Computing Motif Correlations in Proteins

JORNG-TZONG HORNG,[1] HSIEN-DA HUANG,[1] SHIH-HSIEN WANG,[1] MING-YOU CHEN,[1]
SHIR-LY HUANG,[2] JENN-KANG HWANG[3]

[1]*Department of Computer Science and Information Engineering, National Central University, Taiwan*

[2]*Department of Life Science, National Central University, Taiwan*

[3]*Department of Biological Science and Technology, National Chiao-Tung University, Taiwan*

Received 24 January 2003; Accepted 28 January 2003

Abstract: Protein motifs, which are specific regions and conserved regions, are found by comparing multiple protein sequences. These conserved regions in general play an important role in protein functions and protein folds, for example, for their binding properties or enzymatic activities. The aim here is to find the existence correlations of protein motifs. The knowledge of protein motif/domain sharing should be important in shedding new light on the biologic functions of proteins and offering a basis in analyzing the evolution in the human genome or other genomes. The protein sequences used here are obtained from the PIR-NREF database and the protein motifs are retrieved from the PROSITE database. We apply data mining approach to discover the occurrence correlations of motif in protein sequences. The correlation of motifs mined can be used in evolution analyses and protein structure prediction. We discuss the latter, i.e., protein structure prediction in this study. The correlations mined are stored and maintained in a database system. The database is now available at **http://bioinfo.csie.ncu.edu.tw/ProMotif/.**

 J Comput Chem 24: 2032–2043, 2003

Key words: protein; motif; structural genomics; data mining; database

Introduction

The essence of structural genomics is to start from the gene sequence, produce the proteins, and determine their 3D structure. The challenge, once the sequence is determined, is to extract useful biologic information about the biochemical and biologic role of the protein in the organism. People always use motifs to identify distant relationships in novel sequences and hence for inferring protein function. The information about domains may be useful knowledge to structural genomics. The domain sharing in proteins can be used to analyze the evolution in the human genome or other genomes.

How can we obtain information about domain sharing of proteins? There are a lot of motif databases, and each database provides the information of domain and the tool for scanning domain on proteins. One can simply make use of the provided information, i.e., researchers may be using such system to know which domain the protein has. However, the domain information we gain is based on a protein or the protein information is based on a specific domain. There is no information about the domain sharing in proteins.

Protein family means proteins that can be grouped on the basis of similarities of their sequences; domain means proteins belonging to a particular family in general share functional attributes and are derived from a common ancestor. Motif means homologous sequences gathered together in multiple alignments with gap insertion or deletion. Motifs are usually expressed using regular expression or profile matrix. These conserved regions, motifs, reflect the core structural or functional elements of the protein. Motifs are typically from 10–20 residues and are sometimes referred to as "blocks," "segments," and "features." It is based on the observation that there are a huge number of different proteins, which can be grouped into a limited number of families based on the local similarities in their sequences. Proteins or protein domains belonging to a particular family in general share functional attributes and are derived from a common ancestor.

It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These regions are in general important for the function of a protein and/or for the maintenance of its 3D structure. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain that distinguishes its members from all other unrelated

Correspondence to: J.-T. Horng; e-mail: horng@db.csie.ncu.edu.tw

Table 1. Amount of Proteins in Different Organisms in PIR-NREF.

Taxonomy	Number of proteins
Homo sapiens	76,880
Human immunodeficiency virus type 1	56,310
Mus musculus	42,590
Arabidopsis thaliana	40,218
Caenorhabditis elegans	24,733
Drosophila melanogaster	24,074
Escherichia coli	19,324
Rattus norvegicus	11,290
Saccharomyces cerevisiae	9449
Oryza sativa	9229
Streptomyces coelicolor	8720
Agrobacterium tumefaciens	8518
Mesorhizobium loti	7312
Hepatitis C virus	7185
Pseudomonas aeruginosa	6810
Sinorhizobium meliloti	6701
Schizosaccharomyces pombe	6637
Mycobacterium tuberculosis	6332
Nostoc sp. PCC 7120	6232
Salmonella enterica	6064

proteins. A pertinent analogy is the use of fingerprints by the police for identification purposes. A fingerprint is in general sufficient to identify a given individual. Similarly, a protein signature can be used to assign a newly sequenced protein to a specific family of proteins and thus to formulate hypotheses about its function.

A motif database becomes a vital tool for identifying distant relationships in novel sequences and hence for inferring protein functions. Several pattern recognition methods have evolved to address different sequence analysis problem, resulting in different and independent databases. Profiles are supposed to be more sensitive and more robust than patterns because they provide discriminatory weights not only for the residues already found at a given position of a motif but also for those not yet found. Diagnostically, these resources have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. For example, regular expressions are likely to be unreliable in the identification of members of highly divergent superfamilies; fingerprints perform relatively poorly in the diagnosis of short motifs; and profiles and hidden Markov models (HMMs) are less likely to give specific subfamily diagnoses.[1]

Pfam[2] is a large collection of multiple sequence alignments and profile HMMs. Pfam families correspond with structural domains and improve domain-based annotation. The definition of domain boundaries, family members, and alignment is done semiautomatically based on expert knowledge, sequence similarity, other protein family databases, and the ability of HMM profiles to correctly identify and align the members.[3] Meta-MEME[4] is a software tool for creating HMMs that focus on highly conserved regions, called motifs. Junier et al.[5] developed a search program that can look for the arrangement of motifs specified by users. The advantage of this approach is that it does not depend on a particular motif algorithm.

The ProMotif database proposed in this study is different from these because the approach discovers the occurrence of associations of motifs in a set of protein sequences by applying data mining techniques. The approach searches the comprehensive protein database and find a set of proteins that contains a motif correlation. These proteins are considered together for further analysis by having the same motif occurrence correlation. Because domains of proteins are in general important for the function and/or the maintenance of its 3D structure, the correlation of domain family of protein may be a basis for structural genome or functional evolutionary analysis. We want to discover the correlation of domain family in proteins to provide a comprehensive description of domain family relationships and facilitate knowledge discovery.

The PIR-NREF[6] is a nonredundant reference protein database designed to provide a timely and comprehensive collection of all protein sequence data, keeping pace with the genome sequencing

Table 2. Unspecific Motifs in PROSITE.

Doc ID	Description	Pattern
PDOC00001	*N*-glycosylation site	N-{P}-[ST]-{P}
PDOC00002	Glycosaminoglycan attachment site	S-G-x-G
PDOC00003	Tyrosine sulfation site	RULE
PDOC00004	CAMP- and cGMP-dependent protein kinase	[RK](2)-x-[ST]
PDOC00005	Protein kinase C phosphorylation site	[ST]-x-[RK]
PDOC00006	Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
PDOC00007	Tyrosine kinase phosphorylation site	[RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y
PDOC00008	*N*-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
PDOC00009	Amidation site	x-G-[RK]-[RK]
PDOC00015	Bipartite nuclear targeting sequence	RULE
PDOC00016	Cell attachment sequence	R-G-D
PDOC00017	ATP/GTP-binding site motif A (P-loop)	[AG]-x(4)-G-K-[ST]
PDOC00029	Leucine zipper pattern	L-x(6)-L-x(6)-L-x(6)-L
PDOC00266	Prenyl group binding site (CAAX box)	C-{DENQ}-[LIVM]-x>

Table 3. Data Statistics in PROSITE and PIR-NREF.

