3

The three-dimensional structure of proteins

Amino acids linked together in a flat 'two-dimensional' representation of the polypeptide chain fail to convey the beautiful three-dimensional arrangement of proteins. It is the formation of regular secondary structure into complicated patterns of protein folding that ultimately leads to the characteristic functional properties of proteins.

Primary structure or sequence

The primary structure is the linear order of amino acid residues along the polypeptide chain. It arises from covalent linkage of individual amino acids via peptide bonds. Thus, asking the question 'What is the primary structure of a protein?' is simply another way of asking 'What is the amino acid sequence from the N to C terminals?' To read the primary sequence we simply translate the three or single letter codes from left to right, from amino to carboxyl terminals. Thus in the sequence below two alternative representations of the same part of the polypeptide chain are given, starting with alanine at residue 1, glutamate at position 2 and extending to threonine as the 12th residue (Figure 3.1).

Every protein is defined by a unique sequence of residues and all subsequent levels of organization (secondary, super secondary, tertiary and quaternary) rely on this primary level of structure. Some proteins are related to one another leading to varying degrees of similarity in primary sequences. So myoglobin, an oxygen storage protein found in a wide range of organisms, shows similarities in human and whales in the 153 residue sequence (Figure 3.2). Most of the sequence is identical and it is easier to spot the differences. When a change occurs in the primary sequence it frequently involves two closely related residues. For example, at position 118 the human variant has a lysine residue whilst whale myoglobin has an arginine residue. Reference to Table 2.2 will show that arginine and lysine are amino acids that contain a positively charged side chain and this change is called a conservative transition. In contrast in a few positions there are very different amino acid residues. Consider position 145 where asparagine (N) is replaced by lysine (K). This transition is not conservative; the small, polar, side chain of asparagine is replaced by the larger, charged, lysine. Regions, or residues, that never change are called invariant.

Secondary structure

Primary structure leads to secondary structure; the local conformation of the polypeptide chain or the spatial relationship of amino acid residues that are close together in the primary sequence. In globular proteins

Proteins: Structure and Function by David Whitford

NH_3-Ala-Glu-Glu-Ser-Ser-Lys-Ala-Val-Lys-Tyr-Tyr-Thr-.....

NH_3---A---E---E---S----S----K----A---V---K---Y----Y----T.....

Figure 3.1 Single- and three-letter codes for amino terminal of a primary sequence

the three basic units of secondary structure are the α helix, the β strand and turns. All other structures represent variations on one of these basic themes. This chapter will focus on the chemical and physical properties of polypeptides that permit the transition from randomly oriented polymers of amino acid residues to regular repeating secondary structure. With 20 different amino acid residues found in proteins there are 780 possible permutations for a dipeptide. In an average size polypeptide of ~100 residues the number of potential sequences is astronomical. In view of the enormous number of possible conformations many fundamental studies of protein secondary structure have been performed on homopolymers of amino acids for example poly-alanine, poly-glutamate, poly-proline or poly-lysine.

Homopolymers have the advantage that either all the residues are identical or in some cases simple repeating units such as poly Ala-Gly, where Ala-Gly dipeptides are repeated along the length of the polymer. In comparison with polypeptides derived from proteins these polymers have the advantage of consistent conformations. This is not to imply that the conformation of homopolymers is regular and ordered. In some instances these polymers are unstructured, however, the common theme is that the polymers are uniform in their conformations and hence are much more attractive candidates for initial studies of structure.

Studies of homopolymers have been advanced by host–guest studies where a polymer of alanine residues is modified by an introduction of a single different residue in the middle of the polypeptide. By measuring changes in stability, solubility, or helical properties these studies allow the effect of the new or guest residue to be accurately defined. Indeed, much of the data reported in tables throughout this book were obtained by the introduction of a single amino acid residue into a polymer containing just one type of residue. The host peptide is usually designed to be monomeric (i.e. non-aggregating), to be soluble and not more than 15 residues in length. By systematically replacing residue 8 with any of the other 19 amino acids these studies have elucidated many of the properties of residues within polypeptide chains. One property that has been extensively studied using this approach is the relative helical tendency of amino acid residues in a peptide of poly-alanine. Results suggest, unsurprisingly, that alanine is the most stable residue to substitute into a poly-alanine peptide whilst proline is the most destabilizing. More revealing is the relative

```
                    *        20         *        40
Human Mb : GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETL :  40
Whale Mb : VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETL :  40

                    *        60         *        80
Human Mb : EKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKG :  80
Whale Mb : EKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKG :  80

                    *       100         *       120
Human Mb : HHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP : 120
Whale Mb : HHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP : 120

                    *       140         *
Human Mb : GDFGADAQGAMNKALELFRKDMASNYKELGFQG : 153
Whale Mb : GDFGADAQGAMNKALELFRKDIAAKYKELGYQG : 153
```

Figure 3.2 The primary sequences of human and sperm whale myoglobin. Identical residues are shown in blue, the regions in yellow show conserved substitutions whilst the red regions show non-conservative changes. Astericks indicate every tenth residue. Single letter codes for residues are used

Table 3.1 The helical propensity of amino acid residues substituted into alanine polymers

Residue	Helix propensity, ΔG (kJ mol^{-1})	Residue	Helix propensity, ΔG (kJ mol^{-1})
Ala	0	Ile	0.41
Arg	0.21	Leu	0.21
Asn	0.65	Lys	0.26
Asp^0	0.43	Met	0.24
Asp^-	0.69	Phe	0.54
Cys	0.68	Pro	3.16
Gln	0.39	Ser	0.50
Glu^0	0.16	Thr	0.66
Glu^-	0.40	Tyr	0.53
Gly	1.00	Trp	0.49
His^0	0.56	Val	0.61
His^+	0.66		

Derived from Pace, C.N. & Scholtz, J.M. *Biophys. J.* 1998, 75, 422–427. The data includes uncharged Glu, Asp and His residues (superscript ^{0}). All residues form helices with less propensity than poly-Ala hence the positive values for ΔG.

helical tendencies or propensities of the other residues when measured relative to alanine (Table 3.1).

When the first crystal structures of proteins became available it allowed a comparison between residue identity and secondary structure. Ala, Leu and Glu are found more frequently in α helices whilst Pro, Gly and Asp were found less frequently than average. Using this analysis of primary sequence a helix propensity scale was derived, and is still used in predicting the occurrence of helices and sheets in folded soluble proteins.

The α helix

The right-handed α helix is probably the best known and most identifiable unit of secondary structure. The structure of the α helix was derived from model-building studies until the publication of the crystallographic structure of myoglobin. This demonstrated that α helices occurred in proteins and were largely as predicted from theoretical studies by Linus Pauling. The α helix is the most common structural motif found in proteins; in globular proteins over 30 percent of all residues are found in helices.

The regular α helix (Figure 3.3) has 3.6 residues per turn with each residue offset from the preceding residue by 0.15 nm. This parameter is called the translation per residue distance. With a translation distance of 0.15 nm and 3.6 residues per turn the pitch of the α helix is simply 0.54 nm (i.e. 3.6×0.15 nm). The pitch is the translation distance between any two corresponding atoms on the helix. One of the major results of model building studies was the realization that the α helix arises from regular values adopted for ϕ (phi) and ψ (psi), the torsion or dihedral angles (Figure 3.4).

The values of ϕ and ψ formed in the α helix allow the backbone atoms to pack close together with few unfavourable contacts. More importantly this arrangement allows some of the backbone atoms to form hydrogen bonds. The hydrogen bonds occur between the backbone carbonyl oxygen (acceptor) of one residue and the amide hydrogen (donor) of a residue four ahead in the polypeptide chain. The hydrogen bonds are 0.286 nm long from oxygen to nitrogen atoms, linear and lie (in a regular helix) parallel to the helical axis. It is worth noting that in 'real' proteins the arrangement of hydrogen bonds shows variation in length and angle with respect to helix axes (Figure 3.5).

Hydrogen bonds have directionality that reflects the intrinsic polarization of the hydrogen bond due to the electronegative oxygen atom. In a similar fashion the peptide bond also has polarity and the combined effect of these two parameters give α helices pronounced dipole moments. On average the amino end of the α helix is positive whilst the carboxyl end is negative. In the α helix, the first four NH groups and last four CO groups will normally lack backbone hydrogen bonds. For this reason very short helices often have distorted conformations and form alternative hydrogen bond partners. The distortion of hydrogen bonds and lengths that occur in real helices are accompanied by the dihedral angles (ϕ and ψ) that deviate significantly from the ideal values of $-57°$ and $-47°$ (see Table 3.2).

Visualization of the α helix frequently neglects the side chains but an ideal arrangement involves the

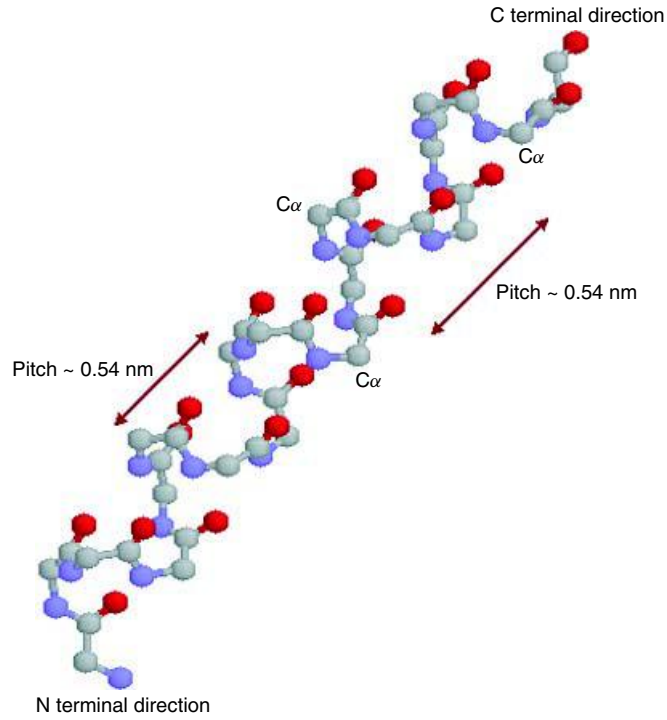


Figure 3.3 A regular α helix. Only heavy atoms (C, N and O, but not hydrogen) are shown and the side chains are omitted for clarity

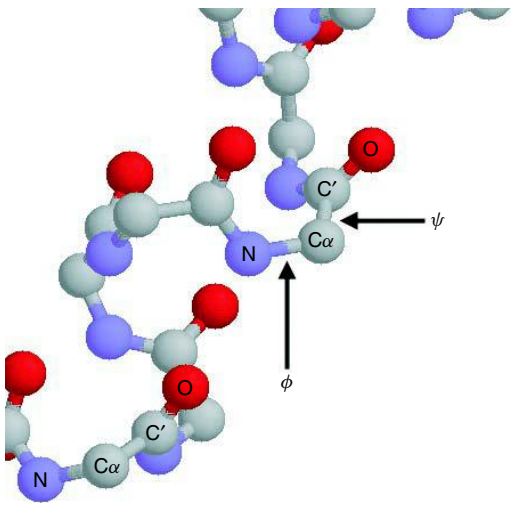


Figure 3.4 ϕ is defined by the angle between the C′–N–Cα –C′ atoms whilst ψ is defined by the atoms N–Cα –C′–N

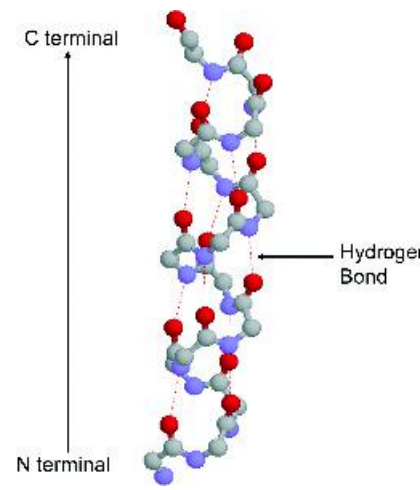


Figure 3.5 Arrangement of backbone hydrogen bonds in a *real* α helix from myoglobin, shows deviations from ideal geometry

Table 3.2 Dihedral angles, translation distances and number of residues per turn for regular secondary structure conformations. In poly(Pro) I ω is 0° whilst in poly(Pro) II ω is 180°

Secondary structure element	Dihedral angle (°)		Residues/turn	Translation distance per residue (nm)
	ϕ	ψ		
α helix	−57	−47	3.6	0.150
3_{10} helix	−49	−26	3.0	0.200
π helix	−57	−70	4.4	0.115
Parallel β strand	−139	+135	2.0	0.320
Antiparallel β strand	−119	+113	2.0	0.340
Poly(Pro) I	−83	+158	3.3	0.190
Poly(Pro) II	−78	+149	3.0	0.312

atoms projecting outwards into solution. Although the side chains radiate outwards there are conformational restrictions because of potential overlap with atoms in neighbouring residues. This frequently applies to branched side chains such as valine, isoleucine and threonine where the branch occurs at the CB atom and is closest to the helix. Steric restriction about the CA–CB bond leads to discrete populations or rotamers (Figure 3.6). The symbol χ_1 (pronounced "KI one") is used to define this angle and is best appreciated by viewing projections along the CA–CB bond.

Alanine, glycine and proline do not have χ_1 angles whilst the χ_2 angle for serine, threonine and cysteine is difficult to measure because it involves determining the position of a single H atom accurately. However, in databases of protein structures the χ_1 angles adopted for all residues (except Ala, Gly and Pro) have been documented, whilst the χ_2 distribution of Arg, Glu, Gln, Ile, Leu, Lys and Met are also well known. This has lead to the idea of rotamer libraries that reflect the most probable side chain conformations in elements of secondary structure (see Table 3.3).

Proline does not form helical structure for the obvious reason that the absence of an amide proton (NH) precludes hydrogen bonding whilst the side chain covalently bonded to the N atom restricts backbone rotation. The result is that proline often locates at the beginning of helices or in turns between two α helical units. Occasionally proline is found in a long helical region but invariably a major effect is to distort the helix causing a kink or change of direction of the polypeptide backbone.

Table 3.3 The range or distribution of χ_1 bond angles in different rotamer populations

Rotamer	χ_1 angle
g^+ (gauche $^+$)	$-120° - 0°$
trans	$120° - 240°$
g^- (gauche $^-$)	$0° - 120°$

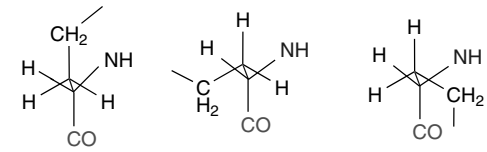


Figure 3.6 The CA–CB bond and different rotamer populations. Instead of looking directly along the CA–CB bond the bond is offset slightly to aid viewing. The front face represents the 'backbone' portion of the molecule. A clockwise rotation is defined as positive (+) whilst a counter clockwise rotation is negative (−) and leads to angles between methylene and CO group of ∼180°, −60° and +60°

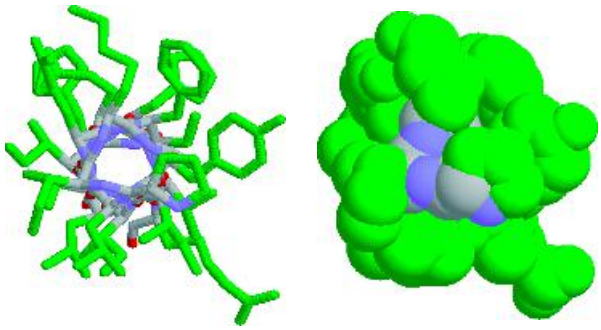


Figure 3.7 Wireframe and space-filling representations of one end of a regular α helix

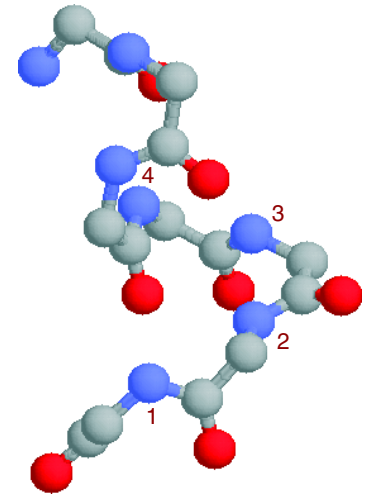


Figure 3.8 A 3_{10} helix shows many similarities to a regular α helix. Closer inspection of the hydrogen bond donor and acceptor reveals i, $i+3$ connectivity, three residues per turn and dihedral angle values of ~−50 and −25 reflect a more tightly coiled helix (see Table 3.2)

The α helix is frequently portrayed in textbooks including this one with a 'ball and stick' or 'wireframe' representation (Figure 3.7). A better picture is provided by 'space-filling' representations (Figure 3.7) where atoms are shown with their van der Waals radii. In the 'end-on' view of the α helix the wireframe representation suggests a hollow α helix. In contrast, the space filling representation emphasizes that little space exists anywhere along the helix backbone. Other representations of helices include cylinders showing the length and orientation of each helix or a ribbon representation that threads through the polypeptide chain. The preceding section views the helix as a stable structure but experiments with synthetic poly-amino acids suggest that very few polymers fold into regular helical conformation.

Other helical conformations

The 3_{10} helix is a structural variation of the α helix found in proteins (Figure 3.8). The 3_{10} helices are often found in proteins when a regular α helix is distorted by the presence of unfavourable residues, near a turn region or when short sequences fold into helical conformation. In the 3_{10} helix the dominant hydrogen bonds are formed between residues $i, i+3$ in contrast to $(i, i+4)$ bonds seen in the regular α helix. The designation 3_{10} refers to the number of backbone atoms located between the donor and acceptor atoms (10) and the fact that there are three residues per turn. With three residues per turn the 3_{10} helix is a tighter, narrower structure in which the potential for unfavourable contacts between backbone or side chain atoms is increased.

Whilst the 3_{10} helix is a narrower structure than the α helix a third possibility is a more loosely coiled helix with hydrogen bonds formed between the CO and NH groups separated by five residues $(i, i+5)$. This structure is the π helix and at one stage it was thought not to occur naturally. However, examples of this helix structure have been described in proteins. Soyabean lipoxygenase has a 43-residue helix containing regions essential to enzyme function and stability running through the centre of the molecule. Three turns of an expanded helix with eight $(i, i+5)$ hydrogen bonds and 4.4 residues per turn make it an example of a π helix containing more than one turn found in a protein (Figure 3.9).

The rarity of this form of secondary structure arises for a number of reasons. One major limitation is that the ϕ/ψ angles of a π helix lie at the edge of the allowed, minimum energy, region of the Ramachandran

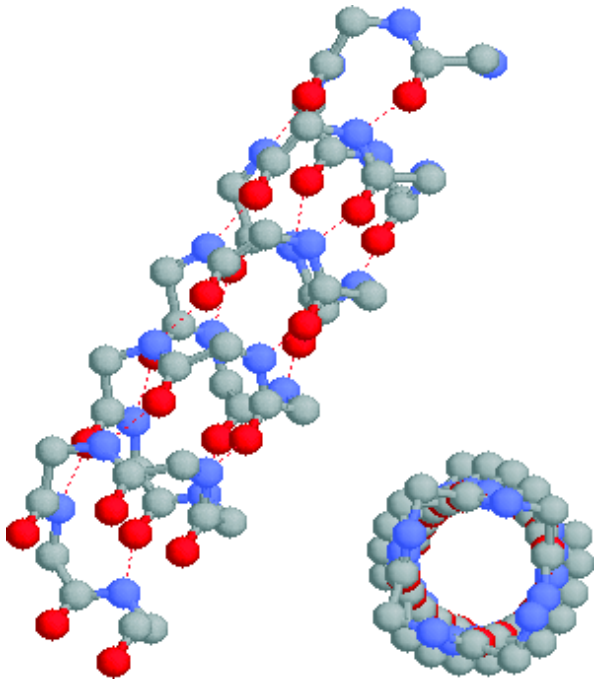


Figure 3.9 The π helix showing i, $i+5$ hydrogen bonds and the end on view of the helix

plot. The large radius of the π helix means that backbone atoms do not make van der Waals contact across the helix axis leading to the formation of a hole down the middle of the helix that is too small for solvent occupation.

