TM-1146

マルチPSIを利用したタンパク質の配列モチーフ抽出

小柳　敏、山岸　晃一、
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

マルチＰＳＩを利用した
タンパク質の配列モチーフ抽出

小柳　敏、山岸晃一、小長谷明彦*
日本電気技術情報システム開発（株）、*日本電気（株）Ｃ＆Ｃシステム研究所

概要
確率的探索アルゴリズムである遺伝アルゴリズムと記述長最小（ＭＤＬ）基準を用いた、タンパク質の配列モチーフ抽出プログラムを、第五世代コンピュータプロジェクトで開発した並列推論マシン「マルチＰＳＩ」上に開発し、そのプログラムを使用して、実際のタンパク質の配列モチーフ抽出を行った。本稿では、タンパク質の配列モチーフ抽出の結果と並列化による台数効果について報告する。

１　はじめに

近年、遺伝子工学の発展により、様々な生物のＤＮＡ配列、アミノ酸配列などの遺伝子情報が分子レベルで収集されつつあり、これとともに、計算機を利用した遺伝子情報解析技術への期待が高まっている。このような遺伝子情報解析技術の一つとして、配列情報からの規則（モチーフ）抽出がある。我々は、これまで、遺伝子情報が生物種の多用性に由来するノイズを含むことに着目し、モチーフ抽出を確率的規則の学習問題として定式化し[1]、遺伝子情報に向いた確率的規則の表現形式として確率的決定述語を提案し[2]、より良い確率的決定述語の選択基準として記述長最小（ＭＤＬ）基準を利用するモチーフ抽出法を提案してきた[3]。ＭＤＬ基準を用いたモチーフ抽出法では、モチーフの良さをモチーフを表現する確率的決定述語の記述長とモチーフの正確さを表す記述長（確率的決定述語の対数尤度）の和（少ないほど良い）で表す。したがって、与えられた配列情報について全ての確率決定述語の記述長を計算すれば、原理的には計算機による自動抽出を行うことができる。

モチーフの自動抽出において、抽出されたモチーフの良否は考慮すべき確率的決定述語の集合、すなわち、仮説空間の設定の仕方と、仮説空間内での探索アルゴリズムの両方に依存する。モチーフ抽出の場合、仮説空間を事前に絞り込むことが困難なこと、必ずしも最適解を求める必要がないことから仮説空間を十分大きくとり、確率的探索アルゴリズムにより準最適な確率的決定述語を求めるという方針をとった。また、確率的探索アルゴリズムとして遺伝アルゴリズム[4]を採用した。

本稿では、遺伝アルゴリズムとＭＤＬ基準を用いた配列モチーフ抽出プログラムを並列推論マシン「マルチＰＳＩ」上に開発し、実際のタンパク質の配列モチーフ抽出を行った結果を報告する。以下、はじめに、２節において遺伝アルゴリズムの基本的な考え方を紹介し、３節で配列モチーフ抽出方法、４節で配列モチーフ抽出結果、５節で「マルチＰＳＩ」上でのプログラムの並列化による台数効果について述べる。

２　遺伝アルゴリズム

遺伝アルゴリズムは、生物の進化の過程をモデルとして発案された確率的探索アルゴリズムの一つである。その特徴は、仮説空間内の候補に０と１のビット列を対応させ、このビット列を各個体の「染色体」と見立てて選択、交差、突然変異などの遺伝子操作を加え、より良い個体へ進化させることにある。例として、アミノ酸配列からのシトクロムＣの配列モチーフ抽出を考える。

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

上記の確率的決定述語は、アミノ酸配列 Sがパターン "CXXCH"（ Xは任意のアミノ酸と照合する）を含めば確率p1でシトクロムＣであり、そうでなければ確率p2で、その他である（確率1-p2でシトクロムＣである）という確率的モチーフを表す。今、このような確率的決定述語のパターンの候補として、

```
"CXXCH","GPXLXG","PGTKW"
```

の３つがあるとする。この場合、それぞれのパターンの組み合わせは、３文字のビット列からなる「染色体」で表現できる。ここで、各ビットが１であるとき、その対応するパターンが含まれることにすると、遺伝アルゴリズムによる解の探索は以下の手順で行われる。

１）初期化

ランダムに「染色体」を選ぶ。ここでは４つ選ぶ。
100, 011, 010, 110

２）選択

「染色体」を適応度に応じて選択する。100,011,010,110の適応度が 2,1,0,1 ならば以下のようになる。
100, 100, 011, 110

３）交差

ある確率で、２つの「染色体」の間でｎビットを交換する。例えば２、３番目の「染色体」の２および３ビット目を交換する。
100, 111, 000, 110

４）突然変異

ある確率で、ビットをランダムに反転させる。例えば３番目の「染色体」の３ビット目を反転させる。
100, 111, 001, 110

以下２から４の処理をｎ回繰り返し、ｎ世代目の「染色体」を得る。

３　モチーフ抽出方法

モチーフ抽出は、２節で述べたように「染色体」の各ビットに一つのモチーフの候補となるパターンを対応させて行った。ただし、ここでモチーフの候補となるパターンの選択方法が問題となる。計算量を減らすために、今回は次のような方法を使ってモチーフの候補となるパターンを選択した。

モチーフ抽出を行うタンパク質について、次の６つで表されるパターンの出現頻度を調べる。

ABC, AXBC, ABXC, AXXBC, AXBXC, ABXXC

ここで、A,B,Cは、ある一つのアミノ酸を表し、Xは任意のアミノ酸を表す。この 48000通りのパターンの中で、出現頻度の高いパターンから順番に30個を選択した。また、今回は遺伝アルゴリズムのパラメータを世代数 50 、個体数 64 、交差確率 0.6、突然変異確率 0.01 に設定した。

