ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Protein Structure and Function Prediction Using Machine Learning Methods – A Review

Hemalatha N.[1], Siddhant Naik[2], Jeason Rinton Saldanha[3]

Assistant Professor, Department of MCA, St. Aloysius College, (AIMIT), Mangalore, India[1]

II M Sc. (Software Technology), St. Aloysius College, (AIMIT), Mangalore, India[2]

II M Sc. (Software Technology), St. Aloysius College, (AIMIT), Mangalore, India[3]

ABSTRACT: Machine learning is a subfield ofcomputer science that includes the study of systems that can learn from data, rather than follow only explicitly programmed instructions. Some of the most common techniques used for machine learning are Support Vector Machine, Artificial Neural Networks, K Nearest Neighbor and Decision Tree. Machine learning techniques are widely used techniques in bioinformatics to solve different type of problems. Protein structure prediction is one of the problems that can be solved using machine learning. The molecules which are important in our cells are Proteins. They are virtually involved in all cell functions. Proteins are categorized on the basis of the occurrence of conserved amino acid patterns which is the feature extraction method. In the post-genomic era Protein function prediction is an important problem. Advancements in the experimental biology have enabled the production of enormous amount of protein-protein interaction data. Thus, to functionally annotate proteins has been extensively studied using protein-protein interaction data. When annotation and interaction information is inadequate in the networks most of the existing network based approaches do not work well. In this paper an attempt has been made to review different papers on proteins functions and structures that are predicted using the various machine learning methods.

I. INTRODUCTION

Proteins represent the most important class of biomolecules in living organisms. They carry out majority of the cellular processes and act as structural constituents, catalysis agents, signaling molecules and molecular machines of every biological system. In all cell functions proteins are virtually involved. Every single protein has specific function within the body. Some of the few proteins are involved in bodily movement, while others are involved in structural support. Proteins differ in functions as well as structures.One of the important goals pursued by bioinformatics and theoretical chemistry is protein structure prediction. It is highly important inbiotechnologyandmedicine.

Proteins are classified according to structural and sequence similarity. The four different levels of protein structure are primary, secondary, tertiary, and quaternary structure. A single protein molecule may contain few of these protein structure types.The structure of protein determines the protein function. The primary structure of a protein is derived from the amino acid sequence of a protein and it is the mostfundamental form of information available about the protein. It plays the most critical role in determiningvarious characteristics of the protein such as its sub-cellular localization, structure and function. Becauseof this, amino acid sequence has tremendous potential to be used extensively for functional annotation ofproteins.

Machine learning focuses on prediction, based on known properties learned from the training data. In the field of biology various application extensively uses methods which are based on machine learning algorithms. These methods have been utilized in diversedomains like genomics, proteomics and systems biology. Specifically, supervisedmachine learning approaches have found immense importance in numerous bioinformatics predictionmethods. In this paper we have put different sections were we have explained how machine learning can be applied to protein structure and function predictions.

II.LITERATURE SURVEY

2.1 Structure Prediction

Hui *et-al.*in their paperhave discussed a computational predictor using a support vector machine which is trained with PDZ domain structure and informationabout the peptide sequence[1][2][4]. They have presented a structure-based predictor of PDZ domain-peptide interactions which can be used to scan C-terminal proteomes to predict PDZ domain mediated protein-protein interactions[3]. To estimate the generality of the predictor, they conducted multiple cross validation tests and summarized the performance. Domain structure features are utilized by the predictor that is derived from the whole domain, which focuses on a core peptide-binding site defined by the highly conserved amino acid positions. One of the important technical results of their work is the use of computationally generated negatives to reduce over-prediction and to supplement training. They showed that the negative interactions in current experimental data sets do not sufficiently cover the negative proteome space resulting in a predictor that returns many hits that are false positives. High cross validation results are achieved by the predictor and findsseveral interactions corresponding to the PDZ mediated protein-protein interactions that are not previously found by their sequence-based predictor. With the help of the predictors they defined a functional map of PDZ domain biology and identified novel PDZ interactors that are involved in different biological processes. Therefore, the predictions will help spread out the coverage of current PDZ mediated protein-protein interaction networks and provide new view into the molecular mechanisms.

