統計数理（2002）
第 50 巻 第 1 号 45–68

特集「DNA 配列の統計解析」
［研究詳解］

分子系統樹法の応用と現状の問題点 ——真核生物の初期進化の解析を例として——*

橋本 哲男[1,2] ・有末 伸子[2] ・長谷川 政美[1,2]

（受付 2001 年 12 月 26 日）

要　　旨

DNA や RNA の塩基配列や蛋白質のアミノ酸配列のデータに基づき，生物の進化系統樹に対する推論を最尤法の枠組みで行うための方法論の概略を述べ，真核生物の初期進化の問題に対するデータ解析の実例を示した．その中で，分子系統樹の推論を誤らせる最も大きな要因として最近注目を集めている Long Branch Attraction アーテファクトについて実例に則して解説した．さらに，それを克服するための手法として，座位間の進化速度の不均質性を Γ 分布の導入により考慮した解析を実例に対して試み，この方法の有効性を示した．その結果，以前のいくつかの分子種の解析で真核生物の根もと近くから分岐するとされていた微胞子虫の位置づけは，解析に用いた分子種において微胞子虫の進化速度が極端に大きいことに伴う Long Branch Attraction アーテファクトであったとの可能性の高いことが明らかとなった．さらに，現在利用しうる全ての分子種のデータに基づいて総合評価の解析を試みると，微胞子虫が真菌に近縁であることが明確に示された．

キーワード：分子系統樹の最尤推定，Long Branch Attraction，座位間の進化速度の不均質性，Γ 分布，真核生物の初期進化，微胞子虫．

1. はじめに

ヒトをはじめとする多くの生物種で，全遺伝情報，すなわちゲノム DNA の全塩基配列を解読するゲノムプロジェクトの作業が急速なピッチで進められており，既に 50 以上の生物種について解読が終了している．これらのデータは，それ自体が個別の生物の基礎研究のためのデータとして重要であるが，そればかりでなく，種間での比較を通して得られる情報は格段に有用なものとなる．たとえば，種間で対応関係にある DNA 塩基配列の違いを解析することにより，形態学的・生理学的特徴の相違が配列上のどのような違いに由来するものであるかをある程度推測することができる．また非常に遠縁の生物種同士でも対応する特定の DNA 配列領域が類似していれば，その部分は生命機能の維持にとって非常に重要なものである可能性がでてくる．

一方，地球上の全ての生物は共通の祖先から進化してきたものであるため，さまざまな生物

[1] 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7
[2] 総合研究大学院大学 先導科学研究科生命体科学専攻：〒240–0193 神奈川県三浦郡湘南国際村

* 本稿は，統計数理研究所 共同研究（13–共研–1021）の研究成果の一部をまとめたものである．また本研究を遂行するにあたり，日本学術振興会科学研究費補助金（10044219, 12554037, 13640709）の資金援助を受けた．

種のゲノム配列データを多くの種間で比較解析することにより，生物の進化の歴史を辿ることができる．配列データに基づいて生物の進化系統樹に関する推論を行う研究分野は「分子系統学」という分野であるが，近年のデータの増大に伴ないその重要性が注目されてきている．分子系統学において系統樹推定の手掛かりを与えるのは，DNA や RNA における塩基置換や蛋白質におけるアミノ酸置換である．共通の祖先から分かれた後のそれぞれの系統における進化の過程で独立に置換が起こるので，生物種によって配列に違いが見られる．こうした違いを異なる生物や遺伝子間で比較することによって，系統樹が推定されるのである．進化の過程は，ランダムな確率過程としてとらえることが妥当である（Kimura（1983））．そのような過程の産物として得られている配列データから系統樹を推定するためには，確率モデルに基づいた統計的な方法が必要である．

本稿では，分子系統樹の最尤推定の方法論の概略を述べるとともに，「真核生物の初期進化の解明」という生物学上の重要な問題への応用例を示す．その中でとくに，分子系統樹の推定方法論において最近とくに深刻な問題点として指摘されつつある Long Branch Attraction（LBA）アーテファクトについて，現実の問題に即して紹介したい．なお，ここで紹介するのと同様に，分子系統樹法の現状の問題点が本特集の曹・長谷川（2002）論文でも別の例を通して取り扱われている．

2. 配列データの取得から分子系統樹推定へ

以下，分子系統樹推定に至るまでのデータ解析のステップについて簡単に述べる．ここに示した各方法のより詳細な内容とそれらの理論についての引用文献は，高木・金久（1996），Hilis et al.（1996），宮田（1998），長谷川・岸野（1996）などに詳述されている．

2.1 配列データの取得とアライメント

データベースから解析の対象とする配列データを取得するために用いられる一般的な方法は，相同性検索（ホモロジーサーチ）である．これは，研究者の手もとにある「問い合わせ配列（query sequence）」とよばれる配列と相同な配列がデータベースに存在するかどうかを検索する方法で，配列比較解析において最も基本的かつ重要な方法である．近年，広く普及している相同性検索のアルゴリズムには FASTA と BLAST があり，いずれもインターネットを通して誰にでも容易に利用できる．たとえば，National Center for Biotechnology Information（NCBI）（http://www.nlm.nih.gov/）や京都大学化学研究所・東京大学医科学研究所による GenomeNet（http://www.genome.ad.jp/）にアクセスすればよい．

相同性検索によって相同な配列が特定できた場合，次に行う作業は複数配列アライメント，すなわち，ギャップを導入することにより進化的に相同な座位の位置合わせを行うことである．アライメントによってはじめて，機能的に重要な部位や特定の高次構造を有する部位に対する生物種間もしくは相同分子種間の比較が可能となる．複数配列の自動アライメントを行うプログラムにはさまざまなものがあるが，よく使われているものはインターネット上でも利用可能である．たとえば，Clustal W は国立遺伝学研究所の SAKURA（http://sakura.ddbj.nig.ac.jp/）や前述の GenomeNet などで利用できる．アライメントが終わると，通常，アライメントに曖昧さを伴わないような座位が選択され，配列データ行列として特定される．たとえば，k 種の生物から n 座位が選ばれた場合，データ行列は，$\boldsymbol{X} = (X_{ij}), (i = 1, \ldots, k; j = 1, \ldots, n)$ である．これをもとに生物ペア間の同一もしくは異なる座位の割合，すなわち類似度行列もしくは距離行列が計算され，生物種間の進化的位置関係を論じる際の基礎データとなる．さらに，より詳しい解析のための方法が分子系統樹法である．

2.2 分子系統樹の最尤推定

分子系統樹を推定するための方法には，大きく分けて 3 つの方法，すなわち，距離行列法，最節約法，および最尤法がある．距離行列法は，距離行列をもとに近縁な配列（もしくは配列群）同士のクラスターを段階的に形成していく方法であり，計算時間がほとんどかからないという望ましい特徴をもつ．一般に距離行列としては，異なる座位の割合からなる行列を進化過程のモデルに基づいて補正したものが用いられる．最節約法と最尤法は，特定数の配列データに対する可能な系統樹のトポロジーを網羅的に探索する方法である．最節約法では，系統樹上の進化的変化数（置換数）の合計を最小にする，すなわち最大に節約する，という原理のもとに最も少ない置換数で説明できるようなトポロジーが真の系統樹の候補として選ばれる．単純で分かりやすい方法ではあるが，系統間で進化（置換）速度が異なる場合には誤りを犯す可能性が高い．最尤法は，ランダムネスを伴う確率過程である進化の過程から生成された配列データを解析する際，統計的に最も標準的な方法である．現実の進化過程に対し，さまざまなモデルを仮定して解析できるため，例えば進化速度の系統間での一定性を仮定しないモデルを用いれば，たとえ現実に進化速度が系統間で異なるような場合にも偏りのない推定ができる．最尤法による実際的な分子系統樹推定の方法論を初めて提案したのは Felsenstein（1981）であり，現実の配列データの解析を通して，方法論の改良・開発を進めてきたのが Hasegawa, Kishino らのグループである（長谷川・岸野（1996））．以下最尤法の概略を簡単に説明する．

ここでは簡単のために 4 種のみからなる根もとのない系統樹を考える．いま，図 1A における各枝の長さを未知パラメータとし，

$$\theta^{(1)} = (t_1, t_2, t_3, t_4, t_5)^T$$

とおき，分子進化に関する他のパラメータを $\theta^{(2)}$ として，$\theta = (\theta^{(1)}, \theta^{(2)})$ とおく．

各分岐点から次の分岐点（または枝の先端）への進化は独立に起こるものと仮定し，時間 t の間に塩基もしくはアミノ酸の状態が i から j に置換する確率（遷移確率）を $P_{ij}(t)$ とする可

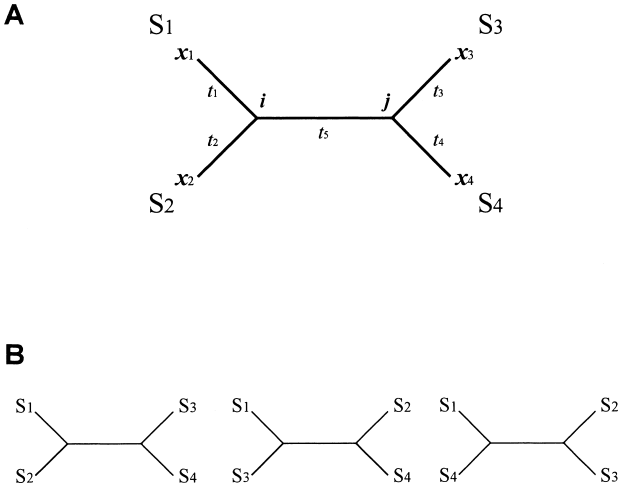

図 1. （A）4 つの系統に対する根のない系統樹． S_1〜 S_4: 現存生物種，x_1〜 x_4: 各座位における現存生物種の塩基またはアミノ酸の観測値，i，j: 祖先生物種の塩基またはアミノ酸の状態，t_1〜 t_5: 枝の長さ．（B）4 つの系統に対する 3 通りの系統樹のトポロジー．

逆な定常マルコフ過程を考えると，ある座位 h で塩基もしくはアミノ酸の観察値が，

$$X_h = (x_{1h}, x_{2h}, x_{3h}, x_{4h})^T$$

となる確率，すなわちある座位 h における尤度は，Chapman-Kolmogorov の等式より，

$$\begin{aligned} f(x_{1h}, x_{2h}, x_{3h}, x_{4h}|\theta) &= \sum_i \sum_j P\{X_0 = i\} P\{X_{t_1} = x_1, X_{t_2} = x_2, X_{t_5} = j | X_0 = i\} \\ &\quad \times P\{X_{t_5+t_3} = x_3, X_{t_5+t_4} = x_4 | X_{t_1} = x_1, X_{t_2} = x_2, X_{t_5} = j, X_0 = i\} \\ &= \sum_i \pi_i P_{ix_1}(t_1) P_{ix_2}(t_2) \sum_j P_{ij}(t_5) P_{jx_3}(t_3) P_{jx_4}(t_4) \\ &\qquad\qquad (i, j = 1, \ldots, 4\ (\text{塩基}); 1, \ldots, 20\ (\text{アミノ酸})) \end{aligned}$$

