様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　4 月　30 日現在

研究種目：基盤研究（C）
研究期間：2006～2008
課題番号：１８５７０２１７
研究課題名（和文）複合尤度を用いた巨大分子系統樹推定技術の開発
研究課題名（英文）Estimation of Large Phylogeny Using Maximum Composite Likelihood.

研究代表者
田村　浩一郎（TAMURA KOICHIRO）
首都大学東京・大学院理工学研究科・教授
研究者番号：00254144

研究成果の概要：
分子系統樹の推定は生命科学の様々な分野で利用される基本的な操作であるが、その方法は理論的側面に長ける最尤法と計算効率に長ける近隣結合法に二分される。そこで、新たに開発した最大複合尤度法を利用し、両者の優れた点を合わせることを試みた。効果はコンピュータ・シミュレーションによって評価した。その結果、最大複合尤度法の適用により、近隣結合法、最尤法のいずれにおいても推定精度、計算時間の改良に成功した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,100,000	0	1,100,000
2007 年度	1,000,000	300,000	1,300,000
2008 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,600,000	450,000	3,050,000

研究分野：　生物学
科研費の分科・細目：　生物科学・進化生物学
キーワード：　複合尤度、最尤法、系統解析、分子系統樹、分子進化、進化距離、近隣結合法

１．研究開始当初の背景

(1) ゲノムプロジェクトを始めとする近年の分子生物学の発展により、大量の DNA・タンパク質配列データが遺伝子データバンクに蓄積されるに至った。分子系統樹の必要性はもはや系統分類学的解析だけにとどまらず、遺伝子ファミリーやゲノムの進化、遺伝子機能の進化学的解析、進化発生学など様々な分野で利用されるようになった。

(2) 分子系統樹推定の方法論は主に 1980 年代に発展し、種々の方法が開発されたが、現在、その中で最尤法(Felsenstein 1981)と近隣結合法(Saitou and Nei 1987)が主に使用されるようになった。最尤法は推定精度が高い、分子進化モデルの自由度が高い、尤度を基準とした統計的評価ができるなど、理論的側面において優れているが、計算時間が極端に長くかかるという応用面での大きな欠点がある。そのため、大量配列データを活用した大規模なデータ解析には事実上使用不可能であった。一方、近隣結合法の最大の利点は計算の高速性で、大量配列データの解析にも十分使用可能であり、精度の高さも最尤法

に比べて大きな遜色はない。しかし、近隣結合法も大量配列データ解析においては問題が生じる。例えば、DNA 進化距離推定法に用いられる対数関数の変数が確率的誤差によって負の値になると計算ができず、系統樹推定に必要な距離行列が得られなくなることがある。また、距離行列作成のために用いられる２配列間の進化距離推定法は、複雑な進化パターンをモデル化する柔軟性に劣るため、単純化したモデルに基づかざるを得ず、大量配列データから得られる詳細な進化情報を生かすことができない。さらに、距離行列の計算は配列の数の二乗に比例し、近隣結合法の計算時間は三乗に比例するため、配列の数が多くなると計算時間のメリットも失われる。

２．研究の目的

(1) 最大複合尤度法の開発

複合尤度(composite likelihood)を用いた最尤法を用いて距離行列を構成する全ての 2 配列間距離を一括推定することにより、近隣結合法（距離行列法）に最尤法の利点を導入する（最大複合尤度法)。距離行列の推定に最尤法の採用により、複雑な進化パターンもモデル化することが可能になり、モデルの評価・選択も行うことが理論的には可能である。一方、最尤法の計算の中でも、複合尤度を用いることにより、計算精度の改良、および計算時間の短縮をはかり、最尤法の欠点を克服する。

(2) 最大複合尤度法の検証

コンピュータ・シミュレーションを用いて通常の距離行列推定法や最尤法と比較することにより、分子系統解析法における複合尤度を用いたパラメータ推定法の有用性を検証する。

(3) 最大複合尤度法の普及

複合尤度を利用して推定精度や計算時間を改善した系統樹推定法を実装したプログラムを開発し、一般公開することにより、多くの研究者が配列データを用いた系統解析を行うことを可能にし、生命科学の一層の発展に寄与する。

３．研究の方法

(1) 最大複合尤度法の開発と検証

本研究を通じて、いろいろな最大複合尤度法を実装した系統解析プログラムを開発し、その推定精度、計算時間などを、以下のコンピュータ・シミュレーションを用いて生成した DNA 塩基配列データを用いて検証した。