Description	Number of records
Number of protein in PIR-NREF beta-release	885,514
Documentation in PROSITE	1121
Motif in PROSITE	1517
Signature in Motif (pattern)	1329
Profile in Motif (matrix)	184
Protein with at least two distinct motifs in PROSITE	46,699

projects and containing source attribution and minimal redundancy. PROSITE[7] is a database of protein families and domains. It consists of biologically significant sites, patterns, and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. PROSITE currently contains patterns and profiles specific for more than 1000 protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins. The Protein Data Bank (PDB)[8] is the single worldwide repository for the processing and distribution of 3D biologic macromolecular structure data. The SCOP[9] database provides a detailed and comprehensive description of the relationships of all known protein structures. The distinction between evolutionary relationships and those that arise from the physics and chemistry of proteins is a feature that is unique to this database. Functional similarity is implied by an evolutionary relationship but not necessarily by a physical relationship. The CE database[10] is derived using the combinatorial extension (CE) algorithm,[11] which compares pairs of protein polypeptide chains and provides a list of structurally similar proteins along with their structure alignments. Using CE, the longest alignment path is evaluated for statistical significance, represented as a *z-score.* This is done by evaluating the probability of finding an alignment path of the same length with the same or smaller number of gaps and distance from a random comparison of structures using a nonredundant set.[12]

L_1 = {large 1 - sequences};
for ($k = 2$; $L_{k-1} \neq 0$; k ++) **do**
  begin
    C_k = New candidates generated from L_{k-1}
    foreach organism - structure sequence s in the database **do**
      Increment the count for all candidates in C_k
        that are contained in c.
    L_k = Candidates in C_k with minimum support.
  end
Answer = Maximal Sequences in $\bigcup_k L_k$;

Notations:
L_k denotes the set of large k - sequences, and C_k the set of candidate k - sequences.

Figure 1. Apriori algorithm.

To face large data, data mining plays a prominent role in knowledge extraction. The enormous number of sequenced genomes, gene identification data, gene expression experimental profiles, and genes categorized in functional classes allows the use of computational techniques to investigate transcriptional regulatory elements in the gene promoter regions and deciphers the mechanisms of gene transcriptional regulation. Frequently used data mining approaches include association rules, statistics, neural networks, clustering, classification, and genetic algorithms, etc. Strikant and Agrawal[13] introduced the problem of mining association rules over basket data. The data mining techniques might mine an enormous number of associations. The enormous number of associations makes it extremely difficult to identify those useful or interesting ones. The chi-square test is one of the approaches to remove insignificant ones. In statistics, chi-square test statistics (χ^2) are extensively applied for testing independence and correlation.[14]

We developed a system to find correlations of domain sharing in proteins. The protein sequences are from the PIR-NREF database and the motifs are from the PROSITE database. The Aproiori algorithm[15] is applied to mine the association of functional domain sharing in protein structures. A friendly user interface to display the mining results is also provided.

Table 4. Partial Rules of Domain Sharing Mined in Proteins.

ID	Hit proteins (ratio %)	Confidence (%)	Body	➜	Head
MC00037	254 (0.5439)	43.72	[PDOC00189] Mitochondrial energy transfer proteins signature	=>	[PDOC00013] Prokaryotic membrane lipoprotein lipid attachment site
MC00038	253 (0.5418)	89.08	[PDOC00559] G-protein coupled receptors family 2 signatures and profiles	=>	[PDOC00013] Prokaryotic membrane lipoprotein lipid attachment site
MC00039	234 (0.5011)	74.05	[PDOC50215] ADAM type metalloprotease domain profile	=>	[PDOC00129] Neutral zinc metallopeptidases, zinc-binding region signature
MC00040	229 (0.4904)	92.71	[PDOC00032] "Homeobox" antennapedia-type protein signature	=>	[PDOC00027] "Homeobox" domain signature and profile
MC00041	222 (0.4754)	47.13	[PDOC00128] Eukaryotic and viral aspartyl proteases signature and profile	=>	[PDOC00013] Prokaryotic membrane lipoprotein lipid attachment site
MC00042	219 (0.469)	98.65	[PDOC00035] "POU" domain signatures	=>	[PDOC00027] "Homeobox" domain signature and profile

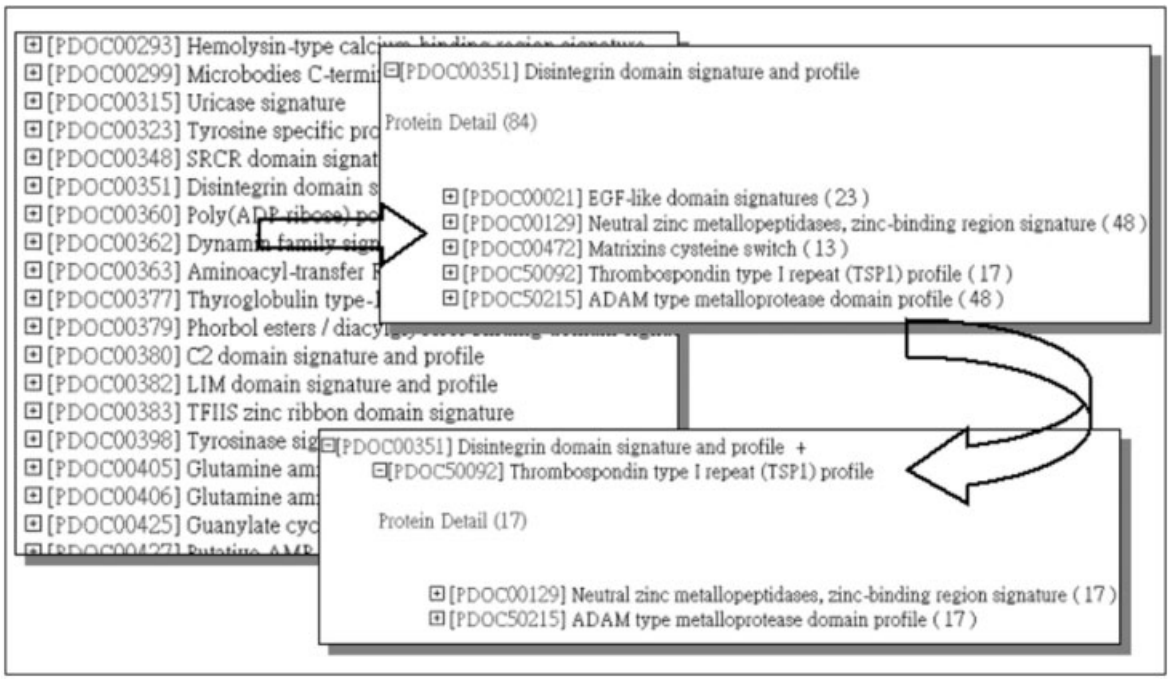

Figure 2. Tree-like interface to show the motif correlation.

