解　説

タンパク質の二次構造予測†

西　川　　建††

1. はじめに

親から子へ伝えられる遺伝情報は，DNA の塩基配列としてコードされた一次元のディジタル情報である．その情報は RNA に転写されたのち，タンパク質のアミノ酸配列へと翻訳される．遺伝情報は「生体の青写真」だといわれることがあるが，一次元のディジタル情報が実際に伝達されるのはここまでであり，それ以上の生体構造を指定するような情報は含まれていない．タンパク質は 20 種のアミノ酸が重合* してできた枝分かれのない高分子である．特定のアミノ酸配列（一次構造ともいう）に従って合成された鎖状のタンパク質分子は生体内で自発的に折りたたまれ，一定の立体構造（三次構造ともいう）をとる．立体構造はタンパク質の機能にとって必須であり，変性によって構造がくずれると機能は消失する．しかし，通常このプロセスは可逆的であり，温度や溶媒条件など外部条件を元に戻せば立体構造は再現し機能的にも再生する．結局，一次構造から三次構造への folding（折れたたみ）のプロセスを支配するのは，最初にインプットされた一次元の配列情報とタンパク質分子の置かれた環境条件である（実際には，生合成過程を含めた途中の“経路”に依存するという場合もあるが，ここでは考えない）．あるいは生理的環境を前提とすれば，一次構造が三次構造を決定するといってもよい．ただし，両者の関係は一対一対応ではなく，いわゆる同族タンパク質ではアミノ酸配列は相当大幅に変異しても立体構造は事実上変わらないことが知られている．つまり立体構造のほうからみると，同じ構造を実現しうるアミノ酸配列は一種類に限られず多数存在するという関係になっている．

では，あるタンパク質のアミノ酸配列が与えられたとして，対応する立体構造を予測することができるだろうか．これがタンパク質の構造予測問題にほかならない．この問題はわれわれの知的好奇心をそそるというばかりでなく，実際面でもその解決が強く望まれている課題である．なぜならタンパク質の一次構造と三次構造に関するわれわれの知識には，次のような大きなギャップが存在するからである．アミノ酸配列データは，DNA 塩基配列の迅速決定法の登場により容易に決められるようになり，タンパク質の配列データベースには現在 10000 件以上のデータが収録されている．一方，立体構造を求めるためにはX線結晶解析法や NMR（核磁気共鳴）法など，いずれにせよ手間と時間のかかる大変な作業を必要とする．これまでに立体構造が決定されているタンパク質は約 200 種類である．アミノ酸配列データはアルファベットの長々とした並びにしかすぎず，そのままでは生きた情報といえない．もしも一次構造から三次構造が予測可能になれば，配列データベース中の莫大なデータを生かすことができるのである．

タンパク質の構造予測の試みはすでに 20 年におよぶ歴史をもつ．その間，いくつもの方法論の開発と多様な試みがなされ，部分的な成功はあった．しかし，決定的な解決策は依然として得られていない．同じ期間に構造予測に欠かせないコンピュータの性能は飛躍的に向上し，また基礎データとなる一次および三次構造に関する実験データは大量に蓄積されてきたにもかかわらずである．たしかにタンパク質は複雑であり，とりうる可能なコンフォメーション（conformation）の数は事実上無限にある*．エネルギ計算を用いた分子シュミレーションなどによって最終的なコンフォメーションを予測することは，超高速コンピュータをもってしてもまだまだ不可能だといわれている．

† Secondary-Structure Prediction of Proteins by Ken NISHIKAWA (Protein Engineering Research Institute).

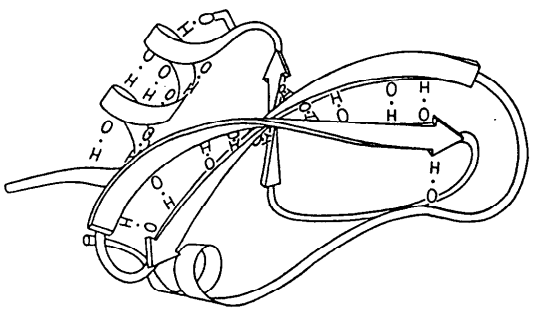
†† 蛋白工研

* 一個のアミノ酸が付加重合するさい 1 分子の水がとれる．したがって，重合したあとはもはや“アミノ酸”ではなくなり，“アミノ酸残基”と呼ばれる．

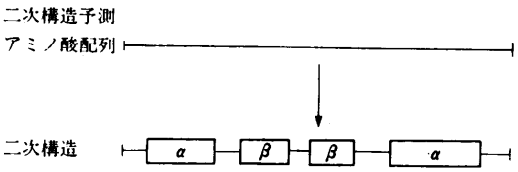
* タンパク質が 100 個のアミノ酸残基からなるとして可能なコンフォメーション（空間形態のこと）の数は N^{100} (N=2〜3) のオーダになる．

もっと現実的なアプローチとして広く試みられているのは，一挙に三次構造を目指すかわりに，一次構造から二次構造をまず予測し，次いで三次構造へ向かうという段階的方法である．ここでタンパク質の「二次構造」とは，α ヘリックスや β シートなど水素結合によって安定化された規則的な構造ユニットをいう(**図-1**(a))．本稿では第一段階の二次構造予測に焦点をあてるが，結論を先どりしていうなら，この問題は単に最初のステップというだけでなく，タンパク質の構造予測全体を代表し，かつそのエッセンスをなすひな型として重要な意味をもっている．二次構造予測を成功させることは立体構造予測に匹敵するほどの価値があるといっても過言ではない（立体構造予測に関しては文献 1)を参照されたい)．

二次構造予測は立体構造予測と比べると問題設定がずっと簡単である．予測すべき二次構造の種類は，α ヘリックス(=A)，β 構造(=B) またはそれ以外のコイル(=C)と呼ばれる状態のいずれかに限られる（さらに第4の状態として turn 構造を考える人もいる)．



