Molecular Conformations

Christopher Wood

bookboon
The eBook company

Christopher Wood

Molecular Conformations

Molecular Conformations
1st edition

ISBN 978-87-7681-545-5

Contents

1 Genetic origins of amino acids

1.1 Constituents and organisation of DNA

DNA (deoxyribonucleic acid) is the genetic coding scheme used by all living organisms to pass on the hereditary programme to the next generation. DNA consists of two cross-linked polynucleotide chains, having an overall length of about 2 metres, which, for eukaryotes, is stored within the nucleus of a cell. The nucleus occupies about 10% of the cell volume and is isolated from the cytoplasm by the nuclear envelope, which consists of inner and outer bi-layer lipid membranes. This double membrane is interspersed with a significant number of pores that allow bi-directional transport of molecules to take place between the nucleus and cytosol. The nucleus contains the vast majority of DNA in a cell, the remainder being within the cytoplasmic mitochondria.

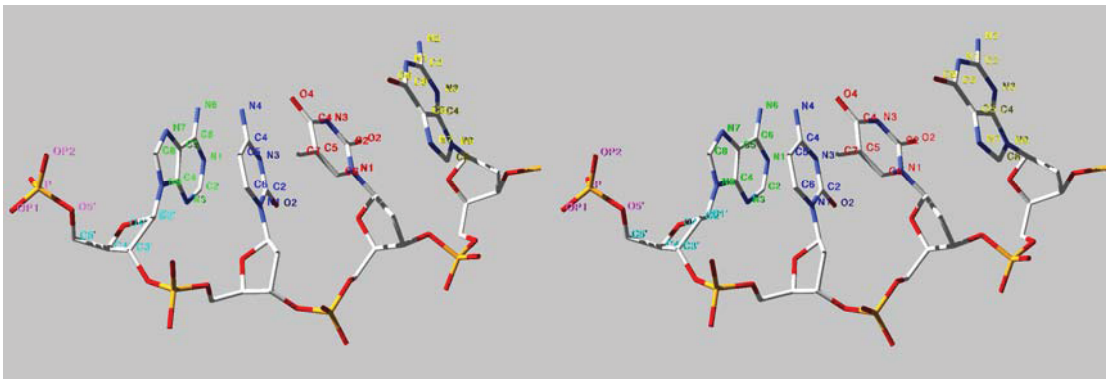


Figure 1.1: The four DNA bases. A stereo image of the four DNA bases of guanine, cytosine, adenine and thymine. Atom colour code: yellow – phosphorous; grey – carbon; blue – nitrogen and red – oxygen. Lettering colour code: magenta – phosphate; cyan – sugar (deoxyribose); thymine – red; guanine – yellow; cytosine – blue and adenine – green.

The polynucleotide backbone of DNA is made up of a repeating pattern of sugar-phosphate. The sugar is a 5-carbon variant known as deoxyribose. Attached to each sugar is a sidechain known as a base, of which there are four types: adenine (A), cytosine (C), guanine (G) and thymine (T). The sugar and phosphate groups of the DNA backbone are linked by covalent bonds (**Fig 1.1**). The DNA is arranged such that two identical strands are arranged in an anti-parallel direction, meaning that if you imagine one strand as going from left to right, then its facsimile will be going from right to left.

To orientate the strands, we make use of the carbons of the deoxyribose sugar. Since the C5 and C3 carbons of the deoxyribose sugar link with neighbouring phosphates, it is possible to talk about a strand being in a C5–C3 direction; the facsimile would then be in a C3–C5 direction. However, this nomenclature is not used and the standard format is to use 5′ (read as 5-prime) and 3′. Any biological system is subject to the laws of thermodynamics. For that reason, these two contra-directional strands of DNA will seek out their minimum energy configuration, which results in the three-dimensional double-helix structure.

In the double-helix the base sidechains are on the inside, with the sugar-phosphate backbone on the outside. The A, C, G and T bases belong to two chemical families known as purine and pyrimidine; A and G belong to the former, with C and T in the latter family. Since the purines are larger than the pyrimidines, it is not possible to have a pyrimidine facing another pyrimidine on the opposing strand of DNA, and still maintain the minimum energy requirement that both strands be parallel; a similar argument goes for two purines. Thus, a purine will always, by complementary base-pairing, pair with a pyrimidine, such that A pairs with T and G pairs with C.

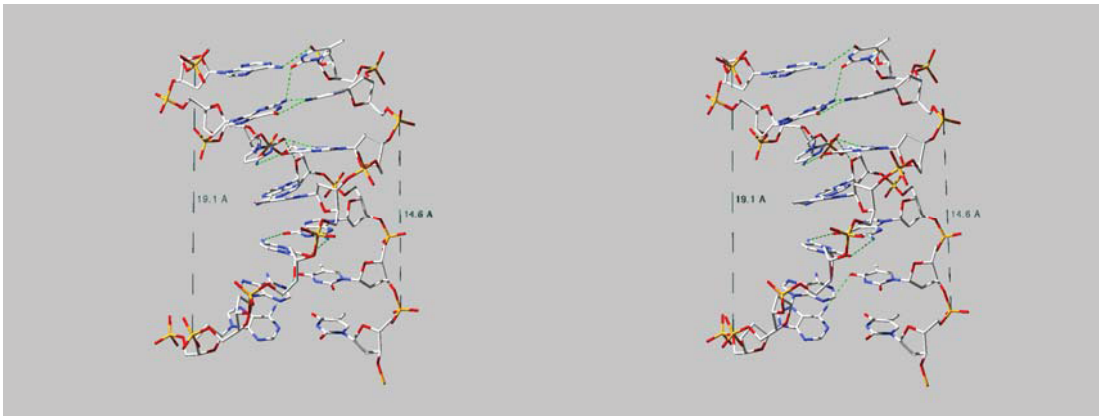


Figure 1.2: Helical Structure of B-DNA. Shown is a stereo image of a section of B-DNA helix. On the left side is the major groove and on the right the minor groove, shown with lengths. One helical turn is about 3.4 nm – here it is 3.7 nm. The colour code is: light green – hydrogen bonds; dark green – distances; yellow phosphorous; red – oxygen; grey – carbon and blue – nitrogen.

1.2 Folding of DNA

The base-pairs within DNA are bound together by hydrogen bonds. Though individually weak, hydrogen bonds are numerous and therefore have the required strength to laterally bind the two opposing strands of DNA. On the proximal and distal sides of the DNA double-helix is a pattern of large-small grooves, more commonly known as major and minor grooves. The length of one helical turn on the DNA double-helix is 3.4 nm and accounts for about ten base-pairs (**Fig 1.2**). If you were to look down the length of a DNA double-helix, it would be seen that the helix extends away from you in a clockwise direction. For that reason, it is known as a right-handed helix.

The pattern of base-pairs within the DNA double-helix contains the program for amino acid and protein synthesis – described in more detail in later sections. Although the two opposing strands of DNA are able to fold themselves into a double-helix, it is not possible to achieve further folding without assistance. In order to get about 2m of DNA into the nucleus, a very high degree of folding is required. The next level of folding is achieved by the use of structural proteins, around which the DNA wraps itself. The proteins that serve this function are collectively known as the *histone* family. These proteins form into a structure known as the *histone octamer* and it is this structure that is fundamental to the next-level folding of the DNA. The combination of DNA and the histone structural proteins is known as *chromatin*.

Around the histone octamer is wrapped 1.65 turns of superhelical DNA, which encompasses 147 base-pairs. This unit is known as the *nucleosome core particle* (NCP). Added to each NCP is a length of H1 linker histone which has an extra length of DNA (about 50 base-pairs) associated with it. The combination of the NCP and the H1 linker histone is known as the *nucleosome* and it is this basic unit that is repeated to achieve folding of the DNA (**Fig 1.3**).

