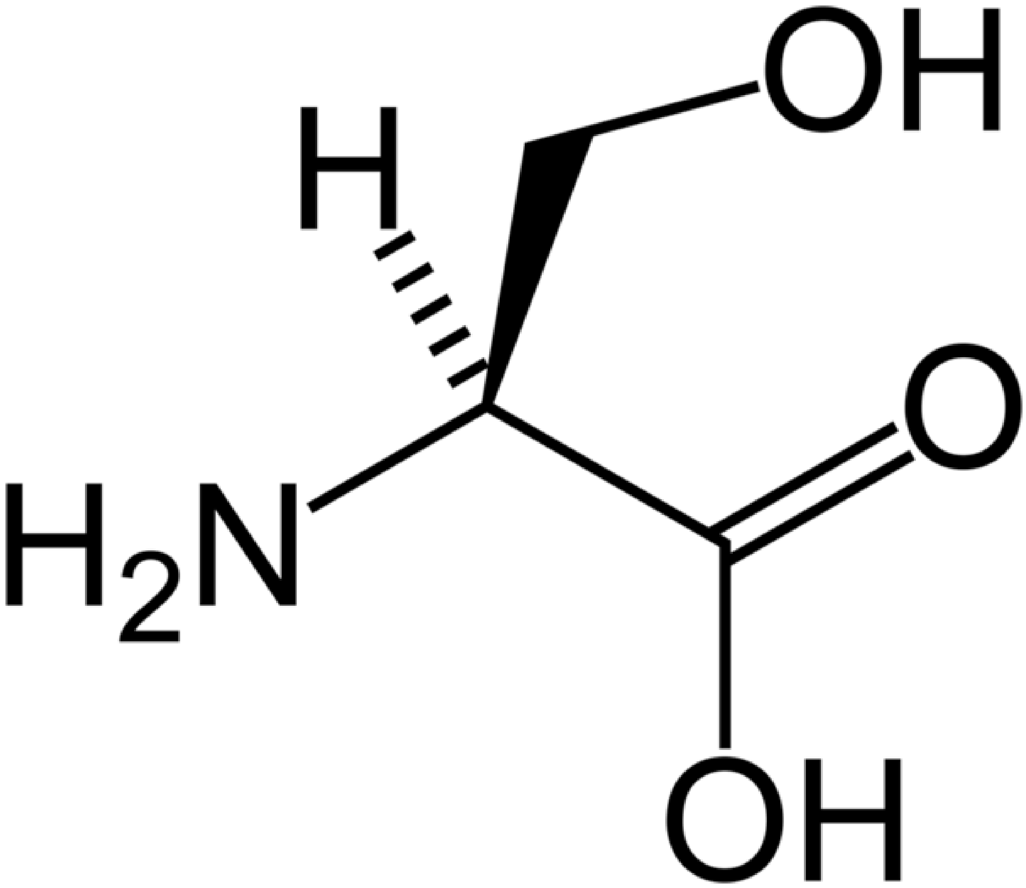
L-Methionine
(Met / M)



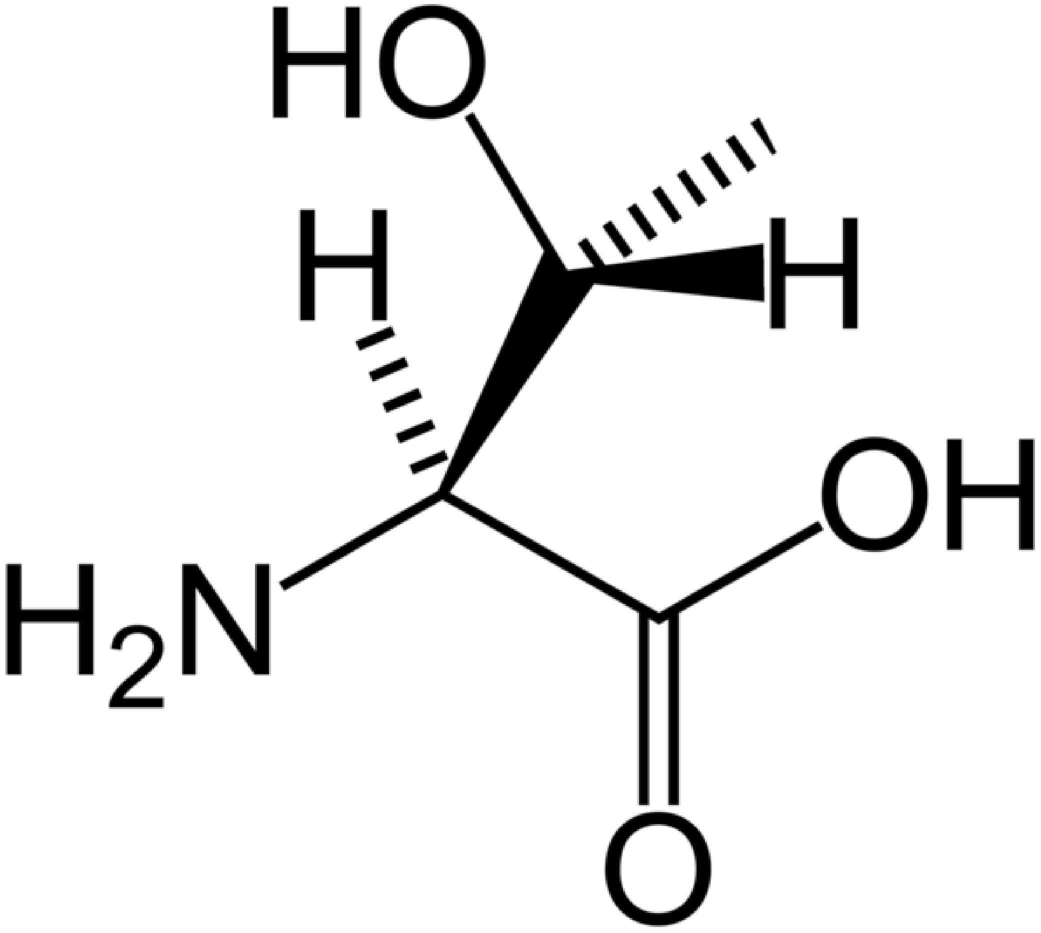
L-Phenylalanine
(Phe / F)



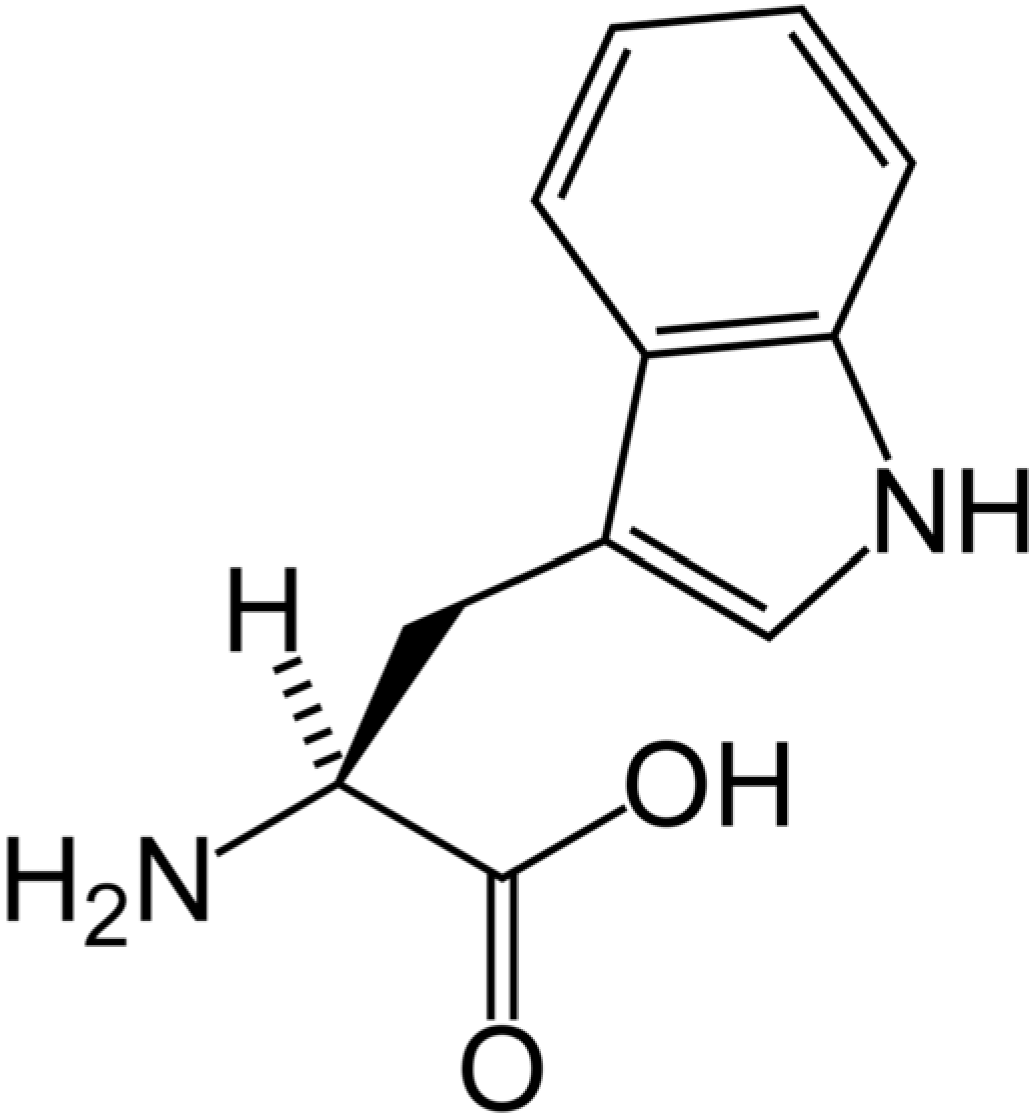
L-Proline
(Pro / P)



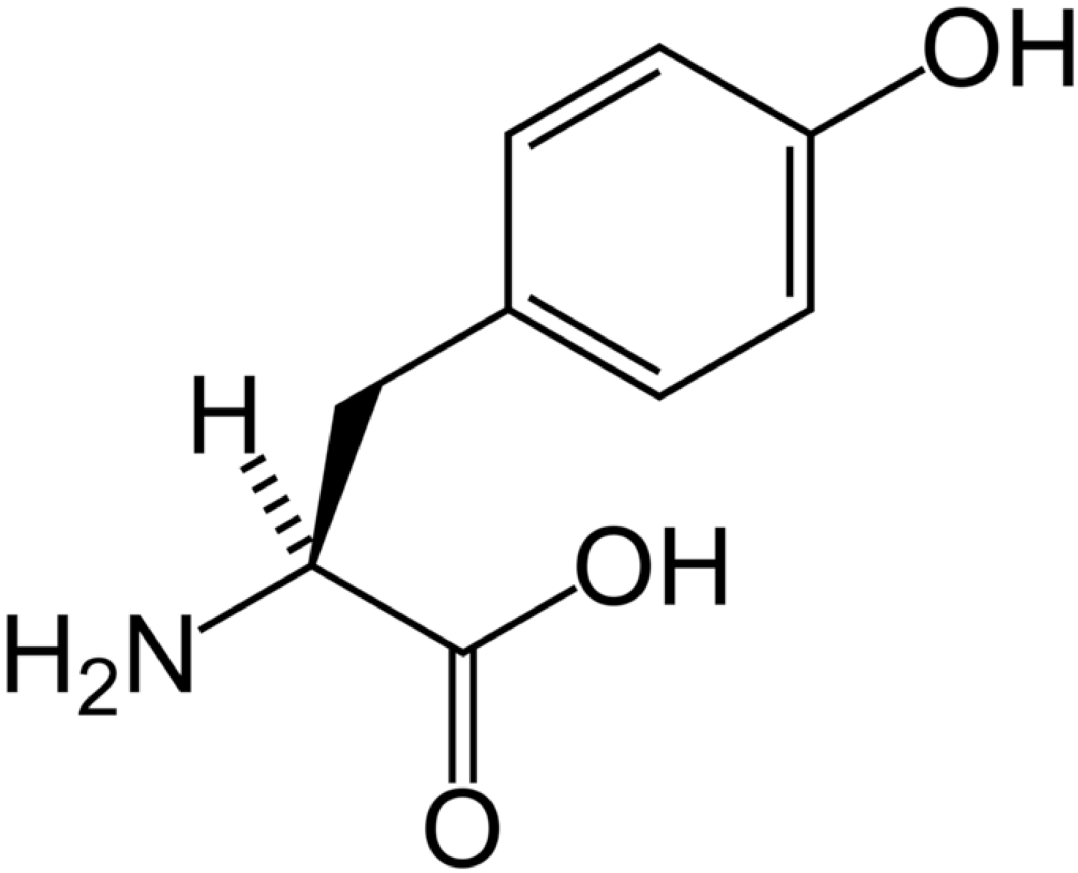
L-Serine
(Ser / S)



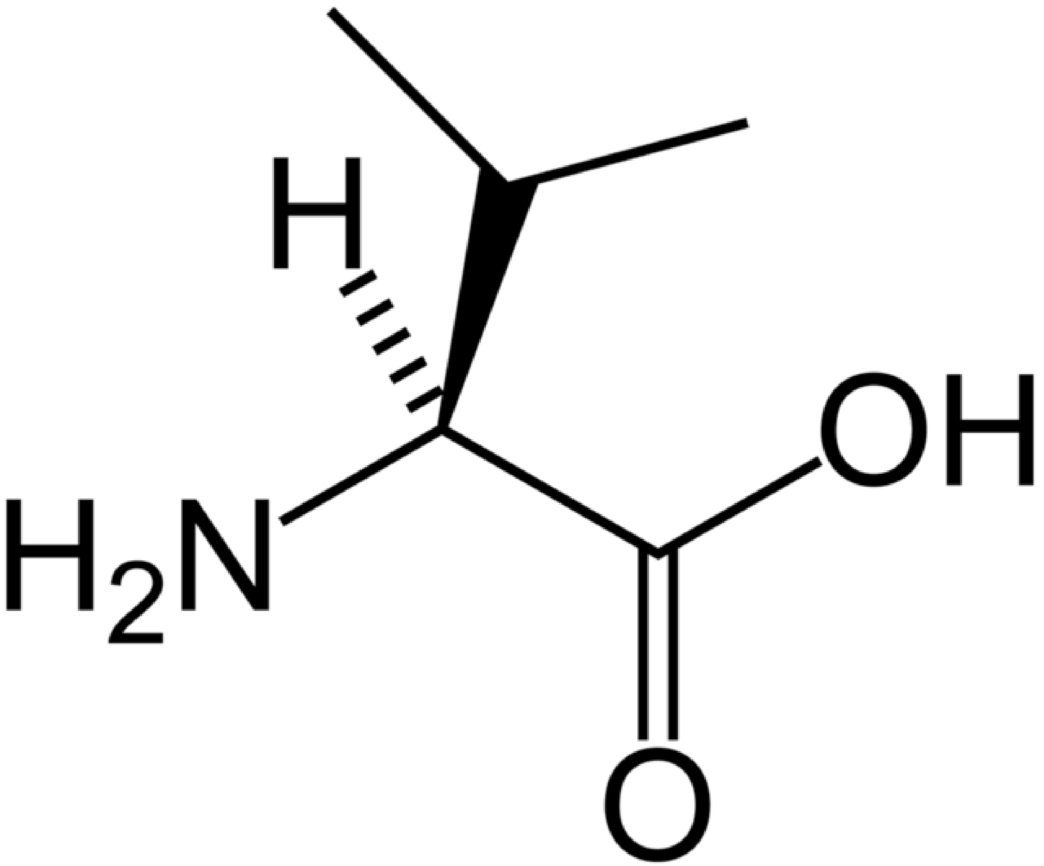
L-Threonine
(Thr / T)



L-Tryptophan
(Trp / W)

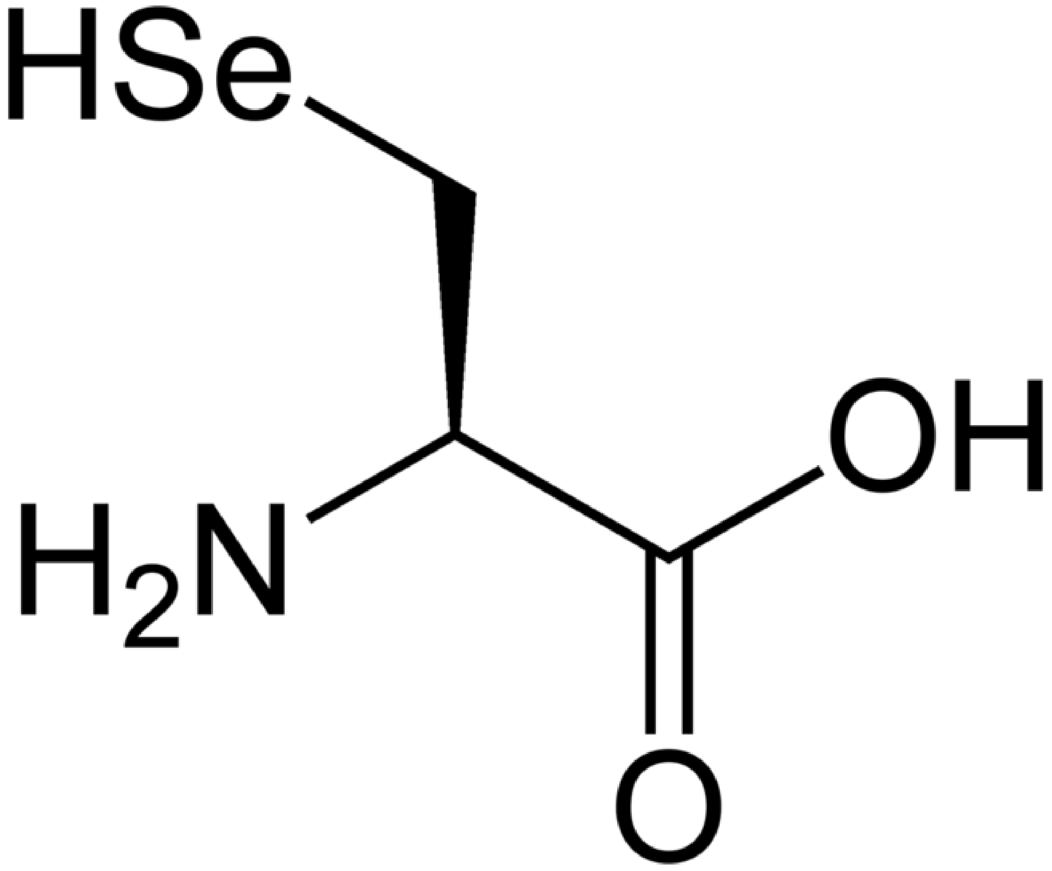


L-Tyrosine
(Tyr / Y)

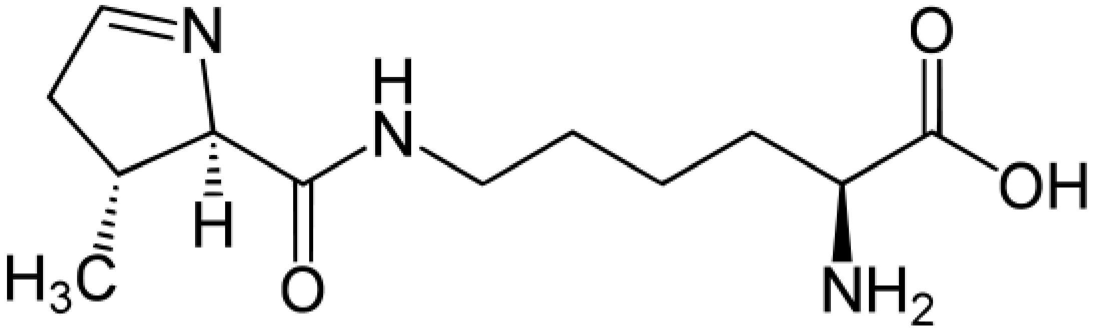


L-Valine
(Val / V)

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



L-Selenocysteine
(Sec / U)



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)** | **$pK_2$ ($\alpha\text{-}^{+}NH_3$)** |
|---|---|---|---|---|---|---|
| **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 | - | - | 12.48 | X | strongl | - | - | - | 148 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $(CH_2)_3NH-C(NH)NH_2$ | | | | y basic | | | | |
| **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).

| | **Abundance** | **ATP cost in synthesis** | **ATP cost in synthesis** |
|---|---|---|---|
| **Amino acid** | **(# of molecules ($\times 10^8$)** | **under aerobic** | **under anaerobic** |
| | **per *E. coli* cell)** | **condition** | **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

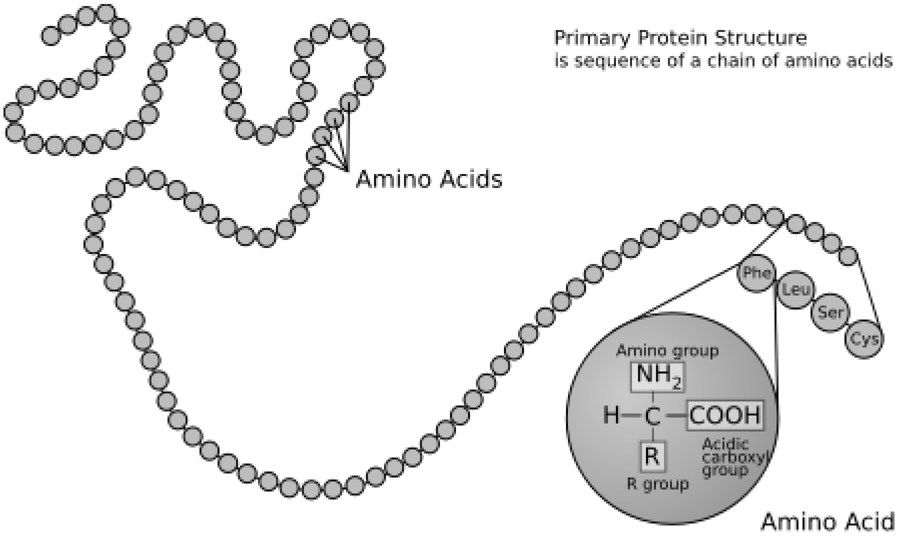
| Amino Acid | Abbrev. |  | Remarks |
|---|---|---|---|
| **Alanine** | A | Ala | Very abundant, very versatile. More stiff than glycine, but small enough to pose only small steric limits for the protein conformation. It behaves fairly neutrally, and can be located in both hydrophilic regions on the 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 3

# Protein Primary Structure



A protein primary structure is a chain of amino acids.

The **primary structure of peptides and proteins** refers to the linear sequence of its amino acid structural units. The term "primary structure" was first coined by Linderstrøm-Lang in 1951. By convention, the primary structure of a protein is reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end.

## *Primary structure of polypeptides*

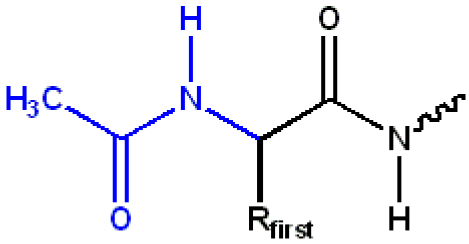
In general, polypeptides are unbranched polymers, so their primary structure can often be specified by the sequence of amino acids along their backbone. However, proteins can become cross-linked, most commonly by disulfide bonds, and the primary structure also requires specifying the cross-linking atoms,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

The chiral centers of a polypeptide chain can undergo racemization. In particular, the L-amino acids normally found in proteins can spontaneously isomerize at the $C^{\alpha}$ atom to form D-amino acids, which cannot be cleaved by most proteases.

