RECOGNIZING PROTEIN BINDING SITES USING STATISTICAL DESCRIPTIONS OF THEIR 3D ENVIRONMENTS

Liping Wei & Russ B. Altman
Section on Medical Informatics
SUMC, MSOB X-215, Stanford, CA 94305-5479
{wei, altman}@smi.stanford.edu

We have developed a new method for recognizing sites in three-dimensional protein structures. Our method is based on our previously reported algorithm for creating descriptions of protein microenvironments using physical and chemical properties at multiple levels of detail (including features at the atomic, chemical group, residue, and secondary structural levels). The recognition method takes three inputs: a set of sites that share some structural or functional role, a set of control nonsites that lack this role, and a single query site. The distribution of properties for the query site is compared to the distributions for both sites and nonsites to determine the group with which it is most similar. A log-odds scoring function, based on Bayes' Rule, computes a score that indicates the likelihood that the query region is a site of interest. In this paper, we apply the method to the task of identifying calcium binding sites in proteins. Cross-validation analysis shows that this recognition approach has high sensitivity and specificity. We also describe the results of scanning four calcium binding proteins (with the calcium removed) using a three-dimensional grid of probe points at 2 Å spacing. The probe points that have high scores cluster around the true calcium binding sites, with the highest scoring points at or near the binding sites. The method fails in only one case where a calcium binding site is created by four proteins in the crystal lattice, and thus not recognizable within the crystallographic asymmetric unit. Our results show that property-based descriptions can be used for recognizing protein sites in unannotated structures.

Introduction

Although they share the same important function of binding calcium ions, calcium binding sites—local regions within proteins that bind calcium ions—differ greatly in their three-dimensional structure. Defining the three-dimensional (3D) structural "motif" of calcium binding is important for understanding the functions of new proteins that possess this motif, and for engineering novel binding capabilities into proteins. Investigations into the geometry and chemistry of calcium binding sites have produced refined models of these sites, and some predictive capabilities. Yamashita and colleagues found that metal binding sites (including calcium binding sites) are centered in a shell of hydrophilic ligands, surrounded by a shell of carbon-containing hydrophobic atom groups (Yamashita et al., 1990). Based on this finding, they defined a hydrophobicity contrast function and used it to predict metal binding sites. Nayal and Di Cera found that a high valence value is both necessary and sufficient to predict calcium binding sites (Nayal & Di Cera, 1994), and they proposed a valence function that can predict calcium binding sites with high spatial accuracy. Other groups have characterized the geometry of interaction of metal ions with coordinating residues in the proteins (Gregory et al., 1993).

Because of the exponential increase of the number of known three-dimensional protein structures, we now have an opportunity to develop statistical approaches for characterizing and recognizing calcium binding sites. The rapidly growing structural database enables us statistically to compare known calcium binding sites with known nonsites, automatically extract features that are useful for distinguishing them, and develop statistical methods for recognizing calcium binding sites in new protein structures. We have previously reported the FEATURE system that computes the spatial distribution of properties in protein structures (Bagley & Altman 1995; Bagley et al. 1995). A comprehensive set of physico-chemical properties are used by FEATURE, ranging in detail from atomic-based properties to those based on chemical group, amino acid type and secondary structure type. Given an interesting environment within a protein structure, FEATURE divides the environment into spatial volumes such as concentric shells (the "shell" option) or a 3D grid of small cubes (the "oriented" option). At each spatial volume, and for each property, FEATURE sums up the property values within the volume to get an overall numerical value, measuring the abundance of the property within the volume. Given a set of sites known to have a specific functional or structural role, as well as a background control set of nonsites, FEATURE can compute and statistically compare their spatial distributions of properties, and determine the local environments and properties that distinguish the sites of interest from the control nonsites. Each of these volume/property pairs is called a distinguishing feature.

