ベイズ統計を用いたタンパク質アミノ酸残基の溶媒接触度と二次構造の同時予測

末廣　藍（指導教員: 由良　敬）

1. はじめに

タンパク質は、アミノ酸が一次元的に連結して構成される直鎖状生体高分子である。生物が用いているタンパク質は、20種類のアミノ酸で構成されている。アミノ酸配列は、DNA に与えられた情報を「転写」と「翻訳」することによって得ることができ、アミノ酸配列が決定すると自発的に立体構造を形成する。

ゲノムプロジェクトによりアミノ酸配列の情報は簡単に得られるようになった。しかし、アミノ酸配列の情報から、立体構造を導出するには、X 線結晶解析などの実験で測定するしか方法がない。立体構造の測定には、時間と費用が膨大にかかる。そこで、計算機を用いてタンパク質のアミノ酸配列から立体構造を推定する研究が、世界中で進められている。しかし現在までに、成功例は数えるほどしかない。

タンパク質は、一般的には球状の構造である。タンパク質はアミノ酸が一次元的に連結して構成される直鎖状生体高分子であるため、アミノ酸配列上の各アミノ酸は、溶媒に接触する表面積が小さい（内部に埋まる）アミノ酸と、溶媒に接触する表面積が大きい（表面に露出する）アミノ酸とに分類することができる。各アミノ酸の埋まり具合を溶媒接触度とよぶ。この特徴とは独立に、タンパク質内部にはアミノ酸配列にそって繰り返し構造が存在することがわかっている。典型的な繰り返し構造としてαヘリックスとβストランドが知られている。これらの構造を二次構造とよぶ。

立体構造をコンピュータで推定する研究において、立体構造の特徴を示す溶媒接触度や二次構造を予測する方法の開発が行われてきた。従来の研究では、アミノ酸配列からの溶媒接触度の予測と、アミノ酸配列からの二次構造の予測はそれぞれ独立に行われてきた。しかし立体構造の形成過程において、アミノ酸溶媒接触度の変化と二次構造の形成は、同時に起こっている現象であり、それらはお互いに関連があると考えられる。そこで、アミノ酸配列を入力して、各アミノ酸の溶媒接触度と二次構造とを同時予測する方法を開発し、それぞれを独立に予測した場合と比較することで、溶媒接触度と二次構造とにどの程度の関連性があり、どの程度予測を向上できるかを検討した。

2. 手法

2.1. 予測ターゲット

タンパク質を構成するアミノ酸の溶媒接触度は 0.0 から 1.0 までの連続値である。各アミノ酸の溶媒接触度の値そのものを予測するのは困難であるので、ここでは値をある閾値で2値に分けて扱うことにした。閾値を 0.3 に設定し、溶媒接触度が 0.3 以下のアミノ酸は埋まっている(b)とし、0.3 よりも大きなアミノ酸は表面に露出している(e)とした。タンパク質の二次構造には、αヘリックスとβストランドがあり、各アミノ酸はαヘリックス（H）、βストランド（E）、またはそれ以外の領域（C）のいずれに含まれる。これらの属性を、アミノ酸配列のみから予測することを試みた。

従来の研究では、アミノ酸配列を入力とし、各アミノ酸の属性が b,e のいずれか、または H, E, C のいずれかを予測していた。今回の研究では、溶媒接触度と二次構造を同時に予測するので、アミノ酸配列を入力とし、各アミノ酸の状態が bH, bE, bC, eH, eE, eC の6種類のいずれかであることを予測する。

