統計数理(2008)
第 56 巻 第 1 号 145–164

特集「分子進化と統計科学」
[研究詳解]

複数遺伝子の結合データに基づく分子系統樹の推測

―真核生物の大系統の解析を例として―

橋本 哲男[1,2] ・有末 伸子[3] ・坂口 美亜子[1] ・稲垣 祐司[1,2]

(受付 2008 年 2 月 1 日)

要　　旨

複数の遺伝子のもつ情報を結合して最尤法により分子進化系統樹に関する推測を行うための方法論の概略を述べ，真核生物の大系統の問題に関するデータ解析の実例を示した．

結合のための統計モデルとして，単に個々の遺伝子(もしくは全データセットを構成する個々の‘区分’)の連結データに対して 1 セットの枝長を推定する「連結モデル」，個別の遺伝子(区分)それぞれについて独立に枝長の推定を行う「分離モデル」，枝長が遺伝子(区分)間で比例しているという仮定を置く「比例モデル」の 3 つのモデルを取り上げ，真核生物 29 種からなる 53 個のリボソームタンパク質全 5,842 座位のデータに適用した．枝長の推定法とデータの分割法に関して，異なる 6 種類のモデルによる解析を AIC により比較した結果，リボソームの大小サブユニット区分による分離モデルの AIC 値が最も低く，このモデルの適合が最も良いことが明らかとなった．遺伝子区分による分離モデルの AIC 値は最も高く，パラメータが過剰であると考えられた．このことから，53 個のリボソームタンパク質間で進化パターンが比較的均質である可能性が示唆された．系統樹の樹型の選択という観点からは，6 種類の解析結果に大差はなく，今回のリボソームタンパク質による解析結果は頑健なものと考えられた．

キーワード： 分子系統樹の推測，最尤法，複数遺伝子による解析，真核生物，大系統，リボソームタンパク質．

1. はじめに

遺伝子やゲノム解析技術の飛躍的な進歩と広範な生物学研究者への浸透を背景として，多様な生物種にわたる膨大な配列データが蓄積されつつある．これらのデータは個別の生物種の生化学的・遺伝学的・分子生物学的研究を展開するための基礎データとして必須であるばかりでなく，これらのデータをもとに生物種間での比較解析を行うことにより，生物の進化の歴史に関する推測が可能となるという点においても非常に重要である．配列データに基づいて生物の進化系統樹に関する推測を行う研究分野は分子系統学とよばれるが，近年のデータ増大に伴いその重要性が認識されてきている．分子系統学において系統樹推測の手掛かりを与えるのは，

[1] 筑波大学大学院 生命環境科学研究科：〒 305–8572 茨城県つくば市天王台 1–1–1
[2] 筑波大学 計算科学研究センター：〒 305–8572 茨城県つくば市天王台 1–1–1
[3] 大阪大学 微生物病研究所：〒 565–0871 大阪府吹田市山田丘 3–1

DNA や RNA における塩基置換やタンパク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，複数の配列をアライメントして座位を揃えて比較すると，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種間で比較することによって，系統樹の樹形と枝の長さ(座位当たりの置換数を単位とする)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)ため，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，系統樹の推測を行うためには，確率モデルに基づいた統計的な方法が必要である．その方法を最初に最尤法の枠組みで定式化したのは，Felsenstein (Felsenstein, 1981)，実際のデータ解析を通じて最尤法を分子系統学の分野に広めるとともに，更なる方法論の開発・改良を行ってきたのは長谷川・岸野(長谷川・岸野, 1996)のグループである．

Felsenstein は 1980 年代から，最尤法のプログラムを PHYLIP パッケージの中に整備していたが，最尤法による解析は最尤系統樹の探索に膨大な計算時間を必要とするため，当時の計算機環境の下ではあまり実際問題に適用されなかった．しかしながら，計算機の性能の向上を背景にシミュレーション研究が進み，最尤法の望ましい性質，すなわち，「系統間での進化速度の一定性が成り立たないような場合にも頑健な推測を与える」という性質が広く知られるようになるにつれて，最尤法による解析の成果が公表されるようになっていった．また，更なる計算機の性能の向上や新しいアルゴリズムの開発・改良に伴い，最尤法でのデータ解析を高速に扱える多くのプログラムが開発されるに至った．たとえば最近，通常の長さ(1,500 塩基や 500 アミノ酸)をもつ数百タクサもの配列データに対して，ヒューリスティックな探索を行い最良な尤度をもつ系統樹を探すのに，普通のパソコンで 1 日以上かからないようなプログラムが開発されている(たとえば，RAxML HPC (Stamatakis et al., 2005; Stamatakis, 2006))．こうした良好な計算環境のもと，現在，最尤法によるデータ解析は分子系統学の研究分野に広く浸透している．

一方，分子系統学研究の現場では，系統マーカーとして複数の遺伝子を用いると，遺伝子間で矛盾した解析結果が得られるという状況の多いことが明らかとなった．その矛盾が統計的誤差の範囲を超えているという場合も少なからず存在することも判明した．しかしこのような状況が生ずるのは当然の帰結とも考えられた．個々の遺伝子にはそれぞれの歴史があり，固有の進化パターンがあるのにもかかわらず，それらを同一の，しかも極めて不十分な進化モデルのもとで比較しているという状況を鑑みれば，誤差自体が過少推定であることは疑う余地はない．また，遺伝子間で統計的誤差の範囲内の矛盾が生ずることは頻繁であるため，一般に 1 つの遺伝子に含まれている，系統に関する情報のうち真の系統を反映するシグナルの総和がそれほど大きくないことも明らかである．こうした状況下，シグナルの増強のために複数の遺伝子の情報を結合して系統の推測を行うという試みがなされるようになってきた．そして現在，大規模データの蓄積や計算環境の飛躍的改善と相俟って，複数遺伝子の結合データ解析は日常的なアプローチとして頻繁に行われている．

地球上の生物は，真核生物(*Eukarya*)，真正細菌(*Bacteria*)，古細菌(*Archaea*)のいずれかの大きな系統的グループに属すると考えられており，これら 3 つのグループを三大超生物界という．真核生物は細胞内に核膜で囲まれた核をもつ生物群で，我々の日常生活に卑近なものとしては，後生動物，菌類(カビ・キノコなど)，陸上植物などが含まれる．しかしながら，真核生物超生物界の系統的多様性の大部分は原生生物(プロティスト)と総称される単細胞の真核微生物によって占められている．真核生物がどのようにして原核生物から進化し，現状の系統的多様性をもつに至ったのかという真核生物の初期進化の問題は進化生物学上，最も重要な問題であるが，研究はあまり進んでおらず未解決な部分が多い．それを解明するためには原生生物の系統的多様性と進化過程を解明する必要がある．かつて，原生生物の遺伝子のデータはほとん

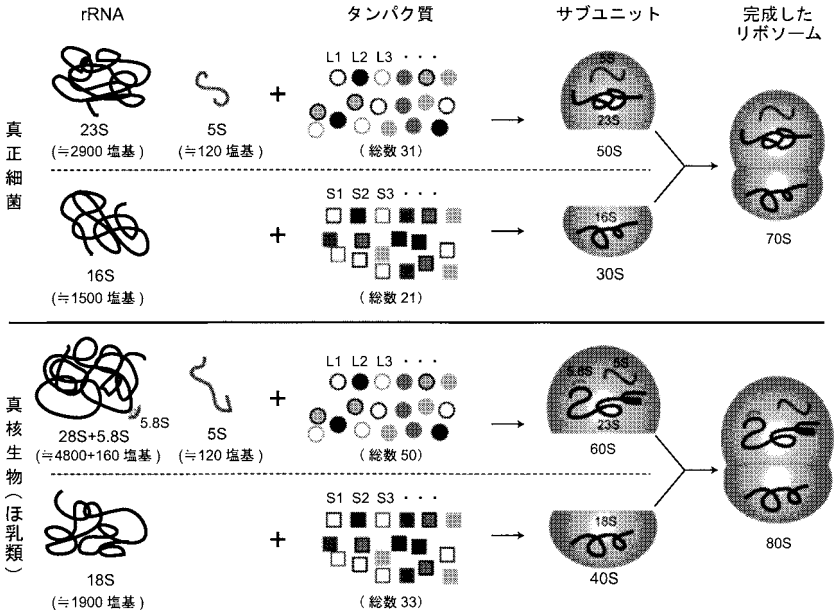

図 1. リボソームの構成成分．S は沈降係数を表す．リボソームタンパク質の番号の表示体系は真正細菌と哺乳類とで異なっており(表 1)，原則として同一番号のものは相同タンパク質ではない．

ど報告されていなかったが，近年の遺伝子解析技術の発展に伴い，さまざまな原生生物に関する遺伝子解析が行われ，配列データが蓄積されつつある．これらをもとに，原生生物の系統進化を探るとともに，真核生物の初期進化における高次の系統群の分岐の過程をたどることが可能となりつつある．

