ICOT Technical Report: TR-657

TR-657

Stochastic Decision Predicates: A Scheme to Represent Motifs

by
A. Konagaya & K. Yamanishi (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

Stochastic Decision Predicates: A Scheme to Represent Motifs

Akihiko Konagaya[†]
Kenji Yamanishi[‡]
C&C Systems Research Labs., NEC Corporation[†],
C&C Information Technology Research Labs., NEC Corporation[‡]

1-1, Miyazaki 4-chome, Miyamaeku, Kawasaki, Kanagawa 216, Japan
TEL:(044)856-2126, E-mail:konagaya@csl.cl.nec.co.jp yamanisi@ibl.cl.nec.co.jp

Abstract

This paper presents a new scheme for classifying genetic sequences, called *Stochastic Decision Predicates*. A stochastic decision predicate consists of Horn clauses and their probability parameters, and represents a (stochastic) motif that denotes a probabilistic mapping from a genetic sequence to a set of categories, such as protein families. For the selection of stochastic decision predicates, quantative evaluation is possible from the viewpoint of predictive performance for unknown sequences as well as discrimination performance for the given genetic sequences. We employ Rissanen's Minimum Description Length (MDL) principle in order to avoid overlearning caused by the statistical fluctuation. Our experimetal results demonstrate that the MDL principle produces motifs with less predictive errors than the maximum likelihood method.

1 Introduction

Recently, several biologists have focused on searching for common patterns in genetic sequences which have been preserved in the evolution process. Such patterns are called "motifs" in the biology community, and are considered to represent special biological functions (e.g. Serine proteinases and Cysteine proteinases) and/or special structures (e.g. Zinc fingers and Leucine zipper consensus)[AA 90].

In this paper, we are focusing on motifs from the viewpoint of computational analysis. From our viewpoint, a motif is considered as a mapping from genetic sequences to certain categories such as globin, cytochrome c, trypsin, ferredoxin, etc. The mapping greatly reduce the search time in genetic databases by using the motifs for the discrimination of genetic sequences instead of dynamic programming matching which has become a time-consuming operation due to the rapid increase of the size of the databases.

As a representation of the motif, one might use an inference rule that discriminates the given genetic sequence, as follows: "If the pattern $\cdots$CAQCH $\cdots$ is included in the sequence, then it corresponds to cytochrome c $\cdots$." However, in actual cases, one would soon notice that it is very hard to find such a deterministic inference rule, because of the existence of noise, or uncertainty, due to the variety of biological species.

To overcome this difficulty, the following type of rule is more appropriate to express the mapping from a genetic sequence to some categories: "If the pattern $\cdots$CAQCH $\cdots$ is included in the sequence, then it corresponds to cytochrome c with probability 4/5 and otherwise with probability 1/5." We call this kind of mapping a *stochastic motif*. Here, it should be noticed that we are interested in extracting a stochastic motif that performs well for unknown sequences rather than a motif with high discrimination performance, that is, with a discrimination probability closer to 1.0 for the given sequences.

In this paper, we will propose a new scheme to represent such stochastic motifs and methodology to select 'good' stochastic motifs, but mention little about the probability structures of the stochastic motifs. See the paper[YK 91] for the formal approach to learning stochastic motifs. First, we propose a new representation of stochastic motifs, which we call *Stochastic Decision Predicates*. Then, we apply the Minimum Description Length(MDL) principle developed in ([Ris 78, 89], ([WB 68]), and ([Sol 64]), to the selection of an optimal hypothesis. This principle gives a strategy of selecting an optimal rule 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 the genetic information processing in [Mil 90], [CK 90], and [BOP 90] for the purpose of 'categorization.' In this paper we will apply the MDL principle to a new 'supervised' learning problem in genetic information processing. Further, we will demonstrate the validity of the MDL priciple in our learning methodology by showing experimental results applied to real genetic data. Our experimental results show that the MDL principle produces motifs with less predictive errors than the maximum likelihood method.

The organization of the rest of this paper is as follows: Section 3 gives a representation for stochastic motifs, which we call *Stochastic Decision Predicates*. Section 4 gives a strategy for finding optimal stochastic decision predicates using the MDL principle. Section 5 presents experimental results on learning genetic motifs based on our proposed strategy. Section 6 gives a discussion on our methodology.

