BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 10 2005, pages 2496–2501
doi:10.1093/bioinformatics/bti340

Systems biology

An evolution based classifier for prediction of protein interfaces without using protein structures

I. Reš, I. Mihalek* and O. Lichtarge*

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston,
TX 77030, USA

Received on December 31, 2004; accepted on February 17, 2005
Advance Access publication February 22, 2005

ABSTRACT

Motivation: The number of available protein structures still lags far behind the number of known protein sequences. This makes it important to predict which residues participate in protein–protein interactions using only sequence information. Few studies have tackled this problem until now.

Results: We applied support vector machines to sequences in order to generate a classification of all protein residues into those that are part of a protein interface and those that are not. For the first time evolutionary information was used as one of the attributes and this inclusion of evolutionary importance rankings improves the classification. Leave-one-out cross-validation experiments show that prediction accuracy reaches 64%.

Contact: ires@bcm.tmc.edu; lichtarge@bcm.tmc.edu

1 INTRODUCTION

Protein–protein interactions play a central role in biology since they mediate the assembly of macromolecular complexes, or the sequential transfer of information along signaling pathways. To tease apart the molecular basis of these functions and of protein networks, it is important to identify individual protein–protein interactions and selectively disrupt them through targeted mutagenesis (Onrust *et al*., 1997; Sowa *et al*., 2001; Madabushi *et al*., 2004). Ideally, a prediction of protein interfaces should start with an available protein structure; many techniques, reviewed below, address this problem. Yet, in most cases, the protein structure is unknown. This makes the prediction of protein–protein interface residues, based on a protein sequence alone, an important problem.

To address this problem, it is instructive to consider that predictions of interacting residues, based on structure information, pool many different types of information (Chotia and Janin, 1975; Jones and Thornton, 1996; Lo Conte *et al*., 1999; Chakrabarti and Janin, 2002; Bahadur *et al*., 2003; Nooren and Thornton, 2003; Ofran and Rost, 2003a; Bahadur *et al*., 2004). These studies consider many potential markers of protein interfaces, including amino acid frequencies, hydrophobicity, interface size, shape and planarity. For example, Jones and Thornton (1997a) have studied protein interfaces by examining the properties of surface residue patches. Based on six parameters (solvation potential, residue interface propensities, hydrophobicity, planarity, protrusion and accessible surface area) they were able to differentiate the interface patch from other surface patches. None of these parameters was better than the others as a discriminator. In a follow up work, Jones and Thornton (1997b) have successfully predicted protein–protein interaction sites for 66% of the structures in their dataset.

Given that machine learning algorithms are designed to learn by example in a multiparameter space, several studies have recently begun to use them to predict interacting surface residues, using neural networks and support vector machines (SVMs). Zhou and Shan (2001) and Fariselli *et al*. (2002) analyzed the composition of residues and their structural neighbors and used neural networks to classify surface residues into interacting and non-interacting ones. This showed the importance of considering structural neighbors while building the classifier. Yan *et al*. (2004a) have trained an SVM to predict whether or not a surface residue is an interface residue, and they have achieved high sensitivity (82.3 and 78.5%) and specificity (81.0 and 77.6%) on two different datasets.

Can similar methods be applied to the proteins of unknown structures? In that case the information on residue composition is still available, but the information on neighboring residues and on surface accessibility is not. Ofran and Rost (2003b) and Yan *et al*. (2004b) have independently shown that the interface residues tend to form clusters in sequence. Based on this observation, Yan *et al*. (2004b) have developed a two-stage classifier. It combines both SVM and Bayesian classifiers to predict which surface residues form interface, and it achieves accuracy of 72% and a correlation coefficient of 0.30. However, they did not try to classify all residues in a protein but only those on its surface (which were determined by using the structure).

In contrast, Ofran and Rost (2003b) attempted to classify residues from protein sequences into interacting and non-interacting ones. Their method uses neural networks based on the sequence clustering of interface residues and interface composition. They report an accuracy of 70%, with 20% sensitivity. The only other work we are aware of that attempts to identify interacting residues from sequence is a study by Gallet *et al*. (2000), where the authors have suggested that the identification of interacting residues is possible based on their hydrophobic moments. However, Yan *et al*. (2004b) tested this method on their dataset and obtained a negative correlation coefficient.

