統計ポテンシャルを用いたタンパク質立体構造予測の検討

天白 進也

1　はじめに

タンパク質は自然環境化で一定の構造を持ち，その形状に即した科学的機能を発現する．そのため，タンパク質機能の理解や予測をするためには，その構造を知る必要がある．タンパク質の構造予測手法は，実験的手法とコンピュータシミュレーション手法に大別される．実験的手法には X 線構造解析，NMR などがあるが，高い精度で構造を予測することができる反面，人的，時間的コストが非常に高い．そのため，実験によらず，コンピュータシミュレーションによりタンパク質の立体構造を予測することが注目されている．

タンパク質は細胞内のリボソームで合成されるため，タンパク質が折りたたまれる過程をコンピュータシミュレーションで再現するには，細胞全体の系を考慮する必要があり，莫大な計算が必要であると考えられる．しかし，1960 年代初頭に行われた Anfinsen の実験により，タンパク質の自然の立体構造が自由エネルギー最小状態に対応することが示唆され，計算機シミュレーションによるタンパク質立体構造予測の可能性が示された．この原理に基づき，タンパク質のポテンシャルエネルギーを表現するさまざまなコンフォメーション関数（エネルギー関数）が開発されており，それを用いて，タンパク質の折りたたみ過程（フォールディング）をコンピュータシミュレーションしたり，タンパク質の構造予測を自由エネルギーの最小化問題ととらえ，最適化手法に関する研究がなされてきた．しかし，現在まで，標準的なサイズ（アミノ酸が 300 程度）のタンパク質を，実用的な精度，時間で予測可能な手法は確立されていない．これは，エネルギー関数の精度とタンパク質構造の系の自由度の大きさに起因するものであると考えられている．

これまで，われわれのグループでは，古典力学に基づくエネルギー関数を用い，自由エネルギー最小化問題としてタンパク質の立体構造予測を行ってきたが，やはり精度の面でよい結果が得られていない．これは，エネルギー関数の精度に起因するところが大きいと考える．そのため，他の評価指標を用いて多面的に構造を評価する必要性が出てきた．そこで，本報告では，二次構造と呼ばれるタンパク質の部分構造の予測結果，既知タンパク質構造の統計情報を基にしたエネルギー関数を用いて，古典力学に基づくエネルギー関数の評価を補間することをめざす．

2　タンパク質の構造

タンパク質は，20 種類あるアミノ酸がペプチド結合により鎖状に連なったポリペプチド鎖（一次構造）であり，タンパク質によってその並びは異なる．タンパク質はポリペプチド鎖が伸びた上体では存在せず，特定の形に折りたたまれて存在する．この折りたたまれた構造が立体構造（三次構造）であり，このときタンパク質は最も安定した状態となる．タンパク質の構造を考える場合，一次構造，二次構造，三次構造までの 3 つの階層で考えることができる．

2.1　二次構造

二次構造とは，10 残基から 20 残基程度のペプチド鎖で構成される規則的な折りたたまれ方のパターンであり，αヘリックス，βシート，ターンなどがある．（Fig. 1 は protein-G と呼ばれるタンパク質であり，αヘリックス，βシート，ターンを形成する）タンパク質の三次構造は二次構造の組み合わせであるといえる．

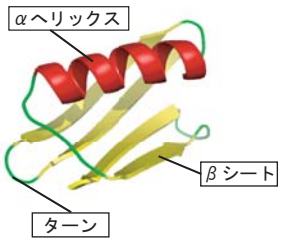
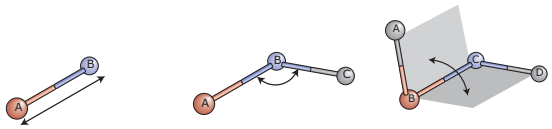


Fig.1　タンパク質の三次構造：Portein-G(出典：自作)

2.2　立体構造を定めるパラメータ

本研究で用いるエネルギー関数では，タンパク質分子を空間座標ではなく相対座標で表現する．したがって，立体構造を表現するパラメータとして，結合長 (Bond Length), 結合角 (Bond Angle), 二面角 (Dihedral Angle) を用いる．結合長とは 2 つの原子の間の長さ，結合角とは 3 つの原子の間の角度，二面角とは 4 つの原子についてはじめの 3 つの原子が作る面と後の 3 つの原子が作る面の間の角度である (Fig. 2 参照)．結合長，結合角はアミノ酸や結合の種類によりほぼ一定であるため，構造最適化の際には，二面角のみを変数として扱う．Fig. 3 に，あるペプチド鎖の立体構造を示す．図中のφ，ψ，ωが二面角にあたる．



