SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit-1 - Protein Engineering and Bioinformatics – SBTA1303

I Protein Structure

The spatial arrangement of atoms in a protein is called a conformation. The term conformation refers to a structural state that can, without breaking any covalent bonds, interconvert with other structural states. A change in conformation could occur, for example, by rotation about single bonds. Of the innumerable conformations that are theoretically possible in a protein containing hundreds of single bonds, one generally predominates. This is usually the conformation that is thermodynamically the most stable, having the lowest Gibbs' free energy (G). Proteins in their functional conformation are called native proteins.

There Are Four Levels of Architecture in Proteins

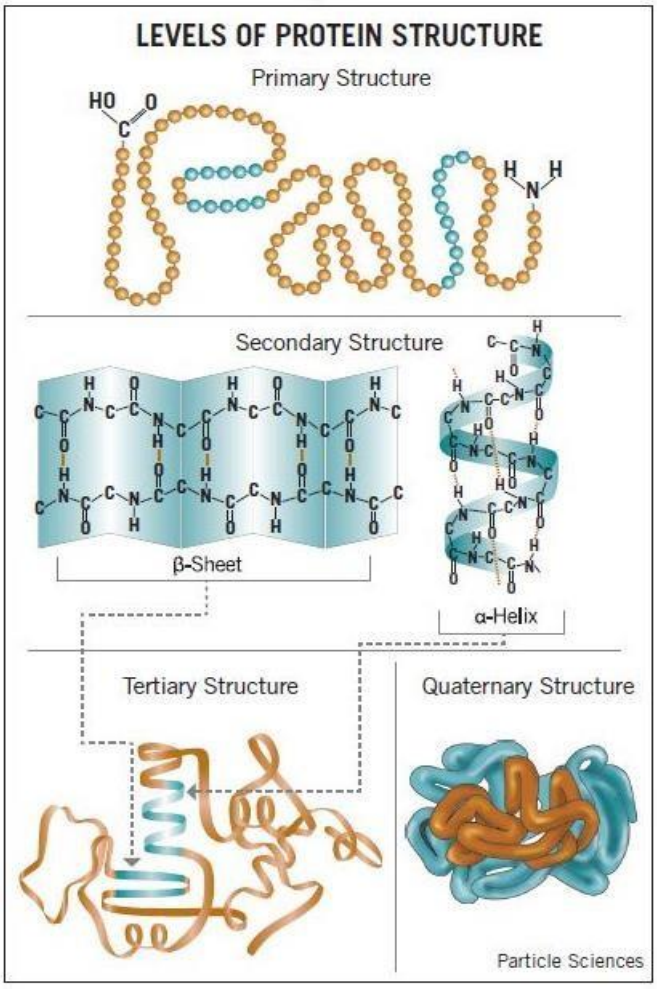


Figure 1 Levels of structure in proteins

Conceptually, protein structure can be considered at four levels (Fig. 1). **Primary structure** includes all the covalent bonds between amino acids and is normally defined by the sequence of peptide-bonded amino acids and locations of disulfide bonds. The relative spatial arrangement of the linked amino acids is unspecified. Polypeptide chains are not free to take up any three-dimensional structure at random. Steric constraints and many weak interactions stipulate that some arrangements will be more stable than others. **Secondary structure** refers to regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain. There are a few common types of secondary structure, the most prominent being the a helix and the β conformation.

Tertiary structure refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional structure of the polypeptide. The boundary between secondary and tertiary structure is not always clear. Several different types of secondary structure are often found within the three-dimensional structure of a large protein. Proteins with several polypeptide chains have one more level of structure: **quaternary structure**, which refers to the spatial relationship of the polypeptides, or subunits, within the protein.

Protein Secondary Structure

Several types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the α helix and β conformations. Using fundamental chemical principles and a few experimental observations, Linus Pauling and Robert Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated.

In considering secondary structure, it is useful to classify proteins into two major groups: fibrous proteins, having polypeptide chains arranged in long strands or sheets, and globular proteins, with polypeptide chains folded into a spherical or globular shape. Fibrous proteins play important structural roles in the anatomy and physiology of vertebrates, providing external protection, support, shape, and form. They may constitute one-half or more of the total body protein in larger animals. Most enzymes and peptide hormones are globular proteins. Globular proteins tend to be structurally complex, often containing several types of secondary structure; fibrous proteins usually consist largely of a single type of secondary structure. Because of this structural simplicity, certain fibrous proteins played a key role in the development of the modern understanding of protein

structure and provide particularly clear examples of the relationship between structure and function; they are considered in some detail after the general discussion of secondary structure.

The Peptide Bond is Rigid and Planar

In the peptide bond, the π-electrons from the carbonyl are delocalized between the oxygen and the nitrogen. This means that the peptide bond has ~40% double bond character. This partial double bond character is evident in the shortened bond length of the C–N bond. The length of a normal C–N single bond is 1.45 Å and a C=N double bond is 1.25 Å, while the peptide C–N bond length is 1.33 Å.

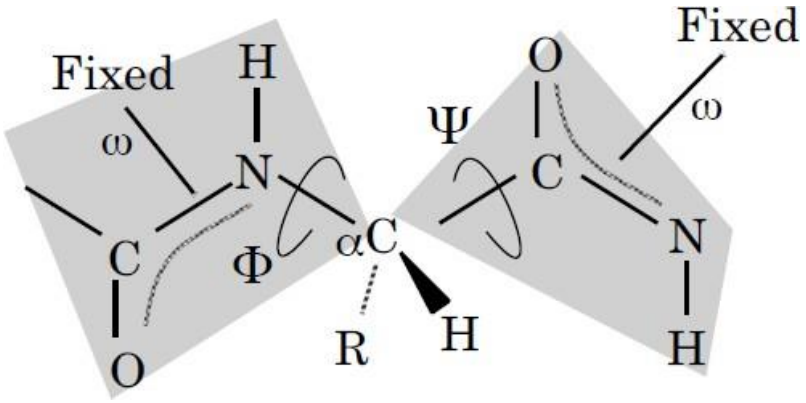


Figure 2 Peptide Bond – Dihedral angles

Because of its partial double bond character, rotation around the N–C bond is severely restricted. The peptide bond allows rotation about the bonds from the α- carbon, but not the amide C–N bond. Only the Φ and Ψ torsion angles can vary reasonably freely. In addition, the six atoms in the peptide bond (the two α-carbons, the amide O, and the amide N and H) are coplanar. Finally, the peptide bond has a dipole, with the O having a partial negative charge, and the N amide having a partial positive charge.

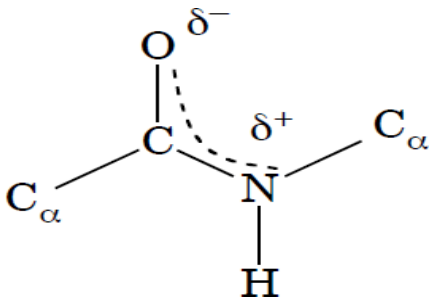


Figure 3 Peptide Bond – Dipole

This allows the peptide bond to participate in electrostatic interactions, and contributes to the hydrogen bond strength between the backbone carbonyl and the Namide proton.

Peptide bond and protein structure

The peptide bond contains three sets of torsion angles (also known as dihedral angles). The least variable of these torsion angles is the ω angle, which is the dihedral angle around the amide bond. As discussed above, this angle is fixed by the requirement for orbital overlap between the carbonyl double bond and the Namide lone pair orbital. Steric considerations strongly favor the trans configuration (i.e. an ω angle of 180°), because of steric hindrance between the alpha carbons of adjacent amino acid residues. This means that nearly all peptide bonds in a protein will have an ω angle of 180°.

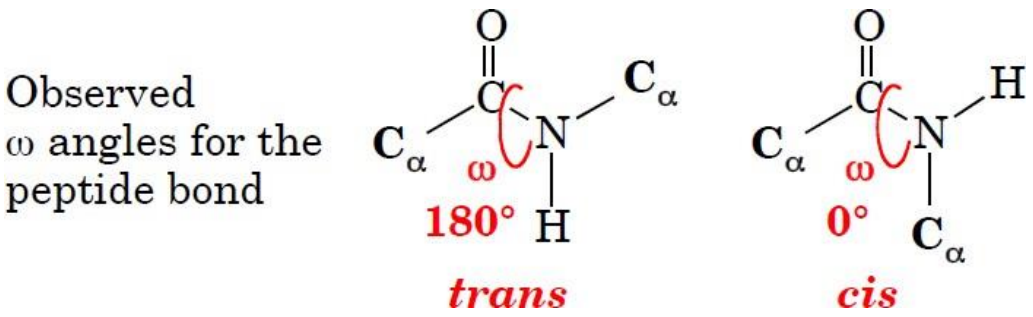


Figure 4 Peptide Bond – ω angle

In considering peptide structures, it is usually much more important to look at the backbone angles that can vary more widely. These angles are the Φ (= phi, Cα–Namide) and Ψ (= psi, Cα–Camide) angles. By definition, the fully extended conformation corresponds to 180° for both Φ and Ψ. (Note that 180° = –180°). Numeric values of angles increase in the clockwise direction when looking away from the α-carbon

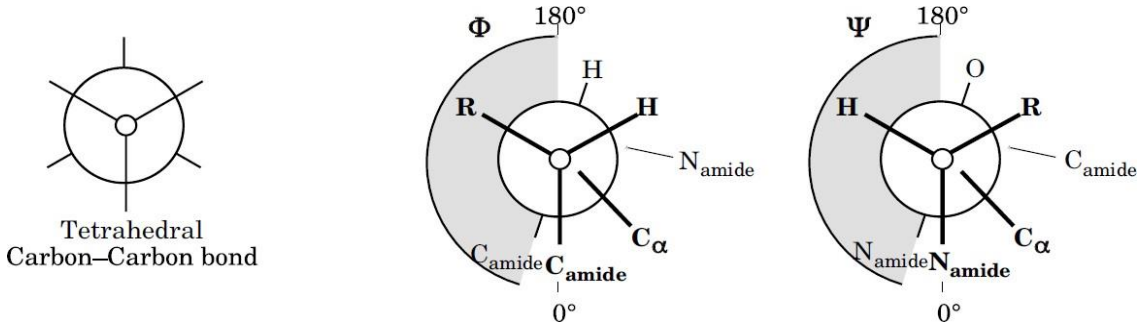


Figure 5 Staggered conformations

By definition, Φ = 0° when the Camide-Namide and Camide-Cα bonds are in the same plane, and Ψ = 0° when the Namide-Camide and Namide-Cα bonds are in the same plane. The (+) direction is clockwise while looking away from the Cα. The torsion angles that the atoms of the peptide bond can assume are limited by steric constraints. Some Φ / Ψ pairs will result in atoms being closer than allowed by the van der Waals radii of the atoms, and are therefore sterically forbidden (for example: 0°:0°, 180°:0°, and 0°:180° are forbidden because of backbone atom clashes).

For tetrahedral carbons, the substituents are typically found in staggered conformations (see figure, above). Peptide bonds are more complicated, because while the α-carbon is tetrahedral, the two other backbone atom types are not. However, the same principle applies: the preferred conformations for peptide bond atoms have the substituent atoms at maximal distances from one another.

A Ψ angle of 180° results in an alignment of the Namide with the carbonyl oxygen from the same residue. This is allowed, although not especially favored. A Ψ angle of 0° places the Namide from one residue very close to the Namide from the previous residue; this results in a steric clash (as well as an unfavorable electrostatic interaction, because both Namide have partial positive charges). The residue side-chains also impose steric constraints. Glycine, because of its small side chain, has a much large ranger of possible Φ / Ψ pairs than any other residue. Proline has a very limited range of Φ angles because its side-chain is covalently bonded to its Namide. Most other residues are limited to relatively few Φ / Ψ pairs (although more than proline). This is especially true for the β-branched residues threonine, valine, and isoleucine, which are the most restricted, because these residues have more steric bulk due to the presence of two groups attached their β-carbon. Allowed values for Φ and Ψ are graphically revealed when Ψ is plotted versus Φ

in a **Ramachandran plot**, introduced by G. N. Ramachandran.

The Ramachandran Plot

In a polypeptide the main chain N-Calpha and Calpha-C bonds relatively are free to rotate. These rotations are represented by the torsion angles phi and psi, respectively.

G N Ramachandran used computer models of small polypeptides to systematically vary phi and psi with the objective of finding stable conformations. For each conformation, the structure was examined for close contacts between atoms. Atoms were treated as hard spheres with dimensions corresponding to their van der Waals radii. Therefore, phi and psi angles which cause spheres to collide correspond to sterically disallowed conformations of the polypeptide backbone.

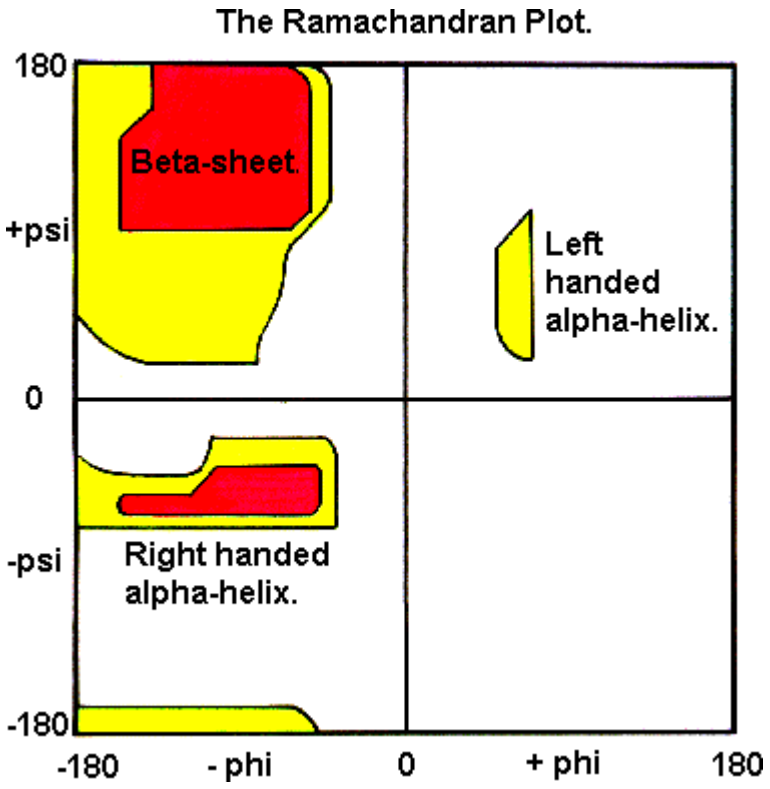


Figure 6 Ramachandran Plot

In the diagram above the white areas correspond to conformations where atoms in the polypeptide come closer than the sum of their van der Waals radii. These regions are sterically disallowed for all amino acids except glycine which is unique in that it lacks a

side chain. The red regions correspond to conformations where there are no steric clashes, ie these are the allowed regions namely the alpha-helical and beta-sheet conformations. The yellow areas show the allowed regions if slightly shorter van der Waals radi are used in the calculation, ie the atoms are allowed to come a little closer together. This brings out an additional region which corresponds to the left-handed alpha-helix.

L-amino acids cannot form extended regions of left-handed helix but occasionally individual residues adopt this conformation. These residues are usually glycine but can also be asparagine or aspartate where the side chain forms a hydrogen bond with the main chain and therefore stabilizes this otherwise unfavorable conformation. The 3(10) helix occurs close to the upper right of the alpha-helical region and is on the edge of allowed region indicating lower stability.

Disallowed regions generally involve steric hindrance between the side chain C-beta methylene group and main chain atoms. Glycine has no side chain and therefore can adopt phi and psi angles in all four quadrants of the Ramachandran plot. Hence it frequently occurs in turn regions of proteins where any other residue would be sterically hindered.

Secondary structure

The term secondary structure refers to the local conformation of some part of a polypeptide. The discussion of secondary structure most usefully focuses on common regular folding patterns of the polypeptide backbone. A few types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the α-helix and β-sheet. Using fundamental chemical principles and a few experimental observations, Pauling and Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated.

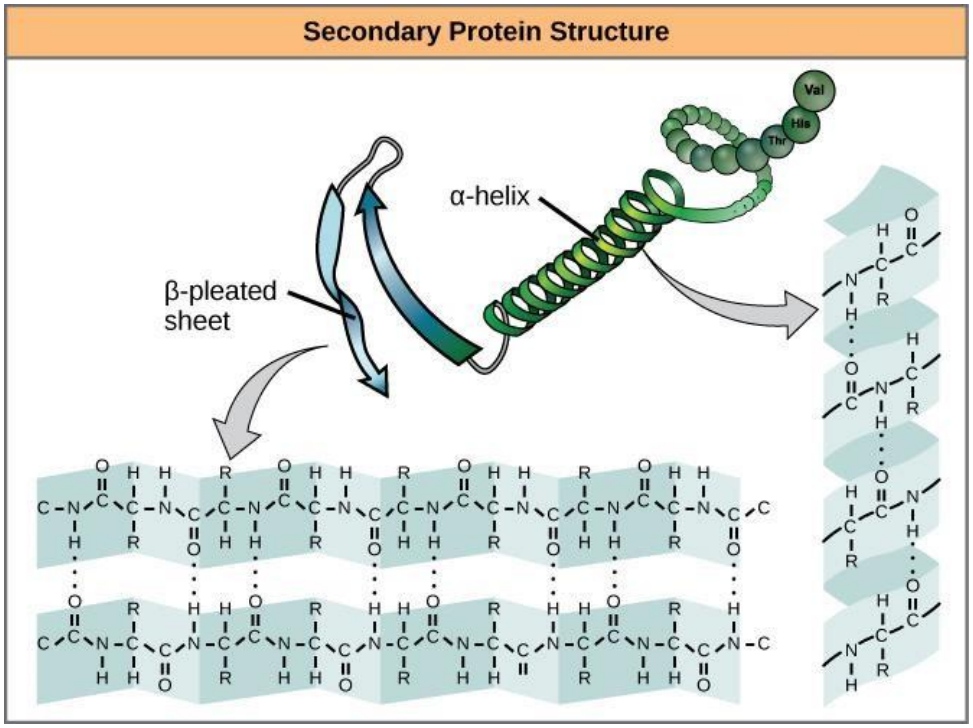


Figure 7 Secondary protein structures

Alpha helix (**α-helix**)

The **alpha helix** (**α-helix**) is a common secondary structure of proteins and is a right hand-coiled or spiral conformation (helix) in which every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid four residues earlier ($i + 4 \rightarrow i$ hydrogen bonding). This secondary structure is also sometimes called a classic **Pauling–Corey–Branson alpha helix** (see below). The name **3.613-helix** is also used for this type of helix, denoting the number of residues per helical turn, and 13 atoms being involved in the ring formed by the hydrogen bond. Among types of local structure in proteins, the α- helix is the most regular and the most predictable from sequence, as well as the most prevalent.

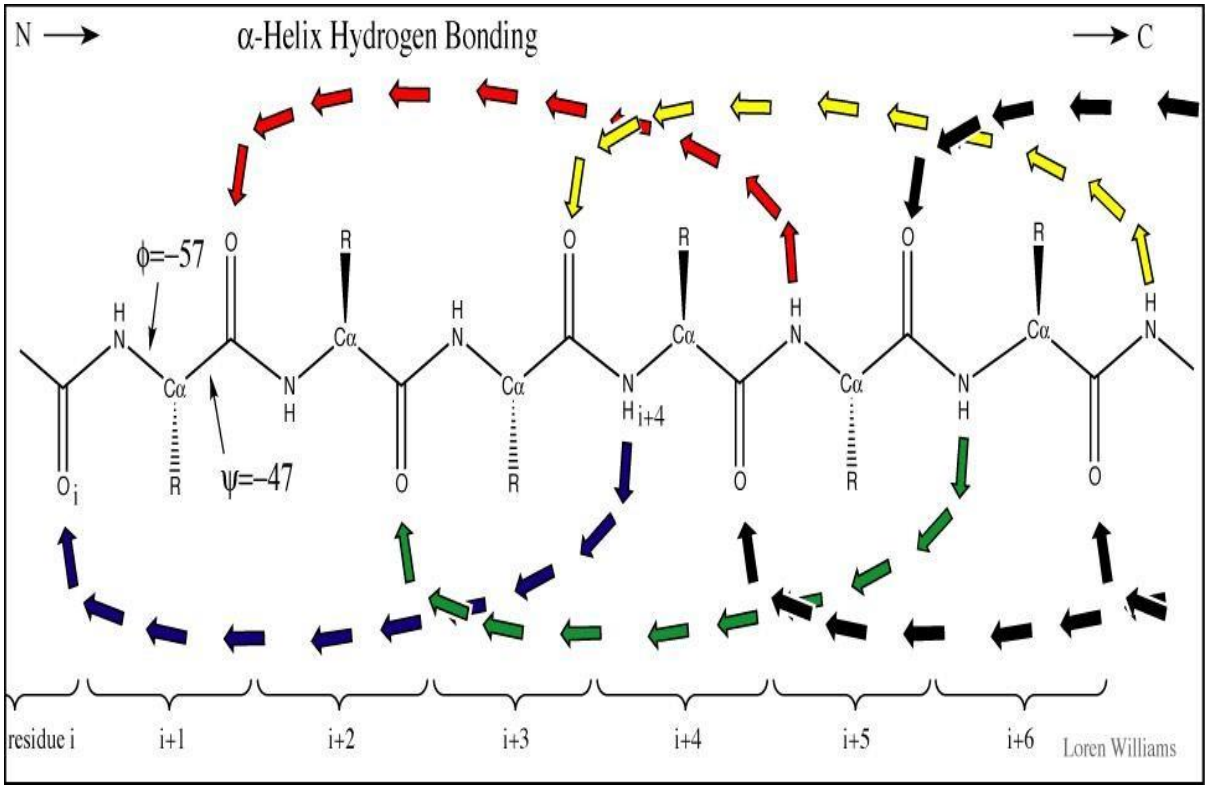


Figure 8 α Helix H-Bonding

PROPERTIES

The amino acids in an α helix are arranged in a right-handed helical structure where each amino acid residue corresponds to a 100° turn in the helix (i.e., the helix has 3.6 residues per turn), and a translation of 1.5 Å (0.15 nm) along the helical axis.

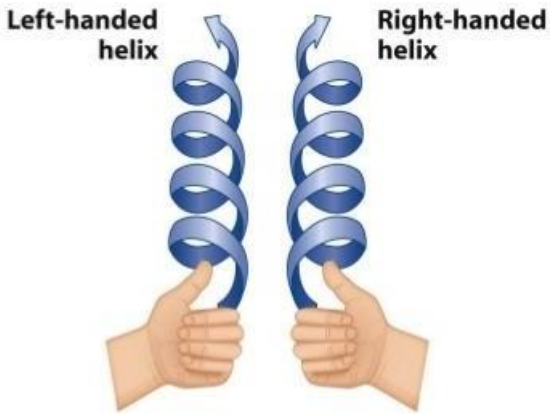


Figure 9 α Helix – Left handed & Right handed

Short pieces of left-handed helix sometimes occur with a large content of achiral glycine amino acids, but are unfavorable for the other normal, biological L-amino acids.

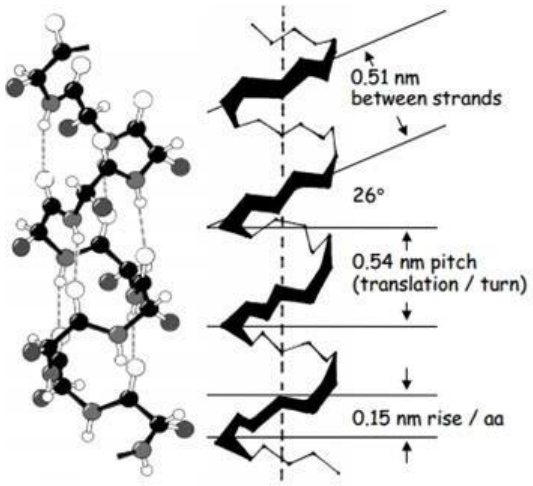


Figure 10 α Helix Pitch

The pitch of the alpha-helix (the vertical distance between consecutive turns of the helix) is 5.4 Å (0.54 nm), which is the product of 1.5 and 3.6. What is most important is that the N- H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid *four* residues earlier; this repeated $i + 4 \rightarrow i$ hydrogen bonding is the most prominent characteristic of an α-helix.

Similar structures include the 3_{10} helix ($i + 3 \rightarrow i$ hydrogen bonding) and the π- helix (i+5→i hydrogen bonding). The α helix can be described as a 3.6_{13} helix, since the i + 4 spacing adds 3 more atoms to the H-bonded loop compared to the tighter 3_{10} helix, and on average, 3.6 amino acids are involved in one ring of α helix. The subscripts refer to the number of atoms (including the hydrogen) in the closed loop formed by the hydrogen bond.

Residues in α-helices typically adopt backbone (φ, ψ) dihedral angles around (-60°, -45°), as shown in the image at right. In more general terms, they adopt dihedral angles such that the ψ dihedral angle of one residue and the φ dihedral angle of the *next* residue sum to roughly - 105°. As a consequence, α-helical dihedral angles, in general, fall on a diagonal stripe on the Ramachandran diagram (of slope -1), ranging from (-90°, -15°) to (-35°, -

70°). For comparison, the sum of the dihedral angles for a 310 helix is roughly -75°, whereas that for the π-helix is roughly -130°.

Table 1- Structural features of the three major forms of protein helices

Geometry attribute	**α-helix**	**310 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)

An α-helix has a dipole, with the partial positive charge toward N-terminus. This is true because all of the partial charges of the peptide bonds are in alignment. The backbone of the helix is ~6 Å in diameter (ignoring side chains).

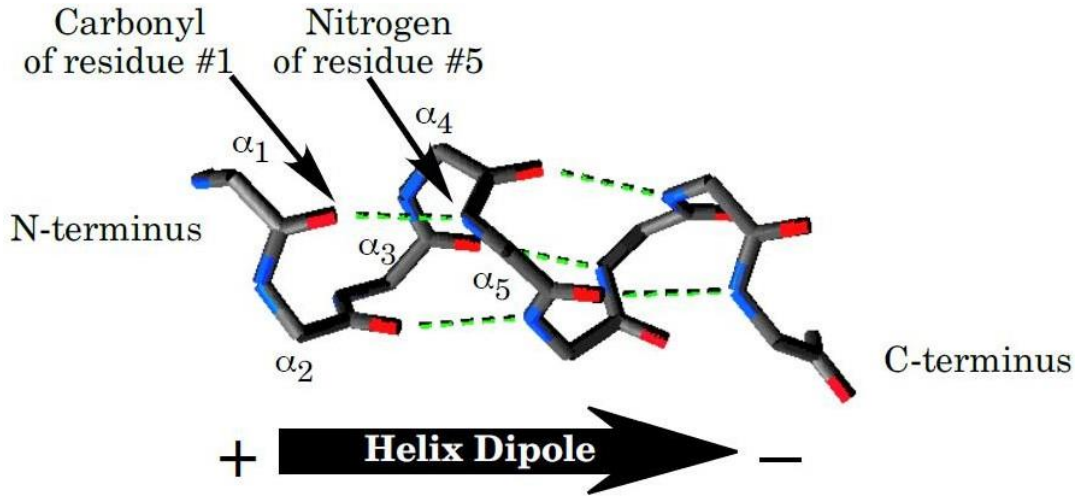


Figure 11 Two-dimensional representations of α-helices

Drawing a three-dimensional helix on paper is difficult. Two types of two dimensional representations (helical wheel and helical net diagrams) are commonly used to simplify

the analysis of helical segments of proteins. The two-dimensional representations are somewhat stylized, but show the major features more clearly than attempting to draw a three- dimensional structure accurately in two dimensions.

The first type of representation is a Helical Wheel diagram. In this diagram, the representation involves looking down the helix axis, and plotting the rotational angle around the helix for each residue. This representation is conceptually easily grasped, but tends to obscure the distance along the helix; residues 0 and 18 are exactly aligned on this diagram, but are actually separated in space by 27 Å.

Helical Wheel

Residue #0 = 0° (by definition)

#1 = 100°

#2 = 200°

#3 = 300°

#4 = 400° = 40°

#5 = 140°

#6 = 240°

#7 = 340°

#8 = 440° = 80°

9# = 900° (from first) = 180°

These angles can be plotted on a circle.

Doing so results in a representation that corresponds to the view looking down the long axis of the helix. (**Note that the rotation is clockwise as the residue number increases**)

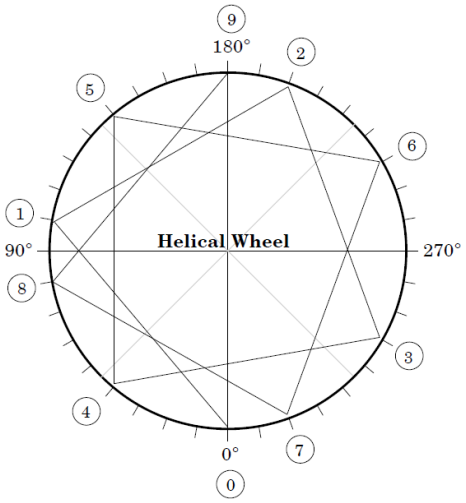


Figure 12 Helical Wheel

Note that residues 0, 3, 4, 7, and 8 are all located on one face of the helix. A helix that has its axis along the border of this region would be expected to have a corresponding amphipathic distribution of polar and non-polar residues. (Amphipathic, meaning "hating both" refers to the presence of both polar and non-polar groups in the helix.)

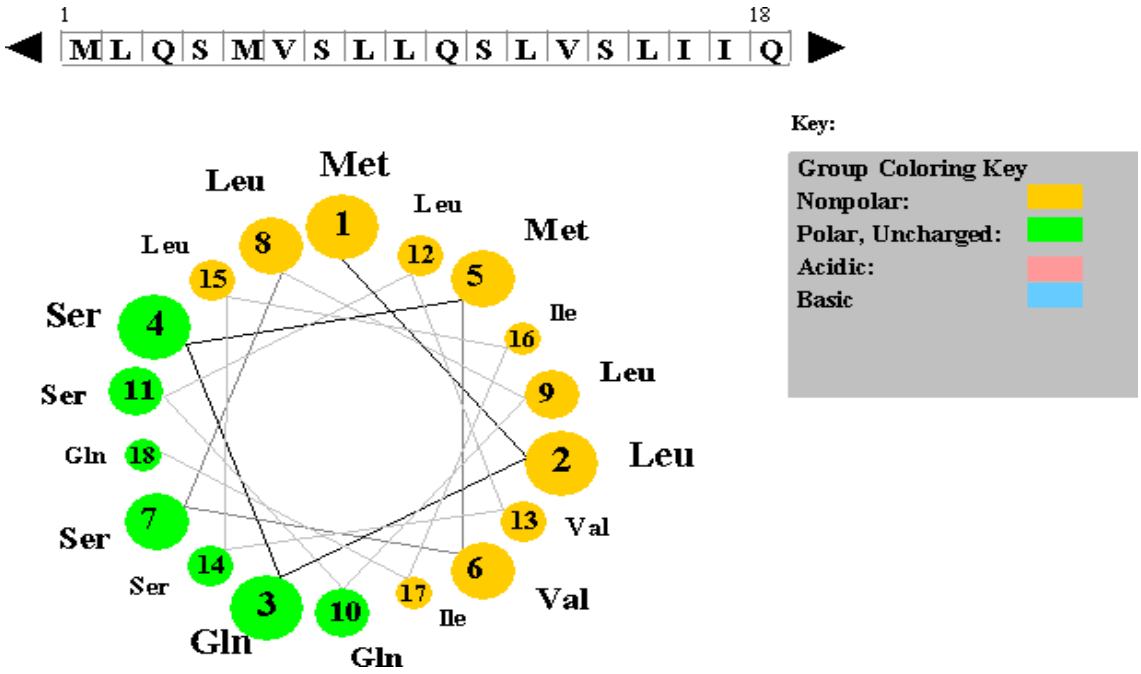


Figure 13 Helical Wheel –Amino acids

The ββ Conformation Organizes Polypeptide Chains into Sheets

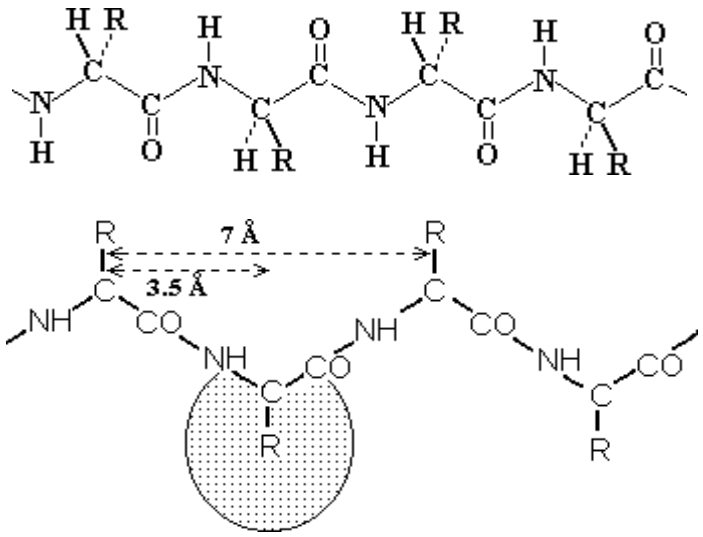


Figure 14 β conformation -an extended state

Pauling and Corey predicted a second type of repetitive structure, the β conformation -an

extended state for which angles phi = -135^{o} and psi = +135^{o}; the polypeptide chain **alternates** in direction, resulting in a zig-zag structure for the peptide chain. Note the shaded circle around R; the extended strand arrangement also allows the **maximum space and freedom of movement for a side chain**. The repeat between identically oriented R-groups is 7.0 Å, with 3.5 Å per amino acid, matching the fiber diffraction data for beta-keratins.

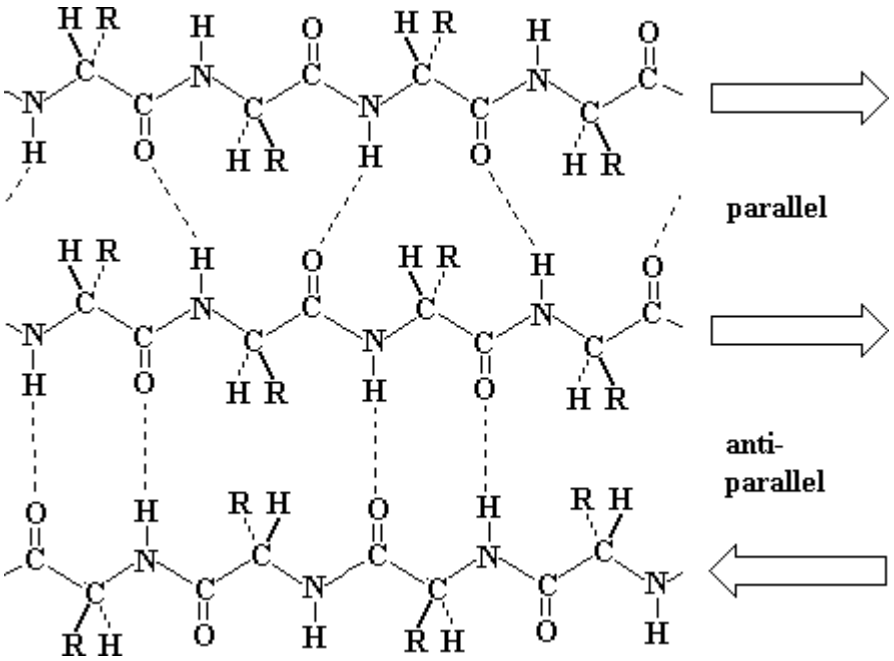


Figure 15 Parallel and Antiparallel beta sheets

Pauling's extended state model matched the spacing of fibroin exactly (3.5 and 7.0 Å). In the extended state, H-bonding NH and CO groups point out at 90^{o} to the strand. If extended strands are lined up side by side, H-bonds bridge from strand to strand. Identical or opposed strand alignments make up parallel or antiparallel beta sheets (named for beta keratin). Antiparallel beta-sheet is significantly more stable due to the well aligned H-bonds.

Table 2 – Dihedral angles in beta sheets

Angle	Antiparallel	Parallel
Φ	–139°	–119°
Ψ	135°	113°

Amino acid preferences for different secondary structure

Alpha helix may be considered the default state for secondary structure. Although the potential energy is not as low as for beta sheet, H-bond formation is intra- strand, so there is an entropic advantage over beta sheet, where H-bonds must form from strand to strand, with strand segments that may be quite distant in the polypeptide sequence.

The main criterion for alpha helix preference is that the amino acid side chain should **cover and protect the backbone H-bonds** in the core of the helix. Most amino acids do this with some key exceptions:

alpha-helix preference: **Ala,Leu,Met,Phe,Glu,Gln,His,Lys,Arg**

The extended structure leaves the **maximum space free** for the amino acid side chains: as a result, those amino acids with **large bulky side chains prefer to form beta sheet structures**:

just plain large: **Tyr, Trp, (Phe, Met)**

bulky and awkward due to branched beta carbon: **Ile, Val, Thr**

large S atom on beta carbon: **Cys**

The remaining amino acids have side chains which **disrupt secondary structure**, and are known as **secondary structure breakers**:

side chain H is too small to protect backbone H-bond: **Gly**

side chain linked to alpha N, **has no N-H** to H-bond; **Pro**

rigid structure due to ring restricts to phi = -60^{o};

H-bonding side chains compete directly with backbone H-bonds **Asp, Asn, Ser**

Clusters of breakers give rise to regions known as **loops or turns** which mark the boundaries of regular secondary structure, and serve to link up secondary structure segments.

β-turn

Turns are the third of the three "classical" secondary structures that serve to reverse the direction of the polypeptide chain.

They are located primarily on the protein surface and accordingly contain polar and charged residues.

Turns were first recognised from a theoretical conformational analysis by Venkatachalam (1968). He considered what conformations were available to a system of three linked peptide units (or four successive residues) that could be stabilised by a backbone hydrogen bond between the CO of residue n and the NH of residue n+3. He found three general types, one of which (type III) actually has repeating , values of -60deg, -30deg and is identical with the 310- helix. The three types each contain a hydrogen bond between the carbonyl oxygen of residue i and the amide nitrogen of i+3. These three types of turns are designated I, II, and III. Many have speculated on the role of this type of secondary structure in globular proteins.

Turns may be viewed as a weak link in the polypeptide chain, allowing the other secondary structures (helix and sheet) to determine the conformational outcome. In contrast (based on the recent experimental finding of "turn-like" structures in short peptides in aqueous solutions, turns are considered to be structure-nucleating segments, formed early in the folding process.

Type I turns occur 2-3 times more frequently than type II. There are position dependent amino acid preferences for residues in turn conformations.

Type I can tolerate all residues in position i to i+3 with the exception of Pro at position i+2. Proline is favoured at position i+1 and Gly is favoured at i+3 in type I and type II turns. The polar sidechains of Asn, Asp, Ser, and Cys often populate position i where they can hydrogen bond to the backbone NH of residue i+2.

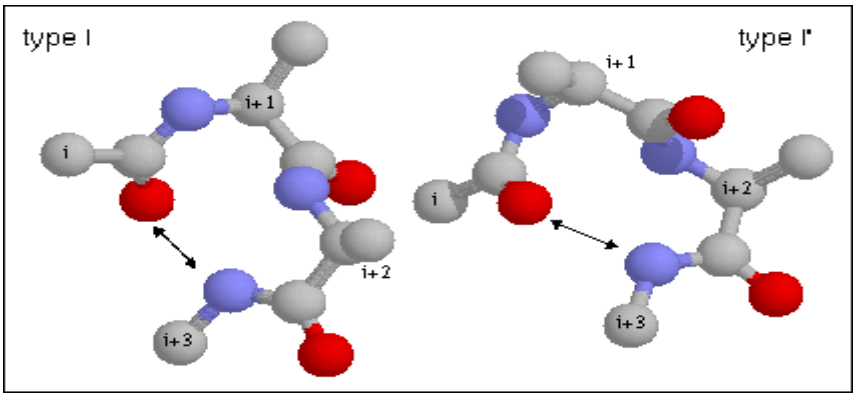


Figure 16 β-turns

Other secondary structures

Random coil

Most proteins have regions in which the Φ and Ψ angles are not repeating. These regions are sometimes referred to as "random coil" although their structures are not actually "random". The non-repeating structures may be considered "secondary structure", in spite of their irregular nature.

Fibrous Proteins Are Adapted for a Structural Function

α-Keratin, collagen, and elastin provide clear examples of the relationship between protein structure and biological function

Table 3 – Secondary structures & Properties of fibrous proteins

Structure	Characteristics	Examples of occurrence
α Helix, cross-linked by disulfide bonds	Tough, insoluble protective structures of varying hardness and flexibility	α-Keratin of hair, feathers, nails
β Conformation	Soft, flexible filaments	Silk fibroin
Collagen triple helix	High tensile strength, without stretch	Collagen of tendons, bone matrix

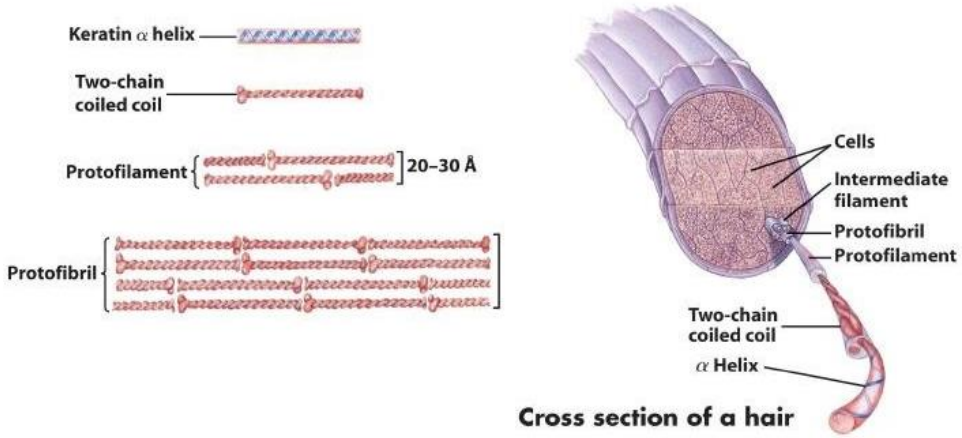
These proteins share properties that give strength and/or elasticity to structures in which they occur. They have relatively simple structures, and all are insoluble in water, a property conferred by a high concentration of hydrophobic amino acids both in the interior of the protein and on the surface. These proteins represent an exception to the rule that hydrophobic groups must be buried. The hydrophobic core of the molecule therefore contributes less to structural stability, and covalent bonds assume an especially important role.

α-Keratin and **collagen** have evolved for strength. In vertebrates, α-keratins constitute almost the entire dry weight of hair, wool, feathers, nails, claws, quills, scales, horns, hooves, tortoise shell, and much of the outer layer of skin. Collagen is found in connective tissue such as tendons, cartilage, the organic matrix of bones, and the cornea of the eye. The polypeptide chains of both proteins have simple helical structures. The α-keratin helix is the right-handed α helix found in many other proteins. However, the collagen helix is unique. It is left-handed and has three amino acid residues per turn. In both α-keratin and collagen, a few amino acids predominate. α-Keratin is rich in the hydrophobic residues Phe, Ile, Val, Met, and Ala. Collagen is 35% Gly, 11% Ala, and 21% Pro and Hyp (hydroxyproline). The

unusual amino acid content of collagen is imposed by structural constraints unique to the collagen helix. The amino acid sequence in collagen is generally a repeating tripeptide unit, Gly-X-Pro or Gly-X-Hyp, where X can be any amino acid. The food product gelatin is derived from collagen. Although it is protein, it has little nutritional value because collagen lacks significant amounts of many amino acids that are essential in the human diet.

In both α-keratin and collagen, strength is amplified by wrapping multiple helical strands together in a superhelix, much the way strings are twisted to make a strong rope. In both proteins the helical path of the supertwists is opposite in sense to the twisting of the individual polypeptide helices, a conformation that permits the closest possible packing of the multiple polypeptide chains.

The superhelical twisting is probably left-handed in α-keratin and right-handed in collagen. The tight wrapping of the collagen triple helix provides great tensile strength with no capacity to stretch: Collagen fibers can support up to 10,000 times their own weight and are said to have greater tensile strength than a steel wire of equal cross section.



-Alpha keratins belong to the intermediate filament (IF) protein family.
-An all α-helix protein.
-Rich in hydrophobic amino acids: Ala, Val, Leu, Ile, Met, Phe

Figure 17 α-Keratin

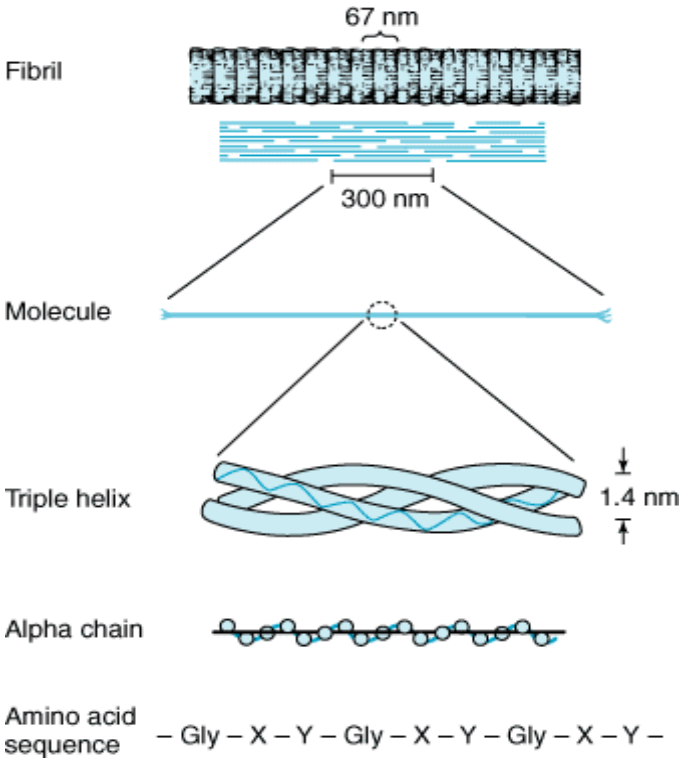


Figure 18 Collagen

The fibroin protein consists of layers of antiparallel beta sheets. Its primary structure mainly consists of the recurrent amino acid sequence (Gly-Ser-Gly-Ala-Gly-Ala)*n*. The high glycine (and, to a lesser extent, alanine) content allows for tight packing of the sheets, which contributes to silk's rigid structure and tensile strength. A combination of stiffness and toughness make it a material with applications in several areas, including biomedicine and textile manufacture.

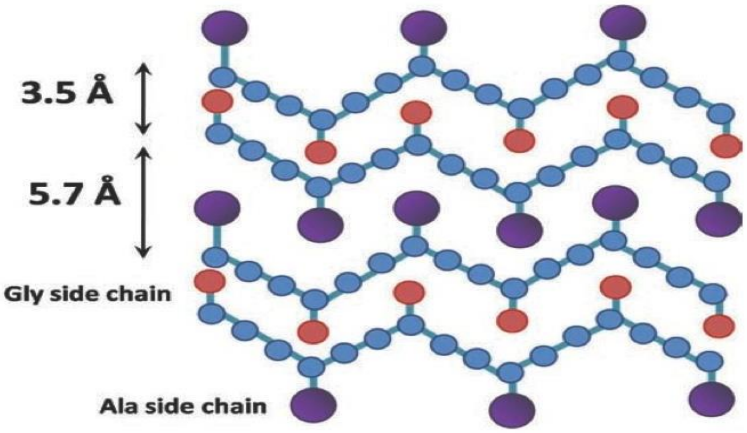


Figure 19 Silk Fibroin

The most characteristic features of a β sheet are the number of strands, their relative directions (parallel or antiparallel), and how the strands are connected. This information can be represented by topology diagrams. They are useful to compare β structures.

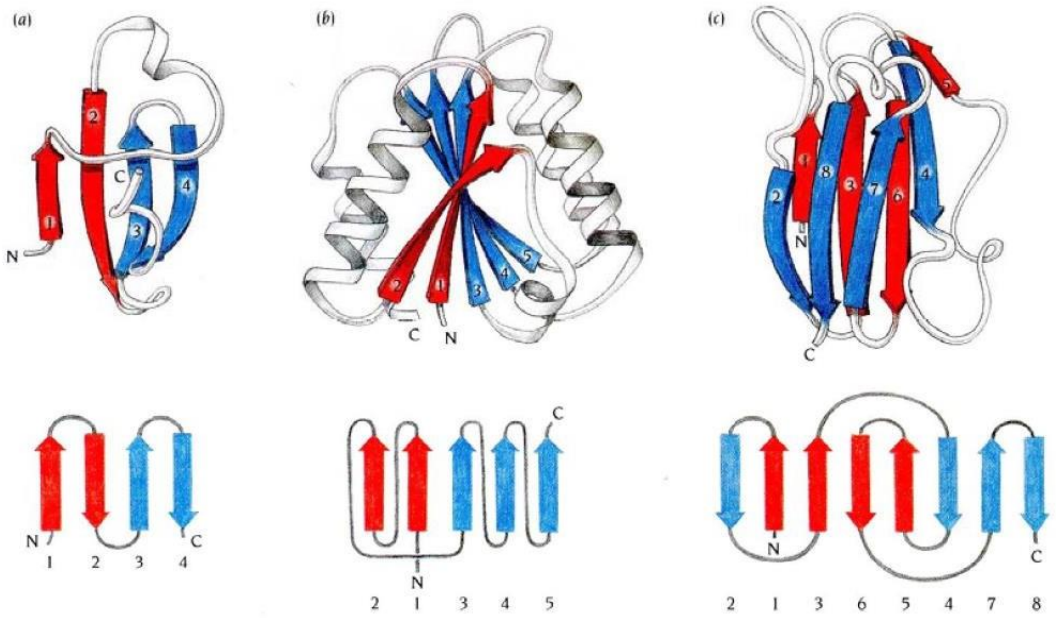


Figure 20 Topology diagrams

Protein Tertiary Structure

Tertiary structure refers to the three-dimensional arrangement of all atoms in a protein. Tertiary structure is formed by the folding in three dimensions of the secondary structure elements of a protein. While the α helical secondary structure is held together by interactions between the carbonyl and amide groups within the backbone, tertiary structure is held together by interactions between R-groups of residues brought together by folding. Disulfide bonds are also counted under the category of tertiary structure interactions. Proteins that are compact are known as globular proteins.

Examination of protein structures resolved by X-ray diffraction and NMR has revealed a variety of folding patterns common to many different proteins. However, even within these folds, distinct substructures or structural **motifs**, i.e. distinctive arrangements of elements of secondary structure, have been described. The term **supersecondary structure** has been coined to describe this level of organisation, which is intermediate between secondary and tertiary.

Motifs or folds, are particularly stable arrangements of several elements of the secondary structure. Supersecondary structures are usually produced by packing side chains from adjacent secondary structural elements close to each other.

Rules for secondary structure

• Hydrophobic side groups must be buried inside the folds, therefore, layers must be created (β−α−β; α− α).

• α-helix and β-sheet, if occur together, are found in different structural layers.

• Adjacent polypeptide segments are stacked together.

• Connections between secondary structures do not form knots.

• The β-sheet is the most stable.

Motif

• Secondary structure composition, e.g. all α, all β, segregated α+β, mixed α/β

• Motif = small, specific combinations of secondary structure elements, e.g. β-α-β loop

Helix super secondary structures

Helix-Turn-Helix Motif

Also called the alpha-alpha type (αα-type). The motif is comprised of two antiparallel helices connected by a turn. The helix-turn-helix is a functional motif and is usually identified in proteins that bind to DNA minor and major grooves, and Calcium-binding proteins.

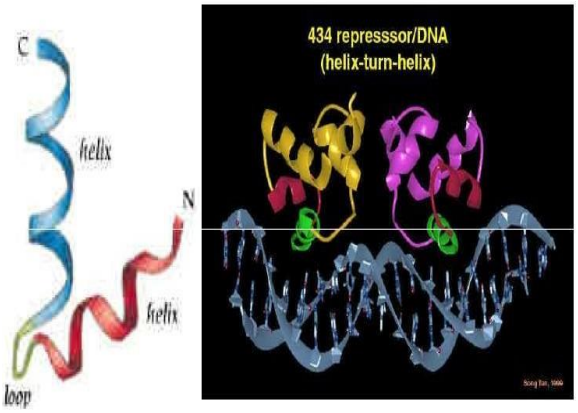


Figure 21 DNA binding Helix-turn-Helix motif

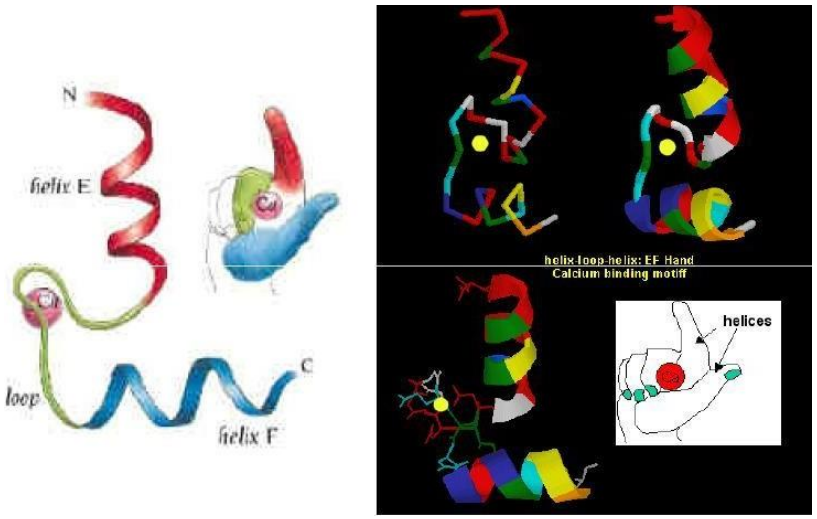


Figure 22 Calcium binding (EF Hand- Calcium binding) motif

Helix-hairpin-helix: Involved in DNA binding

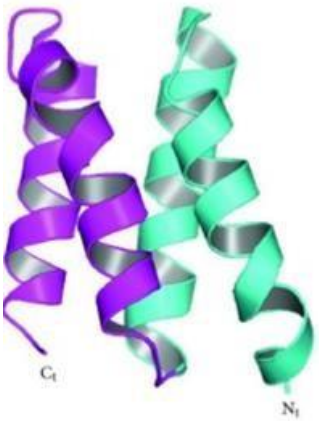


Figure 23 Helix-hairpin-helix

Alpha-alpha corner

Short loop regions connecting helices which are roughly perpendicular to one another

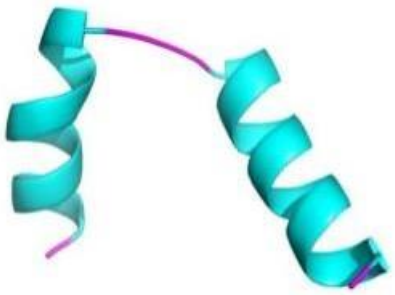


Figure 24 Alpha-alpha corner

Sheet super secondary structures

All beta tertiary structural domains can occur in proteins with one domain (eg. concanavalin A, superoxide dismutase), and occurs at least once in proteins with two domains (eg. chymotrypsin), or three domains (eg. OmpF).

The beta strands making up these domains are all essentially antiparallel and form structures to achieve stable packing arrangements within the protein.

There are presently (as of version 1.39) about 70 subclasses listed in SCOP for this domain, and some examples of these are outlined below.

Beta barrels

This is the most abundant beta-domain structure and as the name suggests the domain forms a 'barrel-like' structure. The beta barrels are not geometrically perfect and can be rather distorted.

There are three main types:

1. Up-and-down barrels
2. Greek key barrels
3. Jelly roll (Swiss roll) barrels

Up-and-down beta-sheets or beta-barrels

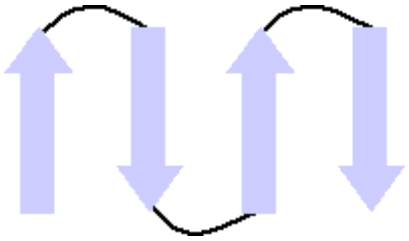


Figure 25 Beta barrels

The simple topology of an up-and-down barrel (named because the beta strands follow each other in sequence in an up-and-down fashion). Usually, the loops joining the beta strands do not crossover the 'ends' of the barrel.

Greek key barrels

These are barrels formed from two, or more, Greek Key motifs. It is a stable structure. The Greek key barrel consists of four anti-parallel Beta strands where one strand changes the topology direction. Hydrogen bonding occurs between strands 1:4, and strands 2:3. Strand 2 then folds over to form the structural motif.

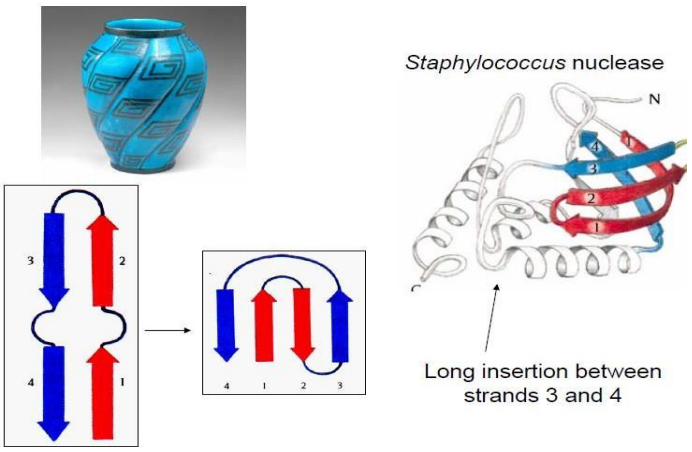


Figure 26 Greek key barrels

Jelly roll barrels

These barrels are formed from a 'Greek Key-like' structure called a jelly roll. Supposedly named because the polypeptide chain is wrapped around a barrel core like a jelly roll (swiss roll). It is a stable structure. This structure is found in coat proteins of spherical viruses, plant lectin concanavalin A, and hemagglutinin protein from influenza virus.

