Published in final edited form as:
Neural Comput Appl. 2004 June 1; 13(2): 123–129. doi:10.1007/s00521-004-0414-3.

Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach

Changhui Yan[1,2,5], **Vasant Honavar**[1,2,4,5], and **Drena Dobbs**[3,4,5]
Changhui Yan: chhyan@iastate.edu

[1] Artificial Intelligence Research Laboratory, Iowa State University, Atanasoff Hall 226, Ames, IA 50011-1040, USA

[2] Department of Computer Science, Iowa State University, Atanasoff Hall 226, Ames, IA 50011-1040, USA

[3] Department of Genetics, Development and Cell Biology, Iowa State University, 2114 Molecular Biology Building, Ames, IA 50011-1040, USA

[4] Laurence H Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011-1040, USA

[5] Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA 50011-1040, USA

Abstract

In this paper, we describe a machine learning approach for sequence-based prediction of protein-protein interaction sites. A support vector machine (SVM) classifier was trained to predict whether or not a surface residue is an *interface residue* (i.e., is located in the protein-protein interaction surface), based on the identity of the target residue and its ten sequence neighbors. Separate classifiers were trained on proteins from two categories of complexes, antibody-antigen and protease-inhibitor. The effectiveness of each classifier was evaluated using leave-one-out (jack-knife) cross-validation. Interface and non-interface residues were classified with relatively high sensitivity (82.3% and 78.5%) and specificity (81.0% and 77.6%) for proteins in the antigen-antibody and protease-inhibitor complexes, respectively. The correlation between predicted and actual labels was 0.430 and 0.462, indicating that the method performs substantially better than chance (zero correlation). Combined with recently developed methods for identification of surface residues from sequence information, this offers a promising approach to predict residues involved in protein-protein interactions from sequence information alone.

1 Introduction

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 applications ranging from rational drug design to analysis of metabolic and signal transduction networks. Because the number of experimentally determined structures of protein-protein complexes is small, computational methods for identifying amino acids that participate in protein-protein interactions are becoming increasingly important (reviewed in [26,28]). This paper addresses the following question: given the fact that a protein interacts with another protein, can we predict which amino acids are located in the interaction site?

Correspondence to: Changhui Yan, chhyan@iastate.edu.

Many investigators have analyzed the characteristics of protein-protein interaction sites to gain insight into the molecular determinants of protein recognition, and to identify characteristics predictive of protein-protein interfaces [4,11,18,22]. In these studies, different aspects of interaction sites such as hydrophobicity, residue propensities, size, shape, solvent accessibility, and residue pairing preferences, have been examined. Although each of these parameters provides some information indicative of protein interaction sites, none of them perfectly differentiates the interface from the rest of the protein surface.

Based on different characteristics of known protein-protein interaction sites, several methods have been proposed for predicting interface residues using a combination of protein sequence and structural information. For example, based on their observation that proline residues occur frequently near interfaces, Kini and Evans [17] predicted potential protein-protein interaction sites by detecting the presence of "proline brackets." Using this strategy, they identified the interaction sites between fibrinogen and 9E9, a monoclonal antibody which inhibits fibrin polymerization. Building on their systematic patch analysis of interaction sites, Jones and Thornton [14,15] successfully predicted interfaces in a set of 59 structures using a scoring function based on six parameters: solvation potential, residues interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area. Gallet et al. [9] identified interacting residues using an analysis of sequence hydrophobicity based on a method previously developed by Eisenberg et al. [6] for detecting membrane and surface segments of proteins. Lu et al. [18] have developed statistical potentials for interfaces and used them in a structure-based multimeric threading algorithm to assign quaternary structures and predict protein interaction partners for proteins in the yeast genome.

Several groups have used neural networks to predict protein-protein interaction sites. Zhou and Shan [32] and Fariselli et al. [7] have independently used neural network algorithms to predict whether or not a residue is located in an interaction site using the spatial neighbors of the target residues as input, and achieved accuracy of 70% and 73%, respectively. Ofran and Rost [23] have successfully predicted protein-protein interaction sites using a neural network method based on their observations that the majority of protein-protein interaction residues are clustered on a sequence and that the protein-protein interfaces differ from the rest of the protein surface in residue composition.

