BACKGROUND AND GENERAL FEATURES

In living organisms, functional proteins rarely are formed by an isolated peptide chain. In order to perform their functions, peptides need to interact and form polypeptide complexes. These interactions are the result of the combination of distinct electrostatic forces such as hydrophobic effect, hydrogen bonds, disulfide bonds… Nowadays, visualizing and modeling protein complexes is possible thanks to the combination of experimental techniques, such as NMR or X-ray crystallization, and bioinformatics tools, which help us to evaluate and refine the complexes.

For this project, through bioinformatic resources, we aim to reconstruct an structural model of a macrocomplex of several units using pairs of domains or chains.

For modeling complexes, a crucial point is knowing how chains are interacting. We define an interaction or contact when a pair of residues from different chains are spatially close. To identify these contacts, we calculate distances between specific atoms (usually carbons alpha or carbons beta) and consider only the pairs whose distance is lower than a cutoff. Depending on the purpose the distance cutoff can change, but, in general, a distance lower than 8Å between Cβs is considered as an interaction. Although, Cβ distances have been proved to reconstruct better the complexes, Cα distances are still widely used, as they are backbone atoms[1].

However, some of these possible interactions can present physical inaccuracy and generate artifacts. One of them are steric clashes. Steric clashes appear due to an unnatural overlap of any two nonbonding atoms in a structure. To remove this kind of artifacts, diverse approaches have been considered. Current state-of-the-art methods also use distances and cutoffs to determine if the possible interaction is a clash. Tools such as MolProbity set this cutoff to 0.4Å, but the distance can also depend on the atoms involved. Other procedures, based on energy parameters of the structures, identify clashes using Van der Waals repulsion energy of the atoms involved[2].

In order to reconstruct a protein complex from several pairs of chains, we also need to identify which chains appear in more than one input pair. To assess it, we decided to perform a pairwise sequence alignment between all the chains from the whole set of input pairs. As we want to work with normalized scores, we divided the score by the length of the longest chain in the alignment[3]. Chains with normalized scores higher or equal to 0.9 will be considered as equal[4].

After the identification of chains and possible contacts, the following step is to locate the components of our complexes into 3D space. One of the most used strategies for this purpose is superimposition, which consists of setting the atoms of a new chain on top of the set of atoms of another chain. In our program, we used the module Superimposer from Bio.PDB[5]. First of all, in this module, the input chains can play different roles: fixed,

moving and alternative. Fixed and moving chains need to have the same atom length and at least 90% of sequence similarity. If their atom lengths are not equal, we will repeat the alignment of both chains and consider only the shared residues in both chains. Moving chain atoms will be rotated and translated on top of the ones of the fixed chain. Rotation and translation movements will be stored in the *rotran* matrix of the Superimposer object. The *rotran* matrix will be applied to the alternative one, resulting in a change of its coordinates, now in function of the fixed chain. However, the addition of the alternative chain can lead to the appearance of steric clashes. For this reason, when a new chain is added to the model, we evaluate the presence of steric clashes. As mentioned above, there isn't an universal cutoff to determine if a contact is a clash or not. In our program, users can choose the desired cutoff, depending on their needs. If a clash is found when the alternative chain is added, this chain will be automatically removed. This process will be repeated as many times as necessary until all the different input chains are placed in the 3D space.

Once the model is generated, we need to evaluate the possible errors of our protein structure. Energy profiles give us a quick overview of the quality of our model. Our program generates a normalized DOPE profile using MODELLER[6]. DOPE, or Discrete Optimized Protein Energy, is a statistical potential used to assess models in protein structure prediction. Areas or residues with a DOPE score higher than zero denote poor quality, so will be the main target of a posterior refinement.
If desired, it is possible to obtain more than one model in our program. To choose conscientiously the best model, we also determined which of the generated models presents the lowest energy. To get this estimation we used the energy function included in MODELLER. This function compares spatial features of selected atoms (in our case, all the atoms of the model) with selected restraints in order to determine the violations. Moreover, a basic MODELLER optimization with restraints is also included in our program. In this step exists also the possibility to obtain a DOPE profile from the non-optimized and optimized models, which can result useful to check if our model has improved after optimizing it. However, restrains applied can vary depending on the case, so we will need to revise and add the most suitable restraints for each case.

REFERENCES

[1] Ramachandran S, Kota P, Ding F, Dokholyan NV. Automated minimization of steric clashes in protein structures. *Proteins*. 2011;79(1):261-270. doi:10.1002/prot.22879

[2] Adhikari B, Cheng. Protein residue contacts and prediction methods. *Methods Mol Biol.* 2016 ; 1415: 463–476. doi:10.1007/978-1-4939-3572-7_24

[3] Peris G, Marzal A. Normalized global alignment for protein sequences. *J Theor Biol.* 2011; 291:22-8. doi:10.1016

[4] Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*. 2008; 71(2): 891-902.

[5] Hamelryck T, Manderick B. PDB parser and structure class implemented in Python. *Bioinformatics*. 2003;19: 2308-2310

[6] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 1993; 234: 779-815