統計ポテンシャルを用いたタンパク質の構造予測

川端　猛

統計ポテンシャルを用いたタンパク質の構造予測

川端 猛

目次

第 1 章

序論

タンパク質の構造予測は構造生物学における依然未解決の難題である。その難しさの原因の一つは、タンパク質の構成要素の複雑さであり、物理シミュレーションでアプローチする際の大きな困難となる。このような複雑な対象を扱う際のもう一つのアプローチとして、統計的アプローチ、すなわち、既知のデータの大域的な特徴が再現されるように単純なモデルを構築する方法がある。近年、データベースに登録されたタンパク質の結晶構造数は増加しており、タンパク質の構造予測においてもこのアプローチが有効となる可能性が高い。本研究では統計的アプローチの一種である「統計ポテンシャル (statistical potential)」と呼ばれる方法を試みる。このポテンシャルは構造の特徴を $\boldsymbol{C}$、配列の特徴を $\boldsymbol{A}$ としたとき、

$$E(\boldsymbol{C}=\boldsymbol{c},\boldsymbol{A}=\boldsymbol{a})=\sum_i E(C_i=c_i,A_i=a_i)=\sum_i -\log\frac{P(C_i=c_i/A_i=a_i)}{P(C_i=c_i)} \qquad (1.1)$$

の式で記述される量のことを指す。C_i , A_i は特徴 $\boldsymbol{C}$、 $\boldsymbol{A}$ の個々の成分であり、お互いに対応するように設定する。予測に用いる場合は、与えられた配列 $\boldsymbol{a}$ に対して可能な全構造群 $\boldsymbol{C}$ の中で最も $E(\boldsymbol{C}=\boldsymbol{c},\boldsymbol{A}=\boldsymbol{a})$ の低い $\boldsymbol{c}$ を予測構造とする。実際にどのような特徴を用いるかは、予測の目的によって異なるが、C_i として各残基の 2 次構造や残基間の距離、A_i としてアミノ酸の 1 残基対や 2 残基対が用いられる場合が多い。統計ポテンシャルは、近年、特に Threading 問題に適用され、大きな成果を挙げてきた[1,2]。また、古典的な 2 次構造予測法の一つである GOR 法[3] は 、この統計ポテンシャルの 2 次構造予測への適用であると考えることができる。

本研究では、このポテンシャルを 2 次構造予測と 3 次構造予測に適用し、その有効性・問題点を明らかにすることを目的とする。序論においては、統計ポテンシャルの理論的根拠とその有用性と問題点を説明し、本研究の概要を説明する。

1.1 統計ポテンシャルの理論的根拠

式1.1の理論的根拠には、大きく分けて (1) 統計力学的な導出と (2) 確率モデルの識別関数としての導出の2つが考えられる。以下、簡単のために、$f(\boldsymbol{C}=\boldsymbol{c}) \rightarrow f(\boldsymbol{c})$, $f(\boldsymbol{C}=\boldsymbol{c}/\boldsymbol{A}=\boldsymbol{a}) \rightarrow f(\boldsymbol{c}/\boldsymbol{a})$, $E(\boldsymbol{C}=\boldsymbol{c},\boldsymbol{A}=\boldsymbol{a}) \rightarrow E(\boldsymbol{c},\boldsymbol{a})$ と書くことにする。

1.1.1 統計力学的な導出

Sippl らは統計ポテンシャルを用いた研究を精力的に行なっており、彼らは式 1.1を以下のような統計力学のボルツマン分布則から導出している[4]。

ある配列 $\boldsymbol{a}$ のタンパク質の系において、構造 $\boldsymbol{c}$ が $E(\boldsymbol{c},\boldsymbol{a})$ によるボルツマン分布則に従うとすると、分布 $f(\boldsymbol{c}/\boldsymbol{a})$ は以下の式で記述される。

$$f(\boldsymbol{c}/\boldsymbol{a}) = \frac{1}{Z}\exp\left[-\frac{E(\boldsymbol{c},\boldsymbol{a})}{kT}\right] \tag{1.2}$$

式中の Z は以下で定義される分配関数である。

$$Z = \int\cdots\int \exp\left[-\frac{E(\boldsymbol{c},\boldsymbol{a})}{kT}\right] d\boldsymbol{c} \tag{1.3}$$

式1.2を変形すると、分布 $f(\boldsymbol{c}/\boldsymbol{a})$ から、$E(\boldsymbol{c},\boldsymbol{a})$ を推定することができる。

$$E(\boldsymbol{c},\boldsymbol{a}) = -kT\log[f(\boldsymbol{c}/\boldsymbol{a})] - kT\log Z \tag{1.4}$$

ここで、参照状態として、「平均的な」アミノ酸配列のタンパク質の系を考え、その構造 $\boldsymbol{c}$ が ポテンシャル $E(\boldsymbol{c})$ によるボルツマン分布に従うとすると、同様にそのポテンシャル $E(\boldsymbol{c})$ は、以下の式で記述される。

$$E(\boldsymbol{c}) = -kT\log[f(\boldsymbol{c})] - kT\log Z' \tag{1.5}$$

$$Z' = \int\cdots\int \exp\left[-\frac{E(\boldsymbol{c})}{kT}\right] d\boldsymbol{c} \tag{1.6}$$

$E(\boldsymbol{c})$ を基準としたポテンシャル $\Delta E(\boldsymbol{c},\boldsymbol{a})$ は以下のようになる。

$$\Delta E(\boldsymbol{c},\boldsymbol{a}) = E(\boldsymbol{c},\boldsymbol{a}) - E(\boldsymbol{c}) = -kT\log\frac{f(\boldsymbol{c}/\boldsymbol{a})}{f(\boldsymbol{c})} - kT\log\frac{Z}{Z'} \tag{1.7}$$

ここで、2つの系の分配関数は等しいという近似、

$$Z \simeq Z' \tag{1.8}$$

を適用すると、式 1.7は

$$\Delta E(\boldsymbol{c},\boldsymbol{a}) \simeq -kT\log\frac{f(\boldsymbol{c}/\boldsymbol{a})}{f(\boldsymbol{c})} \tag{1.9}$$

となる。さらに以下の分布の独立性の近似を行なう。

$$f(\boldsymbol{c},\boldsymbol{a}) \simeq \prod_i f(c_i/a_i), \quad f(\boldsymbol{c}) \simeq \prod_i f(c_i) \tag{1.10}$$

すると、$\Delta E(\boldsymbol{c},\boldsymbol{a})$ は、

$$\Delta E(\boldsymbol{c},\boldsymbol{a}) \simeq \sum_i -kT\log\frac{f(c_i/a_i)}{f(c_i)} \tag{1.11}$$

となり、式1.1の統計ポテンシャルと等価な形が導出される。ここで $f(c_i/a_i)$, $f(c_i)$ は結晶構造データベースの統計から求める。

1.1.2 確率モデルの識別関数としての導出

統計ポテンシャルは、問題に対する確率モデルを用いたベイズ決定として導出することができる。ここでは、構造予測における2つの問題、Threading と2次構造予測について説明する。特に Threading 問題については、統計ポテンシャルは極めて自然な形で導出される。2次構造予測の問題については、相関の程度が異なる事象を独立近似する場合の補正として統計ポテンシャルが導出される。また、第I部で用いる GOR 法のポテンシャルとの関係についても言及する。

Threading の場合

Threading とは、結晶構造データベースの部分構造に配列をのせてポテンシャルを評価し、ポテンシャルの最も低い構造を予測構造とする予測方法である。この Threading 用の最適な識別関数の構成という形で、式 1.1を自然な形で導入することができる。

配列 $\boldsymbol{a}$ に対する N 構造をクラス $N(\boldsymbol{a})$、N 構造以外のデコイ構造をクラス D とする。2つのクラスから確率的に構造が出力されるモデル（図1.1）を考えると Threading とは「データベース内の各構造 $\boldsymbol{c}(\in \boldsymbol{C})$ が、クラス $N(\boldsymbol{a})$ に属するか、クラス D に属するか識別する」ことであると考えられる。ここで、クラス $N(\boldsymbol{a})$ が構造 $\boldsymbol{c}$ を出力する確率 $P(\boldsymbol{c}/N(\boldsymbol{a}))$, クラス D が構造 $\boldsymbol{c}$ を出力する確率 $P(\boldsymbol{c}/D)$ を導入する。このとき、

$$P(\boldsymbol{c}/N(\boldsymbol{a})) > P(\boldsymbol{c}/D) \tag{1.12}$$

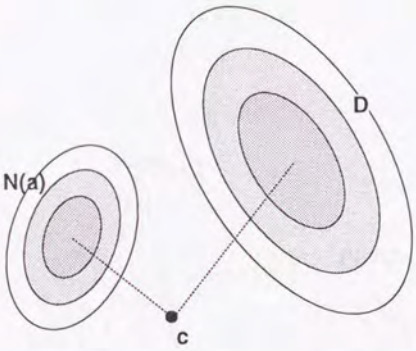


図 1.1　Threading 問題に対する確率モデルの概念図。構造空間を 2 次元として、N 構造の確率分布 $P(c/N(a))$ とデコイ (D) 構造の確率分布 $P(c/D)$ を楕円で表している。構造 c が、N 構造と D 構造のどちらから生じた確率が高いか判定することが Threading 問題であると考えることができる。

が成り立つ $\boldsymbol{c}$ を配列 $\boldsymbol{a}$ の N 構造とすることがベイズ決定となる。図 1.1にこの確率モデルの概念図を示した。式 1.12 を対数の形で表すと以下のようになる。

$$E(\boldsymbol{c},\boldsymbol{a}) = -\log\frac{P(\boldsymbol{c}/N(\boldsymbol{a}))}{P(\boldsymbol{c}/D)} < 0 \tag{1.13}$$

ここで、確率 $P(\boldsymbol{c}/N(\boldsymbol{a}))$,$P(\boldsymbol{c}/D)$ を式 1.10と似た以下の独立の近似をして推定する。

$$P(\boldsymbol{c}/N(\boldsymbol{a})) \simeq \prod_i f(c_i/a_i), \quad P(\boldsymbol{c}/D) \simeq \prod_i f(c_i) \tag{1.14}$$

すると式 1.13は

$$E(\boldsymbol{c},\boldsymbol{a}) = \sum_i -\log\frac{f(c_i/a_i)}{f(c_i)} < 0 \tag{1.15}$$

となり、式 1.1の統計ポテンシャルと同形になる。この導出を前節の統計力学的な導出と比べると、独立の仮定の問題は同じだが、参照状態の意味が明確に説明される。ただし、式 1.13は、$E(\boldsymbol{c},\boldsymbol{a})$ が負の場合、構造 $\boldsymbol{c}$ は N 構造と判定されることを示しているだけで、$E(\boldsymbol{c},\boldsymbol{a})$ が最小となる構造が N 構造と一致することは保証しない。図 1.2にこの関係を模式図で示した。

2 次構造予測の場合

2 次構造予測の場合、構造特徴 $\boldsymbol{c}$ は 2 次構造（ $c \in \{\alpha helix, \beta sheet, coil\}$ ）の列となる。多くの 2 次構造予測では 、 $\boldsymbol{c}$ は独立であるという仮定をして、ポテンシャル $E(c_i = c, \boldsymbol{A} =$

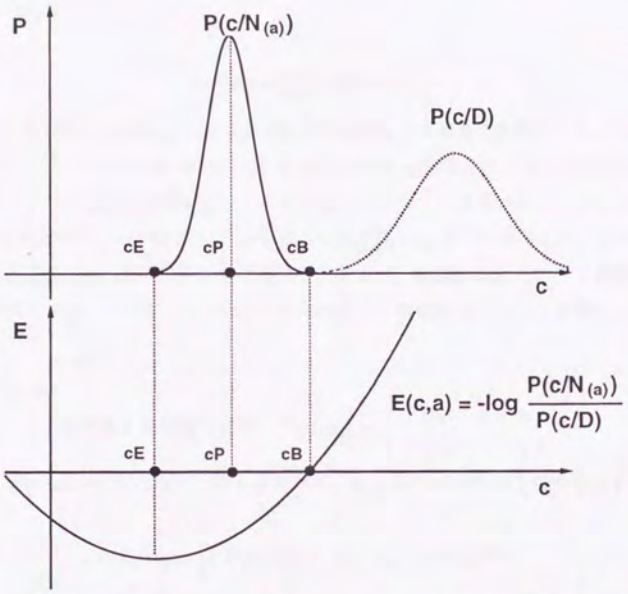


図 1.2　Threading 問題に対する確率モデルと統計ポテンシャルの概念図。上が N 構造の確率分布 $P(c/N(a))$ とデコイ構造の確率分布 $P(c/D)$ 、下がこれらの確率分布から 作成した統計ポテンシャル $E(c,a)$ を示す。横軸は構造空間を 1 次元で表しており $P(c/N(a))$ と $P(c/D)$ は 1 次元の正規分布であるとした。構造 cP は、確率分布 $P(c/N(a))$ が最も高い構造、すなわち配列 a の N 構造を示す。構造 cB は確率分布 $P(c/N(a))$ と $P(c/D)$ が等しいボーダーの構造でこの構造より左にある構造はベイズ決定においては、すべて $N(a)$ 構造であると判定される。構造 cE は統計ポテンシャル $E(c,a)$ が最も低い構造である。構造 cP と構造 cE はほとんどの場合一致しない。このことは統計ポテンシャルのポテンシャル最小構造が必ずしも N 構造と一致しないことを示唆する。

$\boldsymbol{a}$) が最も低い c を順に求めることで、全体の予測構造 $\boldsymbol{c}$ を得る。2 次構造予測をベイズ推定の確率モデルで行なうには、与えられた配列 $\boldsymbol{a}$ に対して、$P(c_j/\boldsymbol{a})$ を想定し、$P(c_j/\boldsymbol{a})$ が最大の c を予測構造とすればよい。ここで、a に対する独立の近似

$$P(c_j/\boldsymbol{a}) \simeq \prod_i P(c_j/a_i) \tag{1.16}$$

を行なうと、以下のポテンシャルを最小にする構造がベイズ推定になる。統計ポテンシャルの式とは一致しない。

$$E(c_j, \boldsymbol{a}) = \sum_i -\log P(c_j/a_i) \tag{1.17}$$

ここで、j 番目の 2 次構造 c_j に対する i 番目の残基 a_i の影響量を考えてみる。もし Zimm-Bragg のヘリックス - コイル転移[5] のような 1 次元の Ising 模型に従って 2 次構造が形成されるなら、$i = j$ のとき最も影響が強く、j から離れるにつれて影響は小さくなるであろう。図 1.3 (a) に現実のタンパク質の 統計から得られた $P(c_j/a_i)$ をプロットした。i が j から離れるにつれ、$P(c_j/a_i)$ は $P(c_j)$ の値に近付く。ここで、極端な例として、2 次構造が他の残基と無相関な場合、つまり、c_j は a_j のみに依存し、他の残基には全く依存しない場合を考える。

$$\text{他残基と無相関な場合：} P(c_j/a_i) = \begin{cases} P(c_j/a_j) & i = j \\ P(c_j) & i \neq j \end{cases} \tag{1.18}$$

この場合、$P(c_j/\boldsymbol{a}) = P(c_j/a_j)$ となるはずだが、もし式 1.16の独立近似を導入すると、

$$P(c_j/\boldsymbol{a}) \simeq \prod_i P(c_j/a_i) = P(c_j/a_j) \times P(c_j)^{|\boldsymbol{a}|-1} \tag{1.19}$$

となり、正しい確率に $P(c_j)^{|\boldsymbol{a}|-1}$ を乗じた数になってしまう。これを補正した近似は以下のようになる。

$$P(c_j/\boldsymbol{a}) \simeq \frac{1}{P(c_j)^K} \prod_i P(c_j/a_i) \tag{1.20}$$

ここで、K は任意の定数であり式 1.18の場合は、$K = |\boldsymbol{a}| - 1$ とすると、正しいベイズ決定となる。以上の例からわかるように、考慮する $\boldsymbol{a}$ の中で無相関な要素が含まれる可能性がある場合、あるいは相関の程度に差がある場合、式 1.20のほうがベイズ推定に近い可能性があると考えられる。しかし、実際に、K の値をどのように設定するかは難しい問題である。ここで、$K = |\boldsymbol{a}|$ とすると、式 1.16は、

$$E(c_j, \boldsymbol{a}) = \sum_i -\log \frac{P(c_j/a_i)}{P(c_j)} \tag{1.21}$$

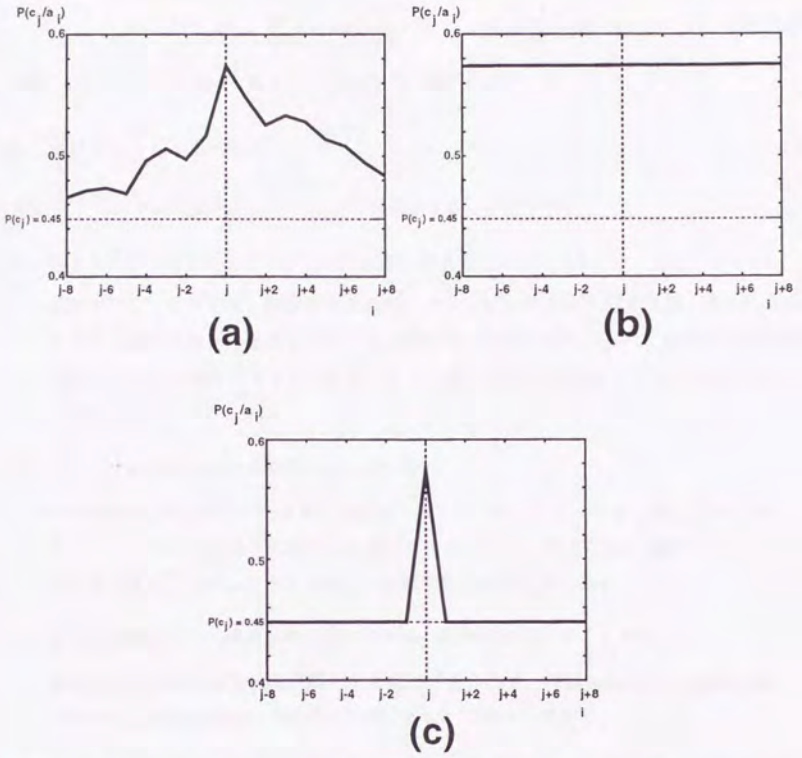


図 1.3　i 残基目のアミノ酸 が a_i であったときの j 番目の 2 次構造 c_j の生じる確率 $P(c_j/a_i)$ の分布。(a) 結晶構造データベースの 統計から得られた分布。$P(c_j = \alpha - helix, /a_i = Ala)$ を示した。i が j から離れるにつれ、$P(c_j/a_i)$ は $P(c_j)$ の値に近付く。(b) 独立近似が成り立つ場合の仮想的な分布。(c) 他残基と無相関であると仮定した場合の仮想的な分布。(a) の実際の統計分布は、(b) と (c) の中間の形状をしている。

となり、式 1.1の統計ポテンシャルと同形になる[1]。

ここで問題を、与えられた $\boldsymbol{a}$ に対して、ある構造クラス c か、そうでないかを判定することであるとすると、ベイズ判定は $P(c/\boldsymbol{a}) > P(\bar{c}/\boldsymbol{a})$ となる。ここで $K = |\boldsymbol{a}|$ とした式 1.20の近似を用いると、識別関数は以下のようになる。

[1]実際の 2 次構造形成は他残基とある程度相関を持つので、$K = |\boldsymbol{a}|$ とするのは大きすぎるかもしれない。本研究の第 I 部で導入される修正 GOR 法においては、大き過ぎる K を補正するために構造ごとのオフセット値を加えるという工夫が行なわれる。

$$E(c_j, \boldsymbol{a}) = \sum_i \left(-\log \frac{P(c_j/a_i)}{P(c_j)} + \log \frac{P(\bar{c_j}/a_i)}{P(\bar{c_j})} \right) < 0 \tag{1.22}$$

この式は GOR 法[3] で用いられている情報量と同形である。

1.2 統計ポテンシャルの長所

統計ポテンシャルの長所として、以下の点を挙げることができる。

1. 様々な相互作用を 1 つのポテンシャルで代表することができる

   統計ポテンシャルには、静電気相互作用、ファンデルワールス相互作用、疎水性相互作用など様々な相互作用がすべて平均的にバランス良く取り込まれている。厳密に物理的にこれらの効果を取り込む場合に比べ、遥かに少ない計算コストで見積もることができる。

2. パラメータ抽出にかかる計算コストが少ない

   他の統計的なアプローチの方法、例えばニューラルネットワークなどの反復的な方法、に比べ、パラメータ抽出にかかる計算コストが少ない。そのため、統計ポテンシャルによる予測は、ジャックナイフ法による厳密な評価が可能である。

3. 特徴の選択により、構造予測の様々な場合に汎用的に用いることができる

   原理的にはどのような特徴に関しても適用できるので、Threading、2 次構造予測、内外予測、配列設計など、様々な予測に関して適用可能である。

1.3 統計ポテンシャルの理論的な問題点

一方、統計ポテンシャルの理論的な問題点も多く存在する。特に統計ポテンシャルの統計力学的な導出は、多くの大胆な近似がなされているため、厳密な自由エネルギーを表しているとは考えにくいという批判がある[6]。特に問題となるのは以下の点であろう。

1. データベース内の分布の同一性の仮定： 統計ポテンシャルは、$f(c)$, $f(c/a)$ などの結晶データベースの統計量が同じ分布則に従うと仮定している。しかし、結晶構造データベースは異なるタンパク質、温度、条件の結晶構造を含んでおり、完全に同一の分布に従っているわけではない。この問題は確率モデルによる導出でも同様に存在する。

2. 独立性の仮定： 例えば、残基間のコンタクトを構造特徴とした場合、式 1.10の独立性の近似が完全に成り立つ系は、(1) アミノ酸が解離した状態で存在し、(2) 希薄な気体

状態にある場合である。実際のタンパク質は全て重合し、密にパッキングしている状態にあるため、この理想的な系とはほど遠い。確率モデルによる場合でも同様な近似（式1.14, 1.16）を行なっており、本質的に同様な問題は存在する。

3. 参照状態の導入: 式1.7のように、平均的なアミノ酸配列のタンパク質の系を参照状態としているが、その物理的な必然性は低い。式1.8の近似は、配列 $\boldsymbol{a}$ が平均的な組成から遠くなればなるほど成り立たなくなる。一方、Threding 問題についての確率モデルについては、この参照状態に明確な説明を与えることができる。が一方、ポテンシャル最小構造がN構造と一致する必然性が失われる。2次構造予測についての確率モデルでは、独立近似の補正の一種として参照状態を導入することは可能である。

これらの理論的問題点はあるものの、統計ポテンシャルの簡潔性はこれらの問題点を上回る長所であると思われる。これらの問題点を承知の上で統計ポテンシャルを用い、その有用性を過信しないことが必要であろう。

1.4 統計ポテンシャルの適用についての問題点

以下に実際の問題に適用する場合の統計ポテンシャルの問題点を挙げる。

- 構造特徴・配列特徴の選択： 統計ポテンシャルは、いかなる特徴を用いても構成することが可能であるが、高い予測精度を得るには、適当な特徴を選択する必要がある。先験的に妥当な特徴を選ぶ方法はなく、物理化学的に考えて妥当な特徴を試行錯誤的に試みるしかない。本論文の第I部は、この試行錯誤を2元符号化関数の探索という形でシステマティックに行なう試みである。第II部においても、Threading に適当なポテンシャルの選択にこの試行錯誤が行なわれる。

- 複雑すぎる特徴を用いると予測精度は下がる： 統計ポテンシャルは、第1原理的なポテンシャルと異なり、あくまで統計的推定として値を用いるので、複雑な特徴を取り上げるとポテンシャルの値の信頼性が低下し、構造予測の精度はかえって低下するという現象が起こる。データベースのサイズに見合った複雑さの特徴を選ぶ必要がある。本論文の第I部の2元語符号の導入は、複雑さをあまり上げずに多残基の相関を取り込む工夫であるといえる。

1.5 本研究の内容

- **第I部 2元符号を用いたタンパク質の2次構造予測：** 第I部では、統計ポテンシャルを2次構造予測に適用する。統計ポテンシャルを用いた2次構造予測法 として GOR 法と

いう古典的な方法がある。本研究では多残基の相関を限られた統計データから効率的に抽出するためにアミノ酸を2種類に分類して考える2元語符号 (binary word encoding) を導入し、予測精度の改善を目指した。

- **第 II 部 統計ポテンシャルを用いたタンパク質の3次構造予測：**第 II 部では、統計ポテンシャルをポテンシャル最小化による3次構造予測に適用する。統計ポテンシャルはその定式化から考えて、異なるタンパク質の結晶構造から N 構造を識別するように設計されており、「タンパク質らしく」ない構造から識別できる保証はない。しかし、ポテンシャル最小化計算においては、そういった構造も予測の対象となる構造空間に含まれてしまう。そこで、本研究では「タンパク質らしい」構造の条件をヒューリスティックに定め、この条件を満たす構造を探索する方法を開発した。その上で、統計ポテンシャルがどのような構造群から N 構造を識別できるかどうかを考察する。

第 I 部

2元語符号を用いたタンパク質の2次構造予測

第 2 章

はじめに

今までに、多くの研究者が、様々なタンパク質の２次構造予測法を提唱してきたが、それらの方法は十分成功しているとは言い難い。その理由として、よく挙げられるのは次の２つである。

1. 配列上遠い残基の影響（遠距離相互作用）が取り込まれていない。

   ２次構造予測法の多くは、局所配列、すなわち予測する残基の前後 10 残基程度の配列を入力とする。しかし、現実のタンパク質は配列上遠くても、立体構造上近接した残基の影響を受ける。特に β シートは遠距離相互作用の影響が大きいと考えられる。

2. 局所配列の情報も統計データ不足のため完全には取り込まれていない。

   多くの２次構造予測では、局所配列の、多数の残基の相関した情報は統計データ不足のため、取り込めていない。

本論文の第 I 部では、この２つの困難のうち後者の「統計データ不足」を解決しようとする試みである。

予測法の複雑さの指標として、Minsky と Papert の定義した予測法の次数について説明する[7]。まず、本研究では「k 残基ポテンシャル（k-residue potential）」を「局所配列の中の k 残基で値が決定されるポテンシャル」と定義する。このとき k をポテンシャルの「台 (support)」と呼ぶ。そして、ある予測法と同じ入出力関係を実現するパーセプトロンをできるだけ小さな台のポテンシャルの和で構築したとする。そのとき、和となるポテンシャルの中で最も大きな台をその予測法の「次数 (order)」と定義する。「次数」が高い予測法ほど局所配列と２次構造の関係をより詳細に記述できると考えられる。

Chou-Fasman 法[8,9] や GOR 法[3] などの初期に開発された予測法は１残基ポテンシャルをベースとしている。つまりこれらの古典的な予測法の次数は１以下である。１残基ポテンシャルは簡単であるが、各残基の２次構造傾向、２次構造の協同性、N-cap/C-cap の傾向など予

測に必要な重要な物理的な特徴を表現することができる。1残基ポテンシャルは2次構造予測に必要不可欠である。

しかし、より高い予測精度の予測法を目指すには、1残基ポテンシャルだけでは不十分である。本研究では「1残基を越える残基によって値が決定されるポテンシャル」のことを「多残基ポテンシャル (multi-residue potential)」と呼ぶこととする。多残基ポテンシャルは、周期性や複数残基の相互作用など残基の組み合わせの効果を記述する場合に必要となる。多残基ポテンシャルを取り込んだ次数の予測法が構築できれば、予測精度も向上すると期待できる。しかし、多残基ポテンシャルの統計的な抽出は、より大きなサイズのデータベースが必要となるという問題がある。例えば、4残基ポテンシャルの場合を考えてみよう。全ての可能な4残基パターンの数は $20^4 = 1,600,000$ あり、現在の結晶構造データベースのサイズ（100,000 残基以下）を遥かに越えてしまう。これでは、1パターンあたりの平均観測数は1を下回ってしまい、ポテンシャルの値を推定することはできない。

