PREDICTING AND CHARACTERIZING METAL-BINDING SITES USING SUPPORT VECTOR MACHINES

L. BOBADILLA AND F. NIÑO

Intelligent Systems Laboratory, Computer Engineering Department,
National University of Colombia
E-mail: {jlbobadillam,lfninov}@unal.edu.co

G. NARASIMHAN*

Bioinformatics Research Group, School of Computer Science,
Florida International University,
Miami FL 33199, USA.
E-mail: giri@cs.fiu.edu

There is a strong need for computational methods to predict and characterize functional binding sites for the functional annotations of protein structures. In this paper, we propose a new method that uses descriptions of the sites based on local properties and support vector machines to predict and obtain key features of metal-binding sites. Previous approaches relied on conserved residues or conserved residue distribution in the protein three-dimensional structure for the predictions. All reported Ca^{2+}, Zn^{2+}, Mg^{2+}, Mn^{2+}, Cu^{2+} binding sites present in a non-redundant PDB (a total of 1144 metal-binding sites in 467 proteins) were used in our experiments. Results from ten-fold cross-validation showed sensitivity and specificity above 95%. Then, using feature selection methods, profiles of critical features were obtained for each metal-binding site. These profiles are consistent with the prior knowledge about metal-binding sites. Furthermore, they provide new insights into the microenvironments of the metal-binding sites.

Keywords: Protein structure, Support Vector Machines, structure-function relationship, metal-binding sites.

1. Introduction

One of the goals of the genome project is to develop tools to compare and interpret genomic information. One fundamental problem is knowing the function of each gene product: Does it bind to another molecule? Is it

*Work partially supported by grant P01 DA15027-01 of the National Institute of Health.

important for regulation of cellular processes? Does it catalyze a chemical reaction? Is it involved in a functional pathway? The importance of answering these questions has led to research efforts aimed at predicting the function of a given protein sequence.

Methods of functional annotation based on sequence data use methods such as sequence alignment and sequence motif detection. However, these methods have their limitations. In the case of sequence alignment, when the sequence similarity goes below 25% to 30% the relationship between the two sequences are hard to detect. Databases such as BLOCKS [8], PROSITE [6], and Prints [2] have sequence patterns that are specific to a given functional family of proteins. While sequence signatures for protein function prediction are very powerful, they still fail to be accurate predictors for function. This is because the factors that determine the functionality of protein active sites are very complex and depend on their three-dimensional structure, and also on the biochemical and biophysical properties of the site. Thus, as sequences diverge, it becomes harder to identify the critical features. And even if two proteins structures are found to be homologous, the relationship between structural and functional similarity is not straightforward [13, 14].

In contrast, a functional annotation method that is based on the tertiary structure and the conserved biochemical and biophysical features of a protein active site should overcome such limitations.

The Protein Data Bank PDB [3] is rapidly growing fueled by the *Protein Structure Initiative* launched by NIGMS [http://www.nigms.nih.gov/psi/]. As of August 31, 2004, PDB contained 26,999 protein structures, which were determined by X-ray crystallography, NMR (Nuclear Magnetic Resonance) spectroscopy and computational methods. For all these proteins one is interested in finding those sites on the protein that are involved in its biochemical and cellular functions. What is needed is a computational method to accurately predict functional sites in protein structures.

In order to devise a predictive method, what is required is a learning procedure that can automatically examine protein molecular structures and can extract useful representations of the key biophysical and biochemical features. Such a general purpose system for producing representations could have medical, pharmaceutical, and industrial applications.

