PROTEIN-PROTEIN INTERACTION CLASSIFICATION TECHNIQUES: A REVIEW

Dilpreet Kaur
Department of Computer Science and Engineering
PEC University of Technology
Chandigarh, India

Shailendra Singh
Department of Computer Science and Engineering
PEC University of Technology
Chandigarh, India

Abstract - **Protein is a very important part of a cell. It carries out its duties as specified by the information encoded in genes. Proteins interact with each other to carry out different biological functions. Information about these interactions will help a lot in understanding different diseases. Different classifiers have been used till now to classify the protein-protein interactions. Some of the classifiers are SVM, SVM-KNN, BP Neural Network. This paper presents the different classifiers that had been used and will also present the future scope in the classification of the protein-protein interactions.**

Keywords: Protein-Protein Interaction, SVM, KNN, BP Neural Network

I. INTRODUCTION

In the coming genomic era, biologists pay more attention on protein structures and functions. However, protein-protein interactions are responsible for the regulation of biological processes and functions in living organisms. The protein-protein interaction datasets can help significantly in identifying protein functions based on recent researches [1]. Protein-Protein interaction is very hot research topic now a days and many algorithms have been proposed on this topic. But still PPI information extraction is a challenging task as best performing PPI information extraction systems are far from user satisfaction [2]. The broad recognition of importance of characterizing the set of all protein interactions in a cell has rendered itself in the development of various experimental and computational techniques [3]. Many computational methods have been proposed for the prediction of protein-protein interactions. There are a few methods that are based on genomic information, such as protein phylogenetic profiles [4], conservation of gene neighborhood [5], gene fusion events [6]. In addition, structural analysis of protein [7] is also considered. Different classifiers have been used for predicting protein-protein interactions. Some of them are SVM [8], SVM-KNN [2], BP Neural Network [1].

In this paper comparison is done between different classifiers that can predict protein-protein interactions that have been developed in past few years. In this paper we will introduce the brief overview of protein-protein interaction classifiers, basics of protein-protein interactions, different classifiers that have been used for the prediction of protein-protein interactions, analysis and results of different classifiers. At the end everything will be explained in brief in conclusion.

II. BASICS OF PROTEIN-PROTEIN INTERACTIONS

Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. Signals from the exterior of a cell are mediated to the inside of that cell by protein–protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases. Proteins might interact for a long time to form part of a protein complex, or a protein may interact briefly with another protein just to modify it. PPI knowledge is the fundamental basis of studying cellular process and mechanism of disease, and is especially useful in predicting unknown functions of protein [12] [13] 14]. PPIs have been studied individually to elucidate the mechanism of focused process. Recently, large quantities of protein interaction data are collected due to the evolution of high-throughput experiments [15].

III. OVERVIEW OF PROTEIN-PROTEIN INTERACTION CLASSIFIERS

Support vector machine (SVM) is a statistical concept for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier. SVM is very useful in predicting protein-protein interactions in that domain information and several protein features including amino acid composition, sequential amino acid usage, and localization, all these are taken into account, whether they are continuous or discrete values, and they can be easily combined into a feature vector [10]. SVM and KNN are combined to form a new classifier for

Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering
Editor In Chief Dr. R. K. Singh. Copyright © 2012 Universal Association of Computer and Electronics Engineers.
All rights reserved. ISBN: 978-981-07-1847-3

SEEK DIGITAL LIBRARY

predicting protein-protein interactions and it also improves the accuracy of SVM classifier [11]. The k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. It was firstly introduced by Covert and Hart. They attributed a test sample the same label class as the label of majority of K nearest neighbors. SVM-KNN classifier performs well on balanced data whereas its performance declines significantly on unbalanced data as it is sensitive to unbalanced training data. [2]. BP Neural Network is supervised learning technique. BP neural network can improve the accuracy of classification. There are several theoretical advantages of BP neural network that make it especially adaptable to be employed in interacting prediction: (a) It can process a batch of testing dataset and gain the predicting results quickly. (b) It is a learning network with teacher. (c) Since the output of the BP neural network is probabilities, we can regard it as a degree of their relationship [1].

