Homology Modeling

Modesto Redrejo Rodríguez

2022-07-07

Table of contents

1 Course Disclaimer

This brief instruction booklet contains the materials for the "Hands-on Protein Modeling" sessions I lectured at the CIVIS Summer School Bioinformatics for non-bioinformaticians, Tübingen, Germany (18-22 July 2022).

Most of the materials are shared with the course *Structural Bioinformatics* that I teach in the Master's Degree in Bioinformatics & Computational Biology @UAM. The contents are also largely inspired in the works of others that shared their course materials, tips and other kind of resources on their own websites, GitHub or Twitter, including Alexandre Bovin, Sergey Ovchinnikov, Martin Steinegger, Carlos Outeiral, among many others. I tried to acknowledge each contribution and I apologize beforehand for those that I may have not mention.

You can reach me by email or Twitter. Please let me know if you find some missing reference. I will also appreciate any suggestion or correction.

Figure 1.1: Link to the website of the Master's Degree in Bioinformatics & Computational Biology at UAM

2 Introduction

2.1 Goals and Warnings

Structural Bioinformatics is a broad discipline that covers structural and computational biology, from visualization and analysis of the structure of biomacromolecules to protein modeling and molecular docking. The field have experienced a great revolution in the last decade. The increase of experimental capacities to analyze structure of proteins and other biological molecules and structures (see Callaway (2020)) and the development of Artificial Intelligence (AI)-assisted structure prediction boosted the capacity of life-science researchers to address a wide variety of questions regarding proteins diversity, evolution and function. The implications of this revolution in biology, biotechnology and biomedicine are still unforeseen.
For a short introductory course on protein modeling, I propose the following three basic objectives:

1. Identify the main applications and limitations of prediction of protein structures in biomedicine and biotechnology.
2. Become familiar with classic and state-of-the-art protein modeling methods.
3. Basic understanding of the result output of a protein modeling experiment and how to evaluate and eventually improve the model quality.

2.2 Warning for future structural biologists

The surrealist Belgian painter René Magritte created a collection of surrealistic paintings entitled ***La trahison des images*** (1928–1929). The most renowned of those paintings show a smoking pipe and the following caption underneath: "Ceci n'est pas une pipe" (This is not a pipe). Indeed! It is the painting of a pipe.

Similarly, a picture of a protein, or a PDF file with the coordinates of a protein structure, is not a protein. It is a representation of ONE structure. Even experimentally determined structures have important limitations that we should always keep in mind: (1) they are a fixed structure whereas proteins *in vivo* are flexible and dynamic and (2) they are subjected to experimental error and they often contain regions of low reliability. That does not mean that protein structures are useless, they can be very useful, but we must be aware of the limitations as well as the applications.

Figure 2.1: Ceci n'est pas une proteine. Source: https://swissmodel.expasy.org/static/course/files/PartIII_quali

3 Before going forward: Protein Structure 101

Although you can make some protein modeling without being an expert in structural biology, a basic understanding of protein structure is strongly advisable. Over the years, I noticed that graduate students in biology, biomedicine and related fields have very different background on protein structure. If you want to review and update your background on protein structure, I recommend you the great review by Stollar and Smith (2020) and the wikipedia articles on protein structures (https://en.wikipedia.org/wiki/Protein_structure), which constituted my main source for this section.

Proteins are key components of life, playing key roles in almost any possible vital function, either as passive, scaffolding elements or as active enzymes that catalyze metabolic reactions. Proteins are built as polymers of amino acids and the sequence of amino acids of a particular protein can be also called **the primary structure** of the protein. Amino acids chains can spontaneously fold up into three-dimensional structures, mostly stabilized by hydrogen bonds between amino acids. The amino acid sequence determine different layers of 3D structure. Each of the 20 natural amino acids has different physicochemical properties that affect its preferred conformation. Thus, a first level of folding is called **secondary structure**, forming common patterns as we will see in a moment.

These stretches of secondary structure patterns can fold in 3D due to interactions between the side chains of amino acids. This is called protein tertiary structure. Finally, two or more individual peptides chains can form a multisubunit proteins that have the so-called **quaternary structure**.

It should be noted that the peptide bond itself cannot rotate as it has double bond-like character. Therefore, rotation can only occur about the bond between the C and the C = O group, (the phi ( ) angle) and the C and the NH group, (the psi ( ) angle). In effect, the polypeptide backbone chain is composed of a repeating series of two rotatable bonds followed by one non-rotatable (peptide) bond. However, not all 360º of the psi and phi angles are possible as neighbouring sidechains can clash due to steric hindrance. In effect, for certain angles and amino acid combinations, the atoms cannot be in the same physical place and this partly explains why some amino acids have a higher propensity (likelihood) to form different types of secondary structure.

