2 個体分散遺伝的アルゴリズムによるタンパク質立体構造のエネルギー最小化

Dual Individual Distributed Genetic Algorithm for Minimizing the Energy of Protein Tertiary Structure

岩橋　崇史
Takashi IWAHASHI

Abstract: This paper describes Genetic Algorithm (GA) for minimizing the energy of protein tertiary structure. In the conventional study, Simulated Annealing (SA) is used to be applied for this problem. In the previous studies, it is also reported that it is difficult to find the optimum solutions by GAs. Dual individual Distributed Genetic Algorithm (Dual DGA) is one of DGAs and is good at global search. The Dual DGA also maintains the diversity of the solutions. Therefore, it can be supposed that they can get a good solution in energy minimization of protein tertiary structure. In this study, Dual DGA is applied to protein tertiary structure. The target protein in this paper is Met-enkephalin that consists of 5 amino acids sequences. The results show that Dual DGA has the higher searching capability than SA.

1　はじめに

タンパク質は生命現象に直接関わるため，その機能の解明は生命現象の仕組みの解明につながる．タンパク質の機能は，その立体構造によって決定される [1] ため，タンパク質の立体構造を解明することが重要となる．タンパク質の最適な立体構造は，そのタンパク質が持つエネルギーの最小状態と対応している．このため，最適化手法によってタンパク質が持つエネルギーが最小となる二面角の組み合わせを求めることで，構造予測が可能となる．

一般的に，遺伝的アルゴリズム (Genetic Algorithm : GA)[2] によるタンパク質立体構造のエネルギー最小化は困難であると報告されている [3]．GA には早期収束により局所解へ収束するという特徴を持っている．本研究では，タンパク質のエネルギー最小化が困難である原因として，GA の早期収束に注目する．そして，早期収束の回避を考慮した 2 個体分散遺伝的アルゴリズム (Dual individuals Distributed Genetic Algorithm : Dual DGA)[4] がタンパク質立体構造のエネルギー最小化に有効な最適化手法であると考えられる．そこで，本研究では小規模なタンパク質である Met-enkephalin のエネルギー最小化に DualDGA を適用する．これにより，小規模なタンパク質立体構造のエネルギー最小化において，Dual DGA の有効性を検証する．

2　タンパク質立体構造のエネルギー最小化

2.1　タンパク質立体構造のエネルギー最小化の研究

タンパク質立体構造エネルギー最小化の研究において，対象とするタンパク質モデルは格子型と全原子型の 2 種類に分けられる．格子型のタンパク質モデルは，設計変数がアミノ酸であり，立体構造が簡略化されているため，精度の良い立体構造を予測することができない．全原子型のタンパク質モデルは，タンパク質を構成する全原子間のエネルギーを考慮している．このモデルでは，膨大な計算時間を必要とする欠点がある．しかし，設計変数である原子間の二面角の組み合わせ最適化することによって，精度の良い立体構造を予測することができる [5]．

本研究では，全原子型のタンパク質モデルにおいて，タンパク質立体構造のエネルギー最小化を行う．一般的に，全原子型のタンパク質モデルにおいて，GA によるエネルギー最小化は困難であると報告されている [3]．そこで，設計変数の依存関係に着目した小林らの実数値 GA による研究 [6] や，局所探索に注目した岡本らのシミュレーテッドアニーリング (Simulated Annealing：SA) による研究 [5] が行われてきた．本研究では，GA によるタンパク質立体構造のエネルギー最小化が困難である原因として，GA の特徴の 1 つとしてあげられる早期収束に注目する．そして，この早期収束を考慮した GA のモデル，Dual DGA を用いる．

