Deep Learning for Prediction of Protein-Protein Interaction

Mohamed Abd Allah Makhlouf
Department of Information Systems Faculty of Computers and Informatics
Suez Canal University, Ismailia, Egypt
m.abdallah@ci.suez.edu.eg

Abstract

Various computational methods are being used to predict Protein-protein interactions (PPI) from many different perspectives in solving different problems which are critical for many biological objectives. One of the important problems developing high-accuracy techniques for identifying PPI to better understand proteins functions, diseases, and therapy design. Deep-learning algorithms have achieved effective results in numerous areas, but their leverage for PPI prediction has not enough. We proposed a hybrid model of deep-learning algorithm and random forest, to study the PPI prediction. The proposed model achieved an average accuracy of 90.04% with Roc area 0.788and Matthew Correlation Coefficient of 0.477 Benchmarking, which are promise to those achieved with previous methods. We think this research is from the first to apply hybrid model of deep-learning and random forest to PPI prediction, and the results demonstrate its potential in this field.

Keywords: *Deep learning, Protein-protein interaction, Random forest, Machine learning*

1. Introduction

PPIs play major roles in many biological processes, such as immune response, cellular organization and signal transduction. Analysis of PPI is a great importance and may focus on drug target detection and aid in therapy [2]. Small-scale experimental methods like chromatography and Biochemical assays have long been used to identify PPIs, but its contribution is low coverage of the huge PPI database due to their low efficacies [1]. High technologies, such as mass spectrometric protein complex identification [5] and yeast two-hybrid screens [4] have generated voluminous data, but, they are expensive and time consuming. Also, these methods may not be applicable to all organisms and most often produce false-positive results [6]. Therefore, high computational techniques are needed to identify PPIs with high quality and accuracy. Newly, many computational techniques have been adopted to solve this problem. Some of these, have attempted to extract new protein information, whereas the others develop a new machine e learning algorithms. In Shen et.al [7] a protein information mining any three continuous amino acids as a unit then calculate the frequencies of those conjoint triads in the protein sequences. Presented in [7] that PPIs could be predicted by sequences alone. Other methods and techniques, such as amino acid index distribution [9] and auto-covariance [8] were developed to extract attributes such as physical chemical, frequencies, and locations of amino acids to represent a protein sequence. Machine-learning algorithms such as support vector machine and its derivatives[10, 11], neural networks [13] and random forest [12], have been applied. However, most studies provided only the results of cross-validation, and did not test prediction results [7, 11, 14, 15].Deep-learning mimic the deep neural connections and learning processes of the human brain, have received considerable attention due to their successful applications in image and speech recognition [16, 17], decision making [19] and natural language understanding [18] .Deep-learning algorithms can handle large data and automatically learn useful and more abstract features [20]. Recently, Deep-learning algorithms have been applied to bioinformatics due to increasing amounts and dimensions of data generated by computational biology [21–24]. Sun, Tanlin, et al.[42] was the first one applied the stacked auto encoder algorithm, to study the sequence-based PPI ,using the prediction for various external datasets.

Xiong et al [25] applied a deep neural network model to predict DNA variants causing aberrant splicing. Their method was more accurate than traditional models. Alipanahi constructed a Deep Bind model using convolutional networks to predict sequence specificities of DNA- and RNA-binding proteins, and identify binding motifs [26]. Identifying the effects of noncoding variants is also a major challenge in genetics. Zhou et al. developed a Deep SEA to learn a regulatory sequence code from large-scale profiling data, predicting a chromatin effects of sequence alterations [27]. Quang and coworkers constructed the DnaQ model achieving more than a 50% improvement compared to other models for predicting the function of non-coding DNA [28]. Spencer et al.[29] exploited a deep belief network (DBN) for protein function prediction to predict protein secondary structures and they reach an accuracy of 80.7% . Then Sheng et.al. [30] Increased the prediction accuracy to 84% using deep convolutional neural. Heffernan et al.[31] predict secondary structures and also predict backbone angles and solvent accessible surface areas. For more detailed of the applications of the deep learning algorithms in computational biology can be found in the review [32].Other machine learning techniques such as random forests and support vector machines have been used to predict interactions from protein sequences (Ben-Hur et al. [46] ; Bock et al. [47] ; Chan et al. [48] et al. [49]. Park et al. [50]establish a comparative study of sequence-based prediction identified three top methods: PIPE2 [51], Sig Prod et al[52], and Auto Correlation Guo et al.[53]. M Alfonse et al. [70] presented a brain tumor diagnostic system. The system classify the type of the tumor which is benign or malignant using support vector machine. H. Mohsen et al. [71] proposed a classification model for Alzheimer's disease based on discrete wavelet transform feature extraction technique and PCA for feature vector selection then features are entered to linear discriminant analysis (LDA) classifier.

