How Proteins Get in Touch: Interface Prediction in the Study of Bio-molecular Complexes

Sjoerd J. de Vries and Alexandre M.J.J. Bonvin*

Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584CH, Utrecht, The Netherlands

Abstract: Protein-protein interface prediction is a booming field, with a substantial growth in the number of new methods being published the last two years. The increasing number of available three-dimensional structures of protein-protein complexes has enabled large-scale statistical analyses of protein interfaces, considering evolutionary, physicochemical and structural properties. Successful combinations of these properties have led to more accurate interface predictors in recent years.

In addition to parametric combination, machine learning algorithms have become popular. In the meantime, assessing the absolute and relative performance of interface predictors remains very difficult: This is due to differences in both the output of the various interface predictors, and in the evaluation criteria used by their respective authors.

This review provides an overview of the state of the art in the field, and discusses the performance of existing interface predictors. The focus is mainly on protein-protein interface prediction, although most issues are also valid for other kinds of interface prediction.

Keywords: protein-protein complexes, interface prediction, binding site prediction

INTRODUCTION

The aim of interface prediction is to identify *a priori* which residues of a protein can interact with another molecule such as a protein, nucleic acid, carbohydrate or small ligand. This generates testable hypotheses about the function of a protein that can be verified by site-directed mutagenesis. In addition, interface prediction can be of assistance in other bioinformatic disciplines, such as function prediction and most notably protein docking, which aims at predicting the three dimensional (3D) structure of a protein complex from the known structures of its individual components.

Most fields are driven by the discovery of new experimental data. In contrast, interface prediction uses mainly data contained in the sequence and 3D structure of proteins. Progress in interface prediction is mainly driven by the development of algorithms to interpret, process, and combine these data. On the other hand, ongoing efforts in genomics provide increasing numbers of protein sequences from a large variety of organisms; at the same time, structural genomics initiatives contribute to populating the database of known 3D structures of proteins and their complexes [1]. These, combined with advances in computer technology, allow protein interfaces to be analyzed and predicted today in a far more systematic way than what was possible in the past. While older methods could only be tested on a case-by-case basis or on a small set of similar complexes, large-scale statistical analysis and validation on non-redundant benchmarks has become the norm. Therefore, interface prediction

is a field that is rapidly developing. At the time of writing, the large majority of protein-protein prediction methods are less than three years old, and new predictors are being published at a substantial rate. Despite this (or perhaps because of this), there is a lack of consensus on how protein interfaces should be predicted, or even what constitutes a good prediction. The aim of this work is to provide a critical overview of the current state of the art, while comparing different methods based on their choice of data, algorithm and evaluation. The focus will be on protein-protein interface prediction, although many of the points discussed are also valid for other kinds of interface prediction.

THE UNDERLYING PROPERTIES OF PROTEIN INTERFACES

A large number of properties have been identified that have some predictive power for interfaces. They can be roughly divided into three groups:

- The type and properties of the residues in the amino acid sequence; these properties include desolvation, interface propensity, and hydrophobicity.

- The evolutionary conservation, which can be derived by comparing a given sequence to sequences of homologuous proteins.

- The information contained in the atomic coordinates of the structure, such as surface accessibility, tertiary structure (spatial neighbors), secondary structure, crystallographic B-factors and properties that describe the chemical composition and the shape of the protein surface.

*Address correspondence to this author at the Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584CH, Utrecht, The Netherlands; Tel: +31.(0)30.2532652; Fax: +31.(0)30.2537623; E-mail: a.m.j.j.bonvin@uu.nl

Unfortunately, there is no single property sufficient for unambiguous identification of the interface[2]. Moreover, considerable disagreement exists on which properties are actually useful. Conservation is an excellent example of a property both widely used and widely debated. First of all, there is a plethora of different methods to compute conservation from multiple sequence alignments (for an excellent review see Valdar [3]). The more sophisticated methods also take into account the evolutionary history of the sequences [4-8]. Conservation has proven its merit in prediction of functional sites [4, 5, 9], such as catalytic sites and residues that bind small ligands, but for protein-protein interfaces, its use has been questioned [10] . Recently, Mihalek *et al.* showed that conserved residues coincide with a well-defined subset of the interface, formed by the residues that become totally shielded from water upon complexation [11]. Reflecting these different findings, some interface predictors implement sophisticated forms of conservation [7, 12, 13], simple forms of conservation [14-16], or no measure of conservation at all [17, 18].

An important problem in interface prediction is the orthogonality of properties, addressed in more detail by Neuvirth *et al.* [19]. Two properties with predictive power may be combined into a better predictor, but only if they are orthogonal, i.e. not too strongly correlated. For example, correlation has been shown to exist between desolvation and hydrophobicity [18]. This may explain some reports about certain properties having little value: new properties are often added to an existing method, having little effect if the new property is not orthogonal to the ones already used.

DIFFERENT TYPES OF COMPLEXES

Protein-protein complexes can be divided into complexes for which the interacting proteins are always bound in the complex (obligate/permanent complexes), and complexes for which the interacting proteins can also exist in the unbound form (non-obligate/transient complexes). Obligate complexes can be further divided into homodimers, heterodimers and multimers; transient complexes are usually heterodimers. Transient complexes can be further divided into enzyme-inhibitor complexes, antibody-antigen complexes and other transient complexes.

It has turned out that the type of complex has considerable consequences for the prediction of the interface. Obligate interfaces are larger, flatter and better conserved than transient interfaces ([10, 17, 20] and references therein). Obligate interfaces consist primarily of side chain-side chain contacts whereas in transient interfaces the backbone plays a more prominent role [17]. Overall interface residue propensities are similar between obligate and transient interfaces: aromatic and hydrophobic residues and atoms are usually enriched in interfaces, whereas charged groups (except arginine) are usually depleted. However, these preferences are stronger in obligate complexes than in transient complexes [17]. In addition to the larger interface-to-surface ratio, this causes obligate complexes to be considerably easier to predict than transient complexes.

The relevance of predictions on obligate complexes can be questioned: by definition, only bound structures are available, from which the interface can be simply extracted instead of predicted. In contrast, predictions on unbound structures generate testable hypotheses on the binding mode of transient complexes. For assessment, interface predictions for transient complexes can be made on either bound or unbound structures, or even homology models. In theory, this may greatly affect the performance of interface prediction; in practice, interface predictors have shown to be robust for small structural differences [7, 17, 19, 21].

