Improving Prediction of Zinc Binding Sites by Modeling the Linkage between Residues Close in Sequence

Sauro Menchetti[1], Andrea Passerini[1], Paolo Frasconi[1], Claudia Andreini[2], and Antonio Rosato[2]

[1] Machine Learning and Neural Networks Group
Dipartimento di Sistemi e Informatica
Università degli Studi di Firenze, Italy
{menchett, passerini, p-f}@dsi.unifi.it
Web: http://www.dsi.unifi.it/neural/
[2] Magnetic Resonance Center (CERM)
Dipartimento di Chimica
Università degli Studi di Firenze, Italy
{andreini, rosato}@cerm.unifi.it
Web: http://www.cerm.unifi.it/

Abstract. We describe and empirically evaluate machine learning methods for the prediction of zinc binding sites from protein sequences. We start by observing that a data set consisting of single residues as examples is affected by autocorrelation and we propose an ad-hoc remedy in which sequentially close pairs of candidate residues are classified as being jointly involved in the coordination of a zinc ion. We develop a kernel for this particular type of data that can handle variable length gaps between candidate coordinating residues. Our empirical evaluation on a data set of non redundant protein chains shows that explicit modeling the correlation between residues close in sequence allows us to gain a significant improvement in the prediction performance.

1 Introduction

Automatic discovery of structural and functional sites from protein sequences can help towards understanding of protein folding and completing functional annotations of genomes. Machine learning approaches have been applied to several prediction tasks of this kind including the prediction of phosphorylation sites [1], signal peptides [2, 3], bonding state of cysteines [4, 5] and disulfide bridges [6, 7]. Here we are interested in the prediction of metal binding sites from sequence information alone, a problem that has received relatively little attention so far. Proteins that must bind metal ions for their function (metalloproteins) constitute a significant share of the proteome of any organism. A metal ion (or metal-containing cofactor) may be needed because it is involved in the catalytic mechanism and/or because it stabilizes/determines the protein tertiary or quaternary structure. The genomic scale study of metalloproteins could significantly

benefit from machine learning methods applied to prediction of metal binding sites. In fact, the problem of whether a protein needs a metal ion for its function is a major challenge, even from the experimental point of view. Expression and purification of a protein may not solve this problem as a metalloprotein can be prepared in the demetallated form and a non-metalloprotein can be prepared as associated to a spurious metal ion. In this paper, we focus on an important class of structural and functional sites that involves the binding with zinc ions. Zinc is essential for Life and is the second most abundant transition metal ion in living organisms after iron. In contrast to other transition metal ions, such as copper and iron, zinc(II) does not undergo redox reactions thanks to its filled D-shell. In Nature, it has essentially two possible roles: catalytic or structural, but can also participate in signalling events in quite specific cellular processes. A major role of zinc in humans is in the stabilization of the structure of a huge number of transcription factors, with a profound impact on the regulation of gene expression. Zinc ions can be coordinated by a subset of amino acids (see Table 2) and binding sites are locally constrained by the side chain geometry. For this reason, several sites can be identified with high precision just mining regular expression patterns along the protein sequence. The method presented in [8] mines patterns from metalloproteins having known structure to search gene banks for new metalloproteins. Regular expression patterns are often very specific but may give a low coverage (many false negatives). In addition, the amino acid conservation near the site is a potentially useful source of information that is difficult to take into account by using simple pattern matching approaches. Results in [9] corroborate these observations showing that a support vector machine (SVM) predictor based on multiple alignments significantly outperforms a predictor based on PROSITE patterns in discriminating between cysteines bound to prosthetic groups and cysteines involved in disulfide bridges. The method used in [9] is conceptually very similar to the traditional 1D prediction approach originally developed for secondary structure prediction [10], where each example consists of a window of multiple alignment profiles centered around the target residue.

