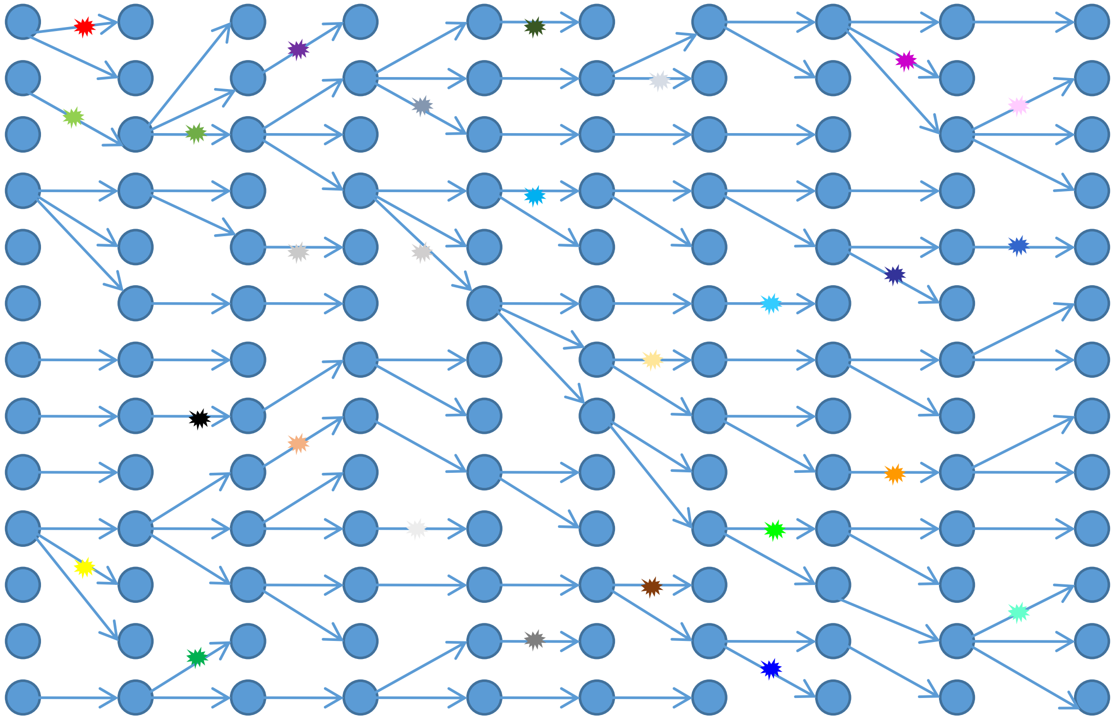
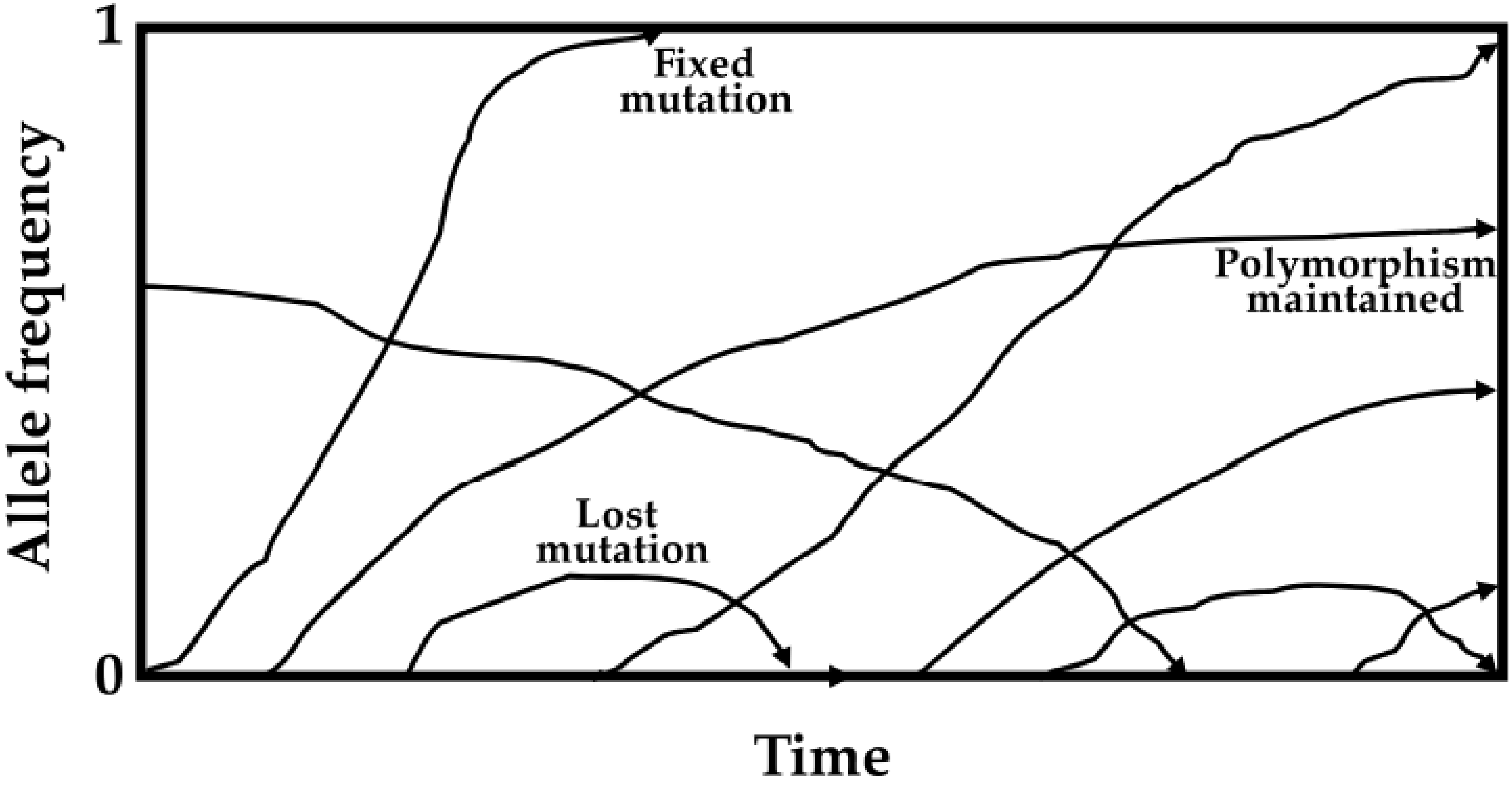
Session 3

Basics of Phylogenetics

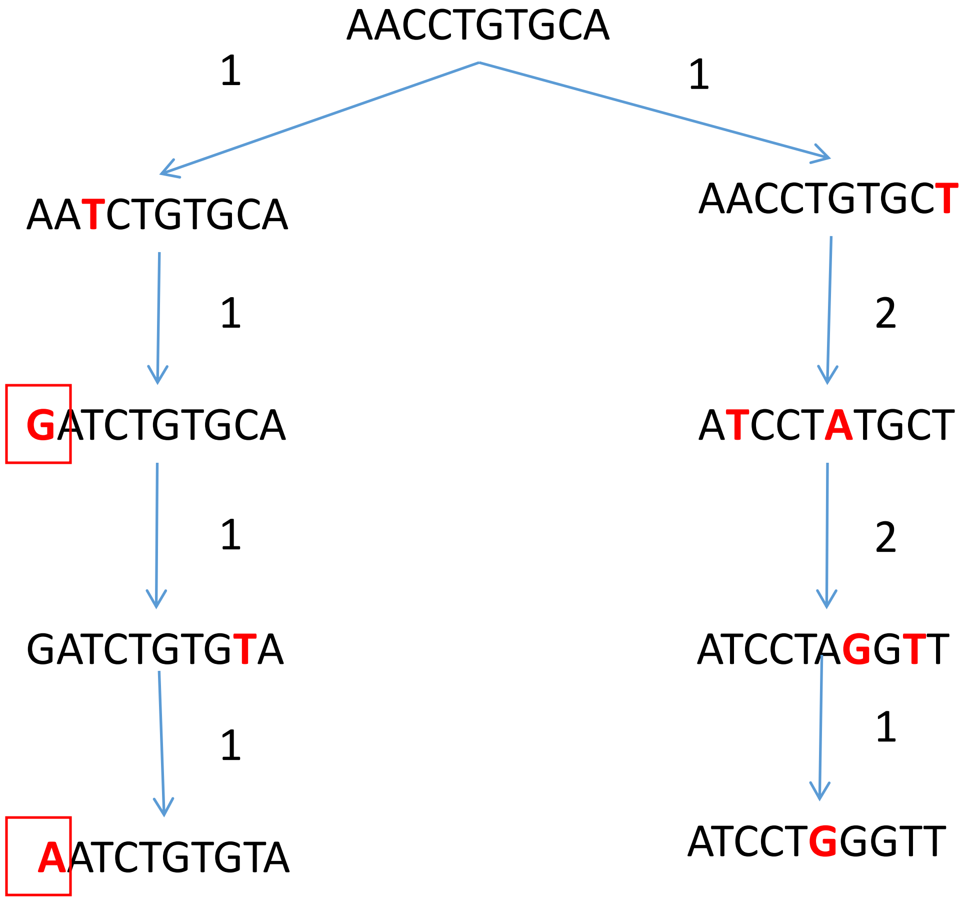
Why do we have to think in trees?



The fate of a mutation in a population



Relationship reflects **TIME** of divergence



AACCTGTGCA

1

1

AA**T**CTGTGCA

1

GATCTGTGCA

1

GATCTGTG**T**A

1

AATCTGTGTA

AACCTGTGC**T**

2

A**T**CCT**A**TGCT

2

ATCCTA**G**G**T**T

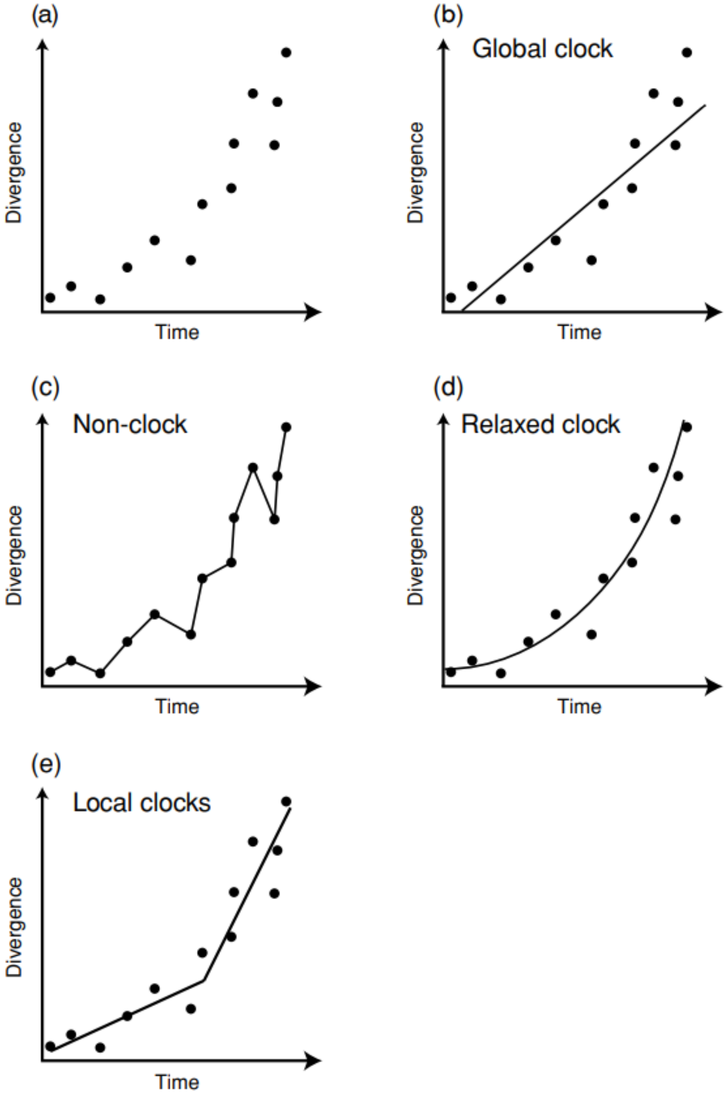
1

ATCCT**G**GGTT

On what depends the distance?

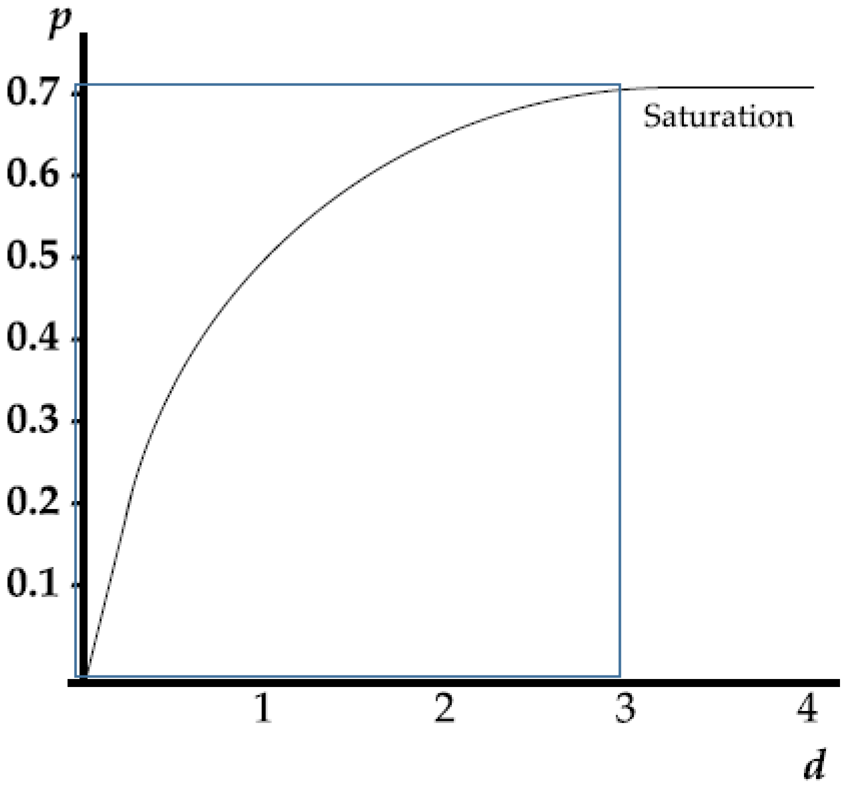
d-distance(Seq1,Seq2) = 10

Molecular clock



Relationship reflects TIME of divergence

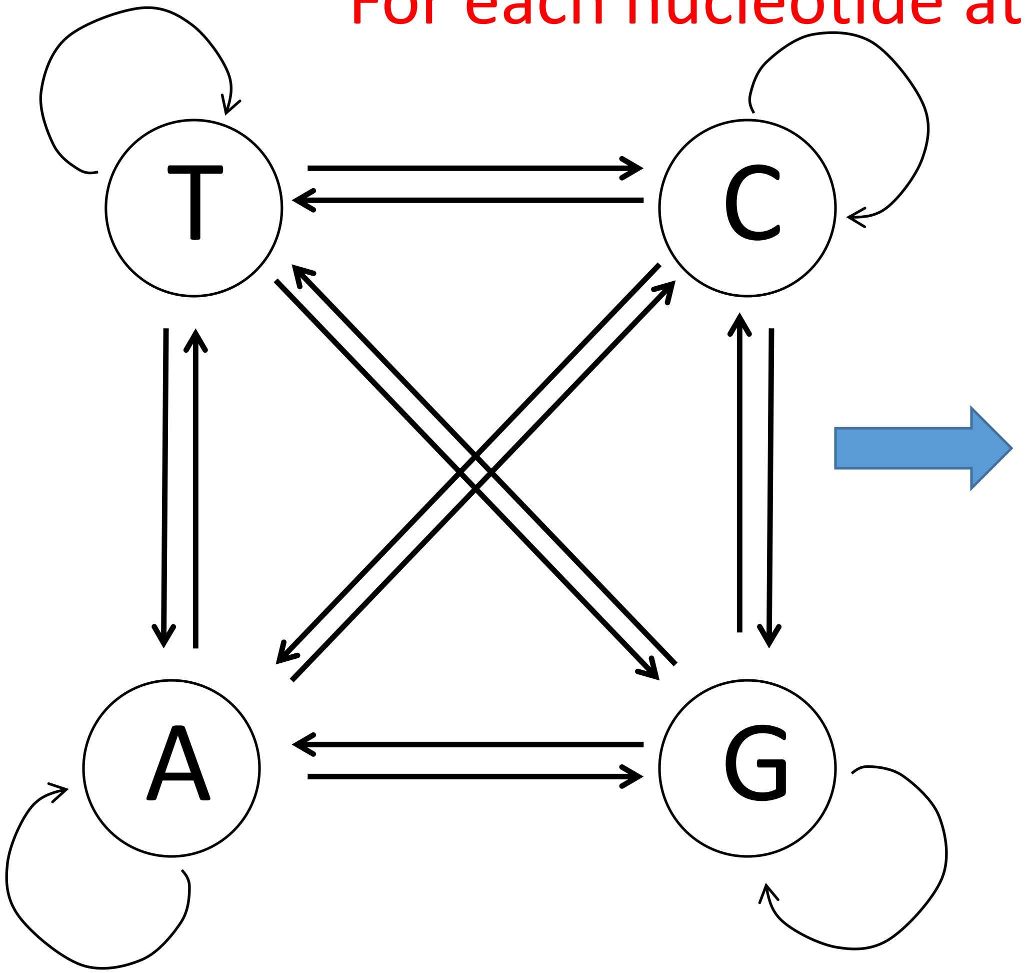
What it is saying about time of divergence and number of substitutions?



