Identification of Surface Residues Involved in Protein-Protein Interaction – A Support Vector Machine Approach

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

[1]Artificial Intelligence Research Labortory, [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,
Iowa State University.
Atanasoff Hall 226, Iowa State University, Ames, IA 50011-1040, USA.
chhyan@iastate.edu

Summary

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 10 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 prediction of residues involved in protein-protein interaction from sequence information alone.

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 for protein-protein complexes is small, computational methods for identifying amino acids that participate in protein-protein interactions are becoming increasingly important (reviewed in Teichmann *et al.*, 2001; Valencia and Pazos, 2002). This paper addresses the question: Given the fact that a protein interacts with another protein; can we predict which amino acids are located in the interaction site?

Based on different characteristics of known protein-protein interaction sites, several methods have been proposed for predicting protein-protein interaction sites using a combination of sequence and structural information. These include methods based on presence of “proline brackets’’ (Kini and Evans, 1996), patch analysis using a 6-parameter scoring function (Jones and Thornton 1997a, 1997b), analysis of hydrophobicity distribution around a target residue (Gallet *et al.*, 2000), multiple sequence alignment (Casarai *et al.*, 1995; Lichtarge *et al.*, 1996; Pazos *et al.*, 1997), structure-based multimeric threading (Lu *et al.*, 2002), analysis of amino acid characteristics of *spatial neighbors* of a target residue using a neural network (Zhou *and* Shan,2001; Fariselli *et al.*, 2002).

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, 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. (Yan *et al.* 2002). 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.

Methods

Protein complexes, proteins and amino acid residues

Proteins of protease-inhibitor complexes and antibody-antigen complexes were chosen from the 115 proteins used in our previous study (Yan *et al.* 2002). From these, we obtained two set of proteins used in this study: 19 proteins from protease-inhibitor complexes and 31 proteins from antibody-antigen complexes. Solvent accessible surface area (ASA) was computed for each residue in the unbound molecule (MASA) and in the complex (CASA) using the DSSP program (Kabsch and Sander, 1983). The relative ASA of a residue is its ASA divided by its nominal maximum area as defined by Rost and Sander (1994). A residue is defined to be a *surface residue* if its relative MASA is at least 25% of its nominal maximum area. 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 (Jones and Thornton, 1996). 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.

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/) (Witten and Frank 1999). The package implements John C. Platt's (1998) sequential minimal optimization (SMO) algorithm for training a support vector classifier using scaled polynomial kernels. The SVM is 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 5 neighboring residues on each side. Following the approach used in a previous study by Fariselli *et al.* (2002), 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 (Dodge *et al.*, 1998). 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.

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 (2000) are investigated to evaluate the performance of the classifier. Detailed definition of these measures can be found in supplementary materials (http://www.public.iastate.edu/~chhyan/isda2003/sup.htm). 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 prediction is correct. The *correlation coefficient* is a measure of how predictions correlate with actual data. It ranges from -1 to 1. When predictions match actual data perfectly, correlation coefficient is 1. When predictions totally disagree with actual data, 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 (Baldi, 2000). 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.

Results

Classification of surface residues into interface and non-interface residues

To evaluate the effectiveness of this approach we used leave-one-out cross-validation (jack-knife) experiments on each category of complexes. 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 averages for the antibody-antigen and protease inhibitor categories, respectively. Detailed results of the experiments are available in supplementary materials (http://www.public.iastate.edu/~chhyan/isda2003/sup.htm).

For proteins from antibody-antigen complexes, the SVM classifies achieved relatively high sensitivity (82.3%), specificity (81.0%), with a correlation coefficient of 0.430 between predicted and actual class labels, indicating that the method performs substantially better than random guessing (which would correspond to correlation coefficient equal to zero). For proteins from protease-inhibitor complexes, the SVM classifiers performed with 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 (Yan *et al.* 2002). 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 that obtained for a single classifier trained using a combined dataset of all six types of protein-protein complexes (0.290).

Table 1. Performance of the SVM classifier

	Antibody-antigen complexes[a]	Protease-inhibitor complexes[a]	Six categories of complexes[b]
CC	0.430	0.462	0.290
SN	82.3%	78.5%	66.9%
SP	81.0%	77.6%	70.8%
FAR	41.0%	35.7%	35.9%

CC correlation coefficient.
SN sensitibity.
SP specificity.
FAR false alarm rate

[a] The SVM classifiers were trained and evaluated separately on two categories of proteins.