Materials and Methods

The protein information and motifs are retrieved from the PIR-NREF[6] and PROSITE databases[7] respectively. The PIR-NREF is a nonredundant reference protein database designed to provide a timely and comprehensive collection of all protein sequence data, keeping pace with the genome sequencing projects and containing source attribution and minimal redundancy. PIR-NREF (release 11-Mar-2002) contains 885,515 entries. PROSITE (release 17) contains 1121 documentation entries that describe 1517 different patterns, rules, and profiles/matrices. The PROSITE database consists of biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can help determine to which known family of proteins a new sequence belongs or which known domain(s) it contains. PROSITE currently contains signatures specific for about 1000 protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins.

The amount of proteins of different organisms stored in PIR-NREF are calculated and listed in Table 1 for more than 6000 entries. The table shows only organisms that have more than 6000 entries. The motifs are then located into all protein sequences to investigate the possible occurrences of protein domains.

Several motifs in PROSITE are considered too unspecific to recognize the protein domains, that is, such motifs are not representative in proteins. For example, the pattern of the "protein

Table 5. Proteins Contain the Motif Correlation.

NREF_ID	Taxonomy	Name	Length	Related PDB_ID
[NF00410087]	*Crotalus adamanteus*	Adamalysin II	201	1IAG; 2AIG:P; 3AIG
[NF00410088]	*C. adamanteus*	Adamalysin II (EC 3.4.24.46) (proteinase II)	203	
[NF00410096]	*C. adamanteus*	Adamalysin (EC 3.4.24.46) II	203	1IAG; 2AIG:P; 3AIG; 4AIG
[NF00410093]	*C. adamanteus*	Adamalysin II	202	1IAG; 4AIG
[NF00410091]	*C. adamanteus*	Adamalysin II (proteinase II) (E.C. 3.4.24.4)	202	2AIG:P; 3AIG; 4AIG
[NF00410105]	*C. atrox*	Atrolysin C (EC 3.4.24.42) precursor (Ht-c)	414	
[NF00410107]	*C. atrox*	Atrolysin C	202	1DTH:A; 1DTH:B
[NF00410115]	*C. atrox*	Vascular apoptosis-inducing protein 1	610	
[NF00410111]	*C. atrox*	Catrocollastatin precursor	609	
[NF00410117]	*C. atrox*	Atrolysin B (EC 3.4.24.41) precursor	414	

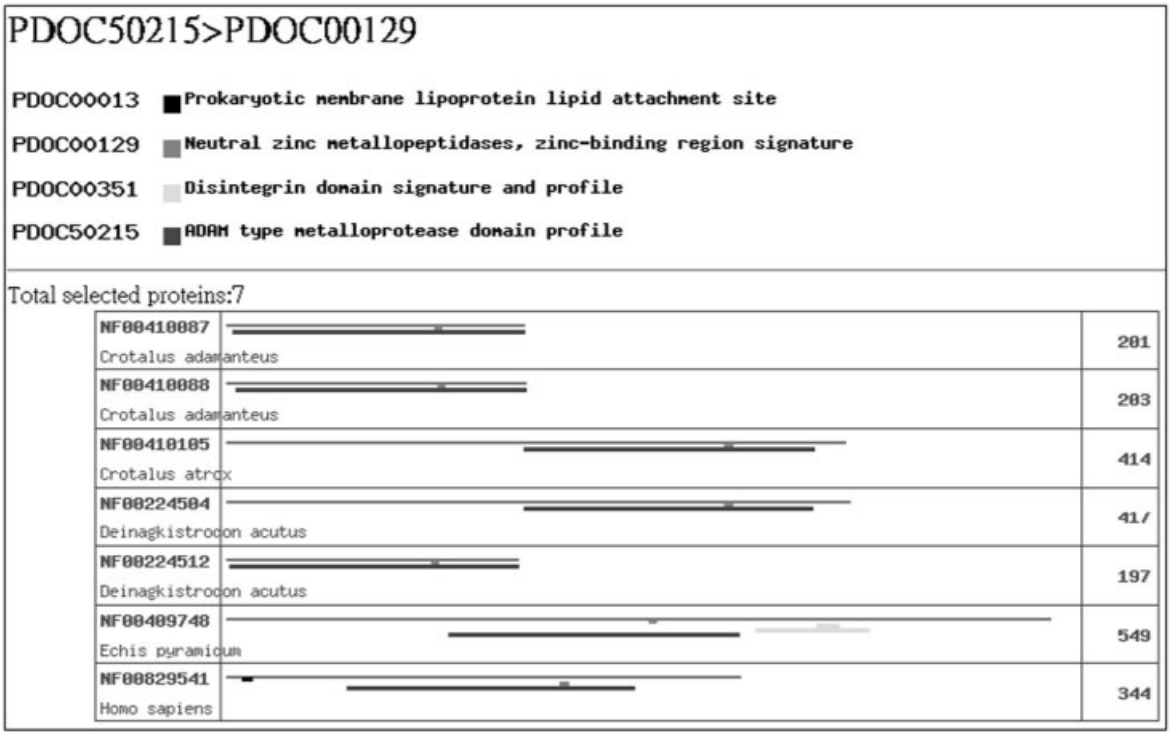

Figure 3. Interface to show the correlations of protein motifs.

kinase C phosphorylation site" in Table 2 is "[ST]-x-[RK]." The pattern is too unspecific so that many proteins are recognized by the motif pattern. Partial unspecific motifs in PROSITE are shown in Table 2. We filter out these short motifs so that our mining result will not be misled.

Each protein in PIR-NREF is mapped to a transaction and motifs in PROSITE are mapped to items of transactions. We then apply a data mining approach to generate associations.