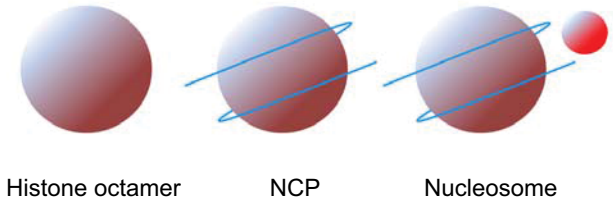


Figure 1.3: Histone proteins and DNA. The histone octamer forms a roughly spherical shape around which are wrapped superhelical coils of DNA, to form the NCP. Finally, the linker histone and some extra DNA (not shown for clarity) form a nucleosome. Consecutive nucleosomes form chromatin. Colours are: purple – histone octamer; blue – DNA and red – linker histone.

The histone octamer itself is formed from two duplicated sets of four histone proteins: H2A, H2B, H3 and H4. These eight histone proteins are formed into the central tetramer (H4–H3)–(H3′–H4′) that is flanked at each side by a (H2A–H2B) dimer to give an overall structure (H2A–H2B)–(H4–H3)–(H3′–H4′)–(H2B′–H2A′).

1.3 The codon structure of DNA

The DNA in the cell nucleus is decoded to form the amino acids that make up proteins. The mechanics of this decoding process are the subject of a later section. There are 20 naturally-occurring amino acids. This number of amino acids cannot be synthesised from one base; instead, the bases are synthesised in groups. The minimum number of bases required to produce twenty amino acids is 3; in fact 3 bases gives 4^3 code words. A code word of three bases is known as a codon. The 64 codons are arranged such that 61 of them encode for amino acids; the remaining three are defined as stop codons, the meaning of which will become apparent. From Table 1.1, it is clear most amino acids have more than one codon, with serine, leucine and arginine having six. The several codons for a particular amino acid are said to be synonymous. Codons form part of a larger DNA sequence known as a gene. Codons will be re-visited when the process of transcription is considered.

Second Codon		T		C		A		G	
First Codon	T	Phe	T	Ser	T	Tyr	T	Cys	T
		Phe	C	Ser	C	Tyr	C	Cys	C
		Leu	A	Ser	A	STOP	A	STOP	A
		Leu	G	Ser	G	STOP	G	Trp	G
	C	Leu	T	Pro	T	His	T	Arg	T
		Leu	C	Pro	C	His	C	Arg	C
		Leu	A	Pro	A	Gln	A	Arg	A
		Leu	G	Pro	G	Gln	G	Arg	G
	A	Ile	T	Thr	T	Asn	T	Ser	T
		Ile	C	Thr	C	Asn	C	Ser	C
		Ile	A	Thr	A	Lys	A	Arg	A
		Met	G	Thr	G	Lys	G	Arg	G
	G	Val	T	Ala	T	Asp	T	Gly	T
		Val	C	Ala	C	Asp	C	Gly	C
		Val	A	Ala	A	Glu	A	Gly	A
		Met/Val	G	Ala	G	Glu	G	Gly	G

Table 1.1: The 3-letter codon codes for amino acids (DNA form).

1.4 Gene structure

All protein-coding genes have essentially similar structures. The parts of the gene that are required to form a functional protein are known as exons. An interim step on the way to protein production is mRNA, which is produced by scanning the DNA. It is now known that there are also a large number of genes which, when transcribed, produce RNA. These RNA-coding genes are currently the subject of intensive research. The gene will have a transcription start site at the 5′-end, where there is a core (or minimal) promoter. There may be additional promoters, known as regulatory promoters, that reside upstream of the core promoter; their function is to allow the binding of transcription factors that will determine the correct expression of the protein.

In addition, there may also be *cis*-acting regulatory elements that determine the correct spatio-temporal expression of the gene. These elements are known as enhancers and repressors and they may lie upstream or downstream of the transcription unit and in introns. They can be some distance away from the transcription unit – up to 1Mb upstream or downstream, and even lie in an adjacent gene. Their function is to assist in the timely suppression or activation of a gene. For example, if a gene is ubiquitously expressed (that is, expressed in all, or a significant number of organs), then it may well contain no regulatory elements. However, if a gene is to be expressed only in the kidney, then repressors will assist in keeping it silent, whilst kidney-specific enhancers will ensure correct and timely expression.

The *cis*-acting elements require the correct chromatin structure, which must be open (or euchromatic). Chromatin which is generally inaccessible to transcription factors is known as heterochromatin.

2 Transcription and translation

2.1 Transcription

The process of transcription involves the production of messenger RNA (mRNA) from DNA. Once the chromatin fibre has been opened up it is accessible to proteins and the process of transcription can then start. The protein which carries out this conversion process (for proteins) is known as RNA Polymerase II. Once a gene's promoter has been exposed, transcription factors will bind to the relevant promoters. This acts as a signal for other transcription-relevant proteins to congregate at these points.

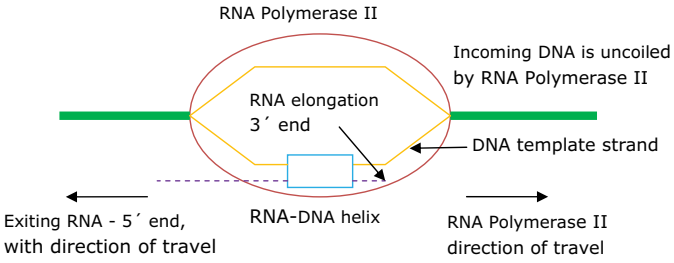


Figure 2.1: Transcription by RNA Polymerase II. The RNA Polymerase II transcription bubble must separate the two strands of DNA before RNA synthesis can begin. Colours are: green – coiled DNA; orange – uncoiled DNA; red – RNA Polymerase II; blue – transcription centre and purple – RNA.

Eventually, a transcription initiation complex is formed (**Fig 2.1**), including RNA Polymerase II and the process of mRNA production can begin. The RNA Polymerase II molecule then separates the two strands of DNA and slides along the transcription unit, producing a strand of mRNA. The exuded mRNA is a complementary copy of the DNA template strand and a duplicate of the non-template strand, such that a DNA cytosine produces a mRNA cytosine, and similarly for A and G. However, a DNA thymine is translated not as itself, but as a uracil (U). So, A, C, G and T in DNA becomes A, C, G and U, respectively, in mRNA. Scanning of a DNA strand by the RNA Polymerase II molecule takes place in a 5´ to 3´ direction.

2.2 Translation

2.2.1 Transfer RNA

When a viable mRNA has been produced, the second of three types of RNA comes into play. This is called transfer RNA (tRNA). tRNA plays a key role in the way the signal encoded in the mRNA codons is translated into protein, hence the process is named translation. In order to complete the translation process, tRNA must first associate with an amino acid and this is facilitated by an enzyme called aminoacyl-tRNA synthetase. Depending on the type of cell there can be 50–100 different types of tRNA molecules. There are 20 types of aminoacyl-tRNA synthetases (one for each amino acid). Thus, an aminoacyl-tRNA synthetase that associates with arginine, for example, will attach that arginine to all tRNAs that are able to recognise arginine codons. Once complexed with an amino-acid, the tRNA then has to locate and bind to mRNA using a specialised loop domain.