We have recently reported that a support vector machine (SVM) classifier can predict whether a surface residue is located in the interaction site using the *sequence neighbors* of the target residue [31]. Interface residues were predicted with specificity of 71%, sensitivity of 67%, and correlation coefficient of 0.29 on a set of 115 proteins belonging to six different categories of complexes: antibody-antigen; protease-inhibitor; enzyme complexes; large protease complexes; G-proteins, cell cycle, signal transduction; and miscellaneous [31]. The results presented in this paper show that the SVM classifiers perform even better when trained and tested on proteins belonging to each category separately, suggesting that the design of specialized classifiers for each major class of known protein-protein complexes will significantly improve sequence-based prediction of protein-protein interaction sites.

2 Methods

2.1 Protein complexes, proteins, and amino acid residues

In our previous study [31], we extracted individual proteins from a set of 70 protein-protein complexes used in the study of Chakrabarti and Janin [4]. After the removal of redundant proteins and proteins with fewer than ten residues, we obtained a data set of 115 proteins belonging to six different categories of complexes. The six categories and the number of proteins in each category are: antibody-antigen (31), protease-inhibitor (19), enzyme complexes (14), large protease complexes (8), G-proteins, cell cycle, signal transduction (22),

and miscellaneous (21). In the study described here, we focused on the proteins from two categories: 19 proteins from protease-inhibitor complexes and 31 proteins from antibody-antigen complexes (the protein list is available at http://www.public.iastate.edu/~chhyan/isda2003/sup.htm). The surface areas of residues in contact with solvent molecules (ASA) were computed for each residue in the unbound molecule and in the complex using the DSSP program [16]. The relative ASA of a residue is its ASA divided by its nominal maximum area, as defined by Rost and Sander [25]. A residue is defined to be a *surface residue* (a residue on a protein surface) if its relative ASA is at least 25% of its nominal maximum area (the overall surface area of the residue that can be contacted by solvent). 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 $1Å^2$ [13]. Using this method, we obtained 360 interface residues and 832 non-interface residues from the 19 proteins from the protease-inhibitor complexes and 830 interface residues and 3370 non-interface residues from the 31 proteins from the antibody-antigen complexes.

2.2 Support vector machine algorithm

Our study used the SVM in the Weka package from the University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/~ml/weka/) [30]. The package implements John C. Platt's [24] sequential minimal optimization (SMO) algorithm for training a support vector classifier using scaled polynomial kernels. The SVM learning algorithm [29] finds a linear boundary, i.e., a hyperplane in a high-dimensional Euclidean space, that separates the training data so that patterns of class 1 fall on one side of the hyperplane and patterns of class −1 fall on the other side of the hyperplane. If the patterns are not separable in the original *n*-dimensional pattern space, a suitable *non-linear kernel* function is used to implicitly map the patterns in the *n*-dimensional input space into a higher (finite or even infinite) dimensional *feature space* in which the patterns become separable. SVM selects the hyperplane that maximizes the margin of separation between the two classes from among all separating hyperplanes. The maximum margin separating hyperplane is fully specified by a weighted combination of the training patterns in the feature space and a bias (threshold term). Suppose the training set consists of a sequence of examples:

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_p, y_p)\},$$

where each $X_i = [x_{i1}; x_{i2}; \ldots; x_{in}]$ is a training sample, and $y_i \in \{-1; 1\}$ its known classification. The classifier constructed by the SVM learning algorithm is of the form:

$$f(x)=\text{sign}\left(\sum_i \alpha_i y_i \langle \Phi(X_i)\Phi(X)\rangle - b\right)$$

where $\Phi(X)$ is the image of an *n*-dimensional pattern vector X in a high-dimensional *feature space* induced by the chosen *kernel function;* $\langle AB\rangle$ denotes the dot product between two vectors **A** and **B**; $X = [x_1; x_2; \ldots; x_n]$ is a pattern to be classified; each X_i is a training sample; $y_i \in \{-1; 1\}$ the corresponding class label; α_i the corresponding *weight* determined by the SVM learning algorithm; and b the *threshold* or *bias* term (also determined by the SVM learning algorithm). Note that sign(Z)=1 if $Z \leq 0$ and sign(Z)=−1 if $Z<0$.