Lena*et-al.* in their paper have introduced a novel deep machine-learning architecture for contact prediction which consists of a multi-dimensional stack of learning modules[11][12]. The stack architecture organizes the prediction in such a way that each level in the stack can receive input, through the temporal feature vectors, and refine the predictions produced by the previous stages in the stack. Here they investigate the learning and generalization capabilities of the DST-NN model, and compared it with plain three-layer Neural Network models, as well as 2D Recurrent Neural Network models, which are two of the most widely used machine learning approaches for contact prediction[5][6][7][8]. Here, the Neural Network model is perfectly equivalent to the NNs implemented in the DST-NN architecture, except for the temporal feature vector (which is missing in the NN implementation). In order to take into account the intrinsic incomparable capabilities of the different DST-NN, NN, and RNN architectures, they have performed the tests by considering a range of exponentially increasing hidden layer sizes (4,8,16,32,64, and 128 units) for each architecture. The proposed architecture is somewhat general and it can be adopted as a starting point for more sophisticate methods for contact prediction or other problems. For instance, while the elementary learning modules of the architecture are implemented using neural networks, it is clear that these could be replaced by other models, such as SVMs. Moreover, here they have considered a simple square neighborhood for encoding the contact predictions in the temporal feature vector more complex relationships could be discovered by exploiting different topologies for such feature vector. While they have used the true contact map as the target for all the levels in the architecture, it is clear that different targets could be used at different levels [9]. DST-NNs of the form NN^{l}_{ijk}, with three spatial and one temporal coordinate, could be applied, for instance, to problems in weather forecasting or trajectory prediction in robot movements[10].

Selbig *et-al.*have presented an approach that reveals the systematic differences in the output of different prediction methods of secondary structurethat allows the derivation of coherent consensus predictions[13]. They built the decision trees from existing data using a machine learning technique. Their method for consensus secondary structure prediction Consensus formation by Decision tree learning is based on machine learning and will be integrated into the Toolbox for Protein alignment system. It relies on applying various secondary structure prediction methods to a training set with given native secondary structure. Cross validation tests are done using the CB396 dataset and a set of 11 CASP3 targets which improves the prediction accuracy in most cases. They have only used the default parameters of the decision tree learning system. By calibrating these parameters to the learning set the accuracy of the learned predictions may be further improved.

Zhou *et-al*. Troyanskaya in their paper presented a new supervised generative stochastic network based method to predict local secondary structure with deep hierarchical representations[14]. Generative stochastic network is a recently proposed deep learning technique to globally train deep generative model. They presented the supervised extension of GSN, that learns a Markov chain to sample from a conditional distribution, and applied it to protein structure prediction. To scale the model to full-sized, high-dimensional data, like protein sequences with hundreds of amino acids, they introduced an architecture, which allows efficient learning across multiple layers of hierarchical representations. This architecture uniquely focuses on predicting structured low-level labels informed with representations learned by the model. It corresponds to labeling the secondary structure state of each amino-acid residue. The model has been trained and tested on unique sets of non-homologous proteins.Their experiments determine supervised generative stochastic network to be an effective algorithm for structured prediction, extending the success of generative stochastic network in capturing complex dependencies in the data. Their model is well suited for low-level structured prediction that is sensitive to local information, while being informed of high-level and distant features. The limitation of their architecture is that the convolutional structure is hard-coded, thus in some cases it may not be peerless to capture the spatial organization of protein sequence, especially for structures formed by long-range interactions.

Titov*et-al.* in their work have proposed a new class of graphical models for structure prediction problems, Incremental Sigmoid Belief Networks where the structural model is a function of the output structure[15]. Based on mean field methods two efficient models are derived, which prove productive in artificial experiments. They demonstrated their effectiveness on a natural language parsing task, where they achieve state-of-the-art accuracy. Also, the model which is a closer approximation to an Incremental Sigmoid Belief Network has better parsing accuracy, suggesting that Incremental Sigmoid Belief Networks are an appropriate abstract model of structure prediction tasks. Exact inference with the proposed class of graphical models is not tractable, but we derive two tractable approximations. First, it is shown that the feed-forward neural network of can be considered as a simple approximation to ISBNs. Second, a more accurate but still tractable approximation based on mean field theory is proposed.

