Principles of Protein Structure, Comparative Protein Modelling and Visualisation

Nicolas Guex and Manuel C. Peitsch

GlaxoWellcome Experimental Research S.A.
16, chemin des Aulx
1228 Plan-les-Ouates / Switzerland
ng45767@GlaxoWellcome.co.uk,mcp13936@GlaxoWellcome.co.uk

The 5 first chapters of this course are adapted from the Internet course on protein structure by

J. Cooper, J. Walshaw & Alan Mills

The following chapters are by MC. Peitsch and N.Guex.

Part I: Introduction to Protein structure

Part II: Protein modelling

Part III: Model quality

Part IV: The SWISS-MODEL modelling environment

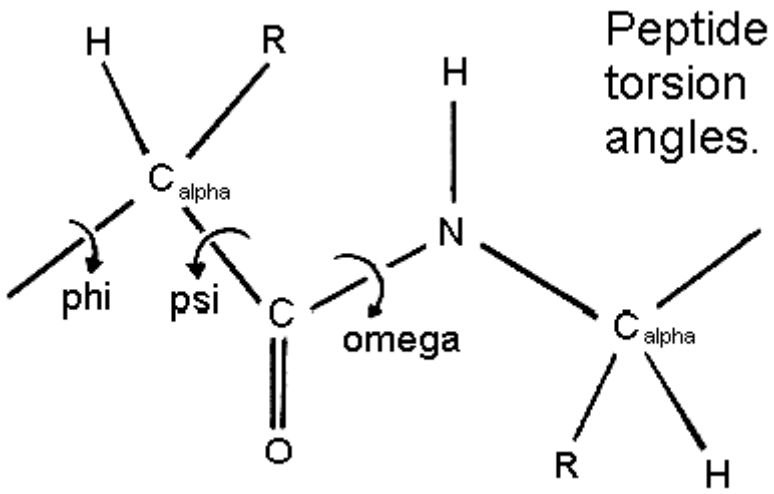
Part V: Examples and case studies

Appendices

1 Secondary structure and backbone conformation

1.1 Peptide Torsion Angles

The figure below shows the three main chain torsion angles of a polypeptide. These are phi (Φ), psi (Ψ), and omega (Ω).

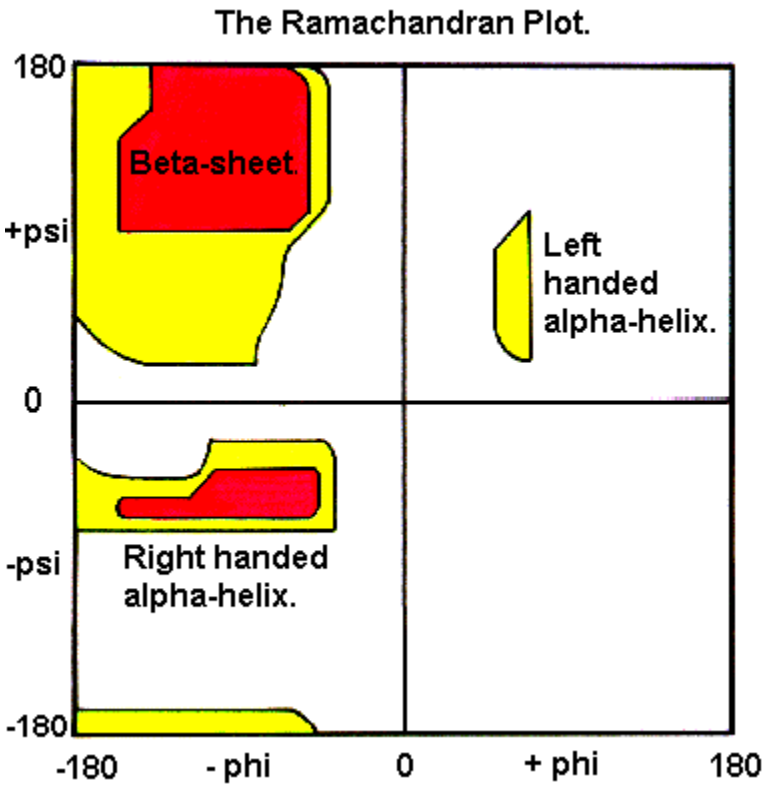


The planarity of the peptide bond restricts to 180 degrees in very nearly all of the main chain peptide bonds. In rare cases = 10 degrees for a *cis* peptide bond which usually involves *proline*.

1.2 The Ramachandran Plot

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

GN Ramachandran used computer models of small polypeptides to systematically vary and 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, and angles, which cause spheres to collide correspond to sterically disallowed conformations of the polypeptide backbone.

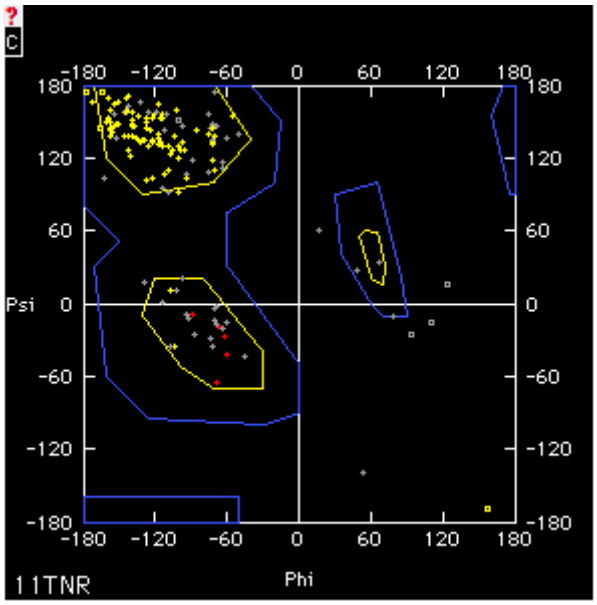


In the diagram above the white areas correspond to conformations where atoms in the polypeptide come closer than the sum of their van der Waals radii. These regions are sterically disallowed for all amino acids except glycine which is unique in that it lacks a side chain. The red regions correspond to conformations where there are no steric clashes, i.e. these are the allowed regions namely the α-helical and α-sheet conformations. The yellow areas show the allowed regions if slightly shorter van der Waals radii are used in the calculation, i.e. the atoms are allowed to come a little closer together. This brings out an additional region which corresponds to the left-handed α-helix.

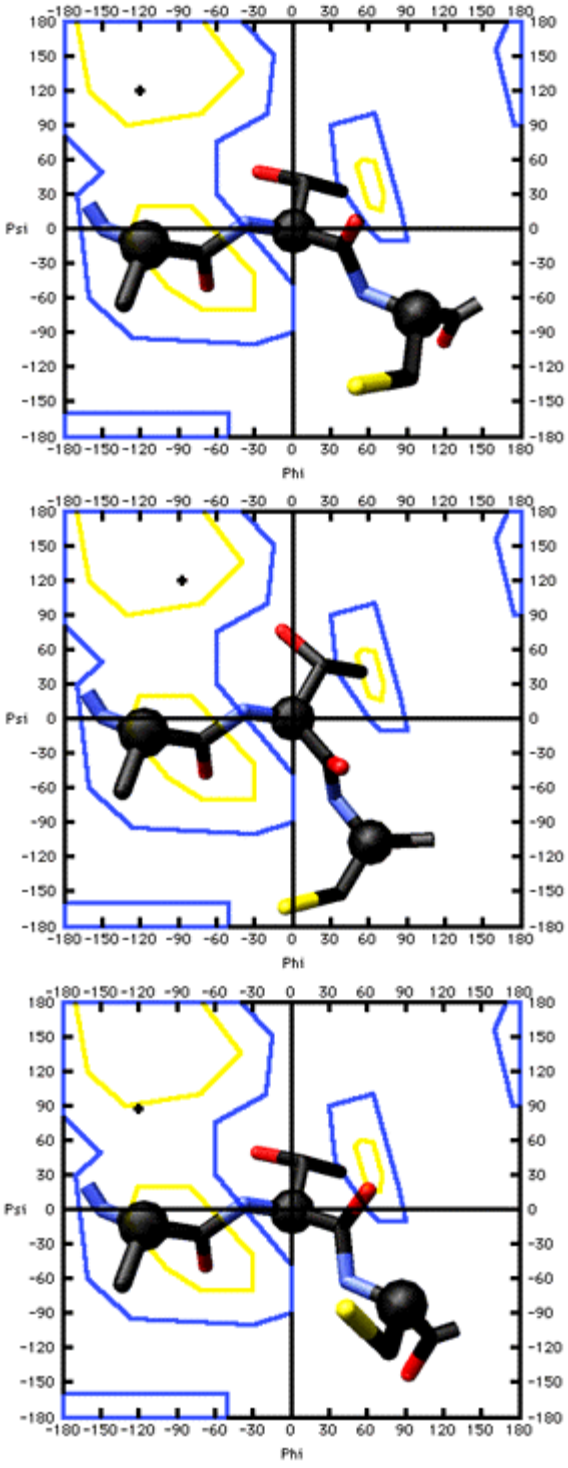
L-amino acids cannot form extended regions of left-handed helix but occasionally individual residues adopt this conformation. These residues are usually glycine but can also be asparagine or aspartate where the side chain forms a hydrogen bond with the main chain and therefore stabilises this otherwise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

Disallowed regions generally involve steric hindrance between the side chain C 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.

Below is a ramachandran plot of a protein containing almost exclusively beta-strands (yellow dots) and only one helix (red dots). Note how few residues are out of the allowed regions; and note also that they are almost all Glycines (depicted with a little square instead of a cross.



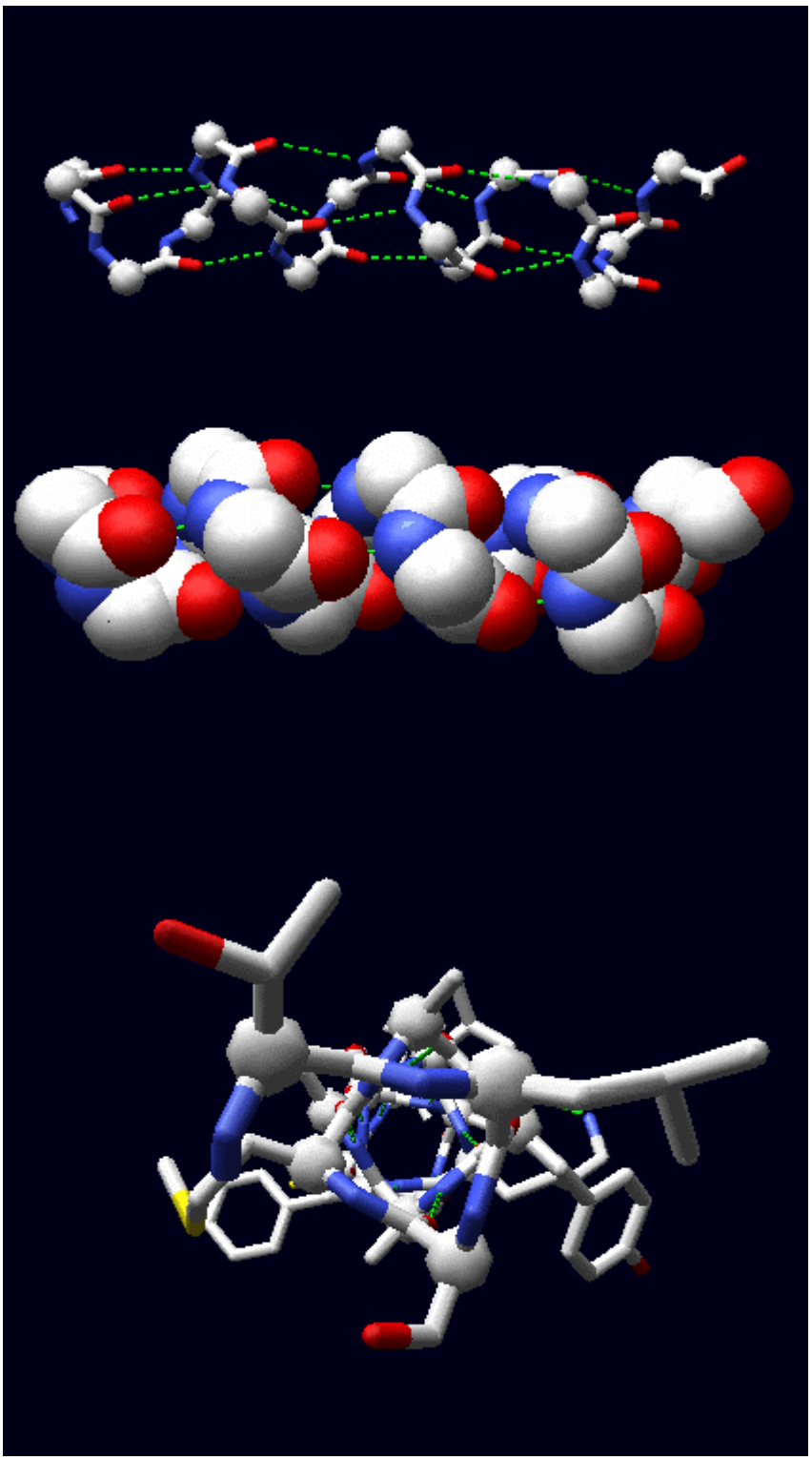
Observe the effect of minor Phi and Psi angle changes:



1.2 The α-helix.

1.2.1 Development of an α-helix structure model.

Pauling and Corey twisted models of polypeptides around to find ways of getting the backbone into regular conformations which would agree with α-keratin fibre diffraction data. The most simple and elegant arrangement is a right-handed spiral conformation known as the 'α-helix'.



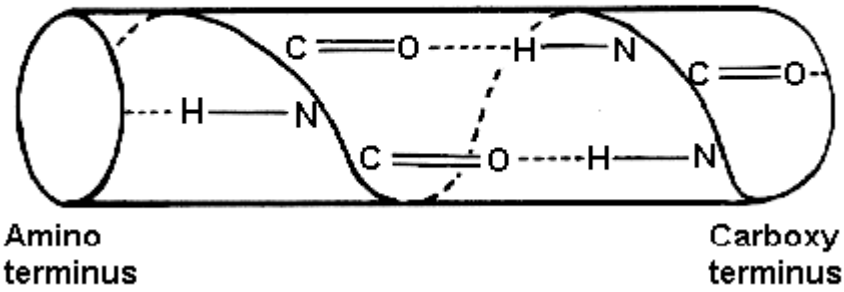
An easy way to remember how a right-handed helix differs from a left-handed one is to hold both your hands in front of you with your thumbs pointing up and your fingers curled towards you. For each hand the thumbs indicate the direction of translation and the fingers indicate the direction of rotation.

1.2.2 Properties of the α-helix.

The structure repeats itself every 5.4 Å along the helix axis, i.e. we say that the α-helix has a pitch of 5.4 Å. α-helices have 3.6 amino acid residues per turn, i.e. a helix 36 amino acids long would form 10 turns. The separation of residues along the helix axis is 5.4/3.6 or 1.5 Å, i.e. the α-helix has a rise per residue of 1.5 Å.

1. Every main chain C=O and N-H group is hydrogen-bonded to a peptide bond 4 residues away (i.e. O_i to N_{i+4}). This gives a very regular, stable arrangement.
2. The peptide planes are roughly parallel with the helix axis and the dipoles within the helix are aligned, i.e. all C=O groups point in the same direction and all N-H groups point the other way. Side chains point outward from helix axis and are generally oriented towards its amino-terminal end.

Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.



All the amino acids have negative phi and psi angles, typical values being -60 degrees and -50 degrees, respectively.

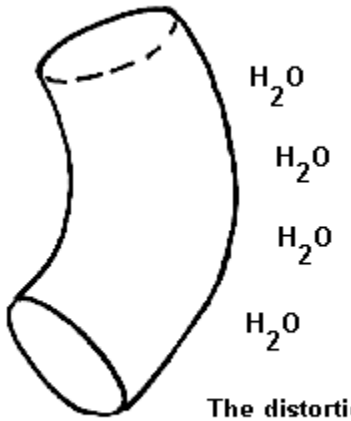
1.2.3 Distortions of α-helices.

The majority of α-helices in globular proteins are curved or distorted somewhat compared with the standard Pauling-Corey model. These distortions arise from several factors including:

1. The packing of buried helices against other secondary structure elements in the core of the protein.

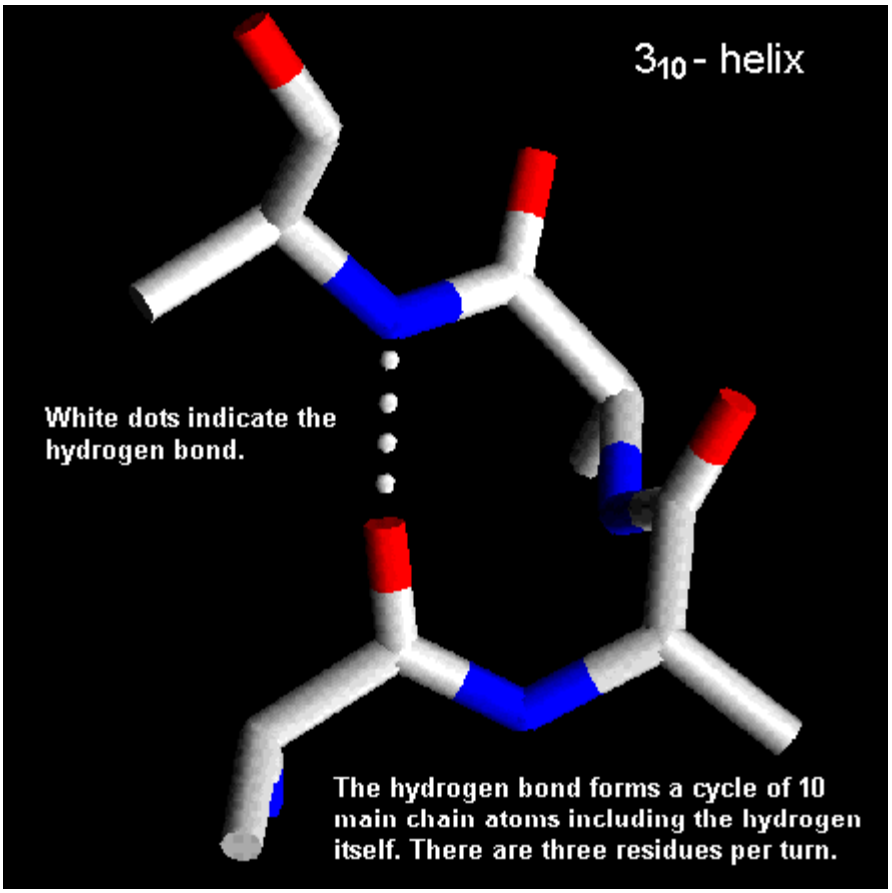
2. Proline residues induce distortions of around 20 degrees in the direction of the helix axis. This is because proline cannot form a regular α-helix due to steric hindrance arising from its cyclic side chain which also blocks the main chain N atom and chemically prevents it forming a hydrogen bond. Janet Thornton has shown that proline causes two H-bonds in the helix to be broken since the NH group of the following residue is also prevented from forming a good hydrogen bond. Helices containing proline are usually long perhaps because shorter helices would be destabilised by the presence of a proline residue too much. Proline occurs more commonly in extended regions of polypeptide.
3. Solvent. Exposed helices are often bent away from the solvent region. This is because the exposed C=O groups tend to point towards solvent to maximise their H-bonding capacity, i.e. tend to form H-bonds to solvent as well as N-H groups. This gives rise to a bend in the helix axis.

Solvent induced distortion of an alpha helix.



The distortion is exaggerated for clarity.

1. **3_{10}-Helices**. Strictly, these form a distinct class of helix but they are always short and frequently occur at the termini of regular α-helices. The name 3_{10} arises because there are three residues per turn and ten atoms enclosed in a ring formed by each hydrogen bond (note the hydrogen atom is included in this count). There are main chain hydrogen bonds between residues separated by three residues along the chain (i.e. O_i to N_{i+3}). In this nomenclature the Pauling-Corey α-helix is a 3.6_{13}-helix. The dipoles of the 3_{10}-helix are not so well aligned as in the α-helix, i.e. it is a less stable structure and side chain packing is less favourable.

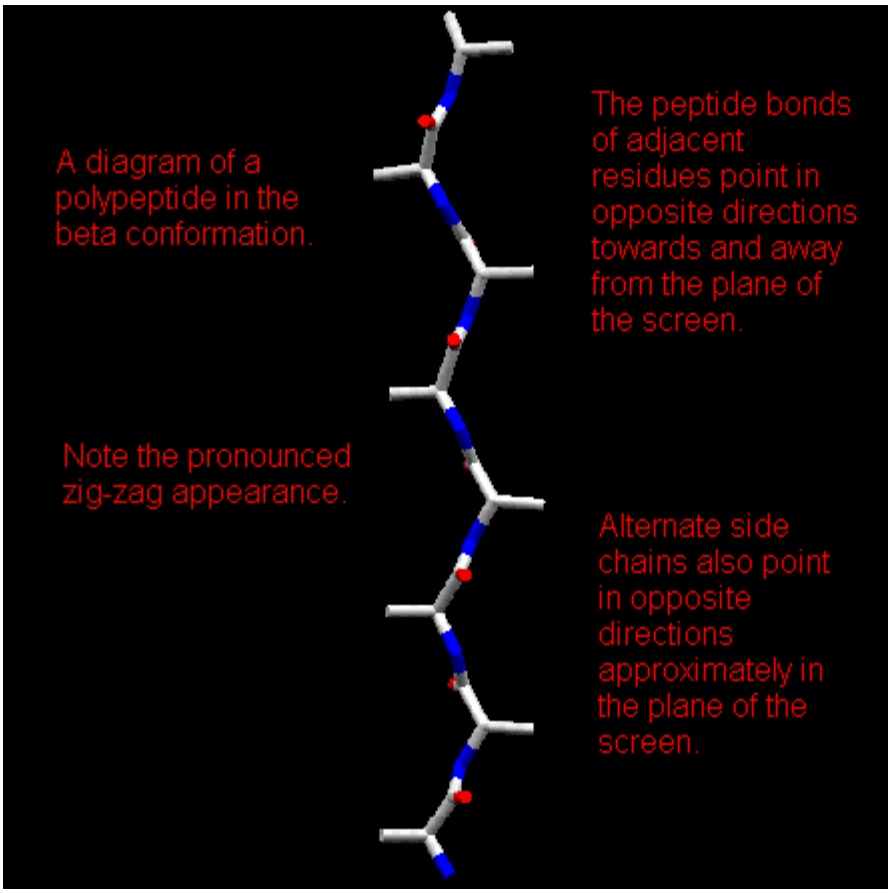


1.3 The β-sheet.

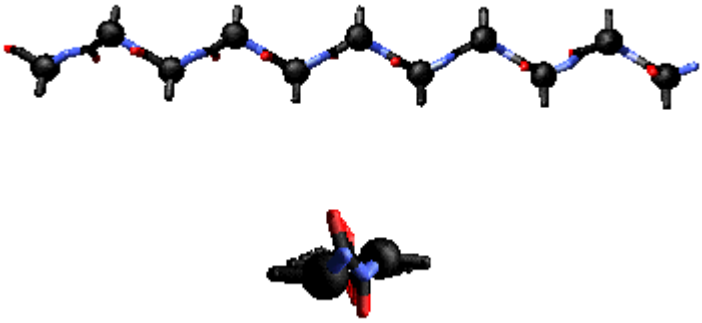
1.3.1 The β–sheet structure.

Pauling and Corey derived a model for the conformation of fibrous proteins known as β-keratins. In this conformation the polypeptide does not form a coil. Instead, it zig-zags in a more extended conformation than the α-helix. Amino acid residues in the β-conformation have negative Φ angles and the Ψ angles are positive. Typical values are Φ = -140 degrees and Ψ = 130 degrees. In contrast, α-helical residues have both Φ and Ψ negative. A section of polypeptide with residues in the β-conformation is referred to as a β-strand and these strands can associate by main chain hydrogen bonding interactions to form a sheet.

In a β-sheet two or more polypeptide chains run alongside each other and are linked in a regular manner by hydrogen bonds between the main chain C=O and N-H groups. Therefore all hydrogen bonds in a α-sheet are between different segments of polypeptide. This contrasts with the α-helix where all hydrogen bonds involve the same element of secondary structure. The R-groups (side chains) of neighbouring residues in a β-strand point in opposite directions.



