New Kernels for Protein Structural Motif Discovery and Function Classification

Chang Wang   CHWANG@CS.UMASS.EDU

Dept. of Computer Science, University of Massachusetts, Amherst, MA 01003, USA

Stephen D. Scott   SSCOTT@CSE.UNL.EDU

Dept. of Computer Science, University of Nebraska, Lincoln, NE 68588-0115, USA

Abstract

We present new, general-purpose kernels for protein structure analysis, and describe how to apply them to structural motif discovery and function classification. Experiments show that our new methods are faster than conventional techniques, are capable of finding structural motifs, and are very effective in function classification. In addition to strong cross-validation results, we found possible new oxidoreductases and cytochrome P450 reductases and a possible new structural motif in cytochrome P450 reductases.

1. Introduction

A goal of structural genomics is to determine proteins' three-dimensional structures from their gene sequences. The challenge, once the structure is determined, is to extract useful biological information about the biochemical and biological role of the protein in the organism. With the rapid expansion in the number of known protein structures, prediction of function based on structure has become one of the major aims of bioinformatics. It provides useful information to biochemical experiments and further improves the performance of genome analysis.

Primary sequence can often be used to infer function. However, some protein functions cannot be identified solely by primary sequence-based methods. In such cases, functional similarities are found from structure comparisons. Many methods, including SSAP (Taylor & Orengo, 1989), DALI (Holm & Sander, 1993), and CE (Shindyalov & Bourne, 1998), have been used for structural comparisons.

There are also methods for predicting function from structure. Many of them compare the structure of a protein with unknown function to the structure of proteins with known function in structural databases, such as CATH (Orengo et al., 1997) and SCOP (Murzin et al., 1995). Other methods, such as SITE (Zhang et al., 1999), FFFs (Fetrow & Skolnick, 1998) and superfamily active site templates (Meng et al., 2004) use structural motif-related information to search for function in an unknown structure.

A structural motif is a conserved sub-structural pattern that is common to a set of proteins sharing similar structures or functions. Most biological actions of proteins depend on structural motifs. Discovery of motifs is a complex process including feature extraction, structure comparison, discovery and evaluation. The feature selection step extracts features to be used for pattern discovery from proteins. Structure comparison is the most difficult step. Many methods have been devised, including pairwise structure alignment using dynamic programming or superposition to minimize RMSD. Other methods, such as geometric hashing (Holm & Sander, 1995) and 3D coordinate templates (Wallace et al., 1996) have also been applied. After structural comparison, patterns matching the input structures are found and evaluated to see whether they are possible structural motifs. Lately, many new methods have been proposed for this problem. For example, SPratt2 (Jonassen et al., 2002) discovers motifs in an unsupervised fashion. Trilogy (Bradley et al., 2002) handles sequence and structure simultaneously and symmetrically in the search process.

We introduce new kernels for three-dimensional structural analysis. Our results have applications in motif discovery and in function classification. As with some other structural methods, we represent a 3D structure as a set of its components in 3D space. We show that

these new methods are sensitive enough to identify some remote structural similarities that are missed by regularly-used approaches.

Our first result is a new method for structural motif discovery. In some cases of motif discovery, the functional motif of a protein can be described by defining the structure's size, shape, etc. But more often, the motif itself is also not completely known, and the researcher has only a more or less rough idea of what to look for (Schmollinger et al., 2004). Thus it is difficult to specify what to look for in advance. Further, often the results of motif discovery are sensitive to the size of the structure (in terms of number of residues) that is specified. If the sought structure size is too small, then one risks missing some of the regulatory patterns in a motif. Conversely, if the structure size is set too large, the motif will likely include some irrelevant parts.

Our approach is different from other methods, in that we do not seek conserved fragments or commonly used geometrically-defined cells. We assume that a simple function is mediated primarily by one amino acid. Thus we focus on identifying small conserved substructures, each centered on a single amino acid. We define the size of the substructure as a fixed-radius ball in 3D space rather than as a fixed number of residues. We use our new kernel[1] $K_{Pattern\_Sim}$ to measure similarity between pairs of substructures. To avoid missing candidate motifs, we examine the substructure centered at each residue. The highly conserved substructures are candidate motifs.

