BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1487–1494
doi:10.1093/bioinformatics/bti242

Structural bioinformatics

Improved prediction of protein–protein binding sites using a support vector machines approach

James R. Bradford and David R. Westhead*

School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK

Received on September 14, 2004; revised on November 18, 2004; accepted on December 16, 2004
Advance Access publication December 21, 2004

ABSTRACT

Motivation: Structural genomics projects are beginning to produce protein structures with unknown function, therefore, accurate, automated predictors of protein function are required if all these structures are to be properly annotated in reasonable time. Identifying the interface between two interacting proteins provides important clues to the function of a protein and can reduce the search space required by docking algorithms to predict the structures of complexes.

Results: We have combined a support vector machine (SVM) approach with surface patch analysis to predict protein–protein binding sites. Using a leave-one-out cross-validation procedure, we were able to successfully predict the location of the binding site on 76% of our dataset made up of proteins with both transient and obligate interfaces. With heterogeneous cross-validation, where we trained the SVM on transient complexes to predict on obligate complexes (and vice versa), we still achieved comparable success rates to the leave-one-out cross-validation suggesting that sufficient properties are shared between transient and obligate interfaces.

Availability: A web application based on the method can be found at http://www.bioinformatics.leeds.ac.uk/ppi_pred. The dataset of 180 proteins used in this study is also available via the same web site.

Contact: westhead@bmb.leeds.ac.uk

Supplementary information: http://www.bioinformatics.leeds.ac.uk/ppi-pred/supp-material

INTRODUCTION

Structural genomics projects are beginning to produce protein structures with unknown function. Such progress requires accurate, automated predictors of protein function to be developed if all these structures are to be properly annotated in reasonable time. Identifying the interface between two interacting proteins provides important clues to the function of a protein and can reduce the search space required by docking algorithms to predict the structures of complexes (for review, see Halperin *et al*., 2002).

Previously, it has been shown that binding sites share common properties that can distinguish them from the rest of the protein (Chothia and Janin, 1975; Jones and Thornton, 1996, 1997a; Nooren and Thornton, 2003a). For example, hydrophobic residues cluster at some interfaces (Glaser *et al*., 2001; Young *et al*., 1994), especially interfaces of obligate complexes (Jones and Thornton, 1996; Lo Conte *et al*., 1999). Other interfaces have a significant number of polar residues (Jones and Thornton, 1996; Lo Conte *et al*., 1999; Larsen *et al*., 1998), usually where interactions are less permanent (Nooren and Thornton, 2003b; Glaser *et al*., 2001), to accommodate electrostatic interactions (Larsen *et al*., 1998). Conserved residues are most likely to be found at important functional sites on a protein (Livingstone and Barton, 1993; Zvelebil *et al*., 1987), although recently it has been argued that residue conservation is rarely sufficient for complete and accurate prediction of a protein–protein interface (Bradford and Westhead, 2003; Caffrey *et al*., 2004). Jones and Thornton (1997a) also implicated shape and solvent accessibility as useful in differentiating binding sites from the rest of the protein surface.

No single parameter absolutely differentiates interfaces from other surface patches (Jones and Thornton, 1997a), so a number of studies have attempted to combine more than one of these physical–chemical properties (Keil *et al*., 2004; Fariselli *et al*., 2002; Zhou and Shan, 2001; Jones and Thornton, 1997b; Neuvirth *et al*., 2004). Using a similar strategy, we have applied an increasingly popular machine-learning approach, the support vector machine (SVM), to the prediction of protein–protein binding sites. SVMs frequently demonstrate high prediction accuracy whilst avoiding over-fitting. They can also handle large feature spaces and condense the information given by the training dataset using support vectors (Hua and Sun, 2001). Molecular biology applications have included gene-expression classification (Brown *et al*., 2000), protein classification (Zavaljevski *et al*., 2002; Dobson and Doig, 2003), protein fold recognition (Ding and Dunchak, 2001) and prediction of protein solvent accessibility (Yuan *et al*., 2002), β-edge strands (Siepen *et al*., 2003), single nucleotide polymorphisms (Krishnan and Westhead, 2003), protein secondary structure (Hua and Sun, 2001; Kim and Park, 2003), protein quaternary structure (Zhang *et al*., 2003) and T-cell epitopes (Zhao *et al*., 2003). SVMs have recently been applied to protein–protein binding site prediction using the profiles of spatially and sequentially neighbouring sequences (Koike and Takagi, 2004), and sequence neighbours of a target residue (Yan *et al*., 2003, 2004).

We trained our SVM to distinguish between interacting and non-interacting surface patches using six of the surface properties discussed above; and then used this SVM as part of our strategy to predict interface surface patches of proteins not included in the training set. Our aim was to develop a general method applicable to all interface types, so we produced our own high-quality, non-redundant dataset containing 180 proteins involved in both transient

*To whom correspondence should be addressed.

and obligate interactions. Even though our dataset was smaller than those generated by automatic methods (Tsai *et al*., 1996; Keskin *et al*., 2004; Preissner *et al*., 1998), we manually checked every complex for evidence in the literature that they actually occur *in vivo*.