４　モチーフ抽出結果

今回は、タンパク質配列データベースPIR 28.0版（7967エントリー）を使用し、次の７つのタンパク質についてモチーフ抽出を行った。

a) cytochrome c
b) cytochrome p450
c) pepsin
d) trypsin
e) globin
f) immunoglobulin C region
g) immunoglobulin V region

以下に、モチーフ抽出を行った結果のある一つの例を示す。ここで、対象配列数は、検索対象となった配列の数、照合配列数は、モチーフを含む配列の数、正例数は、モチーフを含む配列の中で検索対象のタンパク質であった配列の数である。例えば、cytochrome cでは、検索の対象となった7967の配列の中で、モチーフ PXLXG & PXXKM & GXKM を含む配列は90個あり、その中でcytochrome c である配列が 90 個あることを表す。

また、CLはモチーフの複雑さを表す記述長、PLは確率変数の記述長で、CLとPLの和が確率的決定述語の記述長である。DLはモチーフの正確さを表す記述長である。

a) cytochrome c　（データベース登録配列数 ... 129）

モチーフ	対象配列数	照合配列数	正例数
PXLXG & PXXKM & GXKM	7967	90	90
others	7877	7877	7838

記述長　415.066 (CL = 50.541, PL = 9.718, DL = 354.807)

b) cytochrome p450 （データベース登録配列数 ... 22）

モチーフ	対象配列数	照合配列数	正例数
PXXFL & RXEXF & PGXG	7967	20	15
others	7947	7947	7940

記述長 156.674 (CL = 50.541, PL = 8.639, DL = 97.494)

c) 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
YXXFD & DTG	7967	25	14
others	7942	7942	7940

記述長 90.86 (CL = 30.253, PL = 8.800, DL = 51.809)

d) trypsin （データベース登録配列数 ... 64）

モチーフ	対象配列数	照合配列数	正例数
AXHC & DSXXP	7967	54	52
others	7913	7913	7901

記述長 183.917 (CL = 32.253, PL = 9.352, DL = 142.312)

e) globin （データベース登録配列数 ... 429）

モチーフ	対象配列数	照合配列数	正例数
VDP & DPXN & LXVXP & VXPXN	7967	310	307
others	7657	7657	7535

記述長 1003.495 (CL = 65.167, PL = 10.589, DL = 927.739)

f) immunoglobulin C region （データベース登録配列数 ... 69）

モチーフ	対象配列数	照合配列数	正例数
CXVXH	7967	183	53
others	7784	7784	7768

記述長 351.321 (CL = 16.288, PL = 10.221, DL = 324.812)

g) immunoglobulin V region （データベース登録配列数 ... 267）

モチーフ	対象配列数	照合配列数	正例数
AXYXC & DXAXY	7967	215	199
others	7752	7752	7684

記述長 688.424 (CL = 33.575, PL = 10.334, DL = 644.525)

遺伝アルゴリズムは必ずしもそのタンパク質を分類する最適なモチーフを見つけるとは限らない。「染色体」の初期値や交差確率、突然変異確率を変えて実行すると、今回得られたモチーフよりもＭＤＬ基準の意味で最適なモチーフが発見できる可能性がある。また、生物学的見地からは、今回求めたモチーフは、多重アラインメント等により発見できる保存部位とは必ずしも一致していない。これは、今回用いたモチーフ抽出アルゴリズムが分類効率の向上だけを目的としているためであり、現在、生物学的知見を反映させることができるような拡張を検討中である。

5　並列化による台数効果

配列モチーフ抽出プログラムの遺伝アルゴリズムの部分を並列に実行した時の実行速度を計測し、並列化による台数効果について調査した。プロセッサの台数が１、２、４、８台について実行速度を計測した。遺伝アルゴリズムのパラメータは、世代数 50 、個体数 64 、交差確率 0.6、突然変異確率 0.01 に設定した。負荷分散の方法は、全体の個体数をプロセッサの台数で等分した数の個体を各プロセッサに割り当てるようにした。今、全体の個体数は 64 なので、２台のプロセッサで実行するときは、各プロセッサにそれぞれ 32 個の個体が割り当てられる。また、各世代ごとにプロセッサ間でランダムに選択した個体の交換を行っている。

以下に、計測の結果を示す。

プロセッサ数	個体数／プロセッサ数	台数効果
1	64	1.000
2	32	1.775
4	16	2.871
8	8	4.086

この結果から、個体数を固定した場合には、プロセッサ１台で実行した時よりも８台で実行した時の方が約４倍速くなることがわかる。より正確な解を求めるためには、総個体数を数千から数万のオーダーにする必要があるため、プロセッサ数をさらに増やしても台数効果が得られることが期待できる。

6　まとめ

遺伝アルゴリズムを使った、タンパク質の配列モチーフ抽出プログラムを「マルチＰＳＩ」上に開発し、そのプログラムを使って実際のタンパク質の配列モチーフ抽出を行ってみた。得られたモチーフは生物学的な見地からは満足のいくものではなかったが、この方法で配列モチーフ抽出が可能なことが確かめられた。また、並列に実行したときの台数効果もかなりあることが確かめられた。

今後、抽出されたモチーフの生物学的な正当性の評価と並列化による効果の評価、並列化の効果を高めるためのプログラムの改善を行っていく。

謝辞

本研究は第五世代コンピュータプロジェクトの一環として行われたものである。本研究の機会を与えて下さったＩＣＯＴの新田室長に深謝致します。また、本研究をサポートして頂いた遺伝子情報処理プロジェクト関係者に感謝致します。

参考文献

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

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

[3] 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.

[4] Goldberg,D.E.(1989). Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley.