遺伝的アルゴリズムによる最尤系統樹の探索

研究ノート

松　田　秀　雄*

Search for Maximum Likelihood Phylogenetic Tree Using A Genetic Algorithm

Key Words : Phylogenetic tree, genetic algorithm, bioinformatics

```
C.albicans      GGTGEFEAGISKDGQTREHALLAYTLGVKQLIVAVNKMDS--VKWDKNRFEEIIKETSNF
D.discoideum    SPTGEFEAGIAKNGQTREHALLAYTLGVKQMIVAINKMDEKSTNYSQARYDEIVKEVSSF
E.gracilis      STTGGFEAGISKDGQTREHALLAYTLGVKQMIVATNKFDDKTVKYSQARYEEIKKEVSGY
E.histolytica   AGTGEFEAGISKNGQTREHILLSYTLGVKQMIVGVNKMDA--IQYKQERYEEIKKEISAF
P.falciparum    ADVGGFDGAFSKEGQTKEHVLLAFTLGVKQIVVGVNKMDT--VKYSEDRYEEIKKEVKDY
S.acidocaldarius AKKGEYEAGMSAEGQTREHIILSKTMGINQVIVAINKMDLADTPYDEKRFKEIVDTVSKF
S.cerevisiae    GGVGEFEAGISKDGQTREHALLAFTLGVRQLIVAVNKMDS--VKWDESRFQEIVKETSNF
                   *  .        **.**  *. *.*. *..*  ** *          *. *.
```

図1　分子系統樹作成のための入力データの例(一部)

1．は　じ　め　に

遺伝的アルゴリズム(以下，GAと略す)は，生物の進化過程の遺伝学的説明をもとにした探索アルゴリズムの一つであり，機械学習や種々の最適化問題での解探索，パターン認識など数多くの問題に幅広く応用されている[1]．

本稿では，生物の進化過程からの類推で始まったGAを，逆に生物の進化解析の一分野である分子系統樹作成に応用する試みについて紹介する．分子系統樹作成では，現在の生物から抽出したDNA塩基配列(またはタンパク質アミノ酸配列)の集合を入力として，その配列集合が形成されるにはどのような進化過程が存在したかを推定する．分子系統樹を作成する問題は一種の組合せ最適化問題として定式化でき，これを解くための探索アルゴリズムとしてGAを使うことができる．

2．分子系統樹の作成

分子系統樹を作成するための入力データとしては，対象生物の間で共通な遺伝子のDNA塩基配列(またはそれを翻訳した結果，生成されるタンパク質のアミノ酸配列)が使われる．これらの配列データでは，進化の過程での塩基やアミノ酸の欠損/挿入を補正するため，あらかじめ多重アラインメント(multiple alignment)と呼ばれる整列処理を行う．

図1に多重アラインメントを行った後のアミノ酸配列の例を示す．図1は，1種類の古細菌*S. acidocaldarius*(硫黄依存好熱菌)と6種類の真核生物*C. albicans*(カンジタ酵母)，*D. discoideum*(細胞性粘菌)，*E. gracilis*(ミドリムシ)，*E. histolytica*(赤痢アメーバ)，*P. falciparum*(マラリア病原虫)，*S. cerevisiae*(パン酵母)のタンパク伸長因子EF-1αのアミノ酸配列から，

* Hideo MATSUDA
1959年11月7日生
1987年神戸大学大学院自然科学研究科システム科学専攻修了
現在，大阪大学大学院基礎工学研究科，情報数理系専攻，計算機科学分野，助教授，学術博士，生物情報科学
TEL　06－850－6601
FAX　06－850－6602
E-Mail matsuda@ics.es.osaka-u.ac.jp

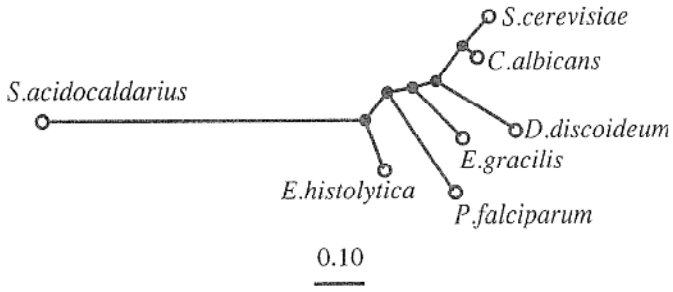

図2　図1の配列から構成した分子系統樹の例

一部を取り出して並べたもので，20種類のアミノ酸を英字1文字のコードで，ギャップを“－”で表している．

図2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．図2で○(葉節点)は前述のアライメントされた配列に対応し，●(内部節点)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．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．また，*S. acidocaldarius*(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，これを群外種(outgroup)として根の位置を決めている．

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いるが[2]，本研究では，最尤法を取りあげた．最尤法は，与えられた配列データから構成可能な系統樹の候補(以下，候補系統樹と呼ぶ)の中から，配列置換をモデル化した確率モデルに基づき計算される尤度が最大のものを選ぶ方法である．現在提案されている中では，最尤法は最も定量的な方法であるが，膨大な候補系統樹の中から尤度最大のものを探索するのが困難なことから，効率の良い探索手法の開発が望まれていた．以下では，この探索にGAを適用した結果について述べる．

3．最尤法へのGAの適用

GAの考え方の基本には積木仮説(building block hypothesis)がある[1]．これは，個体の遺伝情報中に積木，すなわち適合度の向上に寄与する短いコード・パターンがいくつか存在し，これらを組み合わせることでより適合度の高い個体を得ることができるというものである．

