DOI 10.1002/pmic.200401118 *Proteomics* 2005, *5,* 876–884

REGULAR ARTICLE

Effect of training datasets on support vector machine prediction of protein-protein interactions

Siaw Ling Lo[1, 2, 4], Cong Zhong Cai[2], Yu Zong Chen[3] and Maxey C. M. Chung [4]*

[1] Department of Biological Sciences, National University of Singapore
[2] Bioprocessing Technology Institute
[3] Department of Computational Science, National University of Singapore
[4] Department of Biochemistry, National University of Singapore
Singapore

Received: July 3, 2004
Revised: October 7, 2004
Accepted: November 11, 2004

Knowledge of protein-protein interaction is useful for elucidating protein function *via* the concept of 'guilt-by-association'. A statistical learning method, Support Vector Machine (SVM), has recently been explored for the prediction of protein-protein interactions using artificial shuffled sequences as hypothetical noninteracting proteins and it has shown promising results (Bock, J. R., Gough, D. A., *Bioinformatics* 2001, *17*, 455–460). It remains unclear however, how the prediction accuracy is affected if real protein sequences are used to represent noninteracting proteins. In this work, this effect is assessed by comparison of the results derived from the use of real protein sequences with that derived from the use of shuffled sequences. The real protein sequences of hypothetical noninteracting proteins are generated from an exclusion analysis in combination with subcellular localization information of interacting proteins found in the Database of Interacting Proteins. Prediction accuracy using real protein sequences is 76.9% compared to 94.1% using artificial shuffled sequences. The discrepancy likely arises from the expected higher level of difficulty for separating two sets of real protein sequences than that for separating a set of real protein sequences from a set of artificial sequences. The use of real protein sequences for training a SVM classification system is expected to give better prediction results in practical cases. This is tested by using both SVM systems for predicting putative protein partners of a set of thioredoxin related proteins. The prediction results are consistent with observations, suggesting that real sequence is more practically useful in development of SVM classification system for facilitating protein-protein interaction prediction.

Keywords:
Database of interacting proteins / Protein function prediction / Protein-protein interaction prediction / Shuffled sequence / Support vector machine / SVM*light*

1 Introduction

Protein-protein interactions play important roles in various biological events [1] and are the basis for assemblies of molecular machines such as RNA polymerase II. The 'guilt-by-association' concept has been used for elucidating functional roles from pairs of interacting proteins [2]. Identification of its partner of known function may provide a useful clue to the possible role of a protein of unknown function. Knowledge of protein-protein interactions is also useful for probing biological pathways and regulation of signaling, metabolic, gene expression and replication processes. Various experimental approaches have been used for the study of protein-protein interactions. These include yeast two-hybrid systems [3], protein complex purification techniques using mass spectrome-

Correspondence: Dr. Maxey C. M. Chung, Department of Biochemistry, National University of Singapore, MD 7, Level 5, 10 Kent Ridge Crescent, Singapore, 117597
E-mail: bchcm@nus.edu.sg
Fax: +65-6779-1453

Abbreviations: DIP, database of interacting proteins; **RI,** reliability index; **SVM,** support vector machine

* Additional corresponding author: Dr. Yu Zong Chen
E-mail: yzchen@cz3.nus.edu.sg

try [4, 5], protein chip [6], correlated messenger RNA expression profiles [7] and genetic interaction data [8]. However, it is not yet feasible to construct a complete protein interaction map by exhaustive experimental studies. Hence, there is a growing interest in the exploration of computational methods for the prediction of protein-protein interactions.

So far, three different computational approaches have been explored for the prediction of protein-protein interactions. The first is based on genomics that uses phylogenetic profiles of the presence and absence of genes in related species [9], conservation of gene order in different species [10] and gene fusion events [11] for functional prediction. The second is based on the analysis of a variety of structural and physicochemical features including the site of interaction from surface patches [12], sequence and residue neighbor profile [13] and molecular docking [14]. The third is concerned with the prediction of putative protein partner(s) from protein primary structure and associated physicochemical properties [15] or the correlated sequence-signatures that recur in concert in various pairs of interacting proteins [16].

Because of the limited availability of protein 3-D structures, methods that derive information directly from protein primary structure are of particular interest. A statistical learning method, support vector machines (SVM), has recently been explored for the prediction of protein-protein interactions [15] as well as protein secondary structure prediction [17], protein fold recognition [18], analysis of protein solvent accessibility [19] and other biomedical problems including microarray gene expression data analysis [20] and cancer diagnosis [21]. These studies have consistently shown that SVM is usually superior to traditional supervised learning methods.