A recent study by Burgoyne and Jackson [22] assessed the predictive power of a number of properties for protein-ligand interfaces and different types of protein-protein interfaces. They found that protein-ligand interfaces are much easier to predict than protein-protein interfaces, with strong signals for desolvation, electrostatic potential, surface cleft size and conservation. These signals persist to some extent for the prediction of protein-protein interfaces in enzymes (which can overlap with small-ligand binding sites); desolvation and especially conservation become, however, more important and cleft size less important. In contrast, only desolvation and conservation provide meaningful signals for inhibitors and other transient complexes, with a much stronger signal present in inhibitors.

Antibodies and antigens form two extremes in the predictability of interfaces. In antibodies, the interface is nearly always formed by the Complementarity Determining Region (CDR) loops, which are not only easily recognizable from the sequence, but also contain a very strong desolvation signal [17, 22]. Antigens, on the other hand, are notoriously hard to predict, with binding sites occasionally overlapping with binding sites of other proteins, but often formed by a totally different and undistinguishable region. For these, Burgoyne and Jackson [22] found only that desolvation has some predictive power. It is our opinion, and that of several others[2, 17], that neither antibodies nor antigens should be used to evaluate the performance of interface prediction algorithms.

INTEGRATION OF PROPERTIES INTO A PREDICTION

After selecting and extracting relevant properties, an interface predictor algorithm must combine these into a single prediction. Based on the output format of the prediction, interface predictors can be divided into two groups: patch-based or residue-based predictors. In their original interface predictor, Jones and Thornton [23, 24] defined a large number of patches on the protein surface and calculated the average physicochemical properties of the residues in a patch. The patches were ranked according to the combination of the properties, and the top-scoring patch was predicted as the interface. This patch-based approach has been followed by several other interface predictors [18, 25, 26]. In contrast, the residue-based predictors, which include most methods, generate a list of residues that are predicted to be in the interface. These residues are not necessarily contiguous on the protein surface, although clustering algorithms are sometimes used to enforce contiguity [12-14, 21].

The actual combination of properties is independent from the output format. The large variety of different combination methods and their underlying principles has been recently reviewed by Zhou and Qin[2] and an elaborate discussion is

outside the scope of this review. In short, all methods can be grouped into two approaches. The first one is to design a physically relevant parametric function that combines the properties, linearly or nonlinearly [7, 16, 17, 24, 27]. Such a function requires the optimization of a usually small number of parameters. The other approach does not make any assumption about how the different properties should be combined but uses instead a machine learning algorithm to combine them in an optimal way from a large number of parameters. Popular machine learning algorithms are neural networks (NN) [15, 20, 21, 26, 28-30], support vector machines (SVM) [12, 13, 25, 31-35], and Bayesian networks [14, 18].

The parametric approach has the advantage of being more transparent and can perform well in case of a well-designed parametric function. The machine learning approach does not require human intuition and can capture subtle, unexpected trends, but is prone to overfitting if no cross-validation is performed.

A typical example of the difference between these two strategies is how to represent the information from the protein sequence. One method assigns physicochemical property values to each residue and supplies these values to the prediction algorithm. Another method represents each residue as an interface propensity, i.e. the probability of the amino acid to occur in interfaces divided by the probability to be on the surface. A third method represents the residue by a 20-element binary vector, each element coding for the presence or absence of an amino acid. The first two methods are suitable for either strategy, while the third method is exclusive to machine learning methods. Atom-level properties instead of residue-level properties can also be used.

Following their first use by Zhou and Shan [30], the large majority of the machine learning methods makes use of sequence profiles. A sequence profile consists of elements of a Position Specific Scoring Matrix (PSSM), obtained through a PSI-BLAST search of the protein sequence against homoguous sequences [36]. A sequence profile indicates thus the composition of the sequence and of its homologs, providing a measure of conservation. During the training phase, the machine learning predictor can learn that some distributions within the profile columns are more often associated with interfaces than others, providing a profile-level interface propensity score. In fact, the sequence profiles for a residue and its nearest spatial neighbors make up the large majority of the parameters for most machine learning predictors. Still, there is evidence that machine learning algorithms do not fully extract all protein sequence information from sequence profiles. Both Res *et al.* [12] and Chung *et al.* [34] found that adding explicit conservation scores to sequence profiles improved the prediction, although the opposite was found by Porollo and Meller [15]. More interestingly, recent work by Dong *et al.* [35] suggests that interface propensities are not well extracted either: although the addition of residue-level interface propensities had no effect, a dramatic improvement could be observed by the explicit addition of propensities at the profile level. We have also recently shown that interface propensities can be extracted more effectively at a pairwise level than at the residue level [37].

All these observations indicate that there is still room for improvement in the extraction and combination of predictive properties.

ASSESSMENT OF INTERFACE PREDICTION

Fundamental Definitions

The most important reason why interface predictors are so hard to compare is their differing definitions of interface and surface residues. For the definition of interface, two methods are commonly in use: one is based on the loss of surface accessibility upon binding, and the other on the distance of each residue to the partner molecule. For both, the choice of the cutoff to separate interface from non-interface can significantly influence the result. In particular, residues at the core of an interface have been found to be more hydrophobic and more conserved than interface residues in the surrounding rim [11, 38]. In addition, the method of surface definition can be critical. This is especially true for methods exploiting conservation; interface residues are more conserved than other residues, but below a certain surface accessibility this signal is lost, due to conservation of core residues for structural stability. Therefore, most methods use a surface accessibility cutoff that is either absolute (in exposed $Å^2$) or relative (as percentage of maximal accessibility) but this differs widely between methods. The use of a cutoff has the consequence that interface residues below the defined cutoff can no longer be predicted. This can be especially problematic if predictions are made from the unbound (free) form (the most relevant case), where residues that are exposed and make contacts to the partner in the bound form may be buried or missing in an unbound crystal structure. Often, the authors of interface predictors choose to remove such residues or proteins from their benchmark.

Evaluation of Performance

Another obstacle in the objective evaluation of interface predictors is the many different ways in which the authors of interface predictors assess their success rates. Confusingly, the same terms are often applied to different quantities. Here, we will try to shed some light on common descriptive statistics of interface prediction success.

