隠れマルコフモデルと遺伝的アルゴリズムによるDNA配列のシグナルパターン抽出

矢田哲士[†1] 石川幹人[†2]
田中秀俊[†3] 浅井潔[†4]

我々は，DNA配列群からシグナルパターンを自動的に抽出する手法を開発した．本手法では，シグナルパターンを確率論的モデルである隠れマルコフモデル（HMM）によって表現している．HMMは，状態を表すノードとそれらを結合する有向パスで構成されるネットワークとして記述される．HMMをシグナルパターンの表現方法として用いる場合，以下の2点が重要な課題となる．(1) 最も好ましいネットワークトポロジーの決定，(2) HMMに関連するパラメータの最適化．本手法は，遺伝的アルゴリズム（GA）とBaum-Welchアルゴリズム（BWA）で構成される．手法のプロシジャは以下のとおりである．(1) GAによるネットワークトポロジーと初期パラメータ値の生成，(2) BWAによるパラメータ値の最適化，(3) GAによるネットワークトポロジーと最適化されたパラメータ値の評価．評価には，モデルの適合度と複雑性の釣合基準を与える赤池情報量基準（AIC）を適用した．以上のプロシジャを繰り返すことによって，DNA配列群に含まれるシグナルパターンを最も良好に表現するHMMネットワークのトポロジーとパラメータ値が得られる．我々は，本手法を霊長類プロモータ領域に関するシグナルパターンの抽出に適用した．本手法により生成されたHMMは，生物学的に知られている複数のシグナル配列を含んでいた．さらに，このHMMを用いてプロモータ領域の予測を行った結果，84.3%の精度でプロモータ領域を認識することが確かめられた．この値は，本手法で生成されたHMMがプロモータ領域のシグナルパターンを良好に表現していることを示している．

Signal Pattern Extraction from DNA Sequences Using Hidden Markov Model and Genetic Algorithm

TETSUSHI YADA,[†1] MASATO ISHIKAWA,[†2] HIDETOSHI TANAKA[†3] and KIYOSHI ASAI[†4]

We have developed a method for the extraction of signal patterns from DNA sequences. In the method, signal patterns are represented as a stochastic model called Hidden Markov Model (HMM) which is described as a network composed of some nodes representing states and some directed paths connecting them. In using this model, problems are attributed to the following two: (1) determination of the most preferable network topology; (2) optimization of parameters associated with this model. The method consists of a Genetic Algorithm (GA) and a Baum-Welch algorithm (BWA). The procedures of the method are as follows: (1) generation of network topologies and initial parameter values using GA; (2) optimization of the parameter values using BWA; (3) evaluation of the network topologies and optimized parameter values using GA. Akaike Information Criterion (AIC), which gives a criterion for the balance of adaptation and complexity of a model, is applied in the evaluation. By repeating the above procedures, the topology and parameters for the most preferable network are obtained. We have applied the method to the extraction of signal patterns from primate promoters. The method has generated an HMM representing signal patterns in the promoters. To validate the method, we have applied the above results to promoter recognition in primate sequences. We have observed that the HMM can recognize promoters with an accuracy of 84.3%. This indicates that good representation of signal patterns has been obtained by the method.

†1 日本科学技術情報センター
The Japan Information Center of Science and Technology
†2 松下電器産業株式会社
Matsushita Electric Industrial Co., Ltd.
†3 三菱電機株式会社
Mitsubishi Electric Co., Ltd.
†4 電子技術総合研究所

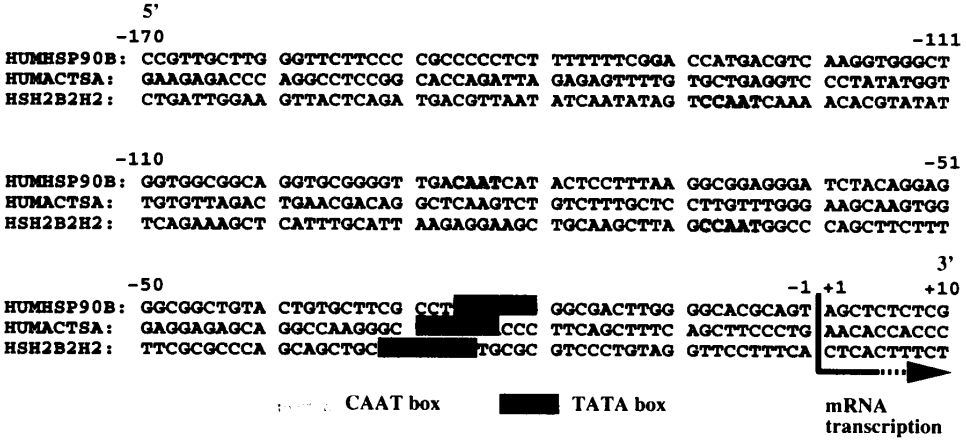

図1 プロモータ領域に見つかっているシグナルパターンの例
Fig. 1 An example of signal patterns found in promoters.

1. はじめに

近年のDNAシークエンシングプロジェクト（DNA sequencing project）の進展にともない，大量のDNA配列群が蓄積されている．このDNA配列群を用いて，生物学的に重要な働きをになうシグナルパターンを自動的に抽出する試みは，興味深い研究課題となっている．DNA配列は，A，T，C，Gの4つのアルファベットで構成される1次元の文字列としてとらえることができる．シグナルパターンとは，異なる配列間に共通に存在する部分文字列（シグナル配列）が形成するパターンを指す．シグナルパターンは，機能あるいは進化に関する生物学的に重要な情報を含んでいることが多い．このため，シグナルパターンの抽出手法は，遺伝情報の発現や制御機構を解明するために欠くことのできない要素技術となっている．また，シグナルパターンに関する生物学的知見が不十分である場合が多いため，抽出されたシグナルパターンを分析することにより，新たな配列情報が発見される可能性が期待されている．

図1に，シグナルパターンの具体例として，遺伝子HUMHSP90B，HUMACTSA，HSH2B2H2のプロモータ（promoter）領域のDNA配列を示す[1]．DNA配列断片は，化学的な方向性を有する．化学的な方向は，配列断片の両端で定義され，一端を5′末端，他端を3′末端と呼ぶ．DNAの配列情報は，5′末端から3′末端の方向に読まれる．このため，ある領域の5′末端側を上流，3′末端側を下流と呼ぶ．図1では，各配列をmessenger RNA（mRNA）の転写開始部位で整列させ，mRNA転写開始部位から上流に170塩基，下流に10塩基のDNA配列を表示している．HUMHSP90B，HUMACTSA，HSH2B2H2のプロモータ領域には，TATA box，CAAT boxと呼ばれる2種類のシグナル配列の存在が知られている．図1では，各DNA配列についてシグナル配列に該当する部分文字列を色塗りで表示している．これらのシグナル配列は，相互に作用し合い，遺伝情報のmRNAへの転写を巧みに制御している．mRNAへの転写は，転写開始部位から下流に向かって行われ，この領域に遺伝子が存在している．mRNAへの転写制御は，遺伝情報の発現制御機構の重要な一部分であることが知られている．一般に，シグナルパターンはシグナル配列の組であり，図1ではTATA boxとCAAT boxが見られる．ここでは，他のプロモータ配列で知られているGC boxやCT signalなどのシグナル配列は報告されていないが，シグナルパターンを抽出することによって，これらのシグナル配列が同定される可能性がある．

シグナルパターンの抽出手法では，どのようにしてシグナルパターンの多様性を表現するかが重要な課題である．シグナルパターンの多様性として，シグナル配列の塩基配列，シグナル配列の位置，シグナル配列の組合せに関する多様性が知られている[2),3)]．これらの多様性について，図1を例にして説明する．TATA boxの塩基配列に着目すると，HUMHSP90BではTATATAG，HUMACTSAではTATATAA，HSH2B2H2ではTATAAAAとなっている．また，mRNA転写開始部位を基準としたTATA boxの位置は，HUMHSP90Bで−27，HUMACTSAで−30，HSH2B2H2で−32となっている．さらに，HUMHSP90Bでは1つのTATA boxと1つのCAAT box，HUMACTSAでは1つのTATA box，HSH2B2H2では1つのTATA boxと2つのCAAT boxが存在している．以上のように，シ

