Protein Engineering, Design & Selection vol. 17 no. 4 pp. 367–373, 2004
Published online May 27, 2004 doi:10.1093/protein/gzh042

Learning to discriminate between ligand-bound and disulfide-bound cysteines

Andrea Passerini[1] and Paolo Frasconi

Dipartimento di Sistemi e Informatica, Università a di Firenze, 50139
Firenze, Italy

[1]To whom correspondence should be addressed.
E-mail: passerini@dsi.unifi.it

We present a machine learning method to discriminate between cysteines involved in ligand binding and cysteines forming disulfide bridges. Our method uses a window of multiple alignment profiles to represent each instance and support vector machines with a polynomial kernel as the learning algorithm. We also report results obtained with two new kernel functions based on similarity matrices. Experimental results indicate that binding type can be predicted at significantly higher accuracy than using PROSITE patterns.

Keywords: disulfide bridges/metal binding sites/prediction tools/support vector machines

Introduction

Non-free cysteines that are not involved in the formation of disulfide bridges usually bind prosthetic groups that include a metal ion, which play an important role in the function of a protein. The discrimination between the presence of a disulfide bridge (DB) or a metal binding site (MBS) in correspondence to a bound cysteine is often a necessary step during the NMR structural determination process of metalloproteins, and its automation may significantly help towards speeding up the overall process. Several proteins are known where both situations are in principle plausible and it is not always possible to assign a precise function to each cysteine. For example, the mitochondrial copper metallochaperone Cox17 contains six cysteines but it is not precisely known which ones are actually involved in metal binding (Heaton *et al.*, 2001). Another example is the inner mitochondrial membrane protein sco1p that contains a CxxxC motif, where experimental evidence could lead us to believe it is involved in copper transport, but whose fold similarity to a thioredoxin fold could conversely suggest catalytic activity (Chinenov, 2000; Balatri *et al.*, 2003).

In addition to the above motivations, the discrimination between DBs and MBSs may help towards a more accurate prediction of the disulfide bonding state of cysteines (Fiser and Simon, 2000). In this task, global descriptors (Mucchielli-Giorgi *et al.*, 2002) and global postprocessing (Martelli *et al.*, 2002) can help to predict the overall tendency of a chain to form or not to form bridges. However, current prediction methods typically fail to identify the type of binding in which each single bound cysteine is involved.

In some well known cases, the presence of a binding site for a prosthetic group can be detected by inspecting the consensus pattern that matches the portion of the protein sequence containing the target cysteine. For example, the 4Fe-4S ferredoxin group is associated with the pattern C-x(2)-C-x(2)-C-x(3)-C-[PEG] (Otaka and Ooi, 1987). Using annotated chains from SWISSPROT and PDB we show that a set rule based on PROSITE (Falquet *et al.*, 2002) patterns can be actually used to separate MBSs from DBs. Many of these rules are highly specific but there are well known examples of false positives. For example, the C-{CPWHF}-{CPWR}-C-H-{CFYW} consensus pattern is often correctly associated with a cytochrome *c* family heme binding site (Mathews, 1985), but the precision of such a pattern is <40%. In addition, there are cases of cysteines involved in an MBS and having no associated pattern.

In this paper we formulate the prediction task as a binary classification problem: given a non-free cysteine and information about flanking residues, predict whether the cysteine can bind to a prosthetic group containing a metal ion (positive class) or is it always bound to another cysteine forming a DB (negative class).

First, we suggest a nontrivial baseline predictor based on PROSITE pattern hits and the induction of decision trees using the program C4.5 (Quinlan, 1993). Secondly, we introduce a classifier fed by multiple alignment profiles and based on support vector machines (SVM) (Cortes and Vapnik, 1995). We show that the latter classifier is capable of discovering the large majority of the relevant PROSITE patterns, but is also sensitive to signal in the profile sequence that cannot be detected by regular expressions and therefore outperforms the baseline predictor.

Materials and methods

Data preparation

The data for cysteines involved in DB formation were extracted from PDB (Berman *et al.*, 2000), We excluded chains if: (i) the protein was shorter than 30 residues; (ii) it had less than two cysteines; (iii) it had cysteines marked as unresolved residues in the PDB file; (iv) the data for metal binding sites were extracted from SWISS-PROT version 41.23 (Boeckmann *et al.*, 2003), since PDB does not contain enough examples of metal ligands. In this case we included all entries containing at least one cysteine involved in metal binding, regardless of the annotation confidence. In this way, examples of bindings with iron–sulfur clusters (2FE2S, 3FE4S, 4FE4S), copper, heme groups, iron, manganese, mercury, nickel and zinc were obtained.