Like other statistical learning methods, the accuracy of SVM classification depends on the relevance of the training dataset to a particular biological problem. Thus it is important to use a reliable training dataset to achieve a better classification. Since experimental conditions and, in some cases, types of proteins are known to affect the accuracy of some of the experimental methods [22], caution needs to be exercised in the interpretation of experimental data. A large-scale comparative assessment of protein-protein interaction data suggested that highest accuracy is achieved for those interactions supported or predicted by more than one method, including *in silico* approach [22]. Hence, to ensure their quality, the dataset of interacting proteins (positive dataset) used in this work is from a subset of the data in the Database of Interacting Proteins (DIP) [23] whose reliability has been tested [24]. Since noninteracting proteins are not readily available, artificial shuffled sequences resembling realistic proteins have been used to construct the dataset of hypothetical noninteracting proteins (negative dataset) for the prediction of protein-protein interactions. Bock and Gough [15] have shown that it gives an average accuracy of 80.9%. However, shuffling sequences artificially may result in sequences with no specific sequence patterns like motifs or domains, while real protein sequences are known to contain these conserved sequence patterns that play an important functional role. It is unclear whether a classification system derived from artificial sequences is sufficiently effective in practical prediction of protein-protein interactions since artificial shuffled sequences, having the possibility of not containing any motifs or domains, are likely to be 'nonfunctional' proteins. It is thus desirable to only use real protein sequences for developing an SVM classification system which might be more relevant to the prediction of protein-protein interactions.

It is of interest to evaluate how the prediction accuracy can be affected by the use of real protein sequences. For such a purpose, real protein sequences are used to construct a negative dataset. These real protein sequences are from an exclusion analysis in combination with subcellular localization information of interacting proteins in DIP. The prediction accuracy of an SVM system trained from this dataset is compared with those artificial shuffled sequences generated from the same principle as described in the literature [15]. The prediction performance of both systems is further evaluated by using them for the identification of putative interacting partners of a set of thioredoxin related proteins.

2 Materials and methods

2.1 Data collection and dataset construction

The positive dataset is downloaded from *Saccharomyces cerevisiae* core subset of DIP database [24]. This dataset is validated by two methods described by Deane and colleagues [24]. The first is to use the expression profile reliability index to estimate the biologically relevant fraction of protein interactions by comparing the RNA expression profiles of the proteins with expression profiles of known interacting and noninteracting pairs of proteins. The second is to use the paralogous verification method to test the reliability of a putative interaction pair by examining whether there is a known paralog that also interacts with its partner protein.

Since a noninteracting protein dataset is not readily available, a hypothetical noninteracting protein dataset is generated based on subcellular localization information and consists of protein pairs that do not colocalize together. The subcellular localization source is retrieved from Munich Information Center for Protein Sequences (MIPS) [25] and only the four main types of localization are considered in this study – cytoplasm, nucleus, mitochondria and endoplasmic reticulum. The yeast proteins used in the positive dataset are assigned with the four types of localization information and those with multiple localizations are removed to minimize the introduction of possible noise in the training process. Four sets of proteins with respect to the four types of localization are generated and proteins from each set are subsequently paired with proteins from a different localization. Due to the enormous amount of possible pairing, 5000 pro-

tein pairs are randomly selected and used in this work. After removing duplication and performing exclusion analysis of the whole DIP yeast interacting proteins, a total of 4660 protein pairs are used as the hypothetical noninteracting dataset.

As a comparison, a second type of negative dataset composed of artificial protein sequences of the hypothetical noninteracting dataset are derived by using the Shufflet program [26] with *k*-let (k = [1, 2]) counts. The *k*-let (the exact words equal to or shorter than a given length *k*) are kept conserved in generating random shuffled sequences. In addition to preserving the amino acid composition which correlates with protein-protein interfaces [27], such a shuffling [26] also maintains the frequencies of dipeptides, tripeptides *etc*. The algorithm ensures that every expected occurrence of each possible *k*-let has the same probability, which is expected to generate datasets with conserved properties that are closer to real protein sequences than pure randomly generated sequences. In general, a *k*-let works well for sequences up to 20^k amino acids in length. Hence in order to maintain the random uniform permutation, only 1-let and 2-let shuffled protein sequences are considered in this study.

Each dataset is further divided in a random fashion into a training set and a testing set while maintaining representatives of distinct protein pairs in each set whenever possible. For example, if the positive dataset has four interacting protein pairs of protein D, then each of the two pairs will be randomly distributed to positive training and testing set respectively. The training dataset is evaluated to remove homologous sequences using BLASTCLUST [28] with identity threshold of 30% and length coverage threshold of 90% to ensure the classifier is not biased to homologous sequences. This gives a positive training set of 2080 interacting proteins, a negative training set of 2331 noninteracting proteins, a positive testing set of 2208 interacting proteins and a negative testing set of 2331 noninteracting proteins.

2.2 Feature extraction and representation

The feature vector of each interacting protein pair is constructed by using encoded representation of tabulated residue properties of the two protein sequences including amino acid composition, hydrophobicity, van der Waals volume, polarity, polarizability, charge and surface tension for each residue in sequences. Each protein sequence is converted into a feature vector using amino acid composition percentage and the feature extraction method based on three descriptors [18]. The first is the composition, which is a percent composition of three constituents/groupings (*e.g.*, polar, neutral and hydrophobic residues for the feature of hydrophobicity). The second is transition, which describes the transition frequencies (polar to neutral, neutral to hydrophobic, *etc*.). The third is distribution, which represents the distribution pattern of a particular property (the position of the first amino acid of a given property and the section in which 25, 50, 75 and 100% of the amino acids with that property are contained).

2.3 Support vector machines

SVM is a relatively new type of supervised learning algorithm for two- or multi-class classification, which was originally developed by Vapnik and coworkers [29, 30]. SVM separates a given known set of {+1, −1} labeled training data *via* a hyperplane that is maximally distant from the positive and negative samples. This optimally separating hyperplane in the feature space corresponds to a nonlinear decision boundary in the input space. Each of the feature vector generated from the protein pairs in the positive dataset and negative dataset are assigned with a label of {+1} and {−1} respectively to indicate if the pair is interacting with each other or not. Details of SVM can be found in the literature [30].