Electrotechnical Laboratory

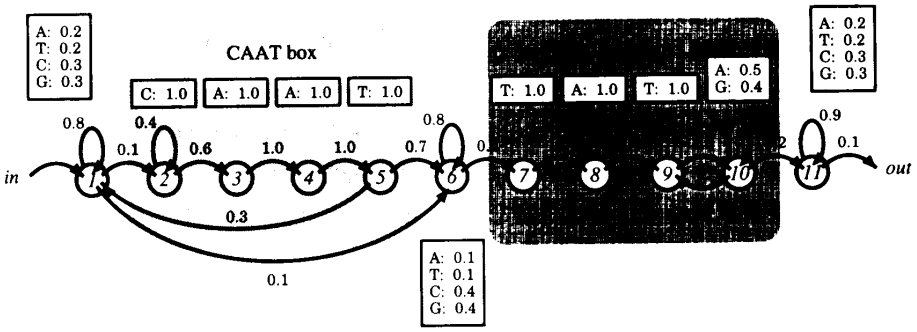

図2 プロモータHMMの例
Fig. 2 An example of promoter HMM.

グナル配列の塩基配列，位置，組合せはプロモータ配列ごとに多様である．

多様性に富むシグナルパターンの表現手法として，確率論的モデルは有効である．我々は，シグナルパターンの表現方法として，確率論的モデルの一種である隠れマルコフモデル（hidden Markov model; HMM）[4]に注目した．HMMでは，シグナル配列の塩基配列，位置，組合せで特徴付けられるシグナルパターンの多様性を，確率を用いることによって柔軟に表現することが可能である．シグナルパターンの確率的な表現モデルは，高い感度でDNA配列中に存在するシグナルパターンを検出することが知られている．このため，DNAやアミノ酸の配列解析にHMMを適用する試みが活発に研究されている[5)~11)]．

HMMは，ノードとノード間を相互に結合する有向パスで構成されるネットワーク構造として記述される．ノードと有向パスは，それぞれ状態と状態遷移を表している．各ノードは出力シンボル確率（output symbol distribution）と初期状態確率（initial state distribution），各有向パスは状態遷移確率（state transition probability）と呼ばれるパラメータで特徴付けられている．HMMを用いてシグナルパターンを表現する場合，HMM中のある状態はDNA配列中のある領域または部位の特徴を表現している．出力シンボルはA，T，C，Gとなり，状態遷移はDNA配列中の位置の移動を表している．

図1に示したプロモータ配列群に含まれるシグナルパターンを，HMMを用いて表現した例を**図2**に示す．各状態に記された数字は状態番号を示す．CAAT boxとTATA boxの塩基配列は，各々，状態遷移 $\underline{S_2} \to S_3 \to S_4 \to S_5$ と $S_7 \to S_8 \to S_9 \leftrightarrow \underline{S_{10}}$ で表されている．下線で記された状態は，自己ループを有することを示す．図1に現れるCAAT boxの塩基配列は，塩基Cの繰り返し回数が配列により異なる．これは，状態 S_2 の自己ループの状態遷移確率で表現されている．また，TATA boxの塩基配列に見られる塩基TAの繰り返し回数の相違や塩基Gの出現は，状態遷移 $S_9 \leftrightarrow \underline{S_{10}}$ や状態 S_{10} の出力シンボル確率で表現されている．シグナル配列の位置は，シグナル配列間の相対的な距離として表現されている．たとえば，CAAT boxとTATA boxの平均距離は状態 S_6 の自己ループの状態遷移確率に反映されている．シグナル配列の組合せは，状態 S_1，S_5，S_6 が関与する状態遷移によって表現されている．1つのTATA boxと1つのCAAT boxが存在するHUMHSP90Bでは，シグナルパターンを表現する状態遷移は $\underline{S_1} \to \underline{S_2} \to \cdots \to S_5 \to \underline{S_6} \to S_7 \to \cdots$ となる．1つのTATA boxしか存在しないHUMACTSAでは，$\underline{S_1} \to \underline{S_6} \to S_7 \to \cdots$ となる．2つのCAAT boxと1つのTATA boxが存在するHSH2B2H2では，$\underline{S_1} \to \underline{S_2} \to \cdots \to S_5 \to \underline{S_1} \to \underline{S_2} \to \cdots \to S_5 \to \underline{S_6} \to S_7 \to \cdots$ となる．

以上のように，図1に示した4本のプロモータ配列に含まれるシグナルパターンは，図2に示したHMMのネットワーク全体で表現されている．本研究では，より多くのDNA配列群を解析の対象とするため，シグナルパターンの多様性はいっそう広がり，より複雑な構造を有するHMMが必要になる．

図2の例でも明らかなように，HMMに存在するパラメータは，シグナルパターンの塩基配列，位置，組合せの特徴を表現している．これらのパラメータ値を，与えられたデータセットに対して最適化するアルゴリズムとして，Baum-Welchアルゴリズム（Baum-Welch algorithm; BWA）[12)]が確立されている．

BWAを適用する場合，HMMのネットワークトポロジーとパラメータの初期値を与える必要がある．HMMを用いてDNA配列群のシグナルパターンを抽出する場合，HMMのネットワークトポロジーはシグナルパターンに大きく依存する．つまり，ネットワークトポロジーを設計するためには，与えられたDNA配列群

に含まれるシグナル配列の特徴（種類数，配列長，周期性など）やシグナル配列間の関係（距離，組合せなど）に関する十分な知見が必要となる．しかしながら，未知のシグナルパターンを抽出する場合，パターンに関する先験的な知見は期待できない．また，BWA は大域的な最適値へパラメータが収束することを保証していないため，パラメータの初期値問題が存在する．以上より，HMM を用いてシグナルパターンの抽出を試みる場合，良好なネットワークトポロジーと初期パラメータ値の決定が重要な問題となる．

ネットワークトポロジーの自動的な設計手法として，successive state splitting アルゴリズム（successive state splitting algorithm; SSS algorithm）[13]と iterative duplication 法（iterative duplication method; ID method）[14]が提案されている．前者は，left-to-right モデルのトポロジーを設計するために音声認識の分野で提案され，アミノ酸配列の分類に応用された[15]．left-to-right モデルはグローバルループを持たない．このため，SSS アルゴリズムを DNA 配列のシグナルパターン抽出に適用した場合，繰り返し DNA 配列などの周期的なシグナルパターンを表現するトポロジーの設計が困難になる．後者は，アミノ酸配列のモチーフ抽出のために提案された．SSS アルゴリズムと比較すると，ID 法はより一般的なトポロジーを扱うことができる．ID 法は，HMM 中で最も状態遷移数の大きいノードを複製することによってトポロジーを生成する．この制限されたトポロジー生成は，少ない計算量で効率的な探索を実現しているものの，トポロジーの局所最適解に陥る可能性が高い．SSS アルゴリズムや ID 法によって，最適なトポロジーが生成されたとしても，パラメータの初期値問題が残されたままである．

本研究の目的は，DNA 配列群に存在するシグナルパターンの表現モデルを，自動的に生成する手法を開発することである．ここでは，シグナルパターンの表現モデルとして HMM を採用し，HMM を用いる際に問題となるネットワークトポロジーと初期パラメータ値の効率的な最適化手法を考察した．我々は，組合せ問題の最適化手法のひとつである遺伝的アルゴリズム（genetic algorithm; GA）[16),17)]の枠組みで，トポロジーと初期パラメータ値の最適化問題をとらえることを試みた．我々の手法は，GA によるヒューリスティックを効果的に取り入れたトポロジーと初期パラメータ値の最適化手法である．本手法では，SSS アルゴリズムや ID 法のようなトポロジー生成の制限がなく，任意のトポロジーの生成が可能である．

