Prediction of Active Site Cleft Using Support Vector Machines

Shrihari Sonavane and Pinak Chakrabarti*

Department of Biochemistry and Bioinformatics Centre, Bose Institute, P-1/12 CIT Scheme VIIM,
Kolkata 700 054, India

Received August 2, 2010

Computational tools are available today for the detection and delineation of the clefts and cavities in protein 3D structure and ranking them on the basis of probable binding site clefts. There is a need to improve the ranking of clefts and accuracy of predicting catalytic site clefts. Our results show that the distance of the clefts from protein centroid and sequence entropy of the lining residues, when used in conjunction with the volume, are valuable descriptors for predicting the catalytic site. We have applied the SVM approach for recognizing and ranking the active site clefts and tested its performance using different combinations of attributes. In both the ligand-bound and the unbound forms of structures, our method correctly predicts the active site clefts in 73% of cases at rank one. If we consider the results at rank 3 (i.e., the correct solution is among one of the top three solutions), the correctly predicted cases are 94% and 90% for the bound and the unbound forms of structures, respectively. Our approach improves the ranking of binding site clefts in comparison with CASTp and is comparable to other existing methods like Fpocket. Although the data set for training the SVM approach is rather small in size, the results are encouraging for the method to be used as complementary to other existing tools.

INTRODUCTION

Data on protein 3D structures are increasing exponentially as a result of various structural genomics projects. The main goal of these projects is to understand functions of all genes in nature. Function itself can be defined at different levels (such as molecule, organelle, cell, tissue, organ, and organism). A lower-level function can be part of different higher-level functions; for example, a protease may be involved in digestion, wound healing, as well as fertilization. Servers are being developed to annotate proteins with Gene Ontology functional terms by extracting features such as 3D fold, sequence, motif, and functional linkages.[1] Even at the molecular level, predicting the function from the 3D structure is a challenging task, and various methods have been developed for this purpose. One can infer the function from an analysis of catalytic residues in enzyme active sites.[2−4] Graph theoretic approach and local spatial similarities have also been used to uncover functional sites[5−13] along with the concept of travel depth.[14] Residue conservation has been an important criterion for the identification of functional sites.[15−26] As ligand-binding sites are located on the protein surface, cleft detection has been an important tool for their recognition.[27−30] Computed electrostatic potential can be used to derive the pK_a values to identify the residues that can perform Brønsted acid−base chemistry.[31−33] Energetic stability considerations[34,35] and combinations of other biophysical properties have also been employed for functional site prediction.[36] Some methods also use the sequence and structural information along with support vector machines to predict the catalytic residues.[37,38]

There is a need for computational tools for predicting enzyme active sites from 3D structure, particularly for the cases where the function cannot be assigned by structure/sequence comparison with other known proteins. Computational tools like VOIDOO,[39] MS package,[40,41] VOLBL,[42,43] CAST (now rechristened as CASTp),[29] a Monte Carlo (MC) procedure,[44] CLIPPERS,[6] etc., are available for the detection and delineation of the clefts and cavities in protein 3D structure. Our goal in this work is to identify the active site cleft where catalysis and/or ligand recognition occurs. We have analyzed the active site clefts, their distance from the protein centroid, and the sequence conservation of cleft residues. In addition, the SVM[45]-based approach, which has gained popularity over other machine learning methods because of its ability to effectively handle noise and large data sets/input spaces, has been used to predict the active site clefts.

MATERIALS AND METHODS

Data Set. A nonredundant set of enzymes was selected from 969 entries in CSA, Catalytic Site Atlas, http://www.ebi.ac.uk, version 2.2.10,[2,4] having literature information for active site residues. These entries were submitted to the PISCES server[46] to identify proteins with low sequence similarity. The input parameters used for culling are the following: sequence percentage identity ≤25%, resolution ≤2.0 Å, R-factor ≤0.2, sequence length 40−10 000 amino acids. A single structure was then taken for each Enzyme Classification (E.C.) group.[47] The resulting data set contained 58 (48 monomers and 10 homodimers) enzymes, and the list is provided in Table S1. For simplicity, only one chain of dimeric proteins was considered in the analysis. Atomic coordinates of the proteins were extracted from the Protein Data Bank (PDB) located at Research Collaboratory for Structural Bioinformatics (RCSB).[48]

* Corresponding author fax: +91-33-2355-3886; e-mail: pinak@boseinst.ernet.in; pinak_chak@yahoo.co.in.

10.1021/ci1002922
Published on Web 11/16/2010

Table 1. Ranking of the Active Site Cleft on the Basis of Volume, Distance from the Protein Centroid, and the Average Sequence Entropy of the Cleft-Lining Residues in the Training Data Set[a]

	ranking based on		
PDB ID	volume	distance from centroid	entropy
1chm	2	1	1
1dae	2	1	1
1kp2	2	1	1
1ra0	6	1	1
1ako	7	1	1
1amp	1	1	2
1hfs	1	1	2
1jdw	1	1	2
1l9x	1	1	2
1oyg	2	1	2
1mrq	1	1	3
1e1a	2	2	1
1trk	2	2	1
1a2t	1	2	2
1b6g	2	3	4
1ofd	3	4	2
1o98	5	10[b]	10

[a] Only those structures (from the ones given in Table S1) are included that have one or more descriptors with rank > 1. [b] The active site cleft is ranked second on the basis of the distance from protein centroid if all the clefts with volume greater than active site cleft are considered in ranking.

