PREDICTING RANKED SCOP DOMAINS BY MINING ASSOCIATIONS OF VISUAL CONTENTS IN DISTANCE MATRICES

PIN-HAO CHI AND CHI-REN SHYU

Medical and Biological Digital Library Research Lab,
Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
E-mail: pinhao@diglib1.cecs.missouri.edu, ShyuC@missouri.edu

Protein tertiary structures are known to have significant correlations with their biological functions. To understand the information of the protein structures, Structural Classification of Protein (SCOP) Database, which is manually constructed by human experts, classifies similar protein folds in the same domain hierarchy. Even though this approach is believed to be more reliable than applying traditional alignment methods in structural classifications, it is labor intensive. In this paper, we build a non-parametric classifier to predict possible SCOP domains for unknown protein structures. With supervised learning, the algorithm first maps tertiary structures of training proteins into two-dimensional distance matrices, and then extracts signatures from visual contents of matrices. A knowledge discovery and data mining (KDD) process further discovers relevant patterns in training signatures of each SCOP domain by mining association rules. Finally, the quantity of rules whose patterns match signatures of unknown proteins determines predicted domains in a ranked order. We select 7,702 protein chains from 150 domains of SCOP database 1.67 release as labelled data using 10 fold cross validation. Experimental results show that the prediction accuracy is 91.27% for the top ranked domain and 99.22% for the top 5 ranked domains. The average response time takes 6.34 seconds, exhibiting reasonably high prediction accuracy and efficiency.

1. Introduction

Protein structure information is known to be more conserved than amino acid sequences and serves as ideal references to study protein structure-to-function relationships. Similar protein folds may suggest similar biochemical functions.[27] In our knowledge, the most reliable structural comparison method is to manually inspect similar protein structures such as SCOP.[17] Proteins with high structural similarity will be classified into the same hierarchical SCOP domain. Even though manual inspection provides more accurate structural classification, it is labor intensive for a large number of protein tertiary structures. Automated structural comparison methods such as Distance Alignment (DALI)[12] and Combinatorial Extension (CE)[22] algorithms globally find a structural alignment between two polypeptide chains such that superimposed segments of amino acids can have a good structural match within a small Root Mean Square Deviation (RMSD) threshold. Due to the huge combination of possible alignments, exhaustively searching a local optimal solution is known to be computationally expensive, proving a complexity of NP-Hard.[9] Therefore, life science researchers and biologists have a great demand on efficient and accurate protein structure classification systems.

Several well-known structural classification databases have been studied in computational molecular biology literatures. Secondary Structure Alignment Program (SSAP) utilizing double phases dynamic programming techniques for optimal structural alignment of two proteins becomes a framework to construct CATH database.[18] The DALI algorithm that applies Monte Carlo heuristics to compare structural similarities from distance matrices is used to conduct structural classifications in FSSP database.[13] Applying specific heuristics for reducing computational complexity, these classical structural alignment algorithms may return variant classification results from the same protein set. To avoid suffering from drawbacks of subjective heuristics, recent classification works[3,6] that maintain higher accuracies than applying each individual method by intersecting multiple intermediate results of existing structural alignment algorithms. Even though structural alignment methods present satisfactory classification accuracies, the process of performing multiple pairwise alignments between an unknown protein and known proteins in databases is still incapable of providing fast predictions.

With the advent of x-ray diffraction and high-resolution nuclear magnetic resonance (NMR) techniques, the amount of newly discovered proteins has grown rapidly in recent years. As July 5th, 2005, Protein Data Bank (PDB), announces 32,107 protein structures and 8,070 of them have not been classified in the latest SCOP release (1.67). This noticeable gap is well-recognized and continues to grow. Hence, there is an urgent need to develop an efficient domain classification method with sufficiently high accuracy to streamline the labor-intensive classification process. It is noteworthy to mention that, instead of replacing human inputs from this classification process, a more realistic approach is to suggest a handful set of top ranked domains for further studies.

In this work, we extend our recent research results in a real-time tertiary structure retrieval system called ProteinDBS[7,14,24] and develop a series of knowledge discovery and data mining techniques to perform fast SCOP domain predictions with reasonably high accuracy. This paper is organized as follows. Section 2 introduces our unique model to cast protein backbone structures into high-dimensional feature vectors. Section 3 describes the algorithm to transform feature values into a set of feature intervals and illustrates the association rule mining using a supervised learning technique. Experimental results of prediction accuracy and efficiency are reported in Section 4. Finally, we conclude this paper and discuss possible future works in Section 5.