2.2　タンパク質のエネルギー関数

タンパク質は 20 種類のアミノ酸がつながったもので，結合する原子間には静電相互作用や水素結合などのエネルギーが存在する．そして，それら全エネルギー和がタンパク質のもつエネルギーとなる．本研究では，全原子型のタンパク質モデルのエネルギーを求める関数として，岡本らが使用しているもの [5] を用いた．式 (1) にそのエネルギー関数を示す．

$$E_{tot} = E_P + E_S \tag{1}$$

式 (1) において，全エネルギー関数 E_{tot} は，タンパク質分子の構造エネルギー E_P と溶媒和の自由エネルギー E_S の和で与えられる (単位は $kcal/mol$)．E_S は溶媒の様々な寄与を想定された項であり，E_P は，式 (2) に示すように，静電相互作用項 E_C，12-6 レナード・ジョーンズ項 E_{LJ}，水素結合項 E_{HB} の分子内すべての原子対についての和に，すべてのボンドの周りの回転角についての和であるねじれエネルギー項 E_{tor} を足したものとなっている．

$$
\begin{aligned}
E_P &= E_C + E_{LJ} + E_{HB} + E_{tor} \\
E_C &= \sum_{(i,j)} \frac{332\, q_i\, q_j}{\epsilon\, r_{ij}} \\
E_{LJ} &= \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\
E_{HB} &= \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\
E_{tor} &= \sum_{i} U_i \left(1 \pm cos(n_i\, \chi^i) \right)
\end{aligned}
\tag{2}
$$

ここで，r_{ij} は i 番目の原子と j 番目の原子間の距離 (単位は $\mathring{A}$) である．また，ϵ は誘電率，χ^i はボンド i における二面角である．各原子ではその重心に相互作用の中心があるとし部分電荷 q_i がそこに集中しているとする．更に，E_C 中の 332 という因子はエネルギーを $kcal/mol$ の単位で表すための係数である．

エネルギー関数内のパラメータおよび分子の幾何情報については，ECEPP/2[7, 8, 9] のものを使用した．気相中のシミュレーションを行うものとしたため，誘電率 ϵ を 2，$E_S = 0$ としている．

3　2 個体分散遺伝的アルゴリズム

本研究で用いる Dual DGA[4] は DGA において 1 島あたりの個体数を 2 としたものである．島内の個体数を極限まで少なくすることによって，島数を増やし，より多様性を維持した探索が期待される．しかし，島内の個体数を極端に少なくすることによって，通常の遺伝的操作では島内の多様性が急速に失われると考えられる．そこで Dual DGA では，遺伝的操作に多様性を維持する機構を組み込んでいる [4]．

DualDGA の手順を以下に示す (Fig. 1)．まず，個体の母集団を，ランダムに，2 個体ずつの島に分割する．個体のビット列はランダムに設定される．そして各島において，次の操作を世代ごとに繰り返す．

1. 一定世代を経過するごとに，移住を行う．まず，2 つの個体のうち，ランダムに一方を選択し，そのコピーを他の島に送る．そして適合度の低い方の個体は，他の島から送られてきた個体に置き換えられる．
2. 2 つの個体を交叉させ，新しい 2 つの子個体を生成する．本論文では，一点交叉を用いている．この段階では，親個体も存在している．
3. 突然変異を行う．交叉で生成された 2 つの子個体をそれぞれ 1 ビット反転させるが，反転する点は，2 つの個体で 1 ビットずらす．これは，島内の 2 つの個体が同一になるのを防ぐためである．
4. 個体の適合度の評価を行う．
5. 2 つの親個体と，2 つの子個体から，それぞれ適合度の高い方の個体を選び，次世代の 2 個体とする．この選択法により，適合度の最も高い個体が必ず次の世代に生き残る．またこのとき，移住個体は選択されない．これにより，初期収束を回避する効果が期待される．

Dual DGA は，従来の DGA と比較して，パラメータ設定の困難さの一部を解消している．1 島あたりの個体数を 2 とすることより，総個体数を決定すれば島数も一意に決まる．よって，Dual DGA において設定すべきパラメータは，個体数と移住間隔のみである．また，1 つの島を粒度の単位とすることで，並列化の粒度を柔軟に変更できるため，並列計算環境に適している．