(a) タンパク質の骨格構造の模式図
(J. S. Richardson[2]より改変)
主鎖のペプチド基の間に形成される水素結合 (C=O…H-N) によって安定化される構造を二次構造という．らせん状のαヘリックスと矢印で描かれた β 構造は規則的な二次構造であり，それ以外の不規則な部分はコイルと呼ばれる．なお，水素結合によってつながった一連の β 構造を β シートという．



(b) アミノ酸配列（一次構造）から対応する二次構造が推定できるだろうか？

図-1

つまり，一次構造と同様にディジタル化された系である．しかも一次元的であり，アミノ酸配列に沿ってαヘリックスと β 構造の領域をそれぞれ表示することができる(**図-1**(b))．したがって形式的にいえば，二次構造予測とは一次元の2組のディジタル変数列の間の対応関係を明らかにすればよい，という問題に帰着する*．図-1(b)の関係をみると何やら DNA とアミノ酸配列との対応にも似て，ごく簡単な問題だと思われるにちがいない．しかし，二次構造はあくまでもタンパク質の空間的な構造に由来する．以下に述べるように，上記の対応関係を明らかにすることは見かけ上の明快さとは裏腹に決して容易ではない．

2. 二次構造予測の実際

ごく素朴な考え方として，タンパク質を構成する20種類のアミノ酸残基はそれぞれ固有の“α ヘリックス形成能”(P_α) および“β 構造形成能”(P_β) をもつと見なすことができる．アミノ酸配列中のある特定の残基，R_i (R はアミノ酸の種類，i は N 末端から数えた残基番号を表す) がαヘリックス状態をとるかどうかは，その周辺の配列，たとえば自分を中心とした5残基の P_α 値の単純平均 $<P_\alpha>$ をとり，その値があらかじめ設定された閾値以上になるか否かによって決まると仮定する．このような計算を (β 構造に対しても同様に) すべての残基について行えば予測された二次構造が得られる．これが最もポピュラな予測法である Chou-Fasman 法[3),4)]の骨子にほかならない．実際にはこの上にいくつかのルールが加わるが，それでもなお全体として簡単な方法であり，計算機プログラムを書くまでもなく電卓だけで十分間に合う**．この方法が現在でもなおよく使われる理由がここにある．もう一つの理由は，予測能力がすぐれていて，他のもっと手の込んだ予測法と比較しても遜色がないばかりか，むしろ相対的に上まわるとされたからである．しかし，予測法の能力に対する安易な評価は禁物であり，初期に開発された予測法の多くはこの点に関し深刻な問題をはらむことが後になって明らかになった[5)]．

予測精度を調べるためには，X線結晶解析法などですでに立体構造（したがって二次構造も）が解明され

* β 構造のうちどれとどれが水素結合を形成し β シート (図-1(a)) をつくっているかというペアリングの問題は伝統的に二次構造予測の範囲外だとして無視されてきた．ここでもこの問題は考えない．

** 開発者の P.Chou はコンピュータをまったく使わず，電卓で計算した数値がびっしり書き込まれたノートをいつも持ち歩いていたそうである．

ているタンパク質を使ってテストする．定量的な評価法は何種類もあるが，以下では最も簡単で分かりやすい尺度，つまり A（α ヘリックス），B（β 構造），C（コイル）の3状態を仮定した上での正答率（正しく予測された残基の割合）を用いることにする．二次構造予測は 1970 年代に新しい方法が次々と開発されたが，どの方法が一番よく当たるのか実はだれにもよく分からなかった．その最大の原因として当時は構造既知のタンパク質の数が限られていたことがあげられる．たとえば，Chou-Fasman が P_α，P_β パラメータを求めるのに使ったタンパク質は 15（後に 29）種類にすぎなかった[3),4)]．実験データがこのように限られていたので，パラメータや予測ルールを抽出するタンパク質サンプルと的中率をテストするサンプルを区別せずに用いることがむしろ当然だった．その結果，多くの予測法が的中率 70% またはそれ以上という高い値を示し，1970年代末には二次構造予測は事実上解決済みだと見なされるに到った[6)]．ところが，その一方でX線結晶解析が徐々にではあるが着実に進展し，結晶構造の解明されたタンパク質は 1980 年代のはじめには100種類程度に達していた（ただし，そのすべてがデータベースに登録されたわけではない）．そのうちの年代的に新しい実験データだけを使うならば，それ以前に開発された予測法の真の能力を評価することができる．このような客観テストの結果，それまでの代表的な予測法はすべて，多種類のタンパク質に対する平均的中率で表して，55% 前後という水準にとどまることが明らかになった[5),7)]．二次構造予測の“意外な難しさ”がこうして認識されるようになり，あらためてこの問題への挑戦が再び行われるようになってきている．現状について述べる前に各種の予測法について次に概観しておきたい．

タンパク質の二次構造予測法は種類が多くすべてをあげることはできないが，代表的だと思われるものをその形式に従って分類すると**表-1** のようになる．第一のタイプは，既知のデータから抽出した統計量あるいは経験則を予測に用いる方法で，経験的方法と総称することができる．統計量を用いる方法の中では，単独のアミノ酸に数値パラメータを与えるもの (singlet)，二つのアミノ酸のペアを考えるもの (doublet)，三つの残基を同時に考えるもの (triplet) がある[6)]．先述の Chou-Fasman 法は singlet 型である．

“half-doublet” は2残基の対を考えるが，**図-2** に示すように予測される側の残基 (X_i) の種類は問わず，パラメータの値はもう一方の残基 (R'_{i+m}) の種類と中心残基 i との間隔 m に依存すると仮定する．残基の種類は 20 とおりであるから，必要なパラメータ数は $20\times(2M+1)$ となり，A，B，Cの3状態を考慮するとさらに3倍になる．ただし M は考慮に入れる片側の残基数（図-2 参照）を表す．Garnier-Osguth-