In our second result, we tune $K_{Pattern\_Sim}$ for application to redox function prediction. Here we leverage known information about the superfamily of thiol/disulfide oxidoreductases. Most oxidoreductases have a CxxC primary sequence motif[2] at their active site. We use this to tune $K_{Pattern\_Sim}$ to oxidoreductases, resulting in a new kernel K_{Redox_Func}. Each substructure we consider consists of all residues that lie in a fixed-radius ball in 3D space. The residue at the center of the ball is called the *central amino acid* and the other residues in the ball are called the *outer amino acids*. For thiol/disulfide oxidoreductases, both the Cs in each CxxC motif are seen as central residues. The outer residues include the residues between two Cs and other amino acids in a fixed-radius ball centered on each C. K_{Redox_Func} measures similarity between substructures by comparing the types of the outer amino

acids, the distances from the outer amino acids to the central amino acids, and distances between the two Cs in the motif's center. We compute similarity between two motif structures using these features.

Our final result is another kernel (K_{3Dball}) designed specifically for tertiary structure comparison. We define the similarity between two protein structures S and T as the sum of structural similarities between any two 3D balls of S and T that have similar constituents. It is similar to DALI, CE, etc., in that we make comparisons between entire three-dimensional structures (i.e. ours are entire structure-based methods as opposed to active site-based methods).

In our experiments, we test our methods on structural superfamilies from CATH and two function superfamilies: thiol/disulfide oxidoreductases and cytochrome P450 reductases. For the two function families, many thiol/disulfide oxidoreductases have a thioredoxin (Trx) fold (Martin, 1995). If a 3D structure is known, one can easily determine whether a given protein possesses a fold. However, some proteins without the fold also have redox function, such as PDB-1d4u. Cytochrome P450 reductase is found in the endoplasmic reticulum of most eukaryotic cells and is an integral component of the monooxygenase system transferring electrons from NADPH to cytochrome P450 via FMN and FAD co-factors. Cytochrome P450 reductase may also donate electrons to heme oxygenase, cytochrome b5, and the fatty acid elongation system, and can reduce cytochrome c. For this family, no conserved motif is known.

We show that our kernels are sensitive to the fold in tertiary structure, although they are not designed for fold identification. They also capture similarities in thiol/disulfide oxidoreductases beyond the Trx-fold that are missed by DALI and CE. As a result, they can be used to find new thiol/disulfide oxidoreductases, since some such proteins that do not possess Trx-fold might be missed by traditional methods. We also successfully apply our kernels to P450 reductases, identifying several possible candidates in PDB. Since K_{3Dball} and $K_{Pattern\_Sim}$ do not require any orientation of the 3D structures or any other prior information about the protein families, our methods should be applicable to many protein families.

Our motif discovery method offers two advantages. First, it doesn't require any prior knowledge. Second, it is very sensitive to small motifs and can also find large motifs by combining small motifs that are close to each other in 3D space. Our kernel-based protein function classification methods also have advantages. First, they are simple and very fast: using

---

[1]While a version of $K_{Pattern\_Sim}$ is positive semidefinite, what we use may not be (Section 2). But for clarity, we use "kernel" to refer to all our similarity measures.

[2]Sometimes a serine replaces one cysteine, but for clarity we will refer to it always as the CxxC motif.

$K_{Pattern\_Sim}$, K_{Redox_Func} and K_{3Dball} are each about 100 times faster than DALI and CE, and can quickly search PDB. Second, they are very sensitive while still maintaining low false positive rates.

The rest of this paper is as follows. In Section 2 $K_{Pattern\_Sim}$ is defined. In Section 3 we introduce K_{Redox_Func}. In Section 4 we define K_{3Dball}. Then in Section 5, we describe how we use the above kernels in motif discovery and function prediction. We summarize our experimental results in Section 6, and we conclude in Section 7.

2. $K_{Pattern\_Sim}$ for Motif Discovery

Recall the definition of central and outer amino acids from Section 1. Each amino acid in a protein is the central amino acid of a set of substructures, where the set comes from varying the radius of the ball. The type of the central amino acid, the types of the outer amino acids and the distances from the outer amino acids to the central amino acid are three major features of a substructure. $K_{Pattern\_Sim}$ computes similarity of two substructures based on these features.

