Comparative Analysis of Amino Acid Sequences based on Rough Sets and Domain Knowledge Hierarchy

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 Industrial Chemistry, Faculty of Engineering,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku Tokyo 113 Japan

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 have a complex combinatorial structure, which makes biochemical experiments very difficult to detect functional components. For this reason, almost all the problems about this field are left unsolved and it is very important to develop a system which assists researchers on molecular biology to remove the difficulties by a combinatorial explosion. In this paper, 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) in order to retrieve the hassles 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 has 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

[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, 東京都文京区湯島 1-5-45
[2]津本浩平, 熊谷泉, 東京大学工学部工業化学科 東京都文京区本郷 7-3-1

very difficult to capture the characteristics of protein, since even simple proteins have complex combinatorial structure. There are 20 amino acids, and even small proteins have about 100 amino-acid sequences, so the search space is almost equal to $20^{100} \simeq 2^{400}$. Actually, molecular biologists are now facing with many problems about experiments caused by combinatorial explosions. For example, even now we cannot exactly determine relations between a sequence and a function only by using physical−chemical knowledge about amino acids, so we have to search for some sequences which is similar to the target sequence and has been already well studied. For this purpose, we apply homological search methods, but what makes the problems difficult is that similar sequences do not always guarantee similarities about functions. Therefore we have to perform many experiments in order to detect the relations by trial and errors. These experiments needs technique of recombinant DNA, but we should focus on the place to substitute normal DNA sequences by non-normal ones because of huge search spaces: the selection of location for substitution may cause combinatorial explosion. Moreover, the selection of non-normal sequences may also generate another type of combinatorial explosion.

For this reason, almost all the problems about this field are left unsolved because of those intractable nature, and it is very important to develop a system which assists researchers on molecular biology to remove the difficulties caused by combinatorial explosion [4].

In this paper, we propose an approach to retrieve the hassles from the experimental environments of molecular biologists. For this purpose, we introduce a rule induction method based on rough sets, which we call PRIMEROSE [10]. However, as shown below, the original PRIMEROSE is too powerless for our purpose. Therefore we also introduce representational hierarchy and hypothesis hierarchy in order to augment our induction methods. Furthermore, we introduce a mechanism which controls the application of domain knowledge to hierarchical representations and which generates hypothesis hierarchy. 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 has not been reported before. Based on these new discovered knowledge, we are now planning some experiments of biochemistry in order to validate our results. Experiments will be started this September, and evaluation of our induced results will be reported when the whole experiments will have been completed.

The paper is organized as follows: in Section 2, we give a brief description about our domain: comparative analysis of lysozyme IIc and α-lactalbumin. In Section 3, we briefly discuss about our rule induction method, which we call PRIMEROSE(Probabilistic Rule Induction Method based on Rough Sets). Section 4 gives discussion on problems on application of empirical learning methods to sequential analysis, and how to use domain knowledge, and Section 5 presents the discovery strategy of MOLA-MOLA and how it works. Finally, in Section 6, we show the results of application of this system to comparative analysis of lysozyme IIc and α-lactalbumin.

2 Lysozyme and α-lactalbumin

Lysozyme IIc is a enzyme which dissolves necrotic tissue in a body space of living things, such as nose. Simply speaking, it transforms dirty trashes difficult to remove into ones easy to clean. All living things have this kind of enzyme, and especially, in the category of vertebrate animals, such as fishes, birds, monkeys, the sequences are almost preserved. That is, this lysozyme IIc

evolves very slowly in terms of molecular evolution. This suggests that almost all the sequences are very important to maintain its function, according to the theories of molecular evolution.

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 mammals, such as monkeys.

This comparative analysis is one of the most interesting subjects in molecular biology because of the following three reasons. 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. In this methodology, even 25 % match is excellent, so the above results suggest that they have the same origin. In addition to this similarity, the global structure of these two proteins are the same, like a soccer ball (called **globular protein**). Second, although the active site of lysozyme, which is defined as the place to determine the function of an enzyme, has been already determined exactly, it has been shown that this site is not only the factor to preserve its function. For example, even if we substitute a few amino acids of α-lactalbumin, which are located at the place corresponding to the active site of lysozyme, by the amino acids specific to lysozyme, we cannot get a lactalbumin product which has the same function as lysozyme. It suggests that some complex interactions between amino acids are indispensable to achieving those functions. Third, the active site of α-lactalbumin has not been found, and it is unknown what parts of the sequences of amino acids are important for function.