と表わせる．ここで，π_i （$i = 1, \ldots, 4$（塩基）; $1 \ldots, 20$（アミノ酸））は種 1 と種 2 の共通祖先において塩基もしくはアミノ酸 i を見出す確率であり，定常的なモデルを考えているので各塩基もしくはアミノ酸の組成値とする．また，$\sum$ は，祖先の塩基またはアミノ酸の状態が不明なため，可能な状態に関してたし合わせることを示している．

さらに，各座位 X_h が互いに独立に同一の確率法則（独立同分布）にしたがって進化するものと仮定すると，n 個の座位に対する全尤度 L は各尤度の積，すなわち，

$$L = \prod_{h=1}^{n} f(X_h|\theta)$$

となり，この対数尤度，

$$\log L = l(\theta|\boldsymbol{X}) = \sum_{h=1}^{n} \log f(X_h|\theta)$$

を最大にするように θ の推定値 $\hat{\theta}$ を求める．

このような推定を図 1B の 3 つの可能なトポロジー全てに関して行い，さらに，θ の各推定値を対数尤度の式に代入して得られる最大対数尤度の値を 3 つのトポロジー間で比較して，その値の一番高いものを真の系統樹の最も良い候補として選択する．これは，統計的モデル選択の問題であり，それぞれのトポロジーが系統学的な仮説に相当する．

最尤法は進化速度が系統間で異なる場合にも，平均的にみて上述の 3 つの方法の中では最も良い推定結果を与えることが知られている（Hasegawa et al.（1991）, Hasegawa and Fujiwara（1993）が，計算時間がかかるというのが欠点である．5 種に対する可能な系統樹のトポロジー数は 15，6 種では 105，7 種では 945 であるが，さらに種数が増えると可能なトポロジーの数は爆発的に増大し，網羅的に最尤系統樹を探索することは事実上不可能となる．このため，実際的な探索法がいくつか考案されている（Adachi and Hasegawa（1996））．

遷移確率 $P_{ij}(t)$ に対するモデルとしては，塩基置換，アミノ酸置換ともにさまざまなものが考案されているが，今回の我々の解析では，塩基置換については HKY85 モデル（Hasegawa et al.（1985），アミノ酸置換については JTT-F モデル（Adachi and Hasegawa（1996）を用いている．HKY85 モデルは，DNA や RNA の塩基置換におけるトランジション型の置換（T⇔C, A⇔G）とトランスバージョン型の置換（T, C⇔A, G）の起こり易さの違いと，4 種類の塩基の頻度が偏っていることを考慮に入れたモデルである．進化的な時間スケールでの微小時間 dt の間に塩基 i が塩基 j に置換する確率 $P_{ij}(dt)$ は，

$$P_{ij}(dt) = \begin{cases} \alpha\pi_j dt & (\mathrm{T} \Leftrightarrow \mathrm{C}, \mathrm{A} \Leftrightarrow \mathrm{G}) \\ \beta\pi_j dt & (\mathrm{T}, \mathrm{C} \Leftrightarrow \mathrm{A}, \mathrm{G}) \end{cases}$$

と表わされる．ここで π_j は塩基 j の組成値であり，α と β はそれぞれトランジションとトランスバージョンの速度に関係したパラメータである．Dayhoff et al.（1978）は互いに近い関係にある生物種のデータからなる蛋白質の多くのグループについて，最節約法によりグループ内に生じたアミノ酸置換を数え上げ，これをもとにアミノ酸の推移確率行列を推定した．Kishino et al.（1990）はこの行列に基づき蛋白質分子系統樹の最尤法による解析を定式化した．その後，Jones et al.（1992）は，多くのアミノ酸配列データの蓄積をもとにこの行列の改訂版を報告した．JTT-F モデルは，Jones et al.（1992）による推移確率行列を解析データセットのアミノ酸組成値で補正して用いるものである．

2.3 分子系統樹の信頼性の評価

一般に，最尤系統樹 m が得られたときには，それが対立仮説としての系統樹 a よりも，真の系統樹の候補としてどの程度良いものであるかを評価する必要がある．そのため，対数尤度の差 $(l_m(\widehat{\theta_m}|\boldsymbol{X}) - l_a(\widehat{\theta_a}|\boldsymbol{X}))$ が漸近的に正規分布に従うことから，その分散の近似的な推定式，

$$\widehat{\mathrm{Var}}[l_m(\widehat{\theta_m}|\boldsymbol{X}) - l_a(\widehat{\theta_a}|\boldsymbol{X})] = \frac{n}{n-1}\sum_{h=1}^{n}\left\{\log\frac{f_m(X_h|\widehat{\theta_m})}{f_a(X_h|\widehat{\theta_a})} - \frac{1}{n}\sum_{h'=1}^{n}\log\frac{f_m(X_{h'}|\widehat{\theta_m})}{f_a(X_{h'}|\widehat{\theta_a})}\right\}^2$$

が求められている（Kishino and Hasegawa（1989））．実際には，この分散の平方根を標準誤差（SE）として，対数尤度の差とともに示す場合が多い．また，2SE や 3SE を基準として差の有意性が論じられる場合もあり，分子系統学の分野では，Kishino-Hasegawa 検定として良く知られている．

ブートストラップ法も系統樹の確からしさを示す指標として非常に頻繁に用いられている．解析に用いる配列データ行列 $\boldsymbol{X}$ の n 個の座位の中から n 個の標本をリサンプリングして仮想的なデータセット（ブートストラップ標本）を多数個（たとえば 10000 個）つくる．すなわち，

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} = (X_1, X_2, \ldots, X_n)$$

をもとに，

$$\boldsymbol{X}^* = \begin{bmatrix} x_{1B_1} & x_{1B_2} & \cdots & x_{1B_n} \\ x_{2B_1} & x_{2B_2} & \cdots & x_{2B_n} \\ \cdot & \cdot & \cdots & \cdot \\ x_{kB_1} & x_{kB_2} & \cdots & x_{kB_n} \end{bmatrix} = (X_{B_1}, X_{B_2}, \ldots, X_{B_n})$$

を作り，これらについて系統樹を推定する作業を繰り返すことにより，特定の系統樹が最尤系統樹として選択される頻度を集計する．この頻度のことをその系統樹のブートストラップ確率という（Felsenstein（1985））．この値が高いほどその系統樹は信頼できそうだということになる．実際には，オリジナルデータに基づく最尤系統樹の各内部枝に対し，その枝を共通祖先とする部分系統樹に含まれる生物が単系統群となるような系統樹それぞれに与えられた頻度の合計を求め，ブートストラップ確率として表示することが多い（図 4 参照）．

ところが，実際上の問題として，ブートストラップ標本についてその都度最尤法の計算を行うのは非常に大変であり，トポロジーや座位の数が多い場合は事実上不可能である．そこで，データからブートストラップ標本をリサンプリングする代わりに，座位の対数尤度をリサンプリングすることにより，近似的なブートストラップ確率を推定することができる．すなわち，m 個のトポロジーの対数尤度関数が，

$$l_i(\theta_i|\boldsymbol{X}) = \sum_{h=1}^{n} \log f_i(X_h|\theta_i)\,, \quad i = 1, \ldots, m$$

と表わされるとき，各トポロジーに対し，$l_i(\hat{\theta}_i|\boldsymbol{X})$ を求めるとともに，各座位の対数尤度 $f_i(X_h|\hat{\theta}_i)$, $(h = 1, \ldots, n)$ を保存しておき，これらをリサンプリングして，

$$l_i{}^*(\hat{\theta}_i|\boldsymbol{X}^*) = \sum_{h=1}^{n} \log f_i(X_{B_h}|\hat{\theta}_i)\,, \quad i = 1, \ldots, m$$

を計算することにより，近似的なブートストラップ確率を求めることができる（Kishino et al.（1990））．この方法は，RELL（Resampling of Estimated Log-Likelihood of sites）法と呼ばれており，実際に n が大きい場合には，この近似の精度は十分に良いことが明らかとなっている（Hasegawa and Kishino（1994））．

系統樹間の比較のための上述の方法は，一般に自由パラメータ数の等しいモデル間での比較のために用いられる．もし，モデル選択の際，自由パラメータ数の異なるモデル間を比較する必要がある場合には，情報量規準 AIC（$= -2$（モデルの最大対数尤度）$+\ 2$（モデルの自由パラメータ数））に基づいて評価し，AIC が最小となるモデルを選択する（Akaike（1974））．

以上，2.2 節と 2.3 節で述べた分子系統樹の最尤推定とモデル選択の解析を行うための基本的なプログラムは，‘MOLPHY’ というパッケージに全て整備されており（Adachi and Hasegawa（1996）），国内外で広く利用されている．

一方，Shimodaira は，リサンプリングを用いて対数尤度の多重比較を行い，モデル選択の信頼性を確率値として評価する方法を提案した（Shimodaira（1998））．これは，Kishino-Hasegawa 検定で見落とされていた選択バイアスを多重比較法で補正したもので，Shimodaiora-Hasegawa 検定として分子系統学の分野で利用され始めている（Shimodaira and Hasegawa（1999））．さらに最近，Shimodaira は，データが仮説を支持するかしないかという情報とブートストラップ法によるリサンプリングだけを使って，クラスタリングによるバラツキを確率値（p 値）として評価する Approximately Unbiased（AU）検定を提案し，分子系統樹のモデル選択への適用を可能にした（下平（2002）（本特集），Shimodaira（2002））．これらの解析のためのプログラムは，‘CONSEL’ という名称で最近公開され（Shimodaira and Hasegawa（2001）），今後，分子系統学の分野で一般的に利用されるようになるものと期待される．

3. 真核生物の初期進化研究の背景

3.1 ミトコンドリアをもたない原生生物は真核生物の祖先型生物か？

地球上の生物は，細胞内に核をもつ真核生物と核をもたない原核生物とに大きく二分される．原核生物のほとんどは，細菌といわれているものでありいずれも単細胞の単純な生物である．一方，真核生物には，動物，真菌（細菌ではなく，カビ，キノコ，酵母などの類），植物などのいわゆる高等な分類群が含まれ，これらのほとんどは多細胞生物である．さらに真核生物の中には，単細胞生物からなるさまざまな分類群が存在しており，それらは原生生物と総称されている（表 1）．

表 1. 真核生物を構成する主な分類群.

