Automated Discovery of Functional Components of Proteins from Amino-acid sequences based on Rough Sets and Change of Representation

Shusaku Tsumoto and Hiroshi Tanaka
Department of Information Medicine, Medical Research Institute
Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku Tokyo 113 Japan
TEL: +81-3-3813-6111 (6159) FAX: +81-3-5684-3618
E-mail:{tsumoto, tanaka}@quasar.tmd.ac.jp

Abstract

Protein structure analysis from DNA sequences is an important and fast growing area in both computer science and biochemistry. Although interesting approaches have been studied, it is very difficult to capture the characteristics of protein, since even a simple protein are made of more than 100 amino acids, which makes biochemical experiments very difficult to detect functional components. For this reason, almost all the problems in this field are left unsolved and it is very important to develop a system which assists researchers on molecular biology to remove the difficulties caused by combinatorial explosions. In this paper we report a system, called MW1 (Molecular biologists' Workbench version 1.0), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically. We apply this method to comparative analysis of lysozyme and α-lactalbumin. The results show that we obtain several interesting results from amino-acid sequences, which have not been reported before.

1. Introduction

Protein structure analysis from DNA sequences is an important and fast growing area in both computer science and biochemistry. Although interesting approaches have been studied, it is very difficult to capture the characteristics of proteins, because even a simple protein has a complex combinatorial structure. For example, let us consider a very small protein made of an one hundred amino-acid sequence (most of the proteins have larger than three hundred sequences). Then, there are $20^{100} \simeq 2^{400}$ kinds of possibilities, because each component of its sequence can take one value from 20 kinds of amino acids. Thus, it is hard to estimate possible structure or chemical properties of proteins from these sequences. What makes the matters worse is that we have only little knowledge about possible function and structure of proteins.

For this reason, almost all the problems in this field are left unsolved because of the above intractable nature caused by complex structure, and it is very important to develop a system which assists researchers on molecular biology to remove the difficulties caused by combinatorial explosions (Hunter 1993). For this purpose, we can introduce a rule induction method, such as AQ15 (Michalski *et al.* 1986) and ID3 (Quinlan 1986). However, applications of such machine learning methods only induce classification rules, which are not sufficient to analyze the functional differences. Therefore we also need to introduce a mechanism which controls the application of domain knowledge in order to analyze the characteristics of induced results and to extract as much information as possible from databases.

In order to incorporate the above control strategy into machine learning methods, we develop a system, called MW1 (Molecular biologists' Workbench version 1.0), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MW1 consists of the following five procedures. First, it exhaustively induces all the classification rules from databases of amino-acid sequences. Second, MW1 changes representation of amino-acid sequences with respect to the main chemical features Then, third, all the rules are induced from each transformed database. Next, fourth, the program estimates the secondary structure of amino-acid sequences via *Chou-Fasman* method (Chou and Fasman 1974). Finally, fifth, MW1 induces all the rules from the databases of secondary structure.

This method is applied to comparative analysis of lysozyme and α-lactalbumin, and the results show that several interesting results are obtained from amino-acid sequences, which has not been reported before. Based on these new discovered knowledge, several experiments are being planned in order to validate discovered results. Interestingly enough, some of them are recently confirmed by biochemical experiments(Tsumoto,K. 1994). The evaluation of other results will be reported when the whole experiments will have been completed.

The paper is organized as follows: Section 2 gives a brief description about our domain: comparative analysis of lysozyme IIc and α-lactalbumin. Section 3 gives discussion on problems on application of empirical learning methods to sequential analysis. Section

Table 1: Primary (Amino-acid) Sequences

Protein	Sequence
α-	KQFTKCELSQLLK@@DIDGYGGIALPELIC
Lactalbumin	TMFHTSGYDTQAIVEN@@NESTEYGLFQIS
	NKLWCKSSQVPQSRNICDISCDKFLDDDIT
	DDIMCAKKIL@DIKGIDYWLAHKALCTEKL
	EOQWL@@CEKL@
Lysozyme	KVFERCELARTLKRLGMDGYRGISLANWMC
	LAKWESGYNTRATNYNAGDRSTDYGIFQIN
	SRYWCNDGKTPGAVNACHLSCSALLQDNIA
	DAVACAKRVVRDPQGIRAWVAWRNRCQNRD
	VRQYVQGC@@GV

"@" denotes "gap" regions after processing multiple alignment procedure.