In this study, we applied deep learning and random forest hybrid model to study sequence-based PPI predictions. Models based on protein sequence achieved the best results on 10-fold cross-validation .The best model had an average accuracy of 90.04% with Roc area 0.788 and Matthew Correlation Coefficient of 0.477 Benchmarking for the whole training benchmark dataset the results were promising and achieved prediction performance that surpassed previous methods.

2. Materials and methods

2.1 Datasets

The dataset in this study adopted from [3] to train-test the proposed model. The dataset contains 151 protein complexes whose key feature is the availability of both bound and unbound structures of the interacting proteins, for67235 number of instances, Savojardo et al [3] focus on the subset of protein complexes that met the following criteria:

- Both bound and identical unbound structures obtained via X-Ray crystallography.
- Interfaces estimated from the bound structure that successfully mapped to unbound structures.

2.2 Features description

The feature descriptors were adopted to perform the classification task. The complete feature set consists of 5 different groups of descriptors encoded in a 34-dimensional real vector for each input residue. Table 1presents a set of the different descriptor sets used in this research adopted form Savojardo et al [3].

Table 1. The descriptors adopted in this study to encode surface residues

Descriptor	Features	Position
Sequence profile	20	1-20
propensity score	1	21
Conservation score	1	22
Residue co-evolution scores	2	23-24
Residue physical-chemical properties	10	25-34

2.2.1 Evolutionary information

Evolutionary information for each position of the protein sequence was extracted in the form of a sequence profile. For a given protein sequence, the BLAST Altschul et al [43] program was exploited to search the Uniprot database for similar sequences and the conforming profile was extracted as additional BLAST output. Then, for each surface residue i, a vector $\mathrm{v_i}$ was computed by averaging sequence profile entries over the surface structural context of the residue i, i.e.

$$\mathbf{v_i} = \frac{\mathbf{1}}{|\mathbf{C(i)}|}\sum\nolimits_{\mathbf{k}\in\mathbf{C(i)}} \mathbf{P_k} \qquad (1)$$

2.2.2 Residue interface propensity

The following log-ratio formula used to score the propensity $\mathrm{p_k}$ of each residue type to be in interaction sites:

$$\boldsymbol{p}_{\boldsymbol{k}=\,\mathbf{log}\frac{f_I\,(k)}{f_S\,(k)}} \qquad (2)$$

Wheref_I $(k)$is the frequency of residue of type k in interaction sitesand f_I $(k)$is the frequency of residue type k in the surface. For each cross validation iteration, propensities and frequencies scores be computed on the training stand kept fixed when encoding the testing set.

2.2.3 Residue conservation

Using the sequence profile obtained from BLAST, a conservation score $\mathrm{c_i}$ calculated for each surface residue position i:

$$\boldsymbol{c}_{\boldsymbol{t}=\,-\frac{1}{\log K}\sum_{j=1}^{K} pij\,\times\,\log P_{ij}} \qquad (3)$$

Where K=20 and $\mathrm{P_{ij}}$ is the frequency of residue type j at position i.

2.2.4 Residue co-evolution scores

There are a lot of methods to extract residue co-evolutionary indexes starting from a MSA. In Savojardo et al. [3] adopted sparse inverse covariance estimation as in the Jones et al., [44] as well as Mutual Information.

2.2.5 Residue physical -chemical properties

Kidera et al. [45] were introduced 10 orthogonal properties used to represent the physical-chemical nature of each residue. The 10 properties were derived with statistical analysis of a range of 188 different physical properties of naturally occurring amino acids. Each residue was represented in the surface according to its type with a 10-dimensional vector.

2.3 The DL-RF prediction Model

In this work, we adopt Deep learning and Random forest to study PPI or not. The model was trained on protein chains in the dataset whose surface residues were represented with the 34-dimensional feature vectors described in Table1in the previous section. The model was trained and tested using the stacking hybrid implementation as shown in figure 1.

