タンパク質の立体構造情報は どのようにアミノ酸配列に 埋めこまれているか？

国立遺伝学研究所 金城 玲

1. はじめに

生物の遺伝情報の実体はDNAの塩基配列であり、その直接的な翻訳がタンパク質のアミノ酸配列である。一方、生物の表現型は、多くの場合タンパク質の機能に還元されるが、その機能の発現はタンパク質の立体構造によって担われている。タンパク質は、遺伝子型から表現型への転換点であるという意味で、特異な存在である¹⁾。したがって、アミノ酸配列と天然構造がどのように対応しているのかを明らかにすることは、遺伝情報の発現を理解する上で基本的なステップである。配列と構造の対応関係を明らかにするための方法の1つとして、タンパク質の立体構造予測法²⁾の開発がある。近年の予測法の改善は目覚ましいが³⁾、一方で、その方法論が複雑に込み入っていて、素人には容易には理解しがたい面があることも事実である。そのため、仮に立体構造予測が完璧にできるようになっても、配列と構造の関係はよくわからないままになりかねない。問題はそれだけにとどまらない。仮に完璧な分子力場パラメータと十分強力な構造探索法があるとして、すると、分子シミュレーションによって立体構造予測はできるはずだが、それでは物理法則の正しさを再確認したことにしかならない。記号のパターン（情報）としてのアミノ酸配列と物理法則を結びつける、あるいは隔てる「何か」を探さなければ生物物理学としては物足りない³⁾。

そこで、筆者らは、より簡便に配列と構造の対応を理解するための1つの手段としてタンパク質の1次元構造に注目した。「1次元構造」というのは、アミノ酸配列の各残基の2次構造や埋もれ度など、立体構造の情報を何らかの形で残基ごとの特徴量または記号に変換

したものである⁴⁾。1次元構造は、立体構造の特徴とアミノ酸残基を1対1に対応させてみるができるため、直観的な把握には適している。しかし、1次元構造は3次元構造をいわば射影（粗視化）したものであるため、1次元構造は3次元構造を断片的にしか反映していないと考えられてきた。ところが最近、かならずしもそうでないことがわかってきた。1次元構造をうまく選べば、もとの3次元構造がほぼ完全に再現できることが示されたのである。本稿では、タンパク質の1次元構造を軸にして、配列と構造の対応関係とは何か、立体構造予測ができるとはどういうことなのか、を考えてみたい。

2. タンパク質の3次元構造は1次元的に表現できる

1次元構造からもとの3次元構造を復元しようとする試みは以前からあった。例えば、Saitohらは、1次元構造の一種である埋もれ度（後述）から天然構造のコンタクトマップを「予測」し、それからもとの天然構造を再現しようと試みたが、うまくいかなかった⁵⁾。

ところが最近になって、Portoらが興味深い結果を報告した⁶⁾。彼らは、タンパク質のコンタクトマップを固有値分解し、その最大固有値に対応するたった1つの固有ベクトル（「最大固有ベクトル」とよぶことにする）だけから、もとのコンタクトマップがほぼ完全に再現できることを示したのだ。コンタクトマップから立体構造を構築することが可能であることは以前から知られていた。最大固有ベクトルはアミノ酸配列と同じ長さの実数のベクトルであるので、これは一種の1次元構造と見なすことができる。つまり、Portoらの結果は、1次元構造から3次元構造が再現できることを示している。これにより、3次元構造の情報を損なうことなく、しかも直観的にアミノ酸配列と構造特徴量を対応づけられるという可能性が示唆された。しかし、まだ問題は残されている。最大固有ベクトルの幾何学的な意味が自明ではないからだ。筆者らが調べたところ、コンタクトマップの最大固有ベクトルの成分は、各残基の埋もれ度と高い相関があることがわかったのだが、最大固有ベクトルと埋もれ度の差が何を意味するのかはわからない。コンタクトマップを再現するためには、最大固有ベクトルを高い精度で与える必要があるのだ、この差は重要である。

そこで筆者らは、もっとわかりやすい1次元構造で3次元構造を再現できないか模索した。その結果、以下

How Is the Three-Dimensional Structure of a Protein Embedded in Its Amino Acid Sequence?