SYSTEMS AND METHODS

Training sets

A comprehensive set of complexes was chosen from the Protein Data Bank (PDB) (Berman *et al*., 2000) and then subjected to a number of stringent filtering steps. Proteins sharing >20% sequence identity with a higher resolution structure (or the most recently determined structure if resolutions were equal) of the same complex type were removed. Evidence in literature had to exist that the complex occurred naturally and was stable as a dimer, i.e. we eliminated interfaces only present as a result of crystal packing. NMR structures were not used, neither were mutant complexes nor structures whose resolution was >3.0 Å. Fragments were allowed unless the interface was severely truncated, but dimers containing a protein of <20 residues were discarded. Complexes whose interfaces were made up of more than one separate chain, or complexes containing more than one binding site of the same type were also removed, as well as complexes containing broken interfaces where one protein contacts the other at two points. Homo-obligomers (from obligate complexes) were classed as such if their subunits shared >80% sequence identity; only the subunit with the largest binding site was retained. A total of 180 proteins taken from 149 complexes survived the filtering process of which 36 were involved in enzyme-inhibitor interactions, 27 in hetero-obligate interactions, 87 homo-obligate interactions and 30 in non-enzyme-inhibitor transient (NEIT) interactions. Interaction type definitions were based on those of Nooren and Thornton (2003a).

Surface generation and interface definition

All the protein surfaces used in this study were solvent excluded surfaces (SES) (Connolly, 1983) generated with a probe sphere of radius 1.5 Å using the MSMS code developed by Sanner and Olson (1996). An atom was defined as part of the interface if it lost >99% of its accessible surface area (ASA) upon complex formation.

Surface patch generation First, the radius of the sphere needed to produce the required patch size (see later) was calculated and this sphere was centred on the surface vertex chosen to be the centre of the patch. Every surface vertex falling within the sphere was labelled as a patch vertex. The irregular topography of most protein surfaces means this procedure often fails to generate a single connected patch. When one large patch was generated with several separate smaller patches only the largest patch was retained. Where the centre of the patch was located at the top of a cavity (or bottom of a protrusion) and sphere diameter was less than the cavity depth (or protrusion height), it was likely that the patch would form a ring excluding a central surface feature; such features were automatically reclassified as part of the patch. At points on the protein surface where sphere radius was greater than protein diameter, surface vertices on the opposite side of the protein to the patch centre could have been enclosed by the sphere and mistakenly labelled as patch vertices. These surface vertices were automatically reclassified as non-patch vertices by eliminating vertices from the patch where the angle between local surface normal and that at the centre of the patch was >110°.

Patch size The size of each patch was based on a study of the relationship between the size of the interface and the sizes of the two proteins within the complex. For each test case the sizes of the proteins and their interface were calculated in terms of number of surface vertices. Using linear regression, we found that the interface size was equivalent to ∼13% of the size of the smallest protein in the complex ($y = 0.13x$, R = Pearson correlation = 0.6) and 12% of the size of its parent protein ($y = 0.12x$, $R = 0.5$). To avoid an excess of non-interacting vertices in the interacting patch and because of the non-circularity of most interfaces, we favoured conservative patch sizes that were less than the average values found above, which was 8% of the size of the smallest protein in the complex when both proteins were known and 6% of the size of the known protein otherwise.

Definition of patch properties

Every surface vertex was labelled with the seven surface properties described below. Properties such as hydrophobicity, conservation and residue interface propensity (calculated for each residue), and solvent accessibility (calculated for each atom) could not be determined for each surface vertex directly. In these cases, vertices were labelled with the properties of the atoms or residues to which they uniquely correspond.

Surface shape We calculated two properties of surface shape called shape index and curvedness (Koenderink, 1991; Duncan and Olson, 1993). Shape index (S) is a number between −1 and +1 that describes the shape of the local surface at any give point and is independent of the scale of the surface as in the following equation.

$$S = -\frac{2}{\pi} \arctan \frac{k_{\max} + k_{\min}}{k_{\max} - k_{\min}} \quad (1)$$

where $k_{\max}$ and $k_{\min}$ are the principle curvatures. Points with a negative index are concave and those with a positive index are convex. Curvedness (R) is defined by

$$R = \frac{\sqrt{(k_{\min}^2 + k_{\max}^2)}}{2} \quad (2)$$

and is a measure of curvature, independent of the nature (convex, concave or saddle) of the surface.

Conservation A BLAST search (Altschul *et al*., 1997) for close homologous sequences of each complex forming protein was carried out against the Swiss-Prot v40.38 database. The homologous sequences (E-values below 10^{-4} and limited to at most 100 sequences) from the BLAST search, together with the query sequence were aligned with CLUSTALW (Thompson *et al*., 1994), and a conservation score at each residue position was then calculated using the Scorecons program (Valdar, 2002).

Electrostatic potential Electrostatic potentials were computed for each individual protein using the Delphi v4 software package (Rocchia *et al*., 2001, 2002). Protons were added to PDB files using the *protonate* and *pol_h* programs distributed with Amber v7 (University of California, San Francisco). *pol_h* was specifically used to place polar hydrogens on lysine, serine and threonine residues. The potential was computed on a discrete cubic grid with 101 points in the x, y and z directions, defined such that the protein filled 50% of the total volume of the cubic grid. We used Amber atomic charges (Weiner *et al*., 1984) and atomic radii derived from the MS program of Connolly (1983). We applied dipolar boundary conditions with a dielectric constant of 2 within the protein and 80 outside. The inside of the protein was defined as any region of space inaccessible to solvent given that the radius of the probe used to map the protein surface was 1.5 Å. Salt concentration was set to 0.5 M with an ion exclusion radius of 2 Å. Sufficient linear/non-linear iterations were performed to give a maximum change of 0.0001 kT e^{-1} at the grid points. Electrostatic potentials were extrapolated to the surface vertices automatically by the Delphi program using the 'site coordinates' option.