In order to compare our results with that obtained in an earlier study [15], the same software SVM^{light} (http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT) [31] as used in that study is employed in this work. A Gaussian kernel function [exp(−γ |a−b|^2)] with an optimized γ parameter is used. Gaussian kernel function has been commonly used and shown to produce higher precision prediction than other kernel functions [32, 33].

As in other statistical learning studies, SVM prediction accuracy can be described by means of the classification accuracy Q, precision and recall.

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{precision} = TP/(TP + FP) \quad (2)$$

$$\text{recall} = TP/(TP + FN) \quad (3)$$

where TP, TN, FP and FN represents true positive, true negative, false positive, and false negative respectively. The SVM classification is further evaluated using five-fold cross-validation and the standard deviation is calculated as an indication of the consistency in the prediction accuracy obtained. In order to better understand the classification by the three models, the distance *d* between the position of the vector of the classified protein and the optimal separating hyperplane in the hyperspace is calculated using the following formula:

$$\boldsymbol{d} = \sum_{i=1}^{l} \alpha_i \boldsymbol{K}(\mathbf{x}, \mathbf{x}_i) + \boldsymbol{b} \quad (4)$$

where the coefficients α_i and *b* are the alpha and threshold values determined by SVM^{light} svm_learn program and $\boldsymbol{K}(\mathbf{x},\mathbf{x}_i)$ is the kernel function.

Scoring of SVM classification can then be estimated by a reliability index (RI) and the *RI* value is defined as *d*/0.2 where *d* is the distance defined above. The relationship between *RI* value and accuracy percentage or statistical *P*-value is shown in Fig. 1a while the Receiver Operator Characteristic (ROC) plot of each *RI* value can be found in Fig. 1b. In general, the absolute value of *d* is in the interval [0,2] and *RI* is a value range from 0 to 10 with $RI >= 7$ corresponding to a rather reliable prediction.

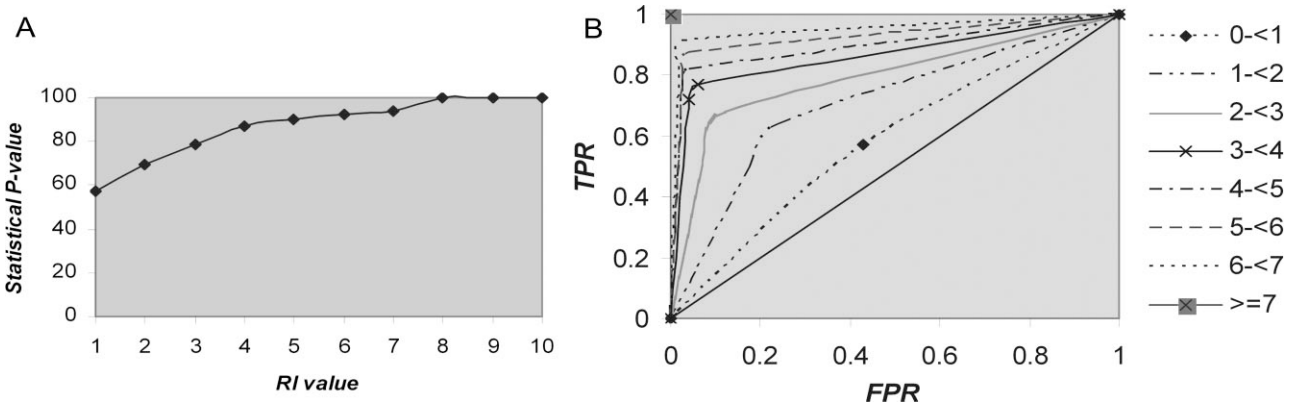

Figure 1. a, Statistical relationship between the RI value and *P*-value (probability of correct classification) derived from analysis of 2208 positive and 2331 negative protein-protein interaction dataset; b, ROC plot of RI value. The legend indicate the range of RI value and its corresponding ROC curve. *TPR* is True Positive Rate (Sensitivity) and *FPR* is False Positive Rate (1-Specificity).

3 Results and discussion

3.1 Prediction accuracy

Table 1 gives the accuracy of SVM prediction of interacting proteins using both artificial shuffled protein sequences and real protein sequences as the negative datasets. It is found that the prediction accuracy using 1-let shuffled protein sequences as a negative dataset is 94.1% while 2-let shuffled protein sequences yield 89.3%, which is comparable to the accuracy of 80.9% from an earlier work [15]. The slight improvement is probably due to the different feature representation and dataset construction methods. In contrast, the prediction accuracy using real protein sequences as negative datasets is 76.9%, which is substantially lower than that derived from the use of shuffled protein sequences as negative datasets.

