BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i371–i378
DOI: 10.1093/bioinformatics/bth920

# A two-stage classifier for identification of protein–protein interface residues

Changhui Yan[1,2,5,*], Drena Dobbs[1,3,4,5] and Vasant Honavar[1,2,4,5,6]

[1]Artificial Intelligence Research Laboratory, [2]Department of Computer Science, [3]Department of Genetics, Development and Cell Biology, [4]Laurence H Baker Center for Bioinformatics and Biological Statistics, [5]Bioinformatics and Computational Biology Graduate Program and [6]Computational Intelligence, Learning, and Discovery Program, Iowa State University, Ames, IA, 50010, USA

Received on January 15, 2004; accepted on March 1, 2004

**ABSTRACT**

**Motivation:** The ability to identify protein–protein interaction sites and to detect specific amino acid residues that contribute to the specificity and affinity of protein interactions has important implications for problems ranging from rational drug design to analysis of metabolic and signal transduction networks.

**Results:** We have developed a two-stage method consisting of a support vector machine (SVM) and a Bayesian classifier for predicting surface residues of a protein that participate in protein–protein interactions. This approach exploits the fact that interface residues tend to form clusters in the primary amino acid sequence. Our results show that the proposed two-stage classifier outperforms previously published sequence-based methods for predicting interface residues. We also present results obtained using the two-stage classifier on an independent test set of seven CAPRI (Critical Assessment of PRedicted Interactions) targets. The success of the predictions is validated by examining the predictions in the context of the three-dimensional structures of protein complexes.

**Contact:** chhyan@iastate.edu

**Supplementary information:** http://www.public.iastate.edu/~chhyan/ISMB2004/list.html

## 1 INTRODUCTION

Protein–protein interactions play a pivotal role in protein function. Completion of many genomes is being followed rapidly by large-scale efforts to identify interacting protein pairs experimentally, in order to decipher the networks of interacting proteins. Experimental proteomics projects have already resulted in complete 'interactomes' (Ho *et al*., 2002; Giot *et al*., 2003; Li *et al*., 2004). While such efforts yield a catalog of interacting proteins, experimental detection of residues in protein–protein interaction surfaces must come from determination of the structure of protein–protein complexes. However, determination of protein–complex structures using X-ray and NMR methods lags far behind the number of known protein sequences. Hence, there is a need for the development of reliable computational methods for identifying protein–protein interface residues (Teichmann *et al*., 2001; Valencia and Pazos, 2002, 2003). Identification of protein–protein interaction sites and detection of specific amino acid residues that contribute to the specificity and strength of protein interactions is an important problem with broad applications ranging from rational drug design to the analysis of metabolic and signal transduction networks.

Protein–protein interfaces have been a topic of study for several years (Chothia and Janin, 1975; Jones and Thornton, 1996; Lo Conte *et al*., 1999; Ofran and Rost, 2003a). Based on the different characteristics of known protein–protein interaction sites, several methods have been proposed for predicting these sites. These include methods based on the presence of 'proline brackets' (Kini and Evans, 1996), patch analysis using a six-parameter scoring function (Jones and Thornton, 1997), properties associated with interface topology (Valdar and Thornton, 2001), analysis of the hydrophobicity distribution around a target residue (Gallet *et al*., 2000), charge distribution on interfaces (Sheinerman and Honig, 2002), multiple sequence alignments (Pazos *et al*., 1997; Valencia and Pazos, 2003), structure-based multimeric threading (Lu *et al*., 2003), docking methods (Halperin *et al*., 2002), using potentials that describe protein–protein interactions (Keskin *et al*., 1998) and analysis of characteristics of spatial neighbors of a target residue using neural networks (Zhou and Shan, 2001; Fariselli *et al*., 2002; Ofran and Rost, 2003b). Our recent work has focused on an analysis of sequence neighbors of a target residue using an support vector machine (SVM) method (Yan *et al*., 2003).

In our previous report, we used an SVM to identify interface residues using sequence neighbors of a target residue

*To whom correspondence should be addressed.

