Protein Engineering, Design & Selection vol. 17 no. 2 pp. 165–173, 2004
Published online January 20, 2004 DOI: 10.1093/protein/gzh020

Prediction of protein–protein interaction sites using support vector machines

Asako Koike[1,2,3] and Toshihisa Takagi[1]

[1]Department of Computational Biology, Graduate School of Frontier Science, The University of Tokyo, Kiban-3A1 (CB01), 1-5-1 Kashiwanoha Kashiwa, Chiba, 277-8561 and [2]Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-koigakubo Kokubunji City, Tokyo, 185-8601, Japan

[3]To whom correspondence should be addressed at the first address.
E-mail: akoike@hgc.jp

The identification of protein–protein interaction sites is essential for the mutant design and prediction of protein–protein networks. The interaction sites of residue units were predicted using support vector machines (SVM) and the profiles of sequentially/spatially neighboring residues, plus additional information. When only sequence information was used, prediction performance was highest using the feature vectors, sequentially neighboring profiles and predicted interaction site ratios, which were calculated by SVM regression using amino acid compositions. When structural information was also used, prediction performance was highest using the feature vectors, spatially neighboring residue profiles, accessible surface areas, and the with/without protein interaction sites ratios predicted by SVM regression and amino acid compositions. In the latter case, the precision at recall = 50% was 54–56% for a homo–hetero mixed test set and >20% higher than for random prediction. Approximately 30% of the residues wrongly predicted as interaction sites were the closest sequentially/spatially neighboring on the interaction site residues. The predicted residues covered 86–87% of the actual interfaces (96–97% of interfaces with over 20 residues). This prediction performance appeared to be slightly higher than a previously reported study. Comparing the prediction accuracy of each molecule, it seems to be easier to predict interaction sites for stable complexes.

Keywords: accessible surface area/hydrophobicity/interaction site ratio/protein interaction site/support vector machine

Introduction

Proteins perform a biological function by interacting with other proteins, compounds, RNA and DNA. Understanding the characteristics of interfacial sites is a requirement for understanding the molecular recognition process. In addition, the ability to predict interfacial sites is important in mutant design and drug design. The physical and chemical aspects of the protein interface have been investigated in a number of studies. As a result, general interfacial sites are widely recognized as being more hydrophobic, flat and protruding than outer surfaces (Chothia and Janin, 1975; Argos, 1988; Jones and Thornton, 1995; Tsai *et al*., 1997). Analysis of a number of complexes also showed that small interfaces contained abundant polar residues (Glaser *et al*., 2001). However, since these characteristics differ between families and/or biological functions, Jones and Thornton (1996) proposed two kinds of complexes: ‘permanent’ and ‘transient’. A ‘permanent complex’ describes multi-chain proteins, and a ‘transient complex’ describes molecules that form a complex only while performing a biological task, such as signal transduction molecules. It was found that ‘permanent complexes’ are more hydrophobic and closely packed in interfaces but less planar, while ‘transient complexes’ are polar and rich in charged groups (Jones and Thornton, 1996, 1997a; Conte *et al*., 1999). More specifically, a ‘strong transient complex’ is characterized as being larger, less planar, and sometimes more hydrophobic than a ‘weak transient complex’ (Nooren and Thornton, 2003). It is also reported that the conformation of a ‘strong transient complex’ often changes considerably with association/dissociation (Nooren and Thornton, 2003). In addition, different ratios for the amino acid composition of whole sequences were found between homo oligomer (homo permanent complexes), hetero oligomer (hetero permanent complexes), homo complexes (homo transient complexes) and hetero complexes (hetero transient complexes). This suggests that predicting interfacial type on the basis of amino acid composition is possible to some extent (Ofan and Rost, 2003a).

Several methods of predicting interaction sites have been reported including non-automated (Korn and Burnett, 1991) and automated methods (Young *et al*., 1994) based on the hydropathy of the structural surface. The surface is divided into patches and the chemical and physical characteristics of the patches, such as hydrophobicity, flatness, protrusion index, and accessible surface area, are calculated. The interfacial patch is predicted using the sum of these values (Jones and Thornton, 1997b). A prediction method using sequence profiles, with/without the accessible surface area, and neural networks for machine learning has also been reported (Zhou and Shan, 2001; Fariselli *et al*., 2002; Ofran and Rost, 2003b). Another study reported that interfacial sites can be predicted using the hydrophobic moment and averaged hydrophobicity, although the application of this method is limited (Gallet *et al*., 2002). Interfacial prediction has been found to be possible to some extent. However, many of the results reported have been obtained using limited, manually selected data or preliminary predictions without recall and precision or without consideration of unpredictable protein types. It is therefore unclear which types of protein are predictable or unpredictable and how precisely their interaction sites can be predicted.

In this study, interfacial sites were predicted using the profiles of spatially/sequentially neighboring sequences and/or surface patches, plus additional information in both hetero and homo complexes. Support vector machines (SVMs) were used in this prediction because they are known to be a powerful technique for making binary decisions. We also discuss predictable and unpredictable interaction sites.

Materials and methods

Each residue was predicted to belong to a particular interaction site on the basis of the profiles of sequentially/spatially neighboring residues and/or surface patch characteristics. SVMs were used as a learning system for these parameters. Interaction site residues and non-interaction residues were used as positive and negative data, respectively.

Collection of data sets

Complexes consisting of multiple protein sequences in the Protein Data Bank (PDB, December 2001) were extracted and those with a resolution of <3.5 Å were selected. Proteins that contact (the distance between any heavy atoms of contacting proteins was within 0.5 nm) other, dissimilar proteins (the threshold of *E*-value of BLAST = 0.01), were extracted as hetero complex proteins and other contacting proteins were collected as homo complex proteins. Small molecules with <100 residues were removed. From these data, sequence/structure redundancy for all sequence pairs was removed by BLAST using a 25% similarity cut-off of >100 amino acid residue regions iteratively. These values correspond to the 'gray zone' of structure preservation (Rost, 1999).

As a result, 324 protein sequences of hetero complexes and 674 protein sequences of homo complexes were obtained. Although all interaction sites were considered, as described below, not all the proteins that comprised the complexes were included in this set to remove sequence redundancy. For the training/testing set for those complexes, hetero complexes with <20 interfacial residues and homo complexes with <30 interfacial residues were not used to exclude, as far as possible, complexes that did not form complexes *in vivo* (that is, false positive). Although this threshold/method may not be enough to remove false-positive complexes, the automatic complete removal of them is quite difficult. As a result, 271 proteins were used as hetero complex validation data and 292 proteins were used as homo complex validation data. Changing these residue cut-off numbers does not significantly affect the prediction performance. The data set used is available as supplementary data at *PEDS* online.

