2種類の部分木交叉オペレータを持つ遺伝的アルゴリズムによる最尤分子系統樹の探索

松田 秀雄，　山下 浩，　金田 悠紀夫
大阪大学基礎工学部情報科学科　神戸大学工学部情報知能工学科

分子系統樹とは，DNA塩基配列やタンパク質アミノ酸配列等の分子データをもとに作成した生物または遺伝子の系統樹である．分子系統樹の作成では，対象生物の数が増えると，構成可能な候補系統樹の数が組合せ的に増大する．このため，最尤法で言う最適な系統樹を作成するには，何らかの発見的または近似的な探索手法が必要となる．我々は以前に，系統樹上の配列間距離に基づく部分木交叉オペレータを持つ遺伝的アルゴリズムにより尤度最大の候補系統樹を探索する手法について報告したが，今回新たな部分木交叉オペレータを加えた遺伝的アルゴリズムを用いた結果について報告する．

Exploring the Maximum Likelihood Phylogenetic Tree Using a Genetic Algorithm with Two Subtree-Crossover Operators

Hideo Matsuda,
Department of Information & Computer Sciences,
Faculty of Engineering Science,
Osaka University

Hiroshi Yamashita and Yukio Kaneda,
Department of Computer & Systems Engineering,
Faculty of Engineering,
Kobe University

Molecular phylogenetic trees are constructed for indicating the evolutionary process of organisms or genes from sequences of DNA nucleotides or protein amino-acids. Since the number of possible alternative trees rapidly increases when the number of sequences becomes large, some type of heuristic search is required to search for the maximum likelihood tree for all alternative trees. In this problem, we present some results using a genetic algorithm that has two crossover operators; an operator combining two subtrees based on inter-sequence distances, which are previously developed by us, and a new operator exchanging similar subtrees between any two alternative trees.

1 はじめに

DNA の解析技術のめざましい発展とともに，DNA 塩基配列やアミノ酸配列などの分子レベルのデータから生物の系統関係を解析する手法が注目されてきている．このような分子レベルのデータに基づく系統分類では，対象生物の進化過程を表す系統樹（従来の表現型による系統樹と区別するため分子系統樹と呼ぶ）の作成が中心となる．特に細菌などの微生物の系統関係では，表現型では区別できない場合があるなどの理由から，分子系統樹による系統解析がさかんに行われている．

分子系統樹の作成は，基本的には，まず分類の対象となる生物の間で共通した機能を持つ分子データ（たとえば，同じ遺伝子の情報を持つ DNA 塩基配列やそれらが翻訳された結果生成されたアミノ酸配列）を選び出し，それらを解析して，対象生物間の系統関係を推定することにより行われる．

分子系統樹の作成法は，現在までに数多く提案されているが [1]，我々はそれらの中で最尤法 [2] について研究してきた [3]．最尤法は，対象生物の DNA 塩基配列（またはアミノ酸配列）を葉とする木として構成された候補系統樹をモデルとして与え，そのモデルのもとで，進化の過程で起こる DNA 塩基（またはアミノ酸）の置換によって，対象生物の配列が実現される確率（これを尤度と呼ぶ）を求める方法である．尤度は候補系統樹ごとに異なるので，尤度最大の候補系統樹を真の分子系統樹の最もよい候補として選ぶ．候補系統樹どうしを尤度により比較できることから，最尤法は現在までに提案されている分子系統樹作成法の中では最も定量的な解析法として知られている（詳細は 2節参照）．

しかし，構成可能な候補系統樹の数は，対象生物の数が増えると急激に増大することが知られており，何らかの発見的探索法が必要になる．さらに，文献 [3] では，候補系統樹を一種の山登り法により探索した結果，多数の局所最適解が存在することが報告されている．

我々は，尤度最大の候補系統樹を，組合せ最適化問題の代表的な近似探索アルゴリズムのひとつである遺伝的アルゴリズム [4]（以下，GA と略す）により求める方法について研究してきた [5, 6]．本稿では，文献 [6] の方法に対して，さらに別の交叉オペレータを加え，探索効率を向上させることを試みたので報告する．

