Unbiased Protein Interface Prediction Based on Ligand Diversity Quantification*

Reyhaneh Esmaielbeiki and Jean-Christophe Nebel

Faculty of Science, Engineering and Computing, Kingston University, London
Kingston-Upon-Thames, Surrey, KT1 2EE, UK
`{r.esmaielbeiki,J.Nebel}@kingston.ac.uk`

Abstract

Proteins interact with each other to perform essential functions in cells. Consequently, identification of their binding interfaces can provide key information for drug design. Here, we introduce Weighted Protein Interface Prediction (WePIP), an original framework which predicts protein interfaces from homologous complexes. WePIP takes advantage of a novel weighted score which is not only based on structural neighbours' information but, unlike current state-of-the-art methods, also takes into consideration the nature of their interaction partners. Experimental validation demonstrates that our weighted schema significantly improves prediction performance. In particular, we have established a major contribution to ligand diversity quantification. Moreover, application of our framework on a standard dataset shows WePIP performance compares favourably with other state of the art methods.

1998 ACM Subject Classification J.3 Life and Medical Sciences (Biology and genetics)

Keywords and phrases Protein-protein interaction, protein interface prediction, homology modeling

Digital Object Identifier 10.4230/OASIcs.GCB.2012.119

1 Introduction

Protein-protein interaction (PPIs) is essential for the functionality of living cells. Alterations of these interactions affect biochemical processes which may lead to critical diseases such as cancer [31]. Therefore, knowledge about protein interactions and their resulting 3D complexes can provide key information for drug design. A number of experimental techniques are available to identify residues involved in PPIs [31]. Although they provide valuable contribution to PPI knowledge, their cost in terms of time and expense limits their practical use [10]. Consequently, computational methods have been proposed to identify protein interfaces. They can be broadly divided in sequence only and structure based approaches.

Sequence based methodologies usually rely on a sliding window which allows calculating specific features associated to each amino acid according to its neighbours [24, 28, 35, 8]. Then, a classifier discriminates between interface and non-interface residues according to residue scores. Those approaches differ mainly in their selection of amino acid properties, such as physico-chemical properties [8, 7], residues distribution [24] or conservation degree [34, 25], machine learning algorithm and scoring functions [38].

When the 3D structure of the query protein (QP) is available, integration of structural information, e.g. residues secondary structure or solvent-accessible surface area, allows

* This work was in part supported by grant 6435/B/T02/2011/40 of the Polish National Centre for Science.

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 119–130
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

better predictions [25, 32]. Three approaches taking advantage of these properties are seen as state of the art [38]. ProMate combines 13 different properties, such as chemical component, geometric properties and information from relevant crystal structures, to generate a quantitative measure [23]. Protein interface residues are then predicted using a clustering process relying on mutual information. Alternatively, Cons-PPISP discriminates between residues using neural networks trained with protein's surface sequence profiles and solvent accessibility of neighbouring residues [6]. Finally, PINUP addresses the problem using an empirical energy function which is based on a linear combination of side chain energy score, interface propensity and residue conservation [22]. Despite fundamental differences, these three approaches display very similar performance [38]. However, each of them seems to capture different important aspects of residue interactions. As a result, a meta-predictor, Meta-PPISP, combining their scores using a linear regression analysis, is able to outperform each of these individual methods in terms of accuracy [27].

With the increasing number of experimentally determined protein 3D structures, they have become the main source of interface prediction methods. First, structurally homologous proteins tend to display similar interaction sites [1, 9]. Secondly, protein's binding sites are evolutionarily conserved among structurally similar proteins (or structural neighbours) [37, 18, 19, 4, 36, 5]. Even remote structural neighbours have been shown to display a significant level of interface conservation [37]. Consequently, structure based methods for interface prediction rely on analysing proteins which are structurally similar to the query protein.

Initial approaches focused on detecting conserved areas among homologous structural neighbours. Carl et al. use graph representation of surface residues [18, 19] to describe homologous binding sites [4]. They then refine their technique by using local structural similarity instead of global similarities of the query protein to detect the structural neighbours [5]. A more general method, PredUs, maps interacting residues from structural neighbours onto QP even if they do not display any homology [36]. Whereas PredUs still dependents on the existence of structural neighbours of QP, PrISE proposes to deal with this limitation by predicting interface residues from local structural similarity only [15]. This is achieved using a repository of structural elements (SE) generated from the Protein Data Bank (PDB) [3]. For each SE of the query protein (consisting of a central residues and it neighbours) a set of similar SEs are extracted from the repository and a weight is assigned to them based on their similarity to QP. The central residues of the query protein's SE are predicted as interface if a weighted majority of its similar SEs are interface residues. Although more general, PrISE displays comparable performance to PredUs [15].