Hydrophobicity We applied the Fauchère and Pliska (1983) scale to all hydrophobicity calculations.

Residue interface propensity Our dataset of 180 proteins was used to derive interface propensities for each of the 20 amino acids. The propensities were calculated as a fraction of the SES that each amino acid contributed to the interface compared to the fraction of the SES that each amino acid contributed to the whole protein surface [Equation (3)]. A propensity >1 indicated that the residue occurred more frequently at the interface than elsewhere on the

SES in our dataset.

$$\text{Interface propensity of amino acid r} = \frac{(A/B)}{(C/D)} \quad (3)$$

where A = number of interface surface vertices associated with r, B = total number of interface surface vertices, C = number of surface vertices associated with r and D = total number of surface vertices.

Solvent accessible surface area (ASA) The ASAs (measured in Å) for each atom of a protein were taken from those calculated by MSMS (Sanner and Olson, 1996) as a part of the protein surface generation process.

Support vector machines (SVMs)

SVMs make predictions by automated learning from existing knowledge (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000). This type of learning requires training data where the answer is known so that rules or other functions that fit the data can be generated. The trained method is then used to predict on new data. In this study, we used the pattern recognition SVM of Joachims (1999) implemented in the mySVM software developed by Rüping (2000); and found that using the radial kernel function with parameters C and γ set to 1.0 and 0.01, respectively, provided both good classification and generalization performance. All other parameters were set to their default values.

SVM Training We generated one protein surface patch involved in interactions (interacting patch) and one patch taken from the non-interacting parts of the surface (non-interacting patch) of equivalent size to the interacting patch for each protein in the dataset under test. The centre of an interacting patch was the centre of geometry of the actual interface and the centre of the non-interacting patch was chosen at random from the set of non-interacting surface vertices. Each surface vertex within a patch was labelled with the properties described earlier; these were subsequently normalized between zero and one, and the mean and standard deviation of each property was calculated across the patch to produce 14 SVM attributes. Based on these data, we trained the SVM to distinguish interacting patches from non-interacting patches (Fig. 1a).

Prediction method For each protein, we generated one patch per surface atom; these patches could then be predicted to be part of or outside the interface using the trained SVM from above. The results of the predictions were indicated by a confidence value assigned to each patch by the SVM. Patches predicted to be part of the interface were assigned a positive value with the highest value given to the patch with properties that best reflected those of most binding sites in the training set (Fig. 1b).

To account for overlapping patches and to reduce the number of false positives, we pooled all the patches with a positive confidence value and applied a novel ranking system. First, all patches are sorted according to their confidence value and the negative patches are removed. The positive patches are then assigned to either a predicted patch set or an overlapping patch set. Initially, only the patch with the highest confidence value is assigned to the predicted patch set and the next patch tested against this patch for overlap. If this next patch shares >10% of its residues with the predicted patch, it is assigned to the overlapping patch set, if not, then it becomes part of the predicted patch set. All subsequent patches are then tested against patches in the predicted patch set in a similar manner. The outcome is a set of non-overlapping patches ranked according to confidence value. These ranked patches were defined as our predicted patches.

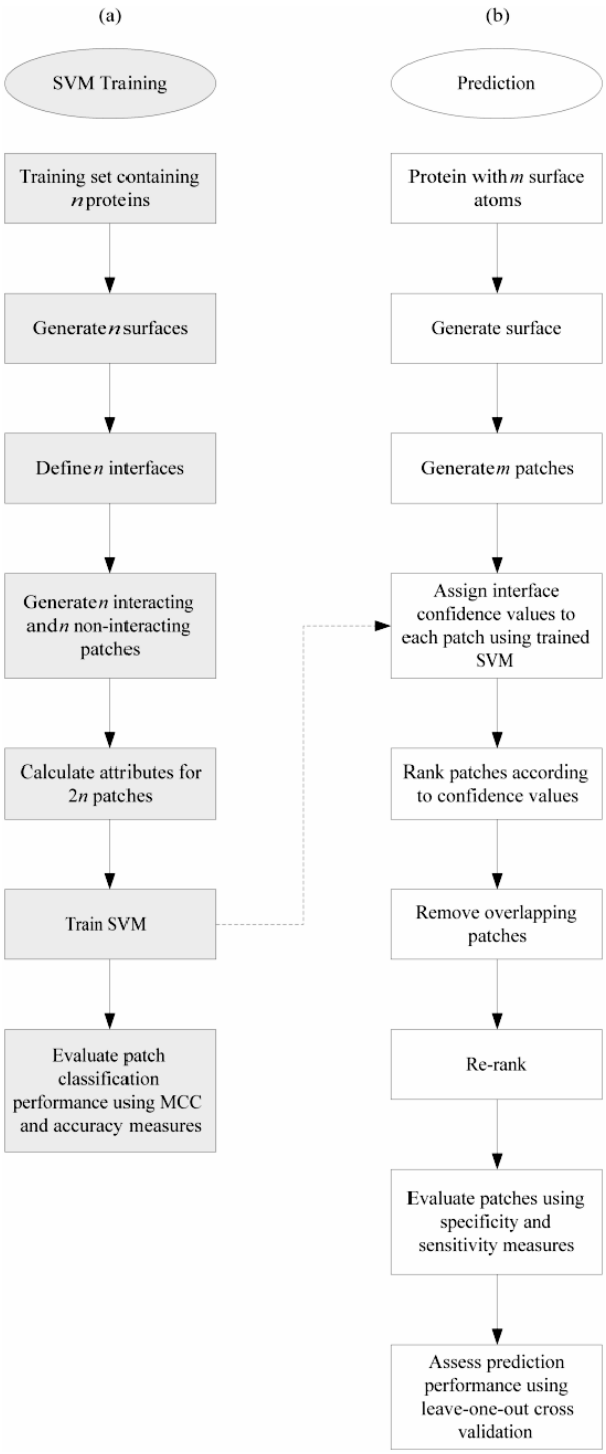

Fig. 1. Flowchart detailing the (**a**) SVM training and (**b**) prediction strategies followed throughout this study.

