1E2-3

プロモータモデリングのための共生進化に基づくHMM生成手法

HMM Generation for Promoter Modeling based on Symbiotic Evolution

大谷 紀子[*1] 新田 克己[*2]
Noriko Otani Katsumi Nitta

[*1]武蔵工業大学環境情報学部 [*2]東京工業大学総合理工学研究科
Musashi Institute of Technology Tokyo Institute of Technology

This paper describes how to generate HMM for promoter modeling using symbiotic evolution that is a kind of GA and can keep diversity in population. We focus on mammalian histone promoters that play important roles in chromosomal functions. Histones are broadly divided into five classes, namely H1, H2A, H2B, H3 and H4. A motif is a nucleotide or amino-acid sequence pattern that is widespread and has a biological significance. We aim to unveil the detailed patterns of motifs in promoter region of all five histone groups and generate HMM that is possible to discriminate them accurately.

1. はじめに

DNA 配列に潜むさまざまな生命現象に関する情報を抽出するために，計算機科学の手法を用いた DNA 配列解析の研究が盛んに行なわれている．抽出した情報の表現技法としては，DNA 配列の多様性の記述が可能な隠れマルコフモデル (hidden Markov model; HMM) が広く利用されており，HMM のトポロジーとパラメータの決定には，遺伝的アルゴリズム (Genetic Algorithm; GA) や遺伝的プログラミング (Genetic Programming; GP) 等を用いる手法が提案されている．

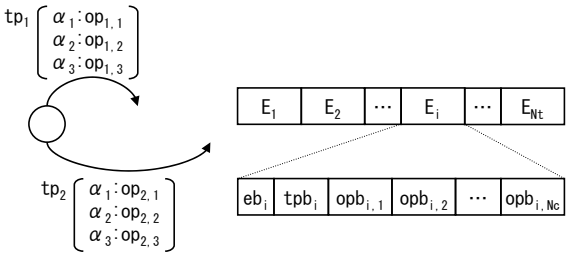
本研究では，GA の中でも多様な解候補からの探索を特長とする共生進化 (symbiotic evolution) に基づいて HMM を生成する手法を提案する．処理対象は，染色体の機能において重要な役割を果たす塩基性たんぱく質ヒストンのプロモータ領域とする．ヒストンは H1，H2A，H2B，H3，H4 の 5 種類に大別される．共通した特徴を持つ部分配列はモチーフと呼ばれ，複数のモチーフの出現パターンが DNA 配列の機能に大きく関与するとされている．各種のヒストンにおけるモチーフの出現パターンの特徴を示すとともに，ヒストンの種類を正確に識別する HMM の生成を目指す．

2. HMM 生成のための共生進化

共生進化は，Moriarty らが提案した GA の 1 手法である [Moriarty 96]．共生進化の特徴は，部分解を個体とする集団と，部分解の組合せを個体とする全体解集団を保持し，両集団を並行して進化させる点にある．部分解集団では解の部分的評価を行ない，最適解に含まれ得る多様な部分解を生成する．部分解のより良い組合せを全体解集団で学習することで，1 集団を進化させる GA よりも多様な解候補からの探索を行なうことができる．帰納論理プログラミングや決定木生成への適用手法が提案されており，有用性が確認されている [大谷 02, 大谷 04]．以下，HMM を生成するための共生進化の手法について述べる．

2.1 部分解 STATE

HMM の 1 つの状態を表すノードと，その状態からの遷移を表すエッジからなる部分を共生進化における部分解とし，STATE と呼ぶ．STATE の表現型の例を図 1(a) に示す．1 つのノードに結合されている 2 本のエッジには，それぞれ状態遷移確率 tp_i とシンボル α_1〜α_3 の出力確率 $op_{i,1}$〜$op_{i,3}$ が付与されている．1 つの状態からの遷移先を最大 N_t 個とするとき，STATE の遺伝子型は図 1(b) に示すように，N_t 個のエッジを表すビット列 E_1〜E_{N_t} から構成される．E_i は，当該エッジの存在を表すビット eb_i，遷移比の値 tpb_i，および出力シンボル α_1〜α_{N_c} の出力比の値 $opb_{i,1}$〜opb_{i,N_c} からなり，各比の値は長さ N_b のビット列で表現される．遺伝子型から表現型に変換する際，eb_i が 1 のときのみ E_i に相当するエッジが作成される．遷移確率 tp_i，出力確率 $op_{i,1}$〜op_{i,N_c} は式 (1)，式 (2) により求められる．



