A Simple Approach for Predicting Protein-Protein Interactions

Mamoon Rashid, Sumathy Ramasamy and Gajendra P.S. Raghava*

Institute of Microbial Technology, Sector-39A, Chandigarh, India

Abstract: The availability of an increased number of fully sequenced genomes demands functional interpretation of the genomic information. Despite high throughput experimental techniques and *in silico* methods of predicting protein-protein interaction (PPI); the interactome of most organisms is far from completion. Thus, predicting the interactome of an organism is one of the major challenges in the post-genomic era. This manuscript describes Support Vector Machine (SVM) based models that have been developed for discriminating interacting and non-interacting pairs of proteins from their amino acid sequence. We have developed SVM models using various types of sequence compositions e.g. amino acid, dipeptide, biochemical property, split amino acid and pseudo amino acid composition. We also developed SVM models using evolutionary information in the form of Position Specific Scoring Matrix (PSSM) composition. We achieved maximum Matthews's correlation coefficient (MCC) of 1.00, 0.52 and 0.74 for *Escherichia coli*, *Saccharomyces cerevisiae*, and *Helicobacter pylori*, using dipeptide based SVM model at default threshold. It was observed that the performance of a prediction model depends on the dataset used for training and testing. In case of *E. coli* MCC decreased from 1.0 to 0.67 when evaluated on a new dataset. In order to understand PPI in different cellular environment, we developed species-specific and general models. It was observed that species-specific models are more accurate than general models. We conclude that the primary amino acid sequence based descriptors could be used to differentiate interacting from non-interacting protein pairs. Some amino acids tend to be favored in interacting pairs than non-interacting ones. Finally, a web server has been developed for predicting protein-protein interactions.

Keywords: Protein interaction, protein sequence, support vector machine, interactome, protein interaction prediction.

BACKGROUND

Proteins are essential macromolecules in living systems. They interact with each other to form protein complexes, which are essential for biological processes and cellular functions. Exploration the of interactome provides the detail about the cellular processes, signal transduction, metabolic pathway, regulatory process, quaternary structure prediction and the basis of biological system [1]. PPIs are important in modifying or designing a drug especially according to the nature of protein- protein interaction in disease associated pathways [2, 3]. Recently, a small molecule inhibitor MI-219 was designed against the p53-MDM2 interaction so as to make p53 functional, leading to induction of cell cycle arrests in all cells and selective apoptosis in tumor cells [4]. There are a number of experimental techniques for determining PPI, which includes co-expression data analysis, pull-down assays, coimmunoprecipitation, tandem affinity purification, two hybrid-based methods [5], Mass spectrometry [6], protein chips [7], binding reaction methods [8] and hybrid approaches [9]. These experimental techniques are costly and time consuming. There is, thus a need to develop computational techniques for predicting PPI on a larger scale.

A large number of computational techniques have been developed for PPI [10-12], which are based on different concepts such as phylogenetic profile [13, 14], conservation of gene neighborhood [15], gene fusion [16, 17], correlated mutations [18]. Other approaches use the signature product method [19] and pair wise kernel methods [20]. Genome context methods have also been used for predicting PPI, which includes phylogenetic profile method, frequency of co-occurrence in predicted operons, and distance between transcriptional start sites of two genes [21]. In some cases, both experimental data and prior knowledge were used for predicting protein interactions [22]. PPIs have also been predicted from the information about the domains, amino acid composition of proteins [23], and conjoint triad feature [24], pseudo-amino acid composition along with gene ontology (GO) annotation [25], and protein structural and physiochemical descriptors from sequence information [26, 27]. For predicting protein-protein interactions, mostly supervised machine learning methods (like support vector machine, random forest method [28] and Bayesian network) have been used. Despite tremendous progress in the field of PPI prediction, there are several major issues yet to be addressed.

Benchmarking

One of the challenges in the field of PPI is benchmarking of existing methods, as most of methods do not follow standard evaluation procedure (e.g. jackknife test or k-fold cross-validation). In addition datasets used in these methods do not have non-interacting pair of proteins (negative examples), which is important for fair evaluation. Ben-Hur and Noble (2005) compare their dataset with GO annotation. However there is a possibility that the GO database may have the same

*Address correspondence to this author at the Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India; Tel: +91-172-2690557 or 2690225; Fax: +91-172-2690632 or 2690585; E-mail: raghava@imtech.res.in; Web: http://www.imtech.res.in/raghava/

1389-2037/10 $55.00+.00

protein which has been used for training in their program. Thus, it is difficult to reflect the actual performance of this method.

Web Server

Implementation of the method/program as a web-server is another important issue from a user point of view. Most of the methods for predicting PPI are based on complex theories like gene fusion technique, correlated mutations, similarity of phylogenetic profiles etc., so it's difficult to develop web server using these methods.

Negative Dataset

Creation of negative data set is much more difficult than that of positive dataset. For example restricting negative examples to non co-localized protein pairs leads to a biased estimate of the accuracy of a predictor of PPI [29].

Redundancy between Training and Test Dataset

Developing independent training and test sets for protein-protein interactions is very challenging. Protein A might interact with 10 other proteins, and protein B might interact with many of those proteins, so protein A and B are not independent, so that all of the interactions with protein A and B would have to be in either the training set or the test set, but not some in one and some in the other; likewise, all 10 of protein A's interactions would have to be in one set, not the other.

Species-Specific or General Method

There is a need to understand whether PPI is environment dependent or independent, as different species have different cellular environment. If protein-protein interactions are independent of cellular surroundings, then we should develop a universal model for all organism otherwise separate model for each organism would be required.

The aim of this study is to develop a fast and reliable method for predicting pair of interacting proteins and to address some of the issues discussed above. In this study, we have developed simple composition based SVM models for discriminating interacting and non-interacting pair of proteins. In the past, composition based SVM-models have been successfully used for predicting subcellular localization of proteins [30-35]. The SVM models for predicting PPI are based on wide range of compositions that include amino acid, dipeptide, and biochemical property. All models developed in this study have been evaluated using 5-fold cross-validation technique and each dataset have positive as well as negative examples. In order to understand nature of PPI in different cellular environments, species-dependent and independent models have been developed. It was observed that model developed for an organism is only valid for that organism; suggesting interaction between proteins is organism dependent. A web server has been developed for prediction of protein interactions. Finally, we predicted interacting pairs of proteins (or interactome) in *S. cerevisiae* and *H. pylori* using our best SVM models at a threshold, which gives more than 90% precision. We also identified amino acids and dipeptides important for protein interactions.