The essential features of a jelly roll barrel are that:

- it is like an inverted 'U' (which is often seen twisted and distorted in proteins)
- it is usually divided into two beta sheets which are packed against each other
- most jelly roll barrels have eight strands although any even number greater than 8 can form a jelly roll barrel

- it folds such that hydrogen bonds exist between strands 1 and 8; 2 and 7; 3 and 6; and 4 and 5

Figure 27 Jelly roll barrels

Beta sandwich

A beta sandwich is essentially a 'flattened' beta barrel with the two sheets packing closely together (like a sandwich!). The first and last strands of the sandwich do not hydrogen bond to each other to complete a 'barrel' structure.

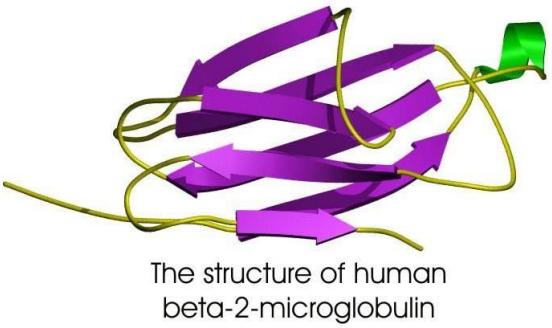


Figure 28 Beta sandwich in beta 2 microglobulin

Aligned or Orthogonal beta strands

Beta strands in barrels or sandwich structures can be orientated in two general ways:

- where the strands in two sheets are almost aligned, and in the same orientation,

to each other and form an 'aligned beta' structure (eg. gamma crystallin)

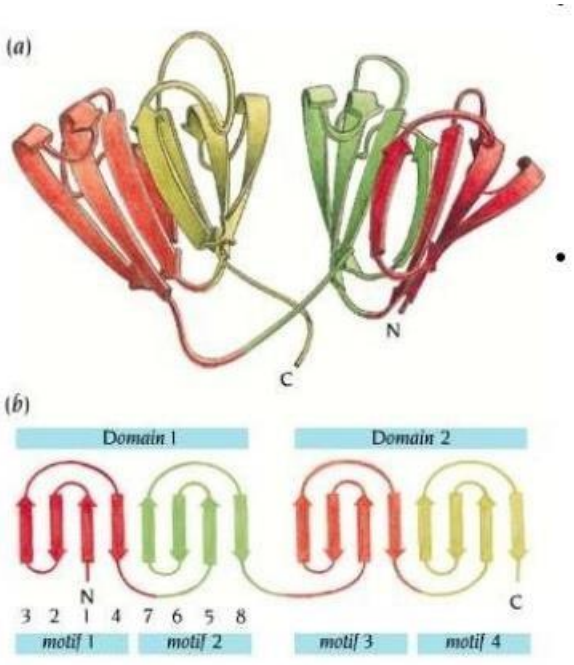


Figure 29 Aligned beta strands

- where the strands, in at least two sheets, are roughly perpendicular to each other and form an 'orthogonal beta' structure.

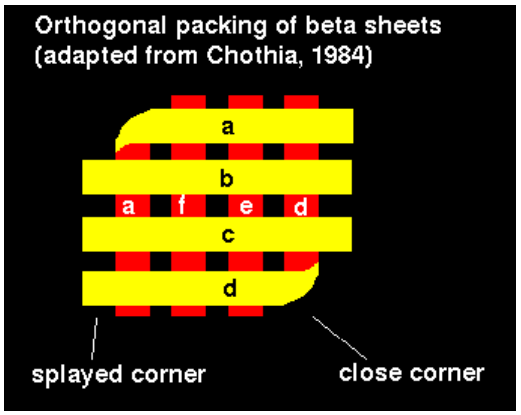


Figure 30 Orthogonal beta sheets

Beta-hairpin: two antiparallel beta strands connected by a “hairpin” bend, i.e. beta-turn 2 x antiparallel beta-strands + beta-turn = beta hairpin

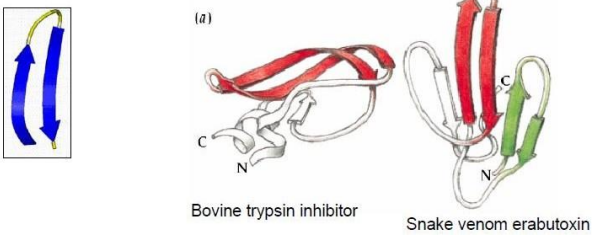


Figure 31 Beta hairpin

Beta-beta corner

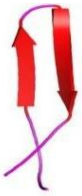


Figure 32 Beta-beta corner

- Two antiparallel beta strands which form a beta hairpin can change direction abruptly. The angle of the change of direction is about 90 degrees and so the structure is known as a 'beta corner'
- The abrupt angle change is achieved by one strand having a glycine residue (so there is no steric hindrance from a side chain) and the other strand having a beta bulge (where the hydrogen bond is broken).
- no known function

α/β Topologies

Beta-Helix-Beta Motif

An important and widespread supersecondary structural motif in proteins is known as the β- α-β motif (Beta-Alpha-Beta motif). The motif consists of two parallel Beta strands that is connected via an alpha helix (with two turns). The motif is found in most proteins that contain parallel beta strands, and the axis of the Helix and the Strands are roughly parallel to each other with all three elements forming a hydrophobic core due to shielding. The β-α-β motif may be structurally or functionally involved. The Loop that connects the C-

terminal of first Beta strand and N-terminal of Helix is frequently involved in ligand binding functions, and the motif itself is frequently found in ion channels.

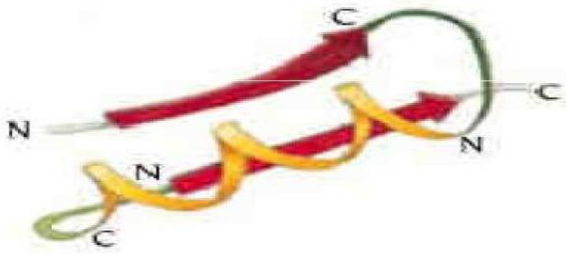


Figure 33 Beta-Helix-Beta Motif

The β - α - β - α - β subunit, often present in nucleotide-binding proteins, is named the **Rossman Fold**, after Michael Rossman

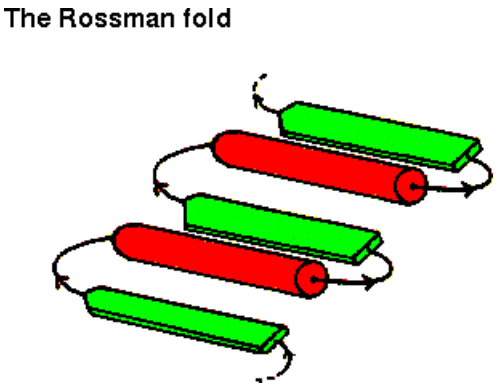


Figure 34 Rossman fold

α/β horse shoe

17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

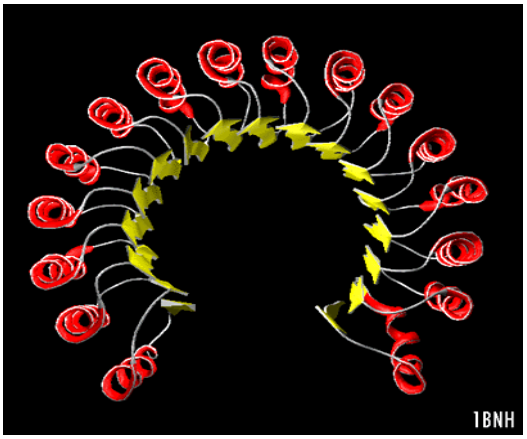


Figure 35 α/β horse shoe

α/β barrels

Consider a sequence of eight α/β motifs:

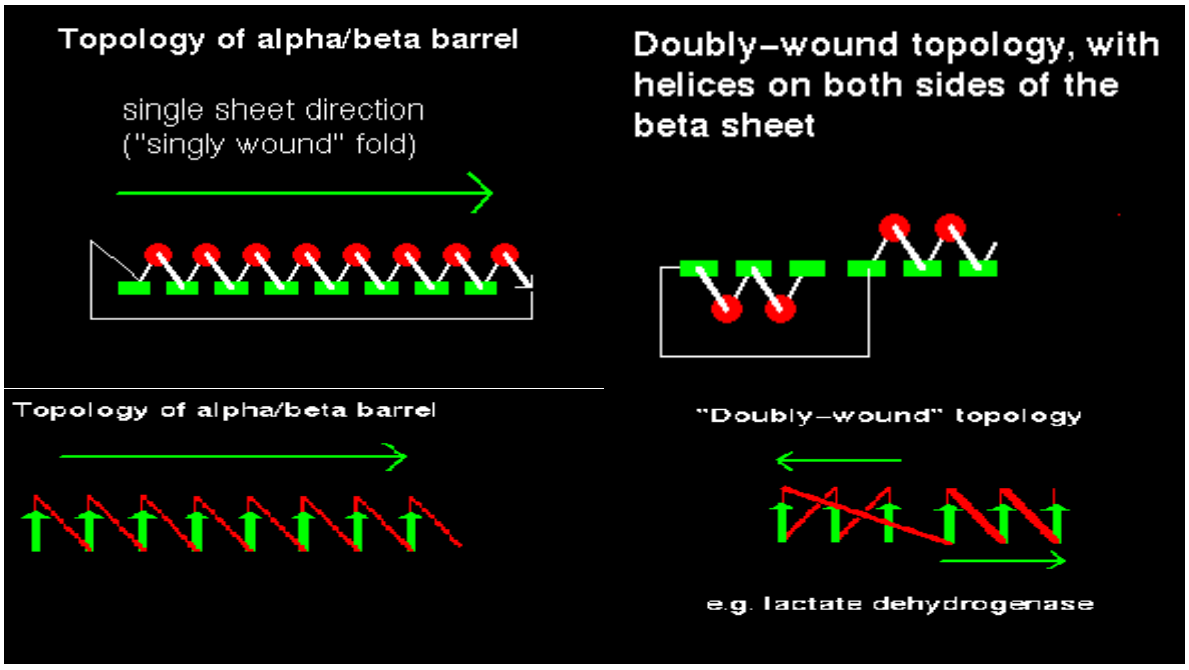


Figure 36: Topology of α/β barrel

If the first strand hydrogen bonds to the last, then the structure closes on itself forming a barrel-like structure. This is shown in the picture of triose phosphate isomerase.

Note that the "staves" of the barrel are slanted, due to the twist of the b sheet. Also notice that there are effectively four layers to this structure. The direction of the sheet does not change (it is anticlockwise in the diagram). Such a structure may therefore be described as **singly wound**.

In a structure which is open rather than closed like the barrel, helices would be situated on only one side of the b sheet if the sheet direction did not reverse. Therefore open a/b structures must be **doubly wound** to cover both sides of the sheet.

The chain starts in the middle of the sheet and travels outwards, then returns to the centre via a loop and travels outwards to the opposite edge:

Doubly-wound topologies where the sheet begins at the edge and works inwards are rarely observed.

Alpha+Beta Topologies

This is where we collect together all those folds which include significant alpha and beta secondary structural elements, but for which those elements are `**mixed**', in the sense that they do NOT exhibit the wound alpha-beta topology. This class of folds is therefore referred to as **α+ β**

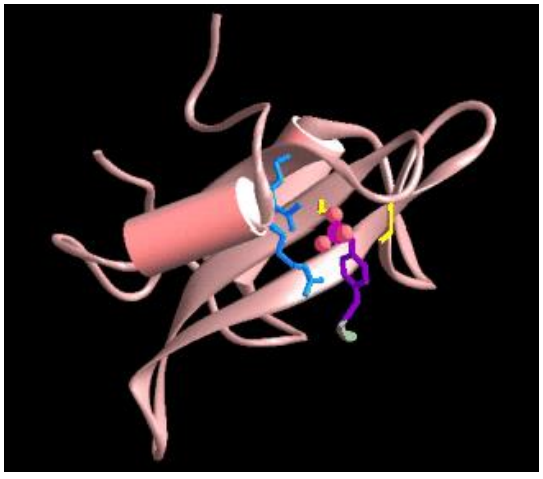


Figure 37 Alpha+Beta Topology

Domains

Domains are stable, independently folded, globular units, often consisting of combinations of motifs vary from 25 to 300 amino acids, average length – 100.

large globular proteins may consist of several domains linked by stretches of polypeptide. Separate domain may have distinct functions (eg G3P dehydrogenase). In many cases binding site formed by cleft between 2 domains frequently correspond to exon in gene

- Some examples of domains:

1. involving α-helix 4-helix bundle globin fold

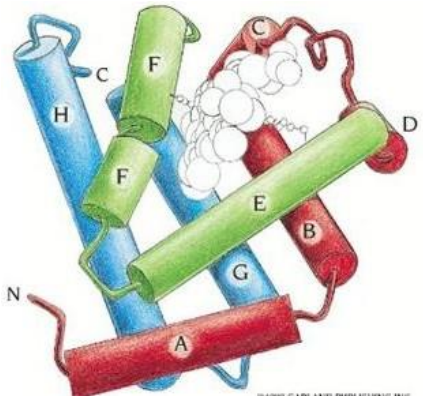


Figure 38 α-helix 4-helix bundle globin fold

The globin fold is found in its namesake globin protein families: hemoglobins and myoglobins, as well as in phycocyanins. Because myoglobin was the first protein whose structure was solved, the globin fold was thus the first protein fold discovered.

2. parallel β-sheets

hydrophobic residues on both sides, therefore must be buried.

□ □ barrel: 8 β strands each flanked by an antiparallel α-helix eg triose phosphate isomerase.)

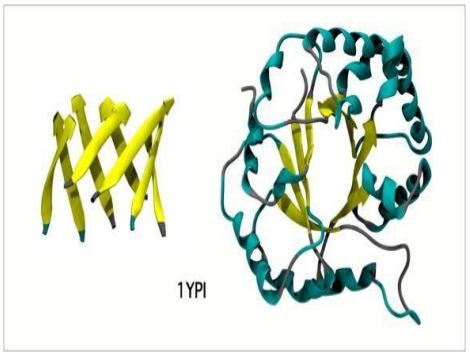


Figure 39 Parallel beta sheets

3. antiparallel β -sheet

Hydrophobic residues on one side, one side can be exposed to environment, minimum structure 2 layers

Sheets arranged in a barrel shape. More common than parallel β - barrels eg. immunoglobulin

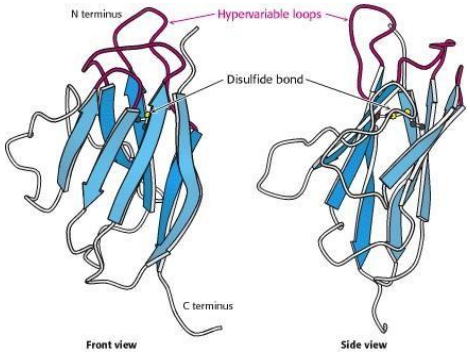


Figure 40 Antiparallel beta sheets

The **immunoglobulin domain** is a type of protein domain that consists of a 2-layer sandwich of 7-9 antiparallel β-strands arranged in two β-sheets with a Greek key topology, consisting of about 80 amino acids.

The backbone switches repeatedly between the two β-sheets. Typically, the pattern is (N-terminal β-hairpin in sheet 1)-(β-hairpin in sheet 2)-(β-strand in sheet 1)-(C-terminal β-hairpin in sheet 2). The cross-overs between sheets form an "X", so that the N- and C-terminal hairpins are facing each other.

Members of the immunoglobulin superfamily are found in hundreds of proteins of different functions. Examples include antibodies, the giant muscle kinase titin, andreceptor tyrosine kinases. Immunoglobulin-like domains may be involved in protein–protein and protein– ligand interactions.

Example of Tertiary Structure: Myoglobin and Hemoglobin

Myoglobin and hemoglobin are hemeproteins whose physiological importance is principally related to their ability to bind molecular oxygen.

Myoglobin

Single polypeptide chain (153 amino acids).

No disulfide bonds 8 right handed alpha helices form a hydrophobic pocket which contains heme molecule protective sheath for a heme group

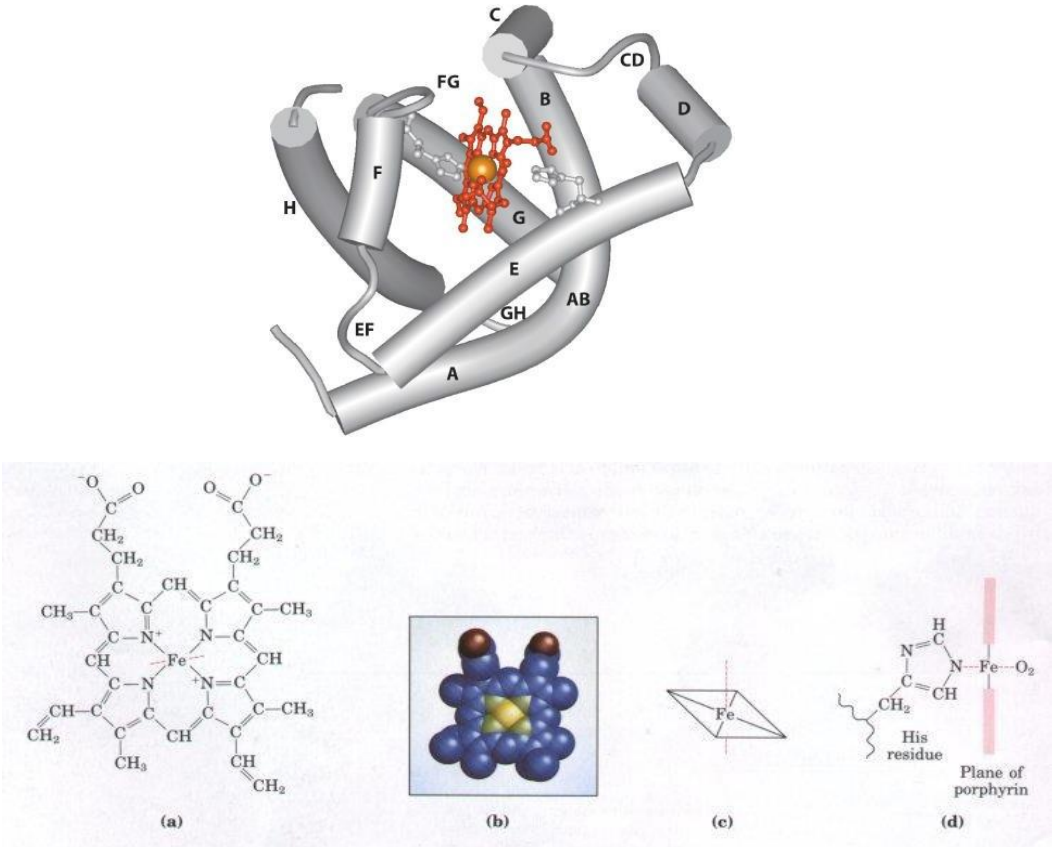


Figure 41 Myoglobin structure

Myoglobin is a monomeric heme protein found mainly in muscle tissue where it serves as an intracellular storage site for oxygen During periods of oxygen deprivation oxymyoglobin releases its bound oxygen which is then used for metabolic purposes The tertiary structure of myoglobin is that of a typical water soluble globular protein Its secondary structure is unusual in that it contains a very high proportion (75%) of α-helical secondary structure A myoglobin polypeptide is comprised of 8 separate right handed a-helices, designated A through H, that are connected by short non helical regions Amino acid R-groups packed into the interior of the molecule are predominantly hydrophobic in character while those exposed on the surface of the molecule are generally hydrophilic, thus making the molecule relatively water soluble.

Each myoglobin molecule contains one heme prosthetic group inserted into a hydrophobic cleft in the protein Each heme residue contains one central coordinately bound iron atom that is normally in the Fe 2+ , or ferrous, oxidation state The oxygen carried by hemeproteins is bound directly to the ferrous iron atom of the heme prosthetic group. The heme group is located in a crevice Except for one edge, non polar side chains surround the heme Fe 2+ is octahedrally coordinated Fe 2+ covalently bonded to the imidazole group of histidine 93 (F8) O 2 held on the other side by histidine 64 (E7)

Hydrophobic interactions between the tetrapyrrole ring and hydrophobic amino acid R groups on the interior of the cleft in the protein strongly stabilize the heme protein conjugate. In addition a nitrogen atom from a histidine R group located above the plane of the heme ring is coordinated with the iron atom further stabilizing the interaction between the heme and the protein. In oxymyoglobin the remaining bonding site on the iron atom (the 6th coordinate position) is occupied by the oxygen, whose binding is stabilized by a second histidine residue Carbon monoxide also binds coordinately to heme iron atoms in a manner similar to that of oxygen, but the binding of carbon monoxide to heme is much stronger than that of oxygen. The preferential binding of carbon monoxide to heme iron is largely responsible for the asphyxiation that results from carbon monoxide poisoning.

Hemoglobin

Oxygen transporter Four polypeptide chains Tetramer Each chain has a heme group Hence four O 2 can bind to each Hb Two alpha (141 amino acids) and two beta (146 amino acids) chains

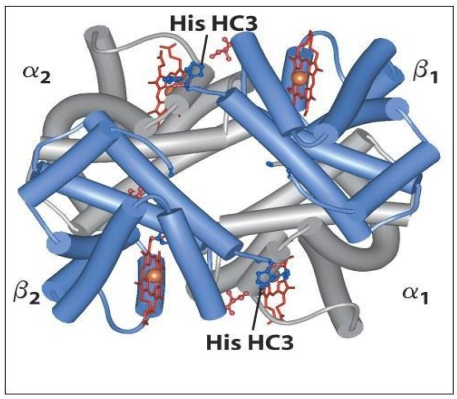


Figure 42 Hemoglobin structure

Hemoglobin is an [α(2):β(2)] tetrameric hemeprotein found in erythrocytes where it is responsible for binding oxygen in the lung and transporting the bound oxygen throughout the body where it is used in aerobic metabolic pathways Each subunit of a hemoglobin tetramer has a heme prosthetic group identical to that described for myoglobin. Although the secondary and tertiary structure of various hemoglobin subunits are similar, reflecting extensive homology in amino acid composition, the variations in amino acid composition that do exist impart marked differences in hemoglobin's oxygen carrying properties In addition, the quaternary structure of hemoglobin leads to physiologically important allosteric interactions between the subunits, a property lacking in monomeric myoglobin which is otherwise very similar to the α-subunit of hemoglobin

Quaternary structure

3-dimensional relationship of the different polypeptide chains (subunits) in a multimeric protein, the way the subunits fit together and their symmetry relationships.

• Only in proteins with more than one polypeptide chain; proteins with only one chain have no quaternary structure.

• Each polypeptide chain in a multichain protein = a subunit • 2-subunit protein = a dimer, 3 subunits = trimeric protein, 4 = tetrameric • homo(dimer or trimer etc.): identical subunits • hetero(dimer or trimer etc.): more than one kind of subunit (chains with different amino acid sequences) • different subunits designated with Greek letters – e.g., subunits of a heterodimeric protein = the "α subunit" and the "β subunit".

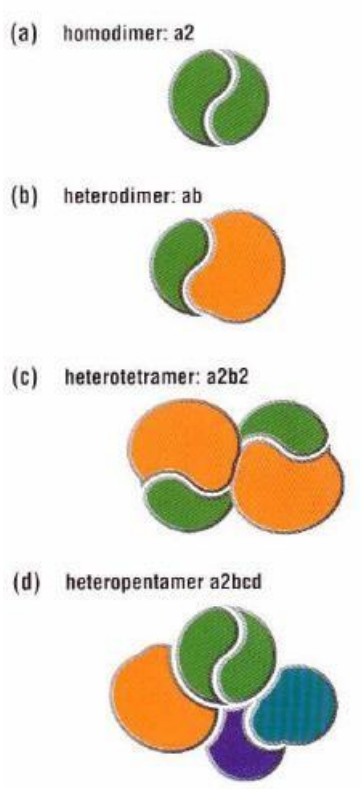


Figure 43 Protein subunits

– NOTE: This use of the Greek letters to differentiate different polypeptide chains in a multimeric protein has nothing to do with the names for the secondary structures α helix and β conformation.

• Some protein structures have very complex quaternary arrangements; e.g., mitochondrial ATP synthase, viral capsids

Symmetry in quaternary structures

• Simplest kind of symmetry = rotational symmetry

• Individual subunits can be superimposed on other identical subunits (brought into coincidence) by rotation about one or more rotational axes.

• If the required rotation = 180° (360°/2), protein has a 2-fold axis of symmetry (e.g., Cro repressor protein above).

• If the rotation = 120° (360°/3), e.g., for a homotrimer, the protein has a 3-fold symmetry axis. Rotational symmetry in proteins: Cyclic symmetry: all subunits are related by

rotation about a single n-fold rotation axis (C2 symmetry has a 2-fold axis, 2 identical subunits; C3 symmetry has a 3-fold axis, 3 identical subunits, etc.)

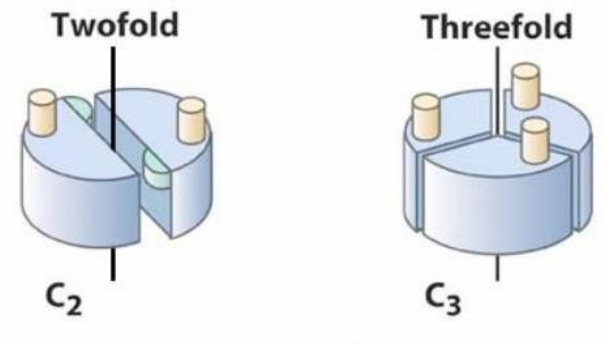


Figure 44 Two common folds of symmetry

Example: Protein Capsid

Viral genomes are surrounded by protein shells known as capsids. One interesting question is how capsid proteins recognize viral, but not cellular RNA or DNA. The answer is that there is often some type of "packaging" signal (sequence) on the viral genome that is recognized by the capsid proteins. A capsid is almost always made up of repeating structural subunits that are arranged in one of two symmetrical structures, a **helix** or an **icosahedron**. In the simplest case, these "subunits" consist of a single polypeptide. In many cases, however, these **structural subunits (also called protomers)** are made up of several polypeptides. Both helical and icosahedral structures are described in more detail below.

1) **Helical Capsids**: The first and best studied example is the plant tobacco mosaic virus (TMV), which contains a SS RNA genome and a protein coat made up of a single, 17.5 kd protein. This protein is arranged in a helix around the viral RNA, with 3 nt of RNA fitting into a groove in each subunit. Helical capsids can also be more complex, and involve more than one protein subunit.

A helix can be defined by two parameters, its amplitude (diameter) and pitch, where pitch is defined as the distance covered by each turn of the helix. **P = m x p**, where m is the number of subunits per turn and p is the axial rise per subunit. For TMV, m = 16.3 and p= 0.14 nm, so P=2.28 nm. This structure is very stable, and can be dissociated and re-associated readily by changing ionic strength, pH, temperature, etc. The interactions that

hold these molecules together are non-covalent, and involve H-bonds, salt bridges, hydrophobic interactions, and vander Waals forces.

Several families of animal virus contain helical nucleocapsids, including the *Orthomyxoviridae* (influenza), the *Paramyxoviridae* (bovine respiratory syncytial virus), and the *Rhabdoviridae* (rabies). All of these are enveloped viruses.

2) **Icosahedral Capsids**: In these structures, the subunits are arranged in the form of a hollow, quasi spherical structure, with the genome within. An icosahedron is defined as being made up of **20 equilateral triangular faces** arranged around the surface of a sphere. They display **2-3-5 fold symmetry** as follows:

- an axis of 2 fold rotational symmetry through the center of each edge.

- an axis of 3 fold rotational symmetry through the center of each face.

- an axis of 5 fold rotational symmetry through the center of each corner. These corners are also called Vertices, and each icosahedron has 12.

Since proteins are not equilateral triangles, each face of an icosahedron contains more than one protein subunit. The simplest icosahedron is made by using 3 identical subunits to form each face, so the minimum # of subunits is 60 (20 x 3). Remember, that each of these subunits could be a single protein or, more likely, a complex of several polypeptides.

Many viruses have too large a genome to be packaged inside an icosahedron made up of only 60 polypeptides (or even 60 subunits), so many are more complicated. In these cases, each of the 20 triangular faces is divided into smaller triangles; and each of these smaller triangles is defined by 3 subunits. However, the total number of subunits is always a multiple of 60. The total number of subunits can be defined as 60 X N, where N is sometimes called the **Triangulation Number**, or T. Values for T of 1,3,4,7,9, 12 and more are permitted.

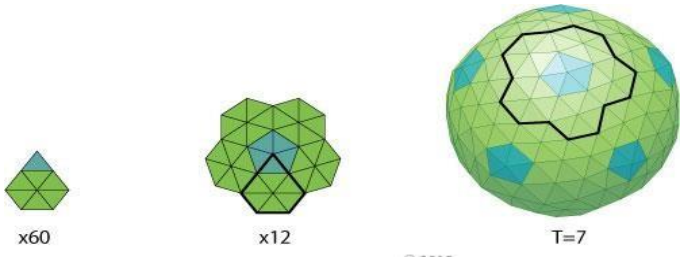


Figure 45 Triangular Number

When virus nucleocapsids are observed in the electron microscope, one often sees apparent "lumps" or clusters on the surface of the particle. These are usually protein subunits clustered around an axis of symmetry, and have been called "morphological units" or **capsomers**.

Forces that stabilize Protein Structure

Proteins are formed of amino acids linked together by the following types of bonds

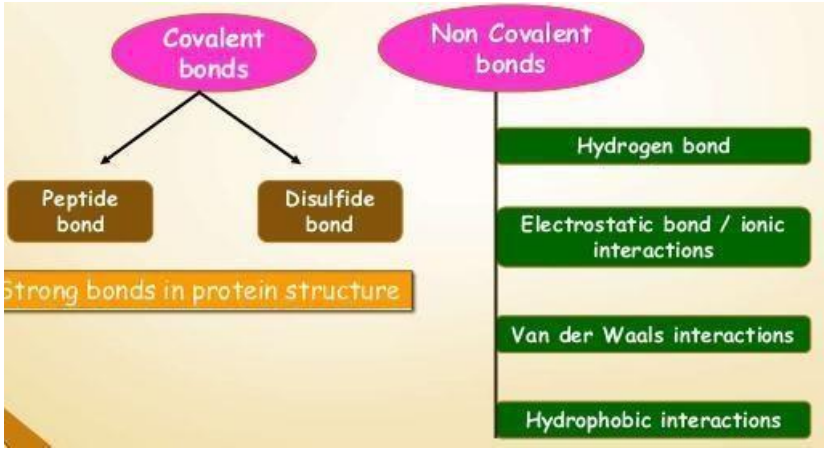


Figure 46 Forces stabilizing protein structure

Covalent Bonds - Disulfide Bridges

Covalent bonds are the strongest chemical bonds contributing to protein structure. Covalent bonds arise when two atoms share electrons.

In addition to the covalent bonds that connect the atoms of a single amino acid and the covalent peptide bond that links amino acids in a protein chain, covalent bonds between

cysteine side chains can be important determinants of protein structure. Cysteine is the sole amino acid whose side chain can form covalent bonds, yielding disulfide bridges with other cysteine side chains: --CH2-S-S-CH2 . A disulfide bridge is shown here:

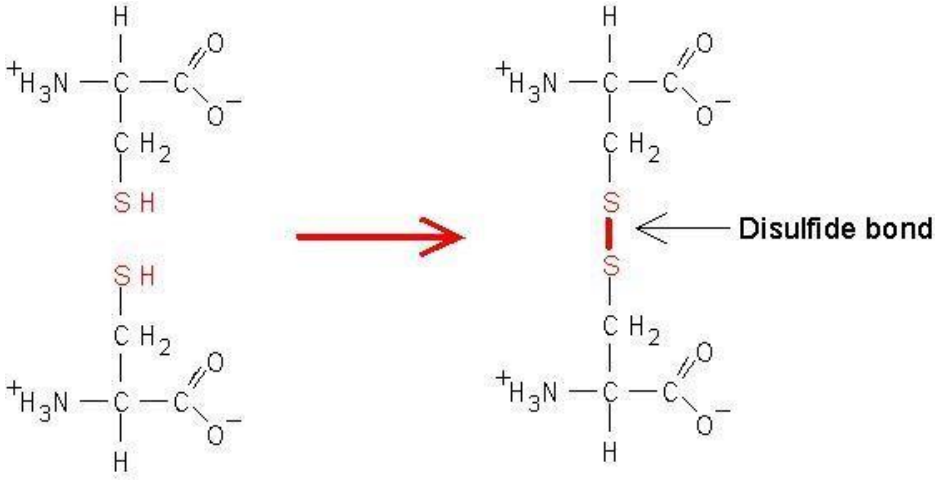


Figure 47 Disulphide bond

Non-covalent bonds Electrostatic Interactions

Ionic Bonds - Salt Bridges

Ionic bonds are formed as amino acids bearing opposite electrical charges are juxtaposed in the hydrophobic core of proteins. Ionic bonding in the interior is rare because most charged amino acids lie on the protein surface. Although rare, ionic bonds can be important to protein structure because they are potent electrostatic attractions that can approach the strength of covalent bonds. An ionic bond-salt bridge between a negatively charged O on the sidechain of glutamic acid lies 2.8 Å from the positively charged N on the amino terminus (lysine) is shown here.

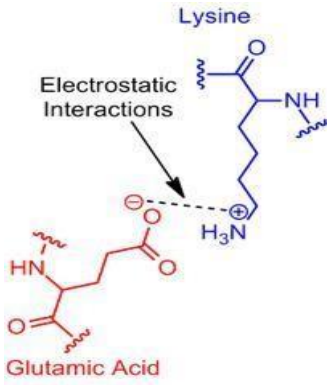


Figure 48 Electrostatic bond

Hydrogen Bonds

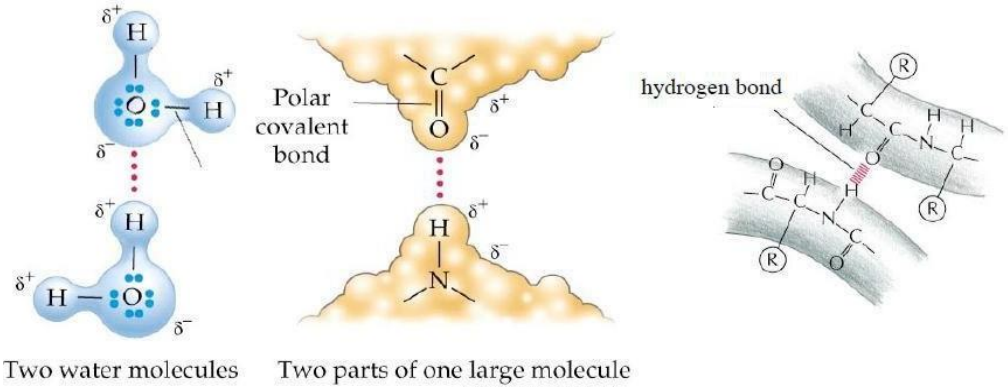


Figure 49 Hydrogen bond

Hydrogen bonds are a particularly strong form of dipole-dipole interaction. Because atoms of different elements differ in their tendencies to hold onto electrons -- that is, because they have different electronegativities -- all bonds between unlike atoms are polarized, with more electron density residing on the more electronegative atom of the bonded pair. Separation of partial charges creates a dipole, which you can think of as a mini-magnet with a positive and a negative end. In any system, dipoles will tend to align so that the positive end of one dipole and the negative end of another dipole are in close proximity. This alignment is favorable.

Hydrogen bonds are dipole-dipole interactions that form between heteroatoms in which one heteroatom (e.g. nitrogen) contains a bond to hydrogen and the other(e.g. oxygen) contains an available lone pair of electrons. You can think of the hydrogen in a hydrogen bond as being shared between the two heteroatoms, which is highly favorable. Hydrogen bonds have an ideal X-H-X angle of 180°, and the shorter they are, the stronger they are. Hydrogen bonds play an important role in the formation of secondary structure. Alpha helices are hydrogen bonded internally along the backbone whereas beta strands are hydrogen bonded to other beta strands. Side chains can also participate in hydrogen bonding interactions. You should be able to list the side chains that can participate in hydrogen bonds now that you know the structures of the side chains. Because hydrogen bonds are directional, meaning the participating dipoles must be aligned properly for a hydrogen bond to form (another w ay of saying it is that the hydrogen bonding angle must be larger than about 135°, with an optimum of 180°), and because unfavorable alignment

of participating dipoles is repulsive, hydrogen bonds between side chains play key roles in determining the unique structures that different proteins form.

Hydrophobic Bonds

Hydrophobic bonds are a major force driving proper protein folding. Burying the nonpolar surfaces in the interior of a protein creates a situation where the water molecules can hydrogen bond with each other without becoming excessively ordered. Thus, the energy of the system goes down.

Therefore, an important factor governing the folding of any protein is the distribution of its polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein such as those belonging to phenylalanine, leucine, isoleucine, valine, methionine and tryptophan tend to cluster in the interior of the molecule (just as hydrophobic oil droplets coalesce in water to form one large droplet). In contrast, polar side chains such as those belonging to arginine, glutamine, glutamate, lysine, etc. tend to arrange themselves near the outside of the molecule, where they can form hydrogen bonds with water and with other polar molecules. There are some polar amino acids in protein interiors, however, and these are very important in defining the precise shape adopted by the protein because the pairing of opposite poles is even more significant than it is in water.

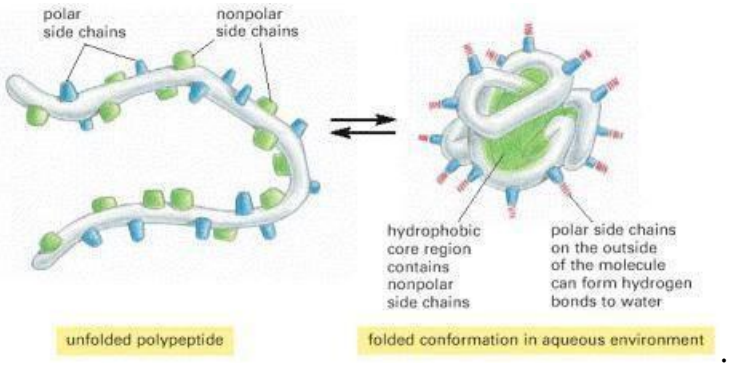


Figure 50 Hydrophobic bonds

Van der Waals Forces

The Van der Waals force is a transient, weak electrical attraction of one atom for another. Van der Waals attractions exist because every atom has an electron cloud that can fluctuate, yielding a temporary electric dipole. The transient dipole in one atom can

induce a complementary dipole in another atom, provided the two atoms are quite close. These short- lived, complementary dipoles provide a weak electrostatic attraction, the Van der Waals force. Of course, if the two electron clouds of adjacent atoms are too close, repulsive forces come into play because of the negatively-charged electrons. The appropriate distance required for Van der Waals attractions differs from atom to atom, based on the size of each electron cloud, and is referred to as the Van der Waals radius. The dots around atoms in this and other displays represent Van der Waals radii.

Van der Waals attractions, although transient and weak, can provide an important component of protein structure because of their sheer number. Most atoms of a protein are packed sufficiently close to others to be involved in transient Van der Waals attractions.

Van der Waals forces can play important roles in protein-protein recognition when complementary shapes are involved. This is the case in antibody-antigen recognition, where a "lock and key" fit of the two molecules yields extensive Van der Waals attractions.

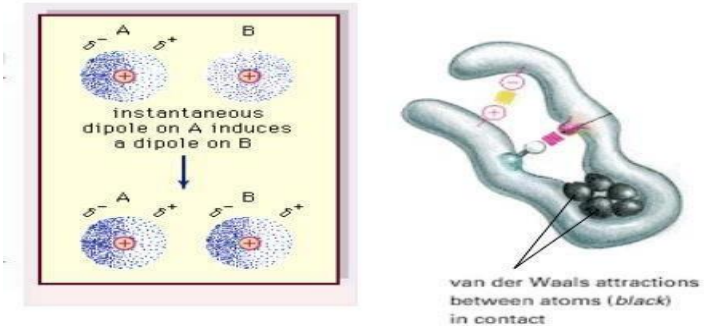


Figure 51 Van der Waals forces

Thermodynamics of protein folding

In contemplating protein folding, it is necessary to consider different types of amino acid side-chains separately. For each situation, the reaction involved will be assumed to be:

$$\text{Protein}_{unfolded} \rightleftharpoons \text{Protein}_{folded}$$

Note that this formalism means that a negative ΔG implies that the folding process is spontaneous.

First we will look at polar groups in an aqueous solvent. For polar groups, the ΔH_{chain} favors the unfolded structure because the backbone and polar groups interact form

stronger interactions with water than with themselves. More hydrogen bonds and electrostatic interactions can be formed in unfolded state than in the folded state. This is true because many hydrogen bonding groups can form more than a single hydrogen bond. These groups form multiple hydrogen bonds if exposed to water, but frequently can form only single hydrogen bonds in the folded structure of a protein.

For similar reasons, the $\Delta H_{solvent}$ favors the folded protein because water interacts more strongly with itself than with the polar groups in the protein. More hydrogen bonds can form in the absence of an extended protein, and therefore the number of bonds in the solvent increases when the protein folds.

The sum of the ΔH_{polar} contributions is close to zero, but usually favors the folded structure for the protein slightly. The chain ΔH contributions are positive, while the solvent ΔH contributions are negative. The sum is slightly negative in most cases, and therefore slightly favors folding. The ΔS_{chain} of the polar groups favors the unfolded state, because the chain is much more disordered in the unfolded state. In contrast, the $\Delta S_{solvent}$ favors the folded state, because the solvent is more disordered with the protein in the folded state. In most cases, the sum of the ΔS_{polar} favors the unfolded state slightly. In other words, the ordering of the chain during the folding process outweighs the other entropic factors.

The ΔG_{polar} that is obtained from the values of ΔH_{polar} and ΔS_{polar} for the polar groups varies somewhat, but usually tends to favor the unfolded protein. In other words, the folding of proteins comprised of polar residues is usually a nonspontaneous process.

Next, we will consider a chain constructed from non-polar groups in aqueous solvent. Once again, the ΔH_{chain} usually favors the unfolded state slightly. Once again, the reason is that the backbone can interact with water in the unfolded state. However, the effect is smaller for non-polar groups, due to the greater number of favorable van der Waals interactions in the folded state. This is a result of the fact that non-polar atoms form better van der Waals contacts with other non-polar groups than with water; in some cases, these effects mean that the ΔH_{chain} for nonpolar residues is slightly negative.

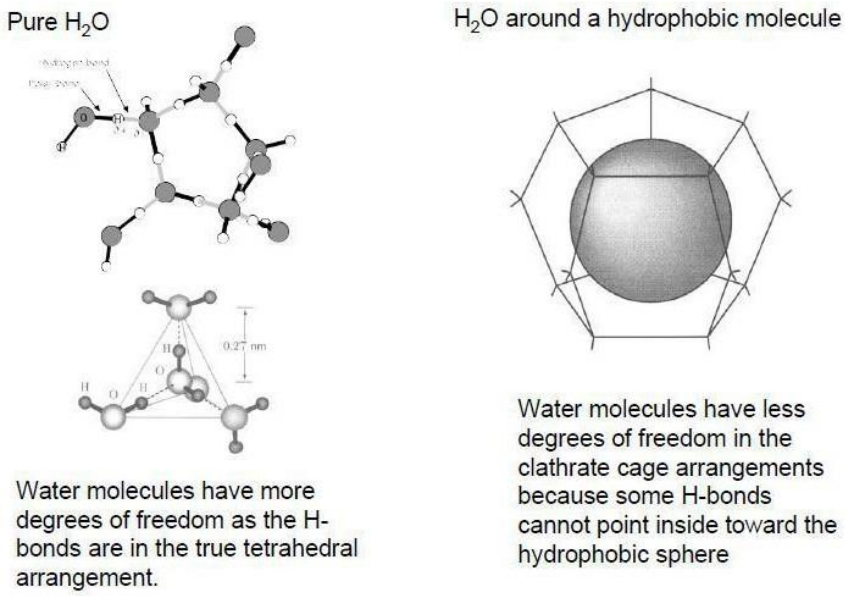


Figure 52 Water environment

As with the polar groups, the $\Delta H_{solvent}$ for non-polar groups favors the folded state. In the case of non-polar residues, $\Delta H_{solvent}$ favors folding more than it does for polar groups, because water interacts much more strongly with itself than it does with non-polar groups.

The sum of the $\Delta H_{non\text{-}polar}$ favors folding somewhat. The magnitude of the $\Delta H_{nonpolar}$ is not very large, but is larger than the magnitude of the ΔH_{polar}, which also tends to slightly favor folding.

The ΔS_{chain} of the non-polar groups favors the less ordered unfolded state. However, the $\Delta S_{solvent}$ highly favors the folded state, due to the hydrophobic effect. During the burying of the non-polar side chains, the solvent becomes more disordered. The $\Delta S_{solvent}$ is a major driving force for protein folding which is called conformational entropy.

The $\Delta G_{non\text{-}polar}$ is therefore negative, due largely to the powerful contribution of the $\Delta S_{solvent}$. Adding together the terms for ΔG_{polar} and $\Delta G_{non\text{-}polar}$ gives a slightly negative overall ΔG for protein folding and therefore, proteins generally fold spontaneously.

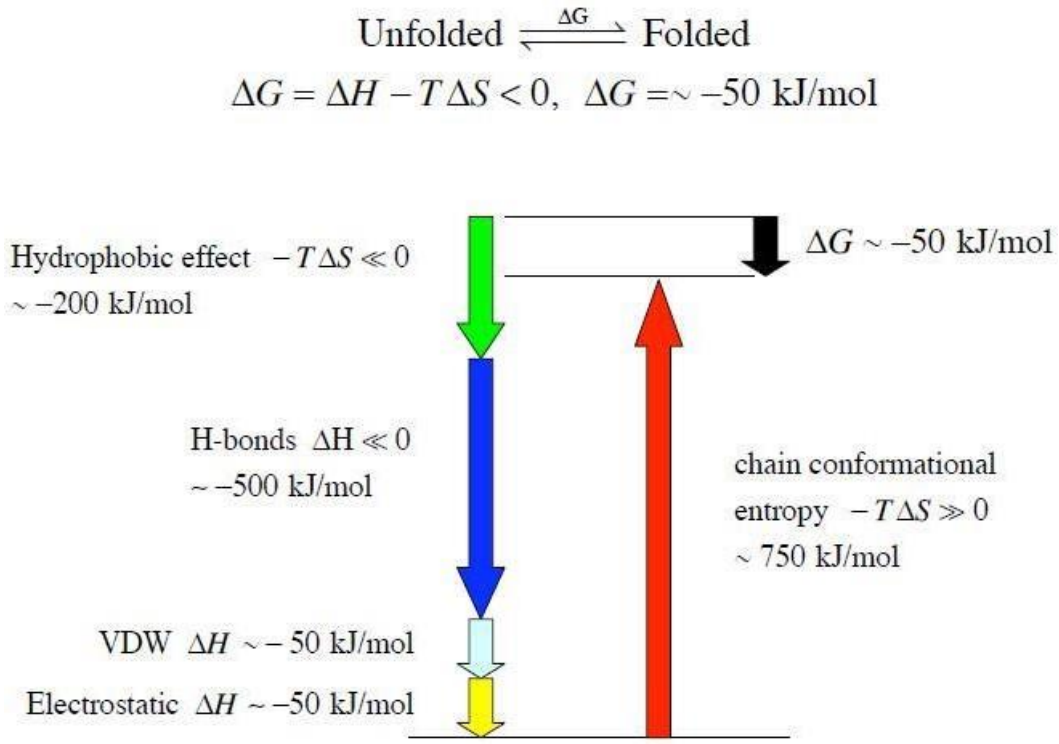


Figure 53 Thermodynamic changes

Raising the temperature, however, tends to greatly increase the magnitude of the $T\Delta S_{chain}$ term, and therefore to result in unfolding of the protein.

The folded state is the sum of many interactions. Some favor folding, and some favor the unfolded state. The qualitative discussion above did not include the magnitudes of the effects. For real proteins, the various ΔH and ΔS values are difficult to measure accurately. However, for many proteins it is possible to estimate the overall ΔG of folding. Measurements of this value have shown that the overall ΔG for protein folding is very small: only about –10 to –50 kJoules/mol. This corresponds to a few salt bridges or hydrogen bonds.

Studies of protein folding have revealed one other important point: the hydrophobic effect is very important, but it is relatively non-specific. Any hydrophobic group will interact with essentially any other hydrophobic group. While the hydrophobic effect is a major driving force for protein folding, it is the constrains imposed by the more geometrically specific hydrogen bonding and electrostatic interactions in conjunction with the hydrophobic interactions that largely determine the overall folded structure of the protein.

PROTEIN FOLDING MECHANISM

Protein Folding

Protein folding is a process in which a polypeptide folds into a specific, stable, functional, three-dimensional structure. It is the process by which a protein structure assumes its functional shape or conformation Proteins are formed from long chains of amino acids; they exist in an array of different structures which often dictate their functions. Proteins follow energetically favorable pathways to form stable, orderly, structures; this is known as the protein's native structure. Most proteins can only perform their various functions when they are folded. The protein folding pathway, or mechanism, is the typical sequence of structural changes the protein undergoes in order to reach its native structure. Protein folding takes place in a highly crowded, complex, molecular environment within the cell, and often requires the assistance of molecular chaperones, in order to avoid aggregation or misfolding. Proteins are comprised of amino acids with various types of side chains, which may be hydrophobic, hydrophilic, or electrically charged. The characteristics of these side chains affect what shape the protein will form because they will interact differently intra molecularly and with the surrounding environment, favoring certain conformations nd structures over others. Scientists believe that the instructions for folding a protein are encoded in the sequence. Researchers and scientists can easily determine the sequence of a protein, but have not cracked the code that governs folding.

Protein Folding theory and experiment

Early scientists who studied proteomics and its structure speculated that proteins had templates that resulted in their native conformations. This theory resulted in a search for how proteins fold to attain their complex structure. It is now well known that under physiological conditions, proteins normally spontaneously fold into their native conformations. As a result, a protein's primary structure is valuable since it determines the three-dimensional structure of a protein. Normally, most biological structures do not have the need for external templates to help with their formation and are thus called self-assembling.

Protein Renaturation

Protein renaturation known since the 1930s. However, it was not until 1957 when Christian Anfinsen performed an experiment on bovine pancreatic RNase A that protein renaturation was quantified. RNase A is a single chain protein consisting of 124 residues. In 8M urea solution of 2-mercaptoethanol, the RNase A is completely unfolded and has its four disulfide bonds cleaved through reduction. Through dialysis of urea and

introducing the solution to O2 at pH 8, the enzymatically active protein is physically incapable of being recognized from RNase A. As a result, this experiment demonstrated that the protein spontaneously renatured.

One criteria for the renaturation of RNase A is for its four disulfide bonds to reform. The likelihood of one of the eight Cys residues from RNase A reforming a disulfide bond with its native residue compared to the other seven Cys residues is 1/7. Futher more, the next one of remaining six Cys residues randomly forming the next disulfide bond is 1/5 and etc. As a result, the probability of RNase A reforming four native disulfide links at random is (1/7 *1/5 * 1/3 * 1/1 = 1/105). The result of this probability demonstrates that forming the disulfide bonds from RNase A is not a random activity.

When RNase A is reoxidized utilizing 8M urea, allowing the disulfide bonds to reform when the polypeptide chain is a random coil, then RNase A will only be around 1 percent enzymatically active after urea is removed. However, by using 2-mercaptoethanol, the protein can be made fully active once again when disulfide bond interchange reactions occur and the protein is back to its native state. The native state of the RNase A is thermodynamically stable under physiological conditions, especially since a more stable protein that is more stable than that of the native state requires a larger activation barrier, and is kinetically inaccessible. By using the enzyme protein disulfide isomerase (PDI), the time it takes for randomized RNase A is minimized to about 2 minutes. This enzyme helps facilitate the disulfide interchange reactions. In order for PDI to be active, its two active site Cys residues needs to be in the -SH form. Furthermore, PDI helps with random cleavage and the reformation of the disulfide bonds of the protein as it attain thermodynamically favorable conformations.

Post translationally Modified Proteins Might Not Renature

Proteins in a "scrambled" state go through PDI to renature, and their native state does not utilize PDI because native proteins are in their stable conformations. However, proteins that are post translationally modified need the disulfide bonds to stabilize their rather unstable native form. One example of this is insulin, a polypeptide hormone. This 51 residue polypeptide has two disulfide bonds that is inactivated by PDI. The following link is an image showing insulin with its two disulfide bonds. Through observation of this phenomena, scientists were able to find that insulin is made from proinsulin, an 84-residue single chain. This link provides more information on the structure of proinsulin and its progression on becoming insulin. The disulfide bonds of proinsulin need to be intact before conversion of becoming insulin through proteolytic excision of its C chain

which is an internal 33-residue segment. However according to two findings, the C chain is not what dictates the folding of the A and B chains, but instead holds them together to allow formation of the disulfide bonds. For one, with the right renaturing conditions in place, scrambled insulin can become its native form with a 30% yield. This yield can be increased if the A and B chains are cross-linked. Secondly, through analysis of sequences of proinsulin from many species, mutations are permitted at the C chain eight times more than if it were for A and B chains.

The Protein Folding Process

Considerable evidence suggests that all of the information to describe the three dimensional conformation of a protein is contained within the primary structure. However, for the most part, we cannot fully interpret the information contained within the sequence. To understand why this is true, we need to take a more careful look at proteins and how they fold.

The polypeptide chain for most proteins is quite long. It therefore has ***many*** possible conformations. If you assume that all residues could have 2 possible combinations (real peptides can have many more than this), a 100 amino acid peptide could have 2^{100} ($\sim 10^{30}$) possible conformations. If the polypeptide tested a billion conformations/second, it would still take over 10^{13} years to find the correct conformation. (Note that the universe is only $\sim 10^{10}$ years old, and that a 100 residue polypeptide is a relatively small protein.) The observation that proteins cannot fold by random tests of all possible conformations is referred to as the Levinthal paradox.

Folding pathways

In classical transition state theory, the reaction diagram for a spontaneous two state system is considered to have a high-energy starting material, a lower energy product, and an energy barrier between them. While the typical diagram that describes the process (such as the one shown below) is useful, it is incomplete.

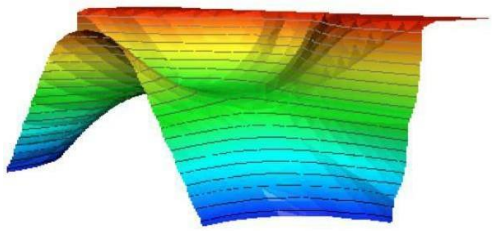


Figure 54

The process for the conversion of S to P could actually take many pathways; the pathway shown is merely the minimum energy route from one state to another. The true situation is described by an energy landscape, with the minimum energy route being the equivalent of a pass between two mountains. Thus, although the pathway involves an energy barrier, other pathways require passing through even higher energy states.

A large part of the reason that single pathways (or small numbers of pathways) exist for chemical reactions is that most reactions involve the cleavage and reformation of covalent bonds. The energy barrier for breaking a covalent bond is usually quite high. In protein folding, however, the interactions involved are weak. Because the thermal energy of a protein molecule is comparable to the typical non covalent interaction strength, an unfolded polypeptide is present in a variety of rapidly changing conformations. This realization led to the Levinthal paradox: because the unfolded protein should be constantly changing its shaped due to thermal motions of the different parts of the polypeptide, it seemed unlikely that the protein would be able to find the correct state to begin transiting a fixed folding pathway.

An alternate hypothesis has been proposed, in which *portions* of the protein self-organize, followed by folding into the final structure. Because the different parts of the protein begin the folding process independently, the shape of the partially folded protein can be very variable. In this model, the protein folds by a variety of different paths on an energy landscape. The folding energy landscape has the general shape of a **funnel**. In the folding process, as long as the overall process results in progressively lower energies, there can be a large variety of different pathways to the final folded state.

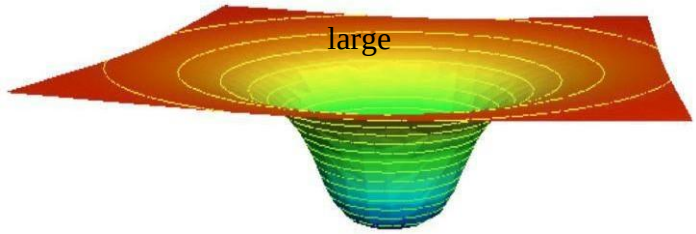


Figure 55

The folding funnel shown above has a smooth surface. Actual folding funnels may be fairly smooth, or may have irregularities in the surface that can act to trap the polypeptide chain in misfolded states. Alternatively, the folding funnel may direct the polypeptide into a ***metastable*** state. Metastable states are local minima in the landscape; if the energy barriers that surround the state are high enough, the metastable state may exist for a long

time – metastable states are stable for **kinetic** rather than **thermodynamic** reasons.
The difficulty in refolding many proteins *in vitro* suggests that the folded state of at least some complex proteins may be in a metastable state rather than a global energy minimum.

Folding process

The lower energies observed toward the depression in the folding funnel are thought to be largely due to the collapse of an extended polypeptide due to the hydrophobic effect. In addition to the hydrophobic effect, de solvation of the backbone is necessary for protein folding, at least for portions of the backbone that will become buried. One method for desolvation of the backbone is the formation of secondary structure. This is especially true for helical structures, which can form tightly organized regions of hydrogen bonding while excluding water from the backbone structure. A general outline for the process experienced by a folding protein seems to look like this:
A general outline for the process experienced by a folding protein seems to look like this:

1. Some segments of a polypeptide may rapidly attain a relatively stable, organized structure (largely due to organization of secondary structural Elements).
2. These structures provide nuclei for further folding.
3. During the folding process, the protein is proposed to form a state called a **Molten globule**. This state readily rearranges to allow interactions between different parts of the protein.
4. These nucleated, partially folded domains then coalesce into the folded protein.

If this general pathway is correct, it seems likely that at least some of the residues within the sequence of most proteins function to guide the protein into the proper folding pathway, and prevent the ―trapping‖ of the polypeptide in unproductive Partially folded states.

Folding inside cells

Real cells contain **many** proteins at a high overall protein concentration. The protein concentration inside a cell is ~150 mg/ml. folding inside cells differs from most experiments used to study folding *in vitro*:

- Proteins are synthesized on ribosomes. The entire chain is not available to fold at once, as is the case for an experimentally unfolded protein in a test tube.
- Within cells, the optimum ionic concentration, pH, and macromolecule Concentration for each protein to fold properly cannot be controlled as tightly as in an

experimental system.

