Evaluating the denoising techniques in protein-protein interaction prediction

Yong-Cui Wang*
Key Laboratory of Adaptation
and Evolution of Plateau Biota,
Northwest Institute of Plateau Biology,
Chinese Academy of Science,
Xining, China, 810001
Email: ycwang@nwipb.cas.cn

Xian-Wen Ren*
State Key Laboratory for Molecular
Virology and Genetic Engineering,
Institute of Pathogen Biology,
Chinese Academy Medical Sciences
and Peking Union Medical College,
Beijing, China, 100730
Email: renxwise@gmail.com

Chun-Hua Zhang
Information School,
Renmin University of China,
Beijing, China, 100872
Email: zhangchunhua@ruc.edu.cn

Nai-Yang Deng†
College of Science,
China Agricultural University,
Beijing, China, 100083
Email: dengnaiyang@cau.edu.cn

Xiang-Sun Zhang†
Academy of Mathematics
and Systems Science,
Chinese Academy of Sciences,
Beijing, China, 100190
Email: zxs@amt.ac.cn

*These two authors contribute equally to the whole work. †Corresponding authors.

***Abstract*—The past decades witnessed extensive efforts to study the relationships among proteins. Particularly, sequence-based protein-protein interactions (PPIs) prediction is fundamentally important in speeding up the process of mapping interactomes of organisms. The composition vectors are usually constructed to encode proteins as real-value vectors, which is feeding to a machine learning framework. However, the composition vector value might be highly correlated to the distribution of amino acids, i.e., amino acids which are frequently observed in nature tend to have a large value of composition vector. Thus formulation to estimate the noise may be needed during representations. Here, we introduce two kinds of denoising composition vectors, which are efficient in construction of phylogenetic trees, to eliminate the noise. When validating these two denoising composition vectors on *Escherichia coli* (*E.coli*) and *Saccharomyces cerevisiae* (*S.cerevisiae*) randomly and artificial negative datasets, respectively, the predictive performance is not improved, and even worse than non-denoised prediction. These results suggest that, the denoising formulation efficient in phylogenetic trees construction can not improve the PPIs prediction, that is, what is noise is dependent on the applications.**

I. Introduction

Identification of the interactions among proteins is crucial to illustrate their functions, and further, can help us to understand the underlying mechanisms of many biological phenomena, such as cell cycles, apoptosis, signal transduction and pathogenesis of diseases. It has became one of the most challenging and important tasks in the post-proteomic researches. Various experimental techniques have been developed for large-scale protein-protein interactions (PPIs) analysis, including yeast two-hybrid systems [1], [2], mass spectrometry [3], [4], protein chip [5] and so on. One computational idea is applying the machine learning methods to learn understandable rules from the available PPIs and furthermore to predict novel interactions. Comparing with costly and time-consuming biochemical experiments, computational methods for PPIs prediction have played an important role [6].

One key issue in machine learning is to extract protein attributes that are highly relevant to prediction of PPIs. Among the various attributes of proteins, the primal sequences are most popular because they are the most basic and the easiest to obtain due to the rapid development of genomic sequencing technologies. In addition, the primary sequences of proteins actually specify their structures that provide the molecular basis for PPIs. Therefore protein primary sequences hold the promise to contain virtually sufficient information to construct the most universal predicting method [6].

The first challenge to construct a universal sequence-based PPIs predictor is how to encode the given proteins as the real-value vectors. Many types of composition vectors have been proposed [6], [7], [8], [9], [10], [11]. However, the composition vector value might be highly correlated to the distribution of amino acids, i.e., amino acids which are frequently observed in nature tend to have a large value of composition vector. Thus formulation to estimate the noise seems to be needed during representations. There are some works that have discussed this problem, for example, Chang, T.H. et.al. have proposed a probability-based mechanism for transforming protein sequences into feature vectors to eliminate the noise of composition vector [12]. With an efficient classification algorithm, the newly designed PPIs predictor is essential for handling highly unbalanced Human PPIs datasets. However,

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

when constructing one protein denoising composition vector, more than 10 thousand times permutation are implemented, the computational consumption is very large. Chan, R.H.F. et al have proposed a low computational cost denoising mechanism, which is based on the principle of maximum entropy, to encode the proteins as real-value vectors [13]. By using the angle-based distance measures on the denoising vectors, they have constructed well-grouped phylogenetic trees.

