

## Chapter 1

---

### Introduction

Non-covalent interactions are weak interactions between atoms or molecules where no chemical reaction takes place. Because no formation or breaking of chemical bonds is induced, non-covalent interactions are often called non-bonded interactions. Formally, we distinguish three types of non-covalent interactions. The most common are the *van der Waals interactions*. They are short range interactions and occur always when two atoms or molecules come close to each other. We define as short range interactions the interactions which become relevant at distances comparable with the size of the interacting atoms. In this way, practically only neighbouring atoms are involved in these interactions. The *Hydrogen bonds* are interactions which are at the boundary between the chemical bonds and non-covalent interactions. They take place between pairs of atoms only if one of them is a proton donor and the other one is a proton acceptor. *Electrostatic interactions* are the third type of non-covalent interactions. In contrast to the other two types, electrostatic interactions are long range ones. This means that electrostatic interactions are also relevant beyond the limits of the closest neighbours. This makes their description somewhat more complicated. Therefore, a special attention will be paid to these interactions.

Proteins became a subject of intensive investigations as a part of the colloid chemistry, since a number of their physical properties, such as sedimentation, diffusion, viscosity, light scattering, and many others are similar to those of the colloid particles. The colloid particles are molecular aggregates kept together by the delicate balance of attractive and repulsive forces, all resulting from the non-covalent interactions between the molecules comprising the colloid system. Let us set aside for

the moment all we know about proteins and glance at the molecule presented in Fig. 1.1. This is Ribonuclease T1, a small protein which binds and splits ribonucleic acids. The similarity of the molecule to a typical colloid particle is manifested in two aspects, at least. First, the molecule looks like an aggregate of atoms. Second, the surface of the molecule is rich of charges; depending on the physical conditions of the solution, the oxygen (red spheres) and nitrogen (blue spheres) atoms may be negatively or positively charged, respectively, or may have partial charges due to delocalisation of their electron clouds. As it will be shown below, the formation of the compact body seen in the figure, as well as the exposure of the charges on the surface of the molecule, are governed by the same forces responsible for the formation of the colloid particles: namely, the complex action of non-covalent interactions of different type.

Proteins are not colloid particles, but molecules with properties, on the basis of which all known forms of living matter exist. Proteins bind and transport organic and inorganic compounds in this way regulating physiological processes or catalysing chemical reactions. These properties of proteins are referred to as *functional properties*. In the lower panel of Fig. 1.1, the complex of Ribonuclease T1 with the inhibitor guanylyl-2'-5'-guanosine is shown. As seen there, the molecule of the inhibitor is situated in a cleft formed by the protein. This cleft is the *active site*, i.e. the site where the substrate ribonucleic acid binds and the catalytic reaction takes place. It has a shape that matches the size and the conformation of the substrate or the inhibitor. In this way, the active site facilitates binding and at the same time makes it specific: compounds with other chemical composition or in "inappropriate" conformation do not bind. Another important feature which is not illustrated in the figure is that the atoms constituting the cleft create a micro environment facilitating the catalytic reaction when the substrate binds. Thus, the active sites, as well as the rest of the molecule, are not just aggregates of atoms, as the first glance at the molecule could suggest, but organised structures. Even small changes of this organisation may diminish or terminate the function of the protein molecule.

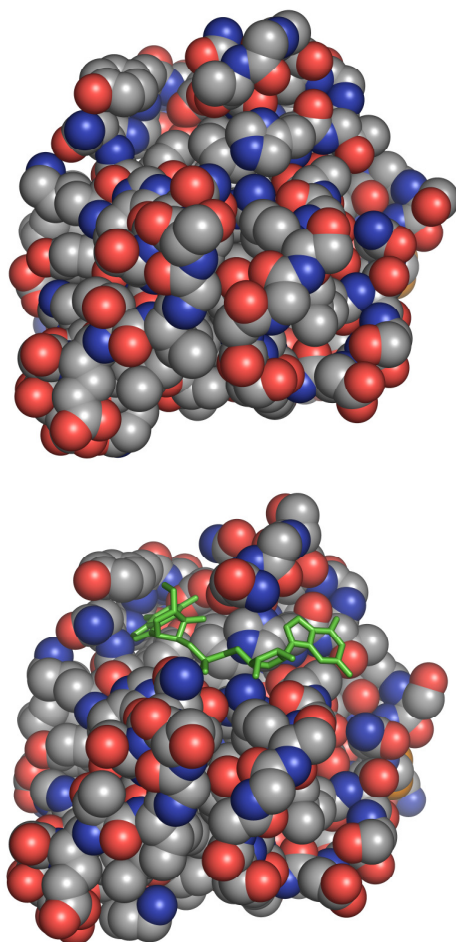


Figure 1.1 Three-dimensional structure of ribonuclease T1 obtained by X-ray crystallography and deposited in Protein Data Bank<sup>1</sup>. The atoms are represented by spheres corresponding to their van der Waals radii and coloured according to their type: grey (carbon), blue (oxygen), red (nitrogen), and yellow (sulphur, partially seen at the right hand side of molecule). These colours will be used in all other figures, unless otherwise stated. The hydrogen atoms are omitted. **Upper panel:** inhibitor free form of ribonuclease T1. **Lower panel:** complex of ribonuclease T1 with the inhibitor guanylyl-2'-5'-guanosine. The inhibitor molecule is represented by sticks and in green in order to make the active site cleft of the protein clearly seen. All colour molecular images are reproduced using The PyMOL Executable Build (2005), DeLano Scientific LLC, South San Francisco, CA, USA, unless otherwise stated.

