Pacific Symposium on Biocomputing 4:190-201 (1999)

Applications of Knowledge Discovery to Molecular Biology: Identifying Structural Regularities in Proteins

Shaobing Su, Diane J. Cook, and Lawrence B. Holder
University of Texas at Arlington
sandy_su@sabre.com, {cook,holder}@cse.uta.edu

Abstract

In recent years, there has been an explosive amount of molecular biology information obtained and deposited in various databases. Identifying and interpreting interesting patterns from this massive amount of information has become an essential component in directing further molecular biology research.

The goal of this research is to discover structural regularities in protein sequences by applying the SUBDUE discovery system to databases found in the Brookhaven Protein Data Bank. In this paper we report the results of applying SUBDUE to several classes of protein structures and discuss the potential significance of these results to the study of proteins.

1 Introduction

The topic of finding biologically meaningful patterns in sequences, secondary, and tertiary structures of proteins and other macromolecules is of interest to many biological and computer scientists. In recent years, the amount of protein structure information obtained has been explosive. This information needs to be analyzed to help understand the structure-function relationship in proteins. The main goal of this research is to apply the SUBDUE knowledge discovery system to identify biologically meaningful patterns in the Brookhaven protein database. In particular, the goal of the discovery is to find distinct structural patterns in categories of proteins or their chains.

1.1 Subdue Knowledge Discovery System

The SUBDUE knowledge discovery system [1,2,3] has been shown to provide an effective means of discovering patterns in structural data sets from several domains. SUBDUE discovers substructures that compress the original data and represent structural concepts in the data. Once a substructure is discovered, the substructure is used to simplify the data by replacing instances of the substructure with a pointer to the newly discovered substructure.

The substructure discovery system represents data as a directed graph with objects in the data represented as vertices in the graph and relationships between objects represented as edges in the graph. Thus, input to SUBDUE is a graph representing the database under inquiry, and the output of SUBDUE is the discovered substructures and the compressed graph described in terms of these substructures.

The substructure discovery algorithm used by SUBDUE is a beam search. The initial substructure corresponds to a unique vertex label in the input graph. As the discovery proceeds, substructure definitions are expanded by adding neighboring edges in all possible ways.

SUBDUE uses the minimum description length (MDL) principle to evaluate each possible substructure, and the top substructures (the exact number is determined by the beam width) are further expanded. The MDL principle states that the best theory to describe a set of data is a theory which minimizes the description length of the entire data set. Thus each substructure is evaluated in terms of how well it can compress the entire data set.

SUBDUE also employs an inexact graph match routine that can be used to find substructures with a slightly different modified definition. A distortion is described in terms of basic transformations such as deletion, insertion, and substitution of the vertices and/or edges. The allowable number of differences can be specified with the *threshold* parameter. For example, when a threshold of 0.2 is specified, SUBDUE will discover all instances of the pattern that have less than 20% difference from the pattern definition.

Graph isomorphism typically requires exponential time to compute. We improve the performance of our inexact graph match by performing a uniform-cost search over the space of partial mappings between two graphs, switching to hill climbing when a user-defined computation limit is exceeded. With these computational constraints in place, SUBDUE's worst-case run time is polynomial in the size of the input graph.

1.2 Structural Hierarchy In Proteins

Proteins are involved in a greater number and greater variety of cellular events than any of the other types of biomolecules[4]. There are three levels of structure that apply to all proteins. They are: (1) the primary level, which refers to the sequence of the amino acids in proteins; (2) the secondary level, which refers to the geometric orientation of the protein backbone; and (3) the tertiary level, which refers to the complete, three-dimensional architecture of the protein. For the secondary structure, there are mainly three types of orientation found in naturally occurring protein chains: helical, sheet, and random. The helical and sheet forms are ordered arrangements, while random forms are random arrangements.

2 Comparison with Other Related Studies

There are several applications of pattern search in proteins on the secondary structure level. Mitchell et al.[7] use an algorithm that identifies subgraph isomorphism in protein structure. They represent the protein structures as an undirected labelled graph, where the secondary structure elements in a protein and the distance and angular relationships between them correspond to the nodes and edges of a graph. Their program allows one to determine whether a query pattern is contained within a complete protein structure (only finding exact matches). The approach of Grindley et al.[8] uses the same representation and finds maximal common substructures between two proteins on the secondary structure level. This approach can therefore highlight areas of structural overlap between proteins. In Koch et al.[9], the graph is considered without explicitly using geometric criteria such as distances and angles in the graph description. The vertices represent the helices and strands assigned by the DSSP algorithm. The edges are calculated on the basis of contacts between the atoms belonging to the respective secondary structure elements. They found this representation could be useful in searching for structurally distantly related proteins. This method has not yet been applied systematically to the PDB database.

