International Journal for Computational Biology (IJCB)
Vol.3, No.2, August 2014, pp. 37-43
ISSN: 2278-8115

Determination of protein-protein interaction through Artificial Neural Network and Support Vector Machine: A Comparative study

Himansu Kumar [1*], Swati Srivastava [2], Pritish Kumar Varadwaj [1]
[1] Indian Institute of Information Technology Allahabad, India
[2] Lovely Professional Universitiy, Jalandhar, Punjab, India

Article Info

Article history:

Received Jul 2nd, 2014
Revised Jul 21th, 2014
Accepted Jul 28th, 2014

Keyword:

Protein-Protein Interaction
Machine Learning
Artificial Neural Network
Support Vector Machine

ABSTRACT

Protein-protein interactions (PPI) plays considerable role in most of the cellular processes and study of PPI enhances understanding of molecular mechanism of the cells. After emergence of proteomics, huge amount of protein sequences were generated but there interaction patterns are still unrevealed. Traditionally various techniques were used to predict PPI but are deficient in terms of accuracy. To overcome the limitations of experimental approaches numerous computational approaches were developed to find PPI. However previous computational approaches were based on descriptors, various external factors and protein sequences. In this article, a sequence based prediction model is proposed by using various machine learning approaches. A comparative study was done to understand efficiency of various machine learning approaches. Large amount of yeast PPI data have been analyzed. Same data has been incorporated for different classification approach like Artificial Neural Network (ANN) and Support Vector Machine (SVM), and compared their results. Existing methods with additional features were implemented to enhance the accuracy of the result. Thus it was concluded that efficiency of this model was more admirable than those existing sequence-based methods; therefore it can be effective for future proteomics research work.

Corresponding Author:

Himansu Kumar
Indian Institute of Information Technology Allahabad.
Email:
himanshu.genetics@gmail.com

How to Cite:

Himansu Kumar *et. al.* Determination of protein-protein interaction through Artificial Neural Network and Support Vector Machine: A Comparative study. IJCB. 2014; Volume 3 (Issue 2): Page 37-43.

1. INTRODUCTION

Protein-Protein Interaction (PPI) have been studied in the prospect of biochemistry, molecular dynamics, quantum chemistry, signal transduction and numerous metabolic or genetic networks [1]. However, PPI are significant for entire interactomics system of all living cell [2]. Abnormalities in the interactions among proteins may causes abnormalities in organism, for example: Huntington Disease, Neurodegenerative Disorder occurs due to repeated poly-glutamine and their interaction in a large protein huntingtin, with unknown function [3]. Analysis of PPI is significant to understand the complex cellular mechanism of an organism, and in searching targets for drug development. PPI prediction is an amalgamated approach of combining bioinformatics and structural biology to determine physical interactions between protein pair's [2]. Clear cut understanding of PPI of various cellular processes like signal transduction, modeling of protein complex structures, various biochemical processes and cellular mechanism is required [3]. It was reported that proteins which are having same functions are closer to interact [4]. If one of the interacting protein's functions is known then it has been proposed that another protein will possess same function.

Drug designing is also an another important area where PPI plays major role in identifying new drugs for disease. To understand drug and cell target interaction, it is crucial to have adequate idea about interaction between two proteins involved during drug target interaction. There are various factors which actively take part during PPI like polarity, hydrophobicity, polarizability, volume of side chain, solvent accessible area, and charge index and these factors are known as descriptors [5]. There are several machine learning approaches like SVM, ANN, Bayesian classification etc. which have been used to understand biological glitches [4,6,7]. Here in this article SVM and ANN approaches were used to predict the PPI. As the amino acid sequence dataset composed of heterogeneous length i.e. Sequences are of varying length therefore prediction is difficult as input parameters are variable as required in machine learning [8]. Therefore in order to have realistic studies for PPIs, heterogeneous length of sequence is demanded to be converted into feature vector information resulting in homogeneous length. Autocorrelation descriptors help in converting numerical vectors of amino acid sequence into uniform matrices containing equal number [8]. .

2. RESEARCH METHOD

To determine protein-protein interaction foremost step is to determine model organism and collect the dataset of its protein sequence [9-13]. Here as model organism Saccharomyces Cerevisiae has been opted because tremendous work has been formerly done on this organism [18]. The complete interacting and non-interacting amino acid sequences pairs of the Saccharomyces Cerevisiae have been downloaded from KUPS (University of Kansas Proteomics Service) database. The dataset present in this database is in FASTA format [10]. Each amino acid have been assigned six descriptors and converted into uniform numerical string with the help of autocorrelation descriptor. It is a class of topological descriptor which converts numerical vectors into homogenous matrices [9].