Finally, the protein can undergo a variety of posttranslational modifications, which are briefly summarized here.

The N-terminal amino group of a polypeptide can be modified covalently, e.g.,

- **acetylation** − C( = O) − $CH_3$



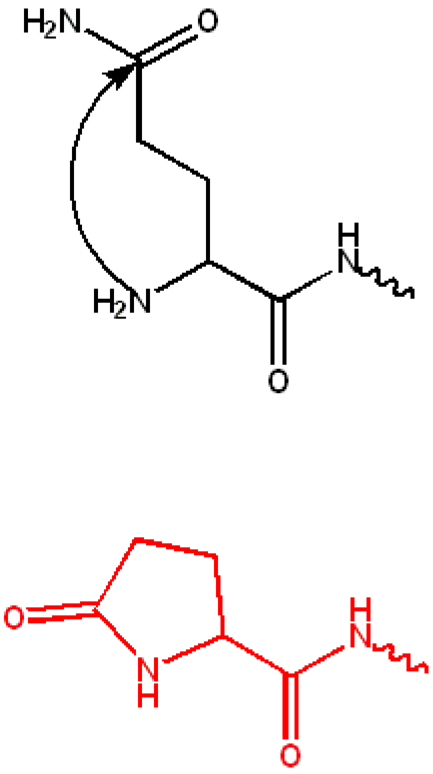
N-terminal acetylation

The positive charge on the N-terminal amino group may be eliminated by changing it to an acetyl group (N-terminal blocking).

- **formylation** − C( = O)H

The N-terminal methionine usually found after translation has an N-terminus blocked with a formyl group. This formyl group (and sometimes the methionine residue itself, if followed by Gly or Ser) is removed by the enzyme deformylase.

- **pyroglutamate**



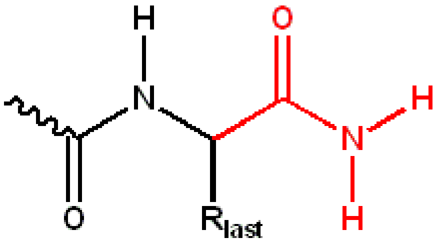
Formation of pyroglutamate from an N-terminal glutamine

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

- **myristoylation** $-\mathrm{C}(=\mathrm{O})-(\mathrm{CH}_2)_{12}-\mathrm{CH}_3$

  Similar to acetylation. Instead of a simple methyl group, the myristoyl group has a tail of 14 hydrophobic carbons, which make it ideal for anchoring proteins to cellular membranes.

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



C-terminal amidation

it

- **amidation**

  The C-terminus can also be blocked (thus, neutralizing its negative charge) by amidation.

- **glycosyl phosphatidylinositol (GPI) attachment**

  Glycosyl phosphatidylinositol is a large, hydrophobic phospholipid prosthetic group that achors proteins to cellular membranes. It is attached to the polypeptide C-terminus through an amide linkage that then connects to ethanolamine, thence to sundry sugars and finally to the phosphatidylinositol lipid moiety.

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

- **phosphorylation**

  Aside from cleavage, phosphorylation is perhaps the most important chemical modification of proteins. A phosphate group can be attached to the sidechain hydroxyl group of serine, threonine and tyrosine residues, adding a negative charge at that site and producing an unnatural amino acid. Such reactions are catalyzed by **kinases** and the reverse reaction is catalyzed by phosphatases. The phosphorylated tyrosines are often used as "handles" by which proteins can bind to one another, whereas phosphorylation of Ser/Thr often induces conformational changes, presumably because of the introduced negative charge. The effects of phosphorylating Ser/Thr can sometimes be simulated by mutating the Ser/Thr residue to glutamate.

- **glycosylation**

  A catch-all name for a set of very common and very heterogeneous chemical modifications. Sugar moieties can be attached to the sidechain hydroxyl groups of Ser/Thr or to the sidechain amide groups of Asn. Such attachments can serve many functions, ranging from increasing solubility to complex recognition. All glycosylation can be blocked with certain inhibitors, such as tunicamycin.

- **deamidation** (succinimide formation)

  In this modification, an asparagine or aspartate side chain attacks the following peptide bond, forming a symmetrical succinimide intermediate. Hydrolysis of the intermediate produces either asparate or the β-amino acid, iso(Asp). For asparagine, either product results in the loss of the amide group, hence "deamidation".

- **hydroxylation**

  Proline residues may be hydroxylates at either of two atoms, as can lysine (at one atom). Hydroxyproline is a critical component of collagen, which becomes unstable upon its loss. The hydroxylation reaction is catalyzed by an enzyme that requires ascorbic acid (vitamin C), deficiencies in which lead to many connective-tissue diseases such as scurvy.

- **methylation**

  Several protein residues can be methylated, most notably the positive groups of lysine and arginine. Methylation at these sites is used to regulate the binding of proteins to nucleic acids. Lysine residues can be singly, doubly and even triply methylated. Methylation does *not* alter the positive charge on the side chain, however.

- **acetylation**

  Acetylation of the lysine amino groups is chemically analogous to the acetylation of the N-terminus. Functionally, however, the acetylation of lysine residues is used to regulate the binding of proteins to nucleic acids. The cancellation of the positive charge on the lysine weakens the electrostatic attraction for the (negatively charged) nucleic acids.

- **sulfation**

Tyrosines may become sulfated on their $O^{\eta}$ atom. Somewhat unusually, this modification occurs in the Golgi apparatus, not in the endoplasmic reticulum. Similar to phosphorylated tyrosines, sulfated tyrosines are used for specific recognition, e.g., in chemokine receptors on the cell surface. As with phosphorylation, sulfation adds a negative charge to a previously neutral site.

- **prenylation** and **palmitoylation** $-C(=O)-(CH_2)_{14}-CH_3$

The hydrophobic isoprene (e.g., farnesyl, geranyl, and geranylgeranyl groups) and palmitoyl groups may be added to the $S^{\gamma}$ atom of cysteine residues to anchor proteins to cellular membranes. Unlike the GPI and myritoyl anchors, these groups are not necessarily added at the termini.

- **carboxylation**

  A relatively rare modification that adds an extra carboxylate group (and, hence, a double negative charge) to a glutamate side chain, producing a Gla residue. This is used to strengthen the binding to "hard" metal ions such as calcium.

- **ADP-ribosylation**

The large ADP-ribosyl group can be transferred to several types of side chains within proteins, with heterogeneous effects. This modification is a target for the powerful toxins of disparate bacteria, e.g., *Vibrio cholerae*, *Corynebacterium diphtheriae* and *Bordetella pertussis*.

- **ubiquitination** and **SUMOylation**

Various full-length, folded proteins can be attached at their C-termini to the sidechain ammonium groups of lysines of other proteins. Ubiquitin is the most common of these, and usually signals that the ubiquitin-tagged protein should be degraded.

Most of the polypeptide modifications listed above occur *post-translationally*, i.e., after the protein has been synthesized on the ribosome, typically occurring in the endoplasmic reticulum, a subcellular organelle of the eukaryotic cell.

Many other chemical reactions (e.g., cyanylation) have been applied to proteins by chemists, although they are not found in biological systems.

## Modifications of primary structure

In addition to those listed above, the most important modification of primary structure is **peptide cleavage** (See: Protease). Proteins are often synthesized in an inactive precursor form; typically, an N-terminal or C-terminal segment blocks the active site of the protein, inhibiting its function. The protein is activated by cleaving off the inhibitory peptide.

Some proteins even have the power to cleave themselves. Typically, the hydroxyl group of a serine (rarely, threonine) or the thiol group of a cysteine residue will attack the carbonyl carbon of the preceding peptide bond, forming a tetrahedrally bonded intermediate [classified as a hydroxyoxazolidine (Ser/Thr) or hydroxythiazolidine (Cys) intermediate]. This intermediate tends to revert to the amide form, expelling the attacking group, since the amide form is usually favored by free energy, (presumably due to the strong resonance stabilization of the peptide group). However, additional molecular interactions may render the amide form less stable; the amino group is expelled instead, resulting in an ester (Ser/Thr) or thioester (Cys) bond in place of the peptide bond. This chemical reaction is called an N-O acyl shift.

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

- Simple hydrolysis will split the polypeptide chain, where the displaced amino group becomes the new N-terminus. This is seen in the maturation of glycosylasparaginase.

- A β-elimination reaction also splits the chain, but results in a pyruvoyl group at the new N-terminus. This pyruvoyl group may be used as a covalently attached catalytic cofactor in some enzymes, especially decarboxylases such as S-adenosylmethionine decarboxylase {SAMDC) that exploit the electron-withdrawing power of the pyruvoyl group.

- Intramolecular transesterification, resulting in a *branched* polypeptide. In inteins, the new ester bond is broken by an intramolecular attack by the soon-to-be C-terminal asparagine.

- Intermolecular transesterification can transfer a whole segment from one polypeptide to another, as is seen in the Hedgehog protein autoprocessing.

## History of protein primary structure

The proposal that proteins were linear chains of α-amino acids was made nearly simultaneously by two scientists at the same conference in 1902, the 74th meeting of the Society of German Scientists and Physicians, held in Karlsbad. Franz Hofmeister made the proposal in the morning, based on his observations of the biuret reaction in proteins. Hofmeister was followed a few hours later by Emil Fischer, who had amassed a wealth of chemical details supporting the peptide-bond model. For completeness, the proposal that proteins contained amide linkages was made as early as 1882 by the French chemist E. Grimaux.

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

Staudinger faced similar prejudices in the 1920s when he argued that rubber was composed of macromolecules.

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

## Primary structure in other molecules

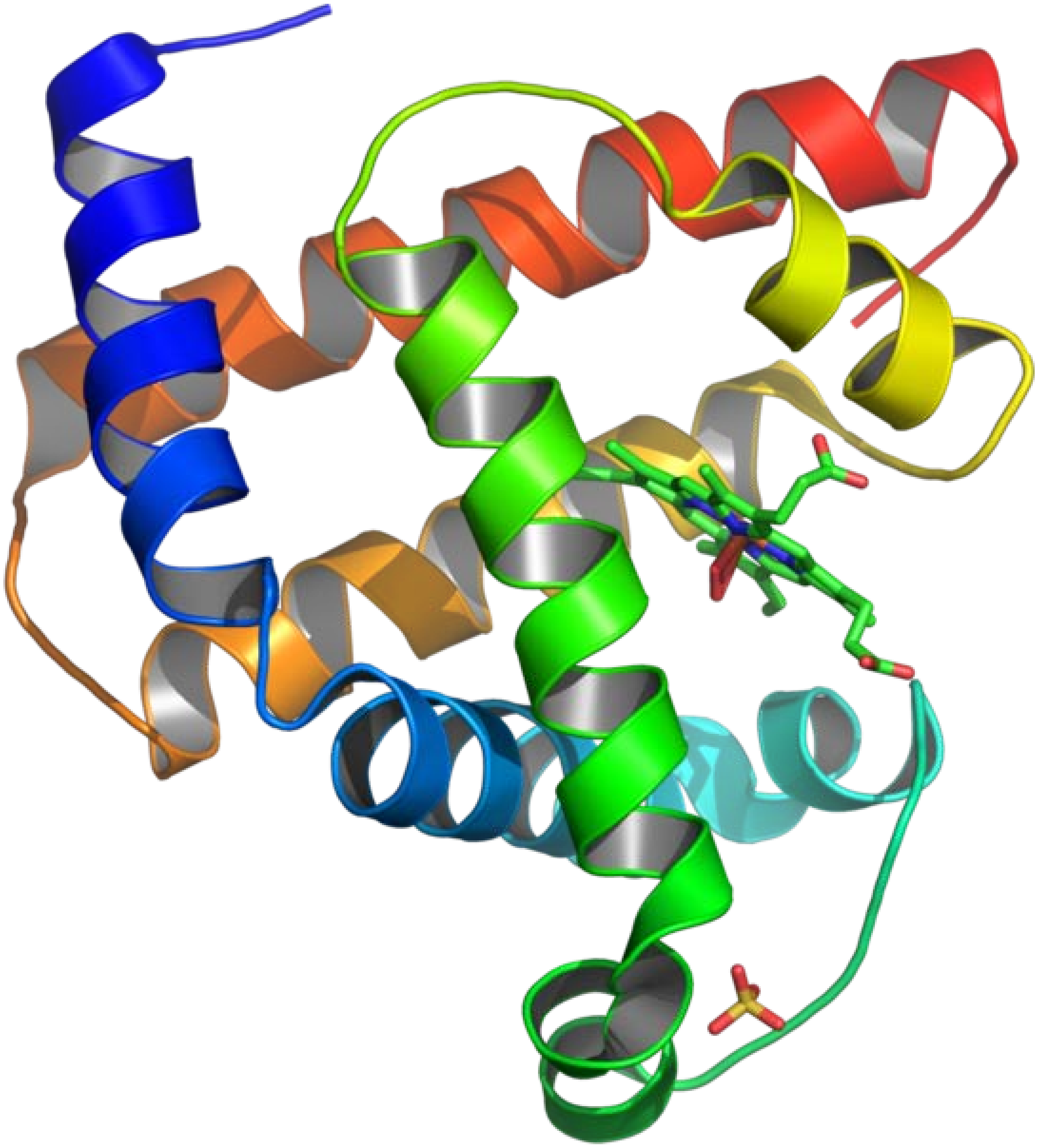
Any linear-chain heteropolymer can be said to have a "primary structure" by analogy to the usage of the term for proteins, but this usage is rare compared to the extremely common usage in reference to proteins. In RNA, which also has extensive secondary structure, the linear chain of bases is generally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

## Relation to secondary and tertiary structure

The primary structure of a biological polymer to a large extent determines the three-dimensional shape known as the tertiary structure, but nucleic acid and protein folding are so complex that knowing the primary structure often doesn't help either to deduce the shape or to predict localized secondary structure, such as the formation of loops or helices. However, knowing the structure of a similar homologous sequence (for example a member of the same protein family) can unambiguously identify the tertiary structure of the given sequence. Sequence families are often determined by sequence clustering, and structural genomics projects aim to produce a set of representative structures to cover the *sequence space* of possible non-redundant sequences.

# Chapter 4

# Protein Secondary Structure



A representation of the 3D structure of the myoglobin protein. Alpha helices are shown in colour, and random coil in white, there are no beta sheets shown. This protein was the first to have its structure resolved by X-ray crystallography by Max Perutz and Sir John Cowdery Kendrew in 1958, which led to them receiving a Nobel Prize in Chemistry in 1962.

In biochemistry and structural biology, **secondary structure** is the general three-dimensional form of *local segments* of biopolymers such as proteins and nucleic acids (DNA/RNA). It does not, however, describe specific atomic positions in three-dimensional space, which are considered to be tertiary structure.

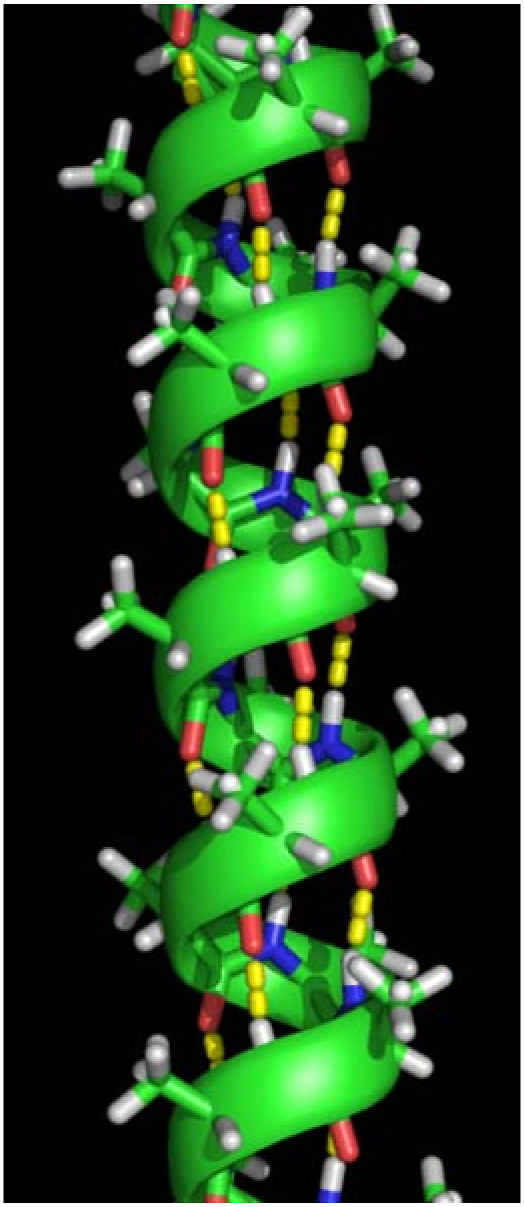
Secondary structure can be formally defined by the hydrogen bonds of the biopolymer, as observed in an atomic-resolution structure. In proteins, the secondary structure is defined by the patterns of hydrogen bonds between backbone amide and carboxyl groups. In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases. The hydrogen bonding patterns may be significantly distorted, which makes an automatic determination of secondary structure difficult.

The secondary structure may be also defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot; thus, a segment of residues with such dihedral angles may be called a helix, regardless of whether it has the correct hydrogen bonds. The secondary structure may be also provided by crystallographers in the corresponding PDB file.

The rough secondary-structure content of a biopolymer (e.g., "this protein is 40% α-helix and 20% β-sheet.") can often be estimated spectroscopically. For proteins, a common method is far-ultraviolet (far-UV, 170-250 nm) circular dichroism. A pronounced double minimum at 208 and 222 nm indicate α-helical structure, whereas a single minimum at 204 nm or 217 nm reflects random-coil or β-sheet structure, respectively. A less common method is infrared spectroscopy, which detects differences in the bond oscillations of amide groups due to hydrogen-bonding. Finally, secondary-structure contents may be estimated accurately using the chemical shifts of an unassigned NMR spectrum.

Secondary structure was introduced by Kaj Ulrik Linderstrøm-Lang at Stanford in 1952.

### *Protein*



Hydrogen bonds (yellow dots) stabilizing an alpha-helix

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

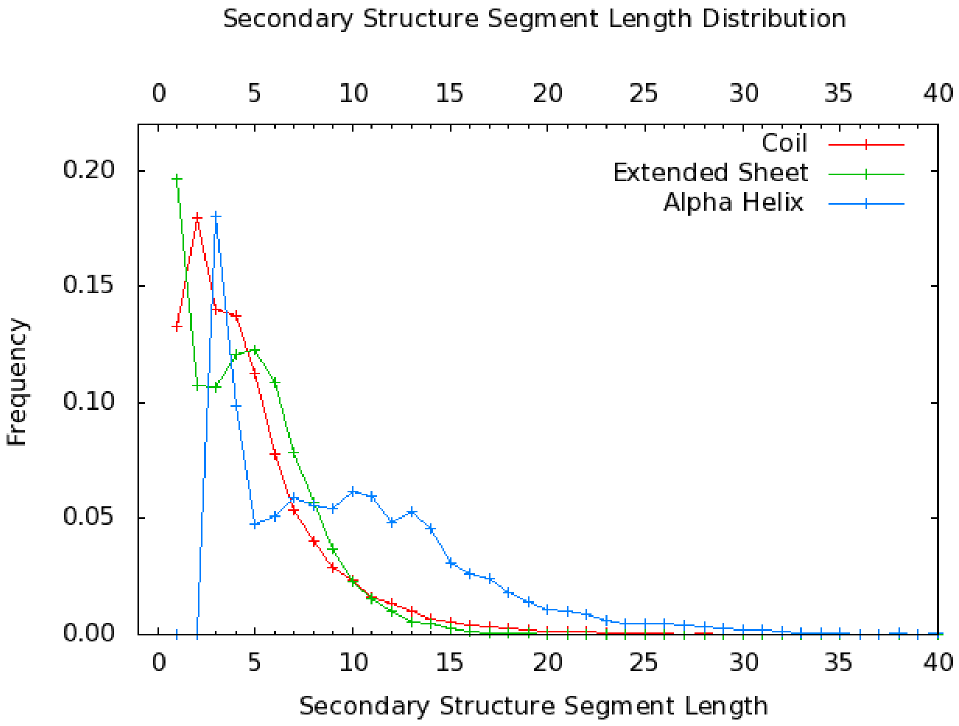
Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as "helix breakers" because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine ("MALEK" in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine and phenylalanine) and $C^{\beta}$-branched amino acids (isoleucine, valine, and threonine) prefer to adopt β-strand conformations. However, these preferences are not strong enough to produce a reliable method of predicting secondary structure from sequence alone.

There are several methods for defining protein secondary structure (e.g. DEFINE, DSSP, STRIDE (protein)).

Structural features of the three major forms of protein helices

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

## The DSSP code



Distribution obtained from non-redundant pdb_select dataset (March 2006); Secondary structure assigned by DSSP; 8 conformational states reduced to 3 states: H=HGI, E=EB, C=STC; Visible are mixtures of (gaussian) distributions, resulting also from the reduction of DSSP states

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

- G = 3-turn helix ($3_{10}$ helix). Min length 3 residues.
- H = 4-turn helix (α helix). Min length 4 residues.
- I = 5-turn helix (π helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment).

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

## DSSP H-bond definition

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

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

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

## Protein secondary-structure prediction

Predicting protein tertiary structure from only its amino acid sequence is a very challenging problem, but using the simpler secondary structure definitions is more tractable and has been the focus for research for a long time.

Although, the 8-state DSSP code is already a simplification from the continuous variation of hydrogen bonding patterns present in a protein the majority of secondary prediction methods simplify further to the three dominant states: Helix, Sheet and Coil. How the conversion is made from 8- to 3-state varies between methods. Early methods of secondary-structure prediction were based on the helix- or sheet-forming propensities of individual amino acids, sometimes coupled with rules for estimating the free energy of forming secondary structure elements. Such methods were typically ~60% accurate in predicting which of the three states (helix/sheet/coil) a residue adopts. A significant increase in accuracy (to nearly ~80%) was made by exploiting multiple sequence alignment; knowing the full distribution of amino acids that occur at a position (and in its vicinity, typically ~7 residues on either side) throughout evolution provides a much better picture of the structural tendencies near that position. For illustration, a given protein might have a glycine at a given position, which by itself might suggest a random coil there. However, multiple sequence alignment might reveal that helix-favoring amino acids occur at that position (and nearby positions) in 95% of homologous proteins spanning nearly a billion years of evolution. Moreover, by examining the average hydrophobicity at that and nearby positions, the same alignment might also suggest a pattern of residue solvent accessibility consistent with an α-helix. Taken together, these factors would suggest that the glycine of the original protein adopts α-helical structure, rather than random coil. Several types of methods are used to combine all the available data to form a 3-state prediction, including neural networks, hidden Markov models and support vector machines. Modern prediction methods also provide a confidence score for their predictions at every position.

Secondary-structure prediction methods are continuously benchmarked, e.g., in the EVA experiment. Based on ~270 weeks of testing, the most accurate methods at present are PSIPRED, SAM, PORTER, PROF and SABLE. Interestingly, it does not seem to be possible to improve upon these methods by taking a consensus of them. The chief area for improvement appears to be the prediction of β-strands; residues confidently predicted as β-strand are likely to be so, but the methods are apt to overlook some β-strand segments (false negatives). There is likely an upper limit of ~90% prediction accuracy overall, due to the idiosyncrasies of the standard method (DSSP) for assigning secondary-structure classes (helix/strand/coil) to PDB structures, against which the predictions are benchmarked.