Although effective, the above approaches are less than perfect and their predictive performance can be further improved. In this paper we identify a specific problem in their formulation and propose an ad-hoc solution. Most supervised learning algorithms (including SVM) build upon the assumption that examples are sampled *independently*. Unfortunately, this assumption can be badly violated when formulating prediction of metal binding sites as a traditional 1D prediction problem. The autocorrelation between the metal bonding state is strong in this domain because of the linkage between residues that coordinate the same ion. The linkage relation is not observed on future data but we show in Section 2.3 that a strong autocorrelation is also induced by simply modeling the close-in-sequence relation. This is not surprising since most binding sites contain at least two coordinating residues with short sequence separation. Autocorrelation problems have been recently identified in the context of relational learning [11] and *collective classification* solutions have been proposed based on probabilistic

learners [12, 13]. Similar solutions do not yet exist for extending in the same direction other statistical learning algorithms such as SVM. Our solution is based on a reformulation of the learning problem where examples formed by pairs of sequentially close residues are considered. We test our method on a representative non redundant set of zinc proteins in order to assess the generalization power of the method on new chains. Our results show a significant improvement over the traditional 1D prediction approach.

2   Data Set Description and Statistics

2.1   Data preparation

We generated a data set of high quality annotated sequences extracted from the Protein Data Bank (PDB). A set of 305 unique zinc binding proteins was selected among all the structures deposited in the PDB at June 2005 and containing at least one zinc ion in the coordinate file. Metal bindings were detected using a threshold of 3Å and excluding carbon atoms and atoms in the backbone. In order to provide negative examples of non zinc binding proteins, an additional set was generated by running UniqueProt [14] with zero HSSP distance on PDB entries that are not metalloproteins. We thus obtained a second data set of 2,369 chains. Zinc binding proteins whose structure was solved in the apo (i.e. without metal) form, were removed from the ensemble of non-metalloproteins.

2.2   A Taxonomy of Zinc Binding Sites and Sequences

Zinc binding sites of zinc metalloenzymes are traditionally divided into two main groups [15]: catalytic (if the ions bind a molecule directly involved in a reaction) and structural (stabilizing the folding of the protein but not involved in any reaction). In addition, zinc may influence quaternary structure; we consider these cases as belonging to a third site type (interface site), which also lacks a catalytic role. Site types can be heuristically correlated to the number of coordinating residues in the same chain. The distribution of site types obtained in this way is reported in Table 1.

Table 2 reports the observed binding frequencies grouped by amino acid type and site type. As expected, cysteines, histidines, aspartic acid and glutamic acid are the only residues that bind zinc with a high enough frequency. It is interesting to note that such residues show different binding attitudes with respect to the site type. While cysteines are mainly involved in structural sites and histidines participate to both Zn4 and Zn3 sites with similar frequency, aspartic and glutamic acids are much more common in catalytic sites than in any other site type.

2.3   Bonding State Autocorrelation

Jensen & Neville [11] define relational autocorrelation as a measure of linkage between examples in a data set due to the presence of binary relations that link

Table 1. Top: Distribution of site types (according to the number of coordinating residues in the same chain) in the 305 zinc-protein data set. The second column is the number of sites for each site type; the third column is the number of chains having at least one site of the type specified in the row. Bottom: Number of chains containing multiple site types. The second row gives the number of chains that contain at least one site for each of the types belonging to the set specified in the first row.

Number of Coordinating Residues	Site Number	Chain Number
1 (Zn1)	37	20
2 (interface - Zn2)	65	53
3 (catalytic - Zn3)	123	106
4 (structural - Zn4)	239	175
Any	464	305

Site types	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,3,4\}$	$\{2,3,4\}$	$\{1,2,3,4\}$
# Chains	14	9	3	21	4	8	7	1	0	2	0

Table 2. Statistics over the 305 zinc proteins (464 binding sites) divided by amino acid and site type. N_a is the amino acid occurrence number in corresponding site type; f_a is the observed percentage of each amino acid in a given site type; f_s is the observed percentage of each site type for a given amino acid. All is the total number of times a given amino acid binds zinc in general.

Site type	Zn4			Zn3			Zn2			Zn1			All
Amino acid	N_a	f_a	f_s	N_a	f_a	f_s	N_a	f_a	f_s	N_a	f_a	f_s	N_a
C	663	69.3	91.8	45	12.2	6.2	10	7.7	1.4	4	10.8	0.6	722
H	220	23.0	45.7	194	52.6	40.3	59	45.4	12.3	8	21.6	1.7	481
D	48	5.0	27.6	83	22.5	47.7	30	23.1	17.2	13	35.1	7.5	174
E	18	1.9	17.5	46	12.5	44.7	28	21.5	27.2	11	29.7	10.7	103
N	5	0.5	83.3	0	0.0	0.0	1	0.8	16.7	0	0.0	0.0	6
Q	2	0.2	33.3	1	0.3	16.7	2	1.5	33.3	1	2.7	16.7	6
Total	956	100	-	369	100	-	130	100	-	37	100	-	1492