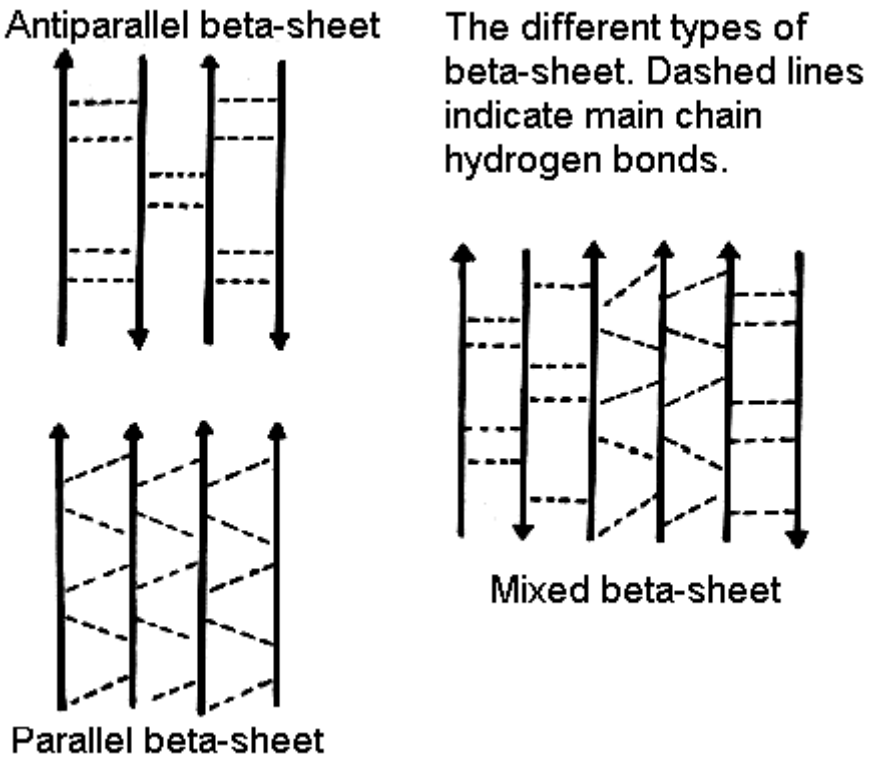
Imagining two strands parallel to this, one above the plane of the screen and one behind, it is possible to grasp how the pleated appearance of the β-sheet arises. Note that peptide groups of adjacent residues point in opposite directions whereas with α-helices the peptide bonds all point one way:



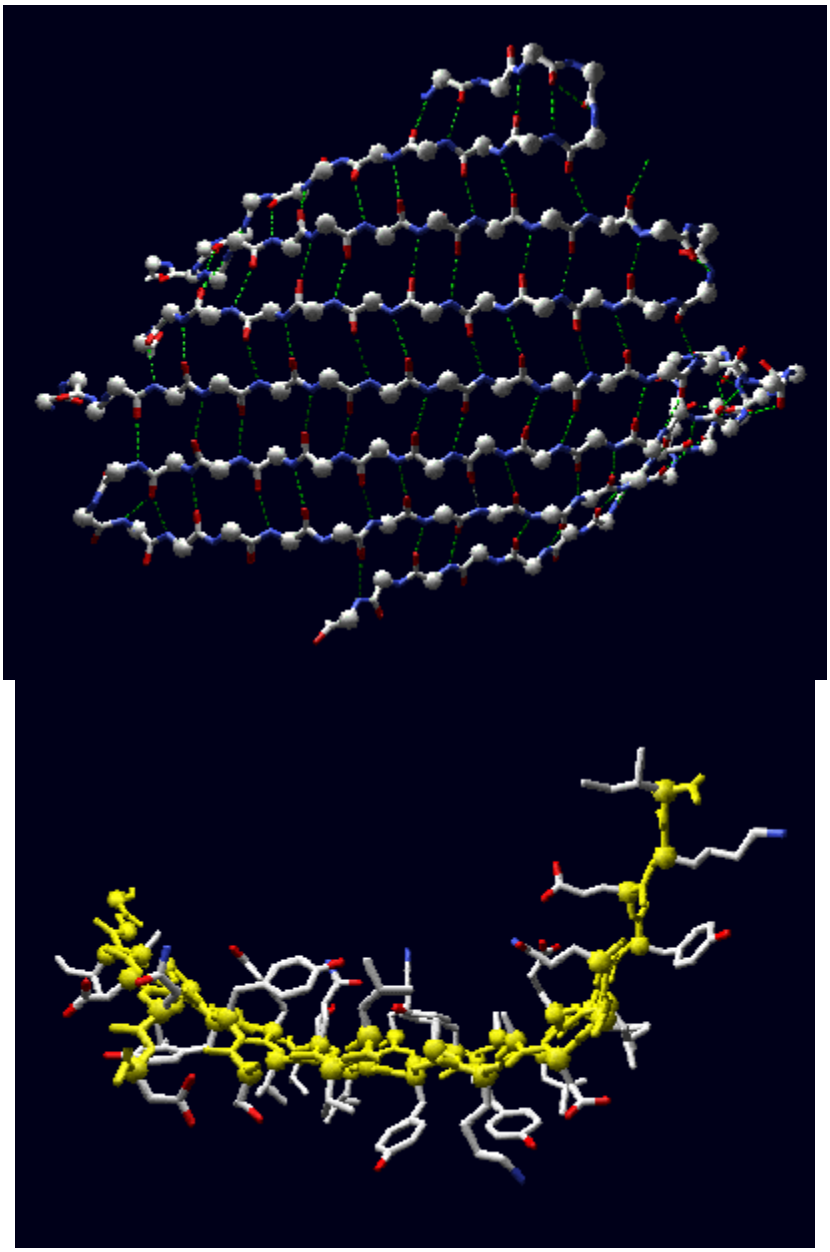
The axial distance between adjacent residues is 3.5 Å. There are two residues per repeat unit which gives the β-strand a 7 Å pitch. This compares with the α-helix where the axial distance between adjacent residues is only 1.5 Å. Clearly, polypeptides in the β-conformation are far more extended than those in the α-helical conformation.

1.3.2 Parallel, anti-parallel and mixed β-sheets.

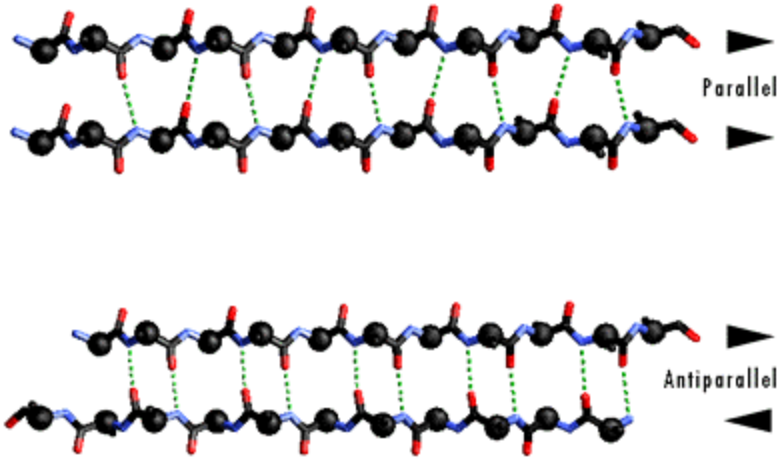
In parallel β-sheets the strands all run in one direction, whereas in anti-parallel sheets they all run in opposite directions. In mixed sheets some strands are parallel and others are anti-parallel.



Below is a diagram of a three-stranded anti-parallel β-sheet. It emphasises the highly regular pattern of hydrogen bonds between the main chain NH and CO groups of the constituent strands.

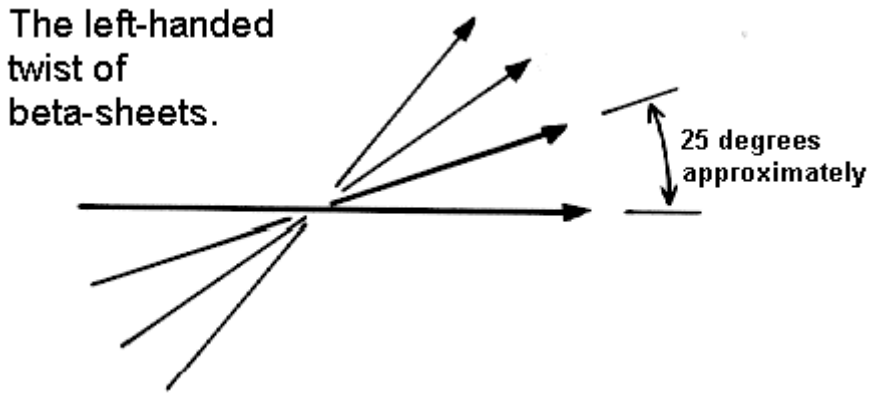


In the classical Pauling-Corey models the parallel β-sheet has somewhat more distorted and consequently weaker hydrogen bonds between the strands:



β-sheets are very common in globular proteins and most contain less than six strands. The width of a six-stranded β-sheet is approximately 25 Å. No preference for parallel or anti-parallel β-sheets is observed, but parallel sheets with less than four strands are rare, perhaps reflecting their lower stability. Sheets tend to be either all parallel or all anti-parallel, but mixed sheets do occur.

The Pauling-Corey model of the β-sheet is planar. However, most β-sheets found in globular protein X-ray structures are twisted. This twist is left-handed as shown below. The overall twisting of the sheet results from a relative rotation of each residue in the strands by 30 degrees per amino acid in a right-handed sense.

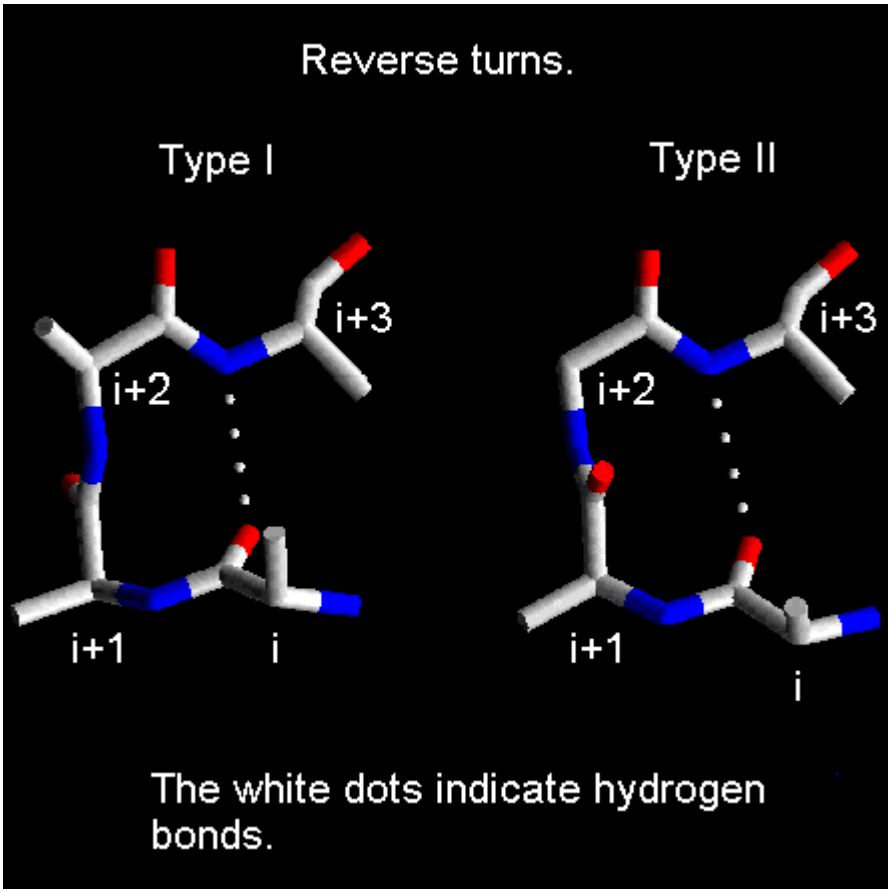


Parallel sheets are less twisted than anti-parallel and are always buried. In contrast, anti-parallel sheets can withstand greater distortions (twisting and β-bulges) and greater exposure to solvent. This implies that anti-parallel sheets are more stable than parallel ones which is consistent both with the hydrogen bond geometry and the fact that small parallel sheets rarely occur (see above).

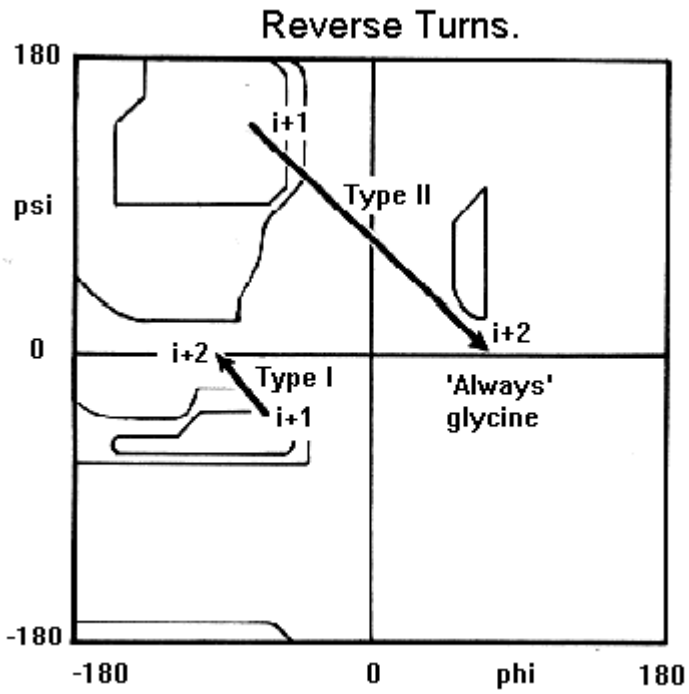
1.4 Reverse turns

A reverse turn is region of the polypeptide having a hydrogen bond from one main chain carbonyl oxygen to the main chain N-H group 3 residues along the chain (i.e. O_i to N_{i+3}). Helical regions are excluded from this definition and turns between β-strands form a special class of turn known as the β-hairpin (see later). Reverse turns are very abundant in globular proteins and generally occur at the surface of the molecule. It has been suggested that turn regions act as nucleation centres during protein folding.

Reverse turns are divided into classes based on the Φ and Ψ angles of the residues at positions i+1 and i+2. Types I and II shown in the figure below are the most common reverse turns, the essential difference between them being the orientation of the peptide bond between residues at (i+1) and (i+2).



The torsion angles for the residues (i+1) and (i+2) in the two types of turn lie in distinct regions of the Ramachandran plot.



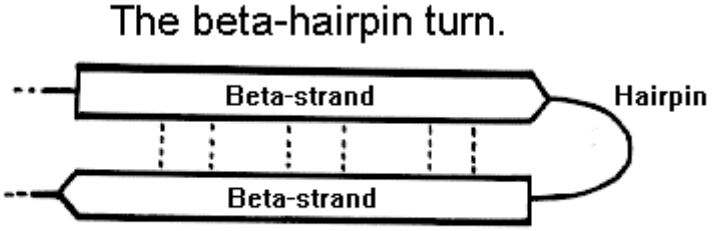
Note that the (i+2) residue of the type II turn lies in a region of the Ramachandran plot which can only be occupied by glycine. From the diagram of this turn it can be seen that were the (i+2) residue to have a side chain, there would be steric hindrance with the carbonyl oxygen of the preceding residue. Hence, the (i+2) residue of type II reverse turns is nearly always glycine.

2 Super-secondary structure

Secondary structure elements are observed to combine in specific geometric arrangements known as motifs or super-secondary structures. In this section we will look at motifs consisting of no more than three secondary structure elements. Larger motifs such as the Greek key will be examined in the sections on tertiary structure and protein folds.

2.1 β-hairpins

β-hairpins are one of the simplest super-secondary structures and are widespread in globular proteins.



The dashed lines indicate main chain hydrogen bonds.

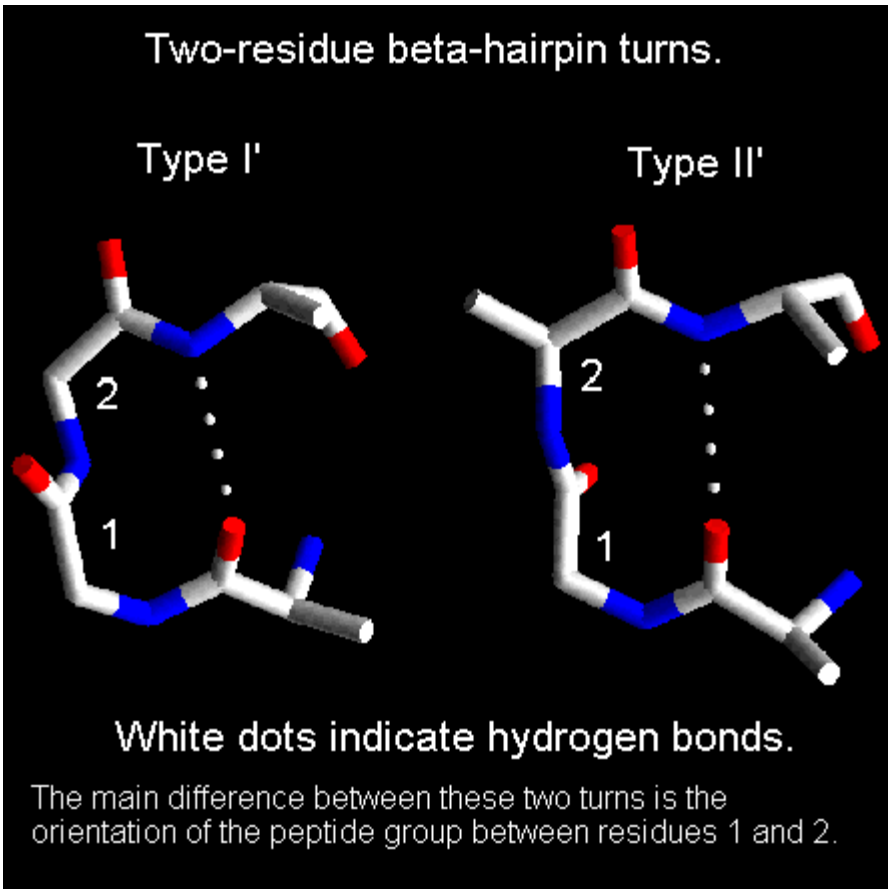
They occur as the short loop regions between anti-parallel hydrogen bonded β-strands. In general a reverse turn (or β-turn, as it they are sometimes called) is any region of a protein where there is a hydrogen bond involving the carbonyl of residue i and the NH group of residue i+3. An alternative definition states that the α-carbons of residues i and i+3 must be within 7.0 Å. The structures of reverse turns are outlined in section 1.4. In this section we will concentrate on those turns which occur between consecutive β-strands, known as β -hairpins. Sibanda and Thornton have devised a system for classifying β-hairpins which is based on two conventions for defining loop regions. In this section we will not go into such details as the objective is indicate the most commonly observed hairpin loop structures.

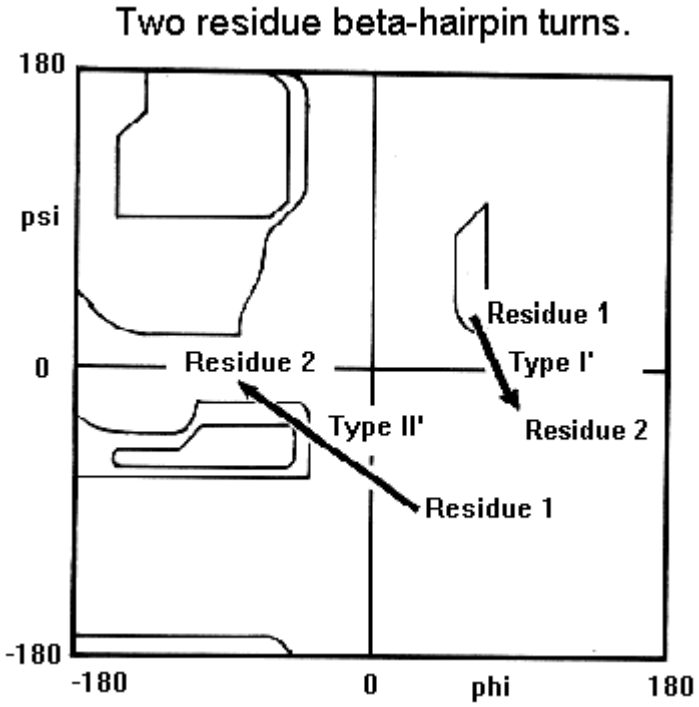
β-hairpin loops adopt specific conformations which depend on their lengths and sequences. Sibanda and Thornton have shown that 70% of β-hairpins are less than 7 residues in length with the two-residue turns forming the most noticeable component. These two-residue β-hairpins all adopt one of the classical reverse turn conformations with an obvious preference for types I' and II'. Type I 2-residue hairpins also occur but with lower abundance. This contrasts with reverse turns where types I and II tend to dominate. In β-hairpins the type I' turn has the correct twist to match the twist of the β-sheet and modelling studies indicated that if either type I or type II turns were to connect the anti-parallel β-strands, they would diverge within a short distance from the turn.

2.1.1 Two-residue β-hairpins

1. **Type I'**.The first residue in this turn adopts the left-handed α-helical conformation and therefore shows preference for glycine, asparagine or aspartate. These residues can adopt conformations with positive Φ angles due to the absence of a side chain with glycine and because of hydrogen bonds between the side chain and main chain in the case of asparagine or aspartate. The second residue of a type I' turn is nearly always glycine as the required Φ and Ψ angles are well outside the allowed regions of the Ramachandran plot for amino acids with side chains. Were another type of amino acid to occur here there would be steric hindrance between its side chain and the carbonyl oxygen of the preceding residue.

Type II'. The first residue of these turns has a conformation which can only be adopted by glycine (see below Ramachandran plot). The second residue shows a preference for polar amino acids such as serine and threonine.

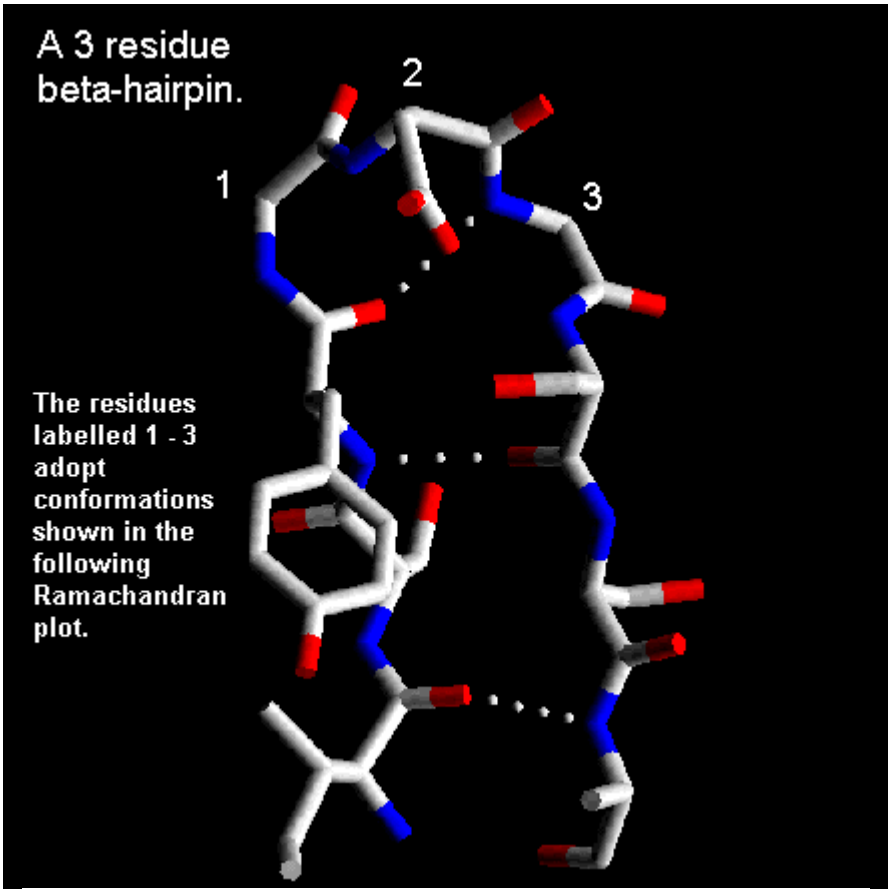


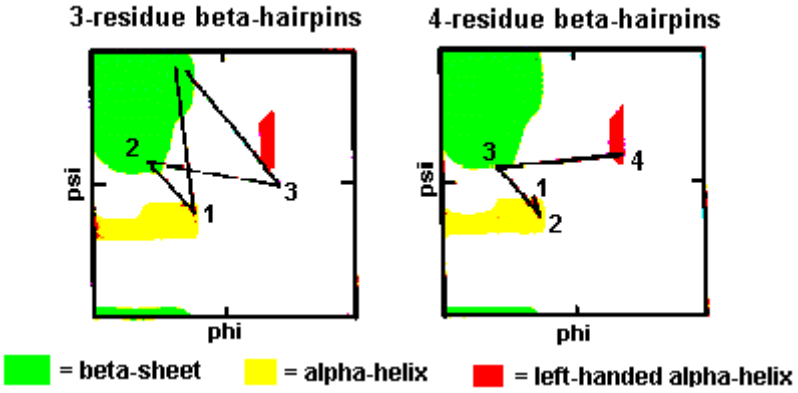


1. **Type I**. Both residues of these turns adopt α-helical conformations.

2.1.2 Three-residue β-hairpins

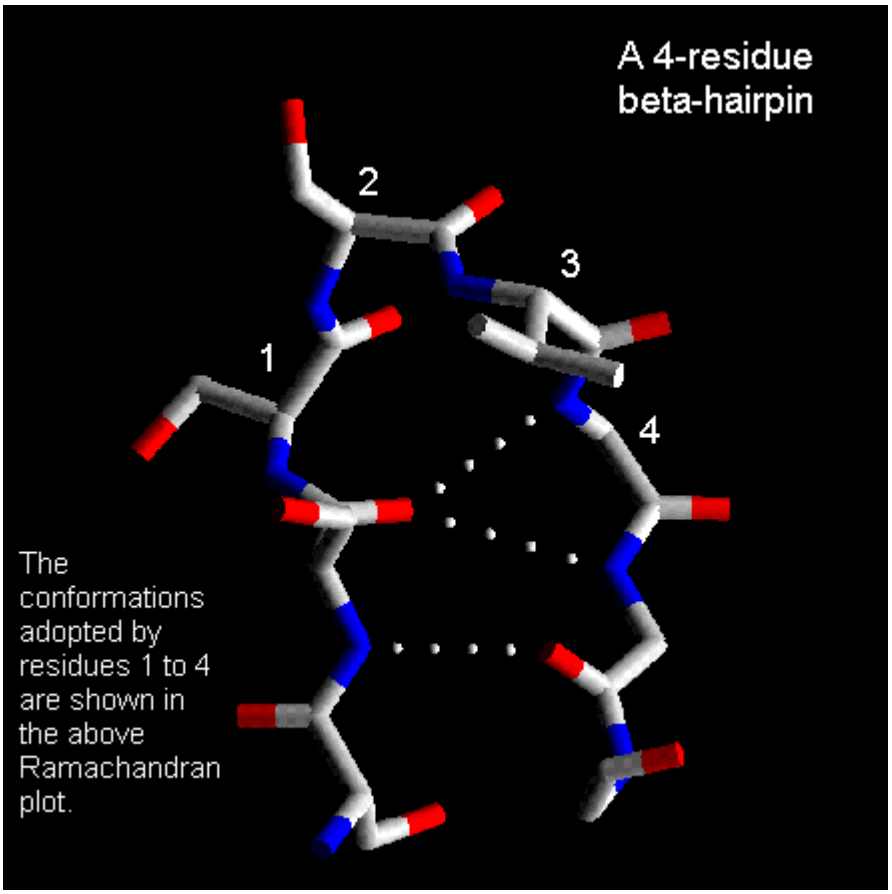
Normally the residues at the ends of the two β-strands only make one hydrogen bond as shown below. The intervening three residues have distinct conformational preferences as shown in the Ramachandran plot. The first residue adopts the right-handed α-helical conformation and the second amino acid lies in the bridging region between between α-helix and β-sheet. Glycine, asparagine or aspartate are frequently found at the last residue position as this adopts Φ and Ψ angles close to the left-handed helical conformation.





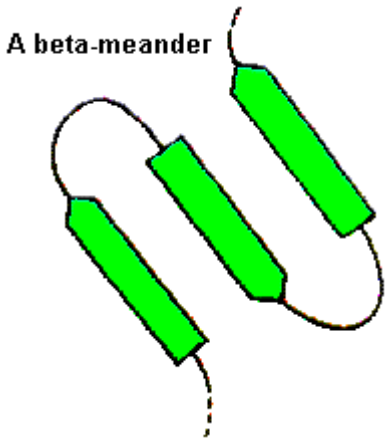
2.1.3 Four-residue β-hairpins

These are also quite common with the first two residues adopting the α-helical conformation. The third residue has Φ and Ψ angles which lie in the bridging region between between α-helix and β-sheet and the final residue adopts the left-handed α-helical conformation and is therefore usually glycine, aspartate or asparagine.



