PROTEINS: Structure, Function, and Bioinformatics 66:630–645 (2007)

Prediction-Based Fingerprints of Protein–Protein Interactions

Aleksey Porollo[1] **and Jarosław Meller**[1,2]*
[1]*Division of Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, Ohio 45229*
[2]*Department of Informatics, Nicholas Copernicus University, 87-100 Toruń, Poland*

ABSTRACT **The recognition of protein interaction sites is an important intermediate step toward identification of functionally relevant residues and understanding protein function, facilitating experimental efforts in that regard. Toward that goal, the authors propose a novel representation for the recognition of protein–protein interaction sites that integrates enhanced relative solvent accessibility (RSA) predictions with high resolution structural data. An observation that RSA predictions are biased toward the level of surface exposure consistent with protein complexes led the authors to investigate the difference between the predicted and actual (i.e., observed in an unbound structure) RSA of an amino acid residue as a fingerprint of interaction sites. The authors demonstrate that RSA prediction-based fingerprints of protein interactions significantly improve the discrimination between interacting and noninteracting sites, compared with evolutionary conservation, physicochemical characteristics, structure-derived and other features considered before. On the basis of these observations, the authors developed a new method for the prediction of protein–protein interaction sites, using machine learning approaches to combine the most informative features into the final predictor. For training and validation, the authors used several large sets of protein complexes and derived from them nonredundant representative chains, with interaction sites mapped from multiple complexes. Alternative machine learning techniques are used, including Support Vector Machines and Neural Networks, so as to evaluate the relative effects of the choice of a representation and a specific learning algorithm. The effects of induced fit and uncertainty of the negative (noninteracting) class assignment are also evaluated. Several representative methods from the literature are reimplemented to enable direct comparison of the results. Using rigorous validation protocols, the authors estimated that the new method yields the overall classification accuracy of about 74% and Matthews correlation coefficients of 0.42, as opposed to up to 70% classification accuracy and up to 0.3 Matthews correlation coefficient for methods that do not utilize RSA prediction-based fingerprints. The new method is available at http://sppider.cchmc.org. Proteins 2007;66:630–645.** © 2006 Wiley-Liss, Inc.

Key words: protein–protein interactions; interaction sites; relative solvent accessibility; machine learning; protein complexes; SPPIDER

INTRODUCTION

Proteins perform their function by interacting with other molecules, such as small ligands, lipids, nucleic acids, and other proteins. Therefore, understanding protein interactions is pivotal for elucidating their function, and for developing explanatory and predictive models of biological systems. Stimulated by the importance of the problem, computational studies on protein–protein interactions encompass a wide array of methods. Examples of such methods range from the prediction of protein interactions based on the analysis of evolutionary relatedness[1] or protein pathways and networks[2] to multimeric threading[3] and protein binding site prediction using docking methods.[4] These latter approaches rely on high resolution structural data to identify protein–protein binding sites and interaction partners, and to model protein–protein complexes.[5,6]

Progress in structural genomics provides an opportunity to further advance the field. For example, analysis of protein complexes provides detailed information regarding amino acid propensities to interaction interfaces. In particular, it has been observed that interacting sites are largely hydrophobic,[7,8] with hot spots consisting of (often conserved) polar residues.[9–11] In addition, amino acid biases and various physical criteria were used to identify distinct types of complexes, including transient versus obligatory and enzyme-binding versus other complexes.[12,13] These statistical biases and structural clues are also being used to develop methods for the prediction of protein–protein interaction (or binding) sites, even when no information about binding partners is available.

Grant sponsors: Computational Medicine Center and Cincinnati Children's Hospital Research Foundation; Grant sponsor: NIH; Grant numbers: AI055338, R01 AR050688, and 5R01GM067823-02.

*Correspondence to: Jarosław Meller, Division of Biomedical Informatics, Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229. E-mail: jmeller@cchmc.org

Received 14 April 2006; Revised 26 July 2006; Accepted 5 September 2006

Published online 6 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21248

The recognition of protein–protein interaction sites can be used to identify functionally important amino acid residues, facilitate experimental efforts to catalog protein interactions, enhance computational docking studies and drug design, as well as enable functional annotation for the growing number of structurally resolved proteins of unknown function.[14,15] In general, the problem of recognition of protein–protein interaction sites (or protein–protein interface recognition) can be cast as a classification problem, that is, each amino acid residue is assigned to one of two classes: interacting (interfacial) or noninteracting (noninterfacial) residues. Consequently, the problem may be solved using statistical and machine learning techniques, such as Neural Networks (NNs)[16–19] or Support Vector Machines (SVMs).[20–23]

From the point of view of the representation (feature space) used to capture characteristic signatures (or fingerprints) of interaction interfaces, one may distinguish two main groups of approaches. The first group of methods attempts to predict interaction sites using just sequence information,[19,20,24] whereas the second group takes available structural information into account as well.[17,25–27] In the latter case, the problem typically involves the identification of specific patches on the surface of an unbound protein structure with residues that are either evolutionarily conserved or have a propensity for interaction interfaces.[7,17,28,29] While having an advantage of being relatively insensitive to structural details, methods that simply map sequence conservation, as encoded by multiple alignments (MA), onto the known structure of an individual protein chain[17,30,31] have recently been shown to achieve rather limited accuracies.[32,33]

On the other hand, structural information derived from a resolved structure of an unbound protein allows one to identify residues that are in contact in 3D and to define potential interacting patches on the surface of a protein. Geometric characteristics and the topology of these potential interacting patches can be taken into account, improving the accuracy of predictions.[23,25,26] Furthermore, structural conservation was found to correlate with propensity to interaction interfaces.[27,34] In this work, we describe a novel approach to the recognition of protein–protein interaction sites in the case when the structure of an isolated protein chain (i.e., unbound structure) is known.

Our general approach builds on recently developed accurate methods for relative solvent accessibility (RSA) prediction.[35,36] These methods use a relatively short sliding window to represent an amino acid and its environment. While amino acid residues may become "buried" because of long-range contacts within the same chain, they might as well be in contact with other chains. We therefore hypothesize that the RSA prediction from short sequence windows should lend itself to the prediction of intermolecular interactions as well. We indeed observe that RSA predictions tend to be consistent with the level of surface exposure in protein complexes, rather than unbound structures of individual protein chains. On the basis of that observation, we propose novel fingerprints of interaction sites that indicate their presence by RSA prediction "errors," that is, the difference between the predicted and observed (in an unbound structure) surface exposure of an amino acid residue.

We use several machine learning approaches, including SVM, NN, and Linear Discriminant Analysis (LDA), to develop and assess a number of classifiers that combine these novel fingerprints with other information derived from sequence and structure. For training and validation, we use several sets of nonredundant protein complexes as well as representative chains derived from these complexes. Rather than treating each interaction interface independently, all known interaction sites are mapped to representative chains from multiple complexes that involve homologs of representative chains. We also evaluate the performance on a set of unbound structures that were resolved independently, providing a more rigorous validation and further assessment of the effects of induced fit (largely neglected in some recent studies that rely exclusively on the coordinates of a single chain obtained from the corresponding complex[26,27]).

We show that the new RSA prediction-based fingerprints yield significantly improved performance, compared with other signatures of protein binding sites, including evolutionary conservation, physicochemical properties, and other structure-based features. Furthermore, we show that among real valued RSA prediction methods assessed in this work, including PHDacc,[37] RVPNet,[38] PROF,[39] and SABLE,[35] the latter provides predictions that are most consistent with the RSA observed in protein complexes, yielding the best discrimination between interacting and noninteracting sites. Finally, we suggest how these systematic biases in RSA prediction may be enhanced by augmenting training sets for RSA prediction methods with data derived from protein complexes.

MATERIALS AND METHODS

Training and Control Sets

All protein complexes used here for training and testing purposes have been derived from the Protein Data Bank (PDB).[40] An initial set of 1695 representative protein complexes was obtained using the PDB as of March 2003 and the following criteria: (i) PDB entry must contain at least two chains; (ii) each chain should be at least 30 residues long (thus excluding complexes with short peptides at this stage); (iii) complexes containing either DNA or RNA sequences were excluded; (iv) each complex should contain at least one nonredundant chain with sequence identity less than 50% to any other chain within the set. These nonredundant chains and associated with them complexes will define, after further filtering and analysis, a set of representative protein chains and their unbound structures with mapped binding sites to be used for the development and assessment of methods considered here.

The initial set of complexes was processed as follows. First, the Protein Quaternary Structure (PQS) server[41] was used to filter out complexes that might contain small

interfaces resulting from crystal packing. The PQS server discriminates crystal packing from the functional protein–protein interaction using primarily the size of solvent exposed area buried during association (with the cutoff of 400 Å^2 per chain), as well as the number of residues buried at the interface, the number of salt and disulphide bridges at the interface, and approximate solvation energy difference upon complex formation.[41] Next, NMR structures, theoretical models, structures identified by PQS as viral units, structures with missing side chain coordinates, and proteins predicted using the MINNOU server[42] to have transmembrane domains were also excluded from the consideration, resulting in a subset of 891 "unproblematic" complexes.