The PROSITE database contains two major types of motifs, i.e., pattern type and matrix type. The pattern is developed in regular

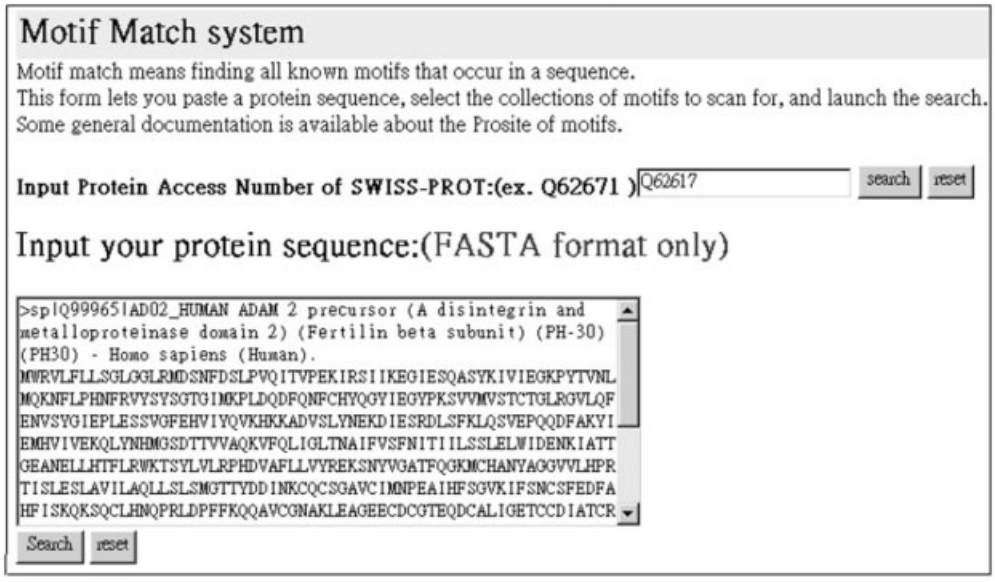

Figure 4. Interface of "motif match" function.

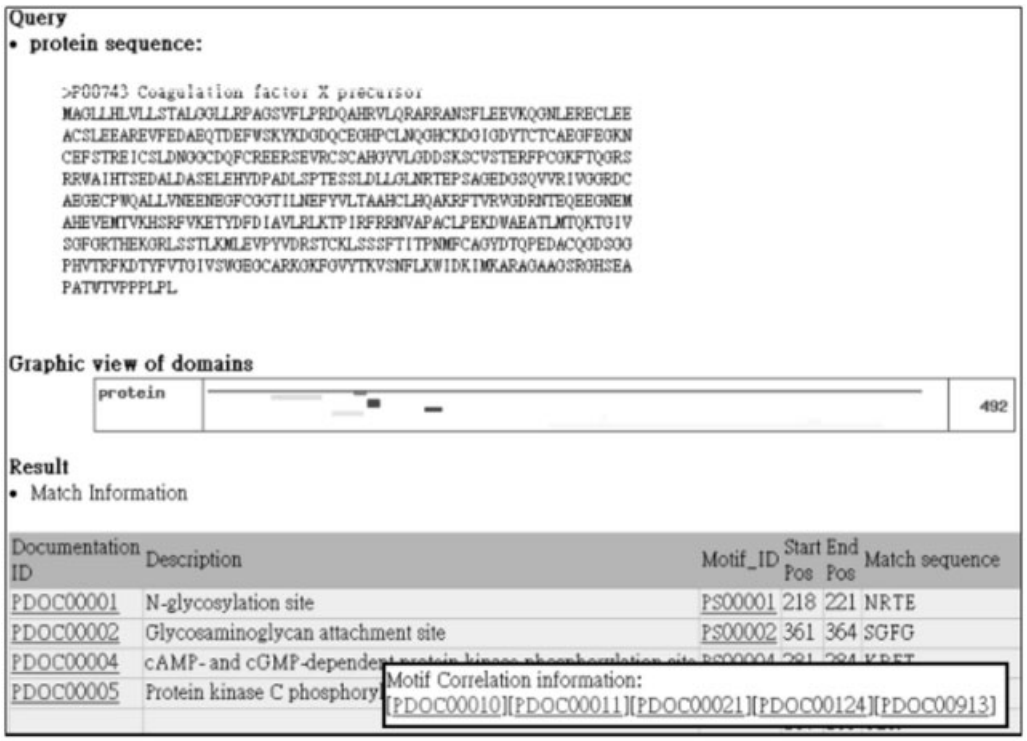

Figure 5. Interface to show query results.

Table 6. Partial Information About 49 Proteins Contained in the Correlation of Four Domains Including "[PDOC00021] EGF-Like Domain Signatures," "[PDOC00010] Aspartic Acid and Asparagine Hydroxylation Site," "[PDOC00124] Serine Proteases, Trypsin Family, Signatures, and Profile," and "[PDOC00913] Calcium-Binding EGF-Like Domain Signature.

NREF ID	Taxonomy	Protein name
[NF00159967]	*Bos taurus*	Prepro-factor X
[NF00163582]	*B. taurus*	Coagulation factor IX (EC 3.4.21.22) (Christmas factor)
[NF00142713]	*Canis familiaris*	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor)
[NF00050443]	*Gallus gallus*	Coagulation factor X precursor (EC 3.4.21.6) (Stuart factor) (virus activating protease) (VAP)
[NF00094351]	*Homo sapiens*	Vitamin K-dependent protein C precursor (EC 3.4.21.69) (autoprothrombin IIA) (anticoagulant protein C) (blood coagulation factor XIV)
[NF00096102]	*H. sapiens*	Coagulation factor X precursor (EC 3.4.21.6) (Stuart factor)
[NF00099200	*H. sapiens*	Factor X peptide
[NF00113984]	*H. sapiens*	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor)
[NF00124360]	*H. sapiens*	Coagulation factor IX; coagulation factor IX (plasma thromboplastic component); factor 9; factor IX; Christmas factor
[NF00128876]	*H. sapiens*	Coagulation factor X precursor
[NF00130101]	*H. sapiens*	Coagulation factor IX
[NF00101550]	*H. sapiens*	Factor VII active site mutant immunoconjugate
[NF00499028]	*M. musculus*	Coagulation factor VII
[NF00517822]	*M. musculus*	Anticoagulant protein C
[NF00515499]	*M. musculus*	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (fragment)
[NF00507321]	*M. musculus*	Coagulation factor X precursor (EC 3.4.21.6)
[NF00511169]	*M. musculus*	Coagulation factor X
[NF00511824]	*M. musculus*	Coagulation factor X precursor
[NF00506394]	*M. musculus*	Coagulation factor VII precursor (EC 3.4.21.21) (serum prothrombin conversion accelerator)
[NF00057391]	*Ornithorhynchus anatinus*	Coagulation factor X
[NF00166611]	*Oryctolagus cuniculus*	Coagulation factor VII
[NF00072237]	*Pan troglodytes*	Coagulation factor XI
[NF00563891]	*Rattus norvegicus*	Coagulation factor Xa (EC 3.4.21.6) precursor

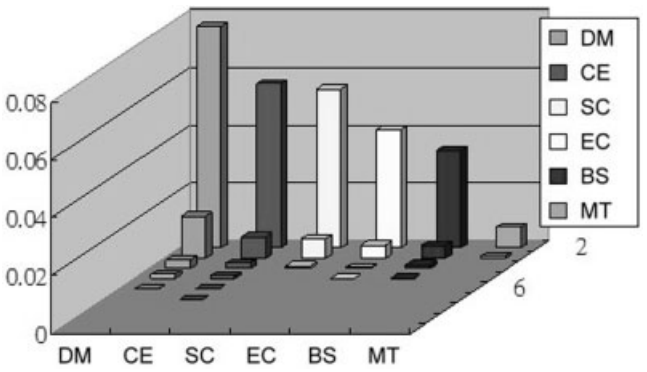

Figure 6. Motif comparison of different organisms.

expression technique and the matrix is developed in the HMM model. Each of the entries in PROSITE is fully documented. The documentations are including a concise description of the protein family that is designed to detect, as well as a summary of the reasons leading to the development of the pattern or profile. Because there are two types of motifs, we must distinguish between them.

