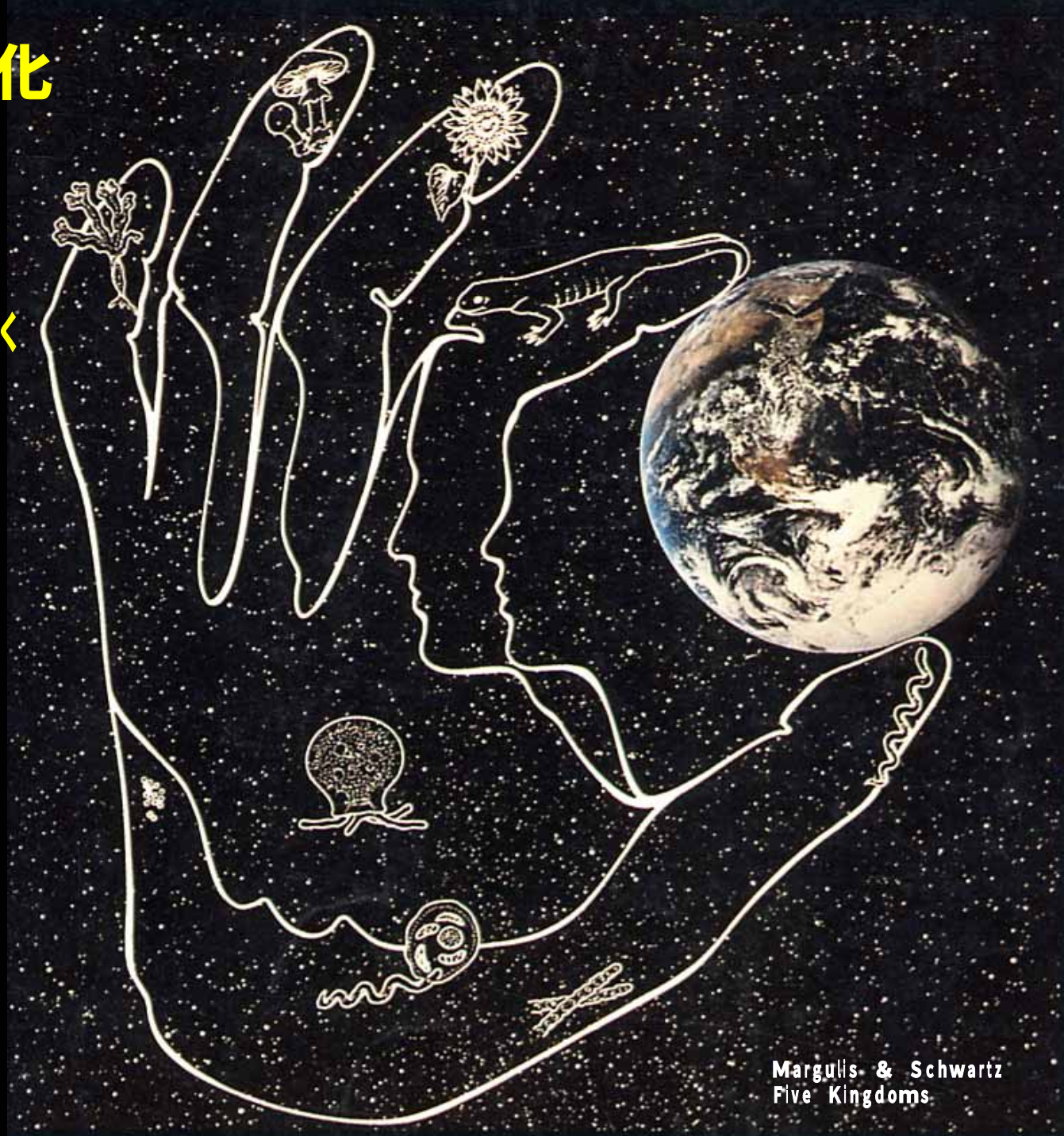
生物分子系統進化プロジェクト

複数遺伝子の
分子系統樹解析に基づく
真核生物の
初期進化の解明



２００５年2月１６日
計算科学シンポジウム

生命環境科学研究科
橋本　哲男

遺伝子配列データ　(アミノ酸アライメント)

```
                                                                      60
生物種1   mvlspadktnvkaawgkvgahageygaealermflsfpttktyfphf-dlshgsaqvkgh
生物種2   mvlspadktnikstwdkigghagdyggealdrtfqsfpttktyfphf-dlspgsaqvkah
生物種3   mvlsaadknnvkgiftkiaghaeeygaetlermfttypptktyfphf-dlshgsaqikgh
生物種4   mslsdkdkaavkglwakispkaddigaealgrmltvypqtktyfahwadlspgsgpvkkh
          * **  **   *    *    *   * * * *     * ***** *  *** **   * *
                                                                     120
生物種1   gkkvadaltnavahvddmpnalsalsdlhahklrvdpvnfkllshcllvtlaahlpaeft
生物種2   gkkvadalttavahlddlpgalsalsdlhayklrvdpvnfkllshcllvtlachhpteft
生物種3   gkkvvaalieaanhiddiagtlsklsdlhahklrvdpvnfkllgqcflvvvaihhpaalt
生物種4   gkvimgavgdavskiddlvgglaalselhafklrvdpanfkilahnvivvigmlypgdfp
          **    *        **    *  ** *** ****** *** *     *      *

生物種1   pavhasldkflasvstvltskyr
生物種2   pavhasldkfftavstvltskyr
生物種3   pevhasldkflcavgtvltakyr
生物種4   pevhmsvdkffqnlalalsekyr
          * ** * ***       *  ***
```

*4生物種で同一のアミノ酸となってる座位

アミノ酸配列は動物種間で良く保存されてる！
〃　　進化的に　　〃

最尤法 (Maximum likelihood method)

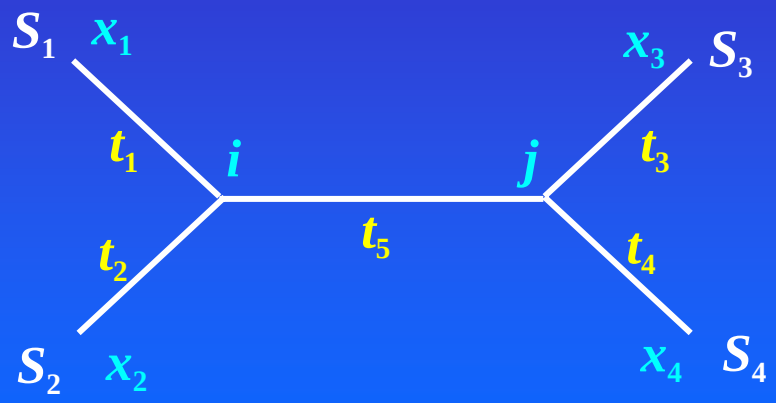
生物種が4つの場合を想定

配列データ

$X = (X_{ij}) \quad (i=1,\ldots 4;\ j=1,\ldots,n)$

h番目の座位:

$X_h=(X_{1h}, X_{2h}, X_{3h}, X_{4h})$



$S_1,\ldots, S_4$: 現存の生物種

$x_1,\ldots, x_4$: 現存の生物種のh番目の座位のデータ

i, j : 共通祖先種のh番目の座位のデータ

$t_1,\ldots, t_5$: 枝の長さ $\quad \theta_1 = (t_1,\ldots, t_5)$; $\quad \theta_2$

仮定1. X_t は置換確率行列が$P_{ij}(t)$で表される連続時間定常マルコフ過程であり、
k から lへの置換確率は

$$P_{kl}(t) = P\{X_{t+s} = l \mid X_s = k\}$$

で、各枝での進化 (置換) は独立に起こる

仮定2. それぞれの座位 X_h $(h = 1,\ldots,n)$ は独立に同一の確率法則に従って進化している

(1) X_h は定常マルコフ過程、各枝での進化は独立（仮定1）

Chapmann-Kolmogorovの式により、

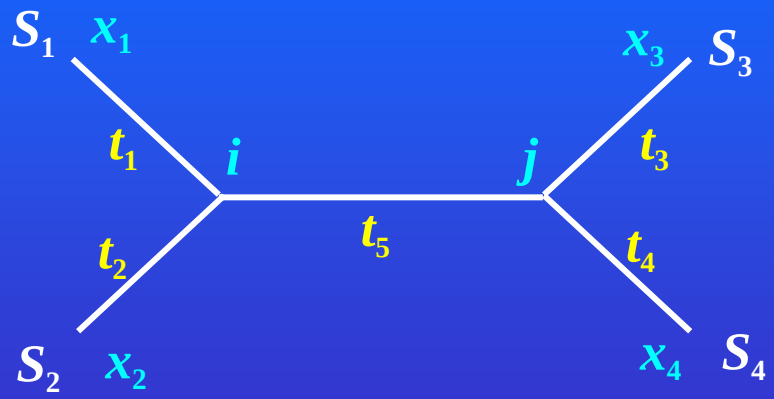
$f(x1, x2, x3, x4|\theta)$ ← 与えられた系統樹のトポロジーと置換確率モデルのもとでh番目の座位のデータが実現する確率

$$= \sum_i \sum_j P\{X_0 = i\} P\{X_{t_1} = x_1, X_{t_2} = x_2, X_{t_5} = j \mid X_0 = i\}$$

$$\times P\{X_{t_5+t_3} = x_3, X_{t_5+t_4} = x_4 \mid X_{t_1} = x_1, X_{t_2} = x_2, X_{t_5} = j, X_0 = i\}$$

$$= \sum_i \left\{ \pi_i P_{ix_1}(t_1) P_{ix_2}(t_2) \sum_j P_{ij}(t_5) P_{jx_3}(t_3) P_{jx_4}(t_4) \right\}$$

π_i は i の組成値



$S_1, \ldots, S_4$ ：現存の生物種

$x_1, \ldots, x_4$ ：現存の生物種のh番目の座位のデータ

i, j ：共通祖先種のh番目の座位のデータ

$t_1, \ldots, t_5$ ：枝の長さ

(2) X_hは独立に同一の確率法則に従って進化している (仮定2)

配列データXが与えられたときの対数尤度は

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

$\theta = (\theta^{(1)}, \theta^{(2)})$

$\theta^{(1)} = (t_1, \ldots, t_5)$ ⟵ 枝の長さ

対数尤度を最大にするθの推定値 $\hat{\theta}$は以下の式を満たすパラメータ空間 Θの中の1点である

$$l(\hat{\theta} \mid X) = \max\{l(\theta \mid X) : \theta \in \Theta\}$$ ⟵ lを最大にするようなθを推定

