確率モデルによる遺伝子情報処理

Genetic Information Processing by Stochastic Model

HMMを用いたタンパク質の2次構造予測

HMM for Secondary Structure Prediction of Protein

浅井 潔 速水 悟 半田剣一 （電子技術総合研究所）
Kiyoshi ASAI Satoru HAYAMIZU Kenichi HANDA （Electrotechnical Lab.）

It is maintained that active use of a stochastic model is essential for analyzing the amino acid sequence of protein or the base sequence of DNA. Here we shall discuss the amino acid sequence of protein. Using HMM, which is a type of stochastic model, an attempt was made to predict the secondary structure of protein. In this process, secondary structures such as helix, sheet and turn are each learned by discrete HMM, and the output probabilities from the stochastic model is used to determine which part of the sequence (whose structure is unknown) corresponds to which structure. With a model in which one amino acid is taken as the output symbol, favorable results could not be obtained, but when two consecutive amino acids were used as the output symbol, good results could be obtained in estimating helix and sheet.

0. まえがき

確率モデルの一種であるHMMを用いて、タンパク質の2次構造予測を試みている。これは、ヘリックス、シート、ターンなどの構造をそれぞれHMMに学習させ、構造が未知の配列に対し、どの部分がどの構造に対応するかを、確率モデルからの出力確率によって推定しようとするものである。

1. 確率モデルによる遺伝子情報処理

タンパク質のアミノ酸配列や、DNAの塩基配列の解析には確率モデルの活用が必要であると考えている。ここでは、タンパク質のアミノ酸配列について述べる。

従来の相関性検査や、モチーフ抽出では、確定的な共通配列や、アミノ酸間の距離の積み上げ（DPなど）による比較が行われている。しかし、確定的な共通配列には、モデルとしての柔軟性がない。より広い意味でのモチーフの表現などは、確率モデルの導入が必要である。

また、ある配列を、特定の配列クラスターと比較する場合には、それぞれの配列と比較するのに時間がかかるし、平均的な代表配列を求めておいてそれと比較するのは少し雑である。クラスターの特徴を、確率モデルに学習させておいて、その確率モデルに対する適合度を考えるのが簡単である。

さらに、構造や機能が既知のデータによって確率モデルを学習すれば、構造や機能が未知のデータがモデルから生成される確率を計算することにより、構造や機能の推定を確率的に行うことができる。

2. 音声認識とタンパク質の2次構造予測

筆者らは、確率モデルの一種であるHMMを用いて、音声認識システムを構築してきた。音声の波形データは10-30 msec単位で周波数分析され、スペクトル、スペクトル変形、ベーク等方式を256-1024程度の大きさの符号値でベクトル量化化されて離散符号になる。音声再構においては、単語のモデルはHMMである音素のモデルと音素の結合順序を規定した辞書から成っている。音素音モデルのHMMは4個程度の状態を持ち、状態遷移の確率を、状態遷移の際電子的に離散符号を出力する分布関数を持っている。このようなにHMMを組み合わせた単語モデルに対して、効率よくデータを学習するアルゴリズムと、特定の離散符号列が、そのモデルから出力される確率を計算するアルゴリズムが知られているので、未知のデータがどの単語であるかを、確率的に認識することができる。

音声認識と2次構造予測にはいくつかの類似性が見られる。

タンパク質は、符号値の大きさが20（アミノ酸の種類）の記号列であり、その1部分が、ヘリックス、シート、ターン等の2次構造を持っている。その様子は、連続するいくつかの離散符号列が音素になっていることと類似して
いる。また、2次構造が組合わさってタンパク質の立体構造をつくる様子は、音素の繋がりが単語を構成する様子に似ている。

3. HMMによるタンパク質の2次構造予測

3-1. 分布のエントロピー

HMMが確率モデルとして有効であるためには、それぞれの出力確率が、適度に異なったビークを持っている必要がある。

音声認識の場合、離散符号化されたデータを用いて分布のエントロピーを計算することにより、それぞれの音素の分布の集中度を知ることができる。

タンパク質の場合の分散度を見るために、ヘリックス、シート、ターンの場合の分布を、全体の分布と比較してみた。勿論、これらの数値は、より多くのデータを使って、すでに生物学者の手によって計算されているであろう。この値は、2次構造予測をするのに、十分な値といえるであろうか？認識実験の結果では、十分ではないのである。