2 Stochastic Decision Predicates

In this section, we propose a new scheme, which we call *Stochastic Decision Predicates*, for representing stochastic motifs. The stochastic decision predicate consists of Horn clauses with probability parameters. The general form is the following.

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

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 clause can be read as S is categorized into C_i with probability p_i if $Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$ are all **true**. We assume sequential interpretation of the 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 given S. Each goal $Q_j^{(i)}$ consists of the disjunction of goals $R_1^{(i)};\cdots;R_{h_j}^{(i)}$ where $R_{h_j}^{(i)}$ represents some predicate that discriminates a category C_i. $R_{h_j}^{(i)}$ is of the form $contain(S,\sigma)$ which is *true* when S contains a pattern σ. In this paper, we deal with a simple pattern that may contain anonymous variables denoted by 'X'. For example, if $\mathbf{A} = \{A,G,C,T\}$, "AAGXCX" and "XCGXT" are patterns. Note that the first X does not necessarily identify with the second X. In addition, our scheme can deal with any patterns and predicates if we can specify the complexity of the patterns and predicates as seen in the next section.

The notation for the stochastic decision predicate is summarized as follows.

[Notation]

- $\mathbf{A}$: alphabet, a set of letters appearing in genetic sequences
  (For example, for nucleic acids, $\mid \mathbf{A_{na}} \mid = 4$, for amino acids, $\mid \mathbf{A_{aa}} \mid = 20, respectively$)
- $\mathbf{S} \stackrel{\text{def}}{=} \{S_1,\cdots,S_n\}$: the set of sequences
- $\mathbf{C} \stackrel{\text{def}}{=} \{C_1,\cdots,C_r\}$: the set of categories
- $S(\in \mathbf{S})$: sequence
- m : positive integer
- $C_i(\in \mathbf{C})$ $(i=1,\cdots,m)$: category in the ith stochastic clause
- $motif(S,C_i)$: 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 $Q_1^{(i)},\cdots,Q_{k_i}^{(i)}$
- $R_l^{(i)}$ $(i=1,\cdots,m,\ l=1,\cdots,h_j,\ j=1,\cdots,k_i)$: predicate of the form : $contain(\sigma,S)$
- $contain(S,\sigma)$: predicate that is **true** if and only if S contains the pattern σ.
- Σ: the set of patterns
- $\sigma(\in \Sigma)$: pattern
- $p_i \in [0,1]$ $(i=1,\cdots,m)$: probability that $motif(S,C_i)$ is **true** for S such that $Q_1^{(i)} \wedge \cdots \wedge Q_{k_i}^{(i)}$ is **true**
- $\theta \stackrel{\text{def}}{=} \{p_1,\cdots,p_m\}$: probability parameter
- M: the set of linearly ordered Horn clauses
- $M(\in \mathcal{M})$: linearly ordered Horn clauses .

2.1 Semantics of Stochastic Decision Predicate

The semantics of stochastic decision predicates is given from the viewpoint of computational learning theory of stochastic rules[Yam 90]. A stochas-

tic decision predicate defines a probablistic mapping from genetic sequences to categories. The probabilistic mapping can be regarded as a conditional probability distribution over the categories when given an sequence, by introducing a probability structure on the sequence–category pairs.

In this interpretation, the problem of extracting motifs in genetic sequences can be regarded as the problem of learning stochastic motifs, which denote conditaional probability distributions, in a probability structure on the sequence–category pairs. In addition, the learning problem can be also reduced to a learning problem in some *hypothesis space* if the hypothesis space contains a good approximation of the target stochastic motif. We consider the set of conditional probability distribution defined by stochastic decision predicates would be one of such hypothesis spaces for learning stochastic motifs. See the paper [YK 91] for the formal approach to learning stochastic motifs.

3 Selection of Stochastic Decision Predicates Using the MDL Principle

In this section, we describe a methodology for applying the MDL principle to learning stochastic decision predicates. On the basis of the MDL principle, the best stochastic decision predicate that one should select in the given examples is the one that minimizes the total description length, that is, the description length of the stochastic decision predicate and the description length of the examples relative to the predicate.