(a) 表現型 (b) 遺伝子型

図 1: STATE の例

$$tp_i = \frac{tpb_i}{\sum_{k=1}^{N_t} tpb_k} \quad (1)$$

$$op_{i,j} = \frac{opb_{i,j}}{\sum_{k=1}^{N_c} opb_{i,k}} \quad (2)$$

STATE の染色体を新たに生成する場合は，各遺伝子をランダムに設定し，遷移比の値の合計，または出力比の値の合計が 0 になる場合は修正する．P_s 個の STATE を生成し，初期の STATE 集団とする．

2.2 全体解 HMM

全体解である HMM は，複数の STATE の組合せで表現される．HMM の生成手順を図 2 に示す．P_h 個の HMM を生成し，初期の HMM 集団とする．複数の HMM から参照される

連絡先: 大谷紀子，武蔵工業大学環境情報学部
〒 224-0015 横浜市都筑区牛久保西 3-3-1，045-910-2938
E-mail: otani@yc.musashi-tech.ac.jp

Step.1	2〜N_s の範囲でランダムに状態数を設定する.
Step.2	設定された状態数の分だけ STATE 集団から無作為に個体を選択する.
Step.3	選択した STATE の遷移確率と出力確率を持つノードを生成する.
Step.4	各エッジの遷移先をランダムに設定する.
Step.5	終了状態に遷移する状態がない場合は，1 つのエッジをランダムに選択して，遷移先を終了状態とする.
Step.6	他の状態から遷移してこない状態を削除する.

図 2: HMM 生成手順

STATE や，いずれの HMM からも参照されない STATE も存在し得る.

2.3 適応度

各個体の評価の際は，訓練例として用意されたシンボル列の出力可否を Viterbi アルゴリズムにより判定する．訓練例集合は，HMM で出力されるべきシンボル列集合 D_p と，出力されるべきでないシンボル列集合 D_n の和集合であり，再現率と適合率を用いて HMM の適応度を算出する．再現率は D_p において出力可能な要素が占める割合を表し，適合率は出力可能な訓練例において D_p の要素が占める割合を表す．HMM の個体 H で出力可能な訓練例の個数を $tfp(H)$，D_p の要素数を $|D_p|$，H で出力可能な訓練例のうち D_p の要素であるものの個数を $tp(H)$ とすると，H の再現率 $rec(H)$，適合率 $pre(H)$，および適応度 $hfit(H)$ は次式で算出される.

$$pre(H) = \frac{tp(H)}{tfp(H)} \tag{3}$$

$$rec(H) = \frac{tp(H)}{|D_p|} \tag{4}$$

$$hfit(H) = \frac{2 \cdot pre(H) \cdot rec(H)}{pre(H) + rec(H)} \times 100 + psum(H) \tag{5}$$

$hfit(H)$ は，F 値と 100 倍と重みつき確率合計 $psum(H)$ の和である．$psum(H)$ は，H における各シンボル列の出力確率を考慮するための指標である．シンボル列 d_k の最適状態系列における出力確率を $p(H, d_k)$ と表すと，$psum(H)$ は次式で求められる.

$$psum(H) = \sum_{d_k \in D_p} p(H, d_k) - \sum_{d_k \in D_n} p(H, d_k) \tag{6}$$

STATE の適応度は，当該個体を参照する HMM のうち，最も適応度の高い HMM の適応度とする.

2.4 世代交代

STATE 集団の世代交代では，[Moriarty 96] と同様にして，P_s 個の個体のうち上位半数をそのまま次世代に残す．下位半数の個体は，上位四半数から選んだ 2 つの個体を親として交叉を行ない，生成された 2 つの子のいずれかと，2 つの親のいずれかで置き換える．交叉により遷移比の値の合計や出力比の値の合計が 0 になった場合は修正する．すべての個体の遺伝子に対して確率 p_m で突然変異を発生させ，次世代の個体とする.

HMM 集団の世代交代モデルとしては，[佐藤 97] で提案されている MGG(Minimal Generation Gap) モデルを採用する．MGG モデルは，局所解収束の回避と進化的停滞の抑制を意図して考案されたモデルである．集団からランダムに非復元抽出された 2 個体を親として子を生成し，親と子の個体のうち，最良個体およびルーレット選択で選ばれた 1 個体の計 2 個体を次世代に残す.