(1) $K_{Pattern\_Sim}(S, T)$ — Similarity of two substructures S and T. Here, the similarity equals zero if the central amino acids of S and T are of different types. If S's and T's central amino acids are the same, then we compute the similarity of S and T by summing the 3D similarities between each amino acid of S and its most similar amino acid from T, where similarity between outer amino acids is based on difference in proximity to the central amino acid. We make our measure symmetric by performing the same operation from T to S. The sum of these two values is used for the similarity of two substructures. Formally, $K_{Pattern\_Sim}(S, T) =$

$$
\begin{cases}
\sum\limits_{i=1}^{|S|} AA\_ssim(S[i], T[i']) + \sum\limits_{j=1}^{|T|} AA\_ssim(T[j], S[j']) \\
When \quad S[1].type = T[1].type \\
\\
\quad 0 \quad \text{otherwise}
\end{cases}
$$

where $S[i]$ is the ith amino acid of S. $S[j']$ is the most similar amino acid in S to $T[j]$, where similarity is determined by AA_ssim, i.e.

$$
j' = \operatorname*{argmax}_{j'':S[j''].type=T[j].type} \{AA\_ssim(S[j''], T[j])\}. \quad (1)
$$

$S[1]$ and $T[1]$ are the central amino acids of S and T.

(2) $AA\_ssim(S[i], T[j])$ — similarity of two amino acids $S[i]$ and $T[j]$ in 3D space. Amino acid 3D similarity is defined as follows: if two amino acids are not

of the same type, then the similarity is zero, else the similarity is computed using the following procedure: first we compute the Gaussian RBF value of the distance from $S[i]$ to $S[1]$ and the distance from $T[j]$ to $T[1]$, then we divide the value by the product of distance from $S[i]$ to $S[1]$ and the distance from $T[j]$ to $T[1]$. The intuition is that amino acids that are close to the central amino acid should have a bigger effect on the central amino acid. $AA\_ssim(S[i], T[j]) =$

$$
\begin{cases}
\frac{RBF(dist(S[i],S[1]),dist(T[j],T[1]))}{dist(S[i],S[1]) \cdot dist(T[j],T[1])} & \text{if } S[i].type=T[j].type \\
0 & \text{if } S[i].type \neq T[j].type
\end{cases}
$$

where $dist(S[i], S[1])$ is the Euclidean distance from $S[i]$ to $S[1]$ and $RBF(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\delta^2}\right)$, where $\delta > 0$ is a parameter.

When computing $K_{Pattern\_Sim}$, if we use all possible values[3] of i' for which $T[i'].type = S[i].type$ (i.e. if we do not restrict i' and j' as in (1)), then it is easy to see that $K_{Pattern\_Sim}(S, T)$ is a positive semidefinite kernel. This is because it is well-known that RBF is a kernel and that sums of kernels are themselves kernels. However, the asymmetry introduced by restricting the values of i' and j' per (1) makes it unclear whether $K_{Pattern\_Sim}$ is a true kernel. Despite this, our results show that $K_{Pattern\_Sim}$ works well in practice.

3. K_{Redox_Func} for Redox Classification

K_{Redox_Func} is a modification of $K_{Pattern\_Sim}$. In many thiol/disulfide oxidoreductases, two cysteines separated by two other residues form a functional motif, which is named the CxxC motif. This motif is conserved in the majority of members in the thiol/disulfide oxidoreductases. The two cysteines are the two central amino acids of this motif. The type of the outer amino acids, positions of the outer amino acids relative to the two central amino acids and distance between the two central amino acids are the three major features of the motif structure. K_{Redox_Func} computes similarity of two substructures based on these features.

Before applying K_{Redox_Func} to thiol/disulfide oxidoreductases, we orient the structures. We first move the protein structure to place the first C in CxxC motif at the origin $(0, 0, 0)$. Then we rotate the protein around two axes to place the second C at $(c, 0, 0)$ for some $c > 0$ and to place the first x in the motif at $(a, b, 0)$ for $a, b > 0$.

---

[3] In experimental results that are omitted, we redefined $K_{Pattern\_Sim}$ to fit this revised definition. Our results in motif identification were adversely affected.

(1) $K_{Redox\_Func}(U, V)$ — Similarity of two structures U and V. We sum the 3D similarities between any amino acid coming from U and any amino acid coming from V to compute the similarity of U and V. Then the sum of the two values is multiplied by the Gaussian RBF similarity of distances between the pairs of central amino acids of U and V. The result is the similarity of the two structures.

$$K_{Redox\_Func}(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} [AA\_redoxsim(U[i], V[j])$$

$$\cdot RBF(dist(U[1], U[2]), dist(V[1], V[2]))] \ ,$$

where $U[i]$ is the ith amino acid of U and $V[j]$ is the jth amino acid of V. $dist(U[1], U[2])$ is the distance from $U[1]$ to $U[2]$. $U[1]$ is the first C in the CxxC motif, and $U[2]$ is the second C. $RBF(x, x')$ returns the Gaussian RBF similarity of x and x'.