Akira R. KINJO

National Institute of Genetics

の3種類の離散的な1次元構造を指定すればよいことがわかった⁷⁾.

2次構造 (secondary structure) ^{*} α ヘリックス, β ストランド, それ以外, の3状態のいずれか.

埋もれ度 (contact number) i 番目の残基の埋もれ度 n_i は以下で定義される:

$$n_i = \sum_j C_{i,j},$$

ここで $C_{i,j}$ は, 残基 i, j が立体構造中でコンタクトしていれば1, そうでなければ0となる量 (コンタクトマップの成分).

残基毎コンタクトオーダー ^{*} (residue-wise contact order)

i 番目の残基の残基毎コンタクトオーダー o_i は以下で定義される:

$$o_i = \sum_j |i - j| C_{i,j}.$$

これらの中で, 2次構造と埋もれ度は古くから知られた量だが, 残基毎コンタクトオーダー (以下「RWCO」と略) は我々が新たに導入した量である. コンタクトオーダーという量はもともと Plaxco ら⁸⁾ がタンパク質のトポロジーとフォールディング速度の相関を調べるために導入した量だが, RWCOはそれを残基毎に分解したような量である. 遠距離コンタクトにかかわっている残基ほどRWCOの値 o_i は大きくなる傾向にある. 埋もれ度とRWCOはよく似た量であるが, どちらか一方だけでは天然構造はうまく再現できない.

長さが56残基から146残基までのさまざまなフォールドのタンパク質16種類について, 天然構造からこれら3種類の1次元構造を計算した後, 逆にこれらの1次元構造を満たす3次元構造を探索したところ, 元の天然構造が正確に (2-4 Å 以下の精度で) 再現された (Table 1) ⁷⁾.

以上の結果から, タンパク質の立体構造は1次元的に表現できることがわかった. これにより, アミノ酸配列と立体構造の関係を直観的に把握するための1つの視点が提供された. さらに, 1次元構造予測ができれば, 3次元構造も (原理的には) 予測できることになる. したがって, 3次元構造予測が不可能に思えた時代に, 改善の策として行われていた1次元構造予測にも, 理屈の上では3次元構造予測と同等の地位が与えられる.

3. 1次元構造予測ができるということは何か?

では, アミノ酸配列と1次元構造はどのように対応しているのだろうか? この問題は1次元構造予測の問題として捉えることができる. アミノ酸配列から1次元構造が予測できれば, その予測法の枠組みやパラメータを通して, 配列と構造がどのように相関しているかが

Table 1 Summary of 3D structures recovered from 1D structures^a

Protein (len) ^b	#/RMSD ^c		RMSD
	[0, 2)	[2, 4)	/min ^d
all α			
1r69 (63)	59	0	0.6
1utg (70)	100	0	0.6
256bA (106)	40	0	0.5
1mba (146)	8	2	0.5
all β			
1shg (57)	19	7	0.7
1csp (67)	13	4	1.2
1ten (89)	2	2	0.9
2pcy (99)	1	5	1.4
α/β			
2gb1 (56)	21	2	0.8
1ctf (68)	88	0	0.7
1vcc (77)	3	18	2.1
2acy (98)	2	2	1.0
135l (129)	0	16	3.4
α/β			
1ay7B (89)	20	7	0.8
1thx (108)	3	5	1.8
3chy (128)	2	2	0.5

^a For each protein, 300 random coil structures were generated which were then optimized so as to satisfy one-dimensional structural restraints. 100 lowest energy structures were selected for the statistics. See ref. 7 for details.

^b PDB identifier with sequence length in parentheses.

^c Number of structures resulted in the given range of RMSD (Å) from the native structure.

^d RMSD (Å) of the structure of the lowest energy.

理解できる. 一般にタンパク質の1次元構造予測では, まず予測の枠組み (線形回帰やニューラルネットワークなど) を定め, それに必要なパラメータを立体構造データベースの統計的学習により決定する. こうして得られたパラメータは, 配列と構造の間の (もしあるならば) 普遍的な関係を表現していると考えられる.

