Introduction to proteins and protein structure

The questions and answers below constitute an introduction to the fundamental principles of protein structure. They are all available at **[link]**.

What are proteins?

Proteins are biological molecules performing a wide variety of functions. For instance, some proteins catalyse a reaction, i.e. they make it go faster than it normally would: those proteins are called *enzymes*. Other proteins transport molecules throughout the body, others yet provide structural support for cells so they have the right shape, etc.

Proteins are involved in nearly every biological process and their function is very often tightly linked to their three-dimensional structure. Therefore, it is crucial to determine the structure of a protein in order to understand fully how it works inside a cell.

The many functions of proteins are reflected by the wide variety of 3D structures they adopt. However, all proteins are made of the same constitutents: amino acids.

What are amino acids?

Proteins are polymers: similar molecules (called *monomers*) are repeated many times to form a chain (the *polymer*). The monomers making up proteins are amino acids, whose general structure is shown in Figure 1.

Amino acids (except glycine) contain a central chiral carbon, i.e. a carbon atom covalently linked to four different groups of atoms, often called *carbon α* (C_α). This central chiral atom is linked to an amino group and a carboxylic acid group, thus the term *amino acid*. It is also attached to a hydrogen atom and a side chain (sometime called R group). An amino acid's side chain is what sets it apart from the other amino acids and is often responsible for the special chemical and biological properties of the amino acid.

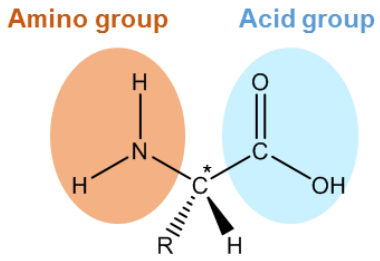


Figure 1 | An amino acid with the amino group and the carboxylic acid group in orange and blue, respectively. The side chain is represented by the letter R and the C_α is shown by the asterisk (*).

In water at pH 7 (which is close to physiological conditions inside cells), the amino group is protonated and positively charged, while the carboxylic acid group is deprotonated and negatively charged, as show in Figure 2.

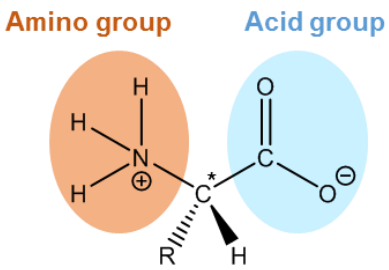


Figure 2 | An amino acid in water at pH 7. R: side chain; *: chiral C_α.

An amino acid in water at pH 7 has a positively charged amino group and a negatively charged acid group, but is neutral overall (as the two charges cancel out). Such compounds are called *zwitterions*.

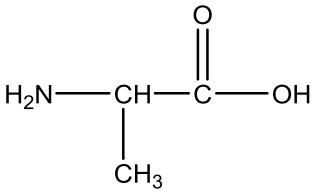
There are twenty standard side chains (R groups), thus there are twenty standard amino acids. Plants can make all of them, whereas animals can only synthesise some of them. The others must be acquired through the diet and are called *essential amino acids*.

What are the 20 standard amino acids?

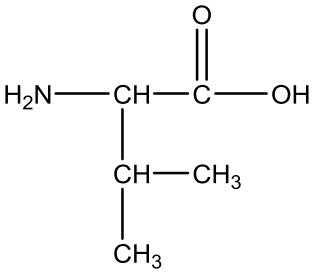
There are twenty standard amino acids, which can be grouped according to their properties, as in Table 1. Each amino acid can be identified by its name (e.g. Alanine), its three-letter code (e.g. Ala) and its one-letter code (e.g. A).

Table 1 | The 20 standard amino acids making up proteins

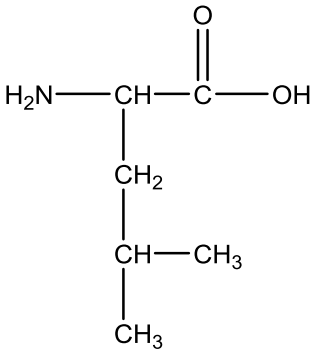
Aliphatic side chains (non-aromatic hydrocarbons)