RESULTS

It has been shown in the past that the subcellular localization of a protein can be predicted from their amino acid composition. In this study, we have extended the same concept to predict interacting and non-interacting pair of proteins from their amino acid composition. In case of amino acid composition, for e.g., a pair of protein is represented by a vector of dimension 40, each protein by a vector of dimension 20. We have developed separate SVM models for each organism (*E. coli*, *S. cerevisiae*, *H. pylori*).

Composition based Methods

We developed SVM based models for predicting interacting pairs of proteins using their amino acid composition. Table **1** shows the percent of correctly predicted pair of interacting proteins (sensitivity) and probability of correct prediction of interacting pairs (PPV-positive predictive value) at different thresholds (Table **S1** for details). SVM assigns a score for each pair of proteins; we assign a pair as interacting pair if it has score more or equal to a value called threshold value. As shown in Table **S1**, we achieved MCC 0.98, 0.39 and 0.63 for *E. coli*, *S. cerevisiae* and *H. pylori* respectively at default threshold 0.0. These results indicate that simple composition based SVM models can be used to discriminate interacting and non-interacting pair of proteins with reasonable accuracy. It has been shown in past that dipeptide composition provides more information than simple amino acid composition and can be used to predict function of a protein [30, 36]. Thus we developed SVM model for predicting PPI using dipeptide composition (see materials and methods for detail). As shown in Table **1** & **S1**, we achieved MCC 1.00, 0.52 and 0.74 for *E. coli*, *S. cerevisiae* and *H. pylori* respectively. The average accuracies of 99.9%, 75.7%, and 86.8% have been achieved for *E. coli*, *S. cerevisiae*, and *H. pylori* respectively. These results indicate that performance of SVM models based on dipeptide composition is better than other composition based models. Receiver operating characteristic (ROC) curves in Fig. (**1**) and Fig. (**2**) show the same pattern for *S. cerevisiae* and *H. pylori* dataset respectively. All models were trained, tested and evaluated using five-fold cross validation technique.

We developed SVM models using split amino acid (SA) composition and achieved performance better than amino acid composition based model and slightly lower than dipeptide based model (Table **2** and Table **S2**). Pseudo-amino acid based SVM model performed comparable to SA, in *E. coli* dataset, and better than SA in *S. cerevisiae* and *H. pylori* datasets (Table **2**).

In order to understand the role of amino acids in interactions, we computed the average compositional biasness (ACB) of each amino acid type (see materials and methods). The magnitude and direction of the ACB value represents how strongly this feature is favored towards interacting or non-interacting proteins. Positive and negative signs of ACB show dominancy of that feature towards interacting and non-interacting pairs, respectively. Some of the values along with feature name (amino acid) have been given in Table **S3**. In a similar fashion we averaged the values of ACB of same amino acid from proteins of a pair (interacting or non-interacting) and represented in Table **S4**. Similarly the im-

Table 1. Performance of PPI Prediction Method based on Amino Acid and Dipeptide Composition Using SVM 5-Fold Cross Validation Technique

Threshold	*E. coli*				*S. cerevisiae*				*H. pylori*			
	AA		DP		AA		DP		AA		DP	
	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV	Sen	PPV
1.0	56.5	100.0	79.0	100	26.3	90.0	28.2	94.9	32.4	95.7	30.0	98.2
0.8	70.9	100.0	88.4	100	35.1	87.5	38.8	93.2	45.2	95.0	48.6	97.5
0.6	81.1	100.0	93.2	100	44.2	83.3	48.9	90.3	55.9	92.4	63.0	95.6
0.4	88.7	100.0	96.2	100	52.7	78.4	57.8	87.4	66.0	88.0	74.7	92.8
0.2	93.0	100.0	98.1	100	60.9	74.0	66.0	82.5	75.3	85.1	82.4	90.2
0.0	95.7	99.9	99.2	100	69.1	69.9	72.8	77.3	82.5	80.5	88.5	85.7
-0.2	97.9	99.9	99.4	100	76.2	65.6	79.1	71.3	88.4	75.4	92.9	80.2
-0.4	99.0	99.7	99.6	100	82.4	61.6	85.0	65.4	93.2	70.0	96.5	74.0
-0.6	99.4	99.4	99.7	100	87.8	58.2	90.0	60.2	95.8	64.6	98.3	66.6
-0.8	99.6	92.1	99.7	100	92.0	55.6	94.1	55.9	97.9	58.5	99.3	58.3
-1.0	100.0	28.4	99.7	36.5	95.2	53.4	96.9	52.8	99.1	54.7	99.7	53.1

Where AA and DP are amino acid and dipeptide respectively; Sen is sensitivity and PPV is positive predictive value.

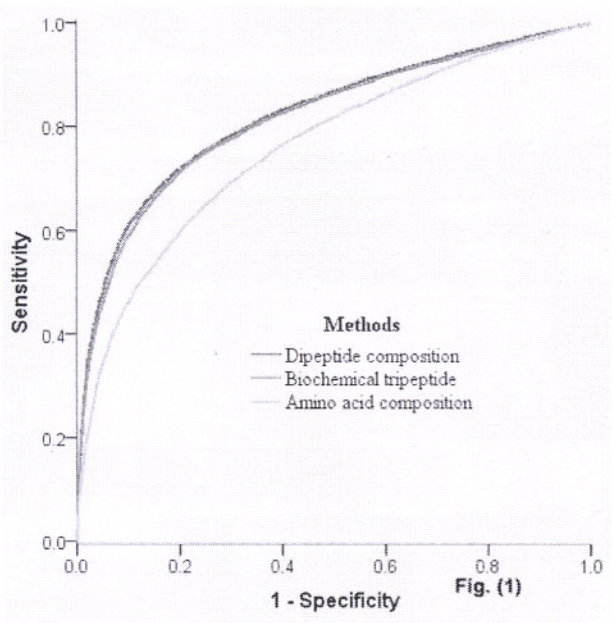

Fig. (1). The ROC curves showing the performance of different methods on *S. cerevisiae* dataset.