4 presents the discovery strategy of MW1 and how it works. Section 5 shows the results of application of this system to comparative analysis of lysozyme IIc and α-lactalbumin. Section 6 compares our method with related work. Finally, Section 7 concludes this paper.

2. Lysozyme and α-Lactalbumin

Lysozyme IIc is a enzyme which dissolves the bacterial walls and suppress the growth of bacteria. All living things have this kind of enzyme, and especially, in the category of vertebrate animals, such as fishes, birds, and monkeys, the sequences are almost preserved.

On the other hand, α-lactalbumin functions as a coenzyme of one reaction which dissolves the chemicals in milk into those easy for babies to take nutrition. So this enzyme only exists in the mammals, such as monkeys, and the marsupials, such as kangaroos.

The comparative analysis of these two proteins is one of the most interesting subjects in molecular biology because of the following two reasons (McKenzie and White 1991). First, α-lactalbumin are thought to be originated from lysozyme IIc, since both of the sequences are very similar (Table 1). According to the results of homological search, about 60 % of the sequences of α-lactalbumin matches with those of lysozyme, which suggests that they are of the same origin [1]. In addition to this similarity, the global three-dimensional structure of these two proteins are almost the same. Second, it is not well known what kinds of sequences mainly contributes to the functions of both enzymes, although many experiments suggests that interactions of several components play an important role in those functions.

[1]In this methodology, a amino-acid sequence of a protein of one species, such as cytochrome c of the mammals, only matches with only 25 % of sequences of different species, such as cytochrome c of the reptiles.

Thus, to develop a system which analyzes the differences between two sequences has the following two contributions to molecular biology and computational biology. First, as to molecular biology, the analysis will make it clear what kind of knowledge biological systems acquired through the evolution from birds to mammals. Second, as to computational biology, the analysis will make it clear what kind of mechanisms is useful to analyze the sequences of similar proteins.

3. Problems of Empirical Learning Methods

It is easy to see that simple application of machine learning methods to DNA or amino-acid sequences without using domain-specific knowledge cannot induce enough knowledge.

For example, simple application of induction of decision trees (Breiman *et al.* 1984; Quinlan 1986) generates only one rule from many possible rules. However, many attributes (exactly, 52 attributes) have the maximum value of information gain. Thus, we have to choose one of such attributes. If simplicity is preferred, that is, if the number of leaves should be minimized, then location 44 will be selected as shown below.

$$\begin{cases} 44 = N & \cdots lysozyme & \cdots (45 cases) \\ 44 = V & \cdots \alpha - lactalbumin & \cdots (23 cases) \end{cases}$$

In this case, we get a simple tree, which consists of one node and two leaves. Unfortunately, this result is not enough, since our objective is not to find a simple rule for classification, but to find as much information as possible.

However, exhaustive induction of possible rules also cause another problem: it is very difficult to interpret all the possible rules without using domain knowledge.

Hence it is very crucial to control application of domain knowledge, according to what problem we want to solve. If we need only some evidential knowledge, we should strictly apply domain knowledge, and focus only on several attributes of training samples. These cognitive aspects of machine discovery system are discussed by researchers on machine discovery (Zytkow 1992).