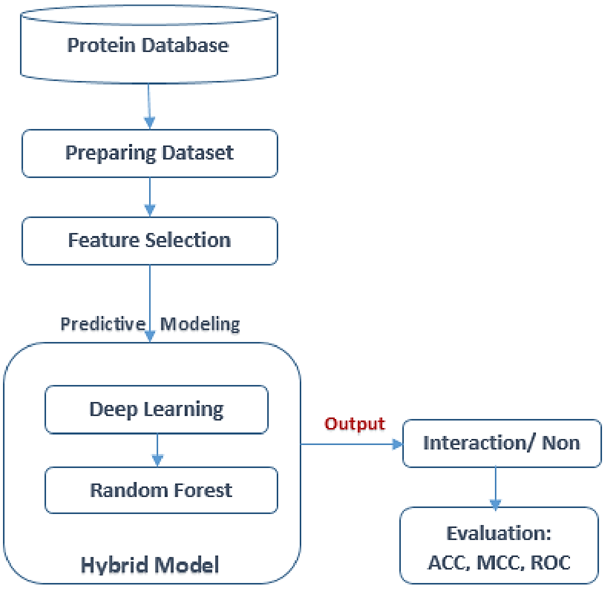


Figure 1. The Proposed Hybrid model

2.3.1 Random Forests

Random Forests developed by Leo Breiman [65] is a method that joins several individual classification trees that operates by constructing a multitude of decision trees at the training time and yielding the final class that is the majority vote of the classes output by individual trees. These trees are created by bootstrap tests of the preparation information and by utilizing arbitrary component choice in the tree age process. It is a discrete classifier, when applied to a test set, it produce a single confusion matrix, which corresponds to a single point on a ROC curve. It is often used in a large training datasets and a very large number of attributes figure 2 show the algorithm steps [65].

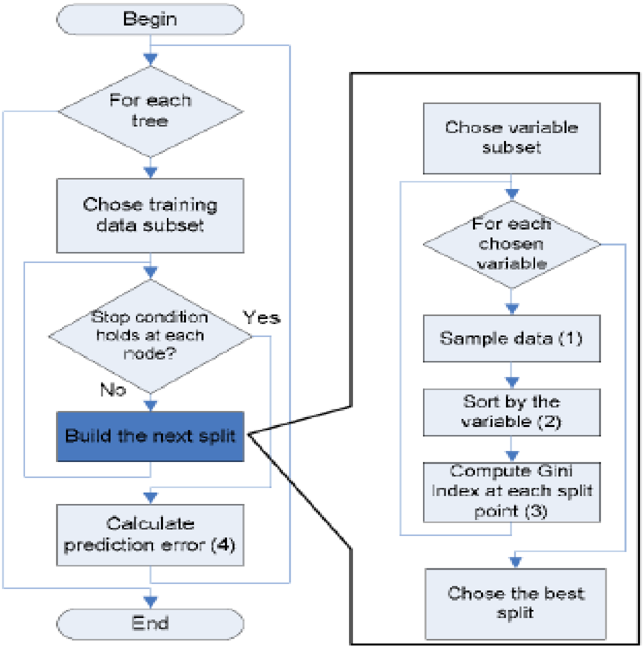


Figure2. Random forest algorithm steps

2.3.2 Deep learning

Deep learning has arisen as a new field of machine learning since 2006 [66] ,[67], [68]. It is a set of machine learning algorithms which aim to learn multiple layered models of inputs, commonly neural networks. The deep neural networks are consists of multiple levels of operations as in figure 3.Deep learning have recently led to progress in classification in various applications in biological data, speech and natural language processing, and computer vision. Using the stochastic Gradient descent (SGD) updater to optimize gradient descent and minimizes the loss function during training. The stochastic of a learning is a form of search. The results of that search are recorded in the form of a weight adjustment, which reduce the search space move toward a position of less error. SGD is used with mini-batches, where parameters are updated based on the average error generated by the instances of a completed batch.

$$\theta_{t+1} = \theta_t - a\delta L(\theta_t) \qquad (4)$$

θ is the weights change according to the gradient of the loss with respect to each theta. **α**is the learning rate. If alpha is very small, convergence on an error minimum will be slow. If it is very large, the model will diverge away from the error minimum, and learning will cease. After each iteration the gradient of the loss (L) changes quickly due to variance among training examples.

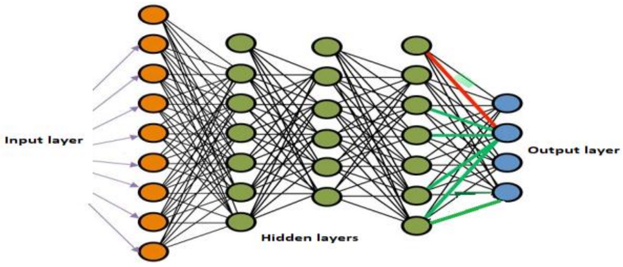


Figure 3.Deep learning architecture

2.4 Evaluation of the Prediction Performance

Standard scoring measures were used to score the method at the level of residue classification .In what follows, True positive (TP), true negative (TN), false positive (FP) and false negative (FN) are calculated for each fold. TP are the actual binding interfaces residues that are predicted correctly. TN are the actual non-interacting residues that are predicted correctly. FP are false predictions of interacting residues. FN are false predictions of non-interacting residues. The following measures were adopted to score interface residue predictions: Recall (true positive rate) of the positive class [Recall (I)], defined as:

$$Recall(I) = \frac{TP}{TP + FN} \qquad (5)$$

Precision of the positive class [Precision(I)], defined as:

$$\text{Precision } (I) = \frac{TP}{TP + FP} \qquad (6)$$

The F1-score of the positive class [F1(I)], defined as:

$$F1(I) = \frac{2\times Recall(I) \times Precision(I)}{Recall(I) + Precision(I)} \qquad (7)$$

