Asian Journal of Information Technology 15 (1): 26-30, 2016
ISSN: 1682-3915

MPPCM: Combing Multiple Classifiers to Improve Protein-Protein Interaction Prediction

[1]A. Hepsiba and [2]R. Balasubramanian
[1]Department of MCA, Karpaga Vinayaga College of Engineering and Technology,
Mother Teresa Women's University, Madhuranthagam, Kodaikanal,
[2]Karpaga Vinayaga College of Engineering Technology, Madhuranthagam, Tamil Nadu, India

Abstract: Determining Protein-Protein Interaction (PPI) in biological systems is of significant importance and prediction of PPI has turn out to be a popular research area. Although, different classifiers have been developed for PPI prediction no single classifier seems to be intelligent to predict PPI with high confidence. Here, it is postulated that by combining individual classifiers the accuracy of PPI prediction could be surely improved. In this research, here developed a method called Modified Protein-Protein Interaction Prediction Classifiers Merger (MPPCM) and this method combines output from two PPI prediction tools, GO2PPI and Phyloprof, using Ada Boost algorithm. The performance of MPPCM was tested by Area Under the Curve (AUC) using an assembled gold standard database that contains both positive and negative PPI pairs. Our AUC test showed that MPPCM significantly improved the PPI prediction accuracy over the corresponding individual classifiers. We found that additional classifiers incorporated into MPPCM could lead to further improvement in the PPI prediction accuracy. Furthermore, cross species MPPCM could achieve competitive and even better prediction accuracy compared to the single species MPPCM. This study established a robust pipeline for PPI prediction by integrating multiple classifiers using Ada Boost algorithm. This pipeline will be useful for predicting PPI in nonmodel species.

Key words: Protein, nonmodel, data mining, accuracy, MPPCM

INTRODUCTION

Protein-Protein Interactions (PPIs) stand for the intentional physical contacts built between multiple proteins for proper biological activities (Gavin *et al*., 2002). Generally, PPIs play vital roles in diverse essential molecular processes including signal transduction, cell metabolism and muscle contraction (Devos and Russell, 2007). With the increasing research attention on PPIs, a number of approaches have been proposed to investigate how they interact (Gavin *et al*., 2006). In the existing literature, the most widely-adopted experimental technologies are Yeast two-Hybrid (Y2H) and Tandem Affinity Purification (TAP). However, the computational process of both of the aforementioned biological techniques is time consuming. In addition, the accuracy of these approaches is still not satisfying. To resolve these two issues simultaneously, efficient computational approaches are required for the effective analysis of PPIs (Franceschini *et al*., 2013; Von Mering *et al*., 2002; Planas *et al*., 2013; Sussman *et al*., 1998; Hart *et al*., 2006).

Thereafter, a number of computational approaches have been proposed and implemented to speed up the predictionprocess of PPIs (Gallone *et al*., 2011). Nevertheless, with the scale of protein sequences getting larger and larger, most of the existing computational approaches become invalid and unsuitable due to the following reasons. These methods are generally proposed to manage with various data types such as protein domain, genomic information and protein structure information and the prior information of protein pairs is needed to properly predict PPIs. However, the data complexity also increases when the data scale gets large such protein pairs are hard to directly obtain and thus, invalidate these computational approaches. Therefore, the protein sequence-based approaches are preferred as they directly derive the necessary information from the amino acid sequence. Recently, Hosur, etc. (Yu *et al*., 2010) proposed a threading-based approach to predict PPIs directly based on protein sequences. Moreover, Guilherme Valente, etc. (Garcia *et al*., 2012) named their method Universal In Silico Predictor of Protein-Protein

Corresponding Author: A. Hepsiba, Department of MCA, Karpaga Vinayaga College of Engineering and Technology,
Mother Teresa Women's University, Madhuranthagam, Kodaikanal, Tamil Nadu, India

Interactions (UNISPPI) which classified PPIs based on the original protein sequence information with a sustaining accuracy.

Protein-Protein Interaction (PPI) networks play important roles in many cellular activities including complex formation and metabolic pathways (Gavin *et al*., 2002) and identification of PPI pairsmay provide important insights into the molecular basis of cellular processes (Alberts, 1998). Several high-throughput experimental approaches have been developed for PPI identification including two-hybrid assays (Devos and Russell, 2007), tandem affinity purification followed by mass spectrometry (Gavin *et al*., 2006) and protein microarrays (Kumar and Snyder, 2002). These high-throughput methods have produced a large amount of PPI data which have been accumulated in the public PPI databases such as DIP (Xenarios *et al*., 2000) and STRING (Franceschini *et al*., 2013). Though, the results generated by these high-through put methods may lack reliability (VonMering *et al*., 2002) and have limited coverage of PPIs in any given organism (Iglesias *et al*., 2013). Further, experimental information for PPI is also available including the X-ray structures of protein complexes in the PDB databank (Sussman *et al*., 1998). Nevertheless, the information from protein structure complexes may belimited compared to the large volume of protein sequences available in the public databases (Hart *et al*., 2006).