3.2 Effect of different negative datasets

This result seems to indicate a correlation between the degree of random shuffling of protein sequences in the negative datasets and the computed classification accuracy. The increasing randomness of the negative dataset tends to give better prediction accuracy, which is expected as increasingly artificial random shuffled sequences are likely to be more easily distinguished from real protein sequences. As shown in Fig. 2, even though shuffled sequences trained classifiers achieve a higher accuracy on shuffled sequences testing datasets, they are not able to perform as well when applied on real sequences testing dataset. This is understandable as the level of difficulty for classifying two datasets of real protein sequences is expected to be higher than that of one set of real protein sequences and one set of shuffled sequences, which partly contributes to the lower classification accuracy derived from the use of real protein sequences. In order to determine the effect of sequence randomness on the performance of the SVM classification, the average distance of support vectors to the respective optimal separating hyperplane for each of the three models is computed. The average distance generated from the negative dataset of real sequence (d*r*) is 0.54, while that of the shuffled 1-let sequences (d*s1*) and shuffled 2-let sequences (d*s2*) is 0.73 and 0.70 respectively. The classification system of the 1-let shuffled protein sequences gives the largest average distance while that of the real protein sequence gives the smallest average distance. Figure 3 explains the effect of using different

Table 1. Prediction accuracy of SVM classification of interacting proteins using shuffled sequences and real protein sequences as negative dataset (dataset for noninteracting proteins).

Negative dataset	TP	FN	TN	FP	Precision (%)	Recall (%)	Prediction accuracy (%)
Shuffled sequences (1-let)	2039	169	2233	98	95.4	92.3	94.1 (1.3)
(2-let)	1935	273	2117	212	90.1	87.6	89.3 (0.7)
Real protein sequences	1527	679	1963	368	80.6	69.2	76.9 (1.7)

TP, TN, FP and FN represents true positive, true negative, false positive, and false negative respectively. Details of the negative datasets construction are given in Section 2.1. A total of 2208 interacting proteins are used as a positive testing dataset, while 2331 noninteracting proteins are in a negative testing dataset. Combined results of five-fold cross-validation are shown. The numbers in parentheses under Prediction accuracy (last column) correspond to the standard deviations with five-fold cross-validation.

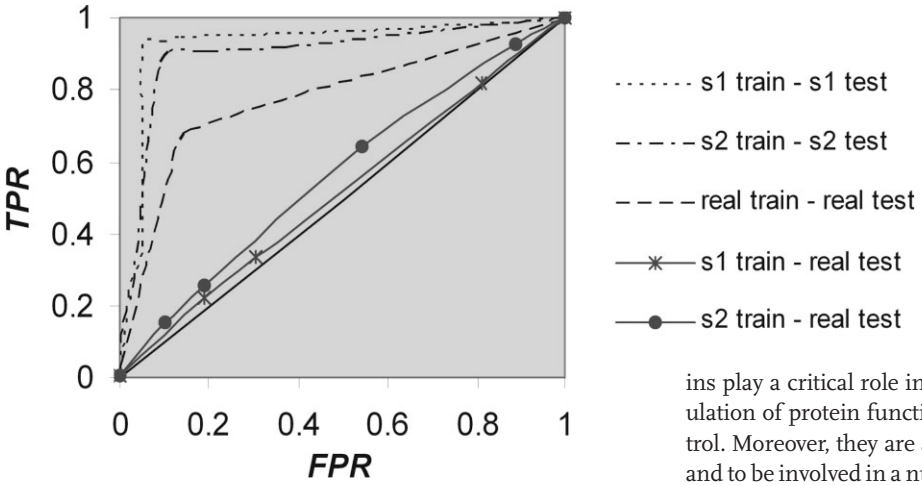

Figure 2. ROC plot of various SVM classifications. s1, s2 and real represents training or testing dataset containing shuffled 1-let, 2-let sequences and real protein sequences as negative dataset respectively, while train and test in the legend indicates SVM training and testing dataset. For example, s2 train – real test in the legend means the ROC curve of the classification using SVM model trained with shuffled 2-let protein sequences as negative dataset on real sequences testing dataset.

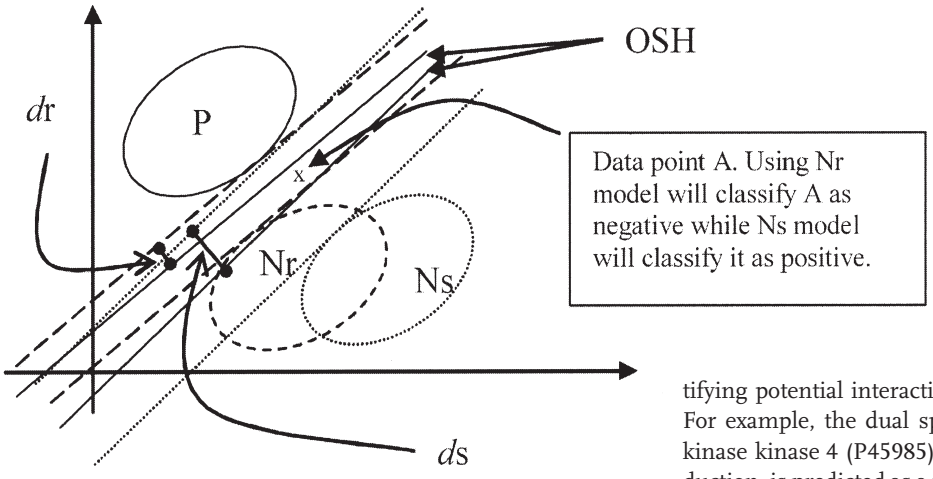

Figure 3. The effect of using different negative datasets in simplified 2-D diagram. The larger margin or the distance ($ds > dr$) between the position of the support vector and the optimal separating hyperplane (OSH) in the hyperspace implies that it is able to distinguish positive real sequence dataset (P) and the shuffled sequence negative dataset (Ns) better than real sequence negative dataset (Nr).

negative datasets in a simplified 2-D diagram. The larger margin in between the two classes of dataset implies that the model is capable of classifying a given test data better than those with a smaller margin. For example, assuming that the data point A is a positive test data, data point A will be classified correctly when 1-let shuffled sequence model is used, but this is not the case when it is classified using the real sequence model.