(a) Bond length　(b) Bond angle　(c) Dihedral angle

Fig.2　立体構造を定めるパラメータ (出典：自作)

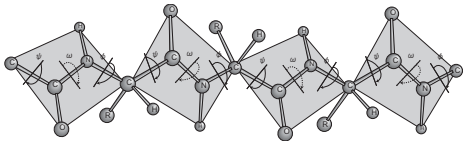


Fig.3　二面角とペプチド結合 (出典：自作)

3　エネルギー関数

本研究では，分子動力学計算プログラムパッケージである TINEKR[2] をもとに，名古屋大学の岡本らが手を加えたものをエネルギー関数として用いてきた．これは古典力学に基づくエネルギー関数である．エネルギー関数としては，この他，量子力学ポテンシャル，統計ポテンシャルを用いたエネルギー関数が開発されている．ここでは，古典力学ポテンシャル，統計ポテンシャルについて述べる．

3.1　古典力学ポテンシャル

分子の立体配座の安定性や配座間のエネルギー差を原子間に働く力によるポテンシャルエネルギーの総和によって計算する手法である．この手法では，分子の持つポテンシャルエネルギー E_{tot} は，タンパク質分子の構造エネルギー E_{conf} と溶媒和の自由エネルギー E_{solv} の和で与えられる（式 1)．

$$E_{tot} = E_{conf} + E_{solv} \tag{1}$$

$$E_{solv} = E_{GB} + E_{SA} \tag{2}$$

$$E_{conf} = E_{BL} + E_{BA} + E_{torsion} + E_{nonbond} \tag{3}$$

ここで，E_{BL} は結合長エネルギー，E_{BA} は結合角エネルギー，$E_{torsion}$ はねじれエネルギー，そして $E_{nonbond}$ はファンデルワールス力と静電相互作用からなる非結合のエネルギー項を示している．それぞれの項の詳細については省略する．また，溶媒和の自由エネルギー計算には generalized-Born/surface area (GB/SA) モデルを用いる．

3.2　統計ポテンシャル

統計ポテンシャルとは，既存の多数のタンパク質立体構造を統計解析することによってエネルギー関数を見積もる手法である．統計ポテンシャルとしては，contact number(埋没度) がよく用いられる．

3.3　埋没度統計ポテンシャル

埋没度とは，ある残基が埋もれている度合いをあらわす量であり，周辺残基数（一定範囲内にあるアミノ酸残基の個数)，あるいは溶媒接触表面積で見積もることができる．

そこで，X 線構造解析や，NMR によって構造同定されたタンパク質立体構造のデータベースである PDB(Protein Data Bank)[1] を用いて，立体構造データベース中の多数の構造を統計解析し，アミノ酸残機種ごとの埋没度の確率密度分布 $P_a(n)$ を求める．ただし a はアミノ酸残基種，n は周辺残基数である．

3.4　Verify 3D

Verify3D は，統計ポテンシャルに基づくタンパク質の構造評価プログラムであり，カリフォルニア大学ロサンゼルス校 (UCLA)DOE 研究室で開発が進められている．[3] Verify3D では，タンパク質のアミノ酸残基が置かれている構造環境を，残基の埋没度と極性原子との接触面積から定義される 6 通りの環境と 3 通りの二次構造（α ヘリックス，β シート，その他）をもとに 18 通りで規定し，構造既知タンパク質のアミノ酸残基の確率密度分布を用いてスコアを計算する．[6]

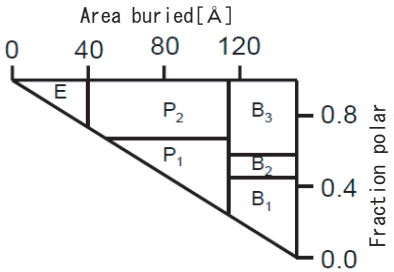


Fig.4　スコアマップ (出典：参考 [6] から引用)

Verify3D の実行ステップを以下に示す．

1. 各残基について，溶媒接触表面積を計算（Area buried)
2. 各残基の表面積のうち，極性原子（溶媒を含む）に覆われている割合を Fraction polar とする
3. Fig. 4 から，各残基 (i) がどの環境クラスに属するか (j) を決定し，環境 j に残基種 i を見出す条件付き確率 $P(i|j)$ を求め，次式により，その残基種のスコアを決定する．

$$Score = P(i|j)/P(i) \tag{4}$$

残基間の影響を無視すると，全体構造の評価値は各残基のスコアの和で表せる．

4　タンパク質の立体構造予測

われわれのグループでは，古典力学に基づくエネルギー関数を用い，自由エネルギー最小化問題としてタンパク質の立体構造予測を行ってきた．最適化手法としては，遺伝的交叉を用いたシミュレーテッドアニーリング (Parallel Simulated Annealing with Genetic Crossover:PSA/GAc)[4] を用いている．