examples to other objects in the domain (e.g. in a domain where movies are the examples, linkage might be due to the fact that two movies were made by the same studio). Here we expect the bonding state of candidate residues be affected by autocorrelation because of the presence of at least two relations causing linkage: `coordinates(r,z)`, linking a residue `r` to a zinc ion `z`, and `member(r,c)`, linking a residue `r` to a protein chain `c`. Unfortunately the first kind of linkage cannot be directly exploited by a classifier as the relation `coordinates` is hidden on new data. However, we may hope to capture some information about this relation by looking at the sequence separation between two candidate residues. In particular, there should be some positive correlation between the bonding state of pairs of residues within the same chain, and it should also depend on the sequence separation between them.

Correlation was empirically measured on the data set described in Section 2. In Figure 1(a) the prior probability of zinc binding for a residue is compared to the same probability conditioned on the presence of another zinc binding residue within a certain separation, for different values of the separation threshold. Figure 1(b) reports the correlation coefficient between the bonding state of pairs of residues, again varying the separation threshold between them. Both curves show a very similar behavior, with the highest peak for a distance of less then three residues, and a small one for a distance of around twenty residues. It can be noted that correlation tends to a non zero residual asymptotic value as distance grows. This effect is due to the relation `member`, by which two residues are linked by the fact of belonging to the same chain.
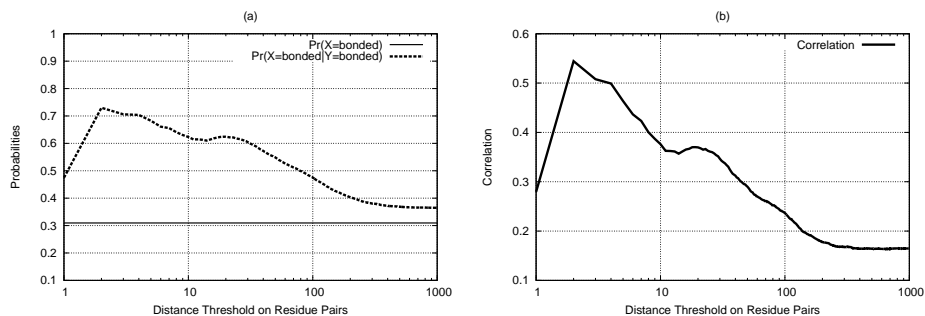
Fig. 1. (a) Probabilities of zinc binding for a given residue: prior and conditioned on the presence of another zinc binding residue within a certain separation. (b) Correlation between the targets of pairs of residues within a given distance.

Patterns of Binding Sites Metal binding sites can be described by patterns characterized by the type of residues coordinating the same ion and their sequence separation. Table 3 reports the most commonly occurring zinc binding patterns together to their frequencies within our data set. Many of these sites, especially the structural ones, contain pairs of coordinating residues whose sequence separation is less than seven residues. In the following, a pattern formed by a single pair of nearby coordinating residues is called a *semi-pattern*. Most structural sites consist of two semi-patterns whose distance ranges between 8 and 29. Catalytic sites typically contain a semi-pattern and a single residue. Finally, interface sites are observed as a single semi-pattern in each chain. Table 4 shows the fraction of sites and zinc proteins containing at least once the semi-pattern [CHDE] x(0-7) [CHDE]. These observations suggest a partial solution to the relational auto-correlation problem based on binary classification of semi-patterns to predict binding sites.

Table 3. Binding site patterns ordered by frequency of occurrence in the 464 sites. Square brackets denote alternative binding residues, x(·) denotes a sequence of residues of an arbitrary length, x($n{-}m$) denotes a sequence of between n and m residues, x($> n$) denotes a sequence of more than n residues. The type column highlights some common binding site patterns: S refers to x(0-7), L refers to x(> 7).