To overcome the limitations in PPI identification using experimental methods, computational approaches have been developed to achieve large-scale PPI prediction in variousorganisms (Gallone *et al*., 2011; Yu *et al*., 2010; Garcia *et al*., 2010; Maetschke *et al*., 2012; Lin *et al*., 2014; Song *et al*., 2014). Out-of-date input features for PPI prediction are mainly from biological data sources which may be divided into four categories: Gene Ontology-(GO-) based, structure-based, network topology-based and sequence-based features (Lin *et al*., 2014). Each individual computational PPI prediction method utilizes only one or few inputsources for PPI prediction. For example, BIPS only takes protein sequences as input for Interolog searching (Smialowski *et al*., 2010). Bio: homology: interolog walk takes protein sequences and well-known PPI networks as input (Gallone *et al*., 2011). Although, thesemethods using single or several features as input can generate fairly accurate results, they are unable to take advantageof other input features that could be helpful for PPI prediction. Thus, machine learning methods (e.g., Bayesian classifiers (Herman *et al*., 2011), Artificial Neural Networks (ANN) (Simonsen *et al*., 2012), SupportVector Machines (SVM) (Zhang *et al*., 2012) and Ada Boost (Augusty and Izudheen, 2013) have been developed to integrate multiple featuresas inputs. Machine learning approaches have shown better per formances compared to some other methods; among them, Ada Boost method seems to show the best performance (Theofilatos *et al*., 2011). In addition, PPI prediction is associated with imbalanced data problem. Zhang *et al*. (2012) proved thatthe imbalanced data problem could be solved by ensemble methods. Augusty and Izudheen (Song *et al*., 2014) further showed that Ada Boost method could improve Zhang's methods indealing with the imbalanced data problem.

MATERIALS AND METHODS

Construction of a gold standard dataset: It is created training and test dataset containing direct interacted proteinpairs of yeast for Protein-Protein Interaction (PPI) prediction using a method. Briefly, 2865 positive PPI pairs were obtained from the DIP database (Xenarios, 2000). These direct interaction protein pairs were tested tobe highly confident PPI pairs by small-scale experiments. Meanwhile, there was insufficient high-confidence negative data, negative PPI pairs were generated by randomly pairing proteins followed by removing the positive PPI pairs. Finally, the positive PPI pairs and the negative PPI pairs were combined by a ratio of 1-100 into a "Gold Standard" dataset. It has been proved that the AUC value is not sensitive to the different positive-to-negative ratios.

Selection of features for PPI prediction: The results of PPI prediction classifiers were used as features of MPPCM. Specifically, Phyloprof has three kinds of input parameters including four PPI prediction methods, eight Reference Taxa Optimization methods and four PPI networks. Without the time consuming PPI prediction method "RUN," there were 102 different classifiers based on different combinations of parameters provided by Phyloprof. As mentioned above, GO2PPI has three kinds of input parameters as well, including two machine learning methods, seven GO terms or terms combinations (BP, CC, MF, BPCC, BPMF, CCMF and BPCCMF) and seven PPI networks. In the same way, there were 98 different combinations of classifiers provided by GO2PPI. It is used combined GO terms in this study because the best accuracy was achieved by the integration of three GO terms in the GO2PPI paper.

PPI prediction using mppcm pipeline: Specifically, a protein pair is first evaluated by classifier sprovided by PPI prediction software such as GO2PPI and Phyloprof. Then, the classification scores from individual classifiers