Definition of protein interaction sites

Surface residues, interaction site residues and inside residues were defined as follows. The solvent accessible surface area (ASA) of each residue was calculated using the DSSP program (Rost and Sander, 1993). The maximum area of each amino acid (X) was calculated using the oligomer GGGXGGG. When the ratio of the surface area of each residue to the maximum area exceeded 10%, it was defined as a surface residue and the remainder was defined as an inside residue. Surface residues were defined as interaction site residues when the distance between any heavy atoms of the interacting proteins was within 0.5 nm. This definition is similar to that of Zhou and Shan (2001). By this definition, ~20% of all residues were defined as interaction sites.

Sequence profiles

Sequence profiles were calculated using PSI-BLAST (Altschul *et al*., 1997). The third iteration alignment, or the converged alignment with the condition *E*-value <0.001 and *h* < 0.001 (*h* is the *E*-value which was used to create a position-specific matrix) was converted into a sequence profile. A profile consisted of amino acid + insertion + deletion + unknown (X) = 23 dimensions for a position. To include the environment of the residue, the profiles of sequentially neighboring residues with *n* windows were also included in the vector. Equation (1) is an example of a vector with 11 windows. Instead of using the profiles of sequentially neighboring residues, we also used the profiles of spatially neighboring residues. In this case, the vector components were arranged in ascending order according to the distance between the spatially neighboring residues. (We called the former, sequentially neighboring residue profiles, and the latter, spatially neighboring residue profiles.)

$$\mathbf{V}_\mathrm{n} = (p_{n\,-\,5\;1}, \ldots, p_{n\,-\,5\;23}, \ldots, p_{n\;1}, \ldots, p_{n\;23}, \ldots, p_{n\,+\,5\;1}, \ldots, p_{n\,+\,5\;23}) \qquad (1)$$

and

$$p_{nj} = \frac{N_{nj}}{\sum_j N_{nj}}$$

where N_{nj} is the number of amino acids j in position n.

Support vector machines

SVMs are supervised learning algorithms proposed by Vapnick (1995). Data examples labeled as positive or negative are projected into a high-dimensional feature space using a kernel, and the hyper-plane in the feature space is optimized to maximize the margin between the positive and negative examples.

We used SVMTorch II (http://old-www.idiap.ch/learning/SVMTorch.html). Only user-defined kernel subroutines were implemented. In this application, linear, polynomial, sigmoid and Gaussian kernels and their sum and product kernels are used. The Gaussian kernel [$\exp(-|a - b|^2/\mathrm{std}^2)$] (std is a parameter) gave the best performance and we therefore report only the results for the Gaussian kernel and its sum and product kernel. In the following discussion, the regularization factor *C* was fixed at 100 and only the parameter 'std' was changed.

Since the SVM optimizes the success ratio for whole sequences but does not optimize the recall and precision (defined below) of interaction sites, prediction performance depends on the ratio of negative and positive data in the learning process. According to the definition, only ~20% of a whole sequence is interaction site residues. If all data are used as learning samples, the prediction result at the default discriminant value (= zero) shows high precision and low recall. Accordingly, half the negative data (non-interaction site residues) were randomly removed from the learning sets when whole sequence residues were used as feature vectors, while a third of the negative data was randomly removed when only surface residues were used as feature vectors in Table I (discussed later). Basically, when the recall–false positive/(false positive + true positive) [recall–FP/(FP + TP)] curves were generated, all the data were used.

Since there were sufficient data for homo–hetero mixed validation (if 3-fold cross-validation was used, the learning time is too long), leave 375 (= 2/3×563) cross-validation was used. For homo and hetero complex validation, 5- and 3-fold cross-validation were used, respectively. In predicting interaction site ratios, 10-fold cross-validation was used for mixed homo and hetero validation data. When no explicit statement is made, 'homo–hetero mixed data' were used.

For homo–hetero mixed validation data, 'filtering by boosting' (Schapire, 1990), which converts a weak learning algorithm into a stronger learning machine, was also applied.

Table I. The recall and precision of each feature vector

Data-type: feature vector[a]	Recall[b] (%)	Precision[c] (%)	Success rate at whole sequence[d] (%)	Success rate at surface[e] (%)
Mix[f]: whole sequence[g] (window 11)	28.8 (22.3)[h]	26.4 (20.0)	69.1 (66.6)	63.5 (60.9)
Mix: whole sequence + boosting by filtering (window 5)	28.8 (21.9)	27.0 (20.0)	69.9 (66.9)	63.7 (61.0)
Mix: whole sequence + actual interaction site ratio (window 5)	35.2 (20.0)	35.8 (20.0)	74.0 (68.0)	68.0 (61.8)
Mix: whole sequence + predicted interaction site ratio (window 5)	28.3 (18.3)	30.7 (20.0)	72.4 (69.0)	65.6 (62.7)
Mix: sequence at surface (window 11)	39.6 (30.4)	40.2 (30.4)	–	63.2 (57.6)
Mix: sequence + ASA (window 9)	41.5 (23.3)	54.9 (30.4)	–	71.4 (60.5)
Mix: spatially neighboring[i] + ASA (15 residues)	44.6 (24.5)	56.1 (30.4)	–	71.0 (60.0)
Mix: spatially neighboring + ASA + actual interaction site ratio (9 residues)	50.4 (26.8)	58.1 (30.4)	–	73.5 (59.1)
Mix: spatially neighboring + ASA + predicted interaction site ratio (9 residues)	42.8 (22.3)	57.8(30.4)	–	73.3 (60.9)
Mix: sequence + ASA + flatness (window 11)	43.2 (24.3)	55.8 (30.4)	–	70.1 (60.1)
Hetero: sequence + ASA (window 9)[j]	45.0 (26.9)	55.9 (32.8)	–	69.7 (57.9)
Homo: sequence + ASA (window 9)[k]	40.3 (21.0)	55.8 (28.9)	–	73.4 (62.2)
Hetero-mixed: sequence + ASA (window 9)[l]	42.4 (24.1)	54.9 (32.8)		71.2 (58.9)
Homo-mixed: sequence + ASA (window 9)[m]	38.4 (21.2)	55.0 (28.9)		72.0 (62.1)

[a] Feature vector = input feature vector of SVM.