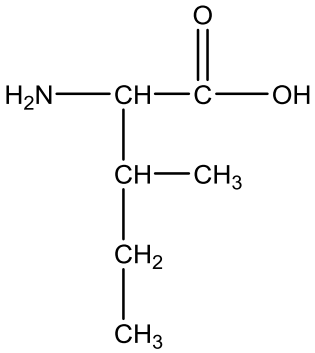
Alanine – Ala – A



Valine – Val – V

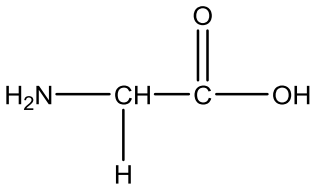


Leucine – Leu – L

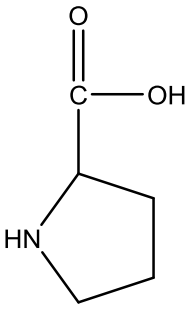


Isoleucine – Ile – I

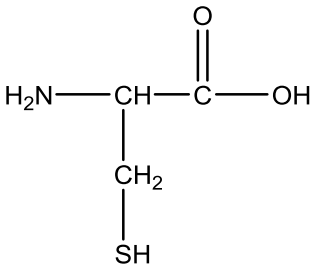
Non-polar side chains



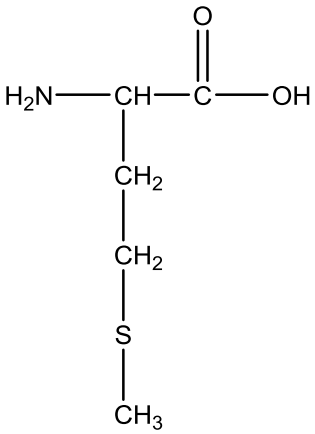
Glycine – Gly – G



Proline – Pro – P

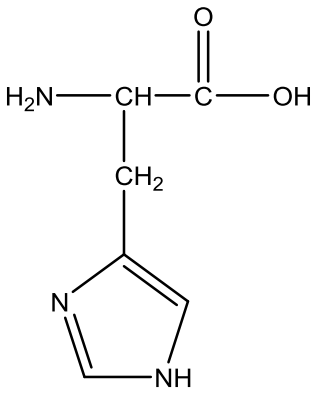


Cysteine – Cys – C

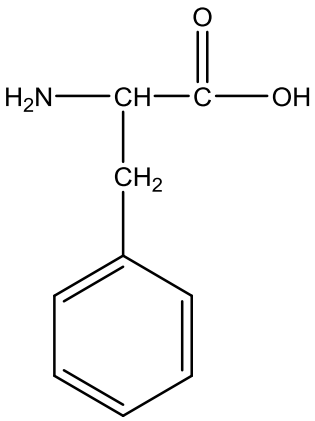


Methionine – Met – M

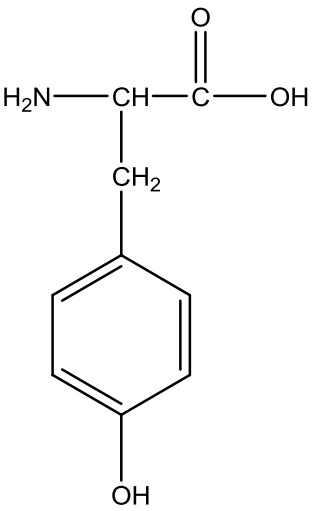
Aromatic side chains



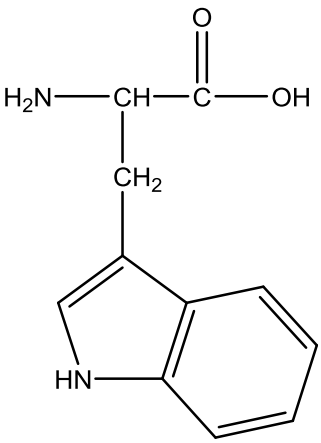
Histidine – His – H



Phenylalanine – Phe – F

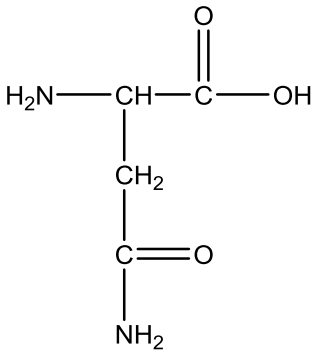


Tyrosine – Tyr – Y

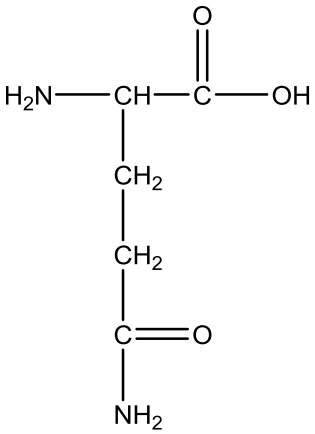


Tryptophan – Trp – W

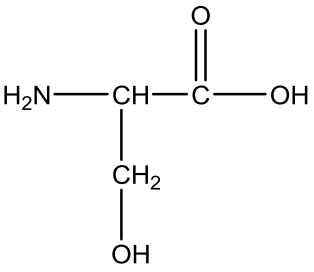
Polar side chains



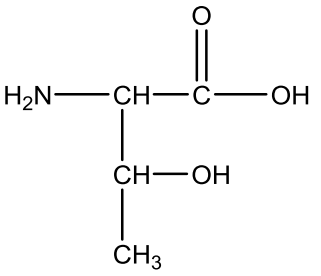
Asparagine – Asn – N



Glutamine – Gln – Q

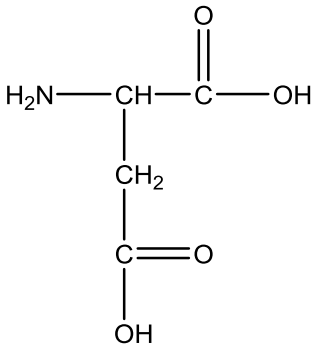


Serine – Ser – S

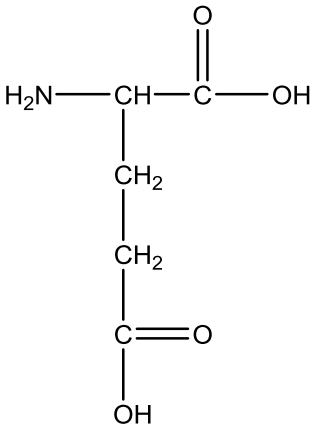


Threonine – Thr – T

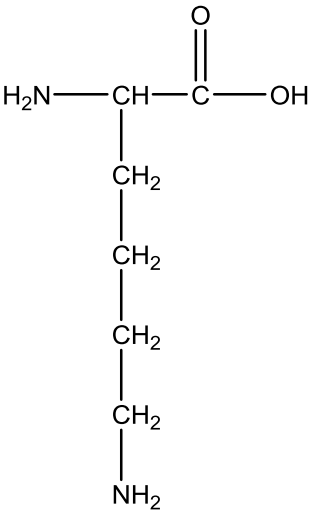
Charged side chains



