Amino Acids (2010) 39:1595–1599
DOI 10.1007/s00726-010-0588-1

SHORT COMMUNICATION

Predicting protein–protein interactions from protein sequences using meta predictor

Jun-Feng Xia · Xing-Ming Zhao · De-Shuang Huang

Received: 24 November 2009 / Accepted: 27 March 2010 / Published online: 13 April 2010

Abstract A novel method is proposed for predicting protein–protein interactions (PPIs) based on the meta approach, which predicts PPIs using support vector machine that combines results by six independent state-of-the-art predictors. Significant improvement in prediction performance is observed, when performed on *Saccharomyces cerevisiae* and *Helicobacter pylori* datasets. In addition, we used the final prediction model trained on the PPIs dataset of *S. cerevisiae* to predict interactions in other species. The results reveal that our meta model is also capable of performing cross-species predictions. The source code and the datasets are available at http://home.ustc.edu.cn/~jfxia/Meta_PPI.html.

Keywords Protein–protein interactions · Support vector machine · Meta approach · Protein sequence · Feature representation

Electronic supplementary material The online version of this article (doi:10.1007/s00726-010-0588-1) contains supplementary material, which is available to authorized users.

J.-F. Xia · D.-S. Huang (✉)
Intelligent Computing Laboratory,
Hefei Institute of Intelligent Machines,
Chinese Academy of Sciences,
P.O. Box 1130, Hefei, Anhui 230031, China
e-mail: dshuang@iim.ac.cn

J.-F. Xia
School of Life Science,
University of Science and Technology of China,
Hefei, Anhui 230027, China

X.-M. Zhao
Institute of System Biology, Shanghai University,
Shanghai 200444, China

Introduction

Protein–protein interactions (PPIs) play important roles in most cellular processes, such as transcription regulation, signal transduction (Zhao et al. 2008a), and recognition of foreign molecules. Knowledge of PPIs can provide insight into protein functions (Zhao et al. 2008b, c), lead to a better understanding of disease mechanisms and suggest novel methods for designing drugs that modulate specific disease pathways. Although high-throughput technologies have opened new prospects to systematically identify PPIs, the interaction networks currently detected are far from complete. In addition, these approaches suffer from both false negatives and false positives, therefore, there is a strong motivation to develop computational approaches to complement experimental methods.

Methods for inferring PPIs (Pitre et al. 2008) using a variety of data sources such as the genomic context (Enright et al. 1999), protein domains (Chen and Liu 2005; Xia et al. 2007), and protein structures (Lu et al. 2002) have been explored. There are also approaches that integrate different types of data such as Bayesian method (Jansen et al. 2003) and kernel based method (Ben-Hur and Noble 2005). However, these methods are not universal, because they depend on the prior information of the protein pairs. Recently, a unique category of prediction methods that use protein sequence information alone to infer PPIs has been put forward, and the results have demonstrated their feasibility (Bock and Gough 2003; Chou and Cai 2006; Guo et al. 2008; Martin et al. 2005; Shen et al. 2007; Shi et al. 2010; Xia et al. 2010; Yang et al. 2010). Nevertheless, most of these methods are not robust and reliable in some cases because they just use one single classification system (Chou and Shen 2007a, b; Nanni 2005; Nanni and Lumini 2006; Shen and Chou 2007; Liu et al. 2008). On the other

hand, with the number of prediction methods available increasing rapidly, the meta method has been proved to be more effective compared to individual classifiers in many fields (Bujnicki et al. 2001; Ishida and Kinoshita 2008; Saini and Fischer 2005).

Here, we report a meta method for improving the quality of PPI prediction. Our method predicts PPIs by integrating the results of six excellent sequence-based methods. Empirical studies have shown that our method yields better prediction accuracy than those individual methods used in the meta predictors when performed on *Saccharomyces cerevisiae* datasets. We also evaluated the final prediction model by preparing the cross-species data as the test set, which further demonstrates the effectiveness of our method.

Materials and methods

We evaluated the performance of the proposed approach using the data (please see supplementary material) as investigated in Guo et al. (2008).

To adopt meta method, six single sequence-based classifiers are constructed by combining different feature representation methods and support vector machine (SVM). Here we used auto covariance (AC) (Guo et al. 2008), conjoint triad (CT) (Shen et al. 2007), local descriptor (LD) (Yang et al. 2010), Geary autocorrelation (GA) (Sokal and Thomson 2006), Moran autocorrelation (MA) (Xia et al. 2010), and Normalized Moreau-Broto autocorrelation (NA) (Feng and Zhang 2000) as feature representation methods. These feature representation methods (see supplementary material) were selected according to their prediction accuracy in previous studies.

