Plant Tissue Cult. & Biotech. 20(1): 37-45, 2010 (June)

PTC&B

Predicting Protein-protein Interaction Using Amino Acid Sequence Information: A Computational Approach

Mohammad Shoyaib, M. Abdullah-Al-Wadud[1], Syed Murtuza Baker[2], Mohammad Nurul Islam[3] and Oksam Chae*

Department of Computer Engineering, Kyung Hee University, 1 Seocheon, Giheung, Yongin, Gyonggi, Korea

Key words: Protein-protein interaction, Amino acid sequence, Computational approach

Abstract

An improved computational approach which implements a protein-protein interaction prediction system based on the sequence information of a protein has been presented. A Support Vector Machine (SVM) is trained with this sequence information to predict the interactions. This interaction prediction technique exhibits 79.81% accuracy over a wide range of data, which is a significant improvement over other conventional computational protein-protein interaction prediction methods.

Introduction

Protein-Protein Interaction (PPI) is the fundamental mechanism that plays a vital role in biological function, DNA replication, immunologic recognition and progression through the cell cycle (Alberts et al. 1989). Predicting this interaction is thus becoming a focal point for researchers. Further understanding the function and the physiological role of proteins is fundamental to the discovery of novel medicinal and protein based products with medical and industrial application. Despite the high importance of recognizing the PPI, very little has been achieved so far, as the experimental approaches for PPI identification are both expensive and laborious. However, an improved computational approach can compliment experimental procedures with increased cost savings and greater confidence in the experimental results. A number of computational methods for predicting PPI, based on sequence or structure information, have already been developed that shows good accuracy. However, they still cannot achieve desired accuracy over a wide range of data. Further, for identifying

*Corresponding author: <oschae@khu.ac.kr>. [1]Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, 89 Wangsan, Mohyun, Cheoin, Yongin, Gyonggi, Korea.<wadud@hufs.ac.kr>. [2]Department of Computer Science & Engineering, East West University, Dhaka, Bangladesh.<galib@ewubd.edu>. [3]Department of Botany, University of Dhaka, Dhaka-1000, Bangladesh. <mnurul@univdhaka.edu>

residues in protein–protein interfaces researchers have attained high levels of accuracy. Although these methods attempt to identify all interface residues, one limitation is that they capture only a small fraction of them.

A large number of experimental techniques have been developed for the systematic analysis of PPI, such as yeast two-hybrid-based methods (Fields and Son 1989), protein chips (Zhu et al. 2001), photo-reactive amino acid analogs, Tandem affinity purification and many more (Tong et al. 2002). However, these experimental techniques are time consuming and very expensive, which is one of the main reasons that computational approaches have also been explored. A computational analysis of phylogenetic profiling has shown some success in predicting the PPI, but the limitations of this method is that this method includes the fact that it can only be used when a complete genome is available (Snitkin et al. 2006). Genes with closely related functions, encoding potentially interacting proteins are often transcribed as a single unit, an operon, in bacteria and are co-regulated in eukaryotes. Gene neighbour and gene clustering methods are being developed to discover details of these closely related interactions (Bowers et al. 2004). *In silico* method, which identifies the interaction, by arranging the two proteins based on the accumulation of signals in the proximity of interacting surfaces has also been reported (Pazos et al. 1997). The limitation here is the need for complete alignment, with a good coverage of species common to the two proteins under study.

Different classification methods have been successfully applied to identify protein interactions. Classifiers are trained with certain protein features and then ran with the test data. This classification approach has resulted in good accuracy being achieved (Qi et al. 2005, Chen and Liu 2005, Bader et al. 2004). Yan et al. (2002) developed an approach for computational prediction of protein-protein interaction sites using a SVM classifier where interface residues and non-interface residues with relatively high specificity (71%) and sensitivity (67%) were identified.

Data mining procedures are emerging for the automatic extraction of information about protein interactions from a large amount of already established protein interaction data. These procedures are applied to extract protein sequence signatures from existing protein-protein interaction data or to discover the stable and significant binding motif pairs from PDB complexes. The extracted data through these procedures is then applied to predict other protein interactions.

Information about the three dimensional (3D) structure of a protein reveals information about interface residues, but most of the time such high resolution information is not available (Recio et al. 2005). So Protein-protein docking, which is the method of determination of the molecular structure of complexes formed by two or more proteins without the need for experimental measurement is used to predict the interactions.

This paper proposes a prediction technique, based on sequence information of amino acid triplets and employing a support vector machine (SVM), covering a large range of organisms. The prediction accuracy achieved through utilizing this binary classifier SVM approach shows a significant improvement.