2.1 Autocorrelation Descriptor:

It assigns physicochemical property of individual amino acid residue contained in protein sequence and compares the autocorrelation between two protein sequences. Autocorrelation Descriptor also considers the local environment of the each residue in the sequence. Now equal length amino acid sequence combined with another equal length interacting amino acid sequence like:

A+B as well as B+A.

2.2 Pre-processing of the dataset:

6000 sequences have been downloaded from KUPS database [10]. Out of 6000 amino acid sequences 4500 sequences were used as training dataset and 1500 sequences as test dataset based on 75:25 principle. Dataset of both interacting and non-interacting amino acid were translated into numerical string via assigning their six physicochemical descriptors to each of the amino acid residue of sequence. Suppose AQGTALP is an amino acid sequence than each individual residue of input amino acid sequence were assigned six numerical value of the descriptors depicting their characteristic features, like A was assigned six descriptors Q was assigned six descriptors and so on.

2.3 Implementation of Autocorrelation Descriptor:

After converting amino acid into their numerical string next task is how to compute interaction among them. The data was huge and heterogeneous therefore it was impossible to compute heterogeneous length data so the dataset was converted to homogenous data with the help of autocorrelation descriptor. Finally, autocorrelation descriptor contains total 30*6 = 180 descriptor values, where 30 is the length of the amino acid sequence and 6 is the number of descriptors. A 180-dimensional vector was built representing protein sequence. An interaction pair is developed through concatenation of two protein sequences [8].

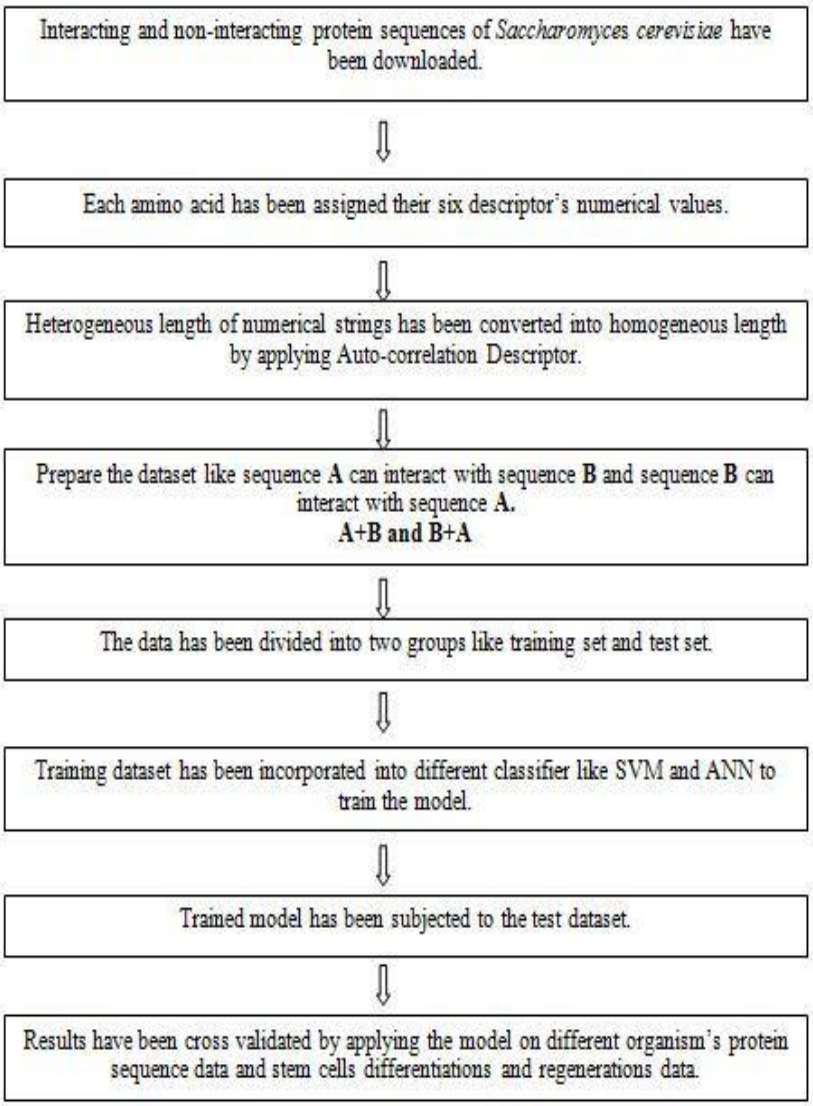

Figure 1: Flow chart of the methodology.

