Research Article

Journal of Molecular Recognition

Received: 26 February 2014, Revised: 19 June 2014, Accepted: 27 June 2014, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jmr.2410

Prediction of protein–protein interaction sites from weakly homologous template structures using meta-threading and machine learning

Surabhi Maheshwari[a] and Michal Brylinski[a,b]*

The identification of protein–protein interactions is vital for understanding protein function, elucidating interaction mechanisms, and for practical applications in drug discovery. With the exponentially growing protein sequence data, fully automated computational methods that predict interactions between proteins are becoming essential components of system-level function inference. A thorough analysis of protein complex structures demonstrated that binding site locations as well as the interfacial geometry are highly conserved across evolutionarily related proteins. Because the conformational space of protein–protein interactions is highly covered by experimental structures, sensitive protein threading techniques can be used to identify suitable templates for the accurate prediction of interfacial residues. Toward this goal, we developed *e*FindSite^{PPI}, an algorithm that uses the three-dimensional structure of a target protein, evolutionarily remotely related templates and machine learning techniques to predict binding residues. Using crystal structures, the average sensitivity (specificity) of *e*FindSite^{PPI} in interfacial residue prediction is 0.46 (0.92). For weakly homologous protein models, these values only slightly decrease to 0.40–0.43 (0.91–0.92) demonstrating that *e*FindSite^{PPI} performs well not only using experimental data but also tolerates structural imperfections in computer-generated structures. In addition, *e*FindSite^{PPI} detects specific molecular interactions at the interface; for instance, it correctly predicts approximately one half of hydrogen bonds and aromatic interactions, as well as one third of salt bridges and hydrophobic contacts. Comparative benchmarks against several dimer datasets show that *e*FindSite^{PPI} outperforms other methods for protein-binding residue prediction. It also features a carefully tuned confidence estimation system, which is particularly useful in large-scale applications using raw genomic data. *e*FindSite^{PPI} is freely available to the academic community at http://www.brylinski.org/efindsiteppi.

Additional supporting information may be found in the online version of this article at the publisher's website.

Keywords: protein-binding site prediction; interfacial site prediction; meta-threading; machine learning; protein models; *e*Thread, *e*FindSite^{PPI}

INTRODUCTION

Proteins often function in conjugation with other proteins, thus an overwhelming number of biological processes are mediated by protein–protein interactions (PPIs) (Rual *et al*., 2005). For example, interacting proteins are routinely involved in signal transduction, protein transport and folding, DNA replication and repair, and cell division, just to mention a few examples. Consequently, significant efforts have been devoted to study PPIs because of their importance in elucidating protein function and molecular recognition processes. Also, PPi sites are attractive targets for therapeutics as the disruption of crucial interactions may attenuate or even impair the function of pharmacologically relevant proteins (Wells and McClendon, 2007; Jubb *et al*., 2012). In recent years, many experimental and theoretical studies have been conducted to discover and characterize these interactions; however, despite evident progress, salient challenges remain. Experimental methods used to identify interface residues are often low-throughput with associated high costs of instruments and experiments. Therefore, many cost-efficient computational approaches have been developed for the prediction of interaction sites to complement experimental efforts. For instance, computationally predicted PPI sites can be used to optimize site-directed mutagenesis experiments by reducing the number of mutations needed to be tested *in vitro* (Sowa *et al*., 2000; Sowa *et al*., 2001; Kortemme *et al*., 2004). Protein–protein docking is another important application of interfacial site prediction. Taking into account even the approximate location of protein interface can, in principle, reduce the search space, improve the accuracy of modeled complexes, and shorten computing time (Halperin *et al*., 2002; Chelliah *et al*., 2006; Li and Kihara, 2012). For instance, Li and Kihara showed that docking results obtained by a docking program PI-LZerD are successfully improved even when the accuracy of supplied PPI restraints is significantly low (Li and Kihara, 2012). On the other hand,

* *Correspondence to: Michal Brylinski, Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA.*
E-mail: michal@brylinski.org

a *S. Maheshwari, M. Brylinski*
Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

b *M. Brylinski*
Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, USA

another study by Shih and Hwang demonstrated that when using bioinformatics-predicted information on interface residues, data-guided protein docking methods perform poorly (Shih and Hwang, 2013), suggesting that PPI restraints should have a certain accuracy in order to improve protein docking.

Until now, a variety of computational methods have been developed for the prediction of PPI sites (Obenauer and Yaffe, 2004; Porollo and Meller, 2007; Pitre *et al*., 2008; Wang *et al*., 2013). Sequence-based methods largely rely on features extracted from sequence profiles constructed by Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (Koike and Takagi, 2004; Chen and Jeong, 2009; Murakami and Mizuguchi, 2010). Other methods extensively utilize remote evolutionary information to detect functionally important sites (Lichtarge *et al*., 1996; Armon *et al*., 2001; Pupko *et al*., 2002; Engelen *et al*., 2009). For example, the Evolutionary Trace algorithm (Lichtarge *et al*., 1996) maps conserved amino acids onto a three-dimensional protein structure and then identifies functional sites by analyzing highly conserved residues in the branches of an evolutionary tree. Identified residues are assumed to be structurally important if they lie in the core of a protein, while those on the surface are relevant for protein function. Finally, as a consequence of the continuously growing structural content in protein databases (Berman *et al*., 2013), a number of structure-based approaches have been developed. These algorithms exploit geometrical and physicochemical features derived from the three-dimensional structures of target proteins (Jones and Thornton, 1997; Liang *et al*., 2006; Jordan *et al*., 2012), for example, the solvent accessibility, secondary structure states, hydrophobicity, B-factors and the local topology. Furthermore, recent studies demonstrate that the interaction sites tend to be conserved among structural analogs (Zhang *et al*., 2010), which stimulate the development of methods for the prediction of PPI sites based on the global structural similarity between query proteins and those with known dimer structures. For example, a recently developed method called PrePPI derives empirical scores from the interfaces of structural neighbors for the prediction of binary PPIs (Zhang *et al*., 2012). The accuracy and coverage of approaches based on the global structural similarity certainly depend on the availability of experimental structures of target proteins as well as the oligomer complexes of their structural neighbors.

PPI sites can be separated from the rest of the surface by various geometric features, for example, accessible surface area, planarity and protrusion (Jones and Thornton, 1997; Nooren and Thornton, 2003), as well as the local structure similarity between query proteins and a repository of known dimers (Jordan *et al*., 2012). Consequently, there is an increasing interest in PPI prediction based on the local similarity; for instance, PrISE detects interaction sites using a local surface similarity between query proteins and a collection of structural elements (Jordan *et al*., 2012). Notwithstanding the evident progress in the structure-based identification of PPI sites in proteins, these methods have not been widely used in proteome-scale applications, primarily because (1) the number of proteins with known structures is far smaller than the number of known sequences; (2) they may require an additional knowledge of interacting partners, which is often unavailable; and (3) their performance depends on the availability of protein dimers structurally similar to query proteins.

In that regard, continuous efforts are directed toward the development of novel approaches for the prediction of protein–protein interfacial sites. In this study, we describe the development and benchmarking of *e*FindSite^{PPI}, a new evolution/structure-based method that can be used to predict PPI sites in proteins with known structures, as well as in gene products whose structures have not yet been solved experimentally. *e*FindSite^{PPI} effectively integrates sensitive meta-threading techniques with structure alignments and machine learning to accurately detect interfacial residues in query proteins. Its unique feature is the capability to predict positions and types of molecular interactions that target proteins are likely to form with their partners. These include many interactions known to stabilize protein–protein complexes, such as hydrogen bonds, salt bridges, as well as hydrophobic and aromatic contacts. Importantly, *e*FindSite^{PPI} makes accurate predictions for protein models with diverse quality, which opens up the possibility for structure-based PPI site identification at the proteome scale. Finally, in comprehensive benchmarks, we demonstrate that *e*FindSite^{PPI} outperforms other methods for the prediction of PPI sites from protein structures.

MATERIALS AND METHODS

Overview of *e*FindSite^{PPI}

*e*FindSite^{PPI} is a new evolution/structure-based approach for the prediction of protein-binding sites, specific interactions as well as the local interfacial geometry. The flowchart shown in Figure 1 illustrates the procedure implemented in *e*FindSite^{PPI}, which starts with the structure of a target protein (Figure 1A). Next, using meta-threading, functionally and structurally related templates are identified in the template library (Figure 1B). For each template, *e*FindSite^{PPI} retrieves its known complexes and maps their interfaces onto the target protein using structure alignments (Figure 1C). Then, the algorithm computes five different attributes for each surface residue in the target protein: the relative accessible area (RSA), generic interface propensity (IP), sequence entropy (SE), position specific interface propensity (PSIP), and the fraction of templates (FT) that have an equivalent residue at the protein–protein interface (Figure 1D). These attributes are combined into probabilistic scores by machine learning using Support Vector Machines (SVMs) and a Naïve Bayes Classifier (NBC) (Figure 1E). Both classifiers are finally used to distinguish between interface and non-interface residues in the target protein (Figure 1F). Below, we describe datasets used in this study, that is, the template library and various

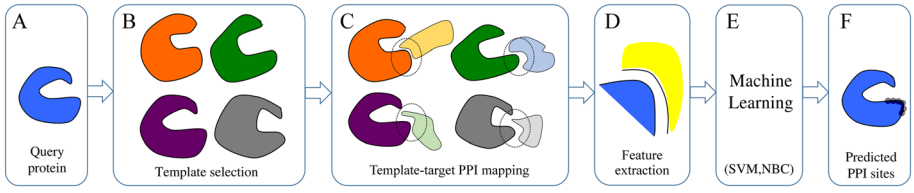

Figure 1. Flowchart for the PPI site prediction using *e*FindSite^{PPI}. Details are given in text.

benchmarking sets provide details on the methods and algorithms implemented in *e*FindSite^{PPI}, and explain evaluation metrics used to assess its performance in PPI prediction.

