

Identifying cysteines and histidines in transition metal binding sites using support vector machines and neural networks

Andrea Passerini^(1,*), Marco Punta^(2,3), Alessio Ceroni⁽¹⁾, Burkhard Rost^(2,3), Paolo Frasconi⁽¹⁾

(1) Università degli Studi di Firenze, Dipartimento di Sistemi e Informatica

Via di Santa Marta 3, 50139 Firenze, Italy

(2) CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th

Street BB217, New York, NY 10032, USA

(3) Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie

Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA

(*) Corresponding author: *passerini(at)dsi.unifi.it*

Tel: +39-0554796361, fax: +39-0554796363

Running Title: CYS and HIS metal bonding state prediction

Keywords: metal binding sites, disulfide bridges, protein function prediction, neural networks, support vector machines

Abstract

Accurate predictions of metal binding sites in proteins by using sequence as the only source of information can significantly help in the prediction of protein structure and function, genome annotation, and in the experimental determination of protein structure. Here, we introduce a method for identifying histidines and cysteines that participate in binding of several transition metals and iron complexes. The method predicts histidines as being in either of two states (free or metal bound) and cysteines in either of three states (free, metal bound or in disulfide bridges). The method uses only sequence information by utilizing position specific evolutionary profiles as well as more global descriptors such as protein length and amino acid composition. Our solution is based on a two-stage machine learning approach. The first stage consists of a support vector machine trained to locally classify the binding state of single histidines and cysteines. The second stage consists of a bi-directional recurrent neural network trained to refine local predictions by taking into account dependencies among residues within the same protein. A simple finite state automaton is employed as a post-processing in the second stage in order to enforce an even number of disulfide bonded cysteines. We predict histidines and cysteines in transition metal binding sites at 73% precision and 61% recall. We observe significant differences in performance depending on the ligand (histidine or cysteine) and on the metal bound. We also predict cysteines participating in disulfide bridges at 86% precision and 87% recall. Results are compared to those that would be obtained by using expert information as represented by PROSITE motifs and, for disulfide bonds, to state-of-the-art methods.

Abbreviations used: SVM, Support Vector Machines; BRNN, Bi-directional Recurrent Neural Networks; 3D, Three-Dimensional; MBS, Metal Binding Sites; DB, Disulfide Bridges; SS, Secondary Structure; AUC, Area Under the ROC Curve.

Availability: The software is available from the corresponding author upon demand.

1 INTRODUCTION

Metal binding sites in proteins. A significant fraction (about one third)^{1, 2, 3} of all known proteins is believed to bind metal ions as cofactors in their native conformation. Metal ions in proteins perform multiple tasks: they help stabilizing protein structure⁴, induce conformational changes^{5, 6, 7} and assist protein function (e.g. electron transfer, nucleophilic catalysis). Metal binding sites (MBS) are generally characterized through the ion (or ions) bound, through the protein atomic groups (ligands) directly involved in binding, through their coordination number (overall number of ligands) and geometry (for a detailed analysis of MBS in the Protein Data Bank structures, see Harding⁸). Metals with a prominent biological role include alkali (K,

Na), alkaline earth (Mg, Ca) and several transition metals in different ionization states, most importantly Mn, Fe, Cu, Zn and Cd, together with the less abundant Mo and W. Given the set of PDB chains containing MBS, about 66% of them bind transition metals, about 37% bind alkaline earth metals, and only about 6% bind alkali metals. Note that in 10% of the cases the same chain simultaneously binds metals of different types (e.g. a transition metal and an alkaline earth one). We are not aware of any estimate on a genomic scale. Different metal groups exhibit different modalities of binding. Alkali and alkaline earth metals bind proteins predominantly through electrostatic interactions, while in the case of transition metals the protein ligands donate an electron pair so to form coordinate covalent bonds (for this reason ligands are also commonly referred to as "donors"). As a consequence, alkali metals have generally low binding affinities, while alkaline earth (twice the charge of alkali) and transition metals can interact more strongly with proteins. Most common ligands are sulfur, nitrogen and oxygen atoms, including backbone carbonyls (particularly, in binding sites that involve alkali and alkaline earth ions). Additionally, water and other small molecules can complete the metal coordination shell. At the amino acid level, the most common metal-binding residues are CYS, HIS, ASP, GLU and, more rarely, MET, ASN, GLN, SER, THR and TYR. Beside the amino acids that participate directly to the chemical bond (first coordination shell), other residues (second and third coordination shells) can be important for structural stabilization of the binding site or for assisting enzymatic activity. Coordination numbers vary considerably throughout the different metals, ranging from a minimum of one to a maximum of about eight. Although most metals have strong preference for a specific coordination number (e.g. six for Ca, four for Zn), they also show a fairly high degree of flexibility (most Ca have from four to seven ligands, most Zn from three to five). The geometry of a binding site results from the optimization of the energetic interactions between ligands and metals and between the ligands themselves. However, even binding sites that share the same metal, coordination number and ligands, can display different geometries due, for example, to a different spacing of the ligands along the protein sequence⁸. Possible geometries include trigonal, tetrahedral, square planar and trigonal pyramidal, trigonal bi-pyramidal and octahedral. To make the situation even more intricate, different metals can fit into the same binding site (sometimes inducing a different MBS geometry, see for example Hertweck and Mueller⁹). Throughout this paper, we will refer to metal-binding amino acids as 'ligands', although the term is generally used to indicate only the atomic groups participating to the covalent bond.

Need for de novo metal binding site predictions. A few recurrent metal-binding sequence-motifs have been identified so far and have been used to predict MBS in proteins for which no direct experimental evidence is available. However, these motifs, while producing fairly accurate predictions, can cover only a small fraction of all known MBS^{10, 11} (see also the Results section) and clearly cannot predict novel binding

sites. Some very well characterized sites have been used in studies of automated sequence design (that is, using known structure scaffolds) (see Lu et al.¹² for a review) as well as of *de novo* protein design (that is, designing the scaffold along with the binding site¹³). While these studies have scored amazing successes, attempts to predict metal-binding sites on a large (typically genomic) scale, when no known sequence motif or 3D structural template is available, have been to date very limited. This is somewhat surprising, given the fact that a method capable of predicting MBS *de novo* would be of invaluable help in protein structure and function prediction, and in experimental structure determination. In particular, it would be extremely relevant for large-scale initiatives, such as Structural Genomics. Structural Genomics ultimate goal is to map the protein sequence space by providing at least one experimental structure for each sequence-structure family¹⁴. As a result of the target selection strategy, targets pursued by Structural Genomics consortia are often proteins for which little or no prior experimental knowledge is available. In this context, knowing that a target protein binds to a metal can be crucial for determining the conditions under which that protein is to be expressed or purified. At the same time, it can provide important clues about the protein function and allow targeting of specific functional families. Recently, Shi et al.¹⁵ have proposed a method for high-throughput identification of metal-binding proteins by means of X-ray absorption spectroscopy. While this methodology provides experimental evidence of metal binding in proteins, it cannot identify the actual ligands within the protein. Hence, the approach appears to be complementary to the method that we introduce here.

