多次元分布の線形基底変換による圧縮表現の提案、及びタンパク質残基間相対位置分布への応用

鬼塚　健太郎† 野口　保†
安藤　誠† 秋山　泰†

多次元での分布を線形基底変換し、さらに変換パラメータ数を大幅に減らすことで、少数のパラメータで正確に多次元分布を記述する方法を提案する。ついで、この手法をタンパク質の同一鎖に含まれる二つの残基の相対位置の分布を表現するために応用し、それを用いてタンパク質立体構造からの残基配列推定問題を解き、多次元分布表現法の有効性を検証する。

隣接する残基間の相対位置を、ほぼ完全に記述する3つの二面角 $(\phi^d, \psi^d, \omega^d)$ の三次元分布に適応し、ついで、隣接しない残基間の相対位置を表す極座標とオイラー角 $(r^p, \theta^p, \phi^p, \theta^e, \phi^e, \psi^e)$ の六自由度の分布に適応し、統計解析して得れた分布を残基配列推定法に応用する。

A compressed representation of multiple-dimensional distribution by linear base transformation, and it's application to residue-pair-relative-distribution of proteins

KENTARO ONIZUKA,† TAMOTSU NOGUCHI,† MAKOTO ANDO† and YUTAKA AKIYAMA †

We propose a representation method to strongly reduce the number of parameters representing a multiple-dimensional distribution. The method reduces the number of parameters by linearly expanding the distribution with orthonormal linear bases.

To evaluate the representation performance of the method, we apply it to the distribution of relative position of two amino-acid-residues in a protein chain. We firstly apply this method to represent the distribution of three dihedral angles $(\phi^d, \psi^d, \omega^d)$ which almost perfectly represents the relative position of adjacent residues, and then apply it to the distribution of the relative position of two residues separated in the sequence, where the relative position is represented by polar coordinates (r^p, θ^p, ϕ^p) and Euler's angle $(\theta^e, \phi^e, \psi^e)$.

1. 研究の背景

分子生物学の分野において、生命体で最も重要な物質であるタンパク質の機能や機能実現の仕組みを解析することは、最も重要な研究課題である。そのためには、できるだけ多くのタンパク質の立体構造を知る必要がある。しかし、タンパク質の立体構造をX線結晶回折、NMR法、電子顕微鏡などで決定するのには、非常に多くの労力と時間がかかる。

タンパク質は、20種類の生体アミノ酸と呼ばれる分子が脱水結合によって一列に鎖のように繋がったものである。タンパク質の一つの鎖については、そのアミノ酸20種類の配列が決定されれば、共有結合に基づく化学式は完全に一意に決定される。タンパク質の残基配列の決定は、現在では極めて短時間で行なうことができる。したがって、タンパク質の立体構造を残基配列から予測する方法が得られれば、実験による立体構造決定を待たずに、タンパク質の機能推定や機能実現の仕組みを解析することができる。

こうして、タンパク質立体構造予測問題が多くの研究者によって研究されてきた。当初は、タンパク質立体構造中に頻繁に見出される規則的な構造である螺旋(ヘリックス)構造と鎖同士が二次元的に膜状の構造を作るシート構造の二つが着目され、残基配列中のどの部分が、ヘリックス構造をとるか、あるいはシート構造の骨格であるストランド構造をとるか、という二次構造予測に関する研究が行なわれた[3)~8)]。

しかし、構造既知のタンパク質の数が少ないこと、また、配列と二次構造との直接的な関係が見出せなかったことから、ヘリックス状態、ストランド状態、それ以外の三状態を予測する精度は、Rostらが、残基配列のみ

† 新情報処理開発機構
Real World Computing Partnership

が知られているタンパク質の配列を構造データと一緒に用いて、データセットを増やす方法[9]を提案して、ようやく70%以上の二次構造予測精度を出す方法も提案された。しかし、最終的な三次元構造を推定するのに必要な二次構造予測精度は得られていない。

1990年代に入って、Chothiaらが、「タンパク質の立体構造は多くても千種類程度しかないであろう」という見解を発表した[10]。このころから、タンパク質の立体構造予測は、二次構造予測から、「構造未知のタンパク質の配列が、構造既知のタンパク質のどれと近い構造をとるか」を推定する縫糸(threading)法による構造推定方法の研究に中心が移ってきた[11]。

