PIER: Protein Interface Recognition for Structural Proteomics

Irina Kufareva,[1] Levon Budagyan,[2] Eugene Raush,[2] Maxim Totrov,[2] and Ruben Abagyan[1,2*]
[1]*Scripps Research Institute, La Jolla, California 92037*
[2]*Molsoft LLC, La Jolla, California 92037*

ABSTRACT **Recent advances in structural proteomics call for development of fast and reliable automatic methods for prediction of functional surfaces of proteins with known three-dimensional structure, including binding sites for known and unknown protein partners as well as oligomerization interfaces. Despite significant progress the problem is still far from being solved. Most existing methods rely, at least partially, on evolutionary information from multiple sequence alignments projected on protein surface. The common drawback of such methods is their limited applicability to the proteins with a sparse set of sequential homologs, as well as inability to detect interfaces in evolutionary variable regions. In this study, the authors developed an improved method for predicting interfaces from a single protein structure, which is based on local statistical properties of the protein surface derived at the level of atomic groups. The proposed Protein IntErface Recognition (PIER) method achieved the overall precision of 60% at the recall threshold of 50% at the residue level on a diverse benchmark of 490 homodimeric, 62 heterodimeric, and 196 transient interfaces (compared with 25% precision at 50% recall expected from random residue function assignment). For 70% of proteins in the benchmark, the binding patch residues were successfully detected with precision exceeding 50% at 50% recall. The calculation only took seconds for an average 300-residue protein. The authors demonstrated that adding the evolutionary conservation signal only marginally influenced the overall prediction performance on the benchmark; moreover, for certain classes of proteins, using this signal actually resulted in a deteriorated prediction. Thorough benchmarking using other datasets from literature showed that PIER yielded improved performance as compared with several alignment-free or alignment-dependent predictions. The accuracy, efficiency, and dependence on structure alone make PIER a suitable tool for automated high-throughput annotation of protein structures emerging from structural proteomics projects. Proteins 2007;67:400–417. © 2007 Wiley-Liss, Inc.**

Key words: protein–protein interaction; structural proteomics; cell signaling and protein recognition; structure–function annotation; alignment-independent interface prediction

INTRODUCTION

As crystallographers continue producing novel protein structures with fully or partially unknown function, the question arises of what aspects of their biological function can be predicted from those structures. Predicting the propensity of a protein to form complexes with other proteins, the location of the interfaces, and possible oligomeric states[1,2] is of particular importance because of the role of protein interactions and associations in molecular biology.[3,4] While modern docking algorithms are getting better at predicting protein association *geometries* (see Refs. 5–9 for reviews), they can only be used when identities and three-dimensional structures of *all* partners are known; and even for those cases, the prediction is further complicated by the induced fit, incompleteness or inadequate quality of available structures, and computer requirements. Most often, however, we either do not know what the second protein is, or do not have its structure. Reliable prediction of protein binding interfaces from a *single* protein with a known 3D structure, therefore, becomes a key computational problem.

Existing methods for protein interface prediction can be divided into two classes: (i) methods incorporating evolutionary information in the form of certain conservation measures derived from multiple sequence alignments

Abbreviations: ASA, solvent accessible surface area; CAPRI, critical assessment of prediction of interactions; MSA, multiple sequence alignment; NEIT, nonenzyme-inhibitor transient interaction; ODA, optimal docking area; PDB, protein data bank; PIER, protein interface recognition; PLS, partial least squares regression; RMSD, root mean square deviation; SVM, support vector machines.

Grant sponsor: NIH; Grant number: 5-R01-GM071872-02.

The PIER predictor is available on the web: http://abagyan.scripps.edu/~kufareva/pier.cgi. The dataset of 748 protein interfaces with the accompanying information can also be downloaded from this web site.

*Correspondence to: Ruben Abagyan, 10550 North Torrey Pines Rd., Mail TPC-28, La Jolla, CA 92037. E-mail: abagyan@scripps.edu

Received 9 December 2005; Revised 2 May 2006; Accepted 16 August 2006

Published online 13 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21233

(MSA) and projected on protein surface, and (ii) those based solely on geometrical, physicochemical, and statistical properties of the surface. Several successful methods of the first class were developed and published recently. Some of them are based on the evolutionary information alone, for example, the Evolutionary Trace method by Lichtarge et al.[10,11] that was further complemented by residue cluster analysis,[12] invariant polar residues mapping,[13] maximum parsimony approach (MP-ConSurf),[14,15] maximum likelihood calculation (Rate4Site),[16] robust incorporation of alignment reliability (REVCOM),[17] and so forth. Other evolutionary methods combine alignment-derived information with the properties of the protein surface, either by obtaining a combined heuristic score (e.g., ProMate by Neuvirth et al.[18]), or by using machine learning approaches such as support vector machines (SVM)[19–22] and neural networks.[23–25]

The reliability of evolutionary conservation scores derived from MSA as determinants of protein interface has been questioned by many authors. In spite of the broad evidence that the interfaces generally mutate at slower rates than the rest of protein surface,[10,26–28] it was argued that conservation score alone is not sufficient for accurate discrimination; moreover, it can be misleading in several ways.[29–32] First, the high variability of alignment composition and extent, unbalanced subfamily representations, and local alignment errors (e.g., shifts) are the issues that need to be taken into account.[33] Alignments can be easily contaminated with paralogs that do not share the same interfaces.[29,34] Furthermore, even orthologous proteins sometimes vary in their quaternary structure and binding interfaces. Second, the prediction greatly depends on the algorithm of deriving scores from the alignment.[11,14,16,17,35–37] Third, even the most sophisticated algorithms break down on the proteins with no or few orthologs. Fourth, many protein interfaces are not expected to be better conserved at all, either because of their function (e.g., the adaptable binding surfaces of the immune system proteins) or because they were formed late in evolution; such interfaces are undetectable with alignment-dependent methods.[30] Finally, the small ligand binding sites are usually better conserved than large protein interfaces,[31,36] and thus, using conservation scores actually leads to increased rate of false positives in automatic protein interface prediction. All these factors, therefore, limit applicability and reliability of the alignment-dependent methods.

Alignment-independent prediction methods rely on an assumption that protein interfaces are different from the rest of the surface by their physicochemical and geometrical properties. While it was demonstrated that the composition of protein interface patches had statistically significant biases,[38–47] the attempts of using the differences for patch discrimination have encountered several difficulties. The physical properties of the interfaces are highly diverse[44,48,49] between various protein families and complex types. Moreover, even within a single interface the binding energy is not distributed evenly among residues; instead, there are so-called "hot spots," which contribute most of the interaction energy, while the other interface residues are of relatively minor importance.[50–52] Finally, the extent and shape of a protein patch in which the small local biases accumulate into a statistically significant signal is not known in advance. In spite of these complications, Jones and Thornton[53] demonstrated that alignment-independent interface identification was possible for 39 out of 59 complexes (66%). Later, it was shown that desolvation is indicative of protein interfaces.[54–56] The optimal docking area (ODA) method[56] based on modified desolvation energies was reported to predict interfaces for 42 out of 53 (80%) heterodimeric transient complexes. The interaction patterns can also be captured by machine learning methods applied to large protein interaction benchmarks: for example, Keil et al.[57] developed a neural network based predictor, which detected 44% of protein interfaces in the set of 7821 structures, which constituted 76% of all PDB structures available at the time of publication.

The purpose of this study is to improve the reliability and accuracy of alignment-independent protein interface prediction, and to evaluate to what extent incorporating evolutionary information may help the interface identification. We developed a method of protein interface prediction from local statistical properties of the protein surface at the atomic-group level that can (but does not have to) be further complemented by evolutionary conservation scores. We used a cross-validated partial least squares (PLS) regression algorithm[58] and evaluated the significance of each protein surface feature in the resulting prediction model. The contribution of the evolutionary signal was as little as 7–10%, with the rest 90–93% being contributed by atomic group composition descriptors, and adding this signal only marginally influenced the prediction performance. Moreover, for certain classes of proteins, using conservation scores actually resulted in deteriorated prediction.

The proposed alignment-independent method demonstrated improved performance over the previously published methods. On a diverse benchmark of 748 proteins known to be involved in homodimeric and heterodimeric interactions, permanent as well as transient, the overall precision at the residue level was 60% at the recall threshold of 50%. The method was also tested on other benchmarks. Using the method, we identified potential new interfaces and corrected mislabeled oligomeric states.

MATERIALS AND METHODS

Data Set

A diverse set of dimeric interfaces was taken from a recent publication,[19] and carefully checked for biological correctness. In several cases we found that the true oligomeric state was different from the assumed dimeric state; these pairs were eliminated from the dataset. Each of the remaining complexes was assumed to be either a permanent dimer or a transient complex. Permanent dimers were classified into homodimers (sequence identity $\geq$90%) and heterodimers (sequence identity <90%). Within the set of transient complexes, enzyme-inhibitor interactions were separated from nonenzyme-inhibitor transient (NEIT) ones.

This produced a dataset of 490 monomers with homodimeric permanent interfaces, 62 monomers with heterodimeric permanent interfaces, and 76 proteins involved in transient interactions. In the latter group, 12 proteins were classified as enzymes with inhibitor-binding interfaces, 12 were inhibitors with enzyme-binding interfaces, and the remaining 52 were molecules involved in other transient interactions. To avoid bias related to underrepresentation of enzyme-inhibitor interactions in the dataset, we had to collect a separate set of enzymes and their protein inhibitors present in PDB.[59] All short chains (less than 70 residues for enzymes and less than 25 residues for inhibitors) were discarded. From the rest of the set combined with the original 24 enzyme-inhibitor interactions, we iteratively removed all sequences sharing more than 50% identity with other sequences. This procedure produced the set of 85 enzymes with protein inhibitor interfaces and 59 protein inhibitors with enzyme-binding interfaces. Among the enzymes, there were 21 serine (EC 3.4.21), 13 cysteine (EC 3.4.22), 13 aspartic (EC 3.4.23), and 4 metallo- (EC 3.4.24) endopeptidases, 1 aminopeptidase (EC 3.4.11), and 1 metallocarboxypeptidase (EC 3.4.17), 53 proteases in total. The obtained set of enzymes and protein inhibitors was added to the dataset, resulting in the total of 748 interfaces. The dataset with accompanying information is available on the web: http://abagyan.scripps.edu/~kufareva/pier.cgi.

