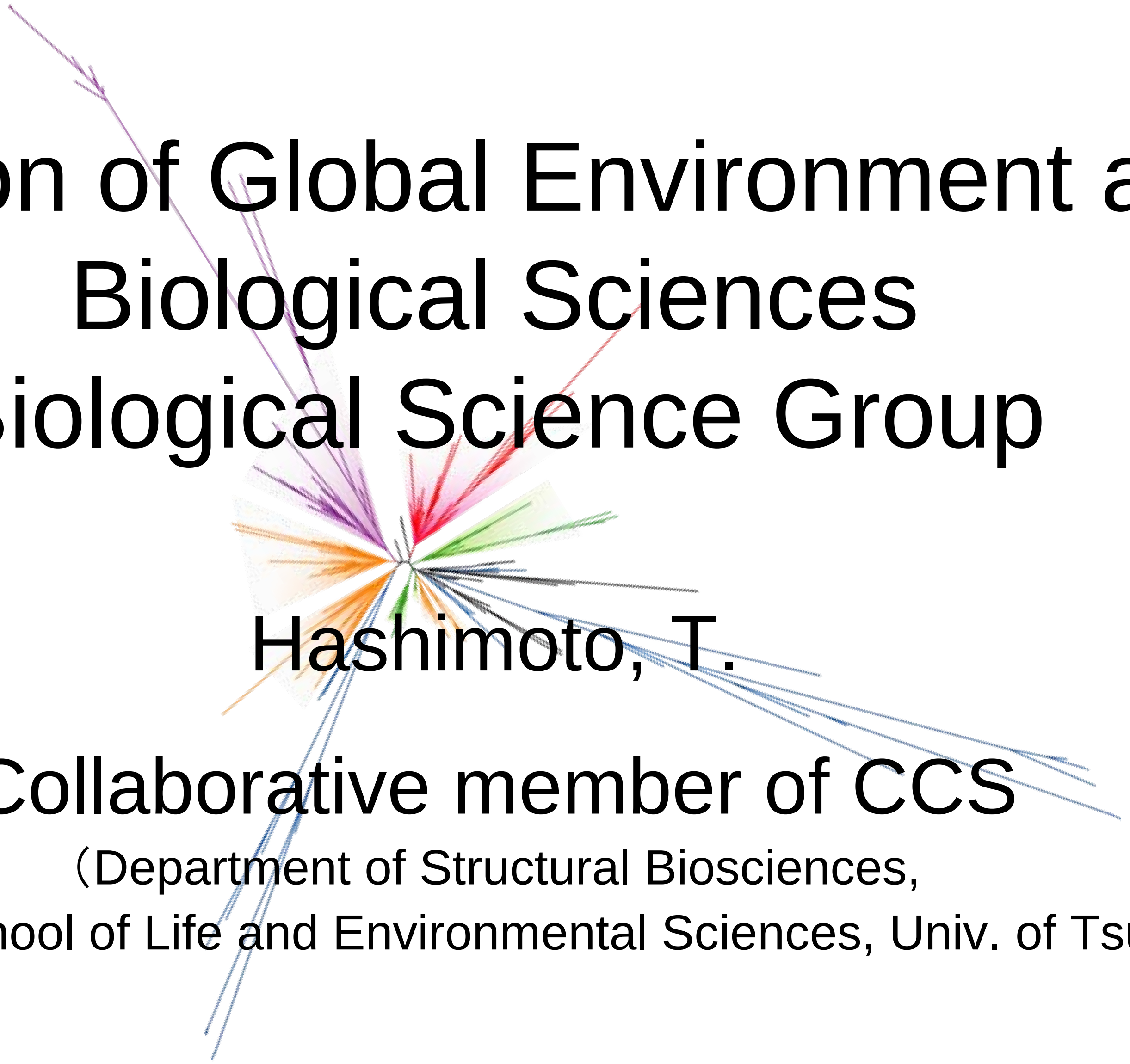


Division of Global Environment and Biological Sciences Biological Science Group

Hashimoto, T.

Collaborative member of CCS

（Department of Structural Biosciences,
Graduate School of Life and Environmental Sciences, Univ. of Tsukuba）

Foundation

Division of Global Environment and Biological Sciences

Biological Science Group

ー Created at Apr. 2004

Hashimoto, T.

Collaborative member

(Dept. Structural Biosci., Univ. of Tsukuba)

Members of Biological Science group

Since Aug. 2005

- Center for Computational Sciences

  - Ass. Prof. Yuji Inagaki

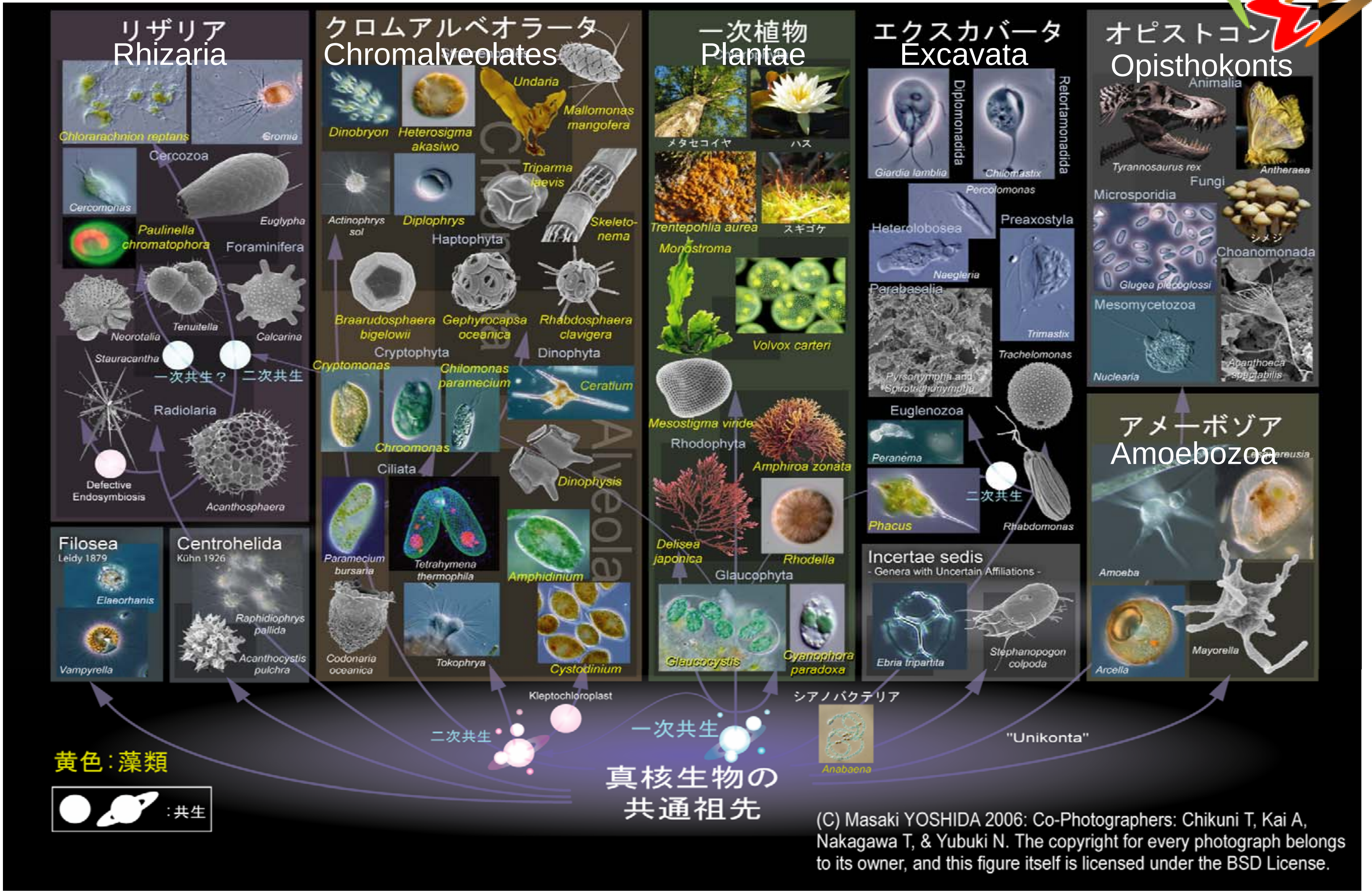
- Dept. Structural Biosciences

  - Prof. Tetsuo Hashimoto
    (Collaborative member)
  - Res. Assoc. Miako Sakaguchi

Research Area

- Molecular Evolution
  - Global phylogeny of eukaryotes
  - Methodological studies on molecular phylogenetic inference

Beautiful creatures



(C) Masaki YOSHIDA 2006: Co-Photographers: Chikuni T, Kai A, Nakagawa T, & Yubuki N. The copyright for every photograph belongs to its owner, and this figure itself is licensed under the BSD License.