Dimer template library

Template library was compiled from all Protein Data Bank (PDB) (Berman *et al*., 2000) entries as of September 2012 with biologically relevant arrangements of two protein chains identified using Protein, Interface, Surfaces, and Assemblies (PISA) (Krissinel and Henrick, 2007). The redundancy was removed at 40% pairwise sequence identity by Cluster Database at High Identity with Tolerance (CD-HIT) (Li *et al*., 2001); however, two homologous dimers were included in the library if they either had structurally dissimilar receptor proteins with a template modeling score (TM-score) of <0.4 (Zhang and Skolnick, 2004), non-overlapping interfacial residues with Matthew's correlation coefficient (MCC) of <0.5, or a different interfacial geometry with an interfacial similarity score (IS-score) of <0.191 (Gao and Skolnick, 2010). Note that an IS-score of 0.191 indicates a significant interfacial similarity at a *p*-value of 0.05. TM-score is a structure alignment quality measure that ranges from 0 to 1 and has a length independent statistical significance threshold of ≥ 0.4, which corresponds to a *p*-value of 3.4×10^{-5} (Zhang and Skolnick, 2004). Here, TM-score is calculated upon structure alignments constructed by Fr-TM-align (Pandit and Skolnick, 2008), whereas the overlap of binding residues and the local structure similarity of binding interfaces (IS-score) are assessed by iAlign (Gao and Skolnick, 2010). The complete template library comprises 17,792 dimer structures.

Benchmarking dataset BM4361

The primary dataset used in *e*FindSite^{PPI} benchmarking, BM4361, consists of complex crystal structures selected from the template library. In each dimer, the longer chain is considered a receptor and the shorter chain is a ligand. We selected those dimers, in which the receptor has 50–600 residues. Furthermore, to avoid ambiguity when assessing the accuracy of interfacial residue prediction, we excluded receptors that interact with different ligands through different binding residues or whose close homologues with $\geq 40\%$ sequence identity form different PPIs. This procedure resulted in a non-redundant dataset of 4,361 protein dimers with unique and biologically relevant interfaces, referred to as BM4361. In addition to benchmarking simulations, this dataset was used to optimize *e*FindSite^{PPI} parameters and to construct machine learning models.

Benchmarking dataset BM1905

This dataset was compiled as a subset of BM4361 to benchmark the accuracy of binding residue prediction against non-native structures. It features three structural forms for each receptor protein: a crystal structure as well as high-quality and moderate-quality protein models. Weakly homologous models were generated by template-based modeling using *e*Thread (Brylinski and Feinstein, 2012; Brylinski and Lingam, 2012) following a procedure described in Supporting Information. *e*Thread is a meta-predictor that integrates several single threading algorithms to improve the recognition of structurally and functionally related templates (Brylinski, 2013). Both models with the preferred accuracy were constructed for 1,905 target proteins, thus the corresponding sets of crystal structures, high-quality, and moderate-quality models are referred to as BM1905C, BM1905H and BM1905M, respectively.

Other datasets

In addition to the BM4361 and BM1905 datasets, we compare the performance of *e*FindSite^{PPI} to other approaches for interfacial residue prediction on datasets used previously in the development and benchmarking of those algorithms. Comparison with PrISE is carried out using bound and unbound receptor conformations from the Benchmark 4.0 dataset (Hwang *et al*., 2010). We note that the accuracy of PrISE is assessed only against crystal structures in their bound conformational state (Jordan *et al*., 2012). We excluded multimeric complexes, in which the receptor is either smaller than 50 or larger than 600 residues, forms multiple interfaces, or the interface is made up of less than 20 residues. This dataset consists of 170 target proteins, 95 in bound and 75 in the unbound conformational state. We also assess the performance of *e*FindSite^{PPI} with respect to ET and iJET predictors (Lichtarge *et al*., 1996; Engelen *et al*., 2009) on the Huang dataset (Caffrey *et al*., 2004), applying similar criteria as described in the previous text. This dataset comprises 52 target proteins including 28 homodimers, 17 heterodimers and 7 transient complexes. When applicable, we modify *e*FindSite^{PPI} parameters to match prediction procedures described in the original publications of PrISE, ET and iJET.

Selection of dimer templates

*e*FindSite^{PPI} is a template-based approach, which employs meta-threading using *e*Thread (Brylinski and Feinstein, 2012; Brylinski and Lingam, 2012) to identify structurally and functionally related proteins in the template library as described previously (Brylinski, 2013). At least one dimer template is required in order to make a prediction. By default, we carry out benchmarking simulations excluding closely related templates, whose sequence identity to the target is $>40\%$. Moreover, we only use templates that structurally align to their targets with a TM-score of ≥ 0.4 (Zhang and Skolnick, 2004) as reported by Fr-TM-align (Pandit and Skolnick, 2008). Note that benchmarking calculations under these conditions are devised to approximate real applications in across-proteome functional annotation, where at most weakly homologous proteins can be identified for the majority of gene products. In addition to the default sequence identity threshold of 40%, we evaluate the performance of *e*FindSite^{PPI} at 30 and 20% as well.

Interfacial probability score

Each residue in the target protein is assigned an interfacial probability score that estimates the likelihood of this residue position to be at the protein–protein interface. These scores are calculated using machine learning and a set of the following residue-level attributes:

Relative surface accessibility

The relative accessible solvent area (ASA) of each residue is calculated using NACCESS (Hubbard and Thomson, 1993). This program implements a method by Lee and Richards (1971), which calculates the atomic accessible surface by rolling a probe of a given size around the van der Waals surface. Residues with a surface accessibility of $<5\%$ are considered buried,

thus non-interfacial. Remaining residues are assigned the relative surface accessibility (RSA) score.

Interface propensity

We use interface residue propensities derived for 20 standard amino acids by Jones and Thornton from a non-redundant set of high-resolution crystal structures of protein–protein complexes (Jones and Thornton, 1996; Jones and Thornton, 1997). Interface propensities (IP) describe the statistical likelihood of different amino acids to be found at protein–protein interfaces. These are calculated for each amino acid (AA_j) as the relative contribution of AA_j to the interfacial ASA compared with the whole surface:

$$IP_j = \frac{\sum_{i=1}^{N_i} ASA_i(j)}{\sum_{i=1}^{N_i} ASA_i} \Bigg/ \frac{\sum_{s=1}^{N_s} ASA_s(j)}{\sum_{s=1}^{Ns} ASA_s} \quad (1)$$

where, $\sum ASA_i(j)$ is the sum of ASA of amino acid residues of type j at the interface, $\sum ASA_i$ is the sum of ASA of all amino acids at the interface, $\sum ASA_s(j)$ is the sum of ASA of amino acid residues of type j on the surface, and $\sum ASA_s$ is the sum of ASA of all amino acids on the surface.

Sequence entropy

Functionally important residues tend to be evolutionarily conserved (Caffrey *et al*., 2004; Guharoy and Chakrabarti, 2005; Mintseris and Weng, 2005); therefore, we include a conservation score estimating the sequence variability for each target residue. First, multiple sequence alignments generated for the target sequence by PSI-BLAST (Altschul *et al*., 1997) are converted to a sequence profile. Then, the conservation score for each residue position (SE) is calculated using the Shannon entropy (Shanon, 1948):

$$SE = \sum_{i=1}^{20} p_i \log_2(p_i) \quad (2)$$

where p_i is the fraction of residues of amino acid type i in a given position in the sequence profile. *SE* ranges from 0 (absolute conservation of a particular residue type) to 4.32 bits (maximum entropy for equally distributed amino acids).

Position-specific interface propensity

The *PSIP* score combines generic interface residue propensities, as described in the previous text, with evolutionary information included in sequence profiles:

$$PSIP = \sum_{i=1}^{20} p_i IP_i \quad (3)$$

where p_i is the fraction of residues of amino acid type i at a given position in the profile and IP_i is the interface propensity for amino acid type i.

Fraction of templates

Finally, we include the fraction of templates (FT) that have an interfacial residue in the equivalent position according to template–target structure alignments constructed by Fr-TM-align.

Individual residue-level attributes, *RSA*, *IP*, *SE*, *PSIP* and *FT*, are combined into a single probabilistic score using machine learning. Two different classifiers, SVMs (Chang and Lin, 2011) and the NBC (Zhang, 2004), are trained to predict interfacial residues according to the assignment by iAlign (Gao and Skolnick, 2010). iAlign assigns interfacial residues based on interatomic contacts, which occur when any two heavy atoms belonging to residues from different chains are within a distance of 4.5 Å. Both machine learning models are twofold cross-validated on the BM4361 dataset. Specifically, dataset proteins are randomly divided into two subsets, *A* and *B*; *A* is used to train a model and then validate it against *B*, and vice versa, the model trained on *B* is validated against *A*. We note that <40% sequence identity between any pair of proteins in the BM4361 dataset ensures that the classifiers are trained and validated using different proteins. Probability thresholds optimized using the BM4361 dataset are 0.202 for the SVM and 0.178 for the NBC predictor. These values were selected to maximize MCC to 0.428, which corresponds to a true positive rate of 0.464 at the expense of 0.076 false positive rate. A given residue in the target protein is predicted to be at the interface when both probabilities are above their threshold values.

