

# From alignment to phylogenetic tree

Jia-Ming Chang

Dep. of Com. Sci., National Chengchi University

<http://www.changlabtw.com>

1996 ~ 2000 Bachelor (推薦甄試入學)  
2002 ~ 2002 Master  
@ Computer Science, National Tsing Hua Uni.



Dr. Chuan Yi Tang



Dr. Ting-Yi Sung



Dr. Wen-Lian Hsu

2002 ~ 2008 military replace service  
@ Institute of Information Sciences  
Academia Sinica



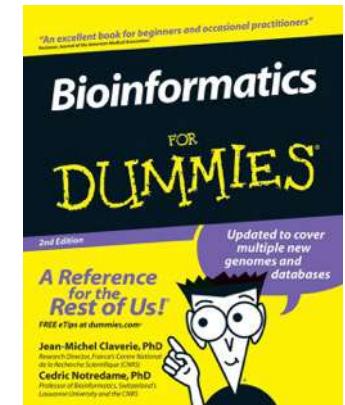
Dr. Giacomo Cavalli

2008~2013 PhD La Caxia fellowship  
@ The Centre for Genomic  
Regulation  
Barcelona, Spain



Dr. Cedric Notredame

2013~2014 Postdoc  
@ Institute of Human Genetics  
Montpellier, France





<https://goo.gl/photos/AT6QkCgfVH8JsmH69>

<https://goo.gl/photos/dBiRxWYxbWSsbCHA8>



Founded in  
**2000**

**30+6**  
groups  
core facilities

**437**  
employees  
(377 scientists  
+ 60 support staff)

**69%**  
foreign researchers

**203**  
peer-reviewed  
publications,  
average  
**IF = 9.01**

Budget:  
**35.55 M€**  
(41.6% core-national  
and regional government;  
58.4% external)

**Position 9 worldwide**  
according to Scimago Institutions Rankings 2014  
(Health sector, Q1 indicator)



# Evolution

- Charles Darwin's 1859 book (*On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*) introduced the theory of evolution.
- To Darwin, the struggle for existence induces a natural selection. Offspring are dissimilar from their parents (that is, variability exists), and individuals that are more fit for a given environment are selected for. In this way, over long periods of time, species evolve. Groups of organisms change over time so that descendants differ structurally and functionally from their ancestors.

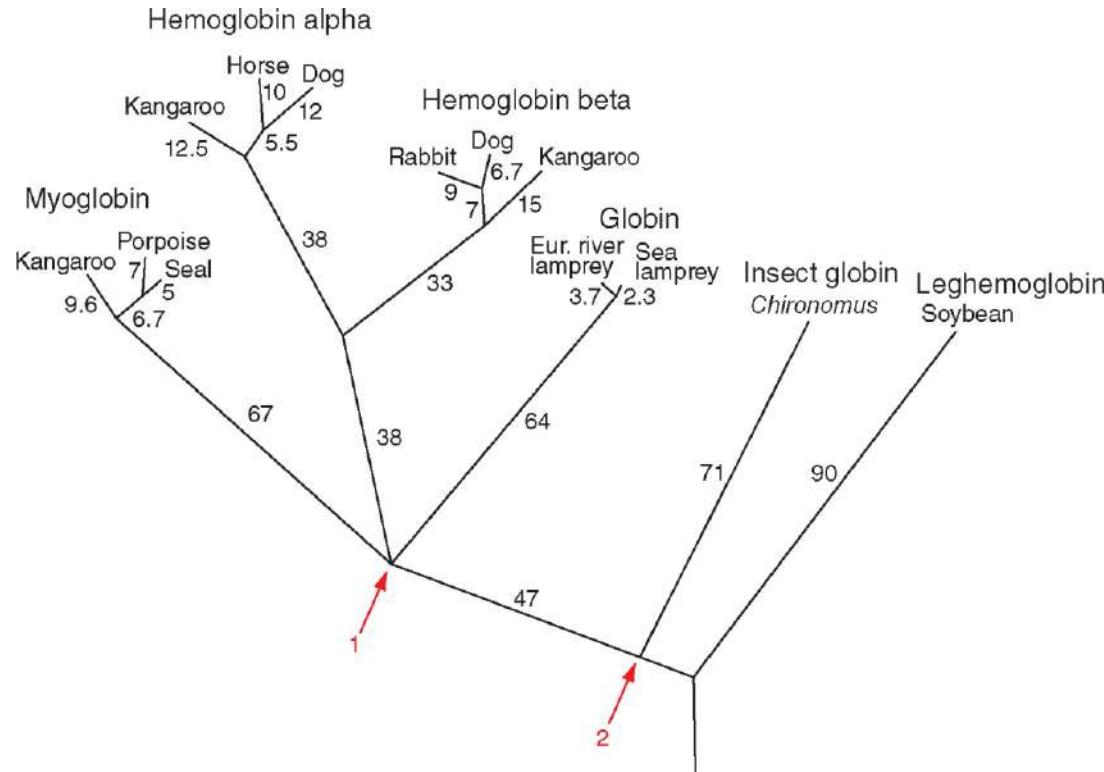
# Evolution

- At the molecular level, evolution is a process of mutation with selection.
- Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life.
- Phylogeny is the inference of evolutionary relationships.
- Traditionally, phylogeny relied on the comparison of morphological features between organisms. Today, molecular sequence data are also used for phylogenetic analyses.

# Goals of molecular phylogeny

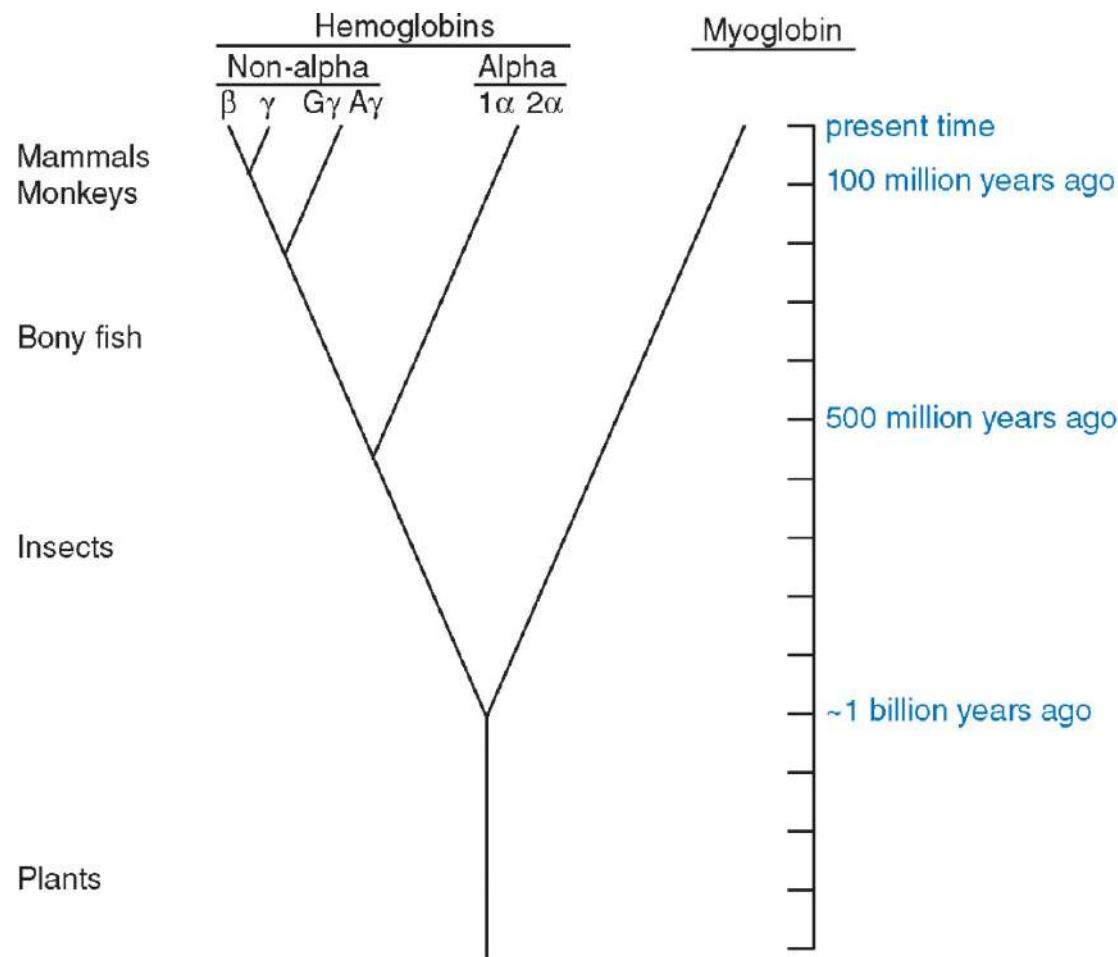
- Phylogeny can answer questions such as:
  - Is my favorite gene under selective pressure?
  - Was the extinct quagga more like a zebra or a horse?
  - Was Darwin correct that humans are closest to chimps and gorillas?
  - How related are whales, dolphins & porpoises to cows?
  - Where and when did HIV originate?
  - What is the history of life on earth?

# 1960s: globin phylogeny (tree of 13 orthologs by Margaret Dayhoff and colleagues)



**Arrow 1:** node corresponding to last common ancestor of a group of vertebrate globins.  
**Arrow 2:** ancestor of insect and vertebrate globins

# 1960s: globin phylogeny (tree of 7 paralogs)



Dayhoff et al. (1972) analyzed related globins in the context of evolutionary time.

# Molecular clock hypothesis

- In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes c, and fibrinopeptides. Some proteins appeared to evolve slowly, while others evolved rapidly.
- Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:
  - For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages

# Positive and negative selection

- Darwin's theory of evolution suggests that, at the phenotypic level, traits in a population that enhance survival are selected for, while traits that reduce fitness are selected against. For example, among a group of giraffes millions of years in the past, those giraffes that had longer necks were able to reach higher foliage and were more reproductively successful than their shorter-necked group members, that is, the taller giraffes were selected for.

# Positive and negative selection

- In the mid-20<sup>th</sup> century, a conventional view was that molecular sequences are routinely subject to positive (or negative) selection.
- Positive selection occurs when a sequence undergoes significantly increased rates of substitution, while negative selection occurs when a sequence undergoes change slowly. Otherwise, selection is neutral.

# Consider using DNA, RNA, or protein for phylogeny

- Four globins are aligned.
  - The DNA contains informative differences in the 5' (and 3') untranslated regions.
  - There are protein changes (top, green arrowheads).
  - There are more DNA changes: note 6 positions having synonymous changes (nucleotides shaded blue) and six positions with nonsynonymous changes (red nucleotides).

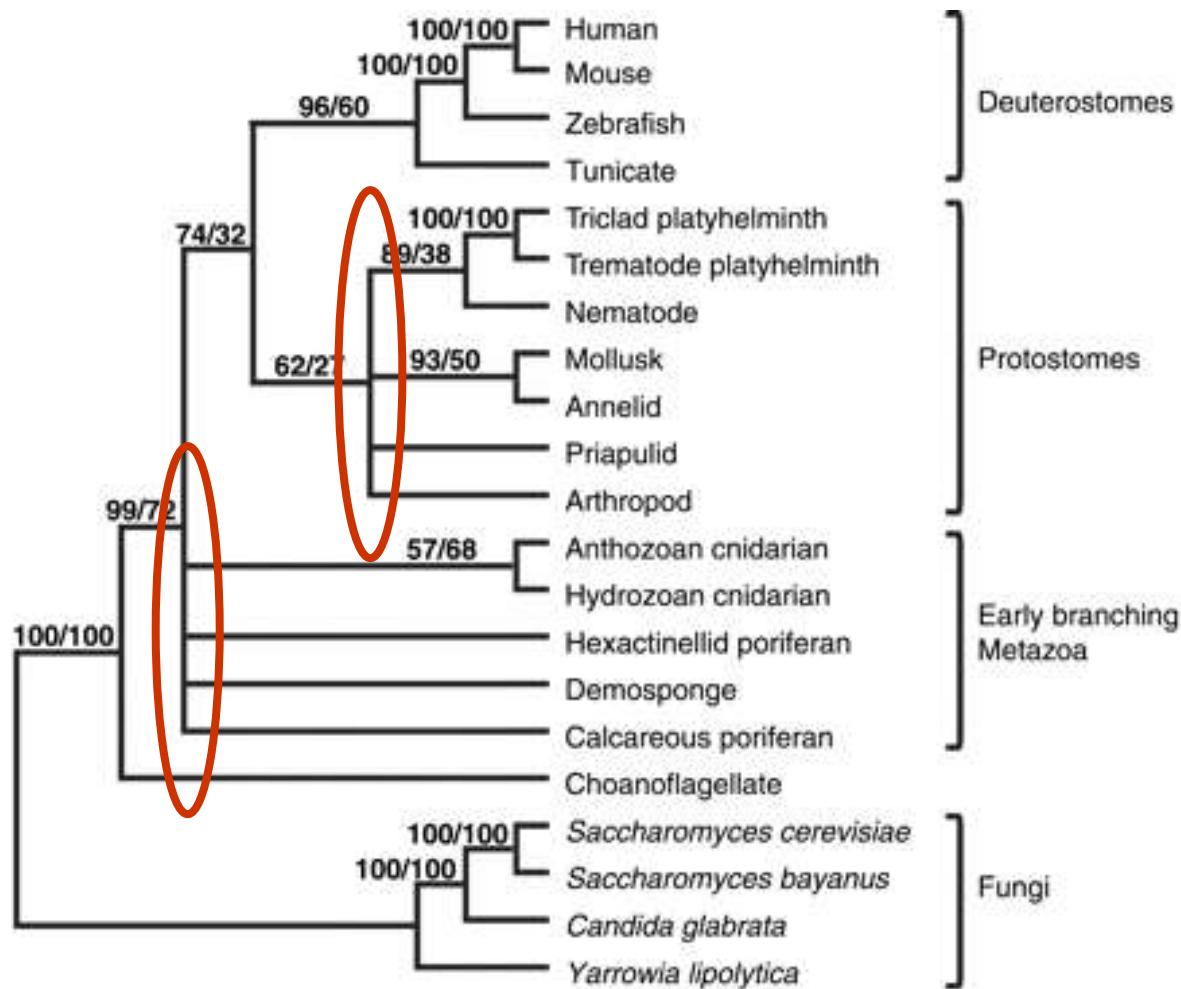
human	M	V	H	L	T	P	E	E	K	S	A	V
chimpanzee	M	V	H	L	T	P	E	E	K	S	A	V
mouse	M	V	H	L	T	D	A	E	K	S	A	V
dog	M	V	H	L	T	A	E	E	K	S	L	V

human	5'	AACAGACACC	ATG	GTG	CAT	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	3'
chimpanzee	5'	AACAGACACC	ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	3'
mouse	5'	AACAGACATC	ATG	GTG	CAC	CTG	ACT	GAT	GCT	GAG	AAG	TCT	GCT	GTC	3'
dog	5'	AACAGACACC	ATG	GTG	CAT	CTG	ACT	GCT	GAA	GAG	AAG	AGT	CTT	GTC	3'
codon															

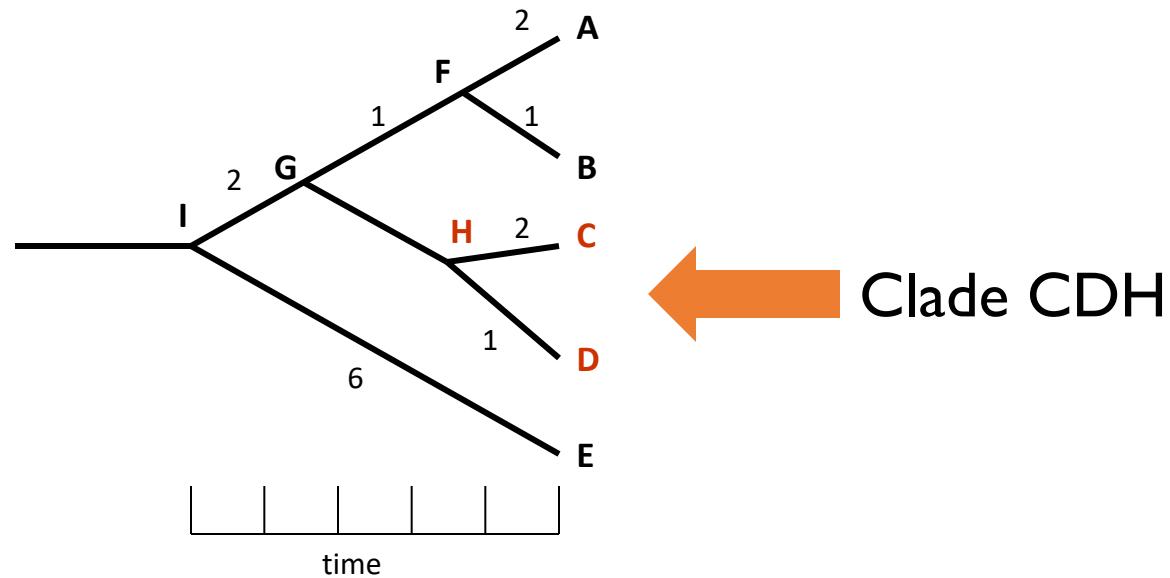


1 2 3 4 5 6 7 8 9 10 11 12

# Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes

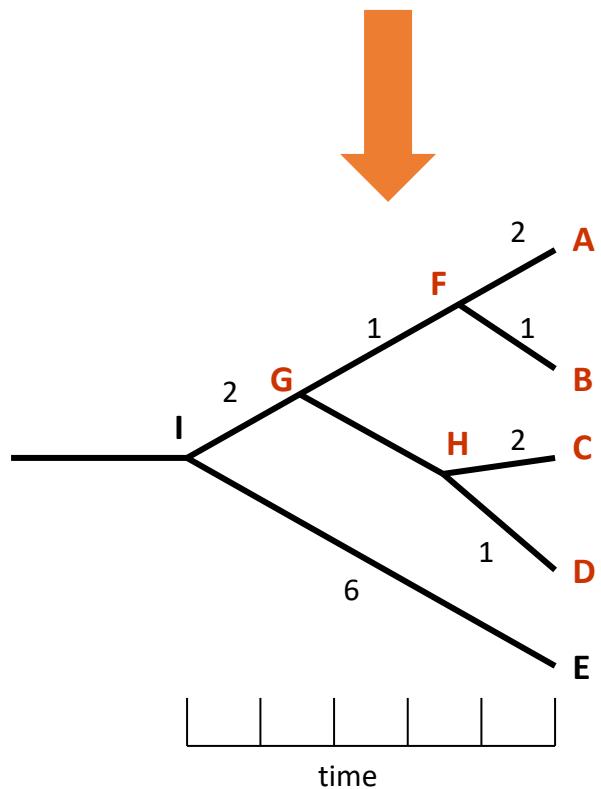


# Tree nomenclature: clades

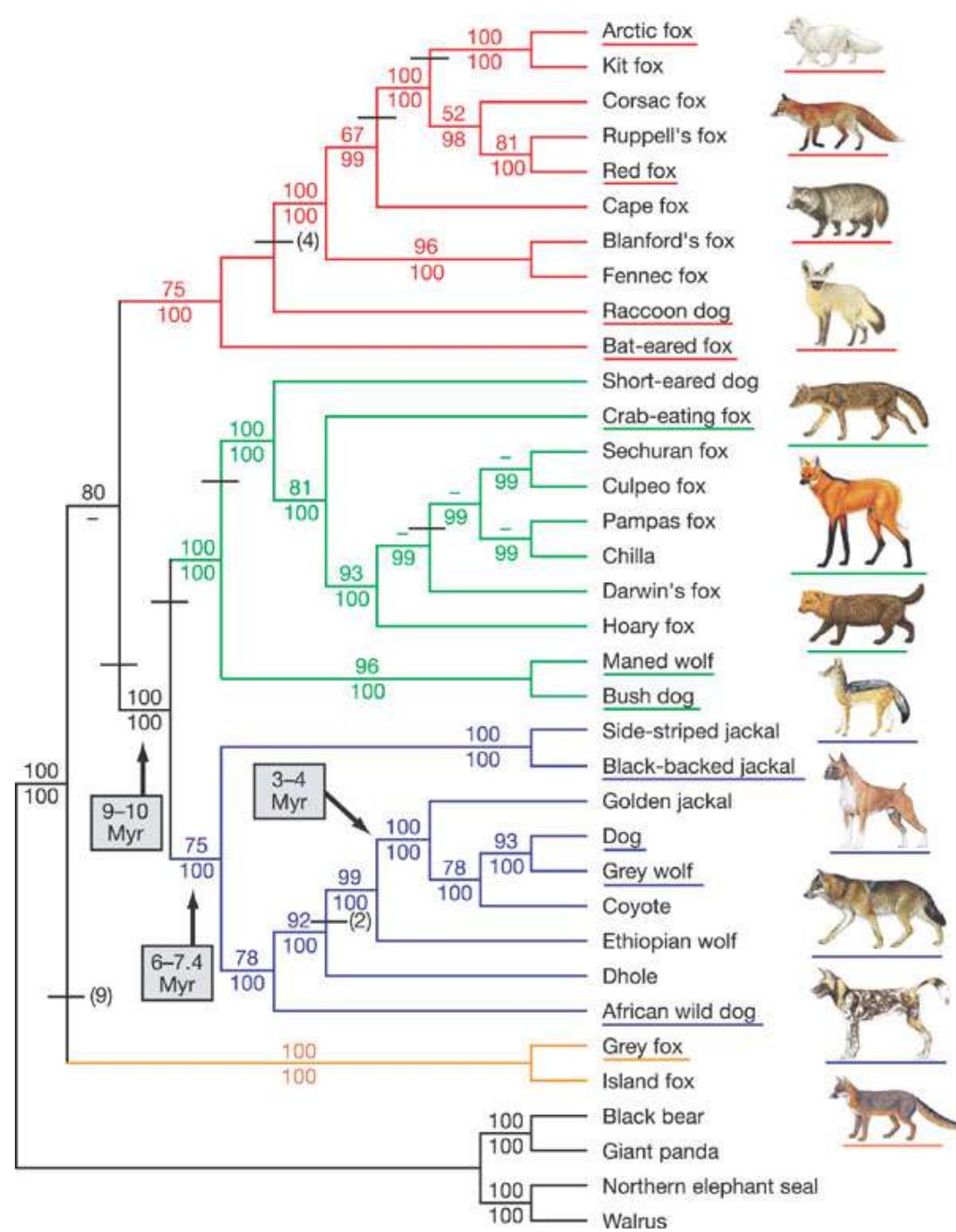


# Tree nomenclature: clades

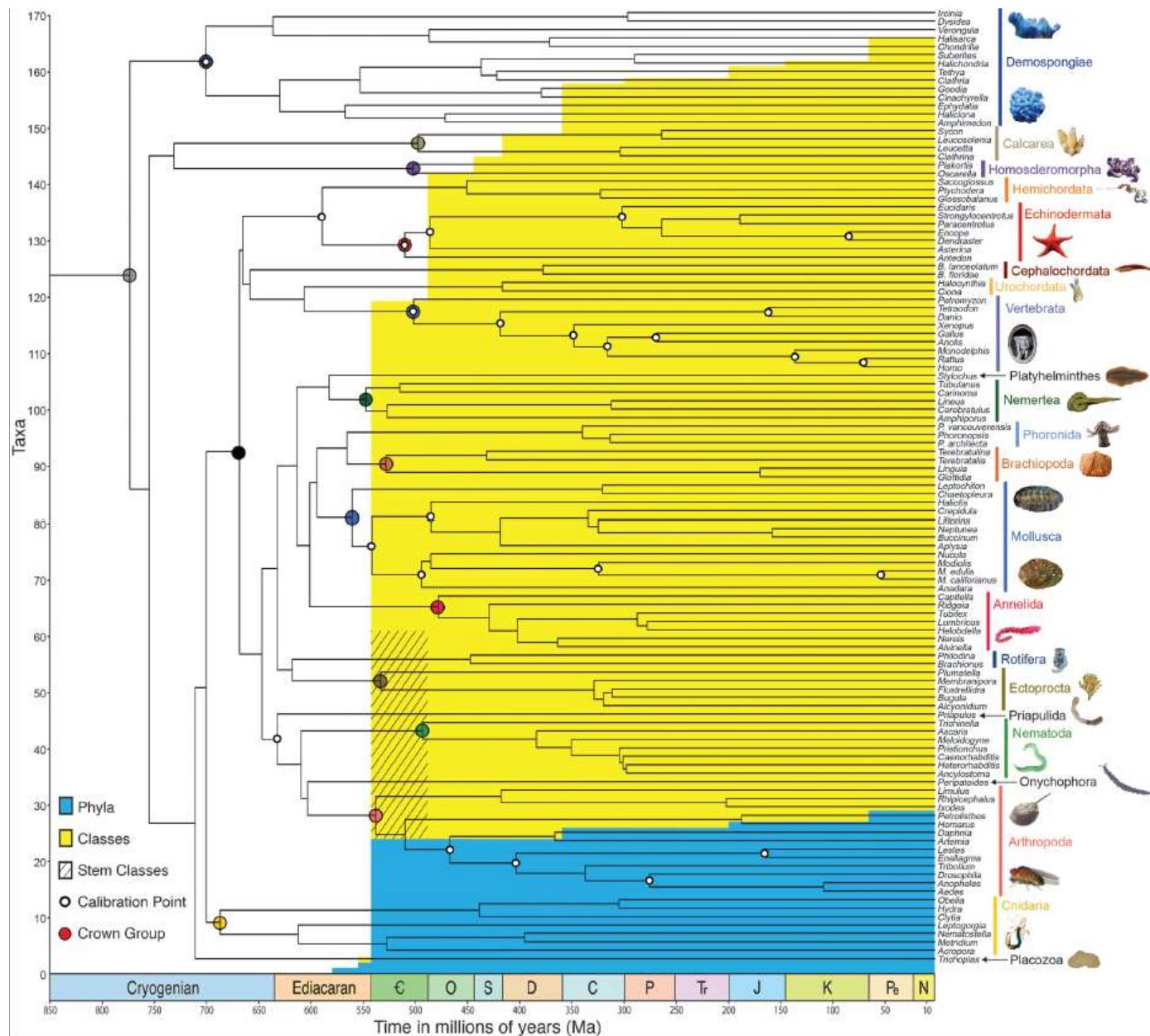
Clade ABF/CDH/G



## Examples of clades



# Diversification of animals (Erwin DH Science 25 Nov. 2011 p.1091, PMID 22116879)

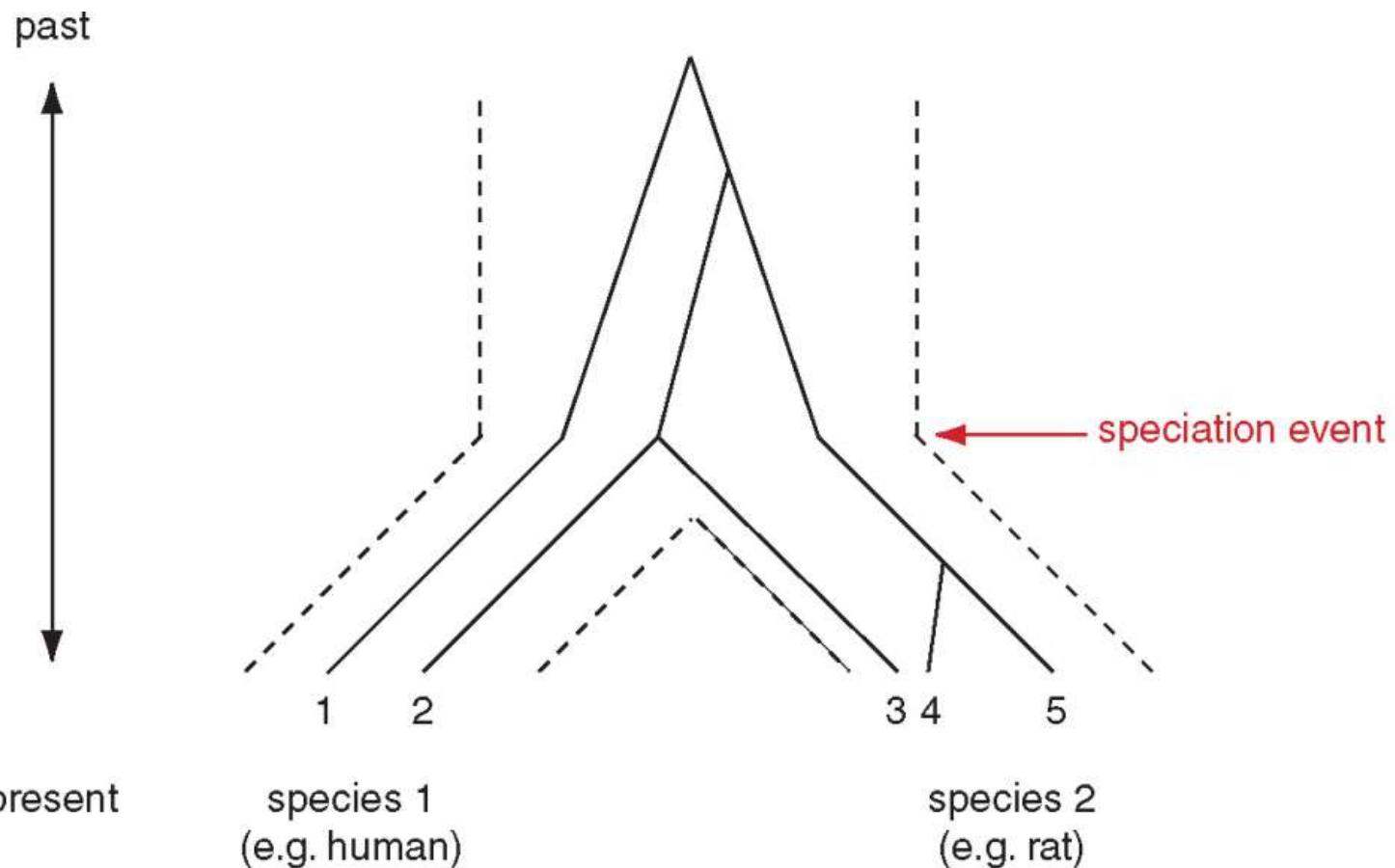


Type of trees (species trees  
vs. gene/protein trees; DNA  
or protein)

# Species trees versus gene/protein trees

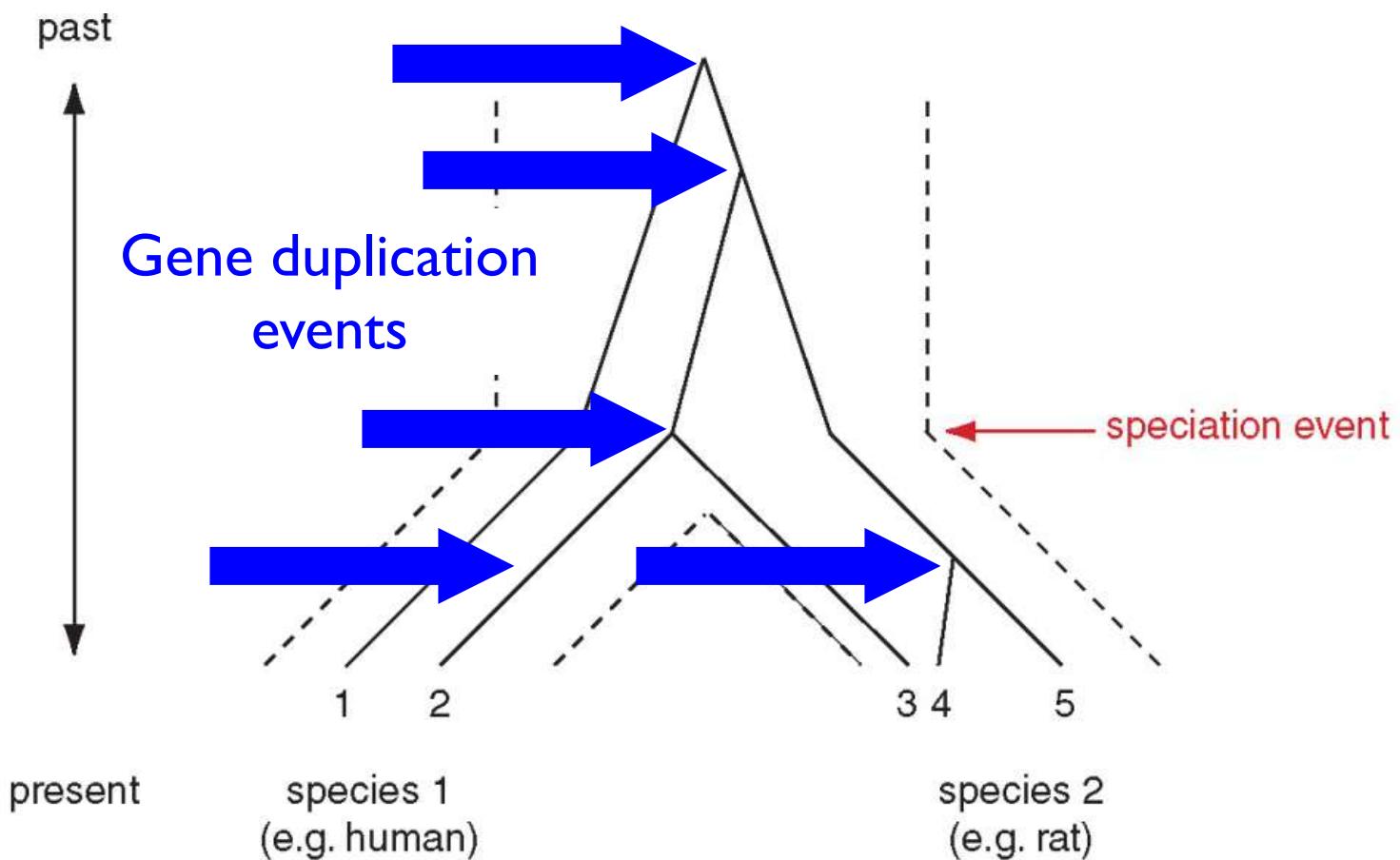
- Molecular evolutionary studies can be complicated by the fact that both species and genes evolve. Speciation usually occurs when a species becomes reproductively isolated. In a species tree, each internal node represents a speciation event.
- Genes (and proteins) may duplicate or otherwise evolve before or after any given speciation event. The topology of a gene (or protein) based tree may differ from the topology of a species tree.

# Species trees versus gene/protein trees



# Species trees versus gene/protein trees

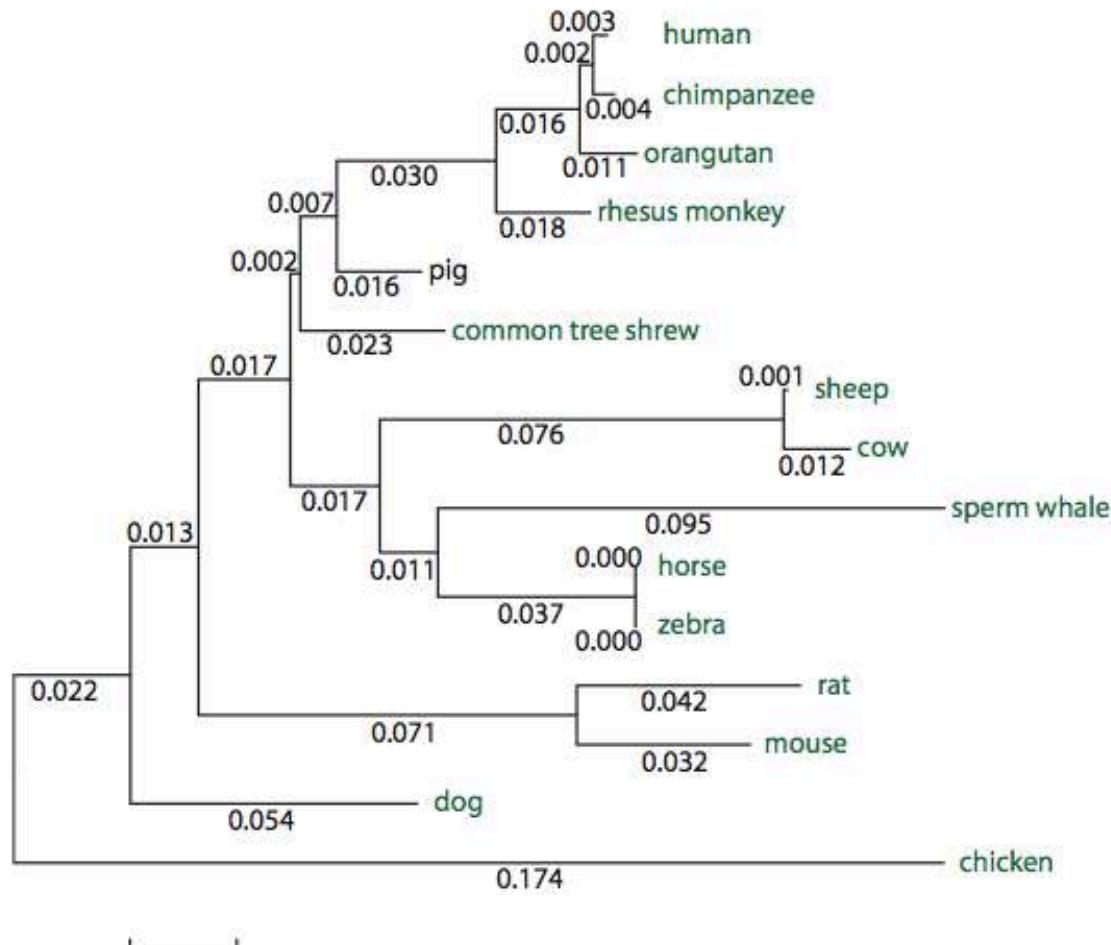
- A gene (e.g. a globin) may duplicate before or after two species diverge!



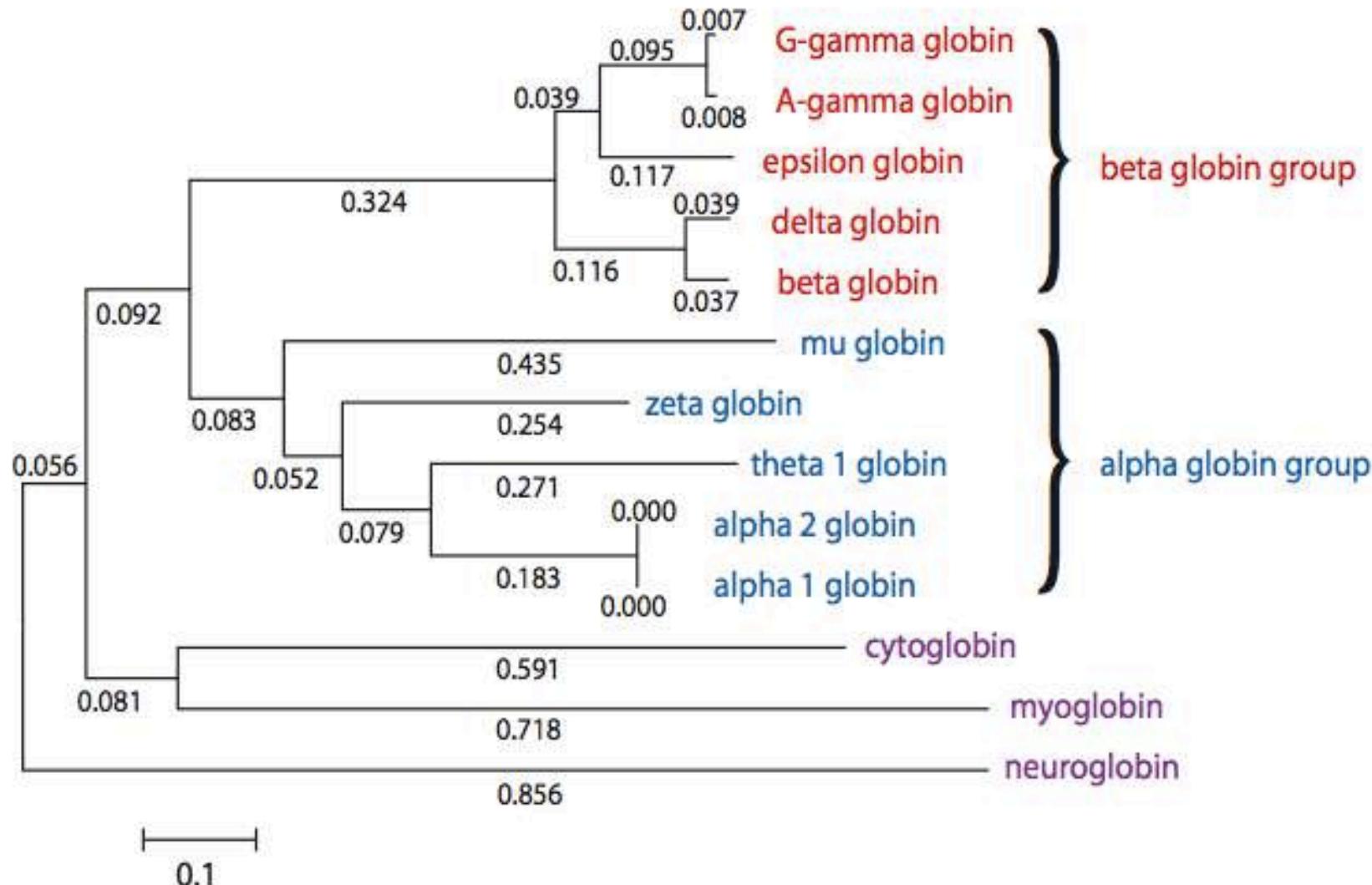
# Definitions: two types of homology

- Orthologs
  - Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.
- Paralogs
  - Homologous sequences within a single species that arose by gene duplication.

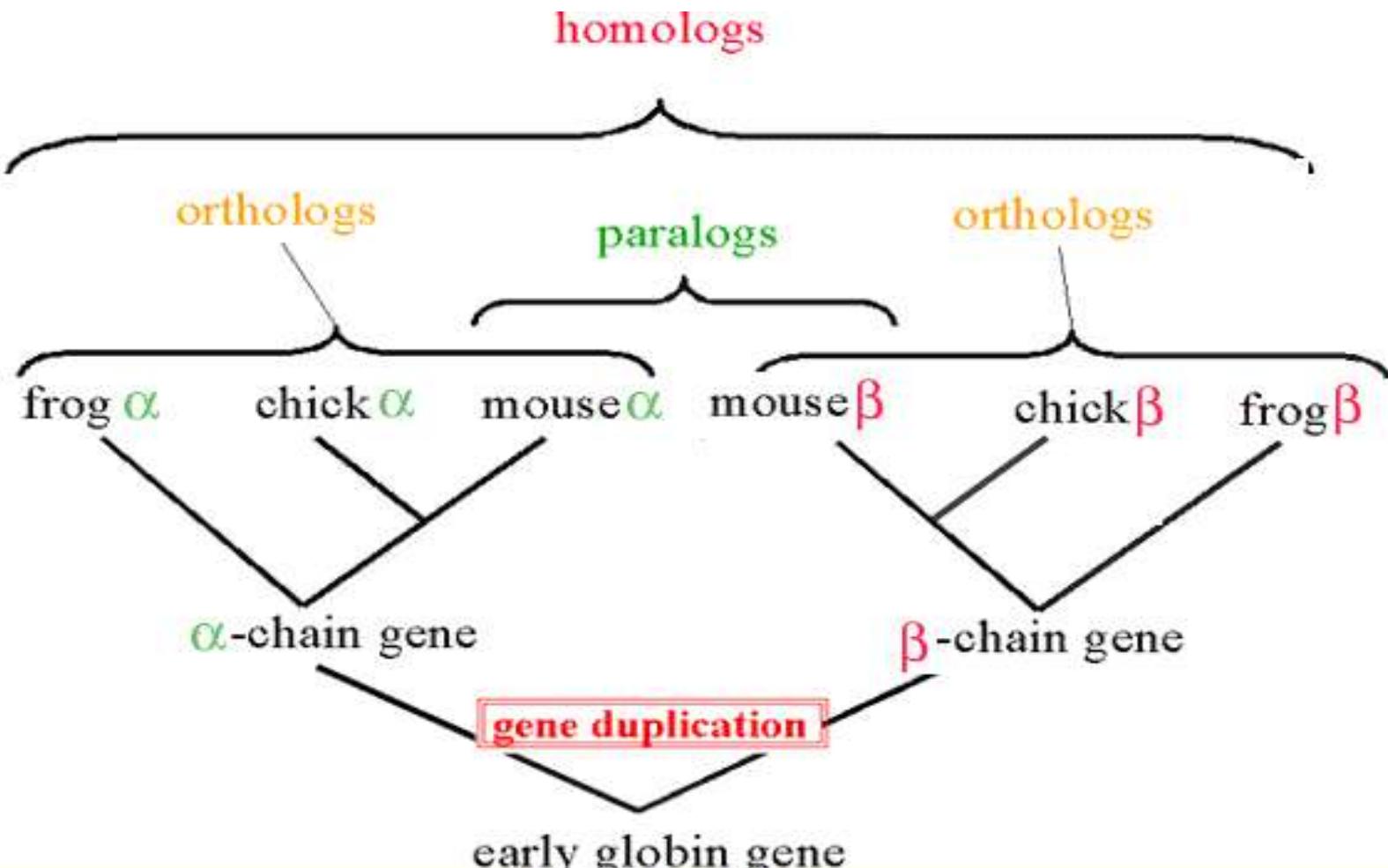
# Myoglobin proteins: examples of orthologs



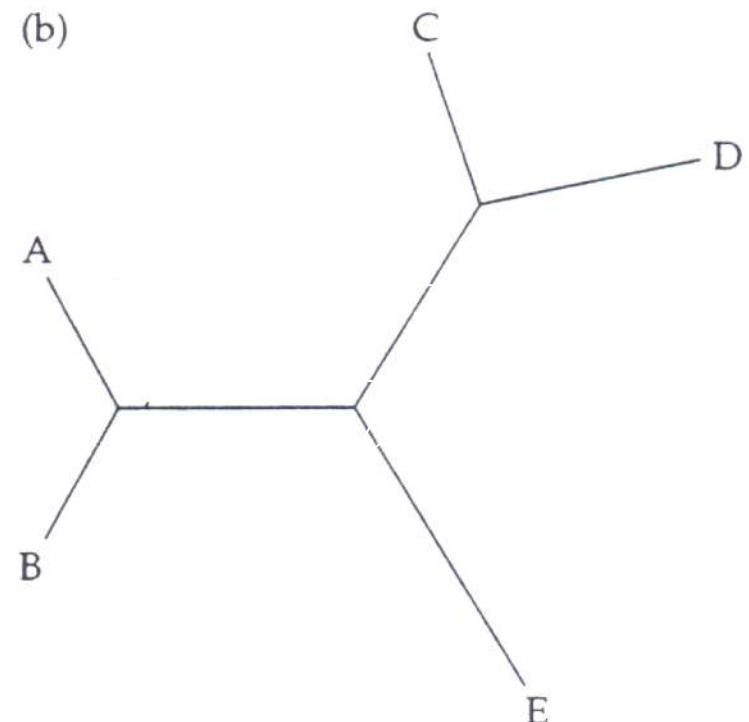
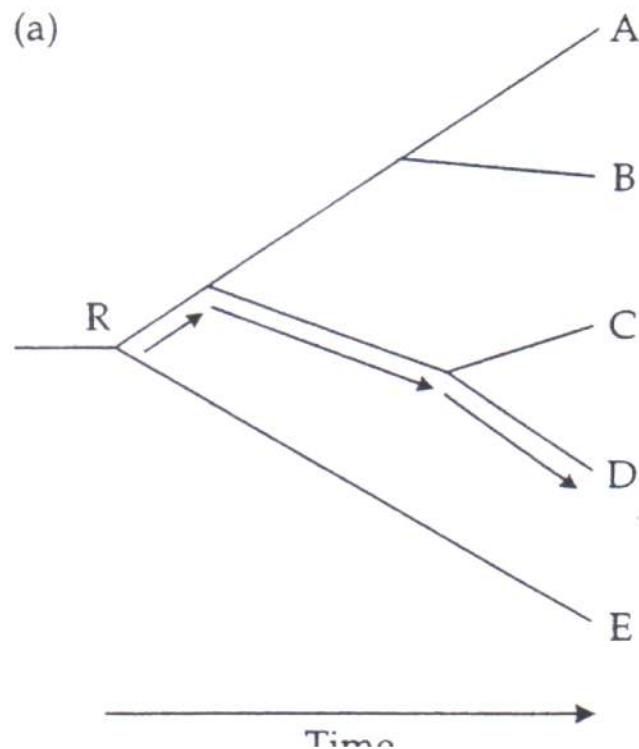
Paralogs: members of a gene (protein) family within a species. This tree shows human globin paralogs.



Orthologs and paralogs are often viewed in a single tree



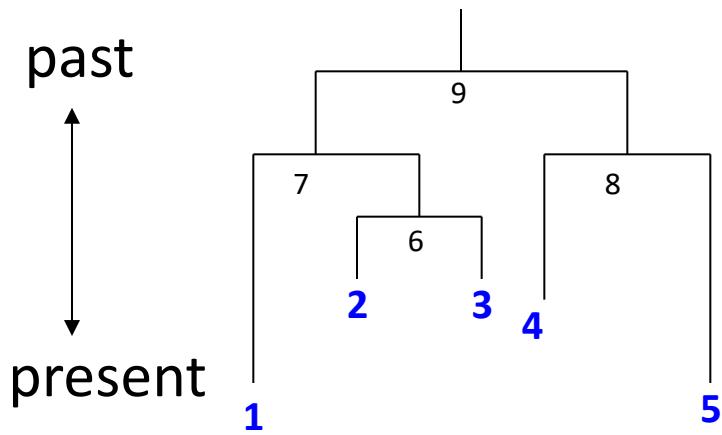
# Outgroup: To root an unrooted tree



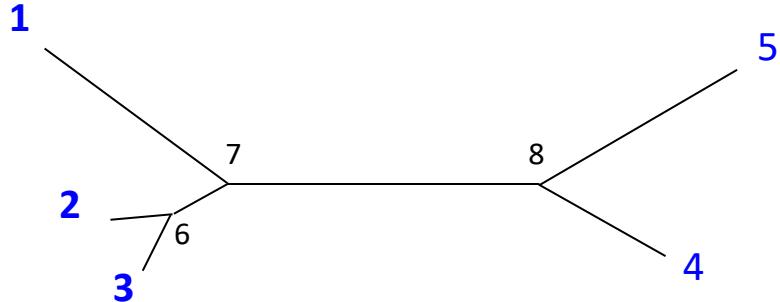
# Tree roots

- The root of a phylogenetic tree represents the common ancestor of the sequences. Some trees are unrooted, and thus do not specify the common ancestor.
- A tree can be rooted using an outgroup (that is, a taxon known to be distantly related from all other OTUs).

# Tree nomenclature: roots

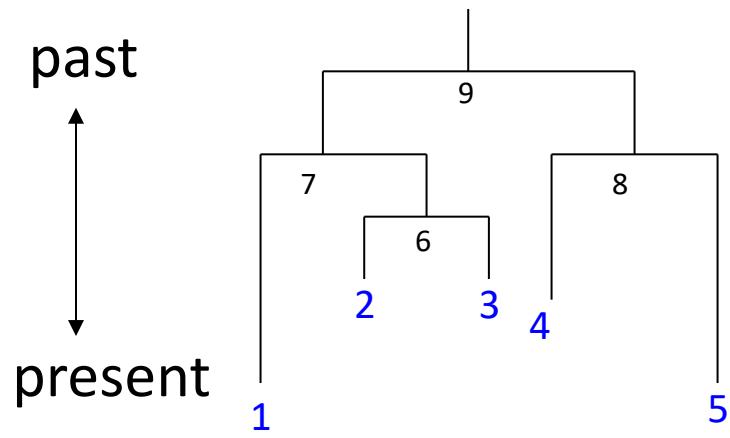


Rooted tree  
(specifies evolutionary path)

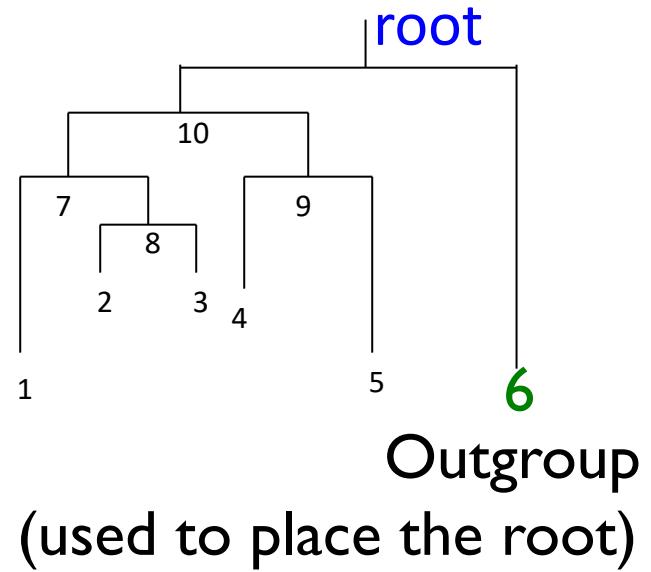


Unrooted tree

# Tree nomenclature: outgroup rooting



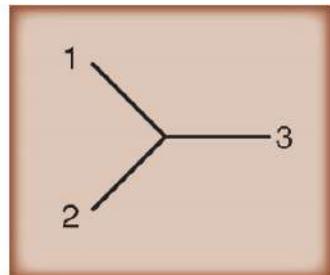
Rooted tree



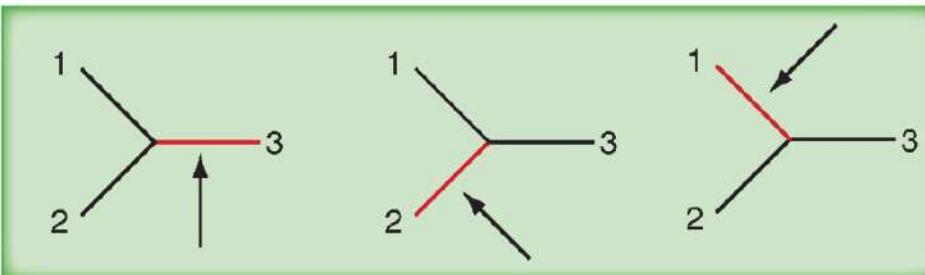
Outgroup  
(used to place the root)

Enumerating trees and  
selecting search strategies

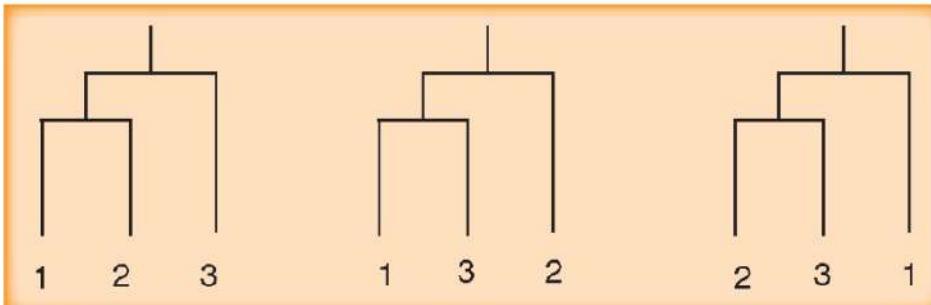
# Numbers of rooted and unrooted trees: 3 OTUs



For three operational taxonomic units (OTUs) there is one possible unrooted tree.

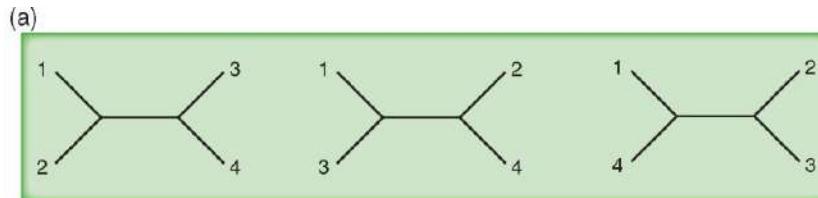


Any of the three edges can be selected to form a root.

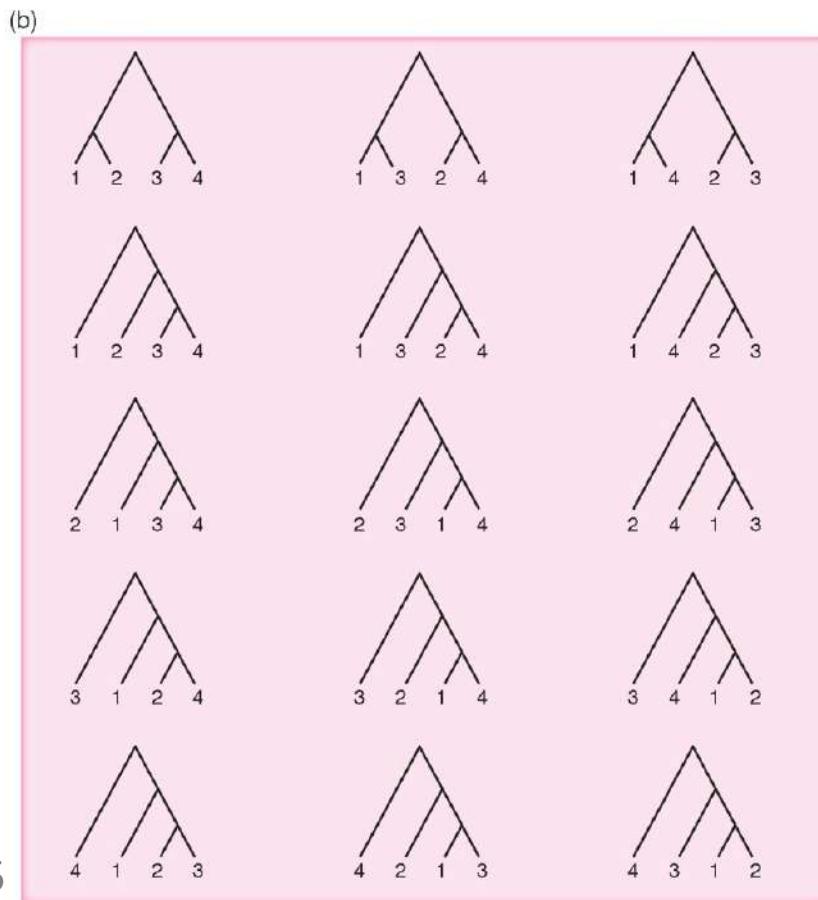


Three rooted trees are possible.

## Numbers of rooted and unrooted trees: 4 OTUs



For 4 OTUs there are three possible unrooted trees.

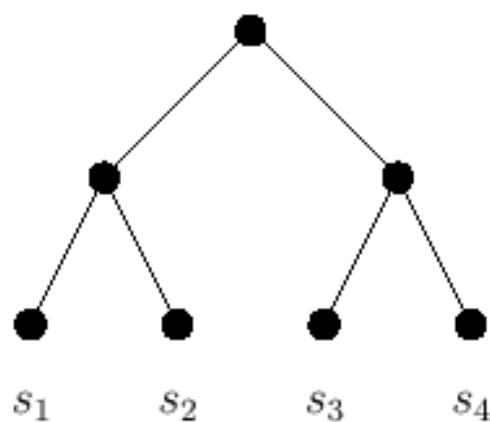


For 4 OTUs there are 15 possible rooted trees.

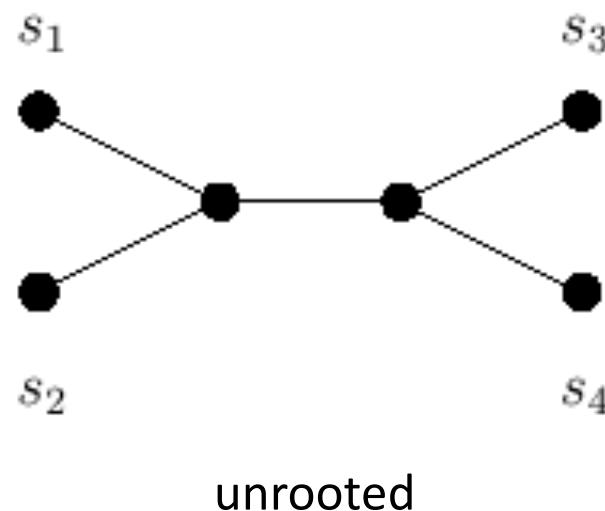
There is only one of these 15 trees that accurately describes the evolutionary process by which these four sequences evolved.

# Evolution Tree

- We construct a evolution tree with the species as the leaf node in order to describe the relationships among species.
  - Rooted evolution tree: The degree of internal node is 3, except the root node. If the evolution tree is ultrametric, then the distances from root to all leaf nodes are the same.
  - Unrooted evolution tree: The degree of internal node is 3.



rooted



unrooted

# The # of unrooted tree

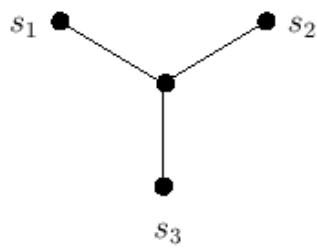
$n = 2$

1

$s_1$  —————  $s_2$

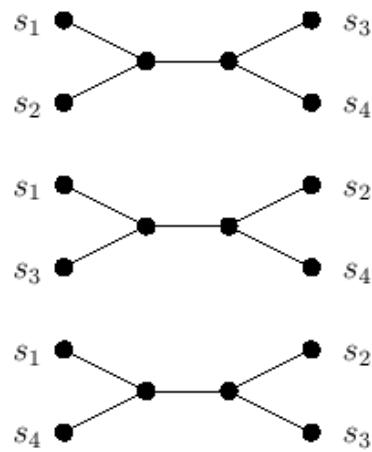
$n = 3$

1



$n = 4$

3



Let  $E(k)$  denote the # of edges in the unrooted tree for  $k$  species.