Main Projects

- Inference on the phylogenetic relationship among large groups of eukaryotes
- Inference on the root of the tree of eukaryotes
- Inference on the mitochondrial evolution
- Inference on the plastid evolution
- Methodological studies on molecular phylogenetic inference

- Establishment of the methodology for large scale phylogenetic analysis based on the maximum likelihood (ML) approach

Collaboration with other Group in CCS

‘Mol-Evol’ group:

- Biological Science Group

  Yuji Inagaki / T. Hashimoto

- High Performance Computing Group

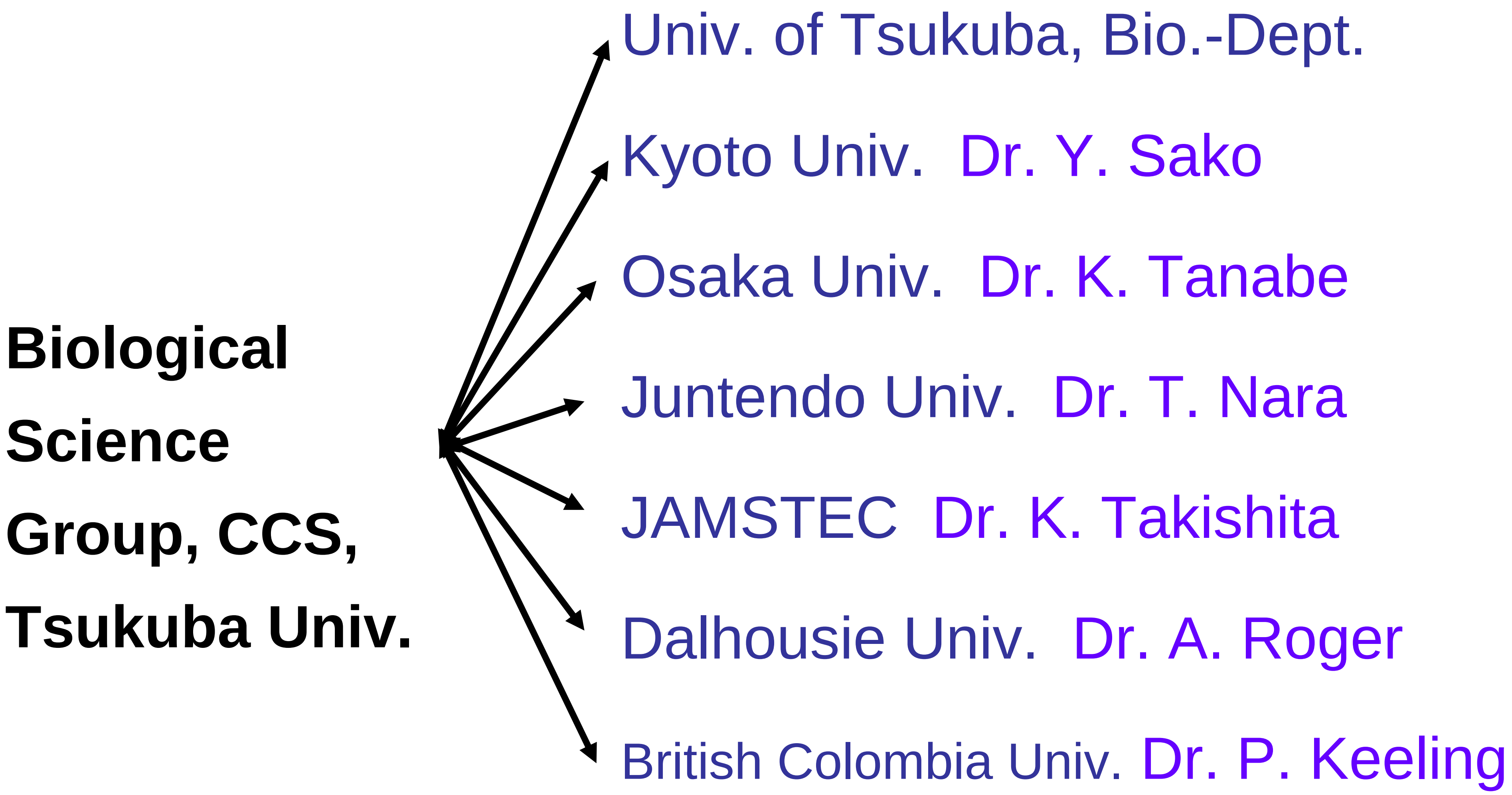
  Prof. Mitsuhisa Sato (Director of CCS)

  (Graduate students)

  Yoshihiro Nakajima

  Akihiro Aida (~Mar. 2007)

Collaboration with other Institutions



Biological Science Group, CCS, Tsukuba Univ.

Univ. of Tsukuba, Bio.-Dept.

Kyoto Univ. Dr. Y. Sako

Osaka Univ. Dr. K. Tanabe

Juntendo Univ. Dr. T. Nara

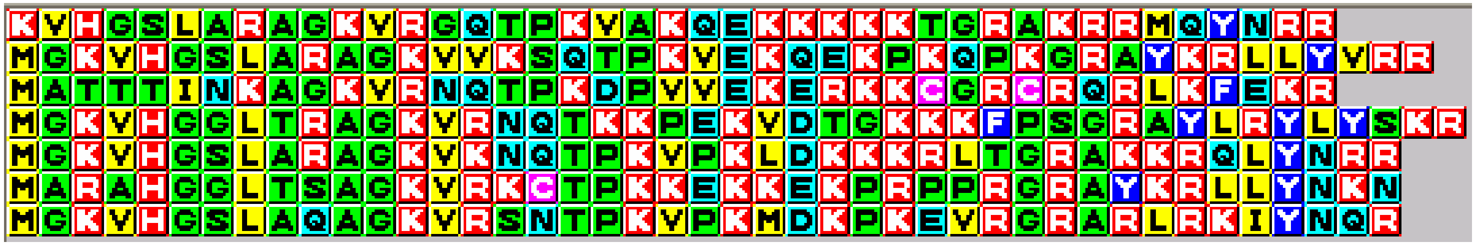
JAMSTEC Dr. K. Takishita

Dalhousie Univ. Dr. A. Roger

British Colombia Univ. Dr. P. Keeling

Sequence Alignment and Data for Phylogenetic Analysis

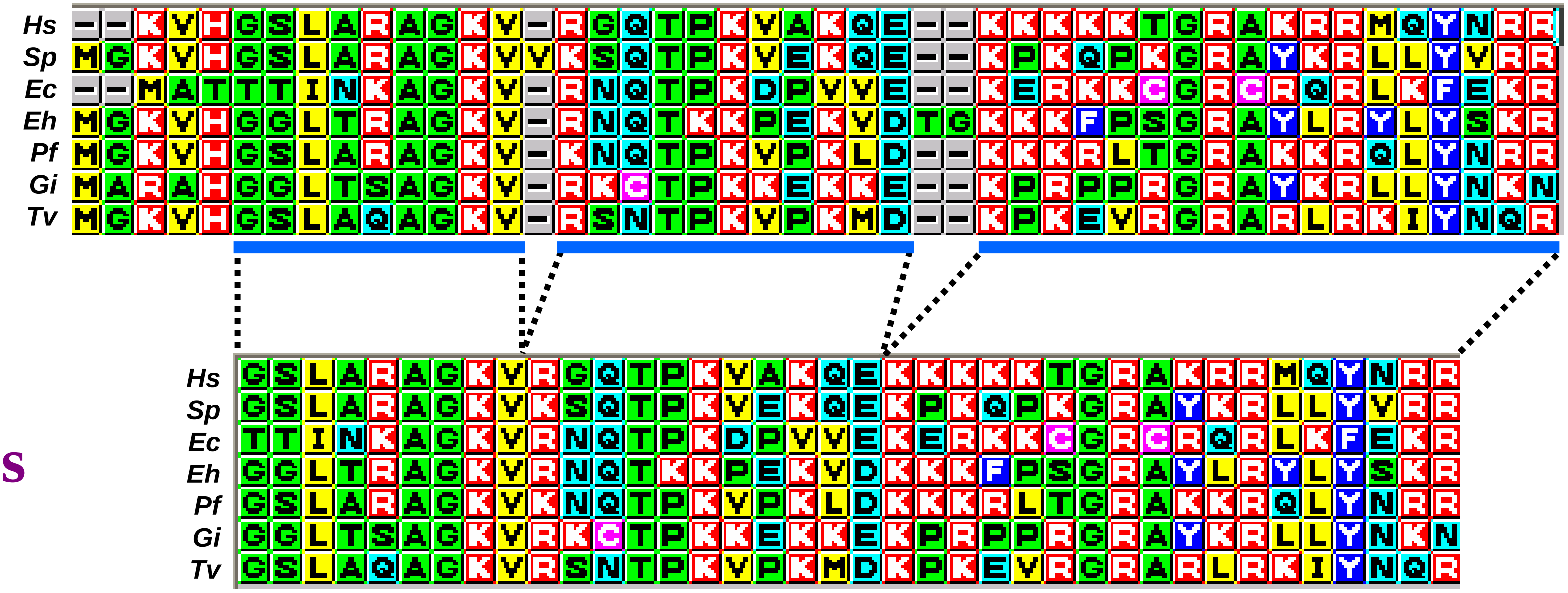
Ribosomal protein mS30



```
Homo sapiens                  KVHGSLARAGKVRGQTPKVAKQEKKKKKTGRAKRRMQYNRR
Schizosaccharomyces pombe     MGKVHGSLARAGKVVVKSQTPKVEKQEKPKQPKGRAYKRLLYVRR
Encephalitozoon cuniculi      MATTTINKAGKVRNQTPKDPVVEKERKKCGRCRQRLKFEKR
Entamoeba histolytica         MGKVHGGLTRAGKVRNQTKKPEKVDTGKKKFPSGRAYLRYLYSKR
Plasmodium falciparum         MGKVHGSLARAGKVKNQTPKVPKLDKKKRLTGRAKKRQLYNRR
Giardia intestinalis          MARAHGGLTSAGKVRKCTPKKEKKEKPRPPRGRAYKRLLYNKN
Trichomonas vaginalis         MGKVHGSLAQAGKVRSNTPKVPKMDKPKEVRGRARLRKIYNQR
```