One type of information that has not been used in these studies is residue conservation and evolutionary information based on phylogenetic trees. The evolutionary trace (ET) method of Lichtarge *et al*. (1996a) ranks residues based on invariance within functional branches of a phylogenetic tree. ET has been successful in finding

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

novel functional sites (Lichtarge *et al*., 1996b; Onrust *et al*., 1997; Sowa *et al*., 2001; Yao *et al*., 2003; Madabushi *et al*., 2004) and in protein structure evaluation (Mihalek *et al*., 2003). Mihalek *et al*. (2004) have recently developed the real value evolutionary trace (RVET), which combines residue conservation (expressed in terms of information entropy) with grouping of related proteins represented by a phylogenetic tree. They have demonstrated that RVET-based ranking of residues increases the sensitivity and the specificity in the prediction of important protein sites.

While the RVET is a method of choice in this work, other methods aim to identify functional sites (Casari *et al*., 1995; Landgraf *et al*., 2001; Armon *et al*., 2001; Aloy *et al*., 2001; del Sol Messa *et al*., 2003), as reviewed recently by Wodak and Méndez (2004). Here, we present an SVM-based prediction of interface residues which, in contrast to prior studies, incorporates evolutionary information as one of the attributes. [The possibility that the use of evolution might improve prediction of interfaces has occurred to Ofran and Rost (2003b) and Yan *et al*. (2004b) but they did not pursue the idea further.] Since we classify all protein residues and require no structure, this work compares best with the study of Ofran and Rost (2003b). Thus, as a reference, we also built a classifier based only on residue composition. We consider this classifier (referred to as composition) to represent the method used in the Ofran and Rost (2003b) work (the main difference is that they used a neural network, while we use an SVM).

To assess performance we adopted the dataset of Yan *et al*. (2004b), and reduced it to 50 protein chains, as explained below. This set was chosen because it has low sequence identity (<30%), which makes it more challenging than the sets used by other groups mentioned above.

2 METHODS

2.1 Real value evolutionary trace

RVET is a method to rank the evolutionary importance of residues in a protein family. It is based on column variation in multiple sequence alignments (MSAs) and evolutionary information extracted from underlying phylogenetic trees. The first step in rank calculation is to form subalignments that correspond to nodes in the tree. Information entropy is calculated for the initial MSA and then corrected with the contributions from subalignment entropies. This subdivision of an MSA into smaller alignments reflects the tree topology, and therefore the evolutionary variation information within it. The rank of a residue belonging to column i in an MSA is given by

$$r_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left\{ -\sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \right\},$$

where f_{ia}^{g} is the frequency of amino acid of type a within a subalignment corresponding to group g, and index n refers to the number of groups. In case $n = 1$ (no evolutionary information included in the form of subalignments) this expression reduces to the information entropy of column i in the MSA (up to an additive factor of 1). Further details can be found in Mihalek *et al*. (2004), but it is important to note that besides evolutionary trees, any other tree that reflects a reasonable functional classification of a protein family may be used as well.

The range of ranks will, in general, vary from protein to protein, depending on the corresponding phylogenetic trees. In order to obtain a uniform range, all ranks for a protein were converted to a scale ranging from 0 to 1, 1 corresponding to evolutionarily most important residues and 0 representing the least important residues.

Sometimes using 20 amino acid types to rank residues may be too restrictive: a hypothetical mutation that swaps two negatively charged residues might not be as drastic as a change, for example, from proline to alanine. We consider this possibility by grouping residues according to their physical and chemical properties into 14 groups: (Ile, Leu, Val), (Ser, Thr), (Arg, Lys), (Asp, Glu), (Asn, Gln), with the remaining 9 residues considered individually. We incorporate this reduced 14 ‘amino acid’ alphabet into the RVET by simply reducing the sum over 20 amino acids in the above expression to a sum over 14 ‘amino acids’. The resulting RVET is termed ‘Similarity’ RVET to distinguish it from the RVET, which distinguishes equally all 20 amino acids, called ‘Rank’ RVET thereafter.

Of course, alternative reduced alphabets are possible. Recently, Elcock and McCammon (2001) have used a reduced alphabet of six amino acid groups, previously employed by Mirny and Shakhnovich (1999), in an information entropy-based work on identification of protein oligomerization states. Here, we did not want to be strict in grouping amino acids, because the reduction of amino acid types inevitably leads to some information loss. However, we did not systematically investigate which alphabet is most optimal to be used by RVET.

In this work, sequences were collected using BLAST search (Altschul *et al*., 1997) on the NCBI Entrez non-redundant protein sequence database, with the E-score of 0.05. MSAs were built using Clustal W v.1.7 (Thompson *et al*., 1994) in the quicktree mode. We used the UPGMA method (Waterman, 2000) to construct the trees.

