International Journal on Advances in Engineering Technology and Science
Volume: 5, Issue: 1, 2024, e-ISSN: 2455-3131
DOI: 10.5281/zenodo.10718952

A Brief Survey on Techniques for Protein Sequence Analysis

Prativesh Pawar[1], Pinaki Ghosh[2]

[1]Research Scholar, [2]Professor
Sanjeev Agrawal Global Educational University, Bhopal

[1]prativesh.acro@mail.com, [2]pinaki.g@sageuniversity.edu.in

***Abstract*-There are currently a lot of biological data available, and data mining is essential in sorting the data. Many research on the use of data mining in bioinformatics have been conducted as a result of the efficacy of data mining techniques in all facets of computational biology. Over the past two decades, a body of literature on data mining methods in bioinformatics analysis has grown. A periodic examination of survey articles is essential, and grouping them makes it easier for the researcher to identify the study. This document also teaches non-specialists how to select among a variety of currently used strategies based on their strengths and weaknesses. In this study, an effort is made to offer a thorough analysis of the algorithms that are optimal for obtaining the desired outcome.**

***Keywords:* Deep learning; natural language processing; protein annotation; protein language model; protein sequence embedding; survey of embedding models.**

I. Introduction

Proteins, the organism's most conspicuous macromolecules, are involved in nearly every biological function. Many critical operations are carried out by protein macromolecules that include structural support for cells; immunological defence; enzymatic catalysis; transmission of cell signals; control of transcription and translation. The diverse three-dimensional structures that distinct protein molecules choose make this feasible. Based on ground-breaking studies from the 1970s, it is believed that a protein's amino acid sequence influences its tertiary structure [1]. Biological research has remained anchored in the sequence-structure- function paradigm of proteins since the early 1980s. The GenBank database now contains more than 2600 million known nucleotide sequences as of 2021, thanks to significant achievements in genome sequencing over the preceding four decades [2-5]. Nearly 200 million nucleotide sequences have been translated into amino acid sequences in UniProt [6].

Despite the abundance of information, it is difficult to deduce the biological activities of proteins only from their amino acid sequences. This is because the three- dimensional structures of proteins are mostly responsible for these activities. Intrinsically disordered proteins, which comprise for up to 30% of the human proteome, are an intriguing exception to this norm [7- 9]. They can function even if they lack well defined tertiary structures. When proteins bind to their binding partners and carry out their biological functions, they can undergo disorder-to-order transitions and gain tertiary structures. Some of the most exact methods for discovering a protein's structure include cryo-electron microscopy [10], NMR spectroscopy [11], and X-ray crystallography [12]. Prediction and identification of protein structures are crucial for biological processes to take place. Protein structures have been experimentally solved significantly more slowly than protein sequences because solving a protein structure requires significant human labour and expenditure [13–18]. The many levels of protein structures are shown in Figure 1.

Prediction approaches for protein structure are classified as either template-based or template-free based on the use of a template structure (FM). TBM generates models by duplicating and enhancing the structural frameworks of other related proteins found in the PDB, whereas FM aims to predict protein structures without reference to global template structures. Methods like FM are also referred to as "ab initio" or "de novo" modelling. To decrease the gap between the amount of known protein sequences and experimentally solved structures, the most effective but difficult method is to use extremely accurate protein structure prediction algorithms. These findings also reveal the fundamental principles that drive the sequence-to structure to function paradigm in proteins.

In order to accurately forecast the structure of a protein, it is necessary to anticipate the secondary structure. To predict protein secondary structures, bioinformatics uses simply the amino acid sequence of the proteins under question. Secondary structural components such as helices and sheets are widely considered to make up

proteins. Secondary structures are densely packed in a hydrophobic environment within the protein core. Prediction of a protein's three-dimensional structure from its amino acid sequence is known as protein structure prediction. From the primary structure, it is possible to extrapolate secondary and tertiary structures. Protein sequences can be solved in part using three- dimensional structures. It has been proven that adopting consensus approaches that integrate complimentary algorithms to increase prediction quality is a valuable tool in the CASP (Critical Assessment of Structure Prediction) competitions [19, 20].