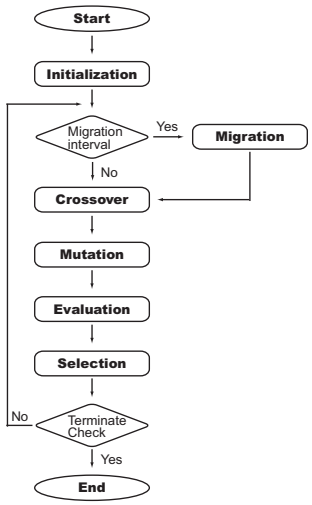

Fig. 1　Flowchart of Dual DGA

4　実験

本節では，タンパク質立体構造のエネルギー最小化における Dual DGA の有効性の検討を行うことを目的として，小規模なタンパク質立体構造のエネルギー最小化を行う．なお，DGA においても同実験を行う．

4.1　対象とするタンパク質

Dual DGA による立体構造予測を行うタンパク質としては，小規模なタンパク質である Met-enkephalin を用いた．Met-enkephalin は Tyr-Gly-Gly-Phe-Met とい

う 5 個のアミノ酸が連なることにより構成されるタンパク質である．そして，ECEPP/2 エネルギー関数 [7, 8, 9] に基づいた気相中において，$E \leq -11kcal/mol$ の領域で最小エネルギー構造をしている [10]．本実験では，Met-enkephalin の主鎖における 10 個の二面角 ϕ_1, ψ_1〜ϕ_5, ψ_5 と，側鎖における 9 個の二面角 χ_1^1〜χ_5^4 をそれぞれ設計変数とした．二面角のとり得る値は $[-180°, 180°)$ の範囲で表現した．

4.2 パラメータ

ここでは，DGA および Dual DGA による Met-enkephalin の立体構造のエネルギー最小化を行う．個体数，島数，移住間隔は DGA，Dual DGA の解探索能力に大きな影響を与えるパラメータである．よって，これらのパラメータをいくつか用意し，実験を行った．用意した個体数と移住間隔を下記に示す．

- 個体数
  800，1600，3200，6400
- 移住間隔
  1，2，3，4，5，6，7，8，9，10

DGA については，1 島あたりの個体数を 4，8，16 の 3 つのパターンを用意した（以下，1 島あたりの個体数が n のときの DGA を DGA(n) と記述する）．その他のパラメータにおいては，Table 1 に示す．なお，終了条件は岡本ら [10] と同一の値としている．試行回数は 30 回とした．

Table 1 パラメータ

model	DGA	Dual DGA
Sub Population Size	16,8,4	2
Number of Design Variables	19	
Chromosome Length	171 (= 19 × 9 Design Variable)	
Selection	Tournament	-
Tournament Size	2	-
Crossover Rate	1.0	
Crossover	1pt. crossover	
Mutation Rate	0.006 (= 1 / 171)	
Migration Rate	0.25	0.5
Number of Elites	1	-
Terminal Criterion	1,900,000 evaluations	

4.3 DGA，Dual DGA の最適解発見率の比較

Dual DGA と DGA の最適解発見率の比較を行った．Fig. 2 にそれぞれの最適解発見率を示す．グラフの横軸は個体数と移住間隔の推移を示している．最適解発見率が最も高かったのは，個体数が 6400，移住間隔が 3 のときの Dual DGA であり，0.93(=29/30) であった．また，個体数が 800，1600，3200 のとき，ほとんどの移住間隔で，Dual DGA の方が DGA よりも最適解発見率が高いことが分かった．しかし，個体数が 6400 のとき，モデル別に最適解発見率の明確な差を確認することができなかった．よって，一概にいずれのモデルが解探索能力が高いとは言えない．