Finally, the BLAST program[43] was used to compute pairwise sequence alignments between all pairs of chains and further exclude (using the E-value of 0.001) redundant chains within and between protein complexes. In addition, sequences redundant with respect to those included in the set developed before for the training of the SABLE method for RSA prediction[35] were also excluded, since SABLE is used to derive new fingerprints of interaction sites. For a final filtering step we used UniProt[44] annotations to exclude complexes that may not represent functional interactions. We would like to comment that while redundant sequences are excluded from considerations when deriving sets of representative protein chains, the corresponding complexes may still be used to identify pairs of interacting chains and to map interaction sites to these representative chains. This step is discussed in details in Definition of an Interaction Site.

The resulting set of nonredundant protein chains, which will be referred to as S435 throughout the paper, consists of 435 protein chains (262 from heterocomplexes and 173 from homocomplexes, referred to as S262 and S173, respectively) and a total of 69509 surface residues, using the threshold of 5% RSA to define exposed residues. For comparison with some methods from the literature, we also used an alternative threshold of 16% RSA, resulting in a decreased number of surface exposed residues (see also Results). We next used the same protocol to derive an independent control set of representative complexes from structures submitted to the PDB server between March 2003 and September 2004. After applying internal redundancy checks, as described above, chains redundant with respect to the S435 set were additionally removed, so as to enable further validation of results for predictors trained using the S435 set. As a result, a set of 149 representative chains (92 from heterocomplexes and 57 from homocomplexes, respectively) and a total of 19,977 surface exposed residues was obtained. This set will be referred to as S149. These data sets, containing both PDB entries with protein complexes and derived from them representative chains, are available from http://sppider.cchmc.org.

For S435 and S149 sets of representative chains, the coordinates of unbound structures are derived from the corresponding complexes by ignoring other chains. Although an additional mapping of interaction sites from alternative complexes involving a representative chain is performed (see Definition of an Interaction Site), the structures used for training may not represent unbound conformations. Therefore, to evaluate the effects of induced fit we also identified a subset of 21 nonredundant protein chains from the S149 set, for which both monomeric (i.e., truly unbound) structures as well as multiple complexes containing their homologs are known. This set will be referred to as S21, if unbound structures are derived from the representative complexes included in S149, or S21a, if truly unbound structures derived from the corresponding monomeric PDB entries are used.

For comparison with methods from the literature, we also used two sets of complexes and interaction interfaces developed before by Fariselli and colleagues[17] and Nooren and Thornton.[45] These sets consist of 226 and 86 chains (representing 113 and 43 interaction interfaces defined by pairs of interacting chains), respectively, and will be denoted as F226 and NT86. In addition, a nonredundant subset of 59 chains was derived from the original F226 set by applying the same sequence redundancy check and the default BLAST E-value. This set will be denoted as F59. Finally, we also used a set of Critical Assessment of PRediction of Interactions (CAPRI) targets to assess the new method and compare it with that of PPI-Pred predictions.[46]

RSA and Its Prediction

The concept of RSA plays an important role in the subsequent definition of an interaction site and novel fingerprints of protein interactions considered here. RSA of the i-th amino acid residue, RSA_i, is defined as the ratio of the solvent exposed surface area of that residue observed in a given structure, SA_i, and some maximum value of the solvent exposed surface area for this kind of amino acid, MSA_i:

$$\mathrm{RSA}_i = \frac{\mathrm{SA}_i}{\mathrm{MSA}_i} 100\%$$

Hence, RSA adopts values between 0 and 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue.[37] Unless specified otherwise (which may be relevant for comparison with RSA prediction methods from the literature), the maximum exposed surface areas are taken from,[47] and correspond to those observed in an extended conformation of a tripeptide, with the residue of interest as the central residue. The DSSP program[48] is used here to compute exposed surface areas. Also, unless specified otherwise, we define surface exposed residues as those that have RSA of 5% and more.

The level of solvent exposure is weakly conserved in families of homologous structures, especially for exposed residues.[35,37] Thus, contrary to the prediction of secondary structures, the highly variable real valued RSA does not support the notion of clearly defined distinct classes of residues and suggests that a regression-based approach is appropriate for this problem. We have recently developed

several real valued RSA prediction methods, using linear Support Vector Regression and NNs-based nonlinear regression models.[35,36] In rigorous tests, following an EVA-like methodology[49] for evaluation of the accuracy of secondary structure prediction methods, the new methods achieved significantly higher accuracy than previous methods from the literature, with mean absolute errors (MAEs) between 15.3 and 15.8% RSA and correlation coefficients between 0.64 and 0.67 on different control sets.[35] In two state projections (e.g., using 25% RSA as a threshold between buried and exposed residues), regression-based methods outperformed current state-of-the-art classification-based approaches.[35,50] These accurate real valued RSA predictions will be further put to the test in this work in the context of protein interactions.

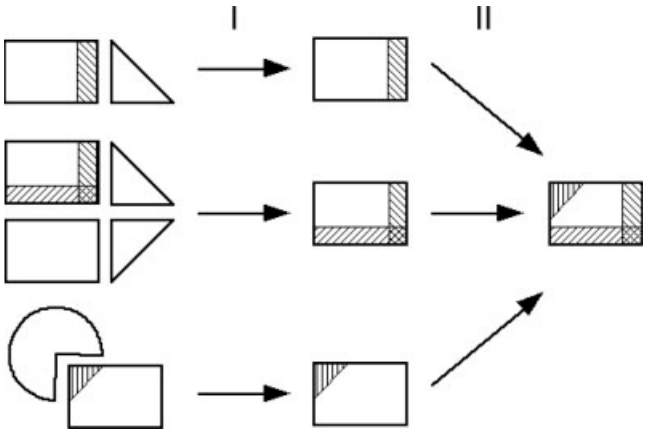

Fig. 1. Schematic representation of the procedure to map known interaction sites (interfaces) from multiple complexes involving representative chains or their close homologs. As an example, three different complexes involving homologous monomeric structures (represented by rectangular shapes) are shown to the left. Shaded areas in the rectangular shape correspond to binding sites identified using the respective complexes (step I), which are then mapped into a structure of a representative chain (step II of the procedure).

Definition of an Interaction Site

In this work, following in the footsteps of previous studies we define interaction sites based on the RSA change upon complex formation, that is, RSA difference between an unbound and bound (complex) structure of an individual chain.[7,51] For each chain considered here, its coordinates are first extracted from the corresponding complex structure and the DSSP program is used to compute the surface exposure of each amino acid residue in an unbound structure of a single protein chain. Subsequently, residues that become buried at the interface upon complex formation may be identified by recomputing the level of surface exposure in the whole complex.

Specifically, an amino acid residue is regarded as an interaction site if (i) it is surface-exposed when considering the structure of an individual protein chain; (ii) the change in its RSA between the isolated chain and the corresponding complex structure is greater than 4% RSA, and the change in its exposed surface area in absolute terms is greater than 5 Å^2. As with other arbitrary thresholds, used in this work, we performed sensitivity analysis to assess the effects of such arbitrary choices and we also followed the literature as much as possible to enable comparison with other methods. In particular, the threshold of 4% RSA for the relative change in surface exposure between an isolated chain and complex structures corresponds to a typical error in RSA prediction for buried residues,[35] which are most relevant for the novel fingerprints of protein interaction sites considered here. Moreover, this choice appears to be qualitatively consistent in terms of the resulting interaction interfaces with the work by Jones and Thornton[7] and the resulting Protein–Protein Interaction Server for the analysis of protein complexes (http://www.biochem.ucl.ac.uk/bsm/PP/server/), as well as a more recent work by Offman and colleagues.[52]

Proteins may be involved in multiple interactions with distinct interaction interfaces. These non (or partially)-overlapping interfaces are oftentimes resolved structurally in a number of different complexes with distinct interacting partners. Such multiple complexes may involve different variants of the same protein, for example, because of processing of the original chain to remove flexible fragments for X-ray crystallography, or orthologous proteins from model organisms that are typically used to study protein interactions. On the other hand, prediction methods considered here start with an unbound protein structure as input and assign all surface residues to one of two classes, that is, potentially predicting multiple interaction patches. Nevertheless, many studies in this field[16,26] consider unique interaction interfaces, as defined by nonredundant pairs of interacting protein chains, independently.

To illustrate the difficulties that can be introduced by this approach, let us consider specific examples of complexes and derived from them interaction interfaces, as included in the training set of Bordner and Abagyan.[26] In particular, it appears that using Bordner and Abagyan approach, some complexes (e.g., 1a2z or 1f51) give rise to several nonredundant pairs (and the corresponding interaction interfaces) involving the same chain, for example, (A,B) and (A,C). As a result, two alternative subsets of residues in chain A may be defined as interfacial, with the remaining residues defined as noninterfacial. In other words, because each interface is considered independently, some residues may be first defined as interacting and then as noninteracting, introducing mutually exclusive class assignments in the training and likely limiting the accuracy of the resulting predictor.

In light of the above, instead of considering each interaction interface separately, we map all known interaction sites to representative chains derived from the initial set of complexes, as illustrated in Figure 1. For the sake of identifying alternative interaction sites we consider all complexes involving close sequence homologs (sequence identity of more than 90% and alignment covering at least 90% of the sequence) of representative chains. These close homologs of representative chains are identified using

sequence alignment to all sequences in PDB (as of September 2005) entries with at least two interacting protein chains. Interaction sites found in any of these complexes are mapped to representative protein chains using the corresponding alignments.