本稿では，最尤法の枠組みにおいて結合データ解析を扱う方法の概略を説明し，真核生物の大系統の解析への適用例について紹介する．さらに，結合データ解析の現状の問題点を指摘する．とくに，細胞質のタンパク質合成装置—リボソーム—を構成する複数のタンパク質の遺伝子データをさまざまな真核生物(大多数は原生生物)間で比較解析し，最尤法により系統樹の推測を行う．リボソームは全ての生物に普遍的に存在し，構成タンパク質や RNA の配列がよく保存されており進化速度が遅いため，真核生物超生物界全体を通した系統樹の解析のために適切な材料であると考えられる(図 1)．

2. 複数遺伝子結合データ解析の統計モデル

最尤法による分子系統樹解析においては，与えられた系統樹の樹型(トポロジー)と置換確率モデルのもとで，解析の対象とするデータ行列の全座位に対する尤度を最大にするように枝の長さなどのパラメータを推定する(Felsenstein, 1981; 長谷川・岸野, 1996)．以下，Pupko et al. (2002)に準じて，最尤法の枠組みにおいて頻繁に用いられる結合データ解析の概略を説明する．

n 種からなる系統樹のある樹型 T における枝(branch)は $2n-3$ 個あるので，それらそれぞれの長さを，$t_1,\ldots,t_{2n-3}$ とおく．また，枝長以外のパラメータを θ とおく．θ には，置換確率モデルに含まれるパラメータ，塩基やアミノ酸の組成値(π)，座位間の進化速度の不均質性をモデル化するための Γ 分布の α パラメータなどが含まれる．いま n 種それぞれに対して m

個に区分(たとえば，遺伝子 m 個)されたデータがあるとする．このとき一番単純な解析法は，n 種それぞれについて，m 個分のデータを単純に連結したデータ行列を作り，それに対して $t_1,\ldots,t_{2n-3}$ と θ を推定するという方法である．このモデルを「連結モデル」といい，その同時確率は，

$$\begin{aligned}&P(\mathrm{data\_1}\&\cdots\&\mathrm{data\_}m|t_1,\ldots,t_{2n-3},\theta,T)\\&\quad=P(\mathrm{data\_1}|t_1,\ldots,t_{2n-3},\theta,T)\times\cdots\times P(\mathrm{data\_}m|t_1,\ldots,t_{2n-3},\theta,T)\end{aligned}$$

という形で表現できる．このモデルでは，全ての区分(遺伝子)で同一の枝長と θ をもつことが仮定されている．一方，各々の区分(遺伝子)で枝長や θ は独立であるという仮定をおくこともできる．すなわち，m 個の区分(遺伝子)のデータを個別に解析するという方法である．このモデルを「分離モデル」といい，同時確率は，

$$\begin{aligned}&P(\mathrm{data\_1}\&\cdots\&\mathrm{data\_}m|t_1^{(1)},\ldots,{t_{2n-3}}^{(1)},\ldots,t_1^{(m)},\ldots,{t_{2n-3}}^{(m)},\theta^{(1)},\ldots,\theta^{(m)},T)\\&\quad=P(\mathrm{data\_1}|t_1^{(1)},\ldots,{t_{2n-3}}^{(1)},\theta^{(1)},T)\times\cdots\times P(\mathrm{data\_}m|t_1^{(m)},\ldots,{t_{2n-3}}^{(m)},\theta^{(m)},T)\end{aligned}$$

と表せる．さらに，枝長は区分(遺伝子)間で比例しているとの仮定を置くこともできる．すなわち，ある区分(遺伝子)の枝長が $t_1,\ldots,t_{2n-3}$ のとき別の区分(遺伝子)の枝長は $rt_1,\ldots,rt_{2n-3}$ と表せるというものである(Yang, 1996)．このモデルでは個々の枝の長さ(進化速度)の比が区分(遺伝子)間で一定であることを仮定している．θ を区分(遺伝子)ごとに推定するものとすると，同時確率は，

$$\begin{aligned}&P(\mathrm{data\_1}\&\cdots\&\mathrm{data\_}m|t_1,\ldots,t_{2n-3},r_1,\ldots,r_m,\theta^{(1)},\ldots,\theta^{(m)},T)\\&\quad=P(\mathrm{data\_1}|t_1,\ldots,t_{2n-3},r_1,\theta^{(1)},T)\times\cdots\times P(\mathrm{data\_}m|t_1,\ldots,t_{2n-3},r_m\theta^{(m)},T)\end{aligned}$$

という形になる．

枝長に関するパラメータ数は，連結モデルでは $2n-3$，分離モデルでは $m(2n-3)$，上記の形の比例モデルでは，$2n-3+(m-1)=2n+m-4$ となり，分離，比例，連結の順にパラメータ数が多い．今，アミノ酸レベルの解析を行うものとし，経験的なアミノ酸置換確率(PAM モデル(Dayhoff et al., 1978)，JTT モデル(Jones et al., 1992)，WAG モデル(Whelan and Goldman, 2001)など)を用い，アミノ酸組成をデータから推定するものとし，連結モデルの解析では連結データに対して 1 つの α (Γ 分布のパラメータ)を，分離，比例モデルでは遺伝子ごとに α を推定するものとすると，各モデルのパラメータ数は以下のようになる．

すなわち，連結モデル：

$$(2n-3)+(20-1)+1=2n+17,$$

分離モデル：

$$m(2n-3)+m(20-1)+m=m(2n+17),$$

比例モデル：

$$(2n+m-4)+m(20-1)+m=2n+21m-4.$$

これらパラメータ数の異なるモデル間での適合の良さを比較するためには，赤池情報量規準(Akaike Information Criterion, AIC) (Akaike, 1974)を用いる．AIC は，

$$AIC=-2\times\text{対数尤度}+2\times\text{パラメータ数}$$

により定義される量であり，異なるモデル間で AIC の値を比較し，AIC が最も小さいモデルを最良のモデルとして選択する．

3. データ解析例：リボソームタンパク質遺伝子に基づく真核生物系統樹の推測

3.1 対象と方法

真核生物の高次の系統群のうち以下の 1)～10)のグループにおける括弧内の生物種(俗名)について，データベース検索により全てのリボソームタンパク質遺伝子の配列データを収集した：1)オピストコント(後生動物：ヒト，ハエ，線虫；菌類：クリプトコッカス，分裂酵母，出芽酵母)；2)アメーボゾア(細胞性粘菌，赤痢アメーバ)；3)緑色植物(シロイヌナズナ，クラミドモナス，オステレオコッカス)；4)紅色植物(シアニディオシゾン，ガルディエリア，オゴノリ)；5)ストラメノパイル(ブラストシスチス，珪藻，卵菌)；6)アルベオラータ(繊毛虫：テトラヒメナ，ゾウリムシ；アピコンプレックス：クリプトスポリディウム，トキソプラズマ，タイレリア，マラリア原虫)；7)ユーグレノゾア(ユーグレナ，リューシュマニア，トリパノソーマ)；8)ヘテロロボサ(ナエグレリア)；9)ディプロモナス(ランブル鞭毛虫)；10)パラバサリア(トリコモナス)．全リボソームタンパク質についてデータベースを精査した結果，現時点で上記 29 生物種のデータが全て揃うものは 53 個あり，それらを表 1 に示した．そのそれぞれについて，Clustal W プログラム(Thompson et al., 1994)を用いてマルチプルアライメントを作成し，マニュアルでそれを修正した．53 個それぞれのタンパク質アライメントから，アライメントに曖昧さを伴わない座位を選択して系統樹推測のためのデータ行列とした．全 53 タンパク質での総計は 5,842 アミノ酸座位となった(表 1)．

分子系統樹の推測のための解析プログラムとしては，ヒューリスティックな系統樹探索のためには，RAxML プログラム(Stamatakis et al., 2005; Stamatakis, 2006)，ユーザー tree による解析には，PAML プログラム(Yang, 1997)を用いた．また，一部，対数尤度の集計のために MOLPHY プログラム(Adachi and Hasegawa, 1996)も併用した．RAxML の解析では 100 個のブートストラップサンプル(Felsenstein, 1985)を用いて内部枝の信頼度を評価した．RAxML の解析以外の解析においては，RELL ブートストラップ法(Kishino et al., 1990)により近似的なブートストラップ値を算出した．複数の系統樹の間で対数尤度の差を比較するためには，CONSEL プログラム(Shimodaira and Hasegawa, 2001)の中の Approximately Unbiased (AU)検定(Shimodaira, 2002; 下平, 2002)を用いた．

