Use of Machine Learning Features to Detect Protein-Protein Interaction Sites at the Molecular Level

Angshuman Bagchi

Abstract Protein-protein interactions (PPI) play pivotal roles in many biological processes like hormone-receptor binding. Their disruption leads to generation of inherited diseases. Therefore prediction of PPI is a challenging task. Machine learning has been found to be an appropriate tool for predicting PPI. Machine learning features generated from a set of protein hetero-complex structures were found to be a good predictor of PPIs. These machine learning features were used as training examples to develop Support Vector Machines (SVM) and Random Forests (RF) based PPI prediction tools. Among the important features the sequence based features related to sequence conservations and structure based features like solvent accessibility were found to have the maximum predictive capability as measured by their Area Under the Receiver Operating Characteristics (ROC) curves (AUC value). The RF based predictor was found to be a better performer than the SVM based predictor for this training set.

Keywords PPI · Machine learning · SVM · RF · Features · ROC

1 Introduction

Protein-protein interactions (PPI) play very significant roles in biological systems. PPI dysfunctions lead to different sets of diseases. However, the experimental methods of PPI identifications are very laborious, expensive and time consuming. There are different computational tools that predict PPIs mainly at the structure and

A. Bagchi (✉)
Department of Biochemistry and Biophysics, University of Kalyani,
Kalyani, Nadia 741235, West Bengal, India
e-mail: angshumanb@gmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 340,
DOI 10.1007/978-81-322-2247-7_6

network level [1–5]. Therefore, an alternative computational approach was considered using machine learning principles to predict PPIs at the molecular level using both sequence and structure information of proteins. The aim of the present work was to build a machine-learning predictor that can predict PPIs just from the sequence information of proteins. For that purpose, a set of more than 300 protein hetero-complex structures were taken from the Protein Data Bank (PDB). Several sequence and structure based features were extracted from this set of proteins. These machine learning features were used to train and evaluate SVM and RF based machine-learning predictors to discriminate between PPI and non-PPI amino acid residues at the molecular level. As inputs, the SVM and RF based predictors would accept amino acid sequence or a PDB formatted protein structure files. Two SVM predictors were developed using linear and Radial Basis Function (RBF) kernels. It was observed that the RF based predictor built on the sequence-based features from the aforementioned protein dataset could distinguish between interface and all other non-interface residues with an accuracy of 76.7 % thereby outperforming the SVM based predictors in all cases in terms of overall accuracy of the prediction process. The RF based predictor built using a combination of sequence- and structure-based features could differentiate between PPI interface and non-interface surface residues with an accuracy of 70.7 %. However, both the SVM predictors built using linear and RBF kernel SVM methods could perform their tasks with 53.3 % and 50.2 % cross-validation accuracies, respectively. It is to be noted that the training set for structure based predictor comprised of PPI interface residues as positive examples and non-interface surface residues as negative examples. But the training set for the sequence-based predictor was made up of PPI interface residues as positive examples and non-interface surface as well as other core residues as negative examples. To test the applicability of our predictor real life examples of PPI data from PDB were used for which there were experimental evidences of protein-protein-interactions. This method is one of such very few tools that provide residue level background of PPIs using only the protein sequence data. The importance of sequence based predictors lie in the fact that more and more protein sequences are added to sequence databases whereas the growth of structure database is very limited. With this tool the researches would be able to have a firsthand knowledge of PPIs just from the amino acid sequences of the proteins.

2 Materials and Methods

2.1 The Dataset Training Dataset for PPI Prediction

The dataset for training purposes was obtained from Chung et al. [6]. It contained X-ray crystal structures of protein hetero-complexes with resolutions less than 3.5 Å. The dataset had 274 non-redundant (as per Chung et al.) chains of protein hetero-complexes with 10,305 interface and 27,172 non-interface residues.

2.2 Feature Generation

The above mentioned dataset was used to extract machine learning features. There were a total of 271 structure based features and 43 sequence based features. The sequence based features were generated from several runs of PSI-BLAST [7]. The structure based features were generated from PDB files.

2.3 Building of Supervised PPI Predictors

The dataset was further subdivided on the basis of the type of features. The sequence based features were used to train sequence based predictors and structure based predictors were trained on features obtained using structural information. In both the cases the datasets were kept balanced by using equal numbers of positive and negative examples. To get SVM based predictors, LibSVM package was used. For developing RF based predictor, R-package was employed. Two SVM based predictors were developed; one with linear kernel and the other with RBF kernel with default values of Regularization parameter C and γ with 10-fold cross validations. C is the penalty factor. γ represents the effectiveness of a training example. It is the reciprocal of the number of features. For RF based predictors, 1,000 trees were generated.

2.4 Evaluation of Predictor Performances

The prediction abilities of the machine learning features were evaluated by calculating the AUC values from the ROC curves. The performances of the different machine learning predictors were compared by calculating the standard measures like accuracy, sensitivity, precision, specificity. 10-fold cross validations were used in each case.

