数理モデル化と問題解決 22−13
（1998. 11. 27）

多次元分布の線形変換による圧縮表現のタンパク質立体構造認識問題への応用

鬼塚 健太郎† 野口 保†
安藤 誠† 秋山 泰†

多次元分布は，その分布を線形基底を用いて基底変換し，その基底の次数打ち切りを行なうことで，少ないパラメータで表現することができる．この方法を，同一タンパク質中に含まれるアミノ酸残基の対の相対位置分布に適応し，タンパク質の縫糸法による自己構造認識問題に応用した．

アミノ酸残基間距離だけを用いた一自由度の場合，相対位置だけを考慮した三自由度の場合，および相対姿勢も考慮した六自由度の場合それぞれについて，タンパク質の残基配列が自分自身の構造を認識する正解率を求め（自己認識率），従来の相対距離にのみ着目した一自由度の統計を用いた場合よりも，多自由度，での統計を用いた方法のほうが，自己認識率が高いことが判明した．

A compressed representation of multiple-dimensional distribution by linear base transformation, and its application to protein 3D structure recognition

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

Multiple dimensional distribution is represented with fewer number of parameters by linearly expanding the distribution and controlling the cut-off orders of expansion. We adopted this method to the distribution of the relative position between two amino-residues in a protein chain, and applied it to the protein fold recognition problem.

We compared the recognition ratio of three cases, adopting the distribution 1) with respect to the distance (one degree of freedom) , 2) with respect to the 3D position (three degrees of freedom), and 3) with respect to the 3D position and the relative orientation (six degrees of freedom). The result is that the self-recognition ratio of multiple dimensional distribution is better than that of the conventional distribution with respect only to the relative distance.

1. 研究の背景

「タンパク質立体構造予測問題」は，数十年にわたって，多くの研究者が取り組んできた，分子生物学における最も重要な未解決問題の一つである．

タンパク質はアミノ酸とよばれる物質が鎖状に結合した分子であり，生体に存在するタンパク質のアミノ酸の種類は 20 種類である (21 種類以上とする考えもある)．近年の技術的進歩により，タンパク質のアミノ酸残基の並び方，すなわちアミノ酸残基配列を決定する技術は飛躍的に進歩し，高速化したため，タンパク質の配列を決定することはそれほど難しくはなくなった．そこで，アミノ酸残基配列から，タンパク質の立体構造を予測することができれば，タンパク質の立体構造を実験的に決定することを待つまでもなく，そのタンパク質の機能の解明などが行なわれるようになる．

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

タンパク質の立体構造予測を研究している研究者たちは，当初は頻繁に現れる特徴的な螺旋 (ヘリックス) 構造やシート構造の骨格となるストランド構造などの二次構造に着目し，配列から二次構造を推定する「二次構造予測問題」についてさまざまな研究を行なった.[1)2)] しかし，実用的な精度に到ることはなかった.[3)] また二次構造を組み上げた三次構造 (すなわち立体構造そのもの) を予測することについては，ほとんど絶望的な状態が長らく続いている．

近年になって，タンパク質の種類が実質的には千種類程度しかないのではないかという見解が出された.[4)] これが契機となって，タンパク質の折り畳み認識 (Fold Recognition) の方法が脚光を浴びるようになった．構造未知のタンパク質の配列が，構造既知のタンパク質と立体構造が類似しているかどうかを判定する方法である．現在ではこの方法は縫糸法 (Threading) と呼ばれることが多い．現状において，タンパク質の立体構造は数百の典型的なパターンがあることが知られている．そこで，構造未知のタンパク質の配列が，これら典型的な立体構