Construction of single classifier

After different types of features for describing amino acids have been built to represent each protein pair, six prediction models are constructed based on SVM. They are termed AC–SVM, CT–SVM, LD–SVM, GA–SVM, MA–SVM and NA–SVM, respectively. The LIBSVM package (Chang and Lin 2001) was employed in this work to implement each individual predictor. The Gaussian Radial Basis Function kernel was exclusively used in the computations based on previous studies.

Meta prediction

Meta approach aims to improve the performance by pooling the predictions made by many individual predictors. The basic architecture of the proposed meta predictor is shown in Supplementary Fig. S1. Our meta prediction is built on the basis of six individual predictors and comprises two main steps. Firstly, the input amino acid sequences of a protein pair are submitted to each component predictor, and the prediction decision values of SVMs from all predictors are collected. In this study, we used the above six individual predictors. Because SVM is a binary classifier, the SVM outputs two-state prediction results. However, the decision values of the SVM can be used as the prediction results for those component predictors. In fact, the decision value is the distance between each input vector and the decision plane. In general, a higher decision value indicates a more reliable prediction (Vapnik 1999). So it can be used to evaluate the reliability of the prediction. As a result, each predictor will perform its own prediction for each protein pair, and then the decision values of the SVMs for these predictors are obtained as the input vector of meta predictor. In the second step, the decision values of each protein pair from different predictors are collected and converted to an input vector for the meta predictor. Thus, the dimension of the input vector for the meta predictor corresponds to the number of individual predictors. Since a simple meta approach such as using a consensus or averaging the results of component predictors would be insufficient in this work, we still adopted SVM as the meta predictor in this study and employed the LIBSVM package to implement the meta predictor.

Results and discussion

Assessment of prediction capability

In order to achieve good experimental results, the capacity parameter C and the kernel parameter γ of the SVMs in the six component predictors and the meta predictor were tried using a grid search method in the range of $C = 2^{-2}$, 2^{-1},..., 2^{8} and $\gamma = 2^{-6}$, 2^{-1},..., 2^{2}, respectively. Considering the numerous samples used in this work, five fold cross-validation was used to investigate the training set, which can minimize the overfitting of the prediction model. The detailed values of C and γ of the six component predictors and meta predictor can be seen in the Supplementary material. Using the optimal parameters of C and γ, the PPI meta prediction model was built using the training set, and the performance of this model was evaluated by the test set. In order to test the robustness of the method, this process has been repeated five times using five training set and five test set as described in supplementary material.

Table 1 illustrates the average prediction performance of the meta predictor and other individual predictors across five runs. The meta method yields a PPI prediction model with the average sensitivity, precision and accuracy of 90.81, 94.38 and 92.70%, respectively. The average sensitivity, precision and accuracy of the individual

Table 1 The average prediction results of the test sets by the meta predictor and other six individual predictors across five runs

Methods	SN (%)	PE (%)	ACC (%)	MCC (%)	AUC (%)
AC–SVM	86.79	90.45	88.81	77.69	95.45
CT–SVM	88.95	92.18	90.70	81.45	96.43
LD–SVM	87.37	89.50	88.56	77.15	95.07
GA–SVM	83.82	86.84	85.55	71.15	92.49
MA–SVM	86.82	90.61	88.90	77.88	95.35
NA–SVM	86.71	89.98	88.52	77.10	94.98
Meta	90.81	94.38	92.70	85.47	97.71

component predictors were also calculated. It can be found that among the methods based on autocorrelation, the performance of MA–SVM method is better than the other two methods (i.e. NA–SVM and GA–SVM methods), which leads to the same conclusion as our previous study (Xia et al. 2010). Compared to other five individual models, the CT–SVM model achieves the best performance. The average prediction sensitivity, precision and accuracy are 88.95, 92.18 and 90.70%, respectively, which indicate that this model is indeed successful in predicting interactions using the conjoint triad method to describe the information of PPIs. However, the performance of meta predictor is better than the most accurate component predictor. The sensitivity, precision and accuracy values obtained from the meta classifier are 1.86, 2.2 and 2% higher than the CT–SVM model, respectively. This result shows the superiority of our meta model compared to any individual predictor. To better investigate whether or not the meta classifier can improve the performance of PPI prediction compared with the individual classifiers in the meta system, we drew the ROC curves of the test set for the meta method and the six individual methods (Supplementary Fig. S2). The curve of the meta predictor shows higher prediction performance on the whole range of FP rates, and it yields an AUC score of 97.71% and a MCC value of 85.47% (as shown in Table 1). These values are superior to those obtained from other prediction methods we examined. In other words, the meta predictor can decrease both false negatives and false positives compared to other methods. The results illustrate that the meta model is an accurate and efficient method for the prediction of PPIs. To sum up, we can readily conclude that as an ensemble classifier, the meta approach outperforms the excellent component classifier such as CT–SVM and AC–SVM with higher discrimination power.