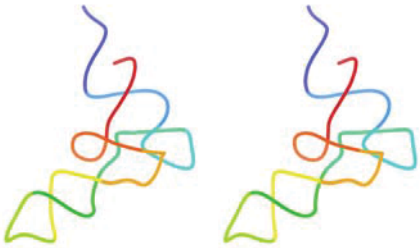


Figure 2.2: Structure of tRNA. Stereo image of tRNA derived from Protein Database entry 1EVV. The structure of tRNA consists of the 5′ and 3′ ends and three loops: DHU, TψC and anticodon. The 5′ end is capped by a phosphate molecule. The 3′ end is capped by a OH molecule and is the point where an amino acid would attach. The anticodon loop will bind to the mRNA. Colours are: orange – DHU loop; cyan – TψC loop; light green – anti-codon loop; red – 5′ end and blue – 3′ end.

A tRNA molecule is 70–80 nucleotides in length and takes the form of a stem-loop configuration (**Fig 2.2**), there being four stems and three loops; one stem is open-ended, being made up of the 5′ and 3′ ends. A specific aminoacyl-tRNA synthetase will catalyse a reaction between a tRNA molecule, an amino acid and ATP, the resulting product being a tRNA molecule covalently bound to an amino acid. The covalent bond is formed between the carboxyl group of the amino acid and the 2′- or 3′-hydroxyl group of an adenosine at the 3′ end of the tRNA molecule. In the anti-codon loop of the tRNA molecule are three anti-codons that have the ability to recognise more than one codon of mRNA.

This capacity for multiple-codon recognition occurs because the tRNA anti-codons employ non-standard base pairing. The third and second bases of the tRNA anti-codon form standard Watson-Crick pairing with the first and second bases in the mRNA codon, respectively. Through non-standard base-pairing, a G as the first base in the anti-codon will recognise a C or U in the mRNA codon, and a U in the said position will recognise an A or G in the mRNA codon.

2.2.2 Protein formation by ribosomes

A ribosome in a eukaryote consists of the large 60S and small 40S subunits. The former contains three ribosomal RNAs (rRNA): the 28S, 5.8S and 5S; the latter a single rRNA: 18S. Both subunits contain a number of ribosomal proteins. All of these elements each have their own well-defined role in the assembly of the pre-initiation complex. The previously-assembled pre-initiation complex is made up of the 40S subunit of the ribosome to which is bound the initiator tRNA-Met (tRNA plus a methionine), along with a molecule of GTP. Attached to the initiator tRNA-Met is eIF-2. This molecule belongs to a family of proteins known as eukaryotic initiation factors (eIF). There are many of these factors involved in the formation of the pre-initiation complex, but only two (eIF2 and eIF4E) are considered herein.

Once formed, the pre-initiation complex attaches to the 5′-end of the mRNA molecule, so that protein assembly may start. To assist with attachment, a molecule called the cap-binding complex will bind to the 5′-end. This molecule is comprised of several eIFs, one of which is eIF4E. eIF4E is the factor that the pre-initiation complex binds to. The initiation complex now scans along the mRNA to the start codon, using hydrolysis of ATP to do so. It does this through the helicase activity of various eIFs; these molecules are able to effectively unwind the mRNA. The initiation complex will, in eukaryotes, be looking for the AUG start codon. Biochemical recognition is facilitated by the fact that the start sequence lies in an ACCAUGG consensus sequence, known as a Kozak consensus. Identification of this location acts as a signal for the large subunit of the ribosome to attach, using hydrolysis of GTP.

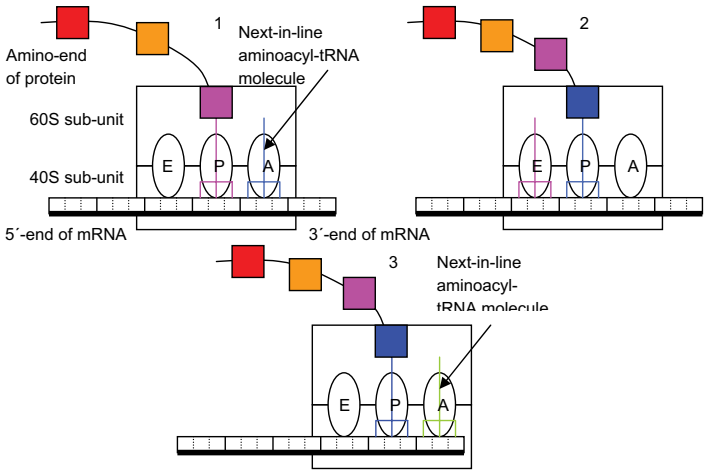


Figure 2.3: Translation by ribosome. The addition of an amino acid to the chain involves three steps. 1 – An appropriate aminoacyl-tRNA binds at the A site of the ribosome. 2 – A peptide bond is formed and the recently-arrived aminoacyl-tRNA moves to the P site, whilst the former P-site occupant is moved to the E site. 3 – The E-site occupant is now ejected from the ribosome and another aminoacyl-tRNA occupies the A site, but only after the ribosome has shuffled along to the next mRNA codon. Colours are: red, orange, magenta and blue – exuded amino acids; green – amino acid being primed.

The now complete ribosome starts to work its way down the mRNA from the start codon; this process is known as *translation elongation*. On the ribosome there are two binding sites for aminoacyl-tRNAs; they are the P and A sites (**Fig 2.3**). The P site is where the initiator tRNA-Met currently resides. The A site sits over the second codon in the open reading frame (considered in the next section), and this is where the next aminoacyl-RNA will locate. Eukaryotic elongation factors (eEF) assist the ribosome in moving along the mRNA molecule.

The next aminoacyl-tRNA is brought to the A site in eukaryotes by the eEF-1 elongation factor and a molecule of GTP. The tRNA molecule is deposited and the eEF-1 elongation factor moves away. The GTP molecule is hydrolized to GDP and then moves away; the energy liberated is used to form a peptide bond between the initiator tRNA-Met and the newly-placed second aminoacyl-tRNA, with the help of a peptidyl transferase enzyme. The GDP molecule also then moves away. The next phase of the process involves translocation. In this process the ribosome moves along three nucleotides so the next codon is placed within the A site. The dipeptide that was in the A site moves to the P site, such that the deacetylated tRNA that was in the P site moves to the E site, or exit site, from where it departs the ribosome. Translocation requires the services of eEF-2 and GTP, which is again hydrolized to GDP. The elongation process then repeats itself until a mature protein emerges.

2.2.3 Open reading frames of mRNA

To initiate protein synthesis, it is necessary to define where the string of bases to be decoded starts and stops; this is done by the use of start and stop codons. The codon for methionine is used as the start codon; decoding stops when one of three stop codons is encountered. Such a stretch of nucleotide bases is known as an open reading frame. There are, potentially, three open reading frames, available to the decoding apparatus, as a given base can be in the first, second or third position of a particular codon (**Fig 2.4**). Although in the first reading frame no stop codon is visible (**Fig 2.4**), one has suddenly appeared in the second reading frame. The consequence of the second reading frame being processed is that a protein will still be synthesised, but the untimely occurrence of the stop codon will cause a sequence-shortened, non-functional protein to be generated.

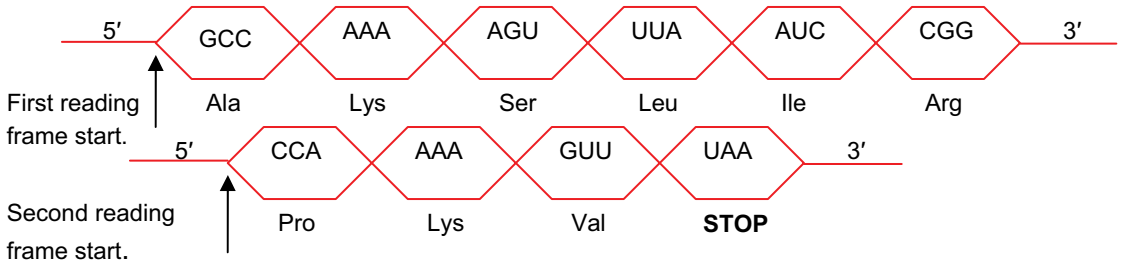


Figure 2.4: Reading frames of mRNA. The same strand of mRNA is shown as having two different upstream sites for the start of translation (not shown). The result is that the same strand can be translated in two ways, to give two open-reading frames. The first gives a functional protein, but the second produces a stop codon, which will result in a non-functional protein. Colour: red – mRNA.