Protein structures are provided in the PDB database.[8] To find the protein structure mapping information of unknown structure proteins, we should find some method to find the relation between known structure PDB information and unknown PIR-NREF protein sequences. The method we used here is the global alignment using FASTA provided by the PIR-NREF database. The PIR-NREF database contains a lot of protein database resources, including SWISS-PROT, PIR-PSD, TrEMBL, and the PDB database. It uses the all-against-all FASTA search to check entries for data redundancy. Because it contains PDB data resources, PIR-NREF uses all-against-all FASTA search to minimize the data redundancy and also provide the all-against-all sequence similarity information. The information it provides about sequence similarity is used here for the relationship between protein sequence and protein structure. After parsing the PIR-NREF database and PDB sequence data, we build the relation between the PDB structure database and PIR-NREF sequence database, and then use the known PDB structure segment to infer the unknown structure of proteins. Further, if the structure is known we are to display on the Web interface and show the related motif position. The information can help biologists know the related position in protein structures.

When we consider the mining association rule algorithm, the transaction with only one item or none is filtered out in applying association rule mining. This is because such a case does not have any correlations of motifs in proteins. In our statistics, around 268,363 proteins have only one domain. These proteins were not taken to be mined for the correlation of domain sharing. Basic statistics of the source data used in this study are shown in Table 3.

Table 7. Pairwise Protein Structure Similarity and Sequence Identity in a Same Group of Motif Correlations.

No.	Rule body => Rule head	Support (%)	Confidence (%)	Total pairs	Sequence identity >20%	Sequence identity < 20% (z-score > 4.0)	Sequence identity < 20% (z-Score < 4.0)
1	PDOC00275 => PDOC00018	6.8	100	253	253	0	0
2	PDOC00183 => PDOC00793	5.92	100	190	190	0	0
3	PDOC00793 => PDOC00183	5.92	100	190	190	0	0
4	PDOC00252 => PDOC00124	5.03	100	136	133	0	3
5	PDOC00398 => PDOC00184	3.25	100	55	55	0	0
6	PDOC00184 => PDOC00398	3.25	100	55	55	0	0
7	PDOC00151 => PDOC00152	3.25	100	55	55	0	0
8	PDOC00152 => PDOC00151	3.25	100	55	55	0	0
9	PDOC00010 => PDOC00913	3.25	100	55	43	1	11
10	PDOC00021+PDOC00913 => PDOC00010	3.25	100	55	43	1	11
11	PDOC00010 => PDOC00021	3.25	100	55	43	1	11
12	PDOC00254 => PDOC00124	3.25	100	55	27	17	11
13	PDOC00010+PDOC00913 => PDOC00021	3.25	100	55	43	1	11
14	PDOC00913 => PDOC00021	3.25	100	55	43	1	11
15	PDOC00010+PDOC00021 => PDOC00913	3.25	100	55	43	1	11
16	PDOC00913 => PDOC00010	3.25	100	55	43	1	11
17	PDOC00018+PDOC50007 => PDOC00380	2.96	100	45	45	0	0
18	PDOC00380 => PDOC50007	2.96	100	45	45	0	0
19	PDOC50215 => PDOC00129	2.96	100	45	38	0	7
20	PDOC00018+PDOC00380 => PDOC50007	2.96	100	45	45	0	0

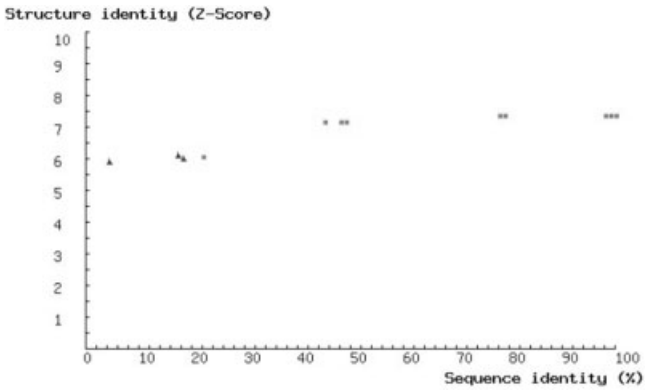

Figure 7. Distribution of protein sequence identity and structure similarity in a protein group with the motif correlation "PDOC50215 => PDOC00129."

In the following, we describe how to mine associations from the combinations of motifs in protein sequences. Consider a large database with transactions, where each transaction consists of a set of items. An association rule is an expression as $A => B$, where A and B are the sets of items. The mining of an association rule is that a transaction in the database that contains A also tends to contain B. For example, 90% of the people who purchase beer also purchase diapers. Here, 90% is called the confidence of the rule. The support of the rule $A => B$ given here in is the percentage of transactions that contain both A and B.

The formal statement of the problem is described below. Let $R = \{r_1, r_2, \ldots, r_n\}$ be a set of motifs from PROSITE. The sets R is called "item set." Let $G = \{g_1, g_2, \ldots, g_m\}$ be a group of proteins with at least two motifs. Each protein is mapped to a transaction containing a set of protein motifs, also called items.

Assume that a protein sequence region S *contains* A, a set of items of I, if $A \subseteq S$. An *association rule* is an implicate of the form $A => B$, where $A \subset I$, $B \subset I$, and $A \cap B = 0$. The rule $A => B$ holds in the set of proteins D with *confidence conf* if $c\%$ of transactions in D contains A and also B. The rule $A => B$ has *support sup* in the motif set D if $s\%$ of proteins in D contained $A \cup B$. In our experiments, the minimum support is set to 0.08%. The association rules are generated if the rule has a higher support and confidence than the user specified. The Apriori and AprioriTid algorithms are then applied to mine association rules. The Apriori algorithm is given in Figure 1.

A huge amount of motif occurrence associations is found in protein sequences and the chi-square test is then used to investigate the correlations of motifs in each association. The motifs in any association are occurring significantly in the proteins if the calculation of chi-square value exceeds the threshold value of 3.84 (with degree of freedom 1 and $\alpha = 0.05$). In statistics, the chi-square test (χ^2) is a widely used method for testing independence and (or) correlation and is applied to discover the significant associations rules in ref. 14. Let f_0 be an observed frequency and f an expected frequency; the chi-square test is used to measure the significance of the deviation from the expected values. The value of χ^2 is defined as $\chi^2 = \Sigma ((f_0 - f)^2/f)$.