縫糸法では、構造既知のタンパク質のアミノ酸残基配列と立体構造との関係を統計的に解析し、立体構造と残基配列との互換性を評価する。構造未知のタンパク質Xの残基配列が与えられたときに、構造既知のタンパク質の立体構造に、そのXの残基配列を糸と通すように当てはめ、その互換性を評価し、その評価が高ければ、そのXは、発見された互換性の高い構造と類似の構造をとると判断する[12]。

この「配列と構造との互換性評価方法」の研究においては、立体構造中の二次構造と、周囲の環境に着目し、アミノ酸残基とそのアミノ酸を取り巻く環境との互換性を統計的に処理する方法が提案された[11]。

後の研究に大きく影響を与えたのは、鎖の中にあるアミノ酸残基二つの立体構造中での相対距離を統計的に処理して、統計力学的なポテンシャルを計算する方法である[13]。数百の構造既知のタンパク質の立体構造と、そのアミノ酸残基配列のデータベースを用い、アミノ酸の種類、配列上での距離、立体構造内での距離をヒストグラム化し、その負の対数に熱力学的系数を掛けたものを、統計的平均力場ポテンシャルとして定義する。そして、ポテンシャルの和がもっとも小さくなるような残基配列が、その立体構造と最も互換性が高いと評価するのである。この方法は、さまざまな研究者によって踏襲されたが、そのほとんどは、同じ鎖に存在する残基対の相対距離だけを考慮したものである。つまり、このポテンシャルは、相対距離のみによって決定される一自由度のポテンシャルなのである。

本論文は、残基対の統計ポテンシャルで、一自由度のポテンシャルでは表現しきれない部分を、できるだけ精密に表現するために、アミノ酸残基間相対位置関係を完全に表す相対三次元位置と相対三次元姿勢の合わせて六自由度全てについて考慮した六自由度のポテンシャルを計算する手法を提案するものである。

まず、一般論として、多次元(あるいは多自由度)の複雑な分布密度を少数のパラメータで比較的正確に表現する方法について述べ、続いて、この方法の直接的な応用として隣接する残基間の相対位置を表現する三つの二面角(ϕ,ψ,ω)で表した場合の分布を表現するために用いる方法を述べる。次に、隣接していない残基間の三次元相対位置の分布で極座標系を用い、三次元相対姿勢の分布にEiler角を用いた本研究の中心となる方法について述べる。さらに、この方法をタンパク質立体構造から残基配列を推定する配列推定問題に適応し、一自由度のポテンシャルに比べて、より正確に分布を表現していることを検証する。最後に、このポテンシャルに関する議論を行ない、今後の研究方針について触れる。

2. 多次元分布の表現法

ここでは一般論として、話を簡単にするために、全ての辺の長さが同一であるN次元の超立方体内に観測される観測点の頻度分布を、少ないパラメータで表現する方法について考える。ここで考える頻度分布とはN個の数値からなるN次元ベクトルで与えられる観測点が、考慮しているN次元の超立方体内に膨大に存在している場合の頻度分布である。

頻度分布の表現方法として、N次元の超立方体の各辺をM等分して、この超立方体をM^N個の部分超立方体に分割し、各部分超立方体にいくつの観測点が存在するかを数え上げ、N次元の度数分布をとる方法がある。この場合は、分布を表すパラメータは、各部分超立方体に存在する観測点の数(あるいはこの数を観測点の全数で割った相対頻度数)であり、この場合のパラメータ数は、部分超立方体の数だけ存在するので、M^N個になる。

しかしこの方法では、空間の次元Nが比較的小さい、たとえば、$N=6$の六次元空間程度であっても、分割数Mを10にすると、部分超立方体の個数は、百万個になり、観測点の数(あるいはサンプル数)Kがこの数字より小さい場合は、ほとんどの部分超立方体内に観測点が発見されず、多くのパラメータは無駄である。

