Chapter 7:
Molecular phylogeny and evolution

Learning objectives

Upon completing this chapter you should be able to:

- describe the molecular clock hypothesis and explain its significance;
- define positive and negative selection and test its presence in sequences of interest;
- describe the types of phylogenetic trees and their parts (branches, nodes, roots);
- create phylogenetic trees using distance-based and character-based methods; and
- explain the basis of different approaches to creating phylogenetic trees and evaluating them.

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

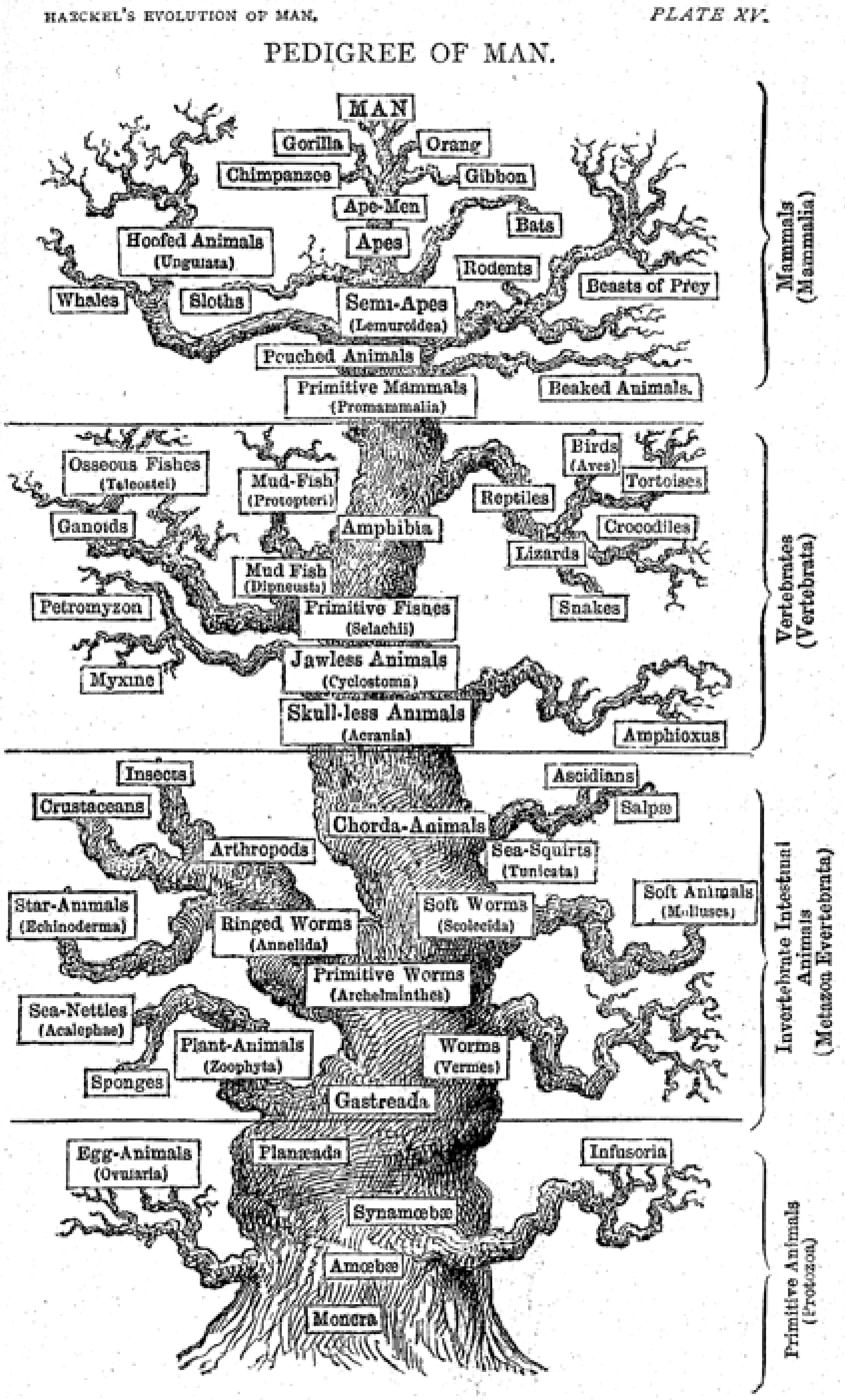
Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

Five kingdom system (Haeckel, 1879)

- animals
- plants
- fungi
- protists
- monera



mammals

vertebrates

invertebrates

protozoa

Introduction

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.

Introduction

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.

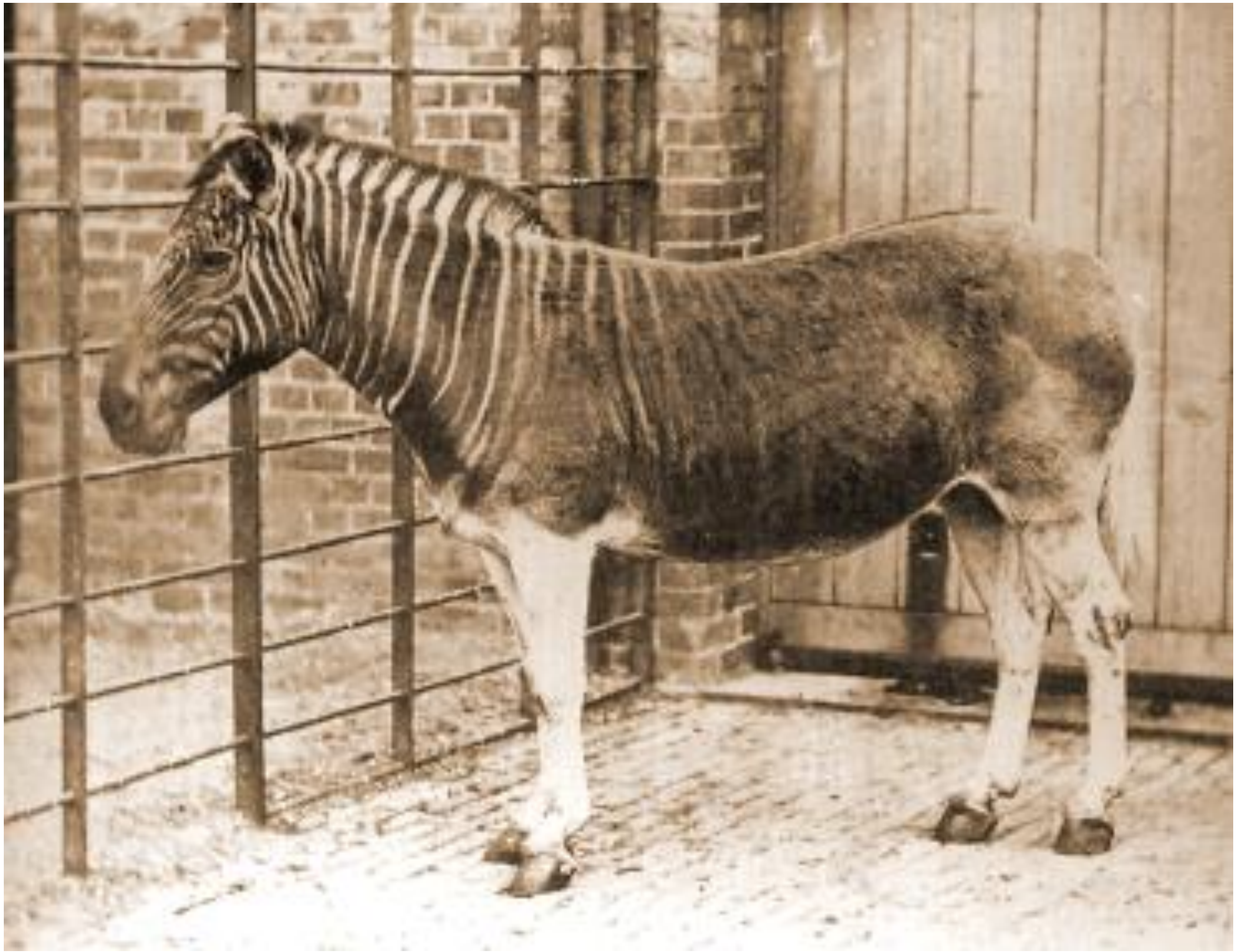
Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny relied on the comparison of morphological features between organisms. Today, molecular sequence data are also used for phylogenetic analyses.

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?

Was the quagga (now extinct) more like a zebra or a horse?



Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

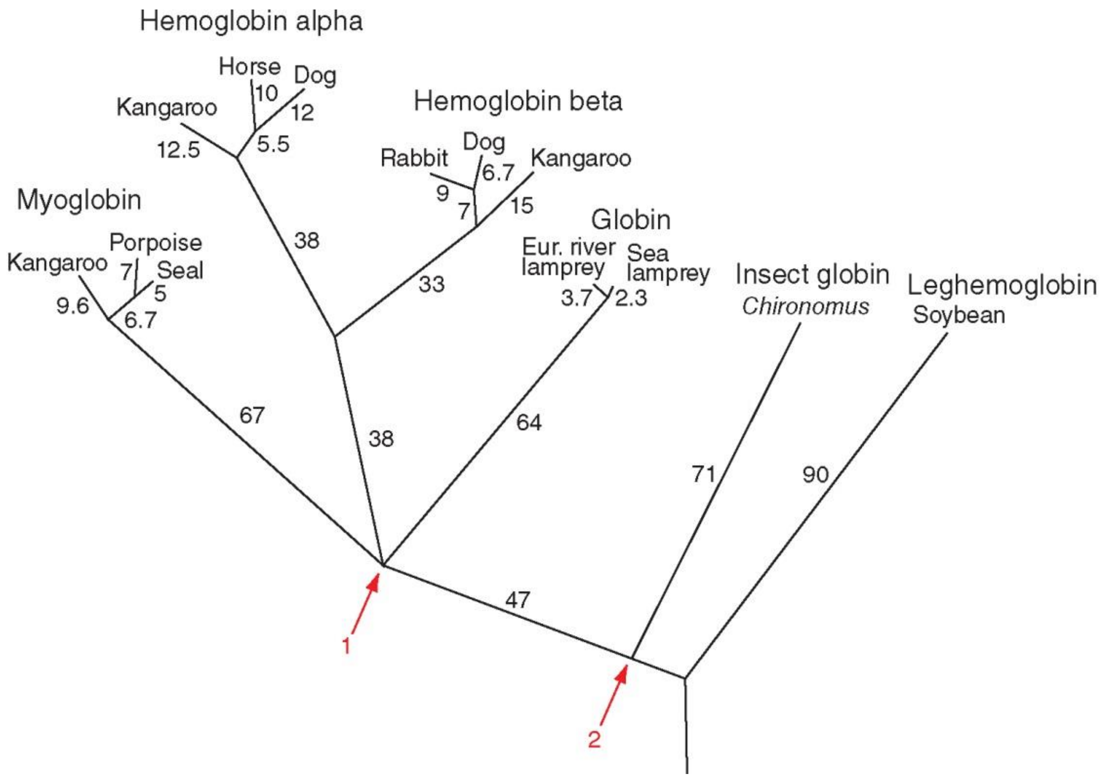
Type of trees (species trees vs. gene/protein trees; DNA or protein)

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

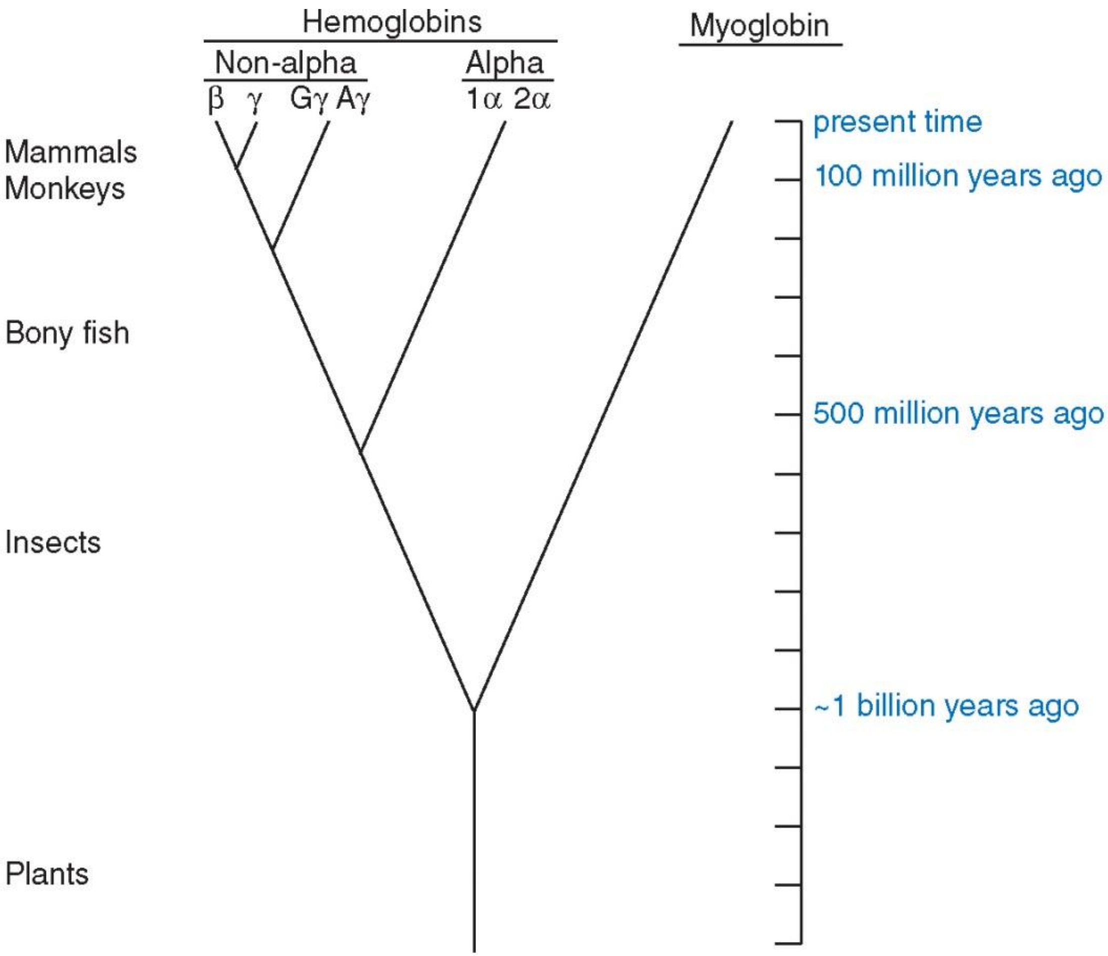
1960s: globin phylogeny (tree of 13 orthologs by Margaret Dayhoff and colleagues)



Arrow 1: node corresponding to last common ancestor of a group of vertebrate globins.

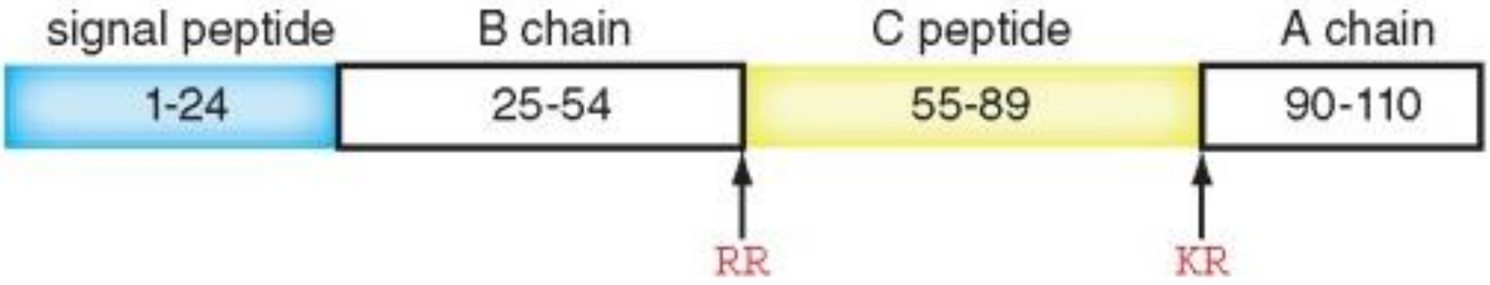
Arrow 2: ancestor of insect and vertebrate globins

1960s: globin phylogeny (tree of 7 paralogs)



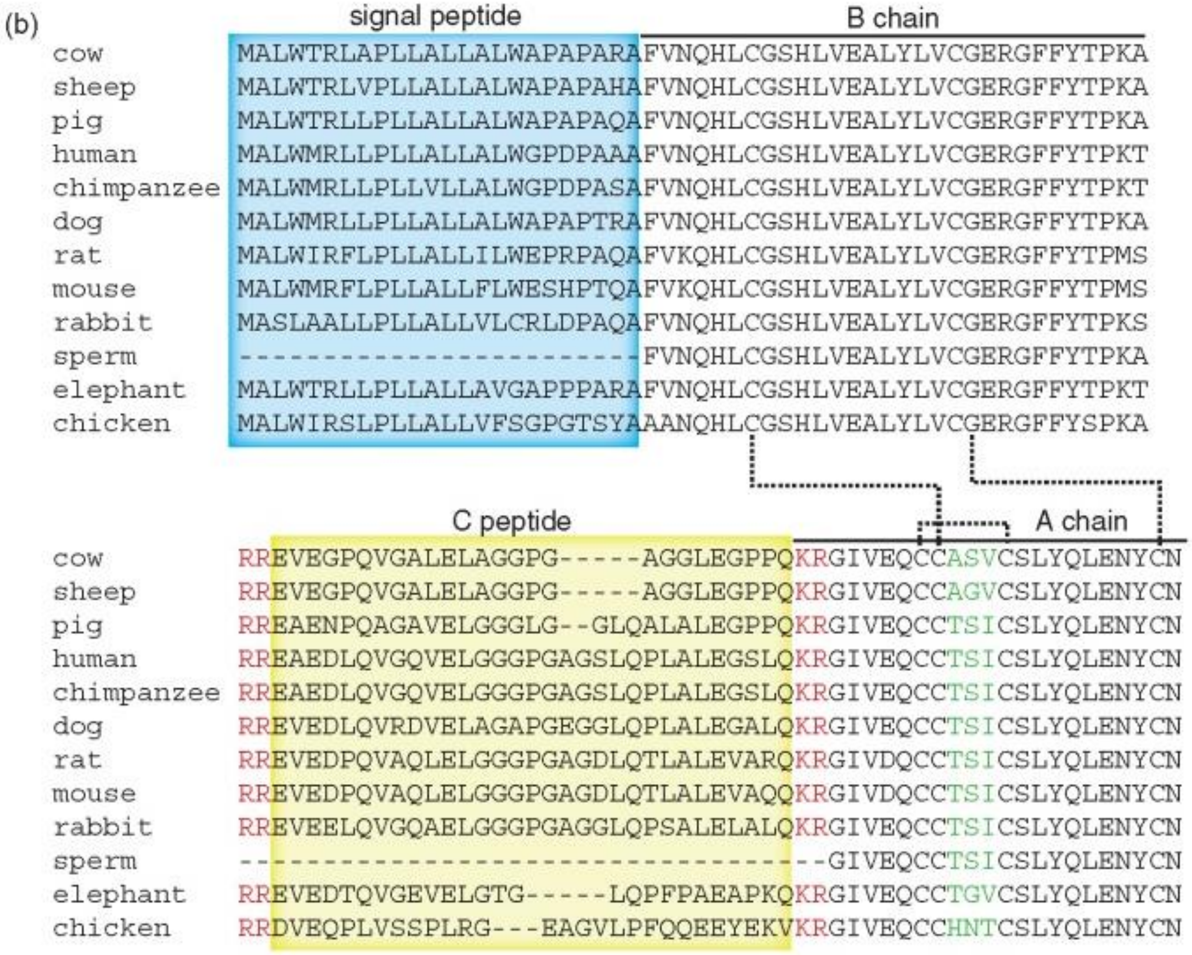
Dayhoff et al. (1972) analyzed related globins in the context of evolutionary time.

Insulin structure



Dibasic residues flank the C peptide which is cleaved and removed.

Insulin structure: conserved blocks



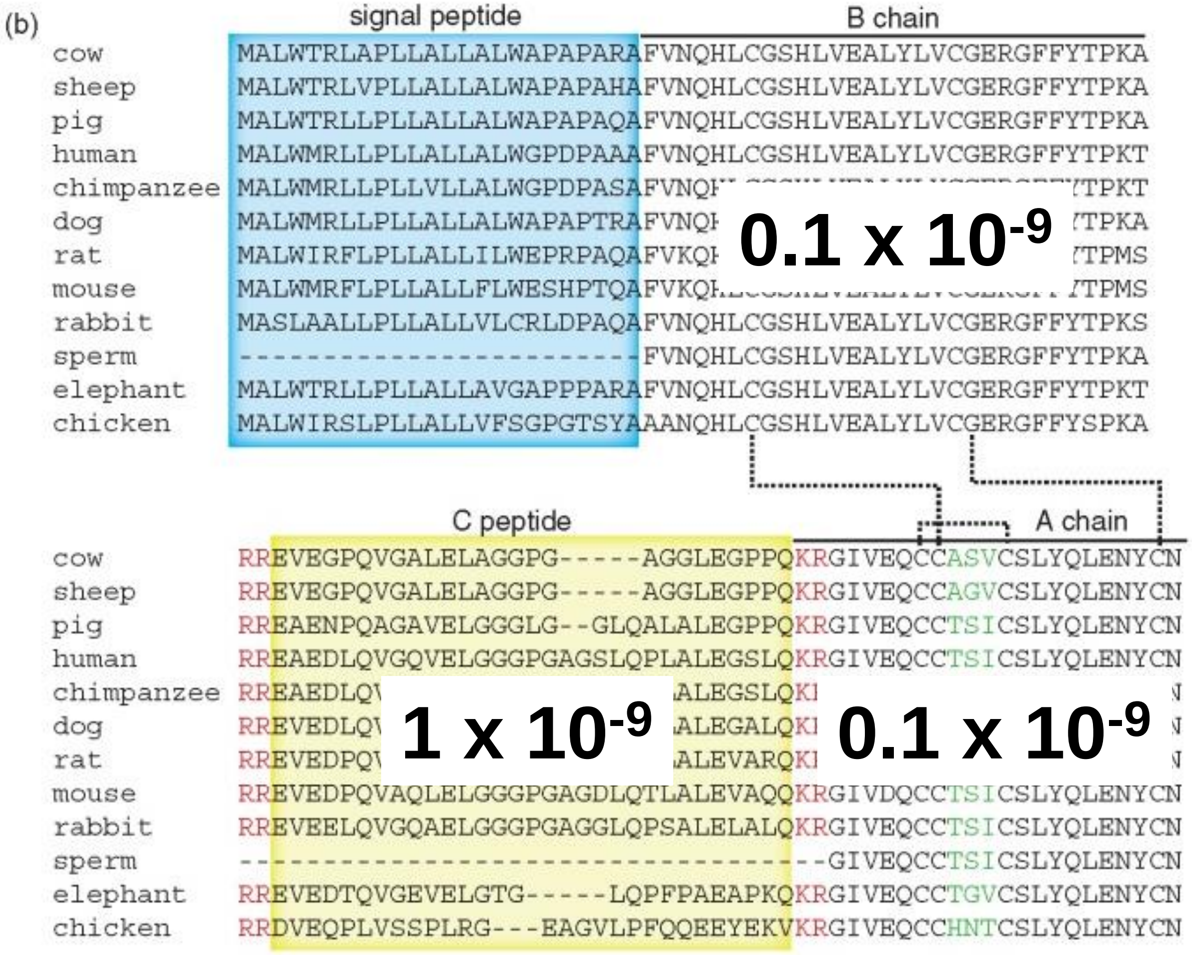
(b)

```
                signal peptide                         B chain
cow        MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
sheep      MALWTRLVPLLALLALWAPAPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
pig        MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
human      MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chimpanzee MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
dog        MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
rat        MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMS
mouse      MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMS
rabbit     MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKS
sperm      ------------------------FVNQHLCGSHLVEALYLVCGERGFFYTPKA
elephant   MALWTRLLPLLALLAVGAPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chicken    MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKA

                     C peptide                          A chain
cow        RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
sheep      RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCAGVCSLYQLENYCN
pig        RREAENPQAGAVELGGGLG--GLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
human      RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
chimpanzee RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
dog        RREVEDLQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN
rat        RREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
mouse      RREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
rabbit     RREVEELQVGQAELGGGPGAGGLQPSALELALQKRGIVEQCCTSICSLYQLENYCN
sperm      -----------------------------------GIVEQCCTSICSLYQLENYCN
elephant   RREVEDTQVGEVELGTG-----LQPFPAEAPKQKRGIVEQCCTGVCSLYQLENYCN
chicken    RRDVEQPLVSSPLRG---EAGVLPFQQEEYEKVKRGIVEQCCHNTCSLYQLENYCN
```

The residues in the B and A chains are highly conserved across species. The rate of nucleotide substitution is 6- to 10-fold higher in the C chain region.