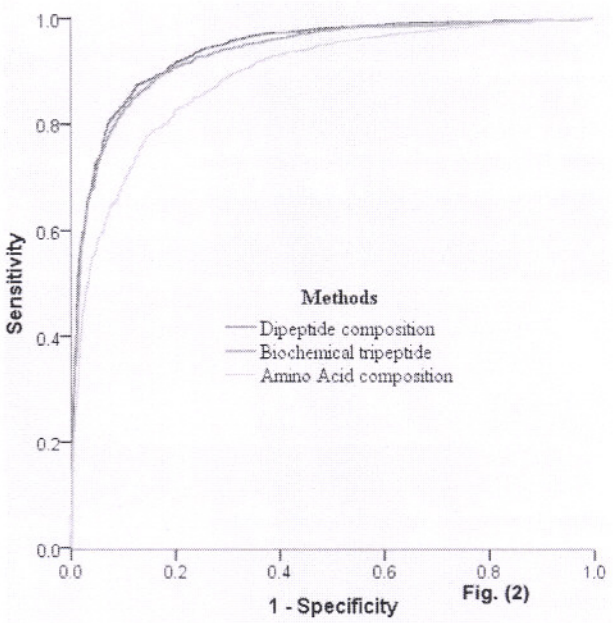

Fig. (2). The ROC curves showing the performance of different methods on *H. pylori* dataset.

portant dipeptides have been shown in Table **S5** and **S6**. The amino acid and dipeptide profile obtained from Table **S4** and S6 respectively suggested that the features which are present in *E. coli* interaction pairs (features corresponding to positive ACB values) are favored in *H. pylori* and *S. cerevisiae* non-interaction data (features corresponding to negative ACB values). Moreover, *S. cerevisiae* and *H. pylori* have similar features among their interacting and non-interacting pairs. To summarize *E. coli* has different amino acid sequence specific signatures in comparison to *S. cerevisiae* and *H. pylori*, whereas the signatures of the latter two are more or less similar.

Table 2. Comparison of Performances of Various SVM Modules

Methods	*E. coli*				*H. pylori*				*S. cerevisiae*			
	Sen	Spe	Acc	MCC	Sen	Spe	Acc	MCC	Sen	Spe	Acc	MCC
SA	94.5	100.0	99.6	0.97	85.3	86.8	86.0	0.72	70.8	76.6	73.7	0.47
PA	96.5	100.0	99.7	0.98	82.5	78.5	80.5	0.61	68.0	72.2	70.1	0.40
BM	92.4	99.9	99.4	0.95	77.1	73.5	75.3	0.51	64.9	65.8	65.3	0.31
BD	97.3	100.0	99.8	0.98	84.4	82.2	83.3	0.67	70.2	72.6	71.4	0.43
BT	97.2	100.0	99.8	0.99	87.8	84.6	86.2	0.72	72.8	78.1	75.5	0.51
PSSM	-	-	-	-	83.0	82.0	82.5	0.65	-	-	-	-
HB	-	-	-	-	86.0	88.8	87.4	0.75	71.2	79.1	75.2	0.50

Where Sen, sensitivity; Spe, specificity; Acc, accuracy; MCC, Matthews correlation coefficient; SA, amino acid composition of four equal parts of sequence; PA, pseudo amino acid composition; BM, biochemical amino acid composition; BD, biochemical dipeptide composition; BT, biochemical tripeptide composition; PSSM, Position Specific Scoring Matrix and HB is dipeptide concatenated with BT

Biochemical Composition and Evolutionary Information

In order to explore the effect of biochemically similar amino acid patches on interaction prediction, we first converted the 20 amino acid residues into six classes [37] for all sequences in the dataset. Further, amino acid, dipeptide and tripeptide compositions have been computed on the converted sequence alphabet. In a similar study [38] tuple of 4 from possible 6^4 (1296) tuple types was used to characterize protein interaction pairs for prediction of interaction. We have reported the biochemical monopeptide (BM), dipeptide (BD) and tripeptide (BT) compositions based methods (Table **2**) in this study. The detailed results of BM and BT have been included in Table **S7**. The result of the method based on BT is comparable to that of classical dipeptide composition, suggesting that local order of amino acids might act as a signature characterizing the interacting protein pairs. Moreover, the lists of important biochemical tripeptides (as estimated by ACB values) have been given in Table **S8** and Table **S9**.

In order to exploit evolutionary information encoded in protein sequences, we have calculated PSSM for each protein in a pair, and presented a binary vector of length 800 for every interacting and non-interacting pairs. The performance of PSSM based method has been depicted in Table **2** and Table **S10**. We have also tried to combine various descriptors mentioned in this study to represent binary vector and developed hybrid SVM based prediction methods. One of the hybrid methods (HB), that concatenated dipeptide and BT, performed comparable to (in *S. cerevisiae*) and slightly better than (in *H. pylori*) dipeptide composition based method (Table **2** and Table **S10**). At the completion of the present study, a high-quality binary interaction data set of the *S. Cerevisiae* interactome was published in literature [39] (Vidal's *S. cerevisiae* data set). We thought to apply our prediction method on Vidal's *S. cerevisiae* data set. After 5-fold cross-validation technique using Vidal's *S. cerevisiae* data set accuracies of 82.4%, 88.1%, and 86.2% were achieved for amino acid, dipeptide, and biochemical tripeptide composition. The detail result has been presented in Table **3**. Comparison of Table **3** with Table **1** and Table **2** shows remarkable increment in the performance using the Vidal's *S. cerevisiae* data set. Fig. (**3**) shows the ROCs plot for Vidal's *S. cerevisiae* dataset using amino acid, dipeptide, and biochemical tripeptide compositions.

Comparison with Existing Methods

We have compared our method directly with three other methods developed in the past. Table **4** shows some statistics on the working datasets and presents the comparison of our method with previous ones on the same dataset. Our method outperformed other existing methods on *E. coli* and *H. pylori* datasets. As evident from an excerpt in introduction section of this article, it is difficult to compare our *S. cerevisiae* model with that of Ben-Hur and Noble (2005). Therefore, we compared the performance of our *S. cerevisiae* model on validation set obtained from Pitre *et al.* (2006) [40] (Table **4**).

All the existing methods (considered here for comparison) used diverse type of descriptors for protein interaction prediction. Yellaboina *et al.* (2007) employed genome context methods (such as distance between transcriptional start site, phylogenetic profile and frequency of co-occurrence in operons), and both Martin *et al.* (2005) and Ben-Hur and Noble (2005) used sequence-based kernel methods (signature product, motif, Pfam, spectrum etc.) implemented within a support vector machine classifier. Therefore, the results in Table **4** suggest that our method is capable of resolving the PPI prediction problem with greater success.

Dataset vs. Performance

Existing literatures in the field of PPI suggested that the accuracy of a prediction method also depends on the type of negative dataset. So far we have shown the performances of our method on the original datasets obtained from their respective sources. Now, in order to explore the variation in performances with variation in datasets, we have designed positive and negative interaction datasets as follows.