2.1.4 Longer loops

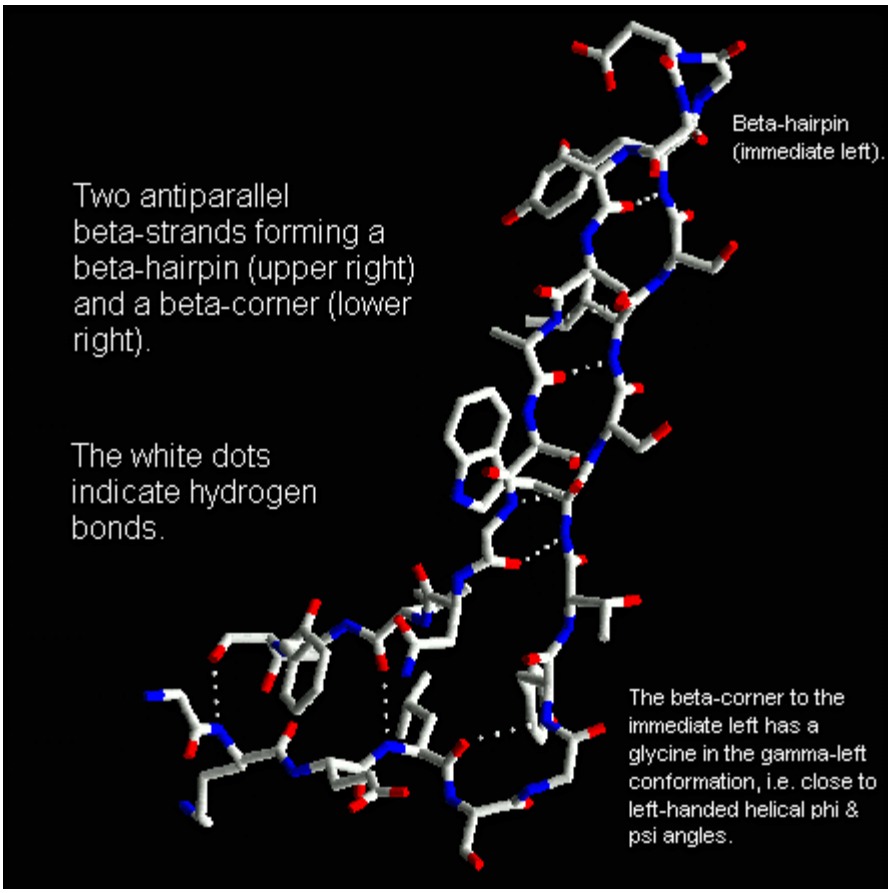
For these, a wide range of conformations is observed and the general term 'random coil' is sometimes used. Consecutive anti-parallel β-strands when linked by hairpins form a super-secondary structure known as the β-meander.



2.2 β-corners

β-strands have a slight right-handed twist such that when they pack side-by-side to form a β-sheet, the sheet has an overall left-handed curvature. Anti-parallel β-strands forming a β-hairpin can accommodate a 90 degree change in direction known as a β-corner. The strand on the inside of the bend often has a glycine at this position while the other strand can have a β-bulge. The latter involves a single residue in the right-handed α-helical conformation which breaks the hydrogen bonding pattern of the β-sheet. This residue can also be in the left-handed helical or bridging regions of the Ramachandran plot.

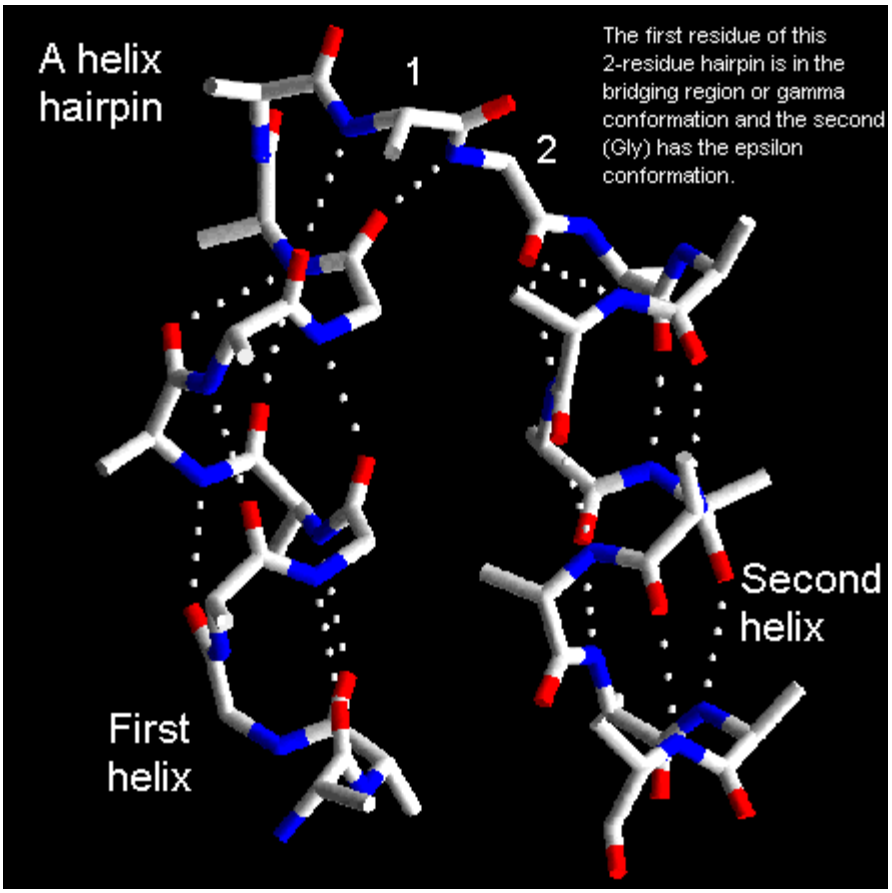
β-corners are observed to have a right-handed twist when viewed from the concave side.



2.3 Helix hairpins

A helix hairpin or αα-hairpin refers to the loop connecting two anti-parallel α-helical segments. Clearly, the longer the length of the loop the greater the number of possible conformations. However, for short connections there are a limited number of conformations and for the shortest loops of two or three residues, there is only one allowed conformation. Anti-parallel α-helices will interact generally by hydrophobic interactions between side chains at the interface. Therefore, hydrophobic amino acids have to be appropriately positioned in the amino acid sequence (one per turn of each helix) to generate a hydrophobic core. Efimov has analysed the conformations of αα-hairpins and some of his results are summarised below.

The shortest α-helical connections involve two residues which are oriented approximately perpendicular to the axes of the helices. Analysis of known structures reveals that the first of these two residues adopts Φ and Ψ αngles in the bridging or α-helical regions of the Ramachandran plot. The second residue is always glycine and is in a region of the Ramachandran plot with positive phi which is not available to other amino acids.

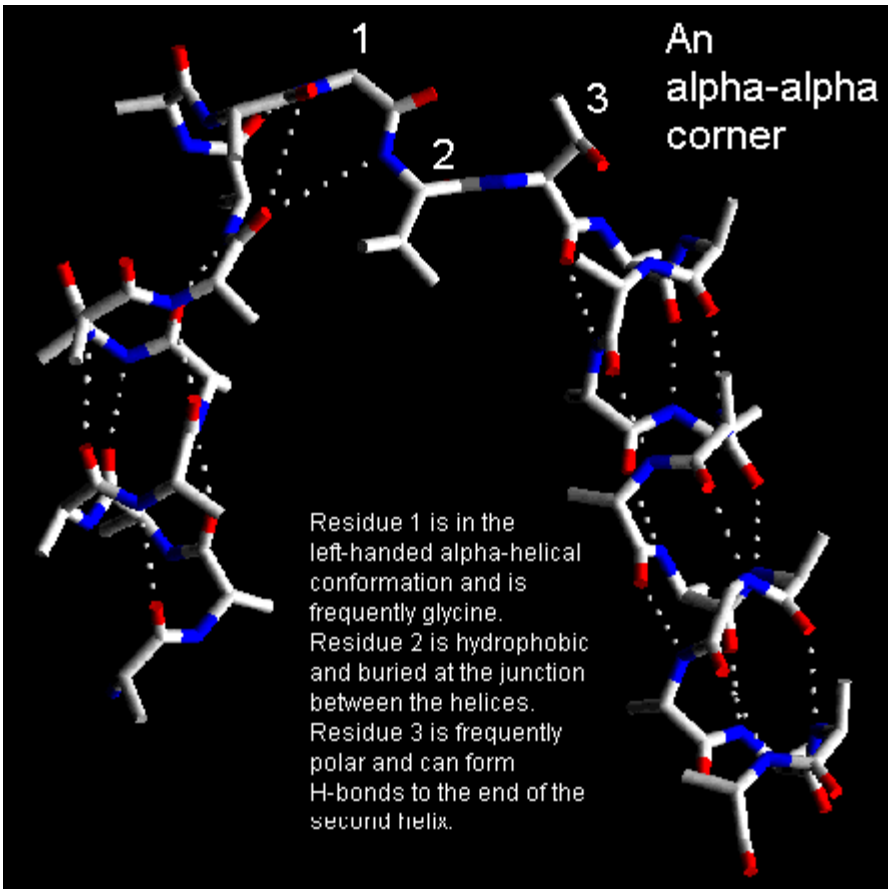


Three residue loops are also observed to have conformational preferences. The first residue occupies the bridging region of the Ramachandran plot, the second adopts the left-handed helical conformation and the last residue is in a β-strand conformation.

Four-residue loops adopt one of two possible conformations. One is similar to the three residue loop conformation described above except that there is an additional residue in the β-strand conformation at the fourth position. The other conformation involves the four residues adopting bridging, β, bridging, and β conformations, respectively.

2.4 The α-α corner

Short loop regions connecting helices which are roughly perpendicular to one another are refered to as αα-corners. Efimov has shown that the shortest αα-corner has its first residue in the left-handed α-helical conformation and the next two residues in β-strand conformations. This conformation can only be adopted when the two helices form a right-handed corner. Indeed, if the helices were linked to form a left-handed corner there would be steric hindrance. This may explain the scarcity of left-handed αα-corners in protein X-ray structures.



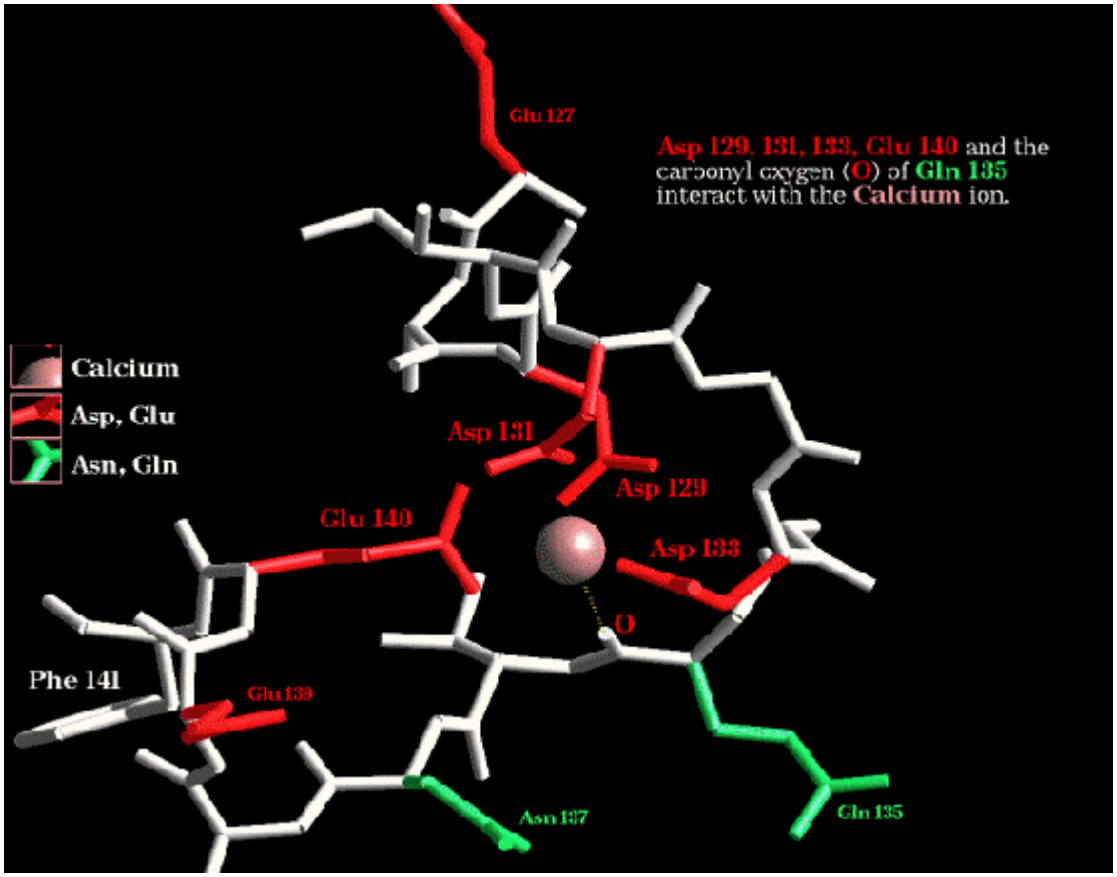
The C-terminal residue of the first helix, which is in the left-handed α-helical conformation, must have a short side chain to avoid steric hindrance and is observed commonly to be glycine. The first residue of the second helix, which is in the β-conformation, frequently has a small polar side chain such as serine or aspartate which can form hydrogen bonds with the free NH groups at the amino-terminal end of the second helix. The central residue of the αα-corner is almost always hydrophobic as it is buried and interacts with other non-polar side chains buried where the ends of the two helices contact each other.

2.5 Helix-turn-helix

The loop regions connecting α-helical segments can have important functions. For example, in parvalbumin there is helix-turn-helix motif which appears three times in the structure. Two of these motifs are involved in binding calcium by virtue of carboxyl side chains and main chain carbonyl groups. This motif has been called the EF hand as one is located between the E and F helices of parvalbumin. It now appears to be a ubiquitous calcium binding motif present in several other calcium-sensing proteins such as calmodulin and troponin C.

EF hands are made up from a loop of around 12 residues which has polar and hydrophobic amino acids at conserved positions. These are crucial for ligating the metal

ion and forming a stable hydrophobic core. Glycine is invariant at the sixth position in the loop for structural reasons. The calcium ion is octahedrally coordinated by carboxyl side chains, main chain groups and bound solvent.

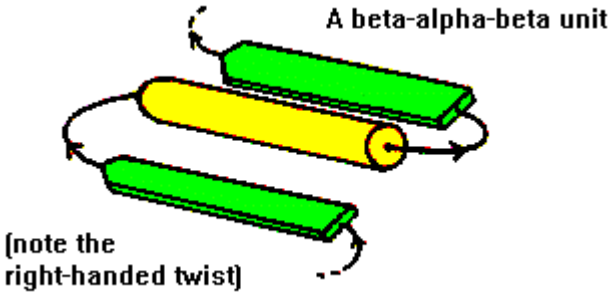


A different helix-loop-helix motif is also common to certain DNA binding proteins. This motif was first observed in prokaryotic DNA binding proteins such as the *cro* repressor from phage lambda. This protein is a homo-dimer with each subunit being 66 amino acids in length. Each subunit consists of an all-anti-parallel three stranded β-sheet with three helical segments inserted sequentially between the first and second β-strands. The two subunits of *cro* associate by virtue of the third β-strands which interact forming a six-stranded β-sheet in the centre of the molecule. Mutagenesis and biochemical work had indicated that residues in the second helix of each *cro* monomer interacted with DNA. Accordingly model building studies indicated that both these helices in the dimeric protein would fit into the major groove of B-DNA. These proteins recognise base sequences which are palindromic, i.e. possess an internal two-fold symmetry axis. The two recognition helices of the *cro* protein are also related by a two-fold axis passing through the central β-sheet region of the dimer. Therefore, the recognition helices of the *cro* dimer fit into the major groove of the DNA and interact with each identical half of the palindrome. Hence, the second helix of the helix-turn-helix motif has an important role in recognising the DNA while the remainder of the structure serves to keep the two helices in the correct relative position for fitting in the major groove of DNA. Many other

helix-turn-helix proteins with different folds exhibit essentially the same mode of binding to DNA.

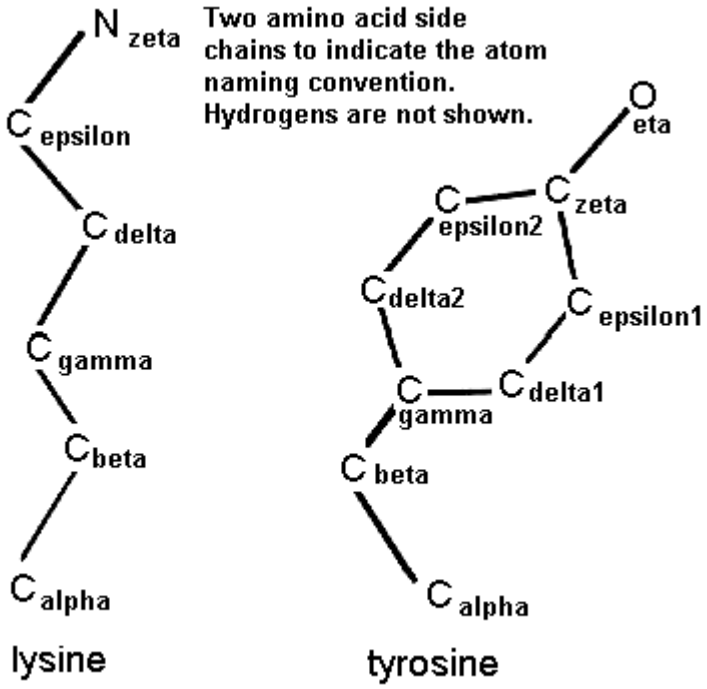
2.6 β-a-β motifs

Anti-parallel β-strands can be linked by short lengths of polypeptide forming β-hairpin structures. In contrast, parallel β-strands are connected by longer regions of chain which cross the β-sheet and frequently contain α-helical segments. This motif is called the β-a-β motif and is found in most proteins that have a parallel β-sheet. The loop regions linking the strands to the helical segments can vary greatly in length. The helix axis is roughly parallel with the β-strands and all three elements of secondary structure interact forming a hydrophobic core. In certain proteins the loop linking the carboxy terminal end of the first β-strand to the amino terminal end of the helix is involved in binding of ligands or substrates. The β-a-β motif almost always has a right-handed fold as demonstrated in the figure.

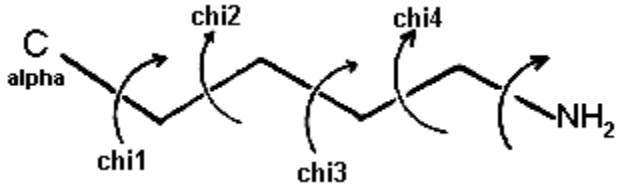


3 Side Chain Conformation.

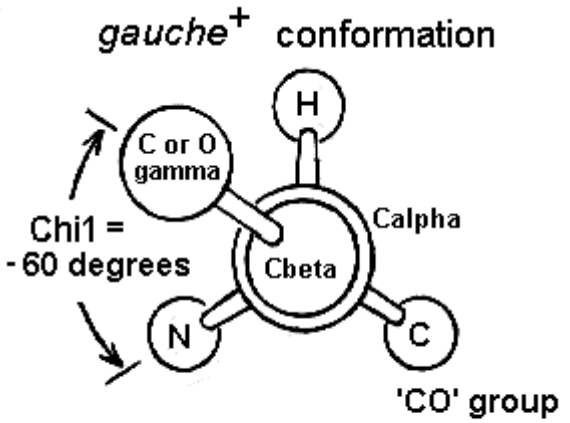
The side chain atoms of amino acids are named in the Greek alphabet according to this scheme.

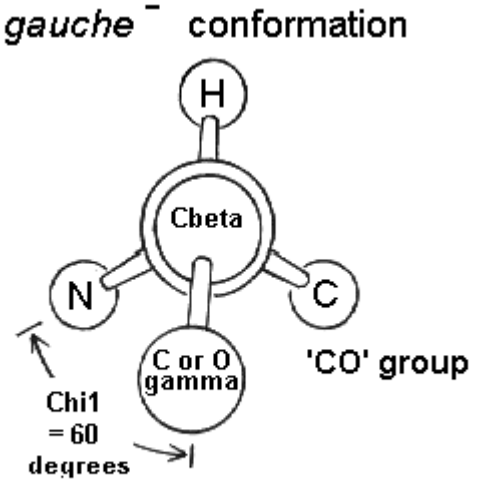


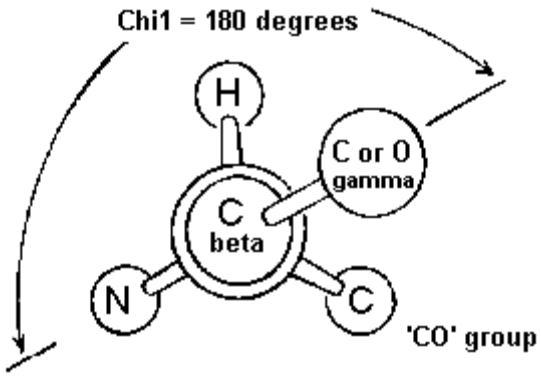
The side chain torsion angles are named χ_1(chi1), χ_2(chi2), χ_3 (chi3), *etc*., as shown below for lysine.



The χ_1 angle is subject to certain restrictions which arise from steric hindrance between the γ side chain atom(s) and the main chain. The different conformations of the side chain as a function of χ_1 are referred to as *gauche*(+), *trans* and *gauche*(-). These are indicated in the diagrams below in which the amino acid is viewed along the Cβ-Cα bond.







trans conformation

The most abundant conformation is *gauche*(+) in which the γ side chain atom is opposite to the residue's main chain carbonyl group when viewed along the Cβ-Cα bond.

The second most abundant conformation is *trans* in which the side chain γ atom is opposite the main chain nitrogen.

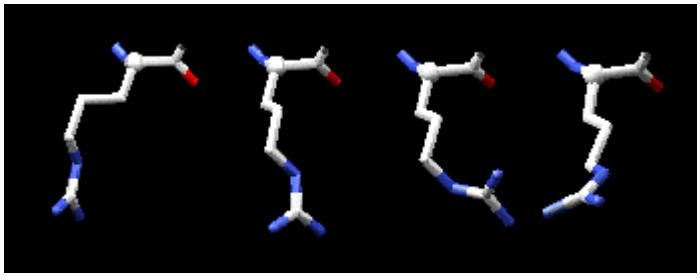
The least abundant conformation is *gauche*(-) which occurs when the side chain is opposite the hydrogen substituent on the Cα atom. This conformation is unstable because the γ atom is in close contact with the main chain CO and NH groups. The *gauche*(-) conformation is occasionally adopted by serine or threonine residues in a helix where the steric hindrance is offset by a hydrogen bond between the γ oxygen atom and the main chain.

With most amino acids the *gauche*(+) and *trans* conformations are adopted with similar abundances although the *gauche*(+) conformation tends to dominate.

Aliphatic amino acids which are bifurcated at Cβ, ie valine and isoleucine, do not adopt the *trans* conformation very often as this involves one of the Cγ atoms being in the unfavourable *gauche*(-) 'position'.

In general, side chains tend to adopt the same three torsion angles (+/-60 and 180 degrees) about χ_2 since these correspond to staggered conformations. However, for residues with an sp2 hydridised γ atom such as phenylalanine, tyrosine, etc., χ_2 rarely equals 180 degrees because this would involve an eclipsed conformation. For these side chains the χ_2 angle is usually close to +/-90 degrees as this minimises close contacts. For residues such as aspartate and asparagine the χ_2 angles are strongly influenced by the hydrogen bonding capacity of the side chain and its environment. Consequently, these residues adopt a wide range of χ_2 angles.

Here are some conformations that can be adopted by Arginines:



4 Tertiary Protein Structure and folds

4.1 Introduction

Chapters 1 and 2 introduced α-helices and β-sheets (*Secondary Struture*), and some common "motifs" composed of 2 or 3 of these elements (*Super-secondary Struture*). Tertiary structure describes the folding of the polypeptide chain to assemble the different secondary structure elements in a particular arrangement. As helices and sheets are units of secondary structure, so the domain is the unit of tertiary structure. In multi-domain proteins, tertiary structure includes the arrangement of domains relative to each other as well as that of the chain within each domain.

There is a blurred distinction between "super-secondary structure" and "tertiary structure". The introduction of the term "super-secondary structure" was necessary when it became clear that certain arrangements of two or three secondary structures are present in many different protein structures, even with completely different sequences.

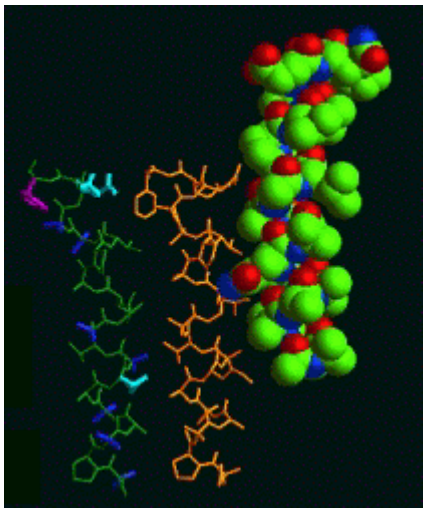
Note that some proteins do not consist of an assembly of these super-secondary motifs. For example, proteins of the globin family consist of eight α-helices in contact, but the helices do not pack against other helices which are adjacent in the sequence, with the exception of the final two, which form an anti-parallel helix-turn-helix motif.

Although the term "motif" is often used to describe super-secondary structures (e.g. Branden and Tooze, 1991), it may also be used to describe a consensus sequence of amino acids identified in a number of different proteins, rather than a repeated three-dimensional conformation. Such a consensus in primary structure generally implies a similarity in tertiary structure. But bear in mind that there are very many protein sequences of which the three dimensional structures are not known for certain, so that the term "motif" strictly applies to primary rather than supersecondary or tertiary structure in these cases.