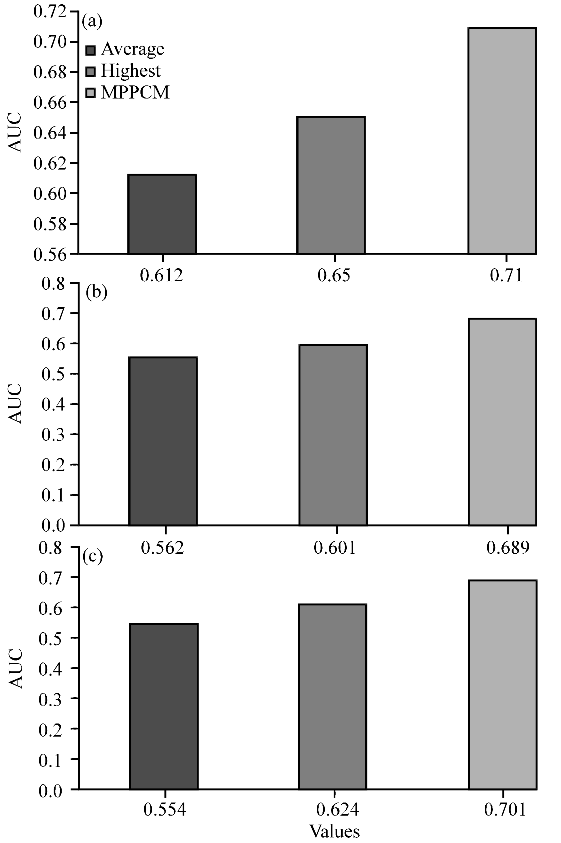

Fig. 1: a)PPI prediction based on classifiers related to SC; b): PPI prediction based on classifiers related to cross species and c) PPI prediction based on classifiers related to all species

are used as input features to produce the final PPI prediction score using Ada Boost algorithm, implemented in the MATLAB. GO2PPI has 98 PPI prediction classifiers, among which 14 are SC-related and 84 are not SC-related (cross species) classifiers. Phyloprof has 96 PPI prediction classifiers, among which 24 are SC-related and 72 are not SC-related (crossspecies) classifiers.

Evaluation of PPI prediction accuracy: The a forementioned gold standard database that contains about 30,000 PPI pairs with a positive-to-negative PPI ratio of 1:100 was used to evaluate the PPI prediction accuracy. The following measures were used to evaluate PPI prediction results: the true positive rate (TPR also called sensitivity), defined as the ratio of correctly predicted positive PPI pairs among all positive PPI pairs, the true negative rate (TNR also called specificity), defined as the ratio of correctly predicted negative PPI pairs among all negative PPI pairs and the False Positive Rate (FPR also called Type I error), defined as the ratio of incorrectly predicted PPI pairs among all negative PPI pairs. The FPR is one minus TNR. The Receiver Operating Characteristic (ROC) curves were created by plotting TPR versus FPR. The Area Under the Curve (AUC) was used asa measure of the prediction accuracy. The AUC value was calculated using the following equation:

$$AUC = \frac{1}{2}\sum_{k=1}^{n}\left(\left(x_k + x_{k-1}\right)\left(Y_k + Y_{k-1}\right)\right) \quad (1)$$

Where:

x_k = The FPR at k pair

y_k = The TPR at k pair in the ranked PPI pair list. The prediction process was repeated 25 times and the average AUC value was reported

In our research, evaluated the PPI prediction accuracy of MPPCMs and the classifiers in GO2PPI and Phylopr of using AUC. Here, it is introduced three categories of MPPCM including GO2PPI, Phylopr of and GO2PPI+Phylopr of with each further divided to three Sub-Categories (SC) cross species and allspecies (i.e., SC plus cross species) (Fig. 1a).

RESULTS AND DISCUSSION

Performance of MPPCM in GO2PPI category: Using the gold standard dataset, the average AUC of the 14 SC-related classifiers in GO2PPI (Table S1) was 0.61 and rf|bpcc|SC was the most accurate classifier with an AUC of 0.63, among these 14 classifiers (Fig. 2). The average AUC of the 84 cross species related classifiers in GO2PPI (Table S1) was 0.59 and rf|bpcc|HS was the most accurate classifier with an AUC of 0.60, among these 84 classifiers (Fig. 2b).

The average AUC of all the 102 (all species) classifiers in GO2PPI (Table S1) was 0.59 and rf|bpcc|SC was the most accurate classifier withan AUC of 0.64, among these 98 classifiers (Fig. 2). The AUCs of MPPCMs are 0.71, 0.70 and 0.69 for SC, cross species and all species MPPCM, respectively (Fig. 2). These results indicate that MPPCMs significantly improved PPI prediction accuracy compared with their corresponding classifiers in GO2PPI category compared with the most accurate classifier in GO2PPI category, the cross species MPPCM improves AUC by 12%.The improvement of MPPCM in SC MPPCM was only 9% (Fig. 2), indicating that the cross species MPPCM had better performance than the SC classifier. The better performance of cross species MPPCM

