BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

LFM-Pro: A Tool for Detecting Significant Local Structural Sites in Proteins

Ahmet Sacan[a,*,†], Ozgur Ozturk[b], Hakan Ferhatosmanoglu[b,c,‡], Yusu Wang[b]

[a]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey,
[b]Computer Science and Engineering Department, The Ohio State University, Columbus, OH
[c]Biomedical Informatics Department, The Ohio State University, Columbus, OH

ABSTRACT

Motivation: The rapidly growing protein structure repositories have opened up new opportunities for discovery and analysis of functional and evolutionary relationships among proteins. Detecting conserved structural sites that are unique to a protein family is of great value in identification of functionally important atoms and residues. Currently available methods are computationally expensive and fail to detect biologically significant local features.

Results: We propose *LFM-Pro* (*Local Feature Mining in Proteins*) as a framework for automatically discovering family specific local sites and the features associated with these sites. Our method uses the distance field to backbone atoms to detect geometrically significant structural centers of the protein. A feature vector is generated from the geometrical and biochemical environment around these centers. These features are then scored using a statistical measure, for their ability to distinguish a family of proteins from a background set of unrelated proteins, and successful features are combined into a representative set for the protein family. The utility and success of LFM-Pro are demonstrated on Trypsin-like Serine Proteases family of proteins and on a challenging classification dataset via comparison with DALI. The results verify that our method is successful both in identifying the distinctive sites of a given family of proteins, and in classifying proteins using the extracted features.

Availability: The software and the datasets are freely available for academic research use at http://bioinfo.ceng.metu.edu.tr/Pub/LFMPro

Contact: ahmet@ceng.metu.edu.tr, {ozturk,hakan,yusu}@cse.ohio-state.edu

1 INTRODUCTION

Rapidly growing protein structure repositories open up new possibilities for discovering functional and evolutionary relationships among proteins, and for elucidating the principles by which a certain structure produces an observed function. The increase in data size, however, also calls for more efficient and accurate methods of comparing proteins and identifying potential functional residues and binding sites.

The classical approaches of structural analysis have focused on global pairwise structural alignment of proteins to detect similarities and help transfer of information about a well-known protein to unknown proteins that can be structurally aligned to it. The structural alignment methods, however, are computationally intensive and do not lend themselves to large-scale comparisons. Moreover, they miss remote homologies, especially when the proteins share only a local region.

Many proteins have a multi-domain nature, and the global similarities alone are not sufficient to identify functional similarities existing in distinct local domains. Inevitably, *local structural motifs* are often required for identification of biological function and homology relationships (Hodgman, 1989; Taylor and Jones, 1991). Manual identification of these regions require intensive genetic and molecular biology experimentation, which may take years of diligent studies. An automated method of detecting potential sites would thus be very much appreciated. We therefore focus, in this study, on automatic discovery of local sites of proteins which have distinguished structural and biochemical features, and may thereby have functional significance.

Previous approaches have assumed that such functional sites are already known (Bagley and Altman, 1995; Wallace et al., 1996), and have focused on building a *description*, rather than *automatic detection* of these sites, with the hope of cataloguing these descriptions as structural motifs, so that unknown proteins could be annotated via comparison with these motifs. The *Local Feature Mining in Proteins* (*LFM-Pro*) framework proposed in this study starts with a group of proteins that share a certain function, and does not assume any prior knowledge about the location or nature of the functional sites. Through comparison of this group of proteins with a background set of unrelated proteins, it is able to detect sites that yield features unique to the family members.

Structural motif search is generally based on graph theoretical algorithms (Spriggs et al., 2003; Huan et al., 2005), geometric hashing (Wallace et al., 1997; Shatsky et al., 2005) and others (Singh and Saha, 2003). In order to discover motifs, these methods search for commonly recurrent local structures in space, based on their specific models. The graph theoretic approaches generally require exponential time in the number of the localities being matched. The computational bottle-neck of these approaches prevent effective automated detection of local motifs. More importantly, these methods analyze the protein at the *residue level*, and fail to handle substitutions of the amino-acids or displacements of the backbone. It has been shown that residues can adopt quite different conformations while managing to conserve the positions of their important

*To whom correspondence should be addressed

†This study was conducted while the author was a Visiting Scholar at The Ohio State University

‡This research was partially supported by US Department of Energy (DOE) Early Career Principal Investigator Award DE-FG02-03ER25573 and US National Science Foundation CAREER Award IIS-0546713

functional atoms (Wallace et al., 1996). Therefore, an *efficient* method that can analyze the protein structures at the finer granularity of *atomic level* is needed.

2 APPROACH

We focus on identification of local sites which are unique to a family of proteins sharing a certain structural or functional property. A site can be defined as a three dimensional location in the protein, and a local spatial neighborhood around this location having a certain structure or function (Bagley and Altman, 1995). In order to mine a protein dataset for possible functional sites, we are faced with three main challenges.