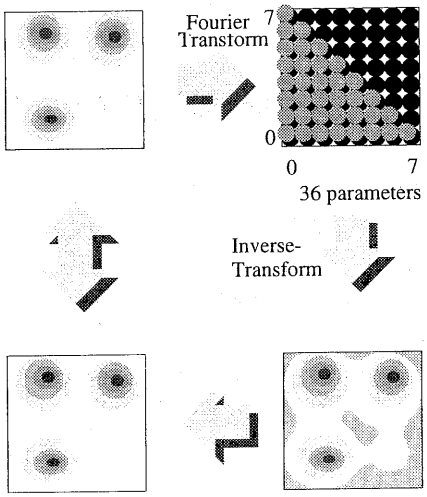

図1 Fourier 展開を利用したパラメータ圧縮

Fig. 1 Parameter compression by Fourier transform

造パターンのいづれか一つをとる可能性があるかどうかを, 構造と配列との互換性尺度を用いて検討するわけである.

Sippl は, 一つのタンパク質の鎖上に存在するアミノ酸残基対の双方の残基種とその三次元的距離の分布を統計計算し, 距離と頻度から得られる統計力学的な平均力場ポテンシャルを用いる互換性尺度を提案した.[6] この方法は非常に注目され, 多くの研究者がこれを踏襲した方法を研究している.[8] Sippl の研究, 及び Sippl らの方法を踏襲した研究においては, その平均力場ポテンシャルは残基対の相対距離にのみに依存する一自由度のポテンシャルである. アミノ酸残基対の相対的な位置関係は, 本来相対位置, 相対姿勢の合わせて六自由度で表現されるなければならない. 従って, Sippl らの研究やそれを踏襲する研究は, かなり大雑把な近似を行なっていることになる.

本研究は, 平均力場ポテンシャルが一自由度, 三自由度, 六自由度それぞれの場合について, タンパク質の立体構造と残基配列との互換性尺度としての能力がどの程度違うかを検討したものである.

多次元ポテンシャルとそのパラメータ数圧縮技術については, 文献[9] において詳細に議論しているので, 本論文では, 次節でそのあらましを紹介し, 続いてこれを用いた残基配列推定問題, 最後に自己構造認識問題への展開を述べる.

2. 多次元分布の線形変換による圧縮表現法

本研究で必要とされるのは, 二つのアミノ酸残基の相対位置関係を完全に表現する相対位置 (三自由度) と相対姿勢 (三自由度) の合わせて六自由度での分布を効率よく

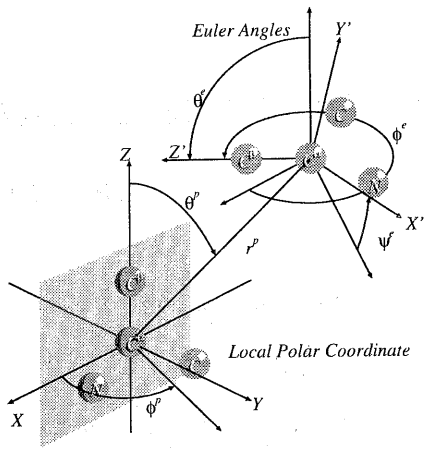

図2 局所座標と残基間相対位置関係

Fig. 2 Polar coordinate and the relative position between two resideues

少ないパラメータで表現する方法である. 分布の表現方法として一般的に考えられる度数分布を用いた場合, 六次元分布を表現するために, 各自由度の軸ついて, たとえば10分割して度数分布をとると, それだけで百万に及ぶパラメータが必要となる. そこで, 分布を線形展開し, その線形展開の打ち切り次数を調節することで, 少ないパラメータで多次元分布を表現する方法を開発した (図1). この方法では, 単純な直交座標系での N 次元分布においては, 度数分布をとるときのパラメータ数に比べて $1/N!$ 程度の数のパラメータで分布を表現できる (後述する極座標系の場合はこれほど効率良くない).

この節では, タンパク質の残基対の相対位置の分布を表現するために用いた方法を簡単に述べる.