Let  $TU(k)$  denote the # of unrooted trees.

$$\begin{aligned} TU(k) &= TU(k-1) \cdot E(k-1) \\ &= (TU(k-2) \cdot E(k-2)) \cdot E(k-1) \\ &= \prod_{i=1}^{k-2} E(k-i) \\ &= \prod_{i=1}^{k-2} (2k - 2i - 3) \\ &= 1 \cdot 3 \cdot 5 \cdots (2k-5) \end{aligned}$$

## The # of rooted tree

- Let  $TR(k)$  denote the # of rooted trees for  $k$  species.

$$TR(k) = (2k - 3) \cdot TU(k)$$

$$= TU(k) \cdot E(k)$$

$$= TU(k + 1)$$

- $TU(15) = 2,027,025$

- $TU(20) \geq 10^{12}$

Number of possible tree topologies for  $n$  taxa

$$N_{trees} = 3 \cdot 5 \cdot 7 \cdots (2n-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

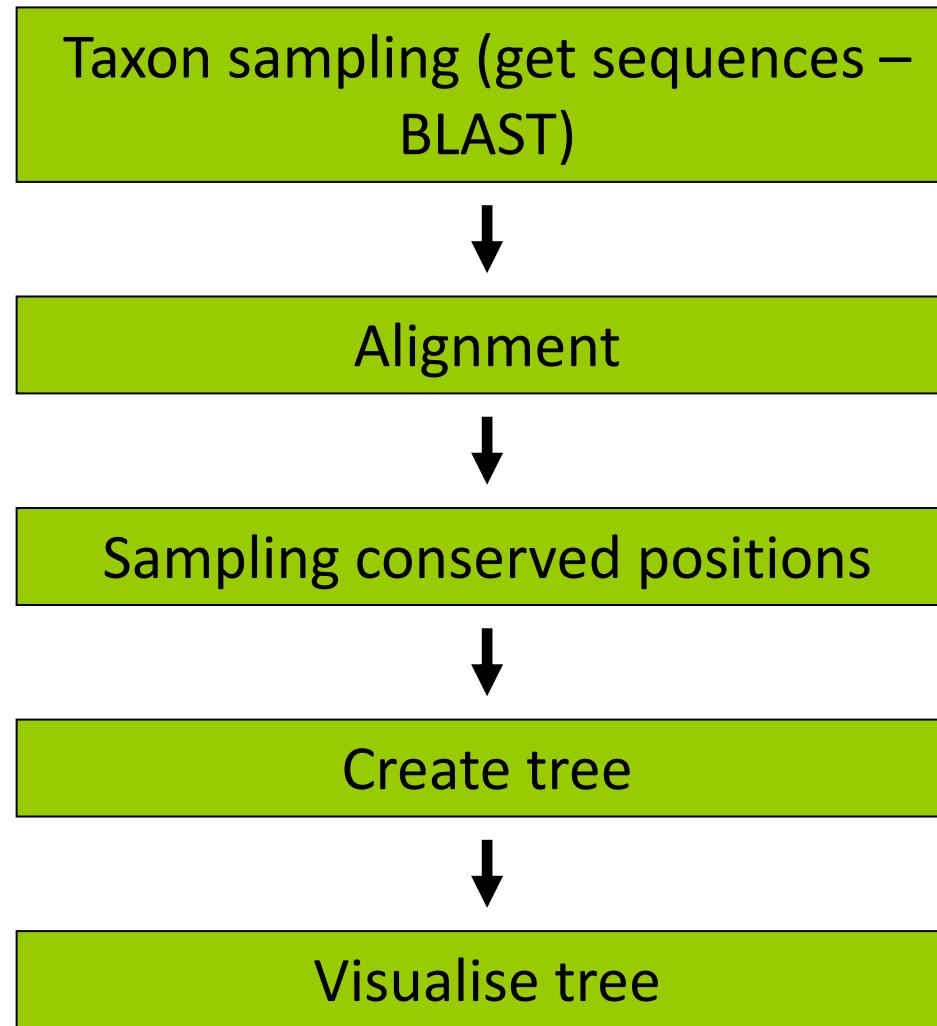
$n$	$N_{trees}$
4	3
5	15
6	105
7	945
...	...
10	2,027,025
...	...
20	$\sim 2 \times 10^{20}$

# Numbers of trees

# of OTUs	# of unrooted trees	# of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
10	105	34,459,425
20	$2 \times 10^{20}$	$8 \times 10^{21}$

# Five stages of phylogenetic analysis

# Flow to build Phylogenetic tree



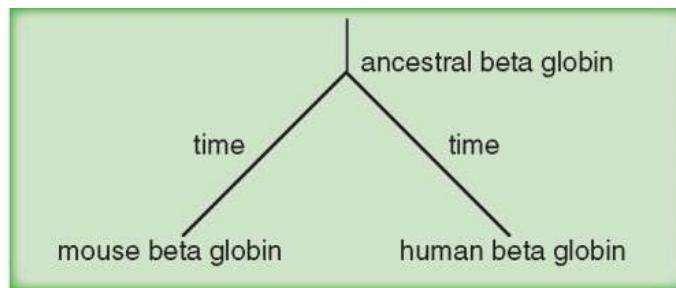
# Stage 1: Use of DNA, RNA, or protein

- If the synonymous substitution rate ( $d_S$ ) is greater than the nonsynonymous substitution rate ( $d_N$ ), the DNA sequence is under negative (purifying) selection. This limits change in the sequence (e.g. insulin A chain).
- If  $d_S < d_N$ , positive selection occurs. For example, a duplicated gene may evolve rapidly to assume new functions.

# Stage 1: Use of DNA, RNA, or protein

- For phylogeny, DNA can be more informative.
- Some substitutions in a DNA sequence alignment can be directly observed: single nucleotide substitutions, sequential substitutions, coincidental substitutions. Additional mutational events can be inferred by analysis of ancestral sequences.

Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time



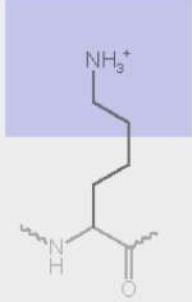
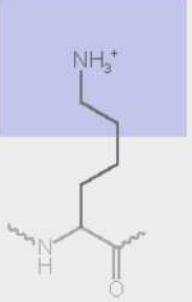
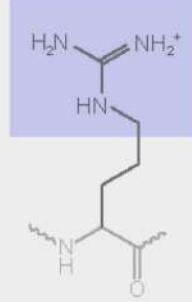
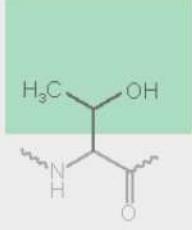
ancestral	M	V	H	L	S	P	V	E	K	S	A	V
human	M	V	H	L	T	P	E	E	K	S	A	V
mouse	M	V	H	L	T	D	A	E	K	S	A	V

protein

ancestral	5'	ATG	GTG	CAT	CTG	AGT	CCT	GTT	CAG	AAG	TCT	GCT	GTT	3'
human	5'	ATG	GTG	CAT	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	3'
mouse	5'	ATG	GTG	CAC	CTG	ACT	GAT	GCT	GAG	AAG	TCT	GCT	GTC	3'

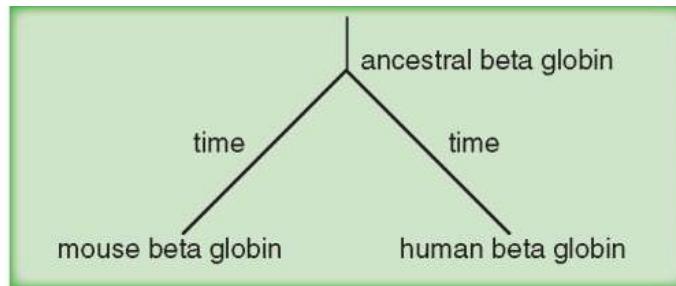
DNA

# Types of point mutations

Point mutations				
No mutation	Silent	Nonsense	Missense	
			conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC      TGC
mRNA level	AAG	AAA	UAG	AGG      ACG
protein level	Lys	Lys	STOP	Arg      Thr
				
				

basic  
polar

Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time



ancestral	M	V	H	L	S	P	V	E	K	S	A	V
human	M	V	H	L	T	P	E	E	K	S	A	V
mouse	M	V	H	L	T	D	A	E	K	S	A	V

ancestral	5'	ATG	GTG	CAT	CTG	AGT	CCT	GTT	CAG	AAG	TCT	GCT	GTT	3'
human	5'	ATG	GTG	CAT	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	3'
mouse	5'	ATG	GTG	CAC	CTG	ACT	GAT	GCT	GAG	AAG	TCT	GCT	GTC	3'

A	A	A	AA	
G	G → C	G → C	CC	parallel substitutions
T	T	T	TT	
C	C	C → G	CG	single substitution
C	C	C → T → A	CA	sequential substitution
T	T	T	TT	
G	G	G	GG	
T	T → A	T → C	AC	coincidental substitutions
T	T → G	T	GT	
C	C → G	C → T → G	GG	convergent substitutions
A	A	A	AA	
G	G → T → G	G	GG	back substitution

ancestral globin (hypothetical)	human globin	mouse globin	observed alignment	Substitution mechanism
------------------------------------	--------------	--------------	--------------------	------------------------

# Step matrices: number of steps required to change a character

(a)

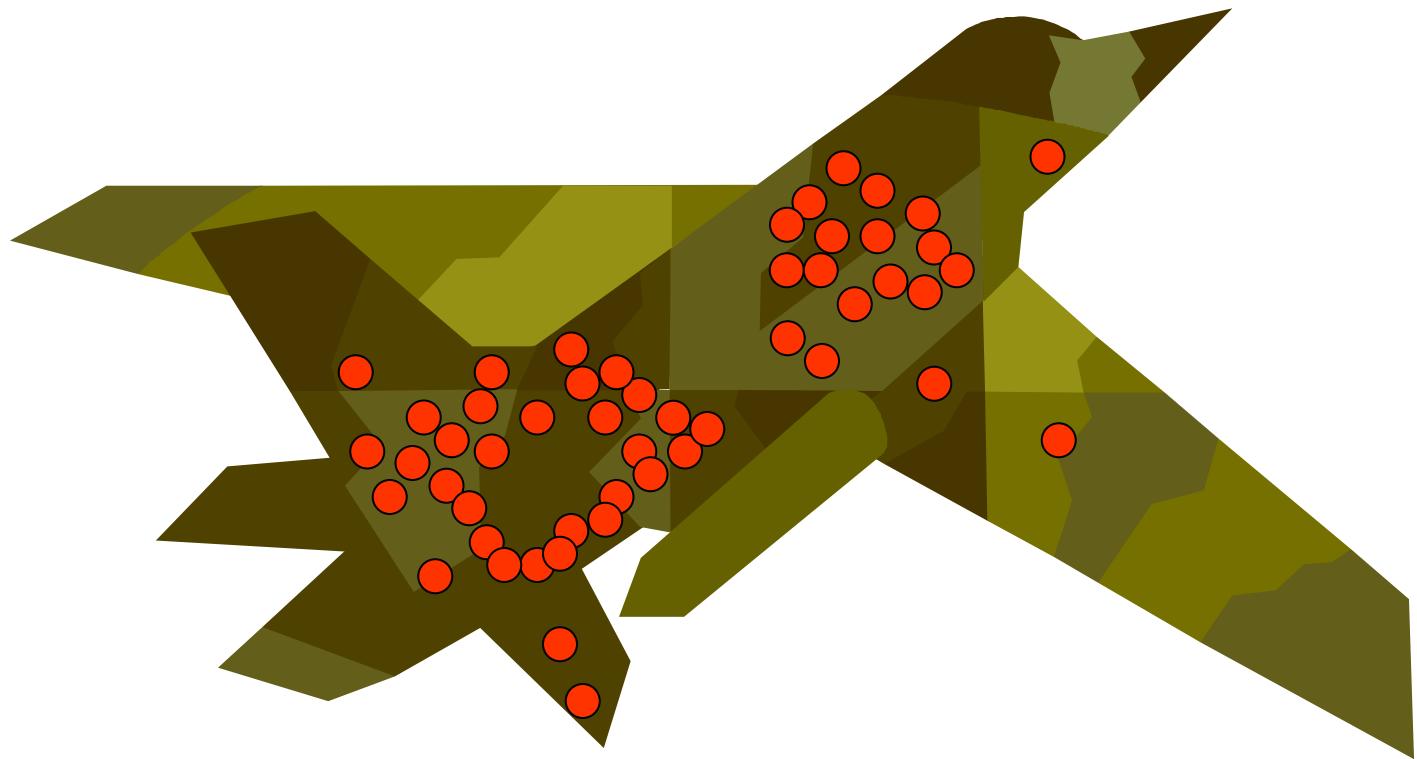
	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

nucleotide step matrix

(b)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C	0	2	3	1	1	2	2	3	2	3	2	2	2	3	1	1	2	2	1	1
D	0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	2	1	3	1	
E	0	3	1	2	2	1	2	2	2	2	2	2	1	2	2	2	1	2	2	
F	0	2	2	1	3	1	2	2	2	2	3	2	1	2	1	2	1	2	1	
G	0	2	2	2	2	2	2	2	2	2	2	1	1	2	1	1	1	1	2	
H	0	2	2	1	3	1	1	1	1	1	1	1	1	2	2	2	2	3	1	
I	0	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	3	2	
K	0	2	1	1	2	1	1	1	2	1	1	2	1	1	2	2	2	2	2	
L	0	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2		
M							0	2	2	2	1	2	1	1	1	2	1	2	3	
N							0	2	2	2	1	1	1	2	3	1				
P							0	1	1	1	1	1	2	2	2	2				
Q							0	1	2	2	2	2	2	2	2					
R							0	1	1	2	1	2	1	2	2	2				
S							0	1	2	1	1	1	2	1	1	1				
T							0	2	2	1	2	1	2	2	2	2				
V							0	2	2	2	1	2	2	2	2	2				
W							0	2	2	2	2	1	2	2	2	2				
Y							0													

amino acid  
step matrix



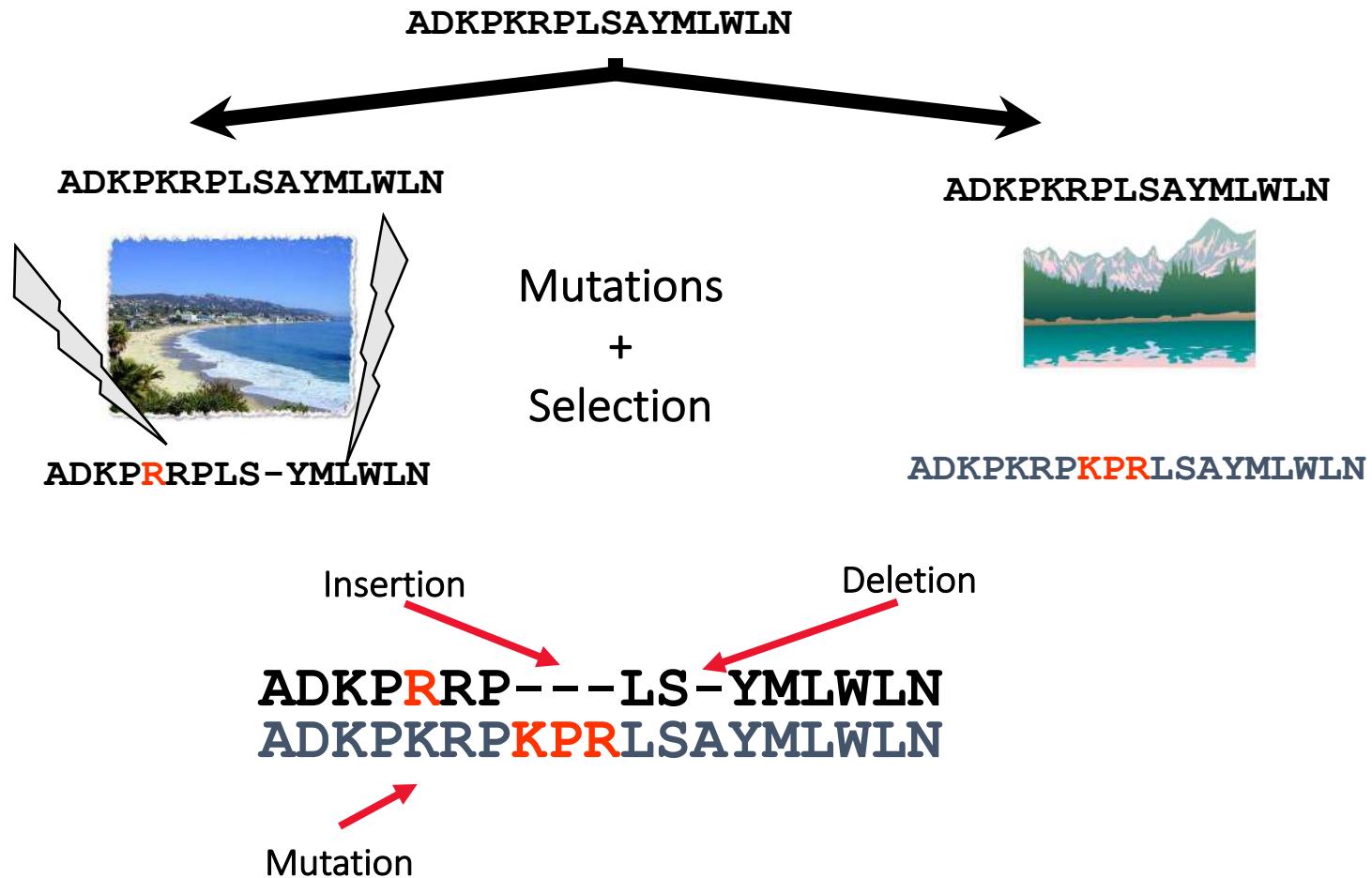
## ON PROTEIN SYNTHESIS

BY F. H. C. CRICK

Medical Research Council Unit for the Study of Molecular Biology,  
Cavendish Laboratory, Cambridge

Biologists should realise that before long we shall have a subject which might be called 'protein taxonomy'—the study of the amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them.

# An Alignment is a STORY



## Stage 2: Multiple sequence alignment

- The fundamental basis of a phylogenetic tree is a multiple sequence alignment.
- (If there is a misalignment, or if a nonhomologous sequence is included in the alignment, it will still be possible to generate a tree.)
- Consider the following alignment of 13 homologous globin proteins (see Fig. 3.2)

# Sequence alignment

**PHYL** 

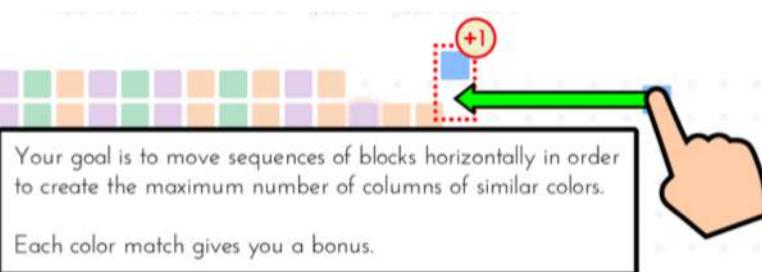
- SOLVE A PUZZLE AND HELP GENETIC DISEASE RESEARCH -



Heart and muscles diseases    Cancers    Metabolic diseases    Digestive and respiratory system diseases  
Blood and immune system diseases    Brain, nervous and sensory system diseases    Infectious diseases    Other diseases



<http://phylo.cs.mcgill.ca/>



Your goal is to move sequences of blocks horizontally in order to create the maximum number of columns of similar colors.  
Each color match gives you a bonus.

#543 | STAGE 1/11

Par 18 **-21** -21 Best Score

matches 18 - mismatches 4 - gaps 8 - gap extends 3



#543 | STAGE 1/11

Par 18 **18** 18 18 Best Score

matches 20 - mismatches 2 - gaps 0 - gap extends 0



However, the sequences are not identical. Thus, color mismatches and gaps are unavoidable and you receive penalties for that.

Your challenge is to find the best trade-off between bonuses and penalties. N.B.: Smaller blocks highlight mismatches.

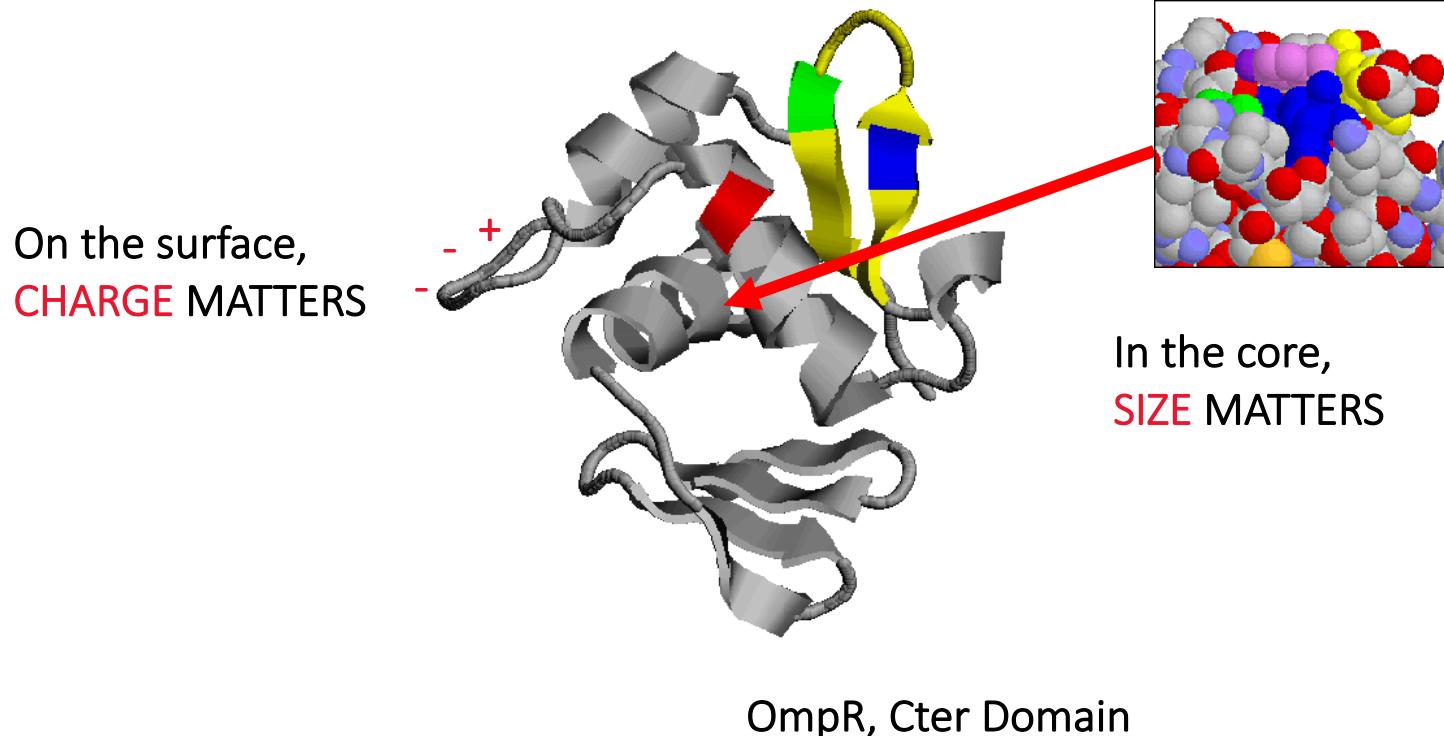
# Pairwise alignment

# pairwise alignment

- The process of lining up two sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

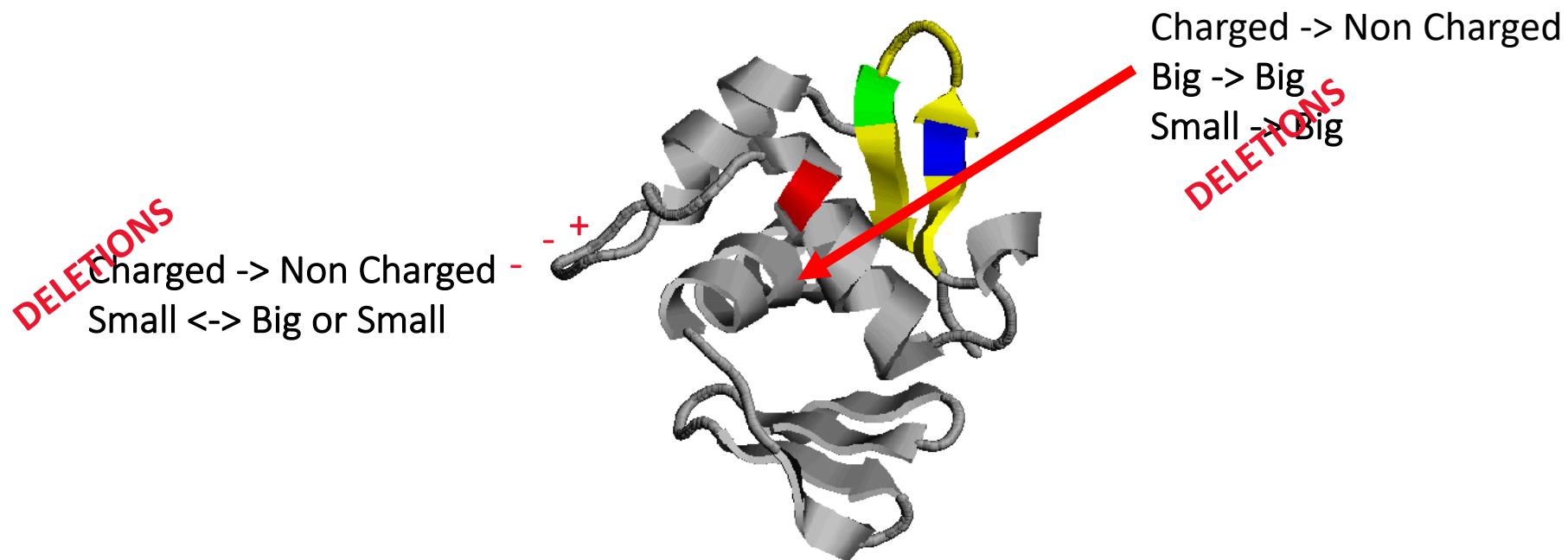
# How Do Sequences Evolve ?

- In a structure, each Amino Acid plays a Special Role



# How Do Sequences Evolve ?

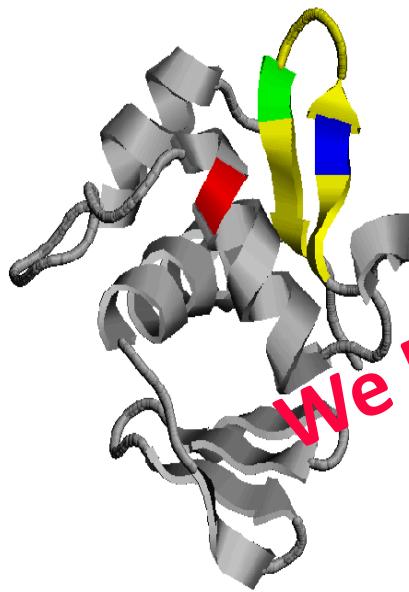
- Accepted Mutations Depend on the Structure



# How Can We Compare Sequences ?

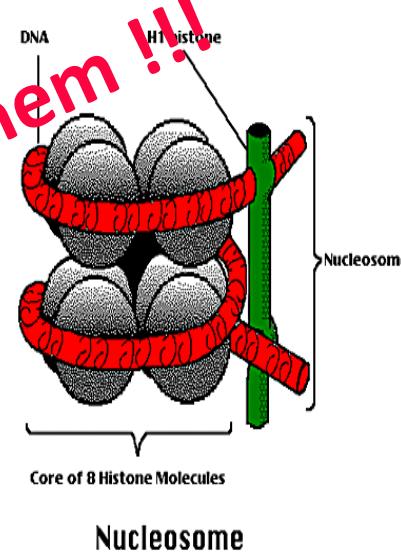
- To Compare Two Sequences, We need:

Their Structure



We Do Not Have Them !!!

Their Function



# Measures of the distance

- Hamming distance
  - agtc
  - cgta
- Edit distance (Levenshtein)
  - ag-tcc
  - cgctca

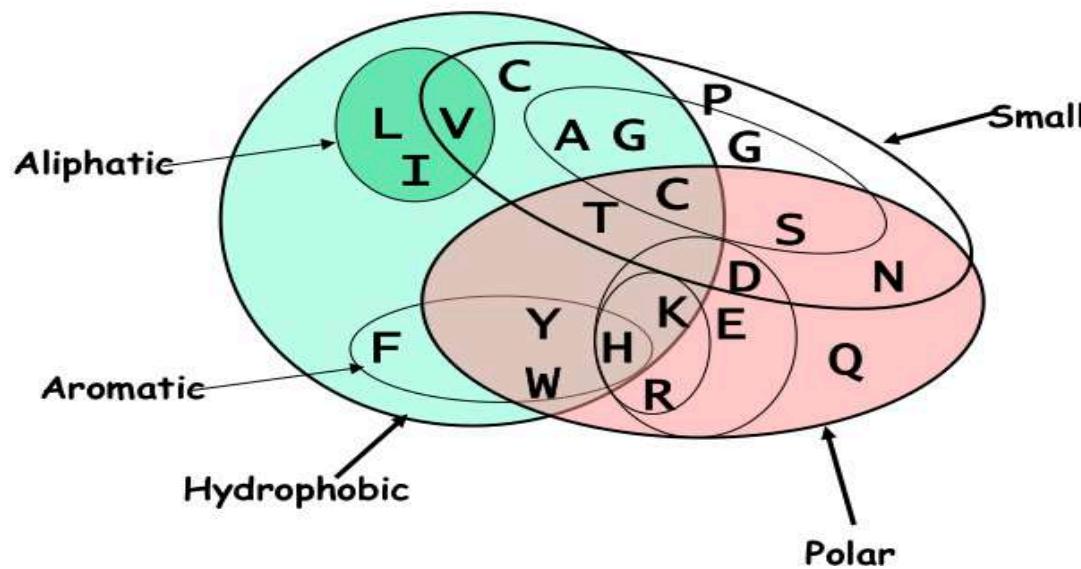
# Substitution Matrix

# How Can We Compare Sequences ?

- To Compare Sequences, We need to Compare Residues
- We Need to Know How Much it **COSTS** to **SUBSTITUTE**
  - an Alanine into an Isoleucine
  - a Tryptophan into a Glycine
- The table that contains the costs for all the possible substitutions is called the **SUBSTITUTION MATRIX**

# How Can We Compare Sequences ?

- Using Knowledge Could Work



But we do not know enough about Evolution and Structure.

Using Data works better.

# Making a Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5						
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-4	-7	-5	-3	-3	0	10			
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

The Diagonal Indicates How Conserved a residue tends to be.  
W is VERY Conserved

Some Residues are Easier To mutate into other similar.

How to derive that matrix?

# Substitution Matrix

- contains values proportional to the probability that amino acid  $i$  mutates into amino acid  $j$  for all pairs of amino acids.
- constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.
- should reflect the true probabilities of mutations occurring through a period of evolution.
- The two major types of substitution matrices
  - PAM
  - BLOSUM

# Making a Substitution Matrix

- Take 100 nice pairs of Protein Sequences,  
easy to align (80% identical).
- Align them...
- Count each mutations in the alignments
  - 25 Tryptophans into phenylalanine
  - 30 Isoleucine into Leucine
  - ....

# Point Accepted Mutations/ Percent accepted mutation (PAMs)

- Margaret Dayhoff and colleagues developed scoring matrices in the 1960s and 1970s.
  - Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978). "A model of Evolutionary Change in Proteins". *Atlas of protein sequence and structure* (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found. pp. 345–358
- A unit of evolutionary distance between 2 amino acid sequences.
- 1 PAM
  - 1 accepted point-mutation (no insertions or deletions) event per 100 aa



Margaret Oakley Dayhoff, 1925~1983

# Dayhoff's protein superfamilies

- Samples to sequences that are sufficiently similar, why?
  - No position has changed more than once
- PAM matrices are based on global alignments of closely related proteins.
- Examined 1572 changes in 71 groups of closely related proteins (>85% amino acid identity).

Some protein families evolve very slowly (e.g. histones change little over 100 million years); others change very rapidly

PROTEIN	PAMS PER 100 MILLION YEARS
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome c	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

accepted point mutations are defined not by the pairwise alignment but with respect to the common ancestor

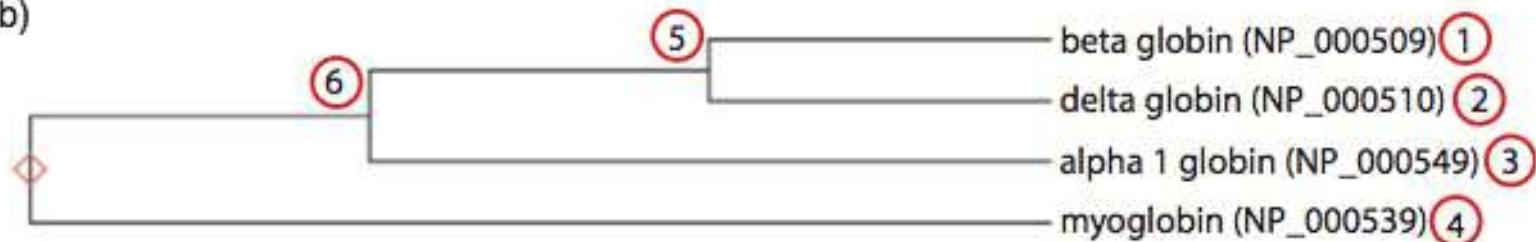
- Dayhoff et al. applied an evolutionary model to compare changes such as **1** versus **2** not to each other but to an inferred common ancestor at position **5**.

(a)

beta globin	MVHLTPEEKSAVTALWGKV
delta globin	MVHLTPEEKTAVNALWGKV
alpha 1 globin	MV.LSPADKT <b>NVKA</b> WGKV
myoglobin	.MGLSD <b>G</b> EWQL <b>V</b> LNVWGKV
5	MVHL <b>S</b> PEEKT <b>A</b> VNALWGKV
6	MVHL <b>T</b> PEEKT <b>A</b> VNALWGKV



(b)



	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
<b>A</b>																				
<b>R</b>	30																			
<b>N</b>	109	17																		
<b>D</b>	154	0	532																	
<b>C</b>	33	10	0	0																
<b>Q</b>	93	120	50	76	0															
<b>E</b>	266	0	94	831	0	422														
<b>G</b>	579	10	156	162	10	30	112													
<b>H</b>	21	103	226	43	10	243	23	10												
<b>I</b>	66	30	36	13	17	8	35	0	3											
<b>L</b>	95	17	37	0	Y	75	15	17	40	253										
<b>K</b>	57	477	322	85	0	147	104	60	23	43	39									
<b>M</b>	29	17	0	0	0	20	7	7	0	57	207	90								
<b>F</b>	20	7	7	0	0	0	0	17	20	90	167	0	17							
<b>P</b>	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
<b>S</b>	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
<b>T</b>	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
<b>W</b>	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
<b>Y</b>	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
<b>V</b>	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	<b>A Ala</b>	<b>R Arg</b>	<b>N Asn</b>	<b>D Asp</b>	<b>C Cys</b>	<b>Q Gln</b>	<b>E Glu</b>	<b>G Gly</b>	<b>H His</b>	<b>I Ile</b>	<b>L Leu</b>	<b>K Lys</b>	<b>M Met</b>	<b>F Phe</b>	<b>P Pro</b>	<b>S Ser</b>	<b>T Thr</b>	<b>W Trp</b>	<b>Y Tyr</b>	<b>V Val</b>

$A_{ij} = j$  replaced by  $i$

**FIGURE 3.8** Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences. Amino acids are presented alphabetically according to the three-letter code. Notice that some substitutions (green shaded boxes) are very commonly accepted (such as V and I or S and T). Other amino acids, such as C and W, are rarely substituted by any other residue (orange shaded boxes).

substitutions are very common (e.g. D → E, A → G) while others are rare (e.g. C → Q, C → E). The scoring system we use for pairwise alignments should reflect these trends.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A								
R	30							
N	109	17						
D	154	0	532					
C	33	10	0	0				
Q	93	120	50	76	0			
E	266	0	94	831	0	422		
G	579	10	156	162	10	30	112	
H	21	103	226	43	10	243	23	10

# 1 PAM - original amino acids (columns) and replacements (rows)

- The diagonals: most residues remain the same about 99% of the time (see shaded entries).
- cysteine (C) and tryptophan (W) undergo few substitutions, and asparagine (N) many.
- The columns are percentages that sum to 100%.

	Original amino acid																			
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.0
N	0.0	0.0	98.2	0.4	0.0	0.0	0.1													
D	0.1	0.0	0.4	98.6	0.0	0.1	0.5													
C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2
K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
P	0.1	0.																		
S	0.3	0.																		
T	0.2	0.																		
W	0.0	0.																		
Y	0.0	0.																		
V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

$$M_{jj} = 1 - \lambda m_j, \text{ where } \lambda \text{ is a proportion constant}$$

# PAM matrices

- PAM1
  - the matrix calculated from comparisons of sequences with no more than 1% divergence.
  - At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.
- Other PAM matrices are extrapolated from PAM1
  - $\text{PAM}_x = \text{multiplied PAM1 by itself}$

# PAM250 matrix: for proteins that share ~20% identity

- 250 changes have occurred for two proteins over a length of 100 amino acids.
- Assumes mutations occur multiple times at any given position
- The diagonal still has high scores but much information content is lost.

	Original amino acid																				
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2	
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3	
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3	
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2	
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3	
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3	
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7	
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2	
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9	
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13	
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5	
M	1	1	1	1	0	1	1	1	2	3	2	6	2	1	1	1	1	1	2		
F	2	1	2	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3		
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4	
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6	
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6	
W	0	2	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0		
Y	1	1	2	1	3	1	1	3	2	2	1	2	15	1	2	2	3	31	2		
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17	

from cysteine (C) to leucine (L) @ PAM250

$$\bullet s_{LC} = 10 \times \log_{10} \left( \frac{f_{LC}}{f_L} \right) = 10 \times \log_{10} \left( \frac{0.02}{0.085} \right) = -6.3$$

	Original amino acid																				
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2	
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3	
D	5	4	8	11	1	7	10	5	6	3	2	5	3	2	4	5	4	2	3	3	
C	2	1	1	1	52	1	1	2	2	2	2	1	1	1	1	1	1	1	1	1	
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	2	4	5	4	2	3	3	
E	5	4	7	11	1	9	12	5	6	3	2	5	3	2	4	5	4	2	3	3	
G	12	5	10	10	4	7	9	27	5	5	4	5	3	2	4	5	4	2	3	3	
H	2	5	5	4	2	7	4	2	15	2	2	5	3	2	4	5	4	2	3	3	
I	3	2	2	2	2	2	2	2	2	2	2	10	6	5	4	5	4	2	3	3	
L	6	4	4	3	2	6	4	3	5	15	34	5	4	3	2	3	4	3	2	3	
K	6	18	10	8	2	10	8	5	8	5	4	5	4	3	2	3	4	3	2	3	
M	1	1	1	1	0	1	1	1	1	1	2	3	4	3	2	3	4	3	2	3	
F	2	1	2	1	1	1	1	1	1	3	5	6	5	4	3	2	3	4	3	2	
P	7	5	5	4	3	5	4	5	5	3	3	3	2	1	0	1	0	1	0	1	
S	9	6	8	7	7	6	7	9	6	5	4	5	4	3	2	3	4	3	2	3	
T	8	5	6	6	4	5	5	6	4	6	4	5	4	3	2	3	4	3	2	3	
W	0	2	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	
Y	1	1	2	1	3	1	1	1	3	2	2	2	1	0	1	0	1	0	1	0	
V	7	4	4	4	4	4	4	5	4	15	10	5	4	3	2	3	4	3	2	3	

**TABLE 3.1 Normalized frequencies of amino acid. These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence.**

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

# What do the numbers mean in a log odds matrix?

- 0: neutral
- +2: indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance
- -10: that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids

Log-odds matrix for PAM10

	7									
-11		7								
-2	-4	12								
-5	-20	-7	9							
-10	-10	-11	-13	8						
-12	-7	-8	-9	-4	7					
-10	-6	-7	-12	-7	-2	8				
-9	-18	-19	-7	-20	-8	-19	13			
-10	-12	-17	-1	-20	-10	-9	-8	10		
-5	-13	-4	-12	-9	-10	-6	-22	-10	8	
L	K	M	F	P	S	T	W	Y	V	

a scoring matrix with “severe” penalties.  
A match of W to W earns +13, but a mismatch  
(e.g. W aligned to T) has a score of -19, far  
lower than in PAM250.

Log-odds matrix for PAM250

6										
-3	5									
4	0	6								
2	-5	0	9							
-3	-1	-2	-5	6						
-3	0	-2	-3	1	2					
-2	0	-1	-3	0	1	3				
-2	-3	-4	0	-6	-2	-5	17			
-1	-4	-2	7	-5	-3	-3	0	10		
2	-2	2	-1	-1	-1	0	-6	-2	4	
L	K	M	F	P	S	T	W	Y	V	

a useful matrix for comparing distantly related proteins.  
two tryptophan (W) residues earn +17 and a  
W to T mismatch is -5.

# BLOcks SUbstitution Matrix (BLOSUM)

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". *PNAS*. 89 (22): 10915–10919

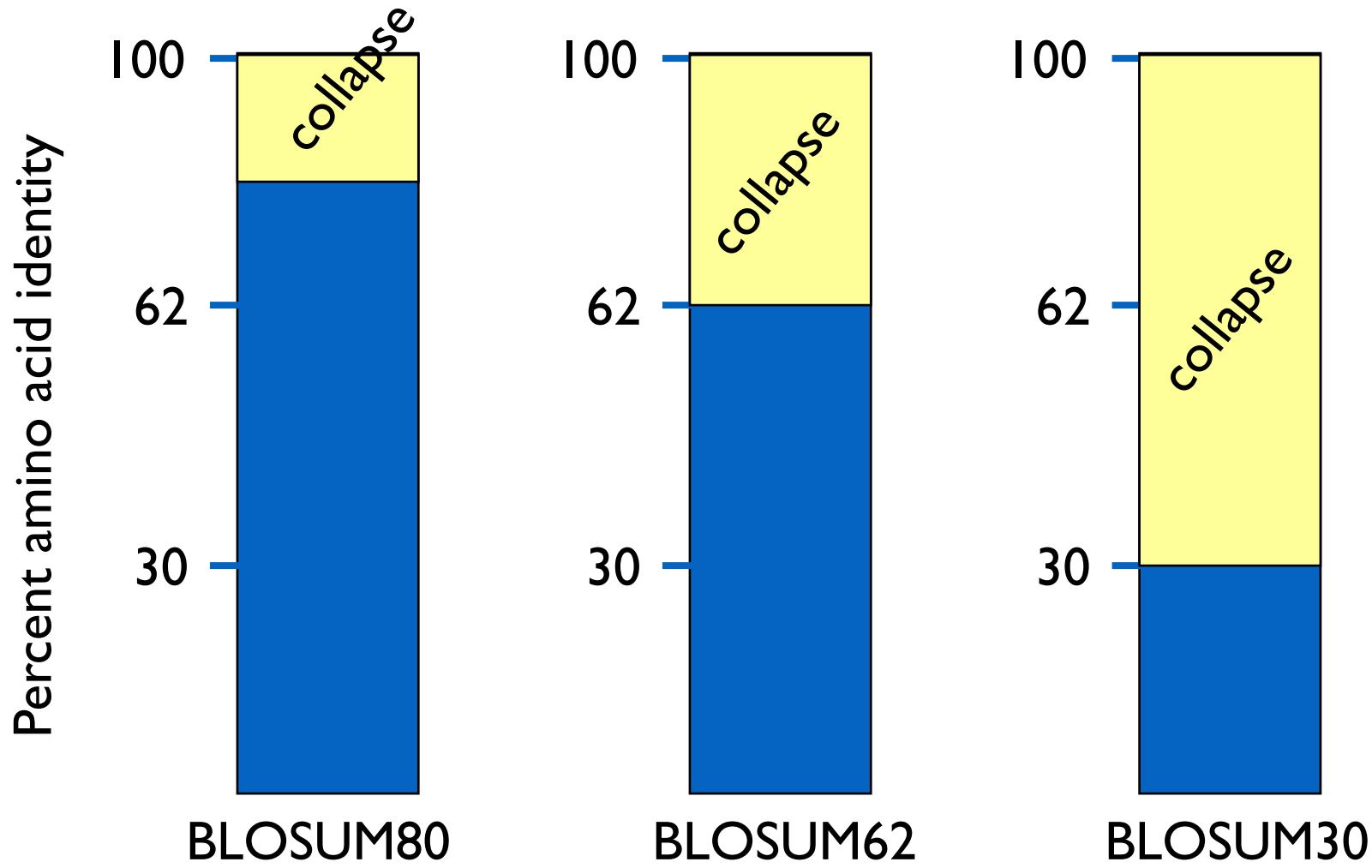
# Idea of BLOSUM

- use aligned ungapped regions of protein families.  
These are assumed to have a common ancestor.
  - based on local alignments
  - Similar ideas but better statistics and modeling. It uses 2000 conserved blocks from 500 families.

# Procedure of BLOSUM

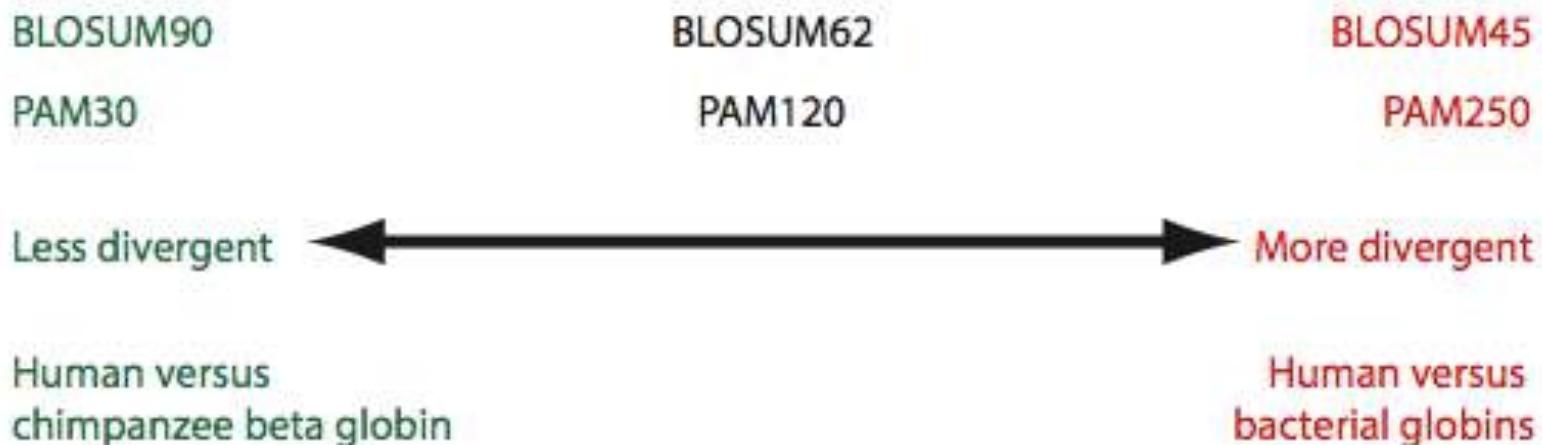
- Cluster together sequences in a family whenever more than L% identical residues are shared, for BLOSUM-L.
- Count number of substitutions across different clusters (in the same family).
- Estimate frequencies using the counts.
- BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

# BLOSUM Matrices



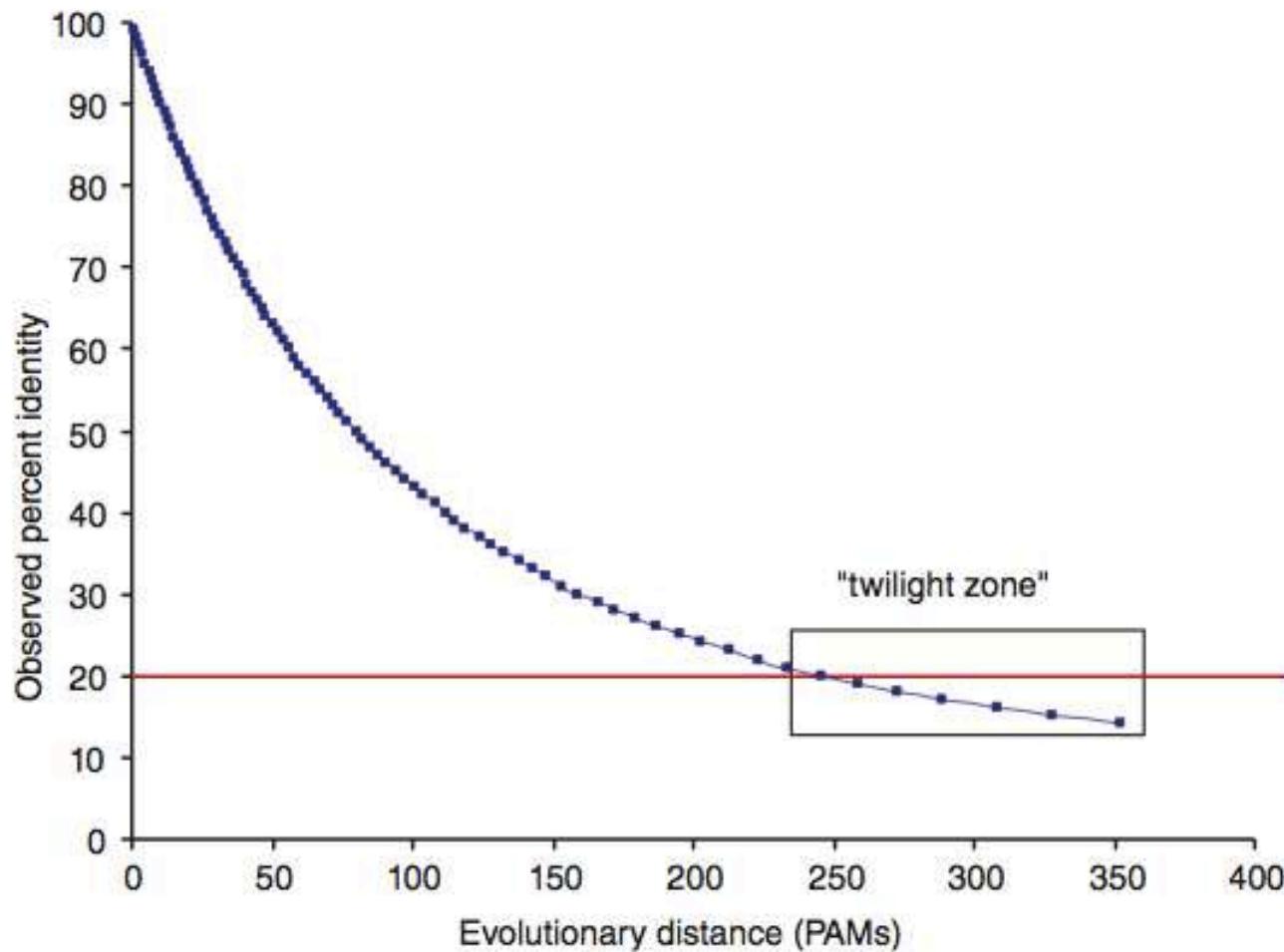
# Summary of PAM and BLOSUM matrices

- A higher PAM number, and a lower BLOSUM number, tends to correspond to a matrix tuned to more divergent proteins.



Could I get some idea of when they  
are likely to do all right?

Two randomly diverging protein sequences change in a negatively exponential fashion



# Pairwise Alignment

# HOW Can we Align Two Sequences?

---

## Different types of pairwise comparisons

---

<i>Method name</i>	<i>Situation</i>
Dot-plot	<b>General exploration of your sequence</b> Discovering repeats Finding long insertion/deletions Extracting portions of sequences to make a multiple alignment
Local alignments	<b>Comparing sequences with partial homology</b> Making high quality alignments Making residue-per-residue analysis
Global alignments	<b>Comparing two sequences over their entire length</b> Identifying long insertion/deletions Checking the quality of your data Identifying every mutation in your sequences

# Dot Matrices

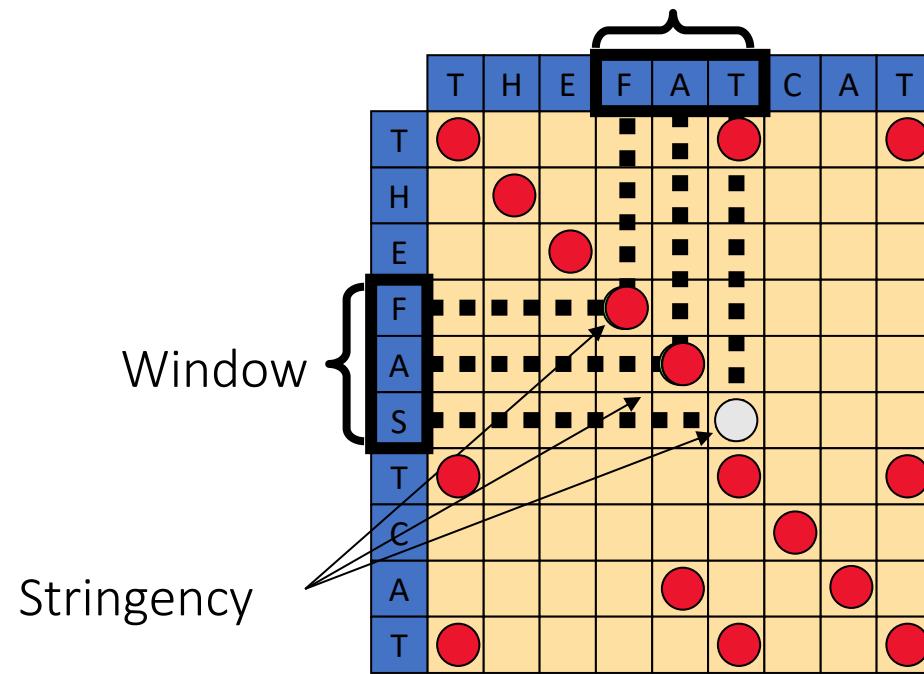
- What are the elements shared by two sequences ?

>Seq1

THEFATCAT

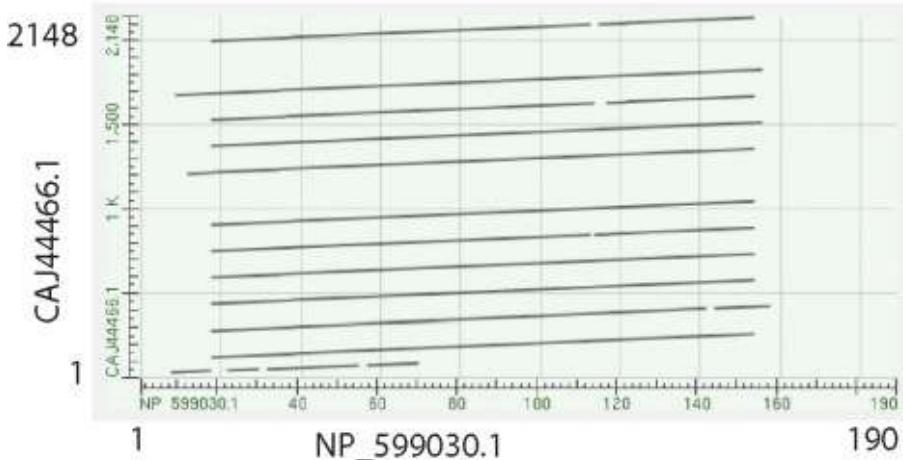
>Seq2

THELASTCAT

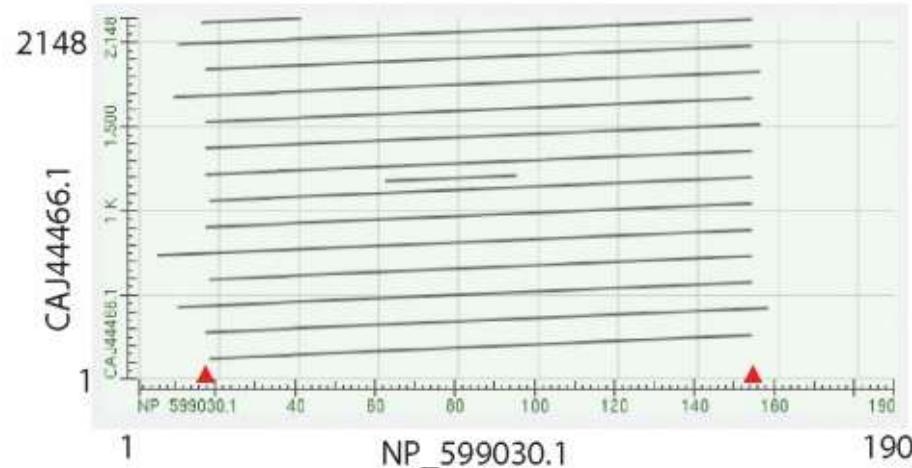


# Pairwise alignment with dotplots

(b) Cytoglobin compared to a snail globin (BLOSUM62)



(c) Cytoglobin compared to a snail globin (PAM250)



Search human cytoglobin against a large snail globin (having many globin repeats). More repeats are observed using PAM250 than BLOSUM62.

To “read” this plot note that cytoglobin (x-axis) matches the snail globin (y-axis) at about a dozen locations across the snail protein. Red arrows indicate that the first few and last few amino acids of cytoglobin do not participate in this repeat structure.

# Dot Matrices - limits

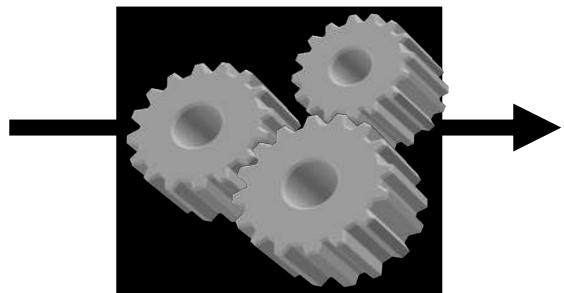
- Visual aid
  - Best Way to **EXPLORE** the Sequence Organisation
  - Does **NOT** provide us with an **ALIGNMENT**

wheat --DPNPKRAMTSFVFFMSEFRSEFKQKHSKLKSIVEMVKAAGER  
| | | | | | | | | | | | | | | | | | | | | | | | | | | |  
????? KKDSNAPKRAMTSFMFFSSDFRS----KHSDL-SIVEMSKAAGAA

# Global Alignments

- Take 2 Nice Protein Sequences
- A good Substitution Matrix (blosum)
- **DYNAMIC PROGRAMMING**

>Seq1  
THEFATCAT  
>Seq2  
THEFASTCAT



DYNAMIC  
PROGRAMMING

# Using Dynamic Programming To Align Sequences

- Understanding the DP concept
- Coding a Global and a Local Algorithm
- Aligning with affine gap penalties

# Dynamic Programming

# THE THEORY OF DYNAMIC PROGRAMMING

RICHARD BELLMAN

**1. Introduction.** Before turning to a discussion of some representative problems which will permit us to exhibit various mathematical features of the theory, let us present a brief survey of the fundamental concepts, hopes, and aspirations of dynamic programming.

To begin with, the theory was created to treat the mathematical problems arising from the study of various multi-stage decision processes, which may roughly be described in the following way: We have a physical system whose state at any time  $t$  is determined by a set of quantities which we call state parameters, or state variables. At certain times, which may be prescribed in advance, or which may be determined by the process itself, we are called upon to make decisions which will affect the state of the system. These decisions are equivalent to transformations of the state variables, the choice of a decision being identical with the choice of a transformation. The outcome of the preceding decisions is to be used to guide the choice of future ones, with the purpose of the whole process that of maximizing some function of the parameters describing the final state.