Identification of Clefts and Cavities. Clefts and cavities for each protein structure were identified using the CASTp (Computed Atlas of Surface Topography of proteins) server[49] located at http://sts.bioengr.uic.edu/castp/. CASTp provides a full description of protein pockets and cavities, including volume, surface area, protein atoms that line the concavity, and features of pocket mouth(s) including identification of mouth atoms as well as measurement of mouth area and circumference. The default probe radius of 1.4 Å has been used for our calculations. The protein volume was calculated using the program ProGeom (server: http://nook.cs.ucdavis.edu/~koehl/ProShape/download.html).

Mapping Active Sites. Information about the residues present in the active site was obtained from CSA.[2,4] The residues were mapped into the clefts obtained from CASTp. The cleft that is lined by all the catalytic residues is considered to be the active site cleft. For most of the cases, all the catalytic residues reside in a single cleft. In cases where they reside in different clefts, we searched for the location of the functional part (main chain or side chain) of the residues, as mentioned in CSA. If they still resided in different clefts, the corresponding entry was removed. The final list is given in Table 1.

Sequence Alignments. The multiple sequence alignments for each enzyme sequence were extracted from the HSSP (homology-derived secondary structure of proteins) database of sequence-structure alignments.[50] The database provides for each PDB entry a list of protein sequences deemed structurally homologous to it on the basis of a homology−threshold curve. The default threshold of 30% sequence identity has been used.

Calculation of Entropy and Mean Sequence Entropies for the Clefts. Entropy is calculated using Shannon's information theoretic entropy to measure the variability at a particular position in a given protein sequence.[51] Sequence entropy is given by the following expression:

$$s(i) = -\sum p(k)\cdot \ln(p(k))$$

where $p(k)$ is the probability that the ith position in the sequence is occupied by a residue of class k. A low value of sequence entropy, $s(i)$, at position i in the multiple sequence alignment implies that the position has been subjected to relatively higher evolutionary conservation than another position in the same alignment having a higher sequence entropy value.

The simple average was taken to calculate the mean entropy value for individual clefts:

$$\langle s\rangle = \sum s(i)/n$$

where the summation is over all (number, n) the cleft-lining residues. We used these average values to rank the clefts, 1 (for the cleft with the lowest entropy) through 9. Any cleft with higher entropy values was assigned the rank 10.

Distance from the Protein Centroid. A threshold of at least one-fourth the volume of the largest cleft for a particular protein was used to select the pockets for that protein. For each cleft with volume above this threshold, the Euclidian distance between the protein centroid and the nearest atom lining the cleft was calculated. Clefts (both pockets and cavities) were then ranked from 1 (nearest to the centroid) through 9. Rank 10 was assigned to all the remaining clefts further out.

Amino Acid, Atom Type, and Secondary Structure Composition. Amino acid composition is given as C_x, where

$$C_x = N_x/N_a$$

N_x is the number of amino acid residue of type X, and N_a is its total number of amino acid residues lining the cleft. Even if one side-chain atom of a particular residue was located in the cleft, it was included; main-chain atoms were not considered in the calculation of composition. The atom type composition considering different types of atoms, as classified by us[52] and the occurrence of secondary structural elements (helix, strand, and the rest, termed "others"), lining the cavities was also calculated in a similar fashion. Secondary structure assignments were made using the DSSP program.[53]

Training and Test Data Sets. From the 58 protein structures, the clefts were selected applying the volume threshold. These clefts (training data set) were used for optimizing the SVM parameters and for training the SVM classifier. The training data set contains 129 clefts (57 positive and 72 negative examples. As discussed under Results, the volume criterion used for identifying the active site would miss the one present in the PDB file, 1o98. As such, this active site was excluded from the list of positive cases). The "PocketPicker" data set of 48 proteins used in a previous study was used as the test set[54] (Table S2). While the structures in the training set compiled by us contained only enzymes, the test set had 39 enzymes and 9 proteins binding various ligands (such as, carbohydrate, fatty acid, biotin, retinol, etc.).

Identifying Binding Clefts in the Test Set. The protein atoms within 3.5 Å of any ligand atom were extracted (the detailed list is provided in Table S3) and mapped into the cleft lining atoms obtained from CASTp. In most of the cases, they reside in the same cleft. Sometimes small clefts share the common atoms with the larger ones; in such cases, the ligand pockets were manually inspected. If the ligand atoms were present in two separate clefts, both were considered as active site clefts. As our goal was to identify the active site cleft by identifying one or more residues in contact with the bound ligand, we used a rather conservative distance of 3.5 Å; the use of a longer distance increases the likelihood of inclusion of additional clefts also.

Ranking Attributes. The attributes used for training the SVM are the distance, surface area, volume, and sequence entropy of the cleft, compositions based on residue (20 attributes), secondary structure (3), and atom type (13), and the volume of all the neighboring clefts that share at least one cleft-lining atom with the cleft under consideration. All 41 attributes used for training were evaluated and ranked using the InfoGainAttributeEval algorithm available in Weka version 3.4.11 using 10-fold cross validation.[55,56] The list and description is provided in Table S4.

SVM Implementation. The freely downloadable LIBSVM package by Chang and Lin was used for the implementation of SVM with the C-SVC SVM type (SVM type for classification) and the widely used Radial Basis Function (RBF) kernel.[57] Two parameters are required for optimizing the SVM classifier: γ, which determines the capacity of the RBF kernel, and the regularization parameter C. Each feature in training and test data sets was scaled in the range of −1 to 1.

