II. Proteins

1. Biological importance

Proteins are the major functional macromolecules of life [1] whose properties recommend them as therapeutic agents, catalysts, vaccines and materials. Among some of their important functions within organisms are: catalyzing metabolic reactions, intracellular molecular transporting, cell signaling and DNA replication. An alteration in any of these functions can lead to major negative consequences to the overall health of the organism.

Mutations in proteins can cause them to lose their function and are the source of many diseases. In some cases, metabolic pathways can be affected by the impaired catalytic activity of a particular protein. In other cases, when structural properties are altered, the loss of a physical function can be experienced. Some misfolded proteins, called infectious prions, can cause normal folded proteins to also become misfolded and can damage neurons, giving the affected brain a spongiform appearance. In a similar way, diseases can stem from proteins that gradually precipitate to form fibrils, long chains of polymerized sheets, a process called amyloidosis. Approximately 50% of human cancers are caused by mutations that lower the stability of a protein that usually has the role to suppress the formation of tumors. In order to restore function or to destroy pathogens or cancers, current therapeutic agents target enzymes and receptors, two different types of proteins with respect to their function [1].

The properties and functions of cells and organisms are determined to a great extent by the proteins that they are able to make. Although the functions of proteins inside the cell are vast and diverse, their common mechanism of action is to bind to a substrate and act upon this interaction. This binding always shows great specificity, meaning that a protein can usually recognize just one or a few molecules out of many thousands that it encounters. This happens because the binding site of the protein has a three-dimensional structure that only matches a specific substrate, like a lock and key. If only a minor change occurs in the amino acid sequence of the protein, this binding site can have a totally different shape and the binding would not be possible [2].

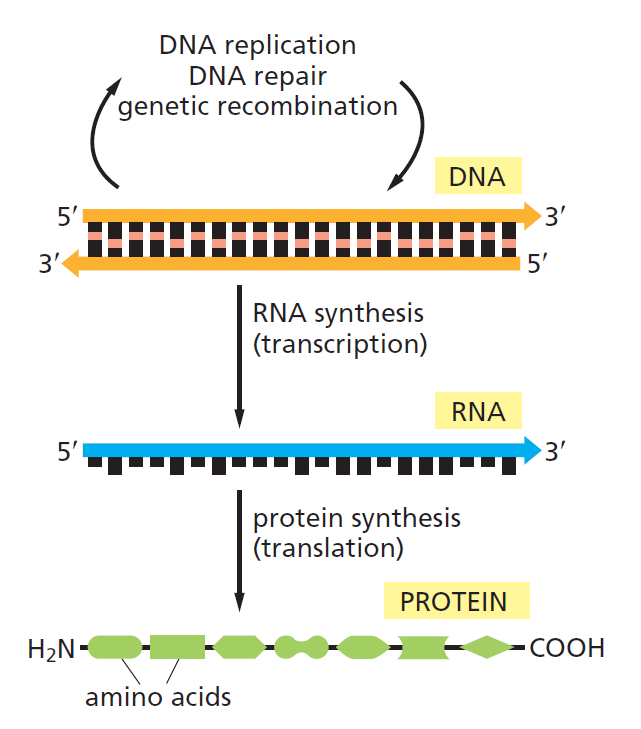
This is one of the most important reasons why the study of protein structure is such an intensely researched domain in the field of bioinformatics and artificial intelligence. If we would know the three-dimensional structure of a protein that we want to target, we could design a molecule that perfectly fits inside its active site and purposefully interacts with the protein, either by enhancing its function or by blocking it, thus restoring the health of the organism. On the other hand, if we would want to act on a specific molecule in the organism that contributes to a disease, we could trace out a protein structure that would attach specifically to that molecule and then synthesize it using currently available techniques.

1. Protein Biosynthesis

The human genome has first been completely sequenced in 2001and has been shown to include approximately 21000 protein-encoding genes which give rise to a much greater number of distinct proteins, but this accounts for only about 1.5% of the total amount of DNA in a human cell [2]. The remaining is considered to be non-coding, regulatory DNA or sequences with functions not yet determined.