2.2 Support vector machines

We use SVM algorithms (Cristianini and Shawe-Taylor, 2000) to address a binary classification problem: residues have to be classified as interacting (‘positive’ examples) or non-interacting (‘negative’ examples). Each instance (residue) is described by an input vector of attributes. The SVMs separate two classes by mapping the input vectors (using a kernel function) into a high dimensional feature space, where a linear separation between the classes with a hyperplane is possible. The implementation of the SVM algorithm used here is the WEKA package (Witten and Frank, 1999), which uses a polynomial kernel. The software can be downloaded freely (http://www.cs.waikato.ac.nz/ ml/weka/).

2.3 Dataset

Our dataset was built from the set of 77 interacting protein chains with sequence identity <30% used by Yan *et al*. (2004b), which itself was extracted from a set of heterocomplexes used in the work of Chakrabarti and Janin (2002). We removed the antibody–antigen complexes (13 chains) because these interfaces are special from the evolutionary standpoint; they are less conserved than the rest of the protein. Furthermore, in two cases one of the partners contained a pair of chains. These were also removed because this would lead to artificial false positives. For example, consider a hypothetical complex of chains AB interacting with chain C. The method would predict not only the interface of AB with C but also the interface between A and B.

Since our method is based on MSAs, our dataset was further limited to chains that could be aligned reasonably well. We eliminated short alignments (≤11 sequences) and those that had sequence identity of ≥80% (this point is elaborated later in the text). Based on these criteria, our final dataset consisted of 50 interacting protein chains, containing a total of 9673 residues.

Interface residues were defined as surface residues that lost relative surface accessible areas (RSAs) upon complex formation. Surface residues were defined as those for which RSA ≥5% (Valdar and Thornton, 2001). The solvent accessibility was calculated using the program NACCESS (Hubbard and Thornton, 1993), which implements the Lee and Richards algorithm, with a probe sphere of radius 1.4 Å (Lee and Richards, 1971). Using these criteria, we obtained 1532 (16%) interface residues (positive examples) and 8141 non-interface residues (negative examples).

2.4 Attributes

We constructed SVMs using either residue composition or evolutionary information, or both. The attributes assigned to residues were calculated from MSAs of homologous protein sequences. For the composition-based classification, each residue is represented by a 20-component vector, which contains the frequencies of 20 amino acid types appearing in the corresponding column in an MSA. For the classifiers which use RVET ranking (or information entropy, as a special case), each residue is described by one number: the rank of the position in an MSA where the residue belongs. Following the literature (Ofran and Rost, 2003b; Yan *et al*., 2004b), the sequence neighbors of the target residue were also included in the attributes, using a 9-residue sequence window. This leads to a 180-component vector of attributes for each instance in the case of a composition based classifier (20 components for the target residue and for each of its 8 neighbors), and a 9-component input vector used by the RVET ranking-based classifier (1 component for each of 9 residues in the window). A classifier which combines composition with evolutionary information requires 189-component vectors. None of the above described attributes require the knowledge of protein structure.

2.5 Cross-validation

The performance of our classifiers was evaluated through a leave-one-out cross-validation method. In general, cross-validation consists of breaking the data into two sets: the ‘training set’ which is used to train the classifier, and the ‘test set’ on which the trained classifier is tested. For the leave-one-out procedure, 1 protein chain was taken out of the dataset and later used for testing, while the remaining 49 chains were used as the training set. This was repeated once for each of the 50 chains.

Only 16% of the data are interacting residues, which leads to highly unbalanced training sets. Using these training sets as such would result in an SVM classifier which classifies all residues as non-interacting. To obtain a balanced training set from each chain that is to be used for training we extracted interacting residues and an equal number of randomly sampled non-interacting residues.

3 RESULTS

The results reported in this work concern the evaluation of residue classification based on the following quantities: the number of true positives (TP) (residues correctly classified as interacting), the number of true negatives (TN) (residues correctly classified as non-interacting), the number of false positives (FP) (non-interacting residues incorrectly classified as interacting), and the number of false negatives (FN) (interacting residues incorrectly classified as non-interacting). We use the following standard measures of performance:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}},$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Correlation coefficient} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

In brief, sensitivity is equal to the fraction of interface residues found, specificity equals the fraction of determined non-interface residues, positive predictive value (PPV) measures the probability that a positive prediction is correct and accuracy gives the percentage of correct predictions. Correlation between the predictions and actual data is measured by the correlation coefficient (CC), which ranges from −1 (perfect anticorrelation) to 1 (perfect correlation). A random classification of a large set of residues as interacting or non-interacting would, for a dataset in which 16% of residues are interacting, result in sensitivity, specificity and accuracy equal to 50%; PPV would be 16%, with the CC equal to 0.