Intra-set redundancy due to sequence similarity was avoided by running the UniqueProt program (Mika and Rost, 2003) with the hssp distance set to zero. Inter-set redundancy however was kept in order to handle proteins with both DBs and metal bindings. It must be remarked that while inter-set redundancy can help the learning algorithm by providing additional data for training, it cannot favorably bias accuracy estimation

Table I. Non homologous sequences obtained by running UniqueProt (Mika and Rost, 2003) with the hssp distance set to zero

	No. of sequences	No. of cysteines
Disulfide bridges	529	2860
Metal bindings	202	758

Sequences containing DBs were obtained from resolved proteins in the PDB (Berman *et al.*, 2000), while those with metal bindings were recovered from SWISSPROT version 41.23 (Boeckmann *et al.*, 2003) by keyword matching.

since redundant cases should be assigned to opposite classes. The number of non-homologous sequences remaining in the data sets are shown in Table I, together with the number of cysteines involved in bridges or metal bindings. Free cysteines were ignored.

PROSITE patterns as a baseline

In this section we establish a procedure to compute a baseline accuracy for the prediction task studied in this paper. In general, the base accuracy of a binary classifier is the frequency of the most common class. For the data set described above the base accuracy is 84.4%. In total absence of prior knowledge a predictor that performs better than the base accuracy is generally considered as successful. However, (i) it must be remarked that, especially for highly unbalanced data sets, precision/recall rates are also needed in order to have a correct view of the classifier performance; (ii) for the task studied in this paper several well known consensus patterns exist that are associated with DBs and with metal binding sites. These patterns partially encode expert knowledge and it seems reasonable, when possible, to make use of them as a rudimentary prediction tool.

Thus, in order to compare our prediction method with respect to a more interesting baseline than the mere base accuracy, we extracted features that consist of PROSITE (Falquet *et al.*, 2002) pattern hits. A PROSITE pattern is an annotated regular expression that describes a relatively short portion of a protein sequence that has a biological meaning. We run the program ScanProsite (Gattiker *et al.*, 2002) on the data set described above searching for patterns listed in the release 18.18 (December 2003) of PROSITE. In this way we found 199 patterns whose matches with the sequences in our data set contain the position of at least one bound cysteine. Approximately 56% of the cysteines bound to a metal ion and ~41% of the cysteines forming DBs matched at least one PROSITE pattern. Many of the patterns associated with DBs have perfect (100%) specificity but each one only covers a very small set of cases. Overall, the fraction of disulfide-bond cysteines matched by a perfectly specific pattern is ~26%. Patterns associated with MBSs have a significantly higher coverage, although their specificity is perfect only ~18% of the time and sometimes is lower. We remark that in our context a metal binding pattern is perfectly specific if every match is actually a MBS, regardless of the metal involved in the bond. Thus, examples of perfectly specific patterns associated with MBSs include PS00198 (4Fe-4S ferredoxins iron-sulfur binding region), PS00190 (cytochrome *c* family heme-binding site), and PS00463 (Fungal Zn(2)-Cys(6) binuclear cluster domain). To further complicate the scenario, 12% of the bound cysteines match more than one pattern. For these reasons, a prediction rule based on pattern matches is difficult to craft by hand and we used the program C4.5 to create rules automatically from data. C4.5 induces a decision trees from labeled examples by recursively partitioning the instance space, using a greedy heuristic driven by information theoretic considerations (Quinlan, 1986). Rules are then obtained by visiting the tree from the root to a leaf. When using C4.5, each bound cysteine was simply represented by the bag of its matching patterns.

Support vector machines algorithm

Support vector machines (SVM) have been introduced in Cortes and Vapnik (1995) and give a well-posed formulation to the supervised learning problem so that a unique solution can be obtained once certain entities have been fixed. Here we briefly sketch the main ideas and notation of the general algorithm to allow proper replication of our experiments. Details can be found in several textbooks including Cristianini and Shawe-Taylor (2000) and Schölkopf and Smola (2002).

Training data are a set of pairs $D_{\mathrm{m}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ where $\boldsymbol{x}_i$ is the input (in our case a real vector of features describing the context around a target cysteine) and y_i is the output class, either $+1$ (indicating a metal-ion binding) or -1 (indicating a DB). The algorithm learns from the data a classification function f that can be used to make predictions about new instances. This function can be written as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} y_i \alpha_i \; k(\boldsymbol{x}, \boldsymbol{x}_i) + \beta_0 \qquad (1)$$

where $\alpha_1, \ldots, \alpha_m, \beta_0$ are parameters adjusted by the learning procedure and K is the kernel, a positive semidefinite symmetric function that measures the similarity between two input vectors and that can be thought of as a generalized dot product. The solution is a minimizer of the following functional:

$$H(f) = \sum_{i=1}^{m} \frac{C}{m} |1 - y_i f(\boldsymbol{x}_i)|_+ + \frac{1}{2} \|f\|_k^2 \qquad (2)$$

where $|a|_+ = a$ if $a > 0$ and zero otherwise, and $\|f\|_K^2$ is the norm of the classification function f induced by the kernel K. The above functional is the sum of two terms: the leftmost one penalizes solutions that do not classify correctly (and with large margin) training data, while the rightmost one is a regularizer that penalizes too complex solutions and allows us to avoid overfitting. The quantity C expresses the relative weight of the two terms and thus controls a trade-off between the memorization of training examples and generalization to new instances.