Consequently, the training and control sets are reduced to representative chains with multiple interaction interfaces mapped to them. Specifically, while initially 18,964 and 6,066 residues were identified as interacting sites in S435 and S149 sets, respectively, the exhaustive mapping using alternative complexes allowed to reassign as binding sites additional residues, resulting in a total of 22,338 and 6,968 interacting residues in these two sets, respectively. Another advantage of this approach is that it allows one to partially account for the induced fit. Namely, even though the coordinates of a single (representative) chain used for training and validation are obtained from the corresponding original complex, sites mapped as interacting from alternative complexes involve residues that are not at the interface in the original complex. Therefore, these residues are more likely to adopt conformations (and specifically the level of surface exposure) consistent with those observed in unbound structures.

Feature Selection and Extraction

In conjunction with machine and statistical learning approaches, we performed an extensive search to derive, optimize, and evaluate features (fingerprints) that can best discriminate between interacting and noninteracting sites. These features can be roughly divided into four groups: (i) single sequence-based attributes; (ii) features derived from evolutionary profiles of protein families; (iii) features based on protein tertiary structure; (iv) novel RSA prediction-based features, including the difference between the predicted and observed in an unbound structure surface exposure of an amino acid residue (denoted dSA throughout the paper). To assess the discriminatory power of individual features we used the F-score, defined as follows:

$$F = \frac{(\bar{x}_{\mathrm{ni}} - \bar{x}_{\mathrm{i}})}{\sigma_{\mathrm{ni}} + \sigma_{\mathrm{i}}}$$

where $\bar{x}_{\mathrm{ni}}$ and $\bar{x}_{\mathrm{i}}$ are the averages (means) over the noninterfacial and interfacial class, respectively; whereas σ_{ni}, σ_{i} are the corresponding standard deviations. In other words, the F-score measures the separation of means for two populations in terms of their variances, and is very closely related to the F-statistics, which is commonly used to evaluate the separation of means for two random variables.[53]

The physicochemical properties of amino acids are taken from the AAIndex database.[54] In particular, hydrophobicity (AAIndex ID = ARGP820101) and the expected number of contacts within 14 Å sphere (AAIndex ID = NISK860101) were found to be most informative. Features derived from multiple sequence alignment (MSA), including position specific scoring matrices, amino acid frequencies, and entropies, were obtained using PsiBLAST[43] with default parameters, three iterations and the nr sequence database as of January 2005.[55] We also consider the MSA-based conservation of charge, small size of the side chain, and hydrophobicity. In that regard, all amino acids were split into two corresponding classes (i.e., charged vs. noncharged, small vs. large side chain, hydrophobic vs. other) according to the classification proposed by Zvelebil et al.[56] Structure-based features, including the level of surface exposure in unbound structures, the number and distances between surface exposed spatial neighbors, were calculated using DSSP[48] and LOOPP.[57]

We also derived new aggregate features using weighted neighbor averages (WNA) over spatial nearest neighbors. Such averages were found to significantly improve the discriminatory power of most of the features considered here. In particular, we define two types of weighted averages of some property P defined for individual residues:

$$P^{\mathrm{surf}}_{\mathrm{WNA}} = \sum_{i=0}^{N} P_i \ \mathrm{RSA}_i \text{ and } P^{\mathrm{dist}}_{\mathrm{WNA}} = P_0 + \sum_{i=1}^{N} \frac{P_i}{d_i}$$

with weights defined by the corresponding surface exposure RSA, as observed in unbound structures, or normalized by the corresponding distance d to the i-th residue (measured between α-carbons), respectively. N is the total number of 3D neighbors located at the molecular surface and within 15 Å sphere centered at the residue of interest ($i = 0$).

Training and Validation Protocols

We used LDA, SVM, and NN methods, as implemented in Tooldiag,[58] LibSVM[59] and SNNS[60] packages, respectively, to learn classifiers from known examples and to combine individual features into final predictors. In particular, linear LDA-based classifiers are compared with that of nonlinear NN and Gaussian kernel SVMs. For the latter ones we estimated, using 10-fold cross-validation on the training set (S435) that the misclassification versus generalization trade off constant, $c = 1$, and the default width of the Gaussian basis function, g, were optimal in terms of classification accuracy. In the case of NNs, the optimal architecture was found to consist of one hidden layer with 5 to 10 nodes, depending on the size (representation) of the problem. The standard backpropagation algorithm with default parameters was used, as implemented in SNNS, and the training was stopped when no significant improvement was observed on a validation set. In 10-fold cross-validation, 10% of vectors were initially put aside as a control set and then the corresponding subset with 90% of vectors were additionally split into an actual training and a validation set (containing 10% of vectors chosen initially as a training subset) for metaparameters optimization. We would like to stress, however, that our sampling of metaparameters was not exhaustive.

To assess the accuracy of the classification methods developed here we used a number of standard performance measures, including the two-class classification accuracy, Q_2, which is defined as the percentage of correct predictions for two-class problem. We would like to comment here that Q_2 measure may not be very informative, especially when classes are not balanced, that is, when one of the classes (here noninterfacial residues) is significantly overrepre-

sented. Therefore, contrasting the classification accuracy with the so-called baseline classifier, which assigns all the points to the larger class is helpful in assessing the quality of a classifier. Another global accuracy measure that is commonly used in this field is the Matthews correlation coefficient (MCC),[61] which is defined as

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

where TP and TN are the numbers of vectors correctly classified as positive (interfacial) and as negative (noninterfacial), respectively, whereas FP and FN denote the number of data points incorrectly assigned to positive and negative classes, respectively. Furthermore, we use the recall (sensitivity), R, and the precision (specificity), P, defined as follows:

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}100\% \text{ and } P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}100\%$$

Another widely used in the machine learning field way of assessing and comparing the performance of classification methods is based on ROC (Receiver Operator Characteristics) curves. The ROC curve represents the correlation between the false-positive rate, defined as FPR = FP/(FP + TN), and the true positive rate (recall), normalized to 1. A large surface area under the ROC curve indicates an overall high accuracy of a classifier.

RESULTS AND DISCUSSION

Assessment of RSA Prediction Methods

We start the discussion of the results from the assessment of biases in RSA prediction methods that are hypothesized here to yield enhanced fingerprints of protein–protein interaction sites. In Table I and Figure 2, we summarize the assessment of four real valued RSA prediction methods in terms of their overall accuracy and biases at interaction sites. We used the S262 set of representative and nonredundant protein chains from heterocomplexes (a subset of the S435 set) and evaluated the overall accuracy of RSA predictions in terms of the MAE. In the first comparison, we use as the true reference the RSA values computed (using the DSSP program and normalized by the maximum accessible areas used by a particular method) from unbound structures of representative chains. For an alternative evaluation, we used the coordinates of an entire complex that contains a representative chain to obtain the "true" RSA values. We observe that (with the exception of the RVPNet method, which is much less accurate compared with other methods) the accuracy of RSA prediction improves significantly if we use as the true reference the RSA values derived from complexes, rather than from individual chains.

In particular, the original SABLE method (denoted here by Un) that was trained using data derived from single chain (unbound) structures achieves MAE of 17.1% when using RSA values observed in unbound structures as the true reference. On the other hand, if the true RSA values are defined as those taken from complexes, SABLE accuracy improves significantly with the estimated MAE of 15.6%. This difference reflects the presence of a significant subset of residues in interaction sites in the S262 set, for which SABLE predicts RSA consistent, in general, with the complexed (rather than unbound) state. These differences, although less informative for discrimination of interaction sites (see next section), are also observed for other RSA prediction methods considered here. In Figure 2, we specifically compare these biases in terms of differences between predicted and actual RSAs for interacting and noninteracting sites.

TABLE I. Biases in RSA Predictions in Terms of Mean Absolute Errors (MAE) for the S262 Set Using Two Alternative Definitions of the Actual RSA

Method	MAE (Un), %	MAE (C), %
RVPNet	28.3	30.7
PHDacc	20.1	17.4
PROF	18.0	16.7
SABLE (Un)	17.1	15.6
SABLE (C)	17.7	15.3

First, the "true" RSA is derived from single chains, that is, unbound structures extracted from the corresponding complexes are used (results are given the second column, denoted as Un). Next, reference RSAs are obtained from bound structures, that is, using entire complexes that contain representative chains (results given in the last column denoted as C). Two variants of SABLE are assessed, one trained on data from unbound structures (Un) and another using an augmented training set with data derived from complexes (C).

We used again the S262 set to derive two populations of (interacting vs. noninteracting) residues. The distributions of averaged (per protein) differences between predicted and observed (in a single chain structure) RSA values are shown for each method. As can be seen from Figure 2, all methods exhibit systematic "errors," with interaction sites being predicted as more buried (shifted toward negative dSA differences) than noninteracting residues. At the same time, however, some interesting overall biases are revealed, with the RVPNet method overpredicting the level of exposure in general, and other methods overpredicting the level of burial to a different degree. The SABLE method is found to have the most desirable properties in that regard, providing the best discrimination between interacting and noninteracting sites. As a result, the discriminatory power of the difference between predicted and observed in an unbound structure RSA, as measured by F-scores, is significantly higher for SABLE-based predictions (see also next section).

