Knowledge Discovery in Molecular Biology: Identifying Structural Regularities in Proteins *

Shaobing Su, Diane J. Cook, and Lawrence B. Holder
Department of Computer Science Engineering
University of Texas at Arlington
Email: {shaobing, cook, holder}@cse.uta.edu
Contact Author: Diane J. Cook
http://cygnus.uta.edu/subdue.html

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 discuss issues relevant to this application including data preparation and representation. We report on the results of applying SUBDUE to several classes of protein structures and discuss the potential significance of these results to the study of proteins.

*Supported by NSF grant ILI-9615272 and THECB grant 003656-045.

1 Introduction

The topic of finding biologically meaningful patterns in sequences, secondary, and tertiary (three-dimensional) structures of proteins and other macromolecules is of interest to many biological and computer scientists. In recent years, there has been an explosive amount of molecular biology information obtained and deposited in various biological databases. The problem of interpreting this information is increasingly becoming the limiting step in many molecular biology projects. Without clues to the probable structure and functions of a new protein, further research is often blocked.

The SUBDUE knowledge discovery system has been shown to provide an effective means of discovering patterns in several domains [4, 5, 7]. The SUBDUE algorithm is based on the Minimum Description Length (MDL) principle with an inexact graph match implementation. The SUBDUE system can discover interesting patterns with either identical instances or instances of slightly different forms.

Progress is being made in the field of sequence and structure matching and prediction for biological systems. However, there has not been sufficient progress to provide a general structure prediction method. Determining the primary structure of a protein is already an automated lab task. On the other hand, determining the secondary and tertiary structure of a protein in the laboratory is still a costly and time-consuming effort. Despite this, the number of protein structures determined experimentally has increased dramatically as a result of recent advances in protein engineering, crystallography, and NMR spectroscopy. As both the sequence and structure databases grow, there is an urgent need for automating the extraction of useful information from such large databases.

The goal of this research is to discover structural regularities in protein sequences by applying SUBDUE to databases found in the Brookhaven Protein Data Bank (PDB). Using the results of this application we will evaluate the potential benefits of applying SUBDUE to PDB and other similar molecular biology databases.

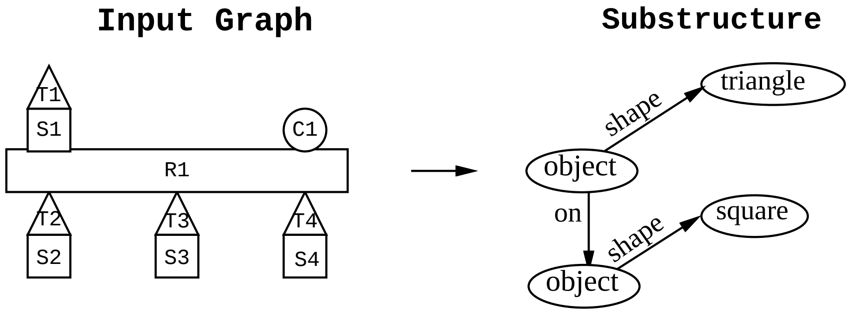

Figure 1: Example substructure in graph form.

2 The Subdue Knowledge Discovery System

We have developed a method for discovering substructures in databases using the minimum description length (MDL), embodied in the SUBDUE system. The minimum description length (MDL) principle introduced by Rissanen [14] states that the best theory to describe a set of data is the theory which minimizes the description length of the entire data set. The MDL principle has been used for decision tree induction, image processing, concept learning from relational data, and learning models of non-homogeneous engineering domains. We define the minimum description length of an input graph to the minimum number of bits necessary to completely describe the graph.

SUBDUE discovers substructures that compress the description length of 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 discovered substructures allow abstraction over detailed structures in the original data. Iteration of the substructure discovery and replacement process constructs a hierarchical description of the structural data in terms of the discovered substructures. This hierarchy provides varying levels of interpretation that can be accessed based on the specific goals of the data analysis.

The substructure discovery system represents structural data as a labeled graph. Objects in the data map to vertices or small subgraphs in the graph, and relationships between objects map to directed or undirected edges in the graph. A *substructure* is a connected subgraph within the graphical representation. This graphical representation serves as input to the substructure

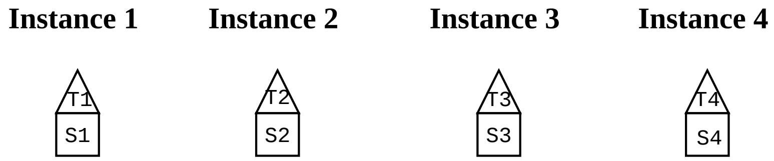

Figure 2: Instances of the substructure.

discovery system. Figure 1 shows a geometric example of such an input graph. The objects in the figure (e.g., `T1`, `S1`, `R1`) become labeled vertices in the graph, and the relationships (e.g., `on(T1,S1)`, `shape(C1,circle)`) become labeled edges in the graph. The graphical representation of the substructure discovered by SUBDUE from this data is also shown in Figure 1.

An *instance* of a substructure in an input graph is a set of vertices and edges from the input graph that match, graph theoretically, to the graphical representation of the substructure. For example, the instances of the substructure in Figure 1 are shown in Figure 2.

The substructure discovery algorithm used by SUBDUE is a computationally-constrained beam search. The algorithm begins with the substructure matching a single vertex in the graph. Each iteration, the algorithm selects the best substructure according to the MDL heuristic and incrementally expands the instances of the substructure. The new unique substructures become candidates for further expansion. The algorithm searches for the best substructure until all possible substructures have been considered or the total amount of computation exceeds a given limit. Evaluation of each substructure is determined by the MDL heuristic; specifically, by how much the description length of the database is reduced when substructure instances are replaced by pointers to the substructure definition.

Because instances of a substructure can appear in different forms throughout the database, an inexact graph match is used to identify substructure instances. Subgraphs are considered to be instances of a substructure definition if the cost of transforming the subgraph into a graph that is isomorphic with the substructure definition does not exceed a user-defined threshold. Transformations between graphs can include addition or deletion of vertices, addition or deletion of edges, vertex label substitutions and edge label substitutions.

SUBDUE discovers substructures that compress the amount of information necessary to conceptually describe the database. To allow SUBDUE to discover substructures of particular interest to a scientist in a given domain, the user can direct the search with expert-supplied background knowledge. Background knowledge can take the form of known substructure models to specifically locate in the database, or graph match rules to adjust the cost of each inexact graph match test. Unlike other existing approaches to graph-based discovery [3, 12, 15, 18, 20], SUBDUE is effective at finding interesting and repetitive substructures in any structural database with or without domain-specific guidance.