The description length of the examples is given by the logarithmic likelihood of categories when the sequences in the examples are given. Given N examples $(S^N, C^N) \stackrel{\text{def}}{=} (S_1, C_1) \cdots (S_N, C_N)$. Let E_j be the set of examples 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 examples which are in E_j and belong to C_j. Then the likelihood of C^N when given S^N with respect to a stochastic decision predicate with probability parameter θ and linear ordered 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 of the examples (S^N, C^N) relative to the stochastic decision predicate P with an estimated parameter $\hat{\theta}$ and M, which we denote $\ell(C^N \mid S^N : \hat{\theta} \prec M)$, is given by $-\log P(C^N \mid S^N : \hat{\theta} \prec M)$, which can be calculated, as follows:

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

where $\tilde{p}_i = N_i^+/N_i$ 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). In addition, $H(\tilde{p}_i)$ and $D(\tilde{p}_i \parallel \hat{p}_i)$ are entropy function and Kullback-Leibler divergence defined as follows.

$$H(\tilde{p}_i) = -\tilde{p}_i \log \tilde{p}_i - (1-\tilde{p}_i)\log(1-\tilde{p}_i)$$

$$D(\tilde{p}_i \parallel \hat{p}_i) = \tilde{p}_i \log \frac{\tilde{p}_i}{\hat{p}_i} + (1-\tilde{p}_i)\log \frac{1-\tilde{p}_i}{1-\hat{p}_i}$$

The description length $\ell(C^N \mid S^N : \hat{\theta} \prec M)$ indicates the number of bits required to encode the distribution of positive examples and negative examples relative to the stochastic decision predicate. The length varies from 0 bits, when $p_i = 0$ *or* 1.0 $(i = 1, \cdots, m)$, to N bits, when $p_i = 0.5 (i = 1, \cdots, m)$. The former occurs when the stochastic decision predicate completely discriminate the target categories in the given examples. The latter occurs when the stochastic decision predicate does not contribute to any discrimination of the given examples.

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

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

Letting $\ell(M)$ be the description length of the set of stochastic clauses M, $\ell(M)$ is given by:

$$\ell(M) = \sum_{i=1}^{m} [\log^*(\sum_{j=1}^{k_i} h_j) + (\sum_{j=1}^{k_i} h_j - 1)$$

$$+\sum_{j=1}^{k_i}\sum_{l=1}^{h_j}\{\log\left(\begin{array}{c} L_l^j(i) \\ X_l^j(i) \end{array}\right) \quad (3)$$
$$+(L_l^j(i)-X_l^j(i))*\log(|\mathcal{A}|-1)\}+\log r\ ]$$

where $L_l^j(i)$ and $X_l^j(i)$ are the number of letters and of variables, respectively, in the $l-$th predicate in the $j-$th disjunction of the $i-$th clause. Other notations follow those defined in Section 2. On the righthand 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 [Ris 83]). 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 predicates in the pattern σ appearing in the predicate '$contain(S,\sigma)$.' The fourth term denotes the description length required to describe letters themselves (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 '$motif(S,C)$'.

The description length $\ell(M)$ denotes the number of bits required to represent a predicate M by means of some encoding scheme. The scheme ought to be designed so that the description length can reflect the complexity of predicates.

By summing (1), (2), and (3), 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 set of stochastic clauses M:

$$\ell(C^N : \hat{\theta} \prec M \mid S^N) \quad (4)$$
$$\stackrel{\text{def}}{=} \ell(C^N \mid S^N : \hat{\theta} \prec M) + \lambda\{\ell(\hat{\theta} \mid M) + \ell(M)\}$$

where λ is the adjustment parameter. The MDL principle asserts that one should select the stochastic clauses $\hat{M}$ which minimize the total description length. That is, $\hat{M}$ is given by:

$$\hat{M} = \arg\min_{M\in\mathcal{M}} \ell(C^N : \hat{\theta} \prec M \mid S^N) \quad (5)$$

We call the hypothesis specified by $\hat{M}$ and the estimated parameter $\hat{\theta}$ for $\hat{M}$—$P(C \mid S : \hat{\theta} \prec \hat{M})$—the *MDL hypothesis* . Notice here that it may be computationally intractable to find $\hat{M}$ that minimizes the total description length when all possible combinations of Horn clauses $\mathcal{M}$ is large. We call the strategy for finding hypotheses with smaller total description length (4) the *MDL strategy*.