2. Preliminaries

Protein Tertiary Structure refers to a single polypeptide chain that is constructed by a long amino acid string. For a protein chain k with n amino acids, its backbone is represented by a n-dimensional vector $\{\vec{C_\alpha^{k,1}}, \vec{C_\alpha^{k,2}}, ..., \vec{C_\alpha^{k,n}}\}$, where the element, $\vec{C_\alpha^{k,i}}$, is the three-dimensional coordinate of the i-th C_α atom. The distance matrix of k is defined as a $n \times n$ symmetric real matrix whose element at i-th column and j-th row is the Euclidean distance between $\vec{C_\alpha^{k,i}}$ and $\vec{C_\alpha^{k,j}}$. A distance matrix is generally sufficient to recover the original three-dimensional backbone structure in polynomial time using distance geometry methods.[11] Several literatures[12,16,26] study comparing similar distance matrices as a equivalent problem to protein tertiary structure comparisons. Our assumption is based on the fact

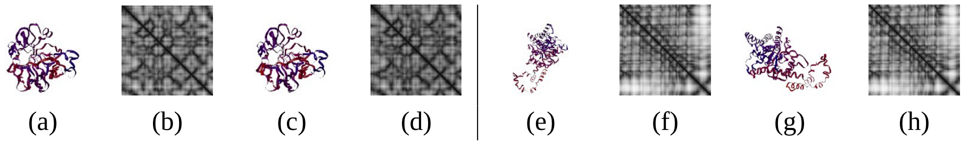


Figure 1. The three-dimensional backbone structures and distance matrices from protein chains selected from the SCOP domain *Carbonic anhydrase*: (a-b)$1am6$, (c-d)$1bic$, and the SCOP domain *D-xylose isomerase*: (e-f)$9xim\_D$, (g-h)$1xlb\_A$

that similar protein folds should have distance matrices with similar visual contents. We also expect that proteins in the same SCOP domain should present high similarities in distance matrices. To pictorially explain our assumption, Figure 1 shows that protein chains from SCOP *Carbonic anhydrase* and *D-xylose isomerase* domains present high similarities in both three-dimensional tertiary structures and two-dimensional distance matrices. Even though similar visual patterns can be identified by manual inspections, it is still a challenging research topic to mimic distance matrix comparisons automatically using computational techniques. Fortunately, there exists a rich body of literatures in the area of content-based image retrieval (CBIR) since early 80's.[5,21,23] The concept of CBIR is to retrieve visually similar images from databases for a query image. This is a perfect fit to the protein distance matrix comparisons. To effectively retrieve similar candidates in a large population of distance matrices, extracting relevant features becomes an important issue to study. In our previous works,[7,24] the distance matrix is divided into six band regions, parallel to its diagonal. In each band, four local features are computed by histograms of four bins of distance ranges: [0-5], [6-10], [11-15], and [16-∞]. We also have extracted nine global features from visual patterns of distance matrices using a suite of standard computer vision algorithms.[10,20,19] After features are extracted, each protein backbone can be transformed into a high-dimensional feature vector and clustered in the feature space. Readers are referred to our previous publications[7,24] for the details of the feature extraction algorithms applied in this work.

The distribution of feature values is expected to have significant correlation to protein domains in SCOP. A set of features with certain ranges could best describe structural patterns of proteins in a specific SCOP domain. Figure 2 depicts a simplified example using three features, namely the 8^{th} localized histogram (The 4^{th} gray-scale level in the 2^{nd} partitioned band region of distance matrix), the 5^{th} texture[10] ($Homogenity$), and the 9^{th} texture ($Cluster\_Tendency$). For proteins in SCOP domains *Carbonic anhydrase* (D_1), *D-xylose isomerase* (D_2) and *Calmodulin* (D_3), these three features are partially overlapped in multiple intervals. From the top range line of Figure 2, it is clear that all database protein structures from D_1 and D_2 mix in the same "Histogram 8" feature interval. Similarly, the "Texture 5" feature is unable to separate proteins in D_2 from those in D_3. Adding association information among feature intervals, the algorithm is able to predict an unknown protein structure to $D_1 : \{f_{Histogram8} \in$ [0.040,0.045) and $f_{Texture9}$ $\in$ [0.005,0.010)$\}$, $D_2 : \{f_{Histogram8} \in$ [0.040,0.045) and $f_{Texture5} \in$ [0.085,0.090)$\}$, or $D_3 : \{f_{Texture5} \in$ [0.085,0.090) and $f_{Texture9} \in$ [0.005,0.010)$\}$.

