タンパク質立体構造予測における交叉の有効性の検証

鍵谷武宏

1　はじめに

自然界のタンパク質はエネルギーが最も低い安定した状態で存在することが知られている．そのためタンパク質の立体構造予測はエネルギー最小化問題として取り扱うことができる．本研究室では，タンパク質のエネルギー最小化問題を解く手法として，「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)」を用いる．また近年のエネルギー関数の改良により，2 次構造が多く含まれる解が生成できるようになった．このことから，全体としては，天然構造には近くないが，部分的に良く似た構造ができている可能性が考えられる．つまり部分解がある問題に対して，ある設計変数の最適値が求まっていれば，GA のオペレータである遺伝的交叉を用いることで，より天然構造に近い構造ができると考えた．そこで本報告では，PSA/GAc により得られたタンパク質の立体構造に対して，様々な交叉手法を適用しその効果を検証する．

2　タンパク質立体構造予測

2.1　タンパク質の立体構造を決定するパラメータ

タンパク質の立体構造は二面角で表現することができる．二面角とは Fig. 2 のように，4 つの連続する原子から成る角度のことで，最適化計算によるタンパク質の立体構造予測では二面角が設計変数として用いられる．例えば Met-enkephalin では Try-Gly-Gly-Phe-Met の五個のアミノ酸から成り，それぞれのアミノ酸は Fig. 2 のように，Try が 6 個，Gly が 3 個，Phe が 5 個，Met が 6 個の設計変数を持つ．またそれぞれのアミノ酸は主鎖と側鎖に別れ，主鎖の二面角と側鎖の二面角が存在する．

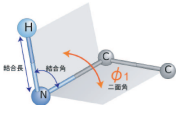

Fig.1　二面角 (出典 : 参考文献 1 より引用)

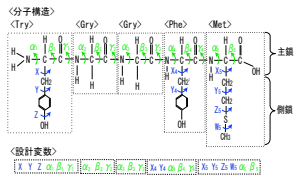

Fig.2　Met-enkephalin(出典 : 自作)

2.2　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．また本研究では，TINKER 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．

2.3　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 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．

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

3　PSA/GAc (Parallel Simulated Annealing with Genetic Crossover)

PSA/GAc とは並列に実行している SA の解の実行時に，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．このモデルでは，Fig. 3 のように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．設計変数間交叉はアミノ酸を一つの固まりとし，各アミノ酸ごとに行う．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．

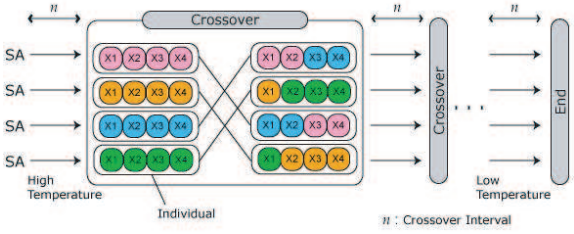

Fig.3　PSA/GAc の模式図 (出典 : 参考文献 1 より引用)

4　タンパク質立体構造予測における交叉の有効性の検討

4.1　対象問題

本実験の対象問題として ProteinG を用いる．ProteinG は 56 残基，321 個の二面角からなるタンパク質で，Fig. 4 のような構造をしている．

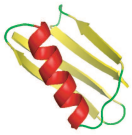

Fig.4　ProteinG(出典：自作)

4.2　実験概要

交叉の有効性を検討するために，3 つの交叉手法について検討した．各交叉手法については以降に記述する．また基本となるプログラムのアルゴリズムは Fig. 5 に示したように，まず親個体 30 個の中から，親となる個体をランダムに 2 個体選択し，交叉を行う．そして子個体を 8 個体生成し，親個体を含む 10 個体の中から最もエネルギーの低い 2 個体を選択し，母集団に返す．そして，この処理を 100 世代まで繰り返し行う．

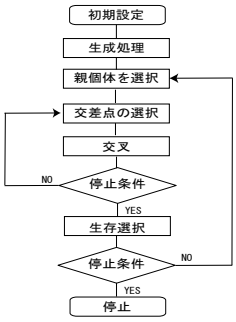

Fig.5　アルゴリズム (出典：自作)

4.3　側鎖を固定した交叉

タンパク質は，隣り合う原子同士が衝突するとエネルギーが高くなる性質がある．そこで交叉の際に隣り合う原子同士が衝突しないように考慮する必要がある．そこで側鎖を固定し，主鎖についてのみ交叉を行った．

4.3.1　結果

母集団の初めのエネルギーと 100 世代後のエネルギーを Fig. 6 に示す．また母集団の初めの RMSD 値と 100 世代後の RMSD 値を Fig. 7 に示す．

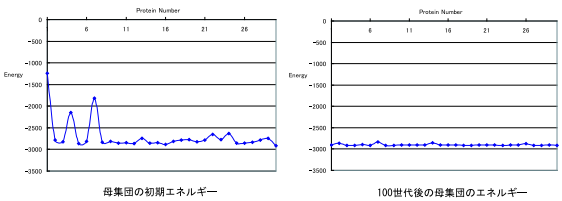

Fig.6　主鎖を交叉する時のエネルギー値の変化
(出典：自作)

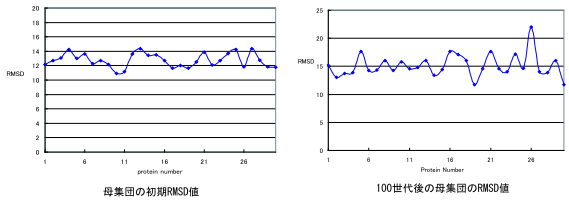

Fig.7　主鎖を交叉する時の RMSD 値の変化
(出典：自作)

4.4　選択の際に最適化を取り入れた交叉

4．3 の手法では，主鎖を入れ替えるだけなので，やはり衝突が生じる場合がある．そこで交叉した後に側鎖を最適化し，衝突を回避するようにした．そこで子固体が生成されると，Fig. 8 のように設計変数毎に「生成」「受理判定」「遷移」という操作を行い，10MCsweep 後に選択するようにした．クーリング率は 0.85，初期エネルギーは 2.5(1200K) とした．

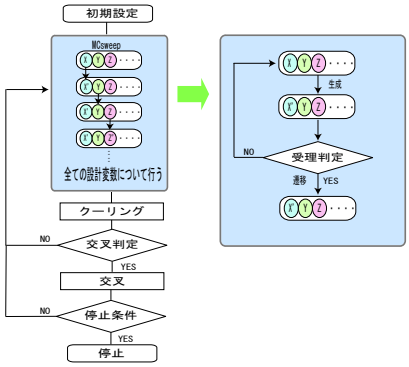

Fig.8　アルゴリズム (出典：参考文献 1 より引用)

4.4.1　結果

母集団の初めのエネルギーと 100 世代後のエネルギーを Fig. 9 に示す．また母集団の初めの RMSD 値と 100 世代後の RMSD 値を Fig. 10 に示す．

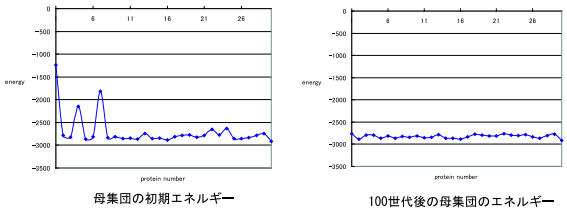

Fig.9　側鎖を SA した時のエネルギー値の変化
(出典：自作)

4.5　エリート個体を固定した交叉

親個体をランダムに選択していたが，1 つの親個体を母集団の中で最もエネルギーの低いものに固定し交叉を行った．

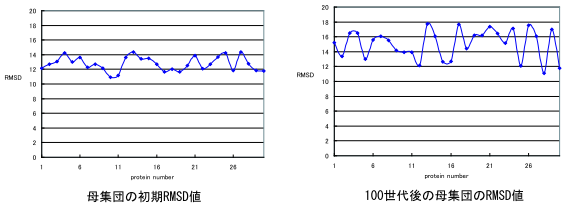

Fig.10　側鎖を SA した時の RMSD 値の変化
(出典：自作)

4.5.1　結果

母集団の初めのエネルギーと 100 世代後のエネルギーを Fig. 11 に示す．また母集団の初めの RMSD 値と 100 世代後の RMSD 値を Fig. 12 に示す．

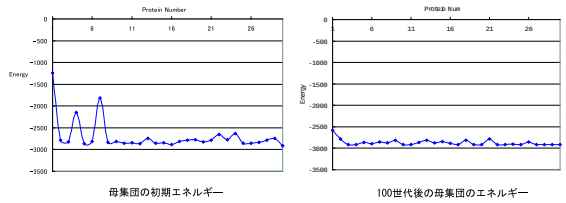

Fig.11　エリートを固定した時のエネルギーの変化
(出典：自作)