These biases may be further enhanced by retraining RSA prediction methods on augmented training sets that include data derived from complexes, rather than individual protein chains. Using the original training set of 860 chains used to develop the SABLE method[35] we identified about 5% of residues as interaction sites (most of the original structures were monomeric and no mapping of other binding sites was used here). Following the protocol used by Adamczak et al.,[35] we retrained a NN-based regression method using RSA values derived from complexes for these residues. We would like to stress again that the

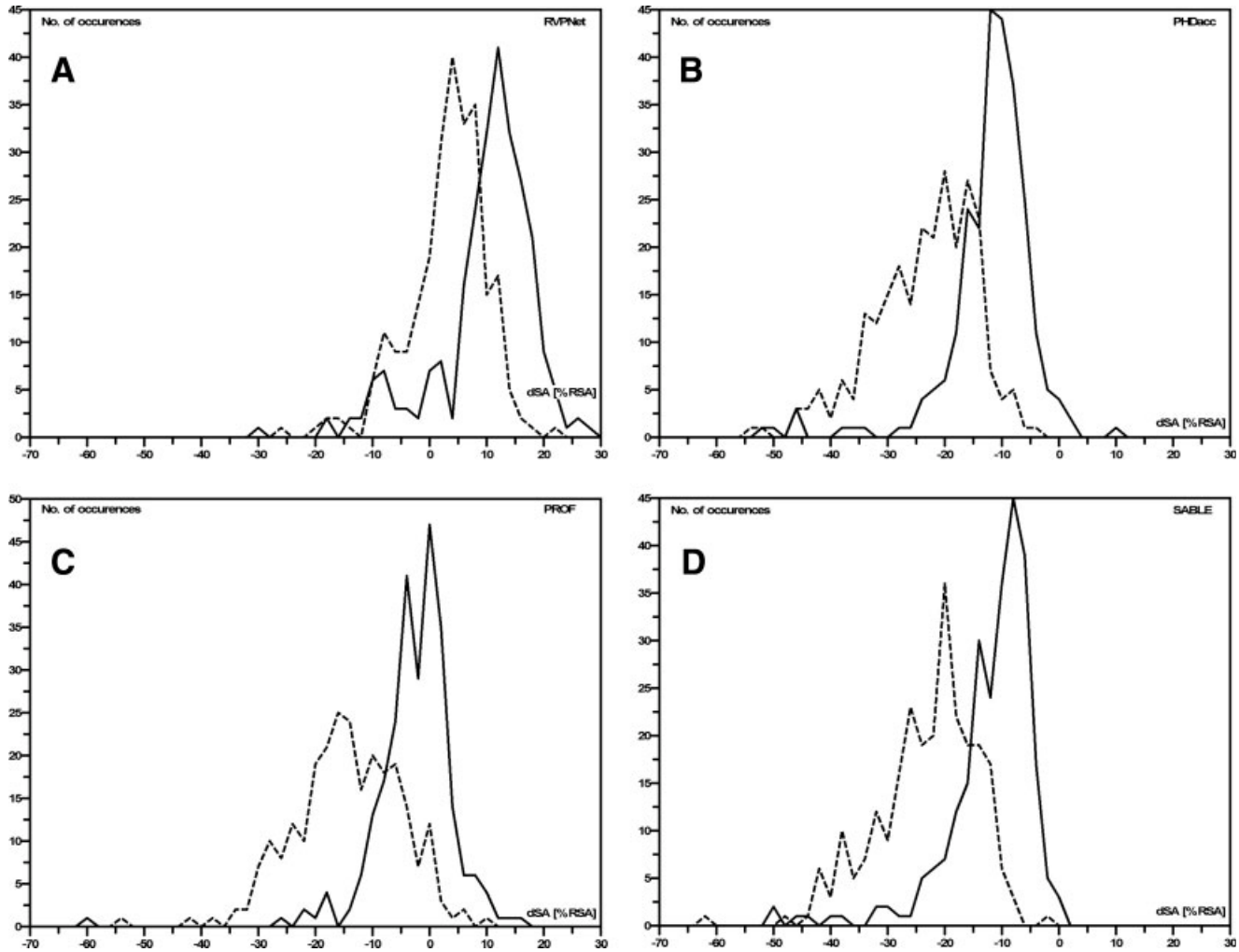

Fig. 2. Distribution of averaged (per protein) differences between predicted and observed RSA values (dSA) for residues within interaction sites (dashed curve) and all other solvent exposed residues (solid curve), in the units of % RSA (hence, dSA of −20 corresponds to residues predicted 20% RSA less exposed than observed, for instance). Four RSA prediction methods are compared, with the results for RVPNet, PHDacc, PROF, and SABLE servers included in panels A, B, C, and D, respectively.

SABLE training set is nonredundant to the S435 set used here for the assessment of RSA predictions. The retrained SABLE method indeed exhibits an increased bias toward interaction sites, with MAE of 17.7% observed when using RSA values from unbound structures as the definition of truth, compared with MAE of 15.3% when RSA values from the corresponding complexes are used as the true reference. We also see a slight increase in the discriminatory power (as measured by F-statistics) when we use the retrained method, which is incorporated in the final protocol for the recognition of protein–protein interaction sites.

Fingerprints of Protein–Protein Interactions

In this section, we discuss the results of the evaluation of individual features in terms of their discriminatory power, as measured primarily using the F-score, defined in Feature Selection and Extraction. Table II summarizes the results for the most informative features (in descending order). Other features, including a variety of amino acid and geometric properties that we tested, proved to be less informative and are not discussed here. The S262 set was used to derive sets of interacting and noninteracting sites and to compute F-scores. As can be seen from the table, the most important features are those that are based on predicted RSA and dSA differences, as well as those that effectively represent the geometry of a putative interaction patch in terms of weighted averages over spatial neighbors (referred to as WNA in Table II and defined in Feature Selection and Extraction; specifically, $P^{\mathrm{surf}}_{\mathrm{WNA}}$ is used for the contact number and hydrophobicity, whereas $P^{\mathrm{dist}}_{\mathrm{WNA}}$ for the remaining aggregate features).

In particular (and in support of our hypothesis), we find that the most informative feature is the neighborhood average for the difference between the experimentally observed (in an unbound structure) RSA and its counterpart predicted by SABLE (i.e., dSA(SABLE-DSSP)), with the F-score of 0.4. Moreover, even without averaging over

TABLE II. Fingerprints of Protein–Protein Interaction Sites and Their Discriminatory Power, as Measured by F-Scores

Feature	F-score
WNA dSA(SABLE – DSSP)	0.40
WNA RSA(SABLE)	0.33
WNA contact number (CN)	0.33
dSA(SABLE – DSSP)	0.29
WNA hydrophobicity (H)	0.27
WNA conservation of charge (CCh)	0.27
WNA conservation of hydrophobicity (CH)	0.26
dSA(PHDacc – DSSP)	0.24
WNA conservation of size (CS)	0.23
WNA conservation of amino acid type (CAA)	0.21
dSA(PROF – DSSP)	0.21
dSA(RVPNet – DSSP)	0.16
Conservation of amino acid type	0.15
Conservation of hydrophobicity	0.14

Note that the difference between the SABLE-predicted and observed (in an unbound structure of an individual chain) RSA is the most informative feature. In addition, structure-based weighted neighbor averaging (indicated by WNA) significantly improves the predictive power of individual features.

3D neighbors, the SABLE derived dSA difference offers significantly higher discriminatory power (F-score of 0.29) compared with the best features that do not involve RSA predictions, that is, conservation of amino acid type and hydrophobicity that result in F-scores of about 0.15. The evolutionary conservation of amino acid properties is measured here by the corresponding entropies at that position, as derived from the underlying MSA. Note that averaging over 3D neighbors improves discriminatory power for these properties, increasing F-scores to 0.27 and 0.26 for WNA entropy of charge and hydrophobicity, respectively. Note also that dSA obtained using SABLE (F-score of 0.29 without WNA averaging) is more informative that those obtained by using other RSA prediction methods (F-score of 0.24 for the second best in that regard PHDacc method). At the same time, dSA using the original SABLE predictor trained on the nonaugmented training set, with data pertaining to unbound structures only, results in somewhat lower discriminatory power (F-score of 0.28).

It is also interesting to note that the actual RSA (not included in Table II), which was used before to enhance interaction site prediction,[26] appears to be less significant (F-score of 0.18 for a weighted average over 3D neighbors). On the other hand, the predicted RSA appears to be quite informative, especially if the neighborhood average is used (WNA RSA(SABLE), F-score of 0.33), potentially boding well for methods that rely on RSA prediction only. However, this is somewhat misleading, since 3D structure is in fact used here to define surface exposed residues (which is laden with additional uncertainty when using RSA prediction), and because the averaging over spatial neighbors (which are not defined unless the structure is known) turns out to be important. In fact, accuracy of predictors that do not utilize structure-derived information is very limited, as discussed in the next section.