A novel sequence-based method for prediction of metal binding sites in proteins. Probably, one of the reasons why progress in the development of *de novo* metal-binding site predictors has been so uncharacteristically (for computational biology) slow is the intrinsic difficulty of the task at hand. Metal binding sites appear as very compact units in 3D and can be predicted with good success once the protein 3D structure is known^{16, 17, 18}. In contrast, predicting MBS from sequence only^{11, 19, 20} is a much more challenging task. Often, ligands are distributed along the protein sequence with no clearly recognizable spacing rule and the same metal can display a high degree of variability as far as the MBS ligand-composition is concerned⁸. In this work, we made an attempt at reducing the complexity of the problem by selecting a subset of "well-behaved" metals and ligands, as described hereafter. As far as ligands are concerned, alkali and alkaline earth MBS are the most diverse. Due to the fact that binding of these metals often involves main-chain carbonyls, virtually any residue can qualify as a ligand. In contrast, transition metals present a more limited spectrum of binding amino acids, mostly comprising CYS, HIS, ASP, GLU and MET. Still, ASP, GLU and MET are rarely found in MBS, when compared to their natural frequency of occurrence in proteins (see Table I). Following these considerations, we decided to restrict our predictions to 1) transition met-

als, with the addition of heme and Fe/S clusters which are pretty common and bind primarily CYS and HIS^{21, 22}(exceptions are, for example, heme binding protein catalases that have TYR as a proximal ligand, i.e. the axial ligand to the iron.) 2) CYS and HIS ligands. In this way, we attempted to predict only ligands that often have side-chain specific interactions with the metal (hence, they are likely to be conserved in the course of evolution) and are enriched in their frequency of occurrence in MBS. Clear drawbacks were that we gave up the chance of predicting binding sites for some of the biologically relevant metals, namely alkali and alkaline earth metals. Also, in MBS comprising other ligands (e.g. ASP and GLU) along with HIS and CYS, we were discarding *a priori* part of the information about the global architecture of the binding site (see Results for a more detailed discussion about these incomplete binding sites). Throughout this work, we made no distinction between mononuclear (one ion) and multinuclear (more than one ion) binding sites and between different ionization states of the same metal. We developed a two-stage architecture based on machine learning techniques predicting HIS as either free or metal-bound and CYS as free, metal bound or part of a DB. We introduced the disulfide bond state for CYS as it presents distinct characteristics with respect to the other two states (free and metal bound) and it is in itself very relevant for the structural and functional characterization of a protein. We compared the predictive performance with the results that would be obtained by using expert information as represented by PROSITE²³ motifs.

Related Works. This paper substantially extends our previous work¹¹ which aimed at discriminating between metal bound and disulfide bound CYS. By adding the free state, it allows the predictor to be employed without any need for *a priori* information on CYS oxidation state. By adding HIS amino acids, it broadens the class of MBS that can be treated. Moreover, it adds a post-processing stage to single-residue raw predictions, which aims at capturing the non-local nature of MBS and DB (for details, see Methods). DB prediction has been extensively studied in recent years and a number of successful machine learning approaches exist^{24, 25, 26, 27}. Conversely, little attention has been focused on predicting MBS. The only exception is a recent work¹⁹ which assumes to know that a protein contains a MBS and mainly focuses on alkali and alkaline earth metals. We are currently working on a specialized method for zinc binding proteins²⁰, which tries to model the linkage between close residues with ad-hoc solutions based on classification of nearby residue pairs.

2 MATERIALS AND METHODS

Protein Data. Protein sequences and structures were collected from the Protein Data Bank²⁸. We obtained a sequence-unique subset of 2,982 protein chains by running UniqueProt²⁹ with HSSP-value³⁰ threshold

equal 0 (i.e. no pair of proteins in the dataset had HSSP-value > 0). When clustering sequences in the PDB, we first selected as cluster seeds proteins containing DB and/or MBS. This allowed increasing the number of such proteins in our final sequence-unique dataset. CYS were labeled as being in DB or in free state according to the DSSP program³¹ assignment. Inter-chain bonded CYS were considered free in order to always have an even number of DB CYS within a single chain²⁴. In an attempt at reducing noise in the data, we visually inspected protein structures in all cases in which two CYS were found within a distance of 2.5 Å, but were not labeled by DSSP as being in DB. As a consequence, in 124 cases we over-ruled the DSSP assignment from free to disulfide bonding. A typical example is the human serum transferrin (PDB code 1a8e), where all CYS are clearly pairwise connected while DSSP labels all of them as free. Metal binding sites were detected by calculations performed directly on the protein 3D coordinate files, according to the following procedure. We parsed the files looking for transition metals and transition metal complexes. Any residue having a heavy atom within 3.0 Å of the metal (or complex) was labeled as a ligand (for ligand and metal-specific distance cutoffs see Harding⁸). In heme and Fe/S complexes, residues binding to the porphyrin ring and to the sulfur atoms were also considered as ligands. Chains containing a single metal binding residue were discarded as outliers. In Table I, we report the list of metals and complexes that we took into consideration, along with the statistics relative to their ligands. As discussed in the Introduction, we focused on biologically relevant transition metals only, further discarding those for which there were too few cases, and limited our predictions to CYS and HIS, the only amino acids that are reasonably enriched in MBS. In the end, we were left (see Table II) with a total of 2,727 chains^a of which 1,722 contained only CYS and HIS in the free state, 687 had at least one DB and 365 had at least one MBS (with 47 chains having both a MBS and a DB).

This dataset was used to train a two-stage machine learning based program, consisting of a local stage implemented with Support Vector Machines (SVM), and a refinement stage consisting of a Bi-directional Recurrent Neural Networks (BRNN) plus a finite state automaton (FSA) enforcing an even number of predicted DB CYS in a single chain (inter-chain DB are ignored). In what follows, we first describe the machine-learning approaches employed and then we discuss the choice of the features that we used as input for their application to the prediction of MBS and DB.

Support Vector Machines. Support vector machine (SVM) learning^{32, 33} is a mature approach for classification and regression and has been applied to several prediction problems in bioinformatics^{34, 35, 36}. While SVMs have been initially conceived for binary classification, different approaches exist in order to extend them to the multiclass case, either by reducing the problem to a combination of binary subproblems, or

^aThe dataset is available at <http://www.dsi.unifi.it/~passe/datasets/mbs06/dataset.tgz>

Table I: Percentage and fraction of times a given amino acid type binds a specific metal ion (or complex) in chains containing a binding site for that ion. For example, in proteins containing a Zn binding site (first line of the Table) 508 out of 1,115 CYS bind to the metal (according to our ligand definition, see Materials and Methods); this corresponds to 46% of the CYS residues in those proteins. Note that since the same protein can have more than one metal bound, the overall number of residues in the line "any" (1,923 in the case of CYS), is always lower-equal than the sum of the number of residues found in the lines for individual metals. On the other hand, there was only one case in our dataset for which one residue was annotated as bound to two metals in the same protein (an ASP residue); as a consequence, the number of residues bound in the line "any" (930 in the case of CYS) is equal to the sum of the number of residues bound found in the lines for individual metals (that is, except for ASP).

Metal	CYS	HIS	ASP	GLU	MET	GLN	ASN
Zn	46 (508/1115)	24 (374/1562)	4 (117/3204)	2 (89/3705)	0 (0/1132)	0 (9/2047)	0 (2/2442)
heme	50 (115/230)	34 (151/450)	1 (5/854)	0 (2/925)	6 (23/367)	1 (7/593)	0 (1/571)
Fe/S	63 (205/326)	3 (10/329)	0 (0/763)	0 (1/886)	0 (0/372)	0 (0/407)	0 (0/485)
Cu	33 (36/108)	32 (86/269)	0 (2/513)	0 (2/455)	4 (9/228)	0 (0/251)	0 (0/391)
Cd	62 (48/77)	32 (25/79)	12 (26/216)	13 (35/262)	0 (0/44)	1 (1/158)	0 (0/176)
Fe	13 (16/122)	18 (59/325)	2 (15/610)	3 (25/745)	0 (0/189)	0 (0/364)	1 (2/381)
Ni	4 (2/46)	16 (18/112)	2 (5/250)	1 (2/271)	0 (0/100)	0 (0/132)	1 (1/153)
Any	48 (930/1923)	25 (723/2942)	3 (169/6045)	2 (156/6836)	1 (32/2290)	0 (17/3781)	0 (6/4339)

Table II: Number of CYS, HIS and entire chains for the three different classes.

	Free	DB	MBS
CYS	4702	3552	933
HIS	12982	0	678
Chains	1722	687	365

by directly defining a multiclass optimization problem (see Hsu Lin³⁷ for a review). We choose the latter approach in the implementation by Crammer and Singer³⁸. Several publicly available packages exist for solving the SVM optimization problem and its multiclass counterpart. In all the experiments reported in this paper we used *SVMLight*^b for binary classification (HIS) and *bsvm*^c for multiclass classification (CYS).

