Discovery of Functional Components of Proteins from Amino-acid sequences based on Rough Sets and Hierarchical Reasoning

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

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

1. Introduction
Protein structure analysis from DNA sequences is an important and fast growing area in both computer science and biochemistry. Although interesting approaches have been studied, it is very difficult to capture the characteristics of protein, because even a simple protein has a complex combinatorial structure. For example, let us consider a case when a protein is made of an one hundred amino-acid sequence. Then, there are \( 20^{100} \approx 2^{100} \) kinds of possibilities, because each component of its sequence can take one value from 20 kinds of amino acids.

Therefore, molecular biologists are now facing with combinatorial explosions. For example, since we cannot exactly determine relations between a sequence and a function only by using physical and chemical knowledge about amino acids even now, we have to search for some well-studied sequences which are similar to the target sequence. For this purpose, we apply homologous search methods, which calculate the similarities between two proteins. However, what makes problems difficult is that sequential similarities do not always guarantee functional similarities. Therefore we have to perform many experiments in order to detect the relations. These experiments need technique of recombinant DNA, which replace DNA components with other components and which is used to study the effect of substitution. However, we should focus on the place to substitute normal DNA components for non-normal ones because of huge search space discussed above.

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

In order to incorporate the above control strategy into machine learning methods, we develop a system, called MOLA-MOLA (Molecular biological data-analyzer and Molecular biological knowledge acquisition tool), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MOLA-MOLA consists of the following six procedures. First, it exhaustively induces all the rules which can be used for classification of proteins from databases of amino-acid sequences. Second, MOLA-MOLA changes representation of amino-acid sequences with respect to the main chemical features of amino acids. Then, third, all the rules are induced from each database transformed by the second procedure and the statistics of each chemical characteristic are calculated. Next, fourth, the program estimates the secondary structure of amino-acid sequences via Chou-
Protein Structure and Our Problem

2.1 Protein Structure

In this subsection, we briefly mention about protein structure and a mechanism of protein synthesis.

It is well known that our bodies are made of proteins, whose codes are described as long DNA sequences. At the first stage of protein synthesis, DNA sequences are translated into amino-acid sequence. In this sequence, a precursor, which is like a header of a database file in computer science, is attached. This precursor is needed to carry DNA information to the factory of protein synthesis, called ribosome. At ribosome, this precursor is removed. The amino-acid sequence without that precursor is called primary structure, which is shown in Table 1. This table shows two primary sequences of human proteins, one is $\alpha$-lactalbumin, and the other is lysozyme. The details of both proteins are mentioned in the next subsection.

Each amino acid has its own chemical characteristics. For example, arginine functions as a base in water since arginine has a basic component. These characteristics make interactions between amino acids, which causes the folding of the primary sequence, which is called secondary structure. Then, interactions between secondary structure makes 3-D structure of protein, called tertiary structure.

Unfortunately, it is difficult to predict tertiary structure from primary and secondary structure, although there have been developed many approaches (Hunter 1993). It is partly because we do not know much about interactions between the units of secondary structure, and partly because estimation of secondary structure may lose information important to predict tertiary structure. Thus, molecular biologists pay much attention to what part is indispensable to protein function from knowledge about amino acids and secondary structure. They always determine focus of attention, and plan some experiments to confirm their hypothesis. However, selecting focus of attention is also very difficult even for molecular biologists. Thus, it is expected that computational methods help to select the focus automatically.

2.2 Lysozyme and $\alpha$-Lactalbumin

Lysozyme IIc is a enzyme which dissolves necrotic tissue in a body space of living things. 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, and monkeys, the sequences are almost preserved. That is, this lysozyme IIc evolves very slowly in terms of molecular evolution (Lewin, 1994). This suggests that almost all the sequences are very important to maintain its function.

On the other hand, $\alpha$-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 three reasons (McKenzie and White 1991). First, $\alpha$-lactalbumin are thought to be originated from lysozyme IIc, since both of the sequences are very similar. According to the results of homologous search, about 60% of the sequences of $\alpha$-lactalbumin matches with those of lysozyme, which Table 1: Primary (Amino-acid) Sequences

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$-Lactalbumin</td>
<td>TMFHTSGYDTQAIVENG09N3ESTYGLFQISVRQVQG0000000</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>KVFRCELARTLRLGMDGYRGIHNLANWMCVRQVQG0000000</td>
</tr>
</tbody>
</table>

"@" denotes "gap" regions after processing multiple alignment procedure.
suggests that they are of 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 this 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 replace a few amino acids of \(\alpha\)-lactalbumin, which are located at the place corresponding to the active site of lysozyme, with 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. Finally, third, the active site of \(\alpha\)-lactalbumin has not been found, and it is unknown what parts of the sequences of amino acids are important for its function.