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 discovered substructures allow abstraction over detailed structures in the original data. Iteration of the substructure discovery and replacement process constructs a hierarchical description of the structural data in terms of the discovered substructures. This hierarchy provides varying levels of interpretation that can be accessed based on the specific goals of the data analysis [4]. In addition to the application described in this paper, SUBDUE has been successfully applied with and without domain knowledge to databases in domains including image analysis, CAD circuit analysis, Chinese character databases, program source code, chemical reaction chains, and artificially-generated databases. Evaluation of these applications is described elsewhere [4, 8].

3 Proteins and the Brookhaven Protein Data Bank

3.1 Overview of Proteins

Proteins are involved in a greater number and greater variety of cellular events than any of the other types of biomolecules. Along with nucleic acids (DNA and RNA), they carry the information that determines what happens in a cell of a living organism [2]. Each protein is different from every other protein in terms of its structure and function. There is also much similarity between proteins, the most common aspect of which is that all proteins are composed from twenty amino acids. Amino acids are therefore the basic building blocks of all proteins. The general structural formula for an amino acid is shown in Figure 3. There are twenty different R groups in the commonly occurring amino acids.

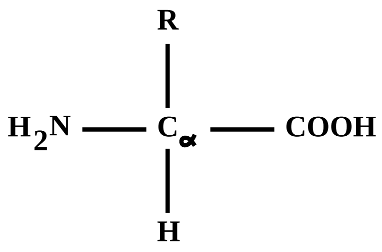

Figure 3: General structural formula for amino acids.

3.2 Structural Hierarchy in Proteins

Familiarity with various aspects of protein structure is indispensable to the eventual understanding of much of the biochemical dynamics of a living organism. There are three levels (aspects) of structure that apply to all proteins. They are: (1) the primary level, which refers to the sequence of the amino acids in the protein; (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.

3.2.1 Primary Structure

The identity and sequence of amino acids are the most fundamental structural characteristics of any protein. Each amino acid residue in a chain is linked to its neighbors in a head-to-tail fashion. The chain starts at the amino terminus (N-terminus) and ends at the carboxyl terminus (C-terminus).

The naturally occurring proteins generally contain varying amounts of the twenty common amino acids and have an average length of 100-150 residues [19]. Functional properties of each protein are a consequence of its amino acid sequence. The identity and total number of amino acid residues are important, but the order in which the residues are linked together is of greatest importance. It is the sequence of residues that determines the overall three-dimensional shape of the molecule, which in turn determines how that molecule will function.

3.3 Secondary Structure

The particular type of orientation assumed by a protein chain is the result of the pattern of free rotation around the bonds of the chain involving the α-carbon atoms. Three major types of

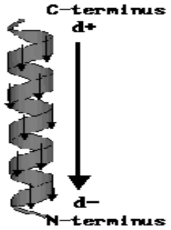

Figure 4: α-helix.

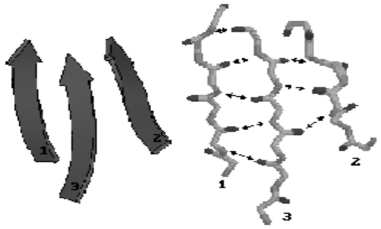

Figure 5: parallel β-sheet.

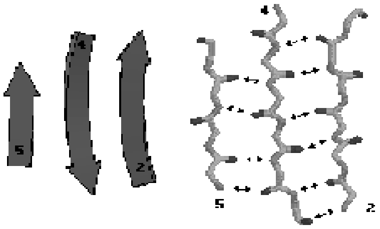

Figure 6: anti-parallel β-sheet.

orientation are found in naturally occurring protein chains: helical, sheet, and random. The helical and sheet forms are ordered arrangements, while the random forms are random arrangements. The percentage of α-helix content in proteins is quite variable, ranging from 0% to 80-90%. In globular proteins (e.g., proteins, in their native state, exist as compact spheroidal molecules), α-helices have an average span of about eleven residues with up to fifty-three residues found in a helix.

There are two arrangements of strands in a sheet. If two neighboring strands are aligned in the same direction from one terminus to the other, the arrangement is termed a *parallel sheet.* If the two are aligned in opposite direction, the arrangement is termed an *anti-parallel sheet.* In globular proteins, β-sheets consist of from two to as many as fifteen strands, with an average of six strands. The number of residues in a β-sheet is up to fifteen residues long, with an average of six residues. Figures 4, 5, and 6 depict a sample α-helix, parallel β-sheet, and anti-parallel β-sheet.

3.4 Tertiary Structure

The tertiary structure of a protein is its three-dimensional arrangement; in particular, the folding of its secondary structure elements, together with the spatial disposition of its R group (or side chain). Each of them is unique and highly complicated. Most proteins have a significant amount of both α-helix and β-sheet, in varying proportions and combinations.

In some proteins, especially in enzymes (e.g., groups of proteins having catalytic functions), there is part of the protein that has a unique three-dimensional structural feature and is crucial for the function of the protein. This part is usually called a site (such as catalytic, cofactor, and regulatory site) in the macromolecule. Understanding the detailed three-dimensional structure of a biomolecule and of these sites crucial for understanding the function of a particular protein.

3.5 Brookhaven Protein Data Bank

The Protein Data Bank (PDB) database is compiled at Brookhaven National Laboratory [1]. It is an archive of experimentally determined three-dimensional structures of biomolecules. The majority of the files represent protein structures determined by X-ray crystallography or NMR spectroscopy.

Every PDB file may be viewed as a collection of record types. We examined the following types of records: SEQRES, HELIX, SHEET, and ATOM. The SEQRES records contain the amino acid sequence of residues in the protein along with other related identification information. The HELIX records are used to identify the position of helices in the protein. The residues where the helix starts and ends are noted, as well as the type of the helices. The SHEET records are used to identify the position of sheets in the protein. The residues where each strand of a sheet starts and ends are noted. The sense (relative orientation) of a strand with respect to the previous strand in the sheet is also provided. The sense is 0 if the strand is the first of a sheet. The sense is 1 if strand n is parallel to strand n-1 in the same sheet, and is -1 if anti-parallel. Finally, the ATOM records contain the orthogonal (X, Y, Z) coordinates (in $\mathring{A}$) for each atom of each residue in the protein. The choice of origin is not consistent between databases and does not effect the coding of the input graph for SUBDUE.