2.1 残基対の相対位置関係の分布

本研究においては, 相対位置の表現のために従来法の相対距離を含む必要性から極座標系を用い, 相対姿勢の表現のために Euler 角を用いる. よって, 単純な直交座標系よりは複雑な処理を必要とする. まず, 線形展開のための基底として, r^p, θ^p, ϕ^p で表現される極座標系において, 角度成分 θ^p, ϕ^p についての線形基底は球面調和関数である. 次に動径成分 r^p については, その積分量 dr^p を考慮して, 球 Bessel 関数 (および対応する球 Neuman 関数) の零次のもので, 考慮している $0 \leq r^d \leq r_{\max}$ で規格直交系をなすものを用いる. そして, 相対姿勢を表現する Euler 角 θ^e, ϕ^e, ψ^e については, まず θ^e, ϕ^e について球面調和関数を用い, ψ^e については, 通常の三角関数を用いることにする (図2). 展開打ち切り次数の調節は, 単純な N 次元空間での場合よりは面倒である. この点については, 文献[9] に詳しい議論がある. 簡単に述べると, 各自由度の展開打ち切り次数は, 全ての自由度での打ち切り次数の和を固定することで得られる. この打ち切り次数の和を $I_{\max}$ とする.

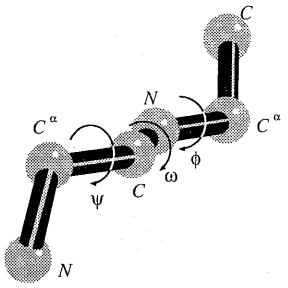

図3 主鎖を構成する原子と二面角 ϕ, ψ, ω
Fig. 3 Dihedral angles ϕ, ψ, ω along the main chain

三自由度 (相対姿勢を表すEuler角を考慮しない場合) の場合は, 後述の六自由度のものと空間解像度が同じである $_{\max} = 6$ のものと (三自由度 (1)), 空間解像度を一自由度のものと一致させた $I_{\max} = 15$ であるものを選び (三自由度 (2)), 六自由度の場合は以下の表からパラメータ数の妥当な $I_{\max} = 6$ であるものを選ぶ.

表中において, k^p, l^p, m^p, l^e, m^e, k^e は r^p, θ^p, ϕ^p, θ^e, ϕ^e, ψ^e のそれぞれの自由度での打ち切り次数の最大値である. またこの表では, $I_{\max}$ を制限しない場合の無圧縮の場合のパラメータ数も提示した. 三自由度の場合は, 無圧縮であっても, 解析不可能なほどのパラメータ数の爆発はないが六自由度の場合は解析不可能なほどの爆発が起こっていることが分かる.

		六自由度				
$I_{\max}$		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	総数	35	330	1717	6233	18023
(無圧縮時)	総数	256	3969	25600	105625	331776

		三自由度				
$I_{\max}$		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
パラメータ	総数	10	37	92	185	326
(無圧縮時)	総数	16	63	160	325	576

2.2 隣接する残基対の相対位置関係の分布

本研究において, タンパク質中で隣接する残基対の相対位置関係については, 一般的に知られている二面角系 ϕ^d, ψ^d, ω^d を用いている. これは, 本来六自由度ある相対位置関係であるが, 化学結合上の制約により, 結合している原子間距離や原子間結合角が概ね一定であることをから, 自由度が大幅に制約を受け, 結果として三つの二面角に自由度が押し込められるためである (図3). ここで ω^d は, 理論的にはラジアンで 0 (*cis* 状態) または, π (*trans* 状態) のみをとることになっている. *cis* 状態になるのは, 残基プロリンである場合にほぼ限定されていて, かつ, 全残基の 1/1000 程度しか存在しない. 本来は, *cis* 状態, *trans* 状態を分けて別の統計をとることで, ϕ^d, ψ^d のみの二自由度の統計をとれば十分であるが, サンプル数などの問題をがあるため, 統計処理を簡単にするために, ω^d についても, 他の二面角と同様に処理した. 本研究においては, 二面各自由度については, $I_{\max}$ を 11 とし, ϕ^d, ψ^d, ω^d の三自由度についてパラメータ数は 364 個である.