2. 手　　法

GA は，生物の進化過程を模倣した探索アルゴリズムである．生物進化は，自然淘汰（natural selection）と遺伝情報の生成により押し進められる．自然淘汰の働きにより，ある世代（generation）において，より環境に適した遺伝情報を持つ個体が，より多くの子孫を次世代に残すこととなる．また，遺伝情報が生成されることにより，集団中には絶えず新しい遺伝情報を保有した個体が産出される．これらの働きにより，環境に対して最適な遺伝情報を持つ個体が，世代とともに集団中に広まることになる．

GA では，個体（individual）で構成される集団（population）が定義される．各個体は，染色体（chromosome）と呼ばれる 1 次元の配列を保有する．染色体中には，問題に対する可能な解がエンコードされている．この解は遺伝情報として取り扱われる．GA では，淘汰の指標として適応度（fitness）と呼ばれる量を定義する．適応度は，個体が保有する解のパフォーマンスを表している．GA における淘汰プロセスは，適応度に応じた頻度で個体間を交配（mating）させ，次世代の集団を生成することによって行われる．遺伝情報の生成は，組換え（recombination）や突然変異（mutation）と呼ばれる操作を，交配時に施すことによって行われる．これらの操作は，組換え率（recombination rate），突然変異率（mutation rate）と呼ばれるパラメータ値で制御される確率的な操作である．

本手法は，GA と BWA で構成される．手法の概要を**図 3** に示す．入力データとして，あるカテゴリーに属する DNA 配列群が与えられる．本手法では，シグナルパターンを HMM の形式で表現し，HMM を個体と見なした集団を定義する．手法のプロシジャを以下に記す．はじめに，GA（組換え操作や突然変異操作）を用いて，各 HMM のネットワークトポロジーとパラメータの初期値を生成する．ただし，初期世代の集団では，乱数によって生成する．続いて，BWA を用い，与えられた DNA 配列群に対して各 HMM のパラメータ値を最適化する．最後に，GA（淘汰プロセス）を用いて各 HMM のトポロジーと最適化されたパラメータ値の評価を行う．本手法は，以上の操作を繰り返すことにより，入力配列群のシグナルパターンを表現するのに最も適したトポロジーとパラメータ値で構成される HMM ネットワークを生成する．

集団中の各個体は，HMM のトポロジーと初期パラメータ値に関する情報を遺伝情報として保有している．

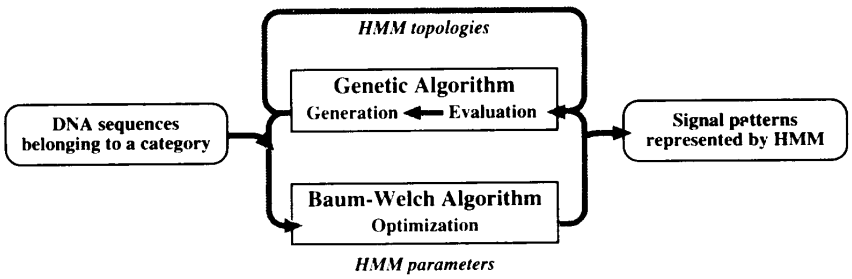

図3 手法の概要
Fig. 3 Schematic diagram of the method.

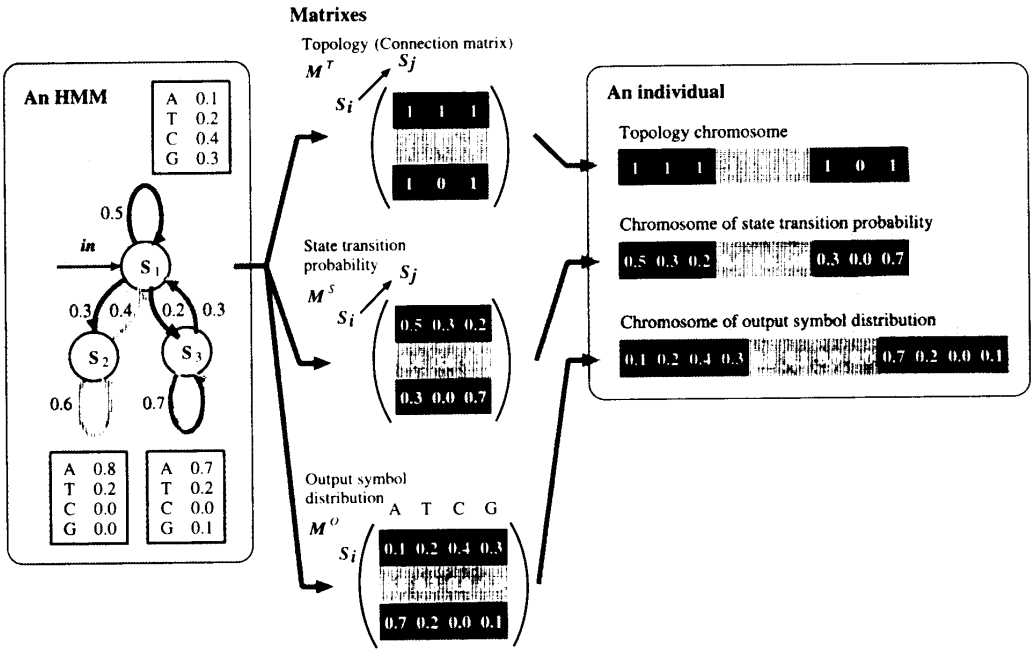

図4 染色体へのHMMのエンコーディング
Fig. 4 Encoding of HMM network based on strong specification method.

各個体は，トポロジー，状態遷移確率の初期値，出力シンボル確率の初期値を指定する3つの染色体で構成される．ここでは，初期状態を固定したため初期状態確率の初期値を指定する染色体は定義しなかった．

我々は，3種類のパラメータセットを各染色体にエンコードする規則として，強指定法（strong specification method）[18]を応用した．強指定法は，ニューラルネットワークのトポロジーに関するエンコーディング規則として提案された．強指定法を応用したHMMのエンコーディング例を図4に示す．HMMトポロジーのエンコーディング手法は以下のとおりである．一般に，任意のトポロジーは結合行列（connection matrix）に変換することが可能である．結合行列中の要素 M_{ij}^T は，ノード i からノード j への有向パスの状態を表している．ここでは，HMMにおける状態 S_i から状態 S_j への状態遷移が存在すれば結合行列の要素 M_{ij}^T は1，状態遷移が存在しない場合は0としている．トポロジー染色体は，結合行列の各列を順次結合することによって組み立てられる．そのため，状態数が m のHMMの場合，状態 S_i から状態 S_j への遷移はトポロジー染色体の $m(i-1)+j$ 番目の要素にマッピングされることになる．強指定法は任意のトポロジーをエンコードすることができるため，我々はトポロジーに関するすべての探索空間を考察することが可能である．状態遷移確率の初期値をエンコードする染色体は，トポロジー染色体と同様の手法によって組み立てられる．ただし，状態遷移確率行列の要素 M_{ij}^S が状態 S_i から状態 S_j への遷移確率を表すことになり，M_{ij}^S の値は実数になる．状態遷移確率は，暗にトポロジーに関する情報を含んでいるが，ここでは，トポロジー染色体において状態遷移が存在する場合のみ状態遷移確率染色体の対応部分を参照することとした．このようにすると，トポロジー染色体と状態遷移確率染色体の対応する要素を掛けた値が，状態遷移確率の

真の値を表すことになる．出力シンボル確率の初期値をエンコードする染色体も，出力シンボル確率行列の各列を順次結合することによって組み立てられる．この行列の要素 M_{ij}^{O} は，状態 i における塩基 j の出力確率を表している．