(Yan *et al.*, 2003). Here, we report a two-stage classifier consisting of an SVM and a Bayesian network classifier that identifies interface residues primarily on the basis of sequence information. The two-stage method achieved 72% accuracy with a correlation coefficient of 0.3 when tested on a set of 77 proteins using 5-fold cross-validations. Experiments on the same dataset demonstrated that the two-stage method outperforms the previously published sequenced-based method of Gallet *et al.* (2000).

CAPRI (http://capri.ebi.ac.uk/) is a community-wide experiment to assess the capacity of protein-docking methods to predict protein–protein interactions. In each round of CAPRI, structures of protein–protein complexes are predicted based on structures of the unbound components. CAPRI targets present interesting test cases for evaluation of computational methods for prediction of interface residues. A two-stage classifier which was trained using the 77 proteins in our dataset was tested on CAPRI targets. The results were evaluated in the context of three-dimensional structures of protein complexes.

## 2 METHODS

### 2.1 Datasets

We extracted individual proteins from a set of 70 protein–protein heterocomplexes used in the study of Chakrabarti and Janin (2002). After removal of redundant proteins and molecules with fewer than 10 residues, we obtained a dataset of 77 individual proteins with sequence identity <30%. These proteins represent six different categories of protein–protein interfaces, classified according to the scheme of Chakrabarti and Janin (2002). The six categories and the number of representatives in each category are: antibody–antigen (13), protease-inhibitor (11), enzyme complexes (13), large protease complexes (7), G-proteins, cell cycle, signal transduction (16) and miscellaneous (17). Because of the low level of sequence identity, the resulting dataset is more challenging than the datasets used in previous studies by our group (Yan *et al.*, 2003) as well as by other authors (Ofran and Rost, 2003b). The list of the 77 proteins is available at http://www.public.iastate.edu/∼chhyan/ISMB2004/list.html.

### 2.2 Definition of surface residue and interface residues

The definition of interface residues used in this study is based on the reduction of solvent accessible surface area (ASA) upon complex formation. ASA was computed for each residue in the unbound molecule (MASA) and in the complex (CASA) using the DSSP program (Kabsch and Sander, 1983). A residue is defined to be a surface residue if its MASA is at least 25% of its nominal maximum area as defined by Rost and Sander (1994). A surface residue is defined to be an interface residue if its calculated ASA in the complex is less than that in the monomer by at least 1 Å$^2$ (Jones and Thornton, 1996). Surface residues were extracted and divided into interface residues and non-interface residues, using structural information from Protein Data Bank (PDB) files. We obtained a total of 2340 positive examples corresponding to interface residues and 5091 negative examples corresponding to non-interface residues.

### 2.3 Analysis of interface residue neighborhoods

Let $P_{\text{actual}}$ be the observed probability that a given neighbor of an interface residue is also an interface residue. Let $P_{\text{background}}$ be the probability that this position has an interface residue by chance. The log-likelihood of the residue for this position belonging to an interface is given by $\log_2(P_{\text{actual}}/P_{\text{background}})$. Positive values for likelihood indicate that the residue under consideration has probability greater than that expected by chance of being an interface residue. Negative likelihood indicates the opposite. A likelihood of 0 indicates that the probability of the residue likely to be an interface residue is the same as what we would expect simply based on the fraction of residues in the dataset that are interface residues.

### 2.4 The two-stage classifier

In designing the two-stage classifier, we exploit the observation that interface residues tend to form clusters on amino acid sequence (Ofran and Rost, 2003b). In the first stage, an SVM classifier is trained to identify interface residues based on the identities of neighboring residues of the target residue. The input to the SVM is an encoding of the identities of nine contiguous amino acid residues, corresponding to a window containing the target residue and four neighboring residues on either side of the target residue. Each of the 9 residues in the window is represented by a 20-bit vector (with 1-bit for each letter of the 20-letter amino acid alphabet). Thus, the SVM classifier accepts $9 \times 20 = 180$-bit vector as input and produces a Boolean output (with 1 denoting an interface residue and 0 denoting a non-interface residue). Our study used the SVM in the Weka package from the University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/~ml/weka/). (Witten and Frank, 1999). The package implements Platt's (1998) sequential minimal optimization (SMO) algorithm for training a support vector classifier using scaled polynomial kernels.