Bi-directional Recurrent Neural Networks and Overall Architecture. As explained above, the individual SVM predictions are obtained *independently* for each target CYS and HIS. While the independence assumption is reasonable for two residues that belong to different proteins (except perhaps when several proteins form a complex and the same metal ion is coordinated by different members of the complex) it is much less likely to hold for some of the residues that belong to the same protein. Metal ions are typically coordinated by several residues, whose bonding states become inevitably correlated. In addition, the formation of disulfide bridges is often a global phenomenon associated with whole chains and two behaviors are dominant in the data: either all or no CYS are disulfide bound³⁹. This again may introduce correlations. In fact, if we know e.g. that one CYS is free, then the probability that another CYS in the same chain is also free would be increased. Ignoring these effects in the model can lead to loss of prediction accuracy. The learning problem in the presence of correlation between instances is often referred to as *collective classification*⁴⁰ and has been tackled with some success using probabilistic graphical models. Presently, however, collective classification is not supported by SVM.

In this paper we attempt to recover some of the data correlation by designing a hybrid solution based on the combination of SVM predictions and bi-directional recurrent neural networks (BRNN). In Passerini et al.²⁰ we experimented specifically on zinc binding sites with a different technique, where data correlation is partially modeled by formulating the task as the classification of sequence close residue pairs. BRNNs were originally developed for learning temporal dependencies in the context of secondary structure prediction⁴¹ and later applied (with some variants) to other tasks such as the localization of disulfide bridges⁴² and the prediction of signal peptides⁴³. In the following we briefly review the basic ideas.

In the sequence learning problem solved by BRNNs, data come in the form of input-output sequence pairs. For the sake of simplicity we focus on a single pair, omitting integer subscripts that index different training sequences. We denote by $\mathbf{z} = [z(1), \dots, z(t), \dots, z(T)]$ the input sequence and by $\mathbf{y} = [y(1), \dots, y(t), \dots, y(T)]$ the corresponding target output sequence. In the present application, sequences have as many elements as CYS and HIS in the chain being considered. As illustrated in Figure 1, the input $\mathbf{z}(t)$ is formed by local predictions from the SVM stage and additional information describing

^b<http://svmlight.joachims.org/>

^c<http://www.csie.ntu.edu.tw/~cjlin/bsvm/>

the local environment around the target residue and the environment that separates it from the following CYS or HIS, as detailed in the BRNN input features section. When cascading two learning machines it is necessary to feed the downstream machine (the BRNN in this case) with the predictions (the bonding state in this case) obtained from the upstream machine (the SVM), rather than using true values. During training, however, feeding the BRNN with the margins calculated by the SVM on the training set data would bring an artifact in the distribution of BRNN inputs since the margins are exactly -1 or +1 for all the bound support vectors and these typically form a considerable subset of the training data. The problem can be avoided by training the BRNN on data not in the SVM training set. In order to achieve this goal and, at the same time not waste examples, we split the training set into three folds. The SVM was then trained on two folds and the predicted margins were obtained on the one left out. The operation was then repeated three times and predicted margins from the left out folds were merged together, thus recovering predictions for all examples in the original training set.

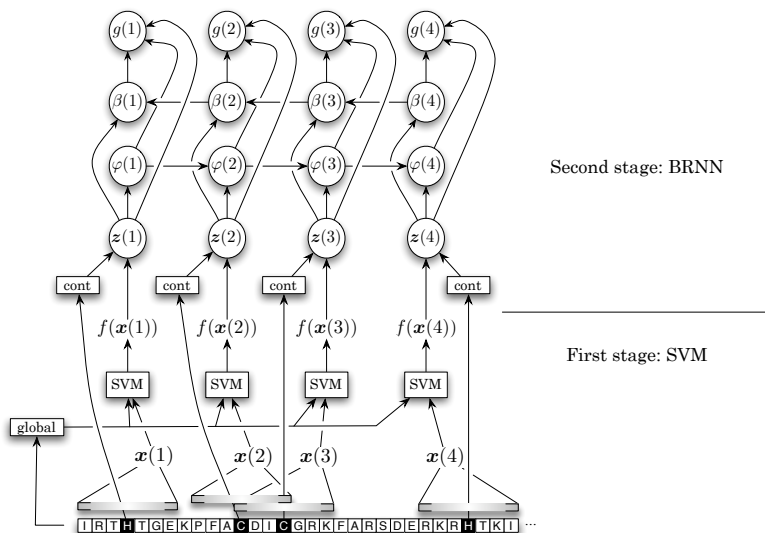


Figure 1: Architecture of the overall predictor. SVM predictions and other context information (boxes labeled “cont”) form the input to the BRNN.

The mapping between the input sequence and the output sequence in a BRNN is *global* in the sense that the predicted output $g(t)$ (actually the predicted output depends on the input sequence x and the position t but we denote it as $g(t)$ to keep notation simple) at every position $t = 1, \dots, T$ depends on BRNN inputs $z(\tau)$ at every other position $\tau = 1, \dots, T$ (see Figure 1). This can be achieved by introducing, for

each position t , two real vectors $\varphi(t) \in \mathbb{R}^d$ and $\beta(t) \in \mathbb{R}^d$ called the *forward* and the *backward* states, respectively. Intuitively the forward state at t , $\varphi(t)$, contains context information about the “past” substring $z(1), \dots, z(t)$; similarly, $\beta(t)$ contains context information about the “future” substring $z(t), \dots, z(T)$. These two vectors, together to the local input $z(t)$, are expected to contain all the information about the sequence that is needed to make a prediction at position t . A standard BRNN can be interpreted as the combination of two discrete-time dynamical systems defined by two recursive nonlinear update equations in which $\varphi(t)$ explicitly depends on $\varphi(t-1)$ and $\beta(t)$ on $\beta(t+1)$. An output function finally calculates $g(t)$ from $\varphi(t)$, $\beta(t)$ and $z(t)$. The output function and the two functions computing $\varphi(t)$ and $\beta(t)$ respectively are implemented by three feed forward neural networks with weights shared over time. This allows to have a fixed number of parameters regardless of the size of the input sequences, thus overcoming the limitation to fixed size inputs of standard feed forward neural networks. A softmax function is used to guarantee that $g(t)$ can be interpreted as (conditional) probabilities that residue at t is free, disulfide bound or metal bound respectively (given the input sequence). This propagation scheme is graphically illustrated in the upper part of Figure 1.

While this scheme does not allow us to model arbitrary correlations between the output targets at any two positions, it may be still useful in the present case since the bonding state at a given position is calculated not just using inputs for that position but also using inputs that may be separated in sequence and close in the 3D space. In this way, correlations between two outputs that are reflected into correlations between the corresponding inputs may be indirectly captured.

BRNNs are trained as typical neural networks by defining a cost function as the negative log likelihood of the data. In the present case the data consists of N proteins, each containing T_j CYS and HIS residues (for $j = 1 \dots, N$) whose binding state is known and denoted as $y_j(t)$, $t = 1, \dots, T_j$. The cost should measure the mismatch between predictions $g_j(t)$ and corresponding targets $y_j(t)$. In the case of classification (as in this application) the proper cost function is the cross-entropy:

$$\mathcal{L}(\theta) = - \sum_{j=1}^{\ell} \sum_{t=1}^{T_j} y_j(t) \log g_j(t; \theta) \quad (1)$$

where θ is the complete vector of parameters (connection weights) for the network. Gradient descent with early stopping is then used to minimize the cost and at the same time control overfitting.