In this study, the SVM was trained to predict whether or not a surface residue is in the interaction site. It is fed with a window of 11 contiguous residues, corresponding to the target residue and five neighboring residues on each side. Following the approach used in a previous study by Fariselli et al. [7], each amino acid in the 11-residue window is represented using 20 values

obtained from the HSSP profile (http://www.cmbi.kun.nl/gv/hssp) of the sequence. The HSSP profile is based on a multiple alignment of the sequence and its potential structural homologs [5]. Thus, in our experiments, each target residue is associated with a 220-element vector. The learning algorithm generates a classifier which takes, as input, a 220-element vector that encodes a target residue to be classified, and outputs a class label.

2.3 Evaluation measures for assessing the performance of classifiers

Measures including correlation coefficient, accuracy, sensitivity (recall), specificity (precision), and false alarm rate, as discussed by Baldi et al. [1], are investigated to evaluate the performance of the classifier. Let TP denote the number of *true positives*-residues predicted to be interface residues that actually are interface residues; TN the number of *true negatives*-residues predicted not to be interface residues that are, in fact, not interface residues; FP the number of *false positives*-residues predicted to be interface residues that are not, in fact, interface residues; FN the number of *false negatives*-residues predicted not to be interface residues that actually are interface residues. Let N=TP+TN+FP+FN. Sensitivity (recall), specificity (precision), and false alarm rate were defined for the positive (+) class as well as the negative (−) class:

$$\begin{aligned}
\text{Sensitivity}^{+} &= \tfrac{\text{TP}}{\text{TP+FN}};\\
\text{Sensitivity}^{-} &= \tfrac{\text{TN}}{\text{TN+FP}};\\
\text{Specificity}^{+} &= \tfrac{\text{TP}}{\text{TP+FP}};\\
\text{Specificity}^{-} &= \tfrac{\text{TN}}{\text{TN+FN}};\\
\text{False alarm rate}^{+} &= \tfrac{\text{FP}}{\text{FP+TN}};\\
\text{False alarm rate}^{-} &= \tfrac{\text{FN}}{\text{FN+TP}}.
\end{aligned}$$

Overall sensitivity, specificity, false alarm rate, and correlation coefficient are calculated as follows:

$$\begin{aligned}
\text{Correlation coefficient} &= \frac{(\text{TP}\times\text{TN})-(\text{FP}\times\text{FN})}{\sqrt{(\text{TP+FN})(\text{TP+FP})(\text{TN+FP})(\text{TN+FN})}};\\
\text{Overall sensitivity} &= \left(\tfrac{\text{TP+FN}}{\text{N}}\right)\text{Sensitivity}^{+} + \left(\tfrac{\text{TN+FP}}{\text{N}}\right)\text{Sensitivity}^{-};\\
\text{Overall specificity} &= \left(\tfrac{\text{TP+FN}}{\text{N}}\right)\text{Specificity}^{+} + \left(\tfrac{\text{TN+FP}}{\text{N}}\right)\text{Specificity}^{-};\\
\text{Overall false alarm rate} &= \left(\tfrac{\text{TP+FN}}{\text{N}}\right)(\text{False alarm rate}^{+}) + \left(\tfrac{\text{TN+FP}}{\text{N}}\right)(\text{False alarm rate}^{-}).
\end{aligned}$$

The *sensitivity* for a class is the probability of correctly predicting an example of that class. The *specificity* for a class is the probability that a positive prediction for the class is correct. The false positive rate for a class is the probability that an example which does not belong to the class is classified as belonging to the class. The *accuracy* is the overall probability that the prediction is correct. The *correlation coefficient* is a measure of how predictions correlate with actual data, ranging from −1 to 1; when predictions match actual data perfectly, the correlation coefficient is 1; when predictions are totally opposite with actual data, the correlation coefficient is −1. Random predictions yield a correlation coefficient of 0. We chose not to emphasize the traditional measure of prediction *accuracy* because it is not a useful measure for evaluating the effectiveness of a classifier when the distribution of samples over different classes is unbalanced [1]. For instance, in the antibody-antigen category, there are 830 interface residues and 3370 non-interface residues in total, 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.

3 Results

3.1 Classification of surface residues into interface and non-interface residues

Leave-one-out cross-validation (jack-knife) was used to evaluate the performance of the SVM classifier in each category of proteins separately. For the antibody-antigen category, 31 such jack-knife experiments were performed. In each experiment, an SVM classifier was trained using a training set consisting of interface residues and non-interface residues from 30 of the 31 proteins. The resulting classifier was used to classify the surface residues from the remaining protein into *interface residues* (i.e., the amino acids located in the interaction surface) and *non-interface residues* (i.e., residues not in the interaction surface). Similarly, 19 jack-knife experiments were performed for the protease-inhibitor category. The results reported in Table 1 represent the averages for the experiments on the antibody-antigen and protease-inhibitor categories. Detailed results for individual proteins are available at http://www.public.iastate.edu/~chhyan/isda2003/sup.htm.

