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，多次元分布を少数のパラメータによって正確に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ことに応用し，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．

まず，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 $(\phi^d, \psi^d, \omega^d)$ の三自由度分布に適用する．ついで，隣接しない残基間の相対位置を表す極座標と相対姿勢を表すオイラー角 $(\theta^e, \phi^e, \psi^e)$ の合計六自由度 $(r^p, \theta^p, \phi^p, \theta^e, \phi^e, \psi^e)$ の分布に適用する．

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

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

We propose a representation method to strongly reduce the 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 power of the method, we apply it to the distribution of the relative position of amino-acid-residue pairs 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 represent the relative position of adjacent residues, and then apply it to the distribution of the relative position of residue pairs in the sequence, where the relative position is represented by six parameters $(r^p, \theta^p, \phi^p, \theta^e, \phi^e, \psi^e)$ three of which are polar coordinates (r^p, θ^p, ϕ^p) and the other three are Euler's angles $(\theta^e, \phi^e, \psi^e)$.

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．タンパク質の機能は，その立体構造を知ることで，解析が容易になる場合が多い．そのためにも，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求めるには，最も強力な方法であるX線結晶回折法 (X-Ray Crystallography) を用いる場合，タンパク質ごとに異なる精製法の開発に概ね数年，結晶化法の開発にさらに数年を要し，X線結晶回折法で得られた回折像の解析にさらに概ね一年の時間を必要とする．結晶化を必要としない核磁気共鳴法 (Nuclear Magnetic Resonance, NMR法) を用いた構造決定法は，複雑で巨大なタンパク質の構造を決定するには不向きであり，また，構造を一意に決定することが難しい．今後期待される構造決定法として，この1, 2年に急速に進歩している電子顕微鏡を用いたSingle Particle Analysis (SPA) 法[1)] がある．SPA法は，原理的に結晶化の必要がなく，また精製についても他の方法ほど完全なものを要求しないが，現状では解像度が悪く (最も良い場合で6Å程度)，精度の高い立体構造 (X線結晶回折では1Å程度の精度がある) を得ることが難しい．

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 (結合したあとのアミノ酸はすでに酸ではなくなって

† 技術研究組合新情報処理開発機構
Real World Computing Partnership

いるので残基と呼ばれる). タンパク質によっては複数の鎖がさらに結合したものもある.

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 (この場合, 立体構造形成時のジスルフィド結合などは除く). タンパク質の残基配列の決定は, 現在飛躍的に発展したDNAの配列解析法などのおかげで, 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.

タンパク質残基配列から, タンパク質立体構造を予測する方法は, 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.[2)~4)] また, このころには, 残基数が数十から二百程度の比較的小さいタンパク質の立体構造が百種程度知られていた. そこで, 構造既知のタンパク質の残基配列と立体構造 (あるいは二次構造) との統計的な関係から, 構造未知のタンパク質の二次構造を推定する方法の研究が盛んに行なわれた.[5)~7)] この時代における構造予測の研究は, 最初に二次構造を精度良く予測し, ついで, 二次構造間の相互作用を考慮して, タンパク質の立体構造全体を推定することを目標としていた.[8)]

しかし, 構造既知のタンパク質の数が少ないこと, また, 二次構造は実際には配列の局所的特徴から決定されるのではなく, 配列上遠い残基同士の相互作用によって決まることが多いことなどが分かり, ヘリックス状態, ストランド状態, それ以外の状態の合わせて三状態のうちのいづれであるかを予測する三状態予測精度は, 60% を越えることはなかった. しかし, Rost らが, 残基配列のみ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)] を提案して以降, 精度は飛躍的に向上し, 70% 以上の二次構造予測精度を出す方法も提案された. Rost らの方法では, 構造既知のタンパク質配列と十分な配列相同性がある構造未知のタンパク質がほぼ類似の構造をもつタンパク質であると仮定して, 構造と配列とを対応させたタンパク質の数を大幅に増やすという方法である. しかし, この方法をもってしても, 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, 立体構造既知のタンパク質が数百種程度知られるようになった. このころ, Chothia らが, 「タンパク質の構造機能から考えられる種類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)] これが切っ掛けとなり, 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類似の構造をとるか」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)] すなわち, 現状で知られている数百種類の立体構造のパターンは, 構造未知のタンパク質の立体構造パターンのかなりの部分と共通であろうと考えるわけである.

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 の残基配列を糸を通すように当てはめ (縫糸し), 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)]

縫糸法の最も単純な方法は, 配列そのものの類似度 (あるいは相同性) を互換性尺度としても用いる方法である. 一般的に知られている事実として, タンパク質残基配列二つを整列 (alignment) させたときに, 相互で25% 程度以上のアミノ酸残基が同一であれば, その両者のタンパク質の立体構造はほぼ類似であると判断できる. この知見により, 構造未知のタンパク質であってもその残基配列が構造既知のタンパク質と十分に相同であれば, 配列相同性に基づく立体構造モデリング (comparative modeling) を行なうことが可能である. 既にこの方法は実用段階に入っており (ただし, 実際にはかなり相同性の高い場合でないと信頼性は乏しい), 創薬などの分野でも応用されている. しかし, 構造未知のタンパク質には, 構造既知のタンパク質と明らかな配列相同性を持たないタンパク質でも, 類似構造をとる場合がある.