Aligned

Selected sites

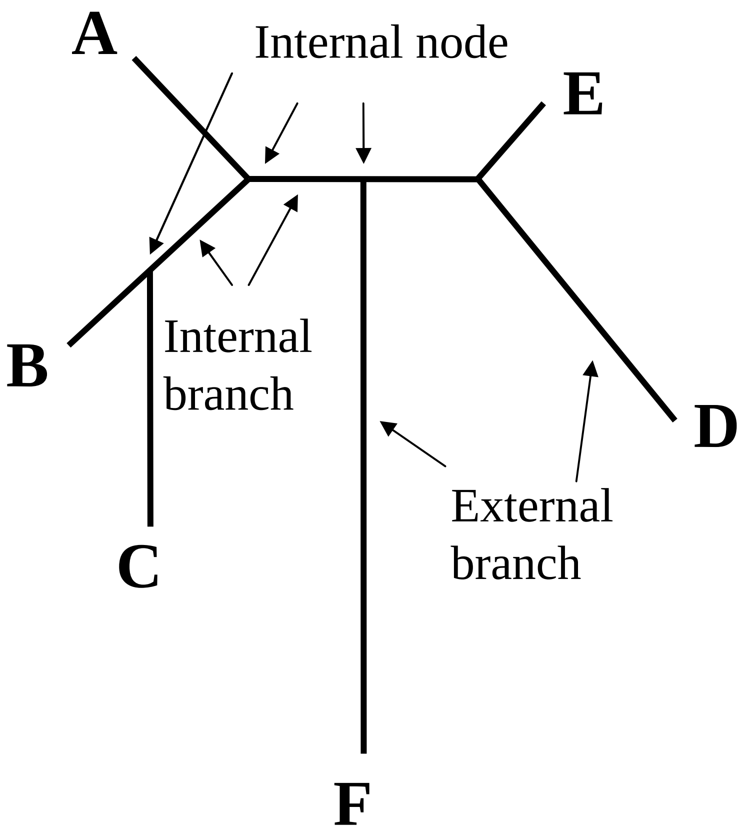


```
Hs --KVHGSLARAGKV-RGQTPKVAKQE--KKKKKTGRAKRRMQYNRR
Sp MGKVHGSLARAGKVVVKSQTPKVEKQE--KPKQPKGRAYKRLLYVRR
Ec --MATTTINKAGKV-RNQTPKDPVVE--KERKKCGRCRQRLKFEKR
Eh MGKVHGGLTRAGKV-RNQTKKPEKVDTGKKKFPSGRAYLRYLYSKR
Pf MGKVHGSLARAGKV-KNQTPKVPKLD--KKKRLTGRAKKRQLYNRR
Gi MARAHGGLTSAGKV-RKCTPKKEKKE--KPRPPRGRAYKRLLYNKN
Tv MGKVHGSLAQAGKV-RSNTPKVPKMD--KPKEVRGRARLRKIYNQR
```

```
Hs GSLARAGKVRGQTPKVAKQEKKKKKTGRAKRRMQYNRR
Sp GSLARAGKVKSQTPKVEKQEKPKQPKGRAYKRLLYVRR
Ec TTINKAGKVRNQTPKDPVVEKERKKCGRCRQRLKFEKR
Eh GGLTRAGKVRNQTKKPEKVDKKKFPSGRAYLRYLYSKR
Pf GSLARAGKVKNQTPKVPKLDKKKRLTGRAKKRQLYNRR
Gi GGLTSAGKVRKCTPKKEKKEKPRPPRGRAYKRLLYNKN
Tv GSLAQAGKVRSNTPKVPKMDKPKEVRGRARLRKIYNQR
```

- Physicochemically similar aa are marked with same color
- Column is called position or site

Phylogenetic Tree



A～F：extant species, groups, etc.

→ one of the relationships between 6 species

→ ((A,(B,C)),F,(D,E))

: tree topology

Branch length: substitution rate

Unit: substitutions/site

Maximum Likelihood (ML) method

- Based on explicit evolutionary models, the analysis uses probability to find a tree that best accounts for the variation in a data set

$$f(\theta \mid \text{data}) = \Pr(\text{data} \mid \underset{\substack{\parallel \\ (\text{tree topology, branch lengths, } \alpha,\ \pi\text{, etc.})}}{\theta}) \rightarrow \text{maximize}$$

$$\Downarrow$$

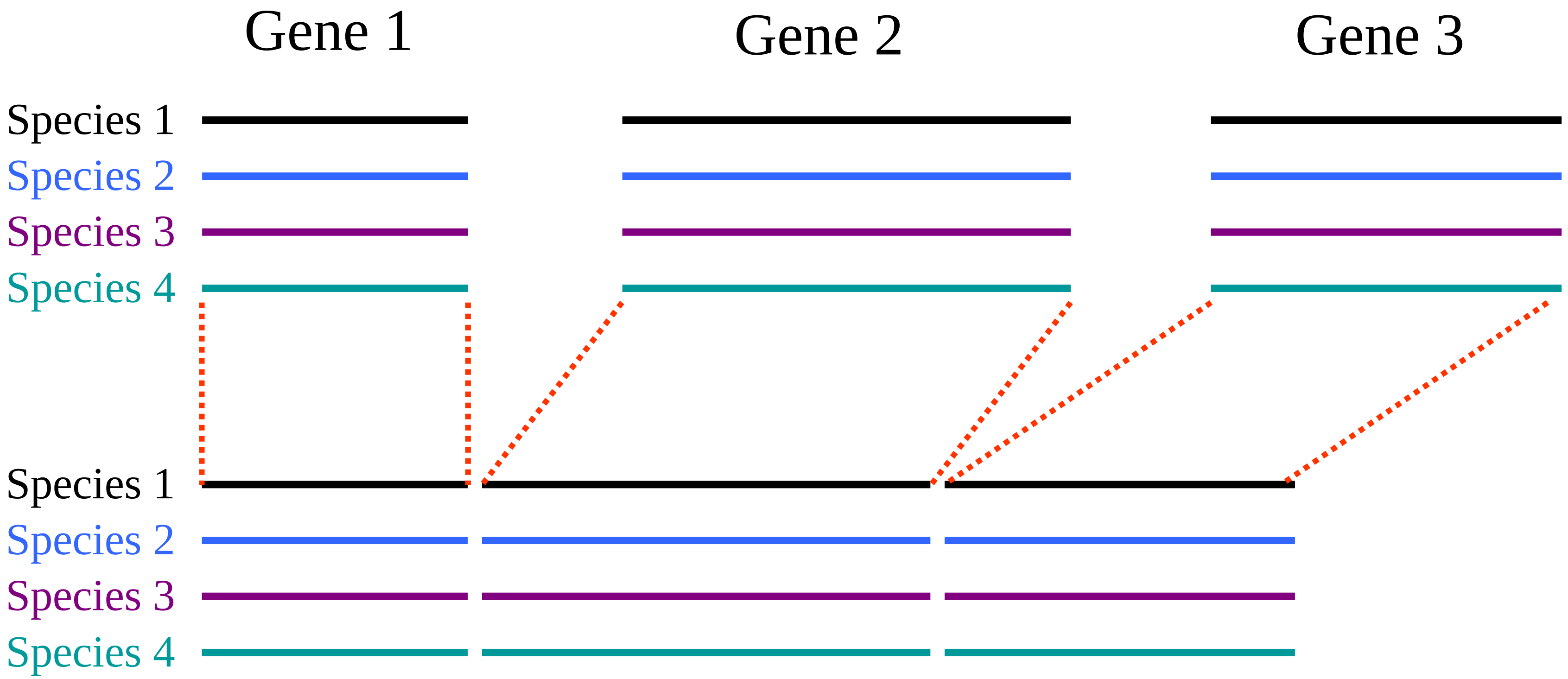
$$\log\{\, f(\theta \mid \text{data})\} \ \ (\text{log-likelihood}) \ \rightarrow \text{maximize}$$

- Calculation is performed on each column of an alignment
- All possible trees are exhaustively examined → Select the ML tree
- Robust against the violation of evolutionary rate constancy between species (sequences) ⇒ applicable for diverse sequences
- Computationally intense

Number of Possible Trees

species	number of tree topologies
1	1
2	1
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
	…
n	$\dfrac{(2n\text{-}5)!}{2^{n-3}(n\text{-}3)!}$

Combined analysis based on Concatenate Model



Gene 1

Gene 2

Gene 3

Species 1

Species 2

Species 3

Species 4

Species 1

Species 2

Species 3

Species 4

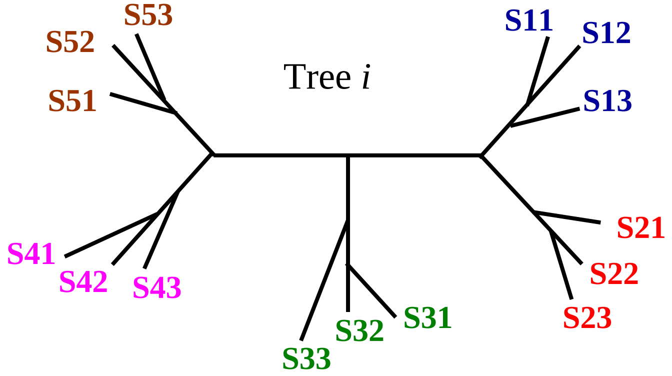
Concatenate sequences of different genes

⇒ regard as 1 gene

ML phylogenetic analysis ⇒ a set of parameters : $\hat{\theta}$

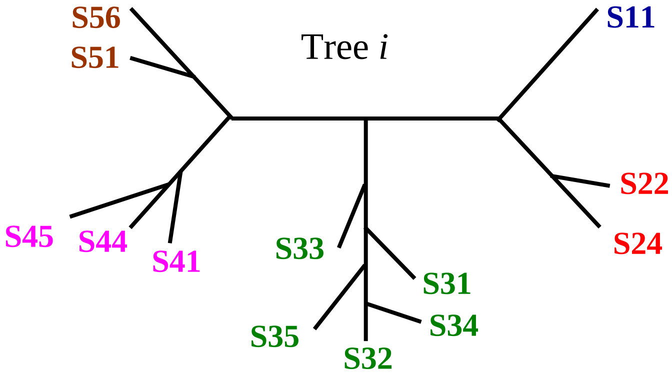
Combined analysis based on Separate Model