There have been several advances in computational protein structure prediction during the last few decades. Proteins fold themselves into three-dimensional structures using only the information stored in their residues. Systems called protein structure predictors, which can use information from the protein sequence to limit possible local and global conformations, are used to guide folding. In consensus techniques, template selection and model averaging are typically done by a majority vote. Primary methods that diverge or converge on the erroneous fold will deviate from the original fold, which can improve quality. Machine learning techniques, particularly artificial neural network models, have long been used in structural bioinformatics. To create larger abstractions, deep learning algorithms are able to dismiss minor changes in input. This is due to the fact that deep learning techniques are increasingly being applied in protein structure prediction because of these two characteristics and the rising availability of protein datasets.

Prediction approaches for protein structure are classified as either template-based or template-free based on the use of a template structure (FM). TBM generates models by duplicating and enhancing the structural frameworks of other related proteins found in the PDB, whereas FM aims to predict protein structures without reference to global template structures. Methods like FM are also referred to as "ab initio" or "de novo" modelling. To decrease the gap between the amount of known protein sequences and experimentally solved structures, the most effective but difficult method is to use extremely accurate protein structure prediction algorithms. These findings also reveal the fundamental principles that drive the sequence-to structure to function paradigm in proteins.

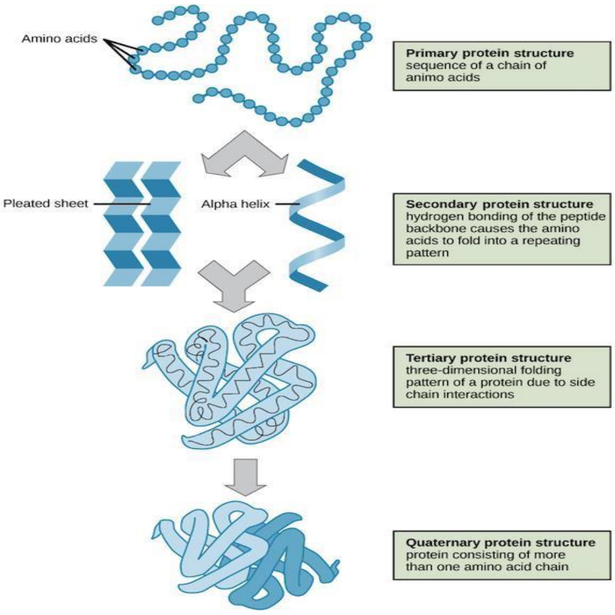
Amino acids
Primary protein structure
sequence of a chain of animo acids
Pleated sheet
Alpha helix
Secondary protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern
Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions
Quaternary protein structure
protein consisting of more than one amino acid chain

Fig-1. Levels of Protein Structure

Primary methods that diverge or converge on the erroneous fold will deviate from the original fold, which can improve quality. Machine learning techniques, particularly artificial neural network models, have long been used in structural bioinformatics. To create larger abstractions, deep learning algorithms are able to dismiss minor changes in input. This is due to the fact that deep learning techniques are increasingly being applied in protein structure prediction because of these two characteristics and the rising availability of protein datasets.

In order to predict a protein's structure, one must first know the amino acid sequence, which is the beginning point for this procedure. The application of deep machine learning algorithms [21] to provide high- quality geometric feature predictions has had a significant impact on the science of protein structure prediction. Many structural elements, such as contact lengths and inter-substance torsion angles, can be forecasted using deep learning. Additionally, multi- layer neural networks provide a high degree of model training accuracy. The application of deep machine learning methods to provide high-quality geometric feature predictions has revolutionized protein structure prediction. When it comes to predicting various structural features like hydrogen bonds and inter- residue interactions, deep learning excels. As an image segmentation problem, the pair-wise contact prediction problem was reformed as a pair of residues representing a pixel in the image. It was AlphaFold that introduced distance maps as a tool for guiding folding. Deep neural networks were used to predict distance maps based on

10query sequences. These maps were then used to guide the assembly of fragments and folding simulations based on gradient descent. AlphaFold also employed deep learning to create new structural components as part of their groundbreaking fragment creation technique. Deep learning-based contact and distance map prediction has raised the topic of what other limitations deep learning can effectively foresee. Inter- residue angle orientation is a simple extension of distance prediction. Just as distance information cannot tell two mirror images of the same structure different, it is impossible to determine the geometry of a structure without first learning about its torsional angle orientations. Inter-residue torsion angle prediction was recently broadened by trRosetta, who used an integrated deep ResNet to predict pairwise residue distances and torsion angles between residues based on co- evolutionary features.