The classification accuracy [ACC], defined as:

$$ACC \quad = \frac{TP+TN}{TP + TN+FP+FN} \qquad (8)$$

The Matthews Correlation Coefficient [MCC], defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)\times(TP+FN)\times(FN+FP) \times(TN+FN)}} \qquad (9)$$

3. Results

The classification power of the different algorithms used in this study is evaluated by training and testing our method on all folds. Table 2 shows detailed results of accuracy, MCC, Roc area, precision, recall and the mean through all the folds. This measures were employed in the state of the art using Naive Bayes.

Table 2. Detailed results for the Naïve Bayes cross-validation

Fold	ACC	MCC	ROC Area	Precision	Recall
Fold0	79.7205	0.120	0.661	0.252	0.222
Fold1	79.7725	0.085	0.563	0.224	0.177
Fold2	74.9915	0.064	0.570	0.240	0.185
Fold3	80.1866	0.132	0.624	0.259	0.233
Fold4	81.3192	0.186	0.684	0.304	0.283
Fold5	82.2309	0.079	0.596	0.213	0.145
Fold6	81.5935	0.172	0.672	0.260	0.296
Fold7	81.4027	0.135	0.642	0.251	0.230
Fold8	85.5409	0.136	0.638	0.224	0.208
Fold9	83.7996	0.121	0.645	0.257	0.169
Average	**81.05579**	**0.123**	**0.6295**	**0.2484**	**0.2148**

Table 3 shows the accuracy, MCC, Roc area, precision, recall and the mean in each fold of neural network cross validation. As can be seen, the results are consistent in all folds. Also, the system is more specific than sensible. Although, the deviations are low for all the folds with a performance balanced around 87%.

Table 3. Detailed results for the neural network cross-validation

Fold	ACC	MCC	ROC Area	Precision	Recall
Fold0	86.9313%	0.314	0.697	0.595	0.231
Fold1	86.2957%	0.253	0.655	0.542	0.177
Fold2	82.812%	0.245	0.693	0.557	0.181
Fold3	87.1416%	0.283	0.686	0.626	0.174
Fold4	88.259%	0.384	0.704	0.663	0.287
Fold5	87.1777%	0.171	0.575	0.491	0.092
Fold6	89.0819%	0.325	0.699	0.612	0.224
Fold7	88.6373%	0.326	0.629	0.707	0.188
Fold8	88.4446%	0.262	0.693	0.367	0.286
Fold9	88.1285%	0.239	0.624	0.565	0.140
Average	**87.291**	**0.2802**	**0.6655**	**0.5725**	**0.198**

Finally, Table 4 presents the results of proposed stacking method between deep learning and random forest and without optimizing the parameters nor applying feature selection. This results show that the proposed method improve the results significantly the performance balanced around 90 %, ROC area 0.79 and MCC 0.48.

Table 4. Detailed results for the DLRF cross-validation

Fold	ACC	MCC	ROC Area	Precision	Recall
Fold0	89.875	0.508	0.832	0.779	0.394
Fold1	88.584	0.51	0.821	0.605	0.549
Fold2	90.088	0.63	0.864	0.827	0.563
Fold3	88.431	0.403	0.734	0.686	0.301
Fold4	91.056	0.569	0.843	0.776	0.486
Fold5	89.476	0.408	0.731	0.736	0.275
Fold6	91.583	0.534	0.835	0.747	0.442
Fold7	90.032	0.442	0.728	0.854	0.265
Fold8	91.765	0.404	0.761	0.653	0.297
Fold9	89.546	0.363	0.734	0.82	0.19
Average	**90.0436**	**0.4771**	**0.7883**	**0.7483**	**0.3762**

3.1 Comparison with Other Method

The performance measures for the proposed model and the others previously developed methods is rather difficult owing to the different data sets (with the exception of Savojardo et al. (2017) [3]), in Table 5 we present the accuracy of DL-RF model with respect to other machine-learning approaches. It shows that DL-RF improves over the recently introduced predictors of interaction sites. This improvement is due to the fact that the input features are relevant. Results in Table 5 indicate that DL-RF scores with the highest ACC, MCC and ROC values on both testing and blind datasets. Recall (the true positive rate) is lower than that of other predictors, indicating that DL-RF labels as ‘interacting’ less residues than other methods, however, with a higher probability to be correct. So, DL-RF is talented with the highest precision and accuracy with respect to the others. Figure 4 presented a comparative diagram for the proposed model and others methods. Figure 5 represents the implementation results and the performance measure for Roc area for the proposed model.

