情 報 学 基 礎 27－6
（1992. 9. 8）

組合せ最適化問題としてのタンパク質 / RNA 構造予測

秋山 泰　金久 實

京都大学 化学研究所

ある種のタンパク質の構造予測、およびRNAの二次構造予測などにおいて、組合せ最適化のアプローチを採用することにより効率的に問題を解ける場合がある。このアプローチを取る上では、扱う問題が組合せ最適化問題として表現できるか否かが最も大きな鍵であり、つぎに強力な枝刈り手法または近似解法が導入できるかどうかが問題となる。組合せの総数を抑制できる場合には総当たり的なアプローチでも成功することがあるが、組合せの総数が多くなった場合でも、組合せ最適化の分野で知られている諸技法を採用することにより、ダイナミクス計算よりも短時間に大域的な安定解を得られる可能性がある。

Combinatorial Approach to Protein/RNA Structure Prediction

Yutaka Akiyama　Minoru Kanehisa

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto-fu 611, Japan

Structure prediction of proteins or nucleic acids is usually performed by means of dynamical simulation. However, in cases of membrane proteins like "bacteriorhodopsin", we can effectively employ combinatorial approach to the searching of feasible conformations. Even in case that the number of possible combinations becomes huge, an optimal (or semi-optimal) solution may be found within a realistic time consumption via applying known combinatorial optimization techniques, including use of Hopfield neural network. RNA secondary structure prediction is also suitable to the combinatorial approach.

1 組合せ最適化問題としての膜タンパク質構造予測

1.1 膜タンパク質構造予測の特殊性

細胞膜や細胞内膜系などの生体膜上に局在してはたらくタンパク質を膜タンパク質と呼ぶ。膜タンパク質は、自らを膜上にとどめておくための必要性から必ず何らかのメカニズムで膜との結合を保っている。膜の主体である脂質二重層との直接的な結合をもたず、膜上に存在する脂質や別の膜タンパクと結合することにより間接的に膜上に留まっているものもあるが、多くの膜タンパク質においては、自らの一部を膜に差し込んで、膜を1回もしくは折り返しながら複数回貫通することによって膜上に留まっている（膜貫通型タンパク[1]）。

このとき膜貫通領域は疎水性のアミノ酸を多く含み、溶液中に出ている領域は親水性のアミノ酸の含有率が高い。また膜貫通領域はαヘリックス構造をとることが多い。これらの特徴を利用すると、あるタンパク質が膜タンパクであるか否かを判定すること、およびどの部分が膜貫通領域であるかを推測する手がかりとなる。

とくに以下で扱うバクテリオロドプシン (bacteriorhodopsin) やハロロドプシン (halorhodopsin) などの場合は、膜を内外に往復するように7回も貫通することにより、タンパク質構造中の大部分が膜貫通領域となっており（図1参照）、これら複数の膜貫通領域が連係して作る膜内での立体構造を定めることがそのタンパク質の機能を推定するための重要なポイントになっている。

このような特殊性をもつ膜貫通型タンパク質の場合には、通常の球状タンパク質の立体構造予測の場合とは全く別のアプローチをとることにより、比較的容易にその構造を予測できるのではないかと期待される。美宅ら[2]は、膜タンパク質の構造予測について以下のような戦略を提案している。

1. 所与のタンパク質が膜タンパク質が否かを判別する。
2. 膜タンパク質である場合、膜貫通領域となる部位を推定する。
3. 膜貫通領域が複数ある場合、それらの相互の位置関係を推定することにより、膜内での立体構造を定める。
4. 膜貫通領域以外については、（必要があれば）球状タンパク質と同様の手法によりその構造推定を行なう。

ここで理想的には各過程は互いに協調しながら進められるべきであるが、以下では構造推定の各過程は逐次的に進めていくものとする。1) 膜タンパク質であるか否かの判定、および2) 膜貫通領域の推測までの過程は、タンパク質の一次構造（鎖に沿ったアミノ酸の並び方）上の特徴に注目したパターン認識的な手法を採用することが適していると考えられる。美宅らは膜タンパク質であるか否かの判定法として、タンパク質中のアミノ酸の疎水性指標が約30残基ごとの長周期波動を持つか否かを調べるフーリエ解析の結果と、タンパク質内の平均疎水性のデータとを組合せて判別分析にかける方法を提案し、約97%の精度[3]で判別を行なうことができた。