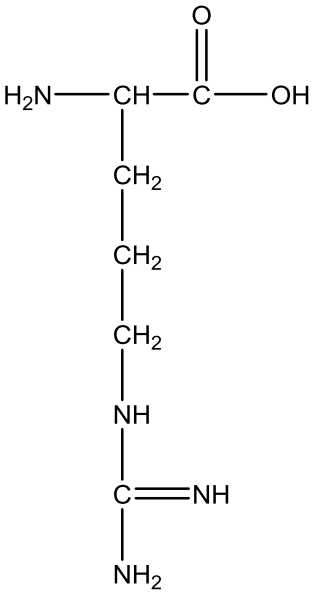
Aspartic acid – Asp – D



Glutamic acid – Glu – E



Lysine – Lys – K



Arginine – Arg – R

At pH 7, aspartic acid and glutamic acid lose the proton on their side chain carboxylic acid, making them negatively charged. Lysine and arginine, on the other hand, gain a proton attached to a nitrogen atom in their side chain, making them positively charged.

Some amino acids may belong to several categories, depending on the conditions in which they find themselves. For instance, at pH 7, a small proportion of histidine molecules will have an extra hydrogen attached to one of the nitrogen atoms in the side chain, making the side chain positively charged.

What is the peptide bond?

Amino acids are joined together to form proteins. The covalent bond between two amino acids is the result of a condensation reaction (where water is released, as shown in Figure 3) and is called the *peptide bond*. Two amino acids joined together form a *dipeptide* and a longer chain of amino acids is called a *polypeptide*, or sometimes simply *peptide*.