B&FG 3e
Fig. 7.3
Page 249

Insulin structure: conserved blocks



(b)

```
                  signal peptide           B chain
cow        MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
sheep      MALWTRLVPLLALLALWAPAPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
pig        MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
human      MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chimpanzee MALWMRLLPLLVLLALWGPDPASAFVNQ                      TPKT
dog        MALWMRLLPLLALLALWAPAPTRAFVNQ                      TPKA
rat        MALWIRFLPLLALLILWEPRPAQAFVKQ                      TPMS
mouse      MALWMRFLPLLALLFLWESHPTQAFVKQ                      TPMS
rabbit     MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKS
sperm      ------------------------FVNQHLCGSHLVEALYLVCGERGFFYTPKA
elephant   MALWTRLLPLLALLAVGAPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chicken    MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKA

                          C peptide                         A chain
cow        RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
sheep      RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCAGVCSLYQLENYCN
pig        RREAENPQAGAVELGGGLG--GLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
human      RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
chimpanzee RREAEDLQ                 ALEGSLQK                      N
dog        RREVEDLQ                 ALEGALQK                      N
rat        RREVEDPQ                 ALEVARQK                      N
mouse      RREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
rabbit     RREVEELQVGQAELGGGPGAGGLQPSALELALQKRGIVEQCCTSICSLYQLENYCN
sperm      ---------------------------------GIVEQCCTSICSLYQLENYCN
elephant   RREVEDTQVGEVELGTG-----LQPFPAEAPKQKRGIVEQCCTGVCSLYQLENYCN
chicken    RRDVEQPLVSSPLRG---EAGVLPFQEEYEKVKRGIVEQCCHNTCSLYQLENYCN
```

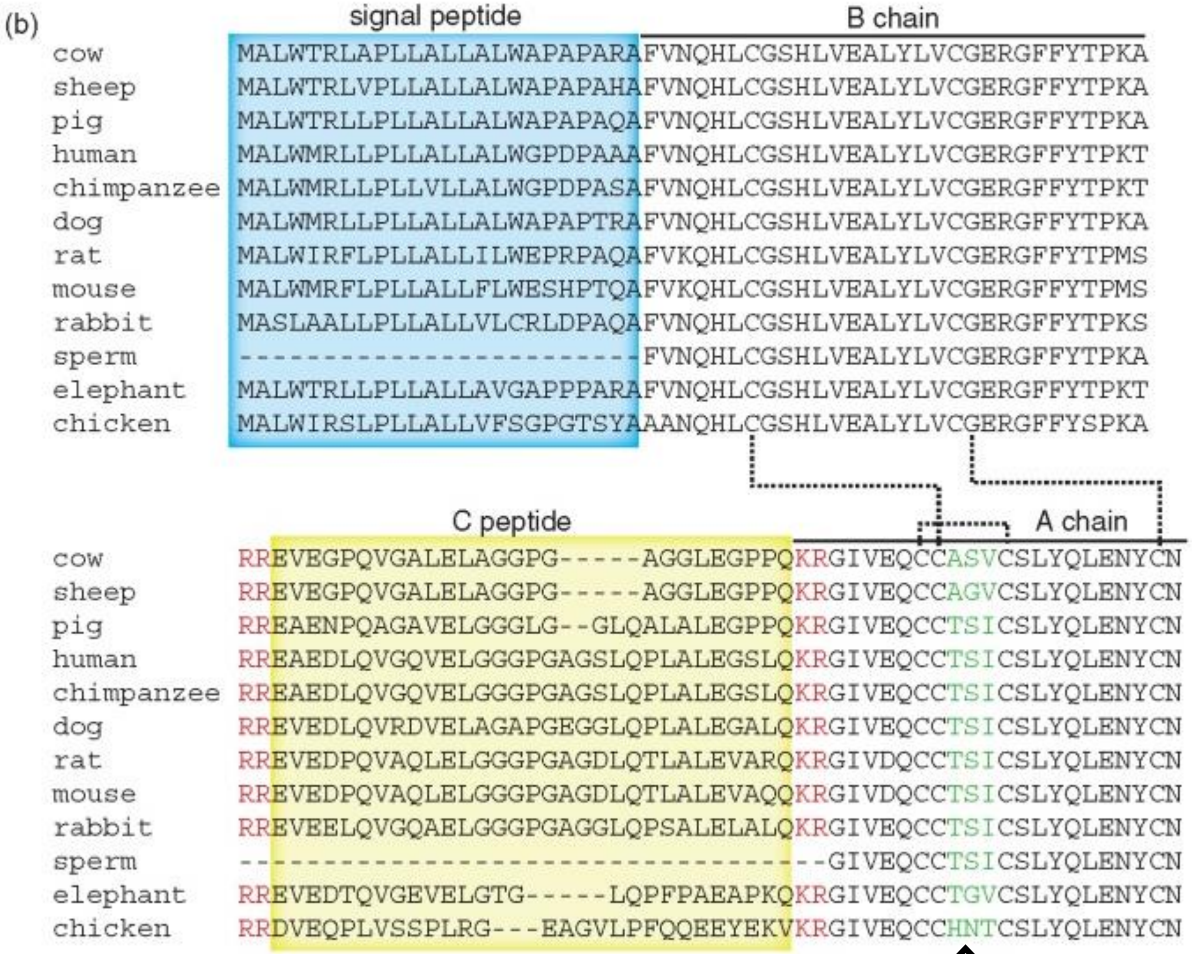
0.1×10^{-9}

1×10^{-9}

0.1×10^{-9}

Number of nucleotide substitutions/site/year

Insulin structure: conserved blocks



(b)

```
                signal peptide          B chain
cow        MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
sheep      MALWTRLVPLLALLALWAPAPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
pig        MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
human      MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chimpanzee MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
dog        MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKA
rat        MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMS
mouse      MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMS
rabbit     MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKS
sperm      ------------------------FVNQHLCGSHLVEALYLVCGERGFFYTPKA
elephant   MALWTRLLPLLALLAVGAPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
chicken    MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKA

                C peptide                               A chain
cow        RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
sheep      RREVEGPQVGALELAGGPG-----AGGLEGPPQKRGIVEQCCAGVCSLYQLENYCN
pig        RREAENPQAGAVELGGGLG--GLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
human      RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
chimpanzee RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
dog        RREVEDLQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN
rat        RREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
mouse      RREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
rabbit     RREVEELQVGQAELGGGPGAGGLQPSALELALQKRGIVEQCCTSICSLYQLENYCN
sperm      -----------------------------------GIVEQCCTSICSLYQLENYCN
elephant   RREVEDTQVGEVELGTG-----LQPFPAEAPKQKRGIVEQCCTGVCSLYQLENYCN
chicken    RRDVEQPLVSSPLRG---EAGVLPFQQEEYEKVKRGIVEQCCHNTCSLYQLENYCN
```

Note the sequence divergence in the disulfide loop region of the A chain. This is a spacer region that is under less evolutionary constraint.

Historical background: insulin

By the 1950s, it became clear that amino acid substitutions occur nonrandomly. For example, Sanger and colleagues noted that most amino acid changes in the insulin A chain are restricted to a disulfide loop region. Such differences are called "neutral" changes (Kimura, 1968; Jukes and Cantor, 1969).

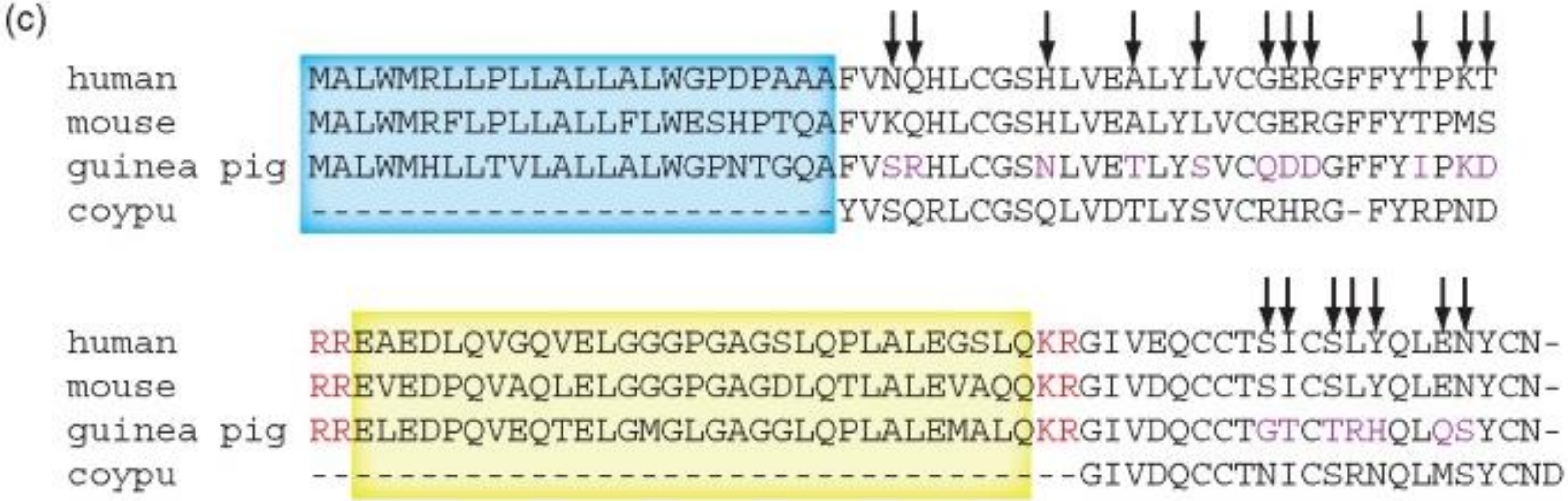
Subsequent studies at the DNA level showed that rate of nucleotide (and of amino acid) substitution is about six- to ten-fold higher in the C peptide, relative to the A and B chains.

Historical background: insulin

Surprisingly, insulin from the guinea pig (and from the related coypu) evolve seven times faster than insulin from other species. Why?

The answer is that guinea pig and coypu insulin do not **bind two zinc ions**, while insulin molecules from most other species do. There was a relaxation on the structural constraints of these molecules, and so the genes diverged rapidly.

Guinea pig and coypu insulins have evolved 7-fold faster than insulin from other species

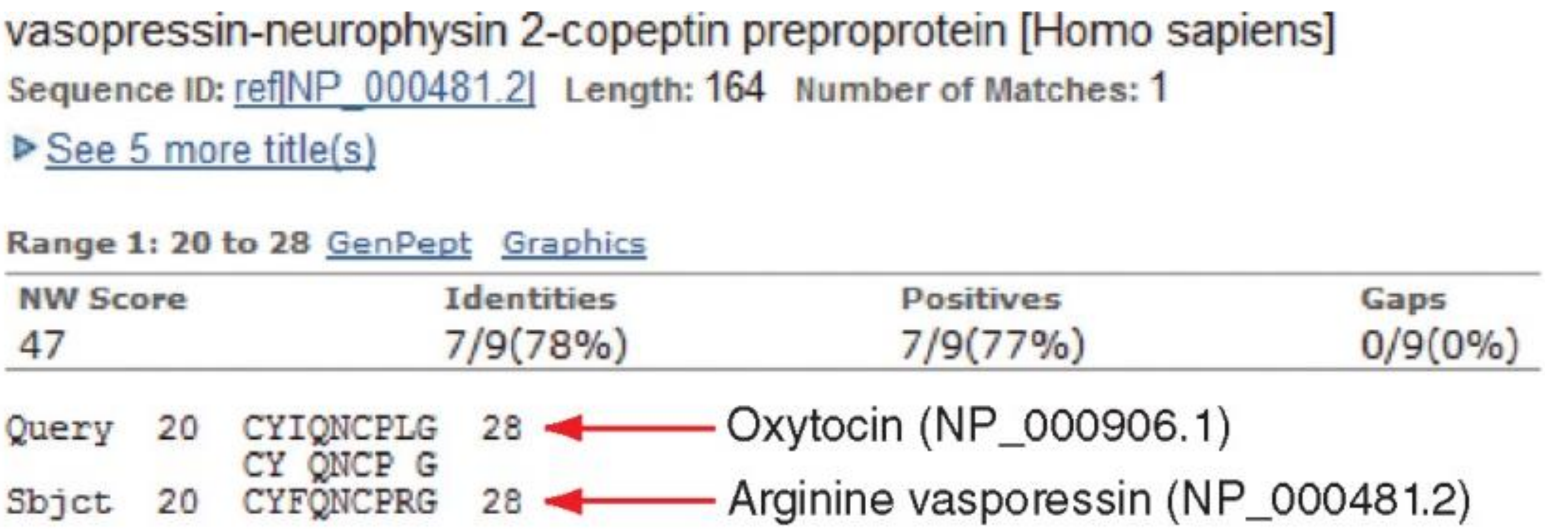


(c)

```
human       MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
mouse       MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMS
guinea pig  MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKD
coypu       ------------------------YVSQRLCGSQLVDTLYSVCRHRG-FYRPND

human       RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN-
mouse       RREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN-
guinea pig  RRELEDPQVEQTELGMGLGAGGLQPLALEMALQKRGIVDQCCTGTCTRHQLQSYCN-
coypu       -----------------------------------GIVDQCCTNICSRNQLMSYCND
```

Arrows indicate 18 amino acid positions at which guinea pig sequences vary from those of human and/or mouse

Early (1960s) insights into protein evolution: oxytocin and vasopressin differ by only two amino acid residues but have vastly different functions



vasopressin-neurophysin 2-copeptin preproprotein [Homo sapiens]

Sequence ID: ref|NP_000481.2| Length: 164 Number of Matches: 1

▶ See 5 more title(s)

Range 1: 20 to 28 GenPept Graphics

NW Score	Identities	Positives	Gaps
47	7/9(78%)	7/9(77%)	0/9(0%)

```
Query  20  CYIQNCPLG  28
           CY QNCP G
Sbjct  20  CYFQNCPRG  28
```

Oxytocin (NP_000906.1)

Arginine vasporessin (NP_000481.2)

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

Molecular clock hypothesis

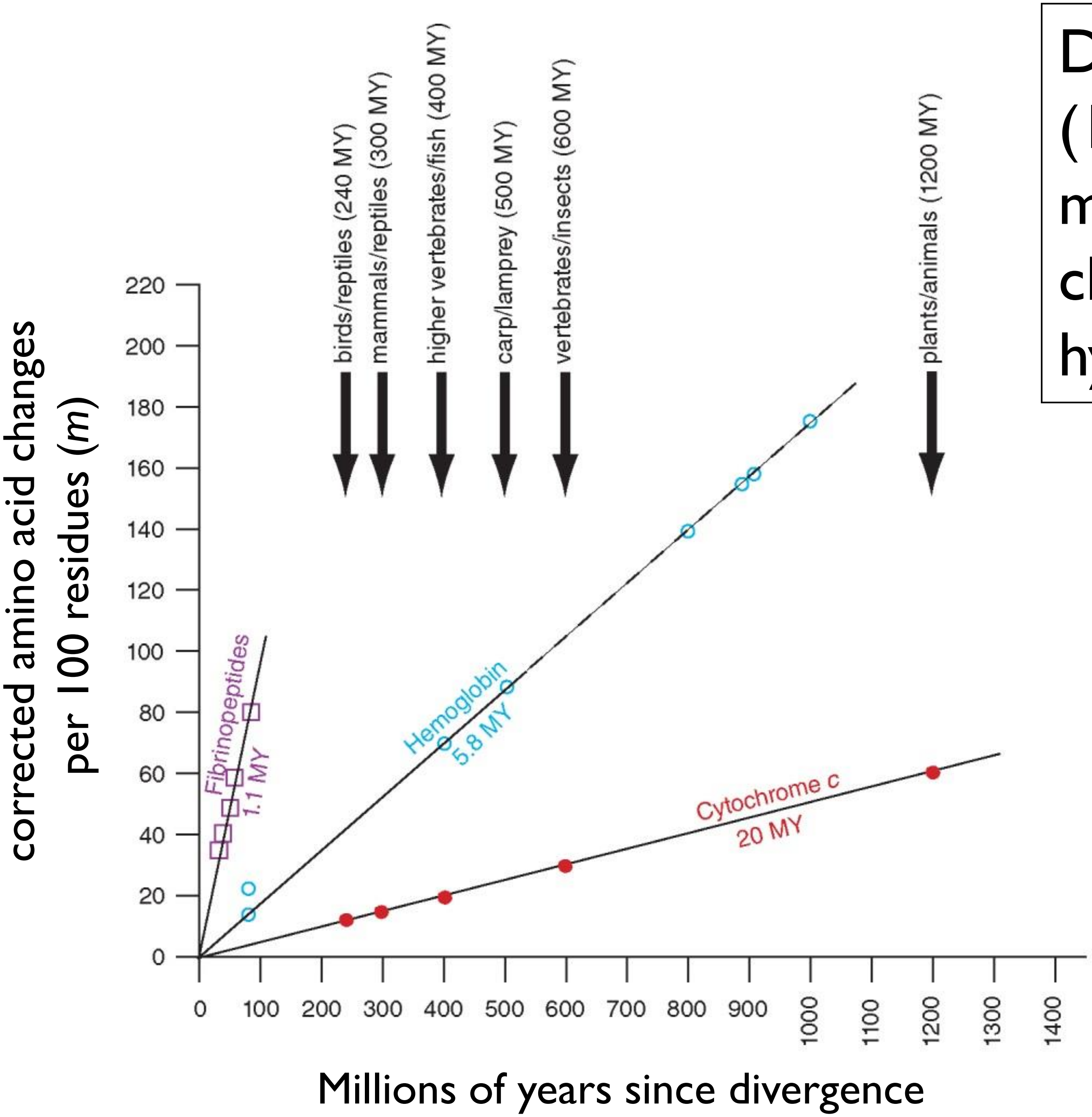
As an example, Richard Dickerson (1971) plotted data from three protein families: cytochrome *c*, hemoglobin, and fibrinopeptides.

The x-axis shows the divergence times of the species, estimated from paleontological data. The y-axis shows *m*, the corrected number of amino acid changes per 100 residues.

n is the observed number of amino acid changes per 100 residues, and it is corrected to *m* to account for changes that occur but are not observed.

$$\frac{N}{100} = 1 - e^{-(m/100)}$$

Dickerson (1971): the molecular clock hypothesis



Molecular clock hypothesis: conclusions

Dickerson drew the following conclusions:

- For each protein, the data lie on a straight line. Thus, the rate of amino acid substitution has remained constant for each protein.
- The average rate of change differs for each protein. The time for a 1% change to occur between two lines of evolution is 20 MY (cytochrome c), 5.8 MY (hemoglobin), and 1.1 MY (fibrinopeptides).
- The observed variations in rate of change reflect functional constraints imposed by natural selection.

Molecular clock hypothesis: implications

If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.

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-20th 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.

Neutral theory of evolution

An often-held view of evolution is that just as organisms propagate through natural selection, so also DNA and protein molecules are selected for.

According to Motoo Kimura's 1968 neutral theory of molecular evolution, the vast majority of DNA changes are **not** selected for in a Darwinian sense. The main cause of **evolutionary change is random drift of mutant alleles** that are selectively neutral (or nearly neutral). Positive Darwinian selection does occur, but it has a limited role.

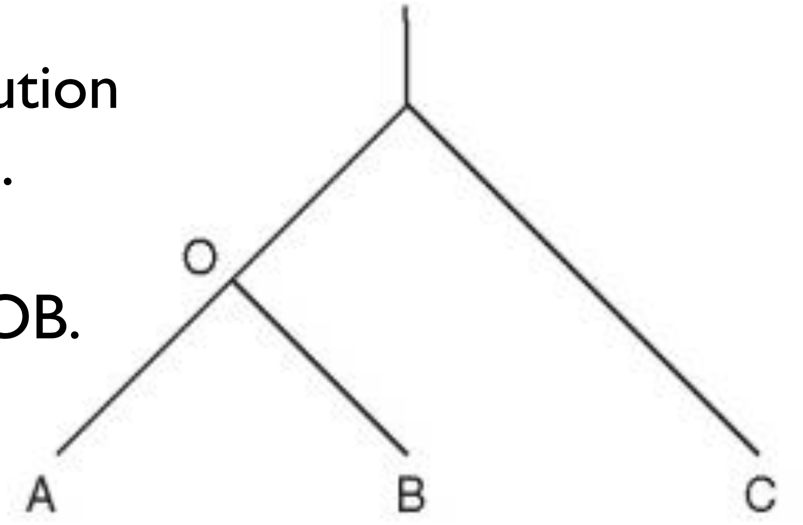
As an example, the divergent C peptide of insulin changes according to the neutral mutation rate.

Relative rate test to test the molecular clock

Test whether protein (or DNA) from organisms A, B evolve at the same rate (Tajima, 1993). Define a common ancestor (O) and select an appropriate outgroup (C).

We will measure substitution rates for AB, AC, and BC.

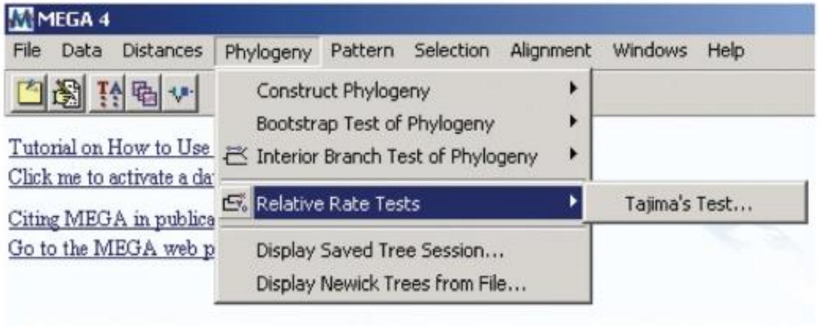
We will infer rates OA, OB.



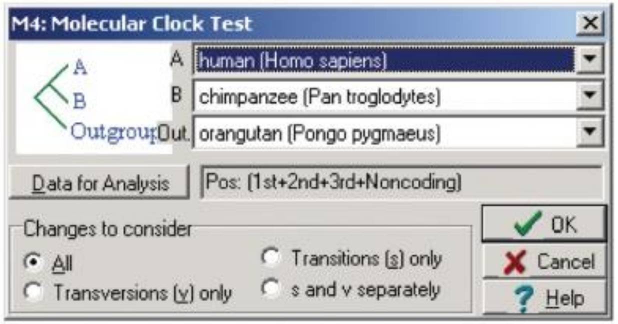
We will perform a chi square (χ^2) test to determine if those rates are comparable (null hypothesis) or whether we can reject the null at a significance level of $p < 0.05$.

Relative rate test to test the molecular clock

Tajima's test is implemented in MEGA (phylogeny pull-down)



In this example
A=human mitochondrial DNA
B=chimp
C=orang-utan (outgroup)



A: human (Homo sapiens)
B: chimpanzee (Pan troglodytes)
Out.: orangutan (Pongo pygmaeus)
Data for Analysis: Pos: (1st+2nd+3rd+Noncoding)
Changes to consider: ● All

The output shows p<0.05. We reject the null hypothesis of equal rates of evolution between human and chimp lineages.

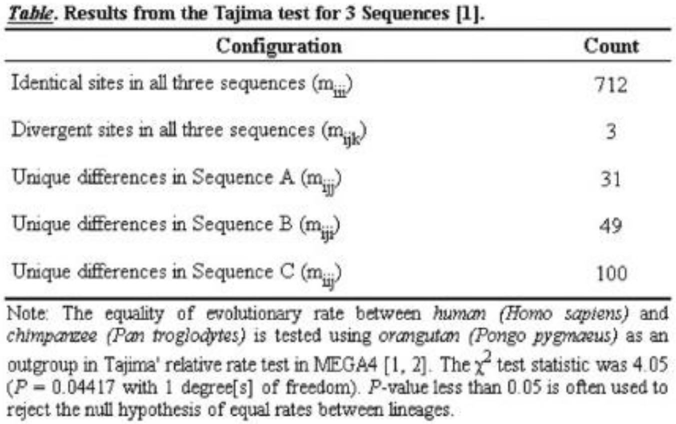
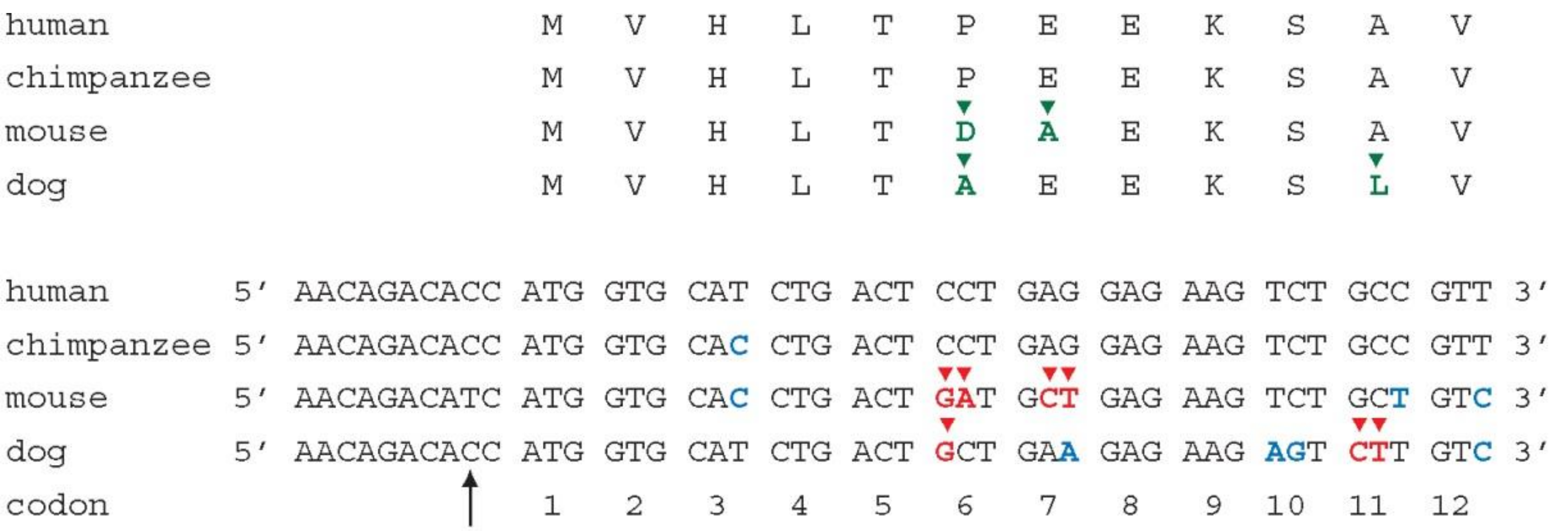


***Table*. Results from the Tajima test for 3 Sequences [1].**

Configuration	Count
Identical sites in all three sequences (m_{iii})	712
Divergent sites in all three sequences (m_{ijk})	3
Unique differences in Sequence A (m_{ijj})	31
Unique differences in Sequence B (m_{iji})	49
Unique differences in Sequence C (m_{iij})	100

Note: The equality of evolutionary rate between *human (Homo sapiens)* and *chimpanzee (Pan troglodytes)* is tested using *orangutan (Pongo pygmaeus)* as an outgroup in Tajima' relative rate test in MEGA4 [1, 2]. The χ^2 test statistic was 4.05 (*P* = 0.04417 with 1 degree[s] of freedom). *P*-value less than 0.05 is often used to reject the null hypothesis of equal rates between lineages.

Consider using DNA, RNA, or protein for phylogeny



```
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
```

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).

