TM-1234

遺伝子的アルゴリズムによる分類規則の学習

小長谷　明彦（日電）

November, 1992

© 1992, ICOT

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

遺伝的アルゴリズムによる分類規則の学習

Learning Classificication Rules by a Genetic Algorithm

○ 小長谷明彦　　(NEC)

Akihiko KONAGAYA, NEC Corporation, 1-1, Miyazaki, 4-chome, Miyamae-ku, Kawasaki

This paper proposes a new genetic algorithm for learning classification rules. The novelty of our algorithm is its use of Rissanen's Minimum Description Length (MDL) principle when selecting "good individuals". Experiments on extracting motifs from protein sequences indicate that this gives an appropriate bias for selecting stable classification rules (motifs). Our experience also shows that the convergence speed largely depends on the adjustment parameter which balances the description length of classification rules and its log-likelihood to the sample protein sequences.
Key Words: Genetic Algorithm, MDL Principle, Machine Learning, Motif Extraction, Genetic Information Processing, Protein Sequence

1　はじめに

遺伝的アルゴリズムの主要な応用の一つとして、分類規則の学習がある。一般に学習データが与えられたとき、これを分類する規則は組合せ的に存在する。このため、最適な分類規則を選択することは極めて困難な問題の一つとされている。このような観点から、分類学習問題は遺伝的アルゴリズムの応用として古くから研究されており、これまでにも様々な分類システム (Classifire System) が提案されている [1]。しかしながら、このような分類システムで利用されている学習アルゴリズムでは分類規則の選択基準があまり明確でないという問題があった。

一方、有効な分類規則の選択基準として、分類規則の学習データに対する対数尤度 (log likelihood) が知られている [2]。対数尤度は学習データと学習で得られたモデル (この場合は分類規則) とのずれを情報量として表したものであり、サンプルの偏りがないという仮定のもとで、分類規則の学習データへの整合性を表す良い指標の一つとなっている。この対数尤度を遺伝的アルゴリズムの個体の選択基準として利用すれば、学習データへの整合性の高い分類規則を高速に学習することが期待できる。本稿では、この学習法を最尤法 GA と呼ぶ。

最尤法 GA では与えられた学習データに最も適合した分類規則を学習することを目標とする。しかしながら、実世界で扱うデータでは多くの場合与えられた学習データにエラー (ノイズ) が含まれており、学習データに最も適合した規則が必ずしも未知のデータの予測に役立たないという問題が生じる。この問題は「過剰適合 (Overfitting)」と呼ばれ、機械学習における大きな課題の一つとなっている。

この過剰適合を避ける機械学習アルゴリズムとして、Rissanen の Minimum Description Length (MDL) 基準を個体の評価基準とする遺伝的アルゴリズム (MDL GA) を提案する。具体的には、分類規則の評価基準として、対数尤度と分類規則を符号化したときの記述長の和で判断する。これにより、学習過程においてできるだけ簡潔な分類規則を抽出する方向にバイアスがかかるため、学習データの持つノイズの影響から必要以上に分類規則が詳細化されるのを防ぎ、より安定的な分類規則の学習が期待できる。

実際、MDL-GA をタンパク質配列からの共通パターン (モチーフ) 抽出に適用したところ、学習データのサンプリングに依存しない安定的なモチーフを抽出できることが確認できた。また、最適解への収束速度は分類規則の記述長と対数尤度との重み付けを決める調節パラメタ (Adjustment parameter) に強く依存し、分類規則の記述長への重みが少なくなるほど収束速度が遅くなることを確認した。例えば、シトクロム c というタンパク質からのモチーフ抽出の例では、分類規則の記述長を考慮しない場合すなわち最尤法 GA では観測時間内 (1 万世代) では最適解への収束は認められなかった。

本稿の構成を以下に示す。はじめに、2 節において学習の対象となるタンパク質配列からのモチーフ抽出問題について述べ、、3 節でモチーフの表

現形式として提案した確率的決定述語 [3] について述べる。次に、4節で、MDL基準を、5節で遺伝的アルゴリズムについて紹介する。そして、6節において、遺伝的アルゴリズムのモチーフ抽出への適用法について述べ、7節で抽出結果の評価について述べる。

