ICOT Technical Memorandom: TM-1263

TM-1263

A Stochastic Approach to Genetic Information Processing

by

A. Konagaya (NEC)

May, 1993

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5

Institute for New Generation Computer Technology

A Stochastic Approach to Genetic Information Processing

Akihiko Konagaya

C&C Systems Research Labs., NEC Corp.,
4-1-1 Miyazaki, Miyamaeku, Kawasaki, Kanagawa 216, Japan
e-mail konagaya@csl.cl.nec.co.jp

Abstract

This paper stresses the importance of stochastic machine learning theory for analyzing genetic information such as protein sequences. It is commonly recognized that machine learning theory would play an essential role to extract important information from the enormous amounts of raw genetic information generated by biologists. However, it is also true that more flexible and robust learning methodologies are required to deal with divergence occurring on the genetic information.
For this purpose, we adopt stochastic knowledge representations and stochastic learning algorithms and show their effectiveness with a stochastic motif extraction system. The system aims to extract stable common patterns conserved in some protein category. In the system, common patterns (stochastic motifs) are represented by stochastic decision predicates, and a genetic algorithm with Rissanen's minimum description length principle is used to select "good stochastic motifs" from the viewpoint of increasing prediction performance.

1 Introduction

Rapid improvement of molecular biology technology has succeeded to generate enormous genetic information, such as nucleic sequences and protein sequences. To analyze the genetic information both computer technology and molecular biology technology are required. This leads to the appearance of new scientific domain named genetic information processing or bio informatics.

The goal of genetic information processing is to extract valuable information from genetic information. To achieve this, various computer-based systems have been developed; homology search systems to retrieve similar genetic sequences from a sequence database, secondary or tertiary protein structure inference systems for unknown genetic sequences, motif extraction systems to find common patterns conserved in protein categories, molecular orbitary or molecular dynamics to simulate the behavior of proteins, etc. To enhance these systems, much attention has been focused on artificial intelligence technology, especially for machine learning theory as well as database, image processing and numeric processing technologies. However, it should be noted that very few machine learning theory have succeeded so far to extract biologically meaningful information.

Let us consider the reasons by focusing on motif extraction from protein sequences. The purpose of motif extraction is to find common patterns in a protein category. Such patterns are important since they are conserved in the evolution process for some reason; in fact, there is a good correspondence between conserved patterns and protein active sites and/or special protein structures such as Zinc-Finger and Luesin-Zipper[1]. From the viewpoint of artificial intelligence, motif extraction can be considered as a kind of inductive learning process which finds rules from given sample sequences. However, extracting valuable motifs is not trivial because (1) almost all motifs have exceptions, (2) overfitting may occur when searching for the best fitting rules for sample sequences, and (3) combinatorial explosion may occur when searching for all motif candidates.

To overcome these difficulties, we adopt a stochastic knowledge representation and a stochastic learning algorithm. That is, we propose a "stochastic motif" that represents stochastic mapping from protein sequences to protein functions or protein structures. A stochastic motif may contain exceptions but is more stable and reliable for discriminating unknown sequences or predicting protein functions or structures. To represent the stochastic motif, we also proposed a stochastic decision predicate, a collection of Horn clauses with a probability to represent reliability of each clause. One of the difficulties of extracting stochastic motifs from protein sequences is overfitting to the given sample sequences. To avoid this, we adopt Rissanen's Minimum Description Length (MDL) principle. We can easily show that the best fitting stochastic motif is unstable in the sense that it depends on the sample sequences. The MDL principle solves this problem by balancing between the complexity of a motif and its classification errors. It gives a strategy of selecting an optimal stochastic motif on the basis of the sum of the bit lengths required to encode a stochastic motif and its logarithmic likelihood to the sample protein sequences.

To avoid the combinatorial explosion in the motif extraction, we use "genetic algorithms", which are a kind of probabilistic search algorithm based on the natural evolution process. The virtue of genetic algorithms is that they offer an efficient generate-and-test search by means of simple genetic operators that simulate "crossover", "mutation" and "selection". Our experimental results demonstrate that a genetic algorithm extracts stable stochastic motifs if the MDL principle is adopted for the design of the selection operator or fitness function.

The organization of the rest of this paper is as follows. Section 2 gives a background of stochastic motif extraction. Section 3 gives a representation for stochastic motifs, which we call *Stochastic Decision Predicates*. Section 4 gives a strategy for selecting a good stochastic motif using the MDL principle. Section 5 gives an algorithm for finding optimal stochastic motifs. Section 6 gives an overview of our stochastic motif extraction system. Section 7 presents experimental results on extracting stochastic motifs based on our proposed methodology. Finally, in section 8 we discuss current difficulties and future works. This work has been done as a part of the fifth generation computer systems project for the evaluation of the parallel inference machines.

