Comparative Analysis of Amino-acid sequences based on Rough Set Theory 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
E-mail: tsumoto.com@tmd.ac.jp, tanaka@cim.tmd.ac.jp

Abstract

Protein structure analysis from DNA sequences is an important and fast growing area in both computer science and biochemistry. One of the most important problems is that two proteins, both of which have the similar three-dimensional structure, have different functions, such as lysozyme and lactalbumin. In such cases, comparative analysis of both amino acid sequences is effective to detect the functional and structural differences. In this paper, we introduce a system, called MW1.5 (Molecular biologists' Workbench version 1.5), which extracts differential knowledge from amino-acid sequences by using rough-set based classification, statistical analysis and change of representation. This method is applied to the following two domain: comparative analysis of lysozyme and α-lactalbumin, and analysis of immunoglobulin structure. The results show that several interesting results from amino-acid sequences, are obtained 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.

One of the most important problems is that two proteins both of which has the similar three-dimensional structure have different functions, such as lysozyme and lactalbumin. In such cases, comparative analysis of both amino acid sequences is effective to detect the functional and structural differences, since local structure should be of primary importance to contribute to the characteristics of theses proteins.

However, in general, only knowledge from sequences is insufficient for analysis, because protein function is thought to be realized by chemical interaction between the components in amino-acid sequences. That is, it is necessary to incorporate domain knowledge, such as chemical knowledge to make comparative analysis be sufficient. Therefore we 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(Zytkow, 1992).

In order to incorporate the above control strategy into machine learning methods, we introduce a system, called MW1.5 (Molecular biologists' Workbench version 1.5), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MW1.5 consists of the following five procedures. First, it exhaustively induces all the classification rules from databases of amino-acid sequences. Secondly, MW1.5 changes representation of amino-acid sequences with respect to the main chemical features. Then, thirdly, all the rules are induced from each transformed databases. Next, fourthly, the program estimates the secondary structure of amino-acid sequences via *Chou-Fasman* method (Chou and Fasman, 1974). Finally, fifthly, MW1.5 induces all the rules from the databases of secondary structure.

This method is applied to comparative analysis of lysozyme and α-lactalbumin, and analysis of structure of immunoglobulin. 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, 1994; Tsumoto, 1995). The evaluation of other results will be reported when the whole experiments will have been completed.

The paper is organized as follows: Section 2 discusses the problems of empirical learning methods when the method is applied to amino acid sequences. Section 3 presents the discovery strategy of MW1.5 and how it works. Section 4 shows the results of application of this system to comparative analysis of lysozyme IIc and α-lactalbumin, and to analysis of structure of immunoglobulin. Section 5 discussed related work, and finally, Section 6 concludes this paper.

2. 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; Quinlna, 1993) 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.

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

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 causes 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).

3. Discovery Strategy

In order to implement discovery strategy of molecular biologists, we develop a system, called MW1.5 (Molecular biologists' Workbench version 1.5), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MW1.5 consists of the following five procedures. First, it applies PRIMEROSE-EX2, which will be discussed in the next subsection, and exhaustively induces all the classification rules from databases of amino-acid sequences. Secondly, MW1.5 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, MW1.5 generates new databases focused on a certain chemical property from original databases. Then, thirdly, PRIMEROSE-EX2 will be 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, fourthly, the program estimates the secondary structure of amino-acid sequences using *Chou-Fasman* method (Chou and Fasman, 1974) (**Secondary Structure Rearrangement**). Finally, fifthly, MW1.5 induces all the rules from the databases of secondary structure, applying PRIMEROSE-EX2.

3.1 PRIMEROSE-EX2

In order to induce rule exhaustively, we introduce a program, called PRIMEROSE-EX2 (Probabilistic Rule Induction Method based on Rough Sets for Exhaustive induction ver 2.0). 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-EX2 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).

Rules of PRIMEROSE-EX2 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_f, which is defined as follows.

Definition 1 (Equivalence Relation) *Let U be a universe, and V be a set of values. A total function f from U to V is called an assignment function of an attribute. Then, we introduce an equivalence relation R_f such that for any $u, v \in U$, $u \equiv R_f v$ iff $f(u) = f(v)$.*

For example, $[a = 1]\&[b = 1]$ will be one equivalence relation, denoted by $R_f = [a = 1]\&[b = 1]$. Secondly, a set of samples which satisfy R_f is denoted by $[x]_{R_f}$, corresponding to a star in AQ terminology. For example, when $\{1, 2, 3\}$ is a set of samples which satisfy R, $[x]_{R_f}$ is equal to $\{1, 2, 3\}$ [1]. Finally, thirdly, U, which stands for "Universe", denotes the whole training samples.