The β strand

The β strand, so called because it was the second unit of secondary structure predicted from the model-building studies of Pauling and Corey, is an extended conformation when compared with the α helix. Despite its name the β strand is a helical arrangement although an extremely elongated form with two residues per turn and a translation distance of 0.34 nm between similar atoms in neighbouring residues. Although less easy to recognize, this leads to a pitch or repeat distance of nearly 0.7 nm in a regular β strand (Figure 3.10). A single β strand is not stable largely because of the limited number of local stabilizing interactions. However, when two or more β strands form additional hydrogen bonding interactions a stable sheet-like arrangement is created (Figure 3.11). These β sheets result in significant increases in overall stability and

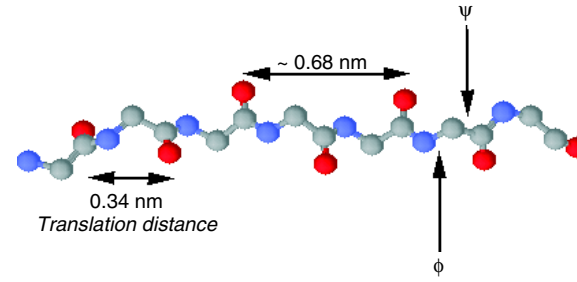


Figure 3.10 The polypeptide backbone of a single β strand showing only the heavy atoms

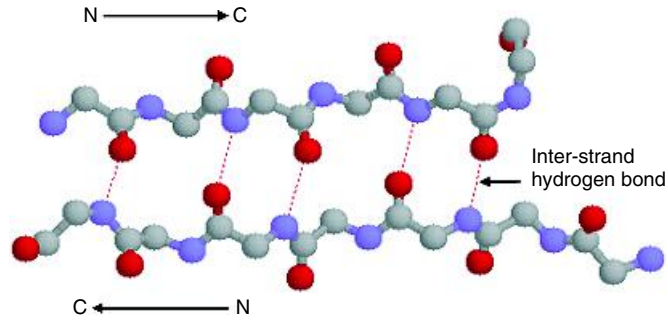


Figure 3.11 Two adjacent β strands are hydrogen bonded to form a small element of β sheet. The hydrogen bonds are inter-strand between neighbouring CO and NH groups. Only the heavy atoms are shown in this diagram for clarity

are stabilized by the formation of backbone hydrogen bonds between adjacent strands that may involve residues widely separated in the primary sequence.

Adjacent strands can align in parallel or antiparallel arrangements with the orientation established by determining the direction of the polypeptide chain from the N- to the C-terminal. 'Cartoon' representations of β strands make establishing directions in molecular structures easy since β strands are often shown as arrows; the arrowheads indicate the direction towards the C terminal. These cartoons take different forms but all 'trace' the arrangement of the polypeptide backbone and summarize secondary structural elements found in proteins without the need to show large numbers of atoms.

Polyamino acid chains do not form β sheets when dispersed in solution and this has hindered study of the formation of such structures. However, despite this observation many proteins are based predominantly on β strands, with chymotrypsin a proteolytic enzyme being one example (see Figure 3.12).

Unlike ideal representations strands found in proteins are often distorted by twisting that arises from a systematic variation of dihedral angles (ϕ and ψ) towards more positive values. The result is a slight, but discernable, right hand twists in the polypeptide chain. In addition when strands hydrogen bond together to form sheets further distortions occur especially with mixtures of anti-parallel and parallel β strands. On average β sheets containing antiparallel strands are more common than sheets made up entirely of parallel strands. Anti-parallel sheets often form from just two β strands running in opposite directions whilst it is observed that at least four β strands are required to form parallel sheets. β strands associate spectacularly into extensive curved sheets known as β barrels. The β barrel is found in many proteins (Figure 3.13) and consists of eight parallel β strands linked together by helical segments.

Turns as elements of secondary structure

As more high-resolution structures are deposited in protein databases it has allowed turns from different proteins to be defined and compared in terms of

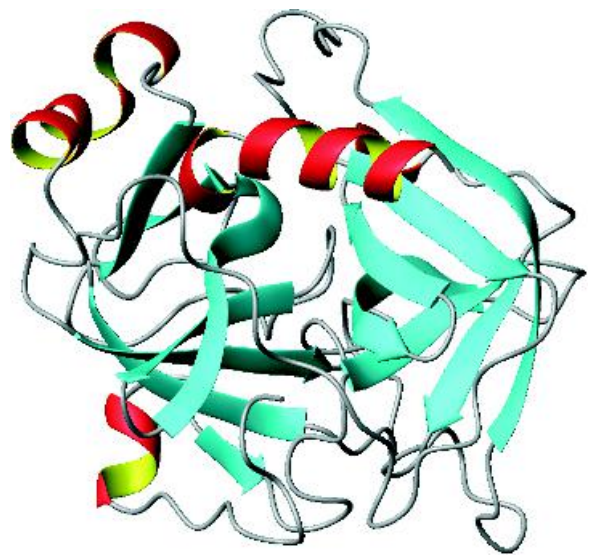


Figure 3.12 Representation of the elements of β strands and α helices found in the serine protease chymotrypsin. (PDB:2CGA). The strands shown in cyan have arrows indicating a direction leading from the N > C terminus, the helices are shown in red and yellow with turns in grey. There is no convention describing the use of colours to a particular element of secondary structure

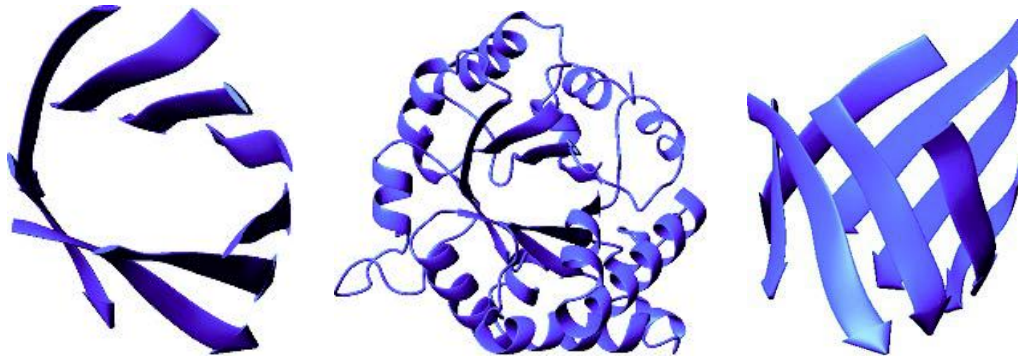


Figure 3.13 The β barrel of triose phosphate isomerase seen in three different views. The eight β strands encompassed by helices are shown in the centre with two different views of the eight parallel β strands shown in the absence of helical elements. Eight strands are arranged at an angle of approximately 36° to the barrel axis (running from top to bottom in the right most picture) with each strand offset from the previous one by a constant amount

residue composition, angles and bond distances. In some proteins the proportion of residues found in turns can exceed 30 percent and in view of this high value it is unlikely that turns represent random structures. Turns have the universal role of enabling the polypeptide to change direction and in some cases to reverse back on itself. The reverse turns or bends arise from the geometric properties associated with these elements of protein structure.

Analysis of the amino acid composition of turns reveals that bulky or branched side chains occur at very low frequencies. Instead, residues with small side chains such as glycine, aspartate, asparagine, serine, cysteine and proline are found preferentially. An analysis of the different types of turns has established that perhaps as many as 10 different conformations exist on the basis of the number of residues making up the turn and the angles ϕ and ψ associated with the central residues. Turns can generally be classified according to the number of residues they contain with the most common number being three or four residues.

A γ turn contains three residues and frequently links adjacent strands of antiparallel β sheet (Figure 3.14). The γ turn is characterized by the residue in the middle of the turn ($i + 1$) not participating in hydrogen bonding whilst the first and third residues can form the final and initial hydrogen bonds of the antiparallel β strands. The change in direction of the polypeptide chain caused by a γ turn is reflected in the values of ϕ and ψ for the central residue. As a result of its size and conformational flexibility glycine is a favoured residue in this position although others are found.

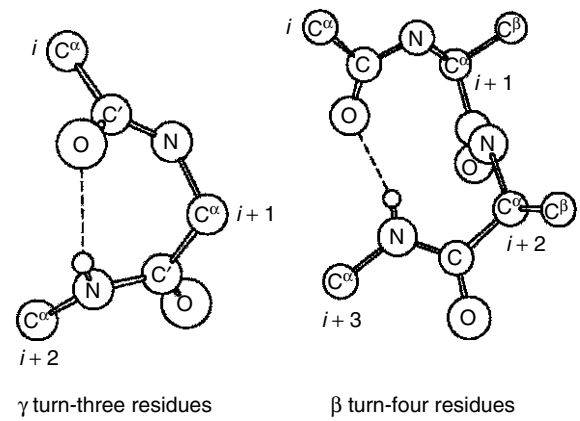


Figure 3.14 Arrangement of atoms in γ and β turns connected to strands of an antiparallel β sheet. γ turns contain three residues whilst β turns have four residues. A large number of variations on this basic theme exist in proteins. Hydrogen bonds are shown by a dotted line whilst only heavy atoms of the polypeptide backbone together with the CB atom are shown

More commonly found in protein structures are four residue turns (β turns). Here the middle two residues ($i + 1, i + 2$) are never involved in hydrogen bonding whilst residues i and $i + 3$ will participate in hydrogen

bonding if a favourable arrangement forms between donor and acceptor. Analysis of structures deposited in protein databases reveals strong residue preferences. In the relatively common type 1β turn any residue can occur at position i, $i+3$ with the exception that Pro is never found. More significantly glycine is often found at position $i+3$ whilst proline predominates at $i+1$. Asn, Asp, Cys and Ser are also found frequently in β-turns as the first residue.

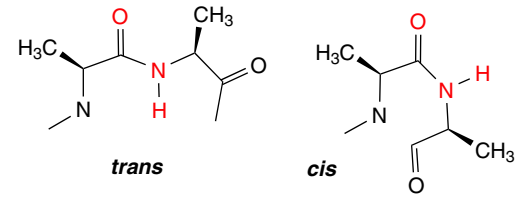


Figure 3.15 *Cis* and *trans* peptide bonds

Additional secondary structure

Both glycine and proline are associated with unusual conformational flexibility when compared with the remaining 18 residues. In the case of glycine there is very little restriction to either ϕ/ψ as a result of the small side chain (H). In the case of proline the opposite situation applies with ϕ, the angle defined by the N–Cα bond, restricted by the five-membered cyclic (pyrrolidine) ring. Proline residues are not suited to either helical or strand arrangements but are found at high frequency in turns or bends.

Additionally polyproline chains adopt unique and regular conformations distinct from helices, turns or strands. Two recognized conformations are called poly(Pro) I and poly(Pro) II. Proline is unique amongst the twenty residues in showing a much higher proportion of *cis* peptide bonds. The two forms of poly(Pro) therefore contain all *cis* (I) or all *trans* (II) peptide bonds. The value of ϕ is restricted by bond geometry to -83° (I) and -78° (II) and the torsion angle restrictions create poly(Pro) I as a right-handed helix with 3.3 residues per turn, whilst poly(Pro) II is a left-handed helix with three residues per turn. Poly(Gly) chains also adopt regular conformation and presents the opposite situation to poly(Pro) chains. Glycine has extreme conformational flexibility. In the solid state poly(Gly) has been shown to adopt two regular conformations designated I and II. State I has an extended conformation like a β strand whilst state II has three residues per turn and is similar to poly(Pro).

The Ramachandran plot

The peptide bond is planar as a result of resonance and its bond angle, **ω**, has a value of 0 or 180°. A peptide bond in the *trans* (Figure 3.15) conformation ($\omega = 180^\circ$) is favoured over the *cis* (Figure 3.15) arrangement ($\omega = 0^\circ$) by a factor of ~1000 because the preferential arrangement of non-bonded atoms leads to fewer repulsive interactions that otherwise decrease stability. In the *cis* peptide bond these non-bonded interactions increase due to the close proximity of side chains and C_α atoms with the preceding residue and hence results in decreased stability relative to the *trans* state. Peptide bonds preceding proline are an exception to this trend with a *trans/cis* ratio of approximately 4.

The peptide bond is relatively rigid, but far greater motion is possible about the remaining backbone torsion angles. In the polypeptide backbone C–N–C_α–C defines the torsion angle ϕ whilst N–C_α–C–N defines ψ. In practice these angles are limited by unfavourable close contacts with neighbouring atoms and these steric constraints limit the conformational space that is sampled by polypeptide chains. The allowed values for ϕ and ψ were first determined by G.N. Ramachandran using a 'hard sphere model' for the atoms and these values are indicated on a two-dimensional plot of ϕ against ψ that is now called a Ramachandran plot (Figure 3.16).

In the Ramachandran plot shown in Figure 3.16 the freely available conformational space is shaded in green. This represents ideal geometry and is exhibited by regular strands or helices. Analysis of crystal structures determined to a resolution of <2.5 Å showed that over 80 percent of all residues are found in this region of the Ramachandran plot. The yellow region indicates areas that although less favourable can be formed with small deviations from the ideal angular values for ϕ or ψ. The yellow and green regions include 95 percent of all residues within a protein. Finally, the purple coloured region, although much less favourable,

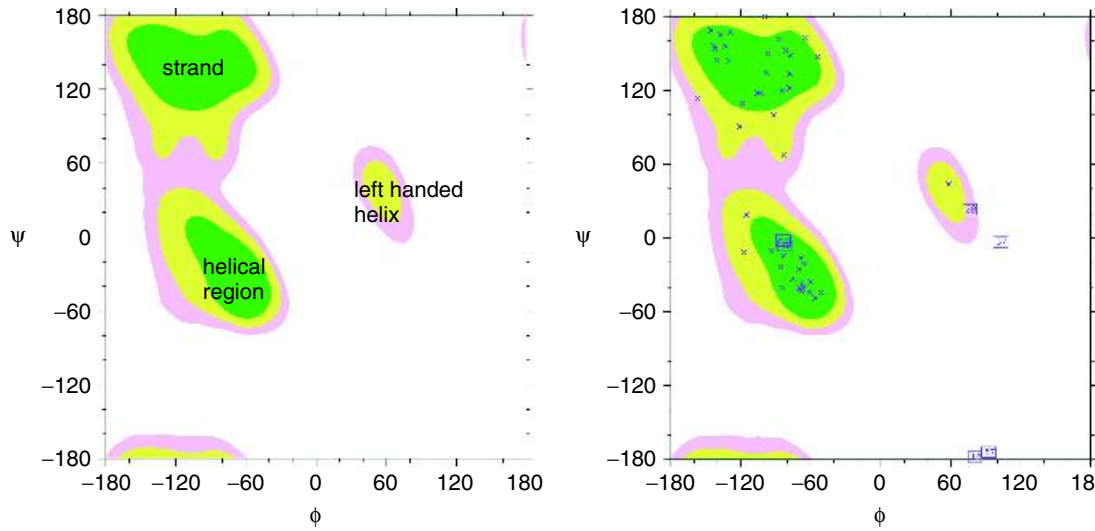


Figure 3.16 Ramachandran plots showing favourable conformational parameters for different ϕ/ψ values. Left: Ramachandran plot showing the ϕ/ψ angles exhibited by regular α helix and β strands. In addition the left-handed helix has a region of limited stability and although not widely shown in protein structures isolated residues do adopt this conformation. Right: Ramachandran plot derived from the crystal structure of bovine pancreatic trypsin inhibitor PDB: 1BPI. The residue numbers for glycine are shown. In BPTI some of these residues do not exhibit typical ϕ/ψ angles

will account for 98 percent of all residues in proteins. All other regions are effectively disallowed with the minor exception of a small region representing left-handed helical structure. In total only 30 percent of the total conformational space is available suggesting that the polypeptide chain itself imposes severe restrictions.

One exception to this rule is glycine. Glycine lacks a C_β atom and with just two hydrogen atoms attached to the C_α centre this residue is able to sample a far greater proportion of the space represented in the Ramachandran plot (Figure 3.17). For glycine this leads to a symmetric appearance for the allowed regions. As expected residues with large side chains are more likely to exhibit unfavourable, non-bonded, interactions that limit the possible values of ϕ and ψ. In the Ramachandran plot the allowed regions are smaller for residues with large side chains such as phenylalanine, tryptophan, isoleucine and leucine when compared with, for example, the allowed regions for alanine.

Similar plots of χ_1 versus χ_2 describe their distribution in proteins and it was found that the distribution

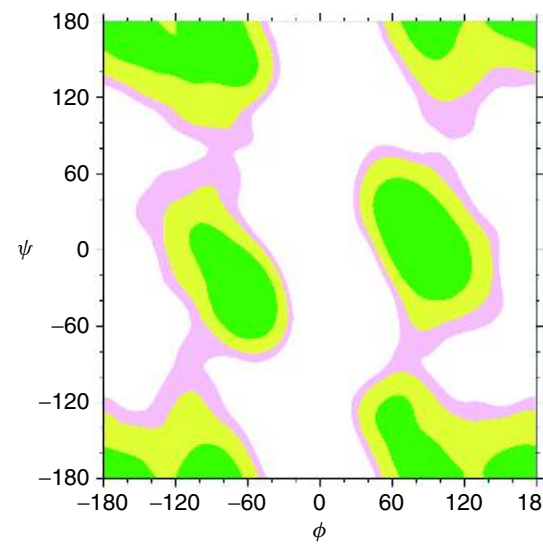


Figure 3.17 Ramachandran plot showing conformational space 'sampled' by glycine residues

of χ_1 and χ_2 side-chain torsion angles followed simple energy-based calculations with preferences for values of approximately $+60^\circ$, 180° and -60° for χ_1 and χ_2 in aliphatic side chains, and $+90^\circ$ and -90° for the χ_2 torsion angle of aromatic residues. Inspection of χ_1/χ_2 distribution reveals that several residues display preferences for certain combinations of torsion angles. Leucine residues prefer the combinations $-60^\circ/180^\circ$ and $180^\circ/+60^\circ$. This approach has led to the derivation of a library of preferred, but not obligatory, side-chain rotamers. These libraries are useful in the refinement of protein structures and include residue-specific preferences for combinations of side-chain torsion angles. Valine, for example, shows the greatest preference for one rotamer (χ_1 is predominantly t) and is unique in this respect amongst the side chains.

Tertiary structure

At the beginning of 2004 ~22 000 sets of atomic coordinates were deposited in databases such as the Protein Data Bank (PDB) and this value increases daily with the deposition of new structures from increasingly diverse organisms (see Figure 3.18). This database is currently maintained at Rutgers University with several mirrors (identical sites) found at other sites across the world (http://www.rcsb.org/pdb). Whilst some of these proteins are duplicated or related (there are over 50 structures of T4 lysozyme and over 200 globin structures) the majority represent the determination of novel structures using mainly X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. Over 1000 different protein folds have been discovered to date with undoubtedly more to follow.

PDB files contain information of the positions in space of the vast majority of atoms making up a protein. By identifying the x, y and z coordinate of each atom we define the whole molecule. However, besides the x, y and z coordinates of atoms PDB files also contain header information describing the primary sequence, the method used to determine the structure, the organism from which this protein was derived, the elements of secondary structure, any post-translational modifications as well as the authors. As a consequence of the great development in bioinformatics PDB files are undergoing homologization to enable easy comparison between structures and to permit widespread use with different computer software packages. Figure 3.19 shows a representative PDB file.