Thus, novel designs such as training gimmicks, weight-optimization strategies, and Recurrent NNs that are commonly utilized to anticipate secondary structure, solvent accessibility, disorder and backbone torsion angles continue to be developed at an accelerated rate. End-to-end differentiability means that during NN training, all parts of the procedure, starting with 3D coordinate prediction, can be altered simultaneously. As an example, the impact of deep reinforcement learning and generative models on CASP has been minimal (Critical Assessment of Structure Prediction).

Computational approaches have dramatically reduced the time and expense of generating new medicines. To deal with challenges of all shapes and sizes, we'll have to use a variety of drug screening and design methodologies. Machine learning and deep learning approaches, which go beyond the constraints of prior studies, are the primary emphasis of this study.

II. LİTERATURE SURVEY

Multi-objective evolutionary techniques that incorporate Rama torsion angle-based sampling, loop- based resampling, stochastic rank-based selection, loop-based crossover, and near native sampling have been devised by Zhang Wei-Li et al. [Nov 2018].' The secondary structural similarity criterion could be used to overcome the energy function's inaccuracy.

For protein secondary structure prediction, Zhou et al. [23] used a combination of deep neural networks and reinforcement learning (CDNN). In addition to the abstraction skills of CNN and the sequence data processing capabilities of LSTM, CDNN also includes a significant classification capability. The cross-entropy error between protein secondary structure labels and dense layer outputs is used to train the CDNN architecture.

Deep ResNet was developed by Xu, Jinbo et al. [24] to predict protein contact/distance and template-free protein folding. According to the findings, deep ResNet has made significant progress in recent years in predicting protein-protein interactions and tertiary structure. Because it does not need evolutionary knowledge to make predictions about natural protein folds, the proposed deep ResNet can correctly fold most of the proteins developed by humans.

It was proposed by Xu and colleagues [25] that a computational method termed "deep structural inference" may be used to predict protein residue/residue interactions using a deep-learning algorithm and template-based structural modelling. Large-scale tertiary structure prediction of more than 1,200 single-domain proteins for the first time reveals much superior predictive capabilities. In addition, it appears that the insights offered by statistical co- evolutionary analyses cannot simply be replaced by providing the unprocessed frequency distributions from several sequence alignments, as CCMPred did to obtain the coupling scores.

End-to-end differing recurrent geometric network (RGN) was developed by Chowdhury et al [26] to predict protein structure from individual sequences. When multiple sequence alignment is not possible for orphan and designer proteins, this is a computationally efficient technique that has numerous advantages. With the help of a simple approach to explain the C backbone's geometry, RGN2 accomplishes this.

Protein secondary structure prediction was improved by Guo et al. [27] by developing a multi-advanced deep belief network-based approach. They were able to boost forecast accuracy by more than 80% as a result of their efforts. The experiment also demonstrated how to forecast secondary structure using hidden Markov model profiles based on emission/transition probability. However, the network's feature set will be imbalanced. By feeding a protein feature vector, which combines the suggested MOS descriptor with AA classification, into a DNN, Wang et al [28] were able to predict the PPIs. The suggested MOS descriptor has the ability to take into account the order connection of the entire AA sequence, in contrast to earlier protein representation like AC, CT, and LD. The ReLU AF, ADAM optimizer, and cross-entropy as the cost function were chosen as network settings for the assignment with the author providing adequate justifications. By adjusting their range, the other parameters, including network depth, width, and the LR, were computed for the specific approach and the best ones were chosen. Finally, the author separately trained the DNN model with AC, CT, and LD and compared their results with the suggested Work.

Jha and Saha [29] carried out another intriguing and novel piece of work utilising an LSTM-based classifier that integrated features produced by two distinct protein modalities, i.e. sequence-based and structure-based information.

First, using the structural representation of the proteins, three separate protein representations based on three distinct attributes were obtained, and then, using a ResNet50 model, corresponding feature sets were obtained. In 2018, Li et al. presented the first study on sequence-based PPI prediction using DNs that was completely based on auto-feature engineering, i.e. without the addition of manually derived features [30]. The input must be numerical for the NN architecture to learn the data. In order to transform the protein sequence, the author randomly assigned a natural number to each AA.