Step.1	STATE の進化
Step.2	HMM 集団から親を選択
Step.3	子を生成
Step.4	子の HMM を評価
Step.5	STATE の評価
Step.6	次世代に残す個体を選択

図 3: 一世代の処理手順

表 1: パラメータ

パラメータ	値
突然変異確率 p_m	0.01
STATE 集団の個体数 P_s	400
HMM 集団の個体数 P_h	400
世代交代回数 G	10000
最大状態数 N_s	50
1 つの状態からの最大遷移数 N_t	3
確率表現ビット数 N_b	2

生成される子は，以下の 4 種類の HMMC_1〜C_4 である.

C_1： P_1 をベースとして，P_1 と P_2 を交叉した HMM

C_2： P_2 をベースとして，P_1 と P_2 を交叉した HMM

C_3： P_1 に STATE の遺伝子を反映した HMM

C_4： P_2 に STATE の遺伝子を反映した HMM

C_1，C_2 は交叉により生成された HMM であり，C_3，C_4 は HMM の各ノードと参照する STATE の遺伝子が一致するよう変更を加えた HMM である．HMM を交叉する際は，2 つの親個体からそれぞれランダムに選択したノードを削除した後，ベースとする親個体に他方の親個体のノードを組み入れる.

子の生成後，C_1〜C_4 の終了状態以外の状態を表すノードに対して確率 p_m で突然変異を発生させる．突然変異では，当該ノードの参照する STATE を変更し，遺伝子に合わせてエッジを変更する．P_1，P_2，C_1〜C_4 のうち，最良個体およびルーレットで選択された 2 個体を次世代に残す.

一世代の処理の流れを図 3 に示す．初期集団を生成した後，図 3 の処理を G 回繰り返し，最も適応度の高い HMM を出力解とする.

3. 評価実験

提案システムにおける一般的な HMM の生成，およびプロモータデータの分類精度に関する評価実験を行なった．各実験で用いたパラメータの値を表 1 に示す.

3.1 一般的な HMM 生成に関する評価

最大状態数 N_s，1 つの状態からの最大遷移数 N_t，出力シンボル 5 種という条件下で，状態数，遷移先，遷移確率，出力確率をランダムに設定して，HMM を 10 個生成する．各 HMM で出力した 100 個のシンボル列に加え，各 HMM では出力不可能なシンボル列を 100 個生成し，data1〜data10 とする.

表 2: ランダムに生成された HMM によるデータでの結果

	元の HMM		システムで生成した HMM			
	状態	遷移	状態	正解率	再現率	適合率
data1	5	5	8.2	0.803	0.910	0.754
data2	4	5	5.6	0.994	0.998	0.989
data3	3	3	4.0	0.975	1.000	0.952
data4	5	5	6.0	0.840	0.970	0.770
data5	6	10	8.0	1.000	1.000	1.000
data6	5	7	10.1	0.817	0.955	0.750
data7	3	4	3.0	0.914	0.909	0.921
data8	3	3	3.8	0.974	1.000	0.950
data9	6	9	4.1	0.876	0.932	0.845
data10	2	2	2.2	0.998	1.000	0.996

各データの状態数と遷移数，各データを訓練例集合として提案システムで HMM の生成を 10 回繰り返したときの各回の状態数，正解率，再現率，適合率の平均を表 2 に示す．ここで正解率とは，各データの出力可否が正解である割合を表す．状態数，遷移数が異なるさまざまなデータで 90%以上の再現率が得られていることがわかる．

適合率が 70%代である data1，data4，data6 の元の HMM には，自己ループを除くと各状態から単一の状態への遷移しかなく，初期状態から終了状態への自己ループを通らないルートは 1 通りしかないという共通点が見られる．新しい構造の個体を生成するための交叉では，2 つの親個体が持つ初期状態から終了状態へのルートを統合しているため，ルートが 1 通りの構造が生成されにくいことが原因の 1 つと考えられる．

状態数，遷移数が最も多い data5 と最も少ない data10 について，元の HMM と提案システムが生成した HMM を図 4 に示す．いずれも正解率が 100%を得た HMM である．data5 では，元の HMM より状態が 1 つ少ない HMM が生成されたが，構造は非常に類似しており，最後に 1 個以上のシンボル “a” が出力されるという特徴が表現されている．