For proteins from the antibody-antigen complexes, the SVM achieved a relatively high sensitivity (82.3%) and specificity (81.0%), with a correlation coefficient of 0.430 between the predicted and actual class labels, indicating that the method performs substantially better than random guessing (which would correspond to a correlation coefficient equal to zero). For proteins from the protease-inhibitor complexes, the SVM classifiers performed with a sensitivity of 78.5% and specificity of 77.6%, and with a correlation coefficient of 0.462. For comparison, Table 1 also summarizes results obtained in our previous study using an SVM classifier trained and tested on a combined set of 115 proteins from six categories [31]. Note that the correlation coefficients obtained in the current study for antibody-antigen complexes (0.430) and protease-inhibitor complexes (0.462), are significantly higher than those obtained for a single classifier trained using a combined data set of all six types of protein-protein complexes (0.290).

3.2 Recognition of interaction sites

We also investigated the performance of the SVM classifier in terms of overall recognition of interaction sites. This was done by examining the distribution of sensitivity$^+$ (the sensitivity for positive class, i.e., interface residues class). The sensitivity$^+$ value corresponds to the percentage of interface residues that are correctly identified by the classifier.

Figure 1a shows the distribution of sensitivity$^+$ values for the 31 experiments in the antibody-antigen category. In 54.8% (17 of 31) of the proteins, the classifier recognized the interaction surface by identifying at least half of the interface residues, and, in 87.1% (27 of 31) of the proteins, at least 20% of the interface residues were correctly identified. Figure 1b shows the distribution of sensitivity$^+$ values for the 19 experiments in the protease-inhibitor category. In 63.2% (12 of 19) of the proteins, the classifier recognized the interaction surface by identifying at least half of the interface residues, and, in 84.2% (16 of 19) of the proteins, at least 20% of the interface residues were correctly identified. Distributions of other performance measures for the experiments are available in supplementary materials (http://www.public.iastate.edu/~chhyan/isda2003/sup.htm).

3.3 Evaluation of the predictions in the context of 3D structures

To further evaluate the performance of the SVM classifier, we examined predictions in the context of the 3D structures of heterocomplexes. In the antigen-antibody category, in the “best” example (correlation coefficient 0.87, sensitivity$^+$ 96%), 22 out of 23 interface residues were correctly identified as such (i.e., there was only one false negative) and five non-interface residues were incorrectly classified as belonging to the interface (false positives).

Figure 2a illustrates results obtained for another example in the antigen-antibody complex category, murine Fab N10 bound to Staphylococcal nuclease (SNase) [3]. Note that the predicted interface residues are shown only for Fab N10, and not for its interaction partner (wireframe) to avoid confusion in the figure. The Fab N10 "target" protein shown in this example ranked ninth out of 31 proteins in the antibody-antigen category in terms of prediction performance, based on its correlation coefficient. True positive predictions are shown in gray. The classifier correctly identified 20 interface residues in Fab N10 (sensitivity^{+} 83.3%), and failed to detect four of them (false negatives, white). Note that several residues that were incorrectly predicted to be interface residues (false positives, black) are located in close proximity to the interaction site. In this example, the SVM classifier correctly identified interface residues from all six complementarity determining regions (CDRs) known to be involved in epitope recognition [3].

Figure 2b, c illustrates results obtained for two proteins from the protease-inhibitor complex category, the "best" example (correlation coefficient 0.83) and "fourth best" (correlation coefficient 0.70). In the best example (Fig. 2b), the target protein is a serine protease, bovine α-chymotrypsin (1acb E), in complex with the leech protease-inhibitor eglin c (1acb I; [8]). Only one interface residue in chymotrypsin was not identified as such (Gly59, white) and only one false positive residue (Leu 123, black) is not located near the actual interface. Figure 2c shows results obtained for the fourth ranked target protein in this category, porcine pancreatic elastase (1fle E) in complex with the inhibitor elafin (1fle I; [27]). In elastase, seven interface residues were not identified (false negatives, white), but there were four false positives (black).