(2) $AA\_redoxsim(U[i], V[j])$ — similarity of two amino acids $U[i]$ and $V[j]$ in 3D space. Formally, if $U[i].type \neq V[j].type$, $AA\_redoxsim(U[i], V[j]) = 0$. If $U[i].type = V[j].type$, $AA\_redoxsim(U[i], V[j]) =$

$$RBF((U[i].x - U[1].x), (V[j].x - V[1].x))$$
$$\cdot RBF((U[i].y - U[1].y), (V[j].y - V[1].y))$$
$$\cdot RBF((U[i].z - U[1].z), (V[j].z - V[1].z))$$
$$\cdot RBF((U[i].x - U[2].x), (V[j].x - V[2].x))$$
$$\cdot RBF((U[i].y - U[2].y), (V[j].y - V[2].y))$$
$$\cdot RBF((U[i].z - U[2].z), (V[j].z - V[2].z)) \ ,$$

where $U[1]$ is the first C in CxxC, $U[2]$ is the second C, and $U[i].x$ is the x coordinate of U's ith residue.

A general procedure to build variants of K_{Redox_Func} for other conserved motif structures can be created by using the residues of the conserved structure as central amino acids and ones near them as outer amino acids, and following the procedure similar to our derivation of K_{Redox_Func}.

4. K_{3Dball} **for Structural Comparison**

We think of a protein as a three-dimensional space filled with 3D balls, where each ball has an amino acid at its center (central amino acid), and includes the outer amino acids that lie within a specified distance from the center. Each amino acid in a protein is the central amino acid of a set of substructures, where the set comes from varying the radius of the ball. Thus for a given radius r, if a protein has m amino acids, it has m 3D balls. By defining a measure of similarity between two balls, we can compare two proteins S and T by summing the similarities of their constituent

balls. The amino acids are encoded by their amino acid type, and a coordinate set (x, y, z) calculated as the mean coordinate of the residue's side chain atoms.

The key ideas of K_{3Dball} are as follows: the more similar 3D balls two proteins share, the more similar the two structures are. (Two balls are similar when they have similar constituents.) Since we consider all pairs of balls between two structures, K_{3Dball} measures similarity of entire structures.

We define similarity of two balls based on the type of the central amino acid and the number of outer amino acids two substructures share. We consider each pair of 3D balls (with a fixed radius r) of the proteins S and T. If two balls have the same type of central amino acid and have at least L outer amino acids match in common, then we say that these two balls have similar constituents. (We do not consider the effect of the distance inside such a ball.) An example is in Figure 1. If the radius r is indicated by the circle and $L = 3$, then 3D balls s and t are similar, since these two balls share the 4 outer amino acids A, A, D and E. In our kernel, r and L are parameters that can be varied to capture various radius length and similarity levels, i.e. we can compare as many or as few residues as we want.
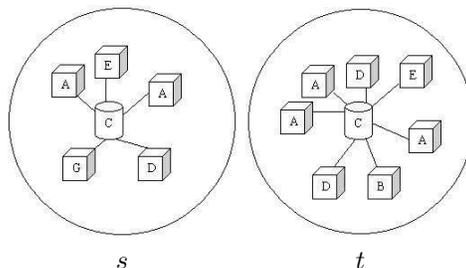

Figure 1. Example of two 3D balls s and t.

(1) $K_{3Dball}(S, T)$ — Similarity of two proteins S and T. Here, we sum the similarities between all pairs of 3D balls from S and T. The result is a measure of the similarity of two entire 3D structures:

$$K_{3Dball}(S, T) = \sum_{i=1}^{|S|} \sum_{j=1}^{|T|} Ball\_sim(Ball_{S[i]}, Ball_{T[j]}) \ ,$$

where $Ball_{S[i]}$ is the ball centered at S's ith amino acid and $|S|$ is the number of amino acids in S.

