Vol.37 No.10　　情 報 処 理　　Oct. 1996

解 説

ゲノム情報

4. タンパク質の立体構造を説明するための確率モデル†

浅 井 潔††

1. まえがき

ゲノム情報は生物進化の蓄積であり，進化の過程は確率的な要素を含んでいるから，膨大な量のゲノム情報を理解するためには，決定論的なルールではなく確率的な「からくり」で説明するのが自然である．

音声認識で盛んに用いられている隠れマルコフモデル(HMM)は，1990年代にはいって分子生物学の分野で急速に広まっている．とくにマルチプル・アラインメントやモチーフ表現，DNAにおける遺伝子構造の認識などにおけるHMMの利用は目覚しいものがある．一方，RNAやタンパク質などの生体高分子の立体構造の複雑な相互作用の仕組みを説明するために確率モデルを利用する試みも注目を浴びている．これは，タンパク質のアミノ酸の1次元的な構造が，折り畳みによって階層的に高次の構造を形成する現象を，記号列上の文法による構文解析と類似の手法によって説明するものである．

本稿では，HMMを中心とした確率モデルのゲノム情報処理への応用と確率文法によるタンパク質立体構造のモデル化について述べる．

2. 隠れマルコフモデル(HMM)

工学的に用いられるHMMは，マルコフ過程の各状態(または状態遷移)に情報の出力確率としての確率分布を割りあてたものである．ゲノム情報の解析では，4種類の塩基や20種類のアミノ酸の記号が出力されると考える．HMMはマルコフ過程のパラメータである状態数N，状態間の遷移確率の行列$A = a_{ij}$，時刻0に各状態にいる初期確率π_iのほかに，各状態の出力確率関数$f_i(x)$(または各状態遷移の出力確率関数$f_{ij}(x)$)を与えることにより決定される．HMMは，確率文法の立場から解釈すると，確率的正規文法である．すなわち，正規文法の各導出規則のそれぞれに確率を付与し，

$$S_i \rightarrow x S_j : p_{ij}^x \tag{1}$$

とするとHMMになる(S_i, S_jは非終端記号，xは終端記号)．前の定式化のパラメータとの間には，$p_{ij}^x = a_{ij} f_{ij}(x)$ の関係がある．

出力信号のデータからHMMのパラメータを学習するアルゴリズムとHMMを用いてデータを構文解析するアルゴリズムは知られているから，20種類のアミノ酸を出力するHMMは，タンパク質の配列解析に用いることができる．筆者は1991年頃からタンパク質モチーフ(構造や機能上意味のあるパターン)の表現とタンパク質の局所構造である2次構造予測に用い始めた[1]．UCSCのHausslerらは，タンパク質のマルチプル・アラインメントを，図-1に示すようなHMMを用いて行った[2]．マルチプル・アラインメントとは，何本かの遺伝的配列を配列の途中にギャップと呼ばれる隙間を挿入することによって，なるべく類似の記号が同じカラムに並ぶように整列させることである．たとえばADHE，AHIE，ADIEの3本の配列を，

```
ADH-E
A-HIE
AD-IE
```

のように整列させるわけである．

任意の2つの記号i,jが同じカラムに並んだときのスコアM_{ij}を定義すると，すべてのカラムで

† Stochastic Models for the Explanation of Tertially Structures of Protein by Kiyoshi ASAI (Genome Informatics Group, Electrotechnical Laboratories).