異なるモデル（系統樹のトポロジー）において $l(\hat{\theta}|X)$ （最大）対数尤度の値を比較

↓

最大の（最大）対数尤度を示すモデル(系統樹のトポロジー)を選択

⟶ **最尤系統樹** (ML-tree)

モデル

- 系統樹のトポロジー
- 置換確率 $Pij\,(t)$

塩基置換
- Poisoonモデル
- Kimura 2 parameter model
- HKY85 model
- TN93 model
- ・・・・

アミノ酸置換
- Poisson model
- Dayhoff model
- JTT model
- ・・・

- 座位間の進化速度の不均質性

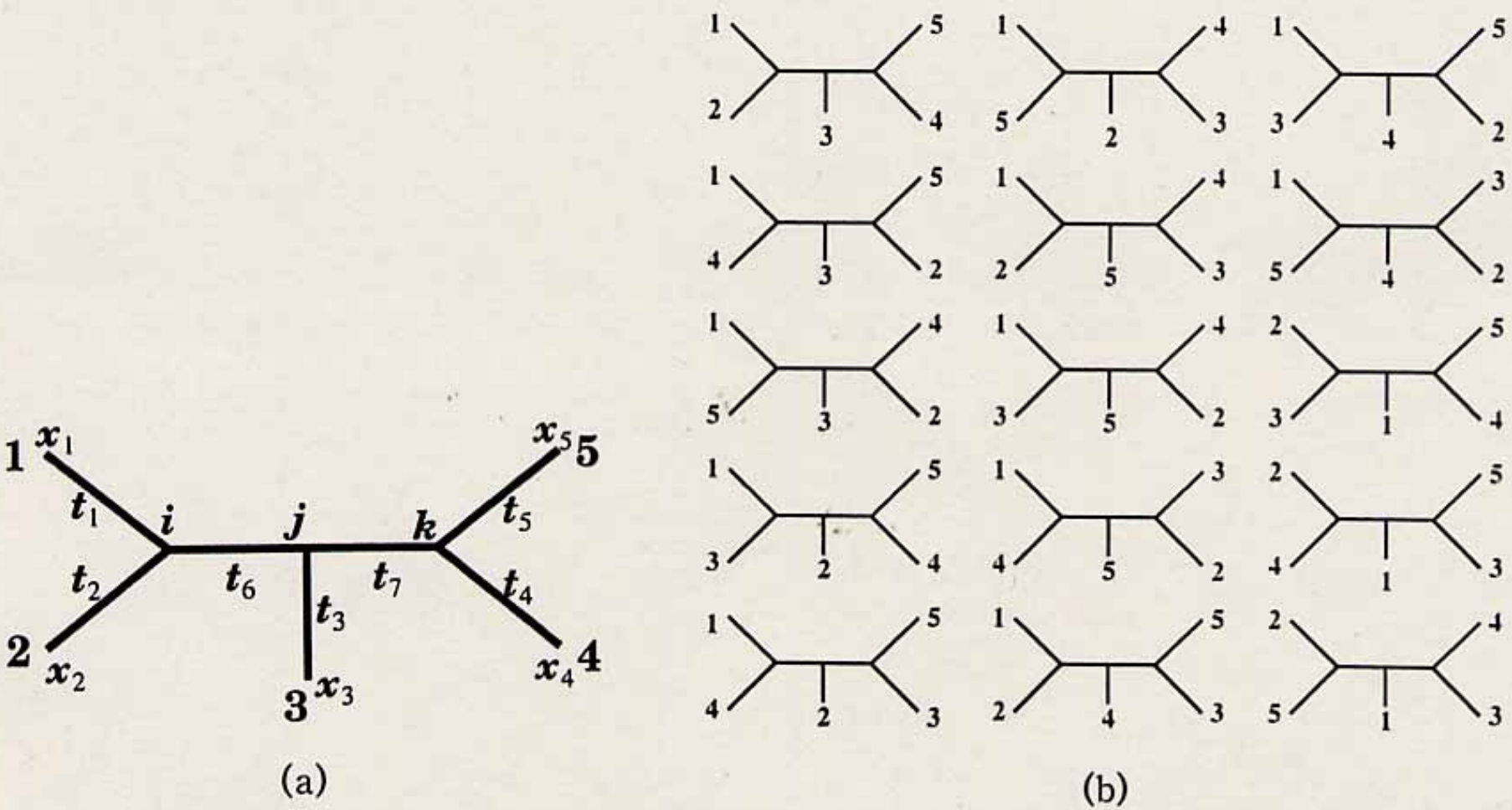


図 7. (a) 5 つの系統に対する根なし系統樹. 1〜5: 現存生物種, x_1〜x_5: 各座位における現存生物種のアミノ酸の観測値, i, j, k: 祖先生物種のアミノ酸の状態, t_1〜t_7: 枝の長さ. (b) 5 つの系統に対する 15 通りの系統樹のトポロジー.

*n*個のOTUからなる系統樹における結節，枝，および，近隣の数

OTU	内部結節	外部結節	総結節	内部枝	外部枝	総枝	近隣
n	*n* - 2	*n*	2*n* - 2	*n* - 3	*n*	2*n* - 3	*n* - 3

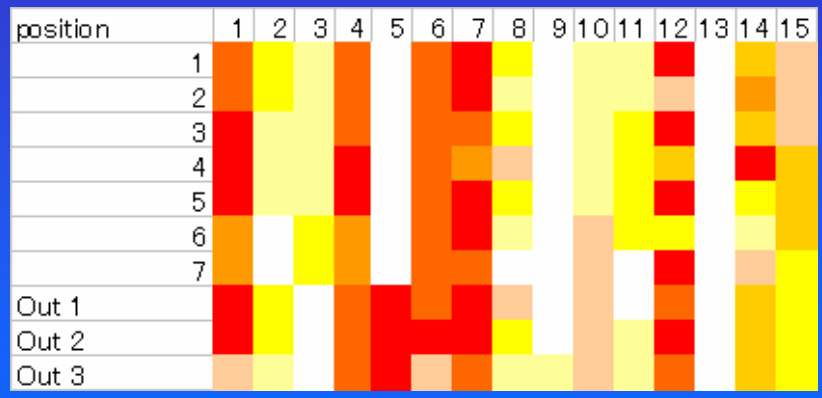
OTUの数に対するすべての可能な二分岐の樹形の数

OTUの数	樹形の数	
	無根系統樹	有根系統樹
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

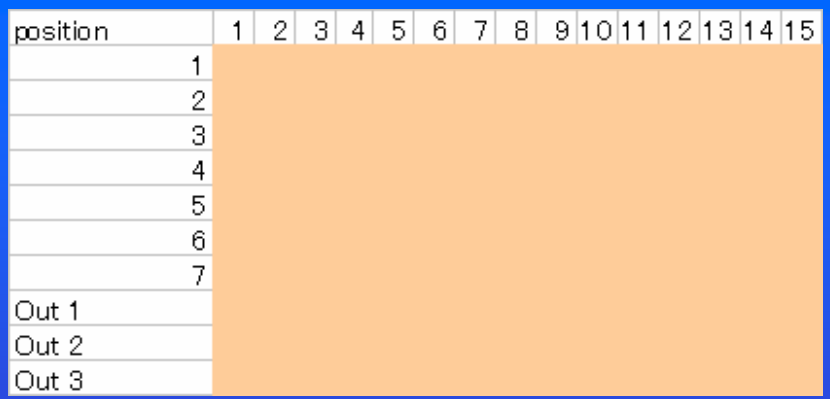
n $(2n-5)! / 2^{n-3} (n-3)!$

座位間の進化速度の不均質性の考慮： RASモデルによる解析

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	G	P	G	D	H	I	E	V	E	A	R	F	W	A	G
2	A	P	G	D	H	V	Q	V	E	S	R	Y	W	V	G
3	K	P	P	N	H	I	D	G	E	A	R	L	W	Y	G
4	G	P	A	Q	H	L	H	V	E	A	K	F	W	V	G
5	G	P	G	K	H	V	K	V	E	A	R	V	W	A	A
6	R	P	S	N	H	V	R	V	E	A	R	I	W	V	V
7	G	S	C	N	H	L	E	G	E	A	K	F	W	V	V
Out 1	R	G	C	D	R	M	N	V	E	S	K	Y	W	T	S
Out 2	T	A	C	D	K	M	D	V	E	S	K	K	W	T	S
Out 3	R	S	C	E	P	L	D	I	Q	A	K	H	W	T	S

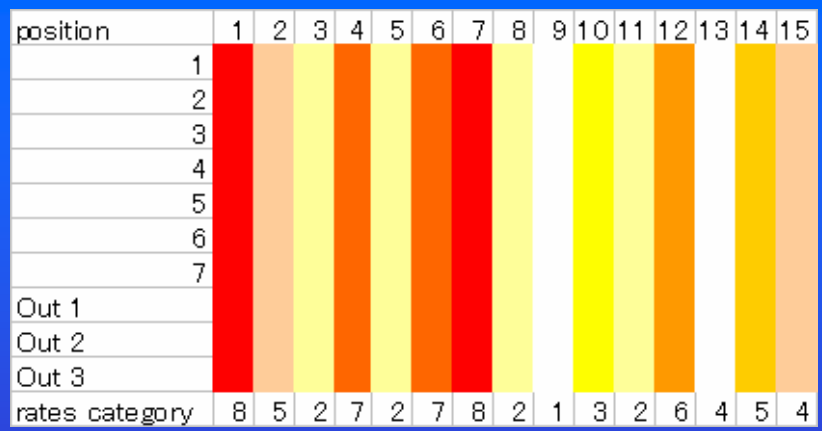


Covarion モデル



均質モデル (Homogeneous model)

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



RASモデル (Rate-Across-Site model)

