遺伝子情報処理と記述長最小（ＭＤＬ）基準

Genetic Information Processing and the Minimum Description Length Principle

小長谷明彦

Akihiko Konagaya

日本電気株式会社 C&C システム研究所

C&C Systems Research Labs., NEC Corporation

This paper stresses the effectiveness of the minimum description length (MDL) principle for genetic information processing. When we extract rules from genetic sequences, stochastic approaches are necessary because of uncertainty due to the intrinsic varieties caused by mutation in the evolutionary process. The MDL principle plays an essential role in selecting good stochastic rules from the viewpoint of predictive performance for unknown sequences as well as discriminatory performance for the given genetic sequences. This paper also presents a new scheme for representing the mapping from genetic sequences to categories called *Stochastic Decision Predicates* and describes a methodogy for applying the MDL principle to the selection of stochastic decision predicates. Our experimetal results demonstrate that the MDL principle produces motifs with less predictive errors than the maximum likelihood method.

1　はじめに

現在、ヒトをはじめとして様々な生物の遺伝子情報（ＤＮＡ配列情報、アミノ酸配列情報）が分子レベルで解明されつつある [DOE 88]。遺伝子情報が蓄積されるにつれ、分子生物学と情報処理の結びつきがクローズアップされてきた。一つの理由は、遺伝子情報のデータ量の多さである。現在、遺伝子情報はすでに５０００万塩基を越えており、計算機の利用が不可欠となっている。もう一つの理由は、遺伝子情報そのものの複雑さである。遺伝子情報は複雑な暗号のようなものであり、その解析には分子生物学に関する広範な知識と高度な情報処理技術の適用が不可欠である。この意味で、遺伝子情報処理は、分子生物学と情報処理技術の両方が合わさってはじめて可能となる未知の境界領域といえよう。

本稿では、このような遺伝子情報処理技術の一つとして、記述長最小（ＭＤＬ）基準が有効なことをアミノ酸配列における確率的な規則抽出を例にして示す。遺伝子情報が何らかの規則性を保持していることは生成されるタンパク質の構造が遺伝子情報より一意に決定されることより明らかである。遺伝子情報処理の目的の一つは、このような規則性を観測された遺伝子データから推論することにある。このような処理は計算論的学習理論における「分類学習」とみなすことができる。例えば、遺伝子情報の一つであるアミノ酸配列から生成するタンパク質を推論する規則（以下、分子生物学の用語にしたがってモチーフ [AA 90] と呼ぶ）の抽出を考える。モチーフは、例えば、「もし、アミノ酸配列がＣＡＱＣＨというアミノ酸のパターンを含めばそれはシトクロムＣである」という推論規則として表現できる。しかしながら、生物は突然変異などの多様性を持つので、ＣＡＱＣＨを含む配列が必ずしもシトクロムＣというタンパク質に分類されるとは限らない。逆に、シトクロムＣでないタンパク質がＣＡＱＣＨというパターンを持つ場合もあろう。一般に、突然変異はどのように生じるのか予測不可能であるから、このような例外を全て考慮した推論規則を抽出するのは極めて困難と言わざるえない。

この問題を解決する一つの手段は、決定的な規則の代わりに確率的な規則を用いることである。例えば、先の例は、「もし、アミノ酸配列がＣＡＱＣＨというアミノ酸のパターンを含めばそれは確率４／５でシトクロームＣであり、確率１／５

でそうでない」という確率的規則として表現できる。このような確率的なモチーフ表現は、細部の例外的な記述を省略することができるため、モチーフ抽出をより容易に行なうことができる [KY 90, KNY 90, YK 91]。

ここで、注意すべき点は、果たしてどのような確率的モチーフが本当に良いかという基準である。配列抽出を計算機で自動的に行うためには、確率的モチーフの評価基準が不可欠である。直観的には例外を含まず、確率1にできるだけ近い確率的モチーフが良いように思える。しかしながら、そのようにして求めた確率的モチーフは与えられた学習セットにしか有効でないモチーフかも知れない。遺伝子情報のモチーフ抽出においては、未知データに対しても有効なモチーフを抽出することが重要である。したがって、確率的モチーフの基準としては、未知データに対する規則の有効性すなわち「予測精度」を重視すべきである。MDL基準 [Ris 78, Ris 89] はまさにこの予測精度を最大にすることを目的とした基準であり、確率的規則の学習においても有効であることが理論的に示されている [Yam 90]。本稿では、このMDL基準を用いて、より予測精度の高いモチーフを抽出できることを実データを用いて示す。

本稿の構成を以下に示す。はじめに、2節において、確率的モチーフの表現形式として提案した確率的決定述語 [KY 91] について紹介する。次に、3節では、MDL基準に基づいて、より良いモチーフを選択するための計算法について示す。そして、最後に、4節において、具体例による評価結果を示す。