Gene 1
5 groups, 15 species



$l_{1(i)}(\hat{\theta}_{1(i)}|\mathrm{X}_1)$

Gene 2
5 groups, 13 species



$l_{2(i)}(\hat{\theta}_{2(i)}|\mathrm{X}_2)$

For each of Tree *i*
(*i* =1,…15)
calculate log-likelihood
of the gene *k*

Gene m
5 groups, x species

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

$+$

$$\sum_{k=1}^{m} l_{k(i)}(\hat{\theta}_{k(i)}|X_k)$$

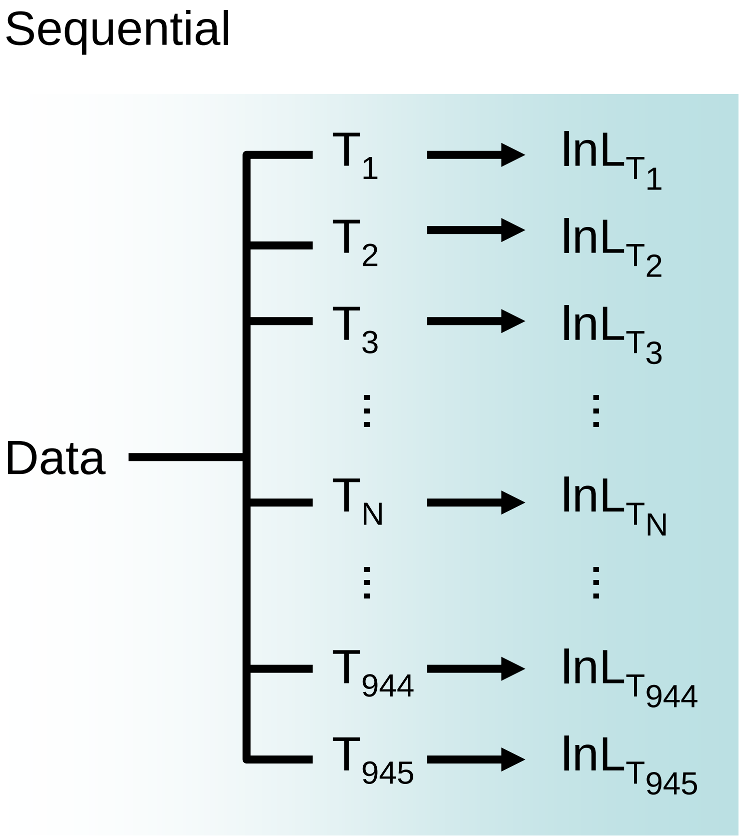
The log-likelihood of the ML tree in total

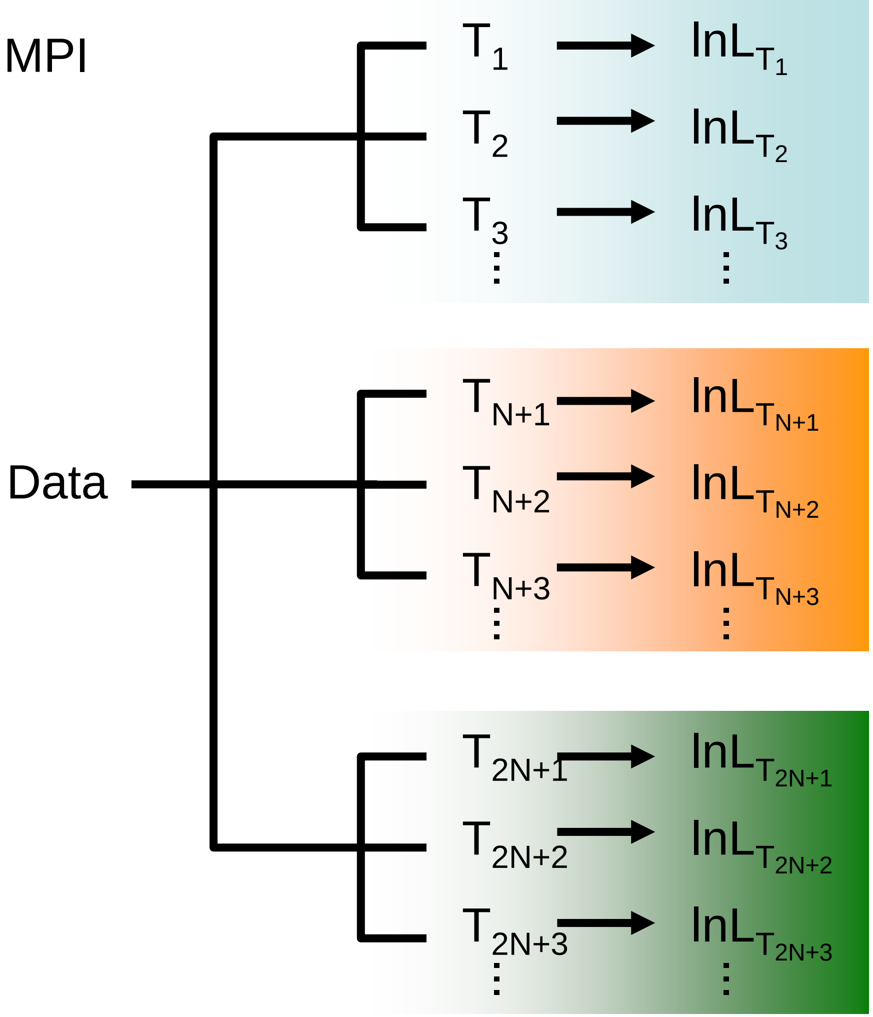
$$\max_{i}\{\sum_{k=1}^{m} l_{k(i)}(\hat{\theta}_{k(i)}|X_k\}=\max_{i}[\sum_{k=1}^{m}\{\sum_{h=1}^{m_k}\log f_{k(i)}(X_{kh}|\hat{\theta}_{k(i)})\}]$$

$(i=1,\ldots,15)$

MPI-Puzzle – For exhaustive tree search

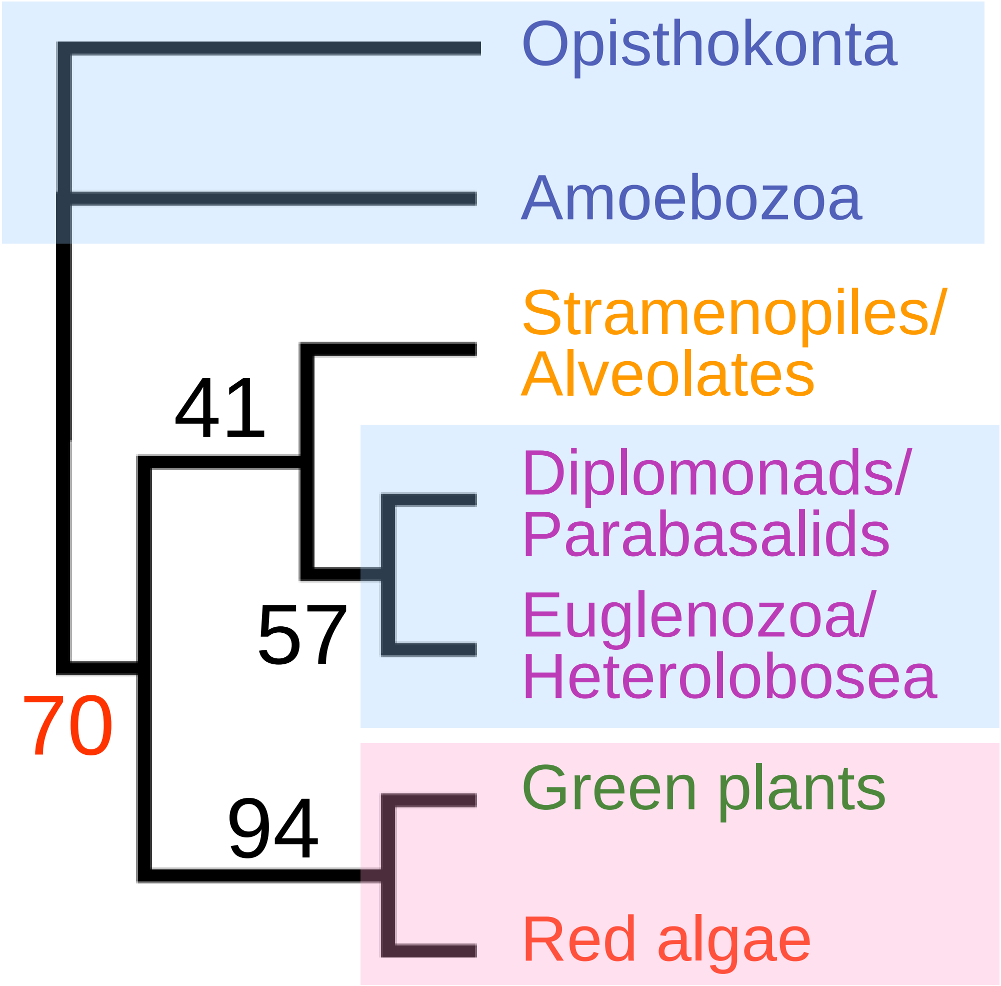
- Calculated lnL by Tree-Puzzle




MPI
Data
T_1 → lnL_{T_1}
T_2 → lnL_{T_2}
T_3 → lnL_{T_3}
⋮ ⋮
T_{N+1} → $lnL_{T_{N+1}}$
T_{N+2} → $lnL_{T_{N+2}}$
T_{N+3} → $lnL_{T_{N+3}}$
⋮ ⋮
T_{2N+1} → $lnL_{T_{2N+1}}$
T_{2N+2} → $lnL_{T_{2N+2}}$
T_{2N+3} → $lnL_{T_{2N+3}}$
⋮ ⋮

Phylogeny of Eukaryotes – Recent results

23-gene analyses – w/o α-tubulin ~10,000 positions
< Separate model > (Exhaustive search for 945 trees)

