ICOT Technical Memorandom: TM-1165

TM-1165

遺伝的アルゴリズムの遺伝子情報処理への適用について

小長谷　明彦、近藤　浩康（日電）

March, 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

遺伝的アルゴリズムの動向と課題 特別講演会 (計測自動制御学会第4回知能工学部会,1992)

遺伝的アルゴリズムの遺伝子情報処理への適用について

小長谷明彦　　近藤浩康

日本電気株式会社 C&C システム研究所

アブストラクト　本稿では記述長最小 (MDL) 基準を用いた遺伝的アルゴリズムを提案し、遺伝子情報処理の研究課題の一つであるモチーフ抽出への適用結果について述べる。遺伝的アルゴリズムは確率的探索を行なうためモチーフ抽出のような組合せ分類学習問題を高速に解くのに適している。しかしながら、組合せ分類学習問題への単純な遺伝的アルゴリズムの適用は与えられた学習データへの過剰適合を引き起こしやすい。この問題を解決するために、モデルのデータへの整合性とモデルの複雑さを考慮して最良のモデルを推定する MDL 基準を導入した。MDL 基準を遺伝的アルゴリズムの適応度の計算に採用することにより、モデルとデータの整合性だけを考慮する最尤法に比べ、過剰適合を避け、より安定なモチーフを抽出できることを実データを用いて示す。

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新たな情報処理技術が求められている。このような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、遺伝子情報読み取りシステムの自動化、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。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。

モチーフ抽出は、共通の性質を持つタンパク質の塩基配列あるいはアミノ酸配列を解析することにより、そのタンパク質を特徴付けている配列パターン (モチーフ) を見つける操作である。もし、このようなモチーフが見つかれば、そのモチーフを含むか否かでタンパク質の分類を行なうことができる。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 [2]。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。このようなモチーフの抽出は計算論的学習理論における「分類学習」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 [3, 4]、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。

第一の問題は、タンパク質の機能、構造を定める配列の並びは一意ではないという生化学的な多様性に由来している。例えば、シトクロムｃというタンパク質が全て CXXCH というモチーフを含む訳ではないし、逆に、CXXCH というモチーフを持つタンパク質が全てシトクロムｃに分類される訳でもない。すなわち、モチーフとタンパク質の機能、構造との関係は必ずしも１対１ではなく、確率的な対応関係と考えるべきである。我々はこのような観点から、モチーフを確率的な規則 (確率的モチーフ) として表現する方法を提案した [5, 7]。確率的モチーフを用いれば、先の例は、「もし、アミノ酸配列が CXXCH というアミノ酸のパターンを含めばそれは確率４／５でシトクロムｃである。」と表現できる。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という特徴を持つ。

第二の問題は、タンパク質の配列の長さが 100 ～ 300 と比較的長いこと、扱うアミノ酸の種類が 20 種と多いことに起因する。単純に５文字のパターンを考えるだけでも、その種類は数百万通りにもなり、実際にはこのパターン同士の組合せを考慮しなくてはならない。また、本稿では言及しないが、アミノ酸はさらに疎水性、極性、大きさなどの属性により分類することができ、これらの属性まで考慮するとパターン候補の総数は天文学的数字となる。このような組合せ問題を解くために、我々は学習アルゴリズムとして、遺伝的アルゴリズムを採用した。遺伝的アルゴリズムでは確率的

探索を行なうため探索時間を大幅に減らすことが期待できる。実際、我々の経験では、組合せ数の対数オーダに近い時間内に満足すべき解を見つけることができている。

第三の問題は、分子生物学者の間でも認識している人はまだ小数であるが、計算機で遺伝子情報を解析する上でエラーを扱う機構が不可欠であることを示している。この問題の要因は大きく分けて2つある。一つはデータそのもののエラーであり、もう一つはデータのサンプリングの偏りである。驚くべきことであるが、現在流通しているアミノ酸配列データベースや塩基配列データベースにはかなりのエラーが含まれており、情報抽出を行なう場合には十分な注意が必要である。これは、もともと、生物実験では100パーセント完全なデータを得ることが非常に難しいということもあるが、それ以上に入力エラーやデータ作成者の思い入れあるいは解釈の違いの影響が大きい。もうひとつのサンプリングの偏りは、実験対象がヒト等の特に重要な生物あるいは大腸菌のような実験しやすい生物に偏っているということに起因する。現在、同一のタンパク質について調べられている生物の種類は多くて高々数百種であり、これは数百万種といわれている全生物種に比べるとあまりに少ない。このような問題に対処するためには、与えられたデータを信用し過ぎないという態度が必要である。我々は、このために、記述長最小(MDL)基準[8, 9]を導入した。MDL基準はモデルとデータとの整合性に、さらに、モデルの複雑さを加味して最良のモデルを推定する基準であり、与えられたデータに偏った学習(過剰適合)を避けることができるという性質を持つ[10]。我々のアプローチでは、遺伝的アルゴリズムの「適応度」として、MDL基準の「記述長」を採用した。これにより、過剰適合を避け、より安定的な確率的モチーフの抽出を実現している。