### 1.1 Some Historical Notes

The first idea for the structuring of proteins was given by Gerardus Johannes Mulder. In his famous paper<sup>2</sup> "Über die Zusammensetzung einiger thierischen Substanzen" ("On the Composition of Some Animal Substances") Mulder investigated the atomic composition of three "albuminous substances", as proteins were then called, noticing that sulphur and phosphorus bind to an organic body with the composition  $C_{400}H_{620}N_{100}O_{120}$ . He named it *protein*, from the Greek πρωτεϊος (primary) and proposed the formulae: **protein + SP** in the case of fibrin and egg albumin, and **protein + 2SP** in the case of serum albumin. In this way, Mulder separated the organic part, the protein, from the inorganic atoms sulphur and phosphorus. At that time, only two amino acids were identified in proteins, glycine and leucine. The sulphur-containing cysteine and methionine were not known, so that separating the sulphur apart from the protein is completely explicable. By the end of the 19th century the list of the amino acids constituting proteins was almost completed. Still, the structure of the protein bodies was unclear. The breakthrough in the understanding of protein structure was made with the hypothesis, independently proposed by Franz Hofmeister<sup>3</sup> and by Emil Fischer<sup>4</sup>, that amino acids in proteins are linked through repeating peptide bonds. It is remarkable that Hofmeister and Fischer reported their ideas in the same day, at the "74th Annual Meeting of the Gesellschaft der deutschen Naturforscher und Ärzte" on September 22, 1902 in Karlsbad (today Karlovy vary, Czech Republic). The hypothesis for the polypeptide nature of proteins became dominant during the next decades. The recognition of all 20 amino acids as protein building blocks was completed with the isolation of threonine in 1938. Still, not very much was known about the spatial organisation of the amino acids in the protein molecules.

The fact that the proteins most commonly used for experimental studies, the globular proteins, exhibit some properties typical for hydrophilic colloid particles — they are water soluble charged particles characterised with a compact structure — stimulated the development of the investigations of electrostatic interactions. In 1924 Linderstrøm-Lang<sup>5</sup> proposed a theory for prediction of hydrogen ion titration curves

of proteins. In this theory, the protein molecule is presented as an impenetrable sphere on the surface of which the charges of the titratable amino acid groups are uniformly distributed. From a present point of view this is a rather rough approximation. Nevertheless, the theory of Linderstrøm-Lang was successfully applied for prediction of titration curves of proteins. The theory of Linderstrøm-Lang became the basis of all following models of electrostatic interactions in proteins known as continuum dielectric models. A valuable contribution to understanding ionisation behaviour of amino acids and proteins was made by Kirkwood with the works on the dissociation constants of organic acids and zwitterions. Based on the same assumption that the protein is an impenetrable sphere, the model was extended by presenting the charges of the titratable groups as point charges<sup>6-8</sup>. Because protein structure was not known, the position of the charge points could not be defined. This made the direct application of the theory to proteins limited to a large extent. In spite of the fact that electrostatic interactions were the first non-covalent interactions that have drawn the attention of the scientists and the following success in developing of comprehensive theoretical approaches, it should be noted that the theoretical description of electrostatic interactions and the evaluation of their role in physical chemical and functional properties of proteins face difficulties even at the present.

The experimental observations convincingly showed that the proteins maintain their compactness at conditions close to the physiological ones, typical for cells or the tissues they are isolated from. At these conditions proteins are in their *native state*. From the concept of the polypeptide nature of protein molecules, it follows that in native proteins the polypeptide chain is folded and in this way forms a compact body. Changing the conditions, for instance by reducing pH or increasing temperature, the proteins lose their compactness and solubility, as well as their activity. Proteins adopt *denatured state*, or simply, they denature. The polypeptide chain in denatured state is not folded any more, it adopts the features of a random coil, so that one can also speak about *unfolded proteins* or *unfolded state*, a term which is more adequate if the state of the polypeptide chain is of interest. Unfolded state can be achieved by adding denaturing agents, such as guanidinium chloride or urea

(chemical denaturation), changing pH of the solution (pH-induced denaturation), or changing the temperature (thermal denaturation). The question arises as to what the forces keeping the polypeptide chain folded are, thereby maintaining the protein molecule in the native state. The simultaneous presence of positive and negative charges in the protein could partially give an answer: the attraction between opposite charges keeps the native proteins compact. This concept was strongly supported by the pH-induced denaturation experiments. Reducing pH, for instance, the negatively charged amino acids become neutral, which leads to the increase of the contribution of the repulsive interactions between the positive charges. As a result of this repulsion the protein molecule denatures. Although the picture of pH-induced denaturation described above is qualitatively correct, it was clear that electrostatic interactions, by themselves, could not give a comprehensive explanation of the stability of the compact structure of native proteins.

It is fashionable nowadays to talk about, and to stimulate, interdisciplinary science. In fact, science becomes spontaneously interdisciplinary if it is needed, and this is not a privilege of the present day. The "young" protein science, inhabiting the room of colloid and organic chemistry, is an example of interdisciplinary research started at least one century ago. The fast advance of the quantum physics brought forth new understanding of the structure of matter and the interactions within and between the molecules. It was found that apart from the chemical bond (sharing of electrons) there are other attractive interactions based on sharing of a hydrogen nucleus: the hydrogen bond. This idea was first proposed by Latimer and Rodebush, as the electrostatic attraction between the unbounded electron pairs and the polar hydrogen atoms was called "weak bond"<sup>9</sup>. Linus Pauling has further developed this idea, introducing also the term hydrogen bond, and implemented it for the explanation of the forces responsible for the compactness of the native proteins<sup>10,11</sup>. In the fundamental work of Pauling and Myrsky<sup>11</sup> the compact structure of proteins was explained in terms of a network of hydrogen bonds between the peptide nitrogen and oxygen atoms, which keep the polypeptide chain in a "uniquely defined configuration". Further stabilisation of the native protein structure comes

from the hydrogen bonds between the side chains of the amino acids\*. Moreover, Pauling and Mirsky explained protein denaturation with breaking of the hydrogen bonds. Thus, together with electrostatic interactions, a new type of non-covalent interactions has been involved in the description of properties of proteins.