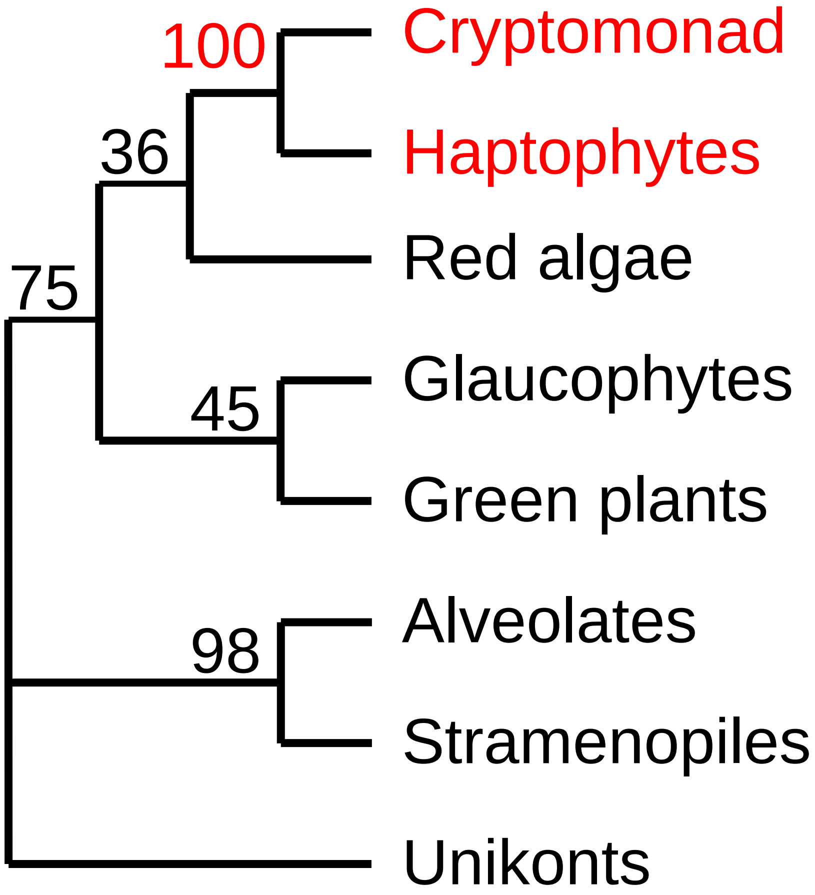


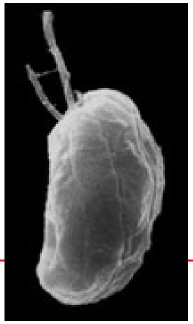
- Recovered three eukaryotic “supergroups”
  - Unikont: Opisthokonta + Amoebozoa
  - Excavata: Dip/Par + Eug/Het
  - Plantae: G+R

Phylogeny of Eukaryotes – Recent results

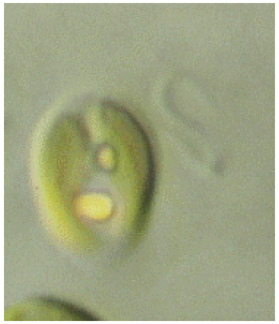
Multi-gene analyses – with ~8,500 positions

< Separate model > (Exhaustive search for 10395 trees)





Guillardia



Pavlova

Multiple Gene Phylogenies Support the Monophyly of Cryptomonad and Haptophyte Host Lineages.

Patron, Inagaki, and Keeling: Curr. Biol. 2007

Summary of Research Activities 2005-2007

― Biological Science Group ―

Publications		Total
Peer reviewed	Review (In Jap.)	
20	4	24

Domestic Meetings			Total
Oral	Poster	Invited	
12	9	9	30

International Meetings			Total
Oral	Poster	Invited	
3	3	2	8

Summary of Research Activities 2005-2007

― Biological Science Group ―

Grants: Grant-In-Aid for Scientific Research from JSPS

2005-2006 (B) ¥14,200,000

2006-2007 (C) ¥3,600,000

Grant-In-Aid for Young Scientists from JSPS

2006-2007 (B) ¥2,600,000

Awards: Yuji Inagaki 2007

An award from the ministor of MEXT for young scientists

Short Term Future Plans

- Phylogenetic relationship among large groups of Eukaryotes
  - Excavata monophyly?
  - Chromalveolates monophyly?

- Methodological studies on molecular phylogenetic inference
  - Imapct of the utilization of different models for combining multiple genes ('concatenate' or 'separate')

Short Term Future Plans

- Establishment of the methodology for large scale phylogenetic analysis based on the maximum likelihood (ML) approach

  - Exhaustive search up to 2,027,025 trees for 10 lineages

Phylogenetic Analysis

Phylogenetic inference: estimation of :

- Tree topology
- Branch length

Methods of phylogenetic reconstruction:

- Distance matrix method
- Maximum parsimony method
- Maximum likelihood method

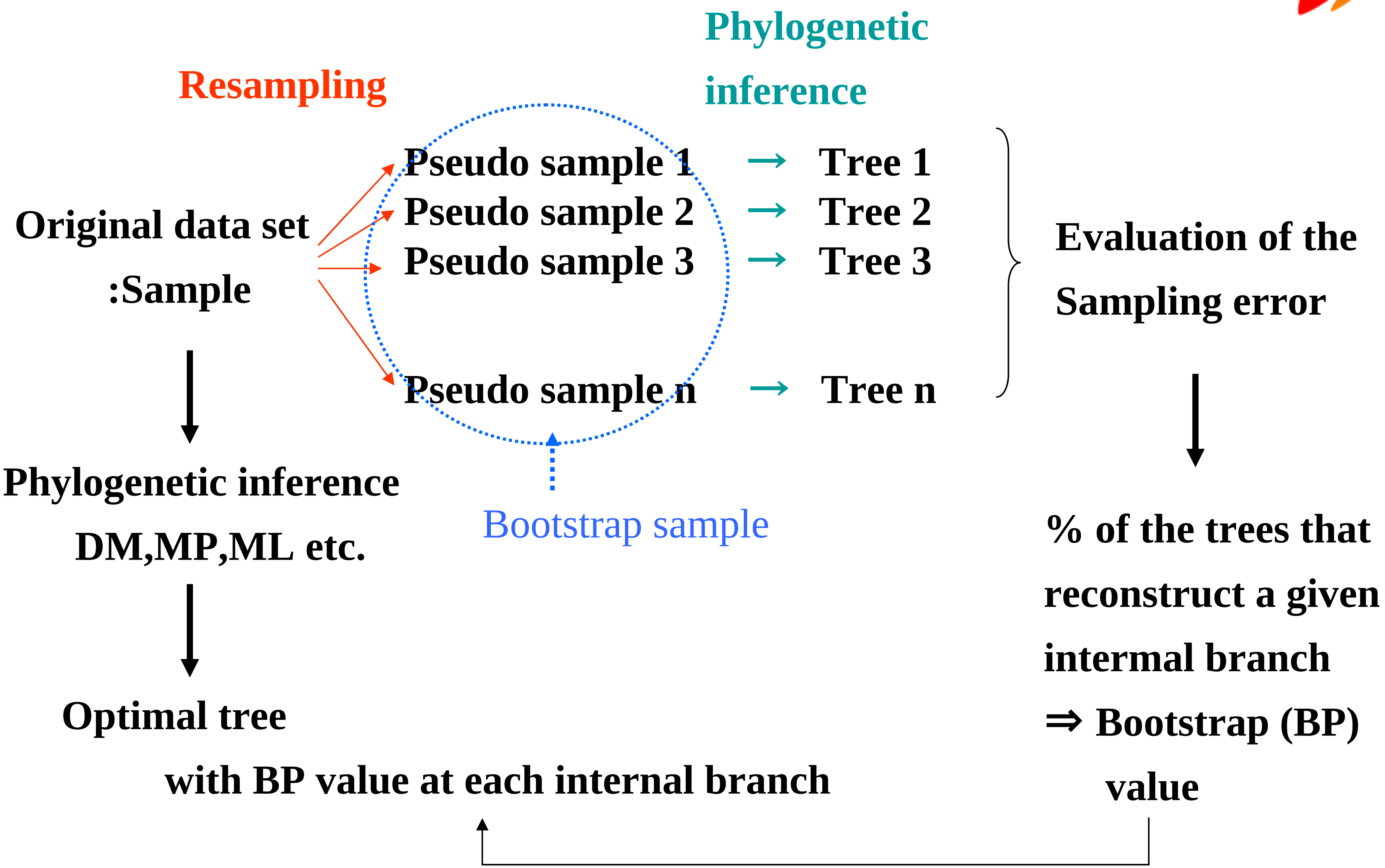
Phylogenetic inference: estimation of :

- Tree topology
- Branch length

Methods of phylogenetic reconstruction:

- Distance matrix method
- Maximum parsimony method
- Maximum likelihood method

Bootstrap analysis



Resampling

Phylogenetic inference

Original data set

:Sample

Pseudo sample 1 → Tree 1

Pseudo sample 2 → Tree 2

Pseudo sample 3 → Tree 3

Pseudo sample n → Tree n

Evaluation of the Sampling error

Bootstrap sample

Phylogenetic inference

DM,MP,ML etc.

% of the trees that reconstruct a given intermal branch

⇒ Bootstrap (BP) value

Optimal tree

with BP value at each internal branch

Maximum likelihood method in brief

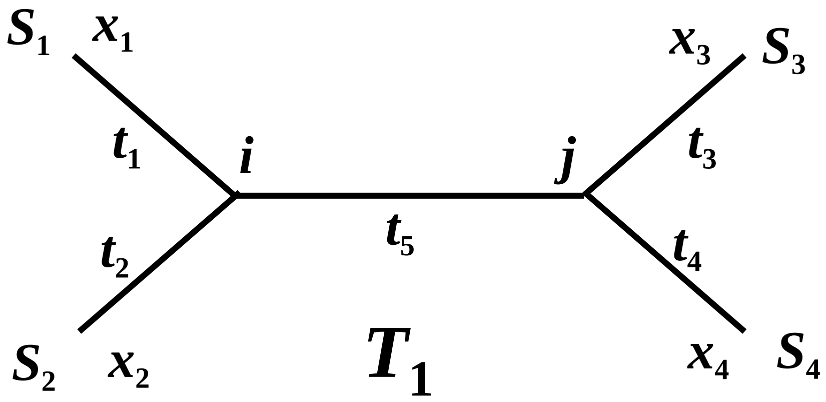
Example data set

(4 species, *n* positions)

***X* = (*Xij*) (*i*=1,…4; *j*=1,…,*n*)**

h'th position

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



$S_1,\ldots, S_4$	**: Extant species**
$x_1,\ldots, x_4$	**: data of *h*'th position in extant species**
i, j	**: data of *h*'th position in ancestral species**
$t_1,\ldots, t_5$	**: branch lengths**

Assumption １. X*t* : Continuous-time stationary Markov process with a transition probability *Pij(t)*
transition probability *k to l* :

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

Evolution (substitutions) of each branch occurs independently.