[b] Recall = True_Positive/(True_Positive + False_Negative).

[c] Precision = True_Positive/(True_Positive + False_Positive).

[d] 'Success rate at whole sequence' and [e] 'success rate at surface' mean the average per residue prediction (interaction site or non-interaction site) accuracy [= True_Positive + True_Negative)/(the total number of residues)] of whole sequence and sequence at surface, respectively.

[f] Mix = hetero complex + homo complex, the mixed data set was learned and tested.

[g] Sequence = sequentially neighboring residue profiles.

[h] Values in parentheses are randomly predicted ones. The recall of random prediction is calculated as the total predicted residue rate (the total number of predicted residues as interaction sites by SVM/the total number of residues) and the precision of random prediction is calculated as the interaction site ratio of test sets (the total number of interaction site residues/the total number of residues). The random success rate is calculated as [1 – random_precision(whole_sequence or surface)]×(1 – random_recall) + random_precision×random_recall.

[i] Spatially neighboring = spatially neighboring residue profiles.

[j] Hetero and [k] homo: hetero and homo data sets were learned and tested, separately.

[l] Hetero-mixed and [m] homo-mixed: the mixed data set was learned and tested. The precision and recall in each complex type were calculated separately.

This consisted of the following steps. First, the SVM learned using *N* samples (abbreviated as SVM-1). Using SVM-1 and a random number, N/2: wrongly predicted (false negative or false positive) samples and N/2: correctly predicted (true positive or true negative) samples were gathered. They became the learning set for SVM-2 (for details see Schapire, 1990). Next, the *N* samples that were predicted differently by SVM-1 and SVM-2 were collected and these became the learning set for SVM-3. The predictions were decided according to the majority of SVM-1, SVM-2 and SVM-3 predictions. Using this method, 10-fold cross-validation was carried out. The number of learning samples for each SVM with boosting (10-fold cross-validation) was set to be the almost the same as that for SVMs without boosting (leave 1/3 data set cross-validation).

Results and discussion

Predicting interaction sites from sequences

In this section, we discuss how accurately interaction sites can be predicted using only sequence information.

Sequence profiles of sequentially neighboring residues. First, we investigated the performance of the method for predicting interaction sites using the sequence profiles and the profiles of sequentially neighboring residues as feature vectors. In each *n*-window length, the feature vectors were 23×*n* dimensions. A Gaussian kernel was used and the parameter 'std' was set to be sqrt(*n*/13)×5 to account for the dimensions of the vector. Although 3–15 window length profiles were tried, the effect of this window length was quite small. This trend was also observed in other vectors combined with other features. The results for the optimum window length are discussed below.

The results for the recall and precision of window length 11 are summarized in Table I as the feature vector, 'whole sequence'. Here, a true positive means that a residue was correctly recognized as an interaction site, while a false negative means it was erroneously recognized as a non-interaction site. As is discussed in previous reports (Jones and Thornton, 1997b; Zhou and Shan, 2001), there is a possibility of the additional interaction surface existence even if high-resolution data are used. However, it is quite difficult to find them. The actual false-positive rate is probably lower than our evaluation. The numbers in parentheses show the results of random prediction. The recall of random prediction is the total predicted residue rate (the total number of predicted residues as interaction sites by SVM/the total number of residues) and the precision of random prediction is the interaction site ratio of test sets (the total number of interaction site residues/the total number of residues). Although the predicted accuracy was >6% higher than that of random prediction, this level of accuracy still seems too low for practical application.

Although SVMs are known to require small learning sets, there are a wide variety of interaction sites. A shortage of learning samples is likely to reduce prediction performance. To shorten CPU time and ensure effective learning, 'boosting by filtering' was also applied. Figure 1 shows the recall–FP/(FP+TP) curves for sequentially neighboring profiles with a window length of 5 for a whole sequence and their 'boosting by filtering'. (To reduce CPU time, we used a window length of 5 instead of 11.) As shown in Figure 1, prediction performance

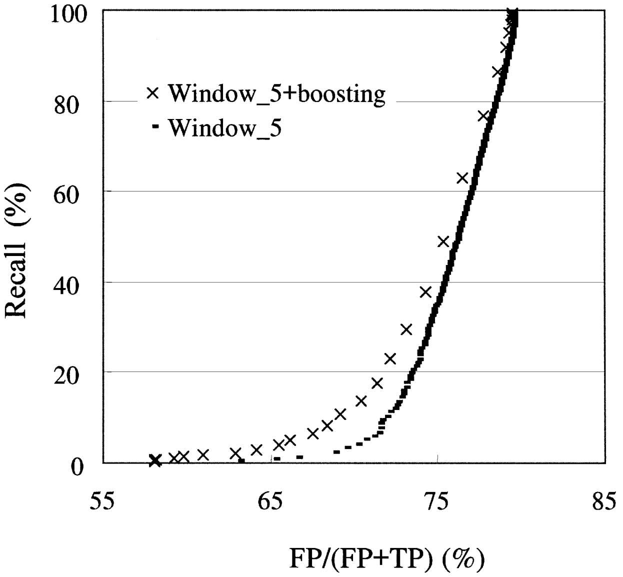

Fig. 1. The recall–FP/(FP+TP) curves of 'sequentially neighboring profiles (window length 5)' and with boosting. The data set is a mix of hetero–homo.

was slightly improved by 'boosting by filtering', especially at a low FP rate. Although the same performance could be achieved using just a single SVM with more learning samples, from the viewpoint of CPU time, 'boosting by filtering' seems useful. The precision of window length 5 + boosting is summarized in Table I ('whole sequence + boosting by filtering'). Unfortunately, since the data set was relatively small, its application to feature vectors with structural information was not carried out in the work described below.

The effect of interaction site ratio. Previous studies showed the different characteristics of interaction sites depending on the interaction surface area (Conte *et al.*, 1999; Glaser *et al.*, 2001). In the small interaction surface area, the hydrophobicity of the interaction sites become closer to or lower than that of the non-interacting surfaces. This is likely to be the reason for the difficulty in predicting interaction sites. In order to consider this trend, interaction site ratio (= 'number of interaction site residues'/'protein sequence length') was added to the sequentially neighboring sequence profiles feature vectors. (As a reference, the relationship between the hydrophobicity of interaction/non-interaction and the interaction site ratio is shown in Supplementary figure 1S available at *PEDS* online.) The results are summarized in Table I as 'whole sequence + actual interaction site ratio'. The kernel was set as K [feature vector = sequentially neighboring profiles + interaction site ratio; Gaussian kernel with std = 5×sqrt(n/13)]. As shown in Table I, this improved the prediction performance. This indicates that taking the interaction ratio into account helps to explain whether or not the residue in focus is on the interaction site. However, in most cases, the interaction site ratio is not known in advance. Accordingly, the prediction of interaction sites using the sequence information is discussed in the next section.