The first challenge is deciding on a data structure for sampling of the 3D distributions of the site locations and determining the size of their spatial neighborhood. For this purpose, a three dimensional grid has previously been utilized (Goodford, 1985; Bagley and Altman, 1995). Although grids offer computational advantages, the protein space has to be sampled in high resolution in order to capture micro-environments, which causes very large grids, defeating the purpose of using a grid-based distribution. Some methods therefore only consider local patterns centered at each residue or at some manually-chosen positions as potential motifs (Jonassen et al., 2001; Liang et al., 2003), possibly missing motifs not centered around such positions. Furthermore, these methods usually miss relatively rare and novel motifs. An automatic method that produces a concise yet complete coverage of the motif space is still missing. The method we present in this paper is able to efficiently sample the motif space for identification of unique structural and functional local motifs. Our method relies on a novel computational geometry method for identification of topologically significant locations and also dynamically adjusts the size of the site based on the residues surrounding the microenvironment.

The second challenge is the characterization of the microenvironment features. Presence of certain amino-acid types as the basic feature (Wako and Yamato, 1998; Singh et al., 1996; Munson and Singh, 1997) does not provide a detailed characterization of the site, and may miss certain motifs because of the similarity and substitutability of amino acids. More detailed characterization of the microenvironment (Bagley and Altman, 1995) consider properties such as hydrophobicity, mobility, and solvent accessibility which can capture the physico-chemical nature of the site at the cost of requiring more time for the computation of these properties. We have found that using the atom frequencies (Li and Parthasarathy, 2001; Milik et al., 2003) is a good tradeoff between accuracy and efficiency in characterizing the microenvironment for the purpose of local motif detection. Moreover, unlike previous studies, we also augment the feature vector to capture the topological information of the backbone surrounding the microenvironment.

The last main challenge is having an efficient and sensitive method for detecting common patterns. Determining which motifs are responsible for an observed function is a difficult task. Graph theoretic approaches try to find common subgraphs, but they are currently not scalable for large space of possible motifs, and they cannot easily handle noise in the data or substitution of residues. Statistical methods have been used (Bagley and Altman, 1995) in characterization of the motif structure while comparing a group of known sites and non-sites, but these methods rely on *a priori* knowledge of the functional sites. Whereas, the method we present uses a data mining approach to discover distinguishing functional sites shared by a family of proteins without requiring prior knowledge of the location or nature of these sites. Moreover, it is robust to noisy patterns, and can handle incorrect initial classification of the data.

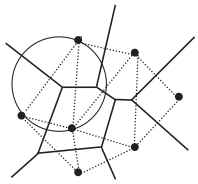

Fig. 2. Delaunay tessellation (dashed lines) and Voronoi diagram (solid lines) of a set of points in 2D. Region enclosed by a Voronoi polyhedron is the area that is closest to the enclosed point than to any other point in the set. Delaunay tesselation is obtained by connecting points that share a boundary. In 3D, Delaunay tessellation would give space-filling tetrahedra. A circle (sphere) can be drawn whose center is a vertex of Voronoi diagram and which passes through the points in the corresponding Delaunay triangle (tetrahedra).

3 METHODS

Figure 1 shows an overall flow-chart of the steps followed in LFM-Pro. We first identify topologically significant local structural centers of each protein, by calculating the critical points of a particular distance field. A ball centered around each critical point defines the spatial neighborhood of these structural centers. Each critical point is then associated with topological and biochemical features of its spatial environment.

Once we generate the feature vectors for each critical point of the proteins, a family of proteins are then searched for shared feature vectors. The aim here is to find critical points unique to a family; therefore, a set of shared feature vectors are chosen such that it is able to distinguish the members of the protein family from a background set of proteins that lack the properties and functions of interest. The group of critical points that are unique to a family are combined to obtain *a representative feature set* for the family. In the following subsections, each of these steps are described in detail.

3.1 Sampling of the Structural Centers

Given a protein P as the set of its alpha Carbon (C_α) atom centers $P = \{p_1, \ldots, p_n\}$, the distance function $\Phi_P : \mathbb{R}^3 \rightarrow \mathbb{R}$ w.r.t. P is defined as follows: $\Phi_P(x)$ is the nearest distance from x to any $p_i \in P$. Φ_P describes the influence of (the backbone atoms of) protein P to its neighboring space via the distance field. Intuitively, if two proteins have similar structure, they should have similar distance fields. In particular, if there are regions in space where proteins display similar local structural patterns, then they should have similar distance fields in and around that region as well.