Assumption ２. Each position evolves independently.

Under the Assumption 1.

Using Chapmann-Kolmogorov equation,

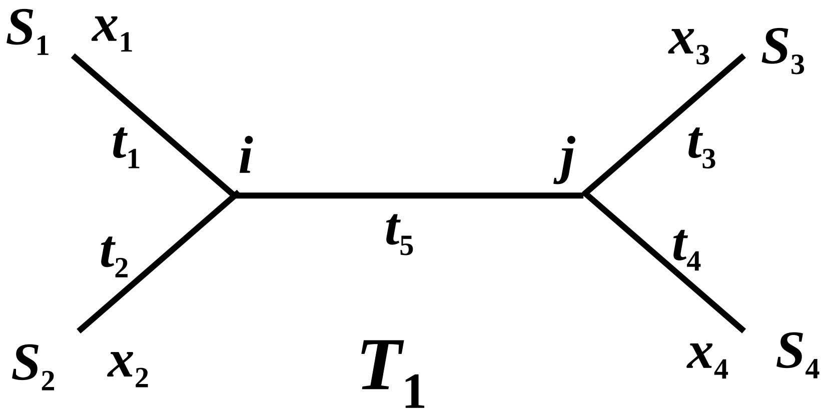
$f\,(x_1, x_2, x_3, x_4|\theta)$ ← **probability of having the data of *h*’th position, given a tree topology, T_1, and a transition matrix Pij (t)**

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

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

$$= \sum_i \left\{ \pi_i Pix_1(t_1)Pix_2(t_2) \sum_j Pij(t_5)\ Pjx_3(t_3)\ Pjx_4(t_4) \right\}$$

π_i **: composition of nucleotide or amino acid *i***



$S_1,\ldots, S_4$	**: Extant species**
$x_1,\ldots, x_4$	**: data of *h*’th position in extant species**
i, j	**: data of *h*’th position in ancestral species**
$t_1,\ldots, t_5$	**: branch lengths**

Under the Assumption 2.

Likelihood function of θ given a data X can be written as:

$$\boldsymbol{L\,(\theta \mid \mathrm{X}) = \prod_{h=1}^{n} f\,(\mathrm{X}h \mid \theta)}$$

Its log-likelihood:

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

$\theta = (t_1, \ldots , t_5)$ **: branch lengths**

$\hat{\theta}$ **can be estimated by maximizing the log-likelihood function,** l

$$\boldsymbol{l\,(\hat{\theta} \mid \mathrm{X}) = \max \{l\,(\theta \mid \mathrm{X}) : \hat{\theta} \in \Theta\}}$$

By comparing the log-likelihood ($l\,(\hat{\theta} \mid \mathrm{X})$ **) values between the different tree topologies,** T_1**,** T_2**, and** T_3**, a tree with the highest log-likelihood value is selected as the ML tree.**

Evolutionary model

	Parameters
-Tree topology, *T*	branch lengths (t_1, …, t_5)
-Transition Probability	those included in $P_{ij}(t)$
-Composition	π_i
-Among site-rate heterogeneity	α (shape parameter of Γ distribution)

$P_{ij}(t)$: for nucleotide substitutions	: for amino acid substitutions
Jukes-Cantor	Poisson
Kimura-2-parameter	PAM
HKY85	JTT
TN93	WAG
……..	……

Bootstrap resampling

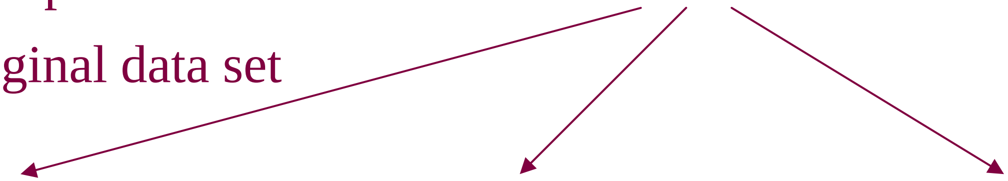
Original data set

: 4 × 10 matrix

Cryptosporidium	**ACTAACTGAG**
Toxoplasma	**ACTATGAGAG**
Theileria	**GCTGTGAAAA**
Plasmodium	**GCTGTGATCA**
	0123456789

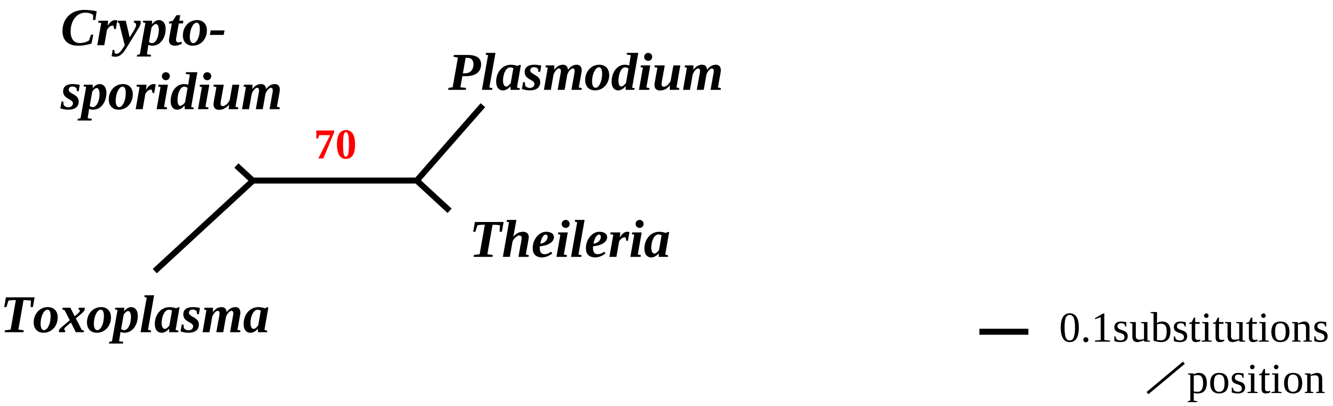
Randomly resample 10 positions out of

10 positions in the original data set

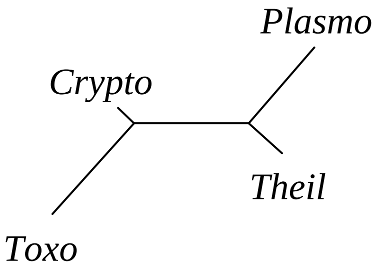


	Pseudo sample 1	Pseudo sample 2	Pseudo sample 3	
Crypto	**CACTGAAGAA**	**AACGTAATAC**	**GGCTAATGAG**	**.....**
Toxo	**CTGTGAAGTA**	**TACGAAATAC**	**GGGTATAGAG**	**.....**
Theil	**CTGTAAGATG**	**TGCAAGATAC**	**AAGTGTAAAA**	**.....**
Plasmo	**CTGTACGTTG**	**TGCAAGCTCC**	**ATGTGTATCA**	**.....**
	1452983740	**4019608281**	**9752346789**	

