

# EVOLUTIONARY INFERENCE:

Some basics of phylogenetic analyses.

Ana Rojas Mendoza  
CNIO-Madrid-Spain.  
Alfonso Valencia's lab.

# Aims of this talk:

- 1.To introduce relevant **concepts of evolution** to practice phylogenetic inference from molecular data.
- 2.To introduce some of the most useful methods and computer programmes to practice **phylogenetic inference**.
- 
- 3.To show some examples I've worked in.

# 1-Concepts of Molecular Evolution

- **Homology vs Analogy.**
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.

# Owen's definition of homology

Richard Owen, 1843



- **Homologue:** the same organ under every variety of form and function (true or essential correspondence).
- **Analogy:** superficial or misleading similarity.

# 1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.

## Similarity $\neq$ Homology

- Similarity: mathematical concept
- **Homology: biological concept**
  - ◆ *Common Ancestry!!!*



70,2%?



# 1. Concepts of Molecular Evolution

- Homology *vs* Analogy.
- Homology *vs* similarity.
- **Orthologous *vs* Paralogous genes.**
- Species tree *vs* genes tree.
- Molecular clock.
- Allele mutation *vs* allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.

## Homologous genes

- **Orthologous genes**

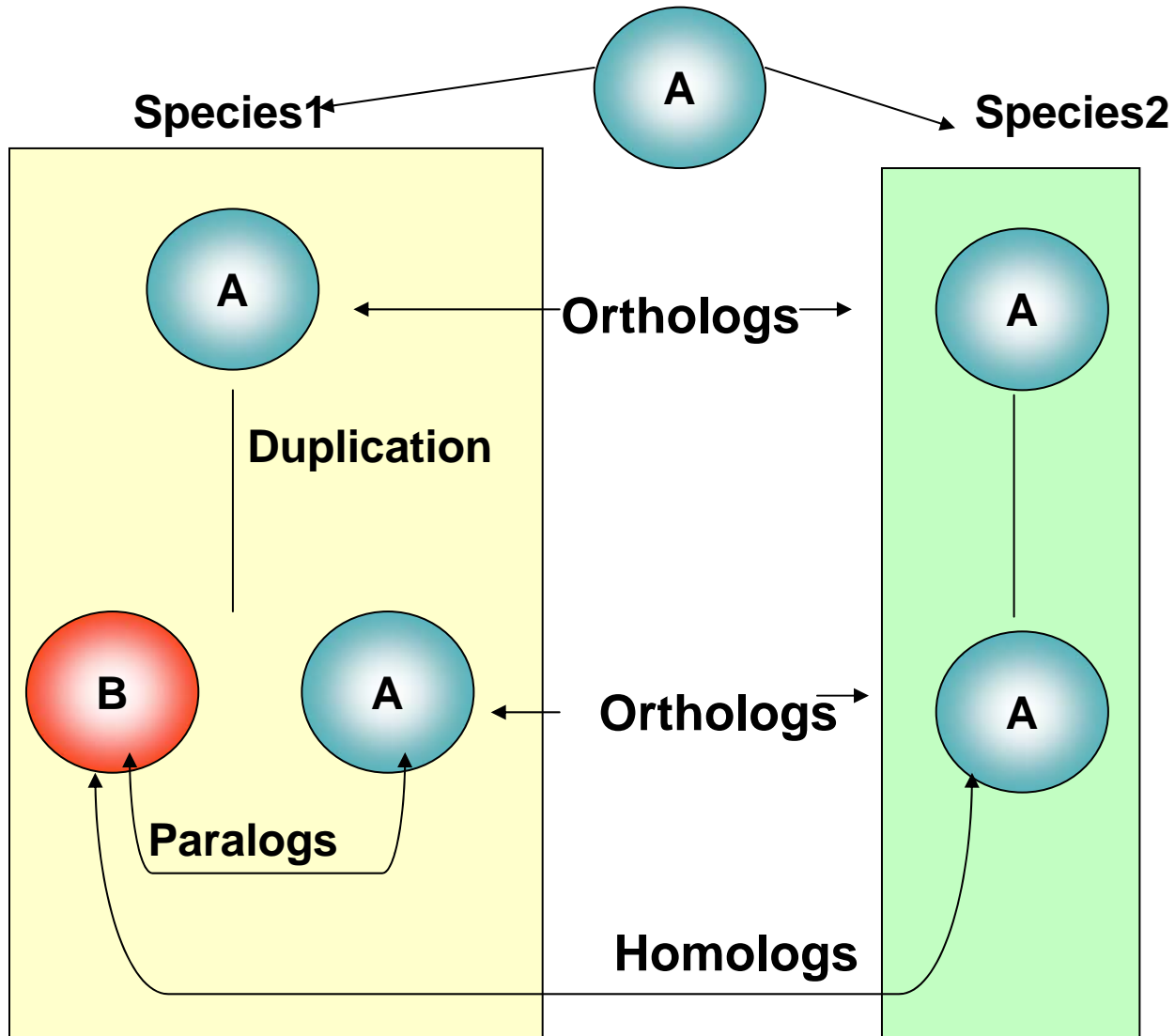
*Derived from a process of new species formation (speciation)*

- **Paralogous genes**

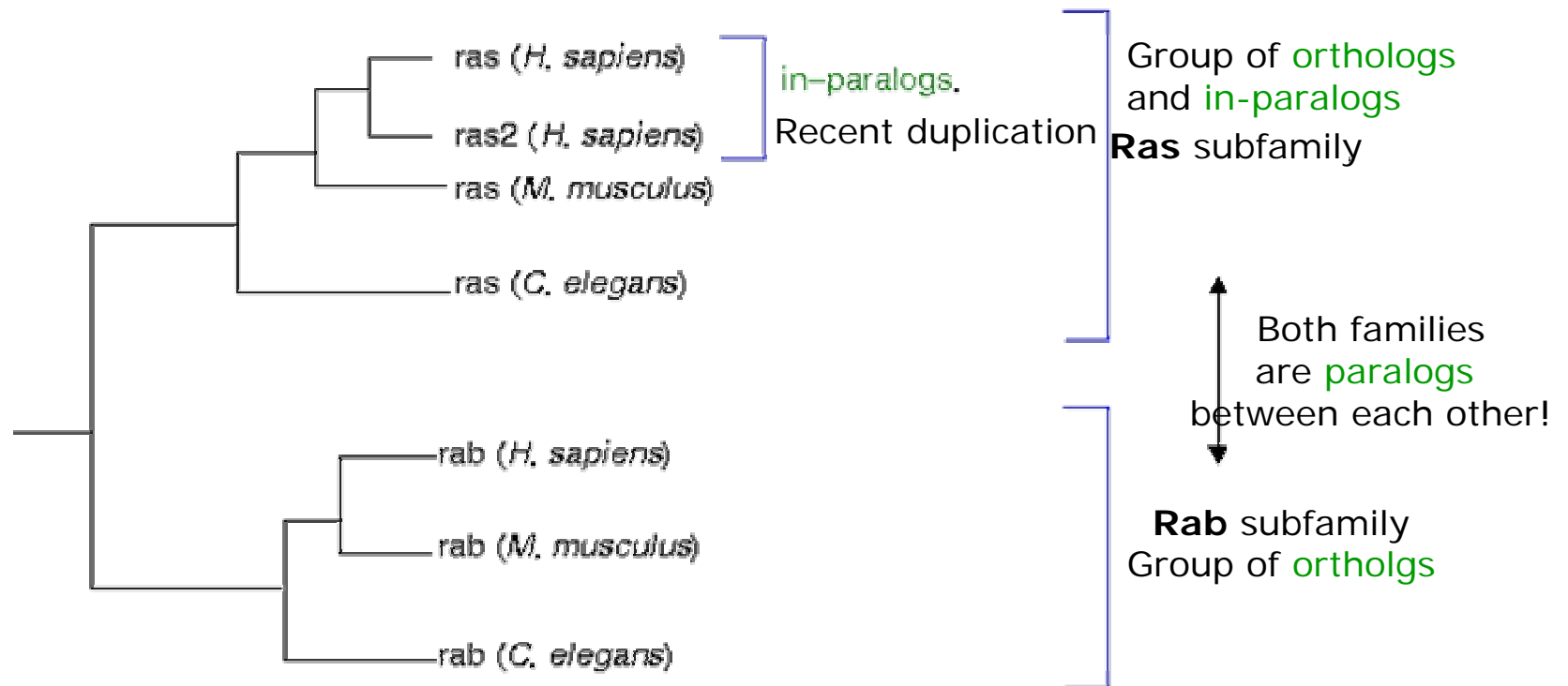
*Derived from an original gene duplication process in a single biological species*



# Homologous genes



# HOMOLOGS/ORTHOLOGS/PARALOGS



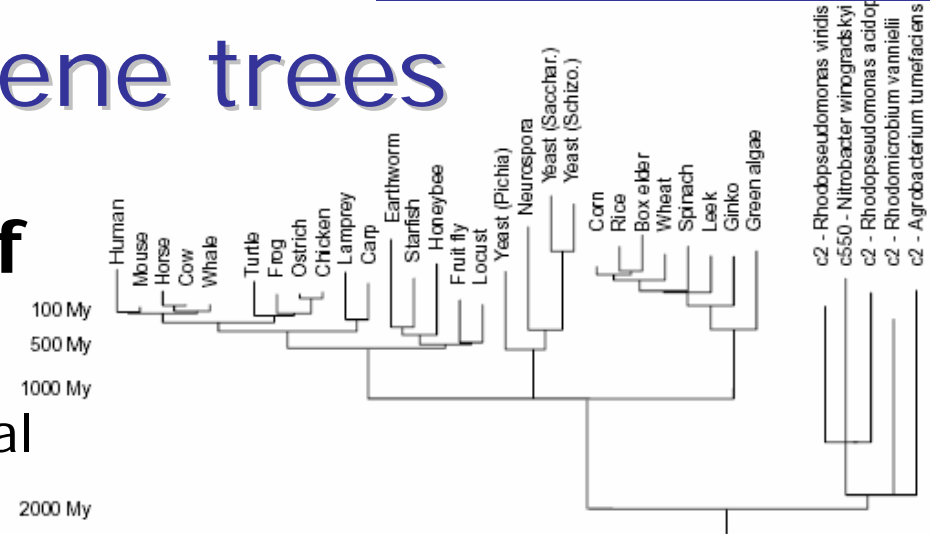
# 1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- **Species tree vs genes tree.**
- Molecular clock .
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.

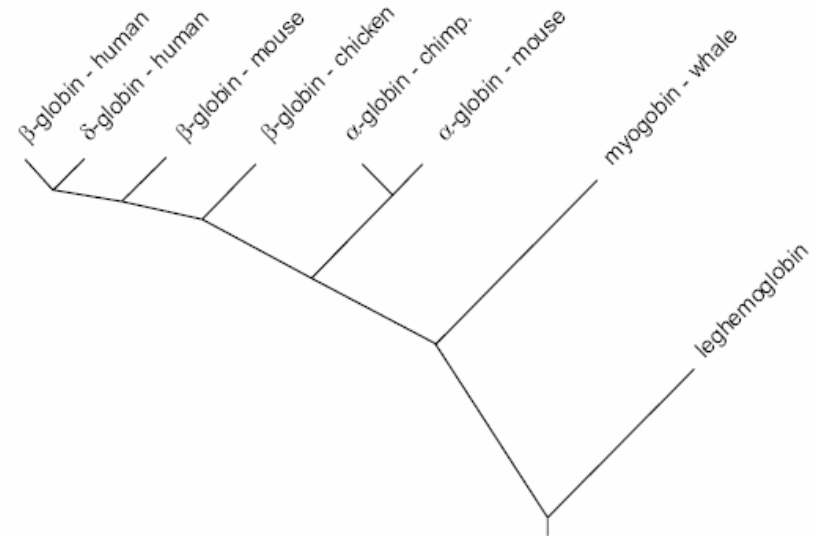
## Species trees vs Gene trees

### Orthologous genes of Cytochrome

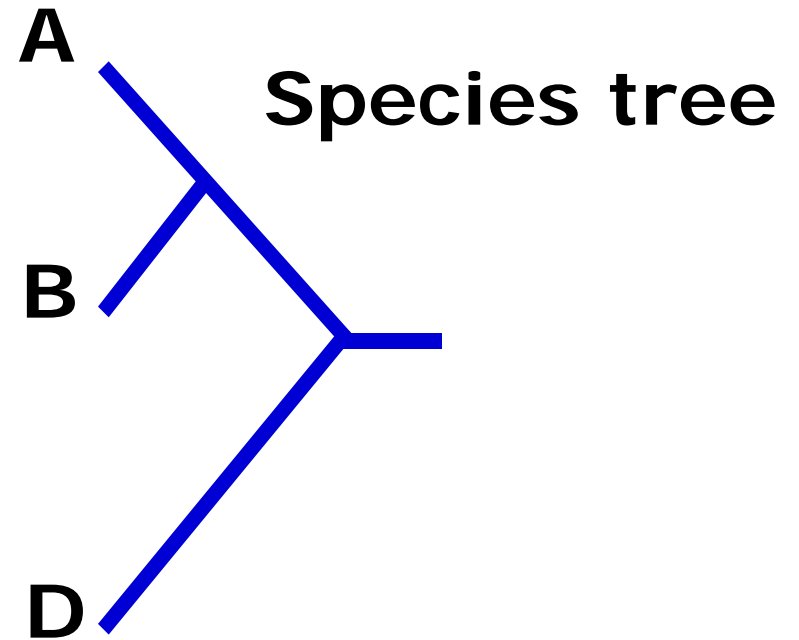
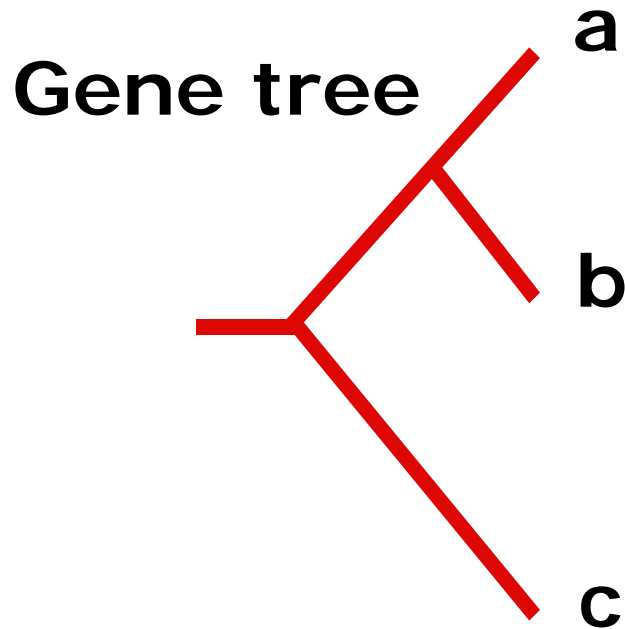
Each one is present in a biological species



- **Paralogous genes of Globin**
- **a, b, d (Glob), Myo y Leg haemoglobin, each originated by duplication from an ancestral gene**



# Species trees and Gene trees



**We often assume that gene trees give us species trees**

# 1. Concepts of Molecular Evolution

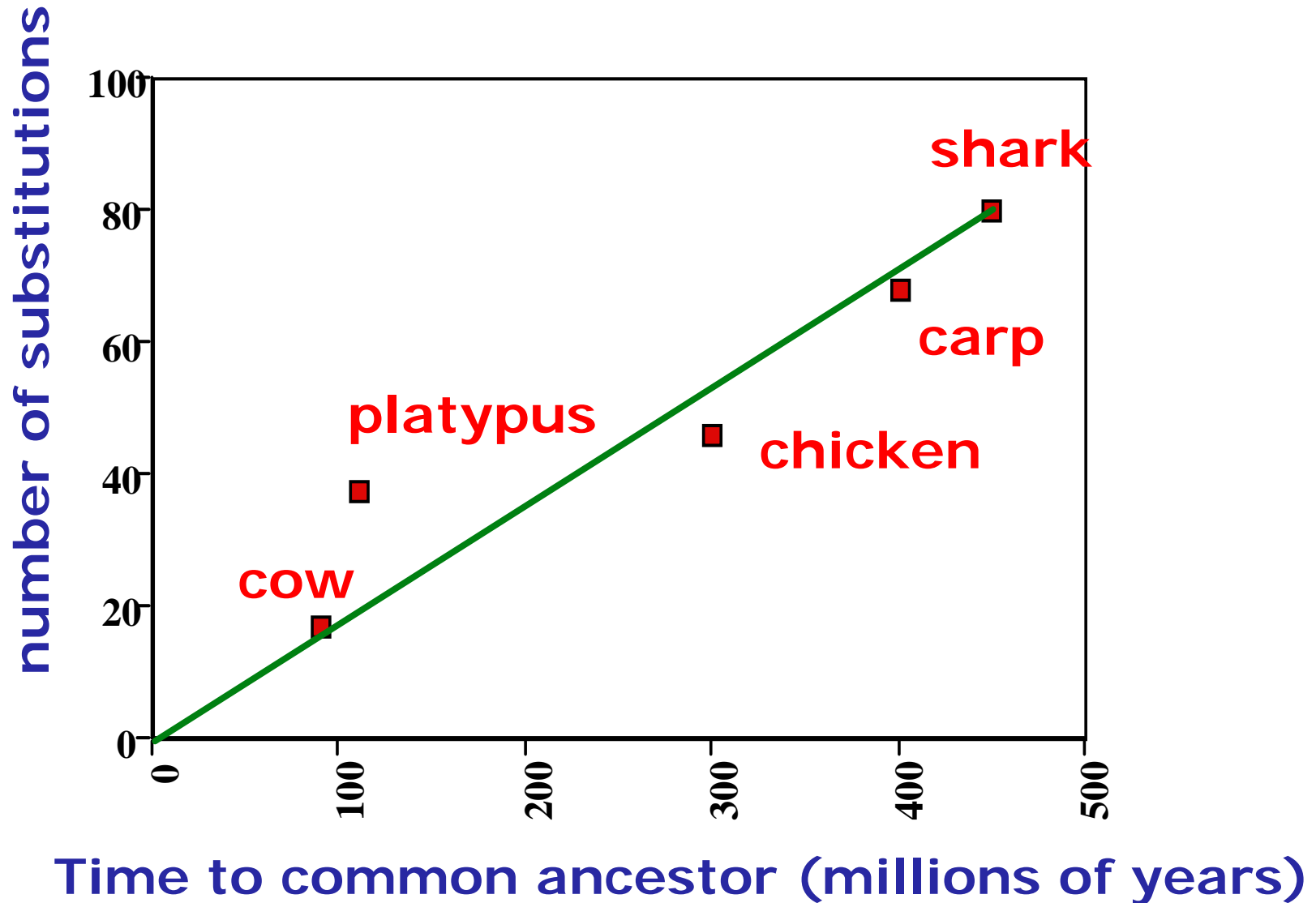
- Homology *vs* Analogy.
- Homology *vs* similarity.
- Ortologous *vs* Paralogous genes.
- Species tree *vs* genes tree.
- **Molecular clock.**
- Allele mutation *vs* allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.

## Is there a molecular clock?

- The idea of a molecular clock was initially suggested by Zuckerkandl and Pauling in 1962.
- They noted that **rates** of amino acid replacements in animal haemoglobins **were roughly proportional** to time- as judged against the fossil record.

# SOME BASICS

The molecular clock for alpha-globin:  
Each point represents the number of substitutions separating each animal from humans



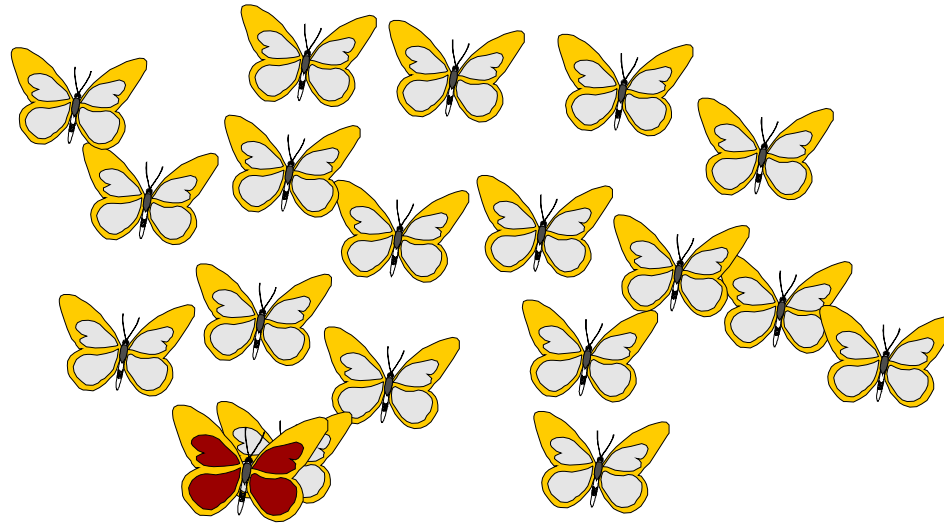


# 1. Concepts of Molecular Evolution

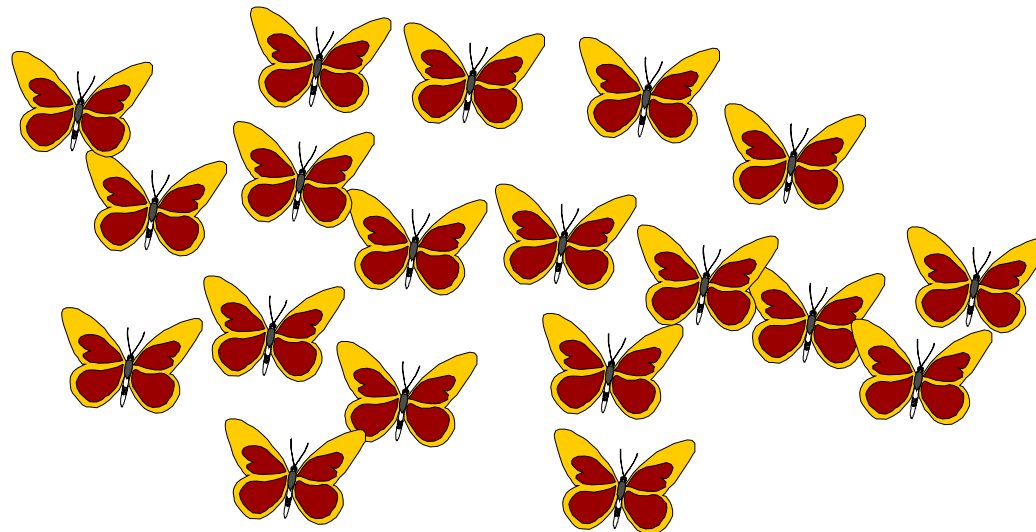
- Homology *vs* Analogy.
- Homology *vs* similarity.
- Ortologous *vs* Paralogous genes.
- Species tree *vs* genes tree.
- Molecular clock .
- **Allele mutation *vs* allele substitution.**
- Rates of allele substitution.
- Neutral theory of evolution.

## ALLELE MUTATION VS. FIXATION

**Mutation**  
=  
**Individual**



**Fixation**  
=  
**Population**



# 1. Concepts of Molecular Evolution

- Homology *vs* Analogy.
- Homology *vs* similarity.
- Ortologous *vs* Paralogous genes.
- Species tree *vs* genes tree.
- Molecular clock .
- Allele mutation *vs* allele substitution.
- **Rates of allele substitution.**
- Neutral theory of evolution.

## Rates of amino acid replacement in different proteins

<b>Protein</b>	<b>Rate (mean replacements per site per 10<sup>9</sup> years)</b>
<b>Fibrinopeptides</b>	<b>8.3</b>
<b>Insulin C</b>	<b>2.4</b>
<b>Ribonuclease</b>	<b>2.1</b>
<b>Haemoglobins</b>	<b>1.0</b>
<b>Cytochrome C</b>	<b>0.3</b>
<b>Histone H4</b>	<b>0.01</b>

- *Evolutionary rates depends on functional constraints of proteins*

## *SUBSTITUTION OR FIXATION RATES IN ESTIMATION*

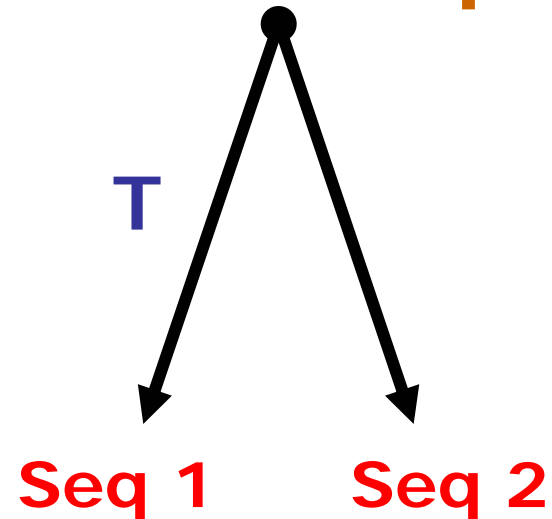
**Nucleotide substitution rate ( $r$ ):**  
**# substitutions per site per year**

$$r = K / (2T)$$

**K** = number of substitutions  
per site between  
homologous sequences.

**T** = Time of divergence.

**Ancestral sequence**



# 1. Concepts of Molecular Evolution

- Homology *vs* Analogy.
- Homology *vs* similarity.
- Ortologous *vs* Paralogous genes.
- Species tree *vs* genes tree.
- Molecular clock.
- Allele mutation *vs* allele substitution.
- Rates of allele substitution.
- **Neutral theory of evolution.**
- Homoplasy.

## Neutral theory of evolution

- At **molecular level**, the most frequent change are those involving **fixation in populations** of neutral selective variants (Kimura, 1968).
  - Allelic variants are functionally equivalent.
  - Neutralism does not deny adaptive evolution.
- Fixation of new allelic variants occur **at a constant rate**, it is equal to mutation rate and independent of population parameters.