[b] The performance of the SVM trained and tested on a mixed set of 115 proteins from six categories (Yan *et al.* 2002)

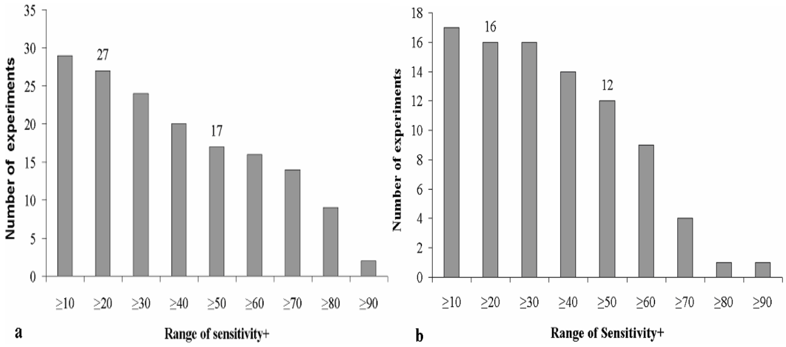

Fig. 1. Interaction site recognition: distribution of Sensitivity$^+$ (sensitivity for predicting interface residues) values**.** The bars in graph illustrate the fraction of the experiments (vertical axis) that fall into the performance categories named below the horizontal axis. **a**. the distribution 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

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.

Fig. 1a shows the distribution of sensitivity$^+$ values for the 31 experiments in 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. **Fig. 1b** shows the distribution of sensitivity$^+$ values for the 19 experiments in 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).

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

To further evaluate the performance of the SVM classifier, we examined predictions in the context of the three-dimensional 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 1 false negative) and only 5 non-interface residue was incorrectly classified as belonging to the interface (false positive).

Fig. 2a illustrates results obtained for another example in the antigen-antibody complex category, murine Fab N10 bound to Staphylococcal nuclease (SNase) (Bossart-Whitaker *et al.*, 1995). Note that predicted interface residues are shown only for Fab N10, and not for its interaction partner (gray) to avoid confusion in the figure. The Fab N10 "target" protein shown in this example ranked 9th 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 red. The classifier correctly identified 20 interface residues in Fab N10 (sensitivity+ 83.3%), and failed to detect only 4 of them (false negatives, yellow). Note that several residues that were incorrectly predicted to be interface residues (false positives, blue) are located in close proximity to the interaction site. In this example, the SVM classifier correctly identified interface residues from all 6 complementarity determining regions (CDRs) known to be involved in epitope recognition (Bossart-Whitaker *et al*., 1995).

Fig. 2b and **c** illustrate results obtained for two proteins from the protease-inhibitor complex category, the "best" example (correlation coefficient 0.83) and "4th 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; Frigerio *et al.*, 1992). Only 1 interface residue in chymotrypsin was not identified as such (Gly59, yellow) and only 1 false positive residue (Leu 123 blue) is not located near the actual interface. **Fig.. 2c** shows results obtained for the 4th ranked target protein in this category, porcine pancreatic elastase (1fle E) in complex with the inhibitor elafin (1fle I; Tsunemi *et al*., 1996). In elastase, only 7 interface residues were not identified (false negatives, yellow), but there were 4 false positives (blue).

Discussion

Protein-protein interactions play a central role in protein function. Hence, sequence-based computational approaches for identification of protein-protein interaction sites, identification of specific residues likely to participate in protein-protein interfaces, and more generally, discovery of sequence correlates 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%). In 54.8% and 62.3% of the proteins tested, the interaction site could be easily recognized because more than half of the interface residues were correctly identified. 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 3-dimensional structures of the complexes suggest 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 (Yan *et al. 2002*) 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 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 antibody-antigen category and protease-inhibitor category. The correlation between actual and predicted labeling of residues as interface versus 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 interaction among proteins that participate in different broad categories of protein-protein interaction. In this context, systematic computational exploration of such sequence features, combined with directed experimentation with specific proteins (e.g., using site-specific mutagenesis) would be of interest. These results also suggest that in building sequence-based classifiers for identifying residues likely to form protein-protein interaction surfaces, a 2-stage approach based on identification of the broad category of interaction the protein is likely to be involved in (say antibody-antigen versus protease-inhibitor), followed by classification of amino acid residues into interface versus non-interface classes may be worth exploring.

Because interaction sites consist of clusters of residues on the protein surface, some false positives (blue residues) in our experiments can be eliminated from consideration if the structure of target protein is known. For example, in Figure 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 likely that Leu 123 is not an interface residue. Thus we can remove Leu 123 from the set of predicted interface residues. Similarly two false positives in Figure 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.* (2002) and Fariselli *et al.* (2002) used neural network-based approaches to predict interaction sites with accuracy of 70% and 73%. 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 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 the 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 recently reported success in predicting surface residues from amino acid sequence (Mandler 1988; Holbrook *et al.* 1990; Benner *et al.* 1994; Gallivan *et al.* 1997; Mucchielli-Giorgi *et al.* 1999; Naderi-Manesh *et al.* 2001). 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 the SVM classifier. The classifier resulting from this combined procedure will be able to predict interaction site using amino acid

