Analysis of Protein-Ligand Interactions Using Localized Stereochemical Features

Reetal Pai[1], James C. Sacchettini[2], and Thomas R. Ioerger[1]
[1]Department of Computer Science, [2]Department Biochemistry and Biophysics
Texas A&M University, College Station, Texas, USA

Abstract ***Computational analyses of protein structure-function relationships have traditionally been based on sequence homology, fold family analysis and 3D motifs/templates. Previous structure-based approaches characterize and compare active sites based on global shape and electrostatic properties. But, these methodologies are unable to capture similarities between diverse active sites that span multiple fold families despite catalyzing the same reaction (convergent evolution). In this work, we extend previous feature-based analyses of active sites by defining a system of localized geometric and electrostatic descriptors that identify localized patterns of protein-ligand interactions. Singular Value Decomposition is used to identify linear combinations of features with maximum information content which are then used to compute the class conditional probability density distribution of active sites using kernel density estimation. We successfully tested our algorithm on a database that contained examples of adenine, citrate, nicotinamide, phosphate, pyridoxal and ribose binding proteins with over 75% accuracy.***

Keywords: Active Site Analysis, SVD, Localized Stereochemical Features

1 Introduction

The binding affinity between a protein and its cognate ligand is determined by their steric and chemical complementarity [16], [19]. Yet, modelling binding site patterns and using them to predict cognate ligands remains challenging. Amongst proteins that catalyze the same reaction, there exists a large diversity in sequence as well as architecture/fold. Automated function prediction algorithms that rely solely on sequence homology or fold family similarity have been unable to capture the diversity amongst active sites. 3D templates/motifs ([1], [5], [22], [24]) that capture the patterns within the active site in the form of identities and relative orientations of the amino acids lining the active site pocket have also had difficulties addressing this diversity. Various studies, including those of adenine active sites [8], the active sites of the DJ-1 superfamily (which consists of kinases involved in the biosynthesis of thiamine) [28] and the active sites of β-carbonic anhydrase from diverse bacterial species [17], have all shown that overall shape and electrostatic complemetarity between active site and ligand is always maintained despite substantial differences in specific residue identities and placements.

An analysis of diverse active sites binding the same ligand that goes beyond specific residue placements and fold analyses is essential in order to better understand and capture the underlying geometric and chemical interaction patterns. While computationally rigorous *Docking* algorithms ([20], [13] etc.) seek to identify high-affinity protein-ligand complexes by maximizing favorable chemical interactions and minimizes steric conflicts, they are not always capable of identifying the correct substrate since accuracy of force fields and scoring functions used in these algorithms are still under debate.

A more computationally-efficient approach capable of capturing patterns within diverse active sites binding the same ligand was developed based on the extraction of descriptive features from active sites [11], [4]. These features allowed for broader similarities between diverse families of protein active sites to be identified since they did not depend on precise location of residues within the active site. But most of the previous feature-based approaches have used globally computed features that do not capture the spatial variations of chemical and geometric properties within an active site and therefore they cannot be used with great accuracy to distinguish between active site clefts belonging to two different ligands. Additionally, previously used features com-

pletely lack geometric information (other than overall active site volume or depth) adding to the difficulties in their use for functional analyses. *FEATURE* [2], [3] incorporated some geometric information into their analysis of active sites by defining distributions of residue properties in radial shells to capture the differences in protein microenvironments between protein active sites and non-sites. They succesfully used this system to identify common biochemical properties within the serine protease active sites and calcium binding sites.