Variation in Non-Interacting Dataset

As the performance of our method on the *E. coli* dataset was exceptionally high (Table **1** and **2**), we thought of exploring this issue in depth. We designed some experiments to

Table 3. Performance of 5-Fold Cross-Validation on Vidal's *S. cerevisiae* Data Set

Threshold	Methods														
	Amino Acid					Dipeptide					Biochemical Tripeptide				
	Sen	Spe	Acc	PPV	MCC	Sen	Spe	Acc	PPV	MCC	Sen	Spe	Acc	PPV	MCC
1	33.9	97.4	65.7	93.0	0.41	31.4	99.6	65.5	98.8	0.42	28.5	99.7	64.1	98.9	0.40
0.8	46.2	95.5	70.9	91.2	0.48	46.4	99.2	72.8	98.4	0.54	43.0	98.9	71.0	97.6	0.51
0.6	57.3	93.5	75.4	89.7	0.54	60.1	98.3	79.2	97.2	0.63	55.9	97.7	76.8	96.1	0.59
0.4	68.1	90.2	79.1	87.4	0.60	72.4	96.2	84.3	95.1	0.71	68.5	95.8	82.2	94.2	0.67
0.2	76.2	86.1	81.2	84.6	0.62	81.1	93.6	87.4	92.7	0.75	78.9	92.1	85.5	90.9	0.72
0.0	83.8	81.0	82.4	81.5	0.65	87.6	88.7	88.1	88.5	0.76	85.8	86.6	86.2	86.5	0.72
-0.2	88.8	73.6	81.2	77.1	0.63	91.2	81.0	86.1	82.7	0.73	91.2	78.0	84.6	80.6	0.70
-0.4	92.6	64.9	78.7	72.5	0.60	95.1	68.4	81.8	75.1	0.66	94.8	63.9	79.4	72.4	0.62
-0.6	95.4	48.7	72.0	65.0	0.50	97.4	51.6	74.5	66.8	0.55	97.4	48.4	72.9	65.4	0:53
-0.8	97.5	33.3	65.4	59.4	0.40	98.8	34.2	66.5	60.0	0.43	98.9	30.5	64.7	58.7	0.40
-1	98.8	21.1	60.0	55.6	0.31	99.6	19.0	59.3	55.2	0.31	99.6	17.6	58.6	54.7	0.30

Sen, Spe, Acc, and MCC are sensitivity, specificity, accuracy, and Matthews correlation coefficient, respectively.

Table 4. Comparison with Existing Methods

Dataset (N/P)	Dataset Source	Measures	Comparison of Methods	
			Source Methods	Our Method
E. coli (13840/1082)	Yellaboina et al. 2007	Sen	79.0	99.2
		Spe	100.0	100.0
		Acc	89.5	99.9
H. pylori (1458/1458)	Martin et al. 2005	Sen	79.9	88.5
		Pre	85.7	85.7
		Acc	83.4	86.8
S. cerevisiae (100/100)*	Pitre et al. 2006 (Table 1 and Table 2)	Sen	61.0	64.0
		Spe	89.0	98.0
		Acc	75.0	81.0

Where N/P, number of non-interacting pairs/number of interacting pairs; Sen, sensitivity; Spe, specificity; Acc, accuracy; and Pre is precision. The dataset marked * was not used for model development in our study rather served as validation set.

show the effect of choosing another negative dataset on the performance of the classifier. The original dataset comprised of negative interaction pairs from two different cellular locations (non-colocalized). It has been shown that restricting negative examples to non co-localized protein pairs leads to a biased estimate of the accuracy of a predictor of PPI [29]. As discussed in detail in materials and methods, two additional versions of *E. coli* negative dataset (random and non-redundant) have been prepared in this study and optimized SVM for performance measurement. The results in Table **5** indicate the marked reduction in performance in case of random and non-redundant negative dataset in comparison to the original (non-colocalized) negative dataset. ROC plots for *E. coli* random negative dataset in Fig. (**4**) also show the same pattern as observed in other ROC plots for *S. cerevisiae* and *H. pylori* datasets. These results are in agreement with those from the previous literature [29].

Non-Redundant Proteins in Interacting Pairs

We have further investigated the effect of distribution of interacting pairs of proteins among 5 sets in 5-fold CV (cross-validation) on the performance of the classifier taking

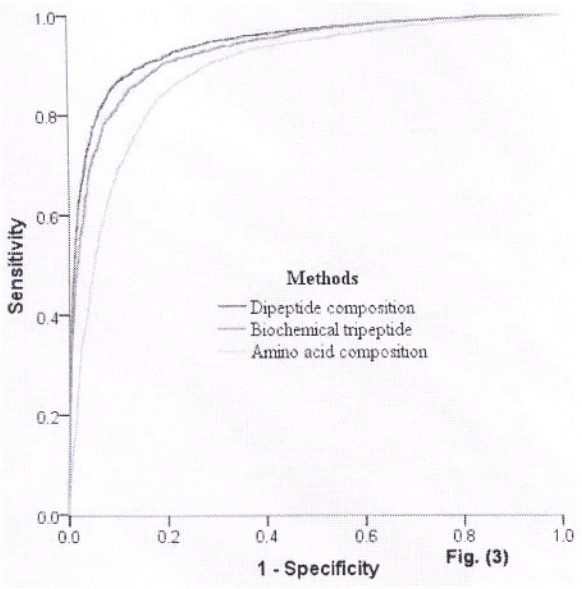

Fig. (3). The ROC curves showing the performance of different methods on Vidal's *S. cerevisiae* dataset.

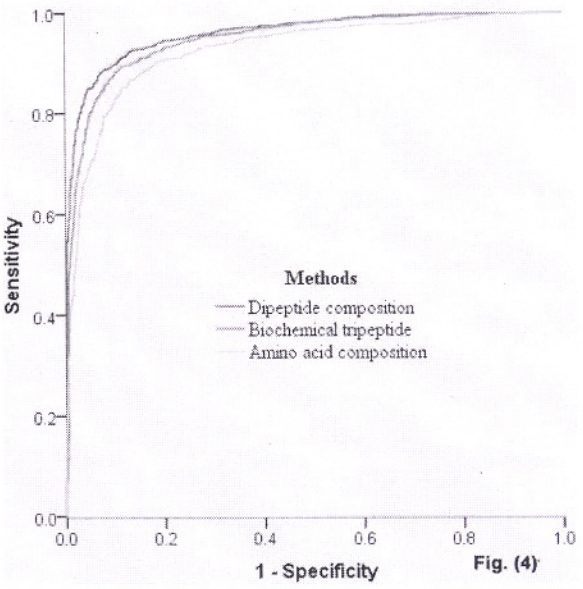

Fig. (4). The ROC curves showing the performance of different SVM models on *E. coli* dataset.