Binding Site Patterns	N	Type
[CHDE] x(·) [CHDE] x(·) [CHDE] x(·) [CHDE]	232	
[CH] x(·) [CH] x(·) [CH] x(·) [CH]	196	
[CHDE] x(0-7) [CHDE] x(·) [CHDE] x(0-7) [CHDE]	161	
[CHDE] x(0-7) [CHDE] x(> 7) [CHDE] x(0-7) [CHDE]	141	SLS
[CHDE] x(·) [CHDE] x(·) [CHDE]	122	
[C] x(·) [C] x(·) [C] x(·) [C]	85	
[CHDE] x(·) [CHDE]	62	
[CHDE] x(0-7) [CHDE] x(> 7) [CHDE]	55	SL
[CH] x(·) [CH] x(·) [CH]	37	
[CHDE] x(> 7) [CHDE] x(0-7) [CHDE]	24	LS
[CH] x(·) [CH]	21	
[CHDE] x(0-7) [CHDE] x(> 7) [CHDE] x(> 7) [CHDE]	17	SLL
[CHDE] x(> 7) [CHDE] x(0-7) [CHDE] x(0-7) [CHDE]	16	LSS
[DE] x(·) [DE]	15	
[DE] x(·) [DE] x(·) [DE]	10	
[CHDE] x(> 7) [CHDE] x(> 7) [CHDE] x(0-7) [CHDE]	10	LLS
[CHDE] x(0-7) [CHDE] x(0-7) [CHDE] x(> 7) [CHDE]	8	SSL
[DE] x(·) [DE] x(·) [DE] x(·) [DE]	1	

3 Methods

3.1 Standard Window Based Local Predictor

Many applications of machine learning to 1D prediction tasks use a simple vector representation obtained by forming a window of flanking residues centered around the site of interest. Following the seminal work of Rost & Sander [10], evolutionary information is incorporated in these representations by computing multiple alignment profiles. In this approach, each example is represented as a vector of size $d = (2k + 1)p$, where k is the size of the window and p the size of the position specific descriptor.

We enriched multiple alignment profiles by two indicators of profile quality, namely the entropy and the relative weight of gapless real matches to pseudo-counts. An additional flag was included to mark positions ranging out of the sequence limits, resulting in an all-zero profile. We thus obtained a position specific descriptor of size $p = 23$. A baseline classifier was constructed using this representation in conjunction with an SVM classifier trained to predict the zinc bonding state of individual residues (cysteine, histidine, aspartic acid and glutamic acid).

Table 4. Site and chain coverage for the [CHDE] x(0-7) [CHDE] semi-pattern. N is absolute, while f is the percentage over the total number of chains/sites of that type.

	Chain Coverage		Site Coverage	
Site Type	N	f	N	f
All	261	85.5	338	72.8
Zn4	168	96.0	227	94.9
Zn3	85	80.1	86	69.9
Zn2	35	66.0	25	38.4
Zn1	13	65.0	0	0.0

Support Vector Machines Support vector machines [16] are a well established machine learning algorithm capable of effectively handling extremely large and sparse feature spaces. Given a training set $D_m = \{(x_i, y_i)\}_{i=1}^m$, where $y_i \in \{-1, 1\}$ is the class label of example x_i, a new instance x is classified as

$$f(x) = \sum_{i=1}^m \alpha_i y_i K(x, x_i) \tag{1}$$

where the sign of $f(x)$ gives the predicted class, and the actual value is a measure of the confidence of such prediction. K is a real valued positive semidefinite kernel function measuring the similarity between pairs of examples, and the weights α_i are learned by a convex optimization function trading off between training errors and complexity of the learned hypothesis. Details on kernel machines can be found in several textbooks [17, 18]. We employed the dot product between example vectors as a baseline linear kernel, to be combined with more complex kernels as described in the experimental section.

3.2 Semi–Pattern Based Predictor

A standard window based local predictor such as the one described in the previous section does not explicitly model the correlation analyzed in Section 2.3, missing a strong potential source of information. Thus, we developed an ad-hoc semi-pattern predictor for pairs of residues in nearby positions within the sequence. A candidate semi-pattern is a pair of residues (cysteine, histidine, aspartic acid or glutamic acid) separated by a gap of δ residues, with δ ranging from zero to seven. The task is to predict whether the semi-pattern is part of a zinc binding site. Each example is represented by a window of local descriptors (based on multiple alignment profiles) centered around the semi-pattern, including the gap between the candidate residues. A semi-pattern containing a gap of length δ is thus encoded into a vector of size $d = (2k + 2 + \delta)p$, where k is the window size and p is the size of the position specific descriptor as described in Section 3.1. In order to address this task, the predictor must be able to compare pairs of semi-patterns having gaps of different lengths. We thus developed an ad-hoc *semi-pattern kernel* in the following way. Given two vectors x and z, of