2 Stochastic Motif

Divergence is one of the characteristics of nature. In practice, it seems difficult to find exact rules in biology. One of such example is a discrimination rule between birds and mammals. Birds can be characterized by simple rules such as having a beak or a bill and wings and laying eggs, and mammals can be characterized by having four legs and childbirth. But, a strange animal *platypus* has a bill and four legs, and lays eggs!

The same thing happens in motif extraction. We can easily find simple common patterns conserved in most sequences in some protein category. However, such simple common patterns almost always have exceptions. The exceptions can be eliminated if we introduce more complex patterns. However, this is not safe because the result may be sample dependent and less effective for the prediction of protein functions and protein structures in unknown sequences. To overcome this difficulty, we pursue stable motifs instead of precise motifs, and propose a "stochastic motif" which inherently includes exceptions, are more stable, and more naturally represent protein functions.

Let us show the example of a stochastic motif using cytochrome c, a protein which plays an important role in the respiratory chain. Figure 1 shows some known cytochrome c sequences for various species. Each character in the sequence corresponds to an amino acid. In most cytochrome c sequences, we can find the common pattern "$CXXCH$" where "C", "X", "H'

Species	Sequence of Cytochrome
Human	..FIMKCSQCHTVEK..
Mouse	..FVQKCAQCHTVEK..
Chicken	..FVQKCSQCHTVEK..
Snake	..FSMKCGTCHTVEE..
Prawn	..FVQRCAQCHSAQA..
Yeast	..FKTRCLQCHTVEK..
Hemp	..FKTKCAECHTVGR..
Tetrahymena	..FDSQCSACHAIEG..
Rhodopila	..FHTICILCHTDIK..
Microbium	..VFKQCKICHQVGP..
Pseudomonas	..VFKQCMTCHRADK..

Figure 1: A part of cytochrome c sequences

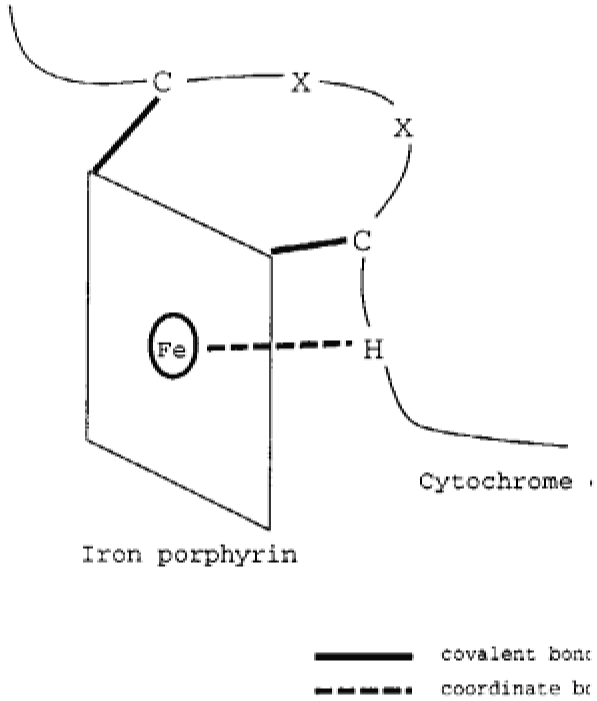

Figure 2: A heme c binding in a cytochrome c

stand for a cysteine, any amino acid, and a histidine, respectively, and the second "X" does not necessary to coincide with the first "X". In fact, the pattern "$CXXCH$" is biologically meaningful because it corresponds to a protein function; the two cysteines and the histidine binds to a heme which cytochrome c holds in the center (Figure 2).

As with other motifs, the pattern "$CXXCH$" also has exceptions. It does not exist in the cytochrome c of Euglena, and the pattern "$CXXCH$" exists in an adrenodoxin of a pig which is a different category from the cytochrome c. A stochastic motif can show the reliability of a pattern by calculating the ratio of the number of target protein sequences containing the pattern and the number of sequences containing the pattern, as follows. "If the pattern ⋯ "$CXXCH$" ⋯ is included in the sequence, then the sequence is cytochrome c with probability 130/227 and otherwise it belongs to other protein categories with probability 8072/8076." Note that the matched sequences for the first clause are eliminated from the total number of sequences to calculate frequency for the second clause.