Therefore, in this paper, we analyze the sequences of both proteins to discover the cause of the difference in functions of these two proteins via computational methods.

3 Rule Induction Method based on Rough Sets

Rough set theory is developed and rigorously formulated by Pawlak[8]. This theory can be used to acquire certain sets of attributes which would contribute to class classification and can also evaluate how precisely these attributes are able to classify data.

For example, let C denote a set whose elements belong to a certain class c and be equal to {1,2,3,4,5}. Then if we have a set {1,2,3} which satisfies an equivalence relation R_1, then we say that R_1 classfies c correctly. Because the relation between {1,2,3} and C corresponds to a proposition, $R_1 \rightarrow c$. We describe these relations in terms of rough set theory as follows:

$$R_1 \rightarrow c \quad iff \quad [x]_{R_1} \subseteq C,$$

where $[x]_{R_1}$ denotes a set which satisfies an equivalence relation R_1.

In the same way, it is possible to formulate partial classification. For example, if we have a set {1,3,4,5,7} which satisfies an equivalence relation R_2 and which include a element(e.g.,7), which does not belong to the class c. However, we have common elements between $[x]_{R_2}$ and C, so we say that R_2 classfies c partially. We can describe this relation in terms of variable precision rough set theory, which is the extension of originial rough set theory [11], as follows:

$$R_2 \xrightarrow{\beta} c \quad iff \quad [x]_{R_2} \bigcap C \neq \phi \quad and \quad \beta = 1 - \frac{card \quad [x]_{R_2} \cap C}{card \quad [x]_{R_2}},$$

where β denotes the misclassfication rate of R_2. So, this means that if a case satisfies R_2, then this case belong to c with the accuracy $1 - \beta$.

In this way, we can develop rule induction(classfication) method based on a set-theoretic apporoach, which is one of the most important features of *rough set theory*. Readers, who would like to know other interesting characteristics of rough sets. could refer to [8].

The above two classfication shows two important characteristics of clasfication: the former is deterministic, and the latter is probabilistic. While the former proposition is desirable because of its certainity, we sometimes have to deal with partial classfication because of the probabilistic nature of a molecular biological domain.

So we use the latter extended definition of the proposition. Furthermore, in order to estimate induced rules, we introduce two statistical measures. Our definition of probabilistc rules is shown as follows:

Definition 1 (Definition of Probabilistic Rules) *Let R_i be an equivalence relation and D denotes a set whose elements belong to a class d and which is the subset of U. A probabilistic rule of D is defined as a tuple, $< R_i \stackrel{\beta}{\rightarrow} d, SI(R_i, D), CI(R_i, D) >$ where $R_i \stackrel{\beta}{\rightarrow} d$ satisfies the following proposition:*

$$R_i \stackrel{\beta}{\rightarrow} d \quad iff \quad [x]_{R_i} \bigcap D \neq \phi \quad and \quad \beta = 1 - SI(R_i, D),$$

and where SI and CI are defined as:

$$SI(R_i,D) = \frac{card\ ([x]_{R_i} \bigcap D)}{card\ [x]_{R_i}}, \qquad and \qquad CI(R_i,D) = \frac{card\ ([x]_{R_i} \bigcap D)}{card\ D}.$$

□

SI corresponds to the accuracy measure defined by Pawlak [8]. For example, if SI of a rule is equal to 0.9, then the accuracy is also equal to 0.9. On the other hand, CI is a statistical measure of how proportion of D is covered by a rule. For example, if CI is equal to 0.5, then half of the elements of a class belongs to the set whose elements satisfy that equivalence relation.

We developed a system, which we call PRIMEROSE (Probabilistic Rule Induction Method based on Rough Sets), which induces the above type of probabilistic rules from databases [10]. While PRIMEROSE is useful to rule induction in probabilisitc domains, it is powerless to apply to sequential analysis in molecular biology as shown in the next section. Therefore we introduce meta-system which control strategy of PRIMEROSE as shown in section 5 and 6.