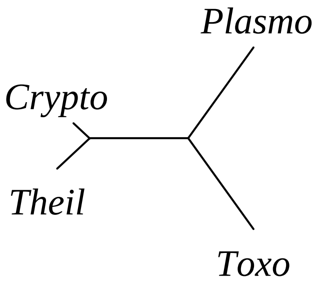
ML tree of the original data set



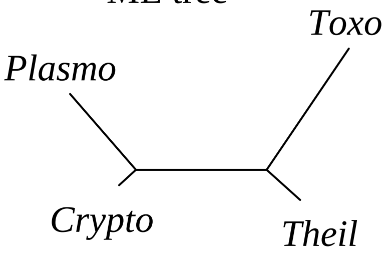
Bootstrap sample 1
ML tree



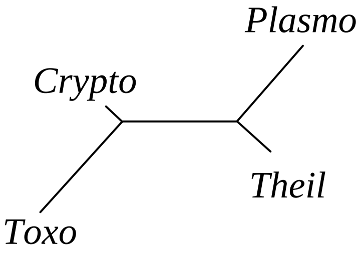
Bootstrap sample 2
ML tree



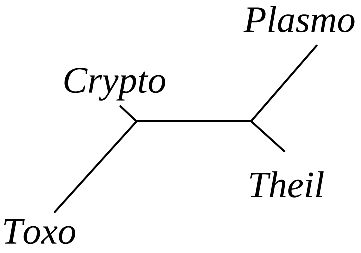
Bootstrap sample 3
ML tree



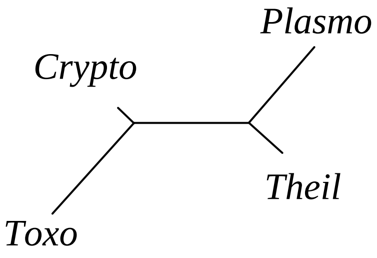
Bootstrap sample 4
ML tree



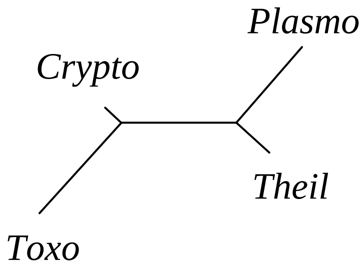
Bootstrap sample 5
ML tree



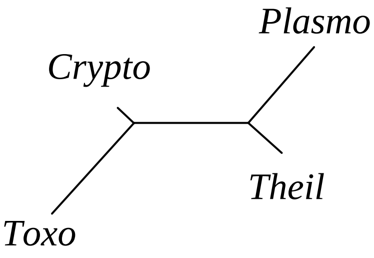
Bootstrap sample 6
ML tree



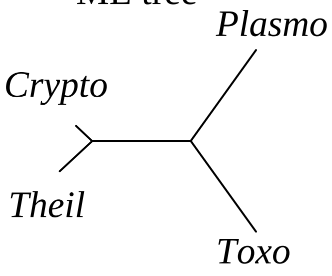
Bootstrap sample 7
ML tree



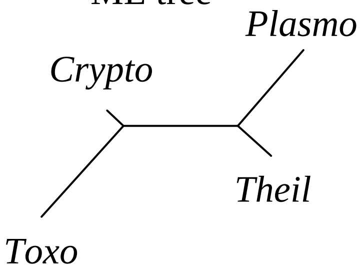
Bootstrap sample 8
ML tree



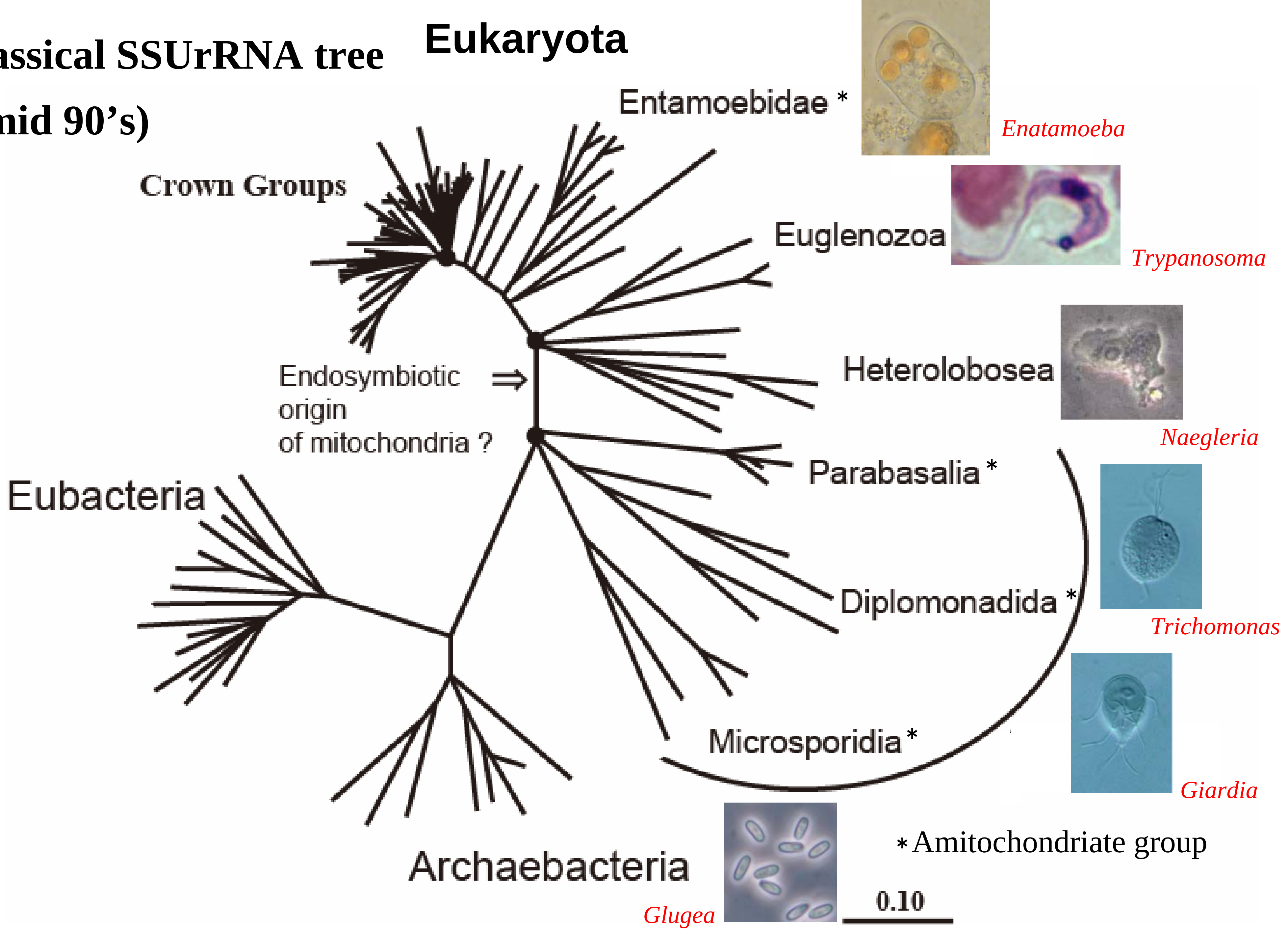
Bootstrap sample 9
ML tree



Bootstrap sample 10
ML tree

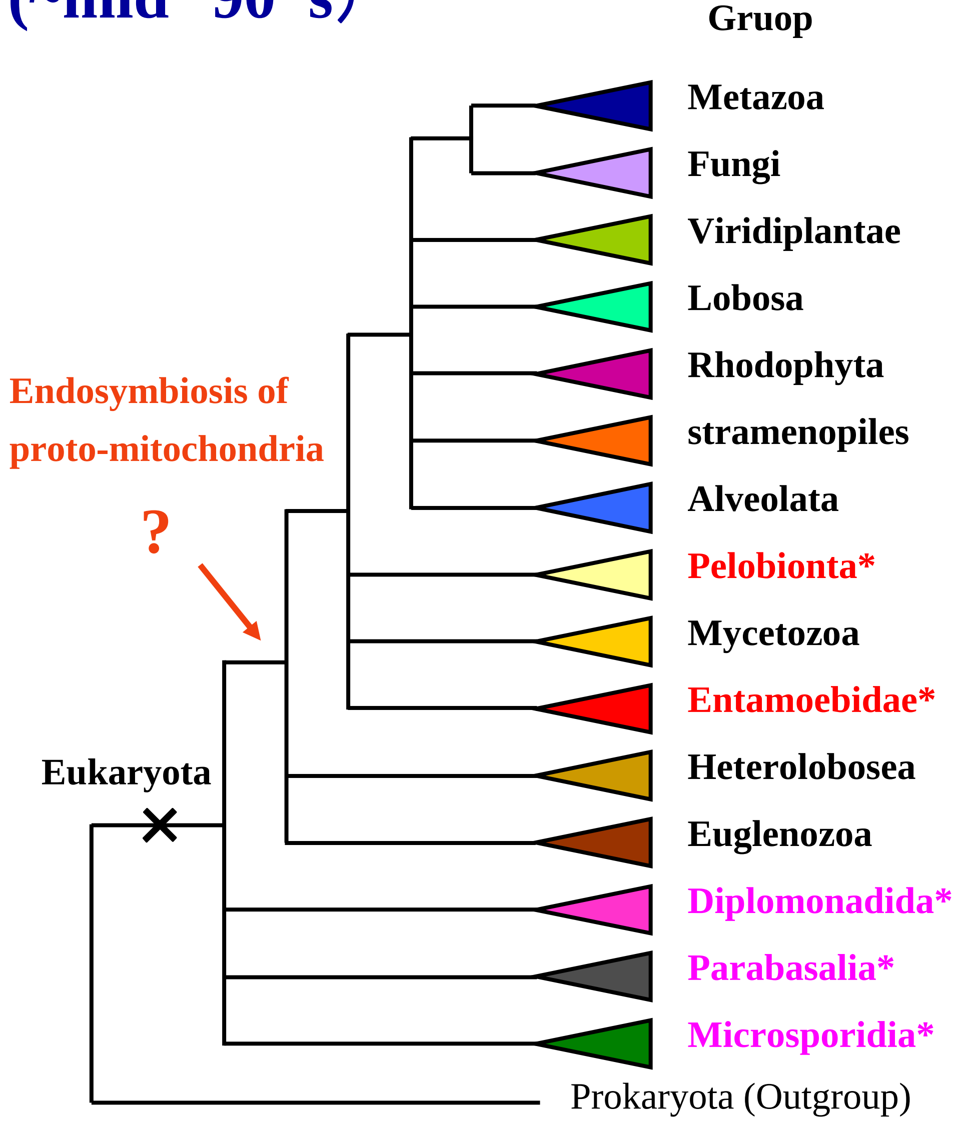


Classical SSUrRNA tree (~mid 90's)



SSUrRNA tree of Eukaryota (~mid 90’s)