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

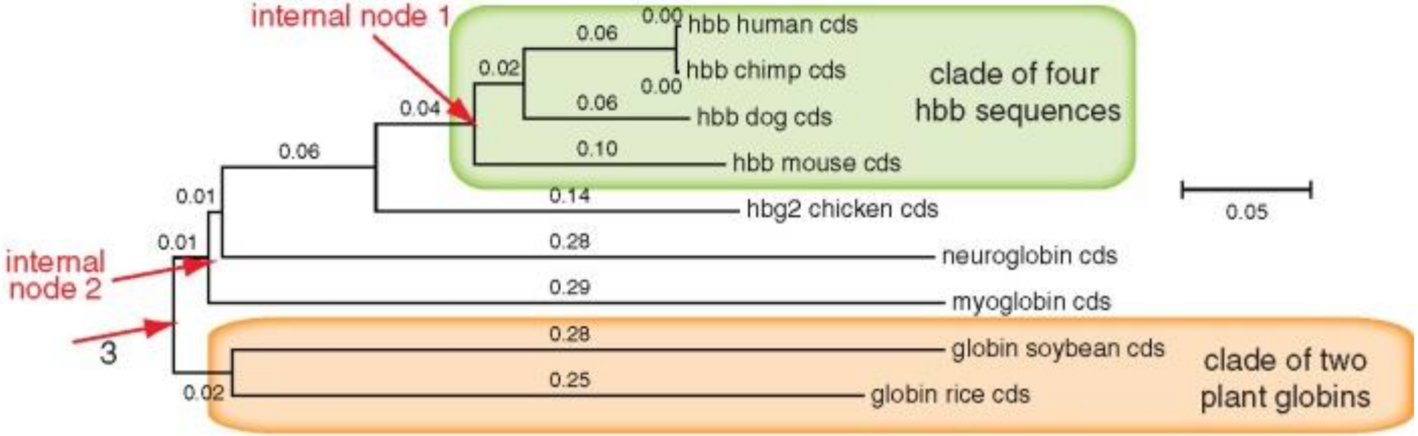
Perspective

Molecular phylogeny: nomenclature of trees

There are two main kinds of information inherent to any tree: topology and branch lengths.

We will now describe the parts of a tree.

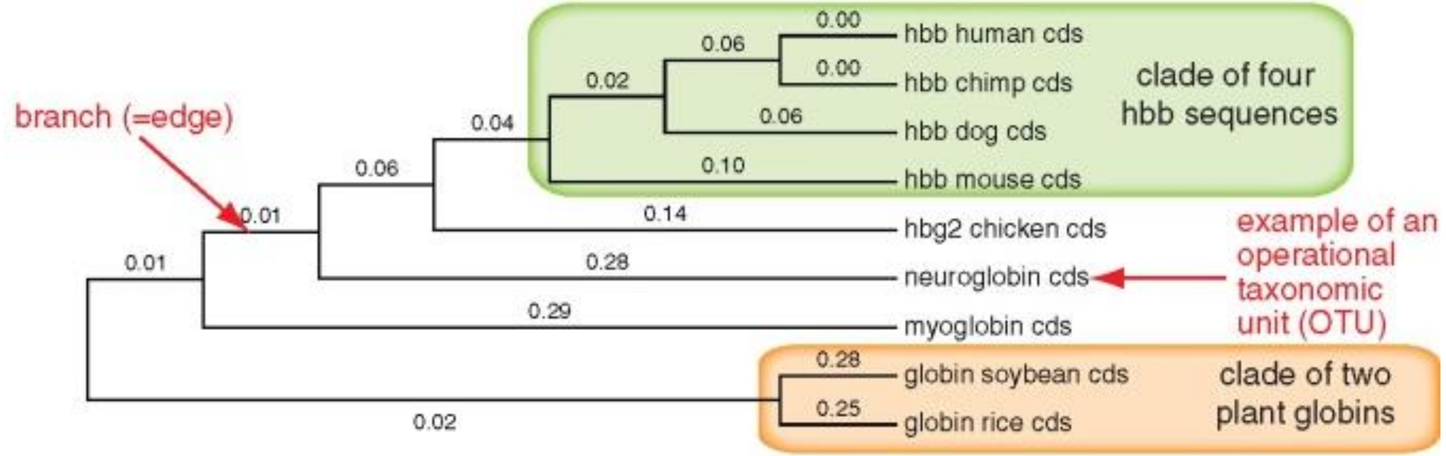
Nine globin coding sequences: neighbor-joining tree (rectangular tree style)



Nine globin DNA coding sequences were imported into MEGA, aligned with MUSCLE, and the branches and nodes are displayed in four different ways.

Note here that there are external nodes (extant sequences at the right) and internal nodes (each represents an ancestral sequence).

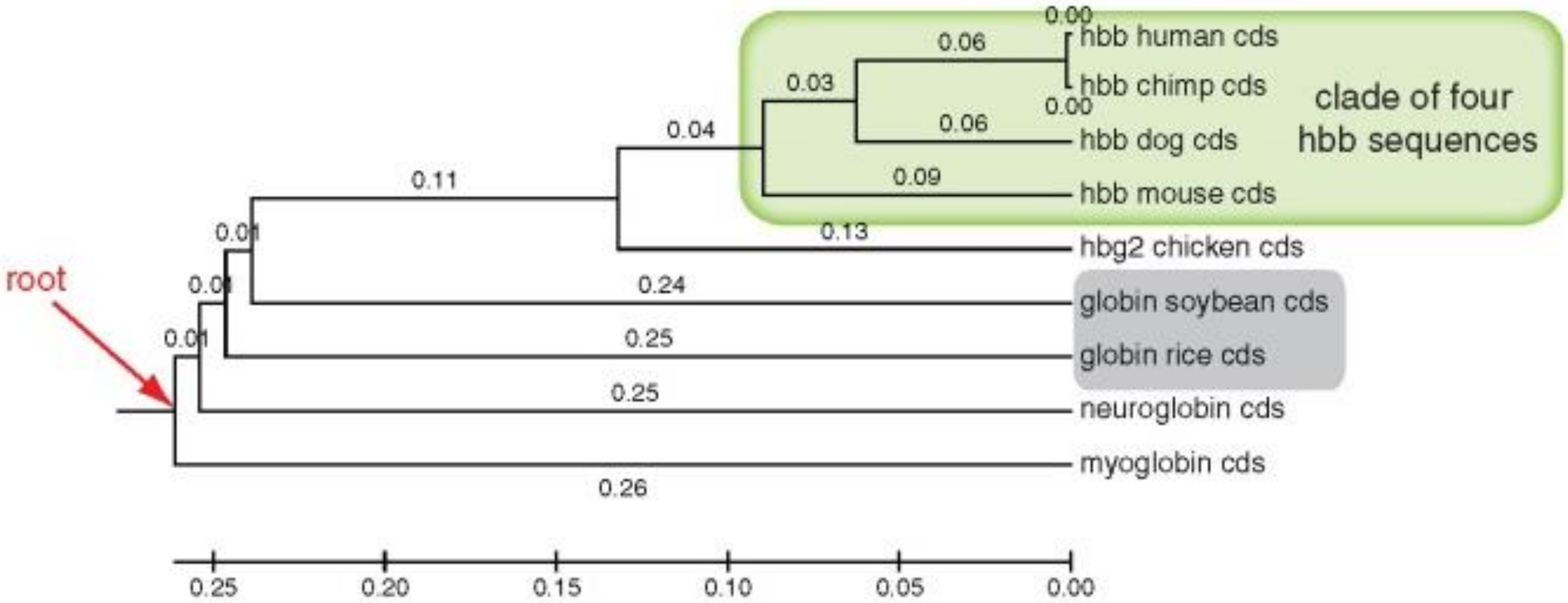
Nine globin coding sequences: neighbor-joining tree ("topology only" tree style)



Advantage of this display format: external nodes are lined up neatly to the right.

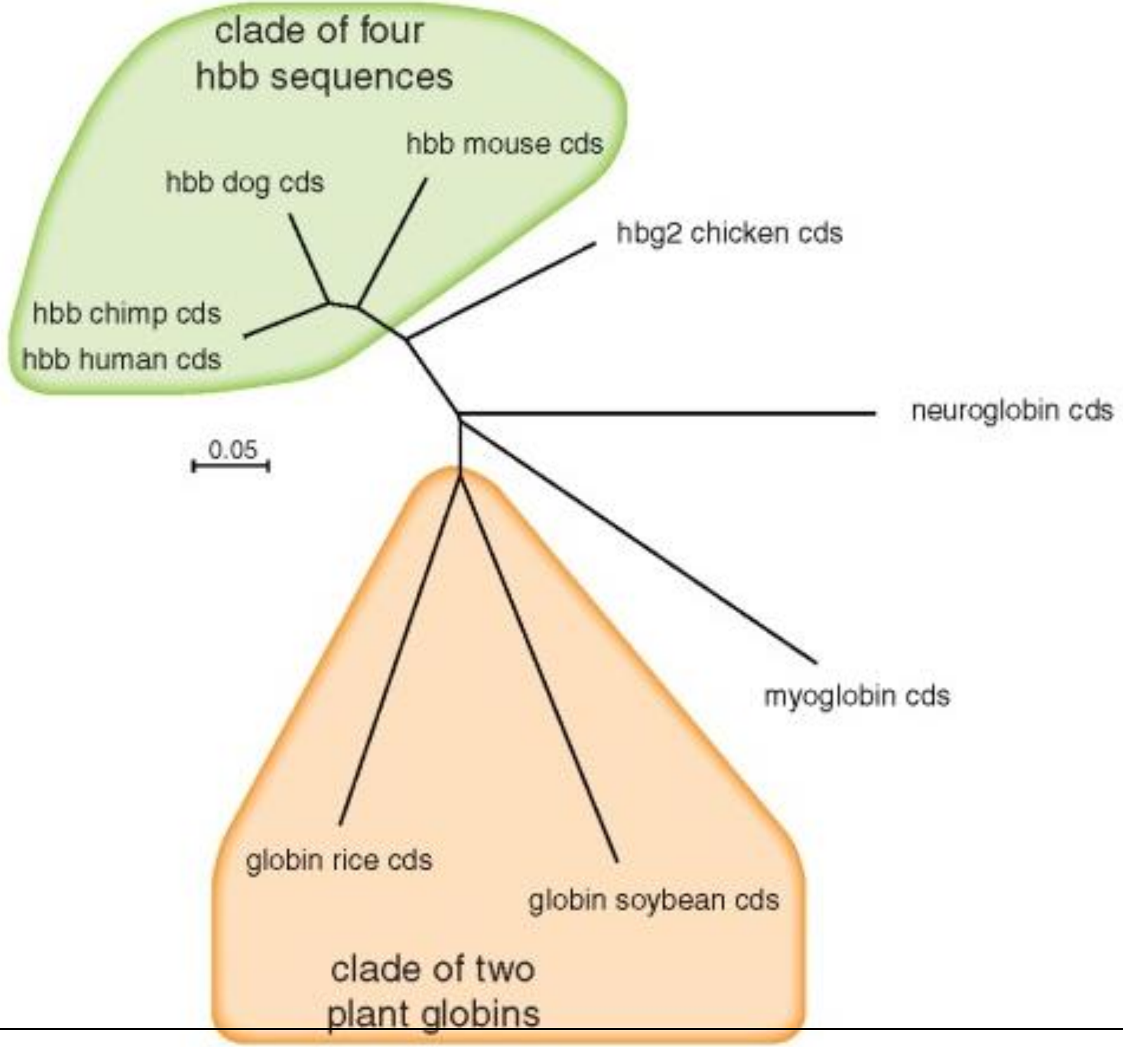
Disadvantage: branch lengths are not proportional to the values (as they were in the previous slide).

Nine globin coding sequences: UPGMA tree



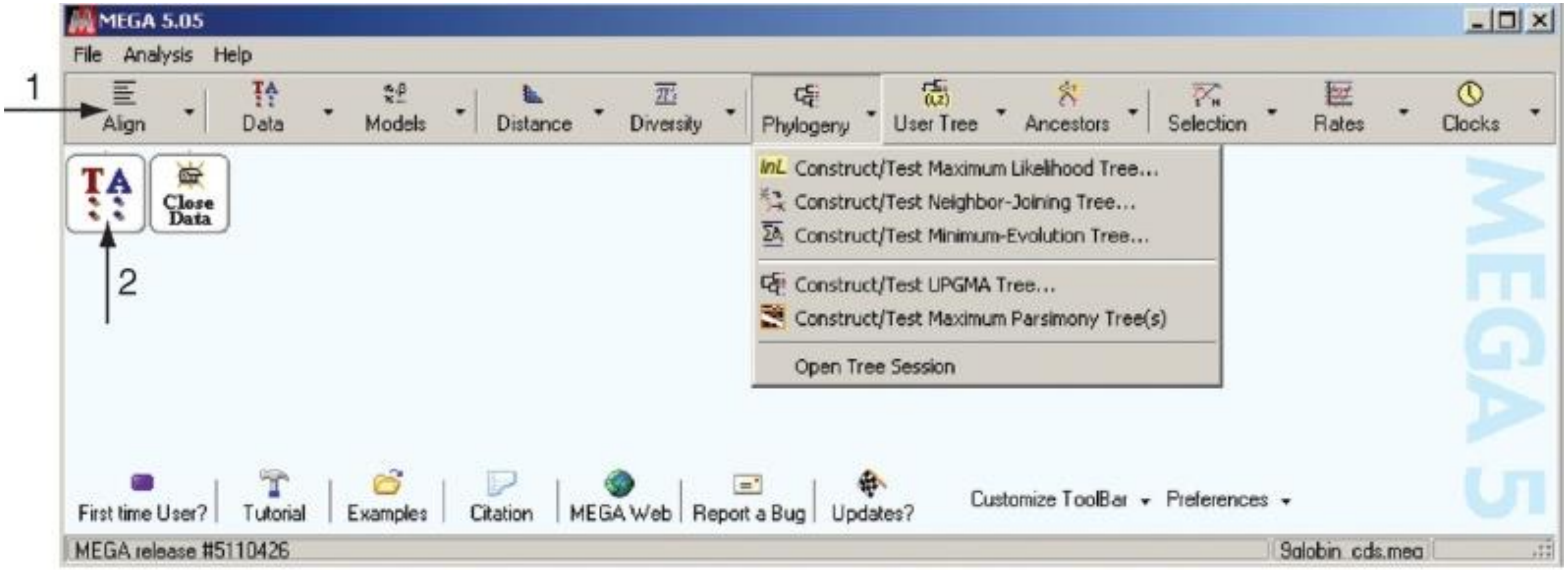
We define UPGMA below. Note that this tree is rooted. The topology of the two plant globins has changed: they now are (unrealistically) members of a clade with vertebrate globins)

Nine globin coding sequences: neighbor-joining tree (radial tree style)



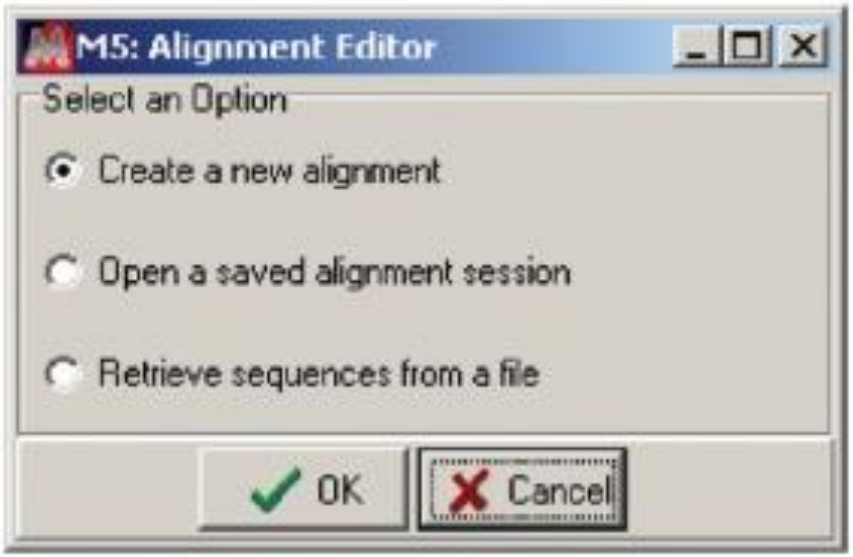
You may choose how to display your data. Be sure to define the scale bar; here it is nucleotide substitutions.

MEGA software for phylogenetic analyses: main dialog box



MEGA is freely available from http://www.megasoftware.net. Visit that site for a manual and publications.

MEGA software for phylogenetic analyses: alignment editor to create or open an alignment



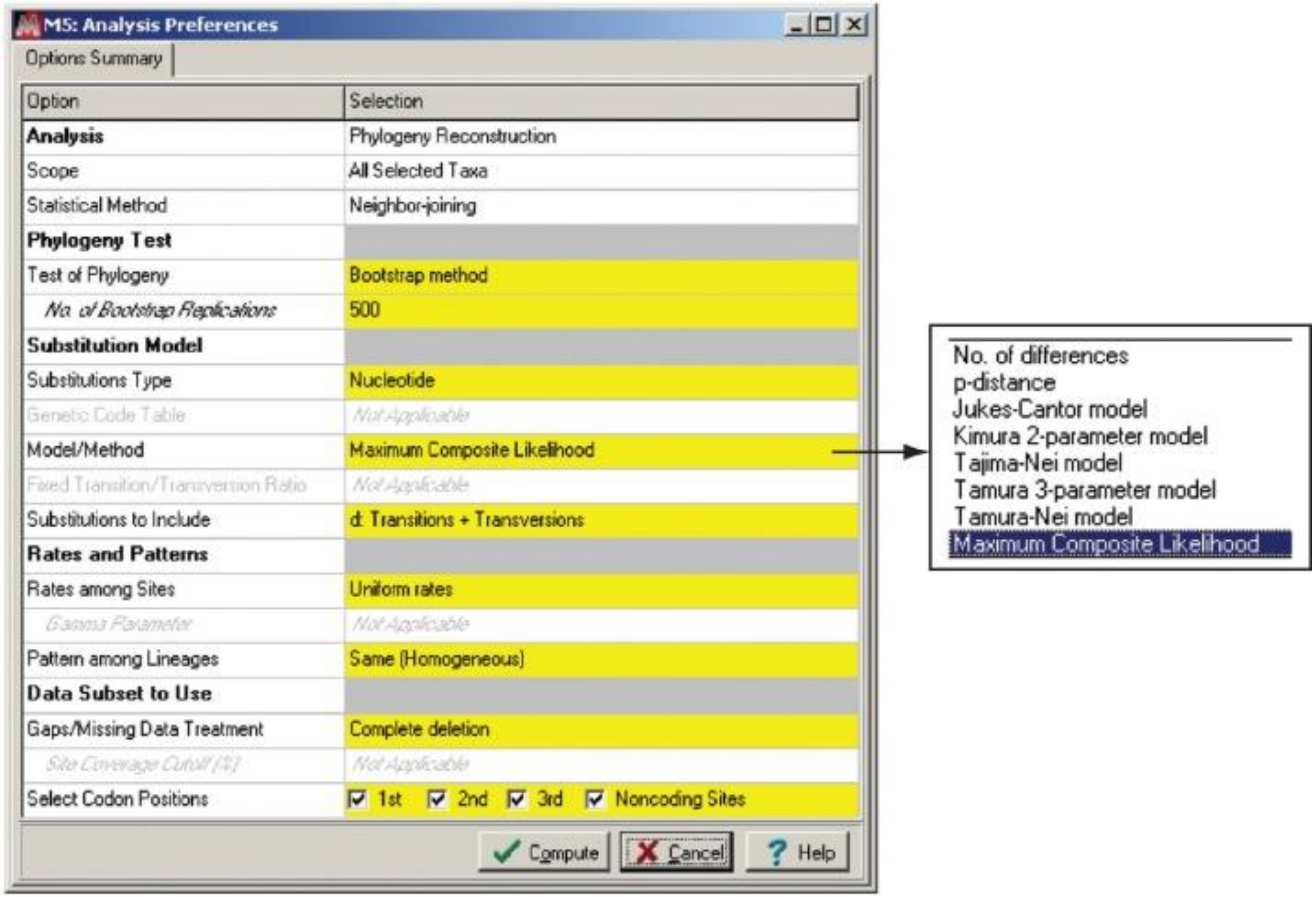
Select an Option

● Create a new alignment

○ Open a saved alignment session

○ Retrieve sequences from a file

MEGA software for phylogenetic analyses: analysis preferences dialog box

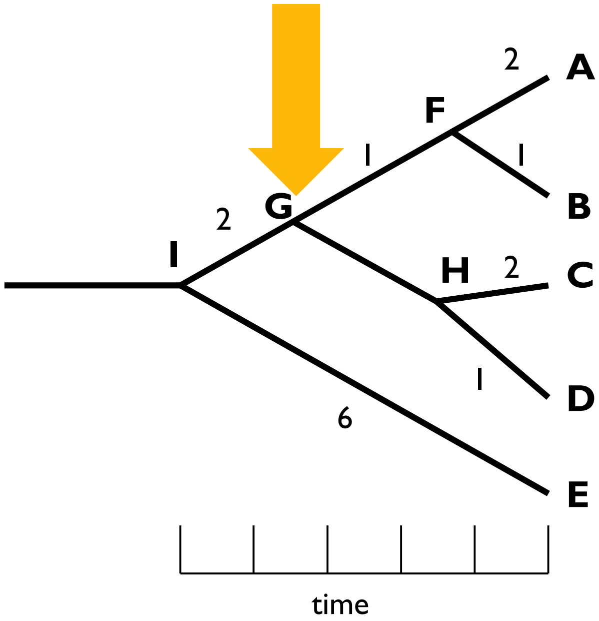


Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Maximum Composite Likelihood
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Uniform rates
Gamma Parameter	*Not Applicable*
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

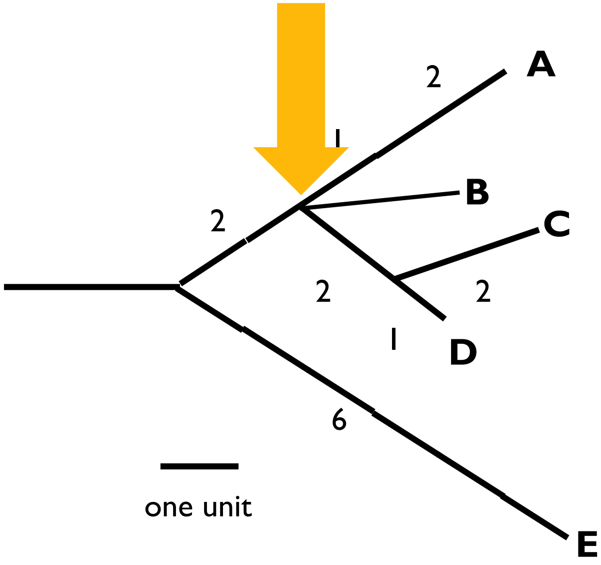
No. of differences
p-distance
Jukes-Cantor model
Kimura 2-parameter model
Tajima-Nei model
Tamura 3-parameter model
Tamura-Nei model
Maximum Composite Likelihood

Tree nomenclature

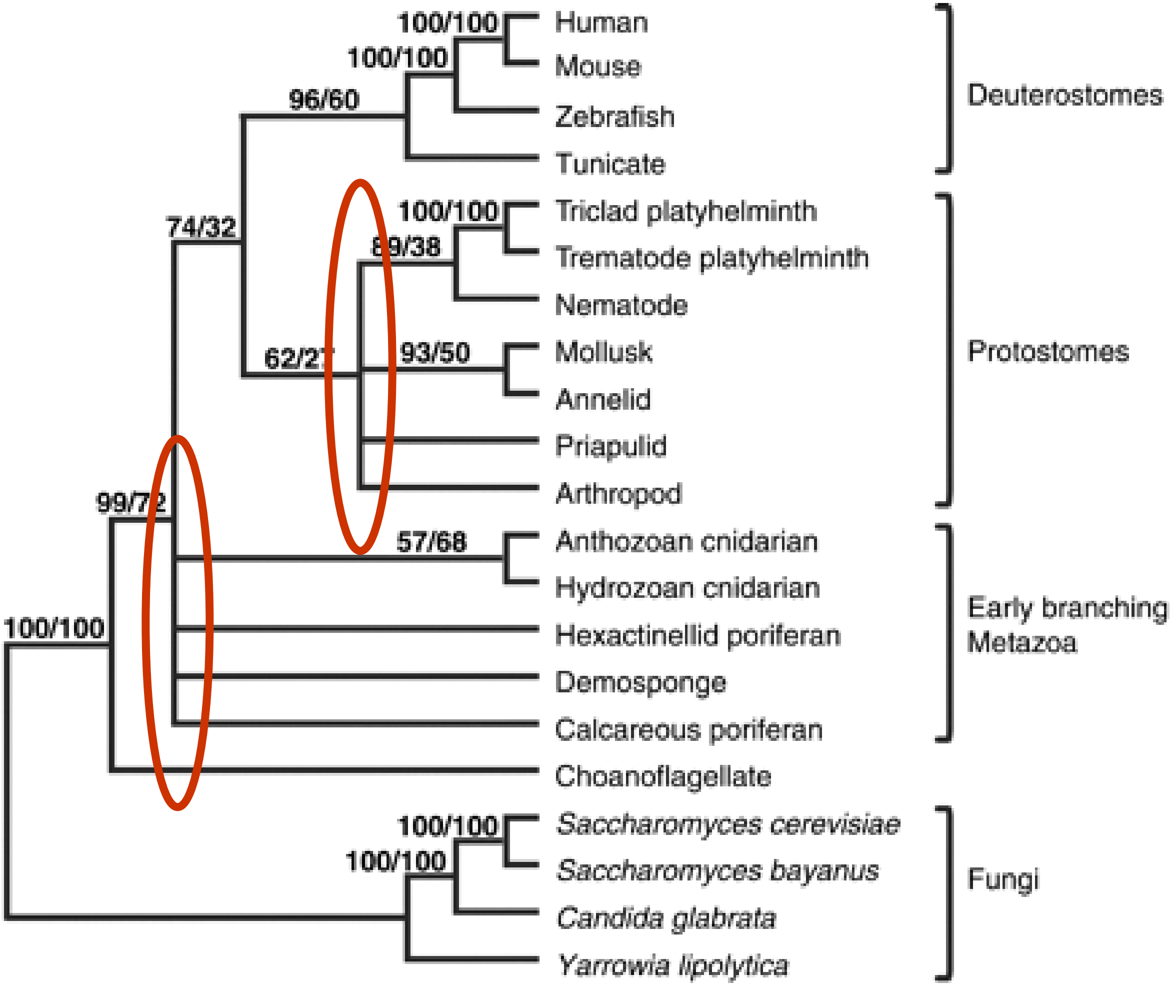
bifurcating
internal
node



multifurcating
internal
node



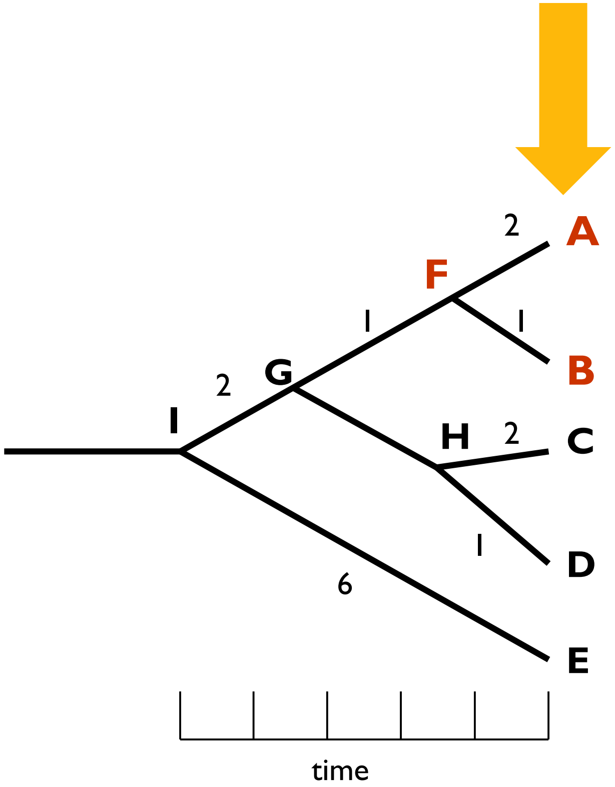
Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes



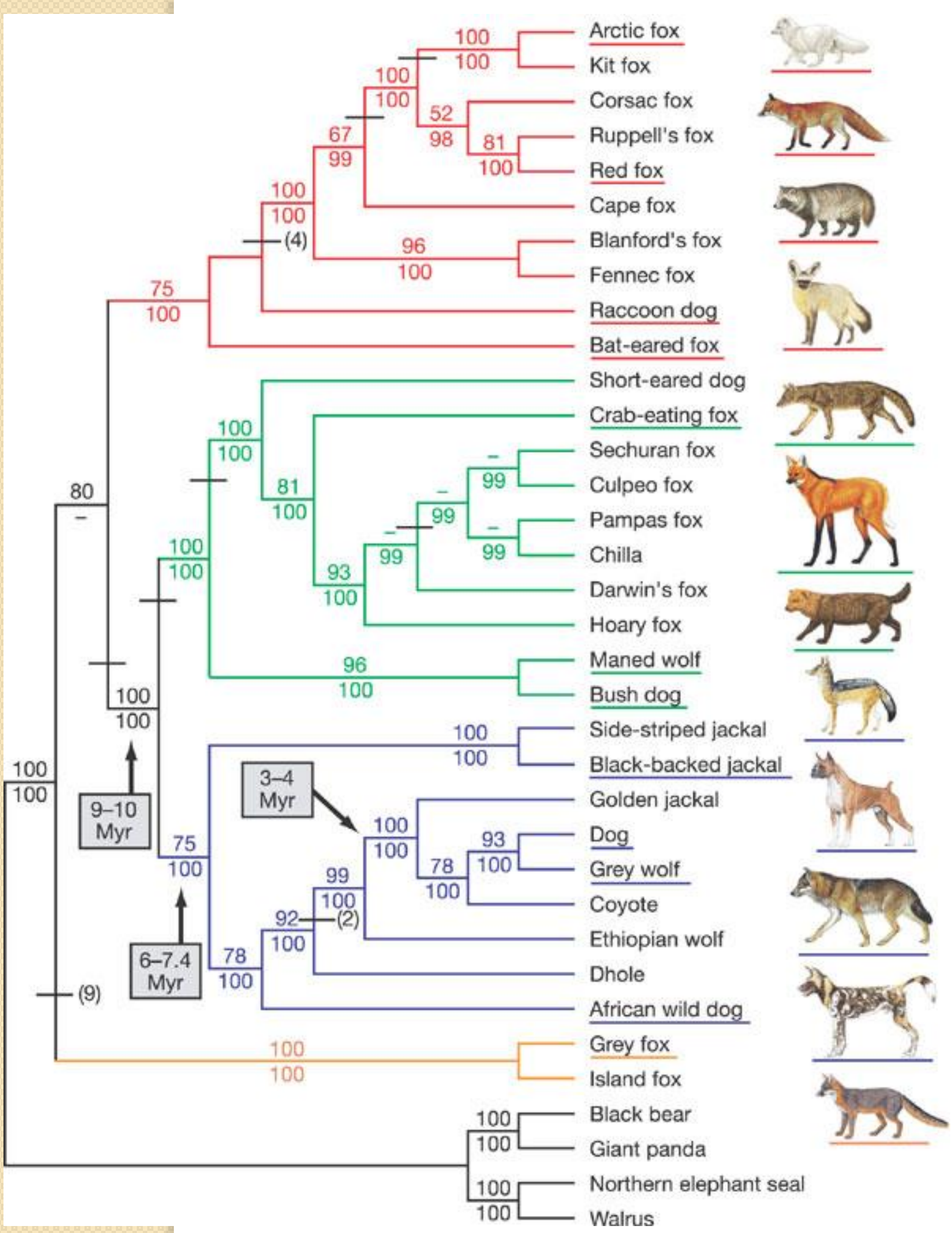
Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933 (2005), Fig. 1.

Tree nomenclature: clades

Clade ABF (monophyletic group)

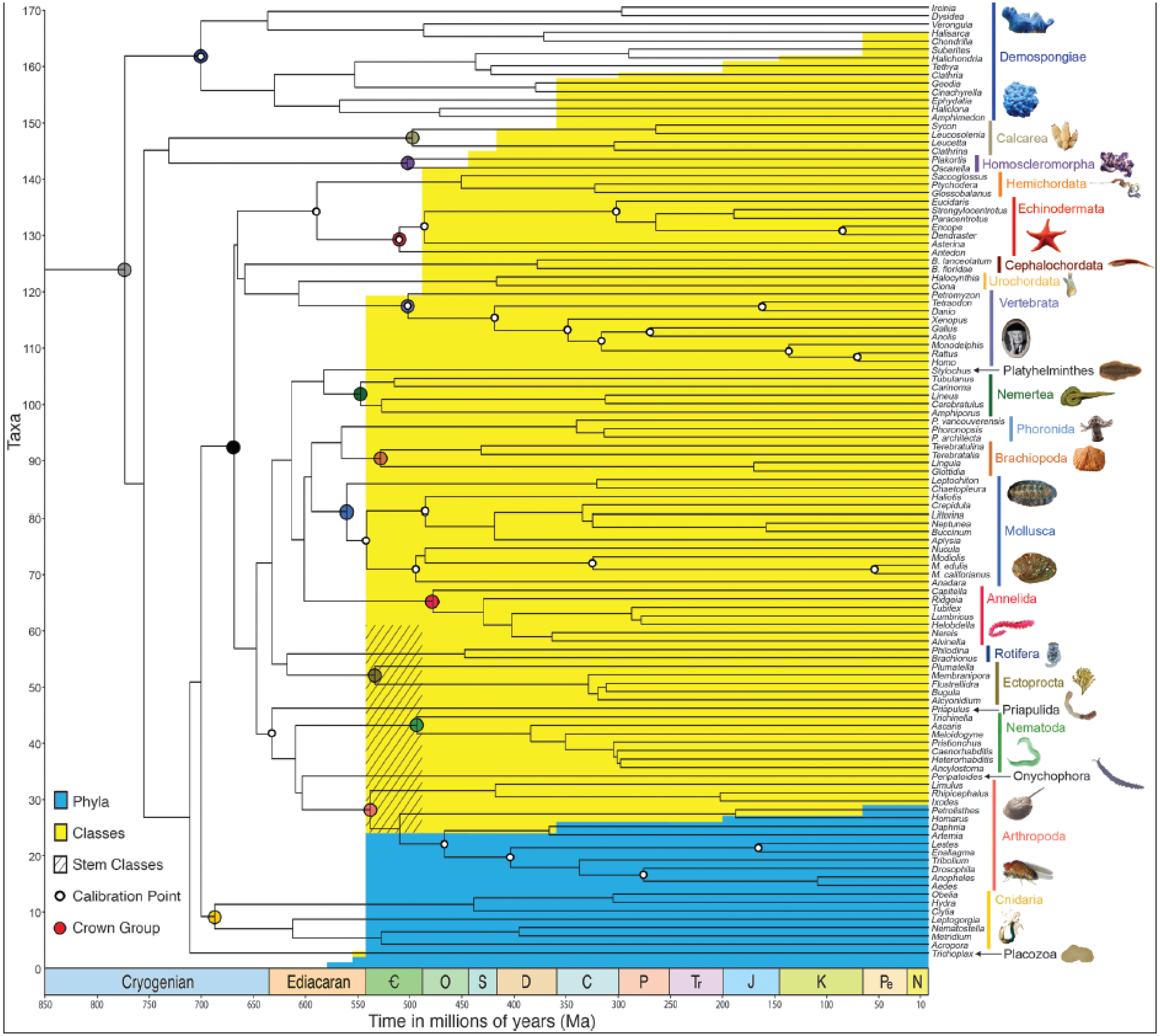


Examples of clades



Lindblad-Toh et al., *Nature* 438: 803 (2005), fig. 10

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



Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

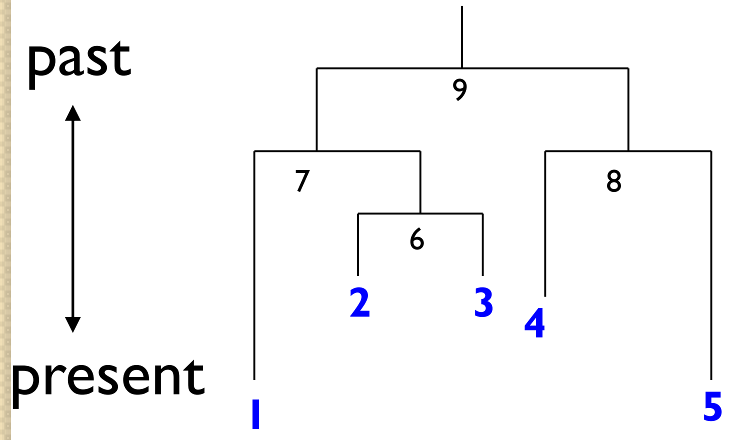
Perspective

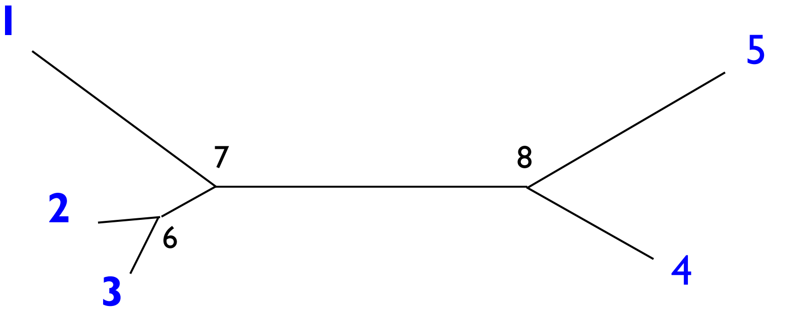
Tree roots

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

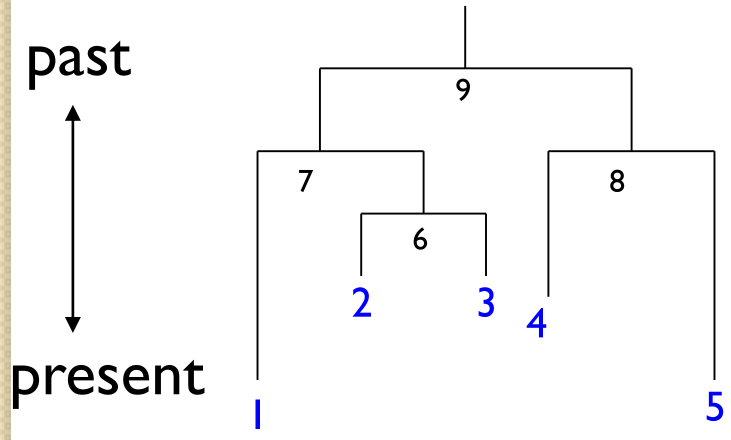




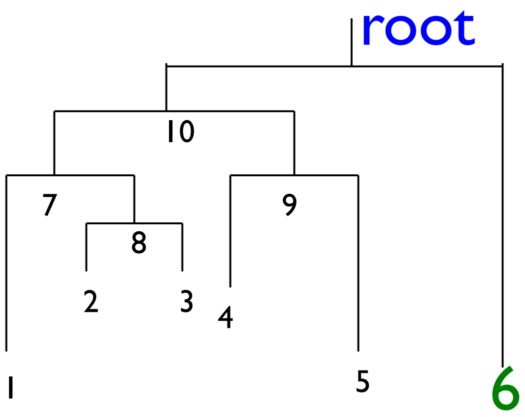
Rooted tree (specifies evolutionary path)

Unrooted tree

Tree nomenclature: outgroup rooting



Rooted tree



Outgroup
(used to place the root)

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

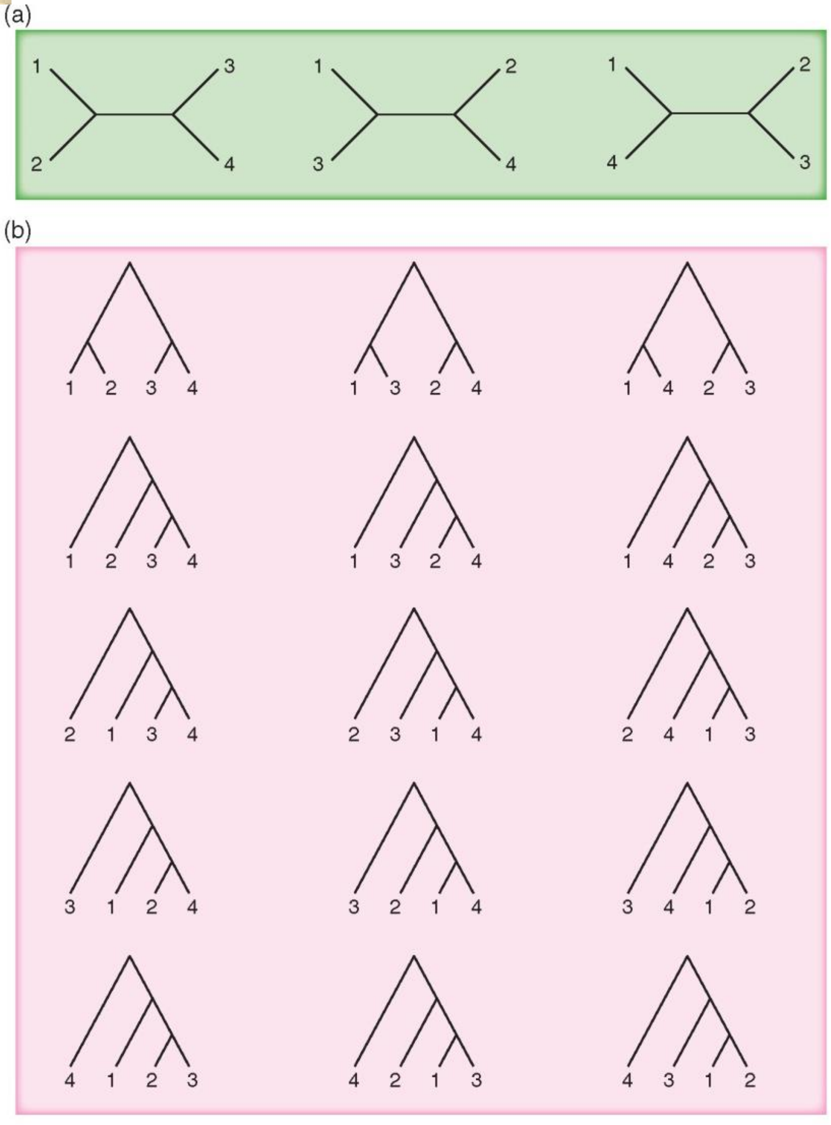
- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

Numbers of trees

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
10	34,459,425	105
20	8×10^{21}	2×10^{20}

Numbers of rooted and unrooted trees: 4 OTUs



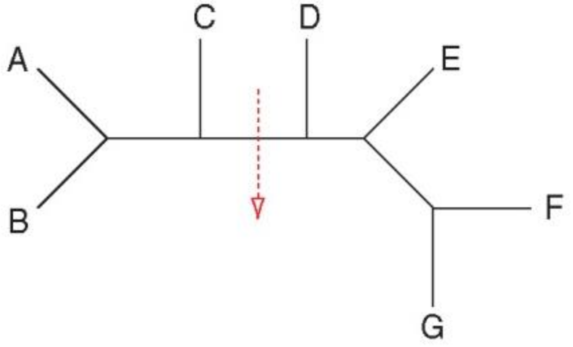
For 4 OTUs there are three possible unrooted trees.

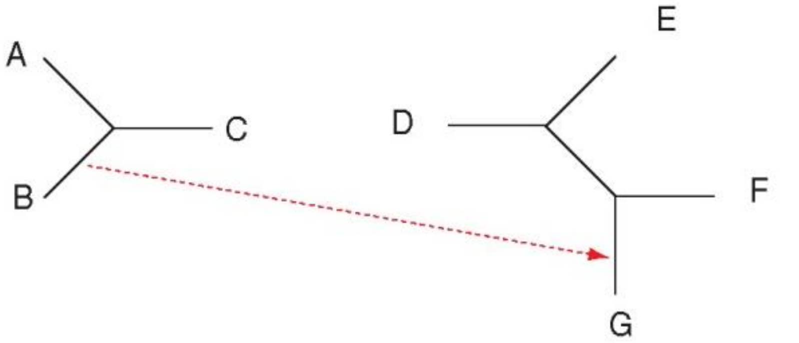
For 4 OTUs there are 15 possible rooted trees.

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

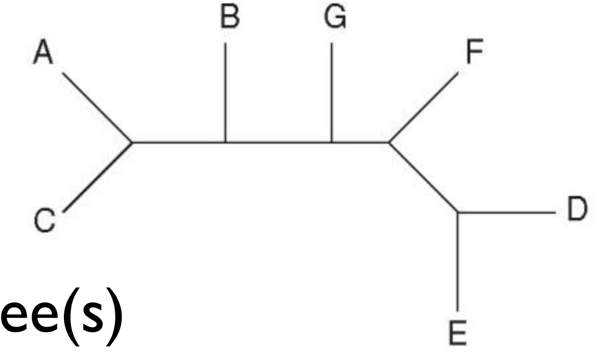
Finding optimal trees: branch swapping

Bisect a branch to form two subtrees





Reconnect via one branch from each subtree; evaluate each bisection



Identify the optimal tree(s)

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

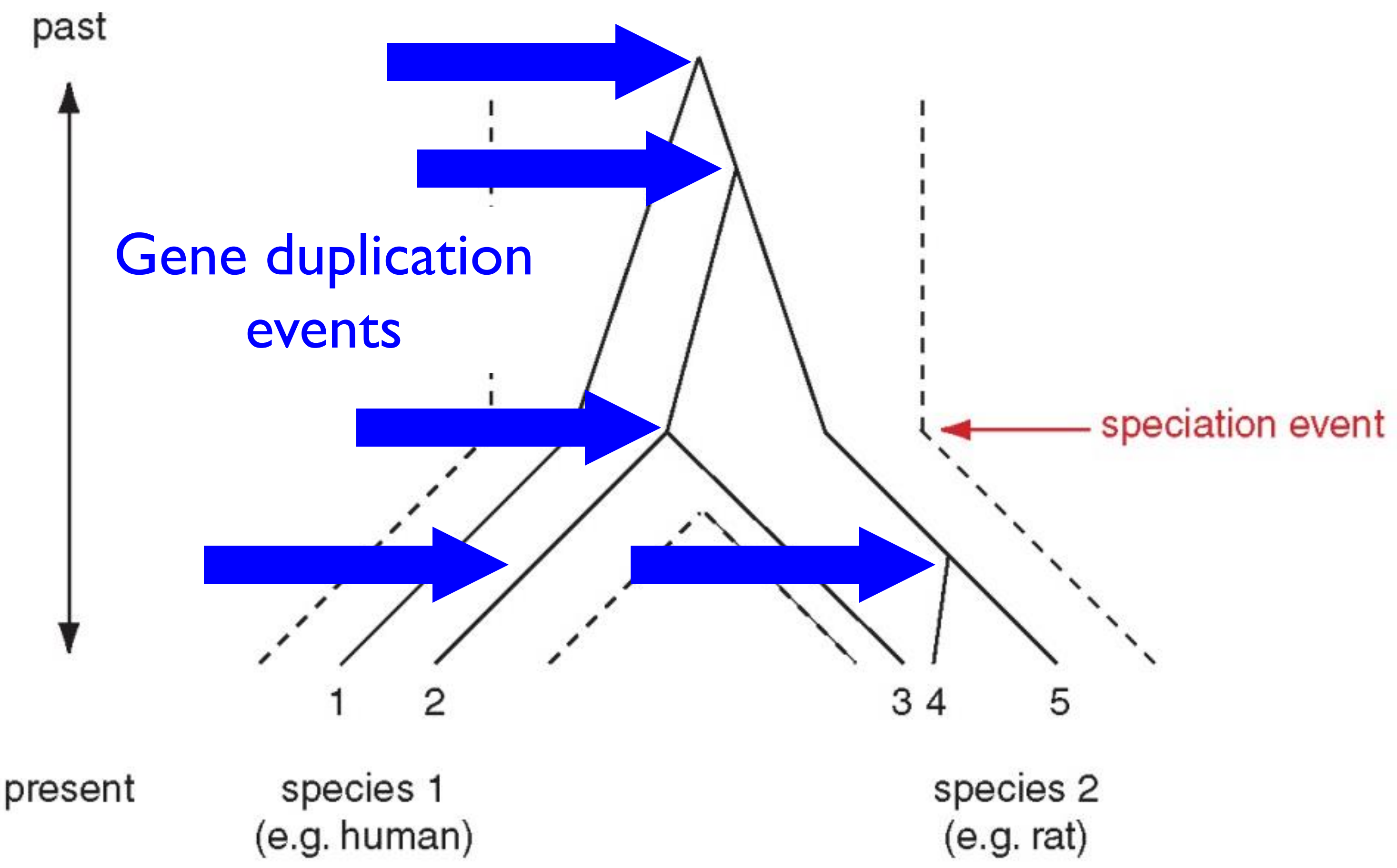
Perspective

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.

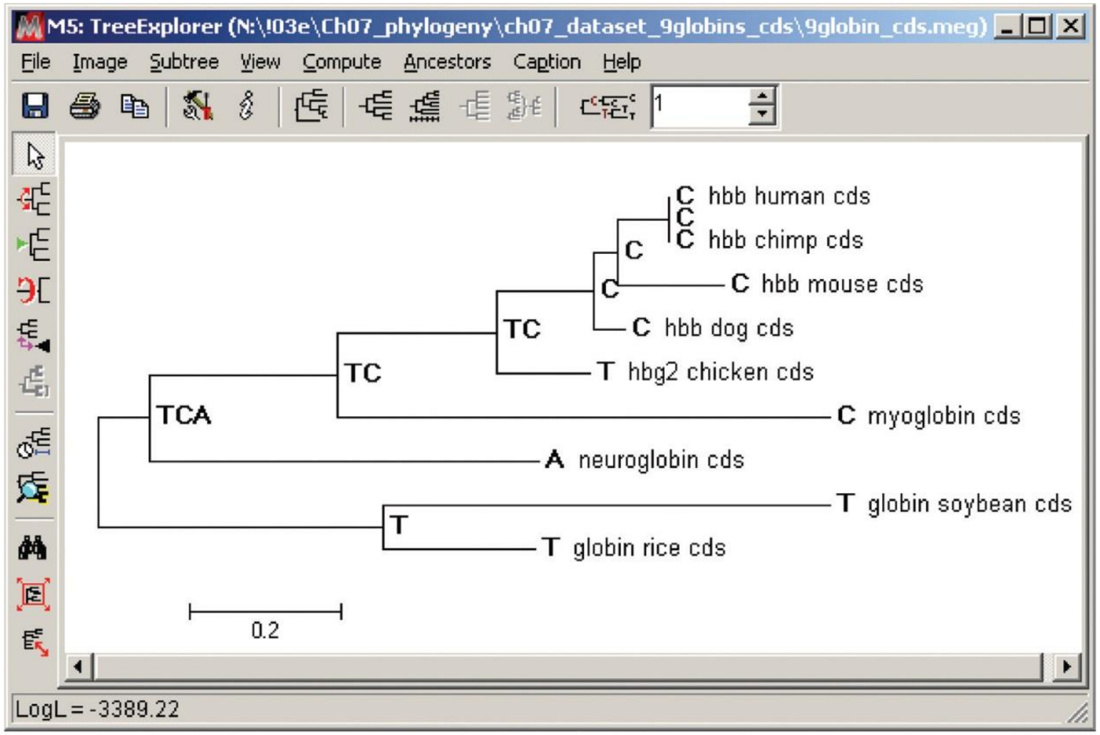
Genes (and proteins) may duplicate or otherwise evolve before or after any given speciation event. The topology of a gene (or protein) based tree may differ from the topology of a species tree.

Species trees versus gene/protein trees



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

Species trees versus gene/protein trees: we can infer ancestral sequences!



Reconstruction of ancestral sequences using MEGA (ancestors tab following creation of a maximum likelihood tree of nine globin sequences).

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

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).

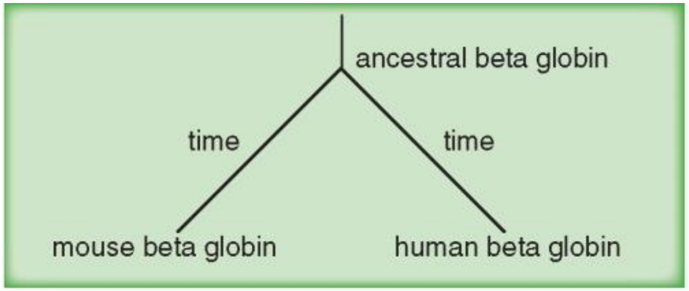
If $d_S < d_N$, positive selection occurs. For example, a duplicated gene may evolve rapidly to assume new functions.

