International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 079-081
e-ISSN:2278-621X

PROTEIN STRUCTURE PREDICTION A STUDY ON DIFFERENT METHODS

AshrithaD'Silva[1], Mallika[2], OshinTheodore[3] & Hemalatha N[4]

Abstract-The inference of the three-dimensional structure of a protein from its amino acid sequence is protein structure prediction, that is, the prediction of its folding and its secondary and tertiary structure from its primary structure.Protein structure prediction methods attempt to determine the native, in vivo structure of a given amino acid sequence. In this paper, we have studied various methods like statistical generation, machine learning which included neural networks, case-based reasoning, swarm intelligence and support vector machine, X-ray crystallography, GOR methods, stochastic methods and evolutionary algorithm.
Keywords- Neural networks, Support vector machine, Evolutionary algorithm, and stochastic method.

1. INTRODUCTION

Proteins are complex organic macromolecules which are essential for the functioning, structure and regulation of body's cells, tissues and organs. Proteins regulate a variety of activities of all organisms, from replication of the genetic code to transporting oxygen and are responsible for regulating the cellular machinery and determining the phenotype.The building blocks of proteins are amino acids. Amino acids are tiny molecules with a common structure. They have a central carbon atom attached to a hydrogen atom, an amino and a carboxyl group, and a fourth functional group (R), which is variable. Amino acids attach to one another through bonds called peptide bonds between the amino nitrogen and the carboxyl carbon. When the bond is formed, a water molecule is released. Using these peptide bonds, amino acids can join together in chains of nearly any sequence, which are known as polypeptides. When a polypeptide is of an appropriate size, structure and sequence, it functionally becomes a protein. A protein's function is ascertained by its structure rather than its sequence of amino acids. However, the sequence of the amino acids is important for determining the end structure of the protein. Functional proteins exhibit a tightly regulated structure, which is held together by hydrophobic interactions, hydrogen bonds and the Vander Waals forces between nearby amino acids, as well as di-sulphide bridges between cysteine residues. The protein structure has four levels of complexity.The primary structure describes the chain of amino acids in the polypeptide chain. The secondary structure describes the huge regular sub-structures. The α-helix and the β-sheet are two major sub-structures that form as secondary structure.The tertiary structure, which forms from the secondary structures, is the final 3D structure of the protein. It is held together by hydrophobic interactions, disulphide bridges, and hydrogen bonds and the Vander Waals forces.Conversely, some proteins can only function when two or more polypeptide chains form dimers or trimers, which are known as oligomers. The arrangement made by the formation of oligomers is called the quaternary structure.The characterization of the structure and function of proteins is of critical importance. As such, innovative methods are continually being developed to carry out this task. The protein structures are determined by techniques such as MRI (magnetic resonance imaging) and X-ray crystallography. These techniques require isolation, purification and crystallisation of the proteins.In the following section, we have studied various methods used for protein structure prediction. Methods user are statistical generation, machine learning, neural networks, evolutionary algorithms, GOR method based on information theory and Bayesian statistics, stochastic methods, minimum distance K-NN and fuzzy K-NN classifier and support vector machines. All the papers produced different results which are discussed in the next section.

2. LITERATURE REVIEW

Moult *et al.*, in their paper, have aimed at establishing the capabilities and limitations of current methods of modelling protein structure from sequence, to determine where progress was being made [1]. There have been five previous CASP experiments in the past.The structure of the experiment was very similar to that of the earlier ones. The prediction teams deposited models of the structures before the experimental results were public. The models were compared with the experiment, using numerical evaluation techniques and human assessment, and a meeting was held to discuss the significance of the results.
Techniques like statistical generation, machine learning which included neural networks, case-based reasoning, swarm intelligence and support vector machine have been used by Hendy *et al.*, in their paper [2]. They have concluded that the

[1]Department of IT & Bioinformatics, AIMIT, Mangalore, Karnataka, India
[2] Department of IT & Bioinformatics, AIMIT, Mangalore, Karnataka, India
[3] Department of IT & Bioinformatics, AIMIT, Mangalore, Karnataka, India
[4] Department of IT & Bioinformatics, AIMIT, Mangalore, Karnataka, India

accuracy protein secondary structure prediction by statistical methods was very low. The accuracy accelerated with the use of intelligent techniques.

Chandonia*et al.*,in their paper, have used and studied a set of 681 chains of proteins of high-resolution structures available in the Brookhaven Protein data bank(PDB) [3]. First, protein chains from all well-resolved structures in the PDB were classified into groups according to sequence homology. The structure with the highest resolution in each group was taken to represent that group. All structures determined by X-ray crystallography had a resolution of 3.0 A˚ or better; 27 chains for which only NMR-determined structures were available were used in addition. Filtering was done to ensure that no pair of protein chains had more than 25% sequence identity. Pairwise alignments were done using global dynamic programming with the identity substitution matrix and a constant gap penalty of 3.

The data used to train the neural networks from the PDBhas been prepared by Peterson *et al*., in their paper [4]. For X-ray structures a resolution cut-off of 2.5 A˚ was used. Sequence profiles were generated with the program PSI-BLAST. Predictions were made using neural networks trained on a set of 1032 high-quality protein chains non-sequence similar to the RS126 set. It was found out that an increase in the accuracy of secondary structure prediction could be obtained by combining many neural network predictions.