4 How to Use Domain Knowledge

4.1 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 our example, AQ-15 (set to save 100 rules) [7] and PRIMEROSE [10] can generate more than 100 rules for classification. It is because there are too many attributes, although the number of target classes is only two, and because many attributes have the same classification power. Furthermore, these rules consist of only one [attribute−value] pair and only show what amino acid can be used for classification. So, from those "fragmental" rules, we have to extract

more structural knowledge. However, these two methods is useful in the sense that they can induce the whole rules from DNA or amino-acid sequences, if we do not use some domain-specific criterion or if there are many optimal attributes under such criterion. On the contrary, in this situation, simple application of induction of decision tree [1, 9] gives us some difficulties. Many attributes(exactly, 52 attributes) have the maximum value of information gain. So 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.

$$\begin{cases} location44 = N & \cdots lysozyme & \cdots (45cases) \\ location44 = V & \cdots \alpha - lactalbumin & \cdots (23cases) \end{cases}$$

In this case, we get a simple tree, which consists of one node and two leaves. This result is much more useless than those of AQ and PRIMEROSE, since our objective is not to find a simple rule for classification. Readers may say that these difficulties will be solved by transforming this simple representation into suitable one. However, in general, choosing suitable representation needs well-defined domain-specific knowledge. As mentioned above, we will face with difficulties caused by combinatorial explosion without domain knowledge. However, if we use domain knowledge strictly, then much interesting information which could be sources of discoveries will be eliminated, and only some evident knowledge will be acquired. So, we cannot fully avoid generating all of the rules which are consistent with training samples.

Hence it is very crucial to control application of domain knowledge, according to what problem we want to solve. If we need only some knowledge, we should strictly apply domain knowledge, and focus only on some attributes of training samples. These cognitive aspects of machine discovery system is discussed by researchers on machine discovery [12]. Here we assume that the cognitive strategy of molecular biologists is mainly modeled by the following process: first, they make all the possible solutions without domain knowledge, second, they apply domain knowledge and interpret these solutions. Then they change representation by applying domain knowledge, and repeat the above first and second procedures, based on this representation.

4.2 Representational Hierarchy

The one is hierarchy of [attribute−value] pair representation, which is based on general domain knowledge of molecular biology.

Molecular biologists use representational hierarchy to describe protein structure, and this hierarchy consists of the following four level. First is called a **DNA-sequence level**, which corresponds to DNA sequences of a protein. Sequences are described by [location−the kind of DNA] pairs as shown in fig.2. Second is called a **primary-structure level**, on which amino-acid sequences are represented by combination of [location−the name of amino acid]. Those amino-acid sequences are determined by codon triplets. Third is called a **secondary-structure level**. This level is represented by the sequence of specific 3-D structure, such as α-helix, β-sheet. These specific 3-D structure are integrated into 3-D structure of proteins, which is the fourth level, called **tertiary-structure level**.

This hierarchy is based on representational issues. There are very useful, since on each level we can use different kind of physical-chemistry knowledge. For example, in the primary-structure level, since chemical characteristics of amino acids classify amino acids into some

categories: since Asparatic acid (D) and Glutamic acid (E) are acidic, so they are included in a set of "acidic" amino acids. Therefore these knowledge is available for comparison of amino-acid sequences.

4.3 Hypothesis Hierarchy

We introduce two axis to represent hierarchy of hypothesis. The first axis is based on representational hierarchy, which is mentioned in the above subsection. Applying PRIMEROSE methods to each sequence of representation, we can induce rules for each hierarchical level. On the other hand, the second axis is based on the level of constraints. As discussed in Section 3, controlling usage of domain knowledge is very important to get suitable hypothesis. And this axis supports that control, which is discussed in the next section. Here, we set four levels of usage of constraints. The first level is no application of domain-specific knowledge. In this level, only syntactical knowledge is induced, so most of those knowledge is meaningless in terms of semantics of biological domains. The second level is usage of primary-type of domain knowledge. Primary-type means knowledge about each unit, and in this level, interactions between units are not considered. The third level is application of second-type of domain knowledge. Second-type means knowledge about interactions between each unit and its neighbors. So, only local interactions between units are considered. In this domain, we define the boundary of neighborhood as 3. For example, location 54 are included in an neighborhood of location 51, but location 55 is not in this neighborhood. Readers may say that this definition of boundary is a little weak, but it is based on heuristics of molecular biologists. The fourth level is consideration of third-type of domain knowledge. At this level, knowledge about remote effect are included in the constraints.