We identify the potential motif centers by finding the critical points of this distance function. Formally, critical points of a smooth function g, are points with vanishing gradients. In our case, for a function defined over $\mathbb{R}^3$, there are four types of critical points: local minima, local maxima, and two types of saddle points. Note that, when distance to backbone atoms is used as function g, it turns out that the set of critical points of Φ_P is the set of intersection points between some Delaunay simplex (a point, edge, triangle, or tetrahedron) with its dual Voronoi elements (a polytope, face, edge, point, respectively) (Figure 2), and can be computed in $O(n^2)$ time where $n = |P|$ (Giesen and John, 2003).

We now collect Π as the set of critical points of the distance function. Some examples of structural motifs that such critical points can capture are illustrated in Figure 3. The *spatial neighborhood* of a critical point is defined as the spherical region centered at the critical point, whose radius is its distance function value.

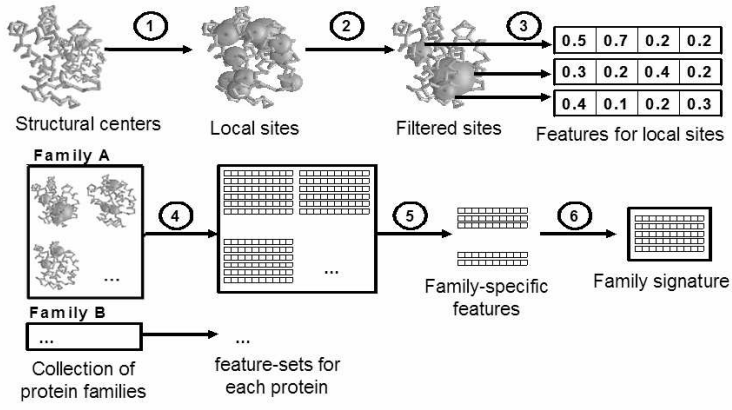

Fig. 1. The general strategy of LFM-Pro. For each protein, 1) location of the critical points of distance field to backbone atoms are identified, 2) the critical points are filtered to remove nonpersistent or unimportant ones, 3) a feature vector that captures the topological and biochemical properties of its spatial neighborhood is associated with each critical point. 4) Feature vectors for the remaining critical points of each protein in the dataset are pooled and 5) those that are generated from family members are assessed for their ability to discriminate the family proteins from the rest of the dataset. 6) the critical points that display the best discriminating behavior in step 5 are combined into a representative feature set of the family.

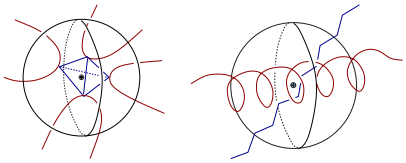

Fig. 3. Two types of motifs captured by critical points of the distance function. In (a), four pieces of protein backbone come close in space, forming a contact as indicated by the tetrahedron in the middle. The double point is a local maximum of Φ. In (b), the cross-point is a saddle point. Local spatial patterns can be captured by taking a ball centered at these critical points.

Following the generation of all critical points of distance, we perform a filtering of these points to eliminate noise. The structural importance of the critical points were assigned using the topological persistence algorithm from (Edelsbrunner et al., 2002), and those with small persistence were removed from Π. This topological method of removing noise is fundamentally different from those that employ clustering of neighboring points, in terms of the type of noise it removes. Roughly speaking, it measures the importance of a feature (critical point) by measuring how persistent this feature remains if the distance field is perturbed. Note that filtering based on persistence effectively eliminates the noise inherent in the crystallography methods used to obtain the atom coordinates. After the filtering step, the number of remaining critical points are roughly the same as the number of the amino acids in the protein.

3.2 Characterizing the Spatial Neighborhood

As a by-product of our structural center sampling method, we have a natural way to decide the neighborhood size, which is better than prefixing some threshold value. For the spatial neighborhood around each critical point, we associate a feature vector, based on both the structural and biochemical nature of the neighborhood. The structural features include: the persistence value of the critical point, the radius of the neighborhood, and the *writhing number*. The biochemical features we use are based on the frequency and location of the constituent atoms within the neighborhood.

The writhing number, or writhe, is originally used to measure the supercoiling phenomenon for a space curve, and has been used to characterize both DNA (Fuller, 1978; Klenin and Langowski, 2000; Swigon et al., 1998) and protein structures (Levitt, 1983; Rogen and Fain, 2003). We compute the writhe of those backbone pieces contained within the spatial neighborhood to measure their relative spatial arrangements.

In order to capture the biochemical nature of the spatial environment, we use the frequencies of each of the side-chain Carbon, Nitrogen, Oxygen, and Sulfur atoms within the spherical region. Furthermore, the location information of these atoms is captured by computing the center of mass for each atom type. Note that our framework can be easily extended to use physicochemical properties such as hydrophobicity, solvent accessibility, Van der Waals radii, or mobility, which can capture more detailed information about the spatial environment (Bagley and Altman, 1995). However, we did not use such extended features in this study, because of the computational cost they incurred.