- Major problems could arise if unfolded or partially folded proteins encountered one another. Exposed hydrophobic regions might interact, and form potentially lethal insoluble aggregates within the cell.

One mechanism for limiting problems with folding proteins inside cells volves specialized proteins called **molecular chaperones**, which assist in folding proteins. Molecular chaperones were first observed to be involved in responses to elevated temperature (*i.e.* ―heat shock‖) to stabilize existing proteins and prevent protein aggregation and were called heat-shock proteins (abbreviated as ―hsp‖). Additional research revealed that heat shock proteins are present in all cells, and that they decrease or prevent non-specific protein aggregation and assist in protein folding.

MOLECULAR CHAPERONES

In molecular biology, **molecular chaperones** are proteins that assist the covalent folding or unfolding and the assembly or disassembly of other macromolecular structures. Chaperones are present when the macromolecules perform their normal biological functions and have correctly completed the processes of folding and/or assembly. The chaperones are concerned primarily with protein folding. The first protein to be called a chaperone assists the assembly of nucleosomes from folded histones and DNA and such assembly chaperones, especially in the nucleus, are concerned with the assembly of folded subunits into oligomeric structures.

One major function of chaperones is to prevent both newly synthesised polypeptide chains and assembled subunits from aggregating into nonfunctional structures. It is for this reason that many chaperones, but by no means all, are heat shock proteins because the tendency to aggregate increases as proteins are denatured by stress. In this case, chaperones do not convey any additional stericinformation required for proteins to fold. However, some highly specific 'steric chaperones' do convey unique structural (steric) information onto proteins, which cannot be folded spontaneously. Such proteins violate Anfinsen's dogma.

Various approaches have been applied to study the structure, dynamics and functioning of chaperones. Bulk biochemical measurements have informed us on the protein folding efficiency, and prevention of aggregation when chaperones are present during protein folding. Recent advances in single-molecule analysis have brought insights into structural heterogeneity of chaperones, folding intermediates and affinity of chaperones for

unstructured and structured protein chains.

Properties

- Molecular chaperones interact with unfolded or partially folded protein subunits, e.g. nascent chains emerging from the ribosome, or extended chains being translocated across subcellular membranes.
- They stabilize non-native conformation and facilitate correct folding of protein subunits.
- They do not interact with native proteins, nor do they form part of the final folded structures.
- Some chaperones are non-specific, and interact with a wide variety of polypeptide chains, but others are restricted to specific targets.
- They often couple ATP binding/hydrolysis to the folding process.
- Essential for viability, their expression is often increased by cellular stress.

Main role: They prevent inappropriate association or aggregation of exposed hydrophobic surfaces and direct their substrates into productive folding, transport or degradation pathways.

Location and Function

Many chaperones are heat shock proteins, that is, proteins expressed in response to elevated temperatures or other cellular stresses. The reason for this behaviour is that protein folding is severely affected by heat and, therefore, some chaperones act to prevent or correct damage caused by misfolding. Other chaperones are involved in folding newly made proteins as they are extruded from the ribosome. Although most newly synthesized proteins can fold in absence of chaperones, a minority strictly requires them for the same. Some chaperone systems work as foldases: they support the folding of proteins in an ATP- dependent manner (for example, the GroEL/GroES or the DnaK/DnaJ/GrpE system). Other chaperones work as holdases: they bind folding intermediates to prevent their aggregation, for example DnaJ or Hsp33.

Macromolecular crowding may be important in chaperone function. The crowded environment of the cytosol can accelerate the folding process, since a compact folded protein will occupy less volume than an unfolded protein chain. However, crowding can reduce the yield of correctly folded protein by increasing protein aggregation. Crowding

may also increase the effectiveness of the chaperone proteins such as GroEL, which could counteract this reduction in folding efficiency.

More information on the various types and mechanisms of a subset of chaperones that encapsulate their folding substrates (e.g. GroES) can be found in the chaperonins. Chaperonins are characterized by a stacked double-ring structure and are found in prokaryotes, in the cytosol of eukaryotes, and in mitochondria.

Other types of chaperones are involved in transport across membranes, for example membranes of the mitochondria and endoplasmic reticulum (ER) in eukaryotes. Bacterial translocation—specific chaperone maintains newly synthesized precursor polypeptide chains in a translocation-competent (generally unfolded) state and guides them to the translocon.

New functions for chaperones continue to be discovered, such as assistance in protein degradation, bacterial adhesin activity, and in responding to diseases linked to protein aggregation (e.g. see prion) and cancer maintenance.

CHEPARONINS

Chaperonins are proteins that provide favourable conditions for the correct folding of other proteins, thus preventing aggregation. Newly made proteins usually must fold from a linear chain of amino acids into a three-dimensional form. Chaperonins belong to a large class of molecules that assist protein folding, called molecular chaperones. The energy to fold proteins is supplied by adenosine triphosphate

GroupI Chaperonins

GroupI Chaperonins are found in bacteria as welas organelles of endosymbiotic origin: chloroplasts and mitochondria. The GroEL/GroES complex in *E. coli* is a Group I chaperonin and the best characterized large (~ 1 MDa) chaperonin complex.

1. GroEL is a double-ring 14mer with a greasy hydrophobic patch at its opening and can accommodate the native folding of substrates 15-60 kDa in size.

2. GroES is a single-ring heptamer that binds to GroEL in the presence of ATP or transition state analogues of ATP hydrolysis, such as ADP-AlF3. It's like a cover that covers GroEL (box/bottle). GroEL/GroES may not be able to undo protein aggregates, but kinetically it competes in the pathway of misfolding and aggregation, thereby preventing aggregate formation.

Group II chaperonins: found in the eukaryotic cytosol and in archaea, are more poorly characterized. TRiC (TCP-1 Ring Complex, also called CCT for chaperonin containing

TCP-1), the eukaryotic chaperonin, is composed of two rings of eight different though related subunits, each thought to be represented once per eight-membered ring. TRiC was originally thought to fold only the cytoskeletal proteins actin and tubulin but is now known to fold dozens of substrates. Mm cpn (*Methanococcus maripaludis* chaperonin), found in the archaea *Methanococcus maripaludis*, is composed of sixteen identical subunits (eight per ring). It has been shown to fold the mitochondrial protein rhodanese; however, no natural substrates have yet been identified. Group II chaperonins are not thought to utilize a GroES-type cofactor to fold their substrates. They instead contain a "built-in" lid that closes in an ATP-dependent manner to encapsulate its substrates, a process that is required for optimal protein folding activity.

Mechanism of action

Chaperonins undergo large conformational changes during a folding reaction as a function of the enzymatic hydrolysis of ATP as well as binding of substrate proteins and cochaperonins, such as GroES. These conformational changes allow the chaperonin to bind an unfolded or misfolded protein, encapsulate that protein within one of the cavities formed by the two rings, and release the protein back into solution. Upon release, the substrate protein will either be folded or will require further rounds of folding, in which case it can again be bound by a chaperonin. The exact mechanism by which chaperonins facilitate folding of substrate proteins is unknown. According to recent analyses by different experimental techniques, GroEL-bound substrate proteins populate an ensemble of compact and locally expanded states that lack stable tertiary interactions. A number of models of chaperonin action have been proposed, which generally focus on two (not mutually exclusive) roles of chaperonin interior: passive and active. Passive models treat the chaperonin cage as an inert form, exerting influence by reducing the conformational space accessible to a protein substrate or preventing intermolecular interactions e.g. by aggregation prevention. The active chaperonin role is in turn involved with specific chaperonin–substrate interactions that may be coupled to conformational rearrangements of the chaperonin. Probably the most popular model of the chaperonin active role is the iterative annealing mechanism (IAM), which focus on the effect of iterative, and hydrophobic in nature, binding of the protein substrate to the chaperonin. According to computational simulation studies, the IAM leads to more productive folding by unfolding the substrate from misfolded conformations or by prevention from protein misfolding through changing the folding pathway.

Human Chaperone Proteins

Chaperones are found in, for example, the endoplasmic reticulum (ER), since protein synthesis often occurs in this area.

Endoplasmic reticulum

In the endoplasmic reticulum (ER) there are general, lectin- and non-classical molecular chaperones helping to fold proteins.

- General chaperones: GRP78/BiP, GRP94, GRP170.
- Lectin chaperones: calnexin and calreticulin
- Non-classical molecular chaperones: HSP47 and ERp29
- Folding chaperones:
- Protein disulfide isomerase (PDI),
- Peptidyl prolyl cis-trans-isomerase (PPI)
- ERp57

Nomenclature and examples of bacterial and archael chaperons.

There are many different families of chaperones; each family acts to aid protein folding in a different way. In bacteria like *E. coli*, many of these proteins are highly expressed under conditions of high stress, for example, when the bacterium is placed in high temperatures. For this reason, the term "heat shock protein" has historically been used to name these chaperones. The prefix "Hsp" designates that the protein is a heat shock protein.

Hsp60

Hsp60 (GroEL/GroES complex in *E. coli*) is the best characterized large (~ 1 MDa) chaperone complex. GroEL is a double-ring 14mer with a hydrophobic patch at its opening; it is so large it can accommodate native folding of 54-kDa GFP in its lumen. GroES is a single-ring heptamer that binds to GroEL in the presence of ATP or ADP. GroEL/GroES may not be able to undo previous aggregation, but it does compete in the pathway of misfolding and aggregation.[19] Also acts in mitochondrial matrix as molecular chaperone.

Hsp70

Hsp70 (DnaK in *E. coli*) is perhaps the best characterized small (~ 70 kDa) chaperone.

The Hsp70 proteins are aided by Hsp40 proteins (DnaJ in *E. coli*), which increase the ATP consumption rate and activity of the Hsp70s. It has been noted that increased expression of Hsp70 proteins in the cell results in a decreased tendency toward apoptosis. Although a

precise mechanistic understanding has yet to be determined, it is known that Hsp70s have a high-affinity bound state to unfolded proteins when bound to ADP, and a low-affinity state when bound to ATP. It is thought that many Hsp70s crowd around an unfolded substrate, stabilizing it and preventing aggregation until the unfolded molecule folds properly, at which time the Hsp70s lose affinity for the molecule and diffuse away. Hsp70 also acts as a mitochondrial and chloroplastic molecular chaperone in eukaryotes.

Hsp90

Hsp90 (HtpG in *E. coli*) may be the least understood chaperone. Its molecular weight is about 90 kDa, and it is necessary for viability in eukaryotes (possibly for prokaryotes as well).Heat shock protein 90 (Hsp90) is a molecular chaperone essential for activating many signaling proteins in the eukaryotic cell. Each Hsp90 has an ATP-binding domain, a middle domain, and a dimerization domain.

Hsp100

Hsp100 (Clp family in *E. coli*) proteins have been studied *in vivo* and *in vitro* for their ability to target and unfold tagged and mis folded proteins. Proteins in the Hsp100/Clp family form large hexameric structures with unfoldase activity in the presence of ATP. These proteins are thought to function as chaperones by processively threading client proteins through a small 20 Å (2 nm) pore, thereby giving each client protein a second chance to fold. Some of these Hsp100 chaperones, like ClpA and ClpX, associate with the double-ringed tetradecameric serine protease ClpP; instead of catalyzing the refolding of client proteins, these complexes are responsible for the targeted destruction of tagged and misfolded proteins. Hsp104, the Hsp100 of Saccharomyces cerevisiae, is essential for the propagation of many yeast prions. Deletion of the HSP104 gene results in cells that are unable to propagate certain prions.

Protein stability

Protein stability is another common problem in protein expression. It is also an important topic in purification, formulation, and storage. Here we will discuss about protein stability in expression only. Properly folded proteins are usually stable during expression and purification. Sufficient amount of intact protein should be obtained. However some proteins appear to be unstable during expression and purification. Some

of them are so unstable that sufficient amount of protein cannot be obtained. Many factors such as amino acid sequence of the protein, protein construction, host cell strain, expression and purification conditions may affect protein stability. Amino acid sequence of a protein itself may be susceptible to degradation. Certain amino acids at the N-terminus of a protein can lead the protein to degradation. These are Arg, Lys, Leu, Phe, Tyr, and Trp residues. Replacing these amino acids with others can greatly increase the protein half-life (N-end rule). Many recombinant proteins are expressed with tags or fusion partners. Amino acid sequences at N-termini of these tags and fusion partners are often optimized for protein yield and stability. Therefore amino acids at N-terminus are not a problem in protein stability for these tagged or fusion proteins. It is reported that regions containing Pro (P), Glu (E), Ser (S), and Thr (T) termed PEST are prone to degradation. It is generally observed that flexible hydrophilic sequences with protease cleavage sites are easily degraded. These sequences may be integral part of a protein. In most cases these sequences cannot be deleted or mutated. Strategies for improving protein stability are needed for these proteins.

Strategies to improve protein Stability: Perform expression in special media containing trace metals, minerals, and vitamins. These chemicals may not be needed for host cell growth, but they may serve as co-factor, prosthetic groups or ligands for recombinant proteins. Therefore they may be critical for correct protein folding and stability. Medium pH should also be balanced near neutral to improve protein stability. There will be no protein degradation caused by nutrition.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit 2 - Protein Engineering and Bioinformatics – SBTA1303

II STRUCTURAL CHARACTERIZATION OF PROTEINS

Frederic Sanger first time achieved complete sequence of protein (bovine insulin) in 1953. For his work, he was awarded the Nobel Prize of Chemistry in (1958).

Protein sequencing refers to the techniques employed to determine the amino acid sequence of a protein. There are several applications of protein sequencing, which are:-

a) Identification of the protein family to which a particular protein belongs and finding the evolutionary history of that protein. Function prediction.

b) Prediction of the cellular localization of the protein based on its target sequence (sequence of amino acids at the N terminal end of the protein which determines the location of the protein inside the cell).

c) Prediction of the sequence of the gene encoding the particular protein.

d) Discovering the structure and function of a protein through various computational methods and experimental methods.

Till date several methods have been utilized for protein sequencing. Two main methods include Edman degradation and Mass Spectrometry. Protein sequence can also be generated from the DNA/mRNA sequence that codes for the protein, which has been explained in details in the recombinant DNA section. Here, we have discussed the most important methods used for protein sequencing and the pros and cons of each method.

Edman degradation

Before sequencing process is initiated, it is necessary to break all non-covalent interactions by denaturants (like high concentration of urea or GuHCl). This process will also separate subunits, in case of oligomeric proteins. Occasionally, subunits of an oligomeric protein are connected by covalent interactions. In that case special treatments are required to separate subunits. The protein is treated with Edman‟s reagent (phenyl isothiocyanate) which reacts with the N-terminal amino acid and under mild acidic condition forms a cyclic compound Phenyl thiohydantoin derivative (PTH–amino acid) of N-terminal amino acid is released. Amino acid of PTH –amino acid derivative is identified by chromatographic property of the PTH –amino acid derivative. In this process N-terminal amino acid is identified after first cycle. ***Since this method proceeds from the N terminal residue, the reaction will not work if that N-terminal of a protein is blocked (generally due to post-translational modification)***.

After first cycle of the reaction, amino group of the second amino acid is free for reaction with Edman‟s reagent and at the end of reaction PTH derivative of second amino acid from N-terminal is released. The process continues till end of sequence or a disulfide bond is encountered in the sequence. PTH-cysteine derivative will remain attached with polypeptide and PTH-cystein will not be released.

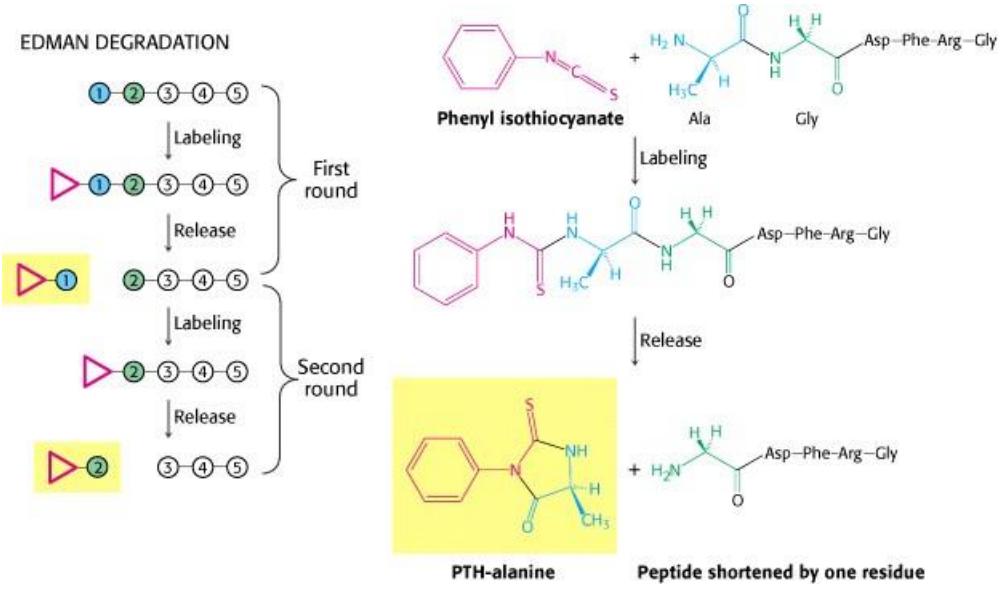


Figure 1 Scheme of protein sequencing by Edman degradation

Thus, reduction of disulfide bond in the polypeptide sequence needed before sequencing process can be initiated. Reduction of cystine can be done by use of Beta-marcaptoethanol (Fig. 2)

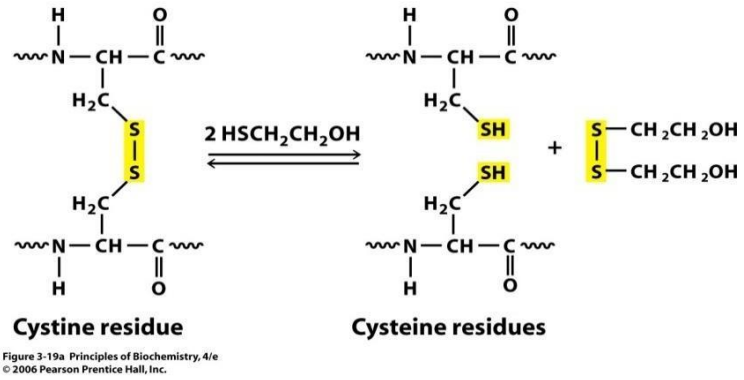


Figure 2 Reduction of cystine

As free cysteine can re-oxidize to form disulfide it is necessary to block free cysteine. This may be done by use of iodoacetic acid or acrylonitrile (free cysteine modification) as shown in Fig3.

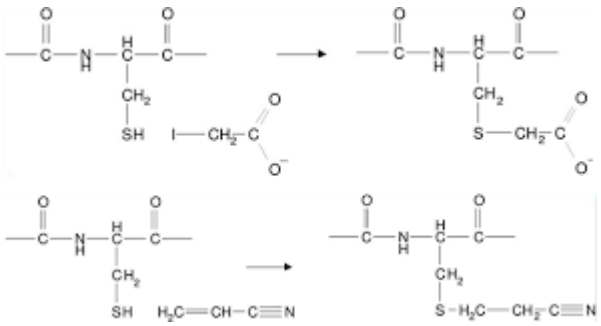


Figure 3 Free cysteine modification

Other method for irreversible oxidation of disulfide bond is use of performic acid. As shown in the figure below, performic acid oxidizes cysteine to negatively charge cysteic acid. Repulsion of negatively charged cysteic acid group prevents re-formation of disulfide and alkylation is not required.

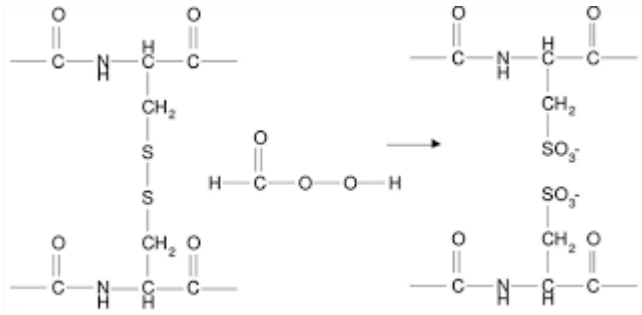


Figure 4 Irreversible oxidation of disulfide bond

Further, the accuracy of each cycle is 98%. So after 60 steps the accuracy is less than 30%. Thus, this method cannot be used for sequencing of proteins larger than 50 amino acids. In case of larger proteins it has to be broken down to short peptide fragments using cleavage proteases such as trypsin (cleaves a protein at carboxyl side of lysine and arginine residues) or chymotrypsin (cleaves at carboxyl side of tyrosine, tryptophan and phenylalanine). Specific cleavage can also be achieved by chemical methods like cyanogen bromide, which always

cleaves at carboxyl side of methionine residue (a protein with 12 methionine will yield 13 fragment polypeptide on cleavage with cyanogen bromide (CNBr).

Protein fragments after a protease (for example trypsin) will be separated and sequenced. Let us assume that the following two peptide sequences are obtained.

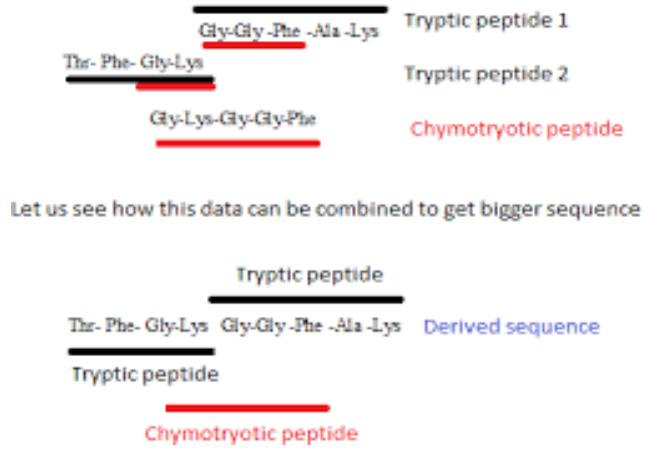


Figure 5 Protease action

2) Protein sequencing using Sanger's reagent and dansyl chloride

Here, the N terminal amino acid of the protein is labeled by dyes like Sanger"s reagent (fluoro-dinitrobenzene) or dansyl chloride. The labeled protein is then hydrolyzed by 6M HCl at 110 ^{o}C by the above mentioned method and loaded in Dowex 50 column and the elution profile is matched with the standard profile obtained from FNB or DNSCl derivative of all the amino acids, to obtain the N terminal amino acid. The reagents produce coloured derivatives which can be easily detected by absorbance (Fig. 6.)

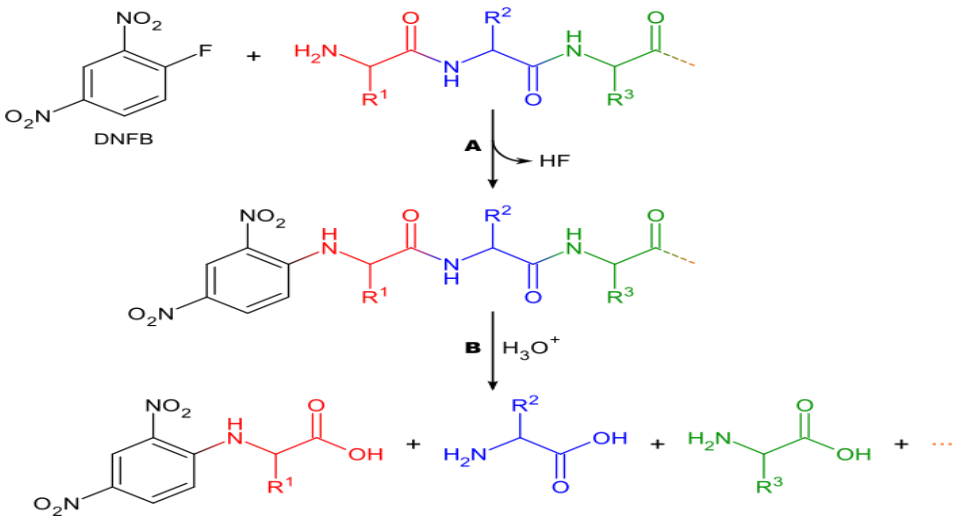


Figure 6 Protein sequencing using Sanger's reagent

Disadvantages of this method include:

 Once we get the N terminal amino acid, the protein is already hydrolyzed in constituent amino acids. Thus we cannot repeat the cycle with same sample. For second amino acid sequencing we require new stock of protein sample and the N-terminal residue need to be cleaved from the protein using an appropriate protease such as amino peptidase. This makes the process very tedious and complicated.

 These dyes selectively labels the amine groups present in the protein and therefore can label the amine groups present in the side chains as well, which may give erroneous results.

Protein sequencing using Molecular Biology techniques

If first few N-terminal amino acid of a protein is known, complete aminoacid sequence can be derived using Molecular Biology techniques. A simple example is as follow:

The genome sequence of *Calotropisprocera,* a plant, or the sequence of procerain B, a novel cystein protease from the plant, gene is not yet known. Thus, the only information for cloning of cDNA we have is the fifteen N-terminal amino acid residues. The double stranded cDNA can be amplified with help of degenerate primer (based of N-terminal amino acid sequence) and oligodT primer. Total RNA can be isolated from young leaf or latex of the plant and first strand of cDNA can be synthesised with oligodT primer by reverse transcription. The second strand of cDNA can be synthesised and the subsequent amplification of double stranded cDNA can be achieved by PCR with degenerate primer as forward and oligodT primer as reverse primmer. The amplified double stranded cDNA of expected size can be subjected to TA cloning and confirmed by sequencing. Once sequence of cDNA is available, it can be translated in protein sequence.

PROTEIN STRUCTURE DETERMINATION X-RAY DIFFRACTION

Historical outline

The method of protein crystallography originates from the discovery of X-rays by Conrad Röntgen, and the subsequent developments by Max von Laue, who was first to observe diffraction of X-rays and revealed the wave nature of X-rays. These discoveries were followed by the experiments by the Brags (father and son), who showed that X-ray diffraction could be used in the determination of the atomic structure of matter. However, the world had to wait for additional 45 years before the first protein structure was determined by protein crystallography. This was the structure of myoglobin, which gave the authors, Max Perutz and John Kendrew the Chemistry Nobel Prize in 1962. Since then several other protein

crystallographic structures have been awarded the Nobel Prize. Among these is the prize awarded to Dorothy Hodgkin for the structures of vitamin B12 and insulin (Chemistry Prize of 1964); Johann Deisenhofer, Robert Huber and Hartmut Michel for the determination of the structure of the first membrane protein, the photosynthetic reaction center (Chemistry Prize of 1988); John E Walker for his role in the determination of the structure of ATP synthase (Chemistry Prize of 1997). Recent prizes related to protein crystallography include those awarded to Peter Agre & Roderick MacKinnon (Chemistry Prize of 2003), Roger Kornberg (Chemistry Prize of 2006), Venki Ramakrishnan, Thomas A. Steitz, Ada Yonath for the elucidation of the ternary structure of the ribosome (Chemistry Prize of 2009), and recently Brian Kobilka and Robert Lefkowitz for functional and structural studies of GPCR proteins (Chemistry Prize, 2012). Protein X-ray crystallography and NMR spectroscopy are currently the only two methods, which provide atomic resolutionary protein structures. Although, with around 90 000 entries in the Protein Data Bank (PDB), of which almost 80 000 were determined by diffraction methods, one could say that the method dominates the field of structural biology. The use of protein structure information is currently widely spread within many areas of science and industry, among which are biotechnology and pharmaceutical industry.

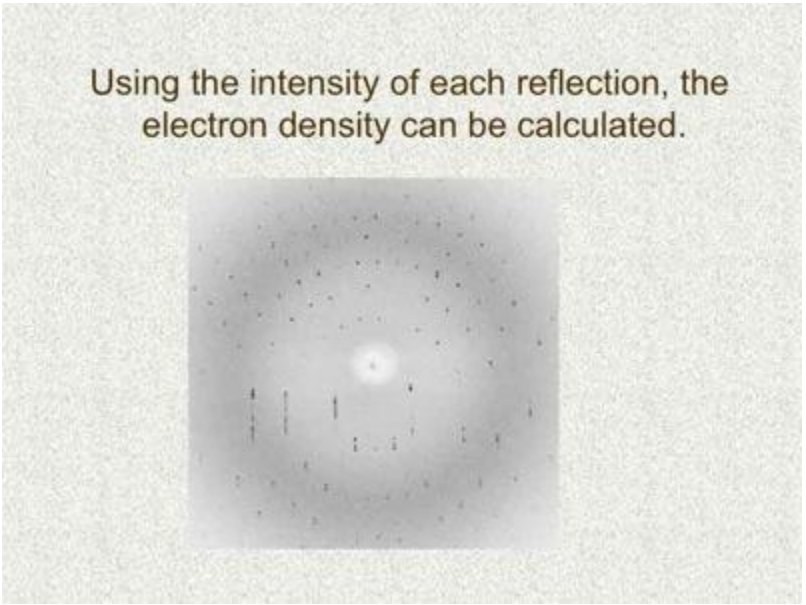


Figure 7

X-ray crystallography makes use of the diffraction pattern of X-rays that are shot through an

object. The pattern is determined by the *electron density* within the crystal. The diffraction is the result of an interaction with the high energy X-rays and the electrons in the atom. The electrons get activated and their relaxation to the initial energy state emits new X-rays. Bundles of such waves can be enhanced if they are in phase, and they get canceled out if they are out of phase. Therefore the diffraction of parallel X-rays from an object containing thousands of unit molecules arranged in a regular lattice results in the enhancement and cancellation of the diffracted waves and a resulting pattern of this vectorial process can be correlated with the distribution of the electrons in the crystal.

X-ray crystallography requires the growth of protein crystals up to 1 mm in size from a highly purified protein source. Crystal growth is an experimental technique and there exists no rules about the optimal conditions for a protein solution to result in a good protein crystal. The protocol has to be established for every new type of protein. Water soluble proteins are easier to crystallize than membrane proteins. The latter tend to precipitate out of solution due to unfavorable protein-protein and protein-solute interactions. To be kept soluble in aqueous solution, membrane proteins need the addition of detergents. The presence of detergents, however, often interferes with regular arrangements of the protein complexes in the crystal resulting in diffuse diffraction pattern. If membrane proteins contain large extra-membranous domains, these water soluble domains can be cleaved off from the membrane buried domain and crystallized individually.

X-rays have a wavelength of 0.2Å to 2.0Å. The wave length, as in an optical microscope, determines the resolution limit of half the applied wave length. X-rays are therefore suited for the atomic distances which reside in the angstrom range. X-rays are high energy electromagnetic radiation and can be recorded on X-ray sensitive film, the normal technique to record diffraction patterns of protein crystals.

X-rays that interact with an electron cause it to oscillate. Oscillating electrons serve as a new source of X-rays that propagate away from the stimulated electron. The waves of neighboring electrons super impose and depending on their being in-phase or out of phase result in a signal or in no signal at all. Diffraction by a crystal can be regarded as the reflection of the primary beam by sets of parallel planes that define the dimensions of the unit cell (the smallest repetitive pattern) of the crystal. The relationship between reflection angle, θ, the distance between the planes, d, and the wavelength, λ, is given by Bragg's law:

$2d\sin\theta = n\lambda$ Bragg's Law

The 2-dimensional distribution of the diffraction pattern can be calculated back into a 3-dimensional space of the electron distribution causing the diffraction. The mathematical formalism to do this is called *Fourier transformation*. The distances between the spots inversely correlates with the distances of the unit cell in the crystal and the intensity of the spots with the density of electrons in the molecular structure. The exact location of the electrons, however, is lost in a single diffraction pattern, because the information of the phase of the diffracted beams is not given. This is called the *phase problem* and is the hardest obstacle to overcome. The phase problem requires at least 3 different protein crystals with identical unit cell geometry and the inclusion of evenly spaced *heavy metals* or derivatives in the protein structure that give information about the relative phase in the individual crystal. The diffraction spots originating from the electron shell of the heavy metals can easily be identified and distinguished from other electron dens centers in the crystal. From the heavy metal location in the unit cell and the phase shift can be determined. The method to solve the phase problem using different crystals with identical protein structures containing regularly but infrequently spaced heavy metals or protein isoforms is known as *multiple isomorphous replacement*.

The amplitudes and phases of the diffraction data are used to calculate an *electron-density map*of the repeating unit of the crystal. This is a step that involves the *interpretation of the raw data*. This step is sensitive to the resolution of the diffraction data, which in turn is determined by the *quality of the protein crystal*, i.e., the regularity of the lattice of the protein in the unit cell and the regularity of the distribution of the heavy atom inclusions. The interpretation of the diffraction data needs information about the amino acid sequence of the protein because depending on the resolution of the data different amino acids can have indistinguishable electron densities (e.g. Tyr and Phe, or Leu and Ile).

Initial models of protein structures due to limits in the resolution have to be refined. This is often achieved by comparing the experimental data with the optimal structure obtained by computer modeling. The difference in experimental structure and hypothetical structure is given as *R-factor*.

Nuclear magnetic resonance, or NMR

Nuclear magnet resonance obtains the same high resolution using a very different strategy. NMR measures the distances between atomic nuclei, rather than the electron density in a molecule. With NMR, a *strong, high frequency magnetic field* stimulates atomic nuclei of the isotopes H-1, D-2, C-13, or N-15 (they have a magnetic spin) and measures the frequency of the magnetic field of the atomic nuclei during its oscillation period back to the initial state. The important step is to determine which resonance comes from which spin. The distance and type of neighboring nuclei determines the resonance frequency of the stimulated atomic nuclei. This dependence on next neighbors known as *chemical shift* (or spin-spin coupling constant) andreflects the local electronic environment and the information contained in *1-D NMR spectra*. For proteins, NMR usually measures the spin of protons. The following reasons make the H-1 NMR spectroscopy the method of choice for biological macromolecules:

- H are present at many sites in proteins, nucleic acids, and polysaccharides

- H have a high abundance for each site

- H nuclei is the most sensitive to detect

1-D spectra contain the information about all the chemical shifts of all the H in the protein. The frequency resolution is often not enough to distinguish individual chemical shifts. 2-D NMR solves this problems by containing information about the relative position of H in molecular structures. 2-D NMR spectra contain information about interaction between H that are covalently linked through one or two other atoms (COSY or *correlation spectroscopy*). Alternatively, pairs of H that can be close in space, even if they are from residues that are not close in sequence (NOE spectra, or *Nuclear Overhauser Effect*). A complete structure can thus be calculated by sequentially assigning cross peak correlations in 2-D spectra. Currently, the size limit for proteins amenable to NMR solution structure analysis is about 200 amino acids. An important feature of the identification of cross peaks is that regular patterns can be recognized that stem from secondary structure elements such as alpha helices and parallel or anti-parallel beta sheets because they contain typical hydrogen bonding networks.

NMR also requires the knowledge of the *amino acid sequence*, but the protein does not have to be in an ordered crystal, yet high concentrations of solubilized protein must be available (NMR

structures are therefor also called *solution structures*). In biopolymers, the primary structure (sequence) logically breaks up the molecule into groups of coupled spins normally one or two groups per residue. This is true not only for proteins, but also for nucleic acids and polysaccharides.

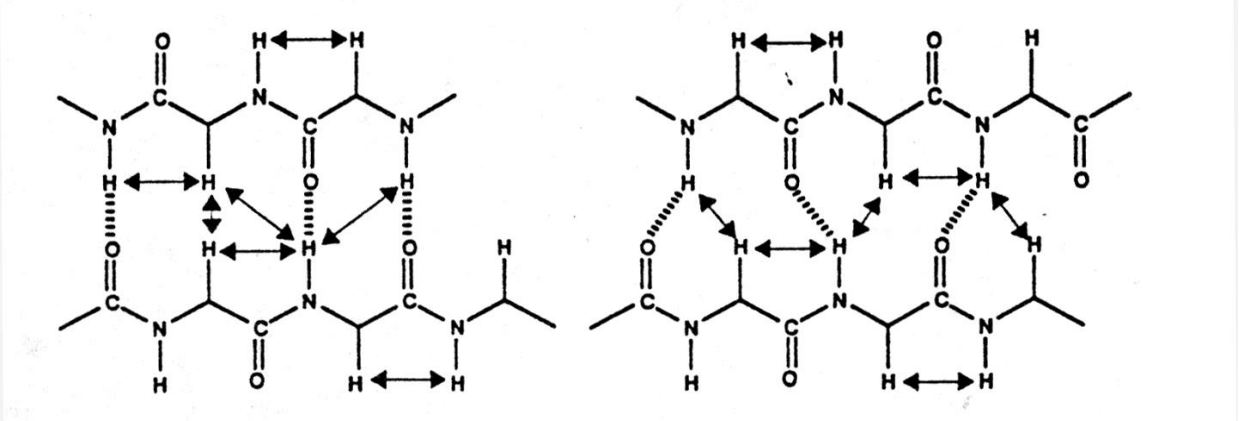


Figure 8 Observed NOEs in antiparallel and parallel b sheets

4. X-ray crystallography and NMR are complementary techniques

Table 1

NMR	X-ray crystallography
short time scale, protein folding	long time scale, static structure
solution, purity	single crystal, purity
< 20kD, domain	any size, domain, complex
functional active site	active or inactive
domains	domains
atomic nuclei, chemical bonds	electron density
resolution limit 2-3.5Å	resolution limit 2-3.5Å
primary structure must be known	primary structure must be know (except if resolution is 2Å or better for every single residue)

CIRCULAR DICHORISM SPECTROSCOPY

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light which arise due to structural asymmetry. The absence of regular structure results in zero CD intensity, while an ordered structure results in a spectrum which can contain both positive and negative signals.

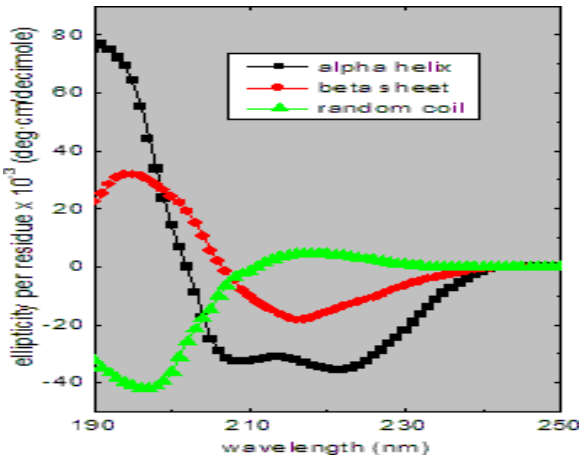


Figure 9

Circular dichroism spectroscopy is particularly good for:

- determining whether a protein is folded, and if so characterizing its secondary structure, tertiary structure, and the structural family to which it belongs
- comparing the structures of a protein obtained from different sources (*e.g.* species or expression systems) or comparing structures for different mutants of the same protein
- demonstrating comparability of solution conformation and/or thermal stability after changes in manufacturing processes or formulation
- studying the conformational stability of a protein under stress -- thermal stability, pH stability, and stability to denaturants -- and how this stability is altered by buffer composition or addition of stabilizers and excipients
  - CD is excellent for finding solvent conditions that increase the melting temperature and/or the reversibility of thermal unfolding, conditions which generally enhance shelf life
- determining whether protein-protein or protein-ligand interactions alter the conformation of protein.
  - If there are any conformational changes, this will result in a spectrum which will differ

from the sum of the individual components. Small conformational changes have been seen, for example, upon formation of several different receptor/ligand complexes.

Determination of Protein Secondary Structure by Circular Dichroism

Secondary structure can be determined by CD spectroscopy in the "far-UV" spectral region (190-250 nm). At these wavelengths the chromophore is the peptide bond, and the signal arises when it is located in a regular, folded environment.

Alpha-helix, beta-sheet, and random coil structures each give rise to a characteristic shape and magnitude of CD spectrum. This is illustrated by the graph to the right, which shows spectra for poly-lysine in these three different conformations. The approximate fraction of each secondary structure type that is present in any protein can thus be determined by analyzing its far-UV CD spectrum as a sum of fractional multiples of such reference spectra for each structural type.

Like all spectroscopic techniques, the CD signal reflects an average of the entire molecular population. Thus, while CD can determine that a protein contains about 50% alpha-helix, it cannot determine which specific residues are involved in the alpha-helical portion.

Far-UV CD spectra require 20 to 200 µl of solution containing 1 mg/ml to 50 µg/ml protein, in any buffer which does not have a high absorbance in this region of the spectrum. (High concentrations of DTT, histidine, or imidazole, for example, cannot be used in the far-UV region.) Note that for many formulated protein samples the absorbance due to the excipients prevents collecting spectra below 200 nm (and even 200 nm is often not possible). When that is true the accuracy reliability of secondary structure calculations (the actual percentages of different structures) is compromised, but the validity of spectral comparisons is not.

Information about Protein Tertiary Structure from Circular Dichroism:

The CD spectrum of a protein in the "near-UV" spectral region (250-350 nm) can be sensitive to certain aspects of tertiary structure. At these wavelengths the chromophores are the aromatic amino acids and disulfide bonds, and the CD signals they produce are sensitive to the overall tertiary structure of the protein.

Signals in the region from 250-270 nm are attributable to phenylalanine residues, signals from

270-290 nm are attributable to tyrosine, and those from 280-300 nm are attributable to tryptophan. Disulfide bonds give rise to broad weak signals throughout the near-UV spectrum.

If a protein retains secondary structure but no defined three-dimensional structure (*e.g.* an incorrectly folded or "molten-globule" structure), the signals in the near-UV region will be nearly zero. On the other hand, the presence of significant near-UV signals is a good indication that the protein is folded into a well-defined structure.

The near-UV CD spectrum can be sensitive to small changes in tertiary structure due to protein-protein interactions and/or changes in solvent conditions.

The signal strength in the near-UV CD region is much weaker than that in the far-UV CD region. Near-UV CD spectra require about 1 ml of protein solution with an OD at 280 nm of 0.5 to 1 (which corresponds to 0.25 to 2 mg/ml for most proteins).

Demonstrating Comparability of Conformation

Often it is necessary to demonstrate that different lots of a protein have equivalent conformations, for example after a scale-up in the purification process or to qualify a new manufacturing site, and CD can be a good tool for this.

The data below show a case where the far-UV spectra show that the recombinant form of an enzyme clearly does not have the same secondary structure as the natural protein (*i.e.* the recombinant protein is not properly folded).

Such cases of significant differences in secondary structure are, however, unusual. More typically subtle differences in conformation do not produce a detectable difference in far-UV CD, but may produce a difference in near-UV CD. One such example, for different lots of a monoclonal antibody, is shown below. This small but reproducible difference at ~240 nm correlates with differences in the stability of different lots of this antibody.

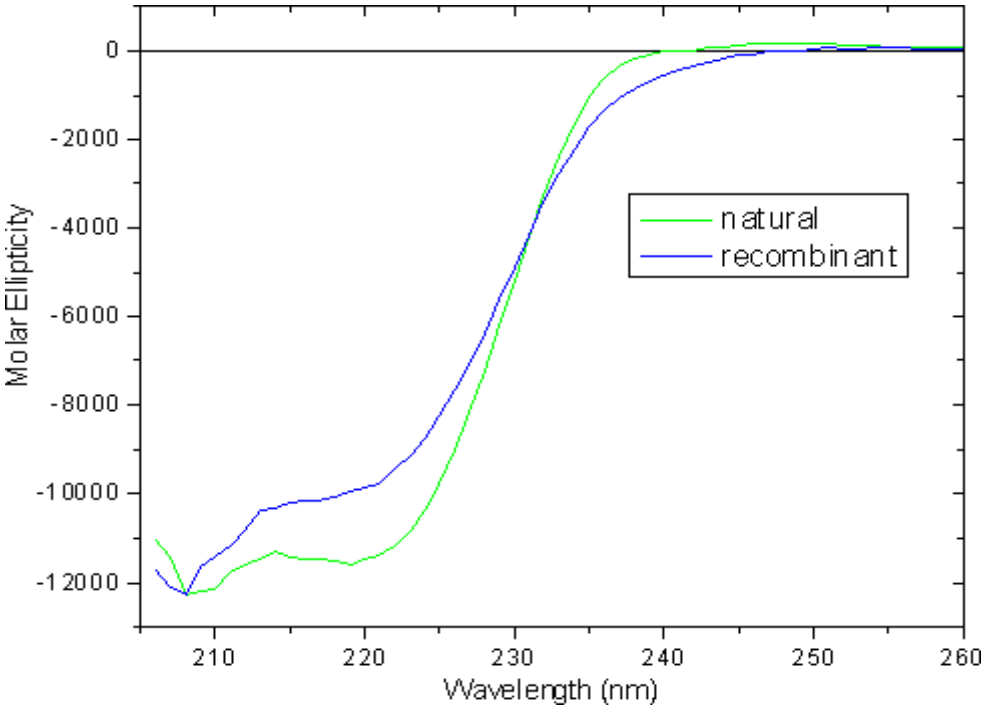


Figure 10

Thermal Stability by Circular Dichroism

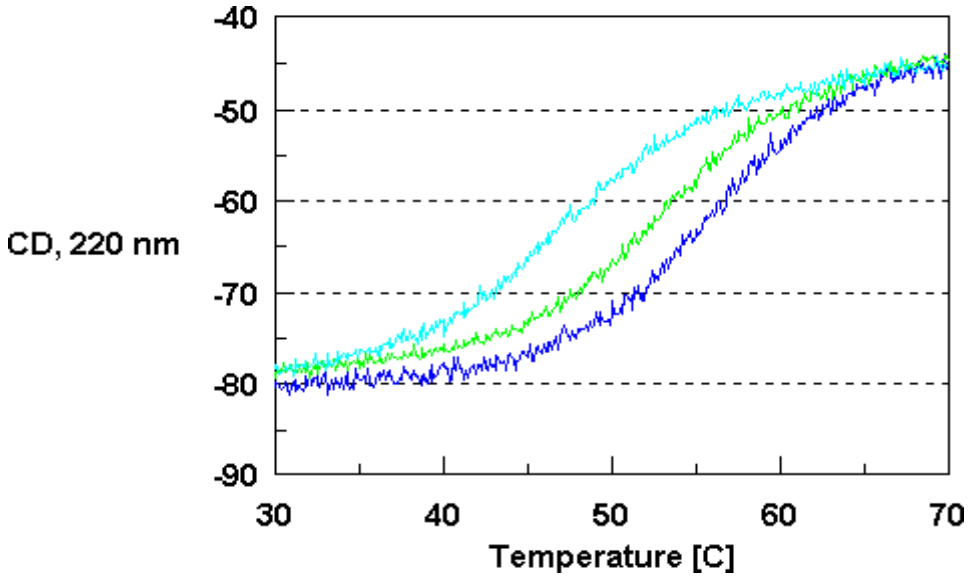


Figure 11

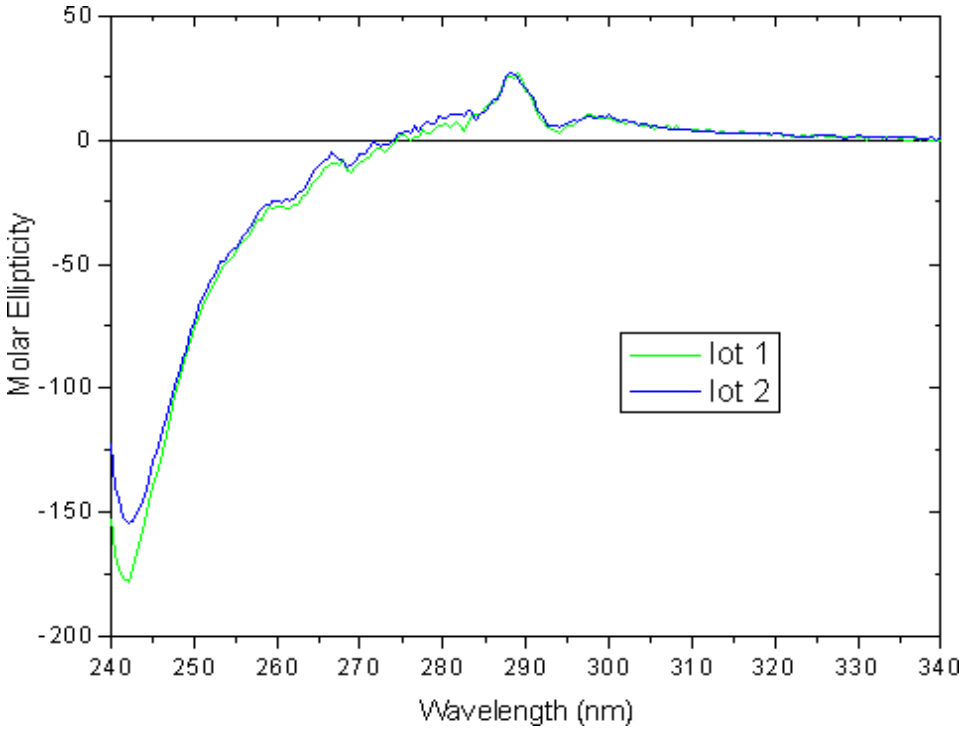


Figure 12

Thermal stability is assessed using CD by following changes in the spectrum with increasing temperature. In some cases the entire spectrum in the far- or near-UV CD region can be followed at a number of temperatures. Alternatively, a single wavelength can be chosen which monitors some specific feature of the protein structure, and the signal at that wavelength is then recorded continuously as the temperature is raised. CD is often used to assess the degree to which solution pH, buffers, and additives such as sugars, amino acids or salts alter the thermal stability.

This graph illustrates thermal scans done in our lab for the same recombinant protein in 3 different buffers. While unfolding is completely reversible under all these conditions, clearly there are quite significant differences in thermal stability.

Many proteins aggregate or precipitate quickly after they are unfolded ("melted"), making unfolding irreversible. The reversibility of the unfolding reaction can be assessed by cooling the sample and then heating again to see if the unfolding reaction is duplicated. Finding solvent conditions that make unfolding reversible may be actually be more important for long-term stability (shelf life) than raising the melting temperature.

If (and only if) the melting is fully reversible, the melting temperature is directly related to conformational stability, and the thermodynamics of protein folding can be extracted from the data. The fact that thermal unfolding can generally be measured by CD at much lower concentrations than by DSC increases the probability of reversible reactions and of thermodynamically interpretable data.

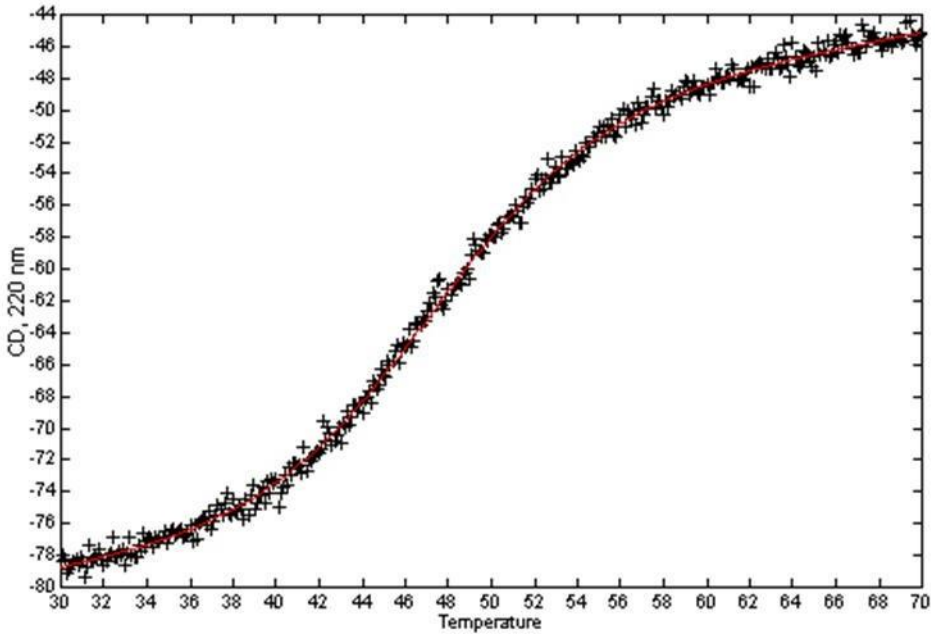


Figure 13

This graph illustrates a detailed analysis of one of the data sets shown above, using custom software developed in our lab. The data (+) were fitted to a simple thermodynamic unfolding model (solid line). The fit returns the melting temperature (midpoint of the transition) as 47.3 +/- 0.1 °C. The width of the transition region is related to the enthalpy of unfolding, □*H*, which the fit returns as 52 +/- 2 kcal/mol. Fitting the data also allows a more reproducible measurement of the onset of unfolding, a temperature which is often more relevant for formulation and shelf-life considerations than the midpoint. The onset (defined as the temperature at which 5% of the protein is unfolded) occurs at 36.1 +/- 0.3 °C in this case.

If the protein precipitates or aggregates as it is unfolded, the melting reaction will be irreversible, and the melting temperature will reflect the kinetics of aggregation and the

solubility of the unfolded form of the molecule as well as the intrinsic conformational stability.

The cooperativity of the unfolding reaction is measured qualitatively by the width and shape of the unfolding transition. A highly cooperative unfolding reaction indicates that the protein existed initially as a compact, well-folded structure, while a very gradual, non-cooperative melting reaction indicates that the protein existed initially as a very flexible, partially unfolded protein or as a heterogeneous population of folded structures.

Melting of Secondary Structure

Changes in secondary structure, monitored in the far-UV CD region, can be determined with as little as 50 µg of protein, at concentrations of 0.2 mg/ml. By following changes over the entire far-UV CD region we can determine whether at high temperatures the protein is losing all of its secondary structure, loses only a portion of its secondary structure, or simply undergoes conformational change involving a change in secondary structure. Occasionally the unfolded form of a protein will possess a defined but totally different secondary structure than the native form (*e.g.*, TNF-alpha contains beta-sheet when folded, but alpha-helix when melted, and many proteins form amyloid-like aggregates following a transition from alpha-helix to beta-strand).

Melting of Tertiary Structure

Changes in tertiary structure can be followed by monitoring changes in the near-UV CD region. Due to the weaker signal in this region this requires 1-3 mg of protein. Such studies will reveal whether the melting of a protein occurs in a single step (with concurrent loss of both secondary and tertiary structure), or in a two-step reaction.

Melting of Protein Complexes

The effect of forming a protein-protein complex (*e.g.* ligand/receptor or antigen-antibody) on the thermal stability of the individual proteins in the complex can also be determined. This works best if the individual proteins have CD spectra which are quite different from each other, such that changes at specific wavelengths can be monitored to follow changes in the corresponding protein. In such cases it is possible to determine whether there is an increase in stability of one or both of the proteins following complex formation.

Infrared spectroscopy of proteins

During the last years the use of Fourier Transform Infrared spectroscopy (FTIR) to determine the structure of biological macromolecules has dramatically expanded. The complete three-dimensional structure of a protein at high resolution can be determined by X-ray crystallography. This technique requires the molecule to form a well ordered crystal which is not possible for all proteins. An alternative to X-ray crystallography is multidimensional nuclear magnetic resonance (NMR) spectroscopy. Using NMR spectroscopy structures of the proteins can be determined in solution. The interpretation of the NMR spectra of large proteins is very complex, so its present application is limited to small proteins (~15-25 kDa). These limitations have led to the development of alternative methods that are not able to generate structures at atomic resolution but provide also structural information on proteins (especially on secondary structure). These methods include circular dichroism (CD) and vibrational (infrared and RAMAN) spectroscopy. The new technique of FTIR spectroscopy requires only small amounts of proteins (1mM) in a variety of environments. Therefore, high quality spectra can be obtained relatively easy without problems of background fluorescence, light scattering and problems related to the size of the proteins. The omnipresent water absorption can be subtracted by mathematical approaches. Methods are now available that can separate subcomponents that overlap in the spectra of proteins. These facts have made practical biological systems amenable to studies by FTIR spectroscopy.

Basic principles of infrared (IR) absorption

IR spectroscopy is the measurement of the wavelength and intensity of the absorption of infrared light by a sample. Infrared light is energetic enough to excite molecular vibrations to higher energy levels.

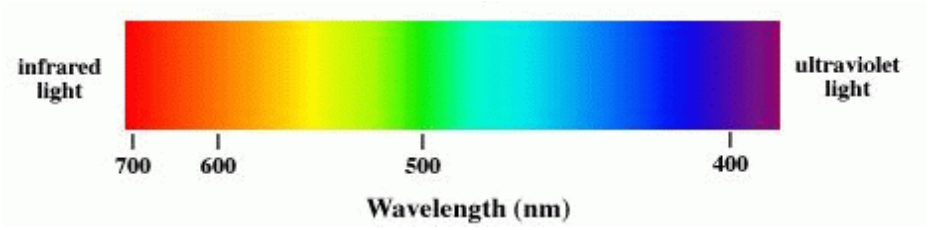


Figure 14

Table 2 Electromagnetic spectrum

frequencyrange (Hz)	wavelength range	type of radiation	type of transition
10^{20} - 10^{24}	10^{-12} - 10^{-16} m	gamma rays	nuclear
10^{17} - 10^{20}	1 nm - 1 pm	x-rays	inner electrons
10^{15} - 10^{17}	400 - 1 nm	ultraviolet light	outer electrons
$4.3x10^{14}$ - $7.5x10^{14}$	700 - 400 nm	visible light	outer electrons
10^{12}-10^{14}	2.5 um - 700 nm	infrared light	vibrations
10^{8} - 10^{12}	1 mm - 2.5 um	microwaves	rotations
10^{0} - 10^{8}	10^{8} - 1 m	radio waves	spin flips

The infrared spectra usually have sharp features that are characteristic of specific types of molecular vibrations, making the spectra useful for sample identification.

Table 3 characteristic IR bands

X-H vibrations	***bond***	***wavenumbers (cm^{-1})***
hydroxyl	O-H	3610-3640
amines	N-H	3300-3500
aromatic rings	C-H	3000-3100
alkenes	C-H	3020-3080
alkanes	C-H	2850-2960
triple bonds		2500-1900
double bonds		1900-1500
deformation/heavy atoms		1500-

For a molecule of N atoms, 3N-6 fundamental vibrations (or normal modes) exist (3N-5 if the molecule is linear). Therefore, for the linear CO2 molecule 4 normal modes have to be expected.