IV. PROTEIN-PROTEIN INTERACTION CLASSIFIERS

Protein-protein interactions play a crucial role in cellular processes. Although these interactions should be determined by stringent experiments and succeeded assays, these steps are time-consuming and labor-intensive. For this reason, many efforts have been made to predict unknown protein-protein interactions [10]. Different researchers used various classifiers to predict and classify protein-protein interactions.
In this paper the accuracy of different classifiers is compared. At the end analysis is done that which classifier performs well in predicting protein-protein interactions.

A. SVM CLASSIFIER BASED PROTEIN-PROTEIN INTERACTIONS

Vapnik and his co workers introduced Support Vector Machines as a supervised machine learning algorithm for binary classification. There are some advantages of SVM over conventional statistical learning algorithms and it also shows high generalization performance independent of feature vector dimensions [2]. The algorithm is chosen to maximize the margin that is the distance from closest pattern. The main aim of SVM is to minimize an upper bound of generalization error by maximizing the margin between the separating hyper plane and data. Fixing the following quadratic equation can help in finding the optimal hyper plane [18]:

$$max \sum_{i=0}^{n} u_i \; -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} u_i \, v_i \; u_j \; v_j \; k\,(w_i, v_i);$$

$$\text{s.t.} \sum_{i=1}^{n} v_i \; u_i = 0, 0 \le u_i \le x, i = 1,2, \dots, n \qquad (1)$$

The function $k\left(w_i, w_j\right) = \phi(w), \phi\left(w_j\right)$ is called kernel function, $\phi(w)$ is the mapping from input space to feature space. If the training data cannot be separated linearly then a non linear boundary can be made using this kernel function. It moves the training data to a high-dimensional space [2]. Whereas, the traditional methods minimize the empirical training error by mapping the input data space to high-dimensional feature data set and apply the structure risk minimization [16].
One of the main challenges in using SVMs for the prediction of PPIs in genome sequence is a suitable encoding of the genome sequences information in some vector space and requires a fixed number of inputs for training. However, there are often unequal length vectors because of protein sequences with different lengths. So a transformation is proposed that converts protein sequence into fixed-dimensional representative feature vectors, where each feature records the correlation of amino acids to the protein sequences of interest [8]. Another challenge in using SVM's is the choice of kernel. This is the reason this technique doesn't provide much of the accuracy.

B. SVM-KNN CLASSIFIER BASED PROTEIN-PROTEIN INTERACTIONS

The k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor.

SVM-KNN classifier does the following[19]:

1) compute distances of the query to all training examples and pick the nearest K neighbors.
2) if the K neighbors have all the same labels, the query is labeled and exit; else, compute the pair wise distances between the K neighbors;
3) convert the distance matrix to a kernel matrix and apply multiclass SVM;
4) use the resulting classifier to label the query

SVM-KNN classifier combined SVM with KNN is presented to improve the accuracy of SVM classifier [11]. In class phase, the algorithm computes the distance from the test sample to the hyper plane of SVM in feature space as described in formula (2); if the distance is greater than the given threshold ε on SVM; otherwise, the KNN algorithm will be used, that is, every support vector is selected as representative point, then the distance between the test sample and every support vector is compared, finally the test sample can be classified by the k nearest neighbors of the sample [2].

$$d(w, w_i) = \|\varphi(w) - \varphi(w_i)\| = \sqrt{K(w,w) - 2K(w, w_{i)} - K(w_i, w_i)} \qquad (2)$$

SEEK DIGITAL LIBRARY

C. PROTEIN-PROTEIN INTERACTION CLASSIFICATION USING BP NEURAL NETWORK

BP (Back Propagation) is a multilayer feed-forward networks according to the training of error reversion propagation algorithm and it can learn and store a large number of input-output models mapping relation without having to reveal the mathematical equation which describes the mapping relation in advance. Its learning rule is to use the steepest descent method to constantly adjust the network weights and thresholds, through the back propagation, so that error sum squares of the network is minimum[20].

BP neural network is composed of three parts- input layer, single or multi-hidden layer (middle layer) and output layer. Fig. 1 gives the model of perception neural network comprises single hidden layer.