3.2 予備的解析

最初に，29 生物種について 53 個のタンパク質のアミノ酸配列データを単純に繋ぎ合わせたデータ(連結モデル)をもとに，ヒューリスティックな系統樹探索を行った．RAxML プログラムにより，アミノ酸置換モデルとして JTT を用い，アミノ酸組成をデータから推定(F オプション)し，座位間の進化速度の不均一性を離散 Γ 分布で近似して解析した結果，図 2 に示す系統樹が最良な系統樹として選択された．さらに，100 個のブートストラップサンプルについても同様に解析し，図 2 の各内部枝に記したブートストラップ値を得た．

図 2 の系統樹では，アメーボゾアを除いて，既に確立された高次系統群それぞれの単系統性が高いブートストラップ値をもって復元されている．すなわち，ユーグレノゾア + ヘテロロボサ(ディスキクリスタータ)，アルベオラータ，ストラメノパイル，緑色植物，紅色植物，およびオピストコントの単系統性である．近年広く認識されている，ランブル鞭毛虫(ディプロモナス)とトリコモナス(パラバサリア)の近縁性(Baldauf et al., 2000; Arisue et al., 2005; Simpson et al., 2006)についても 100%のブートストラップ値で復元されている．さらにこれら高次系統群同士の関係を見てみると，緑色植物と紅色植物の単系統性(Moreira et al., 2000; Rodriguez-Ezpeleta et al., 2005)およびアルベオラータとストラメノパイルの単系統性(Arisue et al., 2002; Rodriguez-Ezpeleta et al., 2005, 2007)が復元されており，これらも近年広く受け入れられている仮説である．ただしブートストラップ値の支持はいずれも低い．一方，アメー

表 1. 解析に用いたリボソーム蛋白質一覧.

番号	真正細菌での番号	哺乳類での番号	進化的保存度区分[a]	解析座位数
大サブユニット (L)				
L1	L1	L10a	EAB	104
L2	L2	L8	EAB	231
L3	L3	L3	EAB	259
L4	L4	L4	EAB	156
L5	L5	L11	EAB	105
L10	L10	P0	EAB	230
L11	L11	L12	EAB	150
L13	L13	L13a	EAB	158
L14	L14	L23	EAB	124
L15	L15	L27a	EAB	104
L18	L18	L5	EAB	167
L22	L22	L17	EAB	124
L23	L23	L23a	EAB	53
L24	L24	L26	EAB	66
L29	L29	L35	EAB	98
L30	L30	L7	EAB	154
mL10		L10	EA	194
mL15		L15	EA	168
mL18		L18	EA	123
mL19		L19	EA	111
mL27		L27	E	79
mL32		L32	EA	93
mL37		L37	EA	64
mL40		L40	EA	49
				(L 小計) 3164
小サブユニット (S)				
S2	S2	Sa	EAB	127
S3	S3	S3	EAB	181
S4	S4	S9	EAB	97
S5	S5	S2	EAB	127
S7	S7	S5	EAB	177
S8	S8	S15a	EAB	126
S10	S10	S20	EAB	67
S11	S11	S14	EAB	112
S12	S12	S23	EAB	135
S13	S13	S18	EAB	130
S14	S14	S29	EAB	31
S15	S15	S13	EAB	133
S19	S19	S15	EAB	61
mS3a		S3a	EA	186
mS4		S4	EA	82
mS6		S6	EA	95
mS7		S7	E	81
mS8		S8	EA	108
mS12		S12	E	62
mS17		S17	EA	50
mS19		S19	EA	99
mS21		S21	E	56
mS24		S24	EA	103
mS25		S25	E	43
mS26		S26	E	34
mS27		S27	EA	69
mS27a		S27a	EA	34
mS28		S28	EA	33
mS30		S30	E	39
				(S 小計) 2678
				(総計) 5842

[a] EAB, 真核生物、古細菌、真正細菌のいずれにも存在するもの；
EA, 真核生物と古細菌に存在するもの； E, 真核生物にのみ存在するもの

ボゾアの単系統性は復元されておらず，赤痢アメーバがディプロモナス＋パラバサリアの姉妹群のところに高いブートストラップ値(88%)で位置づけられている．この原因は定かではないが，比較的枝の長い赤痢アメーバと顕著に枝の長い，ランブル鞭毛虫＋トリコモナスの間で，長い枝同士が間違えて結合し易いという Long Branch Attraction (LBA)のアーテファクト(Felsenstein, 1978; Philippe and Laurent, 1998; 橋本 他, 2002)が起こっているという可能性が考えられる．アメーボゾア，ディプロモナス，パラバサリアのいずれのグループもタクソンサンプリングが不十分であるということが LBA をもたらした要因かもしれない．一方，細胞性粘菌と赤痢アメーバが共通祖先をもつ(アメーボゾア単系統)とし，その枝をオピストコントの共通祖先のところに移動させた系統樹(図 2 の矢印)は，現時点の系統進化学の知見からみて，お

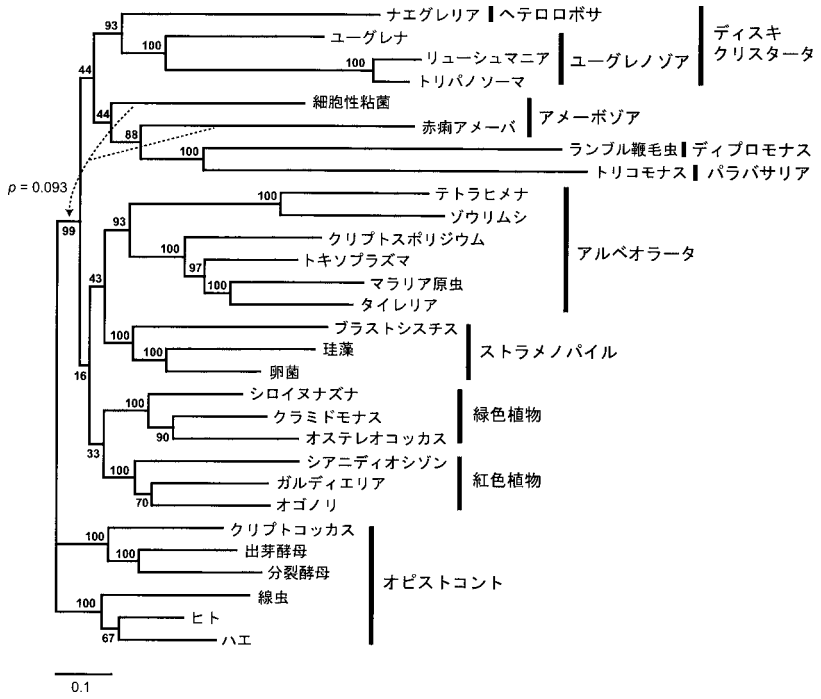

図 2. リボソームタンパク質による真核生物全体の系統樹．RAxML プログラムによる解析結果．29 生物種 5842 座位の連結データ（連結モデル）をもとに 10 個の初期系統樹（最大節約系統樹）を出発点として，ヒューリスティック探索により到達した系統樹のうちで尤度最大（lnL = −194268.6）のものを示した．アミノ酸置換モデルとして JTT + Γ を用い，アミノ酸の組成値はデータから推定した．Γ 分布パラメータは，$\alpha = 0.8092$．枝の長さは推定アミノ酸置換数を表す．内部枝上の数値はブートストラップ値（%）で，100 回のリサンプリングデータに基づく結果を示した．赤痢アメーバと細胞性粘菌が共通祖先をもつとして，その祖先を点線の矢印の位置に移動させたときの系統樹とこの系統樹を比較する AU 検定（本文参照）の p 値を示した．

そらく‘正しい’系統樹である．すなわち，基本的な細胞の体制が 1 本鞭毛である，オピストコントとアメーボゾアが近縁（ユニコント）で，それ以外の 2 本鞭毛を基本体制とする生物群（バイコント）から区別しうる（Stechmann and Cavalier-Smith, 2003）という系統樹であり，複数遺伝子による分子系統樹解析によっても支持されている（Bapteste et al., 2002; Rodriguez-Ezpeleta et al., 2005, 2007）．この‘正しい’系統樹の対数尤度ともとの系統樹の対数尤度の差は有意ではない（$p > 0.05$，AU 検定）ため，図 2 の系統樹が最良なものであったとしても，必ずしもこの解析で尤もらしい結果が得られなかったということにはならない．

次に，前述の解析と同様に連結モデルを用いて，あらかじめ与えた系統樹に対する網羅的探索を行った．まず，図 2 の結果とこれまでに確立されている知見をもとに，①アメーボゾア＋オピストコント（ユニコント），②緑色植物＋紅色植物，③アルベオラータ，④ストラメノパイル，⑤ディプロモナス＋パラバサリア，⑥ディスキクリスタータの単系統性をあらかじめ仮定し，これら 6 つの系統に対して可能な全 105 通りの系統樹を網羅的に探索した（解析 I）．各高次系統群内部の関係としては図 2 の系統樹の関係を用いた．PAML パッケージの中の codeml プログラムを用い，ヒューリスティック探索の際と同様に JTT (F) + Γ モデルにより解析し