そこで、観測点の分布を、Fourier展開することを考える。観測点は、その部分にδ関数が存在すると考えることができるので、K個の観測点が存在する場合、その分布$f(X)$は、以下の式で表される。

$$f(X)=\sum_{k=1}^{K}\delta(X_k-X). \tag{1}$$

この分布をFourier展開する。ここで簡単のためにFourier展開で用いられる規格直交線形基底である三角関数を、問題の定義領域を表す超立方体の一辺の長さLで直交するように以下の形で表現する。

$$g_i(x)=\begin{cases}\frac{1}{\sqrt{2.0\pi}} & (i=0)\\ \sin\left(\frac{(i+1)\pi}{L}x\right) & (i=2n+1)\\ \cos\left(\frac{i\pi}{L}x\right) & (i=2n)\end{cases} \tag{2}$$

さらに、このg_iをN次元に拡張したものを、$G_I(X)$とする。添字jはN次元ベクトルあるいは座標のj番目の成分を表す。

$$G_I(X) = \Pi_j g_{I_j}(X_j) \tag{3}$$

この規格直交線形基底 $G_I(X)$ で、$f(X)$ を展開するので、$f(X)$ は展開された結果、以下のように表現される。

$$f(X) = \sum_I a_I G_I(X) \tag{4}$$

このとき、a_I は、$G_I(X)$ の規格直交性により、以下の計算で求めることができる。

$$a_I = \int f(X) G_I(X) dX \tag{5}$$

$$= \sum_I \int \delta(X - X_I) G_I(X) dX \tag{6}$$

$$= \sum_I G_I(X_I) \tag{7}$$

現実には、I を無限とれば、$f(X)$ は任意の精度で展開できるが、この展開を有限で打ち切ることで、分布 $f(X)$ は、展開した個数だけのパラメータ数で近似的に表現できる。パラメータ数は、打ち切り次数をどのように設定するかで任意に選べる。本研究においては、各軸ごとの展開次数の総和が一定以下であるように打ち切り次数を決定することとし、各軸での打ち切りを $\sum_k I_k \leq I_{\max}$ となるように選ぶ。この場合のパラメータ数 P は、N 次元の超三角錐の体積を考えて、$I_{\max}$ が十分大きいならば、だいたい $P \simeq I_{\max}^N / N!$ 程度になる。$N = 2$ の二次元平面上の関数の近似については、同様の方法が、デジタル画像の情報圧縮手法として広く使われている[14]。これまで述べてきた方法により、多次元の複雑な分布を少数のパラメータで効率よく表現できることがわかる。

3. タンパク質立体構造における主鎖二面角分布への応用

タンパク質の立体構造とアミノ酸残基配列との対応を見る上で、隣接するアミノ酸残基の相対位置を、もっとも少ないパラメータで正確に記述する方法は、隣接するアミノ酸残基の主鎖の共有結合まわりの二面角をパラメータとするものである。

アミノ酸残基のうち主鎖を構成する原子は、N, C^α, C の三つである。隣接する残基も考慮して、$-N-C^\alpha-C-N-C^\alpha-C-$ と結合している。二面角 ϕ^d は、$C-N-C^\alpha$ が作る平面と $N-C^\alpha-C$ が作る平面とがなす角度であり、ψ^d は $N-C^\alpha-C$ の作る平面と $C^\alpha-C-N$ が作る平面とがなす角度であり、ω^d は、$C^\alpha-C-N$ が作る平面と $C-N-C^\alpha$ が作る平面とがなす角度である。

隣接残基間の相対位置自由度は、ϕ^d, ψ^d, ω^d の三つの二面角の自由度で近似的に表現可能である。ϕ^d, ψ^d, ω^d は、角度であるから、2π の周期をもつので、0 から 2π の範囲を、Fourier 展開することで、全節で述べた多次元分布の表現法をそのまま応用することができる。この場合次元数は $N = 3$ である。$I_{\max}$ を 12 にすると、パラメータ数は、455 個になる。