また膜貫通ヘリックスの推測には、ヘリックス部分における局所的な疎水性の高さと、3.6残基ごとの疎水性周期とを手がかりとすることができ、およそ80%の精度で予測ができる[4]。

以上の過程は比較的高い信頼性をもって推定が可能であるが、膜貫通領域が多数存在する場合には、これらの相互作用により膜内に作られる立体構造を推定する第三の過程がもっとも難しくなる。その方法について次節で触れる。

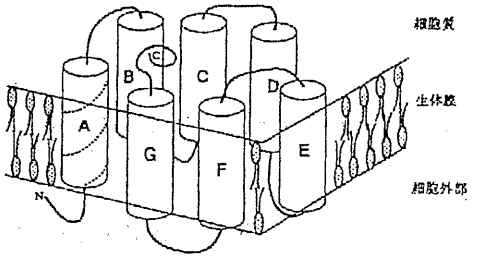

図1: バクテリオロドプシンの構造

1.2 ダイナミクス計算 対 組合せ最適化

複数の膜貫通ヘリックスの組合せで形成される膜内構造を推定するとき、ひとつの方法は各ヘリックスを独立な高分子とみなし、高分子間の相互作用によるダイナミクスを充分に長い時間をかけてシミュレートして、その安定な配置を探ることである。しかし各ヘリックスはアミノ酸が20～30残基、原子数で数百の大きさであるから、これらが多体で相互作用をする系のダイナミクス計算には膨大な計算量が必要となる。さらに初期配置に依存した多数のローカルミニマムの存在が予想され、膜内での最も安定な配置を正確に計算することはたいへんに困難である。

一方、本論文では組合せ最適化の考え方に基づ

く計算手法をとりあげる。ヘリックス間の相互作用のダイナミクスは計算せず、複数のヘリックスがとりうる位置関係を離散化された有限個の候補に限定した上で、それらの候補から最も安定性の高い配置を選択するというものである。この方法の利点は、位置関係の離散化の程度を粗くすればするほど、必要な計算量が減少させられることである。またダイナミクス計算に比べて多くの位置関係（ダイナミクス計算でいえば初期配置）を試すことができるためローカルミニマムによる障害が少なくなる。離散化の程度はある程度粗くしておいて、得られた解を初期配置としてダイナミクス計算により解を精密化することも可能である。

逆に欠点として、位置関係の離散化の程度が細かすぎる場合には、取り得る組合せの数が指数的に増えるため、計算量がむしろ増加することがある。組合せ最適化のアプローチがうまくいくための条件は、以下のようになる。

1. タンパク質が複数の部分構造に分かれており、各部分構造が取り得る状態がそれぞれ有限個の離散的状態として表現可能であること。

   （各部分構造が取り得る状態の候補が、他の部分構造の配置によって大きく影響を受けないことが望ましい）

2. （例えば膜タンパク質の物理的制約などにより）各部分構造が取り得る状態の数が比較的少なく、それらの組合せの総数があまり多くないこと。この場合には（適切な枝刈り技法を伴った）総当たり的探索で解が得られる。

3. 前項の条件が満たされず組合せの総数が極めて多い場合であっても、強力な近似解法や、Hopfield 型ニューラルネットによる探索などが適用可能で、最適解または準最適解を実用的な時間内で計算可能であること。

1.3 膜貫通ヘリックス間相互作用の単純化モデル

各ヘリックスのモデル化

ハロロドプシンにおいては、7つの膜貫通領域は全てαヘリックス構造（タンパク質のポリペプチド鎖の一部が規則的ならせん構造をとるもの。約3.6残基で一周）であると考えられており、それぞれ20～30残基の長さをもつ。これらの膜貫通ヘリックスの各々を、図2に示すような「正k角柱」でモデル化する。（以下のモデル化の原アイデアは美宅[2]によるものである。現在筆者らは美宅らと共同研究を進めており、計算モデルの詳細化と組合せ最適化アルゴリズムの開発を担当している。本節は筆者らが担当した部分の概略を報告するものである。）

