TR-658

Learning Stochastic Motifs from Genetic Sequences

by
K. Yamanishi & A. Konagawa (NEC)

June, 1991

ICOT

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

(03)3456-3191 ~ 5
Telex ICOT J32964

Institute for New Generation Computer Technology

To appear in Machine Learning: Proceedings of the Eighth International Workshop (1991), Morgan Kaufmann

Learning Stochastic Motifs from Genetic Sequences

Kenji Yamanishi
C&C Information Technology Res. Labs.
NEC Corporation
1-1, Miyazaki, 4-chome Miyamae-ku,
Kawasaki, Kanagawa 213 Japan

Akihiko Konagaya
C&C Systems Res. Labs.
NEC Corporation
1-1, Miyazaki, 4-chome Miyamae-ku,
Kawasaki, Kanagawa 213 Japan

Abstract

This paper presents a methodology for learning stochastic motifs from given genetic sequences. A stochastic motif here is a probabilistic mapping from a genetic sequence (which has been drawn from a finite alphabet) to a number of categories (cytochrome c, globin, trypsin, etc.). We propose a new representation of stochastic motifs, *stochastic decision predicates* (SDPs) and reduce our learning problem to that of learning SDPs. We employ Rissanen's Minimum Description Length (MDL) principle in selecting an optimal hypothesis and present a detailed method for calculating description lengths relative to SDPs. Experimental results show the validity of our learning strategy.

1 INTRODUCTION

We consider here the problem of learning the relationship between genetic sequences which have been drawn from a finite alphabet, and their corresponding categories (cytochrome c, globin, trypsin, etc.). While a given group of genetic sequences may at first follow such a general rule as, for example, "If a sequence contains the pattern ···CAQCH ···, then it corresponds to cytochrome c ··· , " in actual situations, however, not all of such sequences will, in fact, belong to that category, because of the existence of noise or uncertainty due to the variety of biological species. The following type of rule would be more appropriate here to express a mapping from a genetic sequence to categories: "If a sequence contains the pattern ···CAQCH ···, then it corresponds to cytochrome c with probability 4/5 and otherwise with probability 1/5." We may call this kind of mapping a *stochastic motif*, which can be regarded as a conditional probability distribution over categories for a given sequence.

The purpose of this study is to give a methodology for learning stochastic motifs from genetic sequences. First, we propose a new representation of the probabilistic structure of stochastic motifs. This representation we call *stochastic decision predicates* (SDPs) and we reduce our learning problem to that of learning SDPs from given genetic sequences. (In this paper, we define SDPs as specific type of probability distributions. See (Konagaya & Yamanishi 91) for a more general definition and detailed discussion of SDPs.)

Next, we apply the Minimum Description Length (MDL) principle developed in (Rissanen 78, 83), (Wallace & Boulton 68), and (Solomonoff 64), to the selection of the best hypothesis. This principle gives a strategy of selecting an optimal hypothesis on the basis of the trade-off between the simplicity of the model and its fitness to the given examples. The MDL principle has been widely applied to learning problems (e.g. (Quinlan & Rivest 89), (Segen 90), (Yamanishi 90) etc.) and also to the genetic information processing for the purpose of 'unsupervised learning' (e.g. (Milosavljevi'c 90), (Cheeseman & Kanefskey 90), and (Babcock, Olson,& Pednault 90) etc.). In this paper we apply the MDL principle to a new type 'supervised' learning problem in genetic information processing, which cannot be longer reduced to the problem of learning decision trees (Quinlan & Rivest 89) or stochastic decision lists (Yamanishi 90), because in classification of genetic sequences, complicated patterns (including variables) appearing in the sequences must be included in the considerations.

Further, we demonstrate the optimality of the MDL strategy for learning SDPs by testing the performance of our method on real genetic data. Our experimental results show that the MDL strategy actually produces motifs with less predictive errors than the maximum likelihood method.

2 A FORMAL DEFINITION OF THE LEARNING PROBLEM

This section gives a formal definition of a probabilistic structure for genetic sequences and of the learning

problem associated with the structure.

In the discussion below, a genetic 'sequence' refers to a string of letters drawn from a fixed alphabet $\mathcal{A}$. For example, for alphabet $\mathcal{A}$ ={a,b,c,d,e,$\cdots$ z}, "asfaeapa$\cdots$" is a sequence. We assume the following about sequences:

(1) Each sequence is independently drawn.