Table 5. Performance of SVM Models on Different *E. coli* Non-Interacting (Negative) Datasets

Descriptors	Type of negative dataset	Sen	Spe	Acc	MCC
Amino Acid Composition	Non-colocalized	97.8	98.9	98.3	0.97
	Random	88.5	83.5	86.0	0.72
	Non-redundant	86.1	81.4	83.8	0.68
Dipeptide Composition	Non-colocalized	99.0	99.7	99.4	0.99
	Random	91.0	88.4	89.7	0.79
	Non-redundant	89.2	85.2	87.2	0.75
Biochemical Tripeptide Composition	Non-colocalized	98.2	99.3	98.8	0.98
	Random	89.6	86.6	88.1	0.76
	Non-redundant	88.3	83.4	85.8	0.72

Sen, Spe, Acc, and MCC are sensitivity, specificity, accuracy, and Matthews correlation coefficient, respectively.

E. coli dataset. Earlier in all experiments we used random equal distribution of interacting pairs in five-fold cross-validation. Now, positive examples (i.e. interacting pairs) have been clustered such that almost all interactions of a protein, suppose A, remain in one set. In this way interacting pairs of any two sets have no common protein, thus facilitating the non-redundancy in training and test sets. This make up of positive dataset, that we have called clustered positive pairs, will certainly reduce the bias in performance during CV and results in Table **6** supported this assumption.

Feature Selection

Though the SVM models were successful in discriminating interacting and non-interacting pairs, they do not provide any information about amino acid residue, dipeptide or biochemical tripeptide involved in interaction. This is a major problem with most machine learning techniques; they work like a black box. For example, in our dipeptide based model we have 800 features (400 for each protein); a user may wish to know the important dipeptides contributing in interaction. This is not only important to understand interaction between two proteins but also for reducing the number of features used in model development.

Based on the ACB (see materials and methods) value, features have been selected for model development. The comprehensive result has been presented in Table **7**. The performance of the method approached its maximum value (when all the features have been included in model develop-

Table 6. Performance of SVM Models on *E. coli* Dataset* where Training and Test Sets Do Not have Redundant Protein in Interacting Pairs

Descriptors	Type of Negative Dataset	Sen	Spe	Acc	MCC
Amino Acid Composition	Non-colocalized	94.7	99.2	97.0	0.94
	Random	75.8	84.8	80.3	0.61
	Non-redundant	74.1	82.8	78.5	0.57
Dipeptide Composition	Non-colocalized	97.3	99.7	98.5	0.97
	Random	80.0	89.9	85.0	0.70
	Non-redundant	77.0	85.4	81.2	0.63
Biochemical Tripeptide Composition	Non-colocalized	97.0	99.3	98.2	0.96
	Random	79.5	87.3	83.4	0.67
	Non-redundant	77.3	84.1	80.7	0.62

Sen, Spe, Acc, and MCC are sensitivity, specificity, accuracy, and Matthews correlation coefficient, respectively. * *E. coli* dataset consist of 1082 interacting and 1082 non-interacting protein pairs.

Table 7. Performance of SVM Models Developed on Selected Features of Different Descriptors Using E. coli Interaction Dataset*

Descriptors	Selected Features		Sen	Spe	Acc	MCC
	ACB +	ACB -				
Amino Acid Composition	5	5	74.8	74.8	74.8	0.50
	10	10	88.0	75.0	81.5	0.64
	15	15	80.2	87.8	84.0	0.68
	All	All	88.5	83.5	86.0	0.72
	Feature Set¶		83.6	84.4	84.0	0.68
Dipeptide Composition	10	10	70.9	83.5	77.2	0.55
	20	20	80.2	86.0	83.1	0.66
	30	30	82.3	85.2	83.7	0.68
	40	40	81.4	81.3	81.4	0.63
	All	All	91.0	88.4	89.7	0.79
Biochemical Tripeptide Composition	10	10	77.4	73.3	75.3	0.51
	20	20	80.5	79.6	80.0	0.60
	30	30	84.5	82.8	83.6	0.67
	40	40	84.8	83.1	83.9	0.68
	All	All	89.6	86.6	88.1	0.76

Where * is dataset having 1082 random non-interacting pairs and 1082 interacting pairs, and Sen, Spe, Acc, and MCC are Sensitivity, Specificity, Accuracy, and Matthews Correlation Coefficient. ACB+ and ACB- are number of features having positive and negative ACB values respectively. "All" means total number of features for that descriptor, given in Table 5. ¶ Features selected by WEKA software using wrapper evaluation and genetic search methods. 17 features have been selected

ment, denoted as "All" in Table 7) as numbers of selected features were increased. For example, in case of dipeptide composition only 80 features (40 having positive and 40 negative ACB values) gained accuracy and MCC of 81.4% and 0.63 with respect to its maximum performance of 89.7% and 0.79 when all 800 features were included for model development. We also applied standard feature selection algorithm using wrapper evaluation and genetic search methods implemented in WEKA (Waikato Environment for Knowledge Analysis). The total 17 amino acids have been selected on which SVM CV performed with 84.0% accuracy ("Feature set" in Table 7).

Cross-Species Prediction

So far we have developed SVM models for each organism separately. Current literature in subcellular location prediction suggested that organism specific prediction methods are better than generalized ones [30, 36, and 41]. We also hypothesized that cellular environment affects the interaction between proteins. Therefore, to full proof this hypothesis we tried to classify interaction dataset of one species on the model developed on another species, and vice-versa. The results (in Table **8**) pointed out that the cross-species classifications have shown remarkable low performance than their corresponding one within the species. By critical observation of Table **S4**, it came to light that some amino acid residues (valine, arginine, glycine, leucine, isoleucine, phenylalanine and proline) which were favored in *E. coli* interaction data (positive ACB value) were present in *S. cerevisiae* and *H. pylori* non-interaction data (negative ACB value). Moreover, *S. cerevisiae* and *H. pylori* interaction data have similar amino acid profile in Table **S4**. These observations can be justified by the results of Table **8**, where the *S. cerevisiae* and *H. pylori* interactions could be predicted poorly on *E. coli* model and to a substantial performance among themselves. Anyway the species specific models (*E. coli*-to-*E. coli*, *H. pylori*-to-*H. pylori*, and *S. cerevisiae*-to-*S. cerevisiae*) performed better than the cross-species ones.

In a similar study Martin *et al.* (2005) also showed that the prediction of protein interaction of one organism on another organism's interaction data resulted in a poor performance. These findings suggested the unique makeup of protein interaction profile maintained in a species.

Interactome Prediction

In order to evaluate any bioinformatics method, normally MCC value is maximized, as it takes care of over- and under-prediction. We computed the performance where sensitivity and specificity come close to each other in order to make balance in prediction while keeping a high MCC value. Though, theoretically this is a logical way to evaluate any prediction model, it is not readily acceptable to experimental biologists. Biologists are much more interested in probability of correct prediction of positive examples (see PPV values in Table **1**) rather than high sensitivity or MCC values. Therefore, for predicting interactome we have selected the threshold at very high PPV value. Based on the SVM models developed exploiting dipeptide composition, we have predicted the interactome of *S. cerevisiae* and *H. pylori*. Since, *E. coli* dataset was comprised of functional interactions; we didn't predict *E. coli* interactome. Out of approximately 17 million, and 1.2 million all possible binary interactions (excluding self interaction) for *S. cerevisiae*, and *H. pylori* respectively, the number of predicted interactions are 196139, and 17233.

Web-based Prediction Server

Some existing web-servers for PPI prediction are Protein-Protein Interaction Prediction [42], InterPreTS [43], Protein-Protein Interaction Prediction Server [23], PIPE [40] etc. that accept protein sequences as input. We have also implemented our method as a web-server called "ProPrint" (http://www.imtech.res.in/raghava/proprint). Some of the existing methods have very low coverage while others take a lot of time (up to 400 hrs in case of PIPE) to predict a binary prediction. We have tested the performance of our prediction server on the validation set provided by Pitre *et al.* (2006) (Table **4**). In comparison "ProPrint" is fast (taking few seconds for a binary prediction), reliable, and accurate.

DISCUSSION

We demonstrated that PPI can be predicted by using simple compositional values, such as amino acid and dipeptide composition in diverse organisms. To the best of our knowledge we used dipeptide composition for the first time for PPI prediction. Also, biochemical classes' composition was also proved to be capable of predicting PPI (Table **2** and Table **S7**). We conclude that the sequence-based protein-protein interaction signature/profile is by and large species specific

Table 8. Comparison of Performance between Prediction within a Species and Cross-Species Prediction Using Dipeptide Composition Feature and SVM

Cross-Species Prediction	Sen	Spe	Acc	MCC
E. coli-to- E. coli	99.2	100.0	99.9	0.99
E. coli-to-S. cerevisiae	52.8	46.4	49.6	-0.01
E. coli-to-H. pylori	48.6	52.5	50.5	0.01
S. cerevisiae-to-S. cerevisiae	72.8	78.6	75.7	0.52
S. cerevisiae-to-E. coli	53.4	66.9	65.9	0.11
S. cerevisiae-to-H. pylori	53.2	54.4	53.8	0.08
H. pylori-to-H. pylori	88.5	85.2	86.8	0.74
H. pylori-to-E. coli	40.6	76.6	74.0	0.10
H. pylori-to-S. cerevisiae	65.8	46.1	55.9	0.12

Sen, sensitivity; Spe, specificity; Acc, accuracy; and MCC is Matthews Correlation Coefficient. Cross-species prediction, let A-to-B is prediction of organism B's interaction on organism A's model.

(Table **8** and Table **S4**). We have also shown that PPI prediction performance greatly depend on the type of negative dataset used for SVM optimization (Table **5**). To reduce the redundancy between any two sets during 5-fold CV, we clustered positive examples such that almost all the interactions of a particular protein remain in one set. By comparing the performance of the method after clustering (Table **6**) with that of non-clustered positive example (Table **5**), we come to a conclusion that clustering reduces the rate of correct prediction of positive examples (sensitivity). Our feature selection experiment concludes that the wrapper based method outperformed our ACB based method (Table **7**). Using amino acid composition, wrapper selected only 17 features compared to 30 features (15 ACB+ and 15 ACB-) by ACB based method while retaining the performance (of 84.0% accuracy) (Table **7**). The detailed results are in Table **7**.

We studied PPI prediction in great detail by utilizing the information from amino acid sequence alone. The performance of our method is better than existing methods (Table **4**) to predict PPI. Moreover, we discussed comprehensively the difficulties in predicting protein interactions- some of them are negative dataset selection, clustering of positive examples in 5-fold CV, cross-species prediction etc.

CONCLUSIONS

The current study suggested a simple and reliable way of predicting PPI. Amino acid sequence based descriptors are efficient in discriminating interacting pair from non-interacting one. Contribution of amino acid residues(s) in protein interaction follow different pattern in various organisms. Model developed on one organism could not be applied to predict interaction in another organism. Some amino acids contribute substantially in interaction while others do not. Our method shows better performance than existing genome context methods. Using our models, genome wide PPI prediction (interactome) has been achieved for *H. pylori* and *Saccharomyces cerevisiae*. Examination of existing PPI prediction methods and their complexity necessitated the development of our simple, efficient, and easy to apply method. ProPrint may be considered as complementary to other protein interaction prediction methods to get a more comprehensive and clear picture of the interactome.

MATERIALS AND METHODS

Our method is alignment free. Alignment free methods have successfully been used in prediction of subcellular location and function of a protein [30, 31, 36, 44-46].

Datasets

Protein interaction datasets have been created separately for three species namely *E. coli*, *S. cerevisiae* and *H. pylori*. We got the *E. coli* functional interaction dataset from the Yellabonia *et al.* (2007); it contains 1082 positive interactions and 13840 negative binary interactions in which negative datasets are formed by combination of one periplasmic and one cytoplasmic protein (non-colocalized). The *S. cerevisiae* physical interaction dataset obtained from the Ben-Hur and Noble (2005), consists of 10517 protein pairs for positive as well as negative examples in which negative dataset was created randomly. *H. pylori* physical interaction dataset was obtained from Martin *et al.* (2005) containing equal number of 1458 interactions and non-interactions (random pairs which do not show interaction).

Feature Extraction

Different features such as amino acid, dipeptide, pseudo amino acid, split amino acid composition and also biochemical descriptors have been extracted from amino acid sequences. These features were calculated separately for each protein sequence in a protein pair and then concatenated the features of each sequence to represent binary vector.

Sequence Composition

Amino acid composition is the frequency of each type of amino acid in a protein sequence. We have generated a dipeptide matrix of size 20×20 from 20 types of amino acids. Dipeptide (n + 1), where n is the position of each residue along the length of protein sequence, composition was calculated as ratio of occurrence of a particular dipeptide (out of 400) by total number of dipeptides in the sequence. Split amino acid composition is simply the amino acid composition of four equal parts of the protein sequence, making feature length of 80 (4*20).