3 Stochastic Decision Predicates

There are many ways to represent stochastic motifs. As a first step for a stochastic representation of motifs, we devised the stochastic decision predicate, a natural extension of a decision list with probabilities. The stochastic decision predicate consists of linearly ordered Horn clauses with probability parameters as follows.

```
motif(S,cytochrome_c) with 130/227.
      :- contain(S,''CXXCH'').
motif(S,others) with 8072/8076.
```

The general form is the following.

$$\begin{array}{lll}
motif(S,C_1) & (\text{with } p_1) & :- Q_1^{(1)} \wedge \cdots \wedge Q_{k_1}^{(1)}. \\
motif(S,C_2) & (\text{with } p_2) & :- Q_1^{(2)} \wedge \cdots \wedge Q_{k_2}^{(2)}. \\
\quad \cdots\cdots\cdots\cdots & & \\
\quad \cdots\cdots\cdots\cdots & & \\
motif(S,C_{m-1}) & (\text{with } p_{m-1}) : & Q_1^{(m-1)} \wedge \cdots \wedge Q_{k_{m-1}}^{(m-1)}. \\
motif(S,C_m) & (\text{with } p_m) & :- Q_1^{(m)} \wedge \cdots \wedge Q_{k_m}^{(m)}.
\end{array}$$

Here we call each "$motif(S,C_i)$ (with p_i) $:- Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$." a *stochastic clause*. The stochastic clause can be read as S is categorized into C_i with probability p_i if $Q_1^{(i)}, \cdots, Q_{k_i}^{(i)}$ are all **true**. We assume sequential interpretation of the stochastic clauses in this paper. That is, $motif(S,C_i)$ is selected after $motif(S,C_1), \cdots, motif(S,C_{i-1})$ are examined. The body goals $Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$ $(i = 1, \cdots, m)$ represent a condition to discriminate a category C_i when S is given. Each goal $Q_j^{(i)}$ consists of the disjunction of goals $R_{1_j}^{(i)}; \cdots; R_{h_j}^{(i)}$ where $R_{h_j}^{(i)}$ represents some predicate that discriminates a category C_i, such as $contain(S, \sigma)$ which is *true* when S contains a pattern σ.

3.1 Semantics of Stochastic Decision Predicate

The semantics of stochastic decision predicates are given from the viewpoint of computational learning theory of stochastic rules[3]. A stochastic decision predicate represents a probabilistic mapping from protein sequences to categories. The probabilistic mapping can be regarded as a conditional probability distribution over the categories when a sequence is given, by introducing a probability structure on the sequence–category pairs. See the paper [4] for the formal approach to learning stochastic motifs.

4 The MDL Principle in Stochastic Motif Extraction

We adopt the MDL principle to avoid overfitting when extracting stochastic motifs. For example, as we have shown in section 2, the pattern "$CXXCH$" has exceptions in the cytochrome c. It is possible to avoid these exceptions by adding more conjunctions and disjunctions of patterns such as "$AAQCH$" and "$PGTKM$". However, care must be taken so that the obtained result does not become sample dependent, that is, overfit to the sample sequences. Therefore, we adopt the MDL principle to extract simple but stable stochastic motifs which may contain exceptions rather than precise motifs without exceptions.

The MDL principle originally comes from coding theory in communication. The basic idea is to optimize the number of bits when sending an information by finding a rule and its exceptions in the information. The MDL principle selects a rule such that minimizes the total bit length of the rule and the exceptions.

For example, suppose there is a binary string "101101100". Sending the string requires 9 bits if we do not use any rule. Less bits are sufficient if we compress the string as three repeats of "10*" and exceptions "110" for the third bit of each repeat instead of * in the rule. Total bits becomes 7.75 where the rule and the exception require 4.75 and 3 bits, respectively. We may find a more complex rule to reduce the number of exceptions, but such a rule might require a longer bit length to be encoded. Therefore, it is important to balance the complexity of rule and the number of exceptions to reduce the total bit length: this is the MDL principle.

In our methodology, we apply the MDL principle for extracting stochastic motifs as the way proposed by Yamanishi for learning stochastic rules: Yamanishi's MDL learning algorithm[3]. In Yamanishi's algorithm, the MDL principle selects a stochastic rule that balances the complexity of the stochastic rule and its likelihood of matching the sample data. We follow his algorithm with slight modification which mainly comes from the difference of stochastic rule representation: stochastic decision lists and stochastic decision predicates, and some practical reasons for applying the MDL learning algorithm to the motif extraction.