4. Discovery Strategy

In order to implement discovery strategy of molecular biologists, we develop a system, called MW1 (Molecular biologists' Workbench version 1.0), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MW1 consists of the following six procedures. First, it applies PRIMEROSE-EX, discussed in the next subsection, and exhaustively induces all the classification rules from databases of amino-acid sequences. Second, MW1 changes representation of amino-acid sequences with respect to the main chemical features

of amino acids, such as the characteristics of electronic charge (i.e., basic, neutral, or acidic) (**Primary Structure Rearrangement**). That is, MW1 generates new databases focused on a certain chemical property from original databases. Then, third, PRIMEROSE-EX is applied again, all the rules are induced from each database generated by the second procedure. Furthermore, the statistics of each chemical characteristic are calculated. Next, fourth, the program estimates the secondary structure of amino-acid sequences using *Chou-Fasman* method (Chou and Fasman 1974) (**Secondary Structure Rearrangement**). Finally, fifth, MW1 induces all the rules from the databases of secondary structure, applying PRIMEROSE-EX.

4.1 PRIMEROSE-EX

In order to induce rule exhaustively, we introduce a program, called PRIMEROSE-EX (Probabilistic Rule Induction Method based on Rough Sets for Exhaustive induction). This method is based on rough set theory, which gives a mathematical approach to the reduction of decision tables, corresponding to the exhaustive search for possible rules. For the limitation of the space, we only discuss the definition of probabilistic rules of PRIMEROSE-EX and an induction algorithm of this system. Readers, who would like to know further information on rough sets, could refer to (Pawlak 1991; Ziarko 1991; Ziarko 1994).

Rules of PRIMEROSE-EX. In the framework of rough set theory, we have several specific notations as follows. First, a combination of attribute-value pairs, corresponding to a complex in AQ terminology, is denoted by an equivalence relation R. For example, $[a = 1]\&[b = 1]$ will be one equivalence relation, denoted by $R = [a = 1]\&[b = 1]$. Second, a set of samples which satisfies R is denoted by $[x]_R$, corresponding to a star in AQ terminology. For example, when $\{1, 2, 3\}$ is a set of samples which satisfy R, $[x]_R$ is equal to $\{1, 2, 3\}$ [2]. Finally, third, U, which stands for "Universe", denotes the whole training samples.

According to this notation, probabilistic rules are defined as follows:

Definition 1 (Probabilistic Rules) *Let R_i be an equivalence relation, D denote a set whose elements belong to a class d, or positive examples in the whole training samples, U, and $[x]_{R_i}$ denote the set of training samples which satisfies an equivalence relation R_i. Finally, let $|D|$ denote the cardinality of D, that is, the total number of samples in D.*

A probabilistic rule of D is defined as a quadruple, $< R \stackrel{\alpha,\kappa,p}{\rightarrow} d, \alpha, \kappa, p >$, where $R \stackrel{\alpha,\kappa,p}{\rightarrow} d$ satisfies the following proposition:

$$R \stackrel{\alpha,\kappa,p}{\rightarrow} d \ \ s.t. \ \ [x]_R \bigcap D \neq \phi,$$

where α and κ are defined as:

$$\alpha = \frac{|[x]_R \bigcap D)|}{|[x]_R|}, \qquad and \qquad \kappa = \frac{|[x]_R \bigcap D)|}{|D|},$$

and where p is a p-value of χ^2-statistics when the relation between $[x]_R$, D, and U is tested as a contingency table. □

The intuitive meaning of the above three variables, α, κ, and p-value is given as follows. First, α corresponds to the accuracy measure. For example, if α of a rule is equal to 0.9, then the accuracy is also equal to 0.9. Second, κ is a statistical measure of how proportion of D is covered by this rule, that is, coverage or true positive rate. For example, when κ is equal to 0.5, half of the members of a class belongs to the set whose members satisfy that equivalence relation. Finally, third, p-value denotes the statistical reliability of a rule $R \stackrel{\alpha,\kappa,p}{\rightarrow} d$. For example, when p is equal to 0.95, the reliability of the rule is 95%.