SVM Optimization. The kernel parameters γ and C have been optimized using a grid search and 5-fold cross validation. In 5-fold cross-validation, the training data set was spilt into five subsets, where one of the subsets was used as the test set, while the other subsets were used for training the classifier. The trained classifier was tested using the test set. The process was repeated five times using a different subset for testing, thereby ensuring that all subsets were used for both training and testing. Matthews correlation coefficient was used during cross-validation instead of percent accuracy as the positive negative ratio in training data set is 57:72.

Performance Measure. The performance was measured by prediction accuracy, and the Matthews correlation coefficient (MCC) was calculated as

$$\text{accuracy} = ((\text{Tp} + \text{Tn})/(\text{Tp} + \text{Fn} + \text{Tn} + \text{Fp}))*100$$

$$\text{MCC} = ((\text{Tp}*\text{Tn}) - (\text{Fp}*\text{Fn}))/\sqrt{((\text{Tp} + \text{Fp})*(\text{Tp} + \text{Fn})*(\text{Tn} + \text{Fp})*(\text{Tn} + \text{Fn}))}$$

Tp, Fp, Tn, and Fn represent the numbers of true positive, false positive, true negative, and false negative, respectively. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and −1 an inverse prediction. MCC can be used for comparison of the results with different positive to negative ratios, whereas accuracy is sensitive to data set imbalance.

RESULTS

To test the applicability of any methodology aimed at identifying the active site, one needs to have a nonredundant data set of enzymes. Such a data set has been curated by us (Table S1). To enable us to compare the result with other methods, we have used a test set (Table S2) that has been used in literature.[54]

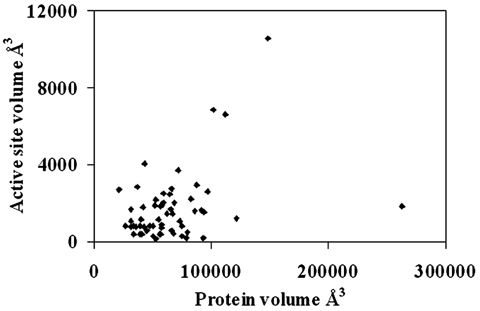

Figure 1. Plot of active site volume against the volume of the protein. The correlation coefficient $r = 0.4$.

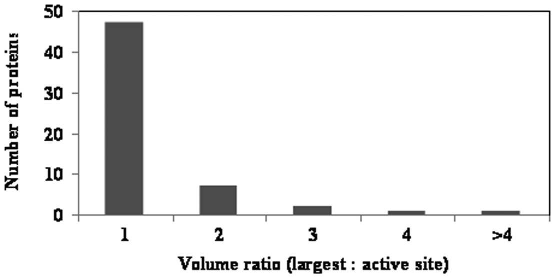

Figure 2. Distribution of the volume ratio: the largest cleft to the one containing the active site.

Active Site Cleft Volume. The volume of the active site clefts ranges from 121 to 10 570 Å^3. In 58 enzymes analyzed, in the majority (47) of the cases the active site resides in the largest cleft; it is the second largest in seven proteins, and in four cases the active site does not belong to the largest or the second largest cleft. The maximum rank of the active site cleft on the basis of volume is 7 from the largest cavity, and only in three cases it is above 4. We also studied if there was any correlation between the volume of the active site cleft and the protein volume. However, the two are poorly correlated (Figure 1).

As the active site pocket is usually one of the largest pockets in the structure, the volume can be used as a selection criterion to choose the likely candidates. Figure 2 plots the distribution of the volume ratio (the largest cleft as compared to the active site cleft). Except for one enzyme (PDB ID 1o98), the ratio is not more than 4.

Cleft Distance from the Protein Centroid and Sequence Entropy. The distance of active site clefts from the protein centroid is calculated as described in Materials and Methods. CASTp detects all the voids as well as clefts in the protein tertiary structure. However, the smaller clefts are not likely to constitute the active site, and Figure 2 shows that the ratio of the volumes of the largest cleft to the active site cleft is usually not more than 4. As such, to minimize the search space and noise created by the inclusion of small clefts, a volume threshold (volume greater than 25% of the largest cleft) was used for selection. When the pockets are ranked on the basis of the distance from the protein centroid (rank 1 indicating the closest), in the majority (52) of the cases the active site cleft is ranked first (independent of the cleft volume), the rank is 2 in three cases, and it is beyond 2 in another three (Table 1). However, it is among the top four in all but one case. (For the exceptional case of 1o98, as its volume ratio was >4 it was not included for the distance

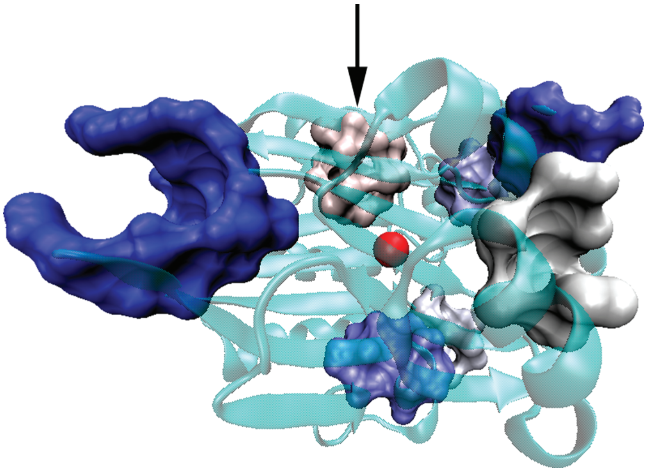

Figure 3. Surface representation of the top seven clefts present in the structure of exonuclease III (PDB file, 1ako). Clefts are colored on the basis of volume from blue (largest) to red, but the ones smaller than the active site cleft are omitted. Arrow indicates the active site cleft. The red sphere is placed at the centroid of the protein, which is displayed in cartoon.

calculation. To reflect this situation, this entry was assigned a rank of 10.) One example of the improvement of the ranking of the active site on the basis of distance from the protein centroid rather than on volume alone is shown in Figure 3, where the rank of the active site cleft improves from 7 to 1.