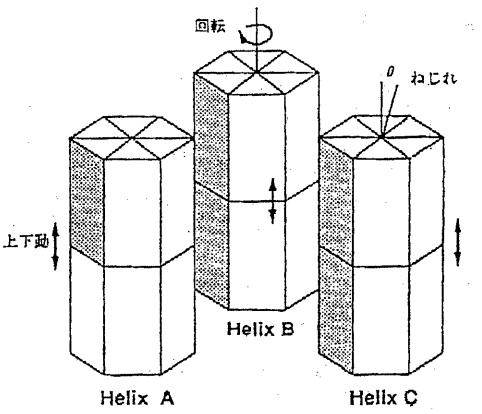

図 2: ヘリックスのモデル化

各ヘリックスを、角度についてk等分（図2では$k=6$）、上限方向にℓ等分（図2では$\ell=2$）した小断片の集まりとして表し、各断片ごとに異なった物理的特性を持つものとする。（各断片はその部分に含まれる原子の平均的特性を表すことになる）

各断片の特性を知るための方法としては、諏訪ら[5]の“Probe Helix 法”を用いる。これはヘリックスの各部分に、例えばセリンだけでできたヘリックス（Probe Helix）をある一定距離に接近させた時に生ずる力場を計算機実験によってもとめ、各断片部分の特性を調べるものである。得られるデータは、極性エネルギーE_{pol}、ファンデルワールス・エネルギーE_{vdw}に関する各断片付近の力場の大きさとなる。図2におけるヘリックスのモデル化の詳細度(k,ℓ)は、この過程でどれだけ細かくデータを取るかに依存している。

以下の計算では、各ヘリックスはあらかじめ定められた位置に置かれるものとし横方向の移動は考えない。また当初は垂直軸に対するねじれ運動や、上下のずれについても、簡単化のために無視するものとする。よって最も単純化されたモデルでは、各ヘリックスは垂直軸のまわりでの回転だけが許され、ヘリックス間の相互作用を通じて最も安定となる回転角度を求めることが構造予測の本質となる。このとき各ヘリックスの回転角度θ_iについても離散化を行ない、$\Delta\theta$ごとの運動に制限をする。（$\Delta\theta$は角柱の精度kとは無関係に定めてもよい）

相互作用のモデル化

各ヘリックスは、静電力とファンデルワールス力により相互作用する。簡単化のため、ヘリックスどうしは互いの最近接点においてのみ点-対-点の作用をするものとする。ただし$\ell > 1$の場合、上下ℓヶ所の力は合算される。このときヘリックス間の相互作用を次の2),3)のように制限することにより、計算時間を短縮することもできる。

1. 全てのヘリックス間に相互作用を許す。距離に応じた力が働く。
2. ある一定距離以下の相互作用のみ計算する。距離に応じた力が働く。
3. 円弧上で左右に隣接するヘリックス間のみに相互作用を許す。（距離は全て一定とする）

ヘリックスi,jの間の相互作用の強さは次式で定義する。ただしrは距離，γは比例係数，θ_i, θ_jは各ヘリックスの回転角である。

$$F(i,\theta_i,j,\theta_j) = \gamma \cdot \frac{E_{tot}(i,\theta_i) \times E_{tot}(j,\theta_j)}{r^2} \quad (1)$$

ここで$E_{tot}(i,\theta_i)$は最近接点の含まれる小断片における、"Probe Helix 法"で得た力場の値（E_{pol}とE_{vdw}の和）である。ふつう最近接点は小断片の中心ではないので、隣接する小断片との間でデータを内挿して、θ_iにおける値を定めている。（rが測定時とは異なる場合は(1)式ではファンデルワールス力の扱いに若干問題がある）

円配置モデル

N本の膜貫通ヘリックスがあるとき、その相対位置に関する知識が特にない場合には、円上に等間隔に配置する円配置モデル（図3）を用いる。図3において、各ヘリックスに描かれている$\vec{AQ}$などのベクトルは、各ヘリックスにおける最初の残基の位置を示し、角柱データの0度位置を表している。系全体が最安定となるときのこれらのベクトルの方向（$\theta_A = \angle PAQ$など）を求める事が計算の目標となる。