As to the calculation of p-value, we view the relation between $[x]_R$, D, and U as a contingency table as shown in the following table.

	d	$\neg d$	Total
R	s	t	$s+t$
$\neg R$	u	v	$u+v$
Total	$s+u$	$t+v$	$s+t+u+v(=n)$

In the above table, $\neg R$ and $\neg d$ denotes the negation of R and d, respectively. Note that each items in the table can be described in the framework of rough set theory, that is, s, t, u, v can be described as $|[x]_R \cap D|(= s)$, $|[x]_R \cap (U - D)|(= t)$, $|D - [x]_R \cap D|(= u)$, and $|(U - D) - [x]_R \cap (U - D)|(= v)$, respectively. It is also notable that $s + t = |[x]_R|$, $s + u = |D|$, and $s + t + u + v = |U|$.

From the above table, χ^2-statistics can be calculated as:

$$\chi^2 = \frac{n(sv - tu)^2}{(s+u)(t+v)(s+t)(u+v)},$$

where n, s, t, u, v is given in the above table. This statistics is a test statistics to check whether R is independent of d. In other words, it indicates whether R is not useful for classification of d or not. From the value of this statistics, p-value is calculated from where this value is located in the χ^2-distribution. For example, when the p-value of χ^2-statistics χ_0 is equal to 0.99, the region whose χ^2-statistics is below χ_0 occupies 99% of the whole distribution. Thus, the probability with which this event will occur is 99%.

According to those values, we classify the induced probabilistic rules into the following four categories:

(1) Definite Rules: $\alpha = 1.0$ and $\kappa = 1.0$,

(2) Significant Rules: $0.5 < \alpha < 1.0$ and $0.9 \leq p < 1.0$

(3) Strong Rules: $0.5 < \alpha < 1.0$ and $0.5 < p < 0.9$,

(4) Weak Rules: $\alpha > 0$.

[2] In this notation, "1" denotes the first(1st) sample in a dataset.

An algorithm for PRIMEROSE-EX. Let D denote training samples of the target class d, or *positive examples*. In the following algorithm, we provide two kinds of specific sets. The one is L_i, which denotes a set of equivalence relations whose size of attribute-value pairs is equal to $i-1$. For example, L_3 includes $[a=1]\&[b=1]$, whereas L_2 includes $[a=1]$ and $[b=1]$. The other is M_i, which denotes a set of equivalence relations for weak rules. For example, when M_2 includes a $[a=1]\&[b=1]$, the accuracy of $[a=1]\&[b=1]$ as to the target concept is lower than 0.5 or the p-value of χ^2-statistics as to the target concept is lower than 0.5. Thus, an equivalence relation in M_i is weak for classification or do not cover enough training samples.

Based on these notations, the search procedure can be described as a kind of the greedy algorithm in the following.

(1) Let L_0 be equal to a set of all the attribute-value pairs $[a_i = v_j]$(selectors in terms of AQ method) and i be equal to 0.

(2) Set M_i to {}, an empty list. Repeat the following three procedures for all the members in a list L_i until L_i is empty. If L_i is empty, goto (7).

(3) Select one pair $R = \wedge[a_i = v_j]$ and check whether $[x]_R \cap D \neq \phi(\alpha > 0)$. If so, then goto (4). Otherwise, remove the pair from L_i, and repeat this procedure again.

(4) Check whether $\alpha > 0.5$. If so, then goto (5). Otherwise, include the pair in a list of weak rules of d, and add this pair to M_i and goto (2).

(5) If $\alpha = 1.0$ and $\kappa = 1.0$, then save this pair as a definite rule of d. Remove the pair from L_i and goto (2). Otherwise, goto (6).

(6) Check the p-value. If $p > 0.9$, register this pair as a significant rule of d. Remove the pair from L_i and goto (2). If $p > 0.5$, register this pair as a strong rule of d. Remove the pair from L_i and goto (2). Otherwise, include the pair in a list of weak rules of d, and add this pair to M_i and goto (2).

(7) If M_i is empty, quit. Otherwise, generate a list of the whole combination of the conjunction formulae in M_i as L_{i+1}. Then increment i(i:=i+1), goto (2).