5 Discovery Strategy of MOLA-MOLA

In order to discover the functional components of two enzymes, lysozyme IIc and α-lactalbumin, we developed a system MOLA-MOLA (MOLecular biology data-Analyzer and MoLecular biology knowledge Acquisition Tool), which supports not only our problems, but also other problems on protein structure analysis, such as detection of the active site.

Discovery Strategy of MOLA-MOLA is shown in Fig. 1. As shown in the above section, this strategy is based on a cognitive model of molecular biologists. First, we apply PRIMEROSE to primary structure of proteins, and induce rules from the sequences without domain knowledge. And then we use domain knowledge to acquire as much knowledge as possible from primary sequences. Second, we estimate secondary structures from primary ones, and transform primary sequences into secondary sequences. Then we repeat the above subprocedures: we apply PRIMEROSE without domain knowledge, and then we induce knowledge with domain knowledge. Third, we again estimate tertiary structure from secondary ones, and repeat the above subprocedures again.

Primary-Structure-Level

From amino-acid sequences, first, we calculate various kinds of statistical measures, such as the composition of amino-acids of two proteins. These are now used for molecular biologists, which

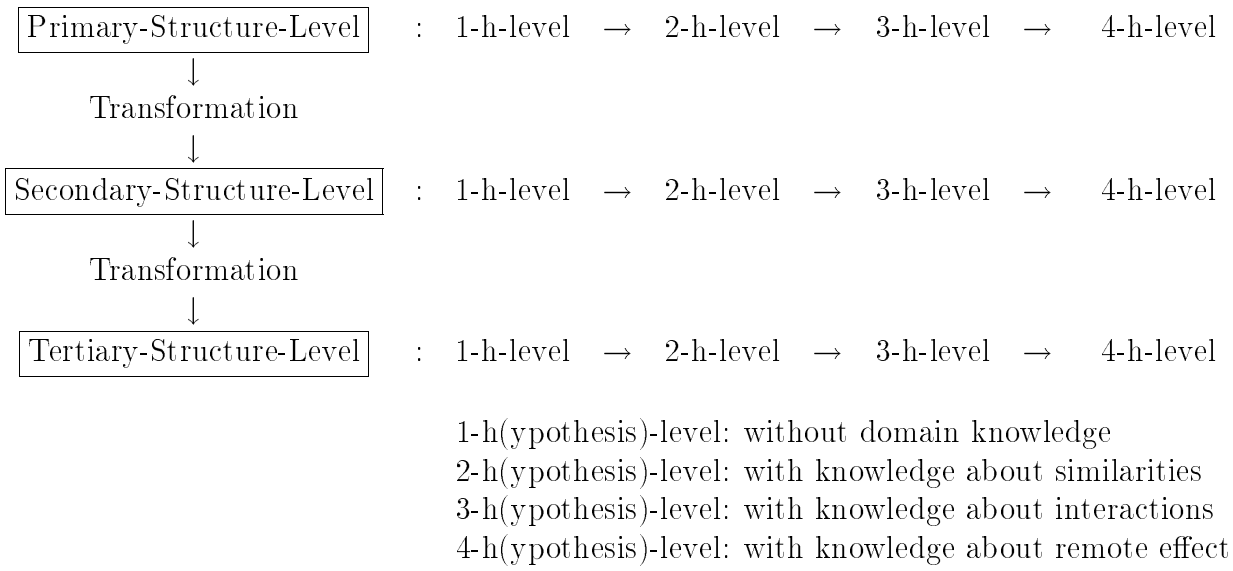

Figure 1: Discovery Strategy of MOLA-MOLA

we are now planning to use those measures to derive global information on these sequences. Second, we apply PRIMEROSE to amino-acid sequences without domain-knowledge. Induced rules are composed of proposition and two statistical measures (SI and CI), as shown in the appendix B. Those Rules are first ordered by the value of CI, and then those which have the same value of CI are ordered by the value of SI. They always include trivial solutions, since similarities between amino acids are not considered (**1-hypothesis-level**). Then we apply primary constraints, such as similarities of amino acids, to these results, and removes trivial ones. These solutions are in the **2-hypothesis-level**. For the above example, PRIMEROSE induces more than hundred rules from the databases of lysozyme and α-lactalbumin. Rules which satisfy SI=1.0 and CI=1.0 are shown in Table 1. These rules are not trivial, however, rules that CI is less than 1.0 have many trivial solutions. For instance, SI and CI of the rule: "[84=L] $\rightarrow$ Lysozyme" are equal to 0.95 and 0.89 respectively, and those of the rule: "[84=F] $\rightarrow$ α-lactalbumin" are equal to 1.0 and 0.86 respectively. These values are very high, but since L and F share the similar characteristics, they will be removed from the 2-hypothesis level. Next, we apply secondary constraints, which includes some knowledge on interaction between neighbors, and generates **3-hypothesis-level**. This procedures is started from rules which have the high values of SI. For example, in the case of the following rules: "[92=A] $\vee$ [92=V] $\rightarrow$ Lysozyme" (SI=1.0, CI=0.89) and "[92=D] $\rightarrow$ α-lactalbumin" (SI=1.0, CI=0.95), the analysis is started from this address, and the following rule is obtained:
$[86 = D]\&[87 = D]\&[88 = D]\&[89 = L]\&[90 = T]\&[91 = D]\&[92 = D]\&[93 = I]\&[94 = M] \rightarrow \alpha - lactalbumin$ (SI=1.0, CI=0.89), and
$[86 = S]\&[87 = D]\&[88 = I]\&[89 = I]\&[90 = A]\&[91 = K]\&[92 = A]\&[93 = V]\&[94 = A] \rightarrow lysozyme$ (SI=1.0, CI=0.62).
Finally, we apply tertiary constraints, which contains knowledge on remote effects, and which restricts our focus of attention. Then **4-hypothesis-level** are obtained. For example, in the case of the above two rules, the former is acidic and the latter is hydrophobic as to the affinity

to water. So the latter region tends to escape from contacting with water.

We store induced results at these four levels, because the assumptions of higher levels may be wrong. If so, we have to proceed the analysis at the lower level.

Secondary-Structure-level

Next, we are going up to a higher representational level. First, we change representation of attributes by applying the *Chou-Fasman* method [2] to primary amino-acid sequences. Then, we obtain prediction of secondary structure for each amino-acid sequence. For example, the 4th to 10th amino acids form α-helix. Based on the above results, we replace the value of each attribute, which is the address of a primary sequence, by the above knowledge on secondary structure. For the above example, we replace the values of the 4 th to 10th attributes by α-helix, α-helix, α-helix, α-helix, α-helix, and α-helix. That is,

primary-structure-level	E	R	C	E	L	A
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
secondary-structure-level	α	α	α	α	α	α

Some attributes have no specific secondary structure. In this case, we replace the value of these attributes by one of the four characteristics: { hydrophobic, polar, acidic, basic }, since they play an important role in making secondary structure.

Then, we apply PRIMEROSE and hypothesis-hierarchy again to these transformed sequences. We obtain rules at each hypothesis level.

For our problem, the induced results at the 1-hypothesis-level are shown in Table 2. Then, we consider the similarities between these sequences. For example, α-helix and hydrophobic region are similar, since these regions are rich in hydrophobic amino acids. So, their behavior is almost the same, except for compactness in 3-D space: α-helix is more compact than general hydrophobic region. Therefore rules about location 107-110 are removed from knowledge in the 2-hypothesis-level. Next, we consider interaction between these regions. While location 83-94, 98-104, 107-110, and 113-117 of lysozyme form a specific complex structure, those of α-lactalbumin make a hydrophobic structure. So these two regions are expected to be very different in tertiary structure. Finally, we consider remote effects of the above region, and then interpret the characteristics of this part. These results of 3-hypothesis and 4-hypothesis level is summarized in Table 3.