分類群	代表的な生物（群）
動物	ヒト、ショウジョウバエ、線虫
真菌	キノコ、カビ、酵母
粘菌	細胞性粘菌、モジホコリカビ
緑色植物	シロイヌナズナ、イネ、クロレラ
紅藻	ノリ、テングサ
アルベオラータ	ゾウリムシ、テトラヒメナ、マラリア原虫
ストラメノパイル	珪藻類、褐藻類、卵菌類
ユーグレノゾア	ユーグレナ、トリパノソーマ
エントアメーバ*	赤痢アメーバ、口腔アメーバ
ペロバイオンタ*	マスティグアメーバ、ペロミクサ
微胞子虫*	グルゲア、エンセファリトズーン
ディプロモナス*	ランブル鞭毛虫、ヘキサミタ
トリコモナス*	膣トリコモナス、口腔トリコモナス

*ミトコンドリアをもたない分類群

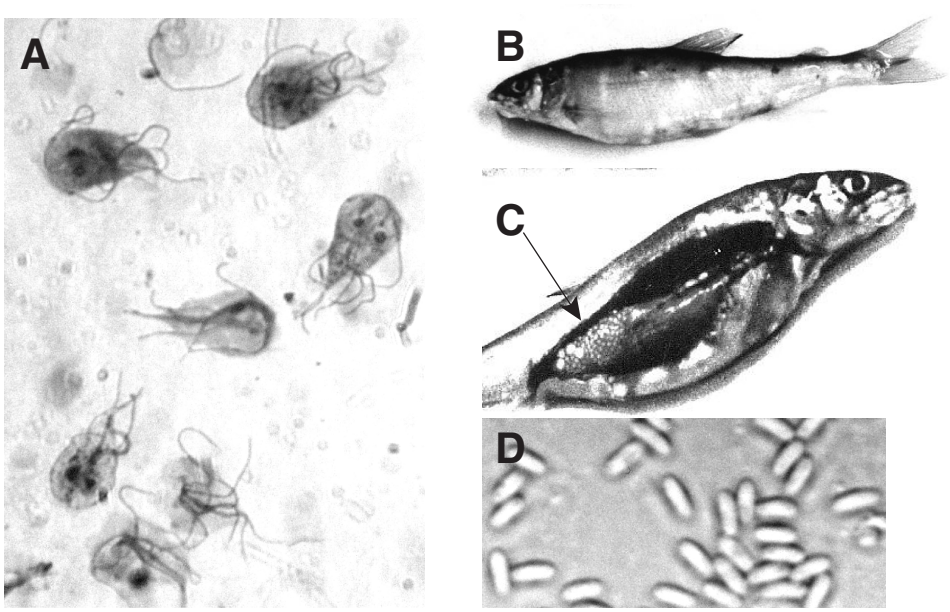

図 2. ミトコンドリアをもたない原生生物.（A）ランブル鞭毛虫（*Giardia intestinalis*, ディプロモナス類）のギムザ染色像. 左右対称で 2 つの等価な核と 8 本の鞭毛をもつ. 大きさは, 長径 9 〜 20 μm, 短径 6 〜 10 μm.（B）グルゲア（*Glugea plecoglossi*, 微胞子虫）が感染したアユ.（C）腹腔内に形成されたグルゲアのシスト（矢印）.（D）グルゲア胞子の光顕像. 大きさは, 長径約 5 μm, 短径約 2 μm.

通常の真核生物の細胞には，酸素呼吸によるエネルギー生成器官であるミトコンドリアが存在している．ミトコンドリアは核とは別の独自の DNA をもっており，自己増殖することができる．また，ミトコンドリア DNA 上の遺伝子の配列は，核 DNA の対応遺伝子の配列よりも，原核生物における対応遺伝子の配列に類似していることが明らかとなっている．これらのことから，ミトコンドリアは，真核生物の祖先型生物に原核生物が細胞内共生することによって生じた器官であるとする説が広く受け入れられている．ところが，原生生物の中には，真核生物であるのにもかかわらずミトコンドリアをもたないものが存在する．図 2 に示したランブル鞭毛虫（ディプロモナス）やグルゲア（微胞子虫）はその一例であり，この他にも，表 1 に示すように，トリコモナス，赤痢アメーバなどの生物にはミトコンドリアはない．このような生物の系統進化学的位置を明らかにすることは，真核生物の初期進化の歴史を解明するうえで重要である．これらミトコンドリアをもたない原生生物の分類群のうちのある種のものが，真核生物全体の系統樹上で根もと近くから分岐しているとすれば，それらの中にミトコンドリアの細胞内共生が起こる以前の真核生物の祖先型に近い生物が存在する可能性がでてくる．一方，ミトコンドリアをもたない分類群に属する生物種の多くは寄生虫であるため，これらが寄生生活に適応してミトコンドリアを二次的に喪失したとする可能性も否定できない．実際，ランブル鞭毛虫や赤痢アメーバは腸管寄生虫，トリコモナスは膣や口腔内の寄生虫，グルゲアは魚の寄生虫である．

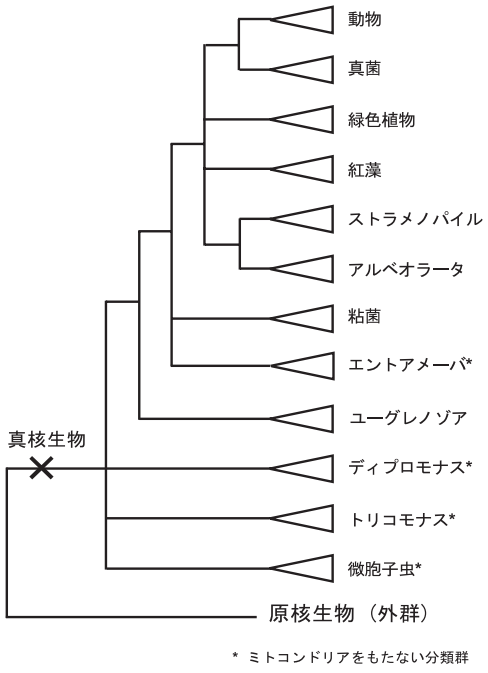

図 3. 小亜粒子リボソーム RNA（SSUrRNA）に基づく真核生物全体の系統樹．Leipe et al.(1993）に基づき，主な分類群相互の関係を模式的に示した． 分岐の順番があまり明確でない部分は多分岐の関係を用いて示している．

表 1 に示した真核生物の大きな分類群相互の系統進化学的関係については，1995 年ぐらいまでは，主として小亜粒子リボソーム RNA（SSUrRNA）の塩基配列比較に基づいて解析されてきた．リボソーム RNA は，細胞内の蛋白質合成装置であるリボソームの構成成分であり，全ての生物に存在することから，真核生物全体を通して系統樹解析を行うのに適した分子であると考えられる．図 3 には，最近に至るまで広く受け入れられていた SSUrRNA の系統樹を模式的に示した（Leipe et al.(1993)）．この系統樹では，真核生物の進化の非常に早い時期に，ミトコンドリアをもたない 3 分類群，微胞子虫，トリコモナス，ディプロモナスが他の真核生物に至る系統から分岐したことが示されている．さらに筆者らは 1996 年に，蛋白質合成に関与する別の分子であるペプチド鎖伸長因子（EF1α 及び EF2）という蛋白質のアミノ酸配列に基

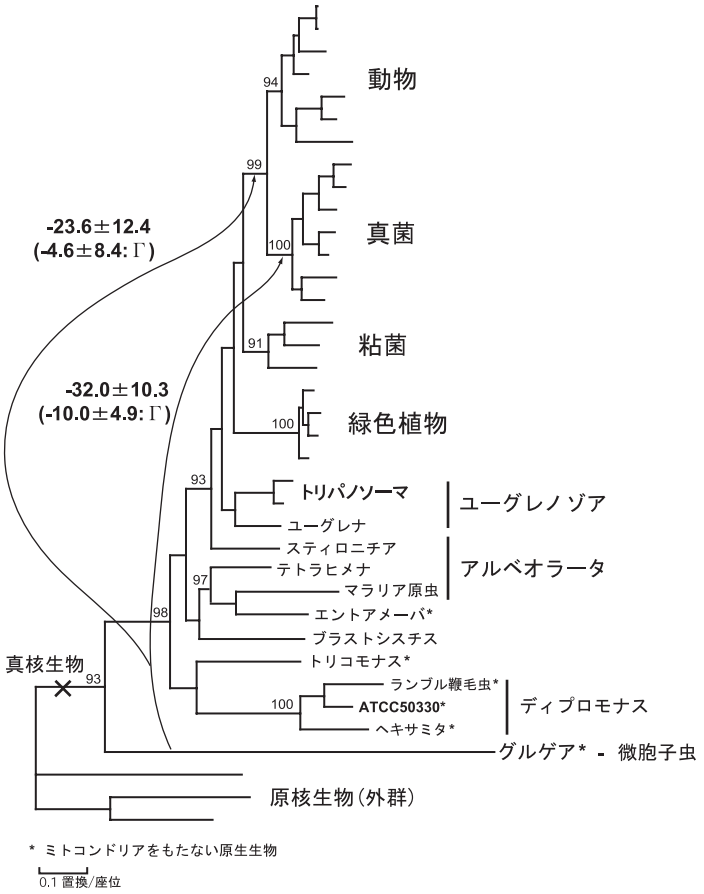

図 4. ペプチド鎖伸長因子 EF1α に基づく真核生物全体の系統樹．アミノ酸置換モデルとして JTT-F を用い，蛋白質分子系統樹の最尤法（Adachi and Hasegawa（1996））によって推定した系統樹．動物，真菌，粘菌，緑色植物，古細菌については属名，種名ともに省略してある．枝の長さは推定アミノ酸置換数に比例している．369 アミノ酸座位を選択して解析に用いた．各内部枝上の数値は，それぞれの枝に連なる 3 つの部分系統樹のなかの分岐が正しいと仮定したもとでのブートストラップ確率（局所ブートストラップ確率）（Adachi and Hasegawa（1996））で，その枝の信頼性の大まかな指標である．90%以上のものについてのみ示してあるが，一部は省略してある．矢印は，Γ 分布により座位間の進化速度の不均質性を考慮した解析での枝の付けかえ操作を示す（本文参照）．

づく解析からも，これら 3 分類群の分岐の早いことを示していた（Hashimoto and Hasegawa (1996); Kamaishi et al.(1996)）（図 4）．こうした結果は，これら 3 分類群の中にミトコンドリアの細胞内共生が起こる以前の祖先型真核生物が存在する可能性を強く示唆した．また，これらの分類群に属する生物がいずれも「原始的」な細胞形態を呈していたことも，これらの祖先型真核生物の候補としての位置づけに大きく荷担した（Cavalier-Smith（1987））．

