```
procedure PRIMEROSE − EX
  var
    i : integer; /* Counter */
    M, L_i : List;
begin
  L_0 := {[a_i = v_j]|[x]_[a_i=v_j] ∩ D ≠ φ};
           /* a set of all the attribute-value pairs */
           /* [a_i = v_j](selectors in terms of AQ method) */
           /* such that α > 0. */
  i := 0;
  M := {};

  while ( i = 0 or M ≠ {} ) do
    begin
      while ( L_i ≠ {} ) do
        begin
          Select one pair R(= ∧[a_i = v_j]) from L_i;
          L_i := L_i − {R};
          if ( α_R = 1.0 and κ_R = 1.0 )
            then Save the quadruple as a Definite rule of d;
          if ( α_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 ≤ 0.5) */
                begin
                  Include the quadruple in a list of Weak rules of d;
                  Append R to M (M := M + {R})
                end
            end
          else /* ( α ≤ 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 conjunction formulae in M)
    end
end {PRIMEROSE − EX}
```

Figure 1: An Algorithm of PRIMEROSE-REX

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

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

[10] Quinlan, J.R. "Induction of decision trees," *Machine Learning*, **1**, pp.81-106, 1986.

[11] Shan, N. and Ziarko, W. "Data-Based Acquisition and Incremental Modification of Classification Rules", Computational Intelligence **11**, 357-370, 1995.

[12] 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.

[13] Tsumoto, S., Tanaka, H., Tsumoto, K., and Kumagai, I. Comparative Analysis of Amino Acid Sequences based on Rough Sets and Domain Knowledge Hierarchy, *Proceedings of Genome Informatics Workshop 1994*, pp. 59-69, Universal Academy Press, Tokyo, 1994.

[14] Tsumoto, S. and Tanaka, H. PRIMEROSE2: Probabilistic Rule Induction Method based on Rough Sets and Resampling Methods. *Computational Intelligence*, **11**, 389-405, 1995.

[15] 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.

[16] 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.

[17] 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.

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

[19] Ziarko, W. and Shan, N. "KDD-R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets," *Proceedings of RSSC-94*, 1995.

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

Furthermore, it is notable that p-value gives the statistical reliability of induced rules. Thus, MW1 can also estimate the reliability of induced results, which is helpful for molecular biologists. For example, if the averaged p-value of the whole rules is lower than 0.5, then the whole results is not so reliable.

In summary, MW1 improves the following two points of MOLA-MOLA: computational complexity and statistical evaluation.

5.2 Ziarko's KDD-R

Ziarko and Shan develop a comprehensive system for knowledge discovery in databases using rough sets, called KDD-R [19]. 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 [11], which is an extention of discerniblity matrix[12][10].

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.

6 Conclusion

In this paper, a system based on combination of a probabilistic rule induction method with domain knowledge is introduced, called MW1 (Molecular biologists' Workbench version 1.0) 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

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

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

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

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

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

[6] 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.

[7] 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.

[10] Due to the limitation of space, we cannot discuss the ideas of decision matrix and discerniblity matrix, readers could refer to [11, 12] for further information.

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[5]. 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[4]. Those IgG 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 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) denote the common amino acids in both types of immunoglobulin sequences.

Interestingly, in the neighbors of (51) and (59), there exist 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) seem 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 [9].

5 Related Work

5.1 MOLA-MOLA

Formerly, we propose a system based on combination of a probabilistic rule induction method with domain knowledge, which we call MOLA-MOLA (Molecular biological data-analyzer and Molecular biological knowledge acquisition tool). Discovery Strategy of MOLA-MOLA, based on a cognitive model of molecular biologists, consists of the following three procedures. First, MOLA-MOLA applies PRIMEROSE[14] to primary structure of proteins, and induces rules from the sequences without domain knowledge. And then it uses domain knowledge to acquire as much knowledge as possible from primary sequences. Second, MOLA-MOLA estimates secondary structures from primary ones, and transforms primary sequences into secondary sequences. Then the system repeats the above subprocedures: it applies PRIMEROSE without domain knowledge, and then it induces knowledge with domain knowledge. Third, MOLA-MOLA again estimates tertiary structure from secondary ones, and repeat the above subprocedures again. While the above procedure is based on a cognitive model of molecular biologists, the whole speed of MOLA-MOLA is very slow: it takes about 24 hours to induce the whole results by SUN Sparc-10.

The main difference between MOLA-MOLA and MW1 is that rule induction module of MW1, called PRIMEROSE-EX, adopts three kinds of statistical measures to induce four kinds of rules, which causes the computational complexity of rule induction to be decreased. Thus, totally, MW1 performs much faster than MOLA-MOLA: it takes about 10 hours to induce rules by SUN Sparc-10.

[9] Recently, our co-authors have got the results which suggest that Tyr(59) and its neighbors play an important role in function of IgG [16].