Table 5. Comparison with several state-of-the-art methods

Method	ACC	MCC	ROC	Precision	Recall
Wang et al. (2006)	NA	0.28	NA	0.69	0.65
Nguyen-Rajapakse (2006)	NA	0.33	NA	0.36	0.93
Deng et al. (2009)	NA	0.35	NA	0.77	0.63
Liu et al. (2009)	0.69	0.33	NA	0.59	0.54
Zhang et al. (2011)	0.67	0.34	NA	0.37	0.76
Jordan et al. (2012)	0.83	0.33	NA	0.42	0.41
Porollo and Meller (2007)	0.72	0.28	NA	0.39	0.54
Li et al. (2008)	0.74	0.27	NA	0.38	0.48
Chen and Zhou (2005)	0.77	0.23	NA	0.46	0.27
Savojardo et al. (2012)	0.47	0.16	NA	0.26	0.80
Huang and Schroeder (2008)	0.79	0.09	NA	0.13	0.34
Savojardo et al. (2017)	0.84	0.48	NA	0.78	0.39
DL-RF	**0.90**	**0.48**	**0.79**	**0.75**	**0.38**

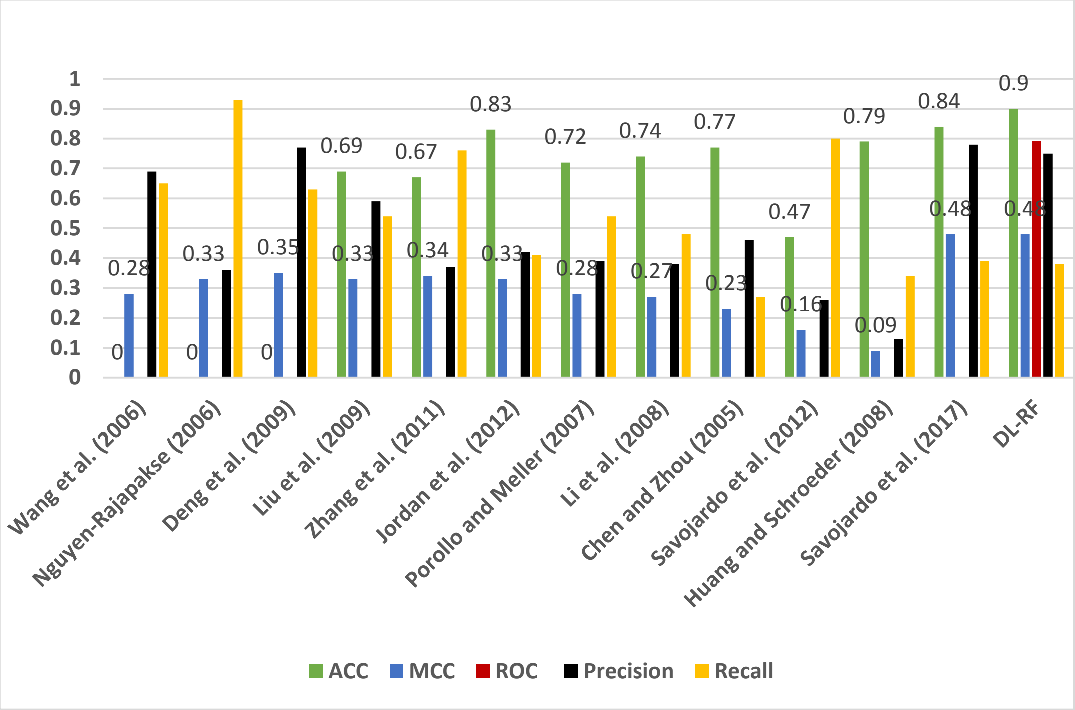
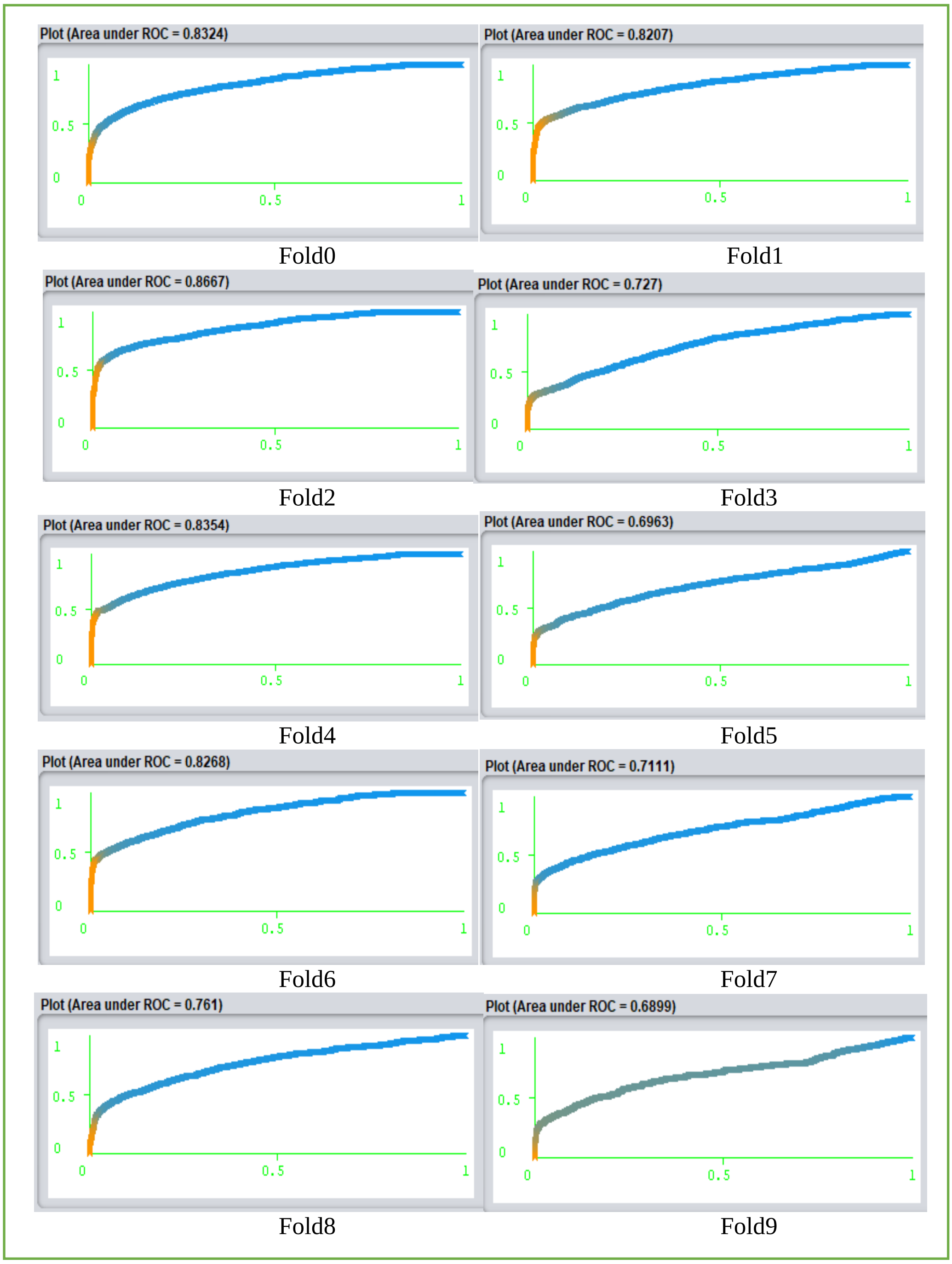