In the second stage, a Bayesian network classifier is trained to identify interface residues based on the class labels (1 for interface or 0 for non-interface) of its neighbors. The inputs for Bayesian classifier are the class labels of the eight residues surrounding the target residue (four on each side). The Bayesian network classifier is trained to output the most likely class label for the target residue given the class labels of its neighboring residues. We used the

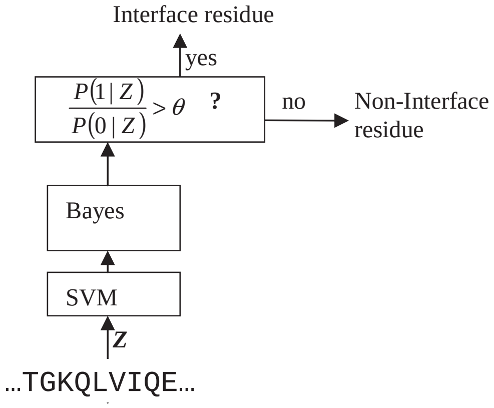

**Fig. 1.** The schematic diagram of the two-stage classifier.

BayesNetB from the Weka package, which implements hill climbing algorithm to learn the Bayesian network structure (Buntine, 1991). (We found that on this dataset, the Naïve Bayes classifier performs as well as a more complex classifier that models the dependencies among the neighboring residues.)

Let $\boldsymbol{C}$ be a Binary random variable that denotes the class label (1 for an interface residue, 0 for a non-interface residue) for the target residue. Let $\boldsymbol{Z}$ be a vector-valued random variable that denotes the input to the two-stage classifier (i.e. a Binary encoding of the target residue and its sequence neighbors). The two-stage classifier classifies the target residue as an interface residue if

$$\frac{P(1|z)}{P(0|z)} > \theta.$$

The schematic diagram of the two-stage classifier is shown in Figure 1. If $\theta = 1$, this procedure corresponds to assigning the most probable class label (maximum a posteriori classification) for the target residue. Varying $\theta$ corresponds to trading off specificity against sensitivity of interface residue prediction (Fig. 4 under Experiments and Results section). We choose $\theta$ so as to maximize the correlation coefficient (see below), which measures the agreement between the actual and predicted class labels on the training data. The resulting classifier is then used to predict whether or not a target residue is likely to be an interface residue based on its identity and the identities of its eight-sequence neighbors.

## 2.5 Performance measures

Let TP is the number of true positives (residues predicted to be interface residues that actually are interface residues); FP the number of false positives (residues predicted to be interface residues that are in fact not interface residues); TN the number of true negatives; FN the number of false negatives; $N = \text{TP} + \text{TN} + \text{FP} + \text{FN}$ (the total number of examples). Then we have:

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

$$\text{Specificity}^{+} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{N}, \text{ and}$$

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

Sensitivity$^{+}$ (sensitivity for interface residue class) measures the fraction of interface residues that are identified as such. Specificity$^{+}$ (specificity for the interface residue class) measures the fraction of the predicted interface residues that are actually interface residues. Accuracy of a classifier measures the estimated probability of correct predictions. Correlation coefficient (CC) is a measure of how well the predicted class labels correlate with the actual class labels. It ranges from −1 to 1 where a correlation coefficient of 1 corresponds to perfect predictions, and a correlation coefficient of 0 corresponds to random guessing. Note that the commonly used measure of accuracy is not a particularly useful measure for evaluating the effectiveness of a classifier when the distribution of samples over different classes is unbalanced (Baldi *et al.*, 2000). Average values of specificity and sensitivity are given by

$$\text{Average specificity} = \tfrac{1}{2}(\text{Specificity}^{+} + \text{Specificity}^{-}),$$

$$\text{Average sensitivity} = \tfrac{1}{2}(\text{Sensitivity}^{+} + \text{Sensitivity}^{-}).$$