Performance on independent dataset

As the meta model produced a good performance, we switched to evaluate the practical prediction capability of our approach. Firstly, we constructed the final prediction meta model using the whole data set (11,188 protein pairs) with the optimal parameters. And then the prediction performance of the final meta predictor was evaluated using a large independent data set (Guo et al. 2008) to obtain a more reliable assessment. These data consist of 10,920 non-interacting pairs in the same subcellular compartments (i.e. cytoplasm, endoplasmic reticulum, golgi apparatus, peroxisome and vacuole) and 11,474 interaction pairs. Supplementary Table S4 shows the comparisons of the results for different methods on independent dataset. The prediction performance, i.e. accuracy, sensitivity, precision, MCC and AUC achieved by the meta predictor are 87.36, 95.87, 78.42, 82.36 and 96.96%, respectively. It can be found that it is capable of improving the predictions of the other five individual methods, by increasing the AUC of 7.38, 5.25, 3.31, 9.02, 10.88 and 11.97% points. Interestingly, the sensitivity of our previous proposed method, i.e. LD model, is 100%, which means that all interaction pairs are predicted correctly. Although LD model could correctly identify interaction pairs, it makes more false positive (i.e. the non-interaction pairs incorrectly predicted as interaction pairs) predictions (precision = 73.90%) compared with the meta predictor (precision = 78.42%). The MCC and AUC of our meta model are 2.26 and 3.31% points higher than that of LD model, respectively. From these analyses, we observed that the meta method gives remarkably better prediction performance when compared to LD model. All these results demonstrate that our proposed method is able to achieve better performance on independent dataset.

Performance on cross-species dataset

In another validation, we tested the ability of our method for predicting PPIs in one species using the interactions from different species. Our model was trained on the *S. cerevisia* core subset in the DIP database; therefore, we chose the other four species in this database as our cross-species test dataset. The performance of the meta method in predicting such samples is summarized in Table 2. The prediction performance in *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens*, *and Mus musculus* achieved by our method are 85.77, 81.48, 90.16 and 88.82%, respectively. It is shown that the meta model can correctly predict the interacting pairs of three species with the accuracy of over 85%, while the *E. coli* subset has a relatively lower accuracy which still outperforms other methods. This result is superior to any other individual predictors of which we are aware. It demonstrates that the meta classifier is also able to achieve better performance towards cross-species dataset. We selected the PPIs data of *S. cerevisiae* to construct the final prediction model, so this model should

Table 2 The prediction results of cross-species dataset by the meta predictor and other six individual predictors

Methods	Test dataset			
	C. elegans (%)	*E. coli* (%)	*H. sapiens* (%)	*M. musculus* (%)
AC–SVM	77.65	80.64	82.79	86.58
CT–SVM	75.50	67.69	78.97	77.96
LD–SVM	75.73	71.24	76.27	76.68
GA–SVM	74.73	67.70	77.20	73.48
MA–SVM	76.55	75.86	80.10	84.03
NA–SVM	75.06	72.17	80.52	82.43
Meta	85.77	81.48	90.16	88.82

represent the features of *S. cerevisiae* PPIs. At the same time, our model can also represent the features of *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus,* which is implied by the generalization ability of this model on these four species.

Performance comparison with other prediction methods

In order to highlight the advantage of our model, it was also tested by *Helicobacter pylori* dataset. The *H. pylori* dataset is composed of 2,916 protein pairs (1,458 interacting pair and 1,458 non-interacting pairs) as described by Martin et al. (2005). This dataset gives a comparison of our method with other predictors, phylogenetic bootstrap (Bock and Gough 2003), signature products (Martin et al. 2005), HKNN (Nanni 2005), ensemble of HKNN (Nanni and Lumini 2006) and boosting (Shi et al. 2010). The results of ten fold cross-validation over six different methods are shown in Table S5. The average prediction performance, i.e. sensitivity, precision, accuracy, MCC and AUC achieved by meta predictor, are 84.02, 90.07, 87.93, 74.06 and 93.98%, respectively. It shows that the prediction results for meta predictor and the ensemble of HKNN, outperforms other state-of-the-art methods, which highlight that a multiple classifier system is more accurate and robust than a single classifier. We also observed that our method clearly achieves better results compared to other multiple classifier systems (i.e. ensemble of HKNN and Boosting), with an improvement of more than 5% in average precision. All these results show that the proposed meta predictor classifier not only achieves accurate performance, but also substantially improves precision in the prediction of PPIs.