Calculation of interfacial interactions

In analyzing interfacial interactions, we consider the following four types of inter-residue contacts: salt bridges, hydrogen bonds, hydrophobic, and aromatic interactions. Salt bridges and hydrogen bonds across protein interfaces are detected by PDB2PQR (Dolinsky and Baker, 2004). Hydrophobic interactions are defined when the distance between any pair of atoms belonging to hydrophobic side chains is ≤5 Å; hydrophobic amino acids include Ala, Ile, Leu, Phe, Pro, Met and Val. Using the same distance threshold, aromatic contacts are identified between the side chains of His, Phe, Trp and Tyr. For each predicted interfacial residue, we calculate the fraction of templates that have a residue in the equivalent position forming a particular type of PPI using template–target structure alignments constructed by Fr-TM-align. These frequency values calculated for all interaction types correspond to the probabilities of various contacts that target residues may form with protein partners. Thresholds optimized on the BM4361 dataset are 0.001 for salt bridges, 0.021 for hydrogen bonds, 0.041 for hydrophobic contacts, and 0.012 for aromatic interactions. Similar to the interface residue prediction, these threshold values maximize the respective MCC.

Confidence estimation system

In proteome-level function inference, reliable predictions cannot be obtained for all targeted gene products; therefore, various predictors are required to provide confidence estimates. Every prediction by *e*FindSite^{PPI} is assigned an overall confidence score (CS) defined as

$$CS = \frac{1}{N} \sum_{i=1}^{N} SVM_i \times NBC_i \quad (4)$$

where N is the total number of predicted binding residues, and SVM_i and NBC_i are the binding probability scores assigned to

*i*th residue by machine learning using SVMs and the NBC, respectively. Calibrated ranges are CS ≥ 0.5 for high, 0.25 < CS < 0.5 for medium, and CS ≤ 0.25 for low confidence predictions.

Performance evaluation metrics

Binding residue prediction by *e*FindSite^{PPI} is assessed using standard evaluation metrics for classification problems:

Sensitivity (true positive rate):

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

Fall-out (false positive rate):

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

Specificity (true negative rate):

$$SPC = \frac{TN}{FP + TN} \tag{7}$$

Precision (positive predictive value):

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

Accuracy:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

Matthew's correlation coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + TN)(FP + FN)(TN + FN)}} \tag{10}$$

where true positives (*TP*), false negatives (*FN*) and false positives (*FP*) is the number of correctly predicted, underpredicted, and overpredicted binding residues, respectively. True negatives (*TN*) is the number of correctly predicted non-interfacial residues. Binding residues in experimental complex structures (positives) are defined as those forming protein–protein interfaces according to iAlign (Gao and Skolnick, 2010). The minimum value is 0, and the maximum value is 1 for all scores, except for MCC that ranges from −1 to 1. MCC quantifies the strength of the correlation between predicted and actual classes; by heavily penalizing both overpredictions and underpredictions, it provides a convenient assessment measure that balances the sensitivity and specificity. In addition to numerical values assessing the classification accuracy, we analyze the prediction results using receiver operating characteristic (ROC) plots. This technique was developed to evaluate the overall performance of a classifier and shows the trade-off between sensitivity and specificity. The area under the ROC curve (AUC) quantifies the performance of a classifier; larger AUC values indicate a better prediction power of the classification model.

The accuracy of interface residue prediction is compared with that of a random, size-independent classifier. First, for a given target protein, we estimate the size of its interface from the number of exposed residues as described by Martin (2014). Next, we randomly select a patch on the target surface whose size is equivalent to the estimated number of interfacial residues. This patch represents a random interface and includes the correction of a size bias, that is, smaller proteins have proportionally more residues within the patch, increasing the chances of overlapping with the correct interface.

RESULTS AND DISCUSSION

Accuracy of template selection

*e*FindSite^{PPI} employs meta-threading and structure alignments to select templates for the prediction of interfacial sites. The prediction accuracy inevitably depends on the quality of the identified set of dimer templates; therefore, using the BM4361 dataset, we first assess the accuracy of template selection. We note that templates used in this study are at most weakly homologous, sharing <40% sequence identity with their targets. Figure 2 shows a series of ROC plots cross-validating the accuracy of template selection with respect to several features. Using template confidence as a variable parameter, Figure 2A (a solid line) shows the performance of *e*Thread in detecting those templates that are structurally similar to the target with a TM-score of ≥0.4. Structure similarity is quantified by the TM-score (Zhang and Skolnick, 2004) calculated for template–target structure alignments constructed by Fr-TM-align (Pandit and Skolnick, 2008). Detecting structurally similar templates yields the maximum accuracy of 0.746 at a true positive rate of 0.642 and a false positive rate of 0.210, resulting in the area under ROC of 0.754.

Next, in addition to the global structure similarity, we also require a template to have a similar location of the PPI interface in order to be considered a positive. Specifically, we measure the interface overlap between the target and a template by calculating MCC over interfacial residues in both structures with residue equivalences taken from structure alignments. MCC values of ≥0.5 indicate that both the target and a template bind their partners at similar locations. Figure 2A (a dashed line) shows that protein templates whose binding interfaces are at similar locations are accurately detected. The corresponding area under ROC is 0.747 with the maximum accuracy of 0.759 obtained at a true positive rate of 0.655 and a false positive rate of 0.215. Finally, we consider the most stringent case, where the interfacial geometry in a template is similar to that in the target with an IS-score of ≥0.191. The IS-score measures interfacial similarity by comparing geometric distances as well as the conservation of contact patterns (Gao and Skolnick, 2010). Encouragingly, the area under ROC is 0.709, with the maximum accuracy of 0.695 at a true positive rate of 0.778 and a false positive rate of 0.419 (Figure 2A, a dotted line). Our results demonstrate that both the interface location and its geometry are conserved across a set of evolutionarily and structurally related proteins, which accords with previous studies (Gao and Skolnick, 2010; Zhang *et al*., 2010). Therefore, threading and meta-threading techniques can be effectively utilized to explore remote relationships between proteins using sensitive sequence profile comparisons. This strategy optimizes the selection of dimer templates for template-based prediction of functional aspects related to PPIs.

Similarity-based approaches to protein docking use dimer templates, in which both monomers are structurally similar to the target monomers (Aloy and Russell, 2003; Zhang *et al*.,

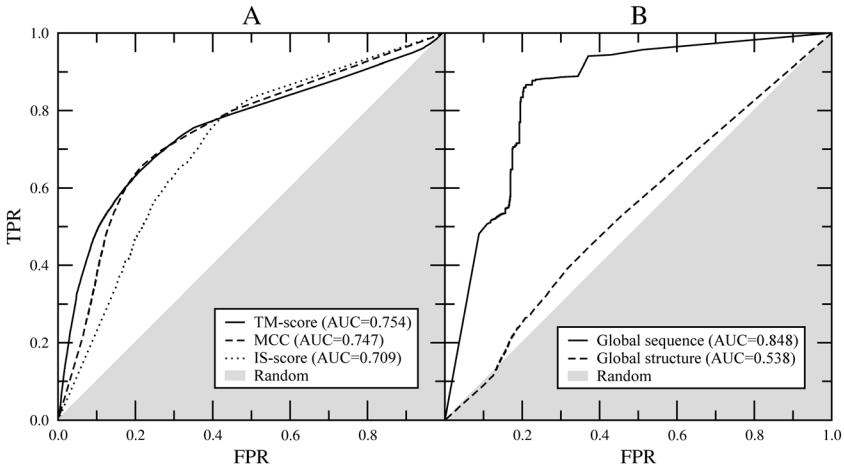

Figure 2. Accuracy of *e*Thread in recognizing templates for PPI site prediction. In (A), correct templates for the receptor (larger subunit) are defined using the global structure similarity with a TM-score of ≥0.4, the overlap of interfacial residues with MCC of ≥0.5, and the local interfacial similarity with an IS-score of ≥0.191. In (B), we evaluate the recognition of those dimer templates in which the ligand (smaller subunit) is globally similar to the target-bound ligand with a sequence identity of ≥40% and a TM-score of ≥0.4, respectively. Combined curves are calculated using a twofold cross-validation against the BM4361 dataset. TPR, true positive rate; FPR, false positive rate. Gray areas correspond to predictions no better than random.

2012). These algorithms employ global structure similarity to construct complex models based on the identified dimer templates. Therefore, we also analyze the capabilities of threading to detect weakly homologous receptor templates that bind globally similar ligands. First, we assess the global structure similarity of template ligands, where the interacting partners with a TM-score ≥0.4 to the target ligand are positives. Figure 2B (a dashed line) shows that binding ligands are not necessarily structurally similar to the target ligand even when they share the same binding location. The corresponding area under ROC is only 0.538, and the maximum accuracy of 0.483 is obtained at a true positive rate of 0.448 and a false positive rate of 0.373. Next, we use global sequence similarity to select interacting partners from the identified dimer templates; here, template ligands whose sequence identity to the target ligand is ≥40% are positives. Interestingly, as shown in Figure 2B (a solid line), receptor templates with similar binding sites tend to bind homologous proteins with respect to the target ligand. The area under ROC is 0.848, and the maximum accuracy of 0.790 is obtained at a true positive rate of 0.866 and a false positive rate of 0.210. We note that structurally similar ligands with a TM-score of ≥0.4 and homologous ligands with a sequence identity of ≥40% were found for 44 and 0.5% of the cases, respectively. This analysis shows that the interface site can be inferred using the global structure similarity when the sequence similarity between the target and template ligands is high. Nevertheless, because of the incompleteness of dimer libraries, the coverage of suitable protein targets is rather low.

