WEEK 5 ASSIGNMENT

Data Systems in the Life Sciences – BIOL 51000 | Fall 2 2020

Christina Morgenstern

MACROMOLECULAR STRUCTURES

1. Discuss different types of protein structures (with a structural hierarchy)

Proteins are macromolecules that consist of amino acids as their fundamental building blocks. Amino acids are organic molecules with a defined structure: a carboxy (C) group, an amino (N) group and a side chain. There are 20 common amino acids that differ in their properties according to the nature of their side chain. The side chain can be unpolar, polar or positively or negatively charged leading to special characteristics of the amino acid. Amino acids are linked together by peptide bonds to form polypeptides which range from a few amino acids to thousands of building blocks. The resulting linear polypeptides are characterized by a directionality that is the amino end or N-terminus on one side and the carboxyl or C-terminus on the other end of the chain. This linear arrangement of a specified order of the 20 amino acids is known as the primary structure of the protein. It represents a unique sequence of amino acids as determined by the information stored in the DNA. In order for a protein to be functional, it needs to be folded into a dedicated structure. The secondary structure is the next hierarchical structure acting on the primary sequence and results from hydrogen bonding within the members of the polypeptide backbone. Two conformations of secondary structure that lead to local sub-structures can be described: α-Helix and β-pleated sheets. The former structure is arranged as a helix and the latter results in planar structures that can be arranged in parallel or anti-parallel. These secondary structures comprise regions in the protein that fulfil special roles like spanning or anchoring the protein to a membrane. A barrel shape of β-sheets can serve as a channel for the protein aiding the transport of substances from one side of the membrane to the other.

Tertiary protein structure is the next level in the hierarchical structural organization of proteins and results from the interaction of the side chains of the amino acids. Those interactions involve hydrogen bonds, ionic bonds, hydrophobic and Van der Waals interactions as well as strong covalent bonds such as disulfide bridges. Tertiary structure folding leads to a three-dimensional structure of a protein subunit or a final protein.

If two or more polypeptide chains are joined together to function as a single unit (multimer) the quaternary structure level is reached. For example, the protein hemoglobin is made up of 4 subunits that are joined together, two α-chains and two β-chains.

The structure of a protein is vital for its function and therefore changes in the primary structure, certain physical and chemical conditions (temperature, pH) can alter the folding. Denaturation refers to the process of unfolding and leads to biologically inactive molecules.

2. How to determine macromolecular structures

The three-dimensional structure of a protein can be determined by X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR). There are also bioinformatic approaches that aim to predict the 3D structure by applying algorithms.
In order to determine the structure of the protein with X-rays the protein needs to be crystallized which is often a challenge. Next, the X-rays which are electromagnetic radiation with a short wavelength of 0.1 nm are shot onto the crystal. This leads to a scattering of the X-rays by the atoms in the sample leading to a diffraction pattern which can be picked up by a detector. The resulting two-dimensional images are further transformed to yield a three-dimensional model.
In contrast to X-ray crystallography, NMR spectroscopy allows the determination of the 3D structure of proteins in a concentrated solution. The protein sample is placed into a strong magnetic field where some of the atomic nuclei will align with the field because of their spin-active isotopes. Sometimes this process is enhanced by the addition of further spin-active isotopes. The alignments are further detected by sending a pulse of radio waves through the sample leading to a characteristic resonance behavior. The resonance frequencies can then be used to determine the chemical and structural qualities. NMR is especially used for smaller protein structures.

3. How 3D coordinates are used to compare the structure

A protein molecule can be described and localized by its atoms in Angstrom-scale space leading to a XYZ triple. Each XYZ triple is further labeled with an atom, residue and chain value. An atom can be identified by a sequential number, a specific atom name, the name and the number of the residue it belongs to, a one-letter code to specify the chain, and the x, y and z-coordinates [1]. For example, the values 54, ALA, C, 35.4, -9.3 and 102.5 describe the 54^{th} atom, the amino acid alanine, the C-chain of the protein and the localization of the atom in 3D space.
The Protein Data Bank (pdb) file holds this information describing the 3D structures of molecules in a text-based format which can be retrieved via the database Protein Data Bank [2]. Newer formats for PDB entries comprise the PDBx/mmCIF and XML formats.
Coordinates can further be stored in other file formats, such as the Molecular Modelling DataBank format (MMDB). The coordinates of proteins can be retrieved and compared to other structures using databases like MMDB or the Research Collaboratory for Structural Bioinformatics (RSCB)

4. How to determine the quality of your protein alignments

In order to determine the quality of protein structure similarity, distance-based measures such as the Root Mean Square Deviation (RMSD) can be used. It is one of the most widely used quantitative measure of similarity between two superimposed molecular structures. It calculates the sum of the distances between the C_α atoms of all pairs in the structure divided by the number of the pairs and the square root taken:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

[3]

The RMSD value assesses the goodness of fit between two sets of coordinates and thus two aligned protein structures. It is represented in Angstrom and a value smaller than 3 is considered as good fit.

References:

[1] 'PDB101: Learn: Guide to Understanding PDB Data: Dealing with Coordinates', *RCSB: PDB-101*. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/dealing-with-coordinates (accessed Nov. 28, 2020).
[2] 'Protein Data Bank (file format)', *Wikipedia*. Nov. 04, 2020, Accessed: Nov. 28, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Protein_Data_Bank_(file_format)&oldid=987019478.
[3] I. Kufareva and R. Abagyan, 'Methods of protein structure comparison', *Methods Mol Biol*, vol. 857, pp. 231–257, 2012, doi: 10.1007/978-1-61779-588-6_10.