3 Methods

We select databases from the Brookhaven 1997 release of the protein data bank for use in this study. Each PDB file is a collection of record types. For example, the SEQRES records contain the amino acid sequence of residues in the protein. The HELIX and SHEET records describe the position of helix or stand in the protein. The ATOM records contain the orthogonal (X, Y, Z) coordinates for each atom of each residue in the protein. Recently, researchers at the MRC laboratory of Molecular Biology and Cambridge Centre for Protein Engineering constructed the Structural Classification of Proteins (SCOP) database. The SCOP database is created mainly by visual inspection.

To accomplish the goal of discovering distinct patterns in categories of proteins, two main groups of data sets are maintained in this particular study. The first one contains all protein PDB files with no duplicate sequences. This represents a global data set. The second data sets contain groups of PDB files for each particular category of proteins or their chains. The compilation of the second data set is mainly based on the classifications listed in the SCOP database. Some of the PDB files contain more than one NMR modeled structure and only the coordinates for the first model are used.

To apply SUBDUE to the PDB, preprocessing programs are used to extract structural information from each the Brookhaven PDB files in the data set and output graph representations of structures which are used as input to the SUBDUE discovery system. For each group of proteins, the primary, secondary, and tertiary structure patterns are identified from the SUBDUE output. These patterns are used as pre-defined patterns in another round of discovery from the global data set. The instances of the patterns identified in each category of proteins are mapped back into individual PDB files from the Brookhaven Protein Data Bank.

3.1 Preprocessing

Primary structure information is extracted from the SEQRES records of the PDB file. For the purpose of identifying primary structure patterns of protein alone, a natural representation would be a linear graph. Each amino acid is represented as a vertex in a graph, the vertex number increments according to the order of the sequence from N-terminus to C-terminus. The vertex label is the name of the amino acid. An edge with a label of "bond" is added between adjacent amino acids in a sequence.

A level of abstraction for the tertiary structure of a protein may be obtained by representing the protein in a form that emphasizes its secondary structures. One way to represent the overall secondary structure contents of a protein is to list the occurrences of helices and strands along the primary sequence. The helix information is extracted from the HELIX records of each PDB file. Each helix in a particular PDB file is represented as a vertex. In the graph representation the vertex label is "h" for helix, followed by the helix type and helix length (number of amino acids involved - 1). The helix length is calculated as the following, where Hlength strands for the length of the helix; SeqNum is the sequence number, and a.a. stands for amino acid.

$$\text{Hlength} = \text{SeqNum(last a.a.) - SeqNum(First a.a)}$$

Similarly, the strand information is extracted from the SHEET records of each PDB file. Each strand in a sheet of a particular PDB file is represented as a vertex. The vertex label is "s" for strand, followed by the orientation (sense) of the strand and the length of the strand. The preprocessing programs then sort the occurrence of the secondary structure elements (helices and strands) from N-terminus to C-terminus. The edge between two consecutive vertices is labelled as "sh" if they belong to the same PDB file. For example, the following simplified PDB lines indicate that the protein has a right-handed helix (with a length of 10), followed by another right-handed helix (with a length of 10), followed by the first strand of the sheet (with a sense of 0 and a length of 7), followed by another right-handed helix (with a length of 10), followed by the second and third strands of the sheet (with length of 8 and 10, respectively). Both strand two and strand three have senses of -1.

HELIX	1	THR 3	MET 13	1
HELIX	2	ASN 24	ASN 34	1
HELIX	3	SER 50	GLN 60	1
SHEET	1	LYS 41	HIS 48	0
SHEET	2	MET 79	THR 87	-1
SHEET	3	ASN 94	LYS 104	-1

The input to SUBDUE for the above example is shown below, where “v” indicates a vertex followed by the vertex number and label, and “e” indicates an edge followed by the connecting vertices and edge label.

```
v 1 h_1_10          — the first right-handed helix
v 2 h_1_10          — the second right-handed helix
v 3 s_0_7           — the first strand of the sheet
v 4 h_1_10          — the third right-handed helix
v 5 s_-1_8          — the second strand anti-parallel to the first
v 6 s_-1_10         — the third strand anti-parallel to the second
e 1 2 sh
e 2 3 sh
e 3 4 sh
e 4 5 sh
e 5 6 sh
```