本研究で行ったコンピュータ・シミレーションでは、主な哺乳類 66 種の系統樹をモデルとし、実際のデータ解析から得られた 448 種類のタンパク質コード遺伝子の進化パターンを模倣した。また、配列数と系統樹推定精度の関係を調べるため､配列数が 64､256、1024、4096 の完全対称樹形も用いた。一つの遺伝子について 100 回繰返し、100 セットの配列データを得た。これらの配列データをもとに、最大複合尤度法によって系統樹推定を行い、進化パラメータと樹形の推定精度、および計算時間を測定し、最尤法による推定の場合と比較した。

また、最尤法の計算に必要なパラメータの初期値を最大複合尤度法によって推定するため、新たに最尤系統樹推定プログラムを開発した。このプログラムを用いて、コンピュータ・シミュレーションによって得た配列データから最尤系統樹を推定し、その計算精度、計算時間を従来の方法（PhyML プログラム）を用いた場合と比較・検討した。

(2) 最大複合尤度法の普及

本研究で開発した最大複合尤度法を用いた系統樹推定法は、MEGA4 ソフトウェアに実装して Web ページで公開した。MEGA4 ソフトウェアは、誰でもインターネットを通じてダウンロードし、研究・教育などに活用できるようにした。同時に、MEGA4 ソフトウェアの使用方法を Web ページにおいて詳細に説明することにより、使用者の便宜をはかった。

４．研究成果

(1) 最大複合尤度法の推定精度

複合尤度を最大化する値を探索することによってパラメータ推定を行う最大複合尤度法を用い、コンピュータ・シミュレーションによって得た 448 遺伝子の仮想配列データについて、トランジション速度パラメータ（α）とトランスバージョン速度パラメータ（β）の比（κ）と、系統樹の各枝の長さの推定を行った。最大複合尤度法によって κ は極めて正確に推定できることが分かった（図１）。推定精度は最尤法とほぼ同一であった。

最大複合尤度法による系統樹の各枝の長さの計算は、2 配列距離を用いた最小二乗法によって行った。最大複合法を用いた場合、枝長の推定は、枝の長さによらず誤差が非常に小さく、100 回の繰り返しの平均はほぼ期待値に等しい（図 2 上)。

一方、PAUP を用いて最尤法で枝長を推定した場合、0.01 以下の枝長は、平均して 0.1～3%程度長く推定されることが分かった(図 2 下)。最尤法では、相互に関連する枝の長さを同時に最適化することができないため、枝長の推定が最も難しい問題の一つとされて