"Interface" is essentially a binary (dichotomous) variable: a residue is either in the interface, or it is not. Therefore, most interface predictors are also binary: a residue is classified as either interface or non-interface. Residues predicted as interface are true positives (TP) if correctly predicted, and false positives (FP) if not. Likewise, residues predicted as non-interface are true negatives (TN) or false negatives (FN).

The following statistics can be used to assess an interface predictor's performance:

- *Accuracy*, also called two-state accuracy: $(TP+TN)/(TP + TN + FP + FN)$

- *Specificity*, also called accuracy, accuracy+, or precision: $TP/(TP+FP)$

- *Sensitivity*, also called recall, sensitivity+, or coverage: $TP/(TP+FN)$

- *Matthew's correlation coefficient*, also called phi, which is equivalent to Pearson's correlation coefficient but can be computed more directly.

For interface prediction, accuracy is a poor evaluator, due to the skewed ratio of interface and non-interface residues. This ratio varies between studies, but typically, residues defined as interface form between 10 % and 30 % of the surface. If the fraction of interface residues is near the lower end of this range, it is trivial to devise a method with 90 % accuracy, simply by predicting all residues as non-interface. This presents a problem in the training of neural networks and SVMs, which is why a fraction of non-interface residues is often removed at random from their training set. Recently, a machine learning interface predictor was developed that is unaffected by this problem [39].

Likewise, it is trivial to achieve a high degree of sensitivity by over-predicting the interface. However, this goes at the cost of specificity. Conversely, the specificity of many interface predictors can be increased by under-prediction, at the cost of sensitivity. However, it is possible to trivially enhance specificity at no cost to sensitivity by assigning a larger fraction of the residues as interface. The fraction of interface residues is equal to the specificity of a random predictor.

The use of Matthew's correlation coefficient (MCC) overcomes this artifact. However, for interface prediction, MCC usually favors over-prediction. When this is not desirable, for example if the goal is a few accurate predictions rather than complete coverage, MCC will have a low value even if the predictions are flawless.

To assess predictions for individual proteins, the hypergeometric distribution can be used [7, 14]. This distribution yields the probability that TP or more interface residues are among (TP+FP) random predictions, sampled independently and without replacement. This is the most rigorous statistical test to prove that a predictor is superior to a random predictor; however, the test becomes easier to pass in case of over-prediction or large proteins, and trivial if applied to a benchmark as a whole. The requirement of independent sampling makes it unsuitable for patch predictors.

In case of patch predictors, the percentage overlap is sometimes used as a criterion for success[24]. If the predicted patches are of similar size as the average interface, the overlap can be regarded as both the specificity and sensitivity, with the following caveats: often, patches are defined as circular whereas interfaces are not, so that the theoretical maximum overlap of the patch predictor is below 100%. Also, patches are by definition contiguous in space whereas the predictions made by a residue predictor do not need to be; therefore, a patch with an overlap of 80% is generally better than a residue list prediction of similar sensitivity/specificity, since in the former case all predicted residues will be very close to the interface, which need not be true for the latter case. Conversely, a residue list prediction of low specificity may still be largely around the interface whereas this is unlikely for a circular patch.

Methods based on parametric combination and several machine learning methods produce a continuous range of residue scores rather than a binary prediction. At the final stage of prediction, these scores are usually converted into binary predictions using a cutoff. Conversely, Bordner and Abagyan [13] converted their SVM, which is a binary classi-fier, into a continuous-score predictor by measuring for each residue the distance of the input vector to the classification hyperplane. Continuous predictors have the advantage that they can be evaluated using a Receiver Operator Characteristic (ROC) plot. On the x axis, the fraction FP/(FP+TN) is plotted, while the sensitivity (TP/(TP+FN)) is plotted on the y axis. Then, for every possible cutoff, the corresponding x and y values are computed and a curve through these points is drawn. This has the advantage that it visualizes a predictor's performance at various desired sensitivities. In addition, the predictor's performance can be easily compared to a random predictor, which always follows a straight line. Finally, the performance can be quantified into a single value using the area under curve (AUC) statistic, which gives a good indication of a predictor's performance compared to random (AUC=0.5). However, the actual specificities cannot be computed from a ROC plot unless the fraction of interface residues is known. Also, a method that aims at generating a few accurate predictions and does not care how the rest of the interface is ranked, may have a poor AUC but still fulfill its goal.

In conclusion, there is no single statistic that can adequately assess or rank interface predictors. Sensitivity and specificity are the most commonly reported values, and their relative importance depends on the goal of the prediction. However, the fraction of interface versus surface is also very important, since this defines the specificity of a random prediction. Other statistics can be useful as well but they are less commonly reported.

Current State of the Art

As illustration, we compared several methods for which a web server is available (cons-PPISP [21], PIER [17], PINUP [16], ProMate [14], SPPIDER [15] and WHISCY [7]) using recent targets from CAPRI, the critical assessment of the prediction of interactions [40]. The resulting predictions, mapped onto the structures, are shown in Fig (**1**). In addition, we collected sensitivity, specificity and fraction of interface statistics from literature for a large number of methods. Tables **1-4** report the performance of all predictors in terms of specificity and sensitivity. The tables are split between mixed (Table **2**) and transient-only benchmarks (Table **3**; these are typically dominated by obligate complexes). Methods evaluated on the benchmark of Ofran and Rost [41], which contains complexes that are predicted to be transient, are listed in both categories. The tables are also split between residue list predictors (Tables **2,3**) and patch predictors (Table **4**). The values were taken from studies performed by the authors themselves, or by authors of other methods in comparison with their own method; no attempt was made to normalize for the various definitions of interface and surface. On the other hand, the definitions of specificity and sensitivity are always as specified above, ignoring alternative measures of success provided by the authors.