Accurate secondary-structure prediction is a key element in the prediction of tertiary structure, in all but the simplest (homology modeling) cases. For example, a confidently predicted pattern of six secondary structure elements βαββαβ is the signature of a ferredoxin fold.

### *Alignment*

Both protein and nucleic acid secondary structures can be used to aid in multiple sequence alignment. These alignments can be made more accurate by the inclusion of secondary structure information in addition to simple sequence information. This is sometimes less useful in RNA because base pairing is much more highly conserved than sequence. Distant relationships between proteins whose primary structures are unalignable can sometimes be found by secondary structure.

## Chapter 5

# Protein Tertiary Structure

In biochemistry and molecular biology, the **tertiary structure** of a protein or any other macromolecule is its three-dimensional structure, as defined by the atomic coordinates.

### *Relationship to primary structure*

Tertiary structure is considered to be largely determined by the protein's primary structure - the sequence of amino acids of which it is composed. Efforts to predict tertiary structure from the primary structure are known generally as protein structure prediction. However, the environment in which a protein is synthesized and allowed to fold are significant determinants of its final shape and are usually not directly taken into account by current prediction methods. Most such methods do rely on comparisons between the sequence to be predicted and sequences of known structure in the Protein Data Bank and thus account for environment indirectly, assuming the target and template sequences share similar cellular contexts. Stanford University's Folding@Home project is a distributed computing research effort which uses its approximately 5 petaFLOPS (~10 x86 petaFLOPS) of computing power to attempt to model the tertiary (and quaternary) structures of proteins, as well as other aspects of how and why proteins fold into the inordinately complex and varied shapes they take. No currently existing algorithm is yet able to consistently predict a proteins' tertiary or quaternary structure given only its primary structure; learning how to accurately predict the tertiary and quaternary structure of any protein given only its amino acid sequence and the pertinent cellular conditions would be a monumental achievement. Although this ambitious goal has yet to be achieved, researchers have discovered how to combine several of the best of Folding@Home's algorithms to accurately predict the folded structure of *some* proteins under certain conditions. The calculations performed by the algorithms are constantly evolving, increasing in complexity and nuance, and involve enormous numbers of variables. These techniques are superficially comparable to weather models that show hurricane storm tracks; each of several algorithms independently models a complex system (the weather, in this case) somewhat differently from each of its sister weather algorithms, and the average of all the algorithms' output is taken to be the most likely "storm track". The shape of proteins can be elucidated through a somewhat similar process.

Researchers are also interested in proteins that can fold into more than one stable configuration; protein aggregation diseases such as Alzheimer's Disease and Huntington's Disease as well as prion diseases such as Mad Cow disease can be better understood by constructing (and deconstructing) disease models; the most common way of doing this is by developing a way of inducing the desired disease state in test animals (administering MPTP to give the animals Parkinson's disease, or knocking out a gene essential for the prevention of certain tumors from the animals' genomes). Folding@Home allows for the modelling of disease states that are not as easily induced, without the need for test animals. Perhaps more importantly, fully human proteins encoded by fully human genes can be used without any of the ethical problems that arise in studying living human beings. Due to its enormous flexibility, which has only briefly been discussed here, coupled with its ability to improve over time, Folding@Home and projects like it are quickly becoming indispensable tools among researchers from a broad variety of disciplines. The possibilities in medicine, biology, pathology, nuclear physics, and other scientific disciplines should a reliable way to accurately model the final tertiary or quaternary structure of human proteins are almost limitless. Proteins, due to the precise conformations they fold into, are nature's original nanomachines; developing an inexpensive and practical way to design and target proteins would completely revolutionize medicine and would have incredibly far-reaching implications. The significance of such a discovery cannot be overstated. To date, over 70 scientific papers have been published on discoveries that relied on Folding@Home.

## *Determinants of tertiary structure*

In globular proteins, tertiary interactions are frequently stabilized by the sequestration of hydrophobic amino acid residues in the protein core, from which water is excluded, and by the consequent enrichment of charged or hydrophilic residues on the protein's water-exposed surface. In secreted proteins that do not spend time in the cytoplasm, disulfide bonds between cysteine residues help to maintain the protein's tertiary structure. A variety of common and stable tertiary structures appear in a large number of proteins that are unrelated in both function and evolution - for example, many proteins are shaped like a TIM barrel, named for the enzyme triosephosphateisomerase. Another common structure is a highly stable dimeric coiled coil structure composed of 2-7 alpha helices. Proteins are classified by the folds they represent in databases like SCOP and CATH.

## *Stability of native states*

The most typical conformation of a protein in its cellular environment is generally referred to as the native state or native conformation. It is commonly assumed that this most-populated state is also the most thermodynamically stable conformation attainable for a given primary structure; this is a reasonable first approximation but the claim assumes that the reaction is not under kinetic control - that is, that the time required for the protein to attain its native conformation before being translated is small.

In the cell, a variety of protein chaperones assist a newly synthesized polypeptide in attaining its native conformation. Some such proteins are highly specific in their function, such as protein disulfide isomerase; others are very general and can be of assistance to most globular proteins - the prokaryotic GroEL/GroES system and the homologous eukaryotic Heat shock proteins Hsp60/Hsp10 system fall into this category.

Some proteins explicitly take advantage of the fact that they can become kinetically trapped in a relatively high-energy conformation due to folding kinetics. Influenza hemagglutinin, for example, is synthesized as a single polypeptide chain that acts as a kinetic trap. The "mature" activated protein is proteolytically cleaved to form two polypeptide chains that are trapped in a high-energy conformation. Upon encountering a drop in pH, the protein undergoes an energetically favorable conformational rearrangement that enables it to penetrate a host cell membrane.

Many serpins (serine protease inhibitors) are metastable, and undergo a conformational change when a loop of the protein is cut by a protease.

## Experimental determination

The majority of protein structures known to date have been solved with the experimental technique of X-ray crystallography, which typically provides data of high resolution but provides no time-dependent information on the protein's conformational flexibility. A second common way of solving protein structures uses NMR, which provides somewhat lower-resolution data in general and is limited to relatively small proteins, but can provide time-dependent information about the motion of a protein in solution. Dual polarisation interferometry is a time resolved analytical method for determining the overall conformation and conformational changes in surface captured proteins providing complementary information to these high resolution methods. More is known about the tertiary structural features of soluble globular proteins than about membrane proteins because the latter class is extremely difficult to study using these methods.

## Interactions stabilizing tertiary structure

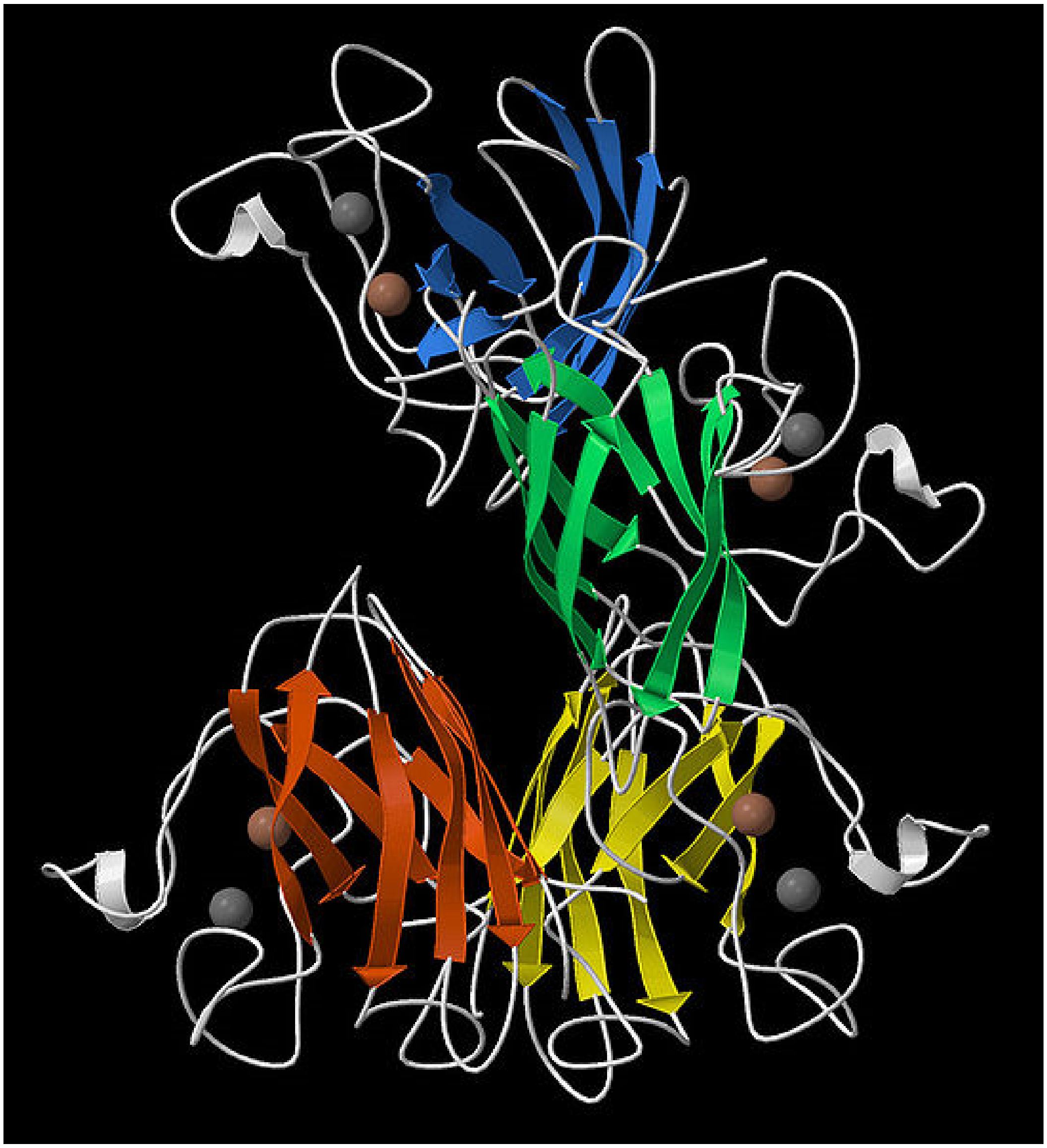
- Disulfide bonds
- Hydrophobic interactions
- Hydrogen bonds
- Ionic bonds

## History

Since the tertiary structure of proteins is an important problem in biochemistry, and since structure determination is relatively difficult, protein structure prediction has been a long-standing problem. The first predicted structure of globular proteins was the cyclol model of Dorothy Wrinch, but this was quickly discounted as being inconsistent with experimental data. Modern methods are sometimes able to predict the tertiary structure *de novo* to within 5 Å for small proteins (<120 residues) and under favorable conditions, e.g., confident secondary structure predictions.

## Chapter 6

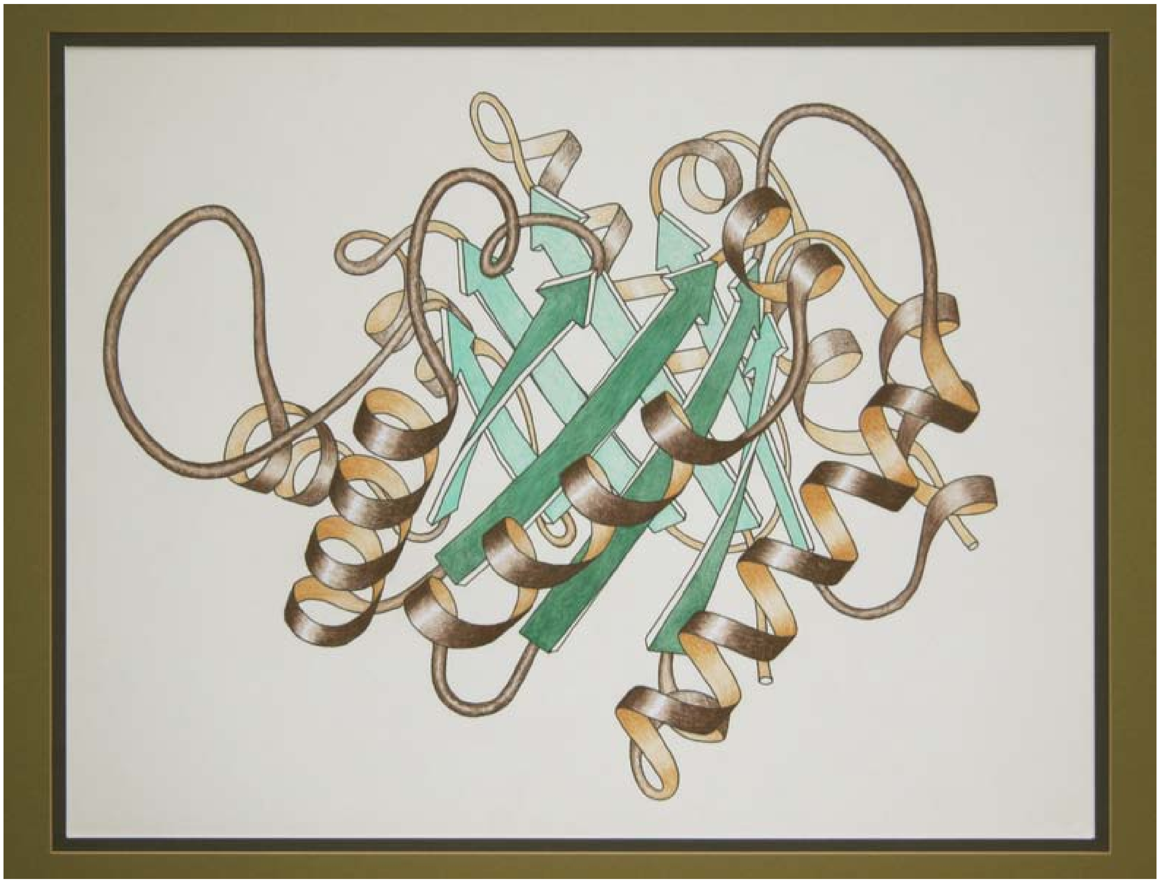
# Ribbon Diagram



Computer-drawn ribbon diagram of two CuZn superoxide dismutase dimers.

**Ribbon diagrams**, also known as **Richardson Diagrams**, are 3D schematic representations of protein structure and are one of the most common methods of protein depiction used today. Ribbon diagrams are generated by interpolating a smooth curve through the polypeptide backbone. α-helices are shown as coiled ribbons or thick tubes, β-strands as arrows, and lines or thin tubes for random coils. The direction of the polypeptide chain may be indicated by a colour ramp along the length of the ribbon.

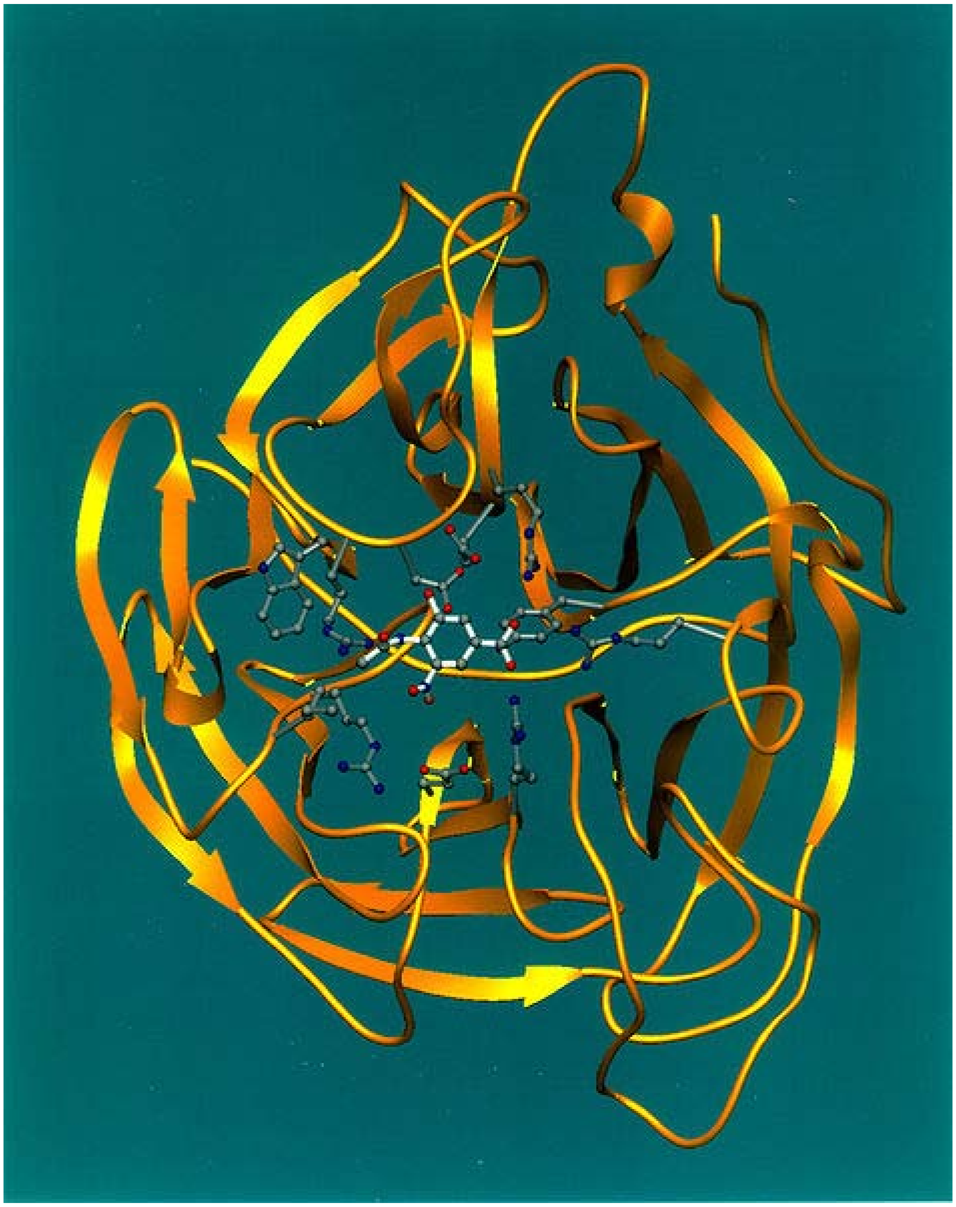
Ribbon diagrams are simple, yet powerful, in expressing the visual basics of a molecular structure (twist, fold and unfold). This method has successfully portrayed the overall organization of the protein structure, reflecting its 3-dimensional information, and allowing for better understanding of a complex object both by the expert structural biologists and also by other scientists, students, and the general public.



Ribbon schematic of triose P isomerase monomer, hand-drawn by Jane Richardson

## *History*

Originally conceived by Jane S. Richardson in 1980 (influenced by some earlier individual illustrations, e.g., see), her hand-drawn ribbon diagrams were the first schematics of 3D protein structure to be produced systematically, to illustrate a classification of protein structures for an article in *Advances in Protein Chemistry* (now available in annotated form on-line at Anatax). These drawings were made in pen on tracing paper over a printout of a Cα trace of the atomic coordinates; they preserved positions, smoothed the backbone path, and incorporated small local shifts to disambiguate the visual appearance. As well as the TIM ribbon drawing at right, other hand-drawn examples are for prealbumin, flavodoxin, and Cu,Zn superoxide dismutase.

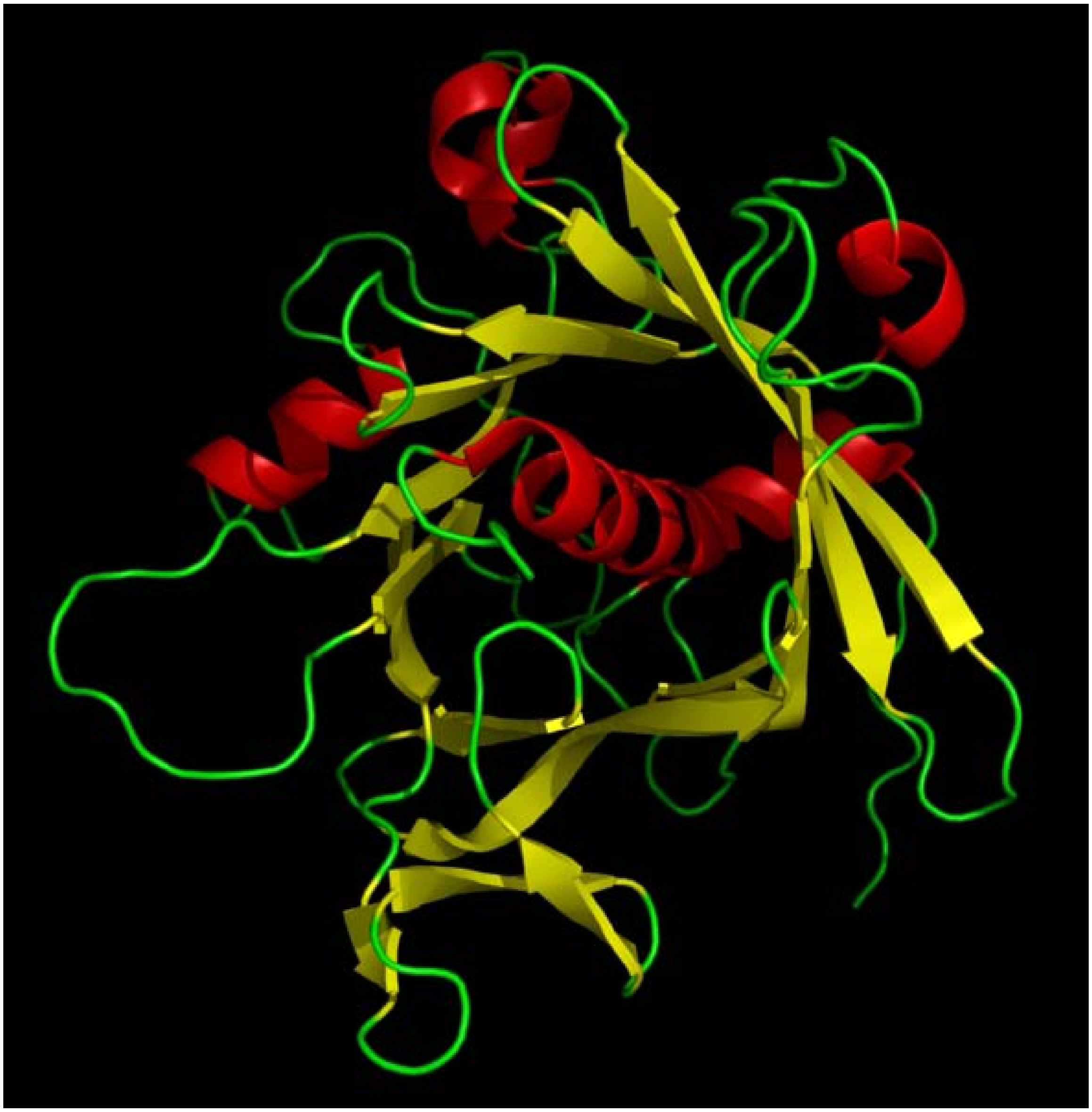


Neuraminidase image, by the Ribbons program

In 1982, Arthur M. Lesk and co-workers first enabled automatic generation of ribbon diagrams through a computational implementation that uses Protein Data Bank files as input. This conceptually simple algorithm fit cubic polynomial B-spline curves to the peptide planes. Most modern graphics systems provide B-splines as a basic drawing primitive. B-Splines are well suited to fitting between data points but not necessarily interpolating through each of those points. To create a line that intersects all data points, Hermite splines can be used. In order to give the right radius for helical spirals while preserving smooth β-strands, B-splines can be modified by offsets proportional to local curvature, as first developed by Mike Carson for his Ribbons program (figure at right) and later adapted by other molecular graphics software, such as the open-source Mage program for kinemage graphics that produced the ribbon image at top right (other examples: 1xk8 trimer and DNA polymerase).

