A COMPARITIVE STUDY ON DIFFERENT METHODS USED FOR BUILDING 3-DIMENSIONAL MODELS OF PROTEIN

Ananya Anurag Anand [1]

[1]*Student, Department of Biotechnology, M.S. Ramaiah University of Applied Sciences, Bangalore, India.*

Corresponding Author: ananyaanurag12@gmail.com

Abstract: - The concept of protein modelling or building of three-dimensional models of proteins using various methods is increasingly gaining sight of the researchers because of the various benefits derived from it that include not only identifying the type but also the function of the protein based on the model predicted. There are two types of methods, template-based methods and non-template-based methods used for modelling the protein structure using various logics. These include homology modelling, threading and ab-initio methods. The ab-initio method is template-independent unlike the other two methods which are template-based. Here, all the three methods have been reviewed in detail and to the best of the knowledge so as to provide better insight into the type of method we must use as per our requirement.

Key Words— ***Homology modelling, Threading, Ab-initio modelling***

I. INTRODUCTION

Since the start of integration of informatics in biology, many researchers have tried to apply different approaches to the determination of protein structure. However, there are three main methods that are currently employed to model out a protein's structure. These include the template-based modelling methods and the non-template based or energy-based modelling method. The template-based modelling method is of two types, one is the template-based homology modelling whereas the other is the template-based threading method. These methods use a template to which structure of the query protein is compared and analysed according to the template. If the two sequences, that is the query and the template sequence have a high enough similarity, they are supposed to be possessing similar structures, as is seen in homology modelling. Threading method also makes use of template protein but it compares the secondary structure of the two to provide us with an output. However, the energy-based modelling does not require a template for the prediction of protein structure. It simply tries to figure out the best suitable model for the query protein based on the energy state of it. The lesser the energy, the more stable is the configuration. Nowadays, there has been increase in attempts to integrate the template-based methods and the energy-based methods.

The template-based methods, nowadays, make use of energy-based model refinement methods. Also, the energy-based methods now incorporate a few protein sampling and machine-learning methods to extract and utilize the information from the protein databases.

II. TEMPLATE-BASED HOMOLOGY MODELLING

Template-based homology modelling, also referred to as comparative modelling, is a protein structure modelling technique based on the principle that the more is the sequence similarity between two proteins, that is the query and the template protein, the more is the likeliness of them having similar structures. This method involves six main steps:

1. Selection of the template

The template selection is done using the PDB (Protein Data Bank), where homologous proteins with determined structures are searched for. This can be done using heuristic programming methods of pairwise alignment, like FASTA or BLAST. SSEARCH or ScanPS can also be used.

There are different zones under which the homology models can be characterised:

- Midnight zone- It denotes less than 20% sequence identity, which means that this particular structure is not a reliable template.
- Twilight zone- It denotes sequence identity of 20%-40%, which means that the sequence may imply some structural identity.
- Safe zone- It denotes more than 40% sequence identity, which means that there is great possibility of structural identity an that this can be used as a reliable template. The sequence with highest homology is used as a template.

2. Sequence alignment

Multiple alignment algorithms, like Praline or T-Coffee are used to align the full length of sequences of the query and the template proteins.

3. Backbone model building

Once the alignment is done, the coordinates of corresponding residues of the identified template protein are copied onto the target or the query protein. If two residues that are aligned are identical, the coordinates of the backbone atoms as well as the side chain atoms are copied but if they differ, then only the coordinates of the backbone atoms are copied.

4. Loop modelling

When an insertion or deletion occurs, it produces a gap in the alignment, that creates holes. These holes or gaps are closed by loop modelling. Loop modelling can be done by two

methods: the database method and the ab initio method. The database searching method finds the spare parts from the known protein structures present in a database to fit it into the two stem regions of the query protein. However, the ab initio method generates random loops and tries to fit the best loop that has low energy and that does not clash with nearby side chains. FREAD (based on database search method), PETRA (based on ab initio method) and CODA (based on a consensus approach utilizing results from both FREAD and PETRA) are important web servers used for loop modelling.