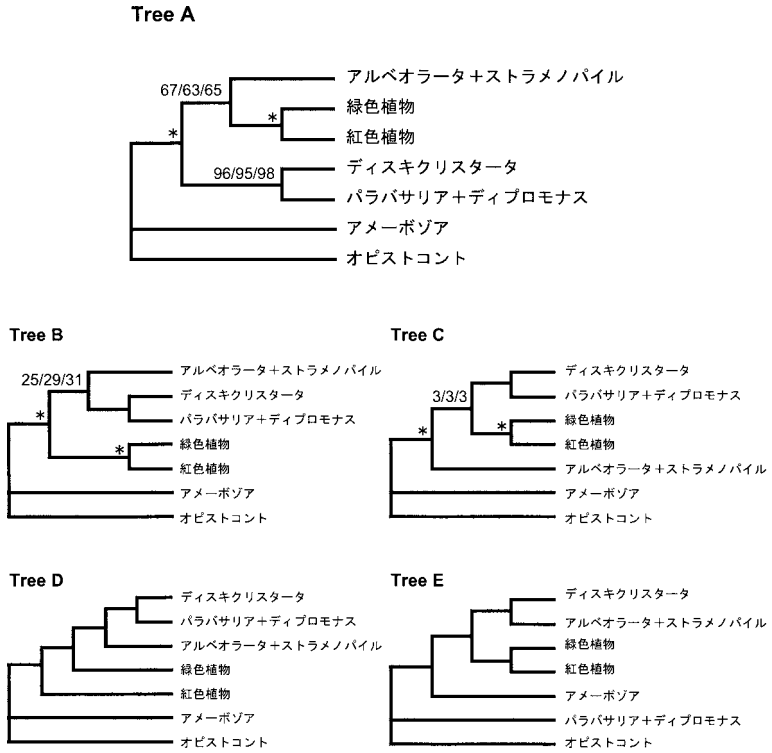

図 3. 対立仮説として検討した 5 系統樹の樹型．表 2 の各系統樹の樹型を描いた．5 つの系統（本文参照）に対する全 15 通りの系統樹の網羅的解析の結果，いずれのモデルの解析においても，最尤系統樹は Tree A，第 2 位，第 3 位の系統樹はそれぞれ Tree B，Tree C となった．Tree D，Tree E は 15 通りに含まれないが，対立仮説として検討した系統樹（本文参照）．Tree A∼C の内部枝上の数値は RELL ブートストラップ値（%）で，分離モデル_LS/連結モデル/分離モデル_遺伝子の順に示し，他のモデルによる結果は省略した．*は 15 通りの解析であらかじめ制約を置いた関係である．

た．さらに，上述の 6 つの系統のうちの②③④を，②緑色植物，③紅色植物，④アルベオラータ＋ストラメノパイルのように置き換えた解析を行った（解析 II）．その結果，I，II いずれの解析も，ディスキクリスタータとディプロモナス＋パラバサリアが近縁で，緑色植物＋紅色植物とアルベオラータ＋ストラメノパイルが近縁であるとする系統樹（図 3 の Tree A）を最尤系統樹として選択した．

3.3 5 つの大系統群間の系統関係の解析

予備解析の結果をもとに，考慮の対象とする大系統群を 5 つに絞り，①アメーボゾア＋オピストコント（ユニコント），②緑色植物＋紅色植物，③アルベオラータ＋ストラメノパイル，④ディプロモナス＋パラバサリア，⑤ディスキクリスタータとした．これら 5 つの系統に対する 15 通りの系統樹について，連結の方法に対するさまざまな統計モデルを仮定して，網羅的探索を行った．まず，連結モデルの解析は上述の連結データをもとに行った．次に 53 個のタンパク質を大サブユニット（LSU）タンパク質と小サブユニット（SSU）タンパク質とに二分し（図 1），LSU と SSU の区分に対して，比例モデルと分離モデルの解析を行った．また，53 個のタンパク質を，進化的保存の程度の別，すなわち *Eukarya*（真核生物），*Archaea*（古細菌），*Bacteria*

表 2.　対立仮説として検討した 5 系統樹のモデル別比較.

	モデル	連結モデル		比例モデル _LS 区分		分離モデル _LS 区分	
	パラメータ数	75		96		150	
	最尤系統樹 / AIC 値	Tree A / 388663.2		Tree A / 388726.8		Tree A / 388494.4	
		Δl	p 値[b]	Δl	p 値[b]	Δl	p 値[b]
Tree A	((Opi,Amo),(EH,PD),((GP,RP),AS))	(-194256.6)[a]	0.795	(-194267.4)[a]	0.793	(-194097.2)[a]	0.812
Tree B	((Opi,Amo),(GP,RP),(AS,(EH,PD)))	-6.7	0.496	-7.0	0.524	-8.9	0.493
Tree C	((Opi,Amo),AS,((GP,RP),(EH,PD)))	-18.6	0.097	-18.0	0.093	-18.7	0.116
Tree D	((Opi,Amo),RP,(GP,(AS,(EH,PD))))	-30.2	0.142	-30.5	0.134	-31.2	0.133
Tree E	((Opi,PD),Amo,((GP,RP),(AS,EH)))	-42.5	0.043	-43.2	0.034	-45.0	0.036

	モデル	比例モデル _EAB 区分		分離モデル _EAB 区分		分離モデル _ 遺伝子区分	
	パラメータ数	117		225		3975	
	最尤系統樹 / AIC 値	Tree A / 388741.2		Tree A / 388710.8		Tree A / 390972.6	
		Δl	p 値[b]	Δl	p 値[b]	Δl	p 値[b]
Tree A	((Opi,Amo),(EH,PD),((GP,RP),AS))	(-194253.6)[a]	0.790	(-194130.4)[a]	0.815	(-191511.3)[a]	0.762
Tree B	((Opi,Amo),(GP,RP),(AS,(EH,PD)))	-8.0	0.527	-8.2	0.482	-8.7	0.532
Tree C	((Opi,Amo),AS,((GP,RP),(EH,PD)))	-20.6	0.068	-18.6	0.128	-29.3	0.082
Tree D	((Opi,Amo),RP,(GP,(AS,(EH,PD))))	-31.8	0.119	-31.9	0.130	-34.7	0.171
Tree E	((Opi,PD),Amo,((GP,RP),(AS,EH)))	-42.3	0.042	-43.4	0.048	-63.2	0.033

[a]（ ）内は最尤系統樹の対数尤度
[b] Approximately Unbiased (AU) 検定

(真正細菌)の三大超生物界に共通に存在するもの(EAB)，*Eukarya* と *Archaea* に共通に存在するもの(EA)，および *Eukarya* だけに存在するもの(E)の 3 つに分け(表 1)，比例モデルと分離モデルの解析を行った．さらに，53 個のタンパク質それぞれを別々に扱うという分離モデルの解析も行った．

これら 6 つのモデルに基づく解析のいずれも，図 3 および表 2 の Tree A を最尤系統樹として選択した．表 2 には，6 つのモデルいずれにおいても同様に 2 番目，3 番目の対数尤度値を示した Tree B および Tree C の解析結果を併記した．さらに，15 通りの中に含まれない対立仮説として，緑色植物と紅色植物の単系統性に対立する仮説として提唱されている Tree D (Nozaki et al., 2003, 2007)，および α-チューブリンを含む解析において頻繁に高い可能性をもって復元される Tree E (Arisue et al., 2005; Simpson et al., 2006)を取り上げ，それらの解析結果も併記した．Tree D の関係は，バイコント生物群のなかで最初に分岐したのは紅色植物であるという仮説である．Tree E では，オピストコントとディプロモナス＋パラバサリアが近縁となっており，ユニコントがまとまっていない．これら 5 つの系統樹に対する対数尤度値の順位は 6 つのモデルにおいて同一であり，系統樹の選択という観点からモデル間の差は認められなかった．すなわち，いずれのモデルも Tree B~D を棄却できず，Tree E を有意水準 5% で棄却した．一方，アミノ酸置換モデルとして，JTT (F)＋ Γ モデルの代わりに WAG (F)＋ Γ モデル，PAM (F)＋ Γ モデルを用いて同様の解析を行った(データ不表示)．全般的に JTT (F)＋ Γ モデルに比べて PAM (F)＋ Γ モデルはやや高い対数尤度の値を与え，WAG (F)＋ Γ モデルはさらに高い対数尤度の値を与えた．しかしながら，系統樹の選択という観点では，いずれのモデルによる結果も JTT (F)＋ Γ モデルによる結果と全く同様であった．このように，置換モデルを変えても結合データ解析のモデルを変えても推測の結果に大差がないということから，リボソームタンパク質による今回のデータ解析の結果は非常に頑健なものであると考えられた．