RESULTS

The results presented here concern the prediction method shown in Figure 1b.

Leave-one-out cross-validation

Leave-one-out cross-validation involved removing one protein from the training set (and the interface residue propensity calculation to avoid bias), training the SVM on the remaining proteins and then predicting the position of the binding site on the selected protein. This process was repeated until all proteins had been left out. Predictions on each individual protein, excluding SVM training, took <10 s on a 1 GHz Pentium III processor. Because non-interacting patches were

Table 1. Summary of mean results from five leave-one-out cross-validations on our dataset

Interaction type	No. of examples	No. of successes[a]	Expected[b]	Rank of successful patch[c]		
				1	2	3
Transients						
Enzyme-Inhibitor	36	23	14	16	3	4
NEIT[d]	30	20	9	12	4	4
Subtotal	66	43	23	28	8	8
Obligomers						
Hetero-obligomer	27	23	10	13	6	4
Homo-obligomer	87	70	30	41	22	7
Subtotal	114	93	39	53	28	12
Total	180	136	65	81	35	19

In all cases, standard deviations were 1–2% of the mean value, they are not shown here for clarity.

[a]Number of proteins where a patch with over 50% specificity and 20% sensitivity was ranked in the top three.

[b]Number of successes expected across the dataset.

[c]Breakdown of the ranks of the patches described above.

[d]NEIT = Non-enzyme-inhibitor transient.

chosen at random for the training step, it was rare for the same set of results to be achieved again with another cross-validation run. Therefore, we repeated the entire cross-validation procedure five times and evaluated average performance.

To each predicted patch, we applied specificity and sensitivity measures relating to those typically used in the field but adapted specifically for this problem. In this study specificity = number of interface residues in patch/number of patch residues, which tells us the proportion of the patch residues that are interface residues and is equivalent to the reliability measure used by Neuvirth *et al*. (2004). Sensitivity = number of interface residues in patch/number of interface residues, which indicates the proportion of interface residues that are included in the patch and is equivalent to the percentage overlap measure used by Jones and Thornton (1997b) and the sensitivity measure used by Neuvirth *et al*. (2004).

In this study, our priority was high specificity with a reasonable level of sensitivity. The reasoning behind this was that complete interface coverage (100% sensitivity) can be guaranteed if the entire surface is treated as a patch, whereas a patch of 100% specificity is often a good indicator of the approximate position of the interface even if sensitivity is comparatively low, owing to small patch size. To this end, we deliberately underestimated patch sizes for each protein (see Methods section). This resulted in 67% of the patches from our dataset being smaller than their corresponding interfaces, with only 4% matching the interface exactly.

A prediction was deemed a success if a patch with over 50% specificity and 20% sensitivity was ranked in the top three. A summary of the mean results of five cross-validation runs based on this criterion is given in Table 1. This table also reports an 'expected' value, which indicates the number of successes one would expect to achieve across a dataset by just making random predictions (see later for how we derive this value). Details of a single run are given for every protein in the dataset individually in Supplementary Table 1. Overall, we were able to predict the location of the interface on 76% (136/180) of the proteins in our dataset. In 60% (81/136) of these instances, a patch with over 50% specificity and 20% sensitivity was the top ranked patch. The stringency of our criterion meant that some proteins were not deemed successful when a reasonable estimate of the interface had been made. For example, in at least four cases during each cross-validation run a patch with over 50% specificity was ranked in the top three even though the 20% sensitivity threshold had not been reached.

A reasonable level of performance extended across all the different interaction types. Success rates ranged from 64% (23/36) for the enzyme-inhibitor interaction type to 85% (23/27) for hetero-obligate interactions. Overall, the SVM achieved a higher success rate, 82% (93/114), with obligate binding sites [Table 1] than transient binding sites, 65% [43/66; Table 1]. This was probably because either the larger number of obligate interactions in the dataset biased the trained SVM towards predicting obligate binding sites, or that transient binding sites were simply more difficult to predict. Performance did not vary significantly between cross-validation runs, as confirmed by a maximum Standard Deviation of only 2% for the success rate on all the protein types.

When patches are located at different parts of a large binding site relative to the size of the protein, then more than one predicted patch with >50% specificity is possible. If these patches are pooled together, most of the interface can be sampled. For example, the four predicted patches produced for 1qax chain A, a 3-hydroxy-3-methylglutaryl-coenzyme A reductase (Tabernero *et al*., 1999), shown in Figure 2A, all achieved >50% specificity but sensitivities of between 9 and 40%. Taken together however, these patches sample >90% of the residues at the dimer interface.

Other datasets

We tested our method against a modified dataset of Jones and Thornton (1997b) and found that our method had performed at least as well as their patch analysis method. We successfully predicted the location of the interface in 72% (34/47) of the dataset using leave-one-out cross-validation as before. With their patch analysis method and using their own criteria based on sensitivity values only, Jones and Thornton achieved a success rate of 64% (30/47).

We also tested our method against the dataset of Neuvirth *et al*. (2004). Further details about this and the comparison with Jones and Thornton (1997b) can be found in Supplementary information.