3.2 パラダイムの転換：ミトコンドリアをもたない原生生物におけるミトコンドリアの 2 次的喪失

ところが，1990 年代後半に入ると，ミトコンドリアをもたない現存の生物群の中に祖先型真核生物が存在するという仮説に大きく反する事実が次々と明らかになった（Roger（1999））．まず，ミトコンドリアで働く熱ショック蛋白質であるシャペロニン（CPN60）の遺伝子が，赤痢アメーバ，トリコモナス，ランブル鞭毛虫において，また，同じくミトコンドリア型の熱ショック蛋白質 70（HSP70）の遺伝子もトリコモナスと微胞子虫において，それぞれの核 DNA の中に見いだされた．しかもこれらの一部については，実際に発現していることも確かめられた．さらに，トリコモナスと赤痢アメーバでは，これらの分子がミトコンドリアとは異なるオルガネラに局在することが示され，これらのオルガネラがミトコンドリアと同一の進化的起源をもつ可能性が示唆された．一方，それぞれの分子系統樹の解析は，ミトコンドリアをもたないいずれの生物種から得られたいずれの配列も明らかにミトコンドリア起源であることを支持した．これらのことから，シャペロニンや HSP70 の遺伝子は，ミトコンドリアをもたない原生生物を含む真核生物全体の共通祖先の段階で，ミトコンドリア DNA から核 DNA の方へ移行し，それ以後，ミトコンドリアを標的として機能を営んできたものと考えられた．さらにミトコンドリアをもたない原生生物においては，ミトコンドリアの喪失後も何らかの機能を保持しつつ残存しているものと考えられた．すなわち，現存のミトコンドリアをもたない原生生物はいずれもかつてはミトコンドリアをもっており，進化の過程でそれらを二次的に失ったのだということとなった．

一方 1990 年代後半には，祖先型真核生物の存在を示唆する根拠となったリボソーム RNA やペプチド鎖伸長因子の系統樹にも疑問が投げかけられるようになった．とくに，微胞子虫が真核生物の根もと近くに位置づけられる点が問題視された．細胞骨格に関連しているチューブリンの解析結果は，微胞子虫の早い分岐を支持しないばかりか，真菌に近縁である可能性を支持したのである（Edlind et al.(1996), Keeling and Doolittle（1996））．さらに，上述の HSP70 の解析からも，微胞子虫が真菌に近縁である可能性の方が強く，分岐が早いことは必ずしも支持されないという結論が導かれた（Germot et al.(1997)）．

ミトコンドリアの二次的喪失の証拠が次々と示される一方，統計的誤差の範囲内の違いも多いとは言え，分子系統樹の解析が用いる分子ごとに互いに矛盾した結論を導くという混乱した状況が続くなか，我々は，ミトコンドリアをもたないものを含む真核生物の主な分類群（表 1）相互の系統進化学的位置関係を再検討し，真核生物の初期進化を解明することを目的として研究を進めてきた．研究を開始した時点では，ミトコンドリアをもたない分類群の配列データはまだあまり多くなく，解析に用いることのできる分子はほんの数種しか存在していなかった．そこで，この目的を達成するためには，(1) ミトコンドリアをもたない分類群におけるさまざまな分子の配列データを蓄積し，(2) 解析方法論上の問題点を十分検討したうえで，さまざまな方法論を適用して詳細なデータ解析を行うとともに，(3) 個々の分子に基づく解析結果を総合評価して結論を導く必要があった．

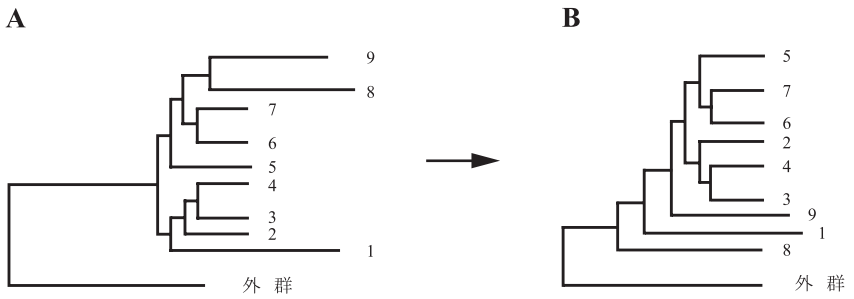

図 5. Long Branch Attraction（LBA）を示す模式図．Philippe and Laurent（1998）に基づき改変．1〜9 は現存生物種を表す．

4. Long Branch Attraction

4.1 系統樹の推論を誤らせる主要因としての Long Branch Attraction（LBA）

研究に先立ってまず検討すべきことは，いったいどうして SSUrRNA や EF1α，EF2 の系統樹と他の分子による系統樹が微胞子虫の位置づけについて大きな矛盾をもたらすのかを明らかにすることであった．その原因として最も考えられそうなことは，微胞子虫のこれらの分子における進化速度（塩基やアミノ酸の置換速度）の極端な増大が系統樹の推定を誤らせている可能性である．一般に，系統樹を推定する際には，関係を明らかにしたい生物群とは系統的にかけ離れていることが既知であるような生物群（外群）を含めて解析を行いその生物群（外群）の共通祖先のところに系統樹の根もとが存在すると仮定する．図 3 や図 4 の例では，真核生物内部の関係を解析するために，原核生物を外群として用いている．その際，真核生物の中に極端に進化速度の大きな系統があると，その系統は本来あるべき位置よりも外群の方へ引っ張られて位置づけられるという傾向をもつ．これは，Long Branch Attraction（LBA）と呼ばれており（Felsenstein（1978)），系統樹の推定を誤らせる大きな要因として近年注目を集めている．Philippe らは，シミュレーション研究により現実の問題として LBA の生ずる可能性の高いことを指摘した（Philippe and Laurent（1998)）．例えば，図 5A のように，外群が遠く内群に他に比べて進化速度が顕著に大きい系統 (1, 8, 9) があるというのが真実であったとしても，推定される系統樹は一般に図 5B のようになり，進化速度の大きい系統が内群の根もと近くから分岐するという傾向が強いのである．彼らは，この根もと付近の系統樹が非対称な部分は LBA によるアーテファクトであるとしている．実際，SSUrRNA の系統樹（Leipe et al.（1993)）や EF1α（図 4)，EF2 の系統樹（Kamaishi et al.（1996)）の真核生物の根もと付近はこのような傾向を示していることから，彼らは，これらの系統樹で微胞子虫が真核生物の根もと近くから分岐するのは，LBA によるアーテファクトであり，チューブリンや HSP70 の系統樹の方が正しく，微胞子虫は真菌に近縁なのであると主張した（Germot et al.（1997), Philippe and Laurent (1998)).

4.2 座位間での進化速度の不均質性を考慮に入れた解析

分岐後の時間が経過するにつれて置換が蓄積され，1 つの枝で 2 回以上の置換（多重置換）が頻繁に起こるようになることは，一般に良く知られている．配列内にほとんど置換が起こり得ない部分と容易に置換が生ずる部分が混在している場合，後者において多重置換が多く起きても，変化しうる部分が限られていることから，全体としての配列間の差異がそれに応じて大きくなるわけではない．すなわち，見かけ上の配列間の関係は，実際よりも近いものであると見

なしてしまうのである．一般に置換の速度は分子の種類によって異なるが，それだけでなく，ある特定の分子種の配列の内部でも座位によって大幅に異なっているというのが現状である．このような不均質性が極端な場合，それを無視して解析すると，多重置換を過小に評価し，進化（置換）速度の大きな系統を外群の方に離すような偏りをもってしまう，すなわち，LBA の効果を生じさせてしまうのである．

そこで，座位間での進化速度の不均質性を考慮することにより，より現実に近い統計モデルを用いて解析し，LBA の効果を抑制する必要が生じてくる．最尤法の枠組みで座位間の不均質性をガンマ分布でモデル化して解析する試みは，Yang（1993）によって定式化され，現実の問題に適用可能な状態になっている（Yang（1996））．以下，長谷川・岸野（1996）に従ってその概略を述べる．

いま，2.2 節の対数尤度の式 $l(\theta|\boldsymbol{X}) = \sum_{h=1}^{n} \log f(X_h|\theta)$ において，$\theta = (\theta^{(1)},\ \theta^{(2)})$ で，$\theta^{(1)}$ が枝の長さであるから，座位間の不均質性は，

$$l(\theta_h^{(1)}, h = 1, \ldots, n, \theta^{(2)}|\boldsymbol{X}) = \sum_{h=1}^{n} \log f(X_h|\theta_h^{(1)}, \theta^{(2)})$$

と，枝の長さを座位ごとに割り当てることによって表現される．さらに，座位間の進化速度の相対比が進化時間を通じて一定であると仮定し，

$$l(\theta^{(1)}, \lambda_h, h = 1, \ldots, n, \theta^{(2)}|\boldsymbol{X}) = \sum_{h=1}^{n} \log f(X_h|\lambda_h\theta^{(1)}, \theta^{(2)})$$

とする．しかし，このままでは推定すべきパラメータ数が座位数の増加とともに増大してしまうので，何らかの方法で，パラメータ数を節約する必要がある．蛋白質コード領域に対応した DNA の塩基座位のように，コドンの 1 番目，2 番目，3 番目とあらかじめ不均質性の構造がわかっている場合には，これらそれぞれのグループに λ_h を割り当てることにより節約できる（Adachi and Hasegawa（1996））．もしこのようなことが不可能な場合には，λ_h が座位ごとに確率的に変化すると仮定し，確率分布 $g(\lambda_h, h = 1, \ldots, n|\phi)$ を導入する．この分布を規定する超パラメータ ϕ は，経験ベイズ法によりデータに基づいて推定される．すなわち，周辺尤度，

$$\begin{aligned} L'(\theta^{(1)}, \phi, \theta^{(2)}|\boldsymbol{X}) &= \int \cdots \int L(\theta^{(1)}, \lambda_h, h = 1, \ldots, n, \theta^{(2)}|\boldsymbol{X}) \\ &\qquad \cdot\ g(\lambda_h, h = 1, \ldots, n|\phi) d\lambda_1 \cdots d\lambda_n \\ &= \int \cdots \int \left\{ \prod_{h=1}^{n} f(X_h|\lambda_h\theta^{(1)}, \theta^{(2)}) \right\} g(\lambda_h, h = 1, \ldots, n|\phi) d\lambda_1 \cdots d\lambda_n \end{aligned}$$

