Prediction of protein-protein interaction based only on coding sequences

Yongcui Wang[1] Jiguang Wang[2] Zhixia Yang[3]
Naiyang Deng[1,*]

[1]College of Science, China Agricultural University, Beijing, China, 100083
[2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science,
Beijing, China, 100080.
[3]College of Mathematics and System Science, Xinjiang University, Urumuchi, China, 830046

Abstract Identification of the interactions among proteins is crucial to illustrate their functions, and further, can help to understand the biological processes and provide insights into the mechanisms of diseases. It has became one of the most challenging and important task in the post-proteomic researches. Comparing with costly and time-consuming biochemical experiments, the computational methods have attracted much attention due to their low costing and competitive performance. In this paper, we develop a new method based only on coding sequence to identify novel protein-protein interactions (PPIs). To deal with the imbalance problem, we select suitable negative training set before implementing support vector machine. The proposed method is validated on PPIs data of $Plasmodium falciparum$ and $Escherichiacoli$, and yields a predictive accuracy of 93.38% and 92.30% respectively. When performed on independent $Plasmodium falciparum$ and $Escherichiacoli$ datasets, our method displays promising generalization ability.

Keywords Protein-protein interactions; Codon usage; Support vector machine; Imbalance problem

1 Introduction

Identification of the interactions among proteins is crucial to illustrate their functions, and further, can help to understand the biological processes and provide insights into the mechanisms 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 PPIs analysis, including yeast two-hybrid systems [5, 11], mass spectrometry [6, 10], protein chip [23] and so on. Comparing with these costly and time-consuming biochemical experiments, the computational methods have attracted much attention due to their low costing and competitive performance.

Current computational methods for PPI prediction require a large amount of genomic data sources, such as, Gene Ontology (GO) annotations, gene expressions, evolution information and son on. However, usually some of them is not available for some important

[*] Corresponding author. E-mail: dengnaiyang@vip.163.com

genes. Sequence-based methods then become popular because they only demand the information of amino acid sequences, but the highest accuracy of these methods is only $\backsim$ 80% [9], such as the methods by Martin et al. [16], Chou and Cai [3]. Shen et al.[19] have developed a conjoint triad feature construction method, and with SVM as the classifier, yields a high prediction accuracy of 83.9%, when applied to predict human PPIs.

The knowledge that codon usage is correlated with expression level has been widely acceptable [12], and the hypothesis of some function-specific codon preferences has been confirmed by experiments [4]. Naiafabadi and Salavati [17] then proposed an sequence-based method by extract the sequence features in the genome instead of the proteome. By using a naïve Bayesian network to combine the frequencies of all codons, the encouraging predictive results were obtained. However, the best results in their work were obtained by incorporating the other genomic data. So we want to develop a computational method for predicting PPIs based only on coding sequence.

As an excellent machine learning method, support vector machines (SVMs), motivated by statistical learning theory [21, 22], have been proven successful on many different classification problems in bioinformatics [18]. 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 codon usage is used to predict PPIs. To deal with the imbalance problem, the suitable negative training set is selected before implementing two-class SVM. The proposed method is validated on the PPIs data of $Plasmodium falciparum$ and $Escherichiacoli$, and yields a predictive accuracy of 93.38% and 92.30% respectively. It is further evaluated on $Plasmodium falciparum$ and $Escherichiacoli$ independent PPIs datasets, and achieves the test sensitivity of 60.49% and 84.4% respectively.

2 Materials and Methods

In this section, we describe the benchmark datasets and the predictive model in this paper.

2.1 Materials

Here, PPIs on two different organisms: $Plasmodium falciparum(P.falciparum)$ and $Escherichiacoli(E.coli)$ are used to validate the performance of the proposed predictive models. $P.falciparum$ is a eukaryote, while $E.coli$ is a prokaryote. The detailed information of these benchmark datasets can be found in Table 1 in [17]. The genome sequences for them can also be downloaded from [17]. Specially, for $P.falciparum$, the benchmark positive and negative sets are the same as the gold standard sets in [17], while for $E.coil$, we exclude the interactions which contain missing proteins in the corresponding genome sequence datasets. Thus the number of interactions is 7689 and 6954 for $P.falciparum$ and $E.coil$ respectively.

