確率的木文法学習を用いたタンパク質二次構造予測

1P－9

馬見塚　拓　安倍　直樹
NEC　C&C研究所

1　はじめに

タンパク質立体構造において、頻繁に見られる規則正しい構造を二次構造と呼ぶ。与えられたタンパク質アミノ酸配列に対し、二次構造領域を予測する問題は、タンパク質の立体構造ひいてはその機能をも理解するための重要な1ステップと考えられており、1960年代より扱われてきた歴史の古い問題である。しかし、いずれの既存手法も局所領域のみからの予測を試みており、3状態(αヘリックス、βシート、その他)予測で60%前後の予測率を挙げるにとどまっている[2]。今後、この二次構造予測の予測率を向上させるためには、一般に配列内の遠距離相互作用の考慮が不可欠とされている。

ここで、代表的な二次構造の一つであるβシートは、βストランドと呼ばれる局所構造が平行に水素結合を保持した構造を指す。特に、βストランドの向きが互い違いに並んだ構造を反平行βシートと呼ぶ。

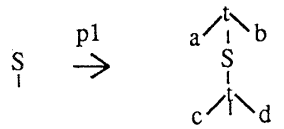
本稿では、βシート(特に反平行βシート)領域を高い精度で予測することを目指し、βストランド内の相対するアミノ酸残基同士の遠距離相互作用を直接扱う新しい木文法システム、及び、その学習・構文解析方法を提案する。さらに、実際のタンパク質データに対する適用例を示す。

2　木文法システム

本稿で用いる木文法システムは、安倍[1]によって提案されているランク付き書き換え文法(Ranked Node Rewriting Grammar (以下、RNRG))と呼ばれるものに、拡張および制限を加えたものである。

RNRGとは、初期木と呼ばれる木構造の集合と、ランク付けされた非終端記号を終端記号および非終端記号によりラベル付けられた同ランクの部分木構造により書き換える規則からなる。ここで、木構造中のノードおよび部分木構造のランクとは、構造外にある子供につながるエッジの数である。RNRGの生成する言語とは、初期木から有限回の書き換え規則の適用により生成される木構造の末端の文字列の集合を指す。

本稿では、RNRGの書き換え規則の各々に適用確率を付加した確率的RNRGに拡張し、また効率化のために文法中のいずれの書き換え規則にも現れる非終端記号は唯一箇所に制限しており、その非終端記号はSとする。なお、ランク1の書き換え規則の例を図1に示す。



S：非終端記号
a,b,c,d：終端記号

図1: 書き換え規則の例

アミノ酸配列への適用においては、終端記号をアミノ酸とみなす。また、複数の類似した規則を一つの規則にまとめるため、各末端ノードにおいて、終端記号であるアミノ酸が確率的にラベル付けられるものとする。

3　学習・構文解析方法

3.1　学習

確率的RNRGの学習は、確率的文脈自由文法や確率的ツリーアジョイニング文法の規則適用確率学習方式[3]の拡張により行える。ただし、本稿では、対象となるRNRG文法に制限を加えているため、一般のRNRGの学習に比べて簡単になっている。以下、ランク1の場合に限って説明する。

与えられたアミノ酸配列σの長さをNとする時に、σの1からi番目、jからk番目、lからN番目のアミノ酸部分列に、末端ノード列が一致するすべての木構造の生成確率の和を内側確率$In[i,j,k,l]$とする。さらに、σのiからj番目、kからl番目のアミノ酸部分列に、末端ノード列が相当するすべての木構造の生成確率の和を外側確率$Out[i,j,k,l]$とする。

ここで、r番目の規則の適用確率を$T(r)$、規則右辺の木構造中の唯一の非終端記号の左上、左下、右下および右上の末端ノードの数を各々$n_f^r(f=1,\cdots,4)$、位置fでのx番目の末端ノードでのアミノ酸αの生成確率を$P_{f,\alpha}^{r,x}(\alpha=1,\cdots,20)$とし、また、$\sigma$の$t$番目のアミノ酸を$\sigma_t(\in[1,\cdots,20])$とする。

内側確率・外側確率の初期化

内側・外側確率の各添え字が、規則を一回のみ適用した木構造の末端ノードに対応する場合に、それらの生成確率の和を各々内側・外側確率の初期値とする。

内側確率の算出

For $i := 1$ to N
 For $j := N$ to i
  For $k := j$ to N
   For $l := N$ to k

$$In[i,j,k,l]=\sum_r\{T(r)\times In[i',j',k',l'] \times\prod_{x=1}^{n_1^r}P_{1,\sigma_{i-x}}^{r,x}\prod_{x=1}^{n_2^r}P_{2,\sigma_{j+x}}^{r,x}\prod_{x=1}^{n_3^r}P_{3,\sigma_{k-x}}^{r,x}\prod_{x=1}^{n_4^r}P_{4,\sigma_{l+x}}^{r,x}\} \quad (1)$$

(ただし,$i'=i-n_1^r, j'=j+n_2^r, k'=k-n_3^r, l'=l+n_4^r$)

Protein Secondary Structure Prediction Using Stochastic Tree Grammer Learning
Hiroshi Mamitsuka and Naoki Abe
C&C Research Labs. NEC Corp.

