International Journal of Engineering & Technology, 7 (4.5) (2018) 168-170

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Deep Learning Approaches for Protein Structure Prediction

Khatri Chandni[1]*, Prof.Mrudang Pandya[2], Dr Sunil Jardosh[3]

[1]*M-Tech Student, Department of Information Technology, CSPIT Charusat- Anand, India*
[2]*Assistant Professor, Information Technology Department CSPIT Charusat-Anand India*
[3]*Principal Software Engineer Progress Software-Hyderabad, India*
**Corresponding author E-mail: 16pgit002@charusat.edu.in ,khatrichandni53@gmail.com*

Abstract

In recent years, Machine Learning techniques that are based on Deep Learning networks that show a great promise in research communities.Successful methods for deep learning involve Artificial Neural Networks and Machine Learning. Deep Learning solves severa problems in bioinformatics. Protein Structure Prediction is one of the most important fields that can be solving using Deep Learning approaches.These protein are categorized on basis of occurrence of amino acid patterns occur to extract the feature. In these paper aimed to review work based on protein structure prediction solve using Deep Learning Networks. Objective is to review motivate and facilitatethese deep learn the network for predicting protein sequences using Deep Learning.

Keywords: *Bioinformatics, Protein Contact Mapping, Protein-Protein Interactions, Protein Structure Prediction, Protein Docking, Protein Folding, DeepLearning*

1. Introduction

Protein Structure Prediction for secondary structure is one of themost important for studying protein structure and function of protein.Understanding of protein secondary structure beyond yields that give benefits to understand human disease and developing thetherapeutic drugs and enzymes. Several attempts at Ab initio secondary structure prediction utilize statistical approaches for employing the data collection from protein with known protein secondary structure of proteins, these methods could not often archive accuracy around 65%[1][2][3]. But recently a novel approach for Ab-Inito approach for protein tertiary structure prediction by Matt spancer et.[1] improve the accuracy around 80.7%. When this sequence (profile) information used for input features, currently the best prediction can be obtained around 80% that improve over past decades. Achieving higher accuracy around 80% is one of the most challenging tasks for researcher.

By using these Neural Network that have effectively used in various variety of classification as well as predicting algorithm including character recognition, speech recognition, weather recognition, face recognition as well for several bioinformatics fields also like Protein Structure Prediction, Protein Docking, Protein Folding, Protein-Protein Interactions etc. These types of network that mainly allows to do prediction for recognizing a complex relationships, as well as if knowledge of these relationship is also more complex which is not known to us. Weights are assigns to these nodes of hidden layers. During the training procedure it will adjust these weights to make output; these output layers is more likely to be reflect the result which is derived from such examples. When

these weights are sets, these information from unknown target input uses as input, allowing these network to predict form unknown properties. Using these multiple hidden units to train these supervised as well unsupervised learning using these Deep Neural Network Architectures[4].

Several Researchers are working of Neural Network based on applied techniques for experimenting such novel Deep learning techniques to stimulate the progress of neural network plays an important role for secondary structure prediction. Using these Conventional Machine Learning techniques- Naïve Bayesian, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine(SVM) and with the use of single hidden layer neural network that have limited in complexity of such function ,these neural network are not efficiently learn. Methods require extracting the hidden features as well complex pattern to be taken out to learn the network. This ability is provided to learn features that learn automatically for such protein. This paper attempt to review and overview of such Deep Learning that contribute to advancement of this fields.

Deep learning or Machine Learning approaches involve several deep learning architectures that learn using high level representation of data[5]. Deep neural network(DNN) is an artificial neural networks having multiple layers of units between input and output layers [6]. To shallow these artificial Neural Networks, Deep Neural Networks model have complex non linear relationshipamong these architectures. DNN architectures have a complex model for object expression as a layer composition.

For observation that can be represented in many ways (for images)such as vector of intensity for each pixel, more abstract way for detection of edges, regions of different shapes etc. several representation are much more easier to learn the model. One of the promising handcrafted features in deep learning is to extract the features algorithm for supervised learning, unsupervised learning, semi-supervised learning also for hierarchical feature extraction[7]. Deep Learning architectures that give best performance in several domains like image processing, face recognition, text recognition also for domains for bioinformatics.

Fields of bioinformatics solve using deep Learning architecturesProtein Structure Prediction, Protein Function Prediction, Protein

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Docking, Protein–Protein Interactions, Protein Contact Predictions, Protein Order-Disorder Prediction[6,8]. Additional several Protein Structure Predictions using Deep Learning have develops using several prediction tools that utilize the global information of such protein sequences [10].