Table 1. Prediction results for evolution and composition based classifiers

	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	CC
Rank	58.1	54.9	24.0	54.0	0.100
Similarity	59.0	53.9	24.3	53.7	0.100
Entropy	54.7	54.9	22.9	53.7	0.074
Composition	56.9	56.0	23.9	56.7	0.097

‘Rank’ and ‘Similarity’ refer to classifiers which use evolutionary ranking of residues calculated by RVET.

Table 2. Prediction results for classifiers with combined attributes

	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	CC
CRank	58.8	58.9	26.0	58.7	0.135
CSimilarity	59.3	58.0	25.5	58.2	0.132
CEntropy	55.1	59.2	24.6	57.9	0.109

CRank (CSimilarity) labels a classifier which combines composition with RVET rank (similarity rank). CEntropy refers to a classifier which combines composition with information entropy.

3.1 Performance of evolution versus composition based SVMs

The results for evolution-only-based classifiers are given in Table 1. We notice that the differences in performance among classifiers are small. Overall, classifier based on composition alone seems to slightly out-perform the others (3% better accuracy). On the other hand, the best sensitivity (59%) is achieved by the Similarity RVET-based SVM, and evidently all classifiers have a low PPV. It is interesting to note that while composition-based classifier uses 180-component vectors as input, the other three classifiers use only 9-component vectors and results are nearly the same.

3.2 Performance of SVMs with combined attributes

In order to test whether the compositional and evolutionary information were complementary, and in an effort to increase performance, we constructed SVMs that combine both types of attributes, using 189-component input vectors. The performance is given in Table 2. Combining the composition with either of the evolution based measures (Rank RVET or Similarity RVET) led to a slight improvement in performance as compared with the composition based SVM (labeled ‘Composition’ in Table 1): 3% increase in specificity, 2% increase in sensitivity, PPV and accuracy. Although these changes are small, they are consistent across all measures of performance used.

Considering the three classifiers compared in Table 2, it seems that combining the information entropy-based SVM with the Composition SVM has the least improvement when compared with the

Table 3. Performance on randomized datasets

	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	CC
Rank	50.5	50.2	20.3	50.2	0.004
Similarity	45.9	52.9	19.6	52.0	−0.010
Entropy	46.3	53.5	19.9	52.8	−0.004
Composition	51.5	49.7	20.5	50.2	0.012
CRank	51.6	47.0	19.2	48.2	−0.014
CSimilarity	51.3	49.1	20.0	49.3	0.001
CEntropy	52.5	47.3	20.1	48.4	0.001

Notation is the same as in Tables 1 and 2.

composition-based SVM. Perhaps this is not surprising, considering the fact that the information provided by the information entropy is already at least partially present in the composition-based attributes. For example, a column in an MSA that consists mostly of one amino acid type will contribute a vector of 20 components to the composition-based classifier, most of which are 0, and one component close to 1. Using this knowledge we can anticipate that the corresponding input to the entropy-based classifier would be a number close to 0 (small entropy), so combining the two does not yield new information. On the other hand, the RVET-based ranks provide information from phylogenetic trees that is not present in the composition-based classifier.

3.3 Inferring significance using randomized datasets

In this work we used a relatively small dataset, hence it is important to infer the significance of our results. This was done by comparing the performance of each classifier with a corresponding classifier trained on a training set in which the labels ‘interacting’ and ‘non-interacting’ were randomly reshuffled. The results are given in Table 3. Comparing Tables 1 and 2 with Table 3 shows that every SVM used in this study performs better in all measures than the randomized classifiers.

3.4 Reducing false positives

Important residues in a protein tend to form spatial clusters (Madabushi *et al*., 2002). Likewise, sequence-based clustering of interacting residues was also observed by Ofran and Rost (2003b) and was used to reduce FP by eliminating positive predictions of fewer than four interacting residues in a window of six residues. Here we follow this lead by implementing a variable filter. After the SVM predictions are obtained, a 9-residue window is moved along the sequence. Predictions of less than N interacting residues in the sliding window, with N ranging from 1 to 6, are considered to be negative predictions.

