BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3386–3387
doi:10.1093/bioinformatics/btm434

Structural bioinformatics

meta-PPISP: a meta web server for protein-protein interaction site prediction

Sanbo Qin[1,2] and Huan-Xiang Zhou[1–3,*]

[1]Institute of Molecular Biophysics, [2]School of Computational Science and [3]Department of Physics, Florida State University, Tallahassee, Florida 32306, USA

Received on May 18, 2007; revised on August 13, 2007; accepted on August 18, 2007

Advance Access publication September 25, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Summary: A number of complementary methods have been developed for predicting protein-protein interaction sites. We sought to increase prediction robustness and accuracy by combining results from different predictors, and report here a meta web server, meta-PPISP, that is built on three individual web servers: cons-PPISP (http://pipe.scs.fsu.edu/ppisp.html), Promate (http://bioportal.weizmann.ac.il/promate), and PINUP (http://sparks.informatics.iupui.edu/PINUP/). A linear regression method, using the raw scores of the three servers as input, was trained on a set of 35 nonhomologous proteins. Cross validation showed that meta-PPISP outperforms all the three individual servers. At coverages identical to those of the individual methods, the accuracy of meta-PPISP is higher by 4.8 to 18.2 percentage points. Similar improvements in accuracy are also seen on CAPRI and other targets.

Availability: meta-PPISP can be accessed at http://pipe.scs.fsu.edu/meta-ppisp.html

Contact: zhou@sb.fsu.edu

Supplementary information: Data sets, linear regression coefficients, and details of prediction results are shown at the site of the meta-PPISP server.

1 INTRODUCTION

It is increasingly recognized that proteins function in the context of multi-component complexes. Interfaces formed in protein complexes carry important structural and functional information. After the publication of the first automated method for predicting residues in protein-protein interfaces in 2001 (Zhou and Shan, 2001), there has been intensive efforts at developing such methods (e.g. Chen and Zhou, 2005a; Fariselli *et al*., 2002; Liang *et al*., 2006; Neuvirth *et al*., 2004; Ofran and Rost, 2003; for a review, see Zhou and Qin, 2007). We are now presented the opportunity to combine different approaches for increasing prediction robustness and accuracy. Such meta-methods have been found to be very effective in structure predictions. We have found enhanced accuracy in predicting solvent accessibility by combining several methods (Chen and Zhou, 2005b). Here we report a metamethod, meta-PPISP, for predicting protein-protein interaction sites.

meta-PPISP is built on cons-PPISP (Chen and Zhou, 2005a), Promate (Neuvirth *et al*., 2004), and PINUP (Liang *et al*., 2006). These methods are chosen for two reasons. First, they are accessible through web servers (at http://pipe.scs.fsu.edu/ppisp.html, http://bioportal.weizmann.ac.il/promate, and http://sparks.informatics.iupui.edu/PINUP/, respectively). Second, they are based on very different approaches using sequence conservation but along with different other attributes as input, and hence may present synergy. cons-PPISP is a neural network predictor that uses sequence profiles and solvent accessibilities of spatially neighboring residues as input. Promate uses a composite probability calculated from properties such as secondary structure, atom distribution, amino-acid pairing, and sequence conservation. PINUP is based on an empirical energy function consisting of a side-chain energy term, a term proportional to solvent accessible area, and a term accounting for sequence conservation.

In meta-PPISIP, the three methods are combined in a linear regression analysis with the raw scores as input. The metamethod is found to consistently outperform the three individual methods.

2 METHODS

For each protein, interface predictions were first obtained from the three individual servers. The Promate and PINUP results have raw scores for each residue, ranging from 0 to 100; higher scores correspond to higher chances of being predicted as an interface residue. cons-PPISP gives consensus results from clustering predictions of a set of neural network models. In the original paper (Chen and Zhou, 2005a), 68 models were used. Here we used a reduced set of 17 models, involving training on small heterodimers. No scores were given in the original cons-PPISP method. Here we assigned scores based on the values of the interface-state output node. The consensus results of cons-PPISP were also taken into consideration: if a residue was predicted to be an interface residue by the consensus, the highest output value among the 17 models was taken as the score; otherwise the lowest among the 17 models was used. As with cons-PPISP, only surface residues (defined as those with at least 10% solvent accessibility) were considered for interface prediction. cons-PPISP scores ranged from 0 to 1; to be consistent, the original Promate and PINUP scores were scaled by 100. Promate and PIPUP also have their own clustering procedures for final predictions. These final predictions are not used for building meta-PPISP.

*To whom correspondence should be addressed.

However, the final predictions of the three individual methods are used below for benchmarking the performance of meta-PPISP.

For each surface residue, the predictor scores (s_{ij}) for itself and its 8 nearest spatial neighbors were used to define a linear function:

$$S = \sum_{i=1}^{3} \sum_{j=0}^{8} c_{ij} s_{ij} + c_0$$

where i refers to the three predictors, $j = 0$ is for the residue itself and increasing j refer to successively farther neighbors. The coefficients c_0 and c_{ij} were optimized on a set of 35 proteins (Enz35), consisting of enzymes and inhibitors in Docking Benchmark 2.0 (Mintseris *et al.*, 2005) filtered at 35% sequence identity. The optimization aimed to generate an S value close to 1 if the residue is known to be an interface residue and close to 0 otherwise. Only unbound structures were used for interface prediction. Optimization and assessment were made against the actual interface residues, defined as those with < 5 Å contacts across the interface in the bound complex. Performance reported on Enz35 was based on a 7-fold cross validation. For other predictions, such as those on CAPRI targets, all the 35 proteins were used for optimization. The optimized coefficients are listed at http://pipe.scs.fsu.edu/meta-ppisp-SI.pdf.