外側確率の算出

For $i := N$ to 1
 For $j := i$ to N
  For $k := N$ to j
   For $l := k$ to N

$$Out[i,j,k,l] = \sum_r \{ T(r) \times Out[i',j',k',l'] \times \prod_{x=1}^{n_1^r} P_{1,\sigma_{i+x}}^{r,x} \prod_{x=1}^{n_2^r} P_{2,\sigma_{j-x}}^{r,x} \prod_{x=1}^{n_3^r} P_{3,\sigma_{k+x}}^{r,x} \prod_{x=1}^{n_4^r} P_{4,\sigma_{l-x}}^{r,x} \}$$

(ただし, $i' = i+n_1^r, j' = j-n_2^r, k' = k+n_3^r, l' = l-n_4^r$)

適用確率・アミノ酸確率分布の算出

r 番目の規則の重み付き頻度を $W_R(r)$、r 番目の規則の位置 f における x 番目のノードでのアミノ酸 $\alpha \in [1,\cdots,20]$ の重み付き頻度を $W_A^{r,f,x}(\alpha)$ と書く。

ここで、内側・外側確率の各添え字 i,j,k,l において、以下の Pr を算出する。

$$Pr[r,i,j,k,l] = In[i,j,k,l] \times T(r) \times Out[i',j',k',l'] \times \prod_{x=1}^{n_1^r} P_{1,\sigma_{i+x}}^{r,x} \prod_{x=1}^{n_2^r} P_{2,\sigma_{j-x}}^{r,x} \prod_{x=1}^{n_3^r} P_{3,\sigma_{k+x}}^{r,x} \prod_{x=1}^{n_4^r} P_{4,\sigma_{l-x}}^{r,x}$$

(ただし, $i' = i+n_1^r+1, j' = j-n_2^r-1, k' = k+n_3^r+1, l' = l-n_4^r-1$、また、ランク 0 の規則に対しては Pr の算出が異なる)

W_R は、Pr より以下のように算出される。

$$W_R(r) = \sum_i \sum_j \sum_k \sum_l Pr[r,i,j,k,l]$$

また、W_A は以下のように算出される。($f = 1$ の時の例を示す。)

$$W_A^{r,1,x}(\alpha) = \sum_i \sum_j \sum_k \sum_l \sum_{\sigma_{i+x}=\alpha} Pr[r,i,j,k,l]$$

頻度 W_R、W_A より、新しい規則適用確率 $T(r)$、アミノ酸の確率分布 $P_{f,\alpha}^{r,x}$ を以下のように算出する。

$$T(r) = \frac{W_R(r)}{\sum_r W_R(r)}$$

$$P_{f,\alpha}^{r,x} = \frac{W_A^{r,f,x}(\alpha)}{\sum_\alpha W_A^{r,f,x}(\alpha)}$$

以上の手順を、適用確率の変化が一定値以下になるまで反復することにより、適用確率、アミノ酸確率分布の学習を行う。

3.2 構文解析

学習により得られた規則を使用し、与えられたテストアミノ酸配列のどの領域が反平行 β シートであるかの予測を構文解析により行う。構文解析は、テスト配列に対して最も高い尤度が得られる書き換え順序の決定に相当し、β シート領域に対応する規則が生成するアミノ酸列の部分を β シート領域と予測する。

このために、構文解析では、学習における内側確率の算出式(1)において、$\sum$ を max とし、最も高い尤度を与えた規則を順次記憶しておく。

4 適用例

立体構造既知のタンパク質、リゾチーム (hen egg lysozyme、残基数 129 : 以下、1LYM) に対し、同一名を持つ他動物種の 28 本のアミノ酸配列を SWISS-PROT データベースから抽出した。

1LYM では、残基番号で、42~46、50~54、57~60 という 3 領域のストランドにより反平行 β シートが構成されている。そこで、28 本の配列をアミノ酸種類ができるだけ一致するように並べ、ほぼ、1LYM の残基番号 42~60 に相当する領域を学習データとして切り出した。各データの長さは、最短で 18、最長で 20 である。

学習により構成された規則の一例を、図 2に示す。図 2

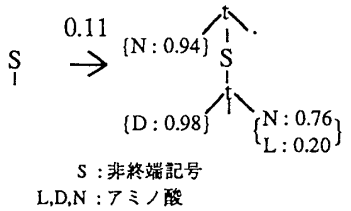


図 2: 学習された規則の例

の各ノードには、1LYM の残基番号で、各 44、52、59 の位置に相当するアミノ酸が高い確率でラベル付けされており、β シート内で互いに相対する位置のアミノ酸分布を、一つの規則として抽出できたことを示している。

5 おわりに

本稿では、タンパク質立体構造における遠距離相互作用の抽出が可能な新しい木文法システム及びその学習・構文解析方法を提案した。また、適用例において、反平行 β シート領域内の相対する位置でのアミノ酸分布を一つの規則として学習できることを示した。本手法は、既存手法と異なり、遠距離に位置する残基同士を直接扱うため、β シート領域を予測するのみならず、β シートの構造をも予測できることを大きな特徴とする。

今後の課題としては、まず学習の高速化がある。現実的な時間でより一般的なアミノ酸配列を学習するには、並列化などによる高速化が必要であろう。また、異なるタンパク質種類間の予測を行うためには、より汎化した規則の抽出が望ましい。アミノ酸種類を分類することによる文字数の低減などの工夫が有効であると予想される。

謝辞

本研究の一部は通産省 [新情報処理プロジェクト] の一環として実施されたものである。

参考文献

[1] N. Abe. Polynomial learnable subclasses of mildy context sensitive language. In *Proceedings of COLING 88*, 1988.

[2] G.D. Fasman. *Prediction of Protein Structure and the Principle of Protein Conformation.* Plenum Press, New York, 1989.

[3] Y. Schabes. Stochastic lexicalized tree-adjoining grammars. In *Proceedings of COLING 92*, 1992.