3.3 Mining for a Representative Feature Set

Each protein p_i now has a set $\Pi = \{c_1, \ldots c_n\}$ of feature vectors generated from its important critical points. Let $F = \{p_1, \ldots p_m\}$ denote a family of proteins that are known to share a common structural or functional property. And let the set G denote the rest of the proteins in the dataset. We wish to determine the critical points that are unique to family F, and assess their ability to discriminate the proteins within the family from the rest of the proteins. Note that the algorithm to detect family-specific critical points has to allow changes in the values of the feature vectors. We utilized a distance-based approach for this purpose.

The dissimilarity $d(c_i, c_j)$ of any given two critical points can be defined in terms of an appropriate distance function between their corresponding feature vectors. We observed that a simple Euclidean distance measure on normalized feature vectors was sufficient in detecting family specific structural centers. A *weighted*–Euclidean distance, that can highlight varying contributions of the individual environment features could also be designed by optimizing the weights against an objective function.

When comparing a critical point c_x to a protein p, we take the distance of c_x to its closest match in p as defined with the distance function:

$$d(c_x, p) = min\{d(c_x, c_1), \ldots, d(c_x, c_n)\}$$

where $c_1, \ldots, c_n$ are the critical points of the protein p. Intuitively, if a critical point c_x is part of a protein p, one would expect a very small value for $d(c_x, p)$.

For each candidate critical point c_x of the proteins in the family F, we calculate its distance to all the proteins in the dataset. For an ideal discriminative critical point, the distances to the proteins in F would be clustered at a minimal, whereas the distances to the rest of the proteins, G, would take upon higher values. We modeled this intuition by defining the *discrimination score s* of a critical point as follows:

$$s(c_x) = \frac{\mu(c_x, G)}{(1 + \mu(c_x, F)) * (1 + \kappa(c_x, F, G))} \quad (1)$$

where $\mu(c_x, F)$ is the average distance of c_x to proteins in the family F,

$$\mu(c_x, F) = avg(d(c_x, p \in F)) \quad (2)$$

and κ is the number of background proteins that have a distance smaller than the maximum within-family distance $d^*(c_x, F) = max(d(c_x, p \in F))$.

$$\kappa(c_x, F, G) = count(d(c_x, p \in G) \leq d^*(c_x, F)) \quad (3)$$

In Equation 1, $\mu(c_x, F)$ and $\mu(c_x, G)$ ensure that those critical points that have small within-family distance and high out-of-family distance get higher discrimination scores. The average distances alone, however, do not guarantee a clear separation of the family proteins from the rest. The term κ favors those critical points that can cluster the family proteins with minimal number of out-of-family proteins. In other words, μ works to select features common to family, while κ works to avoid features that cannot discriminate family proteins from the rest. Each term in the denominator is padded with 1 for numerical stability.

Using the discrimination scores, we obtain a set of critical points ranked by the scores reflecting how representative they are for a given family F. We refer the collection of the critical point features with their associated scores as the *representative feature set* of the family.

Classification Modeling. Let $\Pi = \{c_1, \ldots c_n\}$ be the representative feature set of family F, with corresponding discriminative scores $S = \{s_1, \ldots s_n\}$ and maximum within-family distances $D^* = \{d_1^*, \ldots d_n^*\}$. The *membership score* of a new protein p to the family F is calculated as follows:

$$\psi(p, F) = \frac{1}{n} \sum_{i=1 \ldots n} s_i \frac{d_i^* - d(c_i, p)}{d(c_i, p)} \quad (4)$$

The membership score ψ, is dominated by the matching features that have small distance and high representative scores. The numerator term in the summation in Equation 4 provides a threshold logic based on the maximum within-family distances d^*. Those features that match the protein with a distance smaller than d^* contribute positively in the membership score, whereas those that have a greater distance are penalized in the scoring. The overall membership score reflects how well a protein matches a representative feature set. In a multi-family classification scheme, the membership score $\psi(p, F)$ can be used to assign the protein p to the closest family.

4 RESULTS

4.1 Experimental Setup

All the experiments were conducted on a single processor *Pentium 4* PC with 2.8 GHz CPU and 1 GB main memory. The selection of centers via determination of critical centers of the distance function was implemented in Python and C, using CGAL (CGAL, 2006) computational geometry library; the feature extraction and mining methods were developed under Matlab environment.

The proteins used in this study were selected from the representative ASTRAL (Brenner et al., 2000) dataset of SCOP 1.69 (Murzin et al., 1995) with less than 40% sequence homology. There were a total of 7,237 entries in the ASTRAL dataset.

The one-time-only generation of critical points and their corresponding feature vectors took 49 seconds on the average per protein.