(2) $Ball\_sim(s, t)$ — similarity of balls s and t. If two structures have similar constituents, then the similarity is the number of pairs of outer amino acids shared

by the two structures, else the similarity is zero:

$$Ball\_sim(s,t) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ have} \\ & \quad \text{different central AAs} \\ 0 & \text{if } Num\_pairs(s,t) < L \\ Num\_pairs(s,t) & \text{if } Num\_pairs(s,t) \geq L \\ & \quad \text{and } s \text{ and } t \text{ have same} \\ & \quad \text{type of central AAs} \end{cases}$$

where L is a threshold stipulating minimum similarity.

(3) $Num\_pairs(s,t)$ — number of pairs of outer amino acids shared by s and t. There are 20 amino acid types, so we use the array $V_s[1:20]$ to represent s. $V_s[i] =$ number of type i outer amino acids in 3D ball s.

$$Num\_pairs(s,t) = \sum_{i=1}^{20} Min(V_s[i], V_t[i]) \ .$$

Theorem 1 $K_{3Dball}(S,T)$ *is positive semidefinite.*

Proof: Let k be a constant that is larger than the number of amino acids in any structure that we will analyze with our kernel. Thus we know that each 3D ball can have at most k outer amino acids. Then $V_s[1:20]$ can be represented by $V_s'[1:20][k]$, where $V_s'[i][j]$ is 0 or 1. If s has m type t amino acids, then $V_s'[t][j]=1$ for $j \leq m$ and $V_s'[t][j]=0$ for $j > m$. Obviously

$$Num\_pairs(s,t) = \sum_{i=1}^{20} \sum_{j=1}^{k} (V_s'[i][j] \cdot V_t'[i][j]) \ .$$

Because $Num\_pairs(s,t)$ can be written as an ordinary dot product, it is a positive semidefinite (PSD) kernel. It is well-known that $aK(\cdot)$ is a PSD kernel if $a \geq 0$ and $K(\cdot)$ is a PSD kernel. Therefore $Ball\_sim(s,t)$ is a PSD kernel. It is also well-known that the sum of PSD kernels is also a PSD kernel. Therefore K_{3Dball} is also a PSD kernel. $\qquad\square$

5. Structural Motif Discovery and Protein Function Classification

5.1. Structural Motif Discovery

We use the following procedure to employ $K_{Pattern\_Sim}$ to discover structural motifs. First we select a random set $\{P_1, \ldots, P_n\}$ of proteins from the superfamily in question. We represent each protein P_i as the set $\{S_1^{P_i}, \ldots, S_{m_i}^{P_i}\}$ of all of the substructures in P_i. I.e. the set of all substructures of radius r centered at each amino acid in P_i. For each substructure $S_j^{P_i}$, we use $K_{Pattern\_Sim}$ to compute its similarity to all the substructures in protein $P_{i'}$. The largest such similarity is used to represent the similarity from substructure $S_j^{P_i}$ to protein $P_{i'}$, i.e.

$$strucsim(i,j,i') = \max_{1 \leq j' \leq m_{i'}} K_{Pattern\_Sim}(S_j^{P_i}, S_{j'}^{P_{i'}}) \ .$$

We repeat this for all proteins $P_{i'}$, $1 \leq i' \leq n$, $i' \neq i$ and sum the results to get a fitness for $S_j^{P_i}$:

$$fitness(i,j) = \sum_{i'=1, i' \neq i}^{n} strucsim(i,j,i') \ .$$

For each protein P_i, we sort its substructures by their fitnesses. The most fit substructures are those in P_i that are most highly conserved across the sample $\{P_1, \ldots, P_n\}$. By examining each sorted list for a relatively large "gap" in fitness values, we can identify candidates for structural motifs in each P_i. Denote this set $\mathcal{S}_{P_i} \subseteq \{S_1^{P_i}, \ldots, S_{m_i}^{P_i}\}$. We create a set of global candidates $\mathcal{S} = \bigcup_i \mathcal{S}_{P_i}$ and sort it by fitness. The top substructures in $\mathcal{S}$ are possible structural motifs.

5.2. Protein Function Classification

We used two machine learning techniques with our kernels to model and classify test proteins: support vector machines (using SVM^{light}(Joachims, 1999)) and a variant of k nearest neighbor (kNN). The kNN method we use is slightly different from the traditional kNN method. Given a new (unlabeled) protein S to classify, we first compute the similarities between S and all the positive proteins in the training set and take the mean of the similarities of the top $k\%$ positive proteins most similar to S. We use the same process for negative proteins. If the mean similarity between S and the positives is significantly larger than that for the negatives, then we predict S to be positive, otherwise negative.

