確率的規則を用いたタンパク質αヘリックス領域予測

1N−2

馬見塚　拓，山西　健司
NEC　C&C情報研究所

1　はじめに

本稿では、タンパク質の二次構造の一つであるαヘリックスの存在領域を確率的規則に基づいて予測する新しい方法を提案し、実験結果により、提案方法が従来手法よりも高い精度で予測できることを示す。

タンパク質の二次構造とは、αヘリックス、βシートなどタンパク質の立体的な構造の中で規則的な構造を指す。与えられたアミノ酸配列に対してこれらの領域を予測することは、タンパク質の立体的な構造を理解し、さらに、その機能を解析する上で重要である。

1960年代より、情報理論、機械学習、ニューラルネット等に基づいて多くの二次構造予測方法が提案されているが、これらの手法は、3状態(αヘリックス、βシート、その他)予測において、すべて予測率が60%前後にとどまっており([1])、立体構造を理解するのに十分とは言い難い。

そこで、筆者らは、まず、アミノ酸配列中の特に重要であると思われるαヘリックス領域を、局所的な配列から従来手法より高い精度で予測する新しい手法を提案する。本手法は、確率的規則の学習理論([5])に基づいている。確率的規則は、アミノ酸配列と二次構造との対応関係に含まれる不確実性を表現するのに適しており、予測率を上げるためには、最適な確率的規則を事例データから学習することが最も重要である([3])。

2　訓練データの生成

訓練アミノ酸配列の複数のαヘリックス領域に対し、αヘリックス領域ごとにPDB(Protein Data Bank)などのアミノ酸配列データベースから正例及び負例を用意する。正例とは、元のαヘリックス領域と60%以上の高い相同性を持つ領域を指し、負例とは元のαヘリックス領域と20-30%の相同性がありαヘリックスではない領域を指す。

3　確率的規則の学習

3.1　確率的規則の構造

確率的規則とは、ここでは、任意の与えられた領域Sに対してαヘリックスが対応する条件付き確率分布を指す。

X_iを領域Sの左から数えてi番目の残基とし、$P(\alpha|X_i)$をX_iに対応する二次構造がαヘリックスである確率とする。ここで、Sがαヘリックスである確率$P(\alpha|S)$が、以下のように書けるような確率構造を仮定する。

$$P(\alpha|S)=\prod_{i=1}^{w}P(\alpha|X_i)$$

さらに、各$P(\alpha|X_i)$を有限分割型確率的規則([5])で表現する。

有限分割型確率的規則とは、i番目の残基位置において、残基X_iを2次元座標(分子量、疎水性)で表し、とり得る範囲を重なり合わない有限個の部分領域(以下、これをセルと呼ぶ)に分割し、mを全セル数、$C_j(i)$をj番目のセルとした時、X_iが$C_j(i)$に含まれる場合に、$P(\alpha|X_i)=p_j(i)$とする規則である。$p_j(i)\,(j=1,\cdots,m)$は確率パラメータである。

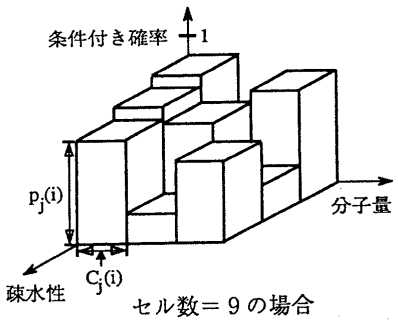

図1: 有限分割型確率的規則の例

3.2　パラメータ推定

有限分割が固定されたもとでの確率パラメータ$p_j(i)$の推定量としては、以下のラプラス推定量$\hat{p}_j(i)$を用いる。ラプラス推定量は、i番目の残基位置において、与えられた事例データの内、$N_j^+(i)$をセル$C_j(i)$に含まれる正例数、$N_j(i)$をセル$C_j(i)$に含まれる全データ数として次式で計算される。

$$\hat{p}_j(i)=\frac{N_j^+(i)+1}{N_j(i)+2}\quad(j=1,\cdots,m)$$

3.3　確率的規則の最適化

有限分割自体として最適なものを事例データから求めるのに、記述長最小(MDL(Minimum Description Length)原理([4])を用いる。ここで、MDL原理とは、規則の記述長と規則を用いて記述される訓練データの記述長との和が最小であるような規則を最適とみなす原理である。これに従えば、各αヘリックス領域において、次式を最小にする有限分割が最適とみなされる。但し、有限分割の構造は領域を形成する全ての残基に対して同一であるとする。

$$-\sum_{i=1}^{W}\sum_{j=1}^{m}\log\{\hat{p}_j(i)^{N_j^+(i)}(1-\hat{p}_j(i))^{N_j(i)-N_j^+(i)}\}+\sum_{i=1}^{W}\sum_{j=1}^{m}\frac{\log N_j(i)}{2}$$

ここで、$\hat{p}_j(i), N_j^+(i), N_j(i)$は前節に従うとし、$m$は全セル数である。また、上式において、第1項は訓練データのShannon符号長を、第2項は$O(\frac{1}{\sqrt{N_j(i)}})$で量子化された各確率パラメータの記述長を表す。

Protein α-helix Region Prediction Using Stochastic Rules
Hiroshi Mamitsuka and Kenji Yamanishi
C&C Information Technology Research Labs. NEC Corp.