2 分子系統樹の作成法

分子系統樹を作成するための入力データとしては，対象生物の DNA または RNA の塩基配列やアミノ酸配列が用いられる．現状では，すべての DNA 塩基配列が解読された生物はわずかなので，対象生物の間でそれらの DNA 全部を比較して系統関係を解析することはほとんど不可能であり，一部分の配列を切り出して解析することになる．ここで，各生物から切り出してくる部分配列は，各生物の進化の過程を反映するような部分配列，すなわちそれらの生物間で共通の祖先種における単一の部分配列から受け継がれた配列になっている必要がある．これは，実際には各生物の間で共通の機能を持つ遺伝子の DNA 塩基配列（またはアミノ酸配列）を選び出すことが広く行われている．

このように共通な機能を持つ DNA 塩基配列やアミノ酸配列であっても，進化の過程での塩基の欠損/挿入によりそれらの長さが多少異なっているので，系統樹作成に先だって，多重アラインメント（multiple alignment）と呼ばれる処理を行い，DNA 塩基配列またはアミノ酸配列中の相同な部分が同じ位置に来るようにギャップを入れて補正する．これにより，分子系統樹の作成の際には，配列間での DNA 塩基またはアミノ酸の置換だけを考えればよいことになる（欠損は塩基またはアミノ酸からのギャップへの置換，挿入はギャップから塩基またはアミノ酸への置換と見なせる）．

図 1に多重アラインメントをかけた後のアミノ酸配列の例を示す．図 1は，1 種類の古細菌（硫黄依存好熱菌，学名 *Sulfolobus acidocaldarius*）と 6 種類の真核生物（カンジタ酵母，タマホコリカビ，ミドリムシの一種，赤痢アメーバ，マラリア病原虫，パン酵母．学名はそれぞれ，

```
カンジタ酵母    GGTGEFEAGISKDGQTREHALLAYTLGVKQLIVAVNKMDS--VKWDKNRFEEIIKETSNF
タマホコリカビ  SPTGEFEAGIAKNGQTREHALLAYTLGVKQMIVAINKMDEKSTNYSQARYDEIVKEVSSF
ミドリムシ      STTGGFEAGISKDGQTREHALLAYTLGVKQMIVATNKFDDKTVKYSQARYEEIKKEVSGY
赤痢アメーバ    AGTGEFEAGISKNGQTREHILLSYTLGVKQMIVGVNKMDA--IQYKQERYEEIKKEISAF
マラリア病原虫  ADVGGFDGAFSKEGQTKEHVLLAFTLGVKQIVVGVNKMDT--VKYSEDRYEEIKKEVKDY
硫黄依存好熱菌  AKKGEYEAGMSAEGQTREHIILSKTMGINQVIVAINKMDLADTPYDEKRFKEIVDTVSKF
パン酵母        GGVGEFEAGISKDGQTREHALLAFTLGVRQLIVAVNKMDS--VKWDESRFQEIVKETSNF
```

図 1: 配列データの例（一部）

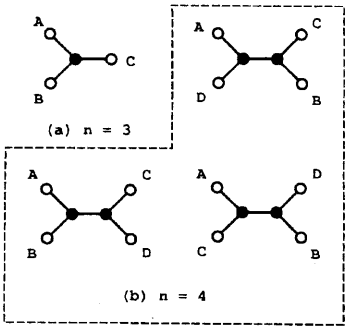


図 2: 無根木で表された系統樹

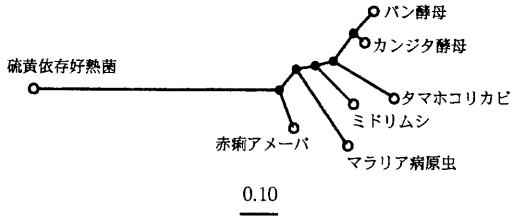


図 3: 図 1 の配列から構成した分子系統樹の例

Candida albicans, Dictyostelium discoideum, Euglena gracilis, Entamoeba histolytica, Plasmodium falciparum, Saccharomyces cerevisiae) のアミノ酸配列のうちの一部を取り出して横に並べたものであり，ギャップが "-" で表されている．

図 2 (a), (b) に，それぞれ，対象生物の数が 3 のときの系統樹と，対象生物の数が 4 のときの分子系統樹を示す．分子系統樹を木とみたときの葉節点を○，葉でない内部節点を●で表している．○は前述のアライメントされた配列に対応し，●は進化の過程で生物種の分岐が起こった位置，すなわち過去に存在したであろう生物（の配列）を表している．また，節点間の枝は，その両端の配列どうしを置換により関連付けている．枝の長さが置換回数を表しており，一方の端の配列からその長さで表された回数の置換が起こると，もう一方の端の配列に変化することを示す．