Significance of predictions

To compare predictions made by the algorithm against results that could have been achieved by simply sampling the protein surface at random we first calculated the probability, p, of finding a patch satisfying our success criterion at random from among the set of patches (one per surface atom) generated for each test case (p = number of patches satisfying criterion/number of patches). This is similar to a strategy used by Jones and Thornton (1997b). When considering the top three patches, we were, in effect, making three attempts at finding a success; so, given p, we calculated the probability of succeeding at least once in these three attempts, $P = 1 - (1 - p)^3$. These P-values allowed us to calculate the number of successes, E, one would expect to achieve across a dataset by just making random predictions, $E = \bar{P}N$, where $\bar{P}$ denotes the average value over the set of N proteins.

Overall, the number of successes achieved with our method was more than twice that of sampling the protein surface at random (Table 1). However, our method also ranked 81 of our 136 successes

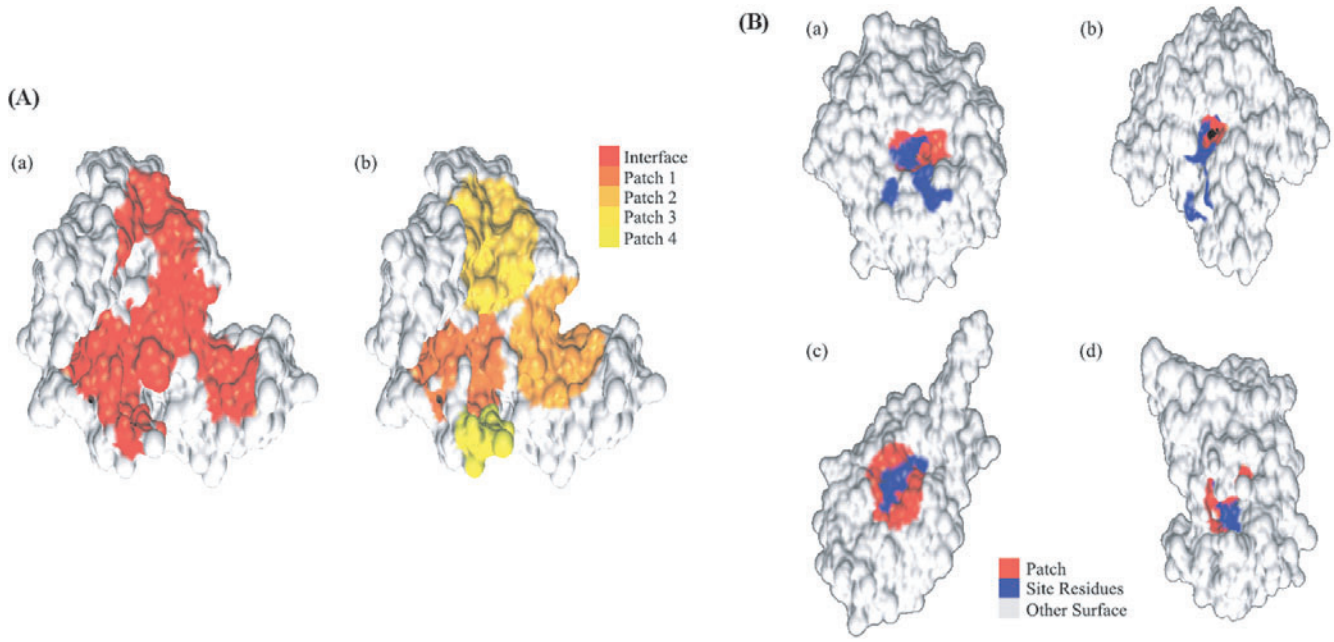

Fig. 2. (**A**) Prediction on 1qax chain A of the 3-hydroxy-3-methylglutaryl-coenzyme A reductase homodimer (Tabernero *et al*., 1999). (a) Dimer interface and (b) the four predicted patches. Together these patches sample >90% of the residues at the dimer interface. (**B**) Functional sites predicted by our SVM as alternatives to the expected binding sites. (a) The catalytic site of nitrite reductase (1aom; Williams *et al*., 1997), (b) the catalytic site of nitric oxide synthase (1nse; Raman *et al*., 1998), (c) the calcineurin binding site of serine–threonine phosphatase (1tco; Griffith *et al*., 1995) and (d) the ATP binding site of phosphatidylinositol phosphate kinase (1bo1; Rao *et al*., 1998).

first. $E = pN$ gives the expected number of patches to be ranked first within the dataset. We calculated this expected number to be 25 for our dataset, less than a third of that achieved by our method.

Heterogeneous cross-validation

We used a second cross-validation strategy on our dataset, which involved training the SVM on the proteins involved in obligate interactions and predicting on the transient (enzyme-inhibitor and NEIT) complex types and vice versa. Whichever interaction type was removed from the full training set was also left out of the interface residue propensity calculation as well. Surprisingly, as shown in Table 2, prediction success was comparable to that of the leave-one-out cross-validation (Table1). This implies that although these transient and obligate interfaces do differ in nature (Nooren and Thornton, 2003b), they share significantly in the properties that enable them to be distinguished from non-interacting parts of the surface. We return to this point later when considering the importance of each patch property for predictions. Overall, we could predict on transient interfaces based on training with obligate interfaces with a success rate of 64% [42/66; Table 2] and on obligate interfaces based on training with transients with a success rate of 83% [95/114; Table 2].

