# The progress in Protein Structure Prediction

Student Name : Bingbing Li
Student ID : 18210705
Email : Bingbing.li@ucdconnect.ie

Computer Science
University College Dublin

## Abstract

With the development of medicine and biotechnology, the demand for protein prediction technology with high accuracy and efficiency is increasing, which makes the correct prediction of protein secondary and tertiary structures become one of the main challenges in bioinformatics. This essay describes the progress in protein structure prediction, giving the comparison of methods used in last two decades with methods exploited nowadays. In the process, it highlights the role of deep learning and machine learning in the development of this technology. In addition, the current most successful protein prediction method will be introduced, as well as its limitations and bottlenecks. From the advancement in protein structure prediction technology, it can be seen that it will play a more important role in the progress of biology, biotechnology, drug design area in the future.

## 1. Introduction

In most biological processes, proteins are at the core. Since Proteins differ mainly from each other in their amino acid sequence, which typically results in different spatial shape and structure, and hence different biological functionality in cells, understanding protein structures has been a huge challenge in biology for decades. However, because some structures are very difficult and time-consuming to solve experimentally and sequence determination in public domain databases far outpaces experimental structure determination, protein structure prediction which is the use of computational techniques to infer the three-dimensional structure of a protein from its amino acid sequence, remains essential.

Since Max Perutz and John Kendrew's pioneering work to determine the structure of globular proteins, this filed started to spring up from the end of 20th century. Another milestone in this field is the thermodynamic hypothesis proposed by

Anfinsen in 1973, in which demonstrates that all the information a protein needs to fold properly is contained in its amino acid sequence. According to physical theory, the amino acid sequence determines the basic molecular composition of a protein, and its natural state corresponds to the most stable conformation with the lowest free energy. However, it is very hard to simulate the folding process by molecular dynamics which is physics-based method because of the difficulties to give such complicated macromolecule an accurate physical description, in addition, the current computing power is not sufficiently capable to search through such a huge amount of conformational space. Whereas according to the experiential rule, that is, proteins with similar amino acid sequences tend to have similar structures, scientists turned their attention to knowledge based methods.

In this viewpoint, this essay will begin by describing the critical process in protein structure prediction and analyzing the basic methods, including essential features, accuracy, and their application to the prediction and understanding of protein function. Then, it demonstrates the progress in protein structure prediction with the injection of machine learning and deep learning. Next, through discussing the current most successful approaches, this paper illustrates the limitations and challenges in it. Finally, it gives the conclusion and predicts the possible role protein structure prediction may play in the growing worldwide effort in biology, biotechnology, drug design in the future.

## 2. Main Approaches in Protein Structure Prediction

There are two main approaches to protein structure prediction. The first class includes comparative modeling which also known as homology modeling, and threading method. Both of them are template-based structure prediction methods, as they depend on measurable similarities which span most of the model sequence and at least one known structure. The second class of methods, de novo or ab initio methods, predict the structure from sequence alone, without relying on any know structures or fold similarity between the modeled sequence. Therefore, “Template-free” is also used when name methods which do not belong to the category of homology modeling and threading.

These are also the main methods that have been used in recent decades. The following paper will analyze these three methods by showing the research results of the paper named “Protein Structure Prediction and Structural Genomics” which published in 2001.

## 2.1 Comparative Modeling

Comparative modeling, refers to the construction of the atomic resolution model of the “target” protein based on the similar amino acid sequence and the experimental three-dimensional structure of its related homologous protein, which is also called “template”. It is focused on finding one or more known protein structures that are likely to resemble the query sequence structure, and on generating an alignment that maps amino acid in the query sequence to amino acid in the template sequence.

Evolution-related proteins have similar sequences, and naturally occurring homologous proteins also have similar protein structures. The three-dimensional protein structure has been shown that it is evolutionarily more conservative than expected based solely on sequence conservation. Moreover, experiment results show that sequences falling below a sequence identity of 20 percent can have very different structures.