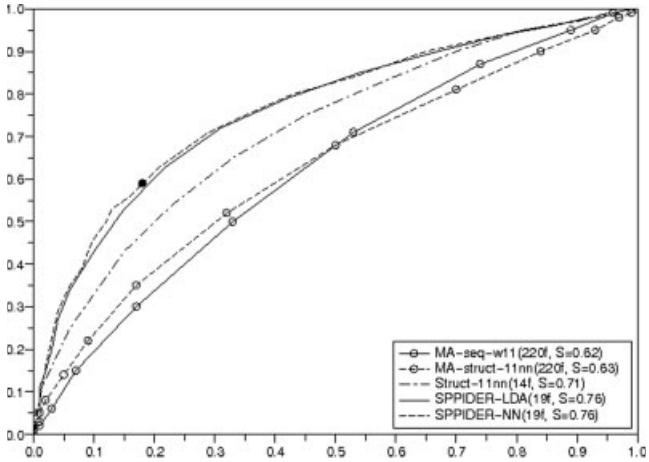

Fig. 3. ROC curves for SPPIDER and other representative methods on the S149 control set. The *y* axis corresponds to the true-positive rate (sensitivity) and the *x* axis to the false-positive rate (see text for details). Note that the surface area under the ROC curve for SPPIDER is significantly larger than for literature-based methods shown for comparison.

Finally, we would also like to comment that the overall discrimination power of individual features is rather limited (note that two distributions with means separated by their average standard deviation result in an F-score of 0.5), especially if no averaging over structure-derived neighborhood is used. This is particularly true about commonly used features derived from patterns of conservation observed in MA, which may explain the rather limited accuracy of methods that simply map evolutionary profiles into the surface of a protein, as indicated in the literature[32,33] and illustrated in Figure 3.

Machine Learning-Based Classifiers

The above findings and novel prediction-based fingerprints of protein interaction sites were incorporated into a new method for enhanced prediction of protein interaction sites, which is referred to as SPPIDER (Solvent accessibility-based Protein–Protein Interaction sites IDEntification and Recognition). After extensive feature selection and extraction, summarized in the previous section and further discussed in the following sections, 19 features were included in the final predictor. Namely, eight features indicated in Table II by the WNA prefix (seven of them also included in Table III below) plus 11 dSA differences for the residues included in the sequence sliding window of length 11 (centered at the residue of interest) were used. To combine these individual features into a classifier we used several machine learning techniques, including LDA, SVMs, and NNs. For training and cross-validation study we used the S435 set and the results are briefly summarized below.

To further support our choice for the final model, we performed leave-one-out feature selection as well as assessment of selected feature subspaces using LDA approach, which enables direct interpretation of the relative importance of individual features. The results for four

TABLE III. Feature Weights in Four Alternative LDA Models Trained Using S435 Set (Contributions to the Norm of the Unit Vector Orthogonal to the Separating Hyperplane) for Several Top Features (Contributions of Remaining Features Are Small and Are Not Shown Here)

	Feature weights in LDA models			
Feature	All features	Without WNA CAA	Without WNA H	No features based on RSA prediction
WNA CN	0.29	0.35	0.30	0.22
WNA CCh	0.04	0.02	0.02	0.34
WNA CH	0.01	0.01	0.02	0.04
WNA CS	0.03	0.00	0.03	0.01
WNA CAA	0.11	X	0.11	0.30
WNA H	0.02	0.02	X	0.10
WNA dSA	0.47	0.56	0.48	X

For the definition of features used, see Table II and the text.

alternative models are shown in Table III: one with all 19 features (first column), one with the neighborhood averaged conservation of amino acid type (WNA CAA) excluded (second column), one with the neighborhood averaged hydrophobicity (WNA H) excluded (third column), and finally a model without features using predicted RSA, that is, consisting of only six features that are explicitly included in Table III. While the first three models achieve similar accuracy (MCC of about 0.4), the latter model performs significantly worse (MCC of about 0.3) on the independent S149 control set. Moreover, WNA dSA is clearly the most important feature in all models that utilize predicted RSA, in consistency with the analysis of discriminatory power of individual fingerprints included in the previous section. It is also interesting to note that when dSA is not used, the conservation of charge (WNA CCh), which otherwise contributes much less to predictions, becomes the most important feature, followed by the conservation of amino acid type (WNA CAA) and the neighborhood averaged contact number (WNA CN). The latter feature captures some geometric characteristics of surface patches and is also important in conjunction with dSA (see Table III).

As discussed in Training and Control Sets, for the S435 set of representative chains we used coordinates of individual chains derived from the corresponding complexes. In other words, the effects of induced fit are initially ignored (see also Comparison With Other Methods). On the other hand, however, we assess the effects of mapping interaction sites from other complexes involving close homologs of representative chains, as described before. An SVM-based classifier is estimated using 10-fold cross-validation to achieve classification accuracy of about 79% and MCC of 0.4, with error bars (standard deviations) of 0.4 and 0.01, respectively. Mapping alternative interaction sites increases the size of the positive class and changes the baseline (the fraction of residues in the larger, i.e. negative or noninteracting class) from about 73% to 68%. At the same time, the estimated classification accuracy for an SVM-based classifier drops to about 75%, thus increasing the differences between the actual classification accuracy and the baseline (this illustrates again the limitations of using the classification accuracy as a global measure of the performance, see Training and Validation Protocols). Performance of NN-based predictors is on par with SVMs, while simple LDA-based classifiers perform slightly worse than more involved nonlinear classifiers.

We further use the S149 set for independent rigorous validation of alternative training protocols and resulting classifiers. All known interaction sites from alternative complexes are mapped to representative chains in the S149 set. Another difficulty that we address here using an independent control set is the uncertainty of the negative class assignment. Predicting whether or not a given amino acid is likely to participate in protein–protein interactions is clouded because experimentally validated data is more abundant for positive cases than negative. In other words, much work exists that positively confirms that a given amino acid is an interaction site; however, data that proves the opposite (that a given residue is *not* an interaction site) is much rarer and more laden with uncertainty. Therefore, special strategies may need to be applied to learn a classifier from data in cases where only "positive" examples (i.e., examples from one class) can be supplied with high confidence and negative examples are unknown or uncertain. One approach to achieving higher accuracy in case of problems with uncertain labels (which we apply here) is based on selection of training examples.

We specifically tested two different strategies of augmenting the training set by removing negative examples that are difficult to classify (and thus not necessarily truly negative). Similar strategies have been proposed in the literature for the prediction of phosphorylation sites, for instance.[62] The first strategy consists of removing data points that are misclassified by some number of individual attributes (such as dSA or entropies) and the second strategy relies on filtering out points that are misclassified by a k-NN classifier in the whole feature space. We explicitly present the results of the second approach. We found that the best results are obtained when 10 nearest neighbors are considered and residues labeled initially as "noninteracting" are excluded from training if at least five of its

TABLE IV. Comparison of Predictors Trained on the Full Training Set (Referred to as "No Filter") of 69,509 Residues Derived From the S435 Set, and on a Subset Obtained by Filtering out About Seven Thousand Residues Initially Classified as Noninteracting and Subsequently Identified as Likely Mislabeled by Using k-NN Approach (Denoted as "k-NN filter")

	Accuracy			
Method	MCC	Q_2, %	R, %	P, %
LDA (no filter)	0.39	73.8	43.0	70.4
SVM (no filter)	0.40	74.4	43.6	72.2
NN (no filter)	0.41	74.5	52.7	67.0
LDA (k-NN filter)	0.41	73.3	59.8	62.3
SVM (k-NN filter)	0.42	74.4	57.4	65.1
NN (k-NN filter)	0.42	74.2	60.3	63.7

Results are given for the whole validation set (S149), without relabeling (or removing) difficult to classify residues. Matthews correlation coefficients (MCC), the two-state classification accuracy (Q_2), recall (sensitivity, R), and precision (specificity, P) are reported for each method.

nearest neighbors belong to the positive class. Using such augmented S435 training set, we were able to achieve a better generalization on the S149 set, as shown in Table IV, even though no filtering is applied to the control set and the initial class assignment based on known interactions is preserved. In particular, significant gains in sensitivity are achieved with somewhat lower specificity (likely to be underestimated though, as many of the false positives correspond to unfiltered, difficult to classify residues that may turn out to be interaction sites).

For the final SPPIDER predictor we chose a consensus-based classifier that combines 10 different NNs obtained in cross-validated training on the augmented S435 set, with k-NN selection procedure used to filter out likely mislabeled points. In fact, the results for NN-based classifiers included in Table IV refer to a consensus of 10 networks from 10-fold cross-validation, with the simple majority voting used to combine individual predictors. As can be seen from the table, NN-based consensus predictor yields overall competitive accuracy (as measured by MCC) and higher than other methods recall, which proved to be advantageous in specific applications. The sensitivity of the final predictor on the S149 set is further assessed in Table V. The fraction of chains (and identified in them interaction interfaces) for which predicted interaction sites overlap with known interacting sites to a different level is given. Note that for about half of 149 representative chains more than 70% of known interacting residues were identified correctly and for about third (29%) more than 90% of known interaction sites were predicted.

TABLE V. Sensitivity of SPPIDER Predictions on the Independent Validation Set (S149)

Overlap with known interaction sites	10%	30%	50%	70%	90%
Fraction of predicted interfaces	0.94	0.83	0.73	0.50	0.29

The number of chains and predicted in them interaction interfaces with an overlap of at least 10, 30, 50, 70, and 90% residues in known interaction sites, respectively, is shown in the bottom row (as a fraction of the total of 149 chains).

