タンパク質の立体構造予測

～近傍構造と距離尺度から構成される高性能交叉手法 dMSXF～

河本敏孝
Toshitaka Kawamoto

Abstract: Recently, Protein research have been wellknown. Protein research generally use Simulated Annealing. But, in this paper, We introduce deterministic multi-step crossover fusion (In this paper, dMSXF) to take parameters. dMSXF is method of crossover that We decide step and neighbor individual. dMSXF is similar to Simulated Annealing. We found if the value of dMSXF parameters are large, search performance is high.

1　はじめに

近年，タンパク質の構造予測が注目されている．タンパク質は 20 種類のアミノ酸が鎖状に連結して作られる物質であり，鎖状のアミノ酸が特定の形に折りたたまれた状態で存在する．また，タンパク質の機能的性質はその立体構造によって決まることが知られており，構造の解明は新薬の開発や病理の解明につながる．

本研究では，探索手法として遺伝的アルゴリズム (Genetic Algorithm，以下 GA) を用いた．GA は生物の遺伝と進化の仕組みを模擬した確率的多点探索アルゴリズムである．

GA の大きな特徴のひとつは，多点探索法であり，解同士の情報交換を探索に有効に利用している点である．子個体と呼ばれる次世代の解候補は親と呼ばれる現世代の解群から交叉オペレータによって生成される．このとき複数 (通常 2 つ) の親の良いところを受け継いだ子孫を生成することが交叉オペレータの目的である．

しかしながら，GA を与えられた問題に適用するに当たってはその問題にとって十分適切な交叉を設計することは困難である．きつい制約条件を持つ問題では実行可能解を生成することが困難であったり，必ずしも親の良いところを受け継ぐ子個体を生成することができないため，GA の性能を十分に引き出すことができない場合が多い．

本研究では，このような GA の課題に対して，問題ごとに近傍構造と距離尺度から容易に構成できる高性能な交叉手法，dMSXF を用いてタンパク質の構造予測を行うことを目的とする．

2　deterministic Multi-step Crossover Fusion(dMSXF)

dMSXF では親個体 1 から親個体 2 に向けて局所探索を行う過程において，まず，Fig1 にあるように親個体 1 と親個体 2 においてビットの違う部分の数をハミング距離とする．このハミング距離をいくつかのステップに区切り，ステップごとに親個体 2 に近づけていく．その際のステップを dMSXF Step という．以下に dMSXF のアルゴリズムを示す．Fig1，Fig2 に dMSXF の探索の様子を示す．

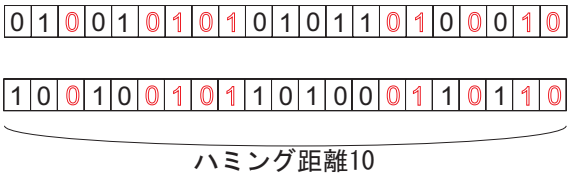

Fig. 1　ハミング距離

- 探索点 $X_1 1$ = 親個体 1 とする
- ステップ k における探索点 X_k の近傍にある μ 個の解群を近傍 $\mathrm{N}(X_k)$ とする．$\mathrm{N}(X_k)$ のすべての近傍解はかならずハミング距離 (近傍解，親個体 2) < ハミング距離 (X_k，親個体 2) を満たさなければならない．$\mathrm{N}(X_k)$ の中でもっともよい近傍解を選択する
- 全ステップの中でもっとも良い解を親個体 1 と置き換える．次に親個体 2 について，他の解，親個体 3 を取り出し，親個体 2 から親個体 3 に向けて探索をすすめる

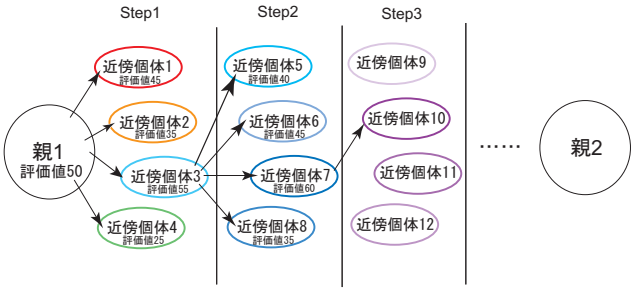

Fig. 2　dMSXF の探索の様子

3　数値実験

3.1　実験概要

交叉手法に dMSXF を用いた GA をタンパク質に適用し，対象問題として Met-enkephalin を用いた．パラメータの中で個体数，dMSXF Step 数，近傍個体数の検討を行った．

3.2　対象問題

本研究では，対象問題として Met-enkephalin を用いた．Table1 に対象問題 (Met-enkephalin) の特徴，Fig3 にその立体構造を示す．