For each monomer in the dataset, solvent accessible surface areas (ASAs) were calculated for all heavy atoms by a modified version of the Shrake and Rupley[60] algorithm implemented in ICM,[61] in three states: monomer alone (unbound state), monomer with all small ligands present in the PDB structure, and monomer with its protein partner (bound state).

A residue was called an internal residue, if the combined ASA of all its heavy atoms did not exceed 3 Å^2. If this was the case, or if more than 3 Å^2 of residue surface participated in an interaction with small ligands, the residue was omitted from the calculations. Each of the other residues was classified as an interface residue, if its ASA differed by more than 20 Å^2 between bound and disjoint states; otherwise it was called a noninterface residue.

Statistical Analysis

All heavy atoms in the 20 naturally occurring amino acids were classified into 32 types according to their chemical element, formal charge, sp-, sp2-, or sp3-hybridization, and the number and the type of covalently bound heavy atoms. Given a protein molecule with known interface, let N be the total number of atoms on its solvent accessible surface, with A_t of them being atoms of type t, $1 \leq t \leq 32$. Also, let n denote the number of interface atoms, with a_t being the corresponding number of atoms of type t.

Assuming the null hypothesis to be a uniform distribution of the atoms over the protein surface, the probability of finding exactly a atoms of type t on the interface would be:

$$P(a) = \frac{\binom{A_t}{a} \times \binom{N - A_t}{n - a}}{\binom{N}{n}}$$

(the so-called *hypergeometric* distribution). Accordingly, the probability of finding at least a_t atoms would be $P(a \geq a_t) = \sum_{a=a_t}^{n} P(a)$ and the probability of finding at most a_t atoms would be $P(a \leq a_t) = \sum_{a=0}^{a_t} P(a)$. If calculated $P(a \leq a_t)$ was less than 0.05, the atoms of type t were said to be *significantly underrepresented* on the interface, and in case of $P(a \geq a_t) < 0.05$ they were *significantly overrepresented*.

Some atomic types are too rare to be qualified as significant upon analysis of single proteins; however, their presence correlates well with the location of interfaces for larger data clusters. For these types, we substituted single proteins by arbitrary sets of 15–20 proteins, and determined the described above probabilities for combined values of N, A_t, n, and a_t within every set.

As expected, the occurrence statistics strongly correlated between some atom types because of the structural constraints; for example, the representational bias of α-carbon and peptide bond nitrogen was the same, since they are always covalently bonded in the protein structure. Such covalently linked atom types were merged into groups, thus avoiding singularities and improving the robustness of the model with respect to rearrangements of small atomic details on protein surface. The obtained groups are listed in Table I.

Patch Generation

A method for surface patch generation was adopted from Ref. 56. First, the solvent accessible surface of the protein was expanded by 3 Å, and a set of evenly distributed surface points was calculated by dividing the surface into triangles with an average side of 5 Å. Next, each point was assigned a surface patch, consisting of all solvent accessible heavy atoms of the protein, located within a certain distance of the point (see Fig. 1). Because of the preliminary surface expansion, the points were located at an average distance of 3 Å from the surface. This preserved the overall protein surface shape while ignoring the small atomic details; it also allowed avoiding, as much as possible, patches "leaking through" the protein interior and including accessible atoms located on the "other side."

For the purpose of interface prediction, the optimal patch size was found to be between 900 and 1000 Å^2. However, to ensure that the patch size was adequate for each individual protein, the distance for patch generation

TABLE I. Normalized PIER Parameters for the 12 Significant Atom Groups

				PIER-value	
	Atom group name	Amino acids	Atoms	Mean	Rmsd
1	AROM6	F	$C_\gamma\ C_\delta^{1,2}\ C_\varepsilon^{1,2}\ C_\zeta$	0.214	0.021
		W	$C_\varepsilon^3\ C_\zeta^{2,3}\ C_\eta^2$		
2	CH3	A	C_β	0.164	0.023
		I	$C_\gamma^2\ C_\delta^1$		
		L	$C_\delta^{1,2}$		
		M	C_ε		
		T	C_γ^2		
		V	$C_\gamma^{1,2}$		
3	Y	Y	$C_\gamma\ C_\delta^{1,2}\ C_\varepsilon^{1,2}\ C_\zeta\ O_\eta$	0.159	0.008
4	S	C	S_γ	0.132	0.009
		M	S_δ		
5	CH	IV	C_β	0.067	0.015
		L	C_γ		
6	AROM5	H	$C_\gamma\ N_\delta^1 C_\delta^2 C_\varepsilon^{1,2}\ N_\varepsilon^1$	0.037	0.012
		W	$C_\delta^1\ N_\delta^2 C_\varepsilon^2$		
7	RCN3+	R	$N_\varepsilon\ C_\zeta\ N_\eta^{1,2}$	0.016	0.016
8	BKBN	ACDEFGHIKL MNPQRSTVWY	$C_\alpha\ C\ O\ N$	0.007	0.003
9	CON	N	$C_\gamma\ O_\delta^1\ N_\delta^2$	−0.003	0.012
		Q	$C_\delta\ O_\varepsilon^1\ N_\varepsilon^2$		
10	CH2	DEFHKLMNPQRWY	C_β	−0.021	0.008
		EKPQR	C_γ		
		I	C_γ^1		
		K	C_δ		
11	KN^+	K	$C_\varepsilon\ N_\eta$	−0.085	0.014
12	COO^-	D	$C_\gamma\ O_\delta^{1,2}$	−0.088	0.023
		E	$C_\delta\ O_\varepsilon^{1,2}$		

See also the diagram in Figure 3.

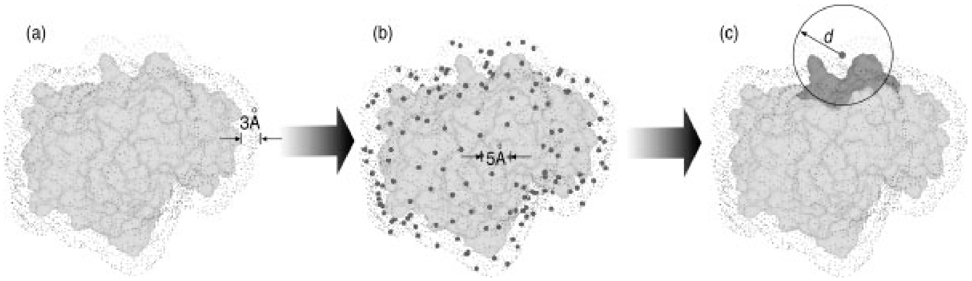


Fig. 1. Surface patch generation steps: (**a**) Protein solvent accessible surface is expanded by 3 Å; (**b**) Evenly distributed points are generated on the expanded surface; (**c**) Each point it assigned a patch consisting of all the solvent accessible atoms within *d* Å from the point, where *d* is calculated from the Eq. (1).

was calculated as a function of the total solvent ASA of the molecule:

$$d = \sqrt{27.235 + 0.018 \times \mathrm{ASA}_{400}^{10000}} \tag{1}$$

where $\mathrm{ASA}_{400}^{10000}$ is the ASA of the isolated molecule, trimmed to fit in the range of [400,10000] Å^2. The coefficients for this equation were obtained by the regression on ASAs and patch sizes. For proteins whose total ASA exceeded 10000 Å^2, using Eq. (1) produced the distance of 14 Å, which resulted in surface patches with average ASA between 900 and 1000 Å^2. The deviations in the values of ASA between different patches reflected the curvature and packing of the surface atoms within the patch.

Using the patch generation procedure for all 748 proteins in our dataset, we generated 232,170 surface patches. The patches were further classified into interface, noninterface, boundary, or small molecule ligand binding patches, using the following cutoffs. A patch was considered an interface patch, if it lost more than 50% of its accessible area upon binding to another protein. If the ratio was less than 5%, the patch was assigned to be a noninterface patch. The patches with the ratio between 5 and 50% were considered boundary patches. If a patch lost some of its accessible area upon binding to a small molecule ligand present in a PDB structure, the patch was classified as a small ligand binding patch.

For the purpose of training a predictive model, boundary and small-ligand binding patches (39% of all patches) were omitted, leaving the set of 141,972 interface and noninterface patches. Of these, 33,082 patches (23%) were interface patches and the rest were noninterface patches.

The described procedure of patch generation was also used at the first stage of the interface prediction algorithm (The PIER Algorithm for Protein IntErface Recognition). No preliminary patch classification or filtering was performed in that case.

Patch Descriptors

Physical descriptors

Given a protein surface patch, for each of the 12 significant atomic groups we calculated the combined solvent accessible area of all atoms of this group within the patch. This produced 12 descriptors for each patch. We refer to these descriptors as *physical descriptors*, since they do not exploit any sequence-derived information.

Alignment-based descriptors

To incorporate evolutionary information, we collected homologous sequences for every protein in the dataset, using BLAST[62,63] search of SwissProt database[64] (release 47.4 of July 05, 2005) with a *P*-value cutoff of 10^{-5}. Only sequences with pairwise identities less than 90% were retained. These sequences were then aligned using the ZEGA algorithm[65,66] as implemented in ICM.[61] For each residue r in the reference sequence, the column was extracted from the alignment, and residue conservation scores $S_{\mathrm{f}}\,(r)$, $S_{\mathrm{m}}\,(r)$, and $S_{\mathrm{e}}\,(r)$ were calculated by three methods:

1. Frequency-based conservation score:

$$S_{\mathrm{f}}(r) = \frac{1}{n}\sum_{j=1}^{n} S_{\mathrm{f}}(r, r_j), \text{where } S_{\mathrm{f}}(r, r_j) = \begin{cases} 1, & \text{if } r_j = r \\ 0, & \text{otherwise} \end{cases}$$

2. Matrix-based conservation score (with Gonnet et al.[67] substitution matrix):

$$S_{\mathrm{m}}(r) = \frac{1}{n}\sum_{j=1}^{n} S_{\mathrm{m}}(r, r_j),$$

$$\text{where } S_{\mathrm{m}}(r, r_j) = \frac{c(r, r_j)}{\sqrt{c(r, r) \times c(r_j, r_j)}}$$

3. Entropy-based conservation score:

$$S_{\mathrm{e}}(r) = \frac{1}{1 - \sum\limits_{a=1}^{20} f_a \log f_a},$$

where f_a is the relative frequency for amino acid type a in the alignment column.