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

⇒ Γ 分布を用いて不均質性をモデル化

Assessing the reliability of individual branches

Bootstrap analysis

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots\cdots & X_{1q} & \cdots\cdots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \cdots\cdots & X_{2q} & \cdots\cdots & X_{2n} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ X_{p1} & X_{p2} & X_{p3} & \cdots\cdots & X_{pq} & \cdots\cdots & X_{pn} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ X_{m1} & X_{m2} & X_{m3} & \cdots\cdots & X_{mq} & \cdots\cdots & X_{mn} \end{bmatrix}$$

$$X_1, \quad X_2, \quad X_3, \quad \cdots\cdots \quad ,X_q, \quad \cdots\cdots \quad ,X_n$$

$$\downarrow$$

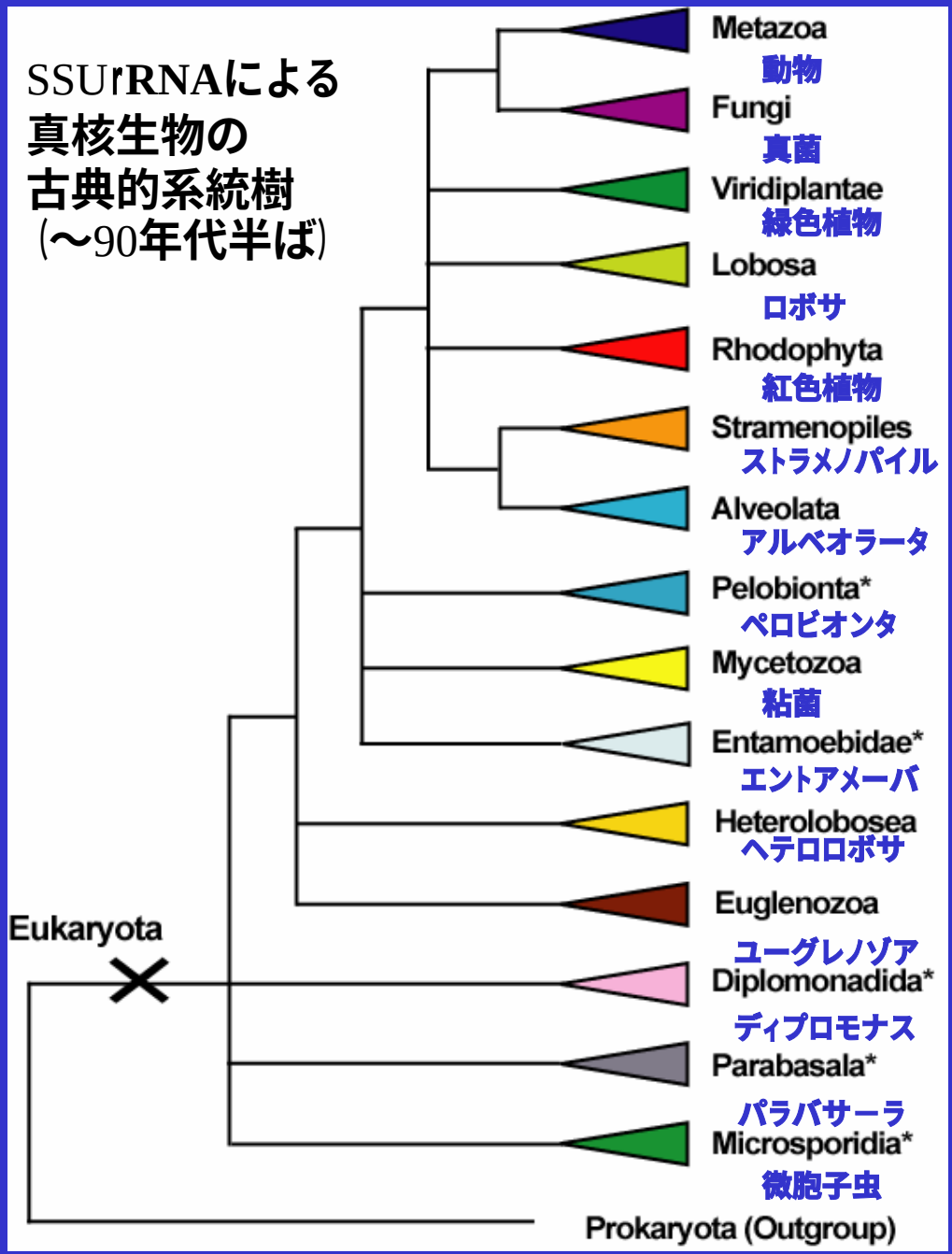
$$X_b^* = \begin{bmatrix} X_1^*, & X_2^*, & X_3^*, & \cdots\cdots & ,X_q^*, & \cdots\cdots & ,X_n^* \end{bmatrix}$$

bootstrap sample

Phylogenetic inference based on bootstrap pseudosamples
（N = 1,000 , 10,000）

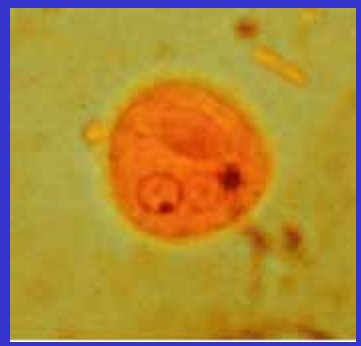
Calculation of the proportion of bootstrap pseudosamples that support a given imternal branch on a tree

SSUrRNAによる
真核生物の
古典的系統樹
(～90年代半ば)



***ミトコンドリアをもたない真核生物**
原始的な細胞構造、
70S型リボソーム　他
⇒アーケゾア仮説

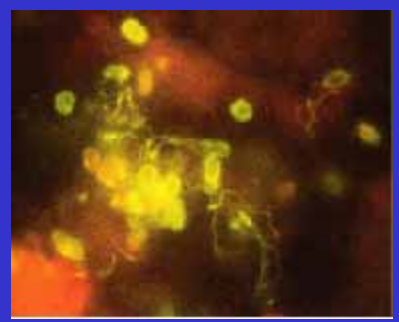
(Cavalier-Smith 1983)



Entamoeba histolytica
(Entamoebidae)

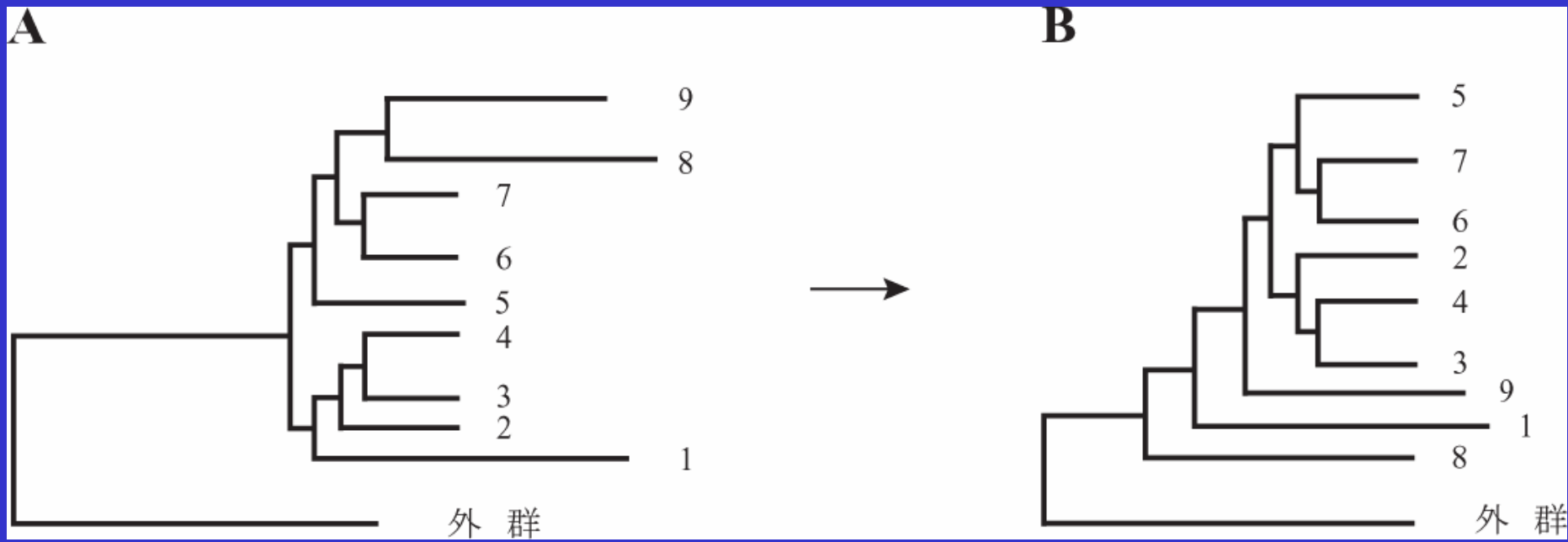
Giardia lamblia
(Diplomonadida)

Trichomonas vaginalis
(Parabasala)