In the PDB file, three-dimensional features of the protein are represented as the X, Y, and Z coordinates of each atom in the protein. The several hundred atoms of even a very small protein make understanding the detailed structure of a protein a considerable effort. To simplify the representation, only the backbone α- carbon coordinates are extracted. The preprocessing program calculates the pair-wise distance of each backbone carbon. If the distance between the two carbons is greater than 6 $\mathring{A}$, this distance information is discarded. The calculated distances are classified as very-short-distance ($\leq 4\ \mathring{A}$) and short-distance ($4\ \mathring{A} <$ distance $\leq 6\ \mathring{A}$). Each amino acid α- carbon is represented as a vertex. Edges between two alpha carbons are labelled either as “vs” for very short distance or “s” for short distance.

4 Results

The SUBDUE knowledge discovery system is used to search for patterns in the input graph of categories of proteins or their chains. Results obtained from the hemoglobin, myoglobin, and ribonuclease A proteins are presented here.

Data Set (# of PDB)	Exp. Parameter	Discovered Pattern (# of instances in sample data set / global)
Hemoglobin (65)	Beam 50	Hemo_sequence1 (63 / 0)
Myoglobin (103)	Beam 50	Myoglo_sequence2 (67 / 0)
Ribonuclease_A (68)	Beam 50	Ribonuclease_A_sequence3 (59 / 0)

Table 1: The discovered sequence patterns in the sample data sets.

4.1 Primary Structure Patterns

The primary structure patterns discovered for the hemoglobin, myoglobin, and ribonuclease A protein categories are summarized in Table 1 and listed in detail below.

Hemo_sequence: THR LYS THR TYR PHE PRO HIS PHE ASP LEU SER HIS GLY SER ALA GLN VAL LYS GLY HIS GLY LYS LYS VAL ALA ASP ALA LEU THR ASN ALA VAL ALA HIS VAL ASP ASP MET PRO ASN ALA LEU SER ALA LEU SER ASP LEU HIS ALA HIS LYS LEU ARG VAL ASP PRO VAL ASN PHE LYS LEU LEU SER HIS CYS LEU LEU VAL THR LEU ALA ALA HIS LEU PRO ALA GLU PHE THR PRO ALA VAL HIS ALA SER LEU ASP LYS PHE LEU ALA SER VAL SER THR VAL LEU THR SER LYS TYR

Myoglo_sequence: VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN ASP ILE LEU ILE ARG LEU PHE LYS SER HIS PRO GLU THR LEU GLU LYS PHE ASP ARG

Ribonuclease_A_sequence: GLY GLN THR ASN CYS TYR GLN SER TYR SER THR MET SER ILE THR ASP CYS ARG GLU THR GLY SER SER LYS TYR PRO ASN CYS ALA TYR LYS THR THR GLN ALA ASN LYS HIS ILE ILE VAL ALA CYS GLU GLY ASN PRO TYR VAL PRO VAL HIS PHE ASP ALA SER VAL

The sequence patterns identified for the hemoglobin, myoglobin, and ribonuclease A proteins are unique to themselves. Notice that the hemoglobin and myoglobin proteins share little sequence similarity. However, we will show they do share a great deal of similarity in their secondary structural patterns.

4.2 Secondary Structure Patterns

The best secondary structural patterns with thresholds 0.0 and 0.3 discovered by SUBDUE for the hemoglobin, myoglobin, and ribonuclease A proteins are shown in Table 2 (thresholds 0.1 and 0.2 are not listed here). The number of instances of each pattern discovered in the specified dataset is indicated along with the number of instances found in other categories of proteins in the global data set. Sample patterns are listed below, described from N-terminus to C-terminus, where edges are represented by "->" and are labeled "sh".

