2.1: Amino Acids and Proteins

François Baneyx
Department of Chemical Engineering, University of Washington
baneyx@u.washington.edu

About 40-to-70% of the dry weight of living matter consists of proteins that are made up of 20 standard α-amino acids (Table 1). At the exception of proline, all amino acids carry a primary amino group and a carboxylic acid group on the same C_α carbon (Fig. 1). This atom is a chiral center and all amino acids (except for glycine) are optically active molecules that exist in two enantiomeric forms known as "L" and "D". Proteins synthesized by ribosomes through translation are exclusively made up of L-amino acids (Fig. 1) and are therefore chiral. This explains why certain enzymes exhibit much higher affinity for a given stereoisomer of their ligand relative to the other. On the other hand, small peptides synthesized in a nonribosomal fashion may contain a combination of L- and D-amino acids. Because the pK_a of α-carboxylic groups is around 2.2 while that of α-amino groups is around 9.4, amino acids are zwitterionic molecules at neutral pH, a characteristic that enhances their solubility in aqueous solvents.

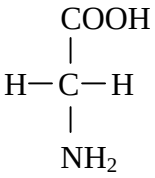
Fig. 1 General structure of an L-amino acid in its neutral and zwitterionic forms. Twenty different R-groups are present in the standard set of amino acids.

Table 1 shows that the side chains (or R groups) of amino acids have very different physico-chemical properties, which confer each amino acid unique characteristics. These include differences in size, polarity, aromaticity, flexibility, charge, reactivity and ability to H-bond and cross-link. Such varied properties are ultimately responsible for the wide range of structure and functions exhibited by proteins. Among the 20 amino acids, glycine and proline are conformationally important. Glycine, which essentially lacks a side chain, can adopt conformations that are sterically forbidden to other amino acids. As a result, glycine confers a large amount of local flexibility to a polypeptide, allowing close packing of the structure. On the other hand, proline is a very rigid amino acid due to the steric constraints imposed by the cyclic nature of its pyrrolidine side group.

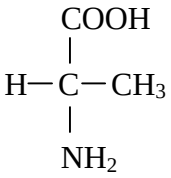
Any two amino acids have the ability to polymerize through a condensation reaction that involves the elimination of a water molecule and the formation of CO-NH linkage known as a peptide bond (Fig. 2). The resulting dipeptide has one free α-amino group and one free, non side chain, carboxyl group (on the left and right of the structure shown in Fig. 2).

Table 1 The twenty standard amino acids.

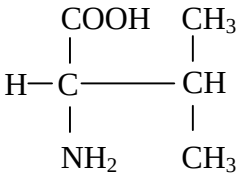
Nonpolar, aliphatic



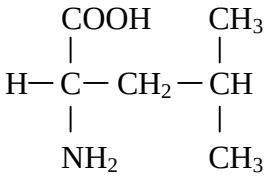
Glycine (Gly, G)
M_r 57.05
Occurrence: 7.2%



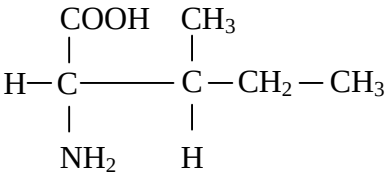
Alanine (Ala, A)
M_r 71.09
Occurrence: 7.8%



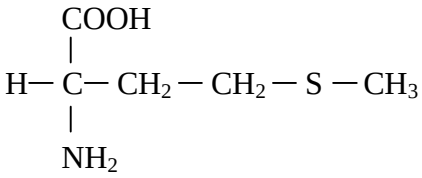
Valine (Val, V)
M_r 99.14
Occurrence: 5.9%



Leucine (Leu, L)
M_r 113.16
Occurrence: 9.1%

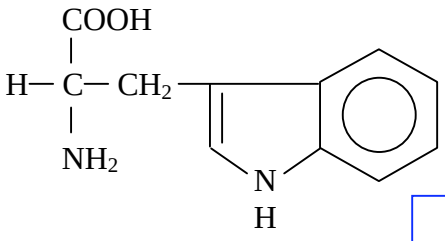


Isoleucine (Ile, I)
M_r 113.16
Occurrence: 5.3%

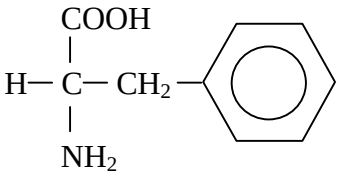


Methionine (Met, M)
M_r 131.19
Occurrence: 2.2%

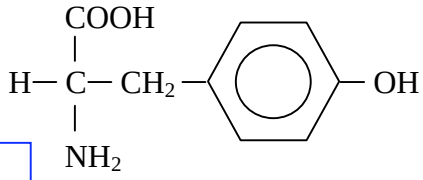
Aromatic



Tryptophan (Trp, W)
M_r 186.21
Occurrence: 1.4%



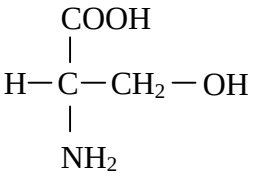
Phenylalanine (Phe, F)
M_r 147.18
Occurrence: 3.9%