Within these restraints, the two principal local conformations that avoid steric hindrance and maximise backbone–backbone hydrogen bonding are the -helix and the -sheet secondary structures. The -helix is a right-handed coil in which backbone NH group hydrogen bonds to the backbone C = O group of the amino acid located four residues earlier along the protein

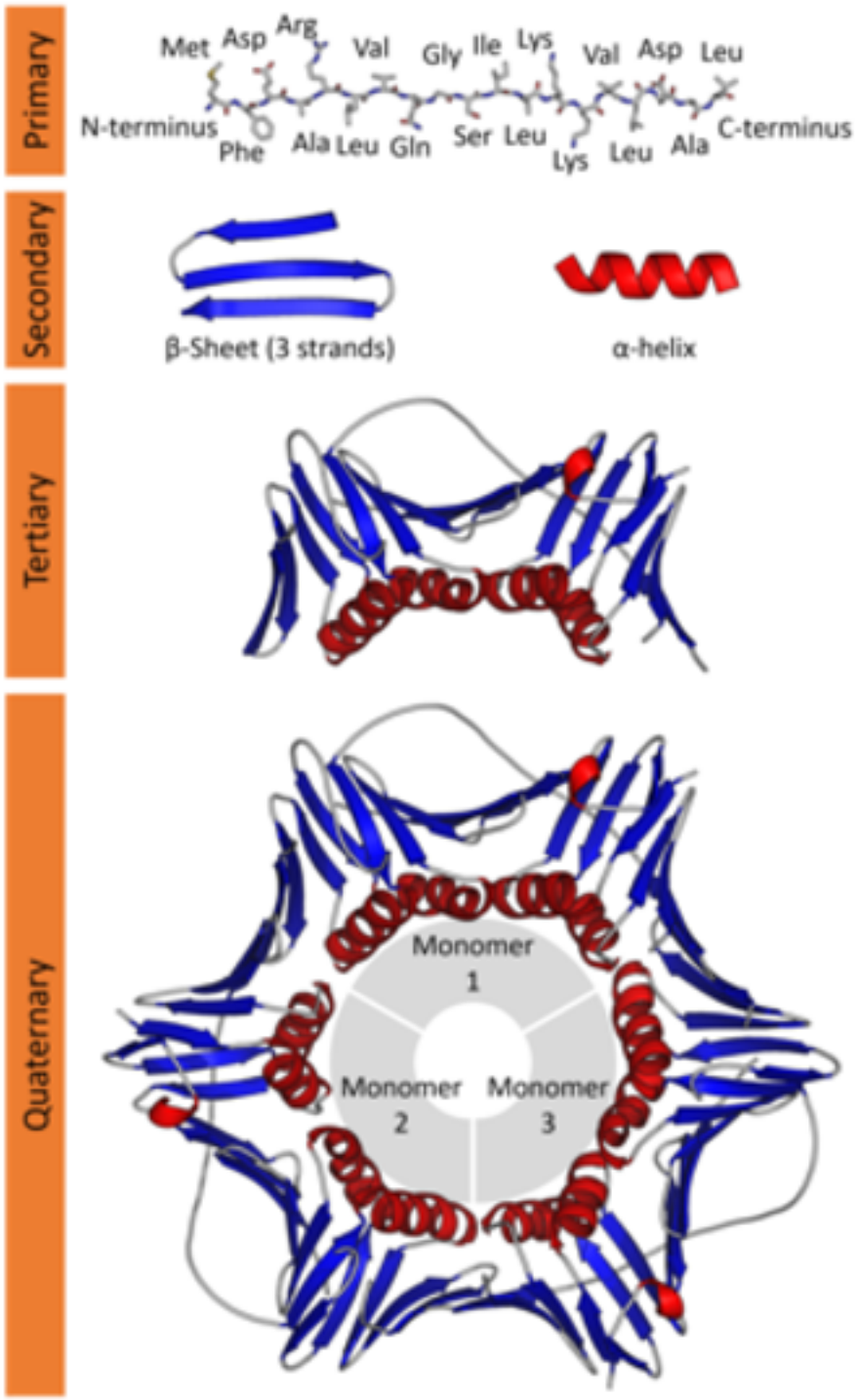

Figure 3.1: Protein structure levels, using human PCNA (PDB 1AXC) as an example.