そこで, 相同性がごく僅かでも, 構造が類似である可能性の高いタンパク質を構造既知タンパク質中から発見可能な方法が必要になる. こうした中で, **配列と構造との互換性評価方法**の研究として, 立体構造中の二次構造に着目し, 特定のアミノ酸残基がどのような二次構造を趣向するか, 及び, 特定のアミノ酸残基がどの程度構造中で埋もれているか, あるいは水溶液に対して露出しているか, といったアミノ酸残基とその残基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)]

しかし, 後の研究に多くの影響を与えたのは, Sippl が1990年に発表した研究である.[13)] これは同じ鎖の中に

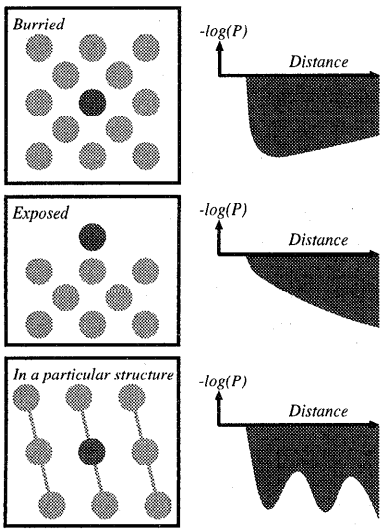


図1 アミノ酸残基を取り巻く環境と平均力場ポテンシャルの模式図(一次元)
Fig. 1 Mean force potential and the environment surrounding an amino-acid residue (1D)

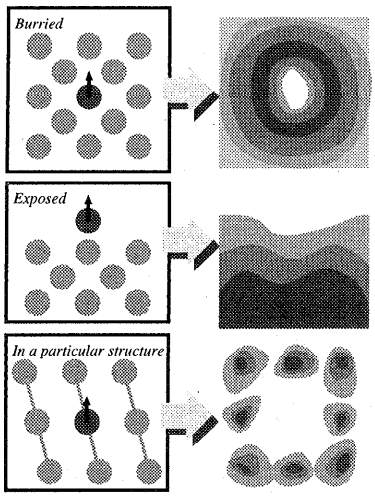


図2 アミノ酸残基を取り巻く環境と多次元ポテンシャルの模式図(二次元)
Fig. 2 Multi-dimensional potential and the environment surrounding an amino-acid residue (2D)

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.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,20種類のアミノ酸残基同士の考えられる400対について,同じ鎖にある残基対の相対距離をヒストグラム化し,その負の対数に熱力学的係数を掛けたものを,統計的平均力場ポテンシャルとして定義する.この定義は,タンパク質の立体構造がボルツマンの熱力学的平衡状態にあると仮定し,統計をとるためのサンプル数が十分大きいと判断できる場合においては,統計力学的に見て意味のある平均力場のポテンシャルの定義になっていると言える.タンパク質立体構造と残基配列との互換性については[14],平均力場ポテンシャルの和が最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(**図1**).

こうして,タンパク質立体構造予測問題は,単純な二次構造予測法の研究から,次第に,立体構造と残基配列との統計的関係に関する研究になり,それが,同じ鎖内に存在する残基対の統計的ポテンシャルの研究へと展開していくことになった.

Sippl らの提案した統計ポテンシャルの方法[13]は,その後,さまざまな研究者によって踏襲された.[15] しかし,それらの研究のほとんどは,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,この場合のポテンシャルは,考慮している残基対の残基間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.このような一自由度のポテンシャルでは,本来全く異なる位置関係にあるものでも,同様の距離にあれば同じ相対位置にあるものとして混同されることになる.特に二次構造においては,その規則性から,距離が同じであっても全く異なる位置関係にある残基ペアは相当数存在する(とくにシート構造の場合など).

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できない部分を,できるだけ精密に表現するために,アミノ酸残基対の相対位置関係の六自由度全て(相対位置三自由度と相対姿勢三自由度)について考慮した統計ポテンシャルを導く方法を提案するものである(**図2**).

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る.続いて,この方法をまず,隣接する残基間の三つの二面角(ϕ, ψ, ω)の分布を表現することに応用する.次に,隣接していない残基間の六自由度の相対位置分布の表現に応用する.さらに,得られた分布表現を用いて,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の解法を提案し,その推定精度をもって,本論文で提案した多自由度分布表現方法の妥当性を検証し,一自由度のポテンシャルに比べて,六自由度の分布表現がより優れていることを示す.最後に,この多自由度統計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.

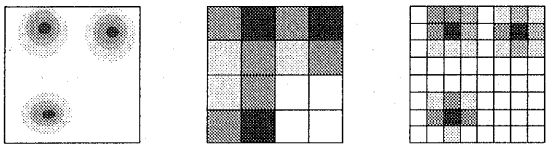


図3 度数分布による二次元分布の表現
Fig. 3 The representation of the two dimentional distribution by a histogram

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分布である.

分布の表現方法として, 最も単純な方法は, 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である (**図3**). 観測点の数が十分に大きければ, M を大きい数にすることができ, 分布の微細構造を表現することもできる. 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) である. したがって, 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 が比較的小さい値, たとえば, $N=6$ (六次元空間) 程度であっても, 各軸の分割数 M を例えば 10 にした場合, 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ない. この場合, 用意したパラメータの多くはゼロになり, パラメータ数が無駄になる. さらに, 観測点が発見される部分超立方体でも観測点を少数しか含まない場合は, 観測誤差が大きく, 統計として無意味になる場合がある.