The following types of interface prediction methods were not included in this study: methods based on a single property; methods that were validated only by case studies rather than benchmarks; methods designed for a specific class of proteins, such as enzymes; methods for which either sensitivity, specificity or fraction of interface residues could not
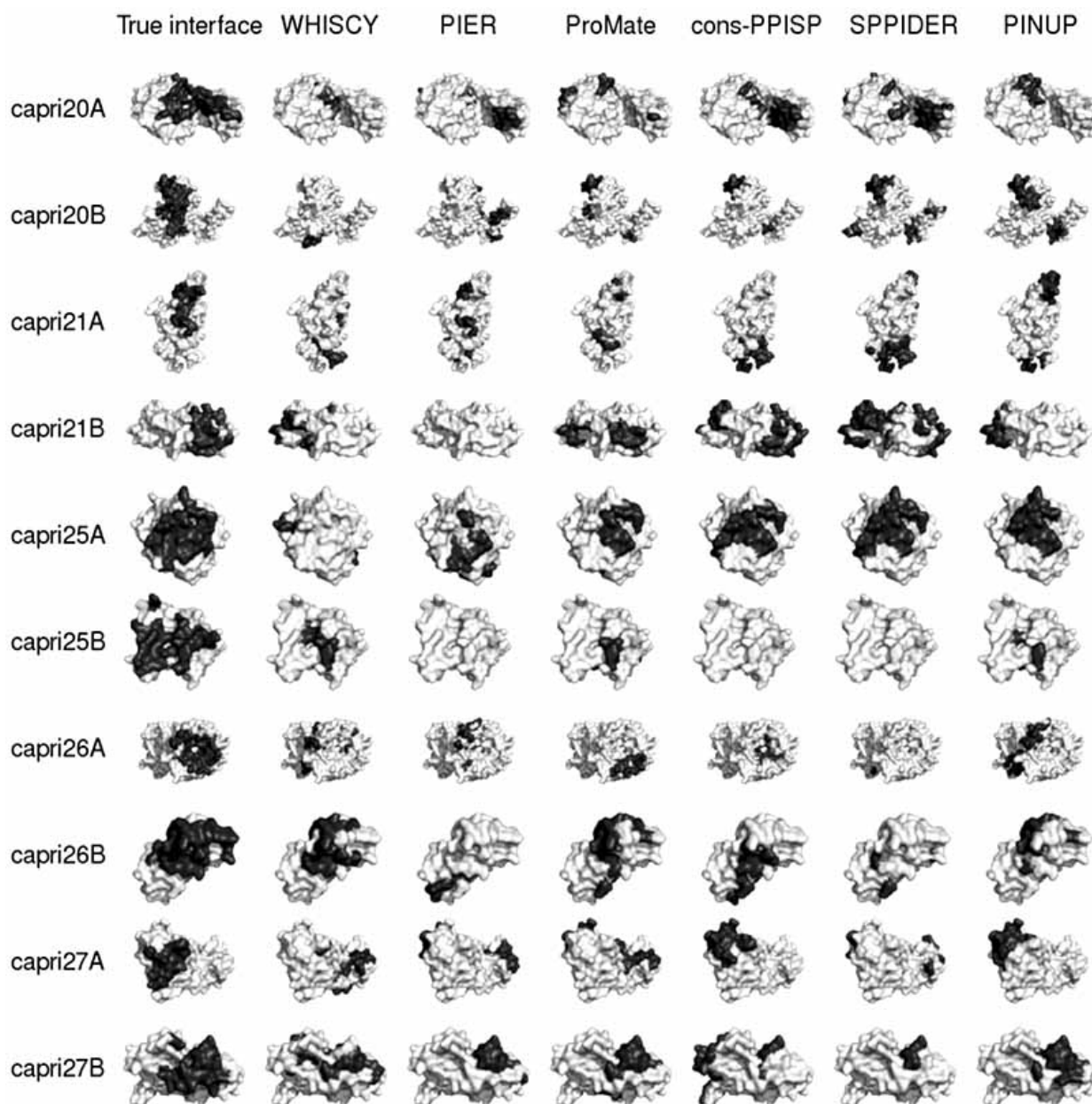
Fig. (1). Performance of various web-server interface predictors (cons-PPISP [21], PIER [17], PINUP [16], ProMate [14], SPPIDER [15] and WHISCY [7]) on recent CAPRI targets [39]. The interface predictions are mapped in dark on the protein surfaces. The first column shows the true interface (in magenta) calculated using a 5Å distance cutoff criterion.

be derived; machine learning methods that did not perform cross-validation; methods tested exclusively on benchmarks from which antibody-antigen complexes were not excluded (if indicated). To the best of our knowledge, there are 22 interface predictors fulfilling these criteria at the time of writing (July 2007) [7, 12-18, 20, 21, 24-35, 39], although new predictors appear regularly and we may have missed some papers. These methods show a large variety of prediction strategies, although some of them [31, 32] are extremely similar in approach.

For every method, the benchmark(s) of complexes on which the method was tested is also indicated, and the various benchmarks are listed in Table **1**. The number of benchmarks is smaller than the number of interface predictors

since several methods were tested on the same benchmarks (although often with different interface/surface definitions). Note that the actual set of proteins is often smaller since many methods have difficulty in predicting some proteins (for example, proteins that are very large, have missing or unusual atoms, are in an inconvenient quaternary state, or for which some properties cannot be obtained). Here it is assumed that such exclusions have no influence on the performance of the method on the benchmark as a whole. Also, no discrimination is made between predictions on bound or on unbound structures.

In addition, we tested four continuous-score interface prediction web servers (PIER, ProMate, SPPIDER and WHISCY) on the docking benchmark 2.0 [42], a non-

Table 1.   Datasets Used in Interface Prediction

Dataset	Source	Description
A	[13]	PDB, 30 % redundancy, no multimers, multiple interfaces possible
B	[12]	hand-curated, 50 heterodimers
C	[33]	PDB, 25 % redundancy
D	[28]	PDB, 30 % redundancy
E	[35]	PDB, 25 % redundancy
F	[38]	PDB, 30 % redundancy, secondary interfaces removed
G	[34]	PDB, 30% redundancy, multiple interfaces possible
H	[30]	PDB, 40% redundancy, mostly heterodimers
I	[20]	PDB, complexes predicted as transient
J	[15]	PDB, blast e=0.001 redundancy
K	[28]	PDB, 30 % redundancy, heterodimers
L	[41]	benchmark 2.0, excluding antibody-antigens
M	[14]	57 hand-curated transient heterodimers
N	[44]	benchmark 1.0, excluding antibody-antigens
O	[21]	8 transient NMR complexes
P	[45]	39 hand-curated transient dimers
Q	[24]	hand-curated
R	[25]	PDB, redundancy 20%
S	[26]	hand-curated

Specific subsets of datasets (table 2-4)
[1]: heterodimers only
[2]: homodimers only
[3]: obligate homodimers only
[4]: obligate heterodimers only
[5]: excluding homodimers and enzyme-inhibitors
[6]: enzymes only
[7]: inhibitors only
[8]: non-enzyme-inhibitor transient complexes only
[9]: 35 non-redundant enzyme-inhibitor complexes
[10]: enzyme-inhibitors only
[11]: difficult complexes (large conformational change) only
[12]: obligates only
[13]: transients only

redundant benchmark of transient complexes, at cutoffs corresponding to 30% and 50% sensitivity. Note that the authors of ProMate recently updated their method [19] after this test was carried out.