我々は，適応度の算出に赤池情報量基準（Akaike information criterion; AIC)[19]を応用した．モデルが複雑になると，与えられた配列に対してモデルが過適合することが報告されている[20]．AIC は，モデルの適合性と複雑性のバランスに関する基準を与えてくれる．HMM 集団における i 番目の HMM の相対適応度 W_i は，次式で与えられる．

$$W_i = w_i / \sum_{j=1}^{N} w_j \qquad (1)$$

with

$$w_i = [-2 \log L(\hat{\theta}_i; f) + 2\lambda p_i]^{-1} \qquad (2)$$

ここで，N は集団中の HMM 数，$\log L(\hat{\theta}_i; f)$ は BWA で算出される i 番目の HMM の推定最大対数尤度，p_i は i 番目の HMM の自由パラメータ数である．最大対数尤度は HMM の適合性を表し，自由パラメータ数は HMM の複雑性を表す．λ は適合性と複雑性のバランスを調整するバランスパラメータである．$\hat{\theta}_i$ は i 番目の HMM の推定パラメータ値セット，f は入力 DNA 配列群を表す．HMM の自由パラメータ数とは，HMM を規定するパラメータ数を指し，各状態に関する自由パラメータ数の和で算出される．各状態における自由パラメータ数は，(パラメータ値が 0.0 以外の状態遷移確率の数−1)+(パラメータ値が 0.0 以外の出力シンボル確率の数−1) で計算される．たとえば，図 4 に示す HMM の自由パラメータ数は 10 である．

我々は，新しいネットワークトポロジーを生成するために組換え操作と 3 種類の突然変異操作を設計した．突然変異には，挿入突然変異（insert mutation)，欠失突然変異（delete mutation)，点突然変異（point mutation）がある．

挿入突然変異操作と欠失突然変異操作は，HMM の各状態について定義され，トポロジーの成長と縮退を引き起こす．HMM トポロジーの成長と縮退の例を**図 5** に示す．上図は，挿入突然変異操作による HMM トポロジーの成長を示す．状態 S_i に挿入突然変異が生じた場合，S_i と同じ状態遷移を有する新しい状態 S_i' を生成する．つまり，S_i が状態 S_j との遷移を有していた場合，S_i' も S_j との遷移を有することになる．S_i' に関する状態遷移確率は，対応する S_i の状態遷移確率と等しいものとなる．また，S_i' に関する出力

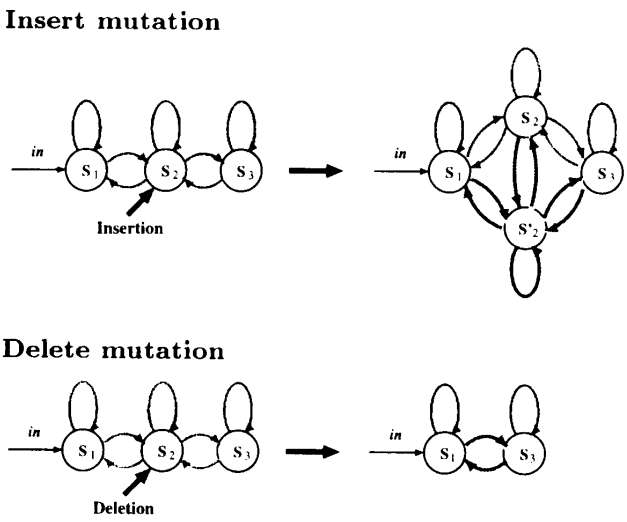

図 5 HMM トポロジーの成長と縮退
Fig. 5 Topology growth and contraction.

シンボル確率は，S_i の出力シンボル確率と等しいものとなる．S_i が自己ループを形成していた場合，S_i' は自己ループを形成し，S_i と S_i' 間に相互状態遷移を形成する．相互状態遷移確率は，乱数により決定される．挿入突然変異操作によって生成されるトポロジーは，必ず元のトポロジーを含んでいる．この操作により，トポロジーは段階的に成長することになる．挿入突然変異操作は，ID 法によるトポロジー生成操作に類似している．ただし，ID 法が最も状態遷移数が大きい状態に対して状態の複製を行なうのに対して，我々の手法はすべての状態に対して一定の確率で状態の複製を行う．これは，多様なトポロジーを生成し，局所最適解に収束することを防ぐためである．下図は，欠失突然変異操作による HMM トポロジーの縮退を示す．状態 S_i に欠失突然変異が生じた場合，S_i と S_i が関係する状態遷移を取り除く．その後，S_i を経て結合していた 2 つの状態間に状態遷移を生成する．生成された遷移に関する状態遷移確率は，乱数により決定される．

組換え操作と点突然変異操作は，トポロジーの遠方探索と近傍探索に対応する操作になっている．HMM トポロジーの変化の例を**図 6** に示す．上図は，組換えによる HMM トポロジーの変化を示す．組換え操作は，状態数が等しい 2 つの HMM 間について定義され，トポロジー染色体の部分要素の交換を行う．染色体中の i 番目の要素に組換えが生じた場合，$i+1$ 番目以降の部分要素を 2 つの HMM 間で交換する．図 6 の上図では，染色体中の 2 点で組換えが生じた場合を示している．下図は，点突然変異による HMM トポロジーの変化を示す．点突然変異操作は，トポロジー染

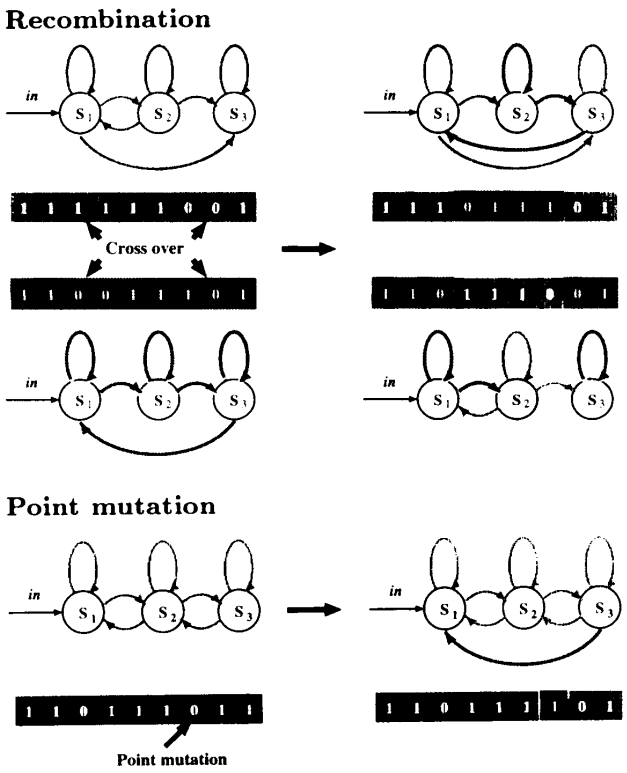

図6 HMMトポロジーの変化
Fig. 6 Topology change.