A decisive step towards revealing protein structure was made shortly after the Second World War. More precisely, the decisive step had been hindered by the war. Two important discoveries should be mentioned. Frederic Sanger developed an experimental methodology for determination of the amino acid sequence of protein molecules. He, and his collaborator Hans Tuppy, were the first to determine the sequence of a protein<sup>12-14</sup>. This was the molecule of insulin (Fig. 1.2).

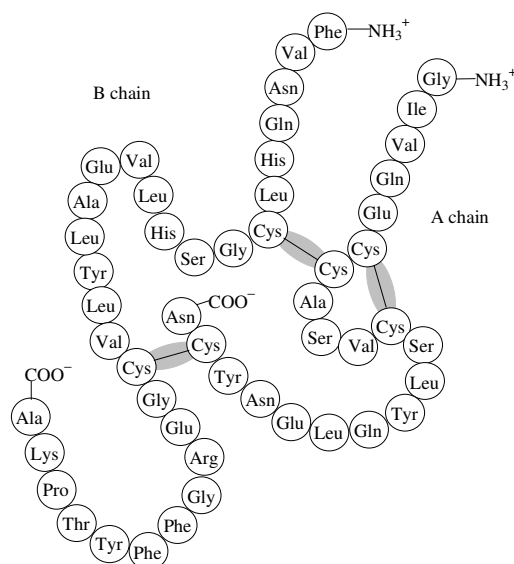


Figure 1.2 The amino acid sequence of insulin determined by Sanger and Tuppy. The protein consists of two polypeptide chains linked by two disulphide bonds. The amino acid names are given in three-letters code (see Table 1.1).

The finding of Sanger and Tuppy can be considered not only as a final proof that the Hofmeister-Fischer hypothesis is correct. It confirmed

\* Terms, such as side chain, main chain, etc. are described in the next section.

that the fundamental element of the protein structure is the polypeptide chain. Thus, the Mulder's "primary" part of the "albuminous bodies" is a sequence of amino acids linked in a polypeptide chain and naturally has the name *primary structure*. The primary structure, i.e. arrangement of the amino acids along the polypeptide chain, is unique for proteins from a given type and given species. This allows the formulation of a new hypothesis, namely that it is the sequence that determines the structural organisation of the protein molecule, and hence its functional properties. This is the so-called *Anfinsen's dogma*<sup>15</sup>, which will be discussed in Section 1.3. Nowadays this hypothesis is beyond any doubt, being continually confirmed, say, by mutagenesis experiments. Mutations of the sequence, i.e. changes caused by adding, removing or substituting amino acids in the polypeptide chain, result in changing the structure and the properties of the protein molecule. Of course, not all changes make the same impact on the structure and the functions of the proteins. Changes in parts of the sequence involved in the formation of the active site are, as a rule, crucial for the functions of the molecule, while other changes may have negligible influence.

One can compare the protein sequence with a text written by means of 20 letters containing the information needed to build a molecule with defined structure and functions. The "letters" in the sequence are the amino acids, which differ from each other by the chemical compositions of their side chain, and hence by their physical chemical properties. Obviously, the information coded in the sequence transforms to a real structure with corresponding properties by means of a mechanism based on the non-covalent interactions between the amino acid side chains, as well as between the atoms of the protein molecule and the surrounding solvent.

The second important discovery is a result of the purposeful work of Pauling on the factor stabilising native proteins. In the year 1951, he published the first model of helical conformations of a polypeptide chain<sup>16</sup>. Based on X-ray data of crystalline amino acid and short peptide, he modelled the well known  $\alpha$ -helix with amazing precision. Later, he proposed another structural organisation, the  $\beta$ -sheet, at which the polypeptide chains are in extended conformation. The polypeptide chains in a  $\beta$ -sheet can be mutually oriented in two ways: parallel, when the

directions form N- to C-termini of adjacent chains are the same, and antiparallel, when N- to C-termini directions are opposite. Examples for a  $\alpha$ -helix and a  $\beta$ -sheet combining parallel and antiparallel orientations of the polypeptide chains are given in Fig. 1.3.