# 3 EXPERIMENTS AND RESULTS

## 3.1 Interface residues tend to form clusters on amino acid sequences

Ofran *et al.* (2003b) investigated the sequence neighborhood of protein–protein interface residues in a set of 333 proteins and reported that 98% of protein–protein interface residues have at least one additional interface residue within 4 positions of N- or C-terminal and 74% have at least 4. Among the 77 proteins we used here, 44 are also in the Ofran dataset. For the 77 proteins, we obtained similar results: 97% of interface residues have at least one additional interface residue, and 70% of the interface residues have at least 4 interface residues within 4 positions on either side. For each interface residue, we analyzed the likelihood that its sequence neighbors are also interface residues. The results are shown in Figure 2. Close neighbors of an interface residue have a high likelihood of being interface residues. The closer a sequence neighbor is to an interface residue, the greater is its likelihood of being an interface residue. When the distance increases to

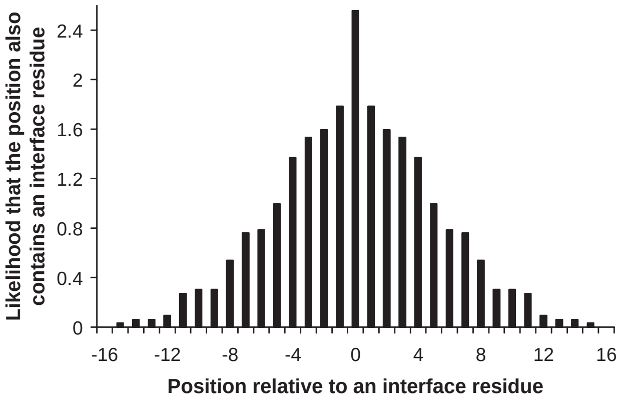

**Fig. 2.** The likelihood that positions neighboring interface residues also contains interface residues. Position 0 is an interface residue. Negative positions are on the N-terminal side of this target residue, positive positions are on the C-terminal. Positive likelihood means that the position has higher probability than random of also being an interface residue.

16 residues, the likelihood drops to 0. The observation that the interface residues tend to form clusters on the primary sequence suggests the possibility of detecting protein–protein interface residues from local sequence information.

Based on the results shown in Figure 2, a window size of nine contiguous residues centered on the target residue was empirically determined to be optimal (data not shown) for constructing the two-stage classifier.

## 3.2 Classification of surface residues from 77 proteins into interface residues and non-interface residues

The two-stage classifier was evaluated using the dataset of 77 proteins in a 5-fold cross-validation experiment. Table 1 shows the classification performance as measured by correlation coefficient, accuracy, specificity$^+$ and sensitivity$^+$. The correlation coefficient was maximized by choosing $\theta = 1$. The resulting classifier achieved an overall accuracy of 72% with a correlation coefficient of 0.30. The SD of accuracy is 2% and that of correlation coefficient is 0.04. Of the residues predicted to be interface, 58% are actually interface residues, and 39% of interface residues are identified as such. We also investigated the fraction of interface residues in each protein that are correctly identified by the classifier. Our results show that in 65 out of 77 (84%) proteins, the classifier can recognize at least 20% of interface residues.

To examine whether the two-stage method learns sequence characteristics that are predictive of target residue functions, we ran a control experiment in which the class labels were randomly shuffled to destroy the attributes–class relationship in the original dataset. The correlation coefficient obtained on the class label shuffled dataset is $-0.01$ (as compared with 0.30 on the original dataset) indicating that the two-stage classifier performs significantly better than a random predictor (correlation coefficient $\approx 0$) (Table 1).

**Table 1.** Classification performance on a dataset of 77 proteins based on 5-fold cross-validation

| Dataset | Two-stage method | | Gallet's method |
|---|---|---|---|
| | Original dataset[a] | Randomized dataset[b] | Original dataset[a] |
| Correlation coefficient | 0.30 | $-0.01$ | $-0.02$ |
| Accuracy | 0.72 | 0.53 | 0.51 |
| Specificity$^+$ | 0.58 | 0.31 | 0.30 |
| Sensitivity$^+$ | 0.39 | 0.37 | 0.44 |

[a] Class labels were not shuffled (i.e. these are original class labels extracted from PDB structure files).

[b] Class labels were randomly shuffled for all the examples before training and testing the classifiers.