多くの研究者が、多残基ポテンシャルを取り込むための努力をしてきたが、完全に成功しているとは言い難い。Gibrat らは、2残基ポテンシャルである「ペア情報 (pair information)」を導入した[10]。しかし、彼らは、データベースのサイズの不足のため、全ての2残基ペアの値を抽出することはできず、統計的に有意なペアの値のみ使用している。長野は3残基ポテンシャルである「3残基項 (triplet term)」を導入した[11]が、データ不足を克服するため、20 種のアミノ酸を7種に分類して用いている。3層ニューラルネットワーク[12,13]や局所ホモロジー法[14,15]は、多残基ポテンシャルを取り込んだ次数の高い予測を構築することが原理的には可能である。しかし、これらの方法で、多残基ポテンシャルが本当に予測に影響を与えているかどうか、はっきりしたことはわからない。3層ニューラルネットワークを適用した2つのグループ[12,13]は、隠れユニットのないニューラルネットワーク（次数1のパーセプトロンと等価）と、最適な数の隠れユニットを持ったニューラルネットワークで、予測精度に大きな差がないことを報告している。このことは、これらのネットワークでは実質的には多残基ポテンシャルは抽出されておらず、予測法の次数は1を越えていないことを示唆する。

予測精度を改善する全く別のアプローチとして、マルチプルアライメントされた配列を入力として用いるという方法がある。Rost と Sander は、マルチプル配列アライメントを入力として用いた方法で、正答率 70% 以上を達成したと報告している[16,17]。このことは進化的な情報が2次構造予測の改善に極めて有用であることを示す。また Rost と Sander は同時に多数の独立したネットワークの集合である"jury decision network" を用いると、単体のネットワークに比べて正答率が 2-3% 向上すると報告している。これは、彼らの"jury decision network" が多残基ポテンシャルを抽出していることを示唆する。

本論文では、多残基ポテンシャルを抽出するための新しい手法として「2元語符号 (binary word encoding)」を提唱する。これは 20 種のアミノ酸からなる局所配列を 0,1 の文字列（2

元語）に変換して扱うことである。2元語の可能なパターンの種類は、20元語に比べ圧倒的に少ないため、2元語符号を用いることで効率的に多残基ポテンシャルが抽出できる。この2元語情報をGOR法と結び付けた方法を開発した。また、マルチプル配列アライメントを入力として用いた予測も行なった。

第 3 章

方法

3.1 2元語符号 (binary word encoding)

前章で述べたように、2次構造予測を改善する有効な方法の一つとして、多残基の相関を考慮した多残基ポテンシャルの導入が挙げられる。多残基ポテンシャルを考慮する場合の問題点は統計データの不足である。本論文では少ないデータ量から有効に多残基の相関を抽出するためにアミノ酸を2種に分類して捉える2元語符号 (binary word encoding) を導入する[18]。

まず、20種のアミノ酸を0か1かに変換する符号化関数 (encoding function) f を導入する。

$$f(a) = b \quad \text{for} \quad a \in \mathcal{A}, \quad b \in \mathcal{B} \tag{3.1}$$

ここで $\mathcal{A}$ は20種のアミノ酸の集合、$\mathcal{B}$ は0と1の集合である。

$$\mathcal{A} = \{A, I, L, M, F, P, V, R, D, E, K, N, C, Q, H, S, T, W, Y, G\} \tag{3.2}$$

$$\mathcal{B} = \{0, 1\} \tag{3.3}$$

符号化関数 f を用いて、アミノ酸の局所配列 $\boldsymbol{x}$ を0と1からなる2元語 (binary word) $\boldsymbol{b}(\boldsymbol{x})$ に変換することができる。

$$\boldsymbol{b}(\boldsymbol{x}) = (f(x_{-N}), \ldots, f(x_0), \ldots, f(x_N)) \in \mathcal{B}^{2N+1} \tag{3.4}$$

$$\boldsymbol{x} = (x_{-N}, \ldots, x_0, \ldots, x_N) \in \mathcal{A}^{2N+1} \tag{3.5}$$

文字列 $\boldsymbol{x}$ と $\boldsymbol{b}(\boldsymbol{x})$ の長さは $2N+1$ であるとする。例えば極性残基のとき0、非極性残基のとき1を返すような符号化関数を用いた場合、長さ7のアミノ酸文字列 $\boldsymbol{x} = (LVEADGF)$ は

2元語 $\boldsymbol{b}(\boldsymbol{x}) = (1101001)$ に変換される。実際にどのような符号化関数を使うべきかは後に議論する。

2元語パターンの総数はある程度長さがあってもさほど大きくならない。例えば、長さ7の20元語のパターンの総数は $1.28 \times 10^9 (= 20^7)$ であるが、同じ長さの2元語パターンの総数はわずか $128 (= 2^7)$ であり、圧倒的に少ない。この程度の大きさであれば、現在利用可能なデータベースのサイズに対して十分小さいため、2元語自体を一つの確率事象と考えた統計量の抽出が可能となると考えられる。

2元語符号は20種のアミノ酸を2種に縮退させているため、多残基ポテンシャルが抽出できる代わりにある種の情報は失われる。そこで、1残基ポテンシャルを基にした古典的な予測法であるGOR法[3]と2元語符号を結合した方法を開発した。

3.2 GOR法との結合

3.2.1 GOR法の基礎

GOR法[3]は1970年代に開発された古典的な2次構造予測法である。同時期に開発されたChou-Fasman法[8,9]やLimの方法[19]に比べ、アルゴリズムが明確で客観性が高いことが特徴である。GOR法は典型的な1残基ポテンシャルを用いた予測法であるが、本研究では多残基ポテンシャルの1種である2元語ポテンシャルを含んだ形に拡張する。

GOR法では、アミノ酸局所配列 $\boldsymbol{x}$ を引数とする識別関数 $G_s(\boldsymbol{x})$ を各2次構造に対して設定し、識別関数の値が最も低い2次構造を予測2次構造 $\hat{s}$ とする。この関係を式で表記すると以下のようになる。

$$\hat{s} = \underset{s \in \mathcal{C}}{argmin}\, G_s(\boldsymbol{x}) \tag{3.6}$$

$\mathcal{C}$ は2次構造クラスの集合であり、本研究では α ヘリックス (α)、β シート (β)、コイル (c) の3種であるとする。

$$\mathcal{C} = \{\alpha, \beta, c\} \tag{3.7}$$

GOR法の識別関数は、統計ポテンシャルの1種で以下の式3.8で表されるポテンシャルの和で構成される[20]。

$$E(X = x : \bar{x}; Y = y) = -\left[\log_2 \frac{P(X = x/Y = y)}{P(X = x)} - \log_2 \frac{P(X = \bar{x}/Y = y)}{P(X = \bar{x})}\right] \tag{3.8}$$

これは統計ポテンシャルをやや修正した形になっており、本研究ではこの形の統計ポテンシャルのことを「GORポテンシャル」と呼ぶこととする。ここで、$P(X = x)$ は事象 $X = x$ が

起こる確率であり、$P(X=x/Y=y)$ は事象 $Y=y$ を知った上で事象 $X=x$ が起こる確率である。$X=\bar{x}$ は $X=x$ の補集合である。このポテンシャルは、事象 $Y=y$ が事象 $X=x$ の生じやすさに与える影響の程度を表している[1]。また第1章で述べたように、GOR ポテンシャルを、ある種の確率モデルにおけるベイズ推定の判別関数と考えることもできる。

2次構造予測においては、式3.8の $X=x$ は、局所配列の中心残基の2次構造 S_0 が $s(s \in \mathcal{C})$ であるという事象 $S_0=s$ に相当する。GOR 法は1残基ポテンシャル $E(S_0=s;\bar{s};R_m=a)$ の和で識別関数が構成される。1残基ポテンシャルにおいては式3.8の $Y=y$ は、局所配列の m 番目の残基が $a(a \in \mathcal{A})$ であるという事象、$R_m=a$ に相当する。1残基ポテンシャル $E(S_0=s;\bar{s};R_m=a)$ は以下の式で定義される。

$$E(S_0=s;\bar{s};R_m=a)=-\left[\log_2\frac{P(S_0=s/R_m=a)}{P(S_0=s)}-\log_2\frac{P(S_0=\bar{s}/R_m=a)}{P(S_0=\bar{s})}\right] \quad (3.9)$$

式中の $P(S_0=s), P(S_0=s/R_m=a)$ 等の確率は統計から推定する。図 3.1 の (a) に1残基ポテンシャルを図示した。GOR 法の識別関数 $G_s^{GOR}(\boldsymbol{x})$ は以下のように1残基ポテンシャルの和で記述される。

$$G_s^{GOR}(\boldsymbol{x})=\sum_{m=-M}^{M}E(S_0=s:\bar{s};R_m=x_m) \quad (3.10)$$

ここで、x_m は局所配列 $\boldsymbol{x}$ の m 番目の残基である。ウィンドウサイズとして原論文では17($M=8$) を用いており、本論文でもそれに従う。GOR 法は単純パーセプトロンに相当しその次数は1を越えない。

3.2.2 2元語ポテンシャルの導入

2元語ポテンシャル (binary word potential) を、2元語を確率事象とした GOR ポテンシャルして導入する。2元語ポテンシャルでは、式3.8の事象 $Y=y$ は、局所配列の2元語 $\boldsymbol{W}$ が $\boldsymbol{w}(\boldsymbol{w} \in \mathcal{B}^{2N+1})$ であること、$\boldsymbol{W}=\boldsymbol{w}$ に相当する。2元語ポテンシャルは以下の式で与えられる。

$$E(S_0=s;\bar{s};\boldsymbol{W}=\boldsymbol{w})=-\left[\log_2\frac{P(S_0=s/\boldsymbol{W}=\boldsymbol{w})}{P(S_0=s)}-\log_2\frac{P(S_0=\bar{s}/\boldsymbol{W}=\boldsymbol{w})}{P(S_0=\bar{s})}\right] \quad (3.11)$$

この2元語ポテンシャルは $2M+1$ 残基に依存して値が決定されるので、$2M+1$ 残基ポテンシャルであるといえる。図 3.1 の (b) に2元語ポテンシャルを図示した。

[1] GOR 法の原論文[20] では、式3.8に -1 を乗じた情報量 $I(X=x:\bar{x};Y=y)=-E(X=x:\bar{x};Y=y)$ を用い、識別関数の値が最大の構造を予測構造とするという形式で記述されている。本論文では、第II部と統一した表記をするために、あえて $E(X=x:\bar{x};Y=y)$ を用いて記述する。

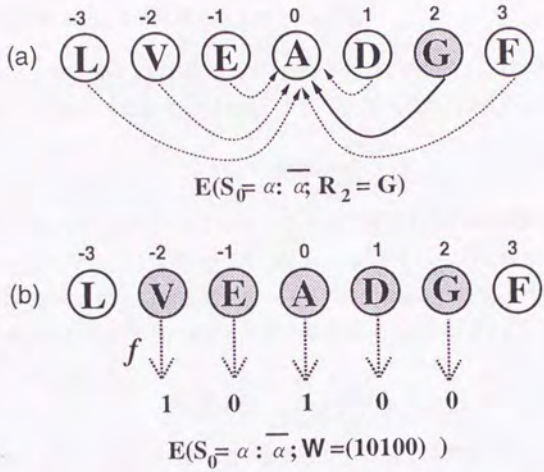


図 3.1 2種のポテンシャル。(a): 1残基ポテンシャル (b): 2元語ポテンシャル

3.2.3 BW-GOR 法

この2元語ポテンシャルと GOR 法の識別関数をカップリングパラメータ λ で組み合わせて、新しい識別関数を導入する。この識別関数を用いた方法を BW-GOR 法と呼ぶことにする。BW-GOR 法の識別関数は以下の式で定義される。

$$G_s^{BW-GOR}(\boldsymbol{x}) = (1-\lambda)\left[\sum_{m=-M}^{M} E(S_0 = s : \bar{s}; R_m = x_m)\right] + \lambda \cdot E(S_0 = s : \bar{s}; \boldsymbol{W} = \boldsymbol{b}(\boldsymbol{x})) \tag{3.12}$$

BW-GOR 法は $2N+1$ 次のパーセプトロンであり、その次数は $2N+1$ を越えない。本研究では2元語の長さを $7(N=3)$ として予測を行なうこととする。

3.3 MGOR 法と BW-MGOR 法

本研究では、より高い予測精度を実現するため、GOR 法を以下の2つの点で修正した修正 GOR(MGOR) 法を導入する。

1. 特別な「スペース」記号を 21 番目のアミノ酸として $\mathcal{A}$ に加える：この記号は N 末端か C 末端であることを表しており、末端の残基はコイルを形成しやすい傾向が表現さ

れやすくなる。Qian と Sejnowski は、この記号の導入によってニューラルネットワーク法の予測精度が向上したと報告している[12]。

2. しきい値 θ_s を識別関数に加える：このパラメータは2次構造 s だけに依存し、局所配列 $\boldsymbol{x}$ に依存しない。本研究ではパラメータ θ_s を以下の式で決定する。

$$\theta_s = -\mu \cdot \log_2 \frac{P(S_0 = s)}{P(S_0 = \bar{s})} \tag{3.13}$$

μ は任意に決定するパラメータである ($\mu \geq 0.0$)。各2次構造の出現頻度は、おおよそ $P(S_0 = \alpha) = 0.28, P(S_0 = \beta) = 0.21, P(S_0 = c) = 0.51$ であるので、しきい値パラメータ θ は $\theta_\alpha = 1.36\mu, \theta_\beta = 1.90\mu, \theta_c = -0.05\mu$ となる。つまり、μ が大きいほどコイルが予測されやすくなり、β シートが予測されにくくなることになる。この工夫は序論で述べた

$$P(c_j/\boldsymbol{a}) \simeq \frac{1}{P(c_j)^K} \prod_i P(c_j/a_i) \tag{3.14}$$

の近似式において、$K = (2M+1) - \mu$ としていることに相当する。

MGOR 法の識別関数は以下の式となる。

$$G_s^{MGOR}(\boldsymbol{x}) = \sum_{m=-M}^{M} E(S_0 = s : \bar{s}; R_m = x_m) + \theta_s \tag{3.15}$$

MGOR 法と2元語情報を組み合わせた BW-MGOR 法の識別関数は以下の式となる。

$$G_s^{BW-MGOR}(\boldsymbol{x}) = (1-\lambda)\left[\sum_{m=-M}^{M} E(S_0 = s : \bar{s}; R_m = x_m) + \theta_s\right] + \lambda \cdot I(S_0 = s : \bar{s}; \boldsymbol{W} = \boldsymbol{b}(\boldsymbol{x})) \tag{3.16}$$

BW-MGOR 法も BW-GOR 法と同様、$2N+1$ 次のパーセプトロンであり、その次数は $2N+1$ を越えない。

3.4 マルチプル配列アライメントデータの使用した場合の識別関数

予測対象となる配列と類縁のタンパク質の配列群をマルチプルアライメントしたデータを使うことで、より高い予測精度が得られることが期待される。配列が複数ある場合の識別関数の構築は、様々な方法が提案されているが、本研究では以下のように単純にそれぞれの配列の識別関数を加算した関数を用いる。

$$G_s^{multi}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{N_A}) = \sum_{p=1}^{N_A} G_s(\boldsymbol{x}_p) \tag{3.17}$$

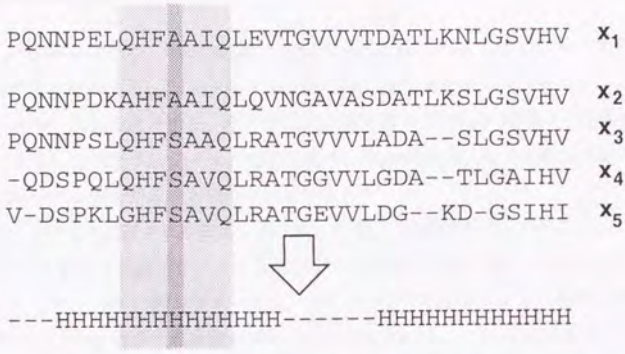


図 3.2 マルチプル配列アライメントデータを使用した場合の予測の概念図。

ここで、$\boldsymbol{x}_1$ は予測対象となる配列、$\boldsymbol{x}_2, \cdots, \boldsymbol{x}_{N_A}$ は類縁のあるタンパク質の配列群、N_A はアライメントされたタンパク質の個数である。図 3.2に、マルチプル配列を用いた場合の予測の概念図を示す。

3.5 使用したデータベースと予測精度の指標

結晶構造のデータベースとして、Rost と Sander が 2 次構造予測の評価に用いた 126 個のタンパク質を使用した[16]。このデータベースは解像度が 2.5Å より良いタンパク質からなり、かつ、相互の同一残基率が 80 残基以上の部分で 25% を越えない。マルチプルアライメントされた配列は、126 個のタンパク質に対応する HSSP データベースを用いた[21]。

予測精度の指標として、従来の 2 次構造予測の研究で良く用いられる 3 状態正答率 Q_3 と相関係数 C_s を用いた。これらの指標は予測結果の表 N_{sp} （＝正しくは 2 次構造 s であって、2 次構造 p と予測された残基の数）だけを用いて計算する。3 状態正答率 Q_3 は以下の式で定義される。

$$Q_3 = \frac{1}{N_{**}} \sum_{s \in \mathcal{C}} N_{ss} \times 100 \tag{3.18}$$

ここで $*$ はその添字に対する和を示す。相関係数 C_s は以下の式で定義される[22]。

$$C_s = \frac{N_{ss}N_{\bar{s}\bar{s}} - N_{s\bar{s}}N_{\bar{s}s}}{\sqrt{(N_{\bar{s}s} + N_{\bar{s}\bar{s}})(N_{\bar{s}\bar{s}} + N_{s\bar{s}})(N_{ss} + N_{\bar{s}s})(N_{ss} + N_{s\bar{s}})}} \quad \text{for} \quad s \in \mathcal{C} \tag{3.19}$$

3.6 予測精度の評価法

GOR 法、BW-GOR 法、MGOR 法、BW-MGOR 法については、ジャックナイフ法を用いて予測精度を評価する。ジャックナイフ法では、予測精度をテストするタンパク質を一つずつデータベースから取り出し、そのタンパク質を除いたデータベースで必要なパラメータの抽出（学習）を行なう。ジャックナイフ法は、他の評価法に比べ、推定の偏りの少ない予測精度が得られる場合が多いと報告されている[23,24]。

ジャックナイフ法で更新される値は、21×17×3 個の 1 残基ポテンシャル $E(S_0 = s; \bar{s}; R_m = a)$ の値、$2^7 \times 3$ 個の 2 元語ポテンシャル $E(S_0 = s; \bar{s}; \boldsymbol{W} = \boldsymbol{w})$ の値、1 個のしきい値 θ_s の値である。ここで、しきい値パラメータ μ、カップリングパラメータ λ、符号化関数 f の値はジャックナイフ法のなかで更新されずに一定の値を用いる。これらの値も更新されることが望ましいが、これらの値を統計的に推定することは極めて困難である。そこで、幾つかの値を用いたジャックナイフ法の評価を繰り返すことでこれらの最適な値を決定した。

3.7 符号化関数の探索法

符号化関数は長さ 20 の 0,1 の文字列であるので、全部で $2^{20} = 1,048,576$ の関数があり、全てを試すのは計算量的に困難である。そこで、モンテカルロ法の一種である、シミュレーテッド・アニーリング法[25] を応用して最適な符号化関数を探索する。最大化する目的関数として Q_3 を用いる。アルゴリズムは以下の通りである。

1. **初期化：** ランダムに f_0 を設定。温度 $T := T_0$、カウンタ $n := 0$ に設定。$Q_3(f_0)$ をジャックナイフ法で評価。
2. **符号化関数の変移：** $f_1 := f_0$ とする。ランダムにアミノ酸 a を選び、$f_1(a)$ を反転する。
3. **予測精度の評価：** ジャックナイフ法で $Q_3(f_1)$ を評価。$\Delta E := -[Q_3(f_1) - Q_3(f_0)]$ とする。
4. **判定：** ランダムに $P \in [0:1)$ を発生。$\Delta E < 0.0$ か $\exp(-\Delta E/T) > P$ なら $f_1 := f_0$
5. **温度変化：** $T := T \times r, n := n + 1$ とする。
6. $n < MAX$ ならステップ 2 に行く。

第 4 章

結果

4.1 GOR 法と BW-GOR 法の予測精度

4.1.1 シングル配列の場合

GOR 法のジャックナイフ法によって評価した予測精度は $Q_3 = 57.8$, $C_\alpha = 0.40$, $C_\beta = 0.35$, $C_c = 0.38$ となった（表 4.1）。この値は原著者の報告している値とほぼ同様である[3,10]。

BW-GOR 法については、シミュレーテッド・アニーリング法を用いて最も正答率 Q_3 が高い符号化関数の探索を行なった。探索に用いたパラメータは $T_0 = 0.01, r = 0.97, MAX = 500$ である。また、$\lambda = 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60$ の 7 種に対して、7 回の符号化関数の探索を行なった。$\lambda = 0.35$ の場合の Q_3 が最も高く、60.6% となった。収束の様子を図 4.1.2に示す。各ステップの予測精度はジャックナイフ法を用いて計算している。得られた最適な符号化関数は以下の通りである。

$$\begin{pmatrix} A & I & L & M & F & P & V & R & D & E & K & N & C & Q & H & S & T & W & Y & G \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.1}$$

GOR 法に比べ、2 元語ポテンシャルを加えた BW-GOR 法の予測精度 Q_3 は、2.8% 予測精度が高かった。相関係数 C_α, C_β, C_c も 0.02 から 0.04 向上した（表 4.1）。

4.1.2 マルチプル配列の場合

マルチプル配列アライメントを用いた場合の GOR 法では $Q_3 = 63.0$ となり、シングルの場合より 5.2% 正答率が上昇した。

マルチプル配列アライメントを用いた場合の BW-GOR 法は、符号化関数として、シングルで得られた式 4.1を用いた。7 種のカップリングパラメータ $\lambda = 0.30, 0.35, 0.40, 0.45,$

表 4.1 GOR 法と BW-GOR 法の予測精度

	Single				Multiple			
	Q_3	C_α	C_β	C_c	Q_3	C_α	C_β	C_c
GOR	57.8	0.40	0.35	0.38	63.0	0.50	0.42	0.43
BW-GOR	60.6	0.44	0.39	0.40	64.5	0.52	0.45	0.44
Improvement	2.8	0.04	0.04	0.02	1.5	0.02	0.03	0.01

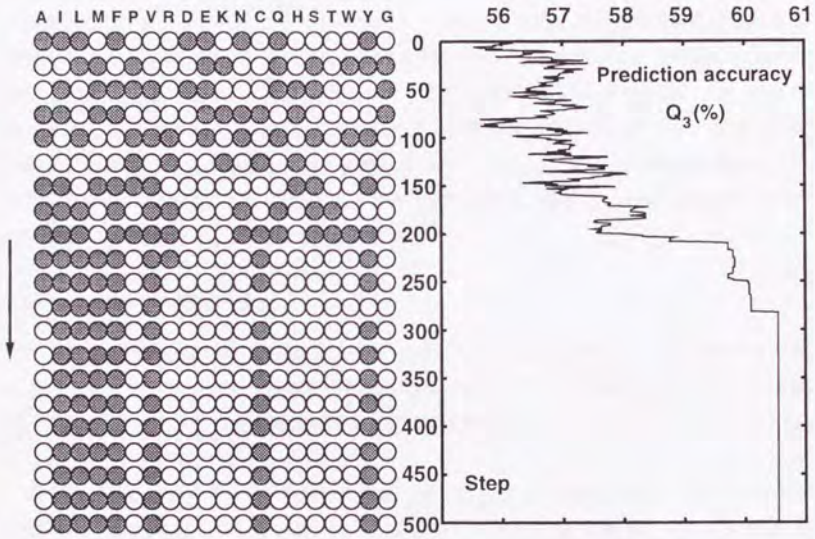


図 4.1　BW-GOR 法の最適な符号化関数の探索の過程。右が正答率 Q_3 とステップ、左がステップに対応する符号化関数である。白い丸は'0'、灰色の丸は'1' を表す。カップリングパラメータ λ は 0.55 を用いた。各ステップの予測精度はジャックナイフ法で計算している。ステップが増すにつれて、予測精度 Q_3 は 60.6% に収束し、そのとき対応する符号化関数が最適符号化関数である。

0.50, 0.55, 0.60 に対し、ジャックナイフ法で予測精度を計算したところ、$\lambda = 0.45$ のときの Q_3 が最も高かった。表 4.1に予測精度を示す。シングルに比べ、Q_3 が 3.9% 向上した。マルチプル配列を用いた GOR 法に比べ、BW-GOR 法は 1.5% 正答率 Q_3 が向上した。

4.2 MGOR 法と BW-MGOR 法の予測結果

4.2.1 シングル配列の場合

MGOR 法には、統計的に決定できないしきい値パラメータ μ が含まれている。このパラメータの決定のために 0.0 から 2.0 まで 0.05 刻みの異なる μ の値でジャックナイフ法で予測精度を評価した。結果を図 4.2 に示す。最も Q_3 が高かった $\mu = 1.30$ を用いることとする。予測精度を表 4.2に示す。$Q_3 = 63.4$ となり、GOR 法と比べて 5.6% 改善された。C_α と C_β は GOR 法に比べ、やや値が向上したが、C_β はやや下がった。これは、しきい値パラメータを加えることで β シートが予測されにくくなったためと考えられる。

BW-MGOR 法について、シミュレーテッドアニーリング法を用いて最適な符号化関数を探索した。しきい値パラメータは MGOR 法と同じ $\mu = 1.30$ を用いた。BW-GOR 法と同様に 7 種の異なる λ について探索を行なったところ、BW-MGOR 法の場合、$\lambda = 0.40$ の場合が最も高い正答率 $Q_3 = 65.4$ % が得られた。収束の様子を図 4.2.2に示す。最適な符号化関数は BW-GOR 法の場合と同様の関数になった（式 4.1）。$\lambda = 0.40$ で最適符号化関数を用いた場合の予測精度を表 4.2に示す。BW-MGOR 法の正答率 Q_3 は MGOR 法に比べ、2.0% 高い。相関係数 C_α, C_β, C_c も同様に改善した。

4.2.2 マルチプル配列の場合

マルチプル配列を用いた MGOR 法についても、シングルと同様にいくつかの異なるしきい値パラメータ μ について予測精度の評価を行なったところ、この場合 $\mu = 0.45$ のときが Q_3 が最大になった。この値を用いた場合の予測精度を表 4.2の左に示す。シングルに比べ Q_3 が 3.4% 向上し 66.8% となった。

マルチプル配列アライメントを用いた BW-MGOR 法では、符号化関数として、シングルで得られた式 4.1を用い、しきい値パラメータはマルチプルの MGOR 法と同じ $\mu = 0.45$ を用いた。7 種のカップリングパラメータ $\lambda = 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60$ に対し、ジャックナイフ法で予測精度を計算したところ、$\lambda = 0.40$ のときの Q_3 が最も高かった。表 4.2の右に予測精度を示す。シングルに比べ、Q_3 が 2.8% 向上し、本研究において試みた方法において最も高い正答率 $Q_3 = 68.2\%$ が得られた。マルチプル配列を用いた MGOR 法に比べ、正答率 Q_3 は 1.4% 向上した。

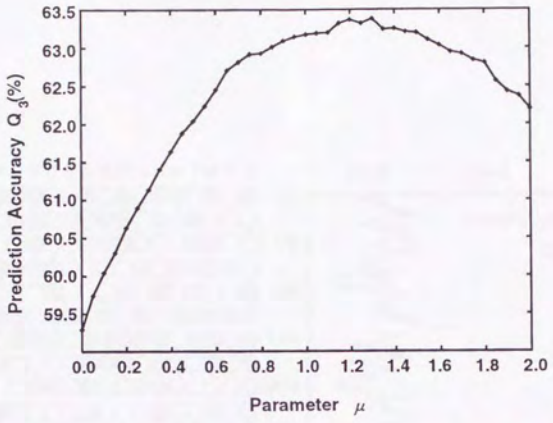


図 4.2　MGOR 法の予測精度 Q_3 としきい値パラメータ μ の関係。$\mu = 1.30$ のとき最も Q_3 が高い。

表 4.2 MGOR 法と BW-MGOR 法の予測精度

	Single				Multiple			
	Q_3	C_α	C_β	C_c	Q_3	C_α	C_β	C_c
MGOR	63.4	0.42	0.32	0.39	66.8	0.52	0.52	0.39
BW-MGOR	65.4	0.46	0.38	0.43	68.2	0.54	0.54	0.44
Improvement	2.0	0.04	0.06	0.04	1.4	0.02	0.05	0.02