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

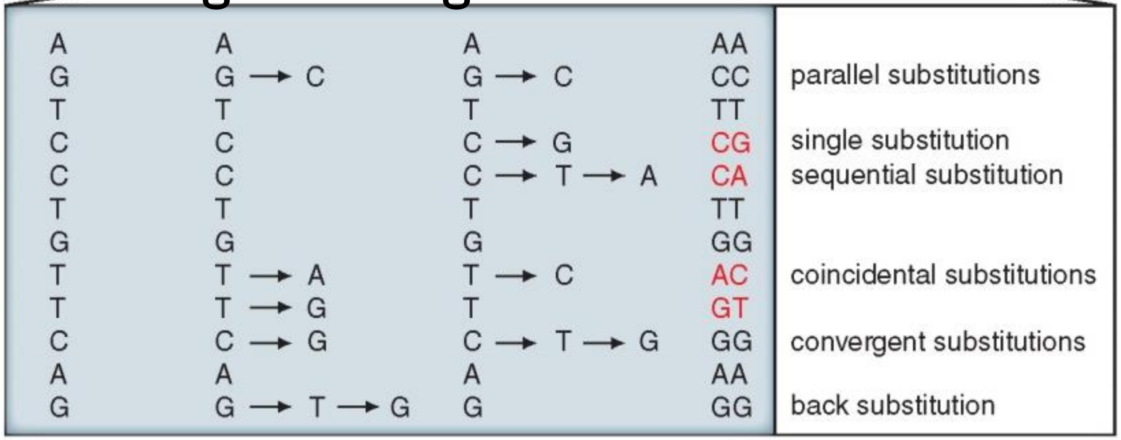


```
ancestral      M   V   H   L   S   P   V   E   K   S   A   V
human          M   V   H   L   T   P   E   E   K   S   A   V
mouse          M   V   H   L   T   D   A   E   K   S   A   V

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′
```

ancestral globin

human globin

mouse globin



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

parallel substitutions

single

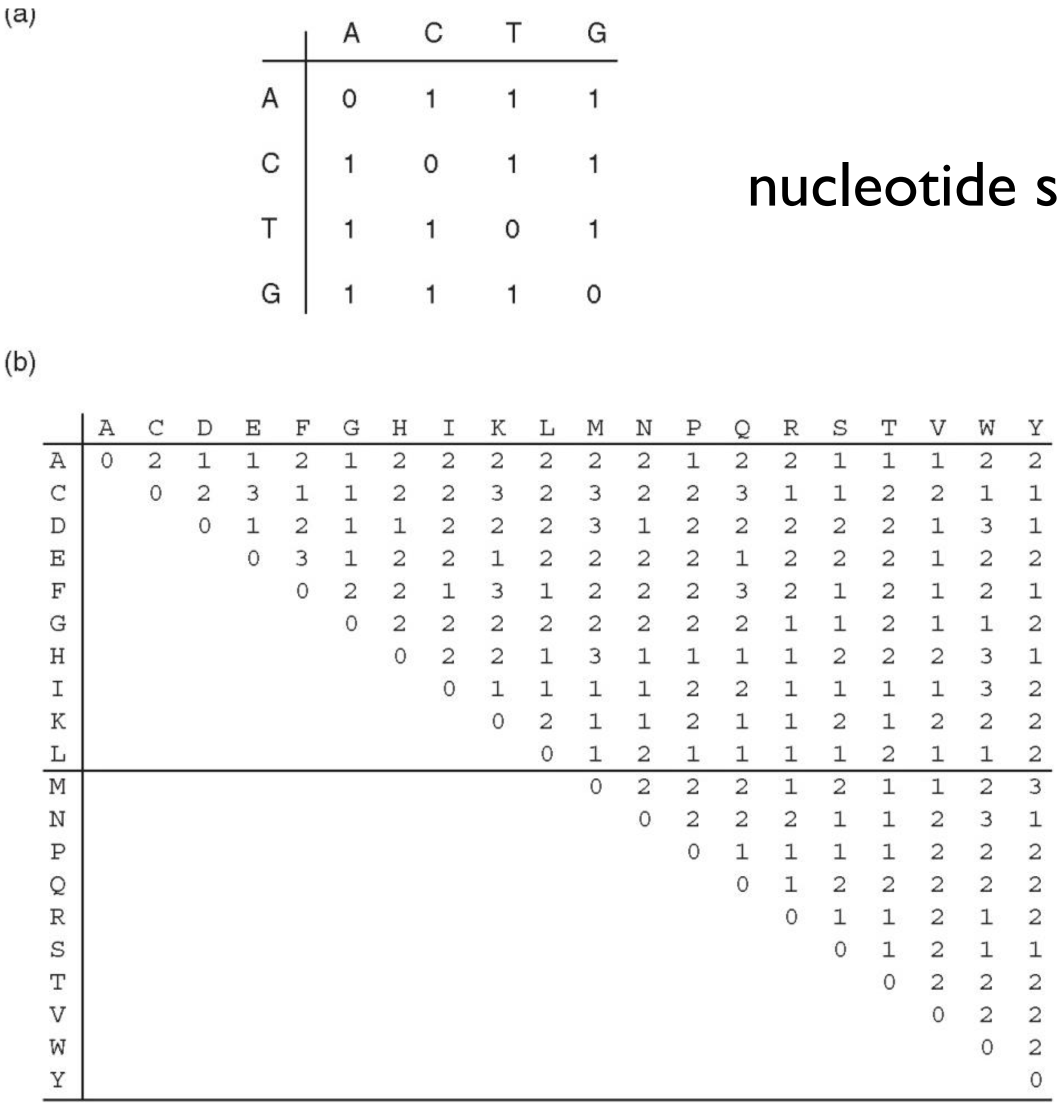
sequential

coincidental

convergent

back substitution

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

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

nucleotide step matrix

amino acid
step matrix

For amino acids, between 1 and 3 nucleotide changes are required to change one residue to another.

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

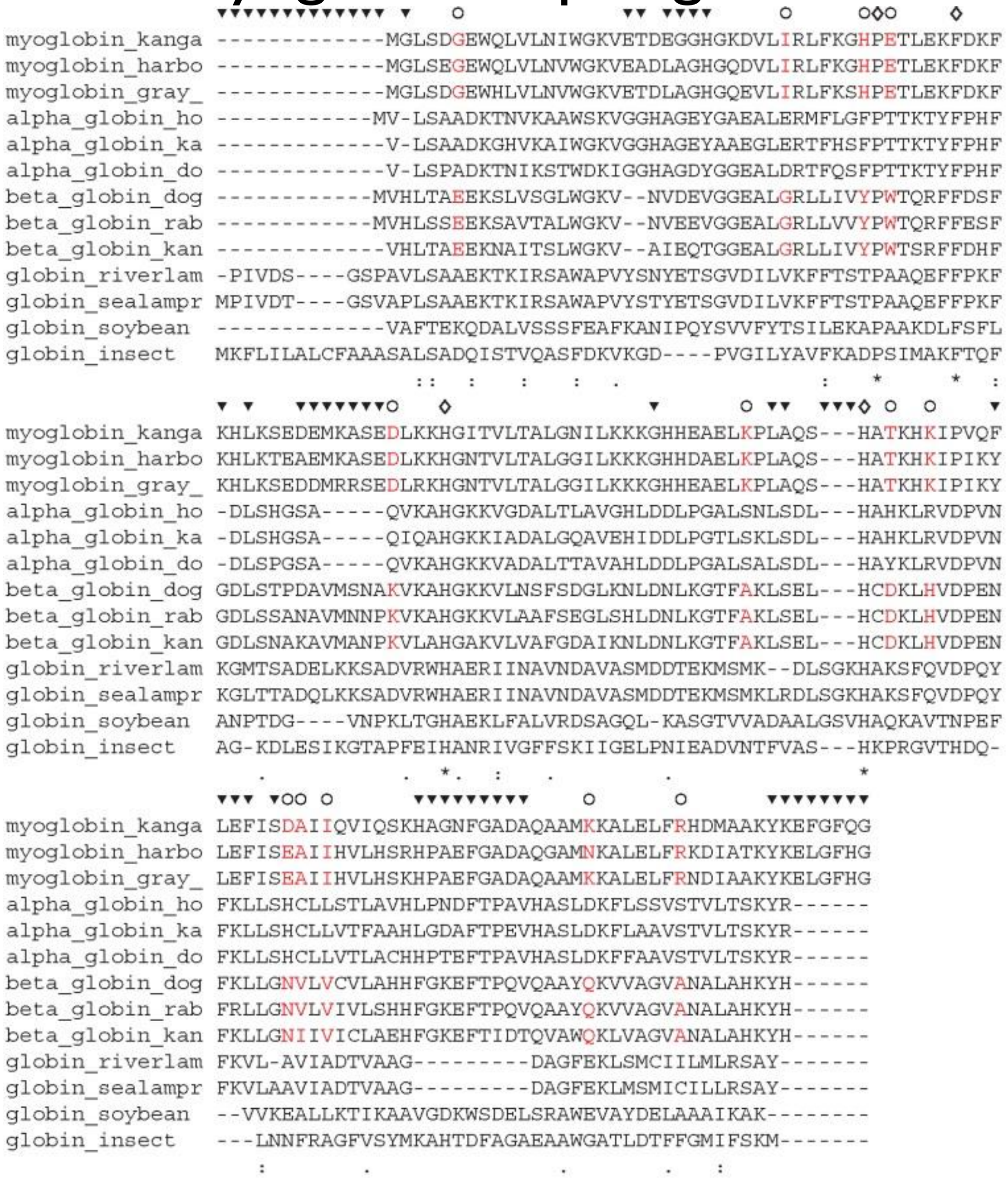
Stage 2: Multiple sequence alignment

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

Multiple alignment of myoglobins, alpha globins, beta globins

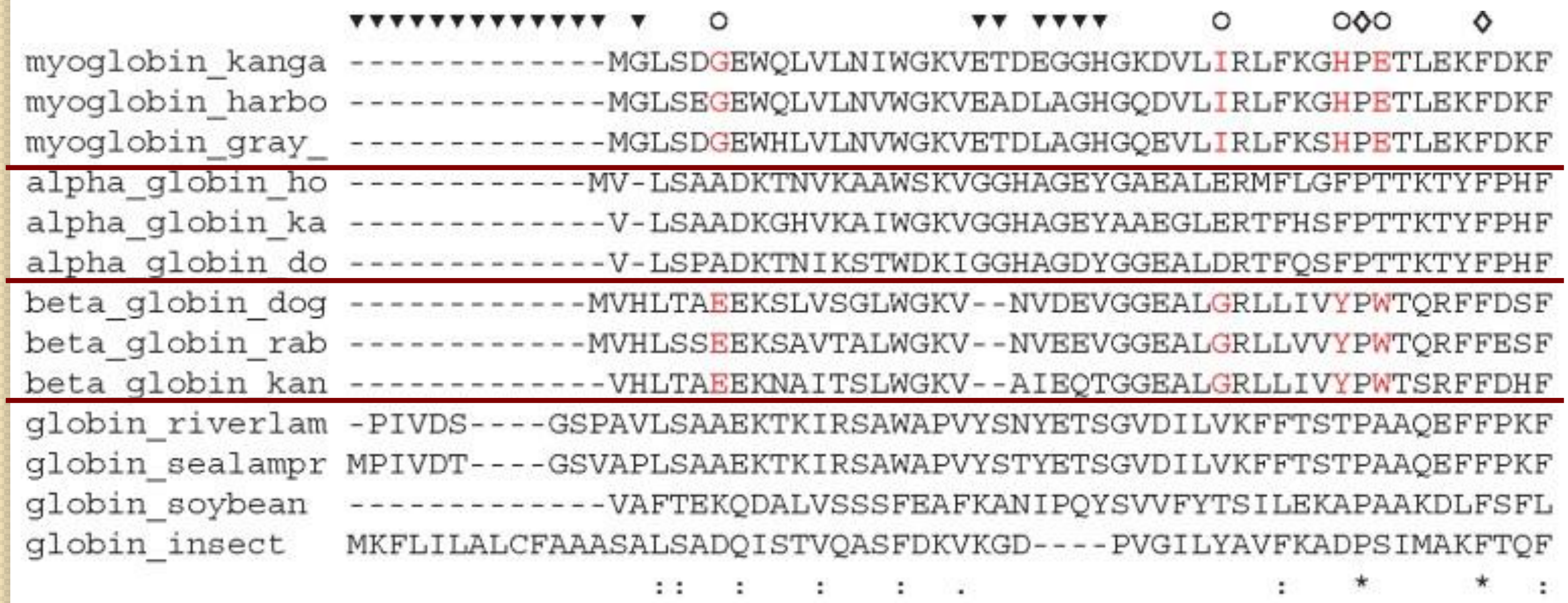


```
                ▼▼▼▼▼▼▼▼▼▼▼▼▼ ▼   ○            ▼▼ ▼▼▼▼     ○     ○◇○    ◇
myoglobin_kanga -------------MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF
myoglobin_harbo -------------MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF
myoglobin_gray_ -------------MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF
alpha_globin_ho ------------MV-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
alpha_globin_ka -------------V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHF
alpha_globin_do -------------V-LSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHF
beta_globin_dog ------------MVHLTAEEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF
beta_globin_rab ------------MVHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF
beta_globin_kan -------------VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF
globin_riverlam -PIVDS----GSPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFFPKF
globin_sealampr MPIVDT----GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKF
globin_soybean  -------------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFL
globin_insect   MKFLILALCFAAASALSADQISTVQASFDKVKGD----PVGILYAVFKADPSIMAKFTQF
                               ::  :   :   :  .               :   *     *  :

                ▼ ▼  ▼▼▼▼▼▼▼▼○   ◇               ▼      ○ ▼▼  ▼▼▼◇ ○  ○    ▼
myoglobin_kanga KHLKSEDEMKASEDLKKHGITVLTALGNILKKKGHHEAELKPLAQS---HATKHKIPVQF
myoglobin_harbo KHLKTEAEMKASEDLKKHGNTVLTALGGILKKKGHHDAELKPLAQS---HATKHKIPIKY
myoglobin_gray_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQS---HATKHKIPIKY
alpha_globin_ho -DLSHGSA-----QVKAHGKKVGDALTLAVGHLDDLPGALSNLSDL---HAHKLRVDPVN
alpha_globin_ka -DLSHGSA-----QIQAHGKKIADALGQAVEHIDDLPGTLSKLSDL---HAHKLRVDPVN
alpha_globin_do -DLSPGSA-----QVKAHGKKVADALTTAVAHLDDLPGALSALSDL---HAYKLRVDPVN
beta_globin_dog GDLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_rab GDLSSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_kan GDLSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL---HCDKLHVDPEN
globin_riverlam KGMTSADELKKSADVRWHAERIINAVNDAVASMDDTEKMSMK--DLSGKHAKSFQVDPQY
globin_sealampr KGLTTADQLKKSADVRWHAERIINAVNDAVASMDDTEKMSMKLRDLSGKHAKSFQVDPQY
globin_soybean  ANPTDG----VNPKLTGHAEKLFALVRDSAGQL-KASGTVVADAALGSVHAQKAVTNPEF
globin_insect   AG-KDLESIKGTAPFEIHANRIVGFFSKIIGELPNIEADVNTFVAS---HKPRGVTHDQ-
                   .          .  *.  :   .        .              *

                ▼▼▼ ▼○○ ○     ▼▼▼▼▼▼▼▼▼▼    ○      ○     ▼▼▼▼▼▼▼▼▼
myoglobin_kanga LEFISDAIIQVIQSKHAGNFGADAQAAMKKALELFRHDMAAKYKEFGFQG
myoglobin_harbo LEFISEAIIHVLHSRHPAEFGADAQGAMNKALELFRKDIATKYKELGFHG
myoglobin_gray_ LEFISEAIIHVLHSKHPAEFGADAQAAMKKALELFRNDIAAKYKELGFHG
alpha_globin_ho FKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
alpha_globin_ka FKLLSHCLLVTFAAHLGDAFTPEVHASLDKFLAAVSTVLTSKYR------
alpha_globin_do FKLLSHCLLVTLACHHPTEFTPAVHASLDKFFAAVSTVLTSKYR------
beta_globin_dog FKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH------
beta_globin_rab FRLLGNVLVIVLSHHFGKEFTPQVQAAYQKVVAGVANALAHKYH------
beta_globin_kan FKLLGNIIVICLAEHFGKEFTIDTQVAWQKLVAGVANALAHKYH------
globin_riverlam FKVL-AVIADTVAAG---------DAGFEKLSMCIILMLRSAY-------
globin_sealampr FKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
globin_soybean  --VVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAAIKAK--------
globin_insect   ---LNNFRAGFVSYMKAHTDFAGAEAAWGATLDTFFGMIFSKM-------
                   :       .              .       .   :
```

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

▼ gaps

◇ 100% conserved



```
                 ▼▼▼▼▼▼▼▼▼▼▼▼▼ ▼   ○            ▼▼ ▼▼▼▼     ○     ○◇○    ◇
myoglobin_kanga  -------------MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF
myoglobin_harbo  -------------MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF
myoglobin_gray_  -------------MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF
alpha_globin_ho  ------------MV-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
alpha_globin_ka  -------------V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHF
alpha_globin_do  -------------V-LSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHF
beta_globin_dog  ------------MVHLTAEEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF
beta_globin_rab  ------------MVHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF
beta_globin_kan  -------------VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF
globin_riverlam  -PIVDS----GSPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFFPKF
globin_sealampr  MPIVDT----GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKF
globin_soybean   -------------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFL
globin_insect    MKFLILALCFAAASALSADQISTVQASFDKVKGD----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).

Outline

- Introduction to molecular evolution
- Principles of molecular phylogeny and evolution
  - Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution
- Molecular phylogeny: properties of trees
  - Topologies and branch lengths of trees
  - Tree roots
  - Enumerating trees and selecting search strategies
- Type of trees (species trees vs. gene/protein trees; DNA or protein)
- Five stages of phylogenetic analysis
  - Stage 1: sequence acquisition
  - Stage 2: multiple sequence alignment
  - Stage 3: models of DNA and amino acid substitution
  - Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
  - Stage 5: evaluating trees
- Perspective

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

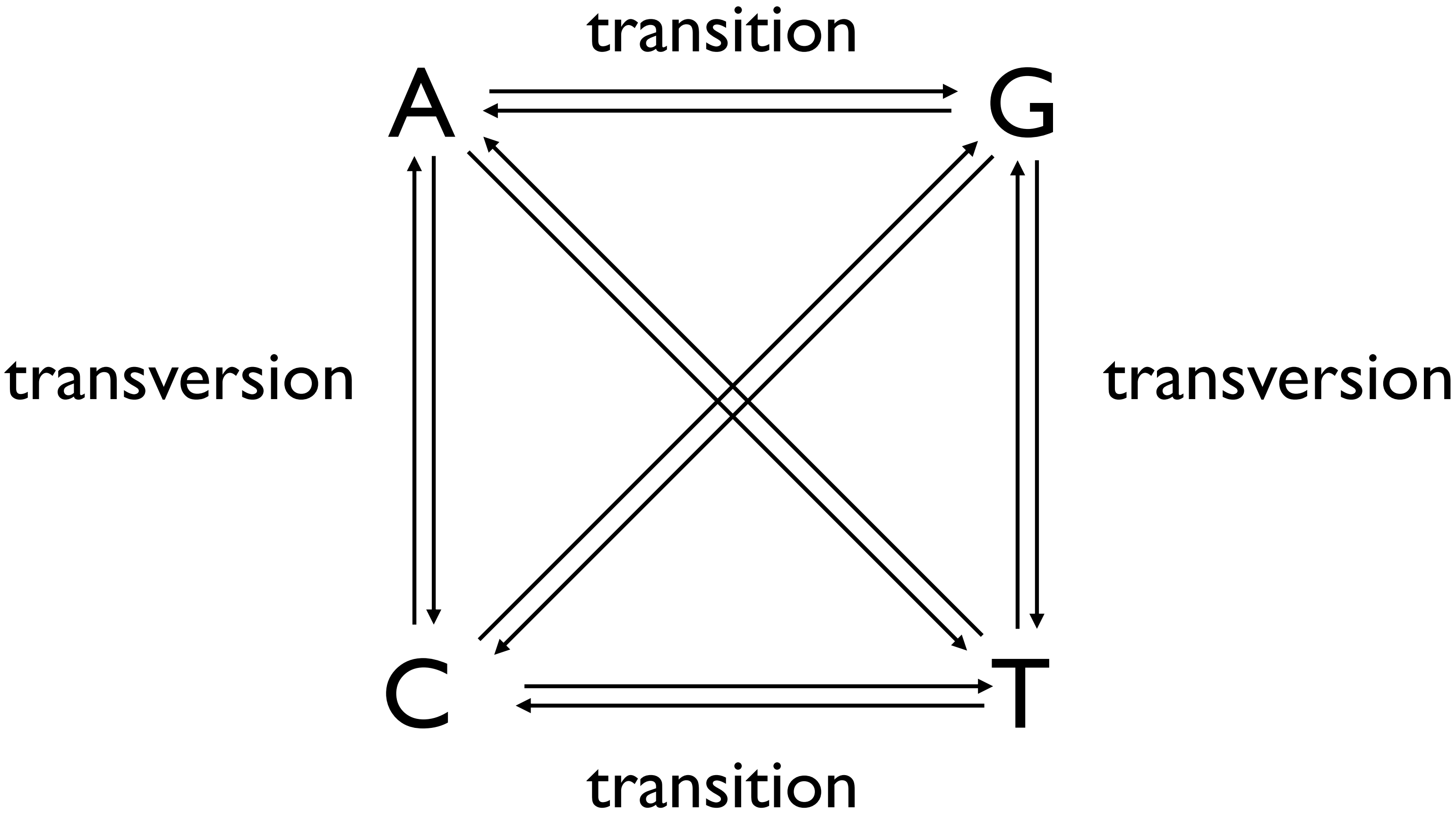
Stage 3: Models of substitution

Jukes and Cantor (1969) proposed a corrective formula:

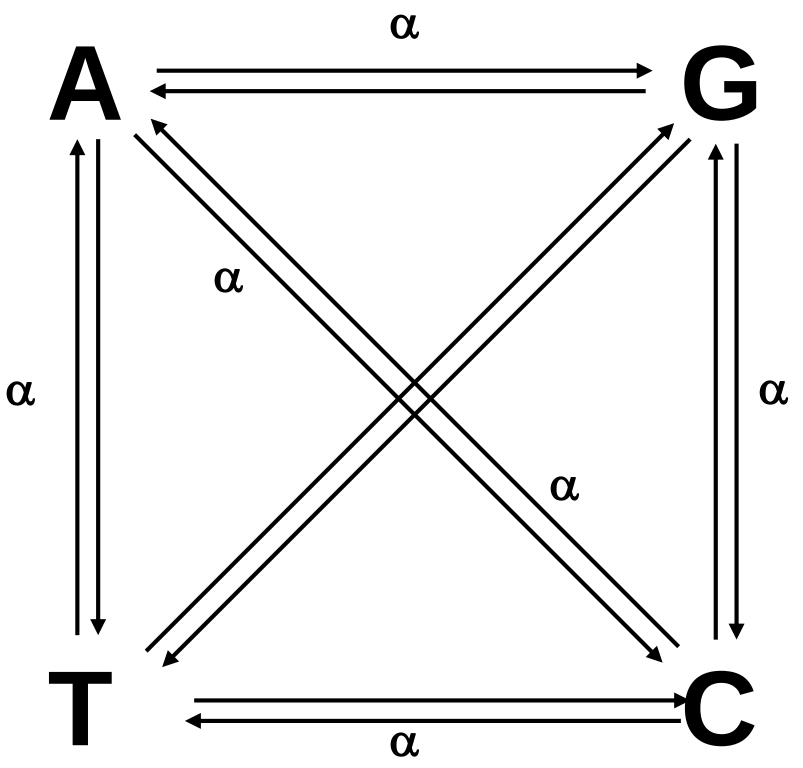
$$\boldsymbol{D = (-\frac{3}{4}) \ln (1 - \frac{4}{3} p)}$$

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 is typically greater than the transversion rate.

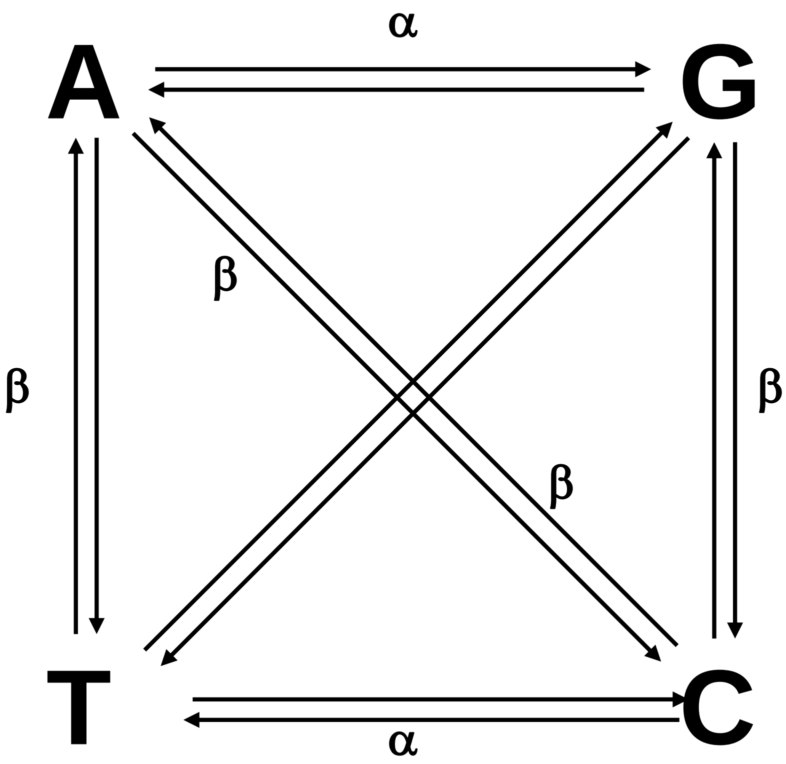
There are dozens of models of nucleotide substitution



Jukes and Cantor one-parameter model of nucleotide substitution (α=β)



Kimura two-parameter model of nucleotide substitution (assumes a ≠ b)



Stage 4: Tree-building methods: distance

Jukes and Cantor (1969) proposed a corrective formula:

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

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

Stage 4: Tree-building methods

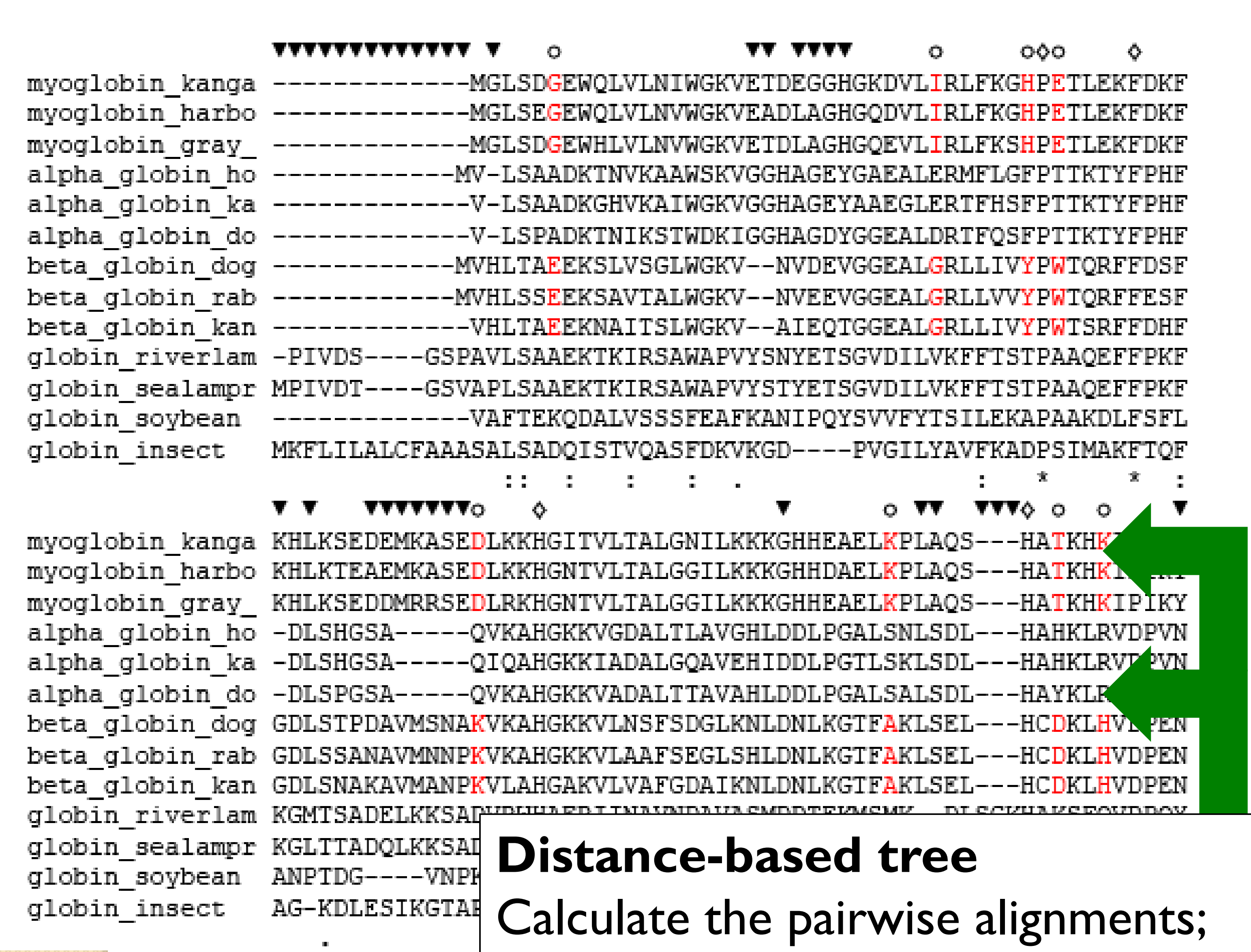
We will discuss several tree-building methods:

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.



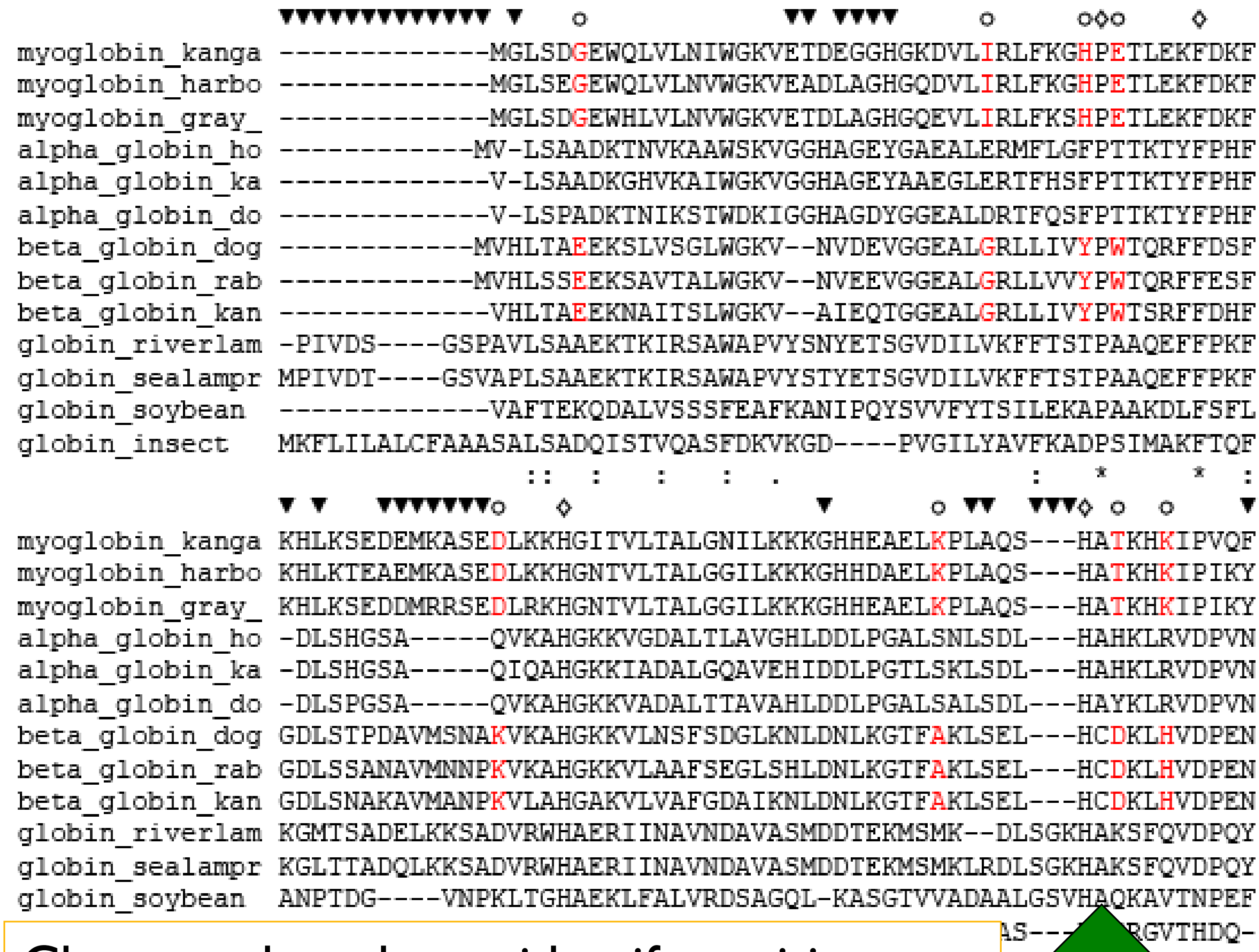
```
                ▼▼▼▼▼▼▼▼▼▼▼▼▼ ▼   ○            ▼▼ ▼▼▼▼     ○     ○◊○    ◊
myoglobin_kanga -------------MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF
myoglobin_harbo -------------MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF
myoglobin_gray_ -------------MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF
alpha_globin_ho ------------MV-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
alpha_globin_ka -------------V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHF
alpha_globin_do -------------V-LSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHF
beta_globin_dog ------------MVHLTAEEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF
beta_globin_rab ------------MVHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF
beta_globin_kan -------------VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF
globin_riverlam -PIVDS----GSPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFFPKF
globin_sealampr MPIVDT----GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKF
globin_soybean  -------------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFL
globin_insect   MKFLILALCFAAASALSADQISTVQASFDKVKGD----PVGILYAVFKADPSIMAKFTQF
                               :: :  :   :  .               :   *     *  :