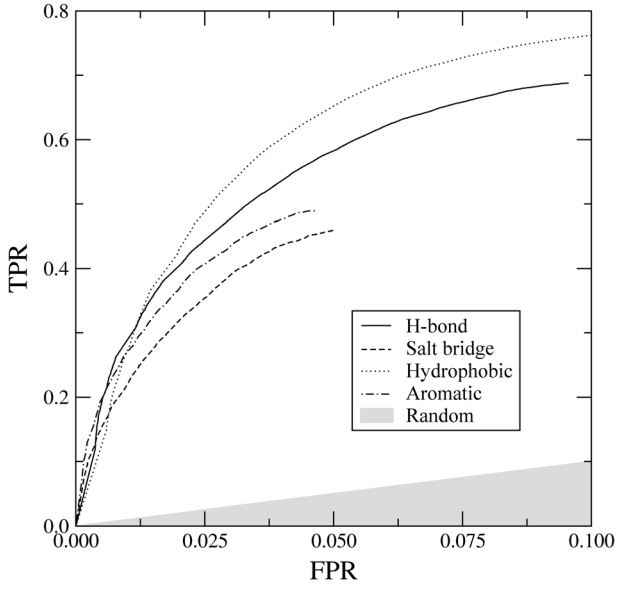

Figure 3. ROC plot evaluating the conservation of different types of protein–protein interactions across sets of evolutionarily weakly related dimer templates. The following non-covalent interaction types are considered: hydrogen bonds, salt bridges, hydrophobic, and aromatic contacts. A variable parameter is the fraction of templates that form the same interactions as the target in structurally equivalent positions. TPR, true positive rate; FPR, false positive rate. Gray area corresponds to interactions found by a random chance.

Conservation of interfacial interactions

Because protein complexes are stabilized by a variety of interactions, we analyze the conservation of interaction patterns across weakly related proteins. For each protein in the BM4361 dataset, interfacial interactions in its dimer templates are mapped to the target residues according to the structure alignments of receptor proteins. ROC plots in Figure 3 show the structural conservation of interfacial hydrogen bonds, salt bridges, aromatic and hydrophobic contacts at protein–protein interfaces. ROC curves end at certain sensitivity values, because we can only take account of those surface residues having an interacting residue at a structurally aligned position in at least one template. The maximum accuracy obtained for hydrogen bonds, salt bridges, hydrophobic and aromatic interactions is 0.900, 0.945, 0.895 and 0.949, at a true (false) positive rate of 0.684 (0.091), 0.459 (0.049), 0.760 (0.098) and 0.488 (0.044), respectively. Comparison of these ROC plots shows that the conservation of interfacial hydrophobic contacts and hydrogen bonds is higher than aromatic interactions and salt bridges. The high conservation of hydrophobic contacts is in line with previous studies suggesting

Journal of Molecular Recognition

that these interactions play a central role in stabilizing protein–protein complexes and the PPIs are dominated by hydrophobic patches (Jones and Thornton, 1996; Jones and Thornton, 1997). Overall, the results suggest that, in addition to binding residues, the interaction conservation patterns detected across structurally and evolutionarily related proteins can be used to predict various interaction types as well. These features can be used to support protein–protein docking simulations by favoring those assembled dimer conformation, in which highly conserved interactions are formed.

Prediction of PPI sites using experimental structures

*e*FindSite^{PPI} extracts PPIs from weakly homologous dimer templates identified by meta-threading for the prediction of protein-binding residues, specific interactions as well as the local interfacial geometry. Most of these features are identified by machine learning techniques. Here, we assess the accuracy of binding residue prediction, that is, the classification of target residues as either interfacial or non-interfacial, using two machine learning algorithms, SVMs and the NBC. As shown in Figure 4, the performance of both classifiers on the BM4361 dataset is fairly comparable. The area under ROC for SVM is 0.737, with the maximum MCC of 0.404 at a true (false) positive rate of 0.573 (0.144). For NBC, the area under ROC is 0.773, with the maximum MCC of 0.339 at a true (false) positive rate of 0.628 (0.209). Encouragingly, combining both classifiers using optimized thresholds, labeled as SVM + NBC in Figure 4, further enhances the discriminatory power. Specifically, MCC improves to 0.428, which corresponds to a sensitivity of 0.464 at the expense of only 0.076 false positive rate.

We also evaluate the performance of *e*FindSite^{PPI} in predicting specific interactions that the target protein is likely to form with its partners. The performance of *e*FindSite^{PPI} in the prediction of interaction types across the BM4361 dataset is shown in Figure 5; note that underpredicted interfacial residues count as false negatives in this analysis. Interestingly, despite the fact that closely homologous templates with a sequence identity of >40% were excluded from benchmarking calculations, the prediction of all interaction types is fairly accurate. True positive rates for hydrogen bonds and aromatic interactions are 0.515 and 0.484, with very small false positive rates of 0.048 and 0.037, respectively. For salt bridges and hydrophobic contacts, the true (false) positive rates are 0.330 (0.031) and 0.306 (0.017). These results demonstrate that eFindSite^{PPI} predicts approximately one half of interfacial hydrogen bonds and aromatic interactions and one third of salt bridges and hydrophobic contacts.

Size and composition of predicted interfaces

In addition to binding residues and interaction types predicted by *e*FindSite^{PPI}, in Figure 6, we analyze the general properties of interfacial sites, such as their size and amino acid composition. Figure 6A shows that the size of interfacial sites predicted by *e*FindSite^{PPI} for the BM4361 dataset correlates well with the size of experimental interfaces identified by iAlign (Gao and Skolnick, 2010); the Pearson correlation coefficient is 0.720 with a standard error of 0.118. In Figure 6B, we compare the amino acid composition of experimental and predicted protein–protein interfaces. The frequencies of amino acids at the predicted interfaces are in good quantitative agreement with the experimental data; the differences are less than 1% on average. Consequently, interfaces predicted by *e*FindSite^{PPI} are predominantly hydrophobic, which is consistent with a previous study conducted by Lijnzaad and Argos showing that interfacial sites often contain the largest or second largest hydrophobic patches on the surface of proteins (Lijnzaad and Argos, 1997). Next, we evaluate the composition of amino acids involved in specific interactions at

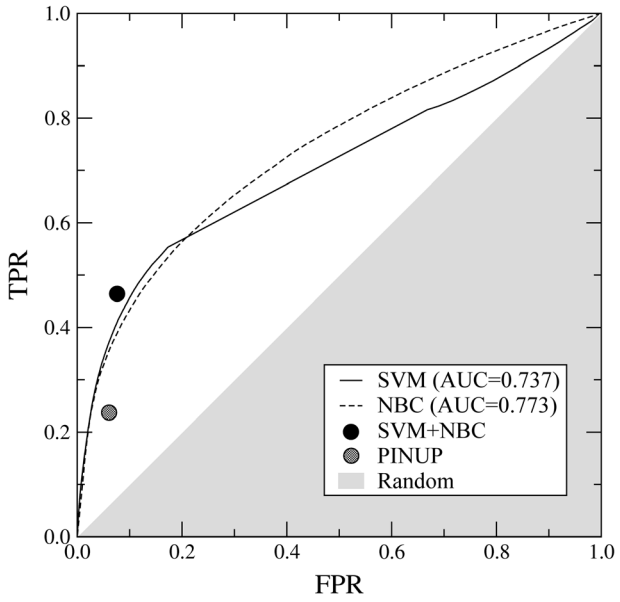

Figure 4. ROC plot assessing the accuracy of interfacial residue prediction across the BM4361 dataset by *e*FindSite^{PPI} compared with PINUP. For *e*FindSite^{PPI}, three prediction protocols are evaluated: SVM only, NBC only and a combination of SVM and NBC. TPR, true positive rate; FPR, false positive rate. Gray area corresponds to predictions no better than random.

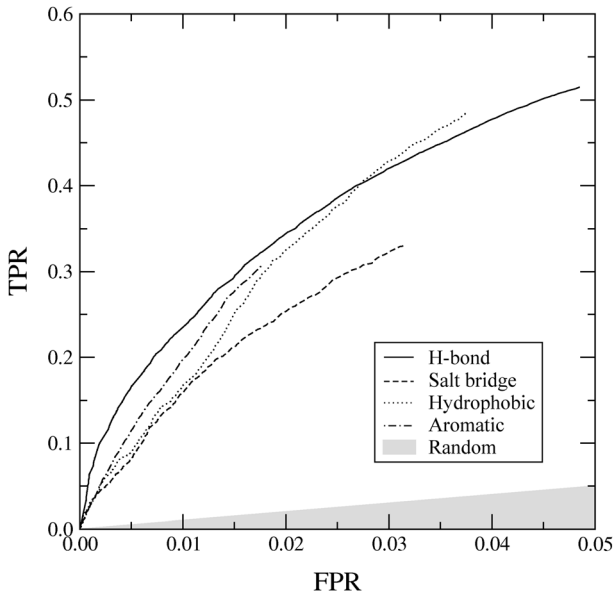

Figure 5. ROC plot for the prediction of various interaction types by *e*FindSite^{PPI} for the BM1905C dataset. The following non-covalent interaction types are considered: hydrogen bonds, salt bridges, hydrophobic, and aromatic contacts. TPR, true positive rate; FPR, false positive rate. Gray area corresponds to predictions no better than random.