# Using Dynamic Programming To Align Sequences

- DP invented in the 1950s by Bellman
  - Programming  $\Leftrightarrow$  Tabulation
- Re-invented in 1970 by Needlman and Wunsch
  - It took 10 year to find out...
  - *Needleman, Saul B. & Wunsch, Christian D. (1970). "[A general method applicable to the search for similarities in the amino acid sequence of two proteins](#)". Journal of Molecular Biology. 48 (3): 443–53*

# Global Alignment

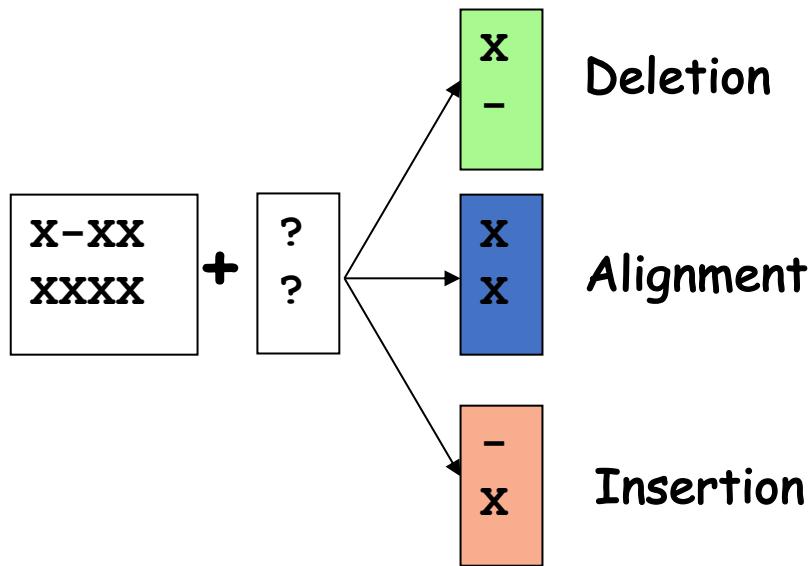
*Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 48 (3): 443–53*

# The Foolish Assumption

- The score of each column of the alignment is independent from the rest of the alignment
- It is possible to model the relationship between two sequences with:
  - A substitution matrix
  - A simple gap penalty

# The Principle of DP

- If you extend optimally an optimal alignment of two sub-sequences, the result remains an optimal alignment

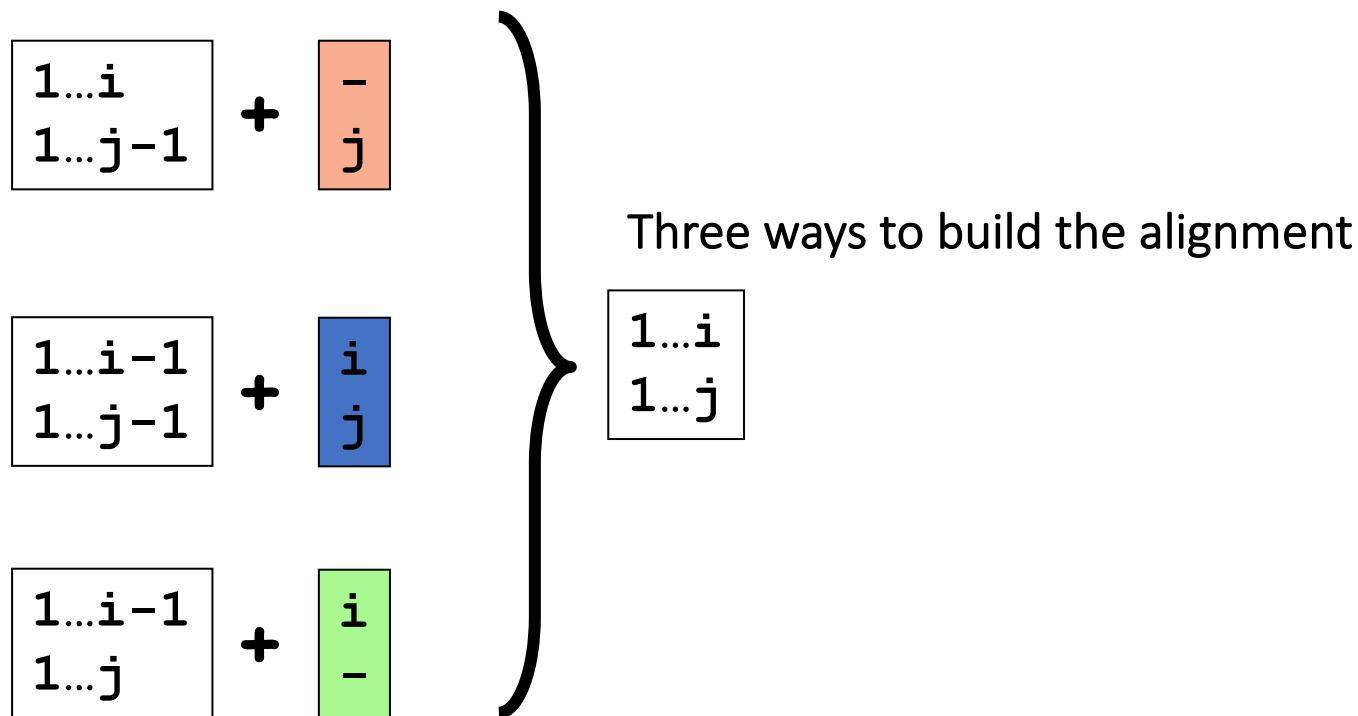


# Finding the score of $i,j$

- Sequence 1:  $[1-i]$
- Sequence 2:  $[1-j]$
- The optimal alignment of  $[1-i]$  vs  $[1-j]$  can finish in three different manners:



# Finding the score of $i,j$



# Formalizing the algorithm

$$\text{score\_m}(i,j) = \text{best} \left\{ \begin{array}{l} \text{score\_m}(i,j-1) + \text{gap\_s} \\ \text{score\_m}(i-1,j-1) + \\ \text{match\_s/mismatch\_s} \\ \text{score\_m}(i-1,j) + \text{gap\_s} \end{array} \right.$$

$1 \dots i$   
 $1 \dots j-1$  + 

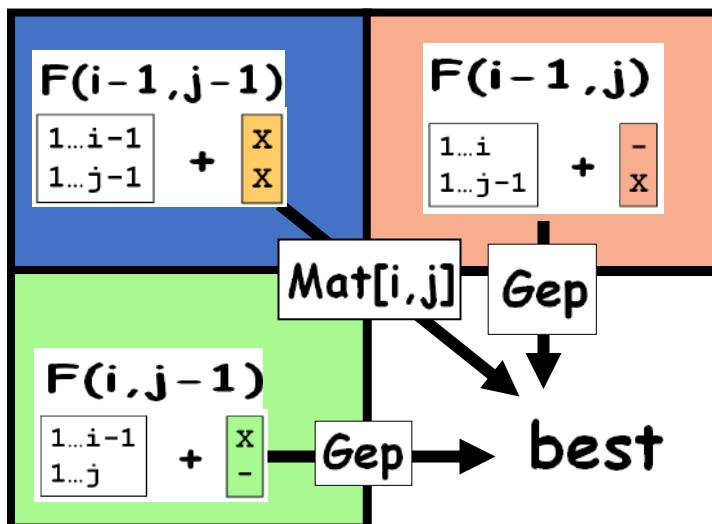
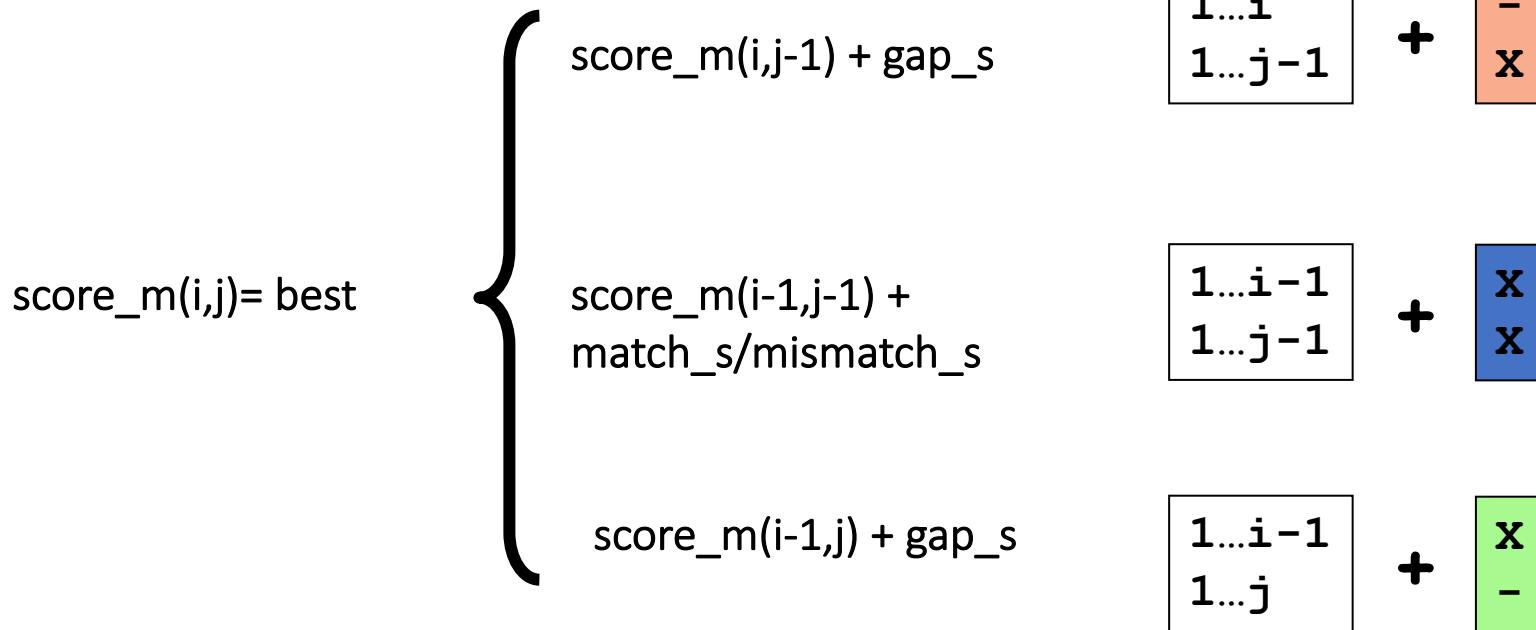
$1 \dots i-1$   
 $1 \dots j-1$  + 

$1 \dots i-1$   
 $1 \dots j$  + 

# Arranging Everything in a Table

	-	F	A	T
-				
F		1... <u>I-1</u> 1... <u>J-1</u>	1...I 1... <u>J-1</u>	
A		1... <u>I-1</u> 1...J	1...I 1...J	
S				
T				

# Filing Up The Matrix

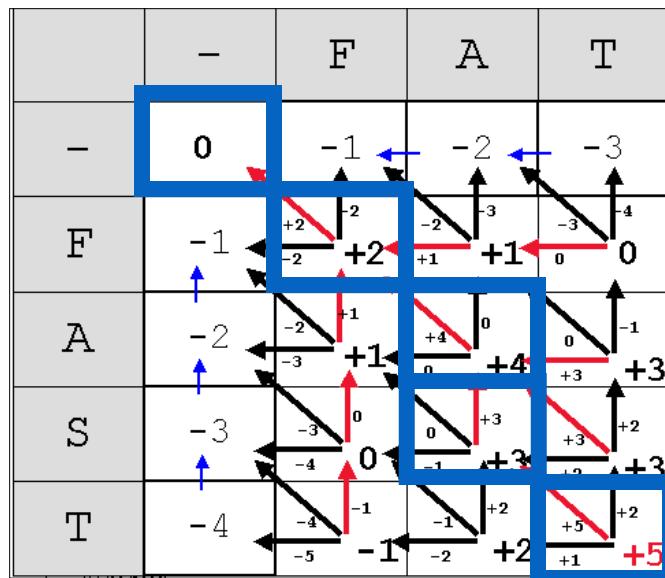


	-	F	A	T
-	0	-1 -2	-2 -3	-3
F	-1 +2	+2 -2 -3 +1	+1 -2 -3 0	0
A	-2 -3 +1	+1 -2 -3 0	+4 0 +4 +3	-1 0 +3
S	-3 -4 0	-3 -4 0	+3 -1 +3 +2	+2 +3 +3
T	-4 -4 -1	-5 -4 -1	-2 -1 +2 +2	+1 +2 +5

The matrix shows payoffs for Player 1 (rows) and Player 2 (columns). Red arrows indicate dominant strategies:

- Player 1's strategy F dominates -.
- Player 2's strategy T dominates -.
- Player 1's strategy S dominates A.
- Player 2's strategy T dominates F.
- Player 1's strategy T dominates S.
- Player 2's strategy T dominates A.

# Delivering the alignment: Trace-back



T  
T  
-  
A  
F  
F

Score of 1...3 Vs 1...4  
 $\Leftrightarrow$   
Optimal Aln Score

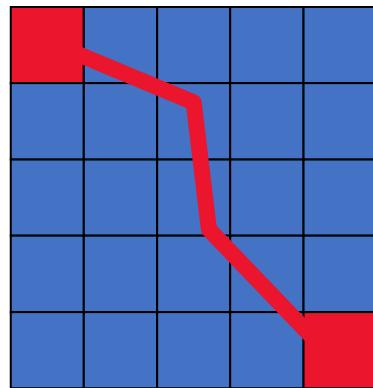
## Trace-back: possible implementation

```
while (! (i==0 && j==0)) :  
    if (direc_m[i][j]== 'sub') :          #SUBSTITUTION  
        aln1[aln_len]=pro1Seq[--i]  
        aln2[aln_len]=pro2Seq[--j]  
    elif (direc_m[i][j]== 'del') :          #DELETION  
        aln1[aln_len]='-'  
        aln2[aln_len]=pro2Seq[--j]  
    elif (direc_m[i][j]== 'ins') :          #INSERTION  
        aln1[aln_len]=pro1Seq[0] [--i]  
        aln2[aln_len]='-'  
    aln_len++  
}
```

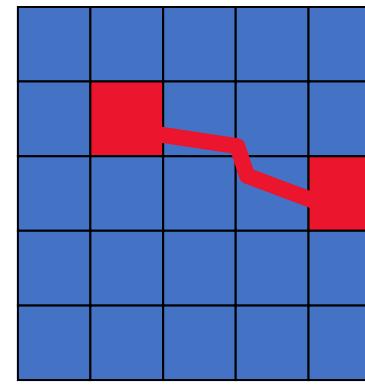
# Local Alignment

Smith, T. F. & Waterman, M. S. Identification of  
common molecular subsequences. *J. Mol.  
Biol.* **147**, 195–7 (1981).

# Local Alignment/Smith And Waterman (SW)



GLOBAL Alignment



LOCAL Alignment

Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–7 (1981).

Adapted from Cedric Notredame

# Global alignment versus local alignment

- Global alignment
  - extends from one end of each sequence to the other.
- Local alignment
  - finds optimally matching regions within two sequences, “subsequences”. => useful to find domains (or limited regions of homology) within sequences
  - almost always used for database searches such as BLAST
  - Smith and Waterman (1981) solved the problem of performing optimal local sequence alignment.
  - Other methods (BLAST, FASTA) are faster but less thorough.

Global alignment (top) includes matches ignored by local alignment (bottom)

(a)

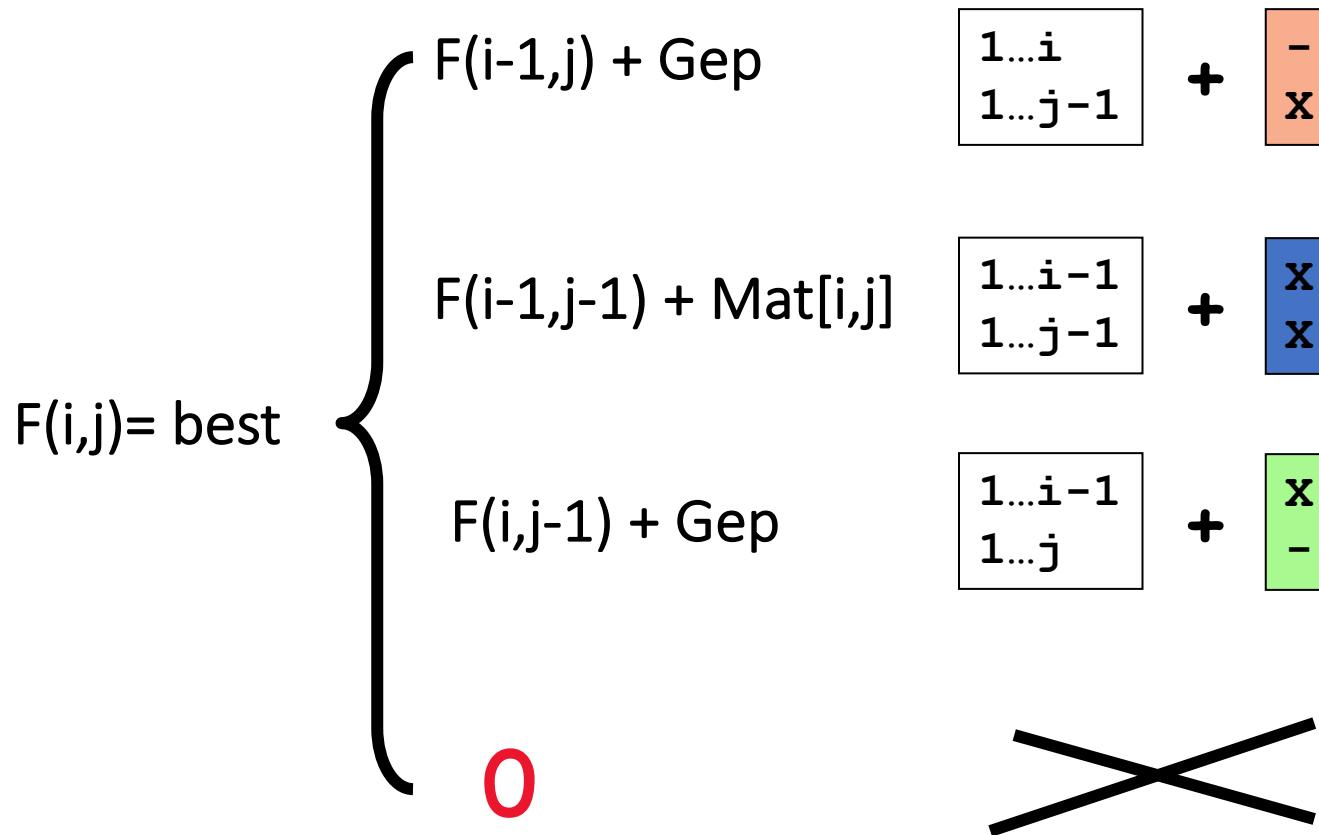
Global:  
5% identity

(b)

NP_824492.1	113	TLYAWAGGAEAFARLTERVFYEKVLKDDVLAPVFEGLMAPEH-----AAHVA : .   . .     : . . . .   . : .   :   .   :   .   . . . .	157
NP_337032.1	10	SFYDAVGGAKTFDAIVSRFYAQVAEDEVLRRVY----PEDDLAGAEERLR	55
NP_824492.1	158	LWLGEVFGGPAAAYSETQGGHGHMVKHLGKNITEVQRRRWVNLLQDAADD : :   . . .       .     .     . . . .   . . . .   : . :   . . . .   . . . .	207
NP_337032.1	56	MFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISLIERDAWLRCMHTAVAS	104
NP_824492.1	208	AGLPT-DAEFRSAFLAYAE 225 . . . .   .   .   . . . .   .   .	
NP_337032.1	105	IDSETLDDEHRRELLDYLE 123	NP 82

Local:  
0% identity

# The Smith and Waterman Algorithm



# The Smith and Waterman Algorithm

0



Ignore The rest of the Matrix



Terminate a local Alignment

# Filing Up a SW Matrix

$$F(i,j) = \text{best}$$

$$\left\{ \begin{array}{l} F(i-1,j) + \text{Gep} \\ F(i-1,j-1) + \text{Mat}[i,j] \\ F(i,j-1) + \text{Gep} \end{array} \right.$$

$$F(i-1,j-1) + \text{Mat}[i,j]$$

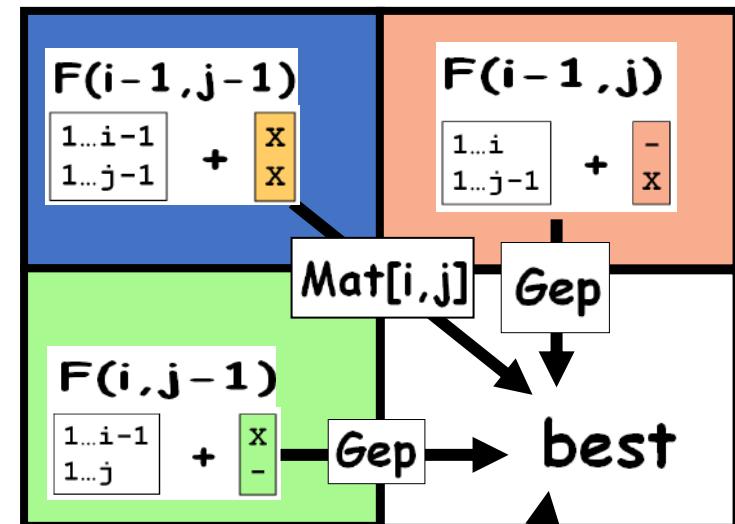
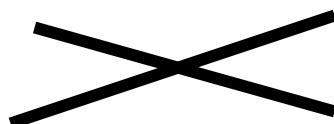
$$F(i,j-1) + \text{Gep}$$

0

$$\begin{matrix} 1 \dots i \\ 1 \dots j-1 \end{matrix} + \begin{matrix} - \\ x \end{matrix}$$

$$\begin{matrix} 1 \dots i-1 \\ 1 \dots j-1 \end{matrix} + \begin{matrix} x \\ x \end{matrix}$$

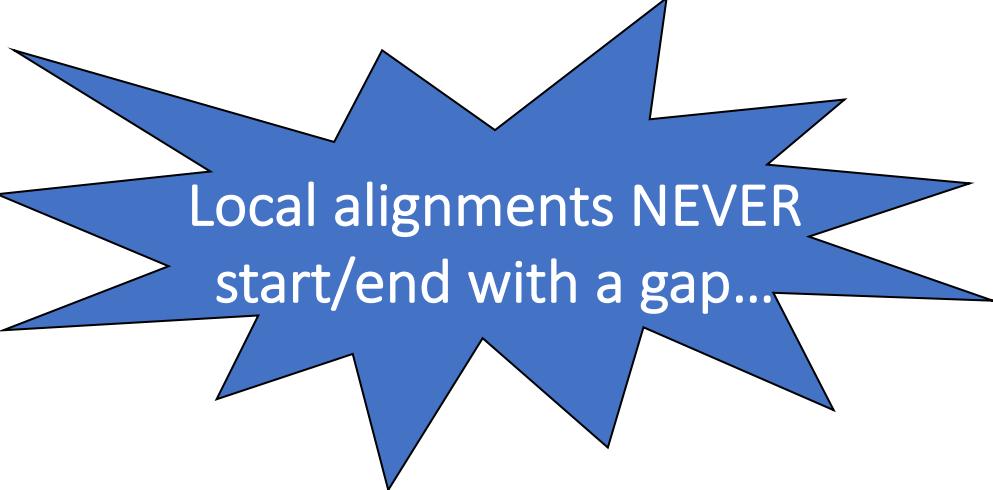
$$\begin{matrix} 1 \dots i-1 \\ 1 \dots j \end{matrix} + \begin{matrix} x \\ - \end{matrix}$$



0

# Filling up a SW matrix: borders

*	-	A	N	I	C	E	C	A	T
-	0	0	0	0	0	0	0	0	0
C	0								
A	0								
T	0								
A	0								
N	0								
D	0								
O	0								
G	0								



Local alignments NEVER  
start/end with a gap...

# Filling up a SW matrix

- Best Local score  $\Leftrightarrow$  Beginning of the trace-back

*	-	A	N	I	C	E	C	A	T
-	0	0	0	0	0	0	0	0	0
C	0	0	0	0	2	0	2	0	0
A	0	2	0	0	0	0	0	4	0
T	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	0	0	4
N	0	0	4	2	0	0	0	0	2
D	0	0	2	2	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0

```

for ($i=1; $i<=$len0; $i++) {
    for ($j=1; $j<=$len1; $j++) {
        if ($res0[0][$i-1] eq $res1[0][$j-1])
            {$s=2;}
        else
            {$s=-1;}

        $sub=$mat[$i-1][$j-1]+$s;
        $del=$mat[$i][$j-1]+$gep;
        $ins=$mat[$i-1][$j]+$gep;

        if ($sub>$del && $sub>$ins && $sub>0)
            {$smat[$i][$j]=$sub;$tb[$i][$j]=$subcode;}
        elsif($del>$ins && $del>0 )
            {$smat[$i][$j]=$del;$tb[$i][$j]=$delcode;}
        elsif( $ins>0 )
            {$smat[$i][$j]=$ins;$tb[$i][$j]=$inscode;}
        else { $smat[$i][$j]=$zero;$tb[$i][$j]=$stopcode; }

        if ($smat[$i][$j]> $best_score) {
            $best_score=$smat[$i][$j];
            $best_i=$i; $best_j=$j;
        }
    }
}

```

Turning  
NW into SW

Prepare Trace back

# The Gotoh Algorithm

## Adding Affine Gap Penalties

### Forcing a bit of Biology into your alignment

Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–8 (1982).

# Gaps

- Positions at which a letter is paired with a null are called gaps.
- Gap scores are typically negative.
- Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is ascribed more significance than the length of the gap.
- Thus there are separate penalties for gap open and gap extension.
- In BLAST, it is rarely necessary to change gap values from the default.

# Insertions and Deletions

- Gap Penalties
    - Opening
    - Extension
  - Opening a gap is more expensive than extending it

# Gap Opening Penalty

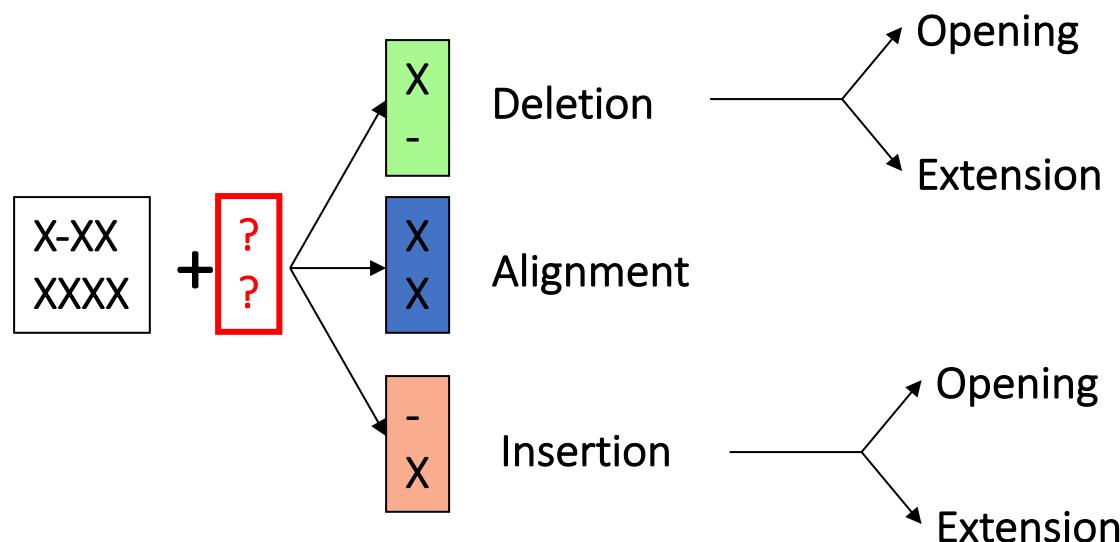
# Gap Extension Penalty

Seq A GARFIELDTHE---CAT

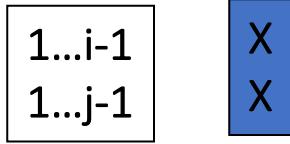
Seq BGARFIELDTHELASTCAT

# But Harder To compute...

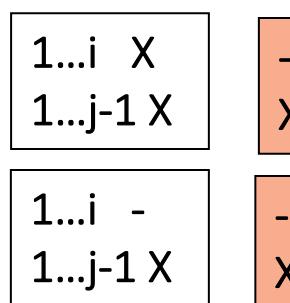
- More Than 3 Ways to extend an Alignment



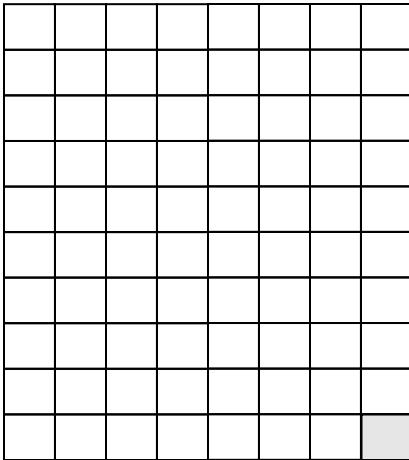
# The Algorithm

$$M(i,j) = \text{best} \quad \left\{ \begin{array}{l} M(i-1,j-1) + Mat(i,j) \\ Ix(i-1,j-1) + Mat(i,j) \\ ly(i-1,j-1) + Mat(i,j) \end{array} \right.$$


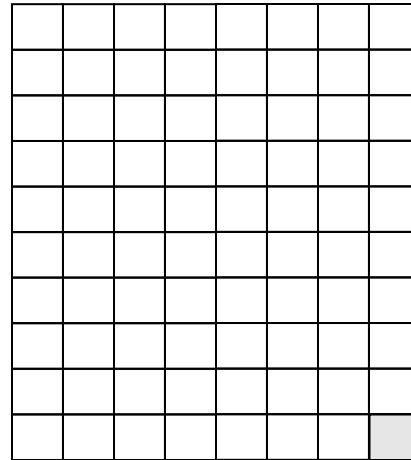
$$Ix(i,j) = \text{best} \quad \left\{ \begin{array}{l} M(i-1,j) + gop \\ Ix(i-1,j) + gep \end{array} \right.$$


$$Ly(i,j) = \text{best} \quad \left\{ \begin{array}{l} M(i,j-1) + gop \\ Ly(i,j-1) + gep \end{array} \right.$$


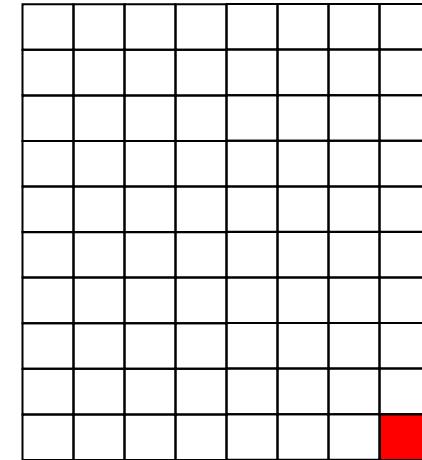
# Trace-back?



$l_x$



$M$



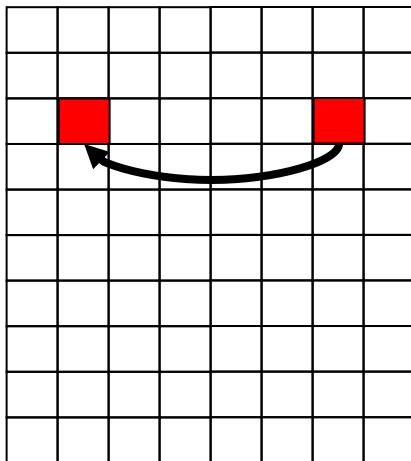
$l_y$

Start From BEST

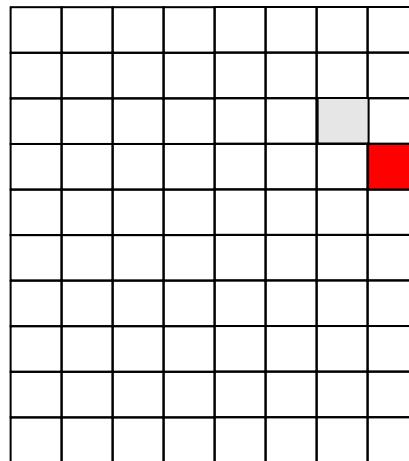
$\left\{ \begin{array}{l} M(i,j) \\ l_x(i,j) \\ l_y(i,j) \end{array} \right.$

# Trace-back?

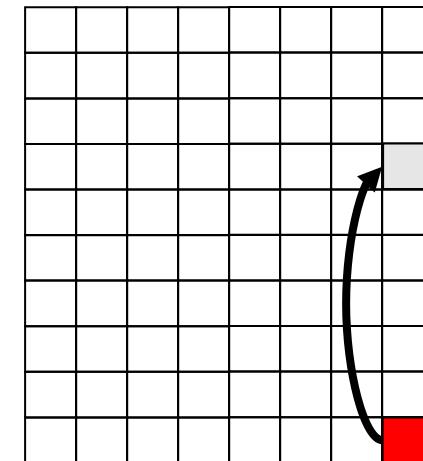
- Navigate from one table to the next, knowing that a gap always finishes with an aligned column...



$\mathbf{lx}$



$\mathbf{M}$



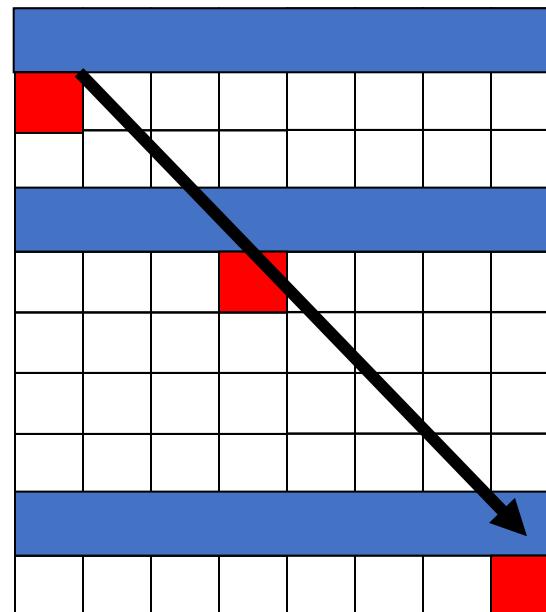
$\mathbf{ly}$

# Remember Not To Run Out of Memory

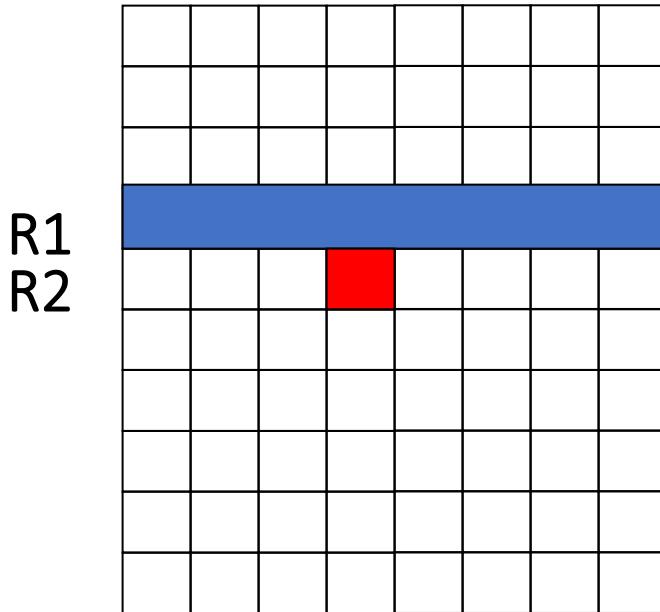
The Myers and Miller Strategy

# A Score in Linear Space

- You never Need More Than The Previous Row To Compute the optimal score

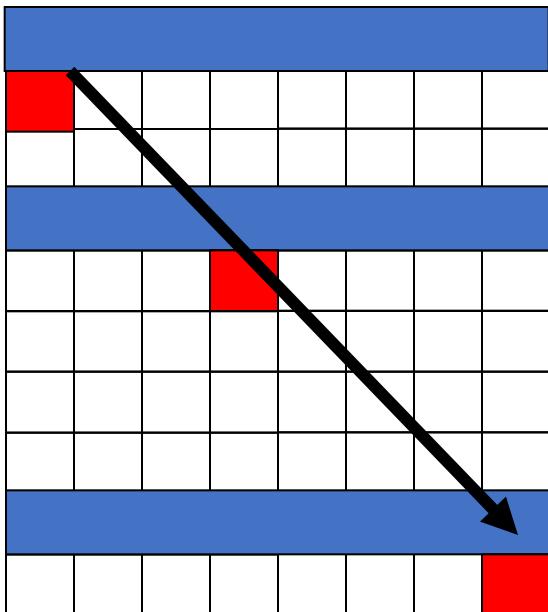


# A Score in Linear Space



```
for i=1:I  
  for j=1:J  
    R2[i][j]=best  
      R2[j-1], +gep  
      R1[j-1]+mat  
      R1[j]+gep  
  for J,  
    R1[j]=R2[j]
```

# A Score in Linear Space



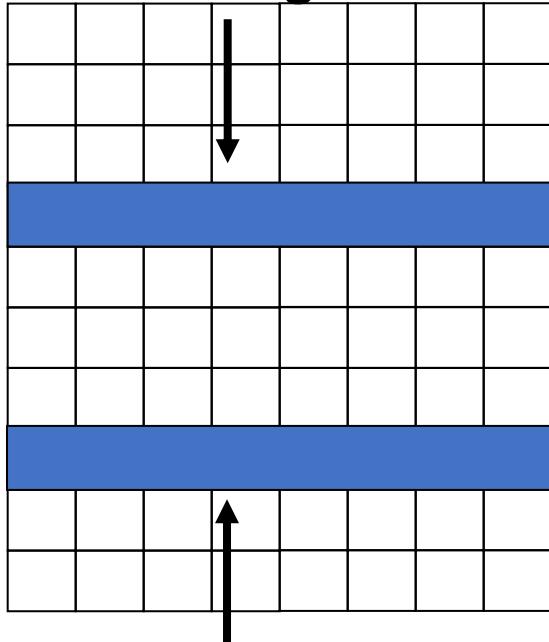
You never Need More Than The Previous Row To Compute the optimal score

You only need the matrix for the Trace-Back,

Or do you ????

# An Alignment in Linear Space

Forward Algorithm



Backward algorithm

$F(i,j)$ =Optimal score of  
0...i Vs 0...j

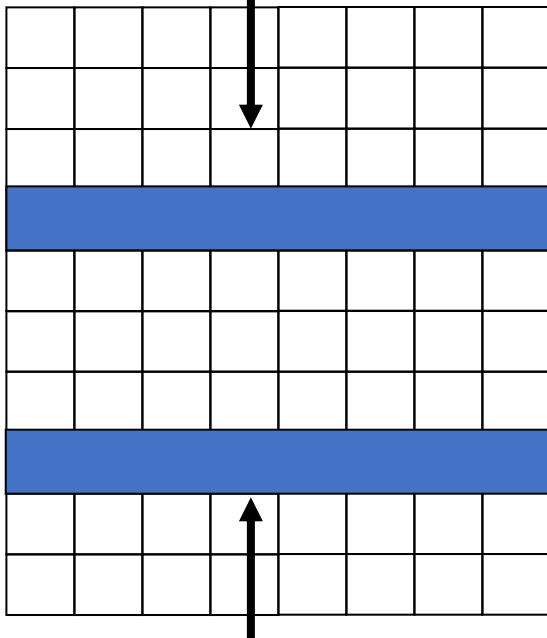
$B(i,j)$ =Optimal score of  
M...i Vs N...j

$B(i,j)+F(i,j)$ =  
Optimal score of the alignment that  
passes through pair i,j

Myers, E. W. & Miller, W. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–7 (1988).

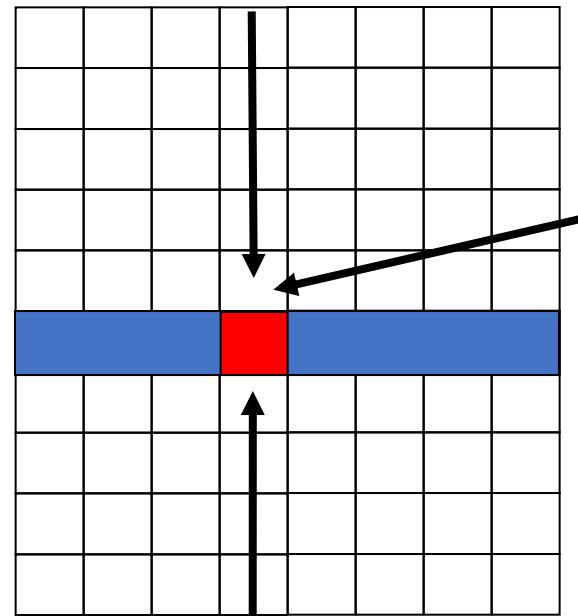
# An Alignment in Linear Space

Forward Algorithm



Backward algorithm

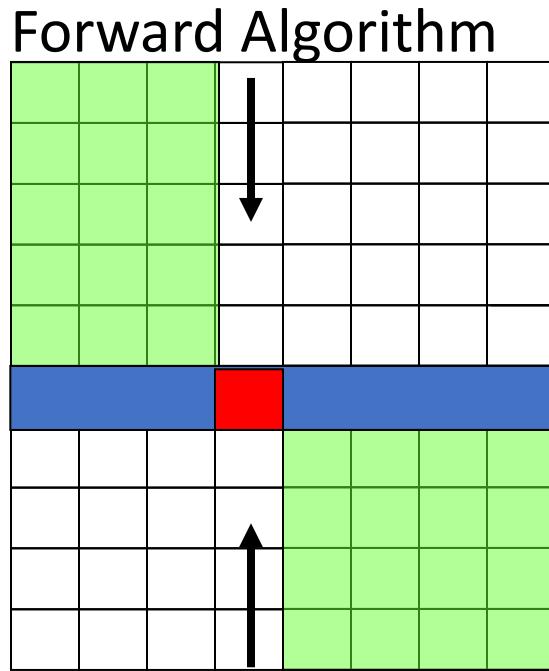
Forward Algorithm



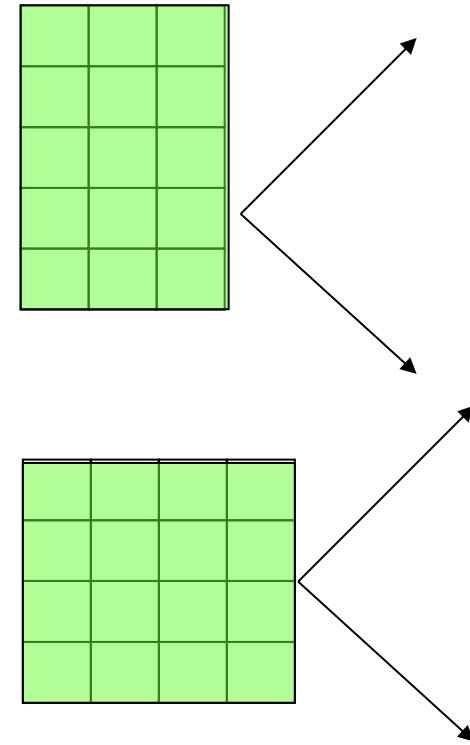
Backward algorithm

Optimal  $B(i,j)+F(i,j)$

# An Alignment in Linear Space



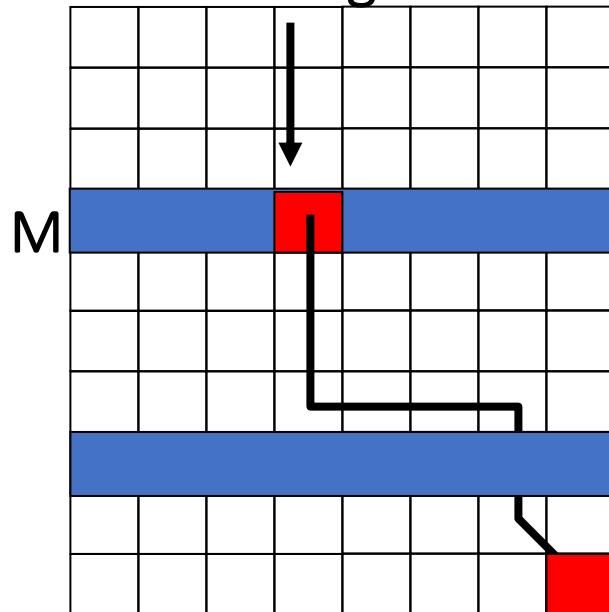
Backward algorithm



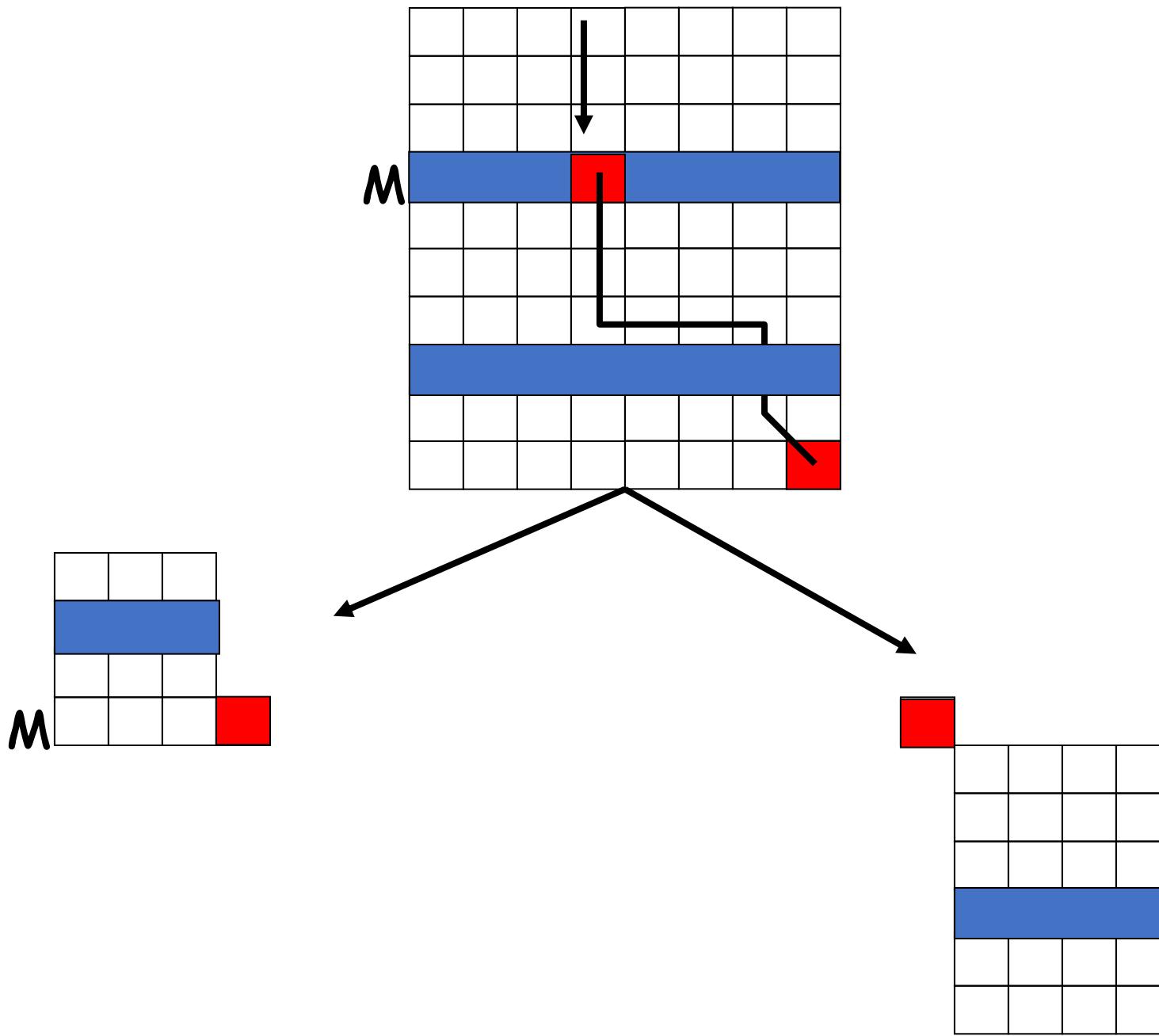
Recursive divide and conquer strategy:  
Myers and Miller (Durbin p35)

# A Forward-only Strategy (Durbin, p35)

## Forward Algorithm



- Keep Row  $M$  in memory
- Keep track of which Cell in Row  $M$  lead to the optimal score
- Divide on this cell



Adapted from Cedric Notredame

# Remember Not To Run Out of Memory

- A survey paper
  - Chao, K.-M., Hardison R. C. and Miller, W., 1994, Recent Developments in Linear-Space Alignment Methods: a Survey, *Journal of Computational Biology*, 1: 271-291.



趙坤茂 (Kun-Mao Chao)  
台大資工系

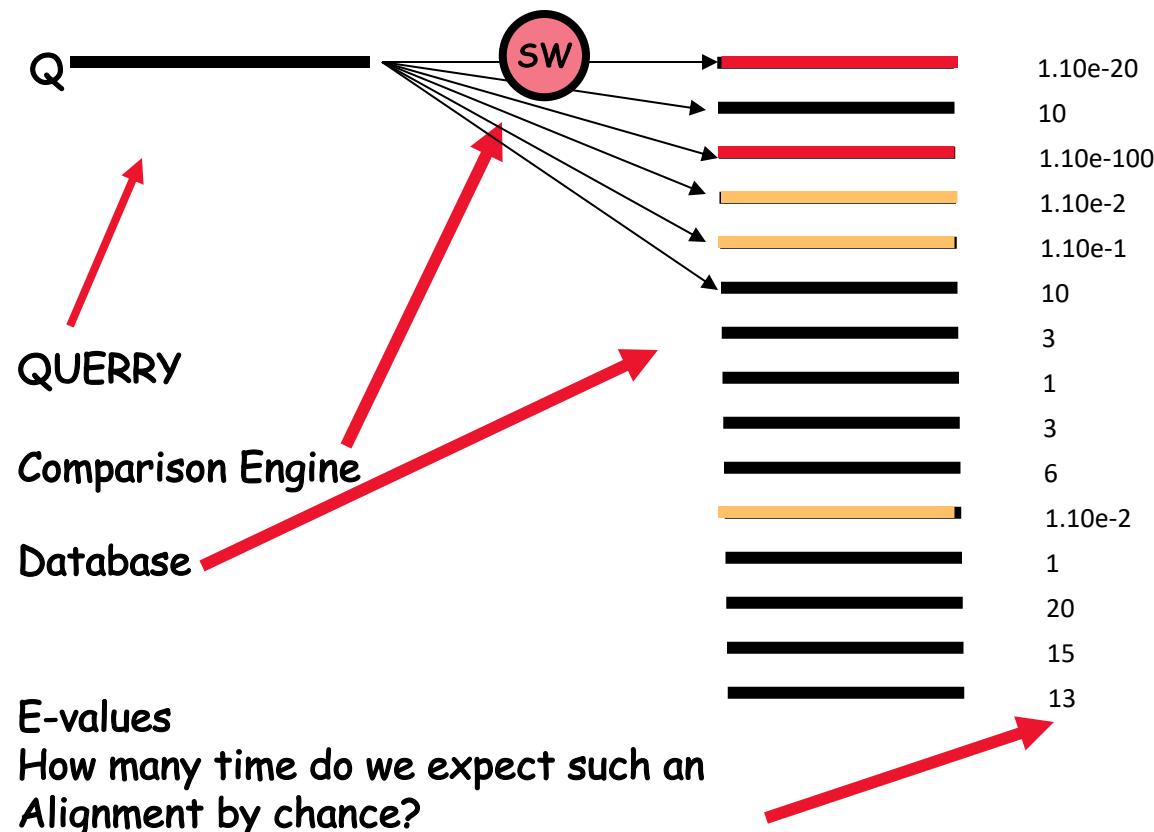
# Basic Local Alignment Search Tool (BLAST)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. J.  
*Mol. Biol.* **215**, 403–410(1990).  
Altschul, S. F. *et al.* *Nucleic Acids Res.* **25**, 3389–3402 (1997).

# Rapid, heuristic versions of Smith-Waterman

- Smith-Waterman (1981) is very rigorous and it is guaranteed to find an optimal alignment.
- But Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database).

# Database Search



Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215**, 403–410(1990).

Altschul, S. F. *et al.* *Nucleic Acids Res.* **25**, 3389–3402 (1997).

- The rapid expansion of genetic sequencing since Sanger's contribution has helped to boost the ranking of papers describing ways to analyse the sequences. A prime example is BLAST (Basic Local Alignment Search Tool), which for two decades has been a household name for biologists wanting to work out what genes and proteins do. Users simply have to open the program in a web browser and plug in a DNA, RNA or protein sequence. Within seconds, they will be shown related sequences from thousands of organisms — along with information about the function of those sequences and even links to relevant literature. So popular is BLAST that versions<sup>8,9</sup> of the program feature twice on the list, at spots 12 and 14.
- But owing to the vagaries of citation habits, BLAST has been bumped down the list by **Clustal** ...

BLAST is a Heuristic Smith and Waterman

three phases

1. Decide who will be compared
2. Check the most promising Hits
3. Compute the E-value of the most interesting Hits

# Phase 1-Decide who will be compared

- compile a list of word pairs ( $w=3$ ) above threshold T
- Example: for a human RBP query  
...FSG**TW**YA... (query word is in green)
- A list of words ( $w=3$ ) is:

FSG SGT GTW TWY WYA

YSG TGT ATW SWY WFA

FTG SVT GSW TWF WYS

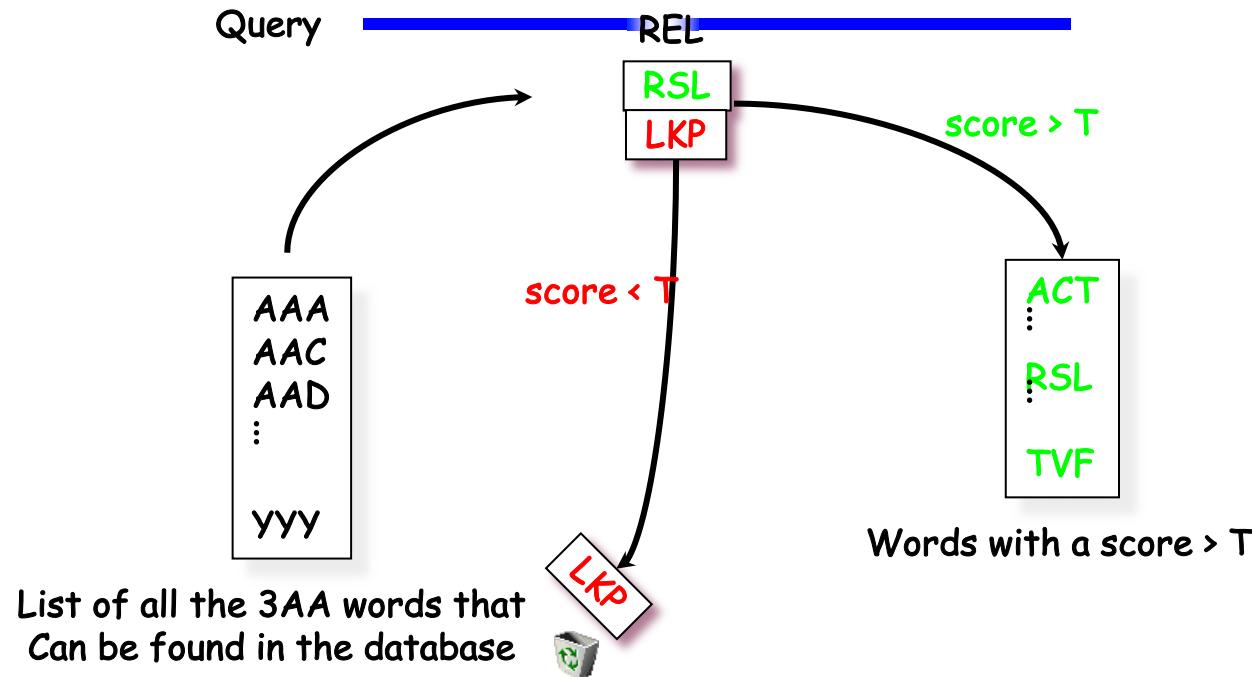
...

## Phase 1: compile a list of words (w=3)

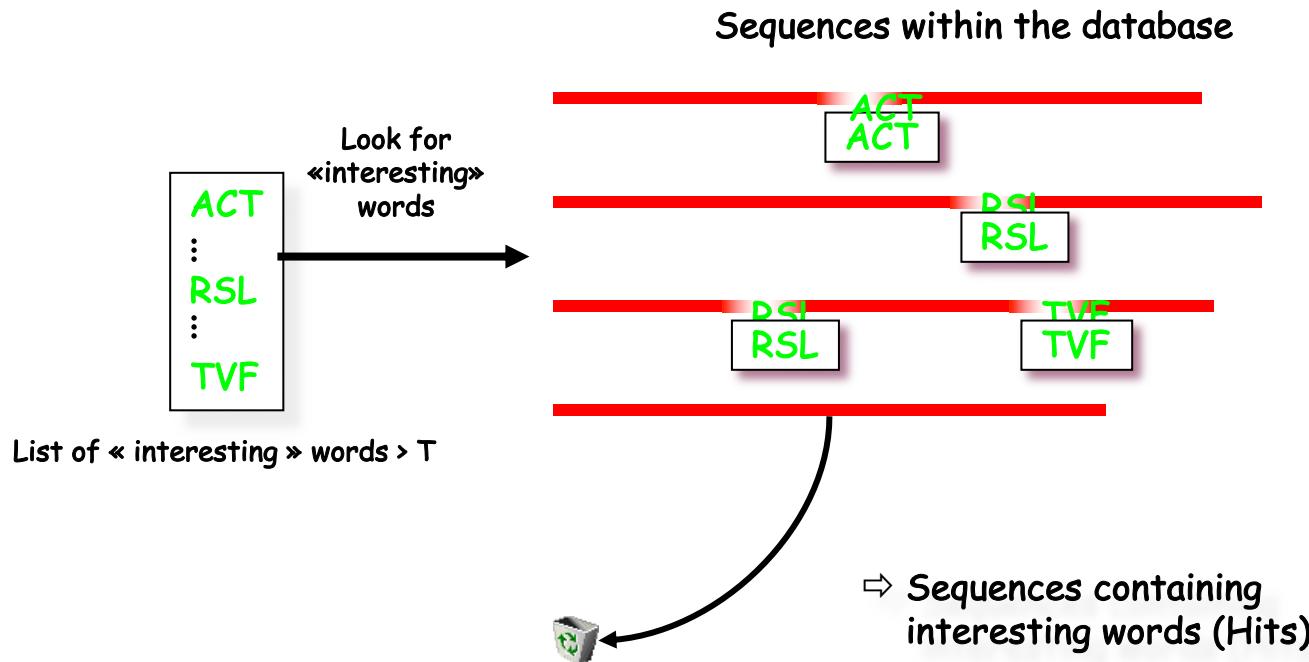
- The default value for BLASTP is 11.