Figure 4. Comparative diagram for the proposed model



Plot (Area under ROC = 0.8324)

Fold0

Plot (Area under ROC = 0.8207)

Fold1

Plot (Area under ROC = 0.8667)

Fold2

Plot (Area under ROC = 0.727)

Fold3

Plot (Area under ROC = 0.8354)

Fold4

Plot (Area under ROC = 0.6963)

Fold5

Plot (Area under ROC = 0.8268)

Fold6

Plot (Area under ROC = 0.7111)

Fold7

Plot (Area under ROC = 0.761)

Fold8

Plot (Area under ROC = 0.6899)

Fold9

Figure 5. The implementation and performance measure for Roc area

4. Conclusion

In this paper a deep learning and random forest are used as stacking hybrid model for the prediction of protein-protein interactions. Accurate methods to identify PPI to discover protein function and identify functionally important residues on protein surfaces is crucial. In this paper, we present DL-RF, an improved predictor of PPI sites. As a classification method, DL-RF adopts a combination of Deep learning and Random forest performing a cross-validation experiments on dataset derived from the Docking Benchmark [69] and consisting of 151 high-resolution protein complexes. The obtained results demonstrates that the input features vector can provide useful information for the training set in order to enhance the quality of the classification. After DL-RF trained and tested when compared with other approaches, DL-R Fout-performs other methods and, become one of the best approaches for PPI prediction.

As future work, this model can be used with the inclusion of an optimization method like PSO and apply feature selection techniques to improve the performance of the model together with physical-chemical characterization.

Acknowledgements

The author thank Prof. Rita Casadio, Prof. Castrense Savojardo and Prof. Pier Luigi Martelli members of the Biocomputing Group, Department of Biological and Environmental Sciences (BiGeA), University of Bologna, Italy for their hospitality during the short staff mobility supported by Erasmus Mundus – EU-METALIC II.

References

[1]. Zhang SW, et al. Some Remarks on Prediction of Protein-Protein Interaction with Machine Learning. Med Chem. 2015;11(3):254–64.