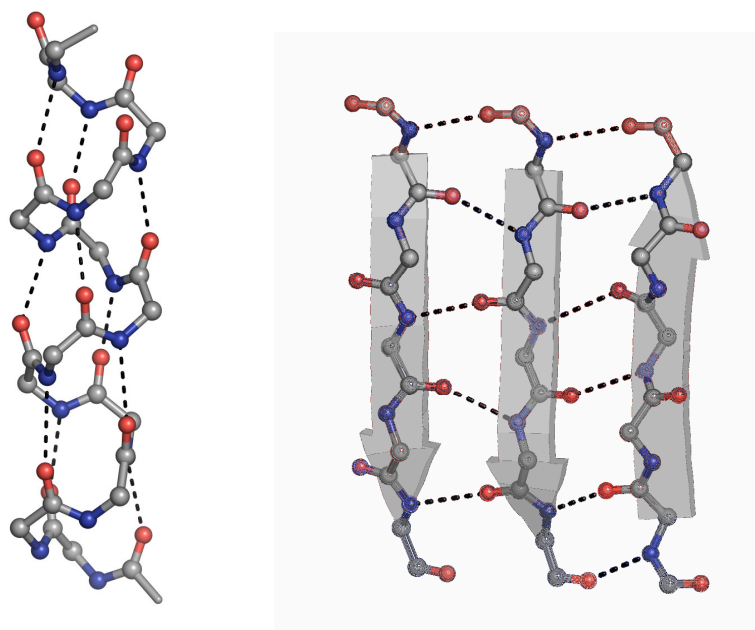


Figure 1.3 The most common secondary structural elements. The hydrogen bonds ( $>C=O \cdots H-N-$ ) are given with dashed lines. The hydrogen atoms are not shown. **Left:**  $\alpha$ -helical segment from *Helicobacter pylori* cysteine rich protein *b*. **Right:** fragment of a  $\beta$ -sheet from antithrombin. The first and the second polypeptide chains from left hand side of the  $\beta$ -sheet are in parallel orientations, whereas the second and the third are in antiparallel mutual orientation. The arrows indicate the parallel and the antiparallel orientations.

The  $\alpha$ -helices and the  $\beta$ -sheets, predicted by Pauling, proved to be the most common conformational pattern found in proteins. There are also other structural organisations of the polypeptide chain found in proteins later. Such are the  $3_{10}$ -helices and two other types of bends, as well as the rarely observed  $\pi$ -helices. All these structural elements are united by the name *secondary structure*.

Electrostatic interactions and hydrogen bonds were considered as the main factors responsible for protein properties, including structural organisation, stability of the native state, as well as functional properties. It has been noticed, however, that the structure of native proteins is related to the surrounding solvent. The pioneering investigations on X-ray diffraction of protein crystals have shown, for instance, that during evaporation of water the diffraction loses its sharpness. This phenomenon could be interpreted in terms of breaking of hydrogen bonds between water molecules and the polar groups of the amino acid side chains of the protein molecule. This example is important with the fact that not only electrostatic interactions and the hydrogen bonds within the molecule, but also the interactions between the solvent molecules and the protein play a role in the stabilisation of the structure of the protein molecules.

It was also known that some organic compounds, such as hydrocarbons, are hydrophobic, i.e. they do not dissolve in water, but rather prefer to form aggregates or to stay at the water/air interface. This effect suggests that there is a kind of attractive interactions between the hydrophobic compounds when they are surrounded by water. Proteins contain amino acids that have hydrophobic side chains, so that this type of interactions should be present also in proteins. Kauzmann was not the first who paid attention to this fact, but he was the first who related it to the stability of native proteins. In an article published in 1959<sup>17</sup> he emphasised the fact that about half of the amino acids found in proteins have non-polar, hydrophobic, side chains. He introduced the term *hydrophobic bond* to describe the attractive interactions the non-polar amino acid side chains are involved in. The hydrophobic interactions cause a seeming attraction between the non-polar amino acid side chains, accompanied by a reduction of their contact area with the water molecules. As an end effect, this leads to the formation of a *hydrophobic core* of the protein molecule. The energetic evaluations of the hydrophobic core formation have shown that hydrophobic interactions are the dominant factor for the stabilisation of the native protein structure.

We speak about hydrophobic interactions, semantically emphasising on the appearance of the phenomenon: the attraction between non-polar

compounds in water medium. As we shall see in Chapter 4, hydrophobic interactions do not exist as a separate type of interactions. Rather, they are an effect of the behaviour of water molecules surrounding a non-polar compound.

The determination of the three-dimensional structure of proteins entirely confirmed the concept of the role of non-covalent interactions in spatial arrangement of the protein molecule. The predicted  $\alpha$ -helical organisation of the polypeptide chain made by Pauling was found in the first high-resolution three-dimensional protein structure, that of myoglobin<sup>18</sup>. It was also found that the titratable groups, i.e. the titratable amino acid side chains, are predominantly of the protein surface, as presumed by Linderstrøm-Lang. The hydrophobic side chains are buried in the protein interior, which follows the concept of Kauzmann.

The three-dimensional structure of proteins is often called *tertiary structure*. Shortly after the first three-dimensional protein structure had been solved, the structure of the haemoglobin was solved. This is a protein molecule which is an assembly of more than one polypeptide chains. If a protein constituted by an assembly of polypeptide chains, it is characterised by *quaternary structure*. The individual polypeptide chains in this case are called *subunits*. The individual subunits can be identical, similar, or completely different, each characterised by a different fold.

The hierarchy of protein structures (primary, secondary, tertiary, quaternary) has been introduced by Linderstrøm-Lang before their unambiguous experimental determination by X-ray crystallography<sup>19</sup>. One can say that ideas about the organisation of the protein molecules were developing with forestalling rates in comparison with their experimental confirmation. It seems that the situation now is the opposite. At present, more than 31,000 protein structures are deposited in the Protein Data Bank\* and this number continuously grows. Still, there is a wide spectrum of questions related to the stability and the functionality of proteins, whose answer is unknown. Of course, the character of these questions is different. We would like to know the molecular mechanisms of enzymatic catalysis; or we would like to predict the change of concrete functional properties of a protein molecule

---

\* <http://www.rcsb.org/pdb/Welcome.do>

caused by mutations. Moreover, we would like to manipulate protein properties; to reduce or increase stability at given conditions, to design proteins, which are active at external conditions (temperature, pH, etc.) distant from those typical for their natural environment. Without knowledge of protein structure and a deeper understanding of non-covalent interactions, these ambitions are not realistic. Therefore, structural biologists, chemists, physicists and theoreticians work together making the modern protein science a very interesting research area.

## 1.2 Overview of Protein Structural Elements and Basic Definitions

In the previous sections, a number of terms were introduced without being precisely defined. This mainly concerns the structural elements of proteins. We will confine ourselves only to the terms relevant for the matter of interest, the non-covalent interactions. A detailed survey on protein structure with typical examples and appropriate illustrations can be found in the books of Lesk<sup>20,21</sup>, as well as in any biochemistry textbook.