†† 電子技術総合研究所知能情報部遺伝子情報グループ

のスコアの合計が最も小さくなる最良の整列はDP(Dynamic Programming)によって厳密に求めるられるが，配列平均の長さがL，配列の本数がnのとき計算時間が$O(L^n)$となるから，本数の多いマルチプル・アラインメントを求めることは難しい．各配列を用いて学習されたHMMに対して，各配列データのHMMにおける最適状態遷移列を求めれば，結果として各配列のマルチプル・アラインメントが得られるが，計算時間は$O(nL^2)$程度ですむ．田中らは，HMMによるマルチプル・アラインメントはBerger-Mansonの逐次法と本質的に同じものであることを明らかにした[3]．

2.1 HMMのネットワーク形状決定法

HMMパラメータは，学習アルゴリズムによって，局所最適な値を得ることができるが，HMMのネットワーク形状はあらかじめ与える必要がある．しかし，最適なHMMのネットワーク形状をあらかじめ知ることは難しい．ゲノム情報処理においては，遺伝配列のもつ構造に応じたHMMのネットワークの設計が必要であり，構造予測とも関係しているので，ネットワーク形状を決定する手法の研究が行われている．

田中らは，タンパク質のアミノ酸配列のモデル化のためにSSS(Successive State Splitting)アルゴリズムという連続分布HMMのネットワーク形状決定法を用いた[4]．これは，1状態から始め，出力分布の分散のもっとも大きい状態を次々に分割しながらネットワークを成長させていく方法である．藤原らは，タンパク質モチーフ抽出をする際，HMMのネットワーク形状の決定法として，ID(Iterative Duplication)法を開発し[5]，図-2のようなHMMを構成した．この方法では，もっとも結合数の多い状態を分割しすべての結合関係をコピーする．そのあと学習データでHMMのパラメータを学習し，弱い(遷移確率の低い)結合を削除することによって，任意の形のネットワークを形成することができる．矢田らは，DNA配列のモデル化を行うのに，GA(Genetic Algorithm)を用いてHMMのネットワークを最適化した[6]．HMMのネットワークは，図-3にあるように結合行列を1次元に展開して疑似遺伝子にマッピングされる．突然変異や交差は，HMMの状態の挿入・削除や，ネットワークの部分的な入れ換えに

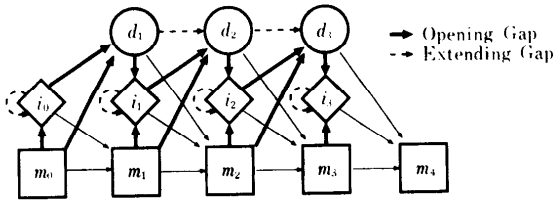

図-1 HMMによるマルチプル・アラインメント．この図ではモデルの長さは4だが配列の長さによって，もっと長いモデルを使う．

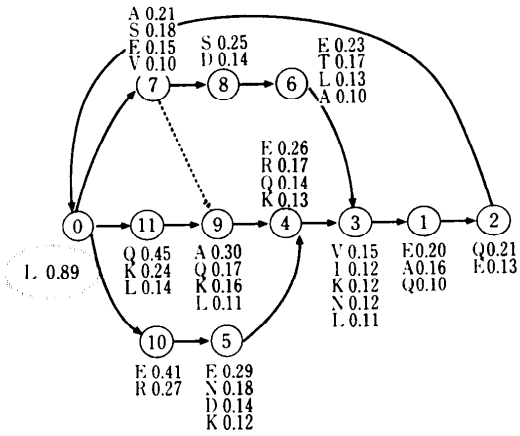

図-2 Leucine ZipperのHMMによる表現

対応している．

3. タンパク構造文法

タンパク質の局所構造である2次構造をアミノ酸配列から予測する精度は，近年改善されて70％を越えた．しかし，タンパク質は複雑に折れ曲がっているから，1次構造上は離れたアミノ酸が，空間的には近くにあって相互作用を生じる(遠距離相互作用)．局所構造である2次構造も，長距離相互作用まで考えないと，正確には決定できないと考えられる．したがって，全体の整合性をみて，局所構造と大局構造を同時に最適化するような手法が必要である．

近年流行している3D－1D法では同時最適化の困難を避け，すでに立体構造のわかっているタンパク質のテンプレートに，アミノ酸配列を並べ，その整合性を比較し，最適のテンプレートを検索する．整合性は，立体構造上の環境(タンパク質の表面にあるか，内部にあるかなど)に基づくスコアと，統計に基づくアミノ酸ペアの遠距離相互