Following the previous works, in this paper, to improve the performance of PPIs prediction, we introduce two types of low costly denoising formulas, which are successfully used in phylogenetic tree construction. By applying Signal-to-Noise Ratio as the input vector, the given protein is converted to a real-value vector, which is feeding to the machine learning approaches. To test whether the predictive results can be improved by denoising, we introduce the support vector machine (SVMs) as the PPIs predictor. SVMs, which are motivated by statistical learning theory [14], [15], [16], have been proven successful on many different classification problems in bioinformatics [17]. Identification of PPIs can be addressed as the two-classification problem: determining whether a given pair of proteins is interacting or not. Inspired by that, in this paper, two-class SVM with the composition vectors and denoising composition vectors are used to predict $Escherichia\ coli$ ($E.coli$) and $Saccharomyces\ cerevisiae$ ($S.cerevisiae$) PPIs, respectively. As a result, both on $E.coli$ and $S.cerevisiae$ randomly and artificial negative datasets, the predictive performance are not improved by introducing the denoising formulations. These results suggest that, the denoising formulation efficient in phylogenetic trees construction can not improve the PPIs prediction, that is, what is noise is dependent on the applications.

II. Materials and Methods

Sequence-based attributes become popular in PPIs prediction not only because that the primal sequences are most basic and the easiest to obtain, but also owing to the assumption that knowledge of the amino acid sequence alone might be sufficient to estimate the evolutionary history, overall structure and function, and the interacting propensity between two proteins. Especially, Shen et.al. have proposed a much simple feature encoding method, called conjoint triad feature (CTF) to represent the protein sequences [6]. The authors have shown that SVM with the CTF outperforms other sequence-based methods in human PPIs prediction. In addition, the CTF can be implemented in an economic way and contains no pre-defined parameters. Inspired by these, here, we first introduce the CTF and then apply the denoising approaches to formulate the denoising CTF vectors. In results section, we test the performance of the denoising CTF vectors on $E.coli$ and $S.cerevisiae$ randomly and artificial negative datasets.

A. Input feature vectors

We give the description on the CTF now. Firstly, based on the dipoles and volumes of the side chains, the 20 amino acids are classified into seven classes: $\{A, G, V\}$, $\{I, L, F, P\}$, $\{Y, M, T, S\}$, $\{H, N, Q, W\}$, $\{R, K\}$, $\{D, E\}$, $\{C\}$. Secondly, a 343 $(7 \times 7 \times 7)$-dimension vector is used to represent a given protein, where each element of this vector is the frequency of the corresponding conjoint triad appearing in the protein sequence. More detailed description for the CTF can be found in [6].

1) Denoising vectors: The CTF considers the frequency of each conjoint triad type. However, the value of CTF's element might be highly correlated to the distribution of amino acids, i.e., triads that consist of amino acid groups frequently observed in nature (e.g., group 1 and 2) tend to have a large value of frequency. To deal with this problem, we introduce two types of denoising formulas, which are efficient in phylogenetic tree construction, to remove noises.

Specially, given a conjoint triad type $\alpha_1\alpha_2\alpha_3$, the following two formulas proposed by Hao et al. [18] and Yu et al. [19], are applied to estimate the noise of $\alpha_1\alpha_2\alpha_3$:

Hao's Formula:

$$q^{Hao}(\alpha_1\alpha_2\alpha_3) = \begin{cases} \dfrac{f(\alpha_1\alpha_2)f(\alpha_2\alpha_3)}{f(\alpha_2)}, & \text{if } f(\alpha_2) \neq 0 \\ 0, & otherwise, \end{cases} \tag{1}$$

where $f(u)$ represents the frequency of any string u appearing in the sequence. The formula (1) reveals the functional and evolutionary relatedness of word sequence, and was successfully applied for the phylogenetic analysis of prokaryotes, based on whole genome sequences [18].

Yu's Formula:

$$q^{Yu}(\alpha_1\alpha_2\alpha_3) = \frac{f(\alpha_1)f(\alpha_2\alpha_3) + f(\alpha_1\alpha_2)f(\alpha_3)}{2}, \tag{2}$$

where $f(u)$ represents the frequency of any string u appearing in the sequence. The formula (2) was commonly introduced in the area of complex and dynamic systems, and was successfully applied for the phylogenetic analysis of prokaryotes, chloroplasts and other phylogenetic problems, based on whole genome sequences [19].