← We need a correction factor!

How to model this process?

For each nucleotide at generation t



Substitution rate matrix

Goes to

Q=

Comes from

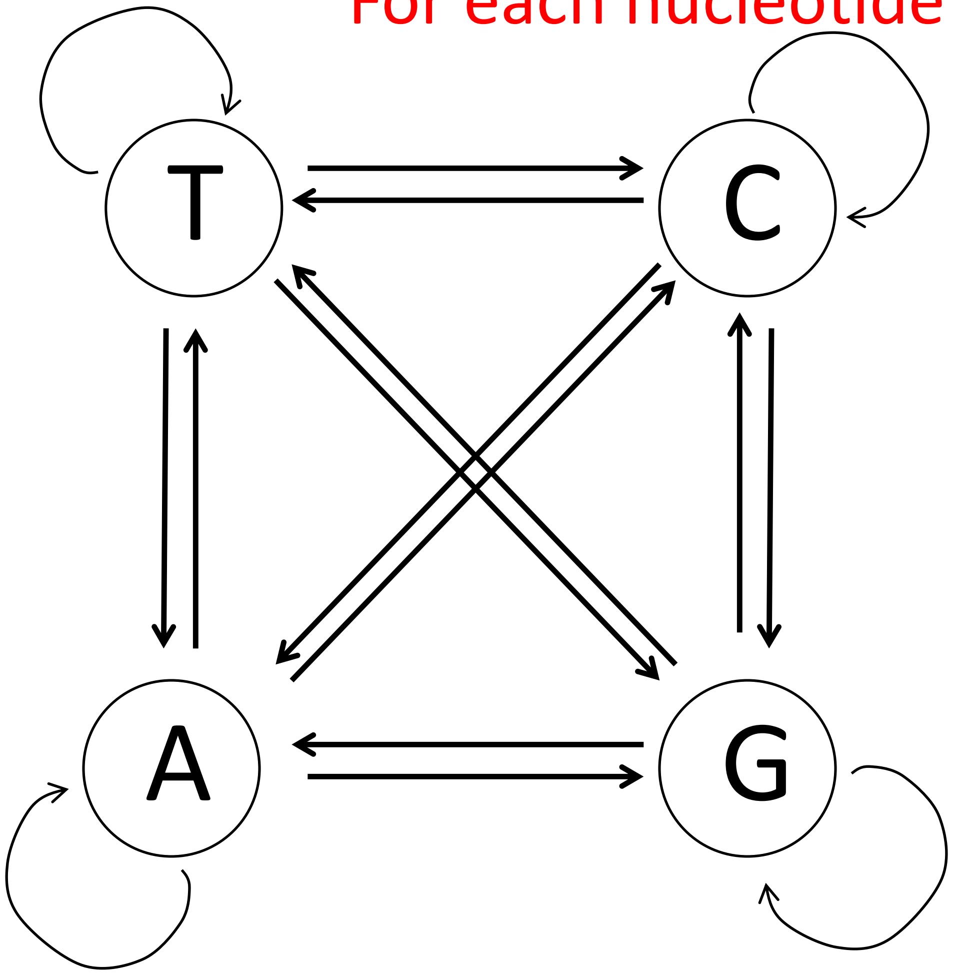
	A	C	T	G
A				
C				
T				
G				

$$\sum_{N\in\{A,T,C,G\}} q_{i,N} = 0$$

$$-\sum_{j\neq i} q_{i,j} = q_{ii}$$

How to model this process?

For each nucleotide at generation t



Probability change matrix

Goes to

P=

Comes from

	A	C	T	G
A				
C				
T				
G				

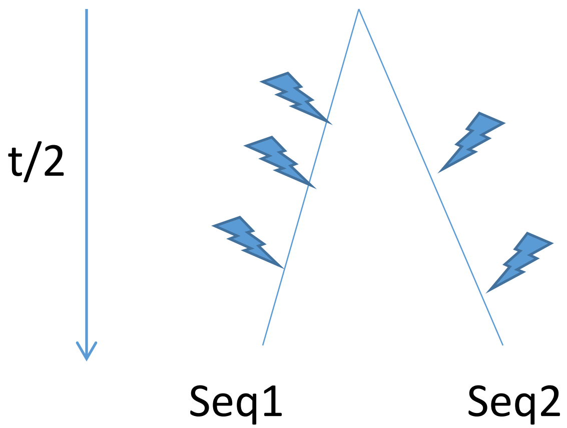
$$P(t) = e^{Qt}$$

$$Q = U \Lambda U^{-1},$$

$$P(t) = \mathrm{e}^{Qt} = U \operatorname{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \exp(\lambda_3 t), \exp(\lambda_4 t)\} U^{-1}.$$

How to model this process?

The Jukes-Kantor 69 model (JK69)



Frequency of observed changes x over n nucleotides

$$\hat{p} = \frac{x}{n}$$

Total substitution rate at any nucleotide

Goes to

$$Q= \text{Comes from} \begin{array}{|c|c|c|c|c|} \hline & \textbf{A} & \textbf{C} & \textbf{T} & \textbf{G} \\ \hline \textbf{A} & -3\lambda & \lambda & \lambda & \lambda \\ \hline \textbf{C} & \lambda & -3\lambda & \lambda & \lambda \\ \hline \textbf{T} & \lambda & \lambda & -3\lambda & \lambda \\ \hline \textbf{G} & \lambda & \lambda & \lambda & -3\lambda \\ \hline \end{array}$$

$$= 3\lambda$$

$$d = 3\lambda \text{t}$$

$$p = 3 * p_1(t) = \frac{3}{4} - \frac{3}{4} e^{-\frac{4d}{3}} = \hat{p}$$

$$\hat{d} = -\frac{3}{4} log\left(1 - \frac{4}{3}\hat{p}\right)$$

How to model this process?

- The Jukes-Kantor 69 model (JK69)

$Q=$

Goes to

Comes from

	A	C	T	G
A	-3λ	λ	λ	λ
C	λ	-3λ	λ	λ
T	λ	λ	-3λ	λ
G	λ	λ	λ	-3λ

Assumptions

- Equal mutation rate for all nucleotides
- In equilibrium, all nucleotide types have the same proportion

P=

Goes to

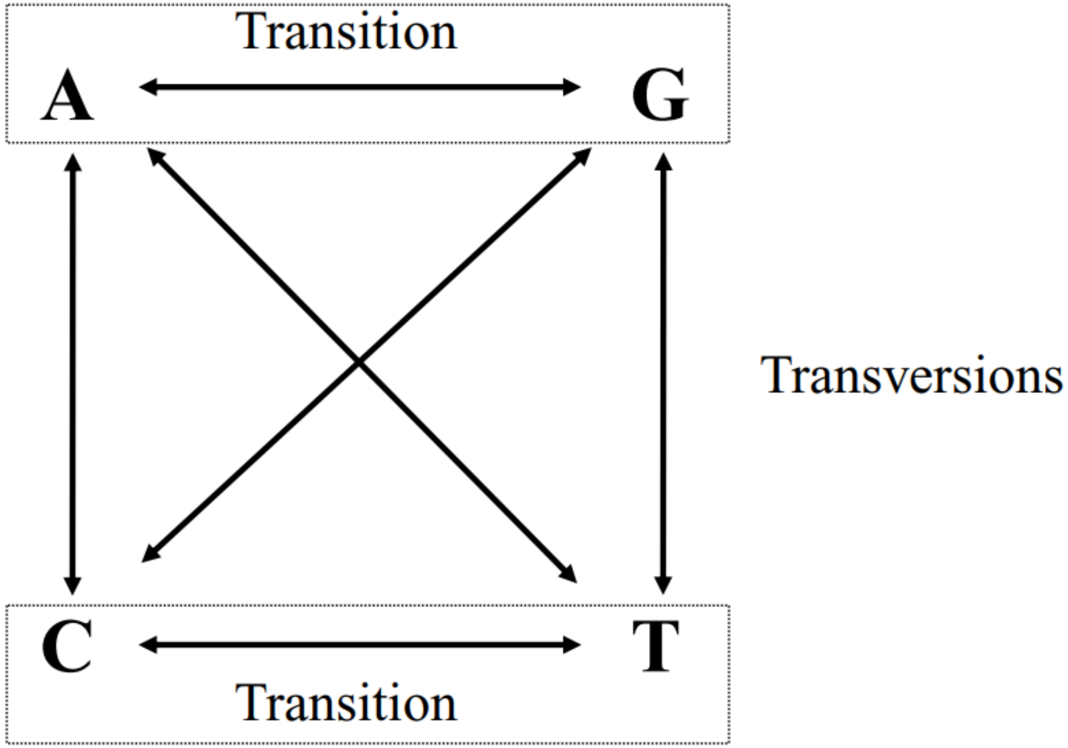
Comes from

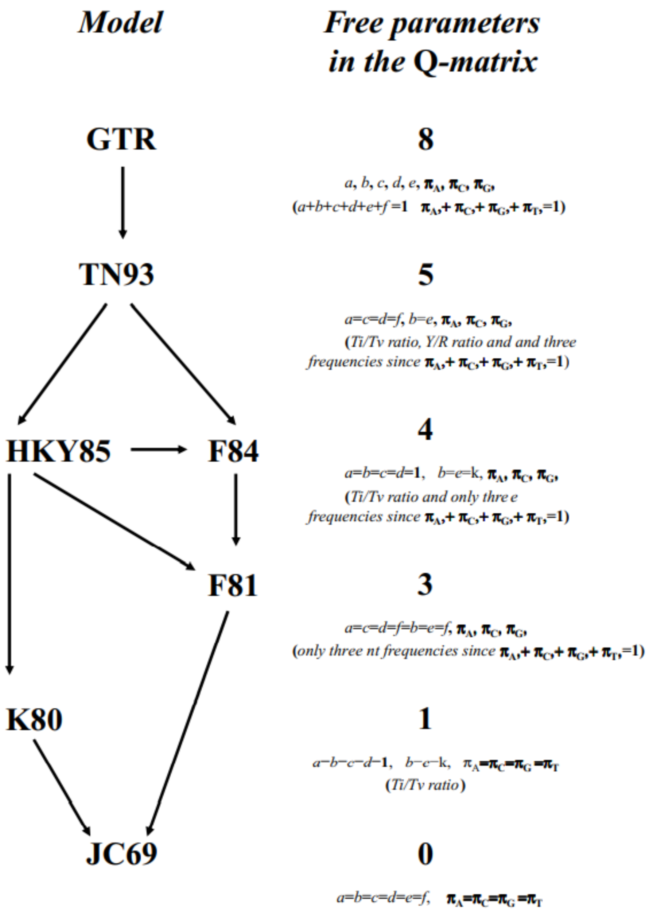
	A	C	T	G
A	p0(t)	p(1)	p(1)	p(1)
C	p1(t)	p(0)	p(1)	p(1)
T	p1(t)	p(1)	p(0)	p(1)
G	p1(t)	p(1)	p(1)	p(0)

$$p_0(t) = \frac{1}{4} + \frac{3}{4} e^{-4\lambda t}$$

$$p_1(t) = \frac{1}{4} - \frac{1}{4} e^{-4\lambda t}$$

Models of evolution





Free parameters in the Q-matrix

8

a, *b*, *c*, *d*, *e*, π_A, π_C, π_G,
(*a*+*b*+*c*+*d*+*e*+*f* =1 π_A,+ π_C,+ π_G,+ π_T,=1)

5

a=*c*=*d*=*f*, *b*=*e*, π_A, π_C, π_G,
(*Ti/Tv ratio, Y/R ratio and and three frequencies since* π_A,+ π_C,+ π_G,+ π_T,=1)

4

a=*b*=*c*=*d*=1, *b*=*e*=k, π_A, π_C, π_G,
(*Ti/Tv ratio and only thre e frequencies since* π_A,+ π_C,+ π_G,+ π_T,=1)

3

a=*c*=*d*=*f*=*b*=*e*=*f*, π_A, π_C, π_G,
(*only three nt frequencies since* π_A,+ π_C,+ π_G,+ π_T,=1)

1

a−*b*−*c*−*d*−1, *b*−*e*−k, π_A=π_C=π_G =π_T
(*Ti/Tv ratio*)

0

a=*b*=*c*=*d*=*e*=*f*, π_A=π_C=π_G =π_T

How to model this process?