必要とされるのは, 多自由度の分布を少数のパラメータでできるだけ正確に表現する方法の開発である. 本研究では, 観測点の分布を, 少ないパラメータで表現するために, Fourier 展開し, その展開次数の制限により, パラメータ数を圧縮する方法を考えた.

一つの観測されたサンプルは, その観測された点に δ 関数が存在すると考えることができるので, K 個のサンプルが存在する場合, その分布 $f(X)$ は, 以下の式で表される. ここで, X は, N 次元空間での座標 (あるいはベクトル) である.

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

この分布を Fourier 展開する. まず分布の範囲として考えている超立方体の一辺の長さを L とし, 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, 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. ここで, 添字 I 自身も N 次元のベクトルであり, 変数 X も N 次元空間の座標であり, 添字 j は N 次元ベクトルあるいは座標の j 番目の成分を表すとする.

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

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.

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

$$=\sum_J\int\delta(X-X_J)G_I(X)dX \tag{6}$$

$$=\sum_J G_I(X_J) \tag{7}$$

現実には, a_I を無限にたくさんとることで, $f(X)$ は精密に展開されることになる. しかし, $f(X)$ は本来観測点 $\delta(X-X_I)$ の和であるから, 平滑なものではない. そこで, $f'(X)$ という平滑化された分布関数を考える. この $f'(X)$ は, $f(X)$ において無限に観測点が得られた場合の分布と似たものであると考えられる. そしてこの関数は, G_I による展開を有限で打ち切ることで得られる. このとき分布 $f(X)$ は, 展開した個数だけのパラメータ数で表現される.

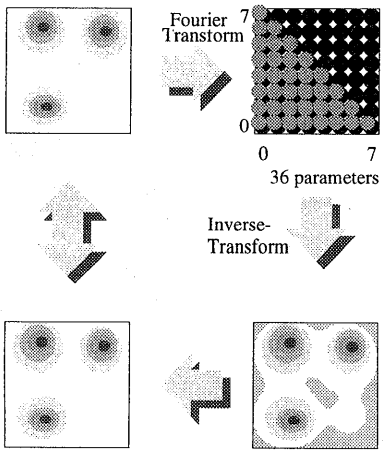


図4 Fourier 展開を利用したパラメータ圧縮
Fig. 4 The parameter compression by Fourier transform

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, パラメータ数が多いほど $f(X)$ の微細な構造が表現できる. これは, 部分超立方体内の観測点を数える度数分布を用いたものにくらべるとパラメータ数を選ぶ上で自由度が高い. 本研究においては, 打ち切り次数を, 各軸ごとの展開次数の総和が一定以下であるようにとることにする. N 次元でそれぞれの軸ごとの展開次数 I_k の和が $I_{\max}$ を越えないように設定することになる. すなわち, $\sum_k I_k \leq I_{\max}$ となるように選ぶ (図 **4**).

この場合, 特定の $k=l$ について, $I_k = I_{\max}$ であれば, 他の $k \neq l$ については, $I_k = 0$ である. したがって, $k=l$ である軸については, $I_{\max}$ 次で打ち切られるだけの解像度, つまり大体 $L/I_{\max}$ で表現されていることになるが, 他の軸については, 解像度が L であることになる.

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. これは, 各軸で, $I_{\max}$ まで展開したときよりも, $1/N!$ 程度少ない (図 **5**). しかし, この場合, 各軸での解像度は, $L/I_{\max}$ だけあり, また, 直交する二つの軸を考えたときに, それぞれ $I_{\max}/2$ までの次数については, 展開されているので, ある程度の解像度が保たれている. $N=2$ の二次元平面上の関数の近似については, ほぼ同様の方法として離散余弦変換 (DCT) による画像圧縮法[16] が, デジタル画像処理において利用されている.

このようにして, 線形展開とその展開打ち切り次数の制御により, 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. しかし, 一つ問題がある. a_I という展開系数で表現された分布は, 実際に利用するときには, もう一度元の基底 (つまり分布を表しているもとの座標系) での表現に直す必要がある. つまり, 逆変換をしなければならない. 本来, 分布は定義域内のいたるところで, 0 より大きな正値をとらなければならない. しかし, 展開は有限で打ち切られているので, 逆変換によって得られた $f'(X)$ は, 負値をとることがある. したがって, $f'(X)$ が負値をとる場合は, これを 0 あるいは, 0 に近い十分小さな正値になるように修正する必要がでてくる. これは, 応用問題によっては, 分布をサンプル量で割って確率密度とし, さらにこれの対数をとって平均場近似のエネルギー (あるいはポテンシャル) として評価する際に問題となる. 今後の研究の方向としては, 分布から推定される確率密度 $\rho(X)$ に対して, $\rho(X) = |\psi(X)|^2$ となる確率振幅 $\psi(X)$ を考え, この $\psi(X)$ を線形展開することが考えられる. 今回は, この問題に関して, 後述のようにポテンシャルの比を用いているので, 負値をとるときには単純に 0 とし, この結果として比を計算するときに分母が 0 になった場合は, その計算を無効とする (中立である 1 にする) ことで問題を解決している.