In this paper, we extend previous feature-based approaches and introduce a new system of fine-grained features (shape and electrostatic descriptors) that describe active sites, allow for substrate recognition and discriminate between active sites that bind different ligands. These features make a compromise between highly-specific 3D templates and general higher-level features. In particular, we capture geometric information through features such as the three eigenvalues of the coordinate variance-covariance matrix (representing the three dimensions of an active site), concavity, curvature and finer-grained features representing a width-profile across a profile axis etc. Additionally, we incorporate spatial distribution of electrostatic information about the site in the form of features based on pairwise distances between different types of groups lining the site (positive, negative, hydrophillic, hydrophobic etc.). These features encode information about local geometry and electrostatic properties necessary for discriminating ligand-binding, without relying on specific contacts at specific coordinates. *Singular Value Decomposition* is used to identify the linear combination of features that best capture the similarity between diverse active site patterns and reduce the dimensionality of the feature vectors. The feature vectors in the reduced dimension SVD space are then used to classify each active site based on the ligand it binds. To this end, kernel density estimation combined with Bayes Theorem is used to find the posterior probability of the various classes for each active site and the native ligand is identified as the class with the highest posterior probability. We show that our methodology is able to identify the underlying binding pattern between diverse active sites binding the same ligand without relying on residue identities/placements, secondary structure information etc when tested on 6 classes of ligands: adenine, citrate, nicotinamide, phosphate, pyridoxal and ribose. Our methodology recognized the correct ligand with over 75% accuracy in each ligand class.

2 Methods

In this section, we will describe the definition of the active site surface, the active site characterization based on a combination of global and local geometric and electrostatic features and the algorithms used to classify active site surfaces based on these features. We will also discuss the methodology used to apply the analysis to apo-proteins and will finally describe the construction of the fragment database.

2.1 Molecular Surface and Active Site Surface Generation

For each of the examples in our database, the protein coordinates were used to compute a molecular dot surface similar to the solvent-accessible surface, defined by Richards [23] and later implemented by Connolly [6], using *Calcsurf*, an in-house program. *Calcsurf* simulates the contacts a water molecule (probe sphere of radius 1.4Å) would make with the protein molecule. Considering the radius of the water molecule and the Van der Waals radii of protein atoms, a grid representing the dot molecular surface is drawn at a distance equal to the sum of these two radii from the protein molecule. The grid points are spaced 1Å apart allowing a fine-grained representation of the solvent-accessible surface. In the case of proteins where the active site is at the interface of multiple chains, the molecular surface was drawn over all the chains that participate in the active site creation thus allowing for the analysis of such active sites. This molecular surface was then used to define the active site surface for a protein as all those atoms of the molecular surface which lay within 3Å of any non-H ligand atom. This active site surface is used in subsequent geometric and electrostatic feature calculations.

$$S = \{a_i \mid \|a_i - b_j\| < 3\mathring{A}; \text{for any } b_j \in Ligand\} \quad (1)$$

2.2 Feature Descriptions

The global shape features used to describe the active site surface, S, are as follows:

- The eigenvalues of the coordinate variance-covariance matrix are used to define the spread of the pocket in three dimensions. The eigenvalues λ_1, λ_2 and λ_3 of the variance-covariance matrix C are calculated using the following equation:

$$|C - \lambda I| = 0 \quad (2)$$

The eigenvector corresponding to the largest eigenvalue is defined as the direction defining the *profile axis*, $\mathbf{v}$ and is used in localized feature computations.

- The local undulations in the surface are measured by computing the average distance between a active site surface atom and its closest n protein atoms. To maintain locality n is chosen to be a relatively small number, in this case, 3. These local undulations are then averaged to yield the surface concavity metric. This metric allows us to distinguish between an active site that is uniformly smooth and one that has many local undulations on its surface.

$$\Gamma(a_i) = \frac{1}{n}\sum_{j=1}^{n} \|a_i - b_j\| \tag{3}$$

$$\Gamma(S) = \frac{\sum_{i=1}^{A} \Gamma(a_i)}{A} \tag{4}$$

In equation 3, b_j is the j^{th} closest protein atom to active site surface atom a_i and $A = |S|$.

- The curvature of a pocket defined as the spread of the pocket around its center of mass ($C_m(S)$) (metric used in [25]).

$$\mathcal{K}(S) = \frac{\mu_p}{\sigma_p} \tag{5}$$

where

$$\mu_p = \frac{\sum_{i=1}^{A} \|a_i - C_m(S)\|}{A} \tag{6}$$

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^{A} \|a_i - C_m(S)\| - \mu)^2}{A-1}} \tag{7}$$

are the mean and standard deviation of the spread of the active site surface atoms around the center of mass of the site and $C_m(S)$ refers to the center of mass of the active site.