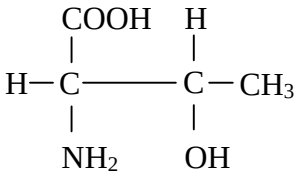
Tyrosine (Tyr, Y)
M_r 163.18
Occurrence: 3.2%

Table 1 (Continued)

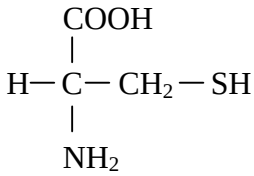
Polar, uncharged



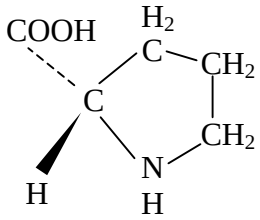
Serine (Ser, S)
M_r 87.08 p$K_a \approx 16$
Occurrence: 6.8%



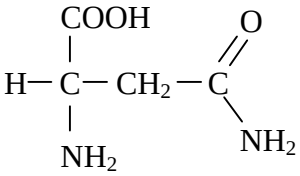
Threonine (Thr, T)
M_r 101.11 p$K_a \approx 16$
Occurrence: 5.9%



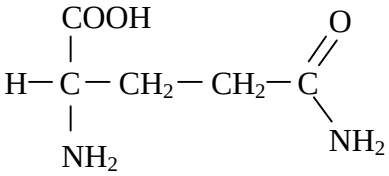
Cysteine (Cys, C)
M_r 103.15 p$K_a \approx 8.35$
Occurrence: 1.9%



Proline (Pro, P)
M_r 97.12
Occurrence: 5.2%

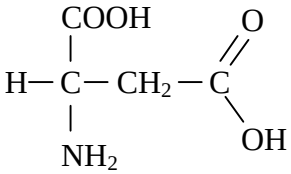


Asparagine (Asn, N)
M_r 114.11
Occurrence: 4.3%

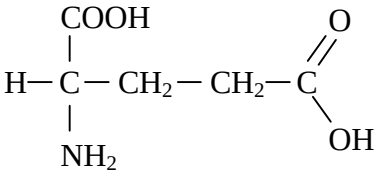


Glutamine (Gln, Q)
M_r 128.14
Occurrence: 4.3%

Negatively charged (acidic)



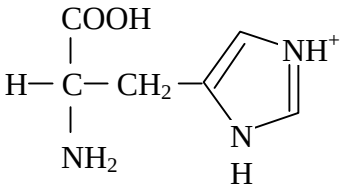
Aspartic acid (Asp, D)
M_r 115.09 p$K_a \approx 3.9$
Occurrence: 5.3%



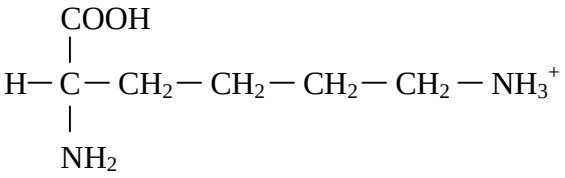
Glutamic acid (Glu, E)
M_r 129.12 p$K_a \approx 4.07$
Occurrence: 6.3%

Table 1 (Continued)

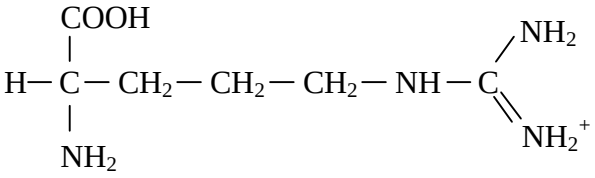
Positively charged (basic)



Histidine (His, H)
M_r 137.14 pK_a ≈ 6.04
Occurrence: 2.3%



Lysine (Lys, K)
M_r 128.17 pK_a ≈ 10.79
Occurrence: 5.9%



Arginine (Arg, R)
M_r 156.19 pK_a ≈ 12.48
Occurrence: 5.1%

The covalent structures, names, three and one-letter abbreviations, molecular mass (in Da), side chain pKa (where appropriate) and likelyhood of occurence of each amino acid in the "average" protein are shown. Amino acids boxed in red are generally external in proteins while those boxed in blue are usually internal. Nonboxed amino acids are ambivalent.