5. Side-chain refinement

The side chain prediction is done using the approach involving the concept of rotamers, which involve the most favourable torsion angles that are extracted from the known crystal structure of proteins. A rotamer library is referred to, where the rotamers are listed and ranked according to their frequency of occurrence. SCWRL is a specialised side chain modelling program.

6. Model refinement

The potential energy calculations are further used to refine the model to its best stable configuration by energy minimization. GROMOS is a program used for performing such energy minimizations.

7. Model evaluation

The final homology model should be evaluated to see if it follows the physiochemical rules, such as checking anomalies in chirality, close contacts and bond lengths. If structural irregularities are found, then it needs refinement. WHAT IF is a good quality protein analysis server used for this purpose.

III. THREADING METHOD

The threading method is based on the prediction of the structural fold of a protein by fitting that protein into a structural database and thereby, selecting the best-fitting fold. The emphasis is on the comparisons of secondary structures as they are the most evolutionarily conserved ones. Here, the algorithms are of two types:

1. Pairwise Energy Method:

Energy based criteria is taken into account while matching the two structures. It involves the alignment of the query sequence with each structural fold that is present in the fold library. This alignment is carried out at the level of sequence profile, using heuristic approaches or dynamic programming. The local alignment is adjusted to obtain low energy for a better fitting. Next, a crude model is built for the target sequence by replacement of aligned residues in the template with corresponding residues in the query. Then, the energy terms of the model are calculated, that include pairwise residue interaction energy, hydrophobic energy, and solvation energy. Finally, the models are ranked on the basis of energy to find out the lowest energy fold that is corresponding to the structurally most stable fold.

2. Profile method:

In this method, a profile is built for a group of related protein structures, by superimposing the structures and extracting statistical information from them. The profile consists of scores that detail out the propensity of each of the twenty amino acids to be present at each profile position. The prediction of the structural fold of an unknown protein sequence requires the prediction of query sequence for its secondary structure, polarity and solvent accessibility. This prediction is compared with the propensity profiles of the known structural folds to find out the fold that represents the predicted profile in the best manner.

Thus, threading and fold recognition helps in assessing the compatibility of an amino acid sequence with a known protein structure present in the fold library. In case, the protein fold that is to be predicted is absent in the library, the method fails. 3D-PSSM< Fugue and GenThreader are important web-based programmes that are used for threading.

IV. RESULTS AND DISCUSSION

The ab initio prediction method can be used to predict the protein structure from the sequence information itself. It aims to produce all-atom protein models based on only the sequence information. Protein folding is modelled on the basis of global free-energy minimization. Since, the protein folding problem has not been solved till now, the ab initio methods are still quite
unreliable.

One of the ab initio methods known as Rosetta has been found to predict 61% of structures 6.0 Å RMSD.

The working of Rosetta is as follows:

1. The fragment libraries for all the segments that are 3 to 8 residues long are extracted from the protein structure database. This is done using the sequence profile-profile comparison method.
2. Then, the tertiary structures are built using MC search of possible combinations of local structures.

V. CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION (CASP)

Be it any discussion related to the prediction of protein structure, the role of CASP can never be ignored. CASP has been conducting community-wide experiments to analyse various protein structure predictions and critically assess them. CASP gives an opportunity to participants worldwide to model out the protein using their own methods and them compare it against the methods used by others. Also, one of its most crucial goal is to promote the template-free models building. It not only assesses the 3D structure but also evaluates things like residue-residue contacts, model structure refinement, and much more. All its results are publicly available. The success of CASP almost fully depend on the researches performed by the experimental community. Also, the experimental community may utilise the CASP for its own benefits by using it to compare what they are doing on a protein with what the other researchers are doing. Therefore, it is necessary for the bioinformaticians to get familiar with the use of CASP and contribute more to its success to gain more success in return.

VI. CHALLENGES

However, one of the biggest challenges lie in the accurate prediction of the 3D structure based on just the primary structure of the protein. There lies a problem in protein-folding. This can be stated under Levinthal's paradox. The paradox states that, "Finding the native folded state of a protein by a random search among all possible configurations can take an enormously long time. Yet proteins can fold in seconds or less."