Table 4 Normal modes *for CO_2*

		cm^{-1}	*IR*	*RAMAN*
stretching (sym.)	-> <- O==C==O	1340	-	+
stretching (asym.)	-> <- <- O==C==O	2349	+	-
deformation	/| /| O==C==O \\|	667	+	-
deformation	+ - + O==C==C	667	+	-

Fourier Transform Infrared (FTIR) spectroscopy

To use the Fourier Transform Infrared Spectroscopy, a continuum source of light (such as a Nernst Globar) is used to produce light over a broad range of infrared wavelengths. Light coming from this continuum source is split into two paths using a half-silvered mirror; this light is then reflected from two mirrors back onto the beamsplitter, where it is recombined. One of these mirrors is fixed, and the second is movable. If the distance from the beamsplitter to the fixed mirror is not exactly the same as the distance from the beamsplitter to the second mirror, then when the two beams are recombined, there will be a small difference in the phase of the light between these two paths. Because of the "superposition principle" constructive and destructive interference exist for different wavelengths depending of the relative distances of the two mirrors from the beamsplitter.

It can be shown that if the intensity of light is measured and plotted as a function of the position of the movable mirror, the resultant graph is the Fourier Transform of the intensity of light as a function of wavenumber. In FTIR spectroscopy, the light is directed onto the sample of interest, and the intensity is measured using an infrared detector. The intensity of light striking the detector is measured as a function of the mirror position, and this is then Fourier-transformed to produce a plot of intensity vs. wavenumber. As radiation source a Michelson Interferometer is used (see the drawing below).

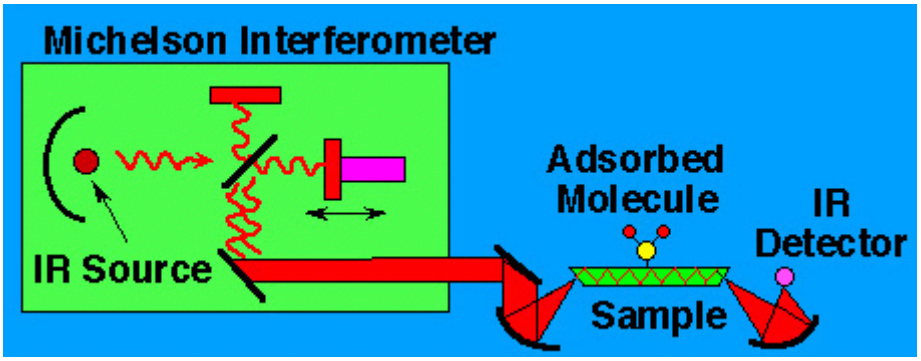


Figure 15

It is necessary to increase the sensitivity somehow, because the absorption due to one monolayer of molecules typically results in a change in intensity of only about one part in 10^5. For semiconductors, one way of increasing the sensitivity is to use multiple internal reflection. In this technique, the edges of the sample are polished, and the light is sent in at an angle. The light bounces around inside the sample, making about 30-50 bounces. This increases the sensitivity by about a factor of 30-50, making it possible to measure the absorption of less than one monolayer of molecules on a surface.

Band assignments Amide vibrations

The peptide group, the structural repeat unit of proteins, gives up to 9 characteristic bands named amide A, B, I, II ... VII. The amide A band (about 3500 cm^{-1}) and amide B (about 3100 cm^{-1}) originate from a Fermi resonance between the first overtone of amide II and and the N-H stretching vibration. Amide I and amide II bands are two major bands of the protein infrared

spectrum. The amide I band (between 1600 and 1700 cm^{-1}) is mainly associated with the C=O stretching vibration (70-85%) and is directly related to the backbone conformation. Amide II results from the N-H bending vibration (40-60%) and from the C-N stretching vibration (18-40%). This band is conformationally sensitive. Amide III and IV are very complex bands resulting from a mixture of several coordinate displacements. The out-of- plane motions are found in amide V, VI and VIII.

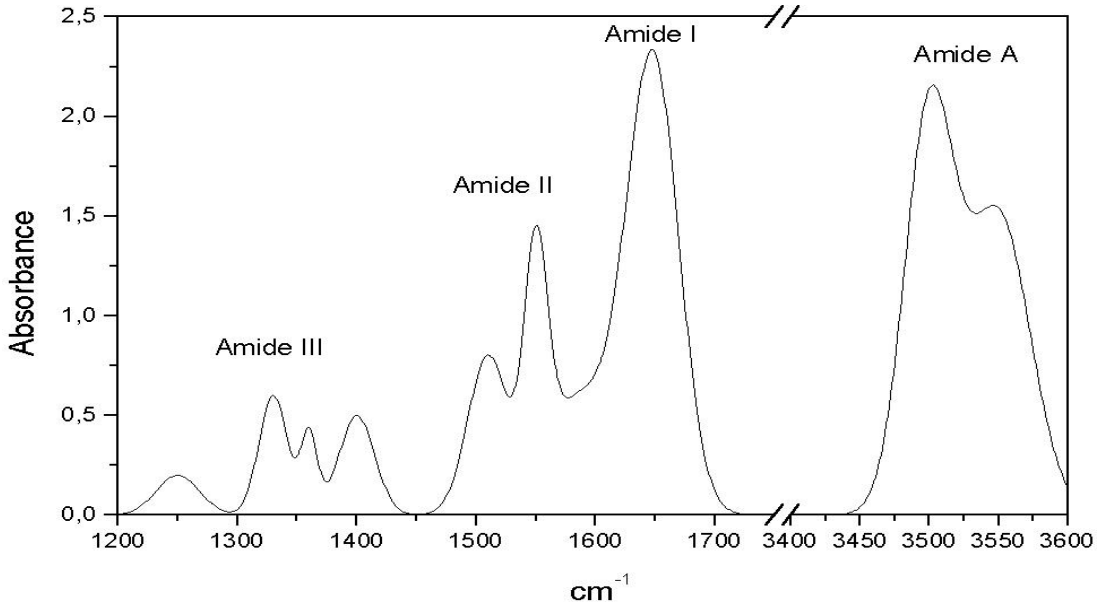


Figure 16

Amide A is with more than 95% due to the N-H stretching vibration. This mode of vibration does not depend on the backbone conformation but is very sensitive to the strength of a hydrogen bond. It has wavenumbers between 3225 and 3280 cm^{-1} for hydrogen bond lengths between 2.69 to 2.85 Å, ***Amide I*** is the most intense absorption band in proteins. It is primarily governed by the stretching vibrations of the C=O (70-85%) and C-N groups (10-20%). Its frequency is found in the range between 1600 and 1700 cm^{-1}. The exact band position is determined by the backbone conformation and the hydrogen bonding pattern. ***Amide II*** is found in the 1510 and 1580 cm^{-1} region and it is more complex than amide I. Amide II derives mainly from in-plane N-H bending (40-60% of the potential energy). The rest of the potential energy arises from the C-N (18-40%) and the C-C (about 10%) stretching vibrations. ***Amide III, V*** are very complex bands dependent on the details of the force field, the nature of side chains and hydrogen bonding. Therefore these bands are only of limited use for the extraction

of structural information.

Amino acid side chain vibrations

The presence of bands arising from amino acid side chains must be recognized before attempting to extract structural information from the shapes of amide I and amide II bands. The contribution of the side chain vibrations in the region between 1800 and 1400 cm^{-1} (amide I and amide II region) has been thor. Among the 20 proteinogenous amino acids, only 9 (Asp, Asn, Glu, Gln, Lys, Arg, Tyr, Phe, His) show a significant absorbance in the region discussed above. The contribution of the different amino acid side chains were fitted by a sum of Gaussian and Lorentzian components.

Table 5

S	**vibration**		**cm^{-1}**	**A_0 (l/mol/cm)**	**FWHH (cm^{-1})**	**surface ($x10^{-4}$ l/mol/cm)**
Asp	-COO st as	pH>pK (~4.5)	1574	380	44	5.5
	-COOH st	pH<pK (~4.5)	1716	280	50	4.1
Glu	-COO st as	pH>pK (~4.4)	1560	470	48	7.1
	-COOH st	pH<pK (~4.4)	1712	220	56	3.6
Arg	$-CN_3H_5^+$ st as		1673	420	40	4.3
	st s		1633	300	40	3.6
Lys	$-NH_3^+$ bd as		1629	130	46	1.8
	bd s		1526	100	48	1.3
Asn	-C=O st		1678	310	32	2.7
	$-NH_2$ bd		1622	160	44	2.5
Gln	-C=O st		1670	360	32	3.1
	$-NH_2$ bd		1610	220	44	3.5
Tyr	ring-OH	pH<pK (~10)	1518	430	8	1.0
	ring-O	pH>pK (~10)	1602	160	14	0.7

			1498	700	10	2.5
His	ring		1596	70	14	0.3
Phe	ring		1494	80	6	0.2
terminal						
	-COO st as		1598	240	47	3.5
	-COOH st		1740	170	50	2.1
	$-NH_3^+$ bd as		1631	210	54	3.8
	bd s		1515	200	60	4.3
	$-NH_2$ bd		1560	450	46	7.5

frequency, absorbance at the maximum (Ao), full width at half height (FWHH), surface of Gaussian band st=stretching vibration bd=bending s=symetrical as=asymetrical

Secondary structure of peptide model compounds

A large number of synthetic polypeptides has been used for the characterization of infrared spectra for proteins with a defined secondary structure content. For example, polylysine may adopt both beta-sheet or alpha-helical structures in dependence on temperature and pH of the solution. Experimental and theoretical work on a large number of synthetic polypeptides has provided insights into the variability of the frequencies for particular secondary structure conformations

Beta sheet

The frequencies of the main absorption bands from synthetic polypeptides adopting an antiparallel chain structure have been compiled by Chirgadze &Nevskaya . From these data it this follows, that the amide I absorption is primarily determined by the backbone conformation and independent of the amino acid sequence, its hydrophilic or hydrophobic properties and charge. The average frequency of the main component is about 1629 cm^{-1} with a minimum of 1615 cm^{-1} and a maximum of 1637 cm^{-1}. The average value for the second frequency is 1696 cm^{-1} (lowest value 1685 cm^{-1}). The parallel beta sheet structure that is not common in synthetic polypeptides leads to an amide I absorption near 1640 cm^{-1}

Helical structures

The alpha-helix: For alpha-helical structures the mean frequency was found to be 1652 cm-1 for the amide I and 1548 cm^{-1} for the amid II absorptions. The half width of the alpha-helix band depends on the stability of the helix. For the most stable helices, the half-width of about 15 cm^{-1} corresponds to a helix-coil transition free energy of more than 300 cal/mole. Other helices display half-widths of 38 cm^{-1} and helix-coil transition free energies of about 90 cal/mole.

The 3_{10}-helix differs from the alpha-helix in that the internal hydrogen bonding occurs between residues i and i+3 instead of i and i+4 in alpha helices.

Turn structures

The beta turn structure involves 4 amino acid residues which form a loop so that the two chain segments separated by the turn adopt an antiparallel orientation and form an i to i+3 hydrogen bond. A number of turn structures have been identified from protein structures: type I (42%, non-helical), type II (15%, non-helical, requires Gly in position 3) and type III (18%, corresponds to one turn of 3_{10} helix). Assignment of beta turns by means of a normal mode analysis for insulin demonstrates a strong overlapping of the different types of beta turns with the alpha-helical absorption. However, an absorption near 1680 cm^{-1} is now clearly assigned to beta turns.

Secondary structure in proteins

The shape of the amide I band of globular proteins is characteristic of their secondary structure. With a publication by Byler & Susi the determination of secondary structures in proteins from FTIR spectra actually started. This had become possible by the availability of high signal-to-noise ratio digitalised spectra obtained by the FTIR spectrometer and by the access to computers and software able to perform many operations on the spectra in a short time.

Deconvolution of the amide I band

The concept of Fourier self deconvolution is based on the assumption, that a spectrum of single bands (each narrow band is characteristic for a secondary structure) is broadened in the liquid or solid state. Therefore, the bands overlap and cannot be distinguished in the amide envelope. A curve fitting procedure can be applied to estimate quantitatively the area of each component representing a type of secondary structure. In the pioneering work by Susi & Byler

the amide I was deconvoluted with a Lorentzian line shape function and a resolution enhancement factor of 2.4 was applied. The deconvoluted spectrum was fitted with Gaussian band shapes by an iterative curve fitting procedure. The results are in good agreement with with the secondary structure information obtained from X-ray crystallographic structures of the proteins under study.

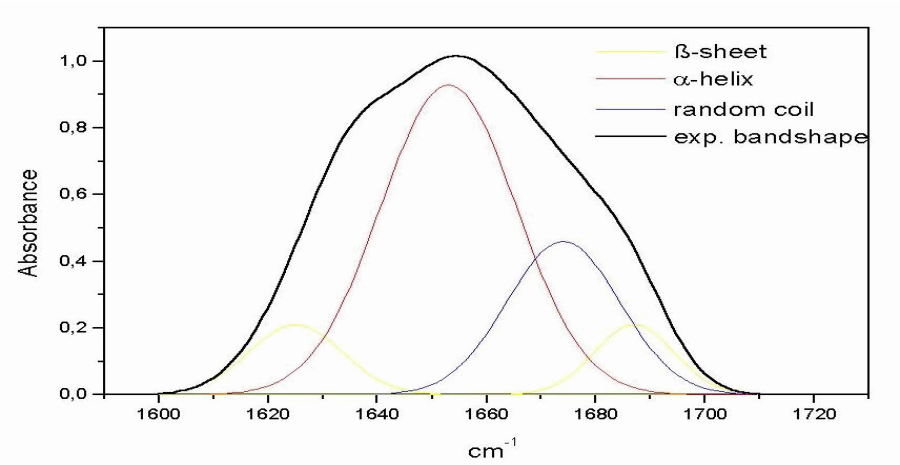


Figure 17

Table 6

	a)			***b)***			
sec. structure	**Mean (cm^{-1})**	**RMS (cm^{-1})**	**Max (cm^{-1})**	**Mean (cm^{-1})**	**RMS (cm^{-1})**	**Max (cm^{-1})**	**Region (cm^{-1})**
turns	1694	1.7	2	-	-	-	
	1688	1.1	2	-	-	-	
	1683	1.5	2	1678	2.1	5	**1682-1662**
	1670	1.4	2	1670	2.9	5	
	1663	2.2	4	1664	1.0	3	
alpha-helix	1654	1.5	3	1656	1.5	3	
				1648	1.6	3	**1662-1645**
unordered	1645	1.6	4	1641	2.0	3	**1645-1637**
beta sheet	1624	2.4	4	1624	2.5	5	

	1631	2.5	3	1633	2.1	4	**1637-1613**
	1637	1.4	3	-	-	-	
	1675	2.6	4	1685	2.1	4	**1689-1682**

MASS SPECTROMETRY IN PROTEIN IDENTIFICATION AND SEQUENCE ANALYSIS

Mass spectrometry is an analytical tool used for measuring the **molecular mass** of a sample. For large samples such as **biomolecules,** molecular masses can be measured to within an accuracy of **0.01%** of the total molecular mass of the sample *i.e.* within a 4 Daltons (Da) or atomic mass units (amu) error for a sample of 40,000 Da. This is sufficient to allow minor mass changes to be detected, *e.g.* the substitution of one amino acid for another, or a post-translational modification.

For small **organic molecules** the molecular mass can be measured to within an accuracy of **5 ppm** or less, which is often sufficient to confirm the molecular formula of a compound, and is also a standard requirement for publication in a chemical journal.

Structural information can be generated using certain types of mass spectrometers, usually those with multiple analysers which are known as **tandem mass spectrometers.** This is achieved by fragmenting the sample inside the instrument and analysing the products generated. This procedure is useful for the structural elucidation of **organic compounds** and for **peptide** or **oligonucleotide** sequencing.

Where are mass spectrometers used?

Mass spectrometers are used in industry and academia for both routine and research purposes. The following list is just a brief summary of the major mass spectrometric applications:

- **Biotechnology:** *the analysis of proteins, peptides, oligonucleotides*
- **Pharmaceutical:** *drug discovery, combinatorial chemistry, pharmacokinetics, drug metabolism*
- **Clinical:** *neonatal screening, haemoglobin analysis, drug testing*
- **Environmental:** *PAHs, PCBs, water quality, food contamination*

- **Geological:** *oil composition*

2. How can mass spectrometry help biochemists?

- **Accurate molecular weight measurements:**

sample confirmation, to determine the purity of a sample, to verify amino acid substitutions, to detect post-translational modifications, to calculate the number of disulphide bridges

- **Reaction monitoring:**

to monitor enzyme reactions, chemical modification, protein digestion

- **Amino acid sequencing:**

sequence confirmation, de novo *characterisation of peptides, identification of proteins by database searching with a sequence "tag" from a proteolytic fragment*

- **Oligonucleotide sequencing:**

characterisation or quality control of oligonucleotides

- **Protein structure:**

protein folding monitored by H/D exchange, protein-ligand complex formation under physiological conditions, macromolecular structure determination

How does a mass spectrometer work?

Introduction

Mass spectrometers can be divided into three fundamental parts, namely the **ionisation source,** the **analyser**, and the **detector.**

The sample has to be introduced into the ionisation source of the instrument. Once inside the ionisation source, the sample molecules are ionised, because ions are easier to manipulate than neutral molecules. These ions are extracted into the analyser region of the mass spectrometer where they are separated according to their **mass (m) -to-charge (z) ratios (m/z)**. The separated ions are detected and this signal sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of a **m/z spectrum**.

The analyser and detector of the mass spectrometer, and often the ionisation source too, are maintained under high vacuum to give the ions a reasonable chance of travelling from one end of the instrument to the other without any hindrance from air molecules. The entire operation of the mass spectrometer, and often the sample introduction process also, is under complete **data system** control on modern mass spectrometers.

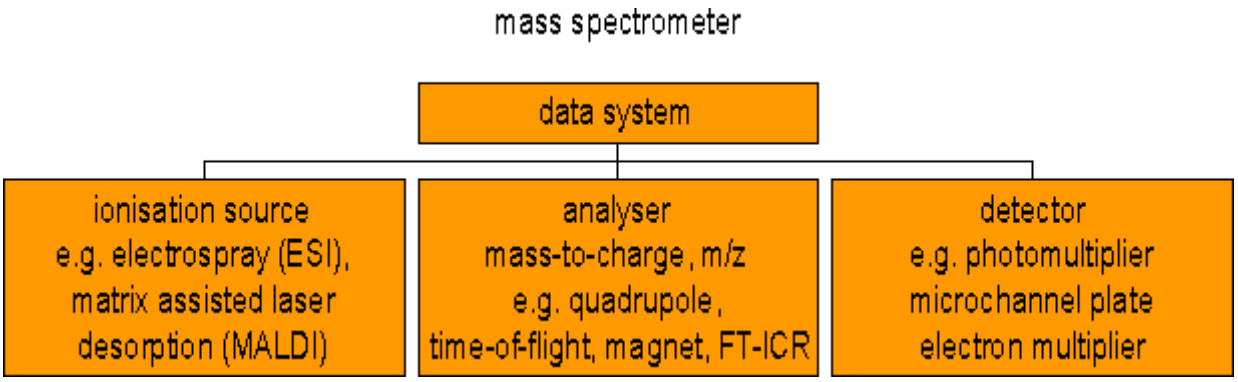


Figure 18 Simplified schematic diagram of a mass spectrometer

Sample introduction

The method of sample introduction to the ionisation source often depends on the ionisation method being used, as well as the type and complexity of the sample.

The sample can be inserted directly into the ionisation source, or can undergo some type of chromatography *en route* to the ionisation source. This latter method of sample introduction usually involves the mass spectrometer being coupled directly to a high pressure liquid chromatography (HPLC), gas chromatography (GC) or capillary electrophoresis (CE) separation column, and hence the sample is separated into a series of components which then enter the mass spectrometer sequentially for individual analysis.

Methods of sample ionisation

Many ionisation methods are available and each has its own advantages and disadvantages (**"Ionization Methods in Organic Mass Spectrometry",** Alison E. Ashcroft, The Royal Society of Chemistry, UK, 1997; and references cited therein).

The ionisation method to be used should depend on the type of sample under investigation and the mass spectrometer available.

Ionisation methods include the following:

Atmospheric Pressure Chemical Ionisation (APCI) Chemical Ionisation (CI)

Electron Impact (EI) **Electrospray Ionisation (ESI)** Fast Atom Bombardment (FAB)

Field Desorption / Field Ionisation (FD/FI)

Matrix Assisted Laser Desorption Ionisation (MALDI)

Thermospray Ionisation (TSP)

The ionisation methods used for the majority of biochemical analyses are **Electrospray Ionisation (ESI)** and **Matrix Assisted Laser Desorption Ionisation (MALDI)**.

With most ionisation methods there is the possibility of creating both positively and negatively charged sample ions, depending on the proton affinity of the sample. Before embarking on an analysis, the user must decide whether to detect the positively or negatively charged ions.

Analysis and Separation of Sample Ions

The main function of the **mass analyser** is to **separate**, or **resolve**, the ions formed in the ionisation source of the mass spectrometer according to their **mass-to-charge (m/z)** ratios. There are a number of mass analysers currently available, the better known of which include **quadrupoles**, **time-of-flight (TOF)** analysers, **magnetic sectors**, and both **Fourier transform** and **quadrupole ion traps** .

These mass analysers have different features, including the m/z range that can be covered, the mass accuracy, and the achievable resolution. The compatibility of different analysers with different ionisation methods varies. For example, all of the analysers listed above can be used in conjunction with electrospray ionisation, whereas MALDI is not usually coupled to a quadrupole analyser.

Tandem (MS-MS) mass spectrometers are instruments that have more than one analyser and so can be used for structural and sequencing studies. Two, three and four analysers have all been incorporated into commercially available tandem instruments, and the analysers do not necessarily have to be of the same type, in which case the instrument is a **hybrid** one.
More popular tandem mass spectrometers include those of the **quadrupole-quadrupole, magnetic sector-quadrupole**, and more recently, the **quadrupole-time-of-flight** geometries.

Detection and recording of sample ions.

The **detector** monitors the ion current, amplifies it and the signal is then transmitted to the data system where it is recorded in the form of **mass spectra**. The **m/z** values of the ions are plotted against their **intensities** to show the **number of components** in the sample, the **molecular mass** of each component, and the **relative abundance** of the various components in the sample.

The type of detector is supplied to suit the type of analyser; the more common ones are the **photomultiplier** , the **electron multiplier** and the **micro-channel plate** detectors.

Electrospray ionisation

Electrospray Ionisation (ESI) is one of the **Atmospheric Pressure Ionisation (API)** techniques and is well-suited to the analysis of polar molecules ranging from less than 100 Da to more than 1,000,000 Da in molecular mass.

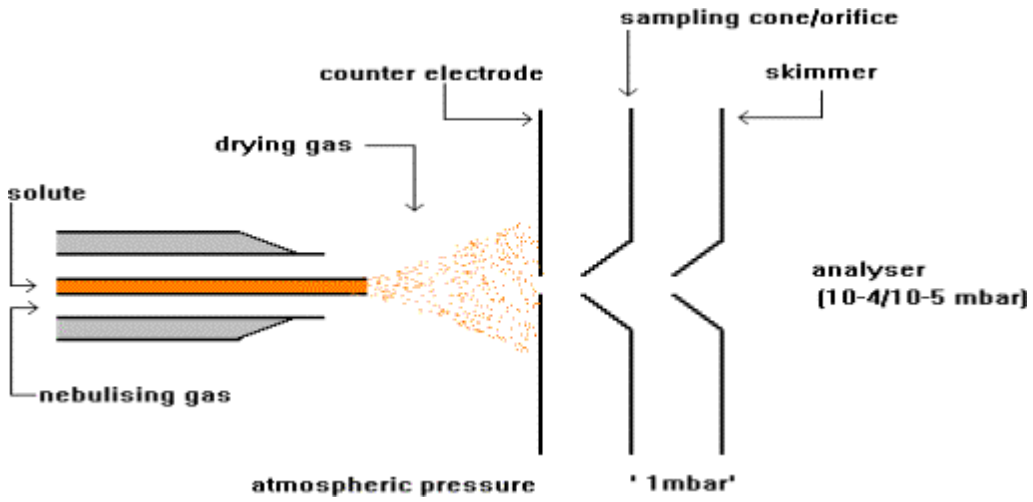


Figure 19

Standard electrospray ionisation source (Platform II)

During standard electrospray ionization, the sample is dissolved in a polar, volatile solvent and pumped through a narrow, **stainless steel capillary** (75 - 150 micrometersi.d.) at a flow rate of 1 mL/min. A **high voltage** of 3 or 4 kV is applied to the tip of the capillary, which is situated within the ionisation source of the mass spectrometer, and as a consequence of this strong electric field, the sample emerging from the tip is dispersed into an **aerosol of highly charged droplets**, a process that is aided by a co-axially introduced **nebulising gas** flowing around the outside of the capillary. This gas, usually nitrogen, helps to direct the spray emerging from the capillary tip towards the mass spectrometer. The charged droplets diminish in size by **solvent evaporation**, assisted by a warm flow of nitrogen known as the **drying gas** which passes across the front of the ionisation source. Eventually charged **sample ions**, free from solvent, are released from the droplets, some of which pass through a **sampling cone** or

orifice into an **intermediate vacuum** region, and from there through a small aperture into the analyser of the mass spectrometer, which is held under **high vacuum**. The lens voltages are optimised individually for each sample.

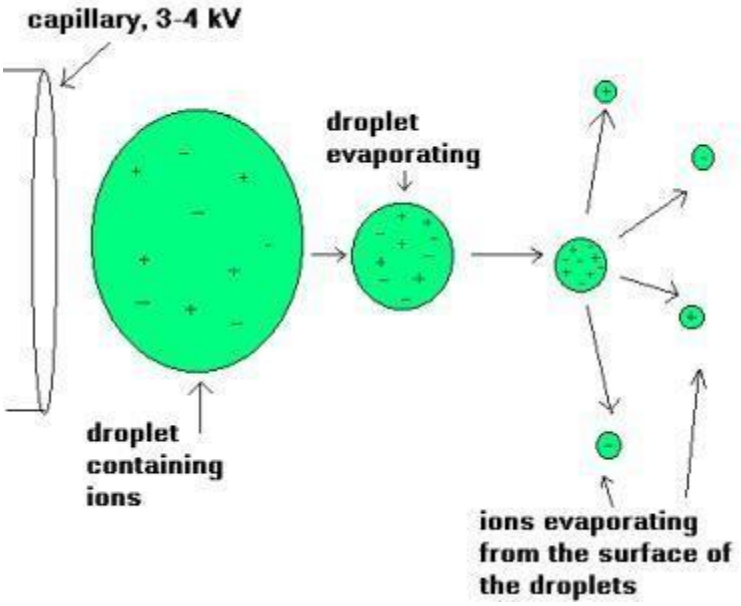


Figure 20 The electrospray ionisation process

Nanospray ionisation is a **low flow rate** version of electrospray ionisation. A **small volume** (1-4 microL) of the sample dissolved in a suitable volatile solvent, at a concentration of ca. **1 - 10 pmol/microL**, is transferred into a miniature **sample vial**. A reasonably **high voltage** (ca. 700 - 2000 V) is applied to the specially manufactured gold-plated vial resulting in **sample ionisation** and spraying. The flow rate of solute and solvent using this procedure is very low, **30 - 1000 nL/min**, and so not only is far less sample consumed than with the standard electrospray ionisation technique, but also a small volume of sample lasts for several minutes, thus enabling **multiple experiments** to be performed. A common application of this technique is for a **protein digest** mixture to be analysed to generate a list of **molecular masses** for the components present, and then each component to be analysed further by **tandem mass spectrometric (MS-MS) amino acid sequencing** techniques.

ESI and **nanospray ionisation** are very sensitive analytical techniques but the sensitivity deteriorates with the presence of non-volatile buffers and other additives, which should be avoided as far as possible.

In **positive ionisation** mode, a trace of formic acid is often added to aid protonation of the

sample molecules; in **negative ionisation** mode a trace of ammonia solution or a volatile amine is added to aid deprotonation of the sample molecules. **Proteins and peptides** are usually analysed under **positive ionisation** conditions and **saccharides and oligonucleotides** under **negative ionisation** conditions. In all cases, the **m/z** scale must be **calibrated** by analysing a standard sample of a similar type to the sample being analysed (e.g. a protein calibrant for a protein sample), and then applying a mass correction.

Data processing

ESI and **nanospray ionisation** generate the same type of spectral data for samples, and so the data processing procedures are identical. In ESI, samples (M) with **molecular masses up to ca. 1200 Da** give rise to **singly charged molecular-related ions**, usually **protonated molecular ions** of the formula $(M+H)^+$ in **positive ionisation mode**, and **deprotonated molecular ions** of the formula $(M-H)^-$ in **negative ionisation** mode.

An example of this type of sample analysis is shown in the m/z spectrum of the pentapeptide leucineenkephalin, YGGFL. The molecular formula for this compound is $C_{28}H_{37}N_5O_7$ and the calculated monoisotopic molecular weight is 555.2692 Da.

The m/z spectrum shows dominant ions at m/z 556.1, which are consistent with the expected protonated molecular ions, $(M+H^+)$. Protonated molecular ions are expected because the sample was analysed under positive ionisation conditions. These m/z ions are **singly charged**, and so the m/z value is consistent with the molecular mass, as the value of z (number of charges) equals 1. Hence the measured molecular weight is deduced to be 555.1 Da, in good agreement with the theoretical value.

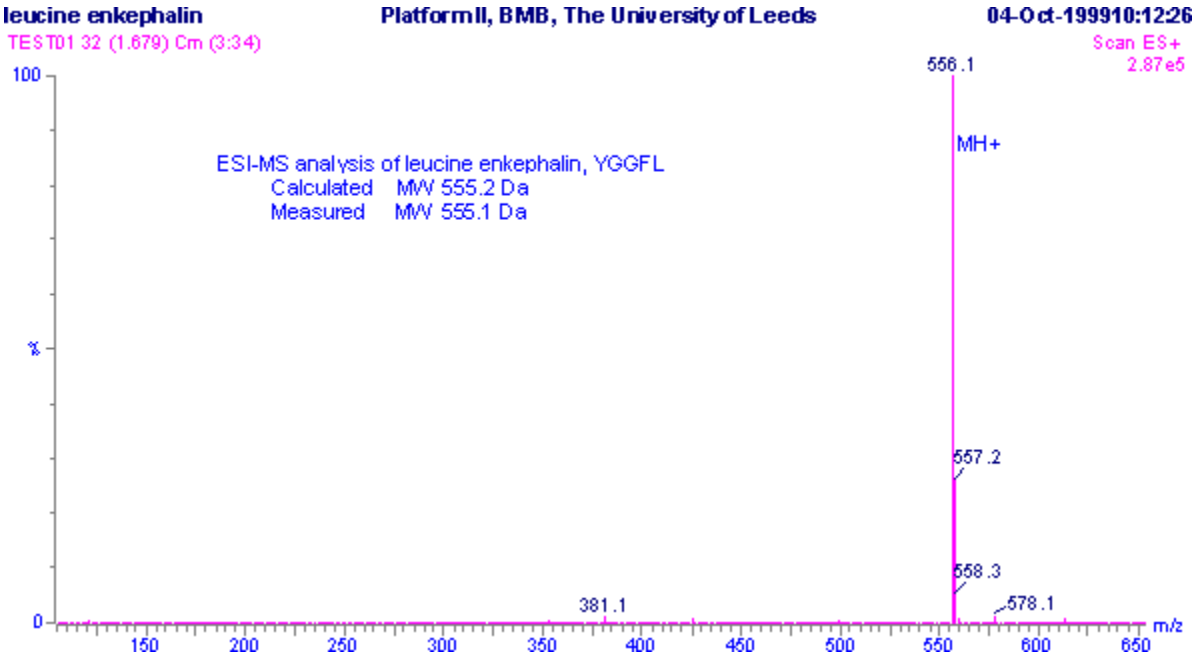


Figure 21 Positive ESI-MS m/z spectrum of leucineenkaphalin, YGGF

The m/z spectrum also shows other ions of lower intensity (ca. 25 % of the m/z 556.1 ions) at m/z 557.2. These represent the molecule in which one ^{12}C atom has been replaced by a ^{13}C atom, because carbon has a naturally occurring isotope one atomic mass unit (Da) higher. The intensity of these isotopic ions relates to the relative abundance of the naturally occurring isotope multiplied by the total number of carbon atoms in the molecule. Additionally the fact that the ^{13}C ions are one Da higher on the m/z scale than the ^{12}C ions is an indication that z = 1, and hence the sample ions are singly charged. If the sample ions had been doubly charged, then the m/z values would only differ by 0.5 Da as z, the number of charges, would then be equal to 2.

The m/z spectrum also contains ions at m/z 578.1, some 23 Da higher than the expected molecular mass. These can be identified as the sodium adduct ions, $(M+Na)^+$, and are quite common in electrospray ionisation. Instead of the sample molecules being ionised by the addition of a proton H^+, some molecules have been ionised by the addition of a sodium cation Na^+. Other common adduct ions include K^+ (+39) and NH_4^+ (+18) in positive ionisation mode and Cl^- (+35) in negative ionisation mode.

Electrospray ionisation is known as a "soft" ionisation method as the sample is ionised by the addition or removal of a proton, with very little extra energy remaining to cause fragmentation

of the sample ions.

Samples (M) with molecular weights greater than ca. 1200 Da give rise to multiply charged molecular-related ions such as **$(M+nH)^{n+}$ in positive ionisation mode** and **$(M-nH)^{n-}$ in negative ionisation mode**. Proteins have many suitable sites for protonation as all of the backbone amide nitrogen atoms could be protonated theoretically, as well as certain amino acid side chains such as lysine and arginine which contain primary amine functionalities.

An example of multiple charging, which is practically unique to electrospray ionisation, is presented in the positive ionisation m/z spectrum of the protein hen egg white lysozyme.

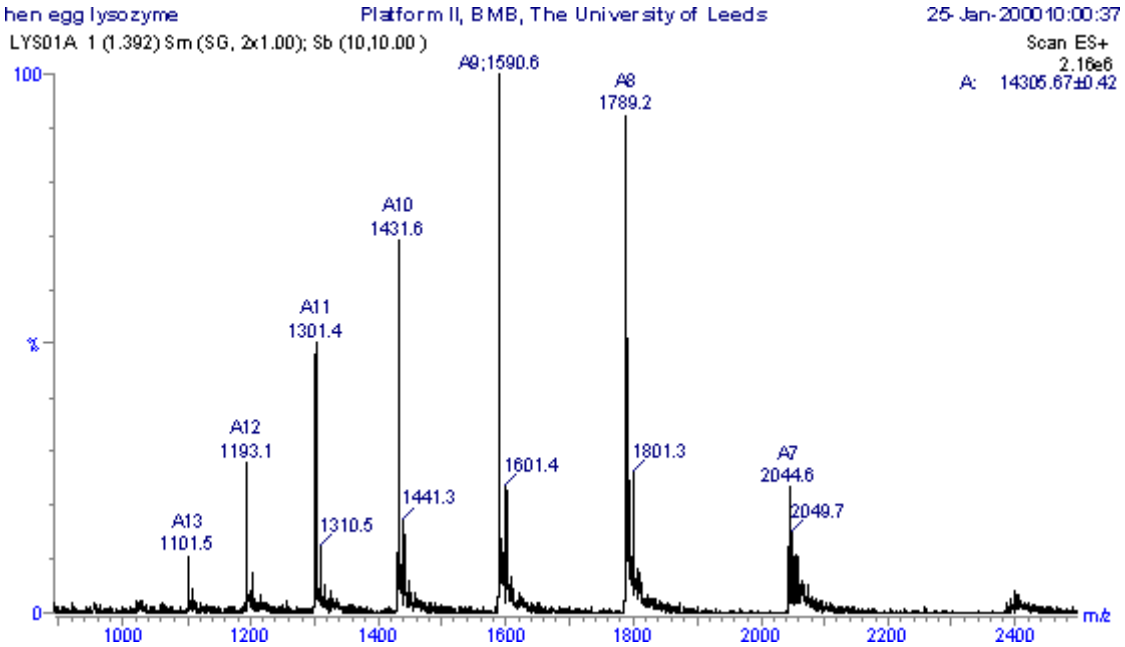


Figure 22 Positive ESI-MS m/z spectrum of the protien hen egg white lysozyme

The sample was analysed in a solution of 1:1 (*v/v*) acetonitrile: 0.1% aqueous formic acid and the m/z spectrum shows a Gaussian-type distribution of multiply charged ions ranging from m/z 1101.5 to 2044.6. Each peak represents the intact protein molecule carrying a different number of charges (protons). The peak width is greater than that of the singly charged ions seen in the leucineenkephalin spectrum, as the isotopes associated with these multiply charged ions are not clearly resolved as they were in the case of the singly charged ions. The individual peaks in the multiply charged series become closer together at lower m/z values and, because the molecular weight is the same for all of the peaks, those with more charges appear at lower m/z values than do those with fewer charges. The m/z values can be expressed as follows: m/z

$= (MW + nH^+)/n$

where m/z = the mass-to-charge ratio marked on the abscissa of the spectrum;
MW = the molecular mass of the sample n = the integer number of charges on the ions H = the mass of a proton = 1.008 Da.

If the number of charges on an ion is known, then it is simply a matter of reading the m/z value from the spectrum and solving the above equation to determine the molecular weight of the sample. Usually the number of charges is not known, but can be calculated if the assumption is made that any two adjacent members in the series of multiply charged ions differ by one charge.

For example, if the ions appearing at m/z 1431.6 in the lysozyme spectrum have "n" charges, then the ions at m/z 1301.4 will have "n+1" charges, and the above equation can be written again for these two ions:

$1431.6 = (MW + nH^+)/n$ and $1301.4 = [MW + (n+1)H^+] /(n+1)$

These simultaneous equations can be rearranged to exclude the MW term:

$n(1431.6) - nH^+ = (n+1)1301.4 - (n+1)H^+$
and so:
$n(1431.6) = n(1301.4) +1301.4 - H^+$
therefore:
$n(1431.6 - 1301.4) = 1301.4 - H^+$
and so:
$n = (1301.4 - H^+) / (1431.6 - 1301.4)$

hence the number of charges on the ions at m/z 1431.6 = 1300.4/130.2 = 10. Putting the value of n back into the equation:
$1431.6 = (MW + nH^+) n$
gives $1431.6 \times 10 = MW + (10 \times 1.008)$ and so MW = 14,316 - 10.08
therefore **MW = 14,305.9 Da**

The observed molecular mass is in good agreement with the theoretical molecular mass of hen egg lysozyme (based on average atomic masses) of 14305.14 Da. The individual isotopes cannot be resolved when the ions have a large number of charges, and so for proteins the average mass is measured.

This may seem long-winded but fortunately the molecular mass of the sample can be calculated automatically, or at least semi-automatically, by the processing software associated with the mass spectrometer. This is of great help for multi-component mixture analysis where the m/z spectrum may well contain several overlapping series of multiply charged ions, with each component exhibiting completely different charge states.

Using **electrospray** or **nanospray ionisation,** a **mass accuracy of within 0.01%** of the molecular mass should be achievable, which in this case represents +/- 1.4 Da.

In order to clarify electrospray/nanospray data, **molecular mass profiles** can be generated from the m/z spectra of high molecular mass, multiply charged samples. To achieve this, all the components are transposed onto a true molecular mass (or **zero charge state**) profile from which molecular masses can be read directly without any amendments or calculations.

The m/z spectrum of lysozyme has been converted to a molecular mass profile using Maximum Entropy processing and the data are shown. The mass profile is dominated by a component of molecular mass 14,305.7 Da, with a series of minor peaks at higher mass, which is usually indicative of salt adducting e.g. Na (M+23), K (M+39), H_2SO_4 or H_3PO_4 (M+98). The molecular masses can be read easily and unambiguously, and a good idea of the purity of the protein is obtained on inspection of the molecular mass profile.

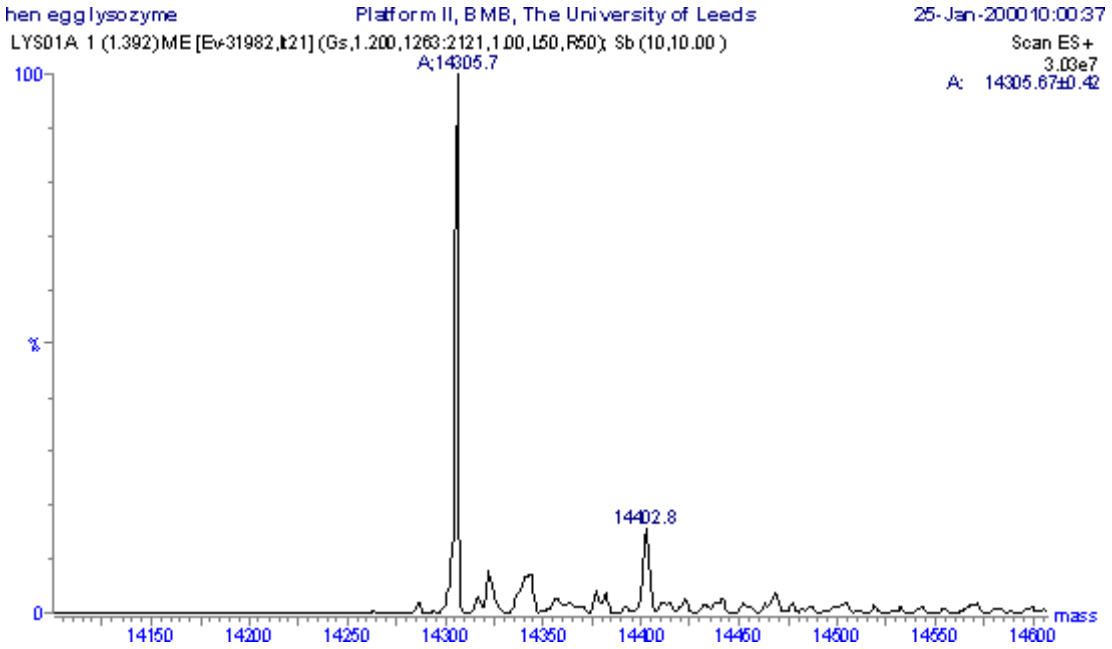


Figure 23 Molecular mass profile of lysozyme obtained by maximum entropy processing of the m/z spectrum

Proteins in their **native state**, or at least containing a significant amount of folding, tend to produce multiply charged ions covering a smaller range of charge states (say two or three). These charge states tend to have fewer charges than an unfolded protein would have, due to the inaccessibility of many of the protonation sites. In such cases, increasing the **sampling cone voltage** may provide sufficient energy for the protein to begin to unfold and create a wider charge state distribution centering on more highly charged ions in the lower m/z region of the spectrum.

The differences in m/z spectra due to the folded state of the protein are illustrated with the m/z spectra of the protein apo-pseudoazurin acquired under different solvent conditions.

Analysis of the protein in 1:1 acetonitrile : 0.1% aqueous formic acid at pH2 gave a Gaussian-type distribution with multiply charged states ranging from n = 9 at m/z 1487.8 to n = 19 at m/z 705.3, centering on n = 15 (lower trace). The molecular mass for this protein was 13,381 Da. Analysis of the protein in water gave fewer charge states, from n = 7 at m/z 1921.7 to n = 11 at m/z 1223.7, centering at n = 9 (upper trace). Not only has the charge state distribution changed, the molecular weight is now 13,444 Da which represents an increase of 63 Da and indicates that copper is remaining bound to the protein. Many types of **protein complexes** can

be observed in this way, including **protein-ligand, protein-peptide, protein- metal and protein-RNA macromolecules**.

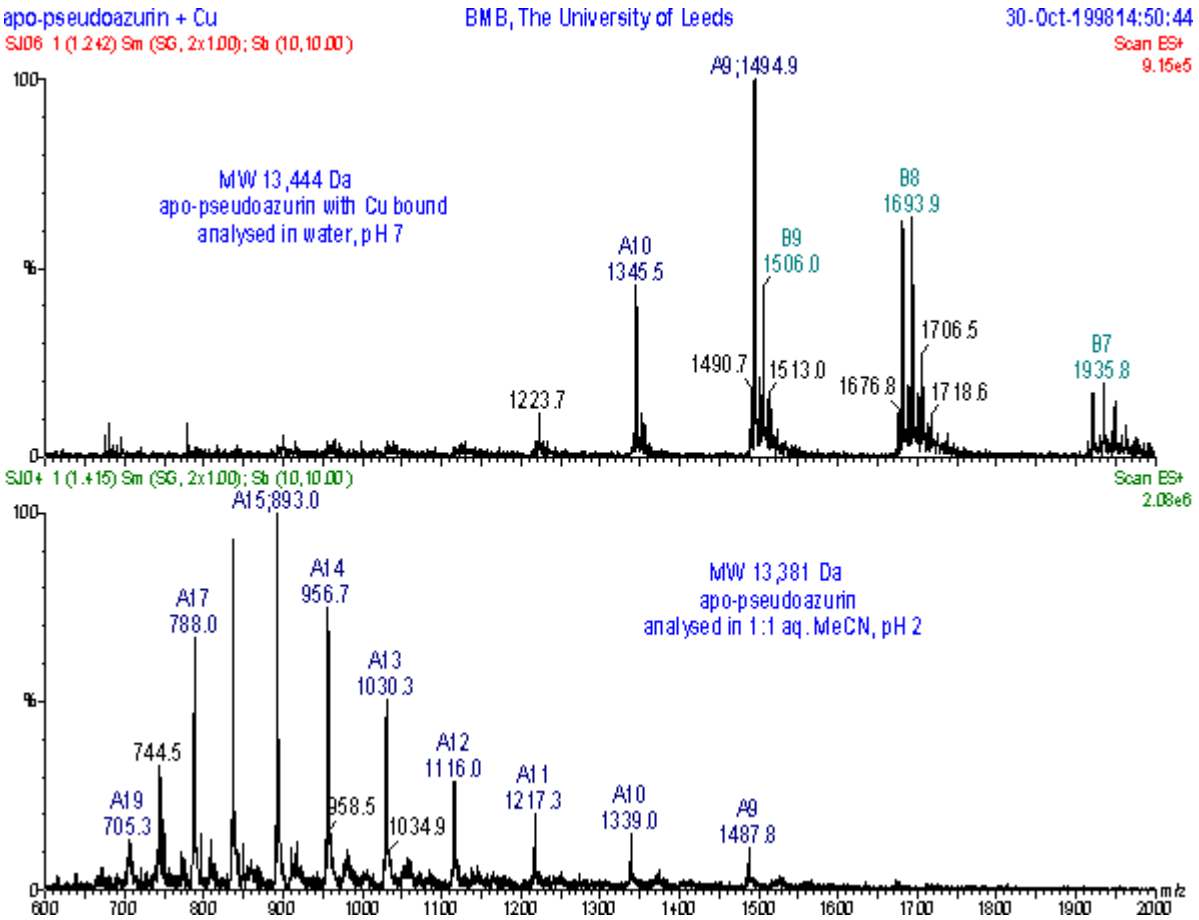


Figure 24 Positive ESI-MS m/z spectra of the protein apo-pseudoazurin analysed in water at pH7 (upper trace) and in 1:1 acetonitrile:0.1% aq. formic acid at pH2 (lower trace)

Matrix assisted laser desorption ionisation

Matrix Assisted Laser Desorption Ionisation (MALDI) deals well with thermolabile, non-volatile organic compounds especially those of high molecular mass and is used successfully in biochemical areas for the analysis of **proteins, peptides, glycoproteins, oligosaccharides, and oligonucleotides**. It is relatively straightforward to use and reasonably tolerant to buffers and other additives. The mass accuracy depends on the type and performance of the analyser of the mass spectrometer, but most modern instruments should be capable of measuring masses to within 0.01% of the molecular mass of the sample, at least up to ca. 40,000 Da.

MALDI is based on the **bombardment** of sample molecules with a **laser** light to bring about

sample ionisation. The sample is pre-mixed with a highly absorbing **matrix** compound for the most consistent and reliable results, and a low concentration of sample to matrix works best. The matrix transforms the laser energy into **excitation energy** for the sample, which leads to sputtering of analyte and matrix ions from the surface of the mixture. In this way energy transfer is efficient and also the analyte molecules are spared excessive direct energythat may otherwise cause decomposition. Most commercially available MALDI mass spectrometers now have a pulsed nitrogen laser of wavelength 337 nm.

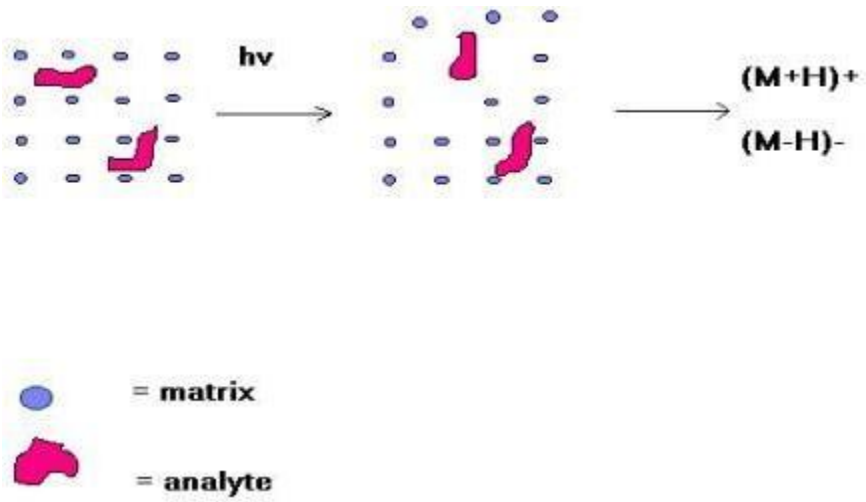


Figure 25 Matrix assisted laser desorption ionisation (MALDI)

The sample to be analysed is dissolved in an appropriate volatile solvent, usually with a trace of trifluoroacetic acid if positive ionisation is being used. A range of compounds is suitable for use as matrices: **sinapinic acid** is a common one for **protein** analysis while **alpha-cyano-4- hydroxycinnamic acid** is often used for **peptide** analysis. The final solution is applied to the sample target which is allowed to dry prior to insertion into the high vacuum of the mass spectrometer. The laser is fired, the energy arriving at the sample/matrix surface optimised, and data accumulated until a m/z spectrum of reasonable intensity has been amassed. The time-of-flight analyser separates ions according to their **mass(m)-to- charge(z) (m/z)** ratios by measuring the time it takes for ions to travel through a field free region known as the flight, or drift, tube. The heavier ions are slower than the lighter ones.

The m/z scale of the mass spectrometer is **calibrated** with a known sample that can either be analysed independently (external calibration) or pre-mixed with the sample and matrix (internal calibration).

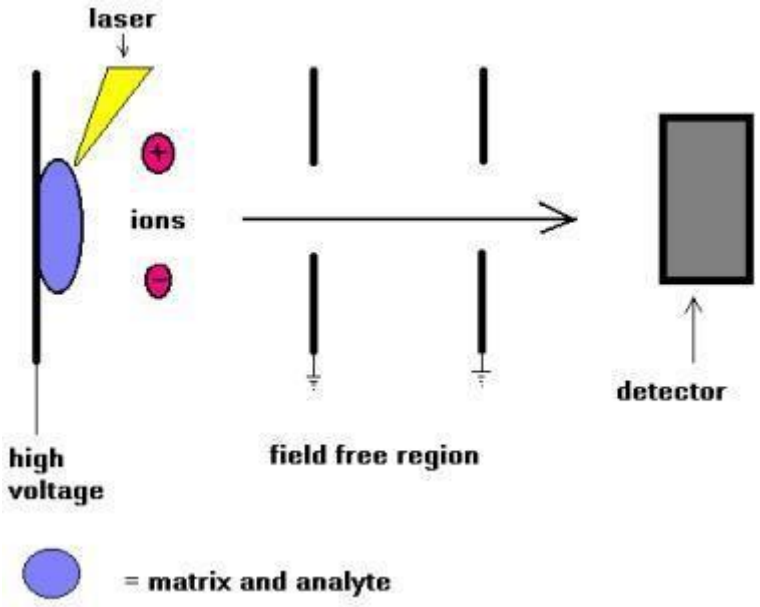


Figure 26 Simplified schematic of MALDI-TOF mass spectrometry (linear mode)

MALDI is also a "soft" ionisation method and so results predominantly in the generation of **singly charged molecular-related ions** regardless of the molecular mass, hence the spectra are relatively easy to interpret. Fragmentation of the sample ions does not usually occur.

In **positive ionisation** mode the **protonated molecular ions ($M+H^+$)** are usually the dominant species, although they can be accompanied by salt adducts, a trace of the doubly charged molecular ion at approximately half the m/z value, and/or a trace of a dimeric species at approximately twice the m/z value. Positive ionisation is used in general for **protein** and **peptide** analyses.

In **negative ionisation** mode the **deprotonated molecular ions ($M-H^-$)** are usually the most abundant species, accompanied by some salt adducts and possibly traces of dimeric or doubly charged materials. Negative ionisation can be used for the analysis of **oligonucleotides** and **oligosaccharides**.

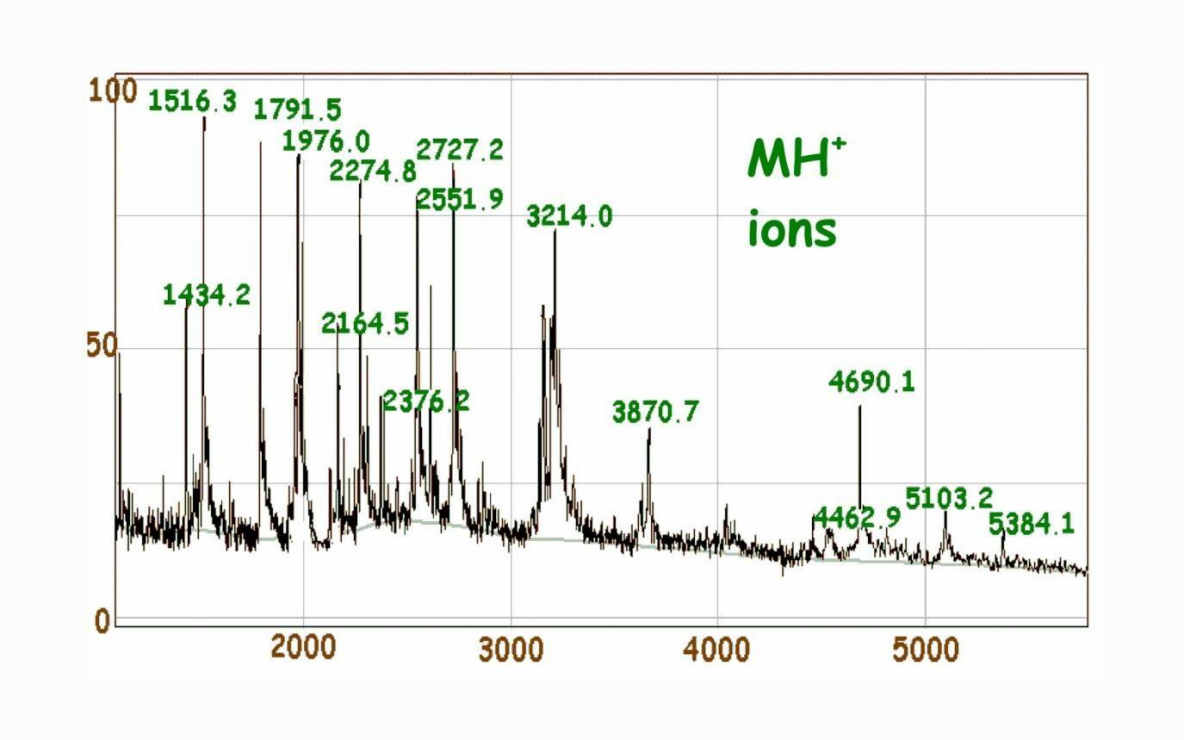


Figure 27 Positive ionisation MALDI m/z spectrum of a peptide mixture using alpha-cyano-4- hydroxycinnamic acid as matrix

Positive or negative ionisation?

If the sample has functional groups that readily accept a proton (H^+) then positive ion detection is used

e.g. amines R-NH_2 + H^+ = R-NH_3^+ as in proteins or peptides.

If the sample has functional groups that readily lose a proton then negative ion detection is used

e.g. carboxylic acids R-CO_2H = R-CO_2^- and alcohols R-OH = R-O- as in saccharides or oligonucleotides

Tandem mass spectrometry (MS-MS): Structural and sequence information from mass spectrometry

Tandem mass spectrometry (MS-MS) is used to produce **structural information** about a compound by fragmenting specific sample ions inside the mass spectrometer and identifying the resulting fragment ions. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic

fragmentation patterns.

A **tandem mass spectrometer** is a mass spectrometer that has more than one analyser, in practice usually two. The two analysers are separated by a collision cell into which an inert gas (e.g. argon, xenon) is admitted to collide with the selected sample ions and bring about their fragmentation. The analysers can be of the same or of different types, the most common combinations being:

- quadrupole - quadrupole
- magnetic sector - quadrupole
- magnetic sector - magnetic sector
- quadrupole - time-of-flight.

Fragmentation experiments can also be performed on certain single analyser mass spectrometers such as ion trap and time-of-flight instruments, the latter type using a post-source decay experiment to effect the fragmentation of sample ions.

Tandem mass spectrometry analyses.

The basic modes of data acquisition for tandem mass spectrometry experiments are as follows:

Product or daughter ion scanning:

the first analyser is used to select user-specified sample ions arising from a particular component; usually the molecular-related (i.e. $(M+H)^+$ or $(M-H)^-$) ions. These chosen ions pass into the collision cell, are bombarded by the gas molecules which cause fragment ions to be formed, and these fragment ions are analysed i.e. separated according to their mass to charge ratios, by the second analyser. All the fragment ions arise directly from the precursor ions specified in the experiment, and thus produce a fingerprint pattern specific to the compound under investigation.

This type of experiment is particularly useful for providing structural information concerning **small organic molecules** and for generating **peptide sequence** information.

Precursor or parent ion scanning:

The first analyser allows the transmission of all sample ions, whilst the second analyser is set

to monitor specific fragment ions, which are generated by bombardment of the sample ions with the collision gas in the collision cell. This type of experiment is particularly useful for monitoring groups of compounds contained within a mixture which fragment to produce common fragment ions, e.g. **glycosylated peptides** in a tryptic digest mixture, **aliphatic hydrocarbons** in an oil sample, or **glucuronide conjugates** in urine.