2 モチーフ抽出問題

モチーフ抽出は、種々の生物における同一のタンパク質の塩基配列あるいはアミノ酸配列を比較し、進化的に保存されている配列パターン (モチーフ) を見つける操作である。例えば、良く知られたモチーフとしては、シトクロムcのCXXCHがある [4]。ただし、ここで、各文字はアミノ酸を表し、Cはシステイン、Hはヒスチジン、Xは任意のアミノ酸を表す。シトクロムcは呼吸鎖の電子伝達に関わる酵素であり、鉄を含有するヘム分子と結合することにより酸化還元反応を行なう。シトクロムcにおいてモチーフ *"CXXCH"* はこのヘム分子との結合部位となっており、このような機能を維持するために進化的に保存されてきたと考えられている。

モチーフ抽出は機械学習における「帰納学習」に相当し、これまでにも計算機による自動抽出が試みられているが [5, 6]、得られた結果の精度ならびに計算速度の観点で十分とは言い難い。この問題の難しさは (1) 例外のないモチーフを見つけることが極めて困難なこと、(2) 解の候補が組合せ的に多くなること、(3) 例外の少ないモチーフが必ずしも分類予測の観点で最適でないこと、が挙げられる。

3 確率的決定述語

モチーフは進化的に保存されているパターンを意味するが、これは絶対的なものではない。また、あるモチーフを持つことが必ず特定の蛋白質であることを保証するものでもない。例えば、シトクロムcにおいても、ミドリムシのシトクロムcは "CXXCH" の代わりに "AAQCH" を持つ。また、豚のアドレノドクシンは "CXXCH" を持つがシトクロムcではない。そのようなモチーフの持つ確率的性格を考慮した表現方法が確率的決定述語である。

以下に確率的決定述語によるモチーフの表現例を示す。

```
motif(S,cytochrome_c) with 137/244
        :- contain(S,''CXXCH'').
motif(S,others) with 9386/9389.
```

この表現は、S が *"CXXCH"* を含めば確率 $\frac{137}{244}$ で S はシトクロムcであり、そうでなければ、確率 $\frac{9386}{9389}$ で *others*(シトクロムc以外のタンパク質) であることを意味する。

4 記述長最小 (MDL) 基準

モチーフを確率的決定述語で表現する場合、述語の条件部におかれる条件の数及び連言／選言の形態、*contain* 述語に含まれる配列パターン、述語に付与される確率等々を変えることにより、多様な表現が可能になる。それらの表現の中で良い表現を選ぶ基準がMDL基準である。

MDL基準は、ノイズを含む学習データから確率モデルの推定を行なう際に有効な基準であり、確率モデルの記述長と確率モデルを用いた時の学習データの記述長の和を最小にする確率モデルを最良の確率モデルであるとする考え方である。確率的決定述語の場合には、

λ (確率的決定述語の記述長) + 対数尤度

を最小にするものが良いと判断される [7, 8]。ここで、λ は記述長の調節パラメターであり、λ が0のときが最尤法に相当する。確率的決定述語の記述長はモチーフ表現が複雑になるほど値が大きくなり、対数尤度はモチーフ表現が対象とするタンパク質を正しく分類できるようにするほど値が小さくなる。したがって、MDL基準では、ノイズに惑わされず、適度な複雑さと正確さを備えた安定的なモチーフを選択することが期待できる。具体的な記述長の計算法を付録Aに示す。より詳細には、文献 [9] を参考にされたい。

5 遺伝的アルゴリズム

遺伝的アルゴリズムは、生物の進化の過程をモデルとして考案された確率的探索アルゴリズムである [1]。ここで採用した単純遺伝的アルゴリズム (*Simple Genetic Algorithm*) は最も基本的な遺伝的アルゴリズムとして知られており、次のような探索を行なうアルゴリズムである。