Prediction of the interaction site ratio. To predict the interaction site ratio, whole sequence amino acid distributions (numbers of each amino acid residues) were used as feature vectors and a Gaussian kernel with std = 40 and C = 200 was used. Figure 2 shows the relationship between the predicted and actual interaction site ratios. The standard deviation between the predicted and actual interaction site ratio was 0.15, 0.13 and 0.15 for the mixed, hetero and homo data, respectively. The Pearson product–moment correlation coefficient between actual interaction site ratio and predicted interaction ratio was 0.52, 0.51 and 0.56 for the mixed, hetero and homo data, respectively. Although the accuracy was not high, it was possible to predict the interaction site ratios to some extent using only the amino acid slant. When continuous amino acid usage (400 dimensions) was used, prediction accuracy did not improve (data not shown). This prediction method may be useful for detecting protein–protein interaction slants due to experimental features.

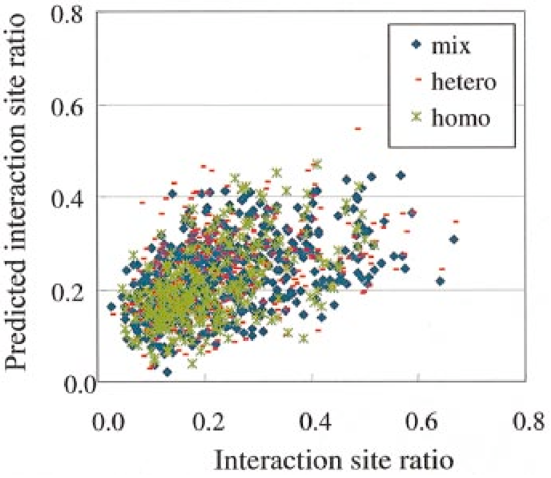

Fig. 2. Relationship between actual interaction site ratio and predicted interaction site ratio.

The effect of predicted interaction site ratios. Since the data set was quite small, 9/10 of the data set was used for the prediction of the interaction site ratio; 1/3 of the data set (included in the 9/10 of the data set) was used for SVM learning for interaction site prediction using the actual interaction site ratio and sequentially neighboring profiles. The remaining 1/10 of the data was used for the test set, i.e. the interaction site ratio for 1/10 of the data was predicted by SVM regression first, and this information was then used to predict the interaction sites. These steps were repeated 10 times. The recall and precision are summarized in Table I as 'whole sequence + predicted interaction site ratio'. The recall–FP/(FP+TP) curves of 'sequentially neighboring profiles', 'sequentially neighboring profiles + actual interaction site ratio' and 'sequentially neighboring profiles + predicted interaction site ratio' are plotted in Figure 3. Although the recall–FP/(FP+TP) curve for the predicted interaction site ratio is lower than that for the actual interaction site ratio, it is higher than that for only sequence profiles. Since the data set was small, the learning for interaction site prediction was carried out using the actual interaction site ratio and the test was carried out using the predicted interaction site ratio. If the predicted interaction site ratios are used for the learning steps, the performance may be improved.

Prediction of interaction sites from structural information

Sequence profiles of sequentially neighboring residues and ASA. Since the accessible surface area is useful for understanding the environmental state of an amino acid

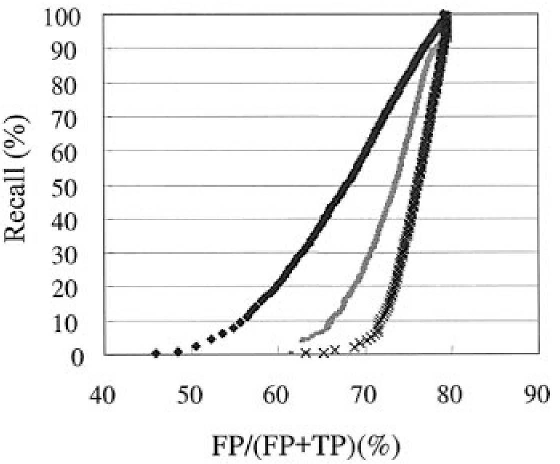

Fig. 3. The recall–FP/(FP+TP) curves of 'sequentially neighboring profiles + actual interaction site' (diamonds), 'sequentially neighboring profiles + predicted interaction site' (–) and 'sequentially neighboring profiles' (×). All feature vectors were calculated by window length 5.

residue, the sequence profile of sequentially neighboring residues and the relative accessible surface areas [ratio of surface area of each residue to maximum area (%); see Materials and methods, henceforward abbreviated to ASA] as calculated by the DSSP (Rost and Sander, 1993) were used as feature vectors. In each n-window length test set, the dimensions of the feature vectors were $23 \times n + n$. The kernel was set at K [feature vector = sequence profiles; Gaussian kernel with std = $5 \times \mathrm{sqrt}(n/13)$]$\times K$ [feature vector = ASA; Gaussian kernel with std = $5 \times \mathrm{sqrt}(n/13)$]. The recall and precision are given in Table I ('sequence + ASA'). Since the total success rate for molecular surfaces was optimized by machine learning, the recall was low in Table I. When the discriminant value (default was zero) of the SVM was changed, the recall rose and the precision fell. For example, when the recall was 50.0%, the precision became 49.0% for homo–hetero mixed types. Many of the residues wrongly predicted as interaction sites were located at sites neighboring interaction sites; 32.1% of wrongly predicted residues at recall = 50% for mixed types were the closest spatially/sequentially neighboring residues of interaction site residues.

The effect of ASA was remarkable compared with the results obtained using the feature vector of sequence profiles for surfaces in Table I ('sequence at surface'). When only ASAs were used as feature vector components, recall and precision were lower than random predictions. This indicates that a combination of sequence profiles and ASA is required. In Table I, 'hetero' and 'homo' show separate training results, while 'hetero-mixed' and 'homo-mixed' show mixed training results. The prediction accuracy changes only slightly, even if hetero and homo complexes are learned individually. The trend is also observed in other feature vectors. This indicates that the interaction site characteristics for hetero and complex are almost the same.