2.2. ベイズ統計の適用方法

アミノ酸配列 a の状態列を s とすると、求める値は、アミノ酸配列が a の際に、このアミノ酸配列が状態列 s_i をとる確率 $P(s_i \mid a)$ である。ベイズの定理を用いると、この確率は次の式で表すことができる[1]。

$$P(s_i \mid a) = \frac{P(a \mid s_i)P(s_i)}{\sum_x P(a \mid s_x)P(s_x)}$$

$P(a \mid s_i)$ は、状態列が s_i である全アミノ酸配列をデータベースから見いだし、その中でアミノ酸配列が a と一致する頻度として求めることができる。また $P(s_i)$ は、状態列が s_i であるアミノ酸配列の頻度として求めることができる。しかし、これでは予測すべきアミノ酸配列と一致するアミノ酸配列がデータベース内に存在しない場合は、予測することができず、予測方法を構築する意味を失う. そこで2つの仮定を導入する。アミノ酸配列 a の k 番目のアミノ酸 a_k の状態 s_k は、k から $\pm(j-1)/2$ までのアミノ酸 $\{a_{k,j}\}$ で決まると仮定する。この仮定は、タンパク質の立体構造形成過程が、原子の近距離相互作用から始まることより妥当であろう。しかし、これでも予測対象のアミノ酸配列において、長さ j の部分配列と一致するアミノ酸配列がデータベースに存在しなければ、予測することができない。そこで、第2の仮定として、アミノ酸 a_k の状態 s_k はアミノ酸 a_k とアミノ酸 a_l の組によって決定されるとする。第1の仮定より a_l は部分配列内のアミノ酸である。つまり部分配列内の各アミノ酸は a_k に直接影響を及ぼし、a_k 以外のアミノ酸がお互いに影響をおよぼすことはないと仮定する。この仮定の正しさは、本予測がどの程度うまくいくかによって評価することにする。

以上の仮定を導入することで、$P(s_i \mid a)$ は次のように表すことができる。

$$P(s_i \mid a_k) = \frac{\prod_{l=-(j-1)/2}^{(j-1)/2} P(a_l, a_k \mid s_i)P(s_i)}{\sum_x \prod_{l=-(j-1)/2}^{(j-1)/2} P(a_l, a_k \mid s_x)P(s_x)}$$

上式のパラメター $P(a_l, a_k \mid s_i)$ と $P(s_i)$ をタンパク質立体構

造データベース Protein Data Bank (PDB)より、求めることができれば、この予測は可能となる。それぞれの値は、PDB に格納されている独立なタンパク質に見いだされる頻度から求めることとした。

3. 結果と考察

3.1. パラメター取得データベース

PDB には共通祖先由来タンパク質の立体構造情報が多く含まれている。また質の悪いデータも多く含まれている。それらを除去した結果 4,739 個のタンパク質立体構造情報が得られた。これらのデータにおいて、6 つの状態は以下のように分布していた。この値を $P(s)$ として用いた。

表1:予測すべき状態のデータベース内分布

状態	bH	bE	bC	eH	eE	eC
頻度	0.183	0.161	0.192	0.145	0.055	0.264

3.2. 様々な状態の予測

4,739 個のタンパク質立体構造情報から、パラメター（$P(a_l, a_k \mid s_i)$ と $P(s_i)$ を抽出し、様々な状態の予測を行った。

まず二次構造のみ（bH+eH, bE+eE, bC+eC）の３状態予測を行ったところ、図１左の結果を得た。図における横軸は j の値を、縦軸は予測精度（%）を意味する。グラフには４本の線がある。一番精度のよい線が、2.方法の項目で記した式を完全に実装した計算方法、２番目に精度のよい方法は、$P(a_l, a_k \mid s_i)$ における a_k を無視した方法、３番目の精度は、従来よく用いられていた頻度のみにもとづく方法 (2.方法の式において $P(s) = 1$ とおく) で計算した結果であり、一番精度の悪い方法は、従来よく用いられる方法でさらに a_k を無視した方法の結果である。2.方法の項で記した式を完全に実装した予測方法の精度(約 65%)は、先行する１本のアミノ酸配列からの SVM やニューラルネットワーク法を用いた二次構造予測研究の方法と同程度の精度である[2]。さらにこの実験の結果、$j = 15$ 程度で精度向上が飽和することがわかった。