- 3D moment invariants which are descriptors of geometric shape that are invariant to rotation and translation [26]. These invariants are calculated as follows:

$$\begin{aligned} J_1 = \;& \mu_{200} + \mu_{020} + \mu_{002} \\ J_2 = \;& \mu_{200}\mu_{020} + \mu_{200}\mu_{002} \\ & + \mu_{020}\mu_{002} - \mu_{110}^2 - \mu_{101}^2 - \mu_{011}^2 \\ J_3 = \;& \mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} \\ & - \mu_{002}\mu_{110}^2 - \mu_{020}\mu_{101}^2 - \mu_{200}\mu_{011}^2 \end{aligned} \tag{8}$$

where

$$\mu_{pqr} = \sum_x \sum_y \sum_z (x-\overline{x})^p (y-\overline{y})^q (z-\overline{z})^r \tag{9}$$

and where $\overline{x}$, $\overline{y}$ and $\overline{z}$ are the coordinates of the center of mass of the active site surface.

Cross sectional features are used for a finer-grained characterization of the active site shape profile along the profile axis. The axis acts as a local frame of reference and places all the examples in canonical positions and the features then capture the spatial variations in shape.

- Cross sections of the pocket at equal spacings (1Å) from the center of mass, $C_m(S)$, and along the profile axis are considered. The distance between any two active site surface atoms in the cross-section is computed and averaged. The cross-section of active site surface S at distance r from the center of mass is defined as the set of vertices:

$$\Omega(S, r) = \{a_i \in S | \overline{a_i p_r} \perp \mathbf{v}\} \tag{10}$$

where $\mathbf{v}$ defines the profile axis and p_r is a point on the profile axis that is r Å away from the center of mass $C_m(s)$; $\|p_r - C_m(S)\| = r$. Now the cross-sectional descriptor of the active site surface at a distance r can be described as the average pairwise distance among vertices in the cross-section:

$$\hat{\Omega}(S, r) = \frac{\sum_{i=1,j=1}^{M_r} \|c_i - c_j\|}{M_r(M_r-1)/2} \tag{11}$$

for $c_i, c_j \in \Omega(S, r)$ where M_r is the total number of active site surface atoms in the cross-section $\Omega(S, r)$.

The electrostatic features capture the spread of charge and hydropathy across the active site surface based on the electrostatic potential across the active site surface. The electrostatic potential at each active site surface atom is based on the partial charges of all the protein atoms. The partial charges used were the same as the ones used by AMBER [7] in their computation of the molecular mechanical force field to compute interaction energies. The potential on an active site surface atom a_i due to a charge q_j placed at a distance d_j from it is given by:

$$V(a_i) = \sum_{j=1}^{N} \frac{q_j}{4\pi\epsilon_0 d_j} \tag{12}$$

where N is the total number of protein atoms. While, in this study we use the Coulomb equation

for potential calculations and do not consider the effects of solvent, this method can be extended using Poisson-Boltzmann solvers such as Delphi [15].

Based on the atom types used in [18], each protein atom was categorized as hydrophobic, hydrophillic or charged. This definition was then extended to define the hydropathy of the active site surface atoms based on the majority classification of its n closest protein atoms. Once again, in order to capture local information, a small number of closest neighbors ($n = 3$), is used. The equation used to categorize the hydropathy (Y) of an active site surface atom a_i is as follows:

$$Y(a_i) = majority(Y(p_j)), j = 1 : n \quad (13)$$

where p_j is the j^{th} closest protein atom to a_i where $Y(p_j) \in \{\text{H, P, C}\}$. The features used to capture the chemical nature of the active site based on the previous definitions of charge and hydrogen bond propensity are as follows:

- The global hydropathy features of the surface S measuring hydrophobicity Y_H, hydrophillicity Y_P and charge Y_C are computed as follows:

$$Y_X(S) = \frac{\sum_{j=1}^{A} y_x(a_j)}{A} \quad (14)$$

$$y_x(a_j) = \begin{array}{ll} 1 & \text{if } Y(a_j) = X \\ 0 & \text{otherwise} \end{array} \quad (15)$$

Pairwise distances between positive, negative and neutral potential points are calculated in order to capture the spatial distribution of potentials while maintaining rotation-invariance.