If the ASA is predicted by regression of the SVM using sequence profiles, the prediction of interaction sites using 'predicted ASA + sequence profiles' is also possible using only sequence information. However, since the ASA prediction performance is quite low (if std = 40 and window length 5 profiles are used, the mean absolute error is 22.3, and the squared standard deviation is 1416 in percent units), the improvement in prediction performance is similar to the effect of 'boosting by filtering', as shown in Figure 1 (data not shown).

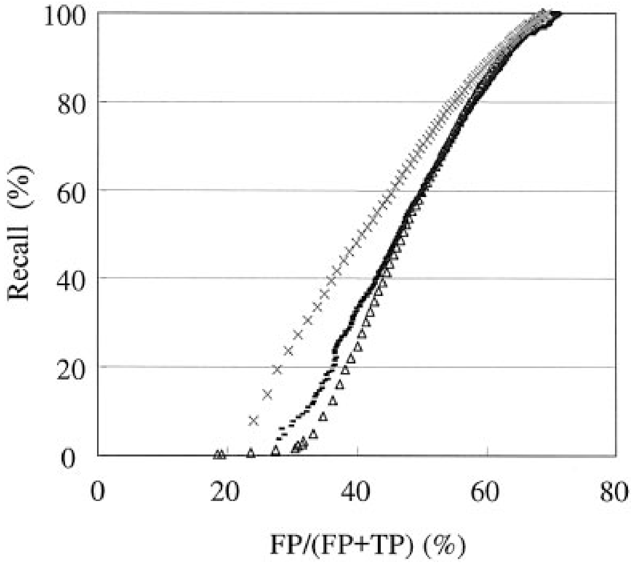

Fig. 4. The recall–FP/(FP+TP) curves of 'spatially neighboring profiles + ASA + actual interaction site ratios' (×), 'spatially neighboring profiles + ASA + predicted interaction site ratios' (–) and 'spatially neighboring profiles +ASA' (triangles). All feature vectors were calculated by spatially neighboring 9 residues.

Sequence profiles of spatially neighboring residues and ASA. When the sequence profiles of spatially neighboring residues are used instead of those of sequentially neighboring residues, the prediction performance is slightly improved. The recall and precision are summarized in Table I as 'spatially neighboring + ASA'. When recall was 50.0%, precision was 54.2% for mixed types. The corresponding recall and precision given by randomly predicted values for mixed types were 28.0 and 30.4%, respectively. Of the wrongly predicted residues for mixed types, 28.1 and 37.1% were located within the closest or next closest sequentially/spatially neighboring residues of the interaction sites, respectively. These predicted interaction sites comprised 86.0% of the actual interfaces (1169 interfaces). If interfaces consisting of <20 residues were ignored, the predicted interaction sites comprised 96.5% of the actual interfaces (626 interfaces). Given these values, this method seems to be useful for actual prediction.

The effect of interaction site ratio. As discussed above, the interaction site characteristics depend on the interaction site ratios and the addition of interaction site ratios to feature vectors improves prediction performance. Also in the case of feature vectors with spatially neighboring profiles + ASA, the prediction performance is increased by the addition of the actual interaction site ratio. The results are summarized in Table I as 'spatially neighboring + ASA + actual interaction site ratios' (hereafter abbreviated as 'actual interaction site ratio'). However, in most cases, the interaction site ratio is not known in advance, and the interaction site ratios were also predicted by SVM regression and amino acid compositions as described above. The 10-fold cross-validation using predicted interaction site ratio was carried out similarly to previous

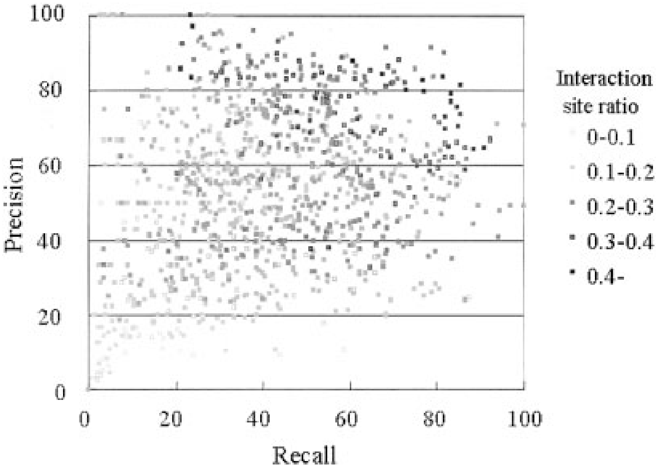

Fig. 5. The precision and recall of each complex (feature vector: spatially neighboring profiles + ASA of 15 residues).

section (see section 'The effect of predicted interaction site ratios'). The results are summarized in Table I as 'spatially neighboring + ASA + predicted interaction site ratios' (hereafter abbreviated as 'predicted interaction site ratio'). Their recall–FP/(FP+TP) curves are plotted in Figure 4. When feature vectors, including actual interaction site ratios, were taken into account, the increase in recall was notable, especially at a low FP rate. When predicted interaction site ratios were used, the prediction performance worsened compared with using the 'actual interaction site ratios', but was slightly higher than using the 'predicted interaction site ratio' at a low FP rate (~30–40%). Unfortunately, the effect of the predicted interaction site ratio was smaller than that of 'sequentially neighboring profiles', as discussed above. The precision of the 'actual interaction site' at recall = 50% (random 26.1%) was 58.3% (30.4%), while that of the 'predicted interaction site' at recall = 50% (random 25.5%) was 56.0% (30.4%). They made up 85.5% (93.5% of interfaces with over 20 residues) (actual interaction site) and 86.7% (97.4%) (predicted interaction site) of the actual interfaces (1169 interfaces), respectively. The 32.7 and 28.7% of wrongly predicted residues for 'actual interaction sites' and 'predicted interaction sites' were the closest spatially/sequentially neighboring residues, respectively.

If there are sufficient data, it may be better to use the predicted interaction site ratio for the learning steps. The larger data set will probably improve the prediction performance for interaction site ratios and interaction sites.