- 個体数の変化による影響
  移住間隔が 1 のときのように，ある特定の移住間隔では，個体数が増すことにより，最適解発見率が増すことを確認した．しかし，すべての移住間隔について，同様のことを認められなかった．

- 移住間隔の変化による影響
  DGA，Dual DGA について，すべての個体数で，移住間隔の大きさにより最適解発見率が異なることが分かった．個体数が 6400 のとき，いずれのモデルにおいても，最適な移住間隔が存在し，それを基準に移住間隔が大きくなるもしくは小さくなるにつれ，最適解発見率が低くなる傾向が確認できた．この傾向は個体数が 6400 のときのみでしか現れず，他の個体数では，移住間隔の変化による特徴的な最適解発見率の変化が見られなかった．

以上のことから，Dual DGA，DGA では個体数が増すにつれ，移住間隔は解探索能力に大きな影響を及ぼし，重要なパラメータになることが分かった．

4.4 解探索履歴の比較

Fig. 3 に，Dual DGA において最も高い最適解発見率を示したパラメータである個体数が 6400，移住間隔が 3 のときの Dual DGA，および移住間隔が 6 のときの DGA の 30 回試行における平均値の解探索履歴を示す．Fig. 3 より，Dual DGA は解探索序盤では多様性を維持しているため，解探索は緩やかであった．そして，解探索終盤にかけては，解収束しており，有効な解探索を行っていることを確認した．また，DGA においては，島内の個体数によって，解探索が異なっていることを確認した．

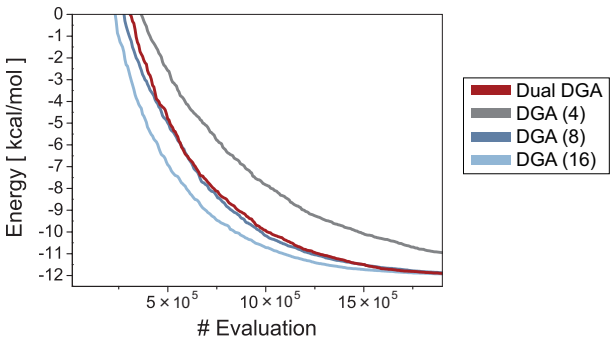

Fig. 3 Energy transition of Met-enkephalin (Populaiton size:6400 Migration Interval:3)

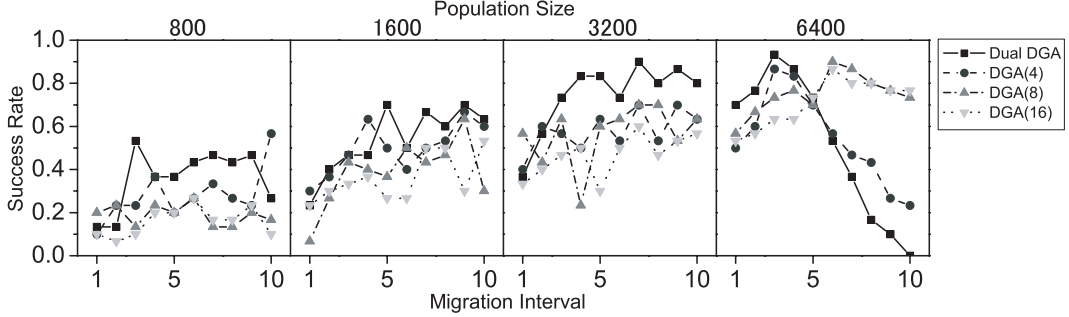

Fig. 2　Success Rate

4.5　他の最適化手法との比較

Dual DGA および DGA と岡本らの結果 [11, 12] と比較する．岡本らは SA によるタンパク質立体構造のエネルギー最小化の研究を行っている [11]．Fig. 4 に各手法による Met-enkepahlin の立体構造のエネルギー最小化における最適解発見率を示す．Dual DGA および DGA については，今回行った実験で，得られた最適解発見率のうち最も高い値を示した．なお，Fig. 4 で示す PSA/GAc とは廣安らが提案した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 using Genetic Crossover : PSA/GAc) である．PSA/GAc はタンパク質立体構造のエネルギー最小化に有効な手法であると報告されている [12]．Fig. 4 より，Dual DGA，DGA は SA よりも最適解発見率が高いことが分かった．また，PSA/GAc と最適解発見率がほぼ等しいことが分かった．以上より，Dual DGA，DGA はタンパク質立体構造のエネルギー最小化において，SA より解探索能力が高いといえる．また，Dual DGA，DGA は PSA/GAc とほぼ等しい最適解発見率を示したため，タンパク質立体構造のエネルギー最小化に有効な手法であると考えられる．

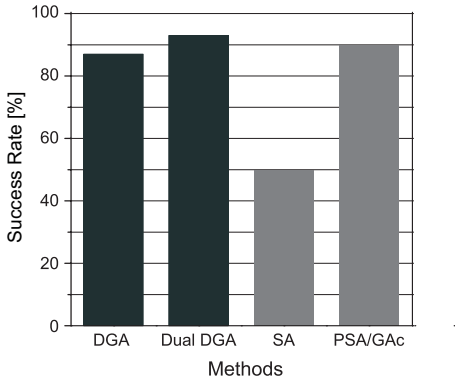

Fig. 4　Success Rate

5　結論

本研究では，多点探索の最適化手法である分散遺伝的アルゴリズム (DGA) と 2 個体分散遺伝的アルゴリズム (Dual DGA) を用いて，5 個のアミノ酸からなる Met-enkephalin を対象に立体構造予測を行った．実験結果より，DGA と Dual DGA の解探索能力はほぼ等しいことが分かった．そして，タンパク質立体構造のエネルギー最小化に用いられている最適化手法の一つである SA よりも DGA，Dual DGA の解探索能力が高いことが明らかとなった．Dual DGA と DGA はほぼ等しい解探索能力ではあったが，Dual DGA は DGA よりもパラメータ設定が簡易であり，並列計算環境に適している．以上より，Dual DGA は小規模なタンパク質立体構造のエネルギー最小化に有効な手法であることが示された．

参考文献

1) 池内俊彦. 生命を学ぶ タンパク質の科学. オーム社出版局, 1999.

2) D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, 1989.

3) Laurence D.Merkle, Gary B.Lamont, Jr. George H.Gates, and Ruth Pachter. Hybrid Genetic Algorithms for Minimization of a Polypeptide Specific Energy Model. *IEEE Conf on Evolutionary Computation*, pp. 396–400, 1996.

4) 廣安知之, 三木光範, 佐野正樹, 谷村勇輔, 濱崎雅弘. 2 個体分散遺伝的アルゴリズム. 計測自動制御学会論文集, Vol. 38, No. 11, pp. 990–995, 2002.

5) 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.

6) 友部修, 小野功, 小林重信. Ga による蛋白質の構造決定に関する実験的考察. 第 25 回知能システムシンポジウム, pp. 35–40, 1998.

7) F.A. Momany, F.A., R.F. McGuire, A.W. Burgess, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 79, pp. 2361–2381, 1975.

8) G. Nemethy, M.S. Pottle, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 87, pp. 1883–1887, 1983.

9) M.J. Sippl, G. Nemethy, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 88, pp. 6231–6233, 1984.

10) Yuko Okamoto, Takeshi Kikuchi, and Hikaru Kawai. Prediction of Low-Energy Structures of Met-Enkephalin by Monte Carlo Simulated Annealing. *CHEMISTRY LETTERS*, pp. 1275–1278, 1992.

11) Ulrich H. E. Hansmann and Yuko Okamoto. Numerical Comparisons of Three Recently Proposed Algorithms in the Protein Folding Problem. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, Vol. 18, No. 7, pp. 920– 933, 1997.

12) 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, Vol. 43, No. 7, pp. 70–79, 2002.