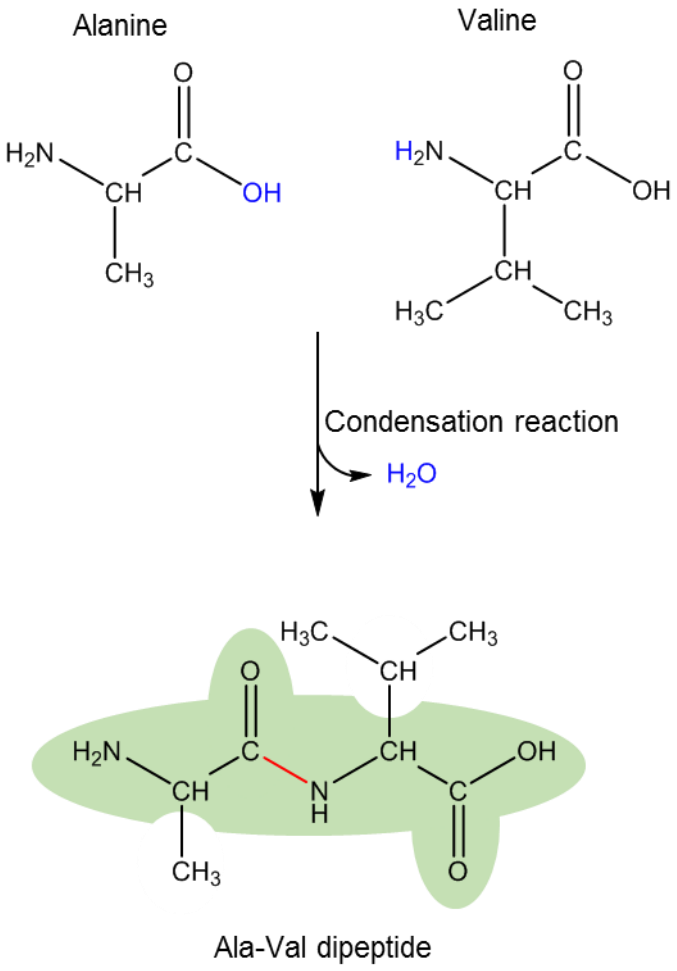


Figure 3 | Formation of an alanine-valine dipeptide. A water molecule (blue) is released by the condensation reaction and the peptide bond (red) is formed. The backbone is highlighted in green.

When they are part of a peptide, amino acids are called *residues*. The peptide bond is formed between the carboxylic carbon of one residue and the nitrogen of the next. The chain made up by the amide nitrogen-C_α-carboxyl carbon of all the residues constitutes the backbone of the peptide (shown in green in Figure 3).

What are the levels of protein structure?

Proteins are made up of polypeptide chains, i.e. polymers of amino acids joined together. The structure of a protein can be studied at four different levels.

Primary structure

⇒ The primary structure of a protein is the sequence of the amino acids that constitute it.

Because of the nature of the peptide bond (cf. above), the backbone of a polypeptide will have a single primary amine at one end and a single carboxylic acid at the other end (at they do not take part in a peptide bond). Those ends are called the N-terminus (primary amine) and the C-terminus (carboxylic acid). The sequence of a polypeptide is always read from the N-terminus to the C-terminus (Figure 4).

N-terminus DIVLTQSPSSLSASLGDTITITCHASQNINVWLSWYQQKPGNIPKLLIYKA
SNLHTGVPSRFSGSGSGTGFTLTISSLQPEDIATYYCQQGQSYPLTFGGGTKLEIKRADAA
PTVSIFPPSSEQLTSGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYS
MSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC **C-terminus**

Figure 4 | The sequence of a polypeptide chain from an antibody, with the N- and the C-termini marked in blue and green respectively.

Secondary structure

The peptide bond between two residues is a single bond, but it is said to have a semi double-bond character. This means that it is particularly rigid for a single bond, forming a planar structure called the amide plane, as shown in Figure 5.

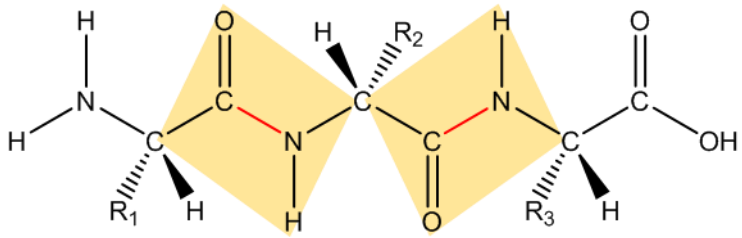


Figure 5 | The amide planes of a tripeptide. Each peptide bond (red) forms a planar structure, the amide plane (yellow), due to is semi double-bond character. $R_{1\text{-}3}$: side chains.