Our methodology selects a stochastic motif that balances the complexity of representation and likelihood of matching the sample sequences. The complexity of a stochastic motif representation is measured by the description lengths to encode the probability parameters and Horn clauses of a stochastic decision predicate. The likelihood of a stochastic motif is measured by the description length of likelihood, that is, by the logarithmic likelihood of categories when the sequences are given to the stochastic motif. The appendix describes the details of calculating the description lengths for a stochastic motif.

5 Genetic Algorithms

To overcome the combinatorial explosion in the motif extraction, we adopt a genetic algorithm, a stochastic search algorithm based on the natural evolution process[6]. Genetic algorithms simulate the survival of the fittest in a population of individuals which represent points in a search space. The individuals are often represented by binary strings. A function, often called a fitness function, gives values to the binary strings. The aim of a genetic algorithm is to find a global optimum of the fitness function when given an initial population of individuals by applying genetic operators in each generation. The genetic operators consist of the following operators: crossover, mutation and selection.

Crossover

The crossover operator produces two descendants by exchanging part of two individuals. This operator aims to make a better individual by replacing a part of an individual with a better part of another individual. For example, crossover of the strings "000110" and "110111" at the

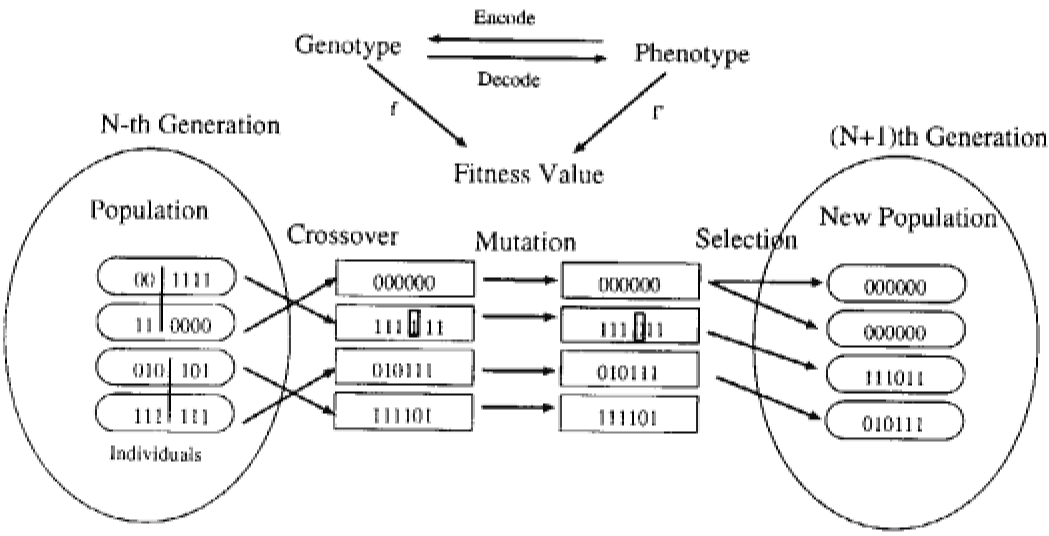

Figure 3: Mechanism of Simple Genetic Algorithms

third position produces the strings "000111" and "110110". The candidates of the crossover operation and the crossover position are randomly chosen.

Mutation

The mutation operator changes certain bit(s) in an individual. For example, the string "000110" becomes "001110" if mutation occurs at the third bit. The operation aims to escape from search spaces from which individuals cannot escape by means of only crossover operators.

Selection

The selection operator chooses good individuals in a population according to their fitness values and the given selection strategy. This operator aims to increase better individuals in the population while maintaining certain diversity. It simulates the survival of the fittest principle. The operator first calculates the relative fitness of all individuals. Then, several lesser individuals are discarded and the same number of better individuals are duplicated according to their relative fitness values. Note that the selection is probabilistic, not deterministic. So, better individuals have a higher chance of remaining or being duplicated but this is not guaranteed.

The performance of genetic algorithms is largely dependent on the design of the fitness function. The most interesting characteristic of our genetic algorithm is in its use of the MDL principle to calculate the fitness value of a stochastic motif. The description length gives the appropriate relative fitness values in the population, although smaller is better in this case.

6 The Stochastic Motif Extraction System

This section gives our overview of the stochastic motif extraction system. The target hypothesis space is the domain of stochastic decision predicates. The search strategy is the MDL principle. The search algorithm is an asynchronous parallel genetic algorithm which consists of the set of subpopulations in which individuals migrate asynchronously. In each subpopulation, individuals represent stochastic motifs in the target hypothesis space, and fitness function calculates the corresponding description lengths of the stochastic motifs represented by the stochastic decision predicates.