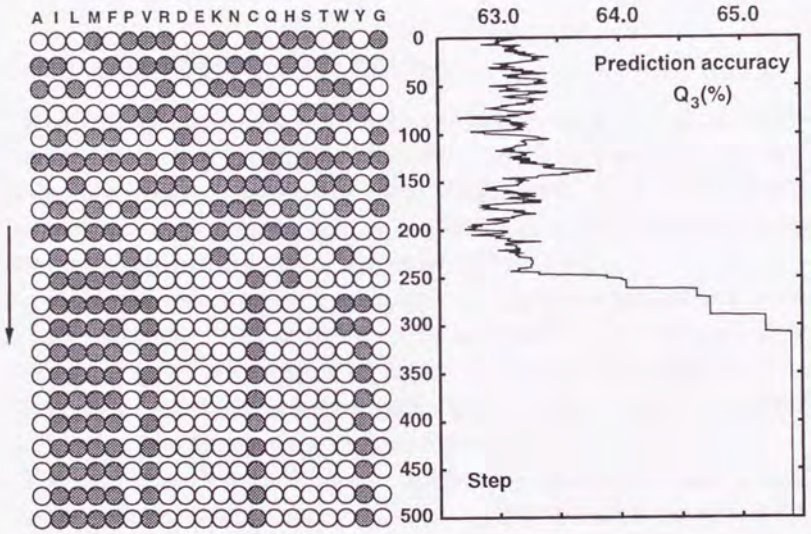


図 4.3　BW-MGOR 法の最適な符号化関数の探索の過程。右が正答率 Q_3 とステップ、左がステップに対応する符号化関数である。白い丸は'0'、灰色の丸は'1' を表す。カップリングパラメータ λ は 0.40 を用いた。各ステップの予測精度はジャックナイフ法で計算している。ステップが増すにつれて、予測精度 Q_3 は 65.4% に収束し、そのとき対応する符号化関数が最適符号化関数である。

第 5 章

考察

5.1 物理化学的な符号化関数を用いた場合の予測

物理化学的な特徴が2次構造の決定にどのような影響を及ぼすか調べるために、いくつかの物理化学的な意味付けが可能な符号化関数を用意し、それを用いた場合の予測精度の差を調べた。用意した関数は、非極性、荷電、極性、正荷電、負荷電、分子量、芳香属、脂肪属の8つである（図5.1）。非極性、荷電、極性の分類はBrandenとToozenに従った[26]。分子量の符号化関数は平均の分子量より大きい残基のみ'1'を返すとした。

BW-MGOR法を用いて、これらの符号化関数を用いた場合の予測精度を調べた。データはシングル配列を用い、しきい値パラメータ μ は1.30で予測を行なった。最適なカップリングパラメータ λ_{opt} は、符号化関数によって異なる。図5.1に5つの符号化関数について λ と Q_3 の関係を示した。例えば、非極性符号化関数の場合、λ_{opt} は0.40でそのときの正答率は64.2%である。各符号化関数の λ_{opt} と対応する正答率を表5.1と図5.1にまとめた。

8つの物理化学的な符号化関数の中では、非極性が最も正答率が高い。polar, positive, negative, positive, weight, aromaticに関しては、2元語情報を加えても予測精度は改善されず、Q_3 は λ について単調減少になる。また、非極性符号化関数は、シミュレーテッド・アニーリング法で得られた最適符号化関数とよく似ている。このことは非極性あるいは疎水性が2次構造の形成に影響を及ぼすことを示唆する。最適ではアラニンとプロリンが極性のグループに、システインとチロシンが極性のグループに分類される点が異なり、非極性というより、分子内部への埋もれやすさが表れていると考えることができる。

5.2 各2次構造に特徴的な2元語

図5.2に各2次構造に特徴的な2元語を示した。ここでは最適符号化関数を用い、2元語ポテンシャル $E(S_0 = s; \bar{s}; \boldsymbol{W} = \boldsymbol{w})$ の値が小さい順に10の語を示した。ただし、データベースでの観測数が10より小さい語は、値に信頼性がないとして除外した。α ヘリックス

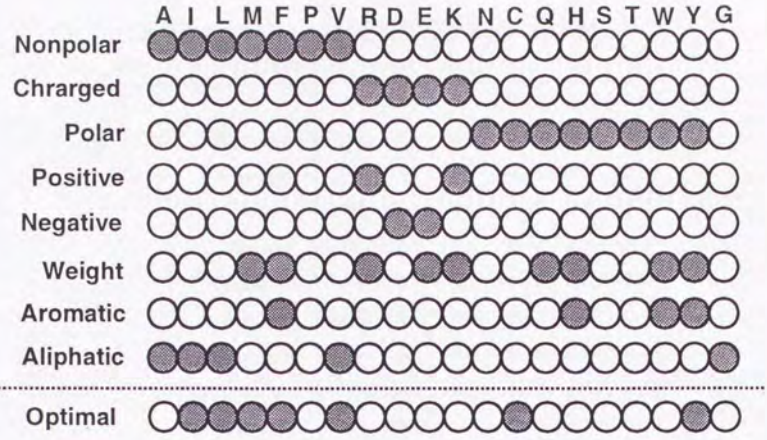


図 5.1　8つの物理化学的な符号化関数。白い丸は'0'、灰色の丸は'1'を表す。最下段の'Optimal'はシミュレーテッドアニーリング法で得られた最適な符号化関数である。

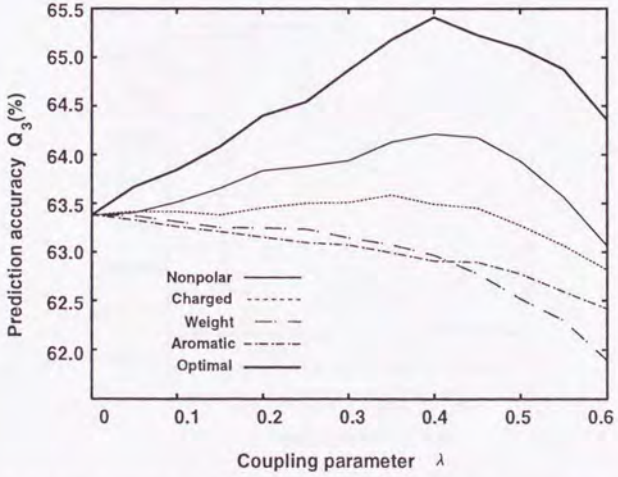


図 5.2　BW-MGOR 法の正答率 Q_3 とカップリングパラメータ λ の関係。4つの物理化学的符号化関数と最適符号化関数について表示した。

表 5.1 各符号化関数に対する BW-MGOR 法の予測精度

方法	符号化関数	λ^{*}_{opt}	Q_3	$\Delta Q_3^{\dagger}$	C_α	C_β	C_c
MGOR	-	-	63.4	-	0.42	0.32	0.39
BW-MGOR	Nonpolar	0.40	64.2	0.8	0.44	0.35	0.41
	Charged	0.35	63.6	0.2	0.43	0.33	0.40
	Polar	0.00	63.4	0.0	0.42	0.32	0.39
	Positive	0.00	63.4	0.0	0.42	0.32	0.39
	Negative	0.00	63.4	0.0	0.42	0.32	0.39
	Weight	0.00	63.4	0.0	0.42	0.32	0.39
	Aromatic	0.00	63.4	0.0	0.42	0.32	0.39
	Aliphatic	0.25	63.4	0.1	0.43	0.33	0.40
	Optimal	0.40	65.4	2.0	0.46	0.38	0.43

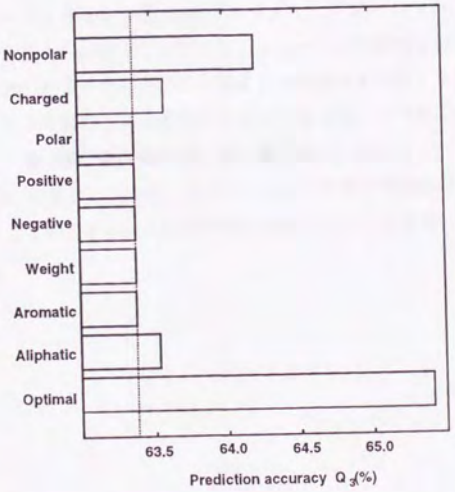


図 5.3　8つの物理化学的な符号化関数と最適符号化関数の正答率 Q_3。それぞれの λ_{opt} を用いて計算した。点線は MGOR 法の正答率 (63.4%) を示している。

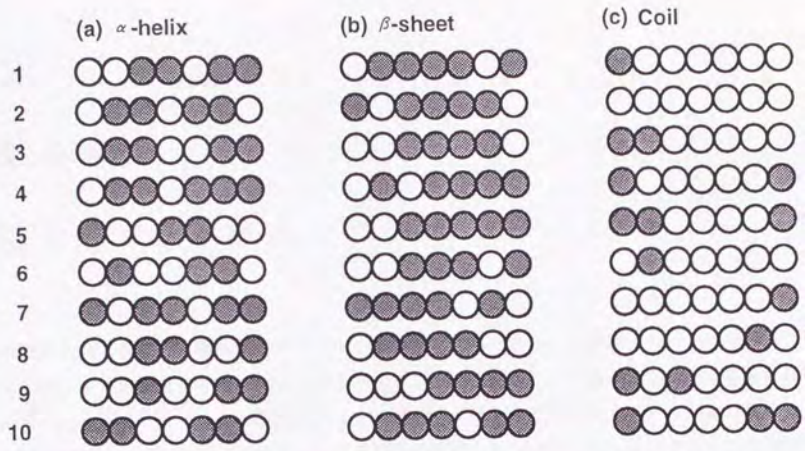


図 5.4 各2次構造に特徴的な2元語

に特徴的な語としては、○○●●○●●のような3～4残基の周期性のある語ばかりがならんでいる。これはヘリカルホイール[27]や疎水モーメント[28]がヘリックス構造を認識するのに有用であるという以前からの報告と一致する。βシートに特徴的な語は非極性基が中心に集中する○○●●●●○のような語が多い。疎水と親水が交互するような語（例えば、○●○●○●○）がβシートを好むという報告が従来からあるが、その傾向はそれほど強く現れてはいない。ただし、●○●○●○●や●○●○●○●などの語は$E(S_0 = \alpha; \bar{s}; \boldsymbol{W} = \boldsymbol{w}) > E(S_0 = \beta; \bar{s}; \boldsymbol{W} = \boldsymbol{w})$にはなっており、$\beta$シートになりやすい傾向は確かにあるが、それほど強くはないということができる。これは同様に疎水パターンを調査したWestとHechtの報告と一致する[29]。

第 6 章

第 I 部 のまとめ

本研究の第 I 部では、多残基ポテンシャル効率的に抽出できる 2 元語ポテンシャルを導入した。統計ポテンシャルを用いた古典的な方法である GOR 法と 2 元語ポテンシャルを組み合わせて用いることにより、予測精度を 1.4 ～ 2.8% 改善することができた。本研究で試みた方法の中では、マルチプルアライメント配列群を使用した BW-MGOR 法の正答率が最高の 68.2 % であった。この正答率は Rost と Sander の 71.6 % には及ばないものの十分比肩する精度であると考えられる。また、2 元語ポテンシャルによる正答率の改善は、多残基相互作用が 2 次構造形成に影響を及ぼすことを示唆する。さらに探索で得られた最適な符号化関数は明らかに非極性・疎水性と関連が深い。このことは疎水性相互作用が 2 次構造の形成に重要な役割を果たしていることを示唆する。

第 II 部

統計ポテンシャルを用いたタンパク質の 3 次構造予測

第 7 章

はじめに

7.1 3 次構造予測法概観

本節では、今までに行なわれてきた 3 次構造予測法を概観し、それらの利点および問題点を論ずる。

7.1.1 2 次構造パッキングによる階層的な方法

この方法は、はじめに 2 次構造予測を行ない、その結果をもとに 2 次構造セグメントを組み合わせることで 3 次構造を得る方法である。2 次構造セグメントを組み合わせる方法は、疎水性コアの形成を基本としながら、ヒューリスティックなルールを加えて行なう場合が多い。Cohen と Kuntz によって開発された方法[30] がこの種の方法で最も精緻に組み立てられたものであろう。彼らの方法は、ルールベースのターン予測をもとに 2 次構造を推定、可能な 2 次構造のパッキングをやはりヒューリスティックなルールで規定し、可能なパッキングを全回探索することで得る。パッキングのルールをまとめたものとしては、Richmond と Richard がグロビンのヘリックスに対して行なった研究[31] が著名である。また、Murzin と Finkelstein の多面体を用いたヘリックスのパッキングのパターン分類[32] は、この種のパッキングのルール化のロシア的な集大成といえるものだろう。ヒューリスティックなルールで可能な配座を制限することは、計算量を実現可能な範囲に押え込むことができるので、1990 年代になっても多くの研究が発表されている。Finkelstein と Reva は、ルールベースで制限した all- β のタンパク質の構造に対して、次に説明する「3D-1D」的な方法を適用することで、N 構造と対応したトポロジーの構造を予測できたと報告している[33]。また、2 次構造セグメントを固定したポテンシャル最小化計算も特に all- α のタンパク質に対して多くなされている[34-36]。

こういった階層的な方法は次のような問題点がある。

- **2 次構造予測の精度の低さ：** 第 I 部でも述べたように、完全な 2 次構造予測を行なう

ことは極めて難しく、最高でも 70 % 程度の精度しか期待できない。信頼性の低い2次構造セグメントを組み合わせて3次構造を組み立てても、その精度には限界がある。

- **階層性は完全には成り立たない**：実際のフォールディングの過程においては、2次構造の形成→2次構造のパッキングという階層性は完全には成り立たたない。特に β シートの多いタンパク質では2次構造の形成とパッキングは非階層的に進行していることを示唆する実験報告が得られている[37]。

7.1.2 ホモロジービルディングによるテンプレート的な方法

一般に同一残基率が30%を越える2つのタンパク質は、全体として類似した構造をとるといわれている。ホモロジービルディング法とは、与えられた配列と配列相同性が高い既知の構造を発見し、その構造を元にして、全体の3次構造を構築していく方法である。この方法はパターン認識でいう「テンプレート・マッチング」を立体構造予測に適用した方法であるといえる。ホモロジービルディングは、似ている構造がデータベース内にある場合は極めて高い精度で予測できるという特長を持つが、似ている構造がない場合は完全に予測不能となってしまう。その一般性の低さが最大の欠点である。しかし、X線結晶構造データベースの数は年々増加していること、タンパク質の全ファミリーの数はさほど多くない数に見積もられること[38]から、その一般性の低さも近い将来解消され、ホモロジービルディング法で全てのタンパク質の立体構造予測が可能になると予想する研究者もいる。しかしながら、少なくとも現在においては類縁の結晶構造が知られていないタンパク質ほど構造予測が望まれる場合が多いこと、全ての自然界のタンパク質の結晶構造が既知となった場合でも人工的に自然界に存在しないタンパク質を設計する場合にはホモロジービルディングでは役に立たないことから、より一般的な対象に適用できる構造予測法の開発は必要であろう。

結晶構造データベースを調べてみると、配列相同性がさほど高くなくても類似した構造をもつ構造対があることがわかる。そこで、こういった「淡い」進化的類似性を評価するため、配列対配列ではなく、配列対構造で評価する方法が、1990年代に入ってから相次いで開発された[39–42,1,2]。この方法は3次元と1次元を比較することから「3D-1D法」、幾つかの構造に配列を載せて評価することから「Threading法」、あるいは構造に当てはまる配列を探すという操作が可能なことから「インバース・フォールディング」法とも呼ばれる。当初、3D-1D法は進化的に無縁で3次構造が似ている「収れん進化」と呼ばれるタンパク質のペアを認識できることが期待されたが、実際には難しく、あくまで進化的に関係のあるタンパク質の認識に有用であるというのが現在での評価である。

3D-1D法では、構造と配列の適合度として第1章で述べた統計ポテンシャルを用いることが普通である。統計ポテンシャルの3次構造予測において最も成功した例が3D-1D法である

ともいえるだろう。

著者は修士過程において、実際に「3D-1D 法」の研究を行ない[43]、その長所は認めつつも、この方法に対して以下のような問題意識を持つに至った。

1. **スコアの意味が不明確：** 3D-1D 法ではアミノ酸配列と構造との適合度を示す統計ポテンシャルに、アライメントの挿入／削除ペナルティのを加えたものをスコアとして用いる。統計ポテンシャルは第 1 章に述べたように、物理的な自由エネルギーあるいは確率モデルの判別関数としての意味を持つ量であるが、挿入／削除ペナルティは進化に由来する確率であり、全く異種の量を加えていることになる。

2. **一般性の低さは本質的に解消されない：** 3D-1D 法はあくまで、「淡い」進化的類縁を認識できるホモロジービルディング法であり、ホモロジービルディング法の最大の問題点である「一般性の低さ」は緩和されるが、本質的に解消されるわけではない。

3. **評価の難しさ：** 3D-1D 法が「配列は似ていないが構造は似ている」タンパク質の認識を目的とするからには、構造の類似度を客観的に定義することが必要であるが、実はこの定義自体非常に難しい問題で、特に遠い類似性の場合コンセンサスとなるような指標が未だ確立されていないのが現状である[44]。

7.1.3 ポテンシャル最小化計算による演繹的方法

配座に対応するポテンシャル関数を定め、そのポテンシャル最小化計算を行なうことで構造予測を行なう方法である。この方法は、3 つの中では最も物理的な立場に近い演繹的な方法である。もっとも、物理的に厳密であろうとするなら、溶媒を含んだ系での自由エネルギー最小状態を統計力学的に推定しなければならないが、構造予測として行なわれているのは計算量を削減するため、もっと単純化されたモデルを用いるのが普通である。一般に (1) 溶媒は平均場として取り込み、(2) 側鎖原子群を一つの作用点で表現した、主鎖のみのモデルを用いる場合が多い。この種の方法で最も初期に行なわれたのは 1975 年の Levitt と Warshel による BPTI のシミュレーションである[45,46]。近年、計算機の性能が向上するに伴い、多くのこの種の研究が試みられている[1]。

現状では、ポテンシャル最小化計算による予測は、良い場合でも、RMS が 5 から 8 Å 程度の構造しか得られず、予測精度としては先に挙げたホモロジービルディング法に及ばない。しかし、(1) ホモロジービルディング法に比べ、一般性の高い予測法が構築できる可能性があ

[1]特に最近では、Dill らのグループを中心に極度に単純化されたモデルのポテンシャル最小化計算による研究が盛んに行なわれているが[47]、これらの研究はフォールディング現象の定性的な理解を主たる目的としており、構造予測より思考実験に近い。

ることと (2) 自由エネルギー最小化という物理的な描像に近い形であるため、フォールディングのメカニズムの解明に貢献できる可能性がある点が注目される。

7.2 ポテンシャル最小化計算による構造予測の3つの要素

本研究の第II部では、最後に述べたポテンシャル最小化計算による方法が最も潜在的な可能性を持った研究の必要な方法であると考え、この方法による構造予測を行ない、その有効性と問題点を明らかにすることを目的とする。そのためにより詳細に今までの研究を述べることとする。ポテンシャル最小化計算による構造予測は (1) 構造表現、(2) ポテンシャル関数、(3) 構造探索の3つの点から特徴付けられる。以下にこの3つについて過去に行なわれてきた研究についてまとめる。

構造表現

ここでは、構造を記述する変数が連続値であるモデルを連続モデル、離散値であるモデルを離散モデルとして分類して説明する。

1. **連続モデル：** 連続モデルとして最も一般的なのは (1) 主鎖の二面角 ϕ, ψ を変数とするモデル[48] と (2) C^α を接続した仮想的な鎖の結合角と二面角を変数とするモデル[49] がある。全原子モデルでよく用いられる xyz 座標による表現は、変数の数の多いことから、構造予測を目的とした研究にはほとんど用いられない。また、2つの角度を一つのパラメータで記述したモデル[46] や距離行列で記述するモデル[50,51] なども提唱されている。連続モデルは詳細に構造が記述できるが、探索すべき構造空間が大きくなることが欠点である。

2. **離散モデル：** 離散モデルは xyz 座標を離散化したモデルと角度を離散化したモデルの2つに分類される。

   xyz 座標の離散化： xyz 座標を離散化したモデルは、単に配座空間を小さくできるだけでなく、残基間の距離が限られた値しかとりえないため、テーブル化により高速な距離計算が可能となるという長所もある。多く使われる方法は、(1) 立方格子[52-55]、(2) ダイアモンド格子[56-59]、(3) 210 格子[60-62] の3つである。これらのモデル共通の欠点は、α ヘリックス・β シートといったタンパク質特有の局所構造が完全には表現できないことである。特に立方格子においては、二面角が0度と180度しかとりえないため螺旋構造の表現ができない。また、遠距離のコンタクトが制限されるという問題もある。立方格子においては i 番目と $i+2k+1$ 番目の残基は近傍で接することができるが、i 番目と $i+2k$ 番目の残基はいかなる配座であっても近傍で接することはで

きない。例えば1番目と200番目の残基は絶対に接触できないことになる。これらの状況はダイアモンド格子や210格子では緩和されるが、本質的な困難は変わらない。

角度の離散化： もう一つの離散化の方法として角度を離散化する方法がある。連続モデルと同様、(1) 主鎖の二面角の (ϕ, ψ) を離散化するモデル[63,35,36,64] と、(2) C^α 鎖の結合角・二面角を離散化するモデル[65,66] の2つがある。角度を離散化した場合の長所としては、局所構造が比較的よく表現できることと、xyz の離散化に比べ離散化の程度を自由に選択できるため、連続モデルとの接続に向いていることの2点が挙げられる。局所構造と大域的構造を同時に正確に表現することが難しいことが欠点である。

ポテンシャル関数

こういった単純化された構造表現のポテンシャル関数を得る方針として、

1. 全原子モデルのパラメータから平均場を抽出する方法
2. アミノ酸ごとの実験的なパラメータを利用する方法
3. 結晶構造データから統計的に抽出する方法
4. 任意に与える方法

の4つがある。1の全原子モデルから抽出する方法は主にファンデルワールス力に用いられる場合が多く[46]、2は溶媒の影響を表面積で見積もる場合に多く用いられる[67]。3の統計的に抽出したポテンシャルが最も多く用いられる方法であり、序論で述べた統計ポテンシャルの定式化で行なわれる場合が多い。(a) 残基間の距離・コンタクト (b) 局所構造 (c) 露出度 の3つの構造特徴がよく用いられる。Miyazawa-Jernigan のコンタクトポテンシャル[68,69] と Sippl の距離ポテンシャル[4] が有名である。Sippl のポテンシャルは序論で述べた $-\log P(c/a)/P(c)$ の統計ポテンシャルであるが、Miyazawa-Jernigan のポテンシャルは、結晶構造データベースの統計を元に構築する点は共通しているが、序論で説明したものとは異なる物理過程を想定しており、定式化がやや異なる。また、同様に統計データを元にするが、パラメータの決定を 反復的な計算で行なう方法[70] や、スピングラス理論による重判別分析で決定する方法[71] も提案されている。4の任意に与える方法は、1,2,3 によって決定できなかったパラメータや、異なる由来のポテンシャルの重み付けのパラメータがこれにあたる。衝突のポテンシャルは任意に与えることが多い。

構造探索

モンテカルロ法[67,58,59]、シミュレーテッド・アニーリング法[54,49]、遺伝的アルゴリズム[35,36] といった確率的探索法で行なう場合が多いが、ポテンシャルの勾配を利用した分子動力学法[45]

や、条件を限定した全回探索[52,56,57,64]、疎水性相互作用による凝集にのみ着目して開発した「疎水ジッパー戦略」を適用した例もある[55]。また、遺伝的アルゴリズムを用いて既知の結晶構造の一部を組み合わせて構築する方法が提案されている[72]。この方法はホモロジービルディングとポテンシャル最小化の中間にあるという意味で興味深い。Scheraga らのグループは全原子のポテンシャル関数を用いて、ポテンシャル最小化法を様々に工夫した方法の開発を精力的に続けている[73]。

7.3 モデルの詳細さと構造探索の効率のトレードオフ

ポテンシャル最小化計算において、モデルの詳細さと構造探索の結果の信頼性はトレードオフの関係にある。詳細な構造表現（例えば全原子モデル）にすると、緻密に N 構造をポテンシャル最小点として記述できるはずだが、構造探索に要する計算量は莫大となり、現実には不十分な構造探索しかできない危険性がある。一方、粗い構造表現（例えば 3 次元立方格子モデル）の場合、構造探索は充分行なうことができ、ポテンシャル最小構造が得られる可能性が高いが、粗い構造表現ではそもそも N 構造を表現できない可能性がある。そこで、現実に計算時間で可能で、しかもある程度の精度を保った中間的な複雑さのモデルを構築することが不可欠である。

しかし、こういった中間的なモデルは、モデルの設計に多くの任意性が含まれるという問題点がある。複雑な全原子モデルであれば、ポテンシャルのパラメータは量子力学計算あるいは低分子の実験データからのフィッティングで求めることができる。また、逆に極度に単純な HP 格子モデル[47] の場合、ポテンシャルのパラメータは疎水基が接触したときの σ のみしかないので、別の意味で任意性が入り込む余地はない。しかし、中間的な複雑さを持つモデルでは、モデルの妥当性が最終的なポテンシャル最小構造のみでしか確認できないため、研究者は、大量のポテンシャル最小化計算を繰り返しながら、少しずつモデルを修正していくという延々と続く試行錯誤を行なうはめになる。

近年提案されたポテンシャルの評価法として、Threading テストがある。これは、結晶構造データベースの部分構造に配列をのせてポテンシャルを評価し、N 構造がポテンシャル最小構造になっているかどうかチェックする方法である。この方法は比較的少ない計算量で行なうことができるため、ポテンシャルの設計の中間評価として有用であると考えられる。しかし、Threading テストは、結晶構造データベースという限定された構造空間のみのテストであるので、あくまでポテンシャルが満たすべき必要条件の一つでしかない。Threading テストは結晶構造データベース内の「タンパク質らしい」構造群から N 構造を認識できるかどうかのチェックであり、そうでない構造群（例えば、コンパクトでない、2 次構造が形成されていない構造など）から N 構造を認識できる保証はない。

表 7.1 モデルの複雑さと構造探索の効率のトレードオフ

モデルの複雑さ	探索空間	ポテンシャルのパラメータ数	例
複雑	大きい	多い（低次から演繹的に決定）	全原子モデル
：	：	：	：
単純	小さい	少ない（任意に指定）	HP 格子モデル

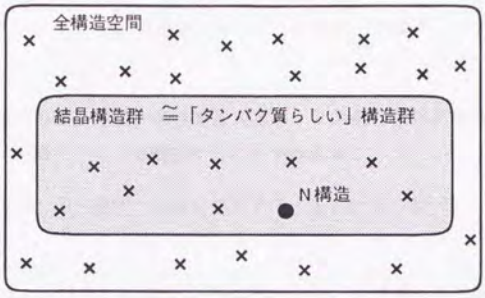


図 7.1 全構造空間、結晶構造群、「タンパク質らしい」構造群の関係の概念図。

7.4 本研究の方法

前述したように、本研究の第 II 部では、進化的類縁のある構造が未知でありホモロジービルディングが不可能であるようなタンパク質を予測対象として想定し、ポテンシャル最小化計算による構造予測を行なう。その際にポテンシャルとしては、第 1 章で述べた統計ポテンシャルを用い、その有効性と問題点を明らかにすることを目的とする。

本研究で採用する統計ポテンシャルは、前述したように 3D-1D 法においては大きな成果を得ているが、ポテンシャル最小化計算においてはそれほど良い結果を得ているわけではない（例えば[74]）。その理由としては、統計ポテンシャルは他の結晶構造群から N 構造を識別するように設計されており、全構造空間から N 構造を識別できる保証はないことが考えられる（図 7.1）。第 1 章の統計ポテンシャルの理論的根拠の部分で述べたように、統計ポテンシャルは Threading 問題における確率モデルの識別関数として自然に導出される。しかし、ポテンシャル最小化計算においては、3D-1D 法と異なり、結晶構造群以外の構造も予測対象となる。そういった全構造空間において、N 構造が統計ポテンシャルの最小構造になる理論的な保証はどこにもない。