6. Experimental Results

6.1. Structural Motif Discovery

We tested $K_{Pattern\_Sim}$ on motif finding in thiol/disulfide oxidoreductases and Cytochrome P450 reductase. We used each amino acid in each protein as the central amino acid of a substructure, with a radius of 6 Å. Since the amino acids that flank the central amino acid are potentially important, we also added to the set of outer amino acids the two that lie immediately upstream and the two immediately downstream from the central amino acid, if they are not already included in the 6 Å ball.

We used all known thiol/disulfide oxidoreductases in PDB with known tertiary structure for our first test set. Following the procedure of Section 5.1, several substructures had sufficiently high fitnesses to be considered structural motifs and all of them were similar to each other, each centered at a cysteine. Evaluation

of the counterparts[4] to the conserved substructure in each protein clearly shows that almost all the counterparts have two cysteines and center on one of them. We also found most of them also have a proline near the two cysteine in 3D space. Such a conserved structure is already known (Fetrow & Skolnick, 1998).

We also tested on Cytochrome P450 reductases. The number of Cytochrome P450 reductases with known tertiary structure is about 10 and no conserved structure motifs are known. Following the procedure of Section 5.1, we found a substructure centered at a glycine that is well conserved. This conserved substructure S also has another glycine as an outer amino acid. Since the data set is so small, it is difficult to draw conclusions about S. But it is interesting to note that S's counterparts in several training proteins (PDB-1AMO, PDB-1B1C, PDB-1JA0) are related to the known Cytochrome P450 reductase docking surface, which is most likely a major portion of the Cytochrome P450's binding surface as evidenced by the inhibition of cytochrome P450 reactions by Cytochrome c (Wang et al., 1997). Since S resembles a docking surface in the above positives, there is some evidence that it is a conserved substructure.

We conclude that our method was sensitive enough to identify the known structural motifs in thiol/disulfide oxidoreductases and selective enough to avoid false positives. It also found something interesting from Cytochrome P450 reductases. Since our method started with no prior information about either superfamily's structure (it only started with the 6 Å radius, which was chosen as a generally reasonable value for the parameter), this method appears to be a good approach to the general structural motif discovery problem.

6.2. Structural Classification

We used a leave one out test to evaluate K_{3Dball} on general-purpose structural classification. Ten superfamilies were retrieved from CATH. CATH clusters proteins at four major levels: class, architecture, topology and homologous superfamily. Homologous superfamilies group together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Proteins in each sequence family have sequence identities $\geq 35\%$. We tested three superfamilies from mainly Alpha class, three from mainly Beta class, and four from mixed Alpha and Beta classes. For each superfamily, we included around 20 proteins. To make sure that the

---

[4]We define protein P's *counterpart* to a substructure S_i^Q in protein Q as the substructure in P that is most similar to S_i^Q using our similarity measure.

Table 1. Summary of leave-one-out test results for 10 superfamilies from CATH. (TP means true positive rate, TN means true negative rate.)

Super Family	TP for kNN	TN for kNN	TP for SVM	TN for SVM
1.10.238.10	80%	95%	66.7%	100%
1.10.760.10	85.7%	93%	71.43%	95%
1.20.120.200	81.25%	94%	50%	97%
2.40.10.10	85%	95%	85%	100%
2.60.40.30	85%	100%	85%	100%
2.60.40.420	89.5%	99%	79%	99%
3.20.20.80	88.5%	95%	84.6%	99%
3.20.20.90	85%	98%	75%	98%
3.40.50.150	77%	89%	60%	98%
3.40.50.300	91%	92%	72.7%	100%

test proteins are not too similar, all test proteins came from different sequence families. The negative set (selected randomly from PDB) consisted of 100 proteins. For each test, we modified the negative set a little by deleting the proteins coming from the test superfamily. In our experiments, we found that different families are sensitive to different 3D ball radii. In general, the range of radii we used was from 7.5Å to 9.0Å. We set $L = 9$ for all the tests.

We used leave one out test for the experiment, where we trained on all but one member of the data set, which was withheld from training and tested. This was done for each member of the training set. The overall performance for the ten tests was generally quite good (Table 1). For kNN, the average true positive rate is about 85% and average true negative rate is about 95%. For SVM, the average true positive rate is about 72.5% and average true negative rate is about 98.5%. In other words, given 20 positive proteins (from one superfamily but different sequence families), we can successfully identify 17 of them at a false positive rate of 5%. The test shows that our method can successfully identify general fold similarities.