2.2 Methods

2.2.1 Construction of input feature vectors

We represent each open reading frame (ORF) by a binary space (V,D) ,where $V = (v_1,\cdots,v_m)$ represents the vector space of the sequence features, and each feature v_i represents a kind of codon; $D = (d_1,\cdots,d_m)$ is the frequency vector corresponding to V, and the value of the d_i is the frequency of type v_i appearing in the corresponding ORF. Because there are 64 kinds of codon, the size of V should be 64; thus, $m = 64$. The detailed definitions for (V,D) are illustrated in [17]. Overall, a 64-dimension vector corresponding with each ORF has been constructed, which is called as codon modes.

Another method for encoding feature vector of ORF is incorporating 64 codons into 20 amino acids, that is, using a 20-dimensional vector to represent ORF, each element of this vector is the frequency of a sort of amino acid appearing in the corresponding ORF, and we call this kind of vector as codon merger modes.

There are two ways to construct the feature vectors which can be applied to represent protein-protein pairs:

1. Concatenating the codon or codon merger modes:

A pair of protein A and protein B is represented by concatenating the codon or codon merger modes D_A and D_B. That is the input feature vector F_{AB} for a protein pair A-B is calculated as follows:

$$F_{AB} = D_A \oplus D_B, \tag{1}$$

where $\oplus$ is the concatenation operator. As the authors do in [20], to make predictive results for protein pair A-B identical to B-A, we train and test on both F_{AB} and F_{BA}, and report the average predictive results in numerical experiments.

2. Distance of protein pairs:

A pair of protein A and protein B is represented by a distance vector. $D_k^{AB} = |d_k^A - d_k^B|$ is used to measure the distance between protein A and B. So the input feature vector F_{AB} for a protein pair A-B is calculated as follows:

$$D_{AB} = (d_1^{AB}, \cdots, d_m^{AB})^{\mathrm{T}},\ d_k^{AB} = |d_k^A - d_k^B|, \tag{2}$$

where $m = 64$ for codon modes, $m = 20$ for codon merger modes.

For codon modes, we use the distance of protein pair to represent a pair of proteins, which is the same as the authors did in [17]. For codon merger modes, the concatenation operator is used. Because the dimension of distance with respect to codon merger modes is 20, it may be not enough to generalize a good predictor with respect to SVM, since SVM would like to deal with high dimensional dataset. We verify this in numerical experiments. Thus SVM_{codon} is used to denote the SVM with codon usage (using 64-dimensional vector to represent protein-protein pairs), and $SVM_{codon\ meger}$ is used to denote the SVM with codon merger usage (using 40-dimensional vector to represent protein-protein pairs).

2.2.2 Predictive model

The two class problem is constructed by using the feature constructing methods described in previous section. The training set is:

$$T = \{(x_i, y_i), i = 1 \cdots, l\},\ x_i \in R^q,\ y_i = \{-1, +1\}, \tag{3}$$

where y_i equals to 1, if there is an interaction between the corresponding protein pair, denoted by positive pair, and equals to -1, if not, denoted by negative pair.

To maintain a balance between positive and negative training data in SVM training procedure, we select a suitable set of training negative data points from the whole negative set firstly, and then perform the two-class SVM. The suitable negative set should be a good representation of the entire negative set, so we select the data points which can embody the main distribution of the whole dataset. Specially, firstly, calculate the mean vector of the whole negative data points; secondly, compute the distance between each data point and the mean vector; finally, select the data points far from the mean vector and make the chosen dataset with the nearly same size of the positive set. After selecting the suitable negative set, we implement two-class SVM to predict PPIs. This method is denoted by $SVM-SN$

We also randomly select the negative set from the whole training negative set, and then use two-class SVM on training positive set and this random negative set to perform the predictive task. It is denoted by SVM-random. We compare the performance of SVM-SN with the average results of SVM-random in experimental section.