4.2 Mining Functional Sites

The success of LFM-Pro could be assessed by applying it to protein families that have well-defined functional sites, and investigating whether the sites detected by LFM-Pro match the known functional sites in these proteins. Serine Proteases are the most studied family of proteins, in the context of structural motif extraction (Bagley and Altman, 1995; Wallace et al., 1996; Milik et al., 2003; Huan et al., 2004, 2005). We follow the tradition and also use Serine Proteases for this study. The proteins were selected from the SCOP superfamily (b.47.1.*) "trypsin-like serine proteases," here on referred as the **SP** superfamily and included both prokaryotic (PSP: 10 SCOP entries) and eukaryotic (ESP: 19 SCOP entries) proteins, which share the same catalytic site.

The local site mining for the SP family took 30 seconds to complete. Note that, with the same number of localities to compare, the subgraph mining methods may take several days to complete (Huan et al., 2005). Figure 4 shows the mapping of the top scoring features on Alpha-lytic protein (1ssx). The top sites obtained by the feature mining algorithm corresponded to the catalytic triad site of the Serine Proteases. The atoms within the immediate neighborhood of the catalytic triad have relatively conserved positions, which is successfully picked up by the mining algorithm. The highest scoring site contained atoms of the residues Ser195, His57, Asp102, Ser214 and Ala55. The residues Ser195-His57-Asp102 form the charge relay system responsible for the hydrolytic cleavage of the appropriate substrate. Ser214 has also been found to be highly conserved in SP (Wallace et al., 1996). We also observed that Ala55 is conserved in SP and we speculate that Ala55 keeps the catalytic triad in its relative orientation via Van der Waals interactions.

The third highest scoring site includes the disulfide bridge Cys189-Cys220, which is distant to the catalytic site, but is nevertheless conserved across Serine Proteases. This disulfide bond keeps the backbone such that Ser195 and Ser214 can remain in close proximity. The next highest scoring site is another disulfide bridge, Cys42-Cys58, which helps keep the His57 and Ala55 residues within the catalytic site 4.

Selection of Background Proteins. One interesting question is whether the use of a background set of proteins is really necessary, i.e., whether it would be possible to detect the functional sites by just finding features common to a family of proteins, without comparison to unrelated proteins. Figure 5 illustrates the effect of the size and nature of the background class of proteins on the detection of functional site in SP. The rank of the first feature that map to the catalytic triad site is used as the basis of evaluation.

We expected that the performance of the algorithm would improve with increasing number of out-family proteins used. As the size of the background set is increased, the contribution of $\mu(c_x, F)$ term in Equation 1 decreases, which translates into *distinguishing* features ranking higher than *common* features. Figure 5 shows that for each type of background set of proteins we used, the algorithm was able to detect the functional site, when given a sufficiently large number of background proteins.

Furthermore, Figure 5 demonstrates that using proteins that share structural features with the family under investigation increases the

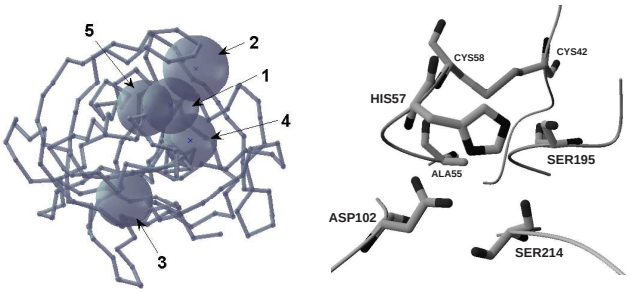

Fig. 4. Mapping of the top scoring sites onto Alpha-lytic protein (1ssx). The features were obtained by mining SP dataset against a random set of 200 background proteins. **Left:** Features 1,2,4,5 span the neighborhood of the catalytic triad, whereas feature 3 contains a distant disulfide bridge CYS189-CYS220. **Right:** A closer look into the catalytic region spanned by features 1,2,4,5. The residues whose side-chain atoms are contained within these sites are shown.

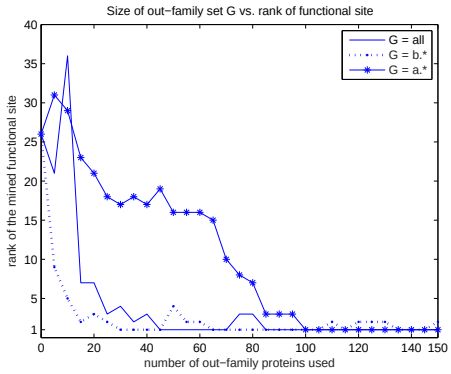

Fig. 5. The effect of the size of the background set G on detection of the functional site. Results given for mining SP dataset against selection of proteins using three sets of proteins: all proteins, only b.* all-beta class, or only a.* all-alpha class. The size of G is shown up to 150 proteins for illustration purposes; the rank of the mined functional site did not change beyond 150 proteins.

accuracy of the mining. When random out-family members were selected from b.* SCOP class of all-beta proteins, the functional triad site is detected among the top-scoring sites, even with only a few out-family proteins. Whereas, significantly more proteins are needed in the out-family set if one uses a.* SCOP class of all-alpha proteins, which share little structural fold similarity with SP. This observation is attributed to the fact that proteins that share structural folds with the investigated family can better prune out insignificant scaffold sites and enhance detection of unique sites.