Results

The mining results are a set of correlations. Each correlation represents the fact that a group of domains usually appears together in a set of proteins. An example of the correlation of motifs mined is shown in Table 4. The first column "ID" indicates the identifier of the motif correlation in our proposed database, i.e., ProMotif. The "Hit Proteins" column indicates the number of

Table 8. Structure Similarity and Sequence Identity of Each Protein Pair.

Protein pair	*z*-score	Sequence identity
1atl_A : 1htd_A	7.4	100
1atl_A : 1dth_A	7.4	100
3aig_ : 4aig_	7.4	100
1htd_A : 1dth_A	7.4	100
2aig_P : 4aig_	7.4	100
3aig_ : 2aig_P	7.4	99
1bsw_A : 1bud_A	7.4	99
1iag_ : 3aig_	7.4	98
1iag_ : 2aig_P	7.4	98
1iag_ : 4aig_	7.4	98
1atl_A : 3aig_	7.4	79
1atl_A : 4aig_	7.4	79
1atl_A : 2aig_P	7.4	79
1htd_A : 3aig_	7.4	79
1htd_A : 4aig_	7.4	79
1htd_A : 2aig_P	7.4	79
1dth_A : 4aig_	7.4	79
1dth_A : 2aig_P	7.4	79
1dth_A : 3aig_	7.4	79
1atl_A : 1iag_	7.4	78
1htd_A : 1iag_	7.4	78
1dth_A : 1iag_	7.4	78
3aig_ : 1bsw_A	7.2	49
3aig_ : 1bud_A	7.2	49
2aig_P : 1bud_A	7.2	49
4aig_ : 1bsw_A	7.2	49
4aig_ : 1bud_A	7.2	49
2aig_P : 1bsw_A	7.2	49
1iag_ : 1bsw_A	7.2	48
1iag_ : 1bud_A	7.2	48
1atl_A : 1bsw_A	7.2	45
1htd_A : 1bud_A	7.2	45
1htd_A : 1bsw_A	7.2	45
1dth_A : 1bsw_A	7.2	45
1dth_A : 1bud_A	7.2	45
1atl_A : 1bud_A	7.2	45
1bsw_A : 1bkc_A	6.1	22
1bud_A : 1bkc_A	6.1	22
1atl_A : 1bkc_A	6.1	18
1htd_A : 1bkc_A	6.1	18
1dth_A : 1bkc_A	6.1	18
3aig_ : 1bkc_A	6.2	17
4aig_ : 1bkc_A	6.2	17
2aig_P : 1bkc_A	6.2	17
1iag_ : 1bkc_A	6	4

proteins that satisfy the motif correlation, and the ratio in the bracket is the percentage of all proteins. The "Confidence" column shows the confidence of an association rule. The body and head parts of an association rule are shown in the fourth and fifth columns. For example, the third tuple, i.e., the ID of MC00039, in Table 4 means 234 proteins have the motif correlation of "ADAM type metalloprotease domain profile" and "Neutral zinc metallopeptidases, zinc-binding region signature" with the confidence value 74.05%. The value indicates that 74.05% of proteins that contain the motif PDOC50215 also contain the motif PDOC00129. The PDOC50215 and PDOC00129 are the IDs in PROSITE.[7]

A tree-like display interface is provided as shown in Figure 2. The tool provides a hierarchical user interface for browsing all the motif correlations mined. One only needs to expand the tree node when interested in some domain family. Each node represents a motif/domain correlation and expands the tree to show the related protein information.

Besides the correlation of motifs in proteins, we also find the proteins that satisfy the correlation. As the example mentioned above, 234 proteins are found to have the "[PDOC50215] ADAM type metalloprotease domain" and "[PDOC00129] Neutral zinc metallopeptidases, zinc-binding region." These 234 proteins are also found in different taxonomies including *Agkistrodon contortrix, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens,* and other taxonomies. Table 5 lists some proteins that satisfy the motif correlations, i.e., [PDOC50205] => [PDOC00129].

A graphical interface is designed and implemented to show the motif position of the related proteins. Using this interface, one can locate the positions of motifs in different proteins. For example, the protein sequence NF00410087 in Figure 3 can be located by the motifs of PDOC00129 and PDOC50215. The length of NF00410087 is 201 residues.

Besides, one further may want to see the related motifs in protein structures. According to the preprocessing of PIR-NREF all-against-all FASTA similarity searching, the relationships between the protein structures and protein sequences are maintained. Using this sequence similarity, we map the sequence to structure segment. A Web interface to display the motifs in protein structure space mapped by FASTA similarity search from PIR-NREF is shown in Table 5 by FASTA search. Many related PDB structures including PDB protein structures with IDs of "1IAG," "2AIG," and "3AIG" have high similarity against FASTA search with protein PIR-NREF ID [NF00410087]. The primary sequences of these PDB structures and the protein sequence, i.e., PIR-NREF ID [NF00410087], have a high similarity score.

For the biologist's convenience, we also provide a "motif match" system. If a biologist wants to know the additional information of a new protein sequence, he inputs the sequence in the "motif match" function to find the corresponding motif and its correlation information. The function is given in Figure 4. The query result of the example in Figure 4 is shown in Figure 5. The results include matching motif ID, motif start position, end position, and the matched sequence of the given protein.

Discussion and Conclusion

In this study, the correlations mined show that many proteins share more than one motif. Such motif correlation suggests that some