3.3 Thioredoxin related proteins

To further evaluate the performance of SVM classification systems trained by using different types of negative datasets, a set of thioredoxin related proteins are used as a preliminary test of the prediction capability of these systems in real case studies. Thioredoxins play a critical role in reduction and oxidation (redox) regulation of protein function and signaling *via* thiol redox control. Moreover, they are also known to facilitate DNA binding and to be involved in a number of functions in defense against oxidative stress, control of growth and apoptosis and if secreted, has chemokine activities [34]. Several human thioredoxin related proteins from the Swiss-Prot database [35] are used in this study. The details of the proteins are listed in Table 2.

A total of 7985 human proteins are extracted from the Swiss-Prot database as the candidates of potential interacting partners of each of these thioredoxin related proteins. Each of the 7985 candidate proteins is paired with each thioredoxin related protein to generate feature vectors which are submitted to the three SVM classification systems by the procedure outlined in Section 2.2.

The results in Table 3 suggest that the SVM classification system using artificial shuffled protein sequences (both 1-let and 2-let shuffling) as the negative training datasets may not be practically useful as their ability in identifying potential interacting protein partners seems limited. For example, the dual specificity mitogen-activated protein kinase kinase 4 (P45985), which is involved in signal transduction, is predicted as a putative partner of TXNL_HUMAN by the SVM system of the 1-let shuffled sequences. However, this prediction result maybe questionable as the same protein is also predicted as a partner of TXN1_HUMAN and TXN5_HUMAN which are not known to be involved in signal transduction. On the other hand, the SVM system of the 2-let shuffled sequences predicts a probable ATP-dependent RNA helicase p54 (P26196) as a potential partner of PDI_HUMAN, which appears to be consistent with the entry 12097 of the Biomolecular Interaction Network Database (BIND) [36] ■. This entry describes a protein-protein complex between *S. cerevisiae* PDI1 and DBP2 ATP-dependent RNA helicase. Besides that, Sepiapterin reductase (P35270) is also shown to be a possible partner of TXN1_HUMAN [37]. To further assess these prediction results, the two sets of putative protein partners are ranked by the reliability index.

Table 2. Details of human thioredoxin related proteins from Swiss-Prot

Entry name	Accession no.	Protein name	Annotated functions
PDI_HUMAN	P07237	Protein disulfide isomerase precursor	Procollagen-proline 4-dioxygenase activity; protein disulfide isomerase activity
TXN1_HUMAN	Q16881	Thioredoxin reductase	Thioredoxin-disulfide reductase activity
TXN5_HUMAN	Q8NBS9	Thioredoxin domain containing 5	Potential redox activity
TXNL_HUMAN	O43396	Thioredoxin-like protein 1	Plays a role in apoptosis; protein-disulfide reduction and signal transduction

As shown in Table 3, the reliability index for the top five protein partners of these two sets is low and thus they may not be confidently predicted as potential partners.

In contrast, the SVM system trained by real protein sequences as the negative training dataset appears to be more capable in identifying potential partners (Table 4). For instance, the proto-oncogene serine/threonine-protein kinase pim-1 (P11309) is predicted as one of the top five potential partners for each of the three thioredoxin related proteins TXN1_HUMAN, TXN5_HUMAN, TXNL_HUMAN and the top potential partner for TXN1_HUMAN. While there is no direct evidence showing thioredoxin related proteins interact with Pim-1 kinase, recent research findings have revealed both proteins are regulated *via* the NF-kB pathway [38–40]. Another protein, mitogen-activated protein kinase 1 (P28482), is also predicted as a potential interacting candidate for TXNL_HUMAN which is consistent with its functional roles in signal transduction and apoptosis [41]. In addition, the 26S proteasome non-ATPase regulatory protein (Q15008) is identified as a putative partner of PDI_HUMAN. It is noted that the same complex has been found in *S. cerevisiae* (entry 12123 in BIND database). Besides that, several proteins with redox functions such as pyruvate dehydrogenase (P08559); 24-dehydrocholesterol reductase (Q15392) and soluble epoxide hydrolase (P34913) are also identified. Pyruvate dehydogenase (P08559) is known to play a role together with thioredoxin in the redox regulation of mitochondria [42] while 24-dehydrocholesterol reductase (Q15392), which is involved in cholesterol biosynthesis, regulates mitochondria initiated apoptotic pathways that are sensitive to the redox environment [43]. Although there may not be a direct interaction between soluble epoxide hydrolase (P34913) and TXN1_HUMAN, a recent publication [44] has shown that the expression of both proteins in the prostate apoptosis pathway may be correlated.