本研究では結晶構造群を「タンパク質らしい」構造の条件を満たす構造群であると仮定し、ポテンシャル最小化の目的を【統計ポテンシャルの最も低い構造を探すこと】ではなく、【統計ポテンシャルの最も低い「タンパク質らしい」構造を探すこと】に設定することで、モデルの設計の試行錯誤を最小限にし、かつ、より精度の高い構造予測を行なうことを目指す。具体的には以下の手続きで予測法を構築していくこととする。

1. 構造表現を設定する。構造表現は、望まれる精度でN構造を表現できるかどうかで評価する。
2. 配列依存ポテンシャル E_{seq} を統計ポテンシャルを用いて設定し、Threadingテストで評価する。
3. 「タンパク質らしい」構造の条件を設定する。この条件は結晶構造データベースの構造がこの条件を満たすことを確認することで評価する。
4. 「タンパク質らしい」構造の条件を満たすほど低くなるようなポテンシャル E_{pro} を設計する。
5. $E = (1-\lambda)E_{pro} + \lambda E_{seq}$ に対するポテンシャル最小化計算を行なう。λ を何通りかに変えて繰り返し計算を行ない、多数のポテンシャル局所最小構造を生成する。
6. 得られたポテンシャル最小構造の中から、「タンパク質らしい」条件を満たす構造を選択し、その中から E_{seq} の最も低い構造を予測構造とする。

第 8 章

方法

8.1 使用する結晶構造データ

第II部では、Kocherらの用いた69のタンパク質のX線結晶構造データを統計データとして用いる[75]。これらは全て20%以下の同一残基率であり、また解像度は2.5 Åより良い。さらに、(1) 同一でない鎖を含むマルチマーは独立した別の鎖と数え、(2) ループ部の座標が一部ないタンパク質（2TS1と1GOX）については、その部分の前後で2つの鎖とし、(3) 繰り返し配列のある7WGAについては最初の44残基のみ使用する。最終的に表8.1に示した75個のタンパク質鎖となる。また、ポテンシャル最小化計算においては、表8.2に示した5つの80残基以下のタンパク質を用いる。ポテンシャル最小化を行なう際に用いるポテンシャルパラメータと二面角代表点の抽出には80残基以下のタンパク質鎖（表8.1の*印の付いたタンパク質）を除いた62個のタンパク質のデータを用いる。

8.2 構造表現

本研究では構造表現法として、離散 (ϕ,ψ) モデルを採用する。序論で述べたように、探索すべき構造空間を限定するために構造表現の離散化はどうしても必要である。離散化の方法としては、2次構造をリアルに表現するためにxyz座標ではなく、主鎖の二面角 (ϕ,ψ) を離散化することとする。このモデルは、主鎖のみで表現され、側鎖の相互作用は全て C^β 原子で代表されるとする（図8.1）。結合長・結合角はすべて理想的な値に固定し、主鎖の二面角 ϕ,ψ のみを変数にする。

二面角 (ϕ,ψ) はいくつかの代表点の値のみ離散的に値をとる。代表点を決定するために結晶構造データベースの角度分布のクラスタリングを行なう。クラスタリングの方法は数多く提唱されているが、本研究ではアルゴリズムの単純さと任意パラメータの少なさからK-mean法を採用した[24]。アルゴリズムは以下の通りである。

- K 個のクラスタの中心点 $\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_K$ をランダムに決める。

表 8.1　Threading テストに用いる 75 のタンパク質鎖。ポテンシャル最小化計算に用いるポテンシャル値は * のついたタンパク質以外から抽出する。

8CATA	3GRS	2YHX	2CTS	1PHH	8ADH	2LBP	1GD1O
4MDHA	1PRCC	6LDH	2APR	1PRCM	1PFKA	3TLN	5CPA
1RHD	2CYP	1CSEE	1PRCL	1PRCH	2CAB	1TIMA	2CNA
1TPP	2ACT	2FB4L	2TS1-1	2ALP	3ADK	1GOX-1	2STV
1GP1A	1GCR	2LZM	1GOX-2	3DFR	2LH4	2SODO	2SNS
4HHBA	4FXN	1LZ1	2CCYA	7RSA	1BP2	2PABA	1HMZA
1CCR	1ACX	4CPV	2CDV	1FD2	2TS1-2	1RNT	1PCY
3FXC	2GN5	1HIP	3B5C	1CC5	451C	1HOE*	2ABXA*
1UTG*	1SN3*	2CRO*	1CSEI*	1NXB*	5PTI*	4RXN*	1CRN*
7WGAA-1*	1PPT*	2MLTA*					

表 8.2 ポテンシャル最小化計算を行なう 5 つのタンパク質

PDBcode	COMPND	残基数	SS 結合数
4RXN	RUBREDOXIN	54	-
5PTI	TRYPSIN INHIBITOR	58	3
2CRO	434 CRO PROTEIN	65	-
3ICB	CALCIUM-BINDING PROTEIN	75	-
1UBQ	UBIQUITIN	76	-

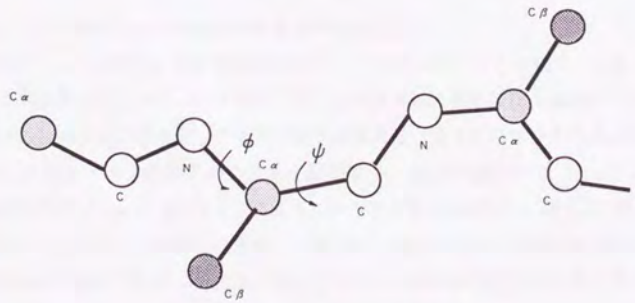


図 8.1 主鎖の二面角による構造表現

表 8.3 クラスタの数とクラスタ内分散の値

K	3	4	5	6	7	8	9	10	11	12
σ_W	1757.9	1339.2	1044.6	798.9	673.7	619.9	566.5	518.5	479.1	443.2

表 8.4 Set6、Set12 における各アミノ酸の代表点の数

	A	I	L	M	F	P	V	R	D	E	K	N	C	Q	H	S	T	W	Y	G
Set6	5	3	4	4	4	2	3	5	7	4	5	6	5	5	5	6	5	3	4	9
Set12	9	5	7	6	7	4	6	8	12	8	10	11	9	9	10	12	10	6	8	18

- クラスタの中心点が変化しなくなるまで以下の処理を繰り返す。収束したときの $\boldsymbol{c}_j$ が代表点となる。

  1. データ $\boldsymbol{x}$ を距離 $(\boldsymbol{x}-\boldsymbol{c}_j)^2$ が最も小さいクラスタ j に分類する。
  2. クラスタの中心点をそのクラスタに属するデータの重心に設定する

$$\boldsymbol{c}_j = \frac{1}{N_j}\sum_{\boldsymbol{x}\in j}\boldsymbol{x}$$

クラスタリングは本質的に任意のパラメータの設定が不可避である。K-mean 法では (1) 初期クラスタ中心点 、(2) クラスタ数 K の 2 つが任意パラメータとなる。本研究では複数回、初期クラスタ中心点 を変えてクラスタリングを行ない、得られたクラスタのクラスタ内分散 σ_W の最も値の小さいクラスタを採用することとする。

$$\sigma_W^2 = \frac{1}{N}\sum_j\sum_{\boldsymbol{x}\in j}(\boldsymbol{x}-\boldsymbol{c}_j)^2 \tag{8.1}$$

図 8.2にクラスタ数が 3,6,12 の場合の代表点の位置を示す。

本研究では、少ないクラスタ数で効率的にタンパク質を表現するためにアミノ酸ごとに異なる数のクラスタ数のデータセットを作成した。クラスタ数の選択には、全体でクラスタリングした場合の級内分散に最も近いクラスタ数を選択する。表 8.4 に 2 つの代表点のセット、Set6 と Set12 の各アミノ酸ごとのクラスタ数を示す。Set6 は級内分散 が $\sigma_W = 798.9$、Set12 は級内分散 が $\sigma_W = 443.2$ となるようにクラスタ数を決定した。図 8.3と図 8.4にアラニン、プロリン、グリシンの場合の Set6 と Set12 の代表点を示す。配座の自由度が高いグリシンには多数の代表点が、逆に自由度の低いプロリンには少数の代表点が割り当てられる。

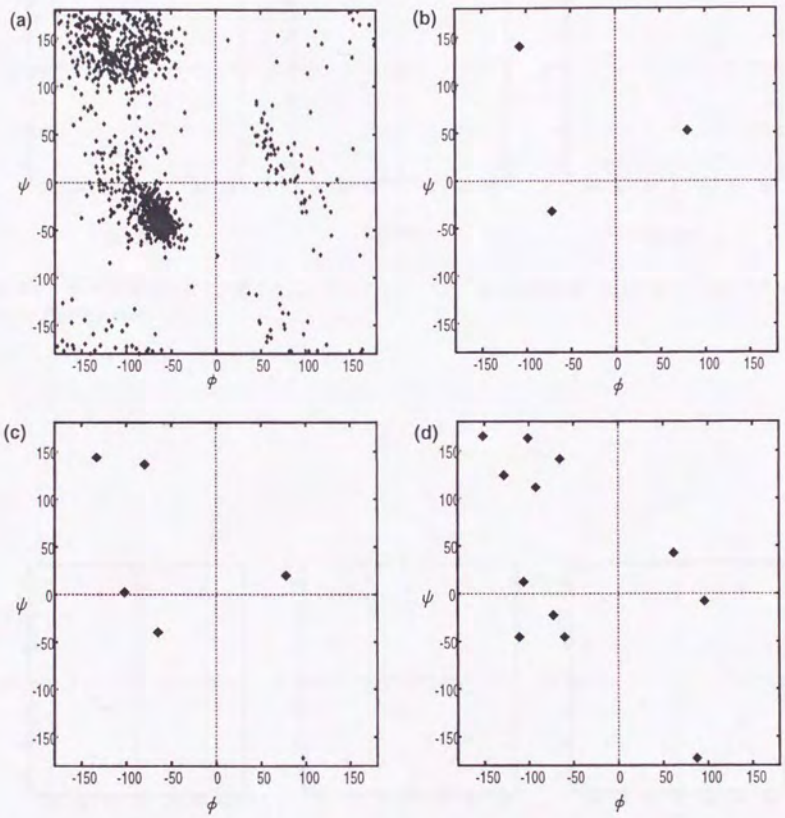


図 8.2　クラスタリングした (ϕ, ψ) 代表点。(a) 結晶構造の統計データ (b) クラスタ数 3 の場合 (c) クラスタ数 6 の場合 (d) クラスタ数 12 の場合

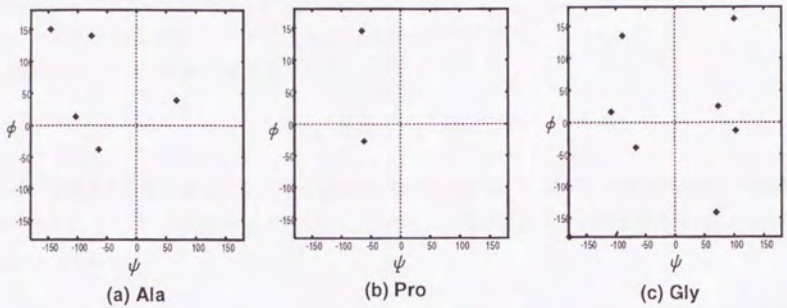


図 8.3　Set6 の場合のアミノ酸ごとにクラスタリングした (ϕ, ψ) 代表点。(a) アラニン (b) プロリン (c) グリシンのみ示す。

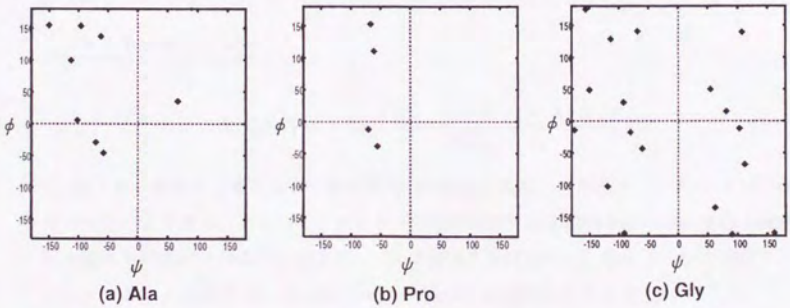


図 8.4　Set12 の場合のアミノ酸ごとにクラスタリングした (ϕ, ψ) 代表点。(a) アラニン (b) プロリン (c) グリシンのみ示す。

8.3 ポテンシャル

本節では、統計ポテンシャルの定式化に従い、いくつかのポテンシャルを導入する。後に導入されると「タンパク質らしい」構造のためのポテンシャル E_{pro} と区別するために、このポテンシャルを配列依存ポテンシャル E_{seq} と呼ぶことにする。

統計ポテンシャルは第1章で述べたように

$$E(c/a) = -\log \frac{P(c/a)}{P(c)} \tag{8.2}$$

の形で計算される量である。ここで c は構造に関する特徴量、a はアミノ酸配列に関する特徴量である。ここで、確率分布 $P(c), P(c/a)$ は次のような統計データの頻度分布 $F(c), F(c/a)$ によって近似する。

$$P(c) \simeq F(c) = N_c/N, \quad P(c/a) \simeq F(c/a) = N_c/N_{ca} \tag{8.3}$$

N は全観測数、N_c は特徴 c であった観測数、N_a は特徴 a であった観測数、N_{ca} は特徴 c かつ特徴 a であった観測数である。Sippl らは、データ数が少ない場合、特に $P(c/a) \simeq F(c/a)$ の値の信頼性に問題があるとして、以下の式で近似することを提唱している[4]。

$$P(c/a) \simeq \left(1 - \frac{m\sigma}{1+m\sigma}\right) F(c) + \frac{m\sigma}{1+m\sigma} F(c/a) \tag{8.4}$$

ここで m は信頼性の目安となる観測数である。

c, a として次のような事象を選びポテンシャルを作成した。

1. コンタクトポテンシャル:

$$EC_{ab}(C) = -\log \frac{P(C_{ij} = 1/A_i = a, A_j = b)}{P(C_{ij} = 1)} \tag{8.5}$$

$C_{ij} = 1$ は i 番目と j 番目の C^β 原子が R_c Å 以内にあることであり、$A_i = a$ は i 番目の残基が a であることを示す。グリシンの場合は C^α 原子座標を用いる。また3残基以上離れた残基ペアのみを考慮する。R_c は何通りか試行して、最もよい値を選択することとする。このポテンシャルは Hinds と Levitt が提唱したものに近い[56,57]。

2. 距離ヒストグラムポテンシャル

$$ED_{ab}(d) = -\log \frac{P(D_{ij} = d/A_i = a, A_j = b)}{P(D_{ij} = d)} \tag{8.6}$$

ここで $D_{ij}=d$ は i 番目と j 番目の C^{β} 原子が d Å 以内にあることであり、 $A_i=a$ は i 番目の残基が a であることを示す。距離は連続ではなく、1 Å ごとのヒストグラムに分ける。このポテンシャルはコンタクトポテンシャルに比べて事象の捉え方が複雑であり、十分なデータ数が確保できない危険性があると考えられるので、8.4式の補正を行なう。m として $D_{ij}=d, A_i=a, A_j=b$ である残基の数 を用い、$\sigma=1/50$ とした。このポテンシャルは Sippl らが提唱したものに近い[4] が、用いる構造表現の粗さからより簡素な形にしている。まず、(1)Sippl らが考慮している Topological Level $(k=|i-j|)$ は無視する。近距離相互作用は次に説明する局所構造ポテンシャルで表現されると考える。次に (2) 距離として 3 Å から R_{max} Å までしか考慮しない。これは Kocher らの方法[75] を踏襲した。R_{max} はいくつかの値を試行して、最も成績の良い値を選択することとする。

3. 局所構造ポテンシャル:

$$EL_{ak}(s)=-\log\frac{P(S_i^{\phi\psi}=s/A_{i+k}=a)}{P(S_i^{\phi\psi}=s)} \tag{8.7}$$

$A_{i+k}=a$ は $i+k$ 番目のアミノ酸が a であることであり、$S_i^{\phi\psi}=s$ は i 番目の二面角 (ϕ,ψ) 領域が s であることである。このポテンシャルは第 I 部で説明した GOR 法の 1 残基ポテンシャルと似た量であるが、2 次構造ではなく二面角領域を構造特徴としている点が異なる。ここで二面角 (ϕ,ψ) 領域は、'A','B','C' の 3 つの領域に設定する。図 8.5に領域の境界を示した。前後 M 残基を加算して用いる (k は $-M$ から M まで)。M の値は、いくつかの値を試行して、最も成績の良い値を選択することにする。

4. コンタクト数ポテンシャル

$$EN_a(n)=-\log\frac{P(N_i=n/A_i=a)}{P(N_i=n)} \tag{8.8}$$

$A_i=a$ は i 番目のアミノ酸が a であることであり、$N_i=n$ は、i 番目の残基とコンタクトしている他の残基の数が n であることである。コンタクトの定義はコンタクトポテンシャルと同一とする。

5. 2元コンタクトポテンシャル:

$$EB_{cd}(C)=-\log\frac{P(C_{ij}=1/B_i=c,B_j=d)}{P(C_{ij}=1)} \tag{8.9}$$

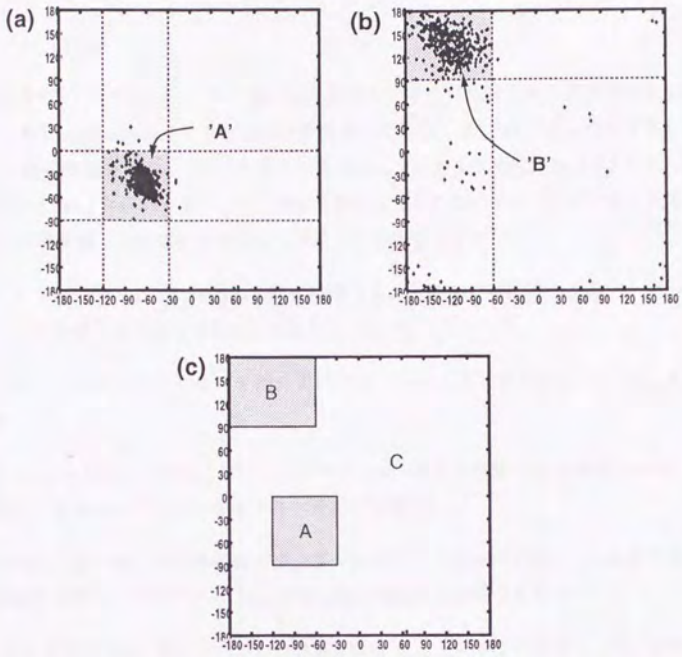


図 8.5　結晶構造データベースの中で DSSP で'H' と判定された残基の二面角分布 (a),'E' と判定された残基の二面角分布 (b)。定義した二面角領域 (c)。

このポテンシャルは、アミノ酸を 2 種類に分類して考慮したコンタクトポテンシャルである。$B_i = c$ は、2 元に分類した i 番目のアミノ酸が c ($c = f(a_i)$) であることである。f は、第 I 部で用いた符号化関数である。コンタクトの定義はコンタクトポテンシャルと同一とする。このポテンシャルは、構造予測の性能はあまり期待できないが、いくつかの符号化関数の比較により Threading に重要な特性の抽出ができる可能性を持つ。

8.4 構造探索

8.4.1 一般的方針

配列依存ポテンシャル E_{seq} は、Threading テストでその性能をある程度評価することができる。もし、Threading テストが良好な結果となれば、E_{seq} は「タンパク質らしい」構造から N 構造を見つけることができることを示唆し、構造探索の問題は【与えられた配列の E_{seq} が最も低い「タンパク質らしい」構造を探すこと】に縮小することができると考えられる。この問題を解くために本研究では次のような手続きを行なう。

1. 「タンパク質らしい」構造の条件を設定する。この条件は実際の結晶データが満たしているかどうかで妥当性をある程度チェックできる。
2. 「タンパク質らしい」条件を満たすほど低くなるようなポテンシャル E_{pro} を設定する。
3. $E = (1-\lambda)E_{pro} + \lambda E_{seq}$ に対するポテンシャル最小化計算を λ を何通りか変えて繰り返し、多数のポテンシャル局所最小構造を生成する。
4. 得られたポテンシャル局所最小構造群の中から、「タンパク質らしい」条件を満たす構造を選択し、その中から E_{seq} の最も低い構造を予測構造とする。

この手続きの長所は、陽に現れるのは評価可能な E_{seq} と「タンパク質らしい」条件の2つだけであり、設定の難しい E_{pro} や、E_{pro} と E_{seq} の重み付けの λ は構造の生成には影響するが、最終的な構造の選択には直接影響を及ぼさないことである。ポテンシャル最小化の方法は極めて多くの方法が提唱されているが、本研究ではアルゴリズムが単純で性能も良いとされるシミュレーテッドアニーリング法[25] を用いる。

8.4.2 「タンパク質らしい」構造の条件

結晶構造が満たしている「タンパク質らしい」構造の条件を設定は、ある程度、主観的になるのはやむを得ない。本研究では自己排除性・コンパクトさ・2次構造の3つが「タンパク質らしさ」であると仮定した。それぞれの条件の詳細を以下に述べる。

1. 自己排除性: 衝突している原子対が N_{ster}^{max} 個を越えないこと。3 Å 以内にある C^{α} 原子対の数 N_{α}^{ster}、C^{β} 原子対の数 N_{β}^{ster} の和を N_{ster} としたとき

$$N_{ster} = N_{ster}^{\alpha} + N_{ster}^{\beta} \leq N_{ster}^{max} \tag{8.10}$$

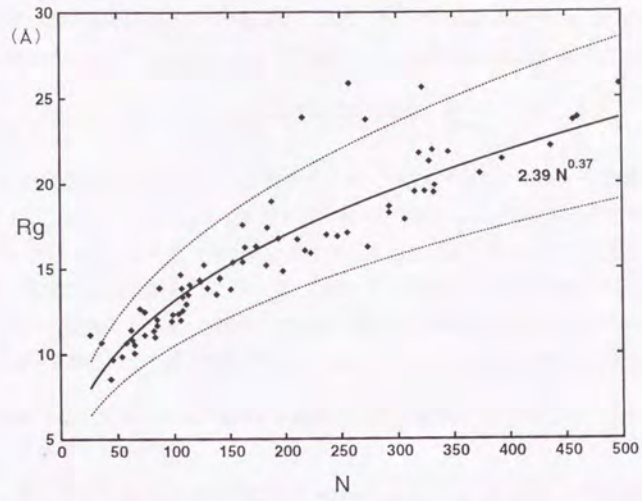


図 8.6　結晶構造の残基数 N と慣性半径 R_g の分布。最小 2 乗法で対数フィッティングした曲線を実線で示す。上限・下限を点線で示す。75 個中 70 個のタンパク質がこの範囲に含まれる。

結晶構造データベースでは衝突している原子対は一つも存在しないため、最大許容衝突数 N_{ster}^{max} は 0 とすることが望ましいが、本研究で採用した二面角を離散化したモデルにおいては、残基の距離関係の表現が制約されるため、ある程度衝突を許さないと、探索効率が著しく落ちてしまう。実際に N_{ster}^{max} をどう設定するかは、次章で述べる。

2. コンパクトさ: 球状タンパク質は、コンパクトな構造に折り畳まれている。実際の X 線結晶構造の慣性半径 R_g を残基数 N で log-log でフィットすると $R_g^{exp} = 2.39N^{0.37}$ となる（図 8.6）。この R_g^{exp} とのずれの比 $D_g = (R_g - R_g^{exp})/R_g^{exp}$ の絶対値が、D_g^{max} 以下であることをコンパクトさの条件とする。

$$|D_g| = \left| \frac{R_g - R_g^{exp}}{R_g^{exp}} \right| < D_g^{max} \tag{8.11}$$

結晶構造データベースのほとんどが含まれるように、$D_g^{max} = 0.2$ に設定する。図 8.6 に示すように 75 個中 70 個のタンパク質がこの範囲に含まれる。この範囲に含まれないタンパク質は 1PPT, 2FB4L, 1PRCL, 1PRCM, 2FB4L の 5 つである。

3. 2次構造の形成: 主鎖の規則的な水素結合パターンによる2次構造が多く形成されることもタンパク質の特徴の一つであるといえる。そこで、全体の R_{sec}^{min} 以上の残基が定常2次構造（α ヘリックスか β シート）であることを2次構造の形成の条件とする。

$$R_{sec} = \frac{(N_{helix} + N_{sheet})}{N} \geq R_{sec}^{min} \tag{8.12}$$

2次構造の定義の方法として水素結合のパターンをもとにしたDSSP[76]が最も一般的である。しかし、この方法だと厳密すぎ、本研究で用いるような粗いモデルでは、特に β シートはほとんど生成できない。そこで、本研究では、粗いモデルに見あった「ゆるい」定義を新たに設定して用いる。DSSPと異なり (1) 二面角領域（局所構造ポテンシャルで定義した図8.5と同じ）と (2)C^{α} 原子の座標の2つだけから以下のように定義する。i 番目の C^{α} 原子の座標を $\boldsymbol{r}_i$、$\boldsymbol{R}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$、$R_{ij} = ||\boldsymbol{R}_{ij}||$ とする。

- α ヘリックス：4つの連続する残基が二面角領域'A'であるとき、その4残基を α ヘリックスとする。
- β シート：以下の条件を満たす6つ $(i-1, i, i+1, j-1, j, j+1)$ の残基を β シートとする（図8.7参照）。
  (a) **二面角**: 6つの残基が全て二面角領域'B'である。
  (b) **距離**: 平行条件（$R_{i-1,j-1}, R_{i,j}, R_{i+1,j+1}$ が全て r_c 以下）か
  逆平行条件（$R_{i-1,j+1}, R_{i,j}, R_{i+1,j-1}$ が全て r_c 以下）を満たす。
  (c) **角度**: $\boldsymbol{b}_k = (\boldsymbol{R}_{k+1} - \boldsymbol{R}_k)/\|\boldsymbol{R}_{k+1} - \boldsymbol{R}_k\|$ としたとき、$|\boldsymbol{b}_i \cdot \boldsymbol{b}_j| > \theta_c$ である。

ここで、β シートの設定に必要なパラメータは 距離のしきい値 R_c と内積のしきい値 θ_c の2つである。X線結晶構造のデータベースでDSSPで β シートのペアであると判定された残基対の距離 R_{ij} と内積 $\boldsymbol{b}_i \cdot \boldsymbol{b}_j$ の分布を図8.8に示した。このヒストグラムから判断して $R_c = 6.0$ Å, $\theta_c = 0.5$ と設定する。

表8.5に、本研究による定義とDSSPによる定義の結晶構造データベースの残基数の比較を示した。β シートの一致はあまり良くないが、全体の89.1 %の残基で新定義とDSSPによる定義が一致している。図8.9に結晶構造データベースの定常2次構造比 R_{sec} と残基数 N の分布を示す。この分布から $R_{sec}^{min} = 0.2$ に設定することにする。この範囲で75個中71個のタンパク質が含まれる。この範囲に含まれないタンパク質は7WGAA1, 2ABXA, 1HIP-, 2GN5-の4つである。

表 8.5　本研究の定義による 2 次構造 (H,E,C) と DSSP による 2 次構造 (Hdssp,Edssp,Cdssp) の比較。数字はデータベースの残基数。

	H	E	C	合計
Hdssp	4192	0	109	4301
Edssp	0	2248	655	2903
Cdssp	430	267	6523	7220
合計	4622	2515	7287	14424

