ICOT Technical Memorandom: TM-1235

TM-1235

GA の遺伝子情報処理への応用

小長谷　明彦（日電）

November, 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

GA の遺伝子情報処理への応用

小長谷明彦

NEC C&C システム研究所

本稿では遺伝子情報処理における遺伝的アルゴリズム (GA) 応用の一つとして、記述長最小 (MDL) 基準を用いた GA によるモチーフ抽出について述べる。モチーフ抽出のような分類学習問題に GA を適用する場合には分類規則 (モチーフ) の選択基準の設計が重要な意味を持つ。選択基準として MDL 基準を採用することにより、過剰適合を避け、安定的なモチーフが抽出できるとともに、GA の適応度関数に有効なバイアスがかかり、最適解への収束速度が早まることを示す。

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 [1]。遺伝子情報が蓄積されるにつれ、得られた遺伝子情報を解析するための遺伝子情報処理技術が求められている。遺伝子情報処理の課題としては、遺伝子解析向きデータベース構築、類似配列検索 (ホモロジー検索)、タンパク構造予測 (二次構造予測、三次構造予測)、共通パターン検索 (モチーフ抽出) 等、多岐に渡るが、本稿では、モチーフ抽出に焦点を当て、同問題に対する「遺伝的アルゴリズム」および「MDL 基準」の有効性を示す。

GA をモチーフ抽出のような分類学習問題に適用する際にもっとも重要なことは分類規則の選択基準の設計である。良く知られた有効な分類規則の選択基準の一つとして、分類規則の学習データに対する対数尤度 (log likelihood) がある [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、我々は 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) を提案した [3]。MDL-GA で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略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、同一世代数までの学習では最尤法 -GA に比べ、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 [4] 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[5]。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[6, 7]、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

MDL基準の考え方は、通信における符号理論に由来する。以下、その基本思想について簡単に紹介する。ある地点から別な地点にあるNビットの情報Aをできるだけ少いビット数で転送することを考える。情報Aをそのまま送ればNビット必要である。ここで、もし、情報Aを規則Rと例外Eで表現できるならば、同じ情報を規則Rと例外Eを符号化するのに必要なビット数で転送できることになる。ここで、規則Rとしてあまり単純な規則を用いると例外Eのビット数が増えてしまい、例外Eが少くなるように複雑な規則を利用した場合には規則Rのビット数が増えてしまう。すなわち、転送する総ビット数を減らすためには、規則Rを符号化するために必要なビット数と例外Eを符号化するために必要なビット数のトレードオフを考慮する必要があり、両者の和を最小化するような規則が望ましい規則となる。

MDL基準の考え方は、ノイズを含む学習データから確率規則の推定を行なうような場合にも適用できる[9]。この場合、確率規則が単純すぎると学習データへの適合性が低くなり、未知データに対する予測率は低下する。逆に、学習データへ適合した確率規則を求めようとすると、ノイズの影響まで含んだ非常に複雑な確率規則を学習してしまい、未知データに対する予測率はやはり低下する。したがって、この場合は、確率規則の複雑さと学習データへの適合性のトレードオフを考慮して確率規則を学習することが重要となる。

MDL基準を用いた確率規則の学習では、このトレードオフを次のようにして実現する。確率規則の複雑さを表現する手法として、確率規則の記述長、すなわち、確率規則を1、0のビット列にエンコードするために必要な長さを利用する。また、確率規則の学習データへの適合性を示す尺度として対数尤度を用いる。

確率的規則の表現法として確率的決定述語を用いた場合には、

$$\lambda\ (\text{確率的決定述語の記述長}) + \text{対数尤度}$$

を最小にする述語がMDL基準の観点から良い確率的決定述語と判断する[10, 11]。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[12.3]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[8]。ここで採用した単純遺伝的アルゴリズム（*Simple Genetic Algorithm:SGA*）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。すなわち、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、すなわち、探索点の集合を世代毎に更新し、適当な世代後において関数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、交叉、突然変異、選択の遺伝的操作を行なう。