[2]. Skrabanek L, et al. Computational prediction of protein–protein interactions.MolBiotechnol. 2008;38(1):1–17.

[3]. Savojardo, Castrense, et al. "ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model." *Bioinformatics* 33.11 (2017): 1656-1663

[4]. Fields S, et al. A novel genetic system to detect protein protein interactions. Nature. 1989;340(6230):245–6.

[5]. Ho Y, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002;415(6868):180–3.

[6]. Collins SR, et al. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics. 2007;6(3):439–50.

[7]. Shen JW, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl AcadSci U S A. 2007;104(11):4337–41.

[8]. Guo YZ, et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008;36(9):3025–30.

[9]. Zhang SW, et al. Prediction of protein-protein interaction with pairwise kernel Support Vector Machine. Int J Mol Sci. 2014;15(2):3220–33.

[10]. Huang YA, et al. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. Biomed Res Int. 2015;2015:902198.

[11]. You ZH, et al. A MapReduce based parallel SVM for large-scale predicting protein-protein interactions. Neurocomputing. 2014;145:37–43.

[12]. You ZH, et al. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS One. 2015;10(5):e0125811.

[13]. Zhao YO. Predicting Protein-protein Interactions from Protein Sequences Using Probabilistic Neural Network and Feature Combination. Int J InfComput Sci. 2014;11(7):2397–406.

[14]. You ZH, et al. Large-scale protein-protein interactions detection by integrating big biosensing data with computational model. Biomed Res Int. 2014;2014:598129.

[15]. Pan XY, et al. Large-Scale prediction of human protein − protein interactions from amino acid sequence based on latent topic features. J Proteome Res. 2010;9(10):4992–5001.

[16]. Hinton G, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag. 2012;29(6):82–97.

[17]. Krizhevsky A, et al. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems. 2012. p. 1097–105.

[18]. Lipton ZC, et al. A critical review of recurrent neural networks for sequence learning, arXiv preprint arXiv:150600019. 2015.

[19]. Silver D, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–9.

[20]. LeCun Y, et al. Deep learning. Nature. 2015;521(7553):436–44. 20. Kuksa PP, et al. High-order neural networks and kernel methods for peptide- MHC binding prediction. Bioinformatics. 2015;31(22):3600–7.

[21]. Li YF, et al. Genome-Wide Prediction of cis-Regulatory Regions Using Supervised Deep Learning Methods. bioRxiv. 2016;2016:041616.

[22]. Xu YJ, et al. Deep learning for drug-induced liver injury. J Chem Inf Model. 2015;55(10):2085–93.

[23]. Zeng HY, et al. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32(12):i121–7.

[24]. Zhang S, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. 2016;44(4):e32-e.

[25]. Xiong HY, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806.

[26]. Alipanahi B, et al. Predicting the sequence specificities of DNA-and RNAbinding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.

[27]. Zhou J, et al. Predicting effects of noncoding variants with deep learning based sequence model. Nat Methods. 2015;12(10):931–4.

[28]. [28] Quang D, et al. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):e107-e.

[29]. Spencer M, et al. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. IEEE/ACM Trans Comput BiolBioin form. 2015;12(1):103–12.

[30]. Sheng W, et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci Rep. 2016;6:18962.

[31]. Heffernan R, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep. 2015;5:11476.

[32]. Angermueller C, et al. Deep learning for computational biology. MolSyst Biol. 2016;12(7):878.

[33]. Smialowski P, et al. The Negatome database: a reference set of noninteracting protein pairs. Nucleic Acids Res. 2010;38suppl 1:D540–4.

[34]. You ZH, et al. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinformatics. 2013;14(8):1–11.

[35]. Guo YZ, et al. PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. BMC Res Notes. 2010;3(1):1.

[36]. [36] Martin S, et al. Predicting protein–protein interactions using signature products. Bioinformatics. 2005;21(2):218–26.

[37]. Nanni L, et al. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics. 2006;22(10):1207–10.

[38]. Nanni L. Hyperplanes for predicting protein-protein interactions. Neurocomputing. 2005;69(1):257–63.

[39]. Zhang YN, et al. Adaptive compressive learning for prediction of protein-protein interactions from primary sequence. J Theor Biol. 2011;283(1):44–52.

[40]. Yu JT, et al. Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics. 2010;26(20):2610–4.

[41]. Park Y, et al. Revisiting the negative example sampling problem for predicting protein-protein interactions. Bioinformatics. 2011;27(21):3024–8.

[42]. Sun, Tanlin, et al. "Sequence-based prediction of protein protein interaction using a deep-learning algorithm." *BMC bioinformatics* 18.1 (2017): 277.

[43]. Altschul,S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., (1997) 25, 3389–3402.