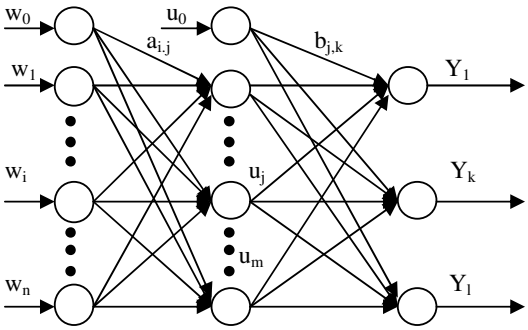

Fig 1: Single hidden layer Neural Network

In this figure, i=0, 1,…, n;

j= 0,1,…,m;

k= 0,1,…,l;

n, m, l is respectively the neuron number about the input layer, hidden layer and output layer of the neural network.

x_0, y_0 is -1. It is used to set the threshold value;

x_i indicates the ith input of these neurons;

$v_{i,j}$ is input weight connecting the ith input and the jth neurons.

Σ indicates neurons deal with the weight disposal about input signal;

f(x) is the activation function of neuron, usually takes single-polarity Sigmoid function,

$f(x) = \frac{1}{1+e^{-x}}$;

y_j is the output of jth neuron.

BP algorithm is divided into two phases-forward propagation and back-propagation. The neurons of input layer are responsible of receiving input information from the outside world and pass it to the neurons of middle layer; the middle layer is the internal information processing layer. The middle layer is responsible for information transformation and it can be designed as a single hidden layer or multi-hidden layer structure according to the requirements of information change capabilities. The last hidden layer passes all information to the neurons of output layer. Information result is exported from the output layer to the outside after the further processing. By the time the forward propagation processing of learning is accomplished. When the actual output and desired output are contradictory, the error back propagation phase begins.

The error modifies weight of each layer according to the mode of error gradient descent through the output layer and goes along back-propagation layer by layer to the hidden layer, input layer. Information forward-propagation and error back-propagation process go round and round. It is not only the process of adjustment the weight value of each layer continuously but also the process of neural network learning and training. This process continues until the error of the network output decrease to an acceptable level or achievement the pre-established learning times.

BP neural network algorithm is quite robust to noisy testing dataset. The learning time is relatively short and evaluating the learning network is typically very fast. It constructs a network with credibility evaluated interaction protein by multilayer feedback mechanism. The algorithm iteratively compares the predicted class of the protein-protein interaction pair with known actual class [1].

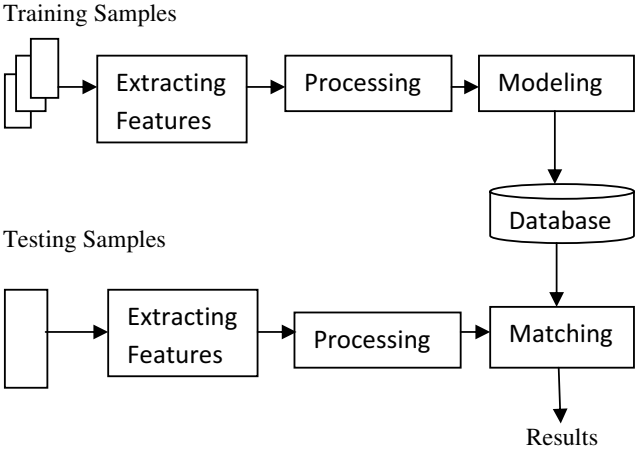

Fig 2: Flow Chat for Protein-Protein interactions

V. ANALYSIS AND RESULTS

A. ANALYSIS

Lishuang LI, Linmei JING, Degen HUANG [2] conducted their experiments on BC-PPI corpus. They extracted 1870 co-occurrences form 1000 sentences out of which 255 were positive instances i.e. they were interacting pairs and 1615 were negative instances i.e. they were non-interacting pairs. They used two main features to predict protein-protein interactions and they are Word Feature (WF), Distance Feature (DF). Based on these two features they calculated Specificity, Sensitivity and Accuracy. These three are defined by them as follows:

$$\text{Specificity} = \frac{TP}{TP+FN}; \quad \text{Senstivity} = \frac{TP}{TP+FP}; \quad \text{Accuracy} = \frac{2\alpha\beta}{\alpha+\beta}; \quad (3)$$

SEEK DIGITAL LIBRARY

Where TP (True Positive) is the number of co-occurrences classified correctly as positive; FN (False Negative) is the number of positive co-occurrences that are classified as negative incorrectly by the classifier; FP (False Positive) is the number of negative co-occurrences that are classified as positive incorrectly by the classifier; F-score is the harmonic value of recall and precision.