2.2 Function Prediction

Qingyao *et-al.*haveproposed an effective Markov chain based Collective Classification algorithm to solve the label deficiency problem in Collective Classification for interrelated proteins from protein-protein interaction networks. The algorithm focuses on how to use unlabeled and labeled data to intensify the classification performance of protein-protein interaction network data. They aimed at modeling the problem using two distinct Markov chain classifiers to make separate predictions with regard to relational features from relational information and attribute features from protein data[16]. The algorithm combines the results of both the classifiers to compute the ranks of labels to indicate the importance of a set of labels to an instance which uses an ICA framework to repeatedly refine the learning models for ameliorate the performance of protein function prediction from protein-protein interaction networks in the insufficiency of labeled data. Protein-Protein Interaction datasets show that the ICAM method which they had proposed is better than the other ICA-type methods given limited labeled training data. This approach can assist as a valuable tool for the study of protein function prediction from protein-protein interaction networks. In future, they will be considering other semi-supervised learning techniques for collective classification in protein-protein interaction network data.

Huang*et-al.*in their paper proposed a novel scoring card method to estimate solubility scores of dipeptides and amino acid residues for predicting solubility of proteins and analyzing the tendency of physicochemical properties[17]. The proposed method scoring card method with solubility scores and dipeptide propensities can be easily applied to the protein function prediction problems that dipeptide composition features play an important role. The scoring card method with solubility scoring matrix performs well in predicting solubility, compared with existing methods using complementary features thatassociate well with solubility. The results approving with the literature reports reveal that the solubility scoring matrices are effective. Since the proposed scoring card method is effective for generating solubility scoring matrix to predict protein solubility, their future work is to apply scoring card method to generate various kinds of scoring matrices of dipeptides for investigating protein function prediction problems where the features of amino acid and dipeptide composition play an important role.

Liang *et-al.*havefocused on the question of how to integrate various data sources to enhance the prediction accuracy[18]. They discussed and evaluated several different integration schemes in their paper. Their pre-CAFA and CAFA results strongly indicate that integrating information from various data sources could enhance protein function prediction accuracy. At the level of sequence similarity-based predictions, they observed that it is beneficial to consider all available annotated proteins, regardless how evolutionary distant they are from a query protein. They used the simple and efficient k-Nearest Neighbor algorithm, coupled with simple integration of prediction scores from various data sources. According to the authors further enhancements that could be done are to include as many available sources of functional and structural protein information. They had used only microarray data from a single, human microarray platform. Information beyond microarray data and protein-protein interaction data, such as chromosomal neighborhood of a gene, mutations, role in various diseases, or protein structure, could certainly be valuable.

Xiong *et-al.*in their paper came up with a new method which combines protein-protein interaction information and protein sequence information to boost the prediction performance based on collective classification[19]. Their method divides function prediction into two phases: First, the original protein-protein interaction network is enriched by adding a number of edges that are inferred from protein sequence information. The added edges are called as implicit edges, and the existing ones are called explicit edges. Second, a collective classification algorithm is employed on the new network to predict protein function. Their key idea was to enrich the protein-protein interactioninformation of PPI networks by adding a number of computed edges, which subsequently improves the prediction performance. They conducted extensive experiments on two real, publicly available protein-protein interaction datasets. Compared to four existing protein function prediction approaches, their method performed better in many situations, which shows that adding implicit edges can indeed improve the prediction performance. The experimental results demonstrate that their method outperforms the existing approaches across a series of label situations, especially in sparsely-labeled networks where the existing approaches do not work well due to protein-protein interaction information Inadequacy. Experimental results also validate the robustness of the proposed approach to the number of labeled proteins in protein-protein interaction networks.