Hemo_s_1_0.0:

```
h_1_14 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
```

Database	Threshold	Pattern 1 (#inst/global)	Pattern 2 (#inst/global)	Pattern 3 (#inst/global)
Hemoglobin	T=0.0	Hemo_s_1_0.0 (50 / 0)	Hemo_s_2_0.0 (52 / 0)	Hemo_s_3_0.0 (50 / NA)
	T=0.3	Hemo_s_1_0.3 (95 / NA)	Hemo_s_2_0.3 (107 / NA)	Hemo_s_3_0.3 (100 / NA)
Myoglobin	T=0.0	Myo_s_1_0.0 (81 / 0)	Myo_s_2_0.0 (82 / 0)	Myo_s_3_0.0 (81 / 0)
	T=0.3	Myo_s_1_0.3 (83 / NA)	Myo_s_2_0.3 (84 / NA)	Myo_s_3_0.3 (84 / NA)
Ribonuclease A	T=0.0	Ribo_A_s_1_0.0 (25 / 0)	Ribo_A_s_2_0.0 (25 / 0)	Ribo_A_s_3_0.0 (25 / 0)
	T=0.3	Ribo_A_s_1_0.3 (36 / NA)	Ribo_A_s_2_0.3 (36 / NA)	Ribo_A_s_3_0.3 (36 / NA)

Table 2: The discovered secondary structure patterns in the sample data sets.

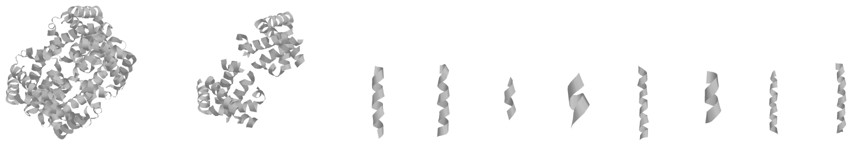

Figure 1: Hemoglobin structure, discovered pattern, and schematic view.

Myo_s_1_0.0:

```
h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_9 -> h_1_18 -> h_1_25
```

Ribo_s_1_0.0:

```
h_1_10 -> h_1_10 -> s_0_7 -> s_0_7 -> h_1_10 -> s_0_3 -> s_0_3 -> s_-1_4 -> s_-1_4
```

4.3 Summary of Results

The Subdue results obtained for the secondary structural pattern discovery in categories of proteins are summarized here. Figures 1 through 3 present an overall view of the protein, the part of the protein where the Subdue-discovered pattern exists, and the schematic views of the best pattern (e.g., pattern 1 with threshold of 0.0) for the hemoglobin, myoglobin, and ribonuclease A proteins, respectively. In each case, the secondary structural elements are listed from N-terminus of the protein to C-terminus.

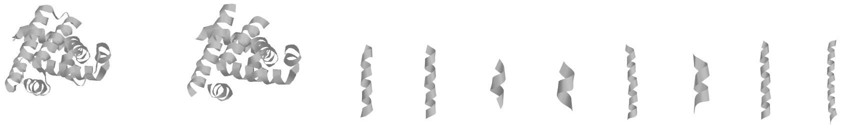

Figure 2: Myoglobin structure, discovered pattern, and schematic view.

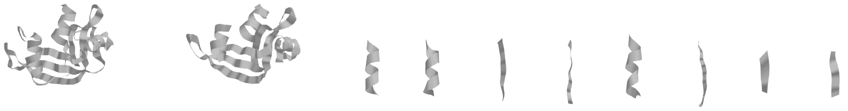

Figure 3: Ribonuclease A structure, discovered pattern, and schematic view.

5 Discussion

5.1 Hemoglobin and Myoglobin Proteins

Hemoglobin and myoglobin are chosen in this study because they have the advantage of familiarity. These proteins are used widely to illustrate nearly every important feature of protein structure, function, and evolution [6]. One molecule of hemoglobin has four protein chains: $\alpha 1$, $\alpha 2$, $\beta 1$, and $\beta 2$ chains (also known as A, C, B, and D chains, respectively). In some species, the two α chains are identical and the two β chains are identical. Myoglobin has one protein chain, of about the same size as each of the four hemoglobin chains [6].

Detailed analysis of the results obtained for the secondary structural patterns of the hemoglobin proteins indicates that there are mainly two types of patterns in the hemoglobin data set. Type 1 includes the two best secondary structural patterns (Hemo_s_1 for thresholds 0.0 and 0.1). They consist of eight helices with various lengths. All the helices are type 1 (e.g., right-handed α - helix). Type 2 patterns include the other two best patterns (Hemo_s_1 for thresholds 0.2 and 0.3). One distinct feature of this type is that one of the helix is very short (length 1).