FEATURE has been used to produce a statistical description of the differences between calcium binding sites and random control nonsites (Bagley & Altman, 1995). It found an abundance of ASP and GLU residues from 2 to 7 Å, a dearth of LEU residues from 4 to 7 Å, an increase in oxygen, amide, and carbonyl groups at 2, 4 and 6 Å, generally low hydrophobicity, high solvent accessibility, high side chain mobility, and a predominance of beta-turns and coils in the surrounding regions. For some protein binding sites, a composite description of the site (recently named "fuzzy recognition templates" (Moodie et al., 1996)), such as produced by FEATURE, may be more useful than deterministic sequence/structure motifs for recognizing sites in new structures.

FEATURE finds for which properties at which corresponding volumes are the known sites and nonsites significantly different. In this sense, FEATURE is essentially a statistical inference system. A further task after characterizing the difference between sites and nonsite, which is probably more important than the inference problem, is to decide whether a new, previously unseen region is likely to be a site. This new task is essentially a supervised classification problem — that of assigning a new individual into one of two possible classes on the basis of a set of features and previously known examples in the classes. In this paper, we present a Bayesian approach to the classification problem. We show that we can recognize

calcium binding sites with high sensitivity and specificity using our statistical description of calcium binding sites and a log-odds scoring function.

Methods

We consider calcium binding sites to be 7 Å spherical regions centered upon crystallographically determined calcium ions. Nonsites are used as explicit background controls, and are 7 Å spherical regions on surface and interior points within proteins that do not bind calcium. We decide whether a query region (the 7 Å sphere around a probe point) is a calcium binding site by comparing the probe region with known calcium binding sites and known nonsites.

The outline of the recognition method is shown in Figure 1. The goal is to compute a score that indicates the likelihood that the query region is a calcium binding site. We start by collecting a set of calcium binding sites and nonsites from the Protein Data Bank (PDB, Bernstein, Koetzle et al. 1977). The sites and nonsites are divided into spatial volumes that are concentric shells of 1 Å thickness. For each site and nonsite, FEATURE computes the count of each property in each of the spatial volumes. It then compares the site counts to nonsite counts, using a Wilcoxon rank-sum test, and reports the distinguishing features. A complete list of distinguishing features for calcium binding sites and the details of the characterization method and its sensitivity analysis have been published elsewhere (Bagley and Altman 1995). The list of distinguishing features form a qualitative model of calcium binding sites. The numerical property counts of sites and nonsites can be used to form a quantitative model. When given a query region in a new structure, we again divide it into concentric shells and apply FEATURE to compute the abundance of the property in the corresponding spatial volumes. The scoring function then compares the values of properties for the query region with the corresponding values of the known calcium sites and nonsites and decides if the evidence supports the hypothesis that the query region is a site.

The scoring function

For each feature, we divide the observed total range of site and nonsite values into k bins ($k = 5$ for these experiments). The value of the property count, v, of the query region must fall in one of the k bins (to account for outliers, values falling out of the range are assigned to the nearest bin.) If v falls in bin i, we can compute the posterior probability that v is drawn from the distribution of the site values (as opposed to the distribution of nonsite values) using Bayes' Rule:

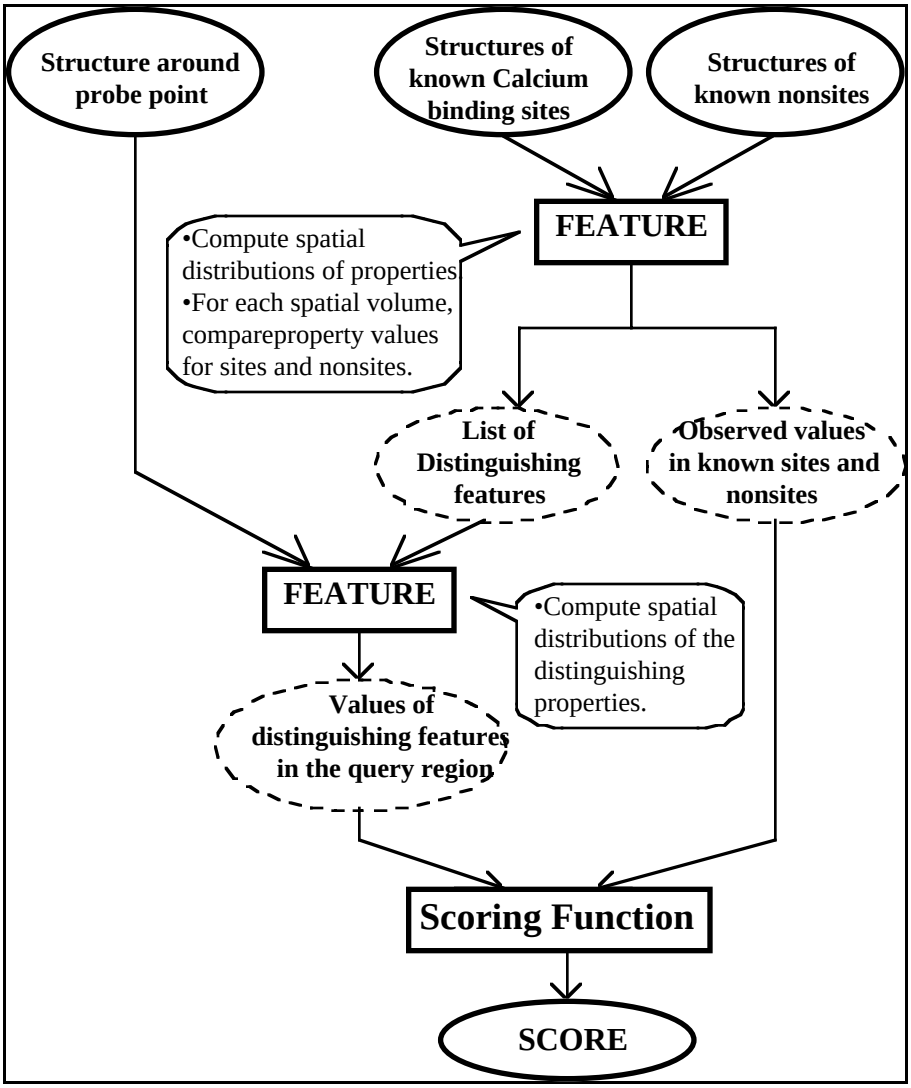


Figure 1: Method outline. Given a probe site in a new structure, and a set of known calcium binding sites and nonsites, the method output a score that indicates the likelihood that the probe site is a calcium binding site.

$$P(site \mid bin\ i) = \frac{P(bin\ i \mid site)P(site)}{P(bin\ i \mid site)P(site) + P(bin\ i \mid nonsite)P(nonsite)},$$

where P(*site*) and P(*nonsite*) are the prior probabilities of being a calcium binding site or nonsite, as supplied by the user. P(*bin i*|*site*) and P(*bin i*|*nonsite*) are the proportions of all feature values that fall in bin *i* for calcium binding sites and nonsites, respectively. They are computed from the property counts for sites and nonsites. To compute the overall likelihood that the probe region is a calcium binding site, the system combines information from all distinguishing features. The overall likelihood score is the sum of the logarithm of the odds ratio:

$$\sum_{\{Distinguishing\ features\}} log \frac{P(site \mid bin\ i)}{P(site)}$$

A logarithm is used so that the ratios can add. This score has the advantage of being simple to interpret: a positive score indicates that there is more evidence that the probe region is a site, whereas a negative score indicates that there is more evidence that it is a nonsite. For any feature we can enumerate all the bins that a new observed value can possibly fall in, and thus can precompute all the log ratios for each feature and for each bin into which a new observed value may fall. The log-ratios are recorded in a score lookup table. When given a new structure, the system computes the spatial distribution of the properties for the new structure, looks up the score table for the individual log-ratios corresponding to the observed feature values, and sums the individual scores to get an overall score. To assist the user in understanding a prediction, the recognition system can also generate an automated prose report of the strongest individual pieces of evidence supporting or refuting the prediction.