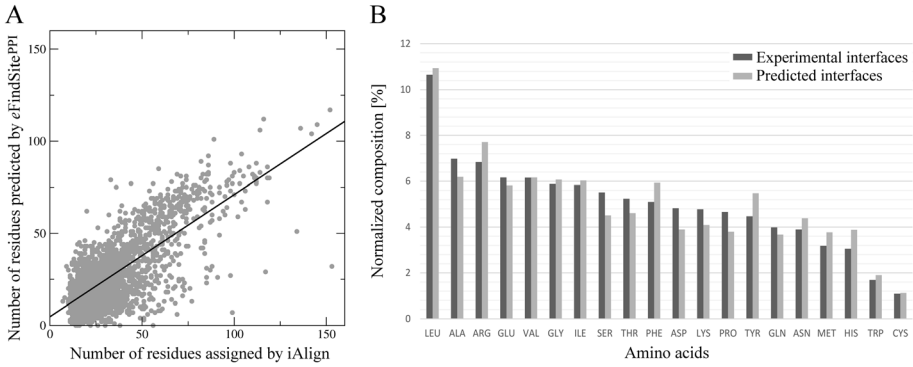

Figure 6. Size and composition of interfaces predicted by *e*FindSite^{PPI}. (A) The correlation between the size of experimental interfaces identified by iAlign and those predicted by *e*FindSite^{PPI}. (B) Amino acid composition of experimental and predicted interfaces.

protein–protein interfaces. The statistics collected for hydrogen bonds are shown in Supporting Information, Figure S3A; we note that because both side-chain and main-chain hydrogen bonds are taken into consideration, all amino acid types are included in this analysis. In general, interfaces are rich in hydrogen bonds, which are the major contributors to electrostatic interactions between proteins (Xu *et al*., 1997). The analysis of the composition of residues involved in the formation of hydrogen bonds at the predicted interfaces reveals that some polar residues are under-represented, for example, Arg, Glu, Asp and Ser (by 3.9, 4.0, 4.5 and 2.3%, respectively), whilst several hydrophobic residues are overpredicted to form hydrogen bonds, for example, Leu, Ala, Ile, Phe, Pro and Met (by 4.8, 2.8, 2.1, 2.6, 2.2 and 1.8%, respectively). The amino acid composition of residues predicted to interact with ligands through salt bridges, hydrophobic and aromatic contacts are comparable to that in the experimental complexes (Supporting Information, Figures S3B-E) except for Arg and Phe, which are slightly overpredicted to form electrostatic and hydrophobic contacts by 5.5 and 5.1%.

Susceptibility to target–template sequence similarity

The accuracy of template-based function inference certainly depends on the target–template sequence similarity; therefore, we analyze the performance of *e*FindSite^{PPI} at different similarity thresholds applied to the selection of evolutionarily related templates. Table 1 summarizes the results obtained at 40, 30 and 20% sequence similarity thresholds. The accuracy of protein interface prediction at 40 and 30% similarity thresholds is comparably high; however, the performance of *e*FindSite^{PPI} starts deteriorating at lower sequence similarity thresholds. For example, MCC is 0.428, 0.381 and 0.177 at 40, 30 and 20% sequence similarity, respectively. This corresponds to a true (false) positive rate of 0.464 (0.076), 0.415 (0.077) and 0.151 (0.042). Thus, excluding templates with >20% sequence identity to the target leads to an approximately twofold drop-off in the prediction accuracy compared with higher sequence identity thresholds. We note that this is a common feature of threading-based approaches to protein function inference from evolutionarily related templates and a similar behavior was observed in ligand-binding site prediction using *e*FindSite (Brylinski and Feinstein, 2013).

Table 1. Performance of *e*FindSite^{PPI} in interface residue prediction across the BM1905C dataset at different target–template sequence similarity thresholds

Similarity threshold	Evaluation metric					
	FPR	TPR	ACC	SPC	PPV	MCC
40%	0.076	0.464	0.835	0.924	0.594	0.428
30%	0.077	0.415	0.824	0.922	0.563	0.381
20%	0.042	0.151	0.800	0.957	0.459	0.177

FPR, false positive rate; TPR, sensitivity; ACC, accuracy; SPC, specificity; PPV, precision; MCC, Matthew's correlation coefficient.

Protein models as targets for PPI prediction

Similar to *e*FindSite, a recently developed algorithm to ligand-binding site prediction, the design of *e*FindSite^{PPI} makes it particularly well suited for structure-based PPI prediction using protein models. Therefore, in addition to target crystal structures, we benchmark *e*FindSite^{PPI} against computer-generated models. The details on model preparation and their structural characteristics are provided as Supporting Information. Benchmarking results for different quality models from the BM1905 dataset compared with experimental structures are presented in Table 2. Because small proteins involve proportionally more residues at interfaces compared with large targets, it is important to eliminate a potential bias caused by this size effect. To address this issue, several techniques for systematic corrections have been recently suggested (Martin, 2014). Table 2 also includes a random background that accounts for the size bias estimated for the BM1905 dataset. Only a fraction of surface residues contribute to PPIs; therefore, most residues assigned by a random classifier are true negatives, resulting in a relatively high accuracy (ACC) and specificity (SPC). However, sensitivity (TPR) and fall-out (FPR) are comparably low and close to the diagonal in a ROC space.

Using the SVM classifier in *e*FindSite^{PPI} yields slightly better performance than NBC, however, combining predictions from both machine learning algorithms (listed as *e*FindSite^{PPI} in Table 2) gives the highest accuracy. For instance, using target crystal structures, MCC for *e*FindSite^{PPI} is 0.428. The performance using protein models is only slightly worse with MCC of 0.371 for high-quality and 0.339 for moderate-quality models. Compared with a random, size-independent classifier, using *e*FindSite^{PPI} yields MCC values higher by 0.417 for target crystal structures, and 0.352 and 0.309 for high-quality and moderate-quality

Journal of Molecular Recognition

Table 2. Comparison of the performance of *e*FindSite^{PPI} and PINUP using different quality target structures

Dataset	Predictor	Evaluation metric					
		FPR	TPR	ACC	SPC	PPV	MCC
BM1905C	*e*FindSite^{PPI} (SVM)	0.150	0.581	0.760	0.850	0.483	0.403
	*e*FindSite^{PPI} (NBC)	0.208	0.627	0.760	0.793	0.421	0.366
	*e*FindSite^{PPI}	0.076	0.464	**0.835**	**0.924**	**0.594**	**0.428**
	PINUP	0.091	0.244	0.748	0.808	0.414	0.189
	Random	0.078	0.086	0.759	0.921	0.209	0.011
BM1905H	*e*FindSite^{PPI} (SVM)	0.161	0.539	0.785	0.838	0.418	0.344
	*e*FindSite^{PPI} (NBC)	0.228	0.590	0.739	0.771	0.357	0.304
	*e*FindSite^{PPI}	0.083	0.428	**0.829**	**0.916**	**0.522**	**0.371**
	PINUP	0.112	0.179	0.722	0.787	0.284	0.080
	Random	0.074	0.087	0.778	0.925	0.201	0.019
BM1905M	*e*FindSite^{PPI} (SVM)	0.169	0.517	0.775	0.839	0.393	0.314
	*e*FindSite^{PPI} (NBC)	0.233	0.571	0.732	0.766	0.341	0.281
	*e*FindSite^{PPI}	0.089	0.402	**0.822**	**0.910**	**0.489**	**0.339**
	PINUP	0.121	0.166	0.709	0.778	0.264	0.053
	Random	0.076	0.097	0.780	0.923	0.212	0.030

For *e*FindSite^{PPI}, three prediction protocols are evaluated: SVM only, NBC only and a combination of SVM and NBC (listed as *e*FindSite^{PPI}). Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly

BM1905C, crystal structures; BM1905H, high-quality models; BM1905M, moderate-quality models.

FPR, false positive rate; TPR, sensitivity; ACC, accuracy; SPC, specificity; PPV, precision; MCC, Matthew's correlation coefficient.

Random performance includes the correction of a size bias.

models. This analysis demonstrates that *e*FindSite^{PPI} is capable of tolerating distortions in modeled target structures.

Prediction confidence

A reliable confidence index is an essential feature to identify those targets, whose interface is likely to be correctly predicted. *e*FindSite^{PPI} uses an average probability score assigned by machine learning to target residues to categorize predictions as either high, medium or low confidence. In Figure 7, we report the prediction accuracy separately for each confidence group using target crystal structures as well as protein models from the BM1905 dataset. In general, confidence estimates correlate well with the actual prediction accuracy assessed by MCC across all datasets, that is, the average MCC for high-confidence predictions is significantly higher than those assigned medium and low confidence. For high-confidence predictions, using targets from the BM1905C, BM1905H and BM1905M datasets yields the median MCC of 0.623, 0.585 and 0.520, whereas for medium (low) confidence predictions, the median MCC is 0.383 (0.128), 0.246 (0.095) and 0.210 (0.086), respectively. As expected, the percentage of high-confidence predictions slightly decreases from 32 to 29% (28%) when high (low) quality protein models are used instead of the target crystal structures; this is shown in Supporting Information, Figure S2. To that end, *e*FindSite^{PPI} offers a reliable confidence index, which can be used to select only accurately predicted interfaces for large-scale protein docking simulations and other applications that may require a high precision.