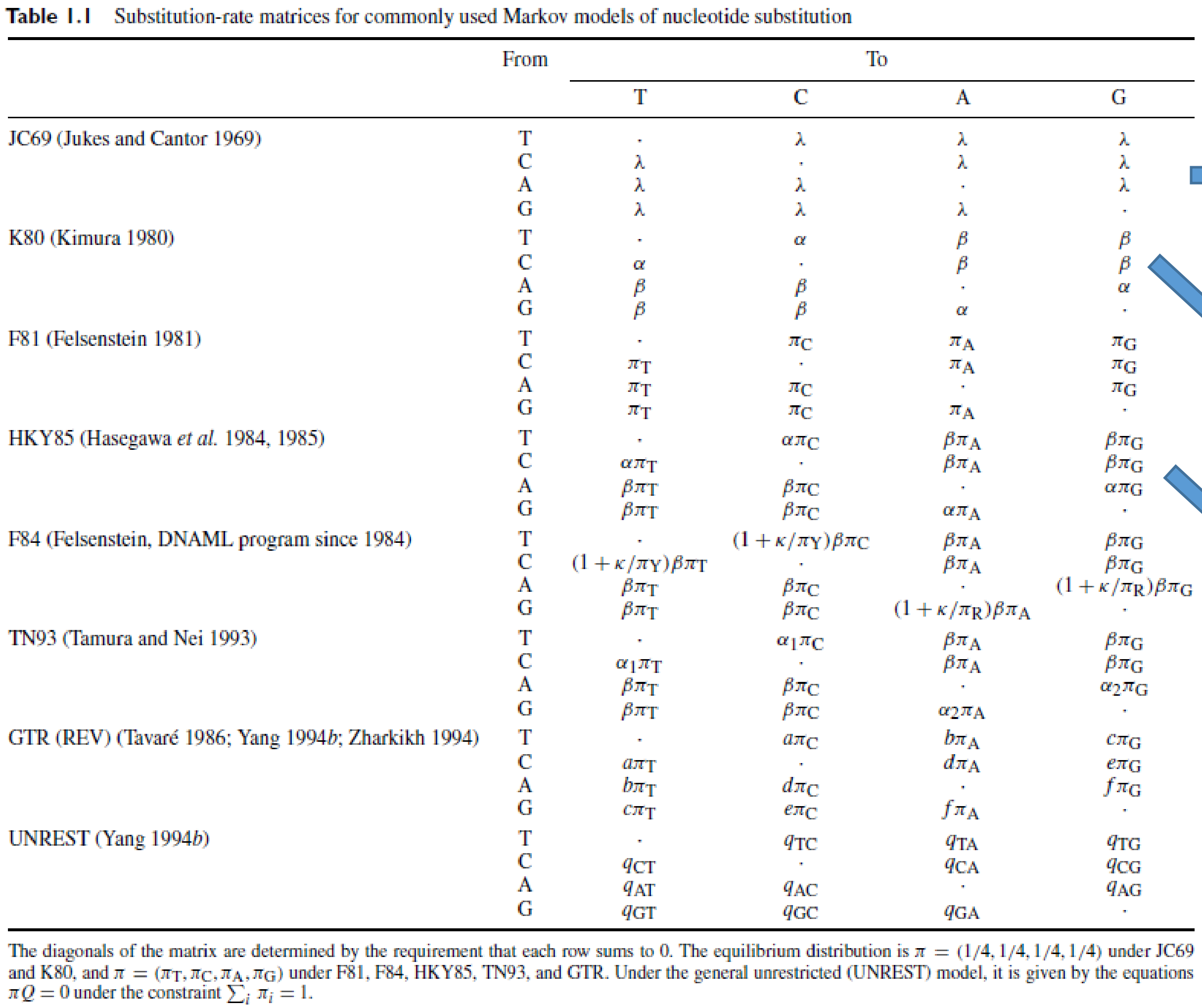
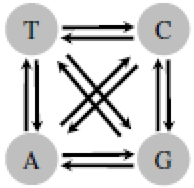


Table 1.1 Substitution-rate matrices for commonly used Markov models of nucleotide substitution

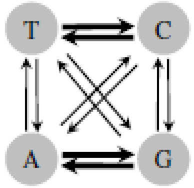
	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	$\cdot$	λ	λ	λ
	C	λ	$\cdot$	λ	λ
	A	λ	λ	$\cdot$	λ
	G	λ	λ	λ	$\cdot$
K80 (Kimura 1980)	T	$\cdot$	α	β	β
	C	α	$\cdot$	β	β
	A	β	β	$\cdot$	α
	G	β	β	α	$\cdot$
F81 (Felsenstein 1981)	T	$\cdot$	π_C	π_A	π_G
	C	π_T	$\cdot$	π_A	π_G
	A	π_T	π_C	$\cdot$	π_G
	G	π_T	π_C	π_A	$\cdot$
HKY85 (Hasegawa *et al.* 1984, 1985)	T	$\cdot$	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha\pi_T$	$\cdot$	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	$\cdot$	$\alpha\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	$\cdot$
F84 (Felsenstein, DNAML program since 1984)	T	$\cdot$	$(1+\kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$(1+\kappa/\pi_Y)\beta\pi_T$	$\cdot$	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	$\cdot$	$(1+\kappa/\pi_R)\beta\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$(1+\kappa/\pi_R)\beta\pi_A$	$\cdot$
TN93 (Tamura and Nei 1993)	T	$\cdot$	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha_1\pi_T$	$\cdot$	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	$\cdot$	$\alpha_2\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	$\cdot$
GTR (REV) (Tavaré 1986; Yang 1994*b*; Zharkikh 1994)	T	$\cdot$	$a\pi_C$	$b\pi_A$	$c\pi_G$
	C	$a\pi_T$	$\cdot$	$d\pi_A$	$e\pi_G$
	A	$b\pi_T$	$d\pi_C$	$\cdot$	$f\pi_G$
	G	$c\pi_T$	$e\pi_C$	$f\pi_A$	$\cdot$
UNREST (Yang 1994*b*)	T	$\cdot$	q_{TC}	q_{TA}	q_{TG}
	C	q_{CT}	$\cdot$	q_{CA}	q_{CG}
	A	q_{AT}	q_{AC}	$\cdot$	q_{AG}
	G	q_{GT}	q_{GC}	q_{GA}	$\cdot$

The diagonals of the matrix are determined by the requirement that each row sums to 0. The equilibrium distribution is $\pi = (1/4, 1/4, 1/4, 1/4)$ under JC69 and K80, and $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ under F81, F84, HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by the equations $\pi Q = 0$ under the constraint $\sum_i \pi_i = 1$.

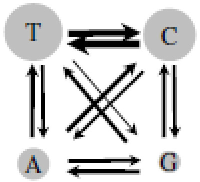
JC69



K80

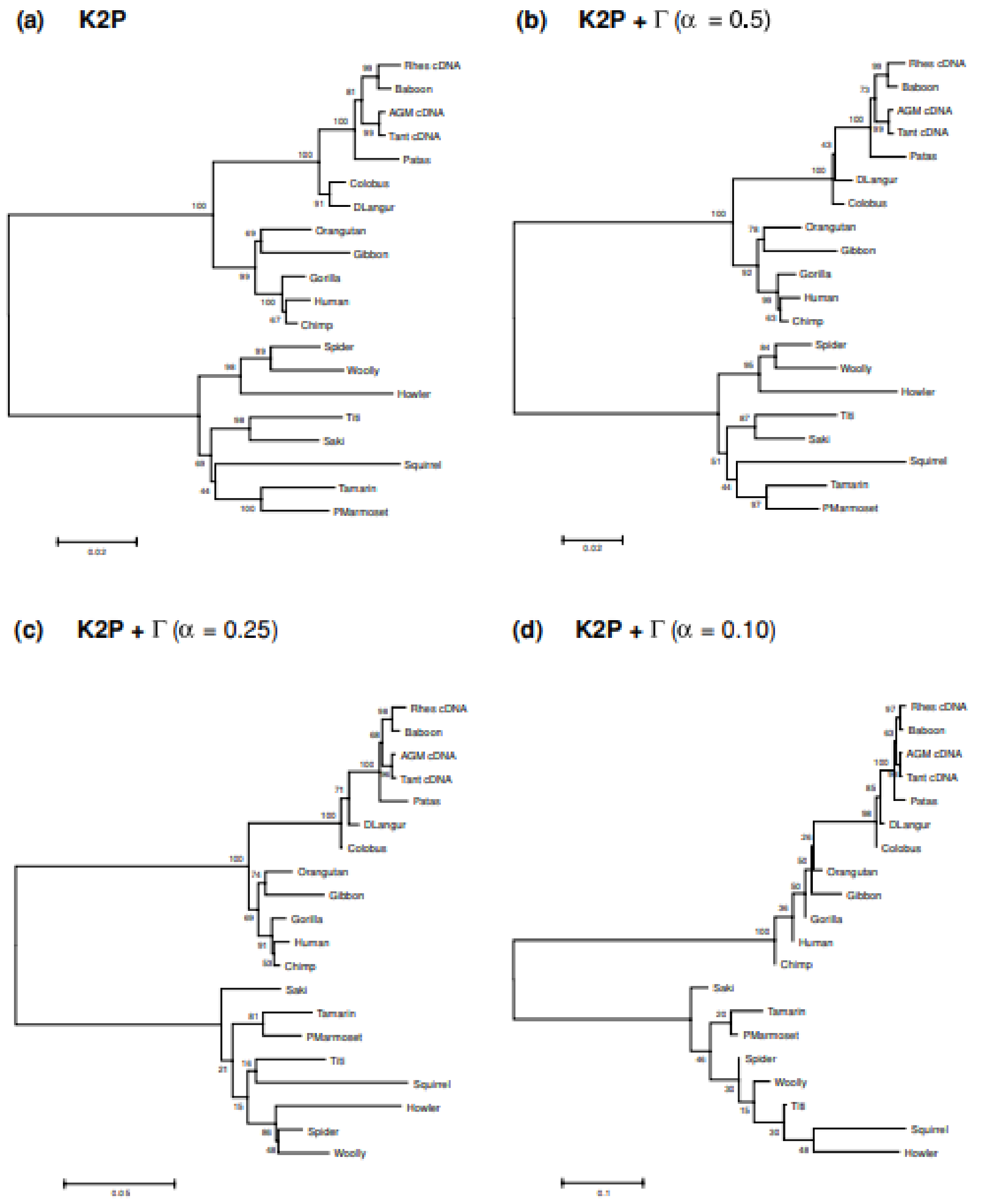


HKY85



Models of evolution

Different lengths can be obtained depending on the evolutionary model!



Which is the best model of evolution?

- “there is a trade-off. More parameters allow a more realistic way of representing the underlying data. But this comes with the danger that too many parameters may **over-fit** the underlying data (overparametrization), resulting in errors during parameter estimation (Sullivan and Joyce 2005). In contrast, simplified models may not realistically represent the data, which can also mislead phylogenetic reconstruction.”

Which is the best one?

- likelihood ratio test

Where M0 is more simple than M1

$$\left.\begin{array}{l}\text{M1 -> P(D| M1)}\\ \text{M0 -> P(D| M0)}\end{array}\right\} \quad LRT = 2 * log\left(\frac{P(D|M_1)}{P(D|M_0)}\right)$$

$$LRT \sim \chi^2\,; df = parameters(M_1) - parameters(M_2)\,;$$

Which is the best one?

$$LRT = 2 * log\left(\frac{P(D|M_1)}{P(D|M_0)}\right) \gg 2$$

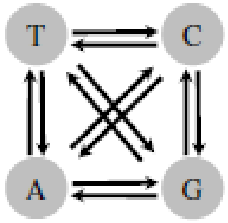
$$LRT \sim \chi^2\ ; df\ = parameters(M_1) - parameters(M_2)\ ;$$

“Including more parameters (M$_1$) substantially improves the likelihood”

Comparison of two nested models

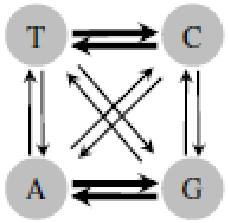
	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	·	λ	λ	λ
	C	λ	·	λ	λ
	A	λ	λ	·	λ
	G	λ	λ	λ	·
K80 (Kimura 1980)	T	·	α	β	β
	C	α	·	β	β
	A	β	β	·	α
	G	β	β	α	·

JC69



One parameter

K80



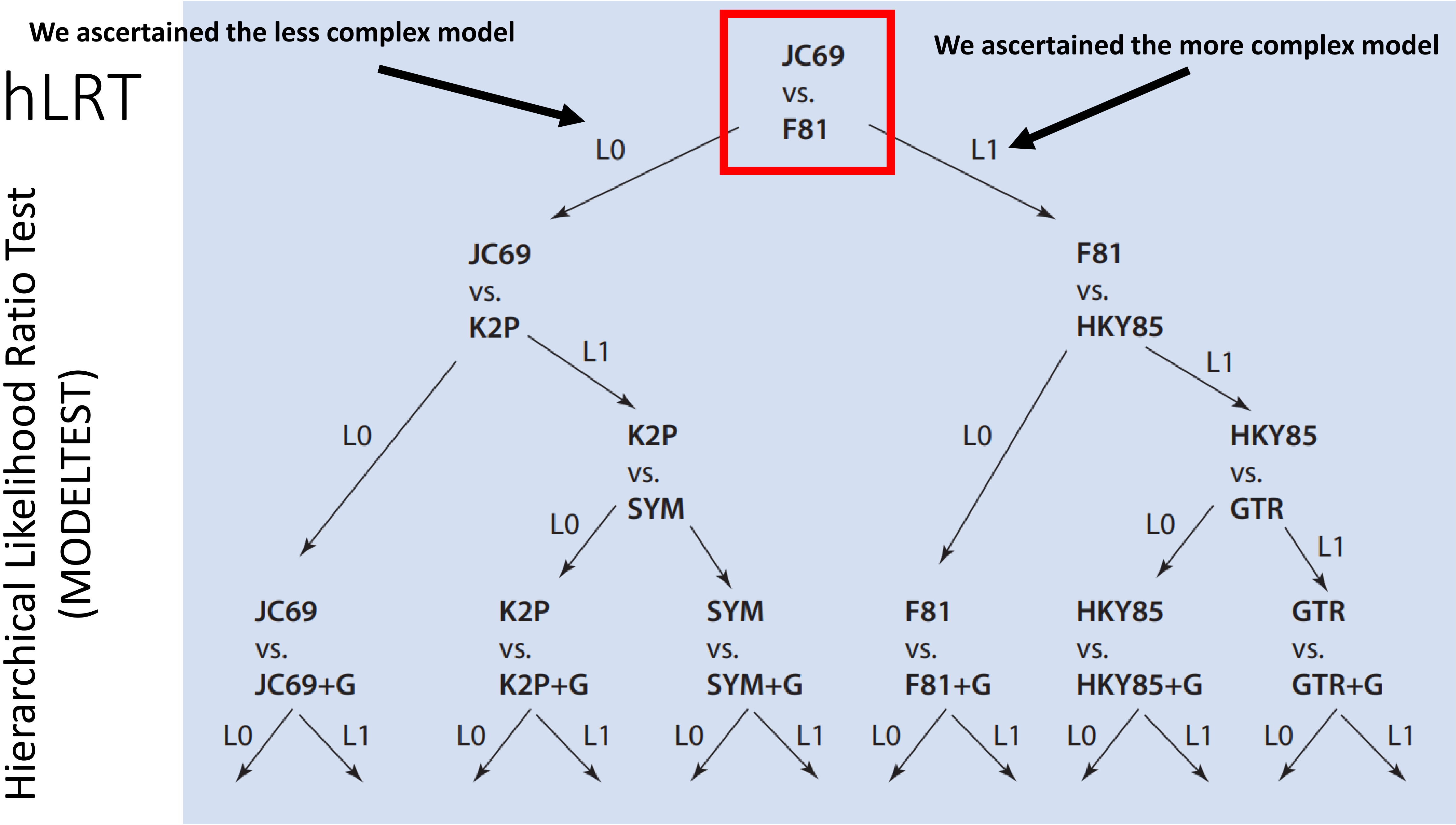
Two parameters

JC69 is a case of *K80* when $\alpha = \beta$

JC69 is *nested* in K80

hLRT

Hierarchical Likelihood Ratio Test (MODELTEST)