Several approaches to recognize functional sites have been proposed. Nayal *et al.* [12] proposed a method that was designed to locate calcium-binding sites. This study was limited since only 32 proteins (with 64 documented calcium-binding sites) were used. A Bayesian technique was used

by Bagley and Altman [1] and Wei and Altman [19, 20] to predict various functional sites. The Bayesian approach assumes independence of the identified features. However, some features are dependent; for instance, an excess of Aspartic acid residues implies an excess of oxygen atoms as well. Approaches like TESS [17] and FFF [7] rely strongly on the existence of conserved residues. Karlin *et al.* [10] showed that zinc-binding sites can be classified into six groups, showing that the same function can be accomplished by different sets of amino acids in the active site. This can be explained by the fact that during the evolution of a site, the selective pressure is on the ability of the molecule to create an effective milieu for the desired structure or function. Amino acids provide a base set of chemical groups that can contribute certain features to the site. But, several features can be realized in multiple ways. If, instead of viewing binding sites as groups of residues, they are perceived as a chemical milieu that accomplish a function, a better insight of the nature of binding sites can be gained.

Yamashita *et al.* [21] have pointed out the local nature of the origin of ion specificity because the hydrophobicity contrast is determined by groups located within 7Å from the metal ion. Other components that are naturally long range, such as the electrostatic properties of the protein, do not seem to affect the rules for recognition and fail to provide a simple algorithm for the prediction of metal binding sites.

In this paper, we introduce a new method that is inspired by the approach of Yamashita *et al.*. This method focuses on local properties of the environment of metal binding sites to predict (Ca^{2+}), (Zn^{2+}), (Mg^{2+}), (Mn^{2+}) and (Cu^{2+}) binding sites. This new approach is independent of conserved residues and conserved residue geometry, and takes advantage of the large number of protein structures available to construct models using a machine learning approach. This method could be the first step in the construction of a library of models to elucidate the function of the overwhelming amount of protein structures expected in the coming years.

Metal ions are critical to the structure of the protein, their stability, and their function. Approximately one third of all proteins have metal ions (PDB contained around 6000 proteins with documented metal binding site); therefore a tool to predict and characterize metal binding sites will be very significant. Although the MDB [5] (The Metalloprotein Database and Browser) uses geometry and residue conservation to provide key quantitative information on the metal-binding sites in PDB protein structures. It does not, however, provide a predictive approach.

The rest of this paper is organized as follows. In Section 2.1 we define a procedure to describe a protein active site. In Section 2.2 the machine learning approach used (Support Vector Machines) for the prediction of protein active sites is briefly explained. Then the results are shown and compared with previously reported approaches. Finally, some experimental results are discussed, some conclusions are drawn and perspectives for future work are made.

2. Methods

We consider metal-binding sites to be spherical regions of radius 7Å, centered at the crystallographically determined metal ions, as suggested by Yamashita *et al.*. Non-sites, used as negative examples, are 7Å spherical regions of the proteins that are known not to have metal-binding sites.

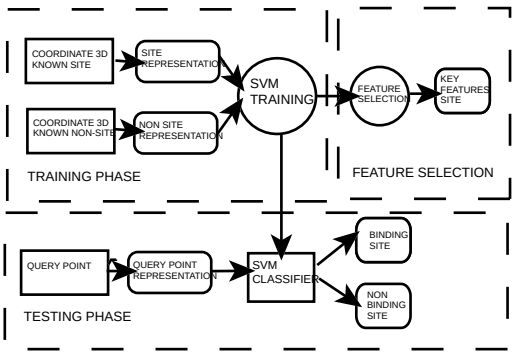

Figure 1. Outline of the algorithm

An outline of the recognition method is shown in Figure 1. The goal is to use a training set of positive and negative examples to obtain a SVM classifier, which can then correctly classify a query region as being a metal-binding site or not. The training set consists of a set of examples of metal-binding sites and non-sites from a non-redundant version of PDB. For each example (both sites and non-sites), a 7Å spherical region is considered and divided into spatial volumes that are concentric shells of thickness 1Å. For each concentric shell, the algorithm computes a measure of each *feature*, thus generating a site description. Then a support vector machine classifier is trained using known sites as positive examples and non-sites as negative examples. Additionally, using feature selection, the most influential fea-

tures of a site description during the training phase are presented as a list of distinguishing features that forms a qualitative model of metal-binding sites. When given a query region in a new structure, it is divided into concentric shells and then a measure of each feature is computed to obtain a description of the site. Then, the representation of the query region is tested using the support vector machine classifier. In the rest of this section, the key elements of our approach, particularly the site description and the support vector machine classifier, are explained in detail.