Supplementary Table **SI** contains a more complete overview of the predictors in term of properties exploited and definitions of interface and surface.

DISCUSSION

The large majority of the methods reported in the various tables are less than three years old, and most of them are actually very recent. In particular, the number of methods

that have been applied to transient complexes has grown explosively in the last years.

The availability of web servers for several methods has allowed us to perform more controlled comparisons: using web servers, we have compared the performance of PIER, ProMate, SPPIDER and WHISCY on the 2.0 protein docking benchmark, excluding antibody-antigens. According to our definition (heavy atom distance less than 5 Å from the partner protein), 14.6% of the surface corresponds to the interface (average value over the benchmark). The specificities of PIER [17], ProMate [14], SPPIDER [15] and WHISCY [7] are 36.6%, 31.5%, 28.9% and 34.3% at a sensitivity of 30 %, respectively, and 28.5%, 25.0%, 24.3% and

Table 2.   Performance of Residue list Predictors on Mixed Benchmarks

Predictor	Year	Type	Dataset	%Int	Spec	Sens	Source
Bordner and Abagyan [13]	2005	SVM	A	15	34	64	
Res et al. [12] *	2005	SVM	B	16	26.3	58.8	
+ clustering into patches			B	16	28.3	55.5	
Koike and Takagi [33] *	2004		C	20	27	28.8	
PPISP (retrained) [21]	2005	NN	D	23	55**	51	
PIER [17]	2007	Parametric	A	25	60	50	
			A, [1]***		69	61	
			A, [2]***		62	55	
Dong et al. [35]	2007	SVM	E, [3]***	27	47.4	61.7	
			E, [4]***	32.8	51.4	59	
Li et al. [38]	2007	CRF	F	27.7	43.5	62.7	
Chung et al. [34]	2006	SVM	G	28	50	67.3	
PPISP [30]	2001	NN	H	29	51	65	
Koike and Takagi [33]	2004	SVM	C	30.4	56.1	44.6	
ISIS [29] *	2007	NN	I	32**	61	20	
SPPIDER [15]	2007	NN	J***	34	63.7	60.3	
PPI-PRED [25]	2004	SVM	J***	34	50.4	27.2	[15]
Wang et al. [32]	2006	SVM	K, [5]	34.8	49.7	66.3	
Wang et al. [31]	2006	SVM	K, [5]	34.8	50.4	65	
Li et al. [27]	2006	Parametric	K, [5]	34.8	49.1	64.2	
Fariselli et al. [28]	2002	NN	K	40	72	56	
Ofran and Rost [20] *	2003	NN	I	40**	62**	0.5**	

Dataset: for letter code and footnotes see table 1; **%Int**: interface fraction, as a percentage of the total number of surface residues; **Spec:** specificity, TP/(TP+FP); **Sens:** sensitivity, TP/(TP+FN); **Source:** the work where the performance was measured; if no source is indicated, it was measured in the method's primary publication.
*: sequence-only; **: estimated; ***: average over all proteins

26.0% at a sensitivity of 50%, respectively. On the same benchmark, the authors of PINUP report a success rate of 37.5% specificity at 36.3% sensitivity, defining 15.5 % of the residues as interface [16]. Moreover, a method based on a novel property, binary profile propensities [35] , achieved a specificity of 38.3 % at 40.3 % sensitivity on the same benchmark, defining only 9.3 % of the residues as interface (Qiwen Dong, personal communication). Therefore, it seems that considerable progress has been made in the field: the four methods that we compared are similar in specificity, but the most recent (PIER) is ranked first, and it seems that two other very recent methods perform even better. Progress seems to occur mostly in the prediction of transient complexes; for mixed benchmarks (which are dominated by obligate complexes), 51%/65 % specificity/sensitivity was achieved already in 2001 [30]; recent methods have improved this to 63.7%/60.3 % using a more liberal interface

fraction (SPPIDER) [15], or 60%/50% using a more tight fraction (PIER) [17]. This is in agreement with the results of Bradford et al. [18]: by using a better integration of properties, they were able to improve upon their PPI-PRED method, making considerable improvement for transient complexes, but almost none for obligate complexes.

Some methods deserve special attention for the novel properties that they introduce. The method of Dong [35] uses explicit interface propensities at the profile level, which results in a dramatic gain in performance. The authors of SPPIDER found that, although surface accessibility has predictive power, its power can be increased by taking the difference between observed accessibility and accessibility predicted by computational means [15]. Finally, two methods should be mentioned that cannot be generally applied (and are therefore not listed in the Tables), but are nevertheless

Table 3.   Performance of Residue List Predictors on Transient Benchmarks

Predictor	Year	Type	Dataset	%Int	Spec	Sens	Source
Ofran and Rost [20] *	2003	NN	I	40**	62**	0.5**	
ISIS [29] *	2007	NN	I	32**	61	20	
Dong *et al.* [35]	2007	SVM	L***	9.3	38.3	40.9	
PINUP [16]	2006	Parametric	L***	15.5	37.5	36.3	
			L, 9		48	50	[2]
			M***	15	44.5	42.2	
PIER [17]	2007	Parametric	L	14.6	36.6	30	this work
			L	14.6	28.5	50	this work
			A, 6***		40	46	
			A, 7***		60	58	
			A, 8***		47	46	
WHISCY [7]	2006	Parametric	L	14.6	35.5	30	this work
			L	14.6	26	50	this work
			N	9.4	33	30	
ProMate [14]	2004	Bayesian	L	14.6	31.5	30	this work
			L	14.6	25	50	this work
			L, 9		38	50	[2]
			N	9.4	22	30	[7]
cons-PPISP [21]	2005	NN	L, 9		36	50	
			N		42.2	38	
			N, 10		49.9	50	
			N, 8		31	28	
			N, 11		37.3	28	
			O		38.5	47	
PPISP (retrained) [21]	2005	NN	N		36.7	35	
			N, 10		47.3	45	
			N, 8		25.2	22	
			N, 11		30.2	33	
SPPIDER [15]	2007	NN	P	23	47	43	
			L	14.6	28.9	30	this work
			L	14.6	24.3	50	this work
			L, 9		27	50	[2]
PPI-PRED [25]	2004	SVM	L, 9		27	50	[2]
Bordner and Abagyan [13]	2005	SVM	P	12	22	67	