**mutation in population**

**probability to fix**

$$- \quad 2 N m \times 1/2 N = m$$

## There is no universal clock

- ~~The initial proposal saw the clock as a Poisson process with a constant rate~~
- Now known to be more complex - differences in rates occur for:
  - different sites in a molecule
  - different genes
  - different base position (synonymous-nonsynonymous)
  - different regions of genomes
  - different genomes in the same cell
  - different taxonomic groups for the same gene
- Molecular Clocks Not Exactly Swiss



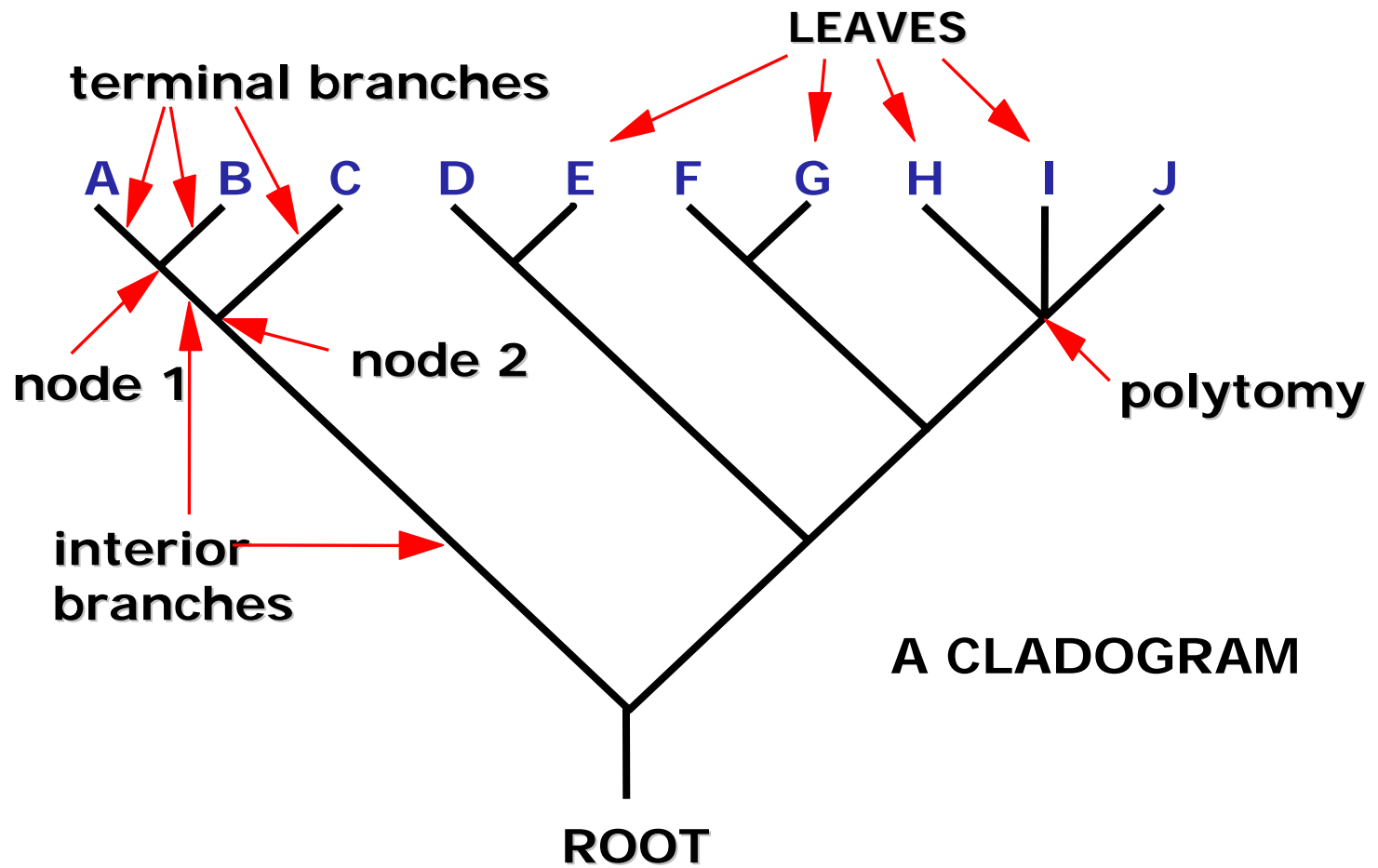
## 2. Concepts of Phylogenetic Systematics

- **What is Phylogenetic systematics?**
- Cladogram and Phylogram.
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.

# Phylogenetic systematics

- Sees **homology as evidence** of common ancestry
- Uses tree diagrams to portray relationships based upon recency of common ancestry
- Monophyletic groups (clades) - contain species which are more closely related to each other than to any outside of the group

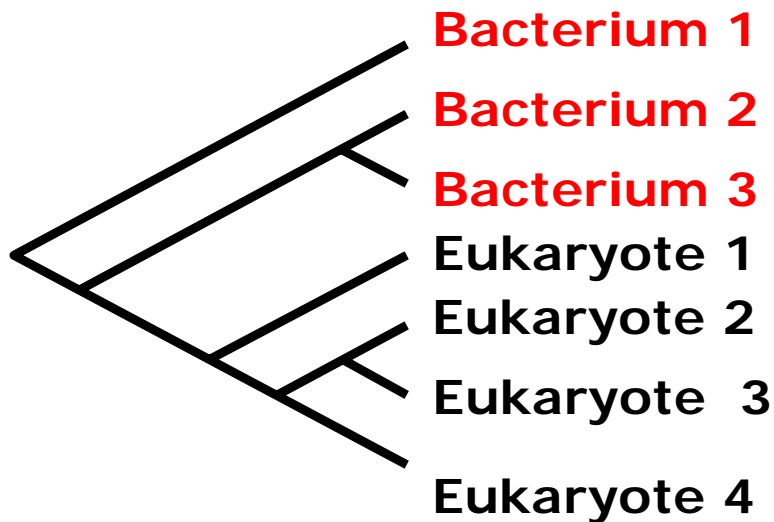
# Phylogenetic Trees



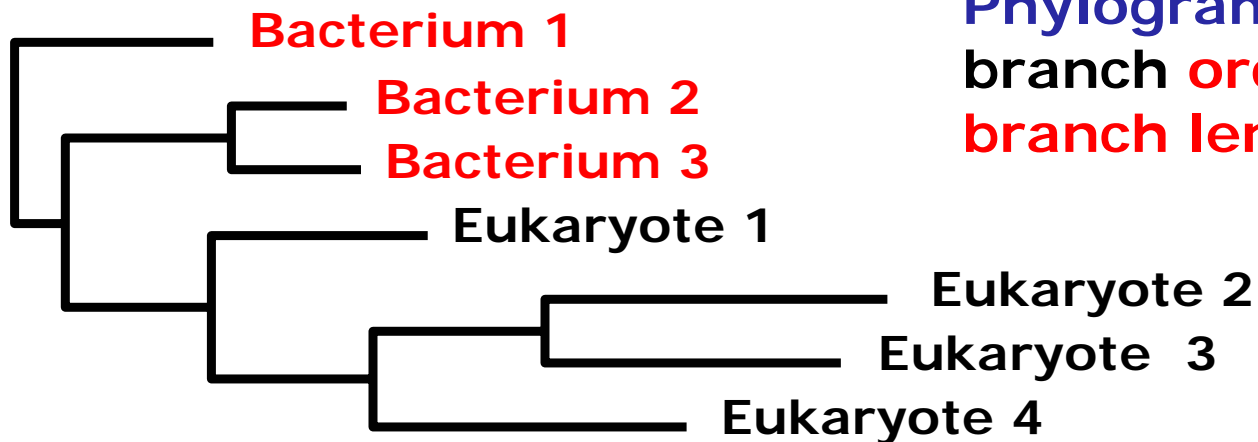
## 2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- **Cladogram and Phylogram.**
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.

# Cladograms and phylograms



**Cladograms** show branching **order** - branch lengths are meaningless

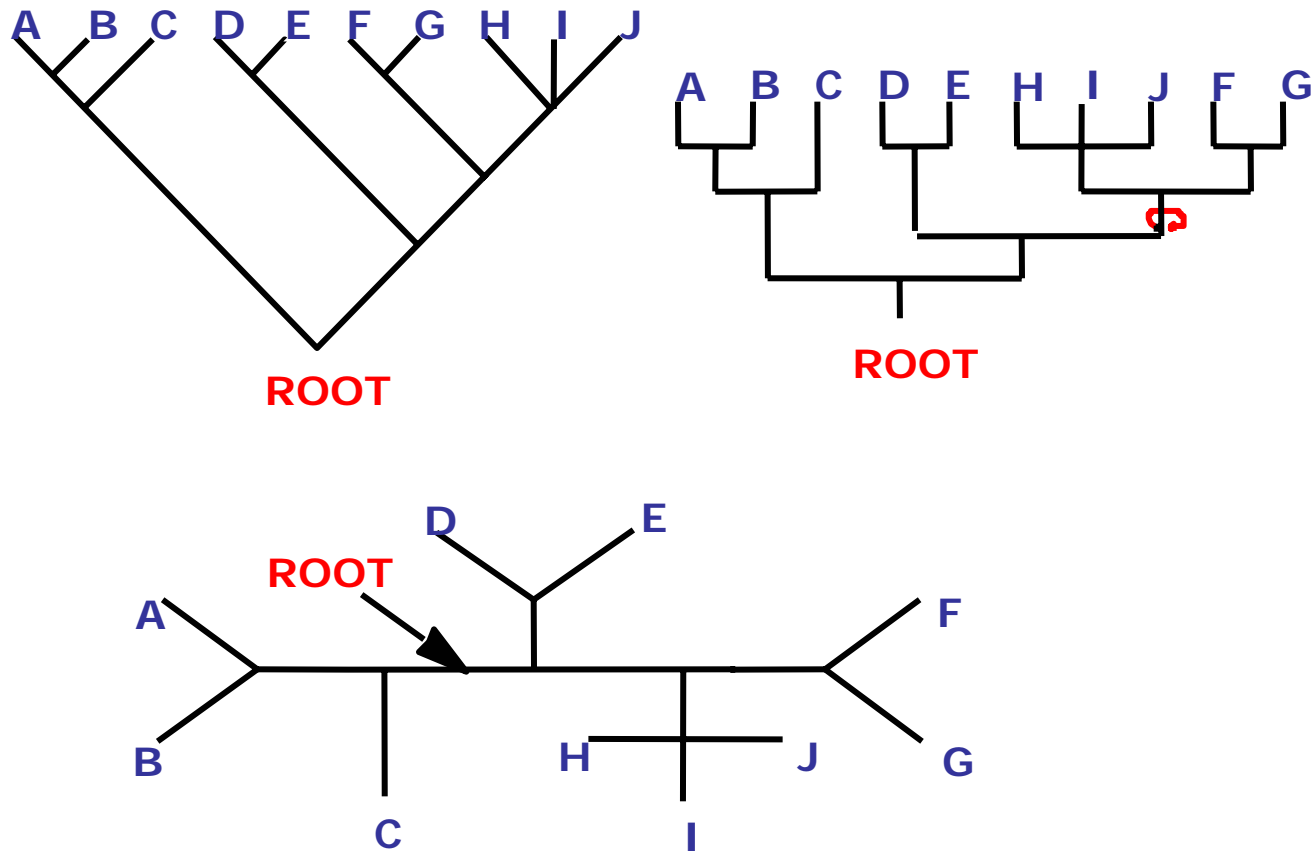


**Phylograms** show branch **order and branch lengths**

## 2. Concepts of Phylogenetic Systematics

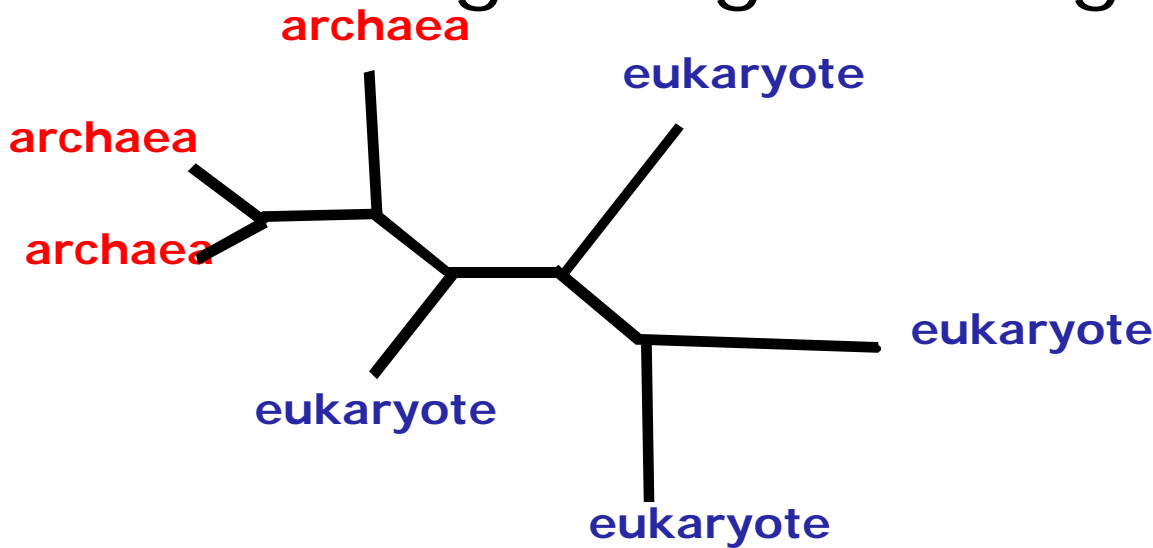
- What is Phylogenetic systematics?
- Cladogram and Phylogram.
- **Monophyletic, Paraphyletic and Polyphyletic groups.**
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.

# Trees - Rooted and Unrooted



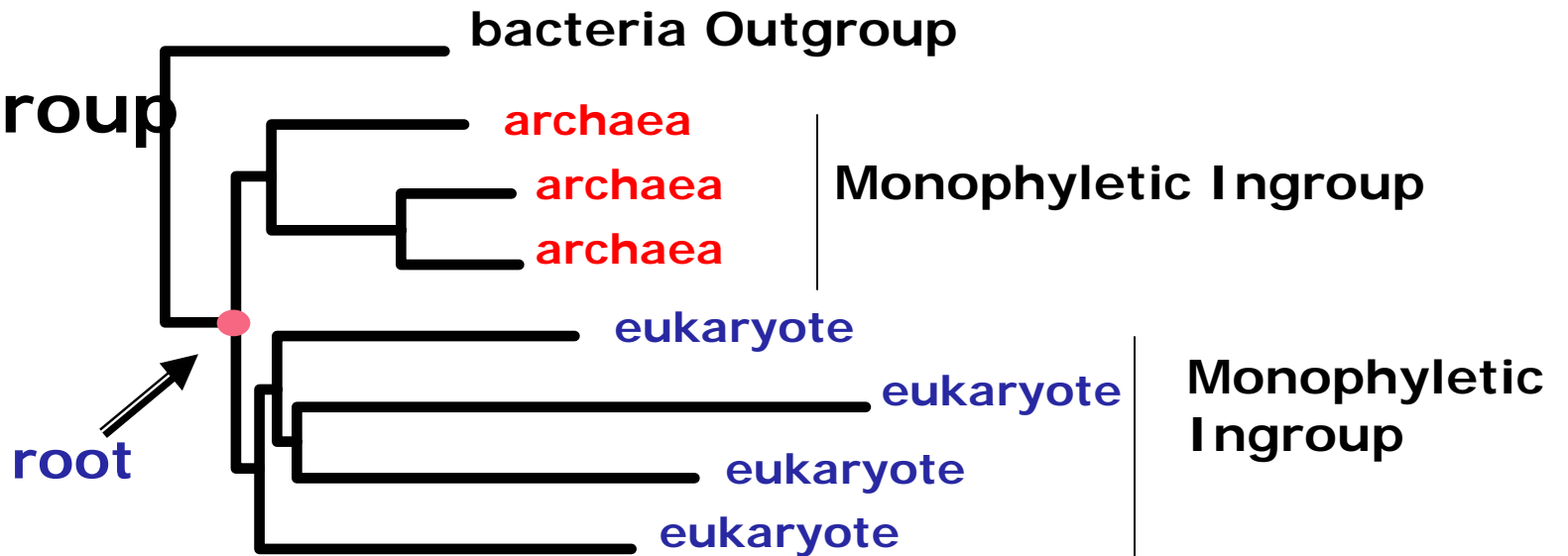
*SOME BASICS*

# Rooting using an outgroup



**Unrooted tree**

## Rooted by outgroup





## 2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- **Character states and evolution.**
- Homoplasy.

# Character:

A descriptor that can have different manifestations in different species. (character states)

## Types of characters

- **Morphological (characteristics of physical attributes).**
- **Behavioral.**
- **Ecological (nest type, host plant, prey type).**
- **Distributional (geographical).**
- **Physiological/chemical .**
- **Molecular.**

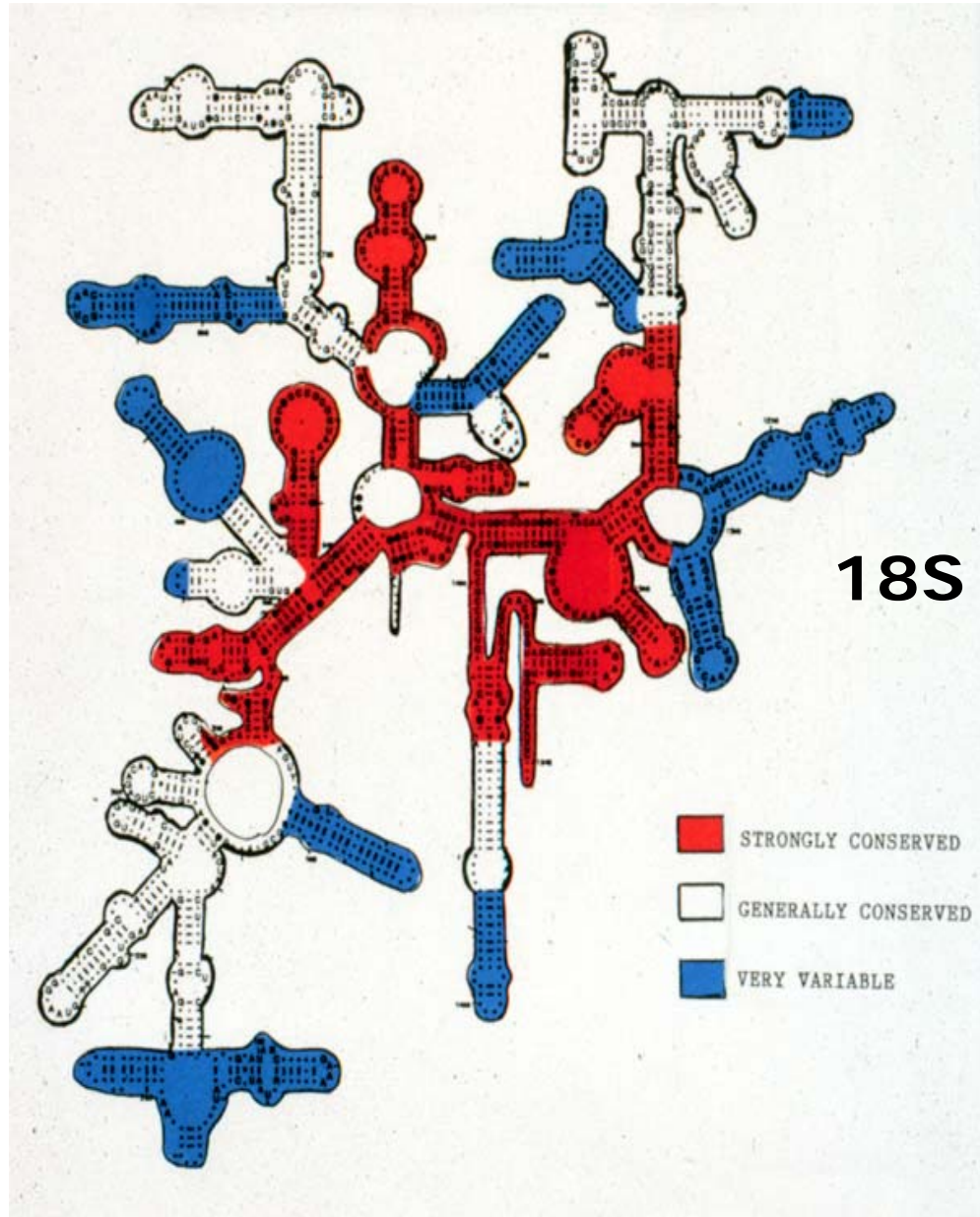
# Character evolution

- Heritable changes (in morphology, gene sequences, etc.) produce different character states.
- Similarities and differences in character states provide the basis for inferring phylogeny (i.e. provide evidence of relationships)
- The utility of this evidence depends on **how often** the evolutionary changes that produce the different character states occur independently.

## Why to use molecular data?

- Molecular data are genetic data:  $V_p = V_g + V_e$
- Molecular data led us to study a huge amount of characters.
- Any kind of homoplastic similarity vanishing at time more characters are considered.
- Indels, duplications and chromosomic rearrangements are rare events with strong weight of homology.
- Molecular data offers a common measure for evolutionary divergence.

Small subunit ribosomal RNA



**18S or 16S rRNA**

# *Molecular characters*

## **1. Protein variation** (1950s-present)

Historically, first molecular characters

### **a.** Isozyme/allozyme variation

- Used mostly at population level, sometimes Phylogenetic.
- Misses lots of underlying variation

### **b.** Amino acid sequencing (1960s, Fitch, etc.)

- Globin genes
- Technically difficult

# *Molecular characters*

## 2. DNA (1970s)

- Has dominated molecular phylogenetics since.
  - Impact of polymerase chain reaction (PCR).
- a. DNA-DNA hybridization (1970s-80s; rare now)
- Famous studies in birds (Sibley and Ahlquist) made some big changes (birds infamous for lack of allozyme variation)
  - **Not character-based**; data are pairwise comparisons between OTUs (suitable only for distance analysis)
  - Advantage of looking at entire genome (single copy DNA anyway)

## Some Common Phylogenetic Methods

		Types of Data	
		Distances	Sites (nucleotides, aa)
Tree building method	Cluster Algorithms	UPGMA NJ	
	Optimality Criteria	Minimum Evolution Least Square	Parsimony Maximum Likelihood Bayesian Inference



## 2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- **Homoplasy.**

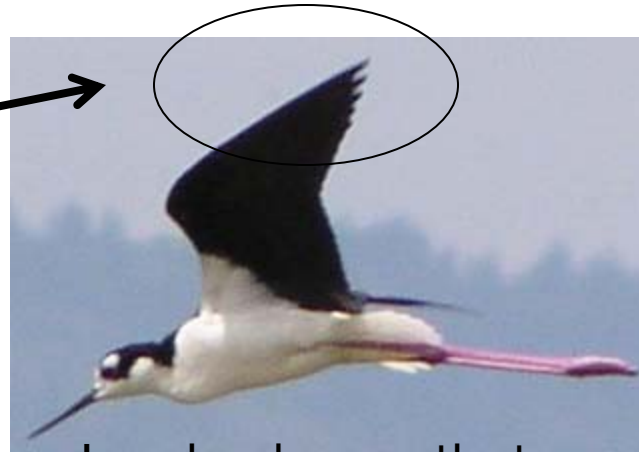
# Homoplasy

## SOME BASICS

- Convergent evolution: similarity due to adaptation, Not to common ancestry!



Involve bones that  
In human will make  
The hands



Involve bones that  
In human will make  
The arms

Both lineages had a huge evolutionary separation before  
They came fliers! They independently became fliers!

Human eye and squid (calamari!) eye...

# PHYLOGENETIC ANALYSIS

**REMINDER!!!**

Trees: **cladograms**- represents only the branching order of nodes  
**phylograms**-represents branching order and branch length  
(number of sequence changes between nodes)

Distance: number of substitutions that have taken place along a branch

## Tree construction:

**algorithmic**: uses an algorithm to construct a tree from data  
(NJ,UPGM: distance methods) Fast, one tree ONLY.

**tree-searching**: builds many trees and then uses a criterion to decide which is the best tree. (Character based)

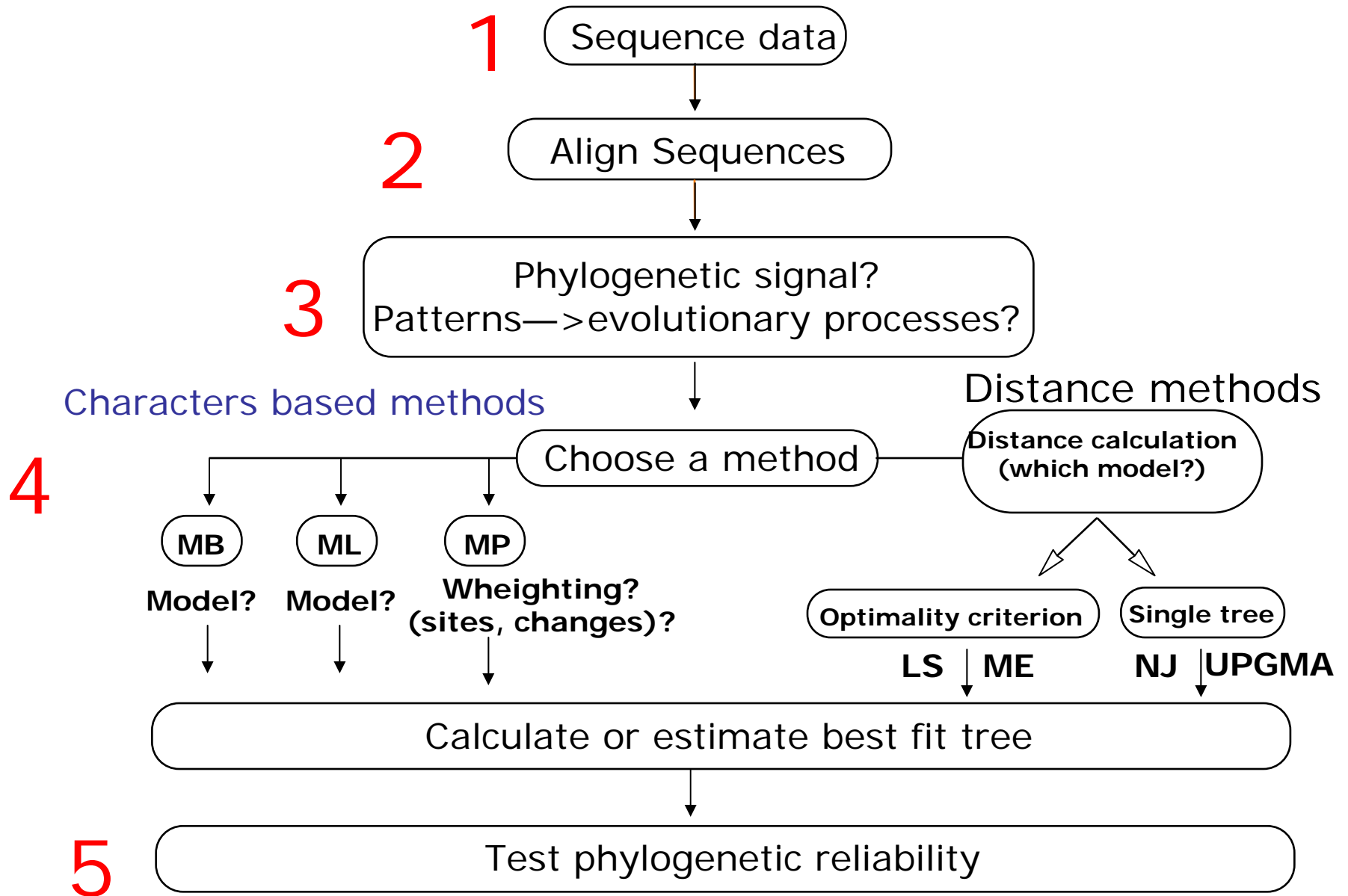
**Parsimony**: several trees. The most likely scenario involves the fewest changes?