4 Discussion

Protein-protein interactions play a central role in protein function. Hence, sequence-based computational approaches for the identification of protein-protein interaction sites, identification of specific residues likely to participate in protein-protein interfaces, and, more generally, the discovery of sequence correlations of specificity and affinity of protein-protein interactions have major implications in a wide range of applications, including drug design, and analysis and engineering of metabolic and signal transduction pathways. The results reported here demonstrate that an SVM classifier can reliably predict interface residues and recognize protein-protein interaction surfaces in proteins of antibody-antigen and protease-inhibitor complexes. In this study, interface and non-interface residues were identified with relatively high sensitivity (82.3% and 78.5%) and specificity (81.0% and 77.6%). With this level of success, predictions generated using this approach should be valuable for guiding experimental investigations into the roles of specific residues of a protein in its interaction with other proteins. Detailed examination of the predicted interface residues in the context of the known 3D structures of the complexes suggests that the degree of success in predicting interface residues achieved in this study is due to the ability of the SVM classifier to "capture" important sequence features in the vicinity of the interface.

Our previous work [31] used a similar approach to predict interaction site residues in 115 proteins belonging to six categories (antibody-antigen; protease-inhibitor; enzyme complexes; large protease complexes; G-proteins, cell cycle, signal transduction; and miscellaneous). In each jack-knife experiment, the classifier was trained using examples from 114 proteins and tested on the remaining protein. The resulting classifier performed with a specificity of 71%, sensitivity of 67%, and with a correlation coefficient of 0.29. In contrast, the results reported in this paper were obtained using separate classifiers for the antibody-antigen category and the protease-inhibitor category. The correlation between the actual and predicted labeling of residues as interface vs. non-interface residues in this case, 0.430 and 0.462, respectively, is substantially better than the correlation of 0.29 obtained using a single classifier trained on the combined data set from all six categories of protein-protein complexes. This indicates that there

may be significant differences in sequence correlates of protein-protein interfaces among proteins that participate in different broad categories of protein-protein interfaces. In this context, systematic computational exploration of such sequence features, combined with directed experimentation with specific proteins, would be of interest.

Because interaction sites consist of clusters of residues on the protein surface, some false positives (black residues) in our experiments can be eliminated from consideration if the structure of target protein is known. For example, in Fig. 2b, Leu 123 is predicted to be an interface residue. From the structure of the target protein, we can see that Leu 123 is isolated from the other predicted interface residues. Thus, it is highly unlikely that Leu 123 participates in the interface; Leu 123 can be removed from the set of predicted interface residues. Similarly, two false positives in Fig. 2c can be removed. Therefore, the performance of the SVM classifier can be further improved if the structure of a target protein (but not the complex) is available. (If the structure of the complex is available, then there is no need to predict interface residues as they can be determined by analysis of the structure of the complex).

Recently, Zhou et al. [32] and Fariselli et al. [7] used neural-network-based approaches to predict interaction sites with accuracies of 70% and 73%, respectively. Ofran and Rost [23] also used a neural network algorithm to predict interaction sites with a precision of 70% and sensitivity of 20%. It would be particularly interesting to directly compare the results obtained in our study and theirs. Unfortunately, such a direct comparison is not possible due to differences in the choice of data sets and methods for accessing performance.

A notable difference between our study and the others is that the only structural information we used is knowledge of the set of surface residues of the target proteins. Knowledge of surface topology and the geometric neighbors of residues used in the other studies were not used in our study. Several authors have reported success in predicting surface residues from the amino acid sequence [2,10,12,19,20,21]. This raises the possibility of first predicting surface residues based on sequence information, and then using the predicted surface residue information to predict the interaction sites using an SVM classifier. The classifier resulting from this combined procedure would be able to predict interaction sites using amino acid sequence information alone. We are also exploring the use of phylogenetic information for this purpose. Other work in progress is aimed at the design and implementation of a server for the identification of protein-protein interaction sites and interface residues from sequence information. The server will provide classifiers that are based on all protein-protein complexes available in the most current release of the PDB.