SubhenduBhusan Rout *et al.,* in their paper, have described that protein structure prediction is the process of prediction of the three dimensional structure of protein from its amino acid sequence [5]. They have found out that the application of bioinformatics provides a gateway to process huge amount of data and that the genetic algorithm is very useful for designing of various drugs, after processing of huge amount data with less amount of time.

Donald Petrey and Barry Honig in their paper, have described that there has been significant progress in the ability to predict the three dimensional structure of protein from their amino acid sequence [6]. They have studied that protein structure prediction has been through of as a "Grand challenge" forsome time. There has been rapid progress in the past few years, much of it made possible by the massive amounts of data that have become available for analysis.

D. Ramyachitra and V. Veralakshmi in their paper, have described the protein structure prediction problem and some of the techniques to predict the structure [7]. They have used evolutionary algorithms forprotein prediction problem to find out the structure from a given amino acid sequence and the algorithm is used to anticipate the structure.

Rajbin Singh *et al.,* in their paper, have described that they have focused on secondary structure prediction of amino acid residues to predict the 2D structure using different input formats of sequences[8]. They have studied the GOR method based on information theory and Bayesian statistics which is quite successful in its accuracy of secondary structure prediction.

The predictions of the three dimensional protein structure using stochastic methods was undertaken by Rout*et al*., in their paper [9]. Stochastic method will help to predict protein structure or study of DNA, RNA in a small amount time. Stochastic method will help to design the drugs. Their work proposed the idea of various real time effects and application of bioinformatics and probability theory on various macro molecules.

O.Day*et al.*,in their paper, have predicted protein structure by applying evolutionary algorithm[10].Their investigation summarizes their progress of using *MofmGA*, modified to scale its efficiency to 4.7 times .The new algorithm has the capabilities to be single and multiple objectives and run with single and multiple competitive templates all configurable. Computational results supports their hypothesis that the MO version provides more acceptable results.

AshishGhoshand BijnanParai in their paper,have made an attempt to map the protein secondary structure prediction problem as pattern classification problem and used three different low cost pattern classification techniques for solving it [11].They used minimum distance K-NN and fuzzy K-NN classifier, among which minimum distance produced best results. The main problem in protein secondary structure prediction is that the data cannot be used directly to classifiers. They had done some patterns in a particular protein are classified correctly by a classifier better than other classifier.

AnureetKaurJohal and Rajbir Singh in their paper, have worked to compare the performance of support vector machines and neural networks in predicting the secondary structure of protein from their amino acid sequences[12].They summarized that neural network provides much better accuracy as compared to Support Vector Machine. They also found out that neural network uses much lesser computational time than support vector machine. Their results also reveal that support vector machine requires much larger memory and powerful processor as compared to neural network.

3. CONCLUSION

In the above papers, various methods have been carried out by different authors to predict the structures of proteins. It was found out that these methods were highly accurate so as to giving suitable results. Protein secondary structure prediction by statistical methods was very low. Procedures were performed on the basis of previously carried out experiments in order to improve the accuracy by using intelligent techniques and neural network predictions. In the future, we aim at working on new methods and comparing the results with those of the existing methods.

4. REFERENCES

[1] J. Moult, K. Fidelis, B. Rost, T. Hubbard and A. Tramontano,"Critical assessment of methods of protein structure prediction (CASP)—round 6", Proteins: Structure, Function, and Bioinformatics, vol. 61, no.S7, pp.3-7, 2005.

[2] H. Hendy, W. Khalifa, M. Roushdy and A.B. Salem,"A study of intelligent techniques for protein secondary structure prediction", Int J Inf Mod Anal, vol. 4, no.1, pp.3-12, 2015.

[3] J.M Chandonia and M. Karplus,"New methods for accurate prediction of protein secondary structure", Proteins: Structure, Function, and Bioinformatics, vol. 35, no.3, pp.293-306, 1999.

[4] T.N. Petersen,C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippertand O. Lund, "Prediction of protein secondary structure at 80% accuracy", Proteins*:* Structure, Function, and Bioinformatics, vol. 41, no. 1, pp.17-20, 2000.

[5] P. Guo,M.Z. Zhang and H.Z. Chen,"A Membrane Computing Model for Genetic Algorithm", Journal of Residuals Science &Technology, vol. 14, no. 3, 2017.

[6] D. Petrey and B. Honig, "Protein structure prediction: inroads to biology", Molecular cell, vol. 20, no. 6, pp.811-819, 2005.

[7] D. Ramyachitraand V. Veeralakshmi,"Computational Analysis of Protein Structure Prediction and Folding", Int. J. Comput. Sci. Inform. Technol. Secur, vol. 4, pp.116-127, 2014.

[8] R. Singh,N. Jain and D.P. Kaur, "GOR Method for Protein Structure Prediction using Cluster Analysis", International Journal of Computer Applications, vol. 73, no.1, 2013.

[9] S.B. Rout and S. Mishra, "Protein Structure Prediction Using Stochastic Process Probabilistic Model".

[10] O.R. Day,G.B. Lamont and R. Pachter,"Protein structure prediction by applying an evolutionary algorithm", In Parallel and Distributed Processing Symposium,Proceedings. International (pp. 8-pp). IEEE, 2003.

[11] A. Ghosh and B. Parai,"Protein secondary structure prediction using distance based classifiers", International Journal of Approximate Reasoning, vol. 47, no1, pp.37-44, 2008.

[12] A.K. Johal and R. Singh,"Protein secondary structure prediction using improved support vector machine and neural networks", International Journal of Engineering and Computer Science, vol.3, no. 1,pp.3593-3597, 2014.