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

3. Problems of Empirical Learning Methods

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

For our example, AQ15 (Michalski et al. 1986) and PRIMEROSE (Tsumoto 1993) 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 classificatory power. Furthermore, these rules consist of only one attribute-value pair and only show what amino acid can be used for classification. Thus, from those "fragmental" rules, we have to extract more structural knowledge. However, these two methods are useful in the sense that they can induce all the possible rules from DNA or amino-acid sequences. On the contrary, in this situation, simple application of induction of decision trees (Breiman et al. 1984; Quinlan 1986) causes another problem. 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

\[ 44 = N \cdot \cdot \cdot \text{lysozyme} \cdot \cdot \cdot (45 \text{cases}) \]

\[ 44 = V \cdot \cdot \cdot \alpha - \text{lactalbumin} \cdot \cdot \cdot (23 \text{cases}) \]

In this case, we get a simple tree, which consists of one node and two leaves. These results are 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 explosions without applying domain knowledge. However, if we use domain knowledge strictly, then much interesting information which could be the sources of discoveries will be eliminated, and only some evidence knowledge will be acquired. Therefore 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 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).

Here we assume that the cognitive strategy of molecular biologists is mainly modeled by the following processes: 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. Discovery Strategy

In this section, first, we overview the strategy of our system, then we describe an algorithm of PRIMEROSE-EX, which applies to exhaustive induction of rules. Finally, we discuss our proposed methods to change representation.

4.1 Overview

In order to implement discovery strategy of molecular biologists, we develop a system, called MOLA-MOLA (Molecular biological data-analyzer and Molecular biological knowledge acquisition tool), which extracts knowledge from amino-acid sequences by controlling application of domain knowledge automatically.

MOLA-MOLA consists of the following six procedures. First, it applies PRIMEROSE-EX, discussed in the next subsection, and exhaustively induces all the rules which can be used for classification of proteins from databases of amino-acid sequences.
Second, MOLA-MOLA 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, MOLA-MOLA generates new databases focused on a certain chemical property from original databases.

Then, third, PRIMEROSE-EX is applied again, all the rules are induced from each database generated by the second procedure. Furthermore, the statistics of each chemical characteristic are calculated.

Next, fourth, the program estimates the secondary structure of amino-acid sequences using Chou-Fasman method (Chou and Fasman 1974) (Secondary Structure Rearrangement).

Fifth, MOLA-MOLA induces all the rules from the databases of secondary structure, applying PRIMEROSE-EX.

Finally, sixth, the system interprets the rules induced by the fifth procedure.

4.2 PRIMEROSE-EX

In order to induce rule exhaustively, we introduce a program, called PRIMEROSE-EX (Probabilistic Rule Induction Method based on Rough Sets for Exhaustive induction). This method is based on rough set theory, which gives a mathematical approach to the reduction of decision tables, corresponding to the exhaustive search for possible rules. For the limitation of the space, we only discuss the definition of probabilistic rules of PRIMEROSE-EX and an induction algorithm of this system. Readers, who would like to know other interesting characteristics of rough sets, could refer to (Pawlak 1991; Ziarko 1993).

4.2.1 Probabilistic Rules. Our definition of probabilistic rules is shown as follows:

Definition 1 (Probabilistic Rules) Let $R_i$ be an equivalence relation and $D$ denote a set whose elements belong to a class $d$, or positive examples. A probabilistic rule of $D$ is defined as a tuple, $< R_i \beta d, SI(R_i, D), CI(R_i, D) >$, where $R_i \beta d$ satisfies the following proposition:

$$R_i \beta d \ s.t. \ [x]_{R_i} \cap D \neq \emptyset,$$

where $\beta = 1 - SI(R_i, D)$, and where $SI$ and $CI$ are defined as:

$$SI(R_i, D) = \frac{\text{card} \ ([x]_{R_i} \cap D)}{\text{card} \ [x]_{R_i}}, \quad \text{and}$$

$$CI(R_i, D) = \frac{\text{card} \ ([x]_{R_i} \cap D)}{\text{card} \ D}.$$ 

SI corresponds to the accuracy measure defined by Pawlak (Pawlak 1991). 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 this rule. For example, when CI is equal to 0.5, half of the members of a class belongs to the set whose members satisfy that equivalence relation.

According to the values of SI and CI, we classify the induced probabilistic rules into the following three categories:

1. Definite Rules: $SI(R_i, D) = 1.0$ and $CI(R_i, D) = 1.0$,
2. Strong Rules: $SI(R_i, D) > 0.5$ and $CI(R_i, D) > 0.75$,
3. Weak Rules: $SI(R_i, D) \geq 0.5$.

4.2.2 An algorithm for PRIMEROSE-EX. Let $D$ denote training samples of the target class $d$, or positive examples. This search procedure is a kind of the greedy algorithm, described as follows.

1. Let $L_0$ be equal to a set of all the attribute-value pairs $[a_i = v_j]$ (selectors in terms of AQ method) and $i$ be equal to 0.
2. Repeat the following three procedures for all the members in a list $L_i$ until $L_i$ is empty. If $L_i$ is empty, goto (6).
3. Select one pair $R = [a_i = v_j]$ and check whether $[x]_{R} \cap D \neq \emptyset (SI(R, D) > 0)$. If so, then goto (4).
4. Otherwise, remove the pair from $L_i$, and repeat this procedure again.
5. Check whether $SI(R, D) > 0.5$. If so, then goto (5).
6. Otherwise, include the pair in a list of weak rules of $d$, and add this pair to $M_i$ and goto (2).
7. Check whether $CI(R, D) > 0.75$. If so, register this pair as a strong rule of $d$. Remove the pair from $L_i$ and goto (2). Otherwise, include the pair in a list of weak rules of $d$, and add this pair to $M_i$ and goto (2).
8. If $M_i$ is empty, quit. Otherwise, generate a list of the whole combination of the conjunction formulae in $M_i$ as $L_{i+1}$. Then increment $i(i := i + 1)$, goto (2).

The above procedure is repeated for all the attribute-value pairs.

4.3 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.
4.3.1 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, electronic charge, inclusion of benzene nucleus, and inclusion of sulfide.

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: \([\text{hydrophobicity} = \text{yes}]\) or \([\text{hydrophobicity} = \text{no}]\). 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: \([33 = \text{hydrophobicity} = \text{yes}]\). This procedure is repeated for all the amino-acids in an original sequence.

4.3.2 Secondary Structure Rearrangement. Next, MOLA-MOLA estimates secondary structure from amino-acid sequences using the Chou-Fasman method (Chou-Fasman 1974), which is the most popular estimation method \(^4\). This Chou-Fasman method outputs the place where specific secondary structures, say \(\alpha\)-helix or \(\beta\)-sheet. According to this estimation, MOLA-MOLA change representation of original databases. For example, the 4th to 10th amino acids are estimated to form an \(\alpha\)-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 4th to 10th attributes are substituted for \(\alpha\)-helix, \(\alpha\)-helix, \(\alpha\)-helix, \(\alpha\)-helix, and \(\alpha\)-helix. That is,

\[
\begin{align*}
\text{Primary Structure} & \quad \text{E R C E L A} \\
\text{Secondary Structure} & \quad \alpha \alpha \alpha \alpha \alpha \\
\end{align*}
\]

It is notable that some attributes may have no specific secondary structure. In these cases, the values 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 33th amino acid of a protein is D (aspartic acid). Because aspartic

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

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

Table 2: Results of Primary Structure Rearrangement

<table>
<thead>
<tr>
<th>Protein</th>
<th>Amino Acid and its Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>lysozyme c</td>
<td>N 27 (A,L 31) K 33</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>E 27 T 31 F 33</td>
</tr>
<tr>
<td>lysozyme c</td>
<td>E 35 N 44 (Y,D 53)</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>(I,S,T 35) V 44 E 53</td>
</tr>
<tr>
<td>lysozyme c</td>
<td>(A,G 76) (A,R 107)</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>I 76 D 107</td>
</tr>
<tr>
<td>lysozyme c</td>
<td>(G,D,Q 117) L 129</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>S 117 E 129</td>
</tr>
</tbody>
</table>

Table 3: Results of Secondary Structure Rearrangement

<table>
<thead>
<tr>
<th>Protein</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>lysozyme</td>
<td>hydrophobic hydrophobic loop</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>polar acidic (\alpha)-helix</td>
</tr>
<tr>
<td>lysozyme c</td>
<td>(\alpha)-helix basic</td>
</tr>
<tr>
<td>(\alpha)-lact</td>
<td>hydrophobic hydrophobic</td>
</tr>
</tbody>
</table>

\(^{5}\text{It is notable that this information can be retrieved from the database generated in the process of primary structure rearrangement.}\)

\(^{6}\text{During the long history of living things, some DNAs are inserted or deleted from the genes, which causes the insertion or deletion of amino acids. Thus, each protein may have the different number of amino acids. Multiple alignment procedures rearrange amino-acid sequences and estimate the effect of the insertion or deletion of amino acids.}\)

\(^{7}\text{All the interpretations of the induced results, shown below, are obtained by discussion with domain experts.}\)
The third exon of lysozyme is different from that of a-\-lactalbumin. Hence these statistics suggest that the third and the fourth exon should be contributed to the functional difference between these two proteins. According to these results, they are now planning to validate these results by the experiments based on technique of recombinant DNA. Since it takes about one to three weeks to study the characteristics of one “mutant” protein, we need more than 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.

7. Conclusion

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 difficulties from the experimental environments of molecular biologists. We apply this method to comparative analysis of lysozyme and a-\-lactalbumin, and the results show that we get some interesting results from amino-acid sequences, which have not been reported before.

References


