Proceedings of the International Conference on Modern Medicine and Global Health (ICMMGH 2023)
DOI: 10.54254/2753-8818/6/20230263

Protein prediction algorithms: Homology modeling, AlphaFold, and Foldit

Songhan Duan[1, †], Runcheng Ke[2, †], Junyi Xiang[3, 4, †]

[1]Jilin University, No.2699 QianjinStreet, Chaoyang District, Changchun, JL, CHN;
[2]No.8 Middle School of Beijing, No.2 XueYuanXiao Road, Xicheng District, Beijing, BJ, CHN, 100000;
[3]Shanghai World Foreign Language Academy, 400 Baihua Avenue, Xuhui District, Shanghai, SH, CHN, 200030
†These authors contributed equally.

[4]1811000419@MAIL.SIT.EDU.CN

Abstract. Protein, one of the most basic structures of biological molecules, have its own four level structure that corresponds with its function. The structures make every protein unique and diverse. Studies of protein must be based on the understanding on protein's structure. Thus, methods must be applied to predict the protein structure. Old methods include homology modeling that are both expensive and time consuming. With the development of modern technology, new methods such as Foldit and AlphaFold was invented. The report would introduce these methods and comparisons would be made between these methods.The introduction aims to improve the understanding about protein prediction for relative researchers.

Keywords: protein structures, algorithms, homology modeling, alphaFold, FOLDIT.

1. Introduction

Homology modeling is the only method to predict protein molecular structure in practice [1]. Studies suggest that proteins with similar sequence may have originated from a common ancestor with similar structure and functions. General homology modeling uses homologous proteins based on one or more known structures as templates. Predict the 3 D structure of the unknown structure protein (target protein). This approach relies on a PDB (Protein Data Bank) that now contains more than 80,000 structures [2]. But out of the more than 80,000 structures. Many are multiple variants of a protein, with only about 4,000 structural families of these structures. Protein sequences with more than 30% general homology can establish quite accurate structure. The higher the homogeneity of the sequence, the higher the accuracy of establishing the model structure. Therefore, homology modeling can infer sequence-similar target structures based on the amino acid sequence of known structures.

The following six steps are required to infer the protein structure by homology modeling: 1. Search for template protein; 2. sequence comparison between target protein and template protein; 3. backbone structure modeling; 4. ring region modeling; 5. side chain modeling and optimization; 6. Overall

DOI: 10.54254/2753-8818/6/20230263

structure optimization and evaluation. These steps can be done with different programs, or by using the homology modeling server or the automatic modeling program module. Currently, there are many servers available on the Internet that can predict protein structures using homology modeling techniques, such as SWISS-MODEL (Geneva), 3D-JIGSAW (London), FAMS (Tokoy) and SDSCI (San Diego), among which SWISS-MODEL network modeling server is the most widely used [3].

2. AlphaFold

Alphafold is a newly invented programs used for protein structure predicting developed by Deepmind [4]. It is a deep learning system, capable of learning from a training set and generate new results from a new input. Deep learning system uses multiple layers to process the original input, and eventually generate the final output. AlphaFold uses a large amount of training set of 29000 proteins, which ensures it to generate reliable result.

2.1. AlphaFold version 1

AlphaFold 1 tries to determine the correlation between residues in a protein. The residues might not be attached on the main sequence. The changes and correlation suggest that there are interaction and contant between the two residues in the protein's 3d structure. Thus, a contact map could be drawn to show the interaction between residues. The algorithm was later developed to determine the possible distance between the two residues.

2.2. AlphaFold version 2

AlphaFold 2 was developed in 2020 [5]. The new version fixed the problem of overfitting in the old algorithm. The old algorithm has modules that was trained independently. Thus, the algorithm would generate more secondary structure than the structure in the reality. The version 2 of the AlphaFold involves the training of all modules together, and eliminating the overfitting of AlphaFold 1.

3. Foldit

Foldit is a multiplayer online game for protein structure prediction, which is invented by a program of Baker's team, UWashington [6]. Hard protein prediction problems can be solved by everyone whether they are scientists or not.

Foldit players interact with protein structures using direct manipulation tools and user-friendly versions of algorithms from the Rosetta structure prediction methodology, while they compete and collaborate to optimize the computed energy.

It shows that top-ranked Foldit players excel at solving challenging structure refinement problems in which substantial backbone rearrangements are necessary to achieve the burial of hydrophobic residues.

4. Comparison

4.1. Prediction speed

The speed of protein prediction is an important factor when evaluating different methods of protein prediction. A faster speed of prediction could enable researchers to quickly apply the results of prediction into research, making the process of research more efficient.