Tree A~C は，網羅的に探索した 15 通りの系統樹のうち，④ディプロモナス＋パラバサリ

アと⑤ディスキクリスタータを近縁であるとする系統樹で，この関係に対する RELL ブートストラップ値は，いずれのモデルにおいても 95%以上にのぼった．この関係は，エクスカベートというグループの一部が単系統であるという関係に相当する．エクスカベートは形態学から単系統性が示唆されている鞭毛虫のグループで，その多くは細胞の腹側に大きな捕食口をもつという特徴をもつ．エクスカベートが単系統か否かということは現在の真核生物の系統進化学上の非常に大きな問題である．今回の解析では，エクスカベートとして位置づけられている系統群の半数以上のグループのデータを含めることができなかったため，現時点でエクスカベートの単系統性をきちんと論じることはできないが，ディプロモナス＋パラバサリアとディスキクリスタータの単系統性が復元されたことは注目に値する．一方，エクスカベート(ディプロモナス＋パラバサリア＋ディスキクリスタータ)と緑色植物＋紅色植物およびアルベオラータ＋ストラメノパイルの 3 者の間の関係は，Tree A∼C の対数尤度差，p 値，BP 値にみられるように，今回の解析からは明確にできなかった．

緑色植物と紅色植物が単系統でないとする説は，これまで複数のグループから提案されてきたが(Stiller et al., 2001; Stiller and Hall, 2002; Nozaki et al., 2003)，近年行われた大規模な複数遺伝子解析によって完全に否定された(Rodriguez-Ezpeleta et al., 2005)．ところが 2007 年に再びこの説が別の複数遺伝子解析の結果をもとに提唱され(Nozaki et al., 2007)，論議を呼んでいる．今回の解析の Tree D は Nozaki らの説に相当するが，いずれのモデルの解析もこの系統樹を有意に棄却せず，少なくとも今回のリボソームタンパク質のデータからはこの問題に決着をつけることはできなかった．一方，ディプロモナス＋パラバサリアとオピストコントが近縁であるという Tree E は今回の解析では否定され，リボソームタンパク質のデータセットは，α-チューブリンに顕著に存在するシグナルをもたないことが明らかとなった．

3.4 モデル間の比較

パラメータ数の異なる 6 つのモデルのデータへの適合の良さを比較するために，Tree A に対する各モデルの AIC 値を表 2 に示した．

最小の AIC 値を与えたモデルすなわち最も適合の良いモデルは分離モデル_LS 区分であった．表 3 には，Tree A に対する各モデルの AIC 値の，分離モデル_LS 区分における最小 AIC 値からの差(ΔAIC)を示した．さらに，対数尤度の差の標準誤差の算出式(Kishino and Hasegawa, 1989)にしたがって，ΔAIC の標準誤差を併記した．AIC の意味で 2 番目に良いモデルは「連結モデル」であった．3 番目以降は，分離モデル_EAB 区分，比例モデル_LS 区分，比例モデル_EAB 区分と続き，AIC 最大のモデルは，分離モデル_遺伝子区分であった．分離モデル_LS 区分以外のいずれのモデルも分離モデル_LS 区分との比較において標準誤差をはるかに超える ΔAIC 値を示した．とくに，分離モデル_遺伝子区分においては ΔAIC 値が 2474.6 ± 370.0 となり，最小 AIC 値との顕著な差を示した．一方，表 2 において Tree A に対する対数尤度の値をモデル間で比較すると，分離モデル_遺伝子区分の対数尤度が最も高い値であり，2 番目の分

表 3. 最尤系統樹 (Tree A) の AIC 値のモデル間比較.

モデル	ΔAIC	±(S.E.)
連結モデル	+168.8	±57.0
比例モデル _LS 区分	+232.4	±40.6
分離モデル _LS 区分	(388494.4)	
比例モデル _EAB 区分	+246.8	±76.8
分離モデル _EAB 区分	+216.4	±83.8
分離モデル _ 遺伝子区分	+2474.6	±370.0

離モデル_LS 区分との間で $\Delta l = 2585.9$ もの大差を示している．3 番目は分離モデル_EAB 区分で，以後比例モデル_EAB 区分，連結モデル，比例モデル_LS 区分の順となっている．分離モデル_遺伝子区分では，遺伝子ごとに枝長，アミノ酸組成，Γ 分布の α パラメータを推定するので，他のモデルよりデータへの適合が良く対数尤度が大きくなるのは当然である．しかしながらパラメータ数が膨大なもの(3,975)となるため，そのペナルティを考慮して AIC で比較すると，最も適合の良くないモデルということになってしまうのである．すなわち，膨大なパラメータを使ったのにもかかわらず，パラメータ数の少ないモデル，たとえば連結モデルに比べて，パラメータ数の増加に見合うほど充分に適合が改善されてないということになる．この結果は，これまでに主として哺乳類の系統に関してミトコンドリアコードタンパク質や核コードタンパク質のデータ解析によって得られた知見とは異なっている．これらの解析では，いずれも分離モデル_遺伝子区分の方が連結モデルよりも良い適合を与えている(たとえば，Cao et al., 2000a, 2000b; Pupko et al., 2002)．最近，哺乳類の解析において膨大な数の核コード遺伝子の結合データ解析が行われた(Nishihara et al., 2007)が，その解析においても分離モデル_遺伝子区分の方が連結モデルよりも低い AIC 値を与えた．この解析では 2,789 個もの遺伝子を用いているため，分離モデル_遺伝子区分のパラメータ数は膨大なものとなるが，その適合は結合モデルに比べて，パラメータ数のペナルティを相殺しさらに相殺分を上回るほどに改善されたという結果となった．このように，複数遺伝子の結合データ解析では，一般に異なる遺伝子間での進化パターンの不均質性が顕著な場合が多いため，分離モデル_遺伝子区分の方が連結モデルよりも低い AIC 値を与える場合が多いのである．一方，リボソームは多くのタンパク質と RNA の分子複合体，すなわち超分子システムである(図 1)．その結果，その構成要素である個々のタンパク質や RNA の進化は協調的に起こっている可能性があり，ミトコンドリアコードタンパク質や核コードの他のタンパク質に比べると，異なるタンパク質間での進化パターンは均質になっているものと考えられる．今回のリボソームタンパク質遺伝子の結合データ解析では，各遺伝子間での進化パターンの不均質性があまり大きくなかったため，連結モデルの方に低い AIC 値が与えられたのであろう．その意味で，リボソームタンパク質は全体で 1 つの大きなタンパク質とみなすことができよう．

次に表 2 の Tree A において，LS 区分について比例モデルと分離モデルを比較してみると，比例モデルのほうの AIC 値は分離モデルのそれに比べて，232.4 ± 40.6 大きい(データ不表示)ことがわかり，このことから，分離モデルの方の適合が良いと結論できる．LS 区分の比例モデルに基づく系統樹の L 区分に相当する Tree A を図 4A に示した．L 区分の総枝長は 10.6 であり，L 区分の枝長を 1 とすると S 区分の枝長に対する比例定数は 0.887 となった．一方，分離モデルにおいて，L 区分，S 区分それぞれのデータに基づく Tree A を図 4 の B，C にそれぞれ示した．総枝長はそれぞれ，10.5, 9.6，S 区分／L 区分の比は 0.914 となり，比例モデルに基づく値とほぼ同様の結果が得られた．このことから，小サブユニットタンパク質(S 区分)の方が大サブユニットタンパク質よりも平均的に進化速度が若干遅いということが明らかとなった．分離モデルの 2 つの系統樹，図 4B と図 4C を比べてみると，各枝のほとんどにおいて，枝長の長さのパターンはほぼ同様の傾向を示したが，ランブル鞭毛虫(Gin)とトリコモナス(Tva)の組とガルディエリア(Gsu)とオゴノリ(Gch)の組に関しては，L 区分と S 区分とで外部枝の長さが逆転している．すなわち，トリコモナスとオゴノリでは L 区分の進化速度が大きく，逆に，ランブル鞭毛虫とガルディエラでは S 区分の進化速度が大きくなっている．分離モデルではこのことが考慮されるが，比例モデルでは枝長の配分比は L 区分と S 区分とで一律に決まってしまうので，区分間でのこのようなパターンの違いは考慮されない．この点が比例モデルの適合を分離モデルのそれより悪くしている原因であると考えられる．