The set of background proteins needed to obtain the most desirable feature-mining results would depend on the specific family being studied. Even though all available proteins can be used as the background set G, it may be desirable to reduce the size of G for efficiency purposes. As a general guideline, we recommend the use of proteins that share the same structural folds, but are missing the target function of interest.

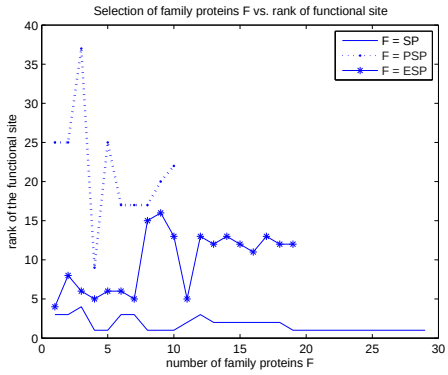

Fig. 6. The effect of the size and composition of the family set F on detection of the functional site. The background set G for this experiment is composed of 200 randomly selected proteins from the b.* SCOP class of all-beta proteins.

Selection of Family Proteins. While seeking features that are distinguishing from unrelated proteins, we also seek that these features be common across the family. For this reason, appropriate selection of the family proteins plays an important role in detection of functional sites. Figure 6 demonstrates the effect of composition and size of the family proteins on detection of the catalytic triad. The region of the catalytic triad is more conserved in Eukaryotic proteins, giving the functional site a higher score. When PSP and ESP proteins are combined (SP), the family set would contain an evolutionarily more diverse set and the algorithm can attribute lower scores to those sites that are unique only to either of these two families, and highlight the functional site that is shared by both protein families.

Appropriate composition of the family proteins was more effective in mining for the functional site than simply increasing the size of the family. In fact, increasing the number of proteins did not give the catalytic triad significantly higher scores in PSP or ESP families. For PSP and ESP families, the high scoring features involved the sites that represent the hydrophobic cores and loops in the secondary structure. These spatial regions show greater variation across proteins, and are detected as representative of the family when a smaller family set is used.

4.3 Binary Classification

To investigate the classification capabilities of LFM-Pro, we used a dataset that was previously utilized under a binary classification scheme (Huan et al., 2004). The first dataset (C_1) includes two families from different SCOP classes: nuclear receptor ligand-binding domain proteins (NB, 16 proteins) from all-alpha class, and the prokaryotic serine protease family (PSP, 10 proteins) from all-beta class. The second dataset (C_2) uses ESP (19 proteins) and PSP families which belong to the same superfamily. Note that PSP and ESP were used together above in the functional-site mining experiments. Whereas, the goal in this section is to evaluate the discrimination power of the representative feature sets for clearly distinct families (C_1) and closely related families (C_2). The proteins were selected from the *Culled-PDB* list (Wang and Dunbrack, 2003) with less than 60% identity.

Table 1. Binary classification results.

Dataset	Method	Features	Dist.Feat	Accuracy
C_1	DT	20,646	934	100%
	AD	23,130–37,394	1,093–1,674	96–100 %
	LFM-Pro	5,282	1	100%
C_2	DT	15,895	20	95%
	AD	18,491–32,569	29–36	93–95 %
	LFM-Pro	2,180	139	100%

The methods Delaunay Tesselation (DT) and Almost Delaunay (AD) are from subgraph mining approach in (Huan et al., 2004); results for the AD entry are given for a range of allowable perturbation values ($\epsilon = 0.1 - 0.75$). The fourth column shows the number of features that have *discrimination power* above 0.75, as defined by the authors; and the number of features required to obtain maximum accuracy in LFM-Pro. Accuracy is defined as the fraction of correct predictions measured by five-fold cross validation.

For families in datasets C_1 and C_2, the feature sets were extracted and scored as described above, and these representative feature sets were used for binary classification of proteins. The subgraph mining approach in (Huan et al., 2004) have achieved perfect accuracy for C_1 dataset, where the two families are from different SCOP classes, but had 5% classification error for the C_2 dataset, in which the two families belong to the same superfamily. LFM-Pro classifies the proteins in both of these datasets with 100% accuracy, when all the extracted features were used in classification (Table 1). We attribute the success of LFM-Pro, in comparison with the graph mining approaches, to the fact that it can accommodate amino acid substitutions and displacements in the backbone, and focuses on the individual atoms within a spatial neighborhood rather than the coarser level information about location of CA atom of the amino acid residues.