- Distribution of potentials across the active site surface: These features calculates the percentage of the active site surface occupied by positive, negative and neutral potential points respectively. The electrostatic nature of each active site surface atom is defined as follows:

$$E(a_i) = \begin{array}{ll} \text{P} & \text{if } V(a_i) \geq \phi_1 \\ \text{N} & \text{if } V(a_i) \leq \phi_2 \\ \text{O} & \text{otherwise} \end{array} \quad (16)$$

where ϕ_1 is set to 0.5 and ϕ_2 is set to -0.5 based on empirical observations of the variation of potentials in the example active site pockets. The electrostatic separation features of the active site surface S measuring spread of positive potentials Δ_P, spread of negative potentials Δ_N and the spread of neutral potentials Δ_O (by a given distance, r) are calculated as follows:

$$\Delta_X(S; r) = \frac{\int_S \int_S u_r(a_i, a_j)\delta_X(a_i, a_j)}{Area(S)^2} \quad (17)$$

where

$$u_r(a_i, a_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(r - \|a_i - a_j\|)^2} \quad (18)$$

$$\delta_X(a_i, a_j) = \begin{array}{ll} 1 & \text{if } E(a_i) = E(a_j) = \text{X} \\ 0 & \text{otherwise} \end{array} \quad (19)$$

where X $\in$ {P, N, O} and $\Delta(S; r)$ gives the average electrostatic property match between all pairs of points on S (double integral) separated by a given distance, r (weighted by the Gausssian kernel u).

There are a total of 37 features: 16 geometric features and 21 electrostatic features:

$$\Phi(\mathbf{x}) = \begin{array}{l} \langle \lambda_{1...3}, \Gamma, \mathcal{K}, J_{1...3}, \hat{\Omega}(r_1), \\ Y_H, Y_P, Y_C, \Delta_P(r_2), \Delta_N(r_2), \Delta_O(r_2) \rangle \end{array} \quad (20)$$

where $r_1 = 2...9\mathring{A}$ and $r_2 = 4...9\mathring{A}$.

2.3 Feature Selection and Active Site Classification Using Kernel Density Estimation

Given a test active site whose cognate ligand is to be identified, we first compute the features described in Section 2.2. The posterior probability of each ligand class given the observed test vector is then computed using Kernel Density Estimation. A *Product kernel* of D single-dimensional Gaussian kernels [9] is used to describe the spread of feature vectors in each ligand class in our database. This probability distribution is given by

$$P_{KDE}(\mathbf{x}|c_i) = \frac{\sum_{j=1}^{N} \frac{1}{h_1 \cdot h_2 \cdot h_3 \ldots \cdot h_d} \prod_{d=1}^{D} K(\frac{x - x^j}{h_d})}{N} \quad (21)$$

$$K(\frac{x - x^j}{h_d}) = \frac{1}{\sqrt{2\pi} h_d} e^{-\frac{1}{2}\left(\frac{x - x^j}{h_d}\right)^2} \quad (22)$$

where h_d gives the optimal bandwidth of each of the Gaussian kernels and is determined using $h_{opt} = 0.9AN^{-\frac{1}{5}}$ where N is the number of examples in the class being considered. A is defined as $A = min(\sigma, \frac{IQR}{1.34})$ where IQR is the *Interquartile Range* for that particular dimension and σ is the sample deviation.

Assuming equal prior probabilities for all ligand classes, this probability density function can be used to estimate the likelihood of a class C_i given a test feature vector ($\Phi(\mathbf{x})$) as ($P(C_i|\ \Phi(\mathbf{x}))$. Each feature vector is classified as belonging to the class with the highest log likelihood. Additionally, this probability can also be used as an indicator of confidence in the class prediction based on the feature analysis.