表現型による進化系統樹と違って，分子系統樹は図 2 のように根（対象生物すべての共通祖先の配列を表す）を持たない無根木で構成されることが多い．これは，DNA 塩基またはアミノ酸の置換だけの解析からでは，ある枝において進化の過程でどちらの方向に置換が起こったかの方向性が決定できないことによる．ただし，無根木の系統樹であっても，対象生物のいずれとも系統関係において大きくかけ離れている生物（これを群外種（outgroup）と呼ぶ）を選ぶことができれば，対象生物と群外種の共通の祖先の位置を対象生物だけからなる部分系統樹の根と見なすことができる．

図 3 に，図 1 に示したアミノ酸配列から作成した分子系統樹を示す．硫黄依存好熱菌は他の生物とは系統的に大きくかけ離れていることが知られているので，図 3ではこれを群外種として根の位置を決めている．なお，図 3では下の 0.10 とラベルがついた線の長さが，配列の各位置あたり平均 0.10 回のアミノ酸置換に対応する長さを示している．

分子系統樹の作成法としては，現在までに様々な方法が提案されており，距離行列法，最大節約法，最尤法に大きく分けられる [1]．距離行列

法とは，与えられた配列から，まずDNA塩基（またはアミノ酸）の置換回数に基づいて配列相互間の距離を計算し，それらの距離の情報から分子系統樹を作成するという2段階の手順をとる方法である．これに対して，最大節約法と最尤法は，配列を距離情報に変換せずにそのまま使って，構成可能な多数の候補系統樹の中から，ある尺度で最良と思われるものを選び出す方法である．その尺度は，最大節約法では候補系統樹での配列の置換回数（最小のものを選ぶ）であり，最尤法ではあらかじめ与えられたDNA塩基（またはアミノ酸）の置換確率に基づき計算された，候補系統樹の実現確率（これを尤度と呼び，尤度が最大のものを選ぶ）である．

これらの方法は，それぞれに一長一短があり，現時点ではどれが一番良いかは決められない．しかも，どの方法も，進化の過程で起こる置換に対してそれぞれ異なる仮定を導入しており，対象となる生物やそれらから切り出した配列によってその仮定が適当かどうかが変わってくる．

たとえば，距離行列法は配列そのものではなく，配列間の置換回数から計算された距離を使って分子系統樹を作成するため，配列のどの位置に置換が起こっているかという情報は考慮しない．つまり，進化の過程で系統によって配列上での置換の位置が大きく異なるような場合には向かない．

また，最大節約法は，大きく系統の異なる生物間で分子系統樹を作成するとき（すなわち，配列間での置換が数多く起こっていると考えられる場合）には，置換回数を過少評価する傾向があるので，このような場合には向かないとされている．

最尤法については，仮定した置換確率が実際の進化における置換の起こり方と異なっている場合に問題が生じる．しかし，最尤法ではこのようなときでも，多くの場合において正しい候補系統樹を選ぶという頑健性（robustness）を持つことが，いくつかのシミュレーション実験から分かっている[7]．

3 分子系統樹作成への遺伝的アルゴリズムの適用

3.1 遺伝的アルゴリズム

GAは，生物進化の遺伝学的説明をもとにした，組合せ最適化問題の解法のひとつである．生物の進化過程においては，ある世代を構成している個体群から次の世代を生成する段階で，遺伝情報の突然変異や交叉によって元とは異なる個体が発生し，環境へより適応する個体が次の世代を構成していく．

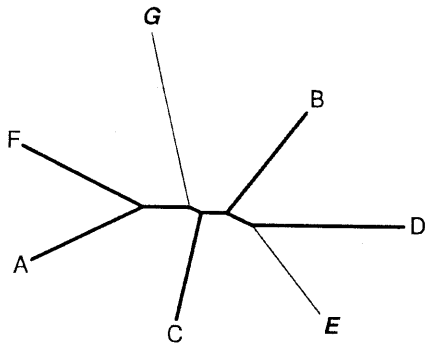
GAでは，各個体は問題に対する解を，遺伝情報として1つ保持している．この遺伝情報は，突然変異や交叉といった操作が行える形でなければならず，また，適合度と呼ばれる関数が定義される．適合度は，問題に対する遺伝情報の優劣の判断に用いられる．このように表される個体の集団が世代交替を繰り返すことによって，最適解の探索が行われる．

各世代では，まず，個体の適合度が計算され，適合度に依存して個体の選択（増殖，淘汰）が行われる．次に，適当な2個体において，遺伝情報の交叉が行われる．さらに，各個体において，遺伝情報の突然変異が行われる．最後に，終了条件の判定が行われ，条件を満たした場合，その時点での適合度が最も高い個体が問題の準最適解となる．

本研究では，選択については単純GAと同じルーレット選択方式を採用した．ルーレット選択方式では，集団における全個体の適合度の和に対する各個体の適合度の割合によって，次の世代の個体として選択される確率を決定している．なお，各世代において集団の中で最も適合度の高い個体は，後の交叉，突然変異で消えてしまうことがないようにしている（エリート保存）．

3.2 距離差最大交叉オペレータ

我々は以前に，最尤法に基づき対数尤度最大の候補系統樹を探索する手法を，単純GA[4]をもとに開発している[6]．この方法では，1個体で1つの候補系統樹を表すことにし，候補系統樹から個体の遺伝情報へのコード化について

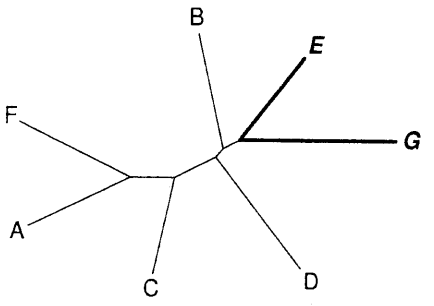


距離行列

	A	B	C	D	E	F
A						
B	0.4548					
C	0.3656	0.3585				
D	0.6495	0.6053	0.5532			
E	0.4119	0.3676	0.3155	0.5232		
F	0.4435	0.5138	0.4246	0.7085	0.4709	
G	0.8621	0.8709	0.7817	1.0656	0.8280	0.9211

対数尤度= -4470.4

(a) 凸凹木

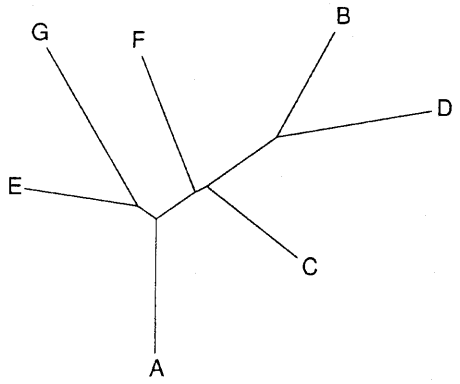


距離行列

	A	B	C	D	E	F
A						
B	0.4673					
C	0.3543	0.3687				
D	0.6467	0.5936	0.5481			
E	0.4243	0.3621	0.3258	0.5507		
F	0.4431	0.5241	0.4111	0.7035	0.4811	
G	0.9037	0.8414	0.8051	1.0300	0.7758	0.9604

対数尤度= -4468.9

(b) 参照木

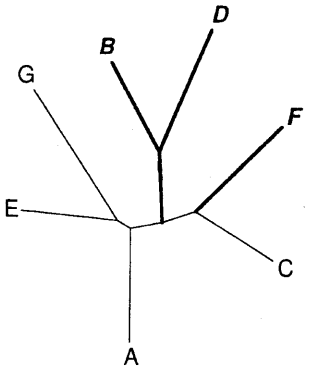


距離行列

	A	B	C	D	E	F
A						
B	0.4565					
C	0.3672	0.3559				
D	0.6593	0.5096	0.5587			
E	0.3579	0.4044	0.3151	0.6072		
F	0.5048	0.4936	0.4043	0.6964	0.4527	
G	0.8520	0.8985	0.8092	1.1013	0.7672	0.9468