This method to reduce false positives will, in general, increase specificity at the cost of decreasing sensitivity, because inevitably we will encounter correct positive predictions that will be converted to negative predictions using the above criteria. This can be seen in Figure 1, where we plot sensitivities and specificities for different filtering values of N. We again notice that combining composition with information entropy does not significantly change the results as compared with just using attributes based on the composition. Furthermore, Figure 1 shows that classifiers can be adjusted, through the choice of parameter N, to maximize sensitivity or specificity.

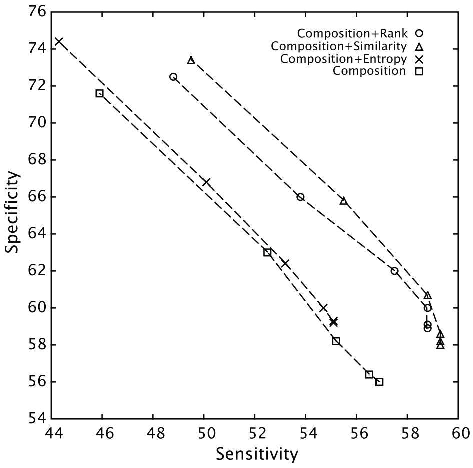

Fig. 1. Specificity versus sensitivity plot obtained by using different values for filtering parameter N. The dashed lines which connect the measured points serve to lead the eye. As N increases from 1 (lower right-hand side of the plot) to 6 (upper left-hand side of the plot), the specificity increases at the cost of lower sensitivity.

Table 4. Reducing false positives using filtering

	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	CC
CRank	57.5	62.0	27.4	60.9	0.151
CSimilarity	55.5	65.8	28.3	63.6	0.168
Composition	55.2	58.2	24.4	58.2	0.101
CEntropy	53.2	62.4	25.8	60.0	0.122

Notation is the same as in Table 2.

A complete picture of performance is shown in Table 4, where for each of the classifiers in Figure 1 we selected a point which in our view represented the best performance that can be obtained by filtering. The best results are achieved with the classifier that combines the composition with the RVET similarity ranks attributes. Its sensitivity is 56%, specificity is 66% and the overall accuracy is 64%.

4 DISCUSSION

This paper addresses the problem of predicting residues involved in protein–protein interaction, using only sequence and MSA information. The biological importance of this problem lies in the fact that the number of known protein sequences is still much larger than the number of available structures. This problem is harder than predicting a functional interface given a protein structure, and predictably, existing methods [the best of which is probably the work of Ofran and Rost (2003b)] perform worse than those that use knowledge of the structure (Zhou and Shan, 2001; Fariselli *et al*., 2002); or which limit classification to the known surface residues (Yan *et al*., 2004a,b).

Motivated by a successful use of evolutionary phylogenetic information to predict functional sites, we have sought to combine evolution with previous approaches (Ofran and Rost, 2003b) by using SVMs. We found that the SVM classifier based on evolution alone was nearly on par with an SVM based only on composition, even though it uses 20-fold fewer parameters. Moreover, these two types of information are not redundant and better results were obtained by combining composition with the RVET ranking of residues. A simple clustering can then be used to further reduce false positives, leading to the best classifier with an accuracy of 64%. This is a 6% increase over the composition-based classifier, and it leads to a positive predictive power of ∼30%.

Why is the improvement in performance obtained by incorporating evolutionary importance of residues not greater? We believe one answer lies in the nature of protein 'hot spots' (Wells, 1991; Bogan and Thorn, 1998; Halperin *et al*., 2004), which show that only a small subset of interface residues contributes the most to the interaction energy. Accordingly, evolutionary trace ranks may best pick out the hot spot residues but be less informative about the rest—the majority of the interface, which is less energetically and evolutionarily important (Halperin *et al*., 2004).

The second answer is technical: RVET ranking adjustments may also have affected the performance. The scale of ranks is a function of tree topology. A tree with many branches will lead to ranks in a different range of values than a tree having only a few branches. This problem was handled by rescaling all ranks to the same scale ([0–1]), but this is an artificial solution. To reduce the effects of this problem on classification we tried to keep only alignments with comparable sequence identities, leading to trees that are similar, so that the rank ranges are also expected to be similar. However, we could not adhere too strictly to this point because the dataset would become too small. As mentioned before, only alignments with identity of <80% were kept, and this led to an increase in performance as compared with keeping all alignments (data not shown).