Since all of the features need not have information equally relevant to classification, it is necessary to select the most relevant features before estimating the probability density function. In this study, we use a standard machine learning approach, Singular Value Decomposition [27], to reduce the dimensionality of our feature vectors. SVD projects the feature vectors onto the directions with maximum variability within the data and helps to increase the accuracy of classification of active sites.

Given a $m \times n$ matrix M that contains the feature vectors for the training data (such that m is the number of features and n is the number of training examples), the Singular Value Decomposition of M is given by:

$$M = U\Sigma V^T \tag{23}$$

where U and V are unitary matrices that contain basis vectors describing the principal directions of variation in M. The matrix Σ contains the singular values of M which are weights for each of the directions of variation (right singular vectors in V.) The directions of variation are linear combinations of features from the feature space. The projection of training data onto SVD space helps to improve separation among clusters belonging to different ligands by emphasizing directions of high variation between classes. Additionally, in this space feature vectors that do not contain any information content have very low singular values (close to 0) and can therefore be ignored. In our case, the transformed axes corresponding to the top 6 singular values were chosen for further computations.

The test example q is projected onto the lower-dimension SVD space as follows:

$$q_{svd} = q^T U \Sigma^{-1} \tag{24}$$

Estimating the probability density functions in this reduced-dimension SVD space helps increase the classification accuracy.

2.4 Analysis of Active Sites in Apo Structures

The analysis described above depends on the knowledge of the active-site shape based on ligand coordinates. Since the ligand identity is unknown in the case of a test active site, it is necessary to extend the present analysis to clefts on the surface of an unliganded protein. This extension will introduce slight variations into the active site shape descriptors and the feature-based analysis has to be robust enough to deal with these inaccuracies. To increase the generality of our approach, we recreated our database to contain uniform-radius active sites. These active sites were created by first choosing a surface vertex closest to the ligand center as the center of the active site. All surface vertices within a chosen radius were then considered to be part of the active site. The choice of the radius depends on the statistical analysis of the fragment pockets in our database (analysis of the average distance of an active site vertex from the center of the ligand). For example, the average distance of an active site vertex from the center of phosphate was found to be 4Å while this distance was 5Å in the case of the other 5 ligands in our database. This definition of the active site as a pocket of a uniform chosen radius, introduces variations in active site shape absent from our earlier definition of the active site pocket based on contact with the ligand. These uniform-radius active sites form the training set during future classification of active sites of unliganded proteins.

2.5 Protein-Ligand Complex Database

A list of proteins complexed with the ligands of interest (adenine, citrate, nicotinamide, phosphate, pyridoxal and ribose) were obtained from the PDBSelect90 [14] list. This was done to ensure that the dataset was non-redundant with sequences with at most 90% identical. The *third column* in Table 1 shows the average homology of the examples of each fold to the other members of the ligand family. The average homology within each class is less than 30% in all cases. The *second column* in Table 1 lists the number of diverse fold families within each ligand class (based on the *SCOP* fold classification by [21]. This table shows that we have a diverse dataset containing examples belonging to various fold families with very low sequence homology to each other. No additional resolution thresholds were applied during database creation (database consists of medium-low resolution structures). In order to maintain class-balance between our six ligand classes, 55 examples of each class were chosen to form the database. The one exception was the citrate dataset which contains only 16 examples. The dataset contains only those cases where the ligand is bound in the active site

and cases where the ligand was bound on the surface (*e.g.* as solvent molecules) were excluded. The protein atoms were separated from the ligand atoms and water molecules were removed in each case and care was taken to ensure that only a single active site of a given protein-ligand complex was considered (when multiple chains of a protein all bound to the same ligand).

3 Results and Discussion

In this section, we examine the ability of our localized stereochemical features to discriminate between active sites that bind adenine, citrate, nicotinamide, phosphate, pyridoxal and ribose. These 6 ligands were chosen since all of them display a wide variation in their active sites. We will also examine the effects of using shape/electrostatic features alone as well as the effects of the choice of a uniform-radius for the active site definition on the overall classification accuracy.