Those two approaches have significantly improved the ability of predicting interface residues; see Table 3. However, they do not deal satisfactorily with the very heterogeneous nature of the PDB. First, the presence of complex duplicates, or homologs, biases predictions towards specific configurations, which can affect negatively performance. Secondly, confidence in the information provided by the interface of a structural neighbour should depend on its degree of homology with QP. Although PrISE acknowledges both issues, it does not address the first one [15]. PredUs deals with these matters in a binary fashion. Complexes involving structural neighbours are clustered and a 40% similarity cut-off is used to choose the representatives which will inform interface prediction. Here, we address those limitations of structure based methods by quantifying, first, homology between QP and its structural neighbours and,second, ligand diversity between the partners, or ligands, of the structural neighbours.

In this study we introduce Weighted Protein Interface Prediction (WePIP) framework, a novel PIP approach based on structural neighbours' information. Its main contribution

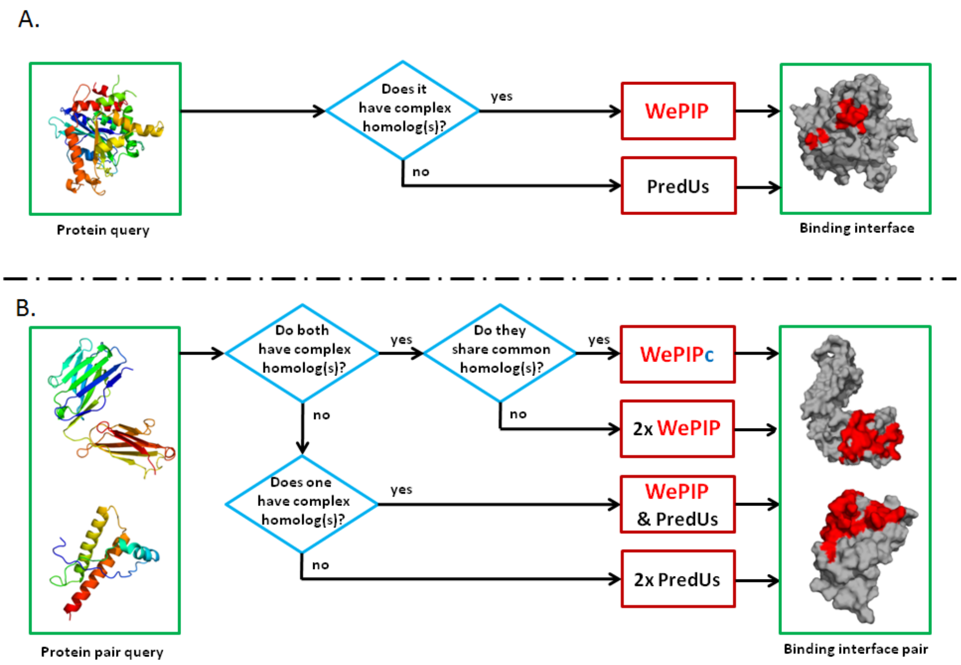


Figure 1 Interface prediction framework for single (A) and pair protein queries (B). A) For a single query, if at least one homologous complex exists, interfaces are predicted using WePIP otherwise PredUs is used. B) For a pair of proteins, depending on the existence of homologous complexes, interfaces are predicted by one or a combination of the WePIPc, WePIP and PredUs methods.

is the weighted score assigned to each residue of QP, which takes into account not only the degree of homology of structural neighbours, but also the nature of their interacting partners. After description of the methodology and validation of the proposed weighted schema, WePIP is evaluated against state of art protein interface prediction methods using a standard benchmark dataset.

2 Methods

2.1 Interface Prediction Principles

When two protein chains form a dimer, they bind through their interaction interfaces. We propose a novel methodology predicting the amino acids which are involved in binding interactions based on the 3D structures of the dimer partners. In this study dimer refers to any two protein chains involved in interaction. Not only does our approach predict the locations of interfaces when both binding partners are known, but it also infers the most likely binding interface of a single protein. Figure 1.A and 1.B describe those interface prediction pipelines for single and pair protein queries respectively.