円配置モデルにおいては、ヘリックス間の最近接点の位置（ヘリックス上での0度位置からの角度）ϕは次式で決まる。ヘリックスAにおけるGとの隣接点を例にとると、

$$\begin{aligned} \phi &= f(\theta) \\ &= \angle QAG = \angle PAG - \angle PAQ \\ &= \frac{N+2}{2N}\pi - \theta_A \end{aligned} \quad (2)$$

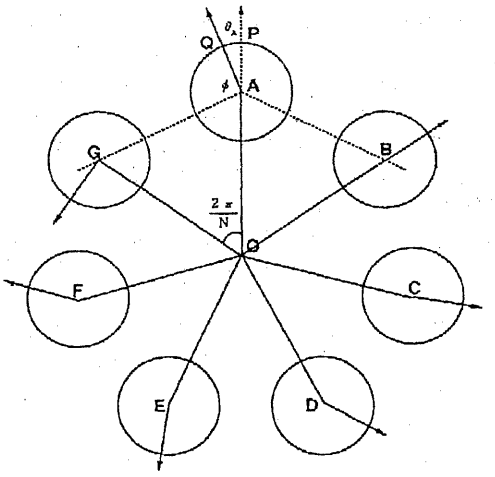

図 3: 円配置モデル

図3においては、円弧上に一次配列上の順序どおりにヘリックスを並べてある。このような仮定（以下、円弧上での順序性仮定と呼ぶ）が導入できるならば、計算は右まわりと左まわりの2通りの配置に対して行なうだけでよい。バクテリオロドプシンの場合はこのように順序性を保った配置をとると考えられている。円弧上での並び方と一次配列上の順序の間に何も仮定ができない場合には、$(N-1)!$通りの並び方のそれぞれについて各々の場合の最適安定解を求め、そらの内から妥当な解を選ぶ必要がある。$N = 7$の時、$(N-1)! = 720$となるから（特に枝刈りをしない場合には）必要な計算量は360倍に増加する。

楕円配置モデル

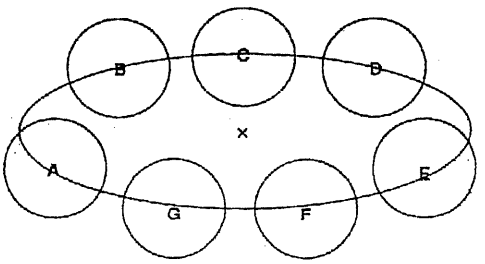

図 4: 楕円配置モデル

立体構造が良く調べられているバクテリオロドプシンの場合には、7本のヘリックスの配置は正円ではなく、一方向につぶれた形をとることが知られている。そこでバクテリオロドプシンの結果を参考

にして適切な離心率をもつ楕円を想定し、楕円上にN本のヘリックスを配置することも考えられる（図4）。楕円配置モデルを導入すると短辺方向と長辺方向で相互作用の距離が変わるので、円配置モデルとはやや違った結果になることも期待できる。円弧上での順序性が仮定できるときは$2N$通り、仮定できない時は$N!$通りの配置を試す必要がある。

1.4 Hopfield型ニューラルネットによる解法

前節で述べたうち最も簡単なモデルを用いても、複数のヘリックスが成す角度の組合せの数はかなり膨大になる。例えば、ヘリックスが7本($N = 7$)で、ヘリックスの回転角度の刻み幅を$1/12\pi$ (15[deg])とすると組合せの総数は、$\{2\pi/(1/12)\pi\}^6 = 4.6 \times 10^9$となる。円弧上の順序性が仮定できない場合にはさらに計算量が増えるから、効率的な枝刈りアルゴリズム、または近似アルゴリズムを導入することが必至である。

そこで著者らはHopfield型ニューラルネットを用いて、この問題の最適解もしくは準最適解を実用的な時間内で得ることをめざしている。Hopfield型ニューラルネットの原理については文献[6, 7, 8, 15]を参照されたい。

ヘリックスの本数がN本、回転角度の刻み幅が$\Delta\theta$のとき、$N \times (2\pi/\Delta\theta)$個のニューロンを図5のように用いて、解を表現する。