### 1.2.1 The amino acids

In Fig. 1.4 the naming of the atoms in the amino acids is illustrated using lysine as an example. It is convenient to start with the atom named  $C\alpha$  and to consider it as a central one. In the free amino acids, i.e. amino acids which are not linked in a polypeptide chain,  $C\alpha$  binds the  $\alpha$ -amino and  $\alpha$ -carboxyl groups ( $-NH_2$  and  $-COOH$  in the left panel of Fig. 1.4, respectively). If the amino acid is part of a polypeptide chain these two groups are transformed to amide ( $-NH-CO-$ ), and form the peptide bonds linking the adjacent amino acids (see also Fig. 1.6). Therefore, in a polypeptide chain the  $\alpha$ -amino and the  $\alpha$ -carboxyl groups can exist only at its ends (see Fig. 1.2). The atoms constituting the polypeptide chain of the protein are denoted as N,  $C\alpha$ , C, and O (the right hand side panel of Fig. 1.4). These atoms form the *main chain* of the protein molecule, or in other words, the *backbone*. The third non-hydrogen atom bound to  $C\alpha$  is  $C\beta$ . This atom is the first in the *side chain* of the amino acid. The atoms

from the side chain are designated by their chemical symbols and successive Greek letters. In the example given in Fig. 1.4, the most distant atom along the side chain is  $N\zeta$ , from the  $\varepsilon$ -amino group (this group is bound to the carbon at position  $\varepsilon$ ,  $C\varepsilon$ ).

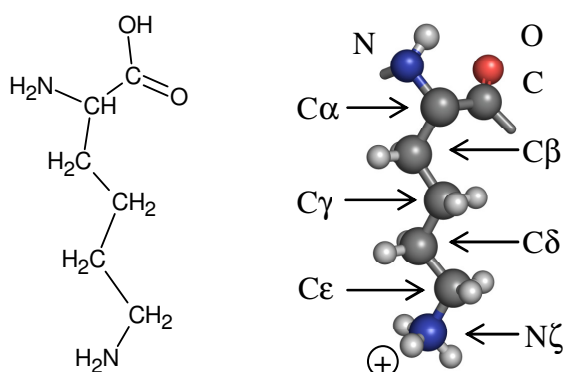


Figure 1.4 Naming of the amino acid residues illustrated with lysine. Note the differences between the left and the right panels. **Left:** The structural formula of the amino acid lysine. The  $\alpha$ - and  $\varepsilon$ -amino groups are in their deprotonated forms, whereas the  $\alpha$ -carboxyl group is protonated; all these groups are neutral in this state. **Right:** Hard spheres model of the amino acid residue lysine. The hydrogen atoms are presented as small spheres in grey, the colour scheme for the other atoms is as in Fig. 1.1. The  $\varepsilon$ -amino group is protonated, i.e. positively charged ( $-\text{NH}_3^+$ ). Note also that only the atoms involved in the peptide bonds are given.

As we have already discussed, the amino acids differ in their side chains. In Table 1.1 the amino acids composing the protein molecules, the *natural amino acids*, are grouped according to the most commonly used criteria. In order to easily distinguish the difference in the chemical composition of the side chains, they are drawn vertically, whereas the atoms participating in the peptide bonds, the  $\alpha$ -amino and the  $\alpha$ -carboxyl groups are horizontally drawn. The aliphatic side chains are entirely hydrocarbon chains, a feature determining their low solubility in water. The shortest aliphatic side chain, that of alanine, contains only one methyl group at position  $\beta$  (see also Fig. 1.6). The other aliphatic side

chains are branched. Valine and isoleucine have two methyl groups on  $\gamma$ -position, whereas leucine has two methyl groups on position  $\delta$ .

Non-polar groups are also poorly soluble in water. Due to their low solubility in water non-polar and aliphatic side chains are also called hydrophobic side chains or *hydrophobic groups*. Glycine does not have a side chain, the position  $\beta$  being occupied by a hydrogen atom. It is classified as a non-polar group because the hydrogen atom at  $\beta$  position is practically not polarised (Chapter 3). The side chain of proline is peculiar with  $C\gamma$  being linked to N from the backbone. The  $-SH$  group of a cysteine side chain is often linked to the  $-SH$  group of another cysteine in the protein molecule, thus forming a *disulphide bridge* (schematically illustrated in Fig. 1.2). The polarity of the S–S cross-link is low. However, if the cysteine side chain is not involved in a disulphide bridge, it is not any more non-polar. Even more, the  $-SH$  group can be deprotonated at alkaline pH making the cysteine side chain negatively charged.

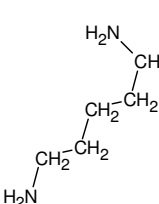
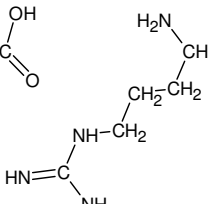
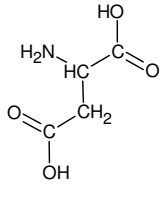
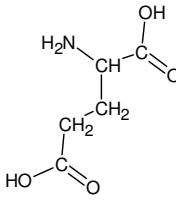
Among the aromatic side chains, those of the histidines and the tyrosines can also be charged depending on the protonation state of the imidazole rings and the phenol hydroxyl groups, respectively. The polar side chains are well soluble in water, they are hydrophilic. They contain amide (asparagine and glutamine) and hydroxyl groups (serine and threonine) which can form hydrogen bonds with other neighbouring polar groups from the protein moiety or with compounds, including water molecules, from the solvent.