3. 残基配列推定問題

本節では, 前節でのパラメータ圧縮法を用いた場合の残基配列推定問題について説明し, 平均力場ポテンシャルが一自由度, 三自由度, 六自由度のそれぞれの場合における残基配列推定精度を検証する. 残基配列推定は, 次節で述べる縫糸法の基礎となるものであり, この推定精度の成否が縫糸法における能力を決定すると考えられる.

この節では, まず残基対の平均力場ポテンシャルの定義, およびその統計処理の方法を述べ, 次に残基配列推定問題への展開を述べる. 最後にポテンシャルが一自由度, 三自由度, 六自由度それぞれの場合について, その残基配列推定精度を検証結果として掲げる. そして, この結果に関する考察を行なう.

3.1 残基対の平均力場ポテンシャル

あるアミノ酸残基の種類 a について, その残基がタンパク質鎖中にある場合, その鎖の配列上で k ずれた場所にアミノ酸残基があり, その種類が b であるとき, 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 $f_k^{ab}(X)$ とする. アミノ酸残基は, 配列上での上流 (N末端) 方向と下流 (C末端) 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.

この残基対をなす二つの残基の間の距離は最大値 $r_{\max}$ 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^{ab} で分布 $f_k^{ab}(X)$ を割った $f_k^{ab}(X)/m_k^{ab}$ を相対頻度分布 $\rho_k^{ab}(X)$ とする. この数の対数に -1 をかけたもの, すなわち, $-\log(\rho_k^{ab}(X))$ が, 残基の種類 a と b が配列上 k 離れている場合の平均力場ポテンシャルである.

つぎに, 対をなす双方の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. ここで, 以下のことがいえる.

$$f_k(X) = \sum_a \sum_b f_k^{ab}(X) \tag{1}$$

$$m_k = \sum_a \sum_b m_k^{ab} \tag{2}$$

このとき, $-\log(\rho_k(X))$ は, 対をなす双方の残基種を考慮しない場合の平均力場ポテンシャルである.

なお, 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．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．本研究では，配列上連続した残基で $r_{\max} = 8$Å 以下の距離に入るのは，ヘリックスの場合 $k = 4$ 程度，ストランドの場合 $k = 2$ 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，$k = 4$ を上限，$k = -4$ 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．

一つ大きな問題点としては，残基対双方の残基種を考慮して統計をとる場合，サンプル数が十分にないということである．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とすることができるようにする必要がある．そこで，残基種を考慮した相対頻度分布 $\rho^{ab}(X)$ と残基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 $\rho^{ab}(X)$ に近くなるような処理を行なう．

$$\varrho^{ab}(X) = \frac{\rho(X) + m_k^{ab}\sigma\rho^{ab}(X)}{1 + m_k^a\sigma} \tag{3}$$

この式で，$\varrho^{ab}(X)$ は，サンプル数 m_k^{ab} が $1/\sigma$ 個程度あったときに，$\rho^{ab}(X)$ と $\rho(X)$ との混合比が 1:1 になり，m_k^{ab} が $1/\sigma$ より十分大きければ $\rho^{ab}(X)$ に近付き，小さければ $\rho(X)$ に近付く．今回の研究では，三自由度 (2)，六自由度の場合にパラメータ数がおよそ三百強になる．そこで，パラメータ数との関係から，$1/\sigma$ を 500 とした．この場合，サンプル数が例え零個であっても，最低 500 個のサンプルが得られたかのような結果が得られることになる．