To evaluate the accuracy of our recognition algorithm, we used two measures: sensitivity (ability to recognize a calcium binding site) and specificity (ability to recognize a site that does not bind calcium). We used the statistical description previously derived from 16 calcium binding sites and 100 random nonsites (Bagley & Altman, 1995). We chose as the independent test set 33 calcium binding sites and 30 random nonsites not previously used in the analysis. The PDB identification number of the protein structures in the test set are (in parentheses are the numbers of calcium binding sites and random nonsites used in each structure): 1ANX(4, 5), 1AYP(2, 5), 1CGV(2, 5), 1CLM(4, 5), 1OMD(3, 5), 3CLN(4, 5), 1SAC(2, 0), 2SCP(6, 0), 3ICB(2, 0), 3PAL(2, 0), and 5CPV(2, 0). We calculated the sensitivity and specificity of the recognition method on the test set. We then pooled

together the original sites and nonsites and the test set, and performed a leave-one-out cross-validation analysis of accuracy.

We performed sensitivity analysis on parameters that may affect the performance of the method. We tested a range of values for each parameter, and calculated the cross-validation sensitivity and specificity:

1. Prior probability of being a calcium binding site: The system requires an estimate of the prior probability of calcium binding sites in order to employ Bayes' Rule. The estimate may not always be accurate, and the method must be robust to a wide range of choices of prior probability. We performed a sensitivity analysis on prior probability ranging from 10^{-6} to 0.8.
2. Radius of the sites/nonsite regions: How big must the local region around the center of interest be? We tested radii ranging from 1Å to 7Å.
3. Number of bins in the scoring function: We performed sensitivity analysis on the number of bins (value of *k*) ranging from 2 to 10.
4. Redundant properties: Because the FEATURE system was designed to give as comprehensive description of sites as possible, some of the properties are redundant. We tested the possible confounding effect of including redundant properties by removing two sets with obvious redundancy: one of the two amino acid residue classifications, and one of the two secondary structural classifications.

In order to comprehensively test the recognition method on thousands of query regions in a realistic test situation, we scanned four calcium binding proteins that were unrelated and not used in the training (each was left out of training when building the model used for testing). The PDB identifiers of the proteins and (in parentheses) the number of calcium binding sites in them as documented in the PDB files are: 1OMD(3), 3 CLN(4), 3ICB(2) and 3PAL(2). For each test structure, we defined a 2 Å search grid. The recognition function was applied at each grid point to compute the likelihood that a calcium ion can bind at that point. Grid points that scored positive were labeled as potential calcium binding points. The high scoring probe points were visualized graphically, and their locations were compared with those of the actual calcium binding sites.

Results

The sensitivity and specificity of calcium binding site recognition in both the independent test set and in the cross-validation analysis are shown in Table 1. Figure 2 shows the histograms of the recognition scores for the calcium binding sites and nonsites in the two analyses. Table 2 shows the results of sensitivity analysis. The structures and the potential calcium binding points found by our

scanning method are shown in Figure 3. Four kinemage files with the results of calcium site scanning are available at:

http://www-smi.stanford.edu/projects/helix/pubs/wei-psb98/.

	Sensitivity (%)	Specificity (%)
Independent test	91	100
Cross-validation	98	100

Table 1: Accuracy of recognition in independent test and cross-validation analysis.

Parameters		Sensitivity (%)	Specificity (%)
Prior of sites	10^{-6}	98	100
	10^{-4}	98	100
	0.01	**98**	**100**
	0.1	96	99
	0.5	94	99
	0.8	94	99
Radius (Å)	1	100	88
	3	98	99
	5	98	100
	7	**98**	**100**
Number of bins	2	96	100
	5	**98**	**100**
	7	98	100
	10	98	100
Remove	Redundancy	98	100

Table 2: Sensitivity analysis on parameters. The sensitivity and specificity are for cross-validation results. The parameters in bold are the ones actually used in the paper.