2 確率的決定述語

本節では、確率的モチーフの表現形式である確率的決定述語について述べる。確率的決定述語は確率変数を備えたホーンクローズからなる。一般形式を次に示す。

$$
\begin{aligned}
&motif(S, C_1) \quad (\text{with } p_1) \quad :- Q_1^{(1)} \wedge \cdots \wedge Q_{k_1}^{(1)}.\\
&motif(S, C_2) \quad (\text{with } p_2) \quad :- Q_1^{(2)} \wedge \cdots \wedge Q_{k_2}^{(2)}.\\
&\qquad \cdots\cdots\cdots\cdots\cdots\\
&\qquad \cdots\cdots\cdots\cdots\cdots\\
&motif(S, others) \quad (\text{with } p_m).
\end{aligned}
$$

各クローズは条件部 $Q_1^i \wedge \cdots \wedge Q_k^i$ が全て真のとき確率 p_i で真となる確率的規則を表す。アミノ酸配列のモチーフでは、S がアミノ酸配列を、C_i がタンパク質のカテゴリを表す。また、特殊なカテゴリとして、「その他」の集合を表す $others$ を用意し、指定された条件以外の場合に対応させる。ここで、各確率変数 p_i の値は各々のクローズ毎に独立に定義されることに注意されたい。また、各クローズの選択は上から逐次的に行う。すなわち、i番目のクローズの条件部は先頭からi-1番目までのクローズの条件部の否定を暗黙に仮定している。

条件部は述語の連言標準形で表現する。すなわち、条件部の各 Q_i は述語のOR結合 $R_1^i \vee \cdots \vee R_j^i$ を許す。OR結合はその他の条件部をコピーすればクローズの形式に展開できるため、クローズの表現能力が変わるわけではない。しかしながら、冗長な条件部のコピーを避けることができるため、3節で述べる確率的決定述語の記述長の増大を抑えることができ、結果として、突然変異情報を含むモチーフを選択しやすくするという効果がある。

また、本稿では、条件を表す述語として配列Sがパターン σ を含むとき真となる contain(S, σ) のみを扱う。これは議論を簡潔するための便宜的な制限であり、モチーフの表現としては3節で述べる符号化が定義できれば任意の述語を扱うことが可能である。

3 記述長の計算法

確率的決定述語の記述長は、データの記述長 (DL)、確率パラメタの記述長 (PL) およびクローズの記述長 (CL) の和によって与えられる。データの記述長は確率的決定述語を通して学習セットを記述する際に必要な記述長であり、対数尤度で与えられる。学習セットとして N 個の配列が与えられたとき、N_j を j 番目のクローズの条件を満たす配列の個数、N_j^+ を j 番目のクローズで定義されたカテゴリに属する配列の個数とする。このとき、求めた確率的決定述語のもとで学習セットを観測する確率、すなわち、確率的決定述語の尤度 (F) は下記の式で表される。

$$F = \prod_{j=1}^{m} p_j^{N_j^+} (1 - p_j)^{N_j - N_j^+}.$$

そして、データの記述長 (DL) は F の逆数の対数をとって、以下の式で与えられる。

$$DL = -\log F = \sum_{i=1}^{m} N_i \{H(\tilde{p}_i) + D(\tilde{p}_i \,\|\, \hat{p}_i)\}$$

ただし、$\tilde{p}_i = N_i^+/N_i$ であり、$\hat{p}_i$ は真の確率変数 p_i^* の推定値であり、N_i^+/N_i(最尤推定値) また

は $\frac{N^+_i+1}{N_i+2}$ (ベイズ推定値) を用いる。さらに、$H(\tilde{p}_i)$ および $D(\tilde{p}_i \parallel \hat{p}_i)$ はそれぞれ、エントロピー関数、Kullback-Leibler 情報量であり、

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

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

で定義される。データの記述長 (DL) は、確率的決定述語に対する正例および負例の分布を符号化するために必要な記述長を表し、その長さは 0 ビット ($p_i = 0$ *or* 1.0 $(i = 1, \cdots, m)$ のとき) から N ビット ($p_i = 0.5 (i = 1, \cdots, m)$ のとき) まで変化する。前者は確率的決定述語が例外無しに完全に分類できる場合であり、後者は確率的決定述語が分類に関して全く貢献していない場合に対応する。

PL を確率的決定述語の確率変数の記述長とする。確率変数の推定値の精度は N を条件を満足する学習配列の個数とすると高々 $O(1/\sqrt{N})$ でしかないので、