The residue conservation scores were projected onto the structure; for each surface patch, the combined patch conservation score was derived as a weighted sum of conservation scores for all residues within the patch. Thus, in addition to the 12 physical descriptors, each patch was assigned three alignment-derived descriptors: frequency-based, matrix-based, and entropy-based ones.

Deriving the Predictor Parameters by PLS Regression

PLS regression[58] is a recent technique that generalizes and combines features from principal component analysis and multiple regression. The ultimate goal is to approximate a dependent (target) variable Y with a linear combination of independent variables X_i (descriptors), that is $Y = w_0 + \Sigma\ w_i\ X_i$. The output of the learning algorithm then consists of a linear coefficient w_i for each of the descriptors X_i and a free term w_0.

Unlike SVMs and other nonlinear machine learning algorithms, PLS regression is transparent and does not have a tendency for generalization errors due to overfitting. While it is known to be sensitive to the noise created by occasional irrelevant outliers, we found that careful selection of the dataset reduces this effect. Furthermore, PLS regression has an advantage of direct numerical assessment, and thus comparison, of contribution of each of the descriptors. Weighed normalized contribution of descriptor X_i is given by

$$\frac{w_i^{\mathrm{rel}}}{\sum\limits_{j} |w_i^{\mathrm{rel}}|} \tag{2}$$

where $w_i^{\mathrm{rel}} = \frac{w_i}{RMSD(X_i)}$ is the normalized weight of the descriptor X_i.

In this study, PLS regression was used to approximate patch function (interface or noninterface) as a linear combination of the combined solvent ASAs of the 12 significant atomic groups within the patch. A recent implementation of PLS in ICM software[68] was used. The training set for PLS consisted of the patches generated as described in Patch Generation, with each patch represented by the set of descriptors X_i from Patch Descriptors. The dependent (target) variable Y was the patch type, equal to 1 for interface patches, and -1 for noninterface patches. This procedure produced the set of 12 linear coefficients, whose signs and magnitudes reflected the tendency of each of the atomic group to participate in protein complex formation, which justified their use for PIER: Protein IntErface Recognition (Fig. 2).

To estimate the importance of evolutionary scores, the PLS regression was trained to distinguish between interface and noninterface patches in two modes: alignment-independent and alignment-dependent. In the first mode, the model was trained solely on the 12 physical descriptors from Physical Descriptors, in the second, alignment-derived descriptors from Alignment-Based Descriptors were also taken into account. This allowed the comparison of the performance between the two modes and numerical assessment of the contribution of conservation scores (see Minor Contribution of Evolutionary Rates).

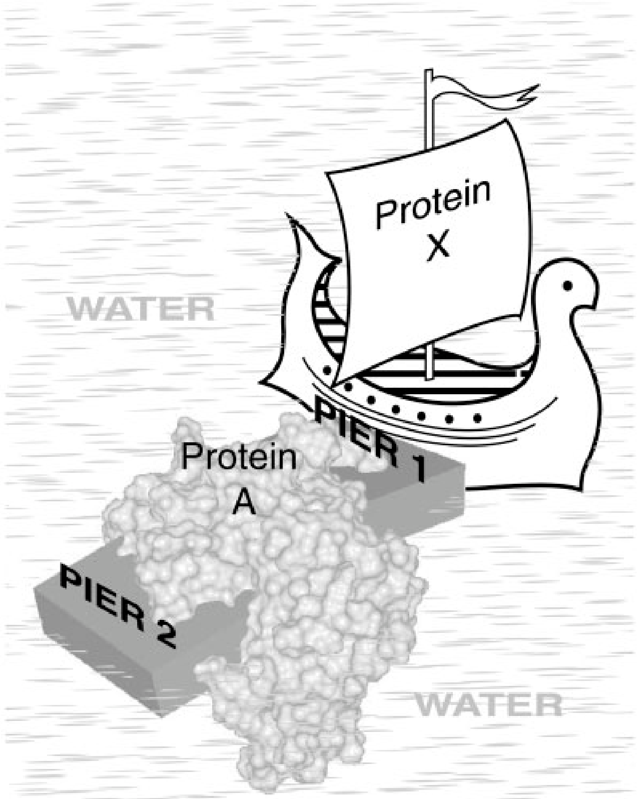


Fig. 2. The Protein IntErface Recognition algorithm detects patches on the surface of the known Protein A, that are likely to serve as "piers" for Protein X, known or unknown.

The PIER Algorithm for Protein IntErface Recognition

The PIER predictive model, which consists of the set of PIER parameters, is used for predicting potential interfaces on individual proteins. Given a protein, the prediction algorithm consists of the following steps:

1. Generate the set of overlapping surface patches for the protein (Patch Generation):
   a. Expand the solvent accessible surface of the protein by 3 Å.
   b. Calculate a set of evenly distributed surface points by dividing the expanded surface into triangles with an average side of 5 Å.
   c. Calculate the adequate distance

$$d = \sqrt{27.235 + 0.018 \times \mathrm{ASA}_{400}^{10000}},$$

   where $\mathrm{ASA}_{400}^{10000}$ is the ASA of the protein (isolated), trimmed to fit in the range of [400,10000] Å^2.
   d. For each point on the expanded surface, calculate the surface patch, consisting of all solvent accessible heavy atoms of the protein, located within d Å of the point (see Fig. 1).
2. For each patch P, calculate values for the patch descriptors X_i (Patch Descriptors):
   a. For every significant atomic group (Table I), calculate the total ASA of all atoms of the group within the patch, thus producing 12 physical descriptors.
   b. In the alignment-dependent mode, additionally calculate three evolutionary descriptors as described in Alignment-Based Descriptors.
3. For each patch P, obtain the patch PIER value as $\mathrm{PIER(P)} = \sum_i w_i X_i$.
4. Transfer patch PIER values to residues:
   a. Calculate the average PIER value for each solvent accessible surface atom as the sum of PIER values of all surface patches including the atom, weighted by the relative ASA of the patches and the relative ASA of the atom within the patches.
   b. Calculate the average PIER value for each surface residue as the arithmetic mean of PIER values of its accessible atoms.

A higher PIER value assigned to a residue suggests that the residue is more likely to be on an interface. Thus, the *continuous* PIER value itself can provide valuable information about strong and weak interfaces, as well as surface spots unfavorable for binding. However, for the purpose of performance evaluation, we need to generate a binary prediction for each residue. This can be achieved by imposing a *decision cutoff* during the following step:

5. Given a decision cutoff C, obtain a binary prediction for residues by

$$PIER_{01}(r) = \begin{cases} 1, & \text{if } \mathrm{PIER}(r) \geq C \ (r \text{ is an interface residue}) \\ 0, & \text{if } \mathrm{PIER}(r) < C \ (r \text{ is a non-interface residue}) \end{cases}$$

Obviously, the prediction result depends on the chosen decision cutoff. Higher cutoff separates highly interface-like residues and can possible lead to undetected weak interfaces, while lower cutoff often results in most of the protein surface being declared interface-like, thus providing no meaningful information. Thus, we suggest considering the raw PIER value for surface residues in real-life situations, whereas binary prediction is only used for performance assessment and comparison.

RESULTS AND DISCUSSION

Atomic Group Preferences on Protein Interfaces

Protein–protein interfaces in general, and the "hot spots" in particular, tend to have physical properties different from the rest of the surface. For example, they have been shown to differ in their hydrophobicity,[19,39,40,45–47] polarity,[43] topography,[42] and so forth. However, since we do not know the exact shape of an interface patch in advance, these differences at the residue level are insufficient to reliably predict protein–protein interfaces. The problem is further complicated for the transient complexes in which the binding patch can also be exposed to solvent and thus the composition differences are even less pronounced.

The statistical analysis of the large representative set of 748 protein interfaces (Data Set and Statistical Analysis) identified atomic types and groups whose representation at the interfaces was significantly biased as compared with the rest of the surface. The following atomic types were found to be significantly overrepresented on the interfaces:

- sp2-hybridized carbon in 6-member aromatic rings;
- sp3-hybridized carbon with 1 or 3 hydrogens in side chains of aliphatic residues;
- sulfur in cysteine, cystine, and methionine;
- sp2-hybridized carbon and nitrogen in 5-member heteroaromatic rings; and
- guanidinium group in arginine.

This result agrees well with the key role that aliphatic and aromatic residues are known to play in protein interaction.[44,47,69–71] The major role of cysteine sulfur atom is well known for heterogenous obligate complexes, for which two or more chains are derived from a single chain precursor and are held together with disulphide bonds. However, sulfur atoms in cystine, which are already involved in disulphide bond within one of the interacting partners, and in methionine, were also found to be significantly overrepresented on all types of interfaces, including homodimeric and transient ones. Indeed, the ability of sulfur-containing residues to form hydrogen bonds and nonhydrogen-bond type interactions was mentioned in literature.[72,73] The guanidinium group of arginine was also found in abundance on protein interfaces. It is able to form as many as four hydrogen bonds with neighboring atoms, thus playing an important stabilizing role in protein complex formation; however, being a charged group, it tends to be exposed to water. This is the reason arginine is often found on the periphery of protein interfaces, while the central part is usually hydrophobic.[69,74]

Additionally, some of atomic types were significantly underrepresented on the interfaces, as compared with the rest of the surface. These were as follows:

- terminal charged nitrogen in lysine;
- sp3-hybridized carbon with 2 hydrogens, as in alkyl portions of large residues;
- carbon and oxygen in carboxyl groups of aspartate and glutamate; and
- carbon, oxygen, and nitrogen in amide functional groups of asparagine and glutamine.

The low representation of carboxyl groups of aspartate and glutamate, the amide group of asparagine and glutamine, and the charged lysine at interfaces is expected, since it is energetically favorable for these charged groups to be exposed to solvent rather than buried. The low occurrence of alkyl (CH) portions of long amino acids, charged as well as aliphatic, is interesting. However, this is not true for transient complexes, which do not display such a clear dislike toward alkyl stems of large residues. Additionally, we found that the α-carbon atoms, the carbonyl groups and nitrogen in peptide bonds are underrepresented on permanent homodimeric interfaces, but not on heterodimeric or transient ones. This means that stable permanent complexes are mostly formed by side chain interactions, while transients extensively exploit backbone/backbone or backbone/side chain interactions. In other words, the permanent interfaces tend to be more "hairy" than the transient ones.

For the purpose of interface recognition, we classified all the protein atoms into 12 atomic groups, according to their interfacial representation bias as well as structural constraints. This improved statistical consistency of the model as well as its tolerance to small-scale side chain movement. The obtained atomic groups are listed in Table I. It is worth noting that these groups are very similar to those selected by Zhang et al.[75] for the purpose of developing effective statistical potentials named atomic contact energies.

PIER Parameters

Table I summarizes mean normalized PIER parameters for the 12 atomic groups derived from the 748 protein interface dataset with PLS regressions (Deriving the Predictor Parameters by PLS Regression). The signs and magnitudes of the parameters match the distribution of corresponding groups between protein interfaces and the rest of the surface, as discussed in Atomic Group Preferences on Protein Interfaces. For example, the parameter for carbon in a six-member aromatic ring is positive with the absolute value as high as 0.214, while the parameter of carboxyl groups of aspartate and glutamate is negative (−0.088). The high positive contribution of aromatic and aliphatic groups together with the negative contribution of some charged groups shows that an "average" interface formation is significantly driven by desolvation.

The PIER parameters appeared to be robust and did not deviate much between several nonoverlapping data subsets obtained by random partitioning of the original dataset (see Table I, column "rmsd"). A comparative diagram for the normalized parameters derived from three nonoverlapping subsets is shown in Figure 3(a), where narrow black, dark grey, and light grey bars correspond to the first, second, or third subset of data. These parameters were derived without separating the interfaces into different interface types; thus, they represent an average interface in each of the three subsets. The low deviation of the parameters on nonoverlapping datasets is an evidence the robustness of PIER model.

To analyze possible differences in atomic group composition and PIER parameter values between different interface group types, we trained the model separately on the sets of permanent homodimers, permanent heterodimers, and transient complexes. Even in this case, the obtained parameters were similar [Fig. 3(b)]. The three

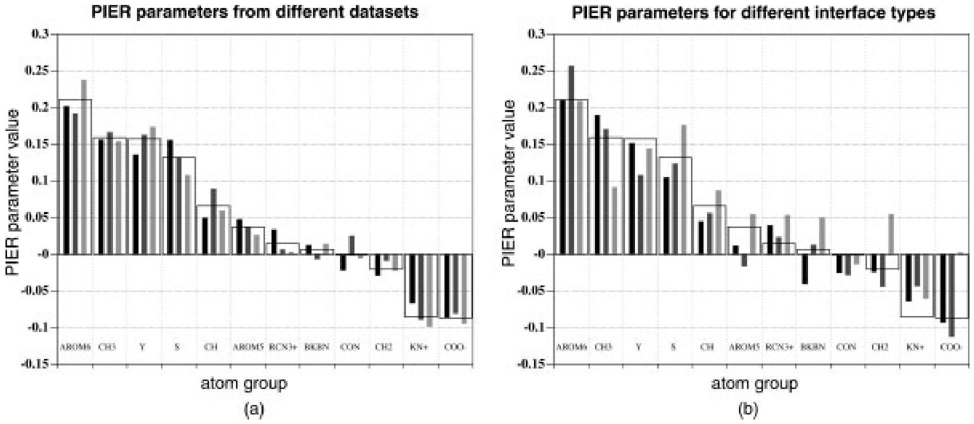


Fig. 3. (**a**) Comparative plot of PIER parameters derived from three nonoverlapping datasets obtained by a random partitioning of the 748 interface dataset. Narrow black, dark grey, and light grey bars represent normalized values derived from the first, second, and third subsets, respectively; wide white bars show average values. (**b**) PIER parameters for the sets of permanent homodimers (black bars), permanent heterodimers (dark grey bars), and transient interfaces (light grey bars). Atom group nomenclature corresponds to that in Table I.

distinguishing features of PIER parameters derived from transient complexes alone are their (i) positive propensity toward alkyl portions of long residues such as Lys, Met, and Arg, (ii) positive propensity toward backbone atoms, and (iii) reduced penalty (or increased "tolerance") to the acidic carboxyl groups of Asp and Glu. While these differences reflect possible compositional diversity between permanent and transient interfaces, as well as between various transient interfaces, they are not critical. In other words, the similarity of PIER parameters derived from the full set of the interfaces and only the transient ones indicates that transient interfaces can be predicted to some extent with a model, trained on permanent dimers, and vice versa.

Protein IntErface Recognition

Identification of atomic groups with biased representation on protein interfaces is of major importance for prediction algorithms. However, at the level of individual atoms or residues, the difference is usually too weakly pronounced to be used for interface/noninterface discrimination. Indeed, *in vivo* the ability of two proteins to form a complex is determined by *cooperative contribution* of many residues located within a certain vicinity of the site of immediate interaction. The weighted averaging implemented in PIER (The PIER Algorithm for Protein IntErface Recognition) approximates such recognition processes *in silico* by aggregation of interface-propensity values from individual atomic groups into values for larger surface patches, which is further decomposed into values for individual residues.

A higher PIER value assigned to a residue suggests that the residue is more likely to participate in a protein complex formation. Thus, the *continuous* PIER values can assist the protein characterization by indicating the location of strong and weak interfaces, as well as surface spots unfavorable for binding. For the purpose of performance assessment and comparison, these values can be further converted into a simpler binary prediction if a certain *PIER decision cutoff* is applied. In this case all the surface residues fall into one of the two classes: interface-like and noninterface-like.

Measuring prediction performance

Because of the inherent complexity of the problem and inevitable dependence of the final patch on the unknown partner, some interface residues are recognized as noninterface-like and some noninterface ones are predicted to be interface-like; these erroneous predictions are called *false negatives* (FN) and *false positives* (FP), respectively. Correctly predicted interface and noninterface residues are referred to as *true positives* (TP) and *true negatives* (TN). For any individual protein TP + FN is the size of the true interface, in residues, and TP + FP is the size of predicted patch.

Recall (sensitivity) and precision (positive predictive value) are the two most commonly used performance measures.[18–21] They are formally defined as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Informally, recall indicates the ratio of true interface residues detected by the predictor, and precision indicates the ratio of detected residues that are true interface residues.

Both measures depend on the selected decision cutoff. Lowering the cutoff results in better recall values with deteriorated precision, and vice versa. The decision cutoff selection in each individual situation is influenced by several factors, one of them being the tolerance ratio to false positives versus false negatives. Some authors purposely selected a higher cutoff, to yield better precision with a still reasonable recall value,[20] while others reported comparable values for the two statistics,[21] or choose higher recall at moderate precision.[19] Another important factor for decision cutoff adjustment is the desired type of interface: for example, transient interfaces usually generate lower signal than permanent ones, thus, for their detection the cutoff must lowered.

In spite of these arguments, we find it potentially confusing to adjust the decision cutoff for each individual protein by selecting, for example, the "top ten" predictions rather than all predictions above a fixed cutoff. Such strategy can be misleading because the fraction of the interface can vary from 0 to 100%. For example, ovalbumin, an amino-acid source, has no (known) protein binding interfaces, while in α-actin and histones H1–H4 most of the surface residues are involved in various protein interactions. For some proteins, a weak transient interface can be overshadowed by another stronger one, forming the top prediction portion (cell division kinase CDK2, for example, has a weak interface with cyclin-dependent kinases regulatory subunit 1,[76] and a strong one with cyclin A[77]).

To compare prediction methods in a more complete and objective way, the following statistics are reported in this study:

1. Overall recall-precision curve for the dataset—values are shown for every possible cutoff, with the overall precision at 50% recall specifically pointed out.
2. Precision and recall for individual proteins in the dataset at a fixed cutoff.
3. Absolute and relative number of proteins with precision exceeding 25% or 50% at 50% recall.

PIER performance analysis and comparison with other models

The model was cross-validated using the following procedure. The set of 748 protein interfaces was randomly divided into three nonoverlapping subsets, approximately equal in size, and the model was trained on each of them separately, yielding three different PIER models (with three corresponding sets of PIER parameters). After that each PIER model derived from a dataset was applied to the interface prediction in two other datasets. The three overall recall-precision curves for such cross validation are shown on Figure 4 in black. From these curves, one can see that on average, the model was able to achieve overall precision of 60% at 50% recall on cross validation.

Why are not these numbers higher? The main reason is that the exact shape of the interaction patch depends strongly on the partner, and we are trying to make a universal, partner-independent prediction. The contacting surface regions of the interacting proteins can extend in different directions around the "hot spot," and are often observed in sites not meant for protein interactions at all. However, on such extended interfaces, only the "hot spots" can usually be detected by partner-independent prediction methods, which results in lower recall with respect to entire interfaces. Also, frequently binding sites for other protein partners, small ligands, or domains missing in the structure are detected by the prediction methods; they are interpreted as false positives, which results in deteriorated precision. In other words, ignoring the nature of the second partner and the ambiguity of interface definition make ideal 100% values unachievable.

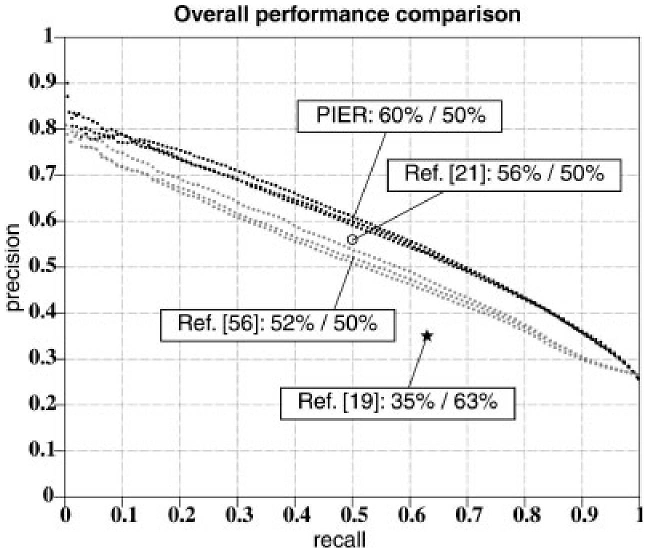


Fig. 4. The recall and precision statistics for four different methods: cross validated PIER on three datasets (black curves), ODA[56] (grey curves; on the same datasets), SVM-based predictors from[19] ★, 35% precision at 63% recall, 99% in common with our dataset), and[21] ○, 56% precision at 50% recall, 26% in common with our dataset).

On the other hand, the precision value obtained by a random partitioning of the dataset ("flip-a-coin" prediction) is expected to be equal to the ratio of interface residues to all residues in the dataset. For most proteins, only a small fraction of surface residues are involved in various protein interactions. The ratio of interface to all surface residues in the considered dataset of 748 proteins was equal 0.25. Therefore, a random "prediction" was expected to yield 25% precision at any recall threshold. The 60%/50% values achieved by PIER for the large dataset are statistically significant at the level below 10^{-300}, according to Fisher's exact test.

For comparison with previous work, we obtained the recall-precision statistics from ODA[56] prediction for the same three datasets. The corresponding curves are shown on Figure 4 in grey. At 50% recall, the ODA method was able to achieve only 52% precision on average. Available recall-precision values from two other similar works are represented on the same plot as black star (★, 35% precision at 63% recall[19]) and white circle (○, 56% precision at 50% recall[21]). In both papers, protein interfaces are predicted with alignment-dependent methods based on SVM. The dataset used by Koike and Takagi[21] consisted of 563 single-chain proteins, with 146 of them identical to proteins in our dataset; in the dataset of Bordner and Abagyan,[19] 627 of 632 interfaces were common with our dataset.

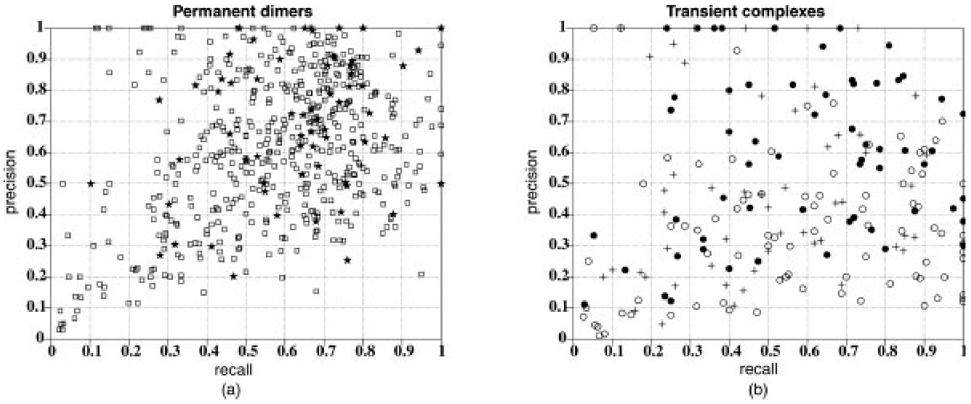


Fig. 5. The PIER predictor performance for all 748 individual proteins in the dataset: (**a**) 552 permanent interfaces, (**b**) 196 transient interfaces. Each mark on the plots represents a pair of recall and precision values for a single protein monomer, with the recall along the horizontal axis, the precision along the vertical axis, and the interface type coded by the mark shape: □ for a permanent homodimeric interface, ★ for a permanent heterodimeric interface, + for a NEIT (nonenzyme-inhibitor transient) interface, ○ for an inhibitor interface on an enzyme, and ● for an enzyme interface on a protein inhibitor. To obtain these plots, the PIER decision cutoff was fixed at the level, corresponding to equally balanced values of overall recall and precision for the entire dataset (56.8%).

TABLE II. Mean PIER Performance for Different Interface Types

Mark	Interface type	# of interfaces	Mean recall, %	Mean precision, %	Precision at 50% recall	
					≥25%	≥50%
□	Permanent homodimers	490	55	62	472 (96%)	368 (75%)
★	Permanent heterodimers	62	61	69	61 (98%)	53 (86%)
○	Enzyme-inhibitor	85	46	40	62 (73%)	37 (44%)
●	Inhibitor-enzyme	59	58	60	57 (97%)	44 (75%)
+	Nonenzyme-inhibitor-transient (NEIT)	52	46	47	44 (85%)	22 (42%)

The marks in the leftmost column correspond to those in Figure 5.

To assess the PIER predictor performance for the *individual proteins* in the dataset, we first fixed the PIER decision cutoff at the level corresponding to equally balanced values of overall recall and precision for the entire dataset (56.8%). Then for each of 748 protein monomers, the surface residues were classified into interface and noninterface residues, based on their PIER value compared with the selected cutoff. Recall and precision values for individual proteins were calculated using Eqs. (3)–(4). The obtained pairs for all 552 permanent and 196 transient interfaces were plotted against each other on a coordinate plane, with the recall along the horizontal axis and the precision along the vertical axis (see Fig. 5). Thus, each mark on the plots represents a pair of recall and precision values for a single protein monomer, with the interface type coded by the mark shape: □ for permanent homodimeric interface, ★ for permanent heterodimeric interface, + for NEIT interface, ○ for inhibitor interface on an enzyme, and ● for enzyme interface on a protein inhibitor. The most successful predictions are obviously located in the top right quadrant of the plot; however, all the marks except ones in the bottom left corner correspond to the interfaces detected with reasonable quality. For many "unsuccessful" proteins the decision threshold adjustment was required, reflecting the fact that their interfaces were weakly pronounced as compared with an average interface in the dataset. Under such relaxed conditions, the binding patch residues were successfully detected with precision exceeding 25 at 50% recall for 696 of 748 proteins (93%); for 524 proteins (70%) the corresponding precision was above 50%. Table II quantifies the variations of prediction quality for different interface types.

The PIER model was also validated by applying it to alternative datasets from related publications. The dataset by Bradford and Westhead[20] is a manually generated, high-quality set of 180 proteins, involved in transient (70) and obligate (110) interactions. The overall precision our model achieved for this dataset at 50% recall was as high as 61.8%. The individual protein analysis showed that the precision at 50% recall exceeded 25% for 169 proteins (94%), and was greater than 50% for 124 proteins (69%).

When tested on the dataset of transient complexes collected by Mintseris and Weng,[32] the PIER model yielded success (precision higher than 25% at 50% recall) for 278 of 340 interfaces (81.7%). Careful examination of unsuccessful predictions revealed that the high rate of false positives is mostly due to the presence of alternative interfaces that are known to interact with other partners, membranes, or domains missing in the structure.

Antibody–antigen and other asymmetric protein interactions

In most considered pairwise protein interactions, both interacting surfaces had a high value of the PIER value. However, in some protein pairs the interface prediction was strongly asymmetric: for one of the interacting protein partners the PIER signal was strong at the interface, while the signal for another partner was nearly undetectable.

In the class of antibody–antigen complexes the asymmetry was the best pronounced: the antigen-binding loops on antibodies had extremely high PIER values, but the epitopes (antibody binding patches) on antigens were virtually indistinguishable from the rest of protein surface in most cases. The rationale for this is the fact that since binding antibodies is not a part of biological function of antigens, the epitopes *cannot* be considered biological interfaces. Therefore, they cannot be detected with a model trained on surface patches suitable for biologically significant protein interactions; they require a separate predictive model. On the other hand, binding antigens *is* the main biological function of antibodies, which is reflected in their strong, as measured by the intensity of the PIER signal, interfaces. The considered examples of antibody–antigen complexes show that under the pressure of natural selection, the evolution of antibodies goes toward formation of highly adhesive interfaces that allow them to attach to various, often nonfunctional, patches on antigen surfaces.

When validating PIER on large datasets, we left all antibody/antigen complexes out as extreme high/low cases. Exclusion of these complexes also allowed adequate comparison with other methods in literature.

Interestingly, examples of similar but milder asymmetry were found in other classes of protein interactions. For example, the average PIER parameters could easily detect the enzyme-binding interfaces on inhibitor proteins, while getting only a weak signal on some of the enzymes. The high prediction accuracy for inhibitors was unexpected: these interfaces are notoriously difficult to detect because of their high evolutionary variability.[30] In fact, several authors suggested that protein inhibitors, required to adapt to a range of proteases, undergo a process of selection, similar to that in immune system proteins.[78–82] The strong PIER signal we found for protein inhibitor surfaces shows that this selection also goes toward formation of well-pronounced "adhesive" enzyme-binding interfaces. However, the interface propensity signal on corresponding inhibitor-binding surfaces on enzymes is often weak, and sometimes absent, suggesting the mechanism that the enzymes might use to escape inhibition.

Figure 6 gives an example of asymmetric enzyme-inhibitor interface: one of bacterial β-lactamase and its inhibitory protein BLIP (PDB entry ***1JTG***). Examples of asymmetric interactions were also found in other classes of regulatory protein–protein complexes.

Consistent with these considerations, the overall prediction performance was lower for enzyme-inhibitor and NEIT interfaces than for permanent interfaces or inhibitor-enzyme ones. To understand the difference between strong and weak interfaces, we tried dividing the dataset into "permanent" and "transient" interfaces, and training the model on the two sets separately, but this did not prove to be very productive. In fact, more than 77% of enzyme-inhibitor and NEIT interfaces were detected well with the average interface PIER parameters proving to have properties similar to permanent and inhibitor–enzyme interfaces. However, careful examination of the weak interfaces led us to identification of a special class of proteins that required a separate prediction model. These protein interfaces possess features making them unusual and undetectable with average PIER parameters: (i) they often have an exposed carboxyl group that is buried upon complexation, and (ii) the ratio of backbone to side chain atoms is much higher than on average. Such interfaces rely mostly on salt bridge and backbone–backbone hydrogen bond formation, as opposed to the majority of interfaces that are in significant degree driven by desolvation.

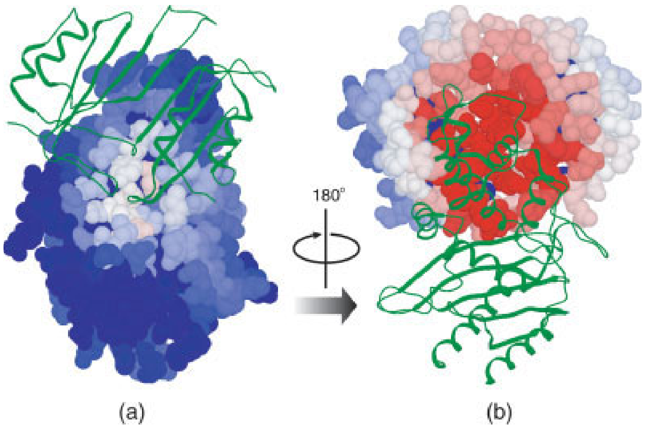


Fig. 6. Asymmetric PIER prediction for the interface of β-lactamase TEM (EC 3.5.2.6, PDB entry 1JTG) and β-lactamase inhibitory protein: (**a**) the β-lactamase molecule is shown in CPK representation and colored according to the PIER prediction (blue—low, red—high), the inhibitory protein backbone in shown in ribbon to indicate the location of the true interface; (**b**) the complex is rotated 180° around the vertical axis, and the partner representation is switched: the inhibitory protein is in CPK, while the enzyme is in ribbon. The PIER signal intensity on the inhibitory protein (b) is much higher than that on the enzyme (a).