4.1　遺伝的交叉を用いたシミュレーテッドアニーリング (PSA/GA)

PSA/GAc では Fig. 5 のように，並列に実行している SA の解の伝達時に遺伝的アルゴリズムのオペレータである遺伝的交叉を用いる．

遺伝的交叉では，親としてランダムに 2 固体を選択し，設計変数間交叉を行う．そして，親固体と生成された子固体を合わせた 4 固体から，エネルギー関数値の良好な

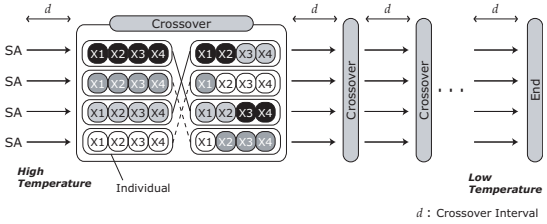
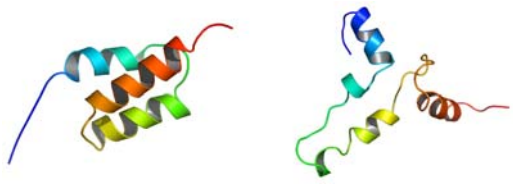


Fig.5　PSA/GAc の概要 (出典：自作)

2 固体を選択し，次の探索点とする．PSA/GAc では小規模なタンパク質（20 残基程度）の構造予測に対する有効性が示されているが，より大規模な問題に対する有効性は示されていない．

4.2　大規模なタンパク質への PSA/GAc の適用

60 残基からなる protein-A に PSA/GAc を適用した．Fig. 6 に Protein-A の天然構造と実験で得られたエネルギー最小構造を示す．また，天然構造のαヘリックス形成残基位置と，30 構造の二次構造 (αヘリックス) 形成率を Fig. 7 に示す．横軸はアミノ酸残基番号，縦軸は 30 試行中のαヘリックス形成数である．なお，protein-A の天然構造は，10 から 19 残基，25 から 37 残基，42 から 55 残基にαヘリックス構造を形成する．



(a) 天然構造　　(b) 最小エネルギー構造

Fig.6　Protein-A を PSA/GAc で最適化した結果 (出典：自作)

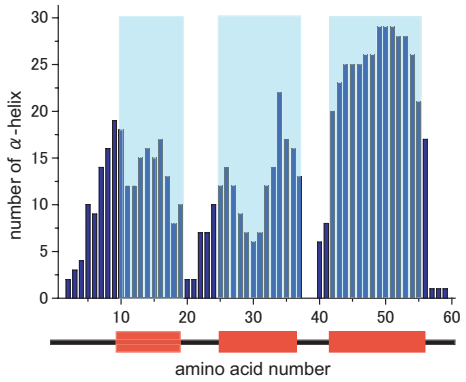


Fig.7　α-helix の出現確立 (出典：自作)

Fig. 6 より，PSA/GAc により最適化した構造は，全体的には天然構造と異なるが，部分的に類似した構造を形成していることがわかる．また，Fig. 7 からも比較的天然構造と同じ残基位置にαヘリックスを形成していることがわかる．

5　提案手法

前章より，大規模なタンパク質に PSA/GAc を適用した場合，全体的には天然構造と異なるが，部分的には類似した構造を形成していることがわかった．

そのため，部分構造最適化によってタンパク質の立体構造を予測する試みもなされている．[5] これは，タンパク質を 10 残基程度の部分に分割して最適化を行った後，それらをつなぎ合わせて全体構造を形作る手法である．二次構造などのタンパク質の部分構造は，距離的に近いアミノ酸残基との相互作用により形成される場合が多いため，この手法により探索の効率化が期待される．

しかし，部分構造を組み合わせることは容易ではなく，最適化した構造をそのままつなぎ合わせると，部分構造どうしが衝突し，古典力学ポテンシャルを用いたエネルギー関数は極端に悪い値を示す．そのため，部分構造最適化による結果を全体構造の形成にうまく反映することができていなかった．

