Macro-Complex Modeling of Biomolecules

Angelo Rosace, Iain Maryanow, Zach Collester
Universitat Pompeu Fabra: MsC Bioinformatics

March 31, 2020

1 The application

This a python-based application that utilizes pairwise interaction data as input to generate macro-complex models of biomolecules. It functions for both protein-protein interactions (PPIs) as well as protein-DNA interactions. The output is a PDB file of the generated model.

2 Biological Background

2.1 Protein Structure

Proteins are vital for the structure, function, and regulation of the body's tissues and organs. Further, protein structure plays an incredibly important role in determining its biological function, and this structure is organized in a hierarchical manner. Primary structure consists of amino acids bound together via peptide bonds to form chains known as polypeptides. The secondary structure arises from hydrogen bonds between amino acids in the polypeptide chain causing the chain to form a spiral (alpha helix) or bend and fold (beta pleated sheet). These intermediate folded structures can then fold together to form three-dimensional protein subunits, known as protein tertiary structure. Most functional proteins have quaternary structure, in which several tertiary domains assemble into multi-subunit complexes via non-covalent bonds (hydrogen bonding, Van der Waals forces, etc.).

2.2 Protein/DNA Interactions

3 Building the Complex

3.1 Data Structure

Once the application receives the pairwise interaction data, it stores the unique data for each chain (chain label, residue order, atom order, atomic locations). Each chain is aligned with one another, and chains with greater than 95% sequence similarity are regarded as the same chain. The logic behind this step is that the number of protein folds that exist in nature is finite and there are a few folds that are extremely common. Therefore, chains with a high degree of sequence similarity are likely to share a very similar structure and can reasonably be recognized as the same chain.

3.2 Chain Superimposition

The application builds macro-complexes via a series of protein chain superimpositions. This functions by fixing one chain in space, moving the other chain, obtaining the corresponding rotation matrix, and then utilizing this matrix to join the moved chain to the fixed chain and adding it the model. This structural superimposition method allows the application to construct complexes in a recursive manner, which will be explained later in more detail.

3.3 Clashing

In order to build complexes that are biologically plausible and obey basic physiochemical laws, it was necessary to ensure that chains did not clash before adding a new one to the model. The application considers the alpha carbon of each residue as the core atoms. Clashing is defined as two core atoms being within a certain distance of another. In the case of this application, clash distance is 2 Å and superimposed chains that have any clashing core atoms are not incorporated in the model.

3.4 Stoichiometry

This program allows the user to input a desired stoichiometry for the model. Stoichiometry refers to the composition of the model, specifically the number of each chain included in the final model. For example, if the inputted protein consists of four unique chains (A, B, C, and D), the user could specific a specific stoichiometry of 1A, 2B, 2C, and 1D, and the model that is built will have 1 A chain, 2 B chains, 2 C chains, and 1 D chain. If the stoichiometric constraints applied by the user do not result in a biologically plausible model, the program will inform the user that the stoichiometry is not valid.

3.5 Recursive Function

To build complexes with multiple subunits, the program utilizes a recursive function. After the first iteration of the recursive function, the model consists of two chains. During the next iteration, the addition of subsequent subunits is dependent upon satisfying the previously mentioned clashing and stoichiometric constraints. This function allows the creation of all plausible models. However, the application has been developed such that only the best model is saved into a PDB file. The best model is defined as the model having the lowest RMSD score. In statistical terms, RMSD (Root Mean Square Deviation) is a measure of the distance between predicted values from a model and actual values. In the case of macro-complex modeling, a low RMSD score is favorable.