Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 62 – 66

Procedia Technology

International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA) 2013

Analysis of features from protein-protein hetero-complex structures to predict protein interaction interfaces using machine learning

Angshuman Bagchi[a*], Matthew Mort[b], Biao Li[c], Fuxiao Xin[d], Carson Carlise[c], Tal Oron[c], Corey Powell[c], Eunseog Youn[e], Predrag Radivojac[d], David N. Cooper[b], Sean D. Mooney[c,f*]

[a]*Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, India*
[b]*Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom*
[c]*The Buck Institute for Age Research, Novato, California-94945*
[d]*School of Informatics, Indiana University, Bloomington, Indiana*
[e]*Department of Computer Science, Texas Tech University, Lubbock, Texas-79409*
[f]*Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana-46202*

Abstract

Protein-Protein-Interactions (PPIs) play the most important roles in most (if not all) of the biological processes. A few such examples include hormone–receptor binding, signal transduction, chaperone activity, antigen-antibody interactions. The disruptions of PPIs may therefore lead to the development of human inherited diseases. There are different analytical techniques to identify amino acid residues in protein interfaces. But they are time consuming, labour intensive and above all very expensive. As an alternative approach to the analytical methods, we have tried to develop machine learning tools to differentiate between protein interface and non-interface amino acid residues. We used sequence- and structure-based features derived from a set of protein hetero-complex structure files from the Protein Data Bank (PDB). We have built supervised predictors based on Random Forests (RF) and Support Vector Machines (SVMs). We have evaluated them with 10-fold cross-validations. Both of our sequence and structure based RF predictors performed better than SVM based ones. The most predictive sequence- and structure-based features are the attributes which measure sequence conservation at a specified amino acid residue and various other measurements of the amino acid residue's neighbouring charge distributions. Our sequence- and structure-based RF classifiers have been validated by evaluating them against the protein complexes with experimentally proven interaction sites. Our predictors are found to detect the protein interface residues in practice.

Availability: http://www.sblest.org/ppi

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Protein-protein interactions; Machine Learning; SVM; RF; Sequence and Structure based Features

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering
doi:10.1016/j.protcy.2013.12.337

1. Introduction

Most of the biological processes are mediated by a complex interplay of protein-protein-interactions (PPIs). Therefore, the understanding of the structural basis of PPIs remains an important endeavor. There are many experimental methods available but they are very expensive both in terms of cost and man power. As an alternative, we have made an attempt to build a computational tool to predict the protein binding and non-binding amino acid residues using machine learning techniques such as Support Vector Machines (SVMs) and Random Forests (RFs) [1-5]. We have used an in-house software suite [6] to extract over 300 sequence- and structure-based features from a non redundant set of protein hetero-complexes in the Protein Data Bank (PDB) [www.pdb.org]. We used those features to build RF and SVM based machine learning predictors. The predictors take the amino acid sequence of a single protein or the three dimensional coordinates of the amino acid residues of a given protein as input. For the SVM predictors we have used Linear and Radial Basis Function (RBF) kernels. We have generated two RF and SVM predictors using separately the sequence and structural features. The sequence based RF predictor can distinguish between interface and all other non-interface amino acid residues with an accuracy of 76.7% outperforming the both the Linear and RBF kernel SVM predictors. On the other hand, the structure based RF predictor is able to discriminate between PPI interface and non-interface surface residues with an accuracy of 70.7% and again has been found to be more accurate than the SVM based ones. In developing the predictors we have used a training-set comprising the PPI interface residues as positives and non-interface surface residues as negative examples for the structure-based predictors. The sequence-based predictors were trained on a dataset made up of PPI interface residues as positive and all other non-interface residues as negative examples. In order to validate our predictors, we have analyzed the PPI data from PDB for which there are experimental evidences of protein-protein-interactions. There are only a very few methods available that are able to produce residue level insight into the PPIs using just protein sequence information. Our approach therefore, represents one such method which is designed to elucidate the residue level PPI interaction schemes using just the protein sequence information. Due to faster growth of protein sequence databases as compared to structure databases we focused on developing a sequence-only PPI predictor as that would help researchers to have a quick idea of the hitherto unknown function(s) of the protein from its sequence. We further have developed a web-server and a web service plug-in for UCSF Chimera. The user can submit a PDB-formatted protein structure file or a fasta-formatted protein sequence file or both as inputs to the web-server with a valid e-mail id. The results will be e-mailed back to the user when it is ready. The user will be able to visualize the prediction results from our structure-based PPI predictor using the UCSF Chimera web plug-in. The users can freely access these tools at http://www.sblest.org/ppi.