Other methods

Akaike Information Criteria

$$\mathrm{AIC} = -2\log_e L_i + 2\boxed{K_i}$$

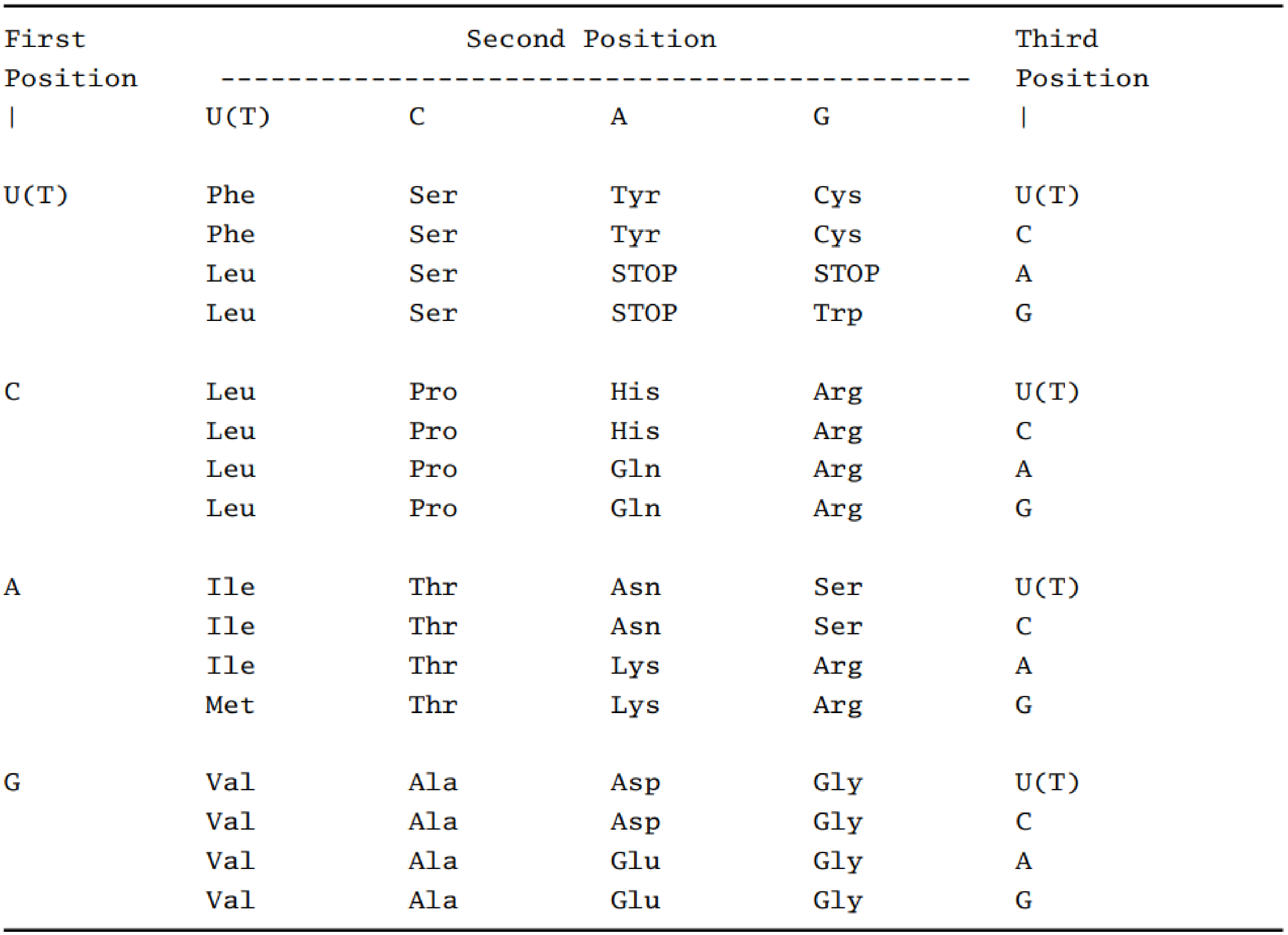
Free parameters

Bayesian Information Criteria

Sequence length

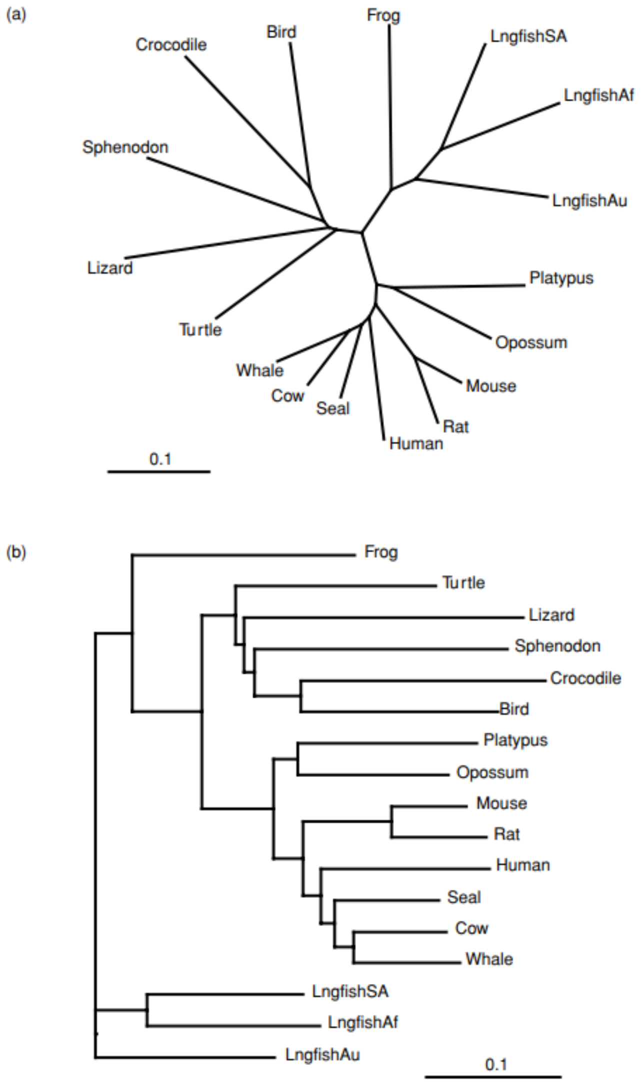
$$\mathrm{BIC} = 2\log_e L_i + K_i \log_e \boxed{n}$$

Remember that the same can be applied to proteins!



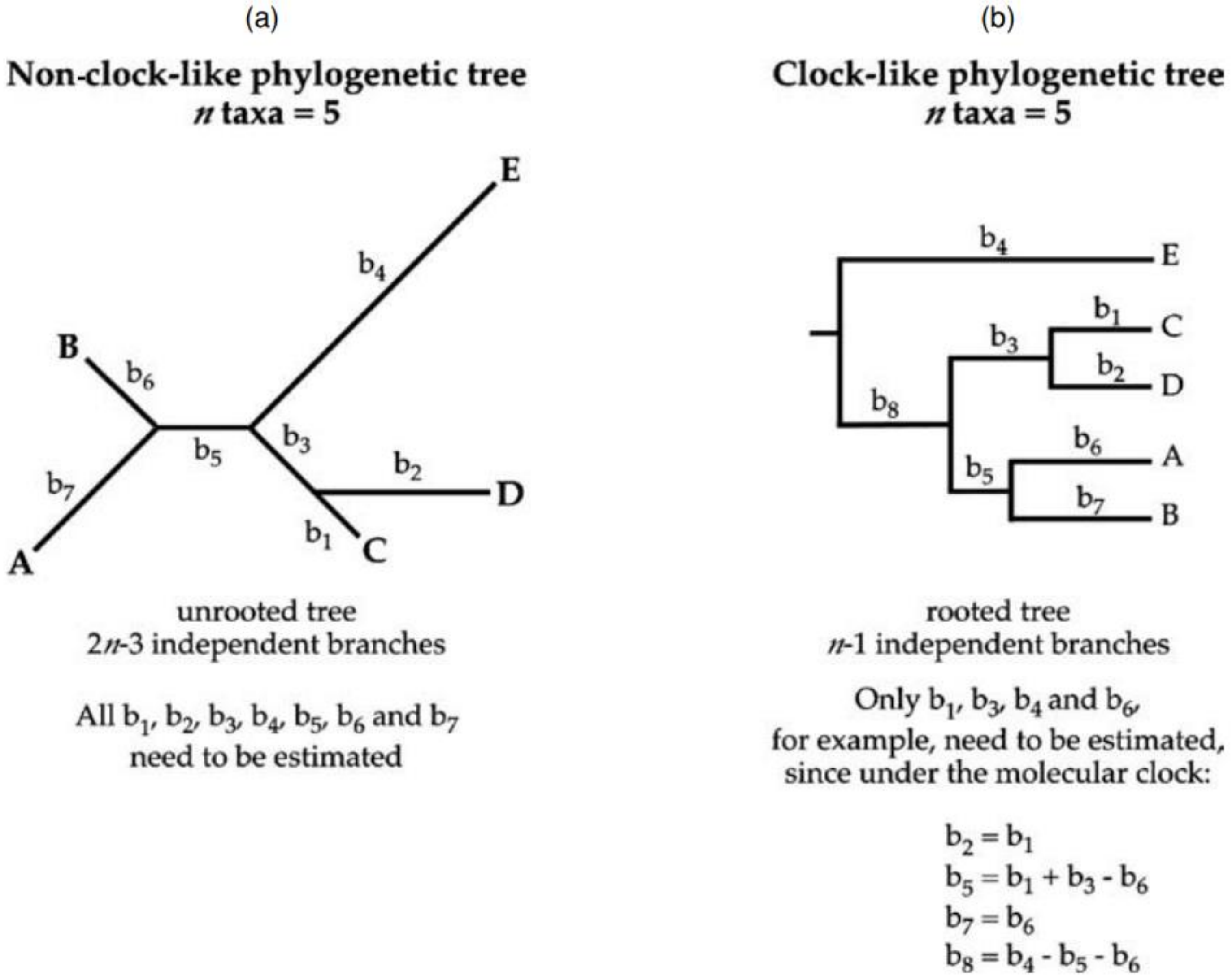
First Position	Second Position				Third Position
|	U(T)	C	A	G	|
U(T)	Phe	Ser	Tyr	Cys	U(T)
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U(T)
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U(T)
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U(T)
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

What is a phylogenetic tree?



Unrooted, rooted and the concept of outgroup

What is a phylogenetic tree?



(a)

Non-clock-like phylogenetic tree
n taxa = 5

unrooted tree
$2n$-3 independent branches

All b_1, b_2, b_3, b_4, b_5, b_6 and b_7 need to be estimated

(b)

Clock-like phylogenetic tree
n taxa = 5

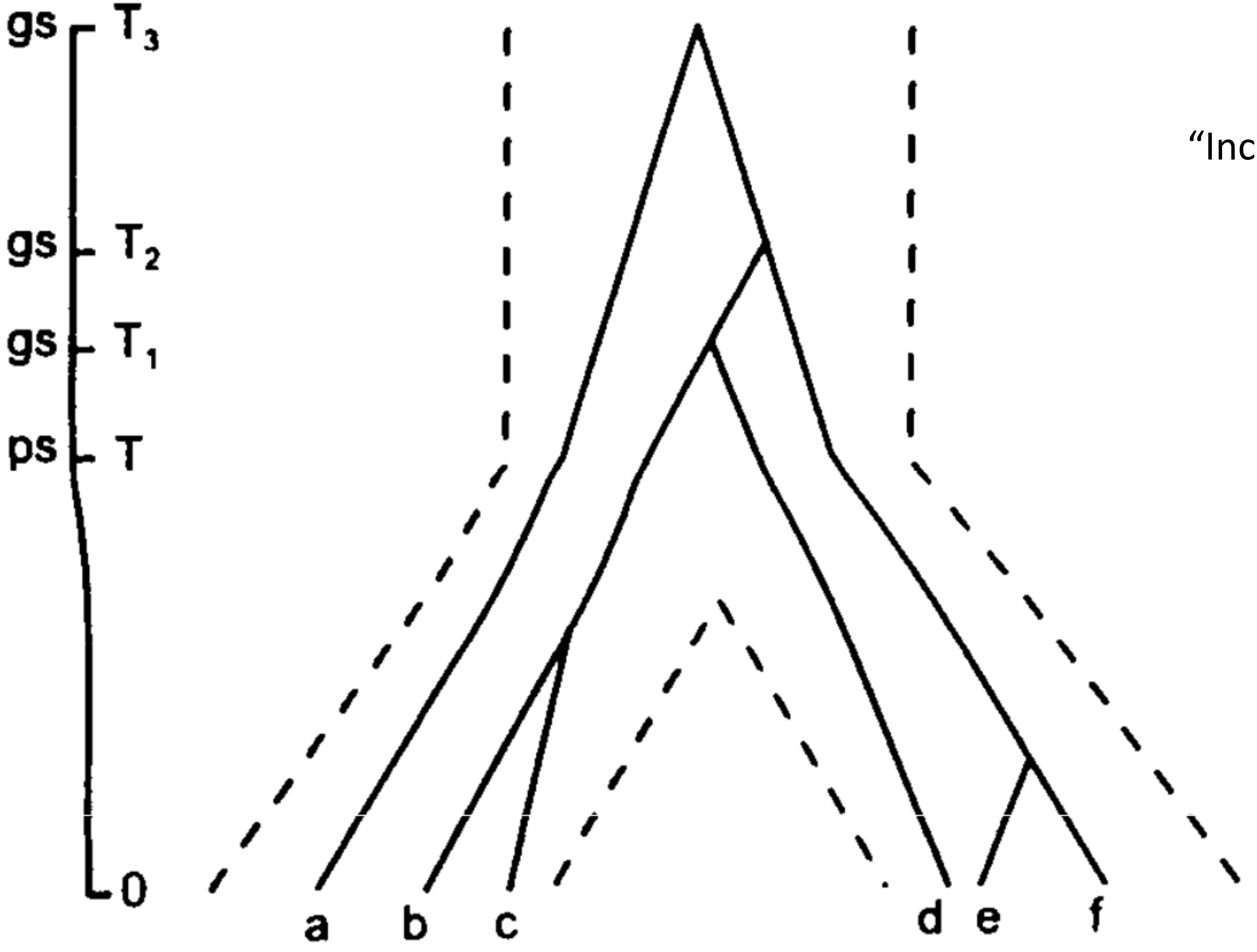
rooted tree
n-1 independent branches

Only b_1, b_3, b_4 and b_6, for example, need to be estimated, since under the molecular clock:

$$b_2 = b_1$$
$$b_5 = b_1 + b_3 - b_6$$
$$b_7 = b_6$$
$$b_8 = b_4 - b_5 - b_6$$

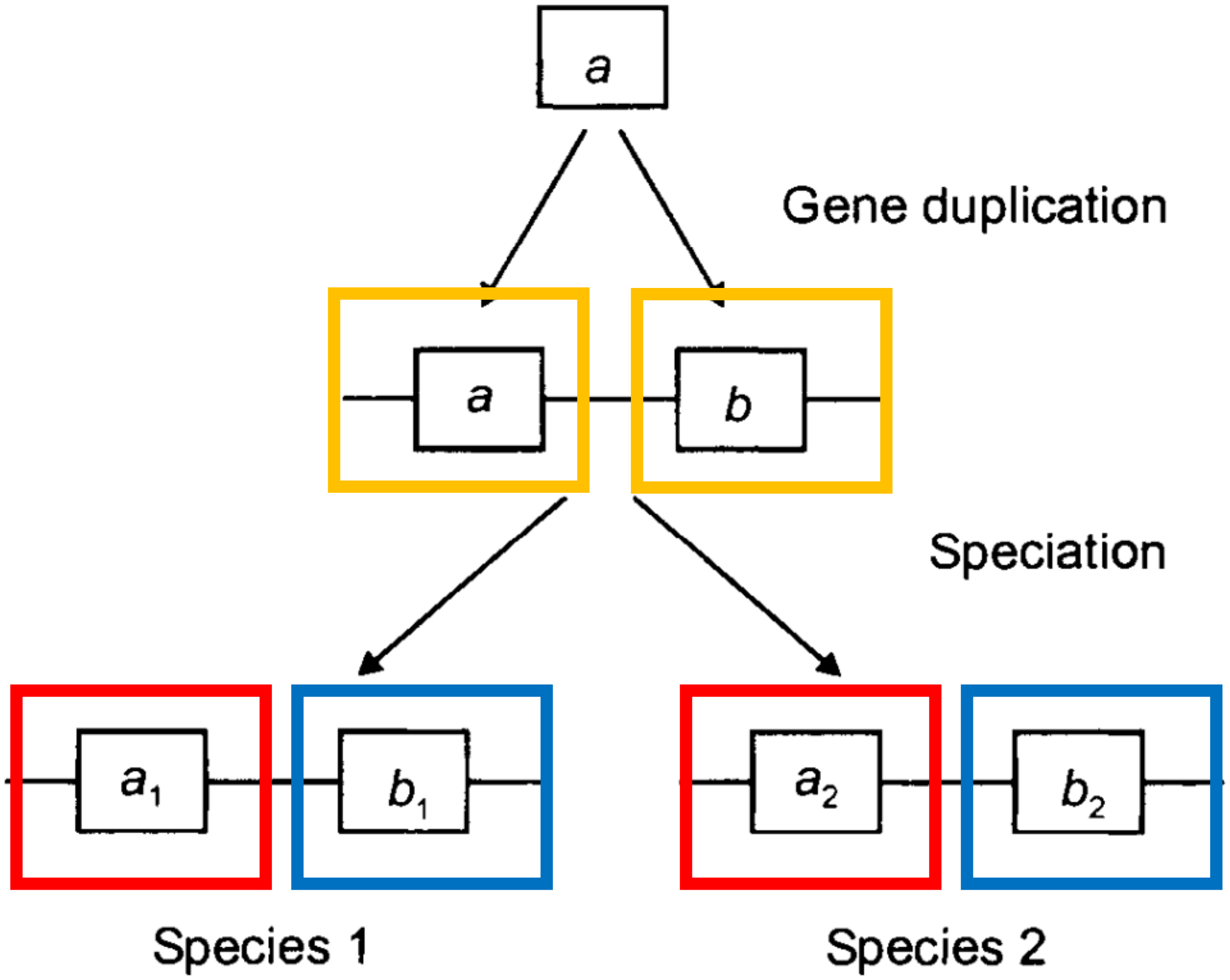
Evolutionary constraints

Species evolution vs phylogenetic trees



“Incomplete lineage sorting”