を最大化することにより推定される．最も基本的な具体的アプローチとしては，各座位に独立に確率分布を割り当てる，すなわち，

$$g(\lambda_h, h = 1, \ldots, n|\phi) = \prod_{h=1}^{n} g_0(\lambda_h|\phi)$$

とするもので，このとき上記の式は，

$$L'(\theta^{(1)}, \phi, \theta^{(2)}|\boldsymbol{X}) = \prod_{h=1}^{n} \left\{ \int f(X_h|\lambda_h\theta^{(1)}, \theta^{(2)}) g_0(\lambda_h|\phi) d\lambda_h \right\}$$

となる．Yang（1993, 1994）は g_0 としてガンマ分布モデルを適用し，さらにこれを離散化することにより数値計算上の負担を大幅に軽減して，一般の系統樹推定に実行可能な方式を提唱した．すなわち，ガンマ分布 $g_0(\cdot|\alpha, \beta)$ のもとでそれぞれの確率が $1/s$ になるように λ_h を s 個の

区間に分け，それぞれの区間内の平均を $\lambda^{(1)}(\alpha,\beta),\ldots,\lambda^{(s)}(\alpha,\beta)$ とする．このとき上式は，

$$L'(\theta^{(1)},\phi,\theta^{(2)}|\boldsymbol{X}) \sim \prod_{h=1}^{n}\left\{\frac{1}{s}\sum_{u_h=1}^{s} f(X_h|\lambda^{(u_h)}(\alpha,\beta)\theta^{(1)},\theta^{(2)})\right\}$$

と近似される．シミュレーションといくつかの解析例により，区間の個数 s は 4 個程度で十分であることが明らかとなっている（Yang（1994））．実際の分子系統樹の解析のためのプログラムとしては Yang（1997）によって‘PAML’というパッケージが作成されている（http://abacus.gene.ucl.ac.uk/software/paml.html）．

4.3 ペプチド鎖伸長因子の系統樹の再解析

図 4 のペプチド鎖伸長因子 EF1α の系統樹の推定に用いられた解析では，座位間の進化速度の不均質性は考慮されていないため，LBA の効果によって微胞子虫が真核生物の一番外側に誤って位置づけられているという可能性が考えられる．そこで，不均質性をガンマ分布により考慮した再解析を行ってみた．図 4 において，座位間の不均質性を考慮しない場合には，微胞子虫（グルゲア）を真菌の共通祖先のところに移動（矢印）させた系統樹の対数尤度は，もとの系統樹（微胞子虫は真核生物の根もとから分岐）の対数尤度よりも，32.0 ± 10.3（$\pm$ は 1SE）低くなる．一方，微胞子虫を動物と真菌の共通祖先のところに移動させた系統樹の対数尤度は，23.6 ± 12.4 低くなる．これらの値を見る限り，微胞子虫が真菌に近縁であるという可能性は否定しうる．ところが，ガンマ分布により不均質性を考慮した場合には，これら 2 つの枝の付け換えに対応する対数尤度差はそれぞれ，10.0 ± 4.9，4.6 ± 8.4 であり不均質性を考慮しない場合に比べて対数尤度差ははるかに減少している．しかも後者については標準誤差の方が大きくなっている．同様の解析を EF2 のデータセットでも試みると，ガンマ分布の考慮により，これらの付け換えに対応する対数尤度差はほぼ 0 に近くなってしまう．すなわち，より現実的なモデル化を行うことにより，微胞子虫が真核生物の根もとから分岐する可能性が低下することが明確に示されたわけである．さらに SSUrRNA の系統樹の再解析の結果も同様の傾向を示している．

5. 微胞子虫の系統的位置と真核生物の初期進化 —— 複数分子による解析

こうしたことから，Philippe らの指摘のとおり，微胞子虫の分岐の早いことを支持していた 3 つの分子の解析結果はいずれもアーテファクトであったとの可能性はもはや否定できないと考えられるが，これら 3 分子はチューブリンの系統樹のように，微胞子虫−真菌近縁説を積極的に支持するわけではない．また，以前この説を支持していた HSP70 の系統樹（Germot et al.（1997））も，その後のデータの増加に伴い必ずしも強い支持をもたらさないことが明らかになってきた．そこで，他の分子ではどの程度この説が支持されるのかを調べる目的で，我々はさらに，微胞子虫を含むさまざまな原生生物について，ヴァリン−及びイソロイシン−tRNA 合成酵素（VRS, IRS）の遺伝子の解析を行い，これらの分子系統樹を推定した（Hashimoto et al.（1998）; Weiss et al.（1999））．その結果，VRS では，微胞子虫−真菌近縁説がある程度支持されたものの，IRS ではやはり LBA の効果のためかそのシグナルは検出されなかった．一方，他の研究グループからも微胞子虫の大亜粒子リボソーム RNA(LSUrRNA) や RNA ポリメラーゼ II（RPOII）のデータが報告され，LSUrRNA では微胞子虫−真菌近縁説を支持もしないが真核生物の根もとから分岐する可能性も高くないという結果（Peyretaillade et al.（1998））が，RPOII ではこの説を強力に支持するという結果（Hirt et al.（1999））が報告された．このように，ここ数年微胞子虫のデータがさまざまな分子について蓄積されたため，これら複数の分子の情報を

結合した解析ができる段階になってきた．そこで以下，現時点で利用可能な全ての分子のデータに基づき，最近我々が行った総合的な解析の結果について紹介する．

5.1 蛋白質による解析とリボソーム RNA による解析の比較

以下の解析では表 1 に示した真核生物の主な分類群のうちの 7 つの系統，すなわち，微胞子虫，真菌，動物，粘菌，緑色植物，アルベオラータ，ユーグレノゾアの間の関係について最尤法により検討する．今回のこの解析では，外群を除いた根のない系統樹を対象とする．現時点において，これらの系統の全てにわたってデータの存在する分子は，EF1α，EF2，VRS，IRS，RPOII，アクチン（ACT），α-チューブリン（TBα），β-チューブリン（TBβ），SSUrRNA，LSUrRNA の 10 分子種であり，前の 8 分子種は蛋白質のアミノ酸配列，後の 2 分子種は rRNA

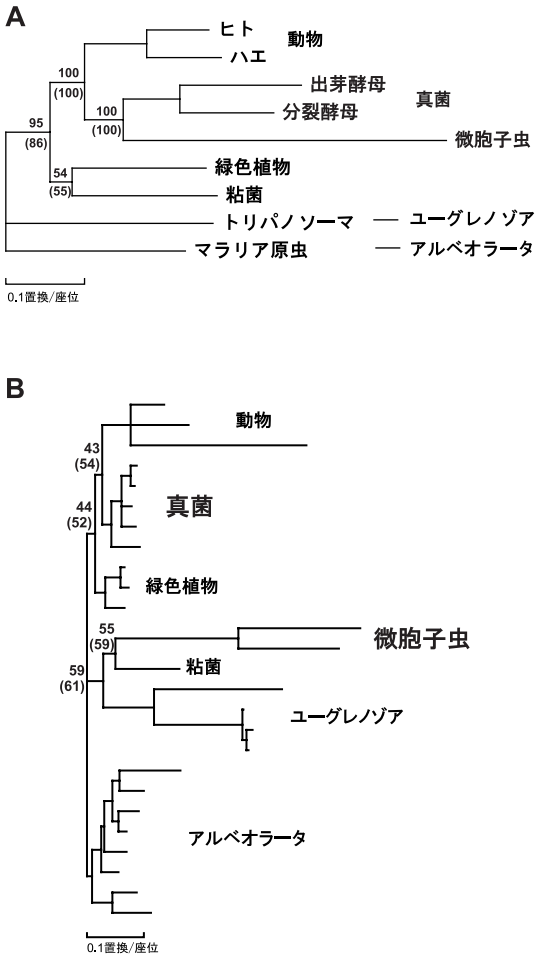

図 6. 結合データに基づく解析の最尤系統樹.（A）蛋白質 8 分子種 9 生物種による解析.（B）リボソーム RNA(rRNA)2 分子種 26 生物種による解析.（B）では動物，真菌，緑色植物，微胞子虫，ユーグレノゾア，アルベオラータ内部の生物種名を省略してある．内部枝上の数値は，結合データに基づく解析でのブートストラップ確率．内部枝下の括弧内の数値は，総合評価に基づく解析でのブートストラップ確率（本文参照）.

表 2. 蛋白質の最尤系統樹と rRNA の最尤系統樹の比較.

		without Γ[a]			
		連結データ		総合評価	
解析対象分子	系統樹[c]	Δl_i[d]	AIC	Δl_i[d]	AIC
蛋白質（8 種）	A	(−44783.5)	89635.0	(−43860.1)	88264.2
	B	−309.3±45.5	90253.6	−337.2±47.6	88938.6
rRNA（2 種）	A	−101.1±23.5	55087.5	−99.3±23.7	54866.8
	B	(−27391.7)	54885.4	(−27232.1)	54668.2

		with Γ[b]			
		連結データ		総合評価	
解析対象分子	系統樹[c]	Δl_i[d]	AIC	Δl_i[d]	AIC
蛋白質（8 種）	A	(−43312.2)	86694.4	(−42518.9)	85597.8
	B	−205.0±31.0	87104.4	−239.6±34.3	86077.0
rRNA（2 種）	A	−29.1±10.7	51640.2	−11.0±9.9	51345.2
	B	(−25739.0)	51582.0	(−25557.6)	51323.2

[a] 座位間での進化速度の不均質性を考慮しない解析.
[b] 座位間での進化速度の不均質性を Γ 分布により考慮に入れた解析. それぞれの分子種における系統樹 A（蛋白質）もしくは系統樹 B（rRNA）での Γ 分布の shape パラメータ α の推定値： EF1α, 0.63; EF2, 0.60; VRS, 0.75; IRS, 0.68; RPOII, 0.74; ACT, 0.58; TBα, 0.90; TBβ, 0.65; 蛋白質（8 種）の連結データ, 0.65; SSUrRNA, 0.42; LSUrRNA, 0.54; rRNA（2 種）の連結データ, 0.51. 実際の解析では, scale パラメータ β は, $\beta = \alpha$ となっており, 平均が 1 になるようにしている.
[c] A, 蛋白質（8 種）の最尤系統樹; B, rRNA（2 種）の最尤系統樹.
[d] 最尤系統樹からの対数尤度の差. ± は 1SE（Kishino and Hasegawa（1989））.（ ）内は最尤系統樹の最大対数尤度の値.