Alternative binding sites

The highest scoring patch should always occupy an area of the protein surface with typical interface properties whether it overlaps with the actual interface or not. It follows that a seemingly unsuccessful patch could actually be occupying another binding site on the protein surface independent of the interface derived from the PDB file. We

Table 2. Summary of mean results from five interaction type cross-validations on our dataset

Interaction type	No. of examples	No. of successes	Expected	Rank of successful patch		
				1	2	3
Transients						
Enzyme-inhibitor	36	22	14	14	5	3
NEIT	30	20	9	13	4	4
Total	66	42	23	27	9	7
Obligomers						
Hetero-obligomer	27	23	10	14	4	4
Homo-obligomer	87	72	30	38	24	10
Total	114	95	39	53	29	14

See footnotes of Table 1 for description of columns.

therefore selected prediction failures where the top ranked patch had achieved 0% specificity and sensitivity with the interface, and where the protein contained another binding site whose residues had been specifically referred to in the literature. In four homo-obligomers, one hetero-obligomer and one protein involved in NEIT interactions, the top ranked patch contained residues of an alternative functional site (Figure 2B). These included potential protein–protein binding sites such as the catalytic sites of nitrite reductase (1aom; Williams *et al*., 1997), nitric oxide synthase (1nse; Raman *et al*., 1998), brefeldin A esterase (1jkm; Wei *et al*., 1999) and Type 1 TGF β-receptor kinase (1b6c; Huse *et al*., 1999), the calcineurin binding site of serine–threonine phosphatase (1tco; Griffith *et al*., 1995) and the binding

Table 3. Results from predictions on ten unbound proteins with >70% sequence identity to proteins within our dataset

Unbound PDB code (_Chain)	Complex PDB code (_Chain)	Surface size (residues)	Interface size (residues)	Details of successful patch Rank	Size (residues)	Sensitivity (%)	Specificity (%)
1d6o_A	1b6c_A	93	21	1	16	67	88
1dks_B	1buh_B	70	13	1	11	62	73
1dqt_A	1i81_C	103	11	2	13	64	54
1f3g	1g1a_F	120	16	1	19	81	68
1he9_A	1he1_A	110	20	2	25	65	52
1hpt	1tgs_I	53	11	2	9	55	67
1rgp	1tx4_A	153	24			No success	
1sup	2sic_E	182	22	2	29	68	52
2ptn	1avw_A	169	22	1	19	68	79
3dni	1atn_D	182	21	1	28	76	57

site of the small molecule ATP on phosphatidylinositol phosphate kinase (1bo1; Rao *et al*., 1998), presumably reflecting some shared properties of small molecule binding sites with protein binding sites.

Unbound proteins

Our SVM was trained and tested on proteins in their bound states where conformational differences between the interface and the rest of the surface are at their most pronounced. In practice, if the binding site is unknown, then the available structure of the protein is more than likely to be that of the unbound state. Shape differences between proteins in their bound and unbound states are usually only apparent at the atomic level but there is potential for our attributes based on shape and ASA to be affected.

In order to test our method on unbound proteins, we first collected a subset of proteins from our dataset containing bound proteins involved in enzyme-inhibitor and NEIT interactions on which our prediction method was most successful, i.e. specificity and sensitivity values of the top ranked patch were frequently >50%. Only ten of these bound proteins had an equivalent unbound structure with >70% sequence identity within the PDB; these ten unbound structures made up our unbound test set.

For each of the ten predictions, we removed the bound protein with >70% sequence identity to the unbound protein from the dataset, trained on the 179 bound proteins that remained and then predicted on the unbound protein (analogous to our leave-one-out method above).

The results of this procedure are shown in Table 3. Our method appears tolerant to most conformational changes involved in complex formation because we were successful in nine out of the ten unbound proteins we tested. All nine of these predictions reached >50% sensitivity, which suggested that our patch sizes, calculated as 6% of the whole protein surface area (see Methods section), were an accurate estimate of interface size. No patch was ranked below two and five were ranked first. We failed to find the RhoA binding site on Rho GTPase-activating protein (RhoGAP), a molecule involved in cell signalling. During complex formation there is a significant conformational change at the binding site on RhoGAP enabling Arg85 to interact with GDP between the two proteins (Rittinger *et al*., 1997). This could account for our failure to find the binding site on unbound RhoGAP.

Validation with CAPRI targets

CAPRI (Critical Assessment of PRediction of Interactions; http://capri.ebi.ac.uk) is a community wide experiment to assess the performance of docking algorithms on targets where only the structures of the unbound components are known. CAPRI targets are also useful to us because they provide a new independent set of interfaces for method validation.

We selected 15 proteins from the 13 targets used in the first four rounds of CAPRI. We omitted proteins that had >20% sequence identity with either another protein in the same complex, another protein in a different target, or with one of the 180 proteins in our own training set. Where possible, we used the interface residues given by CAPRI. In one case, we had to use our own definition of the interface as detailed in the Methods section.

For each prediction, we chose the patch ranked in the top three with the highest specificity and sensitivity values and calculated the probability of obtaining a patch as good or better in x random predictions, where x is the rank of the best patch, using a method related to that described earlier.