いる。この結果は、この最尤法の問題を反映したものと考えられる。

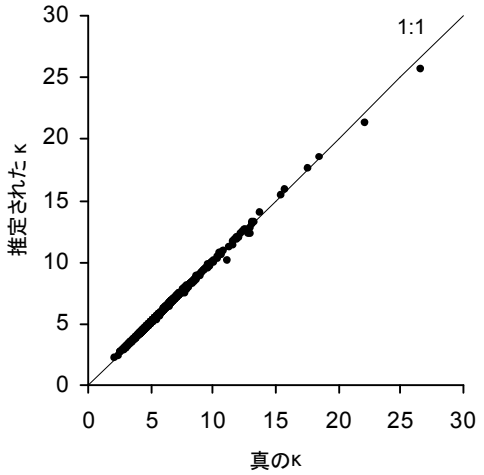

図 1. 最大複合尤度法によって推定したκ
各点は一つの遺伝子について 100 回の繰り返しの平均値を表す。

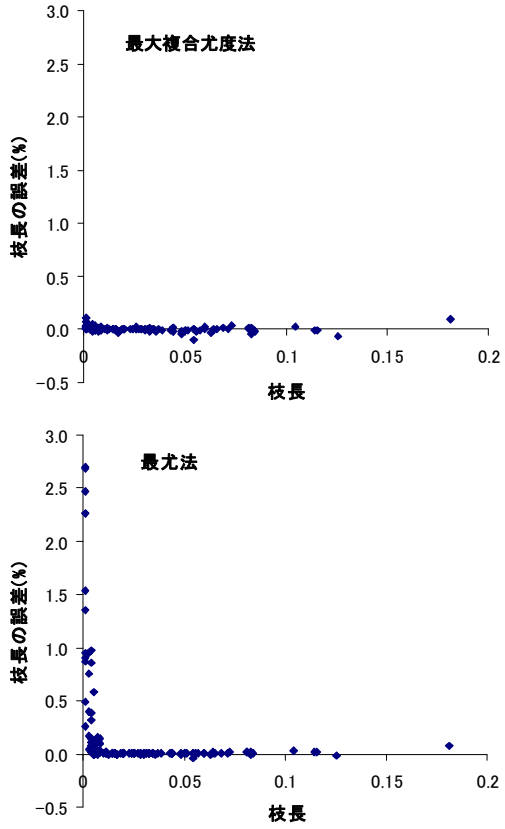

図 2. 最大複合尤度法と最尤法の系統樹枝長推定精度

(2) 最大複合尤度法のパラメータ計算時間

最大複合尤度法では距離行列を求めるため、計算時間は配列数の二乗に比例することが期待される。また、配列の各座位は最初に 1 回だけ比較され配列間の差異が求められるが、その時間は非常に短いため実際上問題とならず、配列の長さは計算時間全体にはほとんど影響しない。一方、最尤法では距離行列を作ることはなく、計算時間は配列数に比例する。また、全ての計算が座位ごとに行われるため、計算時間は配列の座位数にも比例する。結果として、計算時間は配列数と配列の座位数の積に比例することが期待される。

本研究のコンピュータ・シミュレーションでは配列数は一定(66 本)であったため、上記のように期待される通り、計算時間は最大複合尤度法では配列の長さに拠らずほぼ一定で、最尤法では配列の長さに直線比例する関係が得られた（図 3)。配列数が 66 の場合、塩基座数が 500 以上になると計算時間は最大複合尤度法の方が速く、塩基座数が 10,000 になると最大複合尤度法の計算速度は最尤法の 20 倍近くに達した。

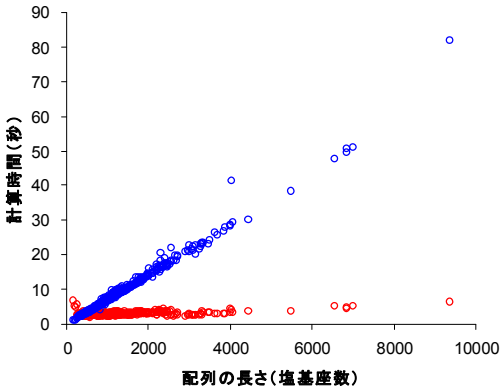

図 3. 最大複合尤度法（赤）と最尤法（青）による
κ および枝長の計算時間

(3) 最大複合尤度法と最尤法の樹形推定精度

コンピュータ・シミュレーションによって得た哺乳類 66 種 448 遺伝子の仮想配列データを用い、最大複合尤度法と最尤法によって系統樹推定を行い、正しく推定された枝の頻度を求めた。図 4 において、対角線より上の領域の点は最尤法の精度の方が高かった場合を示すが、448 遺伝子全てについて最尤法の優位性が示された。

そこで、どのような場合に最尤法の優位性が高いのかを明らかにするため、配列の長さ、進化速度、トランジション／トランスバージョン比、塩基組成の偏りなどと、最尤法と最大複合尤度法の樹形推定誤差の割合との関係を調べた。その結果、配列の長さのみが関連し（図 5)、その他の要因の影響は認められなかった。配列の長さが非常に短い場合、最尤法の樹形推定誤差は複合尤度法に対して 10～20%程度低いだけであるが、配列の長さが 4kb を超えると、誤差は最大複合尤度法の 30～40%になることが分かった。平均的な遺伝子のサイズ（1kb～2kb）においては、最尤法の樹形推定誤差は最大複合尤度法（近隣結合法）の約半分と結論される。