The charged amino acid side chains have groups which at physiological pH values are usually charged. In the aspartic and the glutamic acids these are  $\beta$ - and  $\gamma$ -carboxyl groups, in the arginine this is the guanidine group, and for the lysine the  $\epsilon$ -amino group. These side chains, as well as the histidines, the tyrosines and reduced (not involved in disulphide bridges) cysteines create a charge constellation which changes with the physical conditions, for instance pH, and thus changes the properties of the protein molecule.

Table 1.1 Structural formulae of the amino acids constituting protein molecules. Three- and one-letter codes of the amino acid names are given in parentheses.

Aliphatic			
Alanine (Ala, A)	Valine (Val, V)	Leucine (Leu, L)	Isoleucine (Ile, I)
Non-polar			
Glycine (Gly, G)	Proline (Pro, P)	Cysteine (Cys, C)	Methionine (Met, M)
Aromatic			
Histidine (His, H)	Phenylalanine (Phe, F)	Tyrosine (Tyr, Y)	Tryptophan (Trp, W)
Polar			
Asparagine (Asp, N)	Glutamine (Gln, Q)	Serine (Ser, S)	Threonine (Thr, T)

Table 1.1 (Continued)

Chargeable			
Lysine (Lys, K)	Arginine (Arg, R)	Aspartic acid (Asp, D)	Glutamic acid (Glu, E)
			

As seen from this brief survey, the individual amino acid side chains interact with their neighbourhood predominantly with one or another type of non-covalent interactions, depending on their chemical compositions. This results in a complex interplay between the different non-covalent interactions, which on its side is at the bottom of the mechanisms for transformation of the information coded in the protein sequence into a structure with given properties.

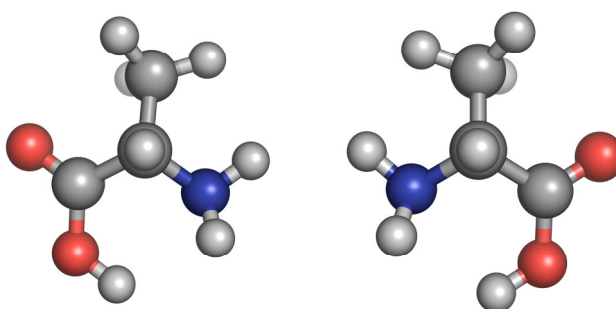


Figure 1.5 Orientation of the atoms in L-amino acid (left hand side) and R-amino acid (right hand side). The amino acids are shown viewed from the hydrogen atom towards the C $\alpha$  atom which is at the centre of the amino acid.

Except for glycine the amino acids are chiral, i.e. they cannot be superimposed on their mirror images. In this way, one distinguishes two

types of amino acid: L-amino acid and their mirror image, the R-amino acid. The usual way for representing the spatial organisation of the atoms connected to  $C\alpha$  is given in Fig. 1.5. Both L- and R-amino acids exist in Nature. However, only the L type is presented in proteins. This fact is still one of the mysteries of Nature still awaiting a plausible explanation.

### 1.2.2 The polypeptide chain

Understanding of the principles for formation of the protein structure cannot be achieved without taking into account the specific properties of the polypeptide chain. In Fig. 1.6 a segment of a polypeptide chain containing the amino acid side chain alanine is shown. The peptide bond connecting two amino acids contains the atoms  $C_{i-1}$  and  $O_{i-1}$  from the previous amino acid,  $i-1$ , and the atoms  $N_i$  and  $H_i$  from the amino acid  $i$  (the alanine). If the group  $i$  is the first one in the polypeptide chain ( $i = 1$ ), the  $\alpha$ -amino group ( $-NH_2$ ) remains free. It is usually called N-terminal amino group, or simply N-terminus. As it has been already mentioned, at physiological pH the  $\alpha$ -amino groups are protonated, i.e. have the form  $-NH_3^+$ . In the same way the  $C_i$ ,  $O_i$ ,  $N_{i+1}$ , and  $H_{i+1}$  form the next peptide bond. If the group  $i$  is the last in the polypeptide chain, the  $\alpha$ -carboxyl group remains free (C-terminus) and is charged at physiological pH values. Thus, the ends of a polypeptide chain, the N- and the C-termini, are usually charged.

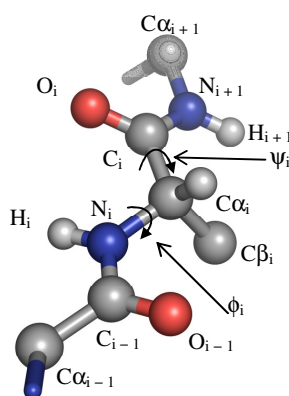


Figure 1.6 The amino acid alanine linked in a polypeptide chain.

The peptide bond is planar. By convention, the angle between the C=O and NH groups has the value  $\omega = 180^\circ$  when the oxygen,  $O_i$ , and the hydrogen,  $H_{i+1}$ , atoms are most distant (*trans* conformation). This is the conformation illustrated in Fig. 1.6. The other conformation that allows planarity of the peptide bond is the *cis* conformation at which  $\omega = 0^\circ$ . This conformation is energetically less favourable than the *trans* conformation and is rarely observed in proteins.

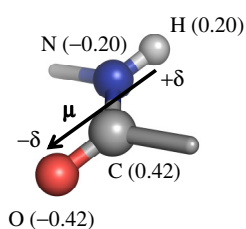


Figure 1.7 Dipole moment of the peptide bond. Values in parentheses are the partial charges of the individual atoms<sup>22</sup>.