These results show that the predicted protein interaction pairs derived from the SVM system of real sequences are more consistent with experimental findings than those from artificial sequences, which suggests that SVM classification systems trained by using real protein sequences may be more practically useful in facilitating the prediction of putative potential interacting partners. Moreover, through the concept of 'guilt-by-association', such systems may also find potential application in facilitating protein function prediction of a novel protein by probing its interaction with other proteins of known function.

It is of interest to note that the four thioredoxin related proteins used in this study have less than 30% sequence identity with each other. The ability of the SVM system trained by the real sequences to both predict protein with redox function for all of the four proteins and identify putative protein partners having specific functions for individual protein, can be partially attributed to the use of feature vectors which are based on physicochemical properties of amino acid sequences rather than sequence similarity. From Table 4, one can see that the false positive rate is not small (indicated by [a)]), which is likely due in part to the limited diversity of the negative datasets used for training the SVM systems.

3.4 *Drosophila melanogaster* interaction database

While the thioredoxin examples have shown the potential of SVM classification system trained using real protein sequences as the negative training dataset, it may be more realistic to apply the three classification systems on a larger and more comprehensive dataset. The *Drosophila melanogaster* interaction dataset from DIP which consists of 20 988 interactions from 7052 proteins is selected as it is the biggest dataset in DIP at the time of writing. Out of the 20 988 interactions, 99.7% are extracted from high-throughput yeast two-hybrid approach [45]. The real sequences classifier predicts 64% as possible interacting protein pairs which is much lower than the shuffled sequences trained classifiers (91.2% and 85.5% for shuffled 1-let and shuffled 2-let sequences respectively). However the recent quality check on DIP yeast dataset (about 8000 interactions) indicates that only 50% of the dataset is reliable [24] while Sprinzak *et al*. [16] have shown that the reliability of high-throughput yeast two-hybrid assays is about 50% which may imply that the false positive rate in the *D. melanogaster* dataset can be close to 50%. This result suggests shuffled sequences trained classifiers are not very capable in differentiating the true positive or real inter-

Table 3. Top five prediction results (in descending order) from SVM classification of putative interacting protein partners of thioredoxin proteins when shuffled 1-let and 2-let sequences are used as negative dataset. Underlined proteins show evidence of being putative protein partners or having similar function.

Thioredoxin proteins (Swiss-Prot ID)	Putative protein partner	(Swiss-Prot ID)	[RI value]
Prediction results using shuffled 1-let sequences as negative dataset			
PDI_HUMAN	Leucine carboxyl methyltransferase	(Q9UIC8)	[4.87]
Protein disulfide isomerase	Desmin	(P17661)	[3.84]
precursor (P07237)	Oxysterols receptor LXR-alpha	(Q13133)	[3.78]
	ATP-dependent CLP protease ATP-binding subunit ClpX-like	(O76031)	[3.59]
	Replication protein A 30 kDa subunit	(Q13156)	[3.56]
TXN1_HUMAN	Leucine carboxyl methyltransferase	(Q9UIC8)	[2.87]
Thioredoxin reductase	Dual specificity mitogen-activated protein kinase kinase 4	(P45985)	[2.86]
(Q16881)	Keratin, type I cytoskeletal 17	(Q04695)	[2.70]
	Oxysterols receptor LXR-alpha	(Q13133)	[2.56]
	Replication protein A 30 kDa subunit	(Q13156)	[2.27]
TXN5_HUMAN	Leucine carboxyl methyltransferase	(Q9UIC8)	[3.65]
Thioredoxin domain	Oxysterols receptor LXR-alpha	(Q13133)	[3.41]
containing 5 (Q8NBS9)	Dual specificity mitogen-activated protein kinase kinase 4	(P45985)	[3.16]
	Replication protein A 30 kDa subunit	(Q13156)	[3.07]
	Desmin	(P17661)	[2.99]
TXNL_HUMAN	Leucine carboxyl methyltransferase	(Q9UIC8)	[5.16]
Thioredoxin-like protein 1	Oxysterols receptor LXR-alpha	(Q13133)	[4.13]
(O43396)	Desmin	(P17661)	[3.98]
	Dual specificity mitogen-activated protein kinase kinase 4	(P45985)	[3.93]
	Replication protein A 30 kDa subunit	(Q13156)	[3.84]
Prediction results using shuffled 2-let sequences as negative dataset			
PDI_HUMAN	Probable ATP-dependent RNA helicase p54	(P26196)	[2.60]
Protein disulfide isomerase	MutS protein homolog 4	(O15457)	[2.20]
precursor (P07237)	Short transient receptor potential channel 6 (TrpC6)	(Q9Y210)	[2.16]
	High-affinity cGMP-specific 3,5-cyclic phosphodiesterase 9A	(O76083)	[2.15]
	Protein-arginine deiminase type II (Peptidylarginine deiminase II)	(Q9Y2J8)	[1.96]
TXN1_HUMAN	Torsin A precursor (Dystonia 1 protein)	(O14656)	[1.15]
Thioredoxin reductase	Ethanolamine kinase (EKI)	(Q9HBU6)	[1.02]
(Q16881)	Pendrin (Sodium-independent chloride/iodide transporter)	(O43511)	[0.34]
	Sepiapterin reductase (SPR)	(P35270)	[0.32]
	MutS protein homolog 4	(O15457)	[0.84]
TXN5_HUMAN	MutS protein homolog 4	(O15457)	[1.91]
Thioredoxin domain	Torsin A precursor (Dystonia 1 protein)	(O14656)	[1.51]
containing 5 (Q8NBS9)	Ethanolamine kinase (EKI)	(Q9HBU6)	[1.31]
	Cholinesterase precursor	(P06276)	[1.11]
	Polycystin 2	(Q13563)	[1.10]
TXNL_HUMAN	MutS protein homolog 4	(O15457)	[2.47]
Thioredoxin-like protein 1	Probable ATP-dependent RNA helicase p54	(P26196)	[2.41]
(O43396)	Ethanolamine kinase (EKI)	(Q9HBU6)	[2.21]
	Polycystin 2	(Q13563)	[2.21]
	Torsin A precursor (Dystonia 1 protein)	(O14656)	[1.93]

