

# Progress and challenges in predicting protein–protein interaction sites

*Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia and Michael L. Tress*

Submitted: 3rd November 2008; Received (in revised form): 11th March 2009

## Abstract

The identification of protein–protein interaction sites is an essential intermediate step for mutant design and the prediction of protein networks. In recent years a significant number of methods have been developed to predict these interface residues and here we review the current status of the field. Progress in this area requires a clear view of the methodology applied, the data sets used for training and testing the systems, and the evaluation procedures. We have analysed the impact of a representative set of features and algorithms and highlighted the problems inherent in generating reliable protein data sets and in the posterior analysis of the results. Although it is clear that there have been some improvements in methods for predicting interacting sites, several major bottlenecks remain. Proteins in complexes are still under-represented in the structural databases and in particular many proteins involved in transient complexes are still to be crystallized. We provide suggestions for effective feature selection, and make it clear that community standards for testing, training and performance measures are necessary for progress in the field.

**Keywords:** *protein–protein interaction; binding sites; protein complexes; prediction; machine learning*

## INTRODUCTION

Interactions between proteins play a crucial part in cellular function and form the backbone of almost all biochemical processes. While many interacting protein pairs have been identified through large-scale experiments on whole genomes [1–5], the residues involved in these interactions are generally not known and the vast majority of the interactions remain to be characterized structurally.

The experimental determination of protein–protein complexes is an expensive and time consuming process and is particularly problematic

for transient complexes, while the prediction of complexes by comparative modelling is suitable only in relatively few cases. One alternative to prediction by comparative modelling is protein–protein docking [6]. Docking procedures use surface complementarity and electrostatics to predict structural complexes, fitting together two or more known structures or reliable 3D structural models via their interacting surfaces. Although there have been successes and advances in the field [7, 8], the methods are hampered by a lack of a complete understanding of the forces involved and by the

Corresponding author: Michael Tress, CNIO, c./ Melchor Fernandez Almagro 3, Madrid 28029, Spain. Tel: +34 91 732 8000; Fax: +34 912 246 980; E-mail: mtress@cnio.es

**Iakes Ezkurdia** is in the final year of a PhD thesis studying machine learning and its application in protein–protein interaction prediction. He has been a postgraduate student at the Centro Nacional de Biotecnología and the Spanish National Cancer Research Centre (CNIO) in Madrid since 2004.

**Lisa Bartoli** is a PhD student in Physics at the University of Bologna. She has been working in the field of structural bioinformatics with the Biocomputing Group since 2005.

**Piero Fariselli** is a senior researcher of the Biocomputing Group of the University of Bologna and has developed machine learning based methods for problem solving in Structural Biology and Protein Sequence Analysis.

**Rita Casadio** is Professor of Biophysics/Bioinformatics at the University of Bologna, group leader of the Biocomputing Group and President of the International Bologna Master in Bioinformatics.

**Alfonso Valencia** is the Director of the Structural Biology and Biocomputation Programme at the Spanish National Cancer Research Centre, Director of the Spanish National Institute for Bioinformatics and senior scientist at the Centre for Astrobiology (CAB-INTA-CSIC) and editor of Bioinformatics.

**Michael Tress** is a staff scientist in the Structural Biology and Biocomputation Programme at the Spanish National Cancer Research Centre and is involved in a range of protein structure and function prediction projects.

conformational changes that often take place upon protein–protein binding.

As the number of proteins with known atomic resolution has grown more groups have addressed the issue of extracting basic features of interacting protein complexes such as shape complementarity [9–15], chemical complementarity [16, 17] and combinations of the two [18–20].

The prediction of the specific amino acid residues that play essential roles in protein–protein interactions is an important step towards deciphering the functional mechanism of proteins. Information about residues that form the interacting surface of a protein are useful for a wide range of applications such as the design of mutants for experimental verification of the interactions, the development of drugs that target protein–protein interactions, understanding the mechanism of the molecular recognition and as an aid to predicting complexes through docking and homology modelling. The prediction of the residues that mediate interactions has taken on a new urgency now that knowledge of protein–protein interactions can help to build detailed maps of metabolic pathways.

Clearly residues present in protein interfaces should be easier to predict if they have distinguishing features. Many studies have attempted to characterize the residues in protein–protein interfaces [21–25]. Earlier works were restricted by the limited subset of oligomeric proteins in the Protein Data Bank (PDB) [26], but more recent works [24, 25] have been able to sub-divide the set of oligomers depending on the strength of the interaction and on whether or not the complexes were homo-complexes.

These later studies suggested that the composition of interacting residues in the interfaces was different in each subset, for example homodimers tended to have more hydrophobic residues in their interfaces than heterodimers and strong transient complexes tended to be larger, less planar and often more hydrophobic than weak transient complexes [24]. Based on these conclusions it was suggested that interacting residues might be predicted from sequence alone [25].

Recent studies have suggested that protein surface ‘hot spots’ (residues that cause a large drop in binding energy when mutated to alanine) have physico-chemical properties that may be predictive [27, 28]. The term ‘hot spots’ has also been applied to conserved residues found in protein-binding sites. Hot spot residues can be predicted and these residues

might be used for predicting protein–protein-binding sites [29–31].

Despite these differences the take-home lesson from these many studies is that protein interfaces do not have characteristics that make them simple to predict. Many groups have developed computational methods for the prediction of interface residues based on either structure [32–46] or on sequence [47–50]. Most prediction methods use features such as the observed and predicted patterns of hydrophobicity, shape and charge of residues on the surface of the protein and resort to using ‘black box’ machine learning methods to predict the interface residues. These methods claim similar success rates, but there has been little independent evaluation of the results.