Other characteristics

To improve prediction performance, we examined the influence of various parameters. To use information about neighboring residues, a patch consisting of a residue and the 10 closest residues was considered. Patch flatness, as defined by Jones and Thornton (1997a) (i.e. the root-mean-square deviation of 11 residues from the least squares plane of 11 residues) was added to the feature vectors of the sequentially neighboring residue profiles + ASA. The results are summarized in Table I ('sequence + ASA + flatness'). The least squares method of finding the surface, which minimizes the sum of the squared distance, from 11 residues in the patch was resolved using Octave (http://www.octave.org). The results were slightly higher than those for 'sequentially neighboring residue profiles + ASA' (Table I, 'sequence + ASA').

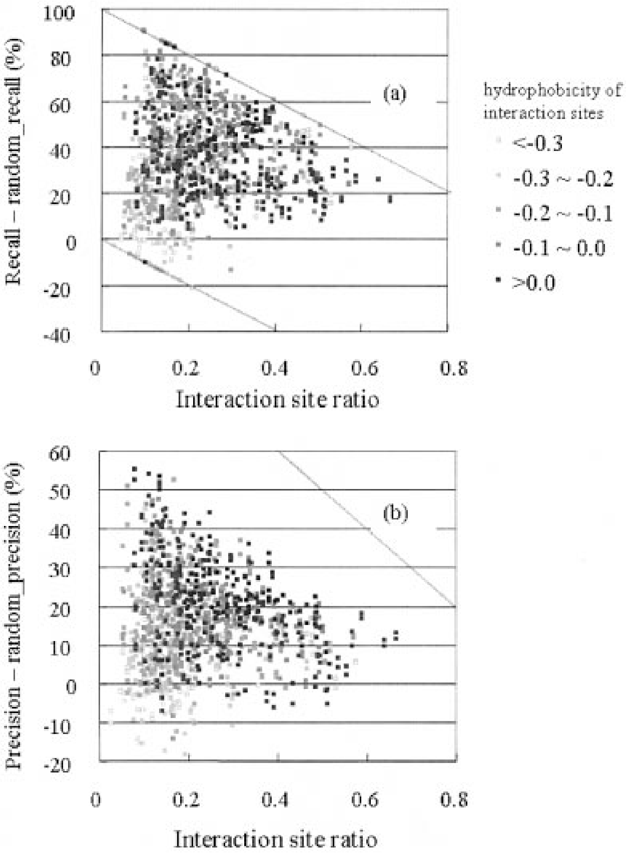

Fig. 6. (**a**) Relationship between interaction site ratio and precision–random_precision of each protein (feature vector: spatially neighboring profiles + ASA of 15 residues). The dark color shows higher hydrophobic interfaces and the light color shows less hydrophobic (hydrophilic) interfaces. The possible range of precision–random_precision (i.e. 1, random precision; 0, random precision) is indicated by the dotted line. (**b**) Relationship between interaction site ratio and recall–random_recall of each protein sequence. The possible range of recall–random_recall is indicated by the dotted line.

Although the total amino acid ratio/length of the target sequence, the conservation score for the residues (Valdar and Thornton, 2001), amino acid ratio in patches and the hydrophobicity of the sequences were added to the feature vectors of sequentially/spatially neighboring residue profiles, there was no increase in prediction accuracy. These values may not be important in interaction site prediction or the kernel used may not be appropriate for this combination of vectors.

To summarize, two feature vectors, 'spatially neighboring residue profiles + ASA + with/without predicted interaction site ratios' and 'sequentially neighboring residue profiles + ASA + patch flatness', were useful in predicting interaction sites and the former was better. When 'actual interaction site ratios' were known in advance, the prediction performance of 'spatially neighboring residue profiles + ASA + actual interaction site ratios' was best.

The relationship between prediction accuracy and protein characteristics

Although the averaged prediction performance is discussed above, the prediction accuracy was quite different among proteins. In this section, we discuss the predictable and unpredictable interaction sites. Figure 5 shows the distribution of the recall and precision of each complex for 'spatially neighboring + ASA'. The light color in Figure 5 indicates complexes with low interaction site ratios. The precision and recall are higher for molecules with higher interaction site ratios. With regard to random predictions, Figure 6a and b shows the relationship between precision–random_precision, recall–random_recall, interaction site ratio and hydrophobicity. The hydrophobicity of interaction sites was calculated using the amino acid index as in the following equation:

$$\sum_{a} p_a \cdot h_a \qquad (2)$$

where p_a is the ratio of amino acid 'a' in the interaction sites and h_a is the hydrophobic value of amino acid 'a' as defined by Sweet and Eisenberg (1983). The light color in Figure 6 indicates low hydrophobic interaction sites. Overall, the prediction of hydrophobic interaction seems relatively easy except for some kinds of proteins. (For example, the averaged recall–random_recall and precision–random_precision for proteins with hydrophobicity >0 are 0.21 and 0.35, respectively. These are larger than the averaged values of the whole data, 0.20 and 0.26.) Strictly speaking, comparisons between the prediction performance for different interaction site ratios are difficult, considering the possible range of recall/precision–random values. There was a trend for proteins with hydrophilic (low hydrophobic) interaction sites and low interaction site ratios to have lower recall–random_recall and precision–random_precision (e.g. the averaged recall–random_recall and precision–random_precision for proteins with hydrophobicity <–0.3 and interaction ratio <0.2, were 0.04 and 0.15, respectively. These are considerably less than the averaged values of the whole data, i.e. 0.20 and 0.26). As this suggests, it seems difficult to locate interaction sites in proteins with hydrophilic interaction and low interaction site ratios with this method. Proteins with low hydrophobic (hydrophilic) interfaces and low interaction site ratios are considered 'transient complexes' based on the previous study (Jones and Thornton, 1996). However, when the automatic method developed by Ofran and Rost (2003a) to distinguish transient and permanent complexes is used, they are not recognized as mainly 'transient complexes'. When the biological functions for the lowest 10 recall–random_recall proteins with an interaction site ratio of <0.3 were investigated manually from references and their single molecule (not complexes) existence was investigated by searching a similar structure according to the FSSP server [http://www.ebi.ac.uk/dari/fssp/, the checked residues are indicated by (–) in the Appendix protein lists], they all seemed to be transient complex. In contrast, when 10 highly (recall–random_recall + precision–random_precision) predicted proteins with high hydrophobic (>0.0) and low interaction site ratio (<0.2) residues were selected and investigated, nine proteins consisted of homo-dimers and at least eight proteins seemed to be permanent complexes *in vivo* [checked residues are indicated by (+) in the Appendix protein lists, available as supplementary material at *PEDS* online]. Furthermore, when 10 highly (recall–random_recall + precision–random_precision) predicted proteins with an interaction site ratio of >0.5 and hydrophobicity of >0.0 were selected and investigated, at least seven proteins were permanent complexes [checked residues are indicated by (++) in the Appendix protein lists]. The prediction of permanent/stable complexes tends to be easier. The experimental reproduction of some complexes, especially in signal transduction, is known to be difficult. The easiness of prediction may be related to the interaction strength.