まず2次構造予測から見てみる. 典型的な2次構造予測法では, ある残基の2次構造を, 配列に沿った前後数残基の影響の重ね合わせの結果として予測を行う. 最近の方法では, 配列プロファイルとニューラルネットワークなどを組み合わせて77-78%程度の精度で予測ができる (Fig. 1a). (実は配列プロファイルを用いるところが曲者なのだがこれについては後述.) 一方, 埋もれ度も局所配列の影響の重ね合わせとしてある程度予測ができる⁹⁾. 2次構造予測と同様に配列プロファイ

Fig. 1 Histograms of prediction accuracy of one-dimensional structures for a database of 750 test proteins. (a) secondary structure (measured in Q_3 , the percentage of correctly predicted secondary structures); (b) contact number (measured in correlation coefficient, Cor); (c) residue-wise contact order (measured in correlation coefficient, Cor). See ref. 10.

ルと機械学習的な方法を組み合わせることで、正解との相関係数が0.73程度と比較的高精度の予測が可能である (Fig. 1b)¹⁰⁾。つまり、2次構造と埋もれ度は、局所配列における各アミノ酸残基の2次構造形成傾向や疎水性指標などといった普遍的な性質の重ねあわせによって、かなりの程度決定されるということになる。ところがRWCOに関しては、正解との相関係数で0.60程度の予測しかできず (統計的には有意ではあるが)、予測するタンパク質によって精度のバラツキが非常に大きいことがわかった (Fig. 1c)。この結果が2次構造や埋もれ度と同様の方法を適用してえられたものであるこ

とを考慮すると、RWCOに関しては、局所配列の普遍的な性質よりも、個々のタンパク質の個性が強く影響していると考えられる。この場合には、本来の意味での (データベースに頼らない) 第一原理からの予測が必要かもしれない。このように、アミノ酸配列から1次元構造が予測できるかできないかということから、配列とその1次元構造に一般的な関係があるかないかがわかる。2次構造と埋もれ度は (局所) アミノ酸配列とのある程度一般的な関係で決まっているが、RWCOはそうではなさそうだ。

ところで、典型的な1次元構造予測では配列プロファイル※ (位置特異的置換行列ともよばれる) と学習機械 (ニューラルネットやサポートベクトル機械など) を用いると書いたが、高精度の予測には、これらのうちどちらがより強く効いているのだろうか？ 少し試してみるとわかるが、配列プロファイルの方が圧倒的に重要である。例えば、2次構造予測では、PSI-BLASTのプロファイルを用いた純粋に線形な予測法 (線形回帰) でも75%程度の精度を得ることができる¹⁰⁾。これに対して、配列プロファイルを用いない場合には、複雑な学習機械を用いても68-69%程度の精度しかえられないことはよく知られている。前述のように、配列プロファイルとニューラルネットの組み合わせによる精度は77-78%なので、ニューラルネットなどの「高級な」方法による効果は高々2-3%程度に過ぎず、質的な改善はあまりないことがわかる。

そこで、配列プロファイルによって高精度の1次元構造予測が可能になるということが何を意味しているのかを考えてみよう。配列プロファイルを作るには、アミノ酸配列のデータベースに対して相同性検索を行い、進化的に類縁関係にある (相同な) アミノ酸配列を同定し、多重アラインメントを作成する。この多重アラインメントをもとにして、各座位でどのようなアミノ酸が出現しやすいかをスコア化したものが配列プロファイルである。アミノ酸残基の置換パターンはタンパク質の機能的制約はもちろん、立体構造によっても部位特異的に制約を受けるので、配列プロファイルは間接的に立体構造を反映していると考えられる。ということは、配列プロファイルを用いて構造を予測することは、はじめから暗に目標の立体構造を与えた上で、その構造を再構築することと見なすことができる*。構造予測の当初の目的は、配列がどのように構造を指定しているかを明らかにすることであったはずな

*このような「カンニング」は、タンパク質のフォールディングの研究で成功を収めている郷モデルと相通ずるものがある。

Fig. 2 Most successful structure prediction methods that are based on sequence profiles can be regarded as “structure reconstruction” methods.