Constant neutral loss scanning:

This involves both analysers scanning, or collecting data, across the whole m/z range, but the two are off-set so that the second analyser allows only those ions which differ by a certain number of mass units (equivalent to a neutral fragment) from the ions transmitted through the first analyser. e.g. This type of experiment could be used to monitor all of the carboxylic acids in a mixture. Carboxylic acids tend to fragment by losing a (neutral) molecule of carbon dioxide, CO_2, which is equivalent to a loss of 44 Da or atomic mass units. All ions pass through the first analyser into the collision cell. The ions detected from the collision cell are those from which 44 Da have been lost.

Selected/multiple reaction monitoring:

Both of the analysers are static in this case as user-selected specific ions are transmitted through the first analyser and user-selected specific fragments arising from these ions are measured by the second analyser. The compound under scrutiny must be known and have been well-characterised previously before this type of experiment is undertaken. This methodology is used to confirm unambiguously the presence of a compound in a matrix e.g. drug testing with blood or urine samples. It is not only a highly specific method but also has very high sensitivity.

Peptide Sequencing by Tandem Mass Spectrometry.

The most common usage of MS-MS in biochemical areas is the **product or daughter ion scanning** experiment which is particularly successful for **peptide** and **nucleotide sequencing**.

Peptide sequencing: H_2N-CH(R')-CO-NH-CH(R")-CO_2H

Peptides fragment in a reasonably well-documented manner. The protonated molecules fragment along the **peptide backbone** and also show some **side-chain fragmentation** with certain instruments.

There are three different types of bonds that can fragment along the amino acid backbone: the **NH-CH**, **CH-CO**, and **CO-NH** bonds. Each bond breakage gives rise to two species, one neutral and the other one charged, and only the charged species is monitored by the mass spectrometer. The charge can stay on either of the two fragments depending on the chemistry and relative proton affinity of the two species. Hence there are six possible fragment ions for each amino acid residue and these are labelled as in the diagram, with the **a, b, and c" ions** having the charge retained on the **N-terminal fragment**, and the **x, y", and z ions** having the charge retained on the **C-terminal fragment**. The most common cleavage sites are at the CO-NH bonds which give rise to the b and/or the y" ions. The mass difference between two adjacent b ions, or y"; ions, is indicative of a particular amino acid residue.

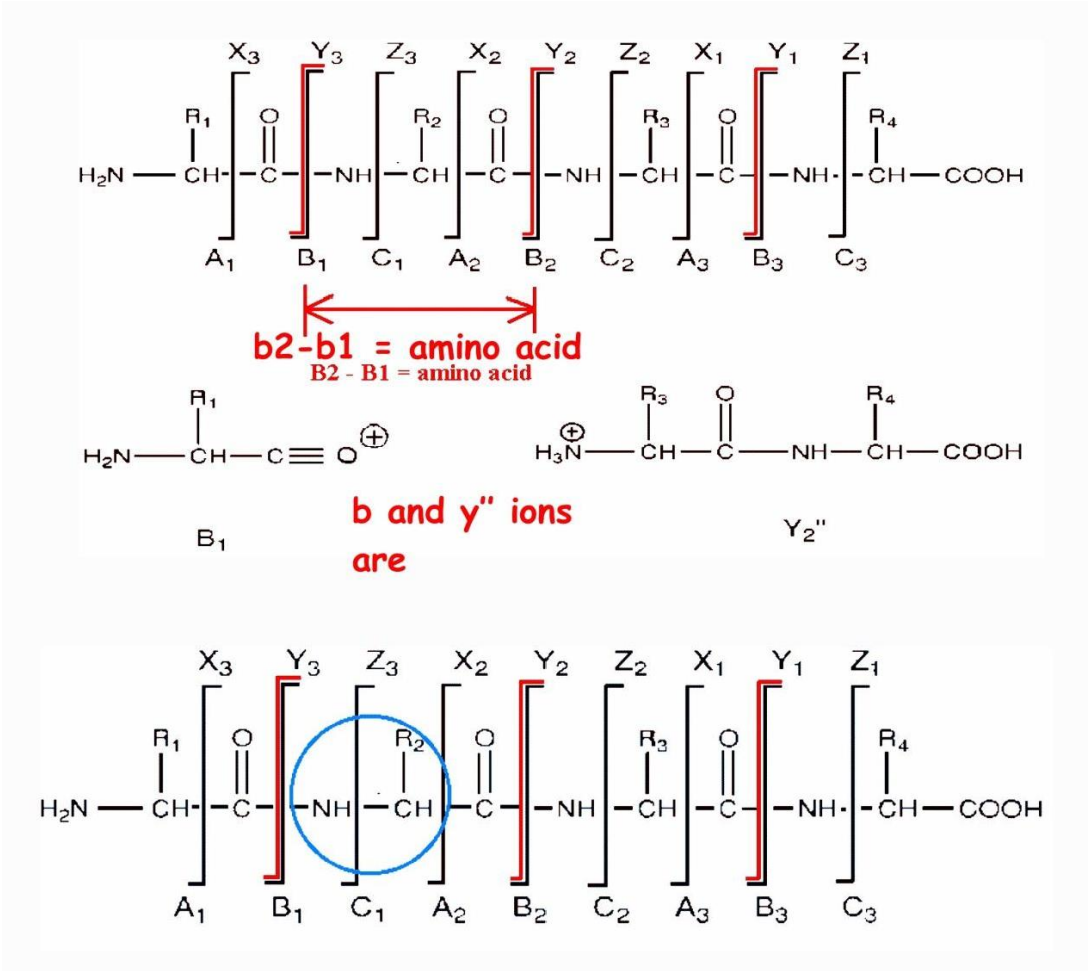


Figure 28 Peptide sequencing by tandem mass spectrometry - backbone cleavages

The extent of **side-chain fragmentation** detected depends on the type of analysers used in the mass spectrometer. A magnetic sector - magnetic sector instrument will give rise to **high**

energy collisions resulting in many different types of side-chain cleavages. Quadrupole - quadrupole and quadrupole - time-of-flight mass spectrometers generate **low energy** fragmentations with fewer types of side-chain fragmentations.

Immonium ions (labelled "i") appear in the very low m/z range of the MS-MS spectrum. Each amino acid residue leads to a diagnostic immonium ion, with the exception of the two pairsleucine (L) and iso-leucine (I), and lysine (K) and glutamine (Q), which produce immonium ions with the same m/z ratio, i.e. m/z 86 for I and L, m/z 101 for K and Q. The immoniumions are useful for detecting and confirming many of the amino acid residues in a peptide, although no information regarding the position of these amino acid residues in the peptide sequence can be ascertained from the immonium ions.

An example of an **MS/MS daughter or product ion spectrum** is illustrated below. The molecular mass of the peptide was measured using standard mass spectrometric techniques and found to be 680.4 Da, the dominant ions in the MS spectrum being the protonated molecular ions ($M+H^+$) at m/z 681.4. These ions were selected for transmission through the first analyser, then fragmented in the collision cell and their fragments analysed by the second analyser to produce the following MS/MS spectrum. The **sequence (amino acid backbone) ions** have been identified, and in this example the peptide is fragmented predominantly at the **CO-NH** bonds and gave both b and y" ions. (Often either the b series or the y" series predominates, sometimes to the exclusion of the other). The b series ions have been labelled with blue vertical lines and the y" series ions have been labelled with red vertical lines. The mass difference between adjacent members of a series can be calculated
e.g. b3-b2 = 391.21 - 262.16 = 129.05 Da which is equivalent to a glutamine (E) amino acid residue; and similarly y4 - y3 = 567.37 - 420.27 = 147.10 Da which is equivalent to a phenylalanine (F) residue. In this way, using either the b series or the y" series, the amino acid sequence of the peptide can be determined and was found to be NFESGK (n.b. the y" series reads from right to left!). The immonium ions at m/z 102 merely confirm the presence of the glutamine (E) residue in the peptide.

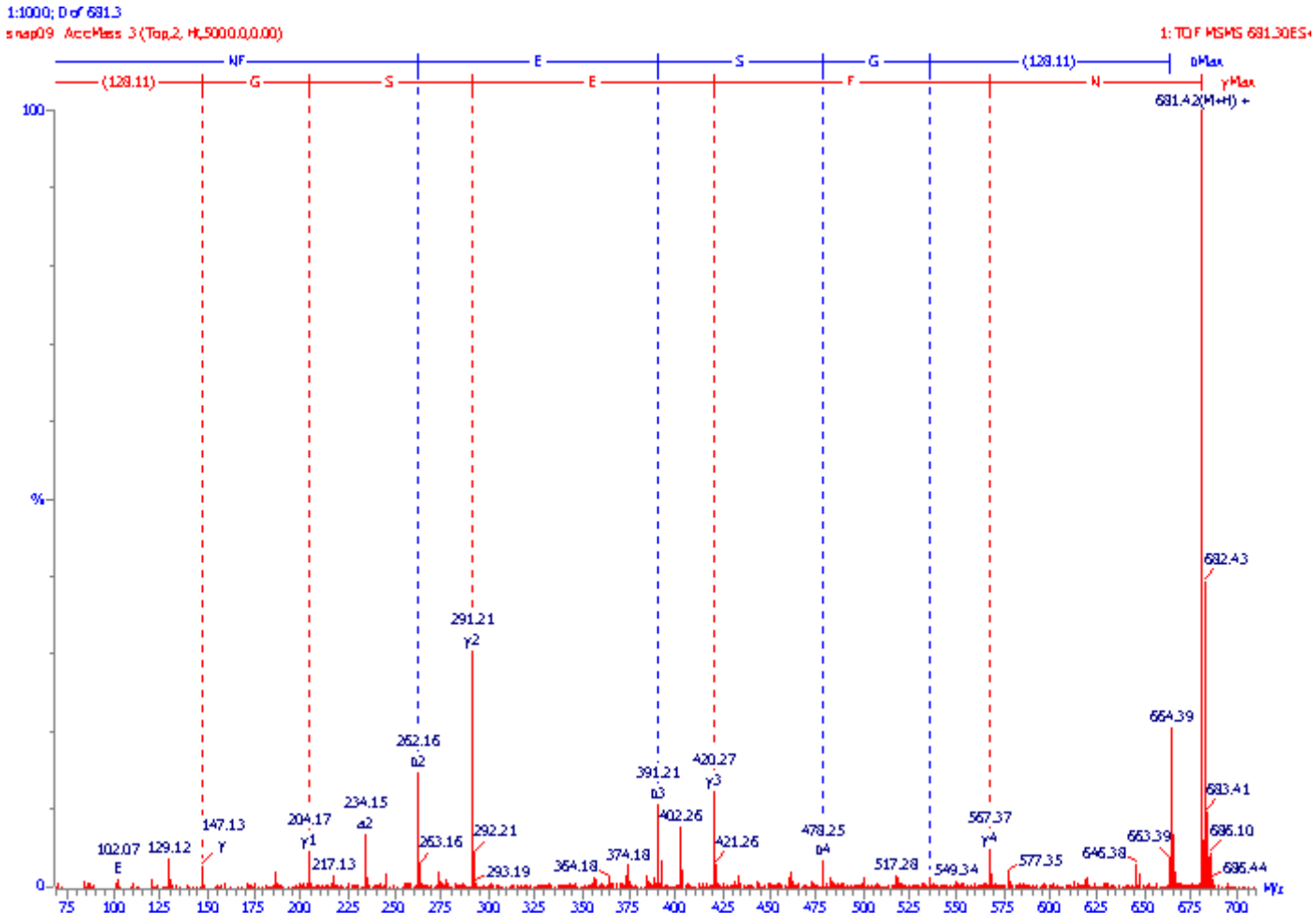


Figure 29 Peptide sequencing by tandem mass spectrometry - an MS-MS daughter or product ion spectrum

A **protein identification study** would proceed as follows:

- a. The **protein** under investigation would be analysed by mass spectrometry to generate a molecular mass to within an accuracy of 0.01%.
- b. The protein would then be **digested** with a suitable enzyme. **Trypsin** is useful for mass spectrometric studies because each proteolytic fragment contains a basic **arginine (R)** or **lysine (K)** amino acid residue, and thus is eminently suitable for positive ionisation mass spectrometric analysis. The digest mixture is analysed without prior separation or clean-up - by mass spectrometry to produce a rather complex spectrum from which the molecular weights of all of the proteolytic fragments can be read. This spectrum, with its molecular weight information, is called a **peptide map**. (If the protein already exists on a **database**, then the peptide map is often sufficient to confirm the protein.)

For these experiments the mass spectrometer would be operated in the **"MS" mode**, whereby the sample is sprayed and ionised from the nanospray needle and the ions pass through the sampling cone, skimmer lenses, Rfhexapole focusing system, and the first (quadrupole) analyser. The quadrupole in this instance is not used as an analyser, merely as a lens to focus

the ion beam into the second (time-of-flight) analyser which separates the ions according to their mass-to-charge ratio.

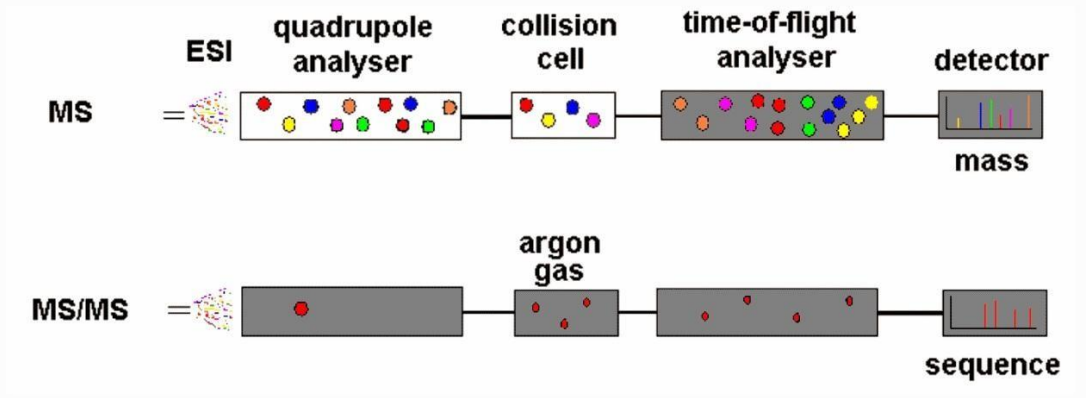


Figure 30 Q-TOF mass spectrometer operating in MS (upper) and MS/MS mode (lower) modes

- c. With the digest mixture still spraying into the mass spectrometer, the Q-Tof mass spectrometer is switched into **"MS/MS" mode**. The protonated molecular ions of each of the digest fragments can be independently selected and transmitted through the quadrupole analyser, which is now used as an analyser to transmit solely the ions of interest into the **collision cell** which lies in between the first and second analysers. An inert gas such as argon is introduced into the collision cell and the sample ions are bombarded by the collision gas molecules which cause them to fragment. The optimum collision cell conditions vary from peptide to peptide and must be optimised for each one. The **fragment (**or **daughter or product**) **ions** are then analysed by the second (time-of-flight) analyser. In this way an **MS/MS spectrum** is produced showing all the **fragment ions** that arise directly from the chosen **parent** or **precursor ions** for a given peptide component.

An **MS/MS daughter** (or **fragment**, or **product**) ion spectrum is produced for each of the components identified in the proteolytic digest. Varying amounts of sequence information can be gleaned from each fragmentation spectrum, and the spectra need to be interpreted carefully. Some of the processing can be automated, but in general the **processing** and **interpretation** of spectra will take longer than the data acquisition if accurate and reliable data are to be generated.

The amount of sequence information generated will vary from one peptide to another, Some peptide sequences will be confirmed totally, other may produce a partial sequence of, say, 4

or 5 amino acid residues. Often sequence "tag" of 4 or 5 residues is sufficient to search a protein database and confirm the identity of the protein.

Peptide Mass Fingerprinting (PMF)

This is another method of protein identification. In this method, 2-D gel electrophoresis is used for protein separation. The separated spots are obtained from the gel and then identified by PMF. The technique is based on the use of a proteolytic enzyme to digest the protein into smaller peptides. The most commonly used enzyme is trypsin, which cleaves lysine and arginine sites. When the digestion is complete, a set of peptides are produced of varying masses that are unique to that protein. The mass of each peptide will be the sum of amino acids present, inducing any modifications that amino acids might have undergone. Once the set of peptides have been obtained, one has to search for peptide sequences.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit 3-Protein Engineering and Bioinformatics – SBTA1303

III PROTEIN DATABASES AND SEQUENCE ANALYSIS

Biological databases

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Why databases?

- Means to handle and share large volumes of biological data
- Support large-scale analysis efforts
- Make data access easy and updated
- Link knowledge obtained from various fields of biology and medicine

Features

- Most of the databases have a web-interface to search for data
- Common mode to search is by Keywords
- User can choose to view the data or save to your computer
- Cross-references help to navigate from one database to another easily

Biological databases can be broadly classified into sequence and structure databases. Nucleic acid and protein sequences are stored in sequence databases and structure database only store proteins. These databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among species in the history of life.

A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. There are two main functions of biological databases:

- **Make biological data available to scientists.**
    - As much as possible of a particular type of information should be available in one single place (book, site, and database). Published data may be difficult to find or access and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).
- **To make biological data available in computer-readable form.**
    - Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

Data Domains

- Types of data generated by molecular biology research:
  – Nucleotide sequences (DNA and mRNA)
  – Protein sequences
  – 3-D protein structures
  – Complete genomes and maps

Sequence Databases

- Nucleic acid sequence databases
- EMBL • GenBank • DDBJ

- Main protein sequence databases
- Swiss Prot
- also TREMBL, GenPept

- Often integrated with other databases

Structure databases

– Experimental data and model coordinates
- NDB, wwPDB, BMRB, CSD, EMDB

Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries. While the initial databases of protein sequences were maintained at the individual laboratories, the development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CDROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

Databases in general can be classified in to **primary**, **secondary** and **composite** databases. A **primary** database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures.

A **secondary** database contains derived information from the primary database. A secondary sequence database contains information like the conserved sequence, signature sequence and active site residues of the protein families arrived by multiple sequence alignment of a set of related proteins. A secondary structure database contains entries of the PDB in an organized way. These contain entries that are classified according to their structure like all alpha proteins, all beta proteins, etc. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary database created and hosted by various researchers at their individual laboratories includes SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford.

Composite database amalgamates a variety of different primary database sources, which obviates the need to search multiple resources. Different composite database use different primary database and different criteria in their search algorithm. Various options for search have also been incorporated in the composite database. The National Center for Biotechnology Information (NCBI) which hosts these nucleotide and protein databases in their large high available redundant array of computer servers, provides free access to the various persons involved in research. This also has link to OMIM (Online Mendelian Inheritance in Man) which contains information about the proteins involved in genetic diseases.

Primary databases

I. Primary database

1. It is also known as archival database
2. Databases consisting of data derived experimentally such as nucleotide sequences and three dimensional structures are known as primary databases.
3. Experimental results are directly submitted into database by researchers across the globe
4. Example: Gen bank, DDBJ, SWISS-PROT

- Contain sequence data such as nucleic acid or protein
- Example of primary databases include :

Protein Databases

- SWISS-PROT
- TREMBL
- PIR

Nucleic Acid Databases

- EMBL
- Genbank
- DDBJ

Secondary databases

II. Secondary database

1. It is also known as curated database
2. Databases consisting of data derived from the analysis of primary data such as sequences, secondary structures etc
3. It contains results of analysis of primary databases and significant data in the form of conserved sequences, signature sequences, active site residues of proteins etc.

- Or sometimes known as pattern databases
- Contain results from the analysis of the sequences in the primary databases
- Example of secondary databases include : PROSITE, Pfam, BLOCKS, PRINTS

Composite databases

- Combine different sources of primary databases.
- Make querying and searching efficient and without the need to go to each of the primary databases.
- Example of composite databases include : NRDB – Non-Redundant DataBase, OWL

UniProt

UniProt is a comprehensive, high-quality and freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. Universal Protein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases.

The UniProt consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). EBI, located at the Welcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy(Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the National Biomedical Research Foundation (NBRF) at the Georgetown University Medical Center in Washington, DC, USA, is heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure, first published in 1965.[2] In 2002, EBI, SIB, and PIR joined forces as the UniProt consortium

SWISSPROT

SWISS-PROT is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton Outstation, the European Bioinformatics Institute (EBI). The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of SWISS-PROT (see http://www.expasy. ch/txt/userman.txt) follows as closely as possible that of the EMBL Nucleotide Sequence Database.

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases (Cross references).

Annotation

In SWISS-PROT two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains, etc.
- Secondary structure. For example alpha helix, beta sheet, etc.
- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT we try as much as possible to merge all these data so as to minimise the redundancy of the database

Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialised data collections. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. For example the sample sequence mentioned above contains, among others, DR (Databank Reference) lines that point to EMBL, PDB, OMIM, Pfam and PROSITE.

TREMBL: A COMPUTER ANNOTATED SUPPLEMENT TO SWISS-PROT

Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed annotation of every entry. This is the rate-limiting step in the

production of SWISS-PROT. On one hand we do not wish to relax the high editorial standards of SWISS-PROT and it is clear that there is a limit to how much we can accelerate the annotation procedures. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDSs) in the EMBL database, except for CDSs already included in SWISS-PROT.

We have split TREMBL into two main sections, SP-TREMBL and REM-TREMBL. SP-TREMBL (SWISS-PROT TREMBL) contains entries (~55 000) which should be incorporated into SWISS-PROT. SWISS-PROT accession numbers have been assigned to these entries. SP- TREMBL is partially redundant against SWISS-PROT, since ~30 000 of these SP-TREMBL entries aie only additional sequence reports of proteins already in SWISS-PROT. REM-TREMBL (REMaining TREMBL) contains those entries (~15 000) that we do not wish to include in SWISS- PROT. This section is organized into four subsections. Most REM-TREMBL entries are immunoglobulins and T-cell receptors. We have stopped entering immunoglobulins and T-cell receptors into SWISS-PROT, because we want to keep only germ line gene-derived translations of these proteins in SWISS-PROT and not all known somatic recombinant variations of these proteins. Another category of data which will not be included in SWISS-PROT is synthetic sequences. A third subsection consists of fragments with less than seven amino acids. The last subsection consists of CDS translations where we have strong evidence to believe that these CDS are not coding for real proteins.

The creation of TREMBL as a supplement to SWISS-PROT was not only for the purpose of producing a more complete and up to date protein sequence collection. Also to achieve a deeper integration of the EMBL nucleotide sequence database with SWISS-PROT + TREMBL.

Structure of a sequence entry

The entries in the SWISS-PROT data bank are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols familiar to biochemists, protein chemists and molecular biologists are used. Each sequence entry is composed of lines.

Different types of lines, each with their own format, are used to record the various data which make up the entry.

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown below:

ID - Identification.
AC - Accession number(s). DT - Date.
DE - Description. GN - Gene name(s).
OS - Organism species. OG - Organelle.
OC - Organism classification. RN - Reference number.
RP - Reference position. RC - Reference comments.
RX - Reference cross-references. RA - Reference authors.
RL - Reference location. CC - Comments or notes.
DR - Database cross-references. KW Keywords.
FT - Feature table data. SQ - Sequence header.
(blanks) sequence data. // - Termination line.

Protein Information Resource

The Protein Information Resource (PIR), located at Georgetown University Medical Center (GUMC), is an integrated public bioinformatics resource to support genomic and proteomic research, and scientific studies. PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers and costumers in the identification and interpretation of protein sequence information. Prior to that, the NBRF compiled the first comprehensive collection of macromolecular sequences in the Atlas of Protein Sequence and Structure, published from 1964-1974 under the editorship of Margaret Dayhoff.

Dr. Dayhoff and her research group pioneered in the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences.

The Protein Information Resource (PIR) produces the largest, most comprehensive, annotated protein sequence database in the public domain, the PIR-International Protein Sequence Database, in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Sequence Database (JIPID).

PIR, MIPS and JIPID constitute the PIR-International consortium that maintains the PIR-International Protein Sequence Database (PSD), the largest publicly distributed and freely available protein sequence database. The database has the following distinguishing features.

- It is a comprehensive, annotated, and non-redundant protein sequence database, containing over 142 000 sequences as of September 1999. Included are sequences from the completely sequenced genomes of 16 prokaryotes, six archaebacteria, 17 viruses and phages, >100 eukaryote organelles and Saccharomyces cerevisiae.

- The collection is well organized with >99% of entries classified by protein family and >57% classified by protein superfamily.

- PSD annotation includes concurrent cross-references to other sequence, structure, genomic and citation databases, including the public nucleic acid sequence databases ENTREZ, MEDLINE, PDB, GDB, OMIM, FlyBase, MIPS/Yeast, SGD/Yeast, MIPS/Arabidopsis and TIGR. Where these databases are publicly and freely accessible and provide suitable WWW access, the cross-references presented on the PIR WWW site are hot-linked so that searchers can consult the most current data.

- The PIR is the only sequence database to provide context cross-references between its own database entries. These cross-references assist searchers in exploring relationships such as subunit associations in molecular complexes, enzyme–substrate interactions, activation and regulation cascades, as well as in browsing entries with shared features and annotations.

- Interim updates are made publicly available on a weekly basis, and full releases have been published quarterly since 1984.

It is split into 4 distinct section (PIR1-PIR4).

PIR1: contains fully classified and annotated entries.

PIR2: includes preliminary entries not been thoroughly reviewed, contain redundancy.

PIR3: contains unverified entries.

PIR4: fall into one of four categories.

- Conceptual translations of art factual sequences.
- Conceptual translations of sequences that are not transcribed or translated.
- Protein sequences or conceptual translations that are extensively genetically engineered
- Sequences that are not genetically encoded and not produced on ribosomes.

Protein data bank

The Protein Data Bank (PDB) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology

The Protein Data Bank (PDB) at Brookhaven National Laboratory (BNL), is a database containing experimentally determined three-dimensional structures of proteins, nucleic acids and other biological macromolecules. The archives contain atomic coordinates, citations, primary and secondary structure information, crystallographic structure experimental data, as well as hyperlinks to many other scientific databases.

Protein Data Bank (PDB) format is a standard for files containing atomic coordinates. Structures deposited in the Protein Data Bank at the Research Collaboratory for Structural

Bioinformatics (RCSB) are written in this standardized format. The complete PDB file specification provides for a wealth of information, including authors, literature references, and the identification of substructures such as disulfide bonds, helices, sheets, and active sites.

Protein Data Bank format consists of lines of information in a text file. Each line of information in the file is called a record. A file generally contains several different types of records, which are arranged in a specific order to describe a structure

Selected Protein Data Bank Record Types

Record Type	
ATOM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in standard residues (amino acids and nucleic acids).
HETATM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.
TER	indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains which are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.
SSBOND	defines disulfide bond linkages between cysteine residues.
HELIX	indicates the location and type (right-handed alpha, *etc.*) of helices. One record per helix.
SHEET	indicates the location, sense (anti-parallel, *etc.*) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.

Table 1 PDB Record Types

The Protein Data Bank (pdb) file format is a textual file format describing the three-dimensional structures of molecules held in theProtein Data Bank. The pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, observed sidechain rotamers, secondary structure assignments, as well as atomic connectivity. Structures are often deposited with other molecules such as water, ions, nucleic acids, ligands and so on, which can be described in the pdb format as well. The Protein Data Bank also keeps data on biological macromolecules in the newer mmCIF file format.

A typical PDB file describing a protein consists of hundreds to thousands of lines like the following (taken from a file describing the structure of a synthetic collagen-like peptide):

HEADER, TITLE and AUTHOR records provide information about the researchers who defined the structure; numerous other types of records are available to provide other types of information.

REMARK records can contain free-form annotation, but they also accommodate standardized information; for records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit.

SEQRES records give the sequences of the three peptide chains (named A, B and C), which are very short in this example but usually span multiple lines.

ATOM records describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Ångströms. The next three columns are the occupancy, temperature factor, and the element name, respectively.

HETATM records describe coordinates of hetero-atoms which are not part of the protein molecule.

Secondary Databases

- A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.
- Based on their contents, biological databases can be either primary database or secondary databases.
- Among the two, secondary databases have become a biologist's reference library over the past decade or so, providing a wealth of information on just any research or research product that has been investigated by the research community.
- Sequence annotation information in the primary database is often minimal.
- To turn the raw sequence information into more sophisticated biological knowledge, much post-processing of the sequence information is needed.
- This begs the need for secondary databases, which contain computationally processed sequence information derived from the primary databases.
- Thus, secondary databases comprise data derived from the results of analyzing primary data.
- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.
- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.
- The amount of computational processing work, however, varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

Importance of secondary databases

- Secondary databases contain information derived from primary sequence data which are in the form of regular expressions (patterns), Fingerprints, profiles blocks or Hidden Markov Models.
- The type of information stored in each of the secondary databases is different. But in secondary databases, homologous sequences may be gathered together in multiple alignments.
- In multiple alignments, there are conserved regions that show little or no variation between the constituent sequences. These conserved regions are called motifs.

- Motifs reflect some vital biological role and are crucial to the structure of the function of the protein. This is the importance of the secondary database.
- So by concentrating on motifs, we can find out the common conserved regions in the sequences and study the functional and evolutionary details or organisms.

Some of the common secondary databases include:

Prosite

- It was the first secondary database developed.
- Protein families usually contain some most conserved motifs which can be encoded to find out various biological functions.
- So by using such a database tool, we can easily find out the family of proteins when a new sequence is searched. This is the importance of PROSITE.
- Within PROSITE motifs are encoded as a regular expression (called patterns).
- Entries are deposited in PROSITE in two distant files. The first file gives the pattern and lists all matches of pattern, whereas the second one gives the details of family, description of the biological role, etc.
- The process used to derive patterns involves the construction of multiple alignment and manual inspection.
- So PROSITE contains documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.
- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.
- The protein motif and pattern are encoded as “regular expressions”.
- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

Prints

- Most protein families are characterized by several conserved motifs.
- All of these motifs can be an aid in constructing the `signatures" of different families. This principle is highlighted in constructing PRINT database.
- Within PRINTS motifs are encoded as unweighted local alignments. So small initial multiple alignments are taken to identify conserved motifs.
- Then these regions are searched in the database to find out similarities.
- Results are analyzed to find out the sequences which matched all the motifs within the fingerprint.
- PROSITE and PRINTS are the only manually annotated secondary databases. The print is a

diagnostic collection of protein fingerprints.

- In the PRINTS database, the protein sequence patterns are stored as "fingerprints". A fingerprint is a set of motifs or patterns rather than a single one.
- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.
- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences, the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

Blocks

- The limitations of the above two databases led to the formation of Block database.
- In this database, the motifs (here called Blocks) are created automatically by highlighting and detecting the most conserved regions of each family of proteins.
- Block databases are fully automated.
- Keyword and sequence searching are the two important features of this type of database.
- Blocks are ungapped Multiple Sequence Alignment representing conserved protein regions.

Pfam

- Pfam contains the profiles used using Hidden Markov models.
- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.
- Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.
- The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.
- The third is the HMM profile.
- The fourth element is the complete alignment of all the sequences identified in that family.

SEQUENCE ALIGNMENT

Pairwise alignment

Pairwise sequence alignment methods are used to find the best-matching piecewise (local)

or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods.

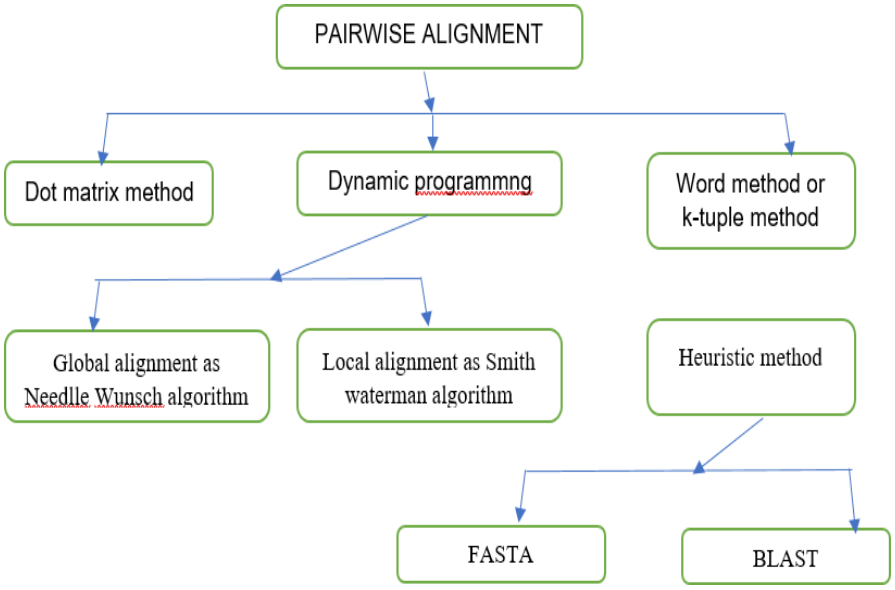


Figure 1

Dot plots

Dot plots are probably the oldest way of comparing sequences (Maizel and Lenk). A dot plot is a visual representation of the similarities between two sequences. Each axis of a rectangular array represents one of the two sequences to be compared. A window length is fixed, together with a criterion when two sequence windows are deemed to be similar. Whenever one window in one sequence resembles another a window in the other sequence, a dot or short diagonal is drawn at the corresponding position of the array. Thus, when two sequences share similarity over their entire length a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner. If two sequences only share patches of similarity this will be revealed by diagonal stretches.

Dynamic programming

The technique of dynamic programming can be applied to produce global alignments via the

Needleman-Wunsch algorithm, and local alignments via the Smith-Waterman algorithm. In typical usage, protein alignments use a substitution matrix to assign scores to amino-acid matches or mismatches, and a gap penalty for matching an amino acid in one sequence to a gap in the other. DNA and RNA alignments may use a scoring matrix, but in practice often simply assign a positive match score, a negative mismatch score, and a negative gap penalty. (In standard dynamic programming, the score of each amino acid position is independent of the identity of its neighbors, and therefore base stacking effects are not taken into account. However, it is possible to account for such effects by modifying the algorithm.) A common extension to standard linear gap costs, is the usage of two different gap penalties for opening a gap and for extending a gap. Typically the former is much larger than the latter, e.g. -10 for gap open and -2 for gap extension. Thus, the number of gaps in an alignment is usually reduced and residues and gaps are kept together, which typically makes more biological sense. The Gotoh algorithm implements affine gap costs by using three matrices. Needleman -Wunsch Algorithm

The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970.[1] The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems and uses the solutions to the smaller problems to reconstruct a solution to the larger problem.[2] It is also sometimes referred to as the optimal matching algorithm and the global alignment technique. The Needleman–Wunsch algorithm is still widely used for optimal global alignment, particularly when the quality of the global alignment is of the utmost importance.

The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings or nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. Like the Needleman–Wunsch algorithm, of which it is a variation, Smith–Waterman is a dynamic programming algorithm. As such, it has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme). The main difference to the Needleman–Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the (thus positively scoring) local alignments visible. Backtracking starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local

alignment. One does not actually implement the algorithm as described because improved alternatives are now available that have better scaling and are more accurate.

Word methods or k-tuple methods or heuristic methods

BLAST - Basic Local Alignment Search Tool

The BLAST algorithm was developed by Altschul, Gish, Miller, Myers and Lipman in 1990. The motivation for the development of BLAST was the need to increase the speed of FASTA by finding fewer and better hot spots during the algorithm. The idea was to integrate the substitution matrix in the first stage of finding the hot spots. The BLAST algorithm was developed for protein alignments in comparison to FASTA, which was developed for DNA sequences.

Different types of BLAST

blastn compares your query nucleotide sequence with database nucleotide sequences

blastp compares your query protein sequence with database of protein sequences that were derived form cDNA of interest blastx first translates your query sequence into amino acids in six reading frames (three forward and three back) then compares the protein sequences with protein databases

tblastn compares your query protein sequence with the database after translating each nucleotide sequence into protein using all six reading frames (This algorithm takes a long time, but is more likely to find distantly related sequences than the blastn, blastx, and blastp.)

tblastx translates both the query nucleotide sequence and the database sequences in all six reading frames and then compares the protein sequences (This, like tblastn, is very time consuming, but finds more results).

Algorithm

1. Remove low-complexity region or sequence repeats in the query sequence.
2. Make a k-letter word list of the query sequence.
3. List the possible matching words.
4. Organize the remaining high-scoring words into an efficient search tree.
5. Repeat step 3 to 4 for each k-letter word in the query sequence.
6. Scan the database sequences for exact matches with the remaining high-scoring words.
7. Extend the exact matches to high-scoring segment pair (HSP).

8. List all of the HSPs in the database whose score is high enough to be considered.

9. Evaluate the significance of the HSP score.

10. Make two or more HSP regions into a longer alignment.

11. Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences.

12. Report every match whose expect score is lower than a threshold parameter E.

In Bioinformatics, **BLAST** for **B**asic **L**ocal **A**lignment **S**earch **T**ool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. The BLAST algorithm and program were designed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman at the National Institutes of Health and was published in the Journal of Molecular Biology in 1990 and cited over 50,000 times.

Background

BLAST is one of the most widely used bioinformatics programs for sequence searching. It addresses a fundamental problem in bioinformatics research. The heuristic algorithm it uses is much faster than other approaches, such as calculating an optimal alignment. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

Before BLAST, FASTA was developed by David J. Lipman and William R. Pearson in 1985.

Before fast algorithms such as BLAST and FASTA were developed, doing database searches for protein or nucleic sequences was very time consuming because a full alignment procedure (e.g., the Smith–Waterman algorithm) was used.

While BLAST is faster than any Smith-Waterman implementation for most cases, it cannot "guarantee the optimal alignments of the query and database sequences" as Smith-Waterman algorithm does. The optimality of Smith-Waterman "ensured the best performance on accuracy

and the most precise results" at the expense of time and computer power.
BLAST is more time-efficient than FASTA by searching only for the more significant patterns in the sequences, yet with comparative sensitivity. This could be further realized by understanding the algorithm of BLAST introduced below.

Examples of other questions that researchers use BLAST to answer are:

- Which bacterial species have a protein that is related in lineage to a certain protein with known amino-acid sequence
- What other genes encode proteins that exhibit structures or motifs such as ones that have just been determined

BLAST is also often used as part of other algorithms that require approximate sequence matching.

The BLAST algorithm and the computer program that implements it were developed by Stephen Altschul, Warren Gish, and David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Webb Miller at the Pennsylvania State University, and Gene Myers at the University of Arizona. It is available on the web on the NCBI website. Alternative implementations include AB-BLAST (formerly known as WU-BLAST), FSA-BLAST (last updated in 2006), and ScalaBLAST.

Input

Input sequences (in FASTA or Genbank format) and weight matrix.

Output

BLAST output can be delivered in a variety of formats. These formats include HTML, plain text, and XML formatting. For NCBI's web-page, the default format for output is HTML. When performing a BLAST on NCBI, the results are given in a graphical format showing the hits found, a table showing sequence identifiers for the hits with scoring related data, as well as alignments for the sequence of interest and the hits received with corresponding BLAST scores for these. The easiest to read and most informative of these is probably the table.

If one is attempting to search for a proprietary sequence or simply one that is unavailable in databases available to the general public through sources such as NCBI, there is a BLAST

program available for download to any computer, at no cost. This can be found at BLAST+ executables. There are also commercial programs available for purchase. Databases can be found from the NCBI site, as well as from Index of BLAST databases (FTP).

Process

Using a heuristic method, BLAST finds similar sequences, by locating short matches between the two sequences. This process of finding similar sequences is called seeding. It is after this first match that BLAST begins to make local alignments. While attempting to find similarity in sequences, sets of common letters, known as words, are very important. For example, suppose that the sequence contains the following stretch of letters, GLKFA. If a BLAST was being conducted under normal conditions, the word size would be 3 letters. In this case, using the given stretch of letters, the searched words would be GLK, LKF, KFA. The heuristic algorithm of BLAST locates all common three-letter words between the sequence of interest and the hit sequence or sequences from the database. This result will then be used to build an alignment. After making words for the sequence of interest, the rest of the words are also assembled. These words must satisfy a requirement of having a score of at least the threshold T, when compared by using a scoring matrix. One commonly used scoring matrix for BLAST searches is BLOSUM62, although the optimal scoring matrix depends on sequence similarity. Once both words and neighborhood words are assembled and compiled, they are compared to the sequences in the database in order to find matches. The threshold score T determines whether or not a particular word will be included in the alignment. Once seeding has been conducted, the alignment which is only 3 residues long, is extended in both directions by the algorithm used by BLAST. Each extension impacts the score of the alignment by either increasing or decreasing it. If this score is higher than a pre-determined T, the alignment will be included in the results given by BLAST. However, if this score is lower than this pre-determined T, the alignment will cease to extend, preventing the areas of poor alignment from being included in the BLAST results. Note that increasing the T score limits the amount of space available to search, decreasing the number of neighborhood words, while at the same time speeding up the process of BLAST.

Program

The BLAST program can either be downloaded and run as a command-line utility "blastall" or accessed for free over the web. The BLAST web server, hosted by the NCBI, allows anyone with a web browser to perform similarity searches against constantly updated databases of proteins and DNA that include most of the newly sequenced organisms.

The BLAST program is based on an open-source format, giving everyone access to it and

enabling them to have the ability to change the program code. This has led to the creation of several BLAST "spin-offs".

There are now a handful of different BLAST programs available, which can be used depending on what one is attempting to do and what they are working with. These different programs vary in query sequence input, the database being searched, and what is being compared. These programs and their details are listed below:

BLAST is actually a family of programs (all included in the blastall executable). These include

Nucleotide-nucleotide BLAST (blastn)

This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.

Protein-protein BLAST (blastp)

This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.

Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)

The program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarises significant features present in these sequences. A query against the protein database is then run using this profile, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is repeated.

By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.

Nucleotide 6-frame translation-protein (blastx)

This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)

This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.

Protein-nucleotide 6-frame translation (tblastn)

This program compares a protein query against the all six reading frames of a nucleotide sequence database.

Large numbers of query sequences (megablast)

When comparing large numbers of input sequences via the command-line BLAST, "megablast"

is much faster than running BLAST multiple times. It concatenates many input sequences together to form a large sequence before searching the BLAST database, then post-analyzes the search results to glean individual alignments and statistical values.

Of these programs, BLASTn and BLASTp are the most commonly used because they use direct comparisons, and do not require translations. However, since protein sequences are better conserved evolutionarily than nucleotide sequences, tBLASTn, tBLASTx, and BLASTx, produce more reliable and accurate results when dealing with coding DNA. They also enable one to be able to directly see the function of the protein sequence, since by translating the sequence of interest before searching often gives you annotated protein hits.

FASTA

FASTA is a DNA and protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.

The original FASTP program was designed for protein sequence similarity searching. FASTA added the ability to do DNA:DNA searches, translated protein:DNA searches, and also provided a more sophisticated shuffling program for evaluating statistical significance. There are several programs in this package that allow the alignment of protein sequences and DNA sequences.

FASTA is pronounced "fast A", and stands for "FAST-All", because it works with any alphabet, an extension of "FAST-P" (protein) and "FAST-N" (nucleotide) alignment.

The current FASTA package contains programs for protein:protein, DNA:DNA, protein:translated DNA (with frameshifts), and ordered or unordered peptide searches. Recent versions of the FASTA package include special translated search algorithms that correctly handle frameshift errors (which six-frame-translated searches do not handle very well) when comparing nucleotide to protein sequence data.

In addition to rapid heuristic search methods, the FASTA package provides SEARCH, an implementation of the optimal Smith-Waterman algorithm. A major focus of the package is the calculation of accurate similarity statistics, so that biologists can judge whether an alignment is likely to have occurred by chance, or whether it can be used to infer homology. The FASTA package is available from fasta.bioch.virginia.edu. The web-interface to submit sequences for

running a search of the European Bioinformatics Institute (EBI)'s online databases is also available using the FASTA programs.

The FASTA file format used as input for this software is now largely used by other sequence database search tools (such as BLAST) and sequence alignment programs (Clustal, T-Coffee, etc.).

FASTA takes a given nucleotide or amino acid sequence and searches a corresponding sequence database by using local sequence alignment to find matches of similar database sequences.

The FASTA program follows a largely heuristic method which contributes to the high speed of its execution. It initially observes the pattern of word hits, word-to-word matches of a given length, and marks potential matches before performing a more time-consuming optimized search using a Smith-Waterman type of algorithm.

The size taken for a word, given by the parameter kmer, controls the sensitivity and speed of the program. Increasing the kmer value decreases number of background hits that are found. From the word hits that are returned the program looks for segments that contain a cluster of nearby hits. It then investigates these segments for a possible match.

There are some differences between fastn and fastp relating to the type of sequences used but both use four steps and calculate three scores to describe and format the sequence similarity results. These are:

Identify regions of highest density in each sequence comparison. Taking a kmer to equal 1 or 2.

In this step all or a group of the identities between two sequences are found using a look up table. The kmer value determines how many consecutive identities are required for a match to be declared. Thus the lesser the kmer value: the more sensitive the search. kmer=2 is frequently taken by users for protein sequences and kmer=4 or 6 for nucleotide sequences. Short oligonucleotides are usually run with kmer= 1. The program then finds all similar **local regions**, represented as diagonals of a certain length in a dot plot, between the two sequences by counting kmer matches and penalizing for intervening mismatches. This way, **local regions** of highest density matches in a diagonal are isolated from background hits. For protein sequences BLOSUM50 values are used for scoring kmer matches. This ensures that groups of identities with high similarity scores contribute more to the local diagonal score than to identities with low similarity scores. Nucleotide sequences use the identity matrix for the same

purpose. The best 10 local regions selected from all the diagonals put together are then saved.

Rescan the regions taken using the scoring matrices. trimming the ends of the region to include only those contributing to the highest score.

Rescan the 10 regions taken. This time use the relevant scoring matrix while rescoring to allow runs of identities shorter than the kmer value. Also while rescoring conservative replacements that contribute to the similarity score are taken. Though protein sequences use the BLOSUM50 matrix, scoring matrices based on the minimum number of base changes required for a specific replacement, on identities alone, or on an alternative measure of similarity such as PAM, can also be used with the program. For each of the diagonal regions rescanned this way, a subregion with the maximum score is identified. The initial scores found in step1 are used to rank the library sequences. The highest score is referred to as init1 score.

In an alignment if several initial regions with scores greater than a CUTOFF value are found, check whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined regions penalising for each gap 20 points. This initial similarity score (initn) is used to rank the library sequences. The score of the single best initial region found in step 2 is reported (init1). Here the program calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. This optimal alignment of initial regions can be rapidly calculated using a dynamic programming algorithm. The resulting score initn is used to rank the library sequences. This joining process increases sensitivity but decreases selectivity. A carefully calculated cut-off value is thus used to control where this step is implemented, a value that is approximately one standard deviation above the average score expected from unrelated sequences in the library. A 200-residue query sequence with kmer 2 uses a value 28.

This step uses a banded Smith-Waterman algorithm to create an optimised score (opt) for each alignment of query sequence to a database (library) sequence. It takes a band of 32 residues centered on the init1 region of step2 for calculating the optimal alignment. After all sequences are searched the program plots the initial scores of each database sequence in a histogram, and calculates the statistical significance of the "opt" score. For protein sequences, the final alignment is produced using a full Smith-Waterman alignment. For DNA sequences, a banded alignment is provided.

FASTA cannot remove low complexity regions before aligning the sequences as it is possible with BLAST. This might be problematic as when the query sequence contains such regions, e.g. mini- or microsatellites repeating the same short sequence frequent times, this increases the score of not familiar sequences in the database which only match in this repeats, which occur quite frequently. Therefore the program PRSS is added in the FASTA distribution package. PRSS shuffles the matching sequences in the database either on the one-letter level or it shuffles short segments which length the user can determine. The shuffled sequences are now aligned again and if the score is still higher than expected this is caused by the low complexity regions being mixed up still mapping to the query. By the amount of the score the shuffled sequences still attain PRSS now can predict the significance of the score of the original sequences. The higher the score of the shuffled sequences the less significant the matches found between original database and query sequence.

The FASTA programs find regions of local or global similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like BLAST, FASTA can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Multiple Sequence Alignment (MSA)

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex. Most multiple sequence alignment programs use

heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive.

1 Dynamic programming - T-Coffee

2 Progressive alignment construction- hill-climbing algorithm, ClustalW

3 Iterative methods - MUSCLE (multiple sequence alignment by log-expectation)

4 Hidden Markov models- HMMER

6 Genetic algorithms and simulated annealing

7 Phylogeny methods

8. Motif finding

AMINO-ACID PROPENSITIES

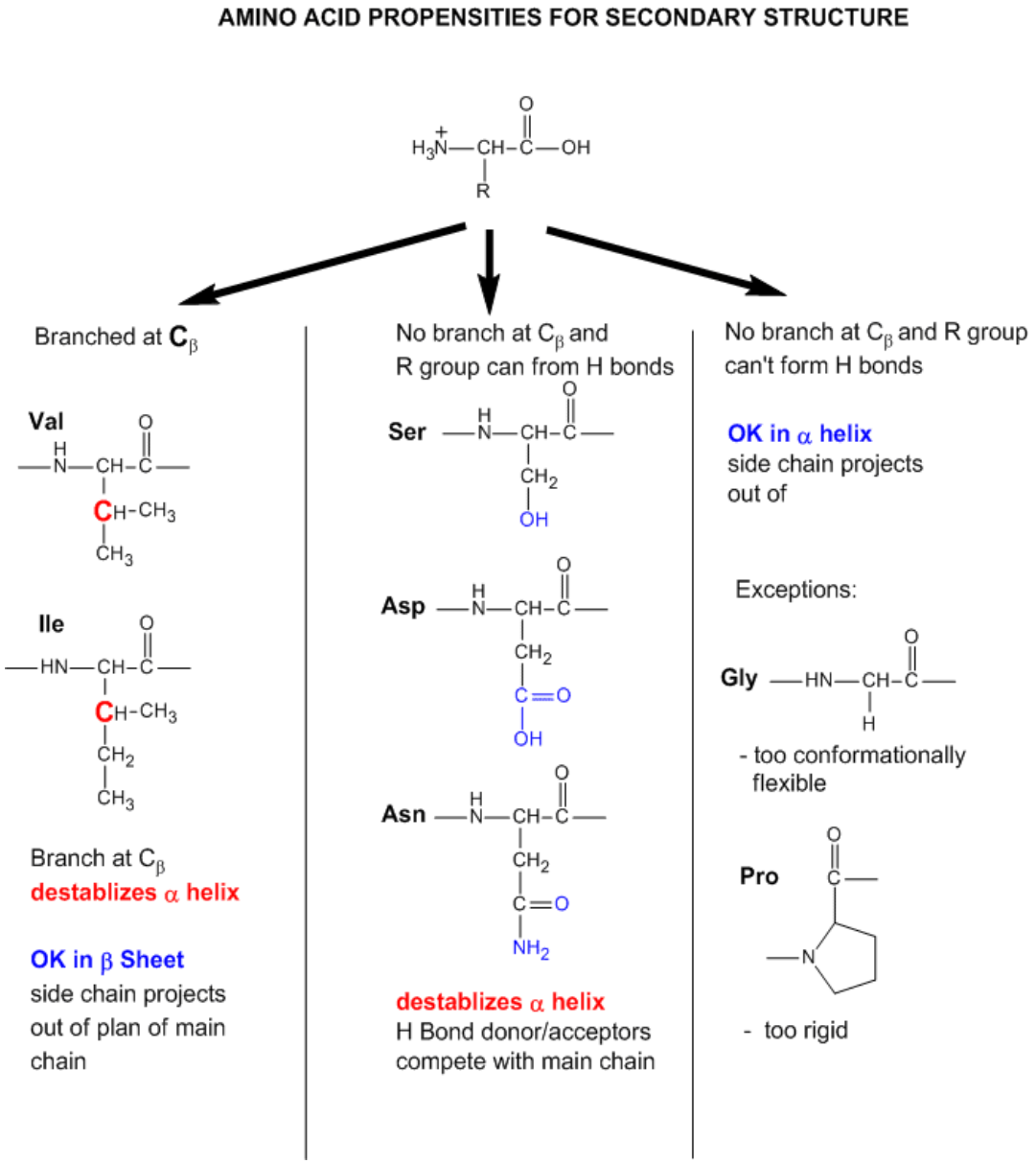


Figure 2

Amino acids vary in their propensity to be found in alpha helices, beta strands, or reverse turns (beta bends, beta turns). These difference can be rationalized from the structure of each amino acid, as described before.

Standard amino acid alpha-helical propensities

Estimated differences in free energy, $\Delta(\Delta G)$, estimated in kcal/mol per residue in an alpha- helical configuration, relative to Alanine arbitrarily set as zero. Higher numbers (more positive free energies) are less favoured. Significant deviations from these average numbers are possible, depending on the identities of the neighbouring residues

Amino Acid	3-Letter	1-Letter	Helical Penalty (Kcal/mol)
Alanine	Ala	A	0
Arginine	Arg	R	0.21
Asparagine	Asn	N	0.65
Aspartic acid	Asp	D	0.69
Cysteine	Cys	C	0.68
Glutamic acid	Glu	E	0.40
Glutamine	Gln	Q	0.39
Glycine	Gly	G	1.00
Histidine	His	H	0.61
Isoleucine	Ile	I	0.41
Leucine	Leu	L	0.21
Lysine	Lys	K	0.26
Methionine	Met	M	0.24
Phenylalanine	Phe	F	0.54
Proline	Pro	P	3.16
Serine	Ser	S	0.50
Threonine	Thr	T	0.66
Tryptophan	Trp	W	0.49
Tyrosine	Tyr	Y	0.53
Valine	Val	V	0.61

Table 2

Chou-Fasman is one of most commonly used algorithms

- Based on observation that certain amino acids tend to be enriched (or depleted) in different secondary structure classes

- The Chou, Fasman and co-workers calculated the propensity of each amino acid to adopt an alpha helix, beta-strand, or coil conformation (later turn as well)

- Propensity values (Pα or Pβ) were calculated using a database of experimentally determined protein structures (first database had 15 proteins)

- Propensity for an amino acid of type i to form an %-helix (P%) is calculated using the following formula:

$$P_{\alpha}^{i} = \frac{\text{fraction of residues of type } i \text{ in } \alpha\text{-helix}}{\text{fraction of all residues in } \alpha\text{-helix}}$$

Example calculation for %-helix propensity (P%) for alanine (A):

- Data from protein structure database (29 proteins):
- Number of alanine residues in the database = 434
- Number of alanine that occur in alpha helix = 234
- Number of all residues in the database = 4741
- Number of all residues that occur in alpha helix = 1798

$$P_{\alpha}^{A} = \frac{234/434}{1798/4741} = \frac{0.539}{0.379} = 1.42$$

- $P_{\alpha}^{i} > 1$ = helix former; $P_{\alpha}^{i} \sim 1$ = indifferent; $P_{\alpha}^{i} < 1$ = helix breaker
- Beta strand propensities are calculated in a similar way:

$$P_{\beta}^{i} = \frac{\text{fraction of residues of type } i \text{ in beta strand}}{\text{fraction of all residues in beta strand}}$$

§ $P_{\beta}^{i} > 1$ = strand former; $P_{\beta}^{i} \sim 1$ = indifferent; $P_{\beta}^{i} < 1$ = strand breaker

P_α			P_β		
E	1.51	H_α	V	1.70	H_β
M	1.45		I	1.60	
A	1.42		Y	1.47	
L	1.21		F	1.38	h_β
K	1.16	h_α	W	1.37	
F	1.13		L	1.30	
Q	1.11		C	1.19	
W	1.08		T	1.19	
I	1.08		Q	1.10	
V	1.06		M	1.05	
D	1.01	I_α	R	0.93	i_β
H	1.00		N	0.89	
R	0.98	i_α	H	0.87	
T	0.83		A	0.83	
S	0.77		S	0.75	b_β
C	0.70		G	0.75	
Y	0.69	b_α	K	0.74	
N	0.67		P	0.55	B_β
P	0.57	B_α	D	0.54	
G	0.57		E	0.37	

Table 3 The Chou-Fasman parameters for the 20 common amino acids.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit 4-Protein Engineering and Bioinformatics – SBTA1303

IV PROTEIN STRUCTURE PREDICTION

Proteins are one of the major biological macromolecules performing a variety functions such as enzymatic catalysis, transport, regulation of metabolism, nerve conduction, immune response etc. The three-dimensional structure of a protein is responsible for its function.

Sequence-Structure Gap and the Need for Structure Prediction

With the advent of recombinant DNA technology it has become possible to determine the amino acid sequences of proteins quite rapidly. However, determining the three dimensional structure of proteins is a time consuming task and hence there exists a vast gap between the number of proteins of known amino acid sequence and that of known structures. This is called as the sequence-structure gap. As the knowledge of the 3-D structure of a protein is very essential to understand its function, it is imperative to develop techniques to predict the structure of a protein from its amino acid sequence.

Basis for Structure Prediction:

The classic experiments carried out by C.B. Afinson in the 60's on the enzyme ribonuclease led to the conclusion that the information to specify the 3-D structure of a protein resides in its amino acid sequence. Within the cell a newly synthesized protein chain spontaneously folds into the compact globular structure to perform its function. Thus nature has an algorithm to fold proteins to their native structures. Efforts have been directed for the past four decades to discover nature's algorithm and computational methods have been developed to predict the structure of proteins from their sequences.

Approaches to Structure Prediction

Prediction of protein structures can be classified into two major categories viz.

(i) Prediction of secondary structure and

(ii) Prediction of tertiary (3-D) structure.

Prediction of secondary structure of proteins attempts to locate segments of the polypeptide chain adopting the α-helical or β-strand structure. Regions that are devoid of these regular secondary structural elements are considered to adopt coil conformation.

In tertiary structure prediction, one attempts to predict the three-dimensional structure of a

protein or the native structure. While so far this has remained an elusive goal, different methods have been developed to press forward to the attainment of this goal.

Secondary structure prediction

What?

- Given a protein sequence (primary structure)

GHWIATRGQLIREAYEDYRHFSSECPFIP

- 1 st step in prediction of protein structure.

(C=Coils H=Alpha Helix E=Beta Strands)

CEEEEECHHHHHHHHHHHHCCCHHCCCCCC

- Technique concerned with determination of secondary structure of given polypeptide by locating the Coils Alpha Helix Beta Strands in polypeptide

Why?