2.1. *Site Description*

To describe a site, a property-based representation of macromolecular structure was used in this work. Similar representations have been shown to facilitate the identification of key features. For example, Bowie *et al.* [4] used a set of base properties (including secondary structure, degree of solvent accessibility, and polarity) to show that these higher level representations are useful to distinguish proper from improper three-dimensional folds. Similarly, Zvelebil *et al.* [22] have shown that properties can be used to characterize the neighborhood of catalytic residues (properties included residue type, mobility, polarity, and sequence conservation). These properties along with others included in the work by Bagley and Altman (detailed secondary structure classification) were included. Also, in order to capture three-dimensional information of a site, a new property was added: the atom density in the three-dimensional structure as described by Karlin *et al.* [11]. A summary of the properties used in this work is given in Table 1.

Table 1. Properties used in the study.

Categories	Properties
Atom properties	Hydrophobicity, Charge Van der Waals volume, B-factor
Chemical group	Hydroxyl, Amide, Amine Carbonyl, Ring System, Peptide
Residue properties	Residue, Hydrophobic, Charged Polar, Non-Polar,Acidic, Basic
Secondary structure	3-Helix, 4-Helix, 5-Helix Bridge, Alpha, Beta, Coil
Tertiary structure	Atom Packing

Our algorithm obtains the spatial distribution of the features in a spherical region of radius 7Å by scanning all the atoms in each of the concentric shells around the active site. The resulting algorithm was implemented and will be referred to as MILIEU.

2.2. *Support Vector Machines*

As mentioned earlier, MILIEU is based on binary *Support Vector Machines* (SVM) [16] classifier. The training of an SVM constructs a hyperplane separating the positive examples from the negative examples in the space of representations [16]. To minimize the number of misclassifications, SVMs map the examples to a point in higher-dimensional *feature* space. To avoid over-fitting, SVMs choose the Optimal Separating Hyperplane (OSH) to maximize the *margin* of error. SVMs are a sophisticated, robust and efficient tool that can model non-linear relationships with ease [15].

3. Experimental Results

3.1. *Data Sets*

PDB contains many protein structures that are identical (or nearly identical) and has many structures that were determined with low resolution methods. Using all the protein structures in PDB as a training set for our algorithm can introduce bias caused by the redundancy. To avoid this bias and to avoid using low resolution structures for the training, our algorithm was trained with a non-redundant high-resolution database called *CulledPDB* [`http://www.fccc.edu/research/labs/dunbrack/pisces/`]. It is a subset of the PDB database in which the protein structures share less than 30% sequence identity and have a resolution of 1.8Å or better with R-factors at most 0.30. Even within CulledPDB, only those proteins with reported metal-binding sites were used.

The amount of sites used in this work outnumbered any of the previous approaches that predicted metal-binding sites in tertiary structures using local properties. In contrast, Yamashita *et al.* used 23 proteins, Nayal used 32 proteins (with 62 Ca^{2+}-binding sites), while Bagley and Altman used 11 proteins. All the 1144 metal-binding sites present in 467 protein structures from CulledPDB were used in our experiments (see Table 2). Non-sites were obtained by taking random points inside a protein structure which were at least 15Å away from any known binding site. Five non-sites were taken from each protein in order to have more negative examples.

To train and test our MILIEU software, all the site descriptions were transformed into vectors. Each such vector can be written as follows: $\mathbf{x}_i = (x_{i\ prop1\ shell1}, x_{i\ prop1\ shell2}, .., x_{i\ propN\ shell7})$, where each component of $\mathbf{x}_i$ represents one of N properties in one of 7 concentric shells.