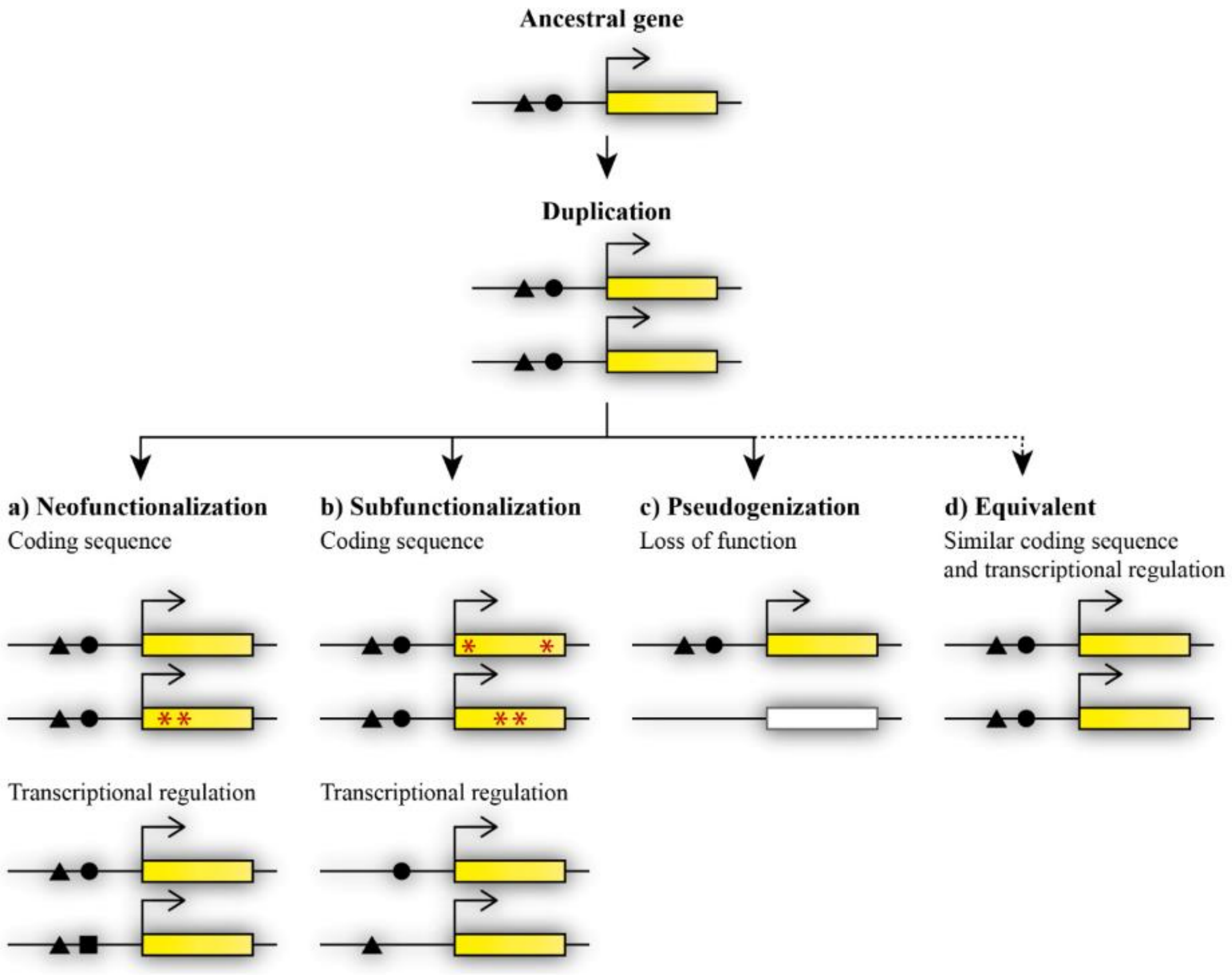
Species evolution vs phylogenetic trees



Paralogous genes

Orthologous genes

Species evolution vs phylogenetic trees



Ancestral gene

Duplication

a) Neofunctionalization

Coding sequence

Transcriptional regulation

b) Subfunctionalization

Coding sequence

Transcriptional regulation

c) Pseudogenization

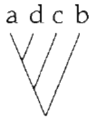
Loss of function

d) Equivalent

Similar coding sequence and transcriptional regulation

How do we get a phylogenetic tree?

OTU	1	2	3	4	5	6
a	A	T	A	T	A	C
b	A	T	C	T	A	C
c	G	T	C	G	A	C
d	T	T	C	G	T	C



Based on input

- Distance matrix based methods
    - Least squares method
    - UPGMA
    - Neighbour-joining
- Character based methods
    - Maximum Parsimony (MP)
    - Maximum Likelihood
    - Bayesian

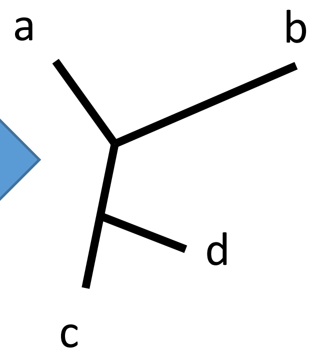
Based on reconstruction method

- Algorithmic
    - Least squares method
    - Neighbour-joining
- Optimality
    - UPGMA
    - Maximum Parsimony (MP)
    - Maximum Likelihood

Distance based methods

OTU	**1**	**2**	**3**	**4**	**5**	**6**
a	*A*	*T*	*A*	*T*	*A*	*C*
b	*A*	*T*	*C*	*T*	*A*	*C*
c	*G*	*T*	*C*	*G*	*A*	*C*
d	*T*	*T*	*C*	*G*	*T*	*C*

OTU	**a**	**b**	**c**	**d**
a	0	.	.	.
b	Db,a	0	.	.
c	Dc,a	Dc,b	0	.
d	Dd,a	Dd,b	Dd,c	0



How do you define “Distance”

How do you build the tree

```
3 T T C A A T C A G G C C C G A
    | |   |           | |
1 T C A A G T C A G G T T C G A
      |       |     | |
2 T C C A G T T A G A C T C G A
    |     |   |     |   |
3 T T C A A T C A G G C C C G A
```

Convert dissimilarity into evolutionary distance by correcting for multiple events per site, e.g. Jukes & Cantor (1969):

$$d_{AB} = -\frac{3}{4}\ln\left(1-\frac{4}{3}\,0.266\right) = 0.328$$

	1	2	3
2	0.266		
3	0.333	0.333	

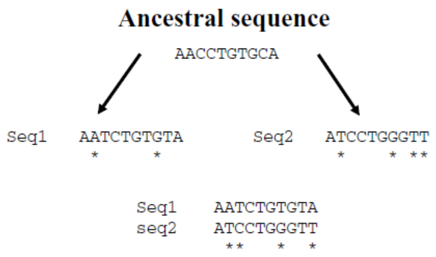
Dissimilarities

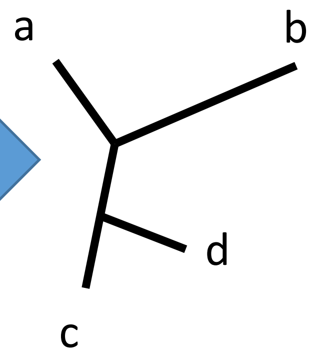
	1	2	3
2	0.328		
3	0.441	0.441	

Evolutionary distances

Character based methods

OTU	1	2	3	4	5	6
a	*A*	*T*	*A*	*T*	*A*	*C*
b	*A*	*T*	*C*	*T*	*A*	*C*
c	*G*	*T*	*C*	*G*	*A*	*C*
d	*T*	*T*	*C*	*G*	*T*	*C*



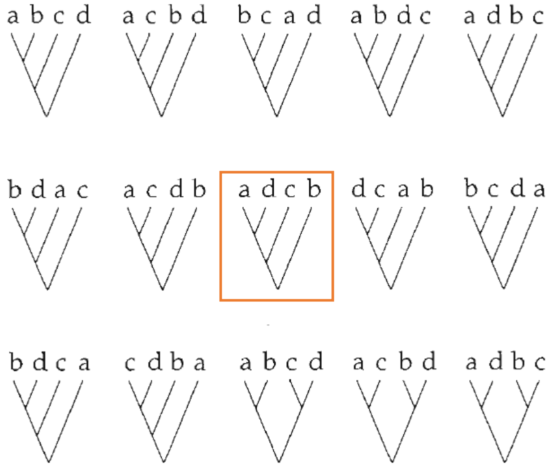


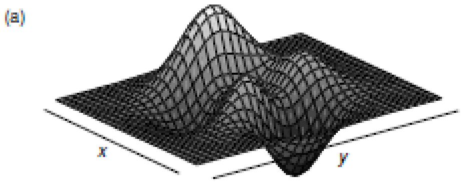
Model that explains how data is generated

How do you build the tree

How do we get THE phylogenetic tree?

OTU	1	2	3	4	5	6
a	A	T	A	T	A	C
b	A	T	C	T	A	C
c	G	T	C	G	A	C
d	T	T	C	G	T	C





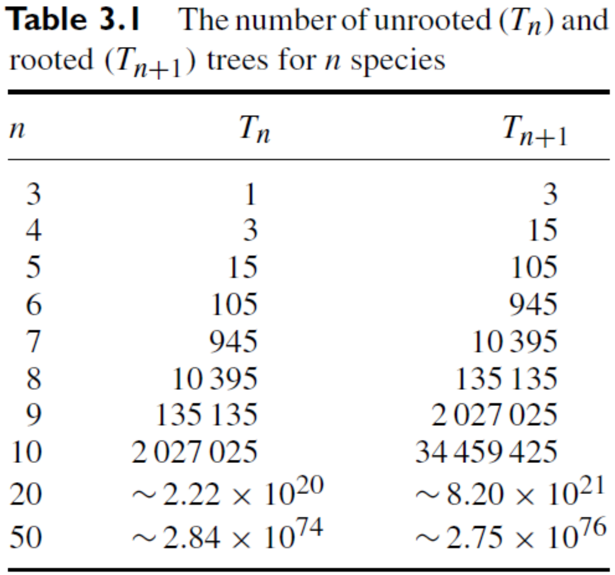
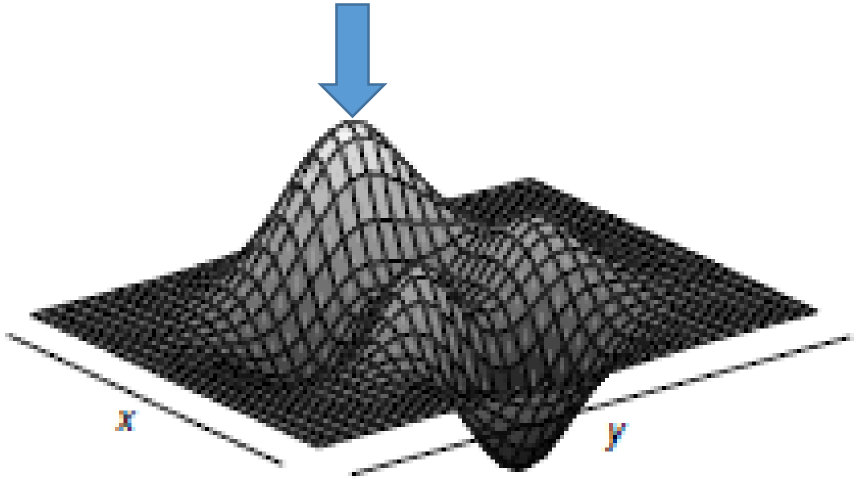


Table 3.1 The number of unrooted (T_n) and rooted (T_{n+1}) trees for n species

n	T_n	T_{n+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
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$

A classical optimization problem

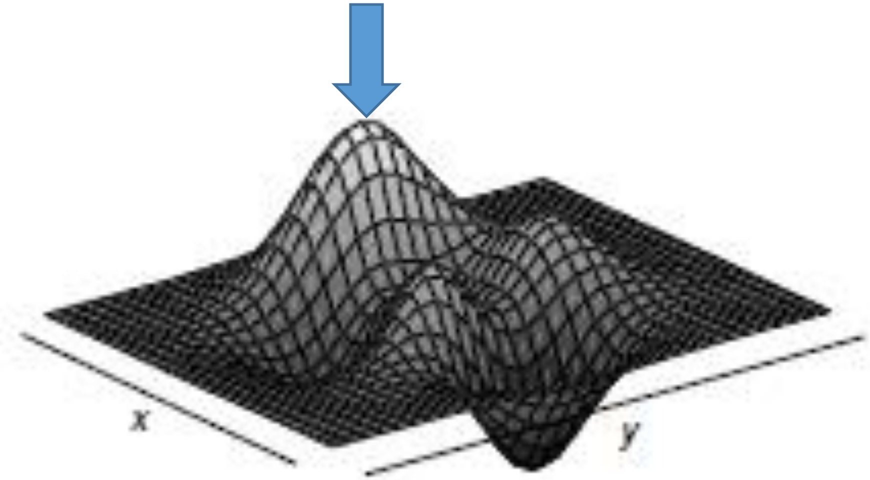
How do we reach the top of the mountain?



$$\frac{dx}{dy} = 0 \qquad \textbf{?}$$

A classical optimization problem

How do we reach the top of the mountain?

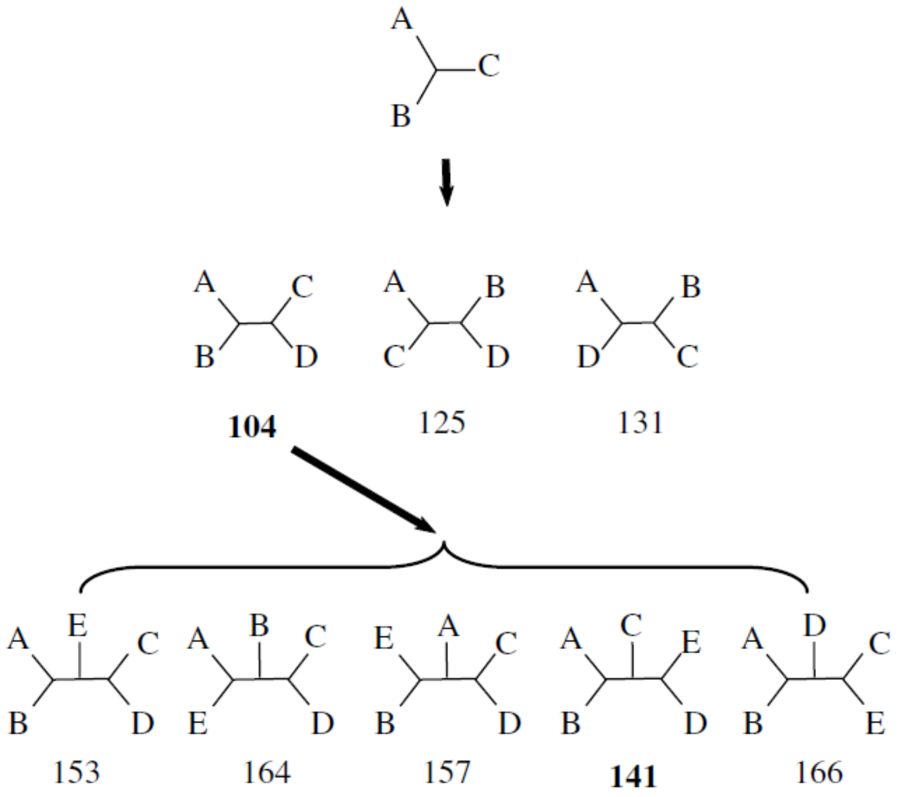


Finding the “best” tree

- Heuristic
  - Hierarchical clustering algorithms
    - Agglomerative
      - Stepwise addition/sequential addition
    - Divisive
      - Star decomposition
  - Tree rearrangement
    - Pruning
    - Regrafting
    - Nearest-neighbour interchange

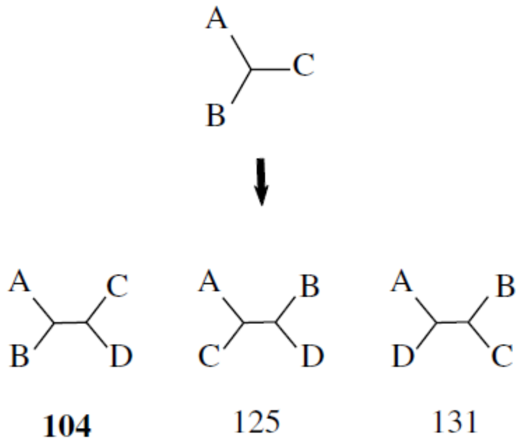
Finding the "best" tree

- Stepwise addition



Finding the “best” tree

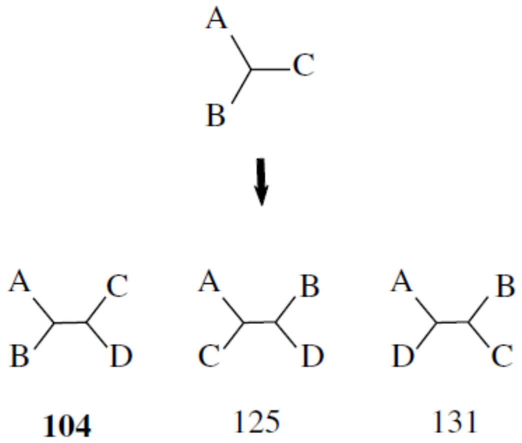
- Stepwise addition



- Which objects do we need for the pseudocode?
- Which functions do we need for the pseudocode?