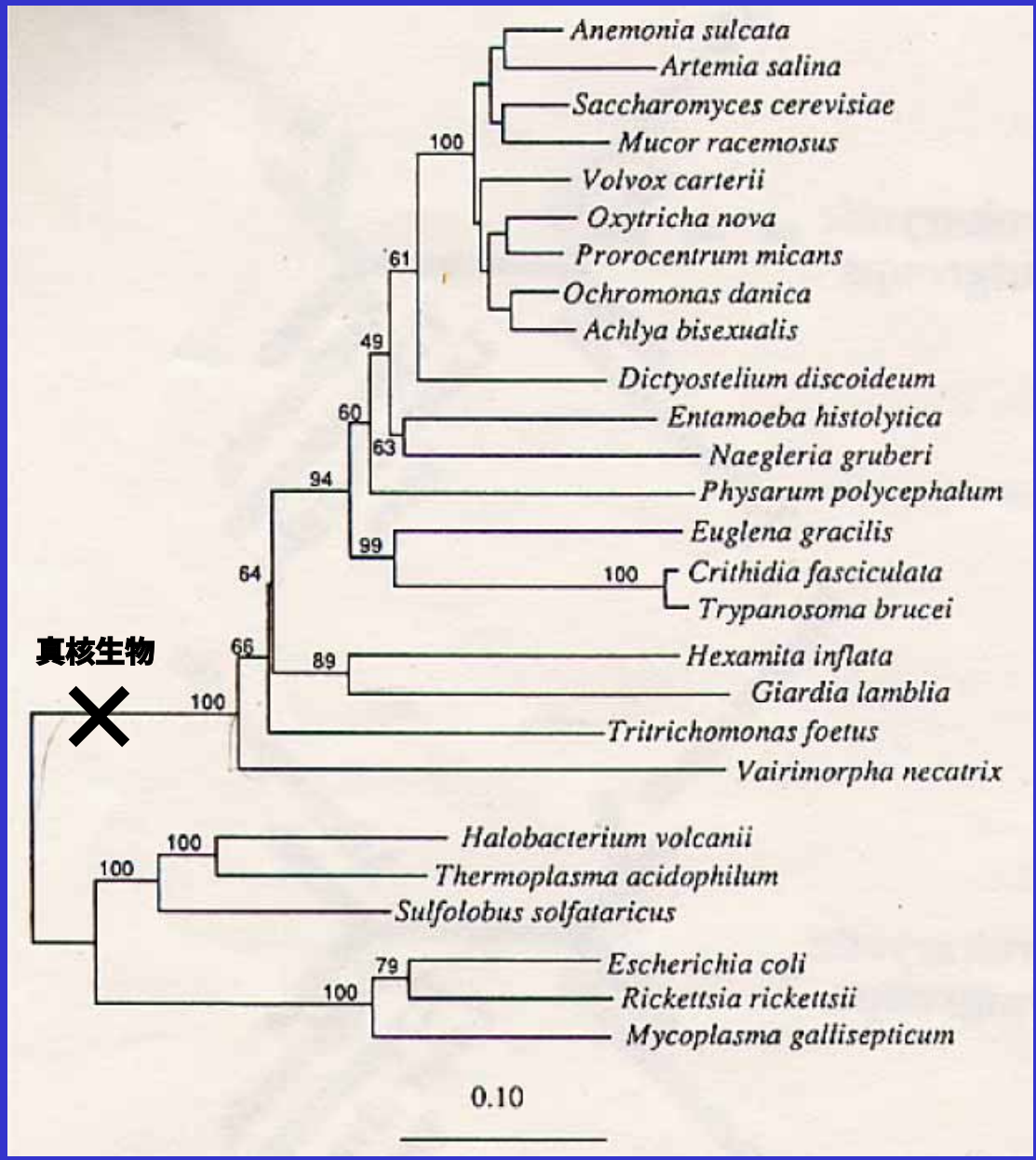


Encephalitozoon hellem
(Microsporidia)

Long Branch Attraction (LBA) アーテファクト



SSUrRNAによる真核生物全体の系統樹 Leipe et al. 1993



←Entamoebidae (エントアメーバ)

←Diplomonadia (ディプロモナス)

←Parabasala (パラバサーラ)

←Microsopridia (微胞子虫)

真核生物の初期進化研究をめぐる90年代後半の情勢

・ミトコンドリアをもたない真核生物のいずれのグループにもミトコンドリア関連遺伝子が存在

→ミトコンドリアの二次的喪失の証拠

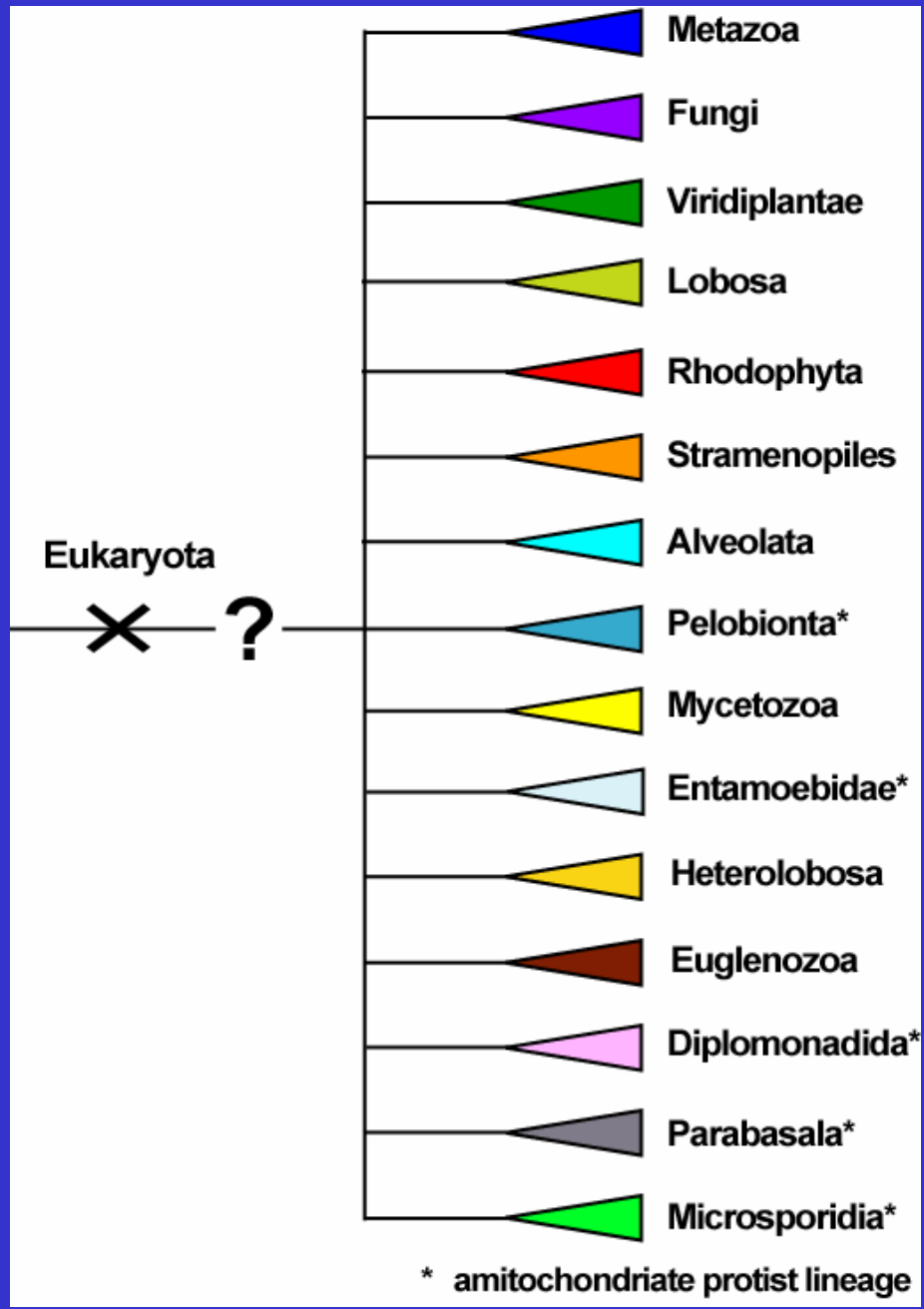
⇒ミトコンドリアをもたないことは必ずしも原始的真核細胞たることを意味しない

・LBAのアーテファクトは極めて深刻

→リボソームRNAの系統樹やそれを支持するEF-1αの系統樹はLBAによるアーテファクトだ

・真核生物の大きなグループの関係についての解析結果は、異なる遺伝子間で大きく異なる

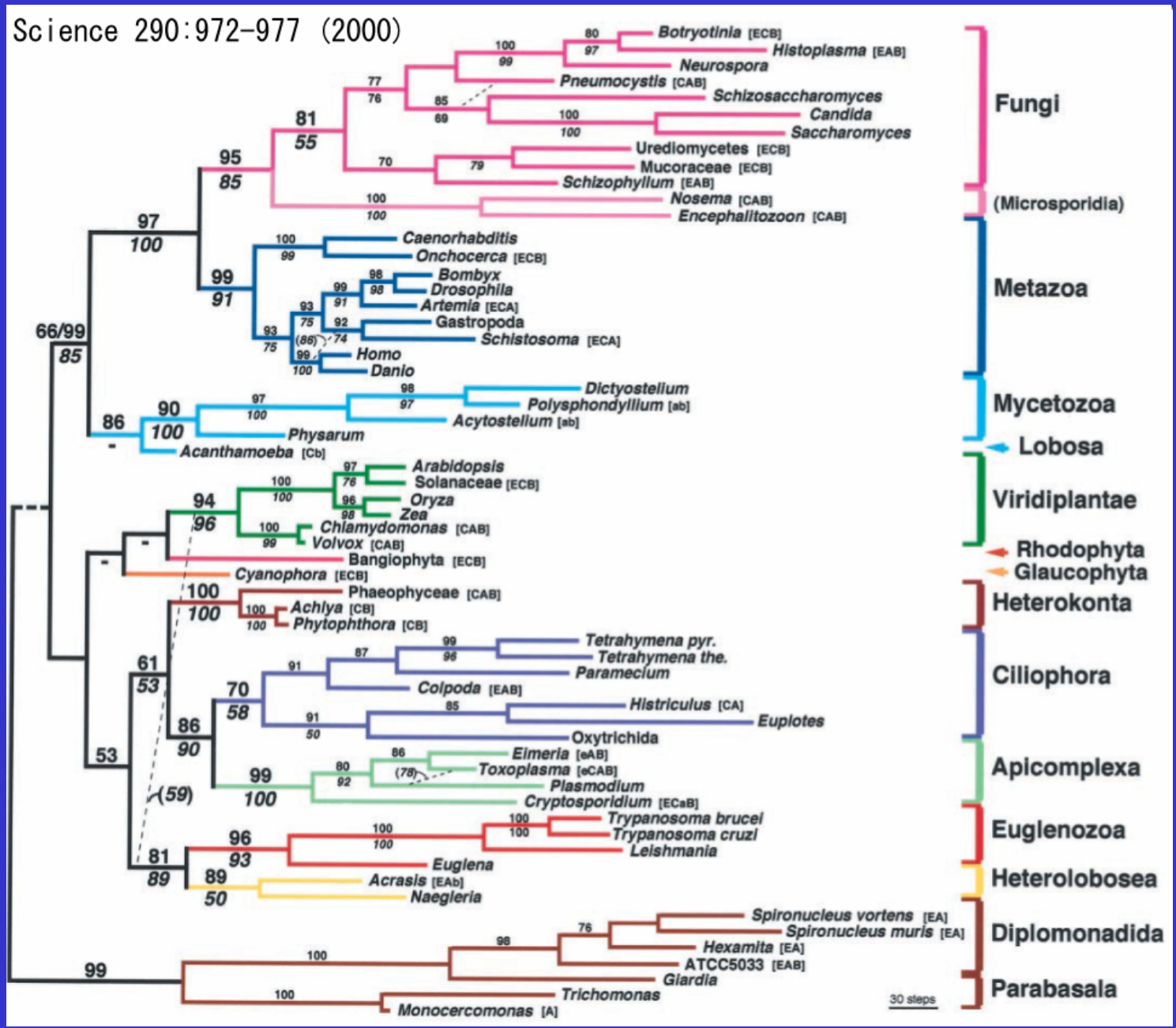
：統計的誤差範囲を超えて有意に異なる場合も多い



真核生物の無根系統樹　(EF-1α、α／βtubulin、actinの結合データ)

(Baldaufらの系統樹)