Acknowledgments

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

References

1. Baldi P, Brunak S, Chauvin Y, Andersen CAF. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–424. [PubMed: 10871264]
2. Benner SA, Badcoe I, Cohen MA, Gerloff DL. Bona fide prediction of aspects of protein conformation: assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. J Mol Biol 1994;235:926–958. [PubMed: 8289328]
3. Bossart-Whitaker P, Chang CY, Novotny J, Benjamin DC, Sheriff S. The crystal structure of the antibody N10-staphylococcal nuclease complex at 2.9 Å resolution. J Mol Biol 1995;253:559–575. [PubMed: 7473734]
4. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343. [PubMed: 11948787]

5. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 1998;26:313–315. [PubMed: 9399862]
6. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984;179:125–142. [PubMed: 6502707]
7. Fariselli P, Pazos F, Valencia A, Casadia R. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361. [PubMed: 11874449]
8. Frigerio F, Coda A, Pugliese L, Lionetti C, Menegatti E, Amiconi G, Schnebli HP, Ascenzi P, Bolognesi M. Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 A resolution. J Mol Biol 1992;225:107–123. [PubMed: 1583684]
9. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000;302:917–926. [PubMed: 10993732]
10. Gallivan JP, Lester HA, Dougherty DA. Site-specific incorporation of biotinylated amino acids to identify surface-exposed residues in integral membrane proteins. Chem Biol 1997;4:739–749. [PubMed: 9375252]
11. Glaser F, Steinberg DM, Vakser A, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89–102. [PubMed: 11276079]
12. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequence. Protein Eng 1990;3:659–665. [PubMed: 2217139]
13. Jones S, Thornton JM. Principles of protein-protein interactions. P Natl Acad Sci USA 1996;93:13–20.
14. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997a;272:121–132. [PubMed: 9299342]
15. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997b;272:133–143. [PubMed: 9299343]
16. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637. [PubMed: 6667333]
17. Kini RM, Evans HJ. Prediction of potential protein-protein interaction sites from amino acid sequence identification of a fibrin polymerization site. FEBS Lett 1996;385:81–86. [PubMed: 8641473]
18. Lu L, Lu H, Skolnick J. Development of Unified Statistical Potentials describing Protein-protein interactions. Biophy J 2003;84:1895–1901.
19. Mandler J. ANTIGEN: protein surface residue prediction. Comput Appl Biosci 1988;4:493. [PubMed: 3208187]
20. Mucchielli-Giorgi MH, About S, Puffery P. PredAcc: prediction of solvent accessibility. Bioinformatics 1999;15:176–177. [PubMed: 10089205]
21. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459. [PubMed: 11170200]
22. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003a;325:377–387. [PubMed: 12488102]
23. Ofran Y, Rost B. Predicted protein-protein interaction sites from local sequence information. FEBS Lett 2003b;544:236–239. [PubMed: 12782323]
24. Platt, J. Fast training of support vector machines using sequential minimal optimization. In: Scholkopf, B.; Burges, CJC.; Smola, AJ., editors. Advances in kernel methods-support vector learning. MIT Press; Cambridge: 1998. p. 185-208.
25. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226. [PubMed: 7892171]
26. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struc Biol 2001;11:354–363.
27. Tsunemi M, Matsuura Y, Sakakibara S, Katsube Y. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 A resolution. Biochemistry 1996;35:11570–11576. [PubMed: 8794736]
28. Valencia A, Pazos F. Computational methods for prediction of protein interactions. Curr Opin Struc Biol 2002;12:368–373.
29. Vapnik, V. Statistical learning theory. Springer; Berlin Heidelberg New York: 1998.

30. Witten, IH.; Frank, E. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kauffman; San Mateo, California: 1999.

31. Honavar, V.; Yan, C.; Dobbs, D. Technical report ISU-CS-TR 02–11. Department of Computer Science, Iowa State University; 2002. Predicting protein-protein interaction sites from amino acid sequence. http://archives.cs.iastate.edu/documents/disk0/00/00/02/88/index.html)