                ▼ ▼   ▼▼▼▼▼▼▼▼○   ◊               ▼      ○ ▼▼  ▼▼▼◊ ○  ○    ▼
myoglobin_kanga KHLKSEDEMKASEDLKKHGITVLTALGNILKKKGHHEAELKPLAQS---HATKHK
myoglobin_harbo KHLKTEAEMKASEDLKKHGNTVLTALGGILKKKGHHDAELKPLAQS---HATKHKI
myoglobin_gray_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQS---HATKHKIPIKY
alpha_globin_ho -DLSHGSA-----QVKAHGKKVGDALTLAVGHLDDLPGALSNLSDL---HAHKLRVDPVN
alpha_globin_ka -DLSHGSA-----QIQAHGKKIADALGQAVEHIDDLPGTLSKLSDL---HAHKLRVDPVN
alpha_globin_do -DLSPGSA-----QVKAHGKKVADALTTAVAHLDDLPGALSALSDL---HAYKLR
beta_globin_dog GDLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_rab GDLSSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_kan GDLSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL---HCDKLHVDPEN
globin_riverlam KGMTSADELKKSADVRWHAERIINAVNDAVASMDDTEKMSMK--DLSGKHAKSFQVDPQY
globin_sealampr KGLTTADQLKKSAD
globin_soybean  ANPTDG----VNPK
globin_insect   AG-KDLESIKGTAP
                   .
```

Distance-based tree

Calculate the pairwise alignments;

if two sequences are related,

put them next to each other on the tree

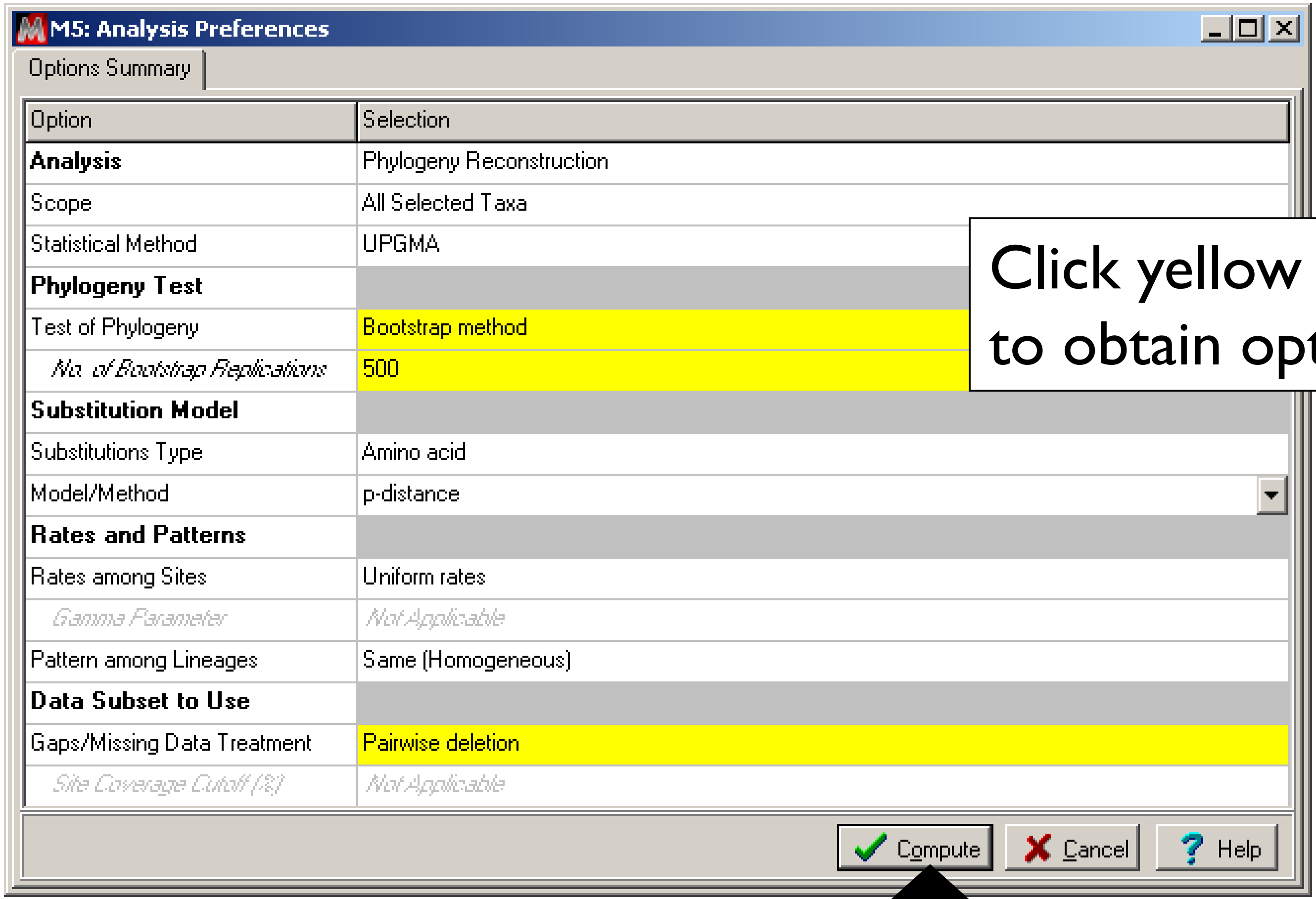