2.3 Cross-validation and parameter selection

To evaluate the performance of SVM-SN and SVM-random, we perform the 10 fold cross-validation procedure as the authors did in [17]: the benchmark dataset is randomly split into 10 subsets with roughly equal size, each subset are taken in turn as the test set, the remaining 9 subsets are used for training.

In SVM-SN and SVM-random, the RBF kernel function is used. The penalty parameter C and the RBF kernel parameter σ are determined by 5-fold cross-validation before implementing each method on PPIs datasets.

2.4 Evaluation criterions

The performance of proposed method is evaluated by using receiver operating curve (ROC) [7]. Furthermore, the other criterions, such as AUC (area under the ROC curve), $sensitivity = TP/(TP+FN)$, $specificity = TN/(TN+FP)$, $precision = TP/(TP+FP)$, $accuracy = (TP+TN)/(TP+TN+FP+FN)$ are also used to display the performance of the proposed predictive methods.

3 Results and discussions

3.1 The performance on $P.falciparum$

We plot the ROC cure and the evaluation criterions for each method on $P.falciparum$ in Figure 1. As shown in Figure 1, the performance of $SVM_{codon} - random$ is comparable with that of PIC, while $SVM_{codon} - SN$ outperforms PIC, and $SVM_{codon\ merger} - SN$ achieves the best predictive performance. $SVM_{codon\ meger} - SN$ outperforms $SVM_{codon} - SN$ with not only high AUC, but also high other criterions except for sensitivity. We want to test whether the performance can be improved by integrating other data sources, such as gene expression data. The microarray data for $P.falciparum$ is downloaded 14 microarray experiments (Affymetrix S98 chipset, GPL 90 GEO platform) from GEO. Missing expression values are filled by the mean of the expression values in other experiments. Genes without corresponding identifiers are discarded. Finally, the 14 experiments covered a total of 4797 unique proteins. We concatenate the gene microarrays and codon merger modes to represent proteins. $SVM_{codon\ merger+gene\ expression} - SN$ is used to denote the $SVM - SN$ with codon merger modes and gene microarrays. From Figure 1, we can

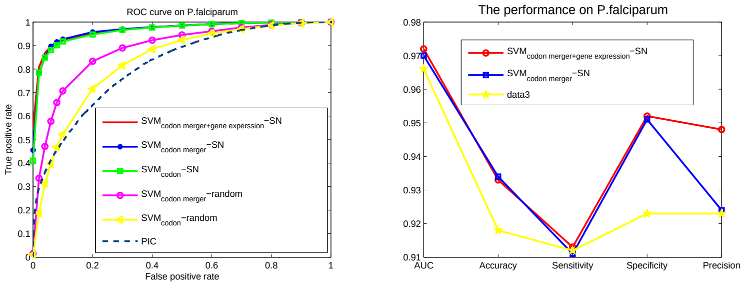

Figure 1: The performance of poposed methods for $P.falciparum$.

imply that if some other genome data sources can be integrated into the predictive model, the performance can be further improved.