The SVM algorithm is capable of handling extremely numerous and sparse features, thus allowing us to exploit a wide local context surrounding the cysteine under investigation. In particular, we provided information in the form of a symmetric window of $2k+1$ residues, centered around the target cysteine, with k varying from 1 to 25. In order to include evolutionary information, we coded each element of the window by its multiple alignment profile computed with psiblast (Altschul *et al.*, 1997).

Preliminary model selection experiments were conducted in order to choose the appropriate kernel, together with its hyperparameters. In subsequent experiments we employed third degree polynomial kernels, with offset equal to one, fixed regularization parameter given by the inverse of the average of $K(x, x)$ with respect to the training set, and distinct penalties (Joachims, 1998) for errors on positive or negative examples, in order to rebalance the different proportion of examples in the two classes (see Table I). Similarity between examples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is therefore computed as:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i^T \boldsymbol{x}_j + 1\right)^3 \tag{3}$$

In order to include information about similarity between different residues, we also implemented a new kernel in which the dot product between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is mediated by a substitution matrix M:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i^T M \boldsymbol{x}_j + 1\right)^3 \tag{4}$$

In order for Equation (4) to be a valid kernel, matrix M has to be symmetric positive definite (Schölkopf and Smola, 2002). We tried the McLachlan amino acid similarity matrix (McLachlan, 1972), which already satisfies such a condition, and the Blosum62 (Henikoff and Henikoff, 1992) substitution matrix, that turned out to be positive definite after normalizing each element as $(M_{\mathrm{rc}} - \min) = (\max - \min)$ where max and min are over the entire matrix. A similar approach (Guermeur *et al.*, 2004) was recently applied to secondary structure prediction.

Results and discussion

Test performances were calculated by 3-fold cross validation: proteins were divided in three groups, maintaining in each group approximately the same distribution of DBs and different kinds of metal binding sites.

The confusion matrix for PROSITE induced rules is shown in Table II. It must be observed that the accuracy in Table II is an optimistic upper bound of the true predictive power of this method with respect to future sequences. This is because PROSITE patterns have been designed to minimize the number of false hits and missed hits by actually inspecting the available sequence databases.

Results for the polynomial kernel are reported in Figure 1. Train and test accuracies are plotted for growing size of the context window, with error bars for 95% confidence intervals, together to the fraction of support vectors over the training examples in the learned models, which is a rough indicator of the complexity of the learned models. The most evident improvement in test accuracy is obtained for a window of size $k = 3$, and corresponds to the global minimum in the model complexity curve with ~56% of training examples as support vectors. Detailed results for such a window are reported in Table III, showing they are obtained for an approximate breakeven point in the precision recall curve. A deeper analysis of individual predictions showed that the vast majority of predictions were driven by the presence of a well conserved *CXXC* pattern, taken as the indicator of metal binding. This explains the high rate of false positives compared to the total number of positive examples, most of them being cysteines containing the pattern but involved in DBs, while most of the false negatives are metal bindings missing it. The learned pattern is actually very common for most bindings involving iron–sulfur, iron–nickel and heme groups, and these kinds of metal binding are actually predicted with the highest recall.

The best accuracy is obtained for a window of size $k = 17$, corresponding to ~87% of examples as support vectors. Detailed results are reported in Tables IV and V, showing a strong reduction of false positives at the cost of a slight increase in the number of metal bindings predicted as DBs. Sequence logos (Schneider and Stephens, 1990) for cysteine contexts in the case of metal bindings (Figure 2, top) show a well conserved *CXXCXXC* pattern, which is common in 4FE4S clusters,

Table II. DB versus metal binding prediction by decision rules learned by c4.5 from patterns extracted from PROSITE

	Precision (%)	Recall (%)	Bridge	Metal	
Bridge	84	99	2845	15	True
Metal	93	27	556	202	
Accuracy	84.2		Predicted		

Precision and recall for both classes, confusion matrix and overall accuracy.