The above procedure is repeated for all the attribute-value pairs. It is notable that the above algorithm is very similar to discovery of association rules developed by Mannila et al. (Mannila et al.1994). We will discuss the comparison of these two methods in Section 6.

In the above algorithm, equivalence relations for significant rules and strong rules in L_i are removed from candidates for generation of L_{i+1}, because they are not included in M_i. Thus, if significant members of L_i are not included in M_i, then computational complexity of generation of L_{i+1} is small. However, when significant members are included in M_i, then the complexity will be very large. This tendency has already been well studied by (Mannila et al. 1994), although in their approaches the complexity will be large when significant members are not included in M_i. According to Mannila's results, the running time would be linear in the size of training samples, but exponential in the size of M_i. We also discuss this issue later in Section 6.

4.2 Change of Representation

We introduce two kinds of change of representation. One is to generate new databases which focus on a certain chemical characteristic from original databases, called *primary structure rearrangement*. The other one is to transform original databases, according to the estimation of the secondary structure, called *secondary structure rearrangement*.

Primary Structure Rearrangement.
The most important chemical characteristics of amino acids which are thought to contribute to determine a protein structure are the following: hydrophobicity, polarity or electronic charge of a side chain, the size of an amino acid, and the tendency of an amino acid to locate the interior of proteins.

For example, in the case of hydrophobicity, which denotes how much an amino acid is intimate with water molecule, there are two kinds of attribute-value pairs: [*hydrophobicity* = *yes*] or [*hydrophobicity* = *no*] [3]. Using these notations, we can change representation of amino-acid sequences. For example, let us consider a case when an attribute-value pair of an original database is $[33 = F]$, which denotes that the 33th amino acid of a protein is F (phenylalanine). Because phenylalanine (F) is hydrophobic, this attribute-value pair is transformed into: [33 = [*hydrophobicity* = *yes*]]. This procedure is repeated for all the amino-acids in an original sequence.

Secondary Structure Rearrangement.
Next, MW1 estimates secondary structure from amino-acid sequences using the *Chou-Fasman* method (Chou-Fasman 1974), which is the most popular estimation method [4]. This *Chou-Fasman* method outputs the place where specific secondary structures: $\alpha - helix$, $\beta - sheet$, and *turn*. According to this estimation, MW1 changes representation of original databases. For example, the 4th to 10th amino acids are estimated to form an α-helix. Based on the above results, the value of each attribute, which is the address of a primary sequence, are replaced by the above knowledge on secondary structure. In the above example, the values of the 4 th to 10th attributes are substituted for α-helix,

[3] In this paper, we only use these qualitative values, although we also have the coefficients of hydrophobicity, which are quantitative values. It would be our future work to deal with quantitative coefficients.

[4] It is notable that our method is independent of this estimation method. Thus, we can replace the *Chou-Fasman* method with the new methods which may gain more predictive accuracy, when such methods are obtained.

Table 2: Results of Primary Structure Rearrangement

Protein	Amino Acid and its Location		
lysozyme c	N 27	(A,L 31)	K 33
α-lact	E 27	T 31	F 33
lysozyme c	E 35	N 44	(Y,D 53)
α-lact	(I,S,T 35)	V 44	E 53
lysozyme c	(A,G 76)	(A,R 107)	
α-lact	I 76	D 107	
lysozyme c	(G,D,Q 117)	L 129	
α-lact	S 117	E 129	

α-helix, α-helix, α-helix, α-helix, and α-helix. That is,

Primary Structure	E	R	C	E	L	A
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Secondary Structure	α	α	α	α	α	α.

It is notable that some attributes may have no specific secondary structure. In these cases, the value of these attributes are replaced by one of the four characteristics: {hydrophobic, polar, acidic, basic}, since they play an important role in making secondary structure, as discussed in the section on primary structure rearrangement. For example, let us consider a case when an attribute-value pair of an original database is [86 = *D*], which denotes that the 86th amino acid of a protein is D (asparatic acid). Because asparatic acid (D) is acidic, this attribute-value pair is transformed into: [86 = *acidic*] [5].