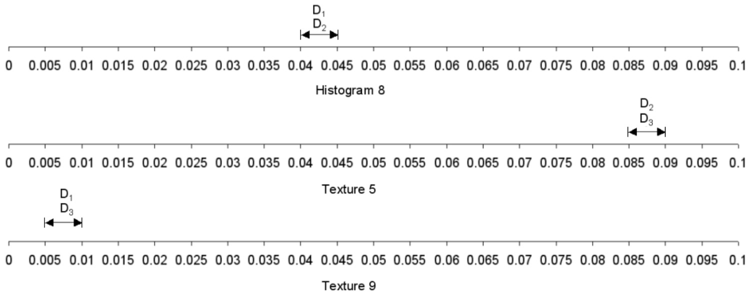


Figure 2. An example of feature intervals for SCOP domains, D_1:*Carbonic anhydrase*, D_2: *D-xylose isomerase* and D_3: *Calmodulin*

Knowledge discovery and data mining techniques have been widely studied in high-throughput data analysis of various aspects such as classification,[15] mining in web usage, spatial data, document indexing,[8] and biological domains.[26] Among data mining techniques, association rule (*AR*) mining is able to retrieve hidden patterns and discover meaningful information from the data. Given a protein chain p_1, it will be preprocessed into an m-dimensional feature vector $\{f_1^{p_1}, f_2^{p_1}, f_3^{p_1}, ..., f_m^{p_1}\}$, where $f_i^{p_1}$ has been normalized in $R[0,1]$ and $1 \leq i \leq m$. Then, the algorithm partitions $R[0,1]$ space of each individual feature of proteins into a set of disjoint intervals $\{[0,\eta_1], (\eta_1,\eta_2], ..., (\eta_n, 1]\}$, where $0 < \eta_1 < \eta_2 < ... < \eta_n < 1$. To discuss data mining algorithms used in this work, each feature interval $(\eta_i, \eta_{i+1}]$ is defined as an *item*. For example, there exist three feature intervals (items) generated from a partition of R[0,1] that are associated with the j^{th} feature of all database proteins such as $I_1 = [0.0, 0.2]$, $I_2 = (0.2, 0.75]$, and $I_3 = R(0.75, 1.0]$. For a protein p_1, the j^{th} feature value, $f_j^{p_1} = 0.5$, will be transformed into *item* I_2. Applying the same item mapping process for m features, each backbone structure is then represented by a set of m items ($m = 33$ in our work). This collection of items forms a *transaction* for mining item associations. In addition, a database D that includes n proteins can be described by n transactions. With a set of items, I, an association rule is defined as an implication rule composed of items with a form $\{X \Rightarrow Y\}$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. *Itemsets* X and Y are called *Antecedent* and *Consequent*, respectively. For an association rule represented by $\{X \Rightarrow Y\}$, the *support* of the rule is the percentage of all *transactions* in D that include $\{X \cup Y\}$ items. The *confidence* of the rule is a ratio of the total amount of *transactions* that contain $\{X \cup Y\}$ to *transactions* with $\{X\}$ items. The association rule mining generates relevant rules in the database with the *support* and *confidence* that can pass *minimal_support* and *minimal_confidence* thresholds, respectively.

3. Method

To precisely predict an unknown protein structure among hundreds or even thousands of SCOP domains, it is critically important to identify appropriate feature intervals, as well as associations among these relevant intervals within each SCOP domain. The way to formulate a partition of a real space $R[0,1]$ has vital impact on determining relevant items.

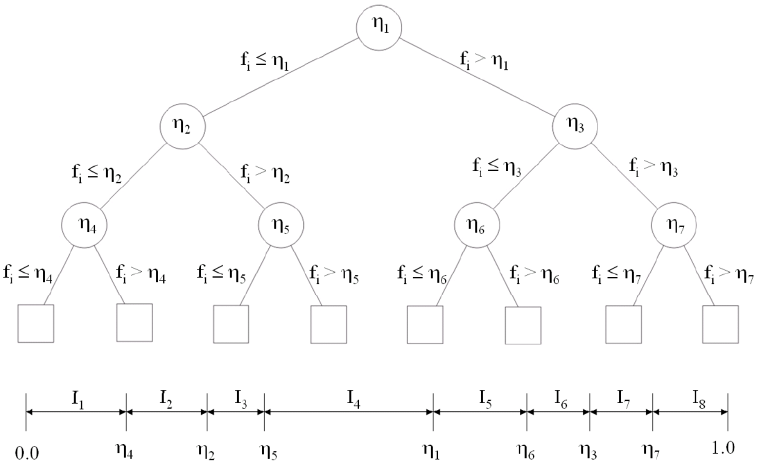


Figure 3. A binary decision tree to determine thresholds for a space partition of feature f_i