3 Results and Discussions

3.1 Identification of the Best Features

The best features with the highest levels of class discriminating abilities were obtained from their values from the Area Under the Receiver Operating Characteristics Curve i.e., the AUC values. Among the sequence based features, the PSSM,

Table 1 Evaluation of the machine learning features with the most class discriminating abilities, obtained by the area under the ROC curve (AUC)

Rank	Feature	AUC
1	PSSM	0.91
2	Solvent accessibility	0.91
3	Information per position	0.88
4	Frequency of Lys residues in a 20 amino acid sequence window	0.86
5	Number of neighboring charged residues (Arg, Asp, Glu, Lys)in Shell 3	0.82
6	Number of Lys residues in Shell 4	0.80
7	Number of neighboring charged residues (Arg, Asp, Glu, Lys) in Shell 2	0.76
8	Number of carbonyl groups in Shell 1	0.73
9	Number of positive ions from His in Shell 4	0.66
10	Secondary structure of the neighboring amino acid residue	0.63

which is the measure of sequence conservation, had the best class discriminating ability. Among the structure based features solvent accessibility ranked the best. Top 10 features were listed in Table 1.

3.2 Comparison of Predictor Performances

Table 2 presents the comparative estimates of the performance measures in terms of sensitivity, specificity, accuracy and the AUC values of the different machine learning predictors. From the table it is quite apparent that the sequence-based RF predictor (Table 2c) was the best among all the sequence based predictors with 77 % overall cross-validation accuracy compared to 61 and 59 % cross-validation accuracies, respectively, for the linear and RBF kernel SVMs. On the other hand, the accuracies for the structure-based RF and SVM-linear and RBF kernel predictors were 71, 53 and 50 %, respectively as presented in Table 2a. The better performance of the sequence based RF predictor as compared to the structure based RF predictor may be attributed to the use of different training sets. The sequence based predictor was trained on interface (positive examples) and non-interface surface and core residues (negative examples) whereas the structure based predictor was built using only the interface and non-interface surface residues excluding the core residues. The worst performer among these predictors was the sequence based predictor trained on interface residues (positive examples) and non-interface residues (negative examples) with overall accuracies of 69, 57 and 58.5 % for RF, SVM-linear and SVM-RBF predictors respectively.

Table 2 Comparative estimate of predictor

Dataset	Method	Accuracy (%)	AUC	Recall/Sensitivity (%)	Precision (%)	Specificity (%)
a *The training set comprises interface residues as positives and non-interface surface residues as negatives*						
Structure	SVM linear	53.3	0.53	22.7	58.4	83.9
Structure	SVM RBF	50.2	0.50	70.7	50.1	29.6
Structure	RF	70.7	0.78	66.3	72.7	75.1
b *The training set comprises interface residues as positives and non-interface surface residues as negatives*						
Sequence	SVM linear	57	0.57	47.1	58.7	66.6
Sequence	SVM RBF	57.4	0.57	49.3	58.8	65.5
Sequence	RF	69.3	0.75	67.3	70.1	71.3
c *The training set comprises interface residues as positives and all non-interface surface and core residues as negatives*						
Sequence	SVM linear	60.5	0.63	57.9	61.1	63.1
Sequence	SVM RBF	58.9	0.59	51.6	60.5	66.3
Sequence	RF	76.7	0.84	74.8	77.8	78.7

4 Conclusions

In this work SVM and RF based machine learning predictors were built using a dataset of more than 300 protein hetero-complex structures. Machine learning features were generated and ranked as per their class discriminating abilities. The sequence based RF predictor ranked the best among all the predictors. Most importantly, this work is one of those works that deal with the prediction of PPIs from protein sequence information only.

Acknowledgments The author is grateful to the BIF Center, Dept of Biochemistry and Biophysics, University of Kalyani for providing workstation to carry out the experiments. The author would like to acknowledge the ongoing DST-PURSE program 2012–2015 for the infrastructural support.

References

1. Park, J., Lee D.-S., Christakis, N.A., Barabasi, A.-L.: The impact of cellular networks on disease comorbidity. Mol. Sys. Biol. **311**, 1–7 (2009)
2. Jones, S., Thornton, J.M.: Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA **93**, 13–20 (2002)
3. Nooren, I., Thornton, J.M.: Diversity of protein-protein interactions. EMBO J. **22**, 3486–3492 (2003)

4. Bogan, A.A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. J. Mol. Biol. **280**, 1–9 (1998)
5. Ofran, Y., Rost, B.: ISIS: interaction sites identified from sequence. Bioinformatics **23**, e13–e16 (2007)
6. Chung, J.L., Wang, W., Bourne, P.S.: Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins: Struct. Funct. Bioinf. **62**, 630–640 (2006)
7. Altschul, S.F., Gish, W., Miller, W., et al.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)