**Table 2.** The performance of two-stage and one-stage classifier

| | SVM method | Two-stage method |
|---|---|---|
| Correlation coefficient | 0.19 | 0.30 |
| Accuracy | 0.66 | 0.72 |
| Specificity$^+$ | 0.44 | 0.58 |
| Sensitivity$^+$ | 0.43 | 0.39 |

## 3.3 Comparison with Gallet's method

Previously Gallet *et al.* (2000) published a method to identify interface residues using an analysis of sequence hydrophobicity based on earlier work of Eisenberg *et al.* (1984). For direct comparison, we evaluated Gallet's method using 5-fold cross validation on the same dataset that was used to evaluate our two-stage classifier. We used an input window size of five for the Gallet method, which is the window size reported to perform best (Gallet *et al.*, 2000). The results shown in Table 1 indicate that the two-stage method achieves a much higher accuracy, correlation coefficient, and specificity$^+$ than Gallet method, thereby outperforming Gallet method in overall classification, although the Gallet method achieves slightly higher value sensitivity$^+$. Notably, the correlation coefficient for the Gallet's method is $-0.02$—very close to that of a random predictor.

## 3.4 Two-stage classifier yields substantially more accurate interface residue predictions than the one-stage SVM classifier

Previously, we reported an SVM method to identify interface residues (Yan *et al.*, 2003). The two-stage method reported here combines an SVM and a Bayesian classifier. Table 2 shows the performance enhancement achieved by the two-stage method. Comparison of the performance shows that

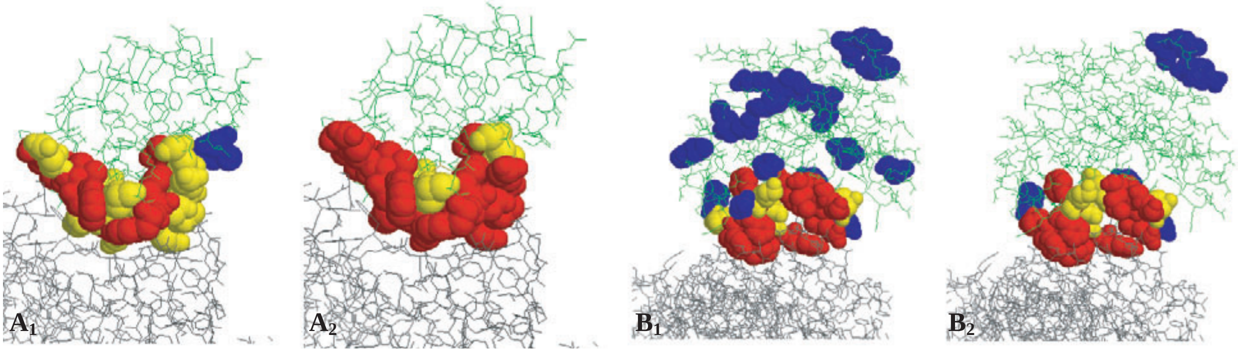

**Fig. 3.** Representative prediction results on the 77 proteins. The target protein (for which the predictions are made) in each complex is shown in green, with residues of interest shown in spacefill and color coded as follows: red, interface residues identified as such by the classifier (TPs); yellow, interface residues missed by the classifier (FPs); and blue, residues incorrectly classified as interface residues (FPs). For clarity, interface residues for the partner protein in each complex (gray wireframe) are not shown. (**A**$_1$) and (**B**$_1$) are the predictions of SVM method. (**A**$_2$) and (**B**$_2$) are the corresponding predictions of two-stage method on the same proteins. A$_1$, A$_2$: predictions on BARSTAR from PDB 1brs; B$_1$, B$_2$: predictions on SEB from PDB 1seb; structure diagrams were generated using RasMol (http://www.openrasmol.org/).

the Bayesian method (the second stage) helps to improve the classification: correlation coefficient increases from 0.19 to 0.30, accuracy increases from 0.66 to 0.72 and specificity$^+$ increases from 0.44 to 0.58; although sensitivity$^+$ decreases slightly from 0.43 to 0.39. Thus, we conclude that exploiting the distribution of interface and non-interface residues in the neighborhood of an interface residue can significantly improve the performance of classifiers for identifying interface residues.