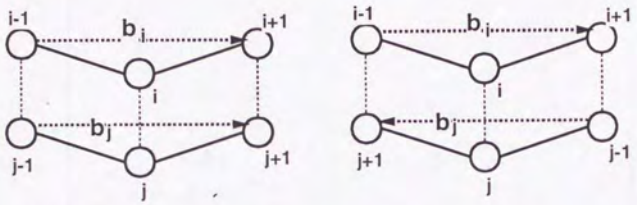


図 8.7 β シートの定義。左図は平行、右図は逆平行。

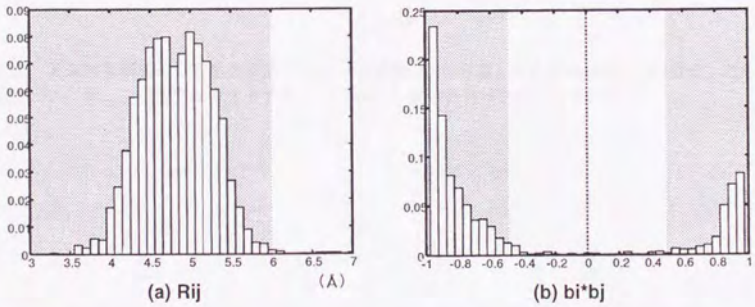


図 8.8　結晶構造データベースの中で DSSP で β シートと判定された残基対の (a) 距離 (b) 内積の分布

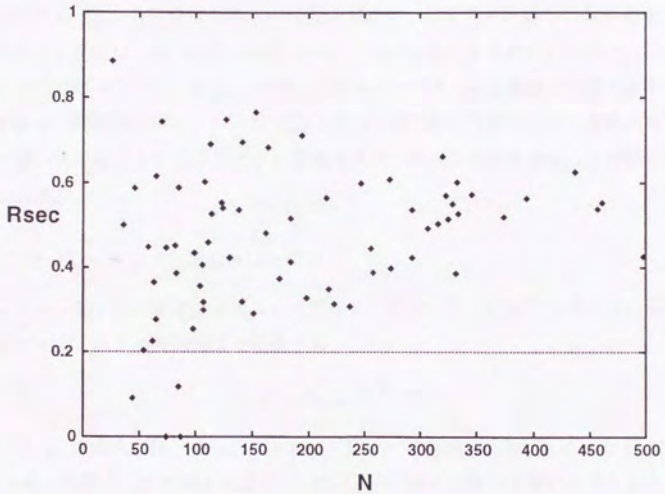


図 8.9　X 線結晶構造の定常 2 次構造比 R_{sec} と残基数 N の分布。2 次構造は本研究で設定した条件で定義した。$R_{sec} \geq R_{sec}^{min} = 0.2$ とすると 75 個中 71 個のタンパク質が含まれる。

8.4.3 「タンパク質らしい」構造のためのポテンシャル

設定した「タンパク質らしい」構造の条件をもとに、この条件を満たすほど低くなるような「タンパク質らしい」構造のためのポテンシャル E_{pro} を以下のように設定する。

$$E_{pro} = N^{\alpha}_{ster} + N^{\beta}_{ster} + |R_g - R^{exp}_g| - N_{helix} - N_{sheet} + E_{reg} \tag{8.13}$$

E_{reg} は2次構造形成の促進のため加えた二面角領域の連続性を表現したポテンシャルである。

$$E_{reg} = \sum_{i}^{N-1} -\log[P(S^{\phi\psi}_{j+1} = s_{i+1}/S^{\phi\psi}_{j} = s_i)] \tag{8.14}$$

ポテンシャル E_{pro} の設定には多くの任意性があり、本研究で設定した以外のポテンシャルもあり得る。例えば、各項の重み付けなどに工夫の余地があるかもしれない。しかし、前述したようにこのポテンシャル E_{pro} は単に「タンパク質らしい」構造を生成するために加えるものであり、配列依存ポテンシャル E_{seq} と異なり値自体に予測としての意味はない。よって、やや粗い決め方でも、探索の効率に影響は及ぼすが、予測結果自体には影響を及ぼさないと思われる。

8.4.4 シミュレーテッド・アニーリング法

ポテンシャル最小化計算は、シミュレーテッド・アニーリング法[25]を用いる。温度降下のスケジュールは、以下の指数関数で決定する。

$$T = T_{max} \times r^{N_{step}} \tag{8.15}$$

ここで、T_{max} は最大温度、N_{step} はステップ数、r は任意の定数 $(0 < r < 1)$ である。r は 最大ステップ数 N^{max}_{step} と最小温度 T_{min} から以下の式で逆算して用いることとする。

$$r = \exp\left[\frac{1}{N^{max}_{step}} \log \frac{T_{min}}{T_{max}}\right] \tag{8.16}$$

全体のアルゴリズムの流れは以下のようになる。ここで、i 残基目の 構造パラメータ（$(\phi\psi)$ の代表点）を x_i、全残基の構造パラメータのベクトルを $\boldsymbol{x}$、$\boldsymbol{x}$ から変換された3次構造のポテンシャルを $E(\boldsymbol{x})$ とする。

1. **初期化：**ランダムに $\boldsymbol{x}$ を設定。温度 $T := T_{max}$、カウンタ $n := 0$ に設定。$E(\boldsymbol{x})$ を計算。

2. **構造の変移：** $\hat{\boldsymbol{x}} := \boldsymbol{x}$ とする。残基番号 i をランダムに選び、i 残基目の配座 $\hat{x}_i$ をランダムに変更する。アミノ酸ごとに代表点の数が異なるので（表8.4）、i を選ぶとき代表点の数に比例して選ばれるようにする。

3. ポテンシャル値の評価： $E(\hat{\boldsymbol{x}})$ を評価。$\Delta E := E(\hat{\boldsymbol{x}}) - E(\boldsymbol{x})$ とする。

4. 判定：ランダムに $P \in [0:1)$ を発生。$\Delta E < 0.0$ か $\exp(-\Delta E/T) > P$ なら $\boldsymbol{x} := \hat{\boldsymbol{x}}$

5. 温度変化： $T := T \times r, n := n + 1$ とする。

6. $n < MAX$ ならステップ 2 に行く。

8.5 使用する構造類似度

予測構造の評価のために以下の構造類似度を用いる。

1. 平均 2 乗誤差： RMS (root mean square deviation)

$$RMS = \left(\frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{r}_i^{\alpha} - \hat{\boldsymbol{r}}_i^{\alpha})^2 \right)^{1/2} \tag{8.17}$$

$\boldsymbol{r}_i^{\alpha}$ は i 番目の C^{α} 原子の位置ベクトルである。$\boldsymbol{r}^{\alpha}$ と $\hat{\boldsymbol{r}}^{\alpha}$ は、重心を合わせ最も 2 つの座標の相関が高くなるように回転を施してから、式 8.17の計算を行なう。最適な回転は特異値分解により解析的に導出される[77]。RMS 値は最も良く用いられる構造の類似度であるが、類似性の低い構造ペアの比較にはあまり向いていない。

2. 距離平均 2 乗誤差： DRMS(distance root mean square deviation)

$$DRMS = \left(\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (r_{ij}^{\alpha} - \hat{r}_{ij}^{\alpha})^2 \right)^{1/2} \tag{8.18}$$

r_{ij}^{α} は、i 番目の j 番目の C^{α} 原子間の距離である。DRMS は RMS ほど頻繁に用いられる指標ではないが、RMS と似た性質を持つ量である。RMS に比べて、特異値分解を用いないためアルゴリズムが単純で計算量も少ないという利点がある。キラリティの異なる構造対を区別できないことが欠点である。計算量の少なさから次章で説明する Build-up 法における距離として用いる。

3. 2 次構造一致率： Q_{sec}

$$Q_{sec} = \frac{1}{N-2} \sum_{i=2}^{N-1} \delta(S_i, \hat{S}_i) \times 100 \tag{8.19}$$

S_i は、i 番目の 2 次構造 ($S_i \in \{H, E, C\}$)、$\delta(a,b)$ は a,b が一致したときのみ 1 を返す関数である。2 次構造は本研究の定義を用いる。1 番目と N 残基目を比較しない

のは、2次構造の定義に必要な (ϕ,ψ) が計算できないからである。この量は第I部で用いた Q_3 とほぼ対応する。

4. (ϕ,ψ) 状態一致率： $Q_{\phi\psi}$

$$Q_{\phi\psi} = \frac{1}{N-2}\sum_{i=2}^{N-1}\delta(S_i^{\phi\psi},\hat{S}_i^{\phi\psi}) \times 100 \tag{8.20}$$

$S_i^{\phi\psi}$ は、i 番目の3状態の (ϕ,ψ) 状態 ($S_i^{\phi\psi} \in \{A,B,C\}$) である。$Q_{\phi\psi}$ は Q_{sec} とよく似た値になる。$S_i^{\phi\psi}$ は 本研究で 採用している構造パラメータ、すなわち離散化された (ϕ,ψ) に 近い量であることから、$Q_{\phi\psi}$ は、Q_{sec} よりも構造パラメータにおける類似度に近い。

第 9 章

結果

9.1 離散 (ϕ,ψ) モデルの表現性

本研究で採用した離散 (ϕ,ψ) モデルは、理想的な結合長・結合角に固定し、(ϕ,ψ) をいくつかの代表点に限定しているため、完全に N 構造と同じ主鎖構造を再現することはできない。そこで、このモデルで表現できる構造の中で最も N 構造に近い構造を計算し、この構造と N 構造と類似度から、構造表現の精度を評価することとする。本研究ではモデルで表現できる N 構造に最も近い構造を Nnear 構造と呼ぶこととする。

RMS など大域的な構造類似度が最も小さい Nnear 構造を得る探索は、それ自体難しい問題で、いくつかの方法が発表されているが[78,65]、全回探索を行なう以外に完全最適解を得る方法はない。本研究では Build-up 法[65] と呼ばれる近似解法を採用する。Build-up 法は分枝限定法 (Branch-and-bound method) の一種で、1 残基目から順々に N 構造に近い構造を N_{keep} 通りの構造を残しながら 組み立てていく方法である。全体の構造を表現する 1 残基から i 残基目までの (ϕ,ψ) 代表点のベクトルを $\boldsymbol{x}^i$、i 残基目の (ϕ,ψ) 代表点を x_i とする。また、$\boldsymbol{x}^i$ を変換した i 残基目までの構造を $\boldsymbol{c}^i(\boldsymbol{x}^i)$ とし、i 残基目までの N 構造を $\boldsymbol{c}^i_{nat}$ とする。このとき、Build-up 法のアルゴリズムは以下のようになる。

1. $i=0$ とする。
2. N_{keep} 個 の $\boldsymbol{x}^i$ に対して、全ての可能な x_{i+1} を加えた、$\boldsymbol{x}^{i+1}$ を生成し、距離 $D(\boldsymbol{c}^{i+1}(\boldsymbol{x}^{i+1}),\boldsymbol{c}^{i+1}_{nat})$ を計算する。
3. 距離 D の小さな順に N_{keep} 個 の $\boldsymbol{x}^{i+1}$ を選択する。
4. $i <$ 残基数 なら、$i := i+1$ として、2 に戻る。

この方法で得られる構造は、距離として $Q_{\phi\psi}$ のように完全に最適性原理が成り立つ距離 D を用いた場合にのみ最適解となり、RMS などの大域的な類似度を用いた場合は近似解となる。

本研究では、距離 D として比較的計算量の少ない DRMS（式 8.18）を用いる。さらに次の 2 つの操作を行なう。(1) C^{α} 原子対と C^{β} 原子対の衝突数の合計 N_{ster} が 2 を越えた構造に対しては $D = D_{max}$（D_{max} は非常に大きな値）とする。(2) キラリティの異なる構造を排除するため、N 構造と二面角 ϕ の符号が異なる構造に対しても $D = D_{max}$ とする。N_{keep} を 500 として、5 つのタンパク質に対して Nnear 構造を作成した。Set6 の場合の N 構造の類似度を表 9.1 に、Set12 の場合を表 9.2 に示した。Set12 の場合のほうが、RMS と DRMS についてかなり良い値になっている。図 9.1 に 3ICB の場合の Set6 と Set12 の Nnear 構造と N 構造を示した。Set12 の場合、RMS = 1.5 Å 程度、2 次構造一致率 $Q_{sec} = 80\%$ 程度の値の Nnear 構造を構築することができ、構造表現として十分な精度を持っていることが示された。本研究では、以後 Set12 の代表点による構造表現を採用することとする。

表 9.1 Build-up 法で作成した Nnear 構造と N 構造の類似度。Set6 の場合。

PDBcode	RMS	DRMS	Q_{sec}	$Q_{\psi\psi}$
4RXN	2.66	1.97	78.8	71.2
5PTI	7.66	2.66	60.7	69.6
2CRO	7.46	5.13	54.0	63.5
3ICB	3.19	2.34	54.8	67.1
1UBQ	3.03	2.26	85.1	75.7

表 9.2 Build-up 法で作成した Nnear 構造と N 構造の類似度。Set12 の場合。

PDBcode	RMS	DRMS	Q_{sec}	$Q_{\psi\psi}$
4RXN	1.66	1.28	78.8	75.0
5PTI	1.45	1.00	85.7	83.9
2CRO	1.02	0.81	92.1	90.5
3ICB	1.70	1.23	87.7	89.0
1UBQ	1.23	0.87	75.7	85.1

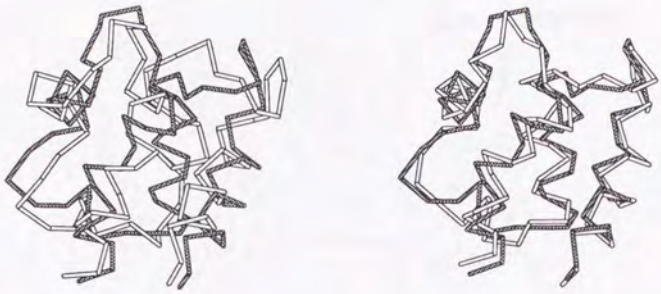


図 9.1　3ICB の N 構造（黒）と Build-up 法で作成した Nnear 構造（白）。左は Set6 の二面角代表点、右は Set12 の二面角代表点を用いた場合。

9.2　Threading テストによる E_{seq} の評価

9.2.1　Threading テストの手続き

以下に Threading テストの手続きを示す。図 9.2にこのテストの概念図を示した。s は配列、c は構造、p はタンパク質（配列 s と構造 c の組）、$E(c,s)$ は構造 c に配列 s をのせたときのポテンシャルの値である。

- 次の処理を結晶構造データベース内の全てのタンパク質 p_i について行なう。

  1. p_i を除いた全てのタンパク質 $\forall p \neq p_i$ からポテンシャル E のパラメータを統計的に抽出する。

  2. 抽出したパラメータを用いて、$E(c,s_i)$ をデータベース内の全て c_i 以外の構造 $\forall c \neq c_i$ に対して計算する。このとき、p_i より残基数が多いタンパク質の構造 c は部分構造に分解して、複数の構造として扱う。平均 $\bar{E}_i$、分散 $\sigma^2_{E_i}$ を求める。また、$E_i^{nat} = E(c_i,s_i)$ を計算する。

各タンパク質に対する評価量としては以下に定義する $Rank_i^{nat}$ と Z_i^{nat} の 2 つの量を用いる。

1. $Rank_i^{nat}$ ： N 構造の順位。すなわち $E_i^{nat} > E(c_j,s_j)$ である構造 c_j の数 $N_{E<E^{nat}}$ に 1 を加えた数。

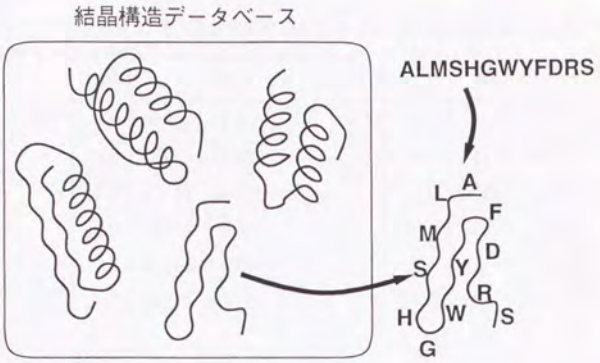


図 9.2 Threading テストの概念図

2. Z_i^{nat}：正規化した N 構造のポテンシャル値。$Z_i^{nat} = (E_i^{nat} - \bar{E}_i)/\sigma_{E_i}$

結晶データベース内の全タンパク質に対する評価は、$Rank_i^{nat}$, Z_i^{nat} を統合した N_{1st}^{nat}、$\bar{Z}^{nat}$ を用いる。

1. N_{1st}^{nat}：$Rank_i^{nat} = 1$ であるタンパク質の数

2. $\bar{Z}^{nat}$：Z_i^{nat} のタンパク質平均。

ここで、本研究で用いた結晶構造データベース内の最大残基数のタンパク質 8CATA は、データベース内に比較する構造が存在しないので、N_{1st}^{nat} と $\bar{Z}^{nat}$ の計算には入れないこととする。

9.2.2 Threading テスト結果

ポテンシャルのパラメータ決定

コンタクトポテンシャルのコンタクトのしきい値 R_c、距離ヒストグラムポテンシャルの遠距離のしきい値 R_{max}、局所構造ポテンシャルのウィンドウ数 $2M+1$ の 3 つのパラメータを決定するために、これら 3 つのポテンシャルのパラメータを変えた Threading 結果を示したのが、表 9.3 と表 9.4である。これらの表から、最も成績の良い、$R_c = 8.0$ Å、$R_{max} = 8.0$ Å、$2M+1 = 3$ 残基を用いることとする。

表 9.5 に 第 I 部で用いたいくつかの符号化関数 f に対する 2 元コンタクトポテンシャルの Threading 結果を示す。表の最下段にある Optimal は BW-MGOR 法による 2 次構造予測に

表 9.3　ポテンシャル EC の R_c を変化させた場合の Threading 結果（左）とポテンシャル ED の R_{max} を変化させた場合の Threading 結果（右）

	R_c	N_{1st}^{nat}	$\bar{Z}^{nat}$
EC	5.0	28/74	-3.02
	6.0	51/74	-4.61
	7.0	58/74	-5.52
	8.0	58/74	-5.80
	9.0	57/74	-5.55
	10.0	54/74	-4.93

	R_{max}	N_{1st}^{nat}	$\bar{Z}^{nat}$
ED	7.0	64/74	-7.75
	8.0	67/74	-8.27
	9.0	67/74	-8.08
	10.0	65/74	-7.50
	11.0	64/74	-7.25
	12.0	64/74	-6.99

最適な符号化関数として得られたものである。Threading 結果の各符号化関数の違いは、2 次構造予測とよく似ており、非極性 (Nonpolar) と最適 (Optimal) の 2 つの成績がよい。

単独で用いた場合と組み合わせて用いた場合の Threading テスト結果

表 9.6に各ポテンシャルの Threading テストの結果を示す。単独で最も成績が良いのは、距離ヒストグラムポテンシャル (ED) であり $N_{1st}^{nat} = 67, \bar{Z}^{nat} = -8.27$ である。コンタクト数ポテンシャルが最も成績が悪い。2 種のポテンシャルを組み合わせた場合では、N_{1st}^{nat} で比較すると、EC+EL と ED+EL の値が同じ 最高の 71 となり、$\bar{Z}^{nat}$ の値で比較すると ED+EL が -7.85 で最もよい。この結果から、本研究では E_{seq} として、距離ヒストグラムポテンシャルと局所構造ポテンシャルの和 $E_{seq} = ED + EL$ を採用して、ポテンシャル最小化計算を行なうこととする。

同じデータを用いた Kocher ら[75] は、仮想側鎖中心の距離ポテンシャルと二面角領域のポテンシャルを組み合わせたときに $N_{1st}^{nat} = 74/74$ を達成している。本研究での結果は彼らの精度よりやや劣るが、十分比肩する精度であると思われる。

表 9.7に各タンパク質に対する $Rank_i^{nat}$ と Z_i^{nat} をそれぞれのポテンシャルに対して示した。これを見ると成績の悪いタンパク質は比較的共通している。2MLTA、1PPTA の 2 つの小さなタンパク質と、定常 2 次構造比の低い 2GN5- はどのポテンシャルでも共通して成績が悪い。

9.3　「タンパク質らしい」構造群（デコイ構造群）の生成

「タンパク質らしい」構造のためのポテンシャル E_{pro}（式 8.13）が、「タンパク質らしい」構造を生成するために有効であること検証するために、E_{pro} のみに対してポテンシャル

表 9.4 ポテンシャル EL の ウィンドウ数 $2M+1$ を変化させた場合の Threading 結果

	$2M+1$	N_{1st}^{nat}	$\bar{Z}^{nat}$
EL	1	54/74	-4.84
	3	57/74	-5.02
	5	55/74	-4.62
	7	57/74	-4.35
	9	53/74	-4.13
	11	53/74	-3.96

表 9.5 2元コンタクトポテンシャル EB の符号化関数による Threading 結果の違い

EB	f	AILMFPVRDEKNCQHSTWYG	N_{1st}^{nat}	$\bar{Z}^{nat}$
	Nonpolar	11111110000000000000	34 / 74	-3.30
	Charged	00000001111000000000	18 / 74	-2.66
	Polar	00000000000111111110	2 / 74	-0.53
	Positive	00000001001000000000	12 / 74	-1.39
	Negative	00000000110000000000	14 / 74	-1.97
	Weight	00011001011001100110	8 / 74	-0.76
	Aromatic	00001000000000100110	4 / 74	-1.30
	Aliphatic	11100010000000000001	24 / 74	-2.54
	Optimal	01111010000010000010	50 / 74	-5.24

表 9.6 各ポテンシャルを組み合わせた場合の Threading の結果

EC	ED	EL	EN	EB	N_{1st}^{nat}	$\bar{Z}^{nat}$
○	-	-	-	-	58/74	-5.80
-	○	-	-	-	67/74	-8.27
-	-	○	-	-	57/74	-5.02
-	-	-	○	-	46/74	-4.65
-	-	-	-	○	50/74	-5.24
○	-	○	-	-	71/74	-7.33
-	○	○	-	-	71/74	-7.85
○	-	-	○	-	61/74	-6.07
-	○	-	○	-	67/74	-7.97
-	-	○	○	-	63/74	-5.89
-	-	○	-	○	67/74	-6.70
-	-	-	○	○	54/74	-5.56

最小化計算を行ない多数の最小構造群を生成した。またその結果から、方法の章で未決定とした自己排除性の条件 $N_{ster} \leq N_{ster}^{max}$ の N_{ster}^{max} の見積もりも行なう。

10,000 ステップ $T_{min} = 10.0, T_{max} = 0.001$ のシミュレーテッド・アニーリング法を用い、異なる初期構造を用いて 4000 個の構造を生成した。図 9.3にこの構造群の N_{ster}, D_g, R_{sec} の頻度分布を示す。

ヒストグラムから、自己排除性の条件 の N_{ster}^{max} はヒストグラムのピークである $N_{ster} = 2$ に設定することとする。$N_{ster} \leq 2$ を満たす構造数は 1612 個 (40.3 %) となる。また、コンパクトさの条件（ $|D_g| \leq 0.2$ ）を満たす構造は 3332 個 (83.3 %)、2 次構造の形成の条件（ $R_{sec} \geq 0.2$）は 3999 個 (99.9 %) であり、この 2 つの条件はほとんどの構造が満たしている。3 つの条件全てを満たすタンパク質は 1354 個 (33.9 %) となった。この 1354 個の「タンパク質らしい」構造の条件を満たす構造群を「デコイ構造群」と呼ぶこととする。

また、3 つの条件以外の「タンパク質らしさ」として、α ヘリックスと β シートの組成比を表 9.8に示した。結晶構造群とデコイ構造群の組成比はかなりよく似ている。図 9.4に各タンパク質の α ヘリックス残基数比 N_{helix}/N と β シート残基数比 N_{sheet}/N によるプロットを示した。また、生成されたデコイ構造群の中からいくつかをリボン図で示した（ 図 9.5 ）。リボン図の描画には MOLSCRIPT を用いた[79] 。

表 9.7 各タンパク質に対する $Rank_i^{nat}$ と Z_i^{nat}

PDBcode	残基数	N_{comp}	$Rank_i^{nat}$ C	D	L	C+L	D+L	Z_i^{nat} C	D	L	C+L	D+L
3GRS-	461	39	1	1	1	1	1	-5.5	-5.7	-5.0	-5.5	-5.6
2YHX-	457	48	1	1	1	1	1	-4.7	-5.1	-5.0	-5.9	-6.0
2CTS-	437	109	1	1	1	1	1	-7.3	-7.5	-7.3	-8.2	-8.1
1PHH-	394	282	1	1	1	1	1	-8.7	-9.7	-6.1	-9.8	-8.9
8ADH-	374	383	1	1	1	1	1	-10.3	-11.6	-7.2	-11.2	-11.2
2LBP-	346	552	1	1	1	1	1	-8.7	-11.3	-6.9	-10.1	-10.6
1GD1O	334	637	1	1	1	1	1	-9.2	-12.2	-8.3	-11.3	-12.1
4MDHA	333	646	1	1	1	1	1	-9.5	-13.1	-6.8	-11.6	-11.8
1PRCC	332	656	1	1	1	1	1	-4.7	-4.7	-6.0	-7.2	-7.0
6LDH-	329	687	1	1	1	1	1	-9.1	-10.9	-6.4	-10.8	-10.8
2APR-	325	732	1	1	1	1	1	-8.9	-11.3	-8.5	-11.5	-11.9
1PRCM	323	757	4	5	1	1	1	-2.9	-2.8	-6.6	-6.4	-6.5
1PFKA	320	797	1	1	1	1	1	-10.3	-13.1	-8.0	-12.2	-12.3
3TLN-	316	854	1	1	1	1	1	-3.6	-7.7	-7.3	-7.6	-9.6
5CPA-	307	990	1	1	1	1	1	-7.5	-11.4	-6.3	-9.4	-10.4
1RHD-	293	1216	1	1	1	1	1	-9.1	-11.2	-7.7	-11.0	-11.1
2CYP-	293	1216	1	1	1	1	1	-5.6	-8.7	-6.3	-7.9	-9.1
1CSEE	274	1559	1	1	1	1	1	-8.6	-14.2	-6.7	-10.2	-12.1
1PRCL	273	1579	37	12	1	1	1	-2.0	-2.6	-5.0	-4.7	-5.2
1PRCH	258	1880	1	1	1	1	1	-5.5	-7.0	-6.5	-7.9	-8.5
2CAB-	256	1923	1	1	1	1	1	-6.5	-10.1	-7.6	-9.6	-10.8
1TIMA	247	2122	1	1	1	1	1	-8.0	-10.3	-6.5	-10.1	-10.4
2CNA-	237	2353	1	1	1	1	1	-5.3	-8.6	-4.7	-6.8	-8.0
1TPP-	223	2690	1	1	1	1	1	-7.1	-11.3	-6.4	-9.1	-10.4
2ACT-	218	2816	1	1	1	1	1	-6.2	-11.5	-8.2	-9.9	-11.9
2FB4L	216	2869	1	1	1	1	1	-7.0	-10.1	-7.2	-9.6	-10.4
2TS1-1	211	3005	1	1	1	1	1	-5.5	-7.5	-4.6	-7.0	-7.2
2ALP-	198	3370	1	1	1	1	1	-6.8	-10.6	-7.6	-9.9	-11.1
3ADK-	194	3487	1	1	1	1	1	-5.9	-9.9	-5.7	-8.0	-9.2
1GOX-1	188	3668	1	1	1	1	1	-4.2	-4.3	-5.3	-6.6	-6.5
2STV-	184	3793	1	1	1	1	1	-4.1	-5.4	-5.2	-6.2	-6.7
1GP1A	183	3826	1	1	1	1	1	-6.5	-9.0	-6.0	-8.5	-9.1
1GCR-	174	4124	1	1	1	1	1	-7.9	-9.7	-5.1	-9.3	-9.3
2LZM-	164	4465	1	1	1	1	1	-7.2	-10.9	-4.8	-8.1	-8.5
3DFR-	162	4537	1	1	1	1	1	-7.3	-9.3	-5.3	-8.5	-8.7
1GOX-2	162	4537	1	1	1	1	1	-7.1	-8.6	-5.9	-8.6	-8.8
2LH4-	153	4871	1	1	1	1	1	-5.0	-7.0	-4.9	-6.6	-7.2