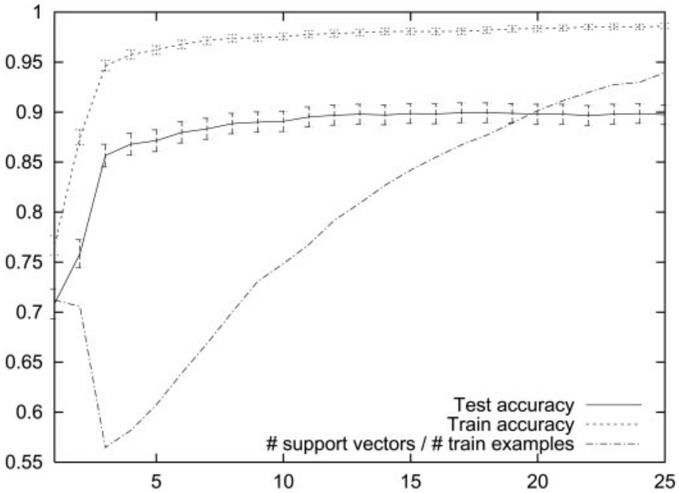

Fig. 1. DB versus metal binding prediction by SVMs with a third degree polynomial kernel. Test and train accuracies with 95% confidence intervals are plotted, together to the fraction of support vectors over the number of training examples, for growing sizes of the window of $2k+1$ residue profiles around the target cysteine, with k going from 1 to 25. Results are averaged over a 3-fold cross validation procedure.

Table III. DB versus metal binding prediction by third degree polynomial kernel SVMs

	Precision (%)	Recall (%)	Bridge	Metal	
Bridge	91	90	2588	272	True
Metal	65	67	247	511	
Accuracy	86		Predicted		

Window of three residue profiles on both sides of the target cysteine. Precision and recall for both classes, confusion matrix and overall accuracy.

Table IV. DB versus metal binding prediction by third degree polynomial kernel SVMs

	Precision (%)	Recall (%)	Bridge	Metal	
Bridge	91	97	2788	72	True
Metal	87	61	292	466	
Accuracy	90		Predicted		

Window of 17 residue profiles on both sides of the target cysteine. Precision and recall for both classes, confusion matrix and overall accuracy.

and the *CXXCH* pattern typical of heme groups, but also the presence of *CG* patterns in positions as distant as nine residues from the target cysteine. DB contexts are much more uniform (see Figure 2, bottom), but show a strong tendency for polar aminos, especially glycine, cysteine and serine all along the window. The model seems capable of exploiting very distant information, and it discovers almost all rules induced by PROSITE patterns (see Table II), as only 13 of 202 metal bindings are lost, while it also corrects eight of 15 false metal bindings. Moreover, the SVM is capable of discovering MBSs where no pattern is available. In order to get some insight into the reasons for these predictions, we collected all the MBSs that do not contain any pattern, and divided them into examples actually predicted as MBSs by the SVM (true positives) and examples wrongly predicted as DB (false negatives). Figure 3 (top and bottom) represents the sequence logos for true positives and false negatives, respectively, where the logos are computed using the average of the PSI-BLAST profiles for all the examples. Part of the correct predictions could be explained by the ability of the SVM to actually recover continuous-valued patterns in the profiles. We conjecture that in some other cases the predictor could have discovered potential consensus patterns that are still not known.

Table V. DB versus metal binding prediction by third degree polynomial kernel SVMs

Ligand	No. of examples	Recall (%)
Cysteine	2860	97.5
Metal	758	61.4
4FE4S	250	71.6
Zinc	156	38.5
Heme	139	87.8
2FE2S	91	58.2
Copper	50	38.0
Iron	26	42.3
3FE4S	25	48.0
Nickel	13	76.9
Mercury	7	00.0
Manganese	1	00.0

Window of 17 residue profiles on both sides of the target cysteine. Recall and number of examples for cysteine (DB) or metal ligand, and details for different kinds of metal binding.

Reported results are quite robust with respect to model regularization, controlled by the cost parameter ‘C’ in SVMs. Figure 4 shows a test accuracy maximization obtained by