```
                ▼▼▼▼▼▼▼▼▼▼▼▼▼▼ ▼   ○            ▼▼ ▼▼▼▼     ○     ○◊○    ◊
myoglobin_kanga -------------MGLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKF
myoglobin_harbo -------------MGLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKF
myoglobin_gray_ -------------MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF
alpha_globin_ho ------------MV-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
alpha_globin_ka -------------V-LSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHF
alpha_globin_do -------------V-LSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHF
beta_globin_dog ------------MVHLTAEEKSLVSGLWGKV--NVDEVGGEALGRLLIVYPWTQRFFDSF
beta_globin_rab ------------MVHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESF
beta_globin_kan -------------VHLTAEEKNAITSLWGKV--AIEQTGGEALGRLLIVYPWTSRFFDHF
globin_riverlam -PIVDS----GSPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFFPKF
globin_sealampr MPIVDT----GSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKF
globin_soybean  -------------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFL
globin_insect   MKFLILALCFAAASALSADQISTVQASFDKVKGD----PVGILYAVFKADPSIMAKFTQF
                               ::  :   :   :  .               :   *     *  :
                ▼ ▼   ▼▼▼▼▼▼▼▼○   ◊               ▼      ○ ▼▼  ▼▼▼◊ ○  ○    ▼
myoglobin_kanga KHLKSEDEMKASEDLKKHGITVLTALGNILKKKGHHEAELKPLAQS---HATKHKIPVQF
myoglobin_harbo KHLKTEAEMKASEDLKKHGNTVLTALGGILKKKGHHDAELKPLAQS---HATKHKIPIKY
myoglobin_gray_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQS---HATKHKIPIKY
alpha_globin_ho -DLSHGSA-----QVKAHGKKVGDALTLAVGHLDDLPGALSNLSDL---HAHKLRVDPVN
alpha_globin_ka -DLSHGSA-----QIQAHGKKIADALGQAVEHIDDLPGTLSKLSDL---HAHKLRVDPVN
alpha_globin_do -DLSPGSA-----QVKAHGKKVADALTTAVAHLDDLPGALSALSDL---HAYKLRVDPVN
beta_globin_dog GDLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_rab GDLSSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSEL---HCDKLHVDPEN
beta_globin_kan GDLSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL---HCDKLHVDPEN
globin_riverlam KGMTSADELKKSADVRWHAERIINAVNDAVASMDDTEKMSMK--DLSGKHAKSFQVDPQY
globin_sealampr KGLTTADQLKKSADVRWHAERIINAVNDAVASMDDTEKMSMKLRDLSGKHAKSFQVDPQY
globin_soybean  ANPTDG----VNPKLTGHAEKLFALVRDSAGQL-KASGTVVADAALGSVHAQKAVTNPEF
                                                            AS---   RGVTHDQ-
```

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

Use of MEGA for a distance-based tree: UPGMA

(an easy method to explain, but not accurate for most purposes)

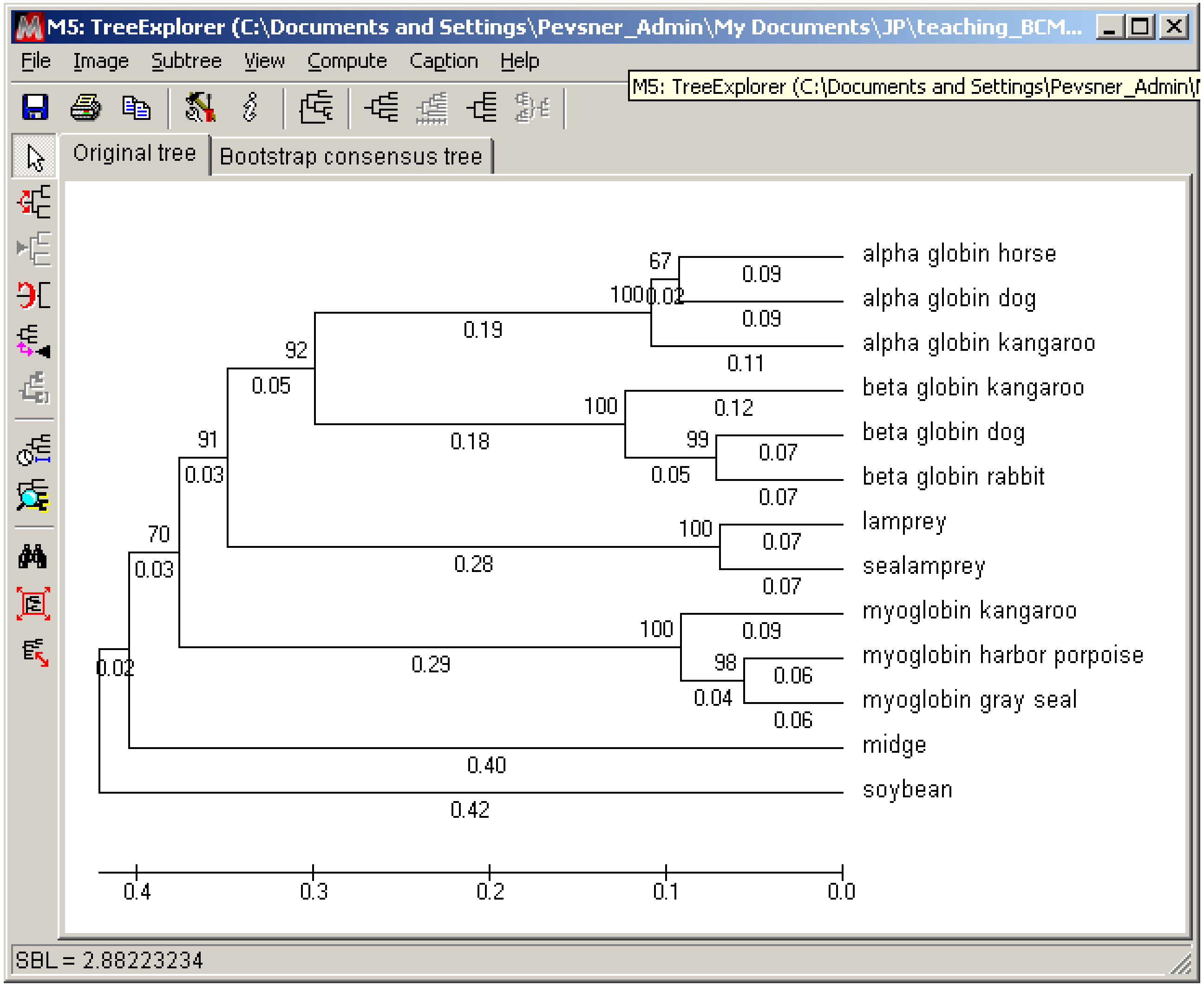


Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	UPGMA
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Amino acid
Model/Method	p-distance
Rates and Patterns	
Rates among Sites	Uniform rates
Gamma Parameter	*Not Applicable*
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Pairwise deletion
Site Coverage Cutoff (%)	*Not Applicable*

Click yellow rows to obtain options

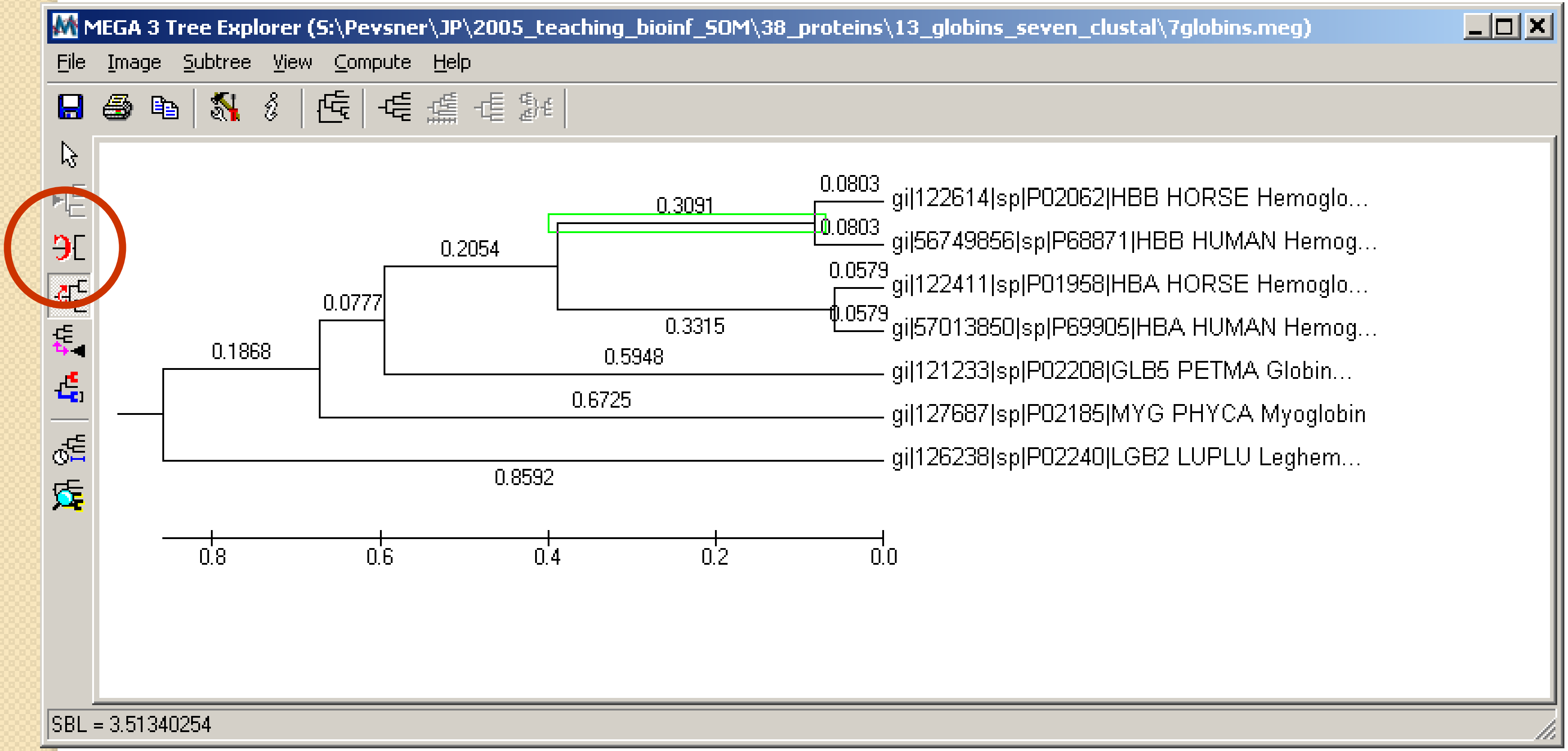
Click compute to obtain tree

Use of MEGA for a distance-based tree: UPGMA



SBL = 2.88223234

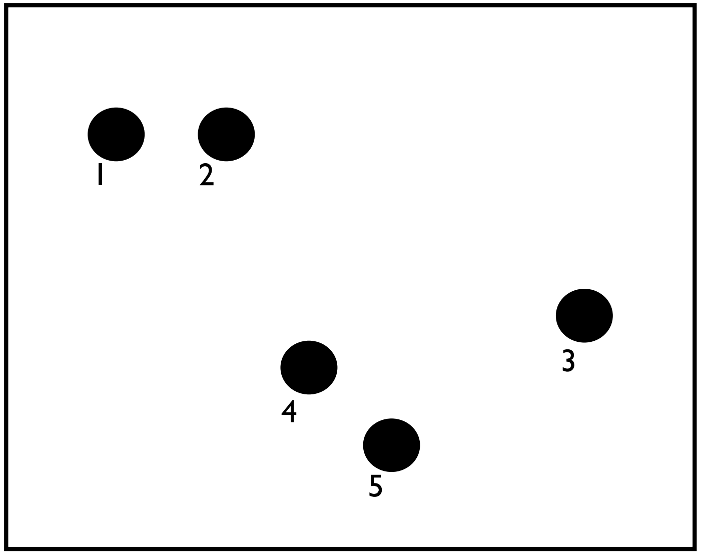
Use of MEGA for a distance-based tree: UPGMA


MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)
File Image Subtree View Compute Help
0.0803
0.3091
0.0803
0.2054
0.0579
0.0777
0.0579
0.3315
0.1868
0.5948
0.6725
0.8592
gi|122614|sp|P02062|HBB HORSE Hemoglo...
gi|56749856|sp|P68871|HBB HUMAN Hemog...
gi|122411|sp|P01958|HBA HORSE Hemoglo...
gi|57013850|sp|P69905|HBA HUMAN Hemog...
gi|121233|sp|P02208|GLB5 PETMA Globin...
gi|127687|sp|P02185|MYG PHYCA Myoglobin
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
0.8
0.6
0.4
0.2
0.0
SBL = 3.51340254

Flipping branches around a node creates an equivalent topology

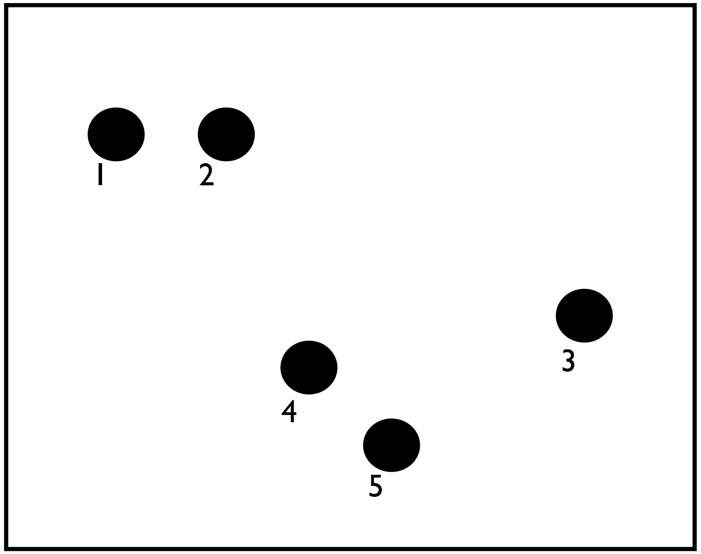
Tree-building methods: UPGMA

UPGMA is
unweighted pair group method
using arithmetic mean



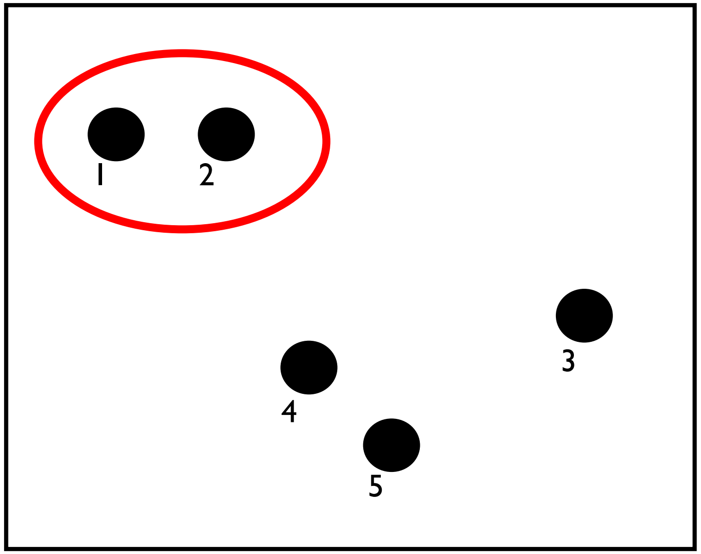
Tree-building methods: UPGMA

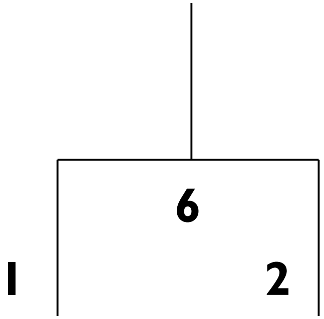
Step 1: compute the pairwise distances of all the proteins. Get ready to put the numbers 1-5 at the bottom of your new tree.



Tree-building methods: UPGMA

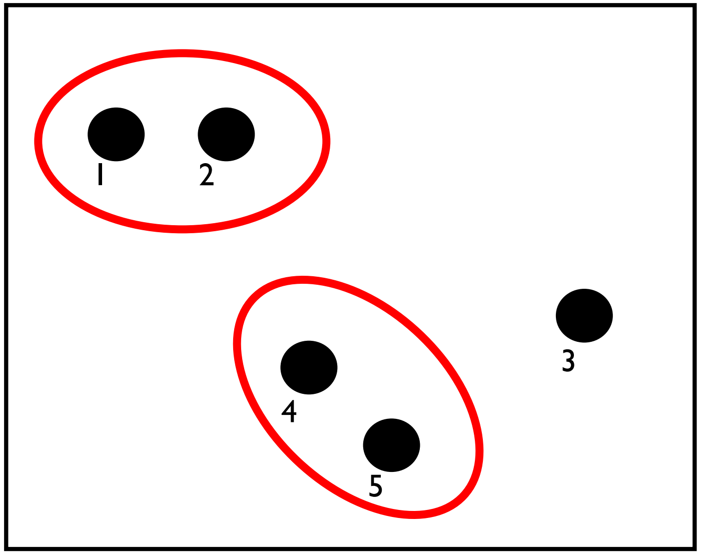
Step 2: Find the two proteins with the smallest pairwise distance. Cluster them.

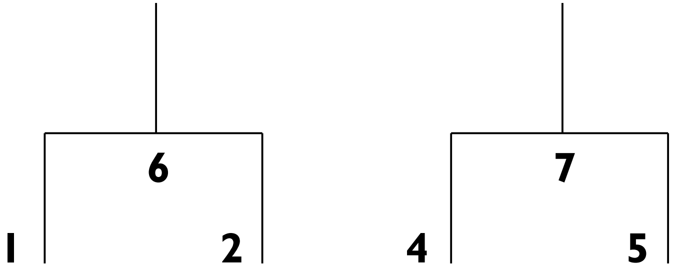




Tree-building methods: UPGMA

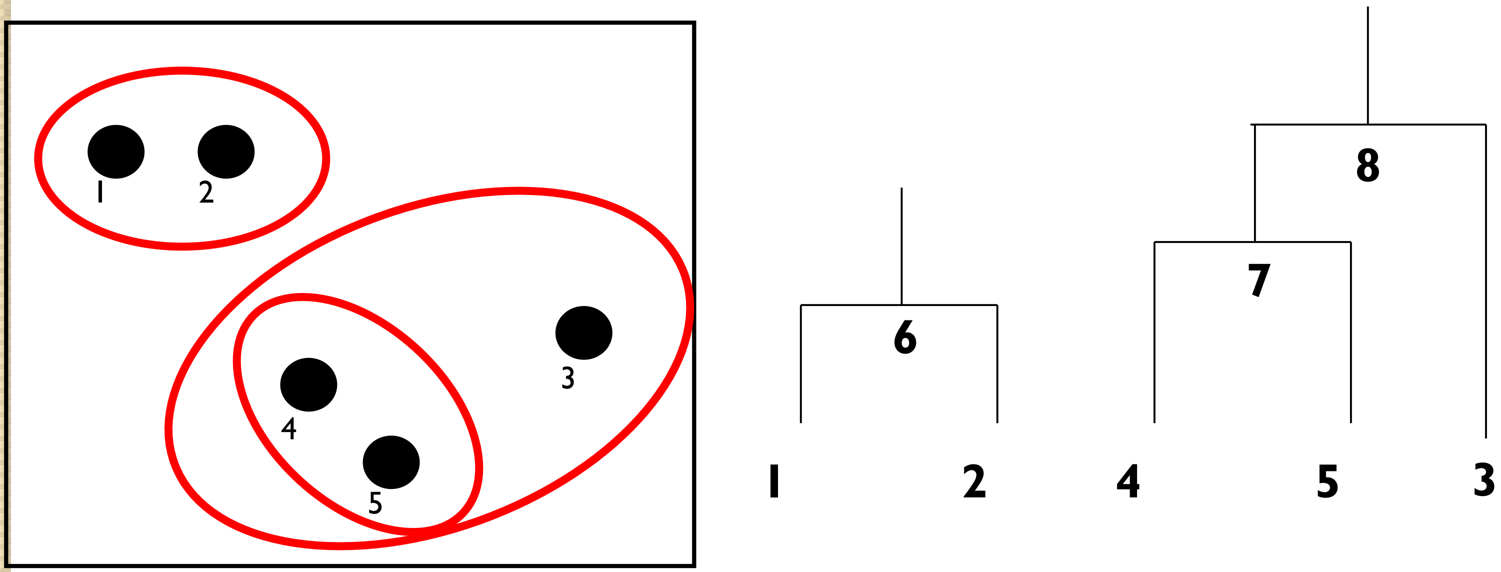
Step 3: Do it again. Find the next two proteins with the smallest pairwise distance. Cluster them.





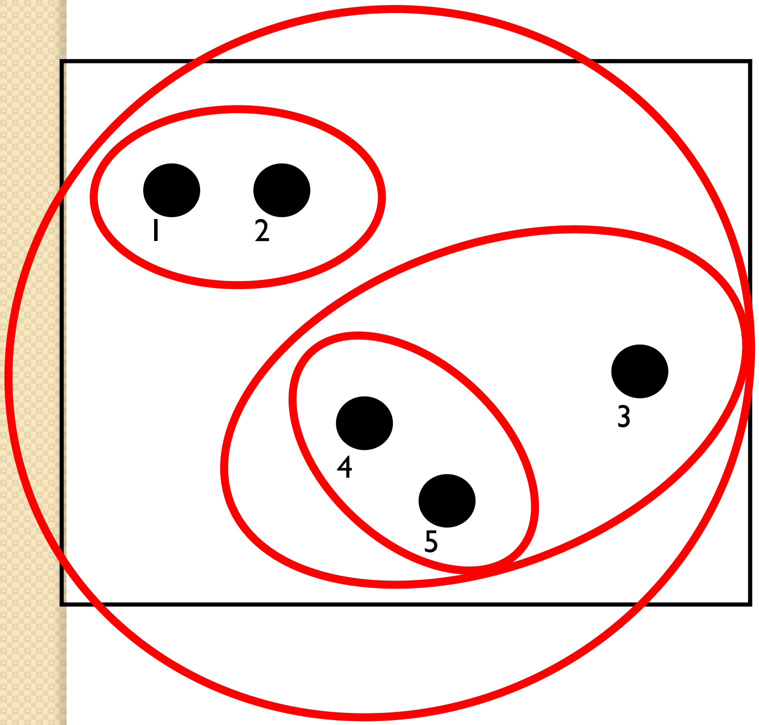
Tree-building methods: UPGMA

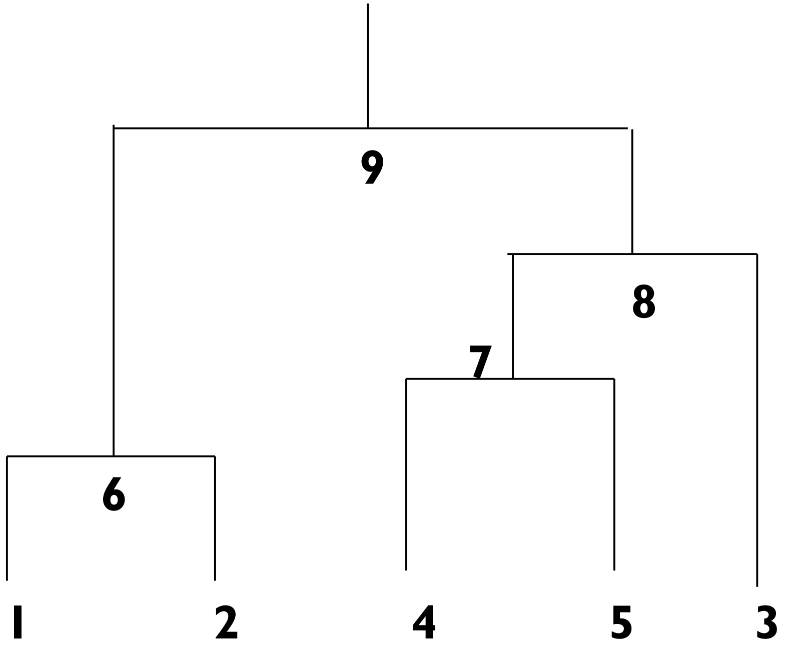
Step 4: Keep going. Cluster.



Tree-building methods: UPGMA

Step 4: Last cluster! This is your tree.





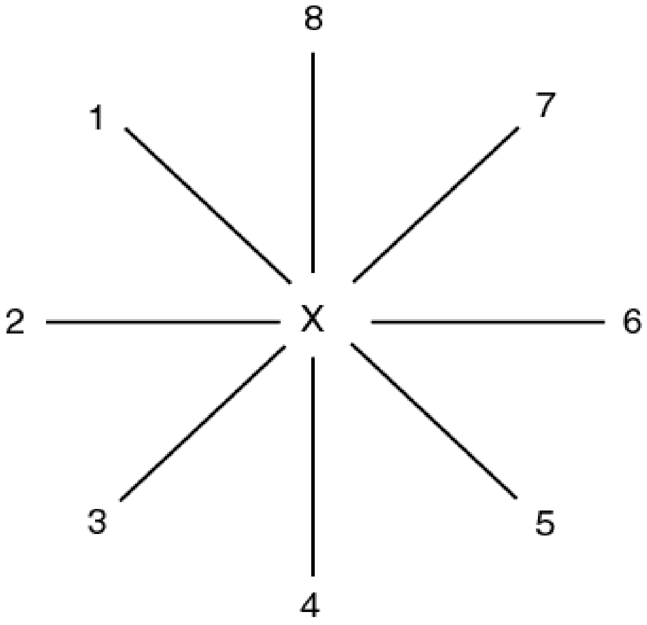
Distance-based methods: UPGMA trees

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).

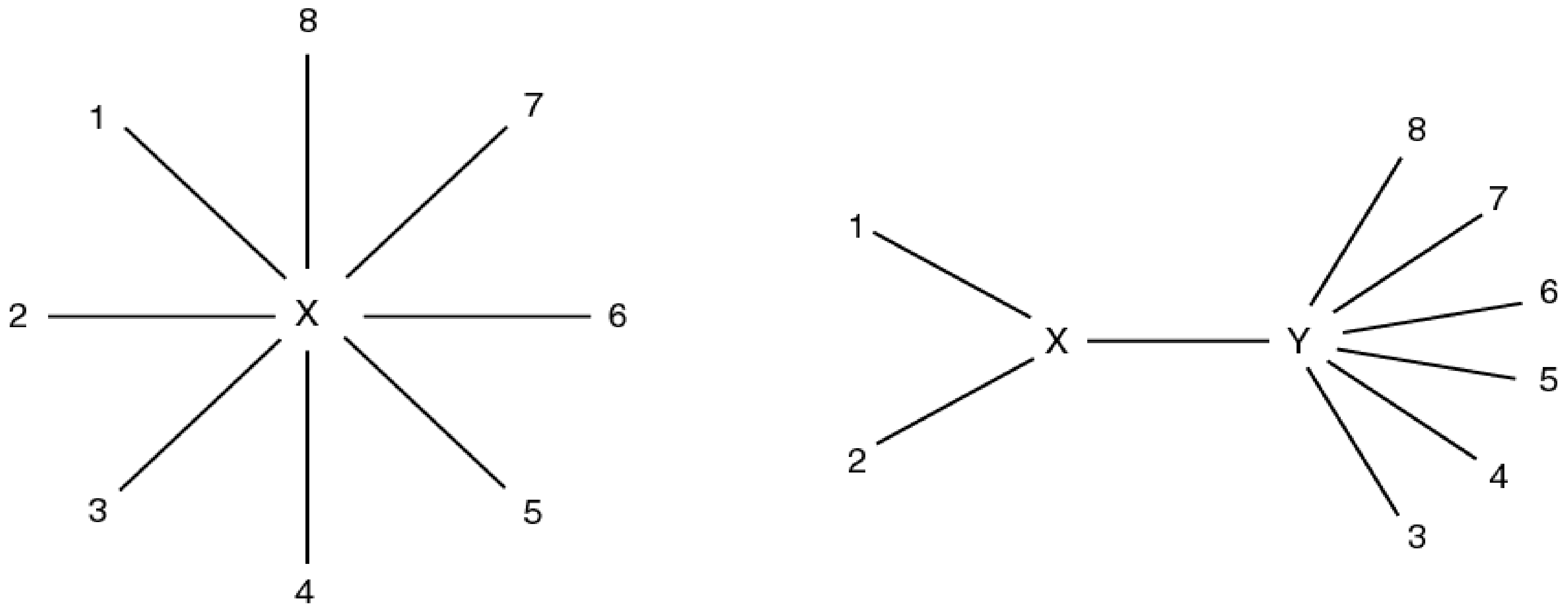
Making trees using neighbor-joining

The neighbor-joining method of Saitou and Nei (1987) Is especially useful for making a tree having a large number of taxa.



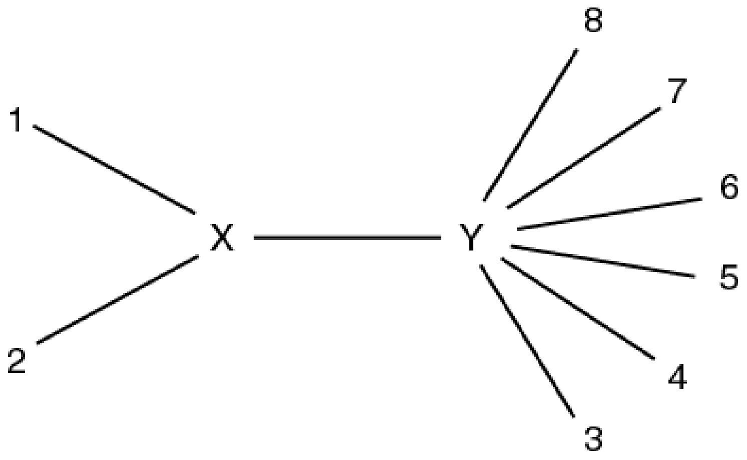
Begin by placing all the taxa in a star-like structure.

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.

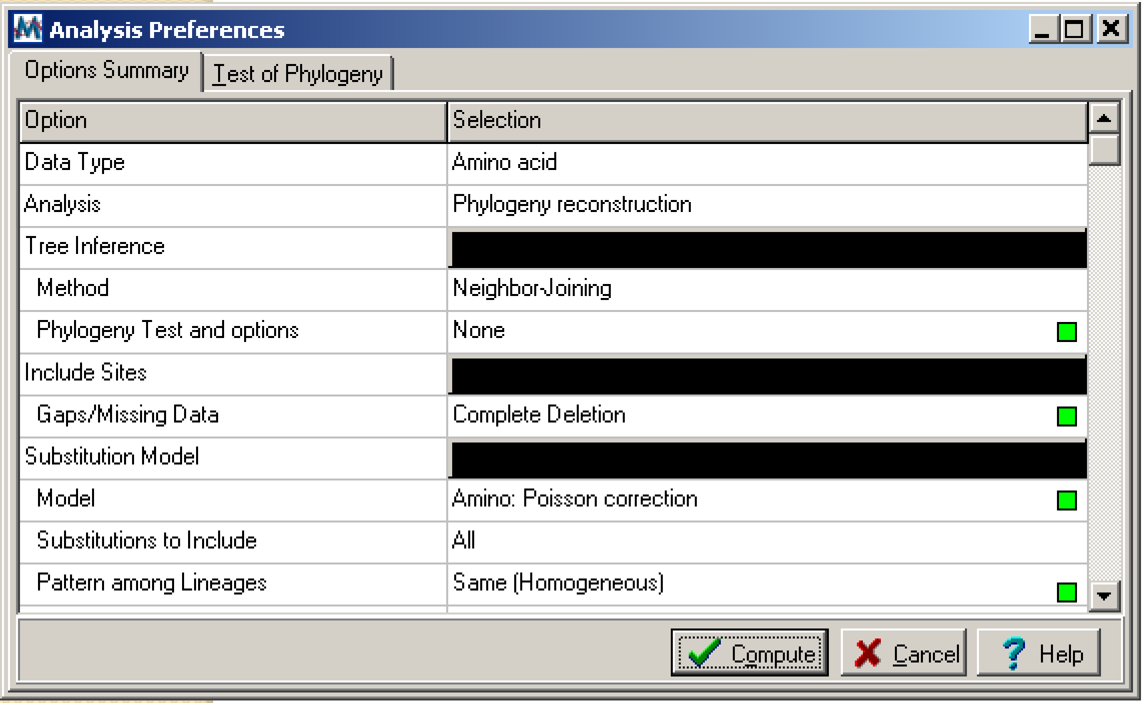
Making trees using neighbor-joining



Define the distance from X to Y by

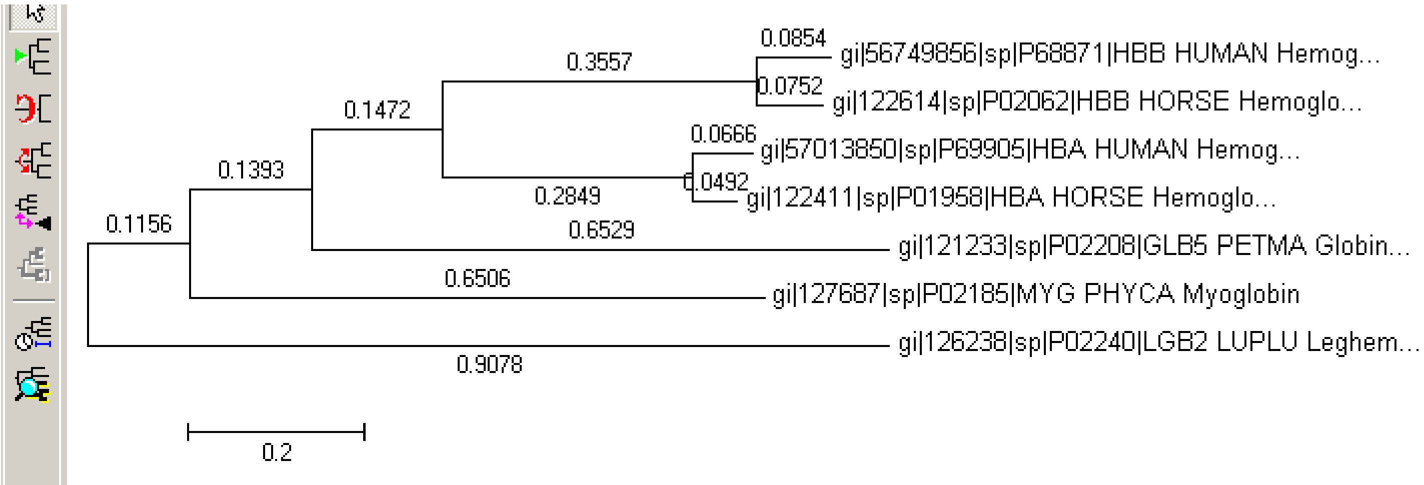
$d_{XY} = 1/2(d_{1Y} + d_{2Y} – d_{12})$

Use of MEGA for a distance-based tree: NJ



Option	Selection
Data Type	Amino acid
Analysis	Phylogeny reconstruction
Tree Inference	
Method	Neighbor-Joining
Phylogeny Test and options	None 🟩
Include Sites	
Gaps/Missing Data	Complete Deletion 🟩
Substitution Model	
Model	Amino: Poisson correction 🟩
Substitutions to Include	All