5. Results and Discussion

We apply MW1 to 23 sequences of α-lactalbumin and 45 sequences of lysozyme from PIR databases, both of which are used as original training samples [6] Then, as inputs of MW1, we use the sequences processed by multiple alignment procedures.

The induced results are shown in Table 2 to 4, where the following three interesting results are obtained [7].

First, Table 2 shows the induced definite rules before change of representation. From the second to sixth columns, alphabets denote amino-acids, and the numbers denote the location in the sequence of a protein.

[5]It is notable that this information can be retrieved from the database generated in the process of primary structure rearrangement.

[6]Readers may say that 68 samples are very small. However, these samples are now all included in Protein databases. As pointed out, most of the datasets collected in genome databases are genes of bacteria, mouse, and other animals which are often used in biochemical experiments. This tendency is one of the difficult problems in genome databases.

[7]The shown results are mainly induced definite rules and significant rules, because including strong and weak rules takes much more space. Thus, due to the limitation of space, we only discuss the results of definite rules and significant rules.

Table 3: Results of Primary Structure Rearrangement with respect to Hydrophobicity

Protein	Location					
	2	4	9	11	33	35
lysozyme c	1	0	1	1	0	0
α-lactalbumin	0	1	0	0	1	1
	44	45	72	73	74	78
lysozyme c	1	0	1	1	0	0
α-lactalbumin	0	1	0	0	1	1
	83	88	92	98	103	106
lysozyme c	1	1	0	1	0	0
α-lactalbumin	0	0	1	0	1	1
	112	114	115	116	118	123
lysozyme c	0	0	0	0	1	0
α-lactalbumin	1	1	1	1	0	1
	129					
lysozyme c	1					
α-lactalbumin	0					

Notations: 1: yes, and 0: no.

Table 4: Results of Secondary Structure Rearrangement

Protein	Location		
	70-77	83-94	98-104
lysozyme c	hydrophobic	hydrophobic	loop
α-lact	polar	acidic	α-helix
	107-110	113-117	
lysozyme c	α-helix	basic	
α-lact	hydrophobic	hydrophobic	

For example, N 27 means that the 27th amino acid of lysozyme IIc is N, or aspargine. These results mean that these amino acids are specific to each protein. In other words, the most characteristic regions are expected to be included. Actually, it is known that E 35, and Y or D 53 are the active site of lysozyme, and also K 33, N 44 and A or R 107 are said to play an important role in its function (McKenzie and White 1991). However, N 27 and L 129 are new discovery results, and no observations or experimental results are reported. Thus, these acids may contribute to the function of lysozyme.

Second, Table 3 shows the results of definite and significant rules after change of representation with respect to hydrophobicity. This table shows that the non-hydrophobic region of 73 to 92th amino acid is specific to α-lactalbumin, and that non-hydrophobic region of 112 to 116th amino acid is specific to lysozyme. The former region corresponds to the binding site of calcium ion, which is a main functional part of α-lactalbumin. However, the function of the latter region is unknown. It may play an important role in the function of lysozyme, because that region easily interacts

with targets and water, which causes the dehydration of targets (Lewin, 1994).

Third, Table 4 shows the results of the definite rules after secondary structure rearrangement. The second row shows the location in sequences, for example, 70-77 means 70th to 77th amino acid in sequences of lysozyme c. Interestingly, although specific amino acids are mainly located at the lower address part (called it N-terminal), specific local structure are mainly located at the higher address part (called it C-terminal). The most significant regions are 98-104 and 113-117, because each secondary structure is very different. Other regions also show that hydrophobic regions of lysozyme correspond to non-hydrophobic regions of α-lactalbumin, and vice versa. Thus, these regions may play an important role in realizing each function [8].

According to these results, they are now planning to validate these results by the experiments based on technique of recombinant DNA. Since it takes about one to three weeks to study the characteristics of one "mutant" protein, we need more that 6 months to confirm our induced results. Readers may say that it takes too much long time for validation, but it is said that we need 10 to 20 years to study the characteristics of the two proteins. Therefore we can save our time to make efficient experiments.