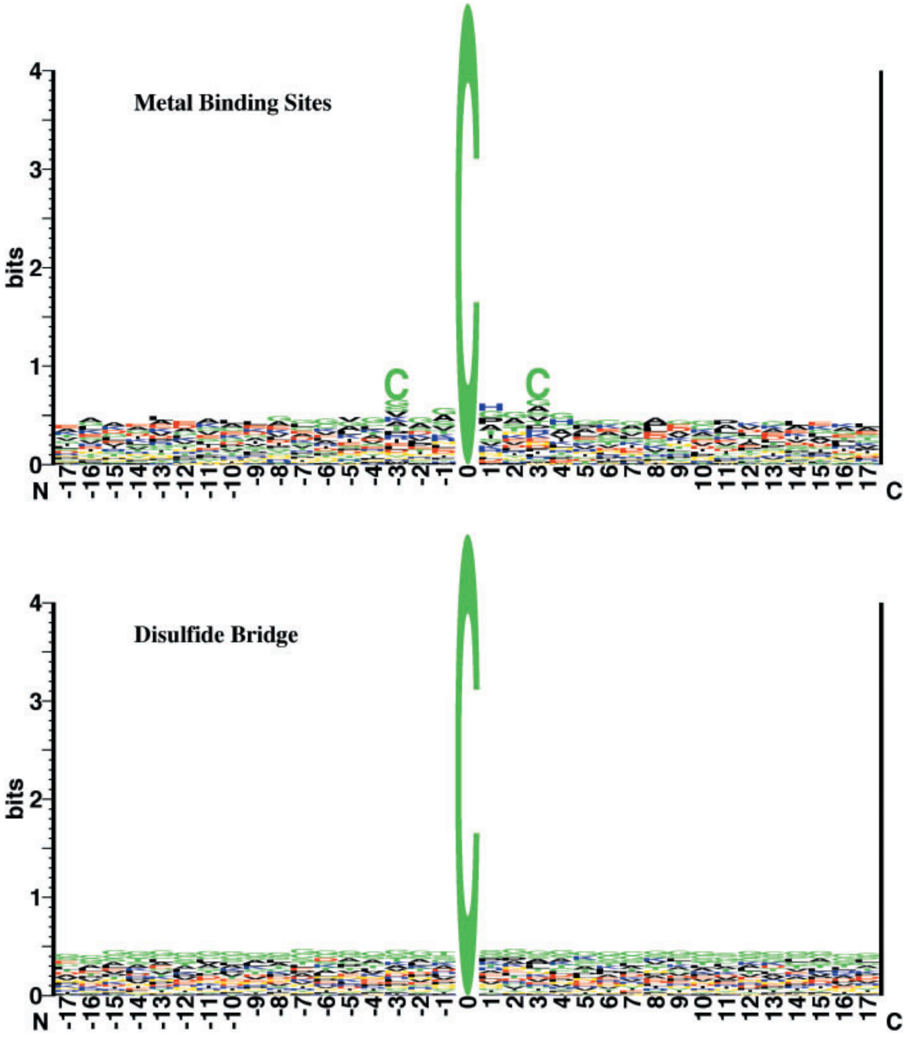

Fig. 2. Sequence logos (Schneider and Stephens, 1990) in the context of cysteines involved in metal binding (top) and DBs (bottom), respectively, with a window of 17 residues on each side of the cysteine bond. Hydrophobic residues are shown in black, positively charged residues are blue and negatively charged residues are red, while uncharged polar residues are green.

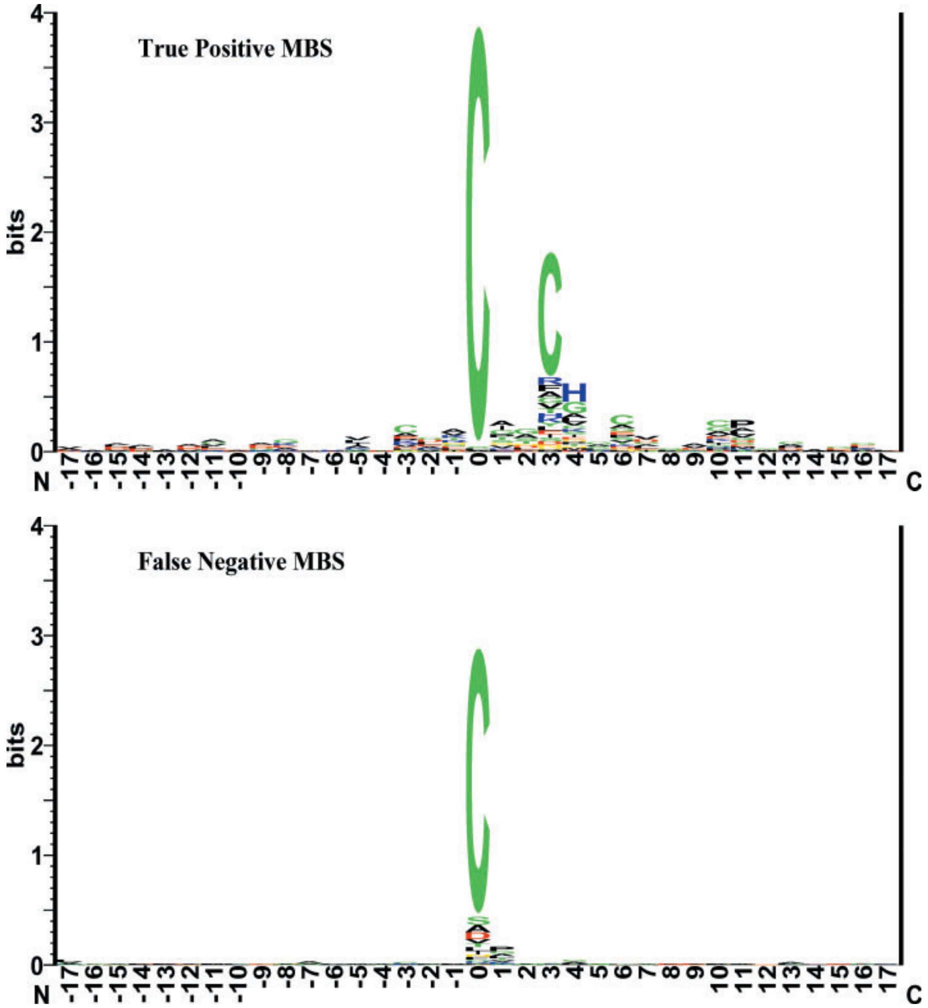

Fig. 3. Sequence logos (Schneider and Stephens, 1990) in the 17 aminos context of cysteines that do not match any PROSITE pattern, and are either truly predicted as MBSs (top) or mistakenly predicted as DB (bottom) by an SVM with a third degree polynomial kernel. Hydrophobic residues are shown in black, positively charged residues are blue and negatively charged residues are red, while uncharged polar residues are green.