Both the HELIX and SHEET records are now being generated automatically by PDB using the Kabsch and Sander algorithm [10]. The algorithm uses a set of simple and physically motivated criteria for secondary structure assignments and provides an unambiguous and physically meaningful definition of secondary structure.

4 Application of Subdue to PDB databases

The main goal of this research is to identify biologically meaningful patterns in the Brookhaven protein database using SUBDUE. In particular, the discovery goal is to find distinct structural patterns in categories of proteins or their chains.

The PDB files used in this study are based on the July 1997 PDB release containing over 6,000 files. The majority of the files represent protein structures determined by X-ray crystallography or NMR spectroscopy. The remaining PDB files contain structures determined for DNA, RNA, and other macromolecules.

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 the protein PDB files with no duplicate sequences (about 4,000 PDB files). This represents a global data set. The second data set contains groups of PDB files for each particular category of proteins (65, 103, and 68 PDB files for hemoglobin, myoglobin, and ribo_nuclease A categories, respectively).

To apply SUBDUE to the PDB, pre-processing programs are used to extract structural information from each PDB file in the data set. The processed files are then input to SUBDUE. For each group of proteins, the primary, secondary, and tertiary structure patterns are identified by SUBDUE. These patterns are used as background knowledge patterns in another iteration of discovery from the global data set.

4.1 Representation of Structure Information

We now turn our attention to methods of representing PDB data as a graph suitable for input to SUBDUE.

4.1.1 Pre-processing PDB files

The first step in our discovery approach is to pre-process the PDB file to extract structural information. The primary structure information is extracted from the SEQRES records of each PDB file. Each line starting with the SEQRES keyword contains the list of amino acids in the order of the sequence from N- to C-terminus. Information other than this sequence is not used in the discovery. A portion of a sample PDB file is listed in the Appendix. For example, the following two sample PDB lines contain the sequence: ALA ASN LYS THR ARG GLU LEU CYS MET LYS SER LEU GLU HIS ALA LYS VAL ASP:

```
SEQRES 1 150 ALA ASN LYS THR ARG GLU LEU CYS MET 1ASH 139
SEQRES 2 150 LYS SER LEU GLU HIS ALA LYS VAL ASP 1ASH 140
```

To convert this information to a graph, each amino acid is represented as a graph vertex. 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 labeled "bond" is added between adjacent amino acids in a sequence.

As the second step, we extract the secondary structure of a protein by listing the occurrences of helices and strands along the primary sequence. The helix information is extracted from the HELIX records of each PDB file. Each line starting with the HELIX keyword contains information about where the helix starts and ends, along with information about the helix type. Our pre-processing program converts this information into a Subdue graph. Each helix in a PDB file is represented with a vertex labeled "h", followed by the helix type and length (number of amino acids in the helix minus one).

The strand information is extracted from the SHEET records of each PDB file. Each line starting with the SHEET keyword contains information about where a particular strand of a sheet starts and ends. The sense or the relative orientation (parallel or anti-parallel) of a strand to the previous one in a sheet is also extracted. The corresponding Subdue graph represents each strand as a vertex labeled "s" followed by the orientation of the strand and the length of the strand (given in the PDB file). The preprocessing program then sorts the occurrence of the secondary structure elements (helices and strands) from N-terminus to C-terminus. The edge between two consecutive vertices is labeled "sh" if they belong to the same PDB file.

For example, the following simplified PDB lines indicate that the described 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 this 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. Each line starting with the ATOM keyword contains information about the amino acid name, sequence number, and X, Y, and Z coordinates of each atom in the amino acid. To simplify the representation of the three-dimensional structure features of a protein, only the backbone α-carbon coordinates are extracted. For example, the following four simplified PDB lines contain backbone α-carbon coordinates for ALA, ASN, LYS, and THR:

```
ATOM CA ALA 1 10.369 0.997 10.519
ATOM CA ASN 2 6.691 0.239 9.830
ATOM CA LYS 3 6.677 1.983 6.389
ATOM CA THR 4 9.693 -0.188 5.372
```

The pre-processing program computes the pair-wise distance between each backbone α-carbon. A Subdue graph is then generated in the following manner: each amino acid α-carbon is represented as a vertex. If the distance between two α-carbons is greater than 6 $\AA$, the information is discarded. Otherwise, edges between two α-carbons are created and labeled as "vs" (very short, distance $\leq$ 4 $\AA$), or "s" (short). The Subdue graph for our sample file is shown below.

v 1 CA_ALA	— α-carbon of ALA
v 2 CA_ASN	— α-carbon of ASN
v 3 CA_LYS	— α-carbon of LYS
v 4 CA_THR	— α-carbon of THR
e 2 1 vs	— very short distance between v 2 (ASN) and v 1 (ALA)
e 3 1 s	— short distance between v 3 (LYS) and v 1 (ALA)
e 4 1 s	— short distance between v 4 (THR) and v 1 (ALA)
e 3 2 vs	— very short distance between v 3 (LYS) and v 2 (ASN)
e 4 2 s	— short distance between v 4 (THR) and v 2 (ASN)
e 4 3 vs	— very short distance between v 4 (THR) and v 3 (LYS)

4.2 Rationale for Representation Choices

The goal of this research is to apply the SUBDUE knowledge discovery system to find biologically meaningful patterns from the PDB database. In order to accomplish this goal, choices must be made to represent the biological information in a graphic form that can be used as input to the SUBDUE discovery system. These representational choices must fulfill the following criteria for the goal of the discovery: (1) The patterns identified by SUBDUE must be representative for each category of proteins; (2) The patterns discovered by SUBDUE should discriminate one category of proteins from those of other categories. The pre-processing steps described in the previous sections are designed to extract useful information from the PDB and represent them in the SUBDUE input format, as simply and yet as completely as possible.

For the purpose of identifying primary structure patterns of proteins alone, a natural representation would be a linear graph with nodes (or vertices) corresponding to the amino acid residue names, and edges corresponding to the peptide bonds between the consecutive residues.

For the helix secondary structure, there are mainly three kinds of information residing in a PDB file. They are: (1) the type of the helix; (2) the amino acid residues involved; and (3) the starting and ending points (e.g., the length of the helix). We can capture this information by a variety of methods, but some representations would place undue emphasis on irrelevant information such as the helix name or type. The representation we choose allows for equal emphasis on each part of the structural definition. Results obtained from this study indicate that the level of abstraction we encode allows for effective identification of distinct secondary structure patterns in categories of proteins. A similar rationale applies to the representation of PDB SHEET information.