4.2 All-α topologies

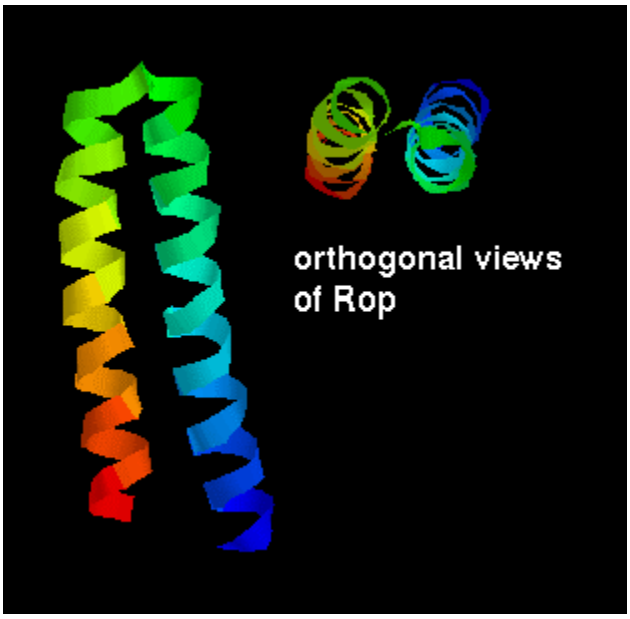
4.2.1 The lone helix

There are a number of examples of small proteins (or peptides) which consist of little more than a single helix. A striking example is alamethicin, a transmembrane voltage gated ion channel, acting as a peptide antibiotic.



4.2.2 The helix-turn-helix motif

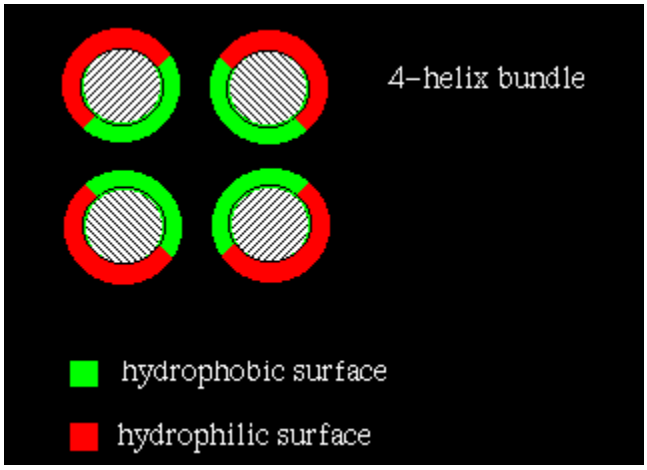
The simplest packing arrangement of a domain of two helices is for them to lie antiparallel, connected by a short loop. This constitutes the structure of the small (63 residue) RNA-binding protein Rop , which is found in certain plasmids (small circular molecules of double-stranded DNA occurring in bacteria and yeast) and involved in their replication. There is a slight twist in the arrangement as shown.



4.2.3 The four-helix bundle

The four-helix bundle is found in a number of different proteins. In many cases the helices part of a single polypeptide chain, connected to each other by three loops. However, the Rop molecule is in fact a dimer of two of the two-helix units shown above.

In four-helix-bundle proteins the interfaces between the helices consist mostly of hydrophobic residues while polar side chains on the exposed surfaces interact with the aqueous environment, as indicated below:

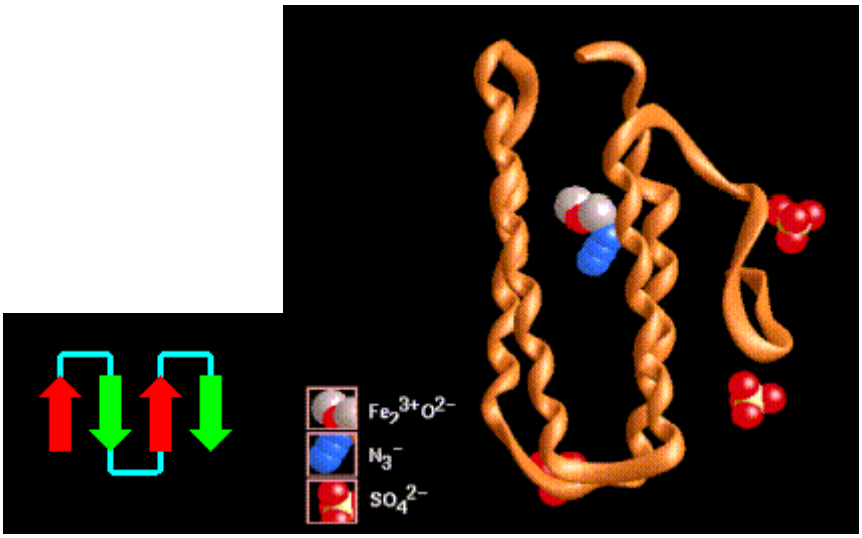


Compare this with the arrangement of residues that would be expected in a membrane-spanning helical domain. The central helices of the photosynthetic reaction centre in fact are arranged similar to the four- helix bundle.

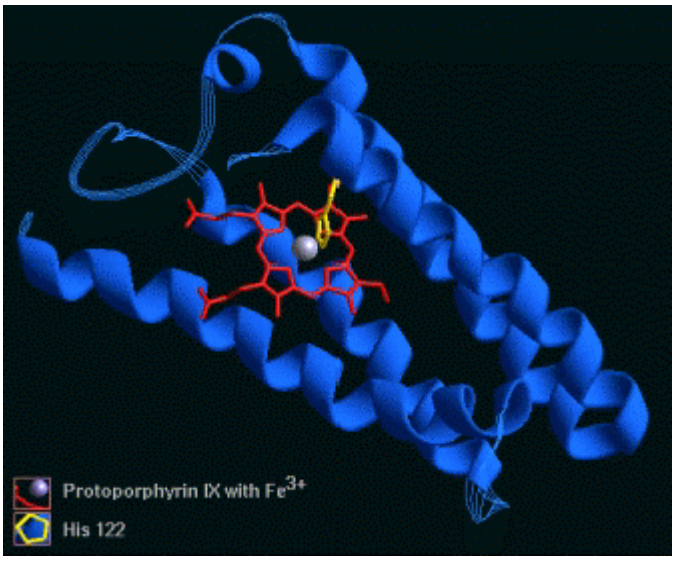
Other examples exhibit a much more open packing arrangement, as in the steroid-binding proteins uteroglobin, and Clara cell 17kDa protein.

4.2.3.1 Topologies

The four helices may be arranged in a simple up-and-down topology, as indicated. A good example is myohemeyrthrin.

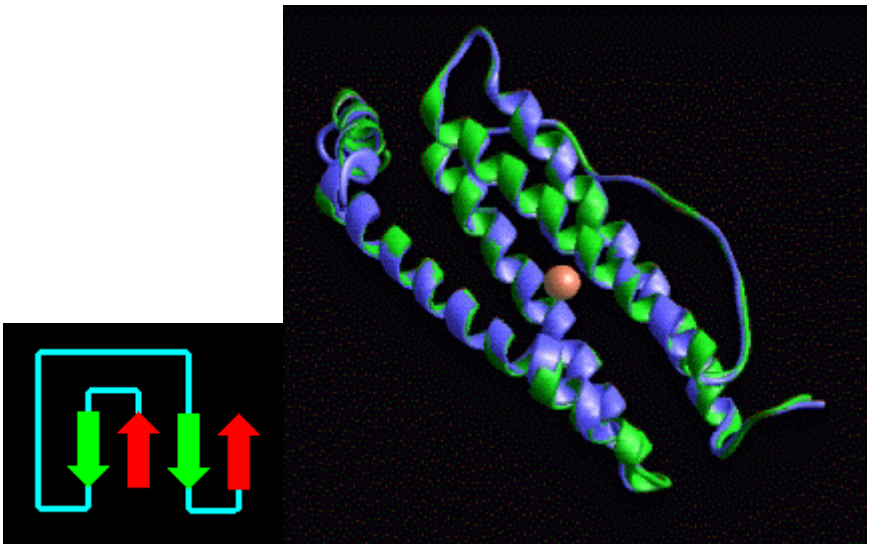


Others are cytochrome c'



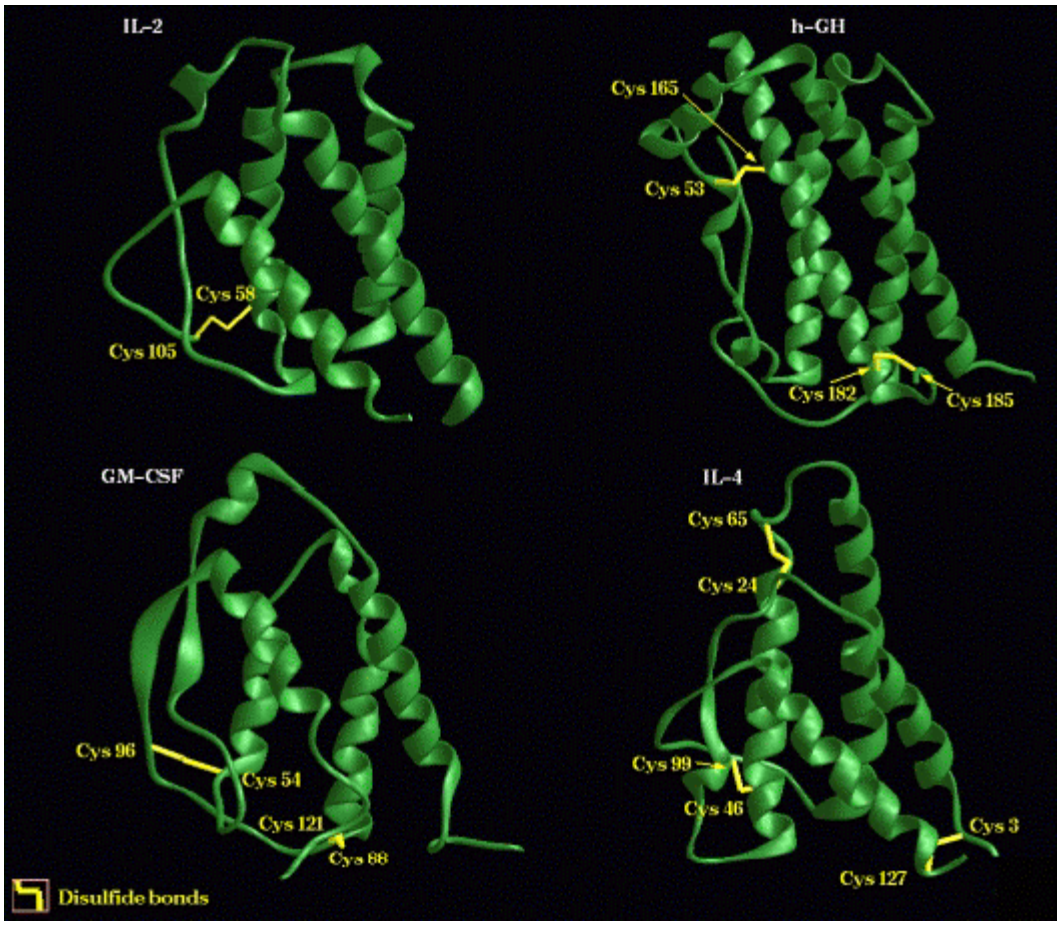
and cytochrome b-562.

A more complex arrangement, such as ferritin is possible:



4.2.3.2 Cytokines

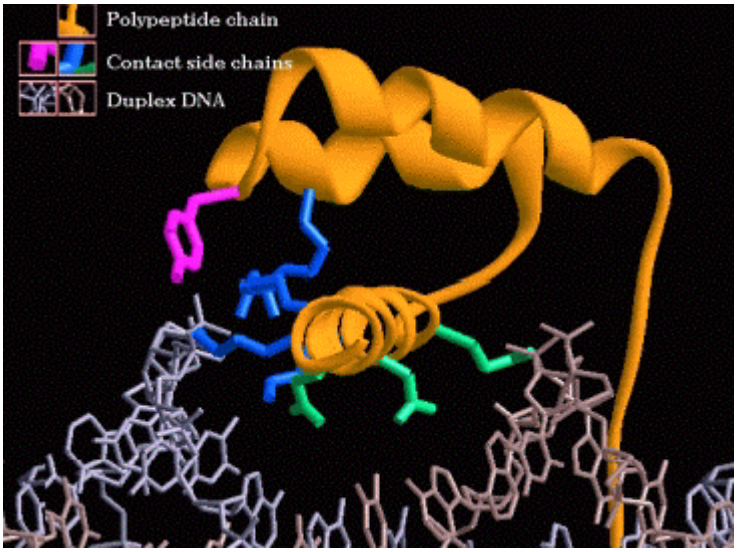
A number of **cytokines** consist of four α-helices in a bundle. Here is a diagram of Interleukin-2, human Growth Hormone, Granulocyte-macrophage colony-stimulating factor (GM-CSF) and Interleukin-4.



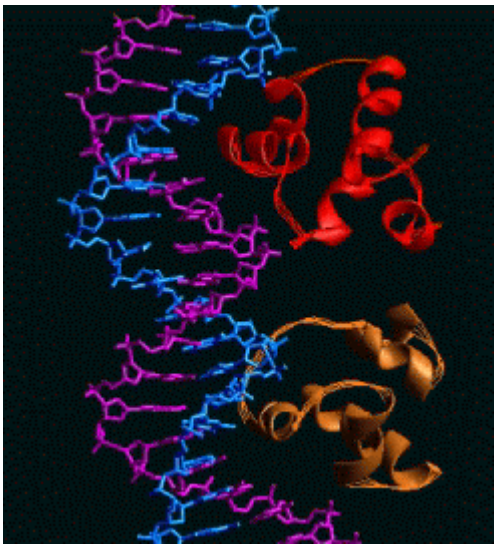
4.2.3.3 α domains which bind DNA

Transcription factors are proteins which bind to **control regions** of DNA. These regions are "upstream" of the structural gene (the sequence which actually codes for a protein) whose transcription they regulate. Transcription factors have a DNA-binding domain and a domain that activates transcription.

The RNA-binding two-helix protein Rop has already been mentioned. A three-helix bundle forms the basis of a DNA-binding domain which occurs in a number of proteins- for example **homeodomain proteins**. Examine the crystal structure of engrailed homeodomain binding to DNA.



And here is the structure of the **cro repressor** from phage 434.



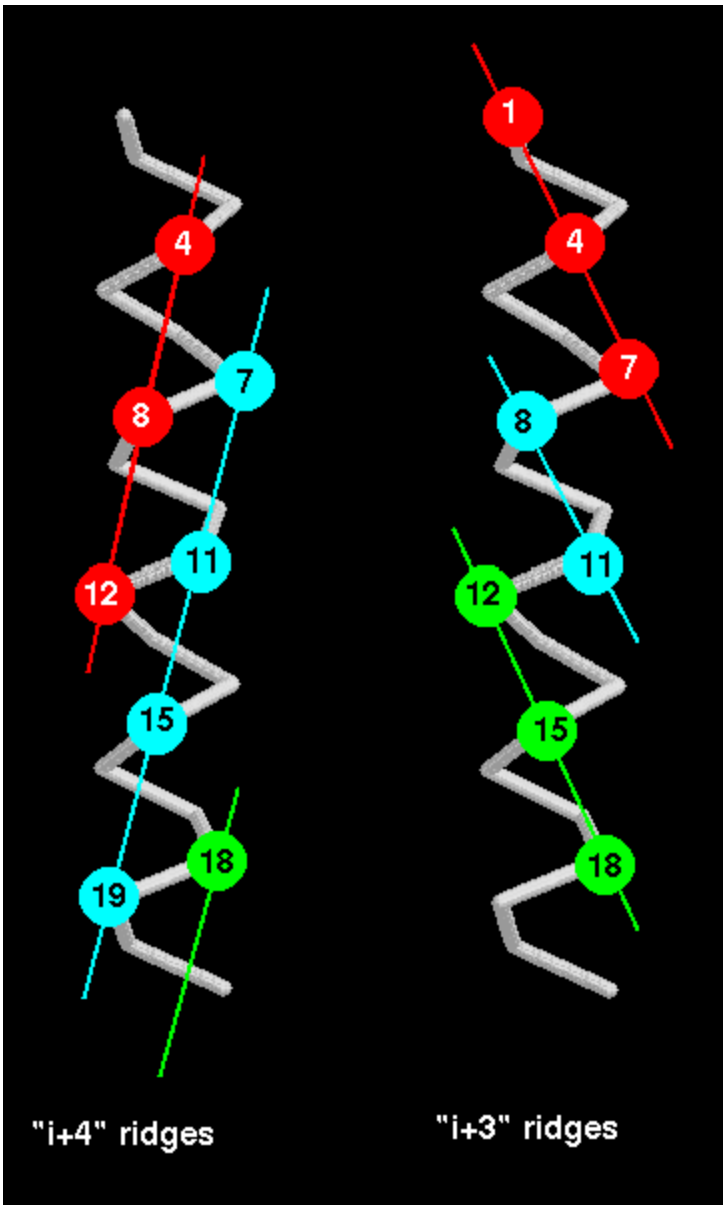
4.2.3.4 Globins

The globin fold usually consists of eight α–helices. The two helices at the end of the chain are anti-parallel, forming a helix-turn-helix motif, but the remainder of the fold does not include any characterised super-secondary structures. These helices pack against each other with larger angles, around 50 °, between them than occurs between antiparallel helices (approximately 20°). See the section below on helix-helix packing. Jane Richardson (1981) describes the globin fold as a "Greek key helix bundle", due to the topological similarity with the Greek key arrangement of anti-parallel β-sheets (see section 4.3 on all β topologies).

In all, fifty-six categories of "mostly α" folds are listed in the Structural Classification of Proteins database. A number of the entries have links to appropriate diagrams by Manuel Peitsch.

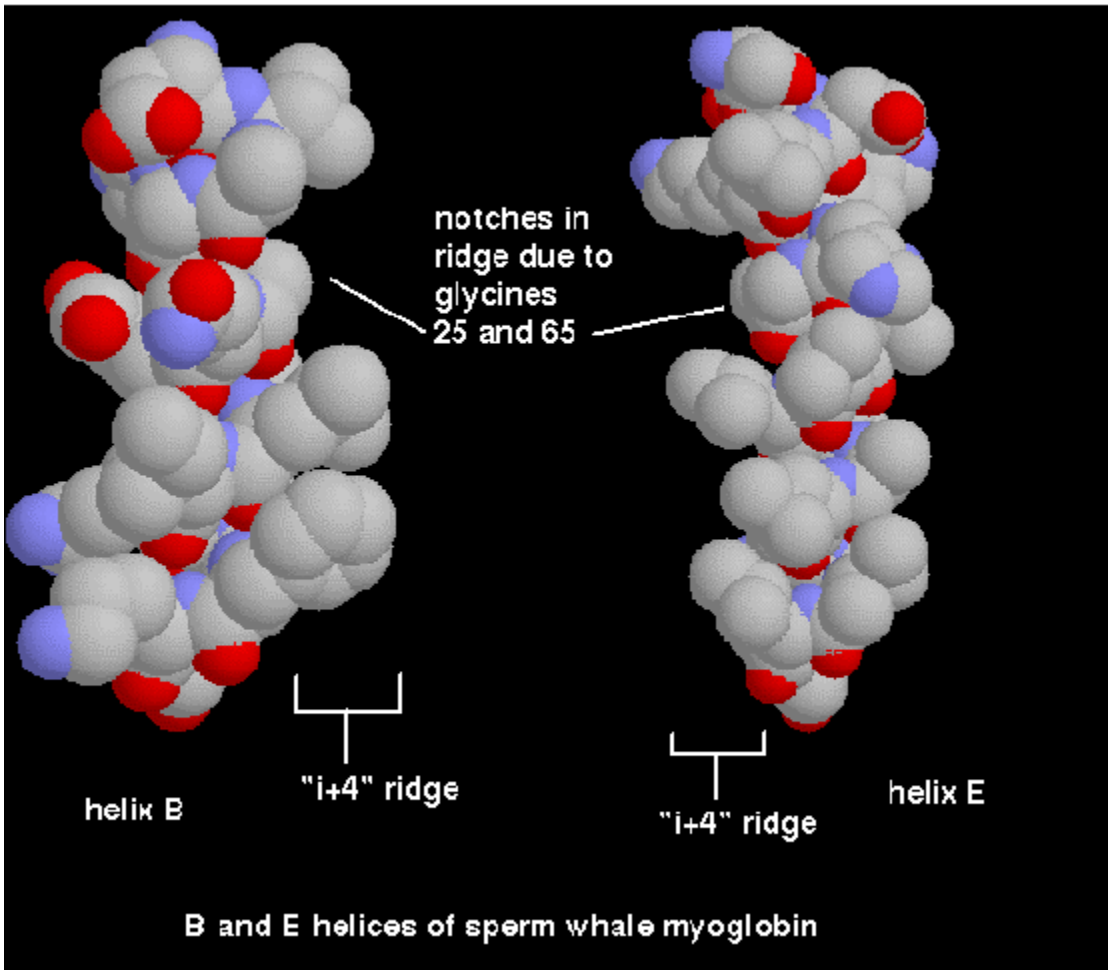
4.2.4 Helix-helix packing

When alpha-helices pack against each other, the side-chains in their interface are buried. The two interface areas should have complementary surfaces. The surface of an α-helix can be thought of as consisting of grooves and ridges, like a screw thread: for instance, the side chains of every 4th residue form a ridge (because there are 3.6 residues per turn). The direction of this ridge is 26° from the direction of the helix axis. Therefore if 2 helices pack such that such a ridge from each fits into the other's groove, the expected angle between the two is 52°. In fact, in the distribution of this angle between packed alpha-helices, there is a sharp peak at 50°. Besides the type of ridge described, ridges can be formed by other stacking patterns of residues, such as every 3rd residue, or indeed every residue. Which ridges are used for packing depends on the size and conformations of the side chains at these relative positions. The "i+4" ridge is believed to be the most common because residues at every 4th position have side-chains which are more closely aligned than in "i+3" or "i+1" ridges as indicated below.

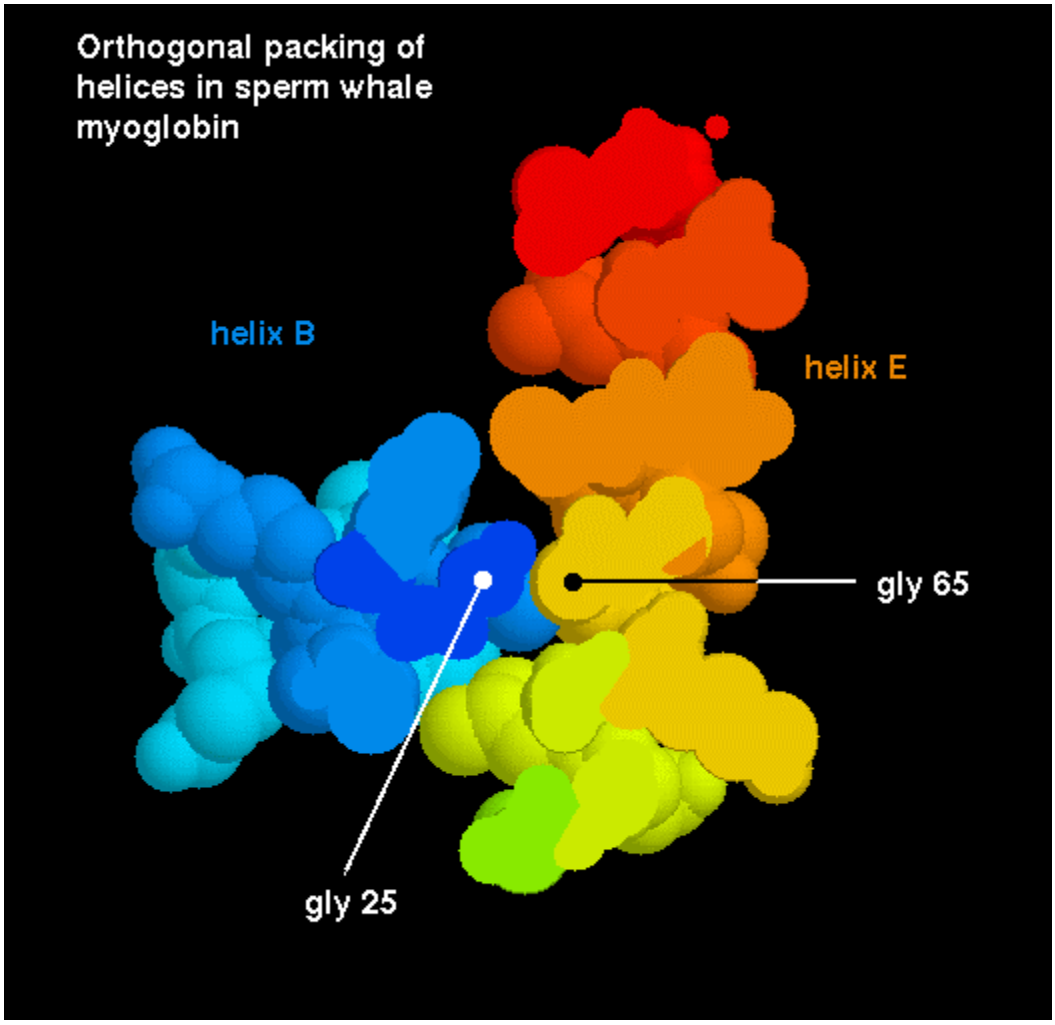


Two other types of packing do occur, however : between an "i+4" ridge and an "i+3" ridge (there is an angle of 23° between the 2 helix axes) and between an "i+4" and an "i+1" ridge (the helices are 105° apart). The "ridges and grooves" model does not describe all the helix-helix packings, as there are examples with unusual inter-axial angles. For instance in the globin fold a pair of helices (B and E) pack such that their ridges cross each other, by means of a notch formed at a pair of glycine residues.

Below is a diagrams of the notch in the ridges of helices B and H:



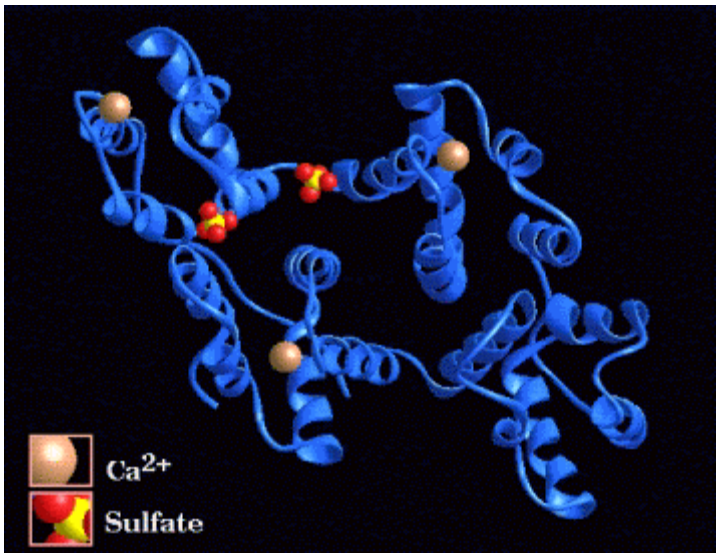
and here is one for a slice through a space-filling model of the two helices packing against each other.