The variation of the largest eigenvalue for the six different ligand classes seen in Panel (a.) of Figure 1 shows that the pyridoxal active sites have the largest values ranging from 7.5 to 15.5Å and the phosphate active sites have the smallest values (3.5Å to 8.5Å). The variation of the cross-sectional shape feature at 4Å for the six ligands seen in Panel (b.) of Figure 1 shows that the phosphate active sites have the largest variation in cross-sectional feature values at 4Å while the citrate active sites show the least variation.

The efficacy of our feature-based method was tested by a leave-one fold family out experiment. In this experiment, each fold family within a ligand class was used as test set and the rest of the examples were used a training data. For example, in order to test the accuracy of features to classify the *citrate synthase* family, it was completely removed from the class of citrate-binding examples. This new subset of citrate-binding sites was then used in training. The probability distribution obtained from kerenel density estimates (based on feature-vectors in reduced SVD space) were then used to classify citrate synthase active site surfaces as citrate-binding. The accuracy of this approach (using active site surfaces defined by contact with known ligand) for all the 6 ligands is shown in the *fourth column* of Table 1. The confusion matrix in Table 2 shows the misclassification between classes. The greatest confusion is between the citrate and the pyridoxal sites (13%), perhaps because of the similarities in size. These tables show that the classification accuracy of the feature-based approach is greater than 75% in the case of all the 6 ligands considered. In our dataset, the phosphate binding sites stand out due to their smaller size and this is reflected in the higher classification accuracy (91%). The high accuracy (81%) in the case of citrate binding sites is also encouraging since this indicates the possibility of accurate classification even in the case of ligands with fewer known binding patterns (only 16 examples as opposed to 55 examples of every other class). These results show that despite the diversity in fold families as well as low homology between the members within a ligand class, the feature-based approach presented in this paper is able to identify the commonality between diverse active sites binding the same ligand. This recognition of similar active sites therefore goes beyond similarity due to sequence homology or fold family similarity.

Our algorithm uses novel local geometric descriptors of active sites capable of capturing some spatial information about the pocket shape and electrostatic nature, and we have argued that it is essential to combine fine-grained characterizations of both shape and chemistry in order to truly capture binding site patterns among diverse active site surfaces. This hypothesis was tested by examining the classification accuracy in two experiments, one using just the geometric features and the other using just the electrostatic features to describe the active sites belonging to various ligand classes (using known ligand coordinates for active site surface definition). The *sixth* and *seventh* columns of Table 1 show that in the case of all ligand classes there is a 30% or greater drop in classification accuracy when only geometric or only electrostatic features are used except in the case of phosphate binding examples. When only geometric features are used, there is a clear distinction between active sites that bind phosphate and those that do not, leading to a 100% accuracy in the classification of phosphate binding sites. Despite this anomaly, these results clearly show that neither shape nor electrostatic descriptors alone are sufficient to describe the active site patterns and that it is necessary to combine these features for an increased accuracy in active site recognition.

Finally, as described in Section 2.4, uniform-radius pockets were created for each of the 6 ligands. The leave-one fold family out analysis was performed on these pockets to verify that our algorithm could accurately classify these feature vectors (noisy but more similar to real-world application). The *fifth column* in Table 1 shows that despite a drop in classification accuracy with these uniform-radius features, the feature-based approach is still able to