Results were very encouraging (Table 4). A significant prediction of the interface was made in 11 of the 15 cases where the P-value for random predictions was <0.25. The best prediction was made on the transient interface on the H chain of *Bacillus subtilis* HPr protein (Target 1) by a patch of rank two with 100% specificity and 52% sensitivity. The interface on the A chain of *Lactobacillus* HPr kinase to which the H chain is bound was predicted by a patch of rank two with 40% specificity and 42% sensitivity. As before, we analysed the four predictions that had failed to find the interface specified by CAPRI and found in two of the cases, Target 2 chain A and Target 3 chain L, that another protein–protein binding site had been predicted with reasonable accuracy. The homodimeric interface in Target 2 between chains A and B was predicted by a patch of rank one with 27% specificity and 54% sensitivity. The heavy–light chain interface on chain L of Target 3 was predicted by the top two patches, the patch of rank one achieved a specificity of 33% and a sensitivity of 77%, whereas that of rank two achieved a specificity of 41% and a sensitivity of 53%.

In ten cases, sensitivity values were higher than specificity values, in contrast to predictions on our own dataset where specificities

Table 4. Evaluation of CAPRI targets

CAPRI Target Number	Chain	Best patch in top three Rank	Specificity (%)	Sensitivity (%)	Probability[a]
1	H	2	100	52	0.01
11	A	2	83	57	0.06
11	B	2	83	26	0.22
10	A	1	49	34	0.09
1	A	2	40	42	0.22
3	A	3	40	52	0.18
3[b]	H	1	40	60	0.10
2	D	2	37	91	0.07
8[c]	A	2	36	55	0.11
8[c]	B	2	34	69	0.05
13	F	1	33	56	0.06
3	L	3	24	38	0.33
7	A	3	11	18	0.68
2	A	2	6	21	0.32
3	C	1	0	0	1.00

[a]The probability of obtaining the predicted patch or better at random.
[b]A homologous protein (1kxq chain H) in the training set was removed prior to prediction.
[c]Binding site residues were obtained from the PDB file 1npe using our definition of an interface (see Methods section).

tended to be higher than sensitivities. This highlights the problem of choosing an appropriate patch size.

Important properties

Our original choice of the seven properties used for training the SVM was based on past studies that have implicated them in distinguishing binding sites from the rest of the protein surface. A posterior analysis of how each property contributed to training the SVM can be found in Supplementary information.

DISCUSSION

We have developed a method for predicting protein–protein binding sites using SVMs. To train the SVM and to test the prediction method we produced our own dataset of 180 proteins—the largest manually produced dataset of its kind containing proteins involved in both transient and obligate interactions that actually occur *in vivo* and are not just a consequence of crystal packing. We were able to successfully predict the location of the binding site on 76% of the 180 proteins in this dataset using a leave-one-out cross-validation procedure. This success rate was achieved using only 14 SVM attributes so prediction performance should be improved when more properties that distinguish between interfaces and the rest of the protein surface become available. Interestingly, when we performed heterogeneous cross-validation by training the SVM on transient complexes and predicting on obligomers (and vice versa) we achieved comparable success rates with obligomers and transients suggesting that the two interface types share some common properties. We have also shown that our method is capable of identifying important functional sites on the protein surface even if the interface specified in the PDB file is not predicted. It would be interesting to learn whether novel functional sites can be found by the SVM on other proteins where prediction has been seemingly unsuccessful.

The method is applicable to both obligate and transient binding sites. This broad specificity represents an improvement over previous patch analysis methods of Jones and Thornton (1997b) who used separate scoring functions for different interaction types; and Neuvirth *et al*. (2004) whose method was only applicable to hetero-transient interactions. Two limitations of patch analysis methods are patch shape (circular) versus interface shape (irregular) and the estimation of patch size. We rarely produced patches that matched interface size and shape, which limited the specificity and sensitivity values that could be achieved with each protein. Our patch size definition seemed to provide a good balance between specificity and sensitivity. Even so, a better way of estimating patch size would have improved the results still further.

ACKNOWLEDGEMENTS

We thank the BBSRC for sponsorship, Dr S. J. Pickering and Dr A. J. Bulpitt for the surface shape code, Dr C. C. Taylor for useful discussions on statistical learning, and the Leeds bioinformatics group.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bradford,J.R. and Westhead,D.R. (2003) Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci.*, **12**, 2099–2103.

Brown,M., Grundy,W., Lin,D., Cristianini,N., Sugnet,C., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci.*, **97**, 262–267.

Caffrey,D.R., Somaroo,S., Hughes,J.D., Mintseris,J. and Huang,E.S. (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.

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

Connolly,M.L. (1983) Analytical molecular surface calculation. *J. Appl. Crystallogr.*, **16**, 548–558.

Cristianini,N. and Shawe-Taylor,J. (2000) *Support Vector Machines and Other Kernel-based learning Methods*. Cambridge University Press, Cambridge, MA.

Ding,C.H. and Dunchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Dobson,P.D. and Doig,A.J. (2003) Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.*, **330**, 771–783.

Duncan,B.S. and Olson,A.J. (1993) Shape analysis of molecular surfaces. *Biopolymers*, **33**, 231–238.

Fariselli,P., Pazos,F., Valencia,A. and Casadio,R. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, **269**, 1356–1361.

Fauchère,J.L. and Pliska,V. (1983) Hydrophobic parameters of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.*, **18**, 369–375.

Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.

Griffith,J.P., Kim,J.L., Kim,E.E., Sintchak,M.D., Thomson,J.A., Fitzgibbon,M.J., Fleming,M.A., Caron,P.R., Hsiao,K. and Navia,M.A. (1995) X-ray structure of calcineurin inhibited by the Immunophilin–immunosuppressant FKBP12–FK506 complex. *Cell*, **82**, 507–522.

Halperin,I., Ma,B., Wolfson,H. and Nussinov,R. (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.

Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.