2. Materials and Methods

2.1 PPI training dataset

We got the dataset from Chung *et al.* [9]. It consists of X-ray crystal structures of protein hetero-complexes with a resolution better than 3.5Å. We used the same definitions of surface and interface residues as proposed by Chung *et al*. The final dataset had 274 non-redundant (as per Chung *et al.*) chains of protein hetero-complex structures containing 10,305 interface residues and 27,172 non-interface surface residues.

2.2 Feature encoding

We used a total of 314 discriminatory features for the development of the machine learning classifiers. There were 271 structure-based features and 43 sequence-based features for each amino acid residue of the input protein. The sequence-based features, taken from [6], measured sequence conservation and were derived from position specific scoring matrices (PSSM) obtained from the output of PSI-BLAST using the amino acid sequence of the input protein. The PSSM score of a position is a log-likelihood ratio for the most likely amino acid substitution at that position, and the weighted observed percentage is calculated from the PSSM score and represents the relative weighted frequency of the most likely amino acid substitution. The information per position (IPP), on the other hand, is the entropy of the position [6]. A window of size 21 (10 residues to the left and 10 residues to the right and

the residue itself) of amino acid frequencies was created from the BLAST output and the distribution of amino acids in this window was encoded as 20 features, one for each amino acid as described in [6]. The PSSM scores for the 20 positions to the left and right of the residue were also captured as features.

2.3 Development of supervised PPI predictors

The aforementioned dataset was further subdivided on the basis of the type of features used. One of the datasets had only sequence-based features while the other one had both sequence and structure-based features. We used balanced datasets in both the cases where equal numbers of positive (interface) and negative (non-interface) examples were present. We used the RF package in R and the LibSVM package to implement separate RF and SVM predictors using the aforementioned datasets with 10-fold cross-validation. Two SVM predictors, one with a linear kernel and the other with a RBF kernel, were built from each dataset. Throughout the experiments, default values of the regularization parameter (C) and γ for linear and RBF kernel SVMs were used. For RF, 1000 trees were generated keeping other parameters to their default values.

2.4 Predictor Performance Evaluation

We used the receiver operating characteristic (ROC) curves and calculated the area under the ROC curve (AUC) to measure the ability of the predictors to discriminate between PPI sites and non PPI sites. We also calculated the standard performance measures including accuracy, sensitivity (recall), specificity and precision with10-fold cross-validation.

3. Results

3.1 Best Features for the Identification of PPI Sites

The class discriminatory abilities of the sequence and structure based features have been measured as the 'Area Under the Receiver Operating Characteristics (ROC) Curves (AUC)' for a predictor using just that feature to distinguish between interface and other non-interface amino acid residues. Table 1 gives the AUC values for the top 5 predictive features. The sequence based feature PSSM is the feature with the highest discriminatory capability.

Table 1: Evaluation of the most informative features, ranked by the area under the ROC curve (AUC).

Rank	Feature	AUC
1	Position Specific Scoring Matrix (PSSM)	0.67
2	Weighted observed percentage	0.65
3	Information per position	0.63
4	Frequency of Lys residues in a 20 amino acid sequence window	0.57
5	Number of neighbouring charged residues (Arg, Asp, Glu, Lys)	0.57

3.2 Performance Comparison of Different Classification Models

The performances of the sequence- and structure-based predictors have been compared by calculating the accuracy, AUC, sensitivity, precision and specificity. The results are provided in Table 2. The performance of the sequence-based RF predictor (Table 2C) has been found to be the best with 76.7 % cross-validation accuracy versus 60.5% and 58.9%, respectively, for the linear and RBF kernel SVM predictors. The same results were obtained when we compared the cross-validation accuracies of the structure based RF predictor with the structure based SVM-linear and -RBF kernel predictors (Table 2A). The better cross-validation accuracy of the sequence-based RF predictor (Table 2A), as compared to the structure-based RF predictor (Table 2C) is because of the exclusion of core amino

acid residues from the training and evaluation of the structure-based predictors (data not shown). Then we evaluated the sequence based predictors trained on a dataset comprising interface residues as positives and non-interface residues as negatives and excluded the core residues from the training dataset. We obtained the worst performances in terms of cross validation accuracies; 69.3%, 57% and 57.4% for RF, SVM-linear and SVM-RBF, respectively, as shown from Table 2B.

