Protein assembler: Theoretical background

Introduction to molecular modeling

Molecular modeling is a computational and biological science. Its aim is predict the conformation of biological molecules using the structural knowledge of previously described molecules and adding chemical, physical and biological information when it is required.

The problem confronted here has been model-building of protein macrocomplexes using pairs of structures or domains. In both cases, we are working with a conserved part of a bigger, three-dimensional structure. For this reason, the model-building process is based on the structural information given by pairs of interactions included in the PDB files. The main procedures implied in this reconstructions structural superimposition and clash analysis.

As commented before, the information used in the model-building is in PDB format, which is the structural format provided by Protein Data Bank. This format is a standard representation for macromolecular structure data derived from X-ray and NMR studies. The main information content of this format are the atomic coordinates description of the structural model. Moreover, it also can include the names of molecules, primary and secondary structure information, sequence database references, ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations. In Biopython we can find a useful PDB parser that allows to access to all this information in an automatized way.

1. Superimposition

Across the evolutionary process, what is more relevant is the functionality, for this reason the protein structure remains longer than the sequence. This is the main interest of structural similarity analysis, but not the only one. For this, the superimposition is very useful in homology modeling, where the structure of a known structure is used to define the structure of a related molecule with an unknown structure. Even that, we haven't use it with this purpose, we have perform structural superimposition to assembly the different subunits of the complex using the interactions information. Once the same sequences are detected, the other subunits are dragged avoiding clashes.

Structural superimposition is the process by which two molecular structures are placed in space minimizing the distance between their backbone atoms. One of the structures is the reference, while the other one is rotated and translated in the space to fit the coordinates of the reference one. Once the two molecular structures are superimposed, we can quantify how different are these two structures though the Root Mean Square Deviation (RMSD). Furthermore, Singular Value Decomposition (SVD) is the algebraic approach used to find the rotation and translation that minimize the distance between the C-alpha coordinates of the pairs of aminoacids.

Some atoms of one structure can be topologically equivalent to the other one, for this reason it is important to use the C-alpha atom to calculate the RMSD. In this case, the atom pair correspondance can be found thanks to the sequence alignment, without complex considerations of internal symmetry.[1]

2. Clash analysis

Check that the alignment is correct is important, but it is also really important to check that there aren't clashes between structures, because different structures have to be in different locations. This is the use that we have give to Neighbor Search, which is an algorithm used to find atom pairs within radius, so it is used to find clashes.

We have choose the clash distance after empirical experiments. The tested parameters have gone from 5 Amstrongs to 2, decreasing them since find the distance that fit better. This distance is a distance of 2 Amstrons. Thanks to this modifications the gamma subuinit of the ATP-syntasa, which had been detected as a clash, could be assembled. On the other hand, we have use a 5% of error, due to the stnadard interval of confidence to the calculus.

References

1. Kufareva, I., & Abagyan, R. (2012). Methods of protein structure comparison. *Methods in Molecular Biology (Clifton, N.J.)*, *857*, 231–257. http://doi.org/10.1007/978-1-61779-588-6_10