Dataset: for letter code and footnotes see table 1; **%Int**: interface fraction, as a percentage of the total number of surface residues; **Spec:** specificity, TP/(TP+FP) (in percent); **Sens:** sensitivity, TP/(TP+FN) (in percent); **Source:** the work where the performance was measured; if no source is indicated, it was measured in the method's primary publication.

*: sequence-only; **: estimated; ***: average over all proteins

Table 4. **Performance of Patch Predictors**

Predictor	Year	Type	Dataset	%Int	Spec	Sens	Top	Top 3	Source
Mixed									
Jones and Thornton [24]	1997	Parametric	Q	15**	>=70			65	
PPI-PRED [25]	2004	SVM	R	15.4	>=50	>=20	45	76	
			R, 12		>=50	>=20	42		
			R, 4		>=50	>=20		85	
			R, 3		>=50	>=20		81	
			Q		>=50	>=20		72	
Bradford *et al.* [18]	2006	Bayesian	R	15.4	>=50	>=20	52	82	
			R, 11		>=50	>=20	52	84	
HotPatch [26]	2007	NN	S, 12	15.6**	>=33		78		
Transient									
ProMate [14] ****	2004	Bayesian	M	18	>=50		63		
			M	18	>=50	>=20	36		[25]
PPI-PRED [25]	2004	SVM	R, 13		>=50	>=20	46		
			R, 10		>=50	>=20		64	
			R, 8		>=50	>=20		67	
			M	18	>=50	>=20	36	53	
Bradford *et al.* [18]	2006	Bayesian	R, 13		>=50	>=20	53	79	
HotPatch [26]	2007	NN	S, 13	7.9**	>=33		39		

Dataset: for letter code and footnotes see table 1; **%Int**: interface fraction, as a percentage of the total number of surface residues; **Spec:** specificity, TP/(TP+FP), indicated is the criterion (in percent) for success; **Sens:** sensitivity, TP/(TP+FN), indicated is the criterion (in percent) for success; **Top:** percentage of proteins for which the specificity and sensitivity criteria are passed by the top-ranking patch; **Top 3:** percentage of proteins for which the specificity and sensitivity criteria are passed by any of the three top-ranking patches; **Source:** the work where the performance was measured; if no source is indicated, it was measured in the method's primary publication.

*: sequence-only; **: estimated; ***: average over all proteins; ****: 10 % of the proteins yielded no prediction

very promising. Hoskins *et al.* [43] developed a method to distinguish interacting from non-interacting beta strands. Even though the method cannot always make a prediction, the predictions that are made are usually very accurate. Fernandez-Recio *et al.* [44] reversed the usual concept that interface predictions can be used in docking: by docking the components of known complexes and studying the most favorable solutions, the authors were able to make accurate predictions of the interface.

Another interesting question is whether protein-protein interfaces can be predicted from sequence alone. It is very difficult to achieve high specificity using sequence alone, since non-surface residues can no longer be trivially eliminated, resulting in a low fraction of interface residues. Koike and Takagi [33] showed that sequence profiles have only limited predicted power, achieving specificity and sensitivity of around 28% on a mixed dataset. Res *et al.* [12] managed to greatly increase sensitivity, but hardly specificity, by adding an explicit conservation term. In contrast, Ofran and Rost [41] succeeded to make accurate predictions based on se-

quence alone (62%), but to the detriment of sensitivity (0.5%). However, the authors recently made tremendous improvements to the performance of their method by including predicted structural properties (surface accessibility and secondary structure) (ISIS), reaching 61% specificity at 20% sensitivity and even 90% specificity for the most confident predictions [29]. These results are all the more impressive since they were obtained on a benchmark of complexes that are predicted to be transient, although the method is yet to be applied to an established benchmark of transient complexes.

Finally, it is important to keep in mind that even predictors with similar overall performance often disagree about the interface in individual proteins [7]. This is well demonstrated in Fig. (**1**) in which interface predictions for various CAPRI targets are mapped onto the structures. It is clear from the figure that predictors do not always agree: the true interface is usually at least partly predicted by one or more predictors, but which predictors make a successful prediction differs from case to case. Moreover, some predictors seem to be better suited for some kinds of complexes than others: for

example, WHISCY performs better on enzymes than on inhibitors while PIER shows a reverse behavior. This suggests that not only properties but also interface predictors can be complementary and can be combined for improved prediction of protein interfaces. Several consensus prediction methods have already been developed [2, 7, 21]. With the advent of new and/or improved methods and of web servers, progress is likely to be made in the near future. Interface predictors are therefore expected to play an increasingly important role in structural biology, proteomic and system biology, both complementing and driving experimental studies of biomolecular interactions.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

One table containing all values used in this work is provided as supplementary material. It contains the various statistics, ordered by interface predictor, including the method of interface and surface definition and the properties used.