横軸のヘリックス番号iに対して、それぞれ縦の列内で1ヶ所のみのニューロンが活性状態になるよう調整されており、出力値が$V_{i,\theta_i} = 1$であるニューロンの位置がそのヘリックスの回転角を表す。

以上の動作をするネットワークを得るために、次のエネルギー関数を設計した。解への収束能力[13]の観点から各ニューロンには連続値出力モデル$0 \leq V_{i,\theta_i} \leq 1$を採用する。

$$
\begin{aligned}
E &= \frac{\alpha}{2}\sum_{i=1}^{N}\sum_{\theta_i=\Delta\theta}^{2\pi}(V_{i,\theta_i}-1)^2 \\
&+ \frac{\beta}{2}\sum_{i=1}^{N}\sum_{\theta_i=\Delta\theta}^{2\pi}\sum_{j=1}^{N}\sum_{\theta_j=\Delta\theta}^{2\pi}\frac{F(i,\theta_i,j,\theta_j)}{F_{average}}V_{i,\theta_i}V_{j,\theta_j} \\
&+ \frac{\alpha}{2}\sum_{i=1}^{N}\sum_{\theta_i=\Delta\theta}^{2\pi}V_{i,\theta_i}(1-V_{i,\theta_i}) \qquad (3)
\end{aligned}
$$

第一項は縦方向に1個だけのニューロンが活性状態であるとき最小値をとり、第二項は安定性の強い解を指向する働きがある。第三項は自己結合削除項[13]と呼び最終的な出力値を0,1に近付ける働きがある。現在のところ各項のバランスをとる係数は$\alpha = \beta$としている。収束性能を向上させるため勾配急峻化・過剰バイアス法等の技法[13]を用いる。

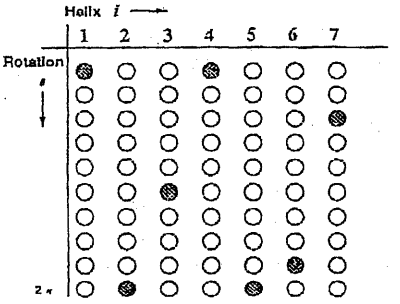

図5: Hopfieldネットの構成

現在このエネルギー関数を使用して膜貫通型タンパク質構造予測の評価を行なっている。その実験結果については報告を別稿にゆずる。

2 組合せ最適化問題としてのRNA二次構造予測

2.1 RNAの二次構造予測

RNAはリボヌクレオチドが連なって一本鎖構造を成した物質である。各リボヌクレオチドがとりえるの4通りの塩基(A・U・G・C)のうち、A-U間とG-C間（まれにG-U間）には水素結合が形成されるため、一本鎖が自然に折り畳まれて独自の物理的構造を作ることがある(folding)。RNAはタンパク質に比べれば構成要素が単純ではあるが、その立体構造予測はタンパク質と同様に困難であるため、それにかわって二次元平面内での安定なfoldingを計算する二次構造予測が広く行なわれている。

RNAの二次構造予測のアルゴリズムとしては、Zukerの方法[9, 10, 11]が有名である。これはRNA全体の安定性が最も高くなるような水素結合の組合せを求めるのに際し、動的計画法を採用したものである。動的計画法ではRNAの塩基長の三乗に比例する時間をかければ、必ず計算モデル上の最適解が得られるという利点があるが、塩基長が長くなると実用的な速度では解くことができず、またpseudo knotを効率的に扱えないなどの欠点があった。そこで筆者らはHopfield型ニューラルネットを用いた高速近似解法を提案した。この提案については既発表[14, 15]であるため、本節ではその概略を紹介するにとどめる。

2.2 RNA 二次構造予測のための Hopfield ネット

筆者らの方法ではまず簡単な文字列マッチ処理により、スタック領域（水素結合が連続的に連なる領域）を形成しうる相補鎖の候補を抽出する。次に Hopfield 型ニューラルネットにより、これらのスタック領域候補から、1) 選択したスタック領域の安定度の総計ができるだけ大きくなり、2) 選択した候補間に矛盾 (図 6) がない、ような候補を選び出す。