The search time depends considerably on the size of the hypothesis space. A large hypothesis space makes it difficult for us to find the optimal stochastic decision predicate in a reasonable time. Therefore, as the first step of motif extraction, we restricted the stochastic predicates to the following forms.

```
motif(S,proteinClass) with p1
      :- contain(S,pattern1) and
         contain(S,pattern2) ...
motif(S,others) with p2.
```

That is, we use a predicate *motif* which discriminates the target protein category *proteinClass* from other proteins (*others*) in the database. The discrimination conditions are represented by the conjunction of a predicate *contain*. As the pattern candidates in the *contain* predicate, we adopt 128 patterns that occur frequently in the target proteins.

The mapping from a stochastic decision predicate to a binary string is the following. Each bit corresponds to one of the 128 patterns. A bit 1 represents the occurrence of the pattern in a discrimination condition, and a bit 0 represents that the pattern does not occur in the discrimination condition. For example, suppose we use 3-bit length binary strings whose first, second, third bits correspond to the pattern "$CXXCH$", "$PXLXG$", "$GXKM$", respectively. Then, the binary string "100" represents the following stochastic decision predicate.

```
motif(S,proteinClass) with p1
      :- contain(S,"CXXCH").
motif(S,others) with p2.
```

The binary sting "011" represents the following stochastic decision predicate.

```
motif(S,proteinClass) with p1
      :- contain(S,"PXLXG") & contain(S,"GXKM").
motif(S,others) with p2.
```

According to this mapping, 128 bits binary strings can express 2^{128} kinds of stochastic decision predicates. As for the genetic operators, we adopt one-point crossover, one-point mutation and roulette wheel selection. Other runtime parameters are the following: the adjustment parameter is 1.0, the number of subpopulations is 63, subpopulation size is 16, the crossover rate is 1.0, the mutation rate is 0.01 and the migration rate is 0.5, that is, one individual per two generations in average.

7 Evaluation

Using the stochastic motif extraction system, we have already extracted 166 stochastic motifs from the protein categories that have more than 10 entries in the Protein Identification Resources (PIR32.0) with currently 9633 entries[1]. Table 1 shows a portion of the results.

In table 1, the line with percent (%) shows the name of protein category, super family number and the number of sequences in the category. The following line shows the common patterns extracted by the system, description lengths and distributions discriminated by the patterns. The column *DL* is the total description length of the extracted stochastic motif. The column *CL*, *PL* and *LL* are the description lengths of Horn clauses, a probability parameter and a logarithmic likelihood to the sample sequences, respectively.

Cytochrome c is a heme-binding protein that carries an electron in respiratory chain. *Cytochrome p450* is a mono-oxygenase containing a proto-heme. *Pepsin* is an acid protease secreted from the stomach. *Trypsin* is a protease secreted from a pancreas. *Globin* is an apo protein that constructs a hemoglobin when binding with a heme molecule. *Immunoglobulin C region*

[1] Annotated and classified entries by homology in pirl.dat.

Table 1: A Portion of Stochastic Motifs obtained from Protein Sequences

Patterns	DL	CL	PL	LL	N_1^+/N_1	N_2^+/N_2
% cytochrome c (1.0, 140) CXXCH	309.544	18.288	10.564	280.693	137/244	9386/9389
% cytochrome P450 (21.0, 33) FXXGXR & GXRXC & RXCXG	127.788	55.523	9.018	63.247	28/28	9600/9605
% pepsin (476.0, 19) FXXXFD & VPXXXC	80.802	38.575	8.700	33.526	17/18	9613/9615
% trypsin (458.0, 40) GWG & CXXDXG	124.490	34.253	9.435	80.802	37/50	9580/9583
% globin (902.0, 456) PXTXXXF & HGXXV	767.740	38.383	10.964	718.392	395/434	9138/9199
% immunoglobulin C region (892.0, 74) VXXFXP & CXVXH	357.216	37.575	9.895	309.746	53/95	9517/9538
% immunoglobulin V region (886.0, 268) DXXXYXC	692.147	20.095	10.871	661.181	237/379	9223/9254

is a constant region of immunoglobulin C. *Immunoglobulin V region* is a variable region of immunoglobulin C.

There are a lot of controversial issues in the biological significance of the obtained results from the view point of genetic information processing. However, the following observations would be more controversial those who are interested in machine learning.

7.1 Comparison of the MDL principle and the Maximum likelihood method

One of our concerns in the stochastic motif extraction is how the MDL principle works in genetic algorithms. To show this, prediction errors are compared to the maximum likelihood (ML) method using the cross validation technique ([7] p.75-76). In the ML method, good individuals are selected using only the description length of likelihood (LL) without consideration for the complexity of a stochastic decision predicate ($CL + PL$).