3 Amino acid geometries and protein folding

3.1 Amino acid structure and bond flexibility

We have seen how DNA is used as a template for the production of mRNA and then how that mRNA is utilized in the production of amino acids and, ultimately, a functional protein. It is now that we can start to look at mechanisms important to conformational changes in a protein. Conformational change refers to fluctuations in the 3-dimensional shape of a protein. The degree of conformational change ranges from small to significant. However, in both cases it is the protein's aim to enhance or decrease its affinity for other molecules, as the case may be.

An amino acid is made up of four elements (**Fig 3.1**): a) the C_α carbon; b) bound to the C_α carbon are H_2N and COOH groups, known as the amino and carboxy terminals, respectively; c) a hydrogen is also bound to the C_α carbon and d) a general R group, where R is known as the sidechain and depends on the amino acid. At pH7 the amino and carboxy groups become $+H_3N$ and COO-, respectively. The carboxy terminal (the α-carboxy group) of one amino acid will bind the amino terminal (the α-amino group) of another, to form a peptide bond – less frequently called an amide bond. This process continues to form a peptide chain, with the creation of one peptide bond resulting in the release of one water molecule. The fact that a water molecule is released means that the two amino acids are "down-sized", such that each is now known as an *amino-acid residue* (simply referred to as *residue*).

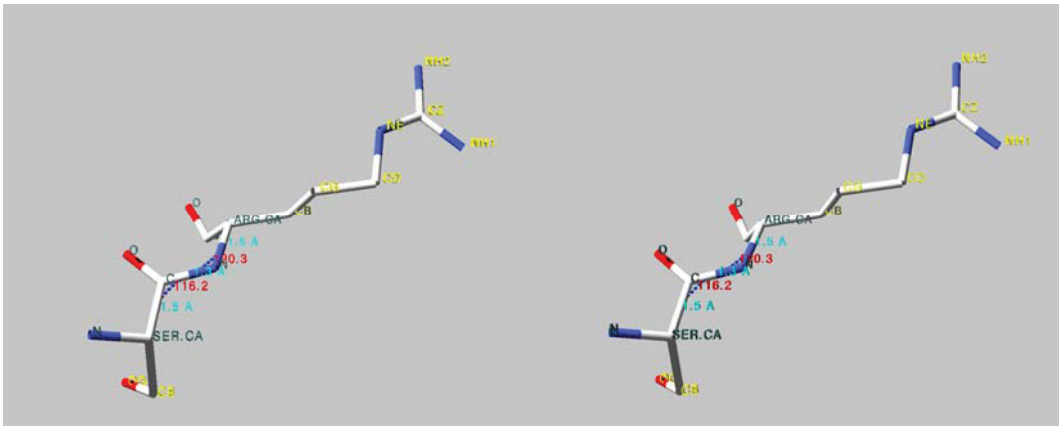


Figure 3.1: Distances and angles of the peptide backbone. This stereo image shows a serine and arginine. Atom colours are: grey – carbon; blue – nitrogen and red – oxygen. Letter colours are: green – atoms of the peptide backbone; yellow – sidechains of the amino acids; cyan – bond lengths and red – angles (with angle marked by a dotted line).

A completed peptide chain has an associated polarity, with an α–amino group at one end and an α–carboxy group at the other. Convention has the amino end as the start of the peptide chain, and the carboxy end as the terminal. The peptide chain, or backbone, has the ability to make numerous hydrogen bonds. Each residue has a carbonyl group that can act as a hydrogen bond acceptor and, with the exception of proline, a NH group that makes a good donor. These hydrogen bonds (see section 3.2) become important in the formation of secondary-structure elements such as α–helices and β–sheets. The peptide bond is covalent in nature and has a certain amount of flexibility associated with it; this becomes important in protein folding. The bonds of the sidechains also have a certain amount of flexibility and, as will be seen, this gives rise to different sidechain configurations.

A covalent bond can be defined in terms of its geometry, with the bond length (L), the bond angle (κ) and the torsion angle (θ). The bond length is measured from the centres of two relevant atoms; the bond angle is that angle formed by three co-planar atoms and two bonds; the torsion angle is the amount of rotation in a bond around some axis. The single bond is formed from a single electron pair – one from each atom – that fills the outer shell of both contributing atoms. The bond is 1.52 Å long (C-C), or 1.45 Å (C-N) and can rotate. A double bond is formed from two electron pairs. Though not relevant to the peptide backbone, double bonds do occur in residue sidechains. The double bond cannot rotate and has a fixed torsion angle; its length is 1.33 Å (C=C) or 1.38 Å (C=N). The disulphide bond occurs between two cysteine amino acid residues when their sulphur atoms come into close proximity. The bond is strong and can dramatically reduce protein flexibility. However, the most important covalent bond is the peptide bond of the protein backbone.

Whilst the length of a peptide bond is important, it allows for the least variation. As peptide bonds are neither single or double bonds, but have a partial double-bond structure, they tend to jump between the two states, and this phenomenon is known as resonance. In the first state there is a single bond between the C′ and its corresponding O and a double bond between C′ and the following residue's N. The second state is a double bond between C′ and O and a single bond between C′ and N. This partial double-bond status means that a C′-N peptide bond is shorter than a standard C-N single bond and longer than a double C=N bond. Two configurations are possible for a planar peptide bond. In the *trans* configuration, the two α-carbon atoms are on opposite sides of the peptide bond. In the *cis* configuration, these groups are on the same side of the peptide bond. The *trans* peptide bond is the most common, occurring in the overwhelming majority of cases. The creation of a *cis* peptide bond is hindered by potential steric clashes.

The peptide bond angles are likewise unvarying at 109.5° for the N-C_α-C′ planar angle, 116° for the C_α-C′-N angle and 122° for the C′-N-C_α angle. However, it is the torsion angle that allows the greatest degree of freedom, and, as such, is the most important parameter. A peptide bond has three torsion angles that have been designated as φ (phi), ψ (psi) and ω (omega). φ is the rotational angle between the N and C_α atoms, ψ is the rotational angle between the C_α and C′ atoms and ω is the rotational angle between C′ and the N of the following residue, and is restricted to the values of 180° and 0°. This limitation means that ω plays little part in a protein's three-dimensional form, this being primarily determined by φ and ψ. The φ and ψ angles are able to move to a much greater extent and, though having that ability, tend not to be uniformly distributed over their range of angles. Rather, they tend to congregate in frequently-occurring φ – ψ pairs. The φ – ψ pairings for the amino-acid residues in a protein, when plotted on a graph with φ on the y-axis and ψ on the x-axis, will clearly be seen to congregate in particular areas for that protein. If this is done for a set of proteins, it will be seen that the φ – ψ pairings congregate in similar areas for all proteins. This fact was discovered by G. N. Ramachandran, and such graphs are now known as Ramachandran plots.

The sidechains of nineteen of the twenty amino acids (strictly-speaking, a glycine has no sidechain – only a hydrogen atom) do not just exist in one set of geometries. Instead, the sidechains can adopt a large number of configurations relative to the backbone. Each configuration is often called a *rotamer*. Whilst rotamer libraries will have rotamers for most of the amino acids, they will have none for glycine and either none or one for alanine. The carbon atoms of the sidechain are labelled in progression of Greek letters from the alpha-carbon, such that a lysine has C_α, C_β, C_γ, C_δ, C_ε. This is the largest number of carbons of all the amino acids. Between each pair of carbons there exists a torsion angle, such that lysine has four. The torsion angles of amino acid sidechains are assigned the Greek letter χ (chi), with a number subscript, so that in the case of lysine the torsion angles are χ_1, χ_2, χ_3, χ_4. To be defined as a rotamer, all the chi angles usually have to be within $\pm 2\Omega$ of the mean angles for that particular rotamer.

We have seen how the peptide backbone is flexible and therefore allows protein folding to take place, and for that reason alone it may seem that the issue of sidechain rotamers is a poor cousin of the former. However, there are occasions in molecular biology when it is very important to establish that you have the correct rotamer. Probably the most important of these is the rotamers of amino acids located in binding domains (the place on one protein where another binds). It is essential that in such an area a sidechain is in the correct orientation.

3.2 Principles of protein folding

3.2.1 Formation of secondary-structure elements

In section 3.1 the ability of a peptide backbone to vary its geometry was discussed. This allows a high degree of flexibility in protein folding and the first step in that process is the creation of secondary-structure elements. Protein folding is the process a chain of amino acid residues undergoes to form the 3-dimensional shape of the protein. Secondary structure is the next hierarchical structure layer, coming above the amino acid sequence. Above this level are the tertiary and quarternary structure levels, the former being the highest level that an individual protein can fold to, the latter arising from an amalgamation of two tertiary structures.

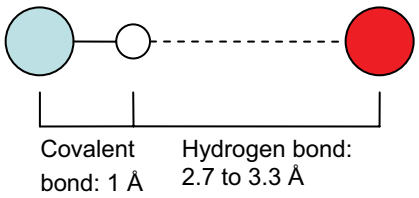


Figure 3.2: Hydrogen bond structure. The hydrogen bond is an important bond in protein structure. Although weak in terms of energy, the many bonds that occur in a protein are such as to ensure that this bond accounts for most of the energy in a protein. Colours are: cyan – nitrogen; white – hydrogen; red – oxygen.

The most common secondary structures are helices and sheets of various types. Whatever the type of secondary structure, no new covalent bonds are created intra-structure during the formation of these elements. Similarly, essentially no new covalent bonds are created inter-structure, save for exceptions such as a disulphide bond. Instead, the formation of these secondary structures relies on the formation of hydrogen bonds. A hydrogen bond will form between a hydrogen atom that is covalently-bound to an oxygen or nitrogen atom and another oxygen or nitrogen that has an unshared electron pair (**Fig 3.2**). A C-H bond is not sufficiently polar to form a hydrogen bond. A S-H bond can form weak hydrogen bonds.

When a secondary structure forms, the atoms of the backbone will come closer together and form numerous hydrogen bonds. Though individually weak, the sheer number of hydrogen bonds ensures that there is sufficient energy to maintain these structures. The α–helix is the most common type of secondary structure, where hydrogen bonds form between the amide nitrogen of amino acid residue *n* and the carbonyl oxygen of residue (*n* + *4*). About 98% of all helices are of the α–type (**Table 3.1**), the remainder being the 3-10 helix and the π–helix. We can characterize these helix types by their φ and ψ angles. All the helix types have associated with them a right-handed chirality. This means that as one looks down the length of the helix, it will be seen to move away from the viewer in a clockwise direction. Conversely, left-handed helices, which are even rarer, advance in an anti-clockwise fashion. Generally, α–helices are straight, but when a proline is present it puts a kink in the helix. The α–helix has a rise of 1.5 Å and has 3.6 residues per turn – meaning its pitch is 5.4 Å.

Helix Form	φ/ψ Angle (°)	Frequency (%)	H-Bond Span	Chirality
α	-57/-47	98	4	Right
3-10	-49/-26	1	5	Right
π	-57/-80	1	3	Right

Table 3.1: Frequency of helix types.

In the β–strand structure, adjacent residues are separated by a distance of 3.5 Å rather than 1.5 Å, and the sidechains of the residues point in opposite directions. However, the β–strand is not a secondary-structure element, and it will interact with another β–strand to form a β–sheet. A β–sheet can take one of two forms, both formed by hydrogen bond cross-linking of the two strands. The two strands can run in the same or opposite directions, known as parallel or anti-parallel β–sheets, respectively. The hydrogen bonding arrangement for the parallel structure is formed by the NH of one residue (A_n) being bonded to the CO group of a residue on the opposite strand (B_n), but the CO group of residue A_n is bound to the NH group of residue $B_{(n+2)}$. The anti-parallel case is much simpler, where the A_n NH and CO groups are bound to the CO and NH groups of B_n, respectively. The anti-parallel β–sheet is by far the most common of the two types. For the case of the anti-parallel β–sheet, the φ and ψ values are –139° and 135°, respectively; while for the parallel β–sheet, the values are –119° and 113°, respectively.

Many texts, when talking about secondary-structure elements, will fail to mention the existence of loops. This is a mistake, because they are important in the area of protein conformational change – the very subject of this book. There are many types of loop, but they all have importance. Some may simply link two secondary-structure elements, but a few have the important property that they act as a hinge joint, opening or closing the protein's structure when a ligand (another molecule) binds, and others may have a role in binding that ligand. Smaller loops are often called turns, the most common of which is the β–hairpin, used to reverse direction of the peptide backbone. In this configuration it is common for the CO group of residue n to be hydrogen bonded to the NH group of residue $(n + 3)$; this serving to stabilize the turn.

Once these secondary-structure elements have formed they will then organize themselves into frequently occurring patterns called *motifs*. There are also more specialized patterns called *domains*. Domains are specifically involved with protein function.

3.2.2 Causes and consequences of misfolding

A protein must function properly once synthesized, with the natural inference that the protein should fold properly.

Improper folding is caused by problems at the genetic level. The genetic problems can be caused by errors at many stages in the course of protein processing, but usually it will be one (or more) of a) problems with the genetic source code, that is the DNA; b) problems in transcription and c) problems with translation. The end result is a disruption in the sequence of amino-acid residues and improper folding. At a genetic level, the aforementioned problems can give rise to, most commonly, insertions, deletions, missense mutations and nonsense mutations. An insertion and deletion is where one or more nucleotides are somehow inserted into the mRNA, or removed, respectively. A missense mutation results in an incorrect amino-acid residue. This is often abbreviated to, for example, Y257K – which is translated as residue number 257, which is ordinarily a tyrosine, is mutated to a lysine. A nonsense mutation means there is a stop codon present in a position where it should not be. The result of this is that the protein is synthesized up to that point and not beyond; in other words the protein is truncated, with consequential loss of function.

Many congenital disorders (defined as being present at birth) have a genetic basis, often involving one gene. Whilst most of these genetic-based syndromes have a mercifully low level of incidence, such as Rett syndrome (involving mutation of the MECP2 gene), others are more common. Genetic disorders can also give rise to conditions that occur later in life, of which there are many specific conditions, but perhaps the most publicised of these – and not just limited to adulthood – is cancer. One of the accepted causes of cancer is when an oncogene acquires a gain of function, or a tumour suppressor gene has a loss of function. It is now coming to light that most tumour suppressors seem to be haploinsufficient – meaning that to assert its proper tumour suppressor role, both alleles are required, and that if one of those alleles is missing, it will give rise to an abnormal phenotype.

3.2.3 Mechanisms of protein folding

Protein folding is one of the most intensively researched areas in the life sciences, with a continuous output of papers. The majority of the research is computer based, concerned with the development of models that simulate the folding process. The problem is that, at best, current computers can only simulate up to about 1 μs of protein folding time. This is a problem when folding times occur from milliseconds to seconds. However, the purpose of this sub-section is not to review something that is nothing less than a proliferation of models and algorithms, but to concentrate on the underlying hydrodynamic and thermodynamic principles.

A protein in the unfolded state is disordered and in the folded state it is ordered. Thus the folding process entails a reduction in entropy. This reduction in entropy must be balanced by an increase in entropy elsewhere. The key to protein folding lies in the properties of water. In bulk water (often referred to as bulk solvent), a water molecule can make up to four hydrogen bonds, two as donor and two as acceptor, as described earlier in this section. These bonds are constantly being broken and re-made, and therefore bulk water is in a highly disordered state. Water molecules that make hydrogen bonds with the surface of a protein tend to be highly ordered and form a stable solvation layer. The dynamics of these water molecules are considerably slower than that of their counterparts in the bulk solvent; that is, the hydrogen bond lifetimes are longer for those water molecules bound to the protein. There is, in fact, an exchange of water molecules between the bulk solvent and solvation layer as described by the *dynamic exchange model*. As well as these fairly strong interactions, there will be non-polar amino-acid residues in the protein.

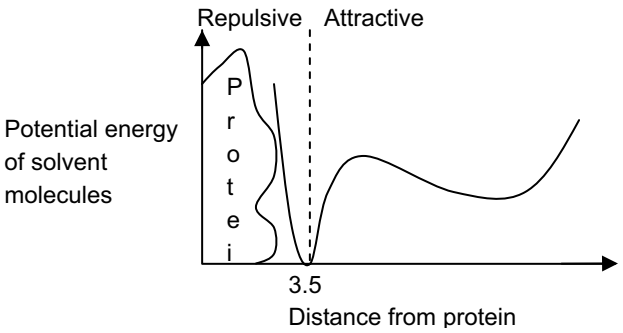


Figure 3.3: Surface region of a protein. Close to the surface of the protein there is high potential energy that is repulsive. This will tend to return these water molecules to the bulk solvent. Others, at about 3.5 Å will become energetic when folding gets under way.

Since these residues are non-polar they will have no bonding with the solvation layer and, as such, they are termed hydrophobic residues. The water molecules that are in the vicinity of these hydrophobic residues will not all be at the same energy – there will be a probability distribution of energies. Close to the hydrophobic residues, because of the probability spread, will be water molecules with high potential energy, which is repulsive. Further out, their energy falls until it reaches zero at about 3.0 to 3.5 Å from the protein surface (**Fig 3.3**); beyond this point there is a mild attraction. The water molecules close to the hydrophobic surface will have sufficient energy to return to the bulk solvent and the hydrophobic residues will tend to clump to minimise exposure of their surface area to the solvent.

This phenomenon of water molecules returning to the bulk solvent increases the entropy of the surrounding solvent, as it becomes more disordered. This compensates for the fall in entropy as the protein folds. This process is called the *hydrophobic effect* and it is what drives the folding process. It should be emphasised that there is no such thing as a hydrophobic bond. The folding times of the protein depend on factors such as the size of the protein and the number of hydrophobic residues in the sequence.

The resulting solvation shell around the protein stabilises the molecule. However, molecular simulations often assume a rigid shell around the protein. This is not strictly the case, since, as we have seen, the solvation shell is highly dynamic environment.

4 Structure-function relationship of proteins

4.1 Relevance of pH and isoelectric point to proteins

In acid-base reactions, it is important to know the concentration of H^+ ions. The measure used to define this concentration is pH:

$$pH = \log\left[\frac{1}{H^+}\right] = -\log\left[H^+\right].$$

Low pH is related to acidic conditions and high pH to alkaline conditions.

The isoelectric point is the pH of a protein at which it has neutral charge. This is a useful facility that can be taken advantage of in protein purification. Specifically, we can use charge in ion-exchange chromatography. Ion-exchange chromatography makes use of positively- or negatively-charged columns to which proteins bind. For example, suppose we want to separate two proteins, A and B that have isoelectric points of A = 7 and B = 9 and we make use of a negatively-charged column. Positively-charged proteins will adsorb (adhere) to the column; negatively-charged proteins will not adsorb. Suppose then that we put the two proteins into a buffer at pH 8. Protein A will be negatively charged and protein B will be positively charged. Hence, B will adhere, but A won't. This example, whilst it puts over the idea, is not a realistic scenario; in practice, you probably wouldn't know what the proteins are – or not all of them – and would have to experiment more with pH.

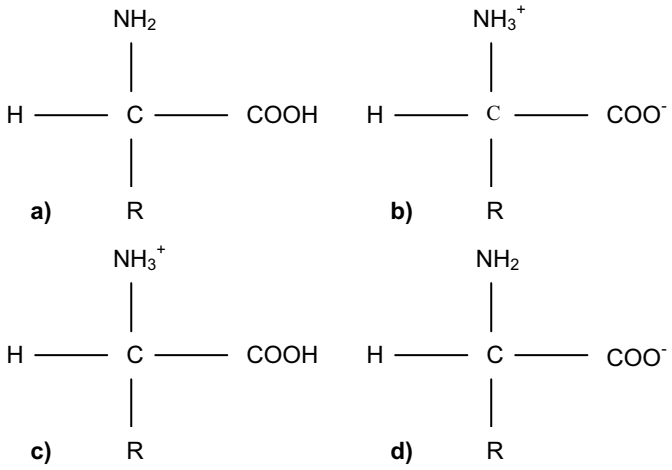


Figure 4.1: Amino-acid modifications with pH. a) Standard representation of an amino acid; b) zwitterions form in a solution at pH 7; c) amino acid at low pH and d) amino acid at high pH.

It will be seen in the next section that amino-acids belong to different charge-based groupings. These groupings can be perturbed slightly by pH and this has important implications for protein function. An amino acid has an acidic carboxyl group (COOH) and a basic amino group (NH_2). It is these two groups that can be manipulated by pH. At neutral pH 7, say in water, an amino acid will surrender the hydrogen ion (H^+) on its carboxyl group, to leave a carboxylate, and the amino group will take up a hydrogen ion (**Fig 4.1**). At low pH there is a predomination of hydrogen ions and so the amino group becomes protonated. Conversely, at high pH, there are few H+ ions around and so the carboxylate group loses its hydrogen. In a protein though, it is not so much the amino and carboxyl groups that are important, as they will be part of the peptide backbone, but the sidechains. Protonation and de-protonation of sidechains are especially important for amino acids that form part of a binding domain, as those two processes allow bonds to be made with a ligand. Very often the location of a ligand in a binding domain is assisted by a metal ion.

	Ala (A)	Arg (R)	Asn (N)	Asp (D)	Cys (C)	Glu (E)	Gln (Q)	Gly (G)	His (H)	Ile (I)
Polar			×		×		×	×		
Non-polar	×									×
Acidic				×		×				
Basic		×							×	
pI	6.0	11.15	5.41	2.77	5.02	3.22	5.65	5.97	7.47	5.94
	Leu (L)	Lys (K)	Met (M)	Phe (F)	Pro (P)	Ser (S)	Thr (T)	Trp (W)	Tyr (Y)	Val (V)
Polar						×	×		×	
Non-polar	×		×	×	×			×		×
Acidic										
Basic		×								
pI	5.98	9.59	5.74	5.48	6.3	5.68	5.64	5.89	5.66	5.96

Table 4.1: Amino acid charge groupings.

4.2 Surface electrostatic potential

Amino acids can be categorized into charge groupings (**Table 4.1**) defined as non-polar, polar, acidic and basic. A non-polar amino acid has no charge associated with it; a polar one ostensibly also has no charge but, under certain circumstances, a dipole can be set up and then it will acquire a pseudo-charge; acidic amino acids are negatively charged and basic ones are positively charged. All these differently charged groups serve to present any ligand with a charged surface. Such surfaces are called surface electrostatic potential maps. The surface generated is a network of charge patches. These become very important in molecular recognition and can form regions of interaction. When imaging surface electrostatic potential, it is common to represent acidic residues as red, basic as blue and grey for polar, and others. The result can be seen in Figure 4.2, which is the electrostatic surface profile of the histone octamer. You should be able to see a medium-sized acidic patch above and to the right of centre. Functionally, we would be looking for a molecule with a corresponding basic patch to interact with this.

To obtain a surface electrostatic potential, the potentials have to be mapped onto a molecular surface. The surface chosen is the solvent-accessible surface. The surface is defined as a percentage of a surface residue that is accessible to solvent. Computer programmes calculate the solvent-accessible surface by rolling a 1.4 Å radius sphere over the surface of the protein. The surface the imaginary sphere rolls over is the Van der Waal's surface; in so doing, a solvent-excluded volume is generated, along with a solvent surface, at a distance of some 2.8 Å.

When a protein undergoes conformational change, it alters its shape. The inference from this is that it will also change its surface electrostatic potential. This may serve to enhance, or deter, binding of another protein. Thus, conformational change is clearly linked to function.

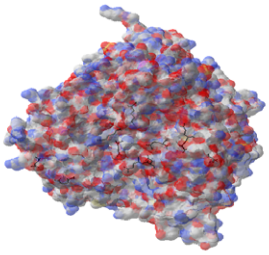


Figure 4.2: Surface electrostatic profile. A surface electrostatic potential map of the histone octamer (protein database code 1TZY). Note the acidic patch slightly above and to the right of centre. Colours are: red – areas of acidic residues; blue – basic residues; grey – uncharged areas.

4.3 Affinity and specificity

Affinity is the strength of binding between two molecules. Specificity is the ability of a protein to bind one substrate in preference to another. If it has high specificity, it will bind only one, or a few substrates; if low, it will bind a range of substrates. Affinity is usually quantified by using the dissociation constant K_D.

Suppose we have a protein A that wants to bind protein B, to form an A-B complex. Initially, at low concentrations of A – $< 0.1K_D$, very little of B is bound to A (**Fig 4.3**). However, at high concentrations – $> 10K_D$ – essentially all of B is bound to A. Imagine that protein A is a molecule in a signalling pathway – the inflammatory response, for example. Ordinarily, the concentrations of A are very low; when the inflammatory response is triggered, the gene for protein A is transcribed and the protein is assembled, with its concentration increasing. We say the protein has increased its expression. Let us assume: that the normal concentration of protein A (with no inflammatory response) is in the picomolar range; that when activated by the inflammatory response, its ***average*** concentration (that is, its concentration taken over the entire nuclear volume) increases to the nanomolar range and that the concentration of protein A required for forming a complex with B is in the micromolar range. The question arises: how does protein A become active in the inflammatory response with apparently too low a concentration?

Before answering the question, it is important to emphasise that this is not an unrealistic scenario – many proteins are at such levels. The answer to the question is that although the average concentration may be low, protein A can reach the required concentration by accruing in local concentrations. This phenomenon is known as colocalization and it is becoming as important a cellular mechanism as transcription and translation – especially in the nucleus. It is now realised that there are many regions of the nucleus where proteins colocalize to increase their effective concentrations. Such areas of the nucleus include: perinuclear compartments, Sam68 bodies, nuclear speckles, cleavage bodies, OPT domains, PML bodies, Cajal bodies and polycomb bodies, amongst others. Proteins also colocalize at transcription centres where dispersed, related genes are pulled together. It is not yet fully understood how chromatin is manipulated in order to get the dispersed, related genes into these transcription centres.

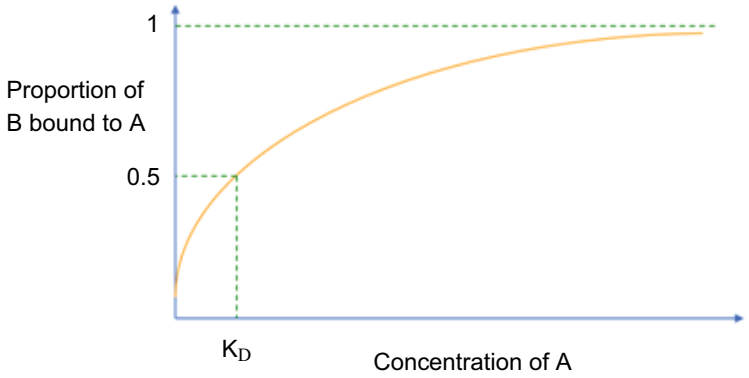


Figure 4.3: Relationship between concentration and binding. The amount of protein B bound to protein A rises sharply with concentration. At a concentration K_D, which is the dissociation constant of the A-B complex, 50% of B is bound to A.

4.4 Allosteric activation

Allostery is the process whereby a disturbance in one part of a protein causes a change in the rest of the protein; in effect, there is a change of shape, which may be small or large. This process should not be confused with conformational change, whereby the protein changes shape of its own accord. The usual cause of allostery is when another molecule, called a *ligand*, binds to the protein. The part of the protein that receives the ligand is called the *binding domain*. The binding domain is made from residues that have a certain degree of flexibility above that of the rest of the protein. This gives the binding domain a higher entropy than the rest of the protein, but upon ligand binding, the entropy falls. This leads to a universal rigidification of the protein with associated entropy loss. This is a thermodynamic interpretation of the traditional model, that says the protein is in the T-state before the ligand binds, and in the R-state afterwards. The binding of the ligand will cause a slight reduction of the rms range over which an atom moves within the protein – perhaps 2%. This may seem small, but a rigidification of 2% upon binding of the ligand equates to an allosteric free coupling energy of 1.5 kcal/mol at room temperature for a protein of 400 residues.

The principle of allostery is very important, as it gives us a spatio-temporal explanation of how proteins can be sequestered and ejected from a functional molecular assembly (**Fig 4.4**). It is important to note that a molecular cluster is not a permanent arrangement. Proteins will join and leave the cluster as and when necessary. In order to function properly, it is necessary that when a protein joins a cluster, it does so for a sufficiently long period to allow some function to be carried out. This is determined by the protein's *residence time* in respect of the cluster. Residence times offer an explanation of how proteins target their cluster. Other inappropriate proteins may well join the cluster; however, their residence will be so short that they will quickly exit the cluster, leaving the way clear for the correct protein to bind.

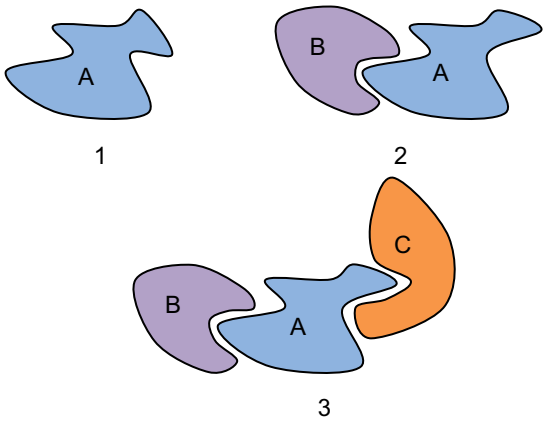


Figure 4.4: Dynamic allostery in molecular formation. Step 1: Protein A has reached its ideal conformation to accept B. Step 2: Proteins A and B now have ideal affinity and bind. Note how this allosterically alters A's distal binding domain, so that protein C can bind. Step 3: C binds and the functional module is complete and can now execute its task. Colours are: blue – protein A; Purple – protein B; orange – protein C.