Since their inception, and continuing in the present, a ribbon diagram is the single most common representation of protein structures and a very common choice of cover image for a journal or textbook.

## Current computer programs

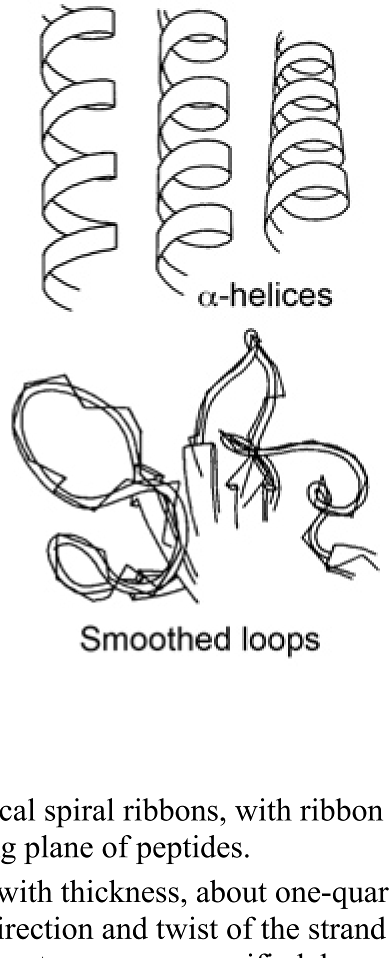


PyMol ribbon of the unusual structure of the "tubby" brain protein

One popular program used for drawing ribbon diagrams is Molscript. Molscript utilizes Hermite splines to create coordinates for coils, turns, strands and helices. The curve passes through all its control points (Cα atoms) guided by direction vectors. The program was built on the basis of traditional molecular graphics by Arthur M. Lesk, Karl Hardman, and John Priestle. Jmol is an open-source Java-based viewer for browsing molecular structures on the web; it includes a simplified "cartoon" version of ribbons. Other graphics programs such as DeepView (example: urease) and MolMol (example: SH2 domain) also produce ribbon images. KiNG is the Java-based successor to Mage (examples: α-hemolysin top view and side view).

UCSF Chimera is a powerful molecular modeling program that also includes visualizations such as ribbons, notable especially for the ability to combine them with contoured shapes from cryo-electron microscopy data. PyMOL, by Warren DeLano, is a very popular and flexible molecular graphics program (based on Python) that operates in interactive mode and also produces presentation-quality 2D images for ribbon diagrams and many other representations.

### *Features of ribbon diagrams*

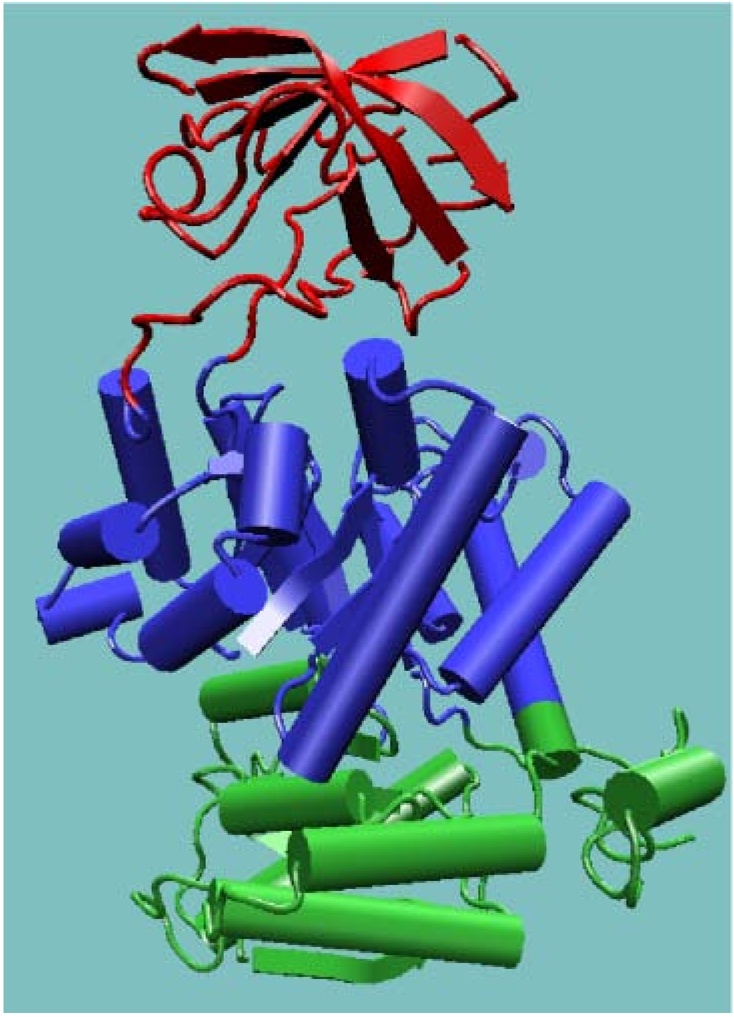


| Secondary Structure | |
|---|---|
| α-Helices | Cylindrical spiral ribbons, with ribbon plane approximately following plane of peptides. |
| β-Strand | Arrows with thickness, about one-quarter as thick as they are wide, shows direction and twist of the strand from amino to the carboxy end. β-sheets are seen as unified, because neighboring strands twist in unison. |

| | |
|---|---|
| **Loops and miscellaneous** | |
| Nonrepetitive loops | Round ropes that are fatter in the foreground and thinner towards the back, following smoothed path of Cα trace. |
| Junctions between loops and helices | Round rope that gradually flattens out into a thin helical ribbon. |
| **Other features** | |
| Polypeptide direction,<br>NH2 and COOH termini | Small arrows on one or both of the termini or by letters. For β-strands, the direction of the arrow is sufficient. Today, the direction of the polypeptide chain is often indicated by a colour ramp. |
| Disulfide bonds | Interlocked SS symbol or a zigzag, like a lightning stroke |
| Prosthetic groups or inhibitors | Stick figures, or ball&stick. |
| Metals | Spheres. |
| Shading and colour | Shading or colour adds dimensionality to the diagram. Generally, the features at the front are the darkest, while becoming lighter and lower in contrast towards the back. |

## Chapter 7

# 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.

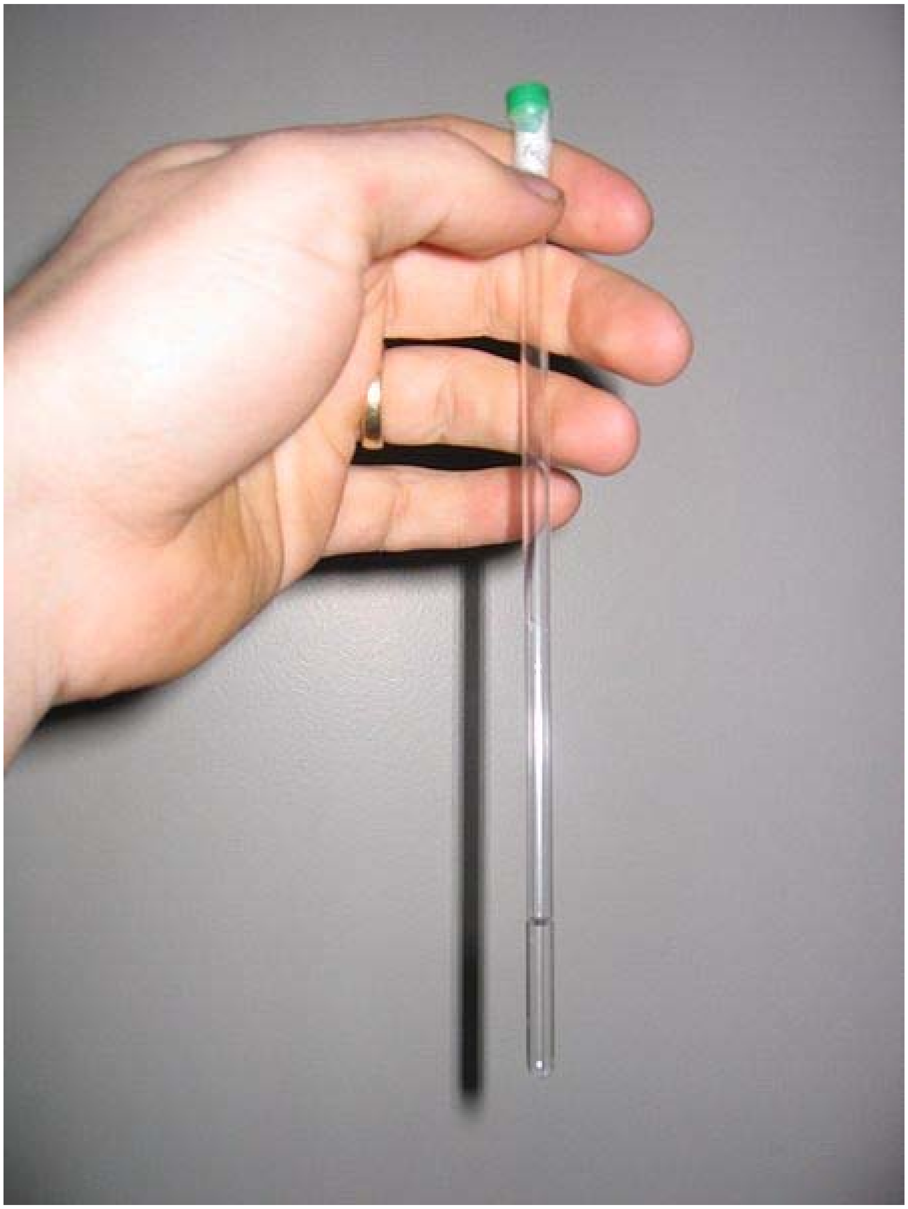
Chapter 8

# Protein Nuclear Magnetic Resonance Spectroscopy

Pacific Northwest National Laboratory's high magnetic field (800 MHz) NMR spectrometer being loaded with a sample.

**Protein nuclear magnetic resonance spectroscopy** (usually abbreviated **protein NMR**) is a field of structural biology in which NMR spectroscopy is used to obtain information about the structure and dynamics of proteins. The field was pioneered by Richard R. Ernst and Kurt Wüthrich, among others. Protein NMR techniques are continually being used and improved in both academia and the biotech industry. Structure determination by NMR spectroscopy usually consists of several following phases, each using a separate set of highly specialized techniques. The sample is prepared, resonances are assigned, restraints are generated and a structure is calculated and validated.

## *Sample preparation*



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

Protein nuclear magnetic resonance is performed on aqueous samples of highly purified protein. Usually the sample consist of between 300 and 600 microlitres with a protein concentration in the range 0.1 – 3 millimolar. The source of the protein can be either natural or produced in an expression system using recombinant DNA techniques through genetic engineering. Recombinantly expressed proteins are usually easier to produce in sufficient quantity, and makes isotopic labelling possible.

The purified protein is usually dissolved in a buffer solution and adjusted to the desired solvent conditions. The NMR sample is prepared in a thin walled glass tube.

## *Data collection*

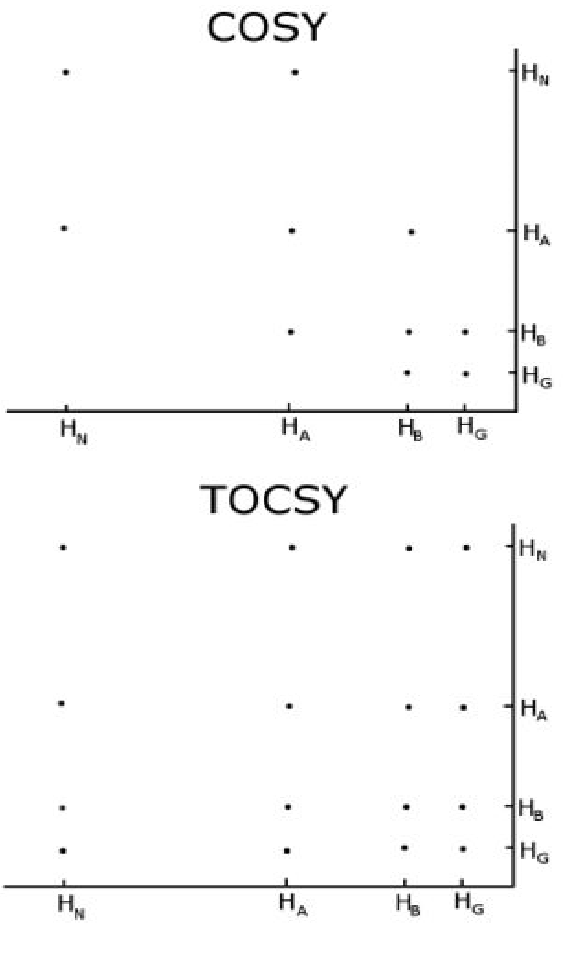
Protein NMR utilizes multidimensional nuclear magnetic resonance experiments to obtain information about the protein. Ideally, each distinct nucleus in the molecule experiences a distinct chemical environment and thus has a distinct chemical shift by which it can be recognized. However, in large molecules such as proteins the number of resonances can typically be several thousand and a one-dimensional spectrum inevitably has incidental overlaps. Therefore multidimensional experiments are performed which correlate the frequencies of distinct nuclei. The additional dimensions decrease the chance of overlap and have a larger information content since they correlate signals from nuclei within a specific part of the molecule. Magnetization is transferred into the sample using pulses of electromagnetic (radiofrequency) energy and between nuclei using delays; the process is described with so-called pulse sequences. Pulse sequences allow the experimenter to investigate and select specific types of connections between nuclei. The array of nuclear magnetic resonance experiments used on proteins fall in two main categories — one where magnetization is transferred through the chemical bonds, and one where the transfer is through space, irrespective of the bonding structure. The first category is used to assign the different chemical shifts to a specific nucleus, and the second is primarily used to generate the distance restraints used in the structure calculation, and in the assignment with unlabelled protein.

Depending on the concentration of the sample, on the magnetic field of the spectrometer, and on the type of experiment, a single multidimensional nuclear magnetic resonance experiment on a protein sample may take hours or even several days to obtain suitable signal-to-noise ratio through signal averaging, and to allow for sufficient evolution of magnetization transfer through the various dimensions of the experiment. Other things being equal, higher-dimensional experiments will take longer than lower-dimensional experiments.

Typically the first experiment to be measured with an isotope-labelled protein is a 2D heteronuclear single quantum correlation (HSQC) spectrum where "heteronuclear" refers to nuclei other than 1H. In theory the heteronuclear single quantum correlation has one peak for each H bound to a heteronucleus. Thus in the 15N-HSQC one signal is expected for each amino acid residue with the exception of proline which has no amide-hydrogen due to the cyclic nature of its backbone. Tryptophan and certain other residues with N-containing sidechains also give rise to additional signals. The 15N-HSQC is often referred to as the fingerprint of a protein because each protein has a unique pattern of signal positions. Analysis of the 15N-HSQC allows researchers to evaluate whether the expected number of peaks is present and thus to identify possible problems due to multiple conformations or sample heterogeneity. The relatively quick heteronuclear single quantum correlation experiment helps determine the feasibility of doing subsequent longer, more expensive, and more elaborate experiments. It is not possible to assign peaks to specific atoms from the heteronuclear single quantum correlation alone.

## *Resonance assignment*

In order to analyze the nuclear magnetic resonance data, it is important to get a resonance assignment for the protein. That is to find out which chemical shift corresponds to which atom. This is typically achieved by sequential walking using information derived from several different types of NMR experiment. The exact procedure depends on whether the protein is isotopically labelled or not, since a lot of the assignment experiments depend on carbon-13 and nitrogen-15.



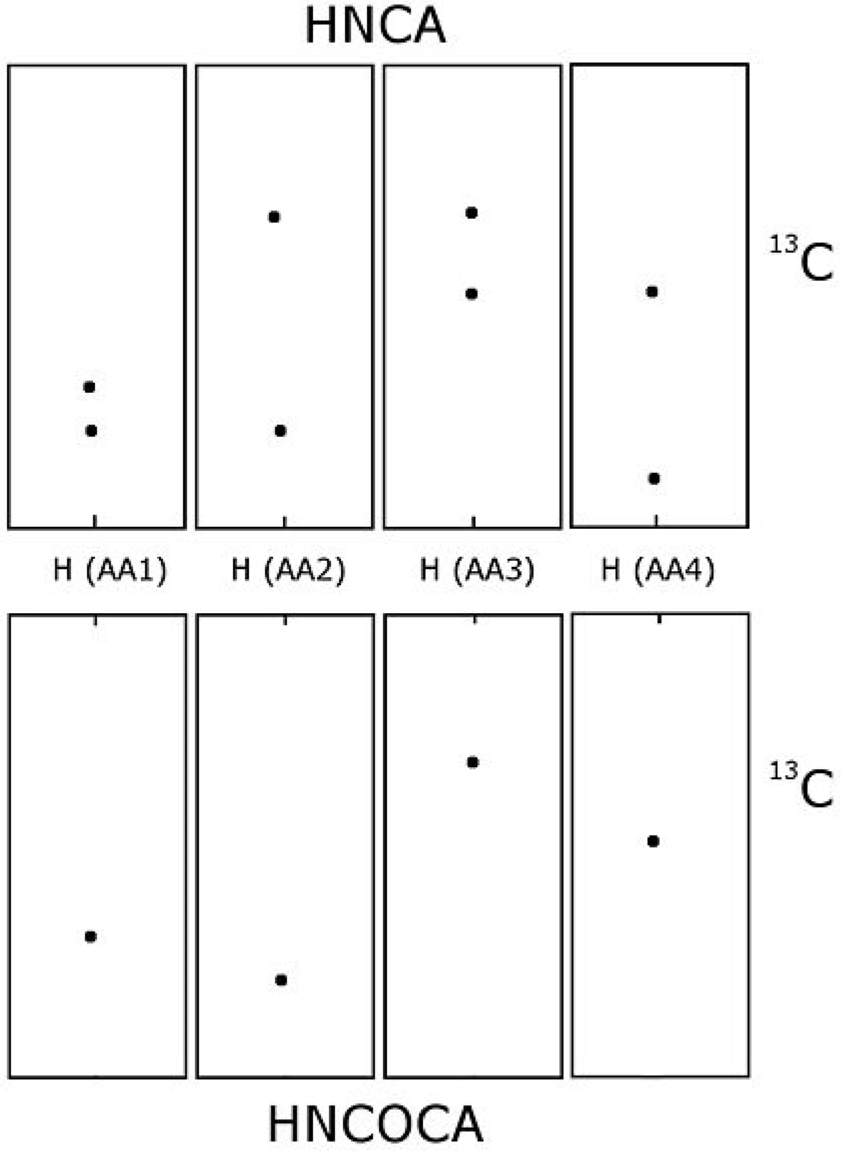
Comparison of a COSY and TOCSY 2D spectra for an amino acid like glutamate or methionine. The TOCSY shows off diagonal crosspeaks between all protons in the spectrum, but the COSY only has crosspeaks between neighbours.

## Homonuclear nuclear magnetic resonance

With unlabelled protein the usual procedure is to record a set of two dimensional homonuclear nuclear magnetic resonance experiments through correlation spectroscopy (COSY), of which several types include conventional correlation spectroscopy, *total correlation* spectroscopy (TOCSY) and nuclear Overhauser effect spectroscopy (NOESY). A two-dimensional nuclear magnetic resonance experiment produces a two-dimensional spectrum. The units of both axes are chemical shifts. The COSY and TOCSY transfer magnetization through the chemical bonds between adjacent protons. The conventional correlation spectroscopy experiment is only able to transfer magnetization between protons on adjacent atoms, whereas in the total correlation spectroscopy experiment the protons are able to relay the magnetization, so it is transferred among all the protons that are connected by adjacent atoms. Thus in a conventional correlation spectroscopy, an alpha proton transfers magnetization to the beta protons, the beta protons transfers to the alpha and gamma protons, if any are present, then the gamma proton transfers to the beta and the delta protons, and the process continues. In total correlation spectroscopy, the alpha and all the other protons are able to transfer magnetization to the beta, gamma, delta, epsilon if they are connected by a continuous chain of protons. The continuous chain of protons are the sidechain of the individual amino acids. Thus these two experiments are used to build so called spin systems, that is build a list of resonances of the chemical shift of the peptide proton, the alpha protons and all the protons from each residue’s sidechain. Which chemical shifts corresponds to which nuclei in the spin system is determined by the conventional correlation spectroscopy connectivities and the fact that different types of protons have characteristic chemical shifts. To connect the different spinsystems in a sequential order, the nuclear Overhauser effect spectroscopy experiment has to be used. Because this experiment transfers magnetization through space, it will show crosspeaks for all protons that are close in space regardless of whether they are in the same spin system or not. The neighbouring residues are inherently close in space, so the assignments can be made by the peaks in the NOESY with other spin systems.

One important problem using homonuclear nuclear magnetic resonance is overlap between peaks. This occurs when different protons have the same or very similar chemical shifts. This problem becomes greater as the protein becomes larger, so homonuclear nuclear magnetic resonance is usually restricted to small proteins or peptides.

### Nitrogen-15 nuclear magnetic resonance



Schematic of an HNCA and HNCOCA for four sequential residues. The nitrogen-15 dimension is perpendicular to the screen. Each window is focused on the nitrogen chemical shift of that amino acid. The sequential assignment is made by matching the alpha carbon chemical shifts. In the HNCA each residue sees the alpha carbon of it self and the preceding residue. The HNCOCA only sees the alpha carbon of the preceding residue.

### Carbon-13 and nitrogen-15 nuclear magnetic resonance

When the protein is labelled with carbon-13 and nitrogen-15 it is possible to record an experiment that transfers magnetisation over the peptide bond, and thus connect different spin systems through bonds. This is usually done using some of the following experiments, HNCO, HNCACO, HNCA, HNCOCA, HNCACB and CBCACONH. All six experiments consist of a HSQC plane expanded with a carbon dimension. In the HNCACO the spectrum contains peaks at the chemical shifts of the carbonyl carbons in the residue of the HSQC peak and the previous one in the sequence. The HNCO only contains the chemical shift from the previous residue, and it is thus possible to assign the carbonyl carbon shifts that corresponds to each HSQC peak and the one previous to that one. Sequential assignment can then be undertaken by matching the shifts of each spin system's own and previous carbons. The HNCA and HNCOCA works similarly, just with the alpha carbons rather than the carbonyls, and the HNCACB and the CBCACONH contains both the alpha carbon and the beta carbon. Usually several of these experiments are required to resolve overlap in the carbon dimension. This procedure is usually less ambiguous than the NOESY based method, since it is based on through bond transfer. In the NOESY-based methods additional peaks that are close in space but not belonging to the sequential residues will appear confusing the assignment process. When the sequential assignment has been made it is usually possible to assign the sidechains using HCCH-TOCSY, which is basically a TOCSY experiment resolved in an additional carbon dimension.