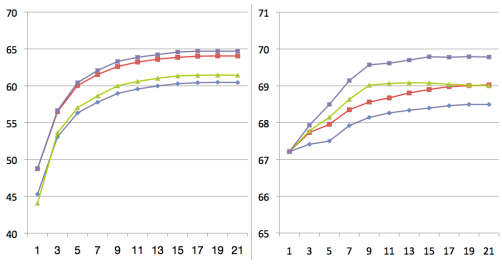

図1:予測の結果。左は二次構造予測、右は溶媒露出度の結果を示す。

次に溶媒接触度のみ（bH+bE+bC, eH+eE+eC)の２状態の予測を行ったところ、図１右の結果を得た。二次構造予測の場合と同様、2.方法の項目で記した式を完全に実装した計算方法が一番高い精度（約 69%）を示した。溶媒接触度の予測においても $j = 15$ 程度で精度向上が飽和することがわかった。しかし、横軸 j に対する精度向上の度合いが、二次構造予測の場合よりも鈍いことがわかった。このことは、二次構造の形成は前後のアミノ酸の種類に強く依存するが、溶媒接触度は、前後のアミノ酸の種類にあまり依存しないことを意味する。

次に二次構造と溶媒接触度を同時に予測(６状態予測）した。その結果を図２に示す。図２には二次構造と溶媒接触度をそれぞれ独立に予測し、それぞれの予測結果を集めて６状態を予測した場合の結果もしめした。一番精度のよい線が、2.方法の項目で記した式を完全に実装した計算方法、２番目に精度のよい方法は、$P(a_l, a_k \mid s_i)$ における a_k を無視した方法、３番目の精度は、二次構造と溶媒接触度をベイズの方法で独立に予測した結果に基づく６状態予測の結果、一番精度の悪い方法は、a_k を無視した方法による独立の予測に基づく６状態予測の結果である。同時予測による精度向上は１%程度であるが、ブートストラップにより精度の分布を調べ、精度の分散を利用して検定を行ったところ、この差は有意であることがわかった。

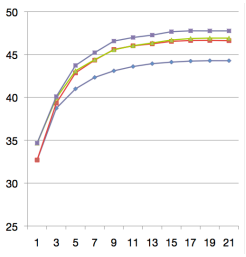

図2:6状態予測の結果。

4. まとめと今後の課題

１本のアミノ酸配列から、二次構造と溶媒接触度を同時に予測する方法は、本研究で新しく開発された方法である。タンパク質の立体構造形成過程を検討すると、二次構造の形成とタンパク質が球状になることはほぼ同時に起こっていることから、二次構造形成と溶媒接触度の違いには何らかの相関関係があると考えられる。よって両性質の特徴量を既知データから同時に抽出し、その特徴量を用いてアミノ酸配列から二次構造と溶媒接触度を同時に予測する方法は、それぞれを独立に予測する方法よりも、高い精度が得られるであろうと予想した。

実験の結果、予測精度の向上は見られたが、その量はわずかであった。このことはタンパク質の立体構造形成過程において、二次構造の形成とタンパク質が球状になることに強い相関がないことを示唆している。この結果は当初の予想に反しており、タンパク質の立体構造形成過程の研究に何らかの示唆を与える可能性がある。

今後は、タンパク質の立体構造形成過程の研究を詳細に調べて、ここで開発した予測方法の改良の可能性を追求し、本手法が、タンパク質立体構造予測に貢献できる手法になるようにしたい。

参考文献

[1] 涌井良幸. 道具としてのベイズ統計. 日本実業出版社, 第 2 版 2009.

[2] Michel J. Thompson and Richard A. Goldstein. Predicting Solvent Accessibility: Higher Accuracy Using Bayesian Statistics and Optimized Residue Substitution Classes, 1996.