The obtained results imply that for some complexes (for example, antibody–antigen and protease–protein inhibitor ones), the protein partners within the complex demonstrate different propensities toward complex formation with each other. The "reluctant" binders have the interface atomic composition that is different from most other protein interfaces, which sometimes allows them to escape complexation or to form only a short-living weak interaction.

Validation on CAPRI targets

One important application of the protein interface patch predictions is to assist protein docking by identifying protein surfaces that can be excluded from sampling. Reducing the extent of surfaces to be sampled may play a critical role in the success of protein docking. To test the ability of the PIER predictor to reliably exclude the unlikely protein interfaces, we tested the algorithm on a set of CAPRI targets. CAPRI[83] is a community-wide experiment on the comparative evaluation of protein–protein docking. Since the experiment started in 2001, there have been 18 confirmed targets in 7 rounds, of which 8 are antibody–antigen and other immune system complexes, 2 are enzyme-inhibitor complexes, and 8 are NEIT complexes. Most targets undergo backbone rearrangements upon complex formation.

The PIER predictor was applied to the 34 proteins from 18 CAPRI targets. The protein surface residues were classified as interface or noninterface based on their PIER value compared with a decision cutoff; they were further checked against the true interacting residues and classified to be true positives, true negatives, false negatives, or false positives. Prediction results are presented in Figure 7 and Table III. The correct interface patch was detected for 27 of 34 proteins, with 6 of 7 unsuccessful predictions being antigens from antibody/antigen complexes. This well illustrates the statement about epitopes being different from biological interfaces, discussed in Antibody–Antigen and Other Asymmetric Protein Interactions. In fact, the only epitope that PIER was able to detect had a significant overlap with the biological (substrate- and inhibitor-binding) interface on the antigen protein (the pig α-amylase, target T06, also PDB entry ***1KXQ***). For 14 proteins the only predicted patch was the correct one; for 10 more two patches were predicted, one of them being correct. Most sites initially marked as "false positives" were found to be alternative biologically meaningful interfaces or sites of interactions with other domains missing in the structure. According to Fisher's exact test of statistical significance, the prediction was significant at a level below 5% for 25 proteins (22 below 1%).

In protein docking, the rigid-body search reduction always implies a certain tradeoff between the increased speed and the danger of discarding the true interface. For the considered CAPRI targets, excluding residues with the PIER value below the cutoff preserved at least part of the correct interface, while reducing the sampling space from 100% to fractions ranging from 13 to 46% for each of the monomers (mean reduction to 29.1%). This corresponds to approximately 10-fold reduction of the search space for the complex. The positive prediction results for CAPRI targets are also evidence of high robustness of the PIER model with respect to potential backbone or side chain rearrangements of protein partners upon complex formation.

Minor Contribution of Evolutionary Rates

To obtain a numerical estimate of the role of evolutionary information in PIER, we used the PLS regression in two modes for each of the several nonoverlapping subsets obtained by random partitioning of the original dataset. In the alignment-independent mode, only combined solvent accessible areas of the significant atomic groups were used as surface patch descriptors. In the alignment-dependent mode, three alignment-based descriptors, namely, frequency, similarity, and entropy scores (see Alignment-Based Descriptors), were added to the accessible areas. The relative contribution of each surface descriptor was calculated using Eq. (2) (Deriving the Predictor Parameters by PLS Regression). In both modes, the contribution of the atomic group descriptors to the resulting interface score was almost the same for the three datasets. In the alignment-dependent mode, the combined contribution of the three alignment-derived descriptors deviated from 7 to 10% between the three datasets. Compared with the total contribution of the atomic composition-based descriptors (90–93%), this number is negligible.

The overall recall-precision curves, obtained with and without evolutionary information, are shown on Figure 8(a). From these curves, it can be easily seen that on average, adding conservation scores only marginally influences the model performance. For example, the overall precision is 60.8% at recall of 50% with the conservation scores, and 60.2% without the conservation scores.

In Figure 8(b), precision at 50% recall from the alignment-independent prediction is plotted against that derived in the alignment-dependent mode, for all individual monomers in the dataset. The plots demonstrate that only with the exception of several enzymes, the prediction quality without conservation scores is virtually indistinguishable from the prediction quality with the conservation scores.

This result appears to be even more significant if we take into account the fact that our dataset did not contain any immune system proteins, undergoing somatic mutations. These proteins were purposely not included to eliminate possible bias toward poorly conserved protein interfaces. Indeed, including such proteins into the dataset would likely lead to increased advantage of the alignment-free prediction over the alignment-dependent one.

Overall, the obtained statistics suggests that using evolutionary conservation scores brings little improvement for protein interface prediction. Therefore, leaving out the evolutionary signal might yield a more transparent, general, and reliable prediction model.

Prediction of Protein Oligomeric States with PIER

Formation of finite size homo-oligomers is a specific case of protein–protein association, and predicting such associations is an important step in annotation of newly solved proteins. The PIER method can also be applied to this task. The conventional approach to the problem takes advantage of the crystallographic symmetry, and assumes that the crystallographic neighbor burying the total solvent accessible surface over a certain threshold represents the biological oligomeric partner. However, this approach is limited and needs to be extended with an independent prediction in some cases. First, the biological symmetry can be incompatible with the crystallographic symmetry (e.g., a five-fold or a seven-fold symme-

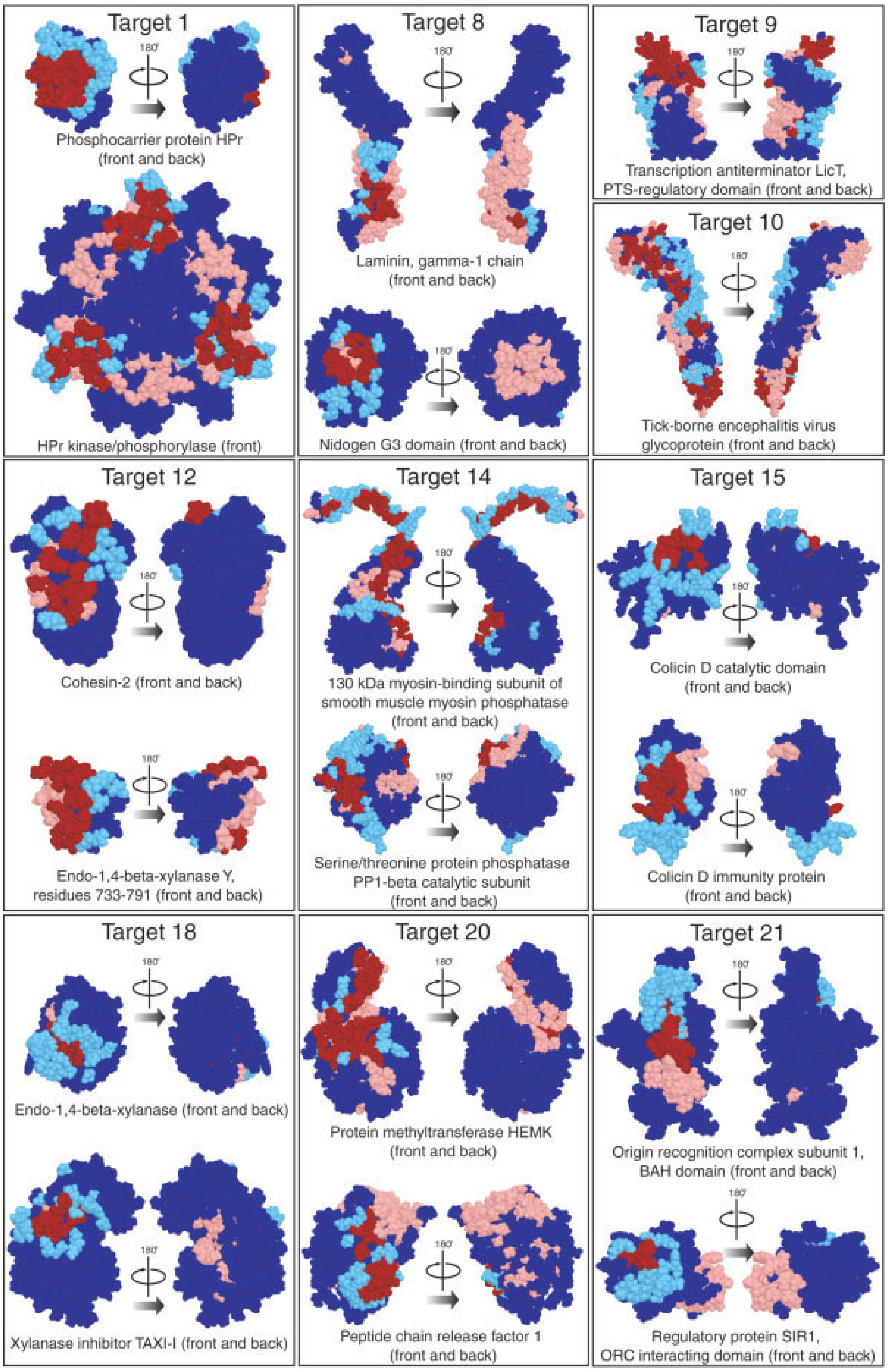

Target 1
180°
Phosphocarrier protein HPr
(front and back)
HPr kinase/phosphorylase (front)
Target 8
180°
Laminin, gamma-1 chain
(front and back)
180°
Nidogen G3 domain (front and back)
Target 9
180°
Transcription antiterminator LicT,
PTS-regulatory domain (front and back)
Target 10
180°
Tick-borne encephalitis virus
glycoprotein (front and back)
Target 12
180°
Cohesin-2 (front and back)
180°
Endo-1,4-beta-xylanase Y,
residues 733-791 (front and back)
Target 14
180°
130 kDa myosin-binding subunit of
smooth muscle myosin phosphatase
(front and back)
180°
Serine/threonine protein phosphatase
PP1-beta catalytic subunit
(front and back)
Target 15
180°
Colicin D catalytic domain
(front and back)
180°
Colicin D immunity protein
(front and back)
Target 18
180°
Endo-1,4-beta-xylanase (front and back)
180°
Xylanase inhibitor TAXI-I (front and back)
Target 20
180°
Protein methyltransferase HEMK
(front and back)
180°
Peptide chain release factor 1
(front and back)
Target 21
180°
Origin recognition complex subunit 1,
BAH domain (front and back)
180°
Regulatory protein SIR1,
ORC interacting domain (front and back)

Fig. 7. PIER prediction results for CAPRI targets from rounds 1–7 (antibody/antigen complexes excluded). The protein surface residues were classified as interface or noninterface based on their PIER value compared with a decision cutoff; they were further checked against the experimentally determined interacting residues and classified to be true positives (red), true negatives (navy), false negatives (cyan), or false positives (pink). See Table III for quantification of these results and results on antibody/antigen complexes.