本稿の構成を以下に示す。はじめに、2節においてモチーフの例を示し、3節で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[6]について紹介する。次に、4節で、MDL基準について、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

モチーフは、遺伝子の塩基配列・蛋白質のアミノ酸配列において、共通の祖先あるいは機能を持つ遺伝子同士・蛋白質同士に共通して見い出すことができる、共通の塩基配列パターン・アミノ酸配列パターンであり、その重要性の故に進化的に保存されるものである。

シトクロムcに見い出されるモチーフ *"CXXCH"* を例に説明する。モチーフ *"CXXCH"* のそれぞれ1文字は一つのアミノ酸に対応し、特に X は任意のアミノ酸に対応する。モチーフ *"CXXCH"* は、長さ5のアミノ酸配列パターンで、1番目のアミノ酸がシステイン(C)、2及び3番目のアミノ酸は任意のアミノ酸、4番目のアミノ酸がシステイン(C)、5番目のアミノ酸がヒスチジン(H)であるものを表している。シトクロムcは呼吸鎖の電子伝達に関わる酵素であり、ヘム鉄を含有していることにより酸化還元反応を行なうことができる。この重要な役割を果たすヘム鉄と共有結合している部位がモチーフ *"CXXCH"* の2つのシステイン(C)であり、その故にこの部位は進化的に保存されている。