Recently Zhou and Qin [51] and de Vries and Bonvin [52] have published excellent and comprehensive reviews on the state of the art of protein–protein interaction prediction and docking. While these two reviews are complementary to this paper, the objective of our review is to assess the current limits of the performance of protein–protein interface residue prediction methods. We review the constraints imposed by the limited structural information available to researchers in the PDB, evaluate the range of features used by most prediction methods and investigate to what extent these methods can be compared and attempt to answer the main questions surrounding the prediction of protein–protein interfaces. For example how useful are sequence-based features such as predicted secondary structure, physico-chemical properties, evolutionary conservation and predicted solvent exposure in predicting protein–protein-binding sites? To what extent do features extracted from the three-dimensional structures of protein complexes help predictions?

## DATA SETS

One of the challenges in creating a method for predicting protein–protein interactions is finding a reliable data set of multimeric proteins. The data needed to train predictors must come from the known structural complexes in the PDB, but while much effort has been put into determining the three-dimensional structures of proteins, few of the structures deposited in the PDB are biological multimers. The relative paucity of data on which to train prediction methods is just one of a number

of challenges that predictors face. In part because of this lack of data there is no standard data set in use. De Vries and Bonvin [52] list 19 different testing and training sets for the 22 different predictors that they consider.

### Complex classification

Complexes can be divided into homo-complexes and hetero-complexes based on sequence identity. Homo-complex interfaces are easier to distinguish because they are typically large and hydrophobic and tend to bury large areas of non-polar residues on binding [22].

Hetero-complexes are more difficult to predict and therefore more interesting, as these include transient complexes where binding sites have to be predicted blindly, from unbound structures. For this reason many prediction groups focus on hetero-complexes [35, 43–50, 53–55]. However, since the vast majority of PDB complexes are homo-complexes of identical chains they are often included in the training and testing sets of a number of methods [36, 38, 39, 41].

Some groups have discriminated between obligate, non-obligate, transient and permanent complexes [56]. The monomers in obligate interactions do not exist as stable structures *in vivo*, while the monomers from non-obligate complexes can exist independently. The distinction between transient and permanent interactions was originally based on the lifetime of association [57].

Although no single physical property definitively distinguishes the interface for all classes of protein complexes, Ofra and Rost [25] showed that homo-obligomer, hetero-obligomer, homo-transient and hetero-transient complexes had different amino acid compositions, suggesting that the prediction of interface type on the basis of amino acid composition might be possible for subsets of complexes.

Obligatory interaction patches tend to have higher shape complementarity and to be characterized by the presence of hydrophobic residues and by tighter packing. In contrast, transient interaction patches are generally smaller and tend to be more polar, with the exception of some enzyme–inhibitor complexes. Furthermore, transient interfaces have lower geometrical complementarity and generate weaker associations [24, 25, 58].

Zhus *et al.* [59] developed an automatic classification method for distinguishing obligate, non-obligate and crystal packing interactions using a

structural model of the complex to determine the interaction types. Although this method is useful for filtering out crystal artefacts from biologically relevant interactions, it cannot be applied when the interacting partner is unknown.

Most protein–protein interaction predictors do not distinguish between sub-classes of complexes when deriving features for training. Although it may be important to take differences between sub-classes into account, most interactions do not readily fall into a definite class and it is difficult to know the complex type in advance of the prediction.

### Redundancy and bias

Training sets are limited to the complexes found in the protein databases. Unfortunately the few hetero-complexes in the PDB are also highly redundant. In addition the PDB has an inherent bias towards certain complexes such as antibody–antigen or enzyme–inhibitor complexes while others, such as membrane complexes, are underrepresented. It is not clear whether any family of complexes is easier to predict, but antigen–antibody complexes, the largest group of complexes, are much more difficult to manage because of antigens highly variable regions. Bias should be removed but retaining sufficient data to train the classifier is a difficult juggling act.

Traditionally redundancy has been tackled by using sequence identity thresholds or by clustering sequences with BLAST [60].

### Biological relevance

The paucity of hetero-complexes means that many predictors include homo-complexes in the training data. In turn this makes the biological relevance of homo-complexes important since predictors are not interested in interactions resulting from crystallization conditions. Predictors generally use the PQS server [53] to predict biological relevance and Swiss-Prot [61] annotation control, to eliminate homo-complexes formed by crystal packing. Not all prediction methods separate out biologically relevant complexes.

### Surface and interface definition

The definition of surface residues is a critical part of creating a data set. Predictor results will depend on how surface residues are defined. The most common definition of surface residues defines exposed residues as those with a relative solvent accessibility (RSA) above 16% [40, 42], in other works use a cut-off

above 10% [62, 63] and above 5% [35]. The higher the threshold, the lower the number of surface exposed residues. This sometimes improves the performance of the classifier, even if some information is lost.

Protein complex interfaces are usually defined by one of three methods. The first simply defines interfaces by distances between interacting residues. Another approach would be to define interacting residues based on differences in the solvent accessible surface area (ASA) when the monomers are separated [21, 23, 57, 64–66]. In this case the coordinates from the monomeric and independently solved unbound structures should be used, since conformational changes upon complex formation can affect the surface patches involved in protein–protein interactions. Interacting regions can also be defined by means of computational geometry, using Voronoi diagrams [67–69].

Based on our calculations and those of previous [70] works the method chosen to define the interface may not be important because the interface sizes and areas defined by the three different approaches tend to be identical or almost identical. When comparing two of the most frequent definitions, distance cut-off for contacts (1.2 nm between CA) and ASA change upon complex formation (fixing a threshold of 4% ASA for the relative change in surface exposure between an isolated chain and complex structures), we found that the interface residues in the data set overlapped in 97% of the cases without affecting predictor performance. Although the method chosen to define the interface is not critical, the threshold chosen for contacts or for accessible surface area differences is important when selecting features since the influence of certain features may differ at the edge of the interface [52].

Another place where different thresholds need to be taken into account is in comparisons between different prediction methods. The higher the proportion of residues in the interface, the easier it is to correctly predict them. This in turn conditions the choice of performance evaluation measures.