(2) The length of each sequence is fixed.

(3) Each sequence belongs to a single category (e.g. cytochrome c, globin, trypsin, etc.).

We have made the second assumption here for simplicity's sake. Although in actual observations, the length of observed sequences may be not fixed, we may adjust for this by taking the lengths of the longest observed sequence as our fixed length and adding an appropriate number of gaps to all sequences which have fewer letters in them.

In addition to these assumptions, we assume that the source from which sequences are drawn is "stochastic." We use the following notation to introduce a probabilistic structure on genetic sequences:

[Notation]
- $\mathcal{A}$: alphabet
- $\mathcal{S} \stackrel{\text{def}}{=} \mathcal{A}^n$ (n: the length of the sequence)
- $\mathcal{C} \stackrel{\text{def}}{=} \{C_1, \cdots, C_r\}$: the set of categories
- $(\mathcal{S} \times \mathcal{C}, \mathcal{B}, \mathcal{P})$ ($\mathcal{P}$: a probability measure on $\mathcal{S} \times \mathcal{C}$, $\mathcal{B}$: a sigma algebra on $\mathcal{P}$.)
- S: *sequence* (random variable on $\mathcal{S}$)
- C: *category* (random variable on $\mathcal{C}$)

For nucleic acids, $|\mathcal{A}| = 5$. For amid acids, $|\mathcal{A}| = 21$. Here the gap '–' is also regarded as a letter.

Note that each $\pi \in \mathcal{P}$ is resolved as follows:

$$\pi(S, C) = Q(S)P(C \mid S) \qquad (1)$$

Here $Q(S)$ is a probability distribution over $\mathcal{S}$. $P(C \mid S)$ is a conditional probability distribution over $\mathcal{C}$ given $S \in \mathcal{S}$, which is called a *stochastic rule*([Yam 90]). In the context of genetic information processing, we call a stochastic rule $P(C \mid S)$ a *stochastic motif* and also call the stochastic motif $P^*(C \mid S)$ defined by the source the *target motif*.

In the above probabilistic setting, we consider the following learning problem:

"Given N examples of pairs (sequence, category): $(S_1, C_1) \cdots (S_N, C_N)$ drawn independently according to an unknown distribution $Q(S)P^*(C \mid S)$, find a good approximation of the target motif $P^*(C \mid S)$ from the hypothesis space $\mathcal{H}$."

Here the *hypothesis space* $\mathcal{H}$ denotes a class of stochastic motifs specified by some constraints.

3 STOCHASTIC DECISION PREDICATES

In this section, we propose a new hypothesis space, *stochastic decision predicates* (SDPs), for representation of stochastic motifs. An SDP probabilistically maps from a genetic sequence to a number of different categories and takes the form of a linearly ordered set of conjunctive normal forms of a number of predicates including "patterns."

First we define a *pattern set* Σ as $\Sigma \stackrel{\text{def}}{=} (\mathcal{A} \cup \{X\})^*$, where $(\mathcal{A} \cup \{X\})^*$ is a set of all strings generated by the elements of $\mathcal{A}$ and X. We call $\sigma \in \Sigma$ a *pattern*. For example, if $\mathcal{A} = \{A, G, C, T\}$, $AAGXCX$ and $XCGXT$ are patterns. Here we regard X as an anonymous variable. That is, for example, in $AAGXCX$, we regard X as a variable but we don't necessarily identify the first X with the second X. In order to define SDPs, we use the following notation. For the sake of simplicity, we assume that the number of categories is two.