GTW	6, 5, 11	22	
GSW	6, 1, 11	18	
ATW	0, 5, 11	16	
NTW	0, 5, 11	16	
(T=11)	GTY	6, 5, 2	13
	GNW		10
	GAW		9

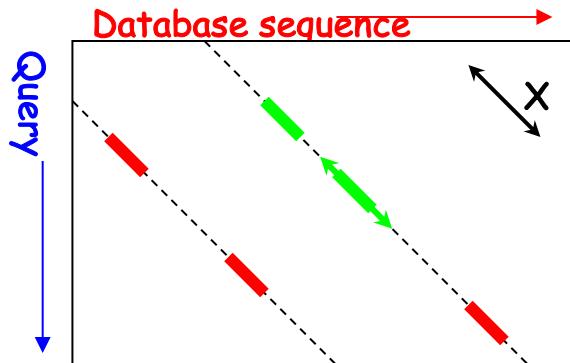
# Phase 1: finding the worthy words



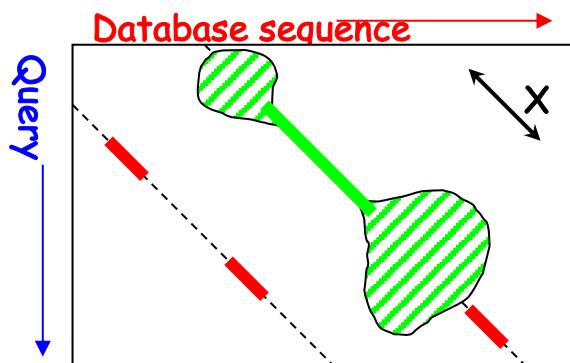
## Phase 2: Eliminate the database sequences that do not contain any interesting word



# Phase 3: Extension of the Hits



2 "Hits" on the same diagonal distant by less than X



Extension by limited Dynamic Programming

# Why Do We Need Multiple Sequence Alignment ?

# Sometimes Two Sequences Are Not Enough...

- The man with TWO watches NEVER knows the time



# Multiple sequence alignment: definition

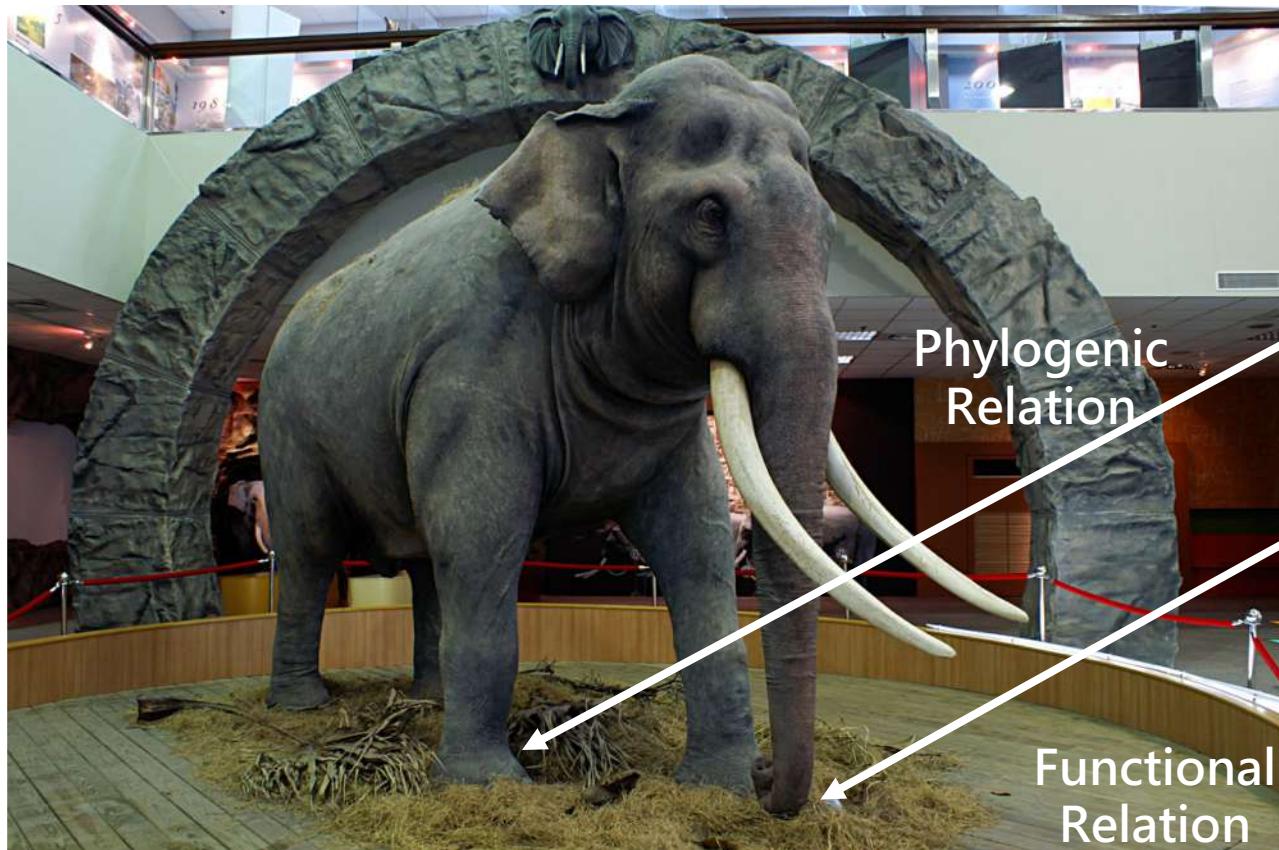
- a collection of three or more proteins (or nucleic acid)
- sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense

# What is A Multiple Sequence Alignment?

- Structural Criteria
  - Residues are arranged so that those playing a **similar role** end up in the **same column**.
- Evolution Criteria
  - Residues are arranged so that those having the **same ancestor** end up in the **same column**.

chite	---	ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGELWRGLKD
wheat	--DPNPKRAPS AFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE	
trybr	KKDSNAPKRAMTSF MFSSDFRS----KHSDL S-IVEMS KAAGAAW KELGP	
mouse	-----KPKRPRSAYNIYVSES FQ----EAKDDS-AQGKLKLVNEAWKNLSP	
	***. :: : .. . : . . * . * : *	

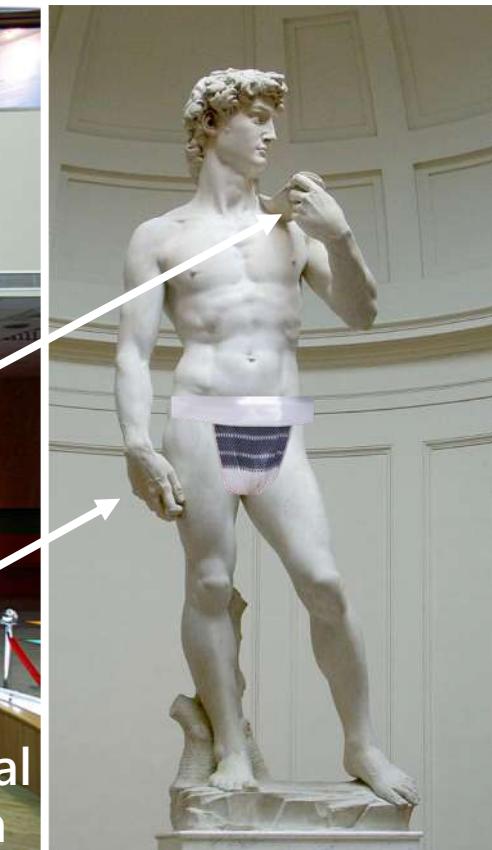
chite	AATAKQNYIRALQEYERN GG-
wheat	ANKLGEYNKAIAAYNKGESA
trybr	AEKDKERYKREM-----
mouse	AKDDRIRYDNEMKSWE EQMAE
	* : . * . :



Phylogenetic  
Relation

Functional  
Relation

[By peellden](#) - 自己的作品



[By Rico Heil \(User:Silmaril\)](#) - private photo

Adapted from Cedric Notredame

## Main Criteria for building a multiple sequence alignment

<i>Criterion</i>	<i>Meaning</i>
<b>Structure similarity</b>	Amino acids that play the same role in each structure are in the same column. Structure superposition programs are the only ones that use this criterion.
<b>Evolutionary similarity</b>	Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
<b>Functional similarity</b>	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
<b>Sequence similarity</b>	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

# MSA algorithms

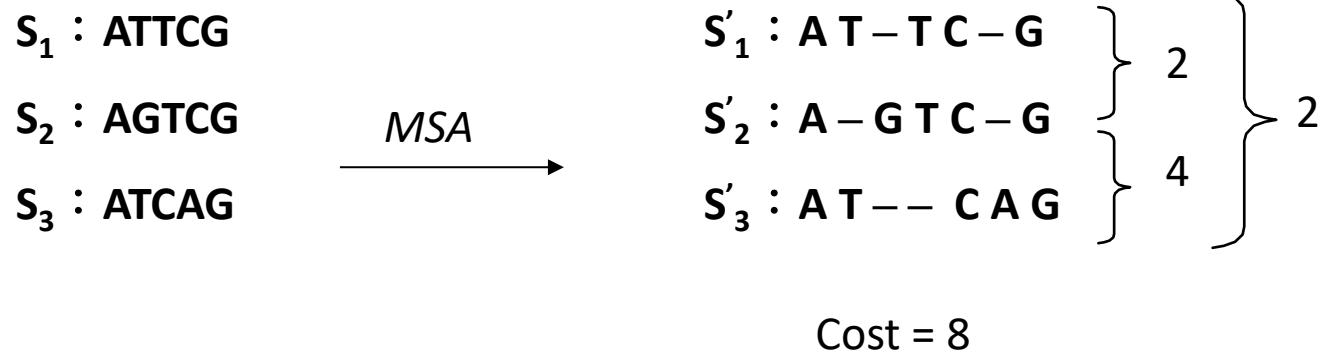
# The COMPUTATIONAL Problem

- A nice set of Sequences
- Substitution Matrix (Blosum)
- Gap Penalties
- An Evaluation Function
- An Alignment Algorithm

# Multiple Sequence Alignment

- Given  $s$  set of sequences => to find an alignment of the sequences such that some object function is minimized
- Scoring function
  - Sum of Pair (SP)
  - Tree Cost: MSA with tree cost will be called tree alignment.
  - Circular Sum(CS)

# Sum of Pair



# MSA with SP-Score: Complexity

- *J Comput Biol* 1994 Winter;1(4):337-48

**On the complexity of multiple sequence alignment.**

**Wang L. Jiang T.**

*McMaster University, Hamilton, Ontario, Canada.*

We study the computational complexity of two popular problems in multiple sequence alignment :

1. multiple alignment with SP-Score => NP-complete(non-metric)
2. multiple tree alignment => MAX SNP-hard

- *Theoretical Computer Science*;259 (2001) 63-79

**The complexity with Multiple sequence alignment with SP-score that is a metric**

*Paola Bonizzoni, Gianluca Della Vedova*

1. multiple alignment with SP-Score => NP-complete(metric)

# MSA with SP-Score : Approximation

- **Performance ratio of  $2-2/k$** 
  - D.Gusfilde,Efficient methods for multiple sequence alignment with guaranteed error bounds, *Bull. Math Bio.*,(1993)
- **Performance ratio of  $2-3/k$** 
  - P.Pevzner, Multiple alignment,communication cost, and graph matching, *SIAM J. Appl. Math.*,(1992)
- **Performance ratio of  $2-l/k$  (assembling  $l$ -way alignments,  $l \leq k$ )**
  - V.Bafna,E.L.Lawler and Pevzner,Approximation algorithms for multiple sequences alignment, *Theor. Comput. Sci.*,(1997)
- **Polynomial Time Approximation Scheme**
  - MSA within a constant band and allows only constant number of insertion and deletion gaps of arbitrary length per sequence on average
    - M. Li,B. Ma. And L. Wang, Near optimal alignment within a band in polynomial time, STOC 2000.

# MSA with SP-Score: Heuristics Algorithm

- Given
  - $k$  : # of Sequences
  - $n$  : Sequences of length
- Heuristics
  - D.F.Feng,R.F.Doolittle, Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25, 351-360., (1987)
  - S.F.Altschul,D.J.Lipman, Trees,star and mutiple biological sequence aligment, *SIAM J. Appl. Math.*,(1989)
  - D.J.lipman,S.F.Altschul, A tool for multiple sequences alignment, *Proc.Nat.Acad. Sci. U.S.A.*,(1989)
  - S.C. Chan,A.K.C. Wang,D.K.Y. Chiu, A survey of multiples sequences comparison methods, *Bull.Math Bio.*,(1992)

# Feng-Doolittle algorithm

D.F.Feng, R.F.Doolittle, Progressive sequence alignment as a  
prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25,  
351-360., (1987)

# Making An Alignment

- Any Exact Method would be **TOO SLOW**
  - We will use a Heuristic Algorithm.
  - **Progressive Alignment** Algorithm is the most Popular

# Multiple sequence alignment: methods

- Progressive methods: use a guide tree (related to a phylogenetic tree) to determine how to combine pairwise alignments one by one to create a multiple alignment.
- Examples
  - CLUSTALW
  - MUSCLE
- - Greedy Heuristic (No Guaranty)
- + Fast

# Progressive Alignment

- Feng and Dolittle, 1988; Taylor 1989



# Feng-Doolittle MSA occurs in 3 stages

- Do a set of global pairwise alignments
  - Needleman and Wunsch's dynamic programming algorithm
- Create a guide tree
- Progressively align the sequences

## generate global pairwise alignments (Progressive 1/3)

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
=====						
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43
=====						

best

score

# Number of pairwise alignments needed

- For  $n$  sequences,  $\frac{n \times n-1}{2}$
- For 5 sequences,  $(5 \times 4) / 2 = 10$
- For 200 sequences,  $(200 \times 199) / 2 = 19,900$

# guide tree (Progressive 2/3)

- Convert similarity scores to distance scores
- A tree shows the distance between objects
- Use UPGMA (defined in the phylogeny chapter)
- ClustalW provides a syntax to describe the tree

## (a) Stage 1: series of pairwise alignments (closely related globin proteins)

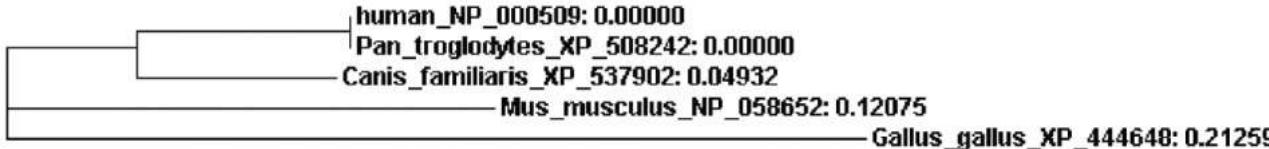
SeqA	Name	Length	SeqB	Name	Length	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100.0
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89.8
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80.27
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69.39
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89.8
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80.27
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69.39
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78.91
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71.43
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66.67

## (b) Stage 2: create a guide tree (calculated from a distance matrix)

```

(
(
(
    human_NP_000509:0.00000,
    Pan_troglodytes_XP_508242:0.00000
    :0.05272,
    Canis_familiaris_XP_537902:0.04932)
    :0.03231,
    Mus_musculus_NP_058652:0.12075,
    Gallus_gallus_XP_444648:0.21259);

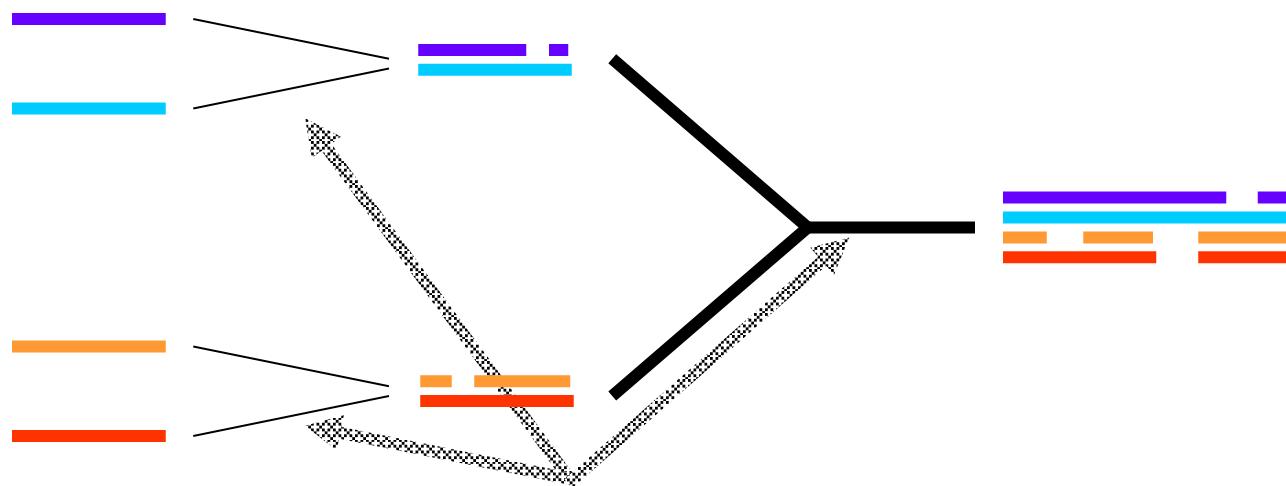
```



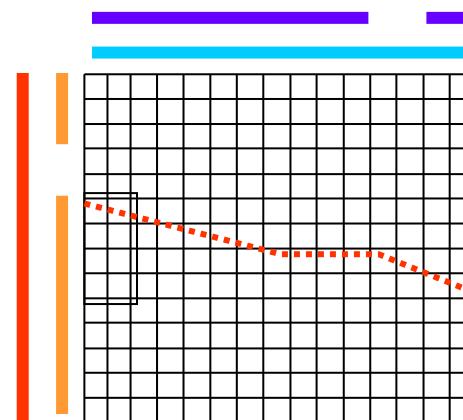
# progressive alignment(Progressive 3/3)

- Make a MSA based on the order in the guide tree
- Start with the two most closely related sequences
- Then add the next closest sequence
- Continue until all sequences are added to the MSA
- Rule: “once a gap, always a gap.”

# Progressive Alignment

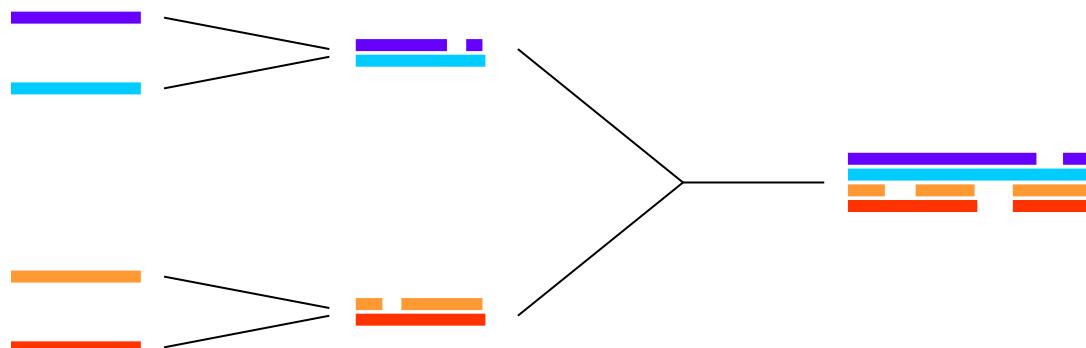


Dynamic Programming Using A Substitution Matrix



# Progressive Alignment

- Depends on the CHOICE of the sequences
- Depends on the ORDER of the sequences (Tree)
- Depends on the PARAMETERS
  - Substitution Matrix
  - Penalties (Gop, Gep)
  - Sequence Weight
  - Tree making Algorithm



# Why “once a gap, always a gap”?

- There are many possible ways to make a MSA
- Where gaps are added is a critical question
- Gaps are often added to the first two (closest) sequences
- To change the initial gap choices later on would be to give more weight to distantly related sequences
- To maintain the initial gap choices is to trust that those gaps are most believable

# ClustalW

Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–80 (1994).

2008~2013 PhD La Caxia fellowship  
@ The Centre for Genomic Regulation  
Barcelona, Spain  
Funded by la Caxia fellowship

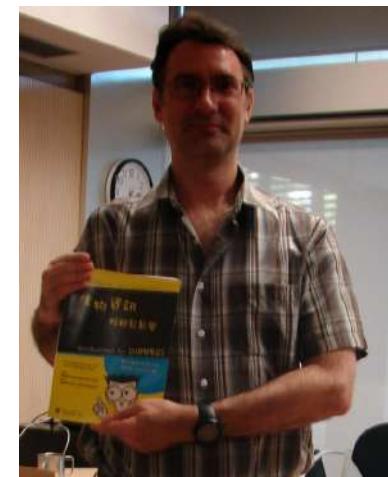
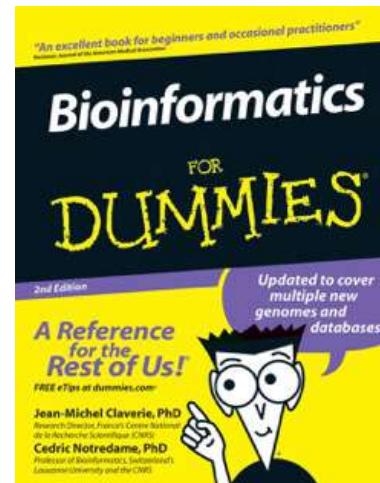


Dr. Cedric Notredame

T-Coffee: A novel method for fast and accurate multiple sequence alignment  
C Notredame, DG Higgins, J Heringa  
Journal of molecular biology 302 (1), 205-217

5606

2000



TITLE

CITED BY

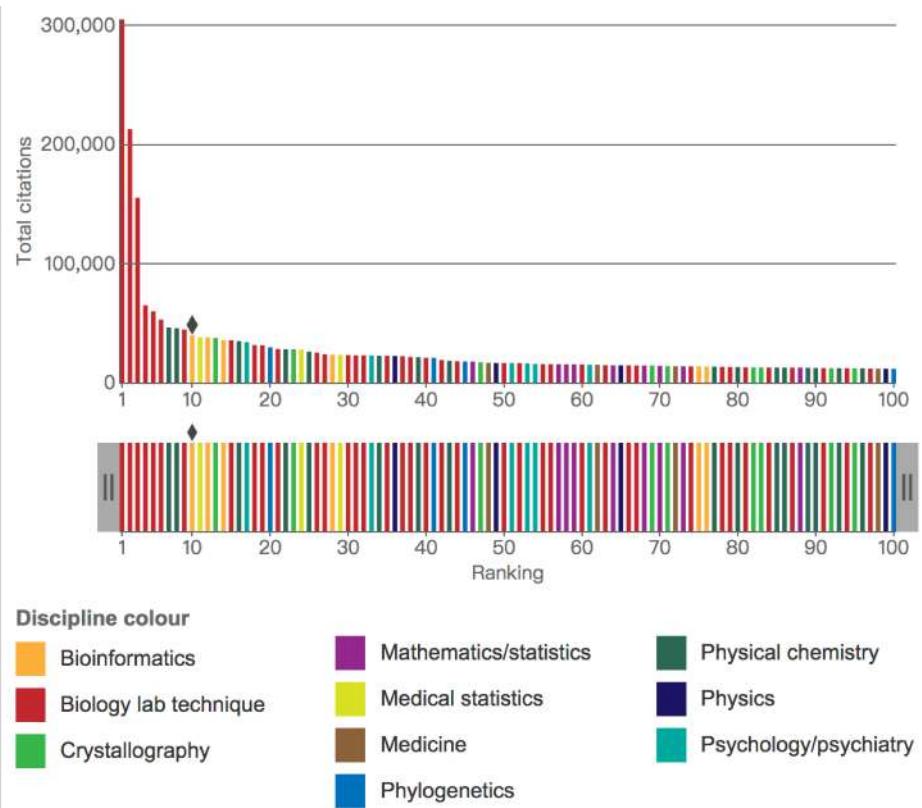
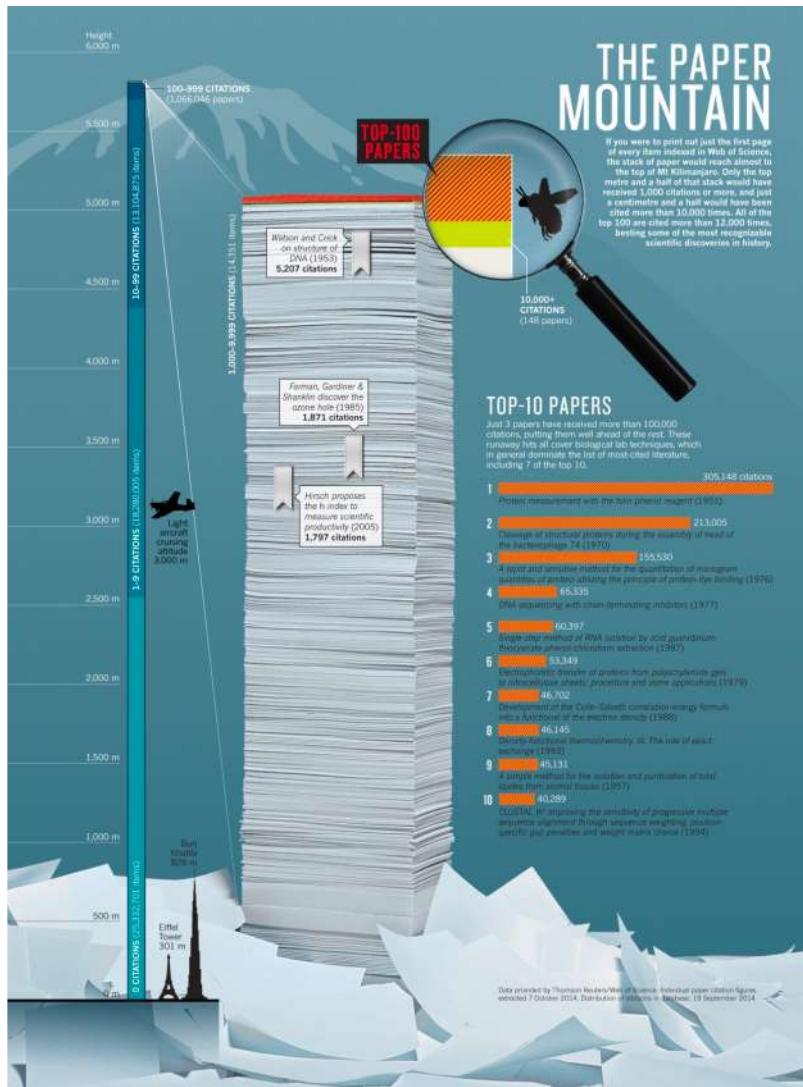
YEAR

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice  
JD Thompson, DG Higgins, TJ Gibson  
Nucleic acids research 22 (22), 4673

58342

1994

# The top 100 papers



## The top 100 papers



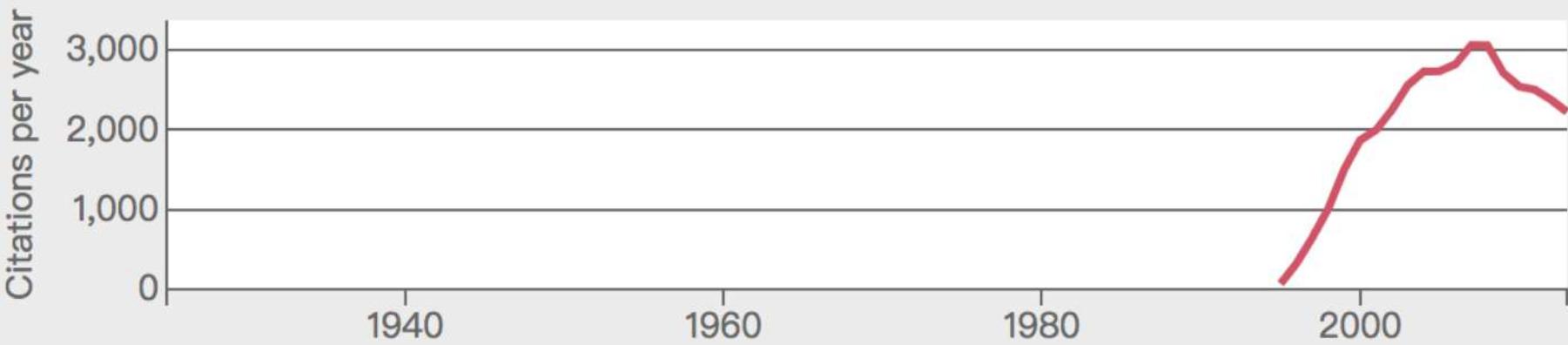
Click through to explore the Web of Science's all-time top-cited papers. (Data provided by Thomson Reuters, extracted on 7 October 2014).

Rank: 10 Citations: 40,289

Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.

Thompson, J. D., Higgins, D. G. & Gibson, T. J

*Nucleic Acids Res.* **22**, 4673–4680 (1994).



# ClustalW

But owing to the vagaries of citation habits, BLAST has been bumped down the list by Clustal, a complementary programme for aligning multiple sequences at once. Clustal allows researchers to describe the evolutionary relationships between sequences from different organisms, to find matches among seemingly unrelated sequences and to predict how a change at a specific point in a gene or protein might affect its function. A 1994 paper describing ClustalW, a user-friendly version of the software, is currently number 10 on the list. A 1997 paper on a later version called ClustalX is number 28.

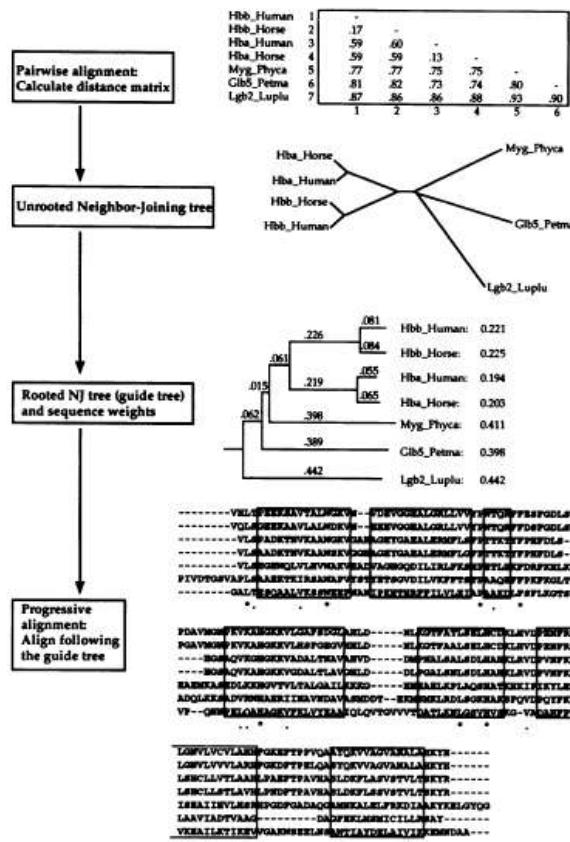
# ClustalW

The team that developed ClustalW, at the European Molecular Biology Laboratory in Heidelberg, Germany, had created the program to work on a personal computer, rather than a mainframe. But the software was transformed when Julie Thompson, a computer scientist from the private sector, joined the lab in 1991. “It was a program written by biologists; I’m trying to find a nice way to say that,” says Thompson, who is now at the Institute of Genetics and Molecular and Cellular Biology in Strasbourg, France. Thompson rewrote the program to ready it for the volume and complexity of the genome data being generated at the time, while also making it easier to use.

The teams behind BLAST and Clustal are competitive about the ranking of their papers. It is a friendly sort of competition, however, says Des Higgins, a biologist at University College Dublin, and a member of the Clustal team. “BLAST was a game-changer, and they’ve earned every citation that they get.”

# Additional features of ClustalW improve its ability to generate accurate MSAs

- Individual weights are assigned to sequences; very closely related sequences are given less weight, while distantly related sequences are given more weight
- Scoring matrices are varied dependent on the presence of conserved or divergent sequences, e.g.:
  - PAM20 80-100% id
  - PAM60 60-80% id
  - PAM120 40-60% id
  - PAM350 0-40% id
- Residue-specific gap penalties are applied



**Figure 1.** The basic progressive alignment procedure, illustrated using a set of 7 globins of known tertiary structure. The sequence names are from Swiss Prot (38): Hba\_Horse: horse  $\alpha$ -globin; Hba\_Human: human  $\alpha$ -globin; Hbb\_Horse: horse  $\beta$ -globin; Hbb\_Human: human  $\beta$ -globin; Myg\_Phyc: sperm whale myoglobin; Glb5\_Petma: lamprey cyanohaemoglobin; Lgb2\_Luplu: lupin leghaemoglobin. In the distance matrix, the mean number of differences per residue is given. The unrooted tree shows all branch lengths drawn to scale. In the rooted tree, all branch lengths (mean number of differences per residue along each branch) are given as well as weights for each sequence. In the multiple alignment, the approximate positions of the 7  $\alpha$ -helices common to all 7 proteins are shown. This alignment was derived using CLUSTAL W with default parameters and the PAM (3) series of weight matrices.

In Figure 1 we give the 7x7 distance matrix between the 7 globin sequences calculated using the full dynamic programming method.

#### The guide tree

The trees used to guide the final multiple alignment process are calculated from the distance matrix of step 1 using the Neighbour-Joining method (21). This produces unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed by a 'mid-point' method (15) at a position where the means of the branch lengths on either side of the root are equal. These trees are also used to derive a weight for each sequence (15). The weights are dependent upon the distance from the root of the tree but sequences which have a common branch with other sequences share the weight derived from the shared branch. In the example in Figure 1, the leghaemoglobin (Lgb2\_Luplu) gets a weight of 0.442, which is equal to the length of the branch from the root to it. The human  $\beta$ -globin (Hbb\_Human) gets a weight consisting of the length of the branch leading to it that is not shared with any other sequences (0.081) plus half the length of the branch shared with the horse  $\beta$ -globin (0.226/2) plus one quarter the length of the branch shared by all four haemoglobins (0.061/4) plus one fifth the branch shared between the haemoglobins and myoglobin (0.015/5) plus one sixth the branch leading to all the vertebrate globins (0.062). This sums to a total of 0.221. In contrast, in the normal progressive alignment algorithm, all sequences would be equally weighted. The rooted tree with branch lengths and sequence weights for the 7 globins is given in Figure 1.

#### Progressive alignment

The basic procedure at this stage is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching order in the guide tree. You proceed from the tips of the rooted tree towards the root. In the globin example in Figure 1 you align the sequences in the following order: human vs. horse  $\beta$ -globin; human vs. horse  $\alpha$ -globin; the 2  $\alpha$ -globins vs. the 2  $\beta$ -globins; the myoglobin vs. the haemoglobins; the cyanohaemoglobin vs. the haemoglobins plus myoglobin; the leghaemoglobin vs. all the rest. At each stage a full dynamic programming (26,27) algorithm is used with a residue weight matrix and penalties for opening and extending gaps. Each step consists of aligning two existing alignments or sequences. Gaps that are present in older alignments remain fixed. In the basic algorithm, new gaps that are introduced at each stage

Thompson et al. (1994) for an explanation of the three stages of progressive alignment implemented in ClustalW

Pairwise alignment:

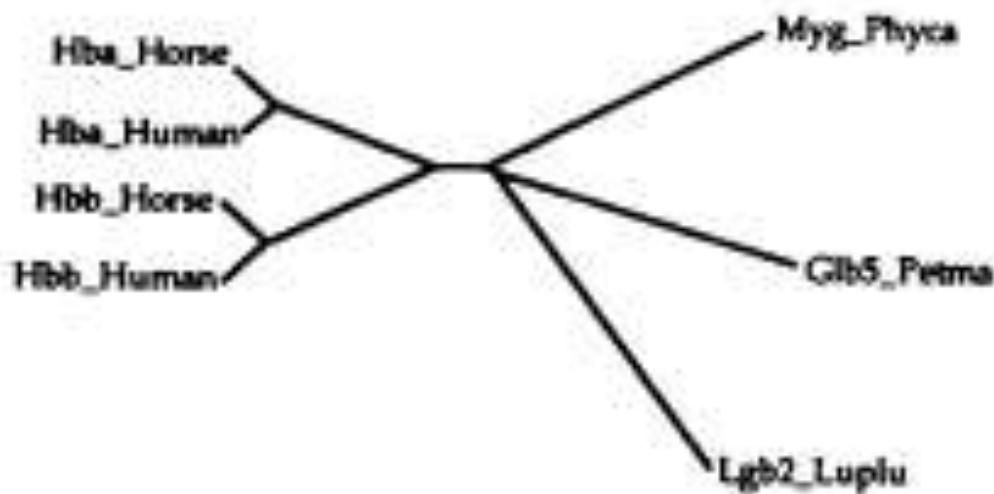
Calculate distance matrix

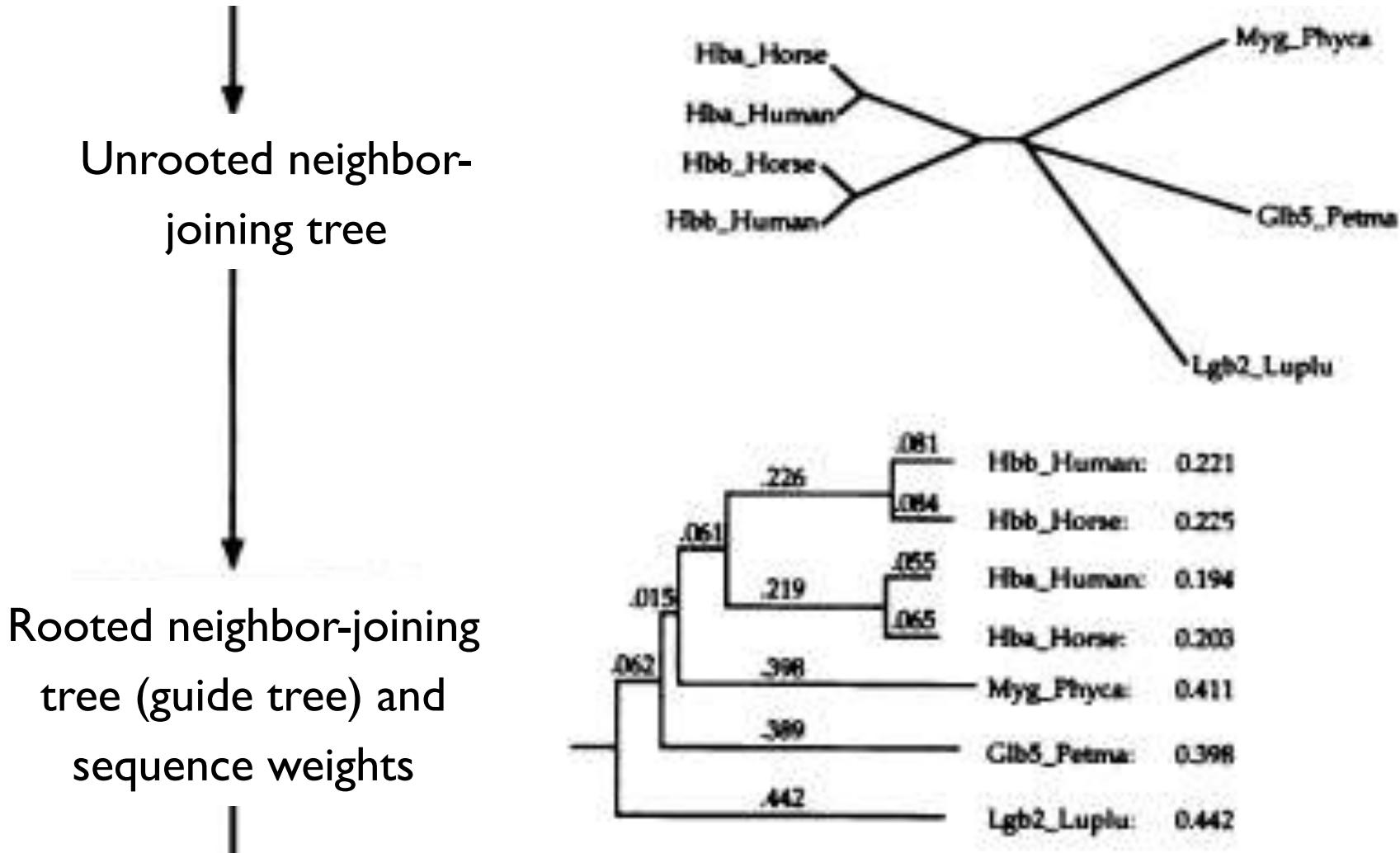


Unrooted neighbor-joining tree

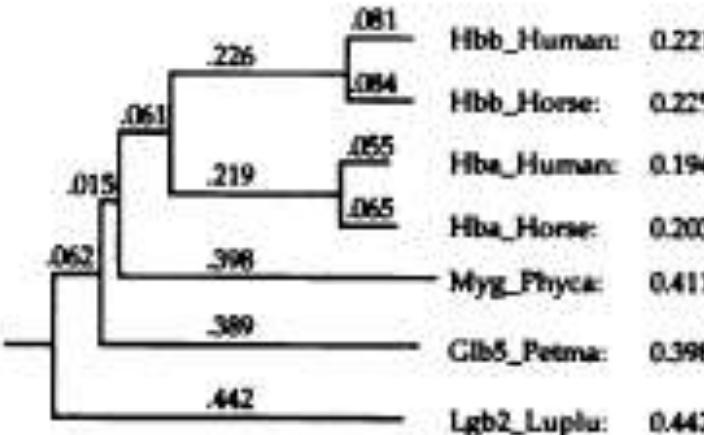


Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phycs	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
	1	2	3	4	5	6	





# Rooted neighbor-joining tree (guide tree) and sequence weights



PDAVMQSPFVKVAKKQKIVYGLP/FLAVSHLD-----NLKQTVTALESLCDELVDPVFIL  
PDAVMQSPFVKVAKKQKIVLSP/FLAVSHLD-----NLKQTVTALESLCDELVDPVFIL  
----BQDQVTKRQKQKIVADAL/THAVAVVD-----IMPHALALSDLMAN/ELVDPVFIL  
----BQDQVTKRQKQKIVGDL/FLAVSHLD-----DLPGALMLSDLMAN/ELVDPVFIL  
EAEHQLAEDLKEEKVTVLTLAGATLKEED-----HEKAELKPLAQHATENK/ELVDPVFIL  
ADQQLKKADVNNHEAMRI/THAVDAVJLMQDT-----EKMMEMLADLQKHAJSPQVDPQIFKV  
VP-----QWPKLQASDMDKTYLTTAEMQLOVTPGIVVVTGATL/ELVDPVFIL-----VPAHNTT

```

LQWVLLVCVLAIRSPKETPPFWQASITDQVVVAGVAAALDQKTYE
LQWVLLVVVLAKRSPKDTPTFELQASPTQKVVVAGVAAALAKTYE
LNSCILLNTLAAILPAEPTPAVHNMSLDQFLASVSTVLTSEKTYR
LNSCILLSTLAVHLPSDFTPAVHNMSLDQFLASVSTVLTSEKTYR
IIRRATIIVLRLRSPGDFGADAOQAMKALELFRDIAARYKEGYO
LAAVIADTVLQ-----DACPFKLMNMICILLRQAY
VKHAIATTTIIVVGAJMMHEELNWTIATDGLATVIGGMDAA

```

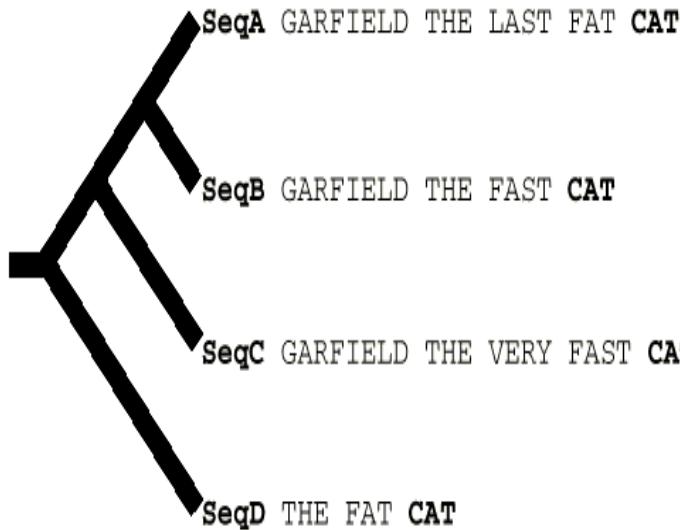
Digitized by srujanika@gmail.com

Credit by B&FG 3e, Jonathan Pevsner

# Progressive Alignment: When Does It Work

- Works Well When Phylogeny is Dense
- No outlayer Sequence

# Progressive Alignment When Doesn't It Work b



CLUSTALW (Score=20, Gop=-1, Gep=0, M=1)

SeqA GARFIELD THE LAST FA-T CAT

SeqB GARFIELD THE FAST CA-T ---

SeqC GARFIELD THE VERY FAST CAT

SeqD ----- THE ---- FA-T CAT

CORRECT (Score=24)

SeqA GARFIELD THE LAST FA-T CAT

SeqB GARFIELD THE FAST ---- CAT

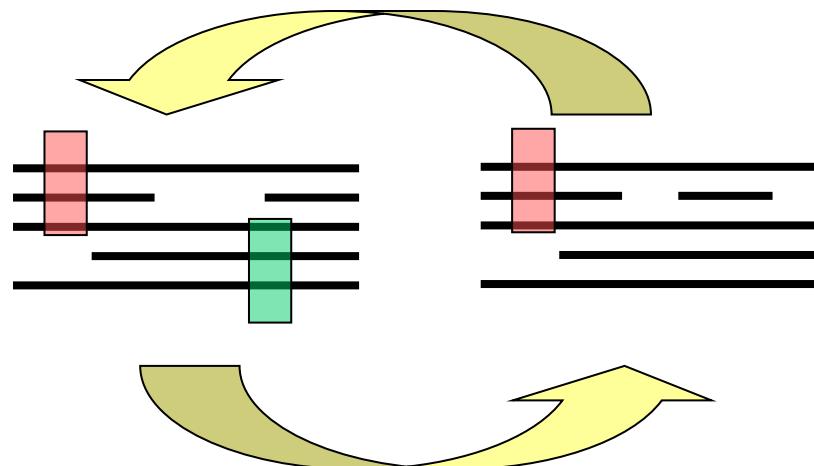
SeqC GARFIELD THE VERY FAST CAT

SeqD ----- THE ---- FA-T CAT

# Iterative approaches

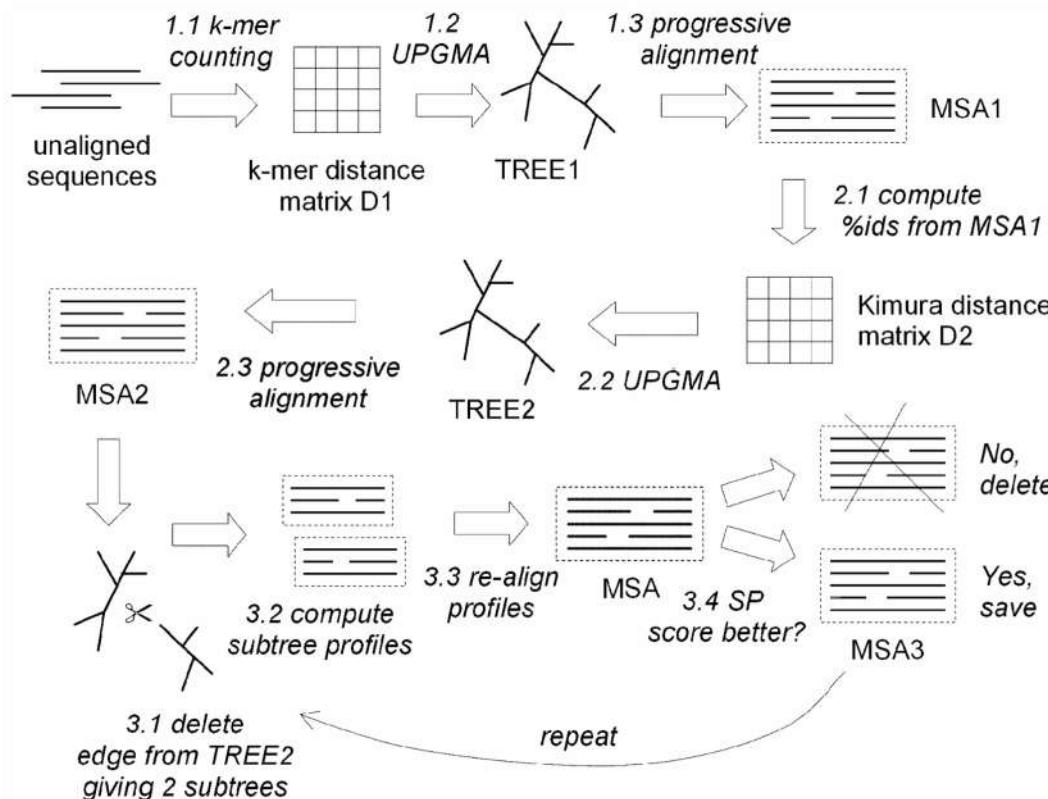
# Iterative methods

- compute a sub-optimal solution and keep modifying that intelligently using dynamic programming or other methods until the solution converges.
- MUSCLE, MAFFT, HMMs, HMMER, SAM,, IterAlign, Praline
- +: Good Profile Generators
- -: Slow, Sometimes Inaccurate



# Muscle

- Edgar, R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics* 5, 1–19 (2004).
- Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792–1797 (2004)

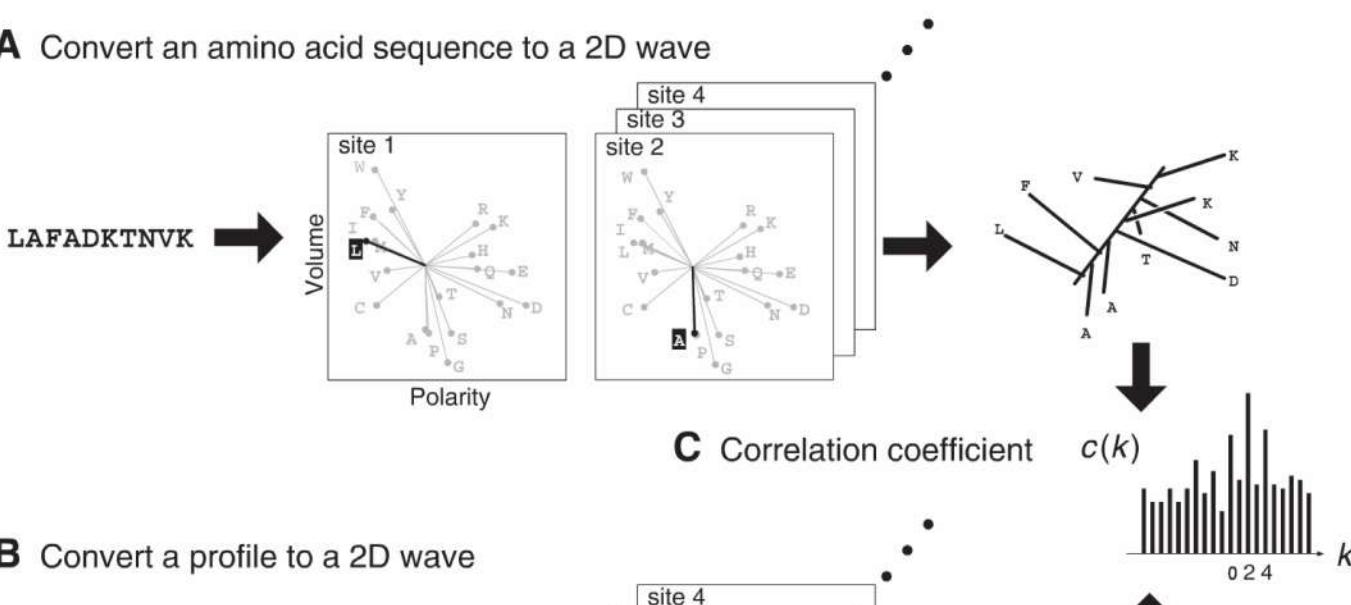


Adapted from Cedric Notredame

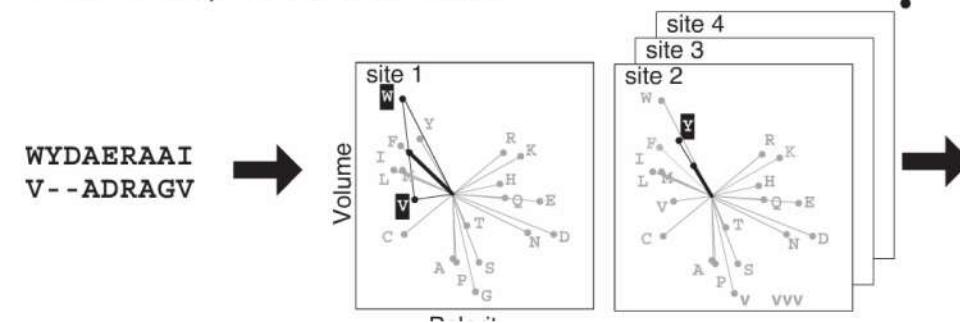
# MAFFT : Fast Fourier Transforme

- Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–66 (2002).
- Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* **9**, 286–98 (2008).

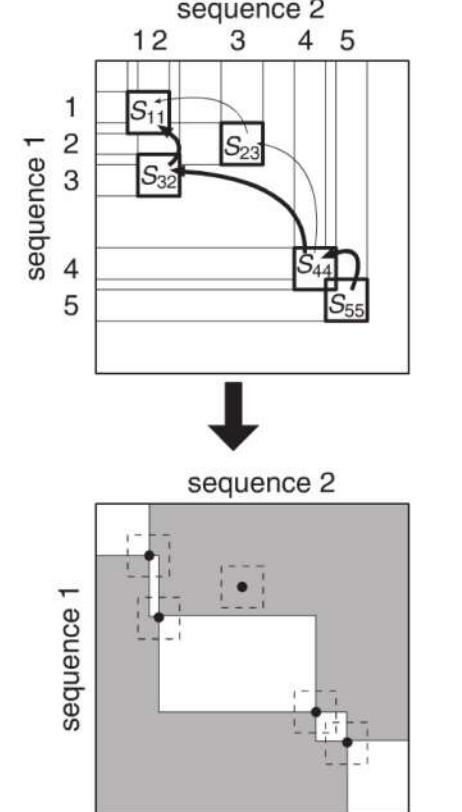
**A** Convert an amino acid sequence to a 2D wave



**B** Convert a profile to a 2D wave



**D** Restrict the area of the DP matrix

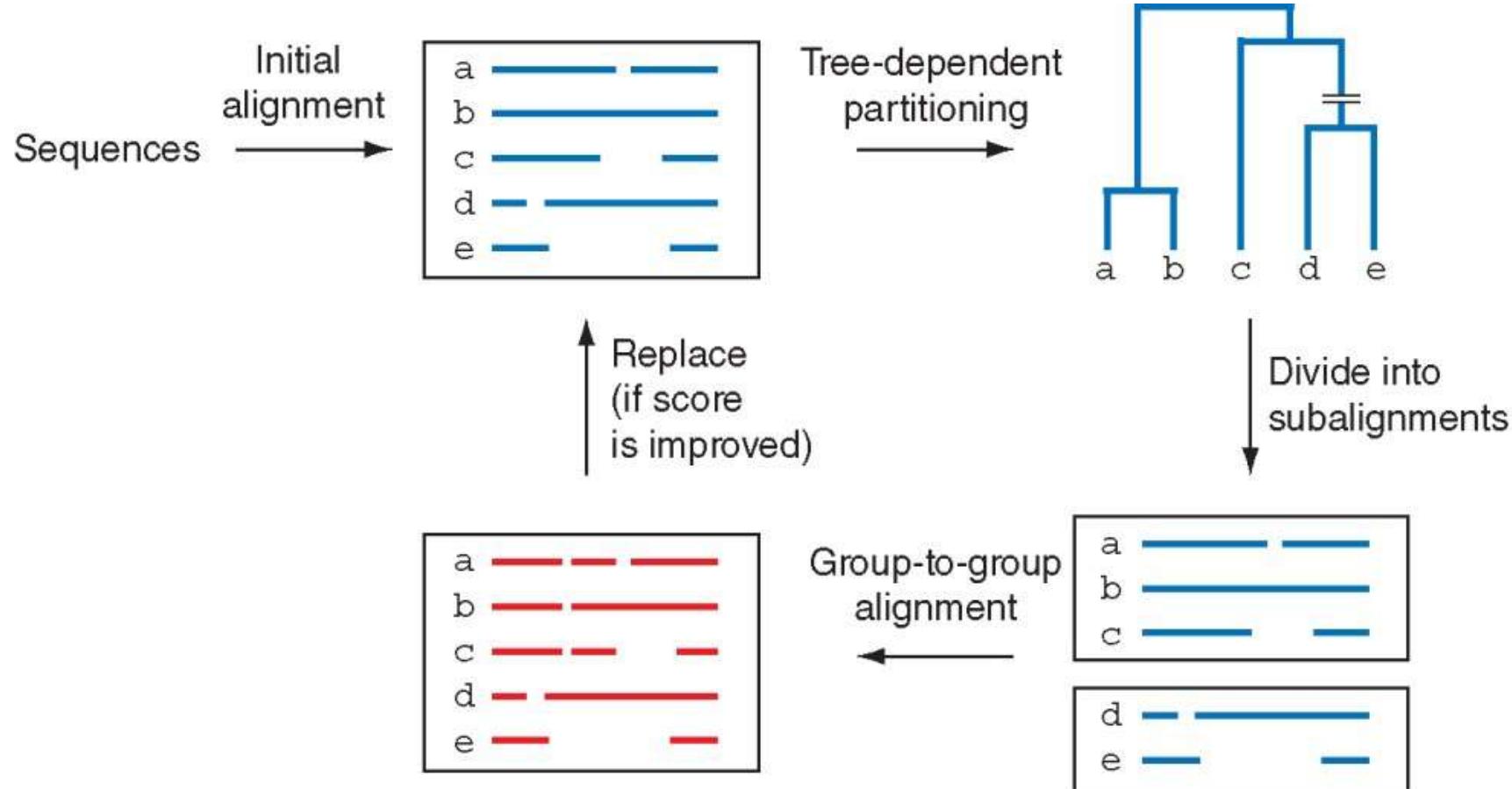


Adapted from Cedric Notredame

# MAFFT

- Uses Fast Fourier Transform to speed up profile alignment
- Uses fast two-stage method for building alignments using  $k$ -mer frequencies
- Offers many different scoring and aligning techniques
- One of the more accurate programs available
- Available as standalone or web interface
- Many output formats, including interactive phylogenetic trees

# Iterative method of MAFFT



# Iterative approaches: MAFFT

**MAFFT version 6**  
Multiple alignment program for amino acid or nucleotide sequences

[Download version](#)  
[Mac OS X](#)  
[Windows](#)  
[Linux](#)  
[Source](#)  
[Usage](#)

**Online version**  
[Alignment](#)  
[Phylogeny](#)  
[Merits and limitations](#)  
[Algorithms](#)  
[Tips](#)  
[Aligning large data](#)  
[Mafft-homologs](#)  
[Benchmarks](#)  
[Feedback](#)

Contact address has changed!!  
kkatoh@kuicr.kyoto-u.ac.jp  
katoh@bioreg.kyushu-u.ac.jp

Multiple sequence alignment and NJ / UPGMA phylogeny

**Input:**  
Paste protein or DNA sequences in fasta format. [Example](#)

```
>gi|55743122|ref|NP_006735.2| retinol-binding protein 4, plasma precursor  
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGITWYAMAKKDPEGLFLQDNIVAEFSVDETGQ  
MSATAKGRRVLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDHIVTDYAVQYSCLRLNLDGTCADSYSFVSRDPNGLPEAQKTVVRQRQEELCLRQYRLIVHNGYCQSGRSERNLL  
>gi|12843160|dbj|BAB25881.1| unnamed protein product [Mus musculus]  
MEWWVALVLLAALCGGSERAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGH  
MSATAKGRRVLLSNWEVACDMVGTFTDTEDPAKFKMKYWGVASFLQKGNDHIVTDYAVQYSCLRLNLDGTCADSYSFVSRDPNGLPETRRLVRQRQEELCLRQYRLIVHNGYCQSGRSERNLL  
>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]  
MVMILLILSALAGLFGAAECQAFHLCKCPNPPVQENFDVNKYLGRWEIEKIPTTFENGRCIQANYSIMEN  
NGKIKVLNLQELRADGTVNQIEGEATPVNLTEPAKLEVKFSTWMPMSAFYWIATDYENYALVYSCTCIIQL  
FIIVDFAWTILARNPNLPPETVDSLKNILTTSNNIDVKKMTVTQVNCPLS
```

or upload a file:  [Browse...](#)

[Use structural alignment\(s\)](#)

**Output order:**  
 Same as input  
 Aligned

**Notify when finished** (optional; recommended when submitting large data):  
Email address:

[Submit](#) [Reset](#)

[Advanced settings](#)

Has about 1000 advanced settings!

# Consistency-based approaches

# Multiple sequence alignment: consistency

- Consistency-based algorithms: generally use a database of both local high-scoring alignments and long-range global alignments to create a final alignment
- These are very powerful and very accurate methods
- Examples: T-Coffee, Prrp, DiAlign, ProbCons

<http://tcoffee.org>



- Make a MSA
- MSA w. structural data
- Compare MSA methods
- Make an RNA MSA
- Combine MSA methods
- Consistency-based
- Structure-based

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

#### T-Coffee Server

Quick links to the most popular T-Coffee modes:



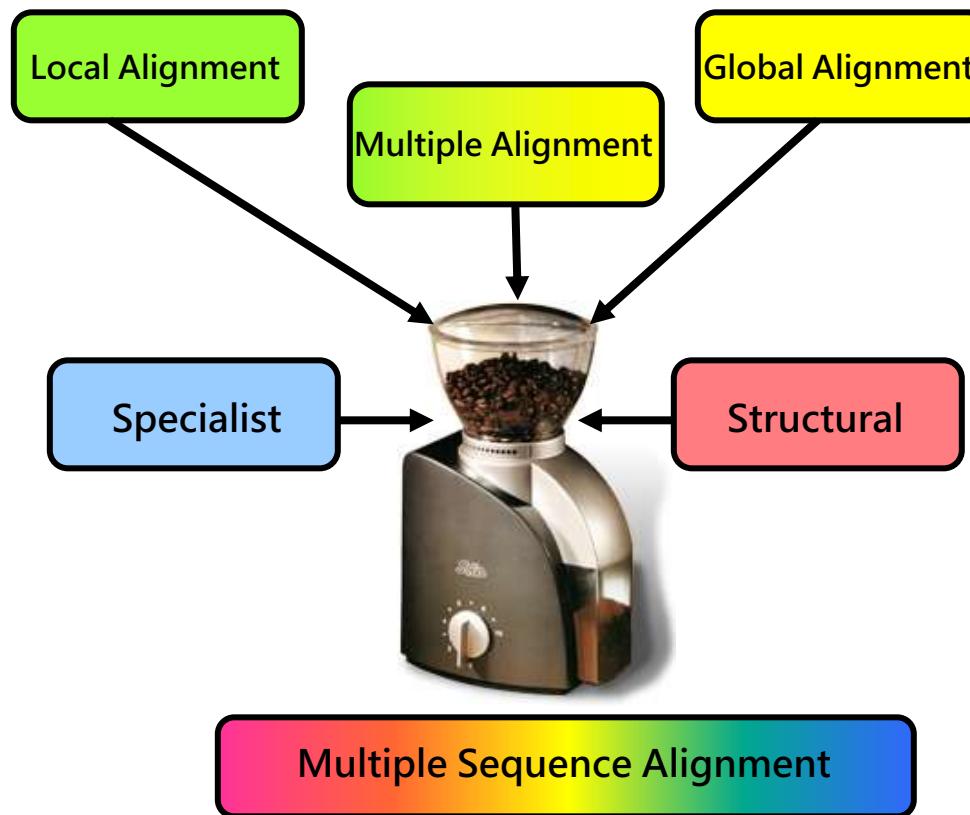
#### Other T-Coffee links

[Documentation](#)

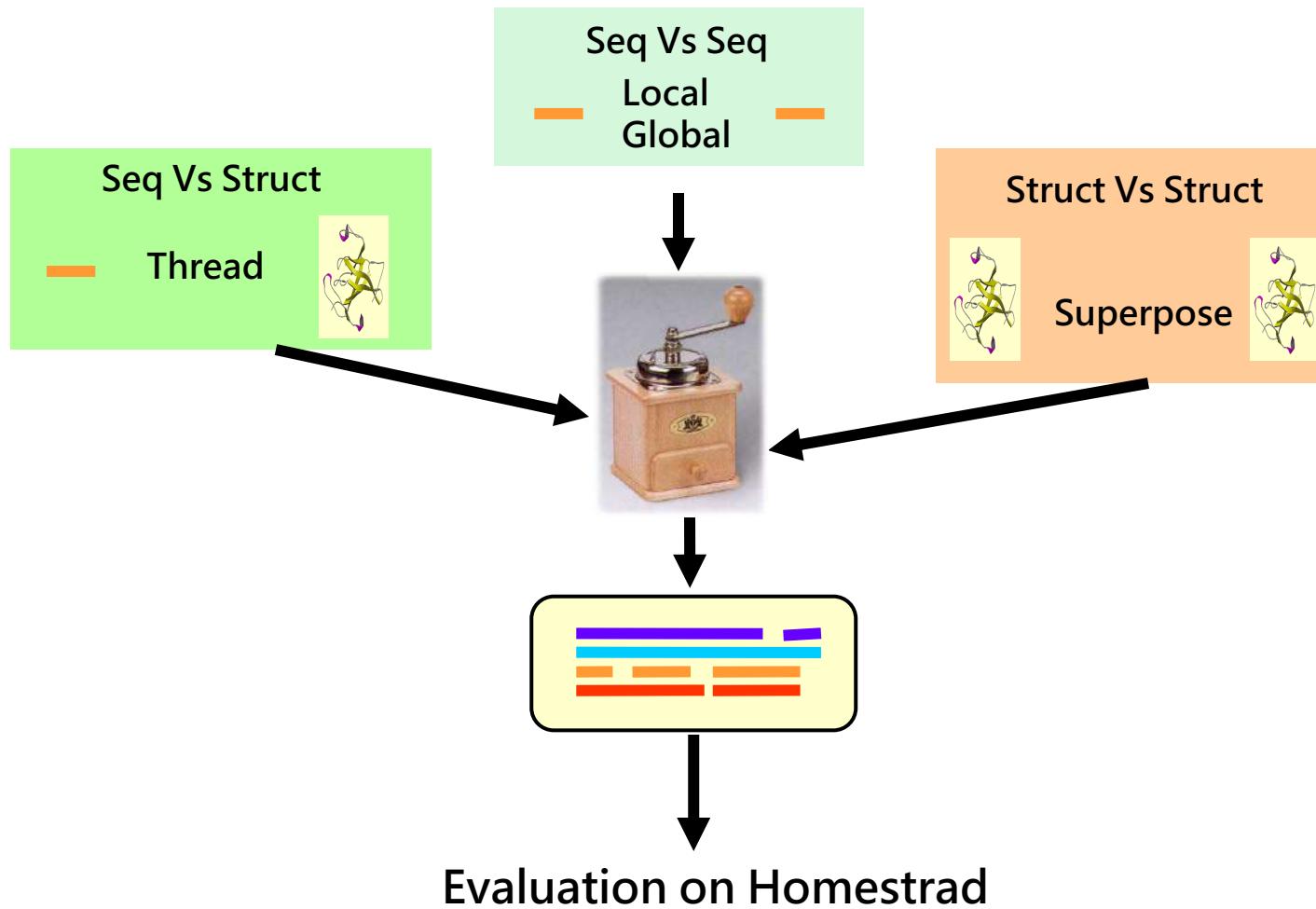
[Downloads](#)

[Support & discussion group](#)

# Mixing Heterogenous Data With T-Coffee



# Mixing Sequences and Structures with T-Coffee



# ProbCons—consistency-based approach

- Combines iterative and progressive approaches with a unique probabilistic model.
- Uses Hidden Markov Models to calculate probability matrices for matching residues, uses this to construct a guide tree
- Progressive alignment hierarchically along guide tree
- Post-processing and iterative refinement (a little like MUSCLE)

# Examples: 5 alignments of 5 globins

- Let's look at a multiple sequence alignment (MSA) of five globins proteins.
- We'll use five prominent MSA programs (each program offers unique strengths)
  - ClustalW, Praline, MUSCLE (used at HomoloGene), ProbCons, and TCoffee.
- We'll focus on a histidine (H) residue that has a critical role in binding oxygen in globins, and should be aligned. But often it's not aligned, and all five programs give different answers.
- Our conclusion will be that there is no single best approach to MSA. Dozens of new programs have been introduced in recent years.