Zhiqiang Ma, Chunguang Zhou, Linying Lu, Yanan Ma et al [3] have shown BP neural network performs well above 87% accuracy rate through 10 cross-validation. According to them 2000 sequences that came from Scerevisiae yeast dataset are classified and out of which 1780 sequences were classified correctly.

B. RESULTS

1) Protein-Protein Interaction Prediction Based on SVM

Lishuang LI, Linmei JING, Degen HUANG [2] obtained their results word feature, word feature and distance feature respectively. The results obtained by them are shown in the table 1 below. It shows that distance feature contributes a lot in improving the performance in comparison to the results of word feature only.

Table1: Results based on SVM

	Specificity	Sensitivity	Accuracy
WF	74.2%	50.6%	60.1%
WF+DF	84.2%	63.9%	72.2%

2) Protein- Protein Interaction Prediction Based on SVM-KNN Classifier

Li R., Ye S. W., Shi Z [11] combined SVM with KNN to improve the accuracy of SVM classifier. Lishuang LI, Linmei JING, Degen HUANG [2] used SVM-KNN classifier to classify Protein-Protein interactions. They combined SVM with KNN algorithm to classify the samples which are near the hyper plane of the SVM in feature space.

Table 2: Results based on SVM-3NN Classifier using WF

	Specificity	Sensitivity	Accuracy
$\varepsilon = 0.1$	74.2%	53.6%	62.2%
$\varepsilon = 0.2$	75.8%	55.8%	64.3%
$\varepsilon = 0.3$	76.7%	60.5%	67.7%
$\varepsilon = 0.4$	77.5%	60.4%	67.9%
$\varepsilon = 0.5$	78.3%	67.1%	72.3%
$\varepsilon = 0.6$	79.2%	67.9%	73.1%
$\varepsilon = 0.7$	77.5%	63.7%	69.9%
$\varepsilon = 0.8$	76.7%	61.7%	68.4%
$\varepsilon = 0.9$	75.8%	61.2%	67.8%

They defined a threshold value of ε, if the distance is greater than ε then the test sample will be classified on SVM, Otherwise KNN algorithm will be used. The results obtained by Lishuang LI, Linmei JING, Degen HUANG in [2] are shown in table 2 (WF) and in table 3 (WF + DF).

The best results were obtained when ε have value between 0.5 and 0.6 with k=3. But when distance feature is also considered along with word feature its performance increases by 10.2% as shown in table 3 below [2].

Table 3: Results based on SVM-3NN Classifier using WF + DF

	Specificity	Sensitivity	Accuracy
$\varepsilon = 0.1$	85.0%	72.3%	78.2%
$\varepsilon = 0.2$	85.3%	73.1%	78.9%
$\varepsilon = 0.3$	85.8%	76.9%	81.1%
$\varepsilon = 0.4$	86.3%	78.1%	82.0%
$\varepsilon = 0.5$	86.7%	78.8%	82.5%
$\varepsilon = 0.6$	87.5%	79.5%	83.3%
$\varepsilon = 0.7$	86.1%	76.7%	81.1%
$\varepsilon = 0.8$	85.6%	75.5%	80.2%
$\varepsilon = 0.9$	84.4%	75.2%	79.5%

3) Protein-Protein Interaction Based on BP Neural Network

Zhiqiang Ma, Chunguang Zhou, Linying Lu, Yanan Ma et al [3] have taken 14511 entries from MIPS database that includes 7000 entries of training samples and 2000 entries of testing samples. They applied 10 testing sample dataset as cross validation. The results obtained by them are shown in table 4. They considered the specificity, sensitivity and accuracy of experimental results. The Specificity and Sensitivity can be obtained by following equations:

$$\text{Specificity} = \frac{TN}{TN+TP}\ ; \qquad \text{Sensitivity} = \frac{TP}{TP+TN}\ ; \qquad (4)$$

Where:
TP: are interacting Pairs
TN: are non-interacting Pairs

Table 4: Results Based on BP Neural Network

Cross-Validation	Specificity	Sensitivity	Accuracy
1	89.9	92.2	91.05
2	90.3	91.2	90.75
3	88.0	89.8	88.9
4	89.9	92.0	90.95
5	86.9	90.3	88.1
6	90.9	92.5	88.1
7	85.0	89.9	88.1
8	85.7	90.1	87.9
9	85.2	90.3	87.75
10	86.0	91.1	88.55