また，data10 のような簡素な構造の HMM 同士では，交叉で統合しても構造が変化しない．したがって，先に述べた「初期状態から終了状態までのルートが 1 通りの HMM は生成されにくい」という問題が発生しないため，遷移確率のみが異なる同一構造の HMM の生成に成功していると考えられる．

3.2 プロモータデータによる評価

[Chowdhary 05] では哺乳動物のヒストンのプロモータ領域のモデリングを行なっている．この分析に用いられている 127 個のプロモータデータにより評価実験を行なった．長さ 250 の塩基配列から MEME/MAST[Bailey 98] によりモチーフが抽出されたデータであり，表 3 に示す 9 種のモチーフとそれぞれの相補配列の番号*1，およびモチーフ以外の塩基を表す`Z` の配列として表現されている．`A`，`G`，`C`，`T` はそれぞれアデニン，グアニン，シトシン，チミンの 4 種の塩基を表す．

ヒストンは H1，H2A，H2B，H3，H4 の 5 つのグループからなることが知られている (表 4)．ある特定のたんぱく質を表す塩基配列のデータを D_p，それ以外の塩基配列のデータを D_n として，提案システムにより HMM を生成する．

各たんぱく質について 10 回ずつ HMM 生成を繰り返したときの状態数，正解率，再現率，適合率の平均を表 5 に示す．いずれのたんぱく質においても，5〜7 個程度の状態数で 80%以上の正解率が得られた．

*1 番号 n のモチーフの相補配列は番号 $-n$ で表されている．

表 3: モチーフ

モチーフ番号	塩基配列
1	`TCTGATTGGTTA`
2	`ATGCAAATGAGG`
3	`CTATAAAAACC`
4	`TTTTCGCGCCCA`
5	`CAATCAGGTCCG`
6	`AACAAACACAA`
7	`CAGCCAATCAGA`
8	`CCATTGGTTAAA`
9	`CCCCGCCCCCCG`
-1	`TAACCAATCAGA`
-2	`CCTCATTTGCAT`
-3	`GGTTTTTATAG`
-4	`TGGGCGCGAAAA`
-5	`CGGACCTGATTG`
-6	`TTGTGTTTGTT`
-7	`TCTGATTGGCTG`
-8	`TTTAACCAATGG`
-9	`CGGGGGGCGGGG`

表 4: プロモータデータ

ヒストンのグループ	データ数	データ内訳		
		ヒト	マウス	ラット
H1	19	9	8	2
H2A	29	15	12	2
H2B	32	17	13	2
H3	23	11	11	1
H4	24	13	10	1

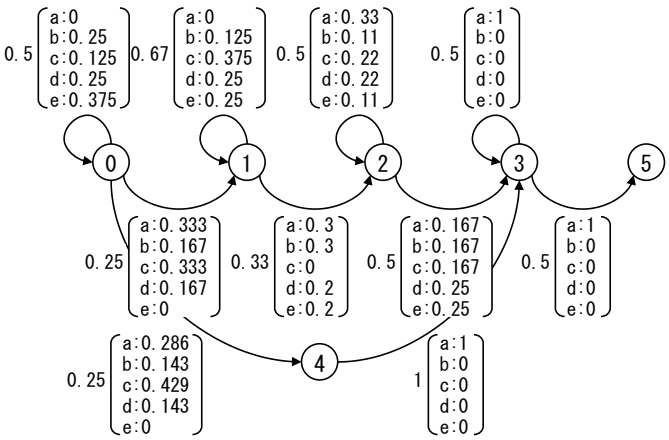
H2B のたんぱく質を識別するために生成された HMM を図 5 に示す．ここでは，各シンボルの出力確率の代わりに，出力されるシンボル，あるいは出力されないシンボルを大括弧の中に記している．括弧のついていないシンボルが出力されるシンボル，括弧のついたシンボルが出力されないシンボルである．この図では状態 1，2，4 はモチーフ番号 (−2) が比較的前の方に出現する場合を表し，状態 3，4 はこのモチーフが後半に出現するか，または，出現しない場合を表す．このように，獲得した HMM が，ヒストンのさらなる分類に利用できる可能性を示唆している．