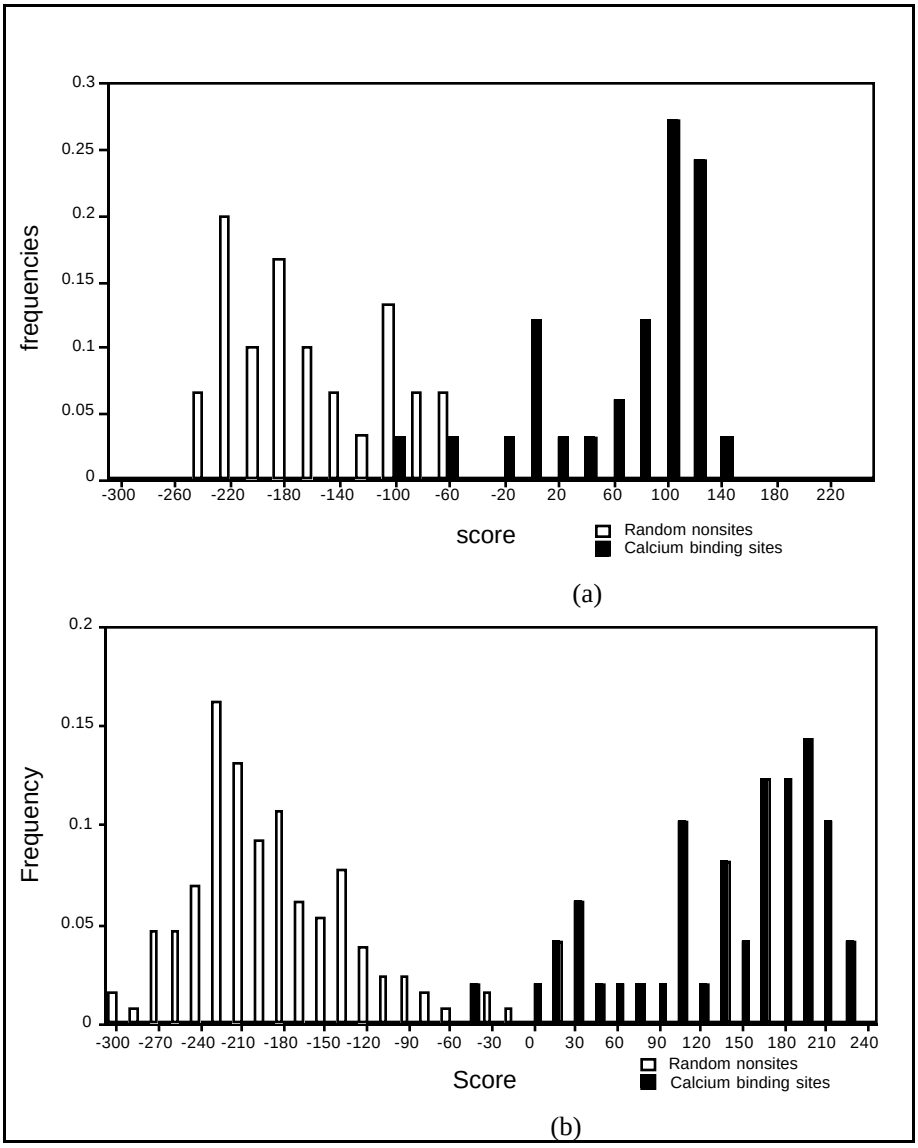
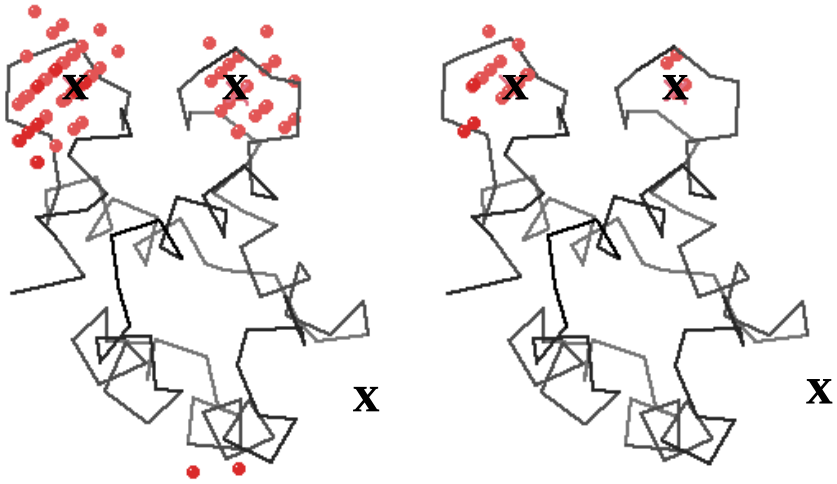


Figure 2: Histogram of recognition scores for the calcium binding sites and random nonsites (a) in the independent test set; (b) in the cross-validation analysis.

