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

Daniel Barry Roche

David Aubert
University of Montpellier
Bio-informatics
email@example.com

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 is it?

It stands for a world championship of predictive structure. This "competition" has started in 1994, and it is define, by themselves their objective to be to help advance the methods of identifying protein structure sequence. They provide the architecture to make these research such as servers, samples and consulting.
With time the CASP became bigger and bigger. It can provides very advanced proteins modelisation tehcnics. A very big consortium has been created since then, with american research group such as Structural Genomics Consortium (SGC), New York Structural Genomics Research Center (NYSGRC).

3. Model quality assessment

3.1. Model quality asessment algorithms

1. ModFOLDclust2
   a) Clustering-based method
   b) Combines structural alignement of multiple models
   c) Producing glabal quality scores and per-residue errors
2. RFMQA
   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

Basically,

3.2. Structurally informed function prediction

There are many predictions ways existing for proteins structures. Most of them are structures observation methods.

1. Geometric methods
2. Energetic methods
3. Homology modelling
4. Surface accessibility
5. Physiochemical properties

Also, others methods do exist. A sequence based method exist. It has significant impact on understanding protein function, elucidating signal transduction networks. This method accentuate the study of the amino acid sequence and his prediction.
This method is particuliary appreciated because the number of sequences to study is cosntantly growing and the sequence study take a consequent amount of time, the prediction will save time and will be able to reveal the protein's sequence.

4. Protein ligand interaction prediction methods

4.1. 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.

4.1.1. 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. Le système de prédiction marche comme ceci:

Un ligand se colle sur une zone de la protéine et y laisse des marques.Ces résidus peuvent être observables mais il y'a une imprécision, le dépot de résidu ce fait à une distance variante. Cette imprécision peut être atténué par la structure de la protéine. En effet, il existe des structures qui favorisent plus l'attachement de ces ligands que d'autre.
La technique FUNFold consiste donc à pouvoir prédire la zone où des ligands vont s'attacher et deviner ainsi la fonction de cette zone en fonction de l'expression du ligand.

4.1.2. Benchmarking

En comparant les valeurs de prédiction retournées par le FUNFold comparées aux valeurs obtenu par le CASP, on voit nettement que la valeur moyenne est bien supérieur pour le FUNFold. Le FUNFold est donc plus éfficace.

4.1.3. 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.

5. Limitations et perspectives

5.1. Limitation

La première limitation est la dépendance au modèle 3D. Sans la modélisation de la protéïne il est impossible d'avancer. Cette dépendance est lié à la base de données utilisée, si la requête à la base de données est nulle, aucune prédiction sera faite.
Chaque séquence ne peut avoir qu'une seule prédiction, dans le cas d'une séquence avec plusieurs expressions nous serons vite limités.

5.2. Perspectives

La première amélioration importante serait l'augmentation de la précision pour la détection des résidus. Cette amélioration permettra directement d'augementer la précision dans la prédiction des protéines. Indirectement cette augmentation de précision passe

par des contraintes logiciel tel que la modélisation 3D de la protéine. Pouvoir atténuer les distances entre les modèles 3D modèles et les protéines ciblées.
Ceci implique également une base de données consequente. En effet cet augmentation de précision va de pair avec volumes de données, plus la précision augmente plus les données pour les representer augmentent.