- secondary structure —tertiary structure prediction
- Protein function prediction
- Protein classification
- Predicting structural change
- detection and alignment of remote homology between proteins
- on detecting transmembrane regions, solvent-accessible residues, and other important features of molecules
- Detection of hydrophobic region and hydrophilic region

Prediction methods

Chou-Fasman method

- Based on the propensities of different amino acids to adopt different secondary structures
- Predictions are made using a rules-based approach to identify groups of amino acids with shared secondary structure propensities

Garnier, Osguthorpe, Robson (GOR) method

- Statistical method of secondary structure prediction based on information theory &

Bayesian probability

Multiple Sequence Alignment (MSA) methods

- Performs secondary structure prediction on a multiple sequence alignment as opposed to a single protein sequence

Neural network-based methods

- Example: **P**rofile network from **H**ei**d**elberg (PHD)

Chou-Fasman method:

1. Alpha Helix Prediction:

A. Nucleate a helix by scanning for groups of 6 residues with at least 4 helix formers (Hα and hα) and no more than 1 helix breaker (Bα and bα).

- Two Iα residues count as one helix former for nucleating a helix

B. Propagate predicted helix in both directions until reach a four residue window with average propensity (Pα) < 1.0

C. The average propensity (Pα) for a predicted helix must be Pα > 1.03 and Pα > Pβ

2. Beta Strand Prediction:

A. Nucleate a β-strand by scanning for groups of 5 residues with at least 3 strand formers (Hβ and hβ) and no more than 1 strand breaker (B$ and b$).

B. Propagate predicted β-strand in both directions until reach a four residue window with average propensity (Pβ) < 1.0

C. The average propensity (Pβ) for a predicted β-strand must be Pβ > 1.05 and Pβ > Pα

3. Resolving conflicting predictions:

(regions with both α-helix and β-strand assignment)

- If average Pα > average Pβ, then the region is alpha helix
- If average Pβ > average Pα, then the region is beta strand

Chou-Fasman algorithm:

- Later versions of the algorithm included predictions for turns
- The original algorithm contained additional rules about the location of certain residues (e.g., proline) in α-helices and β-strands
- More recent versions of the algorithm have used sequential tetrapeptide average propensities to predict secondary structure

• The propensity values have also been variously recalculated with larger protein data sets (original data sets based on 15 and 29 proteins)

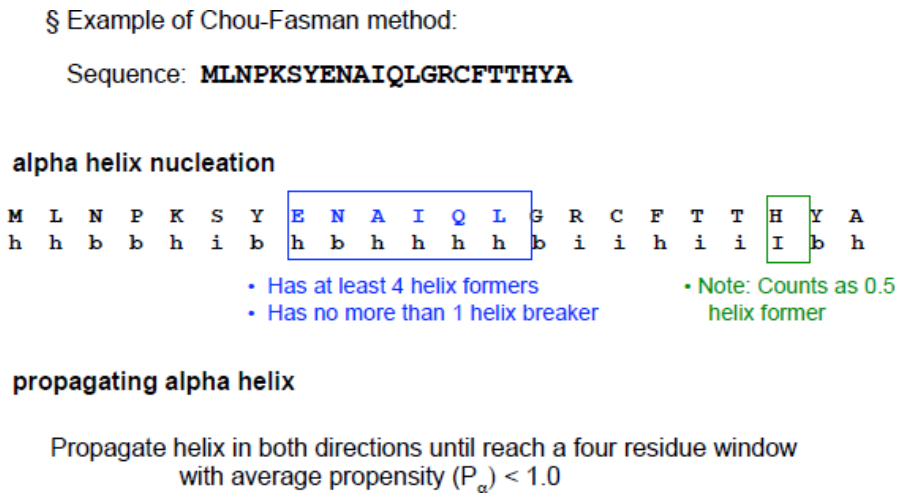


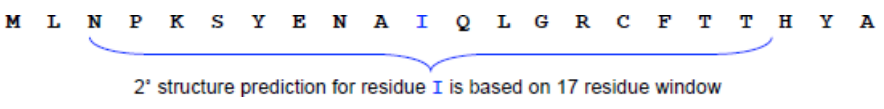
Figure 1

GOR (Garnier,Osguthorpe,Robson) Method

Key difference: Chou-Fasman uses individual amino acid propensities, while GOR incorporates information about neighboring amino acids to make prediction

A 20 x 17 matrix of directional information values for each secondary structure class was calculated from a database of known structures

These matrices are used to predict the secondary structure of the central (9th) residue in a 17 residue window:



The secondary structure class with highest information score over 17 residue window is selected as the prediction for the central residue of the window (e.g., I is predicted to be α-helix)

Multiple sequence alignment method

A multiple sequence alignment arranges protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived from a common ancestor), superposable (in a 3D structural alignment - α helix / β sheet) or play a common functional role (catalytic sites, nuclear localisation signal, protein-protein interaction sites,...). Uses BLAST to identify homologous protein sequence fragments in a protein structure database (PDB).

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTS-NIGS-ITVNWYQQLPG-
LRLS-CSVSGFIFSS-YAMYWVRQAPG
-LS-LTCTVSGTSFDDYYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFN-WYVDG-
A--TLVCTISDFYPGAVTVA-WKADS-
AALGCTVKDYFPEPVTVSWN--SG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

Goal: try to have a maximum of identical/similar residues in a given column of the alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSVSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCTISDFYPGA--VTVAWKADS--
AALGCTVKDYFPEP--VTVSWNSG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

Table 1

Main Criteria for building a multiple sequence alignment	
Criterion	***Meaning***
Structure similarity	**Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	**Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	**Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
Sequence similarity	**Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

What are the applications of multiple sequence alignment

§ Protein structure and function prediction

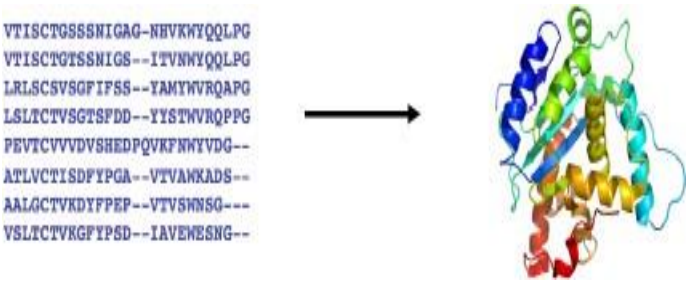


Figure 2

§ Phylogenetic inference

Figure 3

§ Detecting similarities between sequences (closely or distantly related) and conserved regions / motifs in sequences.

§ Detection of structural patterns (hydrophobicity/hydrophilicity, gaps etc), thus assisting improved prediction of secondary and tertiary structures and loops and variable regions.

§ Predict features of aligned sequences like conserved positions which may have structural or functional importance.

§ Computing consensus sequence.

§ Making patterns or profiles that can be further used to predict new sequences falling in a given family.

§ Deriving profiles or Hidden Markov Models that can be used to remove distant sequences (outliers) from protein families.

§ Inferring evolutionary trees / linkage.

How is a multiple sequence alignment used?

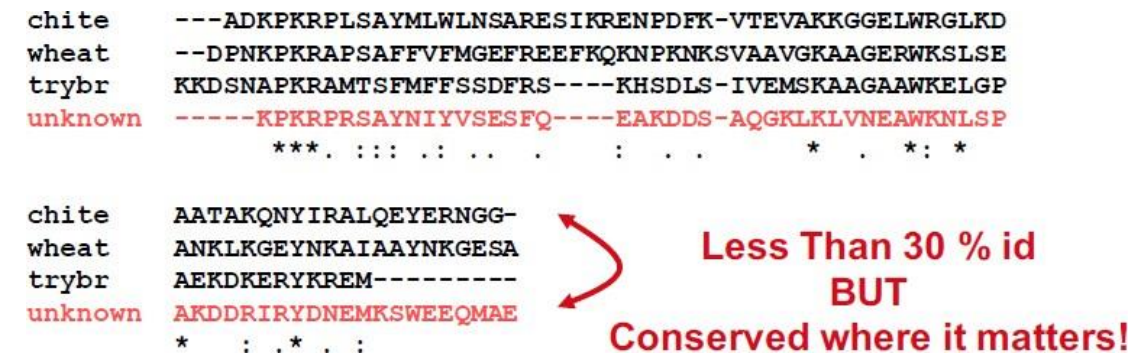


Figure 4

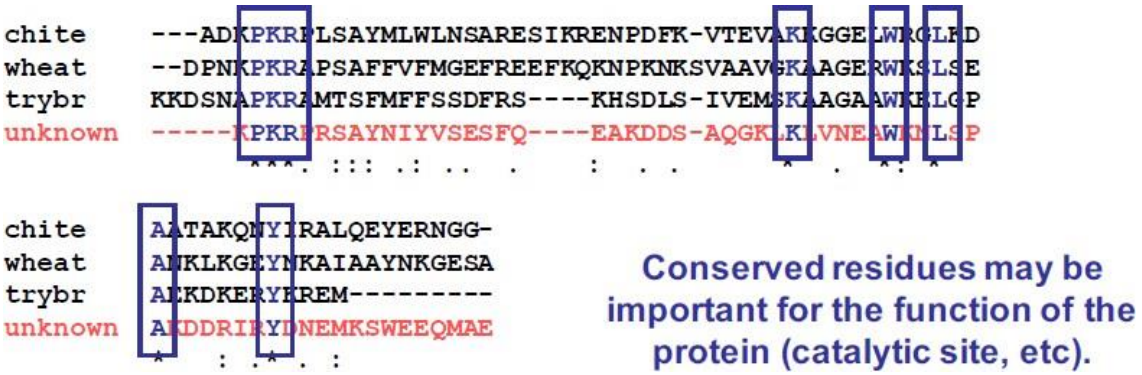


Figure 5

How to score a multiple sequence alignment?

The usual scoring method:

- assumes independance between the columns

$$S(m) = \sum_i S(m_i)$$

S(m) = score of the whole alignment m
$S(m_i)$ = score of column i in this alignmen

- scores each column according a "sum-of-pairs" (SP) function using a substitution scoring matrix.

$$S(m_i) = \sum_{k<l} s\left(m_i^k, m_i^l\right)$$

m_i^k = residue in sequence k in column i
S(a,b) = score from a substitution matrix (PAM or BLOSUM for example)

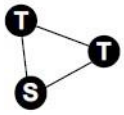
Figure 6

Example

```
123456789...                    L
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSVSGFIFSS--YAMYWVRQAPG
```

$S_6 = s(T,T)+s(T,S)+s(T,S)$

$S_{15} = s(G,-)+s(G,-)+s(-,-)$



A score is calculated for each column, using scoring matrices and gap penalties. Note that here a gap-gap penalty should also be specified.

$S_{alignment} = S_1 + S_2 + S_3 + ... + S_L$

The alignment score is the sum of the column scores.

Neural network secondary structure prediction methods

Artificial neural networks (ANN), with both statistical (linear regression and discriminant analysis) and artificial intelligence roots, are information processing units that that are modeled after the brain and its 100 billion neurons. In a neuron, the distal and proximal dendrites receive signals and communicate to the cell body, which in turn communicates with other neurons via its axon and its terminals.

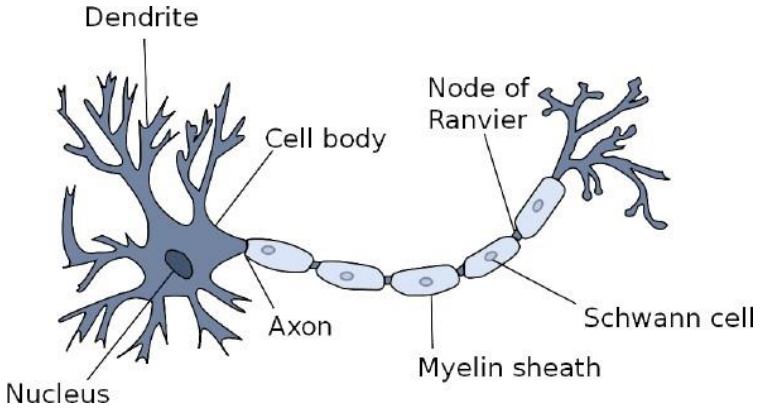


Figure 7

Similarly, an ANN receives inputs (dendrites) that are processed with influence by weights to become outputs (axon).

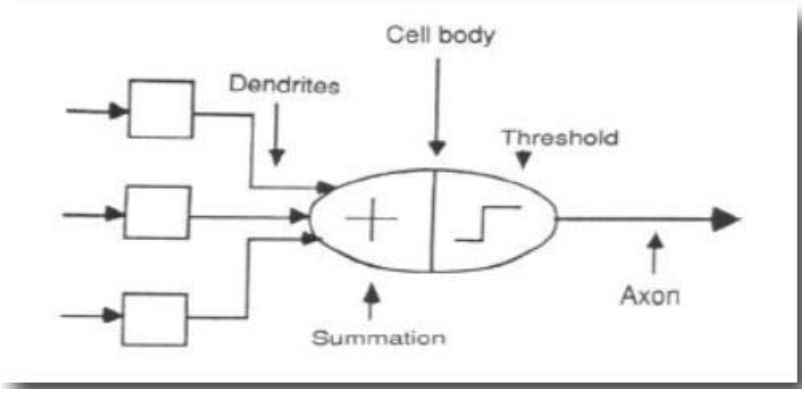


Figure 8

The neurons or nodes interconnect with informational flows (unidirectional or bidirectional) at various weights or strengths. The simplest architecture is the perceptron, which consists of 2 layers (input and output layers) that are separated by a linear discrimination function (10). In a multi-layer perceptron (MLP) model, there are three layers: the input nodes, the hidden nodes layer, and the output nodes.

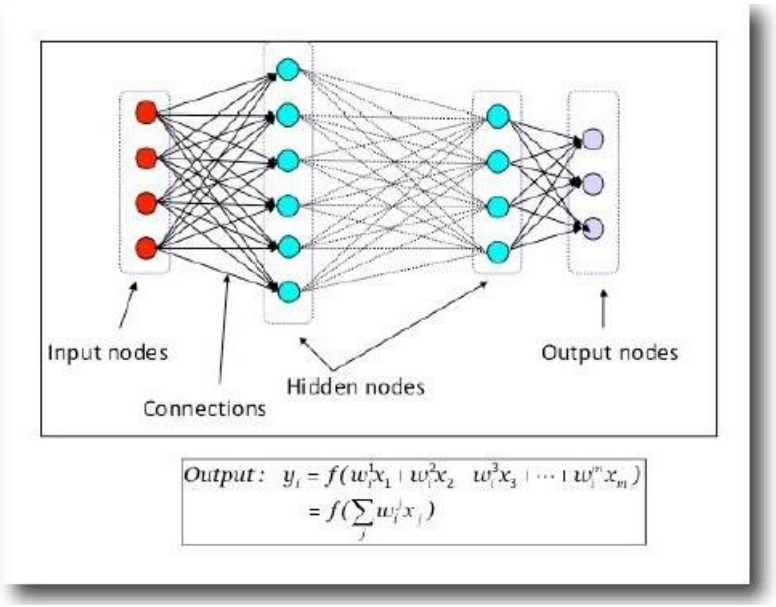


Figure 9

Learning/ Training

In a feed-forward neural network architecture, a unit will receive input from several nodes or neurons belonging to another layer. These highly interconnected neurons therefore form an infrastructure (similar to the biological central nervous system) that is capable of learning by successfully perform pattern recognition and classification tasks. Training of the ANN is a process in which learning occurs from representative data and the knowledge is applied to the new situation.

This training or learning process occurs by arranging the algorithms so that the weights of the ANN are adjusted to lead to the final desired output. The learning in neural networks can be supervised (such as the multilayer perceptron that trained with sets of input data) or unsupervised (such as the Kohonen self-organizing maps which learn by finding patterns). Neural networks can also perform both regression and classification.

The ANN learning process consists of both a forward and a backward propagation process. The forward propagation process involves presenting data into the ANN whereas the important backward propagation algorithm determines the values of the weights for the nodes during a training phase. This latter process is accomplished by directing the errors for input values backwards so that corrections for the weights can be made to minimize the error of actual and desired output data. A recurrent neural network is a series of feed-forward neural networks sharing the same weights and is good for time series data. ANN can therefore extract patterns or detect trends from complicated and imprecise data sets.

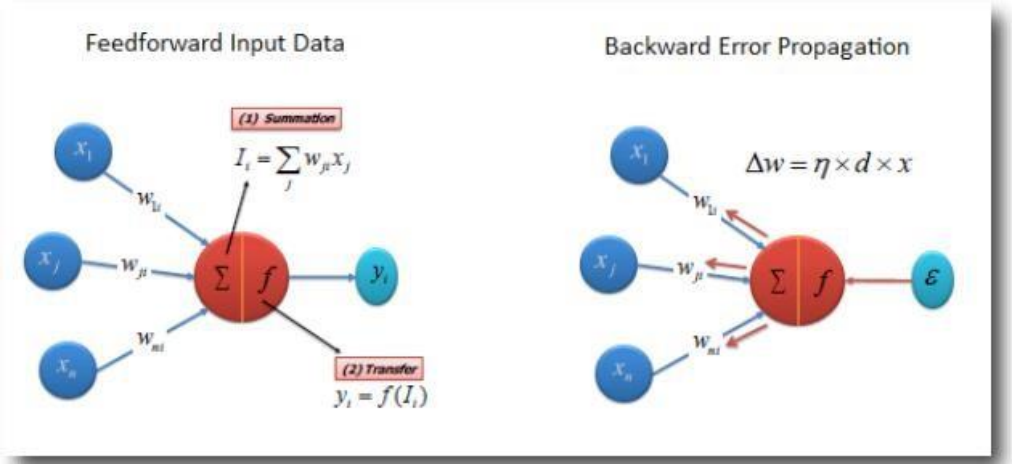


Figure 10

Application of ANN to bioinformatics needs the following strategy:

Extraction of features from molecular sequences to serve as training/prediction data; preprocessing that consists of feature selection and encoding into vectors of real numbers; neural network for training or prediction; post processing that consists of output encoding from the neural network; and finally the myriad of applications (such as sequence analysis, gene expression data analysis, or protein structure prediction).

In secondary structure prediction, neural network methods are trained using sequences with known secondary structure, and then asked to predict the secondary structure of proteins of unknown structure

§ Example: **P**rofile network from **H**ei**d**elberg (PHD) uses multiple sequence alignment with neural network methods to predict secondary structure

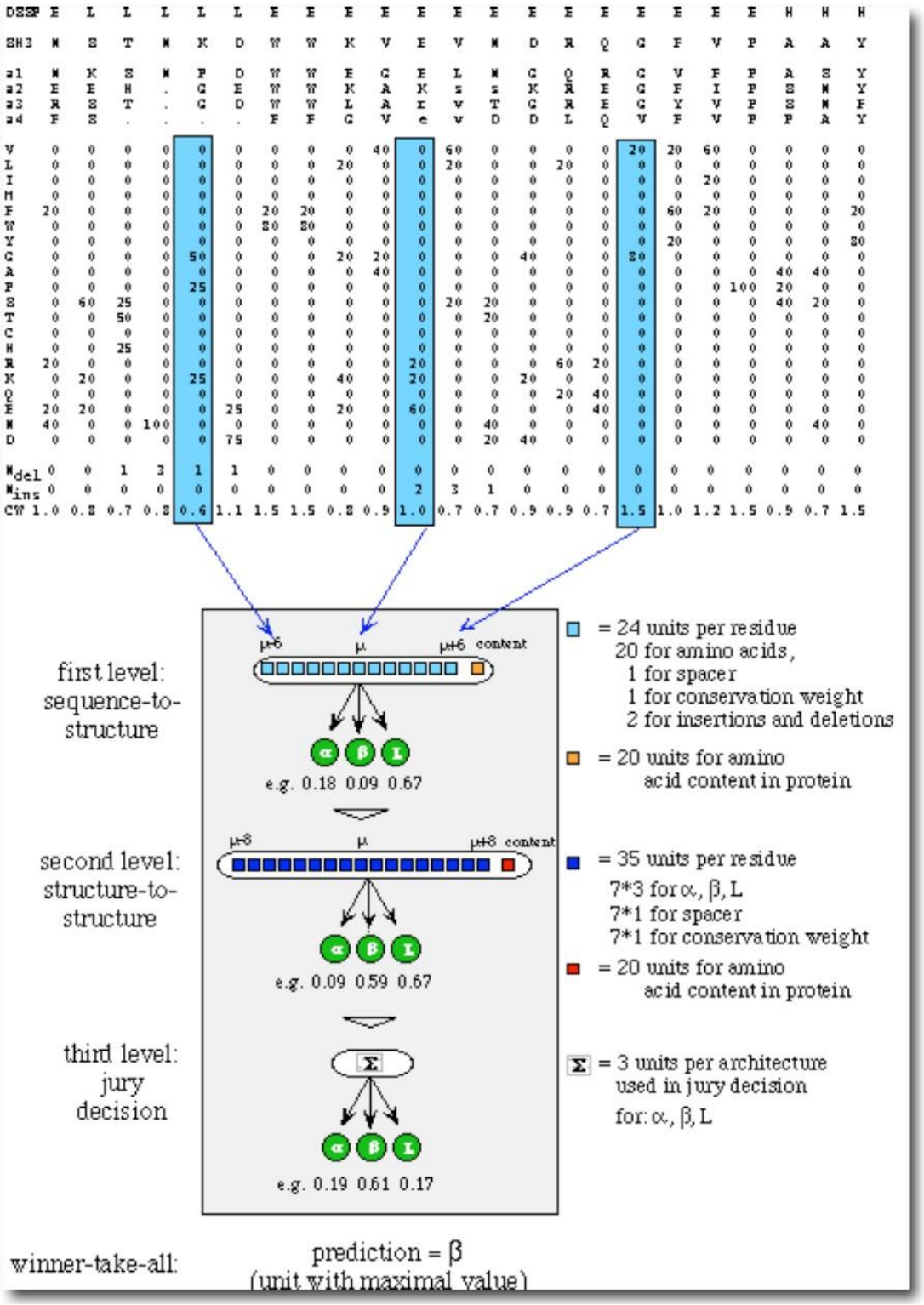


Figure 11

Network architecture (PHD). A profile-based neural network system for protein secondary structure prediction. The multiple alignment is seen at the top with a profile of amino acid occurrences compiled. Then the alignment is fed into the neural network, which consists of 3 layers: 2 network layers and an additional layer for averaging over the independently trained networks

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) andbiotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

Accuracy of Secondary Structure Prediction

§ Prediction accuracy

- Accuracy is usually measured by Q3 (or Qindex) value
- For a single conformation state, i:

$$Q_i = \frac{\text{number of residues correctly predicted in state i}}{\text{number of residues observed in state i}} * 100\%$$

• where i is either helix, strand, or coil. For all three states:

$$Q_3 = \frac{\text{number of residues correctly predicted}}{\text{number of all residues}} * 100\%$$

§ Accuracy of prediction methods

- A random prediction has a Q3 value of ~ 33-38%
- Chou-Fasman method typically has a Q3 ~ 56-60%
- GOR method (depending upon version) has a Q3 ~ 60-65%

- MSA, neural network methods have Q3 ~70%

PROTEIN TERTIARY STRUCTURES: PREDICTION FROM AMINO ACID SEQUENCES

The biological function of a protein is often intimately dependent upon its tertiary structure. X-ray crystallography and nuclear magnetic resonance are the two most mature experimental methods used to provide detailed information about protein structures. However, to date the majority of the proteins still do not have experimentally determined structures available. As at December 2000, there were about 14 000 structures available in the protein data bank (PDB, http://www.pdb.org), and there are about 10 106 000 sequence records sequences in

GenBank (http://www.ncbi.nlm.nih.gov/Genbank). Thus theoretical methods are very important tools to help biologists obtain protein structure information. The goal of theoretical research is not only to predict the structures of proteins but also to understand how protein molecules fold into the native structures. The current methods for protein structure prediction can be roughly divided into three major categories: comparative modelling; threading; and ab initio prediction. For a given target protein with unknown structure, the general procedure for predicting its structure is described below:

Comparative modelling

It is based on two major observations:

1. The structure of a protein is uniquely determined by its amino acid sequence. Knowing the sequence should, at least in theory, suffice to obtain the structure.

2. During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures. This relationship was first identified by Chothia and Lesk (1986) and later quantified by Sander and Schneider (1991). Thanks to the exponential growth of the Protein Data Bank (PDB), Rost (1999) could recently derive a precise limit for this rule, shown in Figure below. As long as the length of two sequences and the percentage of identical residues fall in the region marked as "safe," the two sequences are practically guaranteed to adopt a similar structure

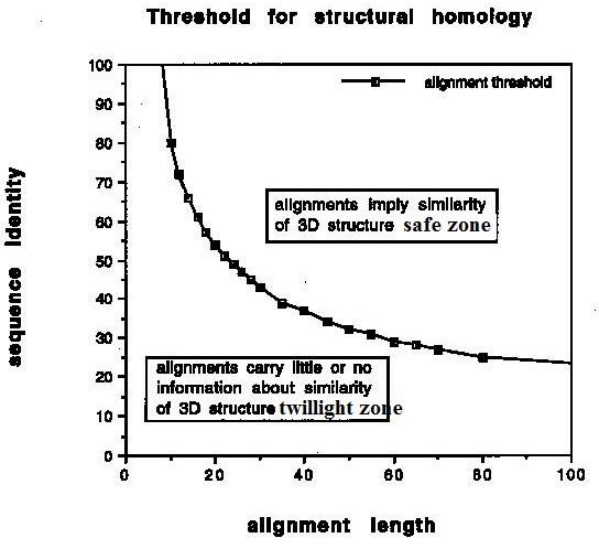


Figure 12

For a sequence of 100 residues, for example, a sequence identity of 40% is sufficient for structure prediction. When the sequence identity falls in the safe homology modeling zone, we can assume that the 3D-structure of both sequences is the same.

The known structure is called the template, the unknown structure is called the target. Homology modeling of the target structure can be done in 7 steps:

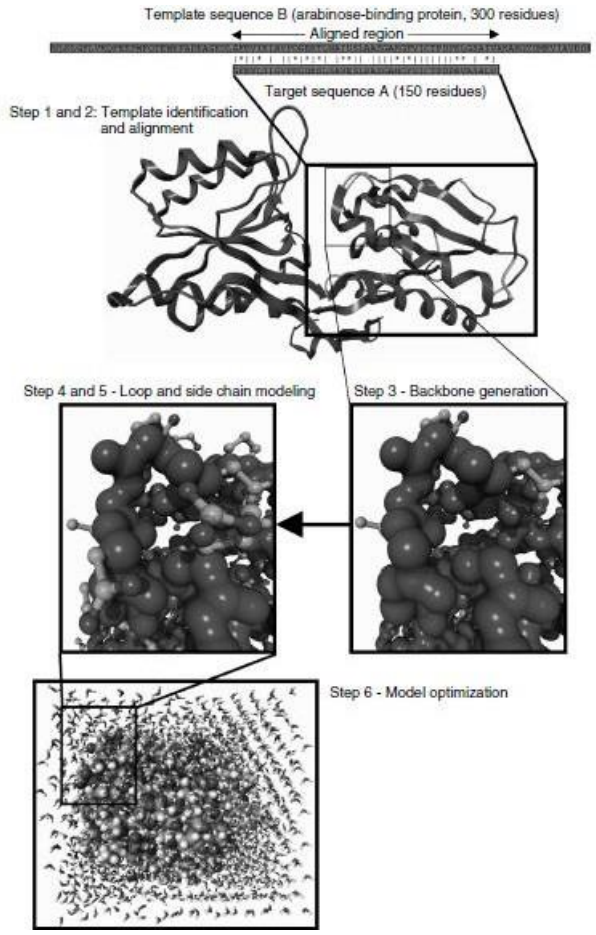
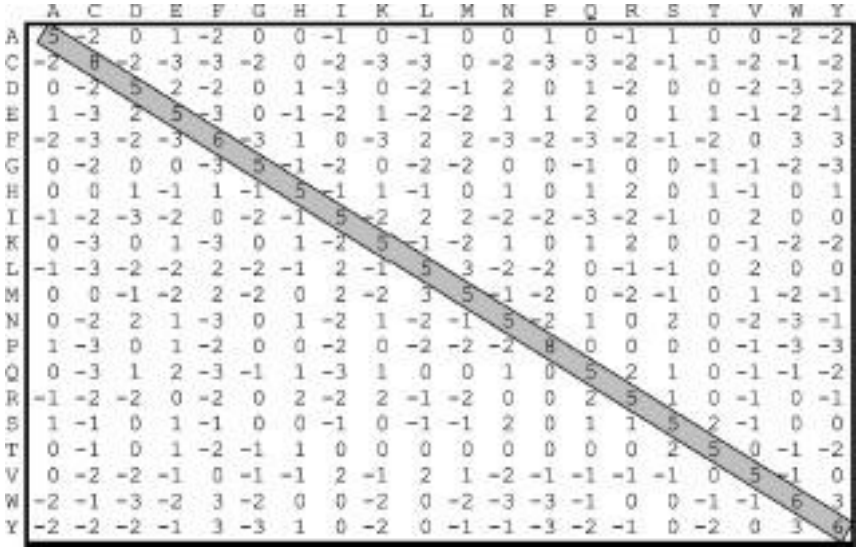


Figure 13

1: Template recognition and initial alignment

In the safe homology modeling zone, the percentage identity between the sequence of interest and a possible template is high enough to be detected with simple sequence alignment programs such as BLAST or FASTA. To identify these hits, the program compares the query sequence to all the sequences of known structures in the PDB using mainly two matrices:

1. A residue exchange matrix (A). The elements of this 20 ∗ 20 matrix define the likelihood that any two of the 20 amino acids ought to be aligned. It is clearly seen that the values along the diagonal (representing conserved residues) are highest, but one can also observe that exchanges between residue types with similar physicochemical properties (for example F → Y) get a better score than exchanges between residue types that widely differ in their properties.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	5	-2	0	1	-2	0	0	-1	0	-1	0	0	1	0	-1	1	0	0	-2	-2
C	-2	8	-2	-3	-3	-2	0	-2	-3	-3	0	-2	-3	-3	-2	-1	-1	-2	-1	-2
D	0	-2	5	2	-2	0	1	-3	0	-2	-1	2	0	1	-2	0	0	-2	-3	-2
E	1	-3	2	5	-3	0	-1	-2	1	-2	-2	1	1	2	0	1	1	-1	-2	-1
F	-2	-3	-2	-3	6	-3	1	0	-3	2	2	-3	-2	-3	-2	-1	-2	0	3	3
G	0	-2	0	0	-3	5	-1	-2	0	-2	-2	0	0	-1	0	0	-1	-1	-2	-3
H	0	0	1	-1	1	-1	5	-1	1	-1	0	1	0	1	2	0	1	-1	0	1
I	-1	-2	-3	-2	0	-2	-1	5	-2	2	2	-2	-2	-3	-2	-1	0	2	0	0
K	0	-3	0	1	-3	0	1	-2	5	-1	-2	1	0	1	2	0	0	-1	-2	-2
L	-1	-3	-2	-2	2	-2	-1	2	-1	5	3	-2	-2	0	-1	-1	0	2	0	0
M	0	0	-1	-2	2	-2	0	2	-2	3	5	-1	-2	0	-2	-1	0	1	-2	-1
N	0	-2	2	1	-3	0	1	-2	1	-2	-1	5	-2	1	0	2	0	-2	-3	-1
P	1	-3	0	1	-2	0	0	-2	0	-2	-2	-2	8	0	0	0	0	-1	-3	-3
Q	0	-3	1	2	-3	-1	1	-3	1	0	0	1	0	5	2	1	0	-1	-1	-2
R	-1	-2	-2	0	-2	0	2	-2	2	-1	-2	0	0	2	5	1	0	-1	0	-1
S	1	-1	0	1	-1	0	0	-1	0	-1	-1	2	0	1	1	5	2	-1	0	0
T	0	-1	0	1	-2	-1	1	0	0	0	0	0	0	0	2	5	0	-1	-2	
V	0	-2	-2	-1	0	-1	-1	2	-1	2	1	-2	-1	-1	-1	-1	0	5	-1	0
W	-2	-1	-3	-2	3	-2	0	0	-2	0	-2	-3	-3	-1	0	0	-1	-1	6	3
Y	-2	-2	-2	-1	3	-3	1	0	-2	0	-1	-1	-3	-2	-1	0	-2	0	3	6

Figure 14

A* A typical residue exchange or scoring matrix used by alignment algorithms. Because the score for aligning residues A and B is normally the same as for B and A, this matrix is symmetric.

2. An alignment matrix (B). The axes of this matrix correspond to the two sequences to align, and the matrix elements are simply the values from the residue exchange matrix for a given pair of residues. During the alignment process, one tries to find the best path through this matrix, starting from a point near the top left, and going down to the bottom right. To make sure that no residue is used twice, one must always take at least one step to the right and one step down. A typical alignment path is shown in Figure B. At first sight, the dashed path in the bottom right corner would have led to a higher score. However, it requires the opening of an additional gap in sequence A (Gly of sequence B is skipped). By comparing thousands of sequences and sequence families, it became clear that the opening of gaps is about as unlikely as at least a couple of nonidentical residues in a row. The jump roughly in the middle of the matrix, however, is justified, because after the jump we earn lots of points (5,6,5), which would have been (1,0,0) without the jump. The alignment algorithm

therefore subtracts an “opening penalty” for every new gap and a much smaller “gap extension penalty” for every residue that is skipped in the alignment. The gap extension penalty is smaller simply because one gap of three residues is much more likely than three gaps of one residue each. In practice, one just feeds the query sequence to one of the countless BLAST servers on the web, selects a search of the PDB, and obtains a list of hits—the modeling templates and corresponding alignments.

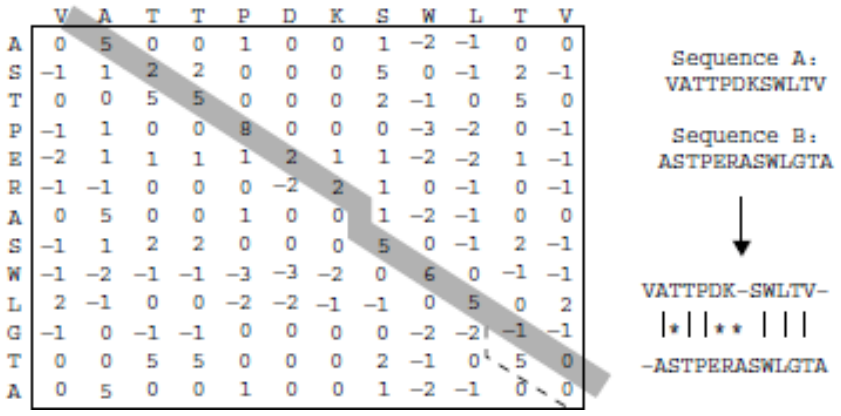


Figure 15

B: The alignment matrix for the sequences VATTPDKSWLTV and ASTPERASWLGTA, using the scores from Figure A. The optimum path corresponding to the alignment on the right side is shown in gray. Residues with similar properties are marked with a star (*). The dashed line marks an alternative alignment that scores more points but requires opening a second gap

2: Alignment correction

Having identified one or more possible modeling templates using the fast methods described above, it is time to consider more sophisticated methods to arrive at a better alignment. Sometimes it may be difficult to align two sequences in a region where the percentage sequence identity is very low.

One can then use other sequences from homologous proteins to find a solution. A pathological example is shown in C:

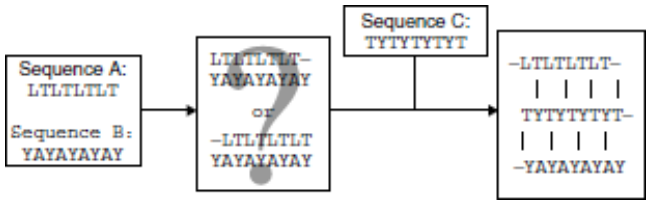


Figure 16

C: A pathological alignment problem. Sequences A and B are impossible to align, unless one considers a third sequence C from a homologous protein.

Suppose you want to align the sequence LTLTLTLT with YAYAYAYAY. There are two equally poor possibilities, and only a third sequence, TYTYTYTYT, that aligns easily to both of them can solve the issue.

The example above introduced a very powerful concept called "multiple sequence alignment." Many programs are available to align a number of related sequences, for example CLUSTALW, and the resulting alignment contains a lot of additional information.

Think about an Ala → Glu mutation. Relying on the matrix in Figure A, this exchange always gets a score of 1. In the 3D structure of the protein, it is however very unlikely to see such an Ala → Glu exchange in the hydrophobic core, but on the surface this mutation is perfectly normal. The multiple sequence alignment implicitly contains information about this structural context. If at a certain position only exchanges between hydrophobic residues are observed, it is highly likely that this residue is buried. To consider this knowledge during the alignment, one uses the multiple sequence alignment to derive position specific scoring matrices, also called profiles. When building a homology model, we are in the fortunate situation of having an almost perfect profile—the known structure of the template. We simply know that a certain alanine sits in the protein core and must therefore not be aligned with a glutamate. Multiple sequence alignments are nevertheless useful in homology modeling, for example, to place deletions (missing residues in the model) or insertions (additional residues in the model) only in areas where the sequences are strongly divergent.

A typical example for correcting an alignment with the help of the template is shown in Figures D and E. Although a simple sequence alignment gives the highest score for the wrong answer (alignment 1 in Fig. D), a simple look at the structure of the template reveals that

alignment 2 is correct, because it leads to a small gap, compared to a huge hole associated with alignment 1.

		1	2	3	4	5	6	7	8	9	10	11	12	13
Template		PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL
Model (bad)	1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model (good)	2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE

Figure 17

D: Example of a sequence alignment where a three-residue deletion must be modeled. While the first alignment appears better when considering just the sequences (a matching proline at position 7), a look at the structure of the template leads to a different conclusion (Figure E)

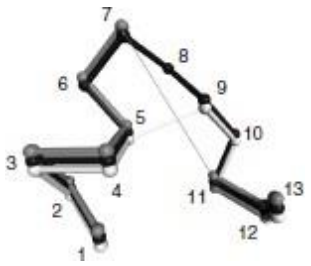


Figure 18

E: Correcting an alignment based on the structure of the modeling template (Cα-trace shown in black). While the alignment with the highest score (dark gray) leads to a gap of 7.5 A between residues 7 and 11, the second option (white) creates only a tiny hole of˚ 1.3 A between residues 5 and 9. This can easily be accommodated by small backbone shifts. (The normal Cα−Cα distance of 3.8 A has been subtracted).

3: Backbone generation

When the alignment is correct, the backbone of the target can be created. The coordinates of the template-backbone are copied to the target. When the residues are identical, the side-chain coordinates are also copied. Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.

4: Loop modeling

Often the alignment will contain gaps as a result of deletions and insertions. When the target sequence contains a gap, one can simply delete the corresponding residues in the template. This creates a hole in the model, this has already been discussed in step 2. When there is an insertion in the target, shown in Figure B, the template will contain a gap and there are no backbone coordinates known for these residues in the model. The backbone from the template has to be cut to insert these residues. Such large changes cannot be modeled in secondary structure elements and therefore have to be placed in loops and strands. Surface loops are, however, flexible and difficult to predict. One way to handle loops is to take some residues before and after the insertion as "anchor" residues and search the PDB for loops with the same anchor-residues. The best loop is simply copied in the model. This is shown in Figure G. The two residues which are colored green in Figure G are used as anchor, the best loop with the inserted resisdues was found in the database and placed in the model.

Figure 19

F: Target sequence (green) with insertion (grey box) results in a gap in the template

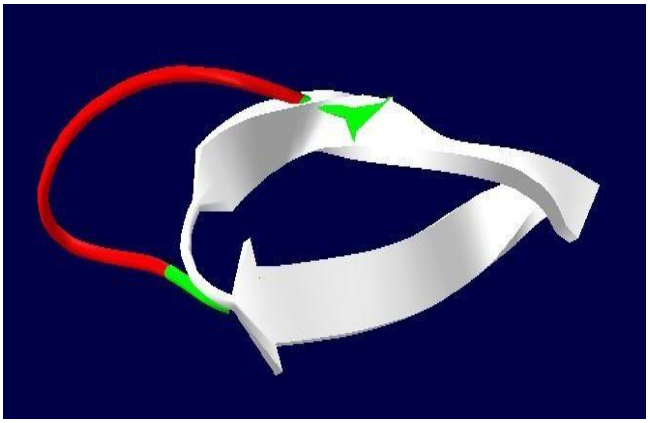


Figure 20

F: The red loop is modeled with the green residues as anchor residues. The insertion of 2 residues results in a longer loop

5: Side-chain modelling

Now it is time to add side-chains to the backbone of the model. Conserved residues were already copied completely. The torsion angle between C-alpha and C-beta of the other residues can also be copied to the model because these rotamers tend to be conserved in similar proteins. It is also possible to predict the rotamer because many backbone configurations strongly prefer a specific rotamer. As shown in Figure G, the backbone of this tyrosine strongly prefers two rotamers and the real side-chain fits in one of them. There are libraries based upon the backbone of the residues flanking the residue of interest. By using these libraries the best rotamer can be predicted. This last method is used by Yasara.

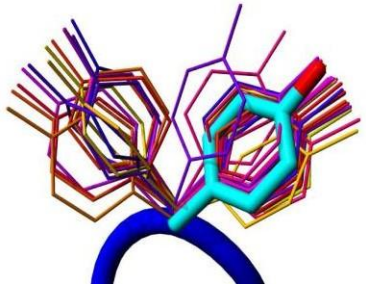


Figure 21

G*:* Prefered rotamers of this tyrosin (colored sticks) the real side-chain (cyan) fits in one of them.

6: Model optimization

The model has to be optimized because many structural artifacts can be introduced while the model protein is being built

- ❑ Substitution of large side chains for small ones
- ❑ Strained peptide bonds between segments taken from difference reference proteins
- ❑ Non optimum conformation of loops

Energy Minimisation is used to produce a chemically and conformationally reasonable model protein structure

Two mainly used optimisation algorithms are

- ➢ Steepest Descent
- ➢ Conjugate Gradients

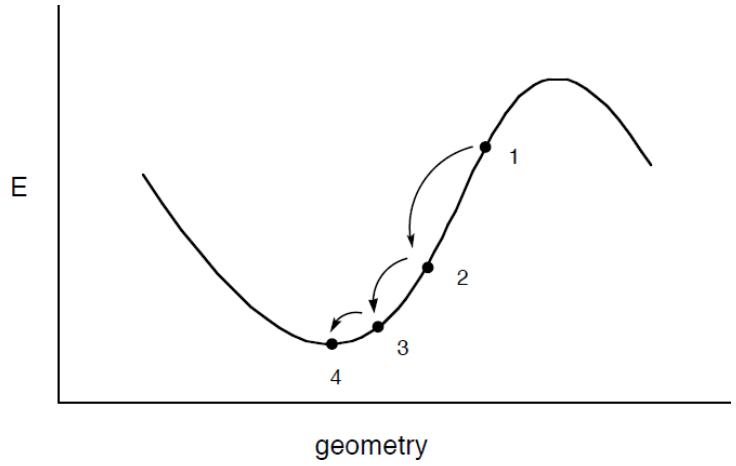


Figure 22

The process of energy minimization changes the geometry of the molecule in a step-wise fashion until a minimum is reached.

Molecular Dynamics is used to explore the conformational space a molecule could visit, Molecular dynamics (MD) is a computer simulation method for studying the physical movements of atoms and molecules

7: Model validation

The models we obtain may contain errors. These errors mainly depend upon two values.

1. The percentage identity between the template and the target.

   If the value is > 90% then accuracy can be compared to crystallography, except for a few individual side chains. If its value ranges between 50-90 % r.m.s.d. error can be as large as 1.5 Å, with considerably more errors. If the value is <25% the alignment turns out to be difficult for homology modeling, often leading to quite larger errors.

2. The number of errors in the template.

   Errors in a model become less of a problem if they can be localized. Therefore, an essential step in the homology modeling process is the verification of the model. The errors can be estimated by calculating the model's energy based on a force field. This method checks to see if the bond lengths and angles are in a normal range. However, this method cannot judge if the model is correctly folded. The 3D distribution functions can also easily identify misfolded proteins and are good indicators of local model building problems.

Modeller

Modeller is a program for comparative protein structure modelling by satisfaction of spatial restraints. It can be described as "Modeling by satisfaction of restraints" uses a set of restraints derived from an alignment and the model is obtained by minimization of these restraints. These restraints can be from related protein structures or NMR experiments. User gives an alignment of sequences to be modelled with known structures. Modeller calculates a model with all non hydrogen atoms. It also performs comparison of protein structures or sequences, clustering of proteins, searching of sequence databases.

THREADING

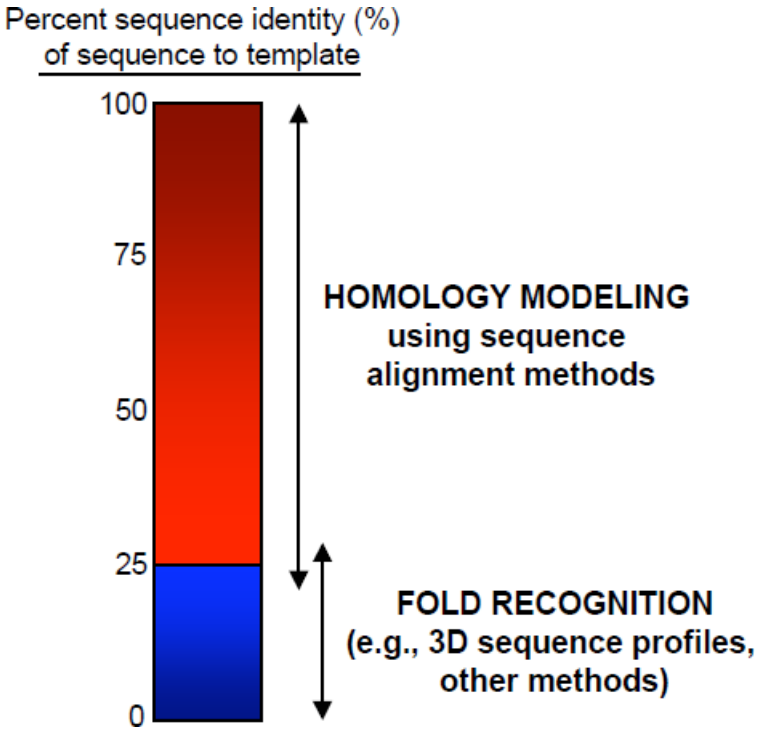


Figure 23

Threading or Fold recognition is a method to identify proteins that have similar 3D structure (fold), but limited or non existent sequence homology. The threading and sequence-structure alignment approachs are based on the observation that many protein structures in the PDB are very similar. For example, there are many 4-helical bundles, TIM barrels, globins, etc. in the set of solved structures.

As a result of this, many scientists have conjectured there are only a limited number of " unique" protein folds in nature. Estimates vary considerably, but some predict that are fewer than 1000 different protein folds. Thus, one approach to the protein structure prediction problem is to try to determine the structure of a new sequence by finding its best fit" to some fold in a library of structures.

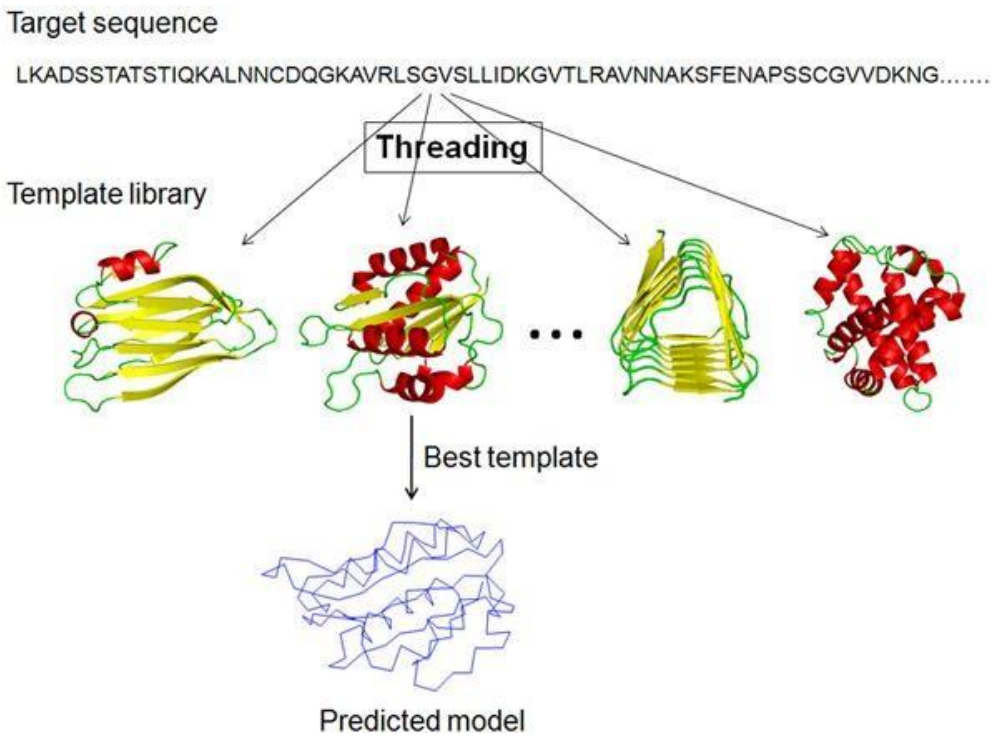


Figure 24

Given a new sequence and a library of known folds, the goal is to _figure out which of the folds (if any) is a good fit to the sequence.

Fold recognition methods include:

• 3D profiles (and protein threading)

- Align sequence to structure

• Profile-based alignment methods that integrate sequence and structural (2D or 3D) information

- e.g., 3D-PSSM or PHYRE software

As a subproblem to fold recognition, we must solve the sequence-structure alignment problem.

Namely, given a solved structure T for a sequence t_1 t_2 ….. t_n = t and a new sequences s_1 s_2 …. s_m = s, we need to find the best match" between s and T. This actually consists of two subproblems:

- Evaluating (scoring) a given alignment of s with a structure T.

- Efficiently searching over possible alignments.

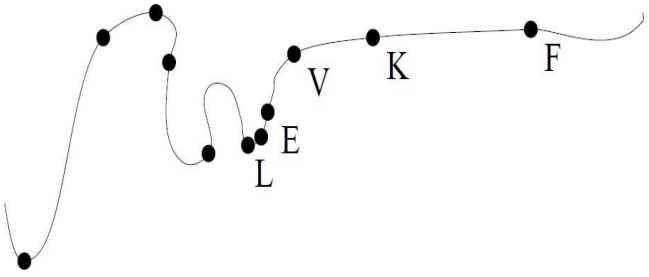


Figure 25

Example: New sequence s=LEVKF, and its best alignment to a particular structure.

There are at least three approaches to the sequence-structure alignment problem.

1. The first method is to just use protein sequence alignment. That is, find the best sequence alignment between the new sequence s and the sequence t with structure T. This is then used to infer the structural alignment: if s_i aligns with t_j , s_i's position in the 3D structure is the same as t_j 's. Scoring in this case is based on amino-acid similarity matrices (e.g., you could use the PAM-250 matrix), and the search algorithm is dynamic programming (O(nm) time). This is a non- physical method; that is, it does not use structural information. The major limitation of this method is that similar structures have lots of sequence variability, and thus sequence alignment may not be very helpful. Hidden Markov model techniques have the same problem.

2. The second method we will describe, the 3D profile method, actually uses structural information. The idea here is that instead of aligning a sequence to a sequence, we align a sequence to a string of descriptors that describe the 3D environment of the target structure. That is, for each residue position in the structure, we determine:

_ how buried it is (buried, partly buried or exposed)

_ the fraction of surrounding environment that is polar (polar or apolar)

_ the local secondary structure (α-helix, β-sheet or other)

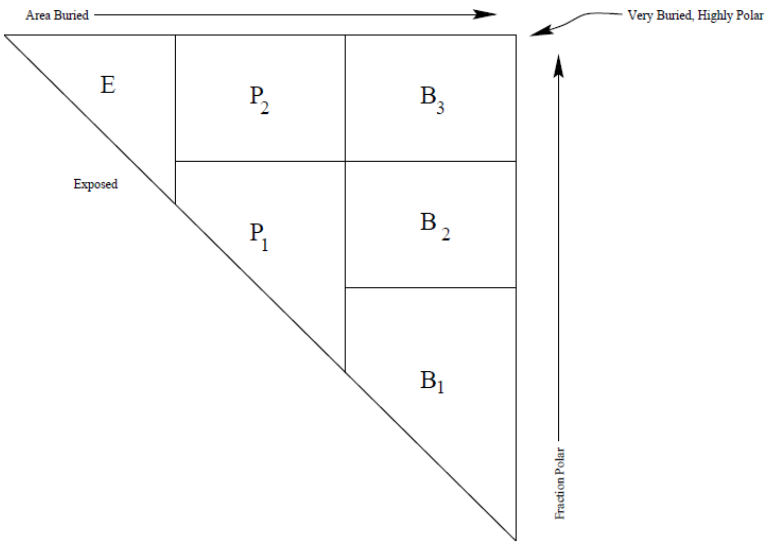


Figure 26

We assign 6 classes of environments to each position in the structure. These environments (E, P1, P2, B1, B2 and B3) depend on the number of surrounding polar residues and how buried the position is. Since there are 3 possible secondary structures for each of these, we have a total of 6 x3 = 18 environment classes.

For each position in the structure, we categorize it into one of 18 environment classes using these characteristics. Because we are using environmental variables, this adds a physical dimension to the problem. The key observation is that different amino acids prefer different environments.

For all proteins in the PDB, we can tabulate the number of times we see a particular residue in a particular environment class, and use this to compute a score for each environment class and each amino acid pair. In particular, we compute a log-odds score of

$$\text{score}_{ij} = \ln\left(\frac{Pr(\text{residue } j \text{ in enviroment } i)}{Pr(\text{residue } j \text{ in any enviroment})}\right)$$

The denominator is obtained from amino acid frequencies present in the PDB. This gives us an 18x20 table as follows:

Table 2

Environment Classes	W	F	Y	⋯
$B_1\alpha$	1.00	1.32	0.18	⋯
$B_1\beta$	1.17	0.85	0.07	⋯
⋮	⋮	⋮	⋮	⋮

Then we can build a 3D profile for a particular structure using this table. Namely, for each position in our structure, we determine its environment class, and the score of a particular amino acid in this position depends on the table we built above.

Thus, for example, if the first position in our structure has environment class B1β, the score of having a tyrosine (Y) in that position is 0.07. Thus, for example, if there are n positions in our structure, we build a table as follows:

Table 3

Position in Fold	Environment Class	W	F	Y	⋯	Gap Penalty
1	$B_1\beta$	1.17	0.85	0.07	⋯	200
2	E loop	-2.14	-1.90	-0.94	⋯	2
⋮	⋮	⋮	⋮			

Then to align a sequence s with a structure, we align the sequence with the descriptors of the 3D environment of the target structure. To find the best alignment, we use a 2D dynamic programming matrix as for regular sequence alignment:

Table 4

	$e_1 e_2 \cdots e_n \leftarrow$ environment classes
s_1	
s_2	
⋮	
s_m	
↑	
new sequence	

Thus, to use the 3D profile method for fold recognition, for a particular sequence we calculate its score (using dynamic programming) for all structures. Signifcance of a score for a particular structure is given by scoring a large sequence database against the structure and calculating

$$z_{-\text{score}} = \frac{\text{score} - \mu}{s}$$

Where μ is the mean score for that structure, and s is the standard deviation of the scores.

The advantages of the 3D profile method over regular sequence alignment is that environmental tendencies may be more informative than simple amino acid similarity, and that structural information is actually used. Additionally, this is a fast method with reasonably good performance. The major disadvantage of this method is that it assumes independence between all positions in the structure.

3. Our third method for sequence-structure alignments uses contact potentials. Most "threading" methods today fall into this category.

Typically, these methods model interactions in a protein structure as a sum over pairwise interactions.

One formalization of the problem is:

Given: a structure P with positions p_1; p_2;…….; p_n, and a sequence s_1; ; sm.

Find: t_1; $t_{2\ldots\ldots}$; tn (where 1< t1 < t2 <_____ < tn ≤ m and t_i indicates the index of the amino acid from s that occupies $_{pi}$) such that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \text{score}\left(i, j, s_{t_i}, s_{t_j}\right)$$

is maximized.

This problem is NP-complete for pairwise interactions. (If the contact graph for a structure is planar, there are approximation algorithms for this problem. However, in practice, they are not used because most graphs would not be planar and heuristics are thought to give better solutions.) One approach commonly used to find threadings is to disallow gaps into core segments (such helices and sheets), and to put lower and upper bounds on distances between core segments. Some algorithms also use exhaustive enumeration and branch and bound techniques to find the best threading. Alternatively, some approaches give up the guarantee of finding the best threading, and use fast heuristics instead.

The score functions come from database-derived pairwise potentials. The general idea is to define a cutoff parameter for contact" (e.g., up to 6 Angstroms), and to use the PDB to count up the number of times amino acids i and j are in contact:

$$\text{score}_{ij} = \ln\left(\frac{Pr(i,j|\text{ contact }\,)}{\text{normalization}}\right).$$

There are several methods to do this normalization. For example, in [2], normalization is by expected frequencies.

Additionally, there are many variations in defining the potentials. For example, in addition to pairwise potentials, some researchers consider single residue potentials as well (e.g., to take into account hydrophobicity or secondary structure), or distance-dependent intervals (e.g., counting up pairwise contacts separately for intervals within 1 Angstrom, between 1 and 2 Angstroms, etc.).

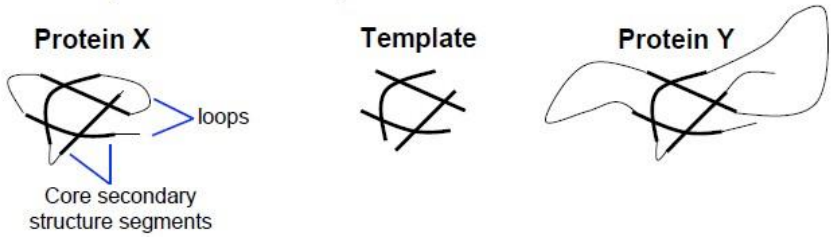
A general paradigm of protein threading consists of the following four steps:

1. Construct a library of core fold templates
2. A scoring (or objective) function is used to evaluate the placement of a sequence in a core template
3. Search for optimal alignments between the sequence and each core fold template
4. Select the core fold template that best aligns (fits) with the protein sequence
    - The 3D model is derived from the optimal alignment (or 'threading') of the sequence to the best scoring structural template

The construction of a structure template database

Select protein structures from the protein structure databases as structural templates. This generally involves selecting protein structures from databases such as PDB, FSSP, SCOP, or CATH, after removing protein structures with high sequence similarities.