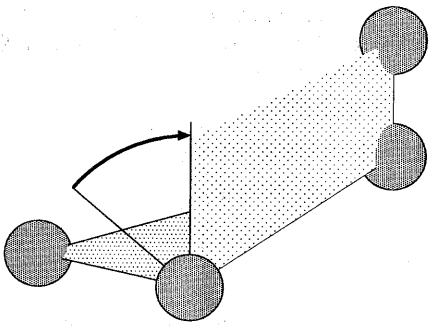

図1　二面角の定義

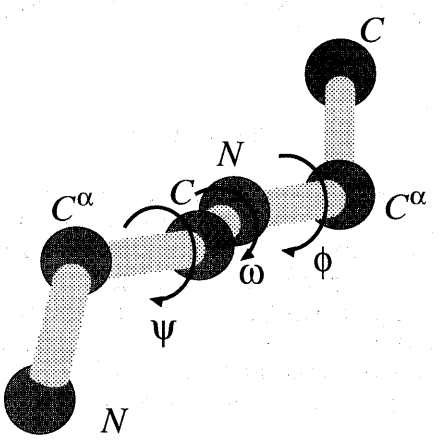

図2　主鎖を構成する原子と二面角 ϕ, ψ, ω

4. タンパク質立体構造におけるアミノ酸残基相対位置分布への応用

タンパク質の立体構造とアミノ酸残基配列との対応を見る上で、同一鎖上にあって配列上隣接していないアミノ酸残基の相対位置を考える。これは、三次元空間内での二つの剛体の相対位置と見ることができる。一方の剛体のもつ固有座標でもう一方を見た場合の三次元相対位置の三自由度と、一方の剛体から見たもう一方の剛体の三次元相対姿勢の三自由度がある。三次元相対位置は、タンパク質立体構造に関する従来研究で、相対位置を決めるためのもっとも重要な自由度として従来から相対距離が利用されてきたので、極座標系を用いることにする。三次元相対姿勢については、一般的な Eiler 角を用いる。ここで、アミノ酸残基の固有座標としては、C^α

原子を原点とし、側鎖を構成する C^β 原子がZ軸上、N 原子が XZ 平面上にあるような固有座標を採用している。

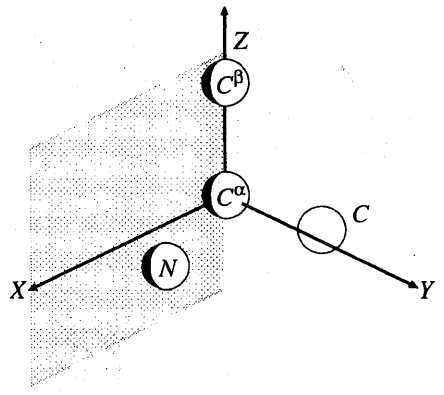

図3 局所座標

極座標系においては、積分量などの問題から、r^p, θ^p, ϕ^p で表現されるそれぞれの極座標方向に対して、まず、θ^p, ϕ^p については、この系での直交基底である球面調和関数を用い、ついで、動径方向 r^p については、軸方向の積分量 dr^p を考慮して、三角関数 $g_i(x)$ を x で割ったものを考える。すなわち、$g_i(x)/x$ である。観測点の測定範囲として、$0 < r^p < r_{\max}$ を考えると、この場合の直交基底は、以下のものになる。

$$\frac{g_i\left(\frac{2\pi r^p}{r_{\max}}\right) r_{\max}}{2\pi r^p} \tag{8}$$

球面調和関数における展開次数打ち切りの方法は、θ^p 方向と ϕ^p 方向では量子数の間に関連があるからであるため。動径方向 r^p の量子数の打ち切り次数 $k^p_{\max}$ と、θ^p 方向の量子数の打ち切り次数 $l^p_{\max}$ と ϕ^p 方向のそれ $m^p_{\max}$ を、$k^p_{\max} = l^p_{\max} + 2m^p_{\max}$ となるようにした。

次に、三次元相対姿勢については、Eiler角 θ^e, ϕ^e, ψ^e を用いる。この場合も、θ^e, ϕ^e の展開については球面調和関数を用い、また、ψ^e については三角関数を用いる。ここでの打ち切り方法は、極座標三次元相対位置の場合と同じである。

最終的に、$r^p, \theta^p, \phi^p, \theta^e, \phi^e, \psi^e$ で表される相対位置については、それぞれの量子数 $k^p, l^p, m^p, l^e, m^e, k^e$ について、$k^p + l^p + 2m^p + l^e + 2m^e + k^e \leq I_{\max}$ となるように各軸方向について打ち切り次数を調整した。本研究では、$I_{\max}$ を6にしてある。この場合の展開パラメータの総数は、330である。この場合の解像度は、タンパク質中のあるアミノ酸残基の周囲の環境で、異なる近傍残基との位置関係をぎりぎりで分離することのできる分解能である。