1OMD



3CLN

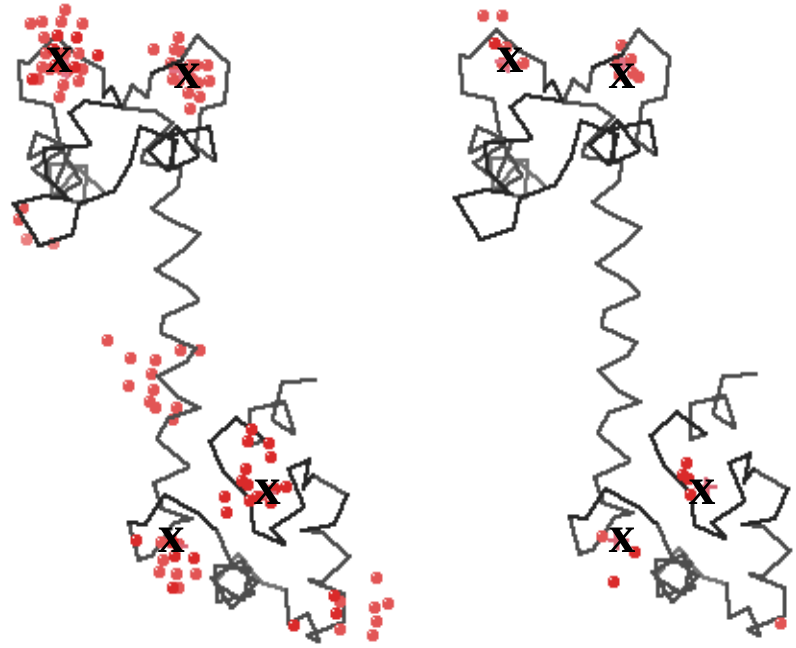


Figure 3 (to be continued)

3ICB

3PAL

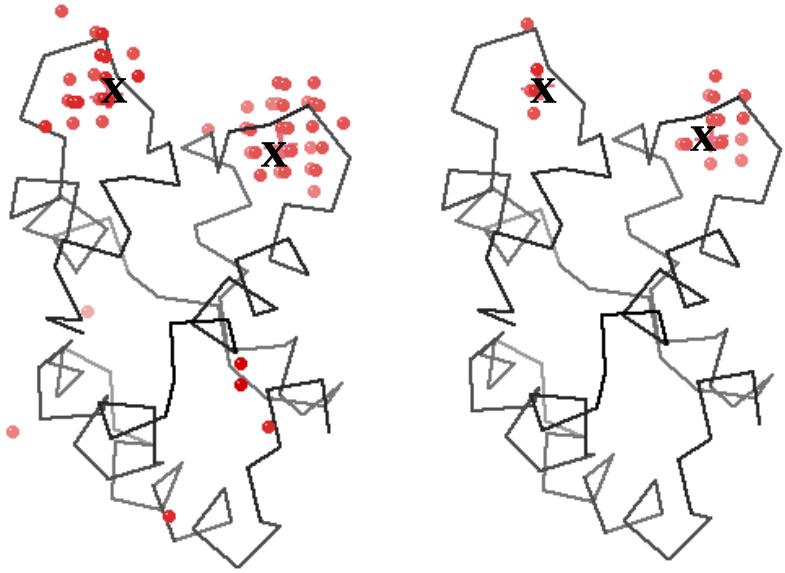


Figure 3 Scanning results. For each protein, the picture on the left show as dots all search points that scored positive, and the picture on the right show as dots the 20 search points that have the top 20 highest scores. The real locations of the calcium ions are marked with "X".