The comparative modeling procedure can be divided into four sequential steps: template selection, target-template alignment, model construction, and model evaluation. The first two steps are often performed essentially combined, as the most common methods of identifying templates depend on sequence alignment production.

When the target protein and the template are closely related, comparative modeling can produce high-quality structural models, which inspired the establishment of a structural genomics consortium. It is dedicated to providing representative experimental structures for all types of protein folding, and organizes Critical Assessment of Techniques for Protein Structure Prediction (CASP) which assesses methods of structure prediction, including comparative modeling, in a biennial large-scale experiment.

## 2.2 Threading

So far, using the homology structure in PDB as a template is the most accurate method to predict protein structure. With the rapid growth of PDB database conversion, homology modeling can be used to predict that the proportion of target proteins has been rapidly increasing. However, when structures in PDB do not have sufficiently similar sequence to the target protein, it is still possible to find a protein with structural similarity to the target protein (Bowie et al., 1991). The method of identifying the template structure from PDB is called threading or folding recognition (Jones, 1999). It actually matches the target sequence with homologous and distant homologous structures according to an algorithm, and uses the best match as the structural template.

The basic principle for threading is that the protein structure is highly conservative in evolution, besides, the number of specific structural folds in nature is limited, and it is often the case that two proteins without any sequence similarity fold into similar folds.

The threading procedure can be divided into four sequential steps: Find proteins of known structure (templates) by the energy of a corresponding coarse model, align sequence and template, build a model, and model assessment (Baker and Sali, 2001).

It has already be mentioned that comparative modeling treats the template in an alignment as a sequence, and it uses only that sequence for prediction, while threading treats the template in an alignment as a structure, and prediction is based on both sequence and structure information from the alignment. When no significant homology is detected, comparative modeling may not be able to produce a significant prediction. However, the threading of proteins is possible to make a prediction based on information about the structure (Torda, 1997). Furthermore, it can sometimes reveal more distant relationships than purely sequence-based approaches. That also explains why in many cases protein threading can be more successful than comparative modeling.

## 2.3 Ab initio Method

There is no guarantee that suitable structural models will always be identified for any target protein. When no sequence similarity with proteins of known structure detected and no fold where threading is possible at acceptable energy levels, the template-free methods are the best choice for the hard target proteins which can not identify a satisfactory template on.

Ab initio methods which is also called De novo methods start from the assumption that the native state of a protein for the given amino acid sequence is at the global free energy minimum and conduct a large-scale search of conformational space for the protein tertiary structure which is extremely low in free energy. The two key components of such methods are the procedure for performing the conformational search efficiently and the free energy function used to assess possible conformations.

It is the most straightforward way to randomly produce the initial conformation of the target protein; however, the burden of the conformational search will be very high in this case. Besides the inadequacy of the current force field, the simulation process with such massive conformation reform is extremely hard to accomplish.

The dynamic and multilevel design of protein structure essentially gives us more choices. First we can predict the basic structural features of the target protein, such as backbone dihedral angle, secondary structure, solvent accessibility, and contact maps which do not necessarily rely any structural template. With the development of computationally intensive highly-sophisticated Machine Learning and Deep Learning algorithms of last decade, the efficiency and accuracy of ab initio methods has been incrementally increasing, and this essay will focus on the contributions made by Machine Learning approaches in the next session.

## 2.4 Accuracy and Application of Models

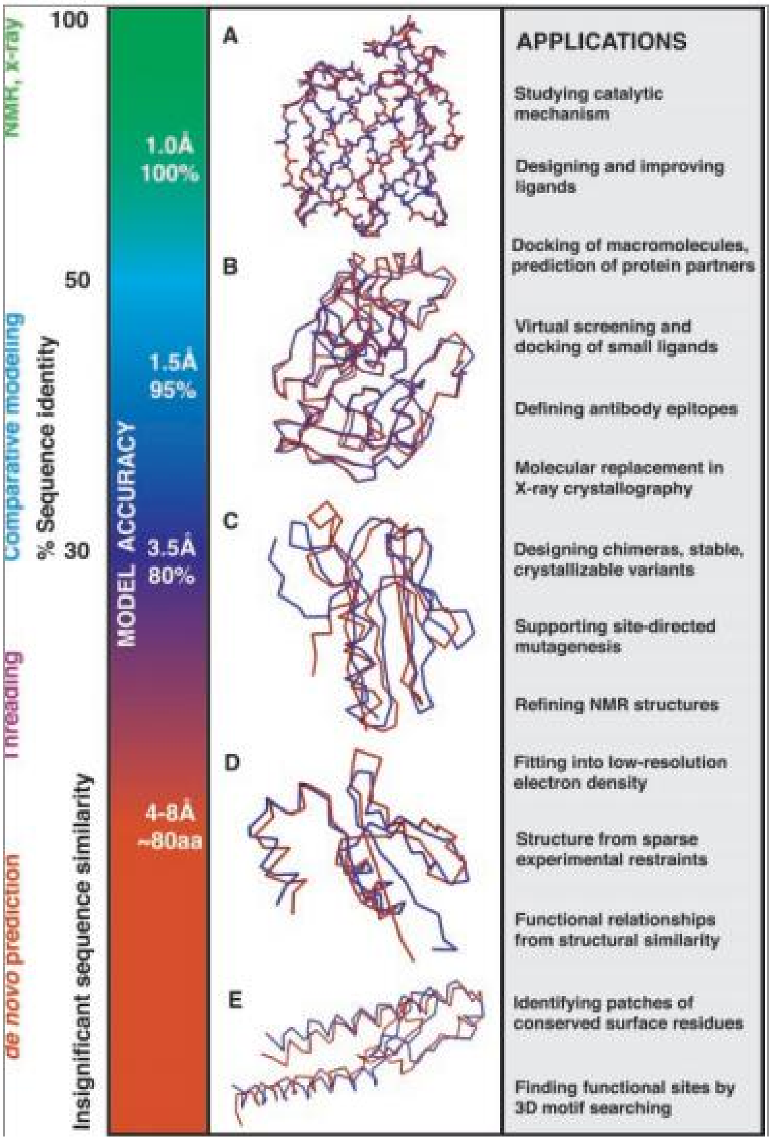
NMR, X-ray
Comparative modeling
Threading
de novo prediction
% Sequence identity
100
50
30
Insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching

Fig 1. Accuracy and application of protein structure models

Fig 1 shows the accuracy of comparative protein structure modeling, threading, and ab initio approaches respectively, as well as the various applicability ranges of these methods. It should be noted that predicted structures are in read, and actual structure are in blue. Sample comparative models based on the identity of their template structure based on about 60 percent (A), 40 percent (B), and 30 percent (C) series. D and E are predictions using Rosetta (de novo method) for the CASP4 prediction experiment.

It can be seen from the results of fig 1. that model accuracy drastically decreases when going from (A) to (E), but the overall predicted structure is still approximately correct. With regard to the comparative model, the accuracy of is correlated with the percentage sequence identity on which it is based, which corresponds with the structural and sequence similarity of two proteins (Martí-Renom et al., 2000). The accuracy of the comparison model decreases significantly as the consistency decreases, and the RMS increases. In addition, when the model is based on an almost insignificant alignment with known structures, it may also have completely wrong folds. Other factors such as template selection and alignment accuracy generally have a greater impact on model accuracy, especially for models based on sequence identity to the template of less than 40%.