Our methodology relies on discovering interaction patterns from the analysis of the 3D structures of complexes involving homologs of the dimer partners, called ‘homologous complexes’. In this work, proteins are defined as homologous if their sequence similarity is expressed by an E-value $\leq 10^{-2}$. First, Blast [2] is used to classify each query partner according to the availability of homologous complexes in PDB [3]. When both interaction partners are known and common homologous complexes exist, the WePIPc method exploits them as templates to predict both interfaces jointly (see Section 2.2). If both partners have homologous complexes, but none of them is common, or, if one deals with only one partner and this partner has homologous complexes, then interfaces are predicted independently from the interaction partner by the WePIP method (see Section 2.3). Finally, when no homologous complex is available, interface prediction is outside the scope of the WePIP/ WePIPc suite. Therefore, a third party PIP software, such as PredUs [36], is required. In this work, we use PredUs when homologous complexes are not available, because not only it is one of the best performing methods, but also it has been implemented as a Web server which can be used free of charge.

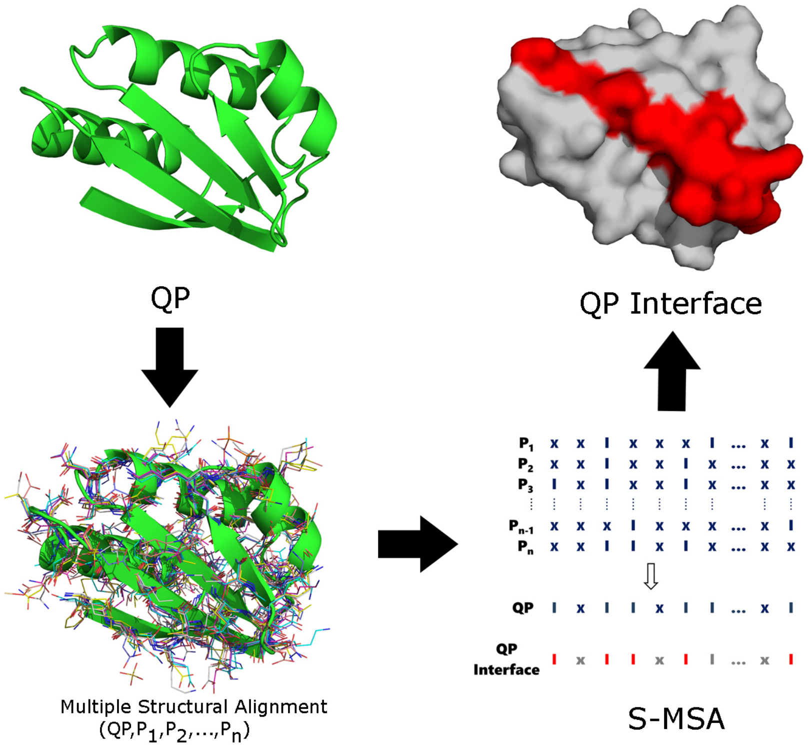
2.2 WePIPc

With rapid increase of experimentally determined structures, homology modelling of the whole 3D structure of a dimer is becoming more and more possible. It has been reported [11] that high quality homologous models could be found for 62% of the protein complexes present in the standard Protein Docking Benchmark 4.0 [13]. Consequently, template-based docking methods have been proposed based on common dimer complexes, i.e. dimers where each chain is homologous to a sequence of the query dimer [11, 21, 20]. For example, 66% of complexes generated by HOMBACOP were categorised as either acceptable or medium-quality models according to CAPRI assessment criteria [21]. Since these approaches have proved particularly accurate, we have included in our PIP framework a module, WePIPc, which infers interfaces based on common complex homologs.

Homologous complexes of each sequence of the protein query pairs are extracted from the PDB using Blast. Common homologous complexes are then selected and ranked by multiplying the E-values associated with the sequential alignments of each query chain with the homologous chain of the common complex. The common complex with the lower score is selected as the template from which the interfaces of the query chains are inferred. This is achieved by mapping the interface residues of the templates on the query chains according to their sequence alignments.