The occurrence of the instances for each category of proteins is mapped back to the PDB file where the pattern exists. When mapped into the individual chains of the PDB, type 1 patterns are found to belong to the β chains of the hemoglobins. Most of the type 2 patterns are from the α chains of the hemoglobins. Table 3 lists a few of the discovered instances mapped back to

PDB Name	Occurrence	Species	PDB Name	Occurrence	Species
pdb2hhb	B, D (0.0); A, C (0.2)	human	pdb1bbb	B (0.0)	human
pdb4hhb	B, D (0.0); A, C (0.2)	human	pdb1thb	A, B, C, D (0.2)	human
pdb3hhb	B (0.0); A (0.2)	human	pdb1cbm	A, B, C, D (0.0)	human
pdb1hbb	B, D (0.0); A, C (0.2)	human	pdb1hba	B, D (0.0); A, C (0.2)	human
pdb2hbc	B (0.0); A (0.2)	human	pdb1cbl	A, B, C, D (0.0)	human
pdb2hbd	B (0.0); A (0.2)	human	pdb1cbl	B (0.0); A (0.2)	human
pdb1hho	B (0.0); A (0.2)	human	pdb1nih	B, D (0.0); A, C (0.2)	human
pdb1coh	B, D (0.0); A, C (0.2)	human	pdb1fdh	G (0.0)	human
pdb2hco	B (0.0); A (0.2)	human	pdb1cmy	B, C (0.0)	human

Table 3: Occurrences of discovered hemoglobin patterns. The chains in which the pattern occurs are listed along with the threshold at which the pattern was discovered.

the PDB files.

Detailed analysis of the secondary structural patterns identified for the myoglobin proteins indicates that there is one dominant pattern (Myo_s_1_0 for thresholds 0.0, 0.1, 0.2, and 0.3). This pattern consists of eight helices with various lengths. All of the helices are type 1. When the pattern is mapped back to the PDB file, it is found that this pattern appears in a majority of the myoglobin proteins in the data set.

The sequence patterns identified for the hemoglobin and myoglobin proteins show much less degree of similarity (Table 1). However, as discussed in the previous paragraphs, they do share great similarity in their overall secondary structure patterns. Actually, as shown below, the patterns of the hemoglobin protein β chains and that of the myoglobin are identical (both type and length) for the middle six helices. The hemoglobin α chain has a much shorter helix in the middle (h_1_1). In the hemoglobin chains (both α and β chains), the last helix is considerably shorter (e.g., five amino acids shorter) than that of the myoglobin protein chain.

Hemoglobin β chains:

```
h_1_14 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
```

Myoglobin chain:

```
h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_9 -> h_1_18 -> h_1_25
```

Hemoglobin α chains:

```
h_1_15 -> h_1_15 -> h_1_6 -> h_1_1 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
```

This is consistent with the results obtained from genetic studies. Genetic studies suggest that the genes of the hemoglobin and myoglobin proteins evolved by divergence from one ancestral gene [6]. The last helix of the hemoglobin chains is shorter than the one in the myoglobin proteins. One of the helices has almost disappeared in the α chains of the hemoglobin proteins. It has been suggested that it may be due to a random evolutionary process, because the absence of the helix was harmless. It may also have some functional reasons for properly positioning the helices for the conformational changes needed in the hemoglobin proteins[6].

5.2 Ribonuclease A proteins

The ribonuclease A proteins are chosen in this study because they play a special role as a model protein to examine the enzyme structure-function relationships. The results obtained for the secondary structural pattern show that there are two types of patterns (Table 2). All patterns include three helices of about the same size (e.g., with a length of 10 or 12). However, it is noted that type 1 patterns (Ribo_A_s_1 at thresholds 0.0, 0.1, and 0.2) have several strands appearing twice (e.g., s_0_7->s_0_7, s_0_3->s_0_3, etc.). We found that these duplicates are the same strand but are observed as participating in the formation of different sheets. Therefore they have duplicated entries in the PDB files. It is not clear why some of the ribonuclease A proteins do not have these duplicates. Possible reasons are the following: (1) These strands may appear to only participate in the formation of one sheet, instead of two, under some experimental conditions; (2) The resolution of the X-ray crystallographic structure may not be high enough to observe the hydrogen-bonding patterns needed to group strands to sheets.

The secondary structural patterns for the ribonuclease A proteins are mapped back into the PDB files. It is observed that ribonuclease S proteins have type 2 patterns (Ribo_A_s_1_0.3) as those in ribonuclease A proteins. Ribonuclease S is a complex consisting of two fragments (S-peptide and S-protein) of the ribonuclease A proteins. The pattern in the ribonuclease S comes from the S-protein fragment.