Table 2: Comparison of machine learning cross-validation results. We defined the dataset as Structure when both structure- and sequence-based information were used for training. Similarly, when we used only the sequence information fro training purpose, the dataset was defined as Sequence.

A. The training set comprises interface residues as positives and non-interface surface residues as negatives.

Dataset	Method	Accuracy (%)	AUC	Recall / Sensitivity (%)	Precision (%)	Specificity (%)
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.

Dataset	Method	Accuracy (%)	AUC	Recall / Sensitivity (%)	Precision (%)	Specificity (%)
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.

Dataset	Method	Accuracy (%)	AUC	Recall / Sensitivity (%)	Precision (%)	Specificity (%)
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

3.3 Case Study

For the validation purposes, we have used known examples from the literature of experimentally determined PPI sites and analyzed them with our predictors [7]. The protein involved in this case study was excluded from the underlying training sets.

***Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2)* (PDB id: 1PD3)**

The influenza A virus, which infects humans and other vertebrates like pigs, waterfowl, etc., is the causative agent of the world flu pandemic [8]. During the infection cycle, viral ribonucleoproteins are replicated in the nucleus, which must then be exported to the cytoplasm to form the mature viral particles. Nuclear export is mediated by the cellular protein Crm1 and putatively by the viral protein NEP/NS2. When we have applied our structure- and sequence-based RF predictors to analyze the viral protein complex our predictors could successfully identify the essential amino acid residues involved in the transport of the virus proteins from the nucleus to the cytoplasm as observed by previous studies. They are Ile76, Gln101, Ala102, Leu105 and Val109 from chain A of the protein [8]. Both our predictors were able to choose those amino acid residues with PPI probability scores $\geq$ 0.61. The interactions are presented in table3.

Table3: Evaluation of RF model using case studies of experimentally derived PPI sites

PDB code	Experimentally derived PPI site	RF prediction probability (structure)	RF prediction probability (sequence)
1PD3	Ile76	0.71	0.61
1PD3	Gln101	0.72	0.74
1PD3	Ala102	0.66	0.73
1PD3	Leu105	0.76	0.76
1PD3	Val109	0.67	0.73

4. Conclusion

We have developed and subsequently evaluated several machine learning classifiers for the identification of PPI interfaces. We have used sequence and structure based features to train the machine learning classifiers. Our classifiers could be applied to protein sequence and structures to have a fair amount of guess about the probable protein binding sites. In other words, our classifiers could have been used for the prediction, prognosis and treatment of inherited disease conditions brought about by disruption of PPI sites. Our sequence only predictors could be used to analyze protein sequences for which no structures are available. We have developed a web resource for the users to predict PPI sites using either sequence alone, or structure and sequence together. This resource can be accessed freely at http://www.sblest.org/ppi.

Acknowledgements

We would like to thank Prof. Philip E. Bourne of UCSD for providing us with the dataset used for training. This research was supported by K22LM009135 (PI: Mooney), R01LM009722 (PI: Mooney), grants from IU Biomedical Research Council, Indiana University, the Showalter Trust and the Indiana Genomics Initiative and the Buck Institute for Age Research. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment.

References

1. J. Park, D.-S. Lee, N.A. Christakis and A.-L. Barabási, The impact of cellular networks on disease comorbidity. Mol. Sys. Biol. 311 (2009) 1.
2. S. Jones and J.M. Thornton, Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93 (1996) 13.
3. I. Nooren and J.M. Thornton, Diversity of protein–protein interactions. EMBO J. 22 (2003) 3486.
4. A. A. Bogan and K.S. Thorn, Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280 (1998) 1.
5. Y. Ofran and B. Rost, ISIS: interaction sites identified from sequence, Bioinformatics 23 (2007) e13.
6. E. Youn , B. Peters, P. Radivojac and S.D. Mooney, Evaluation of features for catalytic residue prediction in novel folds. Protein Sci. 16 (2007) 216.
7. N. Zaki, S. L. Molnar, W.E. Hajj and P. Campbell, Protein-protein interaction based on pairwise similarity. BMC Bioinformatics 10 (2009) 150.
8. H. Akarsu et. al., Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). EMBO J. 22 (2003) 4646.
9. J. L. Chung, W. Wang and P.E. Bourne, Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins: Structure, Function, and Bioinformatics. 62 (2006) 630.