3 確率的決定述語

モチーフは進化的に保存されるが、絶対的に保存されるわけではない、またある配列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。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,mitochondria_cytochrome_c)
   with 129/225
   :- contain("CXXCH",S).
motif(S,others) with 8081/8084.
```

この表現の意味は、S が *"CXXCH"* と一致する部位を含めば確率 $\frac{129}{225}$ で S はミトコンドリアシトクロムcであり、そうでなければ、確率 $\frac{8081}{8084}$ で *others*(シトクロムc以外のタンパク質)である。

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、不確実性を含む学習セット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セット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、

> 確率的決定述語の記述長 + 不確実性の記述長

を最小にするものが良いと判断される。確率的決定述語の記述長はモチーフの表現が複雑になるほど値が大きくなり、不確実性の記述長はモチーフの表現がデータからずれているほど値が大きくなる。それゆえ、与えられたデータに対する適合度とモチーフ表現の複雑さとのトレードオフを考慮したモチーフ表現を良い表現とする基準である。各記述長の計算法の詳細については文献[5, 6]を参照されたい。

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[11, 12]。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0 1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0.6突然変異確率は0.01に設定した。

7 抽出結果

遺伝的アルゴリズムを用いたモチーフ抽出を、蛋白質データベース NBRF PIR29.0 (8309 エントリ) に対して次の7つの蛋白質について行なった。cytochrome c は、呼吸鎖における電子伝達に関わるシトクロムc、cytochrome p450 は、プロトヘムを含有し酸素添加反応を触媒するシトクロム p-450、pepsin は、胃に分泌される酸性プロテアーゼであるペプシン、trypsin は、すい臓などで生合成されるプロテアーゼであるトリプシン、globin は、ヘモグロビンを脱ヘムして得られるアポ蛋白質、immunoglobulin C region は、免疫グロブリンの固定領域、immunoglobulin V region は、免疫グロブリンの可変領域である。

表 1: クロス検定法によるシトクロム C に対する予測エラー確率

	MDL 基準	最尤法
見落しエラー確率	3/71703	57/71703
誤分類エラー確率	96/71703	0/71703
合計	99/71703	57/71703

付録に示す一連の表は、上記の各蛋白質に対してモチーフ抽出を行なった結果である。同表において、データベース登録配列数は、蛋白質データベース中に含まれている対象蛋白質の数、対象配列数は比較対象となったエントリの数、照合配列数はパターンが一致しているエントリの数、正例数は照合配列数の内正しかったものの数、CL はモチーフの複雑さを表す記述長、PL は確率パラメタの記述長で CL と PL の和が確率的決定述語の記述長であり、DL はモチーフの正確さを表す記述長である。

これらの結果は必ずしも生物学的保存部位ではないが、保存性が高く、かつ、安定的なモチーフの確率的決定述語表現である。

さらに、表1に MDL 基準を用いてモチーフを求めたときの予測エラーの発生確率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エラーの発生確率を示す。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。シトクロム c の場合には、合計のエラー発生確率は最尤法の方が少ないが、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と判定した確率(誤分類エラー確率)はMDL基準の方が最尤法より高い。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。これらをさらに分類する方式については現在検討中である。一方、シトクロムcの配列をシトクロムcでないと判定した確率(見落しエラー確率)は最尤法の方がMDL基準よりも高い。これは、逆に、最尤法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基準を用いた場合には、10組中、9組までが同じモチーフCXXCHを抽出しているのに対し、最尤法では抽出したモチーフ10組全てが異なっており、過剰適合の現象が起きていることが確認できる。

以上より、MDL基準は誤分類エラー確率に関しては改善の余地はあるものの、スクリーニングに必要な見落しエラー確率を大幅に削減しており、配列モチーフ抽出により適した判定基準といえる。

8 結論

遺伝的アルゴリズムを用いたモチーフ抽出について述べた。遺伝的アルゴリズムは確率的探索を行なうことにより組合せ問題を効率良く解くことが可能である。また、MDL基準の記述長を遺伝的アルゴリズムの適応度として採用することにより、過剰適合を避け、より安定的な確率的モチーフを抽出できることを示した。

本稿で述べた手法は非常に一般的であり、多くの組合せ分類問題に適用可能と思われる。また、本稿では言及しなかったが、遺伝的アルゴリズムは並列化も容易であり、現在、本システムは並列マシン上で稼働していることを合わせて報告しておく[13]。

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.*

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

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

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

[10] 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.

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

[12] 小長谷,(1991). 確率的モチーフ: 現状と課題, 第2回公開ワークショップ ヒトゲノム計画と情報処理技術, (pp.66-69).

[13] 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, 第2回公開ワークショップ ヒトゲノム計画と情報処理技術, (pp.70-73).

付録

1. cytochrome c （データベース登録配列数 132）

モチーフ	対象配列数	照合配列数	正例数
CXXCH	8309	225	129
others 8084	8084	8081	

記述長 286.89 (CL = 16.28, PL = 10.40, DL = 260.21)

2. cytochrome p450 （データベース登録配列数 21）

モチーフ	対象配列数	照合配列数	正例数
GXXXCXG and PXXFXP	8309	37	20
others	8272	8272	8271

記述長 97.22 (CL = 36.38, PL = 9.11, DL = 97.22)

3. pepsin （データベース登録配列数 16）

モチーフ	対象配列数	照合配列数	正例数
FDXG and SXXXWV	8309	14	14
others	8295	8295	8293

記述長 68.86 (CL = 33.25, PL = 8.41, DL = 27.19)

4. trypsin （データベース登録配列数 38）

モチーフ	対象配列数	照合配列数	正例数
GWG and CXXDXG	8309	48	36
others	8261	8261	8259

記述長 106.69 (CL = 31.25, PL = 9.30, DL = 66.14)

5. globin （データベース登録配列数 430）

モチーフ	対象配列数	照合配列数	正例数
PXTXXXF and HGXXV	8309	403	370
others	7906	7906	7846

記述長 719.70 (CL = 35.38, PL = 10.80, DL = 673.51)

6. immunoglobulin C region （データベース登録配列数 74）

モチーフ	対象配列数	照合配列数	正例数
CXVXH	8309	200	61
others	8109	8109	8096

記述長 343.56 (CL = 16.29, PL = 10.32, DL = 316.96)

7. immunoglobulin V region （データベース登録配列数 268）

モチーフ	対象配列数	照合配列数	正例数
DXXXYXC	8309	363	237
others	7946	7946	7915

記述長 659.66 (CL = 18.10, PL = 10.73, DL = 630.84)