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の構造解析

Parallel Simulated Annealing using Genetic Crossover for Protein Folding Problem

小掠 真貴
Maki OGURA

Abstract: This paper shows Parallel Simulated Annealing using Genetic Crossover for Protein Folding Problem (PSA/GAc). SA is used at protein folding problem, however energy function of protein is so complicated that conventional SA needs much computation time. PSA/GAc that was proposed by authors is the algorithm that can reduce computation time and expand searching ability compared with conventional SA in protein folding problems.

1　はじめに

本研究では，アミノ酸配列情報からのタンパク質の立体構造を解析するため，新たなシミュレーテッドアニーリング (Simulated Annealing : SA) の手法を提案し，小規模なタンパク質の立体構造解析を行った．

タンパク質は生物のさまざまな生命現象を直接担っているという点で重要な物質であり，その構造を解明することは生命現象の仕組みを解明することへとつながっている．タンパク質の構造解析は最適化問題の一つとして考えられ，自然なタンパク質の立体構造はエネルギーの最小状態に対応している．数値計算を用いた構造解析では SA を中心に研究が進められているが，そのエネルギー関数は大域的にいくつかの，局所的には無数の極小値を持っていることから，現在の SA の手法では多くの計算時間を要するという問題点がある．

この問題点を解決するにあたり，局所的な探索が得意な SA に，大局的な探索が得意でかつ部分解の組み合わせで最適解が得られる問題に有効である遺伝的アルゴリズム (Genetic Algorithm : GA) のオペレータを取り入れることで，タンパク質の構造解析に対して有効な手法を提案する．本研究では，SA に GA のオペレータを取り入れた手法として遺伝的交叉を用いた並列 SA(Parallel Simulated Annealing using Genetic Crossover : PSA/GAc) を提案し，小規模なタンパク質 (Met-enkephalin) の構造解析を行うことによってその有効性を検討した．また，PSA/GAc を分散メモリ型並列計算機に実装するための並列モデルを考案し，有効性を検討した．

2　遺伝的交叉を用いた並列 SA の概要

提案手法は Fig. 1 に示すように，一定間隔のアニーリングを繰り返し，遺伝的交叉を行って最適解を求めようとするアルゴリズムである．GA のオペレータである遺伝的交叉を用いた SA であるため，SA の探索点の総数 (並列数) を個体数と呼ぶこととする．

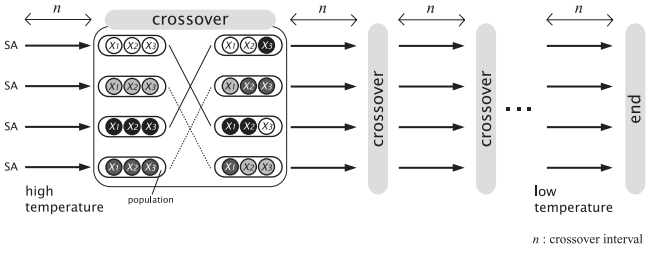

Fig. 1　Simulated annealing with genetic crossover

このモデルでは，解の伝達時に並列に実行している SA からランダムに親として 2 個体を選択し，設計変数間交叉を行う．もとの親と生成した子との 4 個体間のうち評価値の高い 2 個体を選択して，この 2 個体から次の探索を続けるという方法である．ある設計変数の最適値が求まっている場合，遺伝的交叉によってその設計変数の最適値を他の SA 探索に伝達することができるため，アニーリングの収束を早めることができる．

3　タンパク質の構造解析

タンパク質の立体構造はエネルギーの最小状態に対応しており，エネルギーを最小とするような構造を最適化手法を用いて求めることで，構造解析が可能である．これまで，タンパク質の構造解析においては SA が主として使用されてきた [1]．岡本らは，小規模なタンパク質 (Met-enlephalin) を対象として構造解析における SA の有効性を確かめている [2]．

Met-enkephalin は Fig. 2 のように Tyr-Gly-Gly-Phe-Met という 5 個のアミノ酸からなり，$E \leq -11kcal/mol$ の領域で最小エネルギー構造をしている．Met-enkephalin の主鎖における 10 個の二面角と，側鎖における 9 個の二面角をそれぞれ設計変数とする．つ

まりこのタンパク質は 19 個の設計変数を持っており，1MCsweep によって 19 回のアニーリングの処理が行われるものとする．

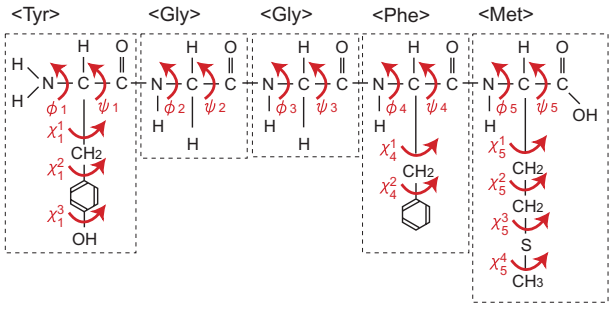

Fig. 2　Met-enkephalin

本実験では PSA/GAc を用いて Met-enkephalin の解析を行い，岡本らの結果と比較した．評価計算回数を同等とするため，PSA/GAc では，20 個体を用い 4992 回の MCsweep を繰り返すこととした．

それぞれ試行は 10 回ずつ行い，最適構造が得られた確率を Table 1 に示した．Success rate とは，試行回数に対して最適解を得られた回数の割合を示している．また，得られた最適構造を Fig. 3 に示す．

Table 1　Comparison between PSA/GAc and Conventional SA

Algorithms	Evaluations	Success rate
PSA/GAc	100005×19	0.90
Conventional SA	100000×19	0.50

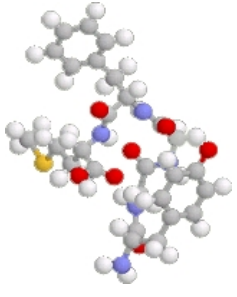

Fig. 3　Low-energy structure of met-enkephalin

Table 1 から，最適構造を得る確率は，従来の SA を用いる場合よりも PSA/GAc を用いた場合の方が高いことが明らかとなった．

4　遺伝的交叉を用いた並列 SA の並列実装モデル

大規模なタンパク質の構造解析では計算量が非常に多くなるため，計測に膨大な時間を要する．今後，大規模なタンパク質に取り組むには計測時間の短縮が重要となる．しかし，PSA/GAc は共有メモリ型並列計算機に対するモデルであり，分散メモリ型並列計算機には適用できない．そこで，分散メモリ型並列計算機に実装するための並列モデルを 4 種考案した．計算時間，解探索能力の各面から 4 種の並列モデルを比較した [3]．ここではその中で最もよい結果を示したモデルの概要を述べる．

この並列実装モデルは，プロセッサを複数用意し，各プロセッサには全個体数をプロセッサ数で分割した数だけの個体を存在させる．交叉時にはそのうちの 1 つを交叉専用のプロセッサとする．

交叉周期まで，すべての個体は並列に SA の処理を行う．交叉周期になると，各プロセッサに存在する個体の中からランダムに 1 個体を選択し，ある 1 つのプロセッサに送信する．交叉プロセッサでは，送信された個体を全て受け取った時点で，それらの個体の中でランダムに 2 個体ずつ設計変数間交叉を行い，子を生成する．もとの親と生成した子との 4 個体のうち評価値の高い 2 個体を残す．そして，交叉プロセッサに残った個体の中から 1 個体ずつランダムに選択し，各プロセッサに送信する．個体を受け取ったプロセッサから SA の処理を再開する．これを終了条件まで繰り返すというモデルである．

IBM RS/6000SP を用いて Met-enkephalin の構造解析を行ったところ，最適構造を 90%以上の確率で得られることが明らかとなった．また，プロセッサを増加させることによって大幅に計算時間を短縮することが可能となった．並列化効率も高く，本並列実装モデルの有効性が証明された．

5　結論と今後の課題

本研究では，遺伝的交叉を用いた並列 SA(PSA/GAc) を提案し，Met-enkephalin という小規模なタンパク質の構造解析を行った．その結果，現在構造解析に用いられている手法よりも良いふるまいをすることが明らかとなった．また PSA/GAc の並列実装モデルを考案し，従来の PSA/GAc よりも計算時間，解探索能力の面で有効であることを確認した．今後は大規模なタンパク質の構造解析を対象とし，良好な結果を得ることを課題としている．

参考文献

1) 木寺詔紀．コンピュータ解析 -最適化をめぐって．蛋白質核酸 酵素, Vol. 39, No. 7, 1994.

2) 岡本祐幸．モンテカルロシミュレーションで探るタンパク質の折り畳み機構．物性研究, Vol. 70, No. 6, pp. 719–742, 1998.

3) 廣安知之, 三木光範, 角美智子, 小掠真貴．遺伝的交叉を用いた並列 SA(分散メモリ型並列計算機への実装モデルの検討)．情報処理学会第 62 回全国大会 講演論文集, 2001.