size d_x and d_z, representing semi-patterns with gap length δ_x and δ_z respectively,

$$
\begin{aligned}
K_{semi-pattern}(x, z) = {}& \langle x[1:w], z[1:w] \rangle \\
& + \langle x[d_x - w : d_x], y[d_z - w : d_z] \rangle \\
& + K_{gap}(x[w+1 : \delta_x p + w], z[w+1 : \delta_z p + w]) \quad (2)
\end{aligned}
$$

where $v[i:j]$ is the sub-vector of v that extends from i to j, and $w = (k+1)p$. The first two contributions compute the dot products between the left and right windows around the semi-patterns, included the two candidate residues, whose sizes do not vary regardless of the gap lengths. K_{gap} is the kernel between the gaps separating the candidate residues, and is computed as:

$$
K_{gap}(u, v) = \begin{cases} K_{\mu gap}(u, v) + \langle u, v \rangle & \text{if } |u| = |v| \\ K_{\mu gap}(u, v) & \text{otherwise} \end{cases}
$$

with

$$
K_{\mu gap}(u, v) = \left\langle \sum_{i=1}^{|u|} u[(i-1)p + 1 : ip], \sum_{i=1}^{|v|} v[(i-1)p + 1 : ip] \right\rangle
$$

$K_{\mu gap}$ computes the dot product between the position specific descriptors within each gap, and if the two gaps have same length, the full dot product between the descriptors in the gaps is added.

3.3 Gating Network

The coverage of the [CHDE] x(0-7) [CHDE] semi-pattern (see Table 4) makes it a good indicator of zinc binding, but a number of binding sites remain uncovered. Moreover, the semi-pattern can match a subsequence which, while not being part of a binding site as a whole, still binds zinc with just one of the two candidate residues. Semi-patterns having a single coordinating residue are considered to be negative examples. This implies that one of the two residues would by construction receive an incorrect label. However, in these cases we can still rely on the local predictor (see Section 3.1) to predict its bonding state. For any given residue we combine the single output from the local predictor, and the (possibly empty) set of outputs from the semi-pattern based predictor, as we get one prediction for each subsequence matching the semi-pattern and containing the residue as one of the two binding candidates. The functional margin calculated by a single SVM (see Eq. (1)) cannot be directly interpreted as a degree of confidence of the predictor, as its magnitude depends on artifacts such as the number of support vectors and the dimension of the feature space. For this reason, in order to combine two predictors, it is preferable to first convert their margins into conditional probabilities using e.g. the sigmoid function approach suggested in [19]:

$$
P(Y = 1|x) = \frac{1}{1 + \exp\left(-Af(x) - B\right)}
$$

where $f(x)$ is the SVM output for example x, and sigmoid slope (A) and offset (B) are parameters to be learned from data. The probability $P(Y_b = 1|x)$ that a single residue binds zinc can now be computed by the following *gating network*:

$$P(Y_b = 1|x) = P(Y_s = 1|x) + (1 - P(Y_s = 1|x))P(Y_l = 1|x) \qquad (3)$$

where $P(Y_l = 1|x)$ is the probability of zinc binding from the local predictor, while $P(Y_s = 1|x)$ is the probability of x being involved in a positive semi-pattern, approximated as the maximum between the probabilities for each semi-pattern x is actually involved in.

4   Experimental Results

We run a series of experiments aimed at comparing the predictive power of the local predictor alone to that of the full gating network. While aspartic and glutamic acids coordinate zinc ions less frequently than cysteines and histidines (see Table 2), they are far more abundant in protein chains. This yields a highly unbalanced data set (the ratios of positive to negative examples were found to be 1:59 and 1:145 for the local and the semi-pattern predictor, respectively). We thus initially focused on cysteines and histidines, bringing the unbalancing down to 1:16 and 1:11 at the residue and semi-pattern level respectively. Moreover, we labelled a [CH] x(0-7) [CH] semi-pattern as positive if both candidate residues bound a zinc ion, even if they were not actually binding the same ion. Preliminary experiments showed this to be a better choice than considering such a case as a negative example, allowing to recover a few positive examples, especially for semi-pattern matches with longer gaps.