varying the regularization parameter: accuracies remain mostly within the confidence interval from the maximum. Table VI shows test accuracies for different rejection rates, when predictions are to be made only if they are over a certain confidence. A rejection of 15% of both positive and negative test examples results in ~94% test accuracy, and predictions on DBs tend to be much more confident than those on metal bindings, having a higher threshold for equal rejection rate. Furthermore, metal bindings are rejected more frequently with respect to their total number in the test set. A finer analysis shows that rejected metal bindings are mostly those for which the model has low recall (see Table V), such as 3FE4S, iron, mercury and zinc, while 4FE4S, heme and nickel are seldom rejected. Figure 5 (left and right) shows precision/recall curves for DBs and metal bindings, respectively, at different rejection rates. The former tends towards the optimal curve at growing rejection rates, while the latter has a complementary behavior with respect to the breakeven point: at higher precisions growing rejection rates result in better performance, while at lower precisions performance for growing rejection rates is affected by the greater quantity of metal bindings rejected with respect to DBs.

Figure 6a shows results for the growing size of the context window, for a third degree polynomial kernel with McLachlan similarity matrix (Equation 4). While train and test accuracies are similar to those obtained without the similarity matrix (Figure 1), the corresponding models have less support vectors, with reductions up to 11% of the training set. This behavior is even more evident for the Blosum62 substitution matrix (Figure 6b) where a slight decrease in test accuracy, still within the confidence interval, corresponds to a reduction up to 30% of the training set. Note that the fraction of support vectors over training examples is a loose upper bound on the leave one out error (Vapnik, 1995), which is an almost unbiased estimate of the true generalization error of the learning algorithm (see, for example, Elisseeff and Pontil, 2003). These kernels are able to better exploit information on residue similarity, thus obtaining similar performances with simpler models. Moreover, the precision/recall rate of the kernel with the Blosum62 matrix is much more balanced with respect to the other two, meaning that it suffers less from the unbalancing in the training set.

Conclusions

We have proposed learning algorithms for predicting the type of binding in which non-free cysteines are involved. The experimental results indicate that learning from multiple alignment profiles data outperforms even non-trivial approaches

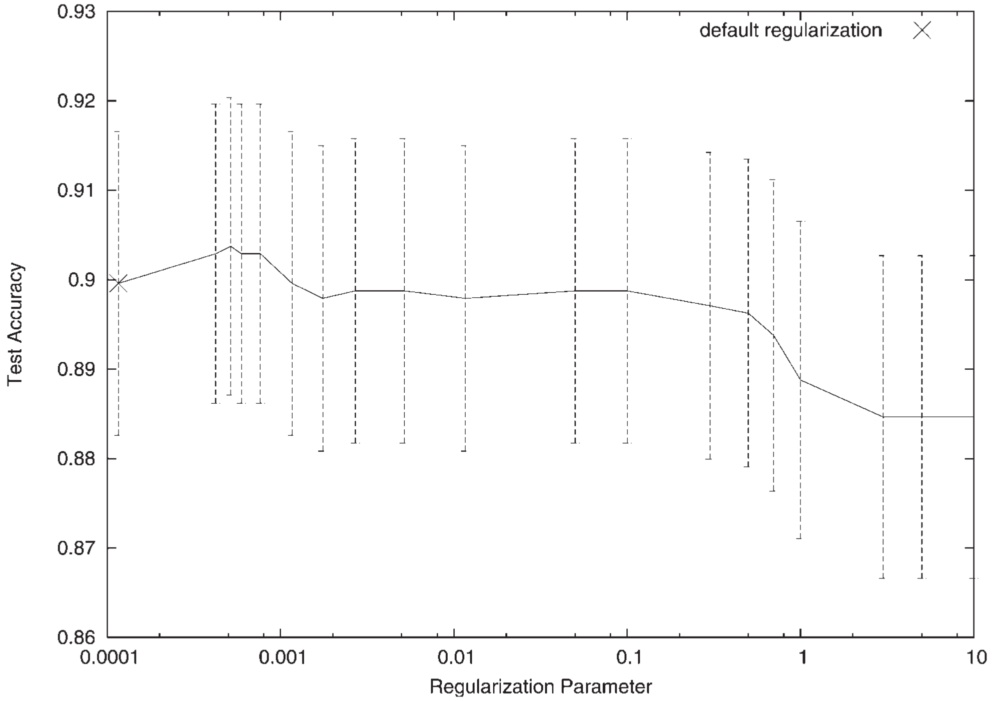

Fig. 4. DB versus metal binding prediction by third degree polynomial kernel SVMs. Window of 17 residue profiles on both sides of the target cysteine. Test accuracies with 95% confidence intervals are plotted versus the 'C' regularization parameter of SVMs. Default regularization is computed as the inverse of the average of $K(x, x)$ with respect to the training set.