In order to enforce an even number of DB CYS we employed the same simple FSA (shown in Figure 2) used in the DISULFIND predictor⁴⁴. Given the sequence of bonding state probabilities (computed by the BRNN) for all cysteines in a given chain, the most likely sequence of bonding states is obtained by running a Viterbi algorithm. Thus a cysteine predicted to be free or metal bonded by the BRNN can be relabeled

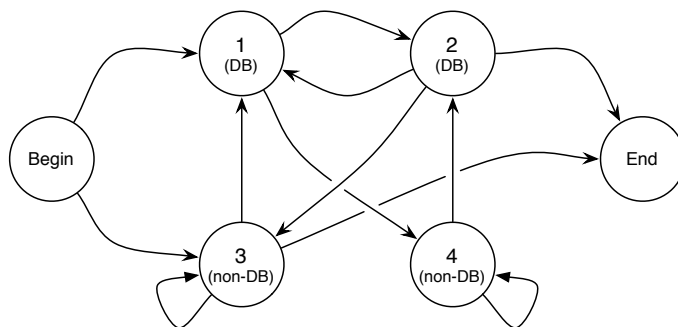


Figure 2: Finite state automaton used as a post-processing stage to enforce an even number of DB CYS.

as disulfide bonded by the FSA, or on the other hand a cysteine predicted to be disulfide bonded can be relabeled as free or metal bonded, depending on which of these two classes got the higher probability at the BRNN stage. Similar ideas (but using a hidden Markov models rather than an automaton) were presented in Martelli et al. (2004)²⁴.

Input Features: Support Vector Machines. Protein sequences were enriched by position specific evolutionary information in the form of profiles. Multiple alignments were obtained by running a single iteration of PSI-BLAST⁴⁵ on the non-redundant (nr) ncbi database using an E value cutoff of $5e - 3$. Individual instances for the first stage SVM predictor (see the SVM section) were then obtained by constructing a window of size $2w + 1$ residues flanking the target residue. Each position in the window was represented by the corresponding profile frequencies, plus a flag indicating positions ranging out of the sequence limits. Overall, this accounted for $(2w + 1) * 21$ input features. One additional feature coded for the relative position of the target residue with respect to the sequence length. We also considered a global descriptor, taking into account the characteristics of the entire protein. It consisted of 33 numeric features. The first 20 features described the normalized amino acid composition of the chain, each entry being computed as $\log(N_i^j / N_j)$ where N_i^j is the number of occurrences of the j -th amino acid in the i -th chain, while N_j is the number of occurrences of the j -th amino acid in the whole training set. More features included: sequence length relative to the average value in the training set, overall number of CYS and overall number of HIS in the chain, relative both to their average values in the training set and to the sequence length itself, and a parity flag for the number of cysteines. Finally, ten features encoded for the average conservation of the CYS and HIS in the chain, as computed from PSI-BLAST profiles. Conservation values were discretized into five equal length bins $([0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1])$ and a one hot encoding of the bins was employed (i.e. a conservation value of 0.57 was represented as 00100).

Input Features: Bi-directional Recurrent Neural Networks. In the refinement stage (see the BRNN section), we enriched SVM predictions for each target residue with additional features describing the residue and its context (box `cont` in Figure 1). Residue features consisted of conservation value, coded using five equally spaced bins as we did for the global descriptor (see previous section), position relative to the sequence length, and a flag indicating the residue type (CYS/HIS). Context features included sequence separation information and secondary structure (SS) predictions. For each target residue, we computed the sequence separation to the following target if any, and discretized such a value into eight bins ($[1, 1], [2, 2], [3, 3], [4, 4], [5, 6], [7, 20], [21, 75], [76, 200], [201, \infty)$). The last bin was also used for terminal target residues (i.e. target residues not followed by any CYS or HIS along the chain). A one hot encoding of the bins was employed as in previous section. Nine real values were employed to encode predicted secondary structure. The first three values represented predictions for the SS of the target residue, one for each of the three SS classes (helix, strand, coil). The next three values encoded average SS predictions in a window of length five on both sides of the target residue. Finally, three more values represented the prediction of the average SS for the region separating the residue from the following target residue. SS Predictions were obtained using an SVM-BRNN architecture similar to the one described in this paper⁴⁶.

Evaluation procedure and performance measures. The architecture was evaluated using a stratified five fold cross validation procedure. Note that the three fold splitting described in the BRNN section was performed in each round of this five fold procedure on the corresponding training set. The secondary structure predictor was also retrained on the five folds in order to assure full independence of the test set with respect to the training data. Standard performance measures include overall *accuracy*, which is the fraction of correct predictions over the total number of predictions, and class-wise *precision* and *recall*. For a given target class C , let true positives (TP) be the number of examples correctly predicted as belonging to C , false positives (FP) the number of examples incorrectly predicted as belonging to C , and false negatives (FN) the number of examples of class C incorrectly assigned to another class. Precision is the number of true positives over the total number of examples assigned to the class, that is $TP/(TP + FP)$. Recall (or coverage) is the number of true positives over the total number of examples belonging to the class, that is $TP/(TP + FN)$. In the case of binary classification, recall for the positive class is also known as *sensitivity* or true positive rate, while recall for the negative class is known as *specificity* or true negative rate. The strong imbalance of the data set (see Table II), however, can make accuracy a misleading parameter as it favors predictors biased towards the most populous class. Moreover, for predictors which output continuous values, such as SVM and BRNN, computing accuracy requires fixing a threshold on the output value, in order to produce a hard decision. A more informed alternative consists in plotting full curves of

the predictor performance when varying the decision threshold. Possible candidates in the case of binary classification are the recall precision curve and the so called Receiver Operating Characteristic (ROC) curve. The ROC curve plots true positive rate against false positive rate, which is computed as one minus the true negative rate. A single performance measure which is invariant to the decision threshold can be obtained by computing the area under such curves. The area under the ROC curve (AUC) has a number of desirable properties⁴⁷ and is becoming increasingly popular as a performance measure. We thus employed AUC for both model selection and final performance measurement. For multiclass outputs, we employed the simple multiclass extension of AUC proposed in Hand and Till⁴⁸. We also plotted recall precision curves in order to visualize the predictor behavior when varying the decision threshold. Whenever a hard decision had to be done, as for confusion matrices and detailed recall values for binding site type and coordination number, we assigned each example to the class achieving the maximum output, once the post-processing FSA stage had corrected possible inconsistencies in the disulfide bonding state predictions.

Model selection. We employed Gaussian kernels for both binary and multiclass SVMs, as they can model complex nonlinear interactions between input features. Model selection thus implied choosing Gaussian width and C regularization parameter, and was conducted in a preliminary phase by running a three fold cross validation procedure on the training set of the first fold. We found out that optimal values for hyperparameters were very stable when varying the size of the input window and across different folds. Therefore we kept them fixed across window sizes and folds ($\gamma = .05$, $C = .1$ for binary SVM, $\gamma = .05$, $C = 5$ for multiclass SVM).

PROSITE-based predictor. We developed a PROSITE-based predictor in a manner similar to our previous work¹¹ in order to compare our method to an informed baseline. A PROSITE²³ motif represents a biological significant portion of a protein sequence. It is described by either a pattern, consisting of an annotated regular expression (possibly enriched by rules written in ordinary English), or a profile, that is a table of position-specific amino acid weights and gap costs. We run the program ScanProsite⁴⁹ on our dataset searching for motifs listed in the release 19.11 (27 September 2005) of PROSITE. In a first experiment, we focused on PROSITE patterns only, excluding the highly probable and least informative ones^d. In this way we found 467 patterns which covered about 8% of the residues. Note that the coverage we obtained is significantly lower than that reported in our previous work¹¹. There are a number of reasons explaining such a difference. First, the highly probable patterns, which we excluded as they did not show performance improvements, would raise the coverage to about 20%. Second, histidine residues and chains without any DB or MBS have far less matches, and were not considered in our previous work. Many patterns (359) were

^dScanProsite `-s` option

perfectly specific with respect to the class label (free, DB, MBS), but they collectively covered just about 3% of the residues. Almost two thirds of the covered residues matched ambiguous patterns with respect to the class label. For this reason, it would have been difficult to hand-craft a prediction rule based on pattern matches. We thus represented each residue by the (possibly empty) bag of its matching motifs, and trained a multiclass SVM predictor with the same evaluation and model selection procedure used for the overall architecture (see previous sections). In a second experiment, we applied the same procedure but considered all available PROSITE motifs (i.e. patterns and profiles), obtaining a coverage of about 32% of the residues, with 9% of the residues covered by perfectly specific motifs (with respect to the class label).