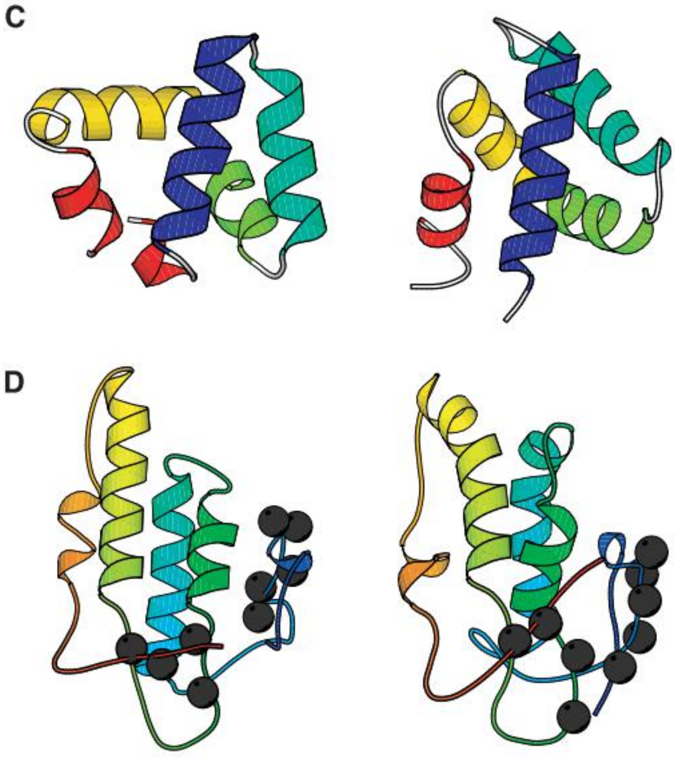

Fig 2. Sample applications of ab initio models

Fig 2 are the results of Rosetta de novo prediction from CASP4. C shows the de novo predicted protein structure which lyses bacteria (left) was found to be similar to a protein structure with similar function (nk-lysin; right). D shows the de novo predicted structure (left) of the signaling protein Frizzled is compared with the experimentally determined structure (right).

From the results, it can be found that the accuracy and reliability of the model generated by the de novo method is much lower than the comparison model based on the alignment with more than 30% sequence alignment, but in some cases, the basic topology of the protein or domain can be well predicted. Although it cannot improve high-resolution information, the resulting low-resolution model can reveal structural and functional relationships between proteins that are not visible from its amino acid sequence, and can be used to analyze evolutionarily conserved amino acid or experimentally. The display provides a framework for the spatial relationship between functionally important residues. Therefore, in favorable circumstances, ab initio prediction can provide some of the most important functional insights that can be obtained from the experimentally determined structure.

## 3. Deep learning methods in protein structure prediction

With the development of advanced DNA sequencing technology, it is much easier to obtain protein sequences than to obtain protein structures (as fig 3 shows), so protein sequences have rapidly accumulated. Compared to the exponentially increasing protein sequence, the number of protein structures grows slowly and steadily. Since not all target proteins can find a satisfactory template in the PDB, and because they are driven by the basic scientific issues of protein folding codes, the template-free approach begins being valued, it relies on a growing database of known sequences and uses the most advanced deep learning techniques to continuously improve algorithms to detect similarities between them. The following will introduce several machine learning methods to reflect their importance in the development of protein prediction.

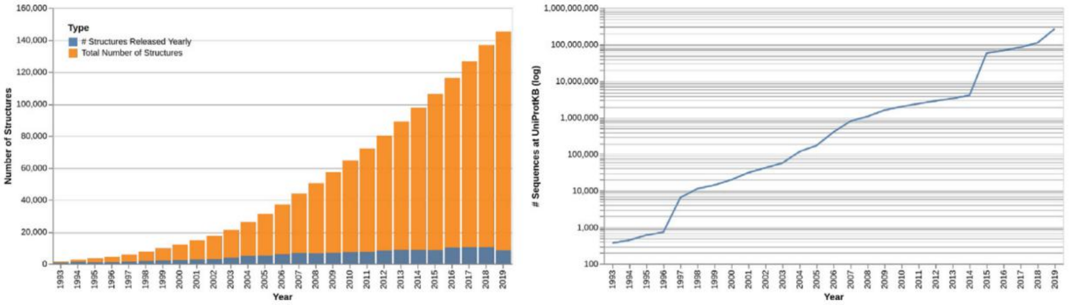

Fig 3. Growth of known structures in the Protein Data Bank (left) and known sequences in Uniprot (right)