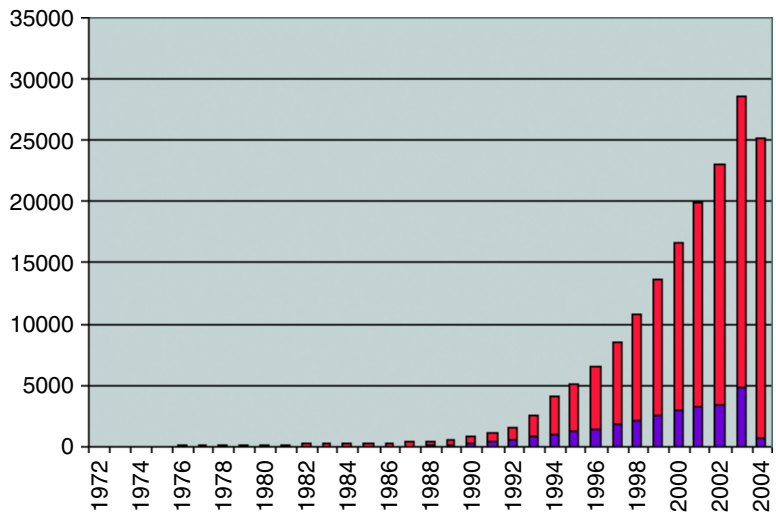


Figure 3.18 Data produced by the Protein Data Bank on the deposition of structures. From Berman, H.M. *et al.* The Protein Data Bank. *Nucl. Acids Res.* 2000 **28**, 235–242. See http://www.rcsb.org/pdb for latest details. Red blocks indicate total structures, purple blocks the number of submissions per year

```
HEADER    HYDROLASE(ZYMOGEN)                      16-JAN-87   2CGA      2CGA   3
COMPND    CHYMOTRYPSINOGEN *A                                           2CGA   4
SOURCE    BOVINE (BOS $TAURUS) PANCREAS                                 2CGA   5
AUTHOR    D.WANG,W.BODE,R.HUBER                                         2CGA   6
>>>>>>
>>>>>>
JRNL        AUTH   D.WANG,W.BODE,R.HUBER                               2CGA   8
JRNL        TITL   BOVINE CHYMOTRYPSINOGEN *A. X-RAY CRYSTAL STRUCTURE  2CGA   9
>>>>>>
REMARK   1                                                              2CGA  14
REMARK   1 REFERENCE 1                                                  2CGA  15
>>>>>>
SEQRES   1 A  245  CYS GLY VAL PRO ALA ILE GLN PRO VAL LEU SER GLY LEU  2CGA  71
SEQRES   2 A  245  SER ARG ILE VAL ASN GLY GLU GLU ALA VAL PRO GLY SER  2CGA  72
SEQRES   3 A  245  TRP PRO TRP GLN VAL SER LEU GLN ASP LYS THR GLY PHE  2CGA  73
>>>>>>
>>>>>>
SEQRES  17 B  245  LEU VAL GLY ILE VAL SER TRP GLY SER SER THR CYS SER  2CGA 106
SEQRES  18 B  245  THR SER THR PRO GLY VAL TYR ALA ARG VAL THR ALA LEU  2CGA 107
SEQRES  19 B  245  VAL ASN TRP VAL GLN GLN THR LEU ALA ALA ASN          2CGA 108
>>>>>>
CRYST1   59.300   77.100  110.100  90.00  90.00  90.00 P 21 21 21    8  2CGA 113
ORIGX1      1.000000  0.000000  0.000000        0.00000                 2CGA 114
ORIGX2      0.000000  1.000000  0.000000        0.00000                 2CGA 115
ORIGX3      0.000000  0.000000  1.000000        0.00000                 2CGA 116
SCALE1       .016863  0.000000  0.000000        0.00000                 2CGA 117
SCALE2      0.000000   .012970  0.000000        0.00000                 2CGA 118
SCALE3      0.000000  0.000000   .009083        0.00000                 2CGA 119
MTRIX1   1   .987700   .155000   .017700        6.21700    1            2CGA 120
MTRIX2   1   .022800  -.031400  -.999200      115.61600    1            2CGA 121
MTRIX3   1  -.154300   .987400  -.034600       -3.74800    1            2CGA 122
ATOM      1  N   CYS A   1     -10.656  55.938  41.808  1.00 11.66       2CGA 123
ATOM      2  CA  CYS A   1     -10.044  57.246  41.343  1.00 11.66       2CGA 124
ATOM      3  C   CYS A   1     -10.076  58.323  42.431  1.00 11.66       2CGA 125
ATOM      4  O   CYS A   1     -10.772  58.097  43.448  1.00 11.66       2CGA 126
ATOM      5  CB  CYS A   1     -10.807  57.718  40.066  1.00 11.66       2CGA 127
>>>>>>
>>>>>>
>>>>>>
ATOM    744  N   ASN A 100     -13.152  77.724  22.378  1.00  8.65       2CGA 866
ATOM    745  CA  ASN A 100     -14.213  76.940  23.011  1.00  8.65       2CGA 867
ATOM    746  C   ASN A 100     -14.134  75.441  22.693  1.00  8.65       2CGA 868
ATOM    747  O   ASN A 100     -13.706  75.062  21.563  1.00  8.65       2CGA 869
>>>>>>
>>>>>>
ATOM   1461  N   VAL A 200      -9.212  70.793  39.923  1.00  9.30       2CGA1583
ATOM   1462  CA  VAL A 200      -9.875  69.689  40.639  1.00  9.30       2CGA1584
ATOM   1463  C   VAL A 200     -10.634  70.148  41.868  1.00  9.30       2CGA1585
ATOM   1464  O   VAL A 200     -10.151  70.985  42.657  1.00  9.30       2CGA1586
>>>>>>
HETATM 3601  O   HOH   601     -20.008  66.224  26.138  1.00 26.69       2CGA3723
HETATM 3602  O   HOH   602     -21.333  66.182  28.756  1.00 18.10       2CGA3724
HETATM 3603  O   HOH   603     -18.000  68.022  22.774  1.00 34.03       2CGA3725
MASTER       60    3    0    0    0    0    0    9 3927    2    0   38  2CGAA  6
END                                                                     2CGA4053
```

Figure 3.19 An abbreviated version of a representative PDB file. The file 2CGA refers to the chymotrypsin. In the above example the structure was determined by X-ray crystallography and the initial lines (header and remarks) of a PDB file give the authors, important citations, source of the protein and other useful information. Later the primary sequence is described and this is followed by crystallographic data on the unit cell dimensions and space group. The important lines are those beginning with ATOM since collectively these lines list all heavy atoms of the protein together with their respective *x*, *y* and *z* coordinates. A line beginning HETATM lists the position of hetero atoms which might include co-factors but more frequently lists water molecules found in the crystal structure. The symbol ≫≫≫≫≫≫≫≫≫ is used here to denote the omission of many lines of text. Where more than one chain exists the coordinates will be listed as chain A, chain B, etc.

Ten years ago the representation of tertiary structures on flat pages of a book represented a major problem as well as creating conceptual difficulty for students. Although tertiary structures remain complex to visualize the widespread development of desktop computers capable of handling these relatively large files coupled with graphical software to represent the structures of proteins in different fashions has revolutionized this area. Many of the images of proteins shown in this book can be viewed as true 'three-dimensional' structures by consulting the on-line version of this book or by downloading the PDB file from one of the many available databases and using this file in a suitable molecular graphics package. An alternative sometimes used is to represent molecules as stereo images that can be viewed with either special glasses or more easily on computers equipped to display such structures again using special glasses. In this book stereo images have not been widely used purely for simplicity but students should endeavour to view such presentations because of their ability to convey space, folding and depth in highly complex structures.

Numerous software packages have been developed for viewing PDB files and each has its own advantages and disadvantages as well as its own supporters and detractors within the scientific community. Many of these packages are public domain software whilst others have been developed as commercial entities. However, there is little doubt that as a result of this important, yet often under-rated, development it has become easier to portray tertiary structures conveying their complexity and beauty.

Detailed tertiary structure

The tertiary structure represents the folded polypeptide chain. It is defined as the spatial arrangement of amino acid residues that are widely separated in the primary sequence or more succinctly as the overall topology formed by the polypeptide (see Figure 3.20). For small globular proteins of 150 residues or less the folded

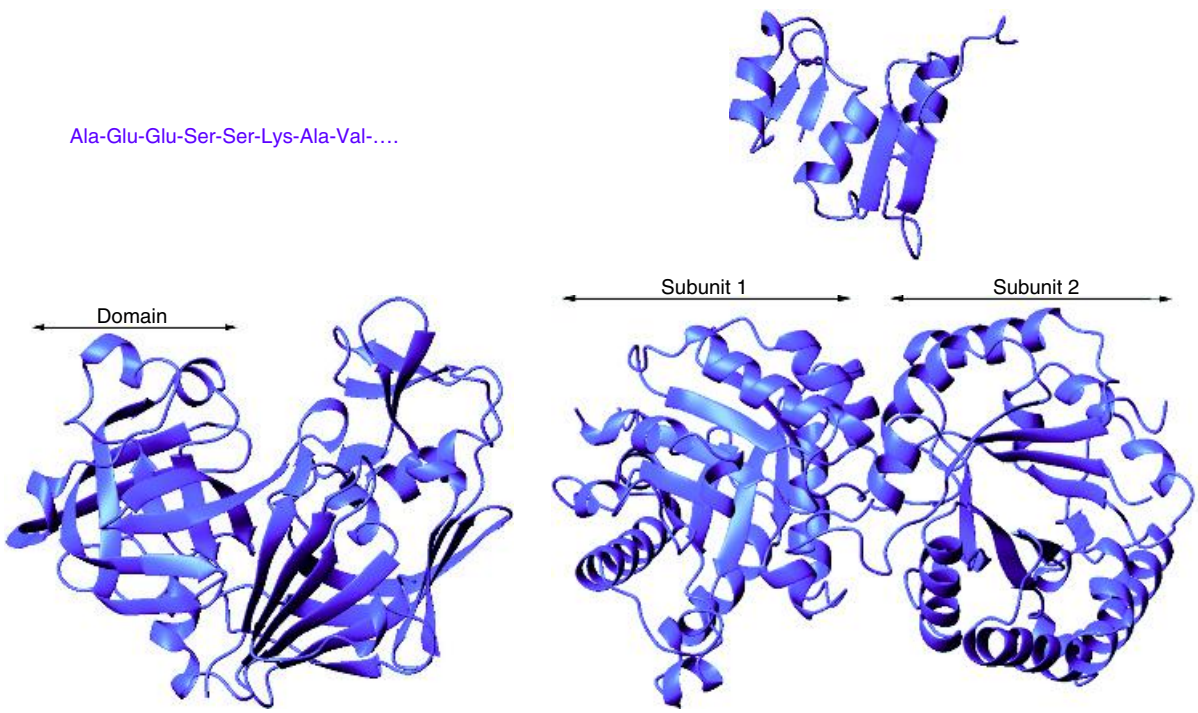


Figure 3.20 Four levels of organization within proteins; primary, secondary, tertiary and quaternary

structure involves a spherical compact molecule composed of secondary structural motifs with little irregular structure. Disordered or irregular structure in proteins is normally confined to the N and C terminals or more rarely to loop regions within a protein or linker regions connecting one or more domains. Asking the question 'what is the tertiary structure of a protein?' is synonymous with asking 'what is the protein fold?.' The fold arises from linking together secondary structures forming a compact globular molecule. Elements of secondary structure interact via hydrogen bonds, as in β sheets, but also depend on disulfide bridges, electrostatic interactions, van der Waals interactions, hydrophobic contacts and hydrogen bonds between non-backbone groups.

Interactions stabilizing tertiary structure

To form stable tertiary structure proteins must clearly form *more* attractive interactions than unfavourable or repulsive ones. The formation of stable tertiary folds relies on interactions that differ in their relative strengths and frequency in proteins.

Disulfide bridges

Disulfide bridges dictate a protein fold by forming strong covalent links between cysteine side chains that are often widely separated in the primary sequence. A disulfide bridge cannot form between consecutive cysteine residues and it is normal for each cysteine to be separated by at least five other residues. The formation of a disulfide bridge restrains the overall conformation of the polypeptide and in bovine pancreatic trypsin inhibitor (BPTI; Figure 3.21), a small protein of 58 residues, there are three disulfide bridges formed between residues 5–55, 14–38 and 30–51. The effect is to bring the secondary structural elements closer together. Disulfide bonds are only broken at high temperatures, acidic pH or in the presence of reductants. In BPTI reduction of the disulfide bonds leads to decreased protein stability that is mirrored by other disulfide-rich proteins.

The hydrophobic effect

The importance of the hydrophobic effect was for a long time underestimated. Charged interactions and hydrogen bonds are not strong *intra*molecular forces because water molecules compete significantly with these effects. However, water is a very poor solvent for many non-polar molecules and this is exemplified by dissolving an organic solute such as cyclohexane in water. Non-polar molecules cannot form hydrogen bonds with water and this prevents molecules such as cyclohexane dissolving extensively in aqueous solutions. As a consequence interactions between water and non-polar molecules are weakened and may be virtually non-existent. The result is an enhancement of interactions between non-polar molecules and the formation of hydrophobic clusters within water. The enhanced interactions between non-polar molecules in the presence of water are the basis for the hydrophobic effect. Since the side chains of many amino acid residues are hydrophobic it is clear that the hydrophobic effect may contribute significantly to intramolecular interactions. The hydrophobic effect can be restated as the preference of non-polar atoms for non-aqueous environments.

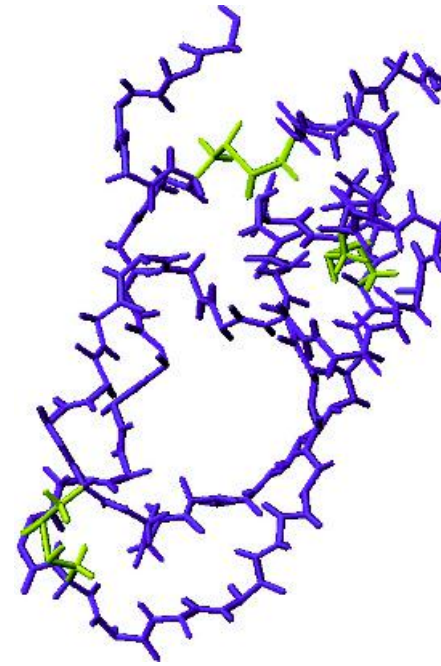


Figure 3.21 Structure of BPTI (PDB: 5PTI) showing the protein fold held by three disulfide bridges (5–55, 14–38 and 30–51)

The magnitude of the hydrophobic effect has proved difficult to estimate but has been accomplished by measuring the free energy associated with transfer of a non-polar solvent into water from the gaseous, liquid or solid states. The thermodynamics governing the transfer of non-polar molecules between phases are complicated but it is worth remembering that the enthalpy change, ΔH, represents changes in non-covalent interactions in going between the two phases whilst the entropy change (ΔS) reflects differences in the order of each system. A summation of these two terms gives ΔG_{tr}, the free energy of transfer of a non-polar or hydrophobic molecule from one phase to another, via the relationship $\Delta G = \Delta H - T\Delta S$, where T is temperature. The overall process involves the transfer of a solute molecule into the aqueous phase by (i) creating a cavity in the water, (ii) adding solute to the cavity, and (iii) maximizing favourable interactions between solute molecules and between solvent molecules.

In ice water molecules are arranged in a regular crystal lattice that maximizes hydrogen bonding with other water molecules, forming on average four hydrogen bonds. In the liquid phase these hydrogen bonds break and form rapidly with an estimated half-life of less than 1 ns. This leads to each water molecule forming an average of 3.4 hydrogen bonds. The capacity for hydrogen bonding and intermolecular attraction accounts for the high boiling point of water and the relatively large amounts of energy required to break these interactions, especially when compared with the interactions between non-polar liquids such as cyclohexane (Figure 3.22). Here the C–H bonds show little tendency to hydrogen bond and this is true for most of the non-polar side chains of amino acids in proteins. The hydrophobic interaction does not derive from the interaction between non-polar molecules or from the interaction between water and non-polar solutes because hydrogen bonds do not form. The driving force for the formation of hydrophobic clusters is the tendency for water molecules to hydrogen bond with each other. Water forms hydrogen bonded networks around non-polar solutes to become more ordered and one extreme example of this effect is the observation of clathrates where water forms an ordered cage around non-polar solutes. Around the cage water

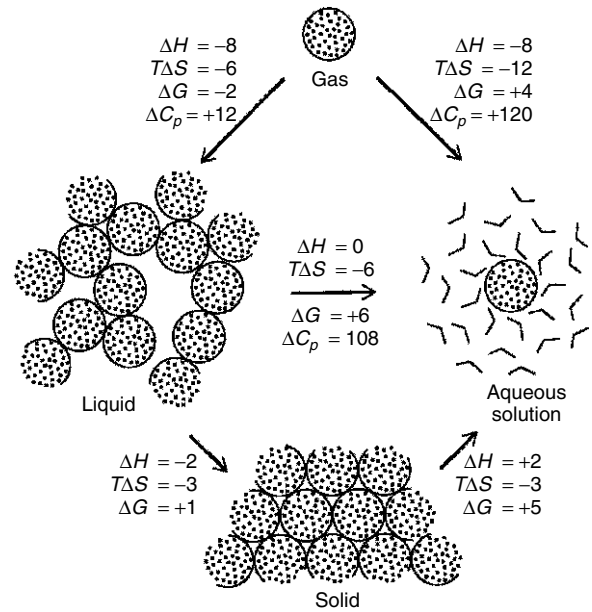


Figure 3.22 Transfer of a non-polar solute (cyclohexane) from gas, liquid and solid phases to aqueous solution around room temperature. The hydrophobic interaction is temperature dependent and the respective enthalpy (ΔH), entropy ($T\Delta S$) and free energy (ΔG) of transfer are given in kcal mol^{-1}. (Reproduced courtesy of Creighton, T.E. *Proteins Structure & Molecular Properties*, 2nd edn. W.H. Freeman, 1993)

forms with maximum hydrogen bonding but each bond has less than optimal geometry.

The energetics of the hydrophobic interaction can now be explored in more detail to provide a physical basis for the phenomena in protein structure. The most interesting transitions are those occurring from liquid phases involving the transfer of a non-polar solute from a non-polar liquid to water. The adverse ordering of water around the solute leads to an unfavourable decrease in entropy, whilst the ΔH_{tr} term is approximately zero. It is the entropic factor that dominates in the transfer of a non-polar solute into aqueous solutions with enthalpy terms reflecting increased hydrogen bonding.

The temperature dependence of the hydrophobic interaction provides still more clues concerning the process. As the temperature is increased water around

non-polar solutes is disrupted by breaking hydrogen bonds and becomes more like bulk water. This process is reflected by a large heat capacity (ΔC_p) that accompanies the hydrophobic interaction and is characteristic of this type of interaction. The large change in heat capacity underpins the temperature dependency of the hydrophobic interaction through its effect on both enthalpic and entropic terms. The magnitude of C_p is related to the non-polar surface area of the solute exposed to water along with many other thermodynamic parameters. As a result a large number of correlations have been made between accessible surface area, solubility of non-polar solutes in aqueous solutions and the energetics associated with the hydrophobic interaction. The hydropathy index established for amino acids is just one manifestation of the hydrophobic interaction described here.

Charge–charge interactions

These interactions occur between the side chains of oppositely charged residues as well as between the NH_3^+ and COO^- groups at the ends of polypeptide chains (Figure 3.23). Of importance to charge–charge interactions[1] are the side chains of lysine, arginine and histidine together with the side chains of aspartate, glutamate and to a lesser extent tyrosine and cysteine.

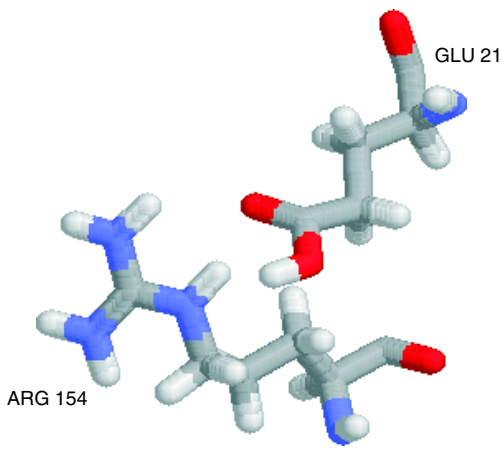


Figure 3.23 An example of electrostatic or charge–charge interactions occurring in proteins. The example shown is from the protease chymotrypsin (PDB: 2CGA). Here the interaction is between arginine and glutamate side chains. The donor atom is the hydrogen of the NE nitrogen whilst the acceptor groups are both oxygen atoms, OE1 and OE2. Reproduced courtesy Creighton, T.E. Proteins Structure & Molecular Properties, 2nd edn. W.H. Freeman, 1993

As a result of their charge the side chains of these residues are found on the protein surface where interactions with water or solvent molecules dramatically weaken these forces. In view of their low frequency and solvated status these interactions do not usually contribute significantly to the overall stability of a protein fold. Coulomb's law describes the potential energy (V) between two separated charges (in a perfect vacuum) according to the relationship

$$V = q_1\ q_2/\varepsilon\ 4\pi\ r^2 \qquad (3.1)$$

where q_1 and q_2 are the magnitude of the charges (normally $+1$ and -1), r is the separation distance between these charges and ε is the permittivity of free space. In other media, such as water this equation is modified to

$$V = q_1\ q_2/\varepsilon_o\ 4\pi\ r^2 \qquad (3.2)$$

where ε_o is the permittivity of the medium and is related to the dielectric constant (ε_r) by

$$\varepsilon_r = \varepsilon/\varepsilon_o \qquad (3.3)$$

When the medium is water this has a value of approximately ~80 whilst methanol has a value of ~34 and a hydrophobic solvent such as benzene has a value of ~2 (a perfect vacuum has a value of 1 by definition). It has been estimated that each charge–pair interaction located on the surface of a protein may contribute less than 5 kJ mol^{-1} to the overall stability of a protein and this must be compared with the much stronger disulfide bridge (~100–200 kJ mol^{-1}). Occasionally charge interactions exist within non-polar regions in proteins and under these conditions with a low dielectric medium and the absence of water the magnitude of the charge–charge interaction can be significantly greater.