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

Definition 2 (Probabilistic Rules) *Let R_f be an equivalence relation specified by some assignment function f, D denote a set whose elements belong to a class d, or positive examples in the whole training samples (the universe), U, and $[x]_{R_f}$ denote the set of training samples which satisfy an equivalence relation R_f. 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, $< R_f \overset{\alpha,\kappa,p}{\rightarrow} d, \alpha, \kappa, p >$, where $R_f \overset{\alpha,\kappa,p}{\rightarrow} d$ satisfies the following conditions:

$$(1) \qquad [x]_{R_f} \bigcap D \neq \phi, \tag{1}$$

$$(2) \qquad \alpha = \frac{|[x]_{R_f} \bigcap D)|}{|[x]_{R_f}|}, \tag{2}$$

$$(3) \qquad \kappa = \frac{|[x]_{R_f} \bigcap D)|}{|D|}, \tag{3}$$

$$(4) \quad p : p\textit{-value of } \chi^2\textit{-statistics}, \tag{4}$$

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

where p is a p-value of χ^2-statistics when the relation between $[x]_{R_f}$, 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. Secondly, κ is a statistical measure of how proportion of D is covered by this rule, that is, coverage or a 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, thirdly, p-value denotes the statistical reliability of a rule $R \overset{\alpha,\kappa,p}{\to} d$. For example, when p is equal to 0.95, the reliability of the rule is 95% [2]

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

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

In the above table, $\neg R_f$ 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 $|[x]_{R_f} \cap D|(=s)$, $|[x]_{R_f} \cap (U-D)|(=t)$, $|D - [x]_{R_f} \cap D|(=u)$, and $|(U-D) - [x]_{R_f} \cap (U-D)|(=v)$, respectively. It is also notable that $s+t = |[x]_{R_f}|$, $s+u = |D|$, and $s+t+u+v = |U|$.

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

$$\chi^2 = \frac{n(sv-tu)^2}{(s+u)(t+v)(s+t)(u+v)}, \tag{5}$$

where n, s, t, u, v is given in the above table. This measure is a test statistic 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 of null hypothesis[3] 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 0.01, the region whose χ^2-statistics is below χ_0 occupies 1% 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: $0 < \alpha \leq 0.5$ or $0 < p \leq 0.5$.

An algorithm for PRIMEROSE-EX2 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 shown in Fig. 1. The above procedure is repeated for all the attribute-value pairs. In the above algorithm, equivalence relations for significant rules and strong rules in L_i are removed from candidates for the 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.

How to Deal with Continuous Data Almost all the chemical charcteristics of amino acids are provided as continous data. For example, a coefficient of hydrophobicity of K(Lysine) is equal to 2.27, which means that it takes 2.27 *kcal/mol* energy to remove water around Lysine. Thus, it is necessary to deal with continous data in order to extract knowledge from the chemical characteristics.

For solution, PRIMEROSE-EX2 transforms continous data into categorical data, some of which is similar to C4.5 (Quinlna, 1993), in the following ways. First, provided the attribute a_i, the system sort database with respect to the value of a_i, such as $\{v_1, v_2, v_3, \cdots, v_j\}$, where $v_1 < v_2 < \cdots < v_j$. Secondly, for each member in the above list, the attribute-value pair is translated into the following binary form: $[a_i \leq v_k]$ and $[a_i > v_k]$ $(1 \leq k \leq j)$. Thus, new j binary attributes will be generated. And thirdly, α, κ, and χ^2-statistics will be calculated for each generated binary attribute. Finally, fourthly, the pair which induces the best rule is selected as a candidate of transformation.

However, this tranlation is only effective to each level[4]. That is, for each level, the above transformation algorithm should be peformed.

[2] This definition is different from that in statistical test. In statistical test setting, p-value denotes the probability that **null hypothesis**, or a negation of a hypothesis to be proved, is true. Thus, p-value calculated from statistical distribution, $\hat{p}$ is equal to the probability that null hypothesis is true. On the other hand, in our setting, p-value is equal to $1-\hat{p}$, which denotes the probability that null hypothesis is false.

[3] As discussed above, null hypothesis is negation of the hypothesis to be proved.