対数尤度 = -4461.9

(c) 距離差最大交叉



距離行列

	A	B	C	D	E	F
A						
B	0.4443					
C	0.3743	0.3624				
D	0.6481	0.5070	0.5662			
E	0.3556	0.3924	0.3225	0.5963		
F	0.5135	0.5016	0.3797	0.7055	0.4617	
G	0.8467	0.8835	0.8135	1.0873	0.7652	0.9528

対数尤度 = -4457.9

(d) 突然変異

図 4: 距離差最大交叉オペレータおよび突然変異オペレータの実行例

は，単純 GA のように 2 進数に変換するのではなく，候補系統樹をグラフ表現でそのまま表している．具体的には，葉節点を正の整数，内部節点を負の整数でそれぞれ番号付けし，それらの間の接続関係を内部接点に隣接した節点番号の組を，内部節点の数だけ並べたリストで表現している．

個体の持つ遺伝情報の表現が 2 進数ではないため，交叉，突然変異のオペレータには，単純 GA で使われるような 2 進数列の部分的入れ換えやビットの反転といった手法は使えない．そこで，文献 [6] では，交叉オペレータとして以下のような独自の操作を導入している（以下，これを距離差最大交叉オペレータと呼ぶ）．

まず，交叉の対象となる 2 つの候補系統樹のそれぞれで各葉節点間の距離を行列の形にまとめる（図 4の (a), (b)）．次に，このようにして得られた 2 つの行列の同一成分同士を比較し，最も値の異なる成分，すなわち交叉の対象となる 2 つの候補系統樹の間で葉節点間の距離が最もくい違っているところを探す（図 4では Eと G）．得られた 2 つの葉節点で距離の近い方を持つ候補系統樹（参照木と呼ぶ）からこの 2 つの葉節点を含む最小の部分木（図 4 (b) の太線部分）を取りだし，もう一方の候補系統樹（凸凹木と呼ぶ）からこの部分木に含まれる葉節点を取り除いてできる部分木（図 4 (a) の太線部分）とを連結して新たな候補系統樹を作る（図 4 (c)）．これが距離差最大交叉オペレータにより得られた候補系統樹となる．

なお，最後の連結操作では 2 つの部分木のどこをつなぐかが問題になるが，ここではあらかじめ基準点となる葉節点を決めておいて，元の凸凹木，参照木における基準点との接続関係が保存されるように連結することにする（図 4 (c) では葉節点 A を基準点にとっている）．

以上の操作で距離差最大の部分木を探すのは以下の理由による．葉節点間の距離が離れると，それらの葉節点に対応している配列の各位置での遷移確率が低くなり，対数尤度が低下する．したがって，このような箇所を見つけてそれらを取り除いた候補系統樹を作ることができれば，より対数尤度の高い候補系統樹が得られる可能性が高いと考えられる．

次に，突然変異オペレータについては，内部節点どうしをつなぐ枝を 1 つ選んで，それにつながっている枝もしくは部分木どうしの入れ換えを行う（これを分枝交換と呼ぶ）．たとえば，図 4 (d) は (c) の葉節点 F につながる枝と葉節点 B，Dを持つ部分木とを入れ換えている．

3.3 類似部分木交叉オペレータ

文献 [6] では，前節の距離差最大交叉オペレータは GA の実行の開始からしばらくの間は，対数尤度の高い候補系統樹の探索に極めて有効なことが示されている．しかし，実行が進むにつれて，世代間での対数尤度の向上が頭打ちになる傾向が見られた．これは，各個体が持つ候補系統樹での葉節点間距離のばらつきが次第に減少していくためと考えられる．

そこで，我々は別の交叉オペレータを考え，これと距離差最大交叉オペレータとを組み合わせて，GA の探索性能を上げることを試みた．この交叉オペレータ（以下，類似部分木交叉オペレータと呼ぶ）は，任意に選んだ 2 つの候補系統樹の間で，3 個以上の葉節点を持つ部分木のうち同じ葉節点かまたは 1 個だけ異なる葉節点を持つ部分木を相互に交換する．

例えば，図 5では 2 つの候補系統樹 (a), (b) はどちらも葉節点 A, B, C を持つ部分木を持っているので，これらを相互に交換して新たな候補系統樹 (a'), (b') を得る．図 5では，全く同じ葉節点を持つ部分木の交換の例であるが，実際には，任意に選んだ 2 つの候補系統樹の間で，このような部分木が必ずしも存在するとは限らないので，1 個だけ葉節点が異なる部分木（例えば，A, B, C と A, B, D）の交換も認めることにした．このときは，どちらかの候補系統樹で，葉節点の付替えを行なって両方の部分木の持つ葉節点を同じにしてから部分木の交換を行う．