色体の各状態遷移について定義される．状態数が m のHMMについて，トポロジー染色体の $m(i-1)+j$ 番目の要素に点突然変異が生じた場合，この要素のビットが反転する．この操作により，状態遷移 $S_i \to S_j$ が存在すれば状態遷移が消滅し，$S_i \to S_j$ が存在しなければ状態遷移が生成される．状態遷移が生成された場合，対応する状態遷移確率染色体の要素が有効となる．図6の下図では，点突然変異により状態遷移が生成された場合を示している．

我々は，HMMパラメータの初期値を最適化するために，状態遷移確率染色体と出力シンボル染色体に対する組換え操作と点突然変異操作を設計した．これらの操作は，上述のトポロジーに関する操作とは独立に行われる．前者は，状態数の等しい2つのHMM間について定義され，染色体の部分要素の交換を行う．後者は，染色体の各要素の値に変化を引き起こす．変化の大きさは，乱数によって決定される．

3. データ

我々は，手法の有効性を検討するために，学習用データとテスト用データを作成した．前者は，霊長類のプロモータ領域のシグナルパターンを表現するHMMを生成するために使用した．後者は，正例と負例で構成され，生成されたHMMがパターンの表現モデルとして妥当かどうかを判断するために使用した．

我々は，GenBank Release 76.0[1]よりfeature tableにTATA boxとCAAT boxが明記されている霊長類のエントリ（entry）を選び出した．続いて，このエントリ群からTATA boxとCAAT boxを含む71塩基から成るプロモータ配列を取り出した．データの統計的な偏りを取り除くために，プロモータ配列群から類似性が高い配列を取り除いた．ここでは，アライメント（alignment）[21]の結果，75%以上の類似性を示した2つの配列のうちの一方を取り除いた．以上の操作により，学習用データとして55本のDNA配列が得られた．

テスト用データの正例と負例の配列群を得るために，GenBank Release 76.0より長さ20000以上のDNA塩基配列を含む霊長類のエントリーを選び出した．短い配列断片は，しばしば複数のエントリに重複して登録されていることがある．そこで，配列データの重複を避けるために，比較的長いエントリだけを選んだ．このエントリ群から，正例として71塩基から成る46本のプロモータ配列が，負例として71塩基から成る54362本の配列が得られた．負例の配列データは，プロモータ領域以外の配列について，71塩基長のウィンドウを20塩基ずつシフトさせながら切り出した配列群である．なお，テスト用データを作成したエントリと学習用データを作成したエントリとの間に重複はなかった．このため，学習用データと正例データとの間に配列の重複は存在しない．

4. 結果

我々は，本手法を霊長類プロモータ領域におけるシグナルパターンの抽出に適用した．ここでは，50個のHMMで構成される集団を定義し，5回の探索を行った．探索世代数は，事前にHMM集団の平均適応度の変化を調べ，平均適応度がほぼ収束していると判断される80世代に設定した．初期状態のHMM集団は，トポロジーとパラメータ初期値をランダムに生成した2状態のHMMで構成される．バランスパラメータ λ（式2）は，0.01~0.10の値について生成されるHMMネットワークの複雑さを事前に調べ，今回は経験的に0.02に設定した．染色体における1座位（染色体配列の要素）あたりの組換え率，点突然変異率，挿入突然変異率，欠失突然変異率は，各々0.10，0.05，0.04，0.02とした．淘汰戦略としては，エリート戦略とルーレット戦略を組み合わせた方式を採用した．

我々の手法で生成された霊長類プロモータ領域のHMMを図7に示す．このHMMは，5回の探索の

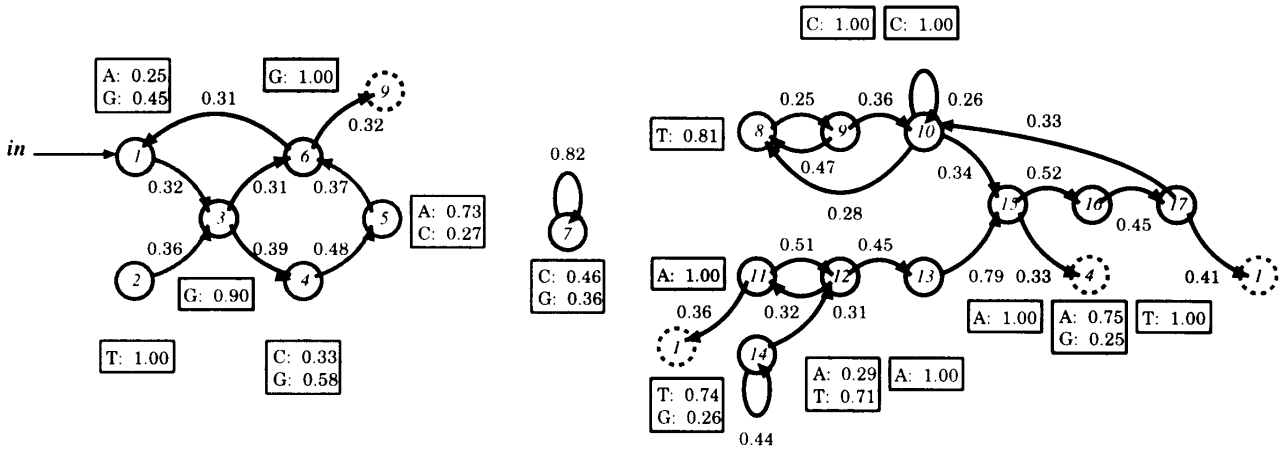

図7 霊長類プロモータ領域のHMM
Fig. 7 A primate promoter HMM obtained by the method.

中で最も高い絶対適応度 w_i（式2）を示したものである．式1の適応度 W_i が相対的なHMMのパフォーマンスの尺度であるのに対し，w_i は絶対的なパフォーマンスの尺度になっている．図7では，確率値が0.25以上の状態遷移と出力シンボルを記している．各状態に記された数字は状態番号を示す．点線で記された状態は，ネットワークを見やすくするために，遷移先の状態を番号で示したものである．

図7では，以下の基準を満たす状態遷移と出力シンボルを重要と見なし，黒実線で記している．ここでは，規則性が高い状態を抽出するために，各状態に関する状態遷移と出力シンボルのエントロピーを計算した．状態 S_i に関する状態遷移と出力シンボルのエントロピー H_i は，次式で与えられる．

$$H_i = -\sum_{j}^{m} P_{i,j} \log P_{i,j} / \log m \qquad (3)$$

状態遷移に関するエントロピーの場合，$P_{i,j}$ は状態 S_i から状態 S_j への状態遷移確率である．m はHMMの状態数で，ここでは17である．出力シンボルに関するエントロピーの場合，$P_{i,j}$ は状態 S_i における塩基 j の出力確率である．m は塩基の種類数で，ここでは4である．$\log m$ で正規化しているため，状態遷移および出力シンボルの両方について $0.0 \leq H_i \leq 1.0$ となる．図7では，閾値を0.5として，状態遷移と出力シンボルのエントロピーがともに $H_i \leq 0.5$ である状態と，対応する状態遷移と出力シンボルとを黒実線で記している．

本手法によって自動的に生成されたHMMがシグナルパターンの表現モデルとして妥当であるかどうかを定量的に議論するために，我々は図7のHMMを用いて霊長類プロモータ領域の認識を行った．ここでは，認識アルゴリズムとしてViterbiアルゴリズム（Viterbi algorithm; VA）[4]を用いた．VAは，ある配列の最大対数尤度と対応する状態遷移を推定することができる．はじめに，学習用データの対数尤度を用いて，プロモータ領域を認識するための閾値を決定した．ここでは，学習用データを92.7%(51/55)で識別する値 -0.772 を閾値とした．続いて，この閾値を用いて，テスト用データの正例と負例に対する認識テストを行った．その結果，図7のHMMは91.3%(42/46)の正例と77.2%(41935/54362)の負例を正しく認識することが確認された．平均認識率は84.3%[(91.3 + 77.2)/2]である．**図8**に霊長類プロモータ領域の認識結果を示す．グラフの上側は正例の対数尤度の分布を，下側は負例の対数尤度の分布を表す．点線で記された中央の垂線は判別のための閾値を表す．

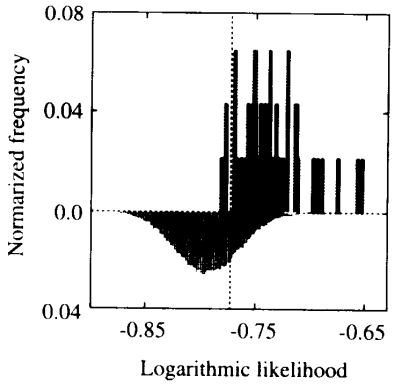

図8 霊長類プロモータ領域の認識結果
Fig. 8 Results of primate promoter recognition using the HMM.