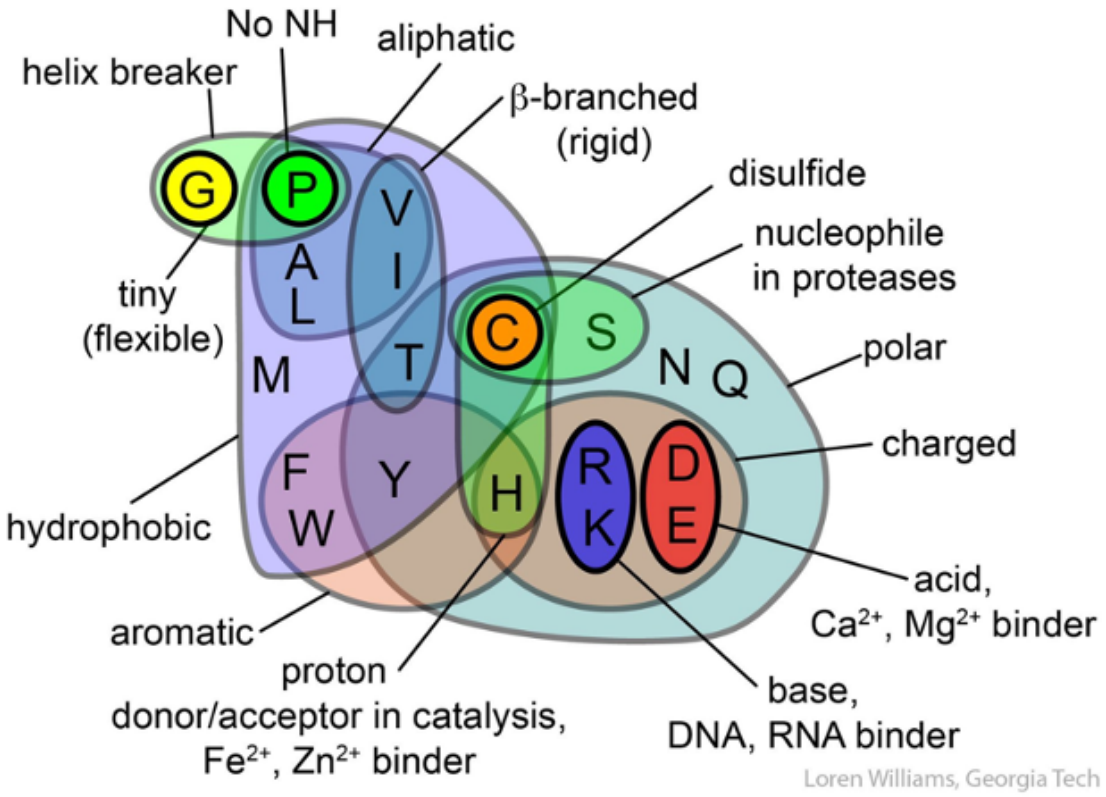

Figure 3.2: Amino acids clasification by type

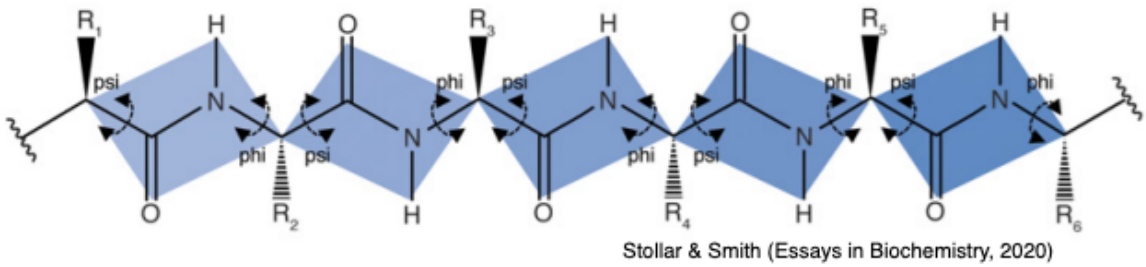

Figure 3.3: Scheme of a generic polypeptide chain. Residue side chains are denoted as R. Coloured rectangles indicate sets of six atoms that are coplanar due to the double-bond character of the peptide bond. Arrows indicate the bonds that are free to rotate with the angle of rotation about the N–C known as phi and about the C –C known as psi. Note that only peptide backbone bonds are labelled, in most cases the R group bond is free to rotate.