The order in which the amino acids are linked to form a specific protein is determined by the sequence of a corresponding gene. The mechanism [2] by which this process takes place has been shown to be universal in all species and it occurs in all living cells. It involves two stages that take place in two different regions of the cell.

The first step in producing proteins occurs in the nucleus of the cell, where the information from the DNA is transferred to another type of molecule capable of holding genetic data, the RNA. This process is called transcription and results in an intermediary product that is able to exit the nucleus and carry the information to the ribosomes, where the second stage takes place. The ribosomes are small structures in the cytoplasm of the cell that read the strand of RNA and produce the proteins by linking specific amino acids together, according to some particular rules. This process is also called translation, because it basically decodes the information from the 4-nucleotides alphabet of the RNA into the 20-amino acids alphabet of the proteins [2]. The entire operation is summarized in the Figure II.1.



**Figure II.1.** The flow of genetic information from DNA to RNA and proteins [2]

Since it is obvious that the translation from nucleotides to amino acids cannot be accounted for by a direct one-to-one correspondence, the scientists tried to group together the nucleotides in order to try to solve this genetic code. It was shown in the early 1960s that a sequence of three consecutive nucleotides was able to represent one amino-acid, each group being called a codon. Since there were four different nucleotides in the RNA, there were 43=64 possible combinations. With only 20 amino acids found in the structure of proteins, it was determined that some combinations are redundant and code the same amino acid. The fantastic feature of this genetic code is its universality, as it is applicable in every cell of every living organism [2].

1. Protein Structure

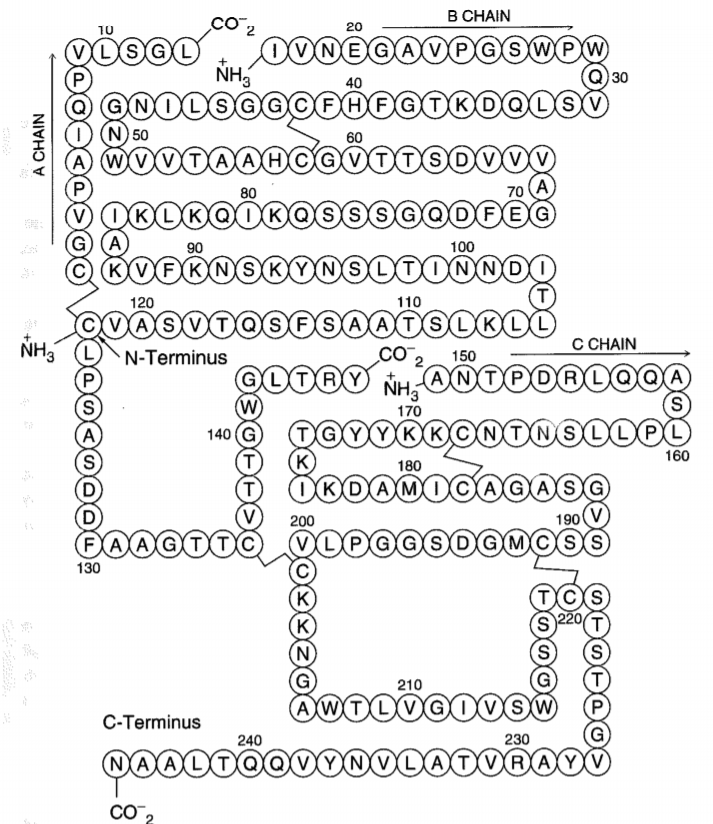
As stated earlier, all proteins contain a linear sequence of amino acids, molecules that contain two types of functional groups: carboxyl group (-COOH) and amino group (-NH2). Each amino acid is linked with the next one by a peptide bond (-CO-NH-) between its carboxyl group and the amino group of the next molecule, giving the main protein two distinct ends: N-terminal end, with the free amino residue, and C-terminal end, with the last carboxyl residue. This is important because the counting of amino acids always starts from the N terminus [3].