Using the cross validation technique, the prediction errors can be counted as follows. Firstly, let S_i be a disjoint subgroup of protein sequences S for certain N where $S = \cup_{i=1}^{N} S_i$. Let S_i' be a sample set which removes the i th subgroup from the original protein sequences ($S_i' = S - S_i$). Then, let M_i be a stochastic motif extracted from the sample set S_i', and count the number of prediction errors E_i^+ and E_i^- using the subgroup S_i as a test set, where E_i^+ shows the number of protein sequences that belong to the target protein category but is not *true* for the first clause of the stochastic motif M_i. E_i^- shows the number of protein sequences that do not belong to the target protein category but is *true* for the first clause of the stochatic motif M_i.

Table 2 shows the prediction errors for cytochrome c by cross validation method when divided into 10 subgroups. The results show that the stochastic motifs obtained using a genetic algorithm with the MDL principle are more stable than the ones obtained using a genetic algorithm with the ML method. As seen in table 2, the stochastic motifs obtained by the genetic algorithm with the ML method is sample dependent. It shows strong discrimination performance for the sample protein sequences ($\sum_{i=1}^{10} E_i^-$), but shows weak predictive performance for the test sequences ($\sum_{i=1}^{10} E_i^+$).

Contrary to our expectations, this result comes from the difference of convergence speed between GA with MDL and GA with ML as shown in figure 4. The upper, middle and lower lines represents the average description lengths of the worst, the average and the best individuals so far in each generation. It arises not from the overfitting caused by the ML method since the

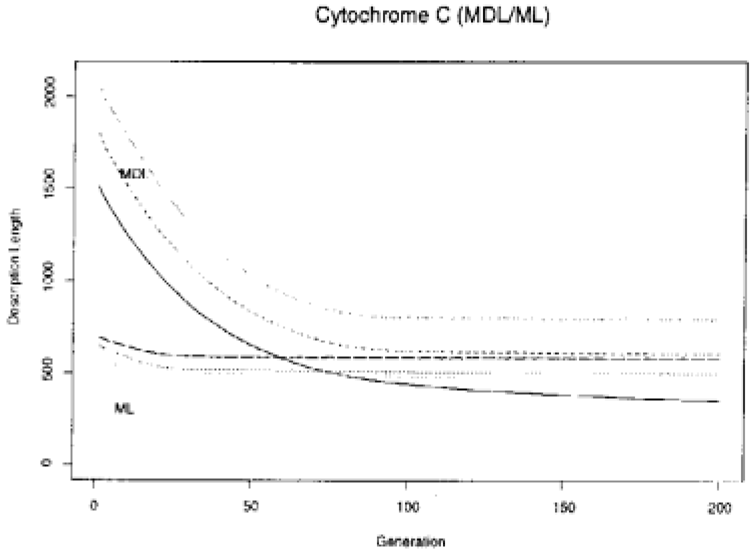

Figure 4: Average description lengths of the best stochastic motif encountered in each generation

Table 2: Prediction errors for Cytochrome C by Cross Validation Method

	MDL	ML
$\sum_{i=1}^{10} E_i^+$	3	57
$\sum_{i=1}^{10} E_i^-$	96	0
Total	99	57

optimal stochastic motif for cytochrome c is *"CXXCH"* in both the MDL principle and the ML method.

The difference of the convergence speed comes from the bias caused by the MDL principle. As shown in figure 5, the number of patterns in the best stochastic motif encountered continuously decrease in case of the MDL principle while it is almost constant in case of the ML method. This is natural since the description length of Horn clauses basically corresponds to the number of patterns. In other words, the MDL principle gives a bias for GA to select individuals with fewer patterns.

One might think it would be possible to reduce the search space if the best stochastic motif can be found in stochastic motifs with fewer patterns. This is true so far as we have examined. The largest stochastic motif has four patterns (Histon H1) and most have two or three patterns. However, it should be noted that we might underestimate the effect of model length (clause length (CL) in this case) and over-simplification may caused by the MDL principle. To show the intrinsic differences between the MDL principle and the ML method, further investigation is required.