Biochemical Descriptors

The 20 amino acid residues are classified into six biochemical similarity classes namely B, J, O, U, X and Z [37]. These classes contain [IVLM], [FYW], [HKR], [DE], [QNTP] and [ACGS] amino acids respectively. The amino acid sequences are decoded based on this classification. Further, this decoded stretch (for e.g. BJOUBJZX….) was used for computation of different compositions. The biochemical monopeptide compositions were similar to amino acid composition but length of the binary vector was 6. For biochemical dipeptide compositions, total 36 (6*6) features were extracted for a protein and 72 features for a pair. Biochemical tripeptide was calculated by total number of each type of tripeptides divided by total number tripeptides in the protein sequence. Totally, 216 (6×6×6) features were extracted from a protein sequence that is 432 for a protein pair.

Pseudo Amino Acid Composition

Pseudo-amino acid (PA) compositions were calculated by using perl script based on the concept of Chou's pseudo-amino acid composition [25, 44]. We considered the parallel correlation type and the hydrophobic parameters for calculating the pseudo amino acid composition. Further, one feature is added to the standard set of features that reflects the sequential order of the protein sequence. In this case, we have dimensionality of 42 for a feature vector representing interaction pair.

Composition of Position-Specific Scoring Matrix (PSSM)

The PSSM profile for each protein was generated using PSI-BLAST [47] by searching the protein against NR database obtained from NCBI. The PSI-BLAST was used with cut-off value 0.001 with three iterations. The PSSM scores were normalized in order to get values between 0 and 1, and then position specific composition of each amino acid was calculated. This way we got composition of amino acids with

evolutionary information in form of 400 values [30] for a protein and 800 for a pair of protein.

Negative Dataset Selection

We selected protein pairs uniformly at random from the set of all protein pairs that are not known to interact [20, 38, and 48]. Moreover, we also excluded those negative pairs which are having at least one partner that is present in positive dataset. We used such a set of non-interacting pairs for further consideration and named it negative pair database. Now we formulated two strategies to make two different negative datasets each of 1082 pairs.

Random Non-Interacting

It is the set of 1082 pairs selected randomly from negative pair database.

Non-Redundant Non-Interacting

It is the set of 1082 pairs from negative pair database such that none of the protein is repeated.

Selection of sequence-based features contributing in PPI prediction

Average compositional bias (ACB) for each feature of a descriptor was calculated (Eq. 1) from pairwise interaction data. Considering amino acid composition as a descriptor we computed 20 ACB values (since maximum 20 natural amino acid residues constitute protein molecule) for a protein and 40 ACB values (20*2) for interacting or non-interacting pairs.

Average Compositional Biasness (ACB)

It is the ratio of difference of summation of composition values for a particular feature of the descriptor from positive examples and negative examples and sum of summation of composition values for a particular feature of the descriptor from positive examples and negative examples. The formula for computing ACB was as follows-

$$\frac{\sum_{i=1}^{L} Cpi - \sum_{i=1}^{M} Cni}{\sum_{i=1}^{L} Cpi + \sum_{i=1}^{M} Cni} \quad \text{Eq.(1)}$$

Where *Cp* and *Cn* are composition values for a particular feature, say alanine, from positive examples and negative examples, and L and M are total number of positive and negative examples respectively. Likewise we have 40 ACB values for interaction dataset for a particular descriptor (here it is amino acid composition).

Then we sorted out these ACB values in descending order in a list. The positive ACB value for a feature stated that particular feature is dominant in interacting pair and vice-versa. Now we trained SVM taking the equal number of features having positive ACB values from the top of the list and equal number of features having negative ACB values from the bottom of the list. We tried to optimize this SVM to approach the performance of the original model that is trained on all 40 compositional values. Similar methodology has been applied on dipeptide and biochemical tripeptide descriptors.

Wrapper Based Attribute Selection

Feature (or attribute) Subset Selection (FSS) is a process of identifying input features which are relevant to the supervised or unsupervised learning (or data mining) problem. We used wrapper evaluator with genetic algorithm based search method to select a feature set for supervised classification oriented problem. Wrappers are a popular type of evaluator: they calculate a score for a subset by inducing a classifier using only those attributes. Wrappers tend to lead to superior accuracy, but need high computational effort, compared to so-called filter methods. Filters use statistical characteristics of the data for evaluation that are independent of the classifier. We exploited wrapper based FSS algorithm implemented in WEKA [49].

AUTHORS' CONTRIBUTIONS

MR retrieved datasets, developed various SVM modules and evaluated all modules. SR processed datasets, run some SVM training and helped in compiling results. GPSR conceived the idea, coordinated it and refined the manuscript drafted by MR and SR. All the authors have read and approved the final manuscript.

ACKNOWLEDGEMENT

We gratefully acknowledge Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology, Government of India, for financial assistance. We sincerely thank Mr. Hifzur Rahman Ansari and Mr Harinder for their help in manuscript formatting. We are grateful to Dr Anand K. Bachawat for his help in improving quality of English of this manuscript. This report has IMTECH communication number 024/2008.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Ge, H.; Walhout, A.J.; Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, **2003**, *19*(10), 551-560.

[2] Ryan, D. P.; Matthews, J.M. Protein-protein interactions in human disease. *Curr. Opin. Struct. Biol.*, **2005**, *15*(4), 441-446.

[3] Peltier, J.M.; Askovic, S.; Becklin, R.R.; Chepanoske, C.L.; Ho, Y.-S.J.; Kery, V.; Lai, S.; Mujtaba, T.; Pyne, M.; Robbins, P.B.; Rechenberg, M.V.; Richardson, B.; Savage, J.; Sheffield, P.; Thompson, S.; Weir, L.; Widjaja, K.; Xu, N.; Zhen, Y.; Boniface, J.J. An integrated strategy for the discovery of drug targets by the analysis of protein-protein interactions. *Int. J. Mass Spectrom.*, **2004**, *238*(2), 119-130.

[4] Shangary, S.; Qin, D.; McEachern, D.; Liu, M.; Miller, R. S.; Qiu, S.; Nikolovska-Coleska, Z.; Ding, K.; Wang, G.; Chen, J.; Bernard, D.; Zhang, J.; Lu, Y.; Gu, Q.; Shah, R. B.; Pienta, K. J.; Ling, X.; Kang, S.; Guo, M.; Sun, Y.; Yang, D.; Wang, S., Temporal activation of p53 by a specific MDM2 inhibitor is selectively toxic to tumors and leads to complete tumor growth inhibition. *Proc. Natl. Acad. Sci. USA*, **2008**, *105*(10), 3933-3938.

[5] Fields, S.; Song, O. A novel genetic system to detect protein-protein interactions. *Nature*, **1989**, *340*(6230), 245-246.