Table 1　Met-enkephalin の特徴

対象タンパク質	Met-enkephalin
二面角数	23
計算終了世代	Tyr-Gly-Gly-Phe-Met

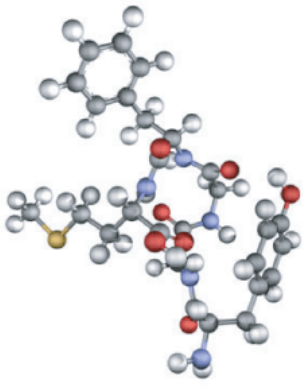

Fig. 3　Met-enkephalin

3.3　パラメータ

実験に用いたパラメータを Table2 に示す．それぞれ平均値と中央値を比較し，交叉手法 dMSXF における最適パラメータについて検討する．

まず，最初に個体数を変化させて実験を行った．個体数パラメータについては Table2 に示すとおりである．

次に，dMSXF Step を変化させて実験を行った．dMSXF Step パラメータについては Table2 に示すとおりである．

最後に，近傍個体数を変化させて実験を行った．近傍個体数パラメータについては Table2 に示すとおりである．

4　実験結果

4.1　個体数

個体数を変化させた場合のエネルギーの中央値の履歴を Fig4 に，平均値の履歴を Fig5 に示す．縦軸をエネルギー値，横軸を世代数とする．

Table 2　GA パラメータ

対象タンパク質	Met-enkephalin
個体数	400，800，1600，3200，6400
染色体長	207
設計変数長	23
計算終了世代	500
dMSXF Step	2，4，8，16，32，64，128
近傍個体生成数	2, 4, 8, 16, 32
力場	OPLS-AA/L
試行数	10

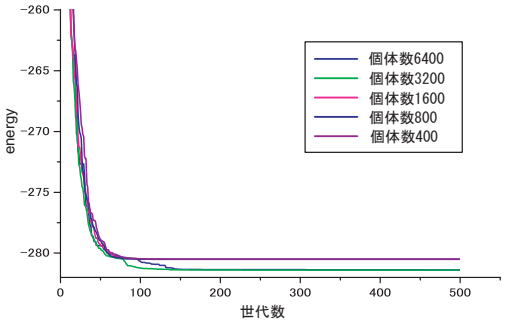

Fig. 4　エネルギーの中央値履歴

この結果を見ると，どの個体数でもほとんど変化は見られないといえる．わずかながらではあるが個体数 3200 の場合がエネルギーが下がっている．

4.2　dMSXF Step 数

dMSXF Step 数を変化させた場合のエネルギーの中央値の履歴を Fig6 に，平均値の履歴を Fig7 に示す．dMSXF Step 数以外のパラメータに関して，個体数は 800，それ以外は Table2 にあるパラメータで実験している．これは，ある程度の精度で評価値を出し，かつ計算時間の短いものを選択した．

この結果を見ると dMSXF Step 数は少ない方が良い効率で最適解を見つけることができるといえる．

4.3　近傍個体数

近傍個体数を変化させた場合のエネルギーの中央値の履歴を Fig8 に，平均値の履歴を Fig9 に示す．近傍個体数以外のパラメータについて，個体数を 800，それ以外は Table2 にあるパラメータで実験した．