4 実行結果

図 6，図 7に EF1α（mRNA からアミノ酸への翻訳過程でのタンパク伸長因子の一つ）のアミノ酸配列 15 本 [10] およびσ因子（細菌におけ

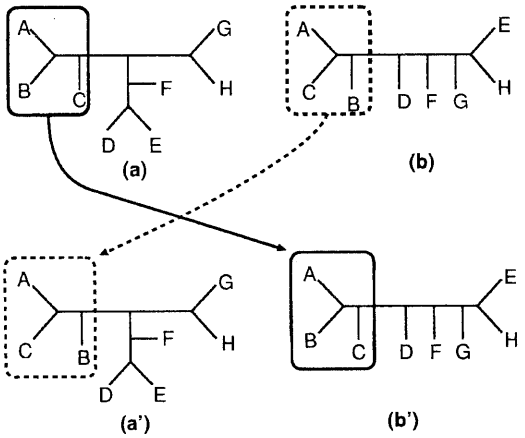


図 5: 類似部分木交叉オペレータの実行例

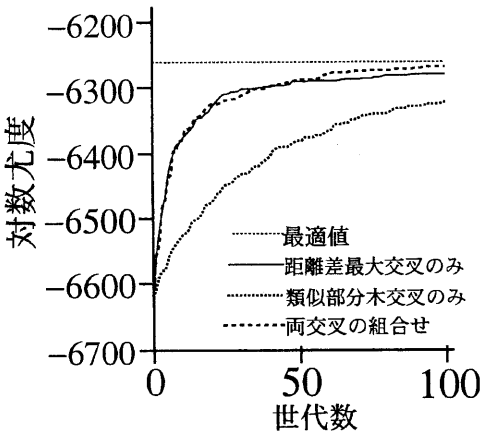


図 6: GA 実行時の対数尤度向上（EF1α）

る DNA から mRNA への転写因子の一つ）のアミノ酸配列 21 本 [11] から，それぞれ GA により対数尤度最大の分子系統樹を探索したときの各世代ごとの対数尤度の向上の様子を示す．

EF1α，σ因子ともに各世代の個体数は 50 とし，EF1αでは交叉率 0.4，突然変異率 0.1 で 100 世代で実行を打ち切り，σ因子では交叉率 0.5，突然変異率 0.1 で 400 世代で実行を打ち切った．なお，どちらの図も，各世代の対数尤度は，乱数の初期値を変えて 10 回実行したときの世代ごとの最大対数尤度の平均値で表している．また，2 つの交叉オペレータの組合せでは，1 世代ごとに交叉オペレータを切替えて実行した．なお，EF1αについては MOLPHY[9] を使った候補系統樹の全数探索により対数尤度の最大値が−6260.59065 であることがわかっているので，図 6ではその値を最適値として示している．

図 6，図 7からわかるように，距離差最大交叉オペレータだけだと早い時点から急速に対数尤度が向上するがすぐに頭打ちになる．これに対して，類似部分木交叉オペレータでは対数尤度の向上は緩やかであるが確実に向上しているのがわかる．図 6では，両交叉を組み合わせることにより，41 世代目以降で距離差最大交叉オペレータだけのときよりも高い対数尤度の候補系統樹を見つけている．また，10 回の実行のう

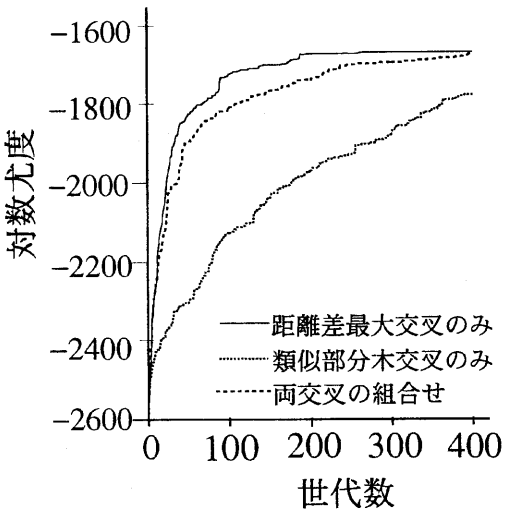


図 7: GA 実行時の対数尤度向上（σ因子）

ち，距離差最大交叉オペレータ，類似部分木交叉オペレータともに単独では100世代目までに最適値に達することはなかったが，両者を組み合わせたときには3回（それぞれ30世代目，82世代目，99世代目）最適値に到達した．