Multiple alignment profiles were computed using PSI-Blast [20] on the non-redundant (nr) NCBI protein database. In order to reduce noise in the training data we discarded examples whose profile had a relative weight less than 0.015, indicating that too few sequences had aligned at that position. This also allowed to discard poly-histidine tags which are attached at either the N- or C-terminus of some chains in the PDB, as a result of protein engineering aimed at making protein purification easier. We employed a Gaussian kernel on top of both the linear kernel of the local predictor and the semi-pattern kernel (Eq. (2)). A stratified 4-fold cross validation procedure was used to tune Gaussian width, C regularization parameter, window size and parameters of the sigmoids of the gating network. Due to the strong unbalancing of the data set, accuracy is not a reliable measure of performance. We used the area under the recall-precision curve (AURPC) for both model selection and final evaluation, as it is especially suitable for extremely unbalanced data sets. We also computed the area under the ROC curve (AUC) to further assess the significance of the results.

The best models for the local predictor and the gating network were tested on an additional stratified 5-fold cross validation procedure, and obtained an AURPC equal to 0.554 and 0.611 respectively. Figure 2 reports full recall precision curves, showing that the gating network consistently outperforms the local predictor. While cysteines are far better predicted with respect to histidines,
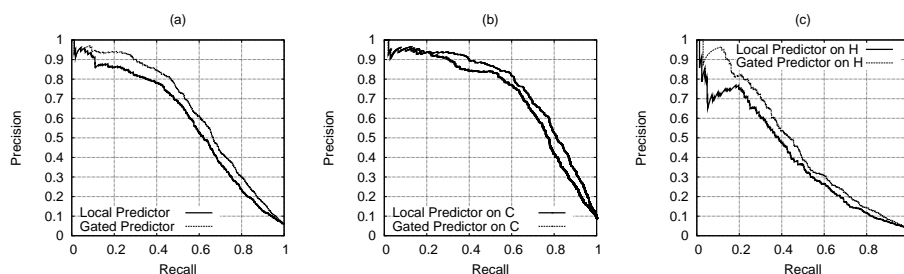
Fig. 2. Residue level recall-precision curves for the best [CH] local and gated predictors. (a) cysteines and histidines together, (b) cysteines only, (c) histidines only.

both predictions are improved by the use of the gating network. AUC values were 0.889 ± 0.006 and 0.911 ± 0.006 for local predictor and gating network respectively, where the method for obtaining the confidence intervals is only available for the AUC computing the standard error of the Wilcoxon-Mann-Whitney statistic, confirming that the gating network attains a significant improvement over the local predictor.
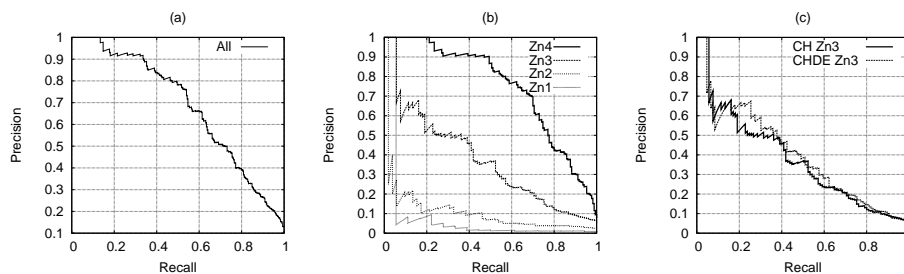
Fig. 3. Protein level recall-precision curves for the best [CH] gated predictor. (a) all proteins together, (b) proteins divided by zinc site type, (c) proteins with Zn3 sites, comparison with the best [CHDE] gated predictor.

Protein level predictions were obtained by choosing the maximum prediction between those of the residues contained in the chain. Figure 3(a) reports the recall precision curve obtained at a protein level for the best gated predictor, while Figure 3(b) shows the results separately for proteins containing different binding site types. As expected, Zn4 sites were the easiest to predict, being the ones showing the strongest regularities and most commonly containing the [CH] x(0-7) [CH] semi-pattern.