尤度最大の候補系統樹を探索するGAにおいて，積木に相当するのは，尤度の向上に寄与するような位置に葉節点(現在の生物の配列)が配置された部分木であると考えられる．そこで，このような部分木を直接的に表現できるコード化手法を考えた．具体的には，個々の候補系統樹をGAでの個体として，各個体を単純GA[1]のように数値列でコード化するのではなく，グラフ表現で表すことにした．すなわち，各節点に番号を付け，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のリストで表現している．これにより，任意の部分木は，その中に含まれている内部節点の各々について，隣接している節点の組を番号で表し，それらを内部節点の数だけ並べたリストで表現できる．

しかし，グラフ表現に基づくコード化で個体を表すことにすると，交叉および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，距離差最大交叉および類似部分木交叉という2種類の交叉オペレータ，および分枝交換に基づく突然変異オペレータを独自に開発した[3),4)]．なお，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．

4．実　行　結　果

系統樹のグラフ表現によるGAの有効性を調べるため，GAと従来の分子系統樹作成法(平均距離法，近隣結合法，最大節約法の各手法と，最尤法では逐次付加法および星状系統樹分割法により候補系統樹を探索するもの)で実際に分子系統樹を作成した結果をそれらの尤度の対数値(対数尤度)により比較した(表1)．

分子系統樹作成のための配列データには，前

表1　対数尤度による手法ごとの結果の比較

分子系統樹作成法	対数尤度(EF-1α)
平均距離法	−6301.2
近隣結合法	−6301.8
最大節約法	−6267.7 .. −6270.3
最尤法(星状系統樹分割法)	−6309.6
最尤法(逐次付加法)	−6260.6 .. −6998.8
最尤法(GA, 距離差最大交叉のみ)	−6260.6 .. −6321.9
最尤法(GA, 類似部分木交叉のみ)	−6283.1 .. −6426.6
最尤法(GA, 両交叉の組合せ)	−6260.6 .. −6288.9

述のEF-1αのアミノ酸配列を15種類の生物から取り出したものを使用した[3].

表1で，最大節約法では置換回数最小の系統樹が複数得られたためそれらの対数尤度の最大値と最小値を示している．また，逐次付加法で配列を付加する順番に結果が依存するため，配列の選択順序を乱数によりランダムに変えて20回実行している．同様に，GAでも選択，交叉，突然変異の各オペレータでいずれも乱数を使うので，乱数の種を変えて10回実行している．なお，GAによる探索では，各世代の個体数を50，交叉率を0.4，突然変異率を0.1にして100世代目で実行を打ち切った．

表1では，2種類の交叉オペレータを組み合わせたGAにより候補系統樹を探索する最尤法が，対数尤度で比較して最も良い結果を出している．なお，この例に関しては候補系統樹の全数探索により対数尤度の最大値が−6260.6であることがわかっている．

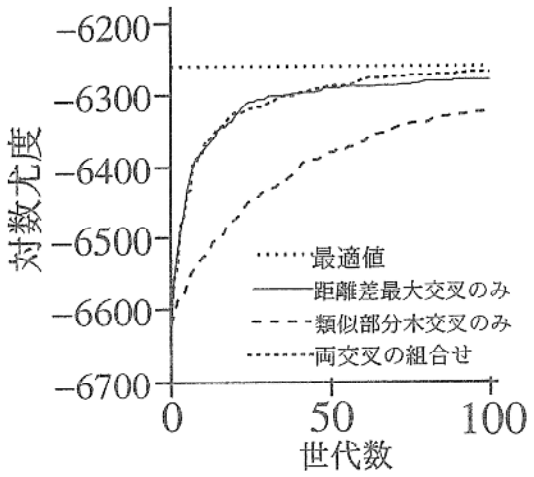

図3　GA実行時の対数尤度向上(EF-1α)

次に，2種類の交叉オペレータの組合せの効果を調べるために，図3に各世代ごとの対数尤度の向上の様子を示す．各世代の対数尤度は，乱数の初期値を変えて10回実行したときの各世代での最大対数尤度の平均値で表している．なお，図3では対数尤度の最大値−6260.6を最適値として示している．

図3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3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(距離差最大交叉オペレータのみの実行で1回だけ最適値と対数尤度で小数第二位が違うだけのものが出ている)が，両者を組み合わせたときは3回，最適値に到達した．

なお，GAによる探索の実行時間については，実行時間の大半(90％以上)が対数尤度の計算で占められるため，入力データとなる配列，世代数，個体数，交叉率，突然変異率が同じであれば，交叉方式の違いによる実行時間の差はほとんどなかった．上記の設定では，DEC Alpha Station 600 (CPU Alpha 21164, Clock 333 MHz)でのユーザCPU時間で，1世代当たり平均して約44秒であった．

5．おわりに

本章では，分子系統樹を作成するための一手法である最尤法に遺伝的アルゴリズム(GA)を組み込み，尤度最大の候補系統樹を探索する方法について述べた．

最尤法は，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が，取り扱う生物種の数が増えると候補系統樹の数が急激に

増大することから，多数の局所最適解が存在することが知られている．

一方，GAは，組合せ最適化問題の代表的な近似解法である．このGAを最尤法に適用することで，最尤法の欠点を補いつつ，従来の手法と比べてより尤度の高い候補系統樹を得ることが可能となった．

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「ゲノムサイエンス」によっている．

参　考　文　献

1) Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addisson-Wesley (1989).
2) 日本生化学会編，分子進化実験法，東京化学同人(1993).
3) 川本芳久，松田秀雄，橋本昭洋，情報処理学会論文誌，37，1107 (1996).
4) 松田秀雄，山下　浩，金田悠紀夫，情処研報96-MPS-10，49(1996).