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \times \left(X_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \quad (1)$$

Where AC=Autocorrelation value, j=one descriptor, i= position in the sequence X, n=length of the sequence, lag= distance between one descriptor to its neighbour, after calculating the AC variables through above equation, concatenate the two sequences of interacting or non-interacting pair.

2.4 Implementation of Classification Approaches:

Machine learning algorithm was used to build a classifier for a given sample to separate data into distinct classes based upon their properties. Newly constructed model is validated through the use of test dataset by predicting their class labels. There are various machine learning approaches like SVM and ANN to classify data. To implement any classification method it requires total number of classes or attributes. The complete dataset must lie into the given classes. Assignment of class labels for each subset dataset should be declared before classification. In our cases there were two classes one is interacting pairs, denoted as '1' and another is non-interacting pairs which is denoted as '0'.

2.5 Classification through Support Vector Machine:

It classifies the data by constructing N-dimensional hyper-plane which separates the dataset into two divisions. In training set there is a class label '+1' and '0', the classifier construct a hyper-plane who maximize the margin between '+1' and '0'. Result of the machine learning approaches mainly depends upon the feature selection or descriptor selection. Descriptors are the features which describes amino acid physiological, physicochemical natures which are responsible for interaction among two proteins. Here six descriptors for each amino acid residue were selected. SVM classify the data in two classes like interacting or non-interacting by constructing two hyper-planes. Hyper-planes should be as far as possible. The points which lie along the hyper-plane were considered as support vector [13-17] . SVM works on the basis of their kernels like linear, sigmoid, gaussian, and radial basis function. In this article radial basis function was used. Literature survey indicates that radial basis function is having better accuracy for binary classification. RBF kernel is most suitable for dataset which is having class-conditional probability distribution and approaching to Gaussian distribution. The dataset can be separated linearly if it is framed into high dimension.

Radial basis function is represented as:

$$K(Xi,Xj)=\exp(\| -\gamma |(|Xi-Xj|)|\|\hat{}2) \quad (2)$$

Where K is kernel function, Xi is input vector, Xj is class like +1 or -1 and γ is kernel parameter.

Classification through Artificial Neural Network:

In this work MATLAB neural network toolbox (nntool) was employed to classify the data. After normalization of the data, it has been grouped into training and test set in ratio of 75:25 containing 6 descriptors value. Bi-classification method is used to get better and unbiased result. If amino acid sequence is interacting than it is having class label 1 and if it is not interacting than it is considered as 0 class label. Feed-forward back propagation network is used to classify the data which contains two hidden layers. The neural network was trained with assist of training data set and a model was constructed. Newly trained model has been subjected to the test dataset and generates the result in bi-classifier format that contains two class labels. Comparison has been done between values of two class's labels. Accuracy of the classifier was calculated with the help of TN (true negative), TP (true positive), FP (false positive), and FN (false negative) [17-22].

For data normalization we have used:

$$x' = \frac{(x - \min)\,(new\_\max - new\_\min)}{(max - \min\ \ ) + \ new\_min} \quad (3)$$

Where x' = Normalized value, x = Original descriptor value, 'max'= maximum value of descriptor, 'min' = minimum value,

After normalization new_min should be 0 and new_max should be 1.

Confusion matrix has been generated by calculating sensitivity, precision and accuracy to validate the result by the following equations:

Sensitivity = TP/(TP+FN) (4)

Precision = TP/(TP+FP) (5)

Accuracy = TP+TN/(TP+TN+FP+FN) (6)

3. RESULTS AND ANALYSIS

3.1. Support Vector Machine

A SVM classifier was modeled, taking 360 attribute (180 for sequence A + 180 for sequence B). The optimal parameter of C and γ were calculated. To determine different combinations of C and γ, the training set has been subjected to 10-fold cross validation and using grid based search. During cross-validation, the training sample was divided into 10 equivalent sets each having approximately 450 protein sequences with 360 attributes. Out of 10 one set is used as 'test set' and rest set of nine were taken as training data. The process continues till all sets were once used for test and training set. The contour plot was generated, depicting C = 32 and γ = 0.125. Through this C and γ we had trained RBF based SVM model using training set of 4500 protein sequence. This model was then subjected to identify a test set of 1500 protein sequence and the prediction accuracy came out to be 70.60 % i.e. (1059 protein sequence were correctly classified in 1500 sequence). A receiver operating curve (ROC) is an analytical tool which is used to check the classifier performance (Figure:2). It is a graph between true positive rate and false positive rate.