The peptide bond is also a highly polar one. The distribution of the partial charges within the peptide bond is given in Fig. 1.7. It should be noted that the values of the partial charges can vary depending on the model and the method used for their calculation. In all cases, however, the charge distribution creates a substantial dipole moment.

An important property of the polypeptide chain is that it is flexible. This flexibility is ensured by the rotation around N-C $\alpha$  and C $\alpha$ -C bonds connecting the amino acid residue with the adjacent peptides. The rotation around N-C $\alpha$  is given by the value of the torsion angle  $\phi$ , whereas the rotation around C $\alpha$ -C is described by the torsion angle  $\psi$  (see Fig. 1.6). The angles  $\phi$  and  $\psi$  are called torsion or dihedral angles. By convention,  $\phi = 0$  if the atoms  $C_{i-1}$ ,  $N_i$ ,  $C\alpha_i$ , and  $C_i$  are co-planar and in *cis* conformation. Similarly,  $\psi = 0$  if the atoms  $N_i$ ,  $C\alpha_i$ ,  $C_i$  and  $N_{i+1}$ , are co-planar and in *cis* conformation. Not all combinations of  $\phi$  and  $\psi$  are possible because at certain combinations the atoms from the side chains collide. In 1968, Ramachandran<sup>23</sup> showed that only  $\phi$  and  $\psi$  are needed to fully describe all conformation of a polypeptide chain and invented a useful representation of the regions where the two angles correspond to sterically allowed conformations (Fig. 1.8).

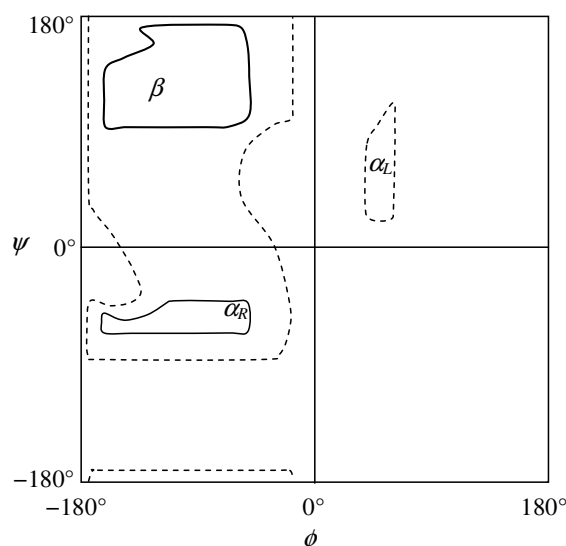


Figure 1.8 Ramachandran plot of the backbone dihedral angles  $\phi$  and  $\psi$ . Sterically allowed combinations are marked with dashed line, energetically most favourable with continuous line. The regions denoted with  $\alpha_R$ ,  $\alpha_L$ , and  $\beta$  correspond to right  $\alpha$ -helix, left  $\alpha$ -helix and  $\beta$ -sheet conformations. The left  $\alpha$ -helix conformation is allowed for glycine only.

As seen from the figure, the regions of values that  $\phi$  and  $\psi$  can have are restricted. However, the number of conformations that the polypeptide chain can adopt is enormously large.

### 1.3 Non-covalent Interactions and Structure-Function Relationships in Proteins

The problem of structure-function relationships in proteins is one of the fundamental problems of protein science. That is, what are the driving forces responsible for the protein molecules to adopt structures with certain biologically relevant properties? Also, can we predict the functional properties of a protein molecule if we know its structure? Further questions can be posed as well. For instance, how are functional properties changed when protein structure is changed? To a certain extent these questions have already been addressed in Section 1.1. We

have pointed out that the mechanism of decoding the information stored in the protein sequence (Anfinsen's dogma) is based on the interplay between the non-covalent interactions. Therefore, it becomes obvious that the problem of structure-function relationships can be approached after a quantitative analysis of non-covalent interactions.

### ***1.3.1 Some comments on Anfinsen's dogma***

Anfinsen's dogma states that the three-dimensional structure of a protein is determined solely by its amino-acid sequence. This conclusion is deduced from the fact that unfolded proteins lose their functional properties adopting the unfolded state: a state at which no secondary or higher levels of structure permanently exist. After removing the denaturing agents proteins spontaneously fold back into the native conformation restoring their biological functions. There are two aspects of this observation. The first one is the simple scheme:

- losing the structure causes losing of function
- restoring the structure causes restoring the function,

which proves the connection between protein structure and protein function. We have to note that the above scheme should not be considered as a strict formula. There are proteins which are unstructured, yet with biological functions. Such are, for instance, the  $\alpha_s$ -casein and the histone H3. Other proteins, such as a number of transcription factors, manifest their biological functions upon the transition from unfolded to folded state. There are also proteins, which are only partially structured. The human  $\gamma$ -interferon can be taken as an example. The polypeptide chain of this protein is well structured, whereas the part containing the last 21 amino acids in the sequence is disordered. The biological functions of the unstructured part of  $\gamma$ -interferon are still unclear<sup>24</sup>. These few examples are not violations of Anfinsen's dogma. They should rather be considered as an extension, making the problem of structure-function relationships in protein more complicated and hence, the task more interesting.

The other aspect concerns the spontaneous refolding after removal of denaturing agent. The interpretation of this observation is that the native conformation of the protein corresponds to the state of global-minimum

of the free energy. This means that the description of the driving force of protein folding obeys the principle of thermodynamics. Although this conclusion is not surprising and, to a certain extent, sounds trivial, it is useful to have it as a standpoint. Protein functions are not a result of forces that are inherent to living matter only; they are result of interactions that follow the known physical laws. Combining structural information, say X-ray data, with the phenomenological data obtained by calorimetric or other indirect experimental methods, one can build a correct understanding of a wide variety of problems united by the term structure-function relationships. One should only add one relevant detail. We need to know more about non-covalent interactions, especially about their specific appearance in proteins.