さらに，EAB 区分について比例モデルと分離モデルを比較してみると(表 2)，比例モデル_EAB

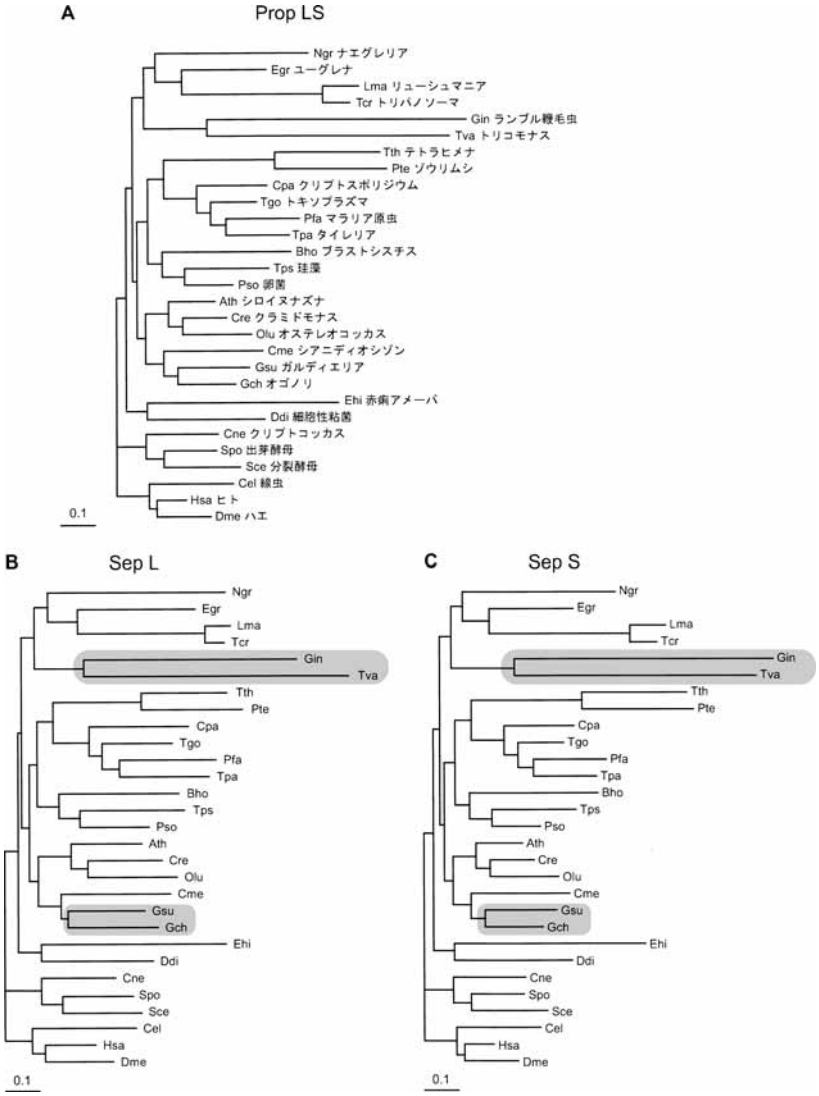

図 4. 大小サブユニット (LS) 区分の解析による Tree A の樹型．A，比例モデルによる L 区分の系統樹．L 区分の枝長：S 区分の枝長=1：0.887．Γ 分布のパラメータは，L 区分の $\alpha = 0.795$，S 区分の $\alpha = 0.782$．B，分離モデルによる L 区分の系統樹で 3,164 座位による解析結果．Γ 分布のパラメータは，$\alpha = 0.801$．C，分離モデルによる S 区分の系統樹で 2,678 座位による解析結果．Γ 分布のパラメータは，$\alpha = 0.770$．パネル B，C において，大小サブユニット (LS) 間で外部枝の長さ (進化速度) のパターンが異なっている部分に影をつけて示した．

区分の AIC 値は，分離モデル_EAB 区分の AIC 値に対して 30.4 ± 34.0 大きく (データ不表示) なっている．しかしながら，この差の絶対値は標準誤差の範囲内に収まっているので，顕著なものとは考えられない．すなわち，2 つのモデルの適合度に大差はなさそうである．図 5 には比例モデルの EAB 区分に対する Tree A をパネル A に，分離モデルの EAB 区分，EA 区分，

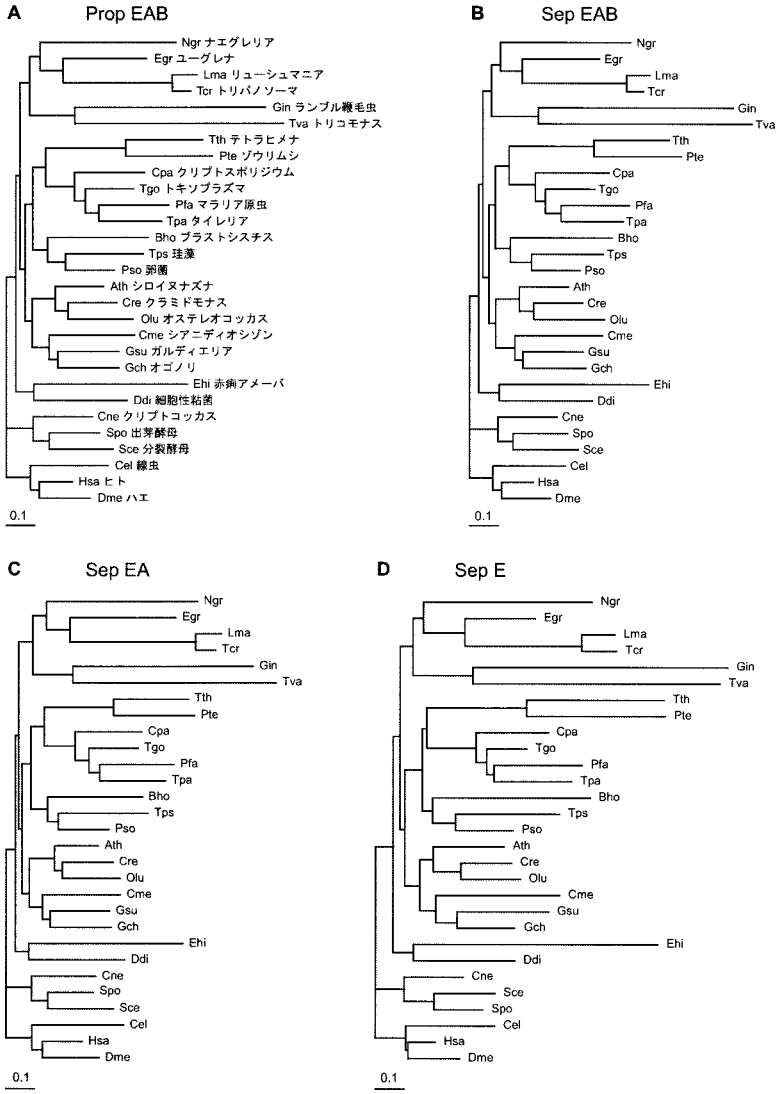

図 5. 進化的保存度(EAB，EA，E)区分の解析による Tree A の樹型．A，比例モデルによる EAB 区分の系統樹．EAB 区分の枝長：EA 区分の枝長：E 区分の枝長＝ 1：1.010：1.225．Γ 分布の α パラメータは，EAB 区分の $\alpha = 0.750$，EA 区分の $\alpha = 0.827$，E 区分の $\alpha = 1.044$．B，分離モデルによる EAB 区分の系統樹で 3,787 座位による解析結果．Γ 分布の α パラメータは，$\alpha = 0.749$．C，分離モデルによる EA 区分の系統樹で 1,661 座位による解析結果．Γ 分布の α パラメータは，$\alpha = 0.826$．D，分離モデルによる E 区分の系統樹で 394 座位による解析結果．Γ 分布の α パラメータは，$\alpha = 1.024$．

E 区分それぞれに対応する Tree A をそれぞれ，パネル B，C，D に示した．比例モデル_EAB 区分の総枝長は 10.0，EA 区分，E 区分に対する枝長の比はそれぞれ 1.01，1.23 であった．一方分離モデルによる各区分の総枝長はそれぞれ 10.0，10.1，12.4 であり，比例モデルでの解析結果とほぼ同様の傾向を示した．分離モデルの 3 つの系統樹を比較すると，区分間で各枝長の

パターンはほぼ同様であったが，E 区分では全体的に枝長が長くなっていた．これらのことから，三大生物界に共通に存在するタンパク質(EAB)と真核生物と古細菌に共通に存在するタンパク質(EA)では，それらの進化速度がほぼ同様の傾向にあるのに対し，真核生物にのみ存在するタンパク質(E)ではその進化速度が増加していることが示唆された．適合度についてみれば，E 区分における進化速度比以外に 3 区分間での進化パターンに大きな差がなかったため，比例モデルと分離モデルの AIC 値に顕著な差が見られなかったものと考えられる．しかしながら，今回の解析では E 区分に属するタンパク質のデータは少なく，5,842 座位中たった 394 座位にすぎない．今後，この座位数をさらに増やして E 区分の進化速度に関する再検討を行う必要がある．

今回，分離・比例モデルの解析においてデータの分割に用いた区分は，LS 区分，EAB 区分，遺伝子区分(分離モデルのみ)であり，分離モデル_LS 区分が最良という結果となったが，LS 区分よりもデータへの適合を良くしうるような区分法が存在するであろうことは明白である．すなわち，すべてのリボソームタンパク質を何らかの方法で，類似した進化パターンをもつ複数のグループに区分できれば，そのような区分による分離もしくは比例モデルの解析はデータへの適合を向上させるであろう．そのためには個々のタンパク質の進化パターンに対する詳細な予備的解析が必須であるとともに，構造や機能に関する広範な分子情報の蓄積も不可欠である．