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

で計算できる。

クローズの記述長 CL は次式で与えられる。

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

ただし、$L_l^j(i)$、$X_l^j(i)$ はそれぞれ i 番目のクローズの j 番目の選言の l 番目の述語のパターン中に現れるパターンの長さおよび変数の個数である。最初の項は、i 番目のクローズにおける contain 述語の個数の記述に必要な記述長を表す。整数 $d > 0$ に対し、$\log^* d$ は $\log d + \log\log d + \cdots$ を表す。ただし、和は正数についてのみ計算する (Rissanen's integer coding scheme [Ris 83]). 2 番目の項は、i 番目のクローズにおける AND-OR 結合の組合せの記述に必要な記述長を表す。3 番目の項は、述語 '$contain(S, \sigma)$' 中に表れるパターン σ における変数の位置を記述するために必要な記述長である。4 番目の項は、パターン σ における変数以外の文字列を記述するために必要な記述長である。最後の項は、確率的決定述語に表れるカテゴリの数を記述するために必要な記述長である。

表 1: ミトコンドリアシトクロム C の分布

モチーフ	N_1 & N_2	N_1^+ & N_2^+	$\hat{p}_1$ & $\hat{p}_2$
SDP I	189	67	0.356
	5969	5966	0.9993
SDP II	73	67	0.906
	6085	6082	0.9993
SDP III	71	67	0.932
	6087	6084	0.9993

表 2: 確率的決定述語の記述長

モチーフ	DL	PL	CL	$Total$
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

DL,PL,CL, and $Total$ はそれぞれデータの記述長、確率変数の記述長、確率的決定述語の記述長および総記述長を示す。

DL,PL,CL の和を求めることにより、確率的決定述語の総記述長 (TL) が求まる。

$$TL = DL + \lambda\{PL + CL\}$$

ここで λ は調整パラメタであり、本稿では 1 として扱う。MDL 基準では、この総記述長 (TL) がもっとも短い確率的決定述語を選ぶ。

4 実験結果

表 1 は、アミノ酸配列データバンク PIR (Protein Identification Resources) の R18.0 版に含まれる 6158 個の配列において、下記のモチーフパターンを含むか否かにより分類した結果である。

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'').

この分類結果より、具体的に確率的決定述語の記述長を求めた結果を表 2 に示す。表において、

表 3: クロス検定法によるミトコンドリアシトクロムCに対する予測エラーの平均値

	MDL 基準	最尤法
予測エラー平均値	0.0008	0.0013

DL は確率的決定述語を用いて PIR のアミノ酸配列を分類した時のアミノ酸配列全体の記述長を、PL は推定した確率変数の記述長を、CL は確率的決定述語を表現するクローズの複雑さを表す。ミトコンドリアシトクロムCの例では、配列モチーフのパターンとして、"CXXCH" だけでは単純すぎ、"CXXCH"and "GPXLXG"and "PGTKM" は複雑すぎ、"CXXCH"and "PGTKM" がこの3つの中では、MDL 基準の意味で一番尤もらしい分類規則であることを示している。

さらに、表3に MDL 基準を用いて配列モチーフを求めたときの予測エラーの平均値と確率的決定述語の複雑さ (PL+CL) を考慮せずにデータの記述長 (DL) だけを用いてモチーフを求める方法 (最尤法) を行なったときの予測誤差の平均値を示す。予測誤差の測定には、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モチーフに対し、残りの 10 分の 1 のデータを未知データとして与えて分類が成功したか否かを調べるクロス検定法を全ての組合せについて行ない、その平均値を求めた。

計算式を次に示す。

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

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

where $N = 6158$.

5 結論

遺伝子情報処理における MDL 基準の利用例を確率的決定述語を用いたモチーフ抽出を例に述べた。遺伝子情報は本質的にあいまいな情報を含んでおり、モチーフ抽出の抽出においては確率的な解析が不可欠である。このような解析手法の一つとして、統計的な裏付けをもち、かつ、計算機での扱いが容易な MDL 基準は極めて有効な手法といえよう。

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システム開発 (株) の山岸氏、小柳氏ならびに C&C システム研究所の疋田部員に感謝の意を表します。

参考文献

[DOE 88] (1988). *Mapping Our Genes, The Genome Projects: How Big, How Fast*, Congress of the United States, Office of Technology Assessment (1988).

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

[KY 90] 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.

[KNY 90] 小長谷, 新田, 山西,(1990).「遺伝子情報知識ベースシステムの構想について」, 情処人工知能研究会 73-9,pp.79-88.

[KY 91] Konagaya,A. & Yamanishi, K. (1991). A Stochastic Desicion Predicate: A Scheme to Represent Motifs, to appear in the AAAI Workshop of Classification and Pattern Recognition in Molecular Biology.

[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.

[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.

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