Although it would be ideal to use only unbound monomers in training predictors, the available set of bound and unbound structures is very small, so it is currently not possible to show the true effect of training with unbound structures. In addition the interaction surfaces of the bound monomers in those few complexes that have bound–unbound pairs tend to be much smaller than the average due to the bias

in the PDB. This bias suggests that it is not even fair to test predictors that have been trained on a standard set of complexes. Testing with unbound complexes is something that may become feasible as the number of known complexes in the PDB expands.

Some authors train and test their predictors using patches [22, 39, 71–74]. Patch analysis methods are limited by shape (patches are generally circular, while interfaces are irregular) and because patch size has to be estimated.

## Multiple interfaces

How to treat multiple protein–protein interfaces is an emerging problem. A number of studies have suggested that proteins may be involved in interactions with a wide range of interaction partners [1–4].

It is possible that many proteins may be transiently involved in several different protein complexes in the same pathway and the lack of experimentally validated complexes in the PDB may mean that we are underestimating the pool of interacting residues by several orders of magnitude. Many and perhaps most proteins may be involved in multiple interactions and have multiple and overlapping interaction surfaces.

We have to ask whether it is really possible to define non-interacting residues based on the currently available data. Although we must consider residues that are not observed to form part of any complex as non-interacting, it is impossible to be sure that these residues do not form part of a complex that has yet to be crystallized.

Dealing with multiple interfaces does present difficulties. If each interface is considered independently, residues may be defined as both interacting and non-interacting, introducing incorrect class assignments in the training of the classifier. Known multiple interaction interfaces should not be considered independently of each other.

While most methods consider dimer interfaces independently, a few groups have taken multiple interactions into account [35, 75].

## Disorder

One further feature to take into account when constructing data sets are disordered regions. Many authors have highlighted the importance of disordered regions in protein–protein interactions and suggested that disorder allows for more interaction partners and modification sites [76–79]. Disorder has

been shown to be important in many interactions [80–82] and there are complexes in the PDB where it is clear that regions that are likely to be disordered in the unbound state become ordered on binding and have an important role to play in stabilizing the interaction. Gunasekaran *et al.* [83] suggested that ordered monomers could be distinguished from disordered monomers on the basis of the per-residue surface and interface areas after analysing the structural characteristics of complexes.

The order–disorder state change that can occur with protein–protein binding represents an added difficulty when training and evaluating predictors with bound complexes.

## DISCRIMINATORY FEATURES IN PROTEIN–PROTEIN INTERACTION INTERFACES

Many groups have attempted to discern the distinguishing features of protein–protein interaction interfaces. For example, Jones and Thornton [57] analysed the surfaces of protein complexes in terms of patches and showed how features as solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area might be a good candidate for the prediction of protein interfaces. Indeed many methods have used the different characteristics of known protein–protein interaction sites [21, 23, 25, 57] to make predictions for residues involved in protein–protein interactions.

Discriminating features used in interface prediction can be divided into two groups, those that require knowledge of the structures of the interacting proteins (for example surface area, B-factors) and those that require no structure input [multiple sequence alignment (MSA) information, amino acid hydrophobicity]. The vast majority of known proteins do not have experimental 3D structures, so discriminatory features that help identify interface residues without the use of structures are very valuable. In the absence of known structure, predicted structure-based features such as secondary structure and accessibility can still be used to predict interface residues.

While sequence and structural features have been much used in predicting protein–protein interaction surfaces, no single feature has been inextricably linked to protein–protein interfaces. Correlations between these features and protein–protein-binding

patches are so subtle that they cannot be predicted with linear statistics alone. This may just reflect the fact that major features of the interaction such as surface area and binding stability can vary substantially in different complexes. For example, the relative contributions of electrostatic and hydrophobic forces involved in complex formation vary between complexes [84].

Many protein structures undergo conformational changes on binding another protein. This means that the use of features derived from static molecular structures may not be enough to describe potential interacting surfaces where conformational flexibility plays a key role. This additional aspect introduces a degree of complexity that is very difficult to take into account with computational methods.

## SEQUENCE-BASED FEATURES

### Residue composition and propensity

Residue interface propensities—the ratios of amino acids contributing to the interface compared to amino acid composition of the whole protein surface—was first used as a feature by Jones and Thornton [22]. They showed that residue frequencies in interfaces vary; for example the mean frequency of tryptophan was higher than that of alanine. Z-scores rather than residue frequencies have also been used and, more recently, interface propensities have been calculated at the profile level [41].

### Hydrophobicity

The hydrophobic effect is often a major contributor to binding affinity and interfaces bury a large extent of non-polar surface area [54]. One characteristic that has been noted is that many interfaces have a hydrophobic core surrounded by a ring of polar residues [85]. Some studies have suggested that interface residues can be predicted to some extent by using the hydrophobic moment and averaged hydrophobicity [86], although other studies have shown that hydrophobic moment and averaged hydrophobicity do not appear to be useful for general interface prediction [33]. It appears likely that the magnitude of the hydrophobic effect is insufficient to identify interfaces [22, 23, 87].

### Predicted structural features

Methods that predict secondary structure are highly reliable [88]. Ofra and Rost [47] have suggested

that methods that incorporate predicted structural features, such as predicted secondary structure, could improve sequence or evolutionary based methods. They found that the prediction of structural features significantly improved their sequence based predictor performance.

### Features derived from MSAs

Characteristics indicative of interaction sites can be captured by sequence profiles such as those generated from PSI-BLAST [55] multiple-sequence alignments. Various methods have been used to calculate conservation from MSAs as it is widely discussed in the review of Valdar [89]. Residue conservation at the interface is observed to be slightly higher than those of general surface residues, although it is not significantly different from those in the protein interior [90–92]. This is because many conserved residues are buried and contribute to protein folding and stability, and conservation is really only discriminatory when surface residues are compared. Despite this Guharoy and Chakrabarti [93] found that the interface core tends to be more conserved than the periphery in both obligate and non-obligate complexes, while other groups have suggested that evolutionary conservation has discriminatory power for obligate and more permanent interactions [75, 94]. Other authors [74] did not use evolutionary signals from multiple-sequence alignments and have claimed that adding evolutionary information only marginally influences the overall prediction performance.