[4] In the subsequent sections, a level l is defined as the number of attribute-value pairs in the premise of a probabilistic rule. For example, a level 2 means that the number

Thus, continuous attributes should be always included in a list M in Fig. 1, which is a list of candidates to generate the whole combination of the conjunctive formulae. For example, at the level 2, let a_1, a_2 and a_3 be attributes whose values are continous. Then, when a_1 is included only in a list of weak rules at the level 1 ($[a_1 \leq v_k]$ and $[a_1 > v_k]$)[5], the tranformation of a_1 at this level will be used to generate the combination. That is, an attribute a_1 is fixed to a binary atrribute. Next, the following combination is considered: $a_1\&a_2$, $a_1\&a_3$. Then, the translation procedure is performed for each combination under the condition which $[a_1 \leq v_k]$ or which $[a_1 > v_k]$.

3.2 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, a coefficient is assigned to each amino acids: a value 3.95 is assigned to R(Arginine), which is the least hydrophobic amino acid, and a value -2.27 is assigned to F(Phenylalanine). Therefore, in the latter case, F will be translated into: $[\textit{hydrophobicity} = -2.27]$. 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: $[33 = [\textit{hydrophobicity} = -2.27]]$. This procedure is repeated for all the amino-acids in an original sequence.

Then, for rule induction, the translation procedure introduced in 3.1.3 is used.

Secondary Structure Rearrangement Next, MW1.5 estimates secondary structure from amino-acid sequences using the *Chou-Fasman* method (Chou and Fasman, 1974), which is the most popular estimation method [6]. This *Chou-Fasman* method outputs the place where specific secondary structures: $\alpha - \textit{helix}$, $\beta - \textit{sheet}$, and *turn*. According to this estimation, MW1.5 changes representation of original databases. For example, the 4th to 10th amino acids are estimated to form α-helix structure. 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, α-helix, α-helix, α-helix, α-helix, and α-helix. That is,

Primary Structure	E	R	C	E	L	A
↓	↓	↓	↓	↓	↓	↓
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: $[86 = \textit{acidic}]$ [7].

4. Experimental Results and Discussion

4.1 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 co-enzyme 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. 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. In addition

of attribute-value pairs is equal to 2, such as $[a = 1]\&[b = 1]$.

[5] If no attribute is in list of weak rules, then the attribute which gives the worst rules will be selected.

[6] 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.

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

Table 1: 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	

Table 2: 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	

to this similarity, the global three-dimensional structure of these two proteins is almost the same. Secondly, it is not well known what kinds of sequences mainly contribute to the functions of both enzymes, although many experiments suggest that interactions of several components play an important role in those functions.

We apply MW1.5 to 23 sequences of α-lactalbumin and 45 sequences of lysozyme from PIR databases, both of which are used as original training samples. Then, as inputs of MW1.5, we use the sequences processed by multiple alignment procedures.

The induced results are shown in Table 1 and 2, where the following three interesting results are obtained [8]. First, Table 1 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. 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. Secondly, Table 2 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 [9].

Table 3: Results of IgG sequences

Protein	Location				
	(51)	52A,B,C	(59)	60	61
Glycosamide	(Ile)	Pro	(Tyr)	Ala	Pro
Protein	(Ile)	Lys	(Tyr)	Asn	Glu

4.2 Structure of Immunoglobulin

The main function of Immunoglobulin G(IgG) is as an antibody to specific chemical agent, such as bacterial wall(Lewin, 1994). There are many kinds of IgG, some of which bind small chemicals, other of which bind large proteins. It is thought that such specificities can be determined by characteristics of "variable" region, called CDR-1, CDR-2, and CDR-3(Kabat, 1991). Those IgGs are classified into two categories: those which bind a chain of glycosamides, which is hydrophobic, and those which bind a protein, which is hydrophilic. Thus, it is expected that these characteristics are coded in the "variable" region.

We apply MW1.5 to 1438 sequences of IgG, which consists of 349 IgG specific to hydrophobic chemical agents, 1089 IgG specific to hydrophilic ones.

In this domain, no definite rules are derived, and the most important results are induced as significant rules, shown in Table 3, where Glycosamide and Protein denotes IgG which bind hydrophobic chemicals and IgG which bind hydrophilic chemicals, respectively and where the second row shows the location in sequences. For example, 52 means 52th amino acid in the sequences of IgG. (51) and (59) denotes the common amino acids in both types of immunoglobulin sequences.

Interestingly, in the neighbors of (51) and (59), there

[8] 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.

