マルチ個体群の並列遺伝的アルゴリズムを用いたタンパク質の配列解析

戸谷智之[†,☆] 石川幹人[†,☆☆]

我々は，効率の良い探索を実現するマルチ個体群方式の遺伝的アルゴリズムを開発し，タンパク質配列の解析問題に応用した．分子生物分野の代表的な配列解析問題であるマルチプルアライメントは，最近，並列反復改善法で効果的に解決できることが示された．そこで使われた並列探索手法は，最良優先探索とマルチ山登り探索であったが，各々は問題点を持っていた．最良優先探索は，スコアの良い解の近傍を集中的に探索するので改善速度が速いが，比較的悪い局所解に陥ることも多い．一方，マルチ山登り探索は，広い範囲を分散的に探索するので比較的良い解へ至りやすいが，解の改善に時間がかかる．マルチプルアライメントの問題は，すでに定評のある評価尺度が確立されており，組み合わせ最適化問題として解決可能である．しかし現時点では，あらゆる観点からの生物学的評価が数値化されているわけではないので，生物学者は，いくつかの準最適解を比較のうえ，そこから生物学的知見を導き出す．そこで，マルチプルアライメントのシステムには，良質の準最適解を高速に生成する機能が必要とされている．我々は，並列反復改善法の解法を遺伝的アルゴリズムの枠組にあてはめ，効率的な探索を行うマルチ個体群方式を考案した．その結果，最良優先探索のように速い改善を行いながら，マルチ山登り探索のように良い準最適解を得られるアライメントシステムを構築できた．

Protein Sequence Analysis Using a Multi-Group Parallel Genetic Algorithm

TOMOYUKI TOYA[†,☆] and MASATO ISHIKAWA[†,☆☆]

We have succeeded in solving protein sequence analysis problems by applying an original multi-group parallel genetic algorithm. It has been shown recently that multiple sequence alignment, a typical sequence analysis problem in molecular biology, can be solved effectively using parallel iterative improvement. In this method, two search strategies, best-first and parallel hill-climbing, were used alternatively. Both of them, however, have problems. The best-first search, which rapidly improves the state of solution, is sometimes trapped in a locally optimal solution having relatively a low score. The parallel hill-climbing method, which distributedly searches for an optimal or sub-optimal solution, is frustrating due to its long execution time. Multiple sequence alignment can be considered as a combinatorial optimization problem. Its score system, however, is still incomplete, so that biologists often compare high-score alignments to extract biological information. Therefore, the multiple alignment method requires a new search strategy to rapidly generate sub-optimal alignments. We have devised a multi-group parallel search strategy which efficiently solves multiple alignment problems, incorporating the iterative improvement method into the genetic algorithm framework.

1. はじめに

代表的な配列解析法であるマルチプルアライメント（Multiple Alignment）は，遺伝子やタンパク質の機能・構造予測，生物種の進化系統樹の作成の際に欠かせない技術である．たとえば，**図1** の (a) のようなタンパク質のアミノ酸配列（出典：PIR データベース）が9本あったとすると，(b) のようにアライメントされる．図で，左側の見出しが配列の名前で，右側の文字が一つ一つのアミノ酸を表現する．図1(b) で，配列のところどころにギャップ "-" を入れることで，**LG.G.FG.V** などの共通文字が同じカラムに並んでいるのがわかる．このように複数の文字が，複数の配列でほぼ共通になっている文字の組を配列モチーフ

† （財）新世代コンピュータ技術開発機構
Institute for New Generation Computer Technology (ICOT)
☆ 現在，シャープ（株）情報商品開発研究所
Presently with Information Systems Product Development Laboratories, Sharp Co.
☆☆ 現在，松下電器産業（株）マルチメディアシステム研究所
Presently with Multimedia Systems Research Laboratory, Matsushita Electric Industrial Co., Ltd.

```
(a) problem
CABL :KLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDY
FER  :LLGKGNFGEVYKGTLKDKTSVAVKTCKEDLPQELKIKFLQEAKILKQYDHPNIVKLIGVCTQRQPVYIIMELVSGGDFLT
TRK  :ELGEGAFGKVFLAECHNLLPEQDKMLVAVKALKEASESARQDFQREAELLTMLQHQHIVRFFGVCTEGRPLLMVFEYMRH
GCPK :SLRGSSYGSLMTAHGKYQIFANTGHFKGNVVAIKHVNKKRIELTRQVLFELKHMRDVQFNHLTRFIGACIDPPNICIVTE
PKC1 :VLGKGNFGKVILSKSKNTDRLCAIKVLKKDNIIQNHDIESARAEKKVFLLATKTKHPFLTNLYCSFQTENRIYFAMEFIG
CGMPK:TLGVGGFGRVELVQLKSEESKTFAMKILKKRHIVDTRQQEHIRSEKQIMQGAHSDFIVRLYRTFKDSKYLYMLMEACLGG
CAMK :ELGKGAFSVVRRCVKVLAGQEYAAKIINTKKLSARDHQKLEREARICRLLKHPNIVRLHDSISEEGHHYLIFDLVTGGEL
FUSED:LVGQGSFGCVYKATRKDDSKVVAIKVISKRGRATKELKNLRRECDIQARLKHPHVIEMIESFESKTDLFVVTEFALMDLH
WEE1 :LLGSGEFSEVFQVEDPVEKTLKYAVKKLKVKFSGPKERNRLLQEVSIQRALKGHDHIVELMDSWEHGGFLYMQVELCENG

(b) result
CABL :KLGGGQYGEVYEGVWK---------KYSLTVAVKTLKED----TMEVEEFLKEAAV---MKEIK-HPNLVQLLGVCTREPPFYIITEFMTYGNLLDY
FER  :LLGKGNFGEVYKGTLK----------DKTSVAVKTCKED--LPQELKIKFLQEAKI---LKQYD-HPNIVKLIGVCTQRQPVYIIMELVSGGDFLT-
TRK  :ELGEGAFGKVFLAECH-NLLP---EQDKMLVAVKALKEA---SESARQDFQREAEL---LTMLQ-HQHIVRFFGVCTEGRPLLMVFEYMRH------
GCPK :SLRGSSYGSLMTAHGKYQIFANTGHFKGNVVAIKHVNKK---RIELTRQVLFELKH---MRDVQ-FNHLTRFIGACIDPPNICIVTE----------
PKC1 :VLGKGNFGKVILSKSK---------NTDRLCAIKVLKKDNIIQNHDIESARAEKKVFLLATKTK-HPFLTNLYCSFQTENRIYFAMEFIG-------
CGMPK:TLGVGGFGRVELVQLK---------SEESKTFAMKILKKRHIVDTRQQEHIRSEKQI---MQGAH-SDFIVRLYRTFKDSKYLYMLMEACLGG-----
CAMK :ELGKGAFSVVRRCVKV---------LAGQEYAAKIINTKK-LSARDHQKLEREARI---CRLLK-HPNIVRLHDSISEEGHHYLIFDLVTGGEL---
FUSED:LVGQGSFGCVYKATRK---------DDSKVVAIKVISKRG-RATKELKNLRRECDI---QARLK-HPHVIEMIESFESKTDLFVVTEFALMD-LH--
WEE1 :LLGSGEFSEVFQVEDP---------VEKTLKYAVKKLKVKF-SGPKERNRLLQEVSI---QRALKGHDHIVELMDSWEHGGFLYMQVELCENG-----
      .LG.G.FG.V......          ......A.K.....   ..........E...   ..... H...................E....
```