Despite these limitations, it is important to appreciate that classification of all protein residues is a harder problem than the classification of surface residues. Many residues might be important for reasons that are not directly related to protein–protein interaction (e.g. protein folding), and this would be a source of noise in this type of prediction. Furthermore, the fraction of residues forming interface will be higher if the classification is done on surface residues only, increasing the probability that a positive prediction is correct. We believe these are the reasons why all SVMs considered in this work have relatively low PPV. For comparison, Yan *et al*. (2004b) report PPV of 58% in classification of surface residues, while the best result obtained here for classification of all residues is 28.3%. If the choice of the representative point in Figure 1 is geared toward achieving maximal PPV, we obtain a PPV of 32%, combining Composition with Similarity or Composition with Rank (corresponding to $N = 6$ in Fig. 1).

In conclusion, we have considered classification of protein residues without using the structural information. We have shown that SVM classifiers which combine residue composition with RVET-based evolutionary information lead to an increase in performance over exclusive use of composition. The improvement is small but significant, and the reasons for this were discussed. Our choice of reduced amino acid alphabet used in the Similarity RVET was heuristic, and a possible way of improving the results might be to optimize the alphabet.

ACKNOWLEDGEMENTS

I.R. thanks Dan Morgan for critically reading the manuscript. We gratefully acknowledge support from the March of Dimes (MOD 03-93), the NSF (DBI-0318415) and the NIH (GM-66099).

REFERENCES

Aloy,P. *et al*. (2001) 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.* **311**, 395–408.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Armon,A. *et al*. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.

Bahadur,R.P. *et al*. (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Bahadur,R.P. *et al*. (2004) A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.* **336**, 943–955.

Bogan,A.A. and Thorn,K.S (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.

Casari,G. *et al*. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.

Chotia,C. and Janin,J. (1975) Principles of protein–protein recognition. *Nature*, **256**, 705–708.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines.* Cambridge Universiy Press, Cambridge, UK.

del Sol Mesa,A. *et al*. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.

Elcock,A.H and McCammon,J.A. (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl Acad. Sci. USA*, **98**, 2990–2994.

Fariselli,P. *et al*. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.

Gallet,X. *et al*. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917–926.

Halperin,I. *et al*. (2004) Protein–protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure*, **12**, 1027–1038.

Hubbard,S.J. and Thornton,J.M. (1993) NACCESS [Computer Program]. Department of Biochemistry and Molecular Biology, University College London.

Jones,S and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Jones,S. and Thornton,J.M. (1997a) Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.

Jones,S. and Thornton,J.M. (1997b) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.

Landgraf,R. *et al*. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Lichtarge,O. *et al*. (1996a) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.

Lichtarge,O. *et al*. (1996b) Evolutionarily conserved $G_{\alpha\beta\gamma}$ binding surfaces support a model of the G protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507–7511.

Lo Conte,L. *et al*. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **185**, 2177–2198.

Madabushi,S. *et al*. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.

Madabushi,S. *et al*. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.* **279**, 8126–8132.

Mihalek,I. *et al*. (2003) Combining inference from evolution and geometric probability in protein structure evaluation. *J. Mol. Biol.* **331**, 263–279.

Mihalek,I. *et al*. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265–1282.

Mirny,L.A. and Shakhnovich,E.I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.

Nooren,I.M.A. and Thornton,J.M. (2003) Structural characterization and functional significance of transient protein–protein interactions. *J. Mol. Biol.* **325**, 991–1018.

Ofran,Y. and Rost,B. (2003a) Analysing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387.

Ofran,Y. and Rost,B. (2003b) Predicted protein–protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236–239.

Onrust,R. *et al*. (1997) Receptor and $\beta\gamma$ binding sites in the α subunit of the retinal G protein transducin. *Science*, **275**, 381–384.

Sowa,M.E. *et al*. (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.* **8**, 234–237.

Thompson,J.D. *et al*. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Valdar,V. and Thornton,J.M. (2001) Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.

Waterman,W.S. (2000) *Introduction to Computational Biology*. Chapman & Hall, London.

Wells,J.A. (1991) Systematic mutational analyses of protein–protein interfaces. *Methods Enzymol.* **202**, 390–441.

Witten,I.H. and Frank,E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, CA.

Wodak,S.J. and Méndez,R. (2004) Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14**, 242–249.

Yan,C. *et al*. (2004a) Identification of interface residues in protease–inhibitor and antigen–antibody complexes: a support vector machine approach. *Neural Comput. Appl.* **13**, 123–129.

Yan,C. *et al*. (2004b) A two-stage classsifier for identification of protein–protein interface residues. *Bioinformatics*, **20**, i371–i378.

Yao,H. *et al*. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.

Zhou,H. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.