Table 1: Results of Primary Structure Rearrangement without Change of Representation

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

Table 2: Results of Secondary Structure Rearrangement

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

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 [7]. 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 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 contribute to the functions of both enzymes, although many experiments suggests that interactions of several components play an important role in those functions.

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. Then, as inputs of MW1, 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 [7]. 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 [7]. 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 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 [8].

[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.

[8] Tsumoto, K. and Kumagai, I. obtain interesting results, which suggest that 98-104th amino acids play important roles in lysozyme function [15].

by [6], 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.

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*.

3.2.1 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 $[hydrophobicity = no]$ [4]. 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 33rd 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.

3.2.2 Secondary Structure Rearrangement

Next, MW1 estimates secondary structure from amino-acid sequences using the *Chou-Fasman* method [2], which is the most popular estimation method [5]. 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, 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: $[86 = acidic]$ [6].

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

[4] 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.

[5] 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.

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

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% [2]

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 $\neg d$ denote the negation of R and d, respectively. Note that each item 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:

$$\chi^2 = \frac{n(sv-tu)^2}{(s+u)(t+v)(s+t)(u+v)}, \tag{1}$$

where n, s, t, u, v is given in the above table. These statistics are test statistics to check whether R is independent of d. In other words, they indicate whether R is not useful for classification of d or not. ¿From the value of these statistics, p-values are calculated from the position where the corresponding value of χ^2-statistics are located in χ^2 distribution. For example, when the p-value of χ^2-statistic χ_0 is equal to 0.99, the region whose χ^2-statistic 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: $0 < \alpha \leq 0.5$ or $0 < p \leq 0.5$.

3.1.2 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 length is equal to $i-1$ [3]. For example, L_3 includes $[a=1]\&[b=1]$, whereas L_2 includes $[a=1]$ and $[b=1]$. The other is M, which denotes a set of equivalence relations for weak rules. For example, when M 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-statistic as to the target concept is lower than 0.5. Thus, an equivalence relation in M 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. It is notable that the above algorithm is very similar to discovery of association rules developed by Mannila et al. [6]. 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

[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] The length of an equivalence relation is defined as the number of attribute-value pairs in a equivalence relation. For example, when $R_f = [a=1]\&[b=1]$, the length of R_f is 2.

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 [2](**Secondary Structure Rearrangement**). Finally, fifth, MW1 induces all the rules from the databases of secondary structure, applying PRIMEROSE-EX.

3.1 PRIMEROSE-EX

In order to induce all the rules 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 [9, 17, 18].

3.1.1 Rules of PRIMEROSE-EX

In the framework of rough set theory, we have several specific notations as follows. First, U, which stands for "Universe", denotes the whole training samples. Second, a combination of attribute-value pairs, which corresponds to a complex of selectors in AQ terminology[8], 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 $f(u) = f(v)$.* □

Finally, third, a set of samples which satisfies R_f is denoted by $[x]_{R_f}$, corresponding to a star in AQ terminology. For example, when $\{1, 2, 3\}$ is the set of samples which satisfy R_f, $[x]_{R_f}$ is equal to $\{1, 2, 3\}$ [1].

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 satisfies 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, $< R_f \stackrel{\alpha,\kappa,p}{\rightarrow} d, \alpha, \kappa, p >$, where $R_f \stackrel{\alpha,\kappa,p}{\rightarrow} d$ satisfies the following conditions:

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

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

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

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

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. 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

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

in order to analyze the characteristics of induced results and to extract as much information as possible from databases[20].

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 databases. Next, fourth, the program estimates the secondary structure of amino-acid sequences via *Chou-Fasman* method [2]. 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 analysis of structure of immunoglobulin. The results show that several interesting results are obtained from amino-acid sequences, which have 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 [15, 16]. 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 discussion on problems on application of empirical learning methods to sequential analysis. Section 3 presents the discovery strategy of MW1 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 6 compares our method with related work. Finally, Section 7 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 [1, 10] 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 [20].

3 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 five 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

Automated Discovery of Protein Functional Units from Amino-acid sequences using Rough-Sets-based Comparative Analysis

S. Tsumoto[1] H. Tanaka[1]
tsumoto@tmd.ac.jp tanaka@tmd.ac.jp
K. Tsumoto[2] I. Kumagai[2]

[1] Department of Information Medicine, Medical Research Institute,
Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113 Japan

[2] Department of Biochemistry and Engineering,
Graduate School of Engineering, Tohoku University

Abstract

Protein structure analysis from DNA sequences is an important and fast growing area in both computer science and biochemistry[3]. 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 (Molecular biologists' Workbench version 1.0), 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.

Keywords: Knowledge Acquisition, Machine Learning, Rough Sets,
Change of Representation, Comparative Analysis

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 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, 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

[1]津本周作, 田中博: 東京医科歯科大学難治疾患研究所医薬情報, 〒 113 東京都文京区湯島 1-5-45
[2]津本浩平, 熊谷泉: 東北大学大学院工学研究科生物工学専攻, 〒 980 仙台市青葉区荒巻字青葉