距離のみを考慮した分布を考える場合、その解像度は、普通0.5Å程度であり、パラメータ数は、ここからおよそ20個程度になる。これに比べて、今回提案する方法では、前述のように展開打ち切りを行なうことで、六自由度の分布を表現しているにも関わらずパラメータ数は330個である。

5. タンパク質のアミノ酸残基配列推定問題への応用

タンパク質中のアミノ酸残基は、その種類によって周囲の環境が異なることが知られている。前節までで述べてきた分布表現は、ある種のアミノ酸残基があった場合、周囲にどのようにアミノ酸が配置されるかを統計的に処理して表現するためのものである。

隣接する残基の相対位置は、残基間に存在する三つの共有結合回りの二面角で決定される。この二面角の分布は、二次構造趣向性と強い関係があり、また、二面角について強い制約を受けるプロリンや、反対に制約をほとんど受けないグリシンについては、他の残基に比べて特異な二面角分布を示す。

残基種ごとに、前節までで述べた近傍残基の分布を統計的に得ることによって、立体構造が与えられたときに、それぞれの部位にどの種類のアミノ酸残基が来ることがもっとも確からしいかをその統計的に得られた分布から推定することができる。これを、**立体構造からの残基配列推定問題**と呼ぶことにする。

まず、あるアミノ酸残基の種類 a について、その残基 a がタンパク質鎖中にある場合、その鎖の配列上で k ずれた場所のアミノ酸残基が a に対して空間的にどのような相対位置に存在するかを、前節までの方法に従って統計処理し、その分布を $f^a_k(x)$ とする。アミノ酸残基は、配列上での上流方向と下流方向が分子構造上対称ではないので、k の値は正値、つまり、下流側の場合と、負値、すなわち上流側の両方の場合を区別しなければならない。

a からの距離は最大値 $r_{\max}$ を越えないもののみを統計対象とし、観測されたサンプル数 m^a_k で分布 $f^a_k(X)$ を割った $f^a_k(X)/m^a_k$ を相対頻度分布 $\rho^a_k(X)$ とする。これに対して、アミノ酸残基の種類を問わずに得られた分布 $f_k(X)$ を考え、これを対応するサンプル数 m_k で割ったものを $\rho_k(X)$ とする。このとき、以下のことがいえる。

$$f_k(X) = \sum_a f^a_k(X) \tag{9}$$

$$m_k = \sum_a m^a_k \tag{10}$$

ここで、アミノ酸残基種別にとられた分布 $\rho^a_k(x)$ と種類を問わない $\rho_k(X)$ の比 $\rho^a_k(X)/\rho_k(X)$ を計算すると、この比が1よりも大きいときは、そのアミノ酸残基種 a については、平均以上の割合で X という相対位置関係にあることが分かり、逆に1より小さいときは、平均

以下の割合でその位置関係 x に来ることになる。そこで、この比の対数に -1 をかけた量、すなわち以下の式で与えられる量を考える。

$$S_k^a(X) = -\ln\left(\frac{\rho_k^a(X)}{\rho_k(X)}\right) \tag{11}$$

この量 $S_k^a(X)$ をスコアとして考え、タンパク質立体構造中のあるアミノ酸残基の部位について、全ての k についてのこのスコアの和を計算する。

$$S^a = \sum_k S_k^a(x) \tag{12}$$

この和が最も小さいアミノ酸残基種 a が、この部位に来る可能性のもっとも高いアミノ酸残基種であると言える。

タンパク質の配列の長さはまちまちであることを考えると、k の上限値および下限値は簡単には定まらない。そこで、k の絶対値が特定の値以上のものは、一括して配列上遠い残基とし、ρ_{far+} と ρ_{far-} として統計処理することにする。$far+$ は k が正の上限値を越えるもので、$far-$ は、k が負の下限値を越える場合である。本論文の研究では、$k=8$ を上限、$k=-8$ を下限としている。なお、$k=\pm 1$ の場合は、二面角 $(\phi^d, \psi^d, \omega^d)$ を自由度として分布を解析し、$k>1$ の場合は、$(r^p, \theta^p, \phi^p, \theta^e, \phi^e, \psi^e)$ を自由度として分布を解析している。