Science (2000) 290:972-977

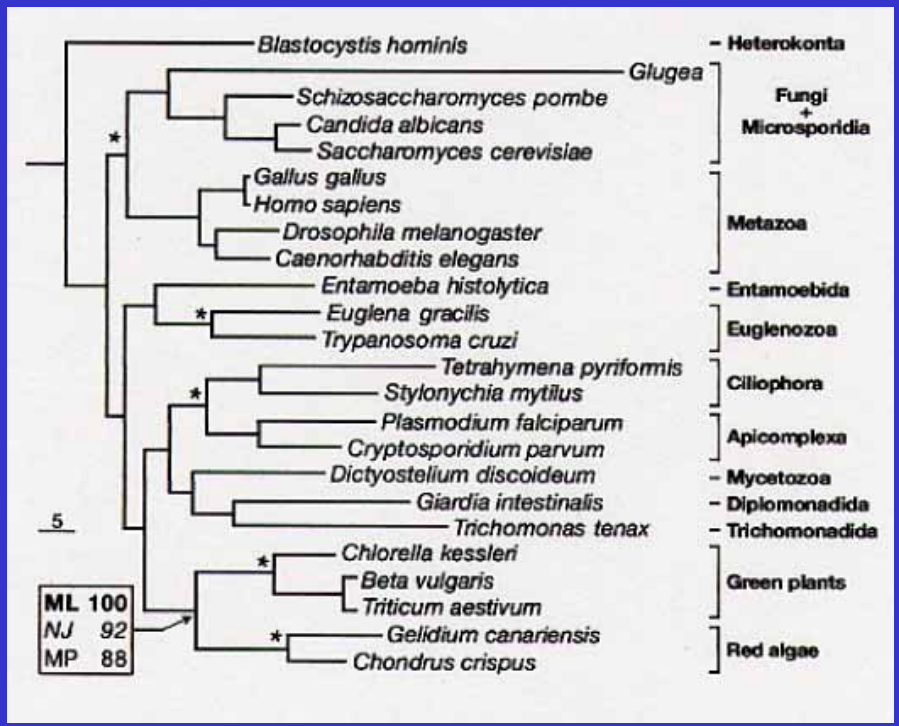


緑色植物と紅色植物は近縁

Viridiplantae　　Rhodophyta

Moreira et al. 2000 Nature 405:69-72

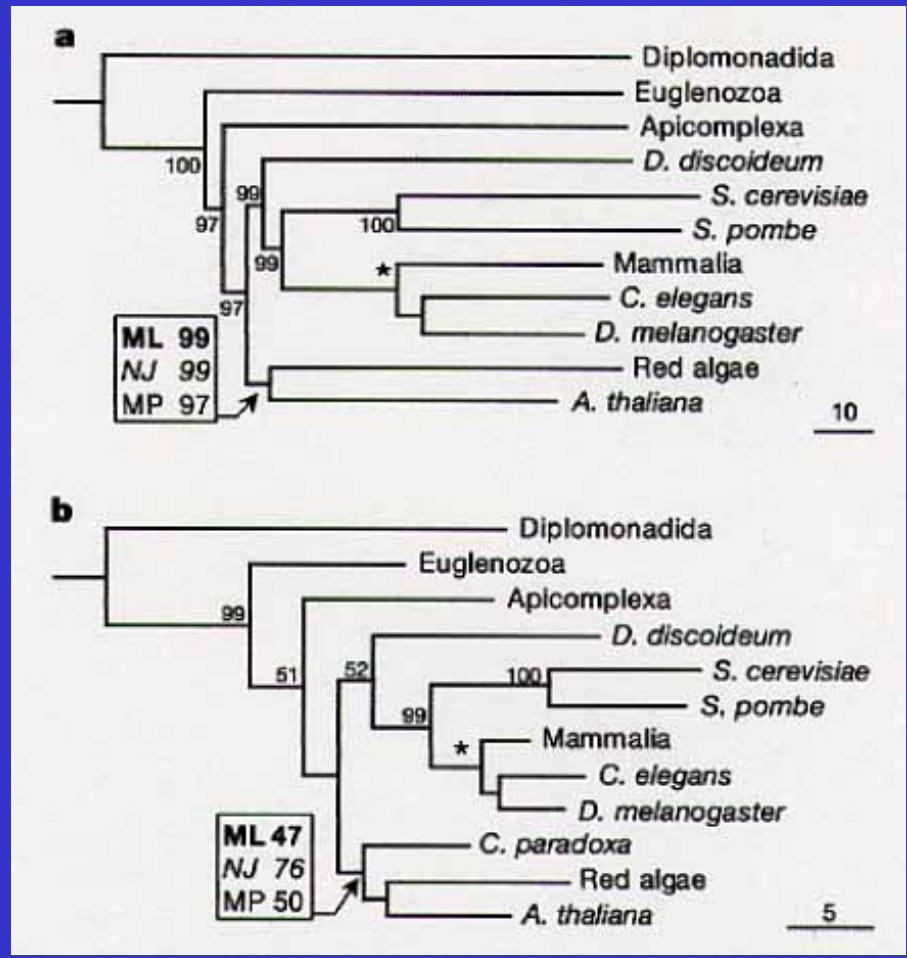
ペプチド鎖伸長因子EF-2による解析



結合データによる解析

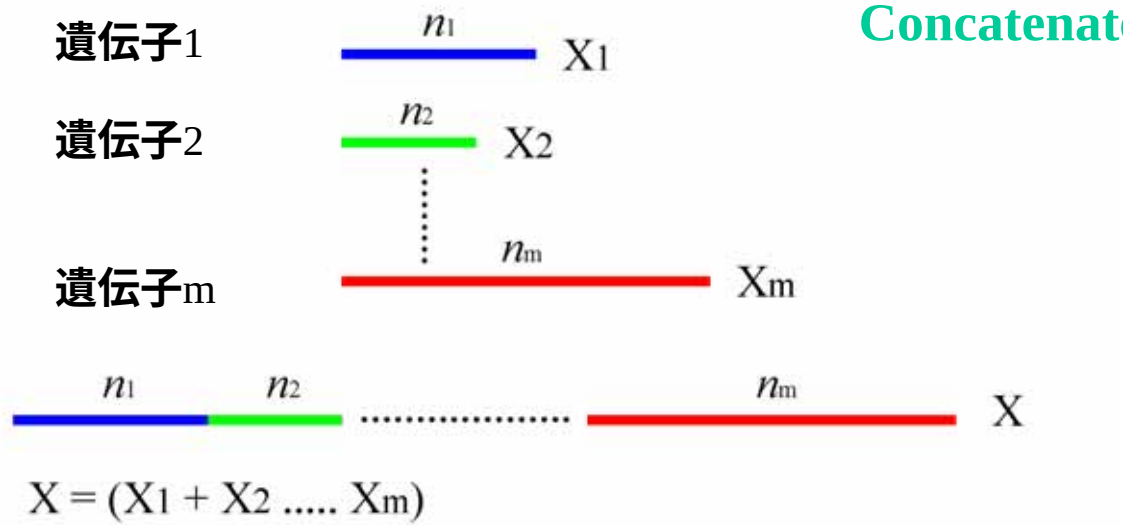
a) 13遺伝子5171アミノ酸座位

b) 6遺伝子1938アミノ酸座位



複数の遺伝子による系統樹の推測 (連結した配列データを用いた解析)

Concatenate model



系統樹 i の対数尤度の推定値

$$l_{(i)}(\hat{\theta}_{(i)}|\mathrm{X}) = \max_{\theta_{(i)}}\{l_{(i)}(\theta_{(i)}|\mathrm{X})\} = \max_{\theta_{(i)}}\left\{\sum_{h=1}^{n1+n2+\ldots nm}\log f_{(i)}(\mathrm{X_n}|\theta_{(i)})\right\}$$

最尤系統樹の対数尤度の推定値

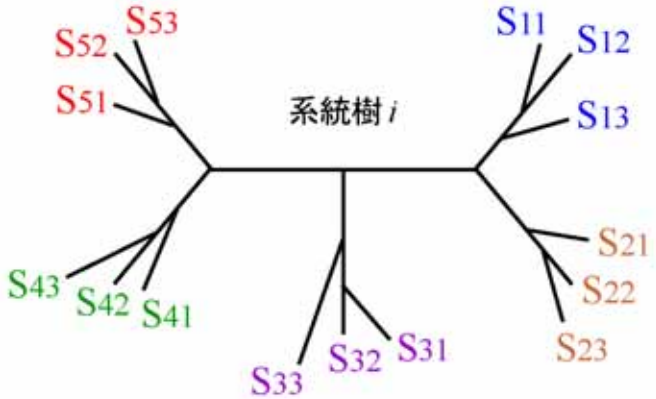
$$\max_{i}\{l_{(i)}(\hat{\theta}_{(i)}|\mathrm{X})\} = \max_{i}\left\{\sum_{h=1}^{n1+n2+\ldots nm}\log f_{(i)}(\mathrm{X}h|\hat{\theta}_{(i)})\right\} \quad (i=1,\ldots,l)$$

それぞれの遺伝子の配列を連結して作成したデータセットに対して1つのパラメータを推定する

複数の遺伝子による系統樹の推測 (対数尤度の連結による解析)

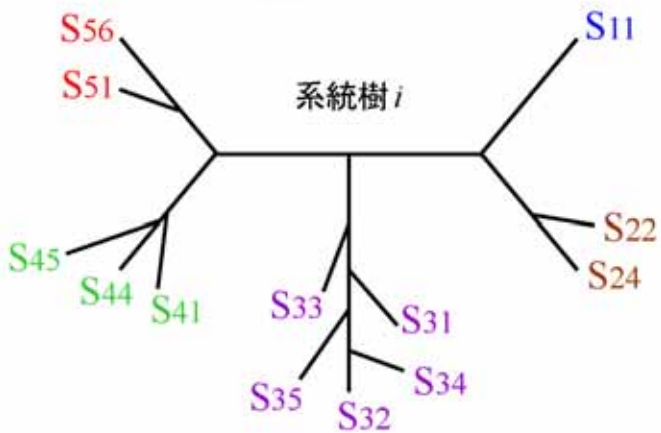
Separate model

遺伝子 1
5 系統, 15 生物種
(5グループ)