3.2 残基配列推定方法の詳細

平均力場ポテンシャルの値は，残基対の相対位置関係が確定し，かつ双方のアミノ酸残基の種類が確定された場合に初めて決定される．よって，配列推定を行なう上で，それぞれの場所での残基種を推定するためのスコアを決定するためには，残基種を推定すべき残基の近傍半径 $r^p < r_{\max}$ 内の全ての残基の残基種が既に決定されている必要がある．しかし，立体構造のみが与えられた場合，これら近傍の残基の種類は決定できないので，残基種を推定すべき残基について残基種ごとのスコアの推定は不可能である．いわゆる凍結近似 (Frozen approximation) により，残基種を推定すべき残基の残基種別スコアを決定する際に，近傍にある (ポテンシャルを形成する対の) 相手側の残基種は (立体構造を与えているタンパク質の配列から参照する) 本来の残基種を用いることにする．したがって，本研究での残基配列推定方法は，構造既知で配列未知のタンパク質の配列推定には利用できない．

対をなすアミノ酸残基双方の種類別にとられた分布 $\rho_k^{ab}(X)$ と双方の種類を問わない $\rho_k(X)$ の比 $\rho_k^{ab}(X)/\rho_k(X)$ 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．

$$S_k^{ab}(X) = -\ln\left(\frac{\varrho_k^{ab}(X)}{\rho_k(X)}\right) \tag{4}$$

このを，ここでは，統計スコアと名付ける．そして，与えられた立体構造に対して，着目している残基について，その残基の種類が a であるスコアを以下の式で計算する．

$$S^a = \sum_k S_k^{ab}(X) \tag{5}$$

凍結近似であるから，このときの対をなすもう一方の残基の種類 b は，立体構造を与えているタンパク質の残基配列そのものから与えられる．こうして，この S^a の値が最も小さい値になる残基種 a が，その場所の残基の種類として尤も相応しいことになる．

3.3 残基配列推定の検証

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配列相同性が 45% を越えないように PDB-REPRDB[10] から 911 個の立体構造 (主鎖) を選びだし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．

正解でき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．以下に検証結果を掲げる．

自由度	残基推定精度	正解スコア比	スコア 0 以下の割合
二面角のみ	18.2%	0.49	0.71
一自由度	18.9%	0.46	0.73
三自由度 (1)	20.4%	0.45	0.74
三自由度 (2)	21.0%	0.45	0.73
六自由度	21.3%	0.48	0.77

まず，相対位置関係のうち，隣接残基からの二面角自

由度に基づくポテンシャルだけを受けて残基配列推定を行なった結果と一自由度のポテンシャルでの推定結果とでほぼ同一の推定精度が得られていることが分かる．それに対して，三自由度 (2) の場合，および六自由度の場合はかなり推定精度が向上していることが明らかになってくる．正解スコア比については，二面角のみの場合が最も良く，ついで六自由度のものが良い．スコア 0 以下の割合については，六自由度のものが最も良い値になっている．

3.4 残基配列推定に関する考察

本研究での推定精度のポテンシャル自由度の違いに関する比較では，自由度が高いほど良い推定精度が得られていることが分かる．同じ程度の空間解像度である一自由度と三自由度 (2) の場合や，三自由度 (1) と六自由度のを比べると，自由度の多いポテンシャルのほうが良い推定結果を与えている．

一自由度での残基配列推定は，基底変換を施しているものの，実質的には Sippl らの提案したポテンシャルと同じものである．本研究から言えることは，残基配列推定問題においては，一自由度のポテンシャルにくらべて，三自由度 (2)，そして六自由度のポテンシャルのほうがより精度の高い推定が行なえるということである．また，同じ数の自由度では空間解像度が高いほど推定精度が高いことが分かる．

4. 縫糸法によるタンパク質構造認識問題

本節では，多自由度の平均力場ポテンシャルを用いた縫糸法を用いて，タンパク質の構造認識問題を解く．

前節の残基配列推定問題を解く上で用いたタンパク質立体構造からそれぞれの場所における残基種のスコアを出す方法を利用し，凍結近似によるプロファイルを作る方法について述べる．続いて，このプロファイルと残基配列との整列 (alignment) 手法について述べ，最後にプロファイルと残基配列との整列で得られるスコアを用いた構造認識手法での認識率についての検証を行なう．そして，最後に，検証で得られた認識率に関する結果について考察を行なう．

