TM-1145

確率的モチーフ：現状と課題

小長谷　明彦（日電）

January, 1992

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

確率的モチーフ: 現状と課題

小長谷明彦

日本電気株式会社 C&C システム研究所
e-mail konagaya@csl.cl.nec.co.jp

遺伝子情報とタンパク質との対応関係を表す概念として提案した確率的モチーフの研究の現状と課題について述べる。確率的モチーフの抽出において、過学習を避け、予測精度の高い確率的モチーフを抽出するために記述長最小 (MDL) 基準を用いている。また、準最適解を効率良く発見する手段として「遺伝アルゴリズム」を採用している。

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が分子レベルで解明されつつある。このような遺伝子に関連する情報 (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。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、遺伝子情報の解析には単なる「計算機の利用」というレベルを越えた「分子生物学と情報処理技術の融合」が必須と思われる。筆者らは、このような観点に基づき、遺伝子情報解析のための技術の一つとして「確率的モチーフ (Stochastic Motif)」なるものを提案してきた [2, 3, 4, 8]。

確率的モチーフは、これまで分子生物学の領域において提案されていた「モチーフ」[1] を計算論的学習理論の立場から見直すことにより得られた概念である。モチーフのこれまでの解釈は特定のカテゴリーを代表する「配列のパターン」である。例えば、「CXXCH という配列パターンはシトクロム C のモチーフである」というように使われる。ここで、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。確かに、多くのシトクロム C に属するタンパク質のアミノ酸配列は CXXCH というパターンを持つ。しかしながら、全てのシトクロム C に属するタンパク質が CXXCH というパターンを持つわけではないし、シトクロム C でないタンパク質が CXXCH というパターンを持つこともある。すなわち、配列のパターンとタンパク質との関係は「確率的」な関係にならざる得ない。ここで、もし、モチーフというものが確率的な対応関係を示すものだとすれば、遺伝子情報からのモチーフ抽出は、「確率的規則の学習 [7]」として扱かうべきである。このとき次の問題が生じる。

確率的規則は学習しすぎると学習データに依存した規則となり、未知データに対する予測能力は却って低下する傾向がある。

これは、いわゆる「過学習」と呼ばれている現象である。過学習は十分なデータ数がない場合や、ノイズを含んでいるようなデータを用いて学習した場合に起きやすい。我々は、この過学習を避けるために記述長最小 (MDL) 基準を利用する。MDL 基準 [6] は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 [7]。また、実際に、予測精度の高い確率的モチーフが抽出できることをクロスバリデーション法を用いることにより、確認した [4]。本稿では、この確率的モチーフ研究の現状と課題について述べる。

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考案した確率的決定述語 [4] 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、4 節で適用例について、5 節でモチーフ抽出システムの簡単な紹介を、6 節で今後の課題について述べる。

2 確率的決定述語

確率的モチーフは遺伝子の配列情報からタンパク質のカテゴリーへの確率的な対応関係を表す。本節では、確率的モチーフの表現形式として考案した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。

$$motif(S, C_1) \quad (\text{with } p_1) \quad :- Q_1^{(1)} \wedge \cdots \wedge Q_{k_1}^{(1)}.$$

$motif(S, C_2)$ (with p_2) $:- Q_1^{(2)} \wedge \cdots \wedge Q_{k_2}^{(2)}$.

.....

.....

$motif(S, others)$ (with p_m).

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。

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

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

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。

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 [5])。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。

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 適用例

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'').

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

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、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。

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

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

ただし $N = 6158$.

5 モチーフ抽出システム

確率的モチーフの最適解を自動的に求めることは、組合せ的爆発を起こすため現在の計算機の能力では困難である。このため、現在、「遺伝アルゴリズム」と呼ばれる確率的探索アルゴリズムを用いて、(確率的決定述語で表現できる範囲で)準最適な確率的モチーフを求めるシステムを構築中である[9]。遺伝アルゴリズムでは、確率的決定述語を0、1のビット列にエンコードし、ビット列を確率的に変化させることにより、「良いビット列」を選択し、増殖させることでより良い確率的決定述語を求めてゆく。良さの基準としては、本稿で述べたMDL基準の記述長を用いる。本システムは第五世代計算機プロジェクトで開発した並列言語KL1で実装されており、現在Multi-PSI上で稼働中である。

6 課題

本手法をさらに発展させるための今後の課題としては以下のものがあげられる。

- 突然変異や実験誤差の扱い: 例えば、アミノ酸配列では、解析が困難なアミノ酸の対を表す記号が含まれている(例、B: アスパラギンとアスパラギン酸, Z: グルタミンとグルタミン酸)。確率的決定述語のOR-結合はこのような問題に対してある程度有効ではあるが、より厳密には、記述長の計算の見直しが必要である。

- カテゴリー間の階層性の取り扱い現在のMDL戦略はカテゴリー間の階層性を全く考慮していない。例えば、MDL戦略は“CXXCH”∧“PGTKM”の代わりに“PGTKM”だけをミトコンドリアシトクロムCの分類条件とする確率的決定述語を選ぶかもしれない。この場合シトクロムC全体で表れるパターン“CXXCH”が分類条件から外されてしまう。単なる分類効率を考えるだけでなく、分子生物学的な知見を反映させるためにはカテゴリー間の階層性を考慮したMDL戦略の確立が不可欠である。

7 結論

確率的モチーフの現状と課題について述べた。確率的モチーフは、遺伝子情報に見られるような確率的な対応関係をより正確に表現することが出来る。さらに、MDL基準を用いることによりより予測精度の高い確率的モチーフを抽出することが可能である。ただし、現在の抽出アルゴリズムでは、分類効率が優先されてしまうため、今後は、生物学的知見を反映できるようなモチーフ抽出システムの研究を進めてゆく予定である。

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の作成ならびにデータの収集をして頂いた遺伝子情報処理プロジェクトの皆様に感謝の意を表します。

参考文献

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

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] 小柳、山岸、小長谷,(1991). マルチPSIを利用したタンパク質の配列モチーフ抽出, ヒトゲノム計画と情報解析技術.