[Notation]
- $S(\in \mathcal{S})$: sequence
- m : positive integer
- $C_i(\in \mathcal{C} = \{C^{(1)}, C^{(2)}\})$: category
- $motif(S, C_i)$: predicate that is true if and only if S belongs to C_i
- $Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$ ($i = 1, \cdots, m$): conjunction of $Q_1^{(i)}, \cdots, Q_{k_i}^{(i)}$
- $Q_j^{(i)} = R_1^{(i)} \vee \cdots \vee R_{h_j}^{(i)}$ ($i = 1, \cdots, m,\ j = 1, \cdots, k_i$): disjunction of $R_1^{(i)}, \cdots, R_{h_j}^{(i)}$
- $R_l^{(i)}$ ($i = 1, \cdots, m,\ l = 1, \cdots, h_j,\ j = 1, \cdots, k_i$) : predicate of the form: $contain(S, \sigma)$
- $contain(S, \sigma)$: predicate that is true if and only if S contains σ.
- $\sigma(\in \Sigma)$: pattern
- $p_i \in [0, 1]$ ($i = 1, \cdots, m$): probability parameter

We define an SDP as a stochastic rule which takes the following general form:

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

The semantics of the SDP is as follows: For given $S \in \mathcal{S}$, C_i is assigned with probability p_i and $\mathcal{C} - \{C_i\}$ is assigned with probability $1 - p_i$ where i is the least index such that S makes $Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$ true (the m–th condition is regarded as true for any S.).

Let $\theta \stackrel{\text{def}}{=} (p_1, \cdots, p_m)$ be a probability parameter vector and $M \stackrel{\text{def}}{=} \{motif(S, C_i) :\!- Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}\}_{i=1,\cdots,m}$ be a countable model. An SDP specified by θ and M defines a stochastic rule, which we denote $P(C \mid S : \theta \prec M)$. Letting the set of all possible $\{M\}$ be $\mathcal{M}$, the set of all stochastic decision predicates, which we denote $\mathcal{H}_{SDP}$, can be written as follows:

$$\mathcal{H}_{SDP} \stackrel{\text{def}}{=} \{P(C \mid S : \theta \prec M) : \theta \in [0,1]^m \text{ and } M \in \mathcal{M}\}$$

where m is the number of clauses in M.

The problem of learning stochastic motifs can be reduced to *estimation* of both of θ^* and M^* specifying the target motif, from given genetic sequences.

4 LEARNING STOCHASTIC MOTIFS USING THE MDL PRINCIPLE

The MDL principle asserts that the best hypothesis is the one that minimizes the total description length (in bits) of the hypothesis plus the description of the examples relative to the hypothesis. In this section, we describe a methodology for applying the MDL principle to learning stochastic SDPs. All logarithms, hereafter, are to the base 2.

For observed examples of $D^N = D_1 \cdots D_N$, $D_i = (S_i, C_i) \in S \times C$ $(i = 1, \cdots, N)$, let $S^N \stackrel{\text{def}}{=} S_1 \cdots S_N$ and $C^N \stackrel{\text{def}}{=} C_1 \cdots C_N$. Let E_j be the set of examples which make the $1, \cdots, (j-1)$ $-$th clauses false and make the $j-$th clause true. Let N_j be the number of elements in E_j and let N_j^+ be the number of examples which are in E_j and belong to C_j $(j = 1, \cdots, m,$ m is the number of clauses in M.)

Then the likelihood of C^N for given S^N with respect to $P \in \mathcal{H}_{SDP}$ with θ and M, which we denote $P(C^N \mid S^N : \theta \prec M)$, is calculated as follows:

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

The description length of S^N for given C^N with respect to the maximum likelihood estimate $\hat{\theta}$ and the countable model M, which we denote $\ell(C^N \mid S^N : \hat{\theta} \prec M)$, is calculated by $-\log P(C^N \mid S^N : \hat{\theta} \prec M)$ as follows:

$$\ell(C^N \mid S^N : \hat{\theta} \prec M) = \sum_{i=1}^{m} N_i \{H(\bar{p}_i) + D(\bar{p}_i \parallel \hat{p}_i)\} \qquad (2)$$

where $\bar{p}_i = N_i^+ / N_i$ and $\hat{p}_i$ is an estimate of the true parameter p_i^*, which is usually set to be N_i^+ / N_i(the maximum likelihood estimator) or $\frac{N_i^+ + 1}{N_i + 2}$(the Bayes estimator). $H(\bar{p}_i) = -\bar{p}_i \log \bar{p}_i - (1 - \bar{p}_i) \log(1 - \bar{p}_i)$. $D(\bar{p}_i \parallel \hat{p}_i) = \bar{p}_i \log \frac{\bar{p}_i}{\hat{p}_i} + (1 - \bar{p}_i) \log \frac{1 - \bar{p}_i}{1 - \hat{p}_i}$ $(i = 1, \cdots, m)$.