表-1 タンパク質二次構造予測法の分類

種　別	開発者
A. 経験的方法	
1. singlet	Chou-Fasman[3),4)]
2. half-doublet (neural network)	Garnier-Osguthorpe-Robson[8)]
	Qian-Sejnowski[9)]
	Holley-Karplus[10)]
3. doublet	Maxfield-Scheraga[12)]
	Gibrat-Garnier-Robson[13)]
4. triplet	Nagano[14)]
5. 配列ホモロジ	Nishikawa-Ooi[17)]
	Levin-Robson-Garnier[16)]
6. 経験則（エキスパートシステム）	Lim[20)]
	Cohen et al.[21)]
B. 統計力学的方法	
	Tanaka-Scheraga[22)]
	Wako-Saito-Scheraga[23)]
	Ptitsyn-Finkelstein[24)]
C. ジョイント法	
	Schulz et al.[25)]
	Argos-Schwarz-Schwarz[26)]

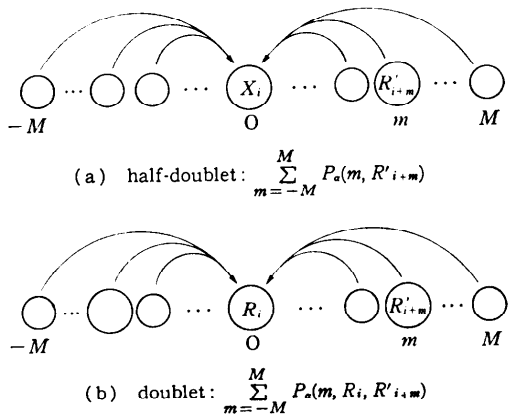


図-2 残基 i の α ヘリックス（β 構造またはコイル）形成能は自分自身とその前後の残基からの寄与の和で与えられる．“half-doublet”(a) では，中心残基から m 個離れた $i+m$ 番目の残基からの寄与は中心残基の種類に関係せず，その残基の種類 (R') と相対位置 ($+m$) にのみ依存すると考える．すなわち，その寄与は m と R'_{i+m} のみに依存した関数 $P_\alpha(m, R'_{i+m})$ で表される．それに対し，“doublet”(b)では，さらに中心残基の種類 (R) にも依存すると仮定する．全体で $(i-M)\sim(i+M)$ の範囲の $2M+1$ 個の部位を考慮する．

orpe-Robson 法[8]は $M=8$ であり，したがってパラメータの総数は $3\times20\times17=1020$ となる．近年，機械学習の論理である neural network 法が相ついで二次構造予測に応用されるようになった[9)~11)]．その中で明らかにされた重要な点は，入力層と出力層の間に挿入される隠れ層 (hidden layer) の効果が現れず，隠れ層を除き入力層と出力層を直接つないでも予測能力は変わらないという事実である[9)]．そして彼らによって最終的に与えられたアルゴリズムをみると，結局その形式はパラメータ数 $20\times(2M+1)$ の half-doublet 型にほかならないことが分かる[9),10)]．

doublet 型は両方の残基の種類を指定する (図-2 (b))のでパラメータ数は half-doublet よりさらに 20 倍多くなり，$M=4$ としても $3\times20\times20\times(2\times4+1)=10800$ となる．こうなると既知タンパク質の総残基数と同程度のオーダになってくる (100 種類のタンパク質で平均鎖長 200 残基とすると総数は 20000) ので，パラメータ値の統計的有意性が疑わしくなる．Maxfield-Scheraga[12] はこの点を考慮してパラメータの各項ごとにデータ数を見て，データ数の多いものは doublet 型，少ないものは half-doublet 型のパラメータ値を与えるようにした．近年発表された Gibrat-Garnier-Robson 法[13]も形式的には Maxfield-Scheraga 法と変わらず，同じように doublet, half-doublet の混合型パラメータセットを用いている．しかし，両者の間には 10 年余りの年月の差があり，その間に既知データ量が増加しているので，その分，Gibrat らの方法は彼ら自身の言うように予測能力の向上が期待できる．Nagano 法[14]は 3 残基からなる triplet 型のパラメータを用いているが，独立変数の数を減らすために 20 種類のアミノ酸を 7 つのグループに類型化し，さらに既知データ数を増やすために構造既知のタンパク質に相同なタンパク質も構造は同じであるとして用いている．

ホモロジ法は構造既知のタンパク質との 10 残基程度の長さの弱い配列上の類似性を利用する．triplet よりもさらに多い残基数を一度に考慮するという意味で表-1 のように分類されている[15]が，経験的パラメータを使う方法とは基礎となる考え方が異なる．ホモロジ法による予測については他の機会に解説した[16]のでここでは省略するが，筆者らの開発したもの[17]のほかに，Levin ら[18]と Sweet[19]の方法が発表されている．経験的方法の中で数値パラメータをまったく使わないものに Lim の方法[20]がある．これは既知データの中から経験的ルールを抽出し，多数のルールの組合せで予測を行う，今でいうエキスパートシステムの先駆的試みだといえる．もっと後になると AI (人工知能) を意識的に適用した Cohen らの方法[21]が登場する．ただし彼らは予測対象を turn 構造だけに絞っている．

以上の経験的な方法に対し，より理論的な予測法として統計力学的方法がある[22)~24)]．これは高分子鎖のヘリックス・コイル転移現象を説明する Ising 理論に基づいており，残基間の協同的なふる舞いが理論式に組み込まれている点ですぐれている．予測に使うときには経験的に求められたパラメータ値が使われる．最後のジョイント法はいくつかの個別の予測法を並行して用いる方法である[25),26)]．これについては次章で詳述する．