図7では，400世代までででは距離差最大交叉オペレータだけの実行が最も対数尤度が高いが，図6と同様の傾向が現れており，さらに世代を進めることにより，両交叉の組合せの方がより対数尤度の高い候補系統樹を探索するようになる可能性が高い．また，10回実行したうちでの対数尤度の最大値は，距離差最大交叉オペレータだけの実行では−1653.78788であったが，2種類の交叉オペレータを組み合わせたときには−1653.33089であり上回っている．

5 おわりに

本稿では，最尤法により求められる対数尤度を評価値としてGAにより最適な分子系統樹を探索する手法について述べた．我々が先に開発した，距離差最大交叉オペレータの欠点であった，実行が進むにつれて対数尤度の向上が緩やかになる状況を改善する試みとして，類似部分木交叉オペレータという新たなオペレータを考え，その実装を行った．

2種類のアミノ酸配列集合からそれぞれ分子系統樹を作成したところ，距離差最大交叉オペレータと類似部分木交叉オペレータを1世代おきに交互に実行する方式は，距離差最大交叉オペレータだけのものと比べ，最適値に近づいたときの対数尤度の向上が加速されており，より早い世代で最適な分子系統樹を発見できる可能性が高いことが示された．

今後の課題としては，2種類の交叉オペレータを単純に世代ごとに切替えるのではなく，状況に応じてより効果的に使いわけ，さらに対数尤度の向上を加速することなどがあげられる．

謝辞

本研究は一部，文部省科学研究費補助金重点領域研究（課題番号 08283103）および奨励研究（課題番号 08780355）によっている．

参考文献

1) 日本生化学会（編）：分子進化実験法，第23章 系統樹作成法，東京化学同人，pp. 373–416 (1993).

2) Felsenstein, J.: Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach, *J. of Molecular Evolution*, Vol. 17, pp. 368–376 (1981).

3) Olsen, G. J., Matsuda, H., Hagstrom, R. and Overbeek, R.: fastDNAml: A Tool for Construction of Phylogenetic Trees of DNA Sequences using Maximum Likelihood, *Computer Applications in Biosciences*, Vol. 10, No. 1, pp. 41–48 (1994).

4) Goldberg, D. E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addisson-Wesley (1989).

5) 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 情報学基礎 95-FI-36-3, pp. 15–22 (1995).

6) 川本芳久，松田秀雄，橋本昭洋：遺伝的アルゴリズムによる分子系統樹の作成，情報処理学会論文誌，Vol. 37, No. 6, pp. 1107–1116 (1996).

7) Fukami-Kobayashi, K. and Tateno, Y.: Robustness of Maximum Likelihood Tree Estimation against Different Patterns of Base Substitutions, *J. of Molecular Evolution*, Vol. 32, pp. 79–91 (1991).

8) Felsenstein, J.: PHYLIP Manual Version 3.3, University herbarium, University of Carifornia, Berkeley (1990).

9) Adachi, J. and Hasegawa, M.: MOLPHY: Programs for Molecular Phylogenetics I – PROTML: Maximum Likelihood Inference of Protein Phylogeny, Computer Science Monographs 27, Institute of Statistical Mathematics, Tokyo (1992).

10) Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N. and Miyata, T.: Early Branchings in the Evolution of Eukaryotes: Ancient Divergence of Entamoeba that Lacks Mitochondria Revealed by Protein Sequence Data, *J. of Molecular Evolution*, Vol. 36, pp. 380–388 (1993).

11) Nakahigashi, K., Yanagi, H. and Yura, T.: Isolation and Sequence Analysis of RpoH Genes Encoding Sigma32 Homologs from Gram-negative Bacteria: Conserved mRNA and Protein Segments for Heat Shock Regulation, *Nucleic Acids Research*, Vol. 23, No. 21, pp. 4383–4390 (1995).