2.3 WePIP

WePIP relies on the observation that interface residues are usually structurally conserved between evolutionary related proteins [36]. Following extraction of homologous complexes from the PDB using Blast, the 3D structure of QP is structurally aligned with its homologs. In this study processing time is reduced by considering at most the 30 homologous complexes involving a chain whose E-value shows closest similarity to the QP. Alignment of multiple protein structures is performed by Multiprot [30], since it is a popular tool [16, 12, 33, 26] that has already be used successfully in interface residue prediction [16]. Using this information, a structure-based multiple sequence alignment (S-MSA) is produced. Then, known interface residues of the homolog complexes are highlighted on the S-MSA, see Figure 2. In agreement with the CAPRI definition [14], an interface residue is defined as an amino acid whose heavy atoms are within $5\AA$ from those of a residue in a separate chain. Using this multiple alignment, an interaction score is calculated for each residue of the query protein (see Section 2.3.1). Finally, the expected number of interface residues, n_{IA}, is predicted from known interfaces (see Section 2.3.2). The n_{IA} residues with the highest scores are then returned as defining the interaction interface.



P1	x	x	I	x	x	x	I	...	x	I
P2	x	x	I	x	x	I	x	...	x	x
P3	I	x	I	x	x	I	x	...	x	x
Pn-1	x	x	x	I	x	x	x	...	x	I
Pn	x	x	I	I	x	I	x	...	x	x
QP	I	x	I	I	x	I	I	...	x	I
QP Interface	I	x	I	I	x	I	I	...	x	I

Figure 2 Application of the WePIP method on a query protein (green). First, it is structurally aligned with its homologous complexes. Then, an S-MSA is produced where X and I represent non-interface and interface residues, respectively. Finally, interaction residues (red) of QP are predicted according to interaction scores and the estimated size of the interface. Note that residues weights are not shown here.

2.3.1 Interaction Score

In order to identify residues likely to be involved in the dimer interaction, we propose a residue scoring function which relies on the S-MSA of QP and its complexed homologs. In principle, any QP amino acid aligned with a residue involved in a dimer interaction is a potential candidate. However, confidence in the association of interaction activity to a residue depends on three factors: the degree of homology between the QP sequence and that of the protein from which the interaction is inferred, the nature of the ligand involved in the interaction with the homologous protein and the number of homologous proteins suggesting interaction. First, the more homologous a complexed protein, k, is to QP, the more informative is that protein regarding which residues are likely to be involved in the dimer interaction. We express this information by the query weight, x_k, (1):

$$x_k = \begin{cases} 1 - 10^{-200}, & \text{if } E_k < 10^{-200} \\ 1 - E_k, & \text{if } 10^{-200} \leq E_k \leq 10^{-2} \\ 0, & \text{if } E_k > 10^{-2} \end{cases} \tag{1}$$

where E_k is the E-value of protein k against QP as estimated by Blast.

Secondly, since none of the homologous complexed proteins interacts with the query ligand, diversity of ligands has to be rewarded given that they increase generalisation of interaction patterns. A second weight conveys this requirement by penalising homologous proteins, whose ligands are similar to each others. This is estimated by the average distance between the sequence of a ligand and all the other as expressed by the arithmetic mean of the pair wise E-values. Given a homologous complex protein, k, interacting with a ligand, L_k, and the other $N-1$ homologous complexed proteins interacting with their respective ligand, L_j, the ligand weight, y_k, is formulated as (2):

$$y_k = \begin{cases} \frac{\sum_{j=1, j\neq k}^{N} E_{(L_k, L_j)}}{N-1}, & \text{if } N > 1 \\ 1, & \text{if } N = 1 \end{cases} \tag{2}$$

where $E_{(L_k,L_j)}$ is set to 1, if $E_{(L_k,L_j)} > 1$, and $E_{(L_k,L_j)}$ is set to 10^{-200}, if $E_{(L_k,L_j)} < 10^{-200}$.

The y_k score is designed so that the presence of complex duplicates does not bias predictions towards their configuration. For example, if a QP has 3 complex homologs, A, B and C where L_A is unrelated to L_B and L_C, but L_B and L_C are identical, $E_{(L_A,L_B)} = 1$, $E_{(L_A,L_C)} = 1$ and $E_{(L_B,L_C)} = 0$. Therefore, $y_A = y_B + y_C$, i.e. interface configurations of A and B/C will have the same weight.

The weighted score of the residue i of protein, k, is expressed by the product of these two weights (3):

$$w_{ik} = \begin{cases} x_k y_k, & \text{if } i \text{ interacts with } C_k \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Finally, since it was shown that usage of non-interface information improves prediction performance [36, 15], the score for residues i of QP is calculated in (4) as the sum of the weights of the interface residues in the homologs over all the interface and non-interface residues which are 3D aligned with i:

$$S_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sum_{j=1}^{N} x_j y_j} \tag{4}$$

Note that for non-interface residues the ligand which is geometrically the closest is used to calculate their weight y_k.