## *Restraint generation*

In order to make structure calculations a number of experimentially determined restraints have to be generated. These fall into different categories, the most widely used is distance restraints and angle restraints.

### Distance restraints

A crosspeak in a NOESY experiment signifies spatial proximity between the two nuclei in question. Thus each peak can be converted in to a maximum distance between the nuclei, usually between 1,8 and 6 angstroms. The intensity of a noesy peak is proportional to the distance to the minus 6th power, so the distance is determined according to intensity of the peak. The intensity-distance relationship is not exact, so usually a distance range is used.

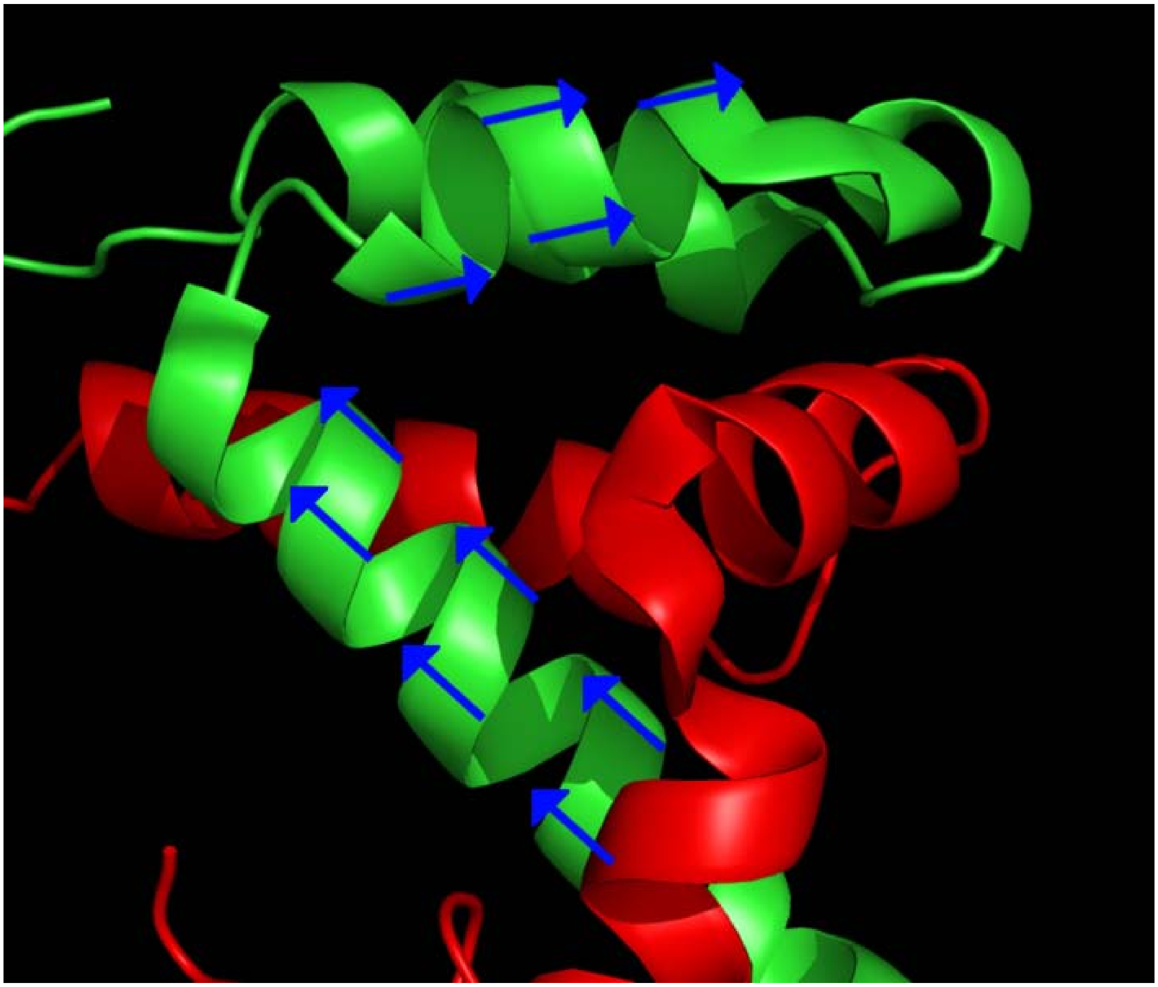
It is of great importance to assign the noesy peaks to the correct nuclei based on the chemical shifts. If this task is performed manually it is usually very labor intensive, since proteins usually have thousands of noesy peaks. Some computer programs such as UNIO, CYANA and ARIA/CNS perform this task automatically on manually pre-processed listings of peak positions and peak volumes, coupled to a structure calculation. Direct access to the raw NOESY data without the cumbersome need of iteratively refined peak lists is so far only granted by the ATNOS/CANDID approach implemented in the UNIO software package  and thus indeed guarantees objective and efficient NOESY spectral analysis.

To obtain as accurate assignments as possible it is a great advantage to have access to carbon-13 and nitrogen-15 noesy experiments, since they help to resolve overlap in the proton dimension. This leads to faster and more reliable assignments, and in turn to better structures.

## Angle restraints

In addition to distance restraints, restraints on the torsion angles of the chemical bonds, typically the psi and phi angles can be generated. One approach is to use the Karplus equation, to generate angle restraints from coupling constants. Another approach uses the chemical shifts to generate angle restraints. Both methods use the fact that the geometry around the alpha carbon affects the coupling constants and chemical shifts, so given the coupling constants or the chemical shifts, a qualified guess can be made about the torsion angles.

## Orientation restraints



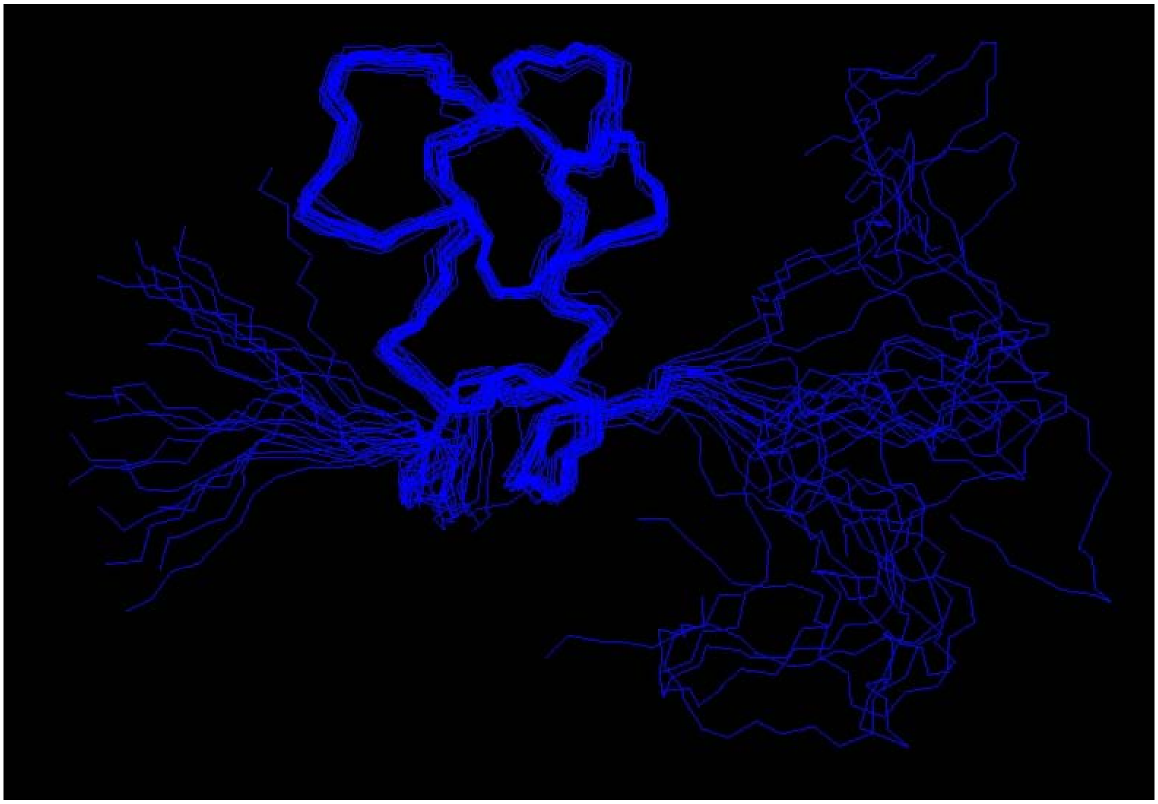
The blue arrows represent the orientation of the N - H bond of selected peptide bonds. By determining the orientation of a sufficient amount of bonds relative to the external magnetic field, the structure of the protein can be determined. From PDB record 1KBH

The analyte molecules in a sample can be partially ordered with respect to the external magnetic field of the spectrometer by manipulating the sample conditions. Common techniques include addition of bacteriophages or bicelles to the sample, or preparation of the sample in a stretched polyacrylamide gel. This creates a local environment that favours certain orientations of nonspherical molecules. Normally in solution NMR the dipolar couplings between nuclei are averaged out because of the fast tumbling of the molecule. The slight overpopulation of one orientation means that a residual dipolar coupling remains to be observed. The dipolar coupling is commonly used in solid state NMR and provides information about the relative orientation of the bond vectors relative to a single global reference frame. Typically the orientation of the N-H vector is probed in a HSQC like experiment. Initially residual dipolar couplings were used for refinement of previously determined structures, but attempts at de novo structure determination have also been made.

### *Hydrogen-Deuterium exchange*

NMR spectroscopy is nuclei specific. Thus it can distinguish between hydrogen and deuterium. The amide protons in the protein exchange readily with the solvent, and if the solvent contains a different isotope, typically deuterium, the reaction can be monitored by NMR spectroscopy. How rapidly a given amide exchanges reflects its solvent accessibility. Thus amide exchange rates can give information on which parts of the protein are buried, hydrogen bonded etc. A common application is to compare the exchange of a free form versus a complex. The amides that become protected in the complex, are assumed to be in the interaction interface.

## *Structure calculation*



Nuclear magnetic resonance structure determination generates an ensemble of structures. The structures will only converge if the data is sufficient to dictate a specific fold. In these structures, it is only the case for a part of the structure. From PDB 1SSU.

The experimentially determined restraints can be used as input for the structure calculation process. Researchers, using computer programs such as CYANA (Software) or XPLOR-NIH, attempt to satisfy as many of the restraints as possible, in addition to general properties of proteins such as bond lengths and angles. The algorithms convert the restraints and the general protein properties into energy terms, and thus tries to minimize the energy. The process results in an ensemble of structures that, if the data were sufficient to dictate a certain fold, will converge.

## *Dynamics*

In addition to structures, nuclear magnetic resonance can yield information on the dynamics of various parts of the protein. This usually involves measuring relaxation times such as $T_1$ and $T_2$ to determine order parameters, correlation times, and chemical exchange rates. NMR relaxation is a consequence of local fluctuating magnetic fields within a molecule. Local fluctuating magnetic fields are generated by molecular motions. In this way measurements of relaxation times can provide information of motions within a molecule on the atomic level. In NMR studies of protein dynamics the nitrogen-15 isotope is the preferred nucleus to study because its relaxation times are relatively simple to relate to molecular motions This however requires isotope labeling of the protein. The $T_1$ and $T_2$ relaxation times can be measured using various types of HSQC based experiments. The types of motions which can be detected are motions that occur on a time-scale ranging from about 10 picoseconds to about 10 nanoseconds. In addition slower motions, which take place on a time-scale ranging from about 10 microseconds to 100 milliseconds, can also be studied. However, since nitrogen atoms are mainly found in the backbone of a protein, the results mainly reflect the motions of the backbone, which is the most rigid part of a protein molecule. Thus, the results obtained from nitrogen-15 relaxation measurements may not be representative for the whole protein. Therefore techniques utilizing relaxation measurements of carbon-13 and deuterium have recently been developed, which enables systematic studies of motions of the amino acid side chains in proteins.

## *NMR spectroscopy on large proteins*

Traditionally nuclear magnetic resonance spectroscopy has been limited to relatively small proteins or protein domains. This is in part caused by problems resolving overlapping peaks in larger proteins, but this has been alleviated by the introduction of isotope labelling and multidimensional experiments. Another more serious problem is the fact that in large proteins the magnetization relaxes faster, which means there is less time to detect the signal. This in turn causes the peaks to become broader and weaker, and eventually disappear. Two techniques have been introduced to attenuate the relaxation: transverse relaxation optimized spectroscopy (TROSY) and deuteration of proteins. By using these techniques it has been possible to study proteins in complex with the 900 kDa chaperone GroES-GroEL.

## *Automation of the process*

Structure determination by NMR has traditionally been a time consuming process, requiring interactive analysis of the data by a highly trained scientist. There has been a considerable interest in automating the process to increase the throughput of structure determination and to make protein NMR accessible to non-experts. The two most time consuming processes involved are the sequence-specific resonance assignment (backbone and side-chain assignment) and the NOE assignment tasks. Several different computer programs have been published that target individual parts of the overall NMR structure determination process in an automated fashion. Most progress have been achieved for the task of automated NOE assignment. So far, only the FLYA and the UNIO approach were proposed to perform the entire protein NMR structure determination process in an automated manner without any human intervention . Efforts have also been made to standardize the structure calculation protocol to make it quicker and more amenable to automation.

## Chapter 9

# Homology Modeling

**Homology modeling**, also known as **comparative modeling** of protein refers to constructing an atomic-resolution model of the "*target*" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "*template*"). Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. It has been shown that protein structures are more conserved than protein sequences amongst homologues, but sequences falling below a 20% sequence identity can have very different structure.

Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. It has been shown that three-dimensional protein structure is evolutionarily more conserved than expected due to sequence conservation.

The sequence alignment and template structure are then used to produce a structural model of the target. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity.

The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The approach can be complicated by the presence of alignment gaps (commonly called indels) that indicate a structural region present in the target but not in the template, and by structure gaps in the template that arise from poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure. Model quality declines with decreasing sequence identity; a typical model has ~1-2 Å root mean square deviation between the matched $C^{\alpha}$ atoms at 70% sequence identity but only 2-4 Å agreement at 25% sequence identity. However, the errors are significantly higher in the loop regions, where the amino acid sequences of the target and template proteins may be completely different.

Regions of the model that were constructed without a template, usually by loop modeling, are generally much less accurate than the rest of the model. Errors in side chain packing and position also increase with decreasing identity, and variations in these packing configurations have been suggested as a major reason for poor model quality at low identity. Taken together, these various atomic-position errors are significant and impede the use of homology models for purposes that require atomic-resolution data, such as drug design and protein-protein interaction predictions; even the quaternary structure of a protein may be difficult to predict from homology models of its subunit(s). Nevertheless, homology models can be useful in reaching *qualitative* conclusions about the biochemistry of the query sequence, especially in formulating hypotheses about why certain residues are conserved, which may in turn lead to experiments to test those hypotheses. For example, the spatial arrangement of conserved residues may suggest whether a particular residue is conserved to stabilize the folding, to participate in binding some small molecule, or to foster association with another protein or nucleic acid.

Homology modeling can produce high-quality structural models when the target and template are closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds. The chief inaccuracies in homology modeling, which worsen with lower sequence identity, derive from errors in the initial sequence alignment and from improper template selection. Like other methods of structure prediction, current practice in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

## Motive

The method of homology modeling is based on the observation that protein tertiary structure is better conserved than amino acid sequence. Thus, even proteins that have diverged appreciably in sequence but still share detectable similarity will also share common structural properties, particularly the overall fold. Because it is difficult and time-consuming to obtain experimental structures from methods such as X-ray crystallography and protein NMR for every protein of interest, homology modeling can provide useful structural models for generating hypotheses about a protein's function and directing further experimental work.

There are exceptions to the general rule that proteins sharing significant sequence identity will share a fold. For example, a judiciously chosen set of mutations of less than 50% of a protein can cause the protein to adopt a completely different fold. However, such a massive structural rearrangement is unlikely to occur in evolution, especially since the protein is usually under the constraint that it must fold properly and carry out its function in the cell. Consequently, the roughly folded structure of a protein (its "topology") is conserved longer than its amino-acid sequence and much longer than the corresponding DNA sequence; in other words, two proteins may share a similar fold even if their evolutionary relationship is so distant that it cannot be discerned reliably. For comparison, the function of a protein is conserved much *less* than the protein sequence, since relatively few changes in amino-acid sequence are required to take on a related function.

## *Steps in model production*

The homology modeling procedure can be broken down into four sequential steps: template selection, target-template alignment, model construction, and model assessment. The first two steps are often essentially performed together, as the most common methods of identifying templates rely on the production of sequence alignments; however, these alignments may not be of sufficient quality because database search techniques prioritize speed over alignment quality. These processes can be performed iteratively to improve the quality of the final model, although quality assessments that are not dependent on the true target structure are still under development.

Optimizing the speed and accuracy of these steps for use in large-scale automated structure prediction is a key component of structural genomics initiatives, partly because the resulting volume of data will be too large to process manually and partly because the goal of structural genomics requires providing models of reasonable quality to researchers who are not themselves structure prediction experts.

## *Template selection and sequence alignment*

The critical first step in homology modeling is the identification of the best template structure, if indeed any are available. The simplest method of template identification relies on serial pairwise sequence alignments aided by database search techniques such as FASTA and BLAST. More sensitive methods based on multiple sequence alignment - of which PSI-BLAST is the most common example - iteratively update their position-specific scoring matrix to successively identify more distantly related homologs. This family of methods has been shown to produce a larger number of potential templates and to identify better templates for sequences that have only distant relationships to any solved structure. Protein threading, also known as fold recognition or 3D-1D alignment, can also be used as a search technique for identifying templates to be used in traditional homology modeling methods. When performing a BLAST search, a reliable first approach is to identify hits with a sufficiently low *E*-value, which are considered sufficiently close in evolution to make a reliable homology model. Other factors may tip the balance in marginal cases; for example, the template may have a function similar to that of the query sequence, or it may belong to a homologous operon. However, a template with a poor *E*-value should generally not be chosen, even if it is the only one available, since it may well have a wrong structure, leading to the production of a misguided model. A better approach is to submit the primary sequence to fold-recognition servers or, better still, consensus meta-servers which improve upon individual fold-recognition servers by identifying similarities (consensus) among independent predictions.

Often several candidate template structures are identified by these approaches. Although some methods can generate hybrid models from multiple templates, most methods rely on a single template. Therefore, choosing the best template from among the candidates is a key step, and can affect the final accuracy of the structure significantly. This choice is guided by several factors, such as the similarity of the query and template sequences, of their functions, and of the predicted query and observed template secondary structures. Perhaps most importantly, the *coverage* of the aligned regions: the fraction of the query sequence structure that can be predicted from the template, and the plausibility of the resulting model. Thus, sometimes several homology models are produced for a single query sequence, with the most likely candidate chosen only in the final step.

It is possible to use the sequence alignment generated by the database search technique as the basis for the subsequent model production; however, more sophisticated approaches have also been explored. One proposal generates an ensemble of stochastically defined pairwise alignments between the target sequence and a single identified template as a means of exploring "alignment space" in regions of sequence with low local similarity. "Profile-profile" alignments that first generate a sequence profile of the target and systematically compare it to the sequence profiles of solved structures; the coarse-graining inherent in the profile construction is thought to reduce noise introduced by sequence drift in nonessential regions of the sequence.

## *Model generation*

Given a template and an alignment, the information contained therein must be used to generate a three-dimensional structural model of the target, represented as a set of Cartesian coordinates for each atom in the protein. Three major classes of model generation methods have been proposed.

### Fragment assembly

The original method of homology modeling relied on the assembly of a complete model from conserved structural fragments identified in closely related solved structures. For example, a modeling study of serine proteases in mammals identified a sharp distinction between "core" structural regions conserved in all experimental structures in the class, and variable regions typically located in the loops where the majority of the sequence differences were localized. Thus unsolved proteins could be modeled by first constructing the conserved core and then substituting variable regions from other proteins in the set of solved structures. Current implementations of this method differ mainly in the way they deal with regions that are not conserved or that lack a template. The variable regions are often constructed with the help of fragment libraries.

### Segment matching

The segment-matching method divides the target into a series of short segments, each of which is matched to its own template fitted from the Protein Data Bank. Thus, sequence alignment is done over segments rather than over the entire protein. Selection of the template for each segment is based on sequence similarity, comparisons of alpha carbon coordinates, and predicted steric conflicts arising from the van der Waals radii of the divergent atoms between target and template.

## Satisfaction of spatial restraints

The most common current homology modeling method takes its inspiration from calculations required to construct a three-dimensional structure from data generated by NMR spectroscopy. One or more target-template alignments are used to construct a set of geometrical criteria that are then converted to probability density functions for each restraint. Restraints applied to the main protein internal coordinates - protein backbone distances and dihedral angles - serve as the basis for a global optimization procedure that originally used conjugate gradient energy minimization to iteratively refine the positions of all heavy atoms in the protein.

This method had been dramatically expanded to apply specifically to loop modeling, which can be extremely difficult due to the high flexibility of loops in proteins in aqueous solution. A more recent expansion applies the spatial-restraint model to electron density maps derived from cryoelectron microscopy studies, which provide low-resolution information that is not usually itself sufficient to generate atomic-resolution structural models. To address the problem of inaccuracies in initial target-template sequence alignment, an iterative procedure has also been introduced to refine the alignment on the basis of the initial structural fit. The most commonly used software in spatial restraint-based modeling is MODELLER and a database called ModBase has been established for reliable models generated with it.

## *Loop modeling*

Regions of the target sequence that are not aligned to a template are modeled by loop modeling; they are the most susceptible to major modeling errors and occur with higher frequency when the target and template have low sequence identity. The coordinates of unmatched sections determined by loop modeling programs are generally much less accurate than those obtained from simply copying the coordinates of a known structure, particularly if the loop is longer than 10 residues. The first two sidechain dihedral angles ($\chi_1$ and $\chi_2$) can usually be estimated within 30° for an accurate backbone structure; however, the later dihedral angles found in longer side chains such as lysine and arginine are notoriously difficult to predict. Moreover, small errors in $\chi_1$ (and, to a lesser extent, in $\chi_2$) can cause relatively large errors in the positions of the atoms at the terminus of side chain; such atoms often have a functional importance, particularly when located near the active site.

## *Model assessment*

Assessment of homology models without reference to the true target structure is usually performed with two methods: statistical potentials or physics-based energy calculations. Both methods produce an estimate of the energy (or an energy-like analog) for the model or models being assessed; independent criteria are needed to determine acceptable cutoffs. Neither of the two methods correlates exceptionally well with true structural accuracy, especially on protein types underrepresented in the PDB, such as membrane proteins.

Statistical potentials are empirical methods based on observed residue-residue contact frequencies among proteins of known structure in the PDB. They assign a probability or energy score to each possible pairwise interaction between amino acids and combine these pairwise interaction scores into a single score for the entire model. Some such methods can also produce a residue-by-residue assessment that identifies poorly scoring regions within the model, though the model may have a reasonable score overall. These methods emphasize the hydrophobic core and solvent-exposed polar amino acids often present in globular proteins. Examples of popular statistical potentials include Prosa and DOPE. Statistical potentials are more computationally efficient than energy calculations.