The same volume threshold was used to select clefts for the calculation of the average sequence entropy. In 48 cases, the active site comes at the top, in six cases at rank 2, and beyond 2 in four cases. As in the case of ranking on the basis of distance from the protein centroid, the active site cleft is also among the top four in terms of entropy.

SVM Training and Predictions. From 58 structures, clefts that satisfy the volume threshold are included in the training data set. The best combinations of γ and C obtained from the optimization process were used for training the SVM classifier. The SVM classifier was subsequently used to predict the test data sets. Normally one gets the result in an affirmative or negative fashion (a pocket corresponds to the active site or not). However, one can also apply a probability model of the SVM, which provides a probability estimation for being the active site for the cleft under consideration. On this basis, we ranked the clefts as the probable active site.

We tested the performance of SVMs resulting from combinations of different top-ranked attributes. The attributes considered for predictions and the optimized values for γ and C are provided in Table 2a. Consideration of MCC values shows that the best results are obtained when the SVM model is trained using the top 10 attributes (Table S4), which gives an accuracy of 73% on the test data set (both bound and unbound forms) (Table 2b) for predicting the active site at rank 1. If the structures in the test set that have sequence similarity >25% with any structure in the training set are excluded (12 cases), the accuracy for the top rank is 72% and within top three is 97%.

The active site cleft prediction flowchart is shown in Figure 4; the number of clefts satisfying conditions at each prediction step using the SVM model (trained using top 10 attributes) is indicated. For a particular structure, only the clefts satisfying the volume criteria can be ranked.

Table 2. SVM Training and Prediction: Based on the Ranking (Table S4), Different Combinations of Top-Ranked Attributes Are Used (a) To Optimize the Values for C and γ (Using the Training Set), and (b) the Corresponding SVM Prediction for the Test Set[a]

(a) attributes	C	γ	MCC
all	1	0.00085	0.87
first 5	1	0.00129	0.83
first 10	1	0.25000	0.89
first 15	1	0.00129	0.88
first 20	1	0.00129	0.86
first 25	1	0.00129	0.86
first 30	1	0.00074	0.87
first 35	1	0.00074	0.87

	ranking[b]			
(b)	bound		unbound	
attributes	rank 1	rank 3	rank 1	rank 3
all together	73	94	73	90
first 5	70	94	73	88
first 10	73	94	73	90
first 15	75	96	77	88
first 20	75	94	75	90
first 25	73	94	71	90
first 30	75	92	73	90
first 35	75	92	73	90

[a] "First X" indicates that the top X attributes from Table S4 have been used. [b] Numbers correspond to the % structures where at least one active site cleft is predicted with rank 1; rank 3 means within ranks 1−3 inclusive. Both ligand-bound and unbound structures are considered.

We checked the individual performance of the top four attributes. The one involving the distance performs the best, followed by entropy (Table S5). Interestingly, these two attributes outperform the SVM models obtained using different combinations (Table 2b). Rank 3 results are comparable for individual, as well as different, combinations. If distance and entropy are used together to train the SVM, the MCC value obtained (0.83) is lower than the values given in Table 2a for all other combinations using more number of attributes.

DISCUSSION

Features of the Active Site Pockets. In the training data set, the average number of atoms and residues in the active site pockets are 141 (±133) and 49 (±41), respectively. The average volume is 1669 (±1795) Å^3. Figure 5 shows that the number of catalytically active residues is usually in the range 2−5, but the cleft containing them can have up to 252 residues (the PDB files having a large number of cleft-lining residues, 252 in 1pj5, 176 in 1lj1, 138 in 1o8a, and 133 in 1jfl, all have channels leading to the active site), and there is no apparent correlation between them.

Volume Threshold. Thornton and co-workers[58] have shown that the active site of an enzyme can be identified using purely geometric criteria. If there is a cleft considerably larger than others, then this largest cleft is most likely to be an enzyme active site. Out of 58 examples, the active site was located in the largest cleft in 81% cases. In 11 cases, the active site was located in clefts other than the largest one. As the largest cleft does not always harbor the active site, we have to consider a few of the large clefts as the