Conclusions

In this work, a meta method built on six sequence-based individual classifiers was developed for the classification of PPIs. As expected, this method can obviously improve the prediction accuracy compared with the current state-of-the-art methods. In addition, a significant improvement in the prediction performance was also observed when performed on cross-species dataset. However, the meta predictor is not easy for biologist to interpret the prediction rules. In order to acquire the interpretable knowledge from experimental data, some methods such as decision tree (Lin et al. 2004) and frequent pattern tree (Wang et al. 2009) have been applied to derive the rule-based knowledge. In future, we will consider how to add these rule-based methods as the component predictor so as to uncover the rules for PPIs. Overall, the proposed meta method is expected to be a powerful tool for expediting the study of protein interaction networks.

Acknowledgments This work is supported by the grants of the National Science Foundation of China (30700161, 60905023, 30900321 and 60975005), the National Basic Research Program of China (2007CB311002), the Young Talent Grant of Hefei Institutes of Physical Science (0823A16121), the National High Technology Research and Development Program of China (2006AA02Z309), Innovation Program of Shanghai Municipal Education Commission (10YZ01), and Shanghai Rising-Star Program (10QA1402700).

References

Ben-Hur A, Noble WS (2005) Kernel methods for the predicting protein–protein interactions. Bioinformatics 21:i38–i46

Bock JR, Gough DA (2003) Whole-proteome interaction mining. Bioinformatics 19:125–134

Bujnicki J, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. Bioinformatics 8:750–751

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

Chen X, Liu M (2005) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21:4394–4400

Chou KC, Cai YD (2006) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734

Chou KC, Shen HB (2007b) Using ensemble of classifier to identify membrane protein types. Amino Acids 32:483–488

Enright A, Iliopoulos I, Kyrpides N, Ouzounis C (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90

Feng Z, Zhang C (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19:269–275

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030

Ishida T, Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344

Jansen R et al (2003) A Bayesian Networks approach for predicting protein–protein interactions from genomic data. Science 302:449–453

Lin N et al (2004) Information assessment on predicting protein–protein interactions. BMC Bioinformatics 5:154

Liu KH, Xia JF, Li XL (2008) Efficient ensemble schemes for protein secondary structure prediction. Protein Pept Lett 15(5):488–493

Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins Struct Funct Genet 49:350

Martin S, Roe D, Faulon J (2005) Predicting protein–protein interactions using signature products. Bioinformatics 21:218–226

Nanni L (2005) Hyperplanes for predicting protein–protein interactions. Neurocomputing 69:257–263

Nanni L, Lumini A (2006) An ensemble of k-local hyperplane for predicting protein–protein interactions. Bioinformatics 22:1207–1210

Pitre S, Alamgir M, Green J, Dumontier M, Dehne F, Golshani A (2008) Computational methods for predicting protein–protein interactions. Adv Biochem Eng Biotechnol 110:247–267

Saini H, Fischer D (2005) Meta-DP: domain prediction meta-server. Bioinformatics 21:2917–2920

Shen HB, Chou KC (2007) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240

Shen JW et al (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 104:4337–4341

Shi MG, Xia JF, Li XL, Huang DS (2010) Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids 38(3):891–899

Sokal R, Thomson B (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 129:121–131

Vapnik V (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999

Wang J, Li C, Wang E, Wang X (2009) Uncovering the rules for protein–protein interactions from yeast genomic data. Proc Natl Acad Sci USA 106:3752–3757

Xia JF, Wang B, Huang DS (2007) Inferring strengths of protein–protein interaction using artificial neural network. In: Proceedings of the international joint conference on neural networks (IJCNN), Orlando, FL, pp 2471–2475

Xia JF, Han K, Huang DS (2010) Sequence-based prediction of protein–protein interactions by means of rotation forest and autocorrelation descriptor. Protein Pept Lett 17(1):137–145

Yang L, Xia JF, Gui J, and Huang DS (2010) Prediction of protein–protein interactions from protein sequence using local descriptors. Protein Pept Lett (in press)

Zhao XM, Wang R, Chen L, Aihara K (2008a) Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res 36(9):e48

Zhao XM, Wang Y, Chen L, Aihara K (2008b) Gene function prediction using labeled and unlabeled data. BMC Bioinformatics 9:57

Zhao XM, Chen L, Aihara K (2008c) Protein function prediction with high-throughput data. Amino Acids 35(3):517–530