Proteins contain an array of 20 different amino acids, listed in Table II.1, along with their abbreviations and the polarity of the side chains. There are an equal number of both polar (hydrophilic) and nonpolar (hydrophobic) molecules, a property that greatly affects the way in which the protein’s three-dimensional shape will look like [2].

**Table II.1**. The 20 amino acids commonly found in proteins [2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Polarity | Amino acid | Abbreviation (three letters) | Abbreviation (one letter) | Type of side chain |
| Polar | Aspartic acid | Asp | D | Negatively charged |
| Glutamic acid | Glu | E | Negatively charged |
| Arginine | Arg | R | Positively charged |
| Lysine | Lys | K | Positively charged |
| Histidine | His | H | Positively charged |
| Asparagine | Asn | N | Uncharged polar |
| Glutamine | Gln | Q | Uncharged polar |
| Serine | Ser | S | Uncharged polar |
| Threonine | Thr | T | Uncharged polar |
| Tyrosine | Tyr | Y | Uncharged polar |
| Nonpolar | Alanine | Ala | A | Nonpolar |
| Glycine | Gly | G | Nonpolar |
| Valine | Val | V | Nonpolar |
| Leucine | Leu | L | Nonpolar |
| Isoleucine | Ile | I | Nonpolar |
| Proline | Pro | P | Nonpolar |
| Phenylalanine | Phe | F | Nonpolar |
| Methionine | Met | M | Nonpolar |
| Tryptophan | Trp | W | Nonpolar |
| Cysteine | Cys | C | Nonpolar |

The folding of a protein chain is also determined by many other interactions between residues from different regions. In Figure II.2 we have the amino acid sequence of the enzyme chymotrypsin, using the one-letter abbreviations from Table II.1. The enzyme is originally synthesized as a long polypeptide chain, but after the formation of the disulfide bridges between different cysteine residues, the initial chain is cleaved in three different pieces. We can see here the importance of these weaker interactions to the overall shape of the molecule [1].



**Figure II.2**. Amino acid sequence of the enzyme chymotrypsin, consisting of three chains linked by weaker bonds [1]

Biologists have studied protein folding in a test tube using highly purified molecules and have found that adding certain solvents that disrupt the interactions between amino acids males the proteins unfold and converts it to a flexible polypeptide chain, losing its conformation. But when removing the solvent, the protein often refolds spontaneously into its original conformation, meaning that the amino acid sequence holds all of the information needed for specifying the three-dimensional shape of a protein [2]. The final folded conformation of any protein chain is generally one that minimizes its free energy.

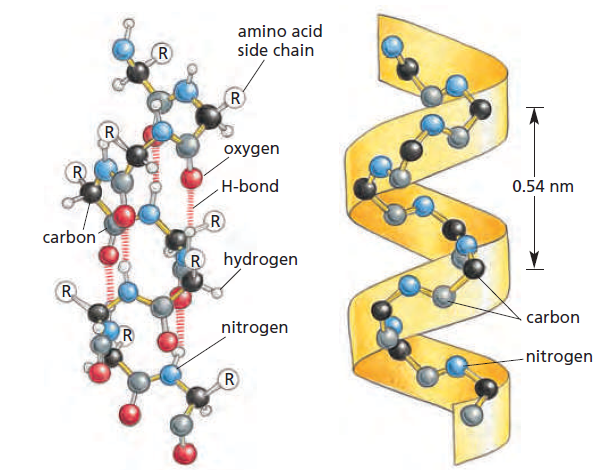
Proteins can be analyzed at four levels:

* primary structure
* secondary structure
* tertiary structure
* quaternary structure

This hierarchy [4] facilitates the description and the understanding of proteins and it does not aim to precisely describe the laws that produce protein structures. It is an abstraction that intends to make the study of protein structures more manageable.