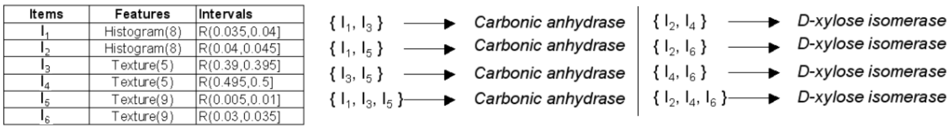
Partitioning a real space too finely will generate many tiny intervals within one domain, resulting in huge amount of association rules. A coarse partition of space will create intervals that mix multiple domains without enough discriminatory power. Instead of randomly or evenly partitioning the real space into intervals, we apply C4.5 decision tree[25] to find relevant intervals for each feature among all database domains.

3.1. *Space Partition Algorithm Using C4.5 Decision Tree*

For each individual feature of all m-dimensional feature vectors, the algorithm constructs a C4.5 decision tree. In total, there are 33 trees for all features used in this work. The splitting criterium to grow the decision tree is based on the minimization of entropy. Let D^t be the set of protein features at a certain node t. The entropy, $H(D^t)$, of node t and the weighted entropy, $H(D^{t'})$, of its child nodes t_l and t_r are computed as follows:

$$H(D^t) = -\sum_{j=1}^{r} p^t_{d_j} \times log(p^t_{d_j}), H(D^{t'}) = \alpha \times H(D^{t_l}) + (1-\alpha) \times H(D^{t_r}) \qquad (1)$$

where $p^t_{d_j}$ denotes the ratio of proteins in domain d_j to the total number of proteins that exist in node t. To compute $H(D^{t'})$, α represents the percentage of protein chains that have been dispatched from a parent node to the left child by the threshold η, which is an optimal threshold and selected based on the maximization of $H(D^t) - H(D^{t'})$. With a top-down iterative node splitting, the algorithm collects sorted thresholds of k internal nodes using in-order traversal, and the space R[0,1] will be partitioned into $k+1$ intervals as a set of *items*. For example, Figure 3 shows that eight *items*, $I_1 = R[0.0, \eta_4]$, $I_2 = R(\eta_4, \eta_2]$, ..., $I_8 = R(\eta_7, 1.0]$, are partitioned by seven threshold values $\{\eta_4, \eta_2, \eta_5, \eta_1, \eta_6, \eta_3, \eta_7\}$. Each protein is then mapped into a 33-item transaction for mining item associations using the intervals selected by the decision trees.



Items	Features	Intervals
I_1	Histogram(8)	R(0.035,0.04]
I_2	Histogram(8)	R(0.04,0.045]
I_3	Texture(5)	R(0.39,0.395]
I_4	Texture(5)	R(0.495,0.5]
I_5	Texture(9)	R(0.005,0.01]
I_6	Texture(9)	R(0.03,0.035]

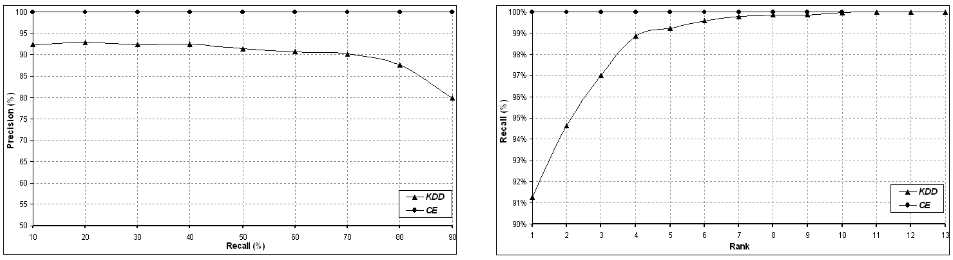
{ I_1, I_3 } → Carbonic anhydrase
{ I_1, I_5 } → Carbonic anhydrase
{ I_3, I_5 } → Carbonic anhydrase
{ I_1, I_3, I_5 } → Carbonic anhydrase
{ I_2, I_4 } → D-xylose isomerase
{ I_2, I_6 } → D-xylose isomerase
{ I_4, I_6 } → D-xylose isomerase
{ I_2, I_4, I_6 } → D-xylose isomerase