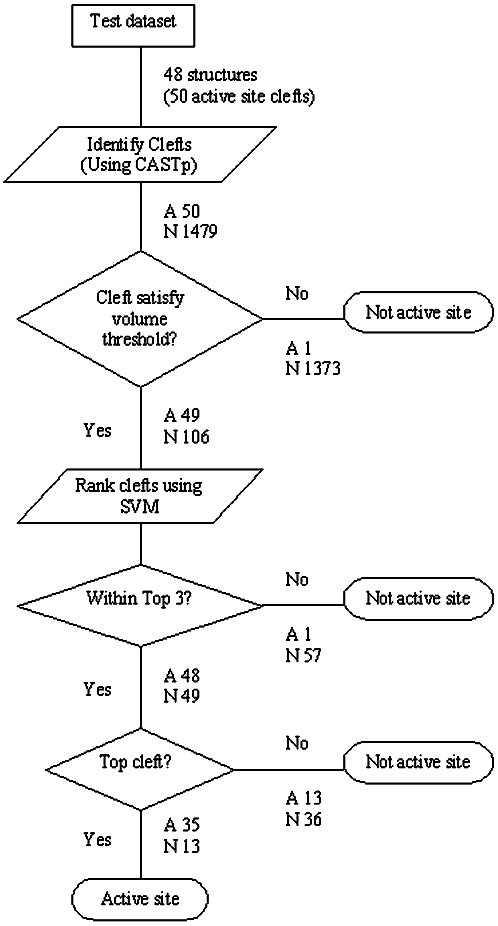

Figure 4. The flowchart for the prediction of the active site cleft in the test set. Numbers at each step correspond to the number of active (preceded by "A") and other ("N") clefts satisfying the corresponding condition (using the bound form of the structures).

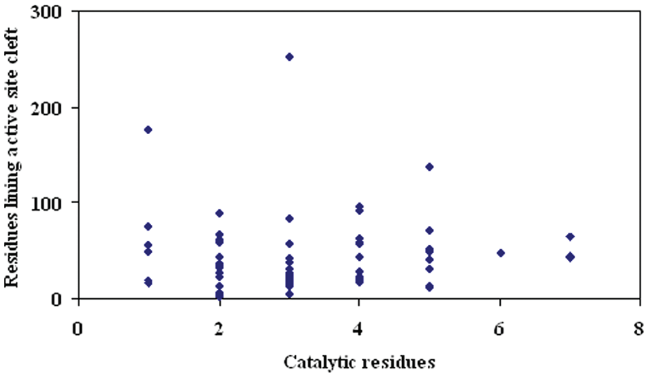

Figure 5. Plot of the number of catalytic residues against the number of residues lining the active site cleft.

likely candidate. We considered all pockets that have a volume of at least 25% of the largest cleft. With this threshold, all the active site clefts get included (a total of 163 clefts containing 57 active sites), except one (PDB file 1o98). If this cleft was to be included, the threshold had to be lowered to ~16%, which would have increased the number of clefts under consideration to 233 (increase by 43%).

Distance from the Protein Centroid and Sequence Entropy of Clefts. It has been reported that the distance of the active site residues and cleft from the protein centroid is a useful property for the prediction of active site clefts.[27] Here, we have used the same property after screening the clefts with volume threshold, and it seems to be a more effective attribute as compared to the volume alone; in 90% cases, the active site comes with the top rank if we use the cleft distance together with the volume criteria, as opposed to using volume alone (81%). Conservation of residues, as measured by sequence entropy, is an important indicator of the involvement of a residue in protein function, in particular ligand binding, catalysis, and protein−protein interaction.[15−20,24,25,21,23,26,22,51] Similar to the distance criterion, the active site comes as rank 1 in 83% cases on the basis of entropy (Table 1).

Features of Cavities in the Test Set. 48 structures compiled by Weisel et al.[54] have been used to judge the performance of the method. All these structures are available in the bound and unbound forms. Percentage changes in the number of atoms, residues, and volume in the active site pocket while going from the bound to unbound form are given in Table 3. It may be noted that in some cases there is a change in the number of active site pockets between the two forms. Only in one case (PDB file 1qpe) does the bound form have the active site spread over two clefts, whereas such cases are much more common (16 cases) in the unbound form. In 31 cases is the active site located in a single pocket in both forms. Based on the sign of the change in volume, for 14 the active site cleft is larger in the bound form, whereas it is the reverse for the remaining 17 cases. Thus, there is no clear trend in the change in the physical characteristics as the ligand binds the apo form of the enzyme. For the test set, the active site is located in 49 and 75 clefts, in the bound and the unbound forms, respectively, which needed to be identified from 155 and 166 clefts (selected on the basis of volume) in the two forms of the structure.

Comparison with Other Studies. The performance of our method in relation to that from the other algorithms can be seen in Table 4. The PocketPicker data set[54] also contains other proteins binding various ligands (such as carbohydrate, fatty acid, biotin, retinol, etc.) along with enzymes, but for the sake of comparison we have not modified this data set. In the bound form of structures, our method correctly predicts the active site clefts in 73% cases at rank one, the corresponding number for the unbound form is 73%. If we consider the results at rank 3 (i.e., the correct solution is among one of the top three solutions), our method outperforms others, except Fpocket that exhibits a slightly better performance for the unbound form. It may be mentioned that SplitPocket[30] reports a success rate of 95% on the same data set. However, unlike our method, which does not need any information on the ligand occupancy by the clefts, SplitPocket explicitly uses the structure with the ligand bound in the pocket so that an analysis can be performed on the changes brought about in the integrity of the surface wall by the presence of the ligand. At the level of rank 3, our method performs comparably to SplitPocket, although the latter does not rank the clefts. Our training data set contains only enzymes; for testing, if we restrict ourselves to only enzymes from the PocketPicker data set, the accuracy at rank 1 increases to 77% for the bound and to 79% for the unbound form, the corresponding values at rank 3 being 97% and 90%, respectively.