3 RESULTS AND DISCUSSION

Comparison with baseline PROSITE-based predictor. The overall best results were obtained for a window of size $W = 15$, giving a multiclass AUC of 0.959 ± 0.002 where the confidence interval is the average standard error of the Wilcoxon-Mann-Whitney⁴⁷ statistic. The multiclass SVM predictors trained on PROSITE matches (see previous section) achieve an AUC of merely 0.608 ± 0.007 and 0.705 ± 0.006 for the pattern only and full motif case respectively, thus proving the significance of our approach. Table III reports confusion matrices and overall accuracies at both residue and protein level for the three predictors. Both SVM-BRNN-FSA and PROSITE motifs achieve quite balanced precision recall values, but the second has much lower values, especially for the DB and MBS classes, showing that available motifs are insufficient to generalized to unseen cases. On the other hand, PROSITE patterns obtain high precision but very low recall for the DB and MBS classes, showing that patterns are very specific but too sparse in order to produce a reasonable coverage of binding sites. If we compare SVM-BRNN and PROSITE patterns at the same level of precision on MBS, we see that at 80% precision, our method gives 56% recall (Fig. 3(a)) vs 19% of the PROSITE-based.

Comparison with state-of-the-art DB predictors. If we restrict ourselves to the binary classification task of discriminating between DB and non-DB CYS, our method achieves state-of-the-art^{24, 25, 26, 27} accuracy at both residue ($Q_r = 90\%$) and protein ($Q_p = 85\%$) levels. Note, however, that direct comparisons are difficult since performances were evaluated on different subsets of the PDB.

DB and MBS recall precision curves for CYS and HIS residues Figure 3(a) reports recall-precision curves for DB and MBS predictions for the overall best predictor. While DB are much easier to predict with respect to MBS, the latter can still be predicted with 67% precision/recall at the break-even, which is the point in the recall precision curve where the two values coincide. Figure 3(b) (right) reports recall-

Table III: Confusion matrices, recall precision for class and overall accuracies for the SVM-BRNN-FSA architecture and the PROSITE pattern based predictors. Confusion matrices have true class on rows, predicted class on columns. Q_r is the fraction of correctly predicted residues, Q_p the fraction of chains for which all residues are correctly predicted.

Predictor	SVM-BRNN-FSA					PROSITE patterns					PROSITE motifs				
Class	Pre (%)	Rec (%)	Free	DB	MBS	Pre (%)	Rec (%)	Free	DB	MBS	Pre (%)	Rec (%)	Free	DB	MBS
Free	95	96	17063	343	278	79	99	17548	68	68	83	88	15497	1650	537
DB	86	87	384	3090	78	83	09	3218	327	7	36	29	2308	1030	214
MBS	73	61	488	148	975	80	19	1309	0	302	45	38	852	148	611
Q_r (%)	92					80					75				
Q_p (%)	77					63					37				

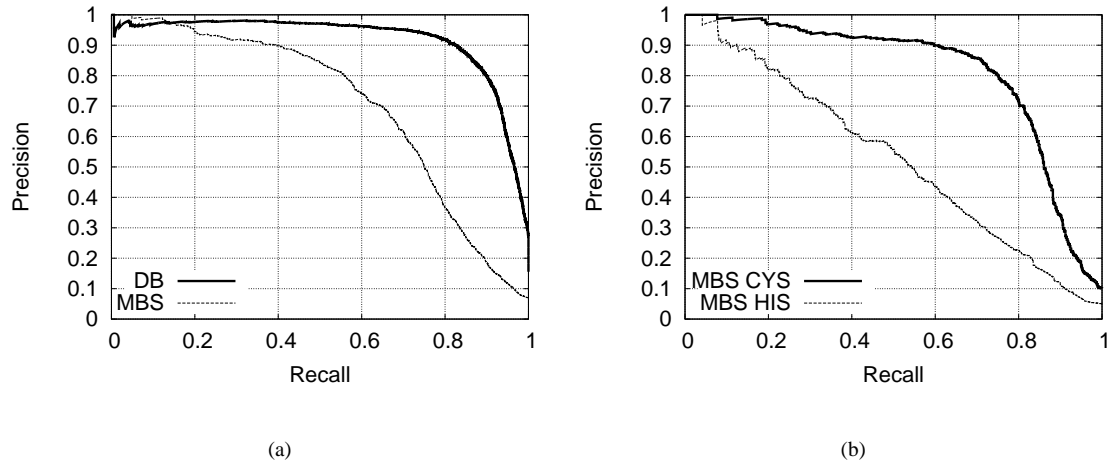


Figure 3: Recall precision curves for the best SVM-BRNN architecture ($W = 15$). (a) DB and MBS predictions. (b) MBS predictions for CYS and HIS residues separately.

Table IV: Recall values (in %) divided by metal and ordered by metal binding frequency, both overall and separate for CYS and HIS.

Metal	CYS+HIS	CYS	HIS
Zn	61 (534/872)	76 (388/512)	41 (146/360)
heme	79 (190/241)	96 (109/114)	64 (81/127)
Fe/S	70 (151/215)	71 (145/205)	60 (6/10)
Cu	41 (50/122)	56 (20/36)	35 (30/86)
Cd	34 (24/71)	50 (24/48)	0 (0/23)
Fe	28 (21/74)	69 (11/16)	17 (10/58)
Ni	25 (4/16)	50 (1/2)	21 (3/14)
Any .	60 (974/1611)	75 (698/933)	41 (276/678)

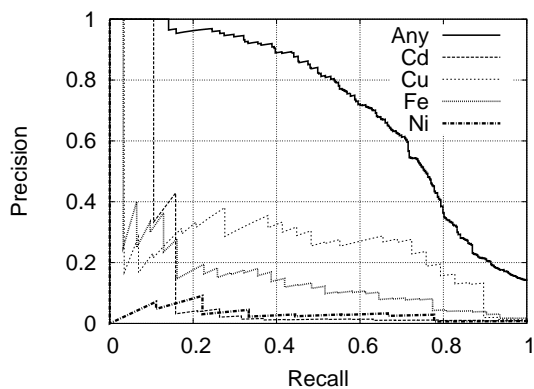
precision curves for MBS predictions for CYS and HIS separately, showing that the former are far better predicted. This is not particularly surprising as CYS profit from having more positive examples and a better balance between positive and negative in our database (see Table II).

Performance depends on metal and coordination number. Table IV reports recall values of MBS predictions separately for each metal, ordered by frequency of occurrence. Fe/S and heme clusters are the easiest to predict, while single ions such as Fe, Cd and Ni can be predicted with reasonable recall when binding CYS, but are extremely difficult to predict when HIS are involved. We further investigated prediction performance as a function of the metal coordination number (Table V). As expected, metals which coordinate with a single residue are virtually always lost (such cases can appear in the dataset if the chain containing them also contains a MBS with higher coordination number), and recall increases with growing coordination number, covering up to 75% and 71% of residues binding tetra- and penta-coordinated metals respectively. It is also interesting to note that there is a significant difference in the recall between 'complete' binding sites (i.e. binding sites that did not comprise any other ligand type besides CYS and HIS) and 'incomplete' ones (i.e. those including additional ligands that we do not predict such as GLU and ASP). The former are almost always far better predicted then the latter. Note that all penta- and hepta-coordinated metals have incomplete binding sites. In particular, of the 24 correctly predicted residues occurring in penta-coordinated metals, 23 bind metals coordinated by four CYS/HIS residues and just one different ligand.