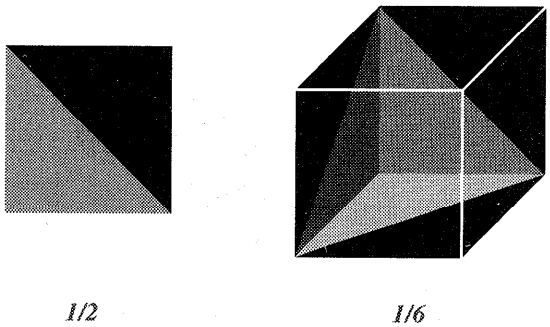


図5 パラメータ数圧縮の比率, 二次元の場合と三次元の場合
Fig. 5 The compression ratio for the two dimensional and three dimensional distribution

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, 最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 (図 **6**).

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 (主鎖二面角に直接関係の無い O 原子を除外している). 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. まず, 二面角 ϕ^d は, $C-N-C^\alpha$ が作る平面と $N-C^\alpha-C$

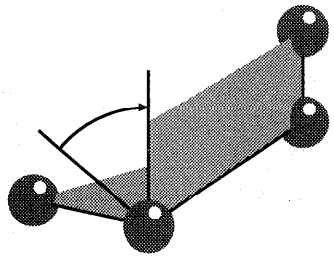


図6 二面角の定義
Fig. 6 The definition of a dihedral angle

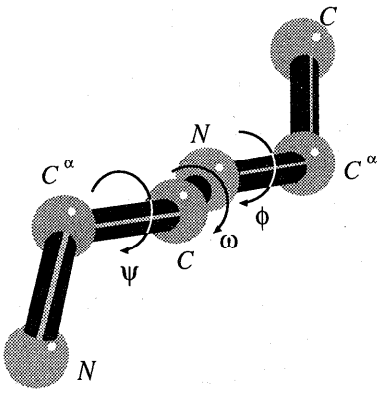


図7 主鎖を構成する原子と二面角 ϕ, ψ, ω
Fig. 7 Dihedral angles ϕ, ψ, ω along the main chain

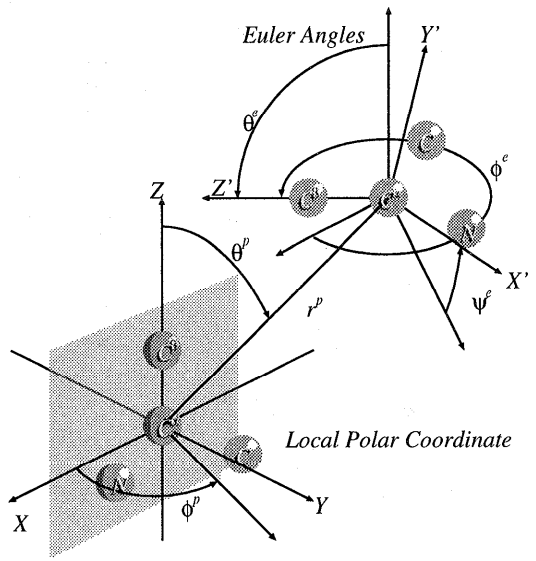


図8 局所座標と残基間相対位置
Fig. 8 The polar coordinate and the relative position of a residue pair

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 (後述の極座標での角度やEuler角と区別するために, ここでは, 上付きの d を用いている). 二面角は, その二面角を構成する結合で, 上流側 (N末端側) の原子から下流側 (C末端側) の原子向きに見て, 時計回りの方向を正方向にとる. 隣接する残基の上流側の残基Aと下流側の残基Bの二面角は, ψ^d, ω^d, ϕ^d とならんでいる (図7).

結合角度, 結合長が不変であると仮定すると, 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のみになる (実際には, 結合長, 結合角度ともに揺らぎがある). このうち, ω^d は, この部分の共有結合の二重結合性から, 理論的にはラジアンで0, あるいは, π のみをとることになっているが, ここでは, 分布計算を簡単にするために, ω^d についても, 他の二つと同様の解析をする. 実際に観測される ω^d の値は, 0 あるいは π の近傍にある程度の広がりで分布している.

ϕ^d, ψ^d, ω^d は, 角度であるから, ラジアンで 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.

例として, $I_{\max}$ を12にして (30度きざみ程度の解像度に対応) 解析すると, パラメータ数は, 455個になる. これは, どの軸も0次から12次まで展開する場合の数 $(12+1)^3=2197$ に比べると, およそ1/5である. $I_{\max}$ が十分大きい場合の推定値である $1/3!=1/6$ よりは多少大きいが, パラメータ数は大幅に少なくなっている.

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と, 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. これには, 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り, 合計で六自由度ある. 三次元相対位置は, デカルト座標系で考えてもよいし, また, 極座標系で考えてもよい. ここでは, タンパク質立体構造に関する研究で, 相対位置を決めるための最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ことを考慮し, 相対距離を陽に含む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Euler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8).

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, 単純な三角関数を用いた線形基底変換を行なうことはできない. この場合は, 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るのが最も妥当であり, 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, 2節で述べた三角関数 $g_i(x)$ 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.

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

これは, $r^p = r_{\max}$ を零点とする球Bessel関数 (及びこれに対応する球Neuman関数) の0次のものである. 分布を解析する目的には, 規格直交系であることが要請されるだけであり, 分布そのものがHelmholtzの方程式を満足する必要はない. したがって, 球面調和関数の量子数と合わせた球Bessel関数の高次のものを使う必要はない.

球面調和関数を用いる上で問題となるのは, θ^p 方向と ϕ^p 方向の展開打ち切り次数である. 三角関数を用いた単純な多次元 Fourier 展開と異なり, 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. ϕ^p 方向は, 単純な三角関数であるが, このときの量子数 m^p は, Legendre陪多項式で与えられる θ^p 方向の量子数 l^p を越えることができない.