の塩基配列のデータである．蛋白質 8 分子種については動物 2 生物種，真菌 2 生物種，他の系統はそれぞれ 1 生物種ずつの合計 9 生物種，rRNA2 分子種については 7 系統にわたり 26 生物種のデータが共通に存在している．そこで，蛋白質と rRNA それぞれについて結合データセットを作成すると，蛋白質では 9 生物種 3,879 座位，rRNA では 26 生物種 2,512 座位からなるデータセットとなった．rRNA のデータについては予備的な解析を行い，その結果とこれまでの知見に基づき 7 系統それぞれの内部の系統関係をあらかじめ特定した．次に両データセットに対し，座位間の進化速度の不均質性を考慮しない解析を行い，7 系統 945 通りの系統樹のトポロジーを探索した．その結果，蛋白質の結合データの最尤系統樹としては図 6A を，rRNA の結合データの最尤系統樹としては図 6B を選択した．蛋白質の最尤系統樹（図 6A）では，微胞子虫は真菌に近縁でそれら共通祖先の姉妹群が動物であるという関係が示されており，これらはいずれも 100%のブートストラップ確率をもって支持されている．これに対し，rRNA の最尤系統樹（図 6B）では，微胞子虫は粘菌と近縁でその姉妹群がユーグレノゾアとなっている．しかしそれらいずれの関係に対しても，ブートストラップ確率による支持は低く（55%, 59%），他の可能性も全く否定できなかった．

さらに別の解析として，連結データをつくるのではなく，各分子種に基づく最尤法の解析を別々に行い，各系統樹のトポロジーに対して，各分子種の解析から得られた対数尤度の和を計算し，それを最大にするトポロジーを最尤系統樹として選択するという「総合評価」の試みを行った．蛋白質の解析，rRNA の解析ともに，図 6 に示したのと同一の系統樹が最尤系統樹として選択され，ブートストラップ確率の値もほぼ同様の傾向にあった（図 6 の括弧内）．

蛋白質の最尤系統樹と rRNA の最尤系統樹の比較をさまざまな解析について示したのが表

2 である．連結データの解析に比べ総合評価の解析では，個々の分子種について個別に枝の長さなどのパラメータの推定を行うため，パラメータ数が多くなっている．このことを考慮に入れてモデル間の比較を行うために AIC を計算して示してある．不均質性を考慮していない解析（without Γ）よりも，考慮した解析（with Γ）の方が明らかに低い AIC 値を示し，いずれの解析においても，総合評価の解析の方が連結データの解析よりも低い AIC 値を示している．これらのことから，不均質性の考慮によりモデルの適合が大きく改善されたことがわかり，パラメータ数の増加というペナルティを考慮に入れたとしても，総合評価の解析の方が良いモデルによる解析であることもわかる．不均質性を考慮しない解析（without Γ）では，連結データ，総合評価いずれの Δl_i の値にも示されているように，蛋白質の解析での rRNA の解析の最尤系統樹（系統樹 B），rRNA の解析での蛋白質の解析の最尤系統樹（系統樹 A）は，それぞれの解析での最尤系統樹に比べて対数尤度の値が非常に低くなっており，明らかに有意に棄却しうる．不均質性を考慮した解析（with Γ）においても，蛋白質の解析での rRNA の系統樹（系統樹 B）はやはり強度に有意に棄却しうる．しかしながら rRNA の解析においては，rRNA の系統樹（系統樹 B）と蛋白質の系統樹（系統樹 A）との対数尤度差の絶対値は，不均質性を考慮しない解析の場合に比べてはるかに小さくなっている．しかも，総合評価の解析ではその差は 11.0 ± 9.9 であり，もはや蛋白質の系統樹（系統樹 A）の可能性を否定できない．すなわち，より良いモデルを用いることにより，rRNA の解析が蛋白質の解析と必ずしも大きく矛盾するものではないことが，明らかになったわけである．これらのことから，微胞子虫が真菌と近縁であるとする蛋白質の系統樹（系統樹 A）の方がより信頼しうるものであることが強く示唆される．表 2 の脚注には，個々の分子種および連結データの解析において推定された Γ 分布の shape パラメータ α の値を示している．0.42（SSUrRNA）から 0.90（TBα）の間の値をとっており，いずれの分子種においても不均質性の度合いは強い．とくに rRNA ではその度合いは顕著である．

5.2 複数分子による個々の解析結果の総合評価

前節の解析では，解析対象とする全ての分子種において，同一の生物種構成，すなわち蛋白質の解析では 9 種，rRNA の解析では 26 種，からなるデータセットを用いて 7 つの系統の関係を検討した．しかし，各系統に対して利用可能な生物種は，用いる分子種間で異なっているというのが一般的な状況である．また，「全ての分子種において同一の生物種の構成」という制約を除けば，各系統内部の生物種数を増やしたり，解析対象としうる分子種の数を増やしたりできる場合があり，より多くの情報が利用可能となる．そのような場合，前節での結合デー

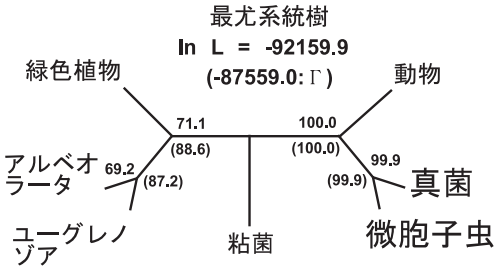

図 7. 10 分子種に基づく総合評価の解析で選択された最尤系統樹．不均質性を考慮しない解析における対数尤度の値とブートストラップ確率を括弧のない数値として示している．不均質性を考慮した解析（with Γ）での各数値は括弧内に示してある．ブートストラップ確率は，表 3 の 29 個のトポロジーについて計算したものである．

タの解析は不可能であるが，総合評価の解析であれば，対数尤度の和という形で各分子種から得られる情報を結合していくことができる．さらに，蛋白質に基づく解析結果と rRNA に基づく解析結果を結合することもできる．

前節では，より良いモデルを適用すれば，rRNA の解析結果が蛋白質の解析結果と大きく矛盾しなくなることをみた．そこで次のステップとして，8 種の蛋白質の解析と 2 種の rRNA の解析を総合評価することにより，合計 10 分子種 6,391 座位のデータに基づき 7 系統の間の関係を検討した．各分子種における 7 系統それぞれの内部の生物種数とその構成は分子種間で異なっているが，それら内部の系統関係は予備的な解析とこれまでの知見に基づきあらかじめ仮定した．図 7 には，座位間の不均質性を考慮しない解析（without Γ）において選択された最尤系統樹を示した．この系統樹では，微胞子虫は真菌と近縁であり（ブートストラップ確率 99.9%），それらの共通祖先の姉妹群が動物になる（100.0%）という関係が示されており，図 6A に示した 8 種の蛋白質（9 生物種）による最尤系統樹の関係に一致している．945 通りの系統樹のトポロジーのうちで，最尤系統樹からの対数尤度差が 3SE（2.3 節参照, Kishino and Hasegawa（1989））以内にあるものもしくは各分子種の個別の解析で最尤系統樹となったものは，全部で 29 個あり，これらについてさらにガンマ分布により座位間の不均質性を考慮した解析（with Γ）を試みた（表 3）．この解析で，29 個の系統樹のうち対数尤度最大および 2 番目となったものは，不均質性を考慮しない解析における最尤系統樹（図 7）および 2 番目の系統樹と一致したが，3 番目以降については若干の入れ替わりが認められた．

いま，これら 29 個の系統樹のそれぞれを対立する系統学的な仮説であるとすると，最尤系統樹からの対数尤度差が有意に大きい系統樹を棄却する，あるいは有意差の認められない系統樹を他の可能性として残しておくことが必要となる．表 3 には，そのための検定の p 値を，Approximately Unbiased（AU）検定（下平（2002）（本特集）），Kishino-Hasegawa（KH）検定（2.3 節参照, Kishino and Hasegawa（1989））, Shimodaira-Hasegawa（SH）検定（Shimodaira and Hasegawa（1999））の別に示した．不均質性を考慮した解析（with Γ）に関して，有意水準を 0.05 とすると，AU 検定では系統樹 2, 4, 5, 6, 7, 8 の 6 個，KH 検定では系統樹 2, 5, 6, 7, 8 の 5 個，SH 検定では 2〜16, 18〜20 の 18 個の系統樹が有意に棄却できないこととなる．これらの数と p 値の絶対値を見ると，多重比較の補正をした SH 検定が補正をしてない KH 検定よりも保守的であることがわかる．さらに AU 検定は，KH 検定よりは保守的であるが，SH 検定ほどには保守的でないこともわかる．いま，AU 検定で棄却できない系統樹を見ると，それらはいずれも微胞子虫と真菌の近縁性を示しておりその姉妹群は動物となっている．SH 検定で棄却されなかった系統樹のうち 2〜14 についてもこの関係は維持されている．したがって，これら検定の結果および図 7 に示したブートストラップ確率の値から見る限り，微胞子虫が真菌と近縁であるとする仮説は強力に支持される．この解析では，微胞子虫と真菌の近縁性を必ずしも積極的に支持しない分子種を含めているため，この結果は，それらの分子種のことを考慮に入れたとしても，「平均的」には微胞子虫-真菌近縁説が正しいのだということを主張している．

ただし，これまでの解析では根もとのない系統樹を対象としているので，この結論は，微胞子虫あるいは真菌に至る枝の上に根もとがないとの仮定のもとでの結論である．これまでの系統学の知見では，動物と真菌の単系統性はほぼ確立した見解となっており，これらいずれかに至る枝の上に根もとが存在しないことは明らかである．もし，微胞子虫のところに根もとがある，すなわち微胞子虫の分岐が最も早いとすると，今回の解析結果は，動物と真菌が単系統にならないことを強力に支持する結果となるため，そのような可能性はほとんど考えられない．したがって，図 7 の系統樹の根もとは，（動物，（真菌，微胞子虫））の内部にはないと考えるのが妥当である．

次に，微胞子虫−真菌近縁説を積極的に支持しない分子種も含め，各分子種による個別の最

表 3. 微胞子虫を含む真核生物の 7 系統群の間の系統関係——10 個の分子種に基づく総合評価 [a].