A variation occurring in normal charge–charge interactions are the partial charges arising from hydrogen bonding along helices. Hydrogen bonding between amide and carbonyl groups gives rise to a net dipole moment for helices. In addition the peptide bonds pointing in the same direction contribute to the accumulative polarization of helices. As a result longer helices

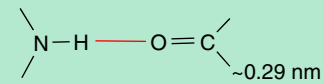
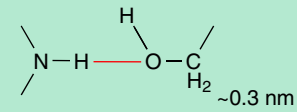
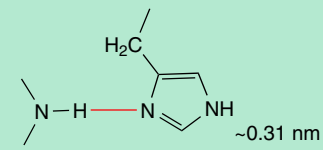
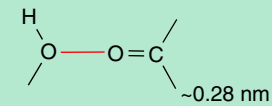
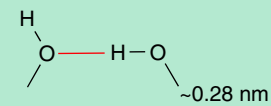
[1]Sometimes called electrostatic, ionic or salt-bridge interactions.

have a greater macrodipole. The net result is that a small positive charge is located at the N-terminal end of a helix whilst a negative charge is associated with the C-terminal end. For a helix of average length this is equivalent to +0.5–0.7 unit charge at the N-terminus and −0.5–0.7 units at the C terminus. Consequently, to neutralize the overall effect of the helix macrodipole acidic side chains are more frequently located at the positive end of the helix whilst basic side chains can occur more frequently at the negative pole.

Hydrogen bonding

Hydrogen bonds contribute significantly to the stability of α helices and to the interaction of β strands to form parallel or antiparallel β sheets. As a result such hydrogen bonds contribute significantly to the *overall* stability of the tertiary structure or folded state. These hydrogen bonds are between main chain NH and CO groups but the potential exists with protein folding to form hydrogen bonds between side chain groups and between main chain and side chain. In all cases the hydrogen bond involves a donor and acceptor atom and will vary in length from 0.26 to 0.34 nm (this is the distance between heavy atoms, i.e. between N and O in a hydrogen bond of the type N–H- - - O=C) and may deviate in linearity by ±40°. In proteins other types of hydrogen bonds can occur due to the presence of donor and acceptor atoms within the side chains. This leads to hydrogen bonds between side chains as well as side chain–main chain hydrogen bonds. Particularly important in hydrogen bond formation are the side chains of tyrosine, threonine and serine containing the hydroxyl group and the side chains of glutamine and asparagine with the amide group. Frequently, side chain atoms hydrogen bond to water molecules trapped within the interior of proteins whilst at other times hydrogen bonds appear shared between two donor or acceptor groups. These last hydrogen bonds are termed bifurcated. Table 3.4 shows examples of the different types of hydrogen bond.

Table 3.4 Examples of hydrogen bonds between functional groups found in proteins

Residues involved in hydrogen bonding	Nature of interaction and typical distance between donor and acceptor atoms (nm)
Amide–carbonyl Gln/Asn -backbone	
Amide-hydroxyl. Backbone-Ser Asn/Gln -Ser	
Amide-imidazole. Asn/Gln -His Backbone -His	
Hydroxyl-carbonyl Ser/Thr/Tyr–backbone Ser-Asn/Gln Thr- Asn/Gln Tyr- Asn/Gln	
Hydroxyl-hydroxyl Thr,Ser and Tyr Tyr-Ser	

Van der Waals interactions

There are attractive and repulsive van der Waals forces that control interactions between atoms and are very important in protein folding. These interactions occur between adjacent, uncharged and non-bonded atoms and arise from the induction of dipoles due to fluctuating charge densities within atoms. Since atoms are continually oscillating the induction of dipoles is a constant phenomena. It is unwise to view van der Waals forces

as simply the interaction between temporary dipoles; the interaction occurs between uncharged atoms and involves several different types. Unlike the electrostatic interactions described above they do not obey inverse square laws and they vary in their contributions to the overall intermolecular attraction. Three attractive contributions are recognized and include in order of diminishing strength: (i) the orientation effect or interaction between permanent dipoles; (ii) the induction effect or interaction between permanent and temporary dipoles; (iii) the dispersion effect or London force, which is the interaction between temporary, induced, dipoles.

The orientation effect is the interaction energy between two permanent dipoles and depends on their relative orientation. For a freely rotating molecule it might be expected that this effect would average out to zero. However, very few molecules are absolutely free to rotate with the result that preferred orientations exist. The energy of interaction varies as the inverse sixth power of the interatomic separation distance (i.e. $\propto r^{-6}$) and is inversely dependent upon the temperature. The induction effect also varies as r^{-6} but is independent of temperature. The magnitude of this effect depends on the polarizability of molecules. The dispersion effect, sometimes called the London force, involves the interactions between temporary and induced dipoles. It arises from a temporary dipole inducing a complementary dipole in an adjacent molecule. These dipoles are continually shifting, depend on the polarization of the molecule and result in a net attraction that varies as r^{-6}.[1]

When atoms approach very closely repulsion becomes the dominant and unfavourable interaction. The repulsive term is always positive but drops away dramatically as the distance between the two atoms increases. In contrast it becomes very large at short atomic distances and is usually modelled by r^{-12} distance dependency, although there are no experimental grounds for this relationship. The combined repulsive and attractive van der Waals terms are described by plots of the potential energy as a function of interatomic separation distance (Figure 3.24). The energy of the van der Waals reactions is therefore described

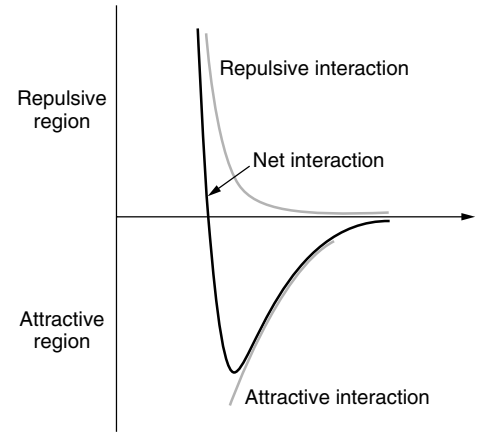


Figure 3.24 The van der Waals interaction as a function of interatomic distance *r*. The repulsive term increases rapidly as atoms get very close together

by the difference between the attractive (r^{-6}) and repulsive terms (r^{-12})

$$E_{\text{vdw}} \propto -\text{A}/r^6 + \text{B}/r^{12} \qquad (3.4)$$

$$E_{\text{vdw}} = \Sigma\ E\ [(r_{\text{m}}/r)^{12} - 2(r_{\text{m}}/r)^6] \qquad (3.5)$$

where the van der Waals potential (E_{vdw}) is the sum of all interactions over all atoms, E is the depth of the potential well and r_{m} is the minimum energy interaction distance. This type of interaction is often called a 6–12 interaction, or the Leonard Jones potential. Although van der Waals forces are extremely weak, especially when compared with other forces influencing protein conformation, their large number arranged close together in proteins make these interactions significant to the maintenance of tertiary structure.

A protein's folded state therefore reflects the summation of attractive and repulsive forces embodied by the summation of the electrostatic, hydrogen bonding, disulfide bonding, van der Waals and hydrophobic interactions (see Table 3.5). It is often of considerable surprise to note that proteins show only marginal stability with the folded state being between 20 and 80 kJ mol^{-1} more stable than the unfolded state. The relatively small value of ΔG reflects the differences in non-covalent interactions between the folded and

[1]The dispersion effect is so called because the movement of electrons underlying the phenomena cause a dispersion of light.

Table 3.5 Distance dependence and bond energy for interactions between atoms in proteins

Bond	Distance dependence	Approximate bond energy (kJ mol^{-1})
Covalent	No simple dependence	~200
Ionic	$\propto 1/r^2$	<20
Hydrogen bond	No simple expression	<10
van der Waals	$\propto 1/r^6$	<5
Hydrophobic	No simple expression	<10

unfolded states. This arises because an unfolded protein normally has an identical *covalent* structure to the folded state and any differences in protein stability arise from the favourable non-covalent interactions that occur during folding.

The organization of proteins into domains

For proteins larger than 150 residues the tertiary structure may be organized around more than one structural unit. Each structural unit is called a domain, although exactly the same interactions govern its stability and folding. The domains of proteins interact together although with fewer interactions than the secondary structural elements within each domain. These domains can have very different folds or tertiary structures and are frequently linked by extended relatively unstructured regions of polypeptide.

Three major classes of domains can be recognized. These are domains consisting of mainly α helices, domains containing mainly β strands and domains that are mixed by containing α and β elements. In this last class are structures containing both alternating α/β secondary structures as well as proteins made up of collections of helices and strands (α + β). Within each of these three groups there are many variations of the basic themes that lead to further classification of protein architectures (see Figure 3.25).

Protein domains arise by gene duplication and fusion. The result is that a domain is added onto another protein to create new or additional properties. An example of this type of organization is seen in the cytochrome b_5 superfamily. Cytochrome b_5 is a small globular protein containing both α helices and β strands functioning as a soluble reductant to methaemoglobin in the erythrocyte or red blood cell. The cytochrome contains a non-covalently bound heme group in which the iron shuttles between the ferric (Fe^{3+}) and ferrous (Fe^{2+}) states (Figure 3.26). The reversible redox chemistry of the iron is central to the role of this protein as an electron carrier within erythrocytes. When united with a short hydrophobic chain this protein assumes additional roles and participates in the fatty acyl desaturase pathway. In mitochondria the enzyme sulfite oxidase converts sulfite to sulfate as part of the dissimilatory pathway for sulfur within cells. Sulfite oxidase contains cytochrome b_5 linked to a molybdenum-containing domain. Further use of the b_5-like fold occurs in nitrate reductase where a flavin binding domain is linked to the cytochrome. The result is a plethora of new proteins containing different domains that allow the basic redox role of cytochrome b_5 to be exploited and enhanced to facilitate the catalysis of new reactions.

A rigorous definition of a domain does not exist. One acceptable definition is the presence of an autonomously folding unit within a protein. Alternatively a domain may be defined as a region of a protein showing structurally homology to other proteins. However, in all cases domains arise from the folding of a *single* polypeptide chain and are distinguished from quaternary structure (see below) on this basis.

Super-secondary structure

The distinctions between secondary structure and super-secondary structure or between tertiary structure and super-secondary structure are not well defined. However, in some proteins there appears to be an intermediate level of organization that reflects groups of secondary structural elements but does not encompass all of the structural domain or tertiary fold. The β barrel found in enzymes such as triose phosphate isomerase could form an element of super secondary structure since it does not represent *all* of the

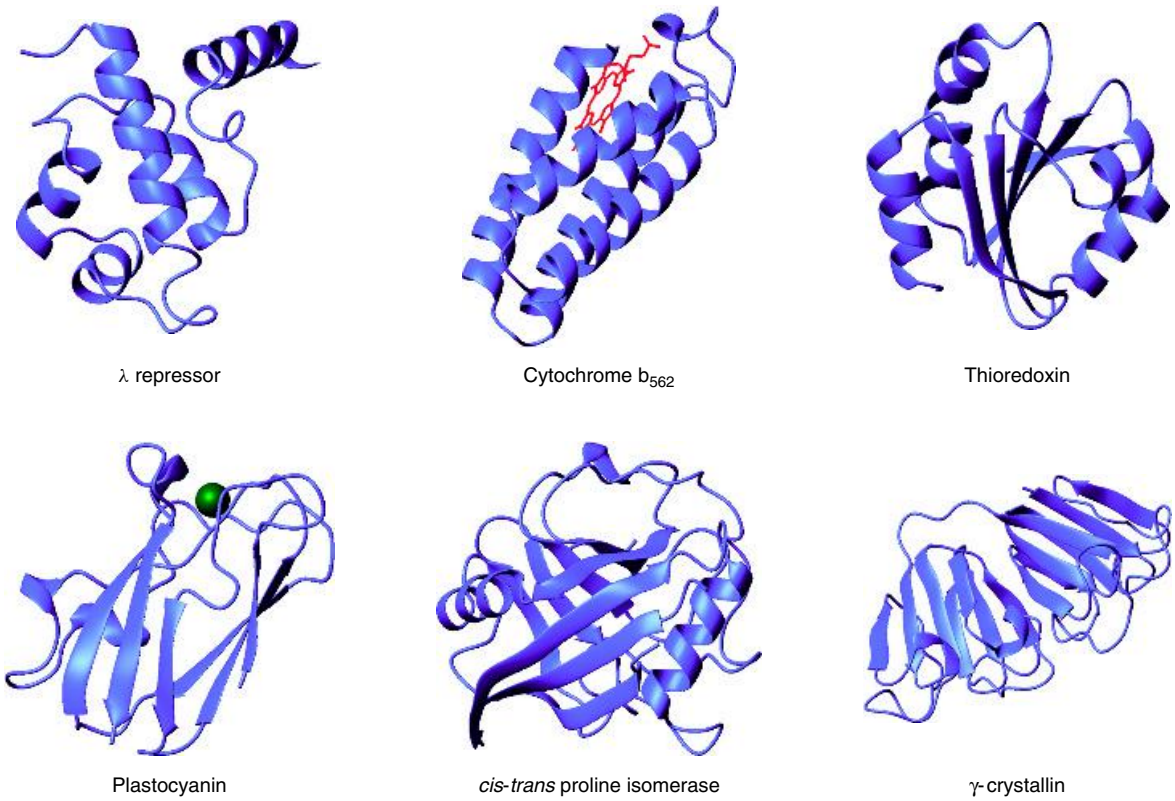


Figure 3.25 The secondary structure elements found in monomeric proteins. The λ repressor protein (PDB: 1LMB) contains the helix turn helix (HTH) motif; cytochrome b-562: (PDB: 256b) is a four-helix bundle heme binding domain; human thioredoxin: (PDB: 1ERU), a mixed α/β protein containing a five-stranded twisted β sheet. spinach plastocyanin: (PDB: 1AG6) a single Greek key motif binds Cu (shown in green); human *cis-trans* proline isomerase (PDB: 1VBS), a small extensive β domain containing a collection of strands that fold to form a 'sandwich'; human γ-crystallin: (PDB: 2GCR), two domains each of which is an eight-stranded β barrel type structure composed of two Greek key motifs

folded domain yet represents a far greater proportion of the structure than a simple β strand. Amongst other elements of super-secondary structure recognized in proteins are the β-α-β motif, Rossmann fold, four-helix bundles, the Greek-key motif and its variants, and the β meander. The cartoon representations in Figure 3.27 demonstrate the arrangement of β strands in some of these motifs and the careful viewer may identify these motifs in some figures in this chapter.

The β meander motif is a series of antiparallel β strands linked by a series of loops or turns. In the β meander the order of strands across the sheet reflects their order of appearance along the polypeptide sequence. A variation of this design is the so-called Greek key motif, and it takes its name from the design found on many ancient forms of pottery or architecture. The Greek key motif shown here links four antiparallel β strands with the third and fourth strands forming the outside of the sheet whilst strands 1 and 2 form on the inside or middle of the sheet. The Greek key motif can contain many more strands ranging from 4 to 13. The Cu binding metalloprotein plastocyanin contains eight β strands arranged in a Greek key motif.

A β sandwich forms normally via the interaction of strands at an angle and connected to each other

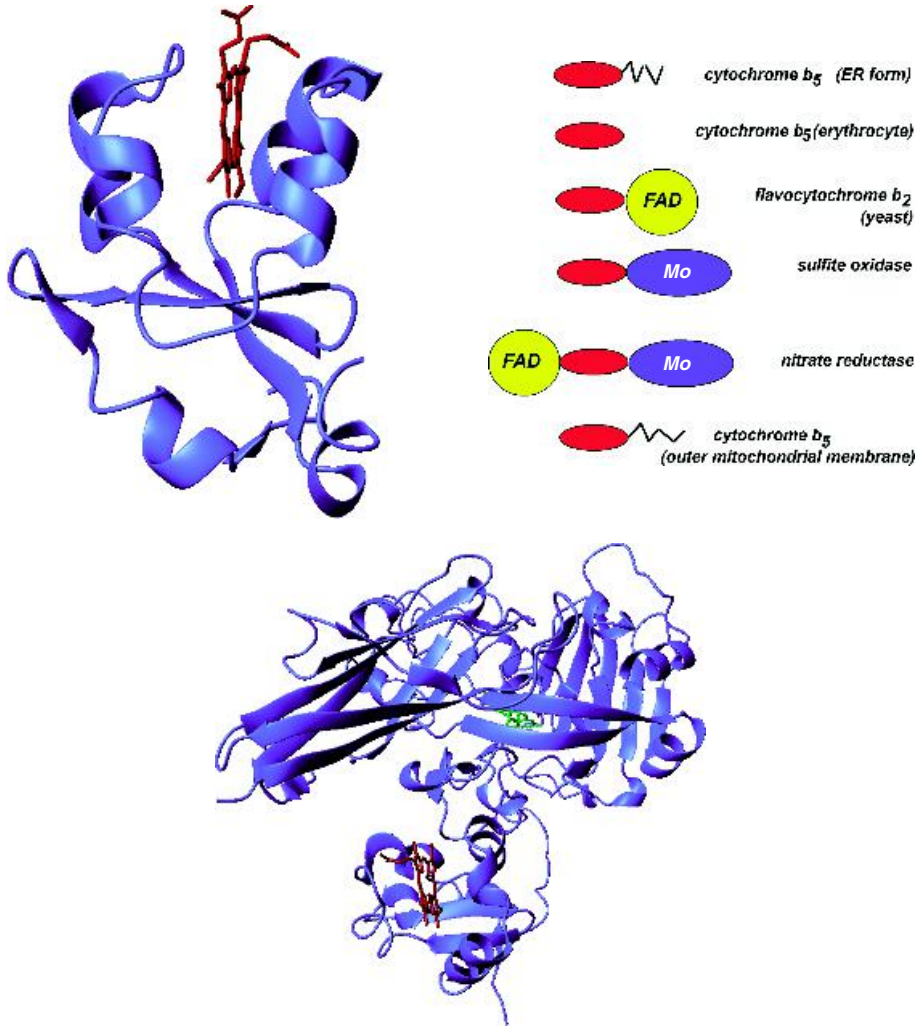


Figure 3.26 The heme binding domain of cytochrome b_5 (PDB:3B5C), a schematic representation of the duplication of this domain (red) to form part of multi-domain proteins by linking it to FAD domains, Mo-containing domains or hydrophobic tails. The arrangement of domains in proteins such as nitrate reductase, yeast flavocytochrome b_2 and sulfite oxidase is shown (left). The structure of sulfite oxidase (bottom) (PDB: 1SOX) shows the cytochrome domain in the foreground linked to a larger molybdenum-pterin binding domain

via short loops. In some cases these strands can originate from a different polypeptide chain but the emphasis is on two layers of β strands interacting together within a globular protein. The layers of the sandwich can be aligned with respect to each other or arranged orthogonally. An example of the second arrangement is shown in human *cis-trans* proline peptidyl isomerase.