For the tertiary structure features of a protein, the orthogonal coordinates for each atom of each residue are given in a PDB file. The several hundred atoms of even a very small protein make understanding the detailed structure of a protein a considerable effort. The most instructive method of representing a protein structure is the backbone of the protein, which can be represented using its α-carbon atoms. One representation would map the absolute coordinate position of each α-carbon to a residue in the primary sequence. However, the absolute coordinates fail to represent the fact that when choosing different origins, the same types of objects will have different coordinates. We overcome this difficulty by using pair-wise distances between all α-carbons in a protein. An encoding of the exact distance is avoided because the accuracy of protein X-ray of NMR structure determination is limited by its resolution, and because the detailed protein geometry can change with varying environmental conditions. Empirical results showed that our "very short" label ($\leq$ 4 $\AA$) captures the distances between all consecutive residues, along with a few other α-carbons in close contact. The "short" label includes all other α-carbons that have spatial proximity. The 6 $\AA$ cut-off is chosen based on experience gained from NMR structural study [17]. For those having distance greater than 6 $\AA$, the long-range interaction is not considered as part of the immediate environment, and therefore the distance information is discarded.

5 Experimental Results

In our experiment, we use SUBDUE to discover patterns within a given category of proteins or their chains. Once the patterns are discovered, they are encoded as background knowledge and used to search the global database for instances of the pattern.

5.1 Discovered Primary Structure Patterns

The results of SUBDUE applied to the primary structures for the hemoglobin, myoglobin, and rubonuclease A proteins are summarized in Table 1, and the discovered sequences are listed in Figure 5.1. For each discovered pattern, the table lists the search beam width and the number of instances found in the sample data set and in the global database (all proteins not included in the sample set). The sequence patterns identified for the hemoglobin, myoglobin, and ribonuclease A proteins are unique to these classes of proteins. Notice that the hemoglobin and myoglobin proteins share little sequence similarity. However, as discussed later, they do share a great deal of similarity in their overall secondary structural patterns.

5.2 Discovered Secondary Structure Patterns

The top three secondary structural patterns discovered by SUBDUE for the hemoglobin, myoglobin, and ribonuclease A proteins are listed in Table 2. A sampling of the discovered patterns is given in Figure 5.2, listed from N-terminus to C-terminus. In this list edges are represented by "->" and are labeled "sh". Patterns are ordered according to how well they compress the original graph: pattern 1 has the highest value, followed by pattern 2 and pattern 3. The number of instances of

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

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)

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

Figure 7: Discovered sequences for hemoglobin, myoglobin, and ribonuclease A.

```
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
Hemo_s_2_0.0: h_1_14 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18
Hemo_s_3_0.0: h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
Hemo_s_1_0.1: h_1_14 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_23
Hemo_s_1_0.2: h_1_15 -> h_1_15 -> h_1_6 -> h_1_1 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
Hemo_s_1_0.3: h_1_15 -> h_1_15 -> h_1_6 -> h_1_1 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
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
Myo_s_1_0.1: h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_9 -> h_1_18 -> h_1_25
Myo_s_1_0.2: h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_23
Myo_s_1_0.3: h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_23
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
Ribo_s_1_0.1: h_1_12 -> s_0_6 -> s_0_6 -> h_1_10 -> s_0_3 -> s_0_3 -> s_-1_4
     -> s_-1_4 -> s_-1_8 -> s_-1_1 -> s_-1_10 -> s_-1_10 -> s_-1_8 -> s_-1_8
     -> s_-1_5 -> s_-1_3
Ribo_s_1_0.2: h_1_10 -> h_1_12 -> s_0_6 -> s_0_6 -> h_1_10 -> s_0_3 -> s_0_3
     -> s_-1_4 -> s_-1_4 -> s_-1_8 -> s_-1_1 -> s_-1_10 -> s_-1_10 -> s_-1_8
     -> s_-1_8 -> s_-1_5 -> s_-1_3
Ribo_s_1_0.3: h_1_10 -> h_1_10 -> s_0_7 -> h_1_10 -> s_0_3 -> s_-1_4 -> s_-1_8
     -> s_-1_8 -> s_-1_6
H_M_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
H_M_s_1_0.1: h_1_15 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_9 -> h_1_18 -> h_1_25
H_M_s_1_0.2: h_1_15 -> h_1_14 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_20
H_M_s_1_0.3: h_1_14 -> h_1_15 -> h_1_6 -> h_1_6 -> h_1_19 -> h_1_8 -> h_1_18 -> h_1_23
```

Figure 8: Discovered secondary patterns for hemoglobin, myoglobin, and ribonuclease A.

each pattern discovered in the specified dataset (e.g., Hemo, Myo, and Ribo_A) is indicated along with the number of instances found in other categories of proteins in the global data set. Several executions of SUBDUE are tested with varying values for allowable graph dissimilarity (T). The significance of these results will be discussed later in the paper.

To identify the degree of similarity between the secondary structural patterns of the hemoglobin and myoglobin proteins, all the PDB files from these two data sets are combined to form a hemoglobin-myoglobin (H_M) data set. The top three patterns discovered by SUBDUE in this combined data set are also shown in Table 2. The number of instances of each pattern in the hemoglobin and myoglobin data sets is also indicated. Notice the great amount of structural similarity between proteins in the hemoglobin and myoglobin databases that is discovered when 20% or 30% of the graph definitions can vary from instance to instance (T=0.2 or T=0.3). In these cases as many common structures are found between these classes of proteins as are found within a single class of proteins.

Table 2: The discovered secondary structure patterns in the sample data sets (NA = Not Analyzed).