4.1 凍結近似によるプロファイル

前節でも述べた通り，残基対双方の残基種を考慮した平均力場ポテンシャルの場合，ある場所での残基種のスコアは構造上近傍にある残基の種類が決定されないとポテンシャルの値が計算できない．よって，与えられた立体構造に対して，アミノ酸残基配列を縫糸することは，理論的に正当な方法は全ての可能な縫糸パターンを生成し，そのパターン一つ一つについて平均力場ポテンシャルの和を計算し，ポテンシャルが最小となるようなパターンを探索する探索問題に帰着する[11]．この問題は，NP ハードな問題であり，厳密解を求めることは計算量からして非常に難しい．そこで，残基配列推定方法で行なったのと同じように，立体構造を与えているタンパク質の配列をそのまま用い (これが凍結近似である)，各場所でのアミノ酸残基種のスコアを計算したプロファイルを用いた方法を考える．このプロファイルと縫糸すべき残基配列との合わせ込みは，動的計画法で最適解を求めることができる．しかし，この方法は正当性が乏しい．

さて，凍結近似を用いて残基配列推定をした結果は 20% を多少越える程度であった．このことは，残基対を形成する双方の残基種を考慮した平均力場ポテンシャルといえども，残基種を決定的にするほどの制約をもっていないことを意味する．すなわち，残基対を形成する片方の残基が明確に決まっていても，あるいはある程度の広がりをもって決定されていても，もう一方の残基種を決定する能力にはそれほど違いはないのである．このことは，凍結近似を使うことをかなり正当化する．凍結近似においては，縫糸する残基配列と，構造を与えているタンパク質の配列が異なっているが，その両者の間の相同性が 20% 程度あるならば，凍結近似を行なっても，厳密な方法で行なってもそれほどの違いはないと判断できるのである．

プロファイルは，それぞれの場所ごとに，それぞれの残基種ごとに S^{α} を計算したものである．ある場所の残基がどの種類である可能性が高いかを示している指標である．

4.2 配列とプロファイルとの整列 (alignment)

凍結近似によって，タンパク質立体構造一つ一つについてプロファイルを作ることができれば，このプロファイルと縫糸すべき残基配列との整列 (alignment) を行なうことで，残基配列と立体構造との互換性を評価することができるようになる．整列は，この場合動的計画法によって高速に行なうことができる．ここでは，Needleman-Wunsch アルゴリズムによる動的計画法を用いた整列を行なう．本論文は，配列の整列に関する詳細は触れない．文献[12][13] を参照されたい．

整列の際，問題となるギャップコストの与え方はアフィンギャップコスト[14] である．ギャップコスト，とくに，配列の両端より外に入る外ギャップコストを高くすると，立体構造と縫糸すべき配列の長さが似ているものの互換性が非常に高くなり，残基配列はほとんどの場合自分自身の構造を最も互換性が高いと判断する．つまり，自己認識率が 100% になる．その他のギャップコストでも，高ければ高いほど自己認識率は高くなり，用いた縫糸法の評価尺度によらず良い評価結果が得られてしまう．本研究においては，ギャップコストは，さまざまな検討の結果，第一ギャップコストを 10，延長ギャップコストを 1，外ギャップコストを 0 とすることにした．

4.3 残基配列の自己構造認識率の検証結果

データセット，及び検証セットのとり方は，残基配列推定におけるものと同じである．ここでの自己構造認識率は，検証セットのタンパク質を含む 911 個のタンパク質の立体構造から作られたプロファイルを用意し，検証セットである 46 個のタンパク質の残基配列が，911 個の