さて，以上のようなさまざまの予測法のうちどれが一番よく当たるのか気になるところである．相互比較を行うためには，まずそれぞれの予測法の計算機プログラムを用意する必要がある．われわれは表-2 に示す 8 種類を用意したが，それらは自前で開発したもの[17]論文を見てプログラムを書いたもの[4),9),20)]，原著者から譲り受けたもの[13),14)]，公開されているもの[8),24)]とさまざまである．テストの対象として PDB データバンク☆から 22 種類のタンパク質を選んだ．それらはいずれも 8 種類の予測法のパラメータやルールの導出には用いられてないものばかりである．22 タンパク質に対する平均値で表したテスト結果を表-2 に示す[27)]．8 つのうち一番成績の良いのは Gibrat らの方法であり，最も悪いのは Chou-Fasman 法であるが，差はそれほど大きくない．むしろ方法論の違いにもかかわらず，多くのものが的中率 60% 程度でほぼ一線に並んでしまうことに驚かされる．ただし，この一致はあくまで平均値についてのみ言えることであり，個々のタンパク質の的中率を比べると予測法ごとにかなり大き

表-2 二次構造予測法の比較

予測法	22タンパク質 (表-3(a)参照) に対する平均正答率
Gibrat-Garnier-Robson[13)]	62.4%
Ptitsyn-Finkelstein[4)]	61.0%
Nagano[14)]	61.0%
Nishikawa-Ooi[17)]	60.6%
Qian-Sejnowski[9)]	60.3%
Lim[20)]	56.5%
Garnier-Osguthorpe-Robson[8)]	56.2%
Chou-Fasman[3)~5)]	54.9%

☆ タンパク質結晶構造データを集めた Protein Data Bank の略称．

いバラツキが存在する．このように3状態評価による平均精度でおよそ 60% というのが二次構造予測の現状だと考えてよいが，この水準は決して十分なものとはいえない．われわれは個別の予測法を総合することにより改善を試みたので次に紹介する．

3. 新ジョイント法

複数の予測法を並用する方法をジョイント法と呼ぶことはすでに述べた．ジョイント法の論理は簡単であり，個別の予測法の与えた予測結果の多数決をとる．たとえば3種類の予測法を並用するとして，i 番目の残基に対する予測結果がそれぞれA，A，Bとするとジョイント法の結果はAとする．このような判断を残基ごとに行うが，3者の答えがA，B，Cとバラバラに分かれたときはあらかじめ決めておいた予測法（たとえば1番目）の結果に従うとする．ジョイント法の予測精度は全体としてみると個別の予測法の精度と大差ないことが知られている[26]が，残基ごとにみると予測の信頼度の高い部分と低い部分が区別できる．たとえば，A，A，Aとすべての方法が同じ答えを出したときは信頼性が高い．つまり，複数の予測法の“満場一致”による予測は“多数決”のそれよりも精度が高いのである．3種類の予測法を使った場合，満場一致とそうでない部分では平均精度で 10% のちがいがあった[17]．問題は満場一致で予測される残基の割合が限られる（タンパク質によってちがうが通常は全体の5割以下[17]）ことであり，もしもなんらかの方法でこの割合を高めることができるならば全体として予測精度は改善されるはずである．この一見して無理な注文に対して，小西[28]の示した解決策はまさにコロンブスのタマゴであった．それは，予測対象であるタンパク質に類縁なタンパク質のアミノ酸配列が既知であればそれらをなるべく多く利用すればよい，というアイデアである．このことを理解していただくには，“同族タンパク質”についての予備知識が必要であろう．

異なる生物種に由来する同種のタンパク質は相互に相同的であり，アミノ酸配列をたがいに並置（alignment）して比較することができる．並置された配列を見ると，ところどころアミノ酸が置換したり，ときには数残基の挿入・欠失が認められ，それらが共通の祖先タンパク質に由来し，分子進化の過程で変異したことが分かる．重要なことは，このようなアミノ酸配列上の変異にもかかわらず，一般に同族タンパク質の立体構造は非常によく保存されているという事実である．これまでに配列比較により同族タンパク質と判断されたもの（通常，全残基の 30% 以上が一致していれば同族だと見なせる）のうちで，立体構造（したがって二次構造も）が明らかに異なるものは一例も見つかっていない．このように一般的に成り立つルールを構造予測に使わない手はない．

具体例で説明すると，図-3 ではウシ膵臓リボヌクレアーゼAを予測対象として，それに類縁の他の哺乳類に由来する同じくリボヌクレアーゼの相同配列が並置されている．上記の一般ルールに従って，これらの二次構造は互いに一致すると仮定してよい．そこで各配列に対してジョイント法による予測を行い，図-3 のよ

```
Res. #          1~10          11~20          21~30          31~40          41~50
SEQUENCE KETAAAKFER QHMDSSTSAA SSSNYCNQMM KSRNLTKDRC KPVNTFVHES
(a)
X-ray 1  CCHHHHHHHH HHHCCCCCCC CCCHHHHHHH HHHHCCCCCC BBBBBBBBCH
(b)
X-ray 2  CCCHHHHHHH HHCCCCCCCC CCCCHHHHHH HHCCCCCCCC CCCBBBCCCC
(c)
Prediction CHHHHHHHH      CCCCCC CCCCCCC  HH HHHCCCCC
(d)
nrbo       HHHHHHH       CCCCCC CCCCC
nrprh      HHEHHHHH      CCCC   CCCC              C
nrgf       HHHHHHH       CC       CCC          C  CCC
nrder      HHHHHHHH      CCCCCC CCCCCCC             C
nrhp                     CCCCCC CCCCCC            CCC
nrgpa                    CCCCCC CCCC     HH HHHH    CC
nrbos      HHHHHHHH      CCCCCC CCCCC
nrwhk      C             CCCCCC CCCCCC
nrpg                     CCCCCC CCCC             CCCC
```