TABLE III. PIER Prediction Results for CAPRI Targets From Rounds 1–7

Target	Name	TP	TN	FP	FN	Recall, %	Precision, %	Significance
T01	HPr kinase/phosphorylase	36	212	98	37	49	27	0.0037
T01	Phosphocarrier protein HPr	16	39	1	16	50	94	1.81×10^{-6}
T02	Bovine rotavirus major capsid protein V6	1	796	56	23	4	2	0.8055
T02	Fab	13	297	27	9	59	32	2.26×10^{-8}
T03	Flu hemagglutinin	0	1079	33	62	0	0	1.
T03	Fab HC63	21	308	14	17	55	60	6.38×10^{-15}
T04	Pig alpha-amylase	0	275	43	26	0	0	1.
T04	Camelid VHH domain	13	72	3	14	48	81	7.44×10^{-7}
T05	Pig alpha-amylase	0	286	43	22	0	0	1.
T05	Camelid VHH domain	22	51	21	13	63	51	0.0009
T06	Pig alpha-amylase	19	281	24	21	48	44	2.67×10^{-9}
T06	Camelid VHH Domain	17	67	4	12	59	81	2.57×10^{-8}
T07	T-cell antigen receptor, extracellular portion of beta-chain	0	158	24	20	0	0	1.
T07	Streptococcal pyrogenic exotoxin A	0	147	9	16	0	0	1.
T08	Laminin, gamma-1 chain	12	82	42	13	48	22	0.1334
T08	Nidogen G3 domain	13	140	44	12	52	23	0.0046
T09	Transcription antiterminator LicT, PTS-regulatory domain	29	78	40	21	58	42	0.0033
T10	Tick-borne encephalitis virus glycoprotein	58	138	76	59	50	43	0.0089
T12	Cohesin-2	20	75	7	9	69	74	8.11×10^{-10}
T12	Endo-1,4-beta-xylanase Y, residues 733–791	13	22	10	5	72	57	0.006
T13	Major surface antigen P30	5	177	11	21	19	31	0.0306
T13	Sag1 Fab	14	307	19	15	48	42	6.29×10^{-9}
T14	130 kda myosin-binding subunit of smooth muscle myosin phophatase	52	134	22	39	57	70	1.81×10^{-12}
T14	Serine/threonine protein phosphatase PP1-beta catalytic subunit	30	102	46	25	55	39	0.002
T15	Colicin D catalytic domain	9	63	2	19	32	82	0.0003
T15	Colicin D immunity protein	11	41	8	16	41	58	0.0201
T18	Xylanase inhibitor TAXI-I	13	223	22	20	39	37	2.26×10^{-5}
T18	Endo-1,4-beta-xylanase I	9	100	4	29	24	69	0.001
T19	Ovine prion	0	57	11	22	0	0	1.
T19	Fab	18	309	10	13	58	64	2.97×10^{-15}
T20	Protein methyltransferase HEMK	41	137	33	11	79	55	8.1×10^{-15}
T20	Peptide chain release factor 1	30	59	90	10	75	25	0.0625
T21	Origin recognition complex subunit 1, BAH domain	10	126	16	16	38	38	0.0015
T21	Regulatory protein SIR1, ORC interacting domain	14	46	35	6	70	29	0.0283

For all 34 proteins, actual numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are given, along with calculated values of recall, precision, and statistical significance as determined by Fisher's exact test.

try). Second, sometimes the biological interfaces cannot be assigned unambiguously because of their marginal size, existence of several comparable patches, etc. The problem of distinguishing true biological interfaces from the artefacts emerging as a result of crystallization conditions has received significant attention in the past years.[24,38,45,84,85] It has been suggested that the total buried area can be complemented by other distinguishing features, such as the presence of water molecules on the interface or evolutionary conservation of interface residues.

The PIER method provides an additional independent evaluation of the assigned oligomerization states without referring to crystallographic symmetry or evolutionary information. Here we demonstrate the applicability of PIER to the problem by predicting quaternary structures of two bacterial enzymes: Tyrosyl-tRNA synthetase and pyruvate, phosphate dikinase.

Aminoacyl-tRNA synthetases are a class of enzymes involved in protein synthesis. Despite functional similarity, tRNA synthetases for different amino acids are extremely diverse in sequence and structure. Being crucial for bacterial development and different from human analogues, bacterial aminoacyl-tRNA synthetases are perspective targets for antibiotics.[86] We applied the PIER predictor to the structure of Tyrosyl-tRNA synthetase from *Staphylococcus aureus*, PDB entry ***1JIK***.[87] In the

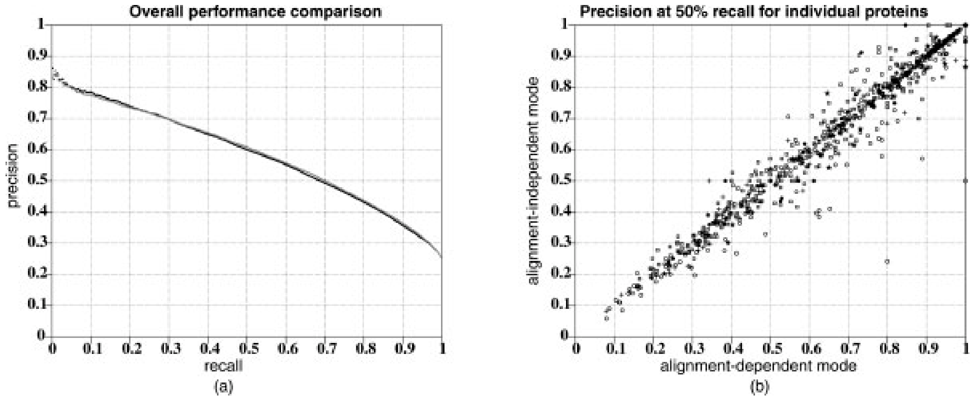
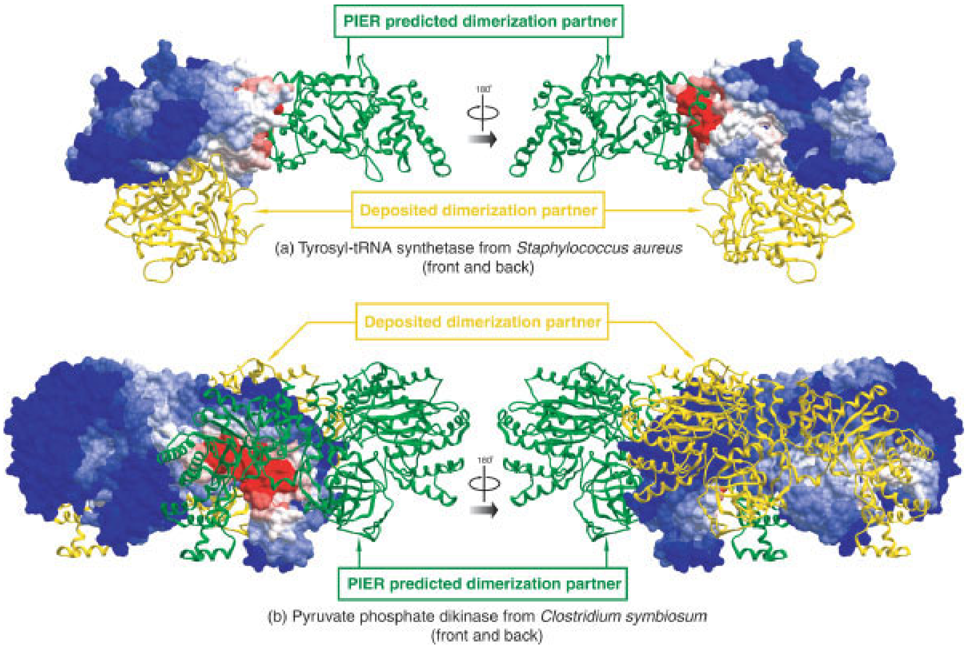


Fig. 8. (**a**) Overall recall and precision curves for the dataset of 748 protein interfaces, derived in alignment-dependent mode (grey) and alignment-independent mode (black). (**b**) Comparative plot of PIER predictor performance in the two modes for the individual proteins in the dataset: for each monomer, the PIER precision at 50% recall obtained in the alignment-dependent mode is plotted along the horizontal axis, and the same quantity obtained in the alignment-independent mode is plotted along the vertical axis. Mark types correspond to the interface types (see Table II). The plots demonstrate that with an exception of a few enzymes, using evolutionary information brings only a marginal advantage in protein interface recognition.


PIER predicted dimerization partner
180°
Deposited dimerization partner
Deposited dimerization partner
180°
PIER predicted dimerization partner

(a) Tyrosyl-tRNA synthetase from *Staphylococcus aureus* (front and back)

(b) Pyruvate phosphate dikinase from *Clostridium symbiosum* (front and back)

Fig. 9. Published (yellow) and PIER-predicted (green) dimerization geometries for (**a**) Tyrosyl-tRNA synthetase from *Staphylococcus aureus* and (**b**) pyruvate phosphate dikinase from *Clostridium symbiosum* (front and back views). In each picture, one of the monomers is displayed in skin representation and colored according to the PIER value (blue—low, red—high). Its dimerization partner as deposited in PDB is shown in yellow ribbon, while the PIER-predicted partner is in green ribbon.

region reported and annotated by the authors as the dimerization interface, no interface-like spot was detected [Fig. 9(a)]. However, the predictor indicated a well-defined interface consisting of residues G74, M77:I78, L90, L128, I131:I134, F136:L137, G141, V144:D147, M149:L150, I163:T169, I172:L173. We suggested that the true dimerization geometry for this protein is different from the one described in the PDB entry [shown in yellow ribbon in Fig. 9(a)], but instead, is mostly created by the middle portion of the chain (the interface formed with the green ribbon subunit). Indeed, at this alternative interface, a two-fold symmetry crystallographic neighbor is present in the same X-ray structure, burying 3178.2 Å of total ASA of the two molecules. Also the orthologous enzyme from *E.coli* (PDB entry ***1X8X***[88]) that shares 52.2% sequence identity and 1.70 Å RMSD with its staphylococcal counterpart when superimposed, has its dimerization interface at the location predicted by PIER, thus providing an additional evidence for the correctness of this prediction.