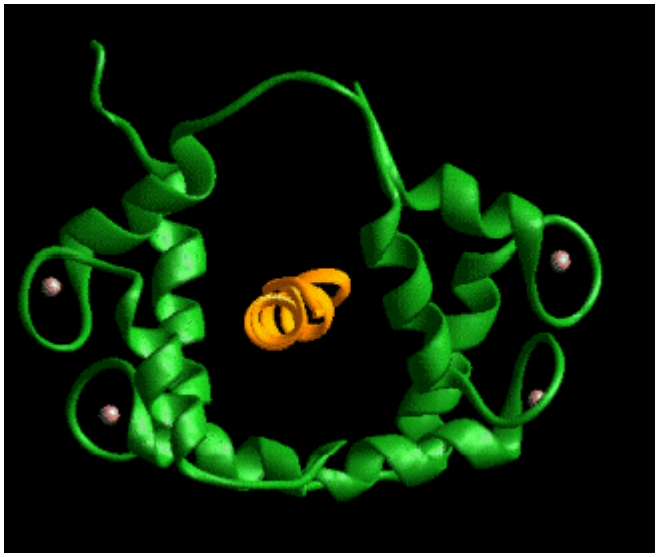
The inter-axial distance between packed helices varies from 6.8-12.0Å, the mean being 9.4 Å; the mean inter-penetration of atoms at the interface is 2.3Å. Therefore it is mainly side chains which make the contacts between the helices.

4.2.5 Other distinctive all-α proteins include :

- Delta-Crystallin
- Annexin V



- Glutathione S-transferase
- Calmodulin- and Parvalbumin-like calcium-binding proteins



4.3 All-β topologies

Protein folds which consist of almost entirely β sheets exhibit a completely or mostly antiparallel arrangement. Many of these anti-parallel domains consist of two sheets packed against each other, with hydrophobic side chains forming the interface. Bearing in mind that side chains of a β-strand point alternately to opposite sides of a sheet, this means that such structures will tend to have a sequence of alternating hydrophobic and polar residues.

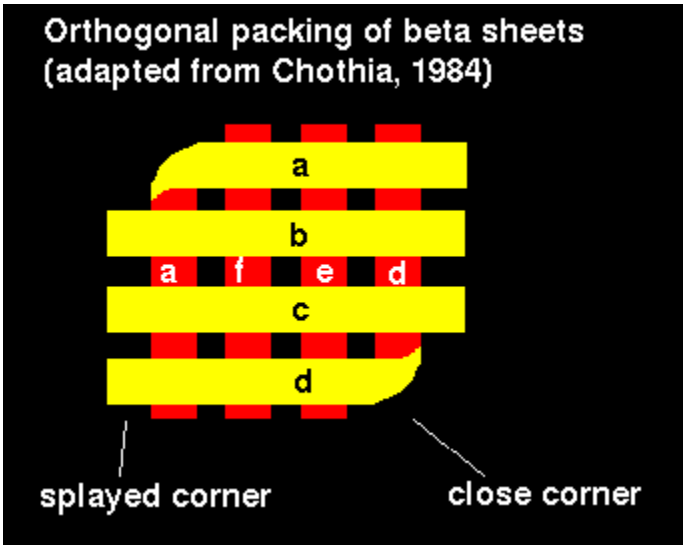
4.3.1 β sandwiches and β barrels

The immunoglobulin fold the strands form two sheets packed against each other, forming a "β **sandwich"**.

4.3.1.1 Aligned and orthogonal β sandwiches

In the immunoglobulin and fibronectin type-3 folds, the two sheets are approximately **aligned**. In fact the mean angle between the 2 sheets is approximately 30° (designated -30° because the uppermost sheet is rotated clockwise with respect to the lower). The two sheets are usually independent in that the linking residues between them are not in β sheet conformation. The angle between the sheets is determined by their right-handed twist. The observed angle varies between -20° and -50°; this is due to variation in the twist. Also side-chains are not always ideally aligned at the interface.

Orthogonal β sheet packings consist of β sheets folded on themselves; the two sheets make an angle of -90°. The strands at one corner or 2 diagonally opposite corners go uninterrupted from one layer to the other. Local coiling at the corner or a β bulge facilitates the right-angled bend. These bends are right-handed, due to permitted Φ and Ψ angles. The figure below illustrates this model.



Only along one diagonal do the two sheets make contact. Large side-chains in loops usually fill the spaces between the splayed corners.

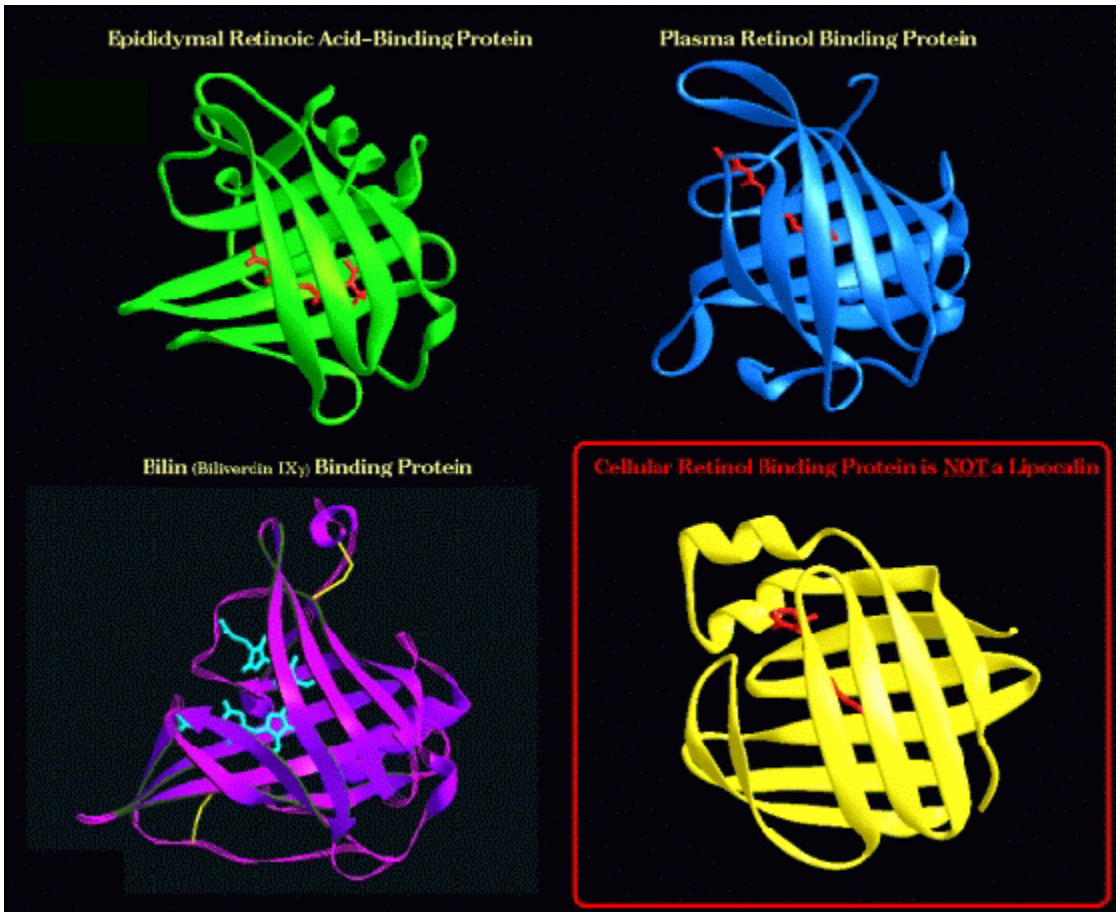
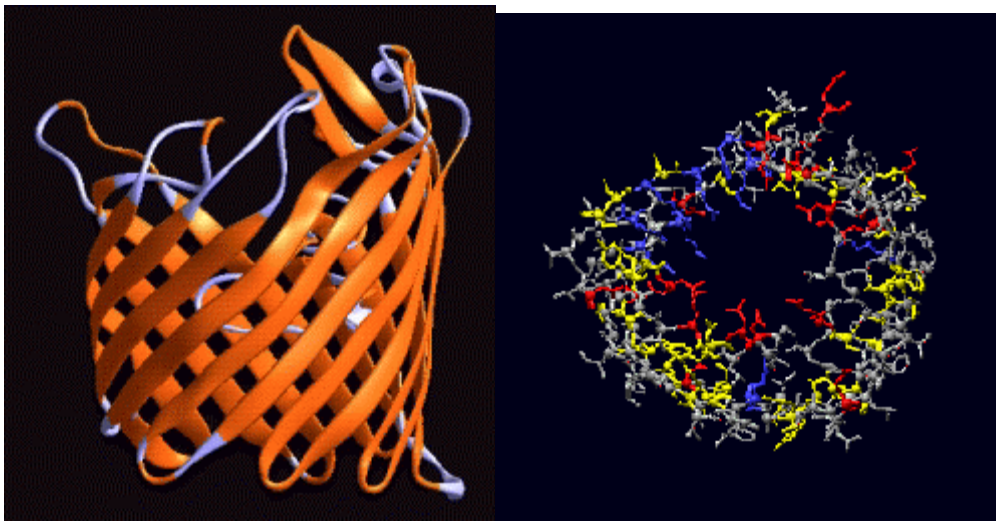


Diagram of this β-sheet arrangement in the Lipocalin family, which bind small molecules between the sheets of the sandwich.

4.3.1.2 β barrels

Some antiparallel β-sheet domains are better described as β-**barrels** rather than β-sandwiches, for example streptavadin and **porin**. Note that some structures are intermediate between the extreme barrel and sandwich arrangements.



4.3.2 Up-and-down antiparallel β sheets

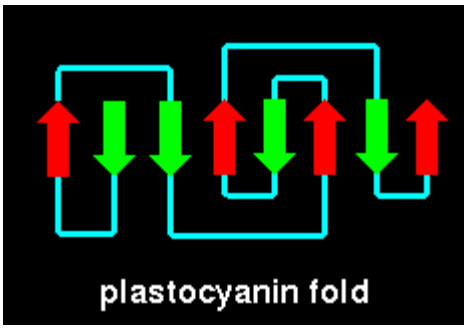
The simplest topology for an antiparallel β-sheet involves loops connecting adjacent strands.

4.3.2.1 The Greek Key topology

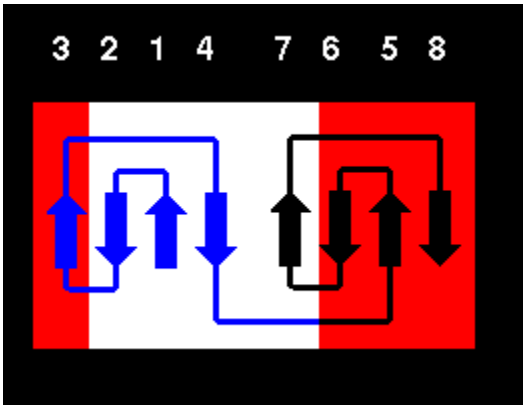
The **Greek Key** topology, named after a pattern that was common on Greek pottery, is shown below. Three up-and-down β-strands connected by hairpins are followed by a longer connection to the fourth strand, which lies adjacent to the first.

Folds including the Greek key topology have been found to have 5-13 strands. An example is given below.

1. **Plastocyanin**. Notice that this has a mixed sheet- there are two parallel pairs of strands.

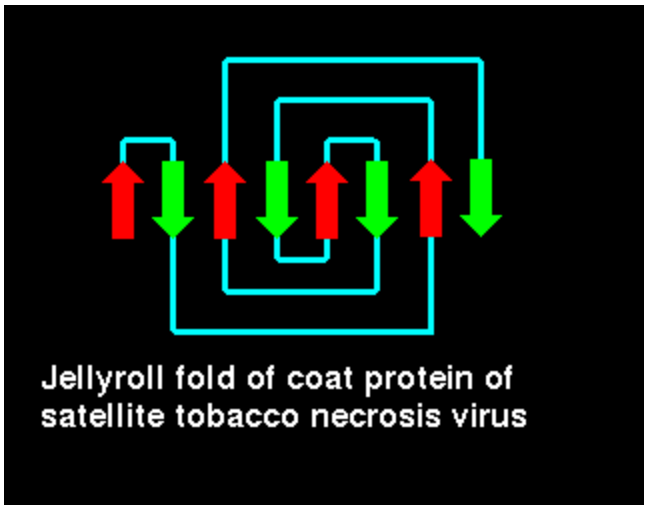


1. **Gamma-crystallin** Gamma-crystallin has two domains each of which is an eight-stranded B-barrel-type structure composed of two Greek keys. In fact, the structure is more accurately described as consisting of two B-sheets, one consisting of strands 2,1,4,7 (white) and the other of strands 6,5,8,3 (red) as indicated in the diagram. Sequence homology has been found between the two Greek key motifs within each domain, and also between the two domains themselves. The latter homology is higher than the former; this implies that the structure evolved from a single Greek key fold by means of a gene duplication to produce a domain of two Greek keys, followed by a second duplication resulting in two similar domains. This is supported by the fact that in some crystallins each Greek key motif is coded by a different exon, with introns between them.



4.3.2.2 The Jellyroll Topology

Richardson(1981) describes the **jellyroll** fold as being formed by the addition of an extra "swirl" to a Greek key:



Click here for a diagram illustrating this fold in the coat protein of satellite tobacco necrosis virus.

4.3.3 β-propellors

One molecule is composed of four of these subunits. Each is a **superbarrel** of six four-stranded antiparallel sheets. The whole structure has a basically up-down topology and is called a β-**propellor**.

4.3.4 β-trefoils

This fold has an approximately 3-fold axis of symmetry.

4.3.5 β-Helix

This dramatically unusual fold was discovered quite recently. The β-strands wind round the structure describing a helical topology.

4.4 α/β topologies

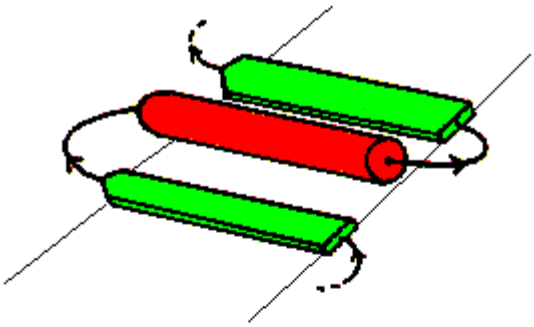
The most regular and common domain structures consist of repeating β-α-β supersecondary units, such that the outer layer of the structure is composed of α helices packing against a central core of parallel β-sheets. These folds are called α/β , or **wound** α β.

Many enzymes, including all those involved in *glycolysis* , are α/β structures. Most α/β proteins are cytosolic.

The β-α-β is always right-handed. In α/β structures, there is a repetition of this arrangement, giving a β-α-β-α.....etc sequence. The β strands are parallel and hydrogen

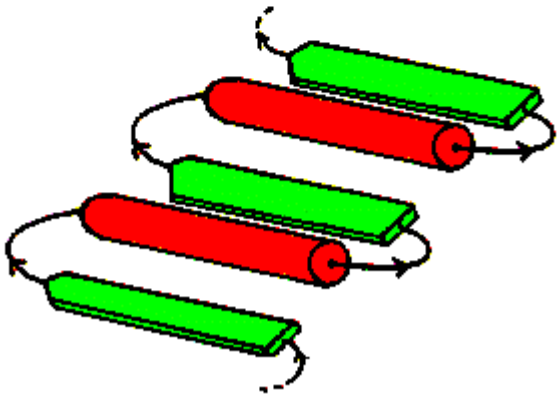
bonded to each other, while the α helices are all parallel to each other, and are anitparallel to the strands. Thus the helices form a layer packing against the sheet.

The β-α-β-α-β subunit, often present in nucleotide-binding proteins, is named the **Rossman Fold**, after Michael Rossman (Rao and Rossman,1973).



The right–handed beta–alpha–beta unit. The helix lies above the plane of the strands.

The Rossman fold

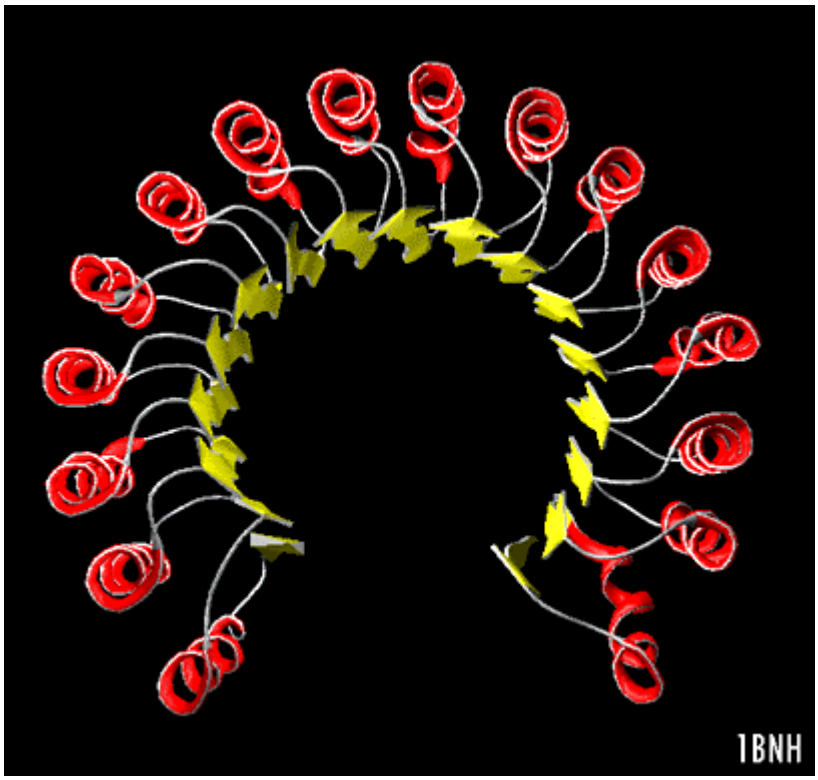


Richardson (1981) names the α/β structures "parallel α/β domains", to denote the fact that each of the 2 secondary structures forms a parallel arrangement. Note that there is no obvious reason why one would not expect to find "parallel all α" (α-α-α subunit) folds, or "parallel all β" (β-β-β) folds in equally large numbers, but these do not occur. However, the marked tendency for helices to pack aligned with sheets has been explained by the "complementary twist" model (Chothia *et al.* , 1977). The right-handed twist of β sheets and the right-handed twist of the row of every 4th residue of the helices (the "i+4" ridges"- see section 4.2.4 on helix-helix packing) mean that the two have complementary

surfaces when aligned. This model is supported by the observation that approximately 90% of the helix residues which interface with a sheet are indeed a multiple of 4 residues apart. Helices packing side by side on a sheet would have helices rotated with respect to each other, due to the sheet twist; the observed interhelical angle is in agreement with this model in 80% of cases. In the other cases the helices are splayed from the sheet, with only one end in contact.

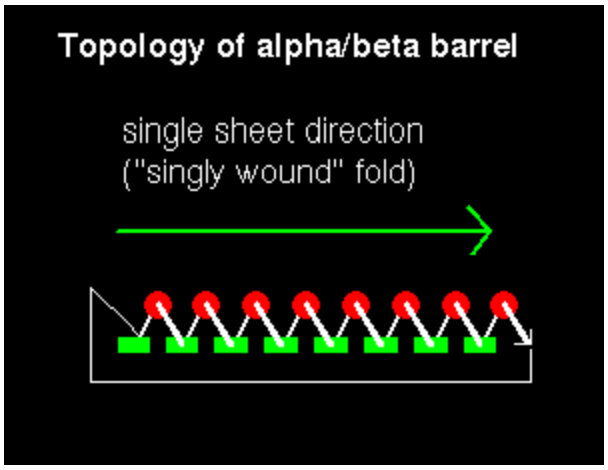
4.4.1 α/β **horseshoe**

The structure of the remarkable placental ribonuclease inhibitor (Kobe, B. & Diesenhofer, J. (1993) *Nature* **V.366**, 751) takes the concept of the repeating α/β unit to extremes. It is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. Look at the image below and you will see the 17-stranded parallel β sheet curved into an open horseshoe shape, with 16 α-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'.

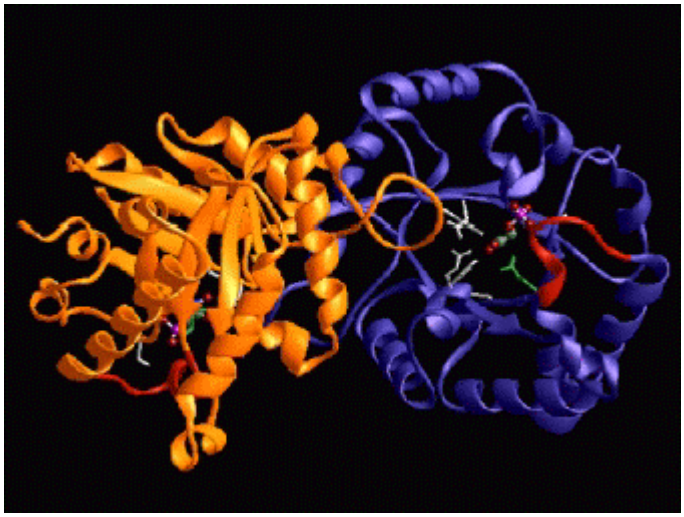


4.4.2 α/β **barrels**

Consider a sequence of eight β-α motifs:

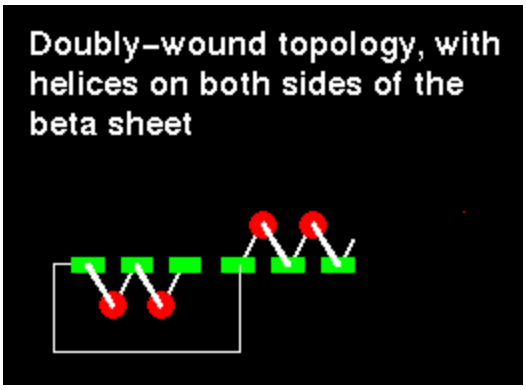


If the first strand hydrogen bonds to the last, then the structure closes on itself forming a barrel-like structure. This is shown in the picture of triose phosphate isomerase.

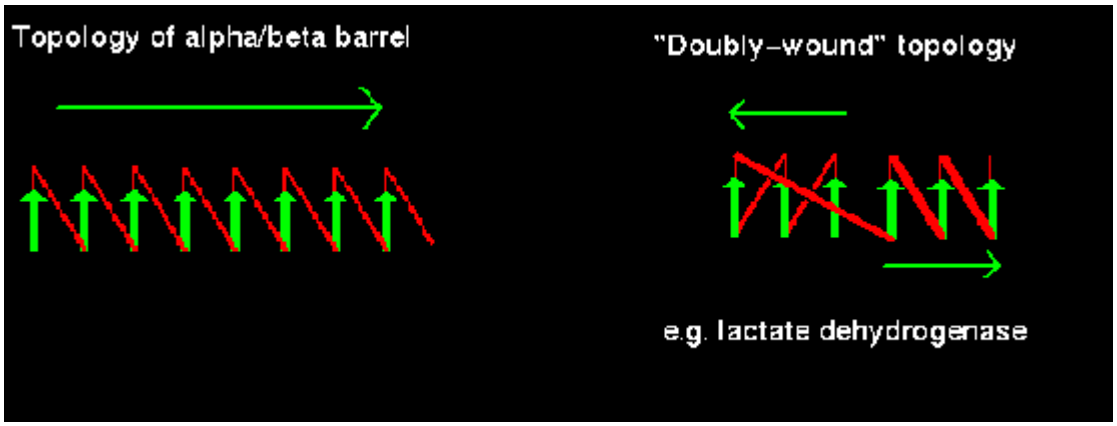


Note that the "staves" of the barrel are slanted, due to the twist of the β sheet. Also notice that there are effectively four layers to this structure. The direction of the sheet does not change (it is anticlockwise in the diagram). Such a structure may therefore be described as **singly wound**.

In a structure which is open rather than closed like the barrel, helices would be situated on only one side of the β sheet if the sheet direction did not reverse. Therefore open α/β 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:



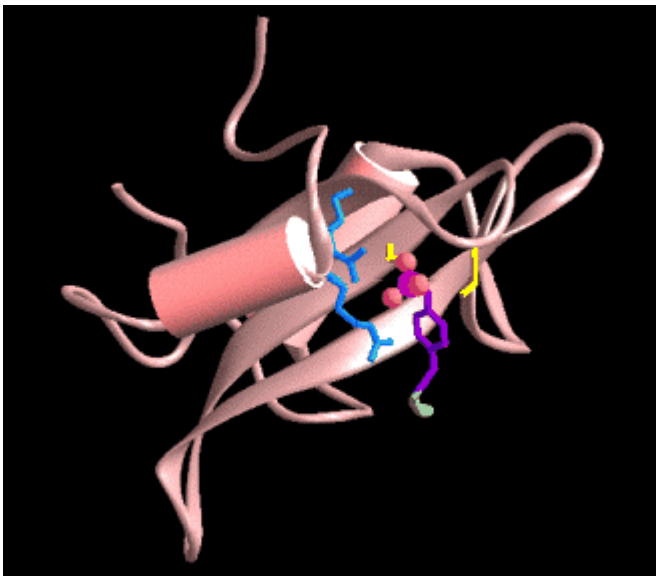
Doubly-wound topologies where the sheet begins at the edge and works inwards are rarely observed.