次にすべての2文字連続に関して、組合わせに分布を調べてみると、強い相関があることがわかる。周期性のある構造に対応できるように、1つの飛び、2つの飛び、ごとの2文字についても同様に考えることができる。3文字以上の相関については、手持ちのデータ数が不十分で、残念ながら、意味のある統計はとれない。

3-2. VQコードとしての20アミノ酸

20種類のアミノ酸は、それぞれ物理的、化学的な特徴を持っている。それらの特徴のうち、目的に合致した特徴を数値化して表現したとしても、それをVQすると、20個以上には決して分離できないし、互いが特徴的に持ったとしても、実は無意味であることはすぐに分かる。1個のアミノ酸で考慮する限り、互いの大きさは決して20個以上にできない。VQとおもな特徴をさらに減らすことも無意味である。

2文字連続をVQコードと考えると、組合わせの大きさは400となる。3文字連続で8000である。これらをそのまま使うとさらに量子化すれば、適当な大きさの符号組を得ることができる。また、1つの飛び、2つの飛び、ごとの2文字でも同様のことができる。

20種類のアミノ酸をそのまま出力記号として用いたとしても、隣接相関や周期性はHMMのネットワーク、遷移確率、出力確率によって表現することができる。しかし、隣接相関や、周期性それ自体に意味があるなら、それをVQコード化する時点で取り込めば、より良いモデルを作成する可能性がある。

3-3. 2次構造予測の方法

現在98のPDBデータを用いて、2次構造予測の計算機実験を行っている。

ヘリックス、シート、ターン、その他の4種類のモデルをHMMで表現し、PDBデータに付いている、HELIX、SHEET、TURNのラベルを用いて、それぞれのモデルを別々に学習させる。HMMのネットワークは、自動的に生成して、学習結果の精度がもっとも高くなるものを採用した。

単語音声認識の場合や単語モデルは、音素のモデルのHMMを辞書に収めた通りの順序にないのでそれを用いるが、未知の構造のタンパク質は、どのような順序で、1次構造が現れるかわからないので、この方法はそのまま使うことはできない。それぞれの組み合わせ（未知の音節列から成る）未知の単語の音声データから、音節列を推定する問題に相当する。音素の数は100以上であるので、この問題は非常に難しい。

しかし、2次構造予測をする場合に、ここでは構造の種類を4種類しか設定していないから、4種類のHMMを順序を規定しない形でうまく組み合わせれば実用的なネットワークを作ることができる。

Fig.1に一例を示す。Fig.1-aにあるようなother、helix、sheet、turnの4つのHMMをまず独立に学習させて遷移確率と出力確率を求めておく。その後、4つのHMMの最初の状態（状態0）と最後の状態（状態5または6）をすべて統合して、新たに状態0とし、Fig.1-bのよう1つのHMMを合成する。遷移確率と出力確率は、新たに合成されたHMMにそのまま引き継がれる。

未知のアミノ酸列に対して、Fig.1-bのHMMはさまざまな状態遷移の順序を使ってそのアミノ酸列と同一の出力を出すことができるが、その出力確率が最も高いなるような状態遷移の順序を求めることができる。それぞれの状態がFig.1-bの中で、元々4種類だったHMMのどの部分にあるかで、4種類の構造とそれぞれの状態の対応がとれるから、その状態遷移の列から未知のアミノ酸配列の2次構造が出力される訳である。Fig.2に出力の例を示す。
Fig. 1-a HMMs of secondary structures.  
Fig. 1-b Combined HMM for prediction.

---

STAGKVIKCAAVLWEEKKPFIESVEVAPFKAKHEVRIKMWATGCIRSDDDHVSGTLVPLVIAGHEAAAGIVESIGEV

76666666601221222222222220777766601089076666076666533440777777666660107760766607680907