Huse,M., Chen,Y.G., Massague,J. and Kuriyan,J. (1999) Crystal structure of the cytoplasmic domain of the type I TGF β-receptor in complex with FKBP12. *Cell*, **96**, 425–436.

Joachims,T. (1999) Making large-scale SVM learning practical. In *Advances in Kernel Methods—Support Vector Learning.* The MIT Press, Cambridge, CA.

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.

Keil,M., Exner,T.E. and Brickmann,J. (2004) Pattern recognition strategies for molecular surfaces: III. binding site prediction with a neural network. *J. Comput. Chem.*, **25**, 779–789.

Keskin,O., Tsai,C-J., Wolfson,H. and Nussinov,R. (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Kim,H. and Park,H. (2003) Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.*, **16**, 553–560.

Koenderink,J.J. (1991) *Solid Shape*. The MIT Press, Cambridge, MA.

Koike,A. and Takagi,T. (2004) Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel.*, **17**, 165–173.

Krishnan,V.G. and Westhead,D.R. (2003) A comparative study of machine learning methods to predict the effects of Single Nucleotide Polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.

Larsen,T.A., Olson,A.J. and Goodsell,D.S. (1998) Morphology of protein–protein interfaces. *Structure*, **6**, 421–427.

Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments—a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9**, 745–756.

Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of Protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: A structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Nooren,I.M. and Thornton,J.M. (2003a) Diversity of protein–protein Interactions. *EMBO J.*, **22**, 3486–3492.

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

Preissner,R., Goede,A. and Frommel,C. (1998) Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. *J. Mol. Biol.*, **280**, 535–550.

Raman,C.S., Li,H., Martasek,P., Kral,V., Masters,B.S. and Poulos,T.L. (1998) Crystal structure of constitutive endothelial nitric oxide synthase: a paradigm for pterin function involving a novel metal center. *Cell*, **95**, 939–950.

Rao,V.D., Misra,S., Boronenkov,I.V., Anderson,R.A. and Hurley,J.H. (1998) Structure of type II β-phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation. *Cell*, **94**, 829–839.

Rittinger,K., Walker,P.A., Eccleston,J.F., Smerdon,S.J. and Gamblin,S.J. (1997) Structure at 1.65 Å of RhoA and its GTPase-activating protein in complex with a transition-state analogue. *Nature*, **389**, 758–762.

Rocchia,W., Alexov,E. and Honig,B. (2001) Extending the applicability of the nonlinear Poisson–Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. B.*, **105**, 6507–6514.

Rocchia,W., Sridharan,S., Nicholls,A., Alexov,E., Chiabrera,A. and Honig,B. (2002) Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson–Boltzmann method. *J. Comp. Chem.*, **23**, 128–137.

Rüping,S. (2000) *MySVM*. University of Dortmund, Dortmund, Germany.

Sanner,M.F. and Olson,A.J. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.

Siepen,J.A., Radford,S.E. and Westhead,D.R. (2003) β edge strands in protein structure prediction and aggregation. *Protein Sci.*, **12**, 2348–2359.

Tabernero,L., Bochar,D.A., Rodwell,V.W. and Stauffacher,C.V. (1999) Substrate-induced closure of the Flap domain in the ternary complex structures provides insights into the mechanism of catalysis by 3-hydroxy-3-methylglutaryl-CoA reductase. *Proc. Natl Acad. Sci. USA*, **96**, 7167–7171.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tsai,C-J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.

Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.

Vapnik,V.N. (1998) *Statistical Learning Theory*. J. Wiley and sons, New York.

Wei,Y., Contreras,J.A., Sheffield,P., Osterlund,T., Derewenda,U., Kneusel,R.E., Matern,U., Holm,C. and Derewenda,Z.S. (1999) Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase. *Nat. Struct. Biol.*, **6**, 340–345.

Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,C., Alagona,G., Profeta,S.Jr. and Weiner,P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**, 765–784.

Williams,P.A., Fulop,V., Garman,E.F., Saunders,N.F., Ferguson,S.J. and Hajdu,J. (1997) Haem-ligand switching during catalysis in crystals of a nitrogen-cycle enzyme. *Nature*, **389**, 406–412.

Yan,C., Dobbs,D. and Honavar,V. (2003) Identification of residues involved in protein–protein interaction from amino acid sequence—a support vector machine approach. In Abraham,A., Franke,K. and Köppen,M. (eds), *Intelligent Systems Design and Applications*. Springer-Verlag, Berlin, Germany, pp. 53–62.

Yan,C., Dobbs,D. and Honavar,V. (2004) A two stage classifier for identification of protein–protein interface residues. *Bioinformatics*, **20**, i371–i378.

Young,L., Jernigan,R.L. and Covell,D.G. (1994) A role for surface hydrophobicity in protein–protein recognition. *Protein Sci.*, **3**, 717–729.

Yuan,Z., Burrage,K. and Mattick,J.S. (2002) Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**, 566–570.

Zachmann,C-D., Heiden,W., Schlenkrich,M. and Brickmann,J. (1992) Topological analysis of complex molecular surfaces. *J. Comp. Chem.*, **13**, 76–84.

Zavaljevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

Zhang,S-W., Quan,P., Zhang,H-C., Zhang,Y-L. and Wang,H-Y. (2003) Classification of protein quaternary structure with support vector machine. *Bioinformatics*, **19**, 2390–2396.

Zhao,Y., Clemencia,P., Valmori,D., Martin,R. and Simon,R. (2003) Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, **19**, 1978–1984.

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

Zvelebil,M.J., Barton,G.J., Taylor,W.R. and Sternberg,M.J. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957–961.