**ML**: seeks for tree that maximizes the likelihood of observing data.

**Bayesian**: seeks from several trees with the greatest likelihoods given the data.

# PHYLOGENETICS DANCING!

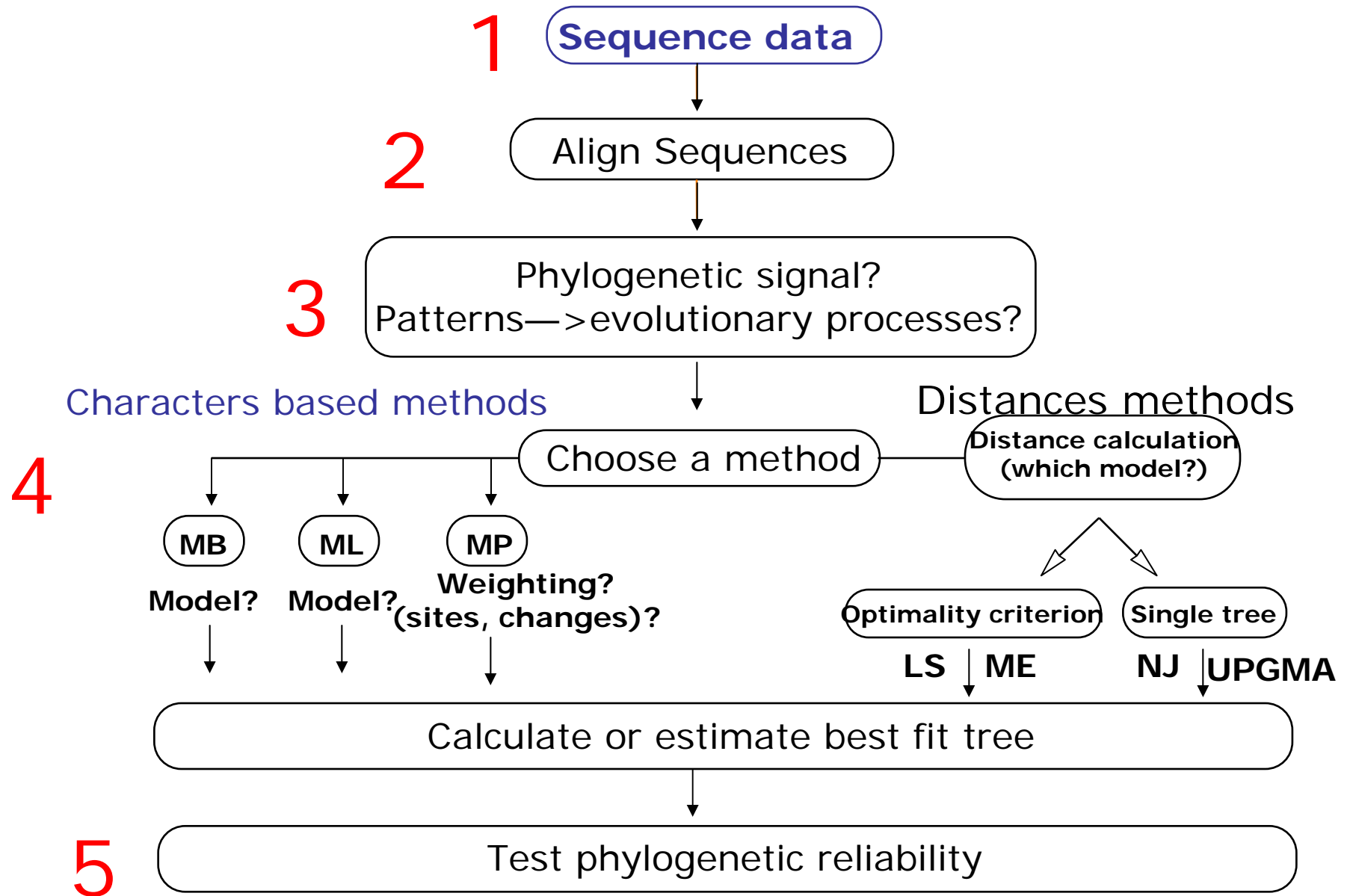
# The five steps in phylogenetics dancing



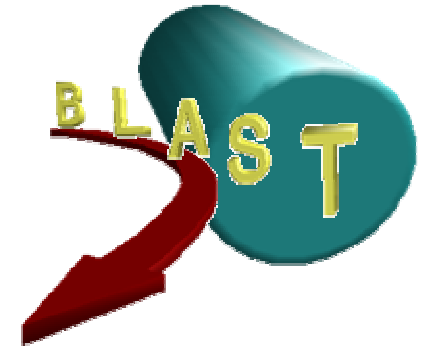


# SEARCHING DB

## The five steps in phylogenetics dancing



# SEARCHING DB



## SEARCHING THE DATABASES

### Pairwise

#### Searching : FASTA

(Lipman & Pearson, 1985, Pearson & Lipman 1988)

#### Basic Local Alignment Search Tool (BLAST)

Altschul, S.F., Gish W., Miller W., Myers E.W., and Lipman D.J.  
J. Mol. Biol. (1990) 215:403-10.

### Profile

#### PSI-BLAST:

Altschul, S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J.  
Nucleic Acids Research (1997) v.25, n.17 3389-3402

### *Iterative search*

**USE of point position specific matrices.**

**Use the matrix to search again!**



# SEARCHING DB

## WHY SEARCHING THE DATABASES?

We want to obtain **all the sequences** related to our query!

OKAY, **but which kind of sequences?**

Am I looking for distant homologs?    → PSI-BLAST

Am I looking for clear orthologs?    → FASTA, BLAST

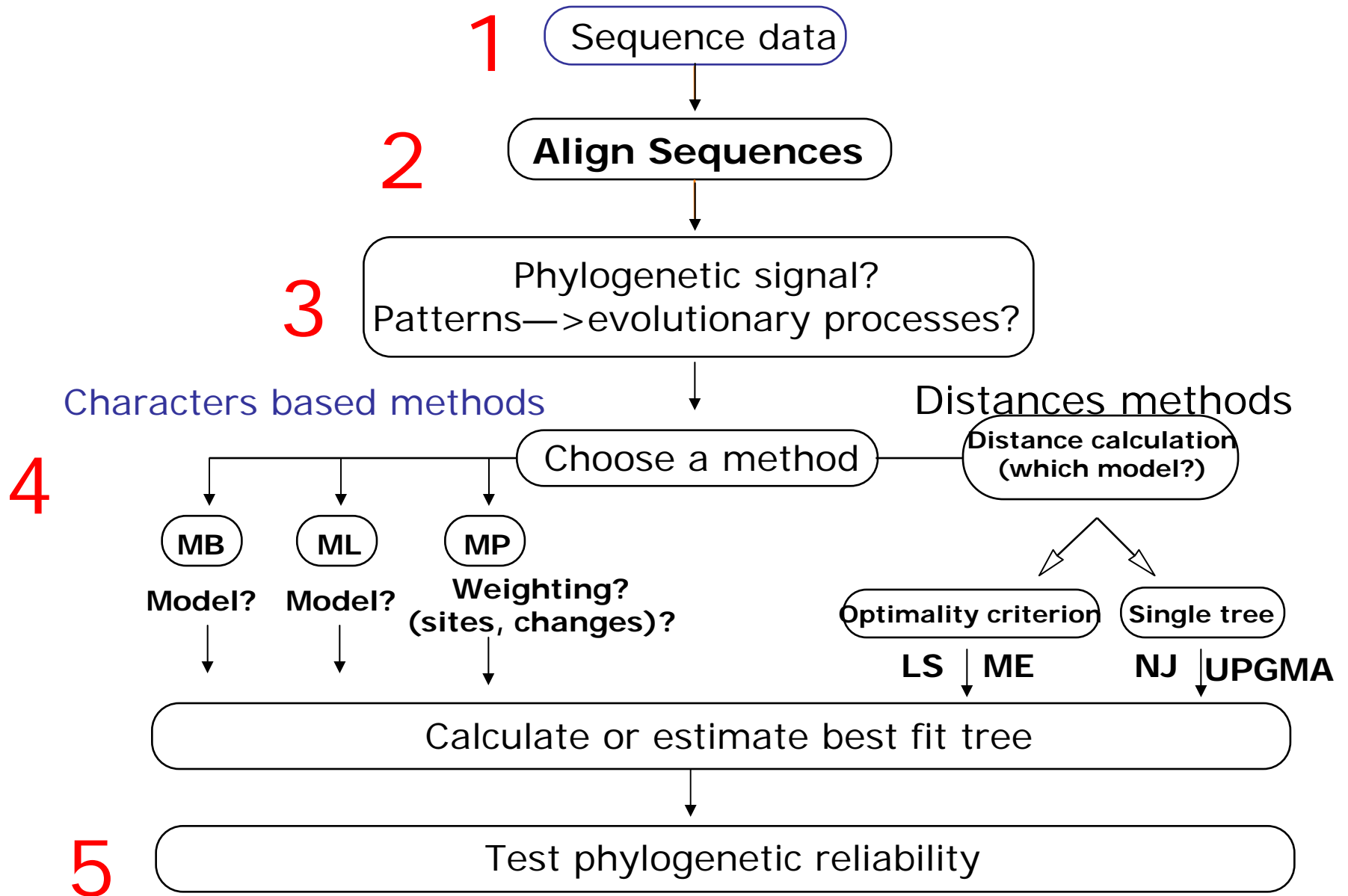
**How many sequences should I use?** Well depends:

**Rule of the thumb:**

Get a representative set of your sequences,  
remove redundancy at lets say 80%!

# The five steps in phylogenetics dancing

## ALIGNING DB



## ALIGNING THE SEQUENCES

We want to align **all the sequences** obtained via searching the databases.

Alignment quality is **CRUCIAL**=> bad alignment=bad tree!!!!

### METHODS:

Greedy approaches: **Progressive alignment** (Feng & Doolittle, 1987-96)

PILEUP, **ClustalW** (improved the Progressive alignment) .. Too greedy! poor when %id <30%

•Then the real improvements:

**T-COFFEE** (Notredame et al, 2000): incorporates local and global information!

**ProbCons**(Do, CB, et al, 2005): like T-Coffee with probabilistic estimations!

# ALIGNING DB

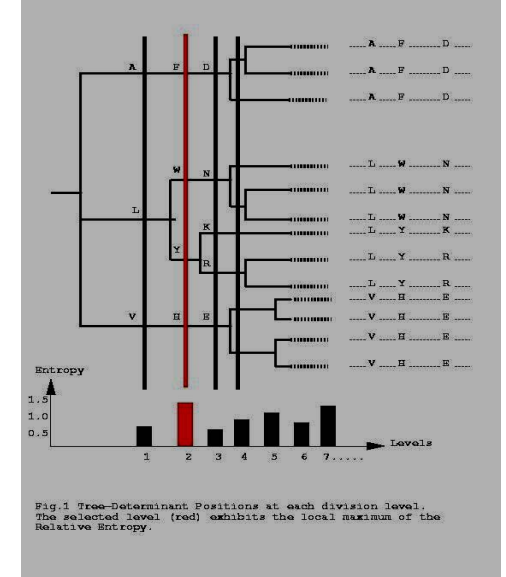
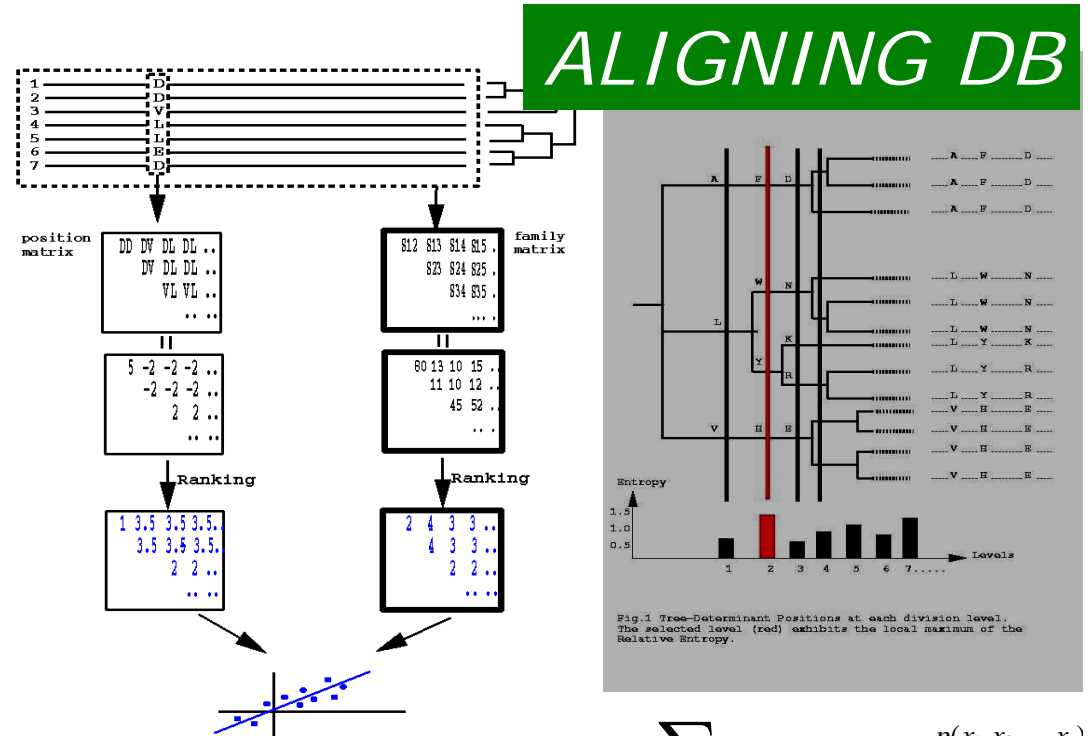
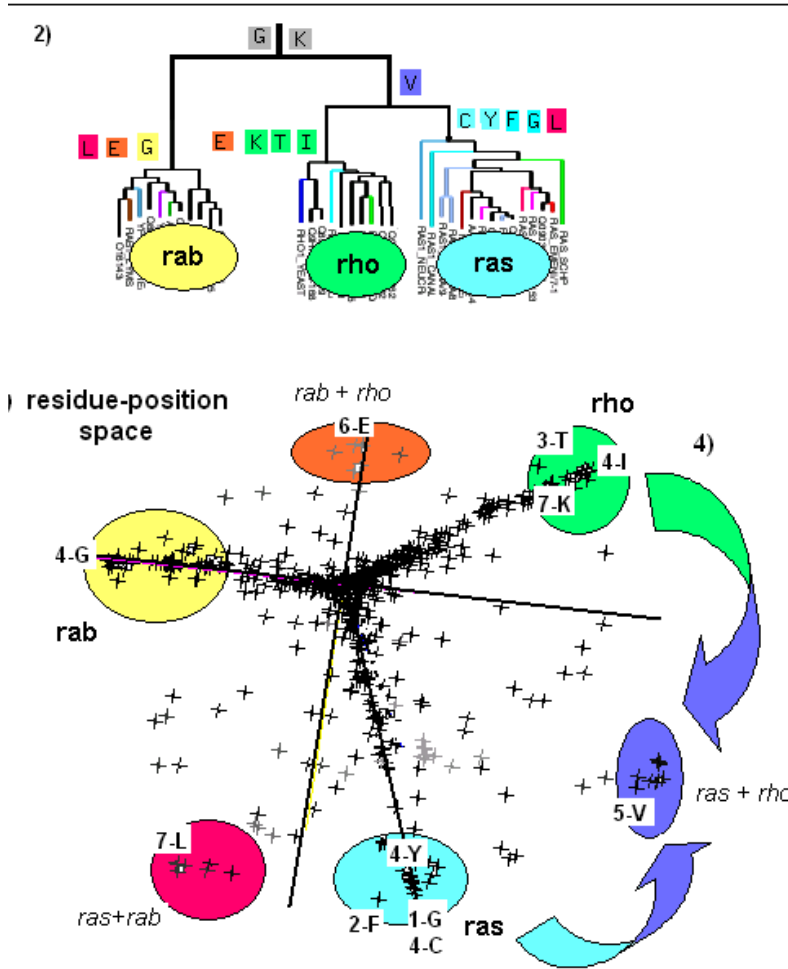
## WHAT CAN I LEARN FROM MY ALIGNMENT?

RASH_HUMAN	5	164	KLWVVGAGGGVVK	AL	TIQLI	NI	FVDE	DA	ET	SYRKQVWIDGETCL	LDI	LOT	AGQE	EYS
RRAS_HUMAN	1	160	KLWVVGGGGVVK	AL	TIQFI	SI	FVSD	DA	ET	SYTKICSVGGIPAR	LDI	LOT	AGQE	EFG
RTC1_HUMAN	1	160	RLWVVGGGGVVK	AL	TIQFI	SI	FVTD	DA	ET	SYTKQCVIDDRAAR	LDI	LOT	AGQE	EFG
RAS2_HYDMA	1	160	KLWVVGGGGVVK	AL	TIQFI	SI	FVQD	DA	ET	SYRKQCVIDDKVAH	LDI	LOT	AGQE	EF
RAS2_DROME	1	160	KLWVVGGGGVVK	AL	TIQFI	SI	FVTD	DA	ET	SYTKQCNIDDVPAK	LDI	LOT	AGQE	EF
RASL_NEUCR	1	160	KLWVVGGGGVVK	CL	TIQLI	NI	FLDE	DA	ET	SYRKQCTIDNEVAL	LDI	LOT	AGQE	EYS
RASL_MOUSE	1	159	KLWVVGAGGGVVK	AL	TIQLI	NI	FVDE	DA	ET	SYRKQVWIDGETCL	LDI	LOT	AGQE	EYS
RAS1_YEAST	1	160	KIVVVGGGGVVK	AL	TIQFI	SI	FVDE	DA	ET	SYRKQVWIDDKVSI	LDI	LOT	AGQE	EYS
RAS_SCHPO	1	160	KLWVVGGGGVVK	AL	TIQLI	SI	FVDE	DA	ET	SYRKKCEIDGEGAV	LDL	LOT	AGQE	EYS
RAS_LENED	1	160	KLWVVGGGGVVK	AL	TIQFI	SI	FVDE	DA	ET	SYRKQCVIDDEVAL	LDV	LOT	AGQE	EYG
RAS2_RHIRA	1	160	KIVMVGGGGVVK	AL	TIQFI	SI	FVDE	DA	ET	SYRKQCLIDSECAM	LDI	LOT	AGQE	EYS
RALA_HUMAN	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RALA_RAT	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RALB_HUMAN	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RALB_RAT	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RALB_XENLA	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RAL_DISOM	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
RALA_DROME	1	157	KVIMVGS GG VVK	AL	TLQFM	DI	FVED	EA	ET	KALSYRKKWVLDGEEVQ	IDI	LOT	AGQE	EYA
CE1393944	1	157	QVIMVGTGGVVK	AL	TLQFM	DI	FVEE	EA	ET	KALSYRKKWVLDGEECS	IDI	LOT	AGQE	EYS
CC42_DROME	1	156	KCVVVG DG AVVK	CL	TIQLI	NI	FPSE	VA	ET	VFENYAVTVMIGGEPYT	LGL	FOT	AGQE	EYD
RH04_YEAST	1	155	KIVVVG DG AVVK	CL	TIQLI	NI	FPTE	IA	ET	VFENYVTNIEGPNQIE	LAL	WOT	AGQE	EYS
RH02_SCHPO	1	156	KLWVVG DG AC GK	SI	TIQLI	NI	FPTE	VA	ET	VFENYVSDCRVDGKSVQ	LAL	WOT	AGQE	EYE
RH02_YEAST	1	156	KLVIIG DG AC GK	SI	TIQLI	NI	FPEQ	HA	ET	VFENYVTDCRVDGKSVL	TL	WOT	AGQE	EYE
RH08_HUMAN	1	156	KIVVVG DS QC GK	AL	LHVFA	DI	FPEN	VA	ET	VFENY TASFEIDTQRIE	L	SLWOT	AGSPY	YD
RH06_HUMAN	1	156	KLVLVGD VQC GK	AL	LQVLA	DI	YPET	VA	ET	VFENY TACLETEEQRVE	L	SLWOT	AGSPY	YD
RH03_YEAST	1	156	KIVILG DG AC GK	SI	TIQLI	NI	FPEV	EA	ET	VFENYIHDIFVDSKHIT	L	SLWOT	AGQE	EFD
RH01_SCHPO	1	156	KLVIIG DG AC GK	CL	TIQLI	NI	FPEV	VA	ET	VFENYVADVEVDGRHVE	L	ALWOT	AGQE	EYD
RH01_YEAST	1	156	KLVIIG DG AC GK	CL	TIQLI	NI	FPEV	VA	ET	VFENYVADVEVDGRRVE	L	ALWOT	AGQE	EYD
RH01_ENTHI	1	151	KIVVVG DG AVVK	CL	TIQLI	NI	IPTA	VA	ET	VFENYFSHVMKYKNEEFI	L	DLWOT	AGQE	EYD
RH0L_DROME	1	156	KITIVG DG GM VK	CL	TIQLI	NI	FPEE	IA	ET	VFENYHACNIAVDORDYN	L	TLWOT	AGQE	EYE

Correlated mutation

Tree-determinant

conserv



$$H(n) = \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log \frac{p(x_1, x_2, \dots, x_n)}{\prod_{i=1}^n p(x_i)}$$

Relative entropy cut,  
del Sol, Valencia 2002

Casari, Sander, Valencia *Nature Str. Biol.* 95

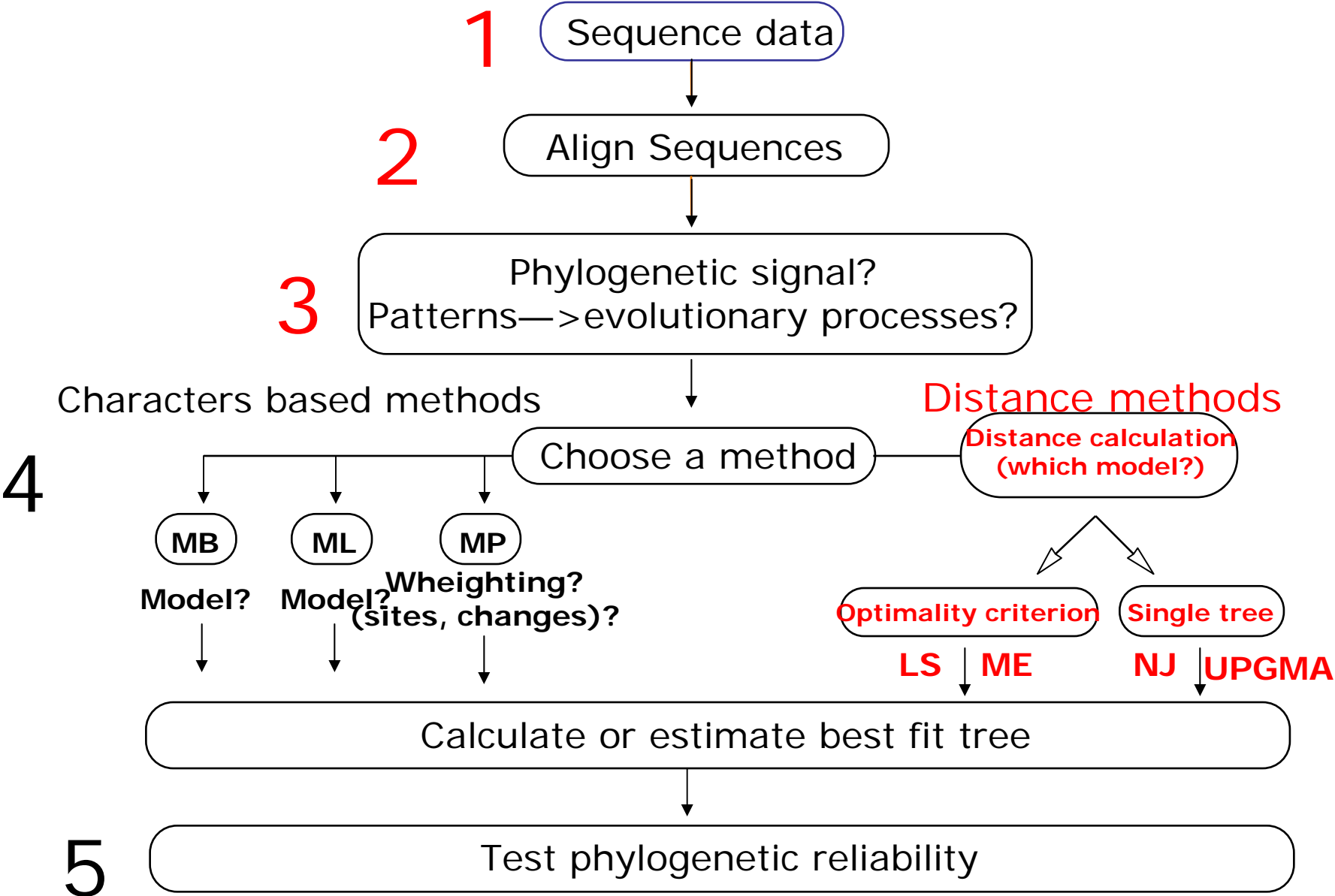
Pazos, Valencia 2003

Romero, Valencia 04

Del Sol, Pazos, Valencia *JMB* 03

# METHODS

## The five steps in phylogenetics dancing



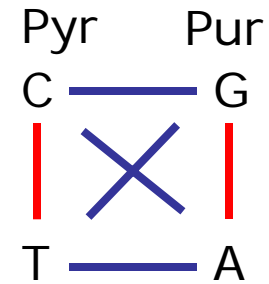
Modified from Hillis et al., (1993). Methods in Enzymology 224, 456-487

# Distance Methods

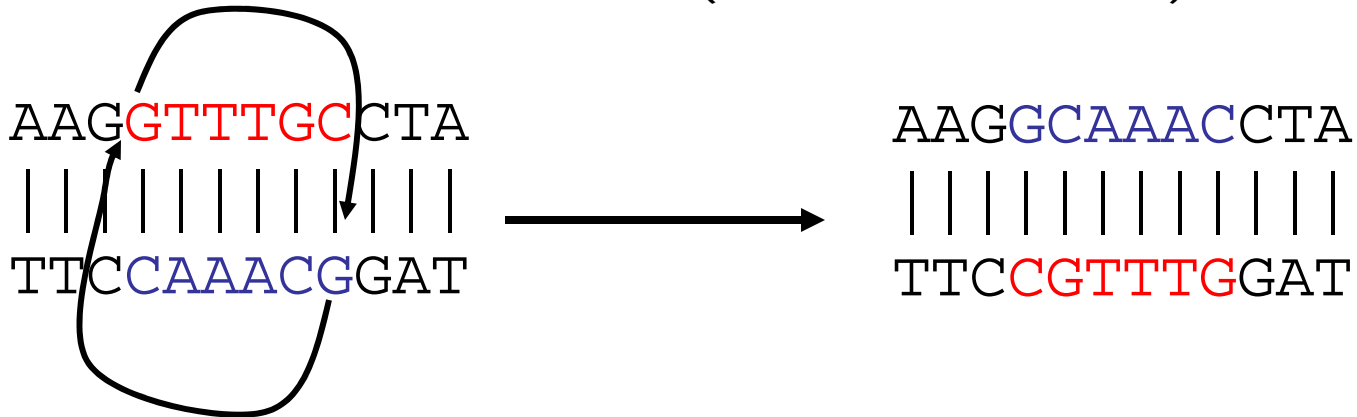
- Distance Estimates attempt to estimate the **mean number of changes** per site since 2 species (sequences) split from each other.
- Simply counting the number of differences (**p** distance) may underestimate the amount of change - especially if the sequences are very dissimilar - because of multiple hits.
- We therefore **use a model** which includes parameters which reflect how we think sequences may have evolved.