# ClustalW

CLUSTAL W (1.83) multiple sequence alignment

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

# Praline

(a) Praline multiple sequence alignment

beta globin	.....MVHLT <b>PEEKSAVTALWGKV..NVDEVGGGEALGRLLVVYPWTQRFFES.FG</b>
myoglobin	.....MGLS <b>DGEWQLVLNWVKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK</b>
neuroglobin	.....MERPE <b>PELIRQSWRAVSRSPLEHGTVLFARLFAL</b> EPDLLPLFQYNCR
soybean	.....MVAFT <b>EKQDALVSSSFEAFKANI</b> PQYSVVFYTSILEK <b>APAAKDLFS..FL</b>
rice	MALVEDNNNAVASF <b>EEQEALVLKSWAILKKDSANIALRFFL</b> KIFEVAPSASQMFS..FL
Consistency	00000000014265438257934573463364343624453686433*35344*50063
beta globin	DLST <b>PDAVMGNPKVKAHGKKVLGAFSDG</b> LAHLDNLKGTF <b>FATLSEL..HCDKLH..VDP</b>
myoglobin	HLKSEDEM <b>KASEDLKKHGATVLTALGGIL</b> KKKGHHEAEI <b>KPLAQS..HATKHK..IPV</b>
neuroglobin	QFSSPEDCLSS <b>PEFLDHIRKVMLVIDAAVTN</b> VEDLSSLEEYL <b>IASLGRKHRAVG..VKL</b>
soybean	A.NGVDP..TN <b>PKLTGHAEKL</b> FALVRDSAGQL. <b>KASGTVVADAA..LGSVH</b> AQKAVTD
rice	R.NSDVPLEKN <b>PKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATH</b> LKYGVGD
Consistency	3166354224776653*4368635424454451335634333542003335440000922
beta globin	<b>ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKV</b> VAGVANALAHKYH.....
myoglobin	<b>KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKEL</b> GFQG
neuroglobin	SSFSTVGESLLYMLEKCL.GPAFT <b>PATRAAWSQLYGAVVQAM</b> SRGWD..GE..
soybean	<b>PQFVVVK</b> EALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKK.....
rice	<b>AHFEVVKFALLDTIKEEV</b> VPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency	43744844498258542305336554454*55465426446754322001000

Note also the changing pattern of gaps within the boxed region in these five different alignments.

# MUSCLE

(b) MUSCLE (3.6) multiple sequence alignment

# Probcons

(c)

PROBCONS

# T-Coffee

(d)

CLUSTAL FORMAT for T-COFFEE Version 5.13

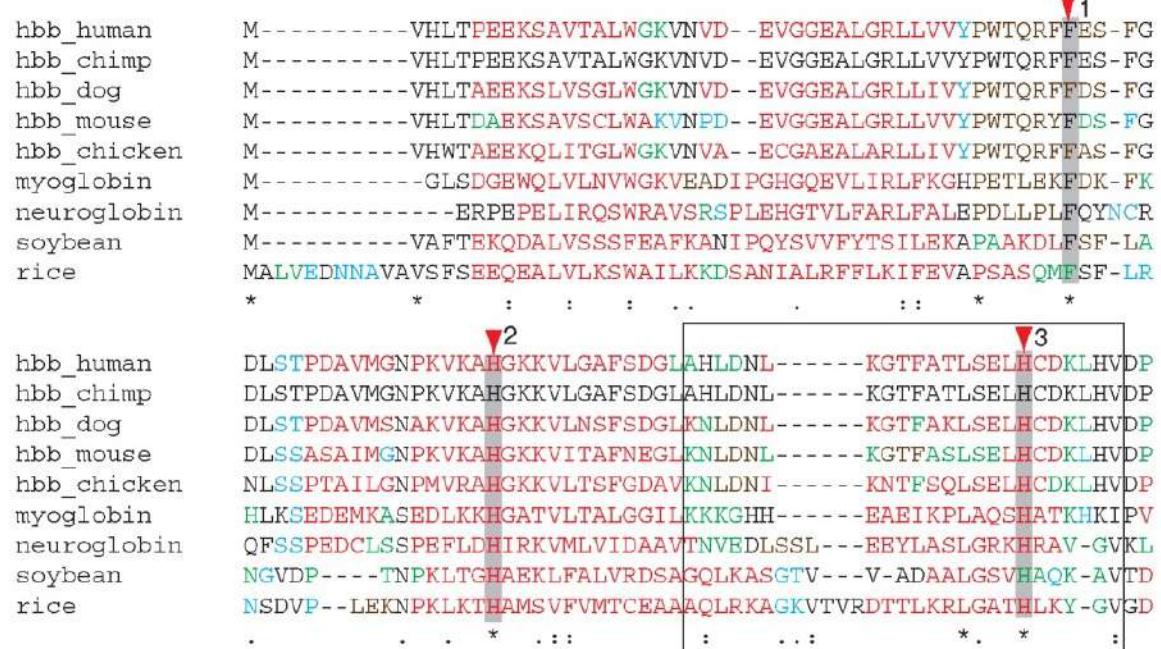
# MAFFT

(a) Alignment of nine globins by MAFFT FFT-NS-2 (v7.058b) (DSSP colors: turn, alpha helix, bend, 3/10 helix)

## MUSCLE

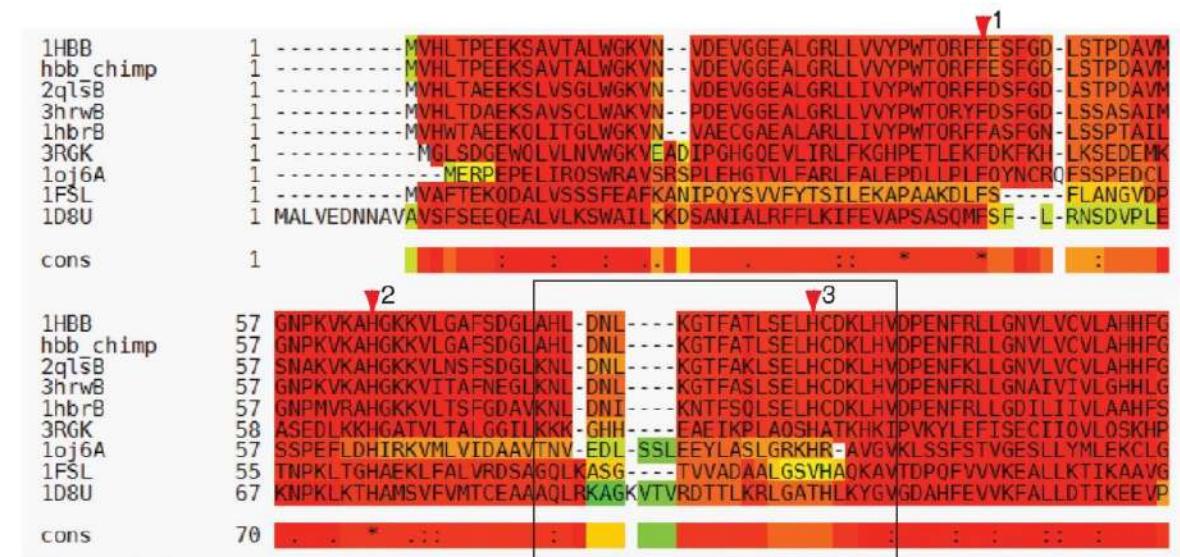
(c) Alignment of nine globins by ProbCons (version 1.12)

## ProbCons



(d) Alignment of nine globins by T-COFFEE (Expresso version\_10.00)

## T-COFFEE



1996 ~ 2000 Bachelor (第一屆推薦甄試入學)

2002 ~2002 Master

@ Computer Science, National Tsing Hua Uni.



Dr. Chuan Yi Tang

TITLE	CITED BY	YEAR
Constrained multiple sequence alignment tool development and its application to RNase family alignment CY Tang, CL Lu, MDT Chang, YT Tsai, YJ Sun, KM Chao, JM Chang, ... Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society, 127-137	91	2002

Oral presentation *Proceedings of the 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS 2002)*, 365-370, Monte Carlo Resort, Las Vegas, USA.

H-RNase3	-----RPPQFTRAQWFAI <b>G</b> HISLNPPRTIAMRA
H-RNase2	-----KPPQFTWAQWFET <b>G</b> HINMTSQGCTNAMQV
BP-RNaseA	-----KETA <b>A</b> AKFER <b>G</b> HMDSSSTAASSSNYCNQMMKS
BS-RNase	-----KE <b>S</b> AAAKFER <b>G</b> HMDSGNSPSSSSNYCNLMMCC
H-RNaseA	MALEKSLVRLLLLVLILLVLGWVQPSLGKE <b>R</b> AKKFQR <b>G</b> HMDSDSSPSSSSTYCNQMMRR
H-RNase4	-----MQDGMY <b>Q</b> RFLR <b>G</b> HVHPEET-GGSDRY <b>C</b> NLMMQR
RC-RNase	-----QNWATFQQ <b>K</b> HIINTPIIN----CNTIMDN
H-RNase3	-INNYRW <b>R</b> CKNQNTFLRTTFANVVNV <b>G</b> NQSIRCPHNRTLNNCHR <b>R</b> F <b>R</b> VPLLHC <b>D</b> LINP
H-RNase2	-INNYQR <b>R</b> CKNQNTFLLTTFANVVNV <b>G</b> NP <small>M</small> T <b>C</b> PSNKTRKNCHHSGSQVPLI <b>H</b> CNLTTP
BP-RNaseA	-RNLT <b>K</b> DR <b>C</b> CKPVNTFVHE <b>L</b> ADVQAVCSQKNVACKNGQT--NCYQSYSTM <b>S</b> ITDC <b>R</b> ETGS
BS-RNase	-RKMTQ <b>G</b> K <b>C</b> CKPVNTFVHE <b>L</b> ADVKA <b>V</b> CSQKKV <b>T</b> CKNGQT--NCYQSKSTM <b>R</b> I <b>T</b> DC <b>R</b> ETGS
H-RNaseA	-RNMTQ <b>G</b> R <b>C</b> CKPVNTFVHEPLVDVQNV <b>C</b> QE <b>K</b> V <b>T</b> CKNG <b>Q</b> G--NCYKSNSSMHI <b>T</b> DC <b>R</b> L <b>T</b> NG
H-RNase4	-RKMTLY <b>H</b> <b>C</b> KRFNTFI <b>H</b> E <b>D</b> IWNIRSI <b>C</b> STTN <b>I</b> Q <b>C</b> CKNG <b>K</b> M--NC <b>E</b> G--VV <b>K</b> VTDC <b>R</b> DTGS
RC-RNase	NIYIVGG <b>Q</b> CKRVNTF <b>I</b> ISSATTVK <b>A</b> <b>I</b> CTG--VINMN-----VLSTTRFQLNT <b>C</b> RTSI
H-RNase3	GAQNISNC <b>T</b> YADRPGRRFYVVA <b>C</b> DNRDPR-DS <b>P</b> RY <b>P</b> VVP <b>V</b> HLD <b>T</b> TI----
H-RNase2	SPQNISNC <b>R</b> YA <b>Q</b> T <b>P</b> ANMFY <b>I</b> VA <b>C</b> DNRD <b>Q</b> RRDPP <b>Q</b> Y <b>P</b> VVP <b>V</b> HLD <b>R</b> II----
BP-RNaseA	S--KYPNCAY <b>K</b> TT <b>Q</b> ANK <b>H</b> I <b>I</b> VA <b>C</b> EGN-----PYVP <b>V</b> HF <b>D</b> ASV----
BS-RNase	S--KYPNCAY <b>K</b> TT <b>Q</b> VE <b>K</b> HI <b>I</b> VA <b>C</b> GGK-----PSVP <b>V</b> HF <b>D</b> ASV----
H-RNaseA	S--RYPNCAY <b>R</b> TS <b>P</b> KER <b>H</b> I <b>I</b> VA <b>C</b> EGS-----PYVP <b>V</b> HF <b>D</b> ASVEDST
H-RNase4	S--RAPNC <b>R</b> YRAIASTR <b>R</b> V <b>V</b> I <b>A</b> <b>C</b> EGN-----PQVP <b>V</b> HF <b>D</b> G-----
RC-RNase	T---PRCP <b>P</b> YSS <b>R</b> TETNYICVK <b>C</b> EN-----QYPVP <b>V</b> HFAGIGRCP-

Fig. 1. The multiple sequence alignment of seven RNases by WorkBench 3.2: The key active site residues homologous to His12, Lys41, and His119 of BP-RNaseA, the cysteine residues responsible for disulfide bond linkage and two matched Gln residues are shown in boxes.

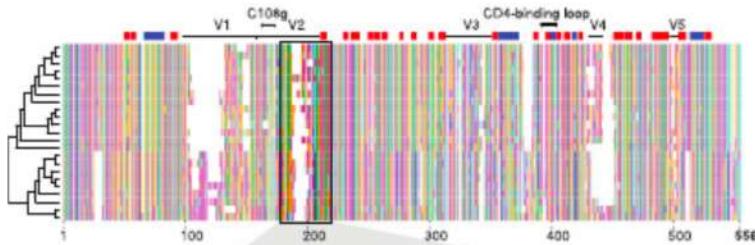
H-RNase3	-RP--PQFTRAQWFAIQHIS-L-NPP---R--CTIAMRAI---NN--Y--RWPCKNQNTF
H-RNase2	MKP--PQFTWAQWFETQHIN-M-TSQ---Q--CTNAMQVI---NN--Y-QR-RCKPNQNTF
BP-RNaseA	-KETAA----AK-FERQHMD-SSTSAAASSNYC-N--QMMKSRN---LTKD-RCKPVNTF
BS-RNase	-KESAA----AK-FERQHMD-SGNSPSSSSNYC-N--LMMCCRK---MTQG-KCKPVNTF
H-RNaseA	-KES-R----AKAFQRQHMD-SDSSPSSSSSTYC-N--QMM-RRRN--MTQG-RCKPVNTF
H-RNase4	-MQDGMY---QR-FLRQHIVHPEET--GGSDRYC-N--LMMQRRK---MTLY-HCKRFNTF
RC-RNase	--XN-W----A-TFQQKHI--I-NT-PIIN--C-N--TIM--DNNIYIVGG-QCKRVNTF
H-RNase3	LRTTFANVVNVCGNQSIRCPHNRTLNNCHRCSRFRVPL-LHC-DLINP-GAQNISNCRYAD
H-RNase2	LLTTFANVVNVCGNPNMTCPSNKTRKNCHHSGSQVPL-IHC-NLTPP-SPQNISNCRYAQ
BP-RNaseA	VHESLADVQAVCSQKNVACK-N-GQTNCYQSYSTMSI-TDC-RET-GSSKYP--NCAY-K
BS-RNase	VHESLADVKAVALCSQKKVTCK-N-GQTNCYQSKSTMRI-TDC-RET-GSSKYP--NCAY-K
H-RNaseA	VHEPLVDVQNVCFQEKVTCR-N-GQGNQCYKSNSSMHI-TDC-RLTNG-SRYP--NCAY-R
H-RNase4	IHEIDIWNIRSICSTTNIQCK-N-GKMINCHE--GVVKV-TDC-RDT-GSSRAP--NCRY-R
RC-RNase	IISSATTVKAIQ--TGV-I--N--M-NVL-STTRFQLNT-QTR-TSI-TP-R--PCPY--
H-RNase3	R-PGR-RFYVVAQDNRD-PRDSPR-YPVVPVHLDTTI----
H-RNase2	T-PAN-MFYIVACDNRDQRRD-PPQYPVVPVHLD-RI----
BP-RNaseA	TTQAN-KHIVACEG-----N---PY--VPVHF DASV----
BS-RNase	TTQVE-KHIVACGG-----K---PS--VPVHF DASV----
H-RNaseA	TSPKE-RHIVACEG-----S---PY--VPVHF DASV----
H-RNase4	AI-ASTRRVVIACEG-----N---PQ--VPVHF D-G----
RC-RNase	SSRTETNYICVKQE-----N---QY---PVHF-AGIGRCP

Fig. 2. The multiple sequence alignment of seven RNases by our CMSA: The key active site residues homologous to His12, Lys41, and His119 of BP-RNaseA, the cysteine residues responsible for disulfide bond linkage, and two matched Gln residues are shown in boxes.

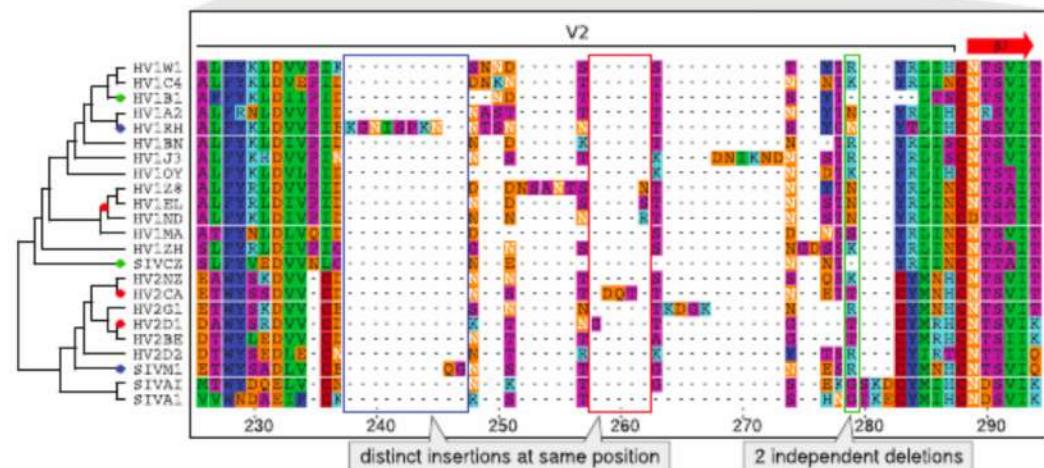
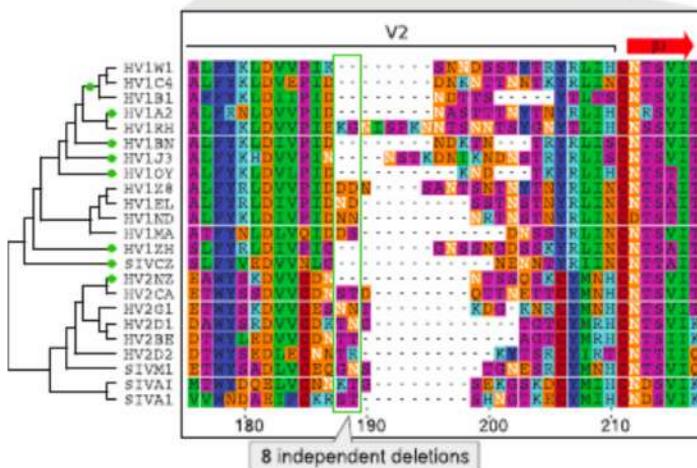
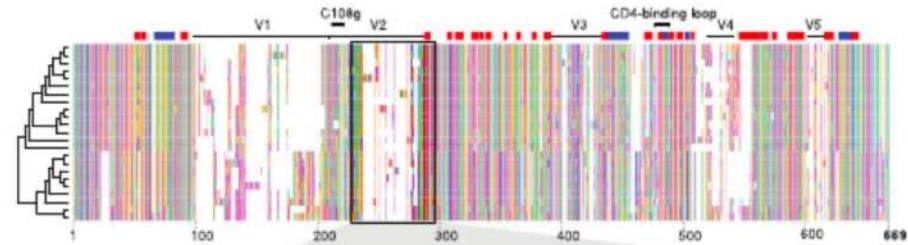
# Prank

- Löytönoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320, 1632–5 (2008).

A



B

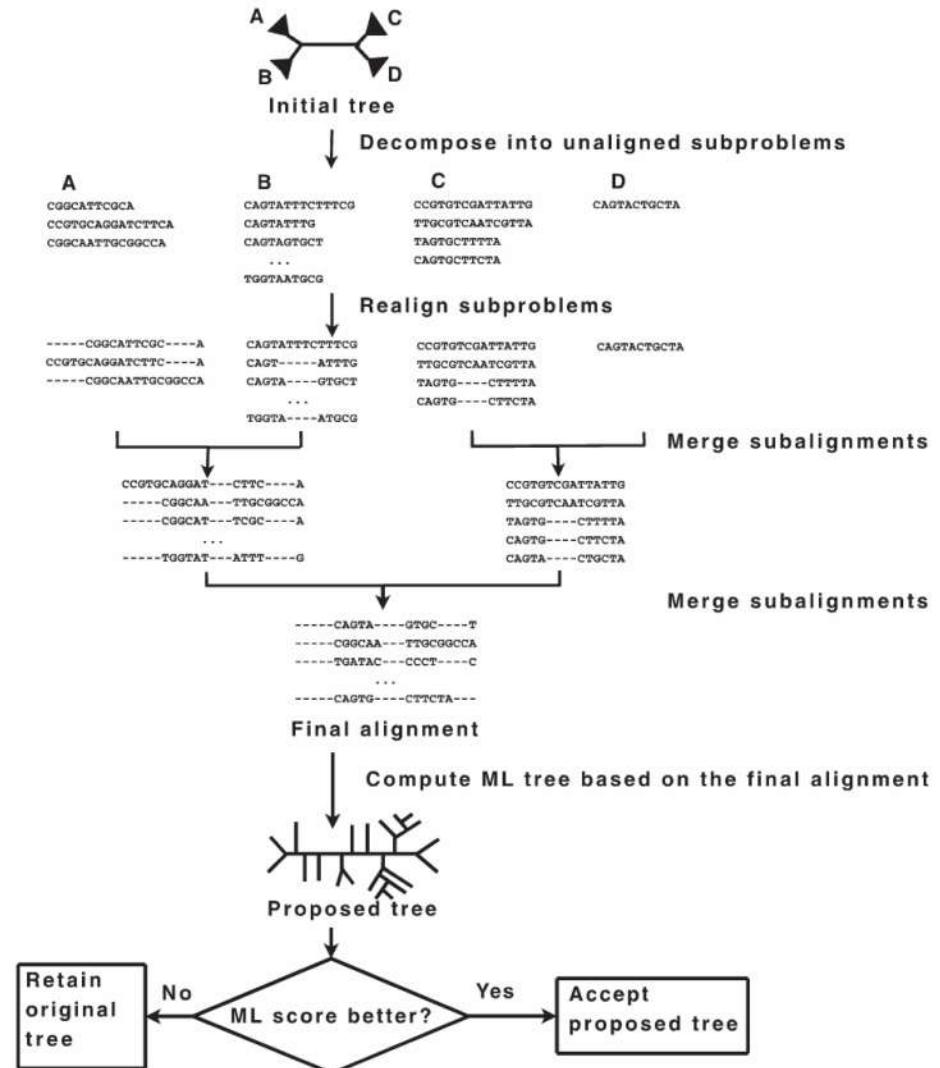
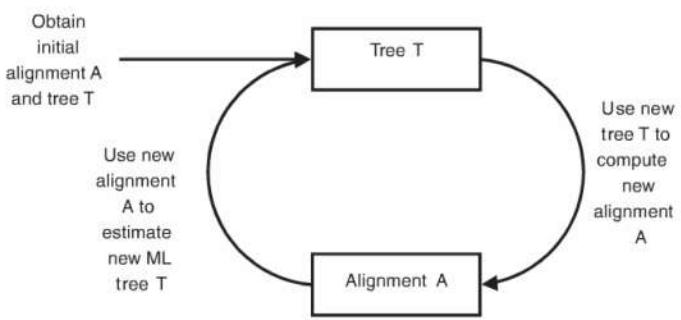


Adapted from Cedric Notredame

# SATe

Liu, K., Raghavan, S., Nelesen, S.,  
Linder, C. R. & Warnow, T. Rapid  
and accurate large-scale  
coestimation of sequence  
alignments and phylogenetic trees.  
*Science* 324, 1561–4 (2009)

**Fig. 1.** SATé's second stage. Beginning with the current best tree/alignment pair, SATé iterates between realigning on the current tree and estimating a RAxML tree for each new alignment. At the end of each iteration, the tree/alignment pair optimizing ML under the GTR+Gamma model of evolution is saved.



# What is the Best Method?

- Kemena, C. & Notredame, C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**, 2455–65 (2009).

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	<b>22.74</b>	
PRANK	Gap	NO	<b>26.18</b>	Science2008
MAFFT	Iterative	NO	<b>26.18</b>	
Muscle	Iterative	NO	<b>31.37</b>	
ProbCons	Consistency	NO	<b>40.80</b>	
ProbCons	MonoPhasic	NO	<b>37.53</b>	
T-Coffee	Consistency	NO	<b>42.30</b>	
M-Coffe4	Consistency	NO	<b>43.60</b>	
<b>PSI-Coffee</b>	<b>Consistency</b>	<b>Profile</b>	<b>53.71</b>	
<b>PROMAL</b>	<b>Consistency</b>	<b>Profile</b>	<b>55.08</b>	
<b>PROMAL-3D</b>	<b>Consistency</b>	<b>PDB</b>	<b>57.60</b>	
<b>3D-Coffee</b>	<b>Consistency</b>	<b>PDB</b>	<b>61.00</b>	<b>Expresso</b>

**Score:** fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

Adapted from Cedric Notredame

# A better Question...

- What is the Best Alignment ?
- What is the best bit of my alignment ?

# Situation <=> Solution

	MUSCLE	MAFFT	PROBCONS	T-COFFEE	CLUSTALW
Accuracy	++	+++	+++	+++	+
<100 Seq.	++	++	+++	+++	+
>100 Seq.	+++	+++	-	+	+
Remote Homologues	++	+++	+++	+++	+
MSA vs Seq.	-	-		+++	+++
MSA vs MSA	-	-	-	+++	+++
>2 MSAs	-	-	-	+++	-
Seq. vs Struc.	-	-	-	+++	+
Splicing Var.	-	+++	-	+++	-
Reformat	-	-	-	+++	++
Phylogeny	-	-	-	+	++
Evaluation	-	-	+	+++	-
Speed	+++	+++	+	+	++

# Purpose <=> Solution

	MUSCLE	MAFFT	PROBCONS	T-COFFEE	CLUSTALW
Dist Based Phylogeny	+++	+++	++	++	++
ML or MP Phylogeny	++	+++	+++	+++	++
Profile Construction	++	+++	+++	+++	++
3D Modeling	++	++	++	+++	+
Secondary Structure P	+++	+++	++	++	++

# Highly Cited Articles

MEGA3	2004	Brief Bioinform.;5(2):150-63. PMID: 15260895	6630
MRBAYES	2001	Huelsenbeck and Ronquist 2001 Bioinformatics;17(8):754-5 PMID: 11524383	5707
CLUSTALW	** 1994	Thompson et al. 1994 Nucleic Acids Res.;22(22):4673-80 PMID: 7984417	29658
BLAST	* 1990	Altschul et al. 1990 J Mol Biol.;215(3):403-10 PMID: 2231712	27660
Neighbor-Joining Algorithm	1987	Saitou and Nei 1987 Mol Biol Evol.;4(4):406-25 PMID: 3447015	20523
Non-Parametric Bootstrap in Phylogenetics	1985	Felsenstein 1985 Evolution;39(4):783-91 PMID: N/A	14566

as of 2006 in the ISI web of knowledge:

§ Source: ISI Web of Knowledge, as of 29.03.2010

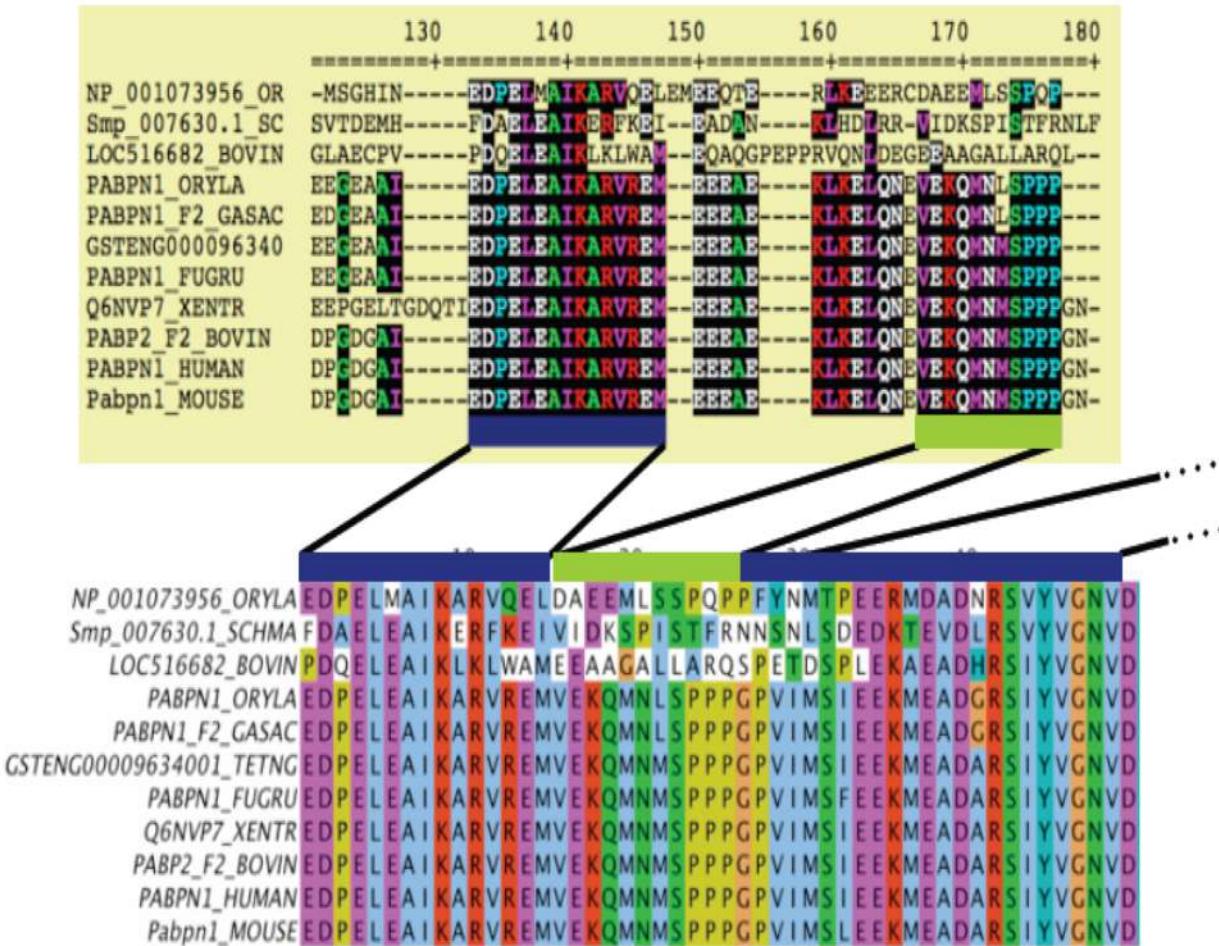
\* most cited paper that year, 26th most cited in the entirety of science

\*\* second most cited paper that year, 31st most cited paper in the entirety of science

# Example Phylogeny Estimation Workflow

- Provide guidance/reference for planning your own analyses
- Show how different tools/stages in an analysis can link together
  - i.e. providing a context in which to place what you learn later on
- Provide a "target" to criticise
  - once you know more, what would you do differently?

# What is a Reduced MSA?

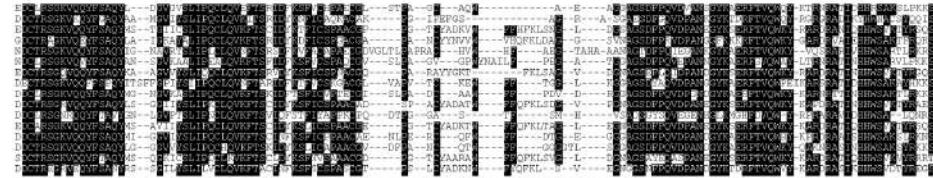


# Preparing Reduced Alignments

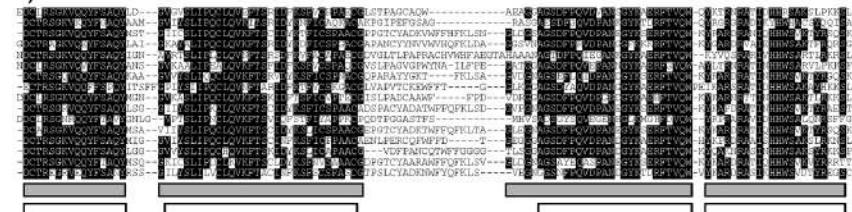
- Choose which columns to remove
  - By eye (using alignment edit, JalView)
  - Automatically
    - Gblocks
    - trimAI

# Gblocks

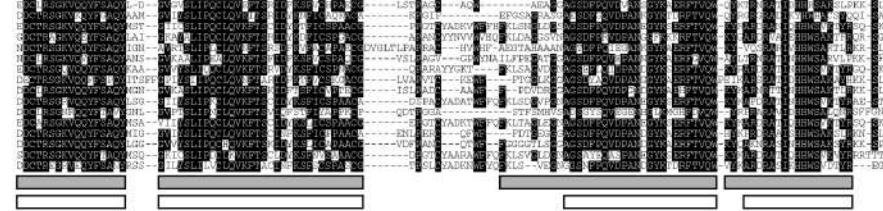
a)



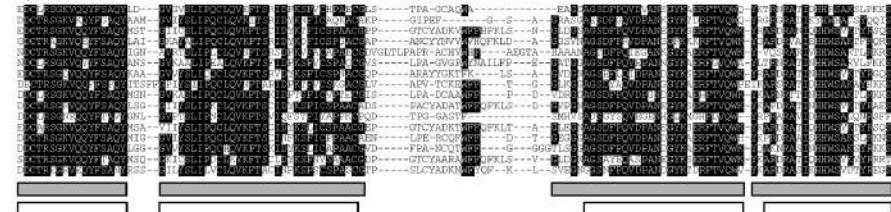
b)



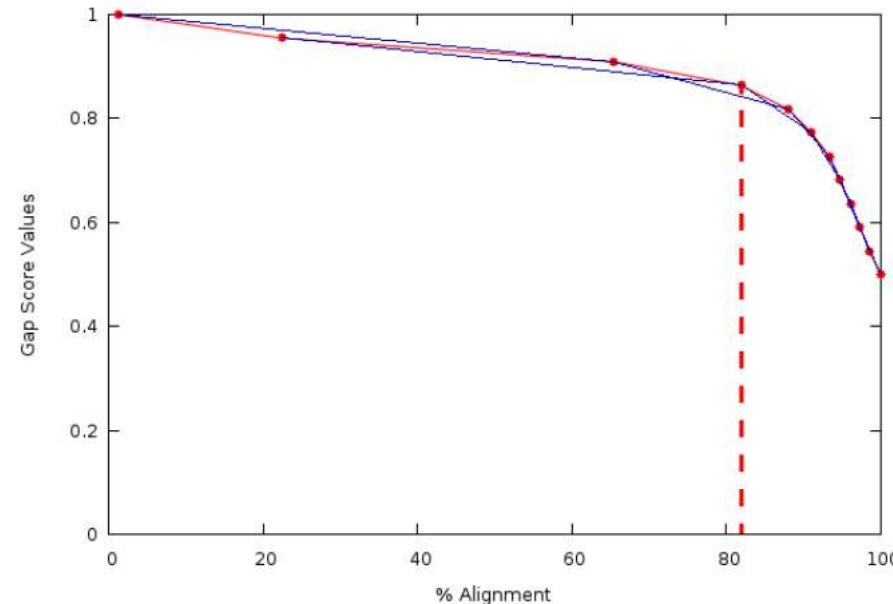
c)



d)



# trimAl



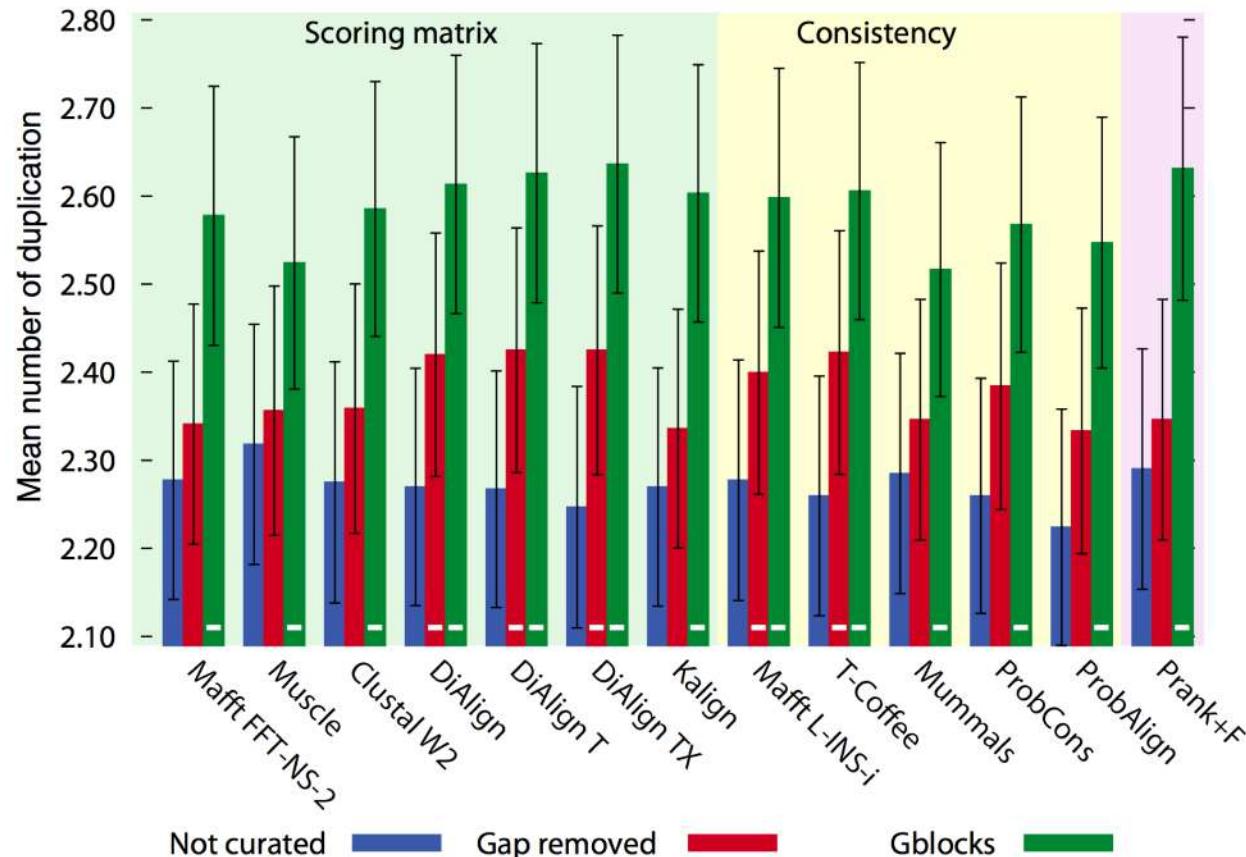
419 citation by Google  
104 citation by Google

Talavera G, Castresana J (2007) Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Syst Biol* 56: 564–577.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

# Effect of excluding gaps and variable regions

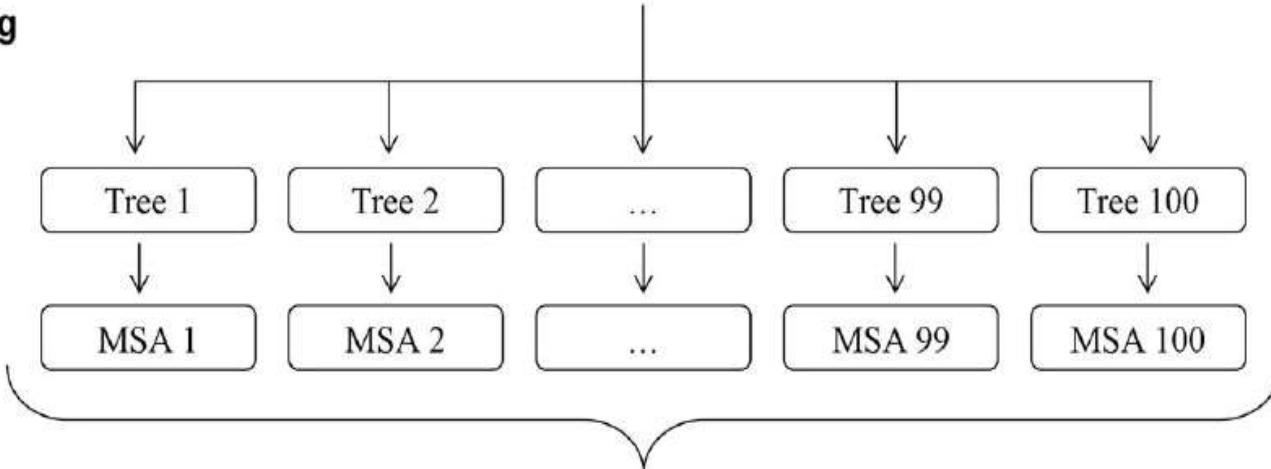
- gaps carry substantial phylogenetic signal, but are poorly exploited by most alignment and tree building programs



## Base MSA

```
RFLILHCSQELENLFSPYCLVKSLQITFQLCLLVFGVS--GT--REVLIRIV---N--QLQYLGTLTIFELLMP  
VHLVSLQNDLNGIFGKSLLSLLTAAVICTVAVYTLI--QGP--TLEGFT--Y--VI--FIGTSVMQVYLVC  
QLLNGLCRKYNDIFKVAFLVSNFVGAGSLCFYLFMLSE--TS--DVLIIA--Q--YILPTLVLVGFTEIC  
ARALDLSEEVNNIFSFLLWNFIAASLVICPAGFQIT--ASN--VEDIGV--Y--FI--FFSASLVQVFVVC  
QRIRSLLTLCQRIVSPYIILSQIILSALIIICFSGYRLQH--VGI-RD---NPGQFISMQLQFVSVMILQIYLPC  
TKVRRLTRECEVLVSPYVLSQVVFSAFIICFSAYRLVH--MGF-KQ---RPGLFVTTVQFVAVMIVQIFLPC  
NLIIDYAAIRPAVTRTIVQFLLIGICLGLSMINLLF--FAD--IWTLGL--A--TVAYINGLMVQTFFFC  
ALCLNLGHFLNEYFRPLIC-QFVAASLHLCLVLCYQL---SAN-ILQPALL--F--YAAFTAAVVQGVSIYC  
QRWVALVALLNRGYGLSMLMQVGNDFLAITSNCYWMFLNF-RQSAASPFIDLQIVASGV-WSAPHLGNGVLVLS
```

## Neighbor joining bootstrap tree reconstruction

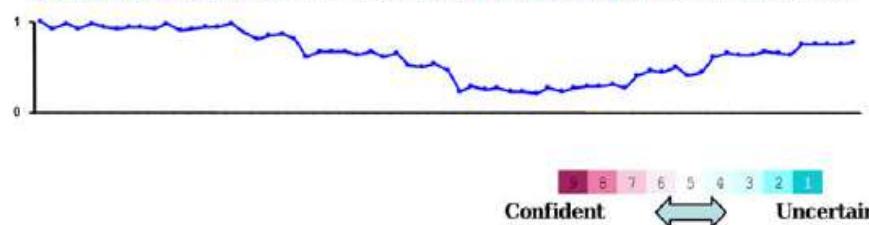


## Progressive alignment

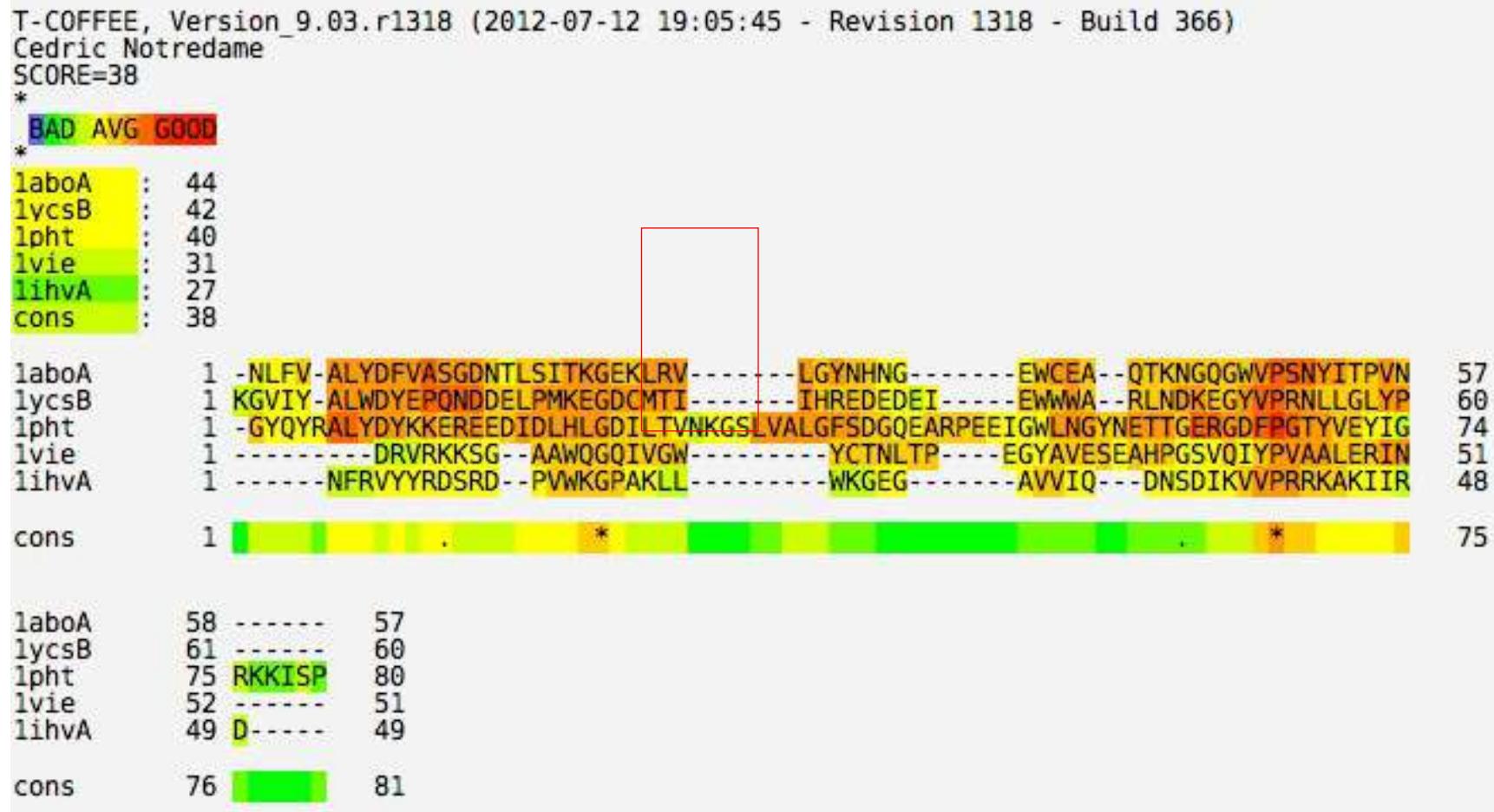
A color-coded sequence alignment is shown, where each residue is highlighted in a different color based on its confidence score. The colors range from dark red (highest confidence) through pink, light blue, cyan, and green to light yellow (lowest confidence). The alignment shows a high degree of conservation across most positions, with significant variation indicated by the color changes.

```
RFLILHCSQELENLFSPYCLVKSLQITFQLCLLVFGVS--GT--REVLIRIV---N--QLQYLGTLTIFELLMP  
VHLVSLQNDLNGIFGKSLLSLLTAAVICTVAVYTLI--QGP--TLEGFT--Y--VI--FIGTSVMQVYLVC  
QLLNGLCRKYNDIFKVAFLVSNFVGAGSLCFYLFMLSE--TS--DVLIIA--Q--YILPTLVLVGFTEIC  
ARALDLSEEVNNIFSFLLWNFIAASLVICPAGFQIT--ASN--VEDIGV--Y--FI--FFSASLVQVFVVC  
QRIRSLLTLCQRIVSPYIILSQIILSALIIICFSGYRLQH--VGI-RD---NPGQFISMQLQFVSVMILQIYLPC  
TKVRRLTRECEVLVSPYVLSQVVFSAFIICFSAYRLVH--MGF-KQ---RPGLFVTTVQFVAVMIVQIFLPC  
NLIIDYAAIRPAVTRTIVQFLLIGICLGLSMINLLF--FAD--IWTLGL--A--TVAYINGLMVQTFFFC  
ALCLNLGHFLNEYFRPLIC-QFVAASLHLCLVLCYQL---SAN-ILQPALL--F--YAAFTAAVVQGVSIYC  
QRWVALVALLNRGYGLSMLMQVGNDFLAITSNCYWMFLNF-RQSAASPFIDLQIVASGV-WSAPHLGNGVLVLS
```

## GUIDANCE scores



# TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction



# Replication instead of filtering

Original align.

1aboA	-NLFV-ALYDFVASGDNTLSITKGEKLRV-----LGYNHNG-----
1ycsB	KGVIY-ALWDYEPQNDDELPMKEGDCMTI-----IHREDEDEI---
1pht	-GYQYRALYDYKKEREEDIDLHLGDILTVNK GSLVALGFSDGQEARPE
1vie	-----DRVRKKSG--AAWQGQIVGW-----YCTNLTP---
1ihvA	-----NFRVYYRDSRD--PVWKGP A KLL-----WKGEG-----



TCS scores

1aboA	-4445-66666676665455566655666-----6565544-----
1ycsB	33444-6666667775556666666666-----655554434---
1pht	-54444776665656655666665554344466666655445555
1vie	-----33344444--555555555-----5555555-----
1ihvA	-----33344444444--4555554433-----33344-----
cons	1333324443434433334444554333111223332221111111



TCS enrich align

1aboA	-NNNLLL	...	-
1ycsB	KGGGVVV	...	-
1pht	-GGGYYY	...	E
1vie	-----	...	-
1ihvA	-----	...	-

# 853 Yeast ToL

RF: average Robinson-Foulds distance respect to Yeast ToL.

TPs: the number of genes whose tree topology is identical with yeast ToL.

	Original		Gblocks relaxed		Gblocks stringent		trimAl gappyout		trimAl strictplus		TCS replicate	
	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs
ClustalW	<b>0.90</b>	643	0.99	629	1.24	584	0.95	628	1.31	561	0.91	<b>649</b>
MAFFT	0.80	665	0.83	653	1.26	573	0.83	657	1.28	562	<b>0.76</b>	<b>669</b>
Muscle	0.95	639	0.91	646	1.26	578	0.96	633	1.29	559	<b>0.84</b>	<b>662</b>
PRANK	<b>0.79</b>	<b>665</b>	0.88	642	1.28	565	0.84	648	1.19	575	0.81	662
SATe	0.86	660	0.87	650	1.28	578	0.85	655	1.25	567	<b>0.79</b>	<b>666</b>
AVE	0.86	654	0.896	644	1.26	575	0.88	644	1.26	565	<b>0.82</b>	<b>661</b>

# Multiple alignment of myoglobins, alpha globins, beta globins

Open circles: positions that distinguish myoglobins, alpha globins, beta globins

▼ gaps    ◇ 100% conserved

myoglobin_kanga	-----MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF
myoglobin_harbo	-----MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF
myoglobin_gray_	-----MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF
<u>alpha_globin_ho</u>	-----MV-LSAADKTNVKAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHF
<u>alpha_globin_ka</u>	-----V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHF
<u>alpha globin do</u>	-----V-LSPADKTNIKSTWDKIGGHAGDYGGAEALDRTFQSFPPTKTYFPHF
<u>beta_globin_dog</u>	-----MVHLTAEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF
<u>beta_globin_rab</u>	-----MVHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF
<u>beta globin kan</u>	-----VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF
globin_riverlam	-PIVDS----GSPAVLSAAEKTIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFPKF
globin_sealampr	MPIVDT---GSVAPLSAAEKTIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFPKF
globin_soybean	-----VAFTEKQDALVSSSFEAFKANIPOYSVVFYTSILEKAPAAKDLFSFL
globin_insect	MKFLILALCFAAASALSADQISTVQASFDFKVKGD---PVGILYAVFKADPSIMAKFTQF
	: : : : . : * * * :

## Stage 2: Multiple sequence alignment

1. Confirm that all sequences are homologous
2. Adjust gap creation and extension penalties as needed to optimize the alignment
3. Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data).

# Stage 3: models of DNA and amino acid substitution

## Stage 3: Models of substitution

- The simplest approach to measuring distances between sequences is to align pairs of sequences, and then to count the number of differences. The degree of divergence is called the Hamming distance.
- For an alignment of length  $N$  with  $n$  sites at which there are differences, the degree of divergence  $D$  is:
  - $D = n / N$

## Stage 3: Models of substitution

- The simplest approach to measuring distances between sequences is to align pairs of sequences, and then to count the number of differences. The degree of divergence is called the Hamming distance.
- For an alignment of length  $N$  with  $n$  sites at which there are differences, the degree of divergence  $D$  is:
  - $D = n / N$
- But observed differences do not equal genetic distance!
- Genetic distance involves mutations that are not observed directly (see Figure 11.11).

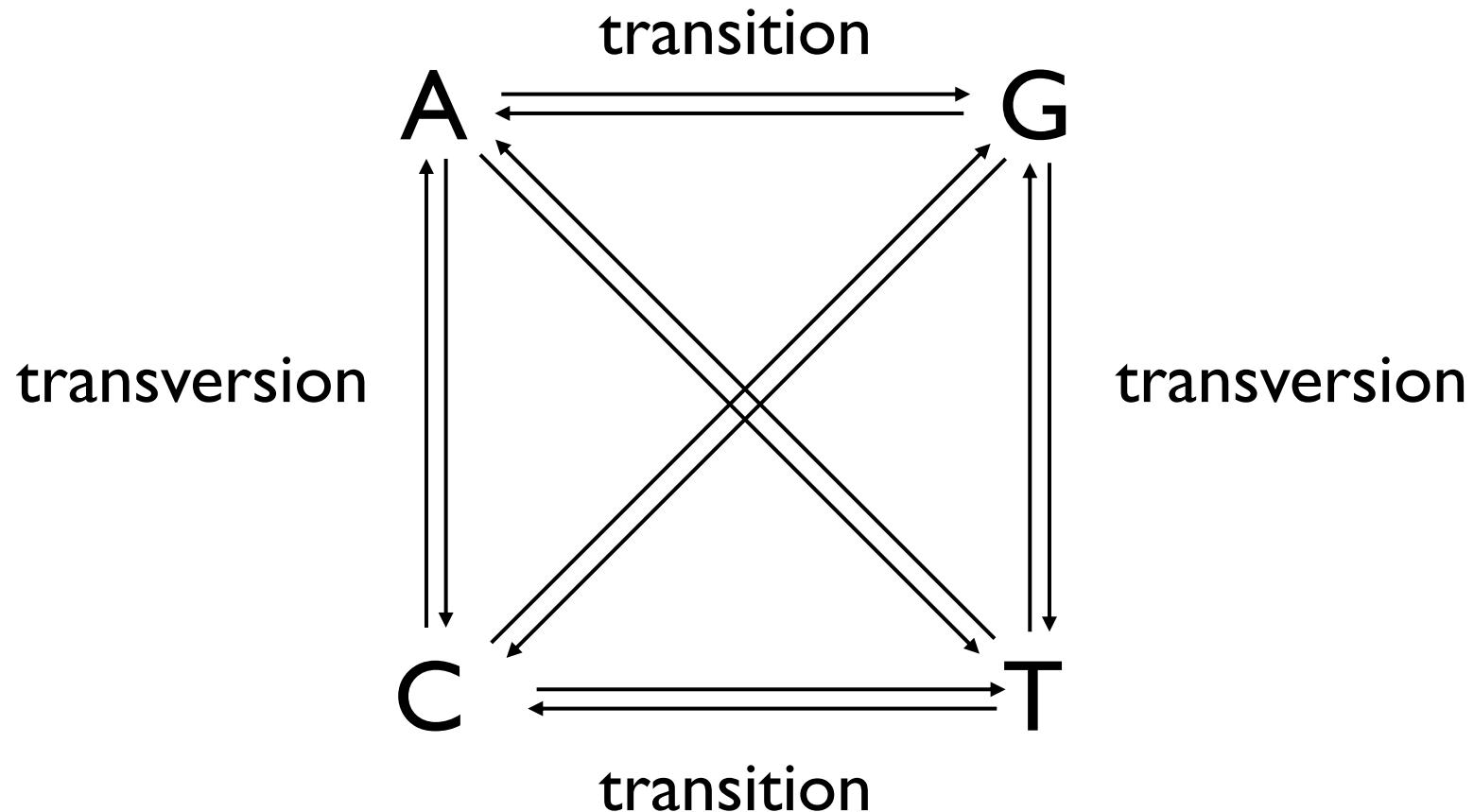
## Stage 3: Models of substitution

- Jukes and Cantor (1969) proposed a corrective formula:

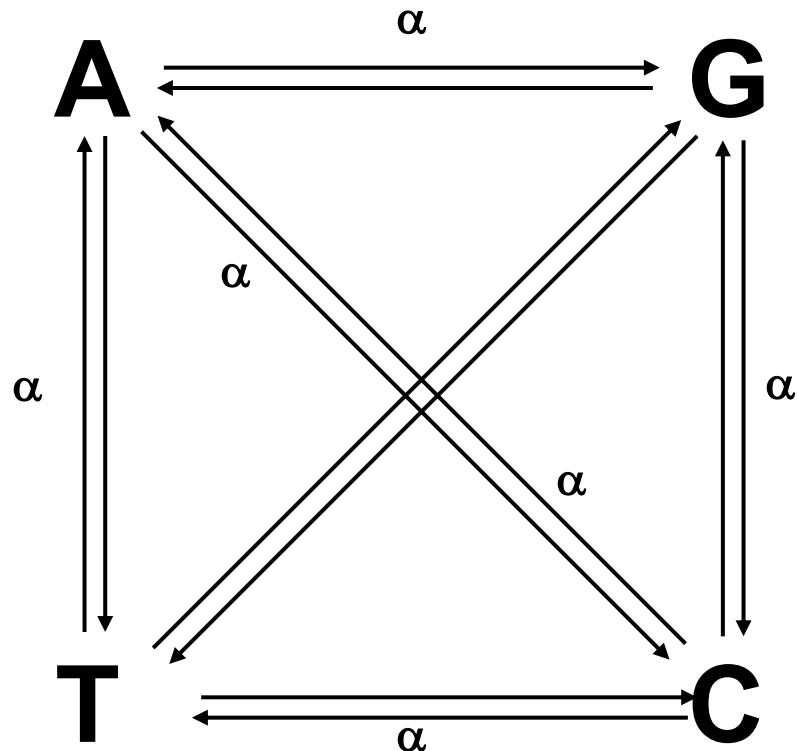
$$D = \left(-\frac{3}{4}\right) \ln \left(l - \frac{4}{3}p\right)$$

- This model describes the probability that one nucleotide will change into another. It assumes that each residue is equally likely to change into any other (i.e. the rate of transversions equals the rate of transitions). In practice, the transition rate is typically greater than the transversion rate.

There are dozens of models of nucleotide substitution



# Jukes and Cantor one-parameter model of nucleotide substitution ( $a=b$ )



## Stage 4: Tree-building methods: distance

- Jukes and Cantor (1969) proposed a corrective formula:

$$D = \left(-\frac{3}{4}\right) \ln \left(1 - \frac{4}{3}p\right)$$

- Consider an alignment where 3/60 aligned residues differ.

- The normalized Hamming distance is  $3/60 = 0.05$ .
  - The Jukes-Cantor correction is

$$D = \left(-\frac{3}{4}\right) \ln \left(1 - \frac{4}{3}0.05\right) = 0.052$$

- When 30/60 aligned residues differ, the Jukes-Cantor correction is more substantial:

$$D = \left(-\frac{3}{4}\right) \ln \left(1 - \frac{4}{3}0.5\right) = 0.82$$

(a) Number of differences

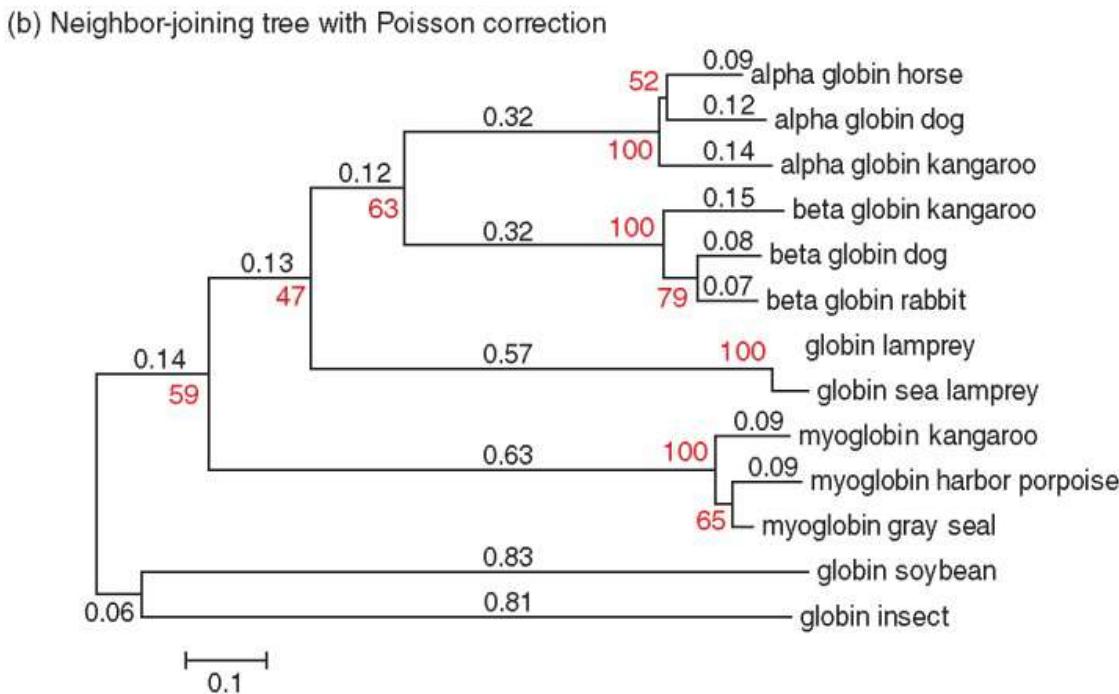
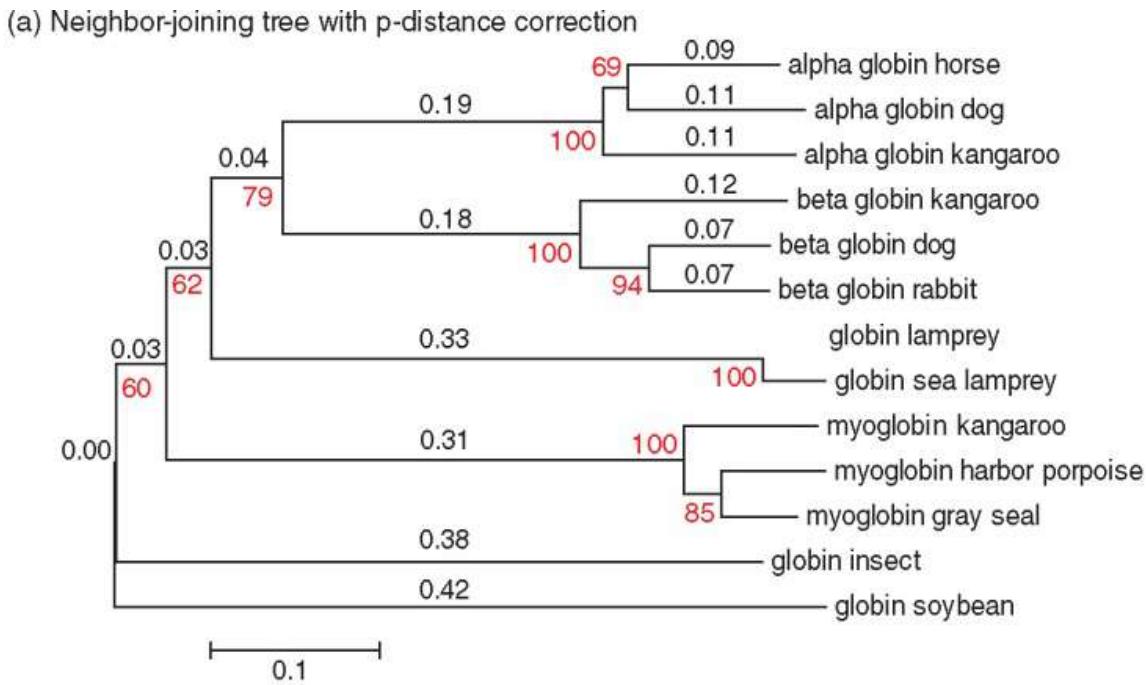
	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbhabor porpoise P68278 <i>Phocoena pho...</i>	19											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	16	12										
4. alphahorse P01958 <i>Equus caballus</i>	84	84	84									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	85	87	84	24								
6. alphadog P60529 <i>Canis lupus familiaris</i>	88	88	86	22	27							
7. betadog XP_537902 <i>Canis lupus familia...</i>	80	79	78	66	63	67						
8. betarabbit NP_001075729 <i>Oryctolagus c...</i>	80	81	78	64	67	65	16					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	83	82	80	68	69	66	25	28				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	88	92	88	77	77	76	83	83	81			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	89	91	89	76	77	76	83	85	81	8		
12. globinsoybean 711674A <i>Glycine max</i> (so...	98	97	97	93	93	93	87	90	90	93	94	
13. globininsect P02229 <i>Chironomus thummi...</i>	87	88	86	92	93	97	92	90	94	98	99	91

(b) p-distance

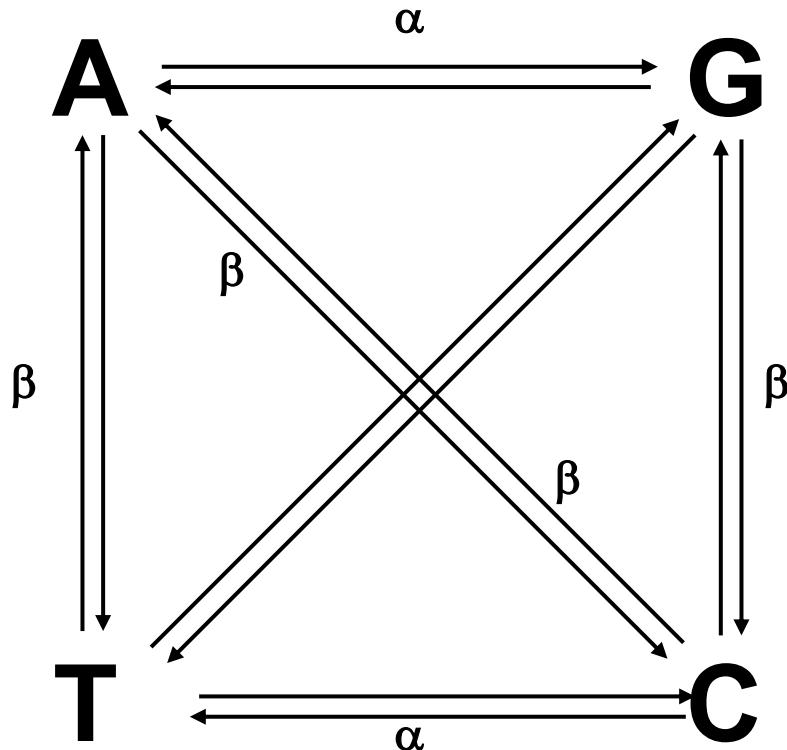
	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbhabor porpoise P68278 <i>Phocoena pho...</i>	0.17											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	0.14	0.11										
4. alphahorse P01958 <i>Equus caballus</i>	0.74	0.74	0.74									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	0.75	0.77	0.74	0.21								
6. alphadog P60529 <i>Canis lupus familiaris</i>	0.78	0.78	0.76	0.19	0.24							
7. betadog XP_537902 <i>Canis lupus familia...</i>	0.71	0.70	0.69	0.58	0.61	0.59						
8. betarabbit NP_001075729 <i>Oryctolagus c...</i>	0.71	0.72	0.69	0.57	0.59	0.58	0.14					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	0.73	0.73	0.71	0.60	0.61	0.58	0.22	0.25				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	0.78	0.81	0.78	0.68	0.68	0.67	0.73	0.73	0.72			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	0.79	0.81	0.79	0.67	0.68	0.67	0.73	0.75	0.72	0.07		
12. globinsoybean 711674A <i>Glycine max</i> (so...	0.87	0.86	0.86	0.82	0.82	0.82	0.77	0.80	0.80	0.82	0.83	
13. globininsect P02229 <i>Chironomus thummi...</i>	0.77	0.78	0.76	0.81	0.82	0.86	0.81	0.80	0.83	0.78	0.79	0.81

(c) Poisson correction

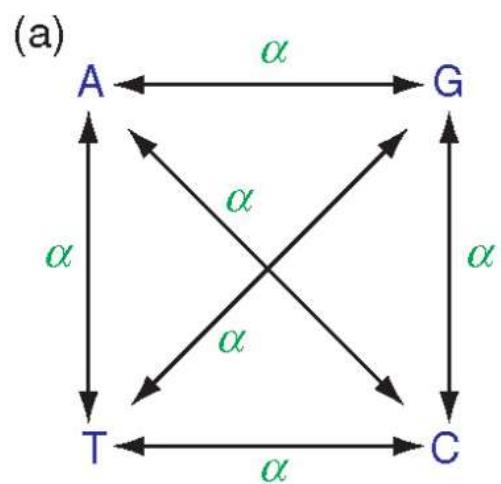
	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbhabor porpoise P68278 <i>Phocoena pho...</i>	0.18											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	0.15	0.11										
4. alphahorse P01958 <i>Equus caballus</i>	1.36	1.36	1.36									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	1.40	1.47	1.36	0.24								
6. alphadog P60529 <i>Canis lupus familiaris</i>	1.51	1.51	1.43	0.22	0.27							
7. betadog XP_537902 <i>Canis lupus familia...</i>	1.23	1.20	1.17	0.88	0.94	0.90						
8. betarabbit NP_001075729 <i>Oryctolagus c...</i>	1.23	1.26	1.17	0.84	0.90	0.86	0.15					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	1.33	1.29	1.23	0.92	0.94	0.88	0.25	0.28				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	1.51	1.68	1.51	1.14	1.14	1.12	1.33	1.33	1.26			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	1.55	1.64	1.55	1.12	1.14	1.12	1.33	1.40	1.26	0.07		
12. globinsoybean 711674A <i>Glycine max</i> (so...	2.02	1.95	1.95	1.73	1.73	1.73	1.47	1.59	1.59	1.73	1.78	
13. globininsect P02229 <i>Chironomus thummi...</i>	1.47	1.51	1.43	1.68	1.73	1.95	1.68	1.59	1.78	1.51	1.55	1.64



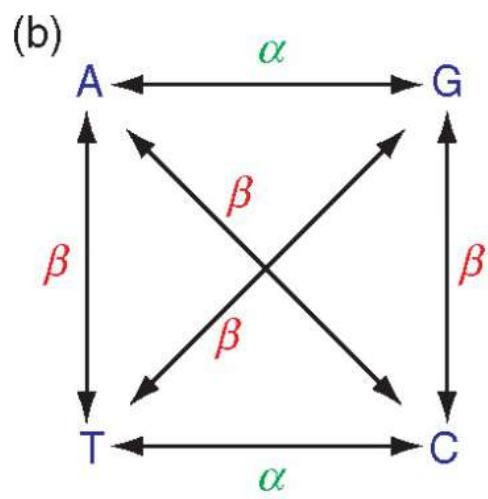
Kimura two-parameter model of nucleotide substitution (assumes  $a \neq b$ )



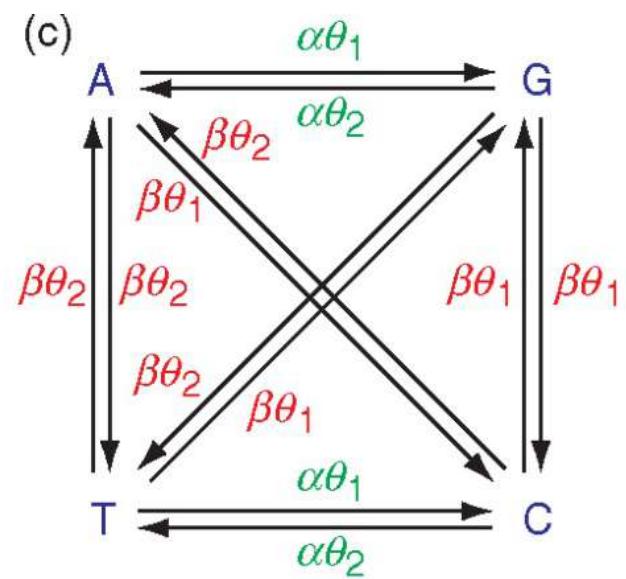
Jukes-Cantor model



Kimura model



Tamura model



## (a) ProtTest lowest- (best-)scoring of 112 models for amino acid substitution (13 globins)

Model	deltaAIC*	AIC	AICw	-lnL
<hr/>				
LG+G+F	0.00	5883.41	0.52	-2898.71
LG+I+G+F	0.37	5883.78	0.43	-2897.89
LG+I+F	5.04	5888.45	0.04	-2901.23
LG+F	10.15	5893.56	0.00	-2904.78
RtREV+I+G+F	23.23	5906.65	0.00	-2909.32
RtREV+G+F	23.90	5907.31	0.00	-2910.65
Dayhoff+G+F	26.95	5910.37	0.00	-2912.18
RtREV+I+F	26.99	5910.40	0.00	-2912.20
DCMut+G+F	27.28	5910.69	0.00	-2912.34
Dayhoff+I+G+F	28.08	5911.49	0.00	-2911.75

## (b) MEGA models for amino acid substitution (13 globins)

Table. Maximum Likelihood fits of 48 different amino acid substitution models

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)	f(A)	f(R)	f(N)
WAG+G	24	4964.195	4836.725	-2393.968	n/a	5.07	0.087	0.044	0.039
WAG+I	24	4965.561	4838.092	-2394.652	0.03	n/a	0.087	0.044	0.039
WAG	23	4967.943	4845.755	-2399.515	n/a	n/a	0.087	0.044	0.039
WAG+G+I	25	4968.408	4835.661	-2392.403	0.02	7.77	0.087	0.044	0.039
Dayhoff+G+I	25	4970.283	4837.535	-2393.340	0.02	6.81	0.087	0.041	0.040
Dayhoff+G	24	4990.568	4863.098	-2407.155	n/a	4.99	0.087	0.041	0.040
JTT+G	24	5003.961	4876.492	-2413.852	n/a	5.25	0.077	0.051	0.043
JTT+I	24	5004.353	4876.884	-2414.048	0.03	n/a	0.077	0.051	0.043
JTT+G+I	25	5005.191	4872.444	-2410.795	0.03	6.48	0.077	0.051	0.043
Dayhoff+I	24	5013.028	4885.559	-2418.385	0.03	n/a	0.087	0.041	0.040

Stage 4: tree-building methods  
(distance-based; maximum parsimony;  
maximum likelihood; Bayesian methods)

# Stage 4: Tree-building methods

- We will discuss several tree-building methods:
  - UPGMA distance-based
  - Neighbor-joining distance-based
  - Maximum parsimony character-based
  - Maximum likelihood character-based (model-based)
  - Bayesian character-based (model-based)

## Stage 4: Tree-building methods

- Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score. Examples of distance-based algorithms are UPGMA and neighbor-joining.
- Character-based methods include maximum parsimony and maximum likelihood. Parsimony analysis involves the search for the tree with the fewest amino acid (or nucleotide) changes that account for the observed differences between taxa.

myoglobin\_kanga -----MGLSDGEWQLVLIWGVETDEGGHGKDVLIRLFKGHPETLEKFDKF  
 myoglobin\_harbo -----MGLSEGEWQLVLNWGVKVEADLAGHGQDVLIRLFKGHPETLEKFDKF  
 myoglobin\_gray\_ -----MGLSDGEWHLVLNWGVKETDLAGHGQEVLIRLFKSHPETLEKFDKF  
 alpha\_globin\_ho -----MV-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFGLGPTTKTYFPHF  
 alpha\_globin\_ka -----V-LSAADKGHVKAIWGVGGHAGEYAAEGLERTFHSFPTTKTYFPHF  
 alpha\_globin\_do -----V-LSPADKTNIKSTWDKIGGHAGDYGGAEALDRTFQSFPPTKTYFPHF  
 beta\_globin\_dog -----MVHLTAEEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF  
 beta\_globin\_rab -----MVHLSSEEKSAYTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF  
 beta\_globin\_kan -----VHLTAEEKNAITSLGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF  
 globin\_riverlam -PIVDS---GSPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFFPKF  
 globin\_sealampr MPIVDT---GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKF  
 globin\_soybean -----VAFTEKQDALVSSFEAFKANI PQYSVVFYTSILEKAPAAKDLFSFL  
 globin\_insect MKFLILALCFAAASALSADQISTVQASFDFKVKGD---PVGILYAVFKADPSIMAKFTQF

:: : : : . : \* \* \* :  
 ▼ ▼ ▼ ▼ ▼ ▼ o ◊ ▼ o ▼ ▼ ▼ ◊ o o o ▼

myoglobin\_kanga KHLKSEDEMKAEDLKHKGITVLTALGNILKKKGHHEAELKPLAQSKATKH  
 myoglobin\_harbo KHLKTEAEMKAEDLKHKGNTVLTALGGILKKKGHDAELKPLAQSKATKH  
 myoglobin\_gray\_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQSKATKH  
 myoglobin\_gray\_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQSKATKH  
 alpha\_globin\_ho -DLSHGSA----QVKAHGKKVGDALTAVGHLDDLPGALSNSSDL---HAHKLRVDPVN  
 alpha\_globin\_ka -DLSHGSA----QIQAHGKKIADALGQAVEHIDDLPGTLSKSDL---HAHKLRVDPVN  
 alpha\_globin\_do -DLSPGSA----QVKAHGKKVADALTAVAHLDLPGALSNSSDL---HAYKL  
 beta\_globin\_dog GDLSTPDAVMSNAVKAHGKKVLSNSDGLKNLDNLKGTFAKLSEL---HCDKLHVDPEN  
 beta\_globin\_rab GDLSSANAVMNNPKVKAHGKKVLAASFSEGLSHLDNLKGTFAKLSEL---HCDKLHVDPEN  
 beta\_globin\_kan GDLSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL---HCDKLHVDPEN  
 globin\_riverlam KGMTSADELKKSAI  
 globin\_sealampr KGLTTADQLKKS  
 globin\_soybean ANPTDG---VNPH  
 globin\_insect AG-KDLESIKGTAF

↗

**Distance-based tree**

Calculate the pairwise alignments:

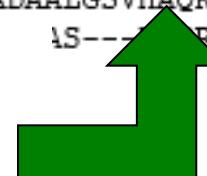
# Distance-based tree

# Calculate the pairwise alignments;

**if two sequences are related,**

put them next to each other on the tree

	▼▼▼▼▼▼▼▼▼▼	○		▼▼▼▼▼	○	○○○	◊
myoglobin_kanga	-----	MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF					
myoglobin_harbo	-----	MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF					
myoglobin_gray_	-----	MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF					
alpha_globin_ho	-----	MV-LSAADKTNVKAWSKVGGHAGEYGAELERMFLGFPPTKTYFPHF					
alpha_globin_ka	-----	V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHFSFPTTAKTYFPHF					
alpha_globin_do	-----	V-LSPADKTNIKSTWDKIGGHAGDYGGAEALDRTFQSFPPTKTYFPHF					
beta_globin_dog	-----	MVHLTAEEKSLVSGLGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF					
beta_globin_rab	-----	MVHLSSEEKSAVTALWGKV--NVEVGGEALGRLLVVYPWTQRFFESF					
beta_globin_kan	-----	VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTTSRFFDHF					
globin_riverlam	-PIVDS---	GSPAVLSAAEKTKIRSAWAPVSNYETSGVDILVKFFTSTPAAQEFPKF					
globin_sealampr	MPIVDT---	GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFPKF					
globin_soybean	-----	VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFL					
globin_insect	MKFLLIALCFAAASALSADQISTVQASFDFVKKGD---	PVGILYAVFKADPSIMAKFTQF					
	: : : : :	: * * :					
	▼ ▼	▼▼▼▼▼○◊		▼	○▼▼▼◊○○		▼
myoglobin_kanga	KHLKSEDEMKAEDLKKHGITVLTALGNILKKKGHHEAELKPLAQ	S---HATKHKIPVQF					
myoglobin_harbo	KHLKTEAEMKAEDLKKHGNTVLTALGGILKKKGHHDAELKPLAQ	S---HATKHKIPIKY					
myoglobin_gray_	KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQ	S---HATKHKIPIKY					
alpha_globin_ho	-DLSHGSA----QVKAHGKKVGDALTLAVGHLDLPGALSNLSDL	--HAHKLRVDPVN					
alpha_globin_ka	-DLSHGSA----QIQAHGKKIADALGQAVEHIDDLPGTLSKSDL	--HAHKLRVDPVN					
alpha_globin_do	-DLSPGSA----QVKAHGKKVADALTAVAHLDLPGALSALSDL	--HAYKLRVDPVN					
beta_globin_dog	GDLSTPDAVMSNAVKVAHGKKVLNSFSDGLKNLDNLKGTFAKLSEL	--HCDKLHVDPEN					
beta_globin_rab	GDLSSANAVMNNPKVKAHGKKVLAFAFSEGLSHLDNLKGTFAKLSEL	--HCDKLHVDPEN					
beta_globin_kan	GDLSSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL	--HCDKLHVDPEN					
globin_riverlam	KGMTSADELKKSAEVRWHAERIINAVNDAVASMDDTEKMSMK	--DLSGKHAKSFQVDPQY					
globin_sealampr	KGLTTADQLKKSAEVRWHAERIINAVNDAVASMDDTEKMSMCLRDLGKHAKSFQVDPQY						
globin_soybean	ANPTDG----VNPKLTGHAEKLFALVRDASAGQL-KASGTVVADAALGSVHAQKAVTNPEF						



Character-based tree: identify positions that best describe how characters (amino acids) are derived from common ancestors

# Definitions: identity, similarity, conservation

- **Homology**
  - Similarity attributed to descent from a common ancestor.
- **Identity**
  - The extent to which two (nucleotide or amino acid) sequences are invariant.
- **Similarity**
  - The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.
- **Conservation**
  - Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

# Main families of Methods for Phylogenetic reconstruction

		COMPUTATIONAL METHOD
		Optimality criterion      Clustering algorithm
DATA TYPE	Characters	PARSIMONY MAXIMUM LIKELIHOOD BAYES INFERENCE
	Distances	MINIMUM EVOLUTION LEAST SQUARES

UPGMA  
NEIGHBOR- JOINING  
FITCH & MARGOLIASH

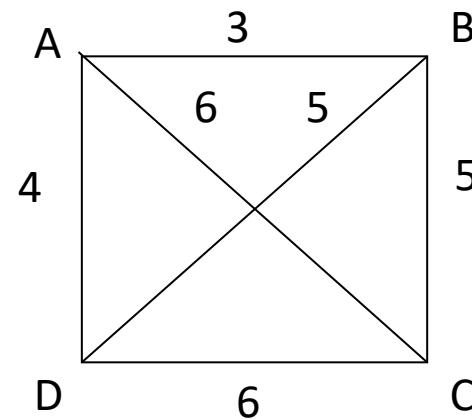
# Construct Tree Algorithm

- Heuristic Algorithm
  - UPGMA
  - Neighbor-Joining

# Ultrametric matrix

- A distance matrix is a *metric* if the distance obey the triangular inequality
  - $M[i, j] + M[j, k] \geq M[i, k]$
- A metric  $M$  is an *ultrametric* if and only if  $M[i, j] \leq \max\{M[i, k], M[j, k]\}$

A	B	C	D	
0	3	6	4	A
0	5	5	B	
0	6	C		
0	D			



# Construct Evolution Tree from Distance Matrix

- In the evolution tree, let
  - $dt(s_i, s_j)$  denote the distance between species  $s_i$  and  $s_j$ .
  - $d(s_i, s_j)$  denote the distance between  $s_i$  and  $s_j$  in the distance matrix.

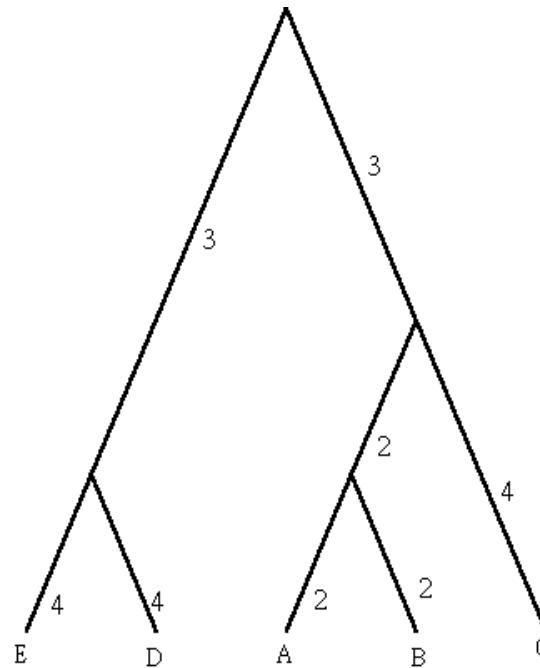
$$dt(s_i, s_j) \geq d(s_i, s_j)$$

## Construct Evolution Tree from Distance Matrix (con't)

- Depending upon different conditions, different evolution tree will be constructed.
- Minimizing
  - Minimax : the maximum of( $dt(s_i, s_j) - d(s_i, s_j)$  )
  - Minisum :  $\sum_{i,j} dt(i, j)$
  - Minisize :  $\sum_{e \in E(T)} w(e)$

## Example for construct Evolution Tree from Distance Matrix

A	B	C	D	E	
0	2	4	8	8	A
0	4	8	8	8	B
0	8	8	8	8	C
0	6	6	14	14	D
0	14	14	14	14	E



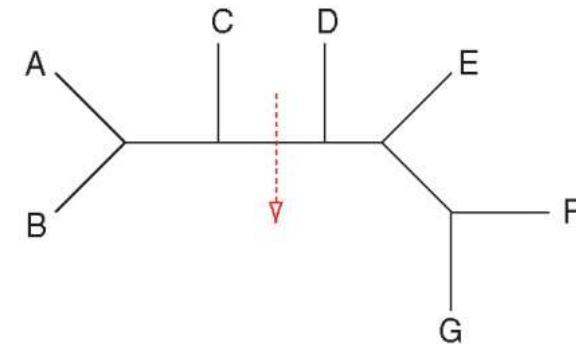
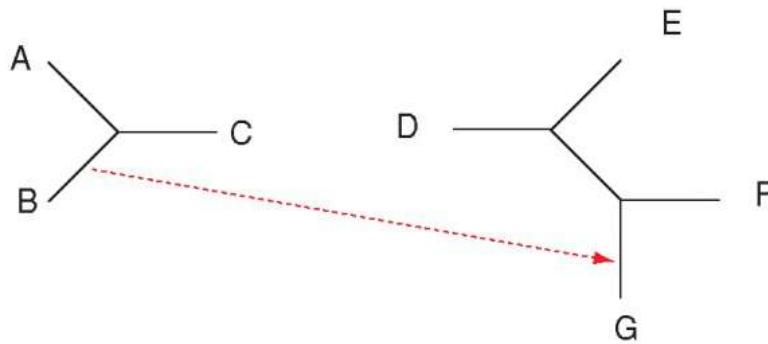
$$\begin{aligned}
 \text{max:} & dt(C,D) - d(C,D) = 6 \\
 \text{sum:} & 4+8+8+14+14+14 \\
 & +14+14+14+8=112 \\
 \text{size:} & 3+4+4+3+2+2+2+4 \\
 & =24
 \end{aligned}$$

## The Complexities of Evolution Tree Problems

	Minimax	Minisum	minisize
Unrooted	NP-Completer	NP-Complete	?
Rooted	$O(n^2)$	NP-Complete	NP-Complete

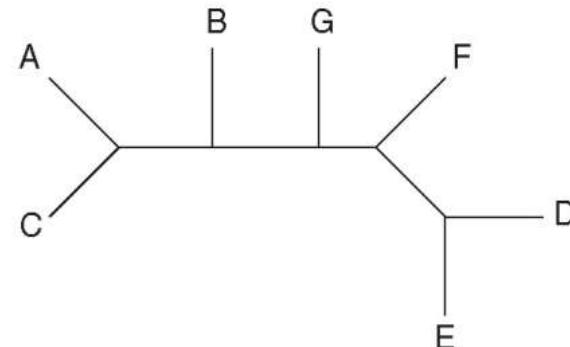
# Finding optimal trees: branch swapping

Bisect a branch to form two subtrees



Reconnect via one branch from each subtree; evaluate each bisection

Identify the optimal tree(s)

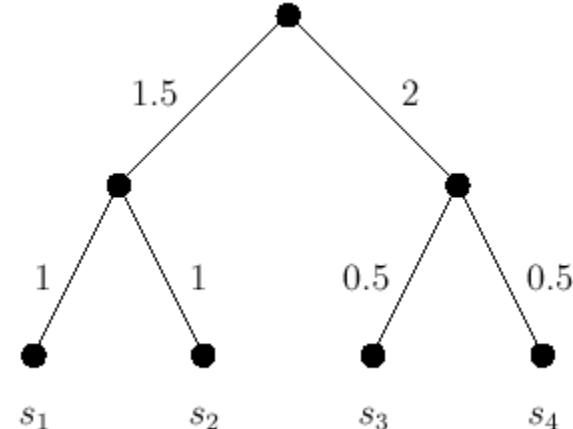
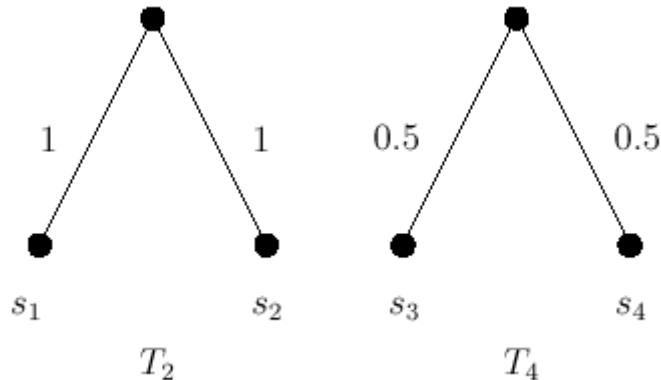
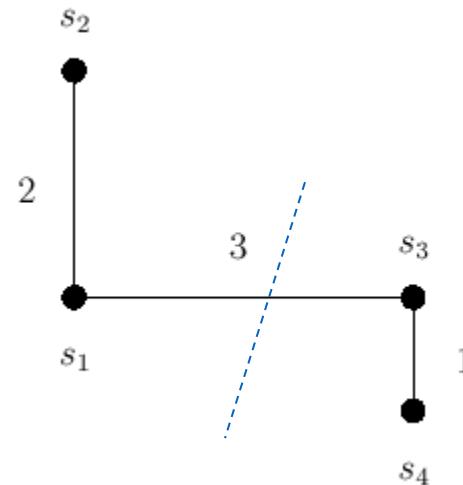


# A Minimax Rooted Evolution Tree Algorithm

Distance Matrix

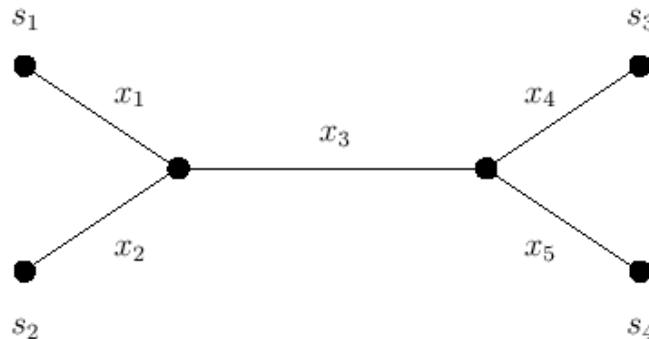
	$s_1$	$s_2$	$s_3$	$s_4$
$s_1$	0	2	3	3.1
$s_2$		0	3.6	5
$s_3$			0	1
$s_4$				0

Minimal Spanning Tree



# When the evolution tree topology is given!

- Take minisize evolution tree for example :



Unrooted Tree Topology

$$\text{Minimize } x_1 + x_2 + x_3 + x_4 + x_5$$

$$x_1 + x_2 \geq d_{12}$$

$$x_1 + x_3 + x_4 \geq d_{13}$$

Subject to

$$x_1 + x_3 + x_5 \geq d_{14}$$

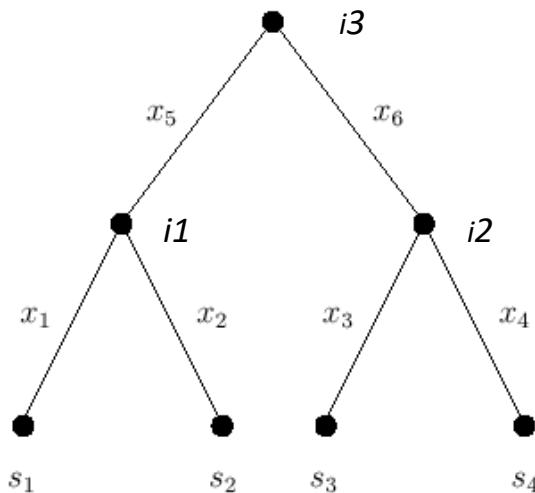
$$x_2 + x_3 + x_4 \geq d_{23}$$

$$x_2 + x_3 + x_5 \geq d_{24}$$

$$x_4 + x_5 \geq d_{34}$$

Linear Programming

## Minimal Ultrametric Tree with a given Topology



$$\text{height}(s) = \max\{\mathcal{M}[i,j] \mid i, j \in L(T_s)\}/2$$

$$\text{height}(i_1) = \mathcal{M}[s_1, s_2]/2$$

$$\text{height}(i_2) = \mathcal{M}[s_3, s_4]/2$$

$$\text{height}(i_3) = \max\{\mathcal{M}[s_1, s_3], \mathcal{M}[s_1, s_4], \mathcal{M}[s_2, s_3], \mathcal{M}[s_2, s_4], \mathcal{M}[s_1, s_2], \mathcal{M}[s_3, s_4]\} / 2$$

Unrooted Tree Topology

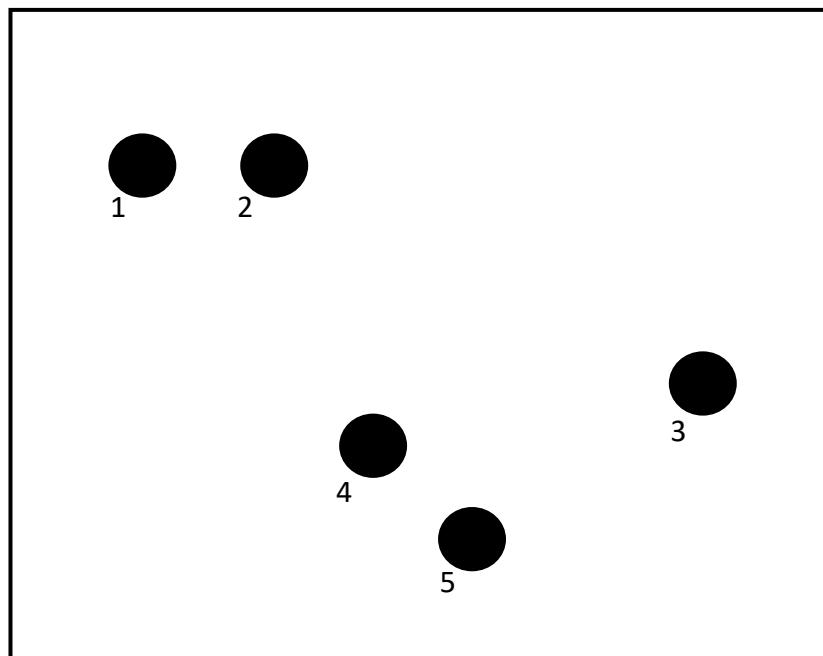
By the postorder traversal of the tree, the heights of all the nodes can be computed in  $O(n^2)$  time.

# Unweighted Pair-Group Method with Arithmetic Mean

UPGMA

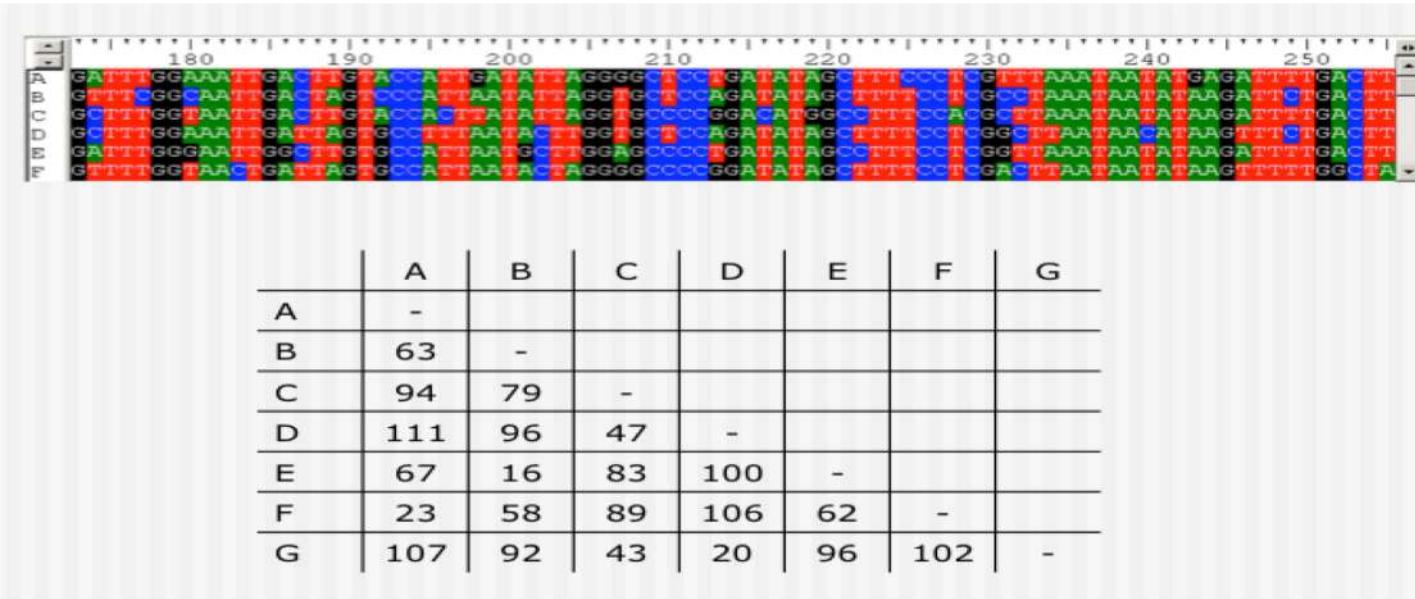
# Tree-building methods: UPGMA

- unweighted pair group method using arithmetic mean



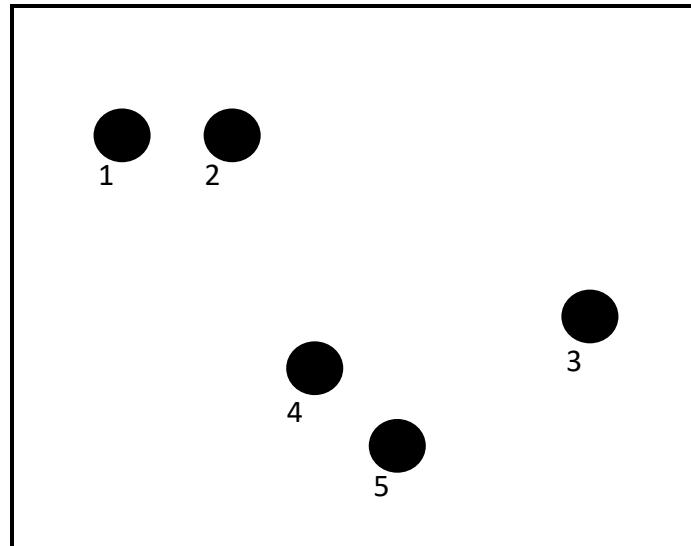
# Unweighted Pair-Group Method with Arithmetic Mean (UPGMA)

- Very Unreliable Method
- Very Simple Principle



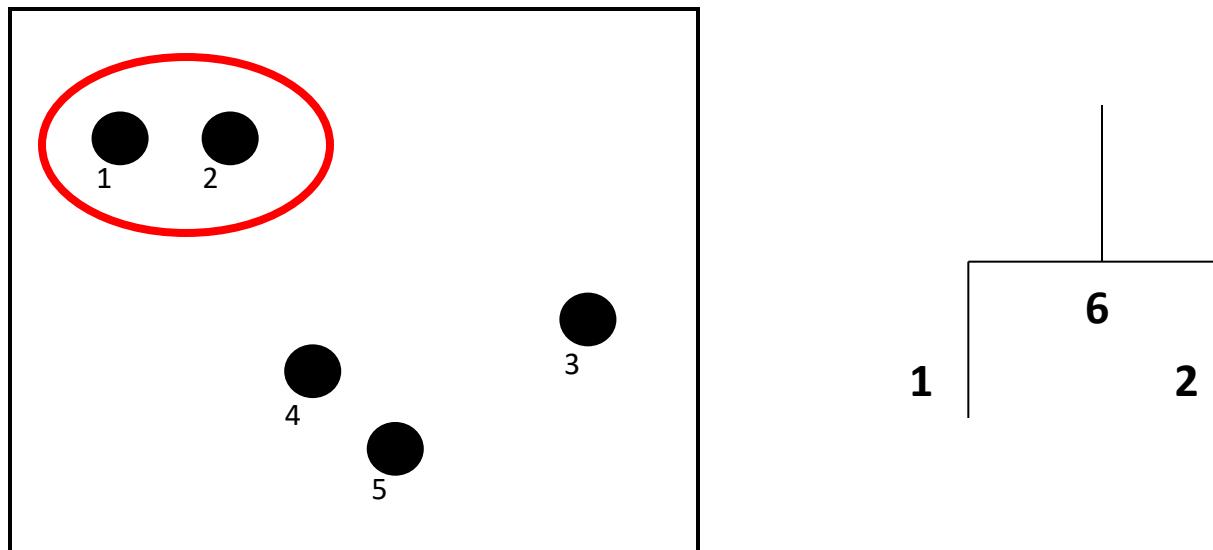
# Tree-building methods: UPGMA

- Step 1: compute the pairwise distances of all the proteins.  
Get ready to put the numbers 1-5 at the bottom of your new tree.



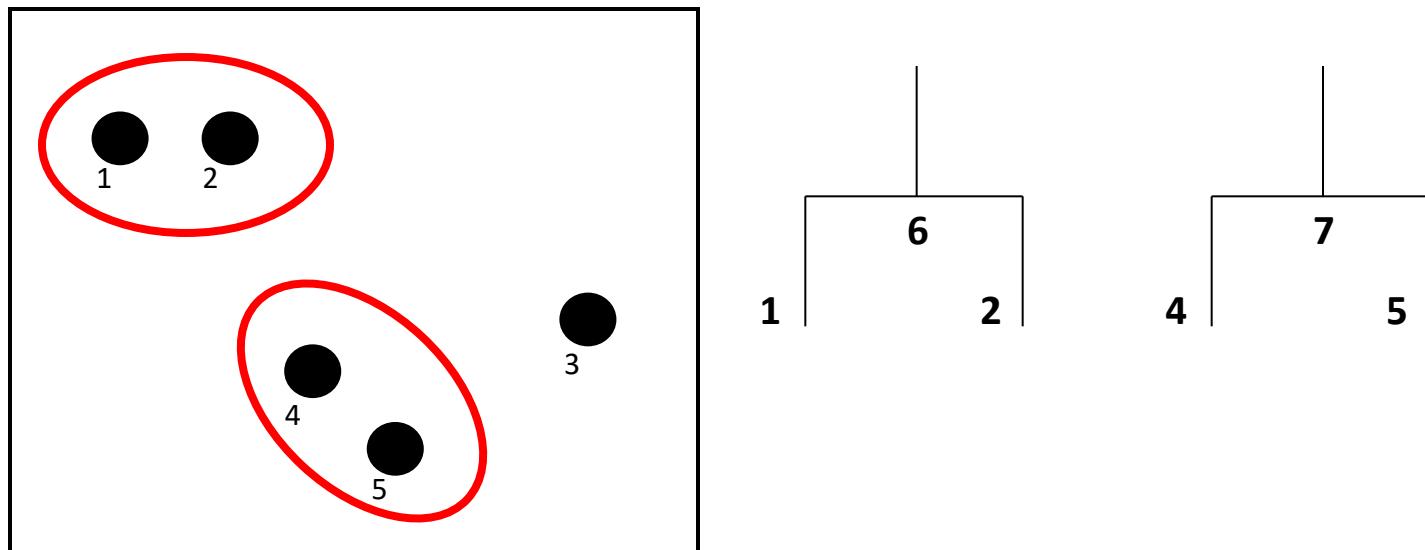
# Tree-building methods: UPGMA

- Step 2: Find the two proteins with the smallest pairwise distance. Cluster them.



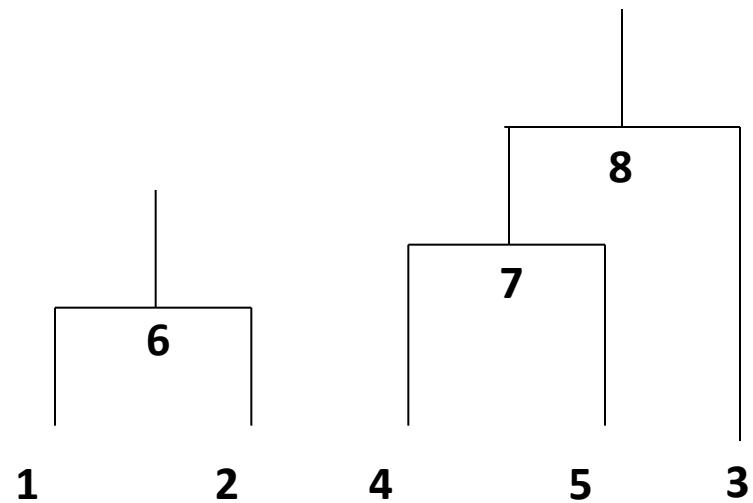
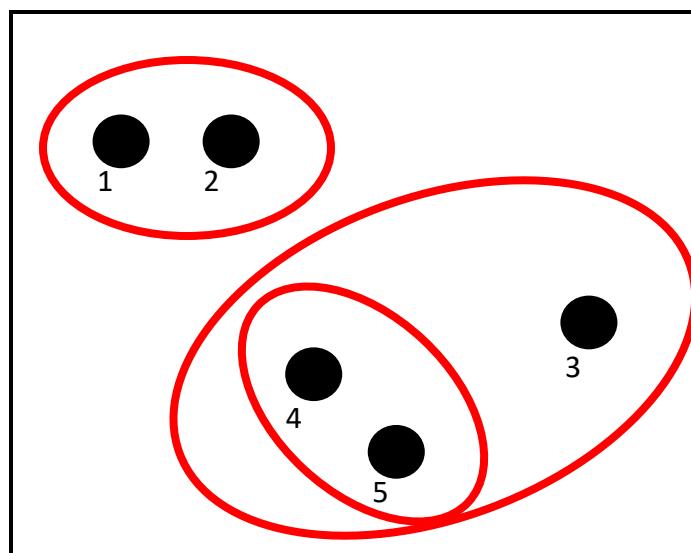
# Tree-building methods: UPGMA

- Step 3: Do it again. Find the next two proteins with the smallest pairwise distance. Cluster them.



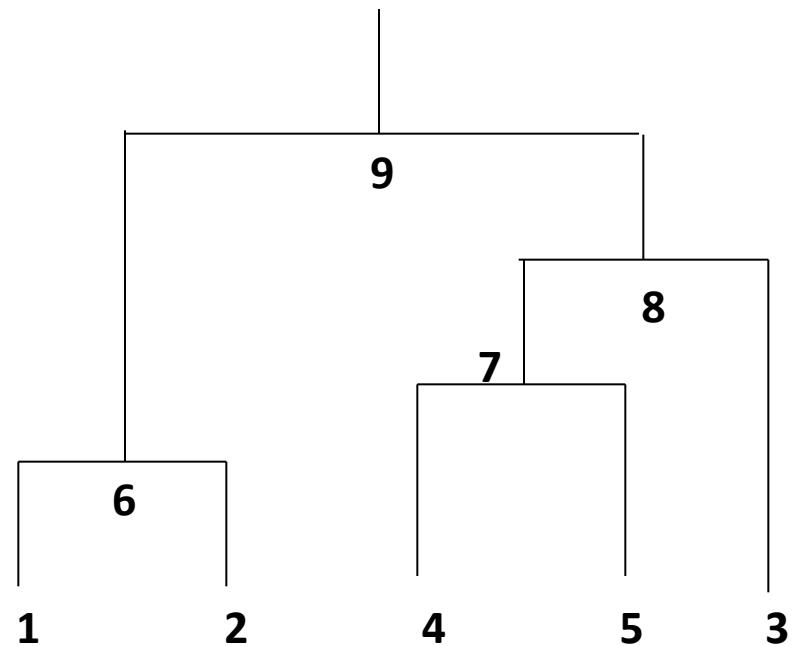
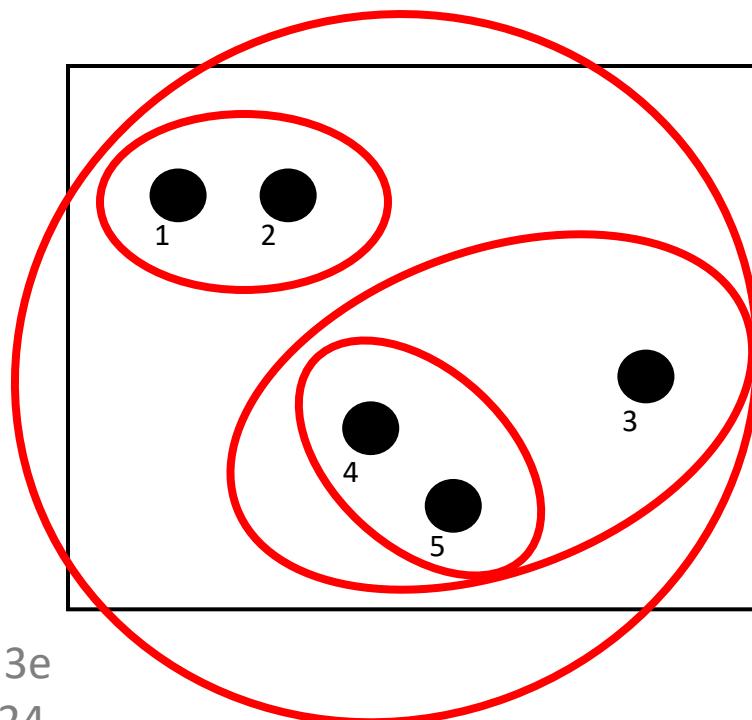
# Tree-building methods: UPGMA

- Step 4: Keep going. Cluster.



# Tree-building methods: UPGMA

- Step 4: Last cluster! This is your tree.

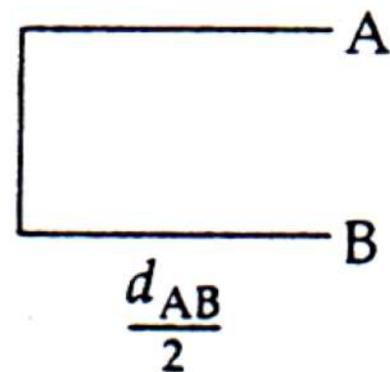


# UPGMA

If  $d_{AB}$  has the smallest value, A and B are treated as a single OTU

OTU	OTU		
	A	B	C
B	$d_{AB}$		
C	$d_{AC}$	$d_{BC}$	
D	$d_{AD}$	$d_{BD}$	$d_{CD}$

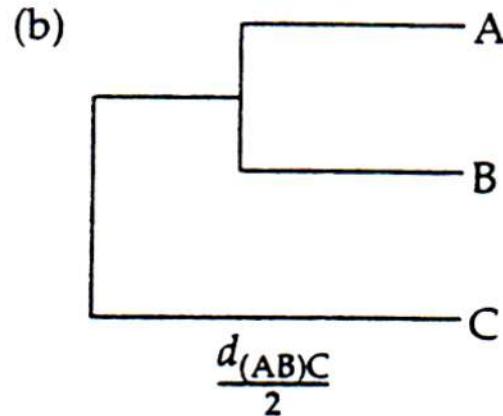
(a)



## UPGMA (con't)

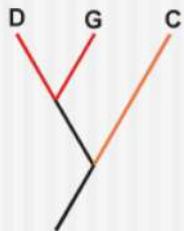
- $d_{(AB)C} = (d_{AC} + d_{BC})/2$ ,  $d_{(AB)D} = (d_{AD} + d_{BD})/2$
- If  $d_{(AB)C}$  has the smallest value
- Repeat

OTU	OTU	
	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	$d_{CD}$

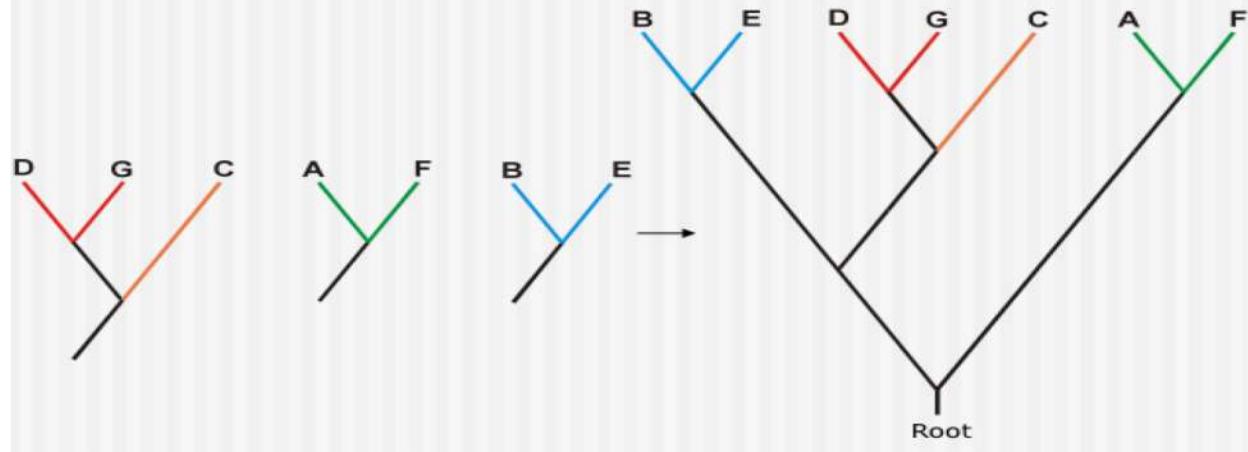


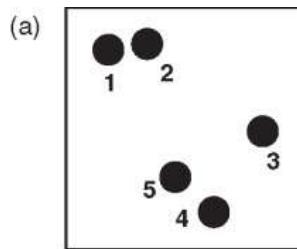
# UPGMA

	A	B	C	E	F	DG
A	-					
B	63	-				
C	94	79	-			
E	67	16	83	-		
F	23	58	89	62	-	
DG	94	84	35	88	94	-

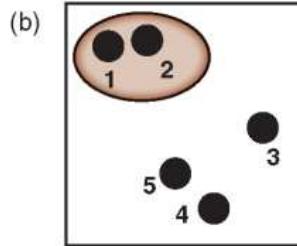


	AF	BE	CDG
AF	-		
BE	188	-	
CDG	112	108	-

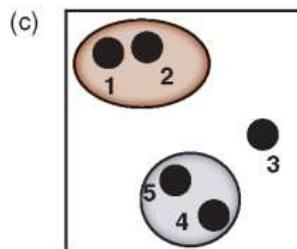
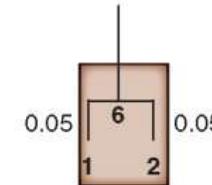




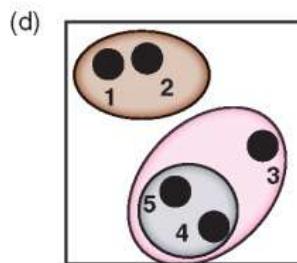
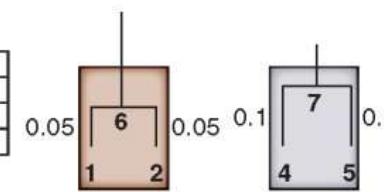
	1	2	3	4	5
1	—				
2	0.1	—			
3	0.8	0.8	—		
4	0.8	1	0.3	—	
5	0.9	0.9	0.3	0.2	—



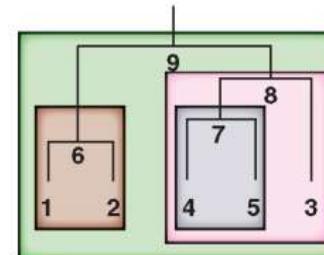
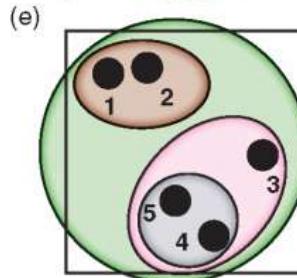
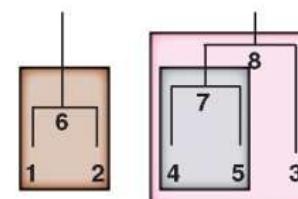
	(1,2)	3	4	5
(1,2)	—			
3	0.8	—		
4	0.9	0.3	—	
5	0.9	0.3	0.2	—



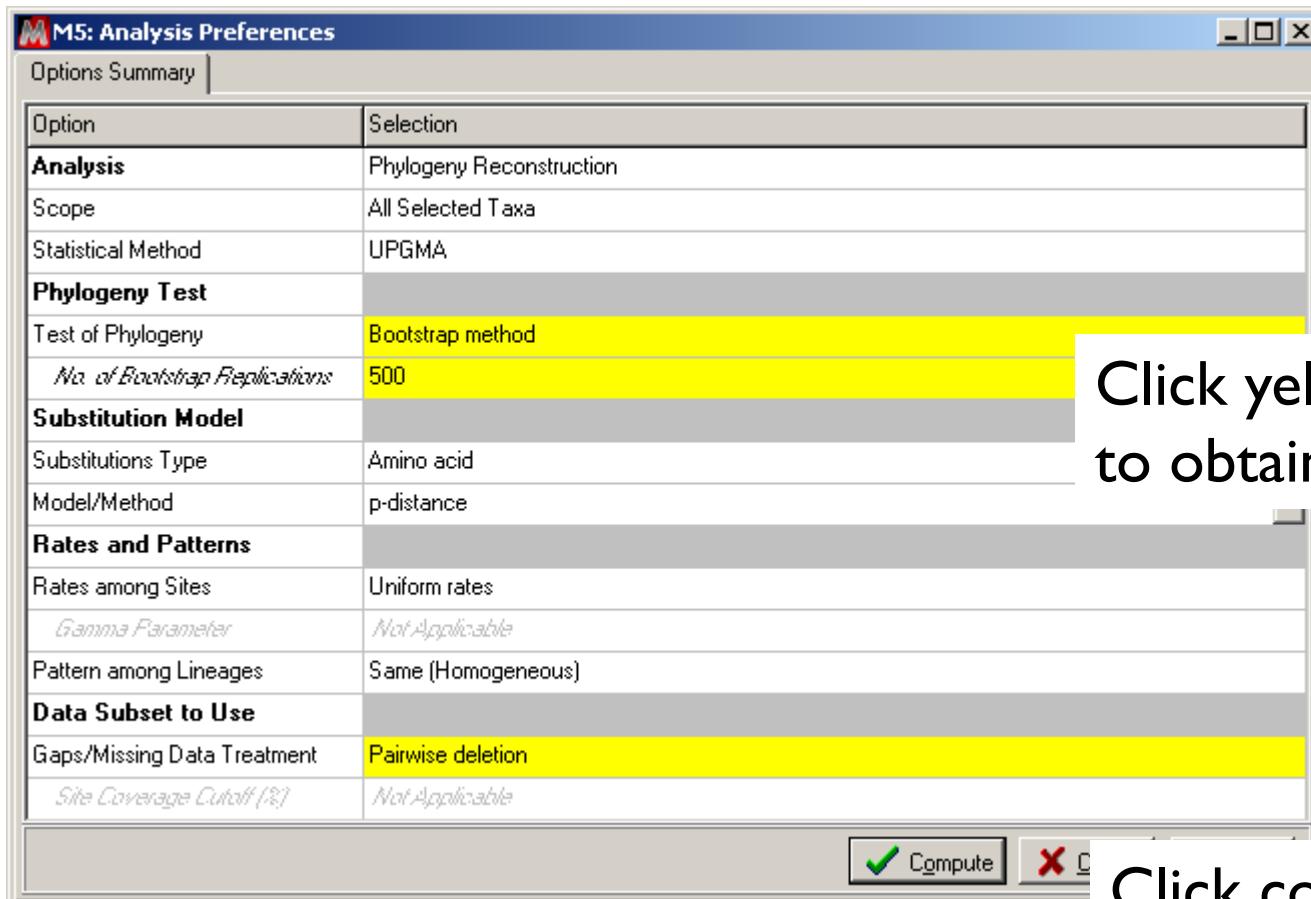
	(1,2)	3	(4,5)
(1,2)	—		
3	0.8	—	
(4,5)	0.9	0.3	—



	(1,2)	[3,(4,5)]
(1,2)	—	
[3,(4,5)]	0.85	—



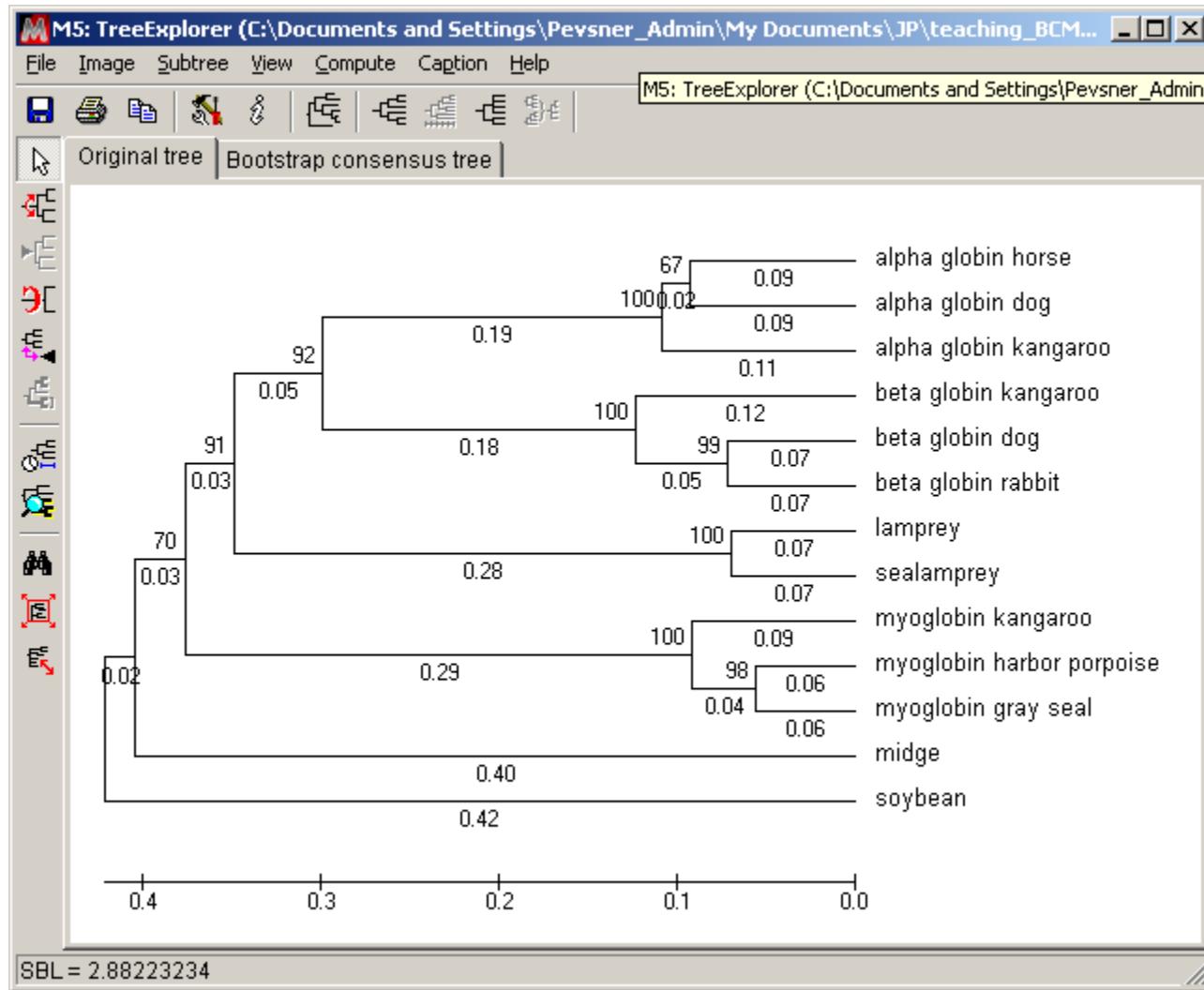
Use of MEGA for a distance-based tree: UPGMA  
(an easy method to explain, but not accurate for most purposes)



Click yellow rows  
to obtain options

Click compute  
to obtain tree

# Use of MEGA for a distance-based tree: UPGMA



# Use of MEGA for a distance-based tree: UPGMA

- Flipping branches around a node creates an equivalent topology



# MEGA provides captions summarizing methods



## Figure. Evolutionary relationships of taxa

The evolutionary history was inferred using the UPGMA method [1]. The optimal tree with the sum of branch length = 2.88223234 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method [3] and are in the units of the number of amino acid differences per site. The analysis involved 13 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 171 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [4].