6. Related Work

6.1 Discovery of Association Rules

Mannila et al.(Mannila et al. 1994) report a new algorithm for discovery of association rules, which is one class of regularities, introduced by Agrawal et al.(Agrawal et al. 1993). Their method is very similar to ours with respect to the following two points.

(1) Association Rules. The concept of association rules is similar to our induced rules. Actually, association rules can be described in the rough set framework.

That is, we say that an association rule over r satisfies $W \Rightarrow B$ with respect to γ and σ, if

$$|[x]_W \cap [x]_B| \geq \sigma n$$

and

$$\frac{[x]_W \cap [x]_B}{[x]_W} \geq \gamma,$$

where n, γ, and σ denotes the size of training samples, confidence threshold, and support threshold, respectively. Also, W and B denotes an equivalence relation and a class, respectively. Furthermore, we also say that W is *covering*, if

$$|[x]_W| \geq \sigma n.$$

It is notable that the above formulae correspond to the formula as to κ and p-value shown in 4.1.1, coverage and the formula as to α, accuracy. The only difference is that we classify rules, corresponding to association rules, into three categories: definite rules, significant rules, and strong rules.

The reason why we classify these rules is that this type of classification can be viewed as the ordering of rules or hypothesis. That is, definite rules correspond to the strongest hypotheses. However, these strongest rules may not be interesting for discovery. Then, significant rules will be considered for the candidates of discovery. If they are not so important, then strong rules will be considered. Finally, all the three kinds rules are found to be not important, then we should search for weak rules. In this way, we simulate the discovery strategy of biochemists by using the classification of classification rules.

(2) Mannila's Algorithm. Mannila et al. introduce an algorithm to find association rules based on Agrawal's algorithm. The main points of their algorithms are database pass and candidate generation. Database pass produces a set of attributes L_s as the collection of all covering sets of size s in C_s. Then, candidate generation calculates C_{s+1}, which denotes the collection of all the sets of attributes of size s, from L_s. Then, again, database pass is repeated to produce L_{s+1}. The effectiveness of this algorithm is guaranteed by the fact that all subsets of a covering set are covering.

The main difference between Mannila's algorithm and our MW1 algorithm is that Mannila uses the check algorithm for covering to obtain association rules, whereas we use statistical analysis to compute and classify rules.

In the discovery of association urles, all of the combination of attribute-value pairs in C_s have the property of covering. On the other hand, our algorithm do not focus on the above property of covering. It removes an attribute-value pair which has both high accuracy and high coverage from L_s and does not include in M_s. That is, PRIMEROSE-EX does not search for regularities which satisfy covering, but search for regularities important for classification.

Thus, interestingly enough, when many attribute-value pairs have the covering property, or covers many training samples, Mannila's algorithm will be slow, although PRIMEROSE-EX algorithm will be fast in this case. When few pairs covers many training samples, Mannila's algorithm will be fast, and our system will not be faster.

6.2 Ziarko's KDD-R

Ziarko and Shan develop a comprehensive system for knowledge discovery in databases using rough sets, called KDD-R (Ziarko and Shan 1995b). Their system consists of the four functional units: data processing unit, a unit for analysis of dependencies, a unit for computation of rules from data, and decision unit.

[8] Recently, our collaborating domain experts have got the results, which suggest that 98-104th amino acid plays an important role in lysozyme function(Tsumoto, K. 1994).

The most important unit is one for computation of rules from data. This unit computes all, or some, approximate rules with decision probabilities, where the probabilities are restricted by lower and upper limit parameters specifying the area of user interest. The rules can be computed for a selected reduct using the method of decision matrix (Ziarko and Shan 1995a), which is an extention of discerniblity matrix (Skowron and Rauzer 1992).