Database	Threshold	Pattern 1	Pattern 2	Pattern 3
		(#instances/global)	(#instances/global)	(#instances/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=0.1	Hemo_s_1_0.1 (51 / NA)	Hemo_s_2_0.1 (58 / NA)	Hemo_s_3_0.1 (52 / NA)
	T=0.2	Hemo_s_1_0.2 (90 / NA)	Hemo_s_2_0.2 (98 / NA)	Hemo_s_3_0.2 (92 / 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=0.1	Myo_s_1_0.1 (81 / NA)	Myo_s_2_0.1 (84 / NA)	Myo_s_3_0.1 (81 / NA)
	T=0.2	Myo_s_1_0.2 (83 / NA)	Myo_s_2_0.2 (84 / NA)	Myo_s_3_0.2 (83 / NA)
	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=0.1	Ribo_A_s_1_0.1 (27 / NA)	Ribo_A_s_2_0.1 (27 / NA)	Ribo_A_s_3_0.1 (27 / NA)
	T=0.2	Ribo_A_s_1_0.2 (27 / NA)	Ribo_A_s_2_0.2 (27 / NA)	Ribo_A_s_3_0.2 (27 / NA)
	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)
H_M	T=0.0	H_M_s_1_0.0 (0 / 81)	H_M_s_2_0.0 (0 / 82)	H_M_s_3_0.0 (0 / 81)
	T=0.1	H_M_s_1_0.1 (0 / 81)	H_M_s_2_0.1 (0 / 82)	H_M_s_3_0.1 (0 / 81)
	T=0.2	H_M_s_1_0.2 (54 / 83)	H_M_s_2_0.2 (51 / 83)	H_M_s_3_0.2 (62 / 83)
	T=0.3	H_M_s_1_0.3 (89 / 83)	H_M_s_2_0.3 (98 / 83)	H_M_s_3_0.3 (98 / 83)

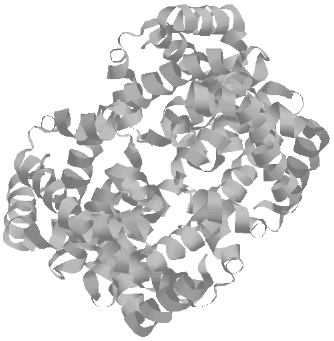

Figure 9: Complete structure of a hemoglobin protein.

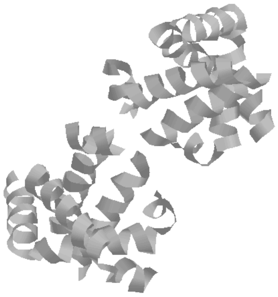

Figure 10: Discovered protein structure of a hemoglobin protein (N-terminus ⟶ C-terminus).

5.3 Discovered Tertiary Structural Patterns

Preliminary results obtained for the tertiary structural pattern discovery indicate that SUBDUE finds small patterns involving two or three residues in the proteins. These patterns are not biologically meaningful and do not fulfill the goal of discovering distinct 3-D patterns in a category of proteins. Future work will focus on discovery in this area.

5.4 Summary of Results

SUBDUE results obtained for the secondary structural pattern discovery in categories of proteins are summarized here. Figure 9 presents an overall view of a hemoglobin protein. Figure 10 shows the part of the proteins where the SUBDUE-discovered pattern for the hemoglobin protein exists, and Figure 11 shows the schematic views of the best pattern (e.g., pattern 1 with threshold of 0.0) discovered by SUBDUE for the hemoglobin proteins. Similarly, Figure 12 presents an overall view of a myoglobin protein, and Figures 13 and 14 show the SUBDUE-discovered pattern within the myoglobin protein and the schematic views of the highest-valued discovered pattern. Figures 15 through 17 present similar results for the ribonuclease A protein. In each case, the secondary structural elements are listed from N-terminus of the protein to C-terminus.

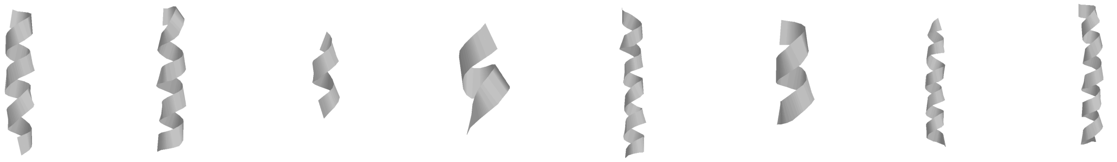

Figure 11: Schematic view of discovered pattern in hemoglobin protein.

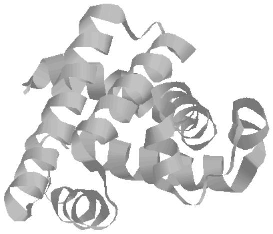

Figure 12: Complete structure of a myoglobin protein.

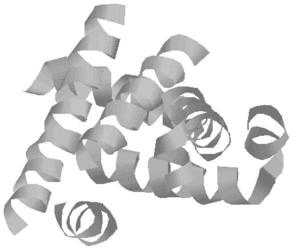

Figure 13: Discovered protein structure in myoglobin protein (N-terminus ⟶ C-terminus).

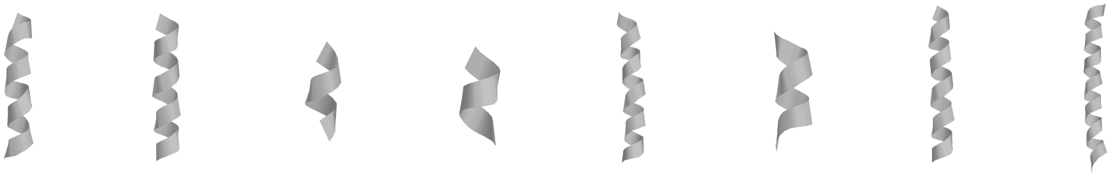

Figure 14: Schematic view of discovered pattern in myoglobin pattern.

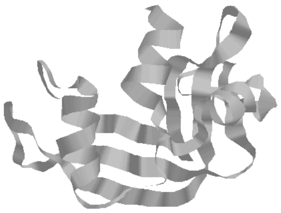

Figure 15: Complete structure of a riboglobin protein.

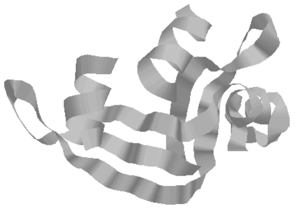

Figure 16: Discovered protein structure of a riboglobin protein (N-terminus ⟶ C-terminus).

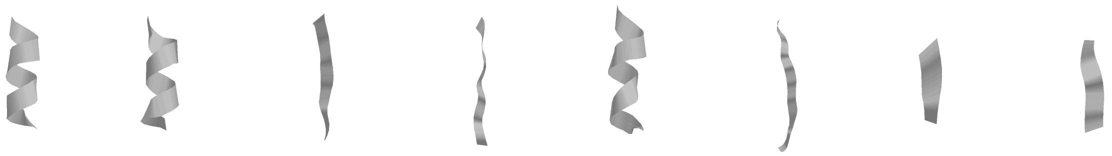

Figure 17: Schematic view of discovered pattern in riboglobin pattern (N-terminus ⟶ C-terminus).