作用のスコアを合計して求める．構造のわかっているタンパク質とアミノ酸配列の類似度の高いタンパク質のマルチプル・アラインメントがあれば，構造上の位置ごとの頻度分布を整合性のスコアとして用いることもできる．同様のことを部分構造について行うことは，以前から行われてきたホモロジーモデリングに相当する．この立場からテンプレートを作るのには，HMM をモデルとして用いればよい．

しかし，HMM では，遠距離相互作用を表現できないから，3D − 1D で用いるような立体構造上の距離に基づく整合性のスコアを実現するためには，より高次の確率モデルが必要である．局所構造が決まらないと大局構造が決まらず，大局構造がわからないと局所構造が予測できないというタンパク質立体構造予測の問題は，連続音声認識の問題と類似している．発声された文章の単語の並びがわからなければ，構文解析や意味の理解ができないが，局所的な音声信号だけから単語を精度よくあてることは難しい．多くの連続音声認識システムでは，局所的な情報からは単語のスコア(確率)を求め，文法規則に従って構文解析をしながらスコアの合計によって認識結果の候補を求めるという方式をとっている．タンパク質立体構造の構文解析を行うためには，自然言語における文法にあたるタンパク質立体構造の相互作用の規則(タンパク質構造文法)を作らなければならない．それにより，対象そのものはまったく異なるにもかかわらず，動的な構文解析や枝刈りをしながらのビームサーチなどの連続音声認識システムの手法を，タンパク質の構造予測問題に用いることができる．

3.1 確率文脈自由文法(SCFG)

HMM はマルコフ過程(正規文法)に基づいているから，配列上隣接した場所の関係についてしか表現できない．文脈自由文法の各導出規則に確率を与えて得られる確率的文脈自由文法(SCFG)を用いることによって，ある種の遠距離相互作用が表現できる．Chomsky の標準形だと，

$$S_i \rightarrow S_j S_k \quad : p_{ijk} \tag{2}$$

$$S_i \rightarrow x \quad : p_i(x) \tag{3}$$

のようになる．RNA の 2 次構造ならば，一般形で

$$S_i \rightarrow a S_j b \quad : p_{ij}(a, b) \tag{4}$$

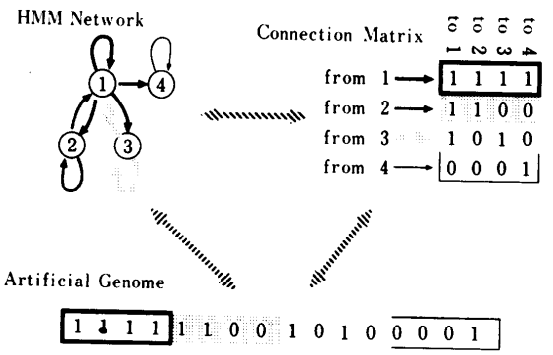

図-3 HMM ネットワークの疑似遺伝子へのマッピング

$$S_k \rightarrow c S_l \quad : p_{kl}(c) \tag{5}$$

$$S_m \rightarrow S_n S_o \quad : p_{mno} \tag{6}$$

のように書くと，塩基対によるステム構造をうまく表現することができる．榊原らは，SCFG を用いて，tRNA の構造をモデル化し，2 次構造の予測と，tRNA の認識を行った[7]．

3.2 高次の確率文法と近似構文解析

遠距離相互作用のうち複雑なものは SCFG によって表現することはできない．RNA の大部分の 2 次構造は SCFG によって表現できるが，pseudo-knot といわれる構造は表現できない．タンパク質の 2 次構造においても，配列上 4 個先のアミノ酸残基と相互作用する α ヘリックスや，並行な β シート間の相互作用，3 本以上の β シート構造などは SCFG では表現できない．これらの遠距離相互作用を表現するには，より高次の文法が必要であり，確率木文法を用いたモデル化が提案されている[8),9)]．これらの高次の文法については，ここでは詳しく触れないので「計算論的言語理論と DNA 計算」の方をご覧いただきたい．