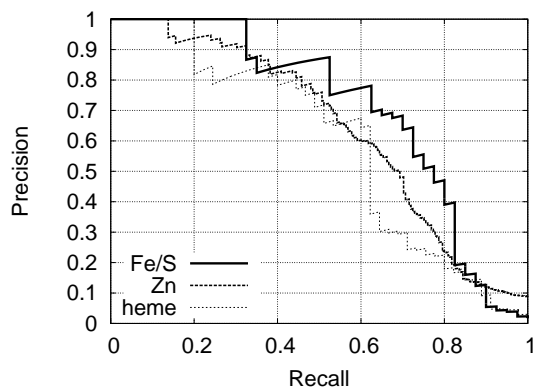
Metalloprotein predictions. At a protein level, we tested the performance of the method in detecting

Table V: Recall values divided by coordination number of MBS. In complete MBS metals bind CYS/HIS residues only, while in incomplete MBS other ligands (e.g. ASP GLU) are also involved.

Coordination	Recall (%)		
	All MBS	Complete MBS	Incomplete MBS
1	6 (2/33)	6 (2/33)	0 (0/0)
2	21 (31/150)	28 (29/104)	4 (2/46)
3	38 (128/334)	50 (109/216)	16 (19/118)
4	75 (777/1037)	79 (680/866)	57 (97/171)
5	71 (24/34)	0 (0/0)	71 (24/34)
6	63 (17/27)	50 (6/12)	73 (11/15)
7	45 (5/11)	0 (0/0)	45 (5/11)



(a)



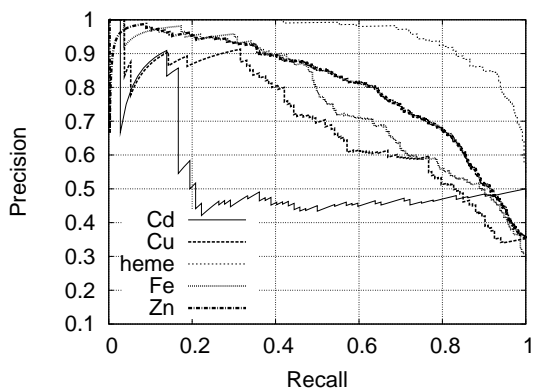
(b)

Figure 4: Recall precision curves for metalloprotein prediction. (a) All metalloproteins together, Cd, Cu, Fe and Ni binding proteins. (b) Fe/S, Zn and heme binding proteins.

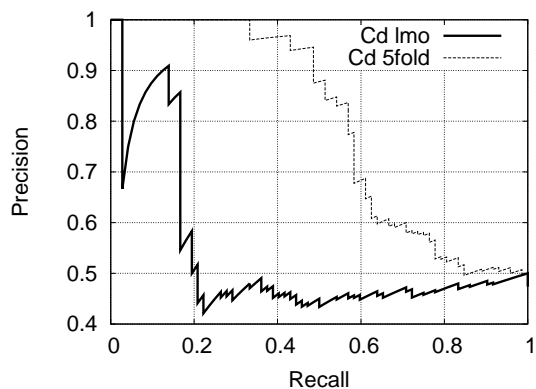
metalloproteins. Protein level predictions were obtained by choosing the maximal MBS prediction among those of the residues contained in the chain. Figure 4(a) reports the recall precision curve (curve *Any*) for such a task, showing a behavior very similar the one obtained at a residue level (see Fig. 3(a)). We also investigated detailed behaviors for different metals (Fig. 4(a-b)). For each metal, positive examples are proteins containing at least one MBS for that metal, and negative examples are non metalloproteins. Fe/S binding proteins are the easiest to predict, followed by Zn proteins and heme binding ones (Fig. 4(b)). Proteins binding other metals (Fig. 4(a)) are much harder to predict, and the scarcity of examples for such MBS (see Table IV) can partially explain this result. Cu proteins are best predicted with high recall, while Fe proteins are best predicted with high precision. Note that the recall precision curve for generic metalloprotein prediction (Fig. 4(a), curve *Any*) is mainly influenced by Fe/S, Zn and heme performance (Fig. 4(b)), as they constitute the vast majority of available MBS (see Table IV).

Generalization over metals. We further tested the ability of the method to generalize over metals, that is, to predict MBS for metals it had not been trained on. We think this may provide some interesting insight on the structural relationship between different MBS. For this purpose we employed a *leave-one-metal-out* procedure, that consisted in dividing the cofactors by metal, and for each of them, training the architecture on all the remaining metals, and testing it on the left out metal. Each test set was limited to proteins containing at least one MBS for the left out metal, thus omitting non metalloproteins in order to focus on performance for the metal under consideration. In Fig. 5(a) we report curves for the different metals. We merged Fe MBS with Fe/S ones, and omitted Ni as there were too few examples to make a significant comparison. In Fig. 5(b-f) we compare leave-one-metal-out results to those obtained when not excluding the metal from the training set. While it is clear that the method is able to generalize over metals, the performance strongly depends on the metal. Heme binding site predictions are almost unaffected by removing explicit examples from the training set; Zn and Fe seem also quite generic, while Cu and especially Cd suffer more for missing explicit training data. The different behavior of Cu and Zn is a little surprising since these two metals have been shown in some cases to bind to the same binding sites (though with different affinities, see Goto et al.⁵⁰). However, only 29 Cu binding chains were available, and their results are thus less reliable. It is interesting to note that the poor performance in Cd prediction is mainly due to the confusion with DB, as more than half of false negatives are actually predicted as DB.

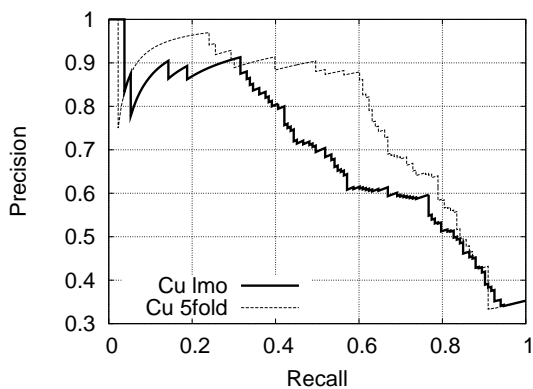
Architecture and Feature Relevance. The local SVM stage alone obtained an AUC of 0.955 ± 0.003 , which is within the standard error of the full architecture AUC (0.959 ± 0.002). However, by plotting recall precision curves for the two stages (Fig. 6(a-b)), we see that while the BRNN refinement only slightly improves over the local curve for DB predictions, it strongly improves MBS ones (Fig. 6(a)), especially



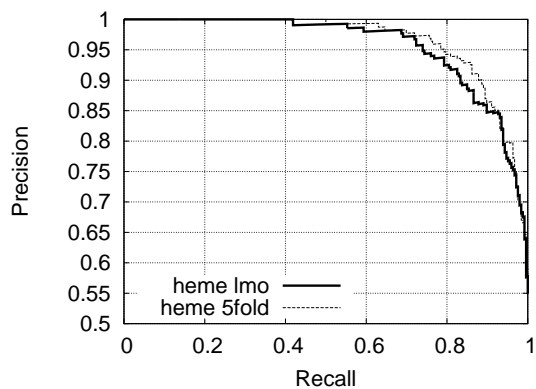
(a)



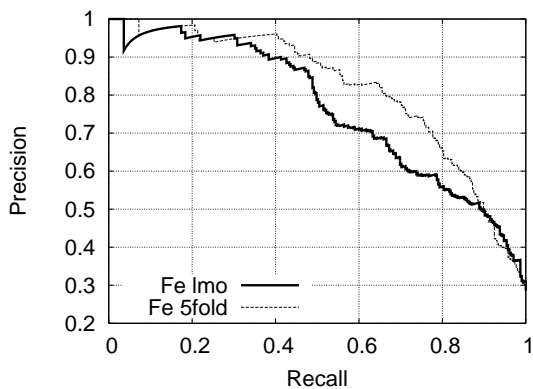
(b)



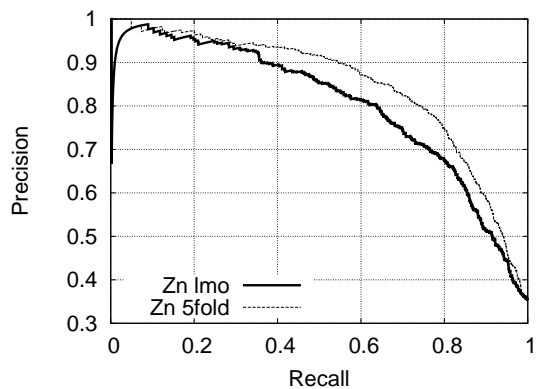
(c)



(d)



(e)



(f)

Figure 5: Leave-metal-out recall precision curves; each curve is computed by training on the binding sites of all metals except the target one, and testing on the target metal binding sites. Figure (a) reports curves for different metals, while the other figures (b-f) compare results for each metal with those obtained by training also on MBS containing that metal (the standard 5-fold cross validation procedure).