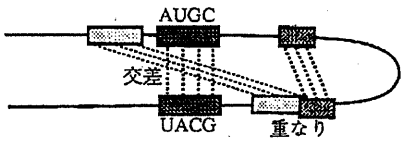

図 6: スタック領域の候補間の矛盾

目的 1) だけを達成するには安定性の高い候補をやみくもに加えていけばよい。しかし制約 2) により組合せが制約される。

この問題を解くためのエネルギー関数を次のように設計した。出力値 $0 \leq v_i \leq 1$ により、i 番目の候補が選ばれる ($v_i = 1$) か否 ($v_i = 0$) かを表現するとき、次式に基づいてネットワークを構成すれば、ニューラルネットは上記の 2 条件が満たされる解において安定する。

$$E = \sum_{i=1}^{n} e_i v_i + \lambda \cdot \frac{\max(|e_i|)}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} v_i v_j \quad (4)$$

e_i = (候補 i の表すスタック領域のエネルギー)

$$c_{ij} = \begin{cases} 1 & \text{候補 } i \text{ と 候補 } j \text{ が互いに矛盾するとき} \\ 0 & \text{候補 } i \text{ と 候補 } j \text{ が互いに矛盾しないとき} \end{cases}$$

n は前処理過程で抽出された候補の数であり、2 つの項のバランスをとる係数には $\lambda = 1$ を用いる。

いくつかの例題を通じて、この方法の性能を実験的に調べたところ [15]、Zuker 法と同等な解を Zuker 法よりも大幅に短い計算時間で求めることができた。筆者らは現在、pseudo knot の効率的な取り扱い、前処理アルゴリズムの高速化、最近の研究に基づくエネルギーパラメータの見直しなどの点で本手法の改良作業を進めている。

3 むすび

ある種のタンパク質の構造予測、および RNA の二次構造予測においては、組合せ最適化のアプローチにより効率的に問題を解ける場合がある。

表現の程度を粗くすることが許されれば、組合せの総数を抑制することができ、総当たり的なアプローチでも成功することがある。また組合せの総数が極めて多い場合でも、組合せ最適化の分野で知られている諸技法を採用することにより、ダイナミクス計算よりも短時間にグローバルミニマムを得られる可能性がある。

このアプローチを取る上では、扱う問題が組合せ最適化問題として表現できるか否かが最も大きな鍵であり、つぎに強力な枝刈り手法または近似解法が導入できるかどうかが問題となる。

謝辞 膜タンパク質の構造予測につき、日頃よりご助言とデータのご提供をいただいている東京農工大学工学部 美宅成樹 助教授、諏訪牧子 助手の両氏に感謝いたします。

参考文献

[1] Alberts *et.al.*: Molecular Biology of the Cell, Garland Pub., (1989).

[2] 諏訪、美宅、斉藤: 第 2 回公開ワークショップ「ヒトゲノム計画と情報解析技術」論文集, 166-169, (1991).

[3] N. Yanagihara, M. Suwa and S. Mitaku: Biophys. Chem., 34, 69 (1989).

[4] S. Mitaku, S. Hoshi and R. Kataoka: J. Phys. Soc. Jpn., 54, 2047 (1985).

[5] M. Suwa, S. Mitaku, K. Shimazaki and T. Chuman: Jpn. J.Appl.Phys.,31, 951(1992).

[6] J. Hopfield.: *Proc. Natl. Acad. Sci. USA*, 79, 2544-2558 (1982).

[7] J. Hopfield: *Proc. Natl. Acad. Sci. USA*, 81, 3088-3092 (1984).

[8] J. Hopfield and D. Tank: *Biol. Cybern.*, 52, 141-152 (1985).

[9] M. Zuker and P. Stiegler: *Nuc. Acids Res.*, 9, 1, 133-148 (1981).

[10] M. Zuker: *Methods in Enzymology*, 180, 262-288 (1989).

[11] M. Zuker.: *Science*, 244, 48-52 (1989).

[12] Y. Takefuji *et.al.*: *Biol. Cybern.*, 63, 337-340 (1990).

[13] 秋山, 古谷: 並列処理シンポジウム JSPP'91 論文集, 469-476 (1991).

[14] 秋山: 電子情報通信学会技術報告 NC90-62, 57-64 (1991).

[15] 秋山, 古谷: 1992 年情報学シンポジウム論文集, 125-134 (1992).