N_{comp}: 比較構造数、C: コンタクトポテンシャル、D: 距離ヒストグラムポテンシャル、L: 局所構造ポテンシャル

表 9.7 各タンパク質に対する $Rank_i^{nat}$ と Z_i^{nat} （続き）

			$Rank_i^{nat}$					Z_i^{nat}				
PDBcode	残基数	N_{comp}	C	D	L	C+L	D+L	C	D	L	C+L	D+L
2SODO	151	4948	1	1	1	1	1	-6.3	-11.4	-6.7	-8.9	-10.9
2SNS-	141	5340	1	1	44	1	1	-4.7	-9.3	-2.4	-5.0	-5.8
4HHBA	141	5340	1	1	1	1	1	-4.2	-6.8	-3.5	-5.3	-6.3
4FXN-	138	5464	1	1	1	1	1	-7.3	-11.8	-4.6	-8.6	-9.5
1LZ1-	130	5801	1	1	1	1	1	-7.0	-11.2	-4.5	-7.8	-8.8
2CCYA	127	5931	2	1	1	1	1	-3.4	-5.4	-3.7	-5.0	-5.3
7RSA-	124	6064	1	1	6	1	1	-6.5	-11.9	-3.4	-7.1	-8.8
1BP2-	123	6110	1	1	1	1	1	-6.0	-11.7	-4.6	-7.5	-10.1
2PABA	114	6525	1	1	1	1	1	-4.3	-4.6	-4.7	-6.1	-6.2
1HMZA	113	6573	4	3	1	1	1	-3.1	-3.8	-3.5	-4.6	-4.5
1CCR-	111	6670	15	1	1	1	1	-2.6	-4.8	-5.2	-5.4	-6.5
1ACX-	108	6819	1	1	1	1	1	-6.4	-9.8	-5.0	-7.3	-8.2
4CPV-	108	6819	1	1	6	1	1	-5.1	-7.2	-2.9	-5.3	-5.2
2CDV-	107	6871	50	74	1	1	1	-2.4	-2.4	-3.5	-4.2	-4.3
1FD2-	106	6925	1	1	1	1	1	-7.2	-8.3	-3.9	-8.1	-7.6
2TS1-2	106	6925	2	1	1	1	1	-3.3	-5.5	-4.4	-5.6	-5.9
1RNT-	104	7034	1	1	1	1	1	-6.8	-8.9	-5.9	-8.6	-8.4
1PCY-	99	7310	1	1	1	1	1	-6.5	-8.8	-5.1	-7.8	-8.2
3FXC-	98	7367	1	1	9	1	1	-7.3	-8.1	-3.0	-7.7	-6.7
2GN5-	87	7995	9	1	3484	159	167	-3.2	-4.4	-0.1	-2.1	-2.1
1HIP-	85	8113	1	1	29	1	1	-5.0	-5.7	-2.6	-5.3	-4.8
3B5C-	85	8113	1	1	1	1	1	-4.4	-6.4	-6.0	-7.5	-8.0
1CC5-	83	8234	37	1	1	1	1	-2.5	-5.3	-5.7	-5.9	-7.3
451C-	82	8296	13	1	1	1	1	-2.8	-4.3	-4.0	-4.8	-5.3
1HOE-	74	8794	75	1	8	1	1	-2.5	-6.4	-3.5	-4.2	-5.7
2ABXA	74	8794	1	1	3895	1	1	-6.0	-8.7	-0.1	-4.0	-3.8
1UTG-	70	9051	79	3	5	1	1	-2.3	-3.9	-3.0	-3.7	-4.3
1SN3-	65	9378	1	1	3	1	1	-7.8	-12.4	-3.3	-7.2	-8.0
2CRO-	65	9378	27	1	2	1	1	-2.7	-7.0	-3.5	-4.6	-5.6
1CSEI	63	9513	3	1	1	1	1	-4.2	-6.2	-3.7	-5.0	-5.5
1NXB-	62	9582	1	1	6	1	1	-7.5	-11.8	-3.0	-6.8	-7.2
5PTI-	58	9859	1	1	3	1	1	-5.7	-9.9	-3.5	-6.2	-7.2
4RXN-	54	10140	1	1	1	1	1	-4.4	-6.7	-3.7	-6.1	-6.4
1CRN-	46	10709	1	1	68	1	1	-6.1	-8.5	-2.4	-5.5	-5.4
7WGAA1	44	10854	1	1	3	1	1	-9.9	-12.1	-5.0	-11.1	-10.2
1PPT-	36	11439	877	19	241	147	21	-1.4	-2.9	-1.9	-2.1	-2.6
2MLTA	26	12180	1490	225	2781	1145	493	-1.1	-2.1	-0.7	-1.3	-1.7

N_{comp}: 比較構造数、C: コンタクトポテンシャル、D: 距離ヒストグラムポテンシャル、L: 局所構造ポテンシャル

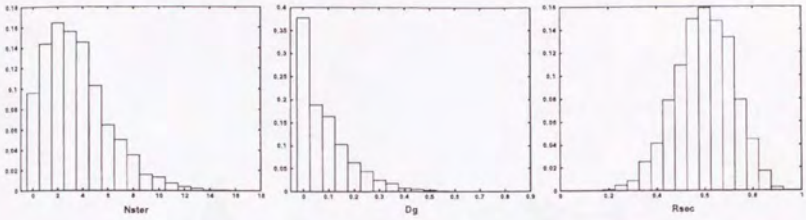


図 9.3　ポテンシャル E_{pro} の最小化で得られた4000個の構造群の N_{ster}（左）、D_g（中）、R_{sec}（右）の頻度ヒストグラム。

表 9.8 結晶構造群とデコイ構造群の2次構造の比

	Helix	Sheet	Coil
結晶構造群	0.34	0.21	0.45
デコイ構造群	0.39	0.25	0.36

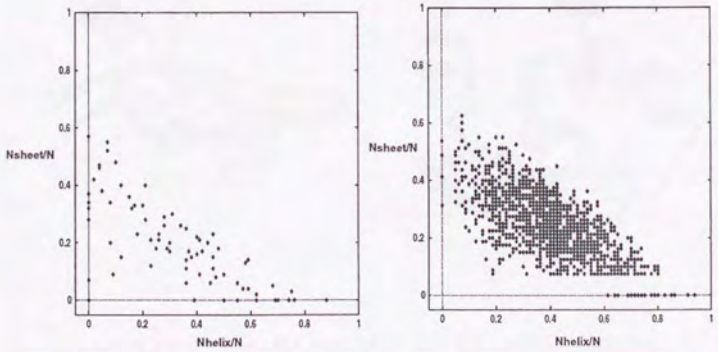


図 9.4　各タンパク質の α ヘリックス残基数比 N_{helix}/N と β シート残基数比 N_{sheet}/N によるプロット。左図は結晶構造データベース、右図は生成した「タンパク質らしい」構造群。

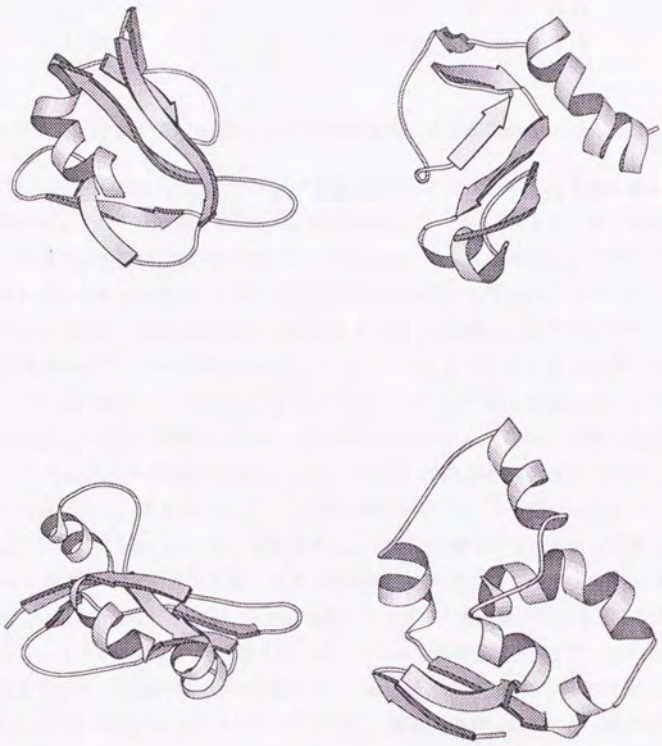


図 9.5 生成した「タンパク質らしい」構造の条件を満たすデコイ構造群の例。

表 9.9 5つのタンパク質のN構造のポテンシャル値

PDBcode	残基数	N_{helix}	N_{sheet}	E_{seq}^{nat}	ED^{nat}	EL^{nat}
4RXN	54	0	11	-21.86	-11.62	-10.24
5PTI	58	12	14	-29.60	-20.14	-9.46
2CRO	65	40	0	-27.63	-16.37	-11.26
3ICB	75	49	6	-35.20	-15.25	-19.94
1UBQ	76	11	27	-29.97	-20.14	-9.83

9.4 結晶構造群に対するThreadingとデコイ構造群に対するThreadingの比較

表8.2に示した5つの80残基以下のタンパク質の配列によるThreadingを結晶構造群とデコイ構造群に対して行ない結果を比較した。結晶構造群は予測を試みるタンパク質の構造は除外する。デコイ構造群はE_{pro}のポテンシャル最小化で得られた構造から「タンパク質らしい」条件を満たす構造を選択した1354個の残基数80の構造を用いた。ポテンシャルE_{seq}として、Jackknife法によるThreading結果が最も良かった距離ヒストグラムポテンシャルEDと局所構造ポテンシャルELの和$E_{seq} = ED + EL$を用いる。また、ポテンシャルのパラメータは80残基以下のタンパク質を除いたデータベースから抽出しているので、予測対象となるタンパク質は統計データから必ず除外されている。表9.11に結晶構造群、表9.12にデコイ構造群のN構造の位置と、E_{seq}の平均、N構造とのRMS, $DRMS$, $Q_{\phi\psi}$, Q_{sec}との平均を示した。表中の$N_{E<E^{nat}}$は、構造群の中でE_{seq}^{nat}より低かったポテンシャルの数、$P(E < E^{nat})$は$N_{E<E^{nat}}$を全構造数N_{comp}で割った値である。どちらの表においても$N_{E<E^{nat}} = 0$であり、結晶構造群、デコイ構造群ともN構造よりポテンシャル値の低い構造は現れていない。4RXN, 5PTI, 2CROに関してはデコイ構造数は結晶構造群の構造数の2倍近くあるにもかかわらず、N構造を凌ぐポテンシャルの構造は出現していない。また、他の平均諸量も2つの構造群ではかなり類似している。図9.6と図9.7に5PTIと3ICBの場合のポテンシャルE_{seq}のヒストグラムを示した。結晶構造群とデコイ構造群のヒストグラムは、分布の概形が非常に良く一致していることがわかる。これらの類似性はデコイ構造群がある意味で現実のタンパク質とよく似た特性を持っていることを示し、本研究で設定したE_{pro}の妥当性を間接的に示しているといえる。

表 **9.10** 5つのタンパク質の Nnear 構造のポテンシャル値

PDBcode	残基数	N_{helix}	N_{sheet}	E_{seq}	ED	EL
4RXN	54	0	0	-15.49	-7.18	-8.30
5PTI	58	10	10	-22.92	-12.51	-10.41
2CRO	65	37	0	-19.55	-6.42	-13.13
3ICB	75	48	0	-24.70	-8.39	-16.31
1UBQ	76	16	14	-19.64	-8.14	-11.50

表 **9.11** 結晶構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	10139	0	0.000	-5.93	10.60	12.40	9.33	34.00	46.61
5PTI	9858	0	0.000	-6.77	6.21	13.31	9.99	35.93	40.43
2CRO	9377	0	0.000	-5.78	5.59	12.50	9.63	39.39	42.24
3ICB	8730	0	0.000	-6.93	13.14	13.32	9.77	40.90	37.41
1UBQ	8668	0	0.000	-7.01	10.36	14.62	10.39	35.29	35.32

N_{comp}: 比較した構造数　$N_{E<E^{nat}}$: E_{seq}^{nat} よりポテンシャルの低い構造数　$P(E<E^{nat}) := N_{E<E^{nat}}/N_{comp}$

$Z^{nat} := (E_{seq}^{nat} - \bar{E}_{seq})/\sigma^{E_{seq}}$　$\bar{E}_{seq}$: 構造群内の E_{seq} の平均　$\overline{RMS}, \overline{DRMS}, \bar{Q}_{\phi\psi}, \bar{Q}_{sec}$: 構造群内の N 構造との $RMS, DRMS, Q_{\phi\psi}, Q_{sec}$ の平均

表 **9.12** デコイ構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	18151	0	0.000	-5.62	8.95	10.81	6.93	34.51	38.87
5PTI	17988	0	0.000	-6.68	5.78	12.00	7.88	39.49	37.92
2CRO	16019	0	0.000	-6.03	4.73	11.03	7.13	38.54	40.36
3ICB	7611	0	0.000	-7.99	13.78	11.92	7.39	42.01	40.32
1UBQ	6437	0	0.000	-6.88	8.29	13.03	8.10	35.36	31.91

表 9.13 結晶構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-10.55	-1.73	-8.82	12.52	6.25	42.31	48.08
5PTI	-14.50	-9.70	-4.80	10.55	8.02	41.07	55.36
2CRO	-17.38	-8.42	-8.96	10.81	7.98	47.62	58.73
3ICB	-11.10	-2.47	-8.63	13.22	7.92	41.10	35.62
1UBQ	-13.88	-8.67	-5.21	12.51	6.60	36.49	40.54

表 9.14 デコイ構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-18.13	-8.89	-9.25	9.16	5.42	48.08	32.69
5PTI	-14.52	-9.04	-5.48	10.77	7.80	50.00	46.43
2CRO	-14.01	-3.75	-10.25	10.00	6.09	57.14	65.08
3ICB	-7.63	-2.38	-5.25	8.04	6.06	65.75	69.86
1UBQ	-11.28	-2.67	-8.61	12.86	7.96	40.54	33.78

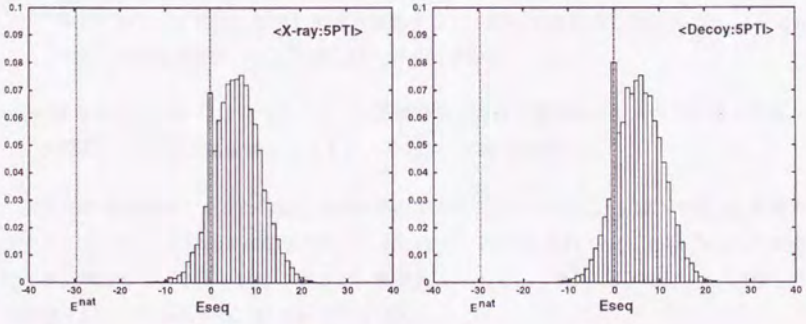


図 9.6 5PTI の場合のポテンシャル E_{seq} のヒストグラム。左が結晶構造群、右図がデコイ構造群。

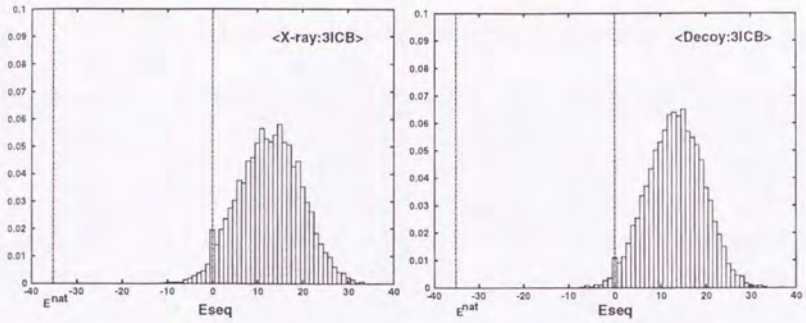


図 9.7 3ICB の場合のポテンシャル E_{seq} のヒストグラム。左が結晶構造群、右図がデコイ構造群。

9.5 5つのタンパク質に対するポテンシャル最小化計算

表 8.2に示した5つのタンパク質に対して、ポテンシャル最小化計算を $E = (1-\lambda)E_{pro} + \lambda E_{seq}$ に対して行なった。E_{seq} は、距離ヒストグラムポテンシャル ED と局所構造ポテンシャル EL の和 $E_{seq} = ED + EL$ とした。$\lambda = 0.4, 0.5, 0.6$ の3通りについてそれぞれ初期値を変えて、100 個ずつ、合計 300 個の構造を生成した。生成した 300 個の中から「タンパク質らしい」構造の条件を満たす構造を選択した構造群を、「最小化構造群」と呼ぶことにする。最小化構造群の中で最もポテンシャル E_{seq} が低い構造を予測構造とする。表 9.15に最小化構造群の統計データを示した。この表から以下のことがわかる。

- 構造数が 100 前後であるにもかかわらず、5つ全てのタンパク質について、N 構造よりポテンシャルの低い構造が相当数存在する。10000 前後の構造数でも $N_{E<E^{nat}} = 0$ であった結晶構造群、デコイ構造群と対照的である。
- 構造群の平均の $\bar{Q}_{\phi\psi}$ と $\bar{Q}_{sec}$ は、結晶構造群・デコイ構造群より大きく改善される。$\overline{RMS}$、$\overline{DRMS}$ は改善はされるが、それほど大きくはない。

表 9.16に予測構造、すなわち最小化構造群の中でポテンシャル E_{seq} が最小であった構造のポテンシャル値と N 構造との類似度を示した。また、図 9.8, 9.10, 9.12, 9.14, 9.16 に N 構造と予測構造のリボン図と距離マップを、図 9.9 , 9.11 , 9.13 , 9.15 , 9.17に (ϕ, ψ) 状態と2次構造を示す。これらから、以下のことがいえる。

- RMS 値は 9 から 13 Å 程度である。

表 9.15 ポテンシャル最小化構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	131	48	0.366	-0.33	-19.16	10.11	6.31	43.33	42.07
5PTI	144	28	0.194	-0.96	-22.42	11.36	7.29	47.71	43.29
2CRO	171	19	0.111	-1.38	-18.41	10.37	6.33	61.69	61.35
3ICB	186	13	0.070	-1.55	-23.43	11.36	6.64	69.12	68.02
1UBQ	87	1	0.011	-2.15	-15.69	12.99	7.88	36.28	33.92

表 9.16 ポテンシャル最小化構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-37.29	-19.78	-17.52	10.15	5.69	48.08	57.69
5PTI	-36.96	-19.75	-17.21	13.58	8.84	60.71	57.14
2CRO	-33.78	-18.00	-15.78	9.94	5.37	61.90	66.67
3ICB	-38.86	-17.30	-21.56	12.05	6.68	80.82	78.08
1UBQ	-30.51	-16.94	-13.58	10.52	5.96	48.65	47.30

- (ϕ,ψ) 状態一致率 $Q_{\phi\psi}$ と 2 次構造一致率 Q_{sec} は 60 % 前後の値であり、局所構造は比較的 N 構造と一致している。
- 距離マップから 3 次構造の部分的な特徴は一致していることがわかる。

図 9.8　4RXN の N 構造（左）と ED+EL による予測構造（中央）と距離マップ（右）。距離マップは上が N 構造、下が予測構造で、8 Å 以下が黒、16 Å 以下が灰色で表示した。

```
Seq  :MKKYTCTVCGYIYDPEDGDPDDGVNPGTDFKDIPDDWVCPLCGVGKDEFEEVEE
PP/N:!BBBBBAACCCBBBAAACCAACCBBCCBBAACBCAABBCAAACBBAACBBBBB!
SS/N:  EEEE     EEE                                    EEEE
PP/P:!BBBCCBBCCBBBBAACCBBCCCBBBCBAACCBBCACBBBCBBBCAAAAAAAA!
SS/P:          EEEE          EEE           EEE EEE HHHHHHHH
```

図 9.9　4RXN のアミノ酸配列 (Seq)、N 構造の二面角領域 (PP/N) と 2 次構造 (SS/N)、予測構造の二面角領域 (PP/P) と 2 次構造 (SS/P)

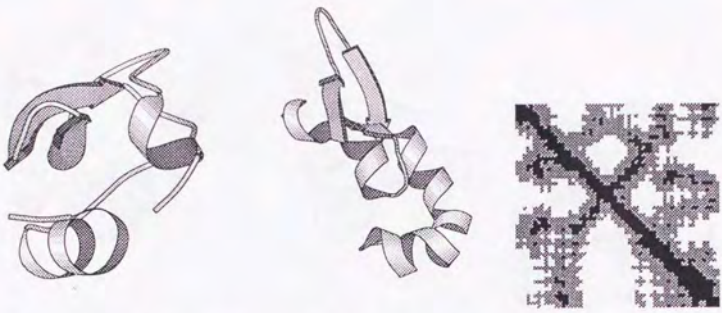


図 9.10　5PTI の N 構造（左）と ED+EL による予測構造（中央）と距離マップ（右）。距離マップは上が N 構造、下が予測構造で、8 Å 以下が黒、16 Å 以下が灰色で表示した。

```
Seq  :RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
PP/N:!BAAAABBBBACABCBCBBBBBBBBAAACBBBBBBBBACBCCBACBBABAAAAAAAACC!
SS/N:  HHHH            EEEEEEE    EEEEEEE            HHHHHHHH
PP/P:!AAAABBBBBCCBAAAAABBBBBBCACCBACBBBBCCAAAAAAAAAAAAAAAAAACC!
SS/P: HHHHEEE      HHHHHEEEE          EEEE  HHHHHHHHHHHHHHHHHH
```

図 9.11　5PTI のアミノ酸配列 (Seq)、N 構造の二面角領域 (PP/N) と 2 次構造 (SS/N)、予測構造の二面角領域 (PP/P) と 2 次構造 (SS/P)

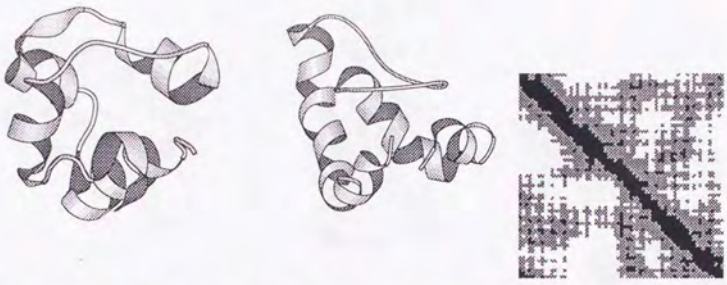


図 9.12　2CRO の N 構造（左）と ED+EL による予測構造（中央）と距離マップ（右）。距離マップは上が N 構造、下が予測構造で、8 Å 以下が黒、16 Å 以下が灰色で表示した。

```
Seq  :MQTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIAMALNCDPVWLQYGT
PP/N:!ABAAAAAAAAAAAACCBBAAAAAAAAACBBAAAAAAAAACCABACBACAAAAAAACCBBAAAAAAC!
SS/N:    HHHHHHHHHHH     HHHHHHHH   HHHHHHHH          HHHHHHH    HHHHHH
PP/P:!AAAAAAAAAAAAAAAAAAAAAAAACCBACBBBBBBACCBCCBAAAAAAAAAAAAACAAAAAAC!
SS/P: HHHHHHHHHHHHHHHHHHHHHHHH                    HHHHHHHHHHHH HHHHHH
```

図 9.13　2CRO のアミノ酸配列 (Seq)、N 構造の二面角領域 (PP/N) と 2 次構造 (SS/N)、予測構造の二面角領域 (PP/P) と 2 次構造 (SS/P)

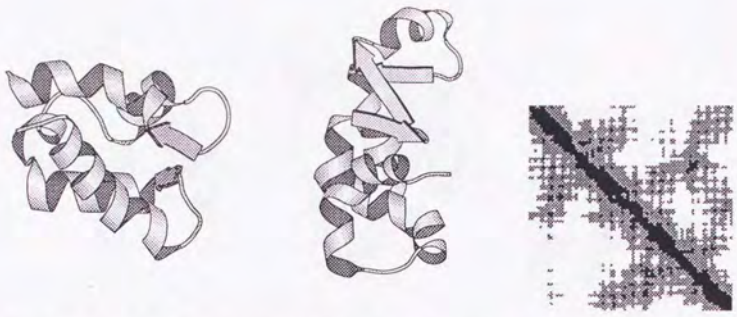


図 9.14　3ICB の N 構造（左）と ED+EL による予測構造（中央）と距離マップ（右）。距離マップは上が N 構造、下が予測構造で、8 Å 以下が黒、16 Å 以下が灰色で表示した。

```
Seq :KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ
PP/N:!BAAAAAAAAAAAAAABCBAABBBAAAAAAAAAAACAAAACCABBAAAAAAAABAACCCBBBAAAAAAAAAAAA!
SS/N:  HHHHHHHHHHHHHH     EEEHHHHHHHHHHH HHHH     HHHHHHHH     EEEHHHHHHHHHHH
PP/P:!BAAAAAAAAAAAAAACCCAAAAAAAAAAAAAAAAAABBBCCBBBBCAAAAAAAACCCCBBBAAAAAAAAAAA!
SS/P:  HHHHHHHHHHHHHH   HHHHHHHHHHHHHHHHHHEEE  EEEE HHHHHHHH    EEEHHHHHHHHHH
```

図 9.15　3ICB のアミノ酸配列 (Seq)、N 構造の二面角領域 (PP/N) と 2 次構造 (SS/N)、予測構造の二面角領域 (PP/P) と 2 次構造 (SS/P)

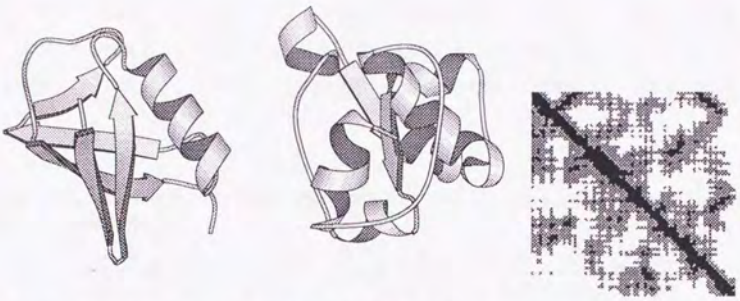


図 9.16　1UBQ の N 構造（左）と ED+EL による予測構造（中央）と距離マップ（右）。距離マップは上が N 構造、下が予測構造で、8 Å 以下が黒、16 Å 以下が灰色で表示した。

```
Seq :MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
PP/N:!BBBBBBACCBBBBBBBBAABBAAAAAAAAAAAACCBCAAABBBBBCCBBBBAABBAAACCBBCCBBBBBBBBBBC!
SS/N: EEEEEE   EEEEEE      HHHHHHHHHH       EEEEE  EEE              EEEEEEE
PP/P:!AAAAAAAACCBBCBBBBBCCAAAAAAAAAAAAACBBAAAAABBBBCCAAAACBBBBCCAAAAACBBBBAAAAAA!
SS/P: HHHHHHHH            HHHHHHHHHHHHH   HHHHHEEE   HHHH       HHHHH  EEEHHHHHH
```

図 9.17　1UBQ のアミノ酸配列 (Seq)、N 構造の二面角領域 (PP/N) と 2 次構造 (SS/N)、予測構造の二面角領域 (PP/P) と 2 次構造 (SS/P)

表 9.17 E_{seq} のみのポテンシャル最小化構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	100	100	1.000	3.87	-42.20	9.63	6.27	45.98	75.46
5PTI	100	100	1.000	2.67	-54.62	10.50	7.88	43.36	54.37
2CRO	100	100	1.000	2.39	-44.82	9.90	6.86	45.76	40.71
3ICB	100	81	0.810	0.86	-40.63	10.82	6.69	51.52	35.07
1UBQ	100	95	0.950	1.56	-42.99	12.03	8.26	41.38	47.80

表 9.18 E_{seq} のみのポテンシャル最小化構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-54.14	-39.00	-15.14	9.22	5.37	51.92	76.92
5PTI	-73.11	-64.83	-8.27	9.42	8.04	44.64	53.57
2CRO	-63.35	-55.43	-7.92	10.74	7.62	46.03	36.51
3ICB	-60.40	-38.78	-21.62	11.90	7.60	58.90	42.47
1UBQ	-62.21	-47.89	-14.32	12.48	8.42	39.19	48.65