Figure 4. Association Rules generating from partitioned feature intervals using Apriori algorithm

3.2. *Mining Training Data and Prediction Model*

After transforming three-dimensional protein backbones into the form of *transactions*, the system then mines associations of the items from training data by applying the Apriori algorithm.[2] The main concept of Apriori algorithm is to generate association rules from frequent itemsets whose *support* is greater than the *minimal_support* threshold. Since any subset of a large transaction is still a frequent itemset, the algorithm finds candidates of frequent itemsets with n_i items from frequent itemsets with $n_i - 1$ items, where $n_i \geq 1$. In Apriori algorithm, *minimal_support* is an important criterium to determine the quantity of association rules. Due to the non-uniformly distributed proteins among all domains, it is inappropriate to mine rules from the entire database using a single *minimal_support*. Therefore, for each domain d, we perform Apriori algorithm and each frequent itemset, I, refers to an association rule $I \Rightarrow d$. For instance, itemsets $\{I_1, I_3, I_5\}$ and $\{I_2, I_5, I_6\}$ are frequent for SCOP domain *Carbonic anhydrase* and *D-xylose isomerase*, respectively. Examples of association rules for domain predictions are shown in Figure 4. After obtaining rules from all SCOP domains, a small portion of rules (2.81%) shared by multiple domains has been pruned out prior to the prediction stage. Our current setting of the *minimal_support* is 90% within each domain. Mining training proteins of 150 SCOP domains populates 2,354 association rules. Discovered rules has been efficiently organized and loaded into main memory for fast predictions.

The next task is to design a scoring function that suggests possible SCOP domains in a ranked order. For an unknown protein t, a complete itemset, $I^t = \{I_1^t, I_2^t, ..., I_m^t\}$, is formed by mapping features into item intervals as discussed in Section 2, where m is the total number of features ($m = 33$ in our work). Given k association rules in domain d, each rule can be represented by $\{I_1^i, I_2^i, ..., I_n^i\} \Rightarrow d$, where $m \geq n \geq 2$ and $k \geq i \geq 1$. Among these rules, we group them into two sets: *matched rules* R_c^d and *mismatched rules* R_m^d, where $|R_c^d| + |R_m^d| = k$. The i-th rule is categorized as *matched rules* when the condition, $\{I_1^i, I_2^i, ..., I_n^i\} \subseteq I^t$, is satisfied. Contrarily, a mismatched rule has at least one item in its antecedent that is not included in I^t for the unknown protein t. The scoring function rewards matched rules and penalizes mismatched rules in each domain. For the i-th matched rule, the scoring function further considers the degree of reward N_i, which is the size of its antecedent. To gauge the degree of penalty for mismatched rules, we use a *discrete distance* measurement, which is demonstrated as follows. Let r_m:$\{I_1^m, I_2^m, ..., I_n^m\} \Rightarrow d$ be a mismatched rule, $f_{ea}(I_i^m)$ be a function that returns which feature maps item I_i^m, and $i_{dx}(I_i^m)$ be a function to return the index value of item I_i^m in integer. As an example, a decision tree for the 3^{rd} feature generates 10 items $\{I_1^{'}, I_2^{'}, ..., I_{10}^{'}\}$, which are sequen-



(a) (b)

Figure 5. (a) A precision-to-recall chart for 10 rounds of experiments (b)An accumulated recall chart for top 13 predicted domains

tially stored in an array of position $\{65, 66, ..., 74\}$. Since item $I_1^{'}$ is partitioned from the 3^{rd} feature, $f_{ea}(I_1^{'})$ is equal to 3 and $i_{dx}(I_1^{'})$ returns 65. For any two items x and y, we define $g(x, y) = 1$ when $f_{ea}(x) = f_{ea}(y)$ and $g(x, y) = 0$ if $f_{ea}(x) \neq f_{ea}(y)$. The *discrete distance* between a mismatched rule r_m and an unknown protein t is defined as: $\sum_{x \in \delta_{r_m}} \sum_{n=1}^{m} |i_{dx}(x) - i_{dx}(I_n^t)|^2 \times g(x, I_n^t)$, where δ_{r_m} is the set of mismatched items in r_m. From the same decision tree, items in the neighborhood of partitioned feature intervals are expected to have structural similarities, resulting in a small *discrete distance*. This penalty is then normalized by M_d, the total number of mismatched items from R_m^d.

$$Score(d) = \frac{\sum_{i=1}^{|R_c^d|} N_i}{(\sum_{j=1}^{|R_m^d|} \sum_{x \in \delta_j} \sum_{n=1}^{m} |i_{dx}(x) - i_{dx}(I_n^t)|^2 \times g(x, I_n^t))/M_d} \tag{2}$$

Taking both reward and penalty into consideration, the scoring function for each domain is defined in Eq. (2). To predict ranked domains of an unknown protein, the algorithm computes and ranks scores for all domains.