32. Zhou H, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343. [PubMed: 11455607]

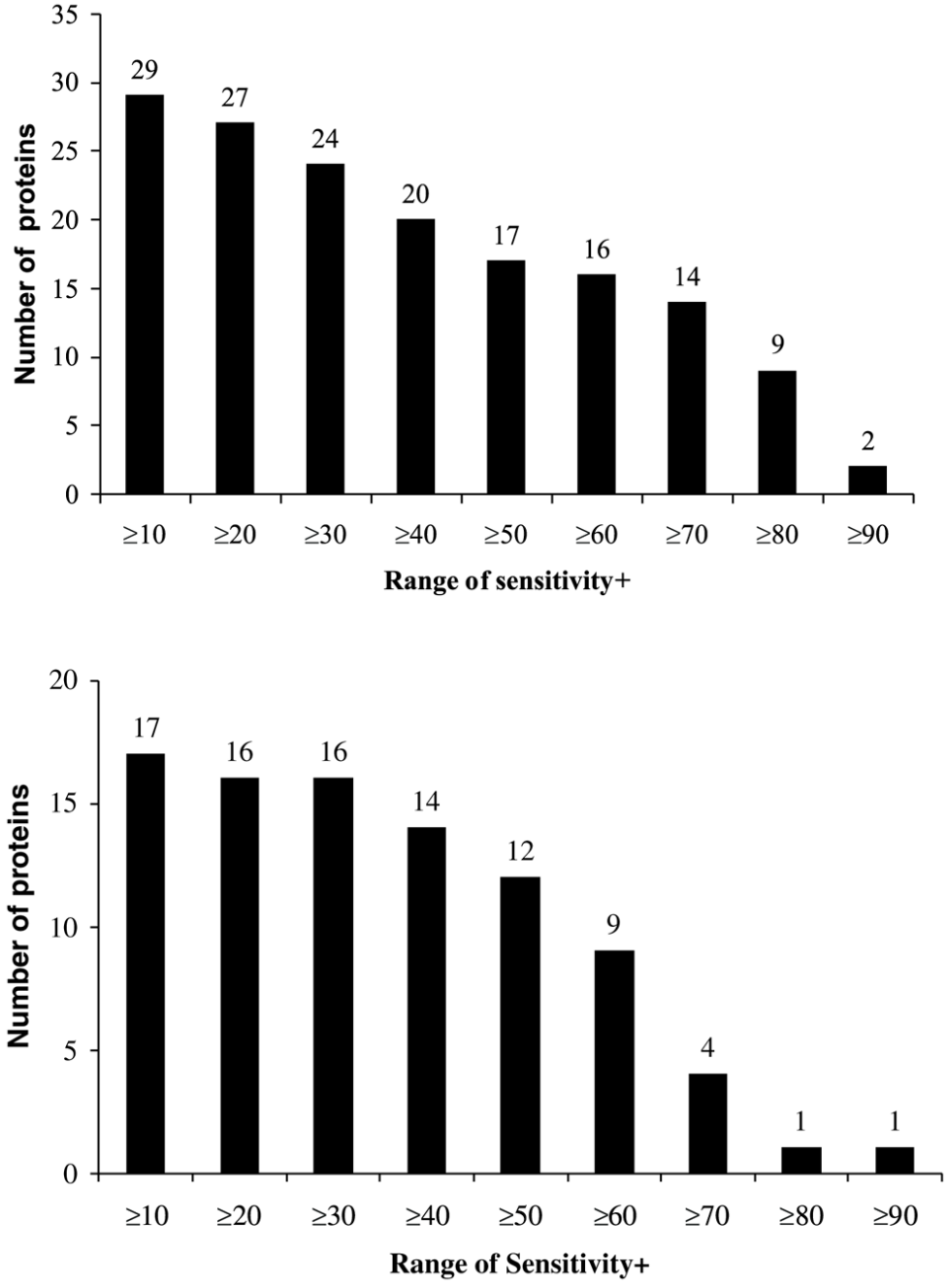

Fig. 1.

a, b Interaction site recognition: distribution of sensitivity^{+} (sensitivity for predicting interface residues) values. The *bars* on the graphs illustrate the fraction of the experiments (*vertical axis*) that fall into the performance categories named below the *horizontal axis*. **a** The distribution of sensitivity^{+} values for 31 experiments in the antibody-antigen category. **b** The distribution of sensitivity^{+} values for 19 experiments in the protease-inhibitor category