[6] Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G.D.; Moore, L.; Adams, S.L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A.R.; Sassi, H.; Nielsen, P.A.; Rasmussen, K.J.; Andersen, J.R.; Johansen, L.E.; Hansen, L.H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R.C.; Gleeson, F.; Pawson, T.; Moran, M.F.; Durocher, D.; Mann, M.; Hogue, C.W.; Figeys, D.; Tyers, M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **2002**, *415*(6868), 180-183.

[7] Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R.A.; Gerstein, M.; Snyder, M. Global analysis of protein activities using proteome chips. *Science* **2001,** 293, (5537), 2101-2105.

[8] Lakey, J.H.; Raggett, E.M. Measuring protein-protein interactions. *Curr. Opin. Struct. Biol.*, **1998**, *8*(1), 119-123

[9] Tong, A.H.; Drees, B.; Nardelli, G.; Bader, G.D.; Brannetti, B.; Castagnoli, L.; Evangelista, M.; Ferracuti, S.; Nelson, B.; Paoluzi, S.; Quondam, M.; Zucconi, A.; Hogue, C.W.; Fields, S.; Boone, C.; Cesareni, G. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **2002**, *295*(5553), 321-324.

[10] Valencia, A.; Pazos, F. Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **2002**, *12*, (3), 368-373.

[11] Salwinski, L.; Eisenberg, D. Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **2003**, *13*, (3), 377-382.

[12] Shoemaker, B.A.; Panchenko, A.R. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **2007**, *3*(4), e43.

[13] Goh, C.S.; Bogan, A.A.; Joachimiak, M.; Walther, D.; Cohen, F.E. Co-evolution of proteins with their interaction partners. *J Mol Biol* **2000**, *299*(2), 283-293.

[14] Pazos, F.; Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **2001**, *14*(9), 609-614.

[15] Dandekar, T.; Snel, B.; Huynen, M.; Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **1998**, *23*(9), 324-328.

[16] 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.

[17] Enright, A.J.; Iliopoulos, I.; Kyrpides, N.C.; Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **1999**, *402*(6757), 86-90.

[18] Pazos, F.; Valencia, A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **2002**, *47*(2), 219-227.

[19] Martin, S.; Roe, D.; Faulon, J.L. Predicting protein-protein interactions using signature products. *Bioinformatics*, **2005**, *21*(2), 218-226.

[20] Ben-Hur, A.; Noble, W.S. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **2005**, *21*(1), i38-46.

[21] Yellaboina, S.; Goyal, K.; Mande, S.C. Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res.*, **2007**, *17* (4), 527-535.

[22] Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N.J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J.F.; Gerstein, M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **2003**, *302*(5644), 449-453.

[23] Dohkan, S.; Koike, A.; Takagi, T. Improving the performance of an SVM-based method for predicting protein-protein interactions. *In Silico Biol.*, **2006**, *6*(6), 515-529.

[24] Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*(11), 4337-4341.

[25] Chou, K.C.; Cai, Y.D. Predicting protein-protein interactions from sequences in a hybridization space. *J. Proteome. Res.*, **2006**, *5*(2), 316-322.

[26] Bock, J.R.; Gough, D.A. Predicting protein--protein interactions from primary structure. *Bioinformatics*, **2001**, *17*(5), 455-460.

[27] Bock, J.R.; Gough, D.A. Whole-proteome interaction mining. *Bioinformatics*, **2003**, *19*(1), 125-134.

[28] Chen, X.W.; Liu, M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **2005**, *21*(24), 4394-400.

[29] Ben-Hur, A.; Noble, W. S., Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **2006**, *7(*1), S2.

[30] Rashid, M.; Saha, S.; Raghava, G.P. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics*, **2007**, *8*, 337.

[31] Bhasin, M.; Raghava, G.P. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **2004**, *32*, (Web Server issue), W414-9.

[32] Bhasin, M.; Garg, A.; Raghava, G.P. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, **2005**, *21*, (10), 2522-2524.

[33] Nishikawa, K.; Ooi, T. Correlation of the amino acid composition of a protein to its structural and biological characters. *J. Biochem.*, **1982**, *91*(5), 1821-1824.

[34] Andrade, M.A.; O'Donoghue, S.I.; Rost, B. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.*, **1998**, *276*(2), 517-525.

[35] Cedano, J.; Aloy, P.; Perez-Pons, J.A.; Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **1997**, *266*(3), 594-600.

[36] Garg, A.; Bhasin, M.; Raghava, G.P. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 2005, 280, (15), 14427-14432.

[37] Taylor, W.R.; Jones, D.T. Deriving an amino acid distance matrix. *J. Theor. Biol.* **1993**, *164*(1), 65-83.

[38] Gomez, S.M.; Noble, W.S.; Rzhetsky, A. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, **2003**, 19(15), 1875-1881.

[39] Yu, H.; Braun, P.; Yildirim, M.A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, J. F.; Dricot, A.; Vazquez, A.; Murray, R. R.; Simon, C.; Tardivo, L.; Tam, S.; Svrzikapa, N.; Fan, C.; de Smet, A. S.; Motyl, A.; Hudson, M. E.; Park, J.; Xin, X.; Cusick, M. E.; Moore, T.; Boone, C.; Snyder, M.; Roth, F. P.; Barabasi, A. L.; Tavernier, J.; Hill, D. E.; Vidal, M., High-quality binary protein interaction map of the yeast interactome network. *Science*, **2008**, *322*(5898), 104-110.

[40] Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N.; Luo, X.; Golshani, A., PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **2006**, *7*, 365

[41] Verma, R.; Tiwari, A.; Kaur, S.; Varshney, G.C.; Raghava, G.P. Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics*, **2008**, *9*, 201.

[42] Chinnasamy, A.; Mittal, A.; Sung, W.K. Probabilistic prediction of protein-protein interactions from the protein sequences. *Comput. Biol. Med.*, **2006**, *36*(10), 1143-1154.

[43] Aloy, P.; Russell, R.B. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **2003**, 19, (1), 161-162.

[44] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43*(3), 246-255.

[45] Chou, K.C.; Shen, H.B. Large-Scale Predictions of Gram-Negative Bacterial Protein Subcellular Locations. *J. Proteome Res.*, **2006**, *5*(12), 3420-3428.

[46] Chou, K.-C.; Shen, H.-B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protocols*, **2008**, *3*(2), 153-162.

[47] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25* (17), 3389-3402.

[48] Zhang, L.; Wong, S.; King, O.; Roth, F. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **2004**, *5*(1), 38.

[49] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, Volume 11, Issue 1.

Received: ???????????????? **Revised:** ???????????????? **Accepted:** ????????????????