acting protein pairs from a false positive, noninteracting protein pairs when applied in real testing dataset. Nevertheless, there is a need to include a reliability check, as suggested by Deane *et al.* [24], in addition to the *RI* value in the real protein sequences trained classifier in order to minimize the false positive rate.

3.5 Potential improvements

In addition to the negative dataset, the diversity of the positive dataset is also important for developing accurate SVM classification systems for protein-protein interactions. At present, the only publicly available and validated positive

Table 4. Top five prediction results (in descending order) from SVM classification of putative interacting protein partners of thioredoxin proteins when real sequences are used as negative dataset.

Thioredoxin proteins (Swiss-Prot ID)	Putative protein partner	(SwissProt ID)	[RI value]
Prediction results using real sequences as negative dataset			
PDI_HUMAN	Alpha-2,8-polysialyltransferase[a)]	(Q92187)	[7.77]
Protein disulfide	24-dehydrocholesterol reductase precursor	(Q15392)	[7.34]
isomerase precursor	Pyruvate dehydrogenase E1 component alpha subunit	(P08559)	[7.15]
(P07237)	Beta-parvin (Affixin) (CGI-56)[a)]	(Q9HBI1)	[7.08]
	26S proteasome non-ATPase regulatory subunit 6	(Q15008)	[7.02]
TXN1_HUMAN	Proto-oncogene serine/threonine-protein kinase pim-1	(P11309)	[9.72]
Thioredoxin reductase	Exostosin-like 3 (Putative tumor suppressor protein EXTL3)[a)]	(O43909)	[9.50]
(Q16881)	Soluble epoxide hydrolase	(P34913)	[9.24]
	Brain mitochondrial carrier protein-1[a)]	(O95258)	[9.20]
	Alpha-2,8-polysialyltransferase[a)]	(Q92187)	[9.04]
TXN5_HUMAN	24-dehydrocholesterol reductase precursor	(Q15392)	[7.17]
Thioredoxin domain	Pyruvate dehydrogenase E1 component alpha subunit	(P08559)	[7.16]
containing 5 (Q8NBS9)	cAMP-dependent 3,5-cyclic phosphodiesterase 4C[a)]	(Q08493)	[6.78]
	Proto-oncogene serine/threonine-protein kinase pim-1	(P11309)	[6.77]
	Angiotensinogen precursor[a)]	(P01019)	[6.75]
TXNL_HUMAN	Alpha-2,8-polysialyltransferase[a)]	(Q92187)	[7.93]
Thioredoxin-like protein 1	Proto-oncogene serine/threonine-protein kinase pim-1	(P11309)	[7.61]
(O43396)	24-dehydrocholesterol reductase precursor	(Q15392)	[7.42]
	Acidic fibroblast growth factor intracellular binding protein[a)]	(O43427)	[7.17]
	Mitogen-activated protein kinase 1 (MAP kinase 2)	(P28482)	[6.91]

Underlined proteins show evidence of being the putative protein partners.

a) Proteins are most probably false positive as they are currently not known to be interacting with thioredoxin related proteins.

interacting protein pairs are those extracted from yeast interaction data of DIP, and these are used in this work. This set of protein pairs may not be representative of all interacting proteins. Hence further improvement in the prediction capability is expected if a more comprehensive training data is used. Recently, cluster analysis of gene expression data has shown that genes with similar functions are likely to be coexpressed [20], hence prediction of protein-protein interactions by combining computer classification with additional information such as coexpression profile is helpful for developing a better tool for predicting putative interacting partners of proteins and for providing clues to the functional roles of a novel or unannotated protein.

4 Concluding remarks

Our study shows that the SVM classification system trained using real protein sequences as the negative training dataset performs better in real testing cases than that using artificial shuffled sequences. Even though the computed prediction accuracy of the former appears to be lower than the latter, the latter may not adequately reflect the true prediction capability because of the intrinsically higher level of difficulty for distinguishing real protein sequences than that for separating real protein sequences from artificial ones. This suggests the importance of using real protein sequences in developing SVM classification systems into a practical tool for protein analysis. Further improvement in the diversity and quality of datasets and classification algorithm may be useful in increasing the prediction accuracy of SVM. These combined with the analysis of additional information such as coexpression profile, may be of help in developing SVM and other classification methods into a useful tool for protein-protein interaction and protein function prediction.

