An overview of protein tertiary structure prediction and structurally informed function prediction

Daniel Barry Roche

David Aubert & Ahmed Rafik
University of Montpellier
Bio-informatics

1 What is a protein structure prediction?

1.1 What is a protein tertiary structure?

The term protein tertiary structure refers to a protein's geometric shape. The tertiary structure will have a single polypeptide chain "backbone" with one or more protein secondary structures, the protein domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular protein determine its tertiary structure. The protein tertiary structure is defined by its atomic coordinates. These coordinates may refer either to a protein domain or to the entire tertiary structure.

The scientific vulgarization is that the tertiary structure is the spacial 3D structur. Her shape can change depending on his structure the pH and the temperature. A consequent number or properties can be found thank to the tertiary structure.
Thus, we can easily conclude why it is a major importance to modelize such data.

1.2 Protein tertiary structure determination

A protein tertiary structure determination allows us to know more on the protein we are currently observing such as:

- Connection between sequence and structures
- Easier than microscopic observation
- Evolution of proteins

1.3 Different protein tertiary structure determination methods

1.3.1 Template-based modelling

1.3.1.a Template-based modelling fold recognition

Similar protein's sequences have the same folds. Thus, we can classify proteins according to their shapes. One of the advantages of such classification is that the number of unique structural folds is very low compares to the number of proteins.
Today, the classification of folds is very advanced and only a few folds are discovered.

1.3.1.b Template-based modelling homology

1.3.2 Template-free modelling

Template-free modeling is the prediction of the proteins structure. Compared to temple-based modelling, no proteins are used as template.
This technique has a lot of advantages compares to template-based ones.

2 Critical Assessments of Techniques for Proteines Structural Prediction

2.1 What and Why?

The aim is to improve the advance on the protein structure identification. The CASP (Critical Assessment of protein Structure Prediction) aim to etablish the actual state of the protein structure prediction methods, to identify progress which were made, and highlight the best directions to take.

it's therefore an official and international competition, which test proteins structure prediction methods. The latter is divided on any categories

2.2 CASP Categories

1. Tertiary structure prediction
2. Template-based modelling
3. Template-free modelling
4. Oligomer prediction
5. Disorder prediction
6. Contact prediction
7. Model quality assessment
8. Function prediction

2.3 CASP's History

The championship was created in 1994, and it occurs all 2 years.
Also, in 2014, we assisted to the eleventh edition.

This permit, now, to automate the models creation, and keep the same quality than the models of humans expert modellers.
Morever, it given the community a benchmark to test the usefulness of their algorithm.

3 Model quality assessment

3.1 How to define the model quality

The model quality idea isn't easy to define because it's very subjective notion.
It was define by comparison before the availability of cristal structure.

Thus, with some evaluation criteria, Research could produce some algorithms which can evaluate models

3.2 Evaluation Algorithm

1. ModFOLDclust2 Create by Daniel Roche and Mc Muffin in 2010, this algorithm operate as follows :
   a) It use Clustering-based method
   b) Combines structural alignement of multiple models with a method using Q-score
   c) Producing glabal quality scores and per-residue errors
2. RFMQA This algorithm is much more recent, it was created en 2014 by Manavalan and Al
   a) It use a single model-based method
   b) Random forest based model quality assessment
   c) Ranks protein models using its structural features and knowledge-based potential energy terms
   d) Produces global model quality score

4 Structurally informed function prediction

4.1 differents methods

There are two types of binding site prediction methods :

1. The sequence based method which identify conserved residues that may be structurally or functionally important

2. The structure based method which is energetic, geometric method and use miscellaneous methods.

4.2 Function prediction in CASP

4.2.1 What is it ?

That consist on make the prediction of ligand binding residues within a protein of unknown structure.

4.2.2 ligand binding site residue prediction methods

1. Predict the location of the protein binding site

2. Predict the ligand and location of the ligand within the binding site

3. Predict the residues that bind to the ligand within the binding site

4.2.3 what's the utility ?

These tools are needed on many fields like annotation of genome, *de novo* drug design, or mutagenesis studies.
It's also used in the elucidation of protein function and to predict ligand binding specificity.

5 Protein ligand interaction prediction methods

5.1 FUNFold

6 Limitations and perspectives