LIST OF ABBREVIATIONS

CRF = Conditional random field

FN = False negative

FP = False positive

NN = Neural network

SVM = Support vector machine

TN = True negative

TP = True positive

3D = Three-dimensional

REFERENCES

[1] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res.,* 28, 235-42.
[2] Zhou, H.X. and Qin, S. (2007) *Bioinformatics*, 23, 2203-9.
[3] Valdar, W.S. (2002) *Proteins*, 48, 227-41.
[4] Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) *J. Mol. Biol.,* 257, 342-58.
[5] Mihalek, I., Res, I. and Lichtarge, O. (2004) *J. Mol. Biol.,* 336, 1265-82.
[6] Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. and Ben-Tal, N. (2005) *Proteins*, 58, 610-7.
[7] de Vries, S.J., van Dijk, A.D. and Bonvin, A.M. (2006) *Proteins*, 63, 479-89.
[8] Bordner, A.J. and Abagyan, R. (2005) *Bioinformatics*, 21, 2315-21.
[9] Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E. and Lichtarge, O. (2002) *J. Mol. Biol.,* 316, 139-54.
[10] Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. and Huang, E.S. (2004) *Protein Sci.,* 13, 190-202.
[11] Mihalek, I., Res, I. and Lichtarge, O. (2007) *J. Mol. Biol.,* 369, 584-95.
[12] Res, I., Mihalek, I. and Lichtarge, O. (2005) *Bioinformatics*, 21, 2496-501.
[13] Bordner, A.J. and Abagyan, R. (2005) *Proteins*, 60, 353-66.
[14] Neuvirth, H., Raz, R. and Schreiber, G. (2004) *J. Mol. Biol.,* 338, 181-99.
[15] Porollo, A. and Meller, J. (2007) *Proteins*, 66, 630-45.
[16] Liang, S., Zhang, C., Liu, S. and Zhou, Y. (2006) *Nucleic Acids Res.,* 34, 3698-707.
[17] Kufareva, I., Budagyan, L., Raush, E., Totrov, M. and Abagyan, R. (2007) *Proteins*, 67, 400-17.
[18] Bradford, J.R., Needham, C.J., Bulpitt, A.J. and Westhead, D.R. (2006) *J. Mol. Biol.,* 362, 365-86.
[19] Neuvirth, H., Heinemann, U., Birnbaum, D., Tishby, N. and Schreiber, G. (2007) *Nucleic Acids Res.,* 35, W543-8.
[20] Ofran, Y. and Rost, B. (2003) *J. Mol. Biol.,* 325, 377-87.
[21] Chen, H. and Zhou, H.X. (2005) *Proteins*, 61, 21-35.
[22] Burgoyne, N.J. and Jackson, R.M. (2006) *Bioinformatics*, 22, 1335-42.
[23] Jones, S. and Thornton, J.M. (1997) *J. Mol. Biol.,* 272, 121-32.
[24] Jones, S. and Thornton, J.M. (1997) *J. Mol. Biol.,* 272, 133-43.
[25] Bradford, J.R. and Westhead, D.R. (2005) *Bioinformatics*, 21, 1487-94.
[26] Pettit, F.K., Bare, E., Tsai, A. and Bowie, J.U. (2007) *J. Mol. Biol.,* 369, 863-79.
[27] Li, J.J., Huang, D.S., Wang, B. and Chen, P. (2006) *Int. J. Biol. Macromol.,* 38, 241-7.
[28] Fariselli, P., Pazos, F., Valencia, A. and Casadio, R. (2002) *Eur. J. Biochem.,* 269, 1356-61.
[29] Ofran, Y. and Rost, B. (2007) *Bioinformatics*, 23, e13-6.
[30] Zhou, H.X. and Shan, Y. (2001) *Proteins*, 44, 336-43.
[31] Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M. and Lyu, M.R. (2006) *FEBS Lett.,* 580, 380-4.
[32] Wang, B., Wong, H.S. and Huang, D.S. (2006) *Protein Pept. Lett.,* 13, 999-1005.
[33] Koike, A. and Takagi, T. (2004) *Protein Eng. Des. Sel.,* 17, 165-73.
[34] Chung, J.L., Wang, W. and Bourne, P.E. (2006) *Proteins*, 62, 630-40.
[35] Dong, Q., Wang, X., Lin, L. and Guan, Y. (2007) *BMC Bioinformatics,* 8, 147.
[36] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res.,* 25, 3389-402.
[37] de Vries, S.J. and Bonvin, A.M. (2006) *Bioinformatics*, 22, 2094-8.
[38] Chakrabarti, P. and Janin, J. (2002) *Proteins*, 47, 334-43.
[39] Li, M.H., Lin, L., Wang, X.L. and Liu, T. (2007) *Bioinformatics*, 23, 597-604.
[40] Janin, J. (2007) *Proteins*, 69, 699-703.
[41] Ofran, Y. and Rost, B. (2003) *FEBS Lett.,* 544, 236-9.
[42] Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. and Weng, Z. (2005) *Proteins*, 60, 214-6.
[43] Hoskins, J., Lovell, S. and Blundell, T.L. (2006) *Protein Sci.,* 15, 1017-29.
[44] Fernandez-Recio, J., Totrov, M. and Abagyan, R. (2004) *J. Mol. Biol.,* 335, 843-65.

Table S1.   All Values in a Single Table, Ordered by Interface Predictor, Including the Method of Interface and Surface Definition and the Properties Used