## STRUCTURE-BASED FEATURES

### Solvent accessible surface area

Interface residues are likely to be accessible to solvent in the unbound state. An analysis of surface patches [22] found that certain classes of complex residues in the unbound interface have a higher solvent-accessible surface area than other surface patches. The solvent accessibilities of individual residues or those averaged over a surface window are used as input by most predictors. The difference between predicted RSA and the observed RSA in detached monomers (dSA) was introduced by Porollo and Meller [35] and was the most discriminative feature for their predictor. They observed that predicted RSA tended to be more consistent with the level of surface exposure in protein complexes than the unbound

structures of individual protein chains and used this as a discriminatory characteristic.

### B Factors

Interface surfaces are less flexible than the rest of the protein surface [95], suggesting that interface residues are ready for the loss of side-chain conformational entropy upon binding. This feature has been used to improve predictor performance. Chung *et al.* [42] weighted conservation scores by a normalized B-factor, thus reducing the conservation scores of the residues in the flexible regions and magnifying those in the rigid regions. Although including B-factors improves the accuracy when using structures detached from complexes, these improvements are much smaller when using independently resolved unbound structures.

### Electrostatic potential

It has been suggested that simple electrostatics could drive the formation of many complexes, while the specificity of the final orientation might be driven by more specific interactions such as hydrogen bonding, salt bridges and interaction between hydrophobic patches [96]. One important finding has been the presence of a significant population of charged and polar residues on protein–protein interfaces [23, 84, 97–99]. Electrostatic potentials have improved predictor performance in various works [42, 71] although they do not play an important role in the funnel concept of protein–protein interactions [100].

### Sensitive sub-family specific methods

These have been developed to uncover functionally important residues in proteins with known structure using information from the differential conservation at the sub-family level [101–103]. Evolutionary Trace has been used with some success to locate protein–protein binding sites [38, 104]. Wang *et al.* [40] also showed that there was little to choose between sequence profiles methods and evolutionary trace methods, but that predictions seemed to improve when the two were combined.

## FEATURE REPRESENTATION

### Sequence windows

Many predictors [36, 47–50, 86, 105] used sequence windows as input rather than single residues because protein features that are proximal in sequence are

often co-located in three-dimensional conformation too [36, 105].

### Structural windows

Structural windows are centred on the surface residue to be predicted and accompanied by a number of nearest neighbours in the patch. The number of nearest neighbours is usually calculated by empirical experiments or based on previous works. Using aggregate features with weighted neighbour averages over spatial nearest neighbours often improves the discriminatory power [32–35, 39, 40, 43, 44, 71, 72, 75].

## CLASSIFICATION AND EVALUATION

### Classification methods

Several methods have been used for the binary classification of interacting and non-interacting residues without any knowledge of binding partners. These include simple statistical functions, scoring functions, supervised learning algorithms and combinations of these algorithms in different steps. Machine learning methods are well suited to the classification of surface residues into interface and non-interface and allow better discrimination than other methods, but the results are difficult to interpret.

Most machine learning methods are based on support vector machines [33, 35, 36, 39–42, 48–50, 71, 75, 86], neural networks [32, 37, 38, 40, 44] or Bayesian networks [39, 72]. In the case of support vector machines (SVM) [106], input data are non-linearly mapped into a hyperspace and optimally separated by a hyper-plane into two classes. Neural networks (NN) combine the input data linearly into nodes then perform a non-linear transformation using hidden intermediate layers. The output data are fed to the final output node. The weights of the linear combinations that form the input to the nodes are optimized on a training data set to minimize the differences between predicted output values. Bayesian networks are probabilistic graphical models that represent joint probability distributions and inference. Conditional random fields [45] and more recently random forest [107] have also been used in the protein–protein interaction sites prediction.

### False positives and post-processing

The performance of the classifier depends to a large extent on how false positives are treated. The simplest way to avoid false positives is to filter the output by omitting isolated predictions [47]. This works because interfaces formed by few residues are rare [38]. The false positive problem can also be avoided by taking into account the distances from each surface residue to all other surface residues in the same chain. The identities of the nearest neighbours are later used for the input to a second predictor. By doing this, information about the shape of the interface is learned and isolated raw false positive predictions are eliminated. For example, Yan *et al.* [36] used the Boolean output of the SVM as input to a Bayesian network classifier that analysed the labels of the neighbours of each predicted residue. Chung *et al.* [42] filtered the output so that a predicted non-interface residue became an interface residue when the distance between its CB atom and the CB atoms of at least three predicted interface residues was lower than 6 Å. Chen and Zhou [63] improved their performance and minimized the problem of over-prediction and under-prediction by combining different neural network results from models with a range of accuracy and coverage. Bradford *et al.* [39] used a Bayesian network for the selection of the patch, improving by 6% on their previous work [71].

From a more general point of view, de Vries and Bonvin [52] suggest that methods disagreeing in patch predictions could be complementary and could be combined for performance improvements. There are currently several consensus prediction methods [108, 62].

### Classifier performance evaluation

Prediction groups have used a number of standard performance measures to assess the accuracy of their classification methods. These include, two-class classification accuracy,  $Q2$ , the percentage of correct predictions for a two-class problem, recall (sensitivity),  $R$  and the precision (specificity),  $P$ , defined as follows:

$$Q2 = \frac{TP + TN}{TP + TN + FP + FN}, R = \frac{TP}{TP + FN},$$

$$P = \frac{TP}{TP + FP}$$

A true positive prediction (TP) is when an observed interface residue is also predicted to be at the

interface. When an observed non-interacting residue is predicted not at the interface we have a true negative (TN) prediction. A false positive assignment (FP) occurs when an observed non-interacting residue is predicted to be at the interface and a false negative one (FN) accounts for an observed interface residue predicted as non-interacting. These measures may not be very informative when classes are not balanced, as is the case with protein–protein interaction residues.