4. 複数遺伝子結合データ解析の問題点

近年，配列データの急速な蓄積と相俟って，分子系統学の多くの研究分野において，複数遺伝子の結合データ解析が行われるようになった．今回取り上げた問題，すなわち，真核生物の初期進化の過程で，大きな系統群が分岐する順番を推測するという問題に関しても，これまでに数多くの結合データ解析が行われてきている．しかしながら，結合データ解析の結果が必ずしも真核生物の大系統に関して一貫した結論を与えてきたというわけではない．用いる生物種(タクサ)や遺伝子の組み合わせが異なっている場合，統計的誤差の範囲を超えて相反する結論が導かれたという場合も存在しており，非常に混乱が生じている．たとえば，2007 年 8 月号の Molecular Biology and Evolution 誌には，Nozaki et al. (2007) と Hackett et al. (2007) という 2 つの論文が掲載されている．前者では，進化速度が遅いと判断される 19 個の核コードタンパク質(うち 10 個はリボソームタンパク質)から選択した 5,216 座位が用いられ，紅色植物は緑色植物とは単系統群を形成せず，バイコント生物群の根もとから分岐する可能性が高い(図 3 の Tree D)との主張がなされている．後者では，前者の解析と同じもの 5 個を含む 16 個の核コードタンパク質(リボソームタンパク質はなし)から選択した 6,735 座位を用いて，一次共生に由来する葉緑体をもつ植物，すなわち，紅色植物，緑色植物，および灰色植物が単系統群を形成するという可能性が高いことが示されている．両者の解析で用いられた遺伝子，タクサは大幅に異なっているため，これらを一概に比較してどちらか一方が信頼しうる結果であるという判断を下すことは困難である．おそらくこの程度の数の遺伝子の解析では，ある遺伝子の組み合わせでは真の系統に関するシグナルが増強されてくるが，別の遺伝子の組み合わせではノイズの方が増強されてシグナルが隠されてしまうという状況なのであろう．したがって，こうした論文の結論はあくまでも解析に用いたデータセットに依存した結論であるということを十分承知しておくことが必要である．遺伝子の数を増加させていくにつれて，ある特定の系統関係が復元される可能性が徐々に高まっていくという結果がある程度の頑健性を伴って示されない限り，その関係が真の系統であるとの判断を下すことはできないであろう．

一方，結合データ解析の方法論に応じて結論が変わりうるという点にも注意が必要である．前述のように，異なる遺伝子の結合データ解析の際には，分離モデルの方が連結モデルよりも

データにより良く適合する場合が多い．また，連結モデルの解析によって強く支持された結論が，より適合の良い分離モデルの解析ではあまり支持されないという場合も見受けられる．このようなときには，分離モデルの結果を尊重してあまり強い主張をすべきではない(たとえば，Cao et al., 2000a; Iida et al., 2007)．また，分離モデルと連結モデルによる推測結果が明らかに異なる場合も観測されている(Takishita et al., 2005)．ところが，一般的に分離モデルによる解析は手間がかかるため，連結モデルによる解析結果だけをもとに結論が下されてしまう場合が多い．前述の膨大な量のデータ(約 100 万座位)を使った哺乳類の解析(Nishihara et al., 2007)においても，ヌクレオチドレベルの解析では連結モデルと分離モデルで最尤系統樹として選ばれる系統樹が異なっており，結合データ解析のモデルの相違が解析結果に大きく影響しうることを示している．このデータはほぼ全ゲノムの配列が決定されている生物種間を対象に，系統樹の推測のために比較可能な遺伝子を全て取り込んだ解析であり，取得可能な全遺伝子のデータに基づく解析であるといえる．そのような膨大なデータ量を用いているのにも関わらずモデル依存的に解析結果が変わるというのが現実である．その意味で，より現実的な進化モデルの開発をたゆまなく追及すると同時に，さまざまなモデル間の相互比較・評価に関わる方法論を現実のデータ解析に即して確立していくことが，今後の分子系統学における結合データ解析にとって非常に重要であると考えられる．

5. 真核生物の初期進化：現時点でわかっていること

最後に，真核生物超生物界を構成する高次の系統群相互の系統関係について，分子系統学的研究により現時点までに広く認められていることを図 6 に示した．

まず，真核生物の起源については諸説があるが(Embley and Martin, 2006)，図 6 では核・細胞質系が古細菌由来であると仮定している．ミトコンドリアの細胞内共生は，現存の全ての真核生物の共通祖先の段階で起こったものと考えられる．色素体の一次共生は，植物(緑色，紅色，灰色)の共通祖先のところで起こったと考えられ，アルベオラータ，ストラメノパイルやユーグレノゾアなどに存在する色素体は，一度色素体を獲得した単細胞真核藻類が他の生物種に二次的に共生することによって生じたものとみなされている．

近年，複数遺伝子の結合データ解析の成果と系統分類学的知見の蓄積により，真核生物はいくつかの非常に大きなグループから構成されると考えられるようになった．それらはスーパーグループと呼ばれ，図 6 に太線で示したオピストコント，アメーボゾア，リザリア，プランテ，クロムアルベオラータ，エクスカベートの 6 つが相当する．これらのうち，リザリア，クロムアルベオラータ，エクスカベートそれぞれの単系統性については複数遺伝子の結合データ解析からは支持されていない．また，プランテについても図 6 には単系統群として示してあるが，前述のように紅色植物の分岐は緑色植物の分岐よりも早いという仮説もある(Nozaki et al., 2003, 2007)．紅色植物由来の二次共生色素体をもつアルベオラータとストラメノパイルの近縁性(Arisue et al., 2002; Harper et al., 2005)は以前から指摘されていたが，近年，同じく紅色植物由来の二次共生色素体をもつハプト植物とクリプト植物の近縁性も明確に示された(Patron et al., 2007; Hackett et al., 2007)．これら 4 つがクロムアルベオラータとして単系統となる可能性は，現時点ではあまり高くない．むしろそれに反する解析結果も提出されている(Hackett et al., 2007; Burki et al., 2007)．エクスカベートについてみると，①ユーグレノゾアとヘテロロボサのディスキクリスタータとしての単系統性，さらにその姉妹群としてのヤコバの位置づけ，②フォルニケータの単系統性とその姉妹群としてのパラバサリアの位置づけ，および③トリマスティクスとオキシモナスの単系統性はいずれも明確に示されているが，エクスカベート全体の単系統性は未だに支持されるに至っていない．一方，有中心粒類太陽虫やアプソゾアなど未

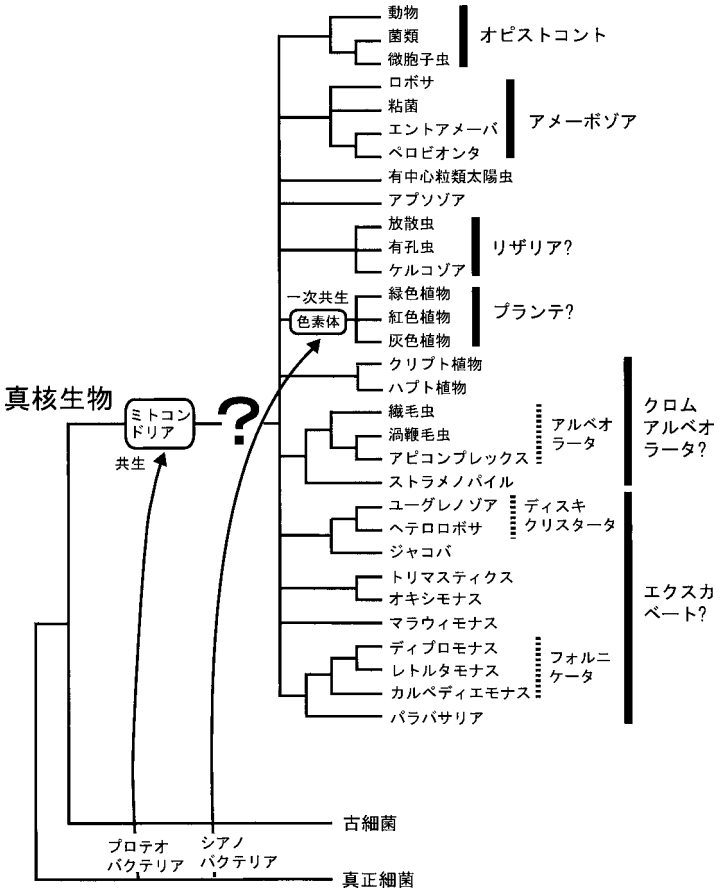

図 6. 真核生物の初期進化に関する現在の知見．現在までに支持されている高次系統群の関係を示した．複数遺伝子の結合データ解析によって単系統性が明確に示されていないスーパーグループ名には？をつけて示した（本文参照）．

だ位置づけの明らかでないグループも存在する（Sakaguchi et al., 2007; Moreira et al., 2007）．図 6 に多分岐として示した部分の分岐順を明らかにするとともに，真核生物系統樹の根もとを決めることが今後の課題であるが，それを実現するためには多くの遺伝子・タクサによる洗練された結合データ解析が必須である．

本稿では，リボソームタンパク質の結合データが真核生物の大系統の解析の素材として有用である可能性を示した．近年，さまざまな真核生物種において，細胞で発現している mRNA の網羅的配列解析が行われるようになってきている．高い発現レベルをもつリボソームタンパク質の配列データはこうした解析から容易に得られるデータである．その意味で，リボソームタンパク質は今後の大規模結合データ解析のための素材の一部として重要である．

謝　辞

本稿は，統計数理研究所共同研究（H06-共研-A59，H18-共研-1013）および筑波大学計算科学研究センター PACS-CS プロジェクト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(17370086)および筑波大学学内プロジェクト(代表 井上 勲)の資金援助を受けた.