Assessment of the Effects of Induced Fit

As mentioned before, most published estimates of the accuracy are based on "unbound" structures conveniently derived from the corresponding complexes by simply disregarding all chains but the chain of interest. Since this approach relies on 3D coordinates derived from the actual complex, the effects of induced fit are ignored. In many cases, however, conformational changes upon complex formation may result in overall structural rearrangement or affect surface patches involved in protein–protein interaction. Therefore, truly unbound coordinates derived from monomeric, independently solved structures should be used for more realistic assessment of the accuracy.

In this section, we present such an assessment for SPPIDER using the S21 subset of S149 control set. For the representative chains included in this set, we were able to identify close homologs for which monomeric structures are known. These truly unbound counterparts of chains in S21 were used to create an alternative S21a set and to assess the decrease in accuracy due to induced fit effects. The average RMSD between structures (measured for alpha carbons) originally included in S21 and their counterparts in S21a is 1.0 (with a standard deviation of 0.5). None of the structures undergoes major conformational change upon complex formation. Thus, our test is limited to relatively small changes induced by binding to another protein. All the interactions were mapped from representative chains from S21 to their counterparts in S21a. Because of slightly different length of some of the chains (between complexes and monomeric structures), the baseline changes from 73% for S21% to 78% for S21a. Here, we used SVM-based classifiers because SVMs are easier to train, enabling comparison of alternative representations with and without B-factors (see below). Moreover, as demonstrated in the previous section, the differences between SVMs and other classifiers are not significant.

As can be seen from Table VI, the overall accuracy using the standard representation consisting of 19 features used to develop SPPIDER (denoted as 19f in Table VI) is significantly lower for the S21 set, compared with the whole S149 (MCC of 0.32 as opposed to 0.42 for the whole set, see Table IV). It is in fact closer to the accuracy observed for transient complexes (see next section), which is consistent with the overall lower averaged surface area buried upon complex formation in chains from S21, compared with the whole S149 set. When truly unbound structures from the S21a set are used instead, the accuracy drops by 0.02 in terms of MCC (to 0.30). Thus, while interaction sites in proteins included in S21 appear to be more difficult to predict, the relative drop in accuracy is small and supports our conclusion that the new fingerprints and prediction methods can be used to improve the recognition

TABLE VI. Analysis of the Effects of Induced Fit and Inclusion of Temperature Factors on a Set of 21 Chains Derived From The S149 set (Denoted as S21) and a Set of Homologs of These Representative Chains, Crystallized Independently as Monomeric Structures (Denoted as S21a)

Control set (representation)	Accuracy			
	MCC	Q_2, %	R, %	P, %
S21 (19f)	0.32	76.5	31.9	63.3
S21a (19f)	0.30	79.1	29.8	56.8
S21 (19f + BF)	0.36	77.4	37.5	64.5
S21a (19f + BF)	0.32	79.7	31.4	59.2

An extended representation consisting of original 19 features plus B-factors is denoted as 19f + BF. Matthews correlation coefficients (MCC), the two-state classification accuracy (Q_2), recall (sensitivity, R), and precision (specificity, P) are reported.

TABLE VII. Performance of SPPIDER in Comparison with Representative Methods From the Literature, Assessed Using a Control Set of 149 Representative Chains (for Details See Text)

Method	Accuracy			
	MCC	Q_2, %	R, %	P, %
SPPIDER	0.42	74.2	60.3	63.7
Struct-11nn	0.29	70.5	33.4	64.9
MA-struct-11nn (16% RSA)	0.27	65.5	55.3	55.0
MA-struct-11nn (5% RSA)	0.17	64.7	35.9	49.1
PPI-Pred	0.16	65.3	27.2	50.4

The overall classification accuracy (Q_2), recall or sensitivity (R), precision or specificity (P), and Matthews correlation coefficients (MCC) are given.

of interaction sites in unbound structures of individual protein chains.

It was suggested that temperature factors may be used to improve the prediction of protein–protein interaction sites. For example, Chung and Bourne proposed to use B-factors to weight contributions from individual structures in their multiple structure-based prediction method that utilizes structurally conserved residues.[27] We evaluated the effects of including B-factors in our own predictors in the context of induced fit. As can be seen from Table VI, while temperature factors improve the accuracy significantly when using unbound structures extracted from complexes (and thus using B-factors that reflect in general lower flexibility of interfacial residues in complexes), the improvements are much smaller when using independently resolved unbound structures. At the same time, the drop in accuracy from MCC of 0.36 to 0.32 due to the induced fit effect is twice as large as the drop from MCC of 0.32 to 0.30 in case of the predictor without B-factors. When setting B-factors to zero to model a situation in which the input structure does not include temperature factors, the accuracy of the results is only slightly better compared with a predictor that does not utilize B-factors at all (MCC of 0.31 vs. 0.30, with somewhat lower sensitivity and slightly improved specificity). Therefore, we decided not to use temperature factors for the final predictor.

We would like to comment, however, that by including these and other features without assessment of induced fit effects, one may easily overestimate the accuracy of predictions. In our case, including B-factors and disregarding the mapping of alternative binding sites in both training (S435) and control (S149) sets, results in a predictor estimated to yield MCC of about 0.5 on the S149 set. We believe that our conservative estimate of 0.42 is more realistic, though.

Comparison With Other Methods

In this section, we further evaluate the effects of different representations and compare SPPIDER with other methods from the literature. We used our own S149 control set as well as several published datasets for this evaluation. Direct comparison with previous methods and published estimates of accuracy is often difficult or impossible, for example, because of differences in the definition of an interaction site, composition of training and control sets as well as validation procedures. Moreover, it appears that none of these methods fully integrates information from multiple complexes to map them into reference chains, elevating the problems of class assignment and largely ignoring the effects of induced fit.

Therefore, we reimplemented several representative methods from the literature. In particular, we assessed the method by Fariselli and colleagues,[17] which represents the residue of interest by identifying 10 surface exposed nearest neighbors in 3D and encoding all 11 residues by the corresponding position specific scoring matrices columns. Our own implementation of this method is referred to as "MA-struct-11nn." As a reference, we also implemented a sequence-based method that does not utilize structural information at all, using an MSA-based representation for a sliding window of 11 sequential neighbors instead (referred to as "MA-seq-w11"). Furthermore, we implemented a method motivated by Bordner and Abagyan approach,[26] with 14 carefully selected structure-based descriptors: weighted average over 11 nearest spatial neighbors for contact number, hydrophobicity and the actual RSA for the central residue, as well as the actual RSA for residues close in sequence (using sliding window of size 11 for comparison with SPPIDER). This approach will be referred to as "Struct-11nn." We also evaluated directly the recently developed PPI-Pred server[46] that utilizes both sequence and structure-based information. The results are summarized in Figure 3 and Table VII.

In accord with some other recent studies,[32] using direct mapping of evolutionary information on the structure of an individual protein chain (represented here by the MA-struct-11nn method) results in a rather limited prediction accuracy. For example, on the S149 set, the classification accuracy of up to 66% (with about 55% recall and 55% precision, respectively) and MCC of up to 0.27 is achieved. On the other hand, when using a carefully designed set of structure-based and evolutionary conservation, as

opposed to straightforward application of MA-based profiles (in general consistency with the approach by[26]), the classification accuracy of about 70.5% (with about 33% recall and 65% precision, respectively) and MCC of up to 0.29 is achieved. In our evaluation, PPI-Pred achieves a classification accuracy of about 65% (with a recall of about 27% and precision of about 50%, respectively) and MCC of 0.16. It should be noted, however, that we use our own definition of an interaction site to define the "true" classification, which may affect the estimated accuracy in the latter case. Also, only the top prediction is taken into account here, as opposed to choosing the most consistent prediction among top three patches predicted by PPI-Pred, as used for the original assessment.[46]

As can be further seen from the ROC curves included in Figure 3, simple mapping of evolutionary information results in a much worse performance (surface area under the ROC curve of 0.63) compared with more refined structure-based approaches (surface area of 0.71 to 0.76). Our estimates of the accuracies are also consistent with recently published results for similar prediction methods.[32,33] We observe that the training set used originally by Fariselli et al. was highly redundant which led to overoptimistic claims. A NN-based predictor trained on the original F226 set, following the protocol by Fariselli and colleagues, yielded a 10-fold cross-validation classification accuracy of 79.3% (with baseline of 72%) and of 0.43, whereas alternative predictor trained on a nonredundant subset F59 was estimated to yield MCC of 0.28 and Q_2 of 75.5% (with baseline of 73%) in 10-fold cross-validation. We would also like to point out that while the accuracy can be numerically somewhat improved when using a different RSA threshold to define surface exposed residues (here we consider 5 and 16% RSA thresholds), predictions with higher thresholds often result in noncontiguous interfaces.

At the same time, SPPIDER, which incorporates RSA prediction-based novel fingerprints, improves significantly on other literature-based methods considered here. For fair comparison of alternative representations and to reduce the effects of training for more involved nonlinear classifiers, all but one ROC curves in Figure 3 were computed for LDA-based predictors. The LDA-based version of SPPIDER with 19 features defined before yields a surface area under the ROC curve of 0.76. Note that SPPIDER-NN, which is used by the SPPIDER server (with the dot indicating the default trade off between sensitivity and specificity), improves only slightly upon its linear counterpart.