$l_{1(i)}(\hat{\theta}_{1(i)}|X_1)$

遺伝子 2
5 系統, 13 生物種



$l_{1(i)}(\hat{\theta}_{1(i)}|X_2)$

遺伝子 m
5 系統, ～生物種

$l_{m(i)}(\hat{\theta}_{m(i)}|X_m)$

最尤系統樹の対数尤度の推定値

それぞれの遺伝子について別のパラメータを推定する

$$\max_i \left\{ \sum_{k=1}^{m} l_{k(i)}(\hat{\theta}_{k(i)}|X_k) \right\} = \max_i \left[\sum_{m}^{k=1} \left\{ \sum_{nk}^{h=1} \log f_{k(i)}(X_{kh}|\hat{\theta}_{k(i)}) \right\} \right] \quad (i=1,\ldots,15)$$

Subtree (グループ)内部の系統関係をあらかじめ仮定

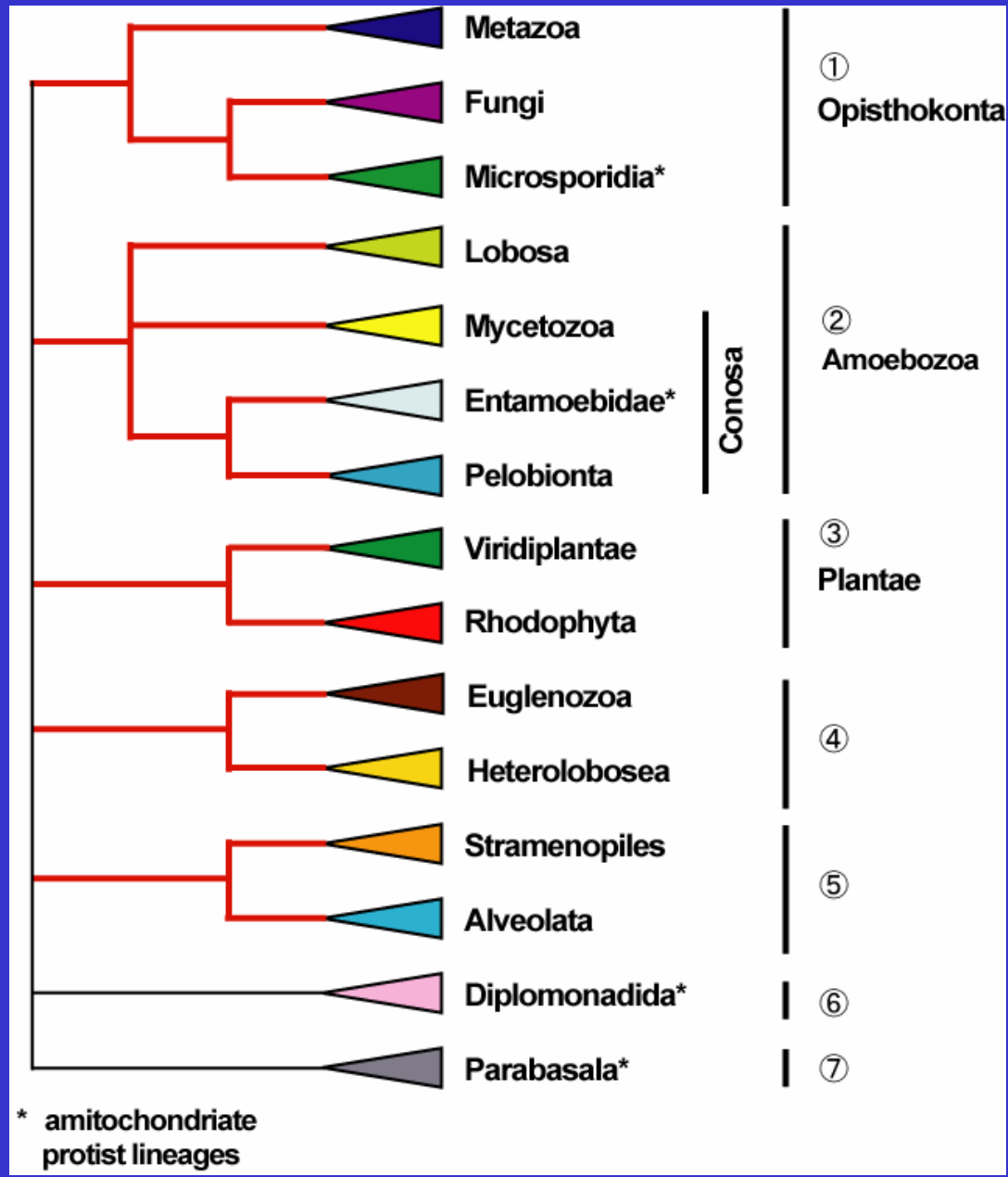
：たとえば Metazoa について

EF-1α	((ヒト,カエル),(ハエ,線虫))
EF-2	(ヒト,ハエ,線虫)
IleRS	(ヒト,ハエ,線虫)
ValRS	(((ヒト,マウス),フグ),ハエ,線虫)
RpS14	(ヒト,ザリガニ,ヒドラ,ハエ,線虫)
RpS15a	(ヒト,ハエ,線虫,ウニ)
RpL5	((((ヒト,ニワトリ),カエル),ホヤ),(ハエ,カイコ,蚊),線虫)
RpL8	((ヒト,カエル),((蚊,ハエ),線虫))
RpL10a	(ヒト,ハエ,線虫)
RPB1	((ヒト,ハエ,線虫), *Monosiga*)
・	・
・	・
・	・
・	・
・	・