Table 3. Comparison of the Active Site Pockets in the Ligand-Bound and Unbound Forms in the Test Set

PDB ID	number of clefts containing active site residues[a]		% change[b] (bound to unbound)		
(bound−unbound)	bound	unbound	no. of atoms	no. of residues	volume
1a6w−1a6u	1	1	28.6	0	31.2
1acj−1qif	1	1	16.5	14.6	15.5
1apu−3app	1	1	−13.9	−7.1	−32.2
1bid−3tms	1	1	6.3	−6.1	9.4
1blh−1djb	1	1	−21.4	−28.6	−58.3
1byb−1bya	1	1	5.5	6.1	−31.8
1cdo−8adh	1	1	28.1	25	34.3
1fbp−2fbp	1	1	75.4	54.2	84.1
1gca−1gcg	1	1	−57.7	−55	−83.7
1hew−1hel	1	1	−120	−57.1	−214.1
1hfc−1cge	1	1	−2.9	0	7
1ida−1hsi	1	1	−50	−6.3	−42.8
1igj−1a4j	1	1	63	66.7	91.6
1imb−1ime	1	1	14.5	8.5	19.2
1ivd−1nna	1	1	−188.5	−125	−283.5
1mrg−1ahc	1	1	34.6	35	65.5
1okm−4ca2	1	1	0	5.9	−10.8
1phd−1phc	1	1	0.8	0	−0.2
1pso−1psn	1	1	−21.7	−21.8	−30
1rne−1bbs	1	1	−19.4	−18.2	−22
1rob−8rat	1	1	−55	16.7	−92.1
1snc−1stn	1	1	37.8	33.3	34.7
1srf−1pts	1	1	0	0	−0.6
1stp−1swb	1	1	−2.3	0	−19.3
1ulb−1ula	1	1	−44.6	−34.1	−36.6
2ctc−2ctb	1	1	−2.5	10	−7.4
2h4n−2cba	1	1	−2.6	5.6	8.1
2ifb−1ifb	1	1	5.4	−3.1	2.6
4dfr−5dfr	1	1	6	2.6	−3.5
4phv−3phv	1	1	44	45	59.8
5cna−2ctv	1	1	45.7	43.8	39.7
1dwd−1hxf	1	2	74.5,−109.8	57.9,−105.3	89.2,−67.2
1hyt−1npc	1	2	71.4,31.0	73.7,21.1	84.1,21.4
1pdz−1pdy	1	2	92.8,82.7	86.3,74.5	98.7,93.8
2pk4−1krn	1	2	−77.8,−111.1	−20.0,0.0	−64.7,−23.7
2sim−2sil	1	2	44.7,−2.1	19.0,4.8	73.2,19.8
3mth−6ins	1	2	14.3,0.0	20.0,40.0	−11.9,1.4
3ptb−3ptn	1	2	36.4,−12.1	21.4,−21.4	55.3,−13.9
5p2p−3p2p	1	2	56.9,36.2	42.1,10.5	81.6,45.6
7cpa−5cpa	1	2	79.5,5.1	80.0,10.0	94.6,−1.3
1inc−1esa	1	3	55.2,55.2,41.4	27.3,45.5,9.1	90.3,83.3
1rbp−1brq	1	3	86.5,81.1,10.8	78.6,82.1,3.6	94.4,91.7,3.0
2tmn−1l3f	1	3	84.6,59.0,23.1	64.7,47.1,17.6	97.5,76.7,21.3
2ypi−1ypi	1	3	13.8,−10.3,−106.9	21.4,0.0,−35.7	36.0,−8.4,−195.1
3gch−1chg	1	3	78.1,75.0,50.0	66.7,60.0,53.3	89.5,84.0,63.1
6rsa−7rat	1	3	45.2,25.8,12.9	42.9,35.7,14.3	75.1,36.8,15.0
1mtw−2tga	1	5	82.9,80.0,65.7,37.1,42.9	57.1,57.1,50.0,21.4,21.4	89.5,93.2,79.0,50.4,63.8
1qpe−3lck	2	3	85.0,22.5,−300.0	71.4,0.0,−285.7	95.4,48.0,−522.8

[a] If there is more than one pocket, all of them are considered, if they satisfy the volume threshold. [b] When more than one pocket is present in the bound form, only the larger one is considered. Values for all the pockets in the unbound form are given separated by comma. The formula used is: 100*[(no. for bound) − (no. for unbound)]/(no. for bound).

Although we have used pockets computed using CASTp, the incorporation of other features in our method leads to a better performance as compared to that by the same server for the identification of active site pockets. The significance of the improved ranking in our method has been confirmed from the sign test. Sign test is a nonparametric test that can be used to test the null hypothesis that the signs of + and − are of equal size, or the population means are equal to the sample mean. When we applied the sign test, the critical values (K) at 5% of the significance obtained for the CASTp and our method on the ligand-bound and the unbound forms are 1.9 and 2.1, respectively. On the other hand, "S" values, representing the number of negative signs (the difference in the ranks between CASTp and our method, i.e., the rank in CASTp minus the rank in our method) for the ligand-bound and the unbound forms are 0 and 3, respectively. The null hypothesis is accepted if the value of S is greater than that of K. Thus, there is significant improvement in result in our method for the bound form (which can also be seen from Table 4). Examples of successfully identified binding site clefts are shown in Figure 6. Figure 6a corresponds to the structure of lobster enolase (PDB code 1pdy). Phosphoglycolic acid is represented in van der Waals spheres, and our approach identified the binding cleft shown as a pink mesh around the ligand. Cleft predicted at rank one by CASTp is shown as a blue mesh representation. Another example