We evaluate the performance of $SVM_{codon} - SN$ and $SVM_{codon\ merger} - SN$ on an independent dataset (Hughes et all, 2005) [13] which is generalized by yeast two-hybrid experiments. This dataset includes 2,823 interactions which contains 1,267 proteins. By training on the benchmark dataset, accuracy on the test dataset are 65.14% and 60.49% by implementing $SVM_{codon} - SN$ and $SVM_{codon\ meger} - SN$ respectively. We generate the negative set by using the positive dataset: for example, AB and IJ are interaction pairs, thus AI, AJ, BI, or BJ could be the negative pairs [19], that is there are 11,288 $(4 \times (2823 - 1))$ non-interactions which can be incorporated into the test dataset. Two test datasets are used to test the performance of our methods: the first dataset contains 2,823 interactions and randomly selected 2,823 non-interactions, while the second one contains 2,823 interactions and the entire 11,288 non-interactions. We train $SVM_{codon\ merger} - SN$ and $SVM_{codon} - SN$ on the benchmark dataset, and test on two test datasets respectively. For $SVM_{codon} - SN$, the AUCs are 0.515 and 0.514 for two test datasets respectively, while for $SVM_{codon\ merger} - SN$, the AUCs are 0.504 and 0.501. That is, although two test datasets contain positive (Pos) and negative (Neg) data points with different ratio (1:1 and 1:4), there have been a little difference on the test results between them for both $SVM_{codon\ merger} - SN$ and $SVM_{codon} - SN$. However, with respect to the low FPR (false positive rate), the TPR (true positive rate) of $SVM_{codon\ merger} - SN$ is higher than that of $SVM_{codon} - SN$ on both two test datasets. For example, for the first test dataset ($Pos : Neg = 1 : 1$), when FPR reaches 0.02, the TPRs are 0.018 and 0.023 for $SVM_{codon} - SN$ and $SVM_{codon\ merger} - SN$ respectively. However, for both $SVM_{codon\ merger} - SN$ and $SVM_{codon} - SN$, the AUCs are just more than 0.5. It implies that both codon and codon merger are not suitable for the physical interaction but prefer to co-pathway interaction prediction.

3.2 Performance on $E.coli$

For $E.coli$, the ROC curves and evaluation criterions are drawn for proposed methods in Figure 2. It shows that, the performance of $SVM_{codon} - random$ is nearly the same as that of PIC, whereas $SVM_{codon} - SN$ performs better than PIC, while $SVM_{codon\ merger} - SN$ outperforms all other methods. Furthermore, SVM with codon merger performs better than that with codon modes not only on randomly selected negative set but also on the

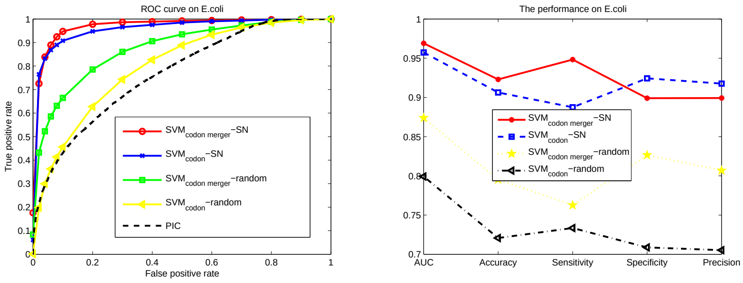

Figure 2: The performance of proposed methods for $E.coli$.

well-chosen negative set.

We test on the independent dataset which is collected by Andres Leon, E. e in [1]. This dataset contains 14,536 experimented physical interactions, by deleting the interactions which is present in the training benchmark dataset, remains 10,529 interactions. We generate the negative set in the same way as in $P.falciparum$ subsection, and 42,112 non-interactions are generated. We use two test datasets to test the performance of our methods: the first test dataset contains 10,529 interactions and randomly selected 10,529 non-interactions, while the second one contains 10,529 interactions and 42,112 non-interactions. We train $SVM_{codon\ merger} - SN$ and $SVM_{codon} - SN$ on the benchmark $E.coli$ dataset respectively, and test on these two test datasets. For $SVM_{codon\ merger} - SN$, the sensitivity is 84.4%, while is 74.6% for $SVM_{codon} - SN$. These results imply that SVM with codon merger modes has the good generalization ability in prediction of PPIs on $E.coli$.