Table 2. Sites used in the study

Site	Structures PDB	CulledPDB structures	Sites in CulledPDB
(Ca^{2+})	2303	125	361
(Zn^{2+})	1957	160	349
(Mg^{2+})	1570	122	278
(Mn^{2+})	584	31	100
(Cu^{2+})	314	29	56
TOTAL		467	1144

3.2. *Predicting metal-binding sites*

To make statistically meaningful validation of our experiments, a k-fold cross-validation method was used. Using cross-validation over all the sites makes the approach independent of which sites are used for training and which ones are used for testing. In k-fold cross-validation, the training set is divided into k subsets of equal size. Each subset is tested using the classifier trained on the remaining $k-1$ subsets. Thus, each instance of the whole training set is predicted once, and the cross-validation accuracy is the percentage of data that is correctly classified. MILIEU was tested with ten-fold cross-validation. To assess the performance of the approach, two measures were used: sensitivity, defined as the ability to recognize a metal-binding site, and specificity, defined as the ability to recognize a non-binding site. The formulae for sensitivity (SN) and specificity (SP) are:

$$SN = \frac{TP}{TP+FP}, \quad \text{and} \quad SP = \frac{FN}{TN+FN},$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN the number of false negatives. The results of the cross-validation experiments (see Table 3) show that the error rates are low.

Table 3. Ten-fold cross validation results.

Site	Sensitivity	Specificity
Calcium (Ca^{2+})	96.71	98.93
Zinc (Zn^{2+})	98.24	97.16
Magnesium(Mg^{2+})	96.42	96.19
Manganese(Mn^{2+})	95.0	94
Copper(Cu^{2+})	96.5	98.8

The prediction of Ca^{2+}-binding sites by MILIEU had a sensitivity of 96.71% and a specificity of 98.93%. In contrast, FEATURE [19] achieved a sensitivity of 98% and a specificity of 100%, but with the weaker *leave-one-out cross-validation* and a much smaller set of proteins (training set:

33 sites and 100 non-sites; test set: 33 sites and 30 non-sites). Additionally, binding sites for Zn^{2+}, Mg^{2+}, Mn^{2+}, Cu^{2+} were predicted by MILIEU quite accurately, with all sensitivities above 95.0% and all specificities above 94.0%.

3.3. *Feature selection to obtain key features.*

The purpose of feature selection is twofold. First, it is equivalent to a dimensionality-reduction step, which can improve the efficiency of the method. More importantly, the resulting description of the metal-binding site helps to identify its key features. As mentioned earlier, each site in the training set was mapped to a vector $\mathbf{x}_i = (x_{i1}, ..., x_{il})$, where l represents the dimensionality of the feature space. In our case, we used 67 properties from each of 7 concentric shells, making $l = 67 \times 7 = 469$.

For a given feature vector $\mathbf{x}_i$, the classifier is given by $sgn[b + \mathbf{w^T x_i}]$, where $\mathbf{w} = \sum_{j=1}^{l} \alpha_j x_{ij}$ is the vector of weights $\mathbf{w} = (w_1, ..., w_l)$.

A simple feature selection approach retains features j for which the value of $|w_j|$ exceeds a defined threshold value, t_d. As t_d is increased, fewer features will be selected. To find the essential features for correct classification of metal-binding sites, a separate experiment was performed by taking 70% of the known examples as the training set. Three sets of experiments were performed for each metal: with (a) no feature selection ($t_d = 0$), (b) threshold of 0.1 ($t_d = 0.1$), and (c) threshold of 0.5 ($t_d = 0.5$) The results are shown in Table 4.

Table 4. Feature Selection Results.

Site	t_d	Number Features	Sensitivity	Specificity
(Ca^{2+})	0	469	100.0	99.54
	0.1	112	99.22	99.074
	0.5	12	96.031	95.9
(Zn^{2+})	0	469	99.00	95.15
	0.1	186	99.0	97.12
	0.5	22	91.753	93.08
(Mg^{2+})	0	469	87.30	96.53
	0.1	175	85.71	95.95
	0.5	26	87.932	94.381
(Mn^{2+})	0	469	100.0	93.75
	0.1	109	100.0	95.74
	0.5	9	93.93	91.48
(Cu^{2+})	0	469	100.0	100.0
	0.1	77	100.0	100.0
	0.5	12	100.0	100.0

It can be seen that the reduction of the features for a site description does not greatly affect the sensitivity and specificity of the method, while making it more efficient. This could be explained by the fact that there are many redundant features in the set of features initially chosen, and that some of the features are noisy.