立体構造から自分自身の構造を正しく認識する正解率である．この検証結果から，三自由度 (2) のポテンシャルを用いた場合が最も自己認識率が高く，ついで六自由度，三自由度 (1)，一自由度，そして隣接二面角のみを用いたものが最も悪いことが分かる．

ポテンシャル自由度	自己認識率 (gap cost=10)
隣接残基二面角のみ	63.0%
一自由度	80.4%
三自由度 (1)	80.4%
三自由度 (2)	84.8%
六自由度	80.4%

自己構造認識率は，今回の検証結果により，必ずしもポテンシャルの自由度が多いだけではなく，空間解像度にも依存している可能性が高いことが示唆される．その意味で，十分な空間解像度をもち，かつポテンシャル自由度が多い三自由度のものが最も良い結果を与えるという結果になった．

5. 全般的な考察

タンパク質の立体構造予測問題における縫糸法による立体構造認識問題において，従来ではアミノ酸残基間の相対距離だけによって決まる一自由度の平均力場ポテンシャルを用いていたのに対して，相対位置関係を表すより多数の自由度をもつ平均力場ポテンシャルを用いて，自己構造認識率を検証した．その結果，ポテンシャルの空間解像度が高く，かつ自由度が 3 である平均力場ポテンシャルを用いた場合が最も良い自己認識率を与えることが分かった．

また，ポテンシャルの質を判断するために残基配列推定を行ない，その残基配列推定精度を求めた結果，自由度の最多である六自由度のポテンシャルを用いたときが最も推定精度が高く，またその際の正解残基種のスコアなども六自由度のものが最も良いことが分かった．

この二つの検証結果は，ともに従来法の一自由度のポテンシャルを用いた方法よりも三自由度，六自由度という自由度の多いポテンシャルを用いた方法が優れていることを立証している．しかしながら，三自由度の場合と六自由度の場合において，解くべき問題によって性能が矛盾していることが分かった．三自由度のポテンシャルと六自由度のそれのどちらが本当の意味で優れているのかの決着はまだ現段階では付けられない．

6. 今後の研究

これまで，多自由度のポテンシャル解析をもっぱら残基推定や構造認識に応用してきたが，今後は，タンパク質の折り畳みシミュレーションにも応用していきたい．

今回の研究によって，三自由度の統計ポテンシャルが良い成績を収めた．この三自由度のポテンシャルは空間解像度が非常に高いという性質を持っている．このことは，タンパク質の折り畳みシミュレーションを行なって，立体構造を縫糸法にたよらず，最初から (*ab initio*) 組み上げる構造予測法において，かなり精密なポテンシャルとして利用できることを意味する．

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. (editor) 1989. *Prediction of Protein Structure and the Principles of Protein Conformation.* New York: Plenum Publishing Corporation.

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

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

5) 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.

6) 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.

7) Sippl, M.J. and S. Weitckus 1992. "Detection of Native-like Models for Amino Acid Sequences," *PROTEINS: Structure, Function, and Genetics 13:* 258-271.

8) Melo, F.; and E. Feytmans 1997. "Novel Knowledge-based Mean Force Potential at Atomic Level," *J. Mol. Biol 267*:207-222

9) 鬼塚健太郎, 野口保, 安藤誠, 秋山泰 1998.「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」情報処理学会「数理モデルと問題解決」研究会論文誌投稿中

10) 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.

11) Geman, S.; and D. Geman 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian," *IEEE PAMI 6*: 721-741.

12) Needleman, S.B.; and C.D. Wunsch 1970 "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol. 48*:443-453.

13) 後藤修 1983.「核酸・蛋白質一次構造の計算機による解析」日本物理学会誌 88 No. 6:477-480.

14) 石川幹人, 十時泰, 戸谷智之, 星田昌紀, 広澤誠 1994.「並列反復改善法によるタンパク質の配列解析」情報処理学会論文誌 35 No.12:2816-2839.