### ***1.3.2 Experimental measurements of non-covalent interactions in proteins***

Non-covalent interactions in proteins cannot be directly measured. Therefore, their quantification is made indirectly, by evaluation of their effect on measurable quantities. Thus, for instance, the magnitude of electrostatic interactions between two titratable groups is deduced from the shift of the ionisation equilibrium constants of these groups with respect to some reference values. The contribution of the hydrogen bonds to structural stability is evaluated on the basis of model compounds or by mutagenesis experiments. In all these cases the above assessments are approximate. The shift of the equilibrium constants depends on electrostatic interactions, but other effects, such as desolvation, have significant influence as well. The use of model compounds for estimation of the contribution of hydrogen bond strength to the stabilisation of secondary structural elements in proteins is *a priori* an approximation, because measurements are not made on proteins. Measurements based on mutagenesis experiments also introduce uncertainty and need careful analysis of experimental data because the chemical composition of the protein molecule is changed.

One can argue that direct experimental measurements can still be done. Indeed, X-ray crystallography, as well as Nuclear Magnetic Resonance (NMR) experiments and electron microscopy, provide direct

structural information. On the basis of X-ray data the spatial coordinates of the protein atoms can be determined, which allows the estimation of a variety of characteristics, such as the geometry of hydrogen bonds, the distribution of the charged groups, the packing, and most importantly: the identification of the active site. One can hardly imagine any progress in the understanding of structure-function relationships without this information. Structural information is substantial but it is not sufficient. In order to understand and hence predict the protein properties — both physical-chemical and biological — one needs to know the magnitude of non-covalent interactions in detail. In other words, one needs to know the interaction energy between the protein atoms. Fortunately, methods for direct measurements of the "energetics" of proteins exist. These are the Differential Scanning Calorimetry (DSC) and the Injection Titration Calorimetry (ITC). By means of DSC one can measure the change of the heat capacity and the enthalpy upon protein denaturation. The change of enthalpy due to ligand binding or protein assembly, for instance, can be directly measured using the method of ITC. These quantities are however thermodynamic ones and characterise the investigated system (the protein molecule and the surrounding solvent) as a whole. They do not provide information about the nature of the interactions that take place in the system. Therefore, in order to understand the connection between the observed phenomena and their molecular nature, we need appropriate theoretical tools. One can say that the bridge between the observation and the understanding is the theory. In the following considerations, we will try to consider the basic tools that can help us understand the extraordinary features of the molecules of life — the proteins.

## References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE, (2000) The protein data bank. *Nucleic Acids Res.*, **28**: 235–242.
2. Mulder GJ, (1839) Ueber die zusammensetzung einiger thierischen substanzen. *J. Prakt. Chem.*, **16**: 129–152.
3. Hofmeister F, (1902) Über den bau des eiweissmoleküls. *Ergeb. Physiol.*, **1**: 759–802.

4. Fischer E, (1902) Ueber die hydrolyse der proteinstoffe. *Chem. Ztg.*, **26**: 939.
5. Linderstrøm-Lang K, (1924) On the ionisation of proteins. *C. R. Trav. Lab. Carlsberg*, **15**: 1–29.
6. Kirkwood JG, (1934) Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *J. Chem. Phys.*, **2**: 351–361.
7. Westheimer FH and Kirkwood JG, (1938) The electrostatic influence of substituents on the dissociation constants of organic acids. II. *J. Chem. Phys.*, **6**: 513–517.
8. Kirkwood JG and Westheimer FH, (1938) The electrostatic influence of substituents on the dissociation constants of organic acids. I. *J. Chem. Phys.*, **6**: 506–512.
9. Latimer WM and Rodebush WH, (1920) Polarity and ionization from the standpoint of the lewis theory of valence. *J. Am. Chem. Soc.*, **42**: 1419–1433.
10. Pauling L, (1928) The shared-electron chemical bond. *Proc. Natl. Acad. Sci. U. S. A.*, **14**: 359–362.
11. Pauling L and Mirsky AE, (1936) On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **22**: 439–447.
12. Sanger F and Yhompson EOP, (1953) The amino-acid identification in the glycol chain of insulin. 1. The identification of the lower peptides from partial hydrolysates. *Biochem. J.*, **53**: 353–374.
13. Sanger F and Tuppy H, (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, **49**: 463–481.
14. Sanger F and Tuppy H, (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymatic hydrolysates. *Biochem. J.*, **49**: 481–490.
15. Anfinsen CB, (1973) Principles that govern the folding of protein chains. *Science*, **181**: 223–230.
16. Pauling L, Corey RB and Branson HR, (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **37**: 205–211.
17. Kauzmann W, (1959) Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, **14**: 1–63.
18. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC and Shore VC, (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, **185**: 422–427.
19. Linderstrøm-Lang KU, (1952) *Proteins and Enzymes The Lane Medical Lectures*. Stanford, CA: Stanford University Press.
20. Lesk A, (2004) *Introduction to Protein Science. Architecture, Function, and Genomics*. New York: Oxford University Press.
21. Lesk A, (2001) *Introduction to Protein Architecture*. New York: Oxford University Press.

22. Hol WGJ, van Duijnen PT and Berendsen HJC, (1978) The  $\alpha$ -helix dipole and the properties of proteins. *Nature*, **273**: 443–446.
23. Ramachandran GN and Sasiskharan V, (1968) *Adv. Protein Chem.*, **23**: 283–437.
24. Tsanev R and Ivanov I, (2001) *Immune Interferon: Properties and Clinical Applications*. Boca Raton, USA: CRC Press.