4. Experiment

We evaluate the performance in accuracy and efficiency for predicting SCOP domains. Experiments are conducted using 10 fold cross validation on a large-scale dataset. With 7,702 protein chains from 150 SCOP domains, 10% of proteins from each domain are randomly selected for blind test. To evaluate the prediction accuracy, we use *Precision and Recall* in the context of machine learning.[4] Given n_r possible SCOP domains, let N_P^d be the number of testing proteins that are predicted to the domain d, N_{TP}^d be the number of testing proteins whose predicted domain d matches its true SCOP domain and N_T^d be the number of testing proteins that are from domain d, where $1 \leq d \leq n_r$. The performance metrics are defined as follows:

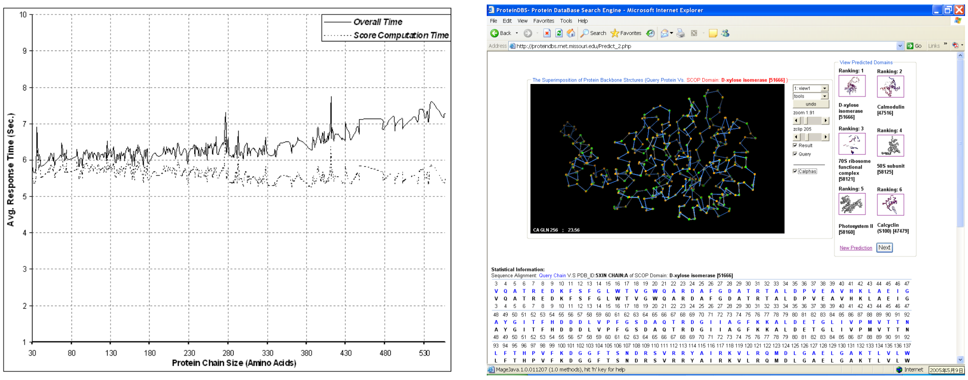
$$Precision = \frac{1}{n_r} \sum_{d=1}^{n_r} \frac{N_{TP}^d}{N_P^d} \ , \ Recall = \frac{1}{n_r} \sum_{d=1}^{n_r} \frac{N_{TP}^d}{N_T^d} \tag{3}$$

Figure 5(a) presents a plot of *Precisions* against *Recall* ranging from 10% to 90%. The ideal case occurs when all testing proteins are predicted correctly, achieving 100%

precision at any recall rate. Our KDD algorithm exhibits 92.42% precision with a 10% recall, 91.35% precision recalling half of them, and 79.77% precision recalling 90% of the entire testing protein set. Normally, the precision will drop by increasing the recall rate. A more practical goal for domain prediction is to suggest a small set of candidate domains to streamline the manual process. To demonstrate the usefulness of our prediction model, we also measure the recall rate by accumulating *True Positives* from the top predicted SCOP domains in the ranked results. In Figure 5(b), our KDD method delivers 91.27% recall rate from the top predicted domain and 99.22% from the top 5 predicted domains. 100% recall rate is achieved by top 13 predictions. What this means is that a human domain classifier only needs to examine 5 domains to guarantee 99.22% coverage of the true domain and 13 domains for 100% coverage.

To evaluate the efficiency of predictions, we measure the average response time. Our system is hosted on a standard Linux Redhat platform with Dual Xeon IV 2.4GHz processors and 2GB RAM. Figure 6(a) shows that the response time of prediction, including feature extractions, itemset generations, and the ranked scores computation. When the protein size increases, it demands more computational resource to extract features on larger distance matrices. This reflects the gaps between two curves in Figure 6(a), where the top curve reports the response time with feature extraction and the bottom curve depicts the response time for computing scores and ranking domains. On the average, predicting an unknown protein to a SCOP domain takes 6.34 seconds. Comparing to a well-recognized structural alignment algorithm, CE,[22] on the same testing data, we conduct pairwise structural alignments for 1 against 7,701 proteins using the *Leave-One-Out* strategy. The SCOP domain of protein with the highest score is specified as the predicted result. We find that CE predicts SCOP domains of all testing proteins correctly. However, pairwise alignments using CE take 15,461.29 seconds. Sacrificing supportable accuracy, our algorithm runs near 2,439 times faster than the CE algorithm. Even though computer algorithms present high prediction accuracy in empirical results, classification by human experts is still believed to be more reliable. Instead of replacing manual classifications, our proposed method assists human experts to make the task of SCOP domain classification achievable and efficient.