5 Conformational change via epigenetics

The solution and publication of the human genome was hailed as one of mankind's greatest scientific achievements – which it is – and that, as a consequence, the fields of genetics and molecular biology would make significant advancements. Whilst advancements have certainly been made in those two fields, they are nowhere near the levels that were being claimed when the human genome was released on the scientific world. The reason for this is that things have turned out to be considerably more complicated than was expected at the time. One of the primary reasons for this is that it was discovered that proteins have many more functions than can be explained by the basic 3-dimensional structure that emerges from the ribosome. As was shown in the previous chapter, function is related to structure; therefore, it was deduced, there must be some mechanism that can somehow modify a protein's structure and thus give it an increased functional capacity. The generic term adopted for such changes is *epigenetics*, and the term *epigenome* is widely used as referring to another functional layer on top of that of the genome.

Epigenetics is generally considered to involve a) small RNAs; b) DNA methylation and c) post-translational modifications (PTMs). Small RNAs are molecules that have the ability to control gene expression. The area is of considerable importance and is the focus of a considerable amount of research. However, further discussion of small RNAs is beyond the scope of this book – because a small RNA is a separate molecule rather than a structural modification of a protein, and, therefore, in this chapter only DNA methylation and PTMs are considered.

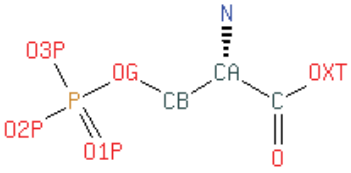
5.1 DNA methylation

Cytosines in human DNA are often methylated when they precede a guanine in a dinucleotide sequence known as a CpG island. Such sequences are up to 2000 bases in length and are usually located close to promoters in a majority of genes in humans. The cytosine is modified by the addition of a methyl group (CH_3) to the 5-carbon of the pyrimidine ring. The methyl group is placed there by a molecule called a DNA methyltransferase (DNMT). The methyl group is surrendered by S-adenosyl methionine.

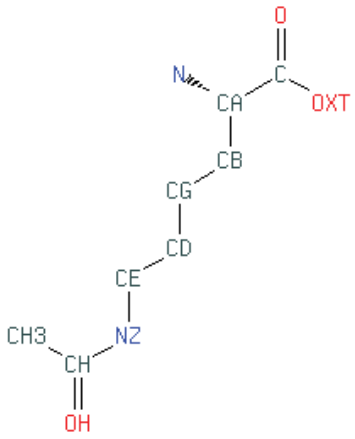
There is quite a lot of non-coding DNA in the human genome and it is desirable to keep that DNA silenced. DNA which is methylated and that is outside CpG islands is never expressed. However, since CpG islands predominantly occur in genes, it is not desirable to have those cytosines permanently methylated. Consequently, if the gene is not expressed, its DNA in the CpG islands will be methylated; if it is to be expressed, the methyl groups are removed. Thus, DNA methylation is a reversible process and is used as a means of controlling gene expression. An exception to this rule is imprinted genes and genes involved with X-chromosome silencing.

5.2 Definition of PTMs

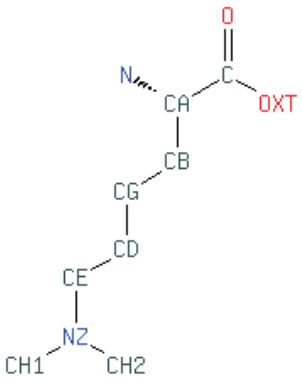
A PTM involves the addition of a small chemical group to a protein that will increase its functional scope. The list of all such modifications is considerable, but the most-commonly researched ones include (**Fig 5.1**): acetylation (lysine), phosphorylation (serine, threonine), methylation (lysine, arginine), sumoylation (lysine) and ubiquitination (lysine). These molecules, when attached to a protein, will allow that protein to have an increased function range – by being able to bind hitherto non-binding substrates. By taking the histones as an example protein, it can be seen how PTMs operate.



(a)



(b)



(c)

Figure 5.1: Modification of various amino-acids by some common PTMs. (a) serine modified by a phosphate group; (b) lysine modified by an acetyl group and (c) lysine modified by a dimethyl group. Colours are: red – oxygen; blue – nitrogen; grey – carbon; orange – phosphorous.

5.3 Bromodomains

The bromodomain is the only domain that will bind an acetylated lysine on histones. The question arises: why will the bromodomain not bind other acetylated lysines? The answer is that those residues within the bromodomain that do not contact the acetylated lysine of the histone protein recognise only that pattern of residues that surrounds the acetylated lysine. This accounts for the remarkable specificity of the bromodomain.

5.4 Chromodomains

The chromodomain will bind histones that have methylated lysines or arginines. The chromodomain consists of an N-terminal, three-stranded, anti-parallel β–sheet that folds against a C-terminal α–helix. The domain has an overall negative charge, and it therefore seems unlikely that it would be involved in binding DNA; it is, however, highly tuned to interactions between proteins. Like the bromodomain, this domain will only bind the histone family of proteins. For example, it has been shown that the chromodomain in the protein will only bind to H3 methylated at lysine 9; if lysine 4 is methylated, there will be no binding.

When an acetylated lysine inserts into a bromodomain it forms bonds with the residues in the binding pocket, including hydrogen bonds. It does this because there is some charge associated with the acetylated lysine. However, there is no charge associated with a methylation and it was therefore wondered how a methylated lysine (or arginine) stays bound to the chromodomain. It is now known that the C-terminal of the chromodomain, which is ordinarily free, wraps around the methylated lysine (or arginine) when it is inserted and it is the residues either side of the methylation that bind to the pocket. In other words, the chromodomain undergoes a conformational change when it binds to a methylated lysine or arginine.

In the histone protein H3 the adjacent residue to lysine 9 – at position 10 – is a serine. When this serine is phosphorylated, the chromodomain in the HP1 protein will dissociate from the methylated lysine at position 9 on H3. The phosphorylation of serine 10 in H3 occurs during mitosis, and so it is this PTM that is responsible for rapid dissociation of HP1 from heterochromatin. The chromodomain has greater affinity for di-methylation than for mono-methylation; similarly, tri-methylation is preferred to di-methylation.

5.5 Domains that bind phosphorylated serines

Thus far, only two domains have been identified as being able to bind to a phosphorylated serine on a histone: the 14-3-3 domain and the tandem BRCT domain.

The 14-3-3 domain is formed from nine α-helices that are in a dimeric form. Helices αA, αC and αD together form the homodimer interface. There are basic residues in the domain pocket that neutralize the charge of the inserted phosphate. Furthermore, solvent (that is water) is excluded from the domain pocket by a number of aromatic residues.

The BRCT domain was first found in the C-terminal of the breast cancer protein BRCA1. These domains are often found in proteins that are involved in DNA damage and control of the cell cycle. When DNA becomes damaged by radiation and a DNA double-strand break occurs, an H2A histone is replaced by H2AX which becomes phosphorylated at serine 139 by ATR or ATM. This particular PTM acts as a signal to a host of DNA repair proteins, and is a good example of how PTMs are involved with gene control.

6 Summary

Proteins are not static 3-dimensional objects. They undergo conformational changes and those changes can be related to a change of function. Even for small conformational changes there can be a significant change in energy.

The binding of one protein to another can allosterically change the latter, such that other proteins are able to bind. This allows the formation of a functional complex. The proteins that make up the complex will not be permanently bound: they come and go as necessary. The determining factor in the formation of a complex is the residence time of a protein.

The affinity of two proteins is determined by their respective concentrations and proteins can sequester in local groups to increase their effective concentrations. This colocalization is now recognised as an important phenomenon.

The functional ability of proteins can be enhanced by PTMs. These modifications of proteins form one of the pillars of epigenetics.