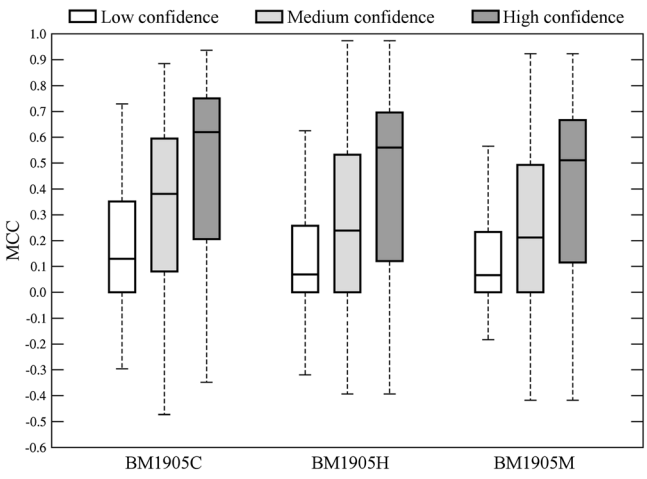

Figure 7. Accuracy of interfacial residue identification for predictions assigned different confidence levels. The accuracy is assessed by Matthew's correlation coefficient; boxes end at the quartiles Q_1 and Q_3 and a horizontal line in each box is the median. Whiskers point at the farthest points that are within 3/2 times the interquartile range.

Comparison with PINUP

We compare the performance of *e*FindSite^{PPI} to several structure-based approaches for protein-binding residue prediction. The first one is PINUP (Liang *et al*., 2006), a method that employs residue-level energy scores, accessible surface area-dependent interface propensities and conservation scores to derive a set of structural and functional constraints. PINUP effectively combines side-chain energy, residue conservation and interface propensity into a single score, which is used to build a consensus region from initial top-ranked patches. The corresponding weight factors were obtained by a linear optimization of the scoring function against a training dataset of 57 protein targets. Figure 4 shows that *e*FindSite^{PPI} is almost twice

as sensitive as PINUP on the BM4361 dataset; a true positive rate for *e*FindSite^{PPI} and PINUP is 0.446 and 0.236, at a comparably low false positive rate of 0.073 and 0.060, respectively. In Table 2, we assess the performance of both methods using experimental structures and different quality protein models from the BM1905 dataset. Consistent with benchmarking results against BM4361, *e*FindSite^{PPI} outperforms PINUP on crystal structures from the BM1905C dataset; for instance, MCC is 0.428 for *e*FindSite^{PPI} and 0.189 for PINUP. More importantly, the prediction accuracy for *e*FindSite^{PPI} against protein models from the BM1905H and BM1905M datasets is much higher than for PINUP. When high (moderate) quality models are used instead of the experimental structures, MCC for PINUP decreases by 0.109 (0.136), whereas for *e*FindSite^{PPI}, MCC decreases only by 0.057 (0.089). Thus, *e*FindSite^{PPI} tolerates structure deformations in protein models more efficiently than PINUP. These unequal performances of *e*FindSite^{PPI} and PINUP can be explained by differences in their prediction techniques. *e*FindSite^{PPI} mainly exploits template–target similarities using global structure alignments, which are fairly insensitive to local distortions in the target proteins, whereas PINUP employs local features, for example, side-chain conformations of individual amino acids as well as solvent accessible surface calculations to predict interface residues. Despite the correct global topology, the local characteristics of computer-generated models may deviate significantly from experimental structures, decreasing the performance of PINUP in binding interface prediction using non-native target conformations.

Next, we compare the performance of *e*FindSite^{PPI} and PINUP separately for 3,896 homodimers and 465 heterodimers identified in the BM4361 dataset. Table 3 shows that both algorithms perform better on homodimers compared with heterodimers; MCC for *e*FindSite^{PPI} (PINUP) is 0.419 (0.187) for homodimer and 0.289 (0.156) for heterodimers. Furthermore, consistent with previous results, *e*FindSite^{PPI} is roughly twice as sensitive as PINUP on both datasets of dimers. We note that the performance of algorithms for PPI site prediction is often different on homodimers and heterodimers; for example, Engelen *et al.* (2009) reported that the average performance of iJET and ET (Lichtarge *et al.*, 1996) were better on homodimers compared with heterodimers. This is because of the fact that homodimers often have a nearly perfect symmetric organization at the interface in contrast to mainly asymmetric interfaces in heterodimers.

Comparison with PrISE

In order to eliminate any potential prediction bias using one dataset, we evaluate the performance of *e*FindSite^{PPI} with respect to other methods on different protein sets. In addition to PINUP, we compare *e*FindSite^{PPI} with PrISE, a recently developed method that exploits local surface similarities to predict protein interfaces (Jordan *et al.*, 2012). This method extracts structural elements from a target protein and scans them through two databases of protein quaternary structures and protein–protein interface residues, ProtInDB (Jordan *et al.*, 2011) and PQS (Henrick and Thornton, 1998). The accuracy of PrISE was previously evaluated using the Protein–Protein Docking Benchmark dataset (Howook Hwang *et al.*, 2009). We ran *e*FindSite^{PPI} on the Benchmark 4.0 dataset following the same procedure as used in PrISE benchmarking (Jordan *et al.*, 2012). In this analysis, we also include results from PINUP reported for the Benchmark 4.0 dataset. Table 4 shows that *e*FindSite^{PPI} outperforms both

Table 3. Comparison of the performance of *e*FindSite^{PPI} and PINUP using homodimers and heterodimers from the BM4361 dataset

Dataset	Predictor	Evaluation metric					
		FPR	TPR	ACC	SPC	PPV	MCC
Homodimer	*e*FindSite^{PPI}	0.088	0.478	**0.820**	**0.911**	**0.574**	**0.419**
	PINUP	0.089	0.239	0.771	0.910	0.414	0.187
Heterodimer	*e*FindSite^{PPI}	0.093	0.354	**0.806**	0.906	**0.456**	**0.289**
	PINUP	0.090	0.217	0.773	**0.909**	0.368	0.156

Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly
FPR, false positive rate; TPR, sensitivity; ACC, accuracy; SPC, specificity; PPV, precision; MCC, Matthew's correlation coefficient.

Table 4. Comparison of the performance of *e*FindSite^{PPI}, PINUP and PrISE on the Benchmark 4.0 dataset. Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly

Dataset	Predictor	Evaluation metric				
		FPR	TPR	ACC	PPV	MCC
Bound	*e*FindSite^{PPI}	0.049	0.399	**0.909**	0.404	**0.352**
	PINUP	0.065	0.347	0.783	0.307	0.246
	PrISE	0.042	0.381	0.790	**0.432**	0.279
Unbound	*e*FindSite^{PPI}	0.047	0.377	**0.909**	**0.499**	**0.338**

Results for PINUP and PrISE are taken from ref. (Jordan *et al.*, 2012).
TPR, sensitivity; ACC, accuracy; PPV, precision; MCC, Matthew's correlation coefficient.

PrISE and PINUP; for example, the accuracy (MCC) is 0.909 (0.352), 0.790 (0.279) and 0.783 (0.246), respectively. Moreover, Benchmark 4.0 also provides apo structures for most of the target proteins; we use these conformations to evaluate the performance of *e*FindSite^{PPI} against unbound experimental structures to complement our previous analysis using protein models from the BM1905 dataset. The accuracy of *e*FindSite^{PPI} against bound and unbound structures is fairly comparable; using apo conformations only slightly decreases the sensitivity by 0.022 and MCC by 0.014. Thus, *e*FindSite^{PPI} performs better than other predictors on the Benchmark 4.0 dataset offering a high prediction accuracy using both bound as well as unbound experimental target conformations.

Comparison with ET and iJET

Finally, we compare *e*FindSite^{PPI} to evolution-based predictors, ET and iJET (Lichtarge *et al*., 1996; Engelen *et al*., 2009). Inspired by the Evolutionary Trace approach (Lichtarge *et al*., 1996), these methods identify PPI interfaces by detecting and analyzing conserved surface patches on target proteins. Evolutionary conservation is the primary feature for the identification of interface residues by both algorithms, as it reflects the evolutionary selection at interfacial sites to maintain the molecular function across protein families. The comparison with ET and iJET is based on the interface residue prediction for 52 protein chains derived from the Huang dataset (Caffrey *et al*., 2004). The targets are experimental structures in their bound conformational state and cover three categories of PPIs: non-transient homodimers, non-transient heterodimers and transient complexes. Table 5 summarizes the performance of *e*FindSite^{PPI}, ET and iJET in terms of sensitivity, specificity, precision and accuracy. Clearly, *e*FindSite^{PPI} produces quantitatively better results than ET and iJET across all targets. For instance, the sensitivity of *e*FindSite^{PPI} is 28.9% (33.8%), 20.8% (14.6%) and 21.2% (7.6%) higher than ET (iJET) on homodimers, heterodimers and transient complexes, respectively. However, despite a lower sensitivity for the transient complexes, iJET gives 7.8% higher precision compared with *e*FindSite^{PPI}. This analysis also shows that similar to ET and iJET, the performance of *e*FindSite^{PPI} decreases from non-transient homodimers to heterodimers to transient complexes. This is consistent with other studies demonstrating that, in contrast to proteins forming transient complexes, the prediction of non-transient interfaces is less complicated, because they are evolutionarily more conserved, larger and flatter (Ofran and Rost, 2003; Caffrey *et al*., 2004).