参 考 文 献

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

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

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

Arisue, N., Hasegawa, M. and Hashimoto, T. (2005). Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data, *Molecular Biology and Evolution*, **22**, 409–420.

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, L. E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müler, M. and Philippe, H. (2002). The analysis of one hundred genes supports the grouping of three highly divergent amoebae, *Dictyostelium, Entamoeba* and *Mastigamoeba*, *Proceedings of the National Academy of Science USA*, **99**, 1414–1419.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S. I., Jakobsen, K. S. and Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups, *PLoS ONE*, **2**, e790.

Cao, Y., Sorenson, M. D., Kumazawa, Y., Mindel, D. P. and Hasegawa, M. (2000a). Phylogenetic position of turtles among amnioites: evidence from mitochondrial and nuclear genes, *Gene*, **259**, 139–148.

Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. and Hasegawa, M. (2000b). Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, *Gene*, **259**, 149–158.

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.

Embley, T. M. and Martin, W. (2006). Eukaryotic evolution, changes and challenges, *Nature*, **440**, 623–630.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading, *Systematic Zoology*, **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.

Hackett, J. D., Yoon, H. S., Li, S., Reyes-Prieto, A., Rümale, S. E. and Bhattacharya, D. (2007). Phylogenetic analysis supports the monophyly of cryptophytes and haptophytes and the association of Rhizaria with Chromalveolates, *Molecular Biology and Evolution*, **24**, 1702–1713.

Harper, J. T., Waanders, E. and Keeling, P. J. (2005). On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes, *International Journal of Systematic and Evolutionary Microbiology*, **55**, 487–496.

長谷川政美，岸野洋久(1996). 『分子系統学』，岩波書店，東京.

橋本哲男，有末伸子，長谷川政美(2002). 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―，統計数理，**50**, 45–68.

Iida, K., Takishita, K., Ohshima, K. and Inagaki, Y. (2007). Assessing the monophyly of chlorophyll-*c* containing plastids by multi-gene phylogenies under the unlinked model conditions, *Molecular Phylogenetics and Evolution*, **45**, 227–238.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences, *Computer Applications in the Biosciences*, **57**, 94–97.

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

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.

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

Moreira, D., von der Heyden, S., Bass, D., Lopez-Garcia, P., Chao, E. and Cavalier-Smith, T. (2007). Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata, *Molecular Phylogenetics and Evolution*, **44**, 255–266.

Nishihara, H., Okada, N. and Hasegawa, M. (2007). Rooting the eutherian tree: The power and pitfalls of phylogenomics, *Genome Biology*, **8**, R199.

Nozaki, H., Matsuzaki, M., Takahara, M., Misumi, O., Kuroiwa, H., Hasegawa, M., Shin-i, T., Kohara, Y., Ogasawara, N. and Kuroiwa, T. (2003). The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids, *Journal of Molecular Evolution*, **56**, 485–497.

Nozaki, H., Iseki, M., Hasegawa, M., Misawa, K., Nakada, T., Sasaki, N. and Watanabe, M. (2007). Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes, *Molecular Biology and Evolution*, **24**, 1592–1595.

Patron, N. J., Inagaki, Y. and Keeling, P. J. (2007). Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages, *Current Biology*, **17**, 887–891.

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

Pupko, T., Huchon, D., Cao, Y., Okada, N. and Hasegawa, M. (2002). Combining multiple data sets in a likelihood analysis: Which models are the best?, *Molecular Biology and Evolution*, **19**, 2294–2307.

Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H. J., Philippe, H. and Lang, B. F. (2005). Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes, *Current Biology*, **15**, 1325–1330.

Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roger, A. J., Gray, M. W., Philippe, H. and Lang, B. F. (2007). Toward resolving the eukaryotic tree: The phylogenetic positions of jakobids and cercozoans, *Current Biology*, **17**, 1420–1425.

Sakaguchi, M., Inagaki, Y. and Hashimoto, T. (2007). Centrohelida is still searching for a phylogenetic

home: Molecular data analyses of seven *Raphidiophrys contractilis* genes, *Gene*, **405**, 47–54.
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.
Shimodaira, H. and Hasegawa, M. (2001). CONSEL: A program for assessing the confidence of phylogenetic tree selection, *Bioinformatics*, **17**, 1246–1247.
Simpson, A. G. B., Inagaki, Y. and Roger, A. J. (2006). Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes, *Molecular Biology and Evolution*, **23**, 615–625.
Stamatakis, A., Ludwig, T. and Meier, H. (2005). RAxML-III: A fast program for maximum likelihood-based inference of phylogenetic trees, *Bioinformatics*, **21**, 456–463.
Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688–2690.
Stechmann, A. and Cavalier-Smith, T. (2003). The root of the eukaryote tree pinpointed, *Current Biology*, **13**, R665–R666.
Stiller, J. W., Riley, J. and Hall, B. D. (2001). Are red algae plants? A critical evaluation of three key molecular data sets, *Journal of Molecular Evolution*, **52**, 527–539.
Stiller, J. W. and Hall, B. D. (2002). Evolution of the RNA polymerase II C-terminal domain, *Proceedings of the National Academy of Science USA*, **99**, 6091–6096.
Takishita, K., Inagaki, Y., Tsuchiya, M., Sakaguchi, M. and Maruyama, T. (2005). A close relationship between Cercozoa and Foraminifera supported by phylogenetic analyses based on combined amino acid sequences of three cytoskeletal proteins (actin, α-tubulin, and β-tubulin), *Gene*, **362**, 153–160.
Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22**, 4673–4680.
Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Molecular Biology and Evolution*, **18**, 691–699.
Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data, *Journal of Molecular Evolution*, **42**, 587–596.
Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood, *Computer Applications in the Biosciences*, **13**, 555–556.

164 Proceedings of the Institute of Statistical Mathematics Vol. 56, No. 1, 145–164 (2008)

Phylogenetic Inference Based on Combined Analysis of Multiple Genes —Illustrative Data Analysis of Higher-order Phylogeny of Eukaryota—

Tetsuo Hashimoto[1,2], Nobuko Arisue[3], Miako Sakaguchi[1] and Yuji Inagaki[1,2]

[1]Laboratory of Microbial Molecular Evolution, Institute of Biological Sciences, University of Tsukuba
[2]Division of Global Environment and Biological Sciences, Center for Computational Sciences, University of Tsukuba
[3]Department of Molecular Protozoology, Research Institute for Microbial Diseases, Osaka University

A maximum likelihood method for phylogenetic inference based on combined analysis of multiple genes is briefly introduced and applied to data analysis of higher-order eukaryotic phylogeny. Three models of branch length estimation are considered assuming that all genes (or partitions for the full data set)have the same branch length (concatenate model), each gene (partition) has a separate set of branch lengths (separate model), and branch lengths are proportional among genes (partitions) (proportional model). Fifty-three ribosomal protein genes from 29 eukaryotic species were used for the analysis. The data set consisted of 5, 842 amino acid positions. Six different models with different methods for estimating branch lengths and for partitioning the data set were compared by Akaike Information Criterion (AIC). Comparison of the AIC values for the maximum likelihood tree demonstrated that a separate model with a partition between large- and small-subunit ribosomal proteins showed the lowest AIC value, while a separate model with a partition among individual genes had the highest AIC value, suggesting that the former model best approximated the data set and the latter model was over-parameterized. It was suggested also that the tempo and mode of sequence evolution was relatively uniform across different ribosomal protein genes. Since no incongruence was observed among the six models for the selection of alternative trees, the present analysis was considered to be robust.

Key words: Phylogenetic inference, maximum likelihood method, combined analysis of multiple genes, eukaryotes, higher-order phylogeny, ribosomal protein.