図-3　ウシ膵臓リボヌクレアーゼAの二次構造予測[28]

最上列のアミノ酸配列はアルファベットの一文字表記，それ以下の列の二次構造はヘリックス（H），ベータ（B），コイル（C）で表す．（a）X 線結晶解析の原著者によって判定された二次構造，（b）Kabsch-Sander[29] の自動判定法によって同定された二次構造，（c）二次構造予測の最終結果，（d）ウシ・リボヌクレアーゼ（nrbo）とそれに相同な8種類の配列データに対する“満場一致”予測の結果（ただし，紙幅の関係で N 末端部（1~50番）のみを示す）．

うに各配列の“満場一致”の部分をつなぎ合わせてゆけば，満場一致で予測される領域を増加させることができるというわけである．こうしてウシ・リボヌクレアーゼだけでは満場一致部分は全体の 24% にすぎなかったものが，全部で 23 種類（図-3 のもの以外に 15 種類を加えて）の相同配列を導入すると 75% まで上昇させることができた．このとき（満場一致部分の）予測精度はそれぞれ 81%，73% であり，予想どおり高水準を保つことが分かった[28]．

以上のように予測対象を類縁タンパク質にまで押し広げたジョイント法を“新ジョイント法”と呼ぶことにする．新ジョイント法の決定版をつくるために，われわれはあらためて個別の予測法の選定からやり直すことにした[27]．前章で述べた 8 種類の予測法の比較テストはその一環として行ったものである．表-2 の結果にもとづき平均精度 60% 以上の上位 5 つの予測法を選ぶことにした．ある対象タンパク質に対し相同配列が 9 本知られているとすると自分を含めて配列は合計 10 本になる．新ジョイント法ではそれぞれについて 5 種類の予測を行うことになるので，全体で 50 とおりの二次構造予測を総合することになる．ところが相同配列の間でたがいに矛盾するような予測結果になる場合も多いので，以下のように“優先度”を定義し優先度の順に予測結果を固定するようにした．

優先度は満場一致（5/5 と記す）予測が一番高く，以下 4/5，3/5（5 つの予測法のうちそれぞれ 4 または三つが一致する場合）の順に低くなり，同時に対象タンパク質に対する予測を優先させ，相同配列の間では対象配列とのホモロジが高いほど優先されるとした．したがって，最高の優先度をもつのは対象配列における 5/5 予測であり，最低は 5 つの方法がバラバラの結果（たとえば二つがA，別の二つがB，最後がC）を与える場合である．こうして最高（=9）から最低（=0）まで 10 段階の優先度を定義した[27]．前出の Konishi-Nishikawa[28] では満場一致部分だけを予測された領域だと見なしていたが，以上のように優先度を導入するとタンパク質の全領域についての予測を行うことができる．新ジョイント法による予測の一例を**図-4** に示す．この場合には既知の相同配列が 11 本あるので 60 とおりの二次構造予測をあらかじめ行ってある．残基ごとにみて優先度の最も高い結果を採用することにより図のような最終結果が得られる．

われわれは新ジョイント法の予測能力を評価するために 2 組のタンパク質データを用意した．サンプルA は前章で用いたものと同じで PDB データバンクから選んだ 22 種のタンパク質であり，サンプルB は PDB に未登録の 15 種類からなる．後者はX線結晶解析によって決定された構造が文献には発表されているが，原子座標データは PDB には登録されていないという意味である．PDB に登録されているものに対しては二次構造判定プログラム DSSP[29] により，水素結合の有無にもとづき，α ヘリックスや β 構造の領域を自動的に同定することができる．一方，PDB に登録されてなくても原著者によって文献中に二次構造の領域が明記されていることがある．その場合は，それを“予測の答え”としてテストに使うことができる．それがサンプルBである．

サンプルA，Bに対する新ジョイント法のテスト結果を**表-3** に示す．サンプルAの成績（**表-3**(a)）は非常によく，22 タンパク質の平均値は 69.0% である．従来のジョイント法と同様に相同配列データを使わなくても ($Q_{(-)}$) 64.8% に達している．われわれは当初目標を平均精度 70% に置いていた．上記の 69.0% という結果を得て事実上目標は達成したと判断した．そして国際会議でも自信をもって発表した[27]．**表-3**(b) のサンプルBを使ったテストは，実は後になって念のためにと思って行ったものである．この結果を見たときはショックだった．何かのまちがいだろうと思って

```
PROTEIN CODE AND NAME            : 1UBQ UBIQUITIN
NUMBER OF AMINO ACID RESIDUES    : 76
NUMBER OF HOMOLOGOUS PROTEINS    : 11

                 +         +         +         +         +         +         +
SEQUENCE  MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
X - RAY   CBBBBBBCCCCBBBBBCCCCCCAAAAAAAAAAAACCCCCCBBBBBCCBBCCCCCCCCCCCCCCCCBBBBBBCCCCC
          * **    **** **** **** *********    *****    * **  ********************  ***
PREDICT   CABBAACCCCCCBBBBBCCCCAAAAAAAAAACCCCCCCCAAAABCCCCCCCCCCCCCCCCCCCCBBBBBBAACCC
PRIORITY  5633337778878999699998899989777999999997747577988889997499999777777777753999
```

図-4 新ジョイント法の適用例

上から順にアミノ酸配列（一文字表記），X 線結晶解析による実測の二次構造，予測された二次構造，0～9 の優先度（本文参照）を示す．＊印は実測と予測が一致した部分．この例では 11 種類の相同配列データが予測に使われている．