5回の探索におけるHMM集団の絶対適応度の平均値（$\bar{w}_i$）の世代変化を**図9**に示す．絶対適応度の平均値は，w_i（式2）の集団内平均値である．

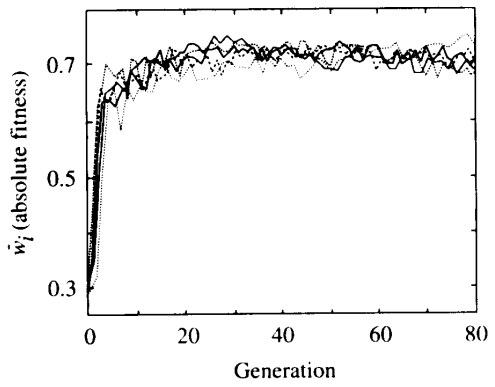

図9 本手法によるHMMの生成過程
Fig. 9 Process of HMM generation by the method.

5. 考　　察

我々は，図7において，TATA boxとCAAT boxを表現している状態遷移を確認した．ここでは，VAを用いて，GenBankに記されているシグナル配列に対応した状態遷移を確認した．その結果，状態遷移 $S_{11} \leftrightarrow S_{12} \rightarrow S_{13} \rightarrow S_{15}$ はTATA box，状態遷移 $S_9 \rightarrow \underline{S_{10}} \rightarrow S_{15} \rightarrow S_{16} \rightarrow S_{17}$ はCAAT boxを表現していることが確認された．下線で記された状態は，自己ループを有することを示す．図7より，これらの状態遷移が高い状態遷移確率を有し，各状態における出力シンボル確率が大きく偏っていることが確認される．状態遷移から出力シンボルを推察すると，TATA boxの代表的なシグナル配列（コンセンサス配列：consensus sequence）はTで始まりATが繰り返された後にAAが続く配列，CAAT boxのコンセンサス配列はCが繰り返された後にAATが続く配列であることが明らかとなる．マルコフ分析法[22]にしたがってHMMの一部を適当な吸収状態と見なすことで，状態 S_{12} から出発して状態 S_{11} を経て再び状態 S_{12} に戻る状態 S_{12} の平均訪問回数と状態遷移 $\underline{S_{10}}$ で出力される平均塩基長を算出することが可能である．計算の結果，前者は1.20，後者は1.35であった．このことは，$S_{11} \leftrightarrow S_{12}$ におけるATの平均的な繰り返しが1.20回，$\underline{S_{10}}$ におけるCの平均的な繰り返しが1.35回であることを示している．以上をまとめると，図7のHMMは，TATA boxのコンセンサス配列がTATAA，CAAT boxのコンセンサス配列がCCAATであることを示している．これらのコンセンサス配列は，一般的なTATA box，CAAT boxのコンセンサス配列[3]，TATAA，CCAATとよく一致している．他の4回の探索においても，これらのコンセンサス配列は確認された．

式3に基づいた状態遷移と出力シンボルを考慮すると，図7のHMMは，より詳細にシグナル配列を表現していることが確認できる．我々は，図7において，$H_i \leq 0.5$ の状態遷移と出力シンボルの組合せで表現されるできるだけ長い塩基配列を探索した．状態遷移 $\underline{S_{14}} \rightarrow S_{12}$ よりTATA boxの5′末端側ではTG-richな部分配列が存在することが，状態遷移 $S_{15} \rightarrow S_{16}$ よりTATA boxの3′末端側では塩基AまたはGが存在することが明らかである．マルコフ分析法[22]より，状態遷移 $\underline{S_{14}}$ が出力する平均塩基長は1.79である．また，状態遷移 $S_3 \rightarrow S_6 \rightarrow S_9$ よりCAAT boxの5′末端側ではG-richな部分配列が存在することが，状態遷移 $S_{17} \rightarrow S_{10}$ よりCAAT boxの3′末端側では塩基Cが存在することが明らかである．状態 S_{10}，S_{16}，S_{17} の状態遷移に関する H_i は，各々0.46，0.48，0.46である．これらの値は，シグナル配列を表現している状態遷移の中では高い値である．VAを用いてシグナル配列の3′側近傍の状態遷移を確認したところ，状態 S_{16} がTATA boxの3′末端，状態 S_{17} あるいは S_{10} がCAAT boxの3′末端の境界であることが明らかとなった．以上のことは，TATAAやCCAATより5′および3′側に拡張された部分配列がTATA boxとCAAT boxのコンセンサス配列として有効であることを示している．拡張部分のコンセンサス配列長は，HMM生成時のバランスパラメータ λ（式2）によって調節されている．拡張されたコンセンサス配列は，iterative weight matrix refinementによって導出されたTATA boxとCAAT boxの詳細なコンセンサス配列[2]とよく一致している．

また，状態 S_1，S_2，S_4，S_5，S_7，S_8 の出力シンボル確率より，霊長類のプロモータにおけるTATA boxやCAAT box以外の領域がGC-richな配列であることが明らかとなる．特に，状態遷移 $S_6 \leftrightarrow \underline{S_7}$ より，プロモータ領域には塩基長5～6のGC-rich部分配列が比較的多く存在することが明らかとなる．マルコフ分析法[22]より，$\underline{S_7}$ で表現されるGC-rich部分配列の平均塩基長は5.56である．状態 S_7 は状態 S_6 と相互状態遷移を有し，遷移確率は，$S_6 \rightarrow S_7$，$S_7 \rightarrow S_6$ ともに0.18で比較的高い．このGC-rich部分配列の特徴は，GC boxの特徴[2]とよく一致している．データベースではGC boxが同定されているプロモータ配列は少ないため，このHMMを用いたVAを行うことによって，プロモータ配列に存在するGC boxが新たに同定される可能性が高い．

認識テストで得られた予測精度（図8）は，本手法によって生成されたHMMがプロモータ領域に関する

シグナルパターンの表現モデルとして妥当であることを示している．霊長類のプロモータ領域予測プログラムとしては，PROMOTER SCAN が有名である[23]．PROMOTER SCAN は，先見的な知見に基づいて作成されたシグナル配列のプロファイルを組み合わせて，霊長類のプロモータ領域を予測する．PROMOTER SCAN の予測精度は，false positive を少なくするように設定された閾値において約 70%であると報告されている．図 7 の HMM の平均認識率は 84.3% であるが，true positive が多くなるように閾値を設定したため，PROMOTER SCAN の予測精度と一概に比較することはできない．しかしながら，図 7 の HMM は，現在報告されている霊長類プロモータ領域の予測モデルとほぼ同等の予測精度を示すと考えることができる．また，生成された HMM の正例のテスト用データに対する認識率は 91.3% と高く，学習用データの認識率 92.7% と比べても認識率の低下は小さい．以上のことは，本手法によって良好なシグナルパターンの表現モデルが生成されたことを示している．

図 9 のグラフは，HMM ネットワークの生成において GA が効果的に機能していることを示している．絶対適応度の HMM 集団内平均値（$\bar{w}_i$）は，すべての探索において，0 ～ 20 世代にかけて急速に立ち上がりが，その後高い値のまま，ある一定の範囲内で推移している．これらの現象は，GA を用いた探索手法の大きな特徴である．

以上の考察により，本手法は霊長類のプロモータ領域に存在するシグナルパターンを表現する HMM を自動的に生成していることが明らかとなる．このことは，GA と BWA の組合せが HMM のトポロジーの生成とパラメータの最適化に有効であり，本手法が実際的な DNA 配列のシグナルパターン抽出問題に十分適用可能であることを示している．本手法の利点は，シグナルパターンの抽出において，シグナル配列に関する先見的な知見（種類数，配列長など）を必要としないことである．