6 モチーフ抽出システム

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の値すなわ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。ただし、探索の候補となる確率的決定述語として任意の表現法を許したのでは探索空間が莫大となり、実用的な時間内に良い確率的モチーフを探索することは困難となる。そこで、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.
```

すなわち、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。

参考までに表1に本システムで抽出した確率的モチーフとその記述長の一部を示す。各行においてタンパク質名の隣に抽出したパターンを示す。また、その下段に左から、全体の記述長（$\ell(T)$）、

表 1: 確率的モチーフ抽出結果の一部

Protein Category				Stochastic Motif Patterns				
$\ell(T)$(	$\ell(M)$,	$\ell(P)$,	$\ell(L)$)	E	N_1^+,	N_1	N_2^+,	N_2
Cytochrome-c				CXXCH				
309.544(	18.288,	10.564,	280.693)	140	137,	244	9386,	9389
Cytochrome p450				FXXGXXXC&CXGXXXA				
95.705(	40.868,	9.179,	45.658)	33	31,	35	9596,	9598
Trypsin				GWG&CXXDXG				
124.490(	34.253,	9.435,	80.802)	40	37,	50	9580,	9583
Pepsin				FXXXFD& VPXXXC				
80.802(	38.575,	8.700,	33.526)	19	17,	18	9613,	9615
Immunoglobulin V region				DXXXYXC				
692.147(	20.095,	10.871,	661.181)	268	237,	379	9223,	9254
Immunoglobulin C region				CXXXXFXP&FXPXXXXXXW&CLXXXXXP				
289.758(	63.079,	9.477,	217.203)	74	53,	53	9559,	9580
Globin				PXTXXXF&HGXXV& PXTXXXXXXF				
703.061(	59.915,	10.878,	632.268)	456	382,	383	9176,	9250

表 2: クロス検定法による予測エラーの評価

	MDL-GA	最尤法 -GA
正例エラー数	3	57
負例エラー数	96	0
合計	99	57

ホーン節の記述長($\ell(M)$)、確率パラメタの記述長($\ell(P)$)、対数尤度($\ell(L)$)、DBに登録されているタンパク質の個数(E)、第1クローズの正例数(N_1^+)、照合数(N_1^-)、第2クローズの正例数(N_2^+)、照合数(N_2^-)を示す。

表1にあるタンパク質の簡単な説明を下記に示す。シトクロムc (Cytochrome c)は呼吸鎖における電子伝達系に関わるタンパク質、シトクロムp450 (Cytochrome p450)はプロトヘムを含有し酸素添加反応を触媒するタンパク質、トリプシン(Trypsin)はすい臓から分泌されるプロテアーゼ、ペプシン(Pepsin)は胃に分泌される酸性プロテアーゼ、Immunoglobulin V regionは免疫グロブリンの可変領域、Immunoglobulin C regionは免疫グロブリンの固定領域、グロビン(globin)はヘムと共にヘモグロビンを構成するタンパク質である。

7 評価