Moreover, Gonzalez-Lopez et al. [31] conducted PPIs prediction without the use of feature engineering by using RNNs and embedding systems. Every triplet in the sequence was given a token (an integer) as part of the tokenization procedure in order to represent the sequence numerically. Each protein's pair representation in the NN was fed to and analysed by two branches with a comparable design independently. The architecture's embedding, recurrent, and FC layers each fulfilled a distinct function. To prevent over-fitting and input standardisation, two crucial parameters Dropout and Branch normalisation were also applied.

III. OBSERVATİONS AND DİSCUSSİON

The outcomes of CNN and a combination of GAN and CNN were compared by Y. Zhao et al [Nov 2020] An improvement in accuracy may be shown when compared to using solely GAN algorithm for feature extraction. Compared to older methodologies, the CASP10 and CASP11 datasets produce more precise results overall.

For protein secondary structure prediction, Zhou et al. [23] used a combination of deep neural networks and reinforcement learning (CDNN). The CDNN technique's efficacy is demonstrated via empirical validation on two different datasets. The forecast, on the other hand, is unaffected by the imbalances. Because of this, forecasts become less accurate.

Deep ResNet was developed by Xu, Jinbo et al. [24] to predict protein contact/distance and template-free protein folding. When it comes to using inter-residue orientation information, the proposed 3D modelling approach is still less advanced and more rudimentary than previous approaches.

End-to-end differing recurrent geometric network (RGN) was developed by Chowdhury et al [26] to predict protein structure from individual sequences. In order to reconstruct the backbone's structure sequentially, this technique must restrict itself to just limiting local dependencies between C atoms (curvature and torsion angles).

Recently, DL technology has become a focal point in both the scientific community and the corporate world, thanks to a number of high-profile studies. Whilst ML has already facilitated remarkable progress in bioinformatics, DL is expected to yield even more substantial and encouraging outcomes in this area. In this study, we conduct a comprehensive evaluation of the three DL architectures—DNNs, CNNs, and RNNs—and their variants in the context of PPI prediction based on sequence information. We also cover the different strategies in terms of data, objectives, and DL architecture structure, and provide optimal values for the relevant parameters.

While it has been established that all of the architectures in question are capable of yielding positive results in this domain, many new challenges have emerged and must be addressed before any of them can be considered fully resolved. These include, but are not limited to, insufficient data and picking an architecture that makes the most of beneficial hyperparameters. Therefore, widespread adoption of DL methods requires extensive study. So, the researchers can benefit from the in-depth analysis presented here, which was developed after painstakingly mining every accessible piece of information. It is hoped that the insights gained from this overview of the literature on the topic of DNs in PPI prediction would be useful to researchers in their ongoing efforts.

IV. Conclusion and Future Trends

Mining data is essential for the organisation of the huge quantities of biological data that are currently available. There have been a multitude of research studying the application of data mining techniques in bioinformatics due to the fact that these methods are so universally beneficial in the field of computational biology. During the course of the previous two decades, research on data mining techniques for bioinformatics analysis has accumulated a substantial body of written material. It is crucial to regularly evaluate survey papers, and doing so is made easier when the articles are categorised into subjects that are conceptually related to one another. This document also advises readers who are not specialists in the industry on how to balance the pros and disadvantages of various approaches and make a decision that is informed by the information gathered. The goal of this study is to produce a detailed analysis of the algorithms that are most effective at achieving the objectives that have been established.

References

[1] Anfinsen, Christian B. "Principles that govern the folding of protein chains." Science 181, no. 4096 (1973): 223-230.

[2] Sanger, Frederick, Steven Nicklen, and Alan R. Coulson. "DNA sequencing with chain- terminating inhibitors." Proceedings of the national academy of sciences 74, no. 12 (1977): 5463-5467.

[3] Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith et al. "The sequence of the human genome." science 291, no. 5507 (2001): 1304-1351.

[4] Metzker, Michael L. "Sequencing technologies— the next generation." Nature reviews genetics 11, no. 1 (2010): 31-46.

[5] Sayers, Eric W., Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt, and Ilene Karsch- Mizrachi. "GenBank." Nucleic acids research 47, no. D1 (2019): D94-D99.