2. Review Study

Matt et-al. have came up with useful research on secondary structure prediction of protein using Deep learning [1].Researcher have presented an Ab-initio methods for predicting the secondary structure of protein that employing deep learning networks architectures train using PSSM(position-specific score matrix)of such protein sequences and Atchley's factors of protein residue. Systematic approach was used that determine effective parameter to train the network which provide a variety of option for input profile, window size and architecture attempt to make these deep network more interactive. Training method that emphasize the improvement of Q3 and SOV score. They have produce the workflow capable of producing the prediction of protein for secondary structure sequences having average SOV score of 74.2% ad Q3 score of 80.7% on fully independent testing set over 198 protein sequences of CASP9 (105 protein) and CASP10((93 protein). They have compared their performance with PSIPRED, SSPRO, PSSpred and RaptorX. Overlay these performance similarity with DNSS having slightly having lower Q3 accuracy and higher SOV accuracy for evolution resulting tools.

DNSS show the same prediction as the other methods can do. Machine Learning techniques dominate in the field of Secondary Structure Prediction that fails to produce significant improvement using sophisticated implementation of machine learning. Perhaps discovery of better features is necessary for secondary structure prediction giving accuracy around ~80%.As a part of investigation they have tested the impact of Atchley's factor as a feature vector, also by adding then to PSSM(Position-Specific Score Matrix) information to increase the accuracy of prediction of protein. However the features including them was quite slight while testing them with other combination of protein. In addition to that Atchley's (FAC) factor appear to decrease the accuracy of the prediction (because these factor was not benefited to all).

Akosu et-al and Navdeep et-al. [9] in their work they have use deep learning techniques for improving the accuracy using novechained convolutional neural architecture with next step conditioningof protein structure prediction. These model achieve 70% of amino acid sequence on CB513 dataset. Using these model that create a state of art of convolution architecture principle to make changes ad to provide the boost the performance using fully connected baseline for providing insight to 8 class fundamental difficulty for protein secondary structure prediction problems. Zhou et-al. & Troyanskaya et-al.[10] in their work they have introduc deep learning approaches for Local Secondary structure of proteins that utilize convolutional generative stochastic network(GSN) on CB513 PDB Databank.
There are mainly two main aspects behind:

1. To construct supervised GSN that infers the distribution of output label for such input data.
2. With the use of convolutional architecture for GSN learning efficiently on high-dimensional large dataset by building the deep hierarchical representation.

By using these Multi-Layer Convolutional Generative Stochastic Networks for capturing the global information among the protein sequences. For predicting the protein secondary structure prediction for 8class is more challenging problems as compare to 3class prediction. These model predict the 4 major states- α-helix, β-stands, loop/irregular and β-turns with higher accuracy. These can predict the performance as low for appearing the structures (bend, 3_{10}-helix, β-bridge, π-helix) due to some imbalance labels present in the data. With these problems for imbalance of data should be improve for further prediction.
Zeming et-al,Yanjun et-al.[11] in their work they have use multilayer shift and stitch convolutional architectures(MUST-CNN) for Protein secondary structure. By taking inspiration from image classification domain these MUST-CNN to predict these protein properties. Author have compare the previous state-of- the-art method for protein structure prediction using multilayer perceptron network(Qi et al.2012[12], Drozdetskiy et al. 2015[13]) For predicting the protein per-position label for each amino acid input sequence it uses window size approach. These architecture mainly have two drawback because of windowing approaches:

1. These type of method need more time for learning to train as well as to test.
2. These type of method use small window size.

In order to overcome these window problems we use Convolutional Neural Network which mainly labels the amino acid entire sequences. In addition to sharing, pooling which reduce the computation so as compare to MLP approach CNN is more beneficial. By using CNN, one issue is to label each position in an input sequence with CNN that mainly pool the resolution.

These issue is being resolve Zeming et .al2016[11] have propose MUST-CNN approach for training end-to-end and per-position architecture for predicting the protein sequence on CullPDB and CB513 PDB databank having 68% of Q8 accuracy.

James Lyons et al.[14] in their work that is basically based on Prediction of protein backbone inter-residue angles using Deep Neural Network. This backbone of protein that provide the information for Ab-inito model of protein prediction. There are mainly two inter- residue angles that's used for representing structural backbone of protein having three consecutive C-α atoms(θ) and a dihedral angle among neighboring C-α bond(τ). Angles reflects the local confirmative backbone over three-four neighboring residues, while the torsion angle above NCα bond(ϕ) and Cα-C bond (ψ) is limited for single residue for protein. Secondary Structure that involve more than three residues θ and τ angles that are provided the local information of the structure complementary to ϕ and τ angles of secondary structure of protein sequences

Zhiyong et al. & Jinbo et al.[15] present a method for predicting the 8class secondary structure prediction using Conditional Neural Fields(CNF). These model captures a non-linear relationship among protein features ad secondary structure, but these interdependency for secondary adjacent residue of proteins. CNF model define the probability distribution using confirmative sampling over local structure of protein sequences. This method not only model such complex relationship between these sequence features among the protein, but also exploit the interdependency among these secondary structure of protein residues. These model give 64.9% Q8 accuracy of CB513 PDB Databank.