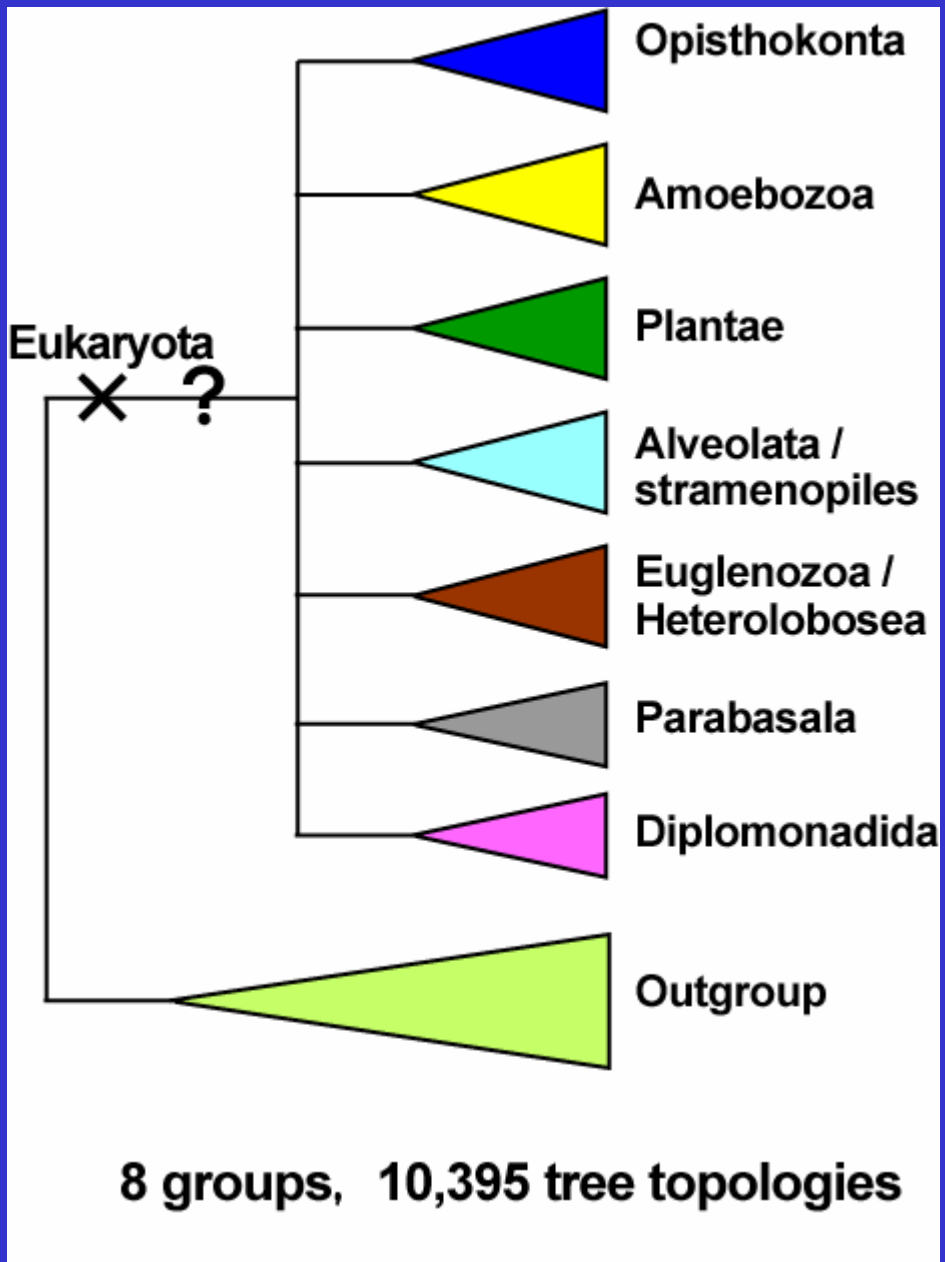
複数遺伝子の配列情報に基づく真核生物のコンセンサス系統樹 2002

Moreira et al. 2000
Baldauf et al. 2000
Bapteste et al. 2002
Arisue et al. 2002

7グループ間の系統的位置関係と系統樹の根もとは不明 ⇒



真核生物7つの大きな単系統群の解析と系統樹のRooting

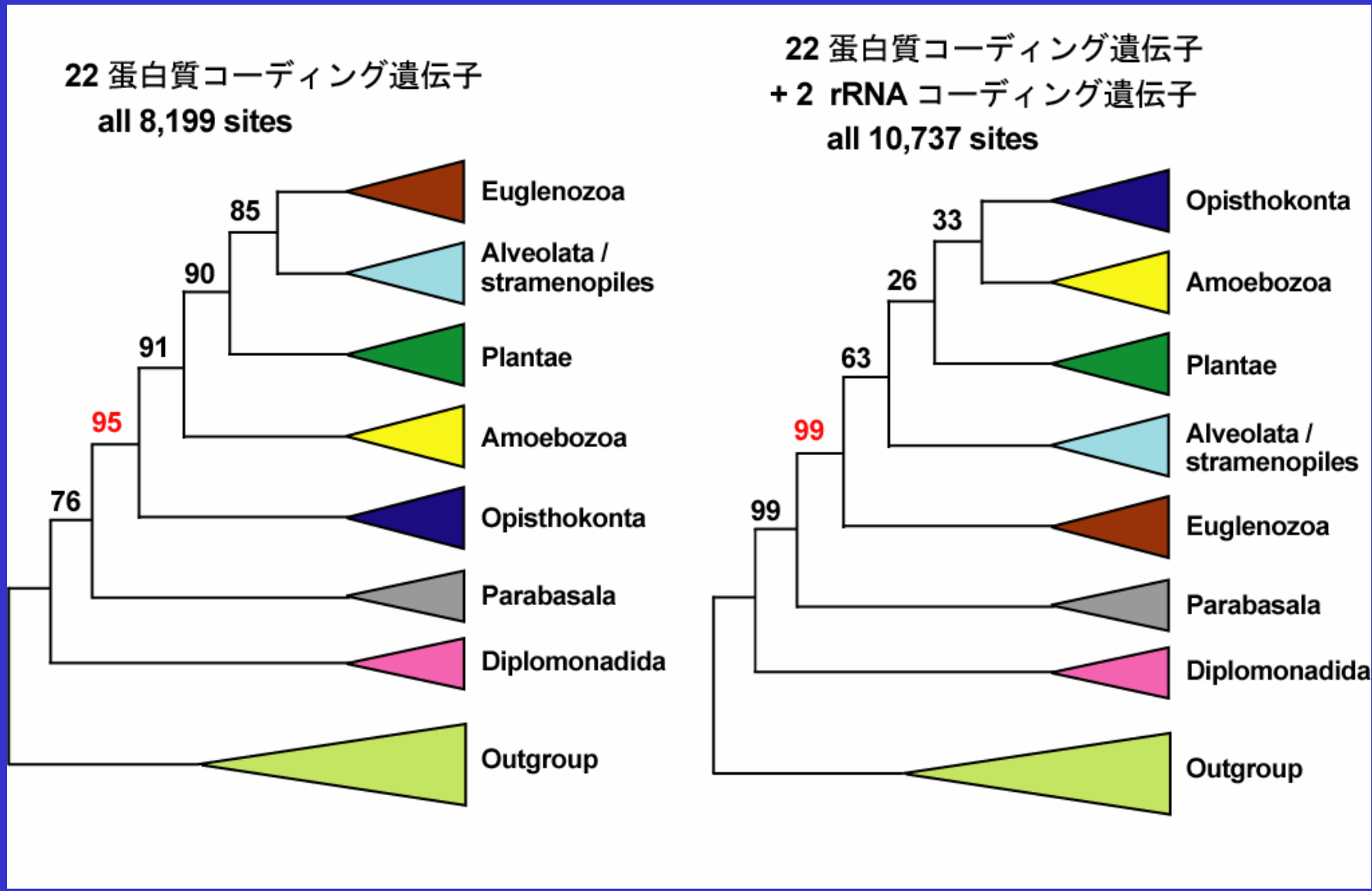


結合データ解析に用いた24遺伝子

Translation:	EF-1α, EF-2, RpS14, RpS15a, RpL5, RpL8, RpL10a, IleRS, ValRS
Transcription:	RNA polymerase II (Rpb1)
Chaperon:	CPN60, HSP70c, HSP70mit, HSP70er, HSP90c, CCTα, CCTγ, CCTδ, CCTζ
Cytoskeleton:	Actin, α-tubulin, β-tubulin
rRNA:	SSUrRNA, LSUrRNA

Outgroup:
原核生物 (*Archaea, Bacteria*)
もしくは paralogous **遺伝子**

全座位 (all) を用いた均質モデルによる解析での the best tree



Parabasala と Diplomonadida の根もと近くへの位置づけに対する強い支持
→ LBA の効果である可能性！ (とくにrRNAを入れた場合)

rRNAを除外し進化速度の大きい座位を除外して RASモデルで解析した際の最尤系統樹 (22 protein genes)