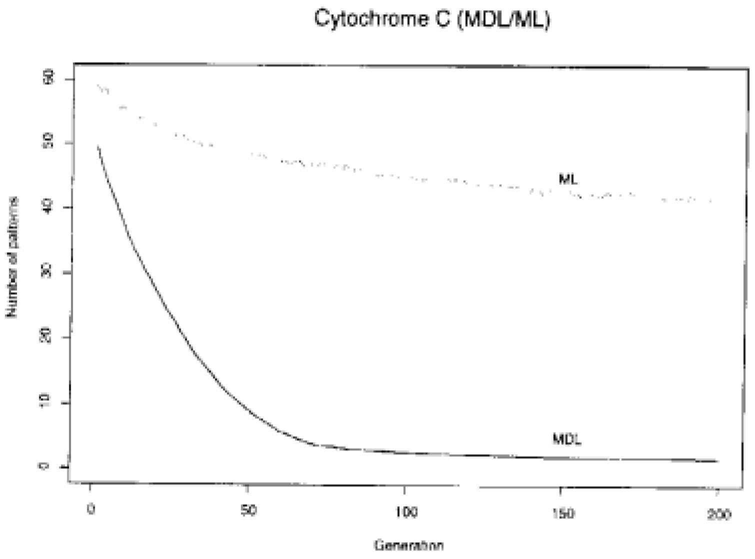

Figure 5: Average number of patterns of the best stochastic motif encountered in each generation

8 Discussion

The following works remain to deal with actual protein sequences on the basis of our methodology.

- The extension of stochastic decision predicate form: In our experience, the number of categories for discrimination is limited two, that is, the target category and the others. A stochastic decision predicate over two categories can be constructed by concatenating the obtained stochastic clauses for each protein category and recalculating the probabilistic parameter, although it causes another combinatorial problem: the order of protein categories. Another interesting extension is providing other predicates, such as a distance between patterns. However, one should be careful that such predicates are really useful for the approximation of protein functions.

- Disjunction of patterns: In the current implementation, no form is provided for the disjunction of patterns on the mapping from stochastic decision predicates to binary strings on the genetic algorithm. For example, the pattern "$CXXCH \vee AAQCH$" may be more appropriate since it eliminates three exceptions caused by Englinae. Finding the pattern "$AAQCH$" is possible if we apply our algorithm to the protein sequences which eliminate the sequences that match "$CXXCH$". However, it should be noted that the pattern "$AAQCH$" is not so reliable since there are only three instances in the protein data base.

- More complex patterns: It is true that the patterns we used in our experiments are too simple to reflect protein functions. For example, it is a well known fact that in the heme-c binding motif "$CXXCH$", no histidine, cysteine, proline nor tryptophan occur in "XX" and that small amino acids tends to occur there. To represent such information, more complex stochastic motifs are required. Our early experience shows that hidden markov models (HMM) seem to be appropriate for this purpose.

- Reducing hypothesis space: Since the MDL principle has a bias against selecting complex patterns, it is possible to eliminate complex patterns, for example, more than five patterns from the hypothesis space. However, we might overbias to the description length of Horn clauses. If this is true, we have to change the adjustment parameter, and also have to search a larger hypothesis space which may include complex patterns, with more than five patterns. In that case, genetic algorithms would be more powerful tools than conventional search algorithms.

9 Conclusion

The importance of stochastic approach for genetic information processing is described using a motif extraction system as an example. Our proposed methodology is characterized by the stochastic representation of motifs using stochastic decision predicates, the MDL principle to avoid overfitting and fast search algorithms using genetic algorithms. Our experimental results show that the methodology actually produces a computationally and biologically meaningful motif for cytochrome c, whose good predictive performance has been statistically proven by the cross validation method. We believe the methodology can also be applied to various kinds of discrimination problems in genetic information processing.

Acknowledgment The author wishes to express his sincere gratitude to Dr. K. Nitta of ICOT and to Dr. N. Koike of NEC Corporation for their encouragement and support. The author thanks Dr. K. Yamanishi and Mr. H. Kondo for their technical advice for the MDL principle and genetic algorithms. The author also thanks Mr. K. Yamagishi, Mr. S. Oyanagi, Mrs. A. Ikeda, Mrs. Y. Kobayashi and Miss K. Hikita for their great contribution to the analysis of real protein sequences.

References

[1] Aitken, Alastair, (1990). *Identification of Protein Consensus Sequences*, Ellis Horwood Series in Biochemistry and Biotechnology.

[2] Rissanen, J.(1978). Modeling by shortest data description. *Automatica, 14*, 465-471.

[3] Yamanishi, K.(1990). A learning criterion for stochastic rules. *Proceedings of the 3-rd Annual Workshop on Computational Learning Theory*, (pp. 67-81), Rochester, NY: Morgan Kaufmann. Its full version is to appear in Jr. on Machine Learning.

[4] Yamanishi, K. & Konagaya, A.(1991). Leaning Stochastic Motifs from Genetic Sequences. *in Proc. of the Eighth International Workshop of Machine Learning.*

[5] Rissanen, J.(1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11*, 416-431.