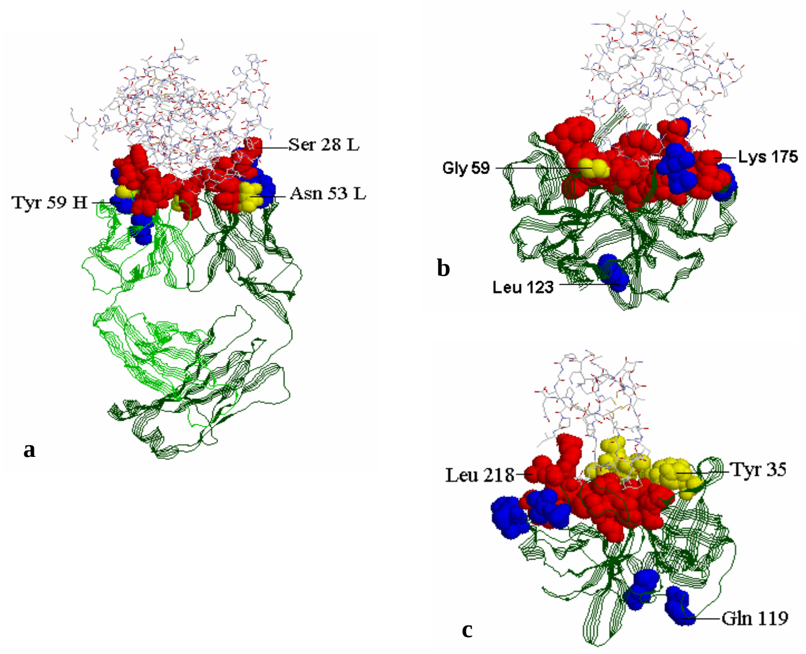

Fig. 2. Interaction site recognition: visualization on three-dimensional structures of representative heterocomplexes. The target protein in each complex is shown in green, with residues of interest shown in *space fill* and color coded as follows: red, true positives (interface residues identified as such by the classifier); yellow, false negatives (interface residues missed by the classifier); blue, 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/）

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 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 PDB.

References

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

Benner SA, Badcoe I, Cohen MA, Gerloff DL (1994) 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* **235**, 926-58

Bossart-Whitaker P, Chang CY, Novotny J, Benjamin DC, Sheriff S (1995) The crystal structure of the antibody N10-staphylococcal nuclease complex at 2.9 Å resolution. *J Mol Biol* **253**, 559-575

Braden BC, Fields BA, Ysern X, Dall'Acqua W, Goldbaum FA, Poljak RJ, Mariuzza RA (1996) Crystal structure of an Fv-Fv idiotope-anti-idiotope complex at 1.9 A resolution. *J Mol Biol***264**:137-51

Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. *Nat Struct Biol* **2**,171-178

Dodge C, Schneider R, Sander C (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res* **26**, 313-315

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

Frigerio F, Coda A, Pugliese L, Lionetti C, Menegatti E, Amiconi G, Schnebli HP, Ascenzi P, Bolognesi M (1992) Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 A resolution. *J Mol Biol* **225**:107-123

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

Gallivan JP, Lester HA, Dougherty DA (1997) Site-specific incorporation of biotinylated amino acids to identify surface-exposed residues in integral membrane proteins. *Chem Biol* **4,** 739-749

Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. *Protein Eng* **3**, 659-665

Jones S,Thornton JM (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci* USA, **93**, 13-20

Jones S, Thornton JM (1997a) Analysis of protein-protein interaction sites using surface patches. *J Mol Boil* **272,** 121-132

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

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

Kini RM, Evans HJ (1996) Prediction of potential protein-protein interaction sites from amino acid sequence identification of a fibrin polymerization site. *FEBS letters* **385**, 81-86

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol***257**, 342-358

Lu L, Lu H, and Skolnick J (2002) MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading *Proteins* **49**, 350-364

Mandler J (1988) ANTIGEN: protein surface residue prediction. *Compute Apple Basic* **4**, 493

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

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

Pazos F, Helmer-Citterich M, Ausiello G, 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. In B Scholkopf C J C, Burges and A J Smola editors, Advances in Kernel Methods - Support Vector Learning, p 185-208, Cambridge, MA, MIT Press

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

Teichmann SA, Murzin AG, and Chothia C (2001) Determination of protein function, evolution and interactins by structural genomics. *Curr Opin Struct Biol* **11**:354-363

Tsunemi M, Matsuura Y, Sakakibara S, Katsube Y(1996) Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 A resolution *Biochemistry* **35**:11570-11576

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

Witten I H, Frank E (1999) Data mining: Practical machine learning tools and techniques with java implementations. San Mateo, CA: Morgan Kaufmann

YanC, Dobbs D, Honavar V (2002) Predicting protein-protein interaction sites from amino acid sequence. Technical report ISU-CS-TR 02-11. Department of computer science, Iowa State University, USA

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