When the outputs are continuous, another measure for comparing the performance of classification methods is the ROC (Receiver Operator Characteristics) curve. The Area Under Curve (AUC) can be computed to give a unique scalar value for comparison. A random classifier corresponds to the diagonal line in a ROC plot and it has an area under that line of 0.5. Thus, all classifiers are expected to show an AUC greater than 0.5, the higher the AUC the better the method performance.

A good measure of classifier performance for problems with unbalanced classes is the Matthews correlation coefficient [109], MCC, which is the correlation coefficient between two dichotomous variables. Defined as follows:

$$\text{MCC} = \frac{\text{TPTN} - \text{FPFN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

The correlation coefficient is 0 for completely random predictions, making comparison with a random baseline easy. Together with the AUC, the correlation coefficient often provides a better-balanced evaluation of predictions.

## COMPARATIVE EVALUATIONS

The majority of the methods cited in this review have used evolutionary information from sequence alignments in addition to residue properties and/or structural properties. However, it is almost impossible to perform an exhaustive comparison of the results obtained from different methods because they use different data sets for testing and training, different definitions of interface residues and different evaluation procedures. Comparison between methods is further complicated by the fact that many methods are not publicly available.

Despite the inherent difficulties in evaluating predictors, two recent reviews [51, 52] have made comparisons between their in-house prediction

method and the currently available web servers. Zhou and Qin, for example, describe the classification methods developed by available web servers and carry out a detailed and thorough assessment using a test set of 60 complexes that came from a docking benchmark set [110] and from the CAPRI docking experiment [8]. The authors were able to rank the servers based on the benchmark set. The complexes from the CAPRI experiment proved to be much more difficult to predict, in part because almost a third of them were antibody–antigen or other immune system complexes.

Unfortunately, as the authors pointed out, the complexes in the test set, or complexes that were homologous to those in the test set, were likely to have been used to varying degrees in developing the tested servers. De Vries and Bonvin has the same problem with their server comparison—some of the protein chains that made up their testing set were used in training the servers that were being tested. The paucity of complexes in the PDB that can be used for training means that predictors often use most or all of the known complexes in training their methods. The only truly fair means of testing interaction prediction servers would be using new complexes, but resolving new complexes is a slow process. Initiatives in establishing community standards and evaluation of progress such as those of ProMateus [111] should help to advance this area of research.

## FEATURE EVALUATION

We performed multiple experiments using an SVM classifier (TIPPI-SVM) developed with the LIBSVM [112] module for the R package [113]. The classifier was built solely with the aim of analysing and verifying results from previous studies and will not be built as a web server. All training and testing was carried out with the data set used by Porollo *et al.* [35] because it is available, manually curated and has been evaluated against other servers. The training set consisted of 262 hetero-complexes and 173 homo-complexes referenced as S435 and there was a non-redundant control set of 92 hetero-complexes and 57 homo-complexes (S149). The data sets are available on the SPPIDER web site (<http://spider.cchmc.org>).

## Feature importance

We tested a comprehensive set of input features in the SVM classifier. The features tested in the classifier included:

- Evolutionarily conserved residues found by Scorecons ([http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons\\_server.pl](http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl)) and the Rate4Site algorithm [114] using MSAs taken from HSSP.
- Simple predicted RSA using SABLE [115], the stand-alone version with default parameters.
- The difference in solvent accessibility (dSA), in an unbound structure between the predicted accessibility with SABLE and observed accessibility calculated with DSSP [116].
- Electrostatic potential, extracted from the STING server [117].
- Residue interface propensity, for each of the 20 amino acids based on the training set calculated as a fraction that each surface amino acid contributed to the interface compared to the fraction that each amino acid contributed to the whole protein surface [71].
- Hydrophobicity taken from the AAIndex database [118].
- Surface curvature extracted from STING server [117].

We also tested other features that are mentioned in the review, but not listed here, such as predicted secondary structure, the relative surface composition, the residue composition of spherical regions (with radius ranging from 0.6 to 1.4 nm) centred on the residue to predict and the protrusion index [119].

For comparison purposes we trained our classifier on combinations of prediction features by leaving one feature out at a time and testing changes in performance.

Feature input was arranged in three different ways to test the impact upon predictor performance. By using a 10-residue sequence window profile, by using the features of the 10 closest spatially neighbouring residues and by weighting the input features based on spatial distance (WNA) [75].

For the structure-based classifiers, a second SVM was used for false positive filtering: the input was the predicted output class from the first SVM along with the initial features. Performance improves as the

**Table 1:** The performance of the classifier with a range of feature representations

Method	Q2	R	P	MCC
SVM-Struct.	69.4	45.3	64	0.32
SVM-WNA	71.5	58.6	63.4	0.37
SVM-Seq	63.2	54.3	4.2	0.09
SVM-WNA-Seq	58.9	49.6	57.8	0.28

The impact of different combinations of input features is shown by the accuracy (Q2), recall (R), specificity (P) and Matthews correlation coefficient (MCC) for each form of representation. SVM-Struct corresponds to the classifier trained using the 10 closest spatially neighbouring residues and all input features, SVM-WNA corresponds to the classifier trained using weighted neighbour averages and all features, SVM-seq corresponds to the classifier trained using a 10-residue sequence window and sequence features only and SVM-WNA-Seq corresponds to a classifier trained using the 10 closest spatially neighbouring residues and taking into account only sequence related features.

classifier is able to filter false positives and learn plausible patch shapes.

Surface residues were defined as those with a relative solvent accessibility greater than 5%; interface residues were defined as those surface residues whose Euclidean distance to at least one residue in the partner chain was below 1.2 nm.

Features derived from sequence only have slightly higher than random class correlation scores\*, but when used along with a structural window correlation significantly improves (Table 1). This shows that the representation of neighbouring residues is crucial for the characterization of interacting residues.