VI. CONCLUSION AND FUTURE SCOPE

Protein-Protein Interactions has been predicted using various classifiers till date. The main classifiers that have been used for predicting protein-protein interactions are SVM, SVM-KNN, BP Neural Network are compared in this paper. The objective

SEEK DIGITAL LIBRARY

of this paper was to analyze the various classifiers that have been used for protein-protein interactions. In conclusion the result of the study suggests that protein-protein interactions can be predicted well using BP Neural Network. This paper also shows the results conducted by different researchers using different classifiers. PPI prediction still remains a challenging task because performance of the best PPI systems is still far from user satisfaction. PPI prediction can be made better by making some changes in the existing classifiers. Hopfield Neural Network can be used for PPI classfication that will improve its accuracy from the previously defined classifiers for PPI prediction.

REFERENCES

[1] Zhiqiang Ma, Chunguang Zhou, Linying Lu, Yanan Ma et al. "Predicting Protein-Protein Interactions Based on BP Neural Network" IEEE 2007.

[2] Lishuang LI, Linmei JING, Degen HUANG "Protein-Protein Interaction Extraction from Biomedical Literatures Based on Modified SVM-KNN" IEEE 2009.

[3] A. Selim Aytuna, Attila Gursoy, Ozlem Keskin, "Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces" BIOINFORMATICS, Vol. 21 no. 12 2005.

[4] M. Pellegrini, E. M. MarCotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.", Proc. Natl. Acad. Sci. USA, April, 1999, vol. 96, pp. 4285-4288.

[5] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia,"Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes." Journal of Molecular Evaluation, 1997, pp. 66-73.

[6] A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events.", Nature, November 4, 1999, vol. 402, pp. 86-90.

[7] U. Ogman, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy, "PRISM: protein interactions by structural matching." Nucleic Acids Research, May 2, 2005, vol. 33, pp. 331-336.

[8] Hong-Wei Liu, "Protein-Protein Interaction Detection By SVM From Sequence Information" The Third International Symposium on Optimization and Systems Biology (OSB'09) Zhangjiajie, China, September 20–22, 2009

[9] Hongbo Zhu, Francisco S Domingues, Ingolf Sommer and Thomas Lengauer, "NOXclass: prediction of protein-protein interaction types" BMC Bioinformatics 2006

[10] Shinsuke Dohkan, Asako Koike, "Support Vector Machines for Predicting Protein-Protein Interactions" Genome Informatics 14: 502{503 (2003)

[11] Li R., Ye S. W., Shi Z., "A effective classified algorithm of support vector machine with multi-representative points based on nearest neighbor principle", Vol 3 IEEE 2001

[12] Vazquez, A., Flammini, A., Maritan A., and Vespignani A., 'Global protein function prediction from protein-protein interaction networks', Nat Biotechnol. 2003, 21, 697-700.

[13] Spirin, V., and Mirny, L. A., 'Protein complexes and functional modules in molecular networks', *PNAS 2003, 100, 12123-12128*

[14] Deng M., Tu, Z., Sun, F., and Chen, T., 'Mapping Gene Ontology to proteins based on protein-protein interaction data', Bioinformatics 2004, 20, 895-902.

[15] Min Su Lee, Seung Soo Park, Min Kyung Kim, " A Protein Interaction Verification System Based on a Neural Network Algorithm" IEEE (CSBW'05), 2005.

[16] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, Bernhard Schölkopf, "An Introduction to Kernel-Based Learning Algorithms" IEEE Transactions on Neural Networks, Vol. 12, NO. 2, March 2001

[17] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," Bioinformatics, vol. 17(5), pp: 455-460, 2001.

[18] Vapnik. V. N., "An overview of Statistical learning theory", Vol 10 Issue 5, IEEE 1999

[19] Hao Zhang, Alexander C. Berg, Michael Maire, Jitendra Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition"

[20] Li Fuliang, Gao Shuangxi, "Character Recognition System Based on Back-propagation Neural Network" DOI 10.1109 IEEE 2010

SEEK DIGITAL LIBRARY