[6] Goldberg,D.E., (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc.

[7] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J.(1984). Classification and regression trees. *Wadsworth Statistics/Probability Series.*

Appendix: How to Calculate Description Lengths of Stochastic Motifs

The description lengths are calculated as follows. Note that "log" denotes logarithm with base 2 in the following calculation. Let LL be description length of likelihood of categories C^N when the sequences S^N are given to the stochastic motif represented by a probability parameter θ and Horn clauses M. Let E_j be the set of sequences which are **false** for the $1, \cdots, j-1$th clauses and are **true** for the jth clause. Let N_j be the number of sequences in E_j and let N_j^+ be the number of sequences which are in E_j and belong to C_j, the category of the $j-$th clause. Then the likelihood of C^N when given S^N with respect to a stochastic decision predicate with a probability parameter θ and Horn clauses M, which we denote $P(C^N \mid S^N : \theta \prec M)$, is calculated as follows:

$$P(C^N \mid S^N : \theta \prec M) = \prod_{j=1}^{m} p_j^{N_j^+} (1-p_j)^{N_j - N_j^+}.$$

The description length LL is given by $-\log P(C^N \mid S^N : \hat{\theta} \prec M)$, which can be calculated, as follows:

$$LL = \sum_{i=1}^{m} N_i \{ H(\tilde{p}_i) + D_{KL}(\tilde{p}_i \parallel \hat{p}_i) \} \tag{1}$$

where $\tilde{p}_i = N_i^+ / N_i$ and $\hat{p}_i$ is an estimate of the true parameter p_i^*, which is set to be $\frac{N_i^+ + 1}{N_i + 2}$(the Bayes estimator). In addition, $H(\tilde{p}_i)$ and $D_{KL}(\tilde{p}_i \parallel \hat{p}_i)$ are entropy function and Kullback-Leibler divergence defined as follows: $H(\tilde{p}_i) = -\tilde{p}_i \log \tilde{p}_i - (1-\tilde{p}_i)\log(1-\tilde{p}_i)$, $D_{KL}(\tilde{p}_i \parallel \hat{p}_i) = \tilde{p}_i \log \frac{\tilde{p}_i}{\hat{p}_i} + (1-\tilde{p}_i) \log \frac{1-\tilde{p}_i}{1-\hat{p}_i}$

Let PL be the description length of the parameter $\hat{\theta} = (\hat{p}_1, \cdots \hat{p}_m)$ for a fixed Horn clauses M. Since the accuracy (variance) of the maximum likelihood estimator is $O(1/\sqrt{N})$, the description length PL is given by:

$$PL = \sum_{i=1}^{m} \frac{\log N_i}{2} \tag{2}$$

Let CL be the description length of the Horn clauses M.
In the motif extraction system, CL is given by:

$$\begin{aligned} CL \;=\; & \sum_{i=1}^{m} [\, \log^*(\sum_{j=1}^{k_i} h_j) + (\sum_{j=1}^{k_i} h_j - 1) \\ & + \sum_{j=1}^{k_i} \sum_{l=1}^{h_j} \{ \log \begin{pmatrix} L_l^j(i) \\ X_l^j(i) \end{pmatrix} \\ & + (L_l^j(i) - X_l^j(i)) * \log(\mid \mathcal{A} \mid -1) \} + \log r \,] \end{aligned} \tag{3}$$

where $L_l^j(i)$ and $X_l^j(i)$ are the number of amino acids and of variables, respectively, in the pattern in the $l-$th predicate in the $j-$th disjunction region of the $i-$th clause. On the right-hand of (3), the first term denotes the description length of the number of *contain* predicates in the $i-$th clause. For any $d > 0$, $\log^* d$ denotes $\log d + \log\log d + \cdots$ where the sum is taken over all positive terms (Rissanen's integer coding scheme [5]). The second term of (3) denotes the description length of the sequence $\vee, \wedge, \wedge, \cdots$ in the $i-$th clause. The third term denotes the description length of the positions of variables in the pattern σ appearing in the predicate '$contain(S, \sigma)$.' The fourth term denotes the description length required to describe amino acids (not variables) included in the pattern σ appearing in the predicate '$contain(S, \sigma)$'. The last term $\log r$ denotes the description length of the category C appearing in the predicate '$motif(S, C)$'.

By summing (1), (2), and (3), we have the following description length DL:

$$DL \overset{\text{def}}{=} LL + \lambda \{ PL + CL \}$$

where λ is the adjustment parameter.