[6] Bairoch, Amos, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger et al. "The universal protein resource (UniProt)." Nucleic acids research 33, no. suppl_1 (2005): D154-D159.

[7] Deiana, Antonio, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. "Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell." PloS one 14, no. 8 (2019): e0217889.

[8] Uversky, Vladimir N. "Unusual biophysics of intrinsically disordered proteins." Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1834, no. 5 (2013): 932-951.

[9] Wright, Peter E., and H. Jane Dyson. "Linking folding and binding." Current opinion in structural biology 19, no. 1 (2009): 31-38.

[10] Maveyraud, Laurent, and Lionel Mourey. "Protein X-ray crystallography and drug discovery." Molecules 25, no. 5 (2020): 1030.

[11] Cavanagh, John, Wayne J. Fairbrother, Arthur G. Palmer III, and Nicholas J. Skelton. Protein NMR spectroscopy: principles and practice. Elsevier, 1995.

[12] Cheng, Yifan. "Single-particle cryo-EM at crystallographic resolution." Cell 161, no. 3 (2015): 450-457.

[13] Yang, Jianyi, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. "The I- TASSER Suite: protein structure and function prediction." Nature methods 12, no. 1 (2015): 7-8.

[14] Ovchinnikov, Sergey, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. "Protein structure determination using metagenome sequence data." Science 355, no. 6322 (2017): 294-298.

[15] Wang, Sheng, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. "Accurate de novo prediction of protein contact map by ultra-deep learning model." PLoS computational biology 13, no. 1 (2017): e1005324.

[16] Zheng, Wei, Yang Li, Chengxin Zhang, Robin Pearce, S. M. Mortuza, and Yang Zhang. "Deep- learning contact-map guided protein structure prediction in CASP13." Proteins: Structure, Function, and Bioinformatics 87, no. 12 (2019): 1149-1164.

[17] Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin et al. "Improved protein structure prediction using potentials from deep learning." Nature 577, no. 7792 (2020): 706-710.

[18] Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. "Improved protein structure prediction using predicted interresidue orientations." Proceedings of the National Academy of Sciences 117, no. 3 (2020): 1496- 1503.

[19] 10Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) Round X. Proteins: Struct., Funct., Genet. 2014, 82 (S2), 1−6.

[20] Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) Round XII. Proteins: Struct., Funct., Genet. 2018, 86 (S1), 7−1.

[21] LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 2015, 521:436-444.

[22] Wang S, Sun SQ, Li Z, Zhang RY, Xu JB: Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 2017, 13.

[23] Zhou, Shusen, Hailin Zou, Chanjuan Liu, Mujun Zang, and Tong Liu. "Combining deep neural networks for protein secondary structure prediction." IEEE Access 8 (2020): 84362-84370.

[24] Xu, Jinbo, Matthew Mcpartlon, and Jin Li. "Improved protein structure prediction by deep learning irrespective of co-evolution information." Nature Machine Intelligence (2021): 1-9.

[25] Xu, Jinbo, Matthew Mcpartlon, and Jin Li. "Improved protein structure prediction by deep learning irrespective of co-evolution information." Nature Machine Intelligence (2021): 1-9.

[26] Chowdhury, Ratul, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M. Church, Peter Karl Sorger, and Mohammed N. AlQuraishi. "Single-sequence protein structure prediction using language models from deep learning." bioRxiv (2021).

[27] Guo, Zhiye, Jie Hou, and Jianlin Cheng. "DNSS2: improved ab initio protein secondary structure prediction using advanced deep learning architectures." Proteins: Structure, Function, and Bioinformatics 89, no. 2 (2021): 207-217.

[28] Wang X, Wu Y, Wang R, Wei Y, Gui Y. A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences. PLoS ONE. 2019;14(6): e0217312.

[29] Jha K, Saha S. Amalgamation of 3D structure and sequence information for protein–protein interaction prediction. Sci Rep. 2020;10(1):1–14.

[30] Li H, Gong XJ, Yu H, Zhou C. Deep neural network based predictions of protein interactions using primary sequences. Molecules. 2018;23(8):1923.

[31] Gonzalez-Lopez F, Morales-Cordovilla JA, Villegas-Morcillo A, Gomez AM, Sanchez V. End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 2344–2350