## 3.5 Evaluation of the predictions in the context of three-dimensional structures

To evaluate further the performance of the classifier, we examined predictions in the context of the three-dimensional structures of heterocomplexes. Two representative prediction results are shown in Figure 3. For comparison, the prediction results for both the SVM method alone (the first stage) and two-stage method are shown. The 1st and 7th best (out of 77 proteins) predictions (in term of correlation coefficient) are shown in Figure 3A and B respectively. Figure 3A$_1$ and B$_1$ are the predictions of SVM method. Figure 3A$_2$ and B$_2$ are the corresponding predictions of two-stage method on the same proteins. Figure 3A$_1$ and A$_2$ show the predictions on BARSTAR from PDB 1brs, which is the complex of BARNASE and BARSTAR. On BARSTAR, the SVM method identified 8 interface residues with 1 FP (Fig. 3A$_1$), whereas two-stage method identified 16 interface residues with 0 FP (Fig. 3A$_2$). Figure 3B$_1$, and B$_2$ show the predictions on SEB from an MHC protein–antigen complex (PDB 1seb), which is the structure of SEB bound by HLA-DR1. On SEB, the SVM method identified 12 interface residues but with 20 FP (Fig. 3B$_1$), whereas the two-stage method identified 13 interface residues with only 7 FP (Fig. 3B$_2$). The results show that the two-stage classifier can successfully identify interface residues with fewer FP than the SVM classifier above. The correctly identified interface residues (residues in red) form contiguous patches on surface. With this level of success, such predictions could be valuable for guiding experimental investigations into the roles of specific residues of a protein in its interaction with other proteins or for limiting search space for docking studies.

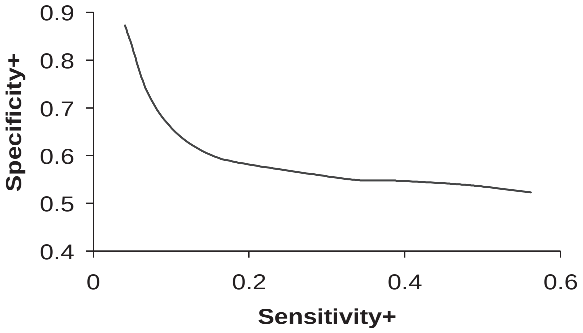

**Fig. 4.** Specificity$^+$ versus sensitivity$^+$ plot of the two-stage method.

## 3.6 Specificity–sensitivity tradeoff

In some situations (e.g. identification of critical interface residues for site-specific mutagenesis), it is desirable to predict interface residues with very high specificity. This requirement can be met by modifying the parameters used by the two-stage classifier. In the results presented so far, the two-stage classifier labels a target residue as an interface residue if $P(1|z)/P(0|z)>1$. As noted above, we can calibrate the cut off to increase the specificity of interface residue predictions (specificity$^+$) at the expense of reduced coverage (sensitivity$^+$). Figure 4 shows the specificity$^+$ versus sensitivity$^+$ plot of the predictions when different cut offs are used. When we increased the cut off to 8, the specificity of interface residue predictions (specificity$^+$) increases to 0.85 and sensitivity$^+$ decreases to 0.05. That is, 85% of the

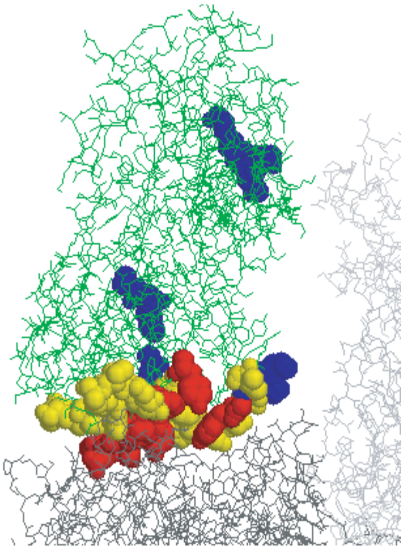

**Fig. 5.** Test results on Fab HC63 in CAPRI target 03. Fab HC63 is shown in green, with residues of interest shown in spacefill and color coded as follows: red, TPs; yellow, FPs; and blue, FPs. For clarity, interface residues for hemagglutinin (gray wireframe) are not shown. Structure diagrams were generated using RasMol (http://www.openrasmol.org/).

residues predicted to be interface residues are actually interface residues although only 5% of the interface residues are identified as such. Alternatively, if it is important to identify more potential interface residues (even at the expense of confidence), 60% interface residues can be identified with 50% specificity$^+$.