In LFM-Pro, each feature in the representative feature set contributes according to its corresponding score, which guarantees that the features that are not as discriminative as the top scoring features do not distort the classification, but only fine-tune it. However, it may be desirable for efficiency and maintenance purposes, to keep only a small fraction of the top-scoring features for classification. Even though perfect accuracy was achieved in C_1 dataset using a single feature; the classification was more stable when more than 20 features are used. Considerably more features were required to distinguish the closely related families in the C_2 dataset.

4.4 Multi-class Classification

In order to further validate our method, we performed a multi-class classification experiment on a more challenging dataset. Namely, the new entries introduced in SCOP 1.69 were classified based on family representations generated from SCOP 1.67. For both SCOP versions, ASTRAL dataset with less than 40% were used. The proteins or families that were re-classified in 1.69 and families that contain a training set less than 5 members were ignored. The final dataset contained 90 families with a total of 1,056 training proteins from SCOP 1.67 and 157 test proteins that were newly added in SCOP 1.69.

For comparison, the test proteins were also classified based on pairwise DALI (Holm and Sander, 1993) scores, such that a query protein is assigned to the family of the protein with highest pairwise Z score. The results of multi-class classification experiment are tabulated in Table 1. The restriction of 40% homology in the dataset makes it particularly challenging. Moreover, an increase in the number of families result in higher number of false positives . DALI could only classify 31.2% of the test proteins correctly, whereas LFMPro obtained a classification accuracy of 37.58%.

Table 2. Multi-class classification results.

Method	Training Accuracy	Test Accuracy
DALI	100%	31.21%
LFMPro	100%	37.58%
DALI and LFMPro	100%	56.05%

The training set is from SCOP 1.67 and test set is the newly added proteins in SCOP 1.69. The last row assumes that an oracle chooses the correct classification given by either method.

Note that the proteins classified correctly by LFMPro are disjoint from those classified correctly by DALI. Combining DALI and LFMPro results and assuming an oracle to decide which one to use for a give protein, 56.05% accuracy is possible. Therefore, a classifier combining the output of these complementary methods would achieve higher accuracy, which is among our future research goals.

5 DISCUSSION

We have presented a data-mining based framework, *Local Feature Mining in Proteins* (*LFM-Pro*), whereby topologically and biochemically conserved regions of a protein family could be automatically discovered. We have demonstrated the success of the method on Serine Protease family of proteins and also on two binary classification datasets. The sites unique to a family of proteins were identified via comparison to a background set of proteins. We have confirmed that the sites detected by our method conforms with the previously reported functional sites. When a background set of proteins is not provided, LFM-Pro scores the local sites based on how common they are across the family proteins.

LFM-Pro gives the most desirable site-mining results when the family being studied contains proteins that are evolutionarily distant but share the site of interest, and when the background family is chosen to contain proteins that share the same structural folds with the family being studied. The objective of maximizing the discriminative scores can be used to determine the optimal size of the background set in feature mining, and the optimal number of features in classification.

LFM-Pro uses feature vectors associated with local neighborhoods that provides comprehensive sampling of the protein space. One of the major advantages of a feature-based approach is the computational efficiency; because the time-consuming graph matching or structural alignment steps are no longer required. Moreover, the feature vectors can be stored in an index structure optimized for range queries, which would further improve the efficiency of the algorithm. A custom filtering step to remove features related to trivial secondary structures can also be performed to reduce the number of candidate features, which would further increase the efficiency of the algorithm.

The framework presented in this study is easily extensible to more sophisticated feature extraction and scoring schemes. One may, for example, augment the features presented here with physicochemical features such as hydrophobicity, solvent accessibility, or mobility. It would also be interesting to investigate critical points of other function fields, such as force fields. Note that we utilized a simple unweighted Euclidean distance function for measuring the dissimilarity between feature vectors, and it was our experience that the algorithm allowed imperfect distance functions. However, fine-tuning the weights of the spatial features may be desirable in order to highlight the contributions of each feature in the representation of local sites. The weights of the distance function can be automatically optimized with the objective of maximizing the discriminative scores of the representative set. We have provided in the software distribution of LFM-Pro, a *simulated annealing* approach for such fine-tuning.

Using *local* structural and biochemical features as opposed to structural alignment of proteins, can potentially yield in identification of very distant evolutionary relationships, and can help discern the function of yet uncharacterized proteins. Local sites of the proteins resist evolutionary modifications if they perform an important biological function, whereas the rest of the protein simply provides a scaffold and is more prone to modifications through mutation, insertion, deletion, and duplication events. Therefore, related proteins can share a common evolutionary ancestry or a common biological function, which may only be identifiable through comparison of these local sites.