We further compared alternative methods considered here on the S21a set of truly unbound structures that was used before to assess the effects of induced fit. As discussed in Assessment of the Effects of Induced Fit, the S21 subset of the S149 set appears to be more difficult to predict for most of the methods considered here. In addition, a further drop in accuracy (of about 0.02 MCC) is observed for SPPIDER when replacing the original (bound) structures by their unbound counterparts. It is therefore of interest to see if the relative advantages of the new method hold on the S21a set. Indeed, as can be seen from Table VIII, the new method significantly outperforms other approaches in terms of all accuracy measures used here. Note that SPPIDER results, which are based on a NN classifier are somewhat different (especially in terms of the trade off between specificity and sensitivity) than the results of an SVM-based predictor included in Table VI. Note also that sequence based methods (represented here by MA-seq-w11) are consistently performing quite poorly, although the advantage of such methods is that they are not dependent on structural details. We would also like to comment that while the accuracy of PPI-Pred is significantly lower than that of SPPIDER, it is in fact somewhat improved compared with the full S149 set as well as the original S21 set (thus, in this test we do not observe a decrease in accuracy due to induced fit for PPI-Pred).

TABLE VIII. Performance of SPPIDER in Comparison With Representative Methods From the Literature, Assessed Using a Subset of Control Set of Representative Chains That Have Crystal Structures in Unbound State (S21a, for Details See Text)

	Accuracy			
Method	MCC	Q_2, %	R, %	P, %
SPPIDER	0.32	77.1	43.8	49.0
PPI-Pred	0.19	73.6	31.2	39.2
Struct-11nn	0.18	76.5	20.4	45.2
MA-struct-11nn (5% RSA)	0.10	70.3	25.7	30.9
MA-seq-w11	0.07	77.0	4.6	39.1

The overall classification accuracy (Q_2), recall (R), and specificity (P) and Matthews correlation coefficients (MCC) are given.

For further assessment of the performance of the new method we used the NT86 set of transient complexes developed by Nooren and Thornton.[45] On this difficult set, SPPIDER achieved an overall classification accuracy of about 74%, with recall of 43% and precision of 47%, and MCC of 0.28. Thus, while the overall accuracy is lower in this case, compared with S149 set, we conclude that SPPIDER can be used to provide useful predictions even for transient complexes. Results on the NT86 set can also be used for further (indirect) comparison with the method by Bordner and Abagyan,[26] who used the same set to test their structure-based method and reported somewhat higher sensitivities and significantly lower specificities compared with our results. As discussed in the Methods, though, the definition of an interaction site in this work appears to be inconsistent, which may lead to unreliable assessment of errors (see discussion in Definition of an Interaction Site).

Finally, we would like to comment that as an alternative strategy to improve upon sequence-based methods, it has been suggested that the predicted RSA can be used to enable a more reliable prediction of interaction sites even when the structural information is not available.[19,31] However, because of the lack of accurate characterization

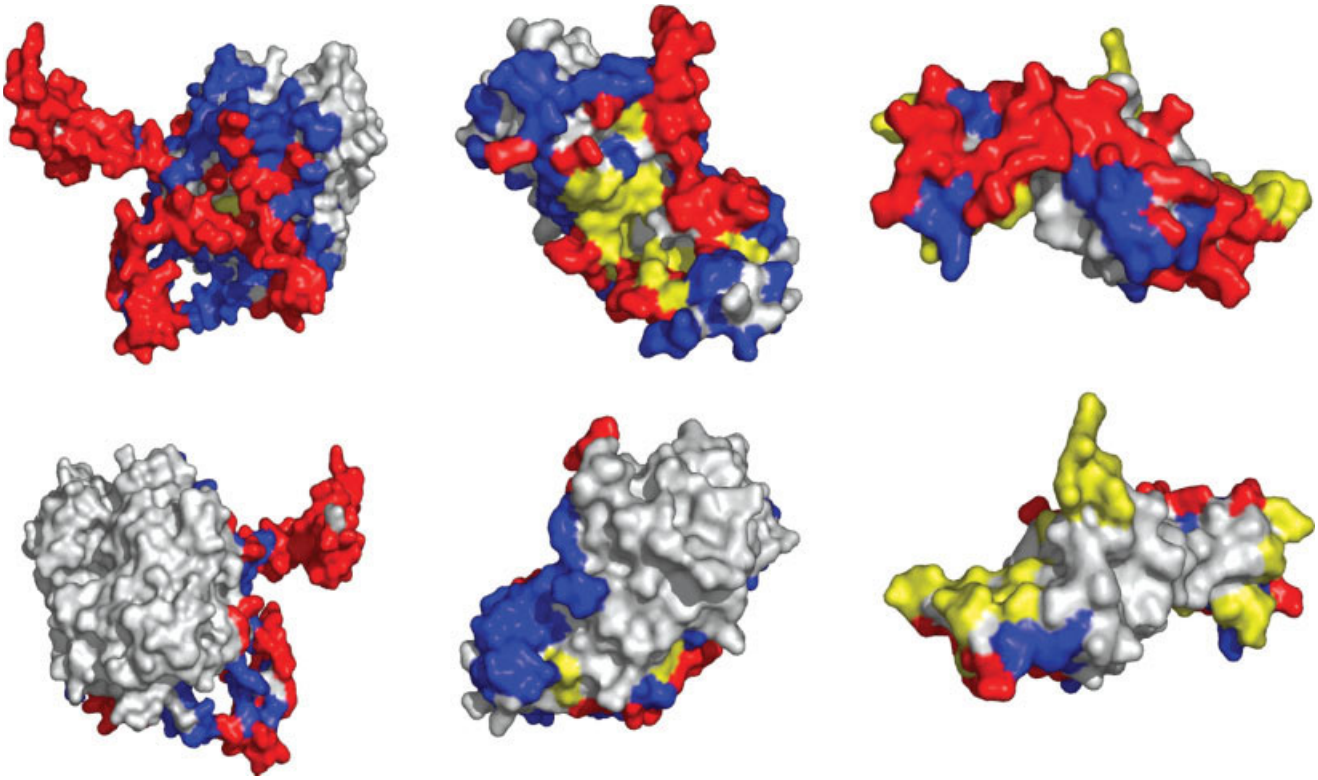

Fig. 4. Examples of SPPIDER predictions of protein interaction sites for: erythrocyte catalase (PDB entry 1f4j, chain A, left panels), cyclin dependent kinase CDK6 (PDB entry 1g3n, chain A, middle panels), and von Hippel-Lindau tumor suppressor protein (1lqb:C, right panels). The following color scheme is used: true positive (known interfacial residues predicted correctly) in red; false negatives (known interfacial residues not predicted by SPPIDER) in blue; false positives (residues predicted as interfacial without support from structural data at present) in yellow. Information about protein binding sites was derived from all complexes containing structural homologs of chains considered here and mapped to representative structures specified above. Panels show front and back views for each protein (animated 3D pictures are available from the POLYVIEW server[63]).

of the surface geometry and the additional uncertainty introduced by RSA predictions,[35,50] it remains to be seen if significant progress can be made without incorporating experimentally derived (or obtained using reliable modeling techniques) high resolution structural information.

Examples of Results for Specific Proteins

Specific examples of current SPPIDER predictions for three different proteins without homology to any of the protein chains used either for the training of SABLE or SPPIDER methods are included in Figure 4. The observed accuracy for these three proteins is close to the overall performance in our tests on large independent control sets (see the previous sections). We only briefly comment here that some of the predicted interaction sites in the VHL tumor suppressor, for which there is no support from structurally resolved protein complexes, that is, technically false positives, are actually indicated as possible interaction sites, based on other experimental data.[64] On the other hand, at least one interaction site in VHL appears to be unknown, and since it is predicted with high confidence, it may be targeted for experimental validation as an example of a false negative in terms of current efforts to map protein interactions. On the other hand, catalase is predicted to have only interaction sites that coincide with known interfaces from existing complexes. Thus, potentially novel interactions for this protein are likely to be either false positives or they may represent a competition for known sites.

In addition, in Table IX we present SPPIDER results on a set of targets from the first four editions of the CAPRI.[65] As before, only predictions for protein chains without sequence homology to any of the proteins used either for the training of SABLE or SPPIDER are included. Unbound structures or structures derived from complexes with other interaction partners, as included in CAPRI target entries are used as input. Interacting residues are identified using our definition of an interacting site and complexes provided for the evaluation of CAPRI predictions. Hence, if multiple interacting chains and the resulting interfaces are included in a given complex they would all contribute to the positive class. This is different than an assessment included in Bradford and Westhead[46] that considers each interface independently, leading to problems with inconsistent definitions of positive versus negative class that were also discussed in Definition of an Interaction Site.