The angles between subsequent amide planes in a polypeptide are called *torsion angles*. They can only adopt certain values, and those values impose certain conformations (or folds) on the backbone.

⇒ The secondary structure of a protein is the local fold of the protein backbone.

Some of those local folds form precise, regular structures, often stabilised by hydrogen bonds. The two most common examples of secondary structure elements are α-helices and β-sheets.

In an α-helix, the polypeptide chain forms a right-handed helical structure with 3.6 residues per turn (Figure 6). The helix is stabilised by hydrogen bonds between the backbone N–H of each residue and the backbone C=O of the amino acid four residues earlier in the sequence. The core of the helix is tightly packed and all the side chains project outward.

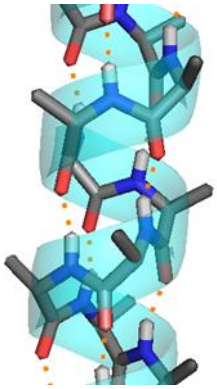


Figure 6 | Example of an α-helix containing alanine residues only. Hydrogen bonds (orange) between backbone atoms four residues apart stabilise the α-helix. Dark grey: carbon; blue: nitrogen; red: oxygen; white: hydrogen. Only the hydrogen atoms involved in hydrogen bonds are displayed.

β-sheets are the other most common type of secondary structure element, shown in Figure 7. They are also stabilised by hydrogen bonds, but between different chains, whereas in an α-helix, the hydrogen bonds are all within the same helix. Similarly to α-helices however, the hydrogen bonds stabilising β-sheets are between backbone N–H and C=O groups.

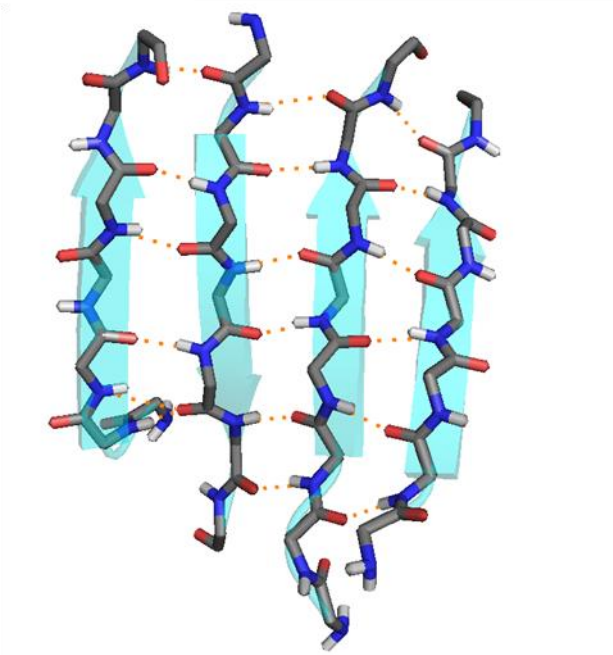


Figure 7 | Example of a β-sheet. All side chain atoms are hidden for simplicity. Hydrogen bonds (orange) between backbone atoms from adjacent β-strands stabilise the β-sheet. Dark grey: carbon; blue: nitrogen; red: oxygen; white: hydrogen. Only the hydrogen atoms involved in hydrogen bonds are displayed.

Proteins also often contain regions of non-repetitive secondary structure. Those regions are called *coils* or *loops*. Although they are not as regular as α-helices or β-sheets, those regions still have a defined structure and should not be confused with the term *random coil*, which is used to describe a protein that has lost its secondary structure (the protein is then said to be *denatured*).

Some amino acids play a specific role in protein secondary structure. For instance, glycine does not have a side chain (it simply has two hydrogen atoms attached to its C_α) and is therefore able to adopt many more folds than other residues. Proline, on the other hand, has a side chain covalently attached to its backbone nitrogen, which means that it cannot adopt as many conformations as other amino acids, and often disrupts secondary structure elements or introduces kinks in α-helices.