Predictor	Year	Type	Surface	Interface	Properties	Dataset	%Int	Acc	Sens	% of prot	
Chung *et al.*	2006	SVM	> 15 %	< 5 A	neighbor list (12), seq profile, ASA, Bfac, conservation	274 hetero >30%, multi-interface	28	50	67.3		
PPISP (re-trained)	2005	NN	>10%	<5A	neighbor list (15), ASA, sequence profile	PDB, redundancy 30 %	23	55 (estimated)	51		
						benchmark 1.0		36.7	35		
						benchmark 1.0, E-I		47.3	45		
						benchmark 1.0, NEIT		25.2	22		
						benchmark 1.0, difficult		30.2	33		
cons-PPISP	2005	NN	>10%	<5A	neighbor list (15), ASA, sequence profile	benchmark 1.0		42.2	38		
			>10%	<5A		benchmark 1.0, E-I		49.9	50		
						benchmark 1.0, NEIT		31	28		
						benchmark 1.0, difficult		37.3	28		
						NMR transient		38.5	47		
						Enz-35 E-I unbound		36	50		
Fariselli *et al.* Bordner and Abagyan	2002 2005	NN SVM	>16%	<12A CA-CA <4A?	neighbor list (10), sequence profile neighbor list(15), REVCOM conservation neighbor list(15), REVCOM conservation patching	heterodimers, >30% redundancy PDB, redundancy 30 %, no multimers, multiple interfaces possible PDB, redundancy 30 %, no multimers, multiple interfaces possible Nooren-Thornton transient	40 15 15 12	72 34 39 22	56 64 57 67		
PPI-PRED	2004	SVM	Vertices >10% >5%	99% loss of ASA <5A 4% RSA loss	patch hydrophobicity, interface propensity, shape index&curvedness, electrostatic potential (Delphi), ASA, conservation (Valdar)	PDB, redundancy 20% PDB, redundancy 20% transient obligate E-I NEIT obligate heterodimers obligate homodimers Jones&Thornton ProMate (bound) Enz-35 E-I unbound PDB, blast e=0.001 redundancy ProMate (bound)	15.4 15.4 18 34 18	>=50 >=50 >=50 >=50 >=50 >=50 >=50 >=50 >=50 >=50 27 50.4 >=50	>=20 >=20 >=20 >=20 >=20 >=20 >=20 >=20 >=20 >=20 50 27.2 >=20	45 76 top 3 46 42 64 top 3 67 top 3 85 top 3 81 top 3 72 top 3 53 top 3 36	performed in Zhou review paper performed in SPPIDER paper
Wang *et al.*	2006	SVM	>16%	<12A CA-CA	neighbor list (10), Rate4Site	Fariselli, excluding homodimers and E-I	34.8	50.4	65		
ProMate (with patching) (without patching)	2004	Bayesian	Vertices >10%	Vertices <5A	atom propensities, chemical character, conservation, secondary structure, hydrophobicity, sequence distances	transient heterodimers, unbound transient heterodimers, unbound Enz-35 E-I unbound	18 18	>=50 >=50 38	>=20 50	63 36	another 10% no predictions performed in PPI-PRED paper performed in Zhou review paper
Li *et al.*	2006	Parametric	>16%	<12A CA-CA	neighbor list (10), Rate4Site	Fariselli	34.8	49.4	65.8		
PPISP	2001	NN	>10%	<5A	neighbor list (20), ASA, sequence profile	mainly heterodimers, redundancy 40%	29	51	65		
Res *et al.*	2005	SVM	all residues	RSA loss	sequence neighbor list, rvET, sequence profile sequence neighbor list, rvET, sequence profile, patching	Heterodimers heterodimers	16 16	26.3 28.3	58.8 55.5		

(Table S1) contd…

Predictor	Year	Type	Surface	Interface	Properties	Dataset	%Int	Acc	Sens	% of prot	
Bradford and Westhead	2006	Bayesian, patch	Vertices	99% RSA loss	patch hydrophobicity, interface propensity, shape index&curvedness, electrostatic potential (Delphi), ASA, conservation (Valdar) (eliminated)	PPI-PRED transient obligate	15.4 15.4	>=50 >=50 >=50 >=50 >=50 >=50	>=20 >=20 >=20 >=20 >=20 >=20	52 82 top 3 53 79 top 3 52 84 top 3	
Jones&Thornton	1997	Parametric	>5%		desolvation,hydrophobicity,ASA,interface propensity, planarity, protrusion	PDB	15 (estimate)	>=70		65 top 3	
PINUP	2006	Parametric	>6% >10%	1A^2 RSA loss <5A	sidechain energy (includes ASA), conservation (self-substitution), interface propensity	ProMate benchmark2.0 Enz-35 E-I unbound	15 15.5	44.5 37.5 48	42.2 36.3 50		average over all proteins average over all proteins performed in Zhou review paper
Dong *et al.*	2007	SVM	>2 A^2	1A^2 RSA loss	sequence profile (13), ASA, interface propensity sequence profile (13), ASA, binary profile interface propensity	benchmark2.0 benchmark2.0 obligate heterodimer obligate homodimer	9.3 9.3 32.8 27	35.8 38.3 51.4 47.4	37.4 40.9 59 61.7		average over all proteins average over all proteins
SPPIDER	2007	NN	>5% >10%	4% RSA loss <5A	hydrophobicity,number of contacts in neighborhood, sequence profile,entropy, conservation of charge/size/hydrophobicity	PDB, blast e=0.001 redundancy Nooren-Thornton transient Enz-35 E-I unbound	34 23	63.7 47 27	60.3 43 50		performed in Zhou review paper
Li *et al.*	2007	CRF	>15%	<5A	sequence profiles, ASA, conservation	PDB, 30 % redundancy, secondary interfaces removed	27.7	43.5	62.7		
Koike and Takagi	2004	SVM	>10% all	<5A <5A	sequence profiles(15), ASA sequence profiles (sequential, 9)	PDB, 25% redundancy PDB, 25% redundancy	30.4 20	56.1 27	44.6 28.8		
PIER	2007	Parametric	3 A^2 ASA	20 A^2 loss	surface- and neighbor-weighted atom propensities	Bordner homodimers heterodimers enzyme inhibitor NEIT	25	60 62 69 40 60 47	50 55 61 46 58 46		average per protein average per protein average per protein average per protein average per protein
HotPatch	2007	NN (patch)		<4A		homodimers transient	15,6 (estimate) 7,9 (estimate)	>=33 >=33		78 39	
ISIS	2007	NN	all	<6A	sequence profiles (sequential, 9), predicted ASA, predicted secondary structure	PDB, complexes auto-classified as transient	32 (estimate)	61	20		
WHISCY	2006	Parametric Parametric Parametric Parametric	 >15% >15% >15%	DIMPLOT DIMPLOT DIMPLOT <5A		Bordner benchmark 1.0 ProMate benchmark 1.0 benchmark 2.0 ProMate benchmark 2 PIER benchmark 2 benchmark 2.0	14 9.4 9.4 14.6 14.6 14.6 14.6	26.4 33 22 34.3 32.4 35.1 25.83	30 30 30 30 30 30 50		

Predictor	Year	Type	Surface	Interface	Properties	Dataset	%Int	Acc	Sens	% of prot	
						ProMate benchmark 2 PIER benchmark 2	14.6 14.6	24.9 27.	50 50		
Ofran&Rost	2003	NN	all residues	<6A	sequence profiles (sequential, 9)	PDB, complexes auto-classified as transient	40	62	0.5		interface fraction is an estimate
Wang *et al.*	2006	SVM	>16%	<12A CA-CA	neighbor list (10), Rate4Site	Fariselli, excluding homodimers and E-I	34.8	49.7	66.3		