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

$$\ell(\hat{\theta} \mid M) = \sum_{i=1}^{m} \frac{\log N_i}{2} \qquad (3)$$

Let $\ell(M)$ be the description length for the countable model M, satisfying the prefix the Kraft inequality: $\sum_{M \in \mathcal{M}} 2^{-\ell(M)} \leq 1$. We calculate $\ell(M)$ by:

$$\begin{aligned} \ell(M) \;=\; & \sum_{i=1}^{m} [\, \log^*(\sum_{j=1}^{k_i} h_j) + (\sum_{j=1}^{k_i} h_j - 1) \\ & + \sum_{j=1}^{k_i} \sum_{l=1}^{h_j} \{\log \begin{pmatrix} L_l^j(i) \\ X_l^j(i) \end{pmatrix} \\ & + (L_l^j(i) - X_l^j(i)) * \log(|\mathcal{A}| - 1)\} + 1 \,] \end{aligned} \qquad (4)$$

where $L_l^j(i)$ and $X_l^j(i)$ are the number of letters and of variables in the $l-$th predicate in the $j-$th disjunction of the $i-$th clause, respectively. On the righthand of (4), the first term denotes the description length for the number of *contain* predicates in the $i-$th clause. Here, for any $d > 0$, $\log^* d$ denotes $\log c + \log d + \log \log d + \cdots$ $(c = 2.865)$ where the sum is taken over all positive terms (Rissanen's integer coding scheme (Rissanen 83)). The second term denotes the description length for the sequence $\vee, \wedge, \wedge, \cdots$ in the $i-$th clause. The third term denotes the description length for the positions of variables in the pattern σ appearing in '$contain(S, \sigma)$.' The fourth term denotes the description length for letters (not variables) specifying the pattern σ appearing in '$contain(S, \sigma)$.' The last 1 bit denotes the description length for the category $C \in \mathcal{C}$ appearing in the predicate '$motif(S, C)$.' See (Konagaya & Yamanishi 91) for the biological meaning of the encoding scheme for $\ell(M)$.

By summing (2), (3), and (4), we have the following total description length $\ell(C^N : \hat{\theta} \prec M \mid S^N)$ relative to the parameter $\hat{\theta}$ and the model M:

$$\begin{aligned} & \ell(C^N : \hat{\theta} \prec M \mid S^N) \qquad (5) \\ & \stackrel{\text{def}}{=} \ell(C^N \mid S^N : \hat{\theta} \prec M) + \lambda\{\ell(\hat{\theta} \mid M) + \ell(M)\} \end{aligned}$$

where λ is the adjustment parameter. The MDL principle asserts that one should select the model $\tilde{M}$ such that minimizes $\ell(C^N : \hat{\theta} \prec M \mid S^N)$. We call the strategy for finding hypotheses with least (or nearly least) total description length the *MDL strategy*.

The optimality of the MDL strategy has been theoretically proven in terms of their convergence to the true model ((Rissanen 78, 83), (Barron 85)). Further, Yamanishi proved in (Yamanishi 90: a full version) that the upper bound on the sample complexity needed for the MDL estimate to converge to the true model within given accuracy and confidence parameters, is less than that for any other estimate (e.g., maximum likelihood estimate, AIC estimate etc.).

Table 1: Distribution of Mitochondria Cytochrome C

Motif	N_1 and N_2	N_1^+ and N_2^+	$\hat{p}_1$ and $\hat{p}_2$
SDP I	189	67	0.356
	5969	5966	0.9993
SDP II	73	67	0.910
	6085	6082	0.9993
SDP III	71	67	0.932
	6087	6084	0.9993

Table 2: Description Lengths relative to Motifs

Motif	L_1	L_2	L_3	Total
SDP I	214.7	10.1	29.7	255.5
SDP II	67.5	9.4	53.4	131.3
SDP III	59.9	9.4	76.2	146.5

Table 3: Estimates of Prediction Error Rates

	MDL	ML
Estimated Error Rate	0.0008	0.0013

5 EXPERIMENTAL RESULTS

In this section, we apply our methodology to learning stochastic motifs for discriminating "mitochondria cytochrome c." Here the mitochondria cytochrome c is a subclass of cytochrome c protein which carries an electron in the respiratory chain. Hereafter, "*mcyt.c*" denotes the mitochondria cytochrome c. Let the domain $\mathcal{S} = \mathcal{A}^{1000}$ and the range be $\mathcal{C} = \{mcyt.c, others\}$, where the alphabet $\mathcal{A}$ is the twenty-letter alphabet of amino acids plus the gap '–' (i.e., $\mid \mathcal{A} \mid = 21$) and the length of each sequence is adjusted to 1000.