1. Sneath P.H.A. and Sokal R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.
2. Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.
3. Nei M. and Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
4. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* (In Press).

# Distance-based methods: UPGMA trees

- UPGMA is a simple approach for making trees.
- An UPGMA tree is always rooted.
- An assumption of the algorithm is that the molecular clock is constant for sequences in the tree. If there are unequal substitution rates, the tree may be wrong.
- While UPGMA is simple, it is less accurate than the neighbor-joining approach (described next).

# Neighbor-Joining

Saitou, N. & Nei, M. *Mol. Biol. Evol.* 4, 406–425 (1987).

Another field buoyed by the growth in genome sequencing is phylogenetics, the study of evolutionary relationships between species.

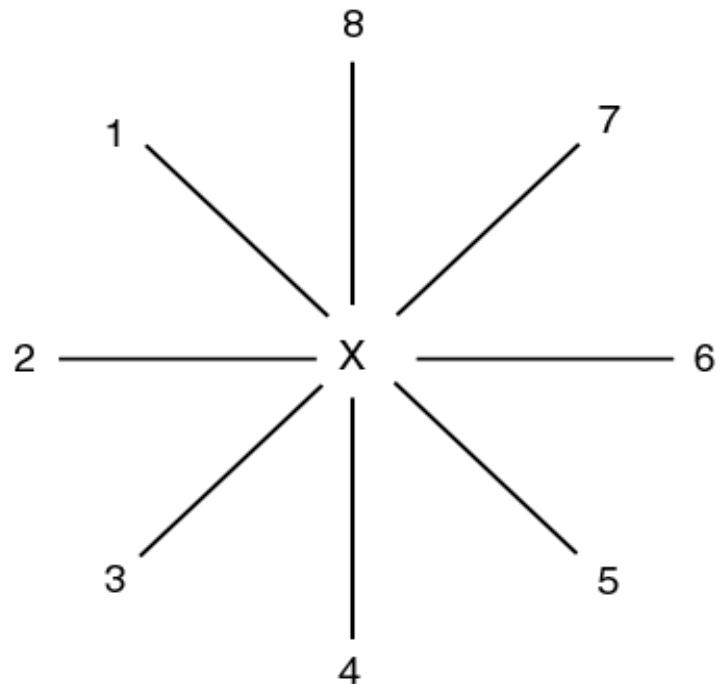
“We physical anthropologists were facing kind of the big data of that time,” says Saitou, now at Japan’s National Institute of Genetics in Mishima. The technique made it possible to devise trees from large data sets without eating up computer resources. (And, in a nice cross-fertilization within the top-100, Clustal’s algorithms use the same strategy.)

Saitou, N. & Nei, M. *Mol. Biol. Evol.* 4, 406–425 (1987).

Number 20 on the list is a paper<sup>12</sup> that introduced the “neighbor-joining” method, a fast, efficient way of placing a large number of organisms into a phylogenetic tree according to some measure of evolutionary distance between them, such as genetic variation. It links related organisms together one pair at a time until a tree is resolved. Physical anthropologist Naruya Saitou helped to devise the technique when he joined Masatoshi Nei’s lab at the University of Texas in Houston in the 1980s to work on human evolution and molecular genetics, two fields that were starting to burst at the seams with information.

# Making trees using neighbor-joining

- The neighbor-joining method of Saitou and Nei(1987) is especially useful for making a tree having a large number of taxa.
- Begin by placing all the taxa in a star-like structure.

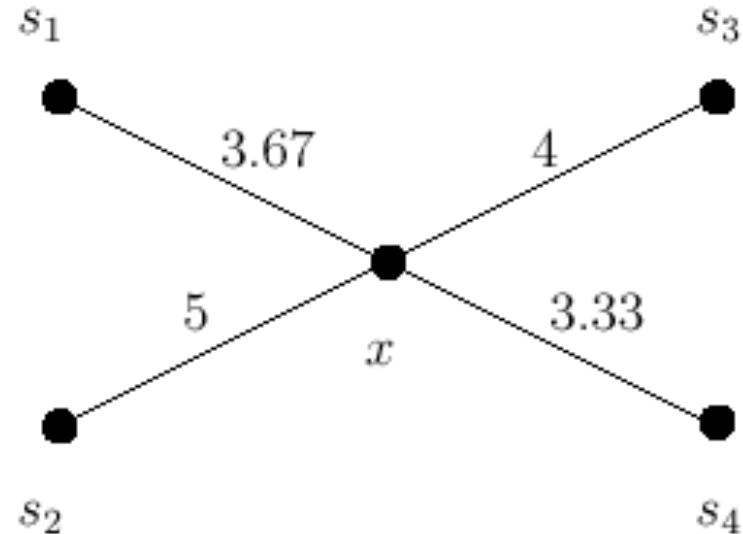


# The Neighbor Joining Method for Unrooted Tree

- A 1-star to initiate the Neighbor Joining method
- i.e.,

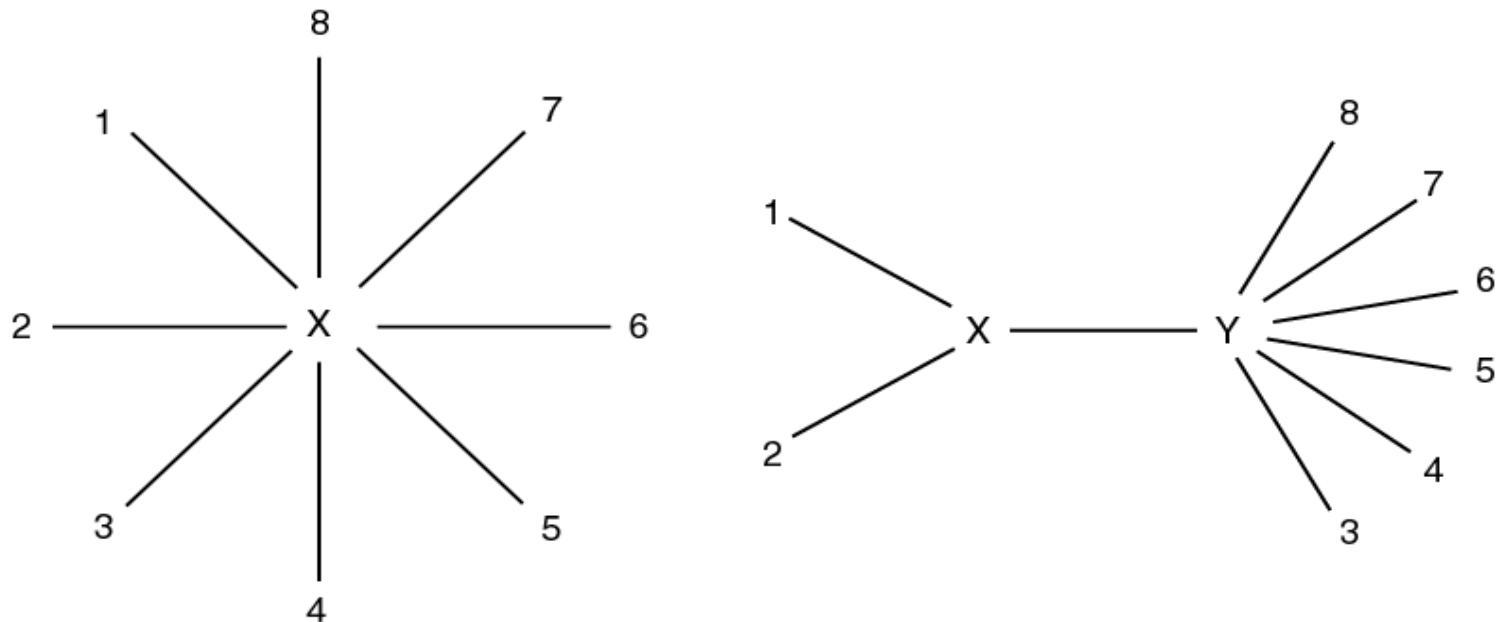
$$W(x, s_1) = \frac{1}{3} (w(s_1 + s_2) + w(s_1 + s_3) + w(s_1 + s_4)) = \frac{1}{3} (4 + 4 + 3) \\ = 3.67$$

	$s_1$	$s_2$	$s_3$	$s_4$
$s_1$	0	4	4	3
$s_2$		0	6	5
$s_3$			0	2
$s_4$				0



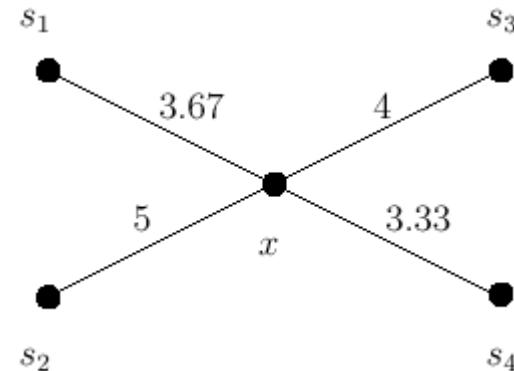
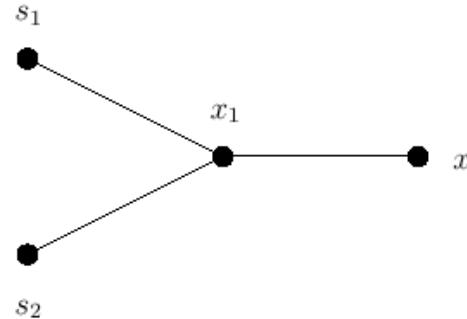
# Making trees using neighbor-joining

- Next, identify neighbors (e.g. 1 and 2) that are most closely related. Connect these neighbors to other OTUs via an internal branch, XY. At each successive stage, minimize the sum of the branch lengths.



# The Neighbor Joining Method for Unrooted Tree

- If  $s_1$  and  $s_2$  are to be paired,



- The New Connection Cost

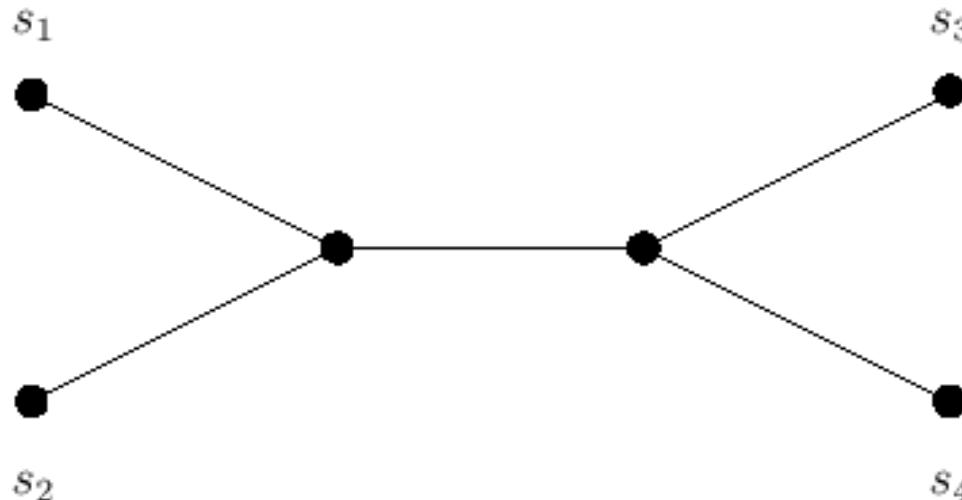
$$NC = \frac{1}{2}(\text{average}(s_1) + \text{average}(s_2) + d(s_1, s_2)) = \frac{1}{2}(3.67 + 5 + 4) = 6.33$$

- The Old Connection Cost

$$OC = (\text{average}(s_1) + \text{average}(s_2)) = 3.67 + 5 = 8.67$$

# The Neighbor Joining Method for Unrooted Tree

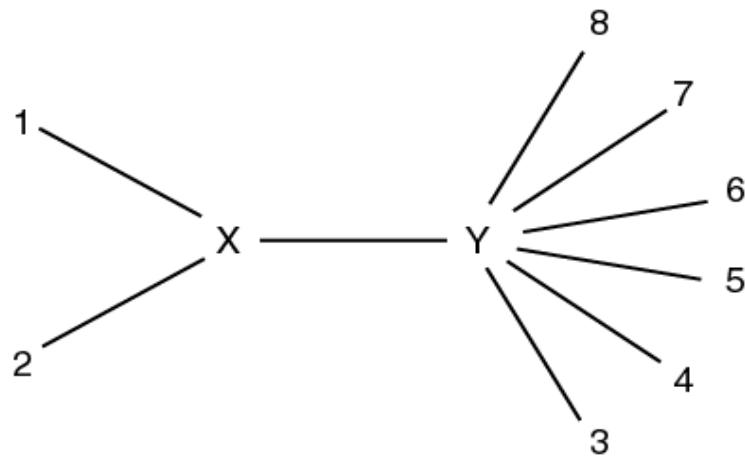
- The cost save by pair  $s_1$  and  $s_2$  =  $OC-NC = 8.67 - 6.33 = 2.34$
- The pairing of  $s_3$  and  $s_4$  produces the largest cost saving.  
Thus we pair  $s_3$  with  $s_4$ .



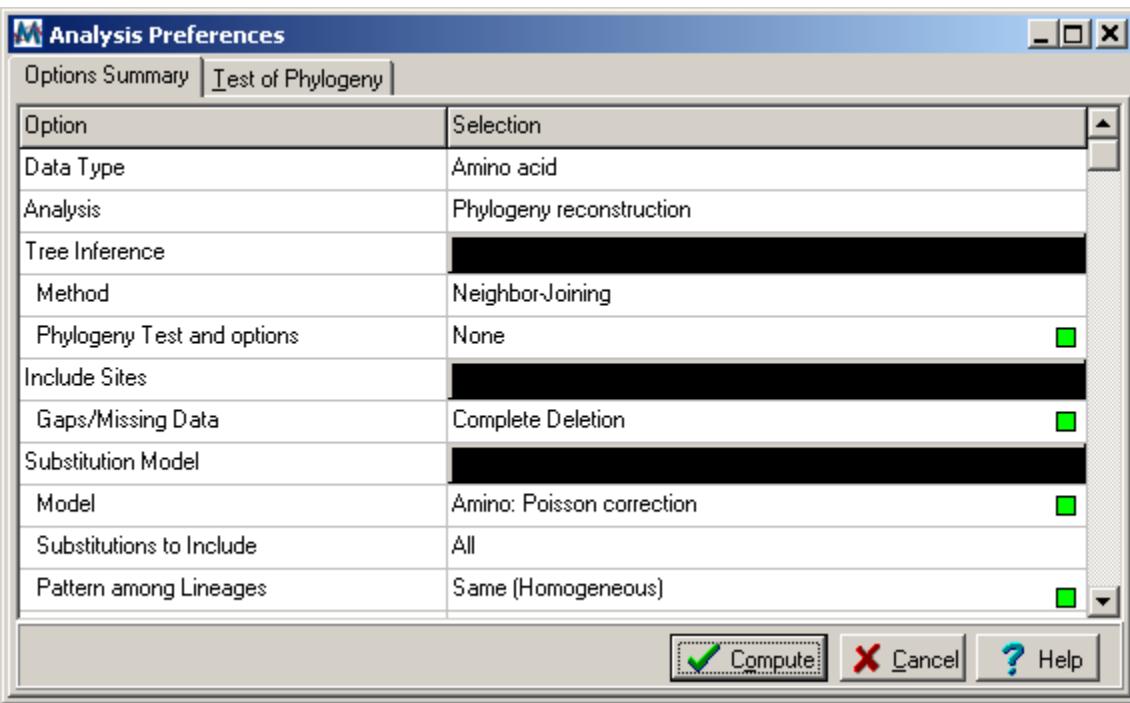
# Making trees using neighbor-joining

- Define the distance from X to Y by

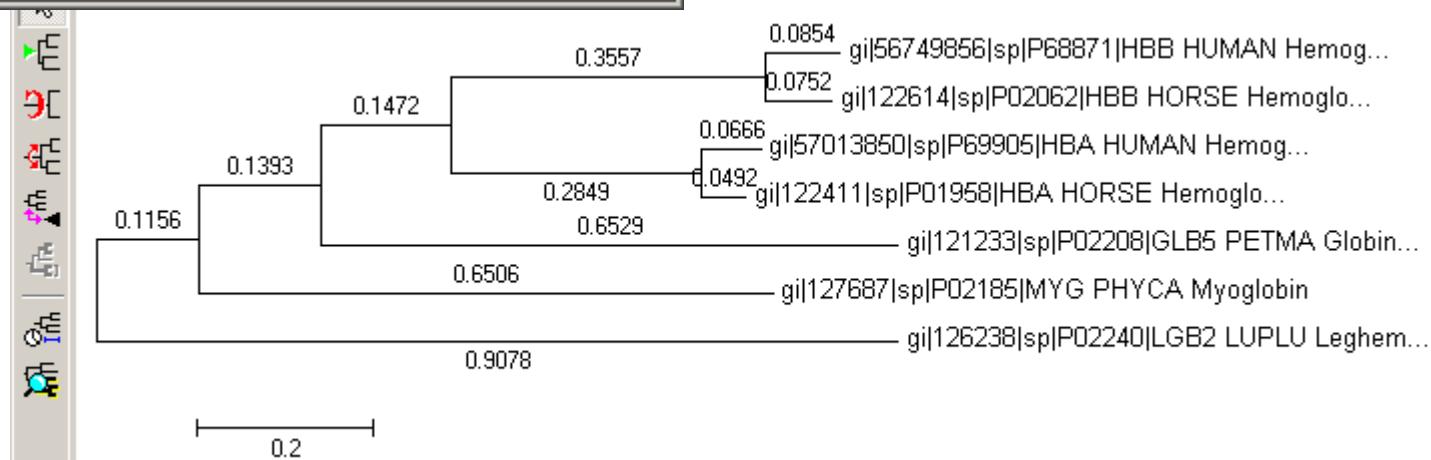
$$d_{XY} = 1/2(d_{1Y} + d_{2Y} - d_{12})$$



# Use of MEGA for a distance-based tree: NJ



Neighbor-joining produces a reasonably similar tree as UPGMA. It is fast, and commonly used (especially for large numbers of sequences).



# Maximum Parsimony Method

# Tree-building methods: character based

- Rather than pairwise distances between proteins, evaluate the aligned columns of amino acid residues (characters).
- Tree-building methods based on characters include maximum parsimony and maximum likelihood.

# Tree-building methods: character based

- The main idea of maximum parsimony is to find the tree with the shortest branch lengths possible. Thus we seek the most parsimonious (“simple”) tree.
- Identify informative sites. For example, constant characters are not parsimony-informative.
- Construct trees, counting the number of changes required to create each tree.
  - ~ 12 taxa or fewer, evaluate all possible trees exhaustively
  - >12 taxa perform a heuristic search.
- Select the shortest tree (or trees).

As an example of tree-building using maximum parsimony

- consider these four taxa:

**AAG**

**AAA**

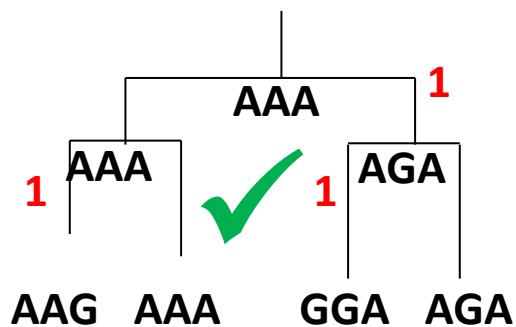
**GGA**

**AGA**

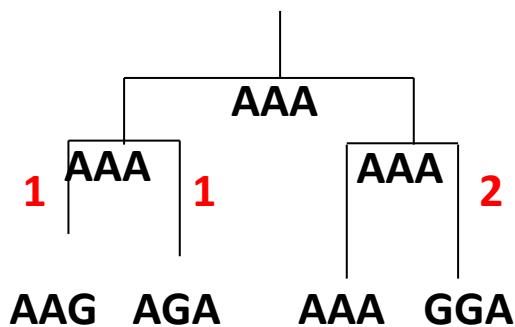
- How might they have evolved from a common ancestor such as AAA?

# Tree-building methods: Maximum parsimony

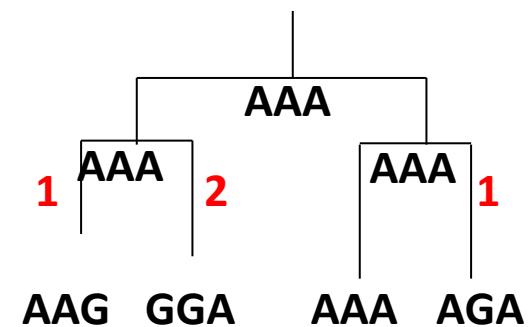
- In maximum parsimony, choose the tree(s) with the lowest cost (shortest branch lengths).



Cost = 3



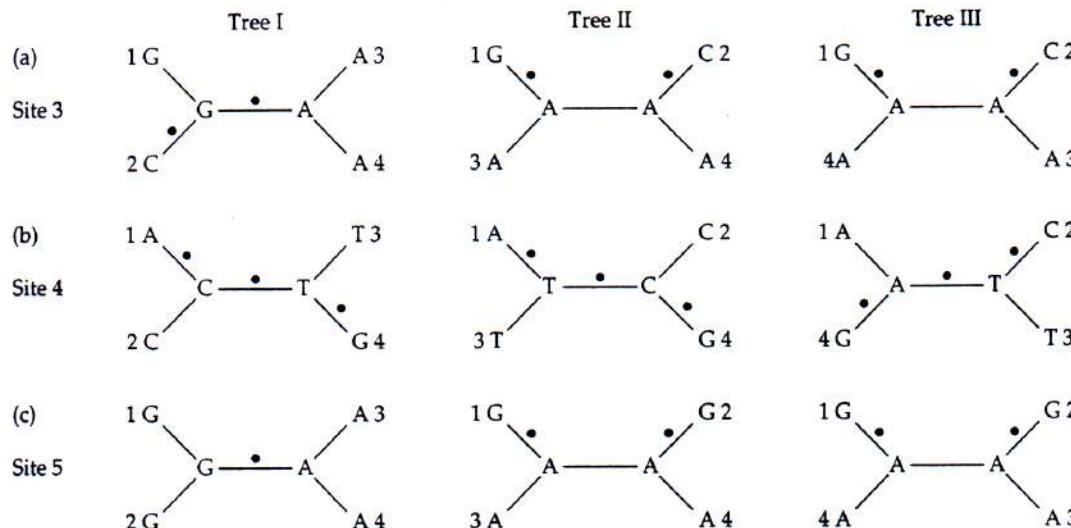
Cost = 4



Cost = 4

# Maximum Parsimony Method

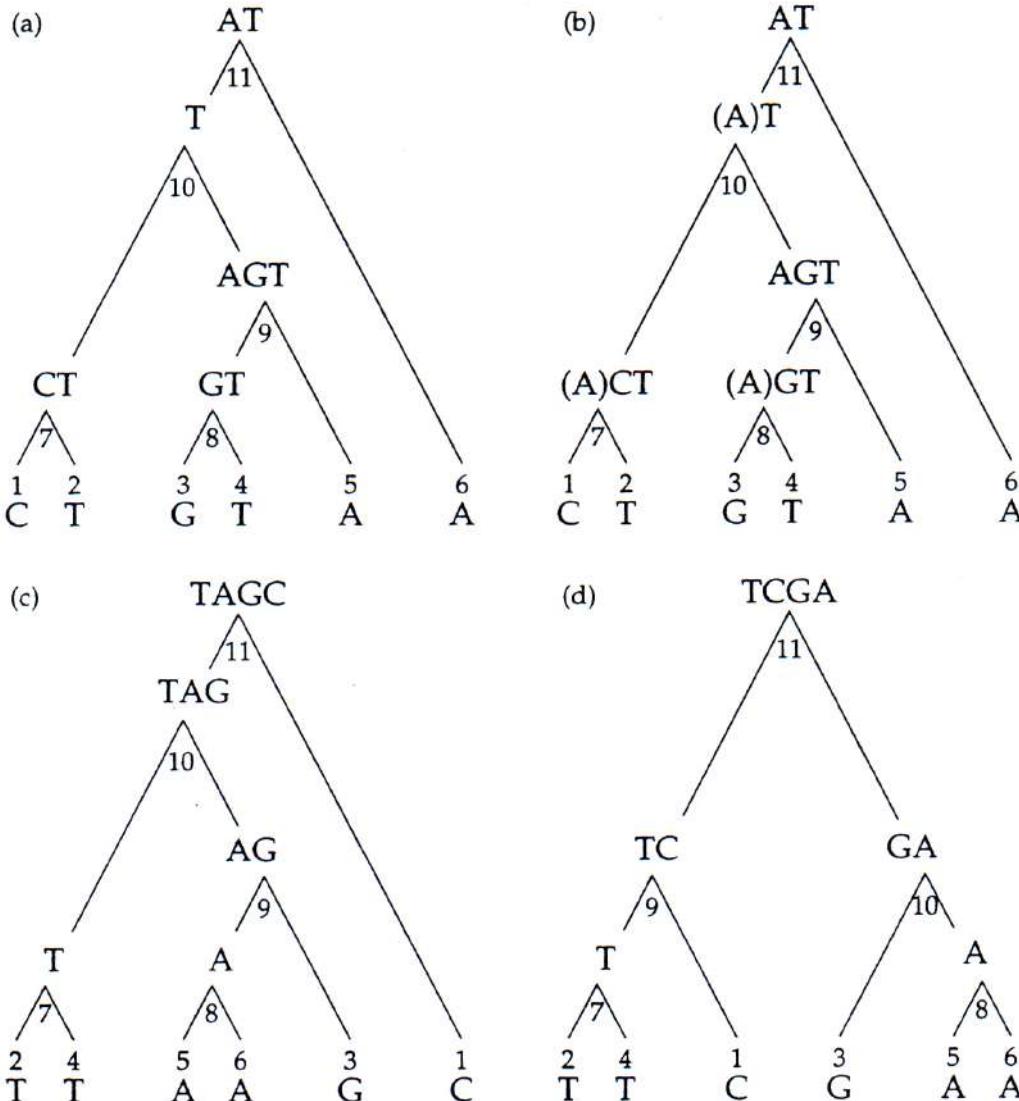
Sequence	Site									Site 5,7 and 9 are <i>informative site</i>
	1	2	3	4	5	6	7	8	9	
1	A	A	G	A	G	T	G	C	A	
2	A	G	C	C	G	T	G	C	G	
3	A	G	A	T	A	T	C	C	A	
4	A	G	A	G	A	T	C	C	G	*
				*		*	*			*



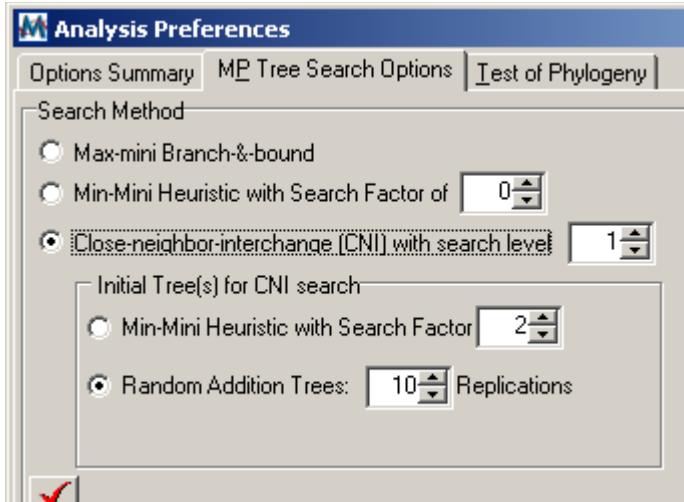
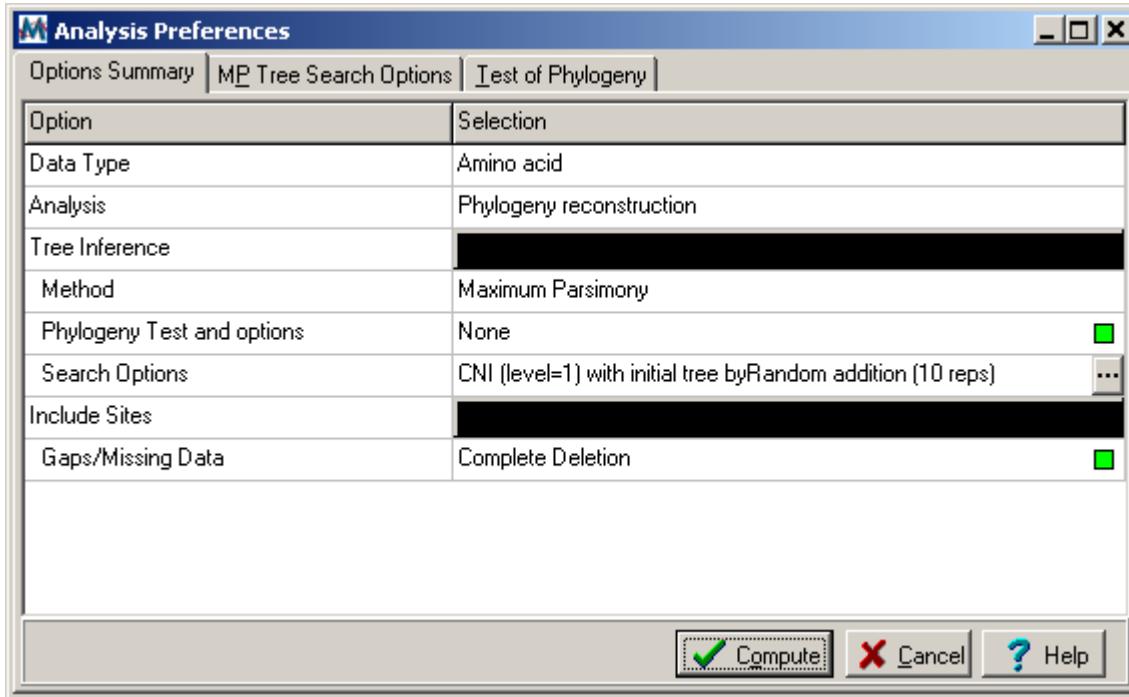
Each three possible trees require 2 changes but Tree I requires 1 change

Select the Tree supported by the *largest number of Informative Site*

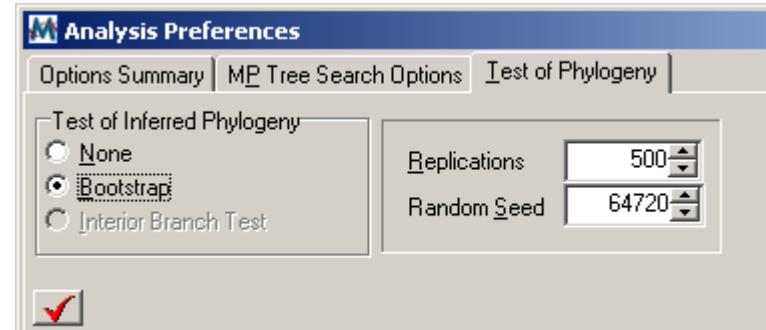
# Maximum Parsimony Method(con't)



# MEGA for maximum parsimony (MP) trees

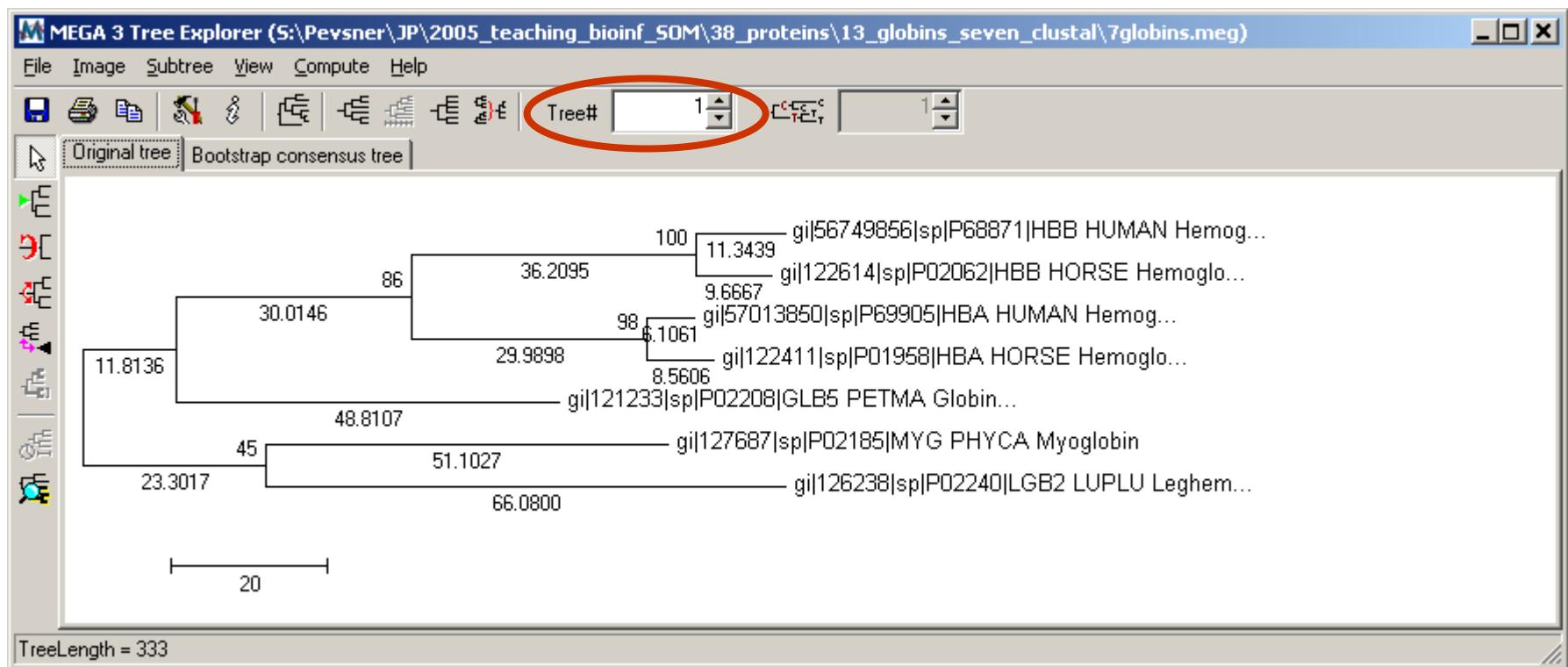


Options include heuristic approaches,  
and bootstrapping

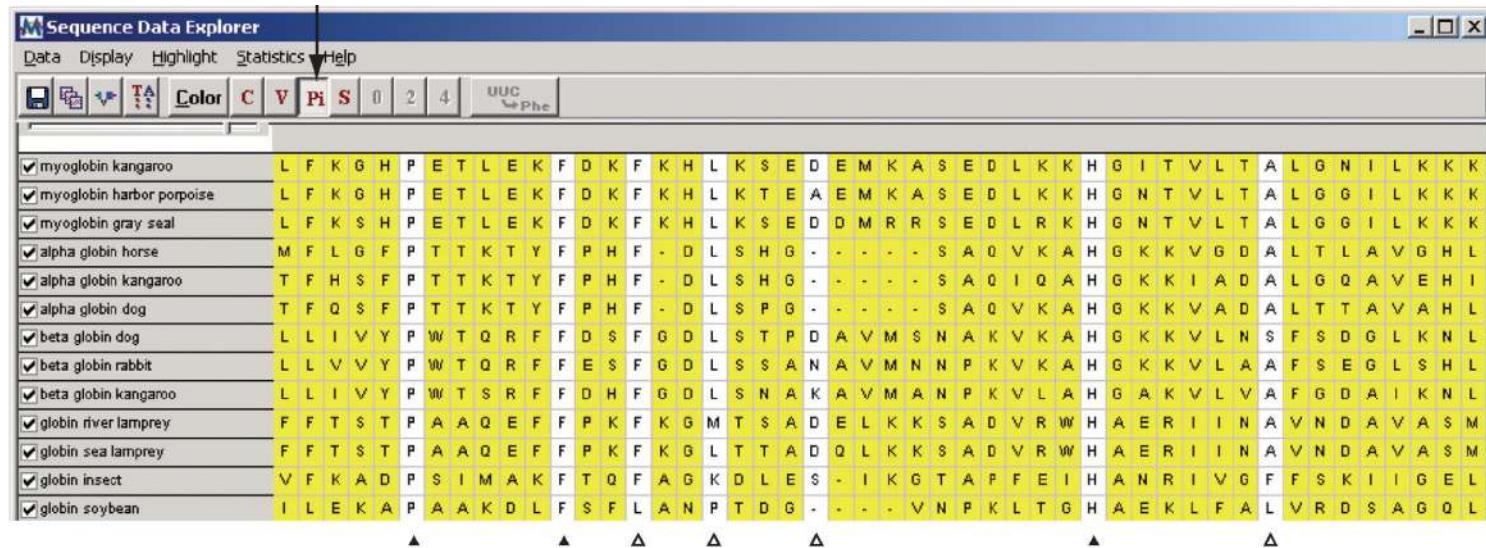


# MEGA for maximum parsimony (MP) trees

- In maximum parsimony, there may be more than one tree having the lowest total branch length. You may compute the consensus best tree.



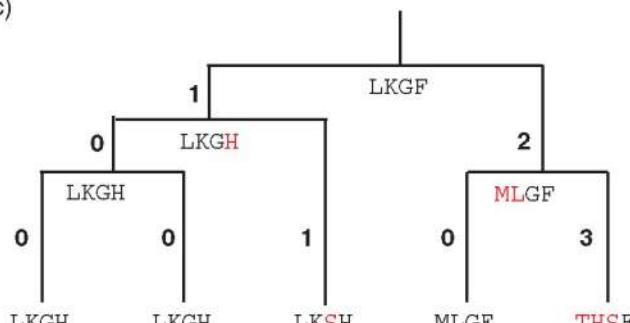
# MEGA displays parsimony-informative sites



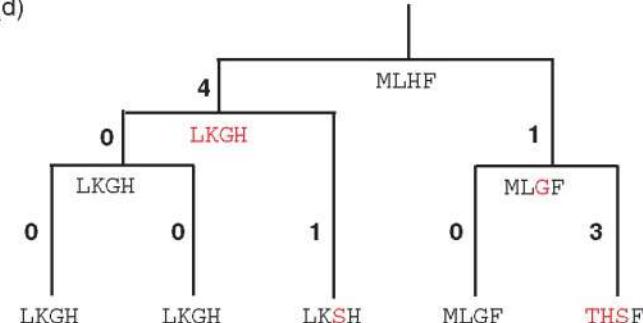
(b)

kangaroo	LKGH
porpoise	LKGH
gray seal	LKSH
horse $\alpha$	MLGF
kangaroo $\alpha$	THSF

(c)

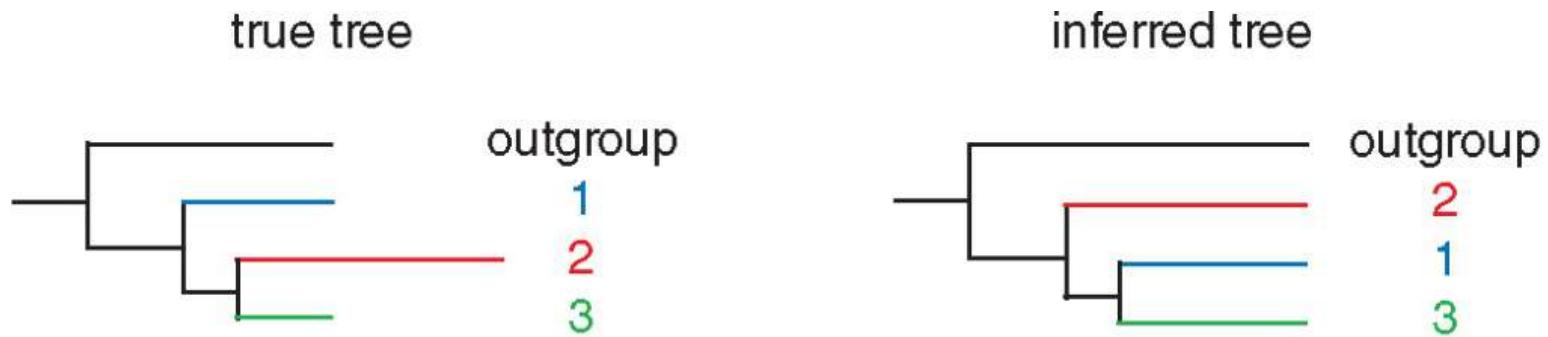


(d)



# Long-branch-chain attraction: an artifact

- The true tree (left) includes taxon 2 that evolves rapidly, and shares a common ancestor with taxon 3.
- The inferred tree (right) places taxon 2 separately because it is attracted by the long branch of the outgroup.



# Maximum Likelihood Method

# Making trees using maximum likelihood

- Maximum likelihood is an alternative to maximum parsimony. It is computationally intensive. A likelihood is calculated for the probability of each residue in an alignment, based upon some model of the substitution process.
- What are the tree topology and branch lengths that have the greatest likelihood of producing the observed data set?
- ML is implemented in the TREE-PUZZLE program, as well as MEGA5, PAUP and PHYLIP.

# Maximum Likelihood Method(con't)

- $s$  homologous sequences each with  $N$  nucleotides
- $X_k = (X_{1k}, \dots, X_{sk})$  the nucleotide configuration at  $k$ th site
- The likelihood function of tree  $T$  at the  $k$ th site
  - i.e.  $L(\theta_1, \dots, \theta_\eta | \mathbf{X}_1, \dots, \mathbf{X}_N, T) = \prod_{\kappa=1}^N f(\mathbf{X}_\kappa | \theta, T)$
- The likelihood function for the entire sequence for tree  $T$

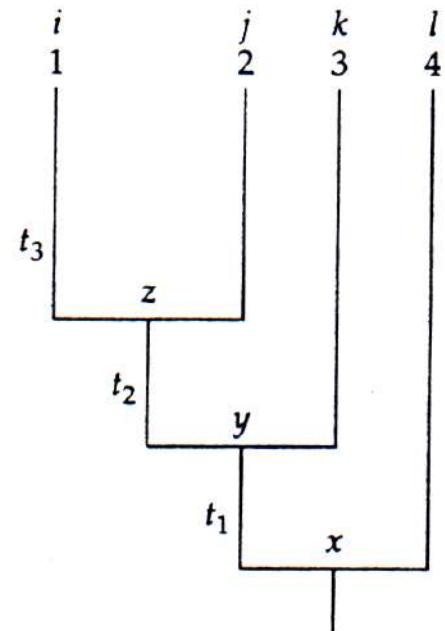
# Maximum Likelihood Method

- The likelihood function depends on the hypothetical tree

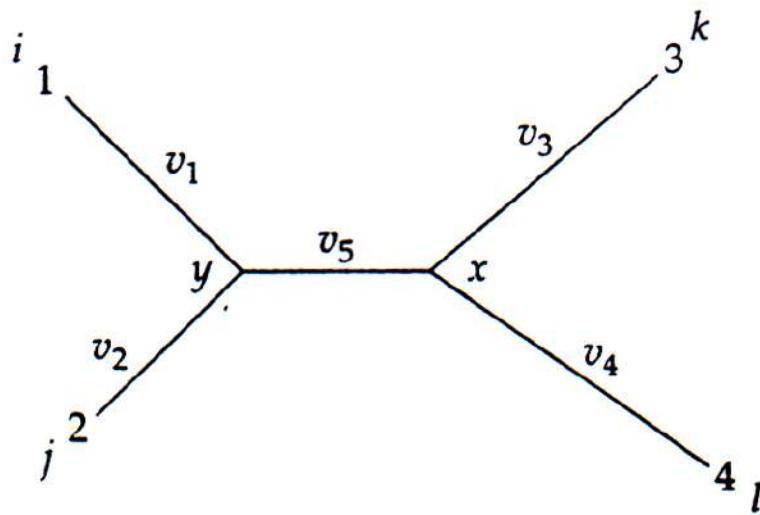
$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\lambda t/3} \quad (5.11)$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\lambda t/3} \quad (5.12)$$

$$\begin{aligned} h(i, j, k, \ell) &= \sum_x g_x P_{x\ell}(t_1 + t_2 + t_3) \\ &\times \sum_y P_{xy}(t_1) P_{yk}(t_2 + t_3) \sum_z P_{yz}(t_2) P_{zi}(t_3) P_{zj}(t_3) \end{aligned} \quad (5.13)$$



# Maximum Likelihood Method(con't)



$$h(i, j, k, \ell) = \sum_x g_x P_{x\ell}(v_4) P_{xk}(v_3) \sum_y P_{xy}(v_5) P_{yi}(v_1) P_{yj}(v_2)$$

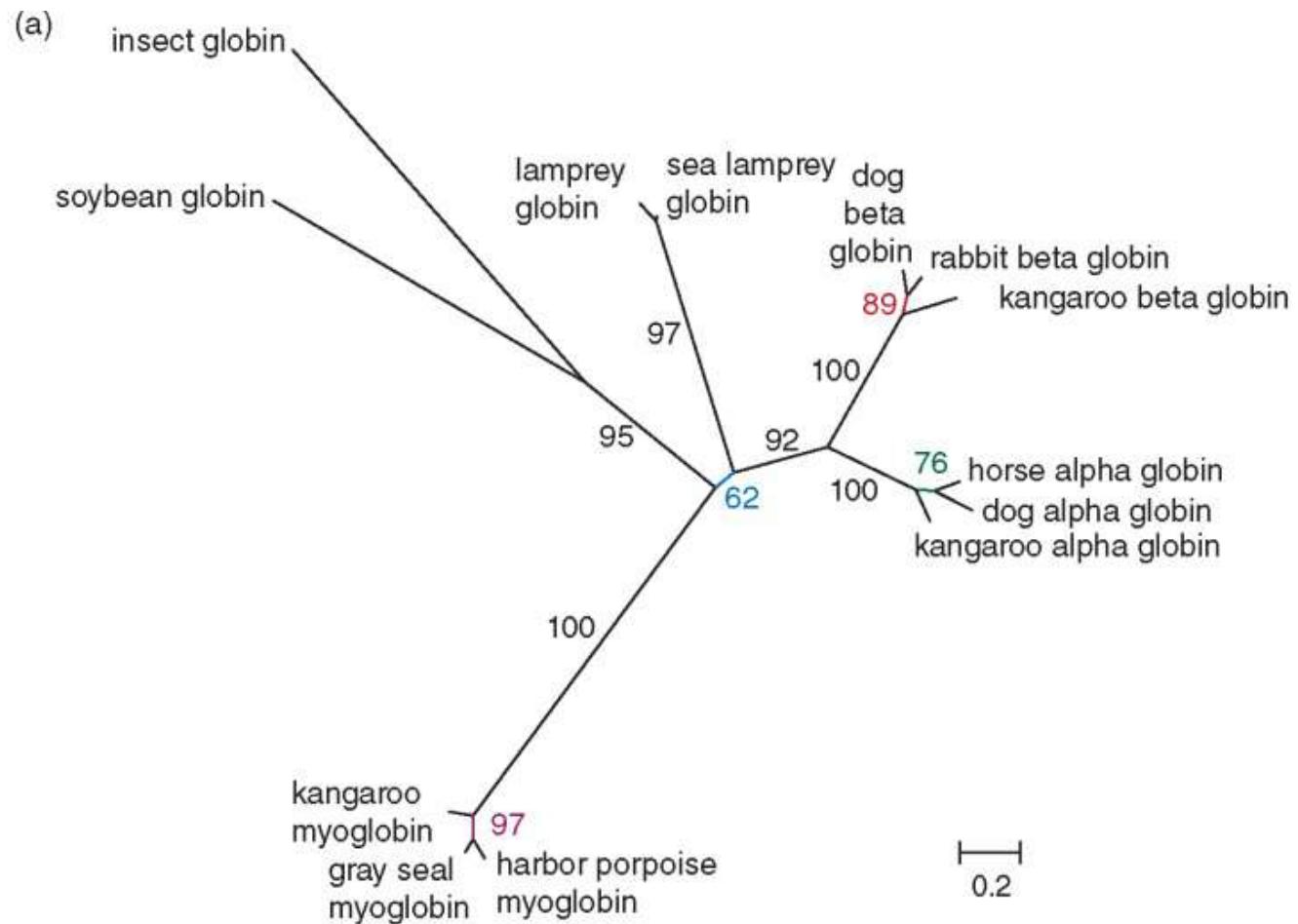
# Maximum likelihood: Tree-Puzzle

- I. Reconstruct all possible quartets A, B, C, D. For 12 myoglobins there are 495 possible quartets.
2. Puzzling step: begin with one quartet tree. N-4 sequences remain. Add them to the branches systematically, estimating the support for each internal branch. Report a consensus tree.

# Maximum likelihood: Tree-Puzzle

- I. Reconstruct all possible quartets A, B, C, D. For 12 myoglobins there are 495 possible quartets.
2. Puzzling step: begin with one quartet tree. N-4 sequences remain. Add them to the branches systematically, estimating the support for each internal branch. Report a consensus tree.

# Maximum likelihood tree



# Bayesian inference of phylogeny with MrBayes

- Bayesian inference is extremely popular for phylogenetic analyses (as is maximum likelihood). Both methods offer sophisticated statistical models. MrBayes is a very commonly used program.
- Notably, Bayesian approaches require you to specify prior assumptions about the model of evolution.

## Bayes' law

$$P(y==T | ev_1) = \frac{P(y==T) \times P(ev_1 | y==T)}{P(ev_1)}$$

$$P(y==F | ev_1) = \frac{P(y==F) \times P(ev_1 | y==F)}{P(ev_1)}$$

# Bayesian inference

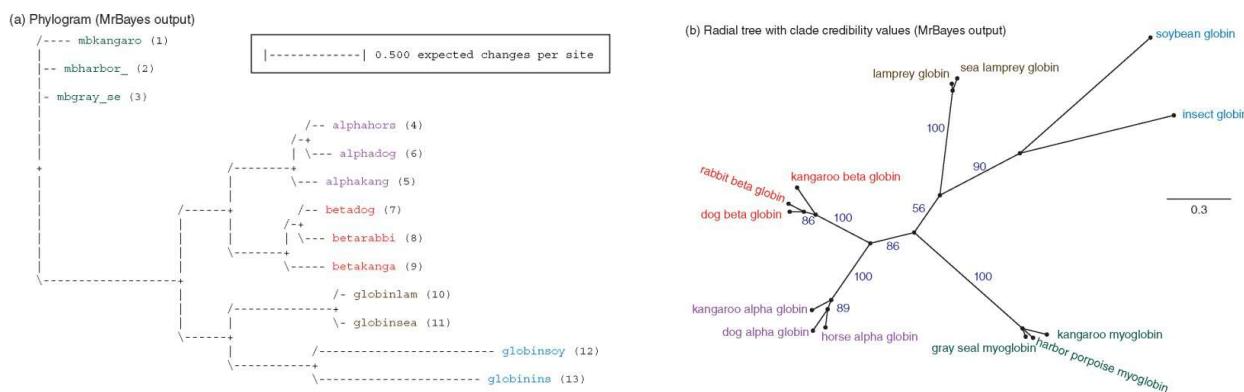
- Calculate:

$$\Pr[Tree|Data] = \frac{\Pr[Data|Tree] \times \Pr[Tree]}{\Pr[Data]}$$

- $\Pr[Tree|Data]$  is the posterior probability distribution of trees. Ideally this involves a summation over all possible trees. In practice, Monte Carlo Markov Chains (MCMC) are run to estimate the posterior probability distribution.

# Bayesian inference of phylogeny

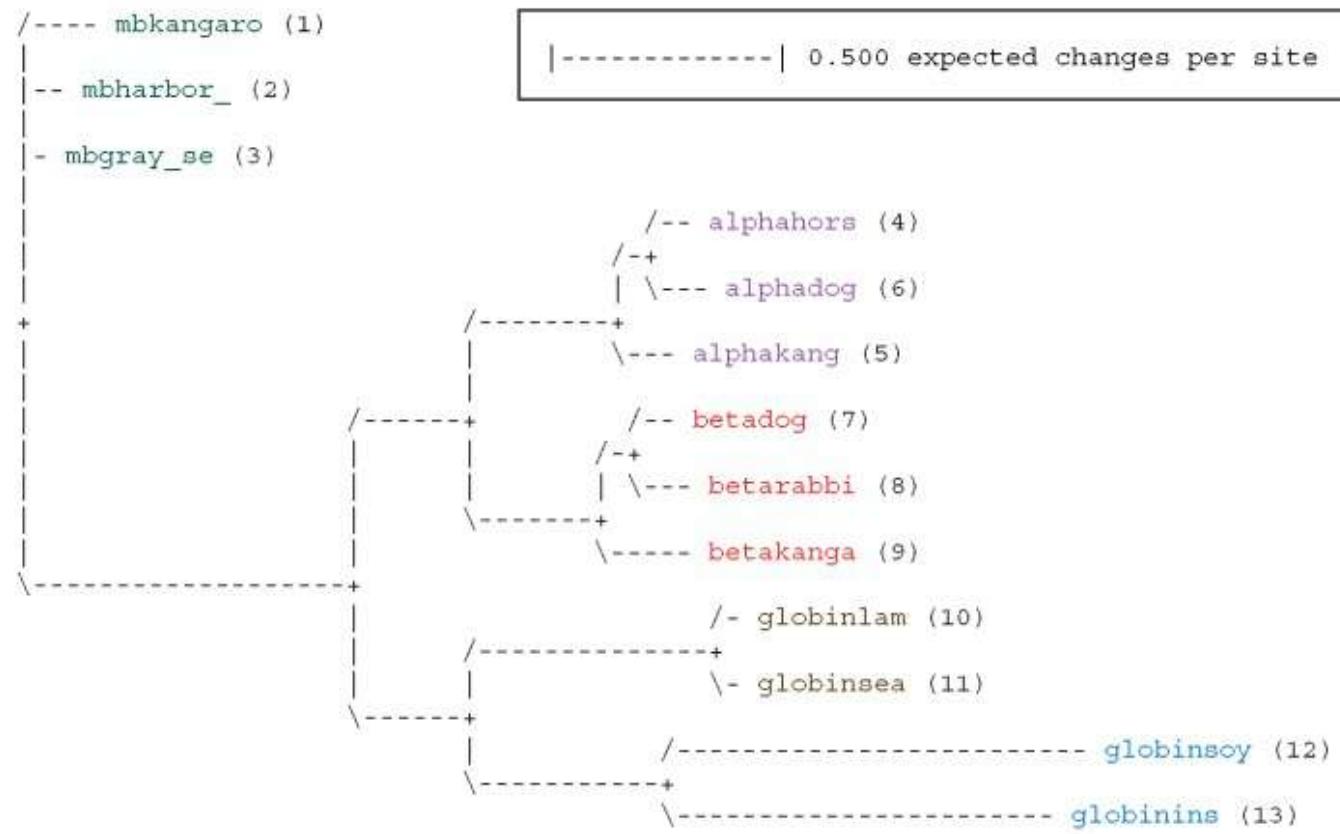
- Align 13 globin proteins with MAFFT.
  - In MrBayes select Poisson amino acid model with equal rates of substitution.
  - Select prior parameters (e.g. equal, fixed frequencies for the states; equal probability for all topologies; unconstrained branch lengths).
  - Run 1,000,000 trials for Monte Carlo Markov Chain estimation of the posterior distribution.
  - Obtain phylogram.
  - Export tree files and view with FigTree software.