The main difference between KDD-R and our system is that PRIMEROSE-EX adopts statistical measures to prune attribute-value pairs. In PRIMEROSE-EX, attribute-value pairs which have high accuracy and high coverage will be used for rule generation and removed from the candidates of complexed rules. On the other hand, KDD-R first removes dependent superfluous attributes using the extension of rough set model, called Variable Precision Rough Set model and then calculates rules using the technique of decision matrix, which is very useful to generate all approximate rules.

Thus, KDD-R focuses mainly on dependencies of attributes with respect to selection of attribute-value pairs, whereas PRIMEROSE-EX focuses on mainly on statistical significance of attribute-value pairs, which is used for selection of attribute-value pairs. Therefore the performance of each system may depend on the characteristics of an applied domain. That is, KDD-R may outperform our method when a dataset has many dependent attributes.

7. Conclusion

In this paper, we propose a system based on combination of a probabilistic rule induction method with domain knowledge, which we call MW1 (Molecular biologists' Workbench version 1.0) in order to retrieve the difficulties from the experimental environments of molecular biologists. We apply this method to comparative analysis of lysozyme and α-lactalbumin, and the results show that we get some interesting results from amino-acid sequences, which have not been reported before.

Acknowledgements

The authors would like to thank Jan Zytkow and the reviewers for insightful comments. This research is supported by Grants-in-Aid for Scientific Research No. 06680343 from the Ministry of Education, Science and Culture, Japan.

References

Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93)*, pp. 207-216.

Breiman, L., Freidman, J., Olshen, R., and Stone, C. 1984. *Classification And Regression Trees.* Belmont, CA: Wadsworth International Group.

Chou, P.Y. and Fasman, G.D. 1974. Prediction of protein conformation. *Biochemistry*, **13**, 222-244.

Dayhoff, M.O. 1972. *Atlas of Protein Sequence and Structure.* Natl. Biom. Res. Foundation, Washington D.C.

Hunter, L.(ed) 1993. *Artificial Intelligence and Molecular Biology*, AAAI press, CA.

Lewin, B. 1994. *Genes V.*, Oxford University Press, London.

Mannila, H., Toivonen, H., Verkamo, A.I. 1994. Efficient Algorithms for Discovering Association Rules, *Proceedings of Knowledge Discovery in Databases (KDD-94)*, pp. 181-192, AAAI press.

McKenzie, H.A. and White, Jr., F.H. 1991. Lysozyme and α-lactalbumin: Structure, Function, and Interrelationships, in: *Advances in Protein Engineering*, pp.173- 315, Academic Press.

Michalski, R.S., et al. 1986. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, *Proceedings of AAAI-86*, 1041-1045, Morgan Kaufmann, CA.

Pawlak, Z. 1991. *Rough Sets*, Kluwer Academic Publishers, Dordrecht.

Quinlan, J.R. 1986. Induction of decision trees, *Machine Learning*, **1**, 81-106.

Skowron, A. and Rauszer, C. 1992. The Discerniblity Matrices and Functions in Information Systems, In Slowinski, R. (ed.) Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory, Kluwer, Dordrecht.

Tsumoto, K. et al. 1994. *Journal of Biochemistry.*

Ziarko, W. 1991. The Discovery,Analysis, and Representation of Data Dependencies in Databases, in: Shapiro,G.P. and Frawley, W.J. (eds.) *Knowledge Discovery in Database*, AAAI press, 1991.

Ziarko, W. (Ed.) 1994. Rough Sets, Fuzzy Sets and Knowledge Discovery. Workshops in Computing, Springer-Verlag, London.

Ziarko, W. and Shan, N. 1995a. A Rough Set-Based Method for Computing All Minimal Deterministic Rules in Attribute-Value Systems, Computational Intelligence **11** (in press).

Ziarko, W. and Shan, N. 1995b. KDD-R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets. *Proceedings of RSSC-94* (in press).

Zytkow, J.M. (Ed.) 1992. *Proceedings of the ML-92 Workshop on Machine Discovery (MD-92).* Wichita, KS: National Institute for Aviation Research.