### 3.7 Evaluation of the two-stage classifier on CAPRI targets

To evaluate further the two-stage classifier, we used our dataset of interface and non-interface residues from the 77 proteins as a training set and used the resulting classifier to identify interface residues in CAPRI targets. At the time this study was performed, 7 CAPRI targets (target 01 through target 07) were available. A representative result is shown in Figure 5: the prediction on Fab HC63 in target 03, which is the complex of Fab HC63 and hemagglutinin. On Fab HC63, the two-stage method identified 10 interface residues with 10 FPs.

## 4 DISCUSSION

Development of accurate and robust computational methods for identification of protein–protein interface residues from amino acid sequence would contribute to elucidation of protein sequence–structure function relationships, with the attendant benefits in a number of applications including drug design. Several approaches for predicting the interface residues from amino acid sequence, protein structure or both have been explored with varying degrees of success. Methods that predict interface residues from amino acid sequence alone, or using amino acid sequence along with the structure of the target protein (but not the structure of the complex it forms with another protein) are of interest because relatively few experimentally determined structures of protein–protein complexes are currently available. In this paper, we have described a machine-learning approach to construct a two-stage classifier for classifying protein surface residues into interface and non-interface residues. The first stage consists of an SVM classifier. A Bayesian classifier is used at the second stage. The Bayesian classifier exploits the observation that interface residues tend to form contiguous or nearly contiguous clusters along the protein sequence. When trained and tested using 5-fold cross-validation on a non-redundant set of 77 proteins (with sequence identity below 30%) selected from heterocomplexes, the method achieved 72% accuracy with a correlation coefficient of 0.3, 66% average specificity and 65% average sensitivity. The specificity of interface residue predictions (specificity$^+$) was 58% and sensitivity (sensitivity$^+$) was 39%. Our results also show that the two-stage classifier that combines the SVM method with the Bayesian network classifier achieves better performance (correlation coefficient = 0.3, accuracy = 0.72) than a single stage SVM classifier (correlation coefficient = 0.19, accuracy = 0.66).

It is worth noting that the two-stage classifier trained using our method, on a subset of 77 proteins, also performed reasonably well in terms of identifying interface residues of CAPRI targets despite the fact that no information from the CAPRI targets was used in training the classifier. Taken together, our experiments show that the two-stage approach, which exploits the observation that interface residues tend to form contiguous or nearly contiguous clusters on protein sequences, significantly outperforms the SVM classifier.

To the best of our knowledge, the methods proposed by Gallet *et al*. (2000) and Ofran and Rost (2003b) represent the only fully sequence-based approaches to prediction of interface residues that have been evaluated on datasets consisting of more than a handful of proteins. These two methods predict interface residues by directly classifying all residues (including surface as well as core residues) into interface residues and non-interface residues whereas the methods reported in this paper classify surface residues into interface residues and non-interface residues. This is especially useful in cases where the structure of the target protein is known although the structure of the complex(es) formed by it with one or more other protein(s) is unknown. For direct comparison, we implemented the Gallet method and used it to classify the same dataset of surface residues used here into interface residues and non-interface residues. The results of our experiments show that the two-stage method presented here outperforms Gallet's method on this dataset. Further comparisons of the methods of Gallet *et al*. and of Ofran and Rose, with and without a second stage Bayesian classifier, with the methods described in this paper, on a broader range of datasets is clearly of interest.