Tertiary-Structure-level

In the same way, we can go up to a higher level: first, sequences at the lower level are processed, and the knowledge needed for analysis at this level are obtained. Second, we change representation of attributes according to those derived results. Third, we apply PRIMEROSE to the new table, and PRIMEROSE induced some rules. In this stage, since we use hypothesis hierarchy, hierarchical hypotheses are calculated.

Unfortunately, we have no accurate procedure that determines tertiary structure from secondary structure, and we only know tertiary structures of proteins which can be crystallized. Hence, in the present version, PRIMEROSE do not have procedures at this level.

Table 1: Results of Primary Structure Level and 1-Hypothesis-Level

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	(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 Level and 1-Hypothesis-Level

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

However, for our problem, tertiary structures of lysozyme and α-lactalbumin have already been analyzed. So these knowledge are given as problem-specific knowledge, and we do not have to execute this process.

6 Induced Results and Evaluation of our system

Third, Table 3 shows the result of secondary-structure-level and 3,4-hypothesis level. So, this result suggests that the higher location play an important role in the function of lysozyme.

We applied MOLA-MOLA to 23 sequences of α-lactalbumin and 45 sequences of lysozyme from PIR databases, both of which are used as training samples. And as inputs of MOLA-MOLA, we use the sequences as inputs which are processed by multiple alignment procedures and in which gaps are inserted.

The induced results are shown in the following four tables. First, Table 1 shows the output of the first procedure, whose induced rules satisfies SI=1.0 and CI=1.0. 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 proteins. That is, 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. However, N 27 and L 129 are new discovery results, and no observations or experimental results are reported. These acids may contribute to the function of lysozyme.

Second, Table 2 shows the output of the second procedure. The second row shows the location in sequences, for example, 70-77 means 70th to 77th amino acid in sequences of lysozyme c. Interestingly, while 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

Table 3: Results of Secondary Structure Level and 3, 4-Hypothesis-Level

Protein	Location
	83-94 & 98-104 & 107-110 & 113-117
lysozyme c	a complex folding structure including one loop
α-lactalbumin	a simple hydrophobic region with a acidic region

Table 4: Statistics of Sequential Analysis

			Hydrophobicity	
Estimated Exon	Match	Rules	lysozyme	α-lactalbumin
1- 22(22)	11(50.0%)	3	6	4
23- 77(55)	25(45.4%)	12	19	18
78-104(27)	7(25.9%)	4	14	9
105- (26)	1(3.8%)	8	9	10

it C-terminal). The most significant regions are 98-104 and 113-117, since 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. So these regions may play an important role in realizing each function. Fourth, the statistics of each estimated exon structure are shown in Table 4. Intuitively, exon denotes semi-global structure of a protein. This semi-global structure is said to be related with functional domain of a protein and with acquiring new function by evolution technique "exon shuffling" [5].

These statistics also support the above three tables. In the second column, the number of match of amino acids between lysozyme and α-lactalbumin are given. From this table, it is notable that third and fourth exon are very different, compared with first and second exon. The third column presents induced rules which satisfies SI=1.0 and CI > 0.5, supporting the result of the second column. The fourth and the fifth column show the averaged number of hydrophobic amino acids, and the third exon of lysozyme is different from that of α-lactalbumin. Hence these statistics suggest that the third and the fourth exon should be contributed to the functional difference between these two proteins.

We 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.

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**, 222-244, 1974.

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

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

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

[6] McKenzie, H.A. and White,JR., F.H. Lysozyme and α-lactalbumin: Structure, Function, and Interrelationships, in: *Advacnces in Protein Engineering*, pp.173- 315, Academic Press, 1991.

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

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

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

[10] Tsumoto, S. and Tanaka, H. PRIMEROSE:Probabilistic Rule Induction based on Rough Sets and Resampling MEthods, *Proc. of RSKD'93*, 1993.

[11] Ziarko,W. Variable Precision Rough Set Model, *Journal of Computer and System Sciences*, **46**, 39-59, 1993.

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