Evolutionary rate category for each site

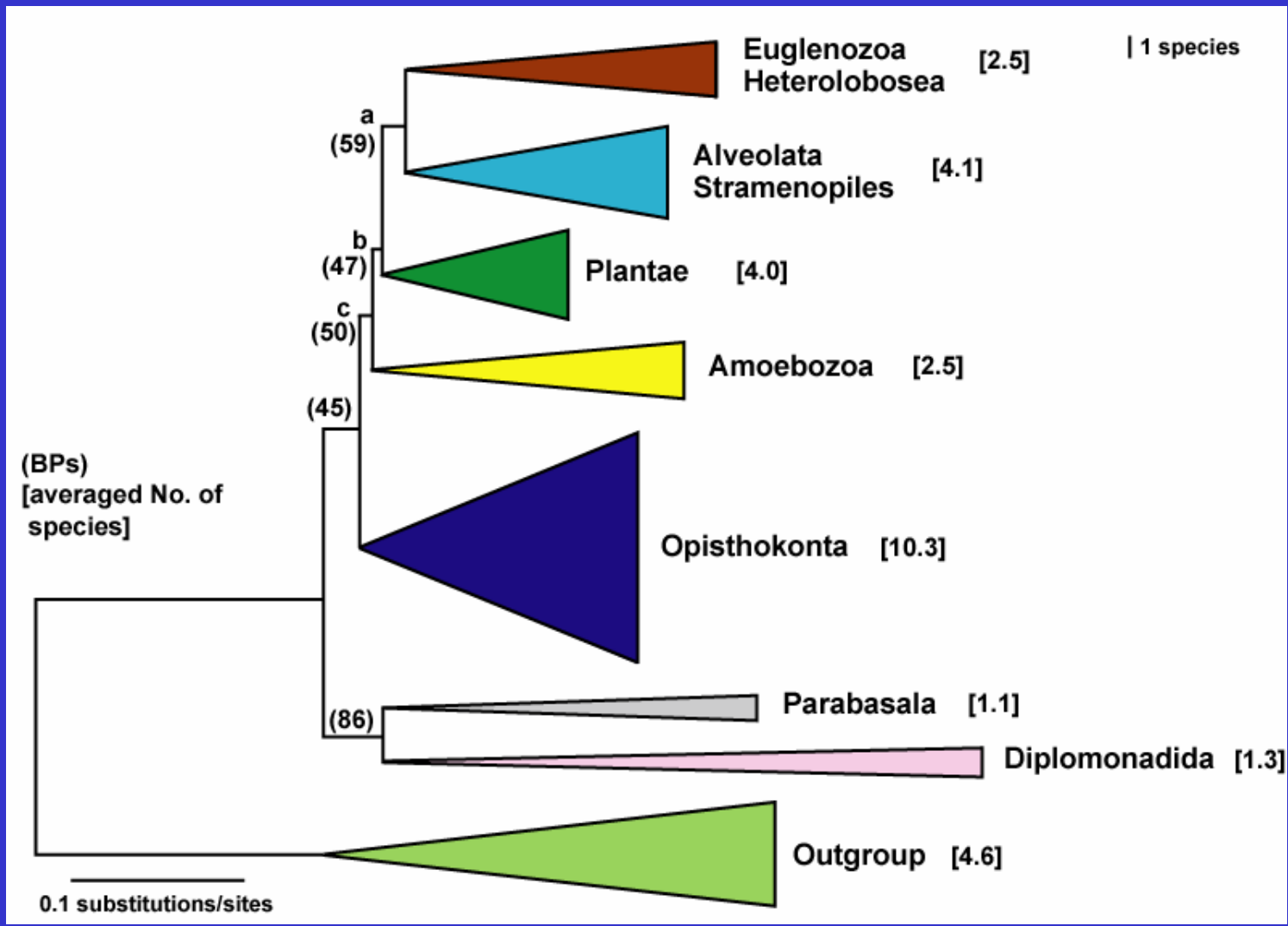
r1
r2
r3
r4
r5
r6
r7
r8

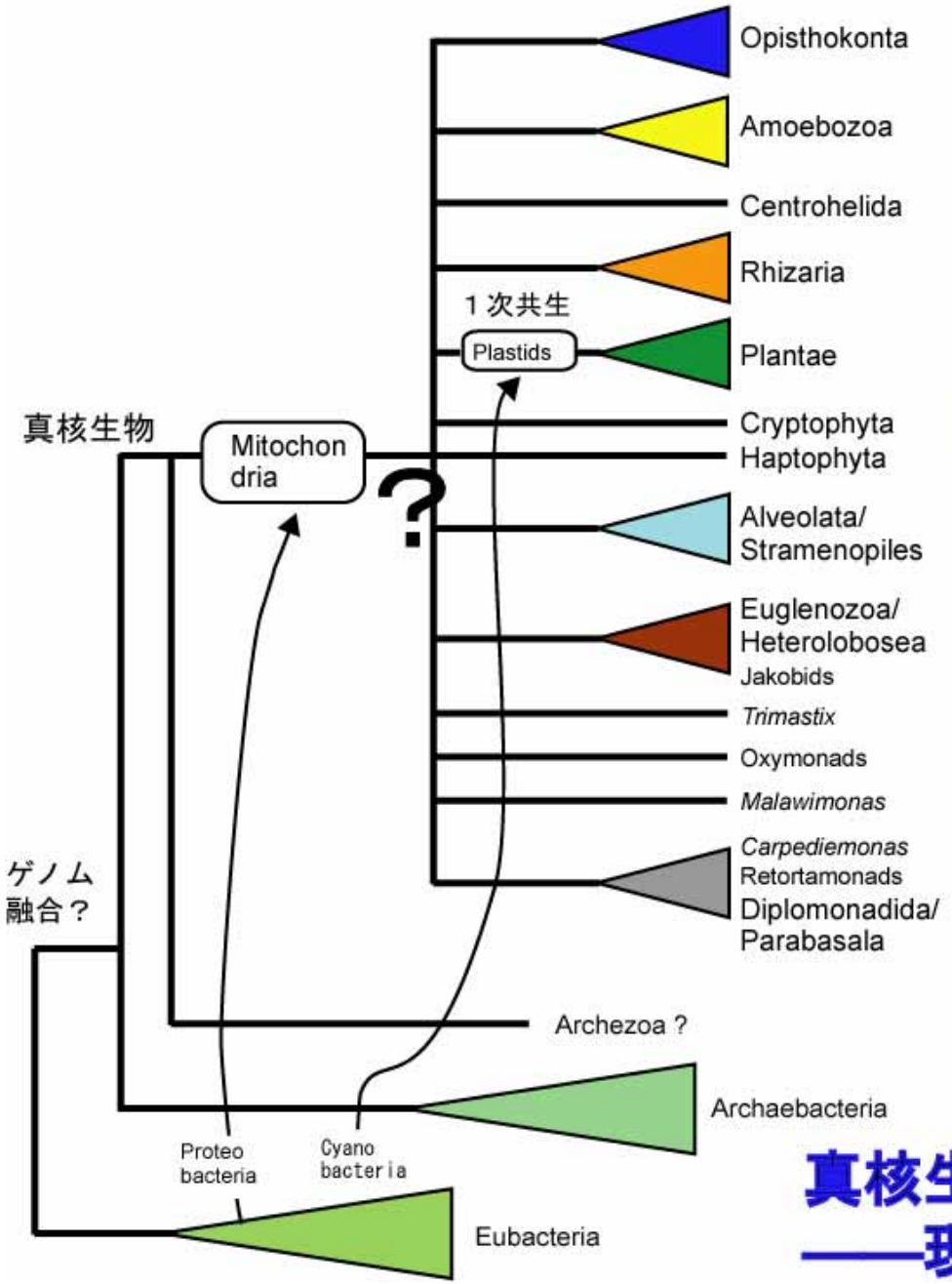
slow

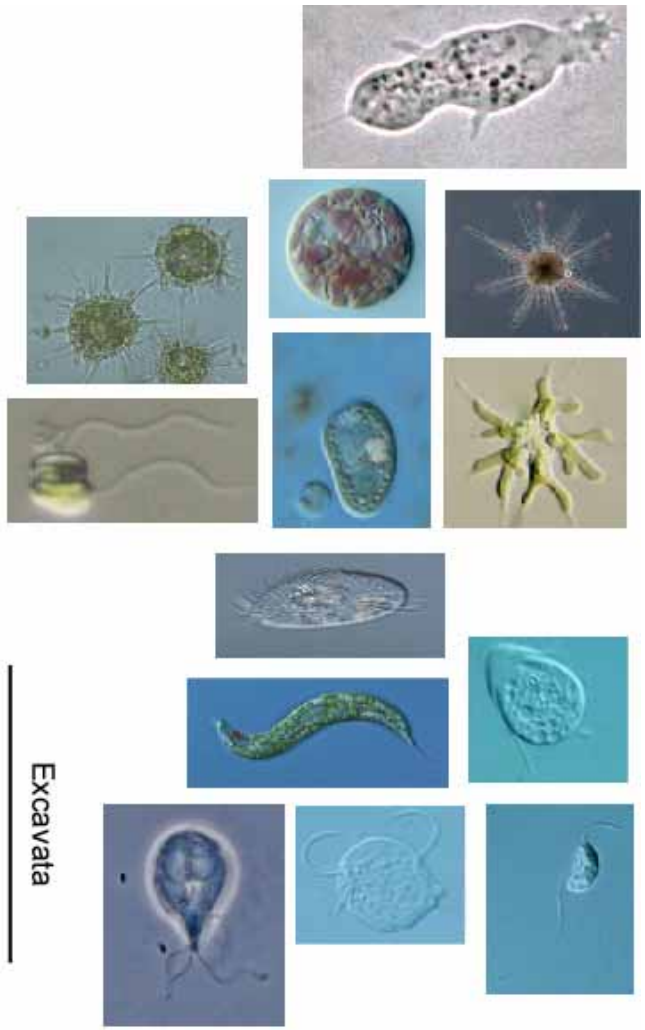
↓

fast

- -r78 data set (excluding r7 and r8) 6047 sites
- JTT-F + Γ model







単細胞真核生物（原生生物）
の解析が重要

真核生物の初期進化
——現在の知見と今後の課題

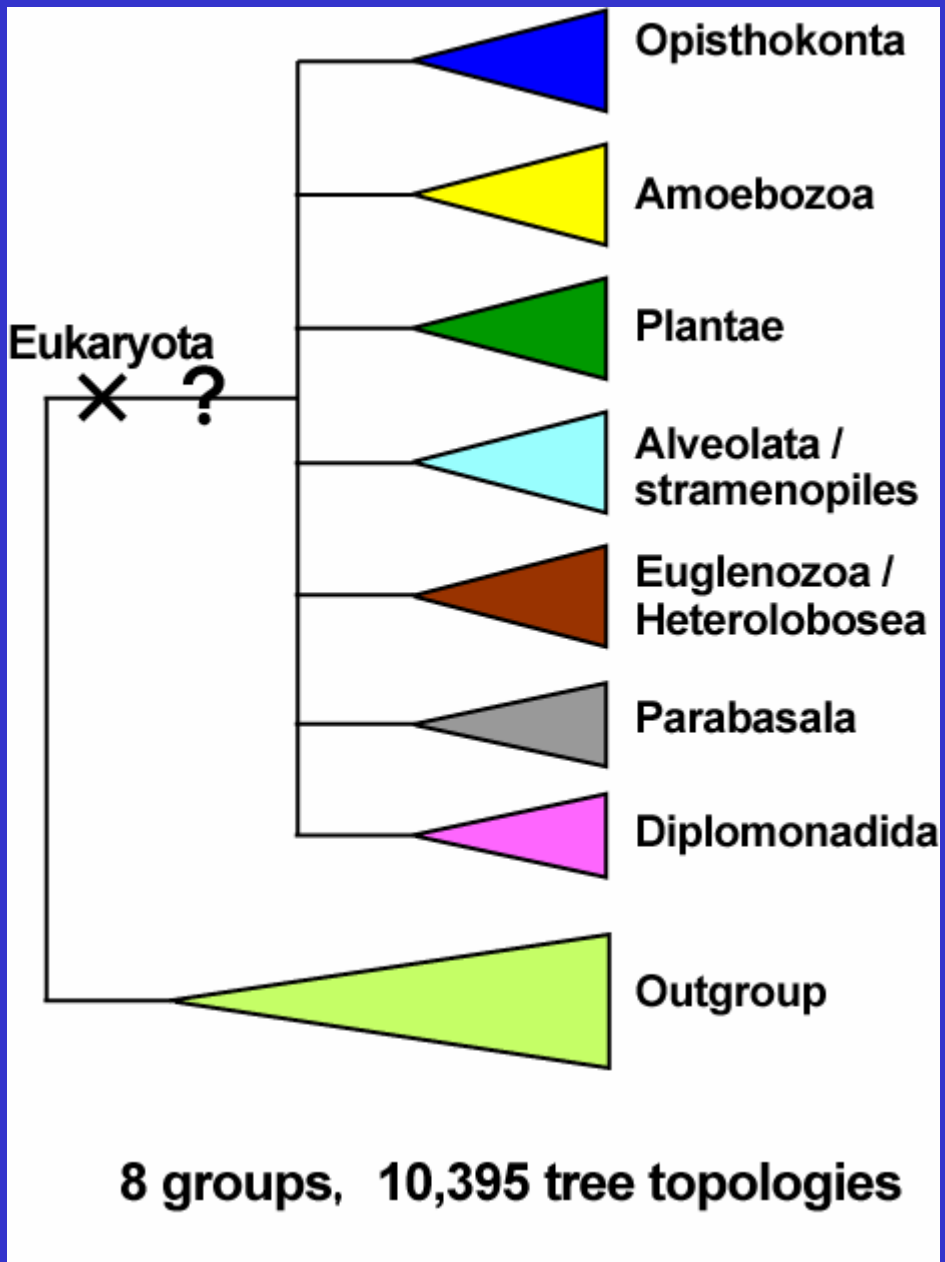
今後の課題:

1. 本研究で用いた遺伝子および他の遺伝子のデータを、十分な種のサンプリングのもとで蓄積
   とくに、Diplomonadida、Parabasala、Amoebozoa、Euglenozoa、Heterolobosea内部の種のサンプリングの充実、ESTによるリボソーム蛋白質遺伝子の網羅的解析さらに、Excavate taxa、Cryptophytes、Haptophytes、Cercozoa、Heliozoa などの位置づけの検討
2. データ解析システムの改良、拡張

   とくに、exhaustive search を行えるOTU数の上限の拡張（9,10,11ぐらいまで）、解析プログラムの並列化
3. より現実的で洗練されたモデルの適用、開発

   とくに、コバリオンモデルのデータ解析への実装

真核生物7つの大きな単系統群の解析と系統樹のRooting



結合データ解析に用いた24遺伝子

Translation:	EF-1α, EF-2, RpS14, RpS15a, RpL5, RpL8, RpL10a, IleRS, ValRS
Transcription:	RNA polymerase II (Rpb1)
Chaperon:	CPN60, HSP70c, HSP70mit, HSP70er, HSP90c, CCTα, CCTγ, CCTδ, CCTζ
Cytoskeleton:	Actin, α-tubulin, β-tubulin
rRNA:	SSUrRNA, LSUrRNA

Outgroup:
原核生物 (*Archaea, Bacteria*)
もしくは paralogous **遺伝子**

Exhaustive Search による最尤系統樹 (Best tree) の探索

例) PAML Ver3.1 の CODEML program

1～2G flops のPC一台を専用に走らせて
N=４０、M＝５００の標準的データセットを用いて
アミノ酸置換のJTT-Fモデルで
rateカテゴリー数８の離散化Γ分布を導入してRASモデルとし
各５種からなる8つの系統 (グループ) 10395通りの
Exhaustive Searchを行うと約200時間必要
同一規模のデータセットからなる遺伝子２０個について
Separate Model にて結合データ解析を試み
２００×２０＝４０００時間＝１６７日 必要
総合評価で近似的なブートストラップ解析を１００００回行うとすると
さらに２０日で合計１８７日 必要

→ PC2000台分の規模でExhaustive Searchを並列化すれば、
2時間で終了
ブートストラップも並列化して２０００台分で計算すれば、14.4分で終了

Exhaustive Search による最尤系統樹 (Best tree) の探索

さらに2系統 (グループ) について各5種ずつを追加し
それぞれのグループ内部の関係を固定し
10グループ2027025通りの Exhaustive Searchを行おうとすると
推定すべき枝の長さは、97本 (50種) ／77本 (40種) ＝1.26倍
Searchすべき系統樹数は、2027025／10395＝195倍
同一規模のデータセットからなる遺伝子20個について
Separate Model にて結合データ解析を試みよう
(遺伝子数、すなわち座位数は変わらない) とすると
167日×1．26×195＝41032日 必要
総合評価で近似的なブートストラップ解析を10000回行うとすると
さらに20日×195＝3900日 必要、 で**合計44932日** 必要

→ PC約2000台分の規模でExhaustive Search、ブートストラップ
ともに並列化すれば、**22．5日**で終了