The Rossmann fold, named after its discoverer Michael Rossmann, is an important super-secondary structure element and is an extension of the β-α-β domain. The Rossmann fold consists of three parallel β strands with two intervening α helices i.e. β-α-β-α-β. These units are found together as a dimer – so the Rossmann fold contains six β strands and four helices – and this collection of secondary structure

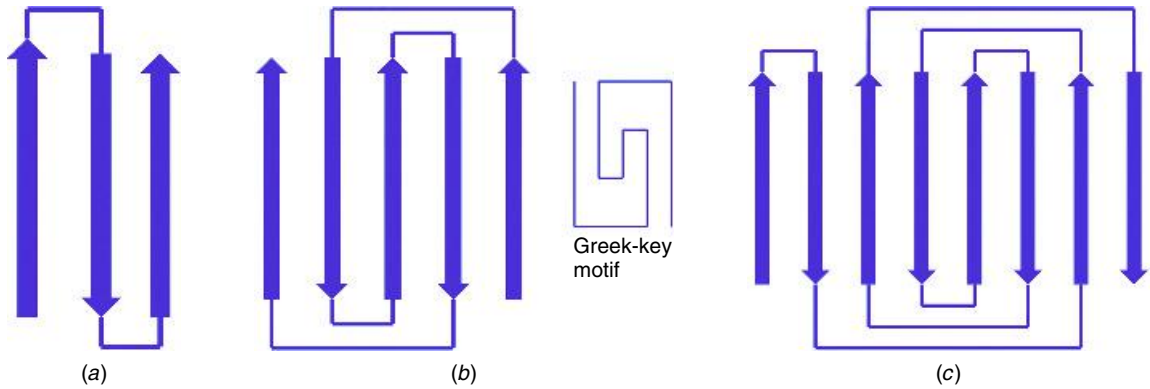


Figure 3.27 Cartoon representations of (a) the β meander, (b) Greek key and (c) Swiss or Jelly roll motifs

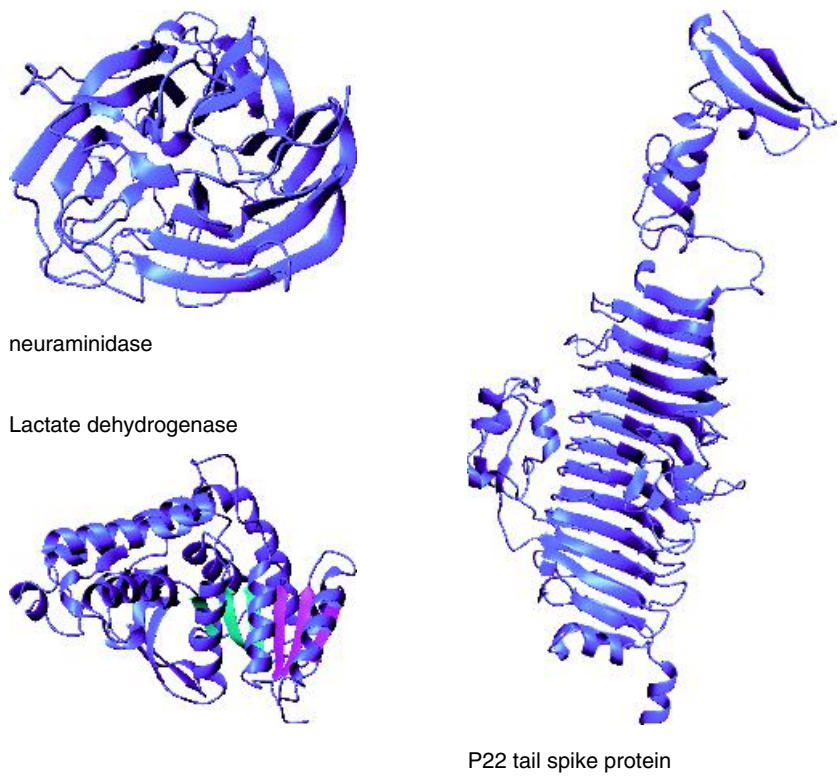


Figure 3.28 The structure of complex motifs involving β strands. Native influenza neuraminidase showing characteristic β propellor (PDB:1F8D). A β helix found in the tailspike protein of bacteriophage P22 (PDB:1TSP). The β-α-β-α-β motif is shown in the Rossmann fold of Lactate dehydrogenase with each set of three strands shown in different colours (PDB:1LDH)

frequently forms a nucleotide-binding site. Nucleotide binding domains are found in many enzymes and in particular, dehydrogenases, where the co-factor nicotinamide adenine dinucleotide is bound at an active site. Examples of proteins or enzymes containing the Rossmann fold are lactate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, alcohol dehydrogenase, and malate dehydrogenase. However, it is clear that this fold is found in other nucleotide proteins beside dehydrogenases, including glycogen phosphorylase and glyceroltriphosphate binding proteins.

Elements of super secondary structure are frequently used to allow protein domains to be classified by their structures. Most frequently these domains are identified by the presence of characteristic folds. A fold represents the 'core' of a protein domain formed from a collection of secondary structures. In many cases these folds occur in more than one protein allowing structural relationships to be established. These characteristic folds include four-helix bundles (cytochrome b_{562} in Figure 3.25), helix turn helix motifs (the λ repressor in Figure 3.25), β barrels, and the β sandwich as well as more complicated structures such as the β propellor and β helix (see Figure 3.28). The β helix is an unusual arrangement of secondary structure – β strands align in a parallel manner one above another forming inter-strand hydrogen bonds but collectively twisting as a result of the displacement of successive strands. The strands all run in the same direction and the displacement result in the formation of a helix. Both left-handed and right-handed β helix proteins have been discovered, and a prominent example of a right handed β helix occurs in the tailspike protein of bacteriophage P22. In this protein the tailspike protein is actually a trimer containing three interacting β helices. The arrangement of strands within a β helix gives any subunit with this structural motif a very elongated appearance and leads to the hydrophobic cores being spread out along the long axis as opposed to a typical globular packing arrangement.

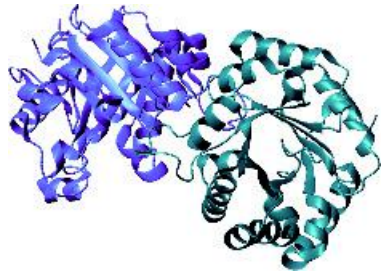
Quaternary structure

Many proteins contain more than one polypeptide chain. The interaction between these chains underscores quaternary structure. The interactions are exactly the same as those responsible for tertiary structure, namely disulfide bonds, hydrophobic interactions, charge–pair interactions and hydrogen bonds, with the exception that they occur between one or more polypeptide chains. The term subunit is often used instead of polypeptide chain.

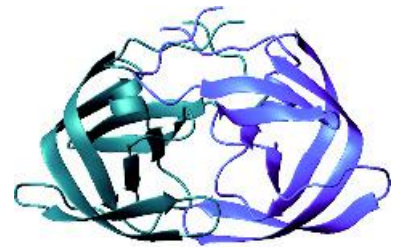
Quaternary structure can be based on proteins with identical subunits or on non-identical subunits (Figure 3.29). Triose phosphate isomerase, HIV protease and many transcription factors function as homodimers. Haemoglobin is a tetramer containing two different subunits denoted by the use of Greek letters, α and β. The protein contains two α and two β subunits in its tetrameric state and for haemoglobin this is normally written as $\alpha_2\beta_2$. Although it might be thought that aggregates of subunits are an artefact resulting from crystal packing it is abundantly clear that correct functional activity requires the formation of quaternary structure and the specific association of subunits. Subunits are held together predominantly by weak non-covalent interactions. Although individually weak these forces are large in number and lead to subunit assembly as well as gains in stability.

Quaternary structure is shown by many proteins and allows the formation of catalytic or binding sites at the interface *between* subunits. Such sites are impossible for monomeric proteins. Further advantages of oligomeric proteins are that ligand or substrate-binding causes conformational changes within the whole assembly and offer the possibility of regulating biological activity. This is the basis for allosteric regulation in enzymes.

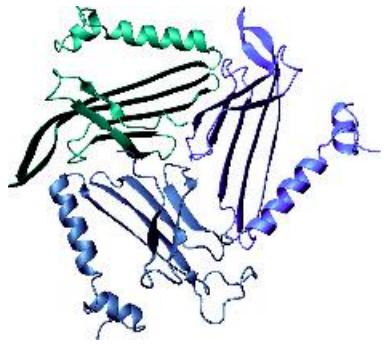
In the following sections the quaternary organization of transcription factors, immunoglobulins and oxygen-carrying proteins of the globin family are described as examples of the evolution of biological function in response to multiple subunits. The presence of higher order or quaternary structure allows greater versatility of function, and by examining the structures of some of these proteins insight is gained into the interdependence of structure and function. A common theme linking these proteins is that they all bind other molecules such as nucleic acid in the case of transcription factors, small inorganic molecules such as oxygen and bicarbonate by the globins and larger proteins or peptides in the case of immunoglobulins.



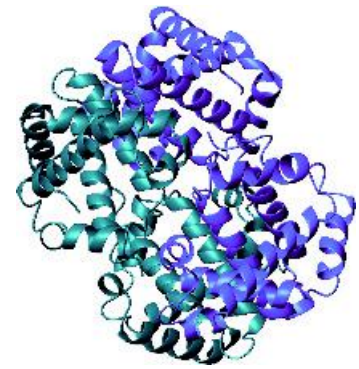
Triose phosphate isomerase (TIM)



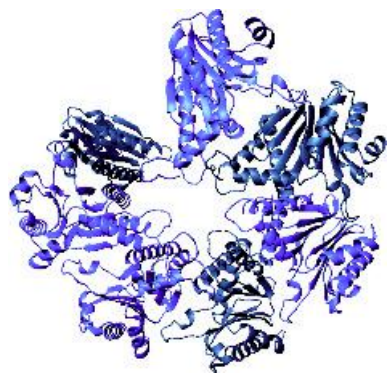
HIV protease



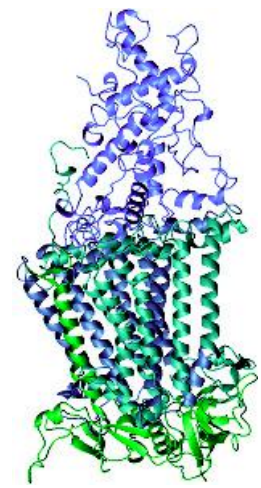
MS2 viral capsid protein



Haemoglobin



Proteasome



Bacterial photosynthetic reaction centre

Figure 3.29 The structures of large oligomeric proteins. Dimers are shown with triose phosphate isomerase and HIV protease. Trimers are represented by the MS2 viral capsid protein. Haemoglobin is a tetramer composed of two pairs of identical subunits. The proteasome consists of four concentric rings each made up of seven subunits. Only one ring is shown in the current view. The bacterial photosynthetic reaction centre contains four different subunits, the H,M,L and C subunits

Dimeric DNA-binding proteins

Many dimeric proteins exist in the proteomes of cells and one of the most common occurrences is the use of two subunits and a two-fold axis of symmetry to bind DNA. DNA binding proteins are a very large group of proteins typified by transcription factors. Transcription factors bind to promoter regions of DNA sequences called, in eukaryotic systems, the TATA box, and in prokaryotes a Pribnow box. The TATA box is approximately 25 nucleotides upstream of the transcription start site and as their name suggests these factors mediate the transcription of DNA into RNA by promoting RNA polymerase binding to this region of DNA in a pre-initiation complex (Figure 3.30).

In prokaryotes the mode of operation is simpler with fewer modulating elements. Prokaryotic promoter contains two important zones called the −35 region and −10 region (Pribnow box). The −35 bp region functions in the initial recognition of RNA polymerase and possesses a consensus sequence of TTGACAT. The −10 region has a consensus sequence (TATAAT) and occurs about 10bp before the start of a bacterial gene.

The cI and cro proteins from phage λ were amongst the first studied DNA binding proteins and act as regulators of transcription in bacteriophages such as 434 or λ. The life cycle of λ bacteriophage is controlled by the dual action of the cI and cro proteins in a complicated series of reactions where DNA binding close to initiation sites physically interferes with gene transcription by RNA polymerase

Figure 3.30 Transcription starts at the initiation site (+1) but is promoted by binding to the consensus TATA box sequence (−25) of the RNA polymerase complex that includes transcription factors. Upstream DNA sequences that facilitate transcription have been recognized (CAT and GC boxes ∼−80 and ∼−90 bases upstream of start site) whilst downstream elements lack consensus sequences but exist for some transcription systems

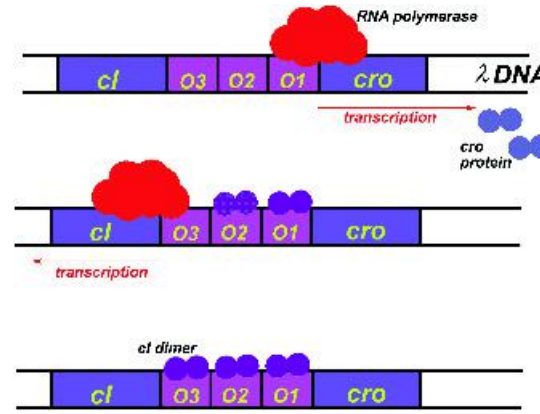


Figure 3.31 The control of transcription in bacteriophage λ by cro and cI repressor proteins. The cI dimer may bind to any of three operators, although the order of affinity is O1 ≈ O2 > O3. In the absence of cI proteins, the cro gene is transcribed. In the presence of cI proteins only the cI gene may be transcribed. High cI concentrations prevents transcription of both genes

(Figure 3.31). For this reason these proteins are also called repressors. When phage DNA enters a bacterial host cell two outcomes are possible: lytic infection results in the production of new viral particles or alternatively the virus integrates into the bacterial genome lying dormant for a period known as the lysogenic phase. Lytic and lysogenic phases are initiated by phage gene expression but competition between the cro and cI repressor proteins determines which pathway is followed. Cro and cI repressor compete for control by binding to an operator region of DNA that contains at least three sites that influence the lytic/lysogenic switch. The bacteriophage remains in the lysogenic state if cI proteins predominate. Conversely the lytic cycle is followed if cro proteins dominate.

Determining the structure of the cI repressor in the presence of a DNA containing a consensus binding sequence uncovered detailed aspects of the mechanism of nucleic acid binding. The mechanism of DNA binding is of considerable interest in view of the widespread occurrence of transcription factors in all cells and the growing evidence of their involvement in many disease states. In the absence of DNA a

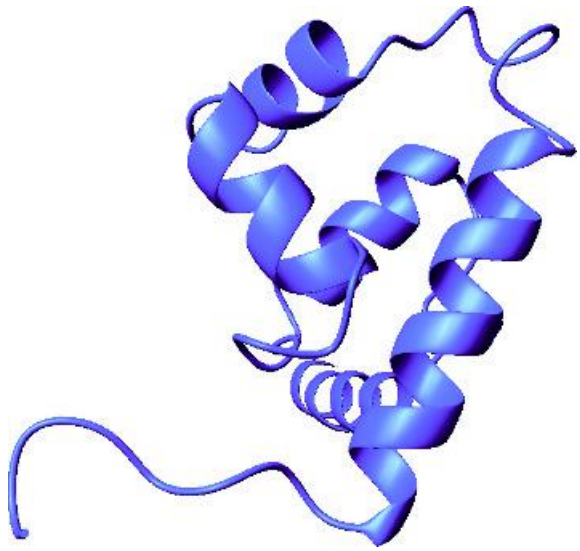


Figure 3.32 The monomer form of the cI repressor (PDB: 1LMB)

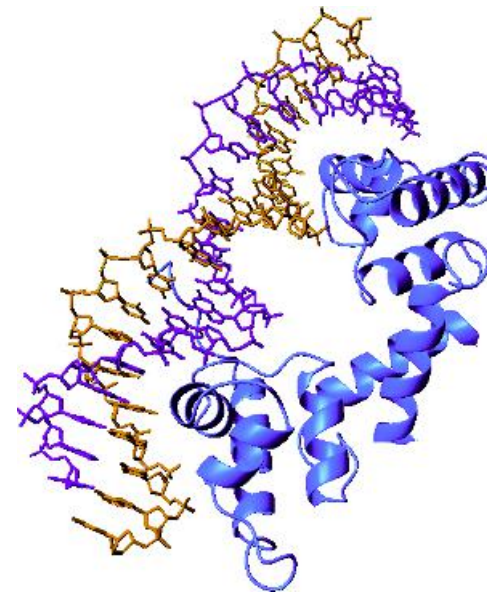


Figure 3.33 The structure of the N terminal domain of the λ repressor in the presence of DNA. The fifth helix forms part of the dimerization domain that allows two monomer proteins to function as a homodimer. In each case helices 2 and 3 bind in the major groove of DNA since the spatial separation of each HTH motif is comparable with the dimensions of the major groove

structure for the cI repressor revealed a polypeptide chain containing five short helices within a domain of ~70 residues (Figure 3.32). Far more revealing was the structure in the presence of DNA and the use of helices 2 and 3 as a helix-turn-helix or HTH motif to fit precisely into the major groove (Figure 3.33).

The N- and C-terminal domains of one subunit are separated by mild proteolysis. Under these conditions two isolated N-terminal domains form a less stable dimer that has lowered affinity for DNA when compared with the complete repressor. The HTH motif is 20 residues in length and is formed by helices 2 and 3 found in the N terminal domain. Although the HTH motif is often described as a 'domain' it should be remembered it is part of a larger protein and does not fold into a separate, stable element of structure. Helix 3 of the HTH motif makes a significant number of interactions with the DNA. It is called the recognition helix with contact points involving Gln33, Gln44, Ser45, Gly46, Gly48, and Asn52. These residues contain a significant proportion of polar side chains and bind directly into the major groove found in DNA. In comparison helix 2 makes fewer contacts and has a role of positioning helix 3 for optimal recognition. Of the remaining structure helices 4 and 5 form part of a dimer interface whilst the C terminal domain is involved directly in protein dimerization. Dimerization is important since it allows HTH motifs to bind to successive major grooves along a sequence of DNA. Determination of the structure of the cro repressor protein showed a very similar two-domain structure to the cI repressor with comparable modes of DNA binding (see Figure 3.34). The overwhelming similarity in structure between these proteins suggested a common mechanism of DNA binding and one that might extend to the large number of transcription factors found in eukaryotes.

Comparison of the sequences of HTH motifs from λ and cro repressors showed sequence conservation (Figure 3.35). The first seven residues of the 20 residue HTH motif form helix 2, a short 4 residue turn extends from positions 8 to 11 and the next (recognition) helix is formed from residues 12 to 20. The sequences of

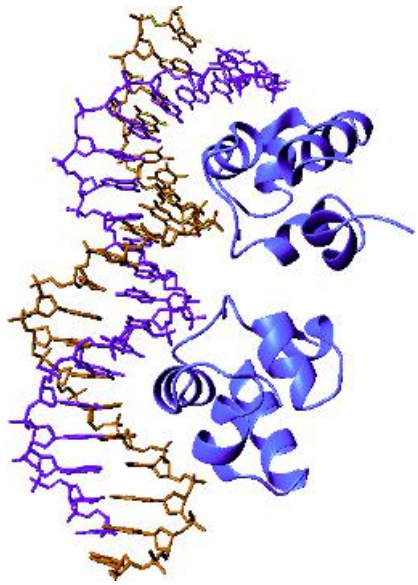


Figure 3.34 The cro repressor protein from bacteriophage 434 bound to a 20-mer DNA helix (PDB:4CRO)

```
Cro     : QTKTAKDLGVYQSAINKAIH
434Cro  : QTELATKAGVKQQSIQLIEA
cI      : QESVADKMGIGQSGVGALFN
```

Figure 3.35 The sequence homology in the cro and cI repressors. Residues highlighted in red are invariant whilst those shown in yellow illustrate conservative changes in sequence

the HTH motif in the cro and λ repressors helped define structural constraints within this region. Residue 9 is invariant and is always a glycine residue. It is located in the 'turn' region and larger side chains are not easily accommodated and would disrupt the turn and by implication the positioning of the DNA binding helix. Residues 4 and 15 are completely buried from the solvent whilst residues 8 and 10 are partially buried, non-polar side chains charged side chains unfavourable at these locations. Proline residues are never found in HTH motifs. Further structural constraints were apparent at residue 5 located or wedged between the two helices. Large or branched side chains would cause a different alignment of the helices and destroy the primary function of HTH motif. DNA binding was critically dependent on the identity of side chains for residues 11–13, 16–17 and 20. Residues with polar side chains were common and were important in forming hydrogen bonds to the major groove.

The globin family and the role of quaternary structure in modulating activity

Myoglobin occupies a pivotal position in the history of protein science. It was the first protein structure to be determined in 1958. Myoglobin and haemoglobin, whose structure was determined shortly afterwards, have been extensively studied particularly in relation to the inter-dependence of structure *and* function. Many of the structure–function relationships discovered for myoglobin have proved to be of considerable importance to the activity of other proteins.

The evolutionary development of oxygen carrying proteins was vital to multicellular organisms where large numbers of cells require circulatory systems to deliver oxygen to tissues as well as specialized proteins to carry oxygen around the body. In vertebrates the oxygen-carrying proteins are haemoglobin and myoglobin. Haemoglobin is located within red blood cells and its primary function is to convey oxygen from the lungs to regions deficient in oxygen. This includes particularly the skeletal muscles of the body where oxygen consumption as a result of mechanical work requires continuous supplies of oxygen for optimal activity. In skeletal muscle myoglobin functions as an oxygen storage protein.

For both haemoglobin and myoglobin the oxygen-carrying capacity arises through the presence of a heme group. The heme group is an organic carbon skeleton called protoporphyrin IX made up of four pyrrole groups that chelate iron at the centre of the ring (Figure 3.36). The heme group gives myoglobin and haemoglobin a characteristic colour (red) that is reflected by distinctive absorbance spectra that show intense peaks around 410 nm due to the Fe-protoporphyrin IX group. The iron group is found predominantly in the ferrous state (Fe^{2+}) and only this form binds oxygen. Less frequently the iron group is found in an oxidized state as the ferric iron

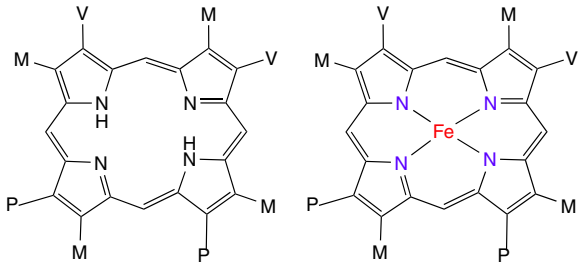


Figure 3.36 The structure of protoporphyrin IX and heme (Fe-protoporphyrin-IX)

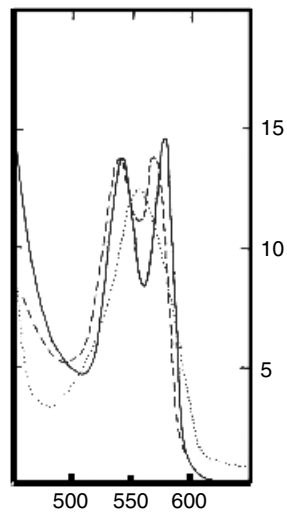


Figure 3.37 The absorbance spectra of oxy, met and deoxyhaemoglobin. The dotted line is the spectrum of deoxyHB, the solid line is metHb whilst the dashed line is oxyHb. DeoxyHb shows a single maximum at 550 nm that shifts upon oxygen binding to give two peaks at 528 and 563 nm. The extinction coefficients associated with these peaks are $\varepsilon_{555} = 12.5$ mM cm^{-1}, $\varepsilon_{541} = 13.5$ mM cm^{-1}, and $\varepsilon_{576} = 14.6$ mM cm^{-1}

(Fe^{3+}), a state not associated with oxygen binding. The binding of oxygen (and other ligands) to the sixth coordination site of iron in the heme ring is accompanied by distinctive changes to absorbance spectra (Figure 3.37).