**Transitions:** changes between Pyrs or purs.

**Tranversions:** changes between Pyrs AND purs (2X more frequent\*)

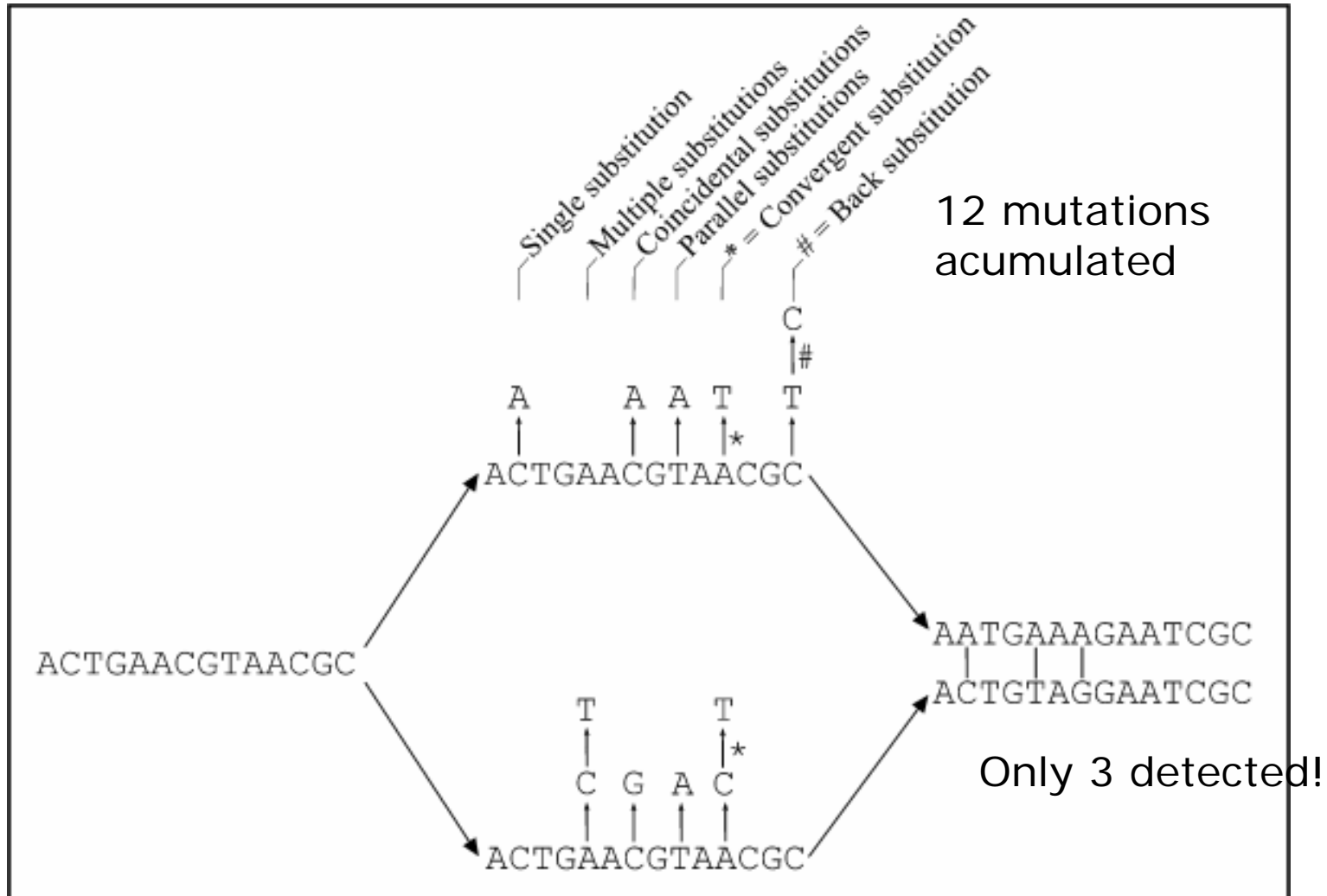


Inversion: 180 rotation ds-DNA (more than 2 bases)



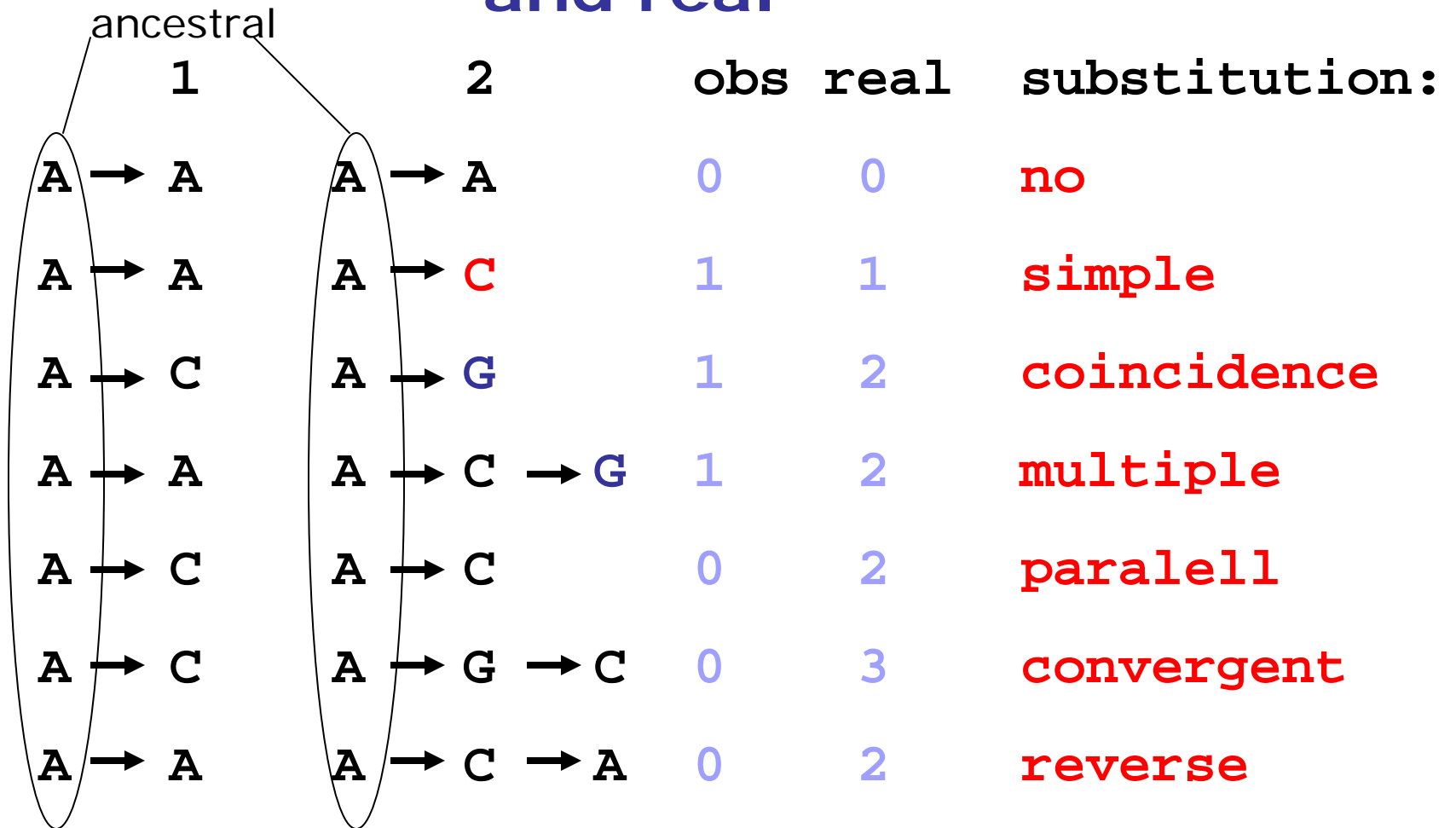


(from Fundamentals of Molecular Evolution,  
Wen-Hsiung Li and Dan Graur, 1991)



# METHODS

## Distances: observed and real



Obs might be <<<< Real changes!

## Distance calculations

- SEQ1    ACGTACGTAA
- SEQ2    ACGTTCGTAT
- SEQ3    TCCATCGTAAA

	S1	S2	S3
S1	0		
S2	0.2	0	
S3	0.4	0.4	0

Similarity

(1-2) 80%

(1-3) 60%

(2-3) 60%

Distance

$1 - 0.8 = 0.2$

$1 - 0.6 = 0.4$

$1 - 0.6 = 0.4$

# Saturation in sequence data:

- Saturation is due to **multiple changes** at the **same site** subsequent to lineage splitting.
- Most data will contain some fast evolving sites which are potentially saturated (e.g. in proteins **often position 3**).
- In severe cases the data becomes essentially random and all information about relationships can be lost

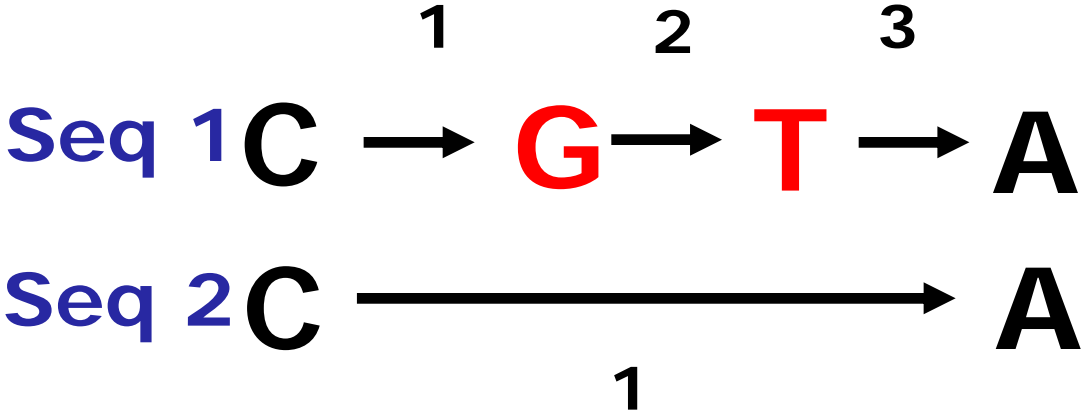
# METHODS

## Multiple changes at a single site - **hidden** changes

Seq 1 AGCG**A**G

Seq 2 GCGG**A**C

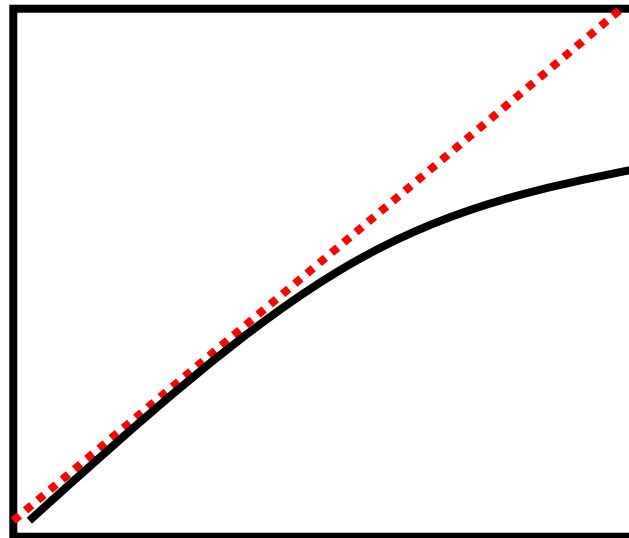
### Number of changes



# *METHODS*

**Substitution  
numbers**

**Observed**



**time**

# METHODS

The simplest model is that of Jukes & Cantor:

$$d_{xy} = -(3/4) \ln (1 - 4/3 D)$$

- $d_{xy}$  = distance between sequence x and sequence y expressed as the number of changes per site.
- (note  $d_{xy} = r/n$  where  $r$  is number of **replacements** and  $n$  is the **total** number of sites. This assumes all sites can vary and when unvaried sites are present in two sequences it will **underestimate** the amount of change which has occurred at variable sites).
- $D$  = is the **observed proportion of nucleotides** which differ between two sequences (fractional dissimilarity).
- $\ln$  = natural log function to correct for superimposed substitutions.
- The  $3/4$  and  $4/3$  terms reflect that there are **four** types of nucleotides and **three** ways in which a second nucleotide may not match a first - with **all types of change being equally likely** (i.e. unrelated sequences should be 25% identical by chance alone).

## METHODS

The natural logarithm  $\ln$  is used to correct for superimposed changes at the same site

- If two sequences are 95% identical, they are different at 5% or 0.05 (D) of sites thus:

$$-d_{xy} = -3/4 \ln (1 - 4/3 \cdot 0.05) = 0.0517$$

- Note that the **observed dissimilarity 0.05** increases only slightly to an estimated 0.0517 - this makes sense because in two very similar sequences one would expect very few changes to have been superimposed at the same site in the short time since the sequences diverged apart
- **However**, if two sequences are **only 50%** identical they are different at 50% or 0.50 (D) of sites thus:

$$-d_{xy} = -3/4 \ln (1 - 4/3 \cdot 0.5) = 0.824$$

- For dissimilar sequences, which may have diverged apart a long time ago, the use of  **$\ln$  infers that a much larger number** of superimposed changes have occurred at the same site



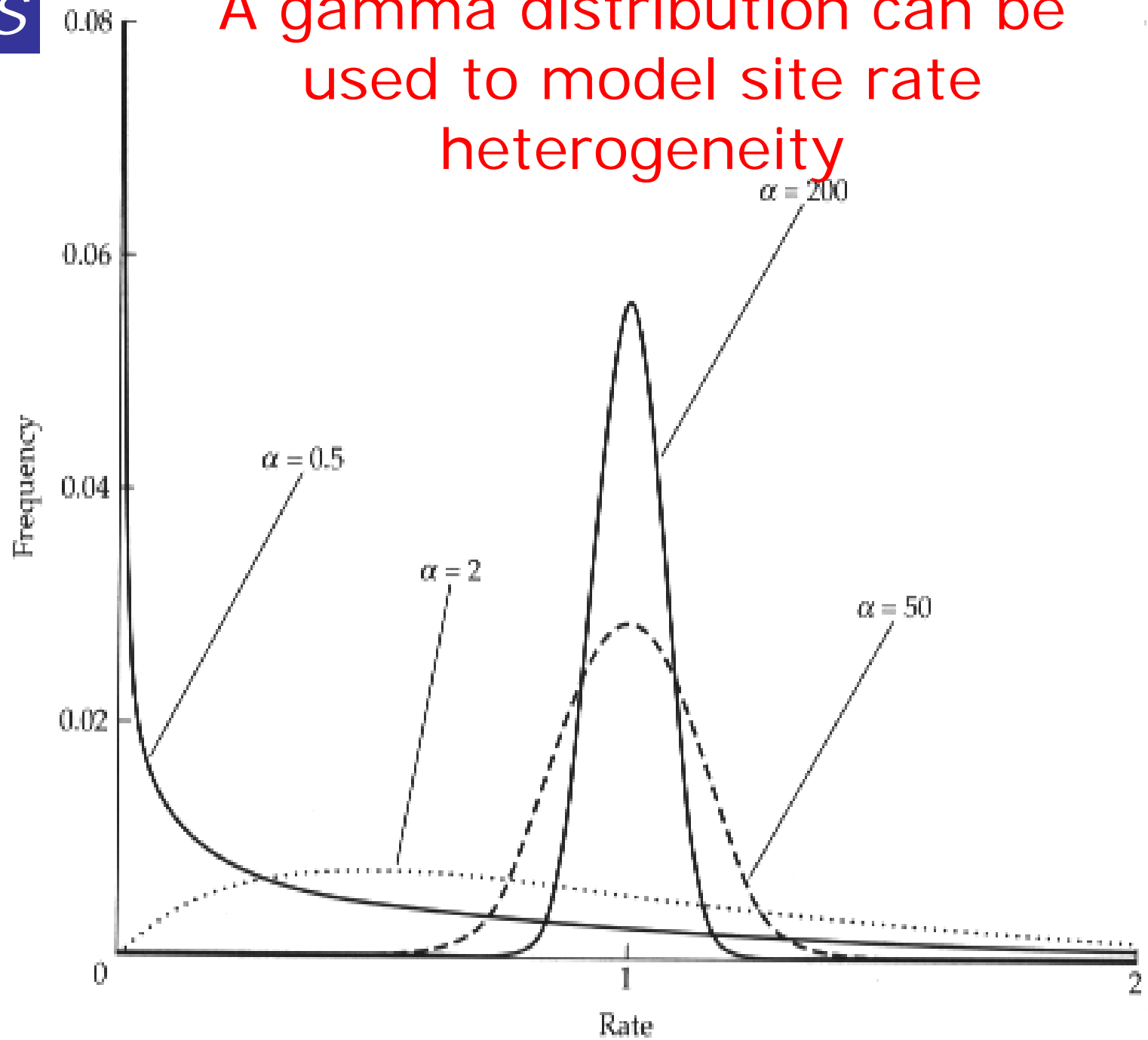
# METHODS

Distance models can be made more parameter rich to increase their realism 1

- It is better to use a model which fits the data than to blindly impose a model on data.
- The most common additional parameters are:
  - A **correction** for the proportion of sites which are **unable to change**.
  - A **correction** for **variable site rates** at those sites which can change.
  - A **correction** to allow different substitution rates for each type of nucleotide change

# METHODS

A gamma distribution can be used to model site rate heterogeneity



### Distances: advantages:

- Fast - suitable for analysing data sets which are too large for ML.
- A large number of models are available with many parameters - improves estimation of distances.
- Use ML to test the fit of model to data.

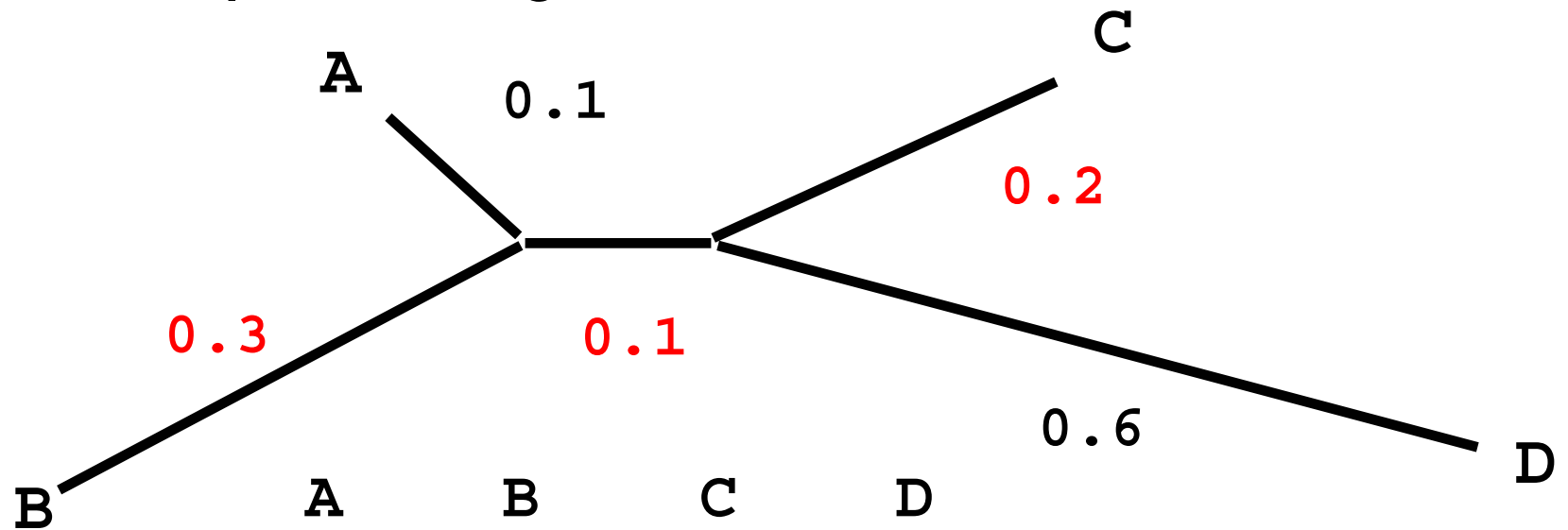
# Obtaining a tree using pairwise distances

## Additive distances:

- If we could determine exactly the true evolutionary distance implied by a given amount of observed sequence change, between each pair of taxa under study, these distances would have the useful property of tree additivity

# METHODS

A perfectly additive tree



	A	B	C	D
A	-	0.4	0.4	0.8
B	0.4	-	0.6	1.0
C	0.4	0.6	-	0.8
D	0.8	1.0	0.8	-

The branch lengths in the matrix and the tree path lengths match perfectly - there is a single unique additive tree

# METHODS

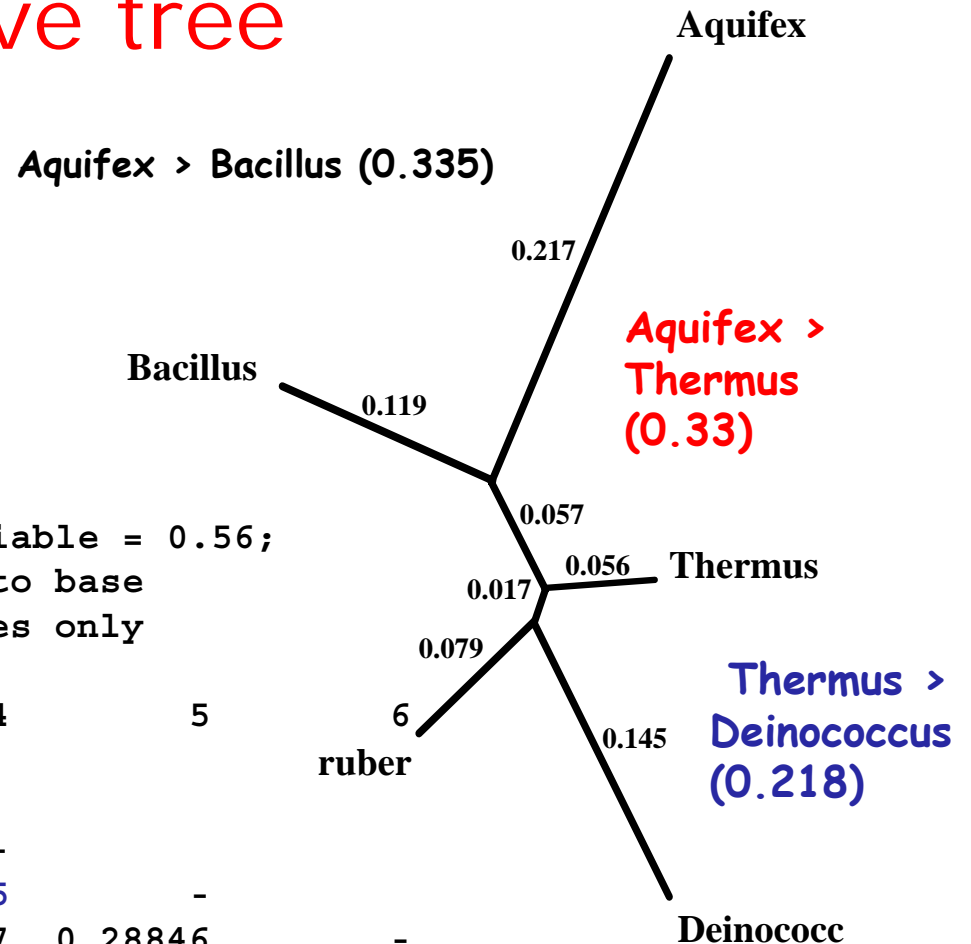
## Distance estimates may not make an additive tree

Some path lengths are longer and others shorter than appear in the matrix

Jukes-Cantor distance matrix

Proportion of sites assumed to be invariable = 0.56;  
identical sites removed proportionally to base frequencies estimated from constant sites only

	1	2	4	5	6
1 ruber	-				
2 Aquifex	0.38745	-			
4 Deinococc	0.22455	0.47540	-		
5 Thermus	0.13415	0.27313	0.23615	-	
6 Bacillus	0.27111	0.33595	0.28017	0.28846	-



## *METHODS*

# Obtaining a tree using pairwise distances

- Stochastic errors will cause deviation of the estimated distances from perfect tree additivity even when evolution proceeds exactly according to the distance model used.
- Poor estimates obtained using an inappropriate model will compound the problem.
- How can we identify the tree which best fits the experimental data from the many possible trees.

# METHODS

## Obtaining a tree using pairwise distances

- We have uncertain data that we want to fit to a tree and find the optimal value for the adjustable parameters (branching pattern and branch lengths).
- Use statistics to evaluate the fit of tree to the data (goodness of fit measures)
  - Fitch Margoliash method - a least squares method
  - Minimum evolution method - minimises length of tree
- Note that neighbor joining while fast **does not** evaluate the fit of the data to the tree.



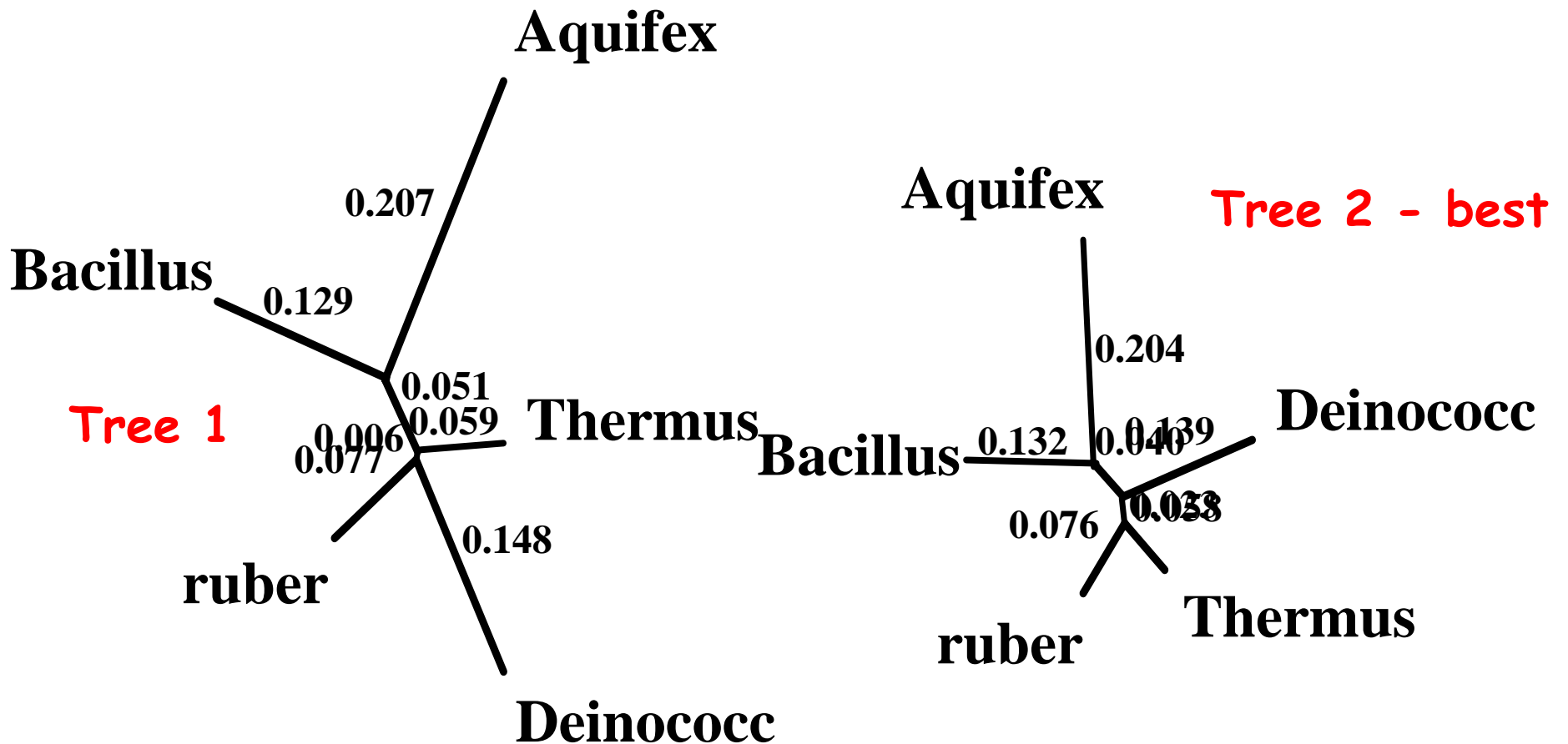
## *METHODS*

### Fitch Margoliash Method 1968:

- Minimises the weighted squared deviation of the tree path length distances from the distance estimates.