In addition to the overall accuracy (note that precision is likely underestimated in this case since we did not perform a further mapping of alternative binding sites from other complexes involving homologs of chains included in

TABLE IX. SPPIDER Prediction for CAPRI Targets That Are Nonredundant to Our Training Set

Target (Chain Label in Target/Result)	MCC	Q_2, %	R, %	P, %	Overlap With CAPRI, %
T01 (A/H)	0.48	77.3	43.5	83.3	37.0
T01 (C/A)	0.44	68.2	91.4	59.6	51.6
T02 (A/A)	0.12	59.2	42.5	44.6	42.9
T02 (H/E)	0.41	67.8	85.1	45.5	68.8
T02 (L/D)	0.27	72.4	41.3	48.7	54.5
T03 (A/A)	0.30	65.4	57.8	62.4	17.4
T04 (A/A)	−0.05	83.7	0.0	0.0	0.0
T07 (A/A)	0.09	90.6	8.3	20.0	5.9
T08 (A/B)	0.13	77.1	27.3	26.1	23.1
T09 (A/A)	0.22	69.3	50.0	36.7	45.8
T10 (A/A)	0.25	71.2	33.7	53.5	26.1
T12 (A/A)	0.31	80.2	20.8	71.4	17.1
T12 (B/B)	0.14	41.7	100.0	33.3	94.7
T13 (A/F)	0.02	86.2	0.0	0.0	0.0

The overall classification accuracy (Q_2), recall or sensitivity (R), precision or specificity (P), and Matthews correlation coefficients (MCC) computed using our own definition of the interface, as well as the overlap with interfaces defined in CAPRI for specific pairs of chains are given. The structure of a chain labeled as target in CAPRI (typically unbound structure or bound to other interaction partners) is used as input.

Table VIII), we also present the overlap between predicted interacting sites and specific interfaces being evaluated in CAPRI. Note that for two targets (alpha amylase interface with the camelaid antibody VH, target T04, and major surface antigen interface with SAG1, target T13), SPPIDER fails to predict any interaction sites within the interfaces of interest. The average overlap for the remaining targets is about 40%, which is similar to the sensitivity reported by Bradford and Westhead for the best out of their three top predictions.

CONCLUSIONS

The importance of protein–protein interactions continues to stimulate the development of both experimental and computational protocols that aim at elucidating protein networks and the underlying physical interactions. The focus of this work is the problem of the recognition of putative protein–protein interaction sites, which is one of the important intermediate steps toward these bigger goals. We have recently developed accurate methods for predicting the extent of solvent exposure of amino acid residues in proteins.[35,36,66] In this work, we demonstrate how these real valued RSA predictions can be used to improve the recognition of protein interaction sites in the case when a structure of an individual protein chain is known (without knowing interacting partners).

In particular, we show that RSA predictions tend to be consistent with the level of exposure observed in protein complexes, rather than unbound structures. We assessed prediction biases for several real valued RSA prediction methods, including PHDacc,[37] RVPNet,[38] PROF,[39] and SABLE.[35] We observe that SABLE predictions are most consistent with RSAs observed in protein complexes, providing the best discrimination of interaction sites. Moreover, we illustrate how these systematic biases in RSA prediction may be enhanced by augmenting training sets for RSA prediction methods with data derived from protein complexes. The proposed RSA prediction-based fingerprints of protein interactions are also shown to yield significantly improved discrimination of interaction sites compared with entropies (conservation of amino acid type and other properties), hydrophobicity, and various structure-based characteristics used before. Using machine learning approaches, these novel fingerprints of protein interactions are combined with sequence-based and other structural features into a new method for an enhanced recognition of protein–protein interaction sites, which is referred to as SPPIDER.

LDA, SVM, and NN-based classifiers are compared, yielding similar results for a given representation. On the other hand, the performance is greatly affected by the choice of the representation. In particular, we demonstrate (in accord with some recent studies[32,33]) that a direct mapping of evolutionary information onto the structure of an individual protein chain results in a rather limited prediction accuracy. On the other hand, geometric features of potential interaction patches and dSA differences are both shown to contribute significantly to the overall accuracy of results; combining these signals improves the results further. Furthermore, the effects of induced fit, that is, conformational changes associated with complex formation and protein–protein interactions,[67] are evaluated using an alternative control set, consisting of proteins for which both monomeric structures and complexes are known. In addition, several alternative strategies are applied to identify and filter out residues that are likely to be involved in protein interactions, even though there is no experimental

data supporting that classification at present, contributing to improved accuracies.

The new method is shown to significantly outperform existing literature-based approaches that we reimplemented for direct comparison. On an independent control set of 149 nonredundant protein chains derived from recently submitted PDB structures, without sequence homology to chains used in the training and with known interactions sites mapped from (in general) multiple complexes, SPPIDER achieves the classification accuracy of about 74% and a MCC of about 0.42. Moreover, for about half of these representative chains more than 70% of known interacting residues were identified correctly, with the overall precision (specificity) of about 64% and recall (sensitivity) of 60%. Therefore, we conclude that novel prediction-based fingerprints of protein interactions are likely to subsequently contribute to improved identification of functionally important residues, facilitating both biochemical and crystallographic studies of protein complexes and protein interactions.

ACKNOWLEDGMENT

The authors thank Dr. Rafal Adamczak for his assistance in the development of biased RSA prediction methods.

REFERENCES

1. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–753.
2. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA 2003;100:8348–8353.
3. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins 2002;49:350–364.
4. Smith GR, Sternberg MJ. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
5. Wodak SJ, Mendez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14:242–249.
6. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–998.
7. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
8. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
9. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
10. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
11. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002;12:14–20.
12. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
13. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325:377–387.
14. Shapiro L, Harris T. Finding function through structural genomics. Curr Opin Biotechnol 2000;11:31–35.
15. Pazos F, Bang J-W. Computational prediction of functionally important regions in proteins. Curr Bioinformatics 2006;1:15–23.
16. Zhou H-X, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
17. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
18. Jeong E, Chung I-F, Miyano S. Prediction of residues in protein–RNA interaction sites by neural networks. Genome Inform 2003; 14:506–507.
19. Ofran Y, Rost B. Predicted protein–protein interaction sites from local sequence information. FEBS Lett 2003;544:236–239.
20. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17:455–460.
21. Yan C, Dobbs D, Honavar V. InIdentification of surface residues involved in protein–protein interaction—a support vector machine approach. Proc. of the Conf. on Intelligent Systems Design and Applications (ISDA-03), Tulsa, Oklahoma, 2003.
22. Zhang S-W, Pan Q, Zhang H-C, Zhang Y-L, Wang H-Y. Classification of protein quaternary structure with support vector machine. Bioinformatics 2003;19:2390–2396.
23. Koike A, Takagi T. Prediction of protein–protein interaction sites using support vector machine. Protein Eng Des Sel 2004;17:165–173.
24. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000; 302:917–926.
25. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338:181–199.
26. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein–protein interfaces. Proteins 2005;60:353–366.
27. Chung JL, Wang W, Bourne PE. Exploiting sequence and structure homologs to identify protein–protein binding sites. Proteins 2006;62:630–640.
28. Valdar WSJ, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
29. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 2003;326:255–261.
30. Aloy P, Querol E, Aviles FX, Sternberg MJE. 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.
31. Berezin C, Glaser F, Rosenberg Y, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 2004;20:1322–1324.
32. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 2004;13:190–202.
33. Yan C, Dobbs D, Honavar V. A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 2004;20 (Suppl 1):i371–i378.
34. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
35. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks based regression. Proteins: Struct Funct Bioinformatics 2004;56:753–767.
36. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 2005;12:355–369.
37. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
38. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
39. Rost B. PROF: predicting one-dimensional protein structure by profile based neural networks, unpublished.
40. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
41. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
42. Cao B, Porollo A, Adamczak R, Jarrell M, Meller J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics 2006;22:303–309.

43. Altschul SF, Madden TL, Schaffer AA. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
44. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:D154–D159.
45. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein–protein interactions. J Mol Biol 2003;325:991–1018.
46. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–1494.
47. Chothia C. The nature of accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.
48. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
49. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
50. Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005;61:318–324.
51. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins 2002;47:334–343.
52. Offman M, Nurtdinov RN, Gelfand MS, Frishman D. No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. BMC Bioinformatics 2004;5:41.
53. Lehman EL. Testing statistical hypotheses, 2nd ed. New York: Wiley; 1986.
54. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res 1999;27:368–369.
55. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.
56. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
57. Meller J, Elber R. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins 2001;45:241–261.
58. Rauber TW.TOOLDIAG 2.1: Pattern Recognition Toolbox. Available at http://www.inf.ufes.br/∼thomas/ (1997).
59. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm (2001).
60. Zell A, Mamier G, Vogt M, Mache N, Hübner R, Döring S, Herrmann K-U, Soyez T, Schmalzl M, Sommer T, Hatzigeorgiou A, Posselt D, Schreiner T, Kett B, Clemente G, Wieland J. SNNS 4.1: Stuttgart Neural Network Simulator. Available at http://www-ra.informatik.uni-tuebingen.de/SNNS/ (1995).
61. Matthews BW. Comparison of predicted and observed secondary structure of T4 ohage lysozyme. Biochim Biophys Acta 1975;405:442–451.
62. Blom N, Gammeltoft S, Brunak S. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999;294:1351–1362.
63. Porollo A, Adamczak R, Meller J. POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. Bioinformatics 2004;20:2460–2462.
64. Czyzyk-Krzeska MF, Meller J. von Hippel-Lindau tumor suppressor: not only HIF's executioner. Trends Mol Med 2004;10:146–149.
65. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. Protein Sci 2005;14:278–283.
66. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:467–475.
67. Goh C-S, Milburn D, Gerstein M. Conformational changes associated with protein–protein interactions. Curr Opin Struct Biol 2004;14:104–109.