In addition, our method is able to predict the SCOP fold of an unknown protein structure from the predicted domain by referencing the known mapping information between domain and fold. For the fold level, our approach exhibits 94.47% prediction accuracy, which is higher than the accuracy of SCOP domain predictions, 91.27%. Due to *one-to-many* relationship between fold and domains, it has a chance to conclude correct folds from incorrectly predicted domains. Therefore, SCOP domain predictions are more challenging than predictions in fold level. For instance, a SCOP fold f_1 contains three domains, such as d_1, d_2, and d_3, respectively. Even though the algorithm predicts a testing protein of SCOP domain d_1 as d_2, the fold is still mapped to f_1. Since the standard testbed of SCOP fold predictions is not available at this moment, we briefly compare to a recent approach in terms of data size, precision, and response time. A prominent work called 3-step scheme(PA+CP+DALI)[1] reports 98.8% accuracy in fold prediction and the average response time is 24,501 seconds. It is noteworthy to mention that their experiments are



Address http://proteindbs.rnet.missouri.edu/Predict_2.php

The Superimposition of Protein Backbone Strctures (Query Protein Vs. SCOP Domain: D-xylose isomerase [51666])

CA GLN 256 ; 23.56

Ranking: 1	Ranking: 2
D-xylose isomerase [51666]	Calmodulin [47516]
Ranking: 3	Ranking: 4
70S ribosome functional complex [58121]	50S subunit [58125]
Ranking: 5	Ranking: 6
Photosystem II [58160]	Calcyclin (S100) [47479]

Statistical Information:
Sequence Alignment: Query Chain V.S PDB_ID:5XIN CHAIN:A of SCOP Domain: D-xylose isomerase [51666]

```
3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
V  Q  A  T  R  E  D  K  F  S  F  G  L  W  T  V  G  W  Q  A  R  D  A  F  G  D  A  T  R  T  A  L  D  P  V  E  A  V  H  K  L  A  E  I  G
V  Q  A  T  R  E  D  K  F  S  F  G  L  W  T  V  G  W  Q  A  R  D  A  F  G  D  A  T  R  T  A  L  D  P  V  E  A  V  H  K  L  A  E  I  G
3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92
A  Y  G  I  T  F  H  D  D  D  L  V  P  F  G  S  D  A  Q  T  R  D  G  I  I  A  G  F  K  K  A  L  D  E  T  G  L  I  V  P  M  V  T  T  N
A  Y  G  I  T  F  H  D  D  D  L  V  P  F  G  S  D  A  Q  T  R  D  G  I  I  A  G  F  K  K  A  L  D  E  T  G  L  I  V  P  M  V  T  T  N
48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92
93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137
L  F  T  H  P  V  F  K  D  G  G  F  T  S  N  D  R  S  V  R  R  Y  A  I  R  K  V  L  R  Q  M  D  L  G  A  E  L  G  A  K  T  L  V  L  W
L  F  T  H  P  V  F  K  D  G  G  F  T  S  N  D  R  S  V  R  R  Y  A  I  R  K  V  L  R  Q  M  D  L  G  A  E  L  G  A  K  T  L  V  L  W
```

MageJava.1.0.011207 (1.0 methods), hit 'h' key for help

2005年5月9日

(a) (b)

Figure 6. (a)Average response times to predict SCOP domains with various protein chain sizes (b) The publicly available domain prediction system based on this our prediction model.

conducted on a comparably small testing set (600 proteins) from 15 SCOP folds.

5. Conclusion

Our automatic SCOP domain ranking and prediction algorithm accelerates the processes of structural recognition for newly discovered proteins. In this paper, we introduce an advanced algorithm to convert high-level features of distance matrices into itemsets for rule mining. The advantage of this KDD approach is to effectively reveal the hidden knowledge from similar protein tertiary structures for ranking and predicting possible SCOP domains. Although a multi-variate decision tree might be able to give comparable performance in classification and response time, the tree approach normally could not provide reasonable ranking results that are more valuable in the real world setting, as discussed previously. From the experimental results, our method can achieve reasonably high prediction performance in both accuracy and efficiency. To extend the scope of SCOP domain predictions, one possible direction is to computationally analyze text-based gene annotations, especially the passages related to gene functions, from structurally similar proteins.