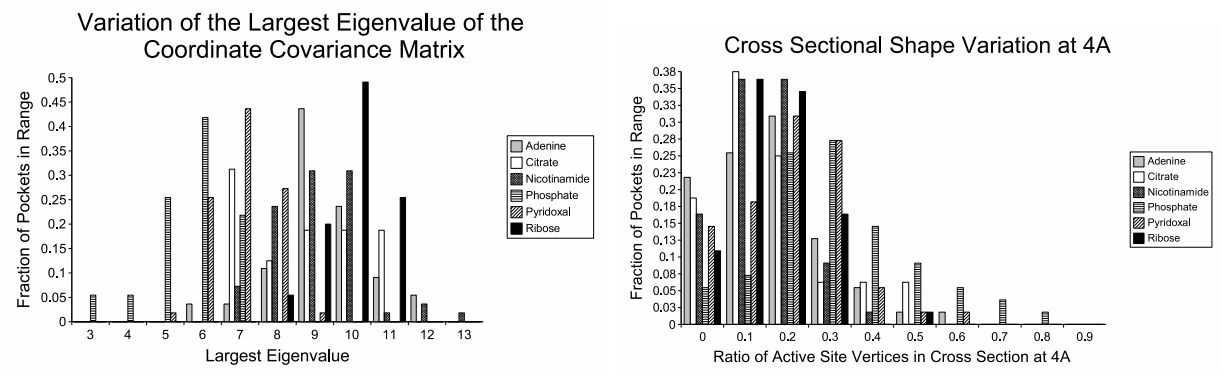

Figure 1: The variation of the largest eigenvalue of the coordinate covariance matrix (panel a) and the variation of cross-sectional feature at 4Å for the uniform-radius active sites belonging to the six ligand classes.

capture the binding patterns associated with the 6 ligands with an accuracy greater than 60% in almost all cases (for citrate binding sites the accuracy drops to 56%). While this is a reduction in accuracy, this accuarcy is still considerably higher than the ones obtained with either shape or electrostatic features considered alone.

4 Conclusions

This paper presents a new approach to active site analysis that goes beyond traditional sequence homology/fold classification and 3D residue templates. We use a feature-based analysis that extends previous global features with more fine-grained features that capture spatial distribution of shape and charge. There exists a great diversity amongst active sites that bind the same ligand and it is necessary to account for this diversity during functional analysis. The sterochemical properties captured by the features are able to capture this diversity without relying on any sequence or secondary structure information.

The use of fine-grained features accurately identified the native ligand for various active site surfaces with upwards of 75% accuracy in almost all cases. This accuracy was maintained even when all examples of the fold families of the test cases were absent during training. This is promising since it allows for the possibility of identification of heretofore unknown binding folds using our methodology. This study has also shown the importance of incorporating both electrostatic and shape features for active site analysis.

References

[1] Artymiuk P., Poirrette A., Grindley H., Rice D. and Willett P. (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, **243**, 327-344.

[2] Bagley SC., Wei L., Cheng C. and Altman RB. (1995) Characterizing oriented protein structural sites using biochemical properties. *In Proc. 3rd Intl. Conf. on Intelligent Systems for Mol. Biol.*, 12-20.

[3] Bagley SC. and Altman RB. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Science* **4(4)**, 622-635.

[4] Bartlett GJ., Porter CJ., Borkakoti N. and Thornton JM. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105-121.

[5] Barker JA. and Thornton JM. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644-1649.

[6] Connolly M. (1983) Analytical Molecular Surface Calculation. *J. Appl. Cryst.* **16**, 548-558.

[7] Cornell WD., Cieplak P., Bayly CI., Gould IR., Merz KM., Ferguson DM., Spellmeyer DC., Fox T., Caldwell JW., and Kollman PA. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.

[8] Denessiouk KA., Rantanen V. and Johnson MS. (2001) Adenine recognition: a motif present in ATP, CoA, NAD, NADP and FAD dependent proteins. *Proteins:Structure, Function and Genetics* **44**, 282-291.

[9] Duda RO. and Hart PE. (1973) Pattern Classification and Scene Analysis. New York: John Wiley & Sons.

[10] Fetrow JS., and Skolnick J. (1998) Method for perdiction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949-968.

[11] Gutteridge A., Bartlett GJ. and Thornton JM. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol* **330**, 719-734.

[12] Hermann JC., Marti-Arbona R., Fedorov AA., Fedorov E., Almo SC., Shoichet BK. and Raushel FM. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* 2007; **448**, 775-779.

[13] Hindle SA., Rarey M., Buning C. and Lengaue T. (2002) Flexible docking under pharmacophore type constraints. *J. Computer Aided Mol. Des.* **16(2)**, 129-149.