5.3 Conclusion of the Results

Results obtained for the hemoglobin, myoglobin, and ribonuclease A protein data sets indicate that the secondary structure patterns discovered by Subdue are representative to each category. The patterns identified for each sample category covered a majority of the proteins in that category. Analysis of those that do not have the pattern indicates many possible reasons. The accuracy of the structure is affected by the quality of the protein sample, the experimental condition, and the human errors. Discrepancies may be due to physiological and biochemical reasons, and structure of the same protein molecule may differ from one species to another. The protein may also be defective. For example, sickle-cell anemia is the classic example of a genetic hemoglobin disease.

The defected protein does not have the right structure to perform its normal function.

Discovered secondary structure patterns are also distinct to each category. Results indicate that no exact instances of the best pattern in one category appear in the other protein categories. However, the current version of SUBDUE has the limitation in that when a predefined pattern is used as a search pattern against the database, only exact matches are located. Therefore, the discovery process may overlook those proteins having similar structure patterns.

The number of protein structures known in atomic detail has increased from 1 in 1960 to more than 6,000 in 1997. A newly determined structure is frequently similar in its secondary and tertiary folds to a known one. The search for common and distinct structure patterns in sets of proteins has become an essential procedure in the investigation of protein structures. The degree of similarity between different categories of proteins may be used for discovering biologically interesting relationships.

Results obtained in this study indicate that the level of abstraction for the tertiary structure which emphasizes its secondary structures is suitable for representing each category of proteins. Through the vertex and edge labelling, essential information on sequential relationships on the secondary structure element is encoded. The structural motifs consisting of secondary structure elements (e.g., helix, sheet) are shown to be responsible for the function of proteins. The secondary structural patterns in each category of proteins can therefore be used as a signature for its class. The inexact graph match algorithm implemented in SUBDUE is useful for finding the similar patterns among different proteins of the same category and across different proteins in related categories.

6 Future Research

The results obtained in this study indicate that the SUBDUE system is suitable for knowledge discovery in the molecular structural database. It should be noted, however, that the results obtained are critically dependent on the secondary structure information used and on the definitions of the structural features and its graph representation. Future work in this area includes the following: (1) A more detailed and consistent description of the secondary structure would be helpful. For example, DSSP can be used to generate more secondary structure classifications than those deposited in the PDB; (2) The relative positions of the secondary structures need to be specified so that the spatial relationships among the secondary structural elements are not limited to that of only sequential; (3) Some of the important interactions such as the hydrogen-bonding and disulfide bonding can be introduced in addition to the secondary structure elements; (4) It is well known that the tertiary or the 3D structure of proteins is extremely complex. Protein tertiary structure comparison still remains a major goal in molecular biology. To apply Subdue for finding tertiary structural patterns, a more suitable representation scheme is needed. This representation scheme should consider the fact that the detailed 3D structures are not identical even for protein pairs that have identical sequences. This deviation is attributed to different crystal forms, to experimental conditions, and to human errors. It may also be a mere reflection of conformational flexibility of protein structures. The tertiary structure comparison for a site (e.g., a catalytic site or other regulatory site) composed of much smaller sets of atoms in proteins is a good starting point.

1. Diane J Cook and Lawrence B Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.
2. D J Cook, L B Holder, and S Djoko. Scalable discovery of informative structural concepts using domain knowledge. *IEEE Expert*, 10:59–68, 1996.
3. Gehad Galal, Diane J. Cook, and Lawrence B Holder. Improving scalability in a scientific discovery system by exploiting parallelism. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 171–174, 1997.
4. R C Bohinski. *Modern Concepts in Biochemistry.* Allyn and Bacon, Inc., 1979.
5. E E Abola, F C Bernstein, S H Bryant, T F Koetzle, and J Weng. Protein data bank. In *Crystallographic Databases-Information Content, Software Systems, Scientific Applications*, pages 107–132. Data Commission of the International Union of Crystallography, 1987.
6. R E Dickerson and I Geis. *Hemoglobin: structure, function, evolution, and pathology.* Benjamin/Cummings Inc., 1982.
7. E M Mitchell, P J Artymiuk, D W Rice, and P Willett. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *Journal of Molecular Biology*, 212:151–166, 1990.
8. H M Grindley, P J Artymiuk, D W Rice, and P Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, 229:707–721, 1993.
9. I Koch, T Lengauer, and E Wanke. An algorithm for finding maximal common subtopologies in a set of protein structures. *Journal of Computational Biology*, 3(2):289–306, 1996.