The standard process for protein structure prediction envisions intermediate prediction steps in which abstraction can be inferred, which is simpler than a complete, detailed 3D structure, but also structurally very informative-we call it protein structure annotation (PSA) (Dill and MacCallum, 2012). The most commonly used PSAs are secondary structure, solvent accessibility, and contact maps. The first two are one-dimensional abstractions describing the arrangement of protein main chains, while the latter are two-dimensional projections of protein tertiary structures.

## 3.1 Methods for 1D Protein Structural Annotations

Predictors of the first generation PSA depended on statistical measurements of single AA propensities towards structural conformations, typically secondary structures,with per-AA accuracies usually less than 60%. In a second generation of predictors, knowledge from more than one AA at a time was fed to different methods, including FFNN , with secondary structure accuracies of about 63 – 64%(Qian and Sejnowski,1988). The third-generation PSA predictor is distinguished by the use of evolutionary information(Rost and Sander,1993), in the form of aligning multiple homologous sequences, as an input to the prediction method, which is almost a universal machine learning or deep learning algorithm. PDH, the early system in this generation achieved to make secondary structure prediction was at over 70% accuracy.

With the later use of PSI-BLAST or HMMER for mining evolution information, and the growing nature of available structures and sequence databases based on PSIPRED, the accuracy of secondary structure prediction continues to reach 76%.

Since the advent of the third-generation, predictor various machine learning such as SVM and k-nearest neighbor, and deep learning algorithms have been widely used for PSA prediction. Common predictors include SPIDER2, SSpro, SPIDER3, and RaptorX-Property, NetSurfP-2.0, etc Deep learning method for convolutional neural networks (Le and Pollastri, 2020).