表-3 新ジョイント法による二次構造予測
$Q_{(-)}$：相同配列データを使わないときの予測精度
$Q_{(+)}$：相同配列データを考慮したときの予測精度

コード(略号)	タンパク質名	残基数	相同配列データ数	$Q_{(-)}$ (%)	$Q_{(+)}$ (%)
(a) サンプルA					
1CTF	Ribosomal protein L7/L12	68	7	50.0	61.8
1LH1	Leghemoglobin	153	15	68.0	73.9
2CDV	Cytochrome C_3	107	4	76.6	72.0
2CTS	Citrate synthase	437	10	73.9	79.6
2WRP	*trp* Repressor	104	0	74.0	74.0
1ACX	Actinoxanthin	108	3	70.4	63.9
1HMG	Haemagglutinin A-chain	328	26	64.9	68.6
1HMG	Haemagglutinin B-chain	175	16	56.6	54.3
1FC1	Immunoglobulin FC fragment	207	62	61.8	62.8
1NXB	Neurotoxin B	62	88	66.1	66.1
1PSG	Pepsinogen	365	49	67.1	74.8
2ALP	α-Lytic protease	198	3	63.1	67.7
4RHV	Rhinovirus shell protein VP2	255	15	61.2	69.8
1ABP	l-Arbinose binding protein	306	3	57.2	60.8
1WSY	Tryptophan synthase β-chain	385	5	70.4	74.0
3PFK	Phosphofructokinase	319	7	66.1	70.0
3GAP	Catabolite gene activator protein	208	20	55.8	63.5
1UBQ	Ubiquitin	76	11	69.7	75.0
2CI2	Chymotrypsin inhibitor 2	65	5	53.8	66.2
2CPP	Cytochrome $P\text{-}450_{cam}$	405	11	68.9	72.3
2OVO	Ovomucoid 3-rd domain	56	33	58.9	62.5
6API	α-1-Antitrypsin	374	35	54.6	60.7
	平均			64.8	69.0
(b) サンプルB					
PPL-C	Phospholipase C	245	0	59.2	59.2
GWH	Porcine growth hormone	191	35	81.2	84.8
ILK-2	Interleukin-2	133	1	49.6	52.6
APP	HIV-1 Aspartyl protease	94	12	58.5	57.4
MDH-L	Methylamine dehydrogenase L-chain	121	1	56.2	59.5
ILK-1B	Interleukin-1 β	153	6	47.7	47.7
BLG	β-Lactoglobulin	162	5	38.3	44.4
ENL	Enolase	436	7	76.4	74.8
XYI	Xylose isomerase	388	5	59.3	59.5
DLH	Dienelactone hydrolase	236	1	60.6	61.0
THS	Thymidylate synthase	316	16	63.6	66.8
MCI	Muconolactone isomerase	96	1	50.0	51.0
PCD	Protocatechuate 3, 4-dioxygenase	200	2	48.5	53.0
HLA(1)	Histocompatibility antigen HLA-A2 chain 1	270	40	55.6	58.1
HLA(2)	Histocompatibility antigen HLA-A2 chain 2	97	9	61.9	69.1
	平均			60.1	61.8

いろいろチェックしたが，操作上のミスは見つからずこの結果は“ほんもの”だということになった．それまでの自信は一ぺんに崩れてしまった．

サンプルA，Bにおける違いは平均精度の大きな差のみならず，タンパク質ごとのバラツキがサンプルBではAよりも大きいことである．つまり，当たりはずれの差が大きく44%から85%（サンプルAでは54～80%）までの幅がある．対象ごとに変化が大きいと予測が不安定になり良くない．一方，各予測法の相対的な順位はサンプルA，Bで共通している．5種類の予測法のサンプルBに対する的中率は示してないが，平均値はすべて54～57%の範囲に入る．したがって，サンプルBに対しても新ジョイント法の成績が一番良いという点は変わっていない．相対的な関係は変わらないまま全体としてレベルが低下したわけである．構造予測はタンパク質ごとにやさしいものと難しいもの

がある．予測しにくいものが偶然かたよったという可能性もあるが，構造未知のタンパク質を相手にするときの状況はサンプルAよりもBに近いと予想せざるをえない．本来は第三者が行うべきテストをどうやら自分自身でやってしまったということらしい．

サンプルBの結果が未知タンパク質に対する現在の二次構造予測の実情を示しているとすると，その水準は相変わらず 60% そこそこにすぎないことになる．タンパク質の構造予測には原理的な無理があるのだろうか，最後にその点を考察しておきたい．

4. なぜ二次構造予測は難しいのか

新ジョイント法は代表的な予測法の中から良いものを選りすぐり，しかも類縁タンパク質の配列データを動員した予測法であり，いうなれば既存の方法とデータを結集した二次構造予測の“総合法”である．ところが，予測能力はどの個別の方法と比較しても相対的に上まわっているものの，すでにみたように的中率は 60% 余りにとどまる．この値はランダムな予測の水準（A，B，Cそれぞれの割合の二乗和として計算すると約 40% になる）と比べてそれほど高くない．図-5 はサンプルBの 15 タンパク質について，予測と実測の二次構造を直接比較して示したものである．α ヘリックスに比べて β 構造の領域は一般に短く，予測しにくいといわれている．しかし，図-5 を見ると予測の良し悪しはかならずしも，α 型，β 型，α/β 型などの folding タイプのちがいだけでは説明できないことが分かる．むしろタンパク質ごとに予測しやすいものと，しにくいものがあるとしか言いようがない．