§ Construct library of core fold templates:



§ A core fold template is an abstract version of a 3D protein structure that represents the common fold of a family of related protein structures

§ Core templates can include information about interacting or neighboring amino acid positions in the structure

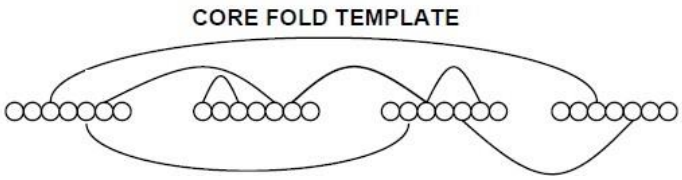


Figure 27

The design of the scoring function

§ Possible sequence/core fold template alignments are scored using a scoring or objective function

§ The scoring/objective function scores the sequence/structure compatibility between a protein sequence and its placement in a core fold template structure

§ The scoring or objective function scores compatibility using parameters such as:

- Amino acid preferences for solvent accessibility
- Amino acid preferences for particular secondary structure

} **Similar to 3D profiles**

- Interactions between neighboring amino acids ('contact' or 'pair' potentials)

Threading alignment

Align the target sequence with each of the structure templates by optimizing the designed scoring function. This step is one of the major tasks of all threading-based structure prediction programs that take into account the pairwise contact potential; otherwise, a dynamic programming algorithm can fulfill it.

§ If interaction terms between neighboring amino acids are not allowed

- Dynamic programming methods will efficiently find the optimal alignment between the sequence and core fold template

§ If interaction terms between neighboring amino acids in the structure are allowed

- Heuristic methods
  - Fast, but may not find optimal alignment
- Exact methods (e.g., branch & bound, Lathrop and Smith (1996) *J. Mol. Biol.* 255:641-665)
  - Will find the optimal alignment, but can take exponential time

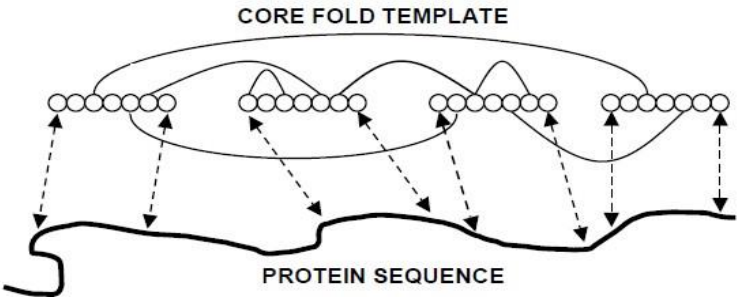


Figure 28

Threading prediction

Select the threading alignment that is statistically most probable as the threading prediction. Then construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template.

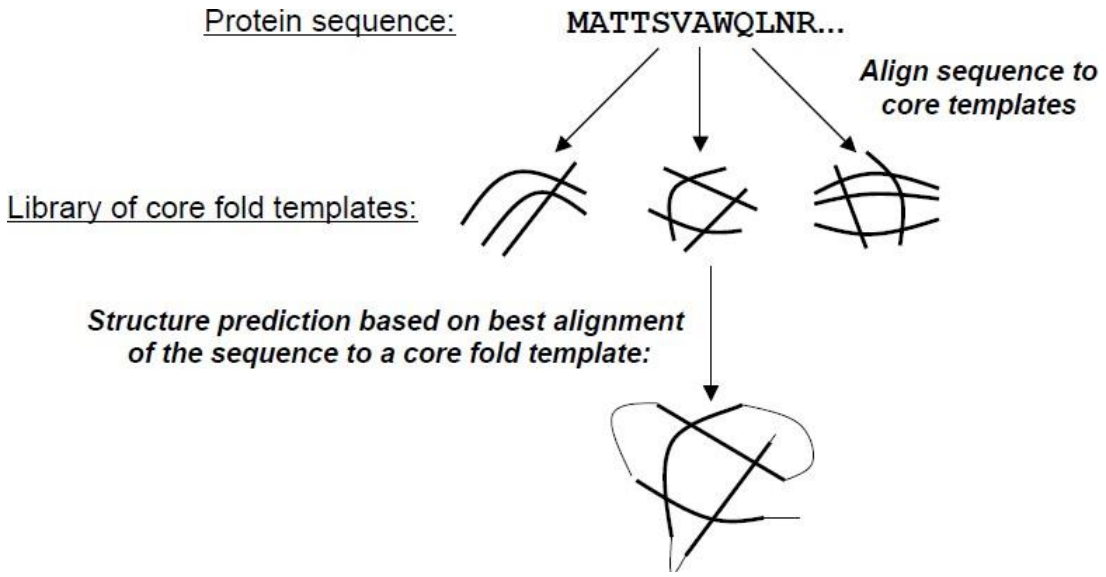


Figure 29

AB INITIO PREDICTION METHOD

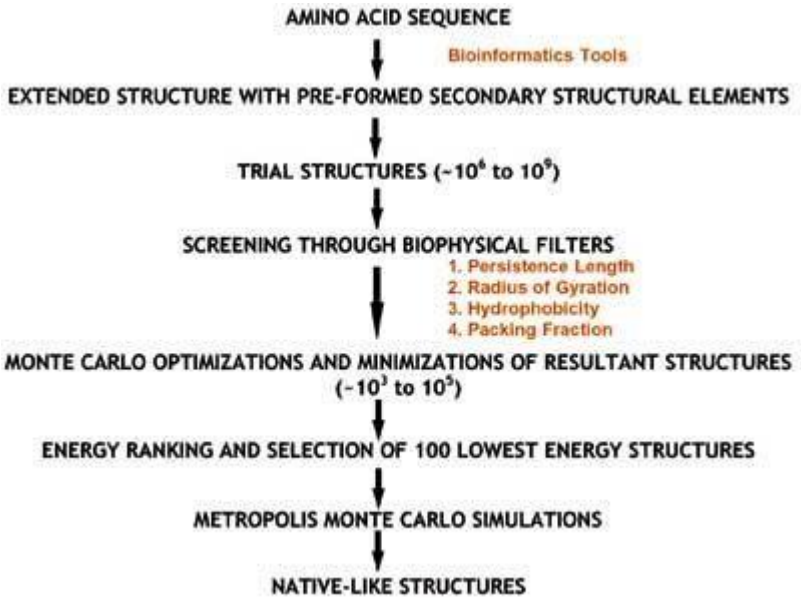


Figure 30

Ab initio, or de novo approaches predict a protein structure and folding mechanism from knowledge only of its amino acid sequence. Often the term ab initio is interpreted as applied to an algorithm based entirely on physico-chemical interactions. On the other hand, the most successful ab initio methods utilize information from the sequence and structural databases in some form. Basic idea of an ab initio algorithm: search for the native state which is presumably in the minimum energy conformation. Usually an ab initio algorithm consists of multiple steps with different levels of approximated modeling of protein structure.

For a consideration of side chains in ab initio predictions, a so-called united residue approximation (UNRES) is frequently used:

- Side chains are represented by spheres ("side-chain centroids", SC). Each centroid represents all the atoms belonging to a real side chain. A van der Waals radius is introduced for every residue type.

- A polypeptide chain is represented by a sequence of Cα atoms with attached SCs and peptide group centers (p) centered between two consecutive Cα atoms.

- The distance between successive Cα atoms is assigned a value of 3.8 Å (a virtual-bond length, characteristic of a planar trans peptide group CO-NH).

- It is assumed that Cα - Cα - Cα virtual bond angles have a fixed value of 90° (close to what is observed in crystal structures). - The united side chains have fixed geometry, with parameters being taken from crystal data.

The only variables in this model of protein conformation are virtual-bond torsional angles γ.

The energy function for the simplified chain can be represented as the sum of the hydrophobic, hydrophilic and electrostatic interactions between side chains and peptide groups (potential functions dependent on the nature of interactions, distances and dimensions of side chains). The parameters in the expressions for contact energies are estimated empirically from crystal structures and all-atom calculations.

An example of the algorithm for structure prediction using UNRES:

1. Low-energy conformations in UNRES approximation are searched using Monte Carlo energy minimization. A cluster analysis is then applied to divide the set of low-energy conformations whose lowest-energy representatives are hereafter referred to as structures. Structures having energies within a chosen cut-off value above the lowest energy structure are saved for further stages of the calculation.

2. These virtual-bond united-residue structures are converted to an all-atom backbone (preserving distances between α-carbons).

3. Generation of the backbone is completed by carrying out simulations in a "hybrid" representation of the polypeptide chain, i.e. with an all-atom backbone and united side chains (still subject to the constraints following the UNRES simulations, so that some or even all the distances of the virtual-bond chain are substantially preserved). The simulations are performed by a Monte Carlo algorithm.

4. Full (all-atom) side chains are introduced with accompanying minimization of steric overlaps, allowing both the backbone and side chains to move. Then Monte Carlo simulations explore conformational space in the neighborhood of each of the low-energy structures.

Monte Carlo algorithms start from some (random) conformation and proceed with (quasi)randomly introduced changes, such as rotations around a randomly selected bond. If the change improves energy value, it is accepted. If not, it may be accepted with a probability dependent on energy increase. The procedure is repeated with a number of iterations, leading to lower energy conformations. A function defining higher energy acceptance probability is usually constructed 25 with a parameter that leads to lower probabilities in the course of simulation ("cooling down" the simulation) in order to achieve convergence and stop the algorithm.

Combinations of approaches

Many of the modern packages for protein structure predictions attempt to combine various approaches, algorithms and features. One of the most successful examples is Rosetta - ab initio prediction using database statistics.

Rosetta is based on a picture of protein folding in which local sequence fragments (3-9 residues) rapidly alternate between different possible local structures. The distribution of conformations sampled by an isolated chain segment is approximated by the distribution adopted by that sequence segment and related sequence segments in the protein structure database. Thus the algorithm combines both ab initio and fold recognition approaches.

Folding occurs when the conformations and relative orientations of the local segments combine to form low energy global structures. Local conformation are sampled from the database of structures and scored using Bayesian logic:

P(structure | sequence) = P(structure) x P(sequence | structure) / P(sequence).

For comparisons of different structures for a given sequence, P(sequence) is constant. P(structure) may be approximated by some general expression favouring more compact structures. P(sequence | structure) is derived from the known structures in the database by assumptions somewhat similar to those used in fold recognition, for instance by estimating probabilities for pairs of amino acids to be at particular distance and computing the probability of sequence as the product over all pairs).

Non-local interactions are optimized by a Monte Carlo search through the set of conformations that can be built from the ensemble of local structure fragments.

In the standard Rosetta protocol, an approximated protein representation is used: backbone

atoms are explicitly included, but side chains are represented by centroids (so-called low-resolution refinement of protein structure). The low-resolution step can be followed by high- resolution refinement, with all-atom protein representation. Similar stepwise refinement protocols can be used to improve predictions yielded by other methods, for instance, in loops (variable regions) of homology-modeling structures.

In recent CASP experiments (Critical Assessment of Structure Prediction), the Rosetta approach turned out to be one of the most successful prediction methods in the novel fold category. Obviously, none of prediction approaches is ideal. Therefore it is reasonable to try to combine the best features of many different procedures or to derive a consensus, meta- prediction. For instance, the 3D-Jury system generated meta predictions using models produced by a set of servers. The algorithm scored various models according to their similarities to each other.

Predictions of coiled coil domains and transmembrane segments

Special algorithms have been developed for domains characterized by special types of interactions. The coiled coil domains are very stable structures formed by regular arrangement of hydrophobic and polar residues in adjacent α- helices. This is possible in the amino acid sequences containing repeats of seven residues (heptads) with hydrophobic residues located at the first and the fourth positions of the heptad and preferences for polar residues at positions 5 and 7. It is possible to design an algorithm that would take into account stabilizing interactions in coiled coils to predict such conformations. Transmembrane proteins contain α-helical segments buried in the membranes. Due to the specific hydrophobic environment in a membrane, protein folding occurs differently as compared to globular proteins folded in the polar water environment. This leads to special folding algorithms, mostly based on known statistics of amino acid frequencies in transmembrane α-helices. Efficient modern algorithms use probabilistic approaches such as Markov models and Bayesian approach.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit 5-Protein Engineering and Bioinformatics – SBTA1303

V PROTEIN ENGINEERING AND APPLICATIONS

Protein engineering is the process of developing useful or valuable proteins. It is a young discipline, with much research taking place into the understanding of protein folding and recognition for protein design principles. It is also a product and services market, with an estimated value of $168 billion by 2017.

There are two general strategies for protein engineering: rational protein design and directed evolution. These methods are not mutually exclusive; researchers will often apply both. In the future, more detailed knowledge of protein structure and function, and advances in high-throughput screening, may greatly expand the abilities of protein engineering. Eventually, even unnatural amino acids may be included, via newer methods, such as expanded genetic code, that allow encoding novel amino acids in genetic code.

APPROACHES

Rational design

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

Directed evolution

In directed evolution, random mutagenesis, e.g. by error-prone PCR or Sequence Saturation Mutagenesis, is applied to a protein, and a selection regime is used to select variants having desired traits. Further rounds of mutation and selection are then applied. This method mimics natural evolution and, in general, produces superior results to rational design. An added process, termed DNA shuffling, mixes and matches pieces of successful variants to produce better results. Such processes mimic the recombination that occurs naturally during sexual reproduction. Advantages of directed evolution are that it requires no prior structural knowledge of a protein, nor is it necessary to be able to predict what effect a given mutation

will have. Indeed, the results of directed evolution experiments are often surprising in that desired changes are often caused by mutations that were not expected to have some effect. The drawback is that they require high-throughput screening, which is not feasible for all proteins. Large amounts of recombinant DNA must be mutated and the products screened for desired traits. The large number of variants often requires expensive robotic equipment to automate the process. Further, not all desired activities can be screened for easily.

Natural Darwinian evolution can be effectively imitated in the lab toward tailoring protein properties for diverse applications, including catalysis. Many experimental technologies exist to produce large and diverse protein libraries and for screening or selecting folded, functional variants. Folded proteins arise surprisingly frequently in random sequence space, an occurrence exploitable in evolving selective binders and catalysts. While more conservative than direct selection from deep sequence space, redesign of existing proteins by random mutagenesis and selection/screening is a particularly robust method for optimizing or altering extant properties. It also represents an excellent starting point for achieving more ambitious engineering goals. Allying experimental evolution with modern computational methods is likely the broadest, most fruitful strategy for generating functional macromolecules unknown to nature.

The main challenges of designing high quality mutant libraries have shown significant progress in the recent past. This progress has been in the form of better descriptions of the effects of mutational loads on protein traits. Also computational approaches have showed large advances in the innumerably large sequence space to more manageable screenable sizes, thus creating smart libraries of mutants. Library size has also been reduced to more screenable sizes by the identification of key beneficial residues using algorithms for systematic recombination. Finally a significant step forward toward efficient reengineering of enzymes has been made with the development of more accurate statistical models and algorithms quantifying and predicting coupled mutational effects on protein functions.

Generally, directed evolution may be summarized as an iterative two step process which involves generation of protein mutant libraries, and high throughput screening processes to select for variants with improved traits. This technique does not require prior knowledge of the protein structure and function relationship. Directed evolution utilizes random or focused

mutagenesis to generate libraries of mutant proteins. Random mutations can be introduced using either error prone PCR, or site saturation mutagenesis. Mutants may also be generated using recombination of multiple homologous genes. Nature has evolved a limited number of beneficial sequences. Directed evolution makes it possible to identify undiscovered protein sequences which have novel functions. This ability is contingent on the proteins ability to tolerant amino acid residue substitutions without compromising folding or stability.

Directed evolution methods can be broadly categorized into two strategies, asexual and sexual methods.

Asexual Methods

Asexual methods no not generate any cross links between parental genes. Single genes are used to create mutant libraries using various mutagenic techniques. These asexual methods can produce either random or focused mutagenesis.

Random Mutagenesis

Random mutagenic methods produce mutations at random throughout the gene of interest. Random mutagenesis can introduce the following types of mutations: transitions, transversions, insertions, deletions, inversion, missense, and nonsense. Examples of methods for producing random mutagenesis are below.

Error Prone PCR

Error prone PCR utilizes the fact that Taq DNA polymerase lacks 3' to 5' exonuclease activity. This results in an error rate of 0.001-0.002% per nucleotide per replication. This method begins with choosing the gene, or the area within a gene, one wishes to mutate. Next, the extent of error required is calculated based upon the type and extent of activity one wishes to generate. This extent of error determines the error prone PCR strategy to be employed. Following PCR, the genes are cloned into a plasmid and introduced to competent cell systems. These cells are then screened for desired traits. Plasmids are then isolated for colonies which show improved traits, and are then used as templates the next round of mutagenesis. Error prone PCR shows biases for certain mutations relative to others. Such as biases for transitions over transversions.

Rates of error in PCR can be increased in the following ways:

1. Increase concentration of magnesium chloride, which stabilizes non complementary base pairing.
2. Add manganese chloride to reduce base pair specificity.
3. Increased and unbalanced addition of dNTPs.
4. Addition of base analogs like dITP, 8 oxo-dGTP, and dPTP.
5. Increase concentration of Taq polymerase.
6. Increase extension time.
7. Increase cycle time.
8. Use less accurate Taq polymerase.

Also see Polymerase chain reaction for more information.

Rolling circle error-prone PCR

This PCR method is based upon rolling circle amplification, which is modeled from the method that bacteria use to amplify circular DNA. This method results in linear DNA duplexes. These fragments contain tandem repeats of circular DNA called concatamers, which can be transformed into bacterial strains. Mutations are introduced by first cloning the target sequence into an appropriate plasmid. Next, the amplification process begins using random hexamer primers and Φ29 DNA polymerase under error prone rolling circle amplification conditions. Additional conditions to produce error prone rolling circle amplification are 1.5 pM of template DNA, 1.5 mM $MnCl_2$ and a 24 hour reaction time. $MnCl_2$ is added into the reaction mixture to promote random point mutations in the DNA strands. Mutation rates can be increased by increasing the concentration of $MnCl_2$, or by decreasing concentration of the template DNA. Error prone rolling circle amplification is advantageous relative to error prone PCR because of its use of universal random hexamer primers, rather than specific primers. Also the reaction products of this amplification do not need to be treated with ligases or endonucleases. This reaction is isothermal.

Chemical mutagenesis

Chemical mutagenesis involves the use of chemical agents to introduce mutations into genetic sequences. Examples of chemical mutagens follow:

Sodium Bisulfate - This mutagenic agent is effective at mutating G/C rich genomic sequences. This is because sodium bisulfate catalyses deamination of unmethylated cytosine to uracil.

Ethyl Methane Sulfonate - This chemical agent alkylates guanidine residues. This alteration causes errors during DNA replication.

Nitrous Acid - This chemical agent causes transversion by de-amination of adenine and cytosine.

The dual approach to random chemical mutagenesis is an iterative two step process. First it involves the *in vivo* chemical mutagenesis of the gene of interest via EMS. Next, the treated gene is isolated and cloning into an untreated expression vector in order to prevent mutations in the plasmid backbone. This technique preserves the plasmids genetic properties.

Targeting Glycosylases to Embedded Arrays for Mutagenesis (TaGTEAM)[ ]

This method has been used to create targeted *in vivo* mutagenesis in yeast. This method involves the fusion of a 3-methyladenine DNA glycosylase to tetR DNA-binding domain. This has been shown to increase mutation rates by over 800 time in regions of the genome containing tetO sites.

Mutagenesis by Random Insertion and Deletion

This method involves alteration in length of the sequence via simultaneous deletion and insertion of chunks of bases of arbitrary length. This method has been shown to produce proteins with new functionalities via introduction of new restriction sites, specific codons, four base codons for non-natural amino acids.

Transposon Based Random Mutagenesis

Recently many methods for transposon based random mutagenesis have been reported. This methods include, but are not limited to the following: PERMUTE-Random Circular Permutation, random protein truncation, random nucleotide triplet substitution, random domain/tag/multiple amino acid insertion, codon scanning mutagenesis, and multicodon scanning mutagenesis. These aforementioned techniques all require the design of mini-Mu transposons. Thermo scientific manufactures kits for the design of these transposons.

Random Mutagenesis Methods Altering the Target DNA Length

These methods involve altering gene length via insertion and deletion mutations. An example is the Tandem Repeat Insertion (TRINS) method. This technique results in the generation of tandem repeats of random fragments of the target gene via rolling circle amplification and concurrent incorporation of these repeats into the target gene.

Mutator strains

Mutator strains are bacterial cell lines which are deficient in one or more DNA repair mechanisms. An example of a mutator strand is the E. coli XL1-RED. This subordinate strain of E. coli is deficient in the MutS, MutD, MutT DNA repair pathways. Use of mutator strains is useful at introducing many types of mutation; however, these strains show progressive sickness of culture because of the accumulation of mutations in the strains own genome.

Focused mutagenesis

Focused mutagenic methods produce mutations at predetermined amino acid residues. These techniques require and understanding of the sequence-function relationship for the protein of interest. Understanding of this relationship allows for the identification of residues which are important in stability, stereoselectivity, and catalytic efficiency. Examples of methods that produce focused mutagenesis are below.

Site saturation mutagenesis

Site saturation mutagenesis is a PCR based method used to target amino acids with significant roles in protein function. The two most common techniques for performing this are whole plasmid single PCR, and overlap extension PCR.

Whole plasmid single PCR is also referred to as site directed mutagenesis (SDM). SDM products are subjected to Dpn endonuclease digestion. This digestion results in cleavage of only the parental strand, because the parental strand contains a GmATC which is methylated at N6 of adenine. SDM does not work well for large plasmids of over ten kilobases. Also, this method is only capable of replacing two nucleotides at a time.

Overlap extension PCR requires the use of two pairs of primers. One primer in each set contains a mutation. A first round of PCR using these primer sets is performed and two double stranded DNA duplexes are formed. A second round of PCR is then performed in which these duplexes are denatured and annealed with the primer sets again to produce heteroduplexes, in which each strand has a mutation. Any gaps in these newly formed heteroduplexes are filled with DNA polymerases and further amplified.

Sequence saturation mutagenesis (SeSaM)

This technique results in the randomization of the target sequence at every nucleotide position. This method begins with the generation of variable length DNA fragments tailed with universal bases via the use of template transferases at the 3' termini. Next, these fragments are extended

to full length using a single stranded template. The universal bases are replaced with a random standard base, causing mutations. There are several modified versions of this method such as SeSAM-Tv-II, SeSAM-Tv+, and SeSAM-III.

Single primer reactions in parallel (SPRINP)

This site saturation mutagenesis method involves two separate PCR reaction. The first of which uses only forward primers, while the second reaction uses only reverse primers. This avoids the formation of primer dimer formation.

Mega Primed and Ligase Free Focused Mutagenesis

This site saturation mutagenic technique begins with one mutagenic oligonucleotide and one universal flanking primer. These two reactants are used for an initial PCR cycle. Products from this first PCR cycle are used as mega primers for the next PCR.

Ω-PCR

This site saturation mutagenic method is based on overlap extension PCR. It is used to introduce mutations at any site in a circular plasmid

PFunkel-Ominchange-OSCARR

This method utilizes user defined site directed mutagenesis at single or multiple sites simultaneously. OSCARR is an acronym for One Pot Simple Methodology for Cassette Randomization and Recombination. This randomization and recombination results in randomization of desired fragments of a protein. Omnichange is a sequence independent, multisite saturation mutagenesis which can saturate up to five independent codons on a gene.

Trimer-Dimer Mutagenesis

This method removes redundant codons and stop codons.

Cassette Mutagenesis

This is a PCR based method. Cassette mutagenesis begins with the synthesis of a DNA cassette containing the gene of interest, which is flanked on either side by restriction sites. The endonuclease which cleaves these restriction sites also cleaves sites in the target plasmid. The DNA cassette and the target plasmid are both treated with endonucleases to cleave these restriction sites and create sticky ends. Next the products from this cleavage are ligated together, resulting in the insertion of the gene into the target plasmid. An alternative form of cassette mutagenesis called combinatorial cassette mutagenesis is used to identify the

functions of individual amino acid residues in the protein of interest. Recursive ensemble mutagenesis then utilizes information from previous combinatorial cassette mutagenesis. Codon cassette mutagenesis allows you to insert or replace a single codon at a particular site in double stranded DNA.

Sexual Methods

Sexual methods of directed evolution involve *in vitro* recombination which mimic natural *in vivo* recombination. Generally these techniques require high sequence homology between parental sequences. These techniques are often used to recombine two different parental genes, and these methods do create cross overs between these genes.

***In vitro* Homologous Recombination**

Homologous recombination can be categorized as either *in vivo* or *in vitro*. *In vitro* homologous recombination mimics natural *in vivo* recombination. These *in vitro* recombination methods require high sequence homology between parental sequences. These techniques exploit the natural diversity in parental genes by recombining them to yield chimeric genes. The resulting chimera show a blend of parental characteristics.

DNA shuffling

This *in vitro* technique was one of the first techniques in the era of recombination. It begins with the digestion of homologous parental genes into small fragments by DNase1. These small fragments are then purified from undigested parental genes. Purified fragments are then reassembled using primer-less PCR. This PCR involves homologous fragments from different parental genes priming for each other, resulting in chimeric DNA. The chimeric DNA of parental size is then amplified using end terminal primers in regular PCR.

Random priming *In vitro* recombination (RPR)

This *in vitro* homologous recombination method begins with the synthesis of many short gene fragments exhibiting point mutations using random sequence primers. These fragments are reassembled to full length parental genes using primer-less PCR. These reassembled sequences are then amplified using PCR and subjected to further selection processes. This method is advantageous relative to DNA shuffling because there is no use of DNase1, thus there is no bias for recombination next to a pyrimidine nucleotide. This method is also advantageous due to its use of synthetic random primers which are uniform in length, and lack biases. Finally this method is independent of the length of DNA template sequence, and requires a small amount of parental DNA.

Truncated Metagenomic Gene-Specific PCR

This method generates chimeric genes directly from metagenomic samples. It begins with isolation of the desired gene by functional screening from metagenomic DNA sample. Next, specific primers are designed and used to amplify the homologous genes from different environmental samples. Finally, chimeric libraries are generated to retrieve the desired functional clones by shuffling these amplified homologous genes.

Staggered Extension Process (StEP)

This *in vitro* method is based on template switching to generate chimeric genes. This PCR based method begins with an initial denaturation of the template, followed by annealing of primers and a short extension time. All subsequent cycle generate annealing between the short fragments generated in previous cycles and different parts of the template. These short fragments and the templates anneal together based on sequence complementarity. This process of fragments annealing template DNA is known as template switching. These annealed fragments will then serve as primers for further extension. This method is carried out until the parental length chimeric gene sequence is obtained. Execution of this method only requires flanking primers to begin. There is also no need for Dnase1 enzyme.

Random Chimeragenesis on Transient Templates (RACHITT)

This method has been shown to generate chimeric gene libraries with an average of 14 crossovers per chimeric gene. It begins by aligning fragments from a parental top strand onto the bottom strand of a uracil containing template from a homologous gene. 5' and 3' overhang flaps are cleaved and gaps are filled by the exonuclease and endonuclease activities of Pfu and taq DNA polymerases. The uracil containing template is then removed from the heteroduplex by treatment with a uracil DNA glcosylase, followed by further amplification using PCR. This method is advantageous because it generates chimeras with relatively high crossover frequency. However it is somewhat limited due to the complexity and the need for generation of single stranded DNA and uracil containing single stranded template DNA.

Synthetic shuffling

Shuffling of synthetic degenerate oligonucleotides adds flexibility to shuffling methods, since oligonucleotides containing optimal codons and beneficial mutations can be included.

***In vivo* Homologous Recombination**

Cloning Performed in Yeast

This method involves PCR dependent reassembly of fragmented expression vectors. These reassembled vectors are then introduced to, and cloned in yeast. Using yeast to clone the

vector avoids toxicity and counter-selection that would be introduced by ligation and propagation in E. coli.

Mutagenic Organized Recombination Process by Homologous *In Vivo* Grouping (MORPHING)

This method introduces mutations into specific regions of genes while leaving other parts intact by utilizing the high frequency of homologous recombination in yeast.

Phage Assisted Continuous Evolution (PACE)

This method utilizes a bacteriophage with a modified life cycle to transfer evolving genes from host to host. The phage's life cycle is designed in such a way that the transfer is correlated with the activity of interest from the enzyme. This method is advantageous because it requires minimal human intervention for the continuous evolution of the gene.

***In Vitro* Non-Homologous Recombination Methods**

These methods are based upon the fact that proteins can exhibit similar structural identity while lacking sequence homology.

Exon Shuffling

Exon shuffling is the combination of exons from different proteins by recombination events occurring at introns. Orthologous exon shuffling involves combining exons from orthologous genes from different species. Orthologous domain shuffling involves shuffling of entire protein domains from orthologous genes from different species. Paralogous exon shuffling involves shuffling of exon from different genes from the same species. Paralogous domain shuffling involves shuffling of entire protein domains from paralogous proteins from the same species. Functional homolog shuffling involves shuffling of non-homologous domains which are functional related. All of these processes being with amplification of the desired exons from different genes using chimeric synthetic oligonucleotides. This amplification products are then reassembled into full length genes using primer-less PCR. During these PCR cycles the fragments act as templates and primers. This results in chimeric full length genes, which are then subjected to screening.

Incremental Truncation for the Creation of Hybrid Enzymes (ITHCY)

Fragments of parental genes are created using controlled digestion by exonuclease III. These fragments are blunted using endonuclease, and are ligated to produce hybrid genes. THIOITHCY is a modified ITHCY technique which utilized nucleotide triphosphate analogs

such as α-phosphothioate dNTPs. Incorporation of these nucleotides blocks digestion by exonuclease III. This inhibition of digestion by exonuclease III is called spiking. Spiking can be accomplished by first truncating genes with exonuclease to create fragments with short single stranded overhangs. These fragments then serve as templates for amplification by DNA polymerase in the presence of small amounts of phosphothioate dNTPs. These resulting fragments are then ligated together to form full length genes. Alternatively the intact parental genes can be amplified by PCR in the presence of normal dNTPs and phosphothioate dNTPs. These full length amplification products are then subjected to digestion by an exonuclease. Digestion will continue until the exonuclease encounters an α-pdNTP, resulting in fragments of different length. These fragments are then ligated together to generate chimeric genes.

SCRATCHY

This method generates libraries of hybrid genes inhibiting multiple crossovers by combining DNA shuffling and ITHCY. This method begins with the construction of two independent ITCHY libraries. The first with gene A on the N-terminus. And the other having gene B on the N-terminus. These hybrid gene fragments are separated using either restriction enzyme digestion or PCR with terminus primers via agarose gel electrophoresis. These isolated fragments are then mixed together and further digested using DNase1. Digested fragments are then reassembled by primerless PCR with template switching.

Recombined Extension on Truncated Templates (RETT)

This method generates libraries of hybrid genes by template switching of uni-directionally growing polynucleotides in the presence of single stranded DNA fragments as templates for chimeras. This method begins with the preparation of single stranded DNA fragments by reverse transcription from target mRNA. Gene specific primers are then annealed to the single stranded DNA. These genes are then extended during a PCR cycle. This cycle is followed by template switching and annealing of the short fragments obtained from the earlier primer extension to other single stranded DNA fragments. This process is repeated until full length single stranded DNA is obtained.

Sequence Homology-Independent Protein Recombination (SHIPREC)

This method generates recombination between genes with little to no sequence homology. These chimeras are fused via a linker sequence containing several restriction sites. This construct is then digested using DNase1. Fragments are made are made blunt ended using S1 nuclease. These blunt end fragments are put together into a circular sequence by ligation. This circular construct is then linearized using restriction enzymes for which the restriction sites are

present in the linker region. This results in a library of chimeric genes in which contribution of genes to 5' and 3' end will be reversed as compared to the starting construct.

Sequence Independent Site Directed Chimeragenesis (SISDC)

This method results in a library of genes with multiple crossovers from several parental genes. This method does not require sequence identity among the parental genes. This does require one or two conserved amino acids at every crossover position. It begins with alignment of parental sequences and identification of consensus regions which serve as crossover sites. This is followed by the incorporation of specific tags containing restriction sites followed by the removal of the tags by digestion with Bac1, resulting in genes with cohesive ends. These gene fragments are mixed and ligated in an appropriate order to form chimeric libraries.

Degenerate Homo-Duplex Recombination (DHR)

This method begins with alignment of homologous genes, followed by identification of regions of polymorphism. Next the top strand of the gene is divided into small degenerate oligonucleotides. The bottom strand is also digested into oligonucleotides to serve as scaffolds. These fragments are combined in solution are top strand oligonucleotides are assembled onto bottom strand oligonucleotides. Gaps between these fragments are filled with polymerase and ligated.

Random Multi-Recombinant PCR (RM-PCR)

This method involves the shuffling of plural DNA fragments without homology, in a single PCR. This results in the reconstruction of complete proteins by assembly of modules encoding different structural units.

User Friendly DNA Recombination (USERec)

This method begins with the amplification of gene fragments which need to be recombined, using uracil dNTPs. This amplification solution also contains primers, PfuTurbo, and Cx Hotstart DNA polymerase. Amplified products are next incubated with USER enzyme. This enzyme catalyzes the removal of uracil residues from DNA creating single base pair gaps. The USER enzyme treated fragments are mixed and ligated using T4 DNA ligase and subjected to Dpn1 digestion to remove the template DNA. These resulting dingle stranded fragments are subjected to amplification using PCR, and are transformed into E. coli.

Golden Gate Shuffling (GGS) Recombination

This method allows you to recombine at least 9 different fragments in an acceptor vector by

using type 2 restriction enzyme which cuts outside of the restriction sites. It begins with sub cloning of fragments in separate vectors to create Bsa1 flanking sequences on both sides. These vectors are then cleaved using type II restriction enzyme Bsa1, which generates four nucleotide single strand overhangs. Fragments with complementary overhangs are hybridized and ligated using T4 DNA ligase. Finally these constructs are then transformed into E. coli cells, which are screened for expression levels.

Phosphoro Thioate-Based DNA Recombination Method (PRTec)

This method can be used to recombine structural elements or entire protein domains. This method is based on phosphorothioate chemistry which allows the specific cleavage of phosphorothiodiester bonds. The first step in the process begins with amplification of fragments that need to be recombined along with the vector backbone. This amplification is accomplished using primers with phosphorothiolated nucleotides at 5' ends. Amplified PCR products are cleaved in an ethanol-iodine solution at high temperatures. Next these fragments are hybridized at room temperature and transformed into E. coli which repair any nicks.

Integron

This system is based upon a natural site specific recombination system in E. coli. This system is called the integron system, and produces natural gene shuffling. This method was used to construct and optimize a functional tryptophan biosynthetic operon in trp-deficient E. coli by delivering individual recombination cassettes or trpA-E genes along with regulatory elements with the integron system.

Y-Ligation Based Shuffling (YLBS)

This method generates single stranded DNA strands, which encompass a single block sequence either at the 5' or 3' end, complementary sequences in a stem loop region, and a D branch region serving as a primer binding site for PCR. Equivalent amounts of both 5' and 3' half strands are mixed and formed a hybrid due to the complementarity in the stem region. Hybrids with free phosphorylated 5' end in 3' half strands are then ligated with free 3' ends in 5' half strands using T4 DNA ligase in the presence of 0.1 mM ATP. Ligated products are then amplified by two types of PCR to generate pre 5' half and pre 3' half PCR products. These PCR product are converted to single strands via avidin-biotin binding to the 5' end of the primes containing stem sequences that were biotin labeled. Next, biotinylated 5' half strands and non-biotinylated 3' half strands are used as 5' and 3' half strands for the next Y- ligation cycle.

Semi-Rational Design

Semi-rational design uses information about a proteins sequence, structure and function, in tandem with predictive algorithms. Together these are used to identify target amino acid residues which are most likely to influence protein function. Mutations of these key amino acid residues create libraries of mutant proteins that are more likely to have enhanced properties.

Advances in semi-rational enzyme engineering and de novo enzyme design provide researchers with powerful and effective new strategies to manipulate biocatalysts. Integration of sequence and structure based approaches in library design has proven to be a great guide for enzyme redesign. Generally, current computational de novo and redesign methods do not compare to evolved variants in catalytic performance. Although experimental optimization may be produced using directed evolution, further improvements in the accuracy of structure predictions and greater catalytic ability will be achieved with improvements in design algorithms. Further functional enhancements may be included in future simulations by integrating protein dynamics

Biochemical and biophysical studies, along with fine-tuning of predictive frameworks will be useful to experimentally evaluate the functional significance of individual design features. Better understanding of these functional contributions will then give feedback for the improvement of future designs.

Directed evolution will likely not be replaced as the method of choice for protein engineering, although computational protein design has fundamentally changed the way protein engineering can manipulate bio-macromolecules. Smaller, more focused and functionally- rich libraries may be generated by using in methods which incorporate predictive frameworks for hypothesis-driven protein engineering. New design strategies and technical advances have begun a departure from traditional protocols, such as directed evolution, which represents the most effective strategy for identifying top-performing candidates in focused libraries. Whole- gene library synthesis is replacing shuffling and mutagenesis protocols for library preparation. Also highly specific low throughput screening assays are increasingly applied in place of monumental screening and selection efforts of millions of candidates. Together, these developments are poised to take protein engineering beyond directed evolution and towards practical, more efficient strategies for tailoring biocatalysts.

Screening and Selection Techniques

Once a protein has undergone directed evolution, ration design or semi-ration design, the libraries of mutant proteins must be screened to determine which mutants show enhanced properties. Phage display methods are one option for screening proteins. This method involves

the fusion of genes encoding the variant polypeptides with phage coat protein genes. Protein variants expressed on phage surfaces are selected by binding with immobilized targets in vitro. Phages with selected protein variants are then amplified in bacteria, followed by the identification of positive clones by enzyme linked immunosorbent assay. These selected phages are then subjected to DNA sequencing.

Cell surface display systems can also be utilized to screen mutant polypeptide libraries. The library mutant genes ate incorporated into expression vectors which are then transformed into appropriate host cells. These host cells are subjected to further high throughput screening methods to identify the cells with desired phenotypes.

Cell free display systems have been developed to exploit *in vitro* protein translation or cell free translation. These methods include mRNA display, ribosome display, covalent and non covalent DNA display, and *in vitro* compartmentalization.

Enzyme engineering

Enzyme engineering is the application of modifying an enzyme's structure (and, thus, its function) or modifying the catalytic activity of isolated enzymes to produce new metabolites, to allow new (catalyzed) pathways for reactions to occur, or to convert from some certain compounds into others (biotransformation). These products are useful as chemicals, pharmaceuticals, fuel, food, or agricultural additives.

An *enzyme reactor* consists of a vessel containing a reactional medium that is used to perform a desired conversion by enzymatic means. Enzymes used in this process are free in the solution.

ANTIBODY ENGINEERING

Natural antibodies consist of an antigen binding site joined to an effector region that is responsible for activating complement and or binding to immune cells. From a biotechnological viewpoint, the incredibly high specificity with which antibodies bind to a target protein is useful for a variety of purposes. Consequently, antibody engineering uses the antigen binding region of the antibody. These are manipulated and are attached to other molecular fragments.

To separate an antigen binding site from the rest of the antibody, gene segments encoding portions of antibody chains are subcloned and expressed in bacterial cells. Bacterial signal sequences are added to the N terminus of the partial antibody chains, which results in export

of the chains into the periplasmic space. Here the VH and VL domains fold up correctly and form their disulfide bonds. The antibody fragments used include Fab, Fv, and **single-chain Fv (scFv).** In a Fab fragment, an interchain disulfide bond holds the two chains together. However, the Fv fragment lacks this region of the antibody chains and thus is less stable. This led to development of the single-chain Fv fragment in which the VH and VL domains are linked together by a short peptide chain, usually 15 to 20 amino acids long. This is introduced at the genetic level so that a single artificial gene expresses the whole structure (VH-linker-VL or VL-linker-VH). A tag sequence (such as a His6-tag or FLAG-tag) is often added to the end to allow detection and purification. Such an scFv fragment is quite small, about 25,000 in molecular weight.

Such scFv fragments are attached to various other molecules by genetic engineering. The role of the scFv fragment is to recognize some target molecule, perhaps a protein expressed only on the surface of a virus-infected cell or a cancer cell. A variety of toxins, cytokines, or enzymes may be attached to the other end of the scFv fragment, to provide the active portion of the final recombinant antibody. In principle, this approach provides a way of delivering a therapeutic agent in an extremely specific manner. At present the clinical applications of engineered antibodies are under experimental investigation.

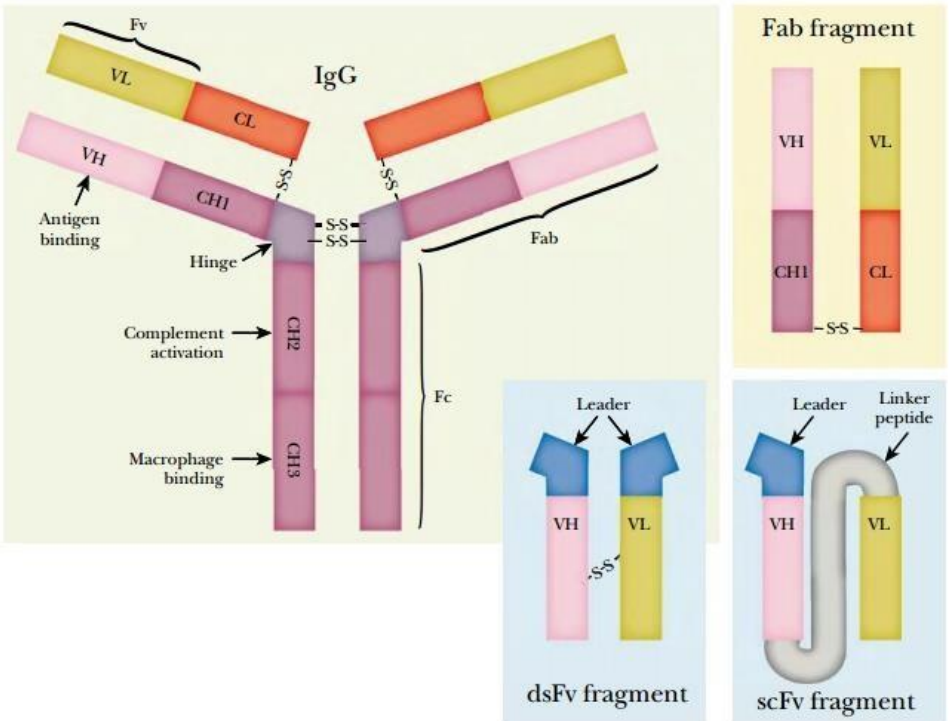


Figure 1

The antigen binding regions used in antibody engineering may be derived from characterized monoclonal antibodies. Alternatively, a library of DNA segments encoding V-regions may be obtained from a pool of B cells obtained from an animal or human blood sample. Such a library should in theory contain V-regions capable of recognizing any target molecule. Using a human source avoids the necessity for the complex humanization procedures described earlier. However, in this case it is necessary to screen the V-region library for an antibody fragment that binds to the desired target molecule. This may be done by the phage display procedure outlined. The library of V-region constructs is expressed on the surface of the phage, and the target molecule is attached to some solid support and used to screen out those phages carrying the required antibody V region.

DIABODIES AND BISPECIFIC ANTIBODY CONSTRUCTS

A variety of engineered antibody constructs are presently being investigated. A **diabody** consists of two single-chain Fv (scFv) fragments assembled together. Shortening the linker from 15 amino acids to five drives dimerization of two scFv chains. This no longer allows intrachain assembly of the linked VH and VL regions. The dimer consists of two scFv fragments arranged in a crisscross manner. The resulting diabody has two antigen binding sites pointing in opposite directions. If two different scFv fragments are used, the result is a bispecific diabody that will bind to two different target proteins simultaneously. Note that formation of such a bispecific diabody requires that VH-A be linked to VL-B and VH-B to VL-A. It is of course possible to engineer both sets of VH and VL regions onto a single polypeptide chain encoded by a single recombinant gene. Bispecific diabodies have a variety of potential uses in therapy, because they may be used to bring together any two other molecules; for example, they might be used to target toxins to cancer cells.

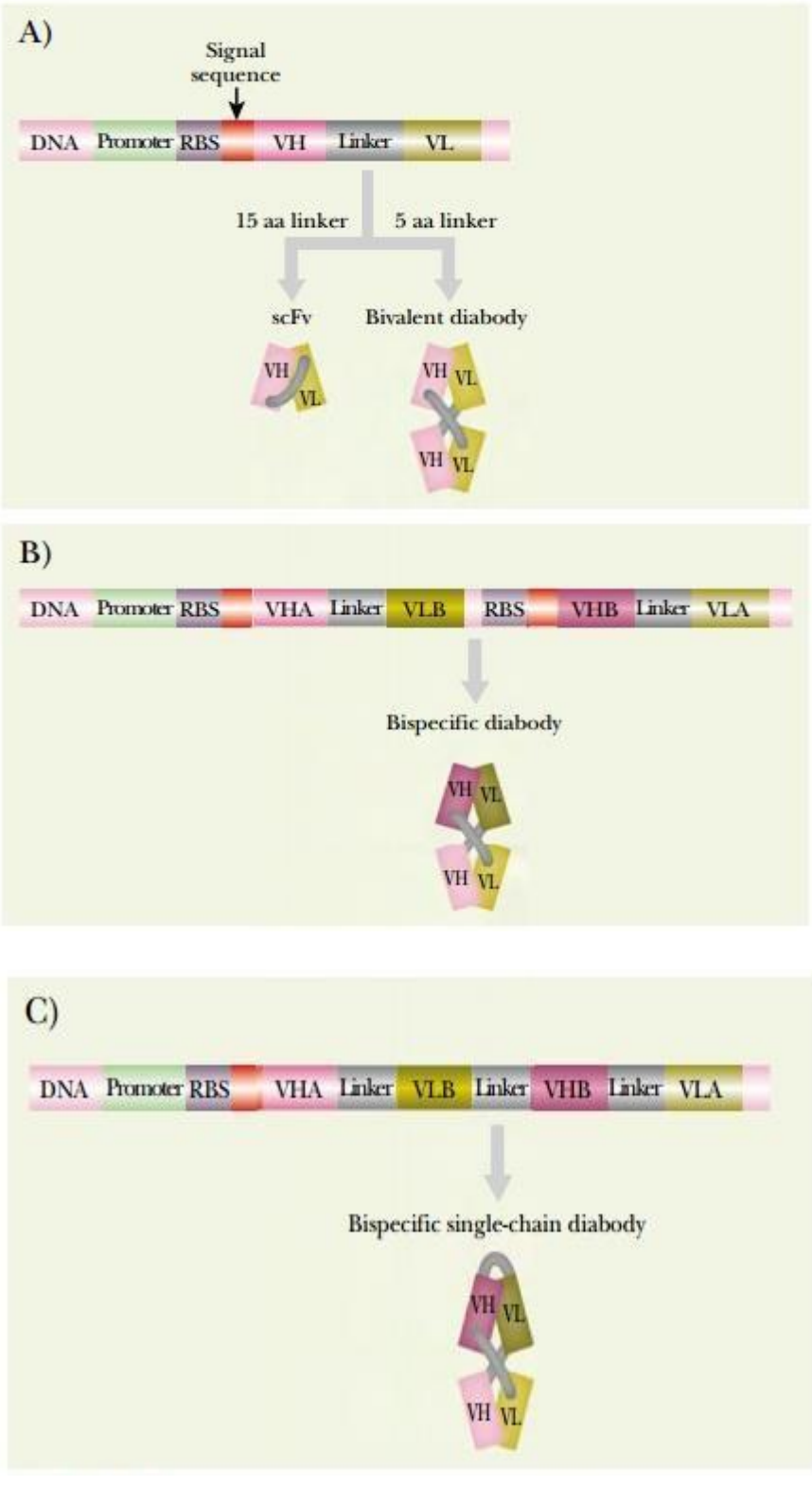

A)
Signal sequence
DNA
Promoter
RBS
VH
Linker
VL
15 aa linker
5 aa linker
scFv
Bivalent diabody
VH
VL
B)
DNA
Promoter
RBS
VHA
Linker
VLB
RBS
VHB
Linker
VLA
Bispecific diabody
C)
DNA
Promoter
RBS
VHA
Linker
VLB
Linker
VHB
Linker
VLA
Bispecific single-chain diabody

Figure 2

Another way to construct an engineered bispecific antibody is to connect the two different scFv fragments to other proteins that bind together. Two popular choices are streptavidin and leucine zippers. Streptavidin is a small biotin binding protein from the bacterium Streptococcus. It forms tetramers, so it allows up to four antibody fragments to be assembled together. Furthermore, binding to a biotin column can purify the final constructs. Leucine zipper regions are used by many transcription factors that form dimers. Often, such proteins form mixed dimers when their leucine zippers recognize each other and bind together. Leucine zipper regions from two different transcription factors that associate (e.g., the Fos and Jun proteins) may therefore be used to assemble two different scFv fragments.

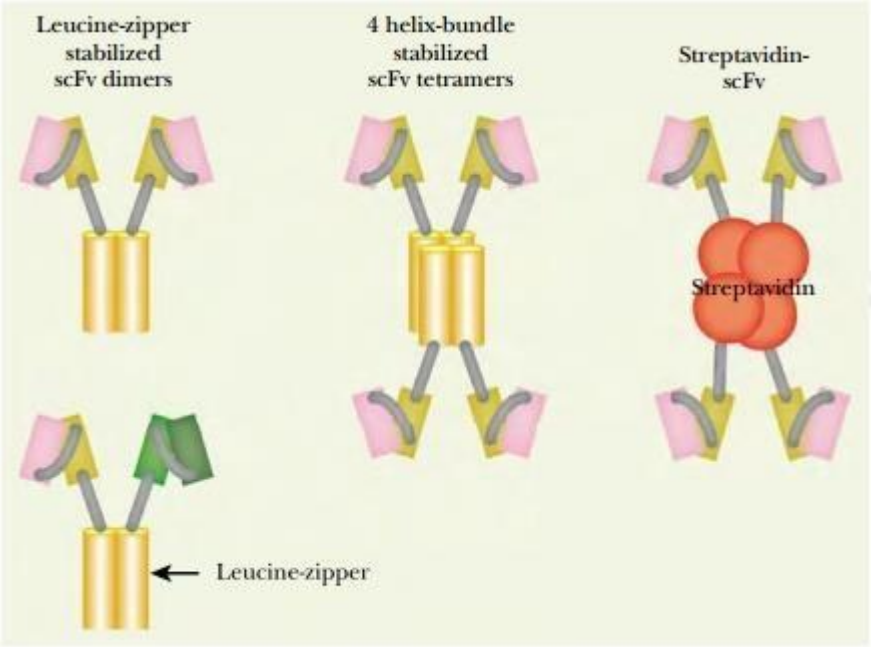


Figure 3

Examples of engineered proteins

Computing methods have been used to design a protein with a novel fold, named Top7 and sensors for unnatural molecules. The engineering of fusion proteins has yielded rilonacept, a pharmaceutical that has secured Food and Drug Administration (FDA) approval for treating cryopyrin-associated periodic syndrome.

Another computing method, IPRO, successfully engineered the switching of cofactor specificity of *Candida boidinii* xylose reductase. Iterative Protein Redesign and Optimization (IPRO) redesigns proteins to increase or give specificity to native or novel substrates and cofactors. This is done by repeatedly randomly perturbing the structure

of the proteins around specified design positions, identifying the lowest energy combination of rotamers, and determining whether the new design has a lower binding energy than prior ones.

Computation-aided design has also been used to engineer complex properties of a highly ordered nano-protein assembly. A protein cage, E. coli bacterioferritin (EcBfr), which naturally shows structural instability and an incomplete self-assembly behavior by populating two oligomerization states, is the model protein in this study. Through computational analysis and comparison to its homologs, it has been found that this protein has a smaller-than- average dimeric interface on its two-fold symmetry axis due mainly to the existence of an interfacial water pocket centered on two water-bridged asparagine residues. To investigate the possibility of engineering EcBfr for modified structural stability, a semi-empirical computational method is used to virtually explore the energy differences of the 480 possible mutants at the dimeric interface relative to the wild type EcBfr. This computational study also converges on the water-bridged asparagines. Replacing these two asparagines with hydrophobic amino acids results in proteins that fold into alpha-helical monomers and assemble into cages as evidenced by circular dichroism and transmission electron microscopy. Both thermal and chemical denaturation confirm that, all redesigned proteins, in agreement with the calculations, possess increased stability. One of the three mutations shifts the population in favor of the higher order oligomerization state in solution as shown by both size exclusion chromatography and native gel electrophoresis.

BIOTECHNOLOGICAL AND BIOMEDICAL APPLICATIONS OF PROTEIN ENGINEERING METHODS

Diversified Applications of Protein Engineering Techniques

Protein engineering principles delivered wide spread applications in many different fields including the fields of biotechnology, nanotechnology and biomedicine. Biotechnological applications mainly include the improvement in the activity/functionality of the industrially important enzymes that are involved in food, detergents, textile industry and also for those being employed to control environmental pollution. Many of the protein engineering techniques are being used in order to fabricate different types of biomaterials for the medical and nanotechnology applications, also in designing the biosensors for their applications in molecular imaging.

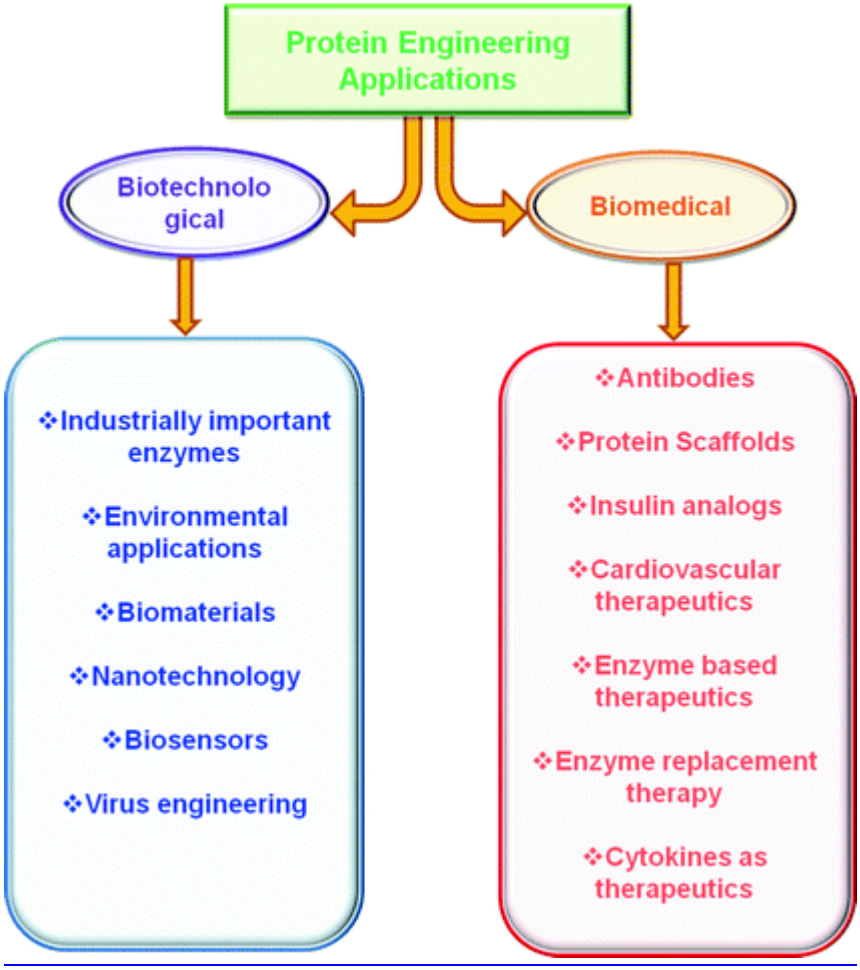


Figure 4

An overview of protein engineering applications in diversified fields of biotechnology and biomedical sciences

Biomedical applications of protein engineering includes development of antibodies, designing of protein scaffolds to inculcate differential medicinal properties in them, designing therapeutics specific to number of diseases like diabetes, cardiac related diseases, modulating the properties of endogenous proteins like cytokines so as to make them viable in the treatment of various inflammatory and infectious diseases.

Industrial Applications

Large numbers of enzymes are being employed in food, detergent and textile industries with the aim to produce the best products at the cheaper cost. Majority of enzymes are being exploited in detergent industry. The cocktails of enzymes are added in the detergents so as to increase the ability of detergents to remove tough stains. A continuous effort is being made to improve the performance of enzymes in the terms of their activity, specificity, catalytic efficiency and stability at wide range of pH and temperature conditions. Protein engineering techniques including directed evolution and rational designing played a critical role in improving the industrially important enzymes including proteases, amylases, lipases, cellulases and xylanases.

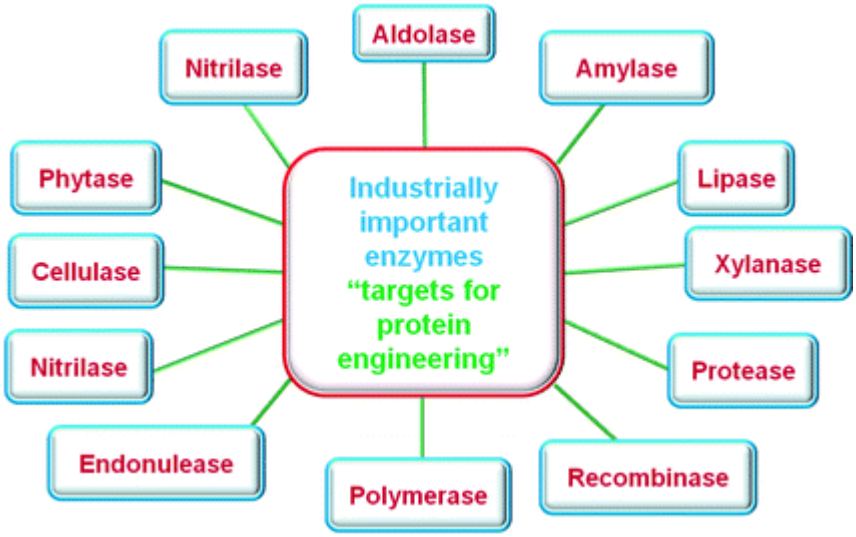


Figure 5

Snapshot of the industrially important enzymes ameliorated using protein engineering techniques

Proteases, constitutes one of the major class of industrially important enzymes with the huge importance in detergent and dairy industry. Proteases can be obtained from animals, plants and microorganisms. Proteases hydrolyze the peptide bonds in the proteins and peptides to produce

small chunks of amino acids. These enzymes with thermo stability and activity at alkaline pH have gained huge importance in detergent industry to remove protein based stains from the clothes. Protein engineering has made it possible to produce the enzymes for the detergent industry that can withstand their activities at wide range of temperatures at alkaline pH. Subtilisins, group of bacterial serine proteases have wider spread use in detergent industry, Subtilisins including Subtilisin bacterial protease nagase (BPN) from *Bacillus subtilis, B. stearothermophilus and B. amyloliquefaciens,* Subtilisin Carlsberg produced by *B licheniformis,* and Subtilisin Novo produced from *B. subtilis* exhibits huge importance in detergent industry. A protein engineering technique, cassette mutagenesis was applied in which Met222 was substituted with all 19 amino acids in the cloned subtilisin gene and the mutants were screened for the subtilis invariants which were stable and resistant towards inactivation by hydrogen peroxide. The comparative dynamic studies of subtilisins, a 269 residue serine protease subtilisin PB92, secreted by B. lentus and an engineered quadruple variant DSAI, showed an improved washing performance due to their more structured substrate binding sites in the engineered variant as compared to the natural protease.