関数 f が与えられた時、その関数 f の最小値を与えるような関数 f の定義域中の点を探索するために、単純遺伝的アルゴリズムを適用する場合は以下のような手順を踏む。

関数 f の定義域は探索空間に対応するが、その各点に対して例えば、*000110*、*110111* などの2

進の表現を与える。即ち、探索空間を固定長の2進の文字列の集合と対応づける。各2進文字列に対してその点の関数fの値が計算可能である。

次に、初期集団（*Initial Population*）を設定する。これは、一定数の2進文字列の集まりであり、探索空間の初期の探索点の集合である。この集団、即ち探索点の集合を世代毎に更新し、適当な世代後の最も良い即ち関数fの値が最も小さい2進文字列に対応する定義域中の点が求める探索点となる。この求められた点が関数fの最小値を与えるという理論的な保証があるわけではないが実験的には良い結果を与えることが多い。

集団を更新する各世代では以下の一連の操作を行なう。

1. 選択（*Selection*）

   集団中に存在する各2進文字列に対してその関数fの値を計算する。この関数fの値がより小さいほどより適応していると考え、より適応しているものが集団中により多くなるように2進文字列の選択・増殖を行なう。このとき、選択・増殖は確率的に行なうため、適応度の低い2進文字列が生き残る可能性も排除されるわけではない。この、選択操作はより良い候補となりそうな探索点を増やすという効果を持つ。

2. 交叉（*Crossover*）

   集団中の2つの2進文字列をとり、その部分文字列を交換した2進文字列を作る。例えば、*000110*と*110111*とを3番目と4番目のビットの間で交叉させるとその結果は、*000111*と*110110*になる。このとき、どの2進文字列をどのくらい交叉させるか（交叉確率）、どのビット間で交叉させるかなどは確率的に決定する。この、交叉操作は複数の探索候補点をマージして新たな候補点を得る操作である。

3. 突然変異（*Mutation*）

   集団中の1つの2進文字列に対し、そのあるビットを反転する。例えば、*000110*の第3ビットを反転させると*001110*となる。このとき、突然変異を起こさせるかどうか（突然変異確率）、どのビットに起こすかどうかなどは、確率的に決定する。この、突然変異操作は新たな候補点を得る操作であるが、交叉操作が特定のビット位置に注目した時集団中でのそのビット位置での０１の存在の多様性に変化を与えないものであるのに対し、この突然変異操作は集団中での特定のビット位置での多様性に変化を与えるところに特徴がある。

6 モチーフ抽出

モチーフ抽出への遺伝的アルゴリズムの適用は次のように行なった。

探索空間は確率決定述語表現であり、関数fは各確率決定述語表現に対して定まる記述長である。関数fの値即ち記述長が最も小さい確率的決定述語表現を探索することが目的である。問題となるのは、探索の候補となる確率的決定述語表現であるがこの表現形態を無制限に自由なものにすると探索空間が莫大となることは避けられない。ここでは、対象とする確率的決定述語表現を以下のようなタイプに制限した。

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

即ち、節の数は、*proteinClass*を表す節とそれ以外（*others*）の2つとし、*proteinClass*を表す節の条件部は、*contain*述語の連言結合とする。また、*contain*述語に含まれるパターンは実際の蛋白質データベースにおいて出現頻度の高い128個を採用した。

確率的決定述語表現の2進文字列表現は、128ビットの2進文字列の各位置にそれぞれ出現頻度の高い128個のパターン1つとを対応づける。そして、各位置のビットが1の時は、対応づけられたパターンの*contain*述語が条件部に存在し、ビットが0の時は存在しないものとする。仮に、3ビットの例で示すと、1番目のビット位置に*"CXXCH"*が2番目のビット位置に*"PXLXG"*が3番目のビット位置に*"GXKM"*が対応づけられているとする。この時、2進文字列が*100*ならば、それに対応する確率的決定述語は、

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

となり、また2進文字列が*011*ならば、対応する確率的決定述語は、

```
motif(S,proteinClass) with p1
 :- contain(S,"PXLXG")
```

```
  & contain(S,"GXKM").
motif(S,others) with p2.
```