Discussion

The accuracy of our recognition method is high, with sensitivity and specificity above 90%. Furthermore, our sensitivity analysis indicates that for a wide range of prior probabilities, neighborhood radii, and bin granularity the method is robust and accurate.

The performance of the method in scanning test proteins is promising. Each protein scan required evaluation of more than 3000 probe positions, and lead to a small number of positive scores. For all four proteins, the method recognized the calcium binding sites as the centers of the highest scoring regions. In only one case was a calcium binding site completely missed. The missed calcium binding site is CA135 as documented in 1OMD. This calcium is at the interface of four molecules in the crystal packing lattice, and its binding site is created by contributions from more than one of these molecules. Since structural data on only one molecule within the asymmetric unit is available to our method, the binding site was missed. If structures of all four molecules were available, there would be high chance that our method be able to find this particular calcium binding site.

The highest scores in each cluster of positively scored probe points are near the actual calcium sites, and are always within 5 Å distance from the precise locations. The very highest scoring points are about 1 Å away from the precise locations. In general, the search points with low positive scores are far away from any of the real binding sites. Because we are using the documentation in PDB files as the gold standard, those points should be considered false positives. Interestingly, most of these false positives are located within 2Å from multiple glutamate and aspartate side chains—amino acids known to interact with metal ions. For example, the two small positive scoring clusters in that appear on the left side of Figure 3 look very much like they could bind a calcium ion. These false positives may actually represent binding sites for other metal ions (or even missed calcium ions) that are not documented in PDB.

Using the statistical model of calcium binding sites (instead of a deterministic one) gives our method the potential to recognize new calcium binding sites which may have different amino acid composition or 3D atomic arrangement from all known calcium binding sites, but which maintain most of the important biochemical and biophysical features.

Our scoring function assumes that all the property/volume combinations are independent and that their contributions can be summed. This independence assumption is clearly false: if the abundance of ASP and GLU is high, then the abundance of oxygen and carbonyl groups are likely to be high as well. Our sensitivity analysis shows a slight increase in performance when we remove one of our redundant secondary structure classifications (the most obviously correlated features in our system). As the quantity of structural data increases, we may be

able to consider correlation effects between properties in a statistically rigorous manner.

Our results demonstrate that there are conserved features in the spatial distributions of properties across a wide range of proteins that share a common structural feature. In the case of calcium binding, the descriptions gathered from a relatively small set of sixteen calcium sites can be used to successfully recognize sites in unrelated proteins. Our recognition method is competitive with functions that use information about calcium valence and the associated geometry of surrounding atoms. Our approach is general, however, and we are endeavoring to create statistical models for other important protein sites. We are also working to improve the efficiency of our scanning code, to allow for large scale, automated structure annotation.

Acknowledgments

We thank Jeffrey Chang for his contributions to the FEATURE system. RBA is supported by NIH LM-05652, LM-06442, and NSF BIR-9600637. We thank M. Corcoran for retyping an accidentally deleted manuscript.

References

Bagley, S. C. & Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci.* **4**, 622-635.

Bagley, S. C., Wei, L., Cheng, C. & Altman, R. B. (1995). *Third International Conference on Intelligent Systems for Molecular Biology,* Cambridge, England, AAAI Press, Menlo Park, CA, ISMB 3, 12-20.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. J., Brice, M. C., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Gregory, D. S., Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1993). The prediction and characterization of metal binding sites in proteins. *Protein Eng.* **6**(1), 29-35.

Moodie, S., Mitchell, J. & Thornton, J. (1996). Protein Recognition of Adenylate: An Example of a Fuzzy Recognition Template. *J. Mol. Biol.* **263**(3), 486-500.

Nayal, M. & Di Cera, E. (1994). Predicting Ca^{2+} binding sites in proteins. *Proc. Nat. Acad. Sci. (USA)* **91**, 817-821.

Yamashita, M. M., Wesson, L., Eisenman, G. & Eisenberg, D. (1990). Where metal ions bind in proteins. *Proc. Nat. Acad. Sci. (USA***) 87**, 5648-5652.