As mentioned above, SVM-SN with codon merger modes is the best predictor on both two kinds of organisms. Although, in [17], the authors also combined codon merger modes with their predictors (called as PI-A), and demonstrated that PI-A had a worse performance than the predictor with codon usage, while SVM with codon merger outperforms with codon modes in our study. The reason behind these results may be that, they applied distance of codon merger modes to represent a pair of proteins, while we use the concatenation operator to formulate the feature vector for a pair of proteins, and SVM is promising in the high-dimensional data. To verify that, we train on the benchmark $P.falciparum$ and $E.coli$ datasets with 5-fold cross-validation by SVM-SN. Under the optimal parameters, for $P.falciparum$, the cross accuracies are 91.02% and 94.72% for distance and concatenation respectively, while for E.coli, the cross accuracies are 91.62% and 94.33% for distance and concatenation respectively.

4 Conclusion

In this paper, the sequence-based methods are proposed to predict PPIs. We extract sequence features in the genome instead of the proteome. Specially, codon and codon merger modes are used to represent proteins, and the distance and concatenation operator are applied to formulate the feature vectors for a pair of proteins. By using $SVM_{codon\ merger} - SN$ in imbalance problem, the significant improvement in prediction can be obtained on both two kinds of organisms. For testing the generalization ability of $SVM_{codon\ merger}-SN$, we train on the benchmark datasets and test on the independent interactions of $P.falciparum$ and $E.coli$ respectively. For $P.falciparum$, the test accuracy on physical interaction generalized by yeast two-hybrid experiments is 60.49%. For $E.coli$, the average test sensitivity is about 84% on the experimental physical interactions. We also test whether the performance can be improved by integrating some other data sources such as gene expression data. By incorporating 14 $P.falciparum$ microarray experiments, the predictive results can be improved. Future work can introduced the classical methods for integration of diverse data, including kernel-level integration [14], ensemble learning [8] and naïve Bayesian network [17].

Efficient feature construction is important in determining the performance of a predictive method, thus future work can focus on how to improve feature extraction method, including using the conjoint triad feature extracted from sequence in proteome like the authors did in [19] to represent proteins. Future work can also be included to use more efficient and simple SVM model on imbalance classification problem to implement prediction task, such as SVM with an offset [15] an so on.

Acknowledgments

This work is supported by the Key Project of the National Natural Science Foundation of China (No. 10631070), the National Natural Science Foundation of China (No. 10801112) and the Ph.D Graduate Start Research Foundation of Xinjiang University Funded Project (No.BS080101).

References

[1] Andres Leon, E., Ezkurdia, I., GarcÍła, B., Valencia, A., Juan, D. (2009) EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Research*, 37, D629-D635.

[2] Ben-Hur, A., and Noble, W. S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21, i38–i46.

[3] Chou, K.C. and Cai, Y.D. (2006) Predicting proteinÍCprotein interactions from sequences in a hybridization space. *Journal of Proteome Research*, 5, 316–322.

[4] Dittmar, K. A., Sorensen, M. A., Elf, J., Ehrenberg, M. and Pan, T. (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO reports*, 6, 151–157.

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

[6] 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.

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

[8] Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.(2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biology*, 9(S3).

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

[10] 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.

[11] 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.

[12] Jansen, R., Bussemaker, H.J. and Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic acids research*, 31, 2242–2251.

[13] LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I.,Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, 10, 103–107.

[14] Lanckriet. G., Deng, M., Cristianini, N., et al. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. Presented at the Pacific Symposium on Biocomputing.

[15] Li, B. Hu, J. Hirasawa, K. Sun, P., Marko, K., 2006. 'Support vector machine with fuzzy decision-making for real-world data classification'. In IEEE World Congress on Computational Intelligence, Int. Joint Conf. on Neural Networks, Canada.

[16] Martin, S., Roe, D. and Faulon, J.L. (2005) Predicting proteinÍC protein interactions using signature products. *Bioinformatics*, 21, 218–226.

[17] Najafabadi, H. S. and Salavati, R. (2008) Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology*, 9, R87–R95.

[18] 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.

[19] 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.

[20] Soong, T., Wrzeszczynski, K.O., Rost, B. (2008) Physical protein-protein interactions predicted from microarrays. *Bioinformatics*, 24, 2608–2614.

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

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

[23] 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.