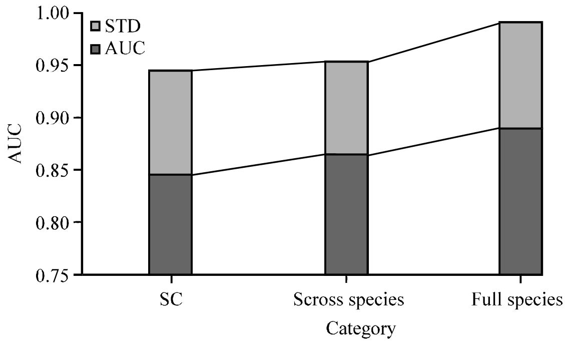

Fig. 2: Comparison of PPI prediction accuracy in the GO2PPI+Phyloprof category

(containing 84 features) than SC MPPCM (containing 14 features) suggests that the larger number offeatures incorporated into MPPCM enhanced PPI prediction accuracy in GO2PPI category.

Performance of MPPCM in the Phyloprof category: Again, using our gold standard dataset, the average AUC of the 24SC-related classifiers in Phyloprof (Table S2) was 0.64 and SC|mi|et was the most accurate classifier with an AUC of 0.71, among these 24 classifiers. The average AUC of the 72 cross species related classifiers in Phyloprof (Table S2) was 0.61 and EC|mi|et was the most accurate classifier with an AUC of 0.72, among these 84 classifiers (Fig. 3b). The average AUC of all the 96 (all species) classifiers in Phyloprof (Table S2) was 0.62 and mi|et |EC was the most accurate classifier with an AUC of 0.72, among these 96 classifiers. The AUCs of MPPCMs are 0.72,0.76 and 0.77 for SC, cross species and all species MPPCM, respectively (Fig. 1). These results indicate that MPPCM ssignificantly improved PPI prediction accuracy compared with their corresponding classifiers in the Phyloprof category. Compared with the most accurate classifier in the Phyloprof category, the cross species MPPCM improves AUC by 6% while the improvement by SC MPPCM is only 1%, indicating that the cross species MPPCM had better performance in AUC improvement. The better performance of cross species MPPCM (containing 72 features) than SC MPPCM (containing 24 features) suggests that more features incorporated into MPPCM could enhance PPI prediction accuracy in the Phyloprof category.

Performance of MPPCM in GO2PPI+Phyloprof category: After separate evaluation of MPPCM in the GO2PPI and Phyloprof categories, we assessed the performance of MPPCM in the GO2PPI+Phyloprof category which combined all the classifiers in both GO2PPI and Phyloprof. The AUCs of MPPCMs in the GO2PPI+Phyloprof category were 0.83, 0.85 and 0.86 for SC, cross species and all species MPPCM, respectively (Fig. 2) which are significantly higher than those of MPPCMs in either GO2PPI or Phyloprof category separately. Compared with the highest AUCs of individual classifiers in GO2PPI and Phyloprof category, the cross species MPPCM improves AUC by 18% and the improvement by SC MPPCM was 18%. These results indicate that MPPCM based on all the 194 classifiers from both GO2PPI and Phyloprof could generate more accurate PPI prediction than MPPCM based on a fewer number of classifiers in GO2PPI or Phyloprof individually, further supporting the aforementioned premise that more features incorporated into MPPCM would enhance PPI prediction accuracy. In summation, based on our combinatorial approach, our cross species MPPCM results yield informative predictions that will help build high-quality PPI networks for nonmodel organisms. Such prediction will be valuable for nonmodel organisms that lack biological data and PPI prediction software for nonmodel organisms (Hosur *et al*., 2010).

Recently, ensemble classifiers for example, LibD3C, were developed based on a clustering and dynamic selection strategy (Maetschke *et al*., 2012). In order to compare the performance of Random Forests method applied by our MPPCM with the latest ensemble classifiers, we performed ensemble classifiers calculationon our all species training and testing datasets of the GO2PPI+Phylopof category by LibD3C in Weka-3.7.12 with default setting. The average AUC by LibD3C was 0.86±0.03 which is in an excellent agreement withour Ada Boost result (0.86±0.02). Therefore, Ada Boost method applied by our MPPCM shows very similar performance with the latest ensemble classifiers (LibD3C).