Case studies

To illustrate the prediction performance of *e*FindSite^{PPI}, we discuss a couple of representative examples. We note that these proteins are not present in the BM4361 dataset, thus have not been used in the construction of machine learning models. The first case study involves a NAD-dependent D-glycerate dehydrogenase (GDH) from *Hyphomicrobium methylovorum* (PDB-ID: 1GDH). This enzyme belongs to the family of oxidoreductases and catalyzes the NADH-linked reduction of 3-hydroxypyruvate to D-glycerate in the serine pathway for the assimilation of one-carbon compounds in methylotrophs (Izumi *et al*., 1990). The GDH molecule forms a homodimer composed of two structurally similar subunits related to each other by a twofold symmetry (Goldberg *et al*., 1994). Figure 8 presents the PPI interface predicted for a GDH monomer by *e*FindSite^{PPI} from remotely homologous templates. 59% of interfacial residues are correctly identified, with 0.992 specificity, 0.951 precision, and 0.909 accuracy (Figure 8A). Moreover, *e*FindSite^{PPI} correctly predicted 7 out of 16 hydrogen bonds as well as two out of five salt bridges present at the GDH interface. Figure 8B illustrates selected correctly identified interactions, including a salt bridge between the side chains of R129-chain A and D277-chain B, and hydrogen bonds between the side chain of R127-chain A and T281-chain B.

The second example is a mouse T cell receptor protein (TCR) (PDB-ID: 1TCR), which is localized on the surface of T cells and is responsible for their activation (Saito *et al*., 1984). These molecules participate in the recognition of antigens bound to major histocompatibility complexes (Wyer *et al*., 1999; van der Merwe and Davis, 2003). TCR is a membrane-anchored heterodimer composed of alpha and beta chains (Garcia *et al*., 1996); we use *e*FindSite^{PPI} to predict interfacial residues separately for both chains. Figure 9 shows that *e*FindSite^{PPI} correctly identified 65% of interfacial residues in chain alpha, with 0.946 specificity,

Table 5. Comparison of the performance of *e*FindSite^{PPI}, ET and iJET using non-transient homodimers and heterodimers as well transient complexes from the ET/iJET dataset

Dataset	Predictor	Evaluation metric				
		FPR	TPR	PPV	SPC	ACC
Homodimer	*e*FindSite^{PPI}	0.049	0.678	**0.657**	**0.951**	**0.917**
	ET	0.058	0.389	0.482	0.856	0.738
	iJET	0.038	0.340	0.552	0.905	0.764
Heterodimer	*e*FindSite^{PPI}	0.071	0.572	**0.614**	**0.929**	**0.871**
	ET	0.065	0.364	0.524	0.854	0.696
	iJET	0.062	0.426	0.575	0.824	0.707
Transient	*e*FindSite^{PPI}	0.048	0.531	**0.460**	**0.952**	**0.922**
	ET	0.032	0.319	0.431	0.906	0.727
	iJET	0.030	0.455	0.538	0.820	0.751

Results for ET and iJET are taken from ref. (Engelen *et al*., 2009).
Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly
TPR, sensitivity; PPV, precision; SPC, specificity; ACC, accuracy.

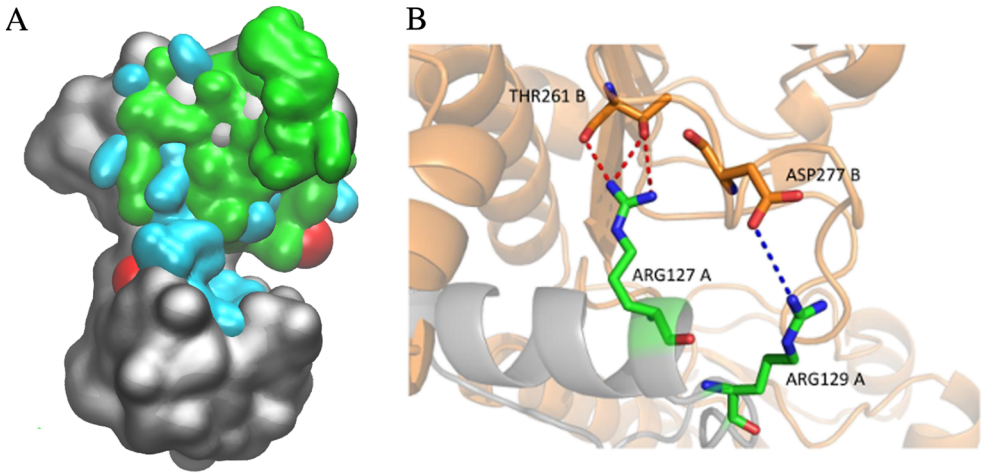

Figure 8. Example of PPI prediction by *e*FindSite^{PPI} for a homodimer (PDB-ID: 1GDH). (A) The surface representation of a monomer chain; true positives, true negatives, false positives, and false negatives are colored in green, gray, red, and cyan, respectively. (B) Interface residues correctly predicted to form specific interactions; dashed blue lines represent salt bridges and red lines represent hydrogen bonds.

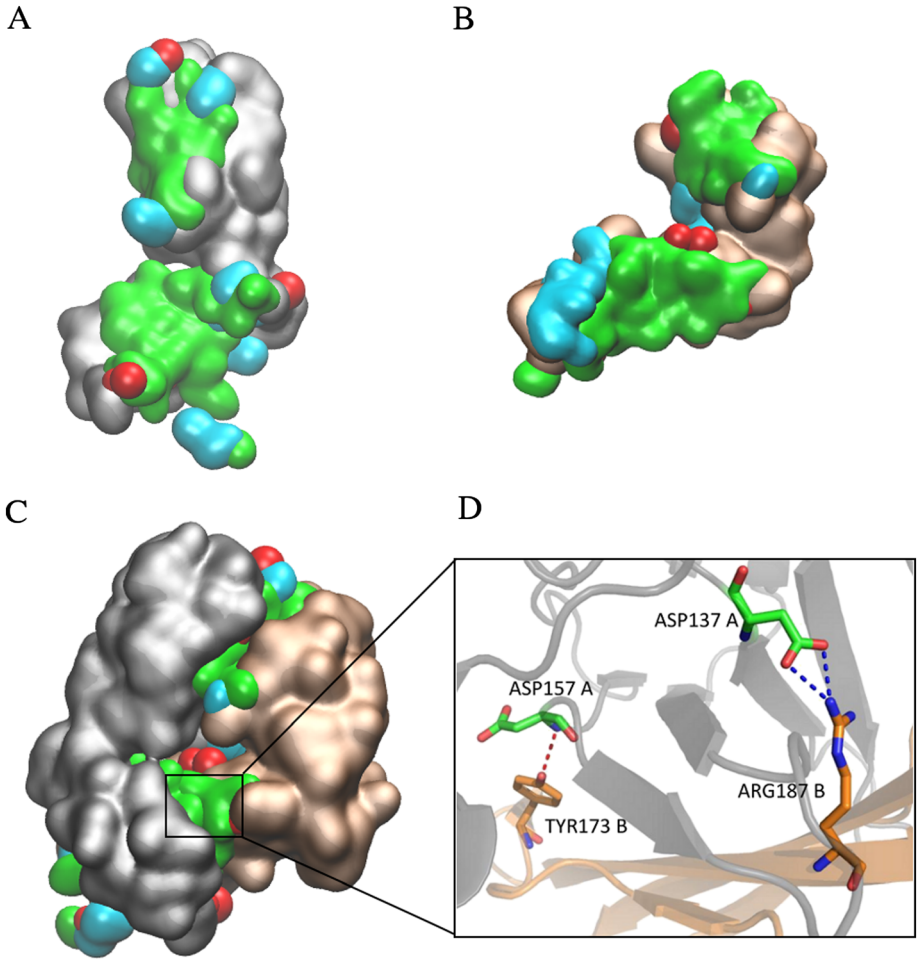

Figure 9. Example of PPI prediction by *e*FindSite^{PPI} for a heterodimer (PDB-ID: 1TCR). The surface representations of alpha and beta chains are shown in (A) and (B); the dimer complex is displayed in (C). True positives, true negatives, false positives, and false negatives are colored in green, gray/tan, red, and cyan, respectively. Interfacial residues in both chains correctly predicted to form specific interactions are shown in (D). Dashed blue lines represent salt bridges and red lines represent hydrogen bonds.

0.420 precision, and 0.929 accuracy (Figure 9A). For chain beta, 46% of interfacial residues are correctly predicted, with 0.815 specificity, 0.959 precision, and 0.817 accuracy (Figure 9B). Importantly, most false positives and false negatives in both chains are located at the rim of interface patches; thus, the prediction of the core interfacial residues is highly accurate. This

is evident in Figure 9C, which shows the heterodimer structure composed of alpha and beta chains interacting via two interfaces. Residues overpredicted and missed by *e*FindSite^{PPI} are mainly positioned either within the interfacial cavity or at the interface edge, whereas those predicted correctly make up the core of the TCR alpha–beta interface. Furthermore, *e*FindSite^{PPI} accurately identified three out of six interfacial hydrogen bonds and one out of two salt bridges stabilizing the dimer complex according to the experimental structure. Figure 9D illustrates two correctly predicted interactions: a salt bridge between the side chains of D137-alpha and R187-beta and a hydrogen bond between the main chain of D157-alpha and the side chain of Y173-beta. These examples demonstrate the capability of *e*FindSite^{PPI} to predict PPI sites, residues, and interaction types for homodimers as well as heterodimers using weakly homologous templates.