# METHODS

## Fitch Margoliash Method 1968:



Optimality criterion = distance (weighted least squares with power=2)

Score of best tree(s) found = 0.12243 (average %SD = 11.663)

Tree #	1	2
Wtd. S.S.	0.13817	<b>0.12243</b>
APSD	12.391	11.663

### Minimum Evolution Method:

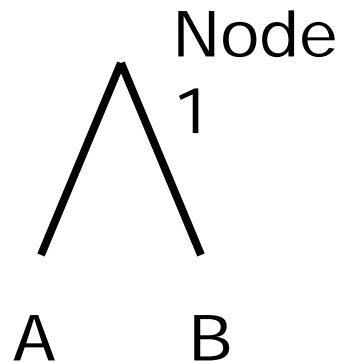
- For each possible alternative tree one can estimate the length of each branch from the **estimated pairwise distances** between taxa and then compute the sum (S) of all branch length estimates. The minimum evolution criterion is to choose the tree with the smallest S value.



## METHODS

# Neighbor joining method

- The neighbor joining method is a greedy heuristic which joins at each step, the two closest sub-trees that are not already joined.
- It is based on the minimum evolution principle.
- One of the important concepts in the NJ method is *neighbors*, which are defined as two taxa that are connected by a single node in an unrooted tree



# METHODS

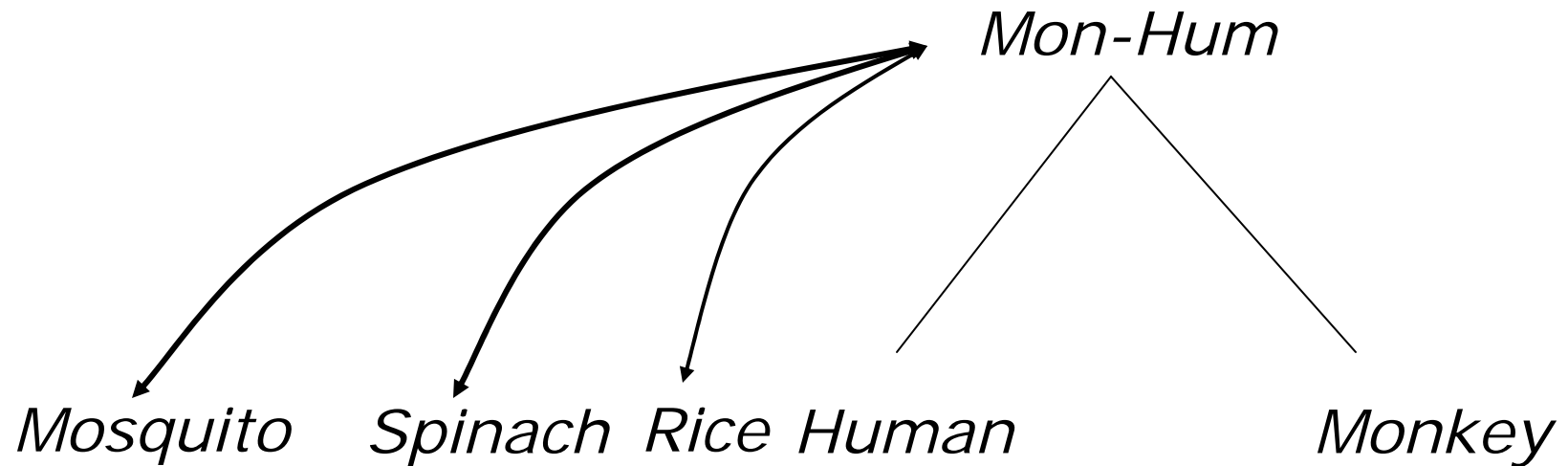
## Distance Matrix

PAM	Spinach	Rice	Mosquito	Monkey	Human
Spinach	0.0	84.9	105.6	90.8	86.3
Rice	84.9	0.0	117.8	122.4	122.6
Mosquito	105.6	117.8	0.0	84.7	80.8
Monkey	90.8	122.4	84.7	0.0	<b>3.3</b>
Human	86.3	122.6	80.8	<b>3.3</b>	0.0

# METHODS

## First Step

PAM distance 3.3 (Human - Monkey) is the minimum. So we'll join Human and Monkey to MonHum and we'll calculate the new distances.



## METHODS

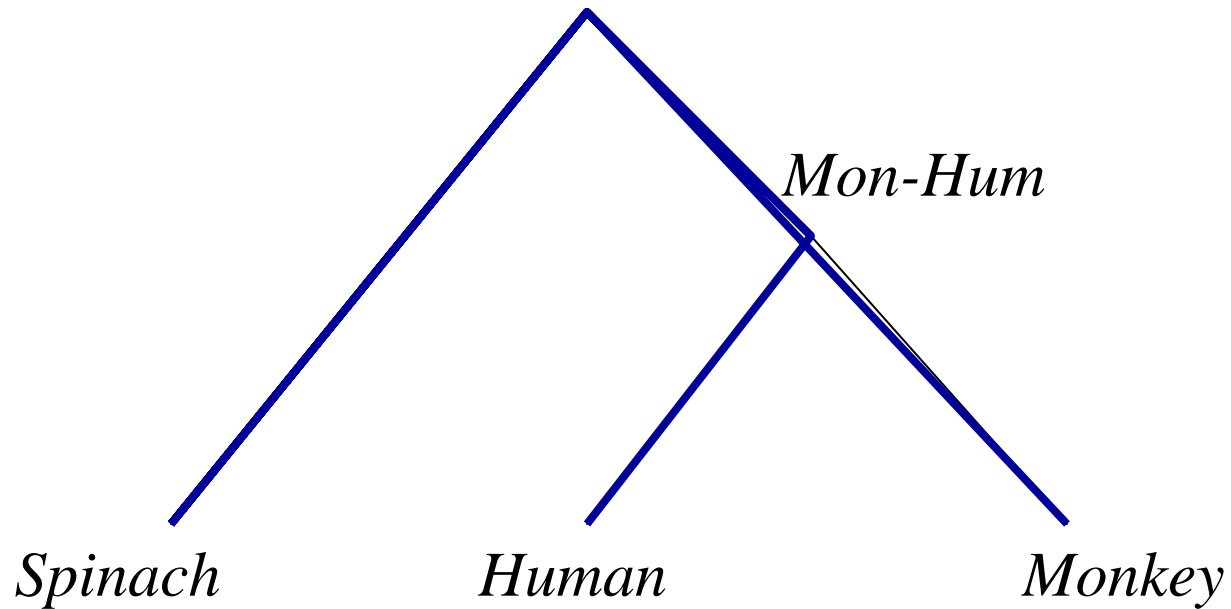
# Calculation of New Distances

After we have joined two species in a subtree we have to compute the distances from every other node to the new subtree. We do this with a simple average of distances:

$Dist[Spinach, MonHum]$

$$= (Dist[Spinach, Monkey] + Dist[Spinach, Human])/2$$

$$= (90.8 + 86.3)/2 = 88.55$$



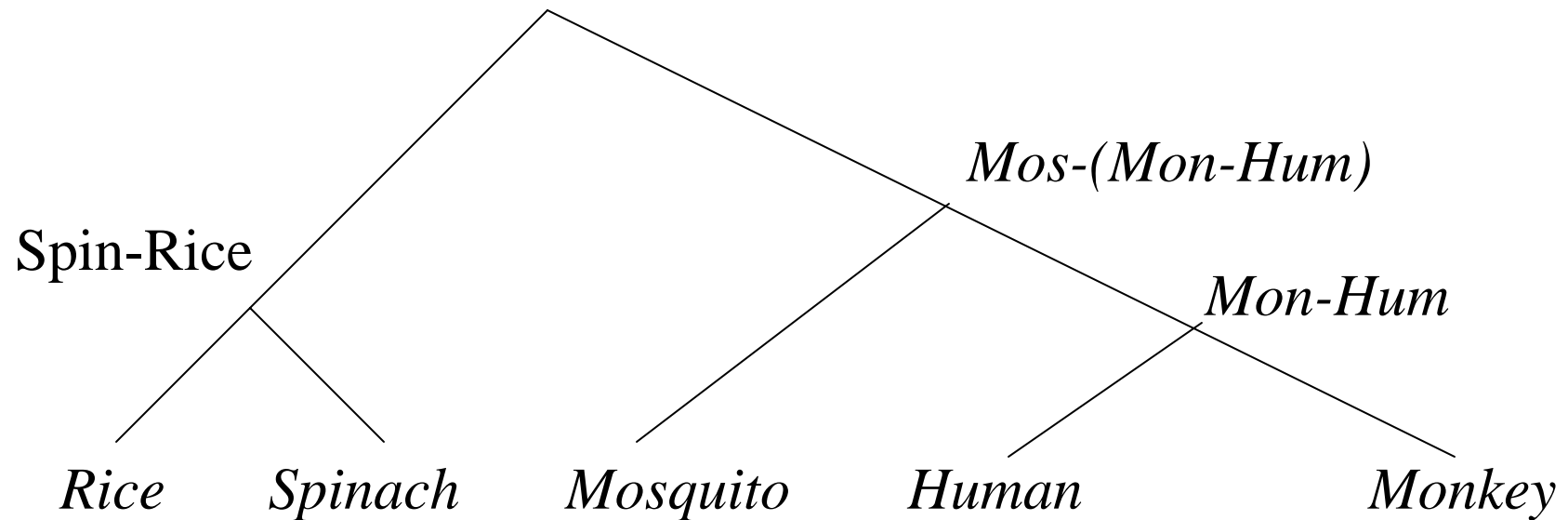


# METHODS

## Last Joining

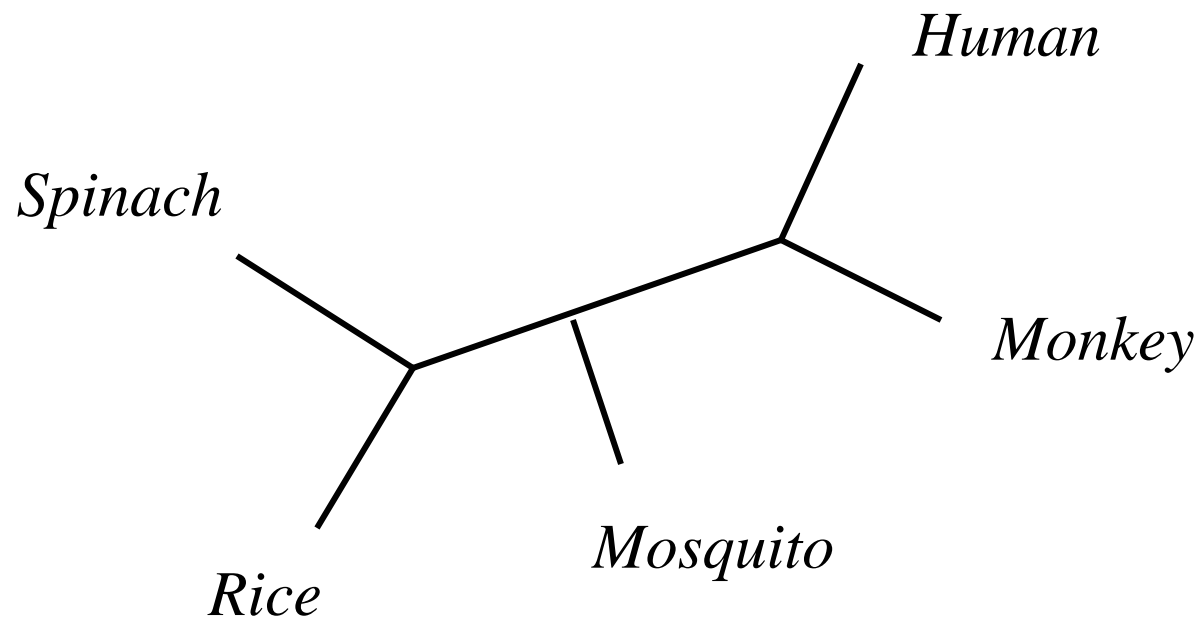
PAM	SpinRice	MosMonHum
Spinach	0.0	108.7
MosMonHum	108.7	0.0

(Spin-Rice)-(Mos-(Mon-Hum))



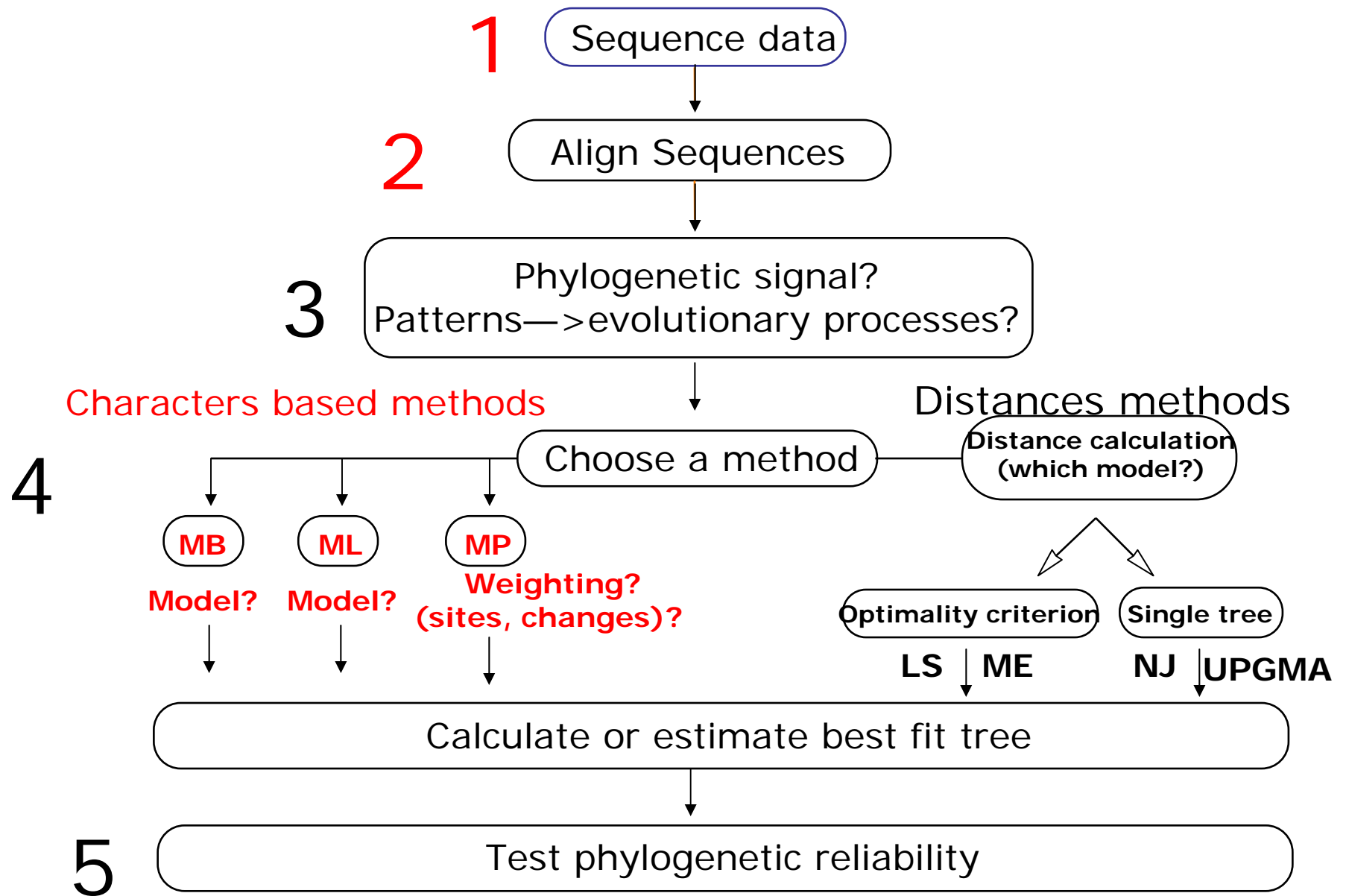
# METHODS

## Unrooted Neighbor-Joining Tree



# METHODS

## the five steps in phylogenetics dancing



## ML: comparison with other methods.

- ML is similar to many other methods in many ways
- In many ways it is fundamentally different.
- ML assumes a model of sequence evolution (so does Maximum Parsimony and so do distance matrix methods).
- ML attempts to answer the question: **What is the probability that I would observe these data (a multiple sequence alignment), given a particular model of evolution (a tree and a process).**

# METHODS

## Maximum Likelihood - goal

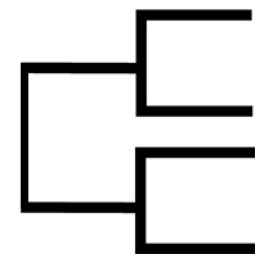
- To estimate the probability that we would observe a particular dataset, given a phylogenetic tree and some notion of how the evolutionary process worked over time.

– P(D/H)

Probability of

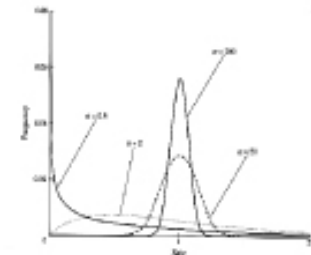
Seq1 aacg  
seq2 accg  
seq3 aaca  
seq4 aatg

given



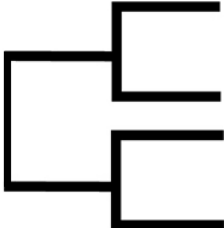
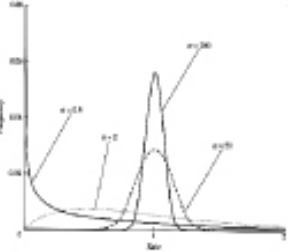
$$\pi = [a, c, g, t]$$

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
<i>b</i>	<i>a</i>	<i>e</i>	<i>f</i>
<i>c</i>	<i>e</i>	<i>a</i>	<i>g</i>
<i>d</i>	<i>c</i>	<i>f</i>	<i>a</i>



## The model

- The two parts of the model are the **tree** and the **process** (the model).
- The model is composed of the composition and the **substitution process -rate of change** from one character state to another character state.

**Model** =  + 
$$\begin{Bmatrix} a & b & c & d \\ b & a & e & f \\ c & e & a & g \\ d & c & f & a \end{Bmatrix} \pi = [a, c, g, t]$$
 

## METHODS

# Does changing a model affect the outcome?

There are different models

**Jukes and Cantor (JC69):**

All base compositions equal (0.25 each), rate of change from one base to another is the same

**Kimura 2-Parameter (K2P):**

All base compositions equal (0.25 each), different substitution rate for transitions and transversions).

**Hasegawa-Kishino-Yano (HKY):**

Like the K2P, but with base composition free to vary.

**General Time Reversible (GTR):**

Base composition free to vary, all possible substitutions can differ.

All these models can be extended to accommodate invariable sites and site-to-site rate variation.

## Strengths of ML

- Does not try to make an observation of sequence change and then a correction for superimposed substitutions. **There is no need to 'correct' for anything**, the models take care of superimposed substitutions.
- Accurate branch lengths.
- Each site has a likelihood.
- If the model is correct, we should retrieve the correct tree\*.
- You can use a model that fits the data.
- ML uses **all the data** (no selection of sites based on informativeness, **all sites are informative**).
- ML can not only tell you about the phylogeny of the sequences, but also the process of evolution that led to the observations of today's sequences.

\*If we have long-enough sequences and a sophisticated-enough model.



## Weaknesses of ML

- Can be inconsistent if we use models that are not accurate.
- Model might not be sophisticated enough (you can 'max-out' on models).
- **Very computationally-intensive.** Might not be possible to examine all models (substitution matrices, tree topologies, etc.).

## Parsimony Analysis

- Given a set of characters, such as aligned sequences, parsimony analysis works by determining the fit (number of steps) of each character on a given tree
- The sum over all characters is called Tree Length
  - Most parsimonious trees (MPTs) have the minimum tree length needed to explain the observed distributions of all the characters

## Results of parsimony analysis

- One or more most parsimonious trees.
- Hypotheses of character evolution associated with each tree (where and how changes have occurred).
- Branch lengths (amounts of change associated with branches).
- Various tree and character statistics describing the fit between tree and data.
- Suboptimal trees – optional.

## Parsimony - advantages

- is a simple method - easily understood operation.
- does not seem to depend on an explicit model of evolution.
- gives both trees and associated hypotheses of character Evolution.
- should give reliable results if the data is well structured and homoplasy is either rare or widely (randomly) distributed on the tree.

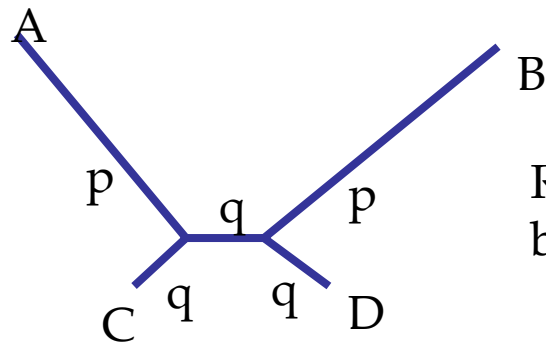
## Parsimony - disadvantages

- May give misleading results if **homoplasy** is common or concentrated in particular parts of the tree, e.g:
  - **thermophilic** convergence
  - base composition biases
  - **long branch attraction**
- Underestimates branch lengths.
- Model of evolution is implicit - behaviour of method not well Understood.
- Parsimony often justified on purely philosophical grounds – we must prefer simplest hypotheses - particularly by Morphologists.
- For most molecular systematists this is unconvincing

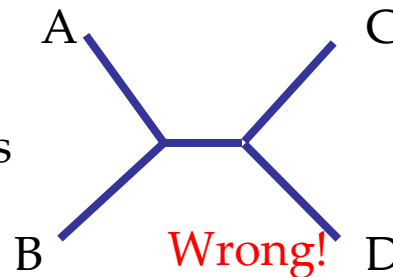
# METHODS

## Parsimony can be inconsistent

- Felsenstein (1978) developed a simple model phylogeny including four taxa and a mixture of short and long branches.
- Under this model parsimony will give the wrong tree



Rates or  
branch lengths  
 $p \gg \gg q$



Long branches are  
attracted but the  
similarity is  
homoplastic

- With more data the certainty that parsimony will give the wrong tree increases - so that parsimony is statistically **inconsistent**.
- Advocates of parsimony initially responded by claiming that Felsenstein's result showed only that his model was unrealistic.
- is now recognised that the **long-branch attraction** in the **Felsenstein Zone** is one of the most serious problems in phylogenetic inference

## Bayesian Inference of Phylogeny

- Clustering methods; UPGMA, NJ
- Parsimony: minimization of cost
- Statistical approaches
  - Maximum Likelihood
  - Bayesian Inference

## Statistical methods

- Maximum likelihood
  - Standard statistical approach
  - Philosophy widely accepted
  - Computationally difficult, especially for confidence intervals
- Bayesian inference
  - Old but marginal statistical approach until recently
  - Philosophy controversial (subjective probability)
  - Computationally efficient numerical solutions to difficult, high-dimensional problems

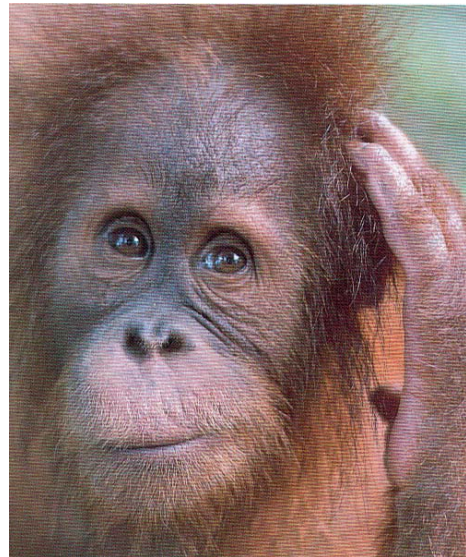


# *METHODS*

Infer relationships among three species:

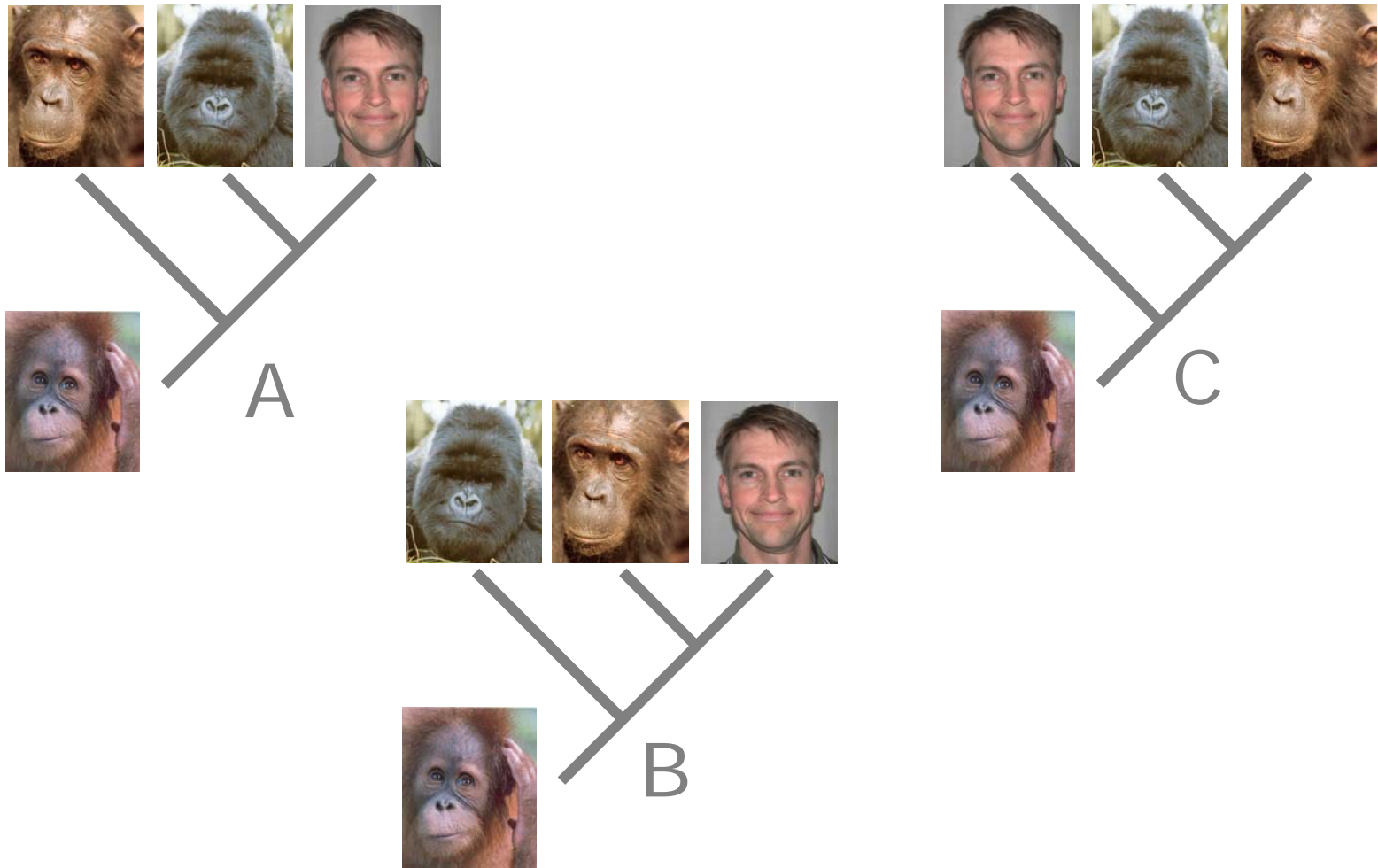


Outgroup:



# METHODS

Three possible trees (topologies):



# Bayes' rule



### Joint probabilities

$\Pr(B) = 0.6$      $\Pr(S) = 0.5$   
 $\Pr(W) = 0.4$      $\Pr(D) = 0.5$

$\Pr(\bullet) = \Pr(B, D) = 0.2$   
 $\Pr(\bullet) = \Pr(B, S) = 0.4$   
 $\Pr(\odot) = \Pr(W, D) = 0.3$   
 $\Pr(\circ) = \Pr(W, S) = 0.1$

### Conditional probabilities

$\Pr(B|D) = \frac{2}{5} = 0.2$

Hide all solid marbles (leaving 5 with dot)  
Of those left, 2 are black

### Bayes' rule

$\Pr(B, D) \xrightarrow{\Pr(D)}$

$$\Pr(D) \Pr(B|D) = \Pr(B) \Pr(D|B)$$

$$\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}$$

$\Pr(B|D) = \frac{\Pr(B) \Pr(D|B)}{\Pr(D)}$   
 $= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}$

### Marginal probability

$$\Pr(B|D) = \frac{\Pr(B) \Pr(D|B)}{\Pr(D)}$$

$\Pr(D) = \Pr(B, D) + \Pr(W, D)$   
 $= \Pr(B) \Pr(D|B) + \Pr(W) \Pr(D|W)$

Pr(D) is termed the "marginal probability of the data"  
It is obtained by "marginalizing over" color

# METHODS

## Bayes' theorem



Posterior  
distribution

Prior distribution

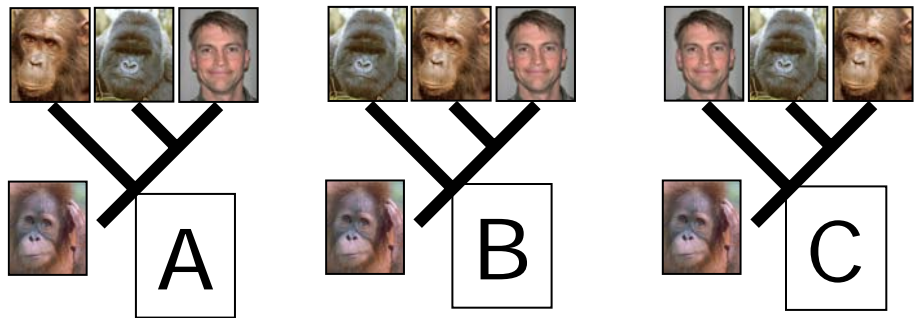
Likelihood function

$$f(\theta | X) = \frac{p(\theta)l(X | \theta)}{\int p(\theta)l(X | \theta)d\theta}$$

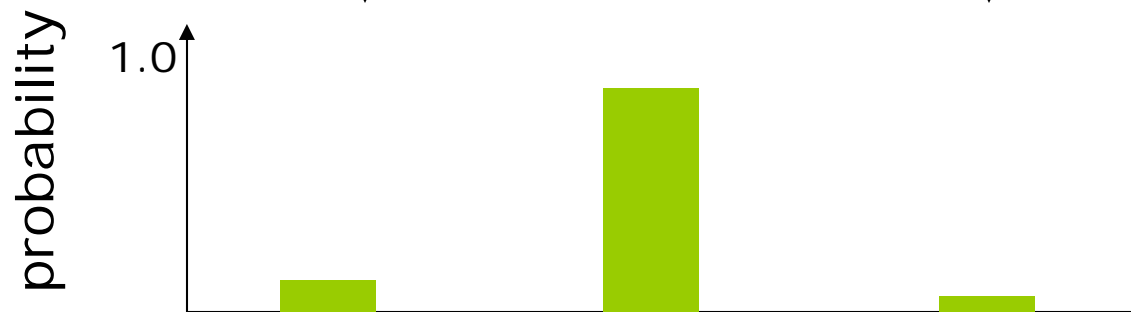
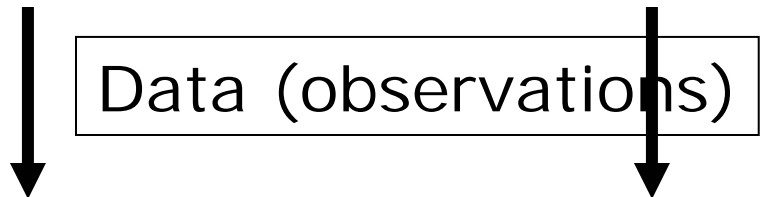
Unconditional probab.

$$\text{Pr [Tree/Data]} = (\text{Pr [Tree]} \times \text{Pr [Data/Tree]}) / \text{Pr [Data]}$$

# METHODS



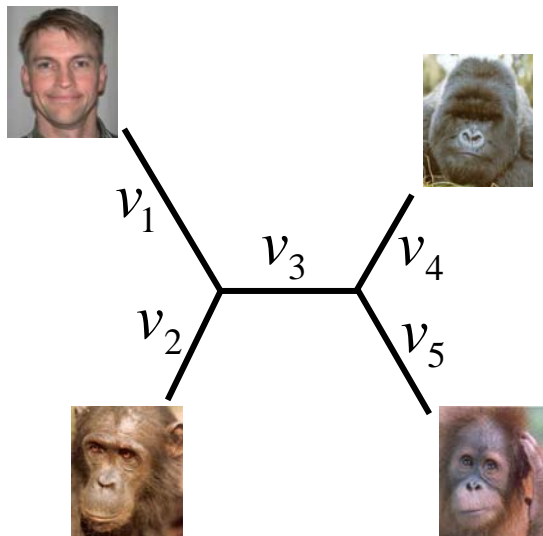
Prior probability distribution



Posterior probability distribution

## Model: tree and branch lengths

### $\theta$ Parameters



topology( $\tau$ )

(branching order)

branch lengths ( $v_i$ )

(expected amount of change per site or character)

$$\theta = (\tau, v)$$



# METHODS

## Data

### X The data

Taxon Characters



ACG TTA TTA AAT TGT CCT CTT TTC AGA



ACG TGT TTC GAT CGT CCT CTT TTC AGA



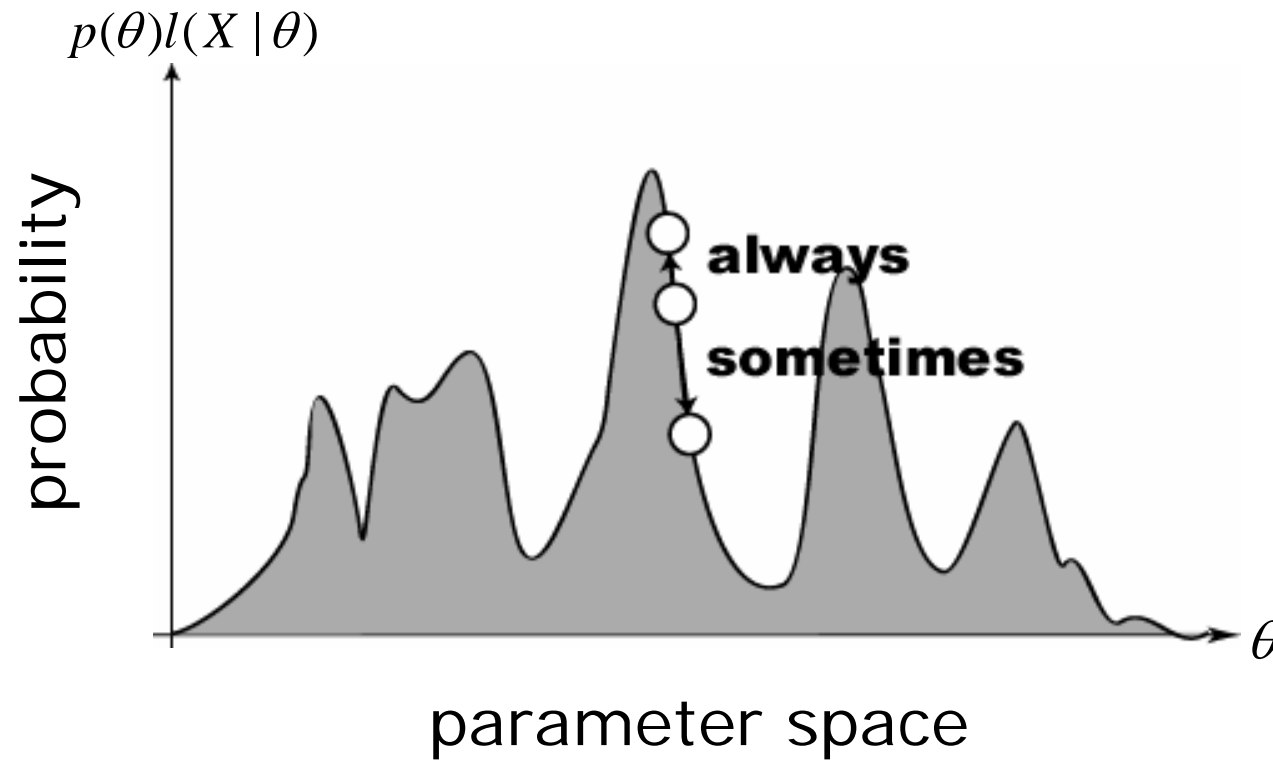
ACG TGT TTA GAC CGA CCT CGG TTA AGG



ACA GGA TTA GAT CGT CCG CTT TTC AGA

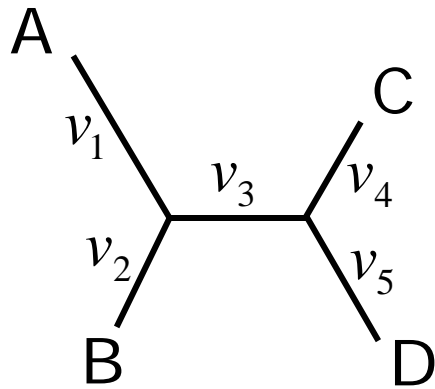
# METHODS

## Markov Chain Monte Carlo (MCMC)





## Model parameters 1

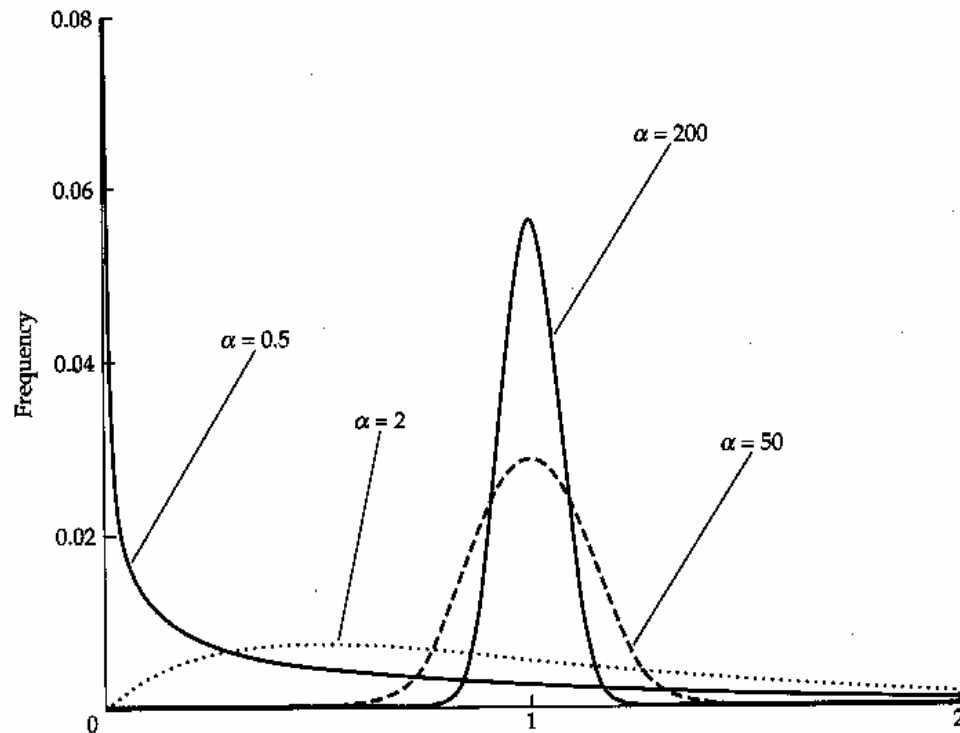


topology ( $\tau$ )  
branch lengths ( $v_i$ )

$$Q = \begin{pmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{pmatrix}$$

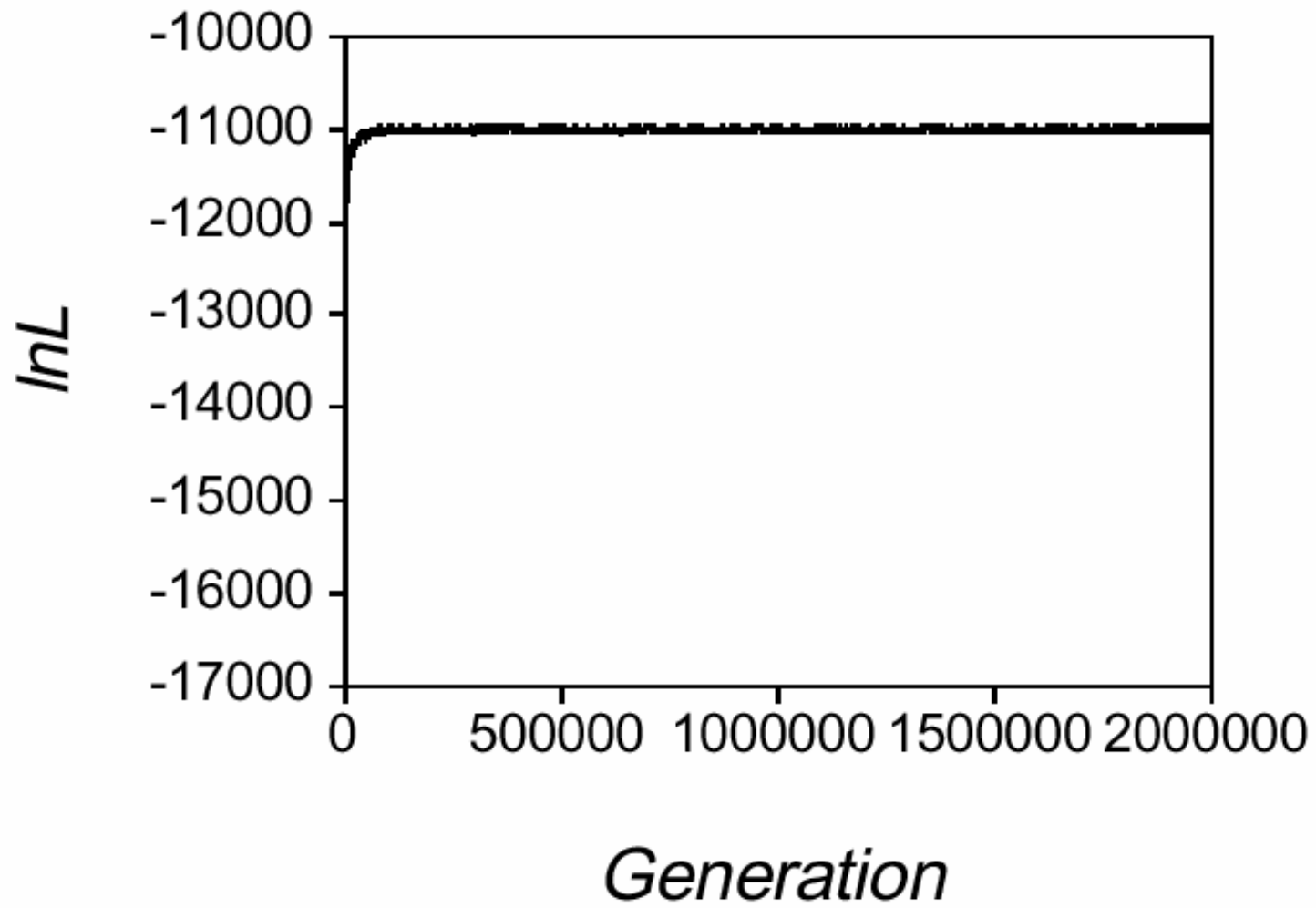
General Time Reversible  
substitution model

## Model parameters 2



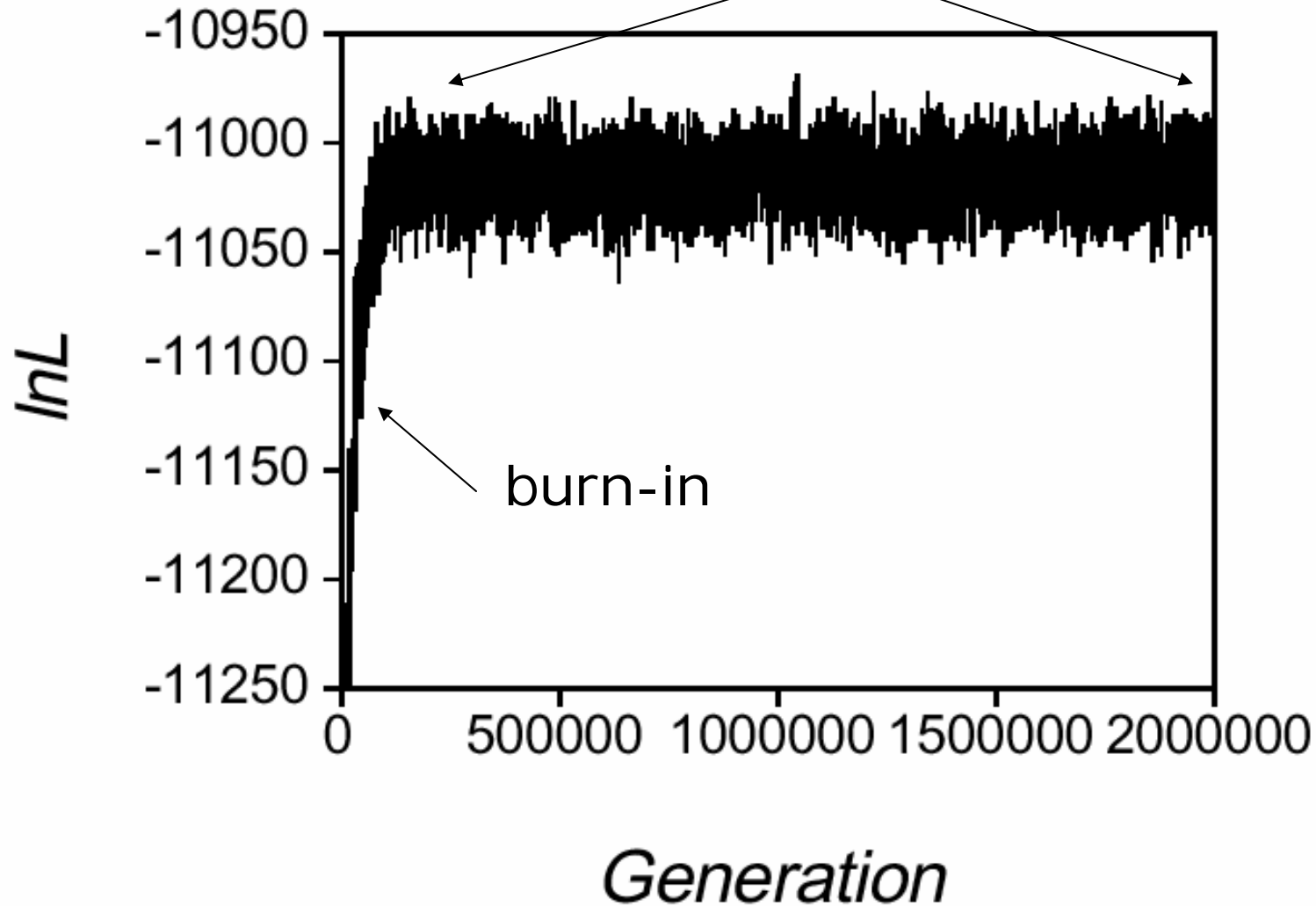
Gamma-shaped rate variation across sites

# METHODS

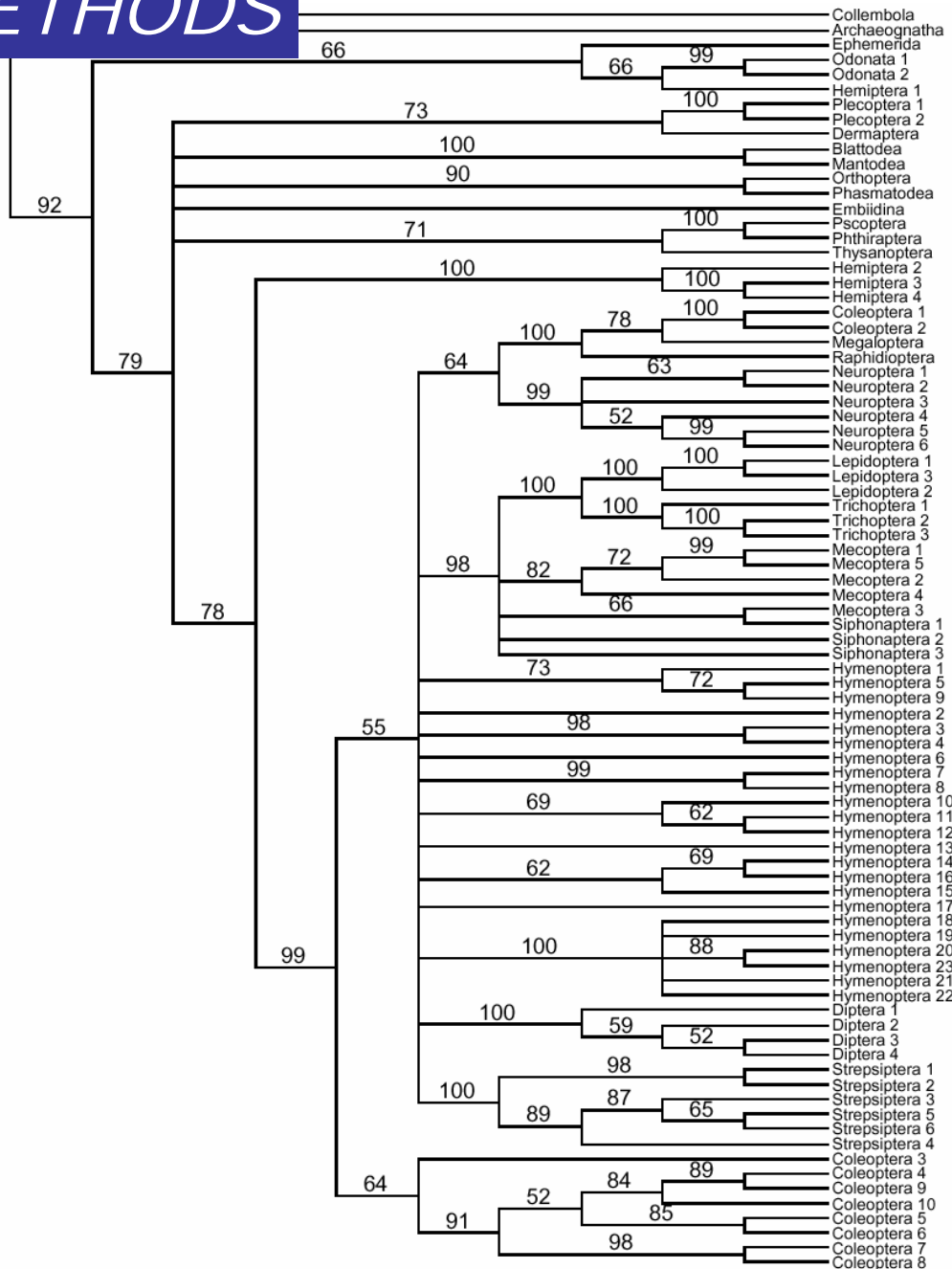


# METHODS

stationary phase sampled with thinning  
(rapid mixing essential)



# METHODS



Majority rule  
consensus tree  
from an MCMC  
run  
(insect **18S** data,  
**GTR** + G)

Frequencies  
represent the  
posterior  
probability of  
the clades

Probability of  
clade being true  
given data and  
model

## Bayesian inference pitfalls

- To what extent is the **posterior** distribution influenced by the prior?
- How do we know that the chains have converged onto the **stationary** distribution?
- Most common approach is to compare **independent** runs starting from different points in parameter space

# PROGRAMS:

*PROGRAMS*

## ML:

PAUP: <http://paup.csit.fsu.edu/about.html> David Swofford. (U-L,M,W)  
UNIX,MAC,Windows

PAML: <http://abacus.gene.ucl.ac.uk/software/paml.html> Ziheng Yang (U-L,M)

PHYLIP: <http://evolution.genetics.washington.edu/phylip.html> Joe Felsenstein

MOLPHY: Jun Adachi and Massami Hasegawa (Pascal)

PASSML: Pietro Lio (Hidden Markov) (U)

## MB:

BAMBE: <http://www.mathcs.duq.edu/larget/bambe.html> Donald Simon &  
B. Larget UNIX, Windows

Mac5: <http://www.agapow.net/software/mac5/> Paul-Michael Agapow  
UNIX,Windows,MAC

## OTHERS!

MEGA2: <http://www.megasoftware.net/> Kumar et al. DOS/Windows

Check out the list of Joe Felsenstein!