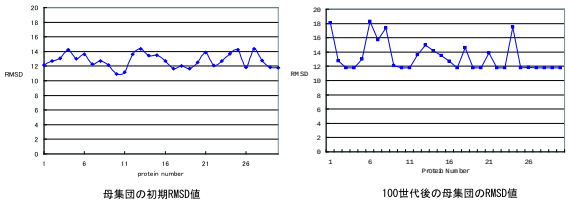

Fig.12　エリートを固定した時の RMSD 値の変化
(出典：自作)

4.6　考察

Fig. 6, Fig. 9, Fig. 11 を見ると，全体的にエネルギーが下がっていることがわかる．しかしながら，Fig. 6, Fig. 9, Fig. 11 の 100 世代後の母集団を見ると，どの交叉手法についても，母集団の最小エネルギーである-2911.24 よりも良い解を生成することはできなかった．また Fig. 11 については，同じ最小エネルギーをとったものが母集団の中に 14 個体存在し，エリート個体に解が収束してしまっていることがわかる．これは 8 個の子個体から 2 個体を選択する際に，エリート個体をベースにした子固体が，選択されると，エリートに非常に良く似た解になってしまう．またエリート個体をベースにした子固体は，エリートに近いため，エネルギーが低くなり，選択されやすいことが原因だと考えられる．このことから，エリート個体を固定した交叉では，解の多様性が維持できず，エリート解に収束してしまうことがわかる．

次に Fig. 7, Fig. 10, Fig. 12 を見ると，初期母集団の RMSD 値に比べて 100 世代後の RMSD 値が全体的に増加していることがわかる．このことから，エネルギーは全体的に下がってはいるが，構造の位置関係は nativ のものと比べると，悪くなっていることがわかる．

最後にそれぞれの交叉手法により生成された解のうち，子個体がどれくらい採択されてるのかを Fig. 13 に示す．Fig. 13 を見ると，SA をすることにより，解の精度が少し上昇することがわかる．またエリート個体を固定した場合と，主鎖を交叉した場合では，どちらも親個体の選択率はほぼ同じであった．しかしながら解の収束はエリートを固定した時のほうが大きかった．このことから，エリート解が及ぼす次世代への生存選択の影響は大きいことがわかる．

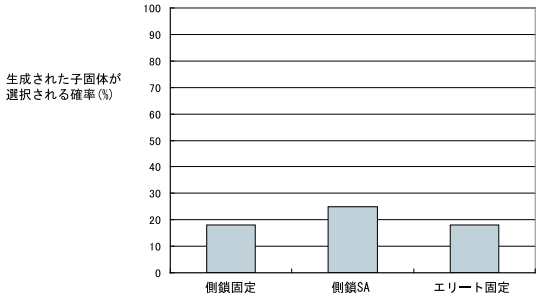

Fig.13　エリートを固定した時の RMSD 値の変化
(出典：自作)

5　まとめ

本報告では，側鎖を固定した交叉，選択の際に最適化を取り入れた交叉，エリート個体を固定した交叉の 3 つの交叉について，その効果を検証した．その結果，エネルギー値の推移を見ると，どの交叉手法においても初期母集団に比べて 100 世代後の母集団のエネルギー値は下がっていることがわかった．しかしながら，母集団の中の最小のエネルギー値は，更新することができなかった．次に RMSD 値について見ると，どの交叉手法においても，初期母集団に比べて 100 世代後の母集団の RMSD 値は増加していることがわかった．このことから，エネルギーは全体的に下がってはいるが，構造の位置関係は nativ のものと比べると，悪くなっていることがわかった．

6　今後の課題

エリートを個体を固定する交叉では，100 世代後には解の収束が見られた．そこで今後は，生存選択の方法についても考え，解がすぐに収束しないような生存選択の方法を検討する必要がある．

参考文献

1) 自作 PSA/GAc の性能検証，宇野尚子
http://mikilab.doshisha.ac.jp/dia/research/report/2003/0719/006/report20030719006.html
2) 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，宇野尚子　http://mikilab.doshisha.ac.jp/dia/monthly/monthly05/20051219/uno.pdf