[44]. Jones,D.T. et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, (2012) 28, 184–190.

[45]. Kidera,A. et al. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem., (1985) 4, 23–55.

[46]. Ben-Hur, A., & Noble, W. S. Kernel methods for predicting protein-protein interactions. *Bioinformatics (Oxford, England)*, *21*(Suppl 1), (2005). i38–i46. doi: 10.1093/bioinformatics/bti1016.

[47]. Bock, J. R., & Gough, D. A. Predicting protein–protein interactions from primary structure. *Bioinformatics*, *17*(5), 455–460. (2001). doi: 10.1093/bioinformatics/17.5.455.

[48]. Yu, C.-Y., Chou, L.-C., & Chang, D. T.-H. Predicting protein-protein interactions in unbalanceddata using the primary structure of proteins. *BMC Bioinformatics*, *11*, 167. (2010). doi: 10.1186/1471-2105-11-167.

[49]. Z.-H., Chan, K. C. C., & Hu, P. Predicting protein-protein interactions from primary protein sequences using a novel multiscale local feature representation scheme and the random forest. *PloS One*, *10*(5), e0125811. (2015). Available at https://www.ncbi.nlm.nih.gov/pub med/25946106. doi: 10.1371/journal.pone.0125811.

[50]. Park, Y., &Marcotte, E. M. Flaws in evaluation schemes for pair-input computational predictions.*Nature Methods*, *9*(12), 1134–1136. . (2012)doi: 10.1038/nmeth.2259.

[51]. Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J. R., Golshani, A. Short Co-occurringpolypeptide regions can predict global protein interaction maps. *Scientific Reports*, *2*, 239. . (2012a) doi:10.1038/srep00239.

[52]. Martin, S., Roe, D., &Faulon, J. L.. Predicting protein-protein interactions using signature products. *Bioinformatics*, *21*, 218–226. (2005) doi:10.1093/bioinformatics/bth483.

[53]. Guo, Y., Yu, L., Wen, Z., & Li, M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research*, *36*(9), 3025–3030. (2008). doi: 10.1093/nar/gkn159.

[54]. Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M., Lyu, M.R.: Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett. 580(2), 380–384 (2006)

[55]. Nguyen, M.N., Rajapakse, J.C.: Protein-Protein Interface Residue Prediction with SVM Using Evolutionary Profiles and Accessible Surface Areas. In: CIBCB 2006, pp. 1–5 (2006)

[56]. Deng, L., Guan, J., Dong, Q., Zhou, S.: Prediction of protein-protein interaction sites using an ensemble method. BMC Bioinformatics 10, 426 (2009).

[57]. Liu, B., Wang, X., Lin, L., Tang, B., Dong, Q., Wang, X.: Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics 10, 381 (2009)

[58]. Zhang,Q.C. et alPredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res., . (2011) 39, 283–287.

[59]. Jordan,R.A. et al. Predicting protein–protein interface residues using local surface structural similarity. BMC Bioinformatics, (2012) 13, 1–14.

[60]. Porollo,A. and Meller,J. Prediction-based fingerprints of protein–protein interactions. Proteins: Struct. Funct. Genet.,(2007) 66, 630–645.

[61]. Li,N. et al. Prediction of protein–protein binding site by using core interface residue and support vector machine. BMC Bioinformatics, (2008) 9, 553.

[62]. Chen,H.L. and Zhou,H.XPrediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins, 61, (2005) 21–35.

[63]. Savojardo,C. et al. Machine-learning methods to predict protein interaction sites in folded proteins. In: Biganzoli, E. et al. (eds.) Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes inComputer Science. Vol. 7548, p. (2012) 127–135.

[64]. Huang,B. and Schroeder,M. Using binding site to improve protein–protein docking.Gene, (2008) 1-2, 14-21.

[65]. Breiman L Random forests. Mach Learn (2001) 45: 5–32.

[66]. Hinton, G. E., S. Osindero, et al. "A Fast Learning Algorithm for Deep Belief Nets." Neural Computation 18(7): (2006). 1527-1554.

[67]. Bengio, Y. (2009). "Learning Deep Architectures for AI." Found. Trends Mach. Learn. 2(1): 1-127.

[68]. Yoshua, B. "Representation Learning: A Review and New Perspectives." IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8): (2013). 1798-1828.

[69]. Vreven,T. et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. J. Mol. Biol. (2015), 427, 3031–3041.

[70]. Alfonse, Marco, and Abdel-Badeeh M. Salem. "An automatic classification of brain tumors through MRI using support vector machine." *Egyptian Computer Science Journal* (2016).

[71]. Mohsen, Heba, et al. "Classification of Brain MRI for Alzheimer's Disease Based on Linear Discriminate Analysis." *Egyptian Computer Science Journal* 41.3 (2017).