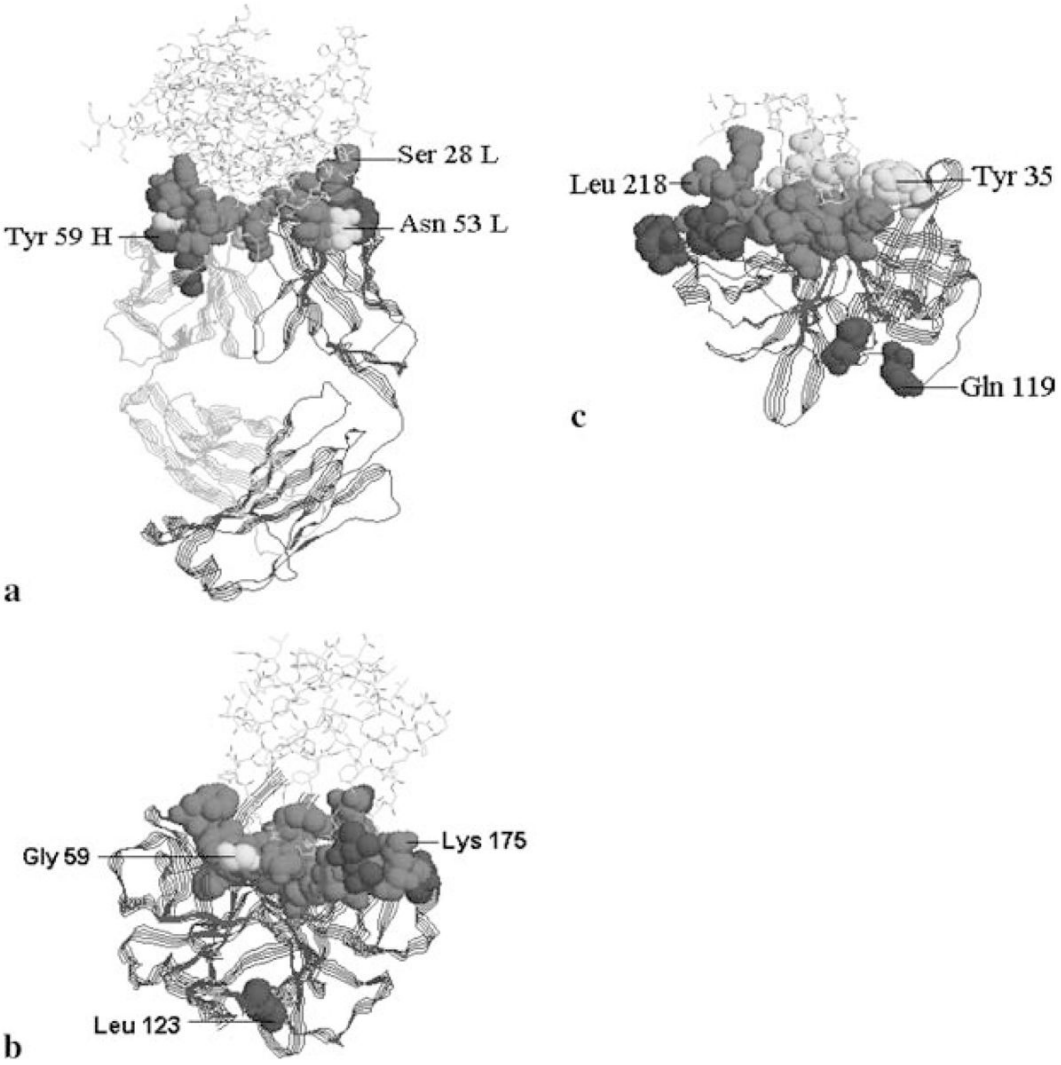

Fig. 2.

a–c Interaction site recognition: visualization of 3D structures of representative heterocomplexes. The target protein in each complex is shown in *strands*, with residues of interest shown in *space fill* and color-coded as follows: *gray*, true positives (interface residues identified as such by the classifier); *white*, false negatives (interface residues missed by the classifier); *black*, false positives (residues incorrectly classified as interface). The interaction partner is shown in *gray* wireframe. **a** FabN10 in the 1nsn complex. **b** α-chymotrypsin in the 1acb complex. **c** Elastase in the 1fle complex. Structure diagrams were generated using RasMol (http://www.openrasmol.org/)

Table 1

Performance of the SVM classifier

	Antibody-antigen complexes[a]	Protease-inhibitor complexes[a]	Six categories of complexes[b]
Correlation coefficient	0.430	0.462	0.290
Sensitivity	82.3%	78.5%	66.9%
Specificity	81.0%	77.6%	70.8%
False alarm rate	41.0%	35.7%	35.9%

[a] The SVM classifiers were trained and evaluated separately on proteins from the antibody-antigen complexes and protease-inhibitors complexes

[b] The performance of the SVM trained and tested on a combined set of 115 proteins from six different categories [31]