6 Discussion

In this study, the SUBDUE knowledge discovery system is applied to the Brookhaven Protein Data Bank (PDB) to identify biologically interesting patterns in categories of proteins. Results obtained from the hemoglobin, myoglobin, and ribonuclease A protein data sets are discussed in this section.

6.1 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]. Hemoglobin is the oxygen carrier of the blood, whereas myoglobin is the oxygen storage protein of the muscle. One molecule of hemoglobin has four protein chains: $\alpha1$, $\alpha2$, $\beta1$, and $\beta2$ chains (also known as A, C, B, and D chains, respectively). In some species, the two α chains (or A and C chains) are identical and the two β chains (or B and D chains) are identical. Myoglobin has one protein chain, of about the same size as each of the four hemoglobin chains.

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_0.0 and Hemo_s_1_0.1 in Table 2). They consist of eight helices with various lengths. All the helices are right-handed α-helix. Type 2 patterns include the other two discovered patterns (Hemo_s_1_0.2 and Hemo_s_1_0.3 in Table 2). One distinct feature of this type is that one 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 (or B and D chains) of the hemoglobins. Most of the type 2 patterns are from the α chains (or A and C chains) of the hemoglobins.

Detailed analysis of the secondary structural patterns identified for the myoglobin proteins indicates that there is one dominant pattern (Myo_s_1_0.0, Myo_s_1_0.1, Myo_s_1_0.2, and Myo_s_1_0.3 in Table 2). 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 patterns identified from the hemoglobin-myoglobin data set indicate that the myoglobin secondary structural patterns share a great deal of similarity with those of the hemoglobin proteins. This is shown in the results obtained using a threshold of 0.2 and 0.3 (H_M_s_1_0.2, H_M_s_2_0.2, H_M_s_3_0.2, H_M_s_1_0.3, H_M_s_2_0.3, and H_M_s_3_0.3 in Table 2).

The primary sequence patterns identified for the hemoglobin and myoglobin proteins show much less degree of similarity. However, as discussed in the previous paragraphs, they do share great similarity in their overall secondary structure patterns. Actually, 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 very short helix in the middle (h_1_1). In the hemoglobin chains, the last helix is considerably shorter (e.g., five amino acids shorter) than that of the myoglobin protein chain.

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 this disappearance may be due to a random evolutionary process, because the absence of the helix was harmless. The disappearance may also have some functional reasons for properly positioning the helices for the conformational changes (e.g., from deoxy- to oxy-) needed in the hemoglobin proteins.

6.2 Ribonuclease A proteins

The ribonuclease A proteins are chosen in this study because they also 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 the patterns all include three helices about the same size (e.g., with a length of 10 or 12). However, it is noted that all these discovered patterns (Ribo_A_s_1_0.0, Ribo_A_s_1_0.1, Ribo_A_s_1_0.2, and Ribo_A_s_1_0.3) have several strands appearing twice. Detailed analysis of the PDB files for which these duplicates exist has been performed. It is 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; or (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 were mapped back into the PDB files. It is observed that several ribonuclease S proteins have the same patterns as those in ribonuclease A proteins. This is consistent with the fact that 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.

6.3 Summary of 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 its category. The patterns identified for each sample category covered a majority of the proteins in that category (33 of the 50 analyzed hemoglobin proteins, 67 of the 89 myoglobin proteins, and 35 of the 52 ribonuclease A proteins contained the discovered patterns). Detailed analysis of those that do not have the pattern indicates that there are many possible reasons. The structure of a protein is affected by many factors. The accuracy of the structure is affected by the quality of the protein sample, experimental conditions, and human error. Discrepancies may also be due to physiological and biochemical reasons. 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 defective protein does not have the right structure to perform its normal function.

Dr. Steve Sprang, a molecular biologist at the University of Texas Southwestern Medical Center, evaluated the patterns discovered by the SUBDUE system. This scientist was asked to review the original database and the discovered substructures, and determine if the discovered concepts were indicative of the data and interesting discoveries. Dr. Sprang indicated that SUBDUE did find an interesting pattern in the data that was previously unknown and suggests new information about the micro-evolution of such proteins in mammals [16].

The secondary structure patterns discovered are also distinct to each protein category. The global data set is searched to identify the possible existence of the discovered pattern from each category. Results indicate that there is no exact match of the best patterns of one category in other category of proteins (Table 2). However, the current version of SUBDUE has the limitation in that when a particular pattern is searched for in the database, only exact size matches are located. Therefore, the discovery process may overlook those proteins having some similar structure patterns.

6.4 Comparison with Related Studies

There are several applications of pattern search in proteins on the secondary structure level. Mitchell et al. [13] use an algorithm that identifies subgraph isomorphism in protein structure. They represent the protein structures as an undirected labeled 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, POSSUM A, allows one to determine whether a query (or predefined) pattern is contained within a complete protein structure. This program is also limited to performing an exact match when search for a specific pattern in the database.

The approach of Grindley et al. [9] uses the same representation scheme 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. [11], 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. By applying these representations of the protein structure, they found that it could be useful in searching for structurally distantly related proteins. This method has not yet been tested systematically on the PDB database. Most of these studies focus on identifying similar patterns in a group of proteins using predefined patterns. SUBDUE will perform similar tasks when the inexact graph match routine is incorporated into its predefined substructure functionality.

7 Conclusions

The number of protein structures known in atomic detail has increased from one in 1960 to more than 6,000 in 1997. More and more frequently, a newly determined structure is similar in its secondary and tertiary folds to a known one. The search for common and distinct patterns in sets of proteins has becoming an essential procedure in the investigation of protein structures.

Finding a signature for a group of proteins allows recognition of such a protein, and provides a basis for inferring functions of similar proteins. The substructures discovered by SUBDUE represent such a signature for a class of proteins, as shown in this paper. The degree of similarity between different categories of proteins may be used for discovering biologically interesting relationships. In addition, generating a hierarchical view of the data using multiple iterations of SUBDUE aids in understanding the protein from the individual building block amino acids and structures.

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.

The results obtained in this study indicate that the SUBDUE system is suitable for knowledge discovery in molecular structural databases. 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. Planned future work applying SUBDUE to the Brookhaven and other related molecular biology databases includes using a more detailed and consistent description of the secondary structure, possibly using resources outside PDB. In addition, 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 discovery of 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 error. 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.