2.3.2 Estimation of the number of interface residues

After calculating S_i for all the amino acids of QP, it is necessary to estimate the expected number of interface residues of its interface, n_{IA}. Studies have shown that despite variability in ligand structures, the binding location between homologous structures and their ligands is conserved [17]. This suggests that the number of interface residues between homologs should remain quite stable even when the binding partners vary. Therefore, WePIP uses the weighted average number of interacting amino acids of all QP's homologs (n_{IA}) to estimate the number of interface residues of QP (5):

$$n_{IA} = \sum_{i=1}^{R} S_i \tag{5}$$

where R is the number of residues in the QP sequence plus the number of gaps added to allow alignment with its homologs. Finally, once n_{IA} has been calculated, the predicted interface is defined as the n_{IA} residues with the highest scores.

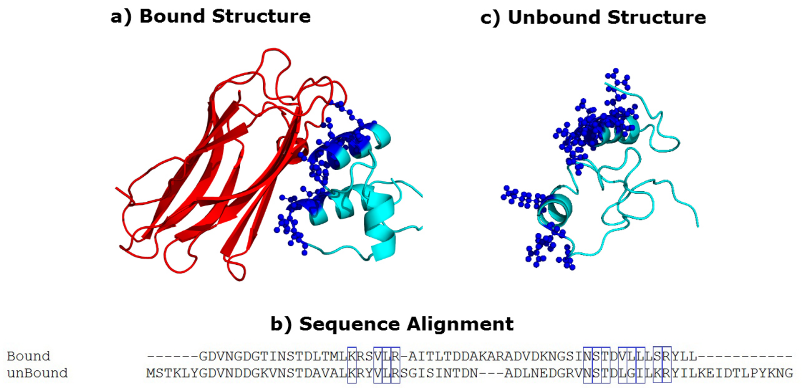

a) Bound Structure
c) Unbound Structure
b) Sequence Alignment
Bound ------GDVNGDGTINSTDLTMLKRSVLR-AITLTDDAKARADVDKNGSINSTDVLLLSRYLL-----------
unBound MSTKLYGDVNDDGKVNSTDAVALKRYVLRSGISINTDN---ADLNEDGRVNSTDLGILKRYILKEIDTLPYKNG

Figure 3 Generation of ground truth interface residues. a) Interfaces residues (blue spheres) are identified on the bound structure (cyan). The interaction partner is in red. b) The unbound and bound sequences are aligned to infer interfaces of the unbound structure. Mapping is shown by blue rectangles. c) Inferred interfaces are shown as blue spheres on the unbound structure (cyan).

3 Experimental results

3.1 Dataset and ground truth

Our interface prediction framework has been evaluated on a standard benchmark dataset, Ds56unbound, to allow comparison of its performance with state of the art methods [37, 15, 38]. The dataset is comprised of 56 unbound chains generated from 27 targets, T01∼T27, investigated during the communitywide experiment CAPRI [14]. The corresponding bound structures (Ds56bound) are used as ground truth. In total, DS56unbound contains 12173 residues including 2112 interacting ones. According to CAPRI's definition, interface residues are defined as amino acids on separate chains which have at least one heavy atom within a cut-off threshold of $5\AA$.

Since interface residues are not explicitly provided in DS56unbound, they were generated from the interface residues in their bound form. The process is illustrated in Figure 3. First, interface residues are detected on the bound complexes. Then, the unbound sequences are aligned with the bound sequences. Finally, the interfaces are mapped from the bound sequences onto the unbound ones.

3.2 WePIP Performance

All chains from Ds56unbound were processed by our interface prediction framework. Following initial homolog search where the Ds56bound complexes were excluded from the Blast results, homologous complexes were returned for 51 chains. Among them, 27 chains (Ds27unbound) displayed common complex(es) with their interacting partner and were further processed by WePIPc. Interfaces of the other 24 chains (Ds24unbound) were estimated by WePIP. Finally, the 5 chains (Ds5unbound) that could not be handled using a homology based approach were submitted to the PredUs server. Table 1 provides detailed performance of our system using standard measures, i.e. precision, recall, F-measure (F1), accuracy, Matthews correlation coefficient (MCC) and area under the receiver operating characteristic – ROC - curve (AUC). As expected, the more the method is able to exploit homology, the better is the interface prediction. Moreover, the table reveals that WePIP is quite conservative in its prediction: it