Physics-based energy calculations aim to capture the interatomic interactions that are physically responsible for protein stability in solution, especially van der Waals and electrostatic interactions. These calculations are performed using a molecular mechanics force field; proteins are normally too large even for semi-empirical quantum mechanics-based calculations. The use of these methods is based on the energy landscape hypothesis of protein folding, which predicts that a protein's native state is also its energy minimum. Such methods usually employ implicit solvation, which provides a continuous approximation of a solvent bath for a single protein molecule without necessitating the explicit representation of individual solvent molecules. A force field specifically constructed for model assessment is known as the Effective Force Field (EFF) and is based on atomic parameters from CHARMM.

A very extensive model validation report can be obtained using the Radboud Universiteit Nijmegen *"What Check"* software which is one option of the Radboud Universiteit Nijmegen *"What If"* software package; it produces a many page document with extensive analyses of nearly 200 scientific and administrative aspects of the model. *"What Check"* is available as a free server; it can also be used to validate experimentally determined structures of macromolecules.

One newer method for model assessment relies on machine learning techniques such as neural nets, which may be trained to assess the structure directly or to form a consensus among multiple statistical and energy-based methods. Very recent results using support vector machine regression on a jury of more traditional assessment methods outperformed common statistical, energy-based, and machine learning methods.

## Structural comparison methods

The assessment of homology models' accuracy is straightforward when the experimental structure is known. The most common method of comparing two protein structures uses the root-mean-square deviation (RMSD) metric to measure the mean distance between the corresponding atoms in the two structures after they have been superimposed. However, RMSD does underestimate the accuracy of models in which the core is essentially correctly modeled, but some flexible loop regions are inaccurate. A method introduced for the modeling assessment experiment CASP is known as the global distance test (GDT) and measures the total number of atoms whose distance from the model to the experimental structure lies under a certain distance cutoff. Both methods can be used for any subset of atoms in the structure, but are often applied to only the alpha carbon or protein backbone atoms to minimize the noise created by poorly modeled side chain rotameric states, which most modeling methods are not optimized to predict.

## Benchmarking

Several large-scale benchmarking efforts have been made to assess the relative quality of various current homology modeling methods. CASP is a community-wide prediction experiment that runs every two years during the summer months and challenges prediction teams to submit structural models for a number of sequences whose structures have recently been solved experimentally but have not yet been published. Its partner CAFASP has run in parallel with CASP but evaluates only models produced via fully automated servers. Continuously running experiments that do not have prediction 'seasons' focus mainly on benchmarking publicly available webservers. LiveBench and EVA run continuously to assess participating servers' performance in prediction of imminently released structures from the PDB. CASP and CAFASP serve mainly as evaluations of the state of the art in modeling, while the continuous assessments seek to evaluate the model quality that would be obtained by a non-expert user employing publicly available tools.

## Accuracy

The accuracy of the structures generated by homology modeling is highly dependent on the sequence identity between target and template. Above 50% sequence identity, models tend to be reliable, with only minor errors in side chain packing and rotameric state, and an overall RMSD between the modeled and the experimental structure falling around 1 Â. This error is comparable to the typical resolution of a structure solved by NMR. In the 30-50% identity range, errors can be more severe and are often located in loops. Below 30% identity, serious errors occur, sometimes resulting in the basic fold being mis-predicted. This low-identity region is often referred to as the "twilight zone" within which homology modeling is extremely difficult, and to which it is possibly less suited than fold recognition methods.

At high sequence identities, the primary source of error in homology modeling derives from the choice of the template or templates on which the model is based, while lower identities exhibit serious errors in sequence alignment that inhibit the production of high-quality models. It has been suggested that the major impediment to quality model production is inadequacies in sequence alignment, since "optimal" structural alignments between two proteins of known structure can be used as input to current modeling methods to produce quite accurate reproductions of the original experimental structure.

Attempts have been made to improve the accuracy of homology models built with existing methods by subjecting them to molecular dynamics simulation in an effort to improve their RMSD to the experimental structure. However, current force field parameterizations may not be sufficiently accurate for this task, since homology models used as starting structures for molecular dynamics tend to produce slightly worse structures. Slight improvements have been observed in cases where significant restraints were used during the simulation.

## Sources of error

The two most common and large-scale sources of error in homology modeling are poor template selection and inaccuracies in target-template sequence alignment. Controlling for these two factors by using a structural alignment, or a sequence alignment produced on the basis of comparing two solved structures, dramatically reduces the errors in final models; these "gold standard" alignments can be used as input to current modeling methods to produce quite accurate reproductions of the original experimental structure. Results from the most recent CASP experiment suggest that "consensus" methods collecting the results of multiple fold recognition and multiple alignment searches increase the likelihood of identifying the correct template; similarly, the use of multiple templates in the model-building step may be less optimal than the use of the single correct template but more optimal than the use of a single suboptimal one. Alignment errors may be minimized by the use of a multiple alignment even if only one template is used, and by the iterative refinement of local regions of low similarity. A lesser source of model errors are errors in the template structure. The PDBREPORT database lists several million, mostly very small but occasionally dramatic, errors in experimental (template) structures that have been deposited in the PDB.

Serious local errors can arise in homology models where an insertion or deletion mutation or a gap in a solved structure result in a region of target sequence for which there is no corresponding template. This problem can be minimized by the use of multiple templates, but the method is complicated by the templates' differing local structures around the gap and by the likelihood that a missing region in one experimental structure is also missing in other structures of the same protein family. Missing regions are most common in loops where high local flexibility increases the difficulty of resolving the region by structure-determination methods. Although some guidance is provided even with a single template by the positioning of the ends of the missing region, the longer the gap, the more difficult it is to model. Loops of up to about 9 residues can be modeled with moderate accuracy in some cases if the local alignment is correct. Larger regions are often modeled individually using ab initio structure prediction techniques, although this approach has met with only isolated success.

The rotameric states of side chains and their internal packing arrangement also present difficulties in homology modeling, even in targets for which the backbone structure is relatively easy to predict. This is partly due to the fact that many side chains in crystal structures are not in their "optimal" rotameric state as a result of energetic factors in the hydrophobic core and in the packing of the individual molecules in a protein crystal. One method of addressing this problem requires searching a rotameric library to identify locally low-energy combinations of packing states. It has been suggested that a major reason that homology modeling so difficult when target-template sequence identity lies below 30% is that such proteins have broadly similar folds but widely divergent side chain packing arrangements.

## *Utility*

Uses of the structural models include protein-protein interaction prediction, protein-protein docking, molecular docking, and functional annotation of genes identified in an organism's genome. Even low-accuracy homology models can be useful for these purposes, because their inaccuracies tend to be located in the loops on the protein surface, which are normally more variable even between closely related proteins. The functional regions of the protein, especially its active site, tend to be more highly conserved and thus more accurately modeled.

Homology models can also be used to identify subtle differences between related proteins that have not all been solved structurally. For example, the method was used to identify cation binding sites on the $Na^+/K^+$ ATPase and to propose hypotheses about different ATPases' binding affinity. Used in conjunction with molecular dynamics simulations, homology models can also generate hypotheses about the kinetics and dynamics of a protein, as in studies of the ion selectivity of a potassium channel. Large-scale automated modeling of all identified protein-coding regions in a genome has been attempted for the yeast *Saccharomyces cerevisiae*, resulting in nearly 1000 quality models for proteins whose structures had not yet been determined at the time of the study, and identifying novel relationships between 236 yeast proteins and other previously solved structures.

## Chapter 10

# Equilibrium Unfolding

In biochemistry, **equilibrium unfolding** is the process of unfolding a protein or RNA molecule by gradually changing its solution conditions, i.e., its environment. Since equilibrium is maintained at all steps, the process is reversible (**equilibrium folding**). Equilibrium unfolding is used to determine the conformational stability of the molecule.

### *Theoretical background*

In its simplest form, equilibrium unfolding assumes that the molecule may belong to only two thermodynamic states, the *folded state* (typically denoted *N* for "native" state) and the unfolded state (typically denoted *U*). This "all-or-none" model of protein folding was first proposed by Tim Anson (1945), but is believed to hold only for small, single structural domains of proteins (Jackson, 1998); larger domains and multi-domain proteins often exhibit intermediate states. As usual in statistical mechanics, these states correspond to ensembles of molecular conformations, not just one conformation.

The molecule may transition between the native and unfolded states according to a simple kinetic model

$$N \rightleftharpoons U$$

with rate constants $k_f$ and $k_u$ for the folding ($U \rightarrow N$) and unfolding ($N \rightarrow U$) reactions, respectively. The dimensionless equilibrium constant $K_{eq} \overset{\text{def}}{=} \frac{k_u}{k_f} = \frac{[U]_{eq}}{[N]_{eq}}$ can be used to determine the conformational stability $\Delta G$ by the equation

$$\Delta G = -RT\ln K_{eq}$$

where *R* is the gas constant and *T* is the absolute temperature in kelvins. Thus, $\Delta G$ is positive if the unfolded state is less stable (i.e., disfavored) relative to the native state.

The most direct way to measure the conformational stability Δ*G* of a molecule with two-state folding is to measure its kinetic rate constants $k_f$ and $k_u$ under the solution conditions of interest. However, since protein folding is typically completed in milliseconds, such measurements can be difficult to perform, usually requiring expensive stopped-flow or (more recently) continuous-flow mixers to provoke folding with a high time resolution. Dual polarisation interferometry is an emerging technique to directly measure conformational change and Δ*G*.

## *Chemical denaturation*

In the less expensive technique of **equilibrium unfolding**, the fractions of folded and unfolded molecules (denoted as $p_N$ and $p_U$, respectively) are measured as the solution conditions are gradually changed from those favoring the native state to those favoring the unfolded state, e.g., by adding a denaturant such as guanidinium hydrochloride or urea. (In **equilibrium folding**, the reverse process is carried out.) Given that the fractions must sum to one and their ratio must be given by the Boltzmann factor, we have

$$p_N = \frac{1}{1+e^{-\Delta G/RT}}$$
$$p_U = \frac{e^{-\Delta G/RT}}{1+e^{-\Delta G/RT}}$$

Protein stabilities are typically found to vary linearly with the denaturant concentration. A number of models have been proposed to explain this observation prominent among them being the **denaturant binding model**, **solvent-exchange model** (both by Schellman, JA) and the **Linear Energy Model** (LEM; Pace, NC). All of the models assume that only two thermodynamic states are populated/de-populated upon denaturation. They could be extended to interpret more complicated reaction schemes.

The **denaturant binding model** assumes that there are specific but independent sites on the protein molecule (folded or unfolded) to which the denaturant binds with an effective (average) binding constant *k*. The equilibrium shifts towards the unfolded state at high denaturant concentrations as it has more binding sites for the denaturant relative to the folded state (Δ*n*). In other words, the increased number of potential sites exposed in the unfolded state is seen as the reason for denaturation transitions. An elementary treatment results in the following functional form:

$$\Delta G = \Delta G_w - RT\Delta n \ln\left(1 + k[D]\right)$$

where $\Delta G_w$ is the stability of the protein in water and [D] is the denaturant concentration. Thus the analysis of denaturation data with this model requires 7 parameters: $\Delta G_w$,Δ*n*, *k,* and the slopes and intercepts of the folded and unfolded state baselines.

The **solvent exchange model** (also called the ‘weak binding model’ or ‘selective solvation’) of Schellman invokes the idea of an equilibrium between the water molecules bound to independent sites on protein and the denaturant molecules in solution. It has the form:

$$\Delta G = \Delta G_w - RT\Delta n \ln\left(1 + (K-1)X_D\right)$$

where *K* is the equilibrium constant for the exchange reaction and $X_d$ is the mole-fraction of the denaturant in solution. This model tries to answer the question of whether the denaturant molecules actually bind to the protein or they *seem* to be bound just because denaturants occupy about 20-30 % of the total solution volume at high concentrations used in experiments, i.e. non-specific effects – and hence the term 'weak binding'. As in the denaturant-binding model, fitting to this model also requires 7 parameters. One common theme obtained from both these models is that the binding constants (in the molar scale) for urea and guanidinium hydrochloride are small: ~ 0.2 $M^{-1}$ for urea and 0.6 $M^{-1}$ for GuHCl.

Intuitively, the difference in the number of binding sites between the folded and unfolded states is directly proportional to the differences in the accessible surface area. This forms the basis for the **LEM** which assumes a simple linear dependence of stability on the denaturant concentration. The resulting slope of the plot of stability versus the denaturant concentration is called the m-value. In pure mathematical terms, m-value is the derivative of the change in stabilization free energy upon the addition of denaturant. However, a strong correlation between the accessible surface area (ASA) exposed upon unfolding, i.e. difference in the ASA between the unfolded and folded state of the studied protein (dASA), and the m-value has been documented by Pace and co-workers. In view of this observation, the m-values are typically interpreted as being proportional to the dASA. There is no physical basis for the LEM and is purely empirical, though it is widely used in interpreting solvent-denaturation data. It has the general form:

$$\Delta G = m\left([D]_{1/2} - [D]\right)$$

where the slope *m* is called the "*m*-value"(> 0 for the above definition) and $[D]_{1/2}$ (also called $C_m$) represents the denaturant concentration at which 50% of the molecules are folded (the *denaturation midpoint* of the transition, where $p_N = p_U = 1 / 2$).

In practice, the observed experimental data at different denaturant concentrations are fit to a two-state model with this functional form for Δ*G*, together with linear baselines for the folded and unfolded states. The *m* and $[D]_{1/2}$ are two fitting parameters, along with four others for the linear baselines (slope and intercept for each line); in some cases, the slopes are assumed to be zero, giving four fitting parameters in total. The conformational stability Δ*G* can be calculated for any denaturant concentration (including the stability at zero denaturant) from the fitted parameters *m* and $[D]_{1/2}$. When combined with kinetic data on folding, the *m*-value can be used to roughly estimate the amount of buried hydrophobic surface in the folding transition state.

## Structural probes

Unfortunately, the probabilities $p_N$ and $p_U$ cannot be measured directly. Instead, we assay the relative population of folded molecules using various structural probes, e.g., absorbance at 287 nm (which reports on the solvent exposure of tryptophan and tyrosine), far-ultraviolet circular dichroism (180-250 nm, which reports on the secondary structure of the protein backbone), dual polarisation interferometry (which reports the molecular size and fold density) and near-ultraviolet fluorescence (which reports on changes in the environment of tryptophan and tyrosine). However, nearly any probe of folded structure will work; since the measurement is taken at equilibrium, there is no need for high time resolution. Thus, measurements can be made of NMR chemical shifts, intrinsic viscosity, solvent exposure (chemical reactivity) of side chains such as cysteine, backbone exposure to proteases, and various hydrodynamic measurements.

To convert these observations into the probabilities $p_N$ and $p_U$, one generally assumes that the observable $A$ adopts one of two values, $A_N$ or $A_U$, corresponding to the native or unfolded state, respectively. Hence, the observed value equals the linear sum

$$A = A_N p_N + A_U p_U$$

By fitting the observations of $A$ under various solution conditions to this functional form, one can estimate $A_N$ and $A_U$, as well as the parameters of $\Delta G$. The fitting variables $A_N$ and $A_U$ are sometimes allowed to vary linearly with the solution conditions, e.g., temperature or denaturant concentration, when the asymptotes of $A$ are observed to vary linearly under strongly folding or strongly unfolding conditions.

## Thermal denaturation

Assuming a two state denaturation as stated above, one can derive the fundamental thermodynamic parameters namely, $\Delta H$, $\Delta S$ and $\Delta G$ provided one has knowledge on the $\Delta C_p$ of the system under investigation.

The thermodynamic observables of denaturation can be described by the following equations:

$$\Delta H(T) = \Delta H(T_d) + \int_{T_d}^{T} \Delta C_p dT$$

$$\rightarrow \Delta H(T) = \Delta H(T_d) + \Delta C_p[T - T_d]$$

$$\Delta S(T) = \frac{\Delta H(T_d)}{T_d} + \int_{T_d}^{T} \Delta C_p dlnT$$

$$\rightarrow \Delta S(T) = \frac{\Delta H(T_d)}{T_d} + \Delta C_p ln\frac{T}{T_d}$$

$$\Delta G(T) = \Delta H - T\Delta S$$

$$\rightarrow \Delta G(T) = \Delta H(T_d)\frac{T_d - T}{T_d} + \int_{T_d}^{T} \Delta C_p dT - T\int_{T_d}^{T} \Delta C_p dlnT$$

$$\rightarrow \Delta G(T) = \Delta H(T_d)(1 - \frac{T}{T_d}) - \Delta C_p[T_d - T + Tln(\frac{T}{T_d})]$$

where $\Delta H$ , $\Delta S$and $\Delta G$indicate the enthalpy, entropy and Gibbs free energy of unfolding under a constant pH and pressure. The temperature, $T$is varied to probe the thermal stability of the system and $T_d$is the temperature at which **half** of the molecules in the system are unfolded. The last equation is known as the Gibbs-Helmholtz equation.

## Determining the heat capacity of proteins

In principle one can calculate all the above thermodynamic observables above from a single differential scanning calorimetry thermogram of the system assuming that the $\Delta C_p$is independent of the temperature. However, it is difficult to obtain accurate values for $\Delta C_p$this way. More accurately, the $\Delta C_p$can be derived from a the variations in $\Delta H(T_d)$vs. $T_d$which can be achieved from measurements with slight variations in $pH$ or protein concentration. The slope of the linear fit is equal to the $\Delta C_p$. Note that any non-linearity of the datapoints indicates that $\Delta C_p$is probably **not** independent of the temperature.

Alternatively, the $\Delta C_p$can be estimated very accurately from the calculation of the accessible solvent area (ASA) of a protein prior and after thermal denaturation as follows:

$$\Delta ASA = ASA_{unfolded} - ASA_{native}$$

For proteins that have a known 3d structure, the $ASA_{native}$ can be calculated through computer programs such as Deepview (also known as swiss PDB viewer. The $ASA_{unfolded}$ can be calculated from tabulated values of each amino acid through the semi-empirical equation:

$$ASA_{unfolded} = a_{polar}*ASA_{polar}+a_{aromatic}*ASA_{aromatic}+a_{non-polar}*ASA_{non-polar}$$

where the subscripts polar, non-polar and aromatic indicate the parts of the 20 naturally occurring amino acids.

Finally for proteins there is a linear correlation between $\Delta ASA$ and $\Delta C_p$through the following equation:

$$\Delta C_p = 0.61 * \Delta ASA$$

### Assessing two-state unfolding

Furthermore, one can assess whether the folding proceeds according to a two-state unfolding as described above. The can be done with differential scanning calorimetry by comparing the calorimetric enthalpy of denaturation i.e. the area under the peak, $A_{peak}$ to the van 't Hoff enthalpy described as follows:

$$\Delta H_{vH}(T) = -R\frac{dlnK}{dT^{-1}}$$

at $T = T_d$the $\Delta H_{vH}(T_d)$can be described as:

$$\Delta H_{vH}(T_d) = \frac{RT_d^2 \Delta C_p^{max}}{A_{peak}}$$

When a two-state unfolding is observed the $A_{peak} = \Delta H_{vH}(T_d)$. The $\Delta C_p^{max}$ is the height of the heat capacity peak.

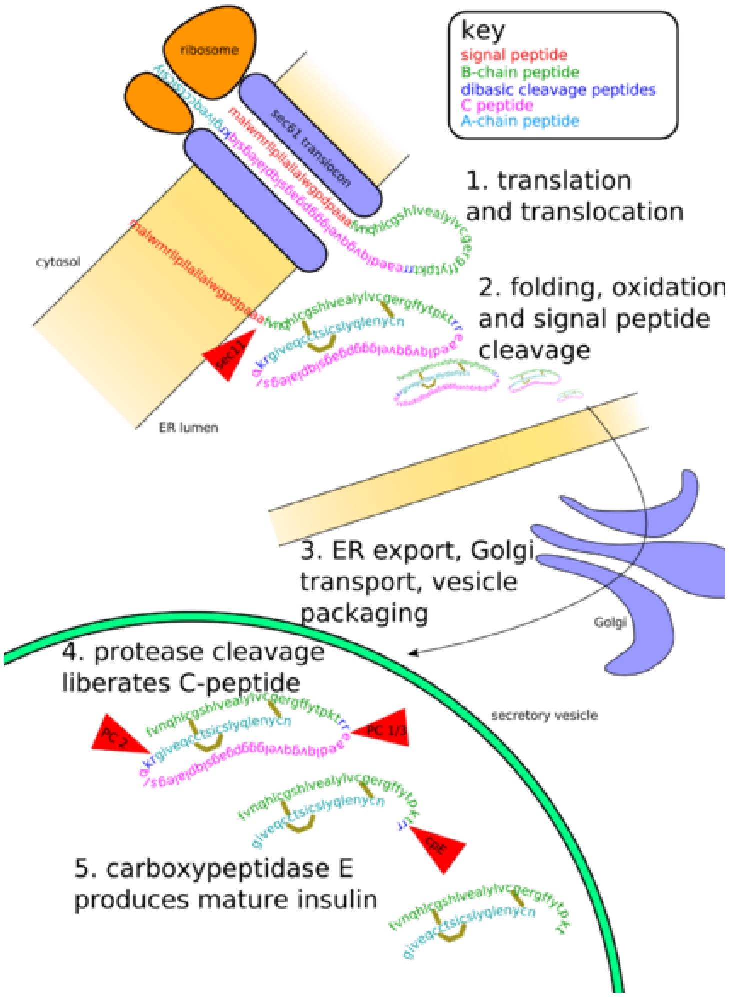
## *Other forms of denaturation*

Analogous functional forms are possible for denaturation by pressure, pH, etc.

## Chapter 11

# Posttranslational Modification

**Posttranslational modification** (PTM) is the chemical modification of a protein after its translation. It is one of the later steps in protein biosynthesis for many proteins.


key
signal peptide
B-chain peptide
dibasic cleavage peptides
C peptide
A-chain peptide
ribosome
sec61 translocon
cytosol
ER lumen
sec11
malwmrllpllallalwgpdpaaa
fvnqhlcgshlvealylvcgergffytpkt
giveqcctsicslyqlenycn
1. translation and translocation
2. folding, oxidation and signal peptide cleavage
3. ER export, Golgi transport, vesicle packaging
Golgi
secretory vesicle
4. protease cleavage liberates C-peptide
PC 2
PC 1/3
CpE
5. carboxypeptidase E produces mature insulin

The bottom of this diagram shows the modification of primary structure of insulin, as described.