Tertiary structure

⇒ The tertiary structure of a protein is its overall 3D arrangement: the folding of secondary structure elements and the position of side chains.

The hydrophobic effect is responsible for most of the tertiary structure of a protein: it is energetically favourable for the protein to fold and bury its hydrophobic residues within its core, away from the surrounding water.

Other bonds and interactions also help the protein fold into the correct tertiary structure. Disulphide bonds are covalent bonds between the sulphur atoms of two cysteine residues. Salt bridges are electrostatic interactions between a negatively charged side chain and a positively charged one. Hydrogen bonds and van der Waals interactions (between hydrophobic residues) are also involved in the tertiary structure. All those forces, interactions and bonds are shown in Figure 8.

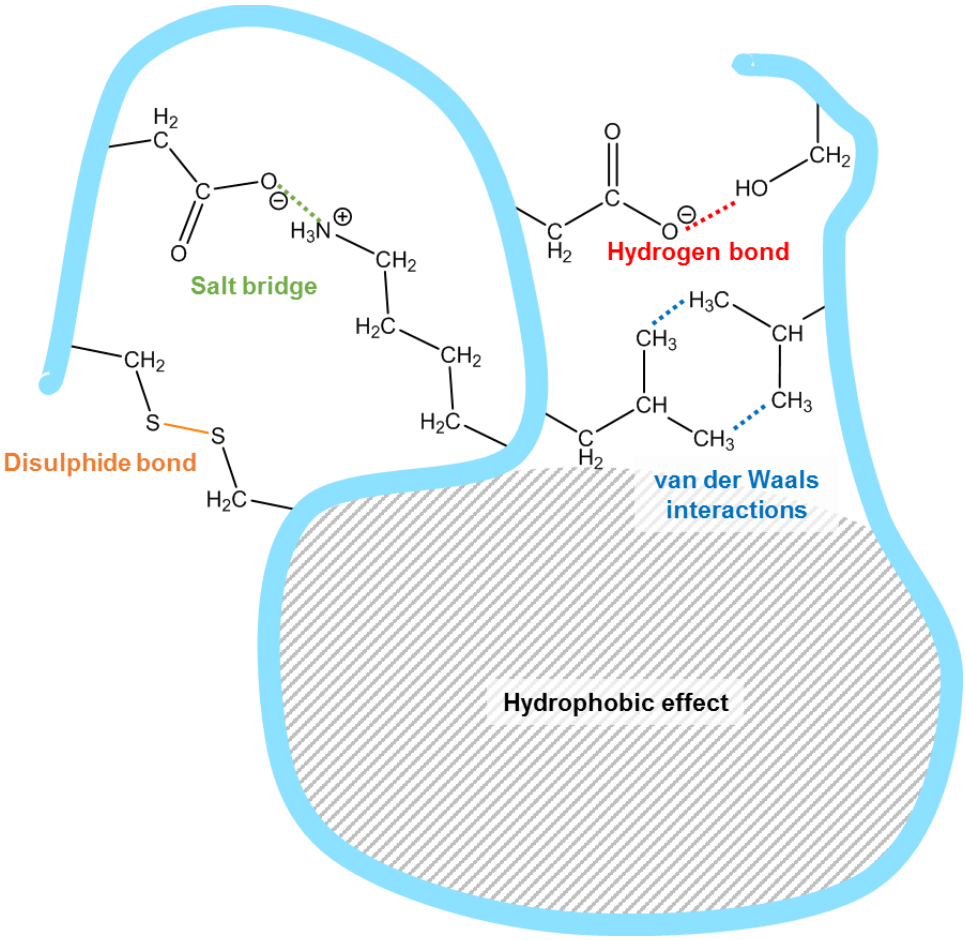


Figure 8 | The forces, bonds and interactions responsible for protein tertiary structure.

Quaternary structure

Some proteins are made up of several polypeptide chains, which assemble once they have adopted their individual tertiary structures. The polypeptide chains may be identical or not: haemoglobin, for instance, has two copies of the same chain and two copies of another, different chain. Antibodies also contain four chains: two heavy chains and two light chains, as shown in Figure 9.

Some proteins are also covalently attached to a non-protein element, e.g. the haem cofactor in haemoglobin (cf. worksheet on haemoglobin).

⇒ The quaternary structure of a protein is the assembly of several polypeptide chains, and sometimes the addition of a non-protein element, to form a functional protein.