Then the input vector feeding to the SVM can be formulated as the Signal-to-Noise Ratio:

$$s(\alpha_1\alpha_2\alpha_3) = \frac{f(\alpha_1\alpha_2\alpha_3) - q(\alpha_1\alpha_2\alpha_3)}{q(\alpha_1\alpha_2\alpha_3)}. \tag{3}$$

Comparing with the CTF, the element of the denoising vector become the Signal-to-Noise Ratio s. For any string of $\alpha_1\alpha_2\alpha_3$, when the value of $q(\alpha_1\alpha_2\alpha_3)$ is zero, it means the corresponding noise is zero. Therefore, in this situation, we let the Signal-to-Noise Ratio $s(\alpha_1\alpha_2\alpha_3)$ be the value of $f(\alpha_1\alpha_2\alpha_3)$.

If the value of noise q is zero for a string $\alpha_1\alpha_2\alpha_3$, we let $s(\alpha_1\alpha_2\alpha_3)$ be the value of $f(\alpha_1\alpha_2\alpha_3)$.

2) Protein pairs vectors: PPIs prediction treats each protein pair as the input, the vectors representing the protein pairs should be proposed. The concatenation operator are common used in protein pairs representation. However, the asymmetry problem will arise due to the fact that the prediction result will be different on protein pair A-B and B-A. To solve this

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

problem, we concatenate the arithmetical and the geometric average of protein vectors to represent the protein pairs, i.e.:

$$F_{AB} = \left(\frac{F_A + F_B}{2} \right) \oplus \sqrt{F_A * F_B}, \qquad (4)$$

where F_A, F_B represent the feature vector of protein A and B, the operator $*$ means the multiplication of the corresponding elements, and $\oplus$ represents the concatenation operator. The above representation method for protein pairs can not only maintain symmetry (A-B identical to B-A), but also make the feature vectors representing proteins constructed uniquely from the protein pair representing vector.

B. Training negative datasets

With the above feature vector construction scheme, the PPIs prediction task is ready to be formalized as a classification problem with the publicly available PPIs as the positive samples, and the others as the negative samples. The training data imbalance problem will arise, because there are only a relatively small number of known PPIs. This situation will make the SVM ineffective in determining the class boundary [20]. To maintain a balance between training positive and negative datasets in SVM training procedure, we introduce two types of negative datasets. The first one is the randomly negative dataset. The randomly negative samples are sampled randomly from the complementary graph of the known PPI network. Comparing with the method for generating the training negative dataset with the help of the functional annotation of proteins, this randomly generating scheme for training negative data can lead to unbiased estimates of prediction accuracy [21]. The second one is the artificial negative dataset. The artificial negative samples are constructed by uShuffle based on the positive datasets, which is successfully used in $S.cerevisiae$ PPIs prediction [23].

C. Benchmark datasets and SVM implementation

Here, PPIs on two different organisms: $E.coli$ and $S.cerevisiae$ are used to validate the performance of the proposed predictive models. The detailed information of these benchmark datasets can be found in Table 1 in [10]. The protein sequences are download from the RefSeq database of NCBI. In addition, the interactions which contain missing proteins in the corresponding proteome sequence datasets are excluded. Thus the number of interactions is 6954 and 6635 for $E.coil$ and $S.cerevisiae$, respectively.

We train the two-class SVM with denoising CTF and CTF by using $LibSVM$ [24]. In the implementation of two-class SVM, the RBF kernel function is used. The penalty parameter C and the RBF kernel parameter γ are optimized by grid search approach with 3-fold cross-validation. To evaluate the performance of our methods, we use the 10-fold cross-validation. The performances of our proposed methods are evaluated by the following evaluation criterions: AUC (area under the receiver operating curve (ROC) curve [25], Accuracy (Acc) = $\frac{TP+TN}{TP+TN+FP+FN}$, Sensitivity (Sn) = $\frac{TP}{TP+FN}$, Specificity (Sp) = $\frac{TN}{TN+FP}$, Precision (Pre) = $\frac{TP}{TP+FP}$, and F-measure= $\frac{2\times Sn\times Sp}{Sn+Sp}$. Here TP is the number of protein pairs correctly predicted to interact, FP is the number of protein pairs predicted to interact but actually not. And TN is the number of protein pairs don't interact and predicted correctly, FN is the number of protein pairs predicted not to interact but actually interact.