TTVRPGDEVILPLFTPGCGKGVKLFPGNCLNDLSMPGTM0DGTGRTGKFHPPFHGTGYSQTVDEISVAKI

776088076660111122222222222220345508901108901078090766608901108907608903453

DAAASPFLKVCILGGCGFSTGCSAVKAVAQTGTSCAVFGCVGVCVIMRSCKAAGAARIG VINDDKFKAKAKEVGATOC

3401080344555508907776666070890776660890345555534453453453345344455340890345

VNPQDYKKFQIVLTEMNSGCVDFSVEVGELDTTMVATLSCQCEAYGVSVIVGVPDQSGLMNPMALLLSGERTWGATFC

30108903445534508901077666034555534555555507776666034555533334550107766608

GFKSKDSVPKILADPHAXKAPDLPILTHVLPFXINEQFDPLLRSGESIRTILT

301089034553333450890890760890345555507766660766607666607680907

---

Fig. 2  Secondary structure prediction result (2 letter case).
Data file is 6ADHA.pdb. 1st line = amino acid, 2nd line = structure label in the database, 3rd line = predicted structure, 4th line = HMM state No., where h: helix, s: sheet, t: turn, u: unlabeled.
3-4. 計算機実験の結果

(a) アミノ酸1個をHMMの出力記号として用いた場合
HELIK、SHEET、TURNの個々のモデルは、それなりにパターンを学習する。自動的に生成したネットワークのうち、学習の結果、出力確率が大きく、かつ、他のパターンの出力確率が小さく（分離度が良いことを意味する）なるものを選び、予測に用いた。
そのうち、HELIK、SHEET、TURNのHMMはそれぞれ4状態、アークの数を順に8、6、6、分布の数を順に5、3、2で実験した結果をTable 1に示す。これは、すべてのタンパク質を学習に用いたクローズドの結果であるが、残念ながら、实用には耐え難しい。

(b) 連続する2個アミノ酸の組を出力記号として用いた場合
連続する2個のアミノ酸の組を1つの離散記号だとみなし、アミノ酸を1つずつずらしながらコード化する。つまり、例えばアミノ酸の系列がPPKAHEならば、「PP」、「PK」、「KA」、「AH」、「HE」の5つの記号だとみなすものである。
Fig.1のHMMを用いて認識した結果をTable 2に示す。ターンの構造推定は余りうまくいっていないが、ヘリックス、シートの推定はうまくいっていることがわかる。

4. 結論および今後の展望

HMMを用いて、タンパク質の2次構造予測を行なった。アミノ酸1個を出力記号とするモデルでは良い結果は得られなかったが、連続するアミノ酸2個を出力記号として用いた場合には、ヘリックスとシートの推定に関して良好な結果を得た。
出力記号の選択については、1つ飛び、2つ飛び、……、n個のアミノ酸についても検討する必要がある。また、より多くのデータで、3個のアミノ酸の組についても検討してみたい。
タンパク質の種類や機能によって、モデルを複数用いる方法も有力である。それにより、どのモデルが最も高い出力確率を出すかによってタンパクの種類や機能を推定し、同時に各モデルの確率から2次構造推定もできる訣である。

参考文献
(3)K.Asai,S.Hayamizu: "Dividing the Distributions of HMM and Linear Interpolation in Speech Recognition", ICASSP 92 (scheduled).

Table 1. Prediction stats(1-letter case).

<table>
<thead>
<tr>
<th>data</th>
<th>guess</th>
<th>h</th>
<th>s</th>
<th>t</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>84.4</td>
<td>7.3</td>
<td>6.6</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>55.3</td>
<td>28.0</td>
<td>13.7</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>55.0</td>
<td>11.6</td>
<td>30.5</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>57.9</td>
<td>13.6</td>
<td>21.3</td>
<td>8.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Prediction stats(2-letter case).

<table>
<thead>
<tr>
<th>data</th>
<th>guess</th>
<th>h</th>
<th>s</th>
<th>t</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>84.3</td>
<td>6.2</td>
<td>5.9</td>
<td>3.6</td>
<td>78.4</td>
</tr>
<tr>
<td>s</td>
<td>18.0</td>
<td>61.0</td>
<td>7.1</td>
<td>13.9</td>
<td>21.2</td>
</tr>
<tr>
<td>t</td>
<td>21.1</td>
<td>13.1</td>
<td>52.7</td>
<td>13.1</td>
<td>23.3</td>
</tr>
<tr>
<td>x</td>
<td>24.1</td>
<td>18.6</td>
<td>13.6</td>
<td>43.7</td>
<td>25.4</td>
</tr>
</tbody>
</table>