4 予測方法

テスト配列に対して、長さ t のウィンドウ W を左 (N 末端側) から動かし、W を訓練配列の α ヘリックス領域の全ての長さ t の部分領域とマッチングを行ない、W の α ヘリックスの尤度を計算する。尤度は学習段階で求めた確率的規則を用いて計算する。

これを、訓練配列の全ての α ヘリックス領域に対して行ない、その中で最大の尤度を W の尤度とする。

さらに、テスト配列中の各アミノ酸残基に対する尤度は、その残基を含む全ての W の尤度の中での最大値を割り当てる。この各残基に対する尤度を 予測曲線 と呼ぶ。

また、あるしきい値 $h(>(\frac{1}{2})^t)$ 以上の尤度を与える残基が α ヘリックスに含まれ、それ以外の残基は α ヘリックスには含まれないとし、予測率 を次式で定義する。

$$Q = \frac{N_\alpha^+ + N_\alpha^-}{N}$$

ここで、N は全残基数、N_α^+ と N_α^- はそれぞれ、正しく予測された実際に α ヘリックスに含まれる残基数、正しく予測された実際に α ヘリックスに含まれない残基数とする。

5 実験結果

今回の実験には、正例としてヘモグロビン α 鎖及び β 鎖の α ヘリックス領域のみを使用した。また、確率的規則を構成する際のアミノ酸の疎水性指標として文献 [2] による指標を使用した。

まず、テスト配列として、PDB から無作為に抽出したタンパク質 (1CDP) を用いた場合の予測曲線を示す。1CDP は、訓練配列の正例との相同性が 10% 以下のタンパク質である。図 2において、実線は予測曲線を、点線は実際の α ヘリックス領域を示す。

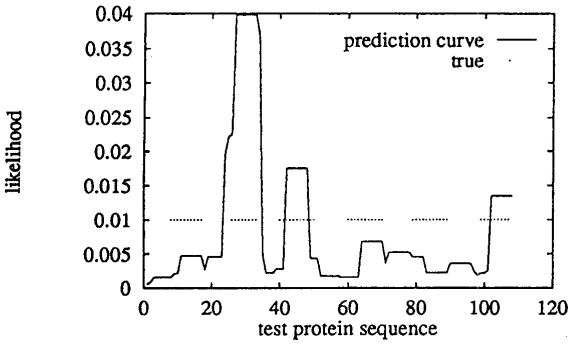

図 2: 1CDP の予測曲線

図 2から訓練データの正例としてヘモグロビンの α ヘリックス領域のみを使用し、1CDP の 6 個の α ヘリックス領域の内の 3 個の領域をほぼ正確に予測できたことがわかる。また、高い尤度が得られなかった 1CDP の他の 3 個の α ヘリックス領域の規則は、使用した訓練配列には含まれていないと予想できる。

さらに、一般に α ヘリックス領域予測の場合には、テストアミノ酸配列の α ヘリックス含有率に予測率が大きく影響される傾向が見られるため、PDB から無作為に抽出した 6 種類のタンパク質 (1CDP、1HRB、3FXC、2SSI、1RN3、3CPA)、全 863 残基に対する予測率を、従来手法の一つである GOR(Garnier-Osguthorpe-Robson) 法による結果とともに表 1に示す。W の長さ t を 7、しきい値 h を 0.01 とした。表 1において、α 残基数は、テスト配列中の α ヘリックス残基数を示す。

	予測率 (%)		アミノ酸残基数	
	本手法	GOR	全残基数	α 残基数
1CDP	60.2	52.8	108	60
1HRB	54.0	46.0	113	81
3FXC	89.8	61.2	98	10
2SSI	86.7	69.0	113	20
1RN3	79.0	66.1	124	33
3CPA	68.1	56.0	307	115
全体	71.7	58.1	863	319

表 1: 予測率

表 1から、本手法は、GOR 法を上回る 70% 以上の平均予測率をあげることが示された。

6 おわりに

タンパク質 α ヘリックス領域を確率的規則を用いて予測する方法を提案した。本手法は、確率的規則を事例データから学習する際の MDL 原理に基づく最適化、アミノ酸の分子量・疎水性の利用、領域単位の予測といった点で特徴付けられる。さらに今回の実験から、提案手法は、訓練配列の正例に一種類のタンパク質のみしか使用しなかったにも関わらず、70% 以上の平均予測率をあげ、予測手法として有効であることが示された。今後、正例のタンパク質の種類を増やし、多くの α ヘリックス領域の規則を構成すれば、予測精度はより向上するだろう。また、本手法を改良し、訓練配列の α ヘリックス領域内の残基間の相関をも表現する確率的規則を学習することが出来れば、さらに精度の高い予測が可能になるだろう。

参考文献

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

[2] J.L. Fauchere and V. Pliska. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *Eur.J.Med.Chem.Chim.Ther.*, 18:369–375, 1983.

[3] H. Mamitsuka and K. Yamanishi. Protein secondary structure prediction based on stochastic-rule learning. In *Proceedings of the Third Workshop on Algorithmic Learning Theory*, 1992.

[4] J. Rissanen. *Stochastic complexity in statistical inquiry*, volume 15 of *Comp. Sci.* 1989. World Scientific.

[5] K. Yamanishi. A learning criterion for stochastic rules. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 67–81. Morgan Kaufmann, 1990. To appear in *Machine Learning*.