表2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。また、一回のGAの試行は100世代までとし、これを100回繰り返し、その中で最も記述長長の短い確率的モチーフを採用した。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

遺伝子情報処理におけるGAの適用例として、タンパク質の配列からのモチーフ抽出を紹介した。モチーフ抽出の結果はモチーフの選択基準の設計に大きく依存しており、同一世代までの実行においては、できるだけ分離精度の高いモチーフを優先する最尤法-GAに比べ、モチーフの複雑さを考慮するMDL-GAの方が、より安定的なモチーフを抽出できることを示した。さらに、モチーフ抽出においてはモチーフの複雑さの比重を高めるほど収束速度が高まることを確認した。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。

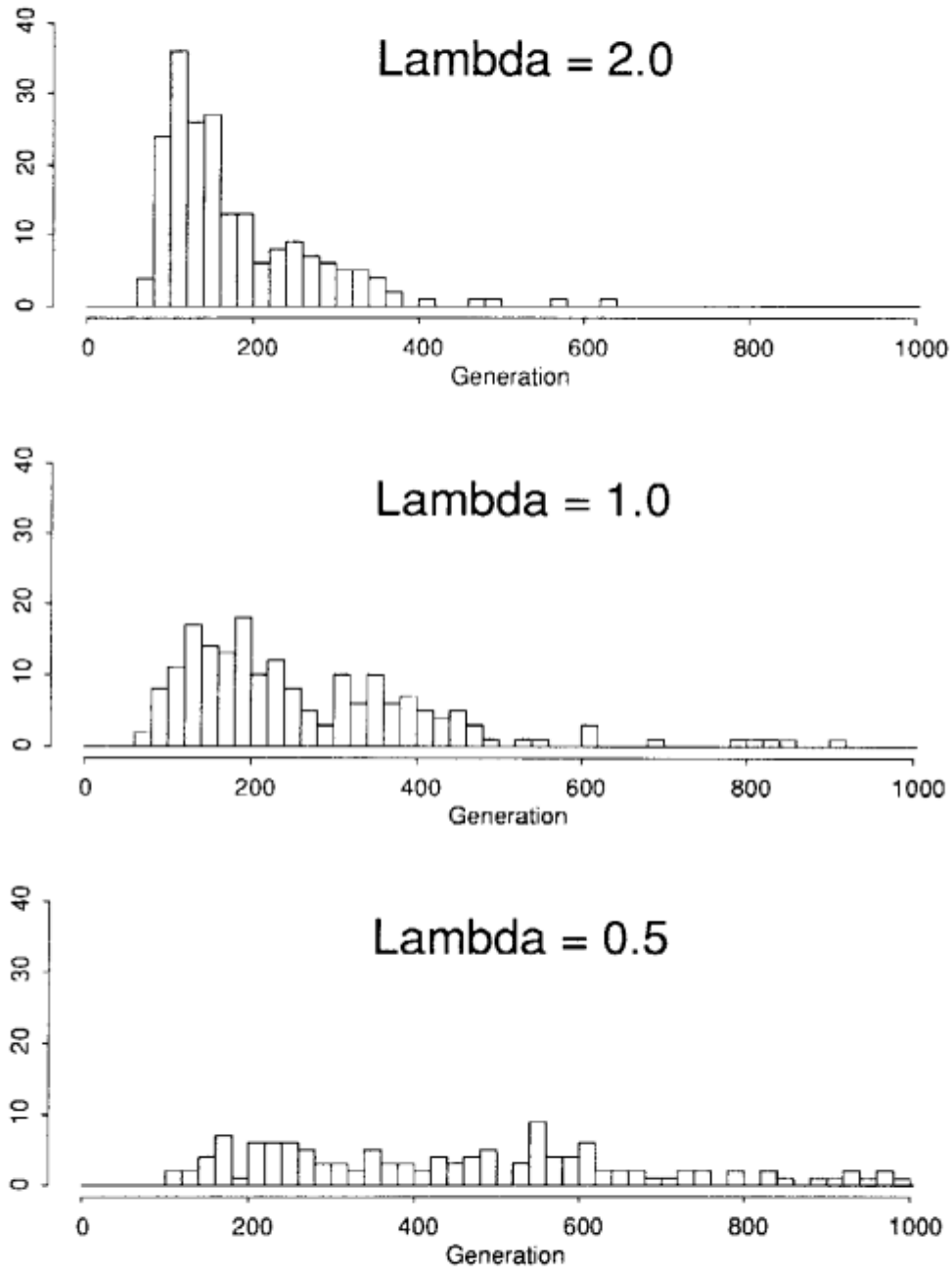

図 1: 最適解が求まるまでの世代数の分布に対する調節パラメタの影響

参考文献

[1] (1988). Mapping Our Genes, The Genome Projects: How Big, How Fast, *Congress of the United States, Office of Technology Assessment* .

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

[3] Konagaya,A. & Kondo, Y. (1993). Stochastic Motif Extraction using a Genetic Algorithm with the MDL Principle, to appear in *Hawaii International Conference on System Sciences.*

[4] 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.*

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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記述長の和で行なう。対数尤度の計算法は以下の通りである。

$$\ell(L) = -\log \prod_{j=1}^{m} p_j^{N_j^+} (1-p_j)^{N_j - N_j^+} \qquad (1)$$

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

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、$\tilde{p}_i = N_i^+/N_i$ であり、$\hat{p}_i$ は p_i 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。

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

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

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。

$$\ell(P) = \sum_{i=1}^{m} \frac{\log N_i}{2} \qquad (3)$$

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。

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

ここで $L_l^j(i)$ と $X_l^j(i)$ 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 [13]、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。