# Bayesian inference of phylogeny

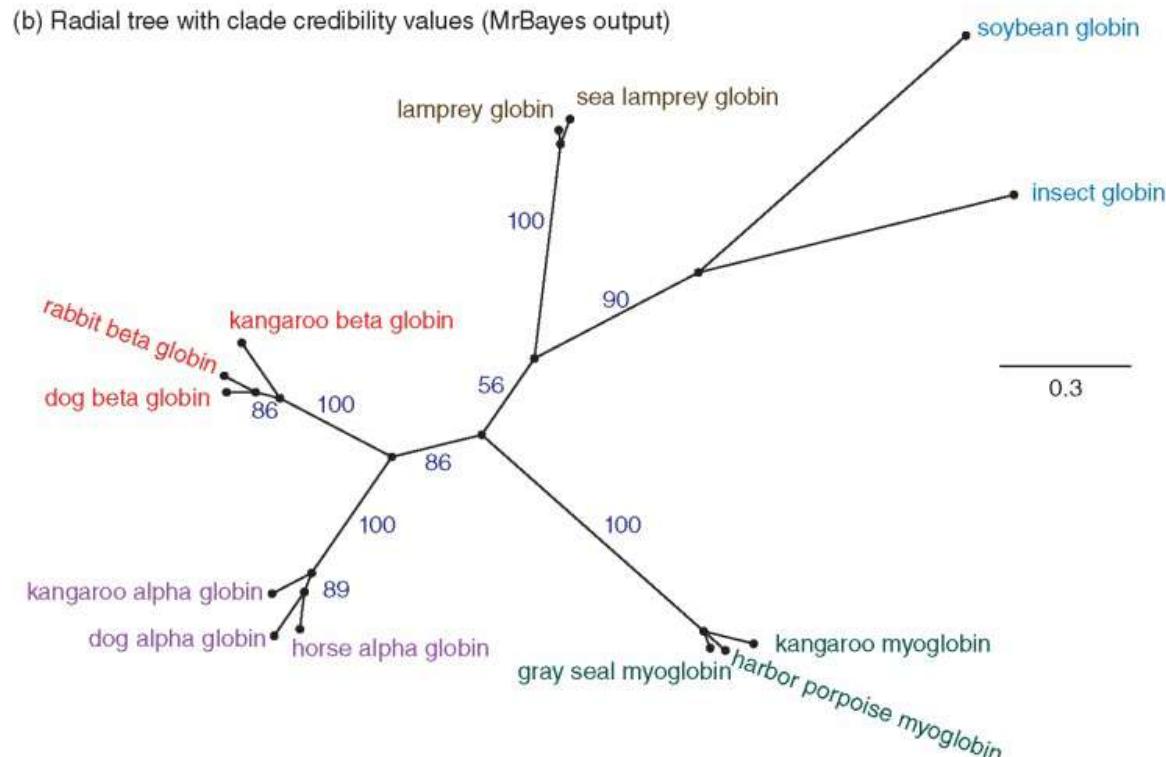
- Phylogram shows clades (note myoglobins are unresolved).

(a) Phylogram (MrBayes output)



# Bayesian inference of phylogeny

- Export tree files and view with FigTree software. Unrooted radial tree is shown. Nodes are given as closed circles. Clade credibility values (along branches) give 100% support for separation of most clades. The node containing the myoglobins is multifurcating.



## Stage 5: Evaluating trees

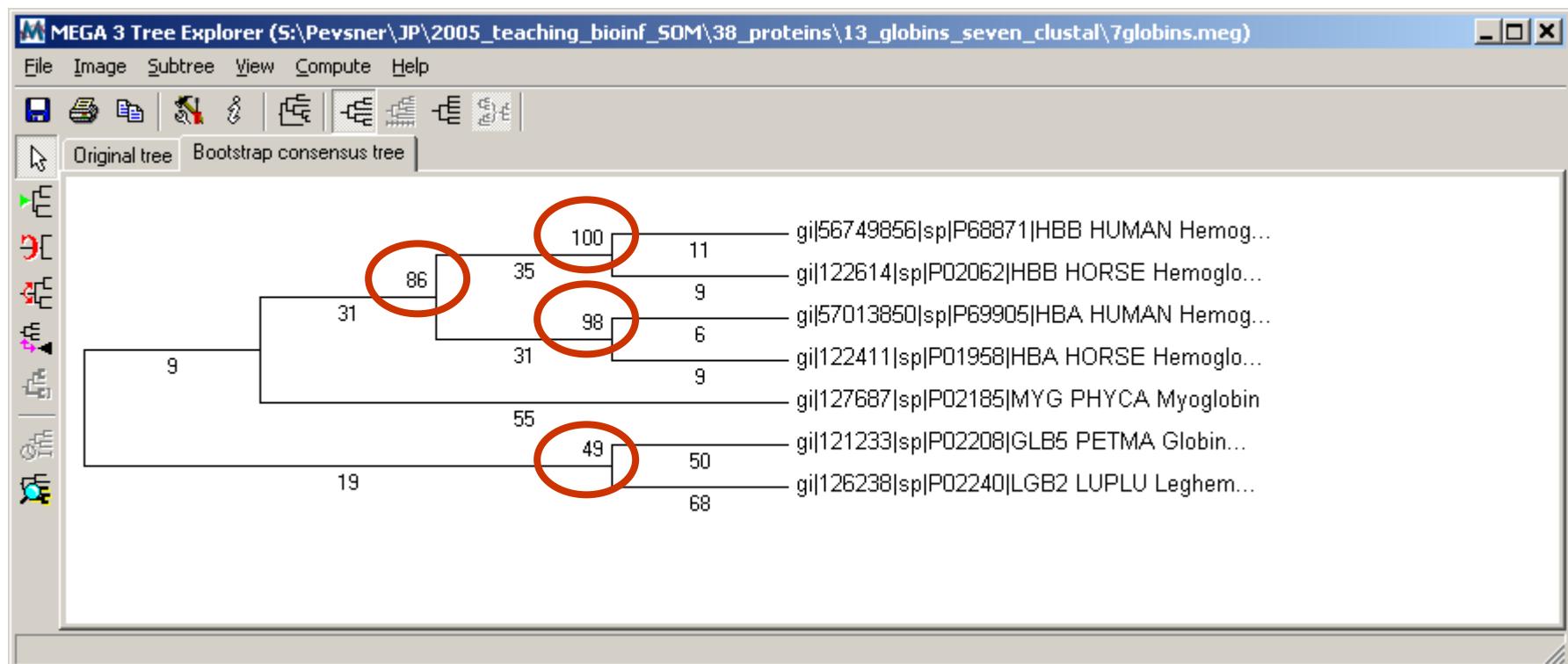
- The main criteria by which the accuracy of a phylogenetic tree is assessed are consistency, efficiency, and robustness. Evaluation of accuracy can refer to an approach (e.g. UPGMA) or to a particular tree.

## Stage 5: Evaluating trees: bootstrapping

- Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?

# MEGA trees display bootstrap values

- Bootstrap values show the percent of times each clade is supported after a large number ( $n=500$ ) of replicate samplings of the data.



## Stage 5: Evaluating trees: bootstrapping

- To bootstrap, make an artificial dataset obtained by randomly sampling columns from your multiple sequence alignment. Make the dataset the same size as the original. Do 100 (to 1,000) bootstrap replicates.
- Observe the percent of cases in which the assignment of clades in the original tree is supported by the bootstrap replicates. >70% is sometimes considered significant.

Felsenstein, J. *Evolution* 39, 783–791 (1985).

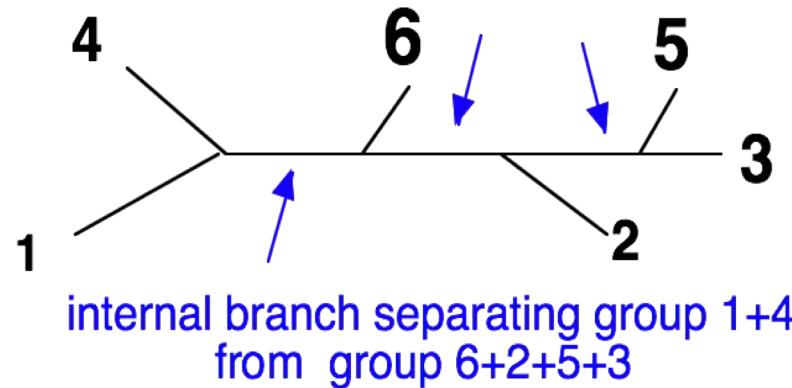
- Number 41 on the list is a description of how to apply statistics to phylogenies.
- In 1984, evolutionary biologist Joe Felsenstein of the University of Washington in Seattle adapted a statistical tool known as the bootstrap to infer the accuracy of different parts of an evolutionary tree. The bootstrap involves resampling data from a set many times over, then using the variation in the resulting estimates to determine the confidence for individual branches. Although the paper was slow to amass citations, it rapidly grew in popularity in the 1990s and 2000s as molecular biologists recognized the need to attach such intervals to their predictions.

Felsenstein, J. *Evolution* 39, 783–791 (1985).

Felsenstein says that the concept of the bootstrap, devised in 1979 by Bradley Efron, a statistician at Stanford University in California, was much more fundamental than his work. But applying the method to a biological problem means it is cited by a much larger pool of researchers. His high citation count is also a consequence of how busy he was at the time, he says: he crammed everything into one paper rather than publishing multiple papers on the topic, which might have diluted the number of citations each one received. “I was unable to go off and write four more papers on the same thing,” he says. “I was too swamped to do that, not too principled.”

# Reliability of phylogenetic trees: the bootstrap

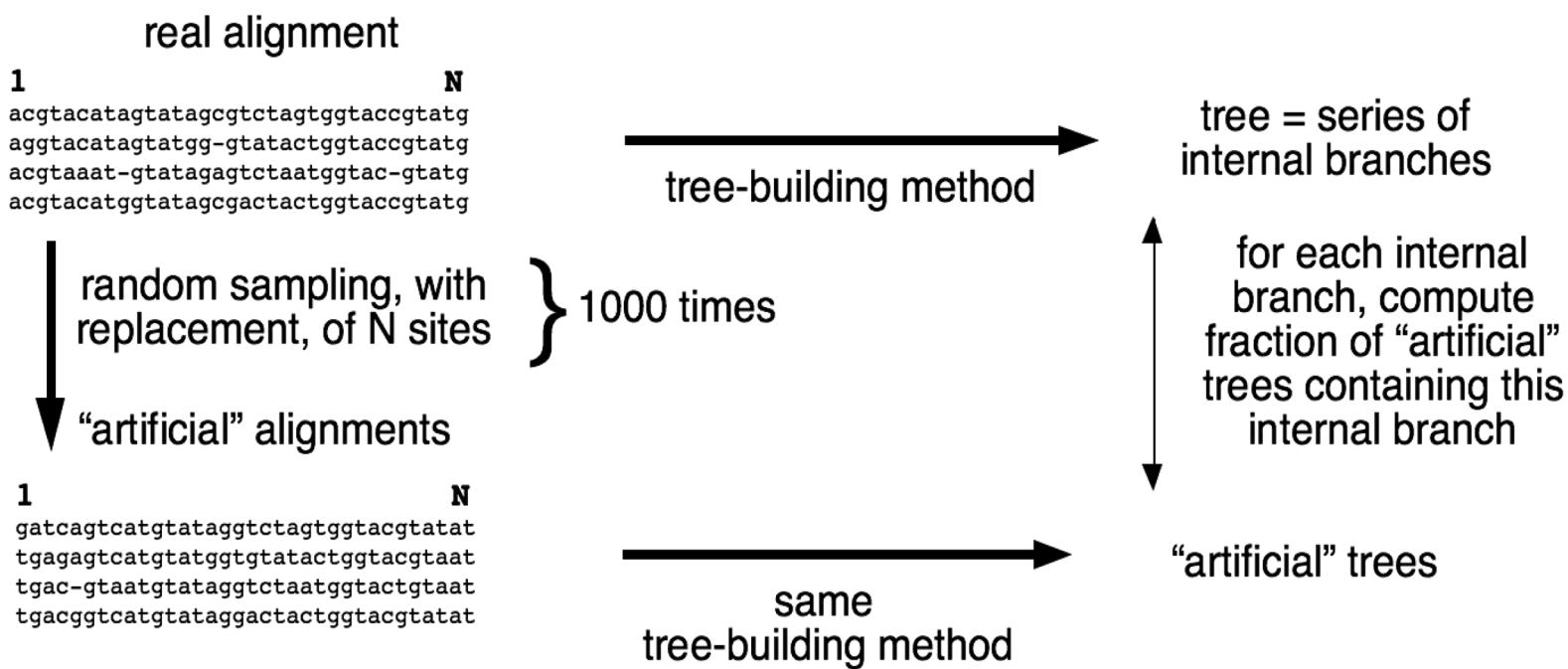
- The phylogenetic information expressed by an unrooted tree resides entirely in its internal branches.



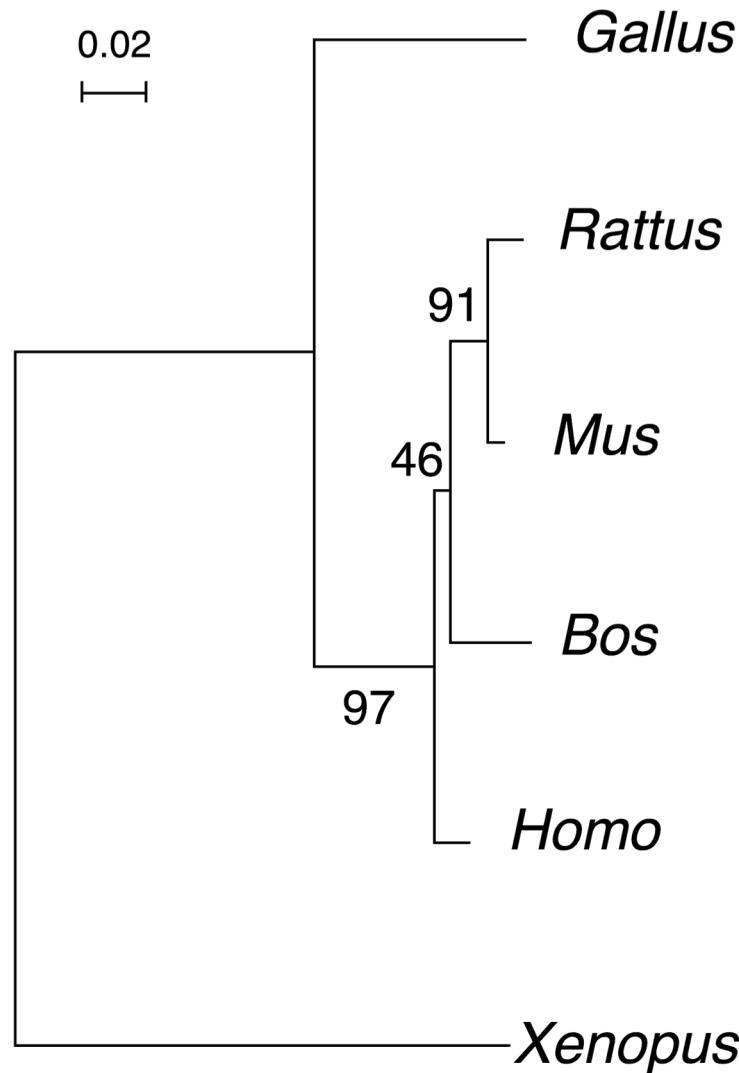
- The tree shape can be deduced from the list of its internal branches.
- Testing the reliability of a tree = testing the reliability of each internal branch.

# Bootstrap procedure

- The support of each internal branch is expressed as percent of replicates.



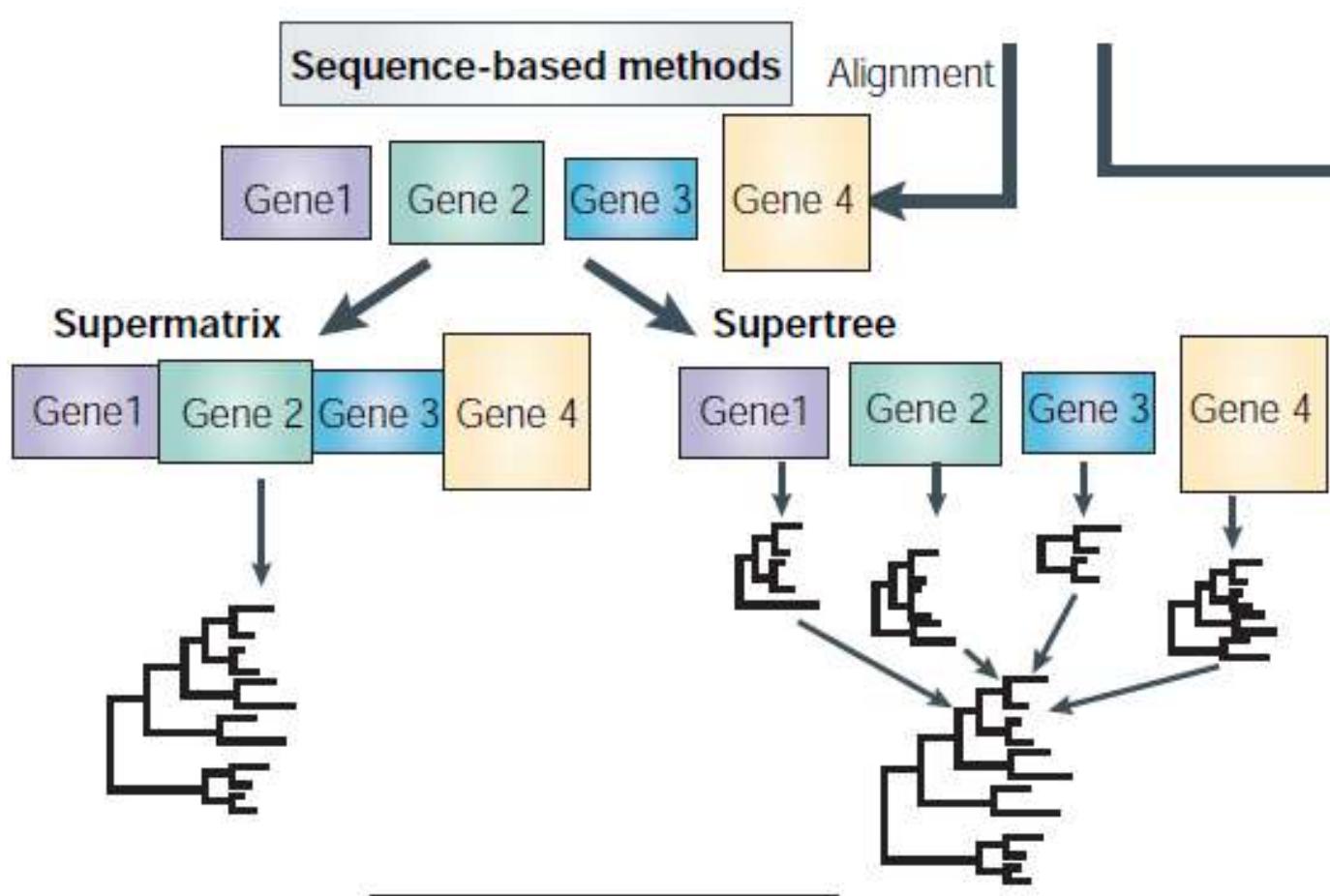
# "bootstrapped" tree



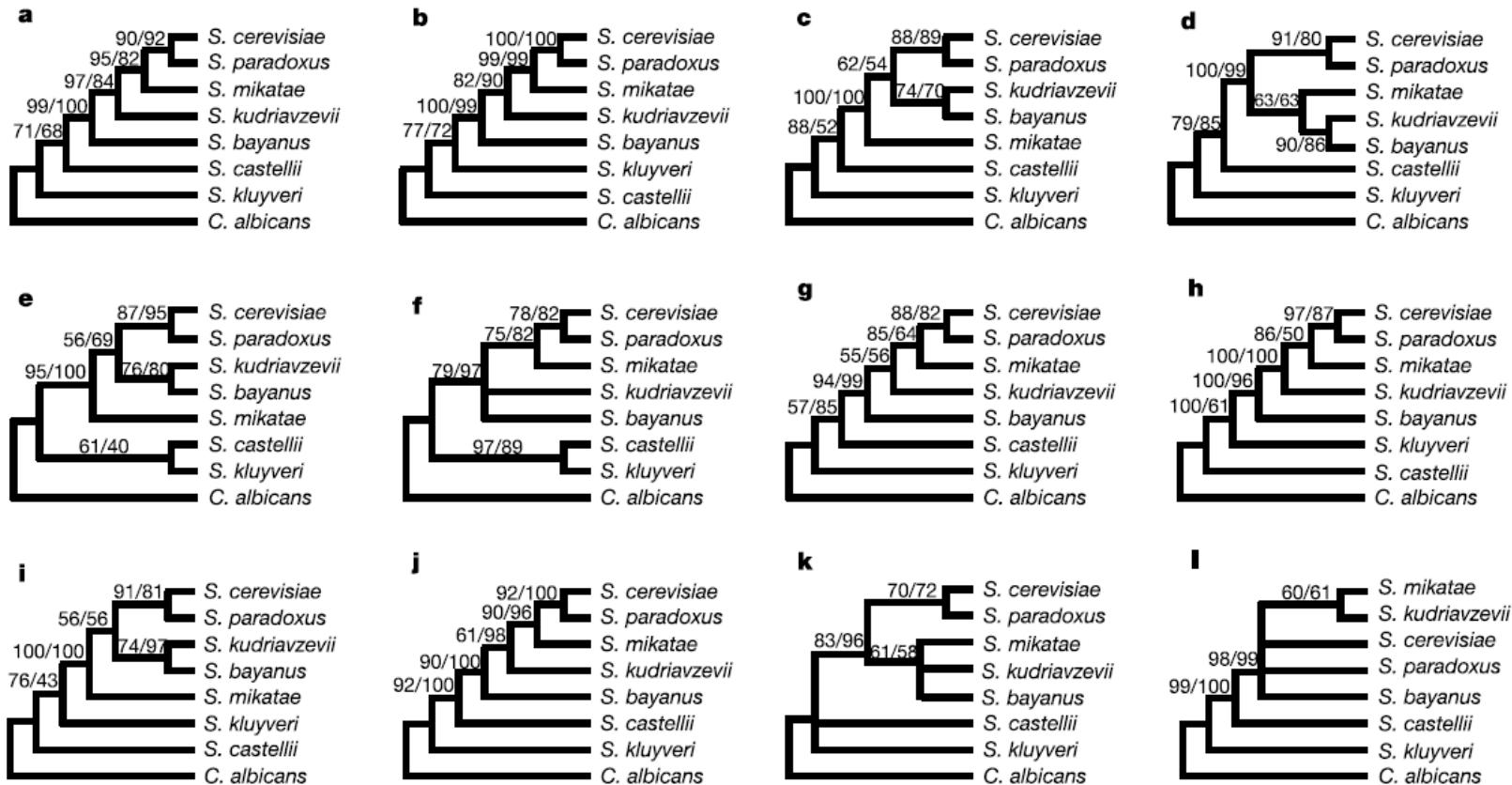
# Bootstrap procedure : properties

- Internal branches supported by  $\geq 90\%$  of replicates are considered as statistically significant.
- The bootstrap procedure only detects if sequence length is enough to support a particular node.
- The bootstrap procedure does not help determining if the tree-building method is good. A wrong tree can have 100 % bootstrap support for all its branches!

# Supermatrix

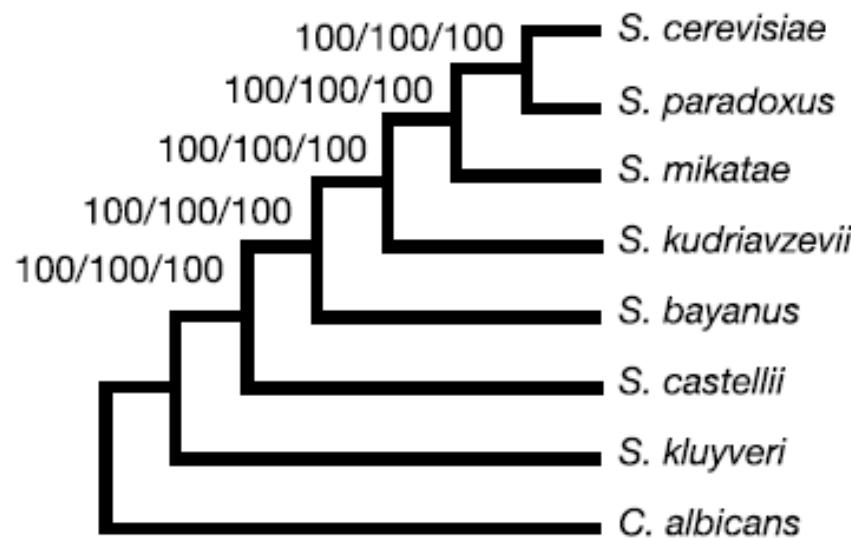


Single-gene data sets generate multiple, robustly supported alternative topologies.



# Concatenation

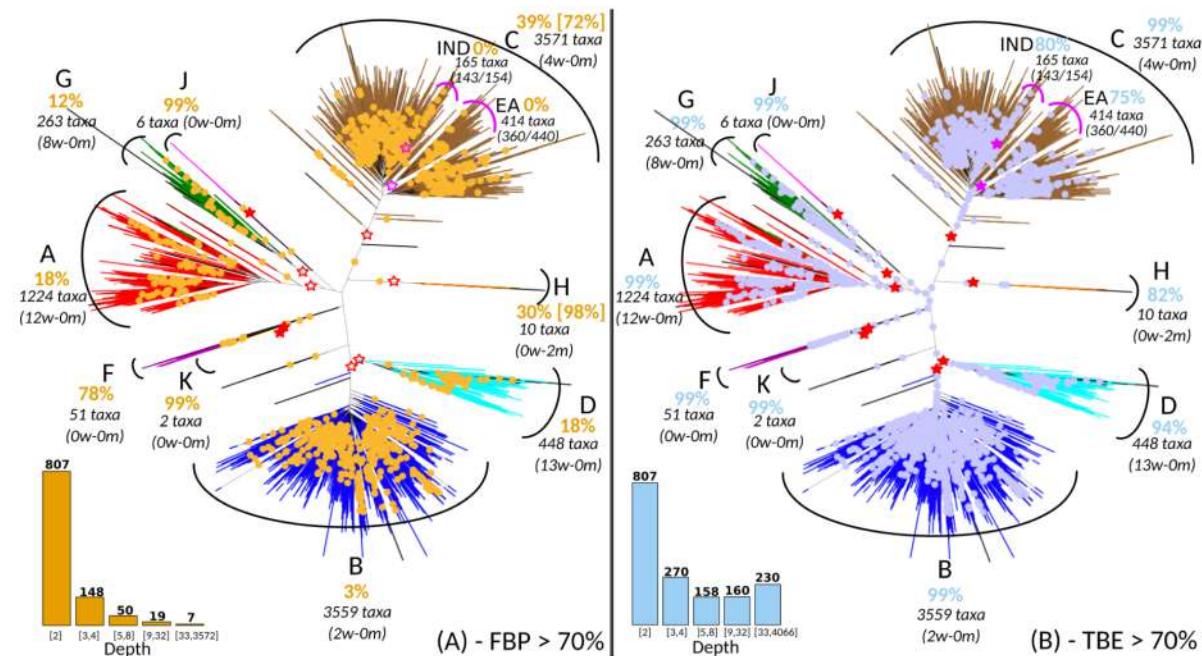
Right Topology + High Bootstrap Support



**Figure 4** Phylogenetic analyses of the concatenated data set composed of 106 genes yield maximum support for a single tree, irrespective of method and type of character evaluated. Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides/MP on amino acids).

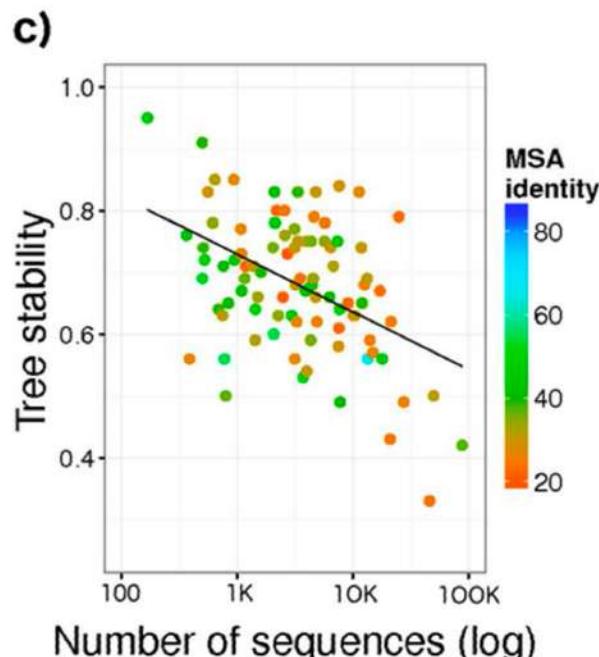
# Boosting Felsenstein Phylogenetic Bootstrap

- *transfer distance, (b,b\*)* : a branch b of the reference tree T and a branch b\* of a bootstrap tree T\* is equal to the number of taxa that must be transferred (or removed), in order to make both branches identical
- Felsenstein (FBP) and transfer (TBE) bootstrap supports on the same tree with 9,147 HIV-1M pol sequences

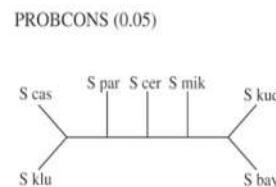
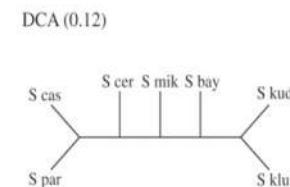
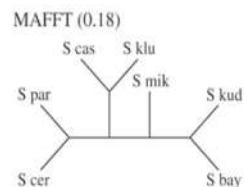
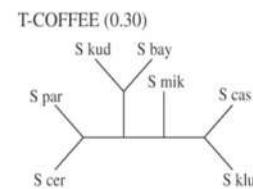
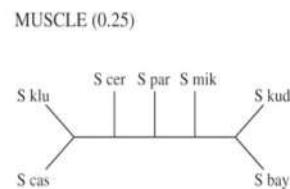
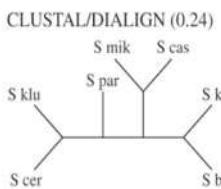


# MSAs and trees are sensitive to sequence input order

- Sequences from HOMFAM protein benchmark dataset
- Shuffled input sequences result in Robinson and Foulds topological variability among FastTree maximum likelihood trees



# Which MSA Method ?



# If you build a tree,



Clustal



Dalign



DCA



Mafft



Muscle



ProbCons



T-Coffee

## Which guy should I trust?

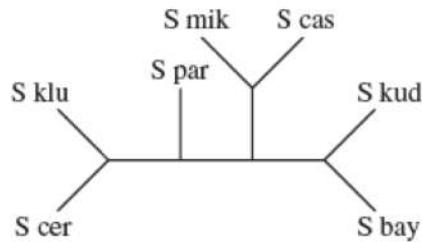
panel D). Most studies ignore that these scores are based on a fixed sequence alignment that supports the tree in the first place; they may thus make us overly confident of its accuracy.

Ari Löytynoja and Nick Goldman, “Uniting Alignments and Trees,” *Science* 324, no. 5934 (June 19, 2009): 1528 -1529.

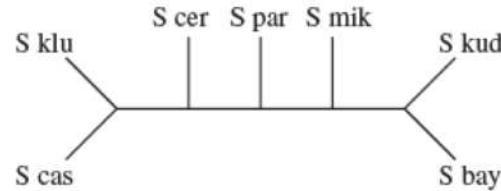
# External alignment uncertainty

## YPL077C with six topologies

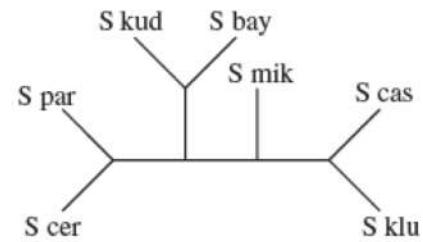
CLUSTAL/DIALIGN (0.24)



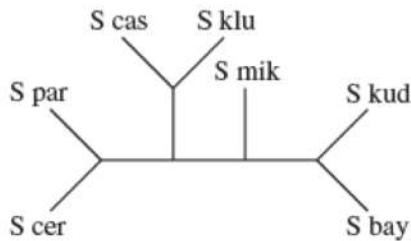
MUSCLE (0.25)



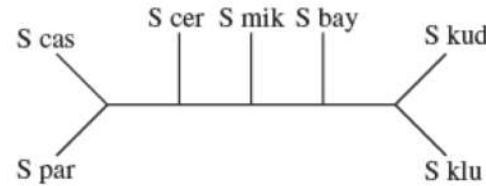
T-COFFEE (0.30)



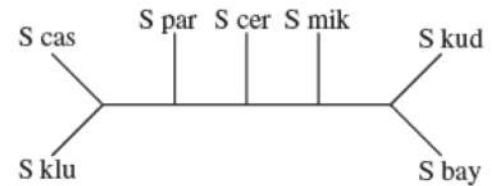
MAFFT (0.18)



DCA (0.12)



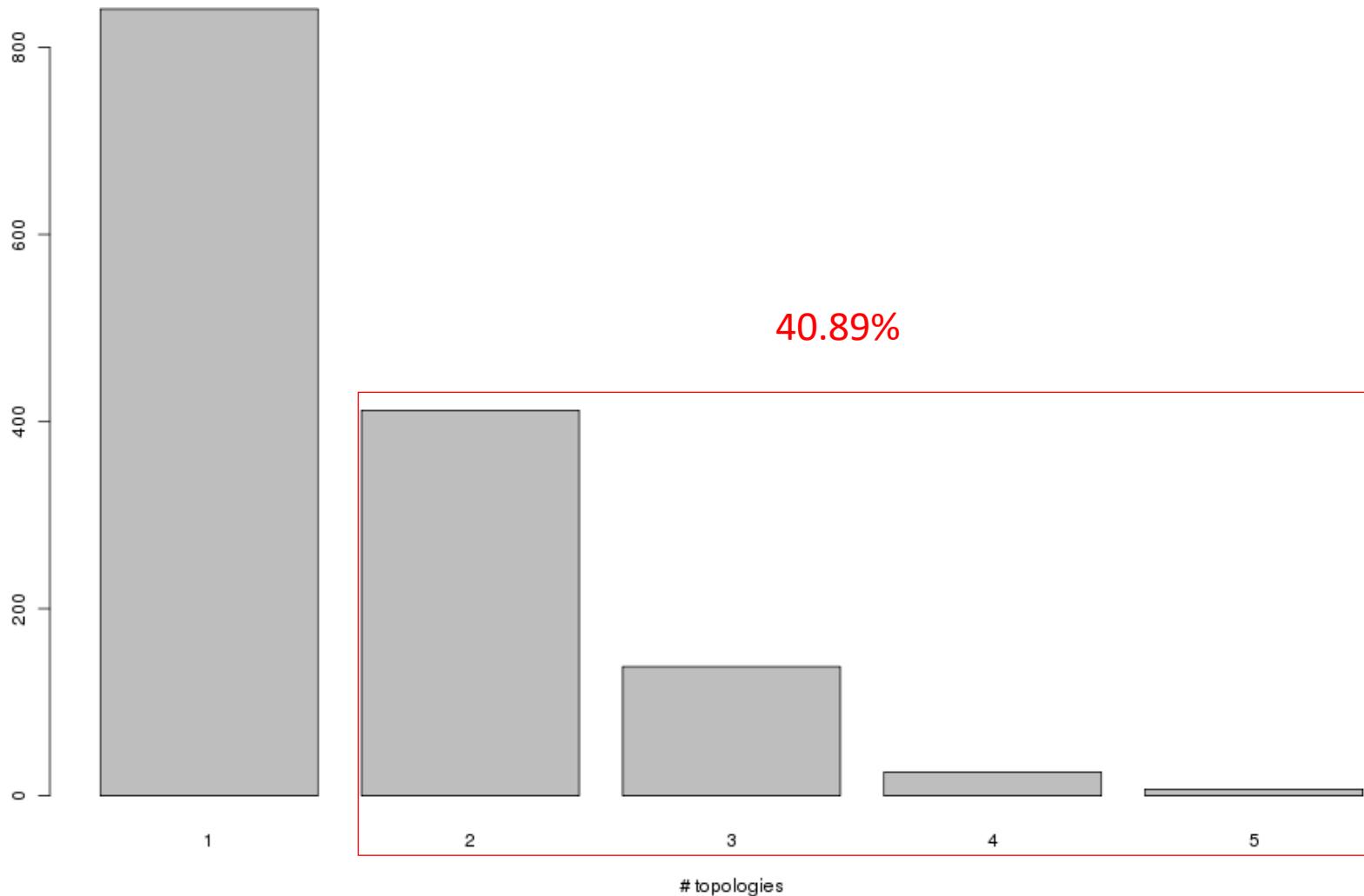
PROBCONS (0.05)



**Fig. 1.** An example, involving ORF YPL077C, in which alignments produced by seven different alignment methods produce six different estimated trees, albeit with low bootstrap support (bootstrap proportions shown parenthetically for each tree).

Karen M Wong, Marc A Suchard, and John P Huelsenbeck, “Alignment uncertainty and genomic analysis”, Science 319, no. 5862 (January 25, 2008): 473-476.

### Topology group size distribution



Karen M Wong, Marc A Suchard, and John P Huelsenbeck, "Alignment uncertainty and genomic analysis", Science 319, no. 5862 (January 25, 2008): 473-476.

# Bootstrap

Original data set  
with  $n$   
characters.

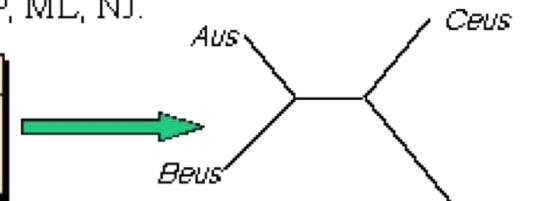
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aur	C	G	A	C	G	G	T	G	G	T	C	T	A	T	A	C	A	C	G	A
Bear	C	G	G	C	G	G	T	G	A	T	C	T	A	T	G	C	A	C	G	G
Ceus	T	G	G	C	G	G	C	G	T	C	T	C	A	T	A	C	A	A	T	A
Deus	T	A	A	C	G	A	T	G	A	C	C	C	G	A	C	T	A	T	T	G

Draw  $n$  characters  
randomly with re-  
placement.  
Repeat  $m$   
times.

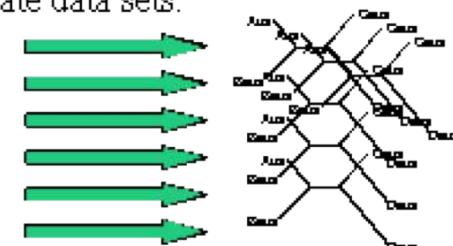
	1	3	13	8	3	19	14	6	20	20	7	1	9	11	17	10	6	14	8	16
Aur	G	A	A	G	A	G	T	Q	A	A	T	C	G	C	A	T	G	T	G	C
Bear	G	G	A	G	G	G	T	T	Q	A	A	C	T	T	T	A	C	G	T	C
Ceus	G	G	A	G	G	T	T	Q	A	A	C	T	T	T	A	C	A	A	G	T
Deus	A	A	G	G	A	T	A	A	G	G	T	T	A	C	A	C	A	A	G	T

$m$  pseudo-replicates,  
each with  $n$  characters.

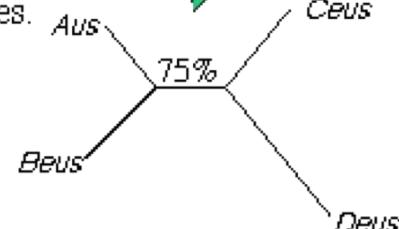
Original  
analysis, e.g.  
MP, ML, NJ.



Repeat original analysis  
on each of the pseudo-  
replicate data sets.



Evaluate the  
results from the  
 $m$  analyses.





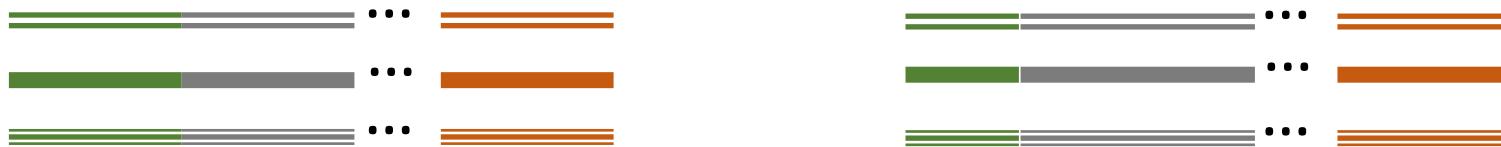
ANDREY ZHARKIKH AND WEN-HSIUNG LI

*Institute for Demographic and Population Genetics, University of  
P.O. Box 20334, Houston, Texas 77225*

Received April 15, 1994; revised September 12, 1994

partial

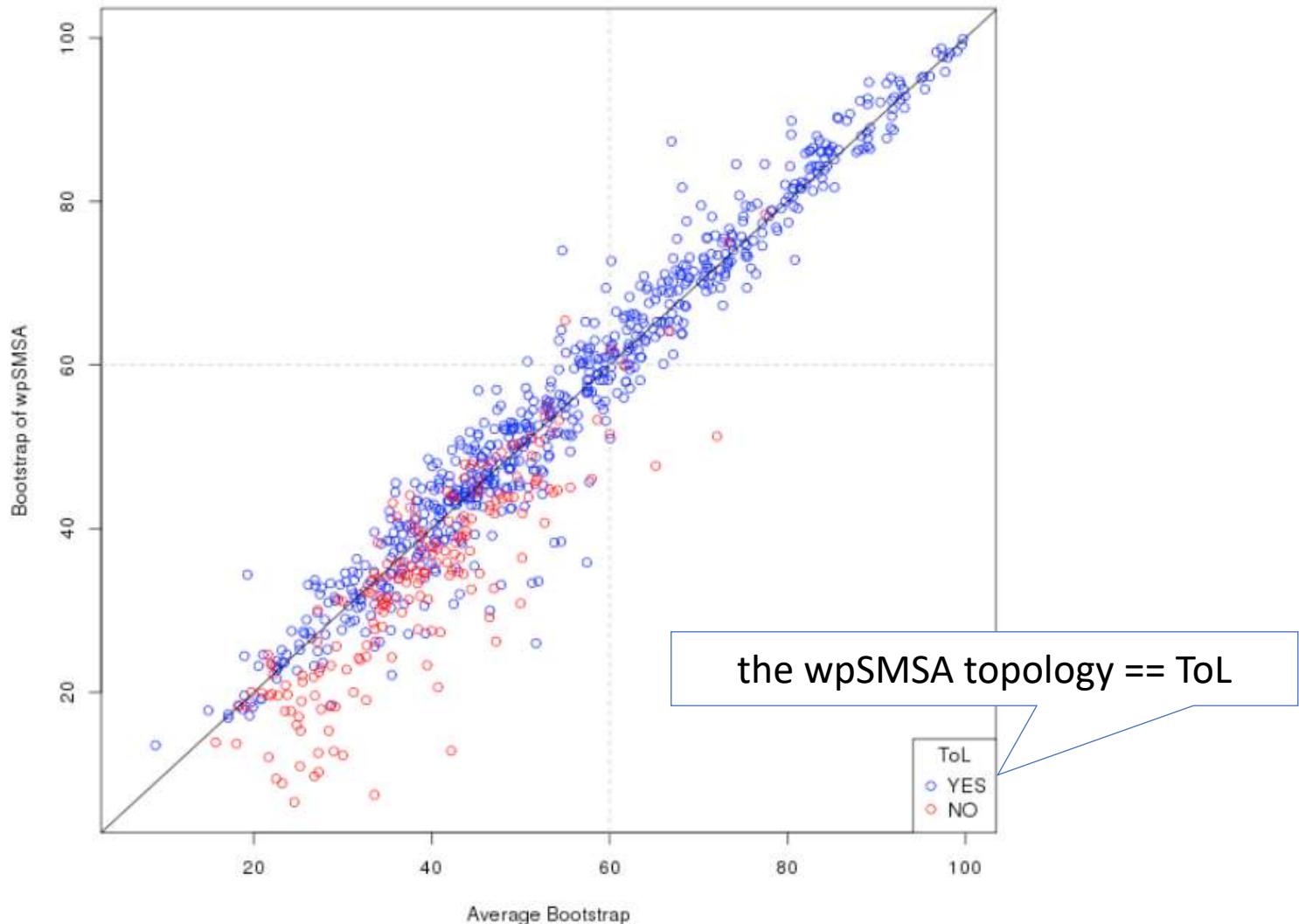
weighted



Average bootstrap, AUC values and the number of TPs for 10 and 25 accepted FPs of each method

Method	ave. Bootstrap	AUC	TPs for 10 FPs	TPs for 25 FPs	TPs in total
Clustal	51.31	0.7521	185	274	643
DCA	50.62	0.7694	194	284	624
DIALIGN	51.94	0.7618	253	340	659
MAFFT	52.82	0.7750	253	359	665
Muscle	52.35	0.7771	224	315	639
Probnt	50.96	0.7790	256	312	642
T-Coffee	51.21	0.7889	234	311	620
M-Coffee	51.41	0.7688	193	325	646
SMSA	77.31	0.8301	329	425	661
pSMSA	50.96	0.8140	342	385	661
wpSMSA	50.86	0.8215	353	423	661

The bootstrap of wpSMSA versus the average bootstrap of individual aligners.



# Perspective

- We have discussed concepts of evolution and phylogeny that address the relationships of protein, genes, and species over time.
- A phylogenetic tree is essentially a graphical representation of a multiple sequence alignment.
- There are many methods for creating phylogenetic trees. Neighbor-joining is a simple trusted method (and is useful for large numbers of taxa). Maximum likelihood and Bayesian methods are commonly used because they are model-based with rigorous statistical frameworks.
- Each method is associated with errors, and it is crucial to begin with an appropriate multiple sequence alignment.

# 2016, 15TH CRG SYMPOSIUM: EVOLUTION AND MEDICINE

- Evolutionary theory forms the basis for our understanding of the natural world. The development, behaviour and physiology of our own species has been shaped by million of years of evolution. Furthermore, the interaction of humans with species that directly affect our health and survival is constantly being modified by evolutionary forces through natural or artificial selection.
- The conference aims to explore the interplay of evolution and medicine, acting either in our own species or on species directly affecting our well being. We will explore topics in cancer evolution, the emergence of resistance to antibiotics and other drugs, the origin and evolution of pathogens, and the evolutionary influence on our own construction and predisposition to various diseases.
- <https://www.youtube.com/watch?v=XPLAREaNKmk>

# Terminology and Concepts of Trees

- Aidan Budd, EMBL Heidelberg
- [http://www.embl.de/~seqanal/courses/commonCourseContent/presentation\\_introPhyloCoME2013Hinxton.pdf](http://www.embl.de/~seqanal/courses/commonCourseContent/presentation_introPhyloCoME2013Hinxton.pdf)
  - P.20~100
  - P.155~183

# References

- Introduction to Phylogeny
  - By Cedric Notredame, CRG
  - [http://webext.pasteur.fr/tekaia/BCGA2012/TALKS/Notredame\\_phylogeny.pdf](http://webext.pasteur.fr/tekaia/BCGA2012/TALKS/Notredame_phylogeny.pdf)
- Inferring Phylogenetic Trees
  - By Olivier Gascuel
  - [http://www.embl.de/~seqanal/courses/commonCourseContent/Gascuel\\_BuildingTrees.pdf](http://www.embl.de/~seqanal/courses/commonCourseContent/Gascuel_BuildingTrees.pdf)
- Mathematical and Computational Phylogenetics
  - By Olivier Gascuel
  - [http://www.embl.de/~seqanal/courses/commonCourseContent/Gascuel\\_Models.pdf](http://www.embl.de/~seqanal/courses/commonCourseContent/Gascuel_Models.pdf)
- Lecture notes of molecular systematics
  - <http://www.bioinf.org/molysys/lectures.html>

# WWW resources for molecular phylogeny (1)

- Compilations

- A list of sites and resources:  
<http://www.ucmp.berkeley.edu/subway/phylogen.html>
- An extensive list of phylogeny programs  
<http://evolution.genetics.washington.edu/phylip/software.html>

- Databases of rRNA sequences and associated software

- The rRNA WWW Server - Antwerp, Belgium.  
<http://rrna.uia.ac.be>
- The Ribosomal Database Project - Michigan State University  
<http://rdp.cme.msu.edu/html/>

# WWW resources for molecular phylogeny (3)

- Sequence alignment editor
  - SEAVIEW : for windows and unix  
<http://pbil.univ-lyon1.fr/software/seaview.html>
- Programs for molecular phylogeny
  - PHYLIP : an extensive package of programs for all platforms  
<http://evolution.genetics.washington.edu/phylip.html>
  - PAUP : a very performing commercial package  
<http://paup.csit.fsu.edu/index.html>
  - PHYLO\_WIN : a graphical interface, for unix only  
<http://pbil.univ-lyon1.fr/software/phylowin.html>
  - MrBayes : Bayesian phylogenetic analysis <http://morphbank.ebc.uu.se/mrbayes/>
  - PHYML : fast maximum likelihood tree building  
<http://www.lirmm.fr/~guindon/phym.html>
  - WWW-interface at Institut Pasteur, Paris  
<http://bioweb.pasteur.fr/seqanalphylogeny>

## WWW resources for molecular phylogeny (4)

- Tree drawing  
NJPLOT (for all platforms)  
<http://pbil.univ-lyon1.fr/software/njplot.html>
- Books
  - Laboratory techniques  
Molecular Systematics (2nd edition), Hillis, Moritz & Mable eds.; Sinauer, 1996.
  - Molecular evolution  
Fundamentals of molecular evolution (2nd edition); Graur & Li; Sinauer, 2000.
  - Evolution in general  
Evolution (2nd edition); M. Ridley; Blackwell, 1996.

<http://changlabtw.com>  
chang.jiaming@gmail.com

A scenic landscape featuring a calm lake in the foreground, surrounded by dark green hills or mountains. In the distance, several colorful tents are scattered across a grassy ridge. The sky is a clear blue with a few wispy white clouds.

Thank You  
Any Question?

# Tree Measurement

- Absolute
  - Minimax
  - Minisum
  - Minisize
- Relative(using distance matrix)
  - Neighbor-point relation
  - Compact set

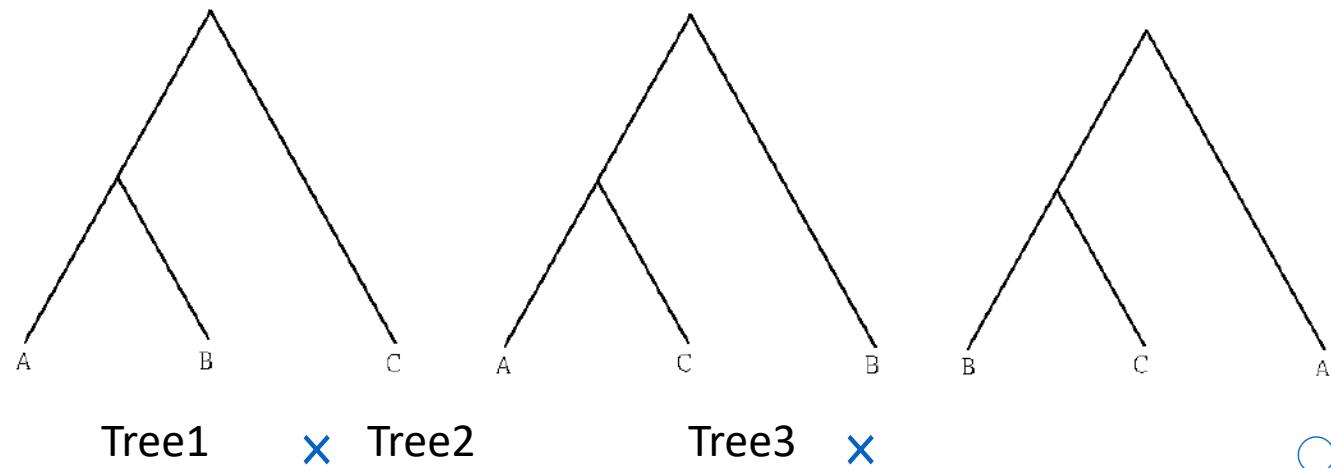
# Neighbor-point Relation Measurement

- For each tree species ,

$$d(i, j) = \min\{d(i, j), d(i, k), d(j, k)\}$$
$$lca(i, j) \leq lca(i, k) = lca(j, k)$$

if and only if

A	B	C
0	8	5 A
0	3 B	
0	C	
Distance matrix		



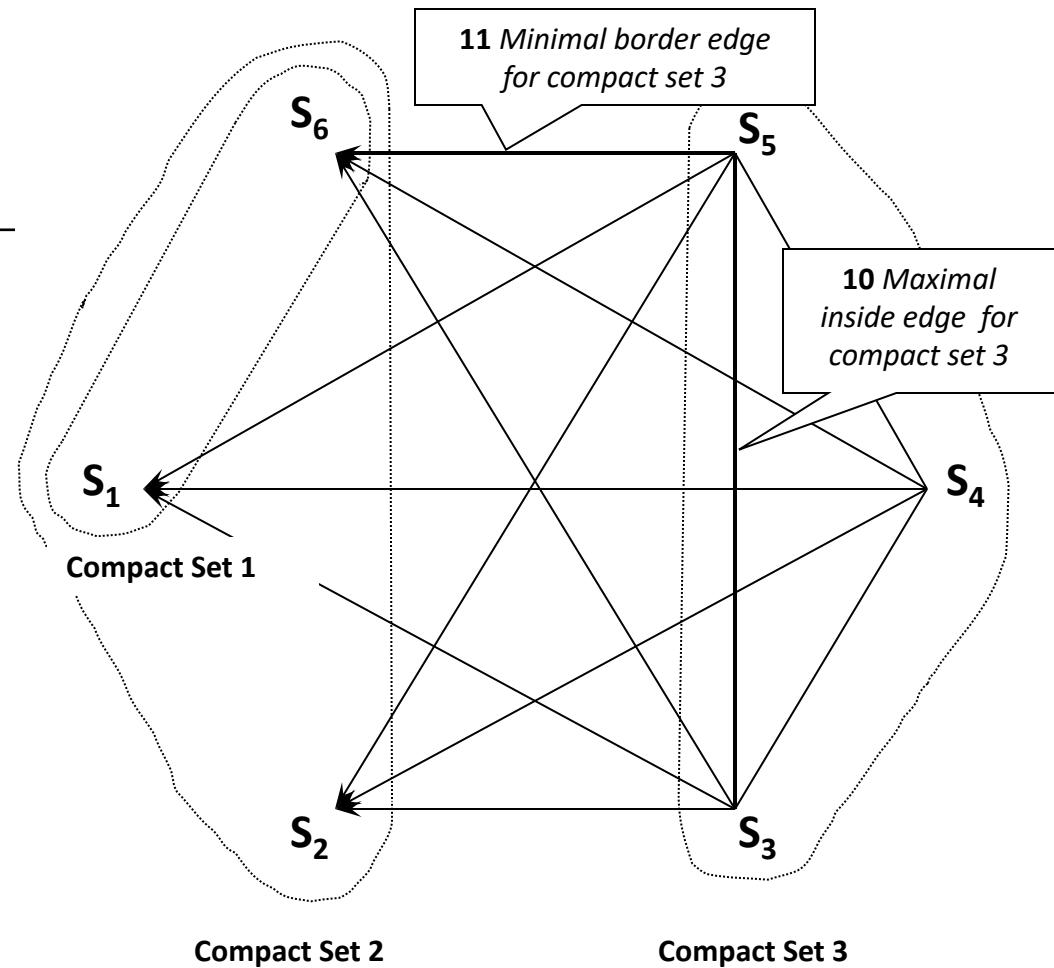
# Compact Set Definition

- Let  $S$  be the set of  $n$  objects  $\{S_1, S_2, S_3 \dots S_n\}$  and  $D(S_i, S_j)$  denote the distance between  $S_i$  and  $S_j$  in the distance matrix  $D$ .
- Consider any  $C$  which is a subset of  $S$ , if the distance between elements in  $C$  and not in  $C$  is larger than the longest distance in  $C$ , then  $C$  is called a **compact set**.
- Property :
  - The entire set  $S$  is a compact set.
  - Each set consisting of a single object is also a compact set.

# Compact Set Example

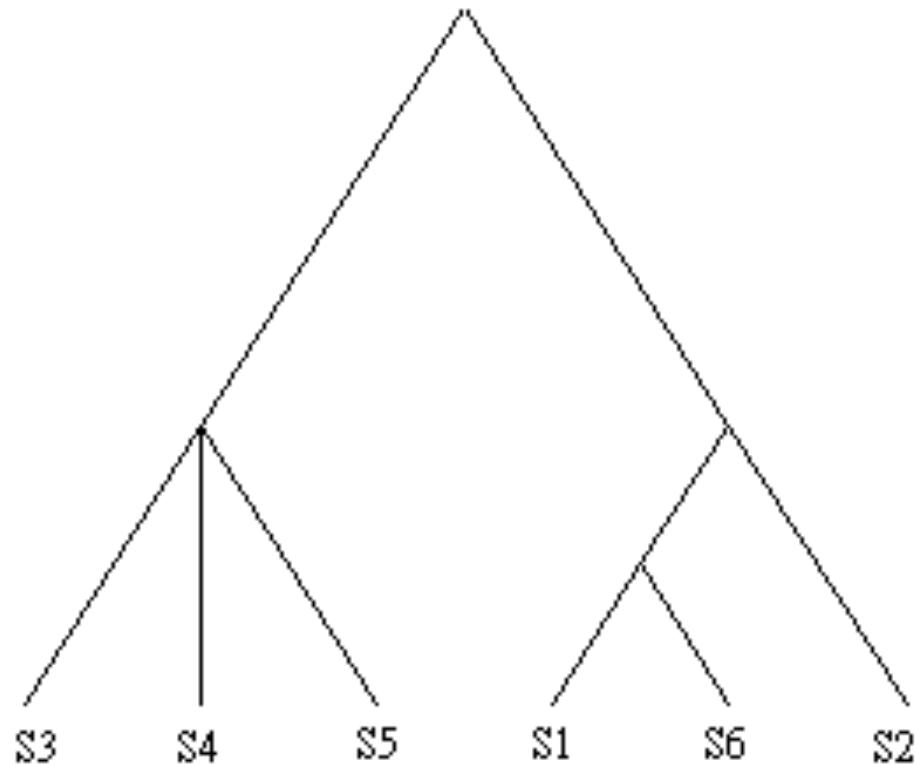
D	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$	$S_6$
$S_1$	0	10	16	18	13	8
$S_2$	0	14	17	15	9	
$S_3$	0	9	10	12		
$S_4$	0	9	19			
$S_5$	0	11				
$S_6$	0					

Distance Matrix



## Compact Set Example(con't)

- Compact Set is hierarchical



# Measure of Compact Set Preservation

- How can we measure the Compact Set Preservation in quantity?

$N1$ : # of the original Compact Set relations

$N2$ : # of the relations preserved after MSA

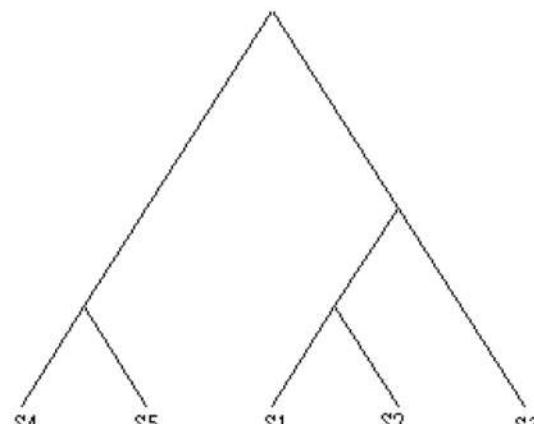
Estimate by Compact Set Preservation =

$$\frac{N2}{N1}$$

## Measure of Compact Set Preservation(con't)

<b>D</b>	<b>S<sub>1</sub></b>	<b>S<sub>2</sub></b>	<b>S<sub>3</sub></b>	<b>S<sub>4</sub></b>	<b>S<sub>5</sub></b>
<b>S<sub>1</sub></b>	<b>0</b>	<b>8</b>	<b>10</b>	<b>13</b>	<b>16</b>
<b>S<sub>2</sub></b>		<b>0</b>	<b>9</b>	<b>14</b>	<b>15</b>
<b>S<sub>3</sub></b>			<b>0</b>	<b>12</b>	<b>13</b>
<b>S<sub>4</sub></b>				<b>0</b>	<b>9</b>
<b>S<sub>5</sub></b>					<b>0</b>

**Distance Matrix**



**Compact Set Tree**

**Original Compact Set relations**

1 2 4

1 2 5

1 3 4

1 3 5

2 3 4

2 3 5

1 2 3

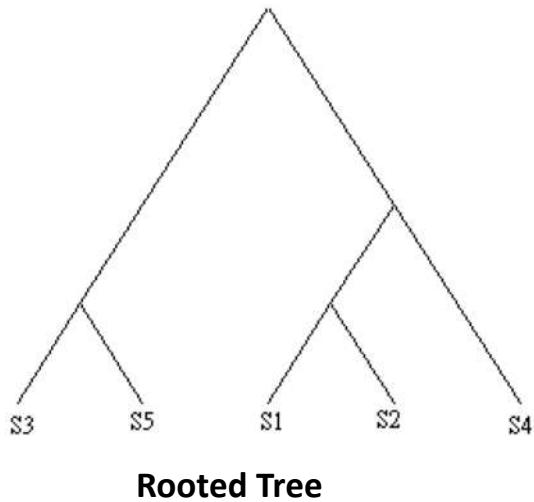
4 5 1

4 5 2

4 5 3

N1 = 10

## Measure of Compact Set Preservation(con't)



The relations preserved after MSA

1 2 4

1 2 5

1 3 4

1 3 5

2 3 4

After MSA

2 3 5

=====>

1 2 3

4 5 1

4 5 2

4 5 3

1 2 4

1 2 5

1 4 3 ×

3 5 1 ×

2 4 3 ×

3 5 2 ×

1 2 3

1 4 5 ×

2 4 5 ×

3 5 4 ×

$$N_2 = 10 - 7 = 3 \Rightarrow$$

Estimate by Compact Set Preservation = 3/10

# Compact Set Evaluation Algorithm

- Step1 : Construct the original Compact Set Tree  $T$  and the Compact Set Tree after MSA  $T'$  [1].
- Step2 : Preorder Traversal  $T'$  to generate the Compact Set relations after MSA  $R'$ ,and mark the entry in the hash table  $H'$  according to  $R'$ .
- Step3 : Preorder Traversal  $T$  to generate the Original Compact Set Relations  $R$ ,and check whether the marked entry in the hash table by  $R$  is a subset of the hash table  $H'$ .
- Total Time Complexity =  $O(n^3)$ ,where  $n$  is the number of sequences
- Reference:
  - 1. E. Dekel,J. Hu and W. Ouyang, An optimal algorithm for finding compact sets, *Inform. Process. Lett.* 44(1992) 285~289