4. おわりに

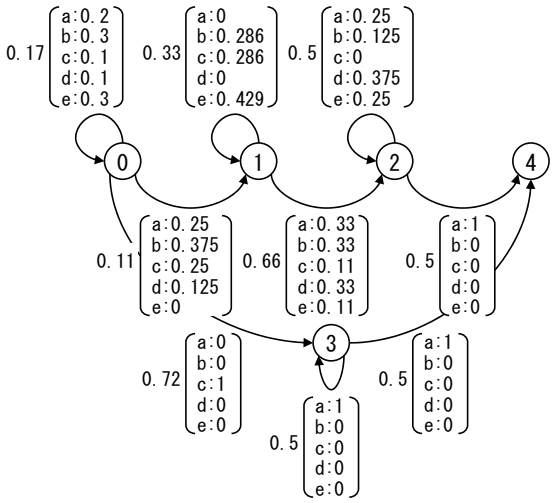
本研究では，各種の塩基性たんぱく質におけるモチーフの出現パターンの特徴を示すことを目的として，共生進化に基づく HMM の生成手法を提案した．1 つの状態に関する性質と，それらの組合せを並行して進化させる共生進化の特長を生

表 5: プロモータデータでの結果

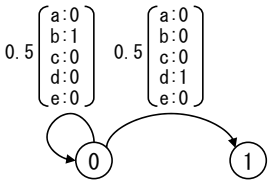
	状態数	正解率	再現率	適合率
H1	7.3	0.884	0.711	0.610
H2A	7.2	0.824	0.645	0.641
H2B	7.0	0.861	0.734	0.741
H3	5.2	0.873	0.717	0.661
H4	5.7	0.884	0.721	0.713



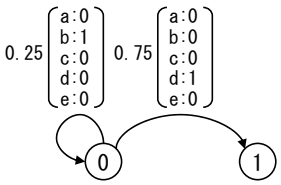
(a) 元の HMM (data5)



(b) 提案システムで生成した HMM (data5)



(c) 元の HMM (data10)



(d) 提案システムで生成した HMM (data10)

図 4: 元の HMM と提案システムで生成した HMM の比較

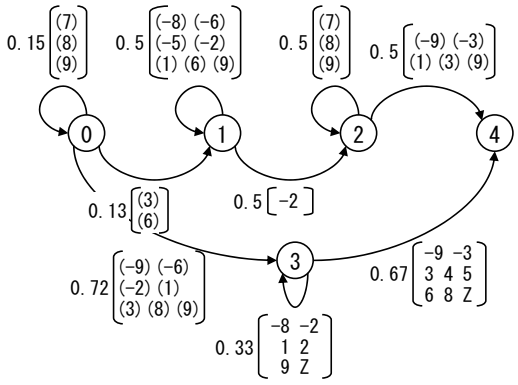


図 5: H2B を識別する HMM

かし，高い精度の HMM が生成できることが示された．しかし，プロモータデータに関してはまだ十分な精度が得られているとはいえない．HMM の交叉法の見直し，およびプロモータデータの特徴を考慮した処理の組込みなど，改良を進める必要がある．

謝辞

ヒストンのデータを利用するにあたり，貴重な助言をいただいた京都大学大学院情報学研究科准教授の矢田哲士博士に感謝いたします．

参考文献

[Bailey 98] Bailey, T. and Gribskov, M.: Combining Evidence using p-values: Application to Sequence Homology Searches, *Bioinformatics*, Vol. 14, pp. 48–54 (1998)

[Chowdhary 05] Chowdhary, R., Ali, R., Albig, W., Doenecke, D., and Bajic, V.: Promoter Modeling: the Case Study of Mammalian Histone Promoters, *Bioinformatics*, Vol. 21, pp. 2623–2628 (2005)

[Moriarty 96] Moriarty, D. and Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, *Machine Learning*, Vol. 22, pp. 11–32 (1996)

[大谷 02] 大谷 紀子, 大和田 勇人：共生進化に基づく帰納論理プログラミングの予測精度の向上, 人工知能学会論文誌, Vol. 17, No. 4, pp. 431–438 (2002)

[大谷 04] 大谷 紀子, 志村 正道：共生進化に基づく簡素な決定木の生成, 人工知能学会論文誌, Vol. 19, No. 5, pp. 399–404 (2004)

[佐藤 97] 佐藤 浩, 小野 功, 小林 重信:遺伝的アルゴリズムにおける世代交代モデルの提案と評価, 人工知能学会誌, Vol. 12, No. 5, pp. 734–744 (1997)