そこで，古典力学ポテンシャルのような厳密な評価関数ではなく，あらい評価指標である統計ポテンシャルを用いれば，部分構造最適化の結果をうまく生かして全体構造を形成できると考えられる．また，統計ポテンシャルではタンパク質らしさを評価するため，古典力学ポテンシャルによる場合よりも解探索性能の向上が見込まれる．

以上より，本研究では，以下のステップにより構造探索を行う．

STEP 1　既探索構造を用いた二次構造の予測
　PSA/GAc を用いてタンパク質構造を最適化する．50 試行ほど試行し，既探索構造アーカイブを作成する．また，残基ごとの二次構造形成率を求める．

STEP 2　部分構造の分割と統計ポテンシャルによる最適化
　STEP1 で得た構造について，10 残基程度の部分構造に分割し，部分構造を組み合わせて全体構造を形成し，統計ポテンシャルによる最適化を行う．なお，統計ポテンシャルの計算には Verify3D を用いる．

STEP 4　古典力学ポテンシャルによる精密化
　STEP3 より得られた構造について，古典力学ポテンシャルで最適化することで精密化する．

STEP 5　STEP 1から4を繰り返す
　STEP 4 で得られた構造を既探索構造アーカイブに追加する．STEP 1から4を繰り返す．

6　予備実験

ここでは，統計ポテンシャルの性能検証のため，PSA/GAc によって最適化した構造と天然構造との評価値の差を比較する．対象問題には，56 残基の protein-G タンパク質を用いた．PSA/GAc により 10 試行計算を行い，エネルギーの低い１０個体を抽出した．また，評価値の計算には Verify3D を用いた．Fig. 8 は，天然構造と PSA/GAc によって最適化した 10 固体について，そ

れぞれアミノ酸残基ごとのスコアの平均をプロットしたものである．縦軸が Verify3D のスコア，横軸がアミノ酸残基番号である．

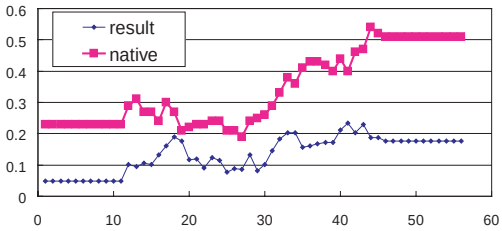


Fig.8　実験結果 (出典：自作)

Fig. 8 より，天然構造がすべての残基について高いスコアを得ていることが確認できる．つまり，統計ポテンシャルを用いて構造を最適化することで天然構造に近い構造を得られる可能性がある．

7　まとめ

古典力学ポテンシャルの最小化による方法のみでは，大規模なタンパク質の構造を正確に予測することが困難である．しかし，二次構造などの部分構造については，比較的天然構造に類似した結果が得られるようになってきた．また，部分構造の最適化結果を用いて，全体構造を予測する手法について，従来の古典力学ポテンシャルによるエネルギー関数では，部分構造どうしが衝突することにより，全体構造をうまく求めることができない．本報告では，古典力学ポテンシャルよりもあらい構造評価関数である統計ポテンシャルを用いたエネルギー関数を用いて，部分構造を組み合わせて全体構造を求める手法を提案した．また，統計ポテンシャルの性能評価のため，天然構造と PSA/GAc により最適化したタンパク質構造について，統計ポテンシャルによるスコアを比較した．その結果，すべてのアミノ酸残基について，天然構造が良好な値を示すことが確認された．そのため，統計ポテンシャルを最小化することができれば，より天然構造に近い構造を予測することが可能である．

その他の統計ポテンシャル関数の調査と検証．また，提案手法の実装および性能検証が今後の課題である．

参考文献

1) The RCSB Protein Data Bank. http://pdb.protein.osaka-u.ac.jp/pdb/index.html.
2) Tinker. http://dasher.wustl.edu/tinker/.
3) Verify3d. http://www.doe-mbi.ucla.edu/Services/.
4) 廣安知之, 三木光範, 小掠真貴．遺伝的交叉を用いた並列シミュレーテッドアニーリング．第 44 回 システム制御情報学会 研究発表講演会講演論文集, pp. 113–114, 2000.
5) 宇野尚子．部分構造最適化の組み合わせによるタンパク質立体構造予測の提案．同志社大学大学院 工学研究科 知識工学専攻, 2004.
6) 東京大学大学院農学生命科学研究科アグリバイオインフォマティクス．　バイオインフォマティクスリテラシー．　http://www.iu.a.u-tokyo.ac.jp/lectures/literacy1/050606.pdf.