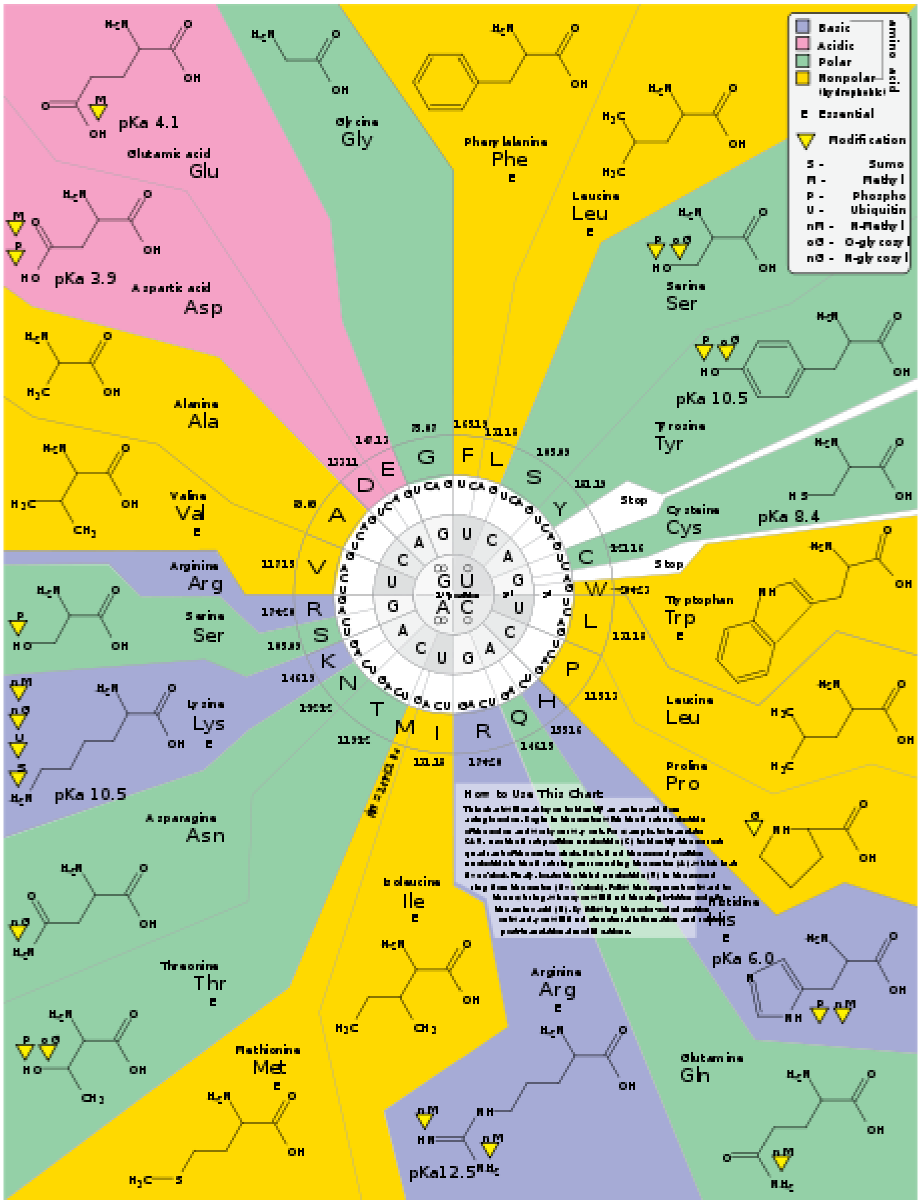
A protein (also called a polypeptide) is a chain of amino acids. During protein synthesis, 20 different amino acids can be incorporated to become a protein. After translation, the posttranslational modification of amino acids extends the range of functions of the protein by attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid (e.g. citrullination) or by making structural changes, like the formation of disulfide bridges.

Also, enzymes may remove amino acids from the amino end of the protein, or cut the peptide chain in the middle. For instance, the peptide hormone insulin is cut twice after disulfide bonds are formed, and a propeptide is removed from the middle of the chain; the resulting protein consists of two polypeptide chains connected by disulfide bonds. Also, most nascent polypeptides start with the amino acid methionine because the "start" codon on mRNA also codes for this amino acid. This amino acid is usually taken off during post-translational modification.

Other modifications, like phosphorylation, are part of common mechanisms for controlling the behavior of a protein, for instance activating or inactivating an enzyme.

Post-translational modification of proteins is detected by mass spectrometry or Eastern blotting.

## *PTMs involving addition of functional groups*



The genetic code diagram showing the amino acid residues as target of modification.

## PTMs involving addition by an enzyme *in vivo*

- acylation, e.g. *O*-acylation (esters), *N*-acylation (amides), *S*-acylation (thioesters)
  - acetylation, the addition of an acetyl group, either at the N-terminus of the protein or at lysine residues. The reverse is called deacetylation.
  - formylation
  - lipoylation, attachment of a lipoate ($C_8$) functional group
  - myristoylation, attachment of myristate, a $C_{14}$ saturated acid
  - palmitoylation, attachment of palmitate, a $C_{16}$ saturated acid
- alkylation, the addition of an alkyl group, e.g. methyl, ethyl

    - methylation the addition of a methyl group, usually at lysine or arginine residues. The reverse is called demethylation.
    - isoprenylation or prenylation, the addition of an isoprenoid group (e.g. farnesol and geranylgeraniol)
        - farnesylation
        - geranylgeranylation
- amidation at C-terminus
- amino acid addition
    - arginylation, a tRNA-mediation addition
    - polyglutamylation, covalent linkage of glutamic acid residues to tubulin and some other proteins.
    - polyglycylation, covalent linkage of one to more than 40 glycine residues to the tubulin C-terminal tail
- diphthamide formation
- gamma-carboxylation dependent on Vitamin K
- glycosylation, the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine, resulting in a glycoprotein. Distinct from glycation, which is regarded as a nonenzymatic attachment of sugars.
    - polysialylation, addition of polysialic acid, PSA, to NCAM
- glypiation, glycosylphosphatidylinositol (GPI) anchor formation
- heme moiety may be covalently attached
- hydroxylation
- hypusine formation (on conserved lysine of [EIF5A] and aIF5a)
- iodination (e.g. of thyroid hormones)
- nucleotides or derivatives thereof may be covalently attached
    - adenylation
    - ADP-ribosylation
    - flavin attachment
- nitrosylation
- S-glutathionylation
- oxidation
- phosphopantetheinylation, the addition of a 4'-phosphopantetheinyl moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis
- phosphorylation, the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine
- pyroglutamate formation
- sulfation, the addition of a sulfate group to a tyrosine.
- selenoylation (co-translational incorporation of selenium in selenoproteins)

## PTMs involving non-enzymatic additions *in vivo*

- glycation, the addition of a sugar molecule to a protein without the controlling action of an enzyme.

## PTMs involving non-enzymatic additions *in vitro*

- biotinylation, acylation of conserved lysine residues with a biotin appendage
- pegylation

## *PTMs involving addition of other proteins or peptides*

- ISGylation, the covalent linkage to the ISG15 protein (Interferon-Stimulated Gene 15)
- SUMOylation, the covalent linkage to the SUMO protein (Small Ubiquitin-related MOdifier)
- ubiquitination, the covalent linkage to the protein ubiquitin.
- Neddylation, the covalent linkage to Nedd

## *PTMs involving changing the chemical nature of amino acids*

- citrullination, or **deimination**, the conversion of arginine to citrulline
- deamidation, the conversion of glutamine to glutamic acid or asparagine to aspartic acid
- eliminylation, the conversion to an alkene by beta-elimination of phosphothreonine and phosphoserine, or dehydration of threonine and serine, as well as by decarboxylation of cysteine
- carbamylation, the conversion of lysine to homocitrulline

## *PTMs involving structural changes*

- disulfide bridges, the covalent linkage of two cysteine amino acids
- proteolytic cleavage, cleavage of a protein at a peptide bond
- racemization of proline by prolyl isomerase

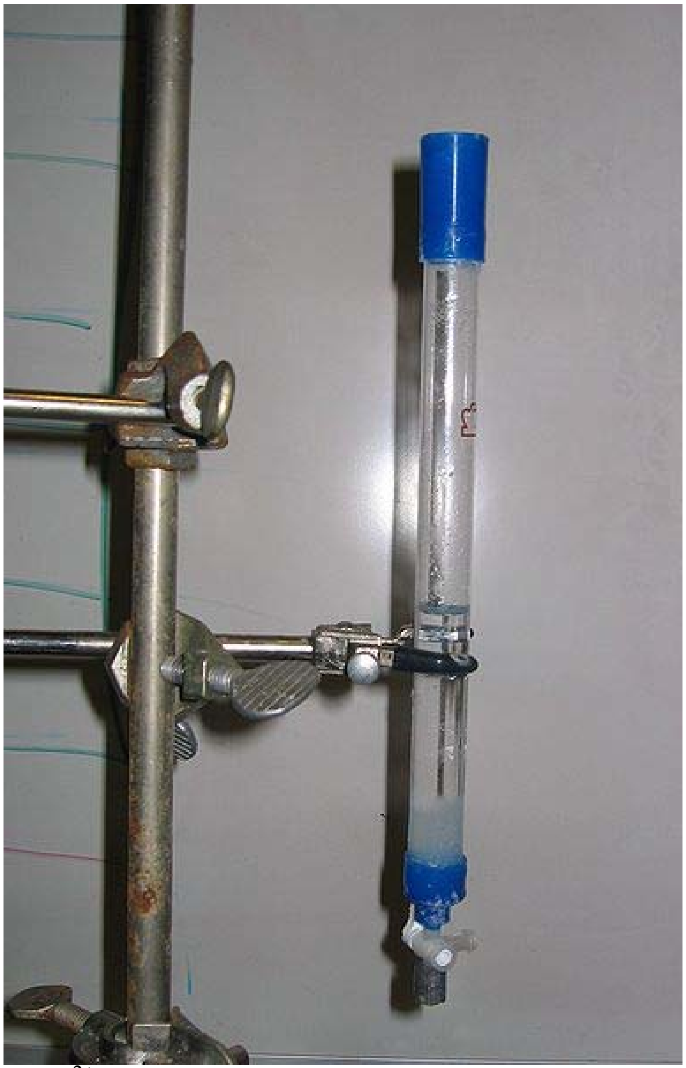
## *Case examples*

- Cleavage and formation of disulfide bridges during the production of insulin
- PTM of histones as regulation of transcription: RNA polymerase control by chromatin structure
- PTM of RNA polymerase II as regulation of transcription
- Cleavage of polypeptide chains as crucial for lectin specificity

Chapter 12

# Polyhistidine-Tag and N-Terminus

## Polyhistidine-tag



A simple column for $Ni^{2+}$-affinity chromatography. The sample and subsequent buffers are manually poured into the column.

A **polyhistidine-tag** is an amino acid motif in proteins that consists of at least five histidine (*His*) residues, often at the N- or C-terminus of the protein. It is also known as **hexa histidine-tag**, **6xHis-tag**, and by the trademarked name **His-tag** (registered by EMD Biosciences). The tag was invented by Roche, although the use of histidines and its vectors are distributed by Qiagen. Various purification kits for histidine-tagged proteins are available from Qiagen, Sigma, Thermo Scientific, GE Healthcare, Macherey-Nagel, Clontech, and others.

The use of the tag for academic users is unrestricted; however, commercial users must pay royalties to Roche. Suitable tag sequences are available free for commercial use; for example, MK(HQ)6 may be used for enhanced expression in *E. coli* and tag removal. The total number of histidine residues may vary in the tag. The his-tag may also be followed by a suitable amino acid sequence that facilitates a removal of the polyhistidine-tag using endopeptidases. This extra sequence is not necessary if exopeptidases are used to remove N-terminal His-tags (e.g., Qiagen TAGZyme). Furthermore, exopeptidase cleavage may solve the unspecific cleavage observed when using endoprotease-based tag removal. Polyhistidine-tags are often used for affinity purification of genetically modified proteins.

## *Applications*

### Protein purification

Polyhistidine-tags are often used for affinity purification of polyhistidine-tagged recombinant proteins expressed in *Escherichia coli* and other prokaryotic expression systems. Bacterial cells are harvested via centrifugation and the resulting cell pellet lysed either by physical means or by means of detergents and enzymes such as lysozyme. At this stage raw lysate contains the recombinant protein among many other proteins originating from the bacterial host. This mixture is incubated with affinity media such as Ni Sepharose, NTA-agarose, His60 Ni, HisPur resin, or TALON resin. Affinity media contain bound metal ions, either nickel or cobalt to which the polyhistidine-tag binds with **micromolar** affinity. The resin is then washed with phosphate buffer to remove proteins that do not specifically interact with the cobalt or nickel ion. Washing efficiency can be improved by the addition of 20 mM imidazole (proteins are usually eluted with 150-300 mM imidazole). Generally nickel based resins have higher binding capacity, while cobalt based resins offer the highest purity. The purity and amount of protein can be assessed by SDS-PAGE and Western blotting.

Affinity purification using a polyhistidine-tag usually results in relatively pure protein when the recombinant protein is expressed in prokaryotic organisms. Depending on downstream applications including the purification of protein complexes to study protein interactions, purification from higher organisms such as yeasts or other eukaryotes may require a tandem affinity purification using two tags to yield higher purity. Alternatively, single-step purification using immobilized cobalt ions rather than nickel ions generally yields a substantial increase in purity and requires lower imidazole concentrations for elution of the his-tagged protein.

Polyhistidine-tagging is the option of choice for purifying recombinant proteins in denaturing conditions because its mode of action is dependent only on the primary structure of proteins. Generally for this sort of a technique, histidine binding is titrated using pH instead of imidazole binding -- at a high pH histidine binds to nickel or cobalt but at low pH (~6 for cobalt and ~4 for nickel) histidine becomes protonated and is competed off of the metal ion. Compare this to antibody purification and GST purification a prerequisite to which is the proper (native) folding of proteins involved.

Polyhistidine-tag columns retain several well known proteins as impurities. One of them is FKBP-type peptidyl prolyl isomerase, which appears around 25kDa (SlyD). Impurities are generally eliminated using a secondary chromatographic technique, or by expressing the recombinant protein in a SlyD-deficient *E. coli* strain. Alternatively cobalt based resins do not bind SlyD from E. coli and can be used for a single step purification .

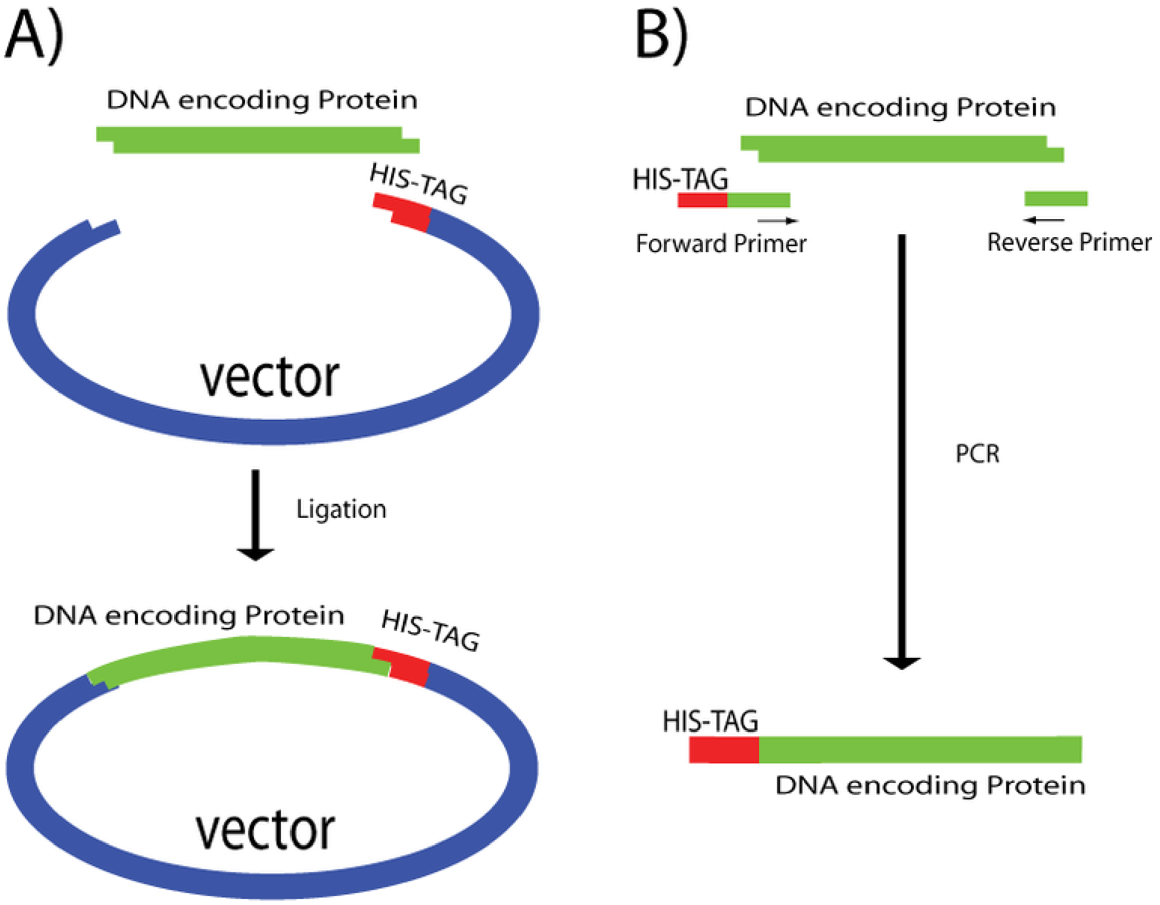
## Binding assays

Polyhistidine-tagging can be used to detect protein-protein interactions in the same way as a pull-down assay. However, this technique is generally considered to be less sensitive, and also restricted by some of the more finicky aspects of this technique. For example, reducing conditions cannot be used, EDTA and many types of detergents cannot be used. Recent advances in dual polarisation interferometry is amenable to EDTA and a wider use of reagents and the use of such site specific tags greatly simplifies the direct measurement of associated conformational change.

## Fluorescent Tags

Hexahistadine CyDye tags have also been developed. These use Nickel covalent coordination to EDTA groups attached to fluorophores in order to create dyes that attach to the polyhistidine tag. This technique has been shown to be effective for following protein migration and trafficking. There has also been recent discoveries that show this technique may be effective in order to measure distance via Fluorescent Resonance Energy Transfer.

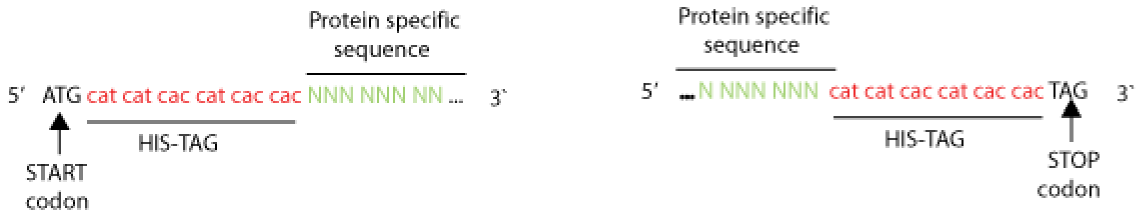
### *Adding Polyhistidine Tags*



*Adding polyhistidine tags*. A) The His-tag is added by inserting the DNA encoding a protein of interest in a vector that has the tag ready to fuse at the C-terminal. B) The His-tag is added using primers containing the tag, after a PCR reaction the tag gets fused to the N-terminal of the gene.

The most common polyhistidine tags are formed of six histidine (6xHis tag) residues which are added at the C-terminal or N-terminal of the protein of interest. The choice of the end where His-tag is added will depend mainly on the characteristics of the protein and the methods chosen to remove the tag. Some ends are buried inside the protein core and others are important for the protein function or structure. In those cases the choice is limited to the other end. On the other hand most available exopeptidases can only remove the His-tag from the N-terminal, therefore removing the tag from the C-terminal will require the use of other techniques.

There are two ways to add polyhistidines. The most simple is to insert the DNA encoding the protein in a vector encoding a His-tag so that it will be automatically attached to one of its ends. Another technique is to perform a PCR with primers that have repetitive histidine codons (CAT or CAC) right next to the START or STOP codon in addition to several (16 or more) bases from one end of the DNA encoding the protein to be tagged.


Protein specific sequence
5' ATG cat cat cac cat cac cac NNN NNN NN ... 3'
HIS-TAG
START codon
Protein specific sequence
5' ...N NNN NNN cat cat cac cat cac cac TAG 3'
HIS-TAG
STOP codon

*Example of primer designed to add a 6xHis-tag using PCR*. Eighteen bases coding six histidines are inserted right after the START codon or right before the STOP codon. At least 16 bases specific to the gene of interest are needed next to the His-tag. With 6 His the protein will have an added 1 kDa of molecular weight.

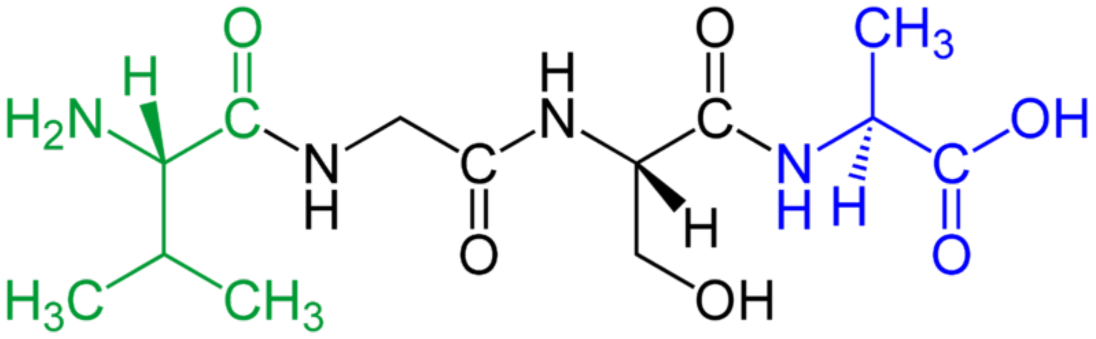
### *Detection*

The polyhistidine-tag can also be used to detect the protein via anti-polyhistidine-tag antibodies or alternatively by in-gel staining (SDS-PAGE) with fluorescent probes bearing metal ions. This can be useful in subcellular localization, ELISA, western blotting or other immuno-analytical methods.

### *Immobilization*

The polyhistidine-tag can be ideally used for the immobilization of proteins on a surface such as on a nickel or cobalt coated microtiter plate or on a protein array.

## N-terminus



A tetrapeptide (example: Val-Gly-Ser-Ala) with **green** highlighted *N*-terminal α-amino acid (example: L-valine) and **blue** marked *C*-terminal α-amino acid (examplel: L-alanine).

The **N-terminus** (also known as the **amino-terminus**, **NH$_2$-terminus**, **N-terminal end** or **amine-terminus**) refers to the start of a protein or polypeptide terminated by an amino acid with a free amine group (-NH$_2$). The convention for writing peptide sequences is to put the N-terminus on the left and write the sequence from N- to C-terminus. When the protein is translated from messenger RNA, it is created from N-terminus to C-terminus.

## *Chemistry*

Each amino acid has a carboxyl group and an amine group, and amino acids link to one another to form a chain by a dehydration reaction by joining the amine group of one amino acid to the carboxyl group of the next. Thus polypeptide chains have an end with an unbound carboxyl group, the C-terminus, and an end with an amine group, the N-terminus.

When the protein is translated from messenger RNA, it is created from N-terminus to C-terminus. The amino end of an amino acid (on a charged tRNA) during the elongation stage of translation, attaches to the carboxyl end of the growing or nascent chain. Since the start codon of the genetic code codes for the amino acid methionine, most protein sequences start with a methionine (more specifically: the modified version *N*-formylmethionine, fMet). However, some proteins are modified posttranslationally, for example by cleavage from a protein precursor, and therefore may have different amino acids at their N-terminus.

## *Function*

### N-terminal targeting signals

The N-terminus is the first part of the protein that exits the ribosome during protein biosynthesis. It often contains signal peptide sequences, "intracellular postal codes" that direct delivery of the protein to the proper organelle. The signal peptide is typically removed at the destination by a signal peptidase. The N-terminal amino acid of a protein is an important determinant of its half-life (likelihood of being degraded). This is called the N-end rule.