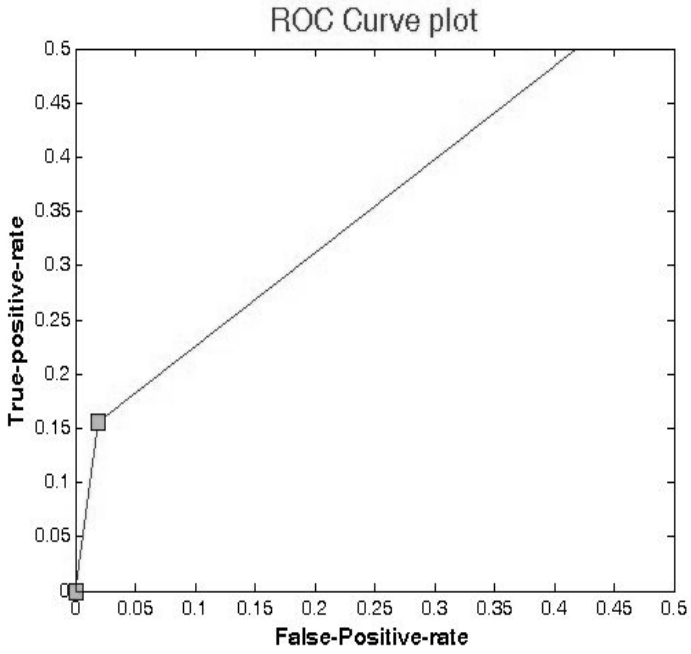

Figure 2: *A ROC curve generated through SVM classifier.*

3.2. Artificial Neural Network

Feed-forward back propagation neural network method was used. Numbers of layers selected were 2 and the model was trained iteratively till it resulted in optimized value i.e. no change in iteration values. The neural network was trained with assist of training data set and a model constructed. Newly trained model has been subjected to the test dataset and generates the result in bi-classifier format that contains two class labels. Comparison has been done between values of two class's labels. The amino acid pairs which is interacting taken as '1' and another which is not interacting taken as '0'. Accuracy of the classifier was calculated with the help of TN, TP, FP, and FN. The final accuracy of ANN classifier predicted was 72.60%. ROC curve has been plotted to check the performance of model (Figure:1).

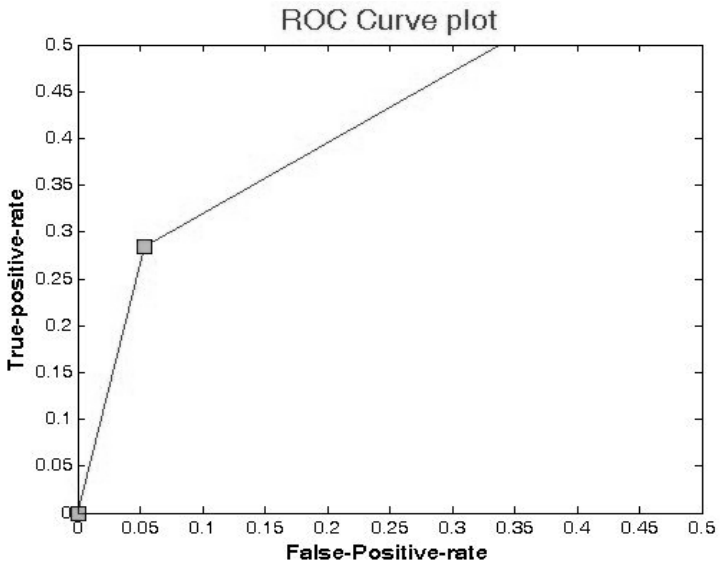

Figure3: *A ROC curve generated through ANN classifier*

Comparative study of two classifiers:

Same datasets were tested on two different classifiers by keeping all the conditions same like same training and test dataset with same number of descriptors. The two classifiers output were analyzed to detect best prediction accuracy for the protein-protein interaction. It was seen that RBF kernel based SVM came out to be 70.06% and feed forward neural network gives 72.6 % accuracy (Table 1). Further comparative study of the two classifiers can be demonstrated via chart. To validate the result various statical analysis like Recall, Mathews, Youden's index and F measures have been calculated and shown in Table 2**.**

Table 1: Accuracy and other parameters of the classifiers

S.N	Classifier	Accuracy	TP	TN	FP	FN
1	ANN	72.6 %	142	947	53	358
2	SVM	70.6 %	78	981	19	422

Table2: Comparative chart of the various parameters of the classifiers

S.N	Classifier	Sensitivity	Specificity	Precision	Recall	Mathews	Youden's	F meas.
1	ANN	0.28	0.94	0.73	0.28	0.32	0.93	0.57
2	SVM	0.15	0.98	0.8	0.16	0.26	0.78	0.31