Presence of large monophyletic groups



* Amitochondriate protist group

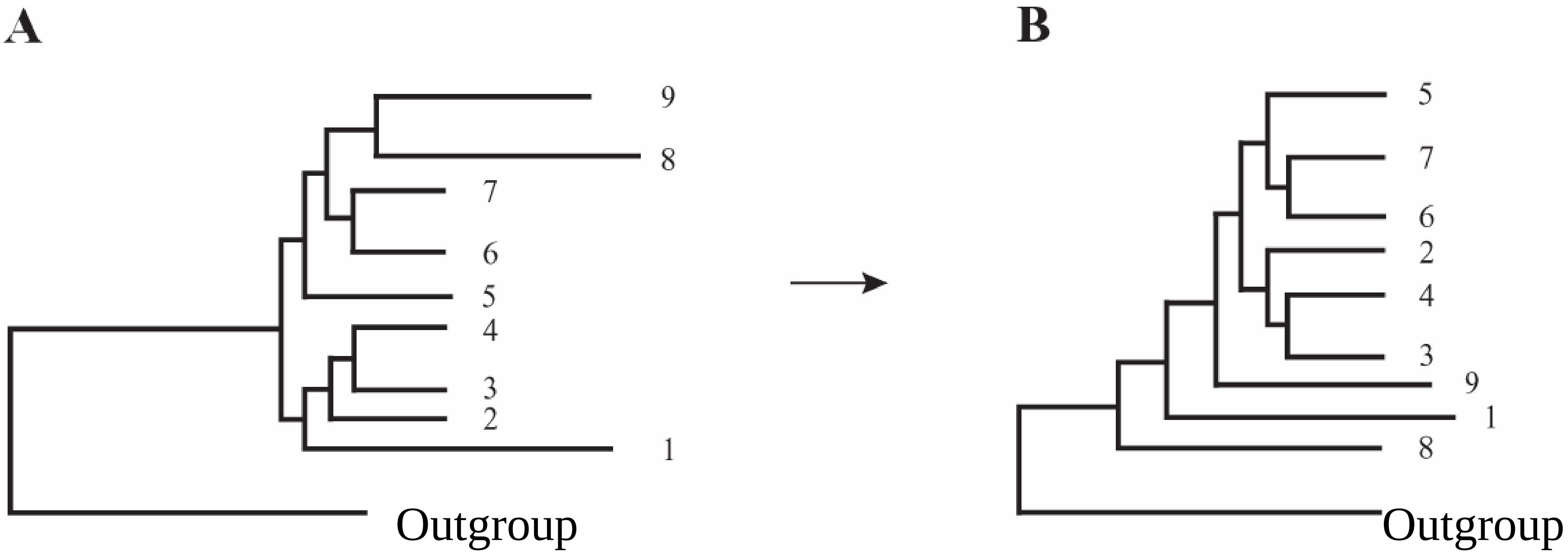
The branching order among these groups was misleading!

Because, long branch attraction (LBA) made the SSUrRNA tree misleading.

LBA— Most serious problem in phylogenetic analysis

Long Branch Attraction (LBA) artefact

: the most serious problem in phylogenetic analyses



Real tree

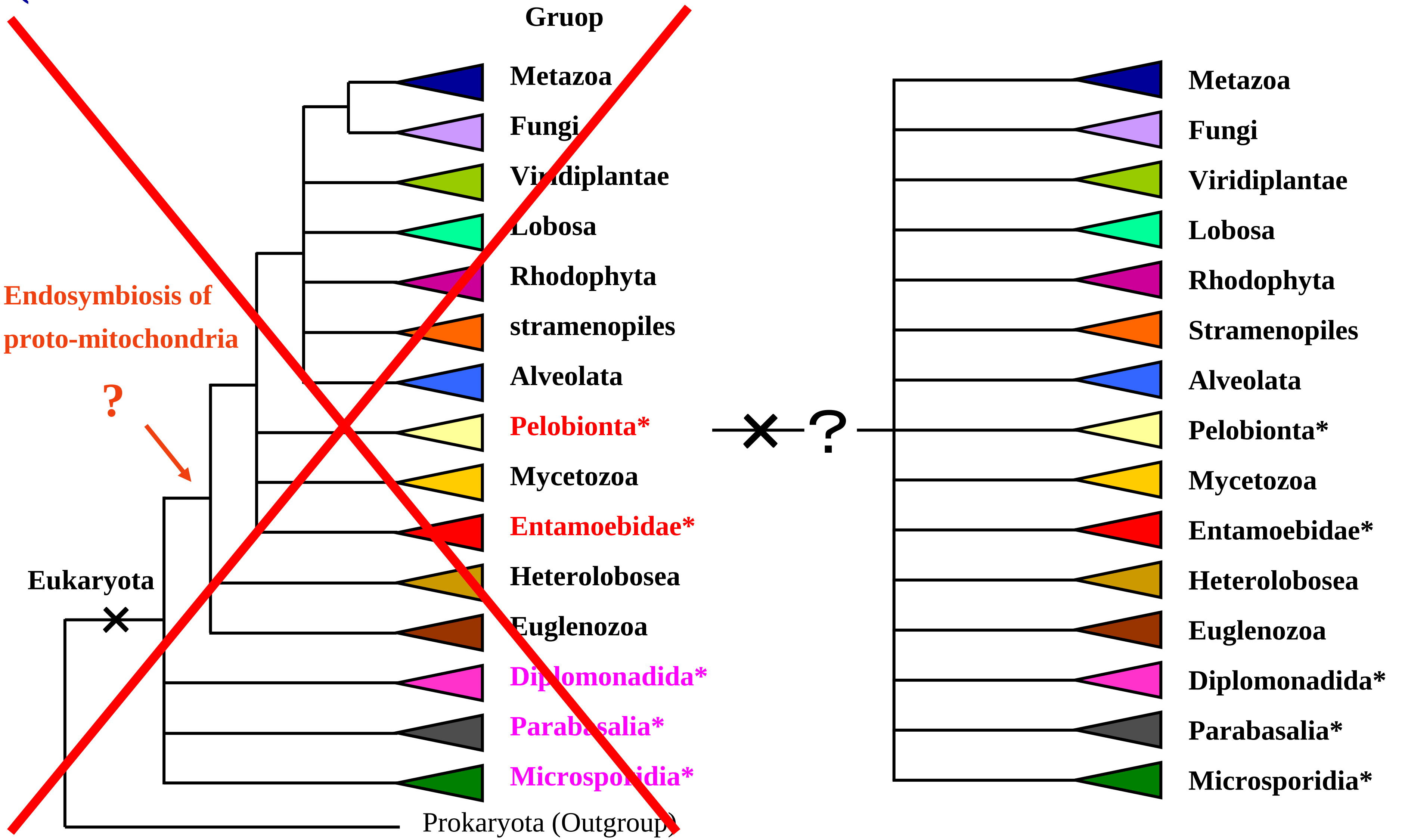
Based on this tree, sequence data at external nods are simulated.

Inferred tree based on the simulated sequence data

Long branches are artificially located at the base of the tree

SSUrRNA tree of Eukaryota (~mid 90’s)

（late 90’s）



* Amitochondriate protist group

No consensus on eukaryotic phylogeny !

Combined phylogeny of multi-genes (2000～)

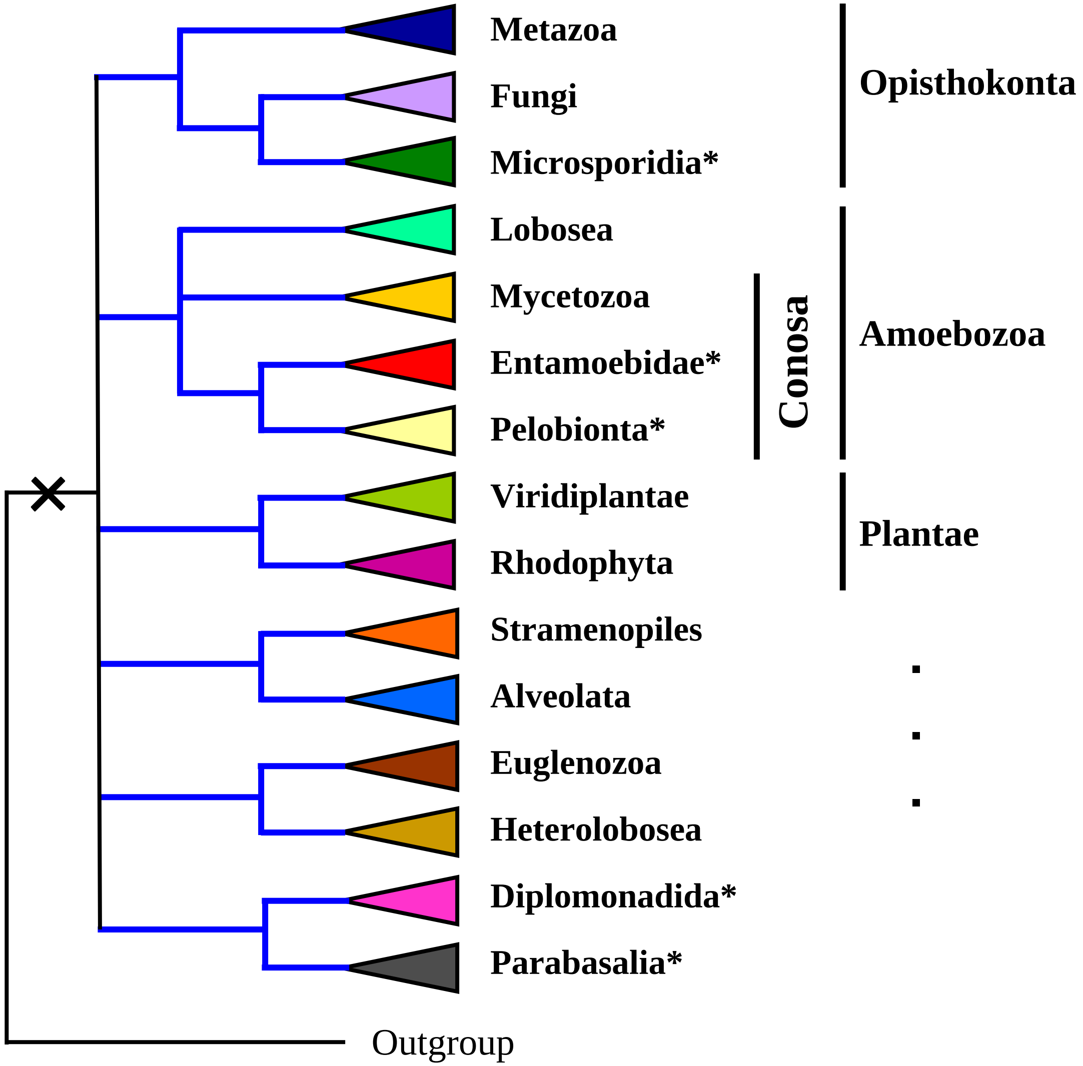
- To accumulate phylogenetic signals of individual genes

↓

Robust inference

Large groups
⇒ Supergroup

Large group ⇒ Supergroup



* Amitochondriate group