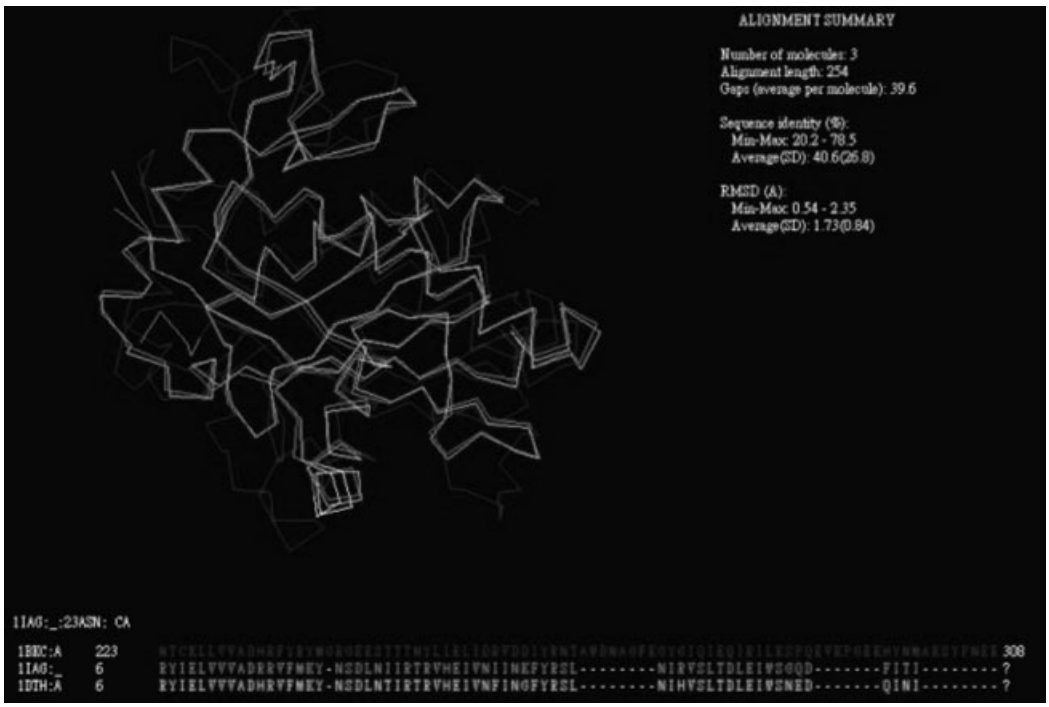

Figure 8. Structure alignment of three proteins.

Table 9. Properties of the Protein Structures with Their Properties Used in the Structure Alignment in Figure 8.

PDB ID	*z*-score	RMSD (Å)	Sequence (%)	Exp.	Title
1bkc_A	—	—	—	X-ray	Catalytic domain of Tnf-Alpha converting enzyme (Tace)
1iag_	6.0	2.33	4	X-ray	First structure of a snake venom metalloproteinase: a prototype for matrix metalloproteinases/collagenases
1dth_A	6.1	2.35	18	X-ray	Metalloprotease

RMSD, root mean square deviation.

biologic functions could be coupled. Such motif couplings should shed new light on the biologic mechanisms and pathways. As a specific example, a case of 49 proteins that contain 4 domains is gone into more detail to explain. These four domain types are "EGF-like domain," "Aspartic acid and asparagine hydroxylation site," "Serine proteases, trypsin family," and "Calcium-binding EGF-like domain," respectively. The accession numbers of these four domain/documentation types in PROSITE are: "[PDOC00021] EGF-like domain signatures," "[PDOC00010] Aspartic acid and asparagine hydroxylation site," "[PDOC00124] Serine proteases, trypsin family, signatures, and profile," and "[PDOC00913] Calcium-binding EGF-like domain signature," respectively. Table 6 lists partial information of these 49 proteins. There are similar proteins in different species. For example, "Coagulation factor IX precursor (EC 3.4.21.22)" is found in human [NF00113894], mouse [NF00515499], and dog [NF00142713]. Besides, we find different functions from their names. For example, [NF00101550] and [NF00094351] are annotated to "Factor VII active site mutant immunoconjugate" and "Vitamin-K dependent protein C precursor," respectively. It is a surprise to find they have different functions; however, they share the four domain types, i.e., "EGF-like domain," "Aspartic acid and asparagine hydroxylation site," "Serine proteases, trypsin family," and "Calcium-binding EGF-like domain." The information mined like above can be used to analyze evolution of the human genome or other genomes.

We further study the motifs in the proteins in several organisms. There are lots of proteins in PIR-NREF, and they allow us to survey and compare the set of domain family combinations present in the archaea, bacterial, and eukaryote taxonomy. The set of organisms we chose are diverse to cover much of the domain family and domain combinatorial space naturally. The organisms include one archaea (*Methanobacterium thermoautotrophicum,* MT), two eubacteria (*Escherichia coli,* EC; *Bacillus subtilis,* BS), one unicellular eukaryote (*Saccharomyces cerevisiae,* SC), and two multicellular eukaryotes (*C. elegans,* CE; *D. melanogaster,* DM). These organisms are from different external environments: For instance, their optimal temperatures range from room temperature (SC) to 85°C in deep marine subsurface oil areas (AF). These genomes cover multicellular and unicellular organisms with different modes of life, from autotrophs (MT) to optional parasites (EC, BS).

We retrieve the protein information of the six organism groups from PIR-NREF and compute the motif distribution in the chosen organism groups. The motif distribution of different groups is shown in Figure 6. From the statistics, we find that more complex organism groups usually have the combination of more than two motifs in proteins. This perhaps means that the biologic function of multicellular organisms is more complex than unicellular organisms.

We are also interested in the structure similarity of the proteins that contain the same motif correlation. As shown in Table 7, some of the motif correlations are mined from the protein sequences having their structures in the PDB database. For example, the first motif correlation "PDOC00275 => PDOC00018" is mined from 23 proteins including 2PSR, 3PSR_A, 1PSR_A, 1MR8_A, 1QLS_A, 1MHO, 1CNP_A, 1UWO_A, 1B4C_A, 1CFP_A, 1SYM_A, 2BCA, 1B1G_A, 4ICB, 3ICB, 1BOD, 2BCB, 1CDN, 1CLB, 1QLK_A, 1A03_A, 1CB1, and 1BOC. The protein sequences of each pair of proteins in the same group are aligned pairwisely by the sequence alignment tool ClustalW[16] and the sequence identities are also computed. At the same time, the CE algorithm[11] is also applied to each pair of protein structures in the same group to compute the structure similarity, represented by the *z*-score. The higher *z*-score means higher structure similarity.

For example, the total number of protein pairs in the motif correlation "PDOC50215 => PDOC00129" is 45, 38 protein pairs

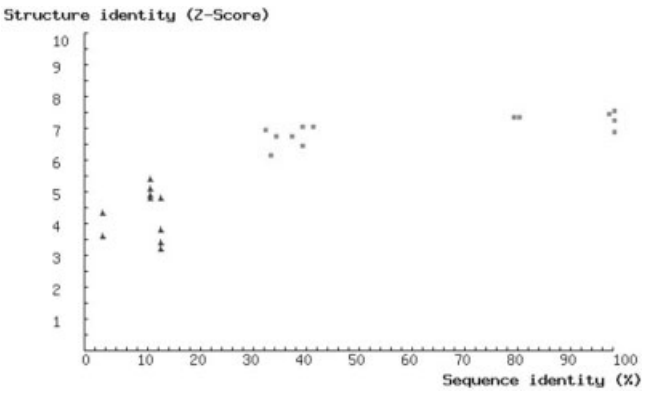

Figure 9. Distribution of protein sequence identity and structure similarity in a protein group with the motif correlation "PDOC00254 => PDOC00124."

Table 10. Structure Similarity and Sequence Identity of Each Protein Pair.