高次の確率文法の問題点は学習および構文解析の計算コストが高いことである．配列の長さを L としたとき構文解析に必要な計算時間は，一般の HMM では $O(L)$，一般の SCFG では $O(L^3)$ 程度になる．確率文脈自由文法を超えるクラスの確率文法の構文解析には，最低でも $O(L^4)$ の計算時間がかかると考えられる．しかも，タンパク質の立体構造は，α ヘリックス 1 本や 3 本の β シート構造だけといった単純なものではなく，これらの相互作用が多数同時に存在するから，相互作用全体を説明できる確率文法の構文解析のコストは非常に高い．

そこで，確率文法のクラスをできるだけ低次におさえ，それらの組合せと近似によって，複雑な相互作用を説明することが考えられる．UCSC の Brown は，2 つの SCFG を組み合わせて，RNA の pseudo-knot 構造の構文解析を逐次的に近似する手法を考案した[10]．SCFG において離れた 2 カ所の位置の同時確率分布を表現することができるのは，限られたパターンの場合だけである．SCFG から外れたパターンの 2 カ所の相互作用は Brown の手法を応用すれば近似的に構文解析することができる．3 カ所以上の同時確率分布を考える必要のある場合(3 本以上の β シート構造など)も SCFG では表現できないが，立体構造のデータの数からいって，3 カ所以上の同時確率分布を推定することはもともと困難なので，2 カ所の同時確率分布の組合せで表現することはそれほど悪い近似ではない．

4. ゲノムデータの偏りと不足

確率モデルを用いてゲノムデータを解析する場合，大きな問題が 2 つある．1 つ目は，遺伝的データは，非常に偏っているということである．遺伝的配列や構造の決定は，興味深いもの，実験的にやさしいものは数多く得られるが，困難なものは後回しにされている．さらに，生物が進化によって関連づけられている以上，いかなる遺伝的配列もデータとして独立ではあり得ない．データの偏りを除くため，類似度の高いデータは除いて用いることもあるが，除かれたデータの情報を捨ててしまうことになる．そこで，データに重みをつける手法がいくつか提案されている．似たデータがたくさんある場合には小さな重みを，類似のデータがない場合には大きな重みを与える．ただし，モデル化しようとしているカテゴリーからはずれたデータが混じっている場合にそれに不当に大きな重みを与えてしまうという危険がある．

もう 1 つの問題はデータの不足である．たとえば，タンパク質配列のデータは多くあっても，立体構造の判明したタンパク質のデータは極端に少なく，統計的に十分なデータが利用可能とはいえない．確率モデルの出力分布について，あらかじめ知識があれば，それを利用してデータの不足を補うことができる．アミノ酸配列のアラインメントに用いるスコア行列も，アミノ酸の置換の頻度についての知識を利用したものである．片方の配列のアミノ酸が i であるときにもう一方の配列(テスト配列)の対応するアミノ酸が j である確率を $P_j(i)$ とし，テスト配列の任意の位置でアミノ酸 j が現れる確率を $P_0(i)$ とする．スコア行列 M_{ij} は，

$$M_{ij}=\log\frac{P_j(i)}{P_0(i)}. \tag{7}$$

を多くの類似のタンパク質について統計をとった結果得られたものである．

HMM でマルチプル・アラインメントを行う場合，単純に出力分布を推定すると，置換しやすいアミノ酸が同一のカラムに揃わない．そこで，各カラムの頻度ベクトルに M_{ij} をかけたものを頻度ベクトルにみたてて出力分布を推定すれば，スコア行列を用いたマルチプル・アラインメントと同様の結果が得られる．ところが，スコア行列は 2 本の配列の整列のために考案されたもので，マルチプル・アラインメントや分布の推定法としては性能が悪い．たとえば，あるカラムの真の出力分布が，1 種類のアミノ酸しか現れないものだったとしても，データ数∞の極限で推定値が異なったものになってしまう．Brown らは，Dirichlet 型の混合分布を事前分布として用い，HMM によるマルチプル・アラインメントを改善した[11]．同様の手法は SCFG およびより高次の確率モデルの分布の推定にも応用可能である．