図 7 および図 8 より，本手法に関する 2 つの課題を指摘することができる．(1) シグナル配列の評価が容易な HMM の生成，(2) 認識精度の高い HMM の生成．図 7 の HMM では，ある状態が複数のシグナル配列やシグナル配列以外の領域を共通に表現している（状態の統合）．たとえば，TATA box と CAAT box の部分配列が状態 S_{15}, S_{16} で共通に表現されている．このため，HMM ネットワークを見ただけで TATA box と CAAT box の 3′末端の境界を評価することが難しく，VA を用いた状態遷移を確認する必要がある．図 8 の認識結果では，負例のテスト用データに関する認識率は 77.2% であり，多くの false positive が観察されている．

前者の課題について，モザイク構造を持つ HMM ネットワークの生成手法を研究することが有効であると考えられる．モザイク構造 HMM は，シグナル配列を表現する部分 HMM 群とシグナル配列以外の領域を表現する部分 HMM 群で構成される．部分 HMM 間は少数の状態遷移で結合される．モザイク構造 HMM の生成は，DNA 配列がモザイク的な構造を持つことから考えても妥当である．モザイク構造 HMM では，各シグナル配列が分離された部分 HMM で表現されるため，シグナル配列の評価が容易になることが期待される．つまり，モザイク構造 HMM のある状態は，あるシグナル配列またはシグナル配列以外の領域を表現することになり（状態の分離）シグナル配列の境界が明確になる．

後者の課題について，モザイク構造 HMM を用いた認識精度の評価は興味深い．図 8 における false positive の配列データを調べたところ，塩基組成に大きな偏りがある配列や規則的な繰り返し部分配列を有する配列が多く含まれていた．具体的には，polyA や polyT 部分配列を含む配列，CT 繰り返し部分配列を含む配列，GT 繰り返し部分配列を含む配列，GC-rich 配列などが含まれていた．polyA 部分配列は状態遷移 $S_{11} \rightarrow S_{12} \rightarrow S_{13} \rightarrow S_{15} \rightarrow S_{16} \rightarrow S_{11}$，($S_{16} \rightarrow S_{11}$ の状態遷移確率は 0.10)，polyT 部分配列は状態遷移 $\underline{S_{14}}$，CT 繰り返し部分配列は状態遷移 $S_8 \leftrightarrow S_9$，GT 繰り返し部分配列は状態遷移 $S_2 \leftrightarrow S_3$ ($S_3 \rightarrow S_2$ の状態遷移確率は 0.23)，GC-rich 配列は状態遷移 $\underline{S_7}$ によって高い対数尤度を示していた．このように，塩基組成に大きな偏りがある配列や規則的な繰り返し部分配列を有する配列は，シグナル配列を表現している状態遷移や出力シンボルの一部を繰り返し利用することによって高い対数尤度を示してしまう．この課題は，HMM を用いた DNA の配列解析に特有の問題を含んでいる．DNA 配列はわずか 4 種類の塩基で構成されているため，ある部分配列に対して高い対数尤度を与える状態遷移と出力シンボルが HMM 中に存在する可能性が高くなる．しかしながら，モザイク構造 HMM では状態遷移に関する制約が強くなるため，塩基組成の偏りや繰り返し配列の影響を受けにくく，認識率の改善が期待される．

本手法において，状態が統合された HMM の生成過程は以下のように推察することができる．DNA 配列群に含まれる複数のシグナル配列を抽出しようとす

る場合，これらのシグナル配列には4種類の塩基がすべて含まれていることが一般的である．このため，本手法でHMMを生成する場合，各塩基を特異的に出力する状態が生成され，これらの状態はシグナル配列にしたがって高い状態遷移確率で結合される．このとき，複数のシグナル配列の5′あるいは3′末端の塩基配列に高い類似性が存在すると，状態の統合が生じることによって，状態遷移に関する自由度を大きく増加させることなく対数尤度を向上させることができる．また，配列の大部分はシグナル配列以外の領域であるため，配列全体の対数尤度を上げるには，この領域における対数尤度を向上させることが有効になる．この領域をHMMで表現する場合，シグナル配列を表現する状態は4種類の塩基について生成されているため，これらの状態を部分的に利用する状態遷移が有効となる．以上のような機構により，HMM中のある状態が複数のシグナル配列やシグナル配列以外の領域に関する同種の塩基を表現することになる．

本手法は，HMMネットワークの生成を繰り返すため，かなりの計算時間を必要とする．上に示した計算機実験の場合，HMMが8〜10状態に成長するまでにSun SparcStation20で2日程度の計算時間が必要となる．さらに，DNAシークエンシングプロジェクトの進展によって，入力に用いられる配列データセットのサイズが急速に増大しているため，より多くの計算時間を必要とする解析が求められている．一方，GAは並列性の高いアルゴリズムとして知られ，並列化GAに関する多くの研究が報告されている[24),25)]．我々は，本手法の開発にあわせて手法の並列化を検証してみた．

本手法では，HMMネットワークの生成を繰り返すためにBWAを数多く実行する必要がある．BWAは，HMMの状態数や入力データセットのサイズが大きくなると，多くの計算時間が必要になることが知られている．このため，本手法では，計算時間の大部分がBWAに費やされている．本手法は，GAを用いているため，ある世代における各HMMに対するBWAを独立に実行することが可能である．我々は，この部分を並列に実行することにより，実計算時間の短縮を試みた．**図10**に本手法における並列アルゴリズムの概要を示す．図中のNは，集団中のHMM数を表す．並列アルゴリズムの実行は，SiliconGraphics POWER CHALLENGE XL（14CPU）で行われた．

図11に並列アルゴリズムの効果を示す．ここでは，各CPU数における並列化の効果を（1CPUにおける実計算時間）/（xCPUにおける実計算時間）で表している．つまり，x個のCPUを用いた場合，理想的な並列化の効果はxになる．14CPUにおける並列化の効果は，10.1であった．このことは，アルゴリズムの並列化によって，実計算時間が良好に短縮されたことを示している．しかしながら，この並列化は，たかだか集団中のHMM数Nの並列度しか期待することはできない．前述したデータセットの増大を考慮すると，より大規模な並列度を有したアルゴリズムの研究・開発も望まれる．

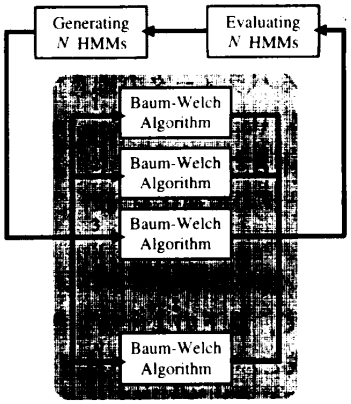

図10 アルゴリズムの並列化
Fig. 10 Parallel processing in the method.

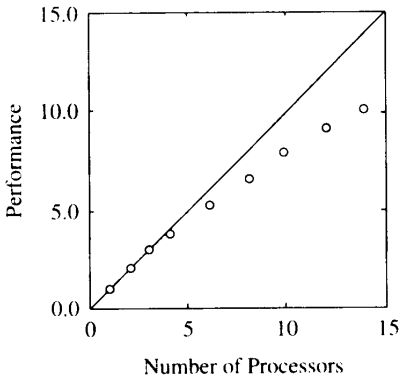

図11 並列アルゴリズムの効果
Fig. 11 Efficiency of parallel processing.