The PIER values also indicated that the dimer structure for pyruvate, phosphate dikinase from *Clostridium symbiosum*, deposited in PDB (PDB entries ***1KBL*** and ***1KC7***), might be incorrect. For comparison, we show the published and the predicted dimerization geometry for this protein in Figure 9(b). This prediction is supported by Herzberg et al.,[89] who mention that “the dimer interface is formed exclusively by the PEP-pyruvate binding domain.”

These examples show that PIER provides a crystal-independent prediction of biological protein assemblies and allows to correct errors and resolve ambiguities resulting from a naive interpretation of the crystallographic interfaces.

PIER Time Requirements

Our implementation of the PIER method is relatively fast. On a Pentium III 700 MHz, the interface calculation only took 10–12 seconds for an average 300-residue protein. We also have estimated the performance on several hundred proteins and derived the effect of protein size. Time dependence of the algorithm on the protein size is polynomial, $T(n) \sim n^{7/4}$, where n is the number of residues in the protein. High calculation speed makes the method applicable for automatic screening and annotation of large sets of protein structures.

CONCLUSIONS

We presented the PIER method that is able to predict protein interfaces from a single protein structure, and does not require any evolutionary signal derived from multiple-sequence alignments. The method is based on empirically derived parameters for protein surface atom groups that reflect common properties of protein interfaces. Exceptions from the PIER rules helped to identify a class of protein interfaces that rely on specialized mechanisms of complex formation. We demonstrated that incorporating the evolutionary conservation only marginally influenced the predictor performance. Fast and reliable, the PIER method may be a useful tool in automatic annotation of known and newly discovered proteins, including identification of novel protein interfaces, prediction of oligomerization states, and explaining the effects of single nucleotide polymorphisms and pathological mutations.

ACKNOWLEDGMENTS

The authors thank Andrew Bordner from Molsoft LLC, Julian Mintseris from Boston University, and James Bradford from University of Leeds, UK, for graciously sharing their datasets of protein interfaces. The authors also thank Joel Janin for valuable discussions. All molecular images were generated with the ICM software.[68]

REFERENCES

1. Szilagyi A, Grimm V, Arakaki AK, Skolnick J. Prediction of physical protein–protein interactions. Phys Biol 2005;2:S1–S16.
2. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.
3. Pagel P, Wong P, Frishman D. A domain interaction map based on phylogenetic profiling. J Mol Biol 2004;344:1331–1346.
4. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. The MIPS mammalian protein–protein interaction database. Bioinformatics 2005;21:832–834.
5. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
6. Méndez R, Leplae R, Maria LD, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. Proteins 2003;52:51–67.
7. Smith GR, Sternberg MJ. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
8. Vajda S, Camacho CJ. Protein–protein docking: is the glass half-full or half-empty? Trends Biotechnol 2004;22:110–116.
9. Wodak SJ, Méndez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14:242–249.
10. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
11. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
12. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
13. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
14. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
15. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–164.
16. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues Bioinformatics 2002;18 (Suppl 1):71–77.
17. Bordner AJ, Abagyan R. REVCOM: a robust Bayesian method for evolutionary rate estimation. Bioinformatics 2005;21:2315–2321.
18. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338:181–199.
19. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein–protein interfaces. Proteins 2005;60:353–366.
20. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–1494.
21. Koike A, Takagi T. Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 2004;17:165–173.
22. Sen TZ, Kloczkowski A, Jernigan RL, Yan C, Honavar V, Ho KM, Wang CZ, Ihm Y, Cao H, Gu X, Dobbs D. Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. BMC Bioinformatics 2004;5:205–205.
23. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
24. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313:399–416.
25. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
26. Halperin I, Wolfson H, Nussinov R. Protein–protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. Structure (Camb) 2004;12:1027–1038.

27. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
28. Valencia A, Pazos F. Prediction of protein–protein interactions from evolutionary information. Methods Biochem Anal 2003;44: 411–426.
29. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99: 5896–5901.
30. Bradford JR, Westhead DR. Asymmetric mutation rates at enzyme–inhibitor interfaces: implications for the protein–protein docking problem. Protein Sci 2003;12:2099–2103.
31. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 2004;13:190–202.
32. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein–protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
33. Rost B, Valencia A. Pitfalls of protein sequence analysis. Curr Opin Biotechnol 1996;7:457–461.
34. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
35. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
36. Mesa AdS, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. J Mol Biol 2003;326:1289–1302.
37. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
38. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein–protein interfaces. J Mol Biol 2004;336:943–955.
39. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
40. Larsen TA, Olson AJ, Goodsell DS. Morphology of protein–protein interfaces. Structure 1998;6:421–427.
41. Conte LL, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
42. MacCallum RM, Martin AC, Thornton JM. Antibody–antigen interactions: contact analysis and binding site topography. J Mol Biol 1996;262:732–745.
43. Nooren IM, Thornton JM. Structural characterisation and functional significance of transient protein–protein interactions. J Mol Biol 2003;325:991–1018.
44. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325:377–387.
45. Rodier F, Bahadur RP, Chakrabarti P, Janin J. Hydration of protein–protein interfaces. Proteins 2005;60:36–45.
46. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
47. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci 1994;3:717–729.
48. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
49. Nooren IM, Thornton JM. Diversity of protein–protein interactions. EMBO J 2003;22:3486–3492.
50. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
51. Clackson T, Wells JA. A hot spot of binding energy in a hormone–receptor interface. Science 1995;267:383–386.
52. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein–protein interactions. Proc Natl Acad Sci USA 2004;101: 11287–11292.
53. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
54. Camacho CJ, Kimura SR, DeLisi C, Vajda S. Kinetics of desolvation-mediated protein–protein binding. Biophys J 2000;78:1094–1105.
55. Kortvelyesi T, Dennis S, Silberstein M, Brown L, Vajda S. Algorithms for computational solvent mapping of proteins. Proteins 2003;51:340–351.
56. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. Optimal docking area: a new method for predicting protein–protein interaction sites. Proteins 2005;58:134–143.
57. Keil M, Exner TE, Brickmann J. Pattern recognition strategies for molecular surfaces. III. Binding site prediction with a neural network. J Comput Chem 2004;25:779–789.
58. Geladi P, Kowalski B. Partial least squares regression: a tutorial. Anal Chim Acta 1986;185:1–17.
59. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
60. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms: lysozyme and insulin. J Mol Biol 1973;79:351–357.
61. Abagyan RA, Totrov MM, Kuznetsov DA. ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 1994;15:488–506.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
63. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
64. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The universal protein resource (UniProt). Nucleic Acids Res 2005;33:154–159. Database issue.
65. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
66. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
67. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–1445.
68. Abagyan R. ICM Manual v. 3.1. 2005.
69. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins 2002;47:334–343.
70. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
71. Samanta U, Chakrabarti P. Assessing the role of tryptophan residues in the binding site. Protein Eng 2001;14:7–15.
72. Duan G, Smith VHJ, Weaver DF. Characterization of aromatic-thiol π-type hydrogen bonding and phenylalanine–cysteine side chain interactions through ab initio calculations and protein database analyses. Mol Phys 2001;99:1689–1699.
73. Pal D, Chakrabarti P. Non-hydrogen bond interactions involving the methionine sulfur atom. J Biomol Struct Dyn 2001;19:115–128.
74. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53:708–719.
75. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
76. Bourne Y, Watson MH, Hickey MJ, Holmes W, Rocque W, Reed SI, Tainer JA. Crystal structure and mutational analysis of the human CDK2 kinase complex with cell cycle-regulatory protein CksHs1. Cell 1996;84:863–874.
77. Russo AA, Jeffrey PD, Pavletich NP. Structural basis of cyclin-dependent kinase activation by phosphorylation. Nat Struct Biol 1996;3:696–700.
78. Borriello F, Krauter KS. Multiple murine α 1-protease inhibitor genes show unusual evolutionary divergence. Proc Natl Acad Sci USA 1991;88:9417–9421.
79. Creighton TE, Darby NJ. Functional evolutionary divergence of proteolytic enzymes and their inhibitors. Trends Biochem Sci 1989;14:319–324.
80. Rheaume C, Goodwin RL, Latimer JJ, Baumann H, Berger FG. Evolution of murine α 1-proteinase inhibitors: gene amplification and reactive center divergence. J Mol Evol 1994;38:121–131.
81. Goodwin RL, Baumann H, Berger FG. Patterns of divergence during evolution of α 1-proteinase inhibitors in mammals. Mol Biol Evol 1996;13:346–358.
82. Hill RE, Hastie ND. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 1987;326:96–99.
83. Méndez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 2005;60:150–169.

84. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.

85. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.

86. Schimmel P, Tao J, Hill J. Aminoacyl tRNA synthetases as targets for new anti-infectives. FASEB J 1998;12:1599–1609.

87. Qiu X, Janson CA, Smith WW, Green SM, McDevitt P, Johanson K, Carter P, Hibbs M, Lewis C, Chalker A, Fosberry A, Lalonde J, Berge J, Brown P, Houge-Frydrych CS, Jarvest RL. Crystal structure of Staphylococcus aureus tyrosyl-tRNA synthetase in complex with a class of potent and specific inhibitors. Protein Sci 2001;10:2008–2016.

88. Kobayashi T, Takimura T, Sekine R, Vincent K, Kamata K, Sakamoto K, Nishimura S, Yokoyama S. Structural snapshots of the KMSKS loop rearrangement for amino acid activation by bacterial tyrosyl-tRNA synthetase. J Mol Biol 2005;346:105–117.

89. Herzberg O, Chen CC, Kapadia G, McGuire M, Carroll LJ, Noh SJ, Dunaway-Mariano D. Swiveling-domain mechanism for enzymatic phosphotransfer between remote reaction sites. Proc Natl Acad Sci USA 1996;93:2652–2657.