2.3 Structure and Function Prediction

Gültas*et-al.*in their work have reported a new method, QCMF, applying principles of quantum information theory[20]. In contrast to the previous method CMF which focused on dissimilar amino acid signals, QCMF simultaneously models similar and dissimilar amino acid pair signals in the detection of functionally or structurally important sites. The result of this study is twofold. First, using the essential sites of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK), they tested the QCMF-method. The QCMF includes two metrics based on quantum Jensen-Shannon divergence to measure both sequence conservation and compensatory mutations. They found that the QCMF reaches an improved performance in identifying essential sites from multiple sequence alignments of both proteins with a significantly higher Matthews's correlation coefficient value in comparison to previous methods. Second, using a data set of 153 proteins, they made a pairwise comparison between QCMF and three conventional methods. To ensure a feasible computation time of the QCMF's algorithm, they leveraged Compute Unified Device Architecture. This comparison study strongly suggests that QCMF complements the conventional methods for the identification of correlated mutations in multiple sequence alignments.

III.CONCLUSION

As has been discussed in the previous sections, machine learning methods have been used extensively in the field of protein function and structure prediction and have significantly contributed in the transformation of huge volume of data into useful knowledge. An attempt has been made in this review paper to provide a glimpse of the vast and ever-expanding realm of machine learning based methods in the area of Bioinformatics and Computational Biology. Distinction of machine learning methods lies in the fact that they do not require explicit knowledge of homology for the purpose of function and structure prediction.

REFERENCES

[1] G. I. V. V. Boser B, "A training algorithm for optimal margin classifiers. In Fifth Annual Workshop on Computational Learning Theory,"

Pittsburgh: ACM Press, 1992.

[2] S.-T. J. Cristianini N, " An introduction to support vector machines and other kernel-based learning methods.," *Cambridge New York: Cambridge University Press.*

[3] A. B. A. S. B. B. B. K. B. E. B. S. C. Y. C. P. C. L. e. a. Hubbard TJ, *Ensembl Nucleic Acids Res* , 2009.

[4] S. X. X. a. G. D. B. Hui, "Predicting PDZ domain mediated protein interactions from structure," *BMC bioinformatics,* 2013.

[5] P. P. G. Baldi, "The Principled Design of Large-Scale Recursive Neural Network Architectures-DAG-RNNs and the Protein Structure Prediction Problem," *Journal of Machine Learning Research* , (2003).

[6] O. O. A. C. R. Fariselli P., "Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins," 2001.

[7] M. R. B. Punta, "PROFcon: novel prediction of long-range contacts," *Bioinformatics,* 2005.

[8] G. K. K. Shackelford, "Contact prediction using mutual information and neural nets.Proteins," 2007.

[9] P. Baldi, "Boolean Auto encoders and Hypercube Clustering Complexity, Designs, Codes, and Cryptography.," 2012.

[10] W. Hsieh, "Machine Learning Methods in the Environmental Sciences," *Neural Networks and Kernels. Cambridge University Press, NY, USA.,* 2009.

[11] N. T. Jetchev, "Proceedings of the 26th Annual International Conference on Machine Learning.," *Trajectory prediction: learning to map situations to robot trajectories.,* 2009.

[12] P. D. K. N. a. P. F. B. Lena, "Deep spatio-temporal architectures and learning for protein structure prediction," *Advances in Neural Information Processing Systems.,* 2012.

[13] J. T. M. a. T. L. Selbig, "Decision tree-based formation of consensus protein secondary structure prediction," *Bioinformatics,* 1999.

[14] J. a. O. G. T. Zhou, "Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction," *ArXiv preprint arXiv* , 2014.

[15] I. a. J. H. Titov, "Incremental bayesian networks for structure prediction," *Proceedings of the 24th international conference on Machine learning. ACM,* 2007.

[16] Q. e. a. Wu, "Collective prediction of protein functions from protein-protein interaction networks.," *BMC bioinformatics,* 2014.

[17] H.-L. e. a. Huang, "Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition.," *BMC bioinformatics,* 2012.

[18] e. a. Lan Liang, "MS-kNN: protein function prediction by integrating multiple data sources," *BMC bioinformatics,* 2013.

[19] W. e. a. Xiong, "Protein function prediction by collective classification with implicit and explicit edges in protein-protein interaction networks.," *BMC bioinformatics,* 2013.

[20] M. e. a. Gültas, "Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming.," *BMC bioinformatics,* 2014.