Bagley and Altman [1] proposed more than 100 significant features for Ca^{2+}-binding sites. Our process identified about 12 highly significant features.

In Figures 2-6, the results of the feature selection procedure are organized and presented as a 3D profile describing the key features in each of the concentric shells. Dark squares imply that the feature is present (after applying feature selection with $t_d = 0.5$).

3.4. *Calcium*

The key features of calcium-binding sites are shown in Figure 2, and are consistent with the conclusions of Yamashita *et al.* [21], where they noted that metal sites in proteins are ligated by a shell of hydrophobic atomic groups (indicated by the presence of oxygen).

Our results also showed the importance of several features in the outer shells including polar residues, mobility, and solvent (indicated as RESIDUE_UNKNOWN), confirming the conclusions suggested by Bagley and Altman [1]. Bagley and Altman [1] also showed the significance of the ASP and GLU residues; our results have grouped them under the RESIDUE_CHARGED feature.

3.5. *Zinc*

The key features of zinc-binding sites obtained by our system are shown in Figure 3 and are consistent with the findings of Yamashita *et al.* [21], as shown by the presence of carbon atoms in the outer shells. Karlin *et al.* [9] identified six types of zinc-binding sites, which was not deducible by our system.

3.6. *Magnesium*

The findings for the magnesium-binding sites clearly indicate the importance of carbon and oxygen atoms, which is in accordance with the conclusions of Yamashita *et al.* [21]. The results are summarized in Figure 4.

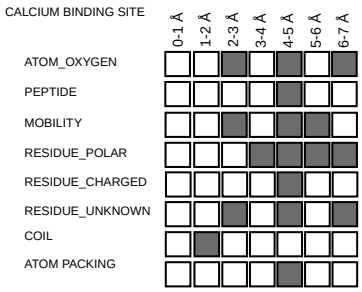

Figure 2. Key findings for calcium-binding sites. Dark squares show features present

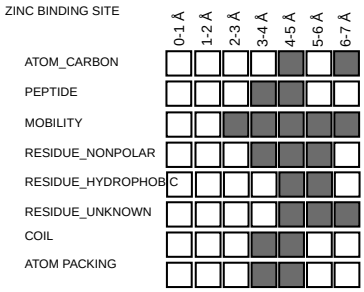

Figure 3. Key findings for zinc-binding sites. Dark squares show features present

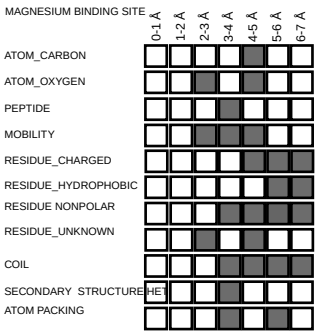

Figure 4. Key findings for magnesium-binding sites. Dark squares show features present

3.7. *Manganese*

The results for manganese-binding sites are similar to that of other metal binding sites and are presented in Figure 5.

3.8. *Copper*

The Histidine and Ring System features, shown in our results in Figure 6, captures the fundamental role of Histidine in the copper-binding sites, as suggested in Karlin [10].

4. Discussion

This paper demonstrates that local properties can be used to train a SVM classifier to obtain an effective system for predicting and characterizing metal-binding sites.

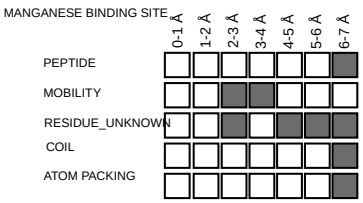

Figure 5. Key findings for manganese-binding sites. Dark squares show features present

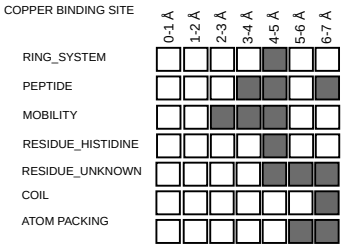

Figure 6. Key findings for copper-binding sites. Dark squares show features present

Our conclusions are consistent with previous findings [21]; we show that all the analyzed metal-binding sites shared a basic set of features. The atom packing feature was crucial in all the cases. Geometry of the binding site is important, but may not be as rigid as suggested by Borkakoti *et al.* [17].