Finding the “best” tree

- Stepwise addition



Classes

Tree contains ***Branch***es

Branch contains Two ***Node***s

Node contains sub-***Tree***s

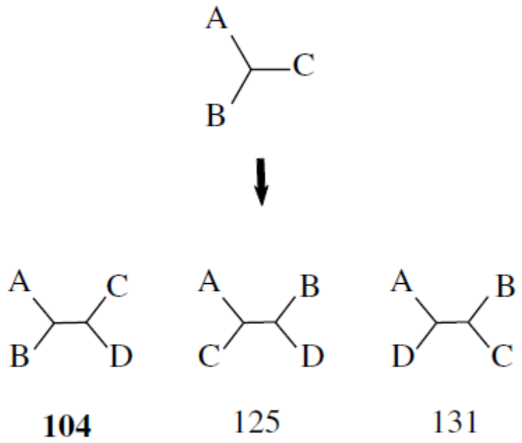
Leaf is a particular type of ***Node***

Functions

insert_Leaf(branch, leaf)
cost(tree)
initialize_tree(list_of_leafs)
copy_tree(test)
next_leaf(list_of_leafs)

Finding the “best” tree

- Stepwise addition

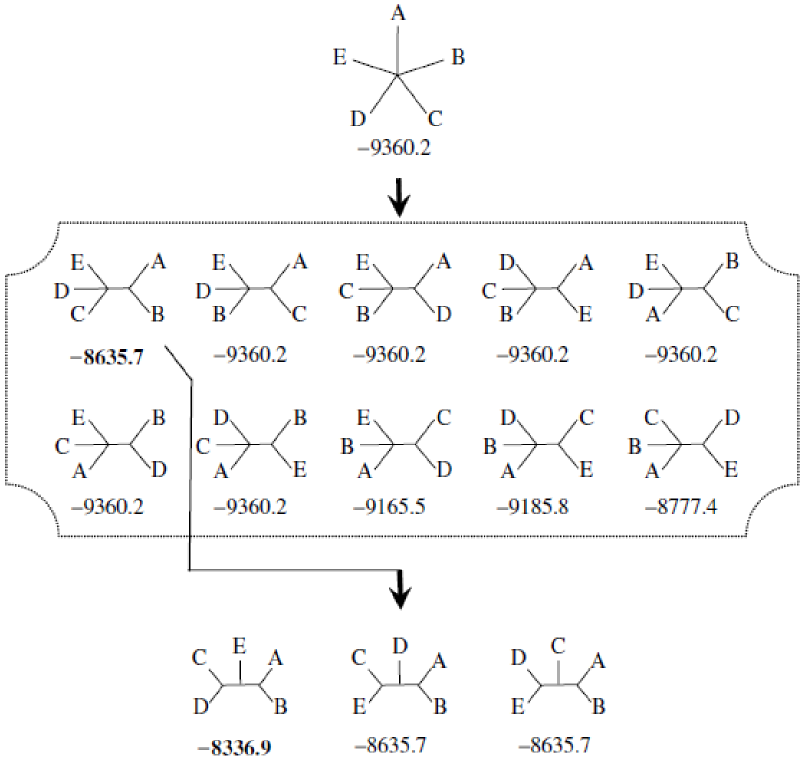


Pseudocode

```
S # is list of Leafs
T <- initialize_tree(S);
WHILE S is not empty DO
      L <- next_leaf(S);
      current_cost <- INF;
      current_Best_T;
      FOR branch in T DO
          T_test <- copy_tree(T)
           insert_Leaf(branch,L)
           cost_t_test <- cost(T_test)
           IF cost_t_test < cost THEN
                   current_Best_T <- T_test;
                   current_cost <- cost_t_test;
            ENDIF
        ENDO
        T <- current_Best_T;
ENDWHILE
```

Finding the “best” tree

- Star decomposition

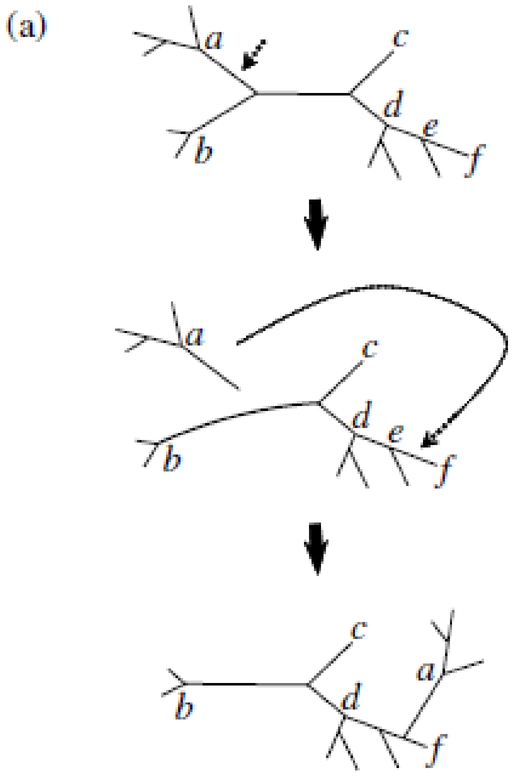


- Which objects do we need for the pseudocode?
- Which functions do we need for the pseudocode?

Finding the “best” tree

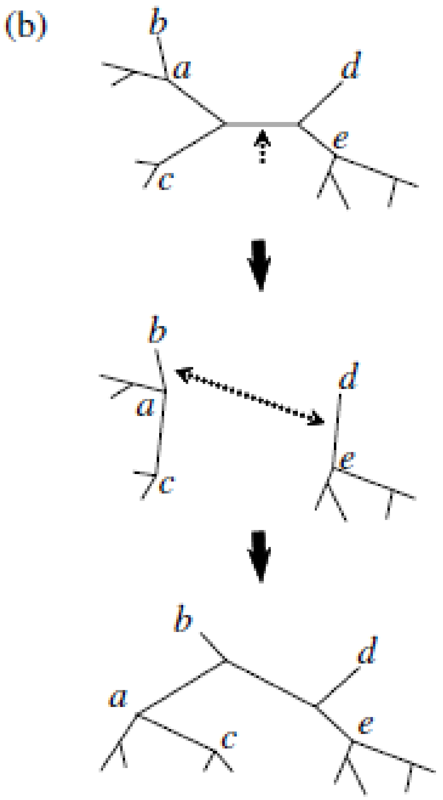
- Pruning and swapping

subtree pruning and regrafting



SPR

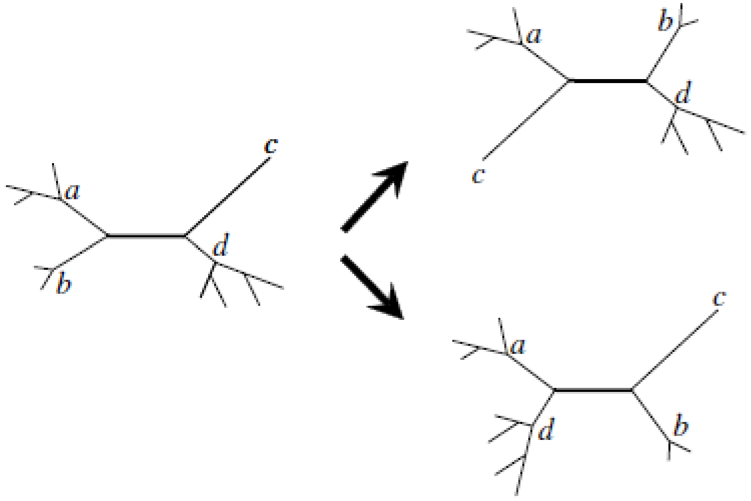
Branch swapping by tree bisection and reconnection

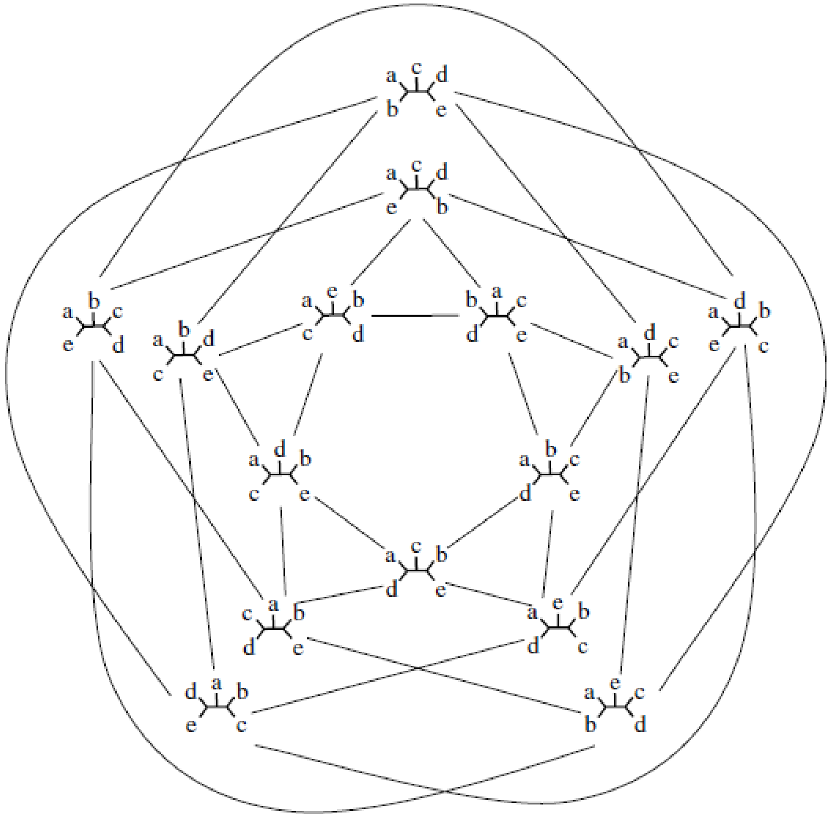


TBR

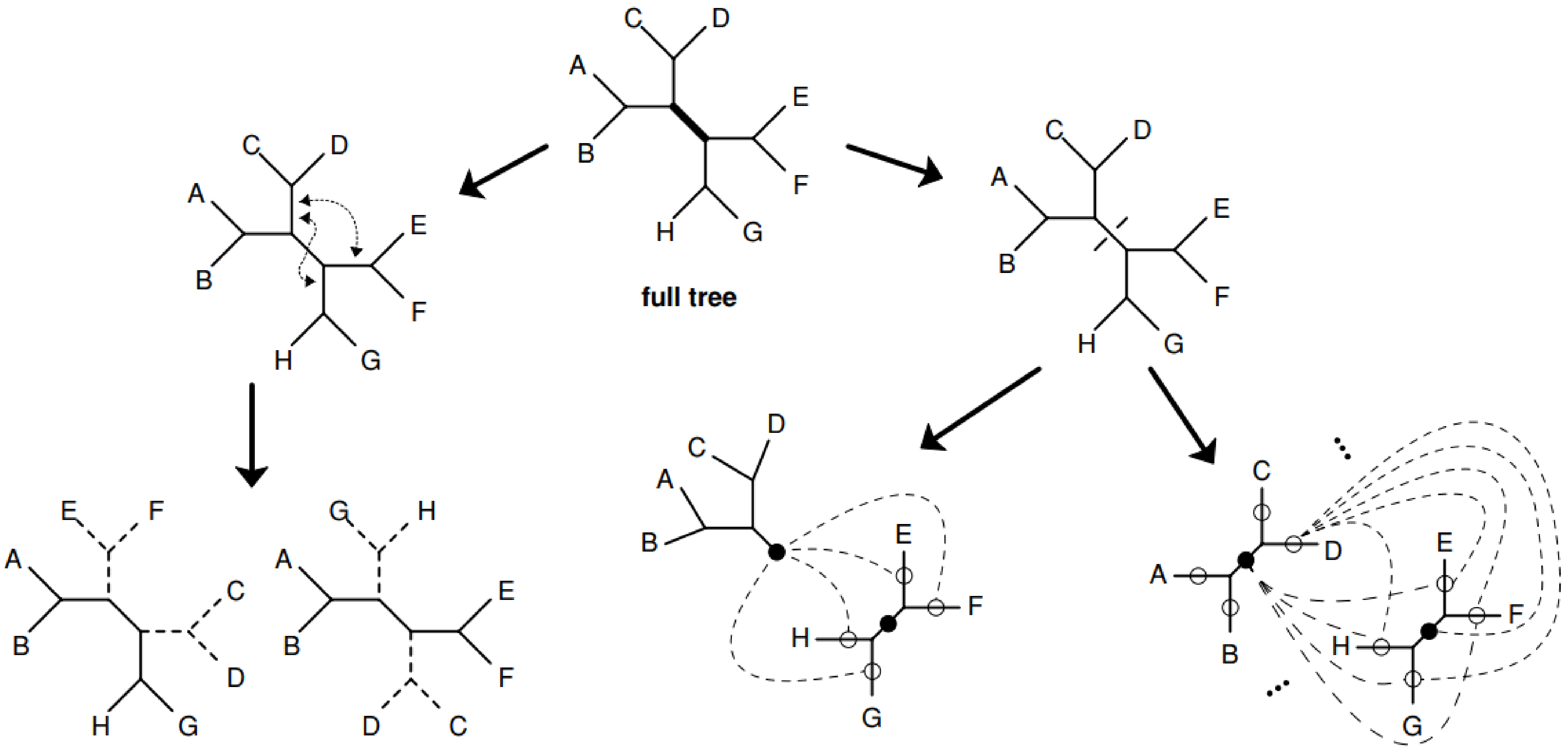
Finding the “best” tree

- Nearest-neighbour interchange (NNI)