	without Γ [c]				with Γ [d]			
系統樹のトポロジー[b]	Δl_i [e]	AU[f]	KH[g]	SH[h]	Δl_i [e]	AU[f]	KH[g]	SH[h]
1 (Alv,E,(P,(S,(A,(M,F)))))	(-92159.9)	-	-	-	(-87559.0)	-	-	-
2 (Alv,P,(E,(S,(A,(M,F)))))	27.8	0.249	0.148	0.821	15.1	0.354	0.214	0.859
3 (Alv,E,(S,(P,(A,(M,F)))))	29.0	0.216	0.107	0.801	31.4	0.038	0.022	0.648
4 (Alv,P,((E,S),(A,(M,F))))	30.0	0.343	0.192	0.799	41.5	0.058	0.042	0.509
5 (Alv,(P,E),(S,((M,F),A)))	30.7	0.223	0.115	0.804	19.5	0.292	0.138	0.821
6 (Alv,E,((P,S),(A,(M,F))))	35.6	0.112	0.062	0.739	23.2	0.145	0.085	0.760
7 (Alv,S,((E,P),(A,(M,F))))	42.4	0.197	0.109	0.688	38.0	0.089	0.061	0.556
8 (Alv,(S,(P,E)),(A,(M,F)))	48.2	0.172	0.087	0.603	36.0	0.121	0.079	0.576
9 (Alv,(P,(E,S)),(A,(M,F)))	61.8	0.072	0.042	0.464	66.0	0.005	0.004	0.236
10 (Alv,P,(S,(E,(A,(M,F)))))	68.1	0.010	0.019	0.414	44.5	0.015	0.034	0.470
11 (Alv,(E,S),(P,((M,F),A)))	72.2	0.000	0.017	0.368	74.4	0.003	0.001	0.172
12 (Alv,(E,(P,S)),(A,(M,F)))	91.0	0.004	0.004	0.233	65.0	0.006	0.006	0.249
13 (Alv,S,(P,(E,(A,(M,F)))))	95.0	0.016	0.003	0.196	70.4	0.001	0.002	0.198
14 (Alv,S,(E,(P,(A,(M,F)))))	97.6	0.000	0.002	0.186	77.3	0.001	0.000	0.152
15 (Alv,(P,E),(S,(M,(F,A))))	107.8	0.004	0.001	0.141	76.2	0.001	0.002	0.163
16 (Alv,P,((E,S),(M,(A,F))))	112.6	0.003	0.005	0.122	101.4	0.008	0.001	0.056
17 (Alv,P,((E,S),(F,(A,M))))	113.3	0.003	0.004	0.127	104.9	0.007	0.000	0.049
18 (Alv,(S,(P,E)),(M,(F,A)))	120.2	0.008	0.003	0.099	92.4	0.001	0.002	0.089
19 (Alv,S,((E,P),(M,(A,F))))	122.9	0.001	0.002	0.083	97.2	0.016	0.001	0.065
20 (Alv,S,((E,P),(F,(A,M))))	127.3	0.003	0.001	0.074	102.1	0.000	0.000	0.054
21 (Alv,(P,(E,S)),(M,(F,A)))	130.1	0.003	0.002	0.074	120.5	0.001	0.000	0.022
22 (Alv,M,(P,(E,(S,(A,F)))))	203.8	0.000	0.000	0.002	193.4	0.000	0.000	0.000
23 (Alv,(P,E),(A,((M,F),S)))	229.7	0.007	0.000	0.001	203.3	0.000	0.000	0.000
24 (Alv,(P,(E,M)),(S,(F,A)))	248.1	0.005	0.000	0.000	219.7	0.000	0.000	0.000
25 (Alv,E,((M,F),(A,(P,S))))	262.4	0.000	0.000	0.000	233.6	0.001	0.000	0.000
26 (Alv,(P,(F,A)),(E,(M,S)))	271.3	0.000	0.000	0.000	254.8	0.003	0.000	0.000
27 (Alv,E,(A,((P,S),(M,F))))	289.3	0.000	0.000	0.000	249.7	0.001	0.000	0.000
28 (Alv,M,(P,(S,(A,(E,F)))))	437.2	0.004	0.000	0.000	368.4	0.000	0.000	0.000
29 (Alv,(M,S),(F,(P,(E,A))))	527.6	0.000	0.000	0.000	446.7	0.000	0.000	0.000

[a] 7 つの系統に対する 945 通りの系統樹のトポロジーのうち 29 通りのみについて示した

[b] Alv, アルベオラータ; E, ユーグレノゾア; P, 緑色植物; S, 粘菌; A, 動物; M, 微胞子虫; F, 真菌

[c] 座位間での進化速度の不均質性を考慮しない解析

[d] 座位間での進化速度の不均質性を Γ 分布により考慮に入れた解析. それぞれの分子の最尤系統樹での Γ 分布の shape パラメータ α の推定値: EF1α, 0.57; EF2, 0.59; VRS, 0.71; IRS, 0.68; RPOII, 0.73; ACT, 0.73; TBα, 0.72; TBβ, 0.64; SSUrRNA, 0.46; LSUrRNA, 0.54

[e] 最尤系統樹からの対数尤度の差. ( )内は最尤系統樹の最大対数尤度の値

[f] Approximately unbiased (AU) 検定 (下平 (2002), Shimodaira (2002)) の p 値

[g] Kishino and Hasegawa (1989) による検定の p 値

[h] Shimodaira and Hasegawa (1999) による検定の p 値

表 4. 各分子の最尤系統樹と総合評価による最尤系統樹との対数尤度差.

	without Γ [a]		with Γ [b]	
分子	系統樹No.[c]	Δl_i [d]	系統樹No.[c]	Δl_i [d]
EF1 α	22	-43.6 ± 15.0	22	-12.7 ± 8.1
EF2	23	-17.7 ± 12.0	23	-11.1 ± 9.5
VRS	25	-7.5 ± 13.6	25	-7.5 ± 9.7
IRS	28	-43.4 ± 22.9	28	-23.3 ± 16.0
RPOII	27	-9.7 ± 13.1	6	-2.9 ± 3.7
ACT	24	-18.3 ± 13.7	24	-9.9 ± 10.7
TB α	2	-7.5 ± 7.1	2	-7.9 ± 6.7
TB β	2	-13.0 ± 8.8	2	-10.6 ± 7.7
SSUrRNA	26	-68.0 ± 17.9	26	-13.2 ± 8.7
LSUrRNA	29	-37.0 ± 23.5	29	-11.2 ± 13.8

[a] 座位間での進化速度の不均質性を考慮しない解析

[b] 座位間での進化速度の不均質性を Γ分布により考慮に入れた解析

[c] 各分子の最尤系統樹のトポロジーの表 3 における No.

[d] 最尤系統樹からの対数尤度の差. ±は 1SE (Kishino and Hasegawa (1989))

尤系統樹と総合評価の最尤系統樹（表 3 の系統樹 1，図 7）とがどの程度食い違っているものであるのかを見るために，各解析において，分子種別の最尤系統樹と総合評価の最尤系統樹との対数尤度差を表 4 に示した．不均質性を考慮しない解析（without Γ）では，EF1α と SSUrRNA においてその差が顕著であるが，これらについても，不均質性を考慮する（with Γ）とその差は大幅に減少していることが見てとれる．すなわち，より現実的なモデルを用いた場合には，いずれの分子種の解析結果も総合評価の最尤系統樹の可能性を容認しているのである．一般に Γ 分布の導入により不均質性を考慮した解析を行うと，考慮しない場合に比べて結論が保守的になる場合がある．これは，不均質性を考慮した解析が，考慮しない場合に誤った結論に達してしまう傾向を軌道修正するような役目となっていることを示している．こうした状況では，個々の分子種の解析から有意な結論を得るのは困難である．しかし，このような場合にこそとくに，総合評価により，個々の分子種がもつ情報を結合していくというアプローチが非常に有効であると考えられる．

5.3 真核生物の初期進化：現時点でわかっていること

我々はこれまでに，系統樹の推論のために有用と考えられるいくつかの分子種について，さまざまな原生生物の遺伝子の解析を行うとともに，それらのデータを含めて分子系統樹の解析と総合評価の解析を継続し，真核生物の初期進化に関する知見を積み重ねてきた．図 8 には，真核生物を構成する主な分類群相互の系統関係について，現時点までに明らかとなっていることを模式的に示した．本稿ではここに示した関係のうち，微胞子虫と真菌が近縁でその姉妹群が動物であるとの解析結果を紹介した．このほかにも，最近明らかになった知見として，ストラメノパイルとアルベオラータの近縁性（Arisue et al.(2002a)），エントアメーバとペロバイオンタが近縁でその姉妹群が粘菌であるという関係，すなわちコノサという分類群（Cavalier-Smith (1998)）が単系統群をなすという関係（Arisue et al.(2002b)）の 2 つを挙げることができる．後者については，我々の解析よりもさらに多くの分子種の結合データに基づく解析結果もこの関係を強力に支持した（Bapteste et al.(2002)）．緑色植物と紅藻の近縁性については，Moreira et

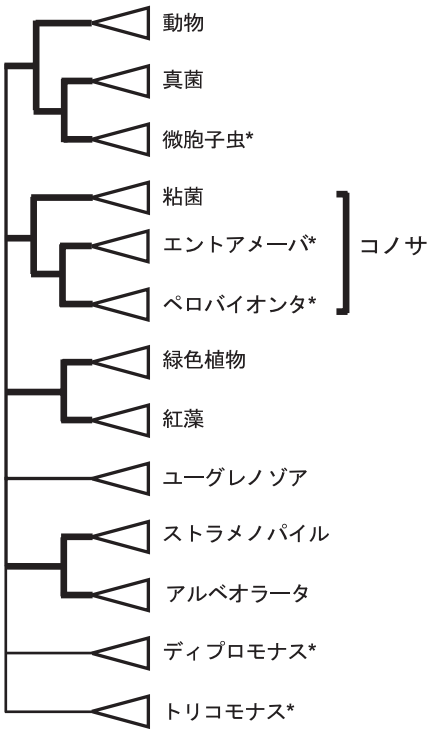

図 8. 真核生物を構成する主な分類群相互の系統進化学的関係．現在までに明らかとなっている関係については，太線で強調して示してある．系統樹の根もとがどこにあるかは未だ不明である．

al.(2000) による複数分子の結合データの解析結果に基づいている．また，Baldauf et al.(2000) には，EF1α，ACT，TBα，TBβ の結合データによる真核生物全体の系統樹が示されているが，その中でも，(動物，(真菌，微胞子虫))，(緑色植物，紅藻)，(アルベオラータ，ストラメノパイル) の三者の関係が復元されている．ただし後二者についてはブートストラップ確率の支持はあまり高くない．図 8 に示した以外の関係については，残念ながら未だはっきりした結論が得られていない．図 8 の不明瞭な部分を明らかにするとともに，真核生物全体の根もとがどこにあるかを解明するのが今後の課題であるが，それを実現するためには，より多くの分子種に基づく解析とそれらの総合評価が必須である．

6. おわりに

本稿では，分子系統樹の推論における問題点の中でとくに深刻な LBA アーテファクトについて紹介し，その克服のために，座位間での進化速度の不均質性を Γ 分布によりモデル化した解析を行うことがある程度有効であることを示した．しかし，このようなモデルでも，現実の分子進化の過程からはかなりかけ離れている．4.2 節では座位間の進化速度の相対比が進化時間を通じて一定であると仮定したが，実際には，ある座位の進化速度は進化時間を通じて変化しうる．これは，Fitch の 'covarion model' とよばれ (Fitch (1971))，いかなる座位も進化系統樹のある部分では変化しうるが，他の部分では変化しえないという状況を仮定する．このモ