Two points should be emphasized in evaluating the significance of these and other interface prediction results. First, it is important to note that the numbers of TP, FP, TN and FN predictions taken together provide all the relevant information for evaluating a classifier. Specificity, sensitivity, accuracy, and the correlation coefficient offer different ways to summarize these four numbers into a single measure of performance. As noted by Baldi *et al*. (2000), each of these measures, taken alone, yields only partial information about classifier performance. This problem is exacerbated when the dataset has unequal numbers of positive examples and of negative examples. For instance, if 80% of the residues are non-interaction residues, then a predictor that always predicts a residue to be a non-interaction residue will have an accuracy of 0.80 (80%). However, such a predictor is useless for correct identification of interface residues. In such a scenario, correlation coefficient is a much better indicator of the performance of a method. In this context, it is worth noting that Gallet's method shows a negative correlation coefficient that is close to zero (random prediction) on the dataset used in this study.

Second, it should be pointed out that because any given protein can interact with multiple partners, some residues identified as FPs in performance assessment of our method, as well as the methods proposed by Gallet *et al*. (2000) and Ofran and Rost (2003b), could in fact be residues that actually participate in contacts with protein(s) other than their known partners in the PDB file (or CAPRI targets).

Mucchielli-Giorgi *et al*. (1999) and Naderi-Manesh *et al*. (2001) have reported an accuracy of 85% in identifying surface residues based on amino acid sequence information using techniques for predicting solvent accessibility of residues. This raises the possibility of coupling our method with surface residue predictions to identify interface residues based on sequence information alone: first, classify all residues into surface residues and core residues; then classify surface residues into interface residues and non-interface residues.

Evolutionary information in sequences has been used in sequence-based methods to identify interface residues (Pazos *et al*., 1997; Valencia and Pazos, 2003). It would be interesting to explore whether methods that exploit evolutionary information along with sequence identity (or biophysical properties of amino acid residues) would yield more accurate identification of interface residues from amino acid sequences. Alternative approaches to exploiting knowledge of the structure (or the predicted structural properties) of the target protein may also result in more accurate prediction of interface residues.

## ACKNOWLEDGEMENTS

This research was supported in part by grants from the National Science Foundation (0219699), the National Institutes of Health (GM066387) and the Iowa State University Plant Sciences Institute.

## REFERENCES

Baldi,P., Brunak,S., Chauvin,Y. and Andersen,C.A.F. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Buntine,W. (1991) Theory refinement on Bayesian networks. *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence*. Morgan-Kaufmann, San Francisco, CA, USA. pp. 52–60.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *J. Mol. Biol.*, **272**, 132–143.

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

Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.

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

Gallet,X., Charloteaux,B., Thomas,A. and Brasseur,R. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, **302**, 917–926.

Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

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.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

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. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Keskin,O., Bahar,I., Badretdinov,A.Y., Ptitsyn,O.B. and Jernigan,R.L. (1998) Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.* **7**, 2578–2586.

Kini,R.M. and Evans,H.J. (1996) Prediction of potential protein–protein interaction sites from amino acid sequence identification of a fibrin polymerization site. *FEBS Lett.*, **385**, 81–86.

Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P., Han,J.J., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.

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

Lu,L., Arakaki,A.K., Lu,H. and Skolnick,J. (2003) Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.*, **13**, 1146–1154.

Mucchielli-Giorgi,M.H., About,S. and Puffery,P. (1999) PredAcc: prediction of solvent accessibility. *Bioinformatics*, **15**, 176–177.

Naderi-Manesh,H., Sadeghi,M., Arab,S. and Movahedi,A.A.M. (2001) Prediction of protein surface accessibility with information theory. *Proteins*, **42**, 452–459.

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

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

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **271**, 511–523.

Platt,J. (1998) *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press, Cambridge, MA.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Sheinerman,F.B. and Honig,B. (2002) On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.*, **318**, 161–177.

Teichmann,S.A., Murzin,A.G. and Chothia,C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.*, **11**, 354–363.

Valdar,W.S. and Thornton,J.M. (2001) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.

Valencia,A. and Pazos,F. (2002) Computational methods for prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.

Valencia,A. and Pazos,F. (2003) Prediction of protein–protein interactions from evolutionary information. In Bourne,P.E. and Weissig,H. (eds), *Structural Bioinformatics*. Wiley Inc., pp. 411– 426.

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

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.

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