Finally, we investigated the viability of training a predictor for all the four amino acids involved in zinc binding, trying to overcome the disproportion issue. On the rationale that binding residues should be well conserved because of their

important functional role, we put a threshold on the residue conservation in the multiple alignment profile in order to consider it as a candidate target. By requiring that $\Pr(D) + \Pr(E) \geq 0.8$, we reduced the unbalancing in the data set for the local predictor to 1:24. At the level of semi-patterns, we realized that such a threshold produced a reasonable unbalancing only for gap lengths between one and three, and thus decided to ignore semi-patterns containing aspartic or glutamic acid with gaps of different lengths. While global performances were almost unchanged, aspartic acid and glutamic acid alone obtained a value of the AURPC of 0.203 and 0.130 respectively. Due to the still high unbalancing, AURPC values for a random predictor are as low as 0.007 for aspartic acid and 0.015 for glutamic acid. AUC values of 0.78 ± 0.03 and 0.70 ± 0.04, respectively (with respect to the 0.5 baseline) confirm that results are significantly better than random. However, results on these two residues are still preliminary and further work has to be done in order to provide a prediction quality comparable to that obtained for cysteines and histidines. It is interesting to note that at the level of protein classification, the only difference that can be noted by using [CHDE] instead of [CH] is a slight improvement in the performances for the Zn3 binding sites, as shown in Figure 3(a). This is perhaps not surprising given that half of [DE] residues binding zinc are contained in Zn3 sites, as reported in Table 2.

5   Conclusions

We have enlightened the autocorrelation problem in the prediction of metal binding sites from sequence information, and presented an improved approach based on semi-pattern classification as a simple linkage modeling strategy. Our results, focused on the prediction of zinc binding proteins, appear to be very promising, especially if we consider that they have been obtained on a non redundant set of chains. Sites mainly coordinated by cysteines and histidines are easier to predict thanks to the availability of a larger number of examples. Linkage modeling allows us to gain a significant improvement in the prediction of the bonding state of these residues. Sites coordinated by aspartic acid and glutamic acid are more difficult to predict because of data sparsity, but our results are significantly better than chance.

The method has been also evaluated on the task of predicting whether a given protein is a zinc protein. Good results were obtained in the case of chains where zinc plays a structural role (Zn4). In the case of chains with catalytic sites (Zn3) the inclusion of D and E targets does allow us to obtain slightly improved predictions. In future work, we plan to test the effectiveness of this method at the level of entire genomes.

References

1. Blom, N., Gammeltoft, S., Brunak, S.: Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol **294** (1999) 1351–1362
2. Nielsen, H., Brunak, S., von Heijne, G.: Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng **12** (1999) 3–9
3. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G.: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng **10** (1997) 1–6
4. Martelli, P.L., Fariselli, P., Casadio, R.: Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. Proteomics **4** (2004) 1665–1671
5. Fiser, A., Simon, I.: Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics **16** (2000) 251–256
6. Fariselli, P., Casadio, R.: Prediction of disulfide connectivity in proteins. Bioinformatics **17** (2001) 957–964
7. Vullo, A., Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics **20** (2004) 653–659
8. Andreini, C., Bertini, I., Rosato, A.: A hint to search for metalloproteins in gene banks. Bioinformatics **20** (2004) 1373–1380
9. Passerini, A., Frasconi, P.: Learning to discriminate between ligand-bound and disulfide-bound cysteines. Protein Eng Des Sel **17** (2004) 367–373
10. Rost, B., Sander, C.: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U.S.A. **90** (1993) 7558–7562
11. Jensen, D., Neville, J.: Linkage and autocorrelation cause feature selection bias in relational learning. In: Proceedings of the Nineteenth International Conference on Machine Learning (ICML2002). (2002)
12. Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann (2002)
13. Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2004)
14. Mika, S., Rost, B.: Uniqueprot: creating sequence-unique protein data sets. Nucleic Acids Res. **31** (2003) 3789–3791
15. Vallee, B.L., Auld, D.S.: Functional zinc-binding motifs in enzymes and DNA-binding proteins. Faraday Discuss (1992) 47–65
16. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning **20** (1995) 1–25
17. Schölkopf, B., Smola, A.: Learning with Kernels. The MIT Press, Cambridge, MA (2002)
18. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge Univ. Press (2004)
19. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D., eds.: Advances in Large Margin Classifiers. MIT Press (2000)
20. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res **25** (1997) 3389–3402