The class of extracellular subtilisin proteases depends on calcium binding for their stability. With the advent of protein engineering, the loops for calcium binding were deleted from the subtilisin BPN from *Bacillus amyloliquefaciens* and destabilized the native structure. This was again restabilized using the directed mutagenesis and selection procedures of protein engineering techniques. The resultant subtilisin showed similar proteolytic activity as that of native enzyme, moreover, it acquired 1000 times more stability in the chelating environment of detergent. Further, DNA shuffling method was employed on 26 subtilisin protease genes, resulting in hybrid genes with improved properties including temperature, stability, activity in organic solvents and activity at high and low pH values. Further, applications of protein engineering resulted in psychrophilic counterpart of mesophilic subtilisin proteases.

Lipases are the other major class of enzymes that are being exploited in both food and detergent industry. Major application of lipase in food industry includes: adds flavor to the dairy products, plays role in processing of other food items like beer, baked foods, milk products, vegetables, meat, used for egg yolk treatment to produce mayonnaise and other emulsifiers, act as biosensors for quantitative analysis of triacylglycerol. Protein engineering has played an important role in designing the efficient lipases for its different applications in food industry. The first thermostable lipase engineered was based on sequence information in mid of 1980s. The first lipase enzyme engineered was *Pseudomonas mendocina* lipase. Many lipases with

improvement in protease stability, oxidative stability and thermostability have been engineered. Many of the lipases gets activated at the substrate-water interface, and undergoes a conformational change mediated by the lid displacement thereby exposing its hydrophobic binding pocket. Mutations at the residues Glu87 and Trp89 in the lid region have been marked to alter the enzymatic activity and thus reported as important sites for hydrolytic activity of lipase from *Humicola lanuginosa*. Further, the engineered *Candida rugosa* lipase (CRL) isoforms were improvised further in the terms of its activity, thermostability, specificity and enantioselectivity by lid swapping and DNA shuffling techniques in order to increase its applications in food industry.

Amylases are the class of industrially important enzymes which hydrolyzes starch into the low molecular weight products such as glucose, maltose and maltotriose. α-amylases have huge importance among many industries namely, food, detergent, paper, fermentation, textile, and pharmaceutical industries. Amylases are also being used in detergents to remove starchy stains from clothes as well as from dishes. Amylases being active at low temperatures and alkaline pH and with the ability to maintain its oxidative stability under detergent conditions is one of the main advantage of their usage in detergents. An increase in thermostability of beta-amylase from barley using random mutagenesis has been reported. Thermostability of α-amylase from *Bacillus licheniformis* has been increased by the mutations at positions 209 and 133, guided by the protein engineering techniques including modeling and site saturation mutagenesis. C-terminal deletions in glucoamylase from *Aspergillus awamori*, resulted in loss of starch binding activity and starch hydrolytic activity but the retention of thermostability and enzymatic activity on soluble starch ensures the individuality of starch binding site and catalytic domains of glucoamylase. Further, thermostability of glucoamylase has been increased by reducing the alpha-helix flexibility by mutating glycines at the helix to alanine.

Cellulases are also accounted as important enzymes in many industries including food, detergent, textile, paper, and pulp. Cellulases mediates the hydrolysis of β-1,4 linkages in cellulose chains. Complete hydrolysis of cellulose occurs by the combination of three main types of cellulases namely: endoglucanase, exoglucanase including cellobiohydrolases (CBHs) and β-glucosidase. These enzymes have been improved individually using the principles of directed evolution and selection strategies. Liu et al. obtained thermostable β-glucosidase mutants using the combinatorial selection and screening strategy. Improved variants of carboxymethyl cellulase have been obtained by DNA shuffling method. Catalytic efficiency of

endo-beta-1,4-glucanase from *Bacillus subtilis* BME-15 has been improved using combination of different directed evolution techniques including error prone PCR and DNA shuffling.

Xylanases are used in paper and pulp industry and their exploitation has replaced the usage of harsh chemicals for bleaching pulp. They are also being used as additives in variety of food including poultry, conditioning of dough, extraction of coffee, starch, and plant oils, also increase nutritional values of agricultural silage and grain feed etc. Xylanases have also been improved to meet the industrial requirements for activity at different conditions, enantioselectivity, substrate specificity, increased tolerance to toxic reagents. Thermostability, thermophilicity (from 75° to 90° C) and alkophilicity (from pH 7.5 to 9) of *T. reesei* endoxylanase II was improved in the version of enzyme engineered by Sung and Taylon by three strategies, one is by the replacement of amino acids at position 10, 27 and 89 by His, Met and Leu respectively, second by the replacement of N-terminal amino acid sequence by the N terminal amino acid sequence of xylanase from *Thermomonospora fusca*, and thirdly by the addition of a tripeptide sequence Gly-Arg-Arg or 10 extra amino acids from N-terminus of *Clostridium acetobutylicum* xynB to the N-terminus. They have also engineered the same enzyme by mutating the residues Val 108, Ser 110, Asn 154 and Ala 158 to Cys and Gln 162 mutated to His. The engineered enzyme had showed an increased activity at 62.5 °C and pH 5.5.

Other important enzymes that have been improved using protein engineering strategies includes aldolases, transaldolases, nitrilases, microbial beta-D-xylosidases, microbial glucoamylases, human butyrylcholinesterase, cholesterol oxidase, phytases, extremozymes, homing endonucleases, Recombinases, DNA polymerases.

Environmental Applications

Protein engineering has also contributed to the wide scale environmental applications including the development of environmental biosensors, development of enzymes with high activities to degrade environmental pollutants and waste management. Many new methods are being employed to create gene expression regulators that results in high expression of enzymes with high catalytic activity under the stress conditions such as in presence of toxic substances or the other environment pollutants.

Oxidative enzymes, one of the important classes of enzymes, exploited for oxidative degradation of toxic organic substances including phenols, azo dyes, and polycyclic aromatic hydrocarbons. The major limitation in using these enzymes includes the rapid denaturation of

enzymes in the presence of organic solvents, low activity of enzymes, high cost, and less availability. Protein engineering played a crucial role in circumventing all these limitations by tailoring of high activity enzymes, with increased stabilities in the organic solvents. These enzymes not only take part in bioremediation, but also helpful in the development of environmental friendly applications.

Fungal peroxidases have also been recognized as important enzymes in controlling the environment pollution, owing to their ability to transform xenobiotics and other polluting agents. For their better industrial and environmental applications, enzyme needs an improvement in its stability and availability. These challenges have been addressed by the protein engineering strategies that have enhanced their operational stability, broaden its substrate range by increasing the enzyme redox potential and also developed the strategies for its heterologous expression and industrial production.

Numerous other bio-degradative enzymes have also been improved to enhance their bioremediation properties. Biphenyl dioxygenase BphA from Burkholderia xenovorans LB400 has been improved by the mutations T335A and F366M, that were incorporated by random mutagenesis method. Keenan et al. employed site saturation mutagenesis method to generate the V350F mutant of 2,4-dinitrotoluene dioxygenase (DDO) of Burkholderia cepacia R34. This mutant showed high activity towards o-nitrophenol (47 times), m-nitrophenol (34 times), and o-methoxyphenol (174 times) and also showed an expanded substrate range including m-methoxyphenol, o-cresol, and m-cresol, for which wild type shows no activity. Canada et al. used DNA shuffling method to enhance the activity of toluene ortho-monooxygenase (TOM) for the oxidation of chlorinated ethenes (contaminant in ground water) and naphthalene (chemical manufacturing intermediate) oxidation. Okuta et al. have used cassette mutagenesis method to obtain chimeric catechol 2,3-dioxygenase. This enzyme showed improved activity against the suicide inhibitor 4-methylcatechol. All these studies exemplify the potential of protein engineering in environmental restoration and green chemistry.

Biomaterial Applications

Protein engineering field had also extended its arms towards the biomaterial science, which has potential applications in biomedicine including drug delivery vehicles, soluble carriers and implantable materials. They also exhibit wide variety of applications in field of nanotechnology and tissue engineering including injectable scaffolds, hydrogels for regenerative medicine

therapies. Many peptide based materials have been designed to serve as components of biosensor/bio-analytical devices, for nucleation of inorganic materials.

Polypeptide based biomaterials exhibits many advantages over the synthetic polymers which includes: (1) Ability of short peptide motifs like RGD, KNEED, IKVAV to mediate cell attachment and spreading, allows these motifs to be incorporated in polypeptide based biomaterials. (2) Property to self-assemble or directed assembly of peptides is used to generate viable or functional biomaterials. (3) Easy degradation of the peptide based biomaterials by the body makes them suitable for drug delivery vehicles.

These polypeptide based materials can be generated either by chemical synthesis or by recombinant DNA technology. Both the strategies have their own pros and cons. In chemical synthesis methods, it is hard to control the various parameters including stereochemistry and chain length. Chemical method is used mainly to synthesize hybrid peptide based materials in which peptide domains are attached to the non-peptide materials. With the advent of recombinant DNA technology, it is easy to precisely control the size, sequence, and stereochemistry of the polypeptides. Further, modifications need to be done after the expression and purification of the polypeptides from the host cell.

Numerous of peptide based biomaterials engineered includes leucine zipper based peptides, coiled-coil domains, beta-sheet forming ionic oligopeptides, beta-hairpin peptides, silk-like proteins, poly-amino acids, elastin-like polypeptides, tropoelastin-based peptides. Some of them are described below in greater detail.

Elastin like Polypeptides (ELPs)

ELPs are the biopolymer repeats of pentapeptide Val-Pro-Gly-X-Gly, where x can be any natural amino acid except proline. ELPs, are being employed in broad range of biomaterials and nanomaterial applications due to their self-assembling properties, biocompatibility, and versatility to fine tune its properties, either by amino acid substitutions or their combination with other polymeric materials. ELPs are widely used as drug carriers in targeted drug delivery systems. ELPs undergo an inverse temperature phase; i.e., they become soluble in aqueous solution below their transition temperature and undergo aggregation above their transition temperature. Such thermally responsive ELPs used in ELP-drug conjugates increases the localization of drug in the regions of tumor that are heated by regional hyperthermia. ELPs are being exploited in intra-articular drug delivery, and also for cartilaginous tissue repair. ELPs are also employed for the sustained-release drug delivery vehicles for the sustained release of

therapeutic agents to the dorsal root ganglion. Temperature-responsive cell sheets coated with elastic protein-based polymer are being used for cell transplantation studies. ELPs used to engineer small diameter vascular grafts, where it provides mechanical strength and includes site for covalent cross-linking. The inclusion of lysine residues in ELPs resulted in their rapid crosslinking with hydroxymethyl phosphines (HMP) at physiological conditions. This crosslinking is used in in situ gelation of ELPs for cell encapsulation. Thus, by altering the number and location of lysine residues in ELPs enables to tune its mechanical properties and the microenvironment they provide to cells. All these studies suggest that ELPs have wide scale applications as it can undergo different types of modifications as per the requirement for resulting biomaterial.

Silk Motifs

Silk is the natural fibrous protein secreted by spiders, which is light weighted and also exhibits high mechanical strength. Because of difficulty in harvesting silk from its natural source therefore silk protein is being synthesized by recombinant DNA technology and protein engineering principles are being employed to produce silk protein variants for their applications in gene delivery, drug delivery and as scaffolds for tissue engineering.

Bini et al. engineered two variants of silk protein, one with RGD motif and other without RGD motif and processed these proteins into fibers to use them as matrices or scaffolds for bone like tissue formation. Haider et al. used silk-elastin like protein polymer SELP-47 K as an injectable matrix for delivery of cell-based therapeutics. They have also concluded that SELP-47 K hydrogel can be employed as scaffold for the encapsulation and chondrogenesis of human mesenchymal stem cells. Haider et al. have also engineered series of SELP variants and used them for preparing hydrogel disks that can be used for the delivery of genes and bioactive agents. SELP hydrogel matrices have shown potentials for the long term controlled gene delivery. Progress is being made in the area of designing of chimeric silk and silk mimetics with high mechanical strength, variable conformations, and high solubility.

Coiled-Coil Motifs

Coiled-coil is a common motif in native proteins, characterized by two right handed α-helices wind around each other to form a left handed superhelix. Coiled coil domain contains heptad repeat sequence (abcdefg), where a and d are hydrophobic amino acids; more commonly leucine, e and g are charged amino acid; most commonly glutamic acid which also accounts for stability of helix by electrostatic interactions. Modification of the amino acid contained in

the heptad sequence can be used to make the electrostatic interactions sensitive to various factors including pH, temperature, denaturants or ligands. This sensitivity can be exploited by the drug delivery systems. pH and temperature sensitive hydrogels have already been made using these coiled coil domains. Xu et al. synthesized series of triblock protein copolymers made of two coiled coil domains with a central water soluble polyelectrolyte segment. These copolymers self assembles into the reversible hydrogels in response to changes in temperature, pH or in presence or absence of denaturant (guanidine hydrochloride). The property of copolymer to reversibly self-assemble into hydrogel makes these polymers as a potential candidates for biomedical field. Petka et al. have also synthesized an artificial self-assembling protein that also undergoes gelation upon temperature or pH changes, the protein consist of terminal leucine zipper motifs and a flanking water soluble polyelectrolyte domain. Leucine zipper also comes in the category of coiled-coil domains containing six heptad repeats that folds in amphiphilic alpha helix and multimerizes through electrostatic interactions that is further mediated by hydrophobic interactions between the nonpolar side chains. Leucine zipper plays an important role in dimerization and promotes the DNA binding of transcription regulatory proteins. Further, many strategies have been employed to stabilize the coiled coil domains and to form the coiled-coil protein based hydrogels including the use of photoreactive amino acids, addition of disulfide bonds, and incorporation of non-canonical fluorinated amino-acid residues. Leucine zipper, with its tunable properties, can be employed for the formation of biologically active scaffolds in various tissue engineering applications.

Calmodulin Motifs

Calmodulin, a 16.5 kDa protein, participates in regulation of $Ca2+$ pathways including neuronal communication and muscle contraction. Calmodulin undergoes conformational change upon binding by 4 $Ca2+$ ions, which allows it to bind one of the 100 different calmodulin binding domains present in other proteins reversibly through the calcium mediated mechanism. Calmodulin motifs are also being utilized in numerous biomaterials and nanomaterials with applications in tissue engineering and nanotechnology. Topp et al. prepared stimuli responsive biomaterials by merging the calmodulin motif with other peptide motifs. Triblock proteins were created by combining different motifs exhibiting different properties. For example, a triblock protein containing calmodulin as the sensory motif, leucine zipper as the self-assembling motif, hydrophilic protein sequences for providing additional flexibility and crosslinking has been created. The modular approach of protein engineering is gaining momentum now-a-days to engineer proteins with diverse functionalities that will further aid in

synthesizing diverse variety of protein based biomaterials with wide range of applications. Hall et al. used the property of calmodulin domain as biosensor to develop a calcium modulated plasmonic switch. These sensors can be used to study the real time dynamics and conformational changes in the proteins present in the cells.

β-Sheet Forming Ionic Oligopeptides

Zhang prepared nanofibers using the short oligonucleotides capable of forming β-sheet structures. Two β-sheets come together to form nanofibers which in turn assembles to form interwoven matrices leading to the formation of hydrogels. The hydrogels thus obtained were highly stable due to the intermolecular interactions including the ionic interactions between oppositely charged residues present on one side of sheet and Vander waal interactions between the hydrophobic residues present on other side of the sheet. Hydrogels obtained by the assembly of such oligopeptides, are sensitive to temperature, salt concentration and pH, so that the properties of these hydrogels can be modulated for their efficient use as biomaterials in tissue engineering. Number of studies have shown the potential of self-assembled hydrogels, scaffolds, nanofibers to support the growth of variety of mammalian cells including chondrocytes, neural cells, osteoblasts, and endothelial cells, whereby they are involved in tissue repair and regeneration.

Collagen, the main component of extracellular matrix and connective tissues is also being used for the creation of collagen based biomaterials, for the tissue engineering applications. Studies suggested the use of collagen based biological scaffolds, their derivatives and biocompatible copolymers for the cell attachment.

Protein engineering techniques, being employed to design new protein modules that can recapitulate the properties of large protein domains, to incorporate non-canonical amino acids to inculcate new chemical functionalities in the protein based biomaterials and thereby paving the way for the new opportunities and challenges to create novel peptide/protein based biomaterials.

Applications in Nanotechnology

With the progress made in the era of generation of peptide based biomaterials, protein engineering has made tremendous progress in the field of nanotechnology, by formulating variety of peptide based nanomaterials. Assembly of nanotechnological systems into functional devices is back breaking task which depends on the materials used in organization of such

nanoscale systems. These systems are synthesized by the integration of various organic, inorganic molecules along with variety of biological macromolecules including lipids, carbohydrates and proteins. Proteins playing several roles in cells, serves as suitable elements for the controlled assembly of nanotechnological systems. Coupling of many protein engineering techniques together to select amino acid sequences that are worth of carrying out specific task in nanosystems is an attractive rationale approach. Peptide sequences that specifically bind inorganic compound surfaces like gold, platinum, and quartz are of great interest. Polypeptides with the ability to aggregate into well-ordered structures of amyloid fibrils serves as the important materials for the nanotechnology applications. Scheibel et al. exploited the polypeptides forming such fibrillar structures for the construction of nanowires. Scheibel et al. have also constructed nanowires using amyloid fibrils thereby suggested the use of specificity of protein functions to generate nanoscale electrical circuits. Ranganathan et al. have designed and synthesized novel cysteine based spirobicyclic peptides that results in the formation of nanotube by vertical stacking of flat spirobicyclic molecules stabilized through the NH–O = C hydrogen bonds. De novo designed peptide MAX3 undergoes thermoreversible self-assembly into a hydrogel network. MAX 3 remains unfolded at ambient temperatures, but it starts folding at higher temperature to form amphiphilic β-hairpin that self assembles to form hydrogel. Hydrogelation occurs as a result of dehydration of non-polar amino acid residues in unfolded protein at higher temperature leading to hydrophobic collapse which initiates the protein folding. At lower temperatures, β-hairpin unfolding leads to the dissolution of hydrogel. Further, alteration in the hydrophobic residues in the peptide can change the temperature at which the peptide self assembles and form hydrogels. Such tunable peptides provide an opportunity for the generation of thermal responsive biomaterials and nanomaterials. McMillan et al. fabricated nanoscale arrays of metal and semiconductor quantum dots by using the chaperonin templates for binding to the preformed nanoparticles. These quantum dots can be modulated and organized into arrays for their use in modern electronic and photonic devices by modifying the chaperonin structure and self-assembling properties. Protein based nanomaterials have also been utilized to form flexible thin lithium ion batteries. Yu et al. have synthesized semi rigid polymers using engineered protein precursors that assembles into rare smectic liquid crystalline phases with scales of tens of nanometres.

Biosensors

Biosensors are the devices that provide the specific quantitative or semi quantitative information about the analytes. Biosensors consisting of receptors that interact with analyte,

and a signal transducer is attached to the receptor, which converts the interaction between the receptor and analyte into useful measurable signal. Diverse variety of biosensors have been developed by integrating nucleic acids, proteins, cells and tissues as receptors coupled with different signal transducing agents. With the need of more specific and sensitive biosensors, stand-alone protein biosensors are now being developed in which both receptor and signal transducers are coupled in a single polypeptide chain. Protein engineering plays an important role in the development of ligand specific receptors and in generation of new receptors for the analytes where no prior receptor or binding information is available. Protein engineering also provides the way to modulate the receptors in such a way that it gives a profound signal upon interaction with the analyte. Many different protein only sensors have been constructed using different protein engineering techniques including site directed mutagenesis, random insertion and deletion.

Benito et al. have inserted the foot and mouth disease virus serotype C1 in beta galactosidase at the sites that are important for the stabilization of active site. Binding of specific antibodies to the antigenic sites caused an increase in the beta galactosidase activity up to 200 %. This indicates that this enzymatic biosensor can be used for the diagnosis of foot and mouth disease. A molecular sensor based on the alkaline phosphate has been designed to detect antibodies. Hybrid protein containing the wild type alkaline phosphatase and its mutant (D101S and D153G) were used. Peptide epitope is inserted between the amino acids 407 and 408 in wild type alkaline phosphatase, which provides the binding site for anti-epitope antibodies. The enzymatic activity was increased by 400 % upon the binding of antibody. They also found that the modulation in the enzymatic activity is not specific for the particular epitope sequence or to particular antibody—epitope combination. Therefore these signaling molecules can be used to tag macromolecules in vivo or for the detection of other macromolecules like receptors, proteins, or hormones. Geddie et al. fabricated P53 based molecular sensors with the peptides recognized by HIV protease or monoclonal antibodies specific to HA, HSV, and LF (lethal factor) epitopes using site directed insertion mutagenesis and heterologous expression. These sensors have wide scale applications in high throughput screens and can act as in vivo sensors for various processes including DNA damage, hypoxia, spindle damage, temperature shock oncogene activation etc., where wild type P53 protein is also involved.

Naturally occurring fluorescence proteins used for live cell imaging exhibit several pitfalls, which have been answered by generating the fluorescence protein variants by the aid of protein engineering and directed evolution techniques. Ai et al. designed the monomeric fluorescence

proteins and increased its brightness by preparing the variants using directed evolution methods.

Further, many FRET based biosensors are also being designed. Zhang fabricated a FRET based biosensor by designing tandem Green fluorescence protein (GFP) fusion proteins that were able to detect proteolytic activity of thrombin. Tandem GFP fusion protein containing a thrombin specific recognition sequence was incorporated in between a cyan-emitting mutant of the green fluorescent protein and an enhanced yellow-emitting fluorescent protein. An increase of 4.6 fold in the fluorescence emission ratio was detected upon the addition of thrombin. This FRET based probe was also tested for dose-dependent effects of thrombin specific inhibitor 'hirudin'. Results showed the sensitivity of fluorescence emission ratios at the sub-nanomolar concentrations of hirudin, indicating that these probes can be used efficiently for high throughput screening of protease inhibitors.

Virus Engineering

Protein engineering plays an important role in engineering virus particles with high stability and activity. Engineered virus particle plays manifold roles in biomedicine, biotechnology and nanotechnology. Potential applications include designing of novel vaccines, vehicles for gene therapy, drug delivery, molecular imaging agents, for construction of nanomaterials. Owing to their wide scale applications in diverse field, it is essential to improve the physical stability of these virus particles to meet the demands for their efficient applications. Protein engineering strategies including rational designing, directed evolution and the combinatorial approaches are being employed to improve the physical properties of virus particles.

Biomedical Applications

Protein engineering strategies have also revolutionized the medical field by the efforts made in the era of protein based therapeutics. Both protein therapeutics and protein engineering fields emerged in early in 1980s. Protein based therapeutics are divided into different groups based on their molecular type. They include; antibody-based drugs, anticoagulants, blood factors, bone morphogenetic proteins, engineered protein scaffolds, enzymes, Fc fusion proteins, growth factors, hormones, interferons, interleukins, and thrombolytics, with antibody based drugs as the largest and fastest growing class of protein therapeutics. Therefore, protein engineering field thankfully returned many diverse variety of protein therapeutics with improved activity, specificity, stability, pharmacodynamics, pharmacokinetics, reduced immunogenicity and improved productivity.

Engineered Proteins as Therapeutics

Insulin—Human insulin is one of the first protein based therapeutic obtained by recombinant DNA technology developed by Eli Lilly at Gentech. Because of several drawbacks of the recombinant insulin, several analogs of insulin have been engineered by replacement of one to three amino acids. These analogs exhibit rapid and prolonged actions and also mimic the properties of endogenous insulin. The rapidly acting analogs include insulin aspart, insulin lispro, and insulin gelusine and the analogs with prolonged actions include insulin detemir and insulin glargine. The commercially most important engineered long acting insulin formulation is lantus®. Based on the lesson learned from these engineered analogs, Supramolecular protein engineering principles were applied to design "zinc" stapled insulin hexamers, in which zinc was stapled between the protein assemblies. His substitutions at ith and i + 4th amino acid residue (Glu → His and Thr → His) pair were introduced at an alpha-helical surface in A chain of Lantus. Thus the crystal structure contained both the conventional axial zinc ion and novel zinc ion at hexamer–hexamer interface. Pharmacological properties of the zinc stapled insulin were enhanced, resulting in long acting insulin depots. This analogue when compared with Lantus, was able to distinguish between insulin receptor and mitogenic insulin-like growth factor receptor, 30 times more stringently than Lantus, which indicates enhanced specificity of the analogue. Both factors including supramolecular assembly and receptor selectivity of the analogue contributes to the safety and efficacy of the insulin therapy. pH sensitive hydrogels have also been investigated for the self-regulating release of insulin for the treatment of diabetes. Glucose oxidase present in the matrix of the gel convert the diffusing glucose into gluconic acid, thereby lowers the local pH of the gel, causes the gel swelling, that finally, leads to the release of insulin. The more detailed information about the insulin analogues and their therapeutic applications has been reviewed well by Berenson et al.

Enzyme based Therapeutics—Enzymes play an important role as therapeutics to target molecules present in the extracellular environment. Amino acid degrading enzymes are primarily being used to develop anticancer agents in cases, where the rapidly growing tumor cells are auxotrophic to particular amino acid, depletion of which, in plasma results in inhibition in the growth of cancer cells. For instance, lymphoid tumor cells, lacking aspargine synthetase activity, are auxotrophic for aspargine. Recombinant PEGylated L-asparginase (Oncaspar®, Enzon) is being used for leukemia treatment. Other enzymes including PEG-arginine deiminase (ADI-PEG 20) for the treatment of arginine-auxotrophic tumors, melanoma and hepatocellular carcinoma are under clinical trials. These enzymes are PEGylated, to

improve their pharmacokinetics and to reduce the risk of immunogenicity. Dornase alfa or recombinant human DNase I (Pulmozyme®, Genentech) is used in the form of aerosol for the treatment of cystic fibrosis. Frequent bacterial infections and lysis of neutrophils in the lungs lead to the release of DNA, which in turn converts into viscous mucus. Recombinant human DNase degrades the DNA to improve the functioning of lungs.

Enzyme replacement therapy involves the treatment of diseases in patients by the replacement of particular enzyme for which they are deficient of. This therapy has been applied for numerous genetic diseases including lysosomal storage disorders. Number of ERT based FDA approved drugs are now available for many diseases. Fabrazyme® (agalsidase beta) for Fabry Disease; Cerezyme® (imiglucerase), VPRIV™ (velaglucerasealfa), Elelyso™ (taliglucerase), for type I Gaucher disease; Lumizyme® (alglucosidasealfa) for Glycogen Storage Disease type II (Pompe disease); Aldurazyme® (laronidase) for MPS I (Hurler, Hurler-Scheie, or Scheie syndrome); Elaprase® (idursulfase intravenous) for MPS II (Hunter disease); Naglazyme™ (galsulfase) for MPS VI (Maroteaux-Lamy syndrome).

Antibodies—Protein designer have lend themselves towards antibody based therapeutics. They aimed at humanizing the therapeutic antibodies. Principles of protein engineering have been applied for improvising the antibodies with respect to various aspects like increasing binding affinity, specificity, stability. Miklos et al. proposed a generalized strategy to design super charged highly thermo resistant antibodies and exemplified this approach by designing a single-chain variable fragment antibody (anti-MS2 scFV) by substituting up to 14 residues with arginine or lysine that showed heightened resistance to thermal inactivation and 30-fold improvement in antigen binding affinity. Fleishman et al. exploited Patch Dock and Rosetta Dock for designing two proteins (HB36 and HB80), that binds to conserved surface patch of the influenza hemagglutinin (HA) from the 1918 H1N1 pandemic virus. These novel proteins promise to serve as templates for future drugs. Recently, one of these HA stem binding proteins have been optimized using deep mutational scanning for their tighter binding and have also been proved to offer protection against influenza virus infection in vivo independent of host immune response. Scientific community is also geared toward designing smaller antibodies with stronger affinity and specificity. Pantazes and Maranas made a contribution in this area by developing a computational method, Optimal Complementarity Determining Regions (OptCDR) for designing antibodies based on complementarity determining regions (CDRs). CDRs, also known as hypervariable regions, present on antibodies where majority of antigen-antibody interactions takes place. Many antibodies including that destined for peptides from

the capsid of hepatitis C, fluorescein, and vascular endothelial growth factor (VEGF), have been designed using optCDR thus opening a way to generate diverse antibody libraries. Advancements have been made in epitope based vaccine design which includes grafting of epitopes onto the protein scaffolds. Scaffolds exploitation in protein based drug designs and therapeutics is being followed for long time such as in immunoglobulin like proteins, DARPins, cysteine knots etc., as discussed in previous sections. Similar strategy is now being extended for epitope based drug design. Correia et al. designed epitope-scaffolds for HIV4E10 epitope, that shared high structural similarity to the epitope, and exhibited high affinity for binding the monoclonal antibody 4E10. Correia et al. developed a strategy known as FFL (fold from loops) for designing epitope scaffolds, and designed an epitope scaffold that triggered respiratory syncytial virus (RSV)—neutralizing antibodies in rhe-sus macaques. Thus these scaffolds can serve as templates for future vaccine development against RSV.

GPCR antibodies—G-protein coupled receptors (GPCRs) are an important class of cell surface proteins that conveys the messages in the external environment to the intracellular effector molecules to carry out cellular signaling processes. They participate in number of biological processes occurring in cell including homeostasis, proliferation, migration of cells, and other sensory functions. They do their job by interacting with large array of molecules including proteins, small peptides, nucleotides, small organic compounds. Due to their critical roles in variety of processes, they are also associated with large number of diseases including infection, inflammation, and cancer, which makes them a fascinating therapeutic target for the treatment of these diseases. Large number of anti-GPCR antibodies has been developed but not even a single antibody targeting GPCR has been approved by FDA. Major hurdles in development of anti-GPCR antibodies includes high variability in GPCR extracellular region, limited exposure of GPCR extracellular epitopes along with difficulty in development of efficient antibody screening tools. Despite these difficulties, several of anti GPCR antibodies have been developed and are under clinical trials.

Cardiovascular therapeutics—Protein engineering field also made tremendous efforts in generating cardiovascular therapeutics. The main target is to engineer proteins that can enhance cardiac microvasculature formation. Proteins with the potential to induce cardiomyocyte proliferation have been identified as a forward direction for protein based approaches for cardiac regeneration. Major limitation in natural/endogenous protein based therapeutics include, insufficient bio-availability and bioactivity, undesirable pharmacokinetics, bio-distribution patterns, and off target effects. All these can be overcome by the usage of protein

engineering principles and techniques to generate reliable and large quantities of protein based therapeutics. Large number of proteins engineered for cardiovascular therapy along with protein engineering techniques exploited for their design and synthesis have been reviewed by Jay and Lee.

Coagulation factors—Advancements have been made in the development of protein based therapeutics for the treatment of bleeding disorders including Hemophilia. Transfusion based treatments gained momentum in 1970s and 1980s, resulted in an increased risk of acquiring blood borne pathogens such as HIV and hepatitis C. With the advent in recombinant DNA technology, molecular biology tools and sequencing, many recombinant clotting factors including rFVIII, rFIX, rVIIa with improved functions, are being expressed and purified, that have been considered as potent gene therapy strategy for blood related diseases. For instance, Recombinant activated protein C (APC), an anticoagulant enzyme, reduced the mortality rate in severe sepsis patients was found to increase bleeding complications due to its anticoagulant activity. In order to reduce the risk of bleeding complications, its anticoagulant activity was reduced by site directed mutagenesis of mutations of residues in two surface loops of APC that resulted in two APC variants, R229A/R230A and KKK191-193AAA. These mutants showed reduced anticoagulant activity but retained their apoptotic activity. Such APC variants have been suggested to decrease the bleeding risks while providing the benefits to the cells.

Cytokines as therapeutic Agents—Cytokines are small soluble proteins secreted mainly by leukocytes and also by some other cell types including fibroblasts, endothelial cells, and epithelial cells. Cytokines plays key roles in modulating immune system by interacting with their receptors present on cell surfaces. Cytokine family of proteins includes chemokines, interferons, colony stimulating factors, and interleukins. Because of their manifold roles in immune system, many of the cytokines have been identified for their therapeutic potentials in treatment of various diseases including inflammatory, autoimmune, malignant and other infectious diseases. IFN-α and IFN-β were the first cytokines that were cloned and synthesized. Many recombinant cytokines have been approved for the treatment of various diseases. For example, recombinant IFN-α has been approved for the treatment of hairy cell leukaemia, recombinant interleukin-2(IL-2) was approved for the treatment of metastatic melanoma. Additional cytokines approved for treating number of diseases have been summarized in the review by Lombardi et al. Cytokines in addition to their protective effects also exhibits adverse effect due to their inherent properties such as: (a), they are pleiotropic, implies that cytokines exert their effects on multiple cell types. (b) They need to be administered at high doses of

cytokines due to their short serum half-life. High doses lead to pleiotropic effects of cytokines that causes adverse effects in treatment of diseases. Thus, there is a high need to modulate their activities that can be accomplished by principles of protein engineering.

Many cytokines have been engineered to optimize their therapeutic potential as well as to overcome their adverse effects. Number of strategies including chemical modification (PEGylation), fusion with other proteins or immuno complexing, mutagenesis, has been exploited to engineer an optimal cytokine with increased serum half-life and enhancement in their specific activity towards their target. Large numbers of cytokines are under clinical and preclinical studies that will result in next generation cytokine therapeutics with improved pharamacokinetic and pharmacodynamic properties.

PEGylated cytokines including PEG-G-CSF, PEG-IFN-α2a and PEG-IFN-α2b have been approved for the treatment of Chemotherapy-induced neutropenia, Chronic hepatitis B/C, Chronic hepatitis C respectively. Recombinant G-CSF also known as filgrastim, used for the treatment of neutropenia has to be administered as daily dose over the course of chemotherapy cycle due to its short serum half-life. This discrepancy has been overcome by PEGylated form of G-CSF (pegfilgrastim), that exhibits prolonged half-life, therefore can be administered only once per chemotherapy cycle. Similarly, the half-life has also been increased in case of IFN-α2a and IFN-α2b along with higher therapeutic potential for the treatment of chronic hepatitis C.

Studies have also reported that cytokines in complex with antibodies showed improvement in their serum half-life and their pharmacological activity. Boyman and Sprent reported that injecting monoclonal antibody specific to IL-2 resulted in proliferation of CD8+ T-cells. But the use of combination of this mAb with recombinant IL-2, resulted in massive increase in proliferation of CD8+ T-cells, indicating that coupling cytokines with antibodies can be utilized to enhance or inhibit the immune responses.

To further improve or modify the functions or biophysical characteristics of cytokines, fusion cytokines are being produced in which cytokines are fused with other biological molecules. For example, fusion to Fc region of antibody, albumin or transferrin to increase their half-life, fusion with cytokine agonists to increase their activity, fusion to bacterial toxins to enhance their cytotoxicity, fusion to antibodies for their localized delivery. Examples of cytokines falling in each of these categories have been shown in Figure and well described in the review

by Lombardi et al. Cytokine mutagenesis is another strategy that has been employed in order to improve the activity, specificity and half-life.

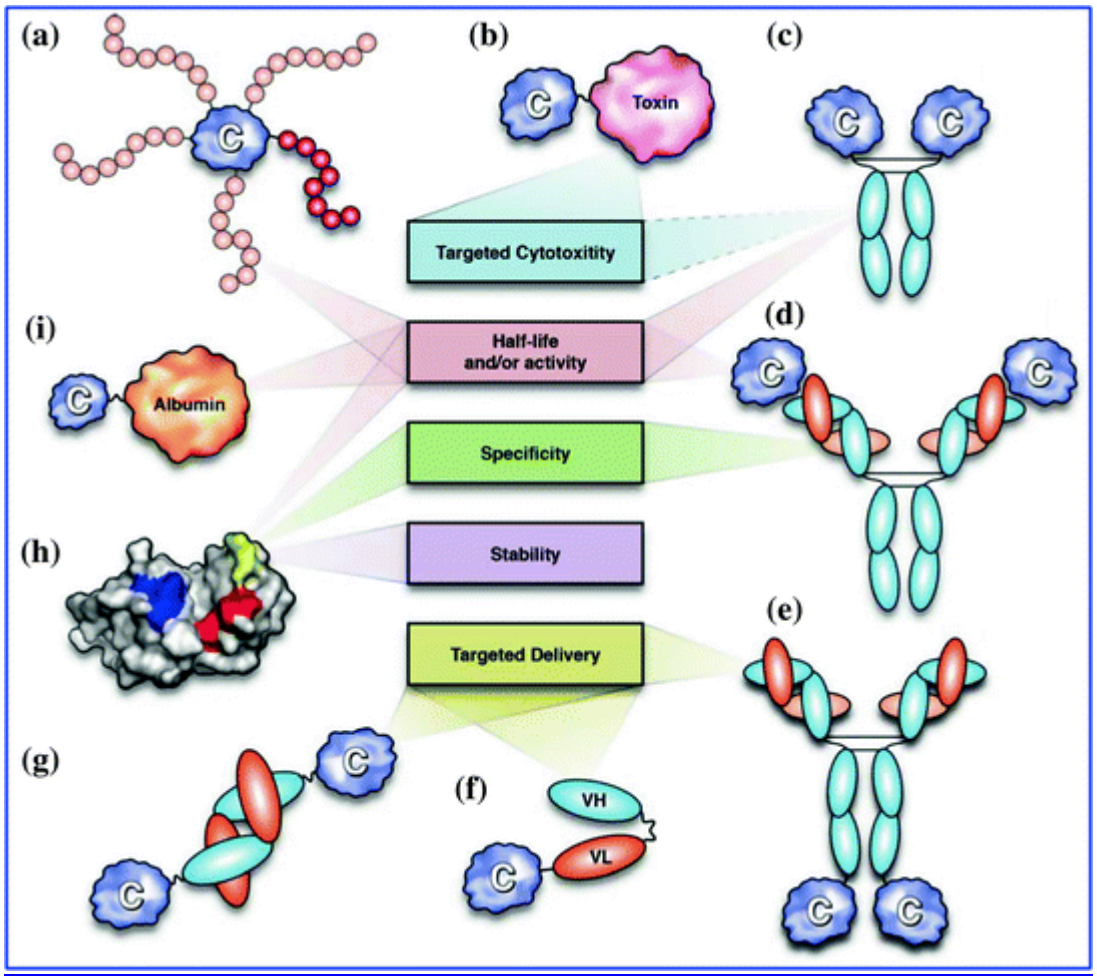


Figure 6

Molecular engineering strategies utilized for cytokine optimization and their major effects. A PEGylation (positional isomers displayed); B cytokine-toxin fusion; C cytokine-Fc fusion; D antibody-cytokine immune complex; E–G immunocytokines; E cytokine-IgG; F cytokine-scFv; cytokine-diabody; H cytokine mutagenesis; I cytokine-albumin fusion (Adapted from Lombardi et al.)

Proleukin®, recombinant form of IL-2, in which free cysteine residues were mutated to serine residues to prevent the formation of unnecessary disulfide bonds leading to protein aggregation. Further, an increase in thermal stability and decrease in protein aggregation of G-CSF was done by mutating helix breaking residues (glycine and proline) to residues encompassing high helical propensity. Variants of G-CSF have also been designed involving mutations of the

receptor binding residues to histidine, process known as “histidine switching” to increase their half-life.

In order to overcome the pleiotropic effects of cytokines, their specificity has been increased by generating the variants using rational designing and directed evolution approaches. As exemplified by superkine 2, a variant of IL-2, developed using in vitro evolution method exhibits nearly 250 fold increased affinity for IL-2Rβ. By the combined effort of both rational design and directed evolution method, IL-4 superkines with increased specificity towards the receptors to perform distinct immunological functions have been developed. These studies imply that cytokines can be redirected towards specific target proteins or cell types to elicit specific actions, thus providing a platform to improve cytokine based therapy.

Cytokine antagonists are also being developed as therapeutic agents. These molecules disrupt the binding of cytokines to their receptors thereby blocking their actions. Pitrakinra, IL-4 variant has been designed rationally inhibited the IL-4/IL-13-mediated proliferative effects in vitro and reduced allergen-induced inflammation in animal models of asthma and skin inflammation. Pitrakinra is under phase II clinical trials for the treatment of allergic asthma and atopic eczema. Similarly, IL-6R antagonist has been designed for the treatment of multiple myeloma and lung fibrosis.

Protein Scaffolds as Therapeutics

Number of protein therapeutic have been developed using engineered protein scaffolds. These scaffolds provide the binding sites that can be modulated according to specific target recognition molecules. Protein scaffolds selected for therapeutic design should be small, soluble, monomeric, highly thermodynamically and chemically stable, without any disulfide bonds/glycosylation sites. They should be readily and highly expressible in microbial host preferably in the cytoplasmic compartment. They should contain surface exposed loops that can serve as binding sites for target molecule. These loop regions can undergo several modifications and should provide a sufficient surface area for highly specific binding and isolation of target molecules. These scaffolds can be monoclonal antibodies or non-antibody scaffolds. Major protein scaffolds used for therapeutics includes:

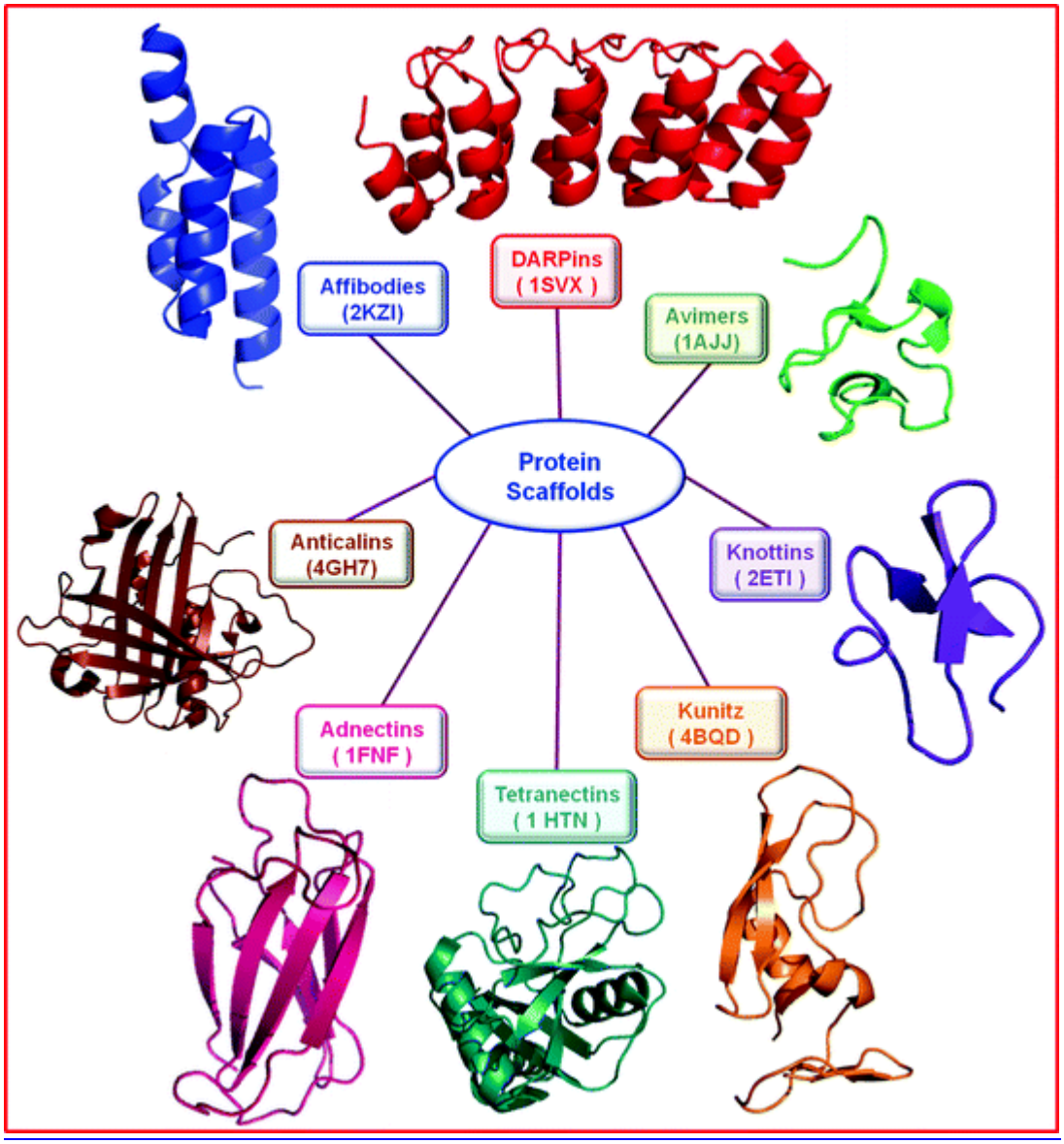


Figure 7

Protein scaffolds for imbibing biomedical applications

Knottins—Knottins, also known as cystine-knot mini proteins, are small proteins characterized by a cystine-knot. These polypeptides carry out diverse functions including ion channel blockade, protease inhibitions, and antimicrobial activity. They are present in plants, animals, fungi and also found in toxins released from spiders, scorpions and snails. These small proteins are approximately 30 residues in length and exhibits common tertiary fold characterized by three antiparallel β-strands connected by loops of variable length and three pairs of intra disulfide bonds. Three pairs of disulfide bonds are between Cys1 and Cys4, Cys2 and Cys5,

and Cys3 and Cys6. Knotted structure is formed as a result of disulfide linkage between Cys3 and Cys6, which penetrates through the macro cycle formed by the two other disulfide bonds and the peptide backbone.

Structure of knottins, held by the covalent linkages, make them chemically, thermally and proteolytically more stable. Knottins maintains their structure and conformation intact even if boiled at high temperatures (above 60 °C), for weeks, placed in 1-N HCl or 1-N NaOH for long time. Such stability in their structures coupled with other properties including small size, non-immunogenic, confers knottins as promising candidates for various therapeutic and diagnostic applications. Interconnecting loops in the structure provides the sites to engineer knottins with diverse molecular recognition properties. Previous studies exploiting cystine knot of Ecballium elaterium trypsin inhibitor II (EETI-II) and truncated forms of human agouti-related protein (AgRP*) serves as promising scaffolds to engineer biologically active proteins. Further, studies have also highlighted that loops are the major determinants for molecular recognition and binding by EFTI-II or AgRP*, therefore, loops can be reengineered to alter their molecular recognition and binding specificities.

Disintegrins, containing RGD or KGD sequences blocks the fibrinogen binding to α(IIb)beta(3) thereby, inhibits platelet aggregation. RGD and KGD peptide sequences were grafted into the cysteine knots of EETI-II and AgRP* and the activity of engineered knottin variants was compared with RGD or KGD motifs alone. Variants were much more potent to inhibit fibrinogen binding, alpha(IIb)beta(3) activation and platelet aggregation as compared to peptides alone which indicates that structural scaffold and amino acid residues in the vicinity of grafted sequence, plays an essential role in the activity of the protein. Silverman et al. replaced a constrained six amino acid loop in AgRP with a nine amino acid loop containing RGD integrin recognition motif and created a library of 20 million variants of AgRP by randomizing the residues in the vicinity of RGD motif. Variants were subjected for screening protocol to isolate variants that binds specifically to the platelet integrin α(IIb)β(3) at nanomolar concentrations. AgRp variants with high affinity and specificity for integrin α(IIb)β(3) were obtained.

Knottins because of their small size and high stability have also been engineered for their molecular imaging applications. Combinatorial methods were used to isolate Knottin variants that are able to bind integrin receptors expressed on tumors and tumor vasculature as promising diagnostic agents to detect integrin expression in living cells. Radiolabeled version of

engineered AgRP peptide has proved to be a promising positron emission tomography (PET) imaging agent for the tumors expressing alpha(v)beta(3) integrin. Engineered EFTI-II knotins have also been conjugated to both near-infrared fluorescence (NIRF) and PET probes for multi-modality imaging and also can be used to detect deep seated tumors in the body.

Affibodies—Affibody molecules are new class of affinity proteins derived originally from B-domain of immunoglobulin binding region of staphylococcal protein A. A relatively short B-domain comprises of 58 amino acids that folds into three helix bundle structure and has been reported to exhibit one of the fastest folding kinetics. Using combinatorial approaches of protein engineering, Z domain with high chemical stability and with intact affinity for Fc part of immunoglobulins and lower affinity for Fab part, has been engineered by mutating several important residues in B-domain. Z-domain exhibit high affinity for their binding proteins is due to surface localization of defined set of amino acid residues. These domains also showed high solubility and expression patterns in several hosts either alone or in conjunction with fusion proteins. Z-domains can be used to engineer affinity binding molecules by altering or randomizing the amino acids present at sites involved in interaction with Fc part of immunoglobulins.

The libraries of affinity binding proteins were constructed by genetically randomizing the 13 surface located amino acids in Z-protein scaffold using the combinatorial protein engineering techniques. All these positions were confined to the first 2 helices with 7 positions in helix-1 and 6 positions in helix-2. Majority of these positions are involved in interaction with Fc domain of human IgG. From these naive libraries, first generation affibody molecules, were selected that were re-randomized to create secondary libraries, and further subjected to more stringent selection criteria.

Affibody molecules specific to different proteins including HER2, EGFR, insulin, transferrin, fibrinogen, tumor necrosis factor-a, IL-8, gp120, CD28, human serum albumin, IgE, IgA, IgM engineered with affinities (KD) ranging from μM to pM. Affinity of some of these molecules further improved either by helix shuffling or sequence alignment in combination with directed combinatorial mutagenesis. For instance, the Taq DNA polymerase specific binding protein (affibody), obtained from combinatorial naive library of Z domain was subjected to further create a hierarchical library by selective randomization of six amino acid positions in one of the two alpha-helices of the domain, that are involved in Taq DNA polymerase binding. Variants selected by monovalent phage display technology, showed Taq DNA polymerase

binding affinities in the range of 30–50 nM as dictated by biosensor assay. Further, improvement in the specificity of affibody molecules for cancer specific ligands resulted in the promising candidates for tumor imaging. For example, an increase of 2200 fold in the affinity has been achieved in (human epidermal growth factor receptor 2) HER2-specific affibody molecule, that can be employed to visualize HER2 expression in tumors using gamma camera.

Affibody molecules have also been generated using chemical synthesis so as to incorporate specific chemical groups inculcating specific chemical activities in the engineered affibody molecules. Site-specifically triple-labelled three-helix bundle affinity proteins have been generated by chemical synthesis, in which three reporter groups namely, 5-(2-aminoethylamino)-1-naphthalenesulfonic acid (EDANS) and 6-(7-nitrobenzofurazan-4-ylamino)-hexanoic acid (NBDX), (constituting a donor/acceptor pair for fluorescence resonance energy transfer), and a biotin moiety, (for surface immobilization) were incorporated at particular sites. CD and biosensor studies showed proper folding and binding specificities in the engineered affinity proteins. These proteins were also shown to act as fluorescence biosensors to specifically detect unlabeled human IgG and IgA. Further, this class of proteins exhibiting manifold biotechnological, diagnostic and therapeutic applications, have been reviewed well by Löfblom et al. and Nygren.

Apart from above described protein scaffolds, several other engineered protein scaffolds are also available in market for various biomedical applications include:

Single-domain antibodies from humans are small 11–15 kDa proteins that comprises of either variable heavy chain or variable light chain domains that carries set of three complementarity determining regions (CDRs) providing the specific binding site to the target antigen.

Small modular immuno-pharmaceuticals (SMIPs) are small artificial proteins, composed of parts of antibodies, and are intended to be used as pharmaceutical drugs. SMIPs are single chain polypeptides containing target binding domain and an effector domain connected by hinge region. Binding domain can be single chain variable fragments that can be modified in different ways to bind wide variety of proteins including soluble proteins and cell surface receptors. Hinge region from immunoglobulin G1, provides flexibility as well as the sites for the association of multiple SMIPs. CD37-SMIP has been designed, which is specific to CD37 expressed on the surface of B-cells, and has been strongly recommended to act as therapeutics for B-cell malignancies. Other SMIPs includesTRU-015, an anti-CD20 IgG fusion protein, developed for the treatment of rheumatoid arthritis and may also be helpful in treatment of B-

cell neoplasms and other autoimmune diseases, TRU-016, an anti-CD37 IgG fusion protein has also been developed for the potential treatment of B-cell malignancies, including chronic lymphocytic leukemia (CLL) and non-Hodgkin's lymphoma (NHL), as well as for autoimmune and inflammatory diseases.

Tetranectins are C-type lectin like, homotrimeric plasma and tissue proteins that were identified by Borean Pharma as protein engineering platforms to create superior quality antibody analogues. Exact biological function of tetranectin is not known but it may be involved in fibrinolysis and proteolysis during tissue remodeling. Tetranectin is 181 residues long single polypeptide chain, comprises a C-terminal binding domain and N-terminal trimerisation domain.

Target binding domain contains five loops whose amino acids can be varied to bind different target molecules, either proteins or oligosaccharides. The monomeric protein, in solution condition, forms a trimer due to the coiled coil formation of trimerisation unit. This trimeric form of tetranectin has prolonged half-life and enhanced stability at physiological conditions without any exchange of trimeric unit among monomers. Human C-type lectin derived TNF antagonist has been developed based on C-type lectin domain (CTLD) library of proteins.

Adnectins are based on 10th fibronectin type III domain, contains three distinct loops analogous to CDRs of antibody that provides the variable regions to generate target binding sites but are much simpler than antibody without any disulfide bond. CT-322, a PEGylated, anti-angiogenic Adnectinis the first therapeutic Adnectin specific for VEGF and is under clinical trials.

A-domain proteins are non-antibody cysteine rich proteins that were first identified in low-density lipoprotein receptor (LDLR)-A module. A-domain proteins contains set of three disulfide bonds, an antiparallel β-sheet, 310 helix and a calcium binding site with both N-and C-terminal folded loops. These proteins bind their targets through multiple sites. Residues at the binding site can be randomly mutated to generate the sites for different target molecules.

Lipocalins are non-antibody secreted proteins, involved in transport of biomolecules, including steroid hormones, vitamins, odorants and several secondary metabolites. Lipocalins are 160–180 residues long proteins, encompasses eight antiparallel β-sheets that forms funnelled barrel like structure along with four loops at one end of the barrel structure. These loops with high flexibility in their structure provide the binding sites for various targets, which makes lipocalins an attractive protein scaffold. Lipocalins were exploited to design anticalin scaffolds which

exhibit potential applications as antidotes, antagonistic protein therapeutics or as target-recognition modules in a new generation of immunotoxins. Anticalins specific for human CTLA4 (cytotoxic T-lymphocyte antigen 4, a CD28-family receptor expressed on mainly CD4+ T cells) and vascular endothelial growth factor are under clinical trials.

Ankyrin Repeat is a 33 amino acid residue protein repeat comprising of β- hairpin-helix–loop—helix structure that was originally identified in cell cycle regulators and in cytosekeletal protein ankyrin. β-turn and loop regions containing non conserved residues serve as the regions to create chemically diverse sites for binding of different target molecules. This scaffold can be used to bind wide variety of target proteins as there is the combination of two variable factors including the number of repeats as per the need and chemical composition of repeats. Based on ankyrin repeat protein, small single domain proteins known as DARPins (Designed Ankyrin repeat proteins) have been designed, exhibiting multiple potential medical applications either alone or in conjugation with other effector moieties that can be either PEG to modulate its serum half-life, low molecular weight cytotoxic agents to kill cells, small peptides or whole proteins such as cytokines, toxins, antibody Fc domains or other DARPins cytokines. MP0112, a DARPin that inhibits all relevant forms of VEGF with high potency is under clinical testing in diabetic macular edema (DME) and wet age-related macular degeneration (wet AMD).

Avimers, are the multimeric binding proteins, engineered by an in vitro exon shuffling and phage display of large family of human extracellular receptor domains, resulting in a multidomain protein with binding and inhibitory properties. Linking multiple binding domains result in creation of higher affinity and specificity as compared to other single epitope binding proteins. AMG220, an avimer is under clinical trial against crohn's disease.

Kunitz domains belong to the class of protease inhibitors which reversibly inhibits trypsin and other serine proteases. These domains were reengineered to modulate their activity towards different proteases, thus making these domains attractive drug candidates against variety of proteases. Kunitz domain based engineered protein, DX-88 (Ecallantide), is potent and selective inhibitor of plasma kallikrein. This molecule has been approved for its therapeutic application against Hereditary angioedema. DX-890 (Depelstat), an inhibitor for neutrophil elastase is under phase II clinical trial for the treatment of acute respiratory distress syndrome. DX-1000, a plasmin inhibitor is under preclinical trials to be exploited as anticancer therapeutic agent.

As summarized from the work of several researchers across the globe, it is undoubtedly evident that protein engineering techniques and their applications have contributed immensely to the development of various industrial, biotechnological and biomedical fields as a next generation vision. Indeed, protein engineering methods have revolutionized the field of medicine as these engineered protein based therapeutics and biomaterials is undoubtedly the unique choice for treating large number of diseases. Expansion of protein therapeutics and all other industrially viable enzymes along with novel protein engineering strategies are in high demand for the betterment of human life. This young hybrid blooming field of protein engineering is still at its infancy and we strongly believe that the greatest achievements of the human intellect in this research area are yet to come for the well-being of all the organisms on the mother Earth.