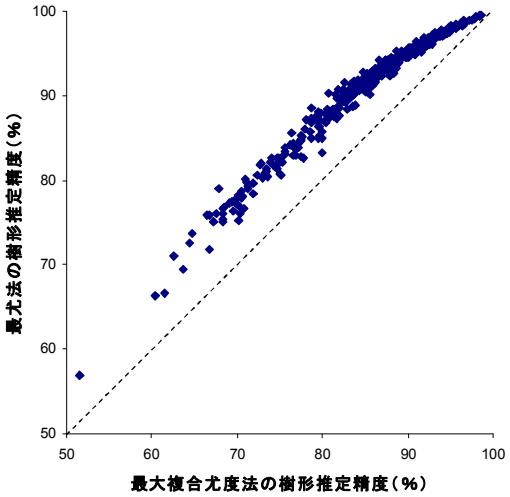

図 4. 最大複合尤度法と最尤法によって正しく推定された枝の割合（%）
各点は合計 448 遺伝子の各遺伝子について 100 回の繰り返しの平均値を表す。

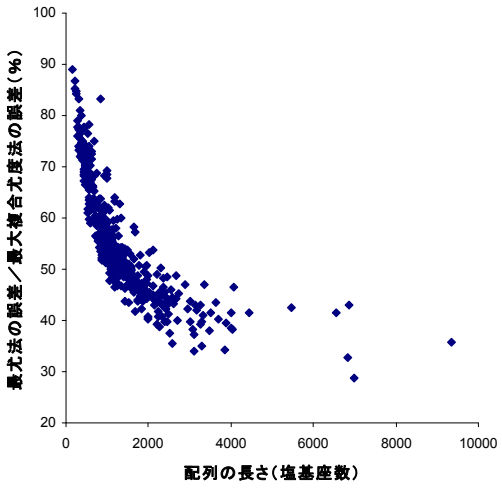

図 5. 最大複合尤度法の誤差に対する最尤法の誤差（%）と配列の長さの関係長の推定時間

(4) 異なる配列数における最大複合尤度法と最尤法の推定精度と計算時間

本研究で行ったコンピュータ・シミュレーションでは、哺乳類 66 種の系統樹をモデルとしたため、配列の数は 66 に固定されていた。そこで、系統樹推定精度と計算時間に及ぼす配列数の影響を調べるため、完全対称樹形をモデル系統樹に用いて配列数が異なる場合のコンピュータ・シミュレーションも行った。配列の長さは 1,000bp とした。

図 6 は、本研究で得られた配列数と樹形推定精度の関係を示すものである。配列数が 64 から 4096 の間では、最大複合尤度法、最尤法のいずれにおいても系統樹推定精度はほぼ一定で、正しく推定された枝の割合は最大複合尤度法では平均約 77%、最尤法では 86%であった。正しく推定されなかった枝の割合（推定誤差）に注目すると、最尤法の推定誤差は最大複合尤度法の約 60%で、哺乳類 66 種の系統樹をモデルとした場合（図 5）の結果と一致する。

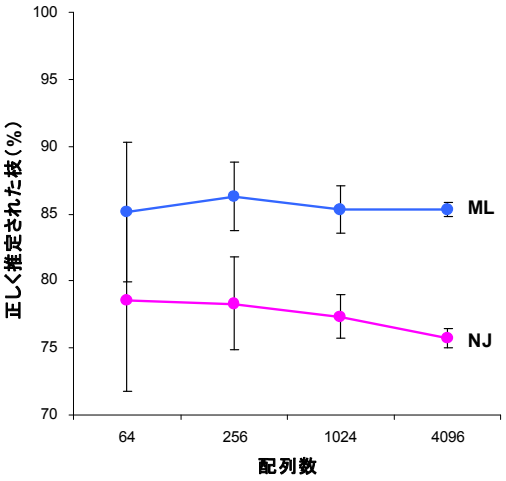

図 6. 配列数と樹形推定精度の関係
ML：最尤法、NJ：最大複合尤度法

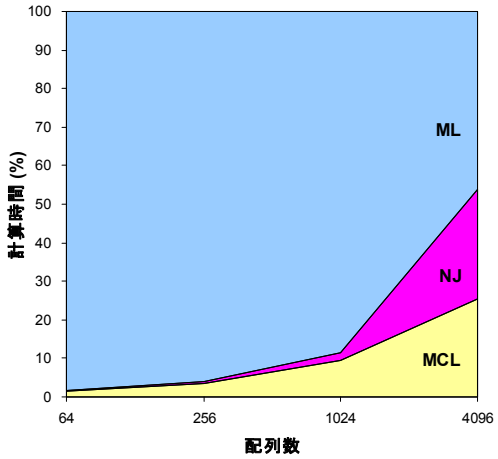

図 7. 配列数と相対計算時間の関係
最大複合尤度法によって初期系統樹を作成し、最近隣枝の交換（NNI アルゴリズム）によって最尤系統樹を探索した。ML：最尤法による最適化の時間、NJ：近隣結合法による初期系統樹作成の時間、MCL：最大複合尤度法による距離行列推定の時間