この結果を見ると，近傍個体数は多い方が収束するスピードが早いことがわかる．

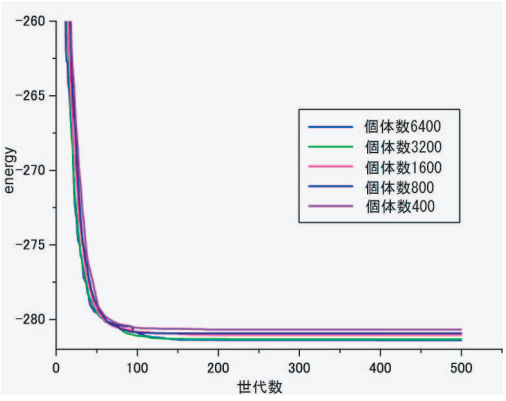

Fig. 5　エネルギーの平均値の履歴

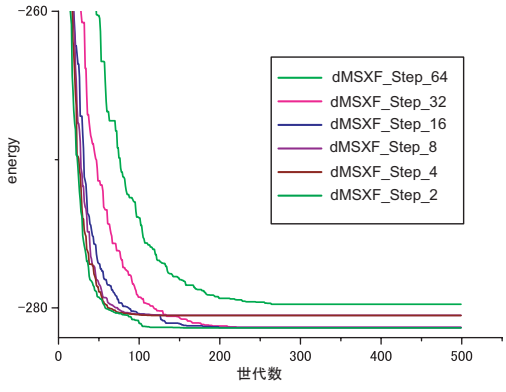

Fig. 6　エネルギーの中央値履歴

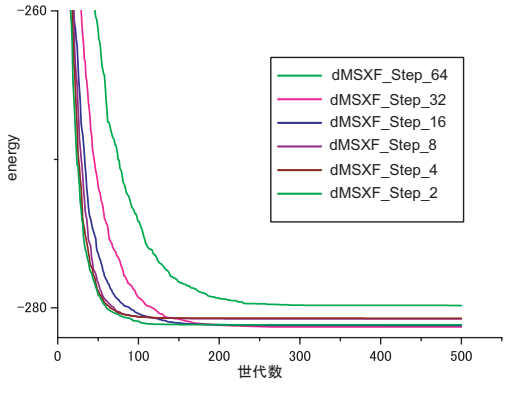

Fig. 7　エネルギーの平均値の履歴

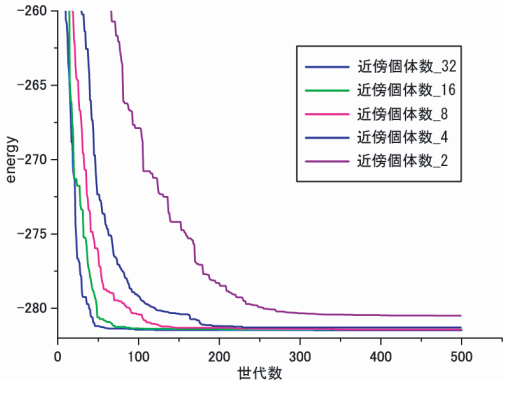

Fig. 8　エネルギーの中央値履歴

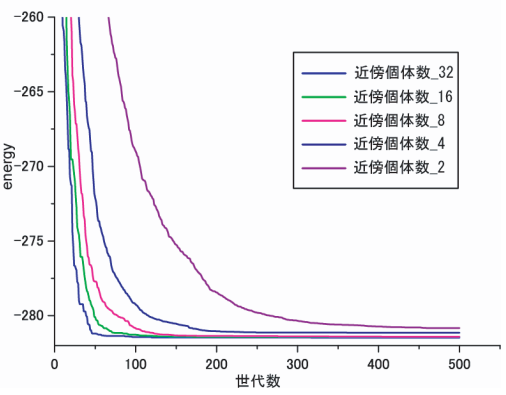

Fig. 9　エネルギーの平均値の履歴

以上より，Met-enkephalin において最適なパラメータは個体数は 3200，dMSXF Step は 2，近傍個体数は 32 にすると良いことが確認できた．

5　SA との比較

次に，交叉手法 dMSXF を用いた GA と逐次 SA の比較を行う．dMSXF のパラメータは前節で得られた最適値 or 最良値を用いる．SA のパラメータは Table3 に示す．

Table 3　SA パラメータ

対象タンパク質	Met-enkephalin
最高温度	2.0
最低温度	0.01
クーリング率	0.99997
総 MCsweep 数	6000
近傍	180 °
力場	OPLS-AA/L
試行数	10

それぞれの手法で探索終了時に得られたエネルギーの中央値，および平均値の結果を Fig10 に示す．縦軸をエネルギー値，横軸を手法とする．

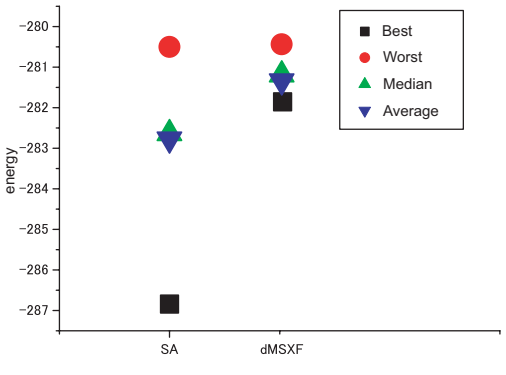

Fig. 10　SA との比較

Fig10 からも明らかなように，SA で解いた場合の方が良いエネルギー値を出していることがわかる．これは，SA という手法が次元ごとに SA 処理を繰り返す MC-sweep であるのに対して，GA では全次元をもとに評価している点にあると思われる．

6　まとめ

本研究では，dMSXF における最適パラメータの検討を行い，最適パラメータを用いたタンパク質の立体構造予測を dMSXF で行った．逐次 SA との比較により dMSXF ではタンパク質のエネルギー最小化問題としては十分な結果が得られないということがわかった．dMSXF は完全に内挿的な交叉であり，単独で用いる場合は初期集団の覆う範囲が最適解を含む，あるいは最適解に必要な要素を初期集団に含むことを前提にしている．今後，この手法を改良，もしくはハイブリッドすることによって新たなタンパク質構造解析手法を考えていきたい．

参考文献

1) Kokoro Ikeda, Shigenobu Kobayashi. Deterministic Multi-step Crossover Fusion: A Handy Crossover for GAs. pp162-171. PPSN7,(2002). ”, 2002