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

Protein structure can be determinate experimentaly, using NMR or X-ray cristallography, and theoricaly, with template based method or template free method.

1.2.1. Why use this methods ?

The experimental method are very exhautive and hard to devellop. So the structure model determination are very important to predict tertiary structure.

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

Homology modelling is different and more complex than fold recognition. Instead of using the structural form of the protein the study of the amino acid sequence is used. Therefor, the complexity is bigger, the manippulation harder but the accuracy is more advance.

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. The protein will not be searched for his structural form or functions, but for some parts of his sequence. All proteins have commun function, like energy functions or signals function for example. This technic is good because less database dependent than the templase-based one. Most functions are recorded, while functions can suffer mutation, transformation the function will not change or will be known.

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

The CASP rewards many type of methods, which are listing. This represents the different fields concerned by the protein structure prediction. On all of this categories, there are many team of researchers who present them method to predict a protein structure.

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) Clustering-based method

b) Combines structural alignement of multiple models with a method using Q-score

c) Producing global quality scores and per-residue errors

So, this algorithm apply on a group of methods and compare their to know which is the best.

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

Also, this algorithm can use only one method to test it quality. it base on un random group of method referency and compare them to the first one to know if it's a good method.

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

The key features to understand FunFOLD is that despite the evolution (mutation, subsitution etc...) the structures, and by extention the folds, are more well-conserved than the sequence. Plus the ligand which allowed operation such as blocking a site, activate or deactivate protein's functions etc... are also well-conserved, this imply that the expression of the proteins depends on his structure and his ligands. Thus if a protein have the same ligand and structure we can imply that they have the same function.

5.1.1. How does it work?

The ligand could be considered as a tiny point of glue, it is first attached on a part of the protein, then it will carry the given information.
The prediction system work like this:
A ligand glue himself on a specific part of the protein and will leave there some marks. These residues can be visible but with some imprecision. The deposit can be realesed at a variable distance. This lack of accuracy can be compensate by the structure of the protein. Indeed, there are some structure that can promote a better accuracy. So the choice of the zone to study is very improtant.
The FUNFold technic consist on the prediction capacity to know where the ligand are going to attached themselves and then to predict the function depending on the ligand expression.

5.1.2. Benchmarking

By comparing the predictions values returned by the FUNFold one compare to others technics used on the CASP, we can clearly see that the mean value is higher for the FUNFold. Thus FUNFold is more efficient.

5.1.3. FUNFOLDQA: quality assement tool

This quality assement for FUNFold is based on protein-ligand site residue. For starter, it is needed to have a 3D structures to analyse. Then the protein structure is analysed. Many methods are used:

1. BDTalign: it's basically form recognition, the closest equivalent residues is choosen from a database.

2. Identify score: It's the same idea as BDTalign but instead of using the 3D form the amino acid sequence is searched and then the closest equivalent is choosen according to the amino acid sequence.

3. Rescaled BLOSUM62, Equivalent residue ligand distance score etc...

Basically most methods presented could be separated in two groups: using the structure of the protein and/or ligand or the composition of the protein/ligand.

6. Limitations et perspectives

6.1. Limitation

The first limitation is also 3D dependency. Without the protein's modelization it is impossible to go further. This dependancy is related to the database, if the query returns a null value, no prediction will be used.
Each sequency can only have one prediction, in that case a multi-expression sequency will be quickly eliminated and there will be a lack of accuracy in the prediction.

6.2. Perspectives

The first upgrade would be a higher ligand residue detection. This ugrade will allow to directly increase the protein's prediction. Indirectly, this accuracy rising will be related to the software constraint, such as the 3D recongnition.
Another upgrade could be a better 3D modelization that will allow a better flexibility in form recognition. This will allow to have better result between the model and the targeted protein.
This implies a consequent architecture. Indeed, this accuracy increasement will be paired with data volume. The higher the precision, the higher the data will be.