Performance was assessed by simultaneous considering two parameters: coverage and accuracy. Coverage is the fraction of the actual interface residues that are predicted as such. Accuracy is the fraction of correct interface predictions among all such predictions.

3 RESULTS

Figure 1 shows the performances of meta-PPISP and the three individual predictors on Enz35 and 25 CAPRI targets. The curve of coverage versus accuracy for each method was obtained by setting the threshold for positive predictions at different levels. Movement toward the upper right corner of the coverage versus accuracy plot signifies better performance. It can be seen that meta-PPISP outperforms all the three individual methods.

The final predictions of cons-PPISP, Promate, and PINUP, based on clustering, have coverages of 37.4, 17 and 45.2%, respectively, on Enz35; correspondingly the accuracies are 40.9, 59.2 and 50%. At the same coverages as the three individual methods, the accuracy of meta-PPISP is 59.1, 69.5 and 54.8, respectively. These accuracy levels are higher than those of the individual methods by 18.2, 10.2 and 4.8% points. On the CAPRI targets, the final predictions of the three individual methods have coverages of 39, 16.7 and 34.7%, respectively; the corresponding accuracies are 38.2, 47 and 41.2%. The uniformly worse predictions on the CAPRI targets are partly due to the fact that about one third of the targets are antibody-antigen and other immune system complexes. Nevertheless meta-PPISP is able to improve the predictions of the individual methods, increasing their accuracies by 3.4, 4 and 7.5% points, respectively. On a third set of targets, consisting of 32 enzymes and inhibitors more recently deposited in the Protein Data Bank, accuracy improvements by meta-PPISP fall within the range spanned by the Enz35 and CAPRI results.

As final predictions of meta-PPISP, we recommend a threshold of $S_{th} = 0.34$ for positive prediction. With this threshold, the numbers of predicted and actual interface residues are roughly equal, and consequently prediction coverage and accuracy are roughly equal. For example, with this threshold, the coverage and accuracy of meta-PPISP for Enz35 are 50.5 and 49.5%, respectively. The numbers of actual and predicted interface residues and true positives for the three sets of targets are listed at http://pipe.scs.fsu.edu/meta-ppisp-SI.pdf. An illustrative comparison between the three individual methods and meta-PPISP on one protein can also be found at http://pipe.scs.fsu.edu/meta-ppisp-SI.pdf). meta-PPISP is accessible at http://pipe.scs.fsu.edu/meta-ppisp.html. Interface predictions are returned to the user by e-mail. For each surface residue, the raw scores of the three individual methods and meta-PPISP are given. In addition, the user is provided a link to a PDB file in which the meta-PPISP scores are stored as B-factors. The user may raise or lower the positive-prediction threshold for different targets or for different purposes.

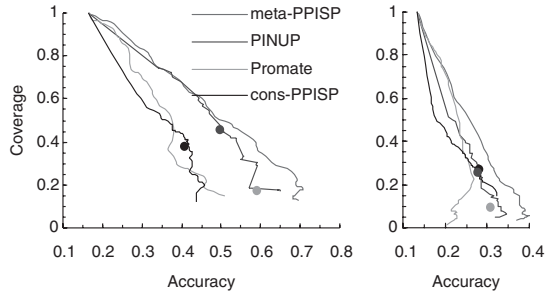

Fig. 1. Performance of meta-PPISP on Enz35 (left) and on CAPRI targets (right). Curves are drawn using raw scores; solid circles with matching colors show the final predictions, based on clustering, of the three individual methods. Promate distinguishes itself among the three individual methods in achieving a significant boost in accuracy by clustering.

In conclusion, we have shown that a metamethod built on three web servers achieves significant improvements in accuracy. Interface predictions have a broad range of applications, including assisting in protein docking (Tjong *et al.*, 2007; van Dijk *et al.*, 2005). The meta-PPISP web server is thus expected to have wide usage and motivate developments of other metamethods.

ACKNOWLEDGEMENTS

This work was support in part by NIH Grant GM058187.

Conflict of Interest: none declared.

REFERENCES

Chen,H. and Zhou,H.-X. (2005a) Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins*, **61**, 21–35.

Chen,H. and Zhou,H.-X. (2005b) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.*, **33**, 3193–3199.

Fariselli,P. *et al.* (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, **269**, 1356–1361.

Liang,S. *et al.* (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.*, **34**, 3698–3707.

Mintseris,J. *et al.* (2005) Protein-protein docking benchmark 2.0: an update. *Proteins*, **60**, 214–216.

Neuvirth,H. *et al.* (2004) ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Ofran,Y. and Rost,B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.*, **544**, 236–239.

Tjong,H. *et al.* (2007) PI^{2}PE: protein interface/interior prediction engine. *Nucleic Acids Res.*, **35**, W357–W362.

Dijk,A.D.J. *et al.* (2005) Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins*, **60**, 232–238.

Zhou,H.-X. and Qin,S.B. (2007) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209.

Zhou,H.-X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.