[9] Tsumoto, K. and Kumagai, I. obtain interesting results, which suggest that 98-104th amino acids play important roles in lysozyme function (Tsumoto, 1994).

exists sequences specific to each type of IgG. As to Glycosamide type, Proline(Pro) seems to play an important role, because Pro is a typical hydrophobic protein. On the other hand, as to Protein type, Lysine(Lys), Asparatic acid(Asp), and Glutamine(Glu) seems to play an important role in its function. These chemical characteristics are also detected by significant rules induced after secondary structure rearrangement: Glycosamide type has a hydrophobic region from 51 to 65 amino acids, but Protein tye has α-helix region in this area [10].

5. Related Work

5.1 Discovery of Association Rules

Mannila et al.(Mannila, 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 (training samples) satisfies $W \Rightarrow B$ with respect to γ and σ, if

$$|[x]_W \cap [x]_B| \geq \sigma n, \qquad (6)$$

and

$$\frac{|[x]_W \cap [x]_B|}{|[x]_W|} \geq \gamma, \qquad (7)$$

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

$$|[x]_W| \geq \sigma n. \qquad (8)$$

It is notable that the left side of the above formulae (6) and (8) correspond to the formula (3) as to κ, coverage, and the left side of the formula (7) corresponds to (2) 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 of 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, the 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 MW1.5 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 rules, all of the combination of attribute-value pairs in C_s have the property of covering. On the other hand, our algorithm does 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.

5.2 Ziarko's KDD-R

Ziarko and Shan develop a comprehensive system for knowledge discovery in databases using rough sets, called KDD-R (Ziarko, 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.

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,1995a), which is an extention of discerniblity matrix (Skoworn 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 re-

[10] Recently, our co-authors have got the results which suggest that Tyr(59) and its neighbors play an important role in function of IgG (Tsumoto, 1995).

```
procedure $PRIMEROSE-EX2$
  var
    $i$ : $integer$; /* Counter */
    $M, L_i$ : $List$;
begin
  $L_1 := \{[a_i = v_j] | [x]_{[a_i=v_j]} \cap D \neq \phi\}$;
            /* a set of all the attribute-value pairs */
            /* $[a_i = v_j]$(selectors in terms of AQ method) */
            /* such that $\alpha > 0$. */
  $i := 1$;
  $M := \{\}$;

  while ( $i := 1$ or $M \neq \{\}$ ) do
    begin
      while ( $L_i \neq \{\}$ ) do
        begin
          Select one pair $R(= \wedge[a_i = v_j])$ from $L_i$;
          $L_i := L_i - \{R\}$;
          if ( $\alpha_R = 1.0$ and $\kappa_R = 1.0$ )
            then Save the quadruple as a Definite rule of $d$;
          if ( $\alpha_R > 0.5$) then
            begin
              Check the $p$-value;
              if ($p > 0.9$), then Register the quadruple as a Significant rule of $d$;
              if ($p > 0.5$), then Register the quadruple as a Strong rule of $d$;
              else /* ($p \leq 0.5$) */
                begin
                  I=nclude the quadruple in a list of Weak rules of $d$;
                  Append $R$ to $M$ ($M := M + \{R\}$)
                end
            end
          else /* ( $\alpha \leq 0.5$) */
            begin
              Include the quadruple in a list of Weak rules of $d$;
              Append $R$ to $M$ ($M := M + \{R\}$)
            end
        end
      $i := i + 1$;
      $L_{i+1}$ := (a List of the whole combination of the conjunctive formulae in $M$)
    end
end {$PRIMEROSE - EX2$}
```

Figure 1: An Algorithm for PRIMEROSE-EX2

moved 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 the 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.

6. Conclusion

In this paper, a system based on combination of a probabilistic rule induction method with domain knowledge is introduced, called MW1.5 (Molecular biologists' Workbench version 1.5) in order to detect the structural differences by using compartive analysis. This method is applied to comparative analysis of lysozyme and α-lactalbumin and to analysis of structure of immunoglobulin. The results show that we get some interesting results from amino-acid sequences, which have not been reported before.

References

Agrawal, R., Imielinski, T., and Swami, A. "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, 1993.

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

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

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

Kabat, E.A. et al.(eds.) *Sequences of Proteins of Immunological Interest*, 5th edition, NIH publication, 1991.

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

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

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

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

Quinlan, J.R. *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, CA, 1993.

Skowron, A. and Rauszer, C. "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, 1992.

Tsumoto, K. et al. "Contribution to antibody-antigen interaction of structurally perturbed antigenic residues upon antibody binding", *Journal of Biological Chemistry*, **269**, 28777-28782, 1994.

Tsumoto, K. et al. "Role of Tyr residues in the contact region of anti-lysozyme monoclonal antibody Hwhel10 for antigen-binding", *Journal of Biological Chemistry*, **270**, 18551-18557, 1995.

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

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