4. CONCLUSION

This work had tried to improve the accuracy of the available classifiers and successfully implemented some new features for the classification of protein-protein interaction. The organism *Saccharomyces Cerevisiae* has been selected because of it's easy availability and it has been model organism for numerous research work. The trained model has been subjected finally to the test dataset and results were validated. In these work two classifiers i.e. SVM and ANN have been selected for the prediction of PPI. It has noticed that ANN is showing better accuracy which is 72.6 % as compared to SVM which is 70.6 %. The model was cross validated by using independent data set of different organism like plasmodium falciperum, stem cell differentiations and regenerations and reported result was meeting the authenticity of the predictive model. The importance of this work can be estimated due to their direct involvement into various biological pathways, and cellular

mechanisms. Certainly this work would be helpful to discover new protein-protein interaction and finally new drug and protein interaction. This work could be extended for increment of accuracy rate by including more descriptors.

5. ACKNOWLEDGEMENTS

The Author would like to thank Indian Institute of information Technology, Allahabad, India for providing valuable infrastructure and research environment to complete this work.

6. REFERENCES

[1] M. Deng, K. Zhang, S. Mehta, T. Chen, F. Sun, Prediction of protein function using protein–protein interaction data, IEEE Computer Society Bioinformatics Conference, 2002;197–206.
[2] J.H. Lakey, E.M. Raggett, Measuring protein–protein interactions, Curr. Opin. Struct. Biol. 8, 1998; 119–123.
[3] P. Legrain, J.Wojcik, J.M. Gauthier, Protein–protein interaction maps: a lead towards cellular functions, Trends Genet.; 2001; 17, 346–352.
[4] A.Valencia, F. Pazos, Computational methods for the prediction of protein interactions, Curr. Opin. Struct. Biol. 12 ; 2002; 368–373.
[5] Charton, M.; Charton, B. I.The structural dependence of amino acid hydrophobicity parameters. J. Theor. Biol.; 1982; 99(4), 629-44.
[6] S.M. Gomez,A. Rzhetsky,Towards the prediction of complete protein–protein interaction networks, Pac. Symp. Biocomput. 7; 2002; 413–424.
[7] R. Bandyopadhyay, K. Maatthews, D. Subramanian, X.X. Tan, Predicting protein-ligand interactions from primary structure, Rice University, Department of Computer Science, Technical Report TR02-387, February 2002.
[8] Marcotte, E. M.; Xenarios, I.; Eisenberg, D. Mining literature for protein-protein interactions. Bioinformatics; 2001; 17(4), 359-363.
[9] Xia J.F., Han K. Sequence-Based Prediction of Protein-Protein Interactions by Means of rotation Forest and Autocorrelation Descriptor. Protein & Peptide Letters; 2010; 17, 137-145.
[10] Xue-wen Chen, Jong Cheol Jeong, and Patrick Dermyer (2010). KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. Nucl. Acids Res. (Database issue): First published online: October 15, 2010.
[11] Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T.O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. Science; 1999; 285(5428), 751-753.
[12] Pazos, F.; Helmer-Citterich, M.; Ausiello, G.; Valencia, A. Correlated mutations contain information about protein-protein interaction. J. Mol. Biol.; 1997; 271(4), 511-523.
[13].C.A.Kumar, M.Ankush, Sungc W, Probabilistic prediction of protein–protein interactions from the protein sequences. Computers in Biology and Medicine; 2006; 36,1143–1154.
[14]. Xuchun Li_, Lei Wang, Eric Sung, AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence; 2008; 21,785–795.
[15] Nanni, L. Fusion of classifiers for predicting protein-protein interactions. Neurocomputing; 2005; 68(3), 289-296
[16] G Yanzhi, Y Lezheng, W Zhining and Menglong. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Research; 2008; 1–6 doi:10.1093/nar/gkn159.
[17] H.Q. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics; 2001; 4 (17) 349–358.
[18] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, T. Takagi, Assessment of prediction accuracy of protein function from protein–protein interaction data,Yeast ; 2001; 18; 523–531.
[19] S. Letovsky, S. Kasif, Predicting protein function from protein/protein interaction data: a probabilistic approach, Bioinformatics; 2003; 19 (1) I197–I204.
[20] J.Weston, S. Mukherjee, O. Chapelle, M. Pontil, V. Vapnik, T. Poggio, Feature selection for SVMs, Adv. Neural Inform. Process. Syst; 2000; 668–674.
[21] Nanni, L Lumini, A. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics; 2006; 22(10), 1207-10.
[22] Chang CC and Lin CJ et al., A practical guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin, 2003.