In our tests, using weighted neighbour averages (WNA), rather than spatial nearest neighbours and a 10 residue structural neighbour window, significantly improves the discriminatory power of the features. Taking into account neighbouring residues and their distances improves correlation performance.

Tests on the importance of features confirmed that the most relevant input feature is dSA (Table 2), with a four-point improvement in MCC when it is included (comparing the predictor with all the features and the predictor with dSA left out). Interface residue propensity and MSA-based features from *Rate4Site* and *Scorecons* are also relevant to the overall performance. Hydrophobicity makes small improvements while the remainder of the tested features (data not shown) did not significantly

**Table 2:** The performance of the classifier with a range of features

Method (all use WNA)	Q2	R	P	MCC
All	71.5	58.6	63.4	0.37
All (-dSA)	64.1	55.3	62.4	0.33
All (-residue interface propensity)	67.3	55.4	61.2	0.34
All (-MSA based features)	66.5	57	61.1	0.34
All (-hydrophobicity)	71.1	57.6	62.8	0.36

Feature importance was measured by evaluating predictor performance after leaving out one feature at a time. Results for overall classification accuracy (Q2), recall (R), specificity (P) and Matthews correlation coefficients (MCC) are shown.

improve classifier performance when they were added as input.

We have limited the performance evaluation here to the assessment of the effect of single features, as an exhaustive combination of features is not within the scope of the evaluation. We cannot exclude the possibility that a specific feature that does not contribute to the performance on its own, may improve predictor performance when it is used in a combination with another input feature that has not been tested here.

## COMPARISON WITH ANOTHER PREDICTOR

In order to compare predictor robustness for different test sets, we compared the results from our SVM predictor with those assessed by Porollo *et al.* SPPIDER is a consensus-based classifier that combines 10 different neural networks (NNs) obtained from cross-validated training on the augmented S435 set, with k-NN selection procedure used to filter out likely mislabelled points.

Table 3 shows how our SVM classifier performs compared to SPPIDER [35] on four independent test sets of 50 randomly chosen complexes. None of the chains in the four test sets shares more than 25% sequence similarity between each other or with any of the chains in the S435 training set (which was used to train the two methods being tested).

The results of the two predictors show that different data sets result in notable variations in performance and highlight that the performance of a given classifier depends to a certain extent on the data set used for evaluation. The effect would have been even greater if the test sets had included very close homologues of the complexes in the

**Table 3:** Performance of our dummy classifier against SPPIDER with the Porollo and Meiler test set and four random test sets

Method	SI49	Dataset 1	Dataset 2	Dataset 3	Dataset 4
SPPIDER	0.42	0.25	0.24	0.26	0.29
TIPPI-SVM	0.37	0.26	0.22	0.25	0.32

TIPPI-SVM and SPPIDER performance is shown against four random data sets of 50 non-overlapping proteins. TIPPI-SVM is the dummy method developed for this review and was trained with the S435 data set using all input features and WNA representation. The publicly available SPPIDER server was developed by Porollo *et al.* [35]. Matthews correlation coefficient (MCC) is shown as the sole performance measure for reasons of clarity. MCC was chosen as the sole measure because it is the only score that takes into account the over-representation of non-interface residues in the data sets.

training set. Though the differences between the four sets of results are substantial, they are not statistically significant. These differences highlight another problem with the lack of non-redundant known complexes—testing sets that do not contain homologues of the complexes used for training have to be small.

## CONCLUSIONS FROM THE TESTING AND TRAINING OF FEATURES

The results obtained here concur largely with the study carried out by Porollo *et al.* [35] and confirm that predictors can make fairly reliable predictions for protein–protein-binding residues based on a limited set of structure-based features. The incorporation of structural information is crucial for the prediction—we found that predictions based solely on sequence features were not much better than random.

This assessment of feature importance demonstrates that although a combination of all relevant features improves the performance of a prediction method, a few features generate quantitative improvements of classification performance. How these features are represented (the structure or sequence window) is also a key point in designing methods to predict interacting residues.

The results do suggest that the prediction of interface residues may have reached a point of saturation. It seems unlikely that there are further improvements to be obtained by additional combinations of the same set of basic input features. Although predictors will have to extract some new, as yet untapped, indicators from the sequence or the

structure in order to move forward in any meaningful way, a slow but steady improvement in prediction is likely as more complexes are deposited in the PDB.

## CONCLUSIONS

An understanding of the mechanisms of protein–protein interactions and the prediction of interacting surfaces requires detailed knowledge of the three-dimensional structures of protein complexes and their unbound monomers. Unfortunately, the dearth of complex structures and the large degree of redundancy in the PDB make it impossible to generate the large data sets that would be required for a reliable training of prediction methods and mean that it is very difficult to test new methods reliably. Indeed the difficulties in working in this field are to a large extent related to the lack of available structural information. As more protein–protein complexes are resolved this will become less of a problem.

From a more general point of view, the publication of methods that have not been evaluated and assessed under consensus standards, such as in other areas of protein structure prediction [8, 120] does not aid scientific progress in protein–protein interaction prediction. This review makes evident the necessity of using common training and testing data sets and common evaluation criteria in order to assess the performance of different prediction methods.

### Key Points

- It is difficult to generate test and training sets that are sufficiently large and unbiased given the low numbers of non-redundant hetero-complexes in the PDB.
- The best predictors of protein–protein interaction sites combine many relevant sequence and structural features to improve their classifiers, though only a few features significantly add to performance on their own.
- The incorporation of distance-weighted structural information has the largest effect on predictor accuracy. Classifiers that use solely sequence features are little better than random.
- Community standards for testing and training sets and evaluation measures are necessary for fair assessment of progress in the field.

### Acknowledgements

We would like to acknowledge the constructive criticisms and inputs of the reviewers who made this a better article.

## FUNDING

Biosapiens Network of Excellence (grant number: LSHG-CT-2003-503265); 2010 Consolider—Ingenio ‘Supercomputing’ project from the Spanish Ministry of Science and Innovation.