本論文の研究では, そのことを考慮して, 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, Euler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. ここでも, θ^e 方向の展開打ち切り量子数 $l^e_{\max}$, ϕ^e 方向の展開打ち切り量子数 $m^e_{\max}$, 及び, ψ^e 方向の展開打ち切り量子数 $k^e_{\max}$ の間には, $k^e_{\max} = l^e_{\max} + 2m^e_{\max}$ なる関係を満たすようにした.

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, 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. このような打ち切り次数の設定では, 最も圧縮率を高くするようにすると $I_{\max}$ は3の倍数であることが要求される.

$I_{\max}$		3	6	9	12	15
k^p	(r^p)	3	6	9	12	15
l^p	(θ^p)	1	2	3	4	5
m^p	(ϕ^p)	2	4	6	8	10
l^e	(θ^e)	1	2	3	4	5
m^e	(ϕ^e)	2	4	6	8	10
k^e	(ψ^e)	3	6	9	12	15
パラメータ数		35	333	1717	6233	18023

表1 展開打ち切り次数とパラメータ数
Table 1 The expansion cut-off-orders and the numbers of parameters

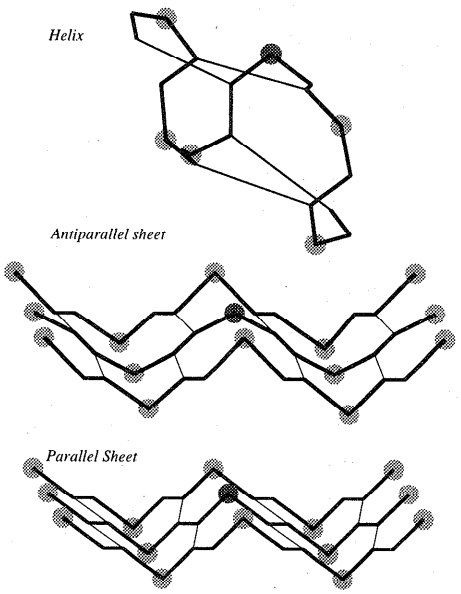


図9 二次構造中の残基相対位置
Fig. 9 The relative position of a residue pair in secondary structures

表1に打ち切り次数の和 $I_{\max}$ とそれぞれの自由度での打ち切り次数 $(k^p, l^p, m^p, l^e, m^e, k^e)$, 及びパラメータ数を掲げる.

本研究では, サンプル数と分布表現の解像度との兼ね合いから, $I_{\max}$ を $I_{\max} = 6$ にしてある. この場合の展開パラメータの総数は, 表にある通り333個である. また, $r^p_{\max}$ をアミノ酸残基間のよく使われる接触距離 ($C^\alpha - C^\alpha$ 間距離で計る) である 10Å として, 10にした. 解像度は, r^p 方向で, 10/7 =1.4Å, 角度方向は, $\theta^p, \phi^p, \theta^e, \phi^e$ については, 概ねラジアンで $2\pi/6$, ψ^e 方向については, $2\pi/7$ 程度になる. なお, これより解像度を上げる場合を考えると, 次に考えられる展開打ち切り次数の組合せは, $I_{\max} = 9$ の場合であるが, このとき展開パラメータ数は, 1717 個になり, 利用できるサンプル

数を大幅に越える場合が出てくる.

この解像度は, タンパク質中のあるアミノ酸残基が, 近傍の他の残基にどのように囲まれているかを考慮する上で, そのアミノ酸残基が二次構造中にあるとした場合, 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. まず動径方向についての解像度 1.44Å は, 分子中の共有結合する原子間距離とほぼ同じであり, また, 角度方向の解像度は, 60 度程度で, これは, たとえば, アミノ酸がシート構造中にある場合, 近傍の水素結合をしているアミノ酸残基の位置のそれぞれをかろうじて区別することのできる角度である, またヘリックス中にある場合も, 配列中隣接していない残基の位置をかろうじて区別できる角度である. しかし, これまで, 残基間距離だけについての分布関数を用いる一自由度の場合では, 残基がシート中に存在する場合, 水素結合している近傍の残基はほぼ同じ距離にあり, これらを区別することができない. さらに, シートにおける並行ストランドと逆並行ストランドでの水素結合している残基間距離は 0.5Å 程度異なるものの, 多少ゆがんだシートなどが多く存在することを考えると, 距離だけで考えた一自由度の分布関数では, これらの違いを区別することはできない. しかし三次元相対姿勢まで考えた六自由度の分布では, 角度 ψ^e がおよそ π 異なり全く逆向きであるから, 角度方向の解像度が低くても, 全く別のものとして区別できる. このことは, ヘリックス中の残基についても言えることで, それぞれの軸ごとの解像度はかろうじて異なる位置関係を分離できる程度のものであるが, 六自由度の分布では, 全く違う位置関係であるように表現されている (**図 9**) したがって本研究において, この展開次数の設定による分解能は十分である.

距離のみを考慮した一自由度の分布をヒストグラムで解析する場合, その解像度は, 0.5Å 程度である.[15] このとき, 分布を記述するパラメータ数は, 距離の範囲を 0Å から 10Å 程度にした場合, 20 個である. これに比べて, 本研究で提案する方法では, 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 333 個である. 表現能力と精度を大幅に向上させつつパラメータ数は 15 倍程度に押え込むことが可能になった.