Table 1: Description of Fragment Database and Analysis of Classification Accuracy

Ligand Name	Num. of Families	Avg. Homology Between Families	Accuracy (Contact Surface)	Accuracy (Uniform Radius) Surface	Accuracy Only Geometric Features	Accuracy Only Electrostatic Features
Adenine	19	7.9%	43/55(78%)	38/55(69%)	25/55(45%)	29/55(52%)
Citrate	12	8.7%	13/16(81%)	9/16(56%)	7/16(44%)	6/16(38%)
Nicotinamide	11	7.3%	46/55(84%)	35/55(64%)	31/55(56%)	22/55(40%)
Phosphate	23	6.9%	50/55(91%)	53/55(96%)	55/55(100%)	25/55(45%)
Pyridoxal	6	10.2%	48/55 (87%)	37/55(67%)	20/55(36%)	26/55(47%)
Ribose	22	8.5%	43/55(78%)	35/55(64%)	24/55(44%)	23/55(42%)

Table 2: Confusion Matrix For the Six Ligand Classes Using Active Sites Definitions Based on Known Ligand Coordinates; I=Adenine, II=Citrate, III=Nicotinamide, IV=Phosphate, V=Pyridoxal, VI=Ribose

True Class	% Accuracy					
	I	II	III	IV	V	VI
I	0.78	0.02	0.11	0.02	0.05	0.02
II	0	0.81	0	0	0.06	0.13
III	0	0.04	0.83	0.05	0.04	0.04
IV	0	0	0.09	0.91	0	0
V	0	0.13	0	0	0.87	0
VI	0	0.09	0.07	0.02	0.05	0.77

[14] Holm L. and Sander C. (1994) Enlarged representative set of protein structures. *Protein Science* **3**, 522.

[15] Honig B., and Nicholls A. (1995) Classical electrostatics in biology and chemistry. *Science*, **268(5214)**, 1144-1149.

[16] Huang C, Smith CV, Glickman MS., Jacobs WR., Sacchettini JC. (2002) Crystal structures of mycolic acid cyclopropane synthases from Mycobacterium tuberculosis. *J.Biol.Chem.* **277**, 11559-11569.

[17] Kimber MS, and Pai EF. (2000) The active site architecture of Pisum sativum β-carbonic anhydrase is a mirror image of that of α-carbonic anhydrases. *EMBO Journal* **19**, 1407-1418.

[18] Li AJ, Nussinov R. (1998) A Set of van der Waals and Colombic Radii of Protein Atoms for Molecular and Solvent-Accessible Surface Calculation, Packing Evaluation and Docking. *Proteins: Structure, Function and Genetics* **32**, 11-127.

[19] Maurice M, Mera PE, Taranto MP, Sesma F, Escalante-Semerena JC and Rayment I. (2007) Structural characterization of the active site of the PduO-type ATP:Co(I)rrinoid adenosyltransferase from Lactobacillus reuteri. *J. Biol. Chem.* **282(4)**, 2596-2605.

[20] Meng EC., Shoichet BK. and Kuntz ID. (1992) Automated docking with grid-based energy evaluation. *J. Comp. Chem.* **13**, 505-524.

[21] Murzin AG., Brenner SE., Hubbard T., and Chothia C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540. **321(5)**, 741-765.

[22] Porter CT., Bartlett GJ. and Thornton JM. (2006) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Acids. Res.* **32**, D129-D133.

[23] Richards FM. (1977) Areas, volumes, packing and protein structure. *Annual Review of Biophysics and Bioengineering* **6**, 151-176.

[24] Laskowski RA., Watson JD. and Thornton JM. (2005) Protein function prediction using local 3D templates. *J.Mol.Biol* **351**, 614-626.

[25] Rosin PL. (2005) Computing Global Shape Measures. *In Handbook of Pattern Recognition and Computer Vision*, 177-196.

[26] Sadjadi FA and Hall EL. (1980) Three-dimensional moment invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **2(2)**, 127-136.

[27] Wall ME., Rechtsteiner A and Rocha LM. (2003) Singular Value Decomposition and Principal Component Analysis. *In A Practical Approach to MicroArray Data Analysis*, 91-109.

[28] Wei Y., Ringe D, Wilson M, Ondrechen M. (2007) Identification of Functional Subclasses in the DJ-1 Superfamily proteins. *PLoS Computational Biology*, **3(1)**.