The system proposed in this paper is certainly not restricted to metal-binding sites, and could be applied to other functional sites. Future work includes designing predictors and descriptors for other binding sites (e.g., sodium, potassium, and other organic molecules).

It is not clear whether our system can be used to address the following questions raised by Karlin [10]: Is there substantial divergent or convergent evolution among metal centers? How do similarities and differences in the metal-binding sites reflect evolutionary and functional processes? According to the results obtained, both process could have taken place. A convergent process could have directed the evolution of some metal-binding sites, as suggested by the fact that there is no fixed set of residues or geometry to accomplish the same function; nature in the course of evolution could have found various ways to obtain similar microenvironments suitable for the metal binding. Divergent evolution is also suggested by some conserved residues. Further studies taking into account more specific evolutionary details are required.

5. Conclusions

We conclude that the metal-binding sites should be seen not as a sequence of residues, but as a set of conserved conformational and environmental features, which can be achieved by different configurations, and which can be characterized by a relatively small number of features. We also showed that these features can be learned by using a machine learning approach.

We used the local features to train a SVM classifier. The resulting software MILIEU showed improved performance in detecting binding sites for calcium, zinc, magnesium, manganese, and copper

References

1. S. C. Bagley and R. B. Altman, *Protein Sci.* **4**, 622-635 (1995).
2. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer, *Nucleic Acids Res.* **28**, 263-266 (2000).
3. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, **30**, 245-248 (2002).
4. J. Bowie, R. Luthy and D. Eisenberg, *Science*, **253**, 164-170 (1991).
5. J. M. Castagnetto, S. W. Hennessy, V. A. Roberts, E. D. Getzoff, J. A. Tainer and M. E. Pique, *Nucleic Acids Res*, **30**, 379-382 (2002).
6. L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, A. L. Bairoch, *Nucleic Acids Res.* **30**, 235-238(2002).
7. J. S. Fetrow and J. Skolnick, *J Mol Biol.* **281**, 949-968 (1998).
8. J. G. Henikoff and E. A. Greene, *Nucleic Acids Res.* **28**, 228-230(2000).
9. S. Karlin and Z.-Y. Zhu, *Proc Natl Acad Sci USA* **93**, 8344-8349 (1997).
10. S. Karlin and Z.-Y. Zhu and K. D. Karlin, *Proc Natl Acad Sci USA* **94**, 14225-14230 (1997).
11. S. Karlin and Z-Y. Zhu and F. Baud, *Proc Natl Acad Sci USA* **96**, 12500-12505 (1999).
12. M. Nayal and E. Di Cera, *Proc Natl Acad Sci USA* **91**, 817-821(1994).
13. C. A. Orengo, A. E. Todd and J. M. Thornton, *Curr Opin Struct Biol* **9**, 374-382 (1999).
14. R. B. Russell, P. D. Sasieni and M. J. Sternberg, *J Mol Biol* **282**, 903-918 (1998).
15. C. Scholkopf, J. C. Burges and A. J. Smola, *Advances in Kernel Methods* (MIT Press, Cambridge, MA,1999).
16. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
17. A. C. Wallace, N. Borkakoti and J. M. Thornton, *Protein Sci.* **6**, 2308-2323 (1997).
18. G. Wang and R. L. Dunbrack, *Bioinformatics.*, **19**, 1589-1591 (2003).
19. L. Wei and R. B. Altman, *Pac. Symp. Biocomput.*, 497-508 (1998).
20. L. Wei and R. B Altman, *Journal of Bioinformatics and Computational Biology*, **1**, 119-138 (1998).
21. M. M. Yamashita, L. Wesson, G. Eisenman and D. Eisenberg, *Proc Natl Acad Sci USA.* **87**, 5648-5652 (1990).
22. M. Zvelebil, M. Sternberg, *Protein Eng*, **2**, 127-138 (1988).