References

[1] 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.

[2] R. C. Bohinski. *Modern Concepts in Biochemistry.* Allyn and Bacon, Inc., 1979.

[3] D. Conklin, S. Fortier, J. Glasgow, and F. Allen. Discovery of spatial concepts in crystallographic databases. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 111–116, 1992.

[4] D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.

[5] D. J. Cook, L. B. Holder, and S. Djoko. Scalable discovery of informative structural concepts using domain knowledge. *IEEE Expert*, 11(5), 1996.

[6] R. E. Dickerson and I. Geis. *Hemoglobin: structure, function, evolution, and pathology.* Benjamin/Cummings Inc., 1982.

[7] S. Djoko, D. J. Cook, and L. B. Holder. An empirical study of domain knowledge and its benefits to substructure discovery. *IEEE Transactions on Knowledge and Data Engineering*, 9(4):575–586, 1997.

[8] G. Galal, D. J. Cook, and L. B. Holder. Exploiting parallelism in a scientific discovery system to improve scalability. *to appear in Journal of the American Society for Information Science*, 1999.

[9] 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.

[10] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymer*, 22:2577–2637, 1983.

[11] 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.

[12] R. Levinson. A self-organizing retrieval system for graphs. In *Proceedings of the National Conference on Artificial Intelligence*, pages 203–206, 1984.

[13] 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.

[14] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, 1989.

[15] J. Segen. Graph clustering and model learning by data compression. In *Proceedings of the Seventh International Machine Learning Workshop*, pages 93–101, 1990.

[16] S. Sprang. Personal Communication, 1998.

[17] S. Su. Structural calculation of an rna hairpin with an adenine bulge. Master's thesis, Louisiana State University, 1996.

[18] K. Thompson and P. Langley. Concept formation in structured domains. In D. H. Fisher and M. Pazzani, editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*, chapter 5. Morgan Kaufmann Publishers, 1991.

[19] D. Voet and J. G. Voet. *Biochemistry.* John Wiley and Sons, Inc., 1990.

[20] K. Yoshida, H. Motoda, and N. Indurkhya. Unifying learning methods by colored digraphs. In *Proceedings of the Learning and Knowledge Acquisition Workshop at IJCAI-93*, 1993.