6.3. Function Classification

6.3.1. Leave One Out Test

Since the number of Cytochrome P450 reductases with known tertiary structure is small, we only tested our method on thiol/disulfide oxidoreductase in a leave one out test. We extracted 21 thiol/disulfide oxidoreductase from the PDB database for our test. Seventeen of them are positive proteins with Trx-fold, four are positive proteins without Trx-fold. The average pairwise

sequence identity in the positive data set is 17%. The negative set (selected randomly from PDB) consists of 100 non-redox proteins having CxxC. Using K_{3Dball} and $K_{Pattern\_Sim}$ does not require prior knowledge. Using K_{Redox_Func} requires knowledge of the location of the CxxC active site. This information is known for the 21 positive proteins in our data set, and it is also known that each CxxC site of each of the 100 negative proteins is not a redox active site. Thus when we trained our classifiers, we used each of the 21 known active sites from the 21 positives as a positive redox structural motif, and we used each CxxC site from each negative (147 substructures total; some proteins have multiple CxxCs) as a negative (non-redox) substructure. Of course, when classifying new (unlabeled) proteins as redox or non-redox, the true active site is unknown. Thus when testing our trained classifiers, we tested on the CxxC site of each test protein, predicting it as positive if at least one site is predicted as positive.

We used a leave one out test. In Table 2, we see that when used with kNN, K_{Redox_Func}, K_{3Dball} ($r = 7.5$ Å and $L = 9$), and $K_{Pattern\_Sim}$ each[5] identified at least 15 of the 17 positives with Trx fold and at least 2 of the 4 without the fold with at most 5% false positive rate. This result shows that our new methods can find 3D similarities of redox proteins. Future work is to investigate why modified kNN performed so much better than the SVM.

We also tested DALI on each entire structure of the proteins in the training set. With DALI, the prediction of test protein S was based on the fraction of positive examples and negative examples that have significant similarity to S in the training set. (We define significant similarity as a z-score ≥ 2.0.) If the fraction of similar positive examples is higher than that of the similar negative examples, the protein is classified as positive, otherwise negative[6]. DALI identified 100% of the redox proteins with the Trx fold, but no positive without the Trx fold was found. We also tested CE on this data with similar results. Thus when measuring similarity with DALI or CE, the positive proteins lacking the Trx fold were more similar to the negative proteins than to the positive ones. Finally, we tried hidden Markov models on the entire primary sequence,

---

[5]To use $K_{Pattern\_Sim}$ for function classification, we use our motif discovery method to find the most conserved substructure in the training set, then use $K_{Pattern\_Sim}$ to test whether a given protein has a substructure that is similar to the motif we found. If it does, we predict it positive.

[6]The intuition behind this rule is that if we simply considered similarity to only positives, then the false positive rate would be unacceptably high. This was corroborated by results not shown.

Table 2. Summary of leave-one-out test results for thiol/disulfide oxidoreductases. (TP means true positive rate, TN means true negative rate)

	TP FOR REDOX WITH FOLD	TP FOR REDOX WITHOUT FOLD	TN
HMM (PRIMARY STRUCTURE)	70.6%	0%	98%
DALI (ENTIRE STRUCTURE)	100.00%	0%	97%
CE (ENTIRE STRUCTURE)	100.00%	0%	98%
$K_{Pattern\_Sim} + kNN$	88.23%	50%	98%
$K_{Pattern\_Sim} + SVM$	82.35%	50%	100%
$K_{Redox\_Func} + kNN$	100.00%	75%	99%
$K_{Redox\_Func} + SVM$	94.12%	50%	98%
$K_{3Dball} + kNN$	94.12%	50%	95%
$K_{3Dball} + SVM$	70.6%	50%	99%

which yielded the worst results.

DALI and CE identified 100% of the positive proteins with the Trx fold, but no positives without the fold. Finally, we note that our methods were each over 100 times faster than DALI and CE. Thus our kernels can very quickly find similarities among thiol/disulfide oxidoreductases beyond the Trx fold.

6.3.2. DATABASE SEARCH

Using the same training set as above, we used kNN with K_{Redox_Func} and K_{3Dball} to search for oxidoreductases in PDB, with 28385 total sequences. K_{Redox_Func} with kNN identified 266 candidates as positive, and K_{3Dball} identified 282 candidates. Over 90% of the known thiol/disulfide oxidoreductases were identified by each method. We also found several candidate thiol/disulfide oxidoreductases. Future work is to examine these for redox function.