4.4.3 Alpha+Beta Topologies

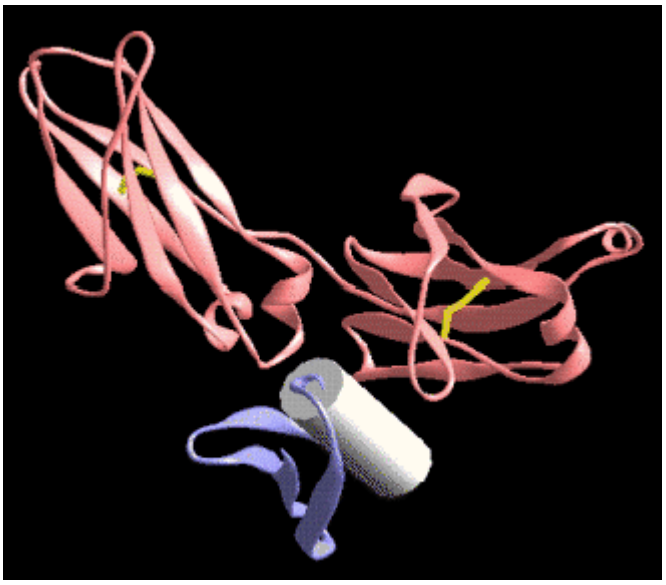
This is where we collect together all those folds which include significant alpha and beta secondary structural elements, but for which those elements are `**mixed**', in the sense that they do NOT exhibit the wound alpha-beta topology. This class of folds is therefore referred to as $\alpha+\beta$

Thus we see that this class includes:-

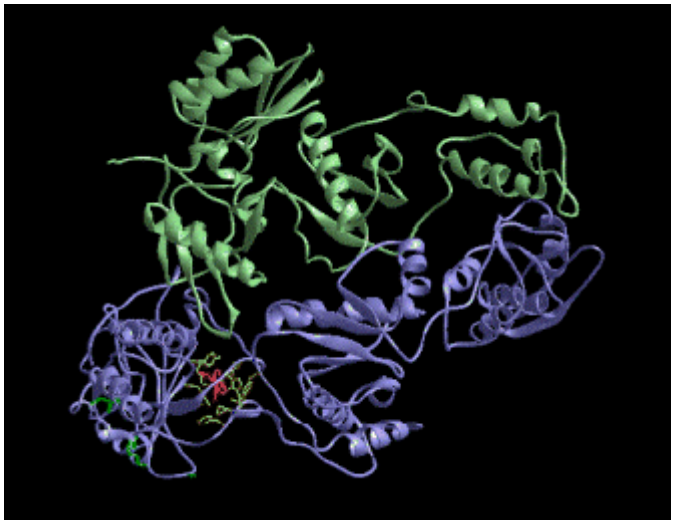
- Bacterial and mammalian pancreatic ribonucleases.
- Lysozome.
- Ubiquitin.
- Histidine-Carrier protein.
- Cysteine proteases such as papain and actinidin.
- Zinc Metallo-proteases.
- SH2 domains.



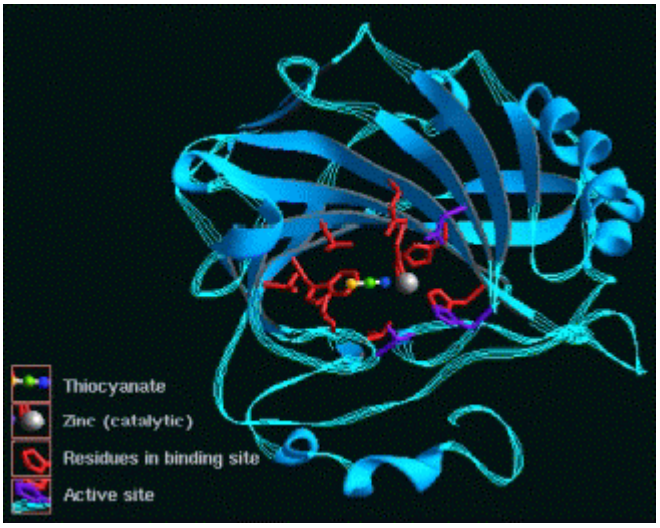
- Protein G (prokaryotic Ig-binding) in blue.



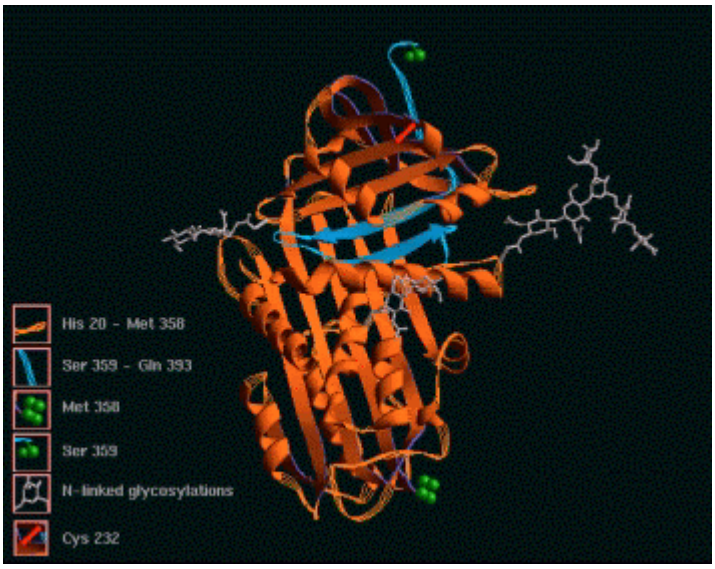
- FK506 binding protein (peptidyl-prolyl isomerase).
- Ribonuclease-H.



- Carbonic anhydrase.



- Serine protease inhibitor (Serpins).



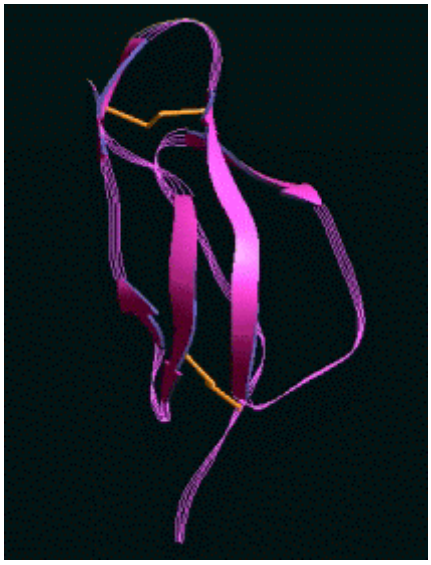
- Thymidylate synthase.

4.5 Small disulphide-rich folds

Here we see a few examples of the main families of small disulphide-rich domains of known structure. The members of these families contain a large number of disulphide bonds which stabilise the fold.

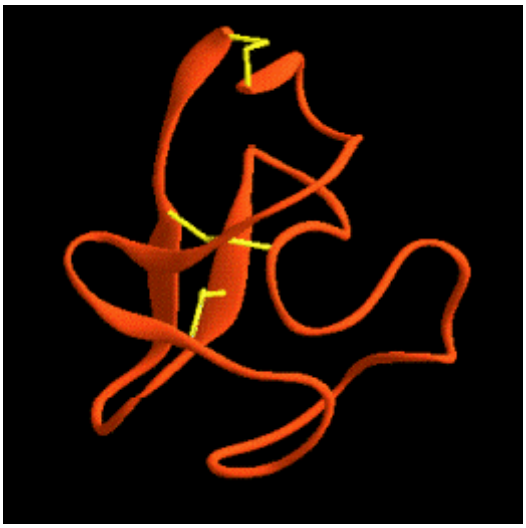
- Serine proteinase inhibitor
- Sea anemone toxin (NMR structure)
- EGF-like domain

- Complement C-module domain



- Wheat Plant Toxin; Naja (Cobra) neurotoxin; green Mamba anticholinesterase.

- Kringle domain



4.4.4 SCOP: Structural Classification of Proteins

Introduction:

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. A knowledge of these relationships is crucial to our understanding of the evolution of proteins and of development. It will also play an important role in the analysis of the sequence data that is being produced by worldwide genome projects.

The scop database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in Protein Data Bank (PDB). It is available as a set of tightly linked hypertext documents which make the large database comprehensible and accessible. In addition, the hypertext pages offer a panoply of representations of proteins, including links to PDB entries, sequences, references, images and interactive display systems. World Wide Web URL **http://scop.mrc-lmb.cam.ac.uk/scop/** is the entry point to the database (MRC site).

Existing automatic sequence and structure comparison tools cannot identify all structural and evolutionary relationships between proteins. The scop classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools to make the task manageable and help provide generality. The job is made more challenging--and theoretically daunting--by the fact that the entities being organized are not homogeneous: sometimes it makes more sense to organize by individual domains, and other times by whole multi-domain proteins.

Classification:

Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold, described below. The exact position of boundaries between these levels are to some degree subjective. Our evolutionary classification is generally conservative: where any doubt about relatedness exists, we made new divisions at the family and superfamily levels. Thus, some researchers may prefer to focus on the higher levels of the classification tree, where proteins with structural similarity are clustered.

The different major levels in the hierarchy are:

1. **Family**: *Clear evolutionarily relationship*
   Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

2. **Superfamily**: *Probable common evolutionary origin*
   Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies.
   For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.
3. **Fold**: *Major structural similarity*
   Proteins are defined as having a common fold if they have same major secondary structures in same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

Usage: We hope that scop will have broad utility that will attract a wide range of users. Experimental structural biologists may wish to explore the region of "structure space" near their proteins of current research, while theoreticians will likely find it most useful to browse the wide range of protein folds currently known. Molecular biologists may find the classification helpful because the categorization assistis in locating proteins of interest and the links make exploration easy. We also hope that scop will find pedegogical use, for it organizes structures in an easily comprehensible manner and makes them accessible from even a simple personal computer.

Table I

Currently (Release 1.48) 9912 PDB Entries were classified into 22140 Domains (excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies
All alpha proteins	126	175
All beta proteins	81	147
Alpha and beta proteins (a/b)	87	135
Alpha and beta proteins (a+b)	151	214
Multi-domain proteins (alpha and beta)	21	21
Membrane and cell surface proteins and peptides	10	16
Small proteins	44	63
Total	520	771

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

4.4.5 CATH: Classification of protein structures

Introduction:

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, class(C), architecture(A), topology(T) and homologous superfamily (H). Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures according to their toplogical connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to toplogy families and homologous superfamilies are made by sequence and structure comparisons.

CATH, can be reach on the Web at this URL:
http://www.biochem.ucl.ac.uk/bsm/cath/index.html

Domains are regions of contiguous polypeptide chain that have been described as compact, local, and semi-independent units (Richardson, 1981). Within a protein domains can be anything from independant globular units joined only by a flexible length of polypeptide chain, to units which have a very extensive interface.

CATH is now a classification of protein domains. Each protein structure in the PDB has been cut into its constituent domains, and each classified separately. The assignment of domain definitions has been made using a consensus procedure (DBS, Jones et al, (1996)), based on three independent algorithms for domain recognition (DETECTIVE (Swindells, 1995), PUU (Holm & Sander, 1994) and DOMAK (Siddiqui and Barton, 1995). This currently allows approximately 53% of the proteins to be defined as single or multidomain proteins automatically. The remaining structures are assigned domain definitions manually, by choosing what was determined to be the best assignment made by one of the algorithms, a new assignment, or an alternative assignment obtained from the literature.

Table II

Version 1.6 of CATH includes 7703 PDB entries (which contain 13 103 chains and 18 557 domains).

CATH Level	Number
Class	4
Architecture	35
Topology Family	672
Homologous Superfamily	1028
Sequence Family	1784
Near Identical Structures	3487
Identical Structures	6274

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH- A Hierarchic Classification of Protein Domain Structures. Structure. Vol 5. No 8. p.1093-1108.

5 Quaternary Structure

The quaternary structure is that level of form in which units of tertiary structure aggregate to form homo- or hetero- multimers. This is found to be remarkably common, especially in the case of enzymes. The prokaryotic biosynthesis of tryptophan provides interesting examples which fall into each of the categories below. (See Branden & Tooze, 1991)

5.1 Covalently-connected tertiary domains

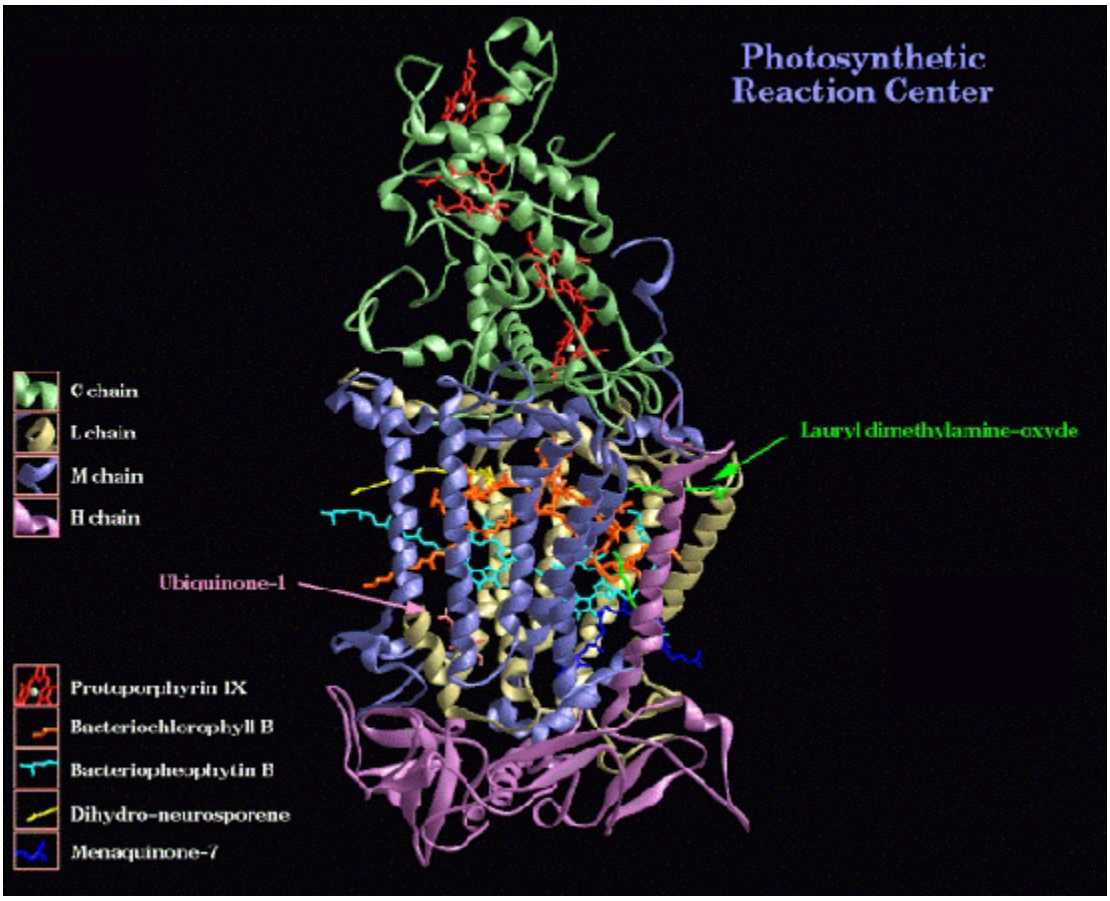
In this class of protein, domains are usually formed as modules covalently "strung together" on a single polypeptide chain. The individual chains of antibodies are like this, strings of immuno-globulin domains. However, light and heavy chains then combine to produce hetero-multimers, which may even associate into higher complexes, as with IgM.

In the case of the single polypeptide chain of pyruvate kinase there are four domains; the central TIM-barrel is the catalytic domain, whereas the other three play no direct role in enzymatic activity. However, the small N-terminal domain of 42 residues is involved in inter-subunit contacts when four copies associate to form a homo-tetramer.

E.coli produces a bifunctional enzyme which performs both the isomerisation of phospho-ribosyl anthranylate AND the synthesis of indole-glyceryl phosphate, two steps in tryptophan biosynthesis. It comprises two very similar eight-stranded α/β barrels, each barrel acting as a separate enzyme.

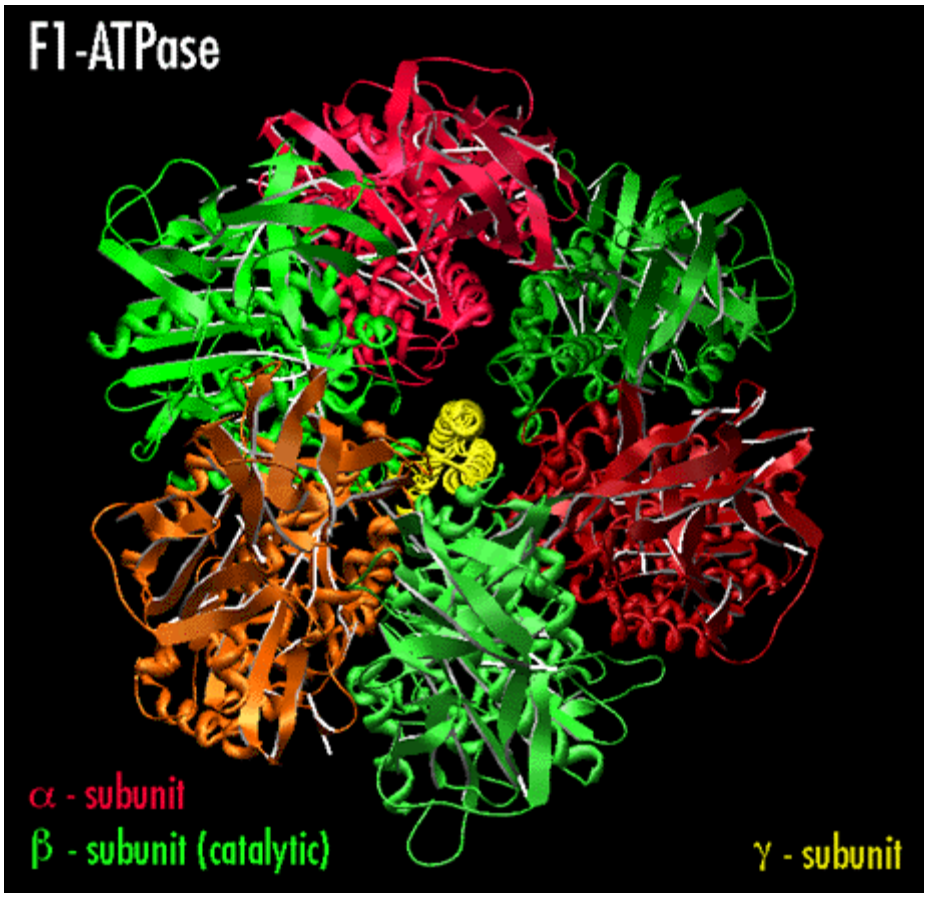
5.2 Hetero-multimers

In this case we see **different** tertiary domains aggregating together to form a unit. The photoreaction centre is a good example.

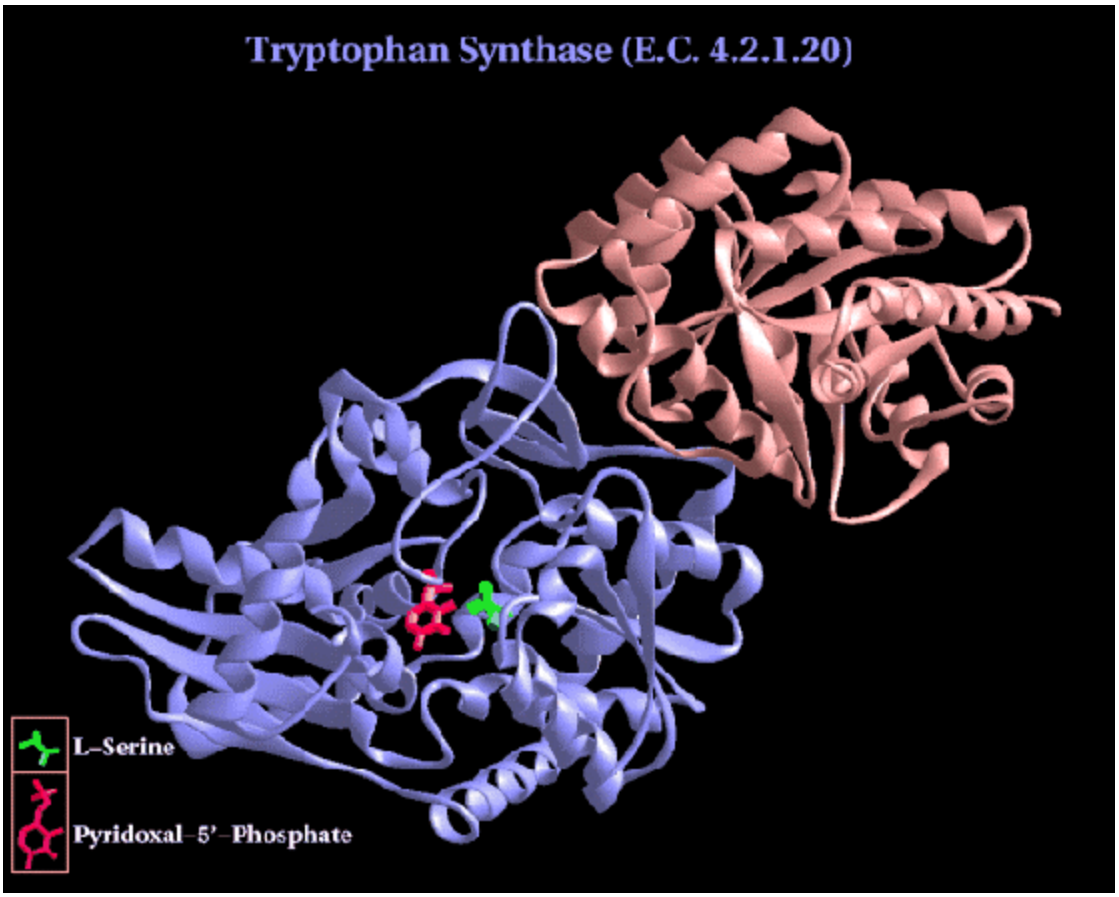


Sometimes, we find that several domains are found in a single enzyme complex, either in a single polypeptide chain, or as an association of separate chains.

Often the domains have related functions, for instance, where one domain will be responsible for binding, another for regulation, and a third for enzymatic activity. Cellobiohydrolase provides anexample of such a protein. It is not uncommon to find more than once the same chain in a protein complex. A good example is the F-1 ATPase.



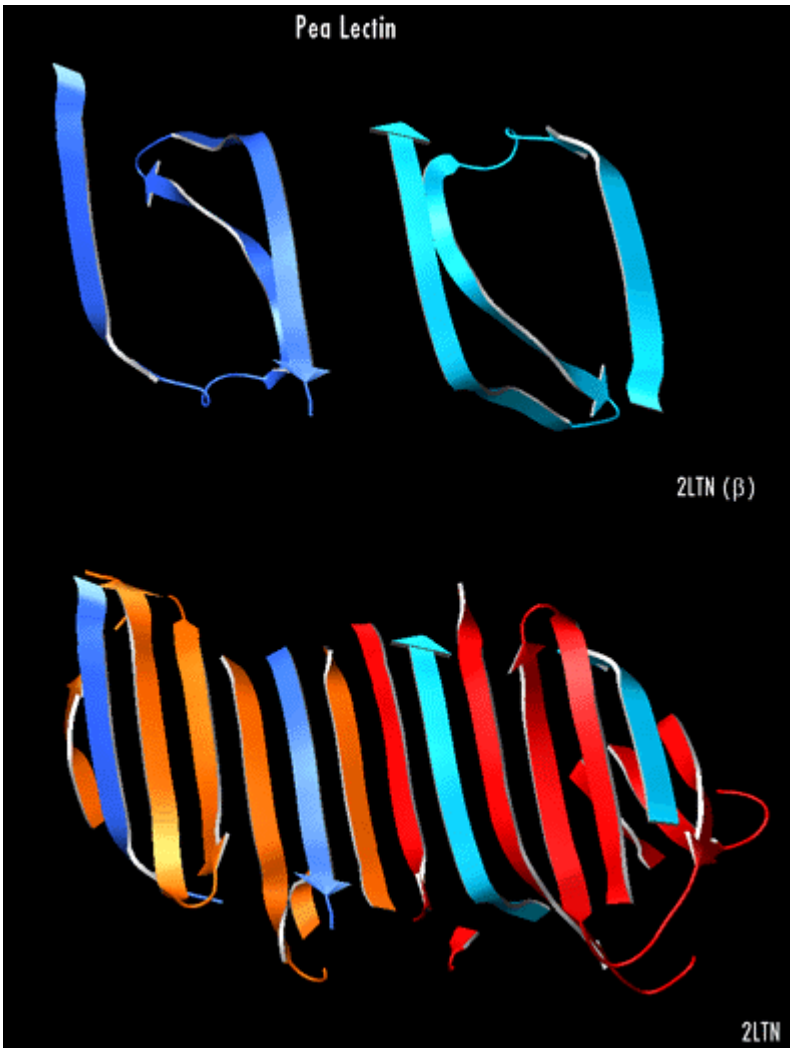
Two (further) steps in the biosynthetic pathway of tryptophan (in *S.typhimirium*) are catalysed by tryptophan synthase which consists of two separate chains, designated α and β, each of which is effectively a distinct enzyme.



The biologically active unit is a hetero-tetramer comprised of 2 α and 2 β units.

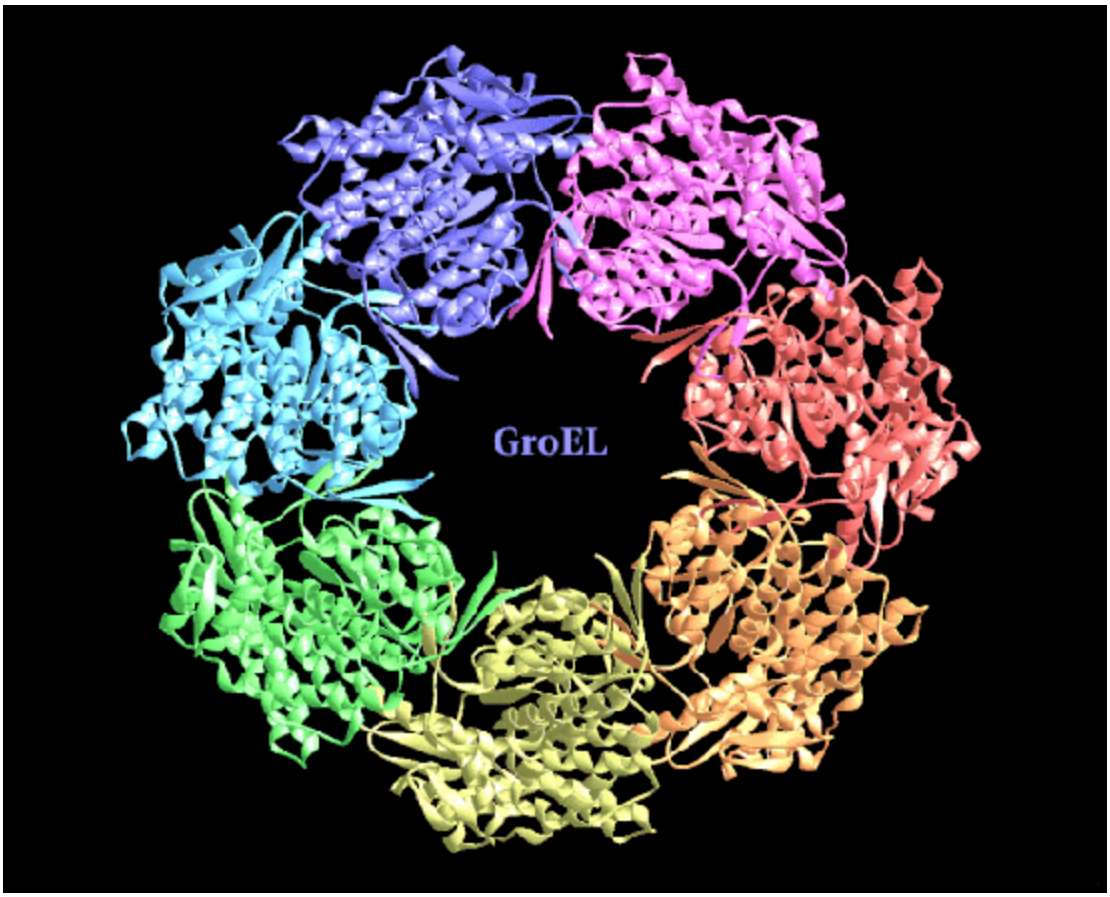
We sometimes find slightly different versions of the same protein associating. Thus, haemoglobin has both an A-chain and a B-chain, which come together to form a hetero-dimer. Two copies of this then associate to form the normal haemoglobin tetramer. Which is equivalent to an A-dimer associating with a B-dimer.