のに、まさに予測しようとする構造によって規定された配列プロファイルを用いる方法が成功を収めているのは興味深い。構造予測問題が「構造再構築」問題になっていたのだ (Fig. 2)。このことは、少し考えれば概念的には明らかなのだが、実際の予測の手続き (進化情報の解析) として顕在化していることはあまり意識されていないように思われる。実は、この流れは1次元構造予測に限った話ではない。フォールド認識法でも現在では配列プロファイル間の比較にもとづく方法が主流である。いわゆる *ab initio* 予測においても、現在最も成功しているフラグメント・アセンブリ法では立体構造の断片を配列プロファイルにもとづいて選択することが多い。そもそも、このような手法がうまくいくのは、冒頭に述べたように、タンパク質が遺伝子型と表現型を兼備していることのひとつの現れなのだろう。

4. まとめ

タンパク質の3次元構造を復元できるような1次元的な表現を利用することにより、アミノ酸配列と立体構造の対応が直観的に把握しやすくなった。例として、各種1次元構造の予測の精度を測ることによって、アミノ酸配列と立体構造の関係には普遍的なもの (2次構造や埋もれ度) と個別的なもの (残基毎コンタクトオーダー) があることを具体的な量として知ることができた。

現在成功している1次元構造予測法の最も重要な処方 は配列プロファイルの使用にあることを確認し、その意味を吟味したところ、配列プロファイルを用いる構造予測法が、実は「構造再構築法」であるという認識に達した。配列プロファイル (あるいはその他の進化的情報) にもとづく構造予測法が究極的に成功するのかわかほかはまだわからない。もし成功するとすれば、タンパク質立体構造の構築原理 (配列と構造の対応関係) の理解には物理の法則と進化の規則が同時に取り扱える枠組みが必要になりそうである。そんなものが可能かどうかも含めて、タンパク質の立体構造の構築原理の理解は、いかにも生物物理学らしい問題である¹¹⁾。

謝 辞

ここで紹介した内容は、西川建教授 (国立遺伝学研究所) と堀本勝久教授 (東京大学) との共同研究、および Ugo Bastolla 博士 (Universidad autonoma de Madrid) との議論にもとづくものです。

文 献

- 1) 西川 建 (1994) 蛋白質—この絶妙な設計物 (赤坂一之編), 4, 79-102, 吉岡書店, 京都.
- 2) 千見寺浄慈, 高田彰二, 本野千恵, 富井健太郎 (2005) 生物物理 45, 165-167.
- 3) 木寺詔紀 (2003) 生物物理学とは何か (曾我部正博, 郷 信広編), 1-2章, 22-30. 共立出版, 東京.
- 4) Rost, B. (2003) Structural Bioinformatics (Bourne, P. E. and Weissig, H., eds.), 28, 559-587, Wiley-Liss, Inc., Hoboken, USA.
- 5) Saitoh, S., Nakai, T. and Nishikawa, K. (1993) *Proteins* 15, 191-204.
- 6) Porto, M., Bastolla, U., Roman, H. E. and Vendruscolo, M. (2004) *Phys. Rev. Lett.* 92, 218101.
- 7) Kinjo, A. R. and Nishikawa, K. (2005) *Bioinformatics* 21, 2167-2170.
- 8) Plaxco, K. W., Simons, K. T. and Baker, D. (1998) *J. Mol. Biol.* 277, 985-994.
- 9) Kinjo, A. R., Horimoto, K. and Nishikawa, K. (2005) *Proteins* 58, 158-165.
- 10) Kinjo, A. R. and Nishikawa, K. (2005) *BIOPHYSICS* 1, 67-74.
- 11) 永山國昭 (2003) 生物物理学とは何か (曾我部正博, 郷 信広編), 2-4章, 191-210. 共立出版, 東京.

タンパク質の立体構造情報はどのようにアミノ酸配列に埋めこまれているか？

金城 玲（きんじょう あきら）

情報システム研究機構国立遺伝学研究所生命情報DDBJ研究センター助手

連絡先：〒411-8540 三島市谷田1111

E-mail: akinjo@genes.nig.ac.jp