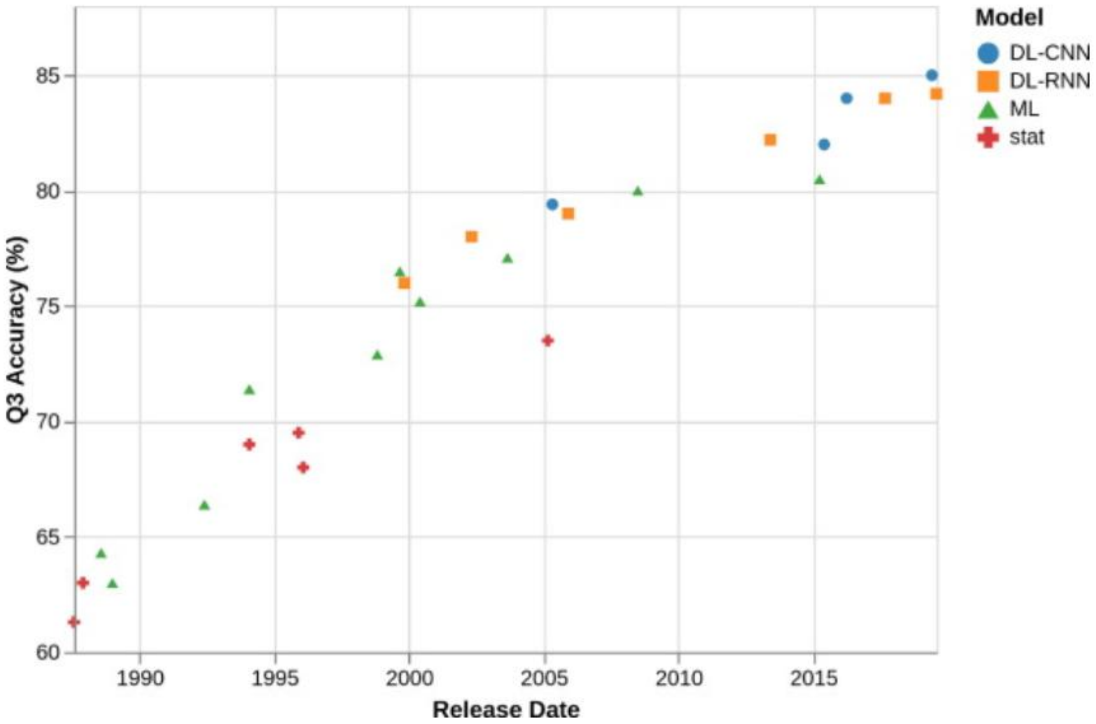

Fig 4. Performances of secondary structure predictors over the years

As fig 4 shows, with the continuous development of deep learning and machine learning methods, the accuracy of secondary structure prediction is increasing year by year, and the DL-CNN (Deep Learning methods based on Convolutional Neural Networksis) and DL-RNN(Deep Learning methods based on Recurrent Neural Networks) achieves much higher than the accuracy of using statistical methods ten years ago.

## 3.2 Methods for 2D Protein Structural Annotations

Usually distance maps and multi-class contact maps result in more reliable 3D structures than binary maps and appear to be more robust. Contact maps have been adopted since the 1990s to reconstruct the complete 3D protein structure. Nonetheless, most of the recent developments in Protein Structure Prediction were driven by Deep Learning methods used to predict contact or distance maps (Cheng et al., 2019).

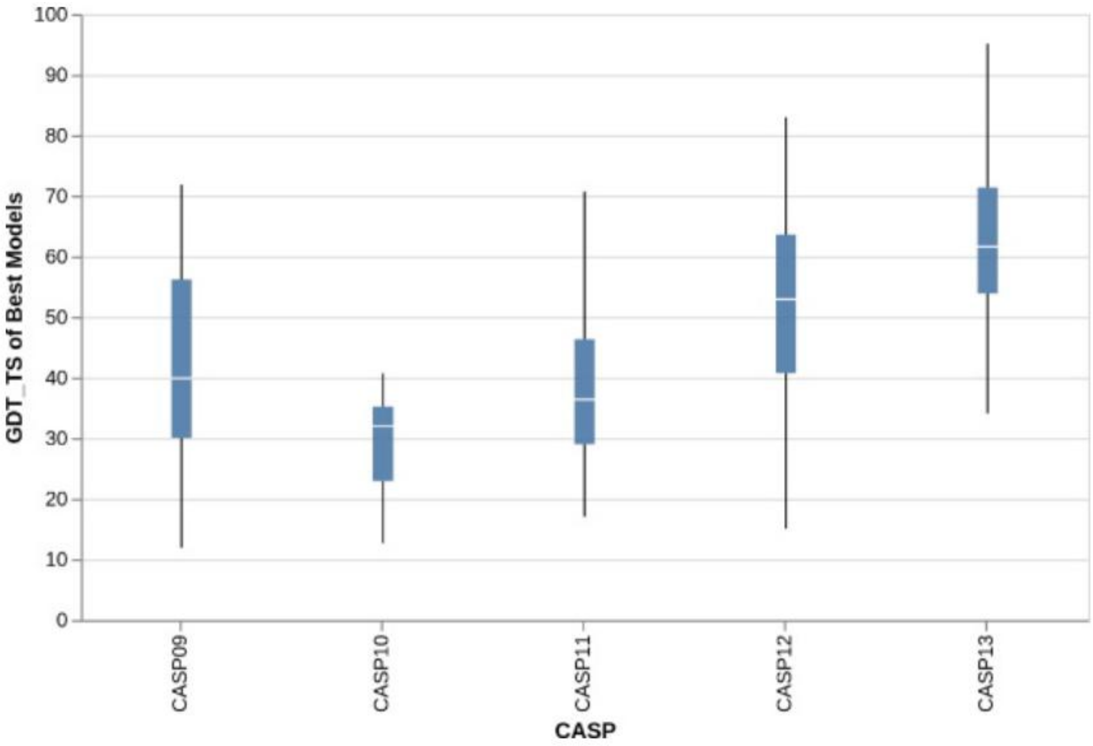

Fig 5. quality of 3D predictions for ab initio targets between CASP9 and CASP13

It can be seen from fig 4 that due to the constant advancement in contact and distant map that, the 3D prediction has been improving over the last few years.