9.6 ポテンシャル E_{seq} のみの最小化構造

参考のために「タンパク質らしい」構造の条件を考慮せず、ポテンシャル E_{seq} のみの最小化計算を同様に 5 つのタンパク質に対して行なった。これは $\lambda = 1.0$ とした場合の最小化計算に相当する。計算は初期構造を変えて 100 個の構造を生成した。「タンパク質らしい」構造の条件による選択は行なわない。表 9.17にこの構造群に対する Threading 結果を示した。ほとんどの構造が E^{nat}_{seq} より低いポテンシャルを持っていることがわかる。また、表 9.18にポテンシャル最小構造の N 構造との類似度を示した。「タンパク質らしさ」を考慮した場合の表 9.16と比較すると、RMS は 9 ～ 12 Å であまり変わらないが、$Q_{\phi\psi}$ が、50 % 前後の低い値になっていることがわかる。また、ポテンシャル最小構造のリボン図を 図 9.18 と図 9.19に示した。これらの構造は、小さく絡まりほとんど 2 次構造を形成しておらず、明らかに「タンパク質らしさ」を備えていない。「タンパク質らしさ」を考慮した図 9.8、図 9.10、図 9.12、図 9.14、図 9.16と比較すると その差は歴然としている。

図 **9.18** E_{seq} のみで最小化した場合の予測構造。4RXN(左), 5PTI(中央), 2CRO(右)。

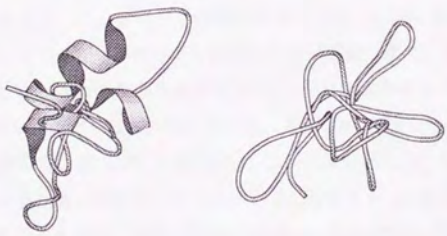


図 **9.19** E_{seq} のみで最小化した場合の予測構造。3ICB(左), 1UBQ(右)。

表 9.19 局所構造を N 構造に固定した場合のポテンシャル最小化構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	23	8	0.348	-0.50	-18.54	9.79	5.93	95.57	91.47
5PTI	83	16	0.193	-0.92	-26.39	9.77	6.39	97.91	91.65
2CRO	254	7	0.028	-2.07	-22.39	10.11	6.19	98.78	99.13
3ICB	216	49	0.227	-0.70	-33.07	10.33	5.79	99.80	91.91
1UBQ	23	0	0.000	-2.38	-19.43	12.54	7.45	98.47	80.49

9.7 局所構造を N 構造に固定した場合の予測構造

次に 2 次構造が完全に予測できた場合を想定し、局所構造を N 構造にほぼ固定してポテンシャル最小化計算を行なった。次のような拘束ポテンシャル $E_{\phi\psi-fix}$ を E_{pro} に加えることで固定する。

$$E_{\phi\psi-fix} = \begin{cases} 0 & Q_{\phi\psi} > 90\% \\ E_{max} & Q_{\phi\psi} \leq 90\% \end{cases} \tag{9.1}$$

つまり、拘束ポテンシャル $E_{\phi\psi-fix}$ は、N 構造との局所状態一致率 $Q_{\phi\psi}$ が 90 % を下回った場合のみ、非常に大きい値のペナルティ E_{max} を返す。この拘束ポテンシャル $E_{\phi\psi-fix}$ を加えた E_{pro} を用いて、方法で述べた探索の手続きに従って探索を行なう。すなわち、$E = (1-\lambda)E_{pro} + \lambda E_{seq}$ に対する最小化を、$\lambda = 0.4, 0.5, 0.6$ について 100 回ずつ行ない、計 300 個の構造を生成、その中で「タンパク質らしい」構造のみ選択する。表 9.19に この構造群に対する Threading 結果を示した。β シートの多いタンパク質 (4RXN, 5PTI, 1UBQ) は、局所構造が N 構造と同一でも β シートのペアが形成されない場合が多く、300 個中 100 個以下しか「タンパク質らしい」構造の条件を満たさない。1UBQ 以外のタンパク質では、E^{nat}_{seq} よりポテンシャルが低い構造がやはり出現している。表 9.20に ポテンシャル最小構造（予測構造）の N 構造との類似度、図 9.20 と 図 9.21に リボン図を示した。局所状態一致率 $Q_{\phi\psi}$ は当然のことながらかなり高い値になっているが、RMS は 6 から 11 Å で、固定しない場合に比べて向上はしているものの、期待したほど N 構造と似た構造は得られなかった。

表 **9.20**　局所構造を N 構造に固定した場合のポテンシャル最小化構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-31.79	-22.65	-9.14	7.57	5.18	96.15	92.31
5PTI	-34.02	-21.78	-12.25	11.95	8.04	98.21	76.79
2CRO	-30.85	-17.90	-12.95	10.04	8.17	96.83	96.83
3ICB	-42.83	-22.89	-19.94	6.42	4.16	100.00	91.78
1UBQ	-27.82	-19.77	-8.05	11.63	7.11	97.30	75.68

図 **9.20**　局所構造を N 構造に固定した場合のポテンシャル最小化計算による予測構造。4RXN(左), 5PTI(中央), 2CRO(右)。

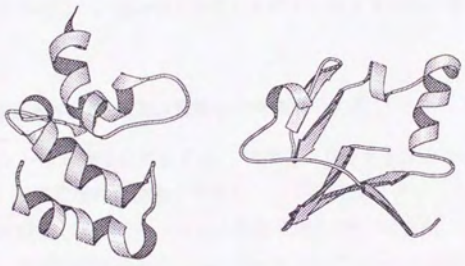


図 9.21　局所構造を N 構造に固定した場合のポテンシャル最小化計算による予測構造。3ICB(左), 1UBQ(右)。

第 10 章

考察

10.1 ポテンシャルと N 構造からの距離の分布の解析

本章では、ポテンシャルと N 構造からの距離・類似度の分布を詳細に解析することで、ポテンシャル曲面の様相を推定し、本研究で設定したポテンシャル関数・構造探索の有効性と問題点を考察する。

10.1.1 N 構造近傍構造群・N 構造近傍最小化構造群の生成

前章で述べたポテンシャル最小化計算では、N 構造よりポテンシャル値が低くかつ N 構造から RMS 値が 10 Å 程度離れた構造が得られた。この構造がグローバルミニマムであるとするなら、本研究で設定したポテンシャルは構造予測に用いるには大きな問題があることを意味する。しかし、本研究で行なったシミュレーテッドアニーリングによる構造探索は必ずグローバル・ミニマム構造が得られることが保証されているわけではない。N 構造から遠くないところに E^{nat} より 極めてポテンシャルの低い構造があったのだが、シミューレーテッド・アニーリングの性能の限界から N 構造に良く似たグローバルミニマム構造が得られなかったということも考えられる。そこで、N 構造近傍の構造群を以下の 2 つの方法で生成し、N 構造近傍にポテンシャルの低い構造があるかどうか調べた。

1. N 構造近傍構造群 (Nnear)

   Build-up 法で得られた Nnear 構造の二面角を 1 から 3 残基ランダムに変移させた構造群の中から、「タンパク質らしい」構造の条件を満たす構造を選択することで得られた構造群。Nnear 構造も含まれる。

2. N 構造近傍最小化構造群 (NnearMin)

   Nnear 構造を初期構造として、$E = (1-\lambda)E_{seq} + \lambda E_{pro}$ に対してポテンシャル最小化を行なって得られた構造群。Nnear 構造から離れ過ぎることを防ぐために、最大ス

表 10.1 N 構造近傍構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	13	0	0.000	-3.59	1.92	9.86	6.00	58.28	45.86
5PTI	26	0	0.000	-5.60	-15.96	5.92	4.28	83.93	82.00
2CRO	623	0	0.000	-3.58	-18.02	4.13	2.98	89.24	90.53
3ICB	384	0	0.000	-4.63	-23.41	3.85	2.67	88.37	86.47
1UBQ	360	0	0.000	-4.45	-17.93	4.61	3.04	84.71	72.60

表 10.2 N 構造近傍最小化構造群に対する Threading 結果

PDBcode	N_{comp}	$N_{E<E^{nat}}$	$P(E<E^{nat})$	Z^{nat}	$\bar{E}_{seq}$	$\overline{RMS}$	$\overline{DRMS}$	$\bar{Q}_{\phi\psi}$	$\bar{Q}_{sec}$
4RXN	112	39	0.348	-0.40	-19.57	6.90	4.39	65.04	73.56
5PTI	25	2	0.080	-1.18	-25.43	10.01	6.09	77.43	74.71
2CRO	248	41	0.165	-0.99	-23.39	7.94	5.42	83.49	83.79
3ICB	204	3	0.015	-2.28	-28.87	6.12	3.74	89.04	92.96
1UBQ	109	23	0.211	-0.77	-27.43	6.51	4.30	86.19	78.69

テップ数 $MAX = 1000$ とし、温度は $T = 0$ に固定したシミュレーテッド・アニーリング法を行なった。$\lambda = 0.4, 0.5, 0.6$ に対して 100 個ずつ構造を生成し、得られた構造群の中から「タンパク質らしい」構造の条件を満たす構造のみ選択する。

表 10.1 と 表 10.2に N 構造近傍群と N 構造最小化構造群に対する Threading 結果を示した。N 近傍構造群で 4RXN、5PTI の構造数が極端に少ないのは、Nnear 構造の定常 2 次構造がかなり少ないため、変移させたときに「タンパク質らしい」構造の条件の 2 次構造の条件を満たさなくなるためである。N 構造近傍群は $N_{E<E^{nat}} = 0$ だが、N 構造近傍ポテンシャル最小化構造群は 相当数の $E < E^{nat}$ である構造が存在する。

表 10.3 と 表 10.4 にポテンシャル最小構造群のデータを示した。特に、表 10.4に注目すると、これらの構造群は 表 9.16に示したランダム構造を初期構造としたポテンシャル最小構造と比べて、同等のポテンシャルの値を持ち、しかも N 構造との類似度ははるかに高い。このことは、N 構造付近にもポテンシャルがかなり低い構造が存在することを示す。

表 10.3　N 構造近傍構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-8.80	-0.81	-7.98	9.24	5.46	65.38	51.92
5PTI	-22.92	-12.51	-10.41	1.45	1.00	83.93	85.71
2CRO	-26.87	-12.79	-14.08	1.34	1.12	92.06	95.24
3ICB	-26.23	-9.36	-16.87	1.91	1.43	87.67	84.93
1UBQ	-22.25	-10.75	-11.50	2.16	1.58	85.14	77.03

表 10.4　N 構造近傍最小化構造群の中でのポテンシャル最小構造のポテンシャル値と N 構造からの距離・類似度

PDBcode	E_{seq}	ED	EL	RMS	$DRMS$	$Q_{\phi\psi}$	Q_{sec}
4RXN	-29.60	-14.88	-14.71	9.25	4.46	63.46	80.77
5PTI	-30.61	-15.76	-14.85	11.84	7.38	80.36	78.57
2CRO	-32.89	-19.16	-13.73	5.49	5.05	82.54	82.54
3ICB	-37.08	-19.70	-17.38	2.48	1.86	90.41	95.89
1UBQ	-36.46	-24.63	-11.83	7.79	5.40	87.84	74.32

10.1.2 ポテンシャルとN構造からの距離の相関

構造予測のためのポテンシャルが満たすべき条件として、今までは「N構造がポテンシャル最小構造となること」のみをチェックしてきたが、いくらこの条件を満たしていても、構造探索においてポテンシャル最小構造を得ることができなければ意味がない。本研究で採用したシミュレーテッド・アニーリング法に限らず、一般にポテンシャル最小構造を探索する場合、ポテンシャル値が低くなるような微小構造変移を繰り返す方法を用いる場合が多い。この場合「ポテンシャルが低くなるほどN構造に近付くこと」つまり、ポテンシャルとN構造からの距離との相関が高いことが条件となる。そこで、様々な構造群に対して、ポテンシャルと距離・類似度との相関係数を調べたのが表10.5である。ここでは、ポテンシャル E_{seq}・ED・EL と RMS・$Q_{\phi\psi}$ との相関係数を5つの構造群に対して示した。RMSは距離、$Q_{\phi\psi}$ は類似度であるので、理想的なポテンシャルの場合、RMSとポテンシャルの相関係数は +1.0、$Q_{\phi\psi}$ とポテンシャルの相関係数は -1.0 となるはずである。この表から以下のことがわかる。

- 一般に $Q_{\phi\psi}$ と E_{seq} の負の相関が高い。この相関は局所構造ポテンシャル EL によってもたらされていることが、$Q_{\phi\psi}$ と EL の相関が高いことからわかる。ただし、タンパク質ごとの差が大きく、α ヘリックスの多いタンパク質、2CRO と 3ICB に対する負の相関が特に高い。
- RMS は、一般にさほど高い正の相関を示さない。4RXN では相関係数が負になってしまっている。N構造近傍構造群 と N構造近傍最小化構造群は例外的に高い正の相関を示すが、このことはこれらの構造群の生成法から考えて当然の帰結かもしれない。

E_{seq} と RMS、$Q_{\phi\psi}$ の分布図を、3ICB（図10.1、10.2）と1UBQ（図10.3、10.4）について示す。これらの図を観察することで、相関係数の数値に現れていることが視覚的・直観的に理解できる。

Park と Levitt は、2次構造をN構造にほぼ固定した構造群において、本研究と同じように、RMSと様々なポテンシャルについて同じような分布図の解析を行なっている[66]。彼らは、Sippl ポテンシャル、Miyazawa-Jernigan のコンタクトポテンシャルなど様々なポテンシャルを試しているが、いずれも本研究で示した図のように RMS とポテンシャルの相関は低い。

10.1.3 構造間の距離の分布

これまでの解析はすべてN構造との距離・類似度を計算していたが、予測構造群として得られた最小化構造群内の構造はお互いにどれだけ似ているのだろうか。表10.6に最小化構造

表 10.5　ポテンシャル E_{seq}・ED・EL と RMS・$Q_{\phi\psi}$ との相関係数。$E_{seq} = ED + EL$ である。

		E_{seq}		ED		EL	
	群	RMS	$Q_{\phi\psi}$	RMS	$Q_{\phi\psi}$	RMS	$Q_{\phi\psi}$
4RXN	X-ray	-0.02	-0.26	-0.04	-0.08	-0.01	-0.25
	Decoy	-0.12	-0.20	0.04	-0.06	-0.15	-0.19
	Min	0.00	-0.30	0.11	-0.14	-0.11	-0.31
	Nnear	-0.31	-0.69	-0.05	-0.47	-0.35	-0.60
	NnearMin	0.14	0.14	0.15	0.09	0.10	0.18
5PTI	X-ray	0.01	-0.30	0.06	0.01	-0.01	-0.35
	Decoy	0.06	-0.32	0.02	-0.01	0.05	-0.37
	Min	0.02	-0.35	0.10	-0.10	-0.06	-0.45
	Nnear	0.45	0.15	0.43	-0.08	0.07	0.62
	NnearMin	0.35	-0.15	0.27	-0.16	0.33	-0.07
2CRO	X-ray	0.12	-0.67	0.09	-0.06	0.10	-0.74
	Decoy	0.15	-0.65	0.03	-0.05	0.16	-0.73
	Min	0.21	-0.51	0.18	-0.28	0.19	-0.58
	Nnear	0.63	-0.83	0.70	-0.58	0.39	-0.91
	NnearMin	0.55	-0.48	0.53	-0.42	0.39	-0.41
3ICB	X-ray	0.04	-0.69	-0.15	-0.11	0.10	-0.73
	Decoy	0.14	-0.60	0.03	-0.07	0.14	-0.65
	Min	0.13	-0.47	0.24	-0.32	0.02	-0.43
	Nnear	0.71	-0.68	0.72	-0.46	0.30	-0.77
	NnearMin	0.28	-0.40	0.34	0.02	-0.02	-0.70
1UBQ	X-ray	0.03	-0.11	-0.02	0.02	0.05	-0.14
	Decoy	0.04	-0.07	0.04	0.01	0.02	-0.10
	Min	0.27	-0.17	0.36	0.05	0.12	-0.26
	Nnear	0.82	-0.31	0.85	-0.26	0.03	-0.22
	NnearMin	0.05	-0.25	0.04	-0.09	0.04	-0.54

＊ X-ray：結晶構造群、Decoy：デコイ構造群、Min：最小化構造群、Nnear：N 構造近傍構造群、NnearMin：N 構造近傍最小化構造群

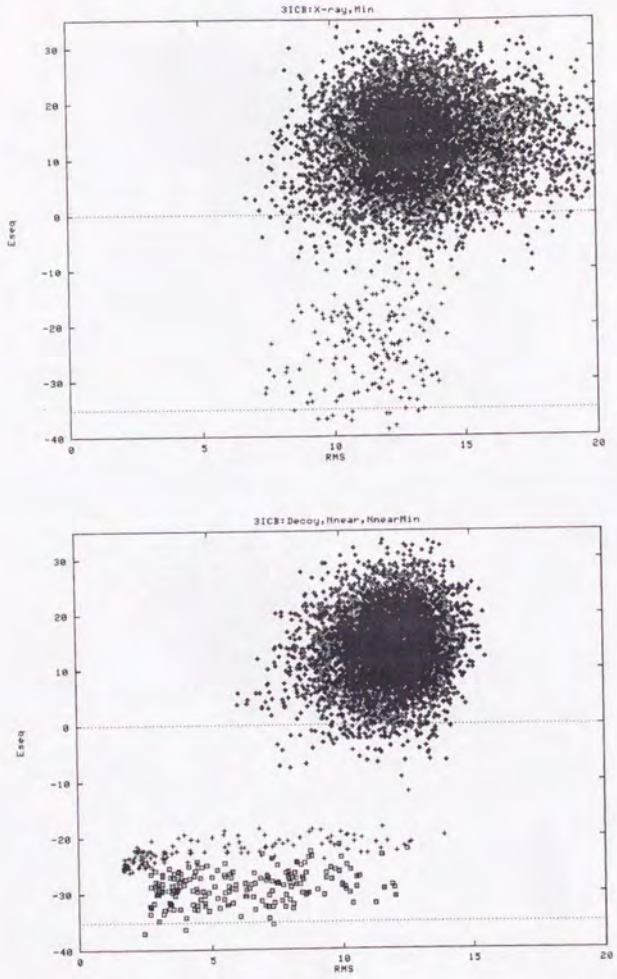


図 **10.1**　3ICB の場合の E_{seq} と RMS の分布図。上図（◇：結晶構造群、＋：最小化構造群）、下図（◇：デコイ構造群、＋：N 構造近傍構造群、□：N 構造近傍最小化構造群）グラフの下側の点線は E^{nat} を示している。

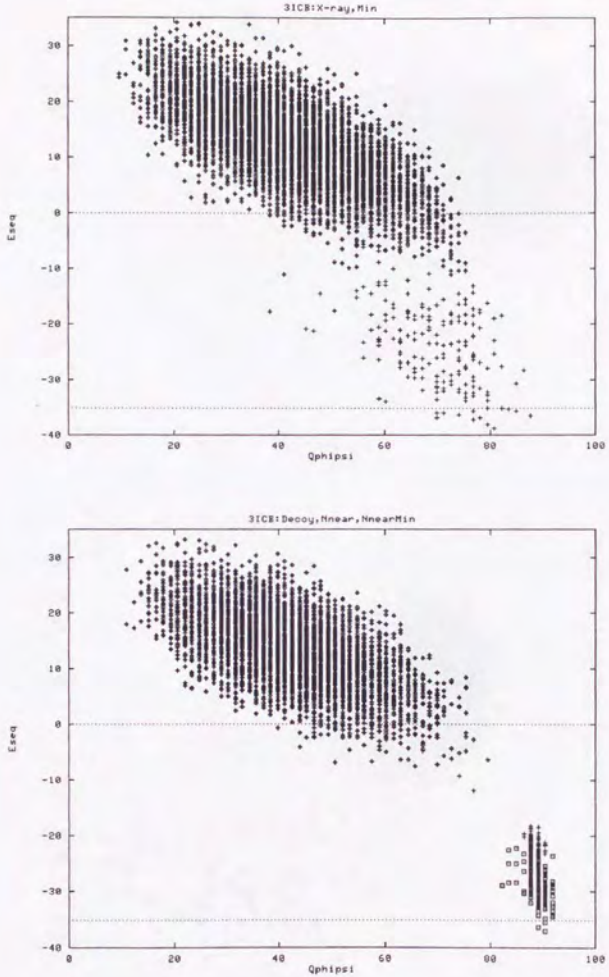


図 **10.2**　3ICB の場合の E_{seq} と $Q_{\phi\psi}$ の 分布図。上図（◇：結晶構造群、＋：最小化構造群）、下図（◇：デコイ構造群、＋：N 構造近傍構造群、□：N 構造近傍最小化構造群）グラフの下側の点線は E^{nat} を示している。

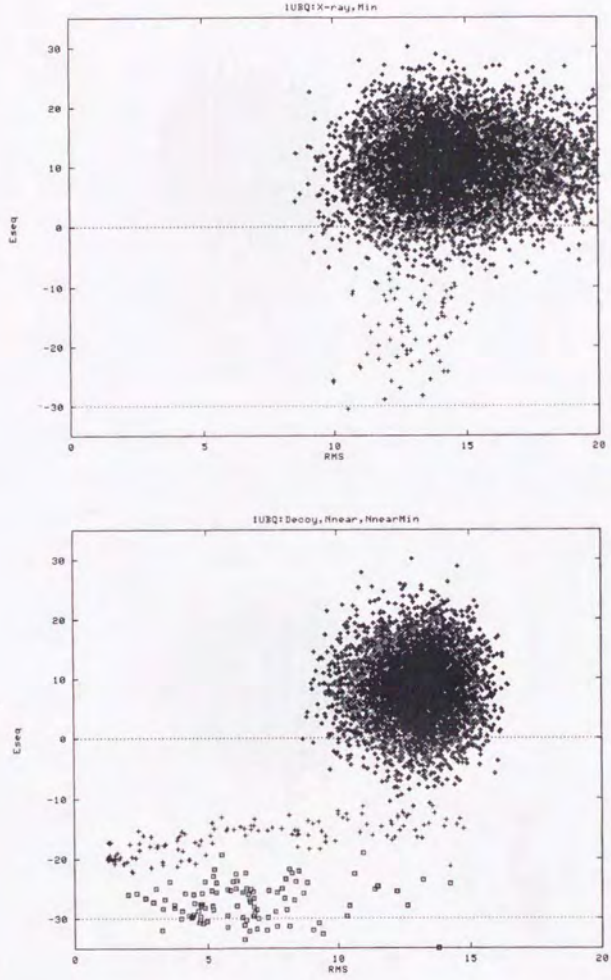


図 **10.3**　1UBQ の場合の E_{seq} と RMS の 分布図。上図（◇：結晶構造群、＋：最小化構造群）、下図（◇：デコイ構造群、＋：N 構造近傍構造群、□：N 構造近傍最小化構造群）グラフの下側の点線は E^{nat} を示している。

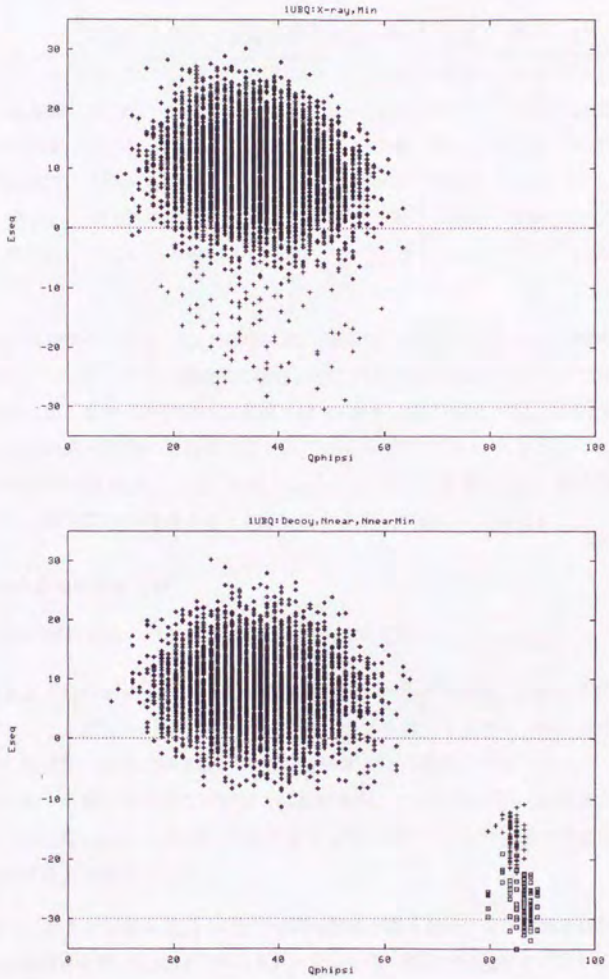


図 **10.4**　1UBQ の場合の E_{seq} と $Q_{\phi\psi}$ の 分布図。 上図（◇：結晶構造群、＋：最小化構造群）、下図（◇：デコイ構造群、＋：N 構造近傍構造群、□：N 構造近傍最小化構造群）グラフの下側の点線は E^{nat} を示している。

表 10.6　デコイ構造群と最小化構造群内の全ての構造間の $RMS \cdot Q_{\phi\psi}$ の平均、標準偏差、最大値、最小値

		$\overline{RMS}$	σ^{RMS}	RMS^{min}	RMS^{max}	$\overline{Q_{\phi\psi}}$	$\sigma^{Q_{\phi\psi}}$	$Q_{\phi\psi}^{min}$	$Q_{\phi\psi}^{max}$
Decoy	-	13.13	1.45	6.74	18.35	42.20	10.29	3.85	96.15
Min	4RXN	10.00	1.44	5.07	14.61	54.38	9.96	19.23	88.46
	5PTI	10.56	1.50	4.87	15.34	57.01	12.06	10.71	94.64
	2CRO	10.99	1.56	5.50	16.46	64.08	12.13	15.87	98.41
	3ICB	11.97	1.60	5.48	17.86	62.67	9.90	28.77	94.52
	1UBQ	12.54	1.54	6.86	16.59	51.63	10.19	13.51	81.08

群内の全ての構造間の $RMS \cdot Q_{\phi\psi}$ の平均値、最小値、最大値を示した。比較のためにデコイ構造群の値も示した。デコイ構造群の場合、ペアの数が莫大になるのでランダムに 1/10 を選択して計算した。最小化構造群では RMS は 10 から 12 Å 程度、$Q_{\phi\psi}$ は 50 から 60 % 程度であり、最小化構造群内の構造はさほど似ているわけではない。しかし、これらの値を、デコイ構造群の平均値 $RMS = 13.13$ Å、$Q_{\phi\psi} = 42.20\%$ と比較すると、最小化構造群内の構造は、デコイ構造群内の構造よりもお互いにやや似ていることがわかる。

10.1.4　分布の解析のまとめ

以上の分布の解析から、わかったことを以下にまとめる。

- N 構造より低いポテンシャル値の構造が出現するかどうかは、どのように構造群をサンプルしたかに大きく依存する。つまり、E_{seq} に対するポテンシャル最小化計算で得られた構造群（最小化構造群、N 構造近傍最小化構造群）では、多く $E < E^{nat}$ の構造が出現するが、そうでない場合（結晶構造群、デコイ構造群、N 構造近傍構造群）では出現しない。これは、図 10.6 のように中心のポテンシャル値が異なる 2 つの分布で表現されると考えられる。

- RMS と ポテンシャル E_{seq} の 正の相関はさほど高くない。さらに最小化構造群でも N 構造近傍最小化群でもほぼ同等のポテンシャル値の構造が出現することから、図 10.6の上図のような水平に近い分布構造になっている。

- $Q_{\phi\psi}$ と ポテンシャル E_{seq} の負の相関は、α ヘリックスの多いタンパク質だと極めて高く、β シートの多いタンパク質だと低い。これをまとめたのが図 10.6の中図と下図である。