図1 マルチプルアライメントの例
Fig. 1 An example of multiple sequence alignment.

(Sequence Motif) と呼び，タンパク質のうちの重要な部分を指し示していると判断される．これは，タンパク質の配列のうち重要である部分に遺伝的変異が起きると，その生物が死滅してしまうので，生き残っている生物は，重要な配列部分を保存する傾向があるからである．

マルチプルアライメントされた結果は次のように利用できる．第一に，アライメントした配列のなかに，構造・機能がよくわかっているものが含まれていれば，アライメントした結果からわかる配列の類似性から，未知の構造・機能を推測することができる．第二に，先に述べたように，重要な配列の部分であるモチーフを見出せる．モチーフは構造や機能を特徴づけたり，データベースから新たな知見を引き出す手がかりになる．第三に，マルチプルアライメントから進化系統樹を描くことができる．系統樹の形成は，アライメントにおける文字置換数から，配列の進化距離を推定して行うのである．このようにマルチプルアライメントは，タンパク質の研究，そして分子生物学の研究の中で重要な役割を担っている．

2. 並列反復改善法

マルチプルアライメントは，長らく熟練した生物学者の手作業に依存してきたが，近年になって計算機による自動化や，支援も徐々に行われてきている．なかでも，並列反復改善法[1]は，アライメント過程で発生するエラーを，反復改善のサイクルで修正できる実用的な手法であった．本章では，この方法について概説する．

2.1 マルチプルアライメントのスコア

マルチプルアライメントの問題は，次の典型的なアライメントスコアを最適化することで，ある程度解決できる．

$$AlignmentScore = \sum_{i<j}^{seq.pair} \sum_{k}^{column} MatchScore(A_{ik}, A_{jk}).$$

$$MatchScore(A_{ik}, A_{jk}) = \begin{cases} Dayhoff(A_{ik}, A_{jk}) & A \text{ がともにアミノ酸のとき,} \\ 0 & A \text{ がともにギャップのとき,} \\ \mathrm{p} & A \text{ がアミノ酸と先頭のギャップのとき,} \\ \mathrm{q} & A \text{ がアミノ酸と 2 番目以降のギャップのとき,} \end{cases}$$

このスコア体系は，ペアワイズアライメントのスコアを総和するタイプであり，SP (Sum of Pairs) 体系と呼ばれている[2]．並列反復改善法は，この体系を用いたアライメントの組み合わせ最適化問題を解く手法である．式のなかで，アミノ酸同士の類似性尺度を与えているのは，Dayhoff の PAM250[3] というマトリックスで，該当アミノ酸対の遺伝的変異が偶然に対して如何に大きいかを数値化したものである．また，ギャップコストは，ギャップの長さに対する一次式で与えている[4]．p, q の値は問題に応じた調整が必要であるが，以下の計算では，$p = -8$, $q = -1$ とした．

2.2 最良優先探索

最良優先探索の並列反復改善法（図2）は，次に示

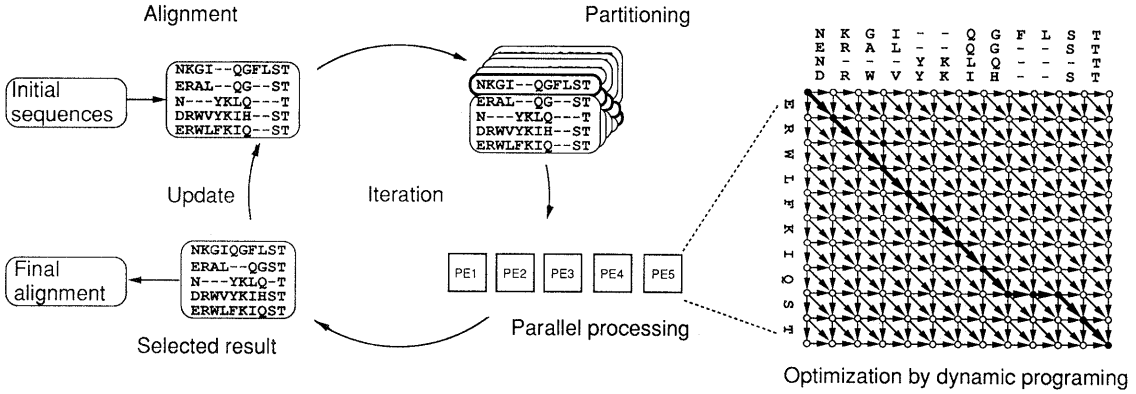

図2 並列反復改善法（最良優先探索）
Fig. 2 Best-first parallel iterative improvement method.

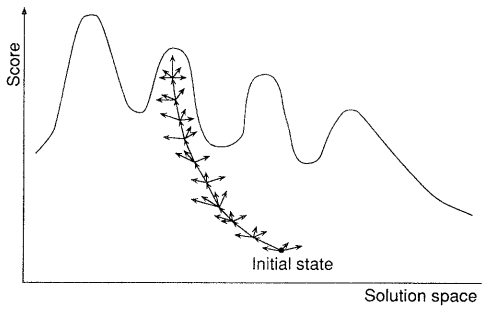

図3 最良優先探索
Fig. 3 Best-first search.

す手順で改善を行う方式である．分割法にいくつかの方法[1]があるが，ここでは1本抜き限定分割に限って話しを進める．

1. ギャップのない n 本の配列群を初期アライメントとする．
2. 初期アライメントに対し，配列1本とそのサブアライメントという，n 通りの分割を生成する．
3. 配列1本とサブアライメントとの，ダイナミックプログラミング（DP）によるアライメントの最適化を，各々のプロセッサで並列に行う．
4. それらの結果を比較し，スコアの最も良い解を，次の改善サイクルの初期アライメントとする．

あるサイクルで，スコアに改善がみられなかったら，その時点のアライメントを最終解とする．このダイナミックプログラミング（DP）は，最適経路問題としてアライメントを最適化する手法[4]である．DPによって，全体のアライメントスコアは，単調に改善される．

図3 に示すように，最良優先探索では，最も良いスコアの解を次々と選んで採用するため，解の改善速度は速いが，比較的悪いスコアの局所解にしばしば陥ることが難点となっている．

2.3 マルチ山登り探索

マルチ山登り探索の並列反復改善法（**図4**）は，以下に示す処理をプロセッサごとに独立に行い，その中で最も良い解を全体の解とする方式である．（ここでも1本抜き限定分割に限って話しを進める．）

1. ギャップのない配列群を初期アライメントとする．
2. 初期アライメントから，ランダムに配列1本を抜き出す．
3. その配列と，残りのアライメントとの間にDPを適用して，アライメントの最適化をする．
4. その結果を，次の改善サイクルの初期アライメントとする．

収束条件に適当なサイクル数を決め，そのサイクル数にわたってスコアに改善がみられなかったら，その時点のアライメントを最終解とする．

マルチ山登り探索では，**図5** に示すように，解空間上での局所解に陥ってしまう傾向を大きく緩和する．しかし，こんどは改善速度が遅くなり，実行時間が増大する問題点があった．

2.4 並列反復改善法の問題点

並列反復改善法で提案された最良優先探索とマルチ山登り探索は，各々，問題点を持っていた．最良優先探索は，集中的な探索のため比較的悪い局所解に陥ることがときどきある．また，マルチ山登り探索は，分散的な探索のため解の改善に時間がかかる．

一方，マルチプルアライメントの問題は，すでに定評のあるSPスコア体系が確立されているものの，現時点では，あらゆる観点からの生物学的評価が，そこに数値化されているわけではない．そのため，生物学者は，スコアの良い数個の解を比較のうえ，生物学的な観点から最も妥当な解を選ぶのが一般的である．

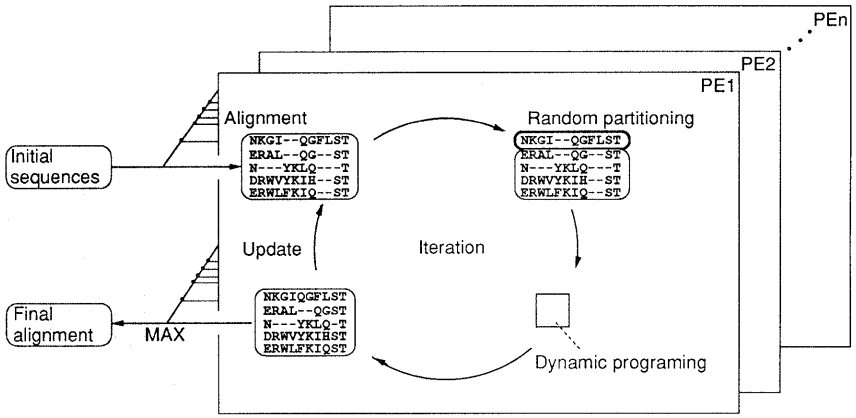

図4 並列反復改善法（マルチ山登り）
Fig. 4 Parallel hill-climbing iterative improvement method.

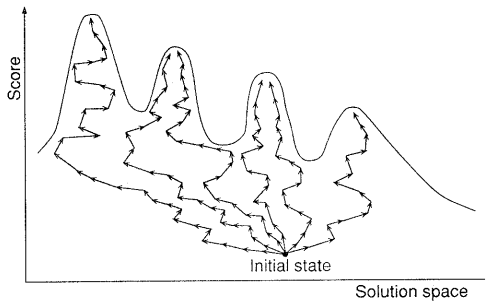

図5 マルチ山登り探索
Fig. 5 Parallel hill-climbing search.

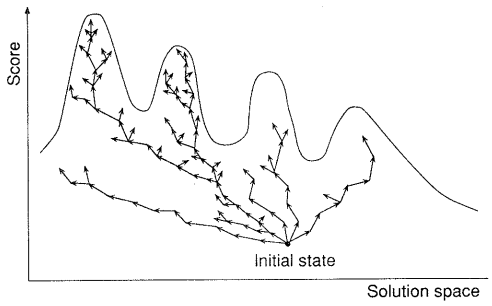

図6 融合的な探索
Fig. 6 Hybrid search.

そこで，並列反復改善法のシステムには，マルチ山登り探索で得られるような良質の準最適解を，最良優先探索のように高速に生成する機能が，必要とされていた．それは，**図6**にあるような，解空間を広域的に探索し，かつ，良質解の近傍は集中探索する，融合的探索手法と考えられる．我々は，遺伝的アルゴリズムのマルチ個体群方式で，そうした探索を実現し，マルチプルアライメントの問題に応用した．

3. 遺伝的アルゴリズムの応用

遺伝的アルゴリズム（以下GAと略す）は，遺伝子の生物学的進化の過程にヒントを得た，計算機上の組み合わせ最適化問題の解法である[5),6)]．最近になって，生物学の問題を組み合わせ最適化問題と捉え，GAを応用する研究が次々に出てきている[7)~10)]．生物界の現象にヒントを得た計算機手法が，生物学分野の問題に応用されているのは，面白い現象である．

本章では，一般的なGAのメカニズムと，GAへのアライメント問題の定式化法を述べる．

3.1 遺伝的アルゴリズムのメカニズム

GAは，生物個体群が，世代交代を繰り返して，環境へ適応していく過程を模擬している．生物個体群は各世代で，遺伝情報に発生する突然変異と，それに伴う自然淘汰，繁殖，そして，雌雄の遺伝情報の交配によって，環境により良く適応する個体群が形成されていく．この過程を解空間内の探索にあてはめたのが，GAである．

図7に具体的なメカニズムの例を示す．各個体が解空間のある一つの状態を保有し，その各々に状態を変更する操作（Modification）を施した後，スコア（適応率）の悪い状態のいくつかを捨てる（Selection）．その失われた状態数に見合う数だけ，生き残った個体がコピーされる（Duplication）．そして，さらに適当に2個体を選び，部分解同士を組み合わせた状態に交換する（Crossover）．この一連の操作を1世代（Generation）として，世代交代を繰り返すと，各個体が持つ解がスコアの良い状態に一様化して行き，解空間の（準）最適解が得られる．通常は，個体群内の最大スコアの解

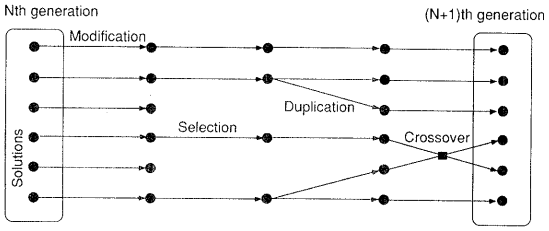

図7 遺伝的アルゴリズム（GA）
Fig. 7 Mechanism of genetic algorithm (GA).

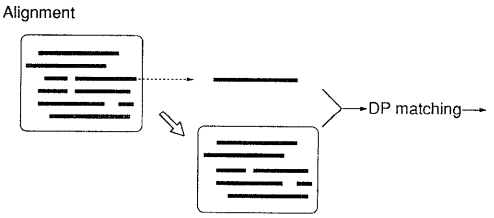

図8 アライメントにおける変異
Fig. 8 Modification of alignment.

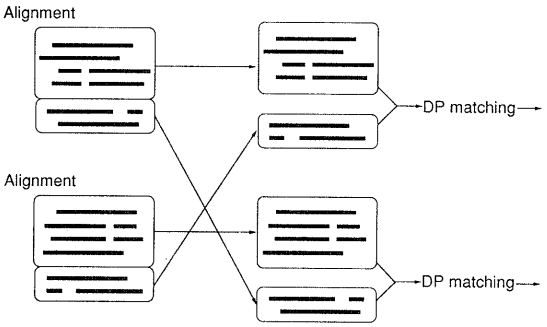

図9 アライメント間の交配
Fig. 9 Crossover of alignments.

を，最終解とする．

一般にGAは，遺伝情報を担うDNAとのアナロジーで，解の状態表現を1次元のビットストリング（1/0の文字列）とする．しかし，それでは大きな問題を解くには効率が悪いことも多い．最近では，ビットストリングでない適当な状態表現を，開発者が問題に応じて，自由に設計する傾向がある．

3.2 基本操作の定式化

我々は，マルチプルアライメントの問題を，次のようにGAの枠組にあてはめた．まず，個体に，ビットストリングでなく，マルチプルアライメントの解そのものをもたせ，適応率に，そのアライメントのスコアを対応させた．そのうえで，変異（Modification），淘汰（Selection），繁殖（Duplication），交配（Crossover）の操作を以下のようにした[11]．

変異： 変異は，各個体に独立に行われる．**図8**のように，個体がもつマルチプルアライメントから，ランダムに1本を抜きだし，残りのアライメントとDPで最適化するのが，一つの変異である．つまり，一つの変異は，反復改善法の1サイクルに相当する．伝統的なGAで突然変異（Mutation）というと，スコアが良くなる場合も悪くなる場合もあるが，本適用法では，スコアは悪くなることはない．

淘汰： 淘汰は，個体群全体を調べ，低いスコアの個体（アライメント解）を上位から指定した割合だけ，捨てる操作である．いわゆるエリート戦略をとっている．

繁殖： 繁殖は，淘汰によって残った個体群から，ランダムに選んだ個体を複製する．淘汰によって失われた個体は，繁殖によって生まれた個体で補われるため，通常は個体群の個体数は一定である．しかし，後で述べるマルチ個体群の方式では，移民などの操作により，各個体群の個体数は増減する．

交配： 交配は，**図9**のように，個体群から二つの個体をランダムに選び出し，部分解を交換する．選ばれた個体のアライメントされた配列群を，ランダムに二つの配列グループに分け，それらの片側の配列グループを交換して，DPを用いて融合する．これにより，アライメントのなかに，比較的良く揃っている部分があれば，他のアライメントにある，別な良く揃っている部分と，互いに融合され，ひときわ高い評価のマルチプルアライメントができる可能性が生まれる．

4. マルチ個体群方式の導入

GAでは，基本操作をどのように設定するかで，かなり特徴の違った探索を実現できる．我々は，独自のマルチ個体群方式を，GAを使ったアライメントシステムに導入し，効果的な探索を実現した．

本章では，一般的なGAのマルチ個体群方式と，我々のアライメント問題への適用法を述べる．

4.1 遺伝的アルゴリズムのマルチ個体群方式

GAは，変異や交配などの設定がある程度自由に行える点で，非常に柔軟な枠組と言える．逆に，GAを効率良く動作させるのもさせないのも，これらの設定の如何にかかっているとも言える．一般にGAは，基本操作をある方針に決めたならば，それに従って最初から最後までGAを実行する．しかし，問題によっては，実行の経過に従って設定を変更した方が良い場合がある．たとえば，実行の初期段階では広い解空間を探索し，終了段階では狭い空間を集中的に探索する戦

略が，多くの問題で有効なことは容易に予想がつく．アライメントの問題でも，シミュレーテッドアニーリングを導入した解法[12),13)]の研究で，その有効性が示されている．

GAでは，実行の経過に沿った設定の変更を，マルチ個体群方式[14),15)]を用いて合理的に行える．設定の異なるいくつかの個体群を準備しておき，各々の個体群の個体数を増減させることで，実行の経過に沿った設定の動的制御が実現できる．たとえば，Mühlenbeinらは，突然変異率の大きく異なる個体群をいくつか設け，それらの間で移民を行う方法を提案している．彼らの方法では，数世代ごとに，成績（群中の最大スコアの解で決まる）の良い個体群が，他の個体群から移民を受け付ける．ほかに少しの無作為な移民を認めることで，実行の初期には，探索の広い（突然変異率の高い）個体群が成績が良く個体数を増やし，実行の終盤では，探索の狭い（突然変異率の低い）個体群が優勢となる．全体として，結果的に，実行の初期段階では広い解空間を探索し，終了段階では狭い空間を集中的に探索する戦略が実現できている．

マルチ個体群方式は，自然界で近隣の生物種が，棲み分けしながらも競合し，環境の変化に柔軟に対応していくモデルに似ている．その意味で，生態学的シミュレーションの観点からも興味深い．

4.2 マルチプルアライメントへの適用

我々は，マルチプルアライメントを解くGAの枠組の検討から，淘汰率の設定により，解探索の戦略が調整できることに気づいた．たとえば，**図10**のように，交配は行わず，淘汰率を0%に設定したうえで，各個体を一つのプロセッサに割り付けると，並列反復改善法のマルチ山登り探索と等価になる．また，淘汰率を上げていくに従って，探索は，最良優先探索の色彩が強くなってくる．なぜなら，**図11**のように，淘汰率が高いと，スコアの高い解のみが残って，かつ，それらのコピーが次々に生まれ，その時点での最良解に近い解空間が集中的に探索される．

我々のマルチ個体群方式の適用では，**図12**に示すように，淘汰率を変えた個体群を四つ用意している．基本的な仕組みは，マルチ山登り法に近い探索をしている（淘汰率が低い）個体群で見つかったスコアの良い解を，最良優先探索に近い（淘汰率が高い）個体群へ移し，集中的な探索をする．その一方で，局所解に陥って来た個体群は個体数を減らし，代わりに，解の多様性が残っている淘汰率の低い個体群の個体数を増やすのである．移民（Migration），消滅（Elimination），生成（Production）の具体的な操作は以下のように

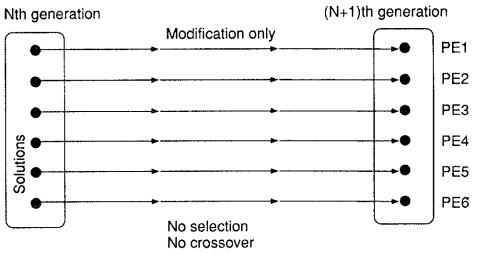

図10 GAにおけるマルチ山登り
Fig. 10 Parallel hill-climbing in GA.

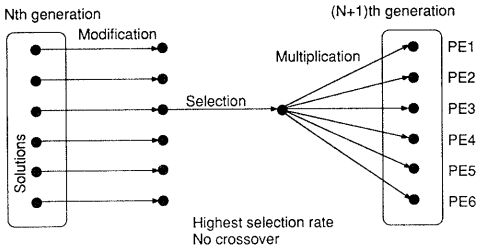

図11 GAにおける最良優先探索
Fig. 11 Best-first search in GA.

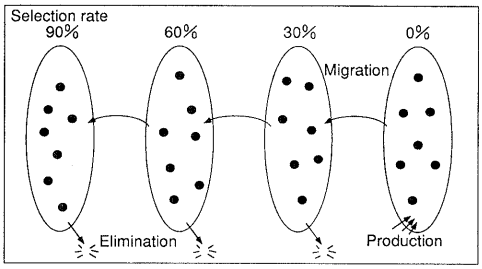

図12 GAのマルチ個体群方式
Fig. 12 Multi-group GA strategy.

する．

移民： 淘汰率が隣接している二つの個体群を比較し，淘汰率の高い個体群の最大スコアの解以上の良いスコアの個体が，淘汰率の低い個体群にあったら，それらをすべて淘汰率の高い方の個体群へ移動させる．

消滅： 最もスコアの良い解と同じスコアの解が個体群の半分を越えたら，そのスコアの解の半分を捨てる．(同じスコアでも異なった解の場合があるが，アライメントの同一性を調べるのは手間がかかるのでスコアで代用する．)

生成： 解の多様性が，最も保存されていると考えられる淘汰率0%の個体群で，繁殖を行い，消滅によって減った個体数相当の個体を新たに生み出す．

このようなマルチ個体群方式により，融合的な探索

が実現できる．つまり，解空間を全体的に多点探索する一方，有望な良い解の周辺は，局所的に集中探索するのである．

5. 実験と結果

上で述べたGAのマルチ個体群方式を利用したシステムを実装し，並列反復改善法と比較する実験を行った．本章では，その実験内容と結果を述べる．

5.1 並列計算機への実装

我々は，マルチプルアライメントの問題を解く，GAのマルチ個体群方式を開発し，その実験システムを，分散メモリ型の並列計算機PIM/m[16]（要素プロセッサ256台構成）上に実装した．使用言語は，プロセス間通信や同期処理の記述が容易な，並列論理型プログラミング言語のKL1[17]である．

実装は次のように行った．PIM/mの多数の要素プロセッサ（PE）のうち，PE0はマスタプロセッサとし，全体の包括的処理にあてる．他のPEには，各個体（アライメント状態解）を一つずつ割り当て，スレーブプロセッサとする．マスタプロセッサは，群間の移民，消滅，生成処理と，群内の淘汰，繁殖処理，そして交配の割り当てを行う．一方，各スレーブプロセッサは，群内の変異処理，割り当てられた交配の具体的な実行を行う．

一つの世代は，いくつかの変異を一定時間（現在80秒にしている）行うことで始まる．各スレーブプロセッサは変異を1回終えるたびにシステムタイマを見て，80秒が経過していたならば，その時点の解の状態をPE0へ送る．PE0は，群間処理（毎世代行う）を行い，続けて群内処理を行う．その結果，スレーブプロセッサによっては，新たなアライメント状態が割り当てられる．この時点でシステムタイマはリセットされる．割り当てが行われたスレーブプロセッサは，保持している解を捨て，新しい解について変異処理を始める．つまり，スレーブプロセッサは，交配が割り当てられない限り，常に変異処理を行っている．

交配が割り当てられたスレーブプロセッサは，変異処理に入る前に交配の実行（二つの部分アライメントのDPによる最適化）を行わねばならない．交配の実行は変異1回（1本配列と残りのアライメントとのDPによる最適化）より少し時間がかかる．80秒の世代時間に，変異は数回行えるが，最初に交配を実行してから変異処理に入ると，変異回数は1，2回少なくなる．

個体群は，各スレーブプロセッサの解に付されたラベルによって区別されており，個体群ごとに，解がまとまって管理されているわけではない．だから，移民がなされても，個体の属する個体群のラベルが変わるのみで，スレーブプロセッサ間を解のデータが移動することはない．よって，マスタプロセッサが集中的な処理をしても，PE間のデータ転送はわずかである．PEの稼働率は，実行全体で98％以上であった．

我々は，GAの実行中の性能をオンラインでモニタするツール（**図13**）も，OSF/Motifを用いて開発した．このツールにより，各個体群の平均（実線），最大（点線），最小（点線）スコアが折れ線グラフ表示できる（拡大/縮小も可能）ほか，世代ごとに，各個体群のスコア分布や詳細情報が表示できる．もちろん，解であるアライメントの内容も表示でき，GAの効率を向上させるための検討に，有効なツールとなっている．こうしたツールと実行プログラムは，ともにICOT無償公開ソフトウェア（`ftp.icot.or.jp` または `http://www.icot.or.jp`）に登録されている．なお，C言語にコンパイル可能で汎用並列計算機上で動作する実行プログラムも，同様に公開されている．

5.2 並列反復改善法との比較

前に述べたように，我々のマルチ個体群の定式化は，従来の並列反復改善法の拡張方式に相当する．そこで，従来法と比較するため，まず最初に，総個体数を253とし，交配なしで実験した．群内個体数の初期状態は，淘汰率90，60，30，0％の群が，それぞれ63，63，63，64個体とした．テストデータには，図1にあるような80文字分のタンパク質配列を，22本一組とした問題，30問を用いた．

図14が，従来の並列反復改善法と，GAのマルチ個体群方式による拡張法とを，同一のテストデータ30問の平均で比較した実験結果である．並列反復改善法の実験結果（▲△）は，論文[1]から，1本抜き最良優先探索（PE 22台），1本抜きマルチ山登り（PE 253台）を転載したものである．

図14の折れ線（●）が，交配なしのマルチ個体群方式の結果を示している．その線は，最良優先探索（▲）と同様な早さで，同程度のスコアの解に達している．さらに，そのスコアを越え，マルチ山登り（△）と同様な高いスコアに至っている．そして，このマルチ個体群方式のスコア改善は，つねにマルチ山登り法のスコア改善より，早期になされている．

マルチプルアライメントの問題は，評価スコアが必ずしも絶対的でないので，準最適な解を早期に得ることが重要であった．図14の結果は，マルチ個体群方式により，最良優先探索よりも良い解を，マルチ山登り探索よりも早期に求められることを示している．こ

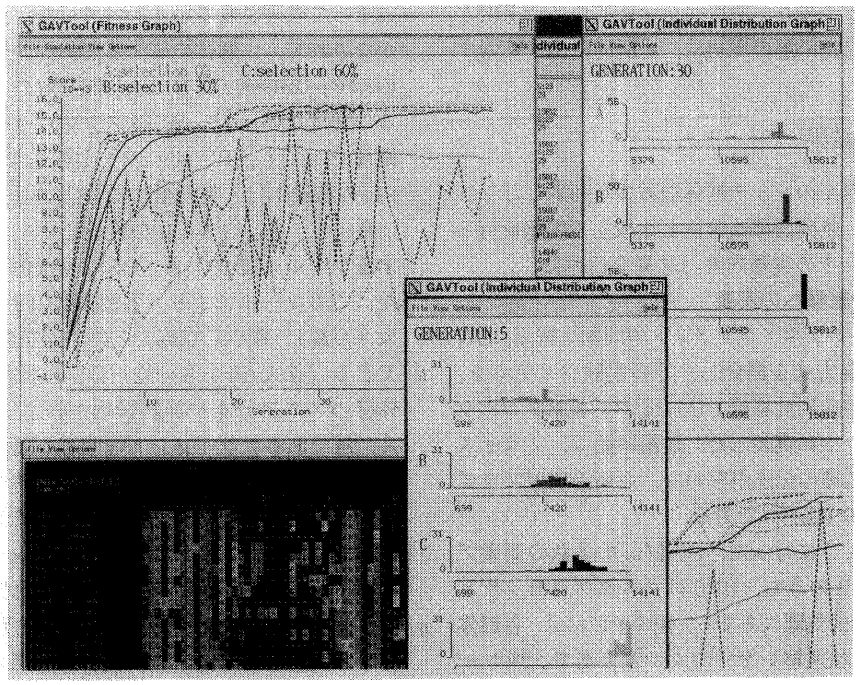

図13 マルチ個体群GAのモニタツール
Fig. 13 Multi-group GA monitoring tool.

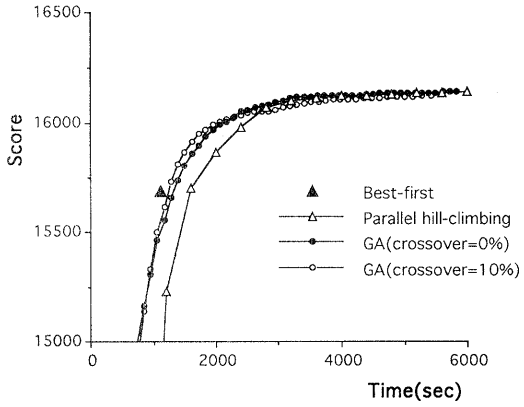

図14 改善履歴比較
Fig. 14 Comparing improvement histories.

れは，従来の並列反復改善法にまさる性能の良い並列探索が，GAで実現できたことを意味する．

次に，各個体群に10％の交配率を導入して，同一のテストデータについて実験を行った．10％の交配率とは，個体群の個体数の10％を越えない最大の偶数だけ，ランダムに個体を選んでペアを作り，交配を適用する操作である．図14の折れ線（○）は，交配を導入した実験の結果を示している．

交配ありの遷移を見ると，交配なしより若干早い立ち上がりではあるが，収束は少し遅いことがわかる．この結果は，交配に関して次の効果を反映したと推測できる．実行の初期段階では，交配の操作により，良い部分解を組み合わせる現象が起き，より良い解が早期に生まれた．その反面，収束段階では，交配が解をシャッフルする作用を及ぼし，収束を遅らせた．収束が遅い分，解の多様性を保持するのであるから，理論的には，収束解のスコアが良くなる可能性があるのだが，今回，収束解に際立った差異は得られなかった．この結果は，交配を導入しなくとも，収束解は十分良かったか，あるいは，交配を導入した試行の実行時間が十分でなかったことが原因に考えられる．

図13では，2時間分のマルチ個体群の実行状況モニタを示している．折れ線グラフの各個体群の平均スコア（実線）を比べると，淘汰率の高い個体群の方が早期に収束傾向を見せている．最大スコア（上側の点線）の変化をみると，淘汰率の低い個体群の最大スコアが，移民によって淘汰率の高い個体群へと移動している．棒グラフで，各個体群のスコア分布を比べると，第5世代では淘汰と移民によって，淘汰率の高い個体群がよりスコアの高いところに分布する傾向が出ている．第30世代になると淘汰率の高い個体群の解が局所解になり，消滅・生成の作用で淘汰率の高い個体群の個体数が減り，代わりに，淘汰率の低い個体群の個体数が増えている．画面には明示されていないが，第

50 世代を越えるあたりから，淘汰率の高い個体群の個体数は数個となって変化がなくなり，淘汰率の低い個体群から良いスコアの解が移民してくるのを，単に待つだけの状態となる．

6. おわりに

我々は，遺伝的アルゴリズムをマルチ個体群の方式で並列計算機上に実装し，実用規模のアライメント問題でも，準最適な解を比較的高速に求めることを可能とした．本システムは，マルチ山登り探索の並列反復改善法より，早期に収束に至り，かつ，最良優先探索のような局所最適解に陥りにくいことが，実験の結果から明らかとなった．また，交配の導入により，初期段階のスコアの向上が早まった．しかし，収束段階においては，交配が解をシャッフルする作用を及ぼして収束を遅らせた．

ここでは，交配の改良法を中心とした，システムの改良法を検討しよう．前に述べたように，交配の操作には次の二つの機能がある．一つは，解の多様性を維持するために，解をシャッフルする機能，もう一つは，良い部分解を組み合わせることで，効率良く解のスコアを向上させる機能である．

本システムの交配では，アライメントをランダムに二つの部分アライメントに分けている．そのため，解のシャッフルは盛んになされているものの，良い部分解の組み合わせは偶然に任されている．良い部分解を積極的に同定して，限定した交配を行うと，最適化の性能が飛躍的に高まる可能性がある．ただ，解のシャッフル機能はどうしても低下するので，注意を要する．

これまで，交配の操作には配列方向（アライメントに対して横方向）の分割のみを考えてきたが，カラム方向（縦方向）の分割も検討すべきであろう．カラム方向の分割は，とくに類似部分間の距離にばらつきがある場合，有効と考えられる．しかし，交配が可能なようにカラム方向に分割するには，工夫も必要である．交配される二つのアライメントにおいて，同一のアライメントがなされているカラムを見つけ，そこで分割する方法[18]が考えられている．

また，マルチ個体群の実装では，今回は淘汰率を変えて行ったが，今後は，個体群によって交配率も変える枠組も検討すると良い．実行が進むに従って交配の効果が変わってくるという，今回得られた実験結果は，効率の良い探索には交配率も変化させることが有効であると暗示している．

謝辞 研究に協力していただいたICOT遺伝子応用グループの方々，研究の機会を与えて下さったICOT内田俊一研究所長，新田克己第2研究部長に感謝いたします．

参考文献

1) 石川幹人，十時 泰，戸谷智之，星田昌紀，広沢 誠：並列反復改善法によるタンパク質の配列解析，情報処理学会論文誌，Vol.35, No.12, pp.2816–2830 (1994).
2) Altschul, S.F. and Lipman, D.J.: Trees, Stars, and Multiple Biological Sequence Alignment, *SIAM J. Appl. Math.*, Vol.49, pp.197–209 (1989).
3) Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C.: A Model of Evolutionary Change in Proteins, *Atlas of Protein Sequence and Structure*, Vol.5, No.3, pp.345–352, Nat. Biomed. Res. Found., Washington DC (1978).
4) Gotoh, O.: Optimal Alignment between Groups of Sequences and Its Application to Multiple Sequence Alignment, *Comput. Appl. Biosci.*, Vol.9, No.3, pp.361–370 (1993).
5) Holland, J.H.: *Adaptation in Natural and Artificial Systems*, Univ. Michigan Press (1975).
6) Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).
7) Konagaya, A. and Kondou, H.: Stochastic Motif Extraction Using a Genetic Algorithm with the MDL Principle, *Proc. 26th Annu. Hawaii Int'l. Conf. Syst. Sci.*, Vol.1, pp.746–755 (1993).
8) Parsons, R., Forrest, S. and Burks, C.: Genetic Algorithms for Sequence Assembly, *Proc. 1st Int'l. Conf. Intelli. Syst. Mol. Biol.*, pp.310–318 (1993).
9) Unger, R. and Moult, J.: On the Applicability of Genetic Algorithms to Protein Folding, *Proc. 26th Annu. Hawaii Int'l. Conf. Syst. Sci.*, Vol.1, pp.715–725 (1993).
10) 松田秀雄，山下 浩，金田悠紀夫：遺伝的アルゴリズムによる分子系統樹の作成，情報処理学会研究報告，95-FI-36, pp.15–22 (1995).
11) Ishikawa, M. et al.: Parallel Iterative Aligner with Genetic Algorithm, *Proc. Genome Informatics Workshop IV*, pp.84–93, Universal Academy Press (1993).
12) Ishikawa, M. et al.: Multiple Sequence Alignment by Parallel Simulated Annealing, *Comput. Applic. Biosci.*, Vol.9, pp.267–273 (1993).
13) Ohya, M., Miyazaki, S. and Ogata, K.: On Multiple Alignment of Genome Sequences, *IEICE Trans. Commun.*, E75-B, pp.453–457 (1992).

14) Mühlenbein, H. and Schlierkamp-Voosen, D.: Predictive Models for the Breeder Genetic Algorithm: Continuous Parameter Optimization, *Evolutionary Computation*, Vol.1, pp.25–49 (1993).
15) 筒井茂義，藤本好司：個体群探索分岐型遺伝的アルゴリズム（Forking GA）の提案，人工知能学会誌，Vol.9, pp.741–747 (1994).
16) Nakashima, H. et al.: Architecture and Implementation of PIM/m, *Proc. Fifth Gener. Comput. Sys. '92*, pp.425–435 (1992).
17) Hirata, K. et al.: Parallel and Distributed Implementation of Concurrent Logic Programming Language KL1, *Proc. Fifth Gener. Comput. Sys. '92*, pp.436–459 (1992).
18) Tajima, K.: Multiple Sequence Alignment Using Parallel Genetic Algorithms, *Proc. Genome Informatics Workshop IV*, pp.183–187, Universal Academy Press (1993).

(平成7年3月22日受付)
(平成7年9月 6 日採録)

戸谷 智之（正会員）

1964年生．1989年大阪大学理学部数学科卒業．同年，シャープ（株）に入社．1990年～95年，(財) 新世代コンピュータ技術開発機構（ICOT）に出向．知識獲得および推論エンジンの研究開発，並列計算機を用いた遺伝子情報処理の研究開発に従事．現在，シャープ（株）情報商品開発研究所勤務．最適化アルゴリズムに興味をもつ．

石川 幹人（正会員）

1959年生．1982年東京工業大学応用物理学科卒業．同大学院を経て，松下電器産業（株）に入社．1989年～95年，(財) 新世代コンピュータ技術開発機構（ICOT）に出向．東京工業大学大学院非常勤講師．現在，松下電器産業（株）マルチメディアシステム研究所勤務．人工知能学会，生物物理学会各会員．元岡賞受賞．博士（工学）．知識情報処理，とくに生物学への応用に興味をもつ．