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の周囲の環境は, その種類によって異なることが知られている. たとえば, アミノ酸残基の中で, その側鎖部分が溶液中で電荷を持つ場合や極性をもつ場合は, 周囲に強い極性をもつ水分子が存在するほうが安定であるから, 水分子との親和性が高く (親水性), よってタンパク質分子の中では, 表面に近い水溶液に接する場所に配置される傾向が強い. 逆に, メチル基などをもち, 極性がなく電荷も帯びていない残基は, 水分子との親和性が悪い (疎水性) ので, タンパク質分子においては, 中心部に配置される傾向がある. また, その分子構造上二次構造形成能力が異なるために, ある種のアミノ酸は, ヘリックスを形成しやすいものもあれば, ストランドを形成しやすいものもある.

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, 周囲にどのように残基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. したがって, 親水性残基の周囲には, 統計的にみて多くの残基が分布することはなく, 逆に疎水性残基の周囲には多くの残基が分布する.

また, 二次構造趣向性の違いにより, ヘリックス内に存在することの多い残基種では理論的にヘリックス構造に対応する相対位置の場所に他の残基が分布するような統計結果を得るし, シート内に存在することの多い残基種では, シート固有の相対位置の場所に他の残基が分布するような統計結果を得る.

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. ヘリックスや, シートを構成するストランドの場合は二面角は理論的に決定されている (実際には揺らぎがある). したがって, この残基間二面角の分布は, 二次構造趣向性と強い関係がある. また, 分子構造上, 二面角について強い制約を受けるプロリンは, 特殊な二面角分布を示し, また側鎖の存在しないグリシンについては, 他の残基に比べて, 側鎖による制約がないので, 二面角の自由度が非常に高く, この場合も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, それぞれの場所にどの種類のアミノ酸残基が来ることが尤も確からしいかを, その統計的に得られた分布から推定することができる. これは, タンパク質立体構造予測問題の**逆問題**であり, ここでは, **立体構造からの残基配列推定問題**と呼ぶことにする.

従来から縫糸法においては, 相対位置を考慮する場合, その対を構成する両残基の残基種を考慮して統計処理するのが一般的である. この場合, 20 種類の残基種があることから, 400 種類の統計処理をする必要がある. しかし, この場合, 考えている残基対の相手側の残基種を特定しないと残基配列推定問題が解けない. 相手側の残基種を実際の残基種 (構造を与えているタンパク質の対応する場所の残基種) に固定する方法 (frozen approximation)

もあるが，これでは配列が未知の場合に拡張することができない．そこで，今回は，相対位置の対の残基種の片方だけを考慮することにした．もちろんこれによって，個々の統計をとるためのサンプル数は，残基対の双方の残基種別で統計をとる場合に比べて約 20 倍利用できることになる．

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 $f_k^a(X)$ 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_k^a で分布 $f_k^a(X)$ を割った $f_k^a(X)/m_k^a$ を相対頻度分布 $\rho_k^a(X)$ 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，以下のことが言える．

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

$$m_k = \sum_a m_k^a \tag{10}$$

ここで，アミノ酸残基種別にとられた分布 $\rho_k^a(X)$ と種類を問わない $\rho_k(X)$ の比 $\rho_k^a(X)/\rho_k(X)$ 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尤も高い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 の上限値及び下限値は簡単には定まらない．また，k の絶対値が大きければ立体構造における相対距離も大きくなる傾向があり，この距離が $r_{\max}$ を越えることも多いので，k が大きくなれば，観測されるサンプル数 m_k は小さくなる．あまりにもサンプル数が少ない場合は，統計として意味をなさない．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，配列上連続した残基で $r_{\max} = 10$ 以下の距離に入るのは，ヘリックスの場合 $k = 6$ 程度，ストランドの場合 $k = 3$ 程度であること，そして他の不規則な構造なども考慮して，$k = 8$ 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．

残基間相対位置について，三次元相対位置と三次元相対姿勢を合わせた六自由度の分布をとるために，統計パラメータは 333 個用いていることを前節までで述べた．また，隣接する残基すなわち，$k = \pm 1$ では，二面角系を用いて三つの二面角で相対位置関係を表現し，そのために，統計パラメータは 455 個用いている．このパラメータ数に対して十分なサンプル数は，パラメータ数の数倍程度は必要であり，1000 程度必要であると考えられる．タンパク質立体構造データで良質なものが数百程度しか得られない現状では，希少アミノ酸であるトリプトファン等では，$k = \pm 8$ などの場合に，数百しかサンプルが存在しないことがある．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たことにできるようにする必要がある．そこで，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.

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

6. 検証実験

統計ポテンシャルの有効性を実証するために, 一般的には自己認識率テストと呼ばれる検証法がある (いわゆる "Sippl テスト" もこの方法の一つである). これは, 検証セットとして選ばれたタンパク質立体構造を含めた多数のタンパク質立体構造に対して, 検証セットの配列を縫糸した場合に, 自分自身の構造との互換性が最も高くなる率を求めるものである. しかし, このテストでは, 配列と構造との整合手法 (alignment), その際のギャップコスト問題など複雑な問題がつきまとい, ポテンシャルそのものの有効性の判断基準とするまでには長い道のりがある. そこで, ここでは, より直接的な方法としてアミノ酸残基配列推定問題を解くことを, 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, 質の高い (高解像度のデータでかつ主鎖原子に欠損がない) 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[17] 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.