Also, it can happen that two different chains associate to form a bigger secondary structure. It is the case of the pea lectin, where a very large β-sheet is nade out of strands coming from different protein chains:



5.3 Homo-multimers

It is far more common to find copies of the **same** tertiary domain associating non-covalently. Such complexes are usually, though not always symmetrical. Because proteins are inherently asymmetrical objects, the multimers almost always exhibit rotational symmetry about one or more axes. The majority of the enzymes of the metabolic pathways seem to aggregate in this way, forming dimers, trimers, tetramers, pentamers, hexamers, octamers, decamers, dodecamers, (or even tetradecamers in the case of the chaperonin GroEL).

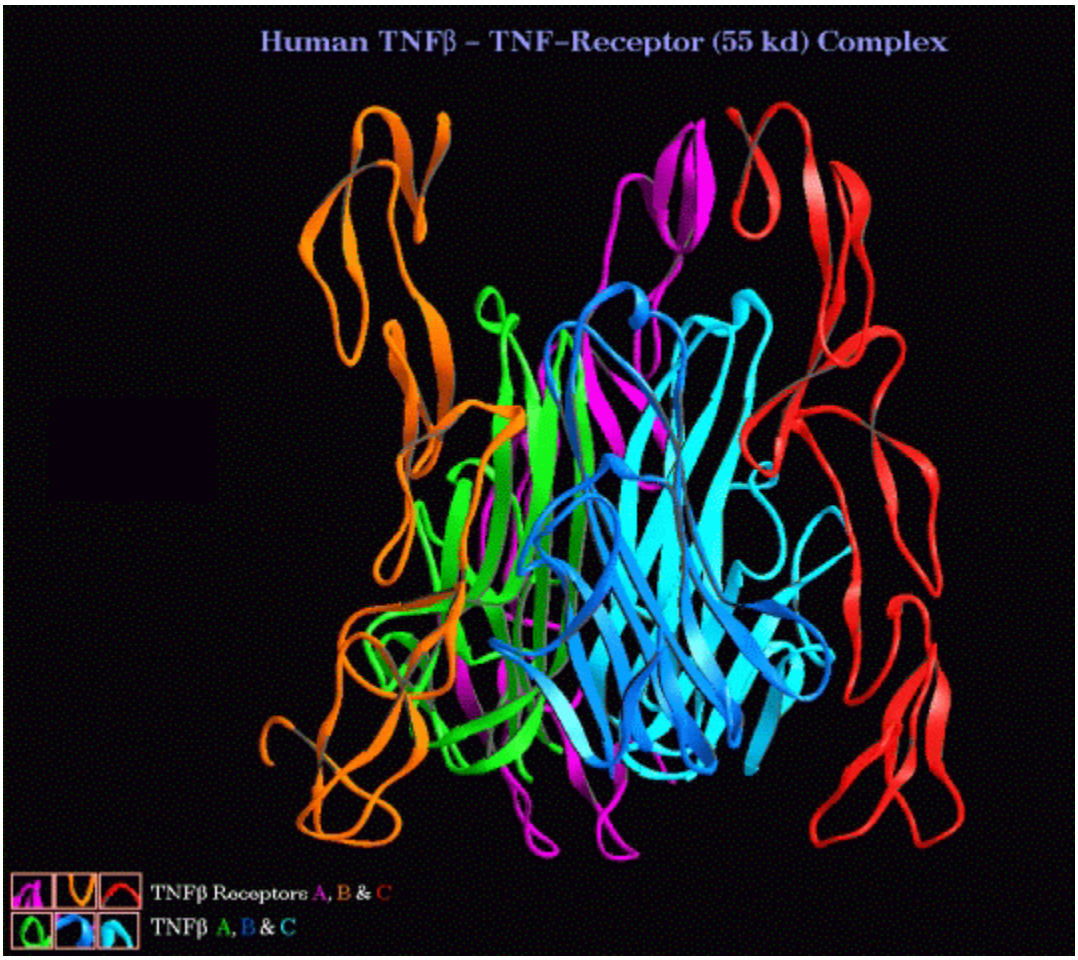


The reason for this is now thought to be the allosteric cooperativity that results in increased catalytic efficiency, effectively a "sharing" of the small conformational changes that accompany substrate binding and catalytic activity. A good well-studied example is the "breathing motion" observed in the haemoglobin tetramer.

Another interesting case study is found with the growth factors where we see dimers formed in 3 different ways, corresponding to two-fold axes in different directions.

There are other examples where dimerisation is necessary to actually create the active site of the enzyme in question. For instances the HIV protease, the viral (aspartic) protease responsible for excising the separate proteins from the single polyprotein that the virus produces once inside the cell.

In some instances, such as cytokines, homo-trimerisation leads to the formation of a functional ligand, which will pull together three single-chain receptors. Upon binding these receptors will trigger intra-cellular signals.



Examples of symmetrical enzyme multimers in the form usually found in cells have been especially prepared and archived at Brookhaven in a directory dedicated to these biological units. It may be accessed by anonymous ftp to **ftp.pdb.bnl.gov** and going to directory **/user_group/biological_units/**. The contents and associated README are worth a look.

5.4 Larger Structures

The molecular machinery of the cell and indeed of assemblies of cells, rely on components made from multimeric assemblies of proteins, nucleic acids, and sugars. A few examples include :-

- Viruses

- Microtubules
- Flagellae
- Fibres of various sorts
- Ribosomes
- Histones
- Gap Junctions

6 Comparative protein modelling

6.1 Introduction

Insights into the three-dimensional (3D) structure of a protein are of great assistance when planning experiments aimed at the understanding of protein function and during the drug design process. The experimental elucidation of the 3D-structure of proteins is however often hampered by difficulties in obtaining sufficient protein, diffracting crystals and many other technical aspects. Therefore the number of solved 3D-structures increases only slowly compared to the rate of sequencing of novel cDNAs, and no structural information is available for the vast majority of the protein sequences registered in the SWISS-PROT database (nearly 60'000 entries in release 34). In this context it is not surprising that predictive methods have gained much interest.

Proteins from different sources and sometimes diverse biological functions can have similar sequences, and it is generally accepted that high sequence similarity is reflected by distinct structure similarity. Indeed, the relative mean square deviation (rmsd) of the alpha-carbon co-ordinates for protein cores sharing 50% residue identity is expected to be around 1Å. This fact served as the premise for the development of comparative protein modelling (also often called modelling by homology or knowledge-based modelling), which is presently the most reliable method. Comparative model building consist of the extrapolation of the structure for a new (target) sequence from the known 3D-structure of related family members (templates).

While the high precision structures required for detailed studies of protein-ligand interaction can only be obtained experimentally, theoretical protein modelling provides the molecular biologists with "low-resolution" models which hold enough essential information about the spatial arrangement of important residues to guide the design of experiments. The rational design of many site-directed mutagenesis experiments could therefore be improved if more of these "low-resolution" theoretical model structures were available.

6.2 Identification of modelling templates

Comparative protein modelling requires at least one sequence of known 3D-structure with significant similarity to the target sequence. In order to determine if a modelling request can be carried out, one compares the target sequence with a database of sequences derived from the Brookhaven Protein Data Bank (PDB), using programs such as FastA and BLAST. Sequences with a FastA score 10.0 standard deviations above the mean of the random scores or a Poisson unlikelyhood probability P(N) lower than 10^{-5}

(BLAST) can be considered for the model building procedure. The choice of template structures can be further restricted to those which share at least 30% residue identity as determined by SIM.

The above procedure might allow the selection of several suitable templates for a given target sequence, and up to ten templates are used in the modelling process. The best template structure - the one with the highest sequence similarity to the target - will serve as the *reference*. All the other selected templates will be superimposed onto it in 3D. The 3D match is carried out by superimposing corresponding Cα atom pairs selected automatically from the highest scoring local sequence alignment determined by SIM. This superposition can then be optimised by maximising the number of Cα pairs in the common core while minimising their relative mean square deviation. Each residue of the reference structure is then aligned with a residue from every other available template structure if their Cα atoms are located within 3.0 Å. This generates a structurally corrected multiple sequence alignment.

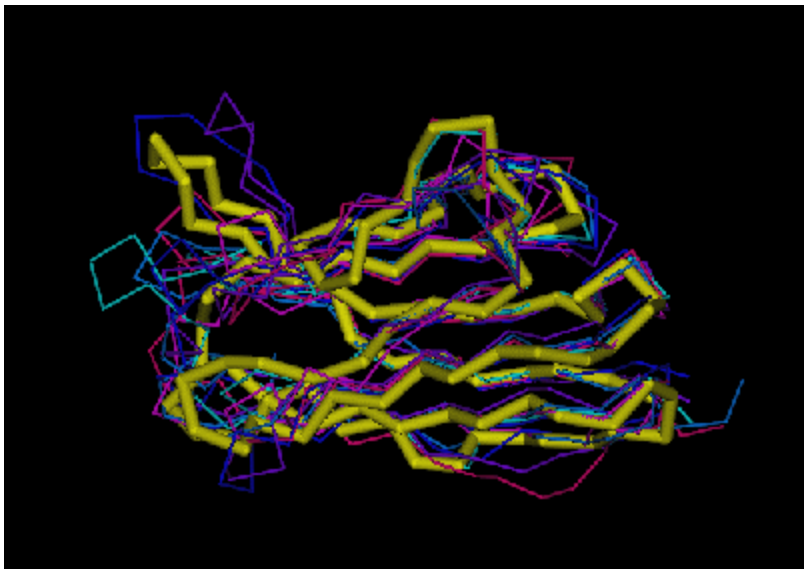
6.3 Aligning the target sequence with the template sequence

The target sequence now needs to be aligned with the template sequence or, if several templates were selected, with the structurally corrected multiple sequence alignment. This can be achieved by using the best-scoring diagonals obtained by SIM. Residues which should not be used for model building, for example those located in non-conserved loops, will be ignored during the modelling process. Thus, the common core of the target protein and the loops completely defined by at least one supplied template structure will be built.

6.4 Building the model

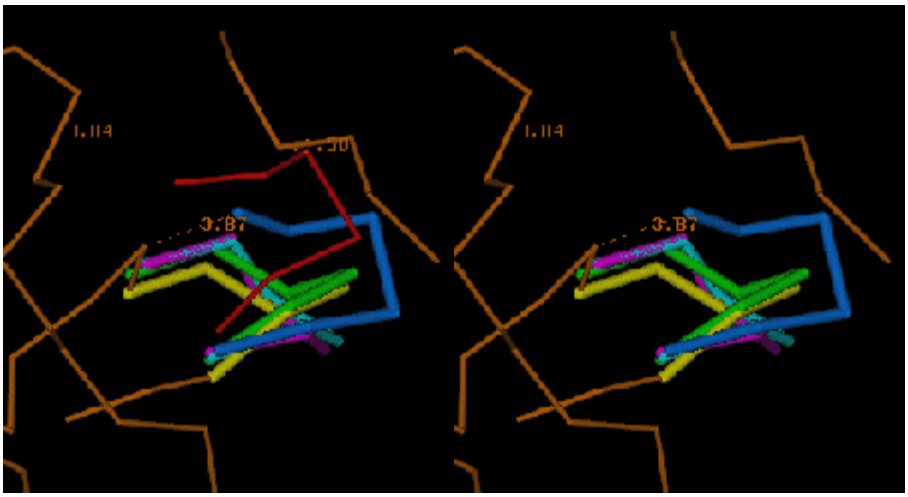
6.4.1 Framework construction

The next step is the construction of a framework, which is computed by averaging the position of each atom in the target sequence, based on the location of the corresponding atoms in the template. When more than one template is available, the relative contribution, or weight, of each structure is determined by its local degree of sequence identity with the target sequence.



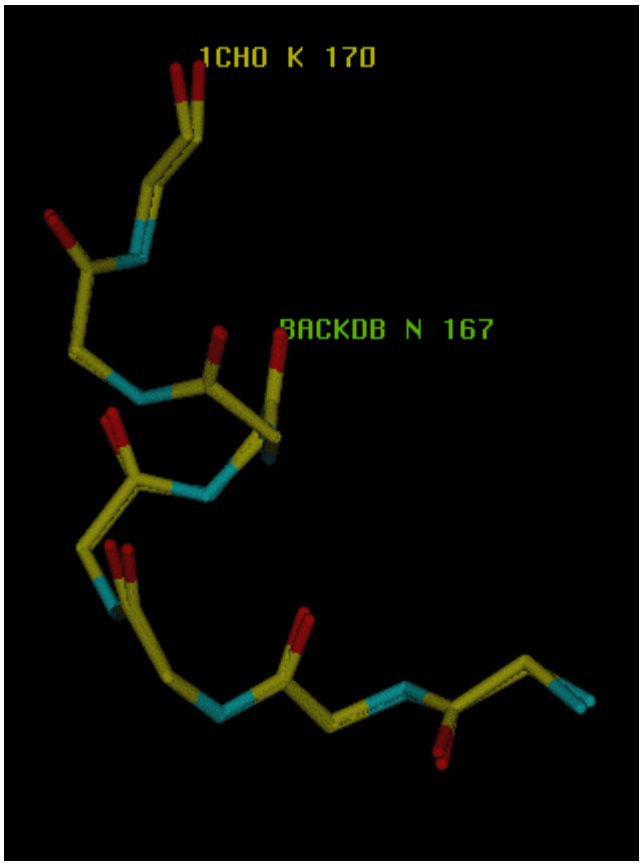
6.4.2 Building non-conserved loops

Following framework generation, loops for which no structural information was available in the template structures are not defined and therefore must be constructed. Although most of the known 3D-structures available share no overall similarity with the template, there may be similarities in the loop regions, and these can be inserted as loop structure in the new protein model. Using a "spare part" algorithm, one searches for fragments which could be accommodated onto the framework among the Brookhaven Protein Data Bank (PDB) entries determined with a resolution better than 2.5 Å. Each loop is defined by its length and its "stems", namely the alpha carbon (Cα) atom co-ordinates of the four residues preceding and following the loop. The fragments which correspond to the loop definition are extracted from the PDB entries and rejected if the relative mean square deviation (rmsd) computed for their "stems" is greater than a specified cut-off value. Furthermore, only fragments which do not overlap with neighbouring segments should be retained. The accepted "spare parts" are sorted according to their rmsd, and a Cα framework based on the five best fragments can be added to the model. In order to ensure that the best possible fragments are used for loop rebuilding, the rmsd cut-off can be incremented from 0.2 onwards until all loops are rebuilt.



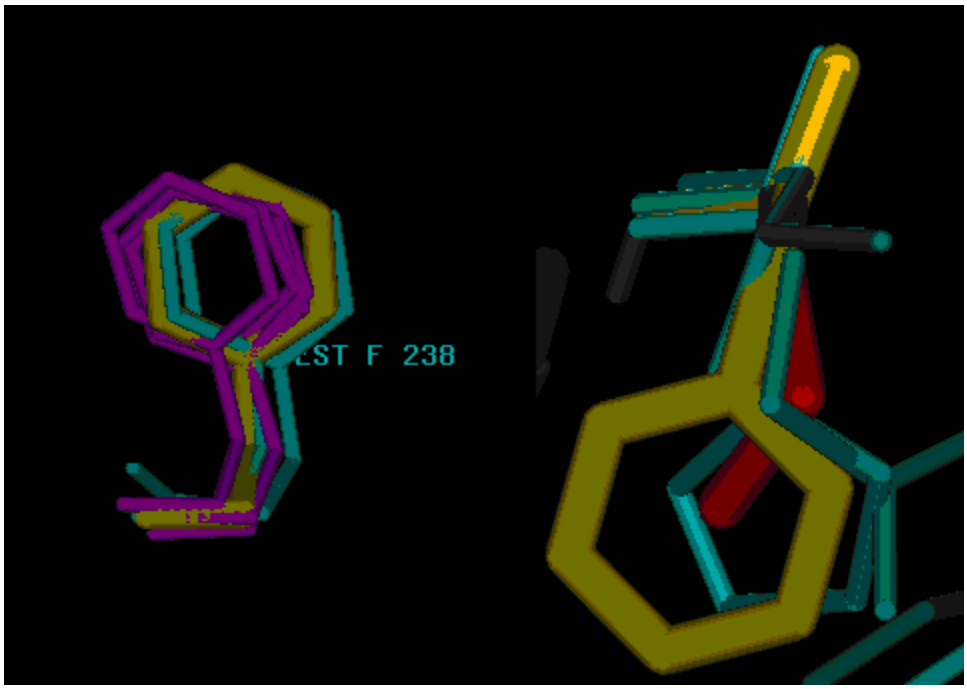
6.4.3 Completing the backbone

Since the loop building only adds Cα atoms, the backbone carbonyl and nitrogens must be completed in these regions. This step can be performed by using a library of pentapeptide backbone fragments derived from the PDB entries determined with a resolution better than 2.0 Å. These fragments are then fitted to overlapping runs of five Cα atoms of the target model. The co-ordinates of each central tripeptide are then averaged for each target backbone atom (N, C, O) and added to the model. This process yields modelled backbones that differ from experimental co-ordinates by approx. 0.2 Å rms.



6.4.4 Adding side chains

For many of the protein side chains there is no structural information available in the templates. These cannot therefore be built during the framework generation and must be added later. The number of side chains that need to be built is dictated by the degree of sequence identity between target and template sequences. To this end one uses a table of the most probable rotamers for each amino acid side chain depending on their backbone conformation. All the allowed rotamers of the residues missing from the structure are analysed to see if they are acceptable by a van der Waals exclusion test. The most favoured rotamer is added to the model. The atoms defining the $\chi 1$ and $\chi 2$ angles of incomplete side chains can be used to restrict the choice of rotamers to those fitting these angles. If some side chains cannot be rebuilt in a first attempt, they will be assigned initially in a second pass. This allows some side chains to be rebuilt even if the most probable allowed rotamer of a neighbouring residue already occupies some of this portion of space. The latter may then switch to a less probable but allowed rotamer. In case that not all of the side chains can be added, an additional tolerance of 0.15 Å can be introduced in the van der Waals exclusion test and the procedure repeated.



6.4.5 Model refinement

Idealisation of bond geometry and removal of unfavourable non-bonded contacts can be performed by energy minimisation with force fields such as CHARMM, AMBER or GROMOS. The refinement of a primary model should be performed by no more than 100 steps of steepest descent, followed by 200-300 steps of conjugate gradient energy minimisation. Experience has shown models optimised that energy minimisation (or molecular dynamics) usually move away from a control structure. It is thus necessary to keep the number of minimisation steps to a minimum. Constraining the positions of selected atoms (such as Cα, or using a B-factor based function) in each residue generally helps avoiding excessive structural drift during force field computations.

7 *De novo* modelling of G-protein coupled receptors

7.1 Introduction

G-protein coupled receptors (GPCR), are seven-helix trans-membrane proteins which are essential in a great variety of physiological events that require the transmission of an external to an intracellular signal. Present in a broad range of organisms, they cause the activation of a guanine nucleotide-binding protein (G-protein) in response to stimuli as diverse as light, odorants, neurotransmitters, peptides and hormones.

Understanding the structure and mechanism of GPCRs is thus central to many aspects of cellular signalling and control. As a result, many multidisciplinary research projects, including those carried out by pharmaceutical companies to find new therapeutic molecules, are aimed at one or another member of this protein family (over 700 known members).

Since no experimentally elucidated 3D-structure is available for GPCRs, and given the high level of interest these proteins attract, it is not surprising that predictive methods to derive information about their 3D-structure are rapidly gaining interest. These efforts have resulted in several theoretical models of the trans-membrane regions that have been successfully used to rationalise site-directed mutagenesis experiments aimed at understanding the interactions between the proteins and their ligands.

7.2 Building GPCR templates

Template models for GPCR can be built using a newly described rule-based technique (Herzyk and Hubbard, 1995) based on the assumption that it is possible to generate models of the trans-membrane regions of GPCRs using (i) experimental data available for at least one member of a particular GPCR subfamily; (ii) the multiple sequences alignment of all known members of this subfamily and (iii) the 2D projection map available for bovine rhodopsin (Schertler *et al,* 1993). The methodology also requires the clear definition of the sequence of the trans-membrane regions. This is normally achieved by combining the helix assignments determined by J. Baldwin (Baldwin, 1993), with those obtained from the multiple sequence alignments using programs such as TMAP (Persson and Argos, 1994) and TopPred (von Heijne, 1992).

Since no experimental structure information is available for the loops connecting the helices of GPCRs, and due to the present lack of reliable *de novo* modelling methods for long loops, the latter will not be predicted and only the trans-membrane helices will be modelled.

Template building is divided into two stages. Firstly, a simplified template is constructed on the basis of experimental and theoretical data (Herzyk and Hubbard, 1995). Secondly, this template is converted into all-atom representation (Peitsch, 1995; Peitsch, 1996).

7.2.1 Template building, Stage 1

Throughout the first stage, the seven trans-membrane helices are represented as rigid, idealised helices. Residues belonging to these helices are represented by one Cα atom and one virtual side-chain atom whose size and position depends on the size and topology of the amino acid side-chain.

The simplified templates, for the way the helices are fitted together, are generated by the global optimisation of a penalty function (Herzyk and Hubbard, 1995). This function is to measure the violations of the structural restraints derived from experimental and theoretical data. The method currently considers restraints imposed on the positions of the helices with respect to the 2D projection map of bovine rhodopsin (XY dimension) and to the membrane plane (Z dimension), as well as orientational restraints which determine the lipid- and interior-facing portions of each helix. Additional distance restraints can be used to describe the relative position of selected residues and a ligand. The restraints on helix orientation may be derived from both mutagenesis experiments as well as from multiple sequence alignments. The relative distances between specific side chains (themselves determined by site-directed mutagenesis) and a ligand are based on their molecular structure and possible non-bonded contact distances (Herzyk and Hubbard, 1995). Each of the restraints is satisfied if its value is between pre-set limits, otherwise a contribution is added to the penalty function.

Monte-Carlo Simulated Annealing (MCSA) (Kirkpatrick *et al,* 1983) is then used to globally optimise the penalty function and produce an optimal model. Calculations for a MCSA trajectory begin with the randomised configuration of rigid body elements, which are then moved sequentially by a random step of a randomly chosen co-ordinate. Each trajectory consist of 25 temperature runs of one thousand Monte-Carlo steps. Using a collection of different starting positions of the model, we compute 50 MCSA trajectories. The final configurations are then averaged and this mean structure is idealised to yield the final template. In order to validate the method we performed the calculations for bacteriorhodopsin using the experimental data available prior to the high resolution electron microscopy (EM) structure (Henderson *et al,* 1990). This produced a simplified template in which Cα atom positions differed from the experimental EM structure by only 1.87Å root mean square deviation. All calculations are performed using a purpose built program "PANDA" developed at York (Herzyk and Hubbard, 1995).

7.2.2 Template building, Stage 2

At the second stage the initial model, is converted to a full atom representation using the backbone rebuilding and side chain reconstitution routines of *ProMod* (Peitsch, 1995; Peitsch, 1996). The stereochemistry of this model is then further optimised by 200 cycles of conjugate gradient energy minimisation using force fields such as CHARMm (Brooks *et al*, 1983).

7.3 Using the GPCR templates

As for the template building stage outline above, the modelling of other members of the GPCR family of proteins requires the clear definition of the sequence of the trans-

membrane helices. This can be achieved by combining the helix assignments determined by J. Baldwin (Baldwin, 1993), with those obtained from the multiple sequence alignments using programs such as TMAP (Persson and Argos, 1994) and TopPred (von Heijne, 1992). As these predictions vary from one method to another and "manual" adjustments are generally necessary to overcome the limitations of the methods, the present version of the server does not provide an automated procedure for helix identification. In this way, the user may also test several helix sequence variations. The next step to obtain a GPCR model consist essentially of a comparative protein modelling attempt using the above described modelling templates.

7.6 Specific references

Baldwin JM (1993) EMBO J **12** 1693-1703

Henderson R Baldwin JM Ceska TA Zemlin F Beckmann E and Downing KH (1990) J Mol Biol **213** 899-929

Herzyk P and Hubbard RE (1995) Biophys J **69** 2419-2442

Kirkpatrick SCD Gellat J and Vecchi MP (1983) Science **220** 671-680

Persson B and Argos P (1994) J Mol Biol **227** 493-509

Schertler GFX Villa C and Henderson R (1993) Nature **362** 770-772

von Heijne G (1992) J Mol Biol **225** 487-494

8 How to evaluate the quality of a model

8.1 General considerations.

- A model is considered wrong if at least part of its structural features a miss-placed relatively to the rest of the model. Errors of that type can very easily slip into a model when erroneous sequence alignments are used during the building procedure. Such models can nevertheless have proper stereochemistry if one gives great care to this aspect during the building procedure.
- In absolute terms a model can be declared inaccurate or imprecise if its atomic co-ordinates are not within 0.5 A rmsd of a control experimental structure. This value comes from the structure/sequences similarity study of Chothia and Lesk [6], in which they demonstrate that different structures of a same protein can deviate by as much as 0.5 A. This criterion can however only be assessed after the fact, and is thereby not usable. In relative terms, however, a model can be considered "*accurate enough*" or as "*accurate as you can get*" when its rmsd is within the spread of deviations observed for experimental structures displaying a similar sequence identity level as the target and template sequences.
  Another source of inaccuracy is the deviation from ideal stereochemical values for bond lengths and angles. Such inaccuracies can be easily detected with the program WhatCheck developed by G. Vriend at the EMBL (Can be reached from the PDB Web site).
- It is crucial to realise that proper stereochemistry as can be assessed with WhatCheck is not a criteria for model correctness. In other terms, it is possible, to build models which would comply with such criteria and have strictly no biological meaning.
- Empirical pair-potentials allow, to some degree, the detection of such errors in models. These algorithms are indeed not sensitive enough to detect subtle differences in conformation but are quite efficient at pointing out regions where sequence and structure do not fit.