デルは今回解析に用いた 4.2 節のモデルよりもはるかに現実的なものである．近年，この方向でのモデル化の研究が始まりつつあり（Tuffley and Steel（1998), Galtier（2001), Penny and Hasegawa（2001), Penny et al.（2001)），今後の進展が望まれる．

分子系統樹の解析には，LBA の他にも数多くの問題点が存在している．たとえば，塩基やアミノ酸組成の系統間での顕著な偏りが系統樹の推定を誤らせる問題，置換の飽和に伴なう情報ロスの問題，置換に関する遷移確率モデルが現実とかけ離れているような場合に生ずる問題などである．とくに，今回のように生物の歴史の非常に古い時代にまで遡って分子系統樹を推定しようとする場合，予期できないものを含めさまざまなノイズがデータに存在しており，問題をより難しいものとしている．そうした状況において，一つもしくは少数の分子種のデータ解析のみから，系統に関する推論を行うのは極めて危険である．今回は，多くの証拠に基づく「平均的」な推論を下すため，個々の分子種の解析結果を対数尤度の和という形で総合評価するアプローチの有効性についても指摘した．今後，このようなアプローチによりさまざまな具体的問題が解決されるものと期待される．

参 考 文 献

Adachi, J. and Hasegawa, M.（1996）. MOLPHY Version 2.3: Programs for molecular phylogenetics based on maximum likelihood, *Comput. Sci. Monographs*, No. 28, The Institute of Statistical Mathematics, Tokyo.

Akaike, H.（1974）. A new look at the statistical model identification, *IEEE Trans. Automat. Control*, AC-**19**, 716–723.

Arisue, N., Hashimoto, T., Yoshikawa, H., Nakamura, Y., Nakamura, G., Nakamura, F., Yano, T. and Hasegawa, M.（2002a）. Phylogenetic position of *Blastocystis hominis* and of Stramenopiles inferred from multiple molecular sequence data, *Journal of Eukaryotic Microbiology*, **49**, 42–53.

Arisue, N., Hashimoto, T., Lee, J. A., Moore, D. V., Gordon, P., Sensen, C. W., Gaasterland, T., Hasegawa, M. and Müller, M.（2002b）. The phylogenetic position of the Pelobiont *Mastigamoeba balamuthi* based on sequences of rDNA and translation elongation factors EF-1α and EF-2, *Journal of Eukaryotic Microbiology*, **49**, 1–10.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. and Doolittle, W. F.（2000）. A kingdom-level phylogeny of eukaryotes based on combined protein data, *Science*, **290**, 972–977.

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M. and Philippe, H.（2002）. The analysis of one hundred genes supports the grouping of three highly divergent amoebae, *Dictyostelium*, *Entamoeba* and *Mastigamoeba*, *Proc. Nat. Acad. Sci. U.S.A.*, **99**, 1414–1419.

曹　纓，長谷川政美（2002）. 分子系統樹推定におけるモデルのミススペシフィケーション ── 脊椎動物の系統進化を例として ──，統計数理，**50**, 69–85.

Cavalier-Smith, T.（1987）. Eukaryotes with no mitochondria, *Nature*, **326**, 332–333.

Cavalier-Smith, T.（1998）. A revised six-kingdom system of life, *Biological Reviews*, **73**, 203-266.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C.（1978）. A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, Vol. 5, suppl. 3（ed. M. O. Dayhoff), 345–352, National Biomedical Research Foundation, Washington, D. C.

Edlind, T. D., Li, J., Visvesvara, G. S., Vodkin, M. H., McLaughlin, G. L. and Katiyar, S. K.（1996）. Phylogenetic analysis of β-tubulin sequences from amitochondrial protozoa, *Molecular Phylogenetics and Evolution*, **5**, 359–367.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.*, **27**, 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap, *Evolution*, **38**, 16–24.

Fitch, W. M. (1971). The nonidentity of invariable positions in the cytochromes c of different species, *Biochemical Genetics*, **5**, 231–241.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model, *Molecular Biology and Evolution*, **18**, 866–873.

Germot, A., Philippe, H. and Le Guyader, H. (1997). Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*, *Molecular and Biochemical Parasitology*, **87**, 159-168.

Hasegawa, M. and Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny, *Molecular Phylogenetics and Evolution*, **2**, 1–5.

Hasegawa, M. and Kishino, H. (1994). Simple methods for estimating bootstrap probability of a maximum likelihood tree, *Molecular Biology and Evolution*, **11**, 142–145.

長谷川政美，岸野洋久（1996）．『分子系統学』，岩波書店，東京.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, 160–174.

Hasegawa, M., Kishino, H. and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics, *Journal of Molecular Evolution*, **32**, 443-445.

Hashimoto, T. and Hasegawa, M. (1996). Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G, *Advances in Biophysics*, **32**, 73–120.

Hashimoto, T., Sánchez, L. B., Shirakura, T., Müller, M. and Hasegawa, M. (1998). Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny, *Proc. Nat. Acad. Sci. U.S.A.*, **95**, 6860-6865.

Hilis, D. M., Moritz, C. and Mable, B. K. (eds.)(1996). *Molecular Systematics*, 2nd ed., Sinauer Associates, Sunderland, Massachusetts.

Hirt, R. P., Logsdon, J. M., Healy, M. W., Dorey, W. F., Doolittle, W. F. and Embley, T. M. (1999). Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins, *Proc. Nat. Acad. Sci. U.S.A.*, **96**, 580-585.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences, *Comput. Appl. Biosci.*, **57**, 94–97.

Kamaishi, T., Hashimoto, T., Nakamura, Y., Masuda, Y., Nakamura, F., Okamoto, K., Shimizu, M. and Hasegawa, M. (1996). Complete nucleotide sequences of the genes encoding translation elongation factors 1α and 2 from a microsporidian parasite, *Glugea plecoglossi*: Implications for the deepest branching of eukaryotes, *Journal of Biochemistry*, **120**, 1095-1103.

Keeling, P. J. and Doolittle, W. F. (1996). Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family, *Molecular Biology and Evolution*, **13**, 1297–1305.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, Massachusetts.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea, *Journal of Molecular Evolution*, **29**, 170–179.

Kishino, H., Miyata, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny

and the origin of chloroplasts, *Journal of Molecular Evolution*, **30**, 151–160.

Leipe, D. D., Gunderson, J. H., Nerad, T. A. and Sogin, M. L. (1993). Small subunit ribosomal RNA^{+} of *Hexamita inflata* and the quest of the first branch in the eukaryotic tree, *Molecular and Biochemical Parasitology*, **59**, 41–48.

宮田　隆 編 (1998). 『分子進化――解析の技法とその応用――』, 共立出版, 東京.

Moreira, D., Le Guyader, H. and Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts, *Nature*, **405**, 69–72.

Penny, D. and Hasegawa, M.(2001). Covarion model of molecular evolution, *Encyclopedia of Genetics* (eds. S. Brenner and J.H. Miller), 473–477, Academic Press, San Diego, California.

Penny, D., McComish, B. J., Charleston, M. A. and Hendy, M. D. (2001). Mathematical elegance with biochemical realism: The covarion model of molecular evolution, *Journal of Molecular Evolution*, **53**, 711–723.

Peyretaillade, E., Biderre, C., Peyret, P., Duffieux, F., Méténier, G., Gouy, M., Michot, B. and Vivarés, C.P. (1998). Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core, *Nucleic Acids Research*, **26**, 3513–3520.

Philippe, H. and Laurent, J. (1998). How good are deep phylogenetic trees?, *Current Opinion in Genetics & Development*, **8**, 616–623.

Roger, A. J. (1999). Reconstructing early events in eukaryotic evolution, *American Naturalist*, **154**, S146–S163.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection, *Ann. Inst. Statist. Math.*, **50**, 1–13.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution*, **16**, 1114–1116.

Shimodaira, H. and Hasegawa, M. (2001). CONSEL: A program for assessing the confidence of phylogenetic tree selection, *Bioinformatics*, **17**, 1246–1247.

下平英寿 (2002). ブートストラップ法によるクラスタ分析のバラツキ評価, 統計数理, **50**, 33-44.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, **51**, 492–508.

高木利久, 金久　實 編 (1996). 『ゲノムネットのデータベース利用法』, 共立出版, 東京.

Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution, *Math. Biosci.*, **147**, 63–91.

Weiss, L. M., Edlind, T. D., Vossbrinck, C. R. and Hashimoto, T. (1999). Microsporidian molecular phylogeny: The fungal connection, *Journal of Eukaryotic Microbiology*, **46**, 17S–18S.

Yang, Z.(1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Molecular Biology and Evolution*, **10**, 1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods, *Journal of Molecular Evolution*, **39**, 306–314.

Yang, Z.(1996). Among-site rate variation and its impact on phylogenetic analyses, *Trends in Ecology & Evolution*, **11**, 367–372.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood, *Computer Applications in the Biosciences*, **13**, 555–556.

Application of Molecular Phylogenetic Inference and Associated Problems: Illustrative Data Analysis on Early Eukaryotic Evolution

Tetsuo Hashimoto

(The Institute of Statistical Mathematics; Department of Biosystems Science, The Graduate University for Advanced Studies)

Nobuko Arisue

(Department of Biosystems Science, The Graduate University for Advanced Studies)

Masami Hasegawa

(The Institute of Statistical Mathematics; Department of Biosystems Science, The Graduate University for Advanced Studies)

The maximum likelihood method of molecular phylogeny, which infers an evolutionary tree based on sequence data of DNA, RNA and proteins, is briefly described and applied to a data analysis on early eukaryotic evolution. Possible existence of a long branch attraction artefact is introduced. This artefact has recently been regarded as one of the most serious problems in making an inferred tree misleading. To overcome this problem, evolutionary rate heterogeneity across sites is taken into consideration by Γ-distribution. With this approach, the phylogenetic position of microsporidia at the basal position of the eukaryotic tree in several previous analyses is shown to be an artefact caused by long branch attraction. The extremely high evolutionary rate of microsporidia in the molecules used in previous analyses may have been a major cause of the artefact. Re-analyses of the currently available molecular data with rate heterogeneity across sites and a combined analysis of these data clearly demonstrate that microsporidia are not early branching eukaryotes but are closely related to fungi.

Key words: Maximum likelihood method of molecular phylogeny, long branch attraction, rate heterogeneity across sites, Γ-distribution, early eukaryotic evolution, microsporidia.