Inference of remote homology is also a key step in evolutionary-based cataloguing of all available protein structures. Assigning a new protein to unique positions in the classification scheme becomes impossible when the homology is not detectable. Using LFM-Pro, it is possible to identify a distinguishing representative feature set for each family, and to quickly assign a new protein to one (or more, for multi-domain proteins) of these families. For instance, using the representative feature set generated by LFM-Pro for Globins family of proteins, we were able to discover proteins 1uby, 1gai, and 1xis to have similar distinctive sites as the Globins. These three proteins were not previously classified to have structural or functional similarities with Globins; however, a multiple alignment revealed that they could indeed be significantly aligned with Globins, confirming the detection by LFM-Pro.

Effective discovery of functional local motifs would have tremendous impact in bioscience research, and would find applications in areas such as multiple structural alignment, protein modeling, drug design and targeting. As a future work, we plan to undertake a large-scale, systematic study where we would extract representative feature sets for all SCOP families, and provide them as a publicly available motif database. The feature vectors extracted from the proteins also lend themselves for an unsupervised learning method where unique functional sites could be automatically discovered without any prior family-membership information.

REFERENCES

Bagley, S. C. and Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci*, 4:622635.

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The astral compendium for sequence and structure analysis. *Nucleic Acids Research*, 28:254256.

CGAL (2006). The cgal project-release 3.1.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533.

Fuller, F. B. (1978). Decomposition of the linking number of a closed ribbon: a problem from molecular biology. In *Proc. Natl. Acad. Sci. USA*, volume 75, pages 3557–3561.

Giesen, J. and John, M. (2003). The flow complex: A data structure for geometric modeling. *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 285–294.

Goodford, P. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, 28:849–857.

Hodgman, T. C. (1989). The elucidation of protein function by sequence motif analysis. *Computer Appl. in the Biosci. (CABIOS)*, 5:1–13.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138.

Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A. (2005). Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology*, 12:6:657–71.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., and Tropsha, A. (2004). Mining family specific residue packing patterns from protein structure graphs. *Proc. of 8th Ann. Intl. Conf. on Research in Comp. Molecular Bio. (RECOMB)*, pages 308–15.

Jonassen, I., Eidhammer, I., Conklin, D., and Taylor, W. R. (2001). Structure motif discovery and mining the PDB. *Bioinformatics*, 18(2):362–367.

Klenin, K. and Langowski, J. (2000). Computation of writhe in modeling of supercoiled DNA. *Biopolymers*, 54:307 – 317.

Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, 170:723–764.

Li, H. and Parthasarathy, S. (2001). Automatically deriving multi-level protein structures through data mining. In *HiPC Workshop on Bioinformatics and Computational Biology*.

Liang, M. P., Banatao, D. R., Klein, T. E., and Brutlag, D. L. (2003). Webfeature: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res*, 31:3324–3327.

Milik, M., Szalma, S., and Olszewski, K. (2003). Common structural cliques: a tool for protein structure and function analysis. *Protein Engineering*, 16:8:543–52.

Munson, P. and Singh, R. (1997). Statistical significance of hierarchical multi-body potentials based on delaunay tessellation and their application in sequence-structure alignment. *Protein Sci.*, 6:14671481.

Murzin, A. G., Brenner, S. E., 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.

Rogen, P. and Fain, B. (2003). Automatic classification of protein structure by using Gauss integrals. *Proc Natl Acad Sci U S A*, 100(1):119–124.

Shatsky, M., Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2005). Recognition of binding patterns common to a set of protein structure. *Lecture Notes in Computer Science*, 3500:440 – 455.

Singh, R. and Saha, M. (2003). Identifying structural motifs in proteins. In *Pac Symp Biocomput*, pages 228–239.

Singh, R., Tropsha, A., and Vaisman, I. (1996). Delaunay tessellation of proteins. *J. Comput. Biol.*, 3:213222.

Spriggs, R. V., Argymiuk, P. J., and Willett, P. (2003). Searching for patterns of amino acids in 3D protein structures. *J Chem Inf Comput Sci*, 43(2):412–421.

Swigon, D., Coleman, B. D., and Tobias, I. (1998). The elastic rod model for DNA and its application to the tertiary structure of dna minicircles in mononucleosomes. *Biophysical Journal*, 74:2515–2530.

Taylor, W. and Jones, D. (1991). Templates, consensus patterns and motifs. *Current opinion in structural biology*, 1:327–323.

Wako, H. and Yamato, T. (1998). Novel method to detect a motif of local structures in different protein conformations. *Protein Engineering*, 11:981–990.

Wallace, A., Laskowski, R., and Thornton, J. (1996). Derivation of 3d coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Sci.*, 5:1001–1013.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*, 6:2308–2323.

Wang, G. and Dunbrack, R. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, 19:1589–1591.