The primary structure describes the sequence of amino acids in a linear order, starting with the N-terminal region of the protein chain. Secondary structure can be described as the local spatial conformation of a polypeptide backbone, excluding the constituent amino acids’ side chains. The major elements of the secondary structure are the α-helix and the β-sheet, with some regions of disorganized amino acids. The tertiary structure refers to the distribution of secondary structures in a three-dimensional space and is greatly influenced by weaker forces and interactions between side chains or with the surrounding medium. The quaternary structure refers to the overall spatial arrangement of polypeptide subunits within a protein composed of two or more polypeptide chains [3, 4].

Although the overall conformation of each protein is unique, when we compare the three-dimensional structures of many protein molecules, two regular folding patterns are often found within them. Both patterns were discovered more than 60 years ago from studies of hair and silk and are particularly common because they involve hydrogen bonds only between the atoms in the polypeptide backbone, and not those in the amino acid side chains. In each case, the protein chain adopts a regular, repeating conformation [2, 5]. These two secondary structure elements are commonly formed because they maximize formation of stabilizing intramolecular bonds and minimize repulsion between adjacent side chain groups, while also being compatible with the rigid nature of the peptide bonds [3].

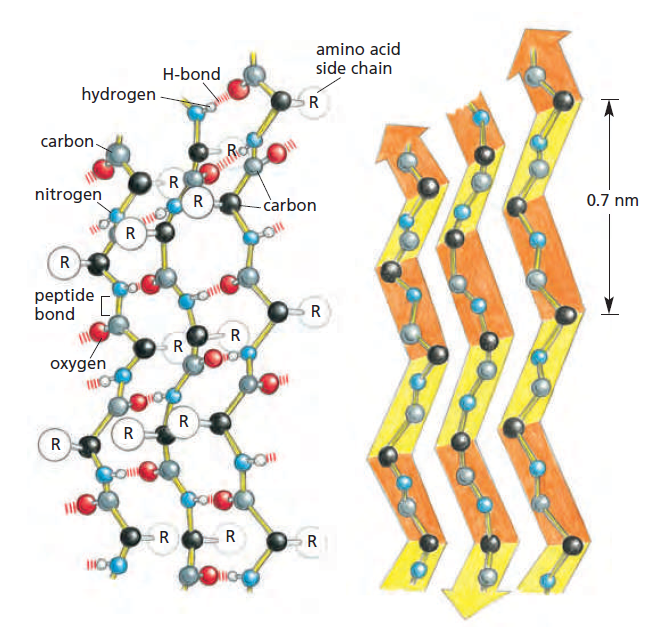


**Figure II.3** The regular conformation of the polypeptide backbone in the α-helix [2]

The first folding pattern was called the **α-helix** and was identified in the protein α-keratin, which can be found in large quantity in the skin, hair and nails. This was able to explain the strength and elasticity of this protein and account for the fiber appearance at the X-ray diffraction. An α-helix is generated when a single polypeptide chain twists around itself to form a rigid cylinder, with a hydrogen bond between every fourth peptide bond and the amino acid side chains protruding outward from the helical backbone [2]. This gives rise to a regular helix with a complete turn every 3.6 amino acids, as can be seen in Figure II.3.

Stretches of α-helix can vary in length from one single helical turn to more than 10 consecutive turns, with the average length being of about three turns, in globular proteins [3]. The proteins located in the cell membrane, having transport and receptor functions, contain extensive regions of α-helix. Those portions of proteins that cross the membrane usually do so as α-helices composed of amino acids with nonpolar side chains. The hydrophilic polypeptide backbone is therefore shielded from the hydrophobic environment of the membrane by its protruding nonpolar side chains [2].

The other major structural element found in globular proteins is the β-sheet and it was first observed in the β form of keratin fibers. Although it was discovered a year after the first element, an approximate understanding of its molecular structure was achieved earlier than for the α structure [5]. The cores of many proteins contain extensive regions of β-sheet, which can be formed from neighboring sections of the polypeptide backbone that run in the same direction (parallel chains) or from a polypeptide backbone that folds back on itself, with each section running in the opposite direction to the one next to it (antiparallel chains), as in the figure below [2]. Both types build a very rigid structure held together by bonds between neighboring chains.



**Figure II.4** The regular conformation of the polypeptide backbone in a β-sheet [2]

Although these are the two major secondary structures that can be identified when looking at protein conformations, most proteins consist of several segments of α-helix and/or β-sheets separated from each other by various loop regions, or coils. These regions can vary in shape and length and allow the overall molecule to fold into a compact tertiary structure [3]. Beside their role in connecting regular secondary elements, loop regions often contribute directly to the biological function of the protein and are exposed to solvent.

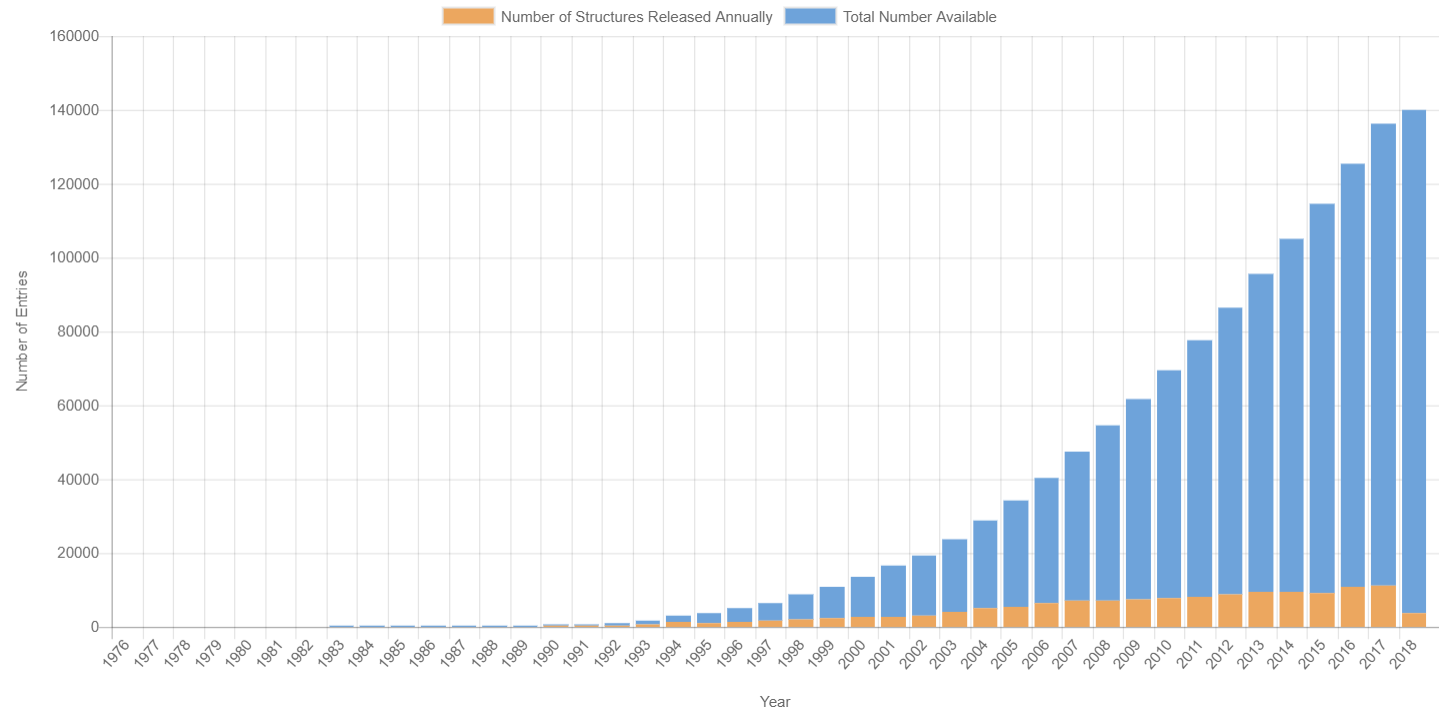
The major driving force for the folding of proteins seems to be hiding and clustering of hydrophobic side chains to minimize their contact with the water around the macromolecule. The basic requirements [1] for folding are that (1) the resulting structures are compact and so have minimal hydrophobic areas in contact with solvent and (2) that the hidden groups that are bound by hydrogen bonds are all paired. The formation of the two secondary structures helps with the second point of the previous statement, as it maximizes the pairing of the hydrogen bonding groups. The helices and sheets pack by stacking their amino acid side chains.