サンプル数は、概ねパラメータ数に対して足りているが、場合によっては足りない場合がある。そこで、サンプル数が少ない場合にも、有効なサンプル数が得られるように、アミノ酸種を考慮した相対頻度分布 $\rho^a(x)$ とアミノ酸種を考慮しない相対頻度分布 $\rho(x)$ を混合し、サンプル数が少ないときには、$\rho(x)$ に近く、サンプル数が十分に大きければ $\rho^a(x)$ に近くなるような処理を行なう。

$$\varrho^a(x) = \frac{\rho(x) + m_k^a \sigma \rho^a(x)}{1 + m_k^a \sigma} \tag{13}$$

この式で、$\varrho^a(x)$ は、サンプル数 m_k^a が $1/\sigma$ 個程度あったときに、$\rho^a(x)$ と $\rho(x)$ との混合比が 1:1 になり、m_k^a が $1/\sigma$ より十分大きければ $\rho^a(x)$ に近付き、小さければ $\rho(x)$ に近付く。今回の研究では、上述のように、パラメータ数との関係から、$1/\sigma$ を 100 とした。

基底による展開打ち切りによって、$\rho(x)$ が負数になる場合は、たとえ、このときは、$\rho^a(x)$ あるいは、補正された $\varrho^a(x)$ が正値であるとスコア $S_k^a(x)$ が発散する。そこで、この場合は、スコアを強制的に 0 にした。

最終的に実際のスコアは、以下の式で与えられる。

$$S_k^a(X) = \begin{cases} -\ln\left(\frac{\varrho_k^a(X)}{\rho_k(X)}\right) & (\rho_k(X) \neq 0) \\ 0 & (\rho_k(X) = 0) \end{cases} \tag{14}$$

6. 検証実験

六自由度のアミノ酸残基相対位置分布の性能を検証するための実験として、この分布を用いたアミノ酸残基配列推定を行なった。タンパク質立体構造データベースである、Brookhaven Protein Data Bank (PDB) の Release 84 から、良質 (高解像度のデータでかつ主鎖原子に欠損がないもの) なタンパク質分子の立体構造データを相互の相同性が 45% を越えないように PDB-REPRDB[15] から 911 個の立体構造 (主鎖) を選びだし統計データセットとした。

ここで、推定精度とは、検証セットのタンパク質の立体構造を与えて、そのタンパク質のアミノ酸残基配列を推定し、正解であった残基の数 N_h を全残基数 N_{all} で割った百分率である。

正解でなかった場合の傾向を調べるために、与えられた立体構造で実際の残基種種のスコアの和 S_{all}^{true} と、推定した残基種のスコアの和 $S_{all}^{predict}$ の比 $S_{all}^{true}/S_{all}^{predict}$ も計算した。

精度の高かった六自由度の場合については、検証セットを統計データに入れたものについても調べた。また、この場合については、σ を 1/100 から 1/500, 1/2000 および 1/10000 に変更した場合についても調べた。

分布自由度	σ	精度	スコア比	正解スコア 0 以下の率
検証セット除外				
1	0.01	17.262%	0.162	57.38%
6	0.01	19.566%	0.113	59.41%
検証セット込				
6	0.01	20.3%	0.137	60.65%
6	0.002	20.6%	0.201	61.65%
6	0.0005	20.7%	0.228	62.22%
6	0.0001	20.6%	0.309	60.65%

この検証実験の結果から分かることは、一自由度の場合と六自由度の場合では、六自由度の場合のほうが推定精度が 2% 以上高く、相対位置分布の表現がより正確に行なわれていることを示していることである。またこの場合、正解スコアが 0 以下である比率も 2% 以上高い。しかし、正解スコア比が 1 次元の場合に比べてかなり低い。σ の値をより小さく変更すると、推定精度、正解スコア比、正解スコア 0 以下の割合ともに良い数値になることが分かる。これは残基種を考慮しない統計の混合を増やすことで、分布が希薄な部分を減少させることができ、これで悪いスコアが出にくくなるためであると考えられる。