4.1.1. Homologous modelling. The speed of homology modeling prediction depends on various factors, such as the size and complexity of the target protein, the quality of the template structure, and the computational resources available to run the modeling software. In general, homology modeling can be a relatively fast process that takes several hours to several days to generate a predicted structure. However, the accuracy of the predicted structure can also vary depending on the quality of the template structure and the alignment between the target sequence and the template sequence.

DOI: 10.54254/2753-8818/6/20230263

Therefore, it is important to carefully evaluate the quality of the predicted structure before undergoing further analysis or experimental validation.

4.1.2. AlphaFold 2. AlphaFold 2 predicts at a much higher speed compared to any other algorithms like Rosetta Fold. With only a conventional GPU, a quick prediction could be made. The algorithm is capable of predicting the results of a chain with 256 residues within only 0.6 minute, 384 residues in 1.1 minute, and 2.1 hours at 2500 residues [7].

4.1.3. Foldit. Because of its principle, Foldit isn't as fast as orther predicting methods. The structures are predicted by players through puzzles and games, so it is amateur. Besides, not all the structures, which were made by players, are definitely correct. Due to all these problems, the speed of prediction is not fixed, and can't be computed accurately.

In conclusion, AlphaFold 2 is better when the factor of speed is compared among the three methods of protein prediction.

4.2. Cost

The cost of prediction is also an important factor to be considered when evaluating different methods of protein structure prediction. Different methods of prediction would require different resources which include apparatus, equipment, and even active players. A lower cost of prediction could foster the application of the methods in research as well as other areas.

4.2.1. Homology modelling. Homology modeling often requires substantial computational power, and it can be expensive to maintain. However, there are many publicly available protein structure databases that serve as templates for homology modeling, which can reduce the cost of the process.

4.2.2. AlphaFold 2. AlphaFold 2 could be operated with only CPU [8]. Though the speed of prediction is rather slow using CPU for prediction, it shows how low the entry cost of predicting a protein structure using AlphaFold. It could be used productively when a consumer level GPU is used. Supercomputer owned by universities could also speed up the prediction. Also, there are no extra cost of predicting more protein structure. The cost of AlphaFold methods of prediction only include the cost of the computing device and the cost of maintenance. With all the factors mentioned above, a database covering a wide range of protein structure could thus be created with a supercomputer used. The data produced by AlphaFold in the database could thus be used by researchers around the world with no cost.

4.2.3. Foldit. Being as a multiplayers online game [9], Foldit isn't the costliest way of prediction. Though the game system needs to be operated and fixed, it costs less than the traditional method. Moreover, this way of prediction doesn't have to use any apparatus except computers. Compared with the value that created by players of Foldit, the cost seems to be insignificant.

4.3. Reliability

The reliability of protein structure prediction methods should also be considered by researchers. As the prediction of protein structure would be used in research, the accuracy should be the most important factor when evaluating different methods.

4.3.1. Homology modelling. The results of homology modeling are very reliable and, in general, it is a highly reliable method to predict protein structure when there is a high-quality template structure and the target and template sequences are highly similar. However, as the similarity of the target sequence to the template sequence decreases, the reliability of the predicted structure also decreases.

4.3.2. AlphaFold 2. The ideal reliability of data about protein structure are 100%. However, without the analysis of protein structure in real life, a protein predicting methods could never reach exactly

DOI: 10.54254/2753-8818/6/20230263

100%. As an algorithm that only relies on the input of residue chain, AlphaFold 2 have considerably high reliability in respect to the actual structure of the protein [10]. According to an article published on nature, the AlphaFold algorithms could generate results that have a median backbone accuracy of 0.96 Å r.m.s.d.95 [11]. The analysis on the known protein structure have shown that the data generated by AlphaFold is trustworthy and could be used for research purposes. However, the actual structure of the protein could never be certain without the analysis on the structure of real protein structure. There are possibilities that the AlphaFold generates data that are inconsistent with the real structure [12]. The possibilities could never be eliminated through the change in algorithms since the nature of machine learning. The research team of the machine learning algorithms adjust the algorithm by evaluating the output of algorithms under a specific input. They have little knowledge about how the inputs are processed to form the output in the algorithms. Limitations should be acknowledged by all researchers who might use the data from AlphaFold database.