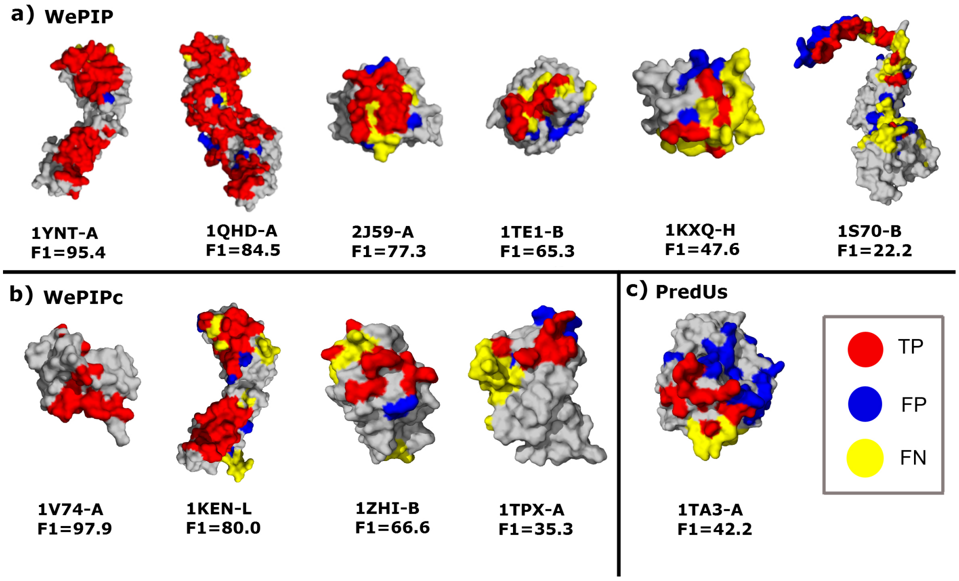


Figure 4 Interface predictions generated by the WePIP framework using either a) WePIP, b) WePIPc or c) PredUs. On each PDB target, true interface residues are coloured in red, whereas false positives and false negatives are shown in blue and yellow respectively. Corresponding F1 scores are also provided.

Table 1 Detailed performance of the WePIP framework. *DS**x**unbound: this means **x** chains out of the 56 unbound chains are solved by this specific predictor.

Predictor*	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**	**AUC**
WePIPc (DS27unbound)	68.8	60.5	63.2	87.0	56.0	77.1
WePIP (DS24unbound)	43.3	35.6	38.2	82.3	28.8	70.1
PredUs (DS5unbound)	23.6	45.5	30.1	75.8	19.4	63.2
WePIP + WePIPc (DS51unbound)	56.8	48.8	51.5	84.8	43.2	73.6
WePIP framework (DS56unbound)	**53.9**	**48.5**	**49.6**	**84.0**	**41.1**	**72.9**

displays relatively low recall value compared to precision. Figure 4 illustrates qualitatively those results by displaying representative predicted interfaces compared to ground truth.

In Figure 5, we provide the receiver operating characteristic curves of WePIP interface predictions for the six targets represented in Figure 4 a). First, curves are in agreements with model ranking based on F1 score. Second, accuracy regarding the number of estimated residues in the interface is highly correlated with the AUC. Finally, actual numbers of residues tend to be close to the curve's optimal cut-off point [29]. This point is located on the ROC curve where the distance is the largest to the random diagonal. This suggests there is scope for improving the estimation of expected interacting residues.

In order to validate the formulation of the weights used by WePIP, interface predictions for Ds24unbound were also estimated by setting those weights to 1. As shown in Table 2, results confirm the added value provided by our weighted schema. While usage of the query weight, x_k, only provides modest improvements when compared to a non weighted approach, the proposed ligand weight, y_k, offers a more significant increase of performance. Finally, when both weights are combined, most performance indicators improve further. These results