From Section 6.1, we found a conserved substructure S in Cytochrome P450 reductases. Each known Cytochrome P450 reductase in our data set has a counterpart to S. We used the counterparts to form a training set for kNN with $K_{Pattern\_Sim}$ to search for other proteins in PDB that have similar substructures. We identified 351 candidates. We also used our set of positives to train a kNN classifier with K_{3Dball}, which found 66 candidates. In both cases, All known positives were found with true negative rates above 99.5%. With $K_{Pattern\_Sim}$, we also found NADPH ferrodoxin reductase, which is one of the two bacterial proteins that fused to give Cyt. P450 reductase. Thus our method identified a protein that was a precursor to one from the P450 superfamily. The other hits need to be further examined.

7. Conclusions

We introduced new approaches for protein tertiary structure comparisons, motif discovery, and function classification. $K_{Pattern\_Sim}$ for motif discovery is different from other methods as it examines all possible substructures that lie in a fixed-radius ball centered at each amino acid in the protein. K_{3Dball} represents a protein as a set of 3D balls in 3-dimensional space. Similarity between proteins is defined by a sum of structural similarities of balls having similar constituents. Since all possible balls are considered, K_{3Dball} quantifies similarity between entire structures. Our kernels are designed to be simple and fast to compute (over 100 times faster than DALI and CE) and are very general. Experiments showed that K_{3Dball} works well for ten structural families from CATH and the two function families thiol/disulfide oxidoreductases and cytochrome P450 reductases. Further, all our methods can find thiol/disulfide oxidoreductases without the Trx fold, which cannot be identified by other popularly-used methods. We also found that $K_{Pattern\_Sim}$ can successfully identify the structural motif in thiol/disulfide oxidoreductases and can capture the functional similarity in those motifs. It also found a candidate structural motif from cytochrome P450 reductases. K_{3Dball} and $K_{Pattern\_Sim}$ do not require orientation of the structures or prior information. Thus they should be applicable to many protein families and offer a viable alternative to other methods of protein tertiary structure comparison.

Acknowledgments

References

Bradley, P., Kim, P. S., & Berger, B. (2002). Trilogy: discovery of sequence structure patterns across diverse proteins. *Proceedings of the National Academy of Sciences* (pp. 8500–8505).

Fetrow, J. S., & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and t1ribonucleases. *J. of Mol. Biology*, *281*, 949–968.

Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. of Molecular Biology*, *233*, 123–138.

Holm, L., & Sander, C. (1995). 3-D lookup: Fast protein structure database searches at 90% reliability. *Proceedings of the Third Int. Conf. on Intelligent Systems for Molecular Biology* (pp. 179–187).

Joachims, T. (1999). *Making large-scale SVM learning practical.* In *Advances in Kernel Methods: Support Vector Learning* (pp. 169–184).

Jonassen, I., Eidhammer, I., Conklin, D., & Taylor, W. (2002). Structure motif discovery and mining the PDB. *Bioinformatics*, *18*, 362–367.

Martin, J. (1995). Thioredoxin—a fold for all reasons. *Structure*, *3*, 245–250.

Meng, E., Polacco, B., & Babbitt, P. (2004). Superfamily active site templates. *Proteins*, *55*, 962–976.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247(4)*, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH—a hierarchical classification of protein domain structures. *Structure*, *5*, 1093–1108.

Schmollinger, M., Fischer, I., Nerz, C., Pinkenburg, S., Götz, F., Kaufmann, M., Lange, K. J., Reuter, R., Rosenstiel, W., & Zell, A. (2004). ParSeq: searching motifs with structural and biochemical properties. *Bioinformatics*, *20(9)*, 1459–1461.

Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, *11*, 739–747.

Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *J. of Molecular Biology*, *208*, 1–22.

Wallace, A. C., Laskowski, R. A., & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Science*, *5(6)*, 1001–1013.

Wang, M., Roberts, D. L., Paschke, R., Shea, T. M., Masters, B. S., & Kim, J. J. (1997). Three-dimensional structure of NADPH-cytochrome P450 reductase: prototype for fmn- and fad-containing enzymes. *Proceedings of the National Academy of Sciences*, *94(16)*, 8411–8416.

Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J., Skolnick, J., & Godzik, A. (1999). From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Science*, *8*, 1104–1115.