The typical predicted interaction sites are shown in Figure 7a and b. Most of the false positives (yellow) are close to interaction sites. With manual investigation, even for permanent complexes with hydrophobic interfaces and high interaction site ratios, interaction site prediction is difficult when the complex consists of multiple entangled molecules. For example, red parts of 2OCC-C in Figure 7c were wrongly predicted as interaction sites. The characteristics of the inside interfaces (domain interfaces) in single molecules may be sometimes similar to that of interaction sites.

Comparison with other studies

There have been several studies on the prediction of interfacial sites (Korn and Burnett, 1991; Young *et al.*, 1994; Jones and Thornton, 1997b; Zhou and Shan, 2001; Fariselli *et al.*, 2002; Gallet *et al.*, 2002; Ofran and Rost, 2003b). This method is similar to the use of neural networks to make predictions (Zhou and Shan, 2001; Fariselli *et al.*, 2002; Ofran and Rost, 2003b). Fariselli *et al.* (2002) show that the recall and precision of interaction sites on the surface are 72 and 56%, respectively, using sequence profiles. When we used the same test set using only sequentially neighboring residue profiles for the feature vectors for the SVM, and when other conditions such as the definition of the interaction site residues were the same, the recall and precision for the residue units were 75.9 and 78.3%, respectively. These values were higher than theirs. They excluded the protease complex to eliminate a strong peculiar signal in the data set. Whether the protease complex was eliminated or not did not affect the prediction performance of our method.

Zhou and Shan (2001) predicted interfacial interaction sites using neural networks in a two-step process. The input values are sequence profiles and ASA of the residue in focus and are the same for the 19 spatially closest surface residues. The recall and precision for the residue units are 50.0 and 51.0%, respectively. (When the four nearest neighbor differences are added to the correct, the modified precision becomes 70%, and using the same evaluation, 76.4% in our results. The corresponding recall may become lower than 50%. With the feature vector, spatially neighboring profiles + ASA, their randomly predicted values are similar to our data.) They removed sequence similarity with a threshold of matched regions of >40%. However, this threshold seems slightly permissive. They counted only one interface, even if the complex consisted of more than two proteins. Their actual recall may be lower than 50% and the precision may be higher than 51%. Although their accuracy cannot be compared directly with our data, our prediction method seemed to perform slightly more accurately than theirs.

Recently, Ofran and Rost (2003b) predicted interaction sites using sequentially neighboring profiles and neural networks. By considering neighboring predicted information (i.e. the contacting residues are sequentially continuous in most cases) and using strongly predicted residues, a high precision (>75%) was achieved at a low recall range (<0.2%). In our methods, when 'spatially neighboring + ASA' and the sum of the positively predicted values within ±3 residues were sorted in descending order, the precision reached 80–85% at a recall <0.2% (in this process of using positively predicted residues as interaction sites, only residues that had more than two positively predicted residues in ±3 sequentially neighboring residues were used). However, a high precision was not obtained using only sequence information. Various studies of the performance of SVM and neural networks may suggest that this might be caused by differences in the definition of interaction sites (their random precision is ~0.4, while our random precision is ~0.2) and/or the size of the data set. By adding the predicted interaction site ratio, the average of our prediction performance (~10% higher than random) seems to be higher than theirs (~4–8% higher than random).

Interfacial site prediction using the hydrophobic moment and averaged hydrophobicity has also been reported (Gallet *et al*., 2002). However, this study did not consider false positives and was applied to a limited number of families. When the hydrophobic moment and averaged hydrophobicity at window lengths 5–11 were used as the input vectors for the SVM, the recall and precision were lower than for random prediction (data not shown). Accordingly, the use of the hydrophobic moment and averaged hydrophobicity does not appear to be useful for general interfacial prediction.

Conclusion

Methods of predicting interaction sites using a SVM were developed. When sequence profiles, sequentially neighboring residue profiles and actual interaction site ratios were used as feature vectors for the SVM, the recall and precision of the interaction sites were >15% higher than randomly predicted values. When predicted interaction site ratios, calculated using SVM regression and amino acid compositions, were used instead of actual interaction site ratios, the recall and precision of the interaction sites were ~10% higher than randomly predicted values and ~4% higher than those calculated using only sequence profiles. Although prediction performance using only sequence information may not be high enough for practical use, it is interesting that the effect of interaction site ratios on prediction performance is remarkable. The performance of this method could be improved if the target families were limited.

When structural information was used, prediction performance was improved. In particular, when 'spatially neighboring residue profiles + ASA + with/without predicted interactions site ratios by SVM regression and amino acid compositions' or 'sequentially neighboring residue profiles + ASA + patch flatness' were used as feature vectors, prediction performance was high; the former was better. When actual interaction site ratios were used instead of predicted interaction site ratios, prediction performance was further improved. Other sequence characteristics, such as the conservation score and amino acid ratio in patches, did not improve prediction performance. When the feature vector 'spatially neighboring residue profiles + ASA + with/without predicted interaction site ratios' was