Materials and Methods

The protein interaction data was collected from http://dip.doe-mbi.ucla.edu a publicly accessible online database of interacting proteins (DIP). In order to test the proposed methodology, a classifier was trained to distinguish between the positive examples of truly interacting protein pairs and the negative examples of non-interacting protein pairs. As this classifier is a binary classifier it considers the interacting protein sequences to be the positive dataset and non interacting protein sequences to be the negative dataset.

The protein interaction refers to the association of protein molecules from the perspective of biochemistry, signal transduction and networks. Interacting sequences are considered as the positive dataset during classifier training. To develop the training dataset, a total number of 7,935 interacting protein sequence pairs, (two interacting proteins), of *D. melanogaster* were collected. Later the model was trained with the dataset of *E. coli, H. sapiens* and *C. elegans* to test the accuracy of the identification of protein interaction on those corresponding species. In these cases, the test data sets were also collected from the same source.

There is no "gold standard" dataset of non-interacting protein sequences, and there is also no database for non-interacting proteins. Therefore, researchers tend to adopt their own techniques in deriving the non-interacting protein sequences. These non-interacting protein sequences are considered as the negative dataset for classifier training. In our work, we picked non-interacting protein pairs from the interacting protein database, for which explicit interaction information is not found. For instance, if AB, BC and CE are three interacting protein pairs, then AC, BE, AE etc. may be considered as the candidates for non-interacting protein pairs (Ben-Hur and Noble 2005, Juwen et al. 2007). The desired number (same as the total number of interacting protein pairs in training set) of non-interacting pairs are selected uniformly at random from the set of all such candidate protein pairs. From the rest of the candidate non-interacting pairs, a random set of pairs are taken as test data. Following the above mentioned process, we picked interacting and non-interacting pairs for both training and test data set for *D. melanogaster*, *E. coli, H. sapiens* and *C. elegans.*

Support vector machine (SVM), a supervised machine-learning technique, has been used in this research to discriminate between interacting and non-interacting protein sequences. SVM classifiers solve multi-class classification problems using the structural minimization principle. In our experiment, we use LIBSVM (Chang and Lin 2001) with RBF kernel. Training is performed in a supervised manner on a collection of interacting and non-interacting protein training sequences. The model developed from the training set is then used to predict the interacting protein sequences from a pair of test sequences.

The main computational challenge in predicting PPI using protein sequences is to describe the important information residing within the sequences of amino acids. To address this problem, we use triad/triplet where any three continuous amino acids are considered as a unit. Thus, differentiation between classes is made according to the triplet, i.e. gamma interferon activation site/gamma interferon activation factor (GAS/GAF) is distinct and belong to different groups. Taking these three amino acids at a time and dividing the amino acids into seven classes a total of 7*7*7 = 343 different combinations are possible.

The properties of PPI can be described using a vector space R as (f_1, f_2, f_3, …, f_{343}), where f_i is defined in Eq. 1.

$$f_i = freq(r_i) \tag{1}$$

where, r_i denotes any specific type of triad/triplet and the function $freq(r_i)$ calculates the frequency of that specific r_i (number of times it occurs) in a protein sequence.

We represent a protein sequence x as a vector in this vector space using Eq. 2.

$$R_x = (f_{x_1}, f_{x_2}, f_{x_3}, \ldots, f_{x_{343}}) \tag{2}$$

To denote a pair of protein sequences ($P_{a,b}$), whether they interact or not, we simply concatenate the vectors representing them, which results in a 686 dimensional vector according to Eq. 3.

$$P_{a,b} = R_a \otimes R_b \tag{3}$$

where, ⊗ represents the concatenation of two vectors.

However, to distinguish interacting and non-interacting pair of protein sequences, we use $P^{+}_{a,b}$ if protein a interacts with protein b, and $P^{-}_{a,b}$ if they do not interact. Such 686 dimensional vectors are used as the positive dataset (interacting protein sequences) and the negative dataset (non-interacting protein sequences) in the SVM.

It is obvious that the frequency of the triads dependent on the length of the protein sequences. In general, a long protein sequence is likely to cause larger frequencies of the triads than a short protein sequence. The variation in lengths creates complications in the prediction accuracy. Hence, normalization is done according to Eq. 4 to neutralize the differences in the lengths of protein sequence. These normalized values are used in the vector representations.

$$\overset{\wedge}{f_i} = \frac{f_i - \min(f_1, f_2, ..., f_{343})}{\max(f_1, f_2, ..., f_{343})} \quad (4)$$

Results and Discussion

The properties of 20 different amino acids play a vital role in protein-protein interaction. This paper manipulates these properties by dividing them into different relevant groups. To extract the key features three consecutive amino acids were taken from the test sequence to make a triplet/triad. Each of those amino acids falls into its own category, depending on its physiochemical properties. First they were grouped into two major classes: hydrophobic, hydrophilic. Thus, any triplet falls into one of the 8 categories (as the amino acids are divided into two categories, total number of groups for a triplet will be 2*2*2=8) as demonstrated in Table 1. In this table, A_1 A_2 and A_3 fall are the first, second and third amino acid respectively in a triplet. G_1 and G_2 are used to indicate groups.