Improvements especially during the last two editions are largely due to improved map predictions.

Next , this essay will introduce one of the most successful approaches to protein structure prediction named AlphaFold, which achieved the best performance in the Ab initio category of CASP13 (Kryshtafovych et al.,2019). It can accurately predict previously unknown folds. Although only using FM technology and not using templates, AlphaFold also scored high in the TBM category in the formula of the evaluator , ranking fourth in the top model and first in the best five models. Most of the accuracy of AlphaFold is because of the accuracy of distance prediction. For each amino acid, AlphaFold predicts $\Phi$ and $\Psi$ angles which are used to construct the predicted initial 3D structure.

The authors of AlphaFold concluded that the model's depth, its large crop size, the large training set of about 29,000 proteins, modern Deep Learning techniques, and the wealth of knowledge from the predicted distance histogram helped AlphaFold achieve high predictive contact map accuracy.

Although AlphaFold can match TBM without using models and can often outperform other ab initio methods, the accuracy for FM targets is still lagging behind that for TBM targets and still can not be relied on for a thorough understanding of hard structures, which is the bottleneck of this method (Senior et al,.2020).

## 4. Conclusion

This essay has described the progress in protein structure prediction, and through comparing the accuracy of protein structure prediction methods from the last two decades with the methods using machine learning and deep learning, emphasizing the significant role of deep learning and machine learning in the process. This essay has also introduced one of the most successful approaches to protein structure prediction, furthermore, through analyzing the limitation of this method, it can be concluded that, even ab initio methods are developing rapidly in the recent year, the most accurate approach is still the template-based approach, which will provide with more profound information about the protein structure. Protein structure prediction plays an indispensable and important role in the progress of biology, biotechnology, and drug design, for instance, contact predictions alone can guide biological insights, and one can predict which molecules or drugs can efficiently bind and how they will bind to protein by using the protein structure prediction.

It can be predicted that through the development of new technologies, protein structure prediction will have a higher level of improvement in the future.

## References

Bowie, J.U., Luthy, R. and Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253(5016), pp.164-170.

Jones, D.T., 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of molecular biology, 287(4), pp.797-815.

Torda, A.E., 1997. Perspectives in protein-fold recognition. Current opinion in structural biology, 7(2), pp.200-205.

MartÍ-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F. and Šali, A., 2000. Comparative protein structure modeling of genes and genomes. Annual review of biophysics and biomolecular structure, 29(1), pp.291-325.

Dill, K.A. and MacCallum, J.L., 2012. The protein-folding problem, 50 years on. science, 338(6110), pp.1042-1046.

Qian, N. and Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), pp.865-884.

Rost, B. and Sander, C., 1993. Prediction of protein secondary structure at better than 70% accuracy. Journal of molecular biology, 232(2), pp.584-599.

Cheng, J., Choe, M.H., Elofsson, A., Han, K.S., Hou, J., Maghrabi, A.H., McGuffin, L.J., Menéndez ‐ Hurtado, D., Olechnovič, K., Schwede, T. and Studer, G., 2019. Estimation of model accuracy in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), pp.1361-1377.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. and Moult, J., 2019. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics, 87(12), pp.1011-1020.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A. and Penedones, H., 2020. Improved protein structure prediction using potentials from deep learning. Nature, pp.1-5.

Le, Q., Torrisi, M. and Pollastri, G., 2020. Deep learning methods in protein structure prediction. Computational and Structural Biotechnology Journal.

Baker, D. and Sali, A., 2001. Protein structure prediction and structural genomics. Science, 294(5540), pp.93-96.