To provide a tool for the research community, we have implemented a web-based interface to predict possible SCOP domains for unknown protein structures. Users are allowed to upload a protein file that follows PDB ATOM format. In Figure 6(b), the superimposition view shows that the query protein is structurally similar to a protein $5xin\_A$ from the top ranked SCOP domain *D-xylose isomerase*. Our system is publicly accessible at http://ProteinDBS.rnet.missouri.edu/Predict.php.

References

1. Z. Aung and K.-L. Tan. Clasifying Protein Folds using Multi-level Information of Protein Strutures. *The Third Asia Pacific Bioinformatics Conference SIG-Structure Meeting*, 2005.
2. R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on knowledge and data engineering*, 5(6):914–925, 1993.

3. T. Can, O. Camoglu, A.K. Singh, and Y.F. Wang. Automated Protein Classification Using Consensus Decision. *Proc. of the 3rd Int. IEEE Comput. Soc. Comput. Syst. Bioinformatics Conference*, 224–235, 2004.
4. R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. *Proc. of the ACM SIGKDD Int. conference on Knowledge discovery and data mining*, 69–78, 2004.
5. S.K. Chang and T.L. Kunii. Pictorial dataBase systems. *IEEE Computer*, 14:13–21, 1981.
6. S. Cheek, Y. Qi, S.S. Krishna, L.N. Kinch, and N.V. Grishin. SCOPmap: Automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics*, 5(1):197–197, 2004.
7. P.H. Chi, G. Scott, and C.R Shyu. A fast protein structure retrieval system using image-based distance matrices and multidimensional index. *Int. J. of Softw. Eng. and Know. Eng.*, 15(3),527–545, 2005.
8. M.H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall, New Jersey, USA, 164–192, 2003.
9. A. Godzik. The structural alignment between two proteins: Is there a unique answer?*Protein Sci.*, 5:1325–1338, 1996.
10. R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. on Syst., Man, and Cybernetics*, SMC-3:610–621, 1973.
11. T.F. Havel, I.D. Kuntz and G.M. Crippen. The theory and practice of geometry. *Bull. Math. Biol.*, 45:665–720, 1983.
12. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1993.
13. L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–602, 1996.
14. M. Leslie. Protein Matchmaking. *Science*, 305:1381, 2004.
15. B. Liu, W. Hsu, Y. Ma. Integrating Classification and Association Rule Mining. *Proc. of the Fourth Int. Conference on Knowledge Discovery and Data Mining*, 80–86, 1998.
16. R. Kolodny and N. Linial. Approximate protein structural alignment in polynomial time. *Proc. Natl. Acad. Sci.*, DOI:10.1073/pnas.0404383101, 12201–12206, 2004.
17. A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP:a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*,247:536–540,1995.
18. C.A. Orengo, F.M.G. Pearl, J.E. Bray, A.E. Todd, A.C. Martin, L. Lo Conte, and J.M. Thornton. The CATH Database provides insights into protein structure/function relationships. *Nucl. Acids. Res.*, 27(1):275–279, 1999.
19. N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. on Syst., Man, and Cybernetics*, SMC-9:62–66, 1979.
20. A. Rosenfeld and A.C. Kak. Digital picture processing. Academic Press, New York, 1982.
21. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern and Machine Intell.*, 2:1349–1380,2000.
22. H.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, 9:739–747, 1998.
23. A.W.M. Smeulders, T.S. Huang, T. Gevers. Special issue on content-based image retrieval. *Int. J. Computer Vision*, 56:5–6, 2004.
24. C.R. Shyu, P.H. Chi, G. Scott, and D. Xu. ProteinDBS - A content-based retrieval system for protein structure databases. *Nucl. Acids. Res.*, 32:w572–575, 2004.
25. J.R Quinlan. C4.5 Programs for Machine Learning. *Morgan Kauffman*, 1993.
26. M.J. Zaki, S. Jin, C. Bystroff. Mining Residue Contacts in Proteins Using Local Structure Predictions. *IEEE Trans. on Syst., Man, and Cybernetics*, 33(5):789–801, 2003.
27. T.I. Zarembinski, L.W. Hung, H.J. Mueller-Dieckmann, K.K. Kim, H. Yokota, R. Kim and S.H. Kim. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Sci. USA*, 95:15189–15193, 1998.