Table 1. Categorization of amino acid triplets when each amino acid is grouped into one of the two groups.

Triplet	Category	Description
$A_1A_2A_3$	$G_1G_1G_1$	All the three amino acids within the triplet fall into group 1.
	$G_1G_1G_2$	The first two amino acids from the triplet fall into group 1 and the last one falls into group 2.
	$G_1G_2G_1$	The first amino acid falls in group 1, the second one falls in group 2 and the third one again falls in group 1.
	$G_1G_2G_2$	The first amino acid falls in group 1 and rest of the two fall into group 2.
	$G_2G_1G_1$	The second amino acid falls in group 2 and the rest fall into group 1.
	$G_2G_1G_2$	The first amino acid falls in group 2, the second one in group.
	$G_2G_2G_1$	The first two amino acids fall in group 2 and the last one in group 1
	$G_2G_2G_2$	All the three amino acids fall in group 2.

Prediction accuracy based on hydrophobic and hydrophilic grouping is 59.88%. The hydrophilic group was further sub-divided into two sub-classes; charged and uncharged, in order to incorporate electrostatic property. In this experiment, the 20 amino acids fall into any one of the three categories. At this point, the accuracy level jumped to 72.08%. Next the charged groups were further divided into positively charged amino acid and negatively charged amino acids since 75% (Voet et al. 2005) of charged residues show strong interaction between oppositely charged members of an ion pair. With these groups 75.59% accuracy was achieved. Finally, the 20 amino acids were divided into seven classes based on seven different properties. These classes are hydrophobic, aromatic, hydrophilic, small hydrophilic, sulphahydral, positively charged and negatively charged. Amino acids within the same class are likely to involve synonymous mutations due to shared characteristics (Yan et al. 2002). The properties of different groups are summarized in Table 2. Based on this classification, the accuracy reached to 79.81% over a wide range of data.

Table 2. Properties of different amino acid groups.

Group characteristics	Group properties
Hydrophobic	Non charged side chain
Aromatic	Side chain contains aromatic ring system
Hydrophobic with long side chains	Side chain contains long uncharged group
Hydrophilic	Side chain contains small charged group
Sulfahydryl	Side chain contains sulfahydryl
Hydrophilic with Negative Charge	Side chain contains negatively charged and polar group
Hydrophilic with positive charge	Side chain contains positively charged and polar group

Table 3. Group based on hydrophobic and hydrophilic.

Group characteristics	Amino acid	Accuracy
Hydrophobic	G, A, V, L, I, M, F, W, P	59.88%
Hydrophilic	S, T, C, Y, N, Q, D, E, K, R, H	

Table 4. Group based on hydrophobic, hydrophilic and charged hydrophilic.

Group characteristics	Amino acid	Accuracy
Hydrophobic	G, A, V, L, I, M, F, W, P	72.08%
Hydrophilic	S, T, C, Y, N, Q	
Hydrophilic with charge	D, E, K, R, H	

To evaluate the prediction accuracy of the proposed methodology a sevenfold cross validation is used instead of Jackknife and bootstrapping (Good 2005) as it is computationally inexpensive and more efficient. Both bootstrapping and jackknife methods estimate the variability of a statistic from the variability of that statistic between sub-samples; as a result it incorporates the effect of self-influence. Whereas the cross validation is free from this self-influence as it splits the data into k subsets; each is held out in turn as the validation test. In sevenfold cross validation, the dataset divided into seven random parts. Each time it trained with six parts and tested with the single part. 8,000 protein sequences of *Drosophila melanogaster* were collected from Database of Interaction Protein (DIP) as the positive dataset and another 8,000 as the negative dataset were created using the method mentioned previously. The result is summarized in Tables 3-6.

Table 5. Group based on hydrophobic, hydrophilic and positively charged hydrophilic and negatively charged hydrophilic.

Group characteristics	Amino acid	Accuracy
Hydrophobic	G, A, V, L, I, M, F, W, P	75.59%
Hydrophilic	S, T, C, Y, N, Q	
Hydrophilic with negative charge	D, E	
Hydrophilic with positive charge	K, R, H	

Many different efforts have been made to predict PPIs in recent years. Jingchun et al. 2007 demonstrated In PrePPI to predict PPI in prokaryotic genome. These authors developed three methods, which they applied to different datasets. In their method, the highest accuracy achieved was 78%. Juwen et al. 2007 proposed PPI prediction technique using sequence information. In their technique they achieve 83.9 ± 1.29% accuracy but they achieved this by using a specific dataset. Sensitivity of 50% and specificity of 98% was achieved by Wan et al. 2002 for large scale statistical prediction of protein-protein interaction by using potentially interacting domain (PID) pairs. In comparison with these methods, the novel method demonstrated in this paper achieves both higher accuracy and coverage across data for diverse organisms.

Table 6. Group based on hydrophobic, aromatic, small hydrophilic, hydrophilic, sulfhydryl, positively charged hydrophilic and negatively charged hydrophilic.

Group characteristics	Amino acid	Accuracy
Hydrophobic	G, A, P	79.81%
Aromatic	F, W, Y	
Small hydrophilic	V, L, I, M	
Hydrophilic	S, T, N, Q	
Sulfhydryl	C	
Hydrophilic with negative charge	D, E	
Hydrophilic with positive charge	K, R, H	

So far we have demonstrated step by step performance improvement while we increase amino acid grouping. This in turn gives us an indication that such grouping is very important for PPI identification. Being inspired with this result, we have tested our idea for PPI detection over three different species across the kingdom, namely *E. coli, H. sapiens* and *C. elegans*. We have achieved similar performances in all these cases (data not shown). The results confirm our intuitive interpretation. Even though, in some cases, our proposed method shows little inferior accuracies compared to some existing works on specific data, it exhibits well generalized and consistent results over different datasets.

Successful and efficient prediction of protein-protein interaction can advance bio-medical research. This paper describes a methodology to predict PPI with high accuracy and good coverage of data types. This method is based on amino acid properties, which are taken as the motif to train the SVM. The main idea is to find all possible patterns grouped into several clusters depending on the physiochemical properties of amino acids, which predominantly appear in the pairs of interacting proteins. A classifier is then trained with these features to predict the association of protein sequences. Further efficiency could be achieved if gold standard negative data is obtained.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K and **Watson JD** (1989) Molecular Biology of the cell, 2nd Edition, Garland, New York.

Bader JS, Chaudhuri A, Rothberg JM and **Chant J** (2004) Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnology **22**: 78–85.

Ben-Hur A and **Noble WS** (2005) Kernel methods for predicting protein-protein interactions, Bioinformatics **21** (suppl. 1): i38-i46.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J and **Yeates TO** (2004) Prolinks: A database of protein functional linkages derived from coevolution. Genome Biol. **5**: R35.

Chang C-C and **Lin C-J**, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen XW and **Liu M** (2005) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics **21**: 4394-4400.

Fields S and **Son O** (1989) A novel genetic system to detect protein-protein interactions. Nature **340**: 245-246.

Gomez SM, Noble WS and **Rzhetsky A** (2003) Learning to predict protein-protein interactions. Bioinformatics **19**: 1875-1881.

Jingchun S, Yan S, Guohui D, **Qi L, Chuan W, Youyu H, Tieliu S, Yixue** L and **Zhongming Z** (2007) In PrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. BMC Bioinformatics **8**: 414.

Juwen S, Jian Z, Xiaomin L, Weiliang Z, Kunqian Y, Kaixian C, Yixue L and **Hualiang J** (2007) Predicting protein-protein interactions based only on sequences information. PNAS **11**(104): 4337-4341.

Pazos F, Citterich MH, Ausiello G and **Velencia A** (1997) Correlated mutations contain information about protein protein interaction. J. Mol. Biol. **271**: 511-523.

Qi Y, Seetharaman JK and **Joseph ZB** (2005) Random forest similarity for protein–protein interaction prediction from multiple sources. Pac. Symp. Biocomput. pp. 531-542.

Recio JF, **Totrov M, Skorodumov C** and **Abagyan R** (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins **58**: 134-143.

Snitkin ES, Gustafson AM, Mellor J, Wu J and **DeLisi C** (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics **7**: 420.

Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B and **Pauluzi S** (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295**: 321-324.

Voet D, Voet JG and **Pratt CW** (2005) Fundamental of Biochemistry, 2nd Edition, John Wiley.

Wan KK, Park J and **Jung KS** (2002) Large Scale Statistical Prediction of Protein-Protein Interaction by Potentially Interacting Domain (PID) Pair. Genome Informatics **13**: 42-50.

Yan C, Honavar V and **Dobbs D** (2002) Predicting Protein-Protein Interaction Sites From Amino Acid Sequence. Technical Report ISU-CS-TR 02-11, Department of Computer Science, Iowa State University.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidingmaier S and **Houfek T** (2001) Global analysis of protein activities using proteome chips. Science **293**: 2101-2105.

Good P (2005) Introduction to Statistics through Resampling Methods and R/S-Plus. John Wiley & Sons. ISBN 0-471-71575-1.