4.3.3. Foldit. The reliability of Foldit depends on diversified factors. The most dependent element is the precision of the atom or molecule models, which is based on the underlying scientific principle, used for predicting protein structures. What's more, the quality of the input data provided by the players is also important. And the capabilities of the players to precisely interpret and manipulate the protein structure also carry weight. Although Foldit has been shown that it does have the ability to generate high-quality protein structures in some cases, the reliability of the predicted structures can also be affected by the technical limitations of the physics-based models and the potential errors of the input data or players' interpretation and manipulations. Besides, the function of collaboration in Foldit can also give rise to the variability into the structures' prediction. For example, different players might view the same problem from different perspectives or have different levels of knowledge. As a result, it is important to carefully assess the quality of the prediction produced by Foldit using diversified verify methods, such as Ramachandran Plots, Energy Minimization Method, and the comparison with the corresponding experimental structures if available [13]. To sum up, Foldit has had the great potential so far, to be one of the most useful and successful methods to predict protein structures. But on account of relying on some specific circumstances and the evaluation, it still has a lot of room for improvement.

5. Conclusion

In conclusion, protein prediction algorithms play a crucial role in the study of protein structures and functions. Homology modeling, AlphaFold, and Foldit are three different methods of protein structure prediction, each with its own advantages and limitations. Homology modeling is a reliable method that requires a high-quality template structure and a highly similar target sequence. AlphaFold is a newly invented program that uses deep learning technology and has a high prediction speed and reliability. Foldit is an amateur method that involves players solving puzzles and games to predict protein structures. Overall, the choice of protein prediction method depends on the specific research needs, available resources, and desired level of accuracy. It is important to carefully evaluate the quality of the predicted structure before undergoing further analysis or experimental validation.

References

[1] Scietti, L. and Forneris, F., "Modeling of Protein Complexes," Methods Mol Biol. Papers 2627, 349–371 (2023).

[2] Huynh, C. B., Nagaarudkumaran, N., Kalyaanamoorthy, S. and Ngo, W., "In Silico and In Vitro Approach for Validating the Inhibition of Matrix Metalloproteinase-9 by Quercetin," Eye Contact Lens. Papers (2023).

[3] Gniado, N., Krawczyk-Balska, A., Mehta, P., Miszta, P. and Filipek, S., "Protein Homology Modeling for Effective Drug Design," Methods Mol Biol. Papers 2627, 329–337 (2023).

[4] Plonski, A. P. and Reed, S. M., "Assessing protein homology models with docking reproducibility," J Mol Graph Model Papaers 121, 108430 (2023).

DOI: 10.54254/2753-8818/6/20230263

[5] Flower, D. R., "To Affinity and Beyond: A Personal Reflection on the Design and Discovery of Drugs," Molecules Papers 27(21), 7624 (2022).
[6] Avramouli, A., Krokidis, M. G., Exarchos, T. P. and Vlamos, P., "In Silico Structural Analysis Predicting the Pathogenicity of PLP1 Mutations in Multiple Sclerosis," Brain Sci Papers 13(1), 42 (2022).
[7] Bansia, H. and Ramakumar, S., "Homology Modeling of Antibody Variable Regions: Methods and Applications," Methods Mol Biol. Papers 2627, 301–319 (2023).
[8] Danneskiold-Samsøe, N. B., Kavi, D., Jude, K. M., Nissen, S. B., Wat, L. W., Coassolo, L., Zhao, M., Santana-Oikawa, G. A., Broido, B. B., Garcia, K. C. and Svensson, K. J., "Rapid and accurate deorphanization of ligand-receptor pairs using AlphaFold," bioRxiv Papers 531341 (2023).
[9] Kok, D. L., Dushyanthen, S., Peters, G., Sapkaroski, D., Barrett, M., Sim, J. and Eriksen, J. G., "Screen-based digital learning methods in radiation oncology and medical education," Tech Innov Patient Support Radiat Oncol. Papers 24, 86–93 (2022).
[10] Tsirigotaki, A., Dansercoer, A., Verschueren, K. H. G., Marković, I., Pollmann, C., Hafer, M., Felix, J., Birck, C., Van Putte, W., Catteeuw, D., Tavernier, J., Fernando Bazan, J., Piehler, J., Savvides, S. N. and Verstraete, K., "Mechanism of receptor assembly via the pleiotropic adipokine Leptin," Nat Struct Mol Biol. Papers (2023).
[11] Willems, P., Huang, J., Messens, J. and Van Breusegem, F., "Functionally annotating cysteine disulfides and metal binding sites in the plant kingdom using AlphaFold2 predicted structures," Free Radic Biol Med. Papers 194, 220–229 (2023).
[12] Roney, J. P. and Ovchinnikov, S., "State-of-the-Art Estimation of Protein Model Accuracy using AlphaFold," bioRxiv Papers.484043 (2022).
[13] Miller, J. A., Khatib, F., Hammond, H., Cooper, S. and Horowitz, S., "Introducing Foldit Education Mode," Nat Struct Mol Biol. Papers 27(9), 769–770 (2020).