となる。

この対応関係により、2^{128} 種類の確率的決定述語表現それぞれが 128 ビットの 2 進文字列に対応づけられる。

初期集団については、ランダムに発生させた 128 ビットの文字列を 64 個使用した。また、交叉確率は 1.0、突然変異確率は 0.01 に設定した。

7 評価

表 1 にシトクロム c に対し、MDL 基準を用いた遺伝的アルゴリズムを用いてモチーフ抽出を行なったときの予測エラーの個数 (MDL-GA) と最尤法を用いたときの予測エラーの個数 (最尤法 -GA) を示す。予測エラーの測定には、クロス検定法を用いた。すなわち、PIR のデータバンクを 10 等分し、10 分の 9 のデータから求めた確率的モチーフに対し、これを決定的なモチーフと解釈し、さらに、残りの 10 分の 1 のデータを未知データとして与え、分類した際のエラーの個数を全ての組合せについて求めた。以下、本結果について考察する。シトクロム c の場合には、合計のエラー個数は最尤法 GA の方が少ないが、その解釈には十分な注意が必要である。まず、シトクロム c 以外の配列をシトクロム c と判定した負例エラーの個数は MDL-GA の方が最尤法 GA より高い。この原因の一つとして、シトクロム c と同様にヘム分子との結合部位を持つシトクロム c' やシトクロム f が CXXCH というモチーフを持つことが挙げられる。実際、分類の対象をシトクロム c でなく、ヘム c 結合タンパクとするだけで負例エラーは 66 に減らせることが確認されている。ただし、CXXCH が分類条件として単純すぎることは事実であり、より精度の高い分類条件の抽出が課題となっている。一方、シトクロム c の配列をシトクロム c でないと判定した正例エラーの個数は最尤法 GA の方が MDL-GA よりも高い。これは、逆に、最尤法 GA で求めたモチーフが与えられた学習データに適合しすぎていることを表している。実際、MDL-GA を用いた場合には、全ての組で同じモチーフ CXXCH を抽出しているのに対し、最尤法 -GA では各組毎に抽出したモチーフが異なっており、安定したモチーフが抽出できていないことを示している。

ここで注意しなくてはならないのは、このような現象が当初想定していたような最尤法による過剰適合からではなく、最尤法 -GA の収束の遅さに起因していることである。実際、シトクロム c の場合、今回の実験で用いた探索空間の中では最尤法も MDL 基準も最適解は CXXCH となることが確認されている。このことを利用して調節パラメター λ の収束速度への影響を測定した。図 1 に示すように、λ が少なくなるほど収束速度は遅くなっている。さらに、λ が 0 の場合は 1 万世代まで実行しても最適解を得ることができなかった。

8 結論

遺伝的アルゴリズムを用いた分類学習において、MDL 基準を用いることにより、ノイズを含む学習データにおいて安定的な解が得られることをモチーフ抽出を例として述べた。また、最尤法と比べ、MDL 基準の方が (最適) 解への収束速度が早いことを確認した。今後は、最尤法 GA における過剰適合の確認、MDL-GA における過剰簡略化の回避策について検討してゆく予定である。なお、本研究は第五世代計算機プロジェクトにおいて ICOT からの再委託により並列推論マシンの評価作業の一環として行なわれた。

謝辞 本研究を進めるにあたって、本研究の機会を与えて頂いた ICOT の Dr. 新田室長ならびに MDL 基準に基づく確率的決定述語の学習に関して助言を頂いた C&C 情報研究所の山西部員に深謝致します。また、本研究に必要なプログラムならびにデータの収集をして頂いた日本電気技術情報株式会社の小柳氏ならびに遺伝子情報処理グループの皆様に感謝の意を表します。

参考文献

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

[2] Quinlan,J.R., (1987). Decision Trees as Probabilistic Classifiers, Proc. 4th Int. Conf. on Machine Learning, pp.31 37.

[3] Konagaya,A. & Yamanishi, K. (1991). A Stochastic Desicion Predicate: A Scheme to Represent Motifs, *in the AAAI Workshop of Classification and Pattern Recognition in Molecular Biology.*

表 1: Evaluation of Prediction Errors by Cross Validation Method

	MDL-GA	Maximumly Likelihood-GA
Positive Example Errors	3	57
Negative Example Errors	96	0
Total Errors	99	57

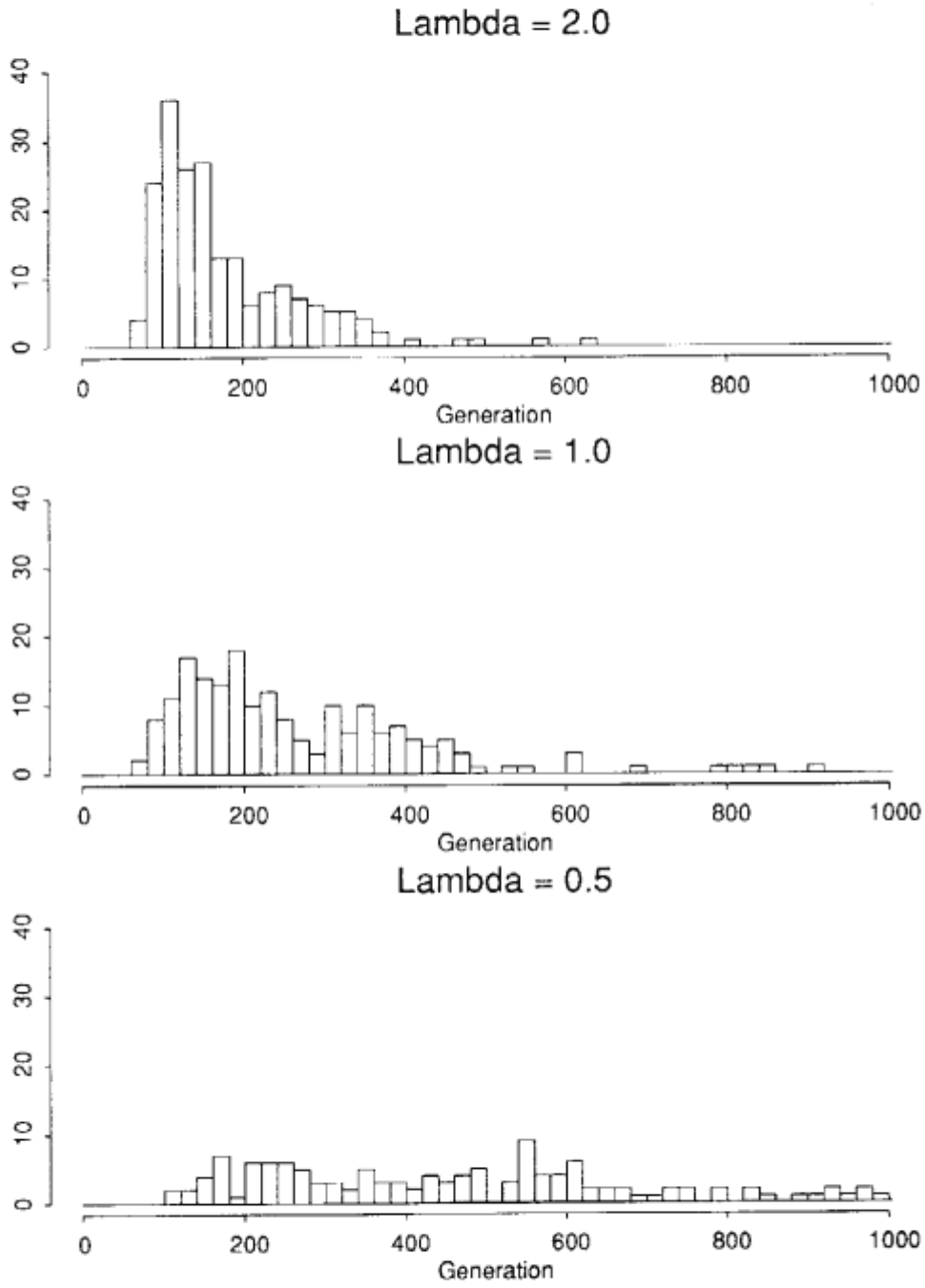

図 1: Comparison of Convergeneee Speed by the Distributions of generations in which the optimal solution is found

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

[5] Rooman,M.J. & Wodak,S.J., (1988). Identification of Predictive Sequence Motifs limited by Protein Structure Data Base Size, *Nature*, vol.335, no.1, pp.45-49.

[6] Smith,H.O., Annau,T.M. & Chandrasegaran, S., (1990). Finding Sequence Motifs in Groups of Functionally Related Proteins, *in Proc. Natl. Acad. Sci. USA*, vol.87, pp.826-830.

[7] 小長谷.山西.(1990). 「記述長最小基準の遺伝子情報処理への適用について」. ソフトウエア科学会第7大会論文集. pp.101-104.

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

[9] Konagaya,A.. (1992). A Stochastic Approach to Genetic Information Processing, *in the Workshop on Algorithmic Learning Theory ALT'92.*

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

付録A 記述長の計算法

分類規則の評価は、対数尤度と確率的決定述語の和で行なう。対数尤度の計算法は以下の通りである。

$$LL = -\log \prod_{j=1}^{m} p_j^{N_j^+} (1-p_j)^{N_j - N_j^+} \tag{1}$$

$$= \sum_{i=1}^{m} N_i \{H(\tilde{p}_i) + D_{KL}(\tilde{p}_i \parallel \hat{p}_i)\} \tag{2}$$

ただし、p_i はモチーフのi番目のクローズのパターン(のAND結合)が特定のタンパク質に出現する確率、$\hat{p}_i = N_i^+/N_i$ であり、$\tilde{p}_i$ は p_i の真の推定量である。本方式では、推定量としてベイズ推定量 $\frac{N_i^+ + 1}{N_i + 2}$ を用いている。N_i はそのパターンを持つ配列の個数、N_i^+ は対象のタンパク質においてそのパターンを持つ配列の個数を表す。

また、H、D_{KL} はそれぞれ、下記の式で定義されるエントロピー関数およびKullback-Leibler情報量である。

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

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

確率的決定述語の記述長は、パラメタの記述長とホーン節の記述長からなる。パラメタの記述長を以下の式で計算する。

$$PL = \sum_{i=1}^{m} \frac{\log N_i}{2} \tag{3}$$

ただし、N_i は、i番目のホーン節の本体部の条件を満たした配列の個数を表す。

ホーン節の記述長は以下の式で計算する。

$$\begin{aligned} CL \;=\; & \sum_{i=1}^{m} [\, \log^*(\sum_{j=1}^{k_i} h_j) + (\sum_{j=1}^{k_i} h_j - 1) \\ & + \sum_{j=1}^{k_i} \sum_{l=1}^{h_j} \{\log \begin{pmatrix} L_l^j(i) \\ N_l^j(i) \end{pmatrix} \\ & + (L_l^j(i) - N_l^j(i)) * \log(20-1)\} + \log r \,] \end{aligned} \tag{4}$$

ここで $L_l^j(i)$ と $N_l^j(i)$ はそれぞれ、i 番目の節における、j 番目のOR結合領域における、l 番目のパターンに含まれるアミノ酸の個数と変数の個数を表す。また、任意の $d > 0$ に対し、$\log^* d$ は、Rissanenの整数符号法[10]、すなわち、$\log d + \log\log d + \cdots$ における正数項の和を表す。2番目の項は、i番目の項における $\vee, \wedge, \wedge, \cdots$ の列を符号化するための記述長を表す。3番目の項は、"$contain(S, \sigma)$" におけるパターン σ 中の変数の個数の記述長を表す。4番目の項は、パターン σ 中の変数以外のアミノ酸20種のパターンを表現するための記述長を表す。最後の項は、"$motif(S, C)$" において、カテゴリ C を表現するための記述長を表す。r は確率的決定述語に現れるカテゴリの個数である。