4 Experimental Results

We will apply our methodology to learning stochastic desicion predicates for discriminating "mitochondria cytochrome c." Here the mitochondria cytochrome c refers to a subclass of cytochrome c protein which carries an electron in the respiratory chain. Let S be the set of all sequences appeared in the Protein Identification Resources R18.0 that contains 6158 amino acid sequences, and C be a set $\{mcyt.c, others\}$. We assume the sequences consist of the twenty-letters, each of which represents an amino acid. Hereafter, "*mcyt.c*" denotes the mitochondria cytochrome c.

Let the following three stochastic decision predicates (SDP, for short) given as candidates for the representation of the optimal motif.

- $motif(S, mcyt.\ c)$ (with p_1) :-
  $contain(S,$ "CXXCH").
  $motif(S, others)$ (with p_2).
- $motif(S, mcyt.\ c)$ (with p_1'):-
  $contain(S,$ "CXXCH")$\wedge contain(S,$ "PGTKM").
  $motif(S, others)$ (with p_2').
- $motif(S, mcyt.\ c)$ (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, genetic algorithm etc.).

Table 1 shows the distribution of the training sequence observed through these three SDPs. In Table 1, the motif patterns described in the first column are those of the SDPs described above. For each pattern on the left, the number of target sequences for discrimination and the number of positive examples that contain the corresponding pattern(s), are shown.

In Table 2, L_1, L_2, L_3, and Total denote the description lengths calculated by (1),(2),(3), and (4), respectively, where λ is set to be 1. $\hat{p}_1$ and $\hat{p}_2$ denote the Bayes estimates of the probability parameters in the first and second clauses, respectively.

Table 1: Distribution of Mitochondria Cytochrome C Sequences

Motif Pattern	N_1 and N_2	N_1^+ and N_2^+		$\hat{p}_1$ and $\hat{p}_2$
CXXCH	189	67		0.356
others	5969	5966		0.9993
CXXCH and PGTKM	73	67		0.906
others	6085	6082		0.9993
CXXCH and GPXLXG and PGTKM	71	67		0.932
others	6087	6084		0.9993

Table 2: Description Lengths for Stochastic Decision Predicate

Motif Pattern	L_1	L_2	L_3	*Total*
CXXCH	214.7	10.1	29.7	255.5
CXXCH and PGTKM	67.5	9.4	53.4	131.3
CXXCH and GPXLXG and PGTKM	59.9	9.4	76.2	146.5

Here, the columns in L_1,L_2,L_3, and *Total* denote the description length of the examples relative to the predicates, the description length of estimated parameters of the predicates, the description lengths for the stochastic decision predicates that contains the pattern in the column *MotifPattern*, and the total length of L_1,L_2,L_3, respectively.

Table 3: Cross Validation Estimates for Prediction Error for Mitochondria Cytochrome C Motifs

	MDL principle	ML method
Aver. Pred. Error Rate	0.0008	0.0013

We can see from Table 2 that the second SDP containing "CXXCH" and"PGTKM" is optimal in the sense of the MDL principle. The first SDP containing only a pattern "CXXCH" is considered as too simple for discrimination of the mitochondria cytochrome c. On the other hand, the third SDP containing "CXXCH," "GPXLXG," and "PGTKM" can discriminate the given sequences best but is considered as too complicated. This suggests that the MDL strategy possibly avoids the overfitting problem; the rule which best fits the given example is sometimes affected by statistical irregularities, and thus, such a rule is not always best in terms of predicting the labeling of future data.

Table 3 shows the estimates of the average prediction error rate for the SDPs selected by the MDL principle and by the maximum likelihood (ML) method. Here the ML method refers to a strategy which selects a stochastic decision predicate such that minimizes the description length of examples ignoring the description of the predicate itself. The estimates of the error rates are obtained by applying the cross validation method ([BFOS 84] p.75-76) to the original training sequences: We split the original set S of all training sequences into 10 subsets $S_1, \cdots, S_{10}$ ($|S_1| = \cdots = |S_{10}|$). For each S_i, we construct a SDP from $S - S_i$, then transform it into a deterministic rule by letting all probability parameters be 1, and test it on S_i. We denote the number of prediction errors on S_i of rules obtained by the MDL method and by the ML method as $Error_{MDL}(S_i)$ and $Error_{ML}(S_i)$. Let R_{MDL} and R_{ML} be the cross validation estimates of error rates for the MDL principle and ML method. 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 SDPs selected by the MDL principle have less prediction errors than those selected by the ML method, in average.