Table 4. Performance of Our Method in Comparison with Others on PocketPicker Data Set[a]

	bound		unbound	
algorithm	rank 1	rank 3	rank 1	rank 3
our method	73	94	73	90
SplitPocket	95		90	
Fpocket	83	92	69	94
PocketPicker	72	85	69	85
LIGSITE(CS)	69	87	60	77
LIGSITE	69	87	58	75
CASTp	67	96	71	85
PASS	63	81	60	71
SURFNET	54	78	52	75
LIGSITE(CSC)	79		71	

[a] Rank is defined in Table 2. Results for CASTp were obtained using the server; those for SplitPocket are taken from Tseng et al.,[30] and those for the other algorithms are taken from Fpocket.[28]

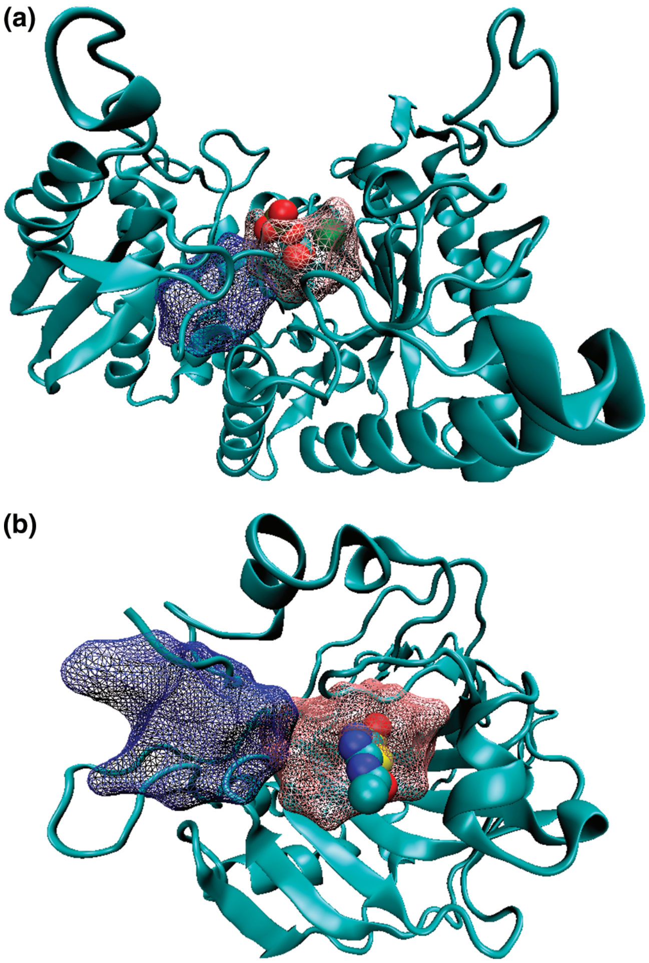

Figure 6. Some examples of successfully predicted binding site clefts: (a) carbonic anhydrase (PDB code 2cba) and (b) lobster enolase (1pdy). Ligand is represented in van der Waals spheres. Catalytic site cleft is represented in a pink maze around the ligand, and cleft predicted at rank one by CASTp[49] is shown in a blue maze.

involves carbonic anhydrase (Figure 6b). In both cases, the binding site cleft is ranked second on the basis of volume, but is the top solution using the criterion of entropy or the distance from the protein centroid.

CONCLUSION

We have compiled a diverse data set of enzymes and analyzed the geometric and the sequence conservation features of the active sites for use in the prediction of active site clefts from their three-dimensional structures. Our results show that the distances of the clefts from protein centroid and sequence entropy, when used in conjunction with the volume, are valuable descriptors for the active site. We have applied the SVM approach for recognizing the active site clefts and tested its performance using different combinations of attributes. Although the data set for training the SVM approach is rather small in size, the results are still encouraging for the method to be used as complementary to other existing tools.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their comments on the manuscript. Discussion with Dr. V. K. Jayaraman on the use of SVM is gratefully acknowledged. S.S. is supported by the Department of Biotechnology, and P.C. is supported by the JC Bose National Fellowship of the Department of Science and Technology.

Supporting Information Available: Additional tables: Tables S1 (details of 58 enzymes that constitute the training set), S2 (test data set used for evaluating SVM performance), S3 (active site residues in the test set), S4 (ranking of attributes), and S5 (performance of the top four attributes applied individually). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Pal, D.; Eisenberg, D. Inference of protein function from protein structure. *Structure* **2005**, *13*, 121–130.

(2) Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **2002**, *324*, 105–121.

(3) Freilich, S.; Spriggs, R. V.; George, R. A.; Al-Lazikani, B.; Swindells, M.; Thornton, J. M. The complement of enzymatic sets in different species. *J. Mol. Biol.* **2005**, *349*, 745–763.

(4) Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **2004**, *32*, D129–133.

(5) Campagna-Slater, V.; Arrowsmith, A. G.; Zhao, Y.; Schapira, M. Pharmacophore screening of the protein data bank for specific binding site chemistry. *J. Chem. Inf. Model.* **2010**, *50*, 358–367.

(6) Coleman, R. G.; Sharp, K. A. Protein pockets: inventory, shape, and comparison. *J. Chem. Inf. Model.* **2010**, *50*, 589–603.

(7) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.

(8) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(9) Tendulkar, A. V.; Wangikar, P. P.; Sohoni, M. A.; Samant, V. V.; Mone, C. Y. Parameterization and classification of the protein universe via geometric techniques. *J. Mol. Biol.* **2003**, *334*, 157–172.