Finding the “best” tree



full tree

Nearest Neighbor Interchange

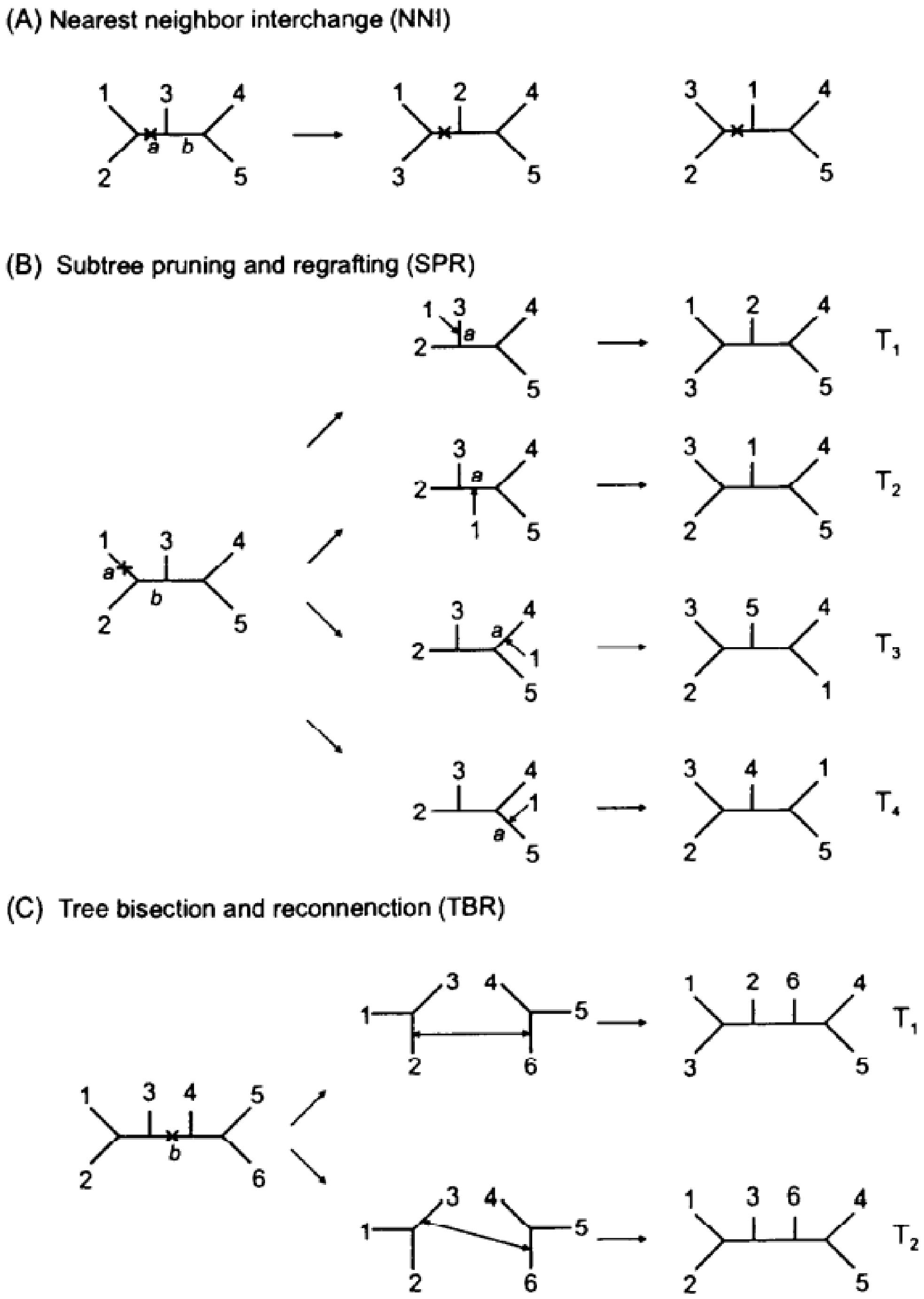
linearly many NNI trees

Subtree Pruning + Regrafting

quadratically many SPR trees

Tree-Bisection + Reconnection

cubic number of TBR trees



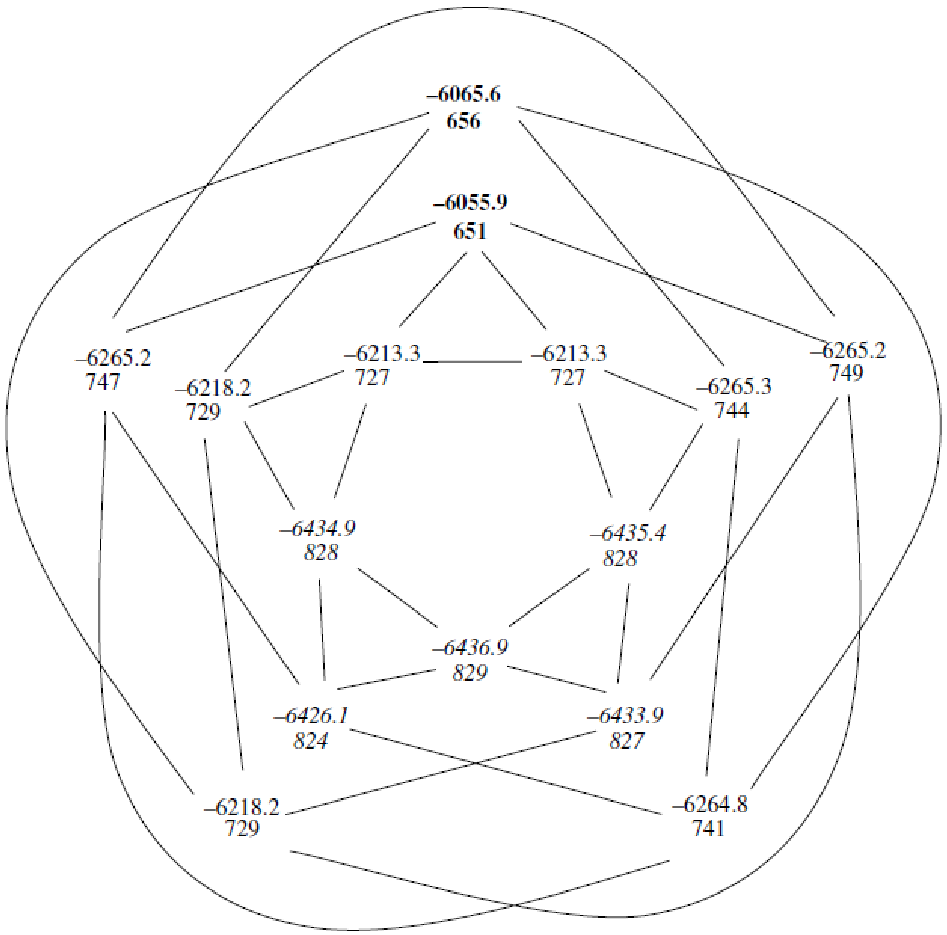
(A) Nearest neighbor interchange (NNI)

(B) Subtree pruning and regrafting (SPR)

(C) Tree bisection and reconnenction (TBR)

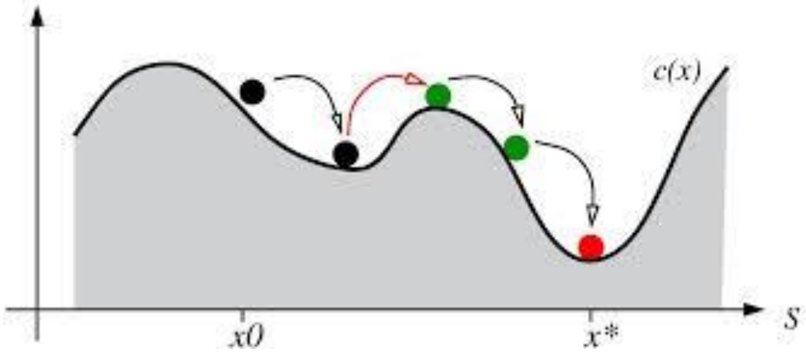
Finding the “best” tree

- Multiple optimal trees



Stochastic search tree

- Simulated annealing/Metropolis algorithm



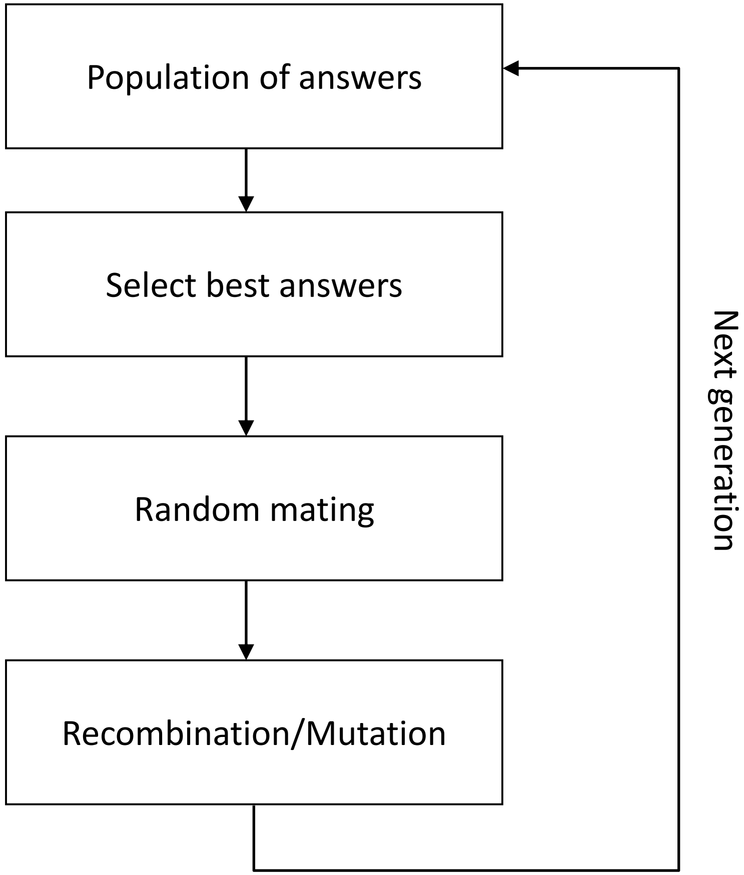
Probability of acceptance of the proposed change

$$p = e^{-\frac{k}{T}\Delta E}$$

```
initialize state E;
V = compute tree statistic of E
FOR G iterations DO
    S = select_neighbour(E);
    L = cost_tree(S);
    compute E = L – V;
    compute probability of acceptance p;
    R = compute random number between (0-1);
    IF R < p THEN
        E = S;
   update T;
END
```

Stochastic search tree

- Genetic algorithm



Stochastic search tree

- Genetic algorithm: crossover

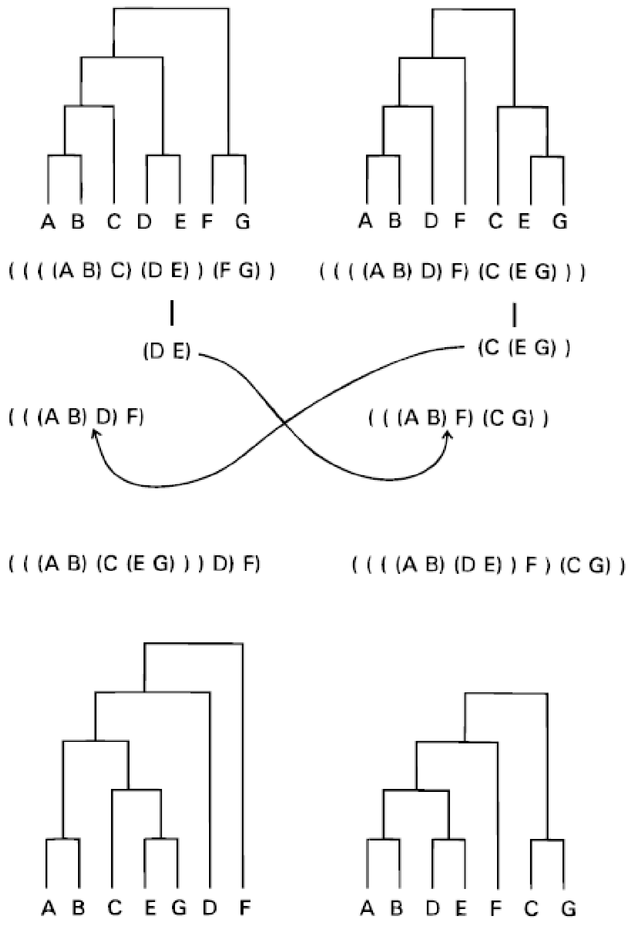
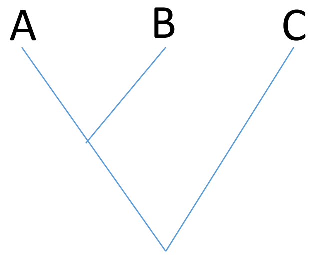

A B C D E F G
A B D F C E G
((((A B) C) (D E)) (F G))
((((A B) D) F) (C (E G)))
(D E)
(C (E G))
(((A B) D) F)
(((A B) F) (C G))
(((A B) (C (E G))) D) F)
((((A B) (D E)) F) (C G))
A B C E G D F
A B D E F C G

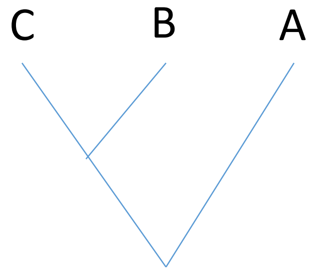
FIG. 1. How the crossover operator works in the evolutionary optimization algorithm. The evolutionary algorithm internally handles tree structures as character strings (top of figure), but for clarity, corresponding trees are also shown. First a crossover fragment (subtree) starting from a randomly picked node (excluding the root) is copied from each parent tree. Then the terminal taxa present in each crossover fragment are pruned from the other parent tree, thus preventing the replication of taxa. Finally the crossover fragments are exchanged between the pruned parent trees by insertion into randomly chosen positions.

How do we assess the robustness of the obtained tree?

Gene dataset 1
Species A
Species B
Species C



Gene dataset 2
Species A
Species B
Species C



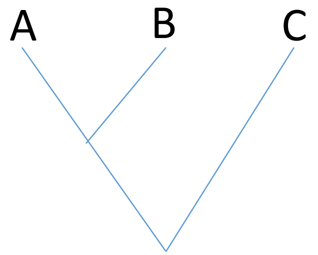
Reasons for discrepancy?

Recent time of speciation
Horizontal gene transfer

OVERFITTING:
We do not generate the tree that summarizes the relationship between species, but the specific tree that summarizes the relationships of the analyzed genes

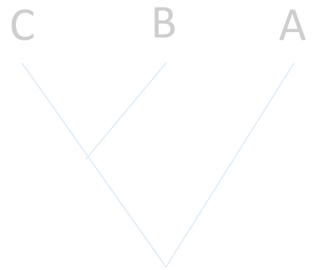
How do we assess the robustness of the obtained tree?

Gene dataset 1
Species A
Species B
Species C



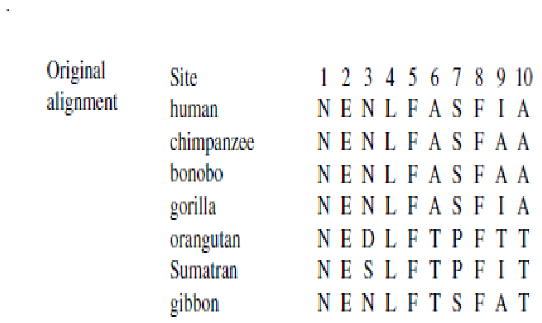
What do we do if we only have ONE dataset???

Gene dataset 2
Species A
Species B
Species C

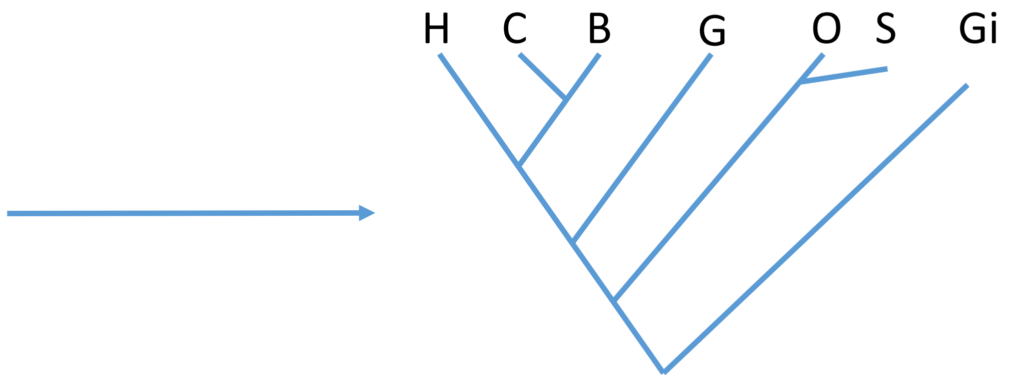


How to assess the robustness of the obtained tree?