8 Appendix – Portion of a Sample PDB File

```
HEADER    COMPLEX (RIBONUCLEASE/DNA)              22-MAY-95   1RBJ      1RBJ   2
TITLE     RIBONUCLEASE B COMPLEX WITH D(TETRA-(DEOXY-ADENYLATE))        1RBJ   3
COMPND    MOL_ID: 1;                                                    1RBJ   4
COMPND   2 MOLECULE: RIBONUCLEASE B;                                    1RBJ   5
COMPND   3 CHAIN: A;                                                    1RBJ   6
COMPND   4 SYNONYM: RNASE B;                                            1RBJ   7
COMPND   5 EC: 3.1.27.5;                                                1RBJ   8
COMPND   6 MOL_ID: 2;                                                   1RBJ   9
COMPND   7 MOLECULE: TETRA-(DEOXY-ADENYLATE);                           1RBJ  10
COMPND   8 CHAIN: B;                                                    1RBJ  11
COMPND   9 SYNONYM: D(PA)4                                              1RBJ  12
SOURCE    MOL_ID: 1;                                                    1RBJ  13
SOURCE   2 ORGANISM_SCIENTIFIC: BOS TAURUS;                             1RBJ  14
SOURCE   3 ORGANISM_COMMON: BOVINE;                                     1RBJ  15
SOURCE   4 ORGAN: PANCREAS;                                             1RBJ  16
SOURCE   5 MOL_ID: 2;                                                   1RBJ  17
SOURCE   6 SYNTHETIC: YES                                               1RBJ  18
EXPDTA    X-RAY DIFFRACTION                                             1RBJ  19
AUTHOR    T.-P.KO,R.WILLIAMS,A.MCPHERSON                                1RBJ  20
REVDAT   1   07-DEC-95 1RBJ    0                                        1RBJ  21
JRNL        AUTH   T.-P.KO,R.WILLIAMS,A.MCPHERSON                       1RBJ  22
JRNL        TITL   THE CRYSTAL STRUCTURE OF A RIBONUCLEASE B + D(PA)4   1RBJ  23
JRNL        TITL 2 COMPLEX                                              1RBJ  24
JRNL        REF    TO BE PUBLISHED                                      1RBJ  25
JRNL        REFN                                                  0353  1RBJ  26
REMARK   1                                                              1RBJ  27
REMARK   2                                                              1RBJ  28
REMARK   2 RESOLUTION. 2.7  ANGSTROMS.                                  1RBJ  29
...
REMARK  18 EXPERIMENTAL DETAILS.                                        1RBJ  70
REMARK  18  DATE OF DATA COLLECTION        : 11-SEP-84                  1RBJ  71
REMARK  18  MONOCHROMATIC (Y/N)            : Y                          1RBJ  72
REMARK  18  LAUE (Y/N)                     : Y                          1RBJ  73
REMARK  18  WAVELENGTH OR RANGE (A)        : 1.54                       1RBJ  74
REMARK  18  DETECTOR TYPE                  : CAD4 DIFFRACTOMETER        1RBJ  75
REMARK  18  DETECTOR MANUFACTURER          : ENRAF-NONIUS               1RBJ  76
REMARK  18  INTENSITY-INTEGRATION SOFTWARE : ORESTES                    1RBJ  77
REMARK  18  DATA REDUNDANCY                : 2.                         1RBJ  78
REMARK  18  MERGING R VALUE (INTENSITY)    : 0.04                       1RBJ  79
REMARK  19                                                              1RBJ  80
REMARK  19 SOLVENT CONTENT (VS)      : 52.  %                           1RBJ  81
DBREF  1RBJ A    1   124  SWS     P00656  RNP_BOVIN       27    150     1RBJ  82
SEQRES   1 A  124  LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS MET  1RBJ  83
SEQRES   2 A  124  ASP SER SER THR SER ALA ALA SER SER SER ASN TYR CYS  1RBJ  84
SEQRES   3 A  124  ASN GLN MET MET LYS SER ARG ASN LEU THR LYS ASP ARG  1RBJ  85
SEQRES   4 A  124  CYS LYS PRO VAL ASN THR PHE VAL HIS GLU SER LEU ALA  1RBJ  86
SEQRES   5 A  124  ASP VAL GLN ALA VAL CYS SER GLN LYS ASN VAL ALA CYS  1RBJ  87
SEQRES   6 A  124  LYS ASN GLY GLN THR ASN CYS TYR GLN SER TYR SER THR  1RBJ  88
SEQRES   7 A  124  MET SER ILE THR ASP CYS ARG GLU THR GLY SER SER LYS  1RBJ  89
SEQRES   8 A  124  TYR PRO ASN CYS ALA TYR LYS THR THR GLN ALA ASN LYS  1RBJ  90
SEQRES   9 A  124  HIS ILE ILE VAL ALA CYS GLU GLY ASN PRO TYR VAL PRO  1RBJ  91
SEQRES  10 A  124  VAL HIS PHE ASP ALA SER VAL                          1RBJ  92
SEQRES   1 B    4    A   A   A   A                                      1RBJ  93
FTNOTE   1                                                              1RBJ  94
FTNOTE   1 CIS PROLINE - PRO A    93                                    1RBJ  95
FTNOTE   2                                                              1RBJ  96
FTNOTE   2 CIS PROLINE - PRO A   114                                    1RBJ  97
HELIX    1   1 ALA A    4  HIS A   12  1                                1RBJ  98
HELIX    2   2 TYR A   25  SER A   32  1                                1RBJ  99
HELIX    3   3 LEU A   51  VAL A   54  1                                1RBJ 100
HELIX    4   4 ALA A   56  SER A   59  5                                1RBJ 101
```

```
SHEET    1   A 3 VAL A  43  VAL A  47  0                                1RBJ 102
SHEET    2   A 3 MET A  79  GLU A  86 -1  N   CYS A  84   O  ASN A  44   1RBJ 103
SHEET    3   A 3 TYR A  97  LYS A 104 -1  N   LYS A 104   O  MET A  79   1RBJ 104
SHEET    1   B 4 LYS A  61  VAL A  63  0                                1RBJ 105
SHEET    2   B 4 CYS A  72  GLN A  74 -1  N   GLN A  74   O  LYS A  61   1RBJ 106
SHEET    3   B 4 HIS A 105  ALA A 109 -1  N   VAL A 108   O  TYR A  73   1RBJ 107
SHEET    4   B 4 HIS A 119  VAL A 124 -1  N   VAL A 124   O  HIS A 105   1RBJ 108
SSBOND   1 CYS A   26    CYS A   84                                     1RBJ 109
SSBOND   2 CYS A   40    CYS A   95                                     1RBJ 110
SSBOND   3 CYS A   58    CYS A  110                                     1RBJ 111
SSBOND   4 CYS A   65    CYS A   72                                     1RBJ 112
CRYST1   44.450   44.450  156.500  90.00  90.00  90.00 P 41 21 2     8  1RBJ 113
ORIGX1      1.000000  0.000000  0.000000        0.00000                 1RBJ 114
ORIGX2      0.000000  1.000000  0.000000        0.00000                 1RBJ 115
ORIGX3      0.000000  0.000000  1.000000        0.00000                 1RBJ 116
SCALE1      0.022497  0.000000  0.000000        0.00000                 1RBJ 117
SCALE2      0.000000  0.022497  0.000000        0.00000                 1RBJ 118
SCALE3      0.000000  0.000000  0.006390        0.00000                 1RBJ 119
ATOM      1  N   LYS A   1      36.389  16.107  -8.387  1.00 32.74      1RBJ 120
ATOM      2  CA  LYS A   1      37.658  16.352  -7.691  1.00 32.62      1RBJ 121
ATOM      3  C   LYS A   1      37.503  16.207  -6.162  1.00 31.98      1RBJ 122
ATOM      4  O   LYS A   1      36.585  15.542  -5.629  1.00 35.79      1RBJ 123
ATOM      5  CB  LYS A   1      38.202  17.767  -8.009  1.00 30.19      1RBJ 124
ATOM      6  CG  LYS A   1      37.227  18.915  -7.617  1.00 27.96      1RBJ 125
ATOM      7  CD  LYS A   1      37.436  20.261  -8.367  1.00 24.01      1RBJ 126
ATOM      8  CE  LYS A   1      36.205  21.169  -8.325  1.00 20.51      1RBJ 127
ATOM      9  NZ  LYS A   1      36.522  22.614  -8.380  1.00 19.21      1RBJ 128
ATOM     10  N   GLU A   2      38.295  17.033  -5.500  1.00 24.81      1RBJ 129
ATOM     11  CA  GLU A   2      38.220  17.139  -4.117  1.00 21.16      1RBJ 130
ATOM     12  C   GLU A   2      36.933  17.817  -3.834  1.00 20.44      1RBJ 131
ATOM     13  O   GLU A   2      36.885  19.052  -3.816  1.00 21.89      1RBJ 132
ATOM     14  CB  GLU A   2      39.345  18.064  -3.654  1.00 24.97      1RBJ 133
ATOM     15  CG  GLU A   2      39.480  18.250  -2.130  1.00 27.00      1RBJ 134
ATOM     16  CD  GLU A   2      39.264  16.995  -1.354  1.00 29.64      1RBJ 135
ATOM     17  OE1 GLU A   2      38.489  16.114  -1.726  1.00 30.83      1RBJ 136
ATOM     18  OE2 GLU A   2      39.986  16.973  -0.245  1.00 29.06      1RBJ 137
ATOM     19  N   THR A   3      35.865  17.059  -3.686  1.00 19.53      1RBJ 138
ATOM     20  CA  THR A   3      34.605  17.736  -3.412  1.00 19.30      1RBJ 139
...
ATOM   1036  C4    A B 204      36.984  33.924  -8.337  0.48 16.81      1RBJ1155
TER    1037        A B 204                                              1RBJ1156
CONECT  194  193  642                                                   1RBJ1157
CONECT  310  309  727                                                   1RBJ1158
CONECT  446  445  842                                                   1RBJ1159
CONECT  496  495  547                                                   1RBJ1160
CONECT  547  496  546                                                   1RBJ1161
CONECT  642  194  641                                                   1RBJ1162
CONECT  727  310  726                                                   1RBJ1163
CONECT  842  446  841                                                   1RBJ1164
MASTER       55    4    0    4    7    0    0    6 1035    2    8   11  1RBJ1165
END                                                                     1RBJ1166
END                                                                     1RBJ1166
```