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によるタンパク質立体構造予測

宇野 尚子
Naoko UNO

1　はじめに

近年，タンパク質の立体構造予測が注目されている．タンパク質はアミノ酸が複数連なって構成される物質で，自然界ではある決まった構造に折りたたまれた状態で存在している．この構造をタンパク質の立体構造と呼び，タンパク質の機能と密接に関わっていると言われている．そのため，立体構造を解明することによって病理の解明や新薬の開発につながることが期待されている．

タンパク質はエネルギーの低い安定した構造に折りたたまれるので，エネルギー最小化問題と捉えることができる．本研究室では「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 (Parallel Simulated Annealing with Genetic Crossover：PSA/GAc[1]）」を用いて立体構造予測を行っている．本報告では，タンパク質の立体構造予測の難しさと，PSA/GAc が抱えている問題点を挙げ，今後の研究方針について述べる．

2　タンパク質立体構造予測

2.1　エネルギー関数

目的関数は，タンパク質の系をモデル化したエネルギー関数を用いる．タンパク質のエネルギー関数は非常に複雑で，Fig. 1 のように大域的にいくつかの，局所的に無数の極小値を持つと考えられている．

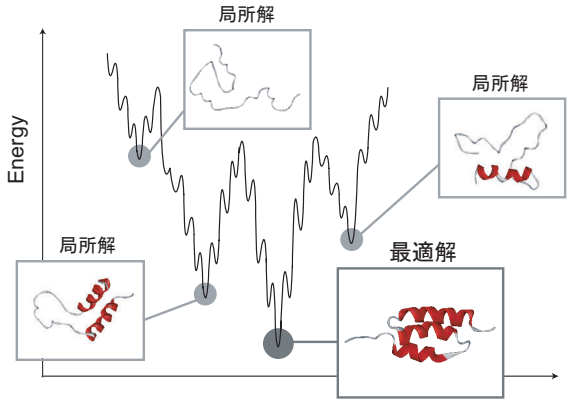

Fig. 1　エネルギー関数の模式図

本研究では，TINKER[2]という分子動力学計算プログラムパッケージを元に名古屋大学の岡本先生が手を加えたものをエネルギー関数として使用する．設計変数は，原子間の回転角である二面角を用いる．

2.2　予測結果の評価基準

タンパク質の立体構造予測を行う際に，評価する基準は 2 つある．1 つはエネルギー値，もう 1 つは構造の形である．構造が既知のタンパク質ならば，シミュレーション結果がどの程度その構造に似ているかが重要となる．

2 つの構造の差異を定量化するために用いられる量が RMSD(Root Mean Square Deviation) である．RMSD は 2 つの分子構造を重ね合わせて，対応する各原子のずれの二乗を平均したものの平方根で定義される．式 (1) に RMSD の求め方を示す．

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - b_i)^2} \qquad (1)$$

RMSD の単位は Å で，値が小さいほど 2 つの構造がよく似ていることになる．

2.3　立体構造予測の現状

タンパク質の立体構造予測問題は，CASP6[3]という世界的なタンパク質の立体構造予測コンテストが行われるほど盛んに行われている研究である．現在の最も大きな問題は，タンパク質のエネルギー関数が完璧ではないということである．現段階では「天然構造のエネルギーは最小にならない」エネルギー関数を使わざるを得ない状況にある．Fig. 2 は Protein-A というタンパク質をエネルギー最小化したときの履歴である．縦軸がエネルギー値 (kcal/mol)，横軸が RMSD(Å) である．

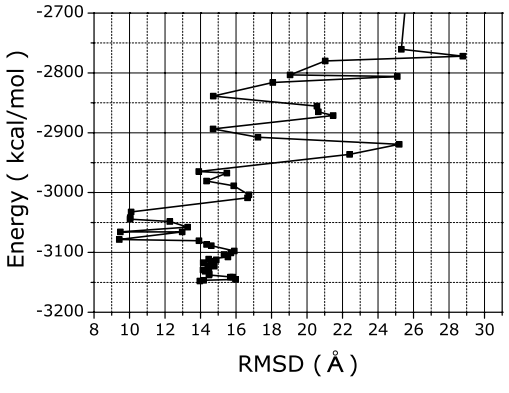

Fig. 2　RMSD とエネルギーの履歴

Fig. 2 のように，全体を見れば徐々に RMSD が小さくなる方向に推移している．しかし，探索終盤では一度 RMSD が小さくなったにも関わらず，それよりエネルギーの低い個体があったために，RMSD が大きくなる方向に推移している．このように，エネルギー最小化を行っても天然構造に近い構造を得るのは難しい．

3 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 (PSA/GAc)

3.1 概要

PSA/GAc は Fig. 3 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たものである．

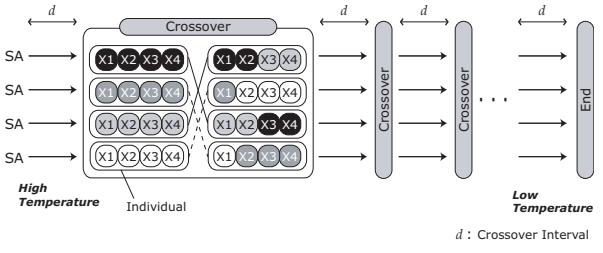

Fig. 3　PSA/GAc の模式図

このモデルでは，解の伝達時に並列に実行している SA から親としてランダムに 2 個体を選択し，設計変数交叉を行う．設計変数間交叉は Fig. 4 のように各設計変数の間でのみ交叉を行う．そして，親個体と生成された子個体を合わせた 4 個体の中から良好な 2 個体を選択し，次の探索点とする．

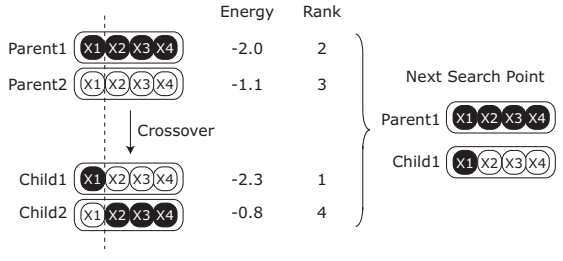

Fig. 4　設計変数間交叉

PSA/GAc では SA を並列化することで探索点が増えるため収束が早くなる．また遺伝的交叉を行うため，部分解がある問題に対して有効な手法であるといえる．

3.2 問題点

タンパク質立体構造予測問題において，PSA/GAc は PSA より性能が向上することが知られている．そのため，遺伝的交叉によってより良い探索が行われていると考えられてきた．しかし，昨年の研究で，交叉によって親個体よりエネルギーの低い子個体が生まれる割合が非常に低いことがわかった [4]．Fig. 5 に交叉後に選択される親個体・子個体の割合を示す．対象問題が Protein-A，交叉間隔 64MCsweep，8 並列で 6000MCsweep した場合の結果である．縦軸が交叉後の選択される個体の割合 (%)，横軸が交叉手法である．

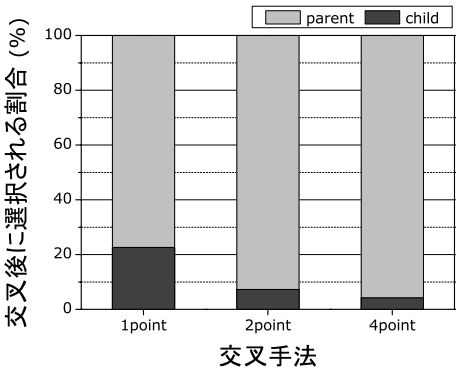

Fig. 5　交叉後の個体の選択割合

このように，子個体が選ばれる確率は最高でも 1 点交叉の約 20 %で，2 点交叉，4 点交叉では 10 %未満である．そのため，交叉で生成された子個体がほとんど選択されておらず，交叉が有効に機能していないという問題があることがわかった．

4 今後の目標

今後の研究方針は，PSA/GAc における遺伝的交叉を有効に機能させるような交叉手法を開発することである．今までの交叉では，大きく構造が異なる親同士で交叉すると，子個体の構造も大きく変わってしまい，エネルギーが高くなる場合が多かった．したがって，今後は構造が大きく変わらないようにしつつ親の形質を受け継ぐような子個体を生成する方法を開発する必要がある．

参考文献

1) 廣安知之, 三木光範, 小掠真貴, 岡本祐幸.
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の検討. 情報処理学会論文誌：数値モデルと応用, Vol.43, No.SIG10(TOM7). 2002.

2) TINKER Home Page.
`http://dasher.wustl.edu/tinker/`

3) CASP6 Home Page.
`http://predictioncenter.llnl.gov/casp6/`

4) 永松秀人.
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によるタンパク質立体構造予測における遺伝的交叉の検討. 同志社大学大学院　修士論文. 2004.