7. 考　　察

タンパク質の立体構造予測問題から派生した、縫糸法によるアミノ酸残基配列の立体構造認識問題において、精密な統計処理を行なうために、同一配列上に存在する残基対の相対位置関係の分布を精密に表現する手法を提案した。

この方法により、利用できるタンパク質立体構造のデータが千前後である現状において、相対位置の六つの

自由度についてその分布を統計処理することを可能にし、従来の相対距離のみによる一自由度の分布で解析する方法にくらべて、大幅に詳細な分布解析が可能になった。

この方法を、タンパク質の与えられた立体構造に関する残基配列推定問題に適応し、推定精度が、従来法にくらべて良いことを示した。与えられた立体構造からそのタンパク質の残基配列を推定する精度がおよそ、20%程度であるということは、一般に「25%程度の配列相同性があれば、同じ立体構造をとる」とされていることを思い起こさせる。

8. まとめと研究の今後

データ数の問題から、現状の六自由度を用いた統計処理では、相対位置関係を解析する残基対の両者のアミノ酸残基種を考慮した統計をとることは難しい。より一層のパラメータ圧縮法を検討するか、データ数を増補する方法を開発して、残基対両者の残基種を考慮した統計をとれるように拡張しなければならない。これを行なうことによって、縫糸法においても、一自由度のポテンシャルと比較した性能評価が可能になる。

六自由度の分布解析で縫糸法が可能になれば、次に行なうことは、縫糸法でえられた立体構造モデルをより一層現実的な立体構造にする方法の開発である。縫糸法で用いた分布から得られるエネルギー関数を用いて局所的構造最適化を行なう方法が一般的である。今回提案した六自由度のエネルギー関数を、これに適応した構造最適化について研究を進めていきたい。

参 考 文 献

1) Chou, P.Y.; and G.D. Fasman 1974. "Prediction of protein conformation". *Biochemistry 13*: 222-244.

2) Fasman, G.D. ed. 1989. *Prediction of Protein Structure and the Principles of Protein Conformation.* New York: Plenum Publishing Corporation.

3) Garnier, J.; D.J. Osguthorpe; and B. Robson 1978. "Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins". *J. Mol. Biol. 120*: 97-120.

4) King, R.D.; and M.J.E. Sternberg 1990. "Machine learning approach for the prediction of protein secondary structure". *J. Mol. Biol. 216*: 441-457.

5) Qian, N.; and T.J. Sejnowski 1988. "Predicting the secondary structure of globular proteins using neural network models". *J. Mol. Biol. 202*: 865-884.

6) Lim, V.I. 1974. "Algorithms for prediction of α-helices and β-structural regions in globular proteins". *J. Mol. Biol. 88*: 873-894.

7) Bohr, H.; J. Bohr; S. Brunek; M.J.R. Cotterill; B. Lautrup; L. Norskov; H.O. Olsen; and B.S.Pertersen 1988. "Protein secondary structure and homology by neural networks". *FEBS Letters 241(1,2)*: 223-228.

8) Cohen, F.E.; R.M. Abarbanel; I.D. Kuntz; and R.J. Fletterrick 1986. "Turn prediction in proteins using a pattern matching approach".

9) Rost, B.; and C. Sander 1993. "Prediction of Protein Secondary Structure at bettter than 70% Accuracy". *J. Mol. Biol. 232*: 584-599.

10) Chothia, C. 1992. "One thousand families for the molecular biologist". *Nature 357 18-JUN* 543-544.

11) Bowie, J. U.; R. Lüthy; and D. Eisenberg 1991. "A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure". *Science 253*: 164-170.

12) Yue, K.; and K. Dill 1991. "Inverse protein folding problem: Designing polymer sequences". *Proc. Natl. Acad. Sci. USA 89*: 4163-4167.

13) Sippl, M.J. 1990 "Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structure in Globular Proteins". *J. Mol. Biol. 213*: 859-883.

14) Wallace, G.K. 1991 "THe JPEG Still Picture Compression Standard" *CACM 34 No.34* 30-44.

15) Noguchi, T.; K. Onizuka; Y. Akiyama; and M. Saito 1997. "PDB-REPRDB, A Database of Representative Protein Chains in PDB (Protein Data Bank)" *Proc. of ISMB'97*: 214-217, The AAAI Press.