<http://evolution.genetics.washington.edu/phylip/software.html>



# Phylogeny Programs

There are some 194 of the phylogeny packages, and 16 free servers, that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or importance. Updates to these pages are made about twice a year (*however, almost no updates have been made since the start of 2001, and this will continue until at least the end of 2002 when I hope to complete a major writing project*). Some of these programs are available over Internet from [ftp server machines](#), or by World Wide Web.

Some programs listed below include both free and non-free ones; in some cases I do not know whether a program is free. I have listed as free those that I knew were free; for the others you have to ask their distributor.

If you discover any inaccuracies, or feel that I have left any important programs or facts out, or if links do not work properly, please e-mail me at: [tomoe@genetics.washington.edu](mailto:tomoe@genetics.washington.edu) ).

**List of packages arranged ...**

by methods available



# PHYLIP



# PROGRAMS

<http://evolution.genetics.washington.edu/phylip.html>

## DNA

**DNAPARS.** Estimates phylogenies by the parsimony method using nucleic acid sequences.

**DNAMOVE.** Interactive construction of phylogenies from nucleic acid sequences, with their evaluation by parsimony and compatibility

**DNAPENNY.** Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search.

**DNACOMP.** Estimates phylogenies from nucleic acid sequence data using the compatibility criterion,

**DNAINVAR.** For nucleic acid sequence data on four species, computes Lake's and Cavender's phylogenetic invariants,

**DNAML.** Estimates phylogenies from nucleotide sequences by maximum likelihood.

**DNAMLK.** Same as DNAML but assumes a molecular clock.

**DNADIST.** Computes four different distances between species from nucleic acid sequences.

## Proteins

**PROTPARS.** Estimates phylogenies from protein sequences using the parsimony method.

**PROTDIST.** Computes a distance measure for protein sequences

**SEQBOOT.** Reads in a data set, and produces multiple data sets from it by bootstrap resampling..

**FITCH.** Estimates phylogenies from distance matrix data under the "additive tree model".

**KITSCH.** Estimates phylogenies from distance matrix data under the "ultrametric" model.

**NEIGHBOR.** An implementation of Saitou and Nei's "Neighbor Joining Method," and of the UPGMA (Average Linkage clustering) method.

**CONSENSE.** Computes consensus trees by the majority-rule consensus tree method,

## Restriction

**RESTML.** Estimation of phylogenies by maximum likelihood using restriction sites data

## Continuous

**CONTML.** Estimates phylogenies from gene frequency data by maximum likelihood.

**GENDIST.** Computes one of three different genetic distance formulas from gene frequency data.

## Discrete characters

**MIX.** Wagner parsimony method and Camin-Sokal parsimony method,

**MOVE.** Interactive construction of phylogenies from discrete character Evaluates parsimony and compatibility criteria.

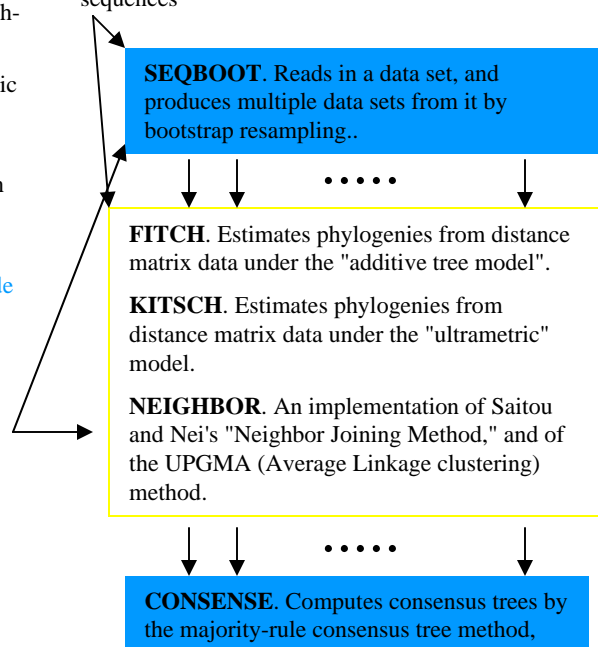
**PENNY.** Finds all most parsimonious phylogenies

**DOLLOP.** Estimates phylogenies by the Dollo or polymorphism parsimony criteria.

**DOLMOVE.** Interactive DOLLOP.

**DOLPENNY.** branch-and-bound method

**CLIQUE.** Finds the largest clique of mutually compatible characters,



# PROGRAMS

Molecular Evolutionary Genetic...  
File Phylogeny Windows Help

Click me to activate a data file  
Go to the MEGA2 web page  
Citing MEGA2 in publications

### Input Data

Data Type

- Nucleotide Sequences
- Protein Sequences
- Pairwise Distance

Missing Data ?

Alignment Gap -

Identical Symbol .

OK Cancel Help

### Confirm

Protein-coding nucleotide sequence data?

Yes No

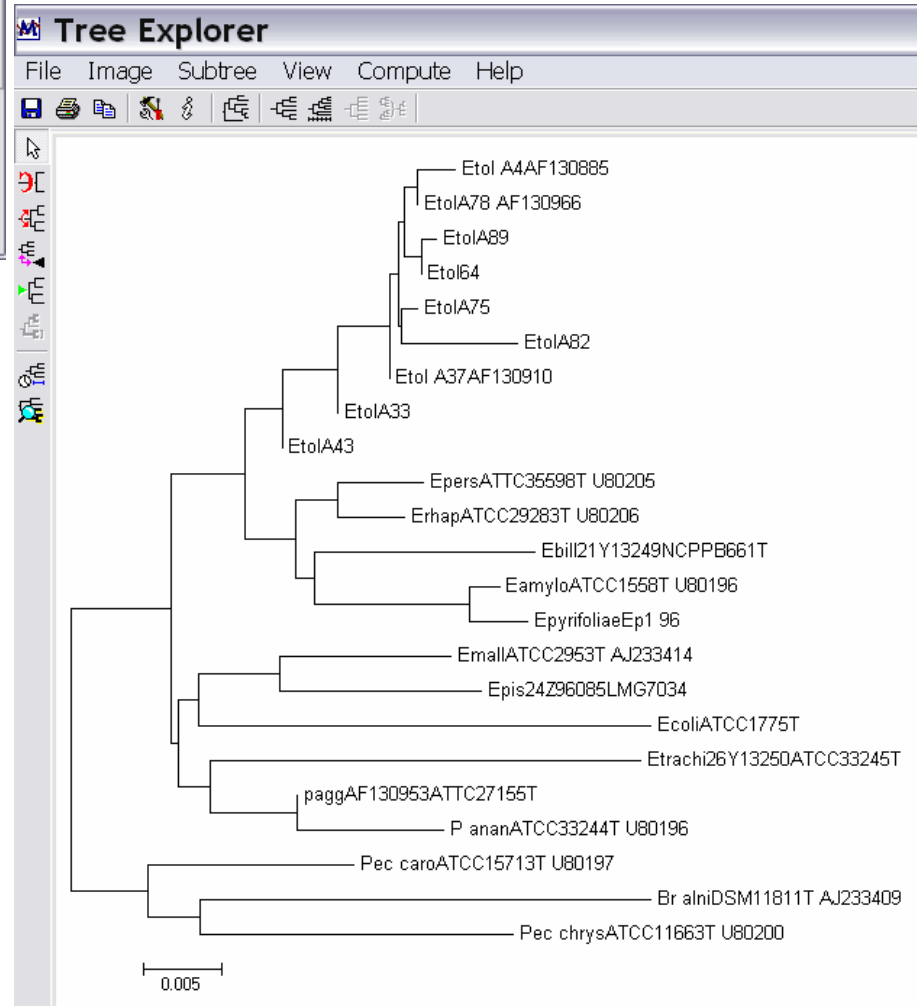
### Molecular Evolutionary Genetic...

File Data Distances Phylogeny Tests Windows Help

Choose Model...  
Compute Pairwise... F7  
Compute Overall Mean...  
Compute Within Group Means...  
Compute Between Groups Means...  
Compute Net Between Groups Means...  
Compute Sequence Diversity

Data File: C:\Documents and Settings\Ana\My Documents\erwinias\ijsmerv\_okacce  
Title: : erw16All.fasta\_aln

7:33:10 PM



## TOOL: MR. BAYES

Based on concept of posterior probabilities: probabilities that are estimated, based on some models (prior expectations), after learning something about the data (Mau et al., 1999). The user postulates a model of evolution, and the program searches for the best trees consistent with both the model, and the data (aln)

Method: Metropolis-coupled Markov Chain Monte Carlo: is a set of independent searches that occasionally exchanges information.

Model for aa replacement: Jones.

Number of markov chains: 4

Number of generations: >900.000

Number of trees generated: 1 tree each 100 generations.

Only trees generated after likelihood convergence are sampled (usually I discard 20% of the initial trees).

MrBayes v3.0B4

(Bayesian Analysis of Phylogeny)

by

John P. Huelsenbeck and Fredrik Ronquist

Section of Ecology, Behavior and Evolution  
Division of Biological Sciences  
University of California, San Diego  
johnh@biomail.ucsd.edu

Department of Systematic Zoology  
Evolutionary Biology Centre  
Uppsala University  
fredrik.ronquist@ebc.uu.se

Type "help" or "help <command>" for information  
on the commands that are available.

MrBayes > execute <filename>

Expecting <name>

MrBayes > sumt filename=<filename.t> contype=allcompat burnin=300

<http://morphbank.ebc.uu.se/mrbayes3/info.php> John Huelsenbeck & Fredrik  
Ronquist.

Linux, windows, mac...

```

Reading MrBayes block
Setting Nst to 6
Setting Rates to Gamma
Successfully set likelihood mode
Setting autoclose to yes
Setting number of generations to
Setting number of chains to 4
Setting print frequency to 100
Setting sample frequency to 10
Setting program to save branch l
Setting chain output file name to
Running Markov chain
MCMC stamp = 1713738510
Model settings:

```

```

Datatype = DNA
Nucmodel = 4by4
Nst = 6
Substitution r
of the rate su
(1.00,1.00,1.00)
Covarion = No
# States = 4
State frequenc
Rates = Gamma
Gamma shape pa
ributed on the
Gamma distribut

```

Active parameters:

Parameters	
Revmat	1
Statefreq	2
Shape	3
Topology	4
Brlen	5

- 1 -- Parameter = Revmat  
Prior = Dirichlet(
- 2 -- Parameter = Statefreq  
Prior = Dirichlet
- 3 -- Parameter = Shape  
Prior = Uniform(0.1
- 4 -- Parameter = Topology  
Prior = All topolog
- 5 -- Parameter = Brlen  
Prior = Branch len

```

Number of taxa = 22
Number of characters = 1463
Compressing data matrix for divis
Division 1 has 264 unique site p
The chain will use the following
With web Chain will change

```

```

Overwrite information in this file (yes/no): yes
Overwriting file "22all.t"

File "22all.p" already exists
Overwrite information in this file (yes/no): yes
Overwriting file "22all.p"

```

Chain results:

1	--	[-6040.883]	(-6095.569)	(-6124.074)	(-60
100	--	(-5553.538)	[-5510.862]	(-5524.556)	(-55
200	--	(-5292.485)	(-5349.851)	(-5366.941)	(-52
300	--	[-5048.705]	(-5196.156)	(-5121.155)	(-50
400	--	(-4977.469)	(-4987.319)	(-4958.716)	(-48
500	--	(-4865.054)	(-4863.180)	(-4889.441)	(-48
600	--	(-4811.581)	[-4713.902]	(-4832.969)	(-47
700	--	(-4780.029)	[-4680.291]	(-4784.629)	(-46
800	--	(-4751.444)	[-4645.695]	(-4745.191)	(-46
900	--	(-4721.009)	[-4631.269]	(-4664.733)	(-46
1000	--	(-4688.981)	(-4607.534)	(-4658.816)	(-45
1100	--	(-4686.537)	(-4585.177)	(-4616.187)	(-45
1200	--	(-4660.167)	(-4550.928)	(-4588.879)	(-45
1300	--	(-4636.410)	(-4532.153)	(-4554.852)	(-45
1400	--	(-4630.269)	(-4519.461)	[-4510.865]	(-45
1500	--	(-4616.658)	(-4517.763)	[-4488.994]	(-44
1600	--	(-4584.277)	(-4491.563)	[-4471.636]	(-44
1700	--	(-4549.489)	(-4493.847)	[-4473.793]	(-44
1800	--	(-4532.018)	(-4470.200)	(-4470.904)	(-44
1900	--	(-4519.365)	(-4463.503)	[-4439.422]	(-44
2000	--	(-4490.951)	(-4461.787)	(-4433.676)	(-44
2100	--	(-4479.080)	(-4447.364)	[-4425.087]	(-44
2200	--	(-4448.220)	(-4452.074)	[-4422.863]	(-44
2300	--	(-4445.675)	(-4433.248)	(-4418.132)	(-44
2400	--	(-4406.824)	(-4431.698)	(-4411.772)	(-44
2500	--	(-4401.989)	(-4403.245)	[-4380.510]	(-44
2600	--	(-4398.370)	(-4389.417)	(-4373.742)	(-43
2700	--	(-4400.827)	(-4386.204)	[-4350.632]	(-43
2800	--	(-4396.801)	(-4373.693)	[-4337.197]	(-43
2900	--	(-4377.571)	(-4361.306)	[-4335.397]	(-43

*REMARKS*

# THE PROBLEM OF THE EUKARYA LINEAGE

## DOMAIN SHUFFLING

## WHAT TO DO THEN?

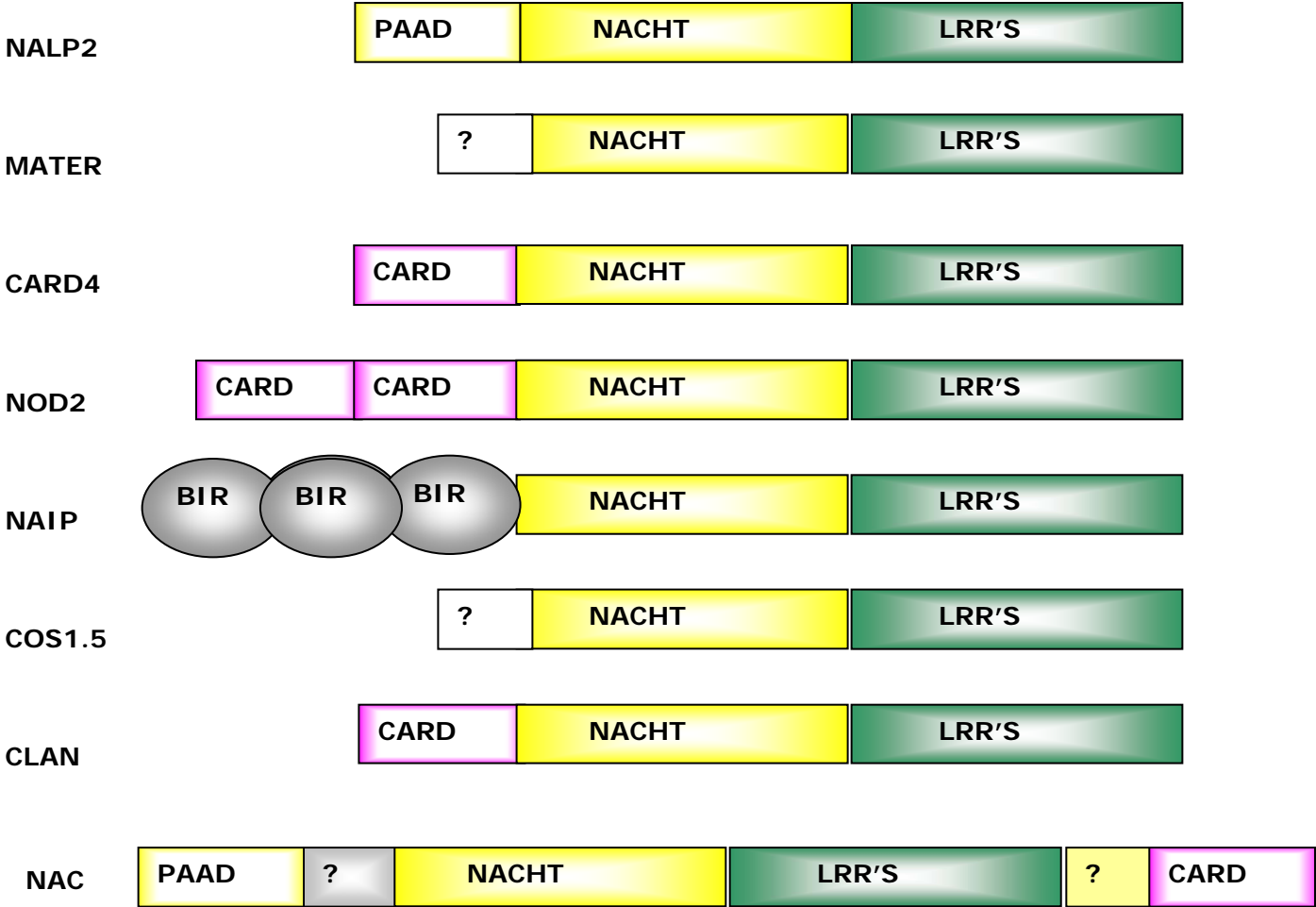
DOMAIN ANALYSES

CHECK CONSISTENCY BETWEEN DOMAIN DISTRIBUTION  
AND PHYLOGENETIC DISTRIBUTION

CHECK IF SHUFFLING IS RECENT OR OLD...

# REMARKS

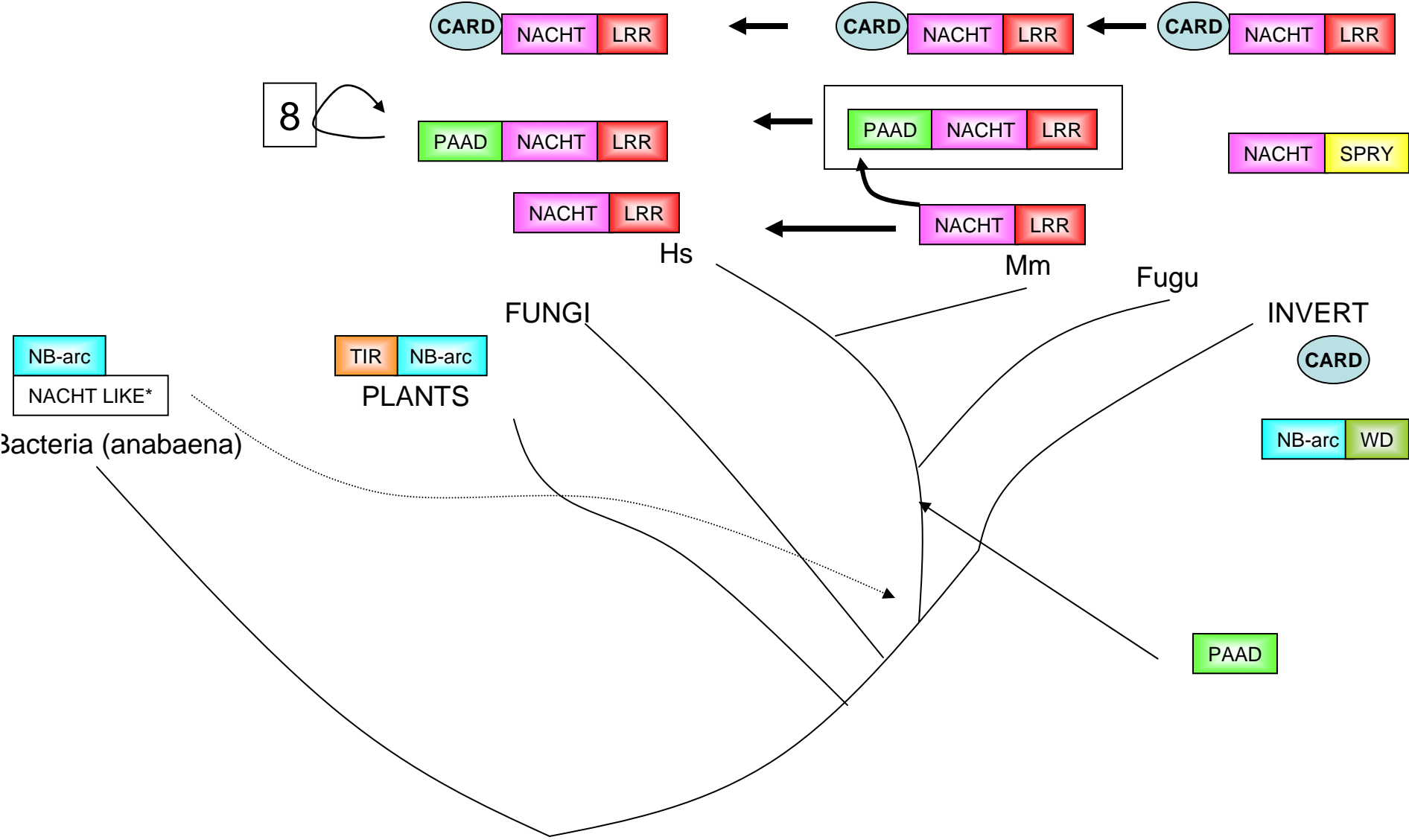
## DOMAIN ARCHITECTURES





# REMARKS

## NACHT DISTRIBUTION: POSSIBLE SCENARIO



## **SOME PRACTICAL EXAMPLES**

- **DESCRIPTION OF NEW SPECIES**

*Erwinia toletana* sp. nov.

- **PLACEMENT OF NEW ISOLATED GENES**

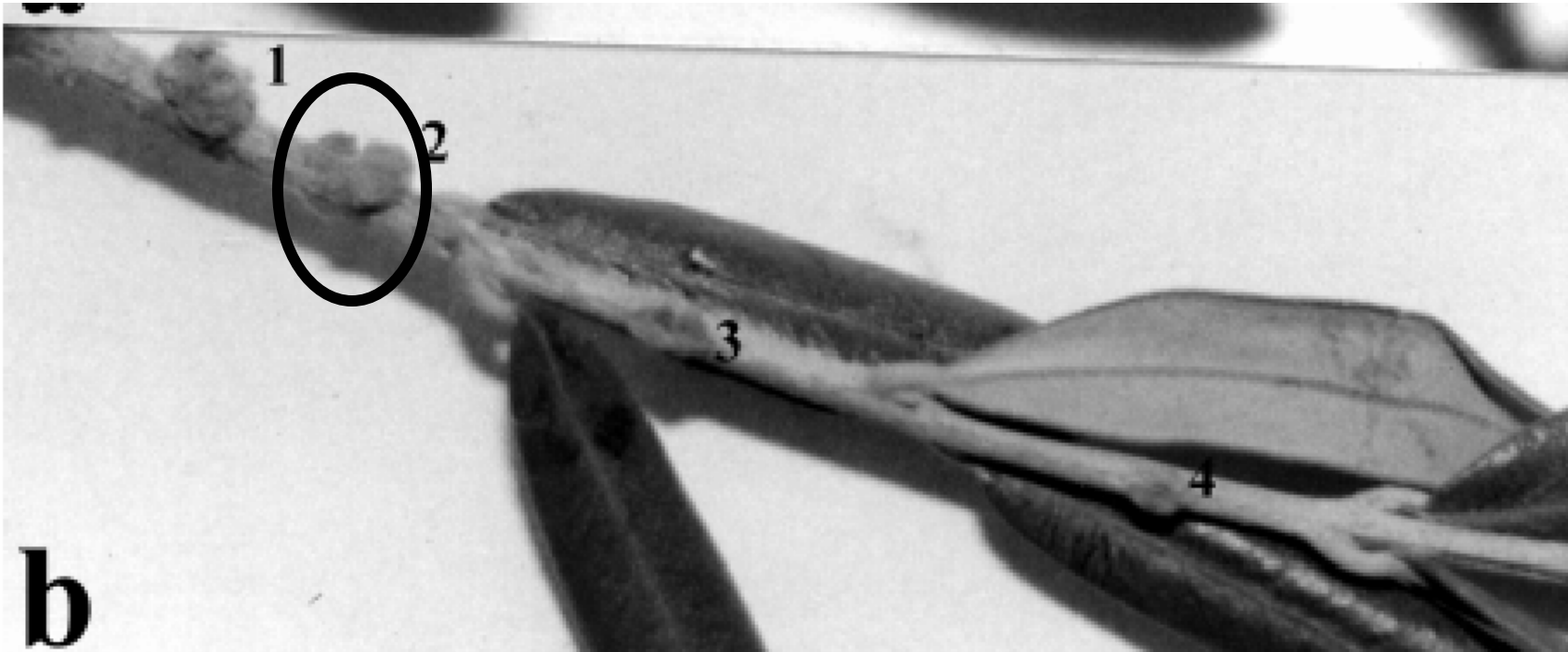
Ocurrence of serin proteases in sponge and jellyfish

## SOME EXAMPLES

- **DESCRIPTION OF NEW SPECIES**

Goal: to obtain a natural antagonist of *P. savastanoi*.

Data: Bacterial species isolated from wild trees' knots (Olives, oleander...)



total of 81 bacterial strains!

The problem: Resemble phenotypically to several..

What to do?:

- Choose an universal conserved marker: i.e. 16SRNA, Extract similar sequences  
Build phylogenetic trees

Gene sequencing:

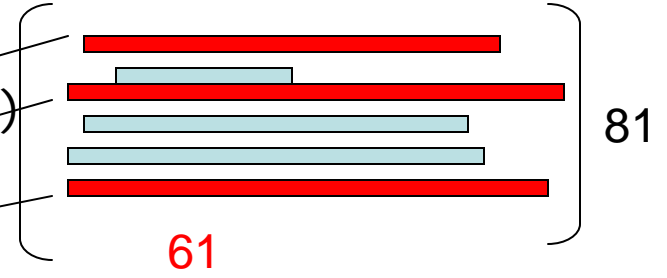
16SRNA, 23SRNA, gnd, mdh

WHY THESE GENES? ???????????

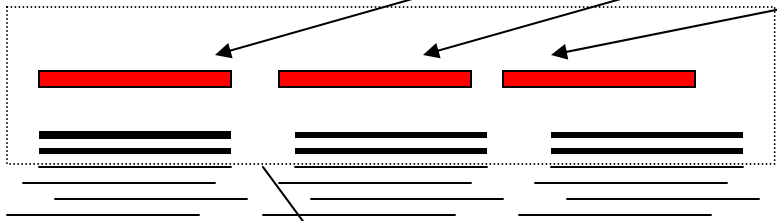
# SOME EXAMPLES

## METHOD FOR 16SRNA

From 81 sequences only the longest retained (61)



stand-alone blasted against a **filtered EMBL DB**



A total of 19,184 sequences retained (from 80,807 initial sequences). .