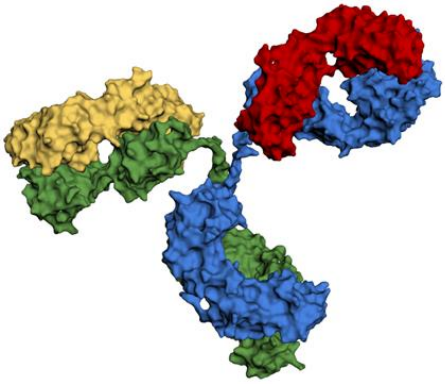


Figure 9 | The quaternary structure of an antibody, comprising two heavy chains (blue and green) and two light chains (yellow and red).

The same forces, bonds and interactions responsible for tertiary structure may be involved in holding different polypeptide chains together.

What is the Protein Databank (PDB)?

When a protein structure is determined experimentally, the 3D coordinates of its constituting atoms are stored in the Protein Databank (PDB), in a PDB file. The Protein Databank is the result of a worldwide effort to collect all known structures of large biological molecules (proteins, DNA and RNA) in standardised files, allowing anyone to visualise them using tools like EzMol. Each PDB file can be easily accessed using its unique, 4-character PDB ID (e.g. 2HHB for deoxyhaemoglobin).

Different national and international entities collaborate to contribute to the global Protein Databank, including the RCSB PDB in the United States or the PDBe in Europe. They both store the same 3D coordinates, but they provide different kinds of information and annotations about each structure.

How are protein structures determined?

There are three main techniques for solving the structure of a protein. The first, which has contributed the most structures to the PDB, is X-ray crystallography. The protein is crystallised, and X-rays are shot at the crystals. The crystals *diffract* the X-rays (i.e. they change their direction), and the way those X-rays are diffracted depends directly on the structure of the protein. Then, a diffraction pattern is recorded, and the structure of the protein can be calculated from it.

The second technique is called nuclear magnetic resonance. The protein is in solution (and not in a crystal) and is placed inside a magnetic field. The protein is irradiated with electromagnetic waves, which will excite the nuclei of its constitutive atoms. After a time, those nuclei relax and, in doing so, produce a signal that reveals information about the other nuclei around them. Then, all that information is pieced together to determine which atoms are near which other atoms in the protein, therefore solving the 3D structure.

The third major technique is cryo-electron microscopy, for which the 2017 Nobel Prize in Chemistry was awarded to Joachim Frank, Richard Henderson and Jacques Dubochet. The protein is neither in a crystal nor in solution, but this time in a thin layer of very cold ice. An electron microscope fires electrons at the protein sample and those electrons are *scattered* (i.e. deflected) when they hit the sample. This produces an image of the protein, which is recorded. This phenomenon is very similar to what happens in a 'normal' light microscope, except that photons (from light) are replaced by electrons, allowing the imaging of much smaller samples. Thousands of images are recorded, with the protein in all possible orientations, and they are then assembled back together to create a 3D model representing the structure of the protein.

What is the resolution of a protein structure?

The resolution of an experimentally determined structure is the smallest distance between two distinguishable features. For instance, if a structure has a resolution of 3Å (0.3nm or 0.0000000003m), it means that we can distinguish two atoms which are 3Å apart or more, but not if they are closer than 3Å. The higher the resolution (i.e. the smaller the number), the better the structure is.

The resolution is often expressed in Ångströms (Å), as it is the most useful unit to describe the length of covalent bonds between atoms (1Å = 10^{-10} m). Table 2 is a general guide of what can be seen at different resolutions.

Table 2 | Examples of features revealed at different resolutions

Resolution	Features
6Å	General shape of the protein and some α-helices.
4Å	Backbone of the protein, secondary structure.
3.5Å	Start to see side chains.
2.7Å	Can see side chains and start seeing water molecules.
1.5Å	Start reaching atomic resolution, where we can make out two covalently bonded carbon atoms.
1.2Å	Can distinguish almost any two covalently linked atoms, except hydrogen (1.2Å is the length of a C=O double bond).

2.7Å is a good resolution for a structure solved by X-ray crystallography, but many structures now achieve much higher resolution: most structures in the PDB are between 1.8Å and 2Å resolution. With cryo-electron microscopy, it is difficult to achieve such high resolutions and 3.5Å is considered good, as it allows the visualisation of side chains.