Di Lena et al [16,17] in their work they have use these Deep Learning Architectures. Introducing a Deep Spatio-Temporal Neural Network (DST-NN) architecture that mainly utilizes temporal as well as spatial features for predicting the protein residueresidue contact mapping. These mainly contain many multiple levels of Neural Network that share the same configuration having single input layers and output unit for such protein contact prediction for protein residue-residue pairs of protein. They all share the weight among the network in the same level.

3. Conclusion

Thus, with the limited amount of research is being done using Deep Learning. It was concluded that by predicting the performance of such Deep Learning methods for improve the

performance with existing state-of-the-art prediction. These Deep Learning approaches mainly provide robust and reliable prediction for such protein sequences. Deep learning network is a revolutionary development that can utilize and create a more powerful predictionfor such Bioinformatics fields

References

[1] Matt Spencer, Jesse Eickholt, and Jianlin Cheng; A DeepLearning Network Approach to ab initio Protein Secondary Structure Predictio IEEE/ACM Trans Comput Biol Bioinform. 2015 Jan-Feb; 12(1): 103–112. Published online2014 Aug 7. doi: 10.1109/TCBB.2014.2343960.

[2] Floudas C, Fung H, McAllister S, Mönnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: A review. Chemical EngineeringScience.2006;61:966–988.

[3] Kopp J, Schwede T. Automated protein structure homologymodeling: a progress report. Pharmacogenomics. 2004;5:405–416

[4] Hinton GE, Osindero S, Teh Y-W. A fast learning algorithmfor deep belief nets. Neural computation. 2006;18:1527–1554.

[5] Arel I et al. Deep Machine Learning- A New Frontier in Artificial Intelligence Research. IEEE Computational Intelligence 2010; 13-18

[6] J. Schmidhuber., "My First Deep Learning System of 1991 +Deep Learning Timeline 1962–2013.".

[7] Song, H.A.; Lee, S. Y. (2013). "Hierarchical Representation Using NMF". Neural Information Processing. Lectures Notes in Computer Sciences 8226. Springer Berlin Heidelberg. pp. 466–473.doi:10.1007/978-3-642-42054- 2_58. ISBN 978-3-642-42053-5.Cadence, "Encounter user guide," Version 6.2.4, March 2008.

[8] J. Schmidhuber., "Learning complex, extended sequences using the principle of history compression," Neural Computation, 4, pp. 234–242, 1992.

[9] Busia, Akosua & Jaitly, Navdeep "Next-Step Conditioned Deep Convolutional Neural Networks Improve Protein Secondary Structure Prediction" arXiv:1702.03865 2017 p1-11.

[10] Jian Zhou and Olga G. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. Journal of Machine Learning Research: W&CP, 32(1):754-762, 2014.

[11] Zeming Lin, Jack Lanchantin, and Yanjun Qi. 2016. MUST-CNN: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press 27-34.

[12] Qi, Y.; Oja, M.;Weston, J.; and Noble,W. S. 2012. A unified multitask architecture for predicting local protein properties.PloS one7(3):e 3235.

[13] Drozdetskiy, A.; Cole, C.; Procter, J.; and Barton, G. J. 2015.JPred4: a protein secondary structure prediction server. Nucleic Acids Research gkv332.
] James Lyons, Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, Kuldip Paliwal, Abdul Sattar, Yaoqi Zhou, and Yuedong Yang. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse autoencoder deep neural network. Journal of Computational Chemistry, 35(28):2040-2046, 2014.

[14] Wang Z, Zhao F, Peng J, Xu J "Protein 8-class secondary structure prediction using conditional neural fields" Proteomics. 2011 Oct;11(19):3786-92. doi: 10.1002/pmic.201100196. Epub 2011 Aug 31.

[15] Pietro Di Lena, Ken Nagata, and Pierre Baldi. Deepspatiotemporal architectures and learning for protein structure prediction. Advances in Neural Information Processing Systems (NIPS) 25, pages 521-529, Lake Tahoe, Nevada, December 3 – 6, 2012.

[16] Pietro Di Lena, Ken Nagata, and Pierre Baldi. Deep architectures for protein contact map prediction. Bioinformatics, 28(19):2449-2457, 2012.