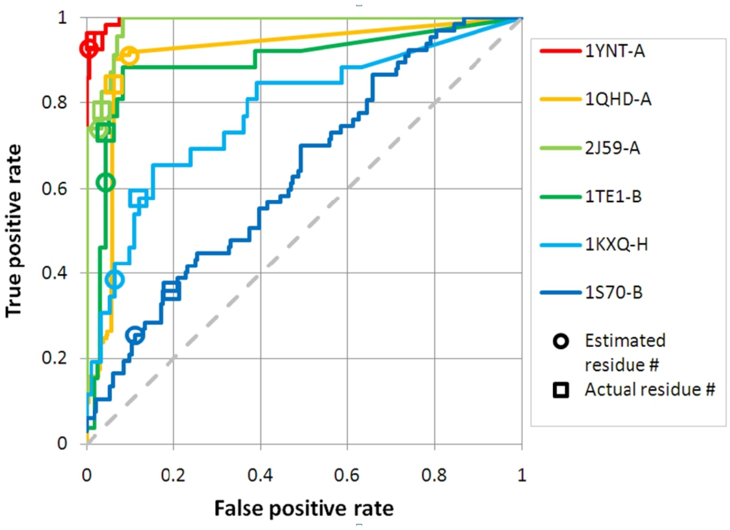


Figure 5 Receiver operating characteristic of WePIP interface predictions.

Table 2 Validation of weights used by WePIP (DS24unbound).

Query weight	Ligand weight	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**	**AUC**
1	1	40.6	32.5	37.0	81.6	25.5	70.3
x_k	1	40.8	32.7	38.7	82.4	26.2	70.4
1	y_k	42.4	34.6	37.3	**82.6**	28.1	**70.9**
x_k	y_k	**43.3**	**35.6**	**41.7**	82.3	**28.8**	70.1

confirm that taking into account both homology of QP and ligands leads to better interface predictions.

Finally, performance of our framework is compared with state of the art methods, see Table 3. Whatever measure is considered, the WePIP framework displays either best or second best performance competing with PrISE [15] and PredUS [36]. Moreover, the two aggregate measures, i.e. F1 and MCC, show that, globally, our system tends to produce better prediction than any other approach. In addition, since WePIP running time for a 300-residue protein with 30 homologous complexes is around 20 seconds on a standard PC, it can be used online. WePIP will be available as a web server in the near future.

Table 3 Comparison of WePIP framework with state of the art methods.

Predictor (DS56unbound)	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**	**AUC**
Promate [15]	28.7	27.3	28.0	76.6	14.0	62.7
PINUP [15]	30.4	30.1	30.2	76.9	16.4	60.0
Cons-PPISP [15]	37.4	34.5	35.9	79.5	23.8	71.2
Meta-PPISP [15]	38.9	24.0	29.7	81.1	20.2	71.5
PrISE [15]	43.7	44.0	43.8	81.2	32.6	**75.5**
PredUs [36]	43.3	**53.6**	47.9	73.2	30.4	72.9
WePIP framework	**53.9**	48.5	**49.6**	**84.0**	**41.1**	72.9

4 Conclusion

Although structure based methods perform best at predicting protein interfaces, they do not deal adequately with biases generated by the heterogeneous nature of the PDB. To address this issue, we have introduced the WePIP framework which associates to each putative interaction residue a confidence score taking into account both the degree of homology of structural neighbours and their ligands. Validation demonstrated that our novel weighted schema significantly improves prediction performance. In particular, we showed the major contribution of ligand diversity quantification. Moreover, application of our framework on a standard dataset shows WePIP performance compares favourably with other state of the art methods.

Despite the fact that our framework is state of the art, prediction of interface residues is still an unsolved problem: predictions remain unreliable for too many protein targets, even when homologous complexes are available. In future work, we intend to refine the proposed framework by integrating within our scoring schema the degree of spatial clustering of interface residues among structural neighbours.

Acknowledgments This work was in part supported by grant 6435/B/T02/2011/40 of the Polish National Centre for Science. The authors sincerely thank Rafael A Jordan at Iowa State University for providing the PrISE prediction results and his helpful assistance.

References

1 Patrick Aloy, Hugo Ceulemans, Alexander Stark, and Robert B Russell. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332(5):989–998, Oct 2003.

2 S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

3 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

4 Nejc Carl, Janez Konc, and Dusanka Janezic. Protein surface conservation in binding sites. *J Chem Inf Model*, 48(6):1279–1286, Jun 2008.

5 Nejc Carl, Janez Konc, Blaz Vehar, and Dusanka Janezic. Protein-protein binding site prediction by local structural alignment. *J Chem Inf Model*, 50(10):1906–1913, Oct 2010.

6 Huiling Chen and Huan-Xiang Zhou. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against nmr data. *Proteins*, 61(1):21–35, Oct 2005.