二次構造予測の限界性として従来から指摘されているのは，予測法の論理には局所的な配列情報しか考慮されていないことである．球状タンパク質においては鎖に沿って遠い残基どうしが接触する，いわゆる長距離相互作用が起こる．すなわち，局所的な配列のみならず大局的な残基間の関係が構造形成と安定性に効いていることは間違いない．このような眼でもう一度，図-5 の結果をみてみると難しさの理由がある程度推定できる．たとえば，β-lactoglobulin（略号，BLG）は的中率 44% で極端に悪い．X線結晶構造を見るとαヘリックス1本と2枚のβシートからなる典型的なβ型のタンパク質であるが，予測ではβ構造の領域がことごとくといってよいほどαヘリックスと判定されている．しかも，その大部分は“満場一致”でヘリックスと予測されているのである．ところで，このタンパク質については興味深い実験が報告されている．変性状態にあるタンパク質に対して，外部条件を急速にジャンプさせて巻き戻り反応をみるという速度論的な folding 実験を行うと，β-lactoglobulin はその初期過程で β 構造に加えて相当量の α ヘリックスが形成され，その後時間とともにヘリックス含量は減少することが測定された[30]．したがって，図-5 の二次構造予測はまんざら“うそ”ではなく，β-lactoglobulin のアミノ酸配列はたしかに強いαヘリックス形成能をもっていると考えられる．しかし，タンパク質全体としてのエネルギ的バランスをみると大局的にはβシートのほうが安定であり，一度構造形成がβ型に傾くとなだれをうって最終構造に向かうものと想像される．

このような関係をもっと一般化すると次のようになる．タンパク質分子内の相互作用を（鎖に沿って近い残基間の）短距離力と（鎖に沿っては遠いが空間的には近い）長距離力に分けてみると，表-4 のように二つの場合が考えられる．一つは両方が矛盾なく協調的に働く場合であり，もう一つは反対に協調的でない場合である．いずれの場合もタンパク質の最終構造は長距離相互作用によって支配されているとすると，ケース1の二次構造予測は当たりやすく，ケース2はまちがった結果を与えることになるだろう．なぜなら二次構造予測は局所的な配列，すなわち短距離相互作用しか考慮していないからである．β-lactoglobulin は短距離力と長距離力が相反する最も極端な例だと思われるが，一般的にはもっと中間的なケースも考えられ，図-5 に見られるような予測しやすいタンパク質から難しいものまでの多様性が生じるのではないか．かつて，球状タンパク質ではアミノ酸残基間の短距離および長距離相互作用はつねに矛盾なく整合的に働くとする，“整合性原理”（consistency principle）とよばれる考え方が提唱された[31]．しかし，もしもそのような原理が一般的に成り立つのであれば，二次構造予測問題はこれほど難しくはないであろう．

分子全体としての大局的なエネルギバランスによってタンパク質の構造は一挙に決まるのだとすると，二次構造予測の難しさとは結局，立体構造予測の難しさと同根であるといえる．部分情報を順々に積み重ね