5 Discussion

Let us mention that the motif extracted by our method is not only computationally meaningful but also biologically meaningful. Actually, in the case of mitochondria cytochrome c, the cysteines (C) appearing in the pattern "CXXCH" denote binding sites for a heme c, and histidine (H) in the "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 tend to be preserved in the evolution process. The motif extracted by our method is sound in the sense that it contains all of these functional sites. However it should be noted that we cannot always extract biologically meaningful motifs by means of statistical analysis of genetic sequences. For this purpose, we have to consider the properties of nucleic acids and amino acids, and the structural information of proteins, such as α–helixes and β–strands. The following work remains to deal with actual genetic sequences on the basis of our methodology.

- An efficient algorithm to find an optimal stochastic decision predicate automatically: A prototype system is now being developed in the fifth generation computer systems project. According to our experience so far, stochastic search algorithms such as the *genetic algorithm* seems to be effective for our purpose.

- The handling of point mutations and experimental ambiguity: For example, actual amino acid sequences contains mutation information and special characters that represent ambiguous elements, such as B for asparagine or asparatic acid, and Z for glutamin and glutamic acid. The disjuction form of stochastic decision predicate may help this to some extent. However, such information should be counted for the calculation of description length of the given examples.

- The handling of category hierarchy: The current MDL strategy might select a motif pattern which has nothing to do with the upper category of the target category. For example, the MDL strategy might select only "PGTKM" instead of "CXXCH"

$\wedge$ "PGTKM" where "CXXCH" represents a motif pattern for cytochrome c, a parent category of mitocondoria cytochrome c. Such selection is tolerable for the purpose of database search, but might be dangerous in the sense that it might lose biological meaning.

6 Conclusion

We have proposed a new methodology for learning stochastic motifs from given genetic sequences. Our proposed methodology is characterized by the new representation of stochastic motifs using stochastic decision predicates (SDP) and by the MDL learning strategy. Our experimental results show that the methodology actually produces a computationary and biologically meaningful motif for mitochondria cytochrome c, whose good predictive performance has been statistically proven by the cross validation method. We believe the methodology can also be applied to the various kind of discrimination problems on genetic sequences.

Acknowledgement The authors wish to express their sincere gratitude to Dr. K. Nitta of ICOT, and to Mr. S. Koike, Dr. M. Yokota, Mr. K. Nakamura and Mr. A. Kaneko of NEC Corporation for their encouragement and support. The authors also thank Mr. K. Yamagishi, Mr. S. Oyanagi of NSIS, and Miss K. Hikita of C&C Systems Research Laboratories, NEC Corporation for their great contribution to our analysis of real genetic sequences.

References

[AA 90] Aitken, Alastair, (1990). *Identification of Protein Consensus Sequences*, Ellis Horwood Series in Biochemistry and Biotechnology.

[BFOS 84] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J.(1984). Classification and regression trees. *Wadsworth Statistics/Probability Series.*

[BOP 90] Babcock, M.S., Olson, W.K., & Pednault, P.D.(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.

[CK 90] Cheeseman, P. & Kanefsky, B.(1990). Evolutionary Tree Reconstruction. *Preprint of Symposium Papers: The Theory and Application of Minimal-Length Encoding, AAAI 1990 Spring Symposium Series*, Stanford, CA.

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

[Ris 78] Rissanen, J.(1978). Modeling by shortest data description. *Automatica, 14*, 465-471.

[Ris 83] Rissanen, J.(1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11*, 416-431.

[Ris 89] Rissanen, J.(1989). *Stochastic Complexity in Statistical Inquiry*, World Scientific, Series in Computer Science, 15.

[Sol 64] Solomonoff, R.J.(1964). A Formal Theory of Inductive Inference. Part 1. *Information and Control, 7*, 1-22.

[WB 68] Wallace, C.S. & Boulton, D.M.(1968). An information measure for classification. *Computer Journal*, 185-194.

[Yam 90] 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.

[YK 91] Yamanishi, K. & Konagaya, A.(1991). Leaning Stochastic Motifs from Genetic Sequences. *Submitted to the Eith International Workshop of Machine Learning.*