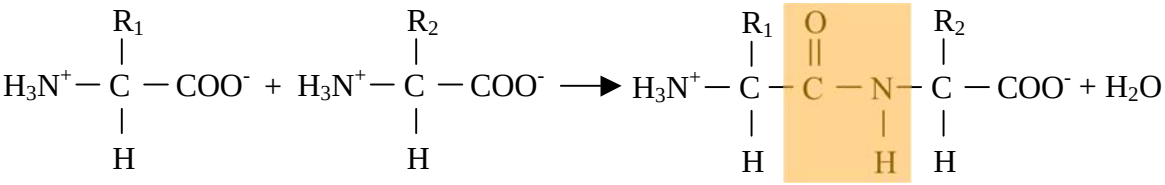


Fig. 2 The condensation of two amino acids leads to the formation of a peptide bond which is a rigid, planar structure (boxed). Bond distances are: 1.53Å for C_α–C; 1.32Å for C–N; 1.46Å for N–C_α; 1.24Å for C=O; and 1.0 Å for N–H.

The reaction can proceed further, leading to tri-, tetra- and ultimately oligo- and polypeptides. When the linear polymer is larger than 50 amino acids (or residues) in length, it is referred to as a protein.

The structure of proteins can be described at four different levels. The *primary structure* refers to the linear sequence of amino acids. It is ultimately responsible for the fold and function of a protein. By convention, the primary sequence is written starting with the free amino terminus and moving towards the carboxy terminus. Thus, an oligopeptide consisting of eight amino acids may be written Ala-His-Arg-Gly-Gly-Ile-Leu-Trp using the three-letter code or AHRGGILW using the one-letter code. With 20 different amino acids to choose from, the diversity of primary structures is staggering. While not all possibilities are used in Nature, it is sobering to realize that a modest, 100 amino acid long protein could be made up of 20^{100} = 1.27 x 10^{130} combinations of the 20 standard amino acids, a number larger than that of all atoms present in the universe ($\approx 10^{79}$)!

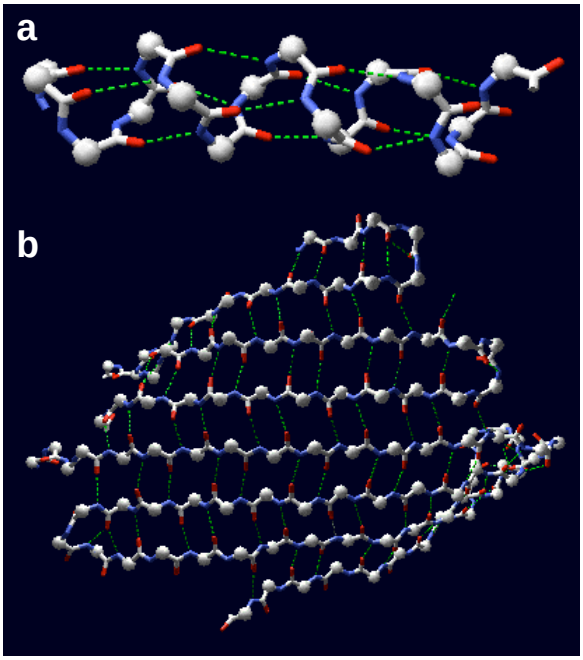


Fig. 3 Structures of an α-helix (a) and a β-sheet (b). Hydrogen bonds between main chain peptide bonds are shown by dashed green lines. Gray balls represent R-group locations.

Proteins do not remain linear and adopt complex three-dimensional shapes that result (mainly) from non-covalent interactions. The *secondary structure* describes the local ordering that stems from hydrogen-bonding within the protein backbone. There are two common secondary structure elements in proteins: α-helices and β-sheets (or β-pleated sheets). Typical globular proteins contain different proportions of α-helices and β-sheets (*see* Fig. 5). Nevertheless, some proteins may be completely devoid of one or the other element of secondary structure. The α-helix (Fig. 3a) is a right-handed spiral whose structure repeats itself every 5.4Å (the pitch) and consists of 3.6 amino acids per turn. This is a very stable arrangement in which every main chain C=O and N–H groups are hydrogen-bonded to a peptide bond located four residues away (e.g., O_i is H-bonded to N_{i+4}). The peptides bonds are essentially parallel to the helix axis and C=O groups and N-H groups point in opposite directions (Fig. 3a). The side chains point outward from the helix axis and generally towards its amino terminus.

In a β-sheet, two or more polypeptide chains, termed β-strands, run side by side and are stabilized by regular hydrogen bonds between main chain C=O and N–H groups (Fig. 3b). Each β-strand exhibits a zig-zag appearance and the side chains of neighboring residues within a strand are separated by 3.5Å and point in opposite directions. This means that a peptide in a β-strand structure will be far more extended than a peptide in an α-helical

structure for which axial separation of adjacent residues is 1.5Å. β-strands can run in the same direction, giving rise to a parallel β-sheet, or in opposite directions, yielding an antiparallel β-sheet (Fig. 4a,b). Antiparallel and parallel β-sheets of less than six strands are common in globular proteins. Mixed β-sheets in which some strands are parallel and others are antiparallel to each other also exist but are seldom encountered in proteins (Fig. 4c). It should be noted that while all hydrogen bonds stabilizing an α-helix involve the same element of secondary structure, β-sheets arise as a result of hydrogen-bonding between different segments of a protein chain.

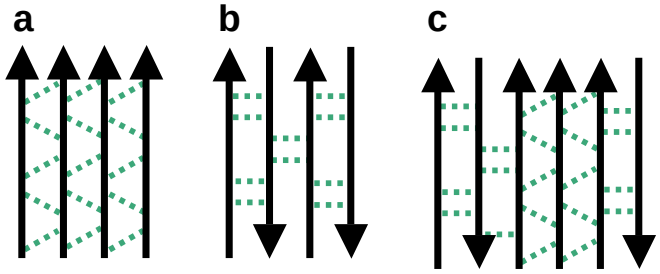


Fig. 4 Three types of β-sheets: parallel (a), antiparallel (b) and mixed (c). β-strands are shown as arrows and hydrogen bonds by dashed green lines.

The *tertiary structure* refers to the global folding of a protein chain, which includes that of secondary structure elements together with the packing of all amino acids side chains. The hydrophobic effect is a major driver of tertiary structure formation. Typically, nonpolar amino acids such as valine, leucine, isoleucine, methionine and phenylalanine will be found in the core of a protein and away from water. On the other hand, charged residues (arginine, histidine, lysine, asparatic acid and glutamic acid) are usually surface exposed. The remaining amino acids are ambivalent and may be found at either location (Table 1). Nevertheless, those buried amino acids that are hydrogen bond donors will almost always be found in a hydrogen-bonded form with an acceptor located in the protein core. Other forces contributing to the adoption of a protein tertiary structure are hydrogen bonding originating from side chains groups, ionic interactions between residues of opposite charge (e.g., glutamic acid and lysine), van der Waals interactions and the formation of disulfide bridges between the sulfhydryl groups of cysteine residues.

Many proteins consist of more than one protein chain. The *quaternary structure* describes how the subunits (or protomers) of oligomeric proteins assemble. Usually, protomers fold into a correct tertiary structure more or less independently and these stable subunits associate with each other to form the final multimeric structure. The regions through which subunits make contact with each other are remarkably similar to the interior of a single-chain protein. Accordingly, subunit association is primarily driven by hydrophobic interactions and hydrogen bond formation. In some cases, interchain disulfide bonds stabilize the structure. Oligomeric proteins may consist of multiple copies of the same subunit (a homo-oligomer; Fig. 5) or multiple copies of two or more different subunits (a hetero-oligomer). There are several reasons for which multimeric proteins have evolved. In the case of enzymes, a larger protein size may help fine-tune the position of the amino acid side chains involved in the architecture of the catalytic active site or serve as an

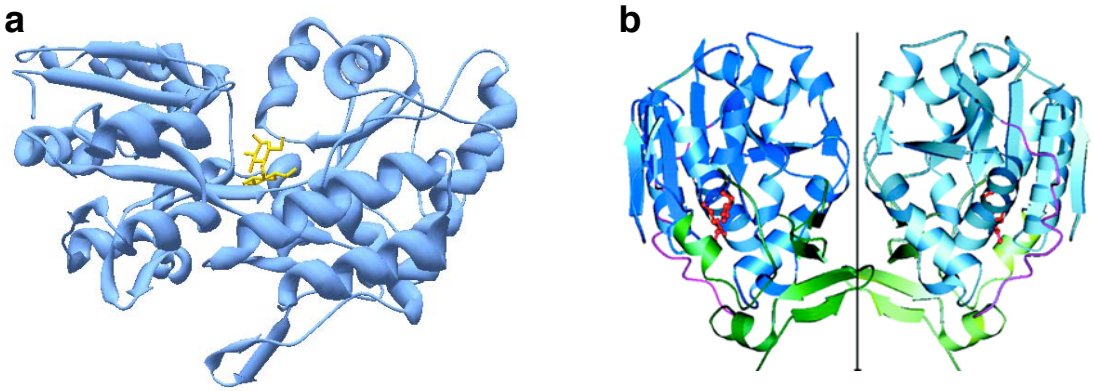


Fig. 5 Ribbon structures of *E. coli* maltose binding protein (a) and Hsp31 (b). Maltose binding protein is a monomeric protein that binds to the small sugar maltose (yellow). Hsp31 is a homodimer composed of two identical subunits. The 2-fold symmetry axis is shown.

efficient way to increase activity if each subunit contains an active site. More importantly, oligomeric arrangements are very important for the structural regulation of protein activity.