*The authors would like to thank Eivind Coward of the Department of Informatics, University of Bergen for sharing the shufflet sequence-randomizing code. The work is supported by the LHK Fund from the Department of Biological Sciences, National University of Singapore, and a Core Competencies grant from the Agency for Science, Technology and Research (A*STAR), Singapore.*

5 References

[1] Pawson, T., Gish, G. D., Nash, P., *Trends Cell. Biol.* 2001, *11*, 504–511.

[2] Oliver, S., *Nature* 2002, *403*, 601–603.

[3] Fields, S., Song, O., *Nature* 1989, *340*, 245–246.

[4] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[5] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[6] Zhu, H., Bilgin, M., Bangham, R., Hall, D. *et al.*, *Science* 2001, *293*, 2101–2105.

[7] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J. *et al.*, *Cell* 2000, *102*, 109–126.

[8] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H. *et al.*, *Science* 2001, *294*, 2364–2368.

[9] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., Yeates, T. O., *Proc. Natl. Acad. Sci. USA* 1999, *96*, 4285–4288.

[10] Dandekar, T., Snel, B., Huynen, M., Bork, P., *Trends Biochem. Sci.* 1998, *23*, 324–328.

[11] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W. *et al.*, *Science* 1999, *285*, 751–753.

[12] Jones, S., Thornton, J. M., *J. Mol. Biol.* 1997, *272*, 121–132.

[13] Zhou, H., Shan, Y., *Proteins* 2001, *44*, 336–343.

[14] Smith, G. R., Sternberg, M. J. E., *Curr. Opin. Struct. Biol.* 2002, *12*, 28–35.

[15] Bock, J. R., Gough, D. A., *Bioinformatics* 2001, *17*, 455–460.

[16] Sprinzak, E., Margalit, H., *J. Mol. Biol.* 2001, *311*, 681–692.

[17] Hua, S. J., Sun, Z. R., *J. Mol. Biol.* 2001, *308*, 397–407.

[18] Ding, C. H. Q., Dubchak, I., *Bioinformatics* 2001, *17*, 349–358.

[19] Yuan, Z., Burrage, K., Mattick, J. S., *Proteins* 2002, *48*, 566–570.

[20] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 262–267.

[21] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2001, *98*, 15149–15154.

[22] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2002, *417*, 399–403.

[23] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P. *et al.*, *Nucleic Acids Res.* 2002, *30*, 303–305.

[24] Deane, C. M., Salwinski, L., Xenarios, I., Eisenberg, D., *Mol. Cell. Proteomics* 2002, *1.5*, 349–356.

[25] Mewes, H. W., Amid, C., Arnold, R., Frishman, D. *et al.*, *Nucleic Acids Res.* 2004, *32*, D41–D44.

[26] Coward, E., *Bioinformatics* 1999, *15*, 1058–1059.

[27] Ofran, Y., Rost, B., *J. Mol. Biol.* 2003, *325*, 377–387.

[28] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., *J. Mol. Biol.* 1990, *215*, 403–410.

[29] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer Verlag, New York 1995.

[30] Burges, C. J. C., *Data Min. Knowl. Disc.* 1998, *2*, 121–167.

[31] Joachims, T., in: Scholkopf, B., Burges, C., Smola, A. (Eds.), *Advances in Kernal Methods: Support Vector Learning*, MIT Press, Cambridge, MA 1999, pp. 42–56.

[32] Burbidge, R., Trotter, M., Buxton, B., Holden, S., *Comput. Chem.* 2001, *26*, 5–14.

[33] Czerminski, R., Yasri, A., Hartsough, D., *Quan. Struct-Act. Relatsh.* 2001, *20*, 227–240.

[34] Amer, E. S. J., Holmgren, A., *Eur. J. Biochem.* 2000, *267*, 6102–6109.

[35] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

[36] Bader, G. D., Betel, D., Hogue, C. W., *Nucleic Acids Res.* 2003, *3*, 248–250.

[37] Schallreuter, K. U., Buttner, G., Pittelkow, M. R., Wood, J. M. *et. al.*, *Biochem. Biophys. Res. Commun.* 1994, *204*, 43–48.

[38] Sakurai, A., Yuasa, K., Shoji, Y., Himeno, S. *et al.*, *J. Cell. Physiol.* 2004, *198*, 22–30.

[39] Zhang, J., Velsor, L. W., Patel, J. M., Postlethwait, E. M., Block, E. R., *Am. J. Physiol.* 1999, *277*, 787–793.

[40] Zhu, N., Ramirez, L. M., Lee, R. L., Magnuson, N. S. *et al.*, *J. Immunol.* 2002, *168*, 744–754.

[41] Shao, L.-E., Tanaka, T., Gribi, R., Yu, J., *Ann. NY Acad. Sci.* 2002, *962*, 140–150.

[42] Bunik, V. I., *Eur. J. Biochem.* 2003, *270*, 1036–1042.

[43] Fernandez-Checa, J. C., *Biochem. Biophys. Res. Commun.* 2003, *304*, 471–479.

[44] Pang, S. T., Dillner, K., Wu, X., Pousette, A. *et. al.*, *Endocrinology* 2002, *143*, 4897–4906.

[45] Giot, L., Bader, J. S., Brouwer, C. *et al.*, *Science* 2003, *302*, 1727–1736.