統計データセット内に精度検証のための検証セットが含まれないようにするため, 911 個から, さまざまな配列長の 46 個のデータをとりだし, これを検証セットにし, これを統計データセットから外した. 残基配列**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 (正解であるはずの) 残基種のスコアの和 S_{all}^{true} と, 推定した残基種のスコアの和 $S_{all}^{predict}$ の比 $S_{all}^{true}/S_{all}^{predict}$ (**正解スコア比**) も計算した. 前節で与えられるスコアが 0 の場合, その場所に残基種が起こる可能性が平均的であることを示し, 負数の場合は, 平均よりもその残基種が起こる可能性が高いことを意味する. したがって, 推定される残基種は必ず負数のスコアを持つ. このスコアの和の比が 1 に近ければ, 正解の残基種を負のスコアとしていることになり, よって, 縫糸法などに応用する場合に有用であることを示し, 0 に近い場合は, 正解残基種についても悪いスコアを出していることになり, 縫糸法に応用する際に注意を必要とすることになる. さらに, 正解となる残基種のスコアが 0 以下である割合 (百分率) も計算した (表中の**スコア 0 以下率**). この比率が 100% であれば, 正解である残基種は, つねに平均以上の分布を示すという判断を推定装置が行なったことになる.

自由度	σ	推定精度	正解スコア比	スコア 0 以下率
1	1/100	17.3%	0.16	57.4%
6_{cross}	1/100	19.6%	0.11	59.4%
6	1/100	20.3%	0.14	60.7%
6	1/500	20.6%	0.20	61.7%
6	1/2000	20.7%	0.23	62.2%
6	1/10000	20.6%	0.31	60.7%

表 2 自由度の違いに対する残基推定精度
Table 2 The residue-prediction accuracy with respect to the degree of freedom

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. もし, 統計をとる際に過学習が行なわれていれば, この精度は検証セットを統計データに入れないもの (表中 "6_{cross}" と表記) に比べて, 推定精度, 正解スコア比ともに大幅に良くなるはずである. また, この場合については, σ を 1/100 から 1/500, 1/2000 及び 1/10000 に変更した場合についても調べた. 結果を**表 2** に掲げる.

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. 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. これは, 分布の表現がより正確であるため, 表現された分布とずれたところに観測点があった場合に, 悪いスコアを出していることを表している. 一自由度の場合は, 分解能が低いのでこれらを混同して, 良いスコアを出してしまうことを意味する.

σ の値をより小さく変更すると, **推定精度**, **正解スコア比**, **スコア 0 以下率**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. しかし, あまりにも混合しすぎると, たとえば, $\sigma = 1/10000$ の場合のように推定精度が上がらなくなることも分かる.

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. これは, 多自由度の分布を正規直交線形基底により展開し, その展開打ち切りを制御することで, 圧倒的に少ないパラメータ数で分布を表現することを可能にする方法を開発し, アミノ酸残基対の相対位置関係を三次元相対位置と三次元相対姿勢の合わせて六つの自由度についてその分布を解析するものである.

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用し, 推定精度が, 従来法にくらべて良いことを示した. 一方で, 推定を誤った場合については, 従来からの一自由度の分布に基づく推定では正しいアミノ酸残基種のスコアがそれほど悪くないのに対して, 六自由度についての推定では, スコアが極端に悪い場合があることが判明した. この傾向は, 推定精度を検証する試験用データを, 統計データをとるためのデータに含めた場合でも同じ傾向が見られた. このことから, タンパク質中では, その立体構造を支えるために重要な残基と, 立体構造を支える上では安定性に寄与しない残基が存在することが示唆される.

なお, 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. すなわち, タンパク質において, 立体構造を支えているのは, その配列中の 20% 程度であり, そのほか, そのタンパク質の機能の実現などのために 5% 程度重要なアミノ酸残基が存在することを示唆する. そして, それ以外のアミノ酸残基は, 構造を破壊する要因にならないかぎりにおいて, ある程度自由に選ぶことができる可能性もある. このことは, 将来において, 相対位置関係を解析する残基対における双方の残基種を考慮した統計をとって残基配列推定を行なった際に, どの程度の推定精度が出るかによって更に詳細な議論ができるものと思われる.

8. まとめと研究の今後

サンプル数の問題から, 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における双方の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, 残基対における双方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.

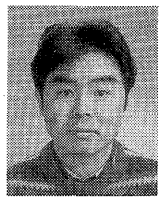
六自由度の分布解析で縫糸法が可能になれば, 次に行なうことは, 縫糸法でえられた立体構造モデルをより一層現実的な立体構造にする方法の開発である. 縫糸法で用いた分布から得られるエネルギー関数を用いて局所的構造最適化を行なう方法が一般的である. 今回提案した六自由度のエネルギー関数を, これに適応した構造最適化について研究を進めていきたい.