5. むすび

タンパク質の立体構造における相互作用は非常に複雑で，しかも立体構造についてのデータは限られているから，それを確率モデルで表現するのは困難な課題である．しかし，本稿で述べたように，高次の確率モデルや近似を用いた高速な構文解析，データの重みづけや事前分布の利用を組み合わせることによって，精密なモデル化が可能になると筆者は考えている．今後のこの分野の発展を期待したい．

参考文献

1) 浅井，速水，半田：確率モデルによる遺伝子情報処理，Genome Informatics Wrokshop II(1991).

2) Haussler, D., Krogh, A., Mian, I. S. and Sjölander, K. : Protein Modeling using Hidden Markov Models:

Analysis of Globins, Proceedings of tne Hawaii International Conference on System Sciences, Vol.1, pp.792-802, IEEE Computer Society Press (1993).

3) Tanaka, H., Ishikawa, M., Asai,K. and Konagaya, A. : Hidden Markov Models and Iterative Aligners:Study of their Equivalence and Possibilities, Proceedings, First International Conference on Intelligent Systems for Molecular Biology (ISMB93), pp.395-401, AAAI Press (1993).

4) Tanaka, H., Onizuka, K. and Asai, K. : Classification of Proteins via Successive State Splitting of Hidden Markov Network, Proceedings of Workshop on Artificial Intelligence and Genome in IJCAI93 (1993).

5) Fujiwara, Y., Asogawa, M. and Konagaya, A. : Stochastic Motif Extraction Using Hidden Markov Model, Proceedings, Second International Conference on Intelligent Systems for Molecular Biology (ISMB94), pp.121-129, AAAI Press (1994).

6) Yada, T., Ishikawa, M., Tanaka, H. and Asai, K. : DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm, Proceedings of Genome Informatics Workshop V (1994).

7) Sakakibara, Y., Brown, M., Hughey, R., Mian, S., Sjölander, K. , Underwood, R. and Haussler, D. : Stochastic Context-free Grammars for tRNA Modeling, Nucleic Acids Research, Vol.22, No.23, pp.5112-5120(1994).

8) Mamitsuka, H. and Abe, N. : Predicting Location and Structure of Beta-Sheet Regions Using Stochastic Tree Grammars, Proceedings, Second International Conference on Intelligent Systems for Molecular Biology (ISMB94), AAAI Press (1994).

9) Kobayashi, S. and Yokomori, T. : Modeling RNA Secondary Structures Using Tree Grammars, Proceedings of Genome Informatics Workshop V (1994).

10) Brown, M. : RNA Pseudoknot Modeling Using Intersections of Stochastic Context Free Grammars with Applications to Database Search, Proceedings of Pacific Symposium on Biocomputing 96 (1996).

11) Brown, M., Hughey, R., Krogh, A., Mian, I.S., Sjölander, K. and Haussler,D. : Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families, Proceedings, First International Conference on Intelligent Systems for Molecular Biology (ISMB93), pp.47-55, AAAI Press (1993).

(平成8年8月5日受付)

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, 現在に至る. この間, 1993 ~ 1995年, ICOTタンパク質立体構造予測ワーキンググループ主査. 1995 ~ 1996年通商産業局機械情報産業局電子機器課に併任. 現在, 遺伝情報グループで確率モデルを用いたタンパク質構造, DNA情報の解析の研究に従事. RWC分子生物情報ワークショップ主査. 主任研究官, 博士(工学).