Let the following three simple SDPs be given as candidates for the representation of the optimal motif.

SDP I
$motif(S, mcyt.\ c)$ (with p_1) : –
$contain(S,$ "CXXCH").
$motif(S, others)$ (with p_2).

SDP II
$motif(S, mcyt.\ c)$ (with p_1) : –
$contain(S,$ "CXXCH") $\wedge$ $contain(S,$ "PGTKM").
$motif(S, others)$ (with p_2)

SDP III
$motif(S, mcyt.\ c)$ (with p_1) : –
$contain(S,$ "CXXCH") $\wedge$ $contain(S,$ "GPXLXG")
$\wedge$ $contain(S,$ "PGTKM").
$motif(S, others)$ (with p_2).

These three SDPs are obtained from the training examples by using some heuristics (e.g. DP matching etc.). In this study, ignoring the computational complexity of finding the three SDPs, we concentrate on the problem of how to select the SDP with the "least" predictive error rate, from among them.

Table 1 shows the distribution of the training sequences observed through the these SDPs and estimates of probability parameters. In Table 1, for each SDP in the first column, the second column shows N_1(: the number of the examples which make the 1st clause true) in the upper row and $N_2(=N-N_1)$ in the lower row. The third column shows N_1^+(: the number of examples which make the 1st clause true and belong to *mcyt.c*) in the upper row and N_2^+(: the number of examples which make the 1st clause false and don't belong to *mcyt.c*) in the lower row. The fourth column shows the estimated parameters: $\hat{p}_1$ in the upper row and $\hat{p}_2$ in the lower row. $\hat{p}_1$ and $\hat{p}_2$ denote the Bayes estimates of the probability parameters in the first and second clauses, respectively.

Table 2 shows the description lengths relative to the three SDPs. In Table 2, L_1, L_2, L_3, and Total denote the description lengths calculated by (2),(3),(4), and (5), respectively. Here λ in (5) is set to be 1.

We can see from Table 2 that the SDP II containing "CXXCH" and "PGTKM" requires the least total description length and thus is optimal in the sense of the MDL principle. The SDP I containing only a single pattern "CXXCH" is considered as too simple for discrimination of mitochondria cytochrome c. On the other hand, the SDP III containing "CXXCH," "GPXLXG," and "PGTKM" can discriminate the given sequences best but is considered as too complicated to give a good model of the target motif.

SDP II, which is selected as the best hypothesis by the MDL strategy, is biologically meaningful. Indeed, in the case of mitochondria cytochrome c, cysteines (C) in the pattern "CXXCH" denote binding sites for a heme c, and both of histidine (H) in the pattern "CXXCH" and methionine (M) in the pattern "PGTKM" denote regands to the iron molecule in the heme c. These amino acids are called *functional sites*, which play the most essential role in protein activity and have been preserved in the evolution process. See (Konagaya & Yamanishi 91) for more detailed discussion on the biological meaning of the motifs extracted by our methodology.

Notice that an SDP can be transformed into a deterministic rule by letting all probability parameters $\{p_i\}$ be 1. Such a deterministic rule can predict whether or not the given sequence belongs to mitochondria cytochrome c. Table 3 shows the estimates of the prediction error rates for the deterministic transformations of the SDPs selected by the MDL principle and by the *maximum likelihood (ML)* method. Here the ML

method refers to the strategy which selects a rule that minimizes only the description length for examples, ignoring the description length for the hypothesis itself.