The structure of myoglobin

Detailed crystallographic studies of myoglobin provided the first three-dimensional picture of a protein and established many of the ground rules governing secondary and tertiary structure described earlier in this chapter. Myoglobin was folded into an extremely compact single polypeptide chain with dimensions of ~4.5 × 3.5 × 2.5 nm. There was no free space on the inside of the molecule with the polypeptide chain folded efficiently by changing directions regularly to create a compact structure. Approximately 80 percent of all residues were found in α helical conformations. The structure of myoglobin (Figure 3.38) provided the first experimental verification of the α helix in proteins and confirmed that the peptide group was planar, found in a *trans* configuration and with the dimensions predicted by Pauling.

Eight helices occur in myoglobin and these helices are labelled as A, B, C etc up to the eighth helix,

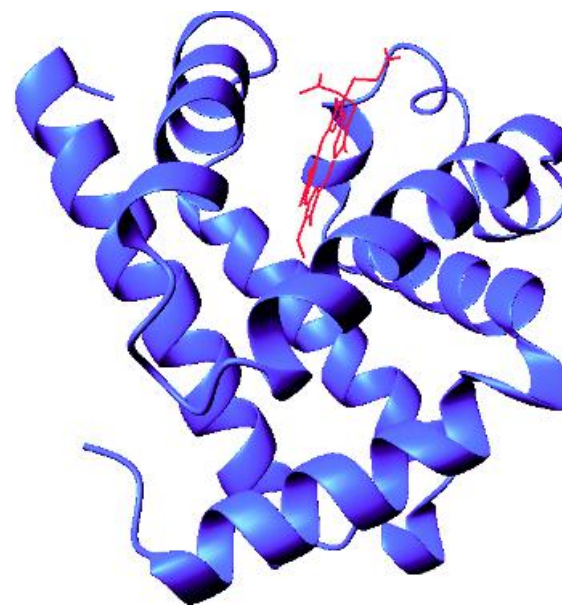


Figure 3.38 The structure of myoglobin showing α helices and the heme group. Helix A is in the foreground along with the N terminal of the protein. The terminology used for myoglobin labels the helices A–H and residues within helices as F8, A2 etc.

helix H. Frequently, although not in every instance, the helices are disrupted by the presence of proline residues as for example occurs between the B and C helices, the E and F helices and the G and H helices. Even in low-resolution structures of myoglobin (the first structure produced had a resolution of ~6 Å) the position of the heme group is easily discerned from the location of the electron-dense iron atom.

The heme group was located in a crevice surrounded almost entirely by non-polar residues with the exception of two heme propionates that extended to the surface of the molecule making contact with solvent and two histidine residues, one of which was in contact with the iron and was termed the proximal (F8) histidine (Figure 3.39). The imidazole side chain of F8 provided the fifth ligand to the heme iron. A second histidine, the distal histidine (E7), was more distant from the Fe centre and did *not* provide the sixth ligand. An obvious result of this arrangement was an asymmetric conformation for the iron where it was drawn out of the heme plane towards the proximal histidine.

The structure of myoglobin confirmed the partitioning of hydrophobic and hydrophilic side chains. The interior of the protein and the region surrounding the heme group consisted almost entirely of non-polar residues. Here leucine, phenylalanine and valine were common whilst hydrophilic side chains were located on the exterior or solvent accessible surface.

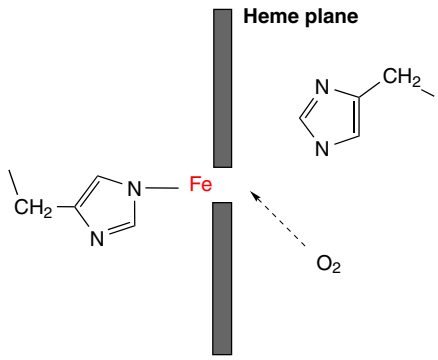


Figure 3.39 Diagram of the proximal and distal histidine side chains forming part of the oxygen binding site of myoglobin with the ferrous iron pulled out of the plane of the heme ring. The vacant sixth coordination position is the site of oxygen binding

There are approximately 200 structures, including mutants, of myoglobin deposited in the PDB but the three most important structures of relevance to myoglobin are those of the oxy and deoxy forms together with ferrimyoglobin (metmyoglobin) where the iron is present as the ferric state. The structures of all three forms turned out to be remarkably similar with one exception located in the vicinity of the sixth coordination site. In oxymyoglobin a single oxygen molecule was found at the sixth coordination site. In the deoxy form this site remained vacant whilst in metmyoglobin water was found in this location. The unique geometry of the iron favours its maintenance in the reduced state but oxygen binding resulted in movement of the iron approximately 0.2 Å towards the plane of the ring (Figure 3.40). An analysis of the heme binding site emphasizes how the properties of the heme group are modulated by the polypeptide. The identical ferrous heme group in cytochromes undergoes oxidation in the presence of oxygen to yield the ferric (Fe^{3+}) state, other enzymes such as catalase or cytochrome oxidase convert the oxygen into hydrogen peroxide and water respectively but in globins the oxygen is bound with the iron remaining in the ferrous state.

Myoglobin has a very high affinity for oxygen and this is revealed by binding curves or profiles that record the fractional saturation versus the concentration of oxygen (expressed as the partial pressure of oxygen, pO_2). The oxygen-binding curve of myoglobin is

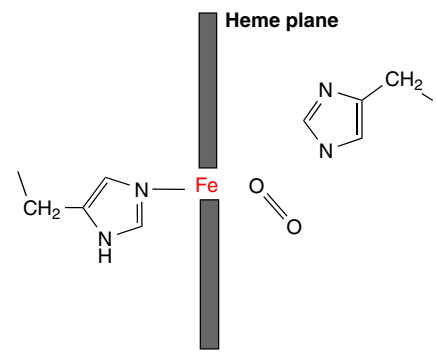


Figure 3.40 Binding of oxygen in oxymyoglobin and the movement of the iron into the approximate heme plane

hyperbolic and rapidly reaches a saturating level seen by the asymptotic line above pO_2 levels of 30 torr. The affinity for oxygen is expressed as an equilibrium

$$MbO_2 \leftrightharpoons Mb + O_2 \tag{3.6}$$

where the equilibrium constant, K, is

$$K = [Mb][O_2]/[MbO_2] \tag{3.7}$$

The fractional saturation of myoglobin (Y) is simply expressed as the number of oxygenated myoglobin (MbO_2) molecules divided by the total number of myoglobin molecules ($MbO_2 + Mb$). Thus the fractional saturation (Y)

$$Y = [MbO_2]/([MbO_2] + [Mb]) \tag{3.8}$$

and substitution of Equation 3.7 into 3.8 yields

$$Y = pO_2/(pO_2 + K) \tag{3.9}$$

where pO_2 reflects the concentration, strictly partial pressure, of oxygen in the atmosphere surrounding the solution. Equation 3.9 may be written by equating the equilibrium constant in terms of the partial pressure of oxygen necessary to achieve 50 percent saturation leading to

$$Y = pO_2/(pO_2 + P_{50}) \tag{3.10}$$

The oxygen-binding curve of haemoglobin revealed a very different profile. The curve is no longer hyperbolic but defines a sigmoidal or S-shaped profile. In comparison with myoglobin the haemoglobin molecule becomes saturated at much higher oxygen concentrations (Figure 3.41). The profile defines a binding curve where initial affinity for oxygen is very low but then increases dramatically before becoming resistant to further oxygenation. This is called cooperativity and these differences are fundamentally dependent on the structures of myoglobin and haemoglobin. In order to understand the basis for these differences it is necessary to compare and contrast the structure of haemoglobin with myoglobin.

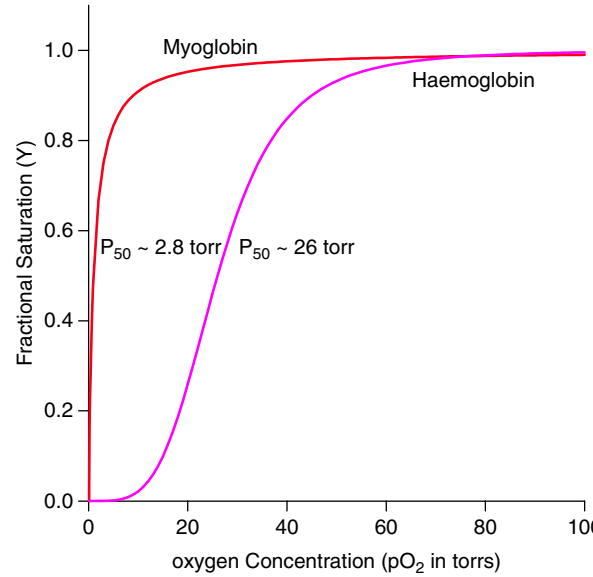


Figure 3.41 The oxygen-binding curves of myoglobin and haemoglobin show different profiles. Whilst the affinity of myoglobin for oxygen is reflected by a hyperbolic curve the affinity curve of haemoglobin is sigmoidal. The P_{50} of myoglobin is 2.8 torr whilst the P_{50} of haemoglobin in red blood cells is 26 torr (760 torr = 1 atm). In the tissues oxygen concentrations of between 20 and 40 torr are typical whilst in the lungs much higher partial pressures of oxygen exist above 100

The structure of haemoglobin

The most obvious difference between haemoglobin and myoglobin is the presence of quaternary structure in the former (Figure 3.42). In mammals adult haemoglobin is composed of four polypeptide chains containing two different primary sequences in the form of 2α chains and 2β chains. The determination of the structure of haemoglobin revealed that each α and β subunit possessed a conformation similar to myoglobin and showed a low level of sequence homology that reflected evolution from a common ancestral protein. The structure of haemoglobin revealed that each globin chain contained a heme prosthetic group buried within a crevice whilst the four subunits packed together forming a compact structure with little free space yet this time with the crucial difference of additional interactions *between* subunits.

Although not immediately apparent the tertiary structures of myoglobin, α globin and β globin chains

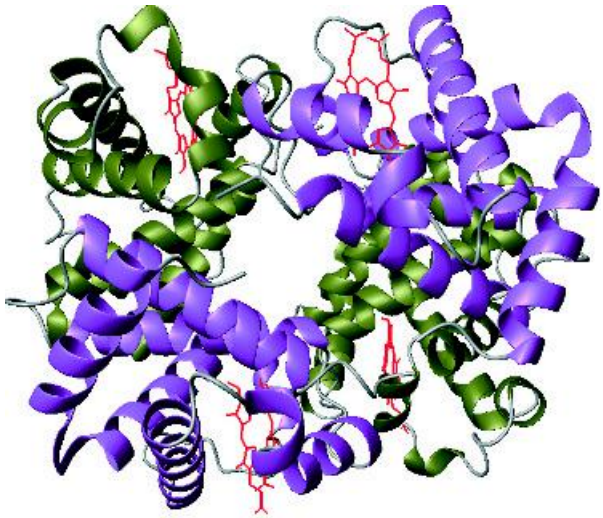


Figure 3.42 The quaternary structure of haemoglobin showing the four subunits each with a heme group packing together. The heme groups are shown in red with the α globin chain in purple and the β globin chain in green

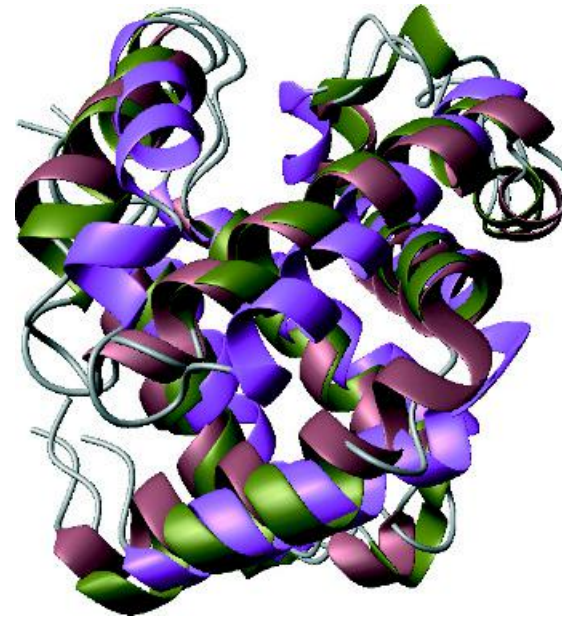


Figure 3.43 Superposition of the polypeptide chains of myoglobin, α globin and β globin. Structural similarity occurs despite only 24 out of 141 residues showing identity. The proximal and distal histidines are conserved along with several residues involved in the heme pocket structure such as leucine (F4) and a phenylalanine between C and D helices

are remarkably similar despite differences in primary sequence (Figure 3.43). With similar tertiary structures for myoglobin and the α and β chains of haemoglobin it is clear that the different patterns of oxygen binding must reflect additional subunit interactions in haemoglobin. In other words cooperativity arises as a result of quaternary structure.

The pattern of oxygenation in haemoglobin is perfectly tailored to its biological function. At high concentrations of oxygen as would occur in the lungs ($pO_2 > 100$ torr or $\sim 0.13 \times$ atmospheric pressure) both myoglobin and haemoglobin become saturated with oxygen. Myoglobin would be completely oxygenated carrying 1 molecule of oxygen per protein molecule. Complete oxygenation of haemoglobin would result in the binding of 4 oxygen molecules. However, as the concentration of oxygen decreases below 50 torr myoglobin and haemoglobin react differently. Myoglobin remains fully saturated with oxygen whilst haemoglobin is no longer optimally oxygenated; sites on haemoglobin are only 50 percent occupied at ~26 torr whereas myoglobin exhibits a much lower P_{50} of ~3 torr.

The physiological implications of these binding properties are profound. In the lungs haemoglobin is saturated with oxygen ready for transfer around the body. However, when red blood cells (containing high concentrations of haemoglobin) reach the peripheral tissues, where the concentration of oxygen is low, unloading of oxygen from haemoglobin occurs. The oxygen released from haemoglobin is immediately bound by myoglobin. The binding properties of myoglobin allow it to bind oxygen at low concentrations and more importantly facilitate the transport of oxygen from lungs to muscles or from regions of high concentrations to tissues with much lower levels.

The cooperative binding curve of haemoglobin arises as a result of structural changes in conformation that occur upon oxygenation. In the deoxy form the sixth coordination site is vacant and the iron is drawn out of the heme plane towards the proximal histidine.

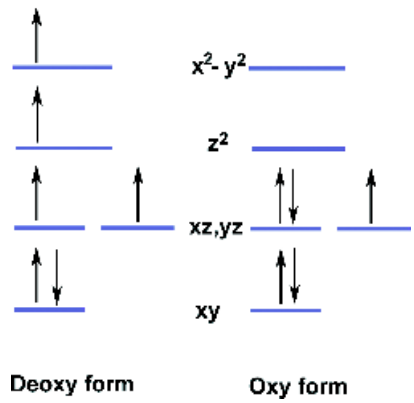


Figure 3.44 The spin state changes that occur in the deoxy and oxy ferrous forms of haemoglobin showing the population of the d orbitals (xy, xz, yz, z^2, x^2-y^2) by electrons. The deoxy form is high spin ($S = 2$) ferrous state with an ionic radius too large to fit between the four tetra pyrrole ring nitrogens. The low spin ($S = 1/2$) state of ferrous iron with a different distribution of orbitals has a smaller radius and moves into the plane of the heme

Oxygenation results in conformational change mediated by oxygen binding to the iron. The events are triggered by changes in the electronic structure of the iron as it shifts from a high spin ferrous centre in the deoxy form to a low spin state when oxygenated. Changes in spin state are accompanied by reorganization of the orbital structure and decreases in ionic radius that allows the iron to move closer to the plane of the heme macrocycle (Figure 3.44).

Changes at the heme site are accompanied by reorientation of helix F, the helix containing the proximal histidine, and results in a shift of approximately 1 Å to avoid unfavourable contact with the heme group. The effect of this conformational change is transmitted throughout the protein but particularly to interactions *between* subunits. Oxygen binding causes the disruption of ionic interactions between subunits and triggers a shift in the conformation of the tetramer from the deoxy to oxy state. By examining the structure of haemoglobin in both the oxy and deoxy states many of these interactions have been highlighted and another significant milestone was a structural understanding for the cooperative oxygenation of haemoglobin proposed by Perutz in 1970.

The mechanism of oxygenation

Perutz's model emphasized the link between cooperativity and the structure of the tetramer. The cooperative curve derives from conformational changes exhibited by the protein upon oxygenation and reflects the effect of two 'competing' conformations. These conformations are called the R and T states and the terms derive originally from theories of allosteric transitions developed by Monod, Wyman and Changeaux (MWC). According to this model haemoglobin exists in an equilibrium between the R and T states where the R state signifies a 'relaxed' or active state whilst the T signifies a 'tense' or inactive form.

The deoxy state of haemoglobin is resistant to oxygenation and is equated with the inactive, T state. In contrast the oxy conformation is described as the active, R state. In this model haemoglobin is an equilibrium between R and T states with the observed oxygen-binding curve reflecting a combination of the binding properties of each state. The shift in equilibrium between the R and T states is envisaged as a two state or concerted switch with only weak and strong binding states existing and intermediate states containing a mixture of strong and weak binding subunits specifically forbidden. This form of the MWC model could be described as an 'all or nothing' scheme but is more frequently termed the 'concerted' model. In general detailed structural analysis has retained many of the originally features of the MWC model.

At a molecular level the relative orientation between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers differed in the oxy and deoxy states. Oxygenation leads to a shift in orientation of $\sim 15^\circ$ and a translation of ~ 0.8 Å for one pair of α/β subunits relative to the other. It arises as a result of changes in conformation at the interfaces between these pairs of subunits. One of the most important regions involved in this conformational switch centres around His97 at the boundary between the F and G helices of the β_2 subunit. This region of the β subunit makes contact with the C helix of the α_1 subunit and lies close to residue 41. In addition the folding of the β_2 subunit leads to the C terminal residue, His146, being situated above the same helix (helix C)

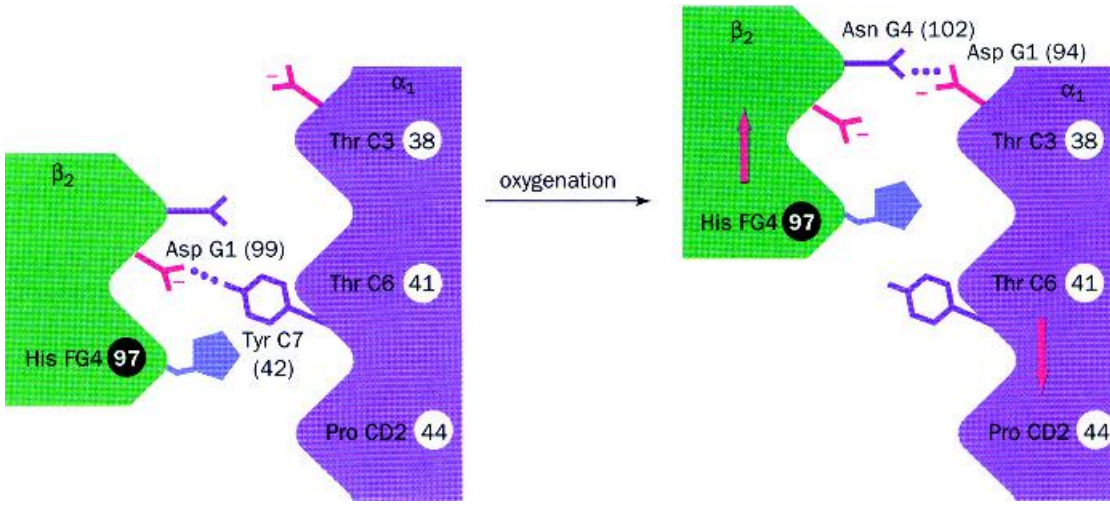


Figure 3.45 Interactions at the $\alpha_1\beta_2$ interface in deoxyhaemoglobin and the changes that occur after oxygenation. In the T state His97 of the β2 subunit fits next to Thr41 of the α1 chain. In the R state conformation changes result in this residue lying alongside Thr38. Although a salt bridge between Asp99 and Tyr42 is broken in the T > R transition a new interaction between Asn102 and Asp94 results. There is no detectable intermediate between the T and R states and the precisely interacting surfaces allow the two subunits to move relative to each other easily. (Reproduced with permission from Voet, D., Voet, J.G & Pratt, C.W., *Fundamentals of Biochemistry*. Chichester, John Wiley & Sons, 1999.)

and it forms a series of hydrogen bonds and salt bridges with residues in this region (38–44). In the deoxy state His 146 is restrained by an interaction with Asp94. As a result of symmetry an entirely analogous set of interactions occur at the $\alpha_2\beta_1$ interface (Figure 3.45).