- **Signal peptide**

The N-terminal signal peptide is recognized by the signal recognition particle (SRP) and results in the targeting of the protein to the secretory pathway. In eukaryotic cells, these proteins are synthesized at the rough endoplasmic reticulum. In prokaryotic cells, the proteins are exported across the cell membrane. In chloroplasts, signal peptides target proteins to the thylakoids.

- **Mitochondrial targeting peptide**

The N-terminal mitochondrial targeting peptide (mtTP) allows for the protein to be imported into the mitochondrion.

- **Chloroplast targeting peptide**

The N-terminal chloroplast targeting peptide (cpTP) allows for the protein to be imported into the chloroplast.

## N-terminal modifications

Some proteins are modified posttranslationally by the addition of membrane anchors that allow the protein to associate with membrane without having a transmembrane domain. The N-terminus (as well as the C-terminus) of a protein can be modified this way.

- **N-Myristoylation**

The N-terminus can be modified by the addition of a myristoyl anchor. Proteins that are modified this way contain a consensus motif at their N-terminus as a modification signal.

- **N-Acylation**

The N-terminus can also be modified by the addition of a fatty acid anchor to form N-acylated proteins. The most common form of such modification is the addition of a palmitoyl group.

## Chapter 13

# HB Plot and Downhill Folding

## HB plot

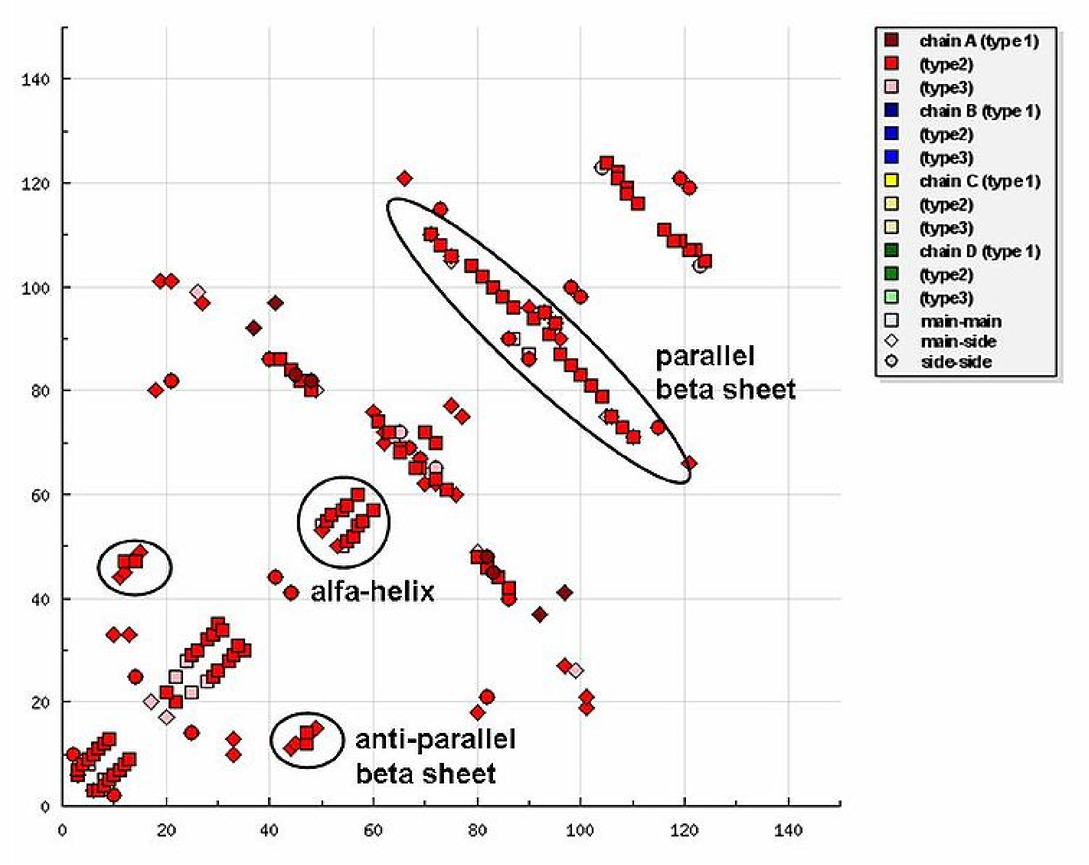
Knowledge of the relationship between a protein's structure and its dynamic behavior is essential for understanding protein function. The description of a protein three dimensional structure as a network of hydrogen bonding interactions (**HB plot**) was introduced as a tool for exploring protein structure and function. By analyzing the network of tertiary interactions the possible spread of information within a protein can be investigated.

HB plot offers a simple way of analyzing protein secondary structure and tertiary structure. Hydrogen bonds stabilizing secondary structural elements (**secondary hydrogen bonds**) and those formed between distant amino acid residues - defined as **tertiary hydrogen bonds** - can be easily distinguished in HB plot, thus, amino acid residues involved in stabilizing protein structure and function can be identified. By analyzing the network of tertiary interactions the *possible spread of information* within a protein can be investigated as well.

### *Features*

The plot distinguishes between main chain-main chain, main chain-side chain and side chain-side chain hydrogen bonding interactions. Bifurcated hydrogen bonds and multiple hydrogen bonds between amino acid residues; and intra- and interchain hydrogen bonds are also indicated on the plots. Three classes of hydrogen bondings are distinguished by color coding; short (distance smaller than 2.5 Å between donor and acceptor), intermediate (between 2.5 Å and 3.2 Å) and long hydrogen bonds (greater than 3.2 Å).

## *Secondary structure elements in HB plot*



Secondary structure elements in HB plot

In representations of the HB plot, characteristic patterns of secondary structure elements can be recognised easily, as follows:

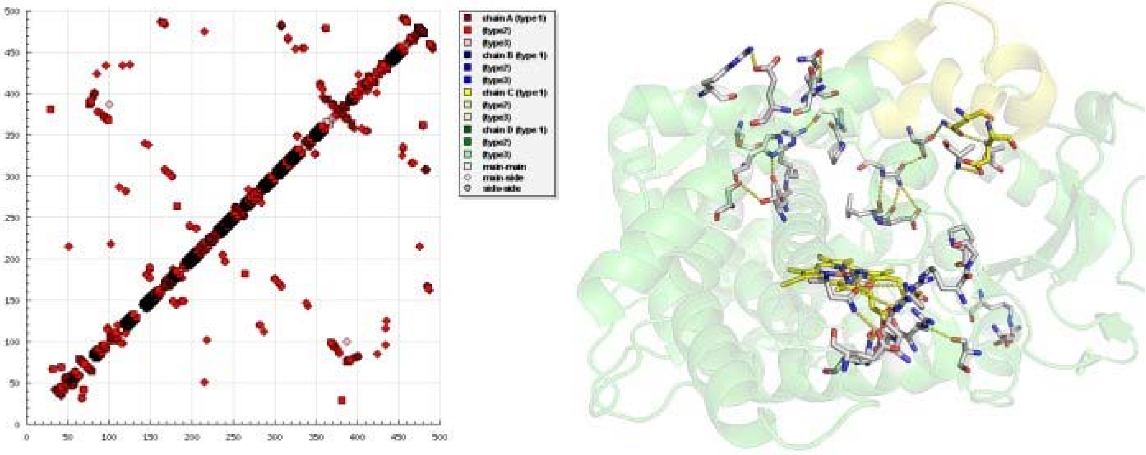
1. Helices can be identified as strips directly adjacent to the diagonal.
2. Antiparallel beta sheets appear in HB plot as cross-diagonal.
3. Parallel beta sheets appears in the HB plot as parallel to the diagonal.
4. Loops appear as breaks in the diagonal between the cross-diagonal beta-sheet motifs.

## *Examples of usage*

### Cytochrome P450s

**The cytochrome P450s (P450s)** are xenobiotic-metabolizing membrane-bound heme-containing enzymes that use molecular oxygen and electrons from NADPH cytochrome P450 reductase to oxidize their substrates. CYP2B4, a member of the cytochrome P450 family is the only protein within this family, whose X-ray structure in both open 11 and closed form 12 is published. The comparison of the open and closed structures of

CYP2B4 structures reveals large-scale conformational rearrangement between the two states, with the greatest conformational change around the residues 215-225, which is widely open in ligand-free state and shut after ligand binding; and the region around loop C near the heme.



HB Plot and structure of Cytochrome P450 2B4 in closed form

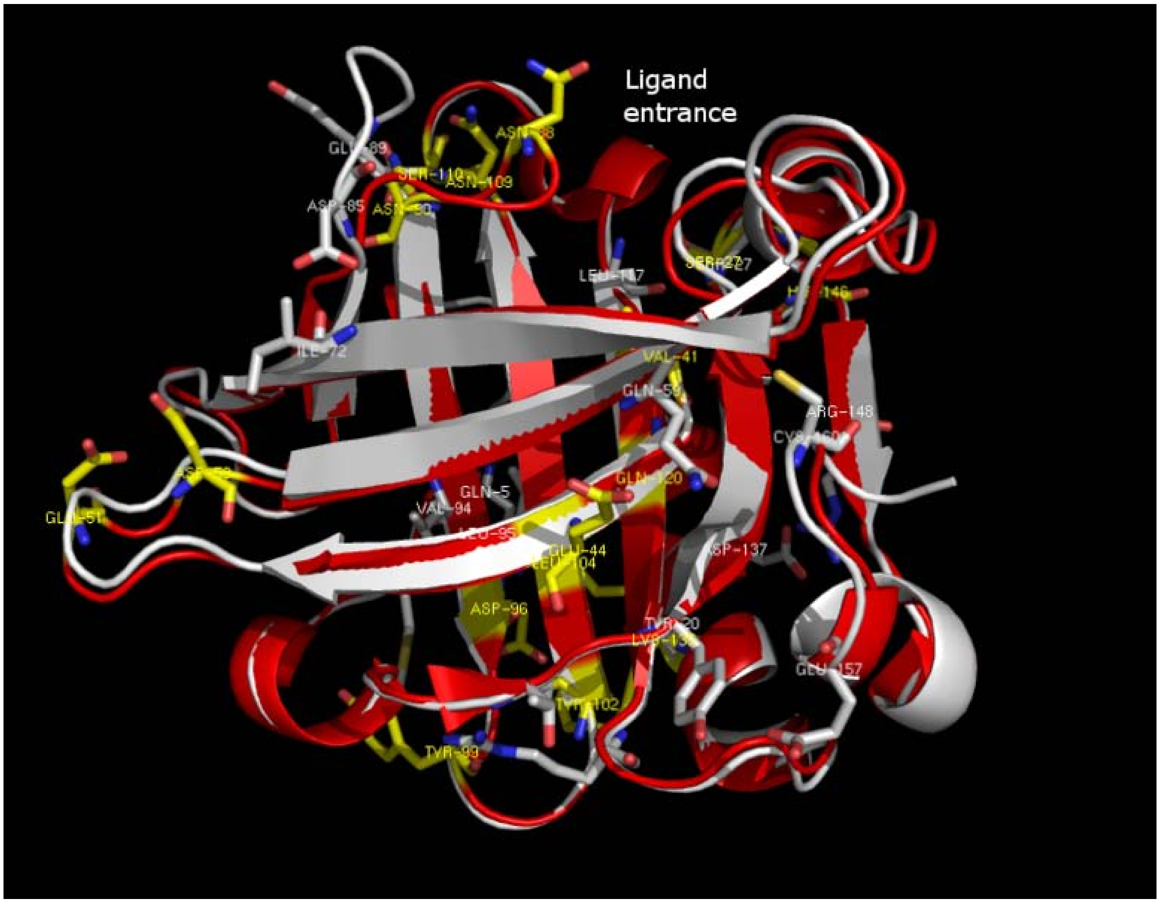
Examining the HB plot of the closed and open state of CYP2B4 revealed that the rearrangement of tertiary hydrogen bonds was in excellent agreement with the current knowledge of the cytochrome P450 catalytic cycle.

The first step in P450 catalytic cycle is identified as substrate binding. Preliminary binding of a ligand near to the entrance breaks hydrogen bonds S212-E474, S207-H172 in the open form of CYP2B4 and hydrogen bonds E218-A102, Q215-L51 are formed that fix the entrance in the closed form as the HB plot reveals.

The second step is the transfer of the first electron from NADPH via an electron transfer chain. For the electron transfer a conformational change occurs that triggers interaction of the P450 with the NADPH cytochrome P450 reductase. Breaking of hydrogen bonds between S128-N287, S128-T291, L124-N287 and forming S96-R434, A116-R434, R125-I435, D82-R400 at the NADPH cytochrome P450 reductase binding site—as seen in HB plot—transform CYP2B4 to a conformation state, where binding of NADPH cytochrome P450 reductase occurs.

In the third step, oxygen enters CYP2B4 in the closed state - the state where newly formed hydrogen bonds S176-T300, H172-S304, N167-R308 open a tunnel which is exactly the size and shape of an oxygen molecule.

## Lipocalin family



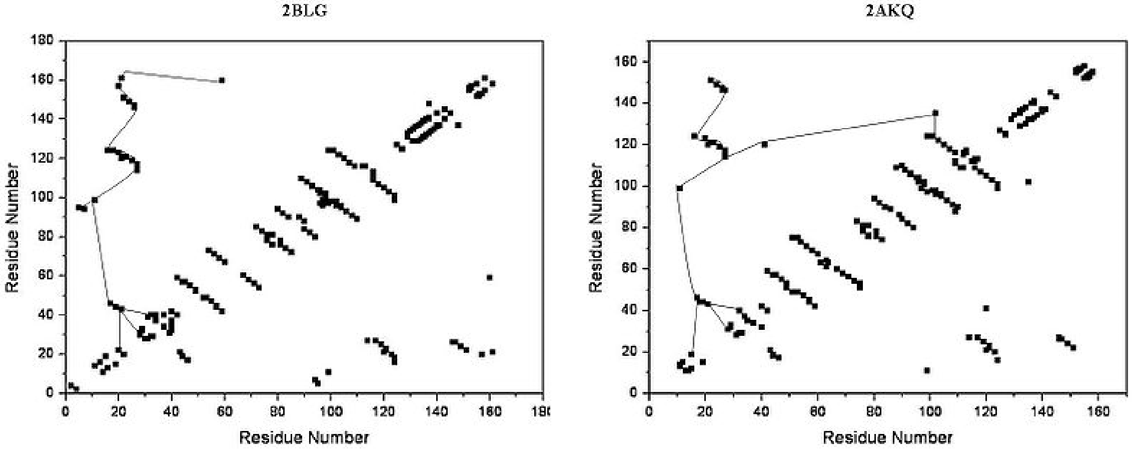
Beta-lactoglobulin in open (white) and ligand-bound (red) form

The **lipocalin family** is a large and diverse family of proteins with functions as small hydrophobic molecule transporters. Beta-lactoglobulin is a typical member of the lipocalin family. Beta-lactoglobulin was found to have a role in the transport of hydrophobic ligands such as retinol or fatty acids. Its crystal structure were determined [e.g. Qin, 1998] with different ligands and in ligand-free form as well. The crystal structures determined so far reveal that the typical lipocalin contains eight-stranded antiparallel-barrel arranged to form a conical central cavity in which the hydrophobic ligand is bound. The structure of beta-lactoglobulin reveals that the barrel-form structure with the central cavity of the protein has an "entrance" surrounded by five beta-loops with centers around 26, 35, 63, 87, and 111, which undergo a conformational change during the ligand binding and close the cavity.

The overall shape of beta-lactoglobulin is characteristic of the lipocalin family. In the absence of alpha-helices, the main diagonal almost disappears and the cross-diagonals representing the beta-sheets dominate the plot. Relatively low number of tertiary hydrogen bonds can be found in the plot, with three high-density regions, one of which is connected to a loop at the residues around 63, a second is connected to the loop around

87, and a third region which is connected to the regions 26 and 35. The fifth loop around 111 is represented only one tertiary hydrogen bond in the HB plot.

In the three-dimensional structure, tertiary hydrogen bonds are formed (1) near to the entrance, directly involved in conformational rearrangement during ligand binding; and (2) at the bottom of the "barrel". HB plots of the open and closed forms of beta-lactoglobulin are very similar, all unique motifs can be recognized in both forms. Difference in HB plots of open and ligand-bound form show few important individual changes in tertiary hydrogen bonding pattern. Especially, the formation of hydrogen bonds between Y20-E157 and S21-H161 in closed form might be crucial in conformational rearrangement. These hydrogen bonds lie at the bottom of the cavity, which suggests that the closure of the entrance of a lipocalin starts when a ligand reached the bottom of the cavity and broke hydrogen bonds R123-Y99, R123-T18, and V41-Q120. Lipocalins are known to have very low sequence similarity with high structural similarity. The only conserved regions are exactly the region around 20 and 160 with an unknown role.



HB Plots of beta-lactoglobulin in open (2BLG) and ligand-bound (2AKQ) form

# Downhill folding

**Downhill folding** is a process in which a protein folds without encountering any significant macroscopic free energy barrier. It is a key prediction of the folding funnel hypothesis of the energy landscape theory of proteins.

## *Overview*

Downhill folding is predicted to occur under conditions of extreme native bias, i.e. at low temperatures or in the absence of denaturants. This corresponds to the *type 0* scenario in the energy landscape theory. At temperatures or denaturant concentrations close to their apparent midpoints, proteins may switch from downhill to two-state folding, the *type 0* to *type 1* transition.

Global downhill folding (or *one-state folding*) is another scenario in which the protein folds in the absence of a free energy barrier under all conditions. In other words, there is a unimodal population distribution at all temperatures and denaturant concentrations, suggesting a continuous unfolding transition in which different ensembles of structures populate at different conditions. This is in contrast to two-state folding, which assumes only two ensembles (folded and unfolded) and a sharp unfolding transition.

Free energy barriers in protein folding are predicted to be small because they arise as a result of compensation between large energetic and entropic terms. Non-synchronization between gain in stabilizing energy and loss in conformational entropy results in two-state folding, while a synchronization between these two terms as the folding proceeds results in downhill folding.

## Experimental studies

Transition state structures in two-state folding are not experimentally accessible (by definition they are the least populated along the reaction co-ordinate), but the folding sub-ensembles in downhill folding processes are theoretically distinguishable by spectroscopy. The 40-residue protein BBL, which is an independently folding domain from the E2 subunit of the 2-oxoglutarate dehydrogenase multi-enzyme complex of E. coli, has been experimentally shown to fold globally downhill. Also, a mutant of lambda repressor protein has been shown to shift from downhill to two-state upon changing the temperature/solvent conditions. However, the status of BBL as a downhill-folding protein, and by extension the existence of naturally occurring downhill folders, has been controversial. The current controversy arises from the fact that the only way a protein can be labeled as two-state or downhill is by analyzing the experimental data with models that explicitly deal with these two situations, i.e. by allowing the barrier heights to vary. Unfortunately, most of the experimental data so far have been analyzed with a simple chemical two-state model. In other words, the presence of a rather large free energy barrier has been pre-assumed, ruling out the possibility of identifying downhill or globally downhill protein folding. This is critical because any sigmoidal unfolding curve, irrespective of the degree of cooperativity, can be fit to a two-state model. Kinetically, the presence of a barrier guarantees a single-exponential, but not vice-versa. Nevertheless, in some proteins such as the yeast phosphoglycerate kinase and a mutant human ubiquitin, non-exponential kinetics suggesting downhill folding have been observed.

A proposed solution to these problems is to develop models that can differentiate between the different situations, and identify simple but robust experimental criteria for identifying downhill folding proteins. These are outlined below.

## *Equilibrium criteria*

### Differences in apparent melting temperatures

An analysis based on an extension of Zwanzig's model of protein folding indicates that global downhill folding proteins should reveal different apparent melting temperatures (Tms) when monitored by different techniques. This was experimentally confirmed in the protein BBL mentioned above. The unfolding followed by differential scanning calorimetry (DSC), circular dichroism (CD), fluorescence resonance energy transfer (FRET) and fluorescence all revealed different apparent melting temperatures. A wavelength-dependent melting temperature was also observed in the CD experiments. The data analyzed with a structure-based statistical mechanical model resulted in a unimodal population distribution at all temperatures, indicating a structurally uncoupled continuous unfolding process. The crucial issue in such experiments is to use probes that monitor different aspects of the structure. For example, DSC gives information on the heat capacity changes (and hence enthalpy) associated with unfolding, fluorescence on the immediate environment of the fluorophore, FRET on the average dimensions of the molecule and CD on the secondary structure.

A more stringent test would involve following the chemical shifts of each and every atom in the molecule by nuclear magnetic resonance (NMR) as a function of temperature/denaturant. Though time-consuming, this method does not require any specific model for the interpretation of data. The Tms for all the atoms should be identical within experimental error if the protein folds in a two-state manner. But for a protein that folds globally downhill the unfolding curves should have widely different Tms. The atomic unfolding behavior of BBL was found to follow the latter, showing a large spread in the Tms consistent with global downhill behavior. The Tms of some atoms were found to be similar to that of the global Tm (obtained from a low-resolution technique like CD or fluorescence), indicating that the unfolding of multiple atoms has to be followed, instead of a few as is frequently done in such experiments. The average atomic unfolding behavior was strikingly similar to that of CD, underlining the fact that unfolding curves of low resolution experiments are highly simplified representations of a more complex behavior.

### Calorimetry and crossing baselines

Baselines frequently used in two-state fits correspond to the fluctuations in the folded or unfolded well. They are purely empirical as there is little or no information on how the folded or unfolded states' property changes with temperature/chemical denaturant. This assumes even more importance in case of DSC experiments as the changes in heat capacity correspond to both fluctuations in the protein ensemble and exposure of hydrophobic residues upon unfolding. The DSC profiles of many small fast-folding proteins are broad, with steep pre-transition slopes. Interestingly, two-state fits to these profiles result in crossing of baselines indicating that the two-state assumption is no longer valid. This was recognized by Munoz and Sanchez-Ruiz, resulting in the development of the variable-barrier model. Instead of attempting a model-free inversion of the DSC profile to extract the underlying probability density function, they assumed a specific free energy functional with either one or two minima (similar to the Landau theory of phase transitions) thus enabling the extraction of free energy barrier heights. This model is the first of its kind in physical biochemistry that enables the determination of barrier heights from equilibrium experiments. Analysis of the DSC profile of BBL with this model resulted in zero barrier height, i.e. downhill folding, confirming the earlier result from the statistical mechanical model. When the variable-barrier model was applied to a set of proteins for which both the rate and DSC data are available, a very high correlation of 0.95 was obtained between the rates and barrier heights. Many of the proteins examined had small barriers (<20 kJ/mol) with baseline crossing evident for proteins that fold faster than 1 ms. This is in contrast to the traditional assumption that the free energy barrier between the folded and unfolded states are large.

## *Simulations*

Because downhill folding is difficult to measure experimentally, molecular dynamics and Monte Carlo simulations have been performed on fast-folding proteins to explore their folding kinetics. Proteins whose folding rate is at or near the folding "speed limit", whose timescales make their folding more accessible to simulation methods, may more commonly fold downhill. Simulation studies of the BBL protein imply that its rapid folding rate and very low energy barrier arise from a lack of cooperativity in the formation of native contacts during the folding process; that is, a low contact order. The link between lack of cooperativity and low contact order was also observed in the context of Monte Carlo lattice simulations These data suggest that the average number of "nonlocal contacts" per residue in a protein serves as an indicator of the barrier height, where very low nonlocal contact values imply downhill folding. Coarse-grained simulations by Knott and Chan also support the experimental observation of global downhill folding in BBL.