6. まとめ

我々は，DNA配列群からシグナルパターンを自動的に抽出する手法を開発した．本手法の利点は，シグナルパターンの抽出において，シグナル配列に関する先見的な知見（種類数，配列長等）を必要としないことである．

本手法では，シグナルパターンの多様性に対応するために，確率論的モデルである隠れマルコフモデル（HMM）によってシグナルパターンを表現している．HMMをシグナルパターンの表現方法として用いる場

合，以下の2点がかねてより重要な課題となっている．(1) 最も好ましいネットワークトポロジーの決定，(2) HMM に関連するパラメータの最適化．我々は，遺伝的アルゴリズム（GA）と Baum-Welch アルゴリズム（BWA）を組み合わせることによって，トポロジーの生成とパラメータの最適化を同時に行うことを実現した．

我々の手法は，他の HMM トポロジーの自動的な設計手法と比較して，以下のような特徴を有する．(1) トポロジーの生成とパラメータの最適化を GA という単一の枠組で実現している．(2) 任意のトポロジーを生成することが可能である．(3) GA によるヒューリスティックを効果的に取り入れることによって，広い探索空間に関する効率的なトポロジーの生成を実現している．

我々は，本手法を霊長類プロモータ領域に関するシグナルパターンの抽出に適用した．本手法により生成された HMM は，生物学的に重要な複数のシグナル配列を抽出していた．さらに，この HMM を用いてプロモータ領域の予測を行った結果，84.3% の精度でプロモータ領域を認識することが確かめられた．このことは，GA と BWA の組合せが HMM のトポロジーの生成とパラメータの最適化に有効であり，本手法が実際的な DNA 配列のシグナルパターン抽出問題に十分適用可能であることを示している．

今後の研究課題として，より高い精度でシグナルパターンの抽出および認識を行う HMM ネットワークの生成手法の開発と，高速に HMM ネットワークの生成を行うアルゴリズムの開発があげられる．前者に対してはシグナル配列を表現する部分 HMM 群とシグナル配列以外の領域を表現する部分 HMM 群で構成されるモザイク構造 HMM の生成手法の開発，後者に対しては大規模な並列化アルゴリズムの開発が有効であると考えられる．

謝辞 本研究を進めるにあたり，遺伝的アルゴリズムの遺伝的操作に関する有益なご助言をいただいた日本電気（株）C&C システム研究所の小長谷明彦博士に深く感謝いたします．また，研究の全般にわたり有益な議論をいただいた（株）三菱総合研究所システム科学部の長阪匡介博士，本論文の執筆にあたり有益なご助言をいただいた（株）三菱総合研究所システム科学部の市吉伸行氏に感謝の意を表します．なお，この研究の初期段階は，(財) 新世代コンピュータ技術開発機構（ICOT）において行われた[26]．

参考文献

1) GenBank: Genetic Sequence Data Bank, Release 76.0, Technical report, BBN Laboratories, U.S.A. (1993).
2) Bucher, P.: Weight Matrix Descriptions of Four Eukaryotic RNA Polymerase II Promoter Elements Derived from 502 Unrelated Promoter Sequences, *J. Mol. Biol.*, Vol.212, pp.563–578 (1990).
3) Levin, B.: *Genes V*, Oxford University Press (1994).
4) Levinson, S.E., Rabiner, L.R. and Sondhi, M.M.: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, *Bell Syst. Tech. J.*, Vol.62, pp.1035–1074 (1983).
5) Churchill, G.A.: Stochastic Models for Heterogeneous DNA Sequences, *Bull. Math. Biol.*, Vol.51, pp.79–94 (1989).
6) Lawrence, C.E. and Reilly, A.A.: An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences, *Proteins*, Vol.7, pp.41–51 (1990).
7) Cardon, L.R. and Stormo, G.D.: Expectation Maximization Algorithm for Identifying Protein-binding Sites with Variable Lengths from Unaligned DNA Fragments, *J. Mol. Biol.*, Vol.223, pp.159–170 (1992).
8) Krogh, A., Mian, I.S. and Haussler, D.: A Hidden Markov Model that Finds Genes in *E.coli* DNA, *Nucleic Acids Res.*, Vol.22, pp.4768–4778 (1994).
9) Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M.A.: Hidden Markov Models of Biological Primary Sequence Information, *Proc. Natl. Acad. Sci. U.S.A.*, Vol.91, pp.1059–1063 (1994).
10) Stultz, C.M., White, J.V. and Smith, T.F.: Structural Analysis Based on State-space Modeling, *Protein Science*, Vol.2, pp.305–314 (1993).
11) Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D.: Hidden Markov Models in Computational Biology, Applications to Protein Modeling, *J. Mol. Biol.*, Vol.235, pp.1501–1531 (1994).
12) Baum, L.E., Petrie, T., Soules, G. and Weiss, N.: A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *Ann. Math. Stat.*, Vol.41, pp.164–171 (1970).

13) Takami, J. and Sagayama, S.: Automatic Generation of the Hidden Markov Network by Successive State Splitting, *Proc. ICASSP* (1991).
14) Fujiwara, Y., Aogawa, M. and Konagaya, A.: Stochastic Motif Extraction Using Hidden Markov Model, *Proc. 3rd Int. Conf. on Intelligent Systems for Molecular Biology*, pp.121–129 (1994).
15) Tanaka, H., Onizuka, K. and Asai, K.: Classification of Proteins via Successive State Splitting of Hidden Markov Network, *Proc. W26 in the 13th Int. Joint Conf. on Artificial Intelligence*, pp.25–30 (1993).
16) Holland, J.: *Adaptation in Natural and Artificial Systems*, MIT Press (1992).
17) Goldberg, D.: *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley (1989).
18) Miller, G.F., Todd, P.M. and Hegde, S.U.: Designing Neural Networks Using Genetic Algorithms, *Proc. 3rd Int. Conf. on Genetic Algorithms*, pp.379–384 (1989).
19) Akaike, H.: Information Theory and an Extension of the Maximum Likelihood Principle, *Proc. 2nd Int. Symp. on Information Theory*, pp.267–281 (1973).
20) Konagaya, A. and Kondo, Y.: Stochastic Motif Extraction Using a Genetic Algorithm with the MDL Principle, *Hawaii Int. Conf. on System Sciences*, pp.746–755 (1993).
21) 石川幹人，十時　泰，戸谷智之，星田昌紀，広沢　誠：並列反復改善法によるタンパク質の配列解析，情報処理学会論文誌，Vol.35, pp.2816–2830 (1994).
22) Feller, W.: *An Introduction to Probability Theory and It's Applications*, 2nd edition, John Wiley & Sons (1957).
23) Prestridge, D.S.: Predicting Pol II Promoter Sequences Using Transcription Factor Binding Sites, *J. Mol. Bio.*, Vol.249, pp.923–932 (1995).
24) Dorigo, M. and Maniezzo, V.: *Parallel Genetic Algorithms: Introduction and Overview of Current Research*, IOS Press (1993).
25) 戸谷智之，石川幹人：マルチ個体群の並列遺伝的アルゴリズムを用いたタンパク質の配列解析，情報処理学会論文誌，Vol.36, pp.2549–2558 (1995).
26) ICOT Free Software No.99: DNA Sequence Analysis Using Hidden Markov Model and Genetic Algorithm, http://www.icot.or.jp/ (1994).

(平成7年9月7日受付)
(平成8年3月12日採録)

矢田　哲士

1965年生．1989年九州大学理学部生物学科卒業．同年（株）三菱総合研究所入社．現在，日本科学技術情報センター客員研究員．生物物理学会会員．計算機生物学，とくにゲノム情報解析の研究に従事．

石川　幹人（正会員）

1959年生．1982年東京工業大学応用物理学科卒業．同大学院を経て，松下電器産業（株）に入社．1989～95年（財）新世代コンピュータ技術開発機構に出向．東京工業大学，明治大学非常勤講師．第8回元岡賞受賞．人工知能学会，生物物理学会各会員．博士（工学）．知識情報処理，とくに生物学への応用に興味を持つ．

田中　秀俊（正会員）

1963年生．1986年東京大学計数工学科卒業．同年三菱電機（株）に入社．1989～95年（財）新世代コンピュータ技術開発機構に出向．データベース，遺伝子情報処理の研究に従事．

浅井　　潔（正会員）

1960年生．1985年東京大学大学院計数工学専門課程修士課程修了．同年通商産業省工業技術院電子技術総合研究所入所．博士（工学）．音声認識，遺伝子情報処理の研究に従事．