1. Protein Structure Online Databases

To facilitate the understanding of, and access to the information available for protein structures, researchers have been gathering and structuring it in online databases, making the data easier to be queried and organized. In order to determine the unique primary structure through quaternary structure of a protein, different physico-chemical methods are employed, such as: X-ray crystallography, NMR spectroscopy or 3D electron microscopy [6].

Fifteen years after the determination of the first protein crystal structure corresponding to myoglobin, the **Protein Data Bank** (PDB) was created in 1971 [6] and initially contained only seven protein structures. The PDB currently archives approximately 130 000 entries and is managed by the Worldwide Protein Data Bank, which contributed to the evolution as the single global archive of macromolecular structure data. But in the first 30 years of its existence, the addition of new molecules was sparse and only by the mid-1990s a boost in the number of entries has been seen, as pictured in Figure II.5. This can be attributed to the advances in computer and information technology, which provided the much required computer power for experiment automation, to the introduction of genetic engineering for easy production of basically any protein using bacterial cells and also to the development of powerful X-ray sources.

The RCSB PDB is the US regional center of the PDB and manages the website (rcsb.org) which offers multiple tools for structure query, browsing, analysis and molecule visualization. It enables users to perform simple searches based on PDB ID, name of the macromolecule, sequence or ligand, but also allows them to build complex search combinations of parameters and criteria. The PDB data is organized in hierarchical trees using external classification and annotation systems and visualization options enable the exploration of three-dimensional structure, structure/sequence information and correspondence between the two [8].



**Figure II.5**. Overall growth of released structures per year for the PDB [7]

Besides the PDB, there are many repositories and databases [6] used in structural biology, chemistry, life sciences and pharmaceutical industry, where they are crucial in the drug discovery process. The growing number of macromolecular structures in the PDB provides a solid foundation and increases the scientific potential of derivative data resources. Fold classification databases such as **CATH** and **SCOP** (Structural Classification of Proteins) aim to classify protein folds in terms of evolutionary relationships as well as sequence similarity, and are references for nonredundant folds and domains used by many structural bioinformatic tools. There are also other specialized data resources [6] that catalog and classify different structural aspects: the **Protein Data Bank of Transmembrane Proteins** (PDBTM), the **KnotProt** database (contains three-dimensional structures of proteins that form knots), **MPStruc** (the database of Membrane Proteins of Known 3D structure).

[1] Fersht, A.: Structure and Mechanism in Protein Science, A Guide to Enzyme Catalysis and Protein Folding, W. H. Freeman and Company, New York, 1999.

[2] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell, 6th Edition, Garland Science, Taylor & Francis Group, New York, 2015.

[3] Walsh, G.: Proteins, Biochemistry and Biotechnology, 2nd Edition, Wiley Blackwell, 2014.

[4] Dorn, M., Barbachan e Silva, M., Buriol, L.S., Lamb, L.C.: Three-dimensional protein structure prediction: Methods and computational strategies, Computational Biology and Chemistry, 53(2014), 251-276.

[5] Richardson, J.S: The Anatomy and Taxonomy of Protein Structure, Advances in Protein Chemistry, 34(1981), 167-339.

[6] Wlodawer, A., Dauter, Z., Jaskolski, M.(editors): Protein Crystallography, Methods and Protocols, Springer, New York, 2017.

[7] <https://www.rcsb.org/stats/growth/overall>

[8] Rose, P.W. et al.: The RCSB protein data bank: integrative view of protein, gene and 3D structural information, Nucleic Acids Research, Database issue, 45(2017), D271-D281.