CONCLUSIONS

The analysis of evolutionarily weakly related dimer proteins reported in this study strongly suggests that the locations of their binding sites are highly conserved, irrespectively of the global structure similarity of protein–protein complexes. Furthermore, the interfacial geometry is preserved as well, thus can be predicted with a high accuracy. This is consistent with previous studies demonstrating that surface regions responsible for protein binding are conserved among structural neighbors (Zhang *et al*., 2010). Exploiting these insights, we developed *e*FindSite^{PPI}, a new approach for the prediction of protein-binding sites using information derived from evolutionarily and structurally related templates. *e*FindSite^{PPI} employs sensitive meta-threading by *e*Thread (Brylinski and Lingam, 2012) to identify evolutionarily related templates and extensively uses various machine learning techniques to detect interfacial residues on a query protein surface. A higher degree of conservation of local interface compared with the global structure of protein complexes forms the basis for an accurate prediction of interfacial binding sites. In addition to these conservation patterns, *e*FindSite^{PPI} also employs other residue-level descriptors to effectively discriminate between interfacial and non-interfacial residues. For instance, it incorporates the relative solvent accessible area and the interfacial propensities of amino acids, which have been already successfully used by several other interfacial site prediction algorithms (Liang *et al*., 2006; Li *et al*., 2008). A high accuracy in extracting structural information from the "twilight zone" templates motivated us to further extend the capabilities of *e*FindSite^{PPI} to predict specific interactions as well. That is, *e*FindSite^{PPI} also detects the types of molecular interactions that target proteins are likely to form with their interacting partners; this is demonstrated for hydrogen bonds, salt bridges as well as hydrophobic and aromatic contacts. Comparative benchmarking calculations on several datasets of protein dimers show that *e*FindSite^{PPI} outperforms other methods for protein-binding residue prediction. Equally important, it is designed to work with protein models so that the interfacial site can be efficiently predicted even when the experimental structure of a query protein is unavailable. Finally, a carefully tuned confidence estimation system identifies those predictions that are likely to be correct. *e*FindSite^{PPI} is freely available to the academic community as a user-friendly web-server and a well-documented stand-alone software distribution at http://www.brylinski.org/efindsiteppi; this website also provides all benchmarking datasets and results reported in this paper.

Acknowledgements

This study was supported by the Louisiana Board of Regents through the Board of Regents Support Fund [contract LEQSF (2012–15)-RD-A-05]. We thank Wei Feinstein and Misagh Naderi who read the manuscript and provided critical comments. Portions of this research were conducted with high performance computational resources provided by Louisiana State University (HPC@LSU, http://www.hpc.lsu.edu).

REFERENCES

Aloy P, Russell RB. 2003. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**(1): 161–162.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17): 3389–3402.

Armon A, Graur D, Ben-Tal N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**(1): 447–463.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**(1): 235–242.

Berman HM, Coimbatore Narayanan B, Di Costanzo L, Dutta S, Ghosh S, Hudson BP, Lawson CL, Peisach E, Prlić A, Rose PW, Shao C, Yang H, Young J, Zardecki C. 2013. Trendspotting in the Protein Data Bank. *FEBS Lett.* **587**(8): 1036–45.

Brylinski M. 2013. Unleashing the power of meta-threading for evolution/structure-based function inference of proteins. *Front. Genet.* **4**(June): 118.

Brylinski M, Feinstein WP. 2012. Setting up a meta-threading pipeline for high-throughput structural bioinformatics: eThread software distribution, walkthrough and resource profiling. *J. Comput. Sci. Syst. Biol.* **6**(1): 1–10.

Brylinski M, Feinstein WP. 2013. eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput. Aided Mol. Des.* **27**(6): 551–67.

Brylinski M, Lingam D. 2012. eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. *PLoS One* **7**(11): e50200.

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**(1): 190–202.

Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**: 1–39.

Chelliah V, Blundell TL, Fernández-Recio J. 2006. Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *J. Mol. Biol.* **357**(5): 1669–1682.

Chen X, Jeong JC. 2009. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **25**(5): 585–91.

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**: W665-W667.

Engelen S, Trojan L a, Sacquin-Mora S, Lavery R, Carbone A. 2009. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.* **5**(1): e1000267.

Gao M, Skolnick J. 2010. iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics* **26**(18): 2259–65.

Garcia KC, Degano M, Stanfield RL, Brunmark A, Jackson MR, Peterson PA, Teyton L, Wilson IA. 1996. An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. *Science* **274**(5285): 209–219.

Goldberg JD, Yoshida T, Brick P. 1994. Crystal structure of a NAD-dependent D-glycerate dehydrogenase at 2.4 A resolution. *J. Mol. Biol.* **236**(4): 1123–1140.
Guharoy M, Chakrabarti P. 2005. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U. S. A.* **102**(43): 15447–15452.
Halperin I, Ma B, Wolfson H, Nussinov R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**(4): 409–443.
Henrick K, Thornton JM. 1998. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**(9): 358–361.
Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. 2009. Protein-Protein Docking Benchmark Version 3.0. *Proteins* **73**(3): 705–709.
Hubbard SJ, Thornton JM. 1993. NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London.
Hwang H, Vreven T, Janin J, Weng Z. 2010. Protein-protein docking benchmark version 4.0. *Proteins* **78**(15): 3111–3114.
Izumi Y, Yoshida T, Yamada H. 1990. Purification and characterization of serine-glyoxylate aminotransferase from a serine-producing methylotroph, Hyphomicrobium methylovorum GM2. *Eur. J. Biochem.* **190**(2): 285–290.
Jones S, Thornton JM. 1996. "Review Principles of protein-protein interactions," vol. 93, no. January, pp. 13–20.
Jones S, Thornton JM. 1997. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**(1): 133–43.
Jones S, Thornton JM. 1997. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**(1): 121–32.
Jordan RA, Wu F, Dobbs D, Honavar V. 2011. "ProtinDb: A data base of protein-protein interface residues," *Iowa State Univ.* [http://protindb.cs.iastate.edu/].
Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. 2012. Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinf.* **13**(1): 41.
Jubb H, Higueruelo AP, Winter A, Blundell TL. 2012. Structural biology and drug discovery for protein–protein interactions. *Trends Pharmacol. Sci.* **33**(5): 241–248.
Koike A, Takagi T. 2004. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* **17**(2): 165–173.
Kortemme T, Kim DE, Baker D. 2004. Computational alanine scanning of protein-protein interfaces. *Sci. STKE* **2004**(219): pl2.
Krissinel E, Henrick K. 2007. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**(3): 774–797.
Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**(3): 379–400.
Li B, Kihara D. 2012. Protein docking prediction using predicted protein-protein interface. *BMC Bioinf.* **13**(1): 7.
Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**(3): 282–3.
Li N, Sun Z, Jiang F. 2008. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinf.* **9**: 553.
Liang S, Zhang C, Liu S, Zhou Y. 2006. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **34**(13): 3698–707.
Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**(2): 342–358.
Lijnzaad P, Argos P. 1997. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins* **28**(3): 333–43.
Martin J. 2014. Benchmarking protein-protein interface predictions: Why you should care about protein size. *Proteins* **82**(7): 1444–1452.
van der Merwe PA, Davis SJ. 2003. Molecular interactions mediating T cell antigen recognition. *Annu. Rev. Immunol.* **21**: 659–684.
Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **102**(31): 10930–10935.
Murakami Y, Mizuguchi K. 2010. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* **26**(15): 1841–1848.
Nooren IMA, Thornton JM. 2003. Diversity of protein-protein interactions. *EMBO J.* **22**(14): 3486–3492.
Obenauer J, Yaffe M. 2004. Computational prediction of protein-protein interactions. *Methods Mol. Biol.* **261**: 445–68.
Ofran Y, Rost B. 2003. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**(1–3): 236–239.
Pandit SB, Skolnick J. 2008. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinf.* **9**: 531.
Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A. 2008. Computational methods for predicting protein-protein interactions. *Adv. Biochem. Eng. Biotechnol.* **110**: 247–67.
Porollo A, Meller J. "Prediction-Based Fingerprints of Protein – Protein Interactions," vol. 645, no. December 2006, pp. 630–645, 2007.
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**(Suppl 1): S71–S77.
Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**(7062): 1173–1178.
Saito H, Kranz D, Takaqaki Y, Hayday A, Eisen H, Tonegawa S. 1984. A third rearranged and expressed gene in a clone of cytotoxic T lymphocytes. *Nature* **312**(5989): 36–40.
Shanon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**: 379–423.
Shih ESC, Hwang M-J. 2013. A critical assessment of information-guided protein-protein docking predictions. *Mol. Cell. Proteomics* **12**(3): 679–86.
Sowa ME, He W, Wensel TG, Lichtarge O. 2000. A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl. Acad. Sci. U. S. A.* **97**(4): 1483–1488.
Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG. 2001. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.* **8**(3): 234–237.
Wang B, Sun W, Zhang J, Chen P. 2013. Current Status of Machine Learning-Based Methods for Identifying Protein-Protein Interaction Sites. *Curr. Bioinf.* **8**(2): 177–182.
Wells JA, McClendon CL. 2007. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**(7172): 1001–1009.
Wyer JR, Willcox BE , Gao GF, Gerth UC, Davis SJ, Bell JI, van der Merwe PA, Jakobsen BK. 1999. T cell receptor and coreceptor CD8 alphaalpha bind peptide-MHC independently and with distinct kinetics. *Immunity* **10**(2): 219–225.
Xu D, Tsai CJ, Nussinov R. 1997. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* **10**(9): 999–1012.
Zhang H. 2004. The Optimality of Naive Bayes. *Mach. Learn.* **1**(2): 3.
Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**(4): 702–710.
Zhang QC, Petrey D, Norel R, Honig BH. 2010. Protein interface conservation across structure space. *Proc. Natl. Acad. Sci. U. S. A.* **107**(24): 10896–901.
Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**(7421): 556–60.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website.