In the T state the iron is situated approximately 0.6 Å out of the plane of the heme ring as part of a dome directed towards HisF8. Oxygenation causes a shortening of the iron–porphyrin bonds by 0.1 Å due to electron reorganization and causes the iron to move into the plane dragging the proximal His residue with it. Movement of the His residue by 0.6 Å is unfavourable on steric grounds and as a result the helix moves to compensate for changes at the heme centre. Helix movement triggers conformational changes throughout the subunits and more importantly between the subunits. Other important conformational switches include disruption of a network of ion pairs within and between subunits. In particular ionic interactions of the C terminal residue of the α subunit (Arg141) with Asp126 and the amino terminal of the α1 subunit and that of the β subunit (His146) with Lys40 (α) and Asp94(β) are broken upon oxygenation. Since these interactions stabilize the T state their removal drives the transition towards the R state.

Cooperativity arises because structural changes within one subunit are coupled to conformational changes throughout the tetramer. In addition the nature of the switch means that intermediate forms cannot occur and once one molecule of oxygen has bound to a T state subunit converting it to the R state all other subunits are transformed to the R state. As a result the remaining subunits rapidly bind further oxygen molecules, leading to the observed shape of the oxygen binding curve seen in Figure 3.46.

Allosteric regulation and haemoglobin

The new properties of haemoglobin extend further than the simple binding of oxygen to the protein. Haemoglobin's affinity for oxygen is modulated by

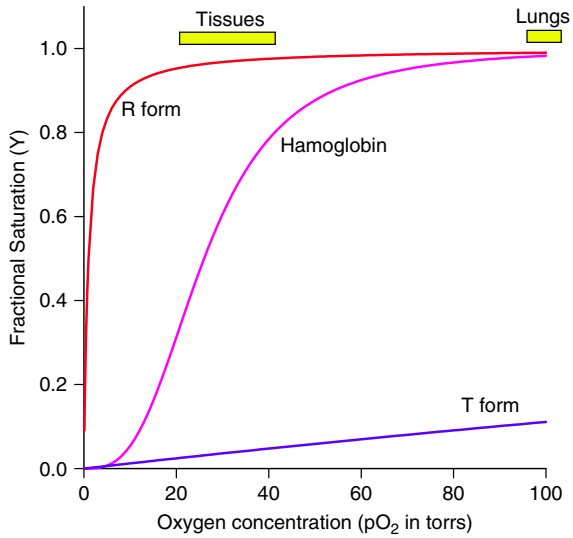


Figure 3.46 Oxygen binding profiles of the R and T states of haemoglobin. The R or active form of haemoglobin binds oxygen readily and shows a hyperbolic curve. The T or inactive state is resistant to oxygenation but also shows a hyperbolic binding curve with a higher P_{50}. The observed binding profile of haemoglobin reflects the sum of affinities of the R and T forms for oxygen and leads to the observed P_{50} of ~26 torr for isolated haemoglobin. Shown in yellow is the approximate range of partial pressures of oxygen in tissues and lungs

small molecules (effectors) as part of an important physiological control process. This mode of regulation is called allostery. An allosteric modulator binds to a protein altering its activity and in the case of haemoglobin the effector alters oxygen binding. The most important allosteric regulator of haemoglobin is 2,3-bisphosphoglycerate (2,3 BPG) (Figure 3.47).[1]

The allosteric modulator of haemoglobin, 2,3 BPG, raises the P_{50} for oxygen binding in haemoglobin from 12 torr in isolated protein to 26 torr observed for haemoglobin in red blood cells (Figure 3.48). Indeed the existence of an allosteric modulator of haemoglobin was first suspected from careful comparisons of the oxygen-binding properties of isolated protein with that found in red blood cells. In erythrocytes 2,3 BPG

[1]Sometimes called 2,3 DPG = 2,3-diphosphoglycerate.

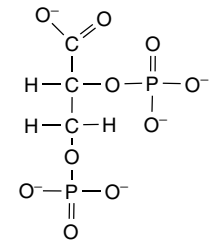


Figure 3.47 2,3-Bisphosphoglycerate is a negatively charged molecule

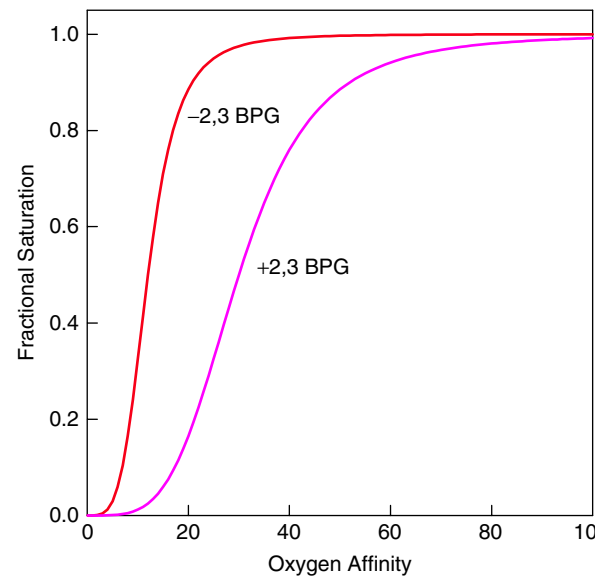


Figure 3.48 Modulation of oxygen binding properties of haemoglobin by 2,3 BPG. The 2,3 BPG modulates oxygen binding by haemoglobin raising the P_{50} from 12 torr in isolated protein to 26 torr observed for haemoglobin in red blood cells. This is shown by the respective oxygen affinity curves

is found at high concentrations as a result of highly active glycolytic pathways and is formed from the key intermediates 3-phosphoglycerate or 1,3 bisphosphoglycerate. The result is to create concentrations of 2,3 BPG approximately equimolar with haemoglobin within cells. In the erythrocyte this results in 2,3 BPG concentrations of ~5 mM.

The effect arises by differential binding – the association constant (K_a) for 2,3 BPG for deoxyhaemoglobin is $\sim 4 \times 10^4$ M^{-1} compared with a value of ~ 300 M^{-1} for the oxy form. In other words it stabilizes the T state, which has low affinity for oxygen whilst binding to the oxy state is effectively inhibited by the presence of bound oxygen. The effect of 2,3 BPG on the oxygen-binding curve of haemoglobin is to shift the profile to much higher P_{50} (Figure 3.48). Detailed analysis of the structures of the oxy and deoxy states showed that 2,3 BPG bound in the central cavity of the deoxy form and allowed a molecular interpretation for its role in allostery. In the centre of haemoglobin between the four subunits is a small cavity lined with positively charged groups formed from the side chains of His2, His143 and Lys82 of the β subunits together with the two amino groups of the first residue of each of these chains (Val1). The result of this charge distribution is to create a strong binding site for 2,3 BPG. In oxyhaemoglobin the binding of oxygen causes conformational changes that lead to a closure of the allosteric binding site. These changes arise as a result of subunit movement upon oxygenation and decrease the cavity size to prevent accommodation of 2,3 BPG. Both oxygen and 2,3 BPG are reversibly bound ligands yet each binds at separate sites and lead to opposite effects on the R→T equilibrium.

2,3 BPG is not the only modulator of oxygen binding to haemoglobin. Oxygenation of haemoglobin causes the disruption of many ion pairs and leads to the release of ~0.6 protons for each oxygen bound. This effect was first noticed in 1904 by Christian Bohr and is seen by increase oxygen binding with increasing pH at a constant oxygen level (Figure 3.49). The Bohr effect is of utmost importance in the physiological delivery of oxygen from the lungs to respiring tissues and also in the removal of CO_2 produced by respiration from these tissues and its transport back to the lungs. Within the erythrocyte CO_2 is carried as bicarbonate as a result of the action of carbonic anhydrase, which rapidly catalyses the slow reaction

$$CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^- \qquad (3.11)$$

In actively respiring tissues where pO_2 is low the protons generated as a result of bicarbonate formation favour the transition from R > T and induce the

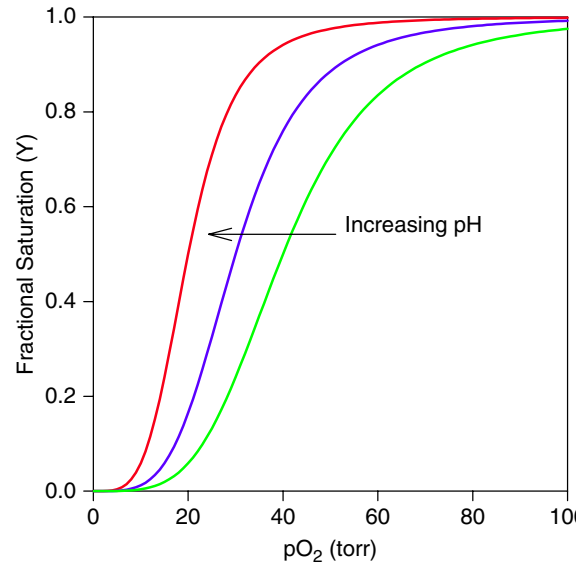


Figure 3.49 The Bohr effect. The oxygen affinity of haemoglobin increases with increasing pH

haemoglobin to unload its oxygen. In contrast in the lungs the oxygen levels are high and binding to haemoglobin occurs readily with the disruption of T state ion pairs and the formation of the R state. A further example of the Bohr effect is seen in very active muscles where an unavoidable consequence of high rates of respiration is the generation of lactic acid. The production of lactic acid will cause a lowering of the pH and an unloading of oxygen to these deficient tissues. Carbon dioxide can also bind directly to haemoglobin to form carbamates and arises as a result of reaction with the N terminal amino groups of subunits. The T form (or deoxy state) binds more carbon dioxide as carbamates than the R form. Again the physiological consequences are clear. In capillaries with high CO_2 concentrations the T state is favoured leading to an unloading of oxygen from haemoglobin.

Immunoglobulins

The immunoglobulins are a large group of proteins found in vertebrates whose unique function is to bind foreign substances invading a host organism. By

breaking the physical barrier normally represented by skin or mucous membranes pathogens enter a system, but once this first line of defence is breached an immune response is normally started. In most instances the immune response involves a reaction to bacteria or viruses but it can also include isolated proteins that break the skin barrier. Collectively these foreign substances are termed antigens.

The immune response is based around two systems. A circulating antibody system based on B cells sometimes called by an older name of humoral immune response. The cells were first studied in birds where they matured from a specialized pocket of tissue known as the bursa of Fabricius. However, there is no counterpart in mammals and B cells are derived from the bone marrow where the 'B' serves equally well to identify its origin. B cells have specific proteins on their surfaces and reaction with antigen activates the lymphocyte to differentiate into cells capable of antibody production and secretion. These cells secrete antibodies binding directly to antigens and acting as a marker for macrophages to destroy the unwanted particle.

This system is supported by a second cellular immune response based around T lymphocytes. The T cells are originally derived from the thymus gland and these cells also contain molecules on their membrane surfaces that recognize specific antigens and assist in their destruction.

Early studies of microbial infection identified that all antigens are met by an immune response that involves a collection of heterogeneous proteins known as antibodies. Antibodies are members of a larger immunoglobulin group of proteins and an important feature of the immune response is its versatility in responding to an enormous range of antigens. In humans this response extends from before birth and includes our lifelong ability to fight infection. One feature of the immune response is 'memory'. This property is the basis of childhood vaccination and arises from an initial exposure to antigen priming the system so that a further exposure leads to the rapid production of antibodies and the prevention of disease. From a biological standpoint the immune response represents an enigma. It is a highly specific recognition system capable of identifying millions of diverse antigens yet it retains an ability to 'remember' these antigens over considerable periods of time (years). The operation of the immune system was originally based around the classic observations reported by Edward Jenner at the end of the 18th century. Jenner recognized that milk maids were frequently exposed to a mild disease known as cowpox, yet rarely succumbed to the much more serious smallpox. Jenner demonstrated that cowpox infection, a relatively benign disease with complete recovery taking a few days, conferred immunity against smallpox. Exposure to cowpox triggered an immune response that conferred protection against the more virulent forms of smallpox, at that time relatively common in England. It was left to others to develop further the ideas of vaccination, notably Louis Pasteur, but for the last 100 years vaccination has been a key technique in fighting many diseases. The basis to vaccination lies in the working of the immune system.

The mechanism of specific antibody production remained puzzling but a considerable advance in this area arose with the postulation of the clonal selection theory by Neils Jerne and Macfarlane Burnet in 1955. This theory is now widely supported and envisages stem cells in the bone marrow differentiating to become lymphocytes each capable of producing a single immunoglobulin type. The immunoglobulin is attached to the outer surfaces of B lymphocytes and when an antigen binds to these antibodies replication of the cell is stimulated to produce a clone. The result is that only cells experiencing contact with an antigen are stimulated to replicate. Within the group of cloned B cells two distinct populations are identified. Effector B cells located in the plasma will produce soluble antibodies. These antibodies are comparable to those bound to the surface of B cells but lack membrane-bound sequences that anchor the antibodies to the lipid bilayer. The second group within the cloned B cell populations are called 'memory cells'. These cells persist for a considerable length of time even after the removal of antigen and allow the rapid production of antibodies in the event of a second immune reaction. The clonal selection theory was beautiful in that it explained how individuals distinguish 'self' and 'non-self'. During embryonic development immature B cells encounter 'antigens' on the surfaces of cells. These B cells do not replicate but are destroyed thereby removing antibodies that would react against the host's own

proteins. At birth the only B cells present are those that are capable of producing antibodies against non-self antigens.

Immunoglobulin structure

Despite the requirement to recognize enormous numbers of potential antigens all immunoglobulin molecules are based around a basic pattern that was elucidated by the studies of Rodney Porter and Gerald Edelman. Porter showed that immunoglobulin G (IgG) with a mass of ~150 000 could be split into three fragments each retaining biological activity through the action of proteolytic enzymes such as papain. Normal antibody contains two antigen-binding sites (Figure 3.50) but after treatment with papain two fragments each binding a single antigen molecule were formed, along with a fragment that did not bind antigen but was necessary for biological function. The two fragments were called the F_{ab} (F = fragment, ab = antigen binding) and F_c (c = crystallizable) portions. The latter's name arose from the fact that its homogeneous composition allowed the fragment to be crystallized in contrast to the F_{ab} fragments.

The IgG molecule is dissociated into two distinct polypeptide chains of different molecular weight via the action of reducing agents. These chains were called the heavy (H) and light (L) polypeptides (Figure 3.51).

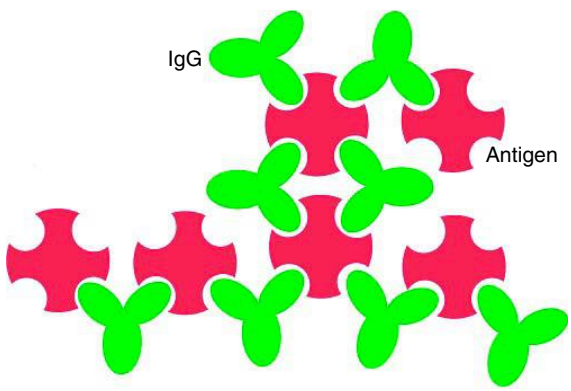


Figure 3.50 Schematic diagram showing bivalent interaction between antigen and antibody (IgG) and extended lattice

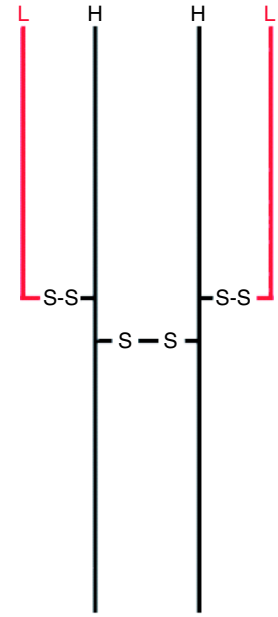


Figure 3.51 The subunit structure of IgG is H_2L_2

Further studies by Porter revealed that IgG was reconstituted by combining 2H and 2L chains, and a model for immunoglobulin structure was proposed envisaging each L chain binding to the H chain via disulfide bridges, whilst the H chains were similarly linked to each other. The results of papain digestion were interpretable by assuming cleavage of the IgG molecule occurred on the C terminus side of the disulfide bridge linking H and L chains. The F_{ab} region therefore contained both L chains and the amino terminal region of 2 H chains whilst the F_c region contained only the C terminal half of each H chain.

The amino acid sequence revealed not only the periodic repetition of intrachain disulfide bonds in the H and L chains but also the existence of regions of sequential homology between the H and L chains and within the H chain itself. Portions of the molecule, now known as the variable regions in the H and L chains, showed considerable sequence diversity, whilst constant regions of the H chain were internally homologous, showing repeating units. In addition the constant regions were homologous to regions on the L chain. This pattern of organization immediately suggested that

immunoglobulins were derived from simpler antibody molecules based around a single domain of ~100 residues, and by a process of gene duplication the structure was assimilated to include multiple domains and chains.

Although the organization of IgG molecules is remarkably similar antigen binding is promoted through the use of sequence variability. These differences are not distributed uniformly throughout the H and L chains but are localized in specific regions (Figure 3.53). The regions of greatest sequence variability lie in the N terminal segments of the H and L chains and are called the V_L or V_H regions. In the L chain the region extends for ~110 residues and is accompanied by a more highly conserved region of the same size called the constant region of C_L. In the H chain the V_H region extends for ~110 residues whilst the remainder of the sequence is conserved, the C_H region. The C_H region is subdivided into three homologous domains: C_H1, C_H2 and C_H3.

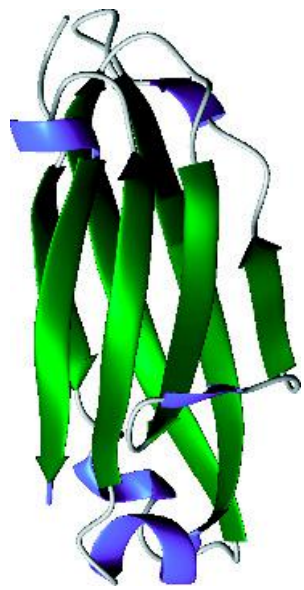


Figure 3.52 The immunoglobulin fold containing a sandwich formed by two antiparallel sets of β strands. Normally seven β strands make up the two sets. In this domain additional sheet structures occur and small regions of a helix are found shown in blue. The important loop and turn regions are shown in grey. In the IgG light chain this fold would occur twice; once in the variable region and once again in the constant region. In the heavy chain it occurs four times

Fractionation studies identified antigen-binding sites at the N terminal region of the H and L chains in a region composed of the V_H and V_L domains where the ability to recognize antigens is based on the surface properties of these folds. Within the V_H and V_L domains the sequences of highest variability are three short segments known as hypervariable sequences and collectively known as complementarity determining regions (CDRs). They bind antigen and by arranging these CDRs in different combinations cells generate vast numbers of antibodies with different specificity.

The light chain consists of two discrete domains approximately 100–120 residues in length whilst the heavy chain is twice the size and has four such domains. Each domain is characterized by common structural topology known as the immunoglobulin fold and is repeated within the IgG molecule. The immunoglobulin fold is a sandwich of two sheets of antiparallel β strands each strand linked to the next via turns or large loop regions, with the sheets usually linked via a disulfide bridge. The immunoglobulin fold is found in other proteins operating within the immune system but is also noted in proteins with no obvious functional similarity (Figure 3.52).

The hypervariable regions are located in turns or loops of variable length that link together the different β strand elements. Collectively these regions form the antigen-binding site at the end of each arm of the immunoglobulin molecule and insight into the interaction between antibody and antigen has been gained from crystallization of an F_{ab} fragment with hen egg white lysozyme.