The estimates of the error rates shown in Table 3 are obtained by applying the cross-validation method to the original training sequences: We randomly divide the original set S of training sequences into 10 subsets $S_1, \cdots, S_{10}$ ($\mid S_1 \mid = \cdots = \mid S_{10} \mid$). Using each $S - S_i$ as a training example, we construct an SDP by the MDL or ML method, then transform it into a deterministic rule by letting all probability parameters $\{p_i\}(i = 1, \cdots, m)$ be 1, then test it on S_i. We denote the number of prediction errors for the MDL motif on S_i as $Error_{MDL}(S_i)$ and that of the ML motif as $Error_{ML}(S_i)$. Let R_{MDL} and R_{ML} be the cross-validation estimates (Breiman, Friedman, Olshen, & Stone 84 p.75-76) of error rates for the MDL principle and ML method, respectively. R_{MDL} and R_{ML} are given by:

$$R_{MDL} = \frac{1}{N} \sum_{i=1}^{10} Error_{MDL}(S_i)$$

$$R_{ML} = \frac{1}{N} \sum_{i=1}^{10} Error_{ML}(S_i)$$

where $N = 6158$. The numerical result shown in Table 3 demonstrates that the MDL principle produces a rule with less prediction errors than selected by the ML method, in average.

6 CONCLUSION

We have proposed a new methodology for learning stochastic motifs from genetic sequences. Our proposed methodology is characterized by the new representation of stochastic motifs, SDP, and by the MDL learning strategy. The experimental results show that the MDL strategy actually produces a motif for discriminating mitochondria cytochrome c with less predictive error than the maximum likelihood method. In this paper, we have concentrated on the issue of how to select the best motif. The issue of how to design an algorithm of finding approximately minimum description length motifs is left for future study.

Acknowledgements

The part of this work has been done in the Fifth Generation Computer Systems Project in Japan. The authors wish to appreciate Dr.K.Nitta of Institute for New Generation Computer Technology for giving opportunities to this work. The authors also wish to express their sincere gratitude to Mr.K.Nakamura, Mr.S.Koike, Mr.A.Kaneko, and Dr.M.Yokota of NEC corporation for their encouragement and support. The authors also thank Mr.K.Yamagishi, Mr.S.Oyanagi of NEC Scientific Information System Development, and Miss K.Hikita of NEC Corporation for their great contribution to genetic data analysis.

References

L.Allison, C.S.Wallace, & C.N.Yee. (1990) Inductive inference over macro-molecules. *Preprint of Symposium Papers: The Theory and Application of Minimal-Length Encoding, AAAI 1990 Spring Symposium Series*, Stanford, CA.

M.S.Babcock, W.K.Olson, & P.D.Pednault. (1990) The use of minimum description length principle to segment DNA into structual and functional domain. *Preprint of Symposium Papers: The Theory and Application of Minimal-Length Encoding, AAAI 1990 Spring Symposium Series*, Stanford, CA.

A.R.Barron. (1985) *Logically Smooth Density Estimation*, PhD dissertation, Dept. of Electrical Eng., of Stanford Univ..

L.Breiman, J.H.Friedman, R.A.Olshen & C.J.Stone. (1984) *Classification and Regression Trees*, Wadsworth Statistics/Probability Series.

P.Cheeseman & B.Kanefsky. (1990) Evolutionary Tree Reconstruction. *Preprint of Symposium Papers: The Theory and Application of Minimal-Length Encoding, AAAI 1990 Spring Symposium Series*, Stanford, CA.

A.Konagaya & K.Yamanishi. (1991) Stochastic decision predicates: a new scheme to represent motifs, to appear in *Proceedings of AAAI Workshop on AI and Molecular Biology*.

A.D.Milosavljević. (1990) Categorization of macromolecular sequences by minimal length encoding. *Technical Report UCSC CRL-90-41*, Univ. of California at Santa Cruz.

J.R.Quinlan & R.L.Rivest. (1989) Inferring decision trees using the minimum description length principle *Information and Computation, 80(3)*, 227-248.

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

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

J.Segen. (1990) Graph clustering and modeling by data compression. *Proceedings of the Seventh International Conference on Machine Learning*, (pp. 93-101), Morgan Kaufmann.

R.J.Solomonoff. (1964) A formal theory of inductive inference. Part 1. *Information and Control, 7*, 1-22.

C.S.Wallace & D.M.Boulton. (1968) An information measure for classification. *Computer Journal*, 185-194.

K.Yamanishi. (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. The full version is to appear in *Machine Learning* (Journal).