CONCLUSION

In our research postulated that by combining individual classifiers the accuracy of PPI prediction could be improved. Here, it is developed a method called Modified Protein-Protein Interaction Prediction Classifiers Merger (MPPCM) and this method combines output from two PPI prediction tools, GO2PPI and Phyloprof, using Ada Boost algorithm. The performance of MPPCM was tested by Area Under the Curve (AUC) using an assembled gold standard database that contains both positive and negative PPI pairs. Our AUC test showed that MPPCM significantly improved the PPI prediction accuracy over the corresponding individual classifiers. It is found that additional classifiers incorporated into MPPCM could lead to further improvement in the PPI prediction accuracy. Furthermore, cross species MPPCM could achieve competitive and even better prediction accuracy compared to the single species MPPCM. This study established a robust pipeline for PPI prediction by integrating multiple classifiers using Ada Boost algorithm. This pipeline will be useful for predicting PPI in nonmodel species.

REFERENCES

Alberts, B., 1998. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell, 92: 291-294.

Devos, D. and R.B. Russell, 2007. A more complete, complexed and structured interactome. Curr. Opin. Struct. Biol., 17: 370-377.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn and M. Simonovic *et al.*, 2013. STRING 9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res., 41: D808-D815.

Gallone, G., T.I. Simpson, J.D. Armstrong and A.P. Jarman, 2011. Bio:: Homology:: InterologWalk a perl module to build putative protein-protein interaction networks through interolog mapping. BMC. Bioinf., 12: 289-289.

Garcia, J.G., E. Guney, R. Aragues, J.P. Iglesias and B. Oliva, 2010. Biana: A software framework for compiling biological interactions and analyzing networks. BMC. Bioinf., 11: 56-56.

Garcia, J.G., S. Schleker, J.K. Seetharaman and B. Oliva, 2012. BIPS: BIANA interolog prediction server, a tool for protein-protein interaction inference. Nucleic Acids Res., 40: W147-W151.

Gavin, A.C., M. Bosche, R. Krause, P. Grandi and M. Marzioch *et al.*, 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415: 141-147.

Gavin, A.C., P. Aloy, P. Grandi, R. Krause and M. Boesche *et al.*, 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature, 440: 631-636.

Hart, G.T., A.K. Ramani and E.M. Marcotte, 2006. How complete are current yeast and human protein-interaction networks. Genome Biol., 7: 120-128.

Herman, D., D. Ochoa, D. Juan, D. Lopez and A. Valencia *et al.*, 2011. Selection of organisms for the co-evolution-based study of protein interactions. BMC. Bioinf., 12: 363-363.

Hosur, R., J. Xu, J. Bienkowska and B. Berger, 2010. iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. J. Mol. Biol., 405: 1295-1310.

Iglesias, J.P., J. Bonet, J.G. Garcia, M.A.M. Lopez and E. Feliu *et al.*, 2013. Understanding protein-protein interactions using local structural features. J. Mol. Biol., 425: 1210-1224.

Kumar, A. and M. Snyder, 2002. Proteomics: Protein complexes take the bait. Nat., 415: 123-124.

Lin, C., W. Chen, C. Qiu, Y. Wu and S. Krishnan *et al.*, 2014. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing, 123: 424-435.

Maetschke, S.R., M. Simonsen, M.J. Davis and M.A. Ragan, 2012. Gene ontology-driven inference of protein-protein interactions using inducers. Bioinf., 28: 69-75.

Simonsen, M., S.R. Maetschke and M.A. Ragan, 2012. Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. Bioinf., 28: 851-857.

Smialowski, P., P. Pagel, P. Wong, B. Brauner and I. Dunger *et al.*, 2010. The negatome database: A reference set of non-interacting protein pairs. Nucleic Acids Res., 38: D540-D544.

Song, L., D. Li, X. Zeng, Y. Wu and L. Guo *et al.*, 2014. nDNA-prot: Identification of DNA-binding proteins based on unbalanced classification. BMC. Bioinf., 15: 298-298.

Sussman, J.L., D. Lin, J. Jiang, N.O. Manning and J. Prilusky *et al.*, 1998. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Crystallogr. Sect. D Biol. Crystallogr., 54: 1078-1084.

Theofilatos, K.A., C.M. Dimitrakopoulos, K.A. Tsakalidis, D.S. Likothanassis and T.S. Papadimitriou *et al.*, 2011. Computational approaches for the prediction of protein-protein interactions: A survey. Curr. Bioinf., 6: 398-414.

Von Mering, C., R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields and P. Bork, 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417: 399-403.

Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron and E.M. Marcotte *et al.*, 2000. DIP: The database of interacting proteins. Nucleic Acids Res., 28: 289-291.

Yu, C.Y., L.C. Chou and D.T.H. Chang, 2010. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. BMC. Bioinf., 11: 167-167.

Zhang, Y., D. Zhang, G. Mi, D. Ma and G. Li *et al.*, 2012. Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. Comput. Biol. Chem., 36: 36-41.