Raising antibodies against lysozyme generated several antibody populations – each antibody recognizing a different site on the surface of the protein. These sites are known as epitopes or antigenic determinants. By studying one antibody–antigen complex further binding specificity was shown to reside in the contact of the V_L and V_H domains with lysozyme at the F_{ab} tip (Figure 3.54). Interactions based around hydrogen bonding pairs involved at least five residues in the V_L domain (Tyr32, Tyr50, Thr53, Phe91, Ser93) and five residues in the V_H domain (Gly53, Asp54, Asp100, Tyr101, Arg102). These donor and acceptors hydrogen

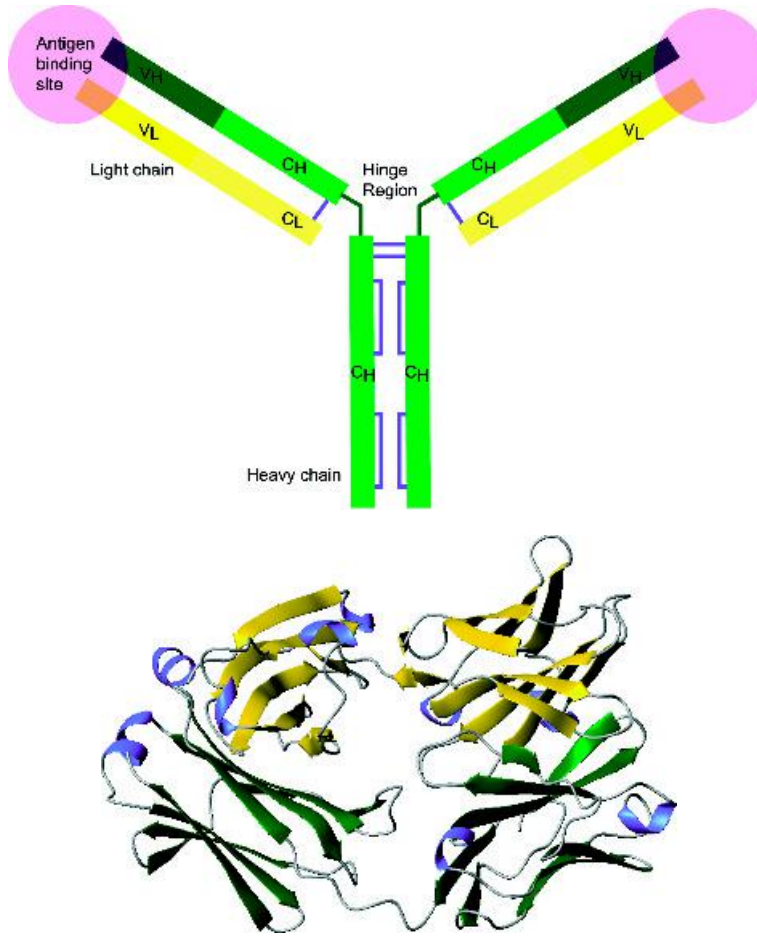


Figure 3.53 The structure and organization of IgG. The location of V_L, V_H and C_H within the H and L chains of an antibody together with a crystal structure determined for the isolated F_{ab} region (PDB: 7FAB). The H chain is shown with yellow strands whilst the L chain has strands shown in green. In each chain two immunoglobulin folds are seen

bonded to two groups of residues on lysozyme with the V_L domain interacting with Asp18, Asn19 and Gln121 whilst those on the V_H domain formed hydrogen bonds with 7 residues on the surface of lysozyme (Gly22, Ser24, Asn27, Gly117, Asp119, Val120, and Gln121).

The epitope on the surface of lysozyme is made up of two non-contiguous groups of residues (18–27 and 117–125) found on the antigen's surface. All six hypervariable regions participated in epitope recognition, but on the surface of lysozyme Gln121 was particularly important protruding away from the surface and forming a critical residue in the formation of a high-affinity antibody–antigen complex. The antibody forms a cleft that surrounds Gln121 with a hydrogen bond formed between the side chain and Tyr101 on the antibody (Figure 3.55). Elsewhere, van der Waals, hydrophobic and electrostatic interactions play a role in binding the other regions of the interacting surfaces between antibody and antigen.

Five different classes of immunoglobulins (Ig) have been recognized called IgA, IgD, IgE, IgG and IgM. These classes of immunoglobulins differ in the composition of their heavy chains whilst the light chain is based in all cases around two sequences identified

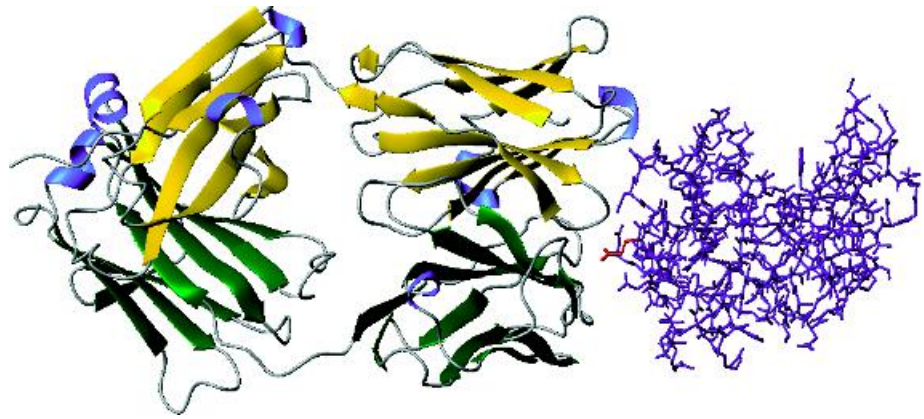


Figure 3.54 Interaction between monoclonal antibody fragment F_{ab} and lysozyme. The lysozyme is shown in blue on right and the prominent side chain of Gln121 is shown in red (PDB:1FDL)

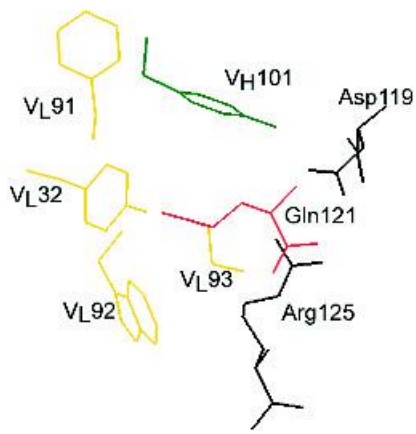


Figure 3.55 Interaction between Gln121 and residues formed by a pocket on the surface of the antibody in a complex formed between lysozyme and a monoclonal antibody

by the symbols κ and λ. The heavy chains are called α, δ, ε, γ and μ by direct analogy to the parent protein (Table 3.6).

As a consequence of remarkable binding affinities exemplified by dissociation constants between $10^4 - 10^{10}$ M antibodies have been used extensively as 'probes' to detect antigen. In the area of clinical diagnostics this is immensely valuable in detecting infections and many diseases are routinely identified via cross-reaction between antibodies and serum containing antigen. Antibodies are routinely produced today by injecting purified protein into a subject animal. The animal recognizes a protein as foreign and produces antibodies that are extracted and purified from sera. One extension of the use of antibodies as medical and scientific 'tools' was pioneered by César Milstein and Georges Köhler in the late 1970s and has proved popular and informative. This is the area of monoclonal antibody production, and although not described in detail here, the methods are widely used to identify disease or infection, to specifically identify a single antigenic site and increasingly as therapeutic agents in the battle against cancer and other disease states.

The humoral immune response involves the secretion of antibodies by B cells and the aggregation of antigen. Aggregation signals to macrophages that digestion should occur along with the destruction of pathogen. The cellular immune response to antigen involves different cells and a very different mechanism. On the surface of cytotoxic T lymphocytes sometimes more evocatively called 'killer T cells' are proteins whose structure and organization resembles the F_{ab} fragments of IgG. T cells identify antigenic peptide fragments that are bound to a surface protein known as the major histocompatibility complex

Table 3.6 Chain organization within the different Ig classes together with their approximate mass and biological roles

Class	Heavy chain	Light chain	Organization	Mass (kDa)	Role
IgA	α	κ or λ	$(\alpha_2\kappa_2)_n$ or $(\alpha_2\lambda_2)_n$	360–720	Found mainly in secretions
IgD	δ	κ or λ	$\delta_2\kappa_2$ or $\delta_2\lambda_2$	160	Located on cell surfaces
IgE	ε	κ or λ	$\varepsilon_2\kappa_2$ or $\varepsilon_2\lambda_2$	190	Found mainly in tissues. Stimulates mast cells to release histamines
IgG	γ	κ or λ	$\gamma_2\kappa_2$ or $\gamma_2\lambda_2$	150	Activates complement system. Crosses membranes
IgM	μ	κ or λ	$(\mu_2\kappa_2)_5$ or $(\mu_2\lambda_2)_5$	950	Early appearance in immune reactions. Linked to complement system and activates macrophages

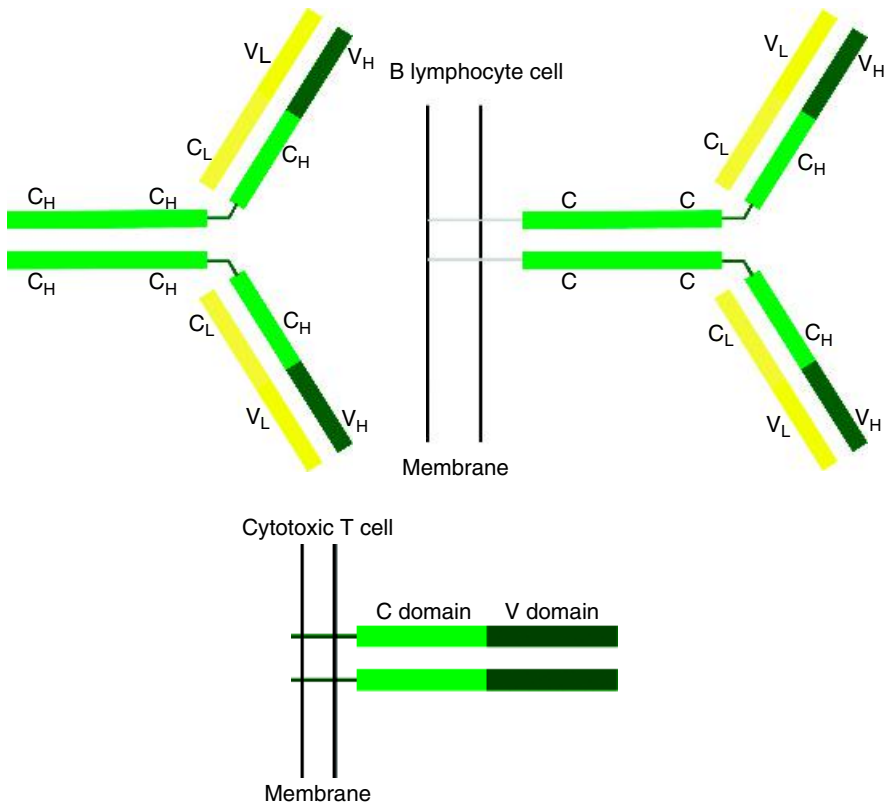


Figure 3.56 The use of the immunoglobulin fold in cell surface receptors by B and T lymphocytes

(MHC). Cytotoxic or killer T cells recognize antigen through receptors on their surface (Figure 3.56) and release proteins that destroy the infected cell. The MHC complex is based around domains carrying the immunoglobulin fold.

Cyclic proteins

Until recently it was thought that cyclic proteins were the result of unusual laboratory synthetic reactions without any counterparts in biological systems. This is now known to be untrue and several systems have been shown to possess cyclic peptides ranging in size from approximately 14 residues to the largest cyclic protein currently known a highly basic 70 residue protein called AS-48 isolated from *Enterococcus faecalis* S-48. Cyclic proteins are a unique example of tertiary structure but employ exactly the same principle as linear proteins with the exception that their ends are linked together (Figure 3.57). In view of the fact that many globular proteins have the N and C terminals located very close together in space there is no conceptual reason why the amino and carboxy terminals should not join together in a further peptide bond.

Cyclic proteins are distinguished from cyclic peptides such as cyclosporin. These small peptides have been known for a long time and are synthesized by microorganisms as a result of multienzyme complex reactions. The latter products contain unusual amino acid residues often extensively modified and are not the result of transcription. In contrast, cyclic proteins are known to be encoded within genomes and appear to have a broad role in host defence mechanisms.

Many cyclic proteins are of plant origin and a common feature of these proteins appears to be their size (~30 residues), a cyclic peptide backbone coupled with a disulfide rich sequence containing six conserved cysteine residues and three disulfide bonds. Unusually the disulfides cross to form a knot like arrangement and the cyclic backbone coupled with the cysteine knot has lead to the recognition of a new structural motif called the CCK motif or cyclic cysteine knot. The term cyclotides has been applied to these proteins. A common feature of all cyclotides is their derivation from longer precursor proteins in steps that involve both cleavage and cyclization. Although the gene sequences for the precursors are known the putative cleaving and cyclizing enzymes have not yet been reported. Many of these proteins have assumed enormous importance with the demonstration that several are natural inhibitors of enzymes such as trypsin whilst others are seen as possible lead compounds in the development of new pharmaceutical products directed against viral pathogens, and in particular anti-HIV activity. Whilst the cyclotide family (see Table 3.7) appear to have a common structural theme based around the CCK motif and the presence of three β strands other cyclic proteins such as bacteriocin AS-48 contain five short helices connected by five short turn regions that enclose a compact hydrophobic core. Despite different tertiary structure, all of the cyclic proteins are characterized by high intrinsic stability (denaturation only at very high temperatures) as well as resistance to proteolytic degradation.

Summary

Proteins fold into precise structures that reflect their biological roles. Within any protein three levels of organization are identified called the primary, secondary and tertiary structures, whilst proteins with more than one polypeptide chain exhibit quaternary levels of organization.

Primary structure is simply the linear order of amino acid residues along the polypeptide chain from the N to C terminals. Long polymers of residues cannot fold into any shape because of restrictions placed on conformational flexibility by the planar peptide bond and interactions between non-bonded atoms.

Conformational flexibility along the polypeptide backbone is dictated by ϕ and ψ torsion angles. Repetitive values for ϕ and ψ lead to regular structures known as the α helix and β strand. These are elements of secondary structure and are defined as the spatial arrangement of residues that are close together in the primary sequence.

The α helix is the most common element of secondary structure found in proteins and is characterized by dimensions such as pitch 5.4 Å, the translation distance 1.5 Å, and the number of residues per turn (3.6). A regular α helix is stabilized by hydrogen bonds

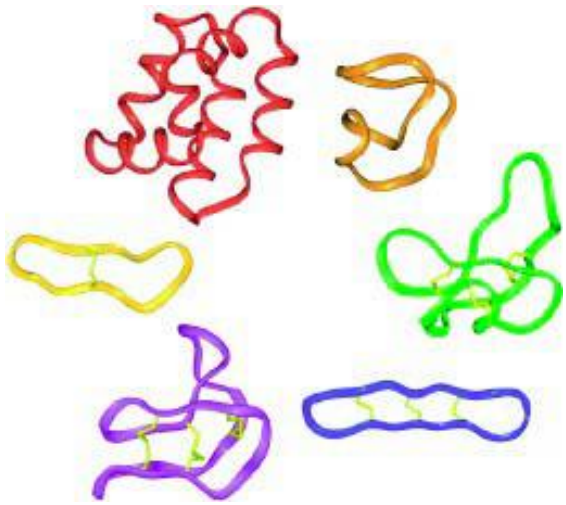


Figure 3.57 The topology of naturally occurring circular proteins. Three-dimensional structures of naturally occurring proteins. Clockwise from the upper left the proteins are: bacteriocin AS-48 (PDB:1E68) from *Enterococcus faecalis*, microcin J25 (PDB:1HG6) from *E. coli*, MCoTI-II (PDB:1HA9, 1IB9) from bitter melon seeds, RTD-1 (PDB:1HVZ) from the leukocytes of Rhesus macaques, kalataB1 (PDB:1KAL) from several plants of the *Rubiaceae* and *Violaceae* plant families, and SFTI-1 (PDB:1SFI, 1JBL) from the seeds of the common sunflower. Disulfide bonds are shown in yellow. (Reproduced with permission from Trabi, M., & Craik, D.J. *Trends Biochem. Sci.* 2002, **27**, 132–138. Elsevier)

orientated parallel to the helix axis and formed between the CO and NH groups of residues separated by four intervening residues.

In contrast the β strand represents an extended structure as indicated by the pitch distance ∼7 Å, the translation distance 3.5 Å and fewer residues per turn (2). Strands have the ability to hydrogen bond with other strands to form sheets – collections of β strands stabilized via inter-strand hydrogen bonding. Numerous variations on the basic helical and strand structures are found in proteins and truly regular or ideal conformations for secondary structure are rare.

Tertiary structure is formed by the organization of secondary structure into more complex topology or folds by interaction between residues (side chain and backbone) that are often widely separated in the primary sequence.

Several identifiable folds or motifs exist within proteins and these units are seen as substructures within a protein or represent the whole protein. Examples include the four-helix bundle, the β barrel, the β helix, the HTH motif and the β propeller. Proteins can now be classified according to their tertiary structures and this has led to the description of proteins as all α, α + β and α/β. The recognition that proteins show similar tertiary structures has led to the concept of structural homology and proteins grouped together in related families.

Tertiary structure is maintained by the magnitude of favourable interactions outweighing unfavourable ones. These interactions include covalent and more frequently non-covalent interactions. A covalent bond formed between two thiol side chains results in a disulfide bridge, but more common stabilizing forces include charged interactions, hydrophobic forces, van der Waals interactions and hydrogen bonding. These interactions differ significantly in their strength and number.

Quaternary structure is a property of proteins with more than one polypeptide chain. DNA binding proteins function as dimers with dimerization the result of specific subunit interaction. Haemoglobin is the classic example of a protein with quaternary structure containing 2α,2β subunits.

Proteins with more than one chain may exhibit allostery; a modulation of activity by smaller effector molecules. Haemoglobin exhibits allostery and this is shown by sigmoidal binding curves. This curve is described as cooperative and differs from that shown by myoglobin. Oxygen binding changes the structure of one subunit facilitating the transition from deoxy to oxy states in the remaining subunits. Historically the study of the structure of haemoglobin provided a platform from which to study larger, more complex, protein structures together with their respective functions.

One such group of proteins are the immunoglobulins. These proteins form the body's arsenal of defence mechanisms in response to foreign macromolecules and are collectively called antibodies. All antibodies are based around a Y shaped molecule composed of two heavy and two light chains held together by covalent and non-covalent interactions.

Table 3.7 Source, size and putative roles for cyclotides characterized to date

Protein	Source	Size (residues)	Role
SFTI-1	*Helianthus annuus*	14	Potent trypsin inhibitor
Microcin J25	*Escherichia coli*	21	Antibacterial
Cyclotide family	*Rubiaceae* and *Violaceae* sp	28–37	Wide range of activities ~45 proteins
McoTI-I and II	*Momordica cochinchinensis*	34	Seed-derived trypsin inhibitor
Bacteriocin AS-48	*Enterococcus faecalis*	70	Hydrophobic antibacterial protein
RTD-1	*Macaca mulatta*	18	Antibiotic defensin from primate leukocytes

Antigen binding sites are formed from the hypervariable regions at the end of the heavy and light chains. These hypervariable regions allow the production of a vast array of different antibodies within five major classes and allow the host to combat many different potential antigens.

The basic immunoglobulin fold is widely used within the immune system with antibody-like molecules found as the basis of many cell surface receptors particularly in cells such as helper and killer T cells and parts of the MHC complex.

Problems

1. Draw a diagram of a typical polypeptide backbone for a pentapeptide. Label on your diagram the following; the α carbon, the side chains, use a box to define the atoms making up the peptide bond and finally identify the torsion angles ϕ and ψ and the atoms/bonds defining these angles.
2. Show how the above pentapeptide changes when the third residue is proline.
3. Poly-lysine and poly-glutamate can switch between disordered structures and helical structures. What conditions might promote this switch and how does this drive formation of helical structure?
4. Using the bond lengths given for C–C and C–N bonds in Chapter 2 together with the dimensions found in α helices and β strands calculate the length of (i) a fully extended polypeptide chain of 150 residues, (ii) a chain made up entirely of one long regular α helix and, (iii) a chain composed of one long β strand. Comment on your results.
5. From Figure 3.2 use the primary sequences of myoglobin to define the extent of each element of helical secondary structure. Explain your reasoning in each case.
6. What limits the conformational space sampled by ϕ and ψ?
7. Some proteins when unfolded are described as a random coil. Why is this a misleading term?
8. List the interactions that stabilize the folded structures of proteins. Rank these interactions in terms of their average ‘energies’ and give an example of

each interaction as it occurs between residues within proteins.

9. How would you identify turn regions within the polypeptide sequence of proteins? Why do turn regions occur more frequently on protein surfaces? What distinguishes a turn from a loop region?

10. What are the advantages of more than one subunit within a protein? How are multimeric proteins stabilized?