10.1.5 ポテンシャル曲面の推察

分布図の解析から、RMS と $Q_{\phi\psi}$ に対するポテンシャル曲面の形状の推察を行なった（図 10.7）。ここで、まず考えなければならないのは、あるRMSあるいは $Q_{\phi\psi}$ に対する可能な構造の数である。「タンパク質らしい」構造群の中でN構造から RMS 離れている構造の数を $N(RMS)$、$Q_{\phi\psi}$ 一致している構造数を $N(Q_{\phi\psi})$ とすると、$N(RMS)$ と $N(Q_{\phi\psi})$ のどちらもある値を頂点とした吊り鐘型の分布になる（図 10.5）。このことは、いかなるポテンシャルを用いても構造空間の性質から、吊鐘のピークの値をとりやすいというバイアスがかかることを意味する。ポテンシャル最小化計算を行なうと、$N(RMS)$ と $N(Q_{\phi\psi})$ の値の高い構造を中心にサンプルしながら、近傍のポテンシャル極小構造に変移するものと考えられる。このとき図10.7のようなポテンシャル曲面を考えると、図 10.6のような分布を定性的に説明できる。RMS の場合極めて凹凸の多いポテンシャル曲面を持ち、N構造付近に限らず至るところに深い極小解があると考えられる。ランダム構造からポテンシャル最小化を行なうと $N(RMS)$ のピーク付近の極小解にトラップされてしまう。α ヘリックスの多いタンパク質についての $Q_{\phi\psi}$ の場合は、滑らかなポテンシャル曲面をもつので、容易にN構造付近までたどりつくことができると考えられる。β シートの多いタンパク質についての $Q_{\phi\psi}$ の場合は、両者の中間的な性質を持ち、途中までは滑らかなポテンシャル曲面であるが、N構造に近付くにつれ、遠距離相互作用が支配的になり、RMS の場合のような凹凸の多い曲面になるのではないかと推察される。

近年、ポテンシャル曲面の形状は漏斗 (funnel) であるという説がWolynesらによって提唱されている[80)81)]。彼らの考えるポテンシャル曲面は 慣性半径を軸として記述される場合が多い。慣性半径が小さくなるまでは比較的なだらかなポテンシャル曲面であり、N構造よりやや大きい慣性半径の状態（モルテン・グロビュール状態[82)]）からN構造に至るまでは凹凸の多いポテンシャル曲面になるというのが、漏斗モデルのポテンシャル曲面の描像である。本研究では「タンパク質らしい」条件を満たす構造群を対象としたポテンシャル曲面を考えた。条件のひとつにコンパクトさがあるため、対象とした構造群の慣性半径はすべてN構造程度にそろっている。これは、モルテン・グロビュール的な条件を満たす構造群の間のポテンシャル曲面を比較していることになり、漏斗モデル的なポテンシャル曲面の描像を正しいならば、フォールディング後期の凹凸の多い曲面となるはずである。これは、本研究で推察したRMSについての E_{seq} のポテンシャル曲面の性質と符合するかもしれない。

10.2 本研究で改良すべき点

本研究では、結晶構造群へのThreadingによって評価したポテンシャル E_{seq} を用いて、「タンパク質らしい」条件を満たす構造を探索して構造予測を行なった。その結果、ある程

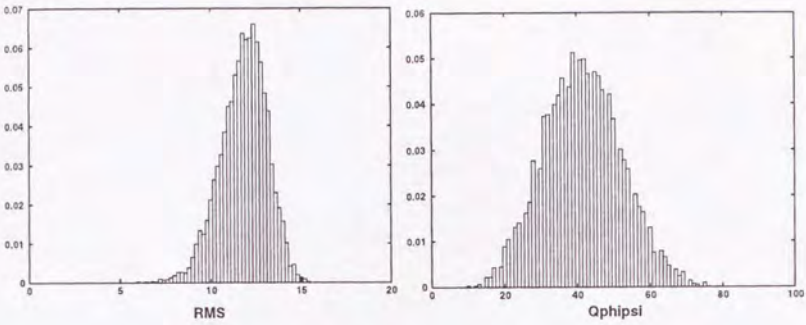


図 10.5　3ICB のデコイ構造群の場合の N 構造からの RMS のヒストグラム（左図）と N 構造との $Q_{\phi\psi}$ のヒストグラム（右図）

度 N 構造と類似した構造が得られたが、満足できるほどの結果は得られなかった。今後、どのような点を改良していけば良いのか考察する。

結晶構造群への Threading テストだけではポテンシャルの評価として十分でない

本研究で用いたポテンシャルは、結晶構造群への Threading テストにおいては、ほとんどのタンパク質において N 構造がポテンシャル最小構造になったが、シミュレーテッド・アニーリングでポテンシャル最小化計算を行なうと、N 構造よりポテンシャルの低い構造が多数存在した。このことは、結晶構造群への Threading テストだけでは、ポテンシャルを評価するのは不十分であることを意味する。より精度の高いポテンシャルを構築するには、多少の計算量の増加は覚悟の上で、ポテンシャル最小化計算を含んだ形の Threading テストを行なう必要があると考えられる。

離散 (ϕ, ψ) モデルは妥当か

離散 (ϕ, ψ) モデルは、比較的良く N 構造を表現することができたが、それでも、Nnear 構造のポテンシャル値は N 構造のポテンシャルよりも高い（表 9.9、表 9.10参照）。実際に得られる可能性のあるモデルは Nnear 構造であり、離散化したことにより N 構造に近い構造が得られる可能性を下げたとも考えられる。しかし、詳細なモデルでは構造探索が困難になる。詳細なモデルを採用するためには、より効率的な構造探索法の開発が不可欠であるといえる。

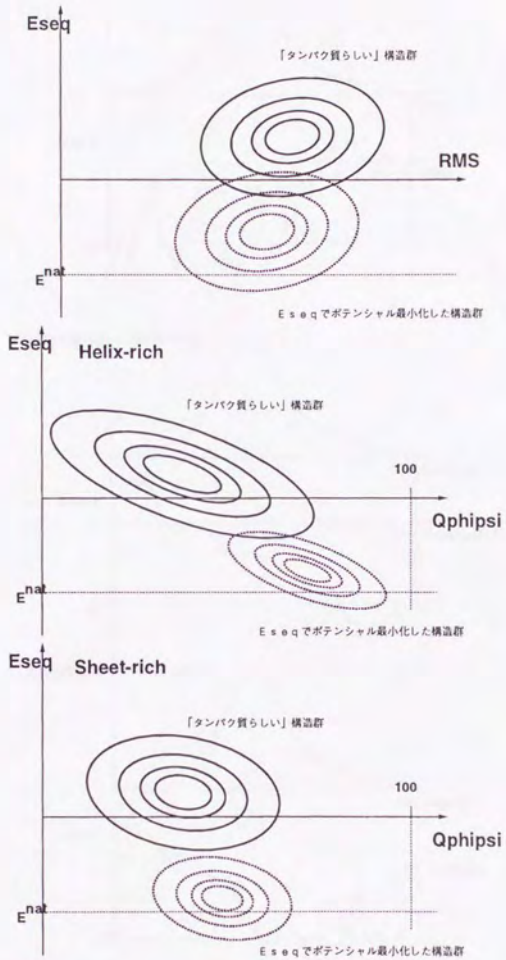


図 10.6　ポテンシャルの分布の概念図。上図：E_{seq} と RMS の分布。中図：α ヘリックスの多いタンパク質における E_{seq} と $Q_{\phi\psi}$ の分布。下図：β シートの多いタンパク質における E_{seq} と $Q_{\phi\psi}$ の分布。

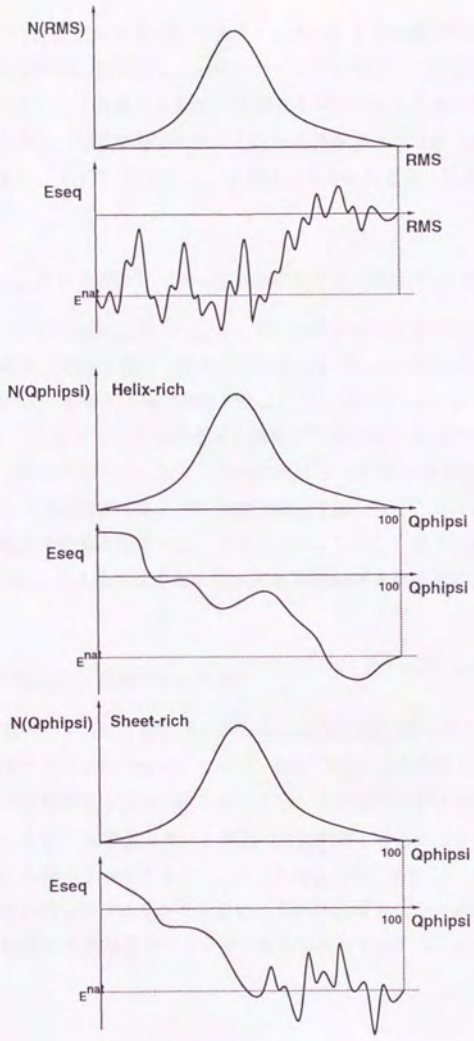


図 **10.7**　可能な構造の数 N と 推察したポテンシャル曲面の概念図。上図は RMS、中図は α ヘリックスの多いタンパク質の場合の $Q_{\phi\psi}$ 下図は β シートの多いタンパク質の場合の $Q_{\phi\psi}$ の場合である。

「タンパク質らしい」構造の条件は妥当なのか

本研究で設定した (1) 自己排除性 (2) コンパクトさ (3) 2次構造の形成の3つの「タンパク質らしい」構造の条件は、極めてヒューリスティックに設定された条件である。この条件はかなり緩い条件であり、より精密な条件がN構造をポテンシャル最小構造とするために必要かもしれない。しかし、条件を精密にするとこの条件を満たす構造の探索自体が困難となる可能性がある。また、「タンパク質らしい」構造の条件から設定した E_{pro} が妥当であるかどうか疑わしい。

$(1-\lambda)E_{pro}+\lambda E_{seq}$ としてポテンシャル最小化することに問題はないのか

本研究では、$E=(1-\lambda)E_{pro}+\lambda E_{seq}$ として、ポテンシャル最小化計算を行ない、「タンパク質らしい」構造の条件を満たす構造のみ有効とした。この方法は 任意性の高い E_{pro} と λ は、陽には現れないが以下の形で影響を及ぼしている。(1) ポテンシャル最小化計算で得られた構造は すべて E についての局所最小構造である。ある構造が局所最小であるかどうかは λ と E_{pro} に依存することになり、その設定によっては N構造近傍に局所最小解 が存在しないことになる可能性がある。(2) 正確な構造予測を行なうには、現実のタンパク質と同じ経路をとる最小化計算を行なうことが望ましい。しかし、最小化計算における構造遷移の経路は、λ と E_{pro} に大きく依存し、その的確な設定が予測結果に大きな影響を及ぼす可能性がある。

統計ポテンシャルの定式化に問題はないのか

本論文の序論で述べたように、統計ポテンシャルは 結晶構造群からN構造を識別できるように設計されたポテンシャルである。よって、結晶構造群と本研究で設定した「タンパク質らしい」構造群の特性が完全には一致しない場合、その違いが予測精度に影響を及ぼしているかもしれない。また、N構造とデコイ構造の判別問題と考えたときには、ポテンシャル値が負になる場合にN構造と判定することはベイズ決定だが、ポテンシャル最小構造がN構造と一致する必然性は理論的には明らかでない。また統計ポテンシャルで用いられている独立の仮定が、より精密にN構造をポテンシャル最小構造とするためには大きな問題となる可能性もある。

第 11 章

第 II 部 の まとめ

第 II 部では統計ポテンシャルを 3 次構造予測に応用するため、次の順序で立体構造予測を行なった。

- 統計ポテンシャルの定式化で作成した配列依存ポテンシャル E_{seq} を複数作成し、結晶構造群への Threading によって評価した。その結果、距離ヒストグラムポテンシャルと局所構造ポテンシャルの組み合わせが最も成績が良く、この 2 つの和を E_{seq} として採用した。
- 「タンパク質らしい」構造の条件を設定し、それを満たすほどポテンシャルが低くなる E_{pro} を設定した。
- 離散的な (ϕ, ψ) の代表点 でタンパク質の構造を表現するモデルを開発した。
- 何通りかの λ で $E = (1-\lambda)E_{pro} + \lambda E_{seq}$ に対して繰り返しポテンシャル最小化計算を行ない、得られた構造の中で「タンパク質らしい」構造のみを最小化構造群とした。最小化構造群の中でもっとも E_{seq} が低い構造を予測構造とした。

これらの手続きで得られた予測構造は、RMS 値はさほど良くなかったが、2 次構造は比較的良く一致しており、3 次構造の大まかな特徴も一致していた。また、「タンパク質らしさ」を考慮せずに E_{seq} だけでポテンシャル最小化を行なった場合に比べ、良好な結果が得られた。

また最小化構造群の中には N 構造よりポテンシャル値の低い構造が相当数存在した。このことは、最小化計算に用いるポテンシャルは、結晶構造群への Threading テストを満たすだけでは、不十分であることを示す。

さらに、ポテンシャルと N 構造からの距離の関係を詳細に調べたところ、局所構造一致率 $Q_{\phi\psi}$ に比べ RMS とポテンシャル E_{seq} の相関は低かった。このことは、3 次構造予測の難しさを示している。また、$Q_{\phi\psi}$ とポテンシャルとの相関は、α ヘリックスが多いタンパク質ほ

ど高く、βシートの多いタンパク質ほど低かった。これはβシートの２次構造予測の難しさと対応すると思われる。

今後より精密な予測を行なうには(1)Threadingテスト以上の精密さのポテンシャル評価法を開発し、ポテンシャルとタンパク質らしい」構造の条件を改良すること、(2)進化的・実験的な情報を用いて有効に構造空間を狭めていくことの２つの方針が考えられる。

第 12 章

結論

本研究では、統計ポテンシャルを2次構造予測と3次構造予測に適用した。2次構造予測においては新たに2元語ポテンシャルを加えることで予測精度を改善することができ、さらに、最適な符号化関数の探索から疎水性相互作用が2次構造形成に影響を及ぼすことが示唆された。3次構造予測においては、統計ポテンシャル E_{seq} の低い「タンパク質らしい」構造を探索する方法を開発した。5つのタンパク質に適用した結果、RMS 値は良い値が得られなかったものの、2次構造は比較的良く一致し、全体の構造の特徴も類似していた。より精度の良い3次構造予測を行なうには、ポテンシャルや「タンパク質らしい」構造の条件をより改善するとともに、実験的・進化的な情報を積極的に採り入れ、配座空間を狭めていく必要があると考えられる。

謝辞

本研究は，著者が東京大学農学生命科学研究科応用生命工学専攻生物情報工学研究室に在籍中に行なったものであり、研究に進めるにあたり親切な御指導を頂いた清水 謙多郎 助教授に心からの謝意を表します。また第 I 部 の研究は、土井 淳多 名誉教授に御指導を頂きました。池口満徳助手，中村周吾助手，田崎康一君，木下 宏君，清水 青史君，広瀬 仁 君，相良 純一君には研究に関して貴重なディスカッションをして頂きました．心から感謝いたします。

参考文献

1) S.J.Wodak and M.J.Rooman. Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, Vol. 3, pp. 247–259, 1993.

2) D.T.Jones and J.M.Thornton. Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, Vol. 6, pp. 210–216, 1996.

3) J. Garnier, D. J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple method for predicting the secondary structure of globular proteins. *J. Mol. Biol*, Vol. 120, pp. 97–120, 1978.

4) M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol*, Vol. 213, pp. 859–883, 1990.

5) B. H. Zimm and J. K. Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *Journal of Chemical Physics*, Vol. 31, No. 2, pp. 526–535, 1959.

6) P.D. Thomas and K.A.Dill. Statistical potentials extracted from protein structures: How accurate are they? *J.Mol.Biol.*, Vol. 257, pp. 457–469, 1996.

7) M. L. Minsky, S. A. Papert (中野 馨／阪口 豊訳). パーセプトロン. パーソナルメディア, 1988.

8) P. Y. Chou and G. D. Fasman. Conformational parameters for amino acids in helical,sheet,and random coil regions calculated from proteins. *Biochemistry*, Vol. 13, pp. 211–222, 1974.

9) P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemisry*, Vol. 2, pp. 222–245, 1974.

10) J. F. Gibrat, J. Garnier, and B. Robson. Further development of protein secondary structure prediction using information theory -new parameters and consideration of residue pairs. *J. Mol. Biol*, Vol. 198, pp. 425–443, 1987.

11) K. Nagano. Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol*, Vol. 109, pp. 251–274, 1977.

12) N. Qian and J. Sejnowski. Prediction the secondary structure of globular proteins using neural network models. *J. Mol. Biol*, Vol. 202, pp. 865–884, 1988.

13) L.H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. *Proc.Natl.Acad.Sci.USA*, Vol. 86, pp. 152–156, 1989.

14) K. Nishikawa and T. Ooi. Amino acid sequence homology applied to the predition of protein secondary structures,and joint prediction with existing methods. *Biochimica et Biophysica Acta*, Vol. 871, pp. 45–54, 1986.

15) T. Yi and E. S. Lander. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol*, Vol. 232, pp. 1127–1129, 1993.

16) B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*, Vol. 232, pp. 584–599, 1993.

17) B. Rost and C. Sander. Combining evolutionary information and neural networks to predict secondary structure. *PROTEINS*, Vol. 19, pp. 55–72, 1994.

18) T. Kawabata and J. Doi. Improvement of protein secondary structure prediction using binary word encoding. *Proteins*, Vol. 26, , 1996 in press.

19) V. I. Lim. Algorithms for prediction of alpha-helical and beta-structural regions in grobular proteins. *J. Mol. Biol*, Vol. 88, pp. 873–894, 1974.

20) B. Robson. Analysis of the code relating sequence to conformation in globular protein. *Biochem.J.*, Vol. 141, pp. 853–867, 1974.

21) L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*, Vol. 233, pp. 123–138, 1993.

22) B. W. Matthews. Comparisons of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, Vol. 405, pp. 442–451, 1975.

23) S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 13, pp. 252–264, 1991.

24) 鳥脇純一郎. 認識工学. コロナ社, 1993.

25) S. Kirkpatrick, C. D. Galatt, and M.P.Vecchi Jr. Optimization by simulated annealing. *Science*, Vol. 220, pp. 671–680, 1983.

26) C. Branden and J. Tooze. タンパク質の構造入門. 教育社, 1991.

27) M. Schiffer and A. B. Edmundson. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.*, Vol. 7, pp. 121–135, 1967.

28) D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, Vol. 81, pp. 140–144, 1984.

29) M. W. West and M. H. Hecht. Binary patterning of polar and nopolar amino acids in the sequences and structures of native proteins. *Protein Science*, Vol. 4, pp. 2032–2039, 1995.

30) F. E. Cohen and I. D. Kuntz. *Prediction of Protein Structure and the principles of Protein Conformation*, chapter 17. Tertiary Structure Prediction. Plenum, 1989.

31) T. J. Richmond and M. Richards. Packing of alpha-helices:geometrical constraints and contact areas. *J. Mol. Biol*, Vol. 119, pp. 537–555, 1978.

32) A. G. Murzin and A. V. Finkelstein. General architecture of the alpha-helical globule. *J. Mol. Biol.*, Vol. 204, pp. 749–769, 1988.

33) A. V. Finkelstein and B.A. Reva. A search for the most stable folds of protein chains. *Nature*, Vol. 351, pp. 497–499, 1991.

34) A. Warshel and M. Levitt. Folding and stability of helical proteins:carp myogen. *J. Mol. Biol.*, Vol. 106, pp. 421–437, 1976.

35) T. Dandekar and P.Argos. Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.*, Vol. 236, pp. 844–861, 1994.

36) T. Dandekar and P.Argos. Identifying the tertiary fold of small proteins with different topologies from sequence and secondary strucutre using the genetic algorithm and extended criteria specific for starnd regions. *J. Mol. Biol.*, Vol. 256, pp. 645–660, 1996.

37) U. Carlsson and B-H. Jonsson. Folding of beta-sheet proteins. *Curr. Opin. Struct. Biol*, Vol. 5, pp. 482–487, 1995.

38) C. Chothia. One thousand families for the molecular biologist. *Nature*, Vol. 357, pp. 543–544, 1992.

39) J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein structure that fold into a known three dimentional structure. *SCIENCE*, Vol. 253, pp. 164–169, 1991.

40) D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, Vol. 358, pp. 86–89, 1992.

41) A. Godzik and J. Skolnick. Sequence-structure matching in globular proteins:application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 12098–12102, 1992.

42) Ken Nishikawa and Yo Matsuo. Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Engineering*, Vol. 6, pp. 811–820, 1993.

43) 川端猛. 情報理論による統計ポテンシャルを用いたタンパク質の構造予測. 修士論文, 1994.

44) K. Mizuguchi and N. Go. Seeking significance in three-dimensional protein structure compparisons. *Curr. Opin. Struct. Biol*, Vol. 5, pp. 377–382, 1995.

45) M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, Vol. 253, pp. 694–698, 1975.

46) M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, Vol. 104, pp. 59–107, 1976.

47) K.A.Dill, S. Brombberg, K. Yue, K. M. Fiebig, D.P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding - a perspective from simple exact model. *Protein Science*, Vol. 4, pp. 561–602, 1995.

48) A. T. Hagler and B.Honig. On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA*, Vol. 75, pp. 554–558, 1978.

49) T. Karasawa, K. Tabuchi, M. Fumoto, and T. Yasukawa. Development of simulation models for protein folding in a thermal annealing process - I.a simulation of bpti folding by the pearl necklace model. *CABIOS*, Vol. 9, pp. 243–251, 1993.

50) A. Aszodi and R. Taylor. Folding polypeptide alpha-carbon backbones by distance geometry methods. *Biopolymers*, Vol. 34, pp. 489–505, 1994.

51) A. Aszodi and R. Taylor. Secondary structure formation in model polypeptide chains. *Protein Engneering*, Vol. 7, pp. 633–644, 1994.

52) D. G. Covel and R. L. Jernigan. Conformation of folded proteins in restricted space. *Biochemistry*, Vol. 29, pp. 3287–3294, 1990.

53) D. G. Covel. Folding protein alpha-carbon chains into compact forms by monte carlo methods. *Proteins*, Vol. 14, pp. 409–420, 1992.

54) D. G. Covel. Lattice model simulations of polypeptide chain folding. *J. Mol. Biol.*, Vol. 235, pp. 1032–1043, 1994.

55) E. E. Lattman, K. M. Fiebig, and K.A.Dill. Modeling compact denatured states of proteins. *Biochemistry*, Vol. 33, pp. 6158–6166, 1994.

56) D.A.Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. *Proc.Natl.Acad.Sci.USA*, Vol. 89, pp. 2536–2540, 1992.

57) D.A.Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, Vol. 243, pp. 668–682, 1994.

58) J. Skolnick and A. Kolinski. Dynamic monte carlo simulations of globular protein folding/unfolding pathways I.six-member, greek key beta-barrel proteins. *J. Mol. Biol.*, Vol. 212, pp. 787–817, 1989.

59) A. Sikorski and J. Skolnick. Dynamic monte carlo simulations of globular protein folding/unfolding pathways II. alpha-helical motifs. *J. Mol. Biol.*, Vol. 212, pp. 819–836, 1990.

60) J. Skolnick and A. Kolinski. Simulations of the folding of a globular protein. *SCIENCE*, Vol. 250, pp. 1121–1125, 1990.

61) J. Skolnick and A. Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.*, Vol. 221, pp. 499–531, 1991.

62) A. Godzick, J. Skolnick, and A. Kolinski. Simulations of the folding pathway of triose phosphate isomerase-type alpha/beta proteins. *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 2629–2633, 1992.

63) M.J.Rooman, J-P A Kocher, and S.J.Wodak. Prediction of protein backbone conformation based on seven structure assignment. *J.Mol.Biol*, Vol. 221, , 1991.

64) K.Yue and K. A. Dill. Folding proteins with a simple energy function and extensive conformational searching. *Protein Science*, Vol. 5, pp. 254–261, 1996.

65) B. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J.Mol.Biol*, Vol. 249, pp. 493–507, 1995.

66) B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J.Mol.Biol*, Vol. 258, pp. 367–392, 1996.

67) A. Wallqvist and M. Ullner. Simplified amino acid potential for use in structure predictions of proteins. *Proteins*, Vol. 18, pp. 267–280, 1994.

68) S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures:quasi-chemical approximation. *Macromolecules*, Vol. 18, pp. 534–552, 1985.

69) S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable hight packing density term, for simulation and threading. *J. Mol. Biol.*, Vol. 256, pp. 623–644, 1996.

70) V. N. Maiolov and G. M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *J.Mol.Biol*, Vol. 227, pp. 876–888, 1992.

71) R.A.Goldstein, Z.A.Luthey-Schulten, and P.G.Wolynes. Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 4918–4922, 1992.

72) J. U. Bowie and D. Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA*, Vol. 91, pp. 4436–4440, 1994.

73) M. Vasquez, G. Nemethy, and H. A. Scheraga. Conformational energy calculations on polyupeptides and proteins. *Chem. Rev.*, Vol. 94, pp. 2183–2239, 1994.

74) A. Elofsson, S.M.Le Grand, and D. Eisenberg. Local moves : An efficient algorithm for simulations of protein folding. *Proteins*, Vol. 23, pp. 73–82, 1995.

75) J-P.A.Kocher, M.J.Rooman, and S.J.Wodak. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J.Mol.Biol.*, Vol. 235, pp. 1598–1613, 1994.

76) W. Kabsh and C. Sander. Dictionary of protein secondary structure:pattern recognition of hydrogen- bonded and geometrical features. *Biopolymers*, Vol. 22, pp. 2577–2637, 1983.

77) A. D. McLachlan. Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.*, Vol. 128, pp. 49–79, 1979.

78) S. Vajda and C. Delisi. *The Protein Folding Problem and Tertiary Structure Prediction*, chapter 13. An Adaptive Branch-and-Bound Minimization Method Based on Dynamic Programming. Birkhauser,Boston, 1994.

79) P. Kraulis. MOLSCRIPT:a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystal.*, Vol. 24, pp. 946–950, 1991.

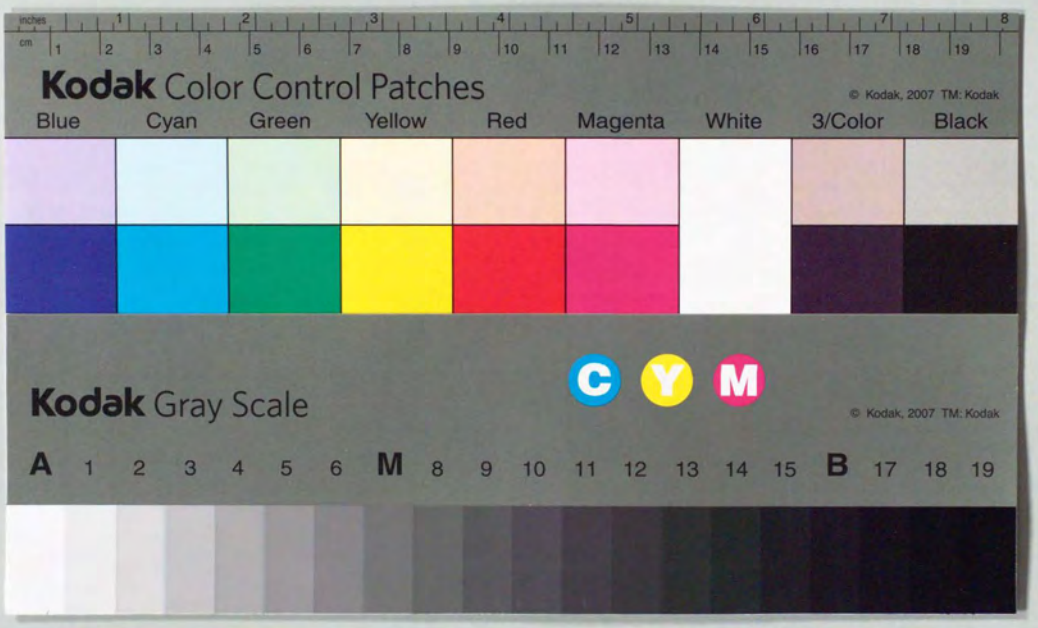
80) P.G. Wolynes, J.N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, Vol. 267, No. 1, pp. 1619–1620, 1995.

81) E.M. Boczko and C.L.Brooks III. First-principle calculation of the folding free energy of a three-helix bundle protein. *Science*, Vol. 269, No. 1, pp. 393–396, 1995.

82) R.H.Pain (崎山文夫監訳)（編）. タンパク質のフォールディング　原理・機構・応用. シュンプリンガー・フェアラーク東京, 1995.

83) G. E. Shultz, R. H.Shirmer(大井龍夫監訳). タンパク質 - 構造 機能 進化 -. 化学同人, 1979.

84) 江口至洋. タンパク質工学の物理化学的基礎. 共立出版, 1991.


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19