III. Results

A. The performance on the randomly negative datasets

We first test the effect of two kinds of denoising CTF ($Denoising^{Hao}$ CTF and $Denoising^{Yu}$ CTF) on $E.coli$ and $S.cerevisiae$ randomly negative datasets, respectively. The evaluation criterions obtained by denoising CTF and CTF on $E.coli$ and $S.cerevisiae$ randomly negative datasets when the corresponding F-measure is the largest are shown in Table 1. From Table 1, we can see that, on both $E.coli$ and $S.cerevisiae$ randomly negative datasets, $Denoising^{Yu}$ CTF outperforms $Denoising^{Hao}$ CTF with high AUC and all other criterions. However, both $Denoising^{Yu}$ CTF and $Denoising^{Hao}$ CTF perform worse than the CTF. For example, on $E.coli$ randomly negative dataset, comparing with the CTF, the AUC and Sn obtained by $Denoising^{Yu}$ CTF decrease by one percent, and the other criterions decrease by more or less to a certain extent. On $S.cerevisiae$ randomly negative datasets, comparing with the CTF, the AUC, Acc, Sp, Pre and F-measure obtained by $Denoising^{Yu}$ CTF decrease by two or three percent, and the Sn drops by seven percent. These results suggest that, on randomly negative dataset, the performance of $Denoising^{Hao}$ CTF and $Denoising^{Yu}$ CTF are not as good as that of CTF, and even worse than it. That is, the denoising procedure can not improve the performance of PPIs prediction.

B. The performance on the artificial negative datasets

We then test the effect of two kinds of denoising CTF on $E.coli$ and $S.cerevisiae$ artificial negative datasets, respectively. The evaluation criterions obtained by denoising CTF and CTF on $E.coli$ and $S.cerevisiae$ artificial negative datasets when the corresponding F-measure is the largest are shown in Table 2. Table 2 show that, on both $E.coli$ and $S.cerevisiae$ artificial negative datasets, $Denoising^{Yu}$ CTF outperforms $Denoising^{Hao}$ CTF with high AUC and all other criterions. However, both $Denoising^{Yu}$ CTF and $Denoising^{Hao}$ CTF perform worse than the CTF. For example, for $E.coli$, on 1-let dataset, comparing with the CTF, the AUC obtained by $Denoising^{Yu}$ CTF decreases by one percent, Acc and F-measure decrease by three percent, Sn drops by six percent, and Sp and Pre are nearly same as the CTF obtained. On 2-let dataset, comparing with the CTF, AUC, Acc, Sn, Pre and F-measure obtained by $Denoising^{Yu}$ CTF decrease by more than one percent, and Pre is nearly same as the CTF obtained. These results suggest that, on $E.coil$ artificial negative dataset, by introducing the denoising formulas, the CTF-based PPIs prediction performance can not be improved.

TABLE I
THE PERFORMANCE COMPARISON OF DENOISING CTF WITH CTF ON RANDOMLY GENERATED NEGATIVE SET.

Organism	Encoding methods	Evaluation criterions					
		AUC	Acc	Sn	Sp	Pre	F-measure
E.coli	CTF	**0.886**	**0.797**	**0.794**	**0.799**	**0.798**	**0.797**
	$Denoising^{Hao}$ CTF	0.849	0.763	0.726	0.799	0.784	0.761
	$Denoising^{Yu}$ CTF	0.877	0.791	0.782	0.799	0.796	0.791
S.cerevisiae	CTF	**0.948**	**0.880**	**0.879**	**0.927**	**0.909**	**0.882**
	$Denoising^{Hao}$ CTF	0.924	0.853	0.806	0.899	0.889	0.851
	$Denoising^{Yu}$ CTF	0.929	0.862	0.824	0.899	0.891	0.861

On $S.cerevisiae$ 1-let datasets, comparing with the CTF, although Acc, Sn, Pre and F-measure obtained by $Denoising^{Yu}$ CTF increase by one percent, the AUC increases only 0.1 percent. On 2-let dataset, comparing with the CTF, although Acc, Sn and F-measure obtained by $Denoising^{Yu}$ CTF increase by two to four percent, the AUC only has 0.1 percent improvement. These results suggest that, on $S.cerevisiae$ artificial negative dataset, the $Denoising^{Yu}$ CTF outperforms the CTF with 0.1 percent AUC improvement. However, this little improvement is insufficient to support the fact that the prediction performance can be improved by introducing the denoising procedure.