Bootstrap

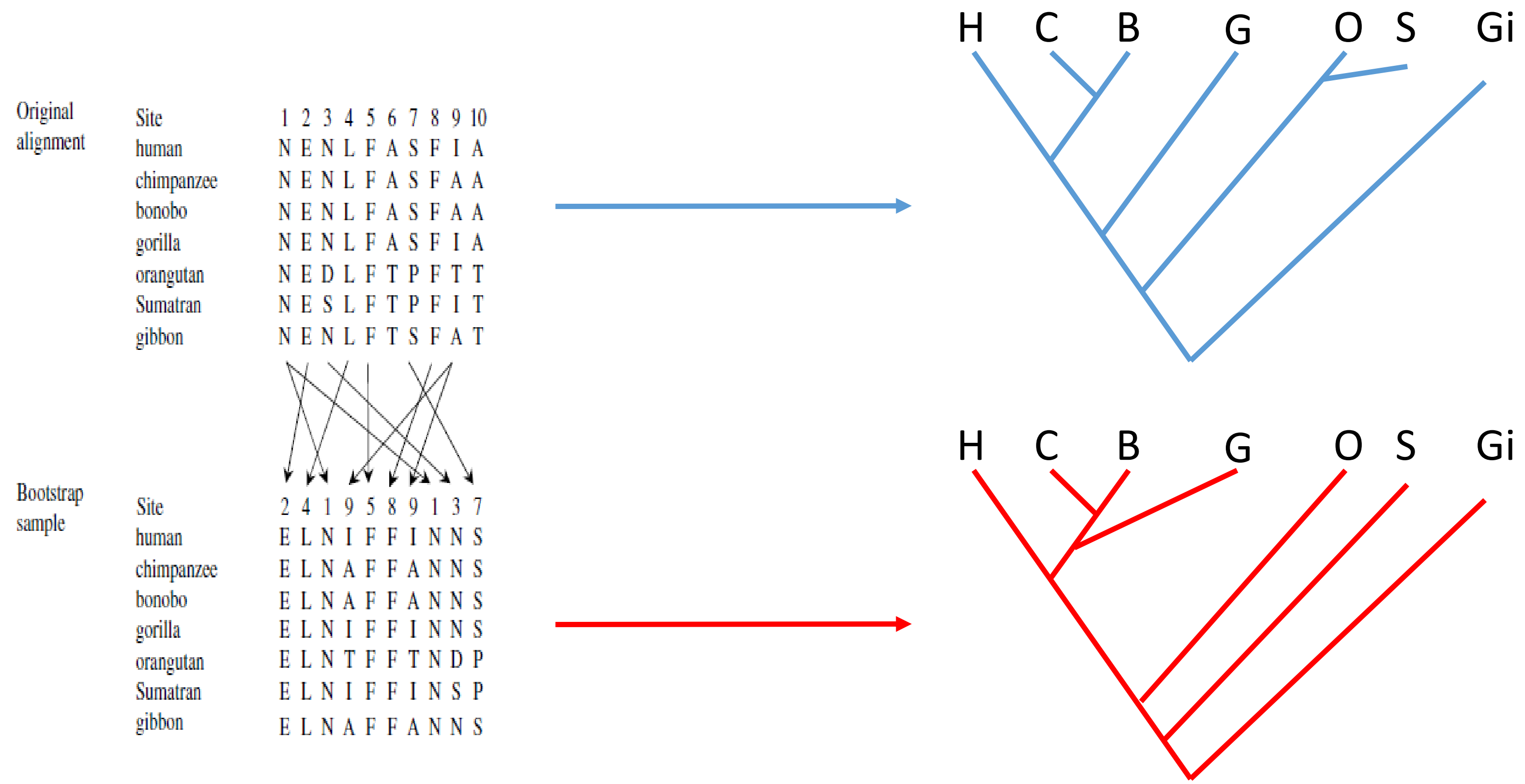


Original alignment	Site	1	2	3	4	5	6	7	8	9	10
	human	N	E	N	L	F	A	S	F	I	A
	chimpanzee	N	E	N	L	F	A	S	F	A	A
	bonobo	N	E	N	L	F	A	S	F	A	A
	gorilla	N	E	N	L	F	A	S	F	I	A
	orangutan	N	E	D	L	F	T	P	F	T	T
	Sumatran	N	E	S	L	F	T	P	F	I	T
	gibbon	N	E	N	L	F	T	S	F	A	T



How to assess the robustness of the obtained tree?

Bootstrap



Original alignment

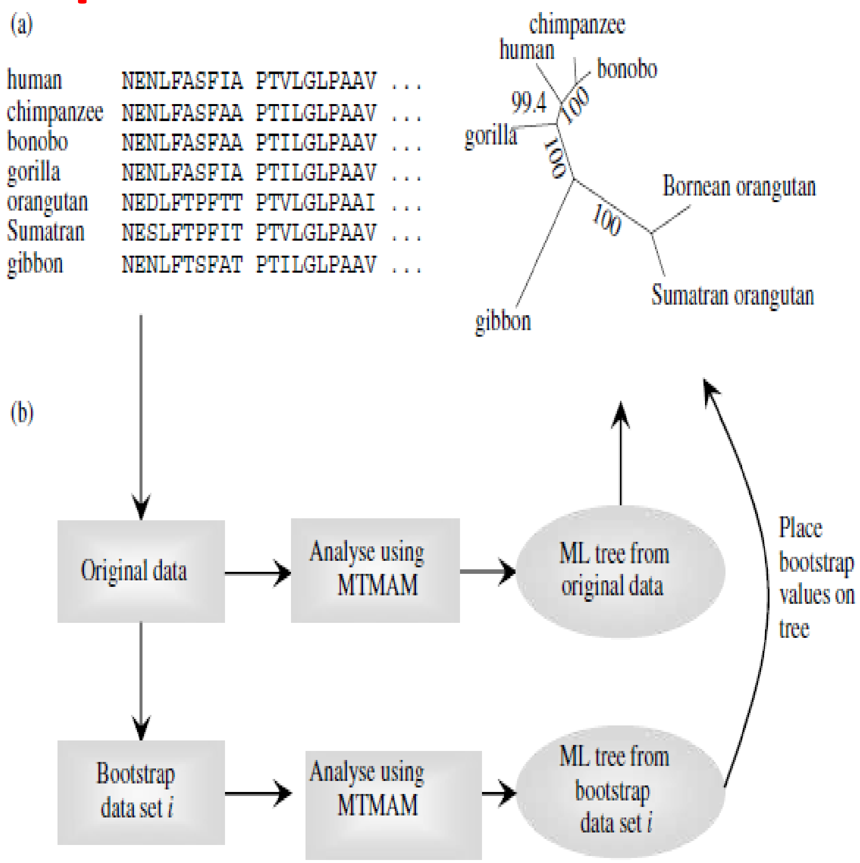
Site	1	2	3	4	5	6	7	8	9	10
human	N	E	N	L	F	A	S	F	I	A
chimpanzee	N	E	N	L	F	A	S	F	A	A
bonobo	N	E	N	L	F	A	S	F	A	A
gorilla	N	E	N	L	F	A	S	F	I	A
orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap sample

Site	2	4	1	9	5	8	9	1	3	7
human	E	L	N	I	F	F	I	N	N	S
chimpanzee	E	L	N	A	F	F	A	N	N	S
bonobo	E	L	N	A	F	F	A	N	N	S
gorilla	E	L	N	I	F	F	I	N	N	S
orangutan	E	L	N	T	F	F	T	N	D	P
Sumatran	E	L	N	I	F	F	I	N	S	P
gibbon	E	L	N	A	F	F	A	N	N	S

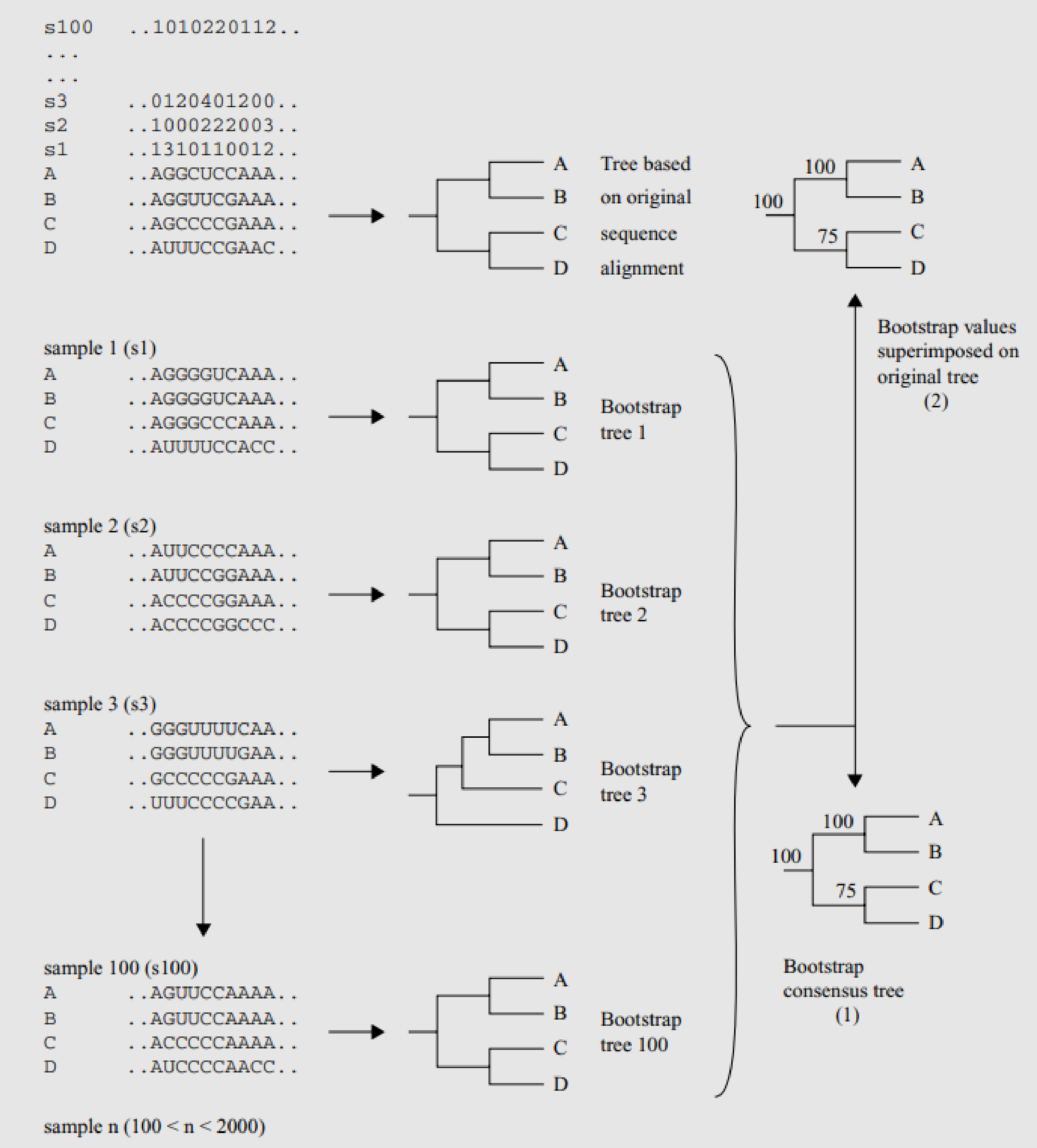
How to assess the robustness of the obtained tree?

Bootstrap



```
human       NENLFASFIA PTVLGLPAAV ...
chimpanzee  NENLFASFAA PTILGLPAAV ...
bonobo      NENLFASFAA PTILGLPAAV ...
gorilla     NENLFASFIA PTILGLPAAV ...
orangutan   NEDLFTPFTT PTVLGLPAAI ...
Sumatran    NESLFTPFIT PTVLGLPAAV ...
gibbon      NENLFTSFAT PTILGLPAAV ...
```

How to assess the robustness of the obtained tree?



Two ways to interpret bootstrap

“How many times in the bootstrap trees we see the same clusters?”

“Average bootstrapped trees”

How to assess the robustness of the obtained tree?

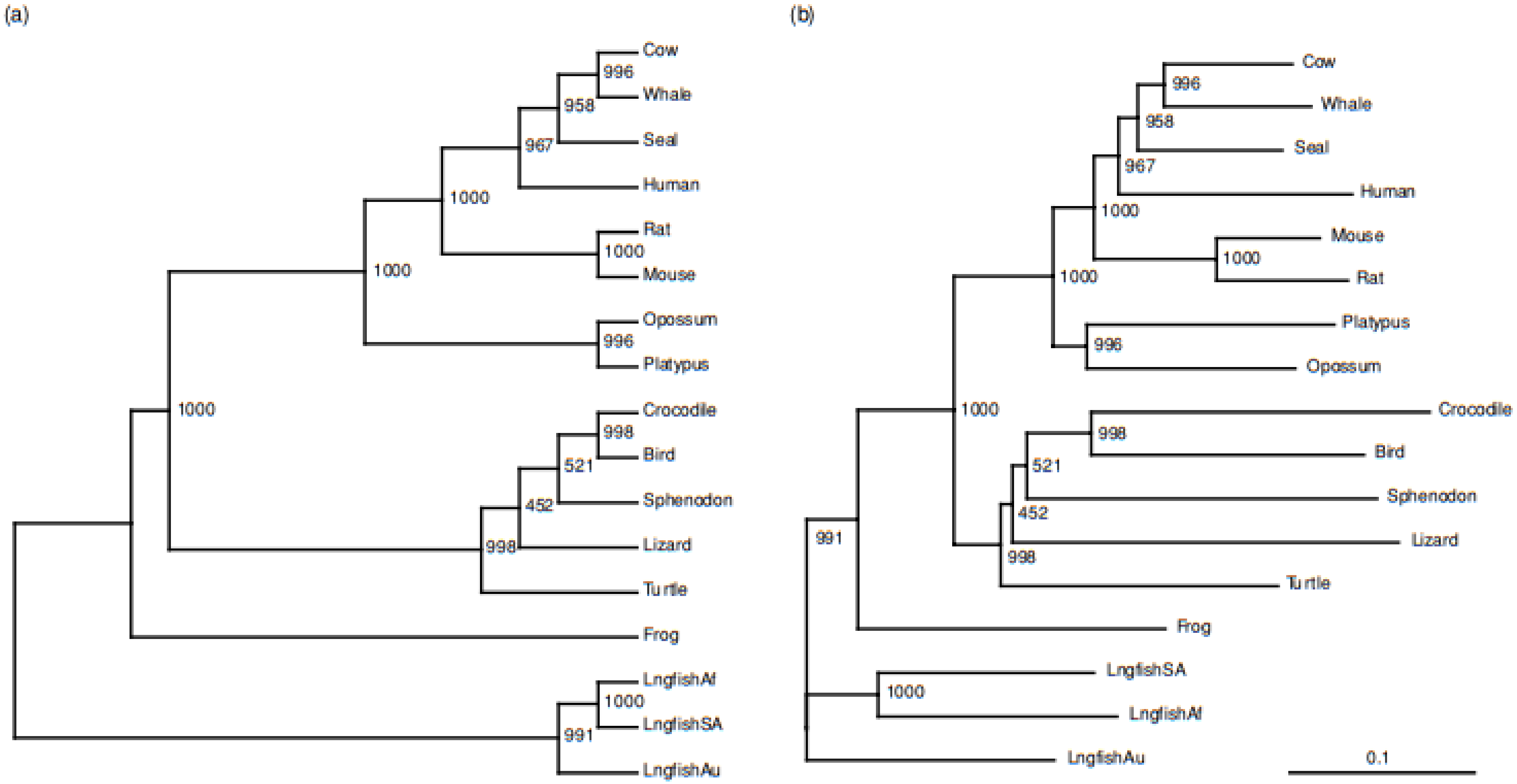


Fig. 5.10 (a) Neighbor-joining consensus tree for 1000 bootstrap replicates of the mtDNA data set as displayed in TREEVIEW. (b) Inferred neighbor-joining tree for the mtDNA data set with bootstrap values. In both cases, the bootstrap values are shown to the right of the node representing the most recent common ancestor of the clade they support.

Which are the main topics to remember?

-
-
-