Table VI. DB versus metal binding prediction by third degree polynomial kernel SVMs

	Threshold			Bridge		Metal		Rejected (%)	
Rejection (%)	Bridge	Metal	Acc	Pre	Rec	Pre	Rec	Bridge	Metal
00	0	0	89.9	90.5	97.5	86.6	61.5	0.0	0.0
05	0.20845	0.0470436	91.8	92.7	97.6	86.9	67.4	02.8	12.9
10	0.340927	0.113824	93.0	93.8	97.9	88.5	71.5	07.1	20.8
15	0.436391	0.159688	94.0	94.8	98.1	89.3	75.2	11.4	28.2
20	0.506477	0.225416	94.4	95.2	98.2	90.3	76.6	16.5	33.0
25	0.56122	0.267722	95.2	95.9	98.4	91.1	79.5	21.3	38.8
30	0.609002	0.337221	95.7	96.1	98.8	93.4	80.8	26.6	42.3

Window of 17 residue profiles on both sides of the target cysteine. Performances for different rejection rates, where rejection percentage is computed separately for examples predicted as bridge or metal, in order to consider the unbalanced distribution of examples between the two classes (i.e. 5% rejection rate indicates that 5% of examples predicted as bridges are to be rejected, as well as 5% of examples predicted as metal). Reported results include rejection thresholds, accuracies (Acc), precision (Pre) and recall (Rec), and percentage of rejected examples belonging to each of the two classes.

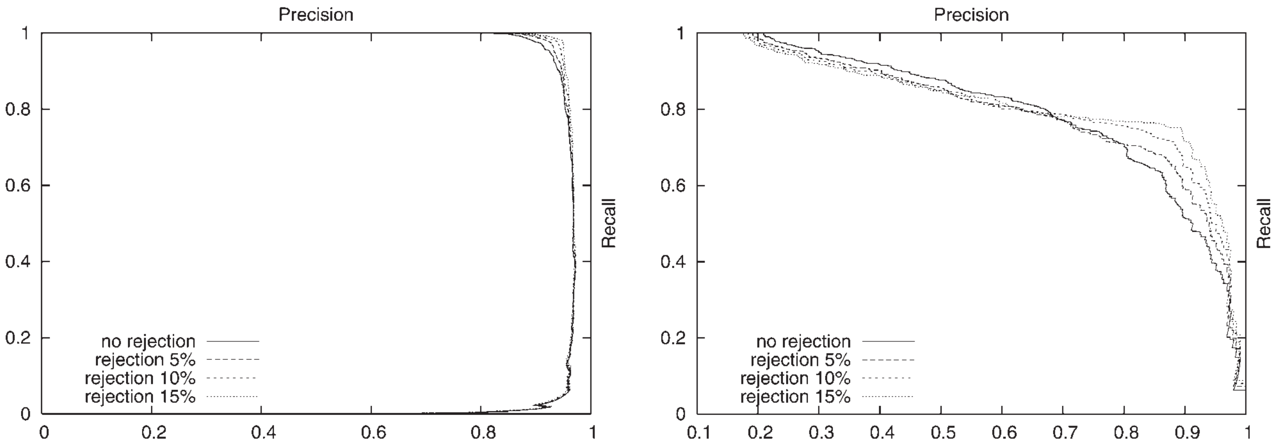

Fig. 5. DB versus metal binding prediction by third degree polynomial kernel SVMs. Window of 17 residue profiles on both sides of the target cysteine. Precision/recall curves over the test set for different rejection rates for the DB (left) and metal binding (right) prediction.