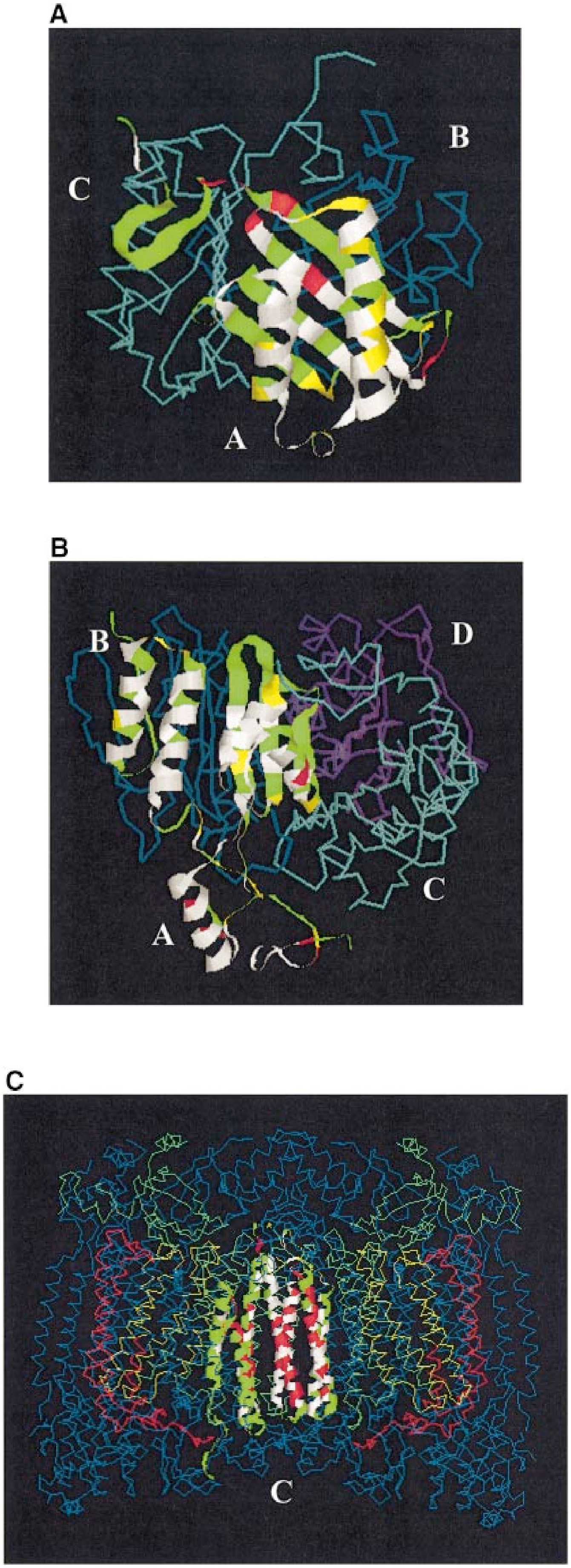

Fig. 7. Overall view of complexes with predicted interaction sites. The protein with predicted interaction sites is shown as a ribbon (green, true positive; white, true negative; yellow, false negative; red, false positive) and their binding proteins are shown as strands. (**A**) 1GD0 (human macrophage migration inhibitory factor): the predicted protein (A), the binding proteins (B and C). (**B**) 1APY (human aspartylglucosaminidase): the predicted protein (A), the binding proteins (B, C and D). (**C**) 2OCC (bovine heart cytochrome *c* oxidase): the predicted protein (C), the binding/non-binding proteins (A–B, D–Z, mol IDs not shown). These figures were prepared by using the program RasMol.

used, the precision was 54–56% for homo–hetero mixed types at recall = 50%, whereas the corresponding recall and precision given by randomly predicted values were 25–28 and 30%, respectively. Approximately 30% of sites wrongly predicted as interaction sites were the closest spatially/sequentially neighboring residues on interaction sites. These predicted interaction sites covered 86–87% (96–97% when interfaces of <20 residues were ignored) of the actual interfaces (1169 interfaces). The prediction performance for this method was slightly better than that of previously reported prediction methods (Zhou and Shan, 2001; Fariselli *et al*., 2002). Considering that only ~30% of protein surfaces consist of interaction sites, this prediction performance seems reasonable. The data set in this study may have been too small for two-step learning of interaction site ratios and interaction sites. The performance of the method using predicted interaction site ratios will probably be improved by increasing the size of the data set.

Prediction accuracy is low for complexes with low interaction site ratios (probably with small interaction surfaces) and hydrophilic (low hydrophobic) interaction sites, and high for complexes with high hydrophobic interaction sites. With manual investigation, permanent complexes tend to be easy to predict and transient complexes difficult. The ease of prediction may be in proportion to the stability of the complex. In relation to this point, elucidation of the relationships between the prediction ease and binding stability such as binding energy will be the subject of future research.

Acknowledgements

This work is supported in part by a Grant-in-Aid for scientific research on priority area (C) Genome Information Science, from the Japanese Ministry of Education, Culture, Sports, Science, and Technology.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Argos,P. (1988) *Protein Eng.*, **2**, 101–113.

Conte,L.L., Chothia,C. and Janin,J. (1999) *J. Mol. Biol.*, **285**, 2177–2198.

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

Fariselli,P., Pazos,F., Valencia,A. and Casadio,R. (2002) *Eur. J. Biochem.*, **269**, 1356–1361.

Gallet,X., Charloteaux,B., Thomas,A. and Brasseur,R. (2002) *J. Mol. Biol.*, **302**, 917–926.

Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) *Proteins*, **43**, 89–102.

Jones,S. and Thornton,J.M. (1995) *Prog. Biophys. Mol. Biol.*, **63**, 31–65.

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

Jones,S. and Thornton J.M. (1997a) *J. Mol. Biol.*, **272**, 121–132.

Jones,S. and Thornton,J.M. (1997b) *J. Mol. Biol.*, **272**, 133–143.

Korn,A.P. and Burnett,R.M. (1991) *Proteins*, **9**, 37–55.

Nooren,I.M.A. and Thornton,J.M. (2003) *J. Mol. Biol.*, **325**, 991–1018.

Ofran,Y. and Rost,B. (2003a) *J. Mol. Biol.*, **325**, 377–387.

Ofran,Y. and Rost,B. (2003b) *FEBS Lett.*, **544**, 236–239.

Rost,B. (1999) *Protein Eng.*, **12**, 85–94.

Rost,B. and Sander,C. (1993) *J. Mol. Biol.*, **232**, 584–599.

Schapire,R.E. (1990) *Mach. Learn.*, **5**, 197–227.

Sweet,R.M. and Eisenberg,D. (1983) *J. Mol. Biol.*, **171**, 479–488.

Tsai,C.J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1997) *Protein Sci.*, **6**, 53–64.

Valdar,W.S. and Thornton,J.M. (2001) *Proteins*, **42**, 108–124.

Vapnick,V. (1995) *The Nature of Satistical Learning Theory*. Springer, New York.

Young,L., Jernigan,R.L. and Covell,D.G. (1994) *Protein Sci.*, **3**, 717–729.

Zhou,H.-X. and Shan,Y. (2001) *Proteins*, **44**, 336–343.

Received October 10, 2003; revised December 25, 2003; accepted December 31, 2003

Edited by Gideon Schreiber