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].

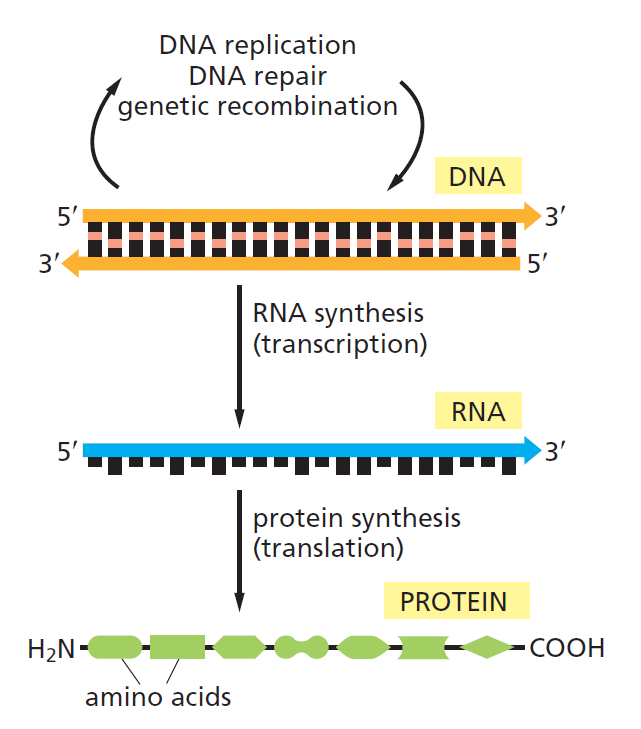
This is one of the most important reasons why the study of protein structure is such an intensely researched domain in the field of bioinformatics and artificial intelligence. If we would know the three-dimensional structure of a protein that we want to target, we could design a molecule that perfectly 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

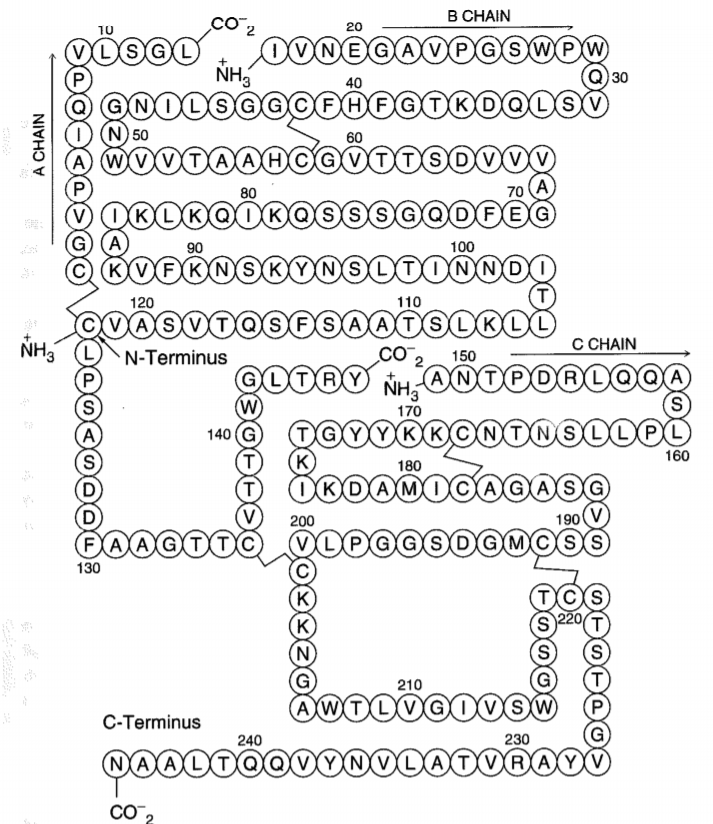
As stated earlier, all proteins contain a linear sequence of amino acids, molecules that contain two types of functional groups: carboxyl group (-COOH) and amino group (-NH2). Each amino acid is linked with the next one by a peptide bond (-CO-NH-) between its carboxyl group and the amino group of the 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 [Dorn2014] 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, Dorn2014].

α-helix

β-sheet

1. Protein Structure Databases

[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.

[Dorn2014] 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.