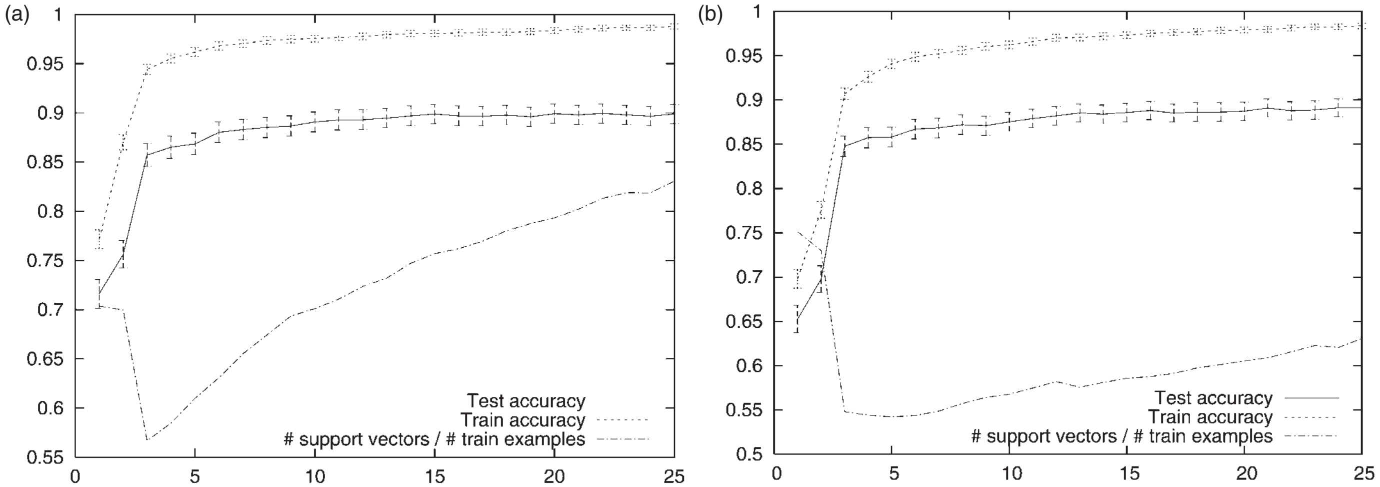

Fig. 6. DB versus metal binding prediction by SVMs with third degree polynomial kernel with McLachlan (**a**) or Blosum62 (**b**) similarity matrix. Test and train accuracies with 95% confidence intervals are plotted, together to the fraction of support vectors over the number of training examples, for growing sizes of the window of $2k+1$ residue profiles around the target cysteine, with k going from 1 to 25. Results are averaged over a 3-fold cross validation procedure.

based on pattern matching, suggesting that relevant information is contained not only in the residues flanking the cysteine but also in their conservation. In some cases the learning algorithm could be actually exploiting in a probabilistic way patterns that are not yet listed in PROSITE, although detecting them with few data points is difficult. In addition, we found that using residue similarity matrices effectively reduces the complexity of the model measured by the number of support vectors, which is also an indication of the expected generalization error. We expect that similar kernel functions could be also useful for other predictive tasks both in one dimension (e.g. solvent accessibility) and two dimensions (e.g. contact maps), especially if task-dependent similarity matrices could be devised.

Acknowledgements

We thank B.Rost and M.Punta for providing part of the data used in the experiments and S.Ciofi-Baffoni for useful discussions. This work is partially supported by Italian MIUR grant No. 2002093941.

References

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

Balatri,E., Banci,L., Bertini,I., Cantini,F. and Ciofi-Baffoni,S. (2003) *Structure*, **11**, 1431–1443.

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

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M. C., Estreicher,A., Gasteiger,E., Martin,M. J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) *Nucleic Acids Res.*, **31**, 365–370.

Chinenov,Y. (2000) *J. Mol. Med.*, **78**, 239–242.

Cortes,C. and Vapnik,V. (1995) *Machine Learning*, **20**, 1–25.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, New York, NY.

Elisseeff,A. and Pontil,M. (2003) In Suykens,J., Horvath,G., Basu,S., Micchelli,C. and Vandewalle,J. (eds), *Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences*. IOS Press, Amsterdam, Vol. 190, pp. 111–130.

Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C., Hofmann,K. and Bairoch,A. (2002) *Nucleic Acids Res.*, **30**, 235–238.

Fiser,A. and Simon,I. (2000) *Bioinformatics*, **16**, 251–256.

Gattiker,A., Gasteiger,E. and Bairoch,A. (2002) *Appl. Bioinform.*, **1**, 107–108.

Guermeur,Y., Lifchitz,A. and Vert,R. (2004) In Schölkopf,B., Tsuda,K. and Vert,J.P. (eds), *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA.

Heaton,D., George,G., Garrison,G. and Winge,D. (2001) *Biochemistry*, **40**, 743–751.

Henikoff,S. and Henikoff,J. G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Joachims,T. (1998) In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA, pp. 169–185.

Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) *Protein Eng.*, **15**, 951–953.

Mathews,F. (1985) *Prog. Biophys. Mol. Biol.*, **45**, 1–56.

McLachlan,A. (1972) *J. Mol. Biol.*, **64**, 417–437.

Mika,S. and Rost,B. (2003) *Nucleic Acids Res.*, 31, 3789–3791.

Mucchielli-Giorgi,M., Hazout,S. and Tuffèry,P. (2002) *Proteins*, **46**, 243–249.

Otaka,E. and Ooi,T. (1987) *J. Mol. Evol.*, **26**, 57–267.

Quinlan,J. (1986) *Machine Learning*, **1**, 81–106.

Quinlan,J. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA.

Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.

Schölkopf,B. and Smola,A. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Received March 5, 2004; accepted May 4, 2004

Edited by Marius Clore