## References

1. Gavin A, Bösch M, Krause R, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**:141–7.
2. Giot L, Bader JS, Brouwer C, *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* 2003;**302**:1727–36.
3. Li S, Armstrong CM, Bertin N, *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* 2004;**303**:540–3.
4. Uetz P, Giot L, Cagney G, *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;**403**:623–7.
5. Ho Y, Gruhler A, Heilbut A, *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002;**415**:180–3.
6. Katchalski-Katzir E, Shariv I, Eisenstein M, *et al.* Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci USA* 1992;**89**:2195–9.
7. Gray JJ, Moughon S, Wang C, *et al.* Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 2003;**331**:281–99.
8. Janin J, Henrick K, Moult J, *et al.* CAPRI: a critical assessment of predicted interactions. *Proteins* 2003;**52**:2–9.
9. Greer J, Bush BL. Macromolecular shape and surface maps by solvent exclusion. *Proc Natl Acad Sci USA* 1978;**75**:303–7.
10. Wodak SJ, Janin J. Analytical approximation to the accessible surface area of proteins. *Proc Natl Acad Sci USA* 1980;**77**:1736–40.
11. Kuntz ID, Blaney JM, Oatley SJ, *et al.* A geometric approach to macromolecule–ligand interactions. *J Mol Biol* 1982;**161**:269–88.
12. Lee RH, Rose GD. Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers* 1985;**24**:1613–27.
13. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983;**221**:709–13.
14. Jiang F, Kim S. ‘Soft docking’: matching of molecular surface cubes. *J Mol Biol* 1991;**219**:79–102.
15. Helmer–Citterich M, Tramontano A. PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J Mol Biol* 1994;**235**:1021–31.
16. Salemme FR. An hypothetical structure for an intermolecular electron transfer complex of cytochromes c and b5. *J Mol Biol* 1976;**102**:563–8.
17. Warwicker J. Investigating protein–protein interaction surfaces using a reduced stereochemical and electrostatic model. *J Mol Biol* 1989;**206**:381–95.
18. Walls PH, Sternberg MJE. New algorithm to model protein–protein recognition based on surface

- complementarity: applications to antibody-antigen docking. *J Mol Biol* 1992;**228**:277–97.
19. Shoichet BK, Kuntz ID. Matching chemistry and shape in molecular docking. *Protein Eng* 1993;**6**:723–32.
  20. Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins* 1994;**20**:320–9.
  21. Chothia C, Janin J. Principles of protein-protein recognition. *Nature* 1975;**256**:705–8.
  22. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* 1997;**272**:133–43.
  23. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol* 1999;**285**:2177–98.
  24. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* 2003;**325**:991–1018.
  25. Ofra Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol* 2003;**325**:377–87.
  26. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucl Acids Res* 2000;**28**:235–42.
  27. Moreira IS, Fernandes PA, Ramos MJ. Hot spots - A review of the protein-protein interface determinant amino-acid residues. *Prot Struct Func Bioinform* 2007;**68**:803–12.
  28. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* 2002;**12**:14–20.
  29. Hu Z, Ma B, Wolfson H, et al. Conservation of polar residues as hot spots at protein interfaces. *Prot Struct Func Genet* 2000;**39**:331–42.
  30. Ma B, Elkayam T, Wolfson H, et al. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci USA* 2003;**100**:5772–7.
  31. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics* 2006;**22**:1335–42.
  32. Fariselli P, Pazos F, Valencia A, et al. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* 2002;**269**:1356–61.
  33. Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel* 2004;**17**:165–73.
  34. Murakami Y, Jones S. SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics* 2006;**22**:1794–5.
  35. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. *Proteins* 2007;**66**:630–45.
  36. Yan C, Dobbs D, Honavar V. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics* 2004;**20**(Suppl 1):371–8.
  37. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 2001;**44**:336–43.
  38. Ofra Y, Rost B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett* 2003;**544**:236–9.
  39. Bradford JR, Needham CJ, Bulpitt AJ, et al. Insights into protein-protein interfaces using a Bayesian network prediction method. *J Mol Biol* 2006;**362**:365–86.
  40. Wang B, Chen P, Huang D, et al. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett* 2006;**580**:380–4.
  41. Dong Q, Wang X, Lin L, et al. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinform* 2007;**8**:147.
  42. Chung J, Wang W, Bourne PE. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* 2006;**62**:630–40.
  43. Liang S, Zhang C, Liu S, et al. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res* 2006;**34**:3698–707.
  44. Chen H, Zhou H. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 2005;**61**:21–35.
  45. Li M, Lin L, Wang X, et al. Protein protein interaction site prediction based on conditional random fields. *Bioinformatics* 2007;**23**:597–604.
  46. Li J, Huang D, Wang B, et al. Identifying protein-protein interfacial residues in heterocomplexes using residue conservation scores. *Int J Biol Macromol* 2006;**38**:241–7.
  47. Ofra Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics* 2007;**23**:13–6.
  48. Bock JR, Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001;**17**:455–60.
  49. Friedrich T, Pils B, Dandekar T, et al. Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics* 2006;**22**:2851–7.
  50. Res I, Mihalek I, Lichtarge O. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics* 2005;**21**:2496–501.
  51. Zhou H, Qin S. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* 2007;**23**:2203–9.
  52. de Vries SJ, Bonvin AMJJ. How proteins get in touch: interaction prediction in the study of biomolecular complexes. *Curr Protein Pept Sci* 2008;**9**:394–406.
  53. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci* 1998;**23**:358–61.
  54. Glaser F, Steinberg DM, Vakser IA, et al. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* 2001;**43**:89–102.
  55. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
  56. Nooren IM, Thornton JM. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. *EMBO J* 2003;**22**:3486–92.
  57. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 1996;**93**:13–20.
  58. Ma B, Elkayam T, Wolfson H, et al. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci USA* 2003;**100**.
  59. Zhu H, Domingues FS, Sommer I, et al. NOXclass: prediction of protein-protein interaction types. *BMC Bioinform* 2006;**7**:27.
  60. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

61. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucl Acids Res* 2007;**35**:193–7.
62. Qin S, Zhou H. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* 2007;**23**:3386–7.
63. Chen H, Zhou H. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Prot Struct Func Bioinform* 2005;**61**:21–35.
64. Janin J, Chothia C. The structure of protein-protein recognition sites. *J Biol Chem* 1990;**265**:16027–30.
65. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol* 1995;**63**:31–65.
66. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins* 2002;**47**:334–43.
67. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol* 1974;**82**:1–14.
68. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. *Structure* 1994;**2**:641–9.
69. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* 1996;**264**:121–36.
70. Gong S, Park C, Choi H, *et al*. A protein domain interaction interface database: InterPare. *BMC Bioinform* 2005;**6**:207.
71. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 2005;**21**:1487–94.
72. 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**:181–99.
73. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Sci* 2006;**15**:1017–29.
74. Kufareva I, Budagyan L, Raush E, *et al*. PIER: protein interface recognition for structural proteomics. *Prot Struct Func Bioinform* 2007;**67**:400–17.
75. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. *Proteins* 2005;**60**:353–66.
76. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 2008;**33**:2–8.
77. Dunker AK, Cortese MS, Romero P, *et al*. Flexible nets. *FEBS J* 2005;**272**:5129–48.
78. Oldfield CJ, Meng J, Yang JY, *et al*. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genom* 2008;**9**:S1.
79. Ekman D, Light S, Björklund ÅK, *et al*. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol* 2006;**7**:R45.
80. Cheng Y, Oldfield CJ, Meng J, *et al*. Mining  $\alpha$ -helix-forming molecular recognition features ( $\alpha$ -MoRFs) with cross species sequence alignments. *Biochemistry* 2007;**46**:13468–77.
81. Oldfield CJ, Cheng Y, Cortese MS, *et al*. Coupled folding and binding with  $[\alpha]$ -helix-forming molecular recognition elements. *Biochemistry* 2005;**44**:12454–70.
82. Bourhis J, Johansson K, Receveur-Brechot V, *et al*. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 2004;**99**:157–67.
83. Gunasekaran K, Tsai C, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* 2004;**341**:1327–41.
84. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol* 2000;**10**:153–9.
85. Larsen TA, Olson AJ, Goodsell DS. Morphology of protein-protein interfaces. *Structure* 1998;**6**:421–7.
86. Gallet X, Charlotiaux B, Thomas A, *et al*. A fast method to predict protein interaction sites from sequences. *J Mol Biol* 2000;**302**:917–26.
87. Korn AP, Burnett RM. Distribution and complementarity of hydrophathy in multisubunit proteins. *Proteins* 1991;**9**:37–55.
88. Rost B. Review: protein secondary structure prediction continues to rise. *J Struct Biol* 00;**134**:204–18.
89. Valdar WSJ. Scoring residue conservation. *Proteins* 2002;**48**:227–41.
90. Grishin NV, Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* 1994;**3**:2455–8.
91. Caffrey DR, Somaroo S, Hughes JD, *et al*. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 2004;**13**.
92. Bradford JR, Westhead DR. Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci* 2003;**12**:2099–103.
93. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. *PNAS* 2005;**102**:15447–52.
94. Valdar WSJ, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Prot Struct Func Genet* 2001;**42**:108–24.
95. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein interfaces. *Protein Sci* 2002;**11**:2860–70.
96. Gabdoulline RR, Wade RC. On the protein-protein diffusional encounter complex. *J Mol Recog* 1999;**12**:226–34.
97. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. *J Mol Biol* 1993;**234**:946–50.
98. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. *J Mol Biol* 1997;**268**:570–84.
99. Xu D, Lin SL, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J Mol Biol* 1997;**265**:68–84.
100. Schlosshauer M, Baker D. Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Sci* 2004;**13**:1660–9.
101. Aloy P, Querol E, Aviles FX, *et al*. 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.

102. Madabushi S, Yao H, Marsh M, *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* 2002;**316**:139–54.
103. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol* 1995;**2**:171–8.
104. Pazos F, Helmer-Citterich M, Ausiello G, *et al.* Correlated mutations contain information about protein–protein interaction. *J Mol Biol* 1997;**271**:511–23.
105. Levitt M, Chothia C. Structural patterns in globular proteins. *Nature* 1976;**261**:552–8.
106. Vapnik V. *Statistical Learning Theory*. John Wiley, New York, 1998.
107. Šikić M, Tomić S, Vlahoviček K. Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput Biol* 2009;**5**:e1000278.
108. Vries SJD, Dijk ADV, Bonvin AM. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Prot Struct Func Bioinform* 2006;**63**:479–89.
109. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;**405**:442–51.
110. Mintseris J, Wiehe K, Pierce B, *et al.* Protein–protein docking benchmark 2.0: an update. *Proteins* 2005;**60**:214–6.
111. Neuvirth H, Heinemann U, Birnbaum D, *et al.* ProMateus—an open research approach to protein-binding sites analysis. *Nucl Acids Res* 2007;**35**:W543–8.
112. Chang C, Lin C. *LIBSVM: A Library for Support Vector Machines*. 2001. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9020>
113. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing in Vienna, 2007.
114. Mayrose I, Graur D, Ben-Tal N, *et al.* Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol Biol Evol* 2004;**21**:1781–91.
115. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Prot Struct Func Bioinform* 2004;**56**:753–67.
116. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**:2577–637.
117. Neshich G, Borro LC, Higa RH, *et al.* The diamond STING server. *Nucl Acids Res* 2005;**33**:W29–35.
118. Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 1999;**27**.
119. Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics* 2002;**18**:980–4.
120. Moul J, Fidelis K, Kryshtafovych A, *et al.* Critical assessment of methods of protein structure prediction - Round VII. *Prot. Struct Func Bioinform* 2007;**69**:3–9.