8.2 What are the sources of errors and inaccuracies?

The quality of a model is determined by two criteria, which will define its applicability (see Part IV):

1. The *correctness* of a model is essentially dictated by the quality of the sequence alignment used to guide the modelling process. If the sequence alignment is wrong in some regions, then the spatial arrangement of the residues in this portion of the model will be incorrect.
2. The *accuracy* of a model is essentially limited by the deviation of the used template structure(s) relative to the (a future) experimental control structure. This limitation is inherent to the methods used, since the models atoms of protein models result from an extrapolation. As a consequence, the C which share 35 to

50% sequence identity with their templates, will generally deviate by 1.0 to 1.5 Å from their experimental counter parts, as do similarly related experimental structures [6]. Furthermore, structural differences between predicted and experimental structures have two sources:

- The errors inherent to the modelling procedures.
- The variations caused by the molecular environment and data collection method incorporated into experimentally elucidated structures which will be used as modelling templates. Indeed, crystallographic structures of identical proteins can vary not only because of experimental errors and differences in data collection conditions (illustrated in [32]) and refinement, but also because of different crystal lattice contacts and the presence or absence of ligands. One of the most interesting examples in which several structures of the same protein, determined by different methods, were compared involves interleukin-4 (IL-4) [33 and references therein]. This cytokine consists of a 130 residue four helix bundle, and its structure was elucidated by x-ray crystallography as well as by NMR. The backbones of three IL-4 crystal structures (PDB entries 1RCB, 2INT and 1HIK) show an rmsd of 0.4 to 0.9 Å, while those of three IL-4 NMR forms (PDB entries 1ITM, 1CYL and 2CYK) give rmsd of 1.2 to 2.6 Å. These values illustrate the structural differences due to experimental procedures and the molecular environment at the time of data collection. Therefore, "*a protein model derived by comparative methods cannot be more accurate than the difference between the NMR and crystallographic structure of the same protein.*" [33].

8.3 Protein core and loops.

Almost every protein model contains non-conserved loops which are expected to be the least reliable portions of a protein model. Indeed, these loops often deviate markedly from experimentally determined control structures. In many cases, however, these loops also correspond to the most flexible parts of the structure as evidenced by their high crystallographic temperature factors (or multiple solutions in NMR experiments). On the other hand, the core residues - the least variable in any given protein family - are usually found in essentially the same orientation as in experimental control structures, while far larger deviations are observed for surface amino acids. This is expected since the core residues are generally well conserved and the conformation of their side chains are constrained by neighbouring residues. In contrast, the more variable surface amino acids will tend to show more deviations since there are few steric constraints imposed upon them.

8.4 Detecting major errors using empirical pair potentials.

Some structural aspects of a protein model can be verified using methods based on the inverse folding approach. Two of them, namely the 3D-1D profile based verification method [15] and *ProsaII* developed by M. Sippl [16], are widely used. The 3D-1D profile of a protein structure is calculated by adding the probability of occurrence for each residue in its 3D-context [15]. Each of the twenty amino acids has a certain probability to

be located in any environmental classes (defined by criteria such as solvent-accessible surface, buried polar, exposed non-polar area and secondary structure) defined by Eisenberg and colleagues. In contrast, *ProsaII* [16] relies on empirical energy potentials derived from the pairwise interactions observed in well defined protein structures. These terms are summed over all residues in a model and result in a more or less favourable energy.
Both methods can detect a global sequence to structure incompatibility and errors corresponding to topological differences between template and target. They also allow the detection of more localised errors such as β-strands that are "out of register" or buried charged residues. These methods are however unable to detect the more subtle structural inconsistencies often localised in non-conserved loops, and cannot provide an assessment of the correctness of their geometry.

8.5 Applicability of model structures.

Protein model obtained with comparative modelling methods can be classified into three broad categories:

1. Models which are based on incorrect alignments between target and template sequences. Such alignment errors, which generally reside in the inaccurate positioning of insertions and deletions, are caused by the weaknesses of the alignment algorithms and can generally not be resolved in the absence of a control experimental structure. It is however often possible to correct such errors by producing several models based on alignment variants and by selecting the most "sensible" solution. Nevertheless, it turns out that such models are often useful as the errors are not located in the area of interest, such as within a well conserved active site.
2. Models based on correct alignments are of course much better, but their accuracy can still be medium to low as the templates used during the modelling process have a medium to low sequence similarity with the target sequence. Such models, as the ones described above, are however very useful tools for the rational mutagenesis experiment design. They are however of very limited assistance during detailed ligand binding studies.
3. The last category of models comprises all those which were build based on templates which share a high degree of sequence identity (> 70%) with the target. Such models have proven useful during drug design projects and allowed the taking of key decisions in compound optimisation and chemical synthesis. For instance, models of several species variants of a given enzyme can guide the design of more specific non-natural inhibitors.

However, nothing is absolute and there are numerous occasions in which models falling in any of the above categories, could either not be used at all or in contrast proved to be more useful and correct than estimated.

9 The SWISS-MODEL server

The aim of the Internet-based SWISS-MODEL server is to provide a comparative protein modelling tool independent from expensive computer hardware and software.

URL: http://www.expasy.ch/swissmod/SWISS-MODEL.html

The SWISS-MODEL server is reachable on the World Wide Web (WWW) and requests can be submitted through easy to fill forms. Since protein modelling is heavily dependent on the alignment between target and template sequences, SWISS-MODEL provides four distinct modes of function accessible through separate forms:

1. The *First Approach Mode*: This mode allows the user to submit a sequence or its Swiss-Prot identification code. In this mode, SWISS-MODEL will go through the complete procedure described above. The *First Approach Mode* also allows the user to define a choice of pre-selected template structures, thereby overruling the automated selection procedure. The results of the modelling procedure will be returned to the user via E-mail.
2. The *Optimise Mode* allows the user to re-compute a model by submitting altered sequence alignments and ProMod command files. The sequence alignment procedure, which is fully automatic in the *First Approach Mode*, may yield sub-optimal alignments and consequently lead to erroneous models. The automated alignment of moderately similar sequences is indeed often imprecise and the boundaries of non-conserved loops are frequently ill defined and miss-aligned. These regions of the sequence alignments must therefore be corrected by hand in order to overcome these weaknesses. *The Optimise Mode* allows the user to do such corrections, and to request the remodelling of the sequence by submitting his own sequence alignment. This is best done by preparing a modelling request within the Swiss-PdbViewer (see chapter 12). These requests can then be saved as "HTML" files and then submitted through a Web browser.
3. The Combine Mode allows the user to combine independently modelled protein chains in a quaternary complex, based on an existing assembly template. In the current version of the server, this assembly template must be a PDB file containing the desired protein complex. The server provides detailed guidelines on how to submit requests to this particular facility. The actions taken by the server include a superposition of each modelled chain onto its respective counterpart on the assembly template and an energy minimisation of the whole complex.
4. The GPCR Mode. To get a GPCR sequence modelled, the first step consists of selecting a template for comparative modelling from the list of available ones. These have been created *a priori* using the methodology described in chapter 7. Then the user types (or uses a "cuts and pastes" function) to introduce the

sequences of the seven trans-membrane helices, along with an E-mail address, a contact name, and a title for the request. The *SWISS-GPCR modelling* pages provide detailed help in Hypertext format, links to other sites relevant to seven trans-membrane receptors, and a demonstration version of a filled in form. Furthermore, the user may query a special index of all GPCR sequences in the SWISS-PROT database, and search their sequences using a BLAST interface (Altschul *et al*, 1990). This allows GPCR sequences to be searched rapidly, and will also tell the user which of the available templates is the most appropriate one on which to model the query sequence. The current version of the SWISS-MODEL server does not allow the user to create a new template using his own experimental data. This can however be achieved upon direct contact with Pawel Herzyk at York.

The SWISS-MODEL Web interface provides help and guidelines to the use of the difference server modes.

10 Swiss-PdbViewer

The aim of Swiss-PdbViewer is to help biologists compare proteins structures, by the mean of a free user-friendly integrated package. As molecular biologists tend to work more with Macintoshes and PCs than with UNIX systems, Swiss-PdbViewer has initially been developed on microcomputers, although a X-Window version is also planned.

A more complete description of the program features, as well as executables, user guide and tutorial can be found on the ExPASy server:

URL: http://www.expasy.ch/spdbv/mainpage.htm

10.1 Using Swiss-PdbViewer as an interface to SWISS-MODEL

As explained in the chapter 6.6, the SWISS-MODEL server functions in two modes. In order to obtain a reliable model, one may need to use the server in *"Optimise mode"*, which requests the user to edit the alignment so that insertions and deletions (indels) are located in reasonable locations (generally in loops, as secondary structure elements tend to be conserved among homologous proteins). Swiss-PdbViewer let the user interactively alter the indels placement, and a visual feedback is provided in real time. Once the best alignment has been decided, a direct submission to SWISS-MODEL can be performed, Swiss-PdbViewer taking care of providing alignments and command files formatted correctly.

PDB files do not necessarily contain all amino-acids, nor all atoms of an amino-acid. Missing atoms are generally part of amino-acids whose side chain points toward the solvent. Indeed, these side chains are more labile and might provide insufficient electronic density to be precisely located.

In some cases alternate atoms positions are given, when electronic density depicts two allowed atom positions for the same atom. SWISS-MODEL can be confused by such incomplete PDB files, therefore all PDB files have been passed through a filter that renumbers all chains from 1, store only one chain by file, and removes uncertain information, as well as the header, solvent molecules and HETATMs. These files are named EXPDB and are the starting point of all comparative modelling. It is therefore very important to prepare your best alignments using such files. Swiss-PdbViewer will not allow you to submit modelling requests if you are not using EXPDB files as templates.

11 Autopsy of a PDB file

Atom coordinates of protein and nucleic acid structures are distributed under the form of PDB files. Those are 80 column formatted text files and present the advantage of being platform independent.

The first six columns are reserved for a keyword describing the type of information that follows on the line (such as HEADER, JRNL, REMARK, ATOM, HETATM, and so on).

A typical PDB file contains a header with information about the entry, literature references, as well as additional remarks that may contain information about how the protein was crystallised, the resolution and so on...

A partial example of PDB file is given below; parts removed are signalled by (...).

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901
234567890
--------------------------------------------------------------------------
---------

HEADER    OXIDOREDUCTASE(NAD(A)-CHOH(D))          12-APR-89   4MDH
4MDH   3
COMPND    CYTOPLASMIC MALATE DEHYDROGENASE (E.C.1.1.1.37)
4MDH   4
SOURCE    PORCINE (SUS $SCROFA) HEART
4MDH   5
AUTHOR    J.J.BIRKTOFT,L.J.BANASZAK
4MDH   6
REVDAT   3   15-APR-92 4MDHB   3       ATOM
4MDHB  1
REVDAT   2   15-JAN-90 4MDHA   1       JRNL
4MDHA  1
REVDAT   1   19-APR-89 4MDH    0
4MDH   7
SPRSDE     19-APR-89 4MDH      2MDH
4MDH   8
JRNL        AUTH   J.J.BIRKTOFT,G.RHODES,L.J.BANASZAK
4MDH   9
JRNL        TITL   REFINED CRYSTAL STRUCTURE OF CYTOPLASMIC MALATE
4MDHA  2
JRNL        TITL 2 DEHYDROGENASE AT 2.5-*ANGSTROMS RESOLUTION
4MDHA  3
JRNL        REF    BIOCHEMISTRY                  V.  28  6065 1989
4MDHA  4
JRNL        REFN   ASTM BICHAW  US ISSN 0006-2960                 033
4MDHA  5
REMARK   1
4MDH  14
REMARK   1 REFERENCE 1
4MDH  15
```

```
REMARK   1  AUTH   J.J.BIRKTOFT,Z.FU,G.E.CARNAHAN,G.RHODES,
4MDH  16
REMARK   1  AUTH 2 S.L.RODERICK,L.J.BANASZAK
4MDH  17
REMARK   1  TITL   COMPARISON OF THE MOLECULAR STRUCTURES OF
4MDH  18
REMARK   1  TITL 2 CYTOPLASMIC AND MITOCHONDRIAL MALATE DEHYDROGENASE
4MDH  19
REMARK   1  REF    TO BE PUBLISHED
4MDH  20
REMARK   1  REFN                                                  353
4MDH  21
                                  (...)
```

The next section provides information on the amino-acid sequence of each chain. The current example contains two chains (A and B).

```
SEQRES   1 A  334  ACE SER GLU PRO ILE ARG VAL LEU VAL THR GLY ALA ALA
4MDH 163
SEQRES   2 A  334  GLY GLN ILE ALA TYR SER LEU LEU TYR SER ILE GLY ASN
4MDH 164
SEQRES   3 A  334  GLY SER VAL PHE GLY LYS ASP GLN PRO ILE ILE LEU VAL
4MDH 165

                                  (...)

SEQRES  24 A  334  VAL GLU GLY LEU PRO ILE ASN ASP PHE SER ARG GLU LYS
4MDH 186
SEQRES  25 A  334  MET ASP LEU THR ALA LYS GLU LEU ALA GLU GLU LYS GLU
4MDH 187
SEQRES  26 A  334  THR ALA PHE GLU PHE LEU SER SER ALA
4MDH 188
SEQRES   1 B  334  ACE SER GLU PRO ILE ARG VAL LEU VAL THR GLY ALA ALA
4MDH 189
SEQRES   2 B  334  GLY GLN ILE ALA TYR SER LEU LEU TYR SER ILE GLY ASN
4MDH 190
SEQRES   3 B  334  GLY SER VAL PHE GLY LYS ASP GLN PRO ILE ILE LEU VAL
4MDH 191

                                  (...)

SEQRES  24 B  334  VAL GLU GLY LEU PRO ILE ASN ASP PHE SER ARG GLU LYS
4MDH 212
SEQRES  25 B  334  MET ASP LEU THR ALA LYS GLU LEU ALA GLU GLU LYS GLU
4MDH 213
SEQRES  26 B  334  THR ALA PHE GLU PHE LEU SER SER ALA
4MDH 214

                                  (...)
```

The next section contains optional information about HET groups (see the HETATM section that will follow for a more detailed description).

```
HET    NAD  A   1      44     NAD CO-ENZYME
4MDH 219
HET    SUL  A   2       5     SULFATE
4MDH 220
HET    NAD  B   1      44     NAD CO-ENZYME
4MDH 221
HET    SUL  B   2       5     SULFATE
4MDH 222
FORMUL   3  NAD    2(C21 H28 N7 O14 P2)
4MDH 223
FORMUL   4  SUL    2(O4 S1)
4MDH 224
FORMUL   5  HOH   *471(H2 O1)
4MDH 225

                                    (...)
```

The next section describe secondary structure elements (HELIX, SHEET and TURN) as they have been provided by the crystallographer. This can be subjective as the definition of these secondary structure elements is loose.

```
HELIX    1 1BA GLY A   13  LEU A   20  1
4MDH 226
HELIX    2 2BA LEU A   20  GLY A   26  1
4MDH 227
HELIX    3  CA MET A   45  ALA A   60  1
4MDH 228
                                    (...)
SHEET    1 S1A 6 LEU A  63  THR A  70  0
4MDH 250
SHEET    2 S1A 6 PRO A  34  ASP A  41  1
4MDH 251
SHEET    3 S1A 6 ILE A   4  GLY A  10  1
4MDH 252
                                    (...)
TURN     1  T1 VAL A   8  ALA A  11
4MDH 274
TURN     2  T2 GLY A  10  GLY A  13
4MDH 275
TURN     3  T3 GLY A  26  PHE A  29
4MDH 276
                                    (...)
```

The next section describe crystallographic information (crystal groups)

```
CRYST1  139.200   86.600   58.800  90.00  90.00  90.00 P 21 21 2     8
4MDH 328
ORIGX1      1.000000  0.000000  0.000000        0.00000
4MDH 329
ORIGX2      0.000000  1.000000  0.000000        0.00000
4MDH 330
ORIGX3      0.000000  0.000000  1.000000        0.00000
4MDH 331
```

```
SCALE1      0.007184  0.000000  0.000000        0.00000
4MDH 332
SCALE2      0.000000  0.011547  0.000000        0.00000
4MDH 333
SCALE3      0.000000  0.000000  0.017007        0.00000
4MDH 334
MTRIX1   1 -0.865540  0.467810 -0.178880       55.21400    1
4MDH 335
MTRIX2   1  0.499790  0.829880 -0.248020       -1.79900    1
4MDH 336
MTRIX3   1  0.032420 -0.304070 -0.952100       89.13300    1
4MDH 337

                                   (...)
```

And finally atom coordinates for amino-acids (or nucleic acids) are provided. Each line starts with the ATOM keyword, and is followed by atom number, atom name, amino-acid name, chain name, amino-acid number, X, Y, Z coordinates, atom weight, and B-factor (this last number can be viewed as an incertitude factor (0-100). A low B-factor meaning that the position of the atom has been determined with accuracy. Typically B-factors of Alpha Carbons (CA) are lower than atoms located at side chain extremities.

```
ATOM      1  C   ACE A   0      11.590   2.938  35.017  1.00 45.90
4MDHB  5
ATOM      2  O   ACE A   0      12.581   2.371  35.517  1.00 28.75
4MDHB  6
ATOM      3  CH3 ACE A   0      10.179   2.477  35.417  1.00 36.75
4MDHB  7
ATOM      4  N   SER A   1      11.648   3.946  34.081  1.00 49.10
4MDH 341
ATOM      5  CA  SER A   1      12.901   4.557  33.573  1.00 52.42
4MDH 342
ATOM      6  C   SER A   1      12.733   5.624  32.482  1.00 48.48
4MDH 343
ATOM      7  O   SER A   1      13.238   5.432  31.363  1.00 57.03
4MDH 344
ATOM      8  CB  SER A   1      13.990   3.553  33.162  1.00 41.45
4MDH 345
ATOM      9  OG  SER A   1      15.105   3.679  34.039  1.00 42.59
4MDH 346
ATOM     10  N   GLU A   2      12.073   6.774  32.772  1.00 37.72
4MDH 347
ATOM     11  CA  GLU A   2      11.948   7.788  31.721  1.00 20.88
4MDH 348
ATOM     12  C   GLU A   2      12.042   9.235  32.169  1.00 28.31
4MDH 349
ATOM     13  O   GLU A   2      11.285   9.654  33.030  1.00 14.56
4MDH 350
ATOM     14  CB  GLU A   2      10.925   7.482  30.621  1.00 18.66
4MDH 351
ATOM     15  CG  GLU A   2      10.188   8.729  30.102  1.00 39.41
4MDH 352
```

```
ATOM     16  CD  GLU A   2       8.693   8.532  30.110  1.00 55.62
4MDH 353
ATOM     17  OE1 GLU A   2       7.885   9.153  29.379  1.00 55.67
4MDH 354
ATOM     18  OE2 GLU A   2       8.352   7.589  30.997  1.00 68.00
4MDH 355
```

(...)

As several enzymes are crystallised in presence of enzymatic cofactors or substrate analogues, that have to be described. As the number of substrates is too large to be described, a generic structure named HETATM regroups all atoms belonging to specific compounds other than amino-acids or nucleotides. In the following example NAD (nicotinamide adenine dinucleotide) and SO4 (sulphate) are described as HETATM. Solvent molecules (H2O) that are seen in the electronic density map also appear in this section.

```
HETATM 5158 AP   NAD B   1      42.641  30.361  41.284  1.00 26.73
4MDH5495
HETATM 5159 AO1  NAD B   1      43.440  31.570  40.868  1.00 20.69
4MDH5496
HETATM 5160 AO2  NAD B   1      41.161  30.484  41.376  1.00 33.73
4MDH5497
HETATM 5161 AO5* NAD B   1      43.117  29.802  42.683  1.00 20.55
4MDH5498
HETATM 5162 AC5* NAD B   1      44.483  29.615  43.002  1.00 17.23
4MDH5499
                                   (...)
HETATM 5202  S   SO4 B   2      44.842  24.424  31.662  1.00 72.77
4MDH5539
HETATM 5203  O1  SO4 B   2      45.916  23.890  32.631  1.00 31.43
4MDH5540
HETATM 5204  O2  SO4 B   2      44.065  23.296  30.916  1.00 26.35
4MDH5541
HETATM 5205  O3  SO4 B   2      45.570  25.307  30.620  1.00 52.53
4MDH5542
HETATM 5206  O4  SO4 B   2      43.834  25.257  32.482  1.00 47.91
4MDH5543
HETATM 5207  O   HOH     0      15.379   1.907   3.295  1.00 58.12
4MDH5544
HETATM 5208  O   HOH     1      58.861   0.984  17.024  1.00 37.58
4MDH5545
HETATM 5209  O   HOH     2      24.384   1.184  74.398  1.00 35.92
4MDH5546
```

(...)

HETATM fields describe only atoms positions, but as they concern non-standard groups, programs don't know which atoms are effectively connected. These information are found in the CONECT fields. In the example provided below, atom number 74 has to be connected to atoms 69 and 75. In absence of CONECT information, atoms are usually connected if they are closer than 2 angstroms.

```
CONECT   74   69   75
4MDH6015
CONECT   77   76
4MDH6016
CONECT   92   90   93
4MDH6017
CONECT   99   98
4MDH6018
```

(...)

12 Case Studies

Selected examples of the manipulation of Swiss-PdbViewer will be presented. Each example demonstrates problems that frequently arise. They are intended to focus the user attention on important details he should be aware of to take the maximum out of Swiss-Model and Swiss-PdbViewer.
Note that a complete user guide containing several other tutorials (in particular on the general manipulation of the program) is available from the Swiss-PdbViewer site:

http://www.expasy.ch/spdbv/mainpage.htm

Selected examples:

- How to superimpose two proteins
- Limitations of the Modelling (how good a model can be)
- How to do Standalone Modelling

Appendix A: Recommended reading

A.A Protein Struture

J-P. Doucet. *Computer-aided molecular design: theory and Applications.* Academic Press Harcourt brace & Company, Publishers (ISBN 0-12-221285-1).

Protein Structure Prediction: A practical Approach. Edited by Michael J.E. Sternberg. Oxford University Press (ISBN 0 19 9634973).

The Protein Folding Problem and Tertiary Structure Prediction. Edited by Kenneth M Merz, Jr and Scott M. Le Grand. Bikhäuser Verlag. (ISBN 3-7643-3693-5).

C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland, New York (1991).

A.V. Efimov. *A novel super-secondary structure of proteins and the relation between the structure and amino acid sequence*. FEBS Lett. **166**, 33-38 (1984).

A.V. Efimov. *Structure of* α-α*-hairpins with short connections*. Protein Engin.**4**, 245-250 (1991).

A.V. Efimov. *Structure of* β-β*-hairpins and* β-β*-corners*. FEBS Lett., **284**, 288-292 (1991).

R.H. Kretsinger. *Structure and evolution of calcium modulated proteins*. CRC Crit. Rev. Biochem. **8**, 119-174 (1980).

D.H. Ohlendorf, W.F. Anderson, M. Lewis, C.O. Pabo and B.W. Matthews. *Comparison of the structures of cro and lambda repressor proteins in bacteriophage lambda*. J. Mol. Biol. **169** 757-769 (1983).

B.L. Sibanda, T.L. Blundell and J.M. Thornton. *Conformation of* β*-hairpins in protein structures*. J.Mol. Biol. **206**, 759-777 (1989).

B.L. Sibanda and J.M. Thornton. B*-hairpin families in globular proteins*. Nature **316**, 170-174 (1985).

C. Chothia, M. Levitt and D. Richardson. *Structure of proteins: Packing of* α *helices and* β *sheets.* Proc. Natl. Acad. Sci. USA**74** 4130-4134 (1977).

C. Chothia. *Principles that determine the structure of proteins.* Ann. Rev. Biochem*.* **53** 537-572 (1984).

A.M. Lesk. *Protein architecture: a practical approach*. IRL press, Oxford (1991).

S.T. Rao and M.G. Rossman. *Comparison of super-secondary structures in proteins*. J. Mol. Biol. **76** 241-256 (1973).

J.S. Richardson. *The anatomy and taxonomy of protein structure* Adv. Prot. Chem. **34** 167-339 (1981).

A.B Protein modelling

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman. *Basic local alignment search tool*. J. Mol. Biol. **215**:403-410 (1990).

J. Bajorath, R. Stenkamp, A. Aruffo *Knowledge-based model building of proteins: Concepts and examples.* Prot. Sci. **2**:1798-1810 (1993).

B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus *CHARMM: A program for macromolecular energy, minimization and dynamics calculation.* J. Comp. Chem. **4**:187-217 (1983).

C. Chothia, A.M. Lesk. *The relation between the divergence of sequence and structure in proteins.* EMBO J. **5**:823-826 (1986).

Guex, N. and Peitsch, M. C. *SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling.* Electrophoresis **18**:2714-2723. (1997)

Guex N, Diemand A and Peitsch M. C. *Protein modelling for all*. TiBS **24**:364-367 (1999)

Guex N, Schwede T and Peitsch M. C. *Protein tertiary structure modelling*. Current Protocols in Protein Science. **Unit 2.8**: 2.8.1-2.8.17 (1999).

X. Huang, M. Miller. *A time-efficient, linear-space local similarity algorithm*. Adv. Appl. Math. **12**:337-357 (1991).

R. Lüthy, J.U. Bowie, D. Eisenberg. *Assessment of protein models with three-dimensional profiles*. Nature **356**:83-85 (1992).

W.R. Pearson, D.J. Lipman. *Improved tools for biological sequence comparison.* Proc. Natl. Acad. Sci. U.S.A. **85**:2444-2448 (1988).

M.C. Peitsch *Protein modelling by E-Mail*. Bio/Technology (Nature) **13**:658-660 (1995).

M.C. Peitsch. *ProMod and Swiss-Model: Internet-based tools for automated comparative protein* modelling. Biochem Soc Trans **24**:274-279 (1996).

Sippl, J.M. (1990) *Calculation of Conformational Ensembles from Potentials of Mean Force: an approach to the knowledge based prediction of local structures in globular proteins***.** J. Mol. Biol. **213**,859-883 (1990).

M.J. Sippl. *Recognition of errors in three-dimensional structres of proteins.* Proteins Struct. Funct. Genet. **17**:355-362 (1993).

Appendix B: A few useful URLs

Server	Internet address
ExPASy	http://www.expasy.ch
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
SWISS-MODEL Repository	http://www.expasy.ch/swissmod/SM_3DCrunch_Search.html
Swiss-PdbViewer	http://www.expasy.ch/spdbv/mainpage.html http://www.pdb.bnl.gov/expasy/spdbv/mainpage.html
Rasmol	http://www.umass.edu/microbio/rasmol
SWISS-3DIMAGE	http://www.expasy.ch/sw3d/sw3d-top.html
Brookhaven Protein Data Bank	http://www.pdb.bnl.gov
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	http://www.biochem.ucl.ac.uk/bsm/cath
PPS Internet Course	http://www.cryst.bbk.ac.uk/PPS