表-4 2種類の球状タンパク質

	ケース1	ケース2
短距離および長距離相互作用の関係	協調的	非協調的
二次構造予測	やさしい	難しい

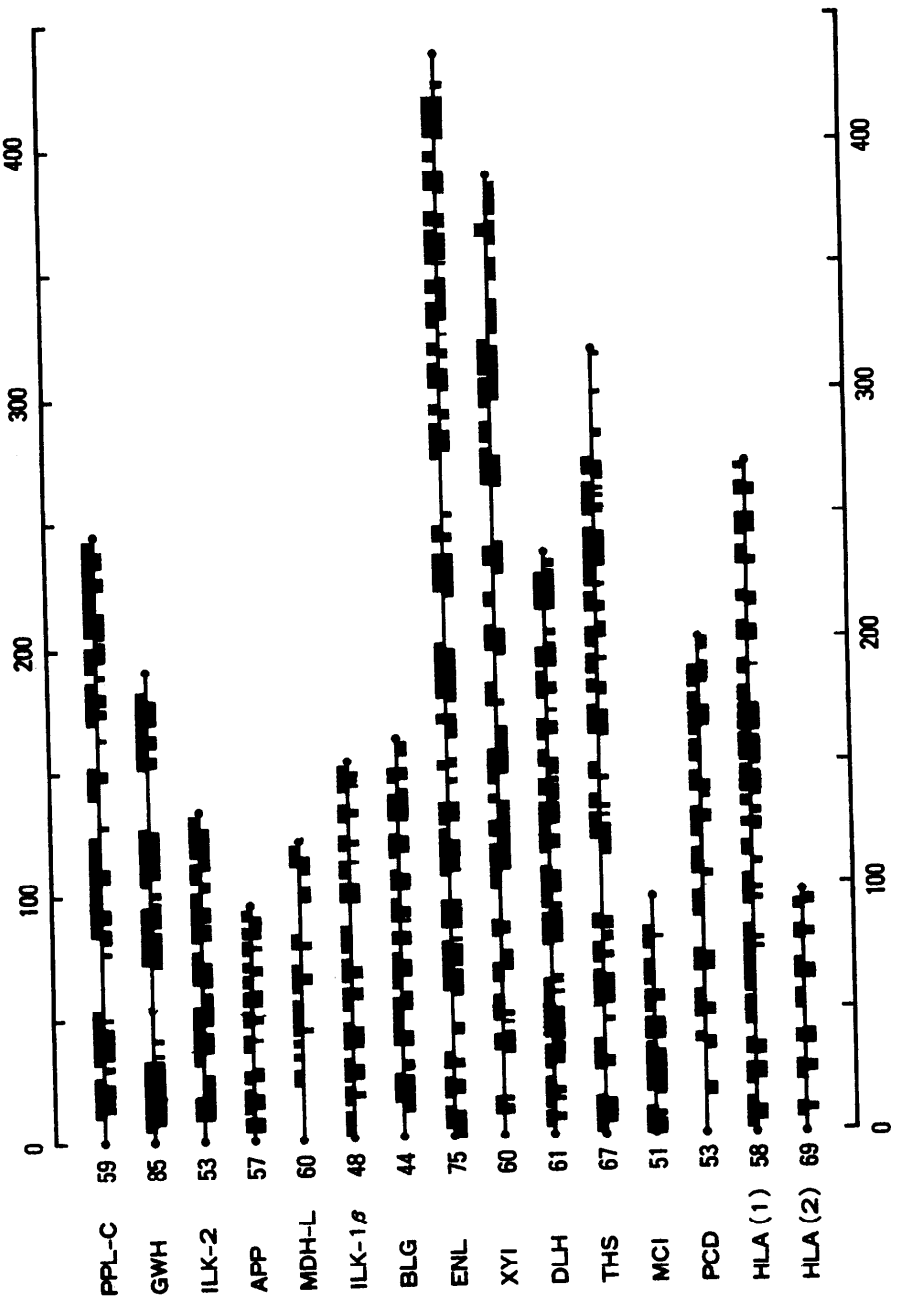


図-5 X線結晶構造と予測二次構造の比較

サンプルBの15種類のタンパク質（表-3(b)）について，α ヘリックス（黒）と β 構造（グレイ）の領域を残基番号に対して表示した．各タンパク質ごとに水平線の上段がX線構造，下段が予測二次構造，左端の数字は予測の正答率(%)を示す．タンパク質は上から順に α 型，β 型，α/β 型，α+β 型という folding タイプごとに大略まとめて並べられている．(α ヘリックスと β 構造が混在するものは α+β 型であるが，そのうち特に β 構造-α ヘリックス-β 構造という繰り返しユニットからなるとき α/β 型という)．

る行き方ではタンパク質には通用しない[32]．タンパク質の構造予測には長距離力を主とする分子全体の力関係をとらえる論理が要求されている．そのためには専門分野にとらわれないフレッシュな考え方の導入が望まれるのである．

参 考 文 献

1) 西川 建：化学増刊，No. 113, pp. 61-75, 化学同人 (1988).
2) Richardson, J. S.: Adv. Prot. Chem., 34, pp. 167-339 (1981).
3) Chou, P. Y. and Fasman, G. D.: Biochemistry, 13, pp. 211-245 (1974).
4) Chou, P. Y. and Fasman, G. D.: Adv. Enzymol 47, pp. 45-148 (1978).
5) Nishikawa, K.: Biochim. Biophys., Acta, 748, pp. 285-299 (1983).
6) Schulz, G. E. and Schirmer, R. H.: Principles of Protein Structure, Chap. 6, Springer-Verlag, New York (1979).
大井龍夫監訳：タンパク質，化学同人 (1980).
7) Kabsch, W. and Sander, C.: FEBS Lett., 155, pp. 179-182 (1983).
8) Garnier, J., Osguthorpe, D. J. and Robson, B.: J. Mol. Biol., 120, pp. 97-120 (1978).
9) Qian, N. and Sejnowski, T. J.: J. Mol. Biol., 202, pp. 865-884 (1988).
10) Holley, L. H. and Karplus, M.: Proc. Natl. Acad. Sci., 86, pp. 152-156 (1989).
11) McGregot, M. J., Flores, T. P. and Sternberg, M. J. E.: Prot. Engineer. 2, pp. 521-526 (1989).
12) Maxfield, F. R. and Scheraga, H. A.: Biochemistry, 15, pp. 5138-5153 (1976).
13) Gibrat, J.-F., Garnier, J. and Robson, B.: J. Mol. Biol., 198, pp. 425-443 (1987).
14) Nagano, K.: J. Mol. Biol., 109, pp. 251-274 (1977).
15) Schulz, G. E.: Ann. Rev. Biophys. Biophys. Chem., 17, pp. 1-21 (1988).
16) 西川 建：蛋白質・核酸・酵素，別冊 No. 29, pp. 79-89, 共立出版 (1986).
17) Nishikawa, K. and Ooi, T.: Biochim. Biophys., Acta, 871, pp. 45-54 (1986).
18) Levin, J. M., Robson, B. and Garnier, J.: FEBS Lett., 205, pp. 303-308 (1986).
19) Sweet, R. M.: Biopolymers, 25, pp. 1565-1577 (1986).
20) Lim, V. I.: J. Mol. Biol., 88, pp. 857-894. (1974).
21) Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. and Fletterick, R. J.: Biochemistry, 25, pp. 266-275 (1986).
22) Tanaka, S. and Scheraga, H. A.: Macromolecules, 9, pp. 168-182 (1976).
23) Wako, H., Saito, N. and Scheraga, H. A.: J. Prot. Chem., 2, pp. 221-249 (1983).
24) Ptitsyn, O. B. and Finkelstein, A. V.: Biopolymers, 22, pp. 15-25 (1983).
25) Schulz, G. E. et al.: Nature, 250, pp. 140-142 (1974).
26) Argos, P., Schwarz, J. and Schwarz, J.: Biochim. Biophys., Acta, 439, pp. 261-273 (1976).
27) Nishikawa, K., Noguchi, T. and Konishi, Y.: Proceedings of the Second International Symposium on Protein Engineering (in press).
28) Konishi, Y. and Nishikawa, K.: Bull. Inst. Chem. Res., Kyoto Univ., 66, pp. 378-385 (1989).
29) Kabsch, W. and Sander, C.: Biopolymers, 22, pp. 2577-2637 (1983).
30) Kuwajima, K., Yamaya, H., Miwa, S., Sugai, S. and Nagamura, T.: FEBS Lett., 221, pp. 115-118 (1987).
31) Go, N.: Ann. Rev. Biophys. Bioeng., 12, pp. 183-210 (1983).
32) 西川 建：生物物理，29, pp. 320-323 (1989).

(平成2年3月24日受付)