(10) Wangikar, P. P.; Tendulkar, A. V.; Ramya, S.; Mali, D. N.; Sarawagi, S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.* **2003**, *326*, 955–978.

(11) Konc, J.; Janezic, D. Protein-protein binding-sites prediction by protein surface structure conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940–944.

(12) Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168.

(13) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* **2007**, *5*, 43.

(14) Coleman, R. G.; Sharp, K. A. Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J. Mol. Biol.* **2006**, *362*, 441–458.

(15) Aloy, P.; Querol, E.; Aviles, F. X.; Sternberg, M. J. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **2001**, *311*, 395–408.

(16) Armon, A.; Graur, D.; Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **2001**, *307*, 447–463.

(17) George, R. A.; Spriggs, R. V.; Bartlett, G. J.; Gutteridge, A.; MacArthur, M. W.; Porter, C. T.; Al-Lazikani, B.; Thornton, J. M.; Swindells, M. B. Effective function annotation through catalytic residue conservation. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12299–12304.

(18) Innis, C. A.; Anand, A. P.; Sowdhamini, R. Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.* **2004**, *337*, 1053–1068.

(19) Landgraf, R.; Xenarios, I.; Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **2001**, *307*, 1487–1502.

(20) Lichtarge, O.; Bourne, H. R.; Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **1996**, *257*, 342–358.

(21) Panchenko, A. R.; Kondrashov, F.; Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **2004**, *13*, 884–892.

(22) Zvelebil, M. J.; Sternberg, M. J. Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* **1988**, *2*, 127–138.

(23) Tseng, Y. Y.; Dundas, J.; Liang, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.* **2009**, *387*, 451–464.

(24) Lichtarge, O.; Sowa, M. E.; Philippi, A. Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol.* **2002**, *344*, 536–556.

(25) Madabushi, S.; Yao, H.; Marsh, M.; Kristensen, D. M.; Philippi, A.; Sowa, M. E.; Lichtarge, O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **2002**, *316*, 139–154.

(26) Yao, H.; Kristensen, D. M.; Mihalek, I.; Sowa, M. E.; Shaw, C.; Kimmel, M.; Kavraki, L.; Lichtarge, O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **2003**, *326*, 255–261.

(27) Ben-Shimon, A.; Eisenstein, M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J. Mol. Biol.* **2005**, *351*, 309–326.

(28) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(29) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897.

(30) Tseng, Y. Y.; Li, W. H. Identification of protein functional surfaces by the concept of a split pocket. *Proteins* **2009**, *76*, 959–976.

(31) Ondrechen, M. J.; Clifton, J. G.; Ringe, D. THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12473–12478.

(32) Ko, J.; Murga, L. F.; Andre, P.; Yang, H.; Ondrechen, M. J.; Williams, R. J.; Agunwamba, A.; Budil, D. E. Statistical criteria for the identification of protein active sites using Theoretical Microscopic Titration Curves. *Proteins* **2005**, *59*, 183–195.

(33) Wei, Y.; Ko, J.; Murga, L. F.; Ondrechen, M. J. Selective prediction of interaction sites in protein structures with THEMATICS. *BMC Bioinf.* **2007**, *8*, 119.

(34) Elcock, A. H. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **2001**, *312*, 885–896.

(35) Ota, M.; Kinoshita, K.; Nishikawa, K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **2003**, *327*, 1053–1064.

(36) Wallach, I.; Lilien, R. H. Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation. *Bioinformatics* **2009**, *25*, i296–304.

(37) Li, N.; Sun, Z.; Jiang, F. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinf.* **2008**, *9*, 553.

(38) Petrova, N. V.; Wu, C. H. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinf.* **2006**, *7*, 312.

(39) Kleywegt, G. J.; Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 178–185.

(40) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709–713.

(41) Connolly, M. L. The molecular surface package. *J. Mol. Graphics* **1993**, *11*, 139–141.

(42) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. Analytical shape computation of macromolecules: II. *Proteins* **1998**, *33*, 18–29.

(43) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins* **1998**, *33*, 1–17.

(44) Chakravarty, S.; Bhinge, A.; Varadarajan, R. A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J. Biol. Chem.* **2002**, *277*, 31345–31353.

(45) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(46) Wang, G.; Dunbrack, R. L., Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **2005**, *33*, W94–98.

(47) Bielka, H.; Dixon, H. B. F.; Karlson, P.; LiebCcq, C.; Sharon, N.; Van Lenten, E.; Velick, S. F.; Vliegenthart, J. F. G.; Webb, E. C. *Enzyme Nomenclature*; Academic Press, Inc.: London, UK,1992.

(48) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(49) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* **2003**, *31*, 3352–3355.

(50) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9*, 56–68.

(51) Guharoy, M.; Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15447–15452.

(52) Sonavane, S.; Chakrabarti, P. Cavities and atomic packing in protein structures and interfaces. *PLoS Comput. Biol.* **2008**, *4*, e1000188.

(53) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

(54) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

(55) Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **2004**, *20*, 2479–2481.

(56) Gewehr, J. E.; Szugat, M.; Zimmer, R. BioWeka--extending the Weka framework for bioinformatics. *Bioinformatics* **2007**, *23*, 651–653.

(57) Chang, C. C.; Lin, C. J. *LIBSVM: a library for support vector machines, Version-2.84 Publisher*, 2001; software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(58) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5*, 2438–2452.

CI1002922