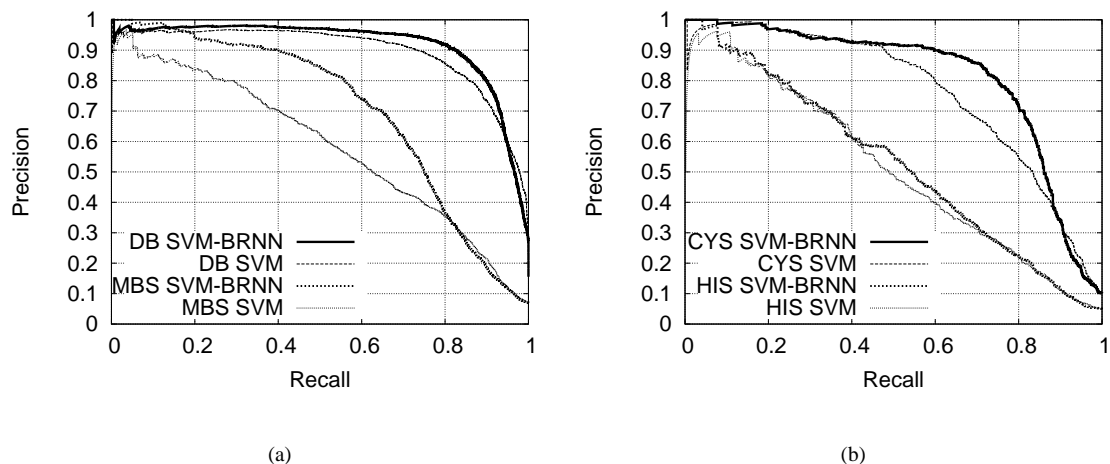


Figure 6: Recall precision curves for the SVM-BRNN architecture compared to the ones for the SVM stage alone. (a) DB and MBS predictions. (b) MBS predictions for CYS and HIS residues separately.

for CYS ligands (Fig. 6(b)). The behavior on DB predictions can be partially explained by observing that all CYS residues within a given protein tend to have a common behavior with respect to disulfide bonding state, due to the folding environment^{51,52} (in the dataset used for the experiments, we had 526 chains with all CYS disulfide bound, 1,521 with none, and only 161 chains with both disulfide bonded and non disulfide bonded CYS). Protein-wise information are in fact already provided to the SVM stage in the form of global descriptors (see the SVM section). Such global descriptors are actually quite important features for the SVM predictor, as removing them results in a significant loss in performance (0.918 ± 0.004 AUC). The importance of global descriptors for disulfide bonding state prediction had already been noted in previous works⁵³. Regarding BRNN features, conservation weight and sequence separation information provided only slight improvements to the network, while secondary structure predictions were more effective in driving up performance of the refinement stage. Table VI reports confusion matrices for different stages of the learning architecture. Results show that the main advantage of the BRNN stage over the local SVM is on the amount of chains with all residues correctly predicted (Q_p). The FSA post-processing stage further improves such performance measure, while leaving accuracy almost unchanged. The main effect of the FSA is that of detecting some free cysteines wrongly predicted as disulfide bonded by the SVM-BRNN or SVM stages, which implies an increase in precision for the DB class (see Table VI). In the binary classification task of discriminating between DB and non-DB CYS, the contribution of the FSA is even

Table VI: Confusion matrices, recall precision for class and overall accuracies for different stages of the learning architecture. Confusion matrices have true class on rows, predicted class on columns. Q_r is the fraction of correctly predicted residues, Q_p the fraction of chains for which all residues are correctly predicted.

Predictor	SVM-BRNN-FSA					SVM-BRNN				
Class	Pre (%)	Rec (%)	Free	DB	MBS	Pre (%)	Rec (%)	Free	DB	MBS
Free	95	96	17063	343	278	95	96	16991	424	269
DB	86	87	384	3090	78	84	87	382	3097	73
MBS	73	61	488	148	975	74	60	484	155	972
Q_r (%)	92					92				
Q_p (%)	77					73				

Predictor	SVM-FSA					SVM				
Class	Pre (%)	Rec (%)	Free	DB	MBS	Pre (%)	Rec (%)	Free	DB	MBS
Free	96	92	16252	342	1090	96	91	16165	438	1081
DB	84	84	445	2997	110	82	85	434	3003	115
MBS	47	66	317	223	1071	47	66	316	232	1063
Q_r (%)	89					89				
Q_p (%)	62					58				

stronger, bringing the performance of the full architecture to $Q_p = 85\%$ with respect to a value of 78% obtained by the SVM-BRNN.

4 CONCLUSIONS

We presented a novel method based on a two-stage machine learning approach for the identification of histidines and cysteines participating in protein metal binding sites or (for cysteines only) in disulfide bridges. In the first stage, a support vector machine predicts the state of individual histidines and cysteines using local information and a protein descriptor, while in the second a bi-directional recurrent neural network models the correlations between histidines and cysteines found within the same protein chain (i.e. potentially belonging to the same metal binding site). For prediction, we considered several transition metals, plus heme and iron/sulfur clusters. Our predictor reached 0.959 overall AUC, with 73% precision and 61% recall on metal binding sites, and 86% precision and 87% recall on disulfide bridges, achieving state-of-the-art results in the subtask of discriminating between disulfide bonded and non-disulfide bonded cysteines. Cysteine

state was generally better predicted than histidine state and differences in performance were also observed among the different metals and complexes, with zinc, heme and iron/sulfur clusters showing the highest accuracy. We believe that our approach can have an impact on protein function and structure prediction, particularly if combined with newly developed high-throughput experimental techniques for identifying metal-binding proteins¹⁵.

Most metals have strong preferences for specific coordination numbers, thus conditioning the number of metal bonded residues which can be found in a given chain. While it would be probably counterproductive to turn such conditioning into hard constraints for the learning algorithm, as we have done for disulfide bonds with the finite state automata, soft constraints on the overall pattern of metal binding residues in a chain could more effectively account for such correlations. We recently²⁰ started to address the problem in the case of zinc binding proteins by jointly classifying pairs of nearby residues, and we expect promising approaches for the task to be developed in the recent field of research on kernel methods for structured output predictions^{54, 55}.

Acknowledgments

MP and BR would like to thank Jinfeng Liu (from Columbia) for computer assistance. MP thanks Matteo Dal Peraro (University of Pennsylvania) for important discussions. The work of MP and BR is supported by the grant R01-GM64633-01 from the NIH. The work of AP and PF is supported by EU STREP APriL II (contract no. FP6-508861) and EU NoE BIOPATTERN (contract no. FP6-508803). We would like to thank the anonymous reviewers whose comments contributed to improve the paper substantially. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

1. Tainer, J. A., Roberts, V. A., and Getzoff, E. D. Metal-binding sites in proteins. *Curr Opin Biotechnol* 2(4):582–591, 1991.
2. Thomson, J. A. and Gray, H. B. Bio-inorganic chemistry. *Curr Opin Chem Biol* 2(2):155–158, 1998.
3. Degtyarenko, K. Bioinorganic motifs: towards functional classification of metalloproteins. *Bioinformatics* 16(10):851–864, 2000.