The 2 most similar are retained to phylogenetic tree reconstructio

Parsimony

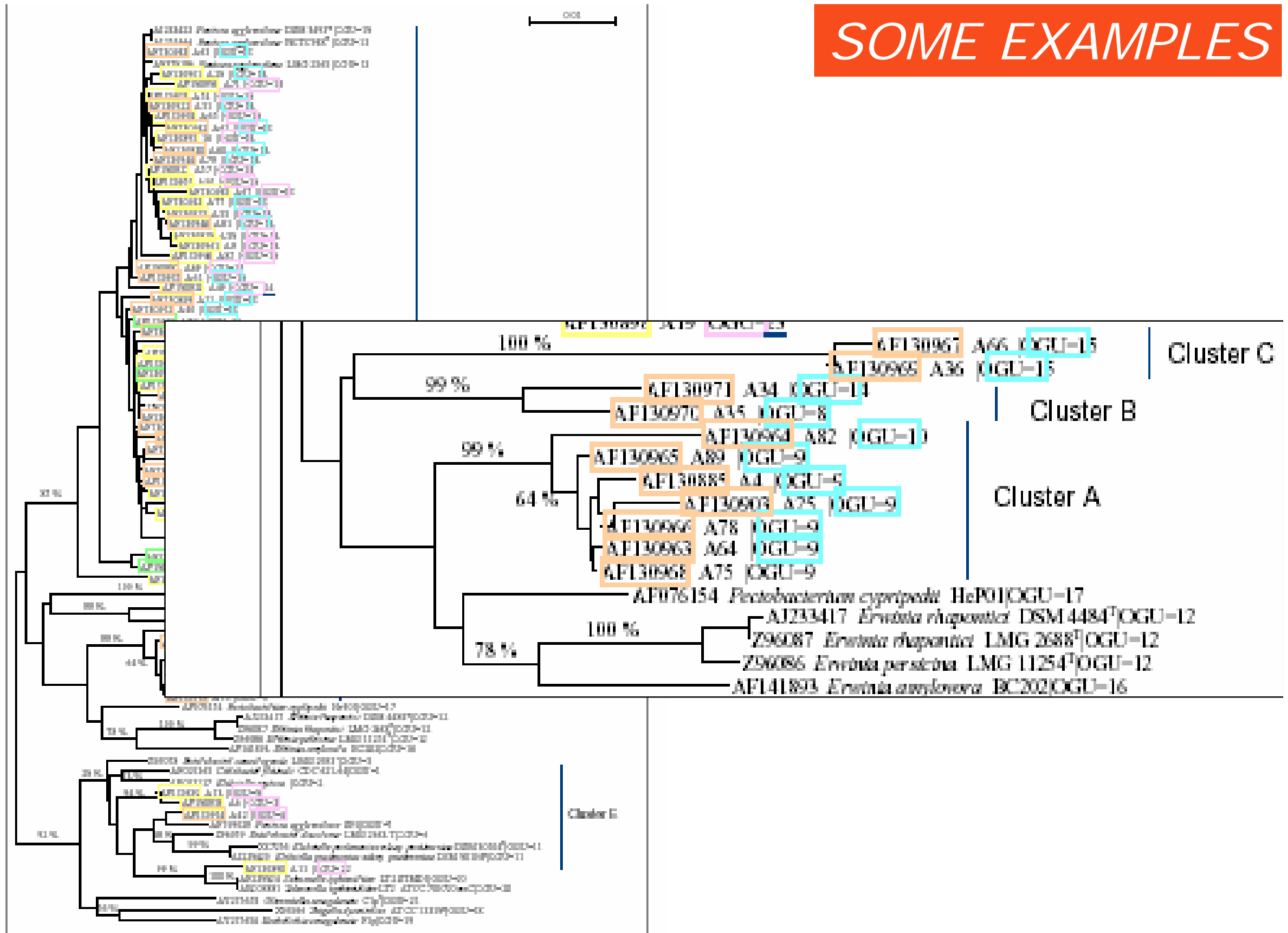
Maximum likelihood

BioNJ

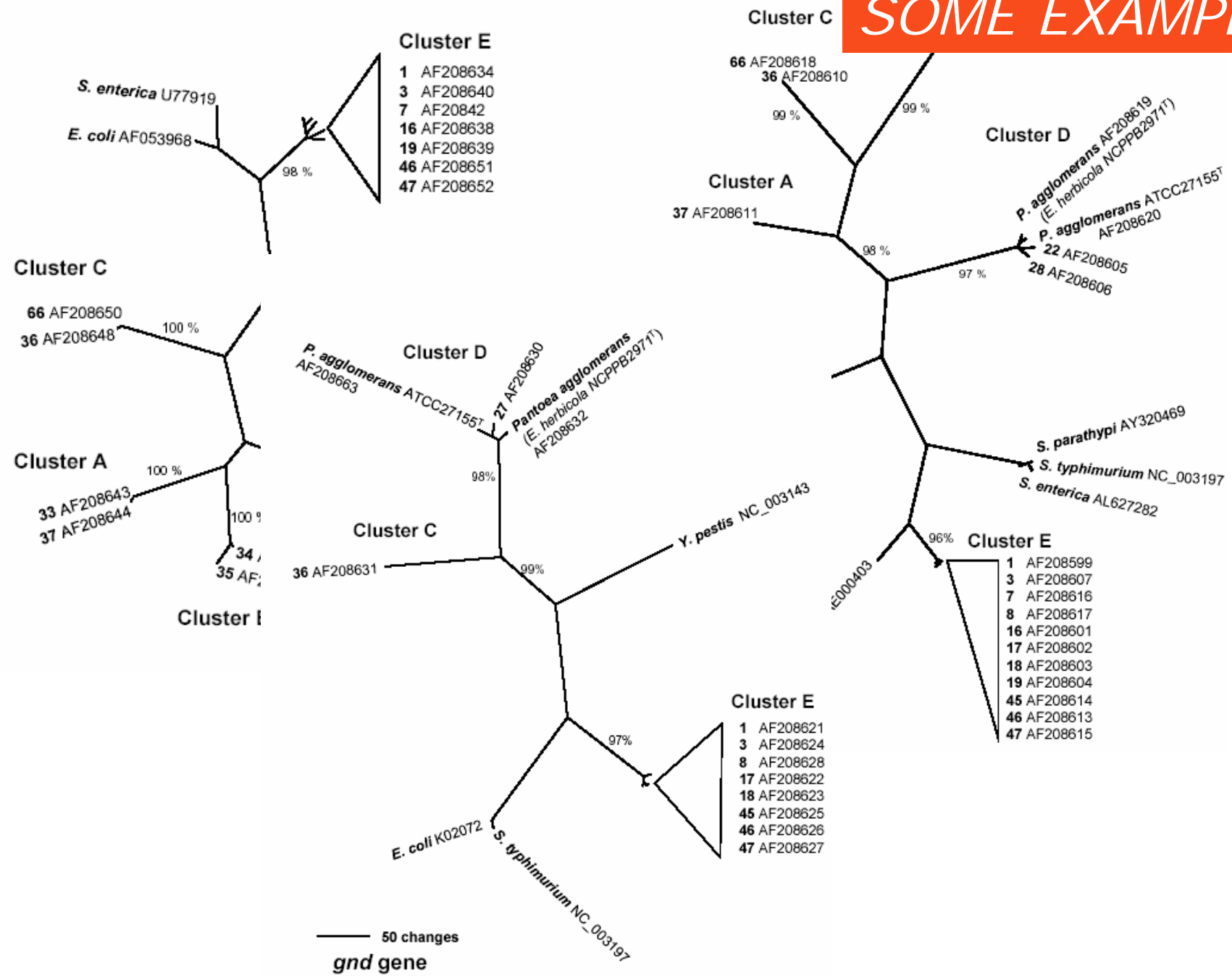
1000 bootstrap

CONSENSUS!

# SOME EXAMPLES



# SOME EXAMPLES



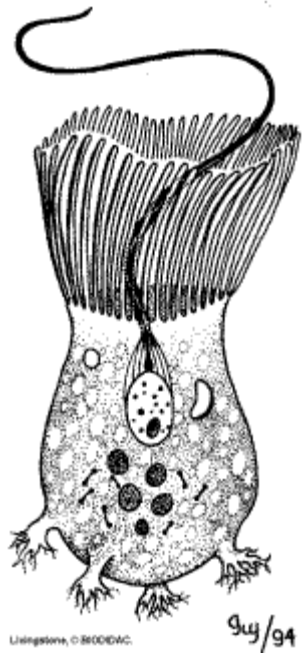
- **PLACEMENT OF NEW ISOLATED GENES**

**Ocurrence of serin proteases in sponge and jellyfish**

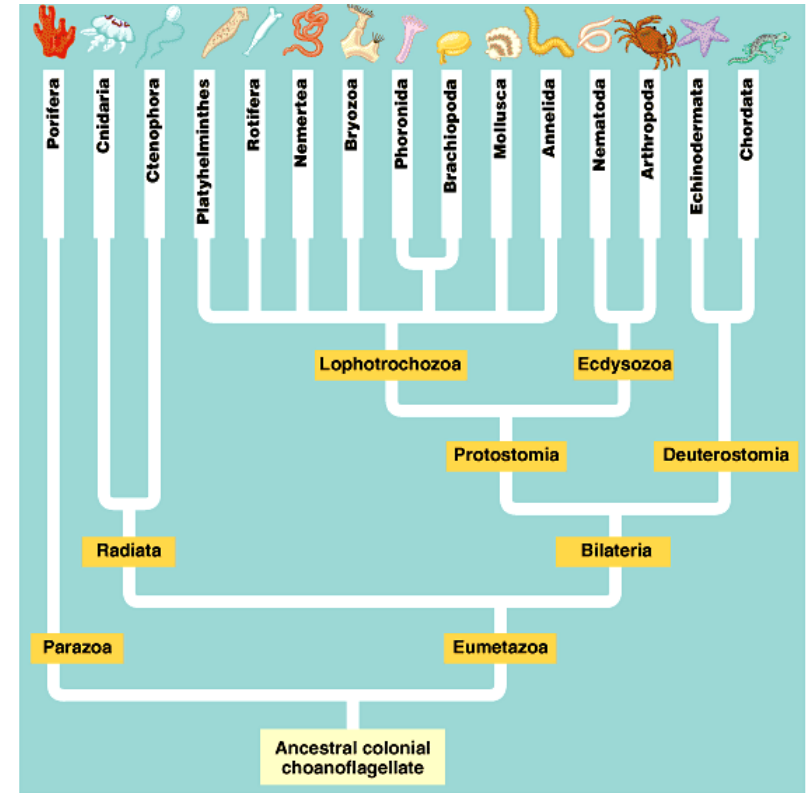
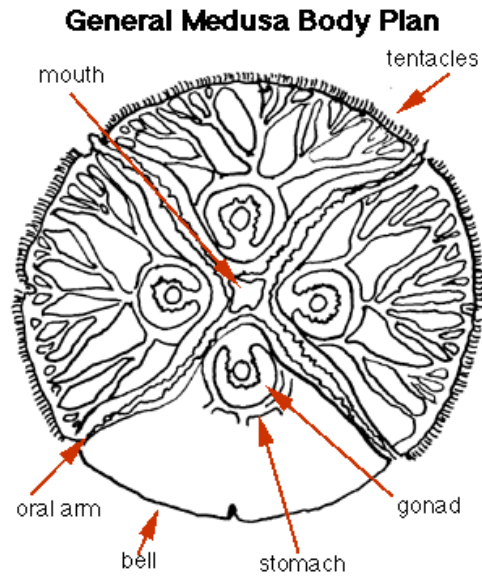
**Goal: Confirm the existence of serin proteases in early-divergent phyla, *cnidaria* and *porifera*.  
Where they come from?**

**Data: SP are absent in plants, and protists and in fungi are restricted to *Streptomyces*. However, there are hundreds in animals!**





Livingstone, © 2000/04/04

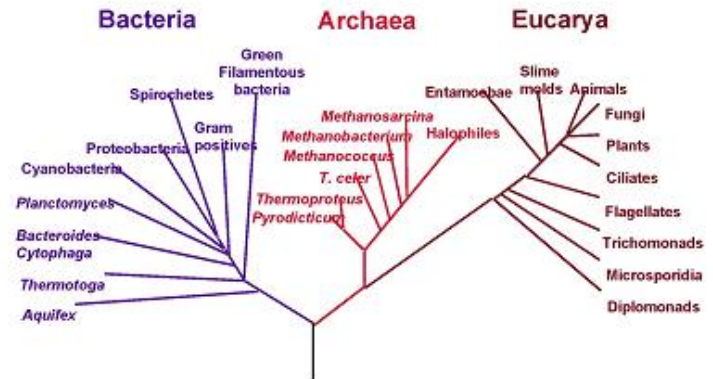


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.



© 2002 CZS

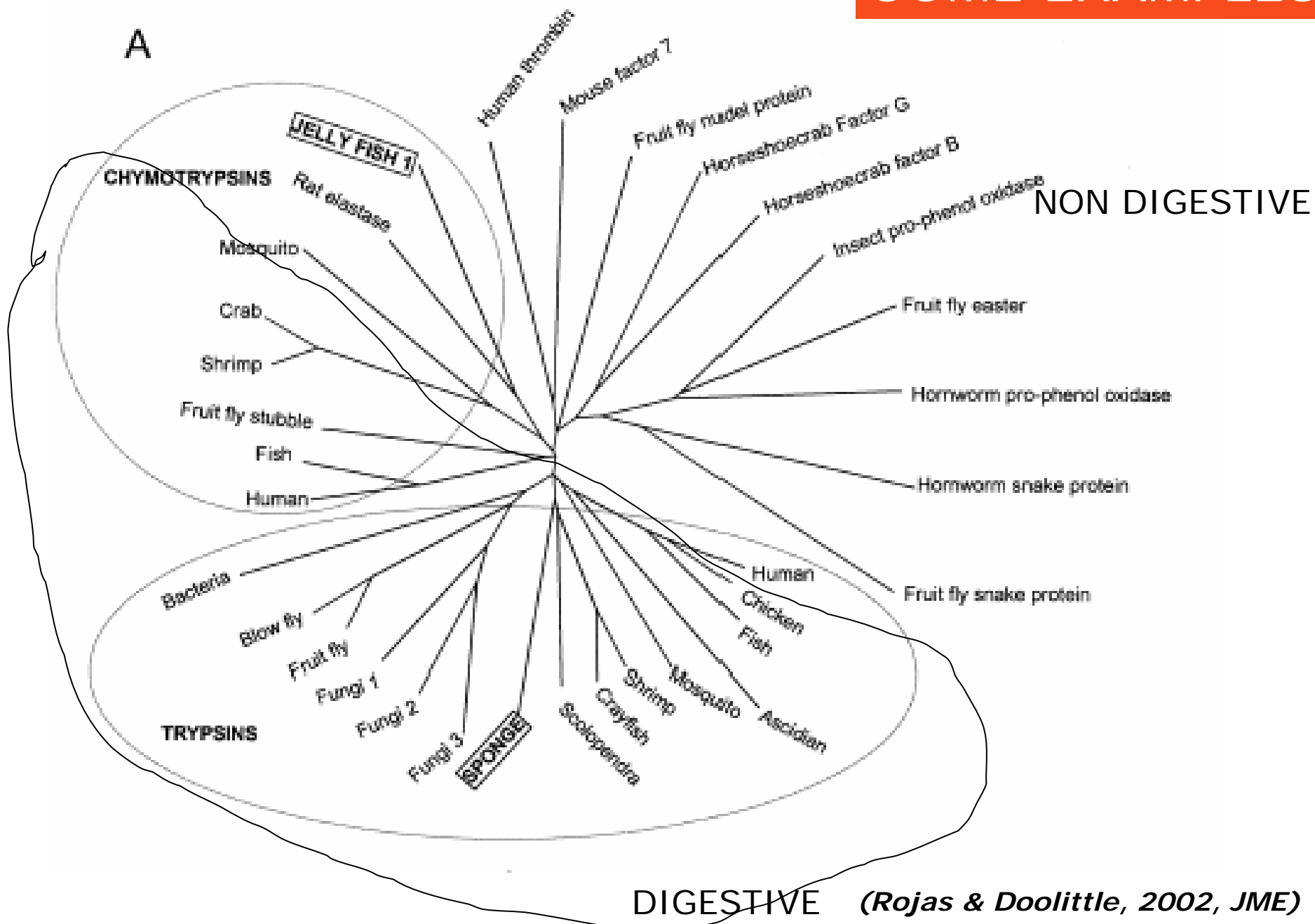
## Phylogenetic Tree of Life



(Rojas & Doolittle, 2002, JME)

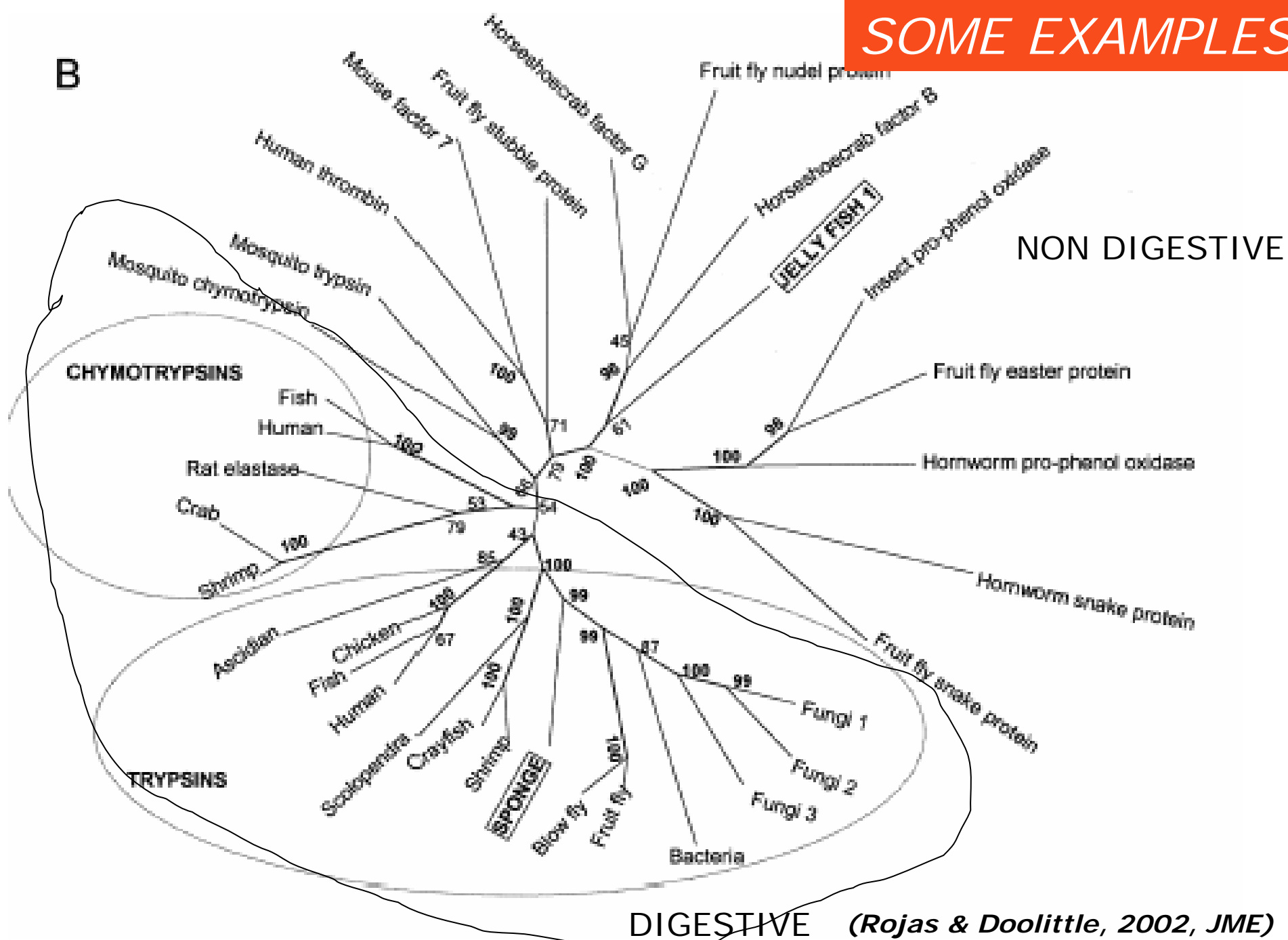


# SOME EXAMPLES



# SOME EXAMPLES

B



## SOME EXAMPLES

WHICH ONE IS THE REAL ONE?

WHAT IS THE ORIGIN OF THE CHYMOTRYPSIN FAMILY?

### ADDITIONAL INFORMATION:

- Sponge has a D189 diagnostic for trypsin (Hannenshalli & Russell, 2000)  
Jelly has N189.
- Codon for Serine at the active site:  
sponge signature for trypsin: TCT  
jelly: AGT,AGC
- When blasted against NR:  
sponge 48% with arthropod trypsin  
jelly 36% with RAT elastase

Disulfide bonds:

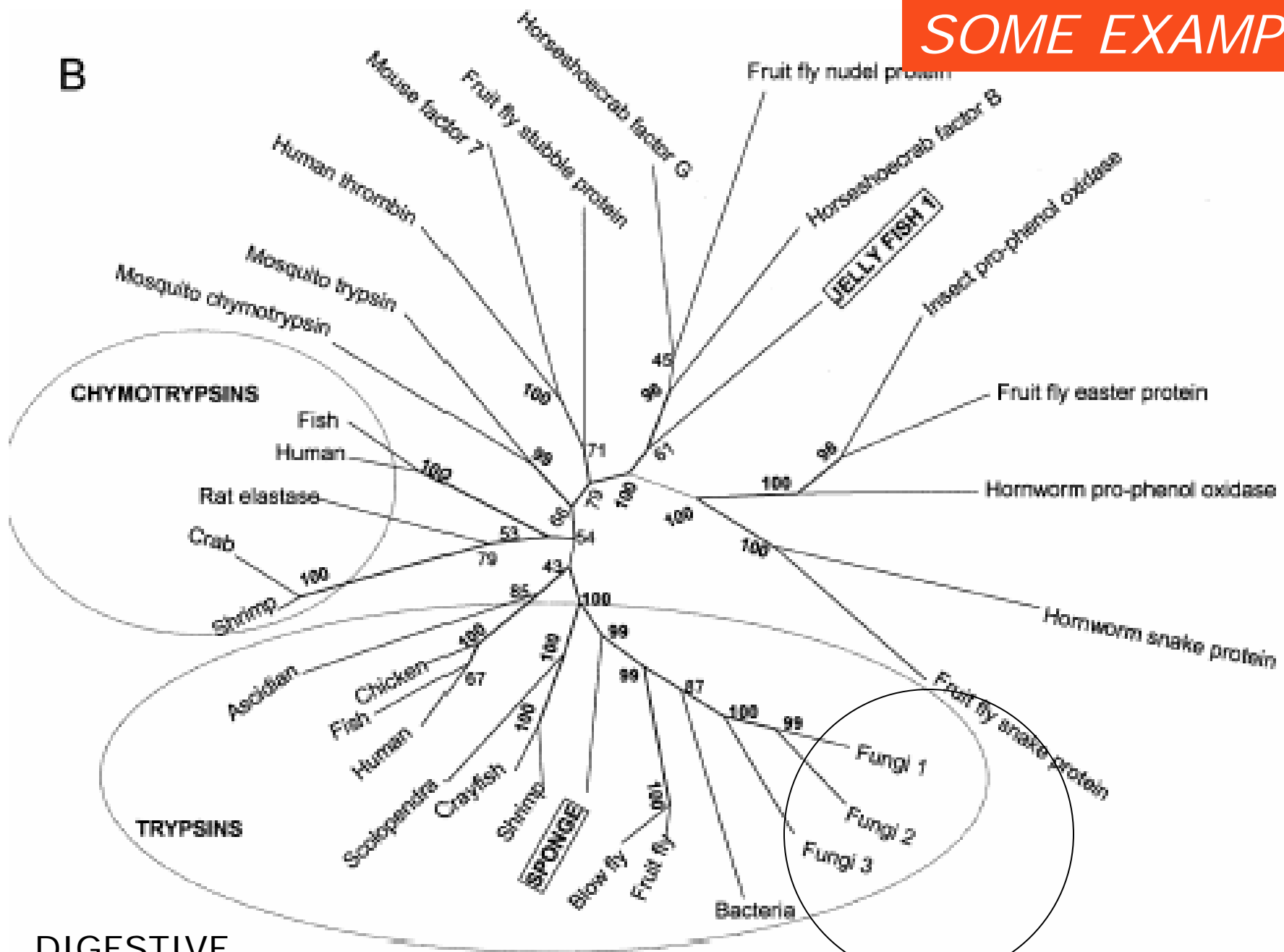
sponge 5 disulfide bonds and cys match with chymotrypsin-elastase (first tree)

Jelly has digestive system with organs, sponge are loose cells.

*(Rojas & Doolittle, 2002, JME)*

# SOME EXAMPLES

B



DIGESTIVE

(Rojas & Doolittle, 2002, JME)

# SOME EXAMPLES

## WHY THE FUNGAL ONES CLADE WITH ANIMALS?

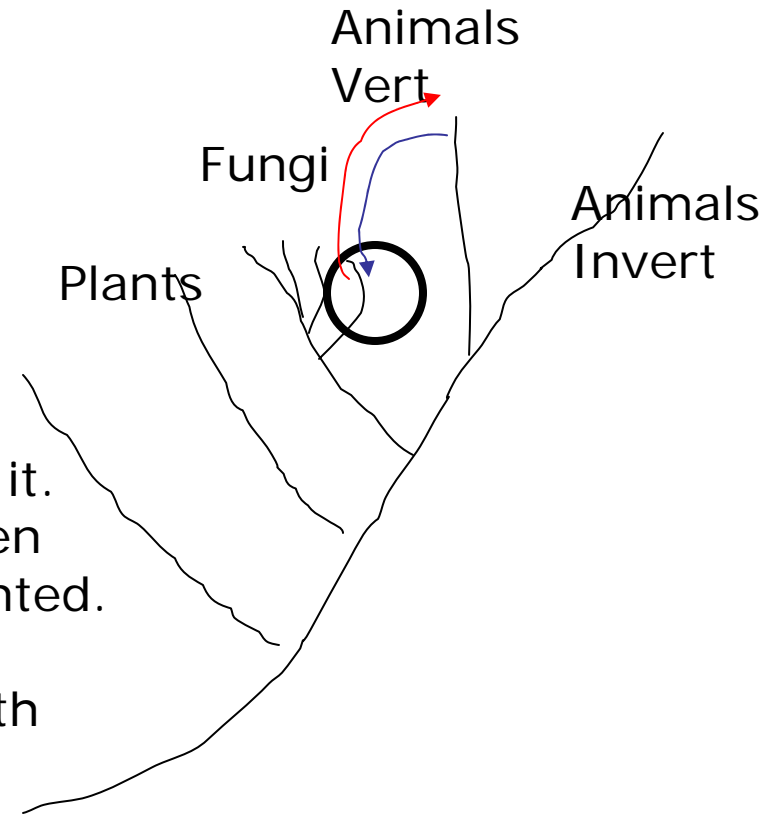
H.G.T!

SCENARIO1

Plants and all fungi-except *Streptomyces* lost it!  
Fungi should be more similar to jelly and sponge

SCENARIO1

then Plants and all fungi never had it. They appeared when digestion was invented. Fungi have them because HGT in both directions.



(Rojas & Doolittle, 2002, JME)

# Acknowledgements:

Frederik Ronquist for slides I borrowed

**THANK YOU!!**