C. The performance of denoising formula on the gene level

Hao's and Yu's formulas were primal proposed on the gene level in [13]. By introducing these denoising formulas, the well-grouped phylogenetic trees have been constructed. Therefore, we test the effect of these two denoising formulas on the gene level on $E.coli$ and $S.cerevisiae$ randomly negative datasets, respectively. That is, we encode protein sequences by codon composition, and introduce the denoising formulas (1) and (2) as the noise of the codon composition, respectively, then apply the Signal-to-Noise Ratio as the input vectors for representing the proteins. The equation (4) is also applied as the representation vector for protein pairs, and it is denoted as denoising codon. The evaluation criterions obtained by denoising codon and codon composition on $E.coli$ and $S.cerevisiae$ randomly negative datasets when the corresponding F-measure is the largest are shown in Table 3. From Table 3, we can see that, on both $E.coli$ and $S.cerevisiae$ randomly negative datasets, $Denoising^{Yu}$ codon outperforms $Denoising^{Hao}$ codon with high AUC and all other criterions. However, both $Denoising^{Yu}$ codon and $Denoising^{Hao}$ codon perform worse than codon itself. For example, on $E.coli$ randomly negative dataset, comparing with codon composition, the AUC, Acc and F-measure obtained by $Denoising^{Yu}$ codon decrease by three percent, Sn drops by more than seven percent, and Pre drops by one percent. On $S.cerevisiae$ randomly negative datasets, comparing with the codon composition, the AUC, Acc and F-measure obtained by $Denoising^{Yu}$ codon decrease by three percent, Sn drops by nearly seven percent, and Pre drops by one percent. These results suggest that, on gene level, the PPIs prediction performance can also not be improved by introducing the denoising methods.

IV. Discussion and Conclusion

In this paper, to improve the performance of PPIs prediction, we introduce the denoising idea which is proved to be useful in construction of phylogenetic trees into the prediction algorithm. Specially, we first encode the given protein sequence by the composition vector, and then introduce two denoising formulas proved to be useful in phylogenetic tree construction as the noise vector, finally apply the Signal-to-Noise Ratio as the input vector for representing the given protein. The concatenation of arithmetical and geometric average of protein vectors is used as the protein pair representation vector, which can not only maintain symmetry, but also make the protein representing vectors constructed uniquely from the protein pair representing vector. We test the effect of the denoising vectors on $E.coli$ and $S.cerevisiae$ randomly and artificial negative datasets, and compare it with the primal composition vectors. The evaluation criterions obtained by both two denoising vectors are not improved. These results suggest that, although the denoising methods can improve the performance of phylogenetic trees construction, it can not improve the performance of PPIs prediction. That is, what is noise is dependent on the applications.

The lower accuracy of denoising methods here may be caused by that the CTF first classifies twenty amino acids into seven classes based on the dipoles and volumes of the side chains, and then apply conjoint triad composition to represent the given protein. That is, the denoising is already done by reducing the classes of amino acids, and further denoising will make information shrink. In the future, we will test the denoising effect based on protein composition vector without fusion amino acids, and further test the conclusions obtained in this article.

Although the dimension reduction is already done by fusing twenty amino acids into seven classes, the most elements of CTF might be redundant and irrelevant, and might not contribute significantly to the PPIs prediction. Therefore, the performance of PPIs prediction might be improved by selecting only relevant elements. Future work will try to do some related works by introducing some wonderful feature selection approaches.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 10801131, No.10971223).

TABLE II
THE PERFORMANCE COMPARISON OF DENOISING CTF WITH CTF ON SHUFFLED NEGATIVE SET

Organism	Encoding	Evaluation criterions					
	methods	AUC	Acc	Sn	Sp	Pre	F-measure
E.coli	1let-CTF	**0.957**	**0.891**	**0.882**	**0.899**	**0.898**	**0.891**
	1let-$Denoising^{Hao}$ CTF	0.910	0.824	0.848	0.799	0.809	0.823
	1let-$Denoising^{Yu}$ CTF	0.940	0.860	0.820	0.899	0.891	0.858
	2let-CTF	**0.936**	**0.856**	**0.892**	**0.899**	**0.890**	**0.853**
	2let-$Denoising^{Hao}$ CTF	0.904	0.818	0.836	0.799	0.807	0.818
	2let-$Denoising^{Yu}$ CTF	0.927	0.841	0.882	0.879	0.887	0.839
S.cerevisiae	1let-CTF	0.956	0.884	0.868	**0.899**	0.896	0.883
	1let-$Denoising^{Hao}$ CTF	0.950	0.885	0.871	0.879	0.897	0.885
	1let-$Denoising^{Yu}$ CTF	**0.957**	**0.899**	**0.879**	0.879	**0.919**	**0.899**
	2let-CTF	0.936	0.850	0.801	0.899	0.888	0.847
	2let-$Denoising^{Hao}$ CTF	0.935	0.866	0.837	0.899	0.893	0.867
	2let-$Denoising^{Yu}$ CTF	**0.937**	**0.872**	**0.845**	**0.899**	**0.894**	**0.871**