4. Banci, L., Bertini, I., Calderone, V., Cramaro, F., Conte, R. D., Fantoni, A., Mangani, S., Quattrone, A., and Viezzoli, M. S. A prokaryotic superoxide dismutase paralog lacking two Cu ligands: from largely unstructured in solution to ordered in the crystal. *Proc Natl Acad Sci U S A* 102(21):7541–7546, 2005.
5. Akke, M., Drakenberg, T., and Chazin, W. J. Three-dimensional solution structure of Ca(2+)-loaded porcine calbindin D9k determined by nuclear magnetic resonance spectroscopy. *Biochemistry* 31(4):1011–1020, 1992.
6. Greenblatt, H. M., Feinberg, H., Tucker, P. A., and Shoham, G. Carboxypeptidase A: native, zinc-removed and mercury-replaced forms. *Acta Crystallogr D Biol Crystallogr* 54 (Pt 3):289–305, 1998.
7. Sun, H., Li, H., and Sadler, P. J. Transferrin as a metal ion mediator. *Chem. Rev.* 99(9):2817–2842, 1999.
8. Harding, M. M. The architecture of metal coordination groups in proteins. *Acta Crystallogr D Biol Crystallogr* 60(Pt 5):849–859, 2004.
9. Hertweck, M. and Mueller, M. W. Mapping divalent metal ion binding sites in a group II intron by Mn(2+)- and Zn(2+)-induced site-specific RNA cleavage. *Eur J Biochem* 268(17):4610–4620, 2001.
10. Andreini, C., Bertini, I., and Rosato, A. A hint to search for metalloproteins in gene banks. *Bioinformatics* 20(9):1373–1380, 2004.
11. Passerini, A. and Frasconi, P. Learning to discriminate between ligand-bound and disulfide-bound cysteines. *Protein Eng Des Sel* 17(4):367–373, 2004.
12. Lu, Y., Berry, S. M., and Pfister, T. D. Engineering novel metalloproteins: design of metal-binding sites into native protein scaffolds. *Chem Rev* 101(10):3047–3080, 2001.
13. Kaplan, J. and DeGrado, W. F. De novo design of catalytic proteins. *Proc Natl Acad Sci U S A* 101(32):11566–11570, 2004.
14. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T., and Rost, B. Automatic target selection for structural genomics on eukaryotes. *Proteins* 56(2):188–200, 2004.
15. Shi, W., Zhan, C., Ignatov, A., Manjasetty, B. A., Marinkovic, N., Sullivan, M., Huang, R., and Chance, M. R. Metalloproteomics: high-throughput structural and functional annotation of proteins in structural genomics. *Structure (Camb)* 13(10):1473–1486, 2005.

16. Gregory, D. S., Martin, A. C., Cheetham, J. C., and Rees, A. R. The prediction and characterization of metal binding sites in proteins. *Protein Eng* 6(1):29–35, 1993.
17. Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L., and Jones, D. T. Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* 342(1):307–320, 2004.
18. Schymkowitz, J. W. H., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F., and Serrano, L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc Natl Acad Sci U S A* 102(29):10147–10152, 2005.
19. Lin, C. T., Lin, K. L., Yang, C. H., Chung, I. F., Huang, C. D., and Yang, Y. S. Protein metal binding residue prediction based on neural networks. *Int J Neural Syst* 15(1-2):71–84, 2005.
20. Menchetti, S., Passerini, A., Frasconi, P., Andreini, C., and Rosato, A. Improving prediction of zinc binding sites by modeling the linkage between residues close in sequence. In *Proceedings of RE-COMB'06*, 309–320, Venice, Italy, April 2-5, 2006.
21. Johnson, D. C., Dean, D. R., Smith, A. D., and Johnson, M. K. Structure, function, and formation of biological iron-sulfur clusters. *Annu Rev Biochem* 74:247–281, 2005.
22. Paoli, M., Marles-Wright, J., and Smith, A. Structure-function relationships in heme-proteins. *DNA Cell Biol* 21(4):271–280, 2002.
23. Hulo, N., Sigrist, C. J. A., Saux, V. L., Langendijk-Genevaux, P. S., Bordoli, L., Gattiker, A., Castro, E. D., Bucher, P., and Bairoch, A. Recent improvements to the prosite database. *Nucleic Acids Research* 32(Database-Issue):134–137, 2004.
24. Martelli, P. L., Fariselli, P., and Casadio, R. Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics* 4(6):1665–1671, 2004.
25. Chen, Y. C., Lin, Y. S., Lin, C. J., and Hwang, J. K. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins* 55(4):1036–1042, 2004.
26. Song, J. N., Wang, M. L., Li, W. J., and Xu, W. B. Prediction of the disulfide-bonding state of cysteines in proteins based on dipeptide composition. *Biochem Biophys Res Commun* 318(1):142–147, 2004.

27. Cheng, J., Saigo, H., and Baldi, P. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins* 62(3):617–629, 2006.
28. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. The protein data bank. *Acta Cryst. D*58:899–907, 2002.
29. Mika, S. and Rost, B. Uniqueprot: creating sequence-unique protein data sets. *Nucleic Acids Res.* 31(13):3789–3791, 2003.
30. Rost, B. Twilight zone of protein sequencealignments. *Protein Engi* 12:85–94, 1999.
31. Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637, 1983.
32. Cortes, C. and Vapnik, V. Support vector networks. *Machine Learning* 20(3):1–25, 1995.
33. Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines*. Cambridge University Press, , 2000.
34. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16(10):906–914, 2000.
35. Jaakkola, T., Diekhans, M., and Haussler, D. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology* 7(1–2):95–114, 2000.
36. Nair, R. and Rost, B. Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol* 348(1):85–100, 2005.
37. Hsu, C. W. and Lin, C. J. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* 13(2):415–425, 2002.
38. Crammer, K. and Singer, Y. On the learnability and design of output codes for multiclass problems. *Machine Learning* 47(2–3):201–233, 2002.

39. Fiser, A. and Simon, I. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics* 16(3):251–256, 2000.
40. Getoor, L., Friedman, N., Koller, D., and Taskar, B. Learning probabilistic models of relational structure. In *Proceedings of ICML'01*, 170–177, Williamstown, MA, USA, June 28 - July 1, 2001.
41. Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15(11):937–946, 1999.
42. Vullo, A. and Frasconi, P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics* 20(5):653–659, 2004.
43. Bodén, M. and Hawkins, J. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics* 21(10):2279–2286, 2005.
44. Ceroni, A., Passerini, A., Vullo, A., and Frasconi, P. Disulfind: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res*, 2006. to appear.
45. Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402, 1997.
46. Ceroni, A., Frasconi, P., Passerini, A., and Vullo, A. A combination of support vector machines and bidirectional recurrent neural networks for protein secondary structure prediction. In *Proceedings of AI*IA'03*, volume 2829, 142–153, Pisa, Italy, Sep 23-26, 2003.
47. Bradley, A. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159, 1997.
48. Hand, D. J. and Till, R. J. A simple generalization of the area under the ROC curve to multiple class classification problems. *Machine Learning* 45(2):171–186, 2001.
49. Gattiker, A., Gasteiger, E., and Bairoch, A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics* 1:107–108, 2002.
50. Goto, J. J., Zhu, H., Sanchez, R. J., Nersissian, A., Gralla, E. B., Valentine, J. S., and Cabelli, D. E. Loss of in vitro metal ion binding specificity in mutant copper-zinc superoxide dismutases associated with familial amyotrophic lateral sclerosis. *J Biol Chem* 275(2):1007–1014, 2000.

51. Fomenko, D. and Gladyshev, V. Genomics perspective on disulfide bond formation. *Antioxid. Redox Signal.* 5(4):397–402, 2003.
52. Rietsch, A. and Beckwith, J. The genetics of disulfide bond metabolism. *Annu Rev Genet.* 32:163–184, 1998.
53. Mucchielli-Giorgi, M., Hazout, S., and Tuffèry, P. Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins* 46:243–249, 2002.
54. Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6:1453–1484, 2005.
55. Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. Learning structured prediction models: a large margin approach. In *Proceedings of ICML'05*, 896–903 , Bonn, Germany, August 7-11, 2005, 2005 .