最大複合尤度法は、距離行列の推定に配列数の二乗、近隣結合法による系統樹作成に配列数の三乗に比例する計算時間を要する。一方、最尤法の計算時間は配列数と配列の長さの積に比例する。すなわち、配列の長さが一定の場合、配列数が増えるにしたがって最尤法の計算時間は、最大複合尤度法に比べて相対的に短くなることが期待される。

本研究では、1000bp の配列データを用い、最大複合尤度法によって初期系統樹を作成

し、それを元に最近隣枝の交換（NNI）アルゴリズムによる最尤系統樹探索を行う場合の計算時間を測定した。結果は図 7 に示す。配列数が 64 の場合、初期系統樹作成時間は全体の 1.5%に過ぎず、ほとんどの計算時間は最尤系統樹探索に費やされることが確認されたが、配列数が 4096 の場合、最大複合尤度法による初期系統樹作成時間とその後の最尤系統樹探索時間はほぼ等しくなり、距離行列法による系統樹推定における計算時間のメリットはほとんどなくなることが明らかとなった。配列数が数千以上に及ぶ巨大系統の推定のためには、距離行列法の改良、最尤法の利用が必要であると結論される。

(5) 最大複合尤度法による最尤法の改良

本研究では、最大複合尤度法はトランジション／トランスバージョン比や枝長の推定においては最尤法と同等の精度があり、一般的な条件では、最尤法に比べて計算時間が短いが、樹形推定精度においては最尤法には及ばないことが明らかとなった。そこで、最大複合尤度法を最尤法における初期条件の設定に利用し、最尤法の精度、計算時間の改良を試みた。最大複合尤度法を取り入れた最尤法を実装した系統樹推定プログラムを新たに開発し、その推定精度、計算時間を、いずれにおいても現在最良と言われる PhyML と比較した（表 1）。その結果、推定精度は PhyML とほぼ同等であるが、計算速度は約半分に短縮することに成功した。

表 1. 最大複合尤度法によって初期値を求めた場合の最尤法の推定精度と計算時間

	推定誤差(%)	*ΔLog L*	計算時間
PhyML	8.93 ± 6.64	―	100
最尤法	9.01 ± 6.54	0.12 ± 0.75	53.3 ± 8.3
最大複合尤度法	14.76 ± 7.75	-16.96 ± 4.46	1.5 ± 0.5

推定誤差：正しく推定されなかった枝の割合(%)
ΔLogL：PhyML による最尤系統樹との対数尤度の差
計算時間：PhyML の計算時間を 100 とした場合の計算時間

(6) 最大複合尤度法の普及

本研究で開発した最大複合尤度法の計算プログラムは、MEGA4 ソフトウェアに実装、http://www.megasoftware.net/ にて公開し、世界中の研究者に利用できるようにした。また、http://evolgen.biol.metro-u.ac.jp/MEGA/ には、MEGA4 ソフトウェアの使用方法を日本語で解説する Web ページを開設し、本研究で得られた研究成果の一般利用と社会貢献に努めた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform, 9:299–306. 2008. 査読有

② Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol. Biol. Evol. 24:1596–1599. 2007. 査読有

〔学会発表〕（計 4 件）

① 田村浩一郎. MEGA4 による分子系統解析. 日本進化学会第 10 回大会. 2008 年 8 月 24 日. 東京

② 田村浩一郎. 巨大系統樹推定のための最尤法の適用. 日本進化学会第 10 回大会. 2008 年 8 月 22 日. 東京

③ Koichiro Tamura. Application of Maximum Likelihood Methods for Large Phylogeny. The international symposium “New Insight of Genome Evolution into Fundamental Activities of Life”2008 年 3 月 28 日. 東京

④ Tamura K. Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Annual Meeting for the Society for Molecular Biology and Evolution. 2006 年 6 月 1 日. Arizona

〔その他〕
MEGA Software に関する Web サイト
1. 研究代表者によるサイト
http://evolgen.biol.metro-u.ac.jp/MEGA/
2. 共同研究者によるサイト
http://www.megasoftware.net/

６．研究組織
(1)研究代表者

田村　浩一郎（TAMURA KOICHIRO）
首都大学東京・大学院理工学研究科・教授
研究者番号： 00254144

(2)研究分担者
なし

(3)連携研究者
なし