TABLE III
THE PERFORMANCE COMPARISON OF DENOISING CODON COMPOSITION WITH CODON COMPOSITION ON RANDOMLY GENERATED NEGATIVE SET.

Organism	Encoding	Evaluation criterions					
	methods	AUC	Acc	Sn	Sp	Pre	F-measure
E.coli	codon	**0.897**	**0.812**	**0.825**	**0.799**	**0.805**	**0.812**
	$Denoising^{Hao}$ codon	0.855	0.766	0.732	0.799	0.785	0.764
	$Denoising^{Yu}$ codon	0.868	0.775	0.751	0.799	0.789	0.774
S.cerevisiae	codon	**0.942**	**0.881**	**0.863**	**0.899**	**0.896**	**0.881**
	$Denoising^{Hao}$ codon	0.887	0.811	0.783	0.899	0.802	0.806
	$Denoising^{Yu}$ codon	0.911	0.847	0.794	0.899	0.888	0.843

REFERENCES

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

[2] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, **98**, 4569–4574.

[3] Gavin, A.C., Boche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A. and Cruciat, C. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

[4] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., 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., Sørensen, B.D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

[5] 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. and Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science*, **193**, 2101–2105.

[6] Shen, J. W., Zhang, J., Luo, X. M., Zhu, W. L., Yu, K. Q., Chen, K. X., Li, Y. X., and Jiang, H. L. (2007) Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, **104**, 4337–4341.

[7] Ben-Hur, A. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**: i38–i46.

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

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

[10] Najafabadi, H., Salavati, R. (2008) Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology*, 9:R87.

[11] Leslie, C., Eskin, E., Noble, W.S. (2002) The spectrum kernel: a string kernel for SVM protein classification. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp 564–575.

[12] Yu, C.Y., Chou, L.C., Chang, D.T.H. (2010) Research article predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics*, 11:167.

[13] Chan, R.H.F. Wang, R.W., Wong, J.C.F. (2010) Maximum Entropy Method for Composition Vector Method. Published Online: 23 DEC 2010, DOI: 10.1002/9780470892107.ch27.

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

[15] Vapnik, V., 1998. Statistical Learning Theory. Wiley.

[16] Deng, N.Y, Tian, Y.J, Zhang, C.H. Support Vector Machines: Theory, Algorithms, and Extensions. CRC Press, 2012. (In Press)

[17] Noble, W.S. (2004) Support vector machine applications in computational biology. In Schoelkopf,B., Tsuda,K. and Vert,J.-P. (eds), Kernel Methods in Computational Biology. MIT Press, Cambridge, MA, pp. 71–92.

[18] Hao, B.L., Qi, J., Wang, B. (2003) Prokaryotic phylogeny based on complete genomes without sequence alignment. *Modern Physics Letters B*, **2**, 1–4.

[19] Yu, Z.G. Zhou, L. Q., Anh, V., Chu, K.H., Long, S.C., Deng, J.Q. (2005) Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment. *Journal of Molecular Evolution*, **60**, 538–545.

[20] Wu, G., Chang, E.Y. (2003) Class-boundary alignment for imbalanced dataset learning. In ICML 2003 Workshop on Learning from Imbalanced Data Sets.

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

[22] Jiang, M., Anderson, J., Gillespie, J., Mayne, M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192.

[23] Guo, Y.Z., Yu, L.Z., Wen, Z.N., Li, M.L. 2008 Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Research*, **36**, 3025–3030.

[24] Hsu, C.W., Chang, C.C., Lin, C.J. (2007) A practical guide to Support Vector Classfication. Available from: *http://www.csie.ntu.edu.tw/ cjlin.*

[25] Gribskov, M., Robinson, N.L. (1996) Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Computers and Chemistry*, **20**, 25–33.

[26] Qi, J., Wang, B., Hao, B.L. (2004) Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach. *Journal of Molecular Evolution*, **58(1)**, 1–11.