DeepMind's AI Program is a small to the Protein Folding Problem that works as follows: The system is first trained on public dataset of the known experimental protein structures, so the experimental approaches will still be needed to withstand biases in training data for algorithm. Also, although most of these predictions are highly accurate, the solution is still not perfect. The AI-based algorithm
encounters difficulties modelling some proteins and also in interaction with other proteins.

Also, there are several challenges in the protein-protein interaction predictions. The experiments used to study these interactions are very time-taking. However, comparative modelling methods have been used for this but they are effective in a relatively few number of cases. One of the alternatives to comparative modelling can be protein–protein docking. The docking procedures make use of surface
complementarity and electrostatics in order to predict the structural complexes, fitting together two known structures or reliable 3D models via their interacting surfaces. But these methods are hampered by a lack of a wholesome understanding of the forces that are involved and also the
conformational changes that often take place upon protein–protein binding.

If we talk about the problems associated with the homology modelling, also known as the comparative modelling, the most challenging regions to model using comparative modelling, are insertions because of the absence of any equivalent region in the template. The problem increases with the length of the segment.

Thus, it is for sure that we need an overall improvement in the understanding of the protein structures and protein folding. Also, by studying protein-protein interactions more closely we can deduce more about the nature of proteins and their behaviour. Not only this, we need to better observe and analyse the behaviour of different proteins in different environments.

VII. CONCLUSION

Tools for prediction of protein structure have advanced considerably in the past decade, but there still remain many challenges. The energy functions that guide the prediction and design still struggle to accurately balance various interactions like the polar and nonpolar interactions and solvation effects. Thus, there is a limited success rate for interface-modelling applications, such as protein docking with backbone flexibility. There are many challenges of interface energetics that include the need to accurately model out conformational preferences of the irregular polypeptide segments. Thus, new approaches are needed to accurately predict and design the protein models. Techniques employed that combine molecular-dynamics trajectories with the analysis of the energy landscapes are required to cover the dynamic aspects of these systems. Protein structure prediction and design are playing important roles in biology, especially medicine. As the protein structure databases are continuing to grow, the availability of new sets of protein
backbones and side-chain packing arrangements is increasing, thereby, opening up new possibilities each time to give them a new direction and use them to identify suitable binding sites and functions. Solutions that are coming up from AI and ML fields are also interesting and it would be interesting to see more new ideas coming up from their side.

REFERENCES

[1] Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170 (1991).
[2] Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755–763 (1998).
[3] Lazaridis, T. & Karplus, M. Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol. 10, 139–145 (2000).
[4] Sadreyev, R. & Grishin, N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326, 317–336 (2003).
[5] Song, Y. et al. High-resolution comparative modeling with RosettaCM. Structure 21, 1735– 1742 (2013).
[6] Bujnicki, J. M. Protein-structure prediction by recombination of fragments. Chembiochem. 7, 19–27 (2006).
[7] Kuhlman, B., Bradley, P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 20, 681–697 (2019).
[8] Edwards, Yvonne & Cottage, Amanda. Bioinformatics Methods to Predict Protein Structure and Function: A Practical Approach. Molecular biotechnology. 23. 139-66. 10.1385/MB:23:2:139 (2003)
[9] Deng, Haiyou et al. "Protein structure prediction." International journal of modern physics. B vol. 32,18 (2018): 1840009.
[10] Fiser, Andras. "Template-based protein structure modeling." Methods in molecular biology (Clifton, N.J.) vol. 673 (2010): 73-94.
[11] Blacklock, K. M., Yachnin, B. J., Woolley, G. A. & Khare, S. D. Computational design of a photocontrolled cytosine deaminase. J. Am. Chem. Soc. 140, 14–17 (2017).
[12] Polizzi, N. F. et al. De novo design of a hyperstable non-natural protein-ligand complex with sub-Å accuracy. Nat. Chem. 9, 1157–1164 (2017).