Protein pair	*z*-score	Sequence identity
1tbq_K : 1tbr_K	7.6	100
1sgp_E : 1sgq_E	7.3	100
1sgp_E : 3sgb_E	7.3	100
1sgp_E : 1sgr_E	7.3	100
1sgq_E : 1sgr_E	7.3	100
1sgr_E : 3sgb_E	7.3	100
1sgq_E : 3sgb_E	7.3	100
1cho_E : 1hja_B	6.9	100
1an1_E : 1ldt_T	7.5	99
1ldt_T : 1tgs_Z	7.4	82
1an1_E : 1tgs_Z	7.4	81
1an1_E : 1cho_E	7.1	43
1ldt_T : 1cho_E	7.1	43
1tgs_Z : 1cho_E	7.1	41
1an1_E : 1hja_B	6.5	41
1ldt_T : 1hja_B	6.5	41
1tgs_Z : 1hja_B	6.5	41
1an1_E : 1tbq_K	6.8	39
1ldt_T : 1tbr_K	6.8	39
1ldt_T : 1tbq_K	6.8	39
1an1_E : 1tbr_K	6.8	39
1tgs_Z : 1tbq_K	6.8	36
1tgs_Z : 1tbr_K	6.8	36
1tbq_K : 1hja_B	6.2	35
1tbr_K : 1hja_B	6.2	35
1cho_E : 1tbq_K	7	34
1cho_E : 1tbr_K	7	34
1tgs_Z : 1sgp_E	4.9	14
1tgs_Z : 1sgr_E	4.9	14
1tgs_Z : 3sgb_E	4.9	14
1tgs_Z : 1sgq_E	4.9	14
1tbr_K : 1sgp_E	3.9	14
1tbr_K : 1sgq_E	3.9	14
1tbr_K : 1sgr_E	3.9	14
1tbq_K : 1sgq_E	3.5	14
1tbq_K : 1sgp_E	3.3	14
1tbq_K : 3sgb_E	3.3	14
1tbr_K : 3sgb_E	3.3	14
1tbq_K : 1sgr_E	3.3	14
1cho_E : 1sgp_E	5.5	12
1cho_E : 3sgb_E	5.5	12
1cho_E : 1sgq_E	5.5	12
1cho_E : 1sgr_E	5.5	12
1an1_E : 1sgp_E	5.2	12
1ldt_T : 1sgp_E	5.2	12
1ldt_T : 1sgr_E	5.2	12
1ldt_T : 1sgq_E	5.2	12
1an1_E : 1sgq_E	5	12
1an1_E : 3sgb_E	5	12
1ldt_T : 3sgb_E	5	12
1an1_E : 1sgr_E	4.9	12
1hja_B : 1sgq_E	4.4	3
1hja_B : 1sgp_E	3.7	3
1hja_B : 1sgr_E	3.7	3
1hja_B : 3sgb_E	3.7	3

of which whose sequence identity are more than 20%. The other seven protein pairs are less than 20%. However, the structure similarity measures, i.e., *z*-score, are more than 4.0. Figure 7 shows the distribution of protein sequence identities and structure similarities of protein pairs in the same protein group. The x-axis denotes the *z*-score value, i.e., the structure similarity, and the y-axis represents the sequence identity. Each dot in the plane is a protein pair with corresponding sequence and structure similarity. More detailed protein pairs with sequence identities and *z*-score are shown in Table 8.

As shown in Figure 8, protein structures 1IAG and 1DTH_A are aligned to protein structure 1BKC_A. The structure alignment result is shown in Figure 8. Detailed information about the three protein structures is listed in Table 9.

Another example from Table 7 is the 12th entry, i.e., "PDOC00254 => PDOC00124." In this example, the number of protein pairs in the motif correlation is 55. It is a surprise to find 17 protein pairs of these whose sequence identities are less than 20%, while the structure similarity measures, i.e., *z*-score, are more than 4.0. We also find the sequence identities of 11 other protein pairs are less than 20% and the *z*-scores are less than 4.0. Figure 9 and Table 10 show the information like Figure 8 and Table 9, respectively.

Orengo et al.[17] statistically and computationally studied the correlation of protein sequence identities and structure similarities of protein pairs and suggested that in general those protein pairs with sequence identities more than 20% also have high protein structure similarities.

The proteins that contain the same motif correlation are said to be in a motif correlation group. By computing the sequence identities and structure similarities of protein pairs in a same group, we find that protein pairs with sequence identities more than 20% also have high structure similarities, i.e., *z*-score > 4.0. In some special motif correlation groups like the examples mentioned above, we find some of the protein pairs in a same group have low sequence identities; however, they have high structure similarity. We will now further study these proteins.

References

1. Eddy, S. R. Bioinformatics 1998, 14, 755–763.
2. Bateman, A.; Birney, E.; Durbin, R.; Eddy, S. R.; Howe, K. L.; Sonnhammer, E. L. Nucl Acids Res 2000, 28, 263–266.
3. Sonnhammer, E. L.; Eddy, S. R.; Birney, E.; Bateman, A.; Durbin, R. Nucl Acids Res 1998, 26, 320–322.
4. Grundy, W. N.; Bailey, T. L.; Elkan, C. P.; Baker, M. E. Comp Appl Biosci 1997, 13, 397–406.
5. Junier, T.; Pagni, M.; Bucher, P. Bioinformatics 2001, 17, 1234–1235.
6. Wu, C. H.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Ledley, R. S.; Lewis, K. C.; Mewes, H. W.; Orcutt, B. C.; Suzek, B. E.; Tsugita, A.; Vinayaka, C. R.; Yeh, L. L.; Zhang, J.; Barker, W. C. Nucl Acids Res 2002, 30, 35–37.
7. Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C. J. A.; Hofmann, K.; Bairoch, A. Nucl Acids Res 2002, 30, 235–238.
8. Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N.; Ravichandran, V.; Gilliland, G. L.; Bluhm, W.; Weissig, H.; Greer, D. S.; Bourne, P. E.; Berman, H. M. Nucl Acids Res 2002, 30, 245–248.

9. Conte, L. L.; Ailey, B.; Hubbard, T. J. P.; Brenner, S. E.; Murzin, A. G.; Chothia, C. Nucl Acids Res 2000, 28, 257–259.
10. Shindyalov, I. N.; Bourne, P. E. Nucl Acids Res 2001, 29, 228–229.
11. Shindyalov, I. N.; Bourne, P. E. Protein Eng 1998, 11, 9, 739–747.
12. Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Protein Sci 1992, 1, 409–417.
13. Strikant, R.; Agrawal, R. In: Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland, September 11–15; Morgan Kaufman, 1995; 407–419.
14. Liu, B.; Hsu, W.; Ma, Y. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August; ACM Press; 1999, 125–134.
15. Agrawal, R.; Imielinski, T.; Swami, A. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 26–28, ACM Press: 1993, 207–216.
16. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Nucl Acids Res 1994, 22, 4673–4680.
17. Orengo, C. A.; Sillitoe, I.; Reeves, G.; Pearl, F. M. G. J Struct Biol 2001, 134, 145–165.