7 Peng Chen and Jinyan Li. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC Bioinformatics*, 11:402, 2010.

8 Xue-wen Chen and Jong Cheol Jeong. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, 25(5):585–591, Mar 2009.

9 Reyhaneh Esmaielbeiki, Declan P Naughton, and Jean-Christophe Nebel. Structure prediction of ldlr-hnp1 complex based on docking enhanced by ldlr binding 3d motif. *Protein Pept Lett*, 19(4):458–467, Apr 2012.

10 Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia, and Michael L Tress. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform*, 10(3):233–246, May 2009.

11 Anisah W Ghoorah, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and David W Ritchie. Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, 27(20):2820–2827, Oct 2011.

12 Inbal Halperin, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. *Structure*, 12(6):1027–1038, Jun 2004.

13 Howook Hwang, Thom Vreven, Joël Janin, and Zhiping Weng. Protein-protein docking benchmark version 4.0. *Proteins*, 78(15):3111–3114, Nov 2010.

14 Joël Janin and Shoshana Wodak. The third capri assessment meeting toronto, canada, april 20-21, 2007. *Structure*, 15(7):755–759, Jul 2007.

15 Rafael A Jordan, Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinformatics*, 13:41, 2012.

16 Ozlem Keskin, Buyong Ma, and Ruth Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–1294, Feb 2005.

17 Ozlem Keskin and Ruth Nussinov. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, 15(3):341–354, Mar 2007.

18 Janez Konc and Dusanka Janezic. Protein-protein binding-sites prediction by protein surface structure conservation. *J Chem Inf Model*, 47(3):940–944, 2007.

19 Janez Konc and Dusanka Janezic. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9):1160–1168, May 2010.

20 Petras J Kundrotas and Emil Alexov. Predicting 3d structures of transient protein-protein complexes by homology. *Biochim Biophys Acta*, 1764(9):1498–1511, Sep 2006.

21 Petras J Kundrotas, Marc F Lensink, and Emil Alexov. Homology-based modeling of 3d structures of protein-protein complexes using alignments of modified sequence profiles. *Int J Biol Macromol*, 43(2):198–208, Aug 2008.

22 Shide Liang, Chi Zhang, Song Liu, and Yaoqi Zhou. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*, 34(13):3698–3707, 2006.

23 Hani Neuvirth, Ran Raz, and Gideon Schreiber. Promate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 338(1):181–199, Apr 2004.

24 Yanay Ofran and Burkhard Rost. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, 544(1-3):236–239, Jun 2003.

25 Yanay Ofran and Burkhard Rost. Isis: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–e16, Jan 2007.

26 Utkan Ogmen, Ozlem Keskin, A. Selim Aytuna, Ruth Nussinov, and Attila Gursoy. Prism: protein interactions by structural matching. *Nucleic Acids Res*, 33(Web Server issue):W331–W336, Jul 2005.

27 Sanbo Qin and Huan-Xiang Zhou. meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23(24):3386–3387, Dec 2007.

28 I. Res, I. Mihalek, and O. Lichtarge. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10):2496–2501, May 2005.

29 Enrique F Schisterman, Neil J Perkins, Aiyi Liu, and Howard Bondell. Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. *Epidemiology*, 16(1):73–81, Jan 2005.

30 Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–156, Jul 2004.

31 Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42, Mar 2007.

32 Mile Sikić, Sanja Tomić, and Kristian Vlahovicek. Prediction of protein-protein interaction sites in sequences and 3d structures by random forests. *PLoS Comput Biol*, 5(1):e1000278, Jan 2009.

33 Christof Winter, Andreas Henschel, Wan Kyu Kim, and Michael Schroeder. Scoppi: a structural classification of protein-protein interfaces. *Nucleic Acids Res*, 34(Database issue):D310–D314, Jan 2006.

34 Li C Xue, Drena Dobbs, and Vasant Honavar. Homppi: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics*, 12:244, 2011.

35 Changhui Yan, Drena Dobbs, and Vasant Honavar. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20 Suppl 1:i371–i378, Aug 2004.

36 Qiangfeng Cliff Zhang, Lei Deng, Markus Fisher, Jihong Guan, Barry Honig, and Donald Petrey. Predus: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res*, 39(Web Server issue):W283–W287, Jul 2011.

37 Qiangfeng Cliff Zhang, Donald Petrey, Raquel Norel, and Barry H Honig. Protein interface conservation across structure space. *Proc Natl Acad Sci U S A*, 107(24):10896–10901, Jun 2010.

38 Huan-Xiang Zhou and Sanbo Qin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209, Sep 2007.