参 考 文 献

1) Sato, C., Sato, M., Iwasaki, A., Doi, T., and Engel, A.: The Na+ channel has four domains surrounding a channel, *J. Struct. Biol.*, Vol. 121, pp. 314-325 (1998).
2) Chou, P. Y., and Fasman, G. D.: Prediction of protein conformation, *Biochemistry*, Vol. 13, pp. 222-244 (1974).
3) Lim, V. I. : Algorithms for prediction of α-helices and β-structural regions in globular proteins, *J. Mol. Biol.*, Vol. 88, pp. 873-894 (1974).
4) Garnier, J., Osguthorpe, D.J. and Robson, B.: Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol.*, Vol. 120, pp. 97-120 (1978).
5) Bohr, H., Bohr, J., Brunek, S., Cotterill, M. J. R., Lautrup, B., Norskov, L., Olsen, H. O. and Pertersen, B. S.: Protein secondary structure and homology by neural networks, *FEBS Letters*, Vol. 241(1, 2), pp. 223-228 (1988).
6) Qian, N., and Sejnowski, T. J.: Predicting the secondary structure of globular proteins using neural network models, *J. Mol. Biol.* Vol. 202, pp. 865-884 (1988).
7) King, R. D. and Sternberg, M. J. E.: Machine learning approach for the prediction of protein secondary structure, *J. Mol. Biol.*, Vol. 216, pp. 441-457 (1990).
8) Fasman, G. D. (editor): *Prediction of Protein Structure and the Principles of Protein Conformation* New York: Plenum Publishing Corporation (1989).
9) Rost, B. and Sander, C.: Prediction of Protein Secondary Structure at bettter than 70% Accuracy, *J. Mol. Biol.*, Vol. 232, pp. 584-599 (1993).
10) Chothia, C.: One thousand families for the molecular biologist, *Nature*, Vol. 357 18-JUN, pp. 543-544 (1992).

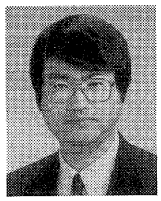
11) Bowie, J. U., Lüthy, R. and Eisenberg, D.: A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure, *Science*, Vol. 253, pp. 164-170 (1991).
12) Yue, K. and Dill, K.: Inverse protein folding problem: Designing polymer sequences, *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 4163-4167 (1991).
13) Sippl, M. J.: 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.*, Vol. 213, pp. 859-883 (1990).
14) Sippl, M. J. and Weitckus, S.: Detection of Native-like Models for Amino Acid Sequences, *PROTEINS: Structure, Function, and Genetics*, Vol. 13, pp. 258-271 (1992).
15) Melo, F. and Feytmans, E.: Novel Knowledge-based Mean Force Potential at Atomic Level, *J. Mol. Biol.*, Vol. 267, pp. 207-222 (1997).
16) Wallace, G. K.:The JPEG Still Picture Compression Standard, *CACM 34*, No. 34, pp. 30-44 (1991).
17) Noguchi, T., Onizuka, K., Akiyama, Y. and Saito, M.: PDB-REPRDB, A Database of Representative Protein Chains in PDB (Protein Data Bank), *Proc. the Fifth International Conference on Intelligent Systems for Molecular Biology*, pp. 214-217 (1997).

(平成10年9月10日受付)
(平成10年10月29日再受付)
(平成10年11月9日採録)



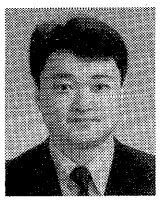
鬼塚健太郎

昭和38年生．平成2年東京都立大学理学研究科物理学専攻修士課程修了．同年松下電器産業（株）入社．（財）新世代コンピュータ技術開発機構（ICOT）に出向，同機構の開発する並列論理型コンピュータの遺伝子情報処理分野への応用に関する研究に従事．平成6年より松下技研（株）において音声認識の研究，平成7年より松下電子工業（株）において画像処理プロセッサの開発に従事，平成8年技術研究組合新情報処理開発機構に出向，並列応用つくば研究室にて並列計算機上でのタンパク質立体構造解析に関する研究に従事．日本生物物理学会会員．



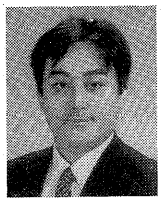
野口　保（正会員）

昭和32年生．昭和58年東京農工大学大学院工学研究科応用物理学専攻修士課程修了．昭和59年富士通（株）入社．同年富士ファコム制御（株）に出向．コンピュータケミストリ分野のシステム開発に従事．平成元年蛋白工学研究所（PERI）に出向．タンパク質立体構造予測の研究に従事．平成3年より富士通（株）にてバイオ分野のシステム開発に従事．平成6年から平成7年にかけて九州工業大学情報工学部受託研究員．平成8年技術研究組合新情報処理開発機構主任研究員．並列応用つくば研究室にて，並列計算機を用いたタンパク質立体構造予測の研究に従事．日本物理学会会員



安藤　誠（正会員）

昭和42年生．平成4年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．同年日本鋼管（株）入社．データベース（関係型，オブジェクト指向など）のデータモデルに関する研究に従事．平成6年同社より米国コンベックスコンピュータ社（現ヒューレットパッカード社）に派遣．並列計算機Exemplarシリーズの OS 開発グループに所属．平成8年技術研究組合新情報処理開発機構に出向．並列応用つくば研究室にて，並列計算機上でのタンパク質立体構造解析等の研究に従事．



秋山　泰（正会員）

昭和36年生．平成2年慶應義塾大学大学院理工学研究科電気工学専攻博士課程修了．工学博士．同年通産省電子技術総合研究所研究官．平成4年京都大学化学研究所助教授．平成8年技術研究組合新情報処理開発機構並列応用つくば研究室室長．現在に至る．並列計算機を用いたタンパク質立体構造及び遺伝子配列情報解析等の研究に従事．電子情報通信学会，日本生物物理学会，分子生物学会，神経回路学会，IEEE 各会員．