Pattern among Lineages	Same (Homogeneous) 🟩

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



0.0854
gi|56749856|sp|P68871|HBB HUMAN Hemog...
0.3557
0.0752
gi|122614|sp|P02062|HBB HORSE Hemoglo...
0.1472
0.0666
gi|57013850|sp|P69905|HBA HUMAN Hemog...
0.1393
0.0492
0.2849
gi|122411|sp|P01958|HBA HORSE Hemoglo...
0.1156
0.6529
gi|121233|sp|P02208|GLB5 PETMA Globin...
0.6506
gi|127687|sp|P02185|MYG PHYCA Myoglobin
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
0.9078
0.2

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. For about 12 taxa or fewer, evaluate all possible trees exhaustively; for >12 taxa perform a heuristic search.

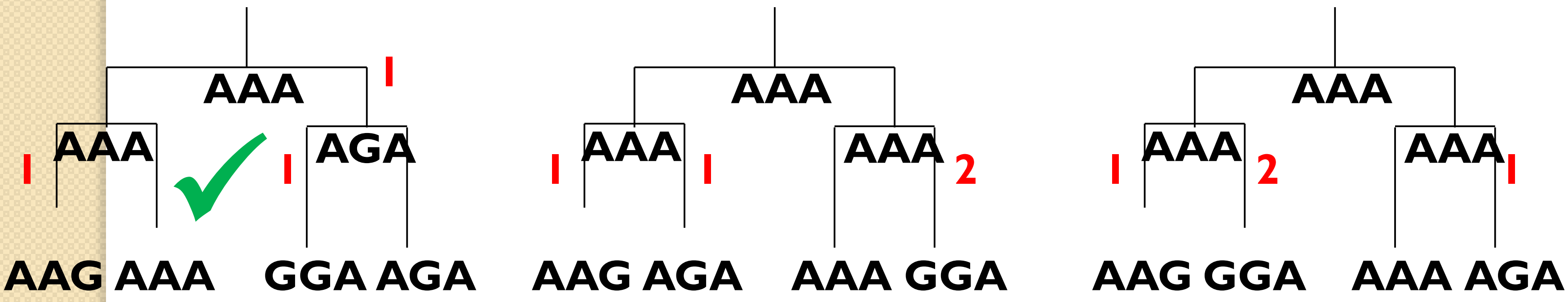
- Select the shortest tree (or trees).

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



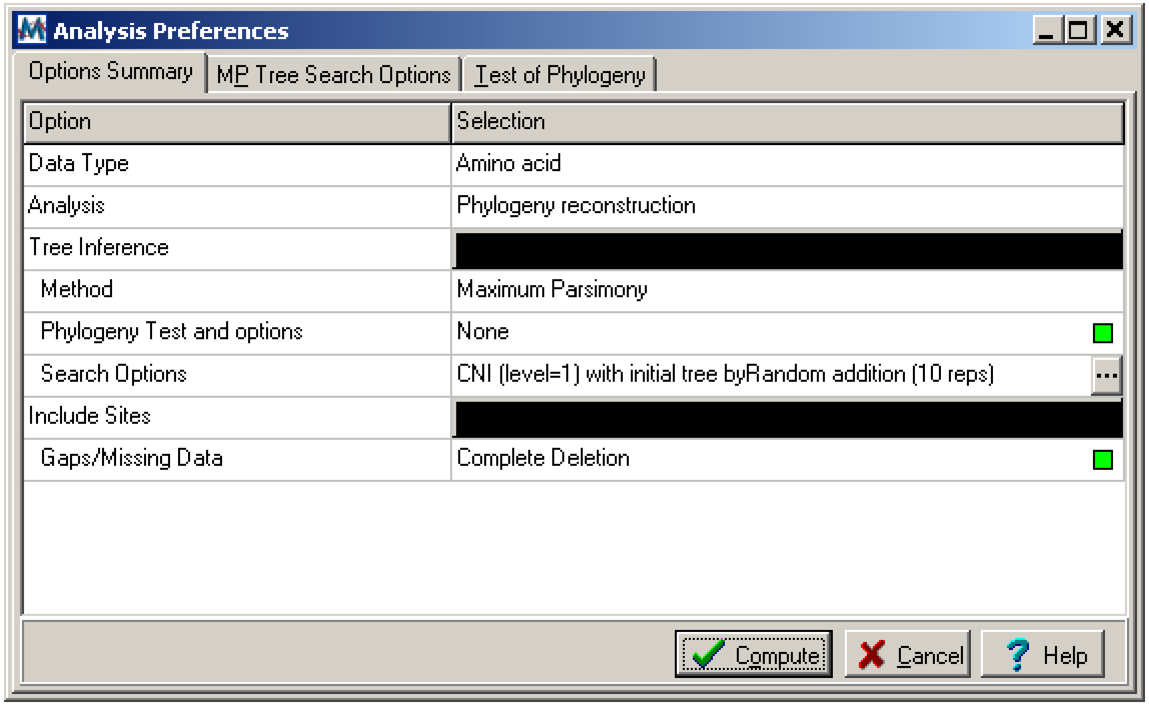
Cost = 3

Cost = 4

Cost = 4

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

MEGA for maximum parsimony (MP) trees

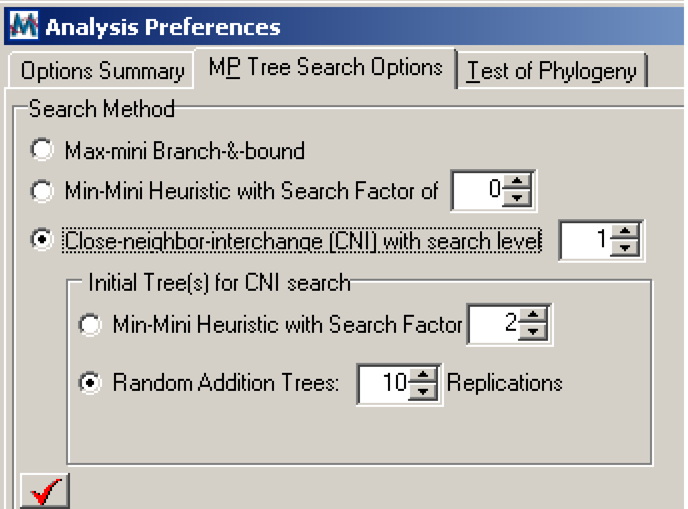


Analysis Preferences

Options Summary | MP Tree Search Options | Test of Phylogeny

Option	Selection
Data Type	Amino acid
Analysis	Phylogeny reconstruction
Tree Inference	
Method	Maximum Parsimony
Phylogeny Test and options	None 🟩
Search Options	CNI (level=1) with initial tree byRandom addition (10 reps) ...
Include Sites	
Gaps/Missing Data	Complete Deletion 🟩

Options include heuristic approaches, and bootstrapping



Analysis Preferences

Options Summary | MP Tree Search Options | Test of Phylogeny

Search Method

○ Max-mini Branch-&-bound

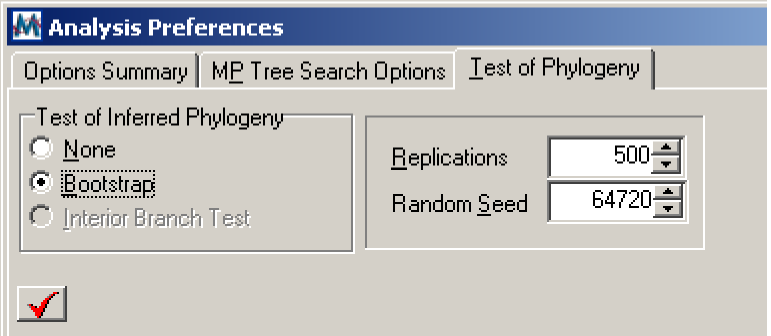
○ Min-Mini Heuristic with Search Factor of 0

● Close-neighbor-interchange (CNI) with search level 1

Initial Tree(s) for CNI search

○ Min-Mini Heuristic with Search Factor 2

● Random Addition Trees: 10 Replications



Analysis Preferences

Options Summary | MP Tree Search Options | Test of Phylogeny

Test of Inferred Phylogeny

○ None

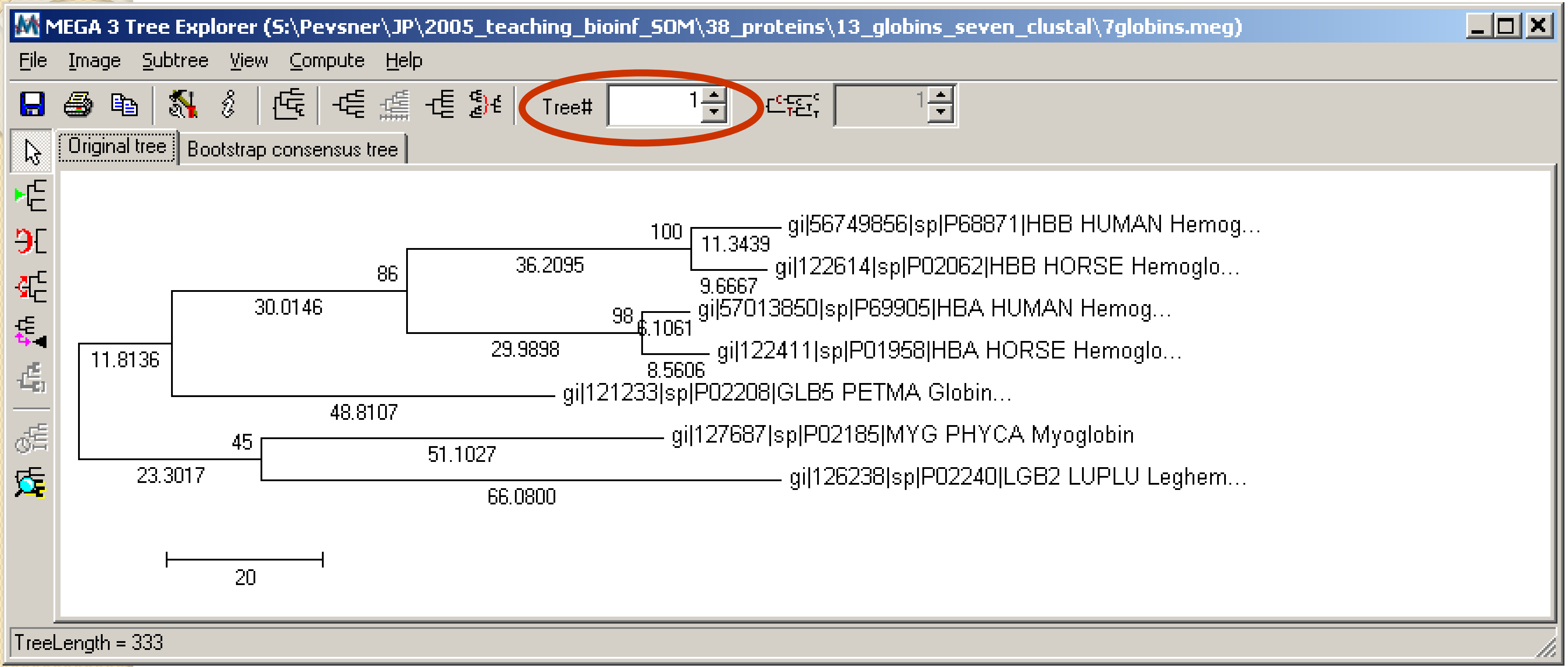
● Bootstrap

○ Interior Branch Test

Replications 500

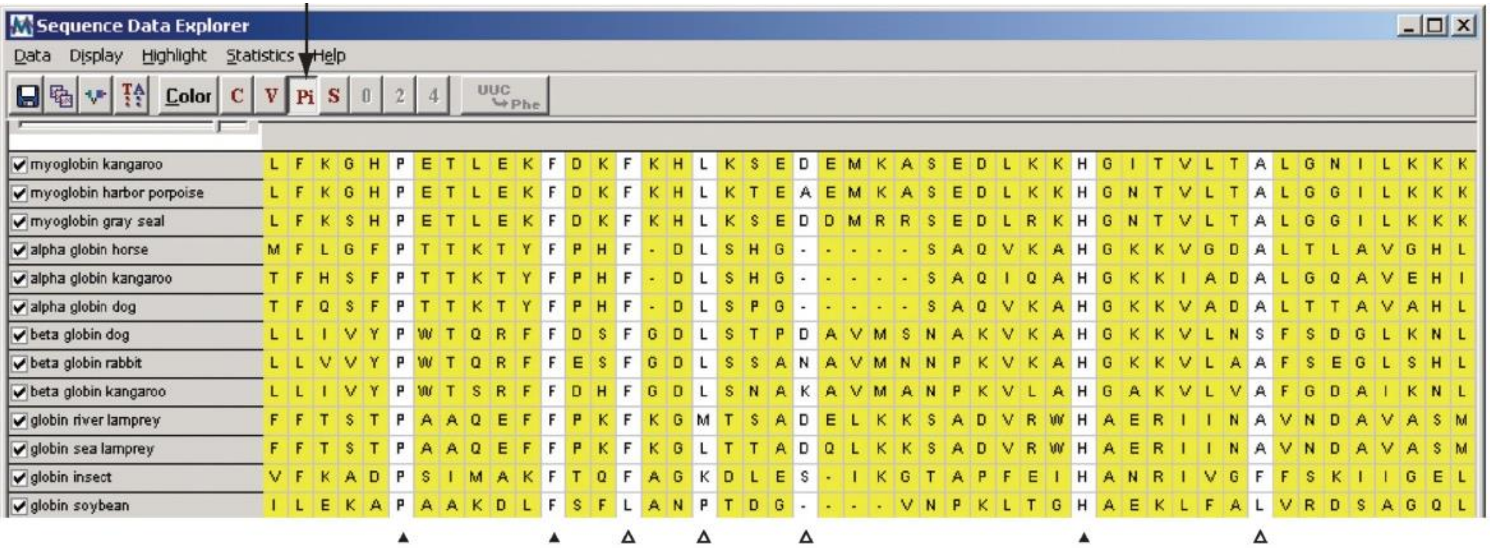
Random Seed 64720

MEGA for maximum parsimony (MP) trees


MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)
File
Image
Subtree
View
Compute
Help
Tree#
1
1
Original tree
Bootstrap consensus tree
100
11.3439
gi|56749856|sp|P68871|HBB HUMAN Hemog...
86
36.2095
gi|122614|sp|P02062|HBB HORSE Hemoglo...
9.6667
30.0146
98
6.1061
gi|57013850|sp|P69905|HBA HUMAN Hemog...
29.9898
gi|122411|sp|P01958|HBA HORSE Hemoglo...
8.5606
11.8136
gi|121233|sp|P02208|GLB5 PETMA Globin...
48.8107
45
gi|127687|sp|P02185|MYG PHYCA Myoglobin
51.1027
23.3017
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
66.0800
20
TreeLength = 333

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

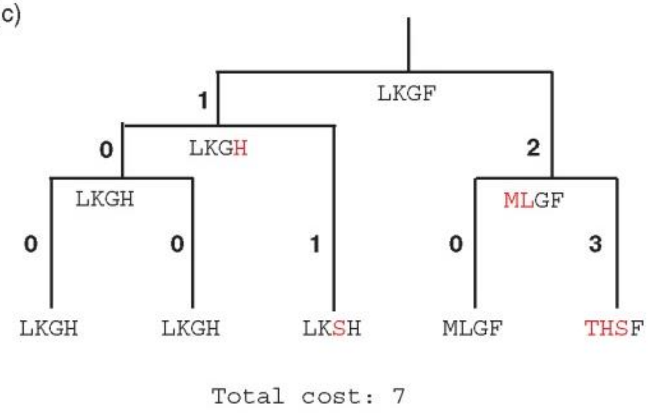


☑ myoglobin kangaroo	L	F	K	G	H	P	E	T	L	E	K	F	D	K	F	K	H	L	K	S	E	D	E	M	K	A	S	E	D	L	K	K	H	G	I	T	V	L	T	A	L	G	N	I	L	K	K	K
☑ myoglobin harbor porpoise	L	F	K	G	H	P	E	T	L	E	K	F	D	K	F	K	H	L	K	T	E	A	E	M	K	A	S	E	D	L	K	K	H	G	N	T	V	L	T	A	L	G	G	I	L	K	K	K
☑ myoglobin gray seal	L	F	K	S	H	P	E	T	L	E	K	F	D	K	F	K	H	L	K	S	E	D	D	M	R	R	S	E	D	L	R	K	H	G	N	T	V	L	T	A	L	G	G	I	L	K	K	K
☑ alpha globin horse	M	F	L	G	F	P	T	T	K	T	Y	F	P	H	F	-	D	L	S	H	G	-	-	-	-	-	S	A	Q	V	K	A	H	G	K	K	V	G	D	A	L	T	L	A	V	G	H	L
☑ alpha globin kangaroo	T	F	H	S	F	P	T	T	K	T	Y	F	P	H	F	-	D	L	S	H	G	-	-	-	-	-	S	A	Q	I	Q	A	H	G	K	K	I	A	D	A	L	G	Q	A	V	E	H	I
☑ alpha globin dog	T	F	Q	S	F	P	T	T	K	T	Y	F	P	H	F	-	D	L	S	P	G	-	-	-	-	-	S	A	Q	V	K	A	H	G	K	K	V	A	D	A	L	T	T	A	V	A	H	L
☑ beta globin dog	L	L	I	V	Y	P	W	T	Q	R	F	F	D	S	F	G	D	L	S	T	P	D	A	V	M	S	N	A	K	V	K	A	H	G	K	K	V	L	N	S	F	S	D	G	L	K	N	L
☑ beta globin rabbit	L	L	V	V	Y	P	W	T	Q	R	F	F	E	S	F	G	D	L	S	S	A	N	A	V	M	N	N	P	K	V	K	A	H	G	K	K	V	L	A	A	F	S	E	G	L	S	H	L
☑ beta globin kangaroo	L	L	I	V	Y	P	W	T	S	R	F	F	D	H	F	G	D	L	S	N	A	K	A	V	M	A	N	P	K	V	L	A	H	G	A	K	V	L	V	A	F	G	D	A	I	K	N	L
☑ globin river lamprey	F	F	T	S	T	P	A	A	Q	E	F	F	P	K	F	K	G	M	T	S	A	D	E	L	K	K	S	A	D	V	R	W	H	A	E	R	I	I	N	A	V	N	D	A	V	A	S	M
☑ globin sea lamprey	F	F	T	S	T	P	A	A	Q	E	F	F	P	K	F	K	G	L	T	T	A	D	Q	L	K	K	S	A	D	V	R	W	H	A	E	R	I	I	N	A	V	N	D	A	V	A	S	M
☑ globin insect	V	F	K	A	D	P	S	I	M	A	K	F	T	Q	F	A	G	K	D	L	E	S	-	I	K	G	T	A	P	F	E	I	H	A	N	R	I	V	G	F	F	S	K	I	I	G	E	L
☑ globin soybean	I	L	E	K	A	P	A	A	K	D	L	F	S	F	L	A	N	P	T	D	G	-	-	-	-	V	N	P	K	L	T	G	H	A	E	K	L	F	A	L	V	R	D	S	A	G	Q	L
						▲						▲			△			△				△											▲							△								

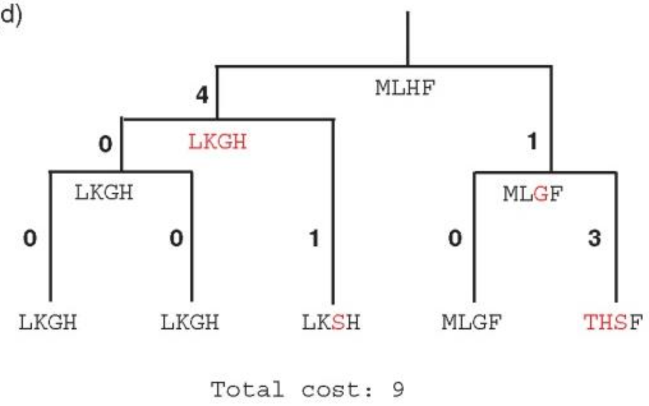
(b)

kangaroo	LKGH
porpoise	LKGH
gray seal	LKSH
horse α	MLGF
kangaroo α	THSF

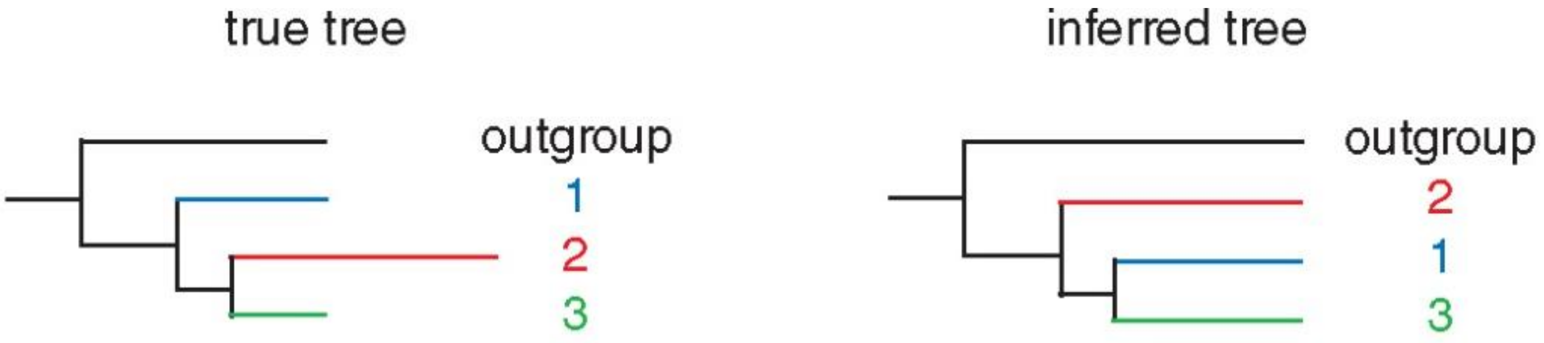
(c)



(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.

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

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?

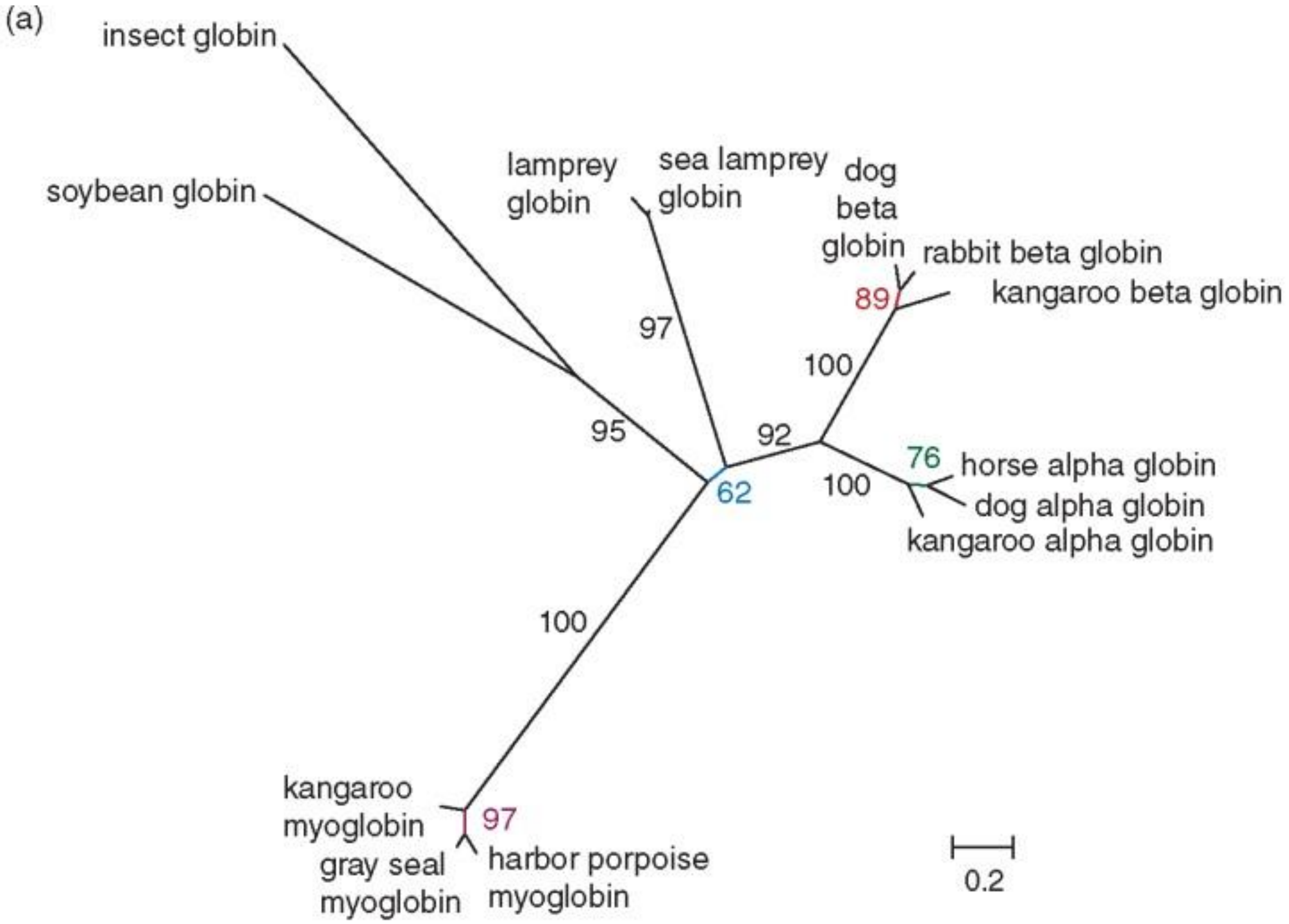
ML is implemented in the TREE-PUZZLE program, as well as MEGA5, PAUP and PHYLIP.

Maximum likelihood: Tree-Puzzle

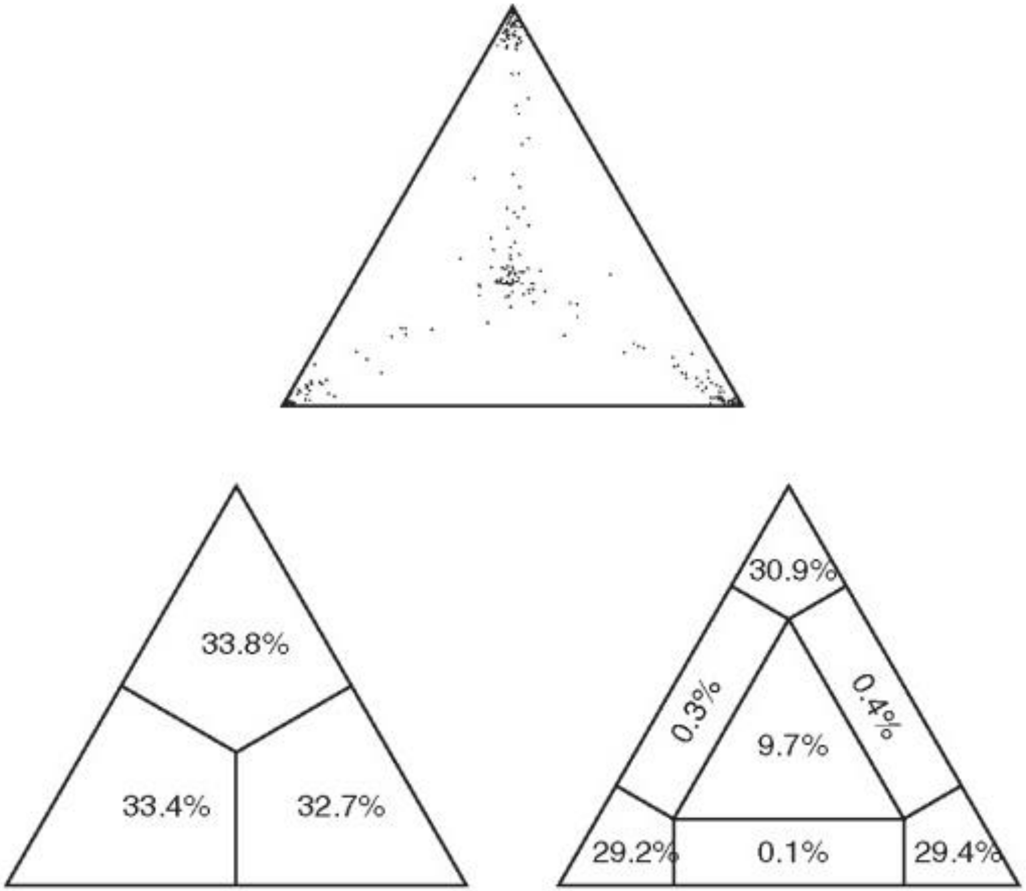
(1) Reconstruct all possible quartets A, B, C, D.
For 12 myoglobins there are 495 possible quartets.

(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



Quartet puzzling: phylogeny by maximum likelihood



Likelihood mapping indicates the frequency with which quartets are resolved. Top: all possible quartets (n=495). Each quartet has 3 posterior weights mapped in triangles. For 13 globins, only 9.7% of quartets are unresolved.

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.

Bayesian inference of phylogeny with MrBayes

Calculate:

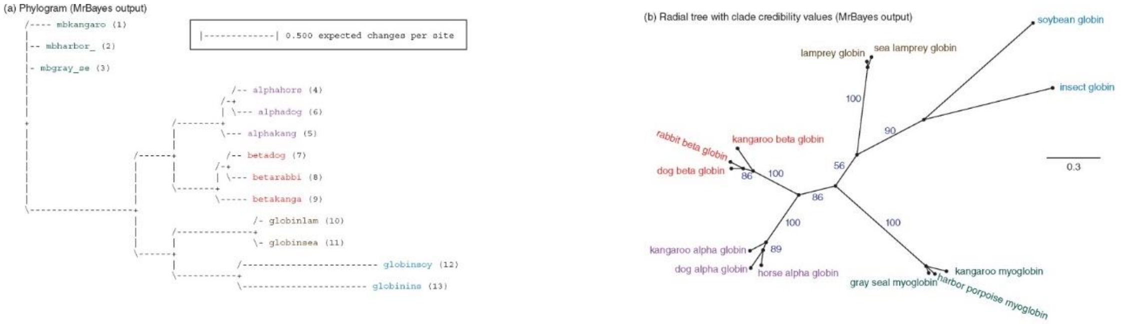
$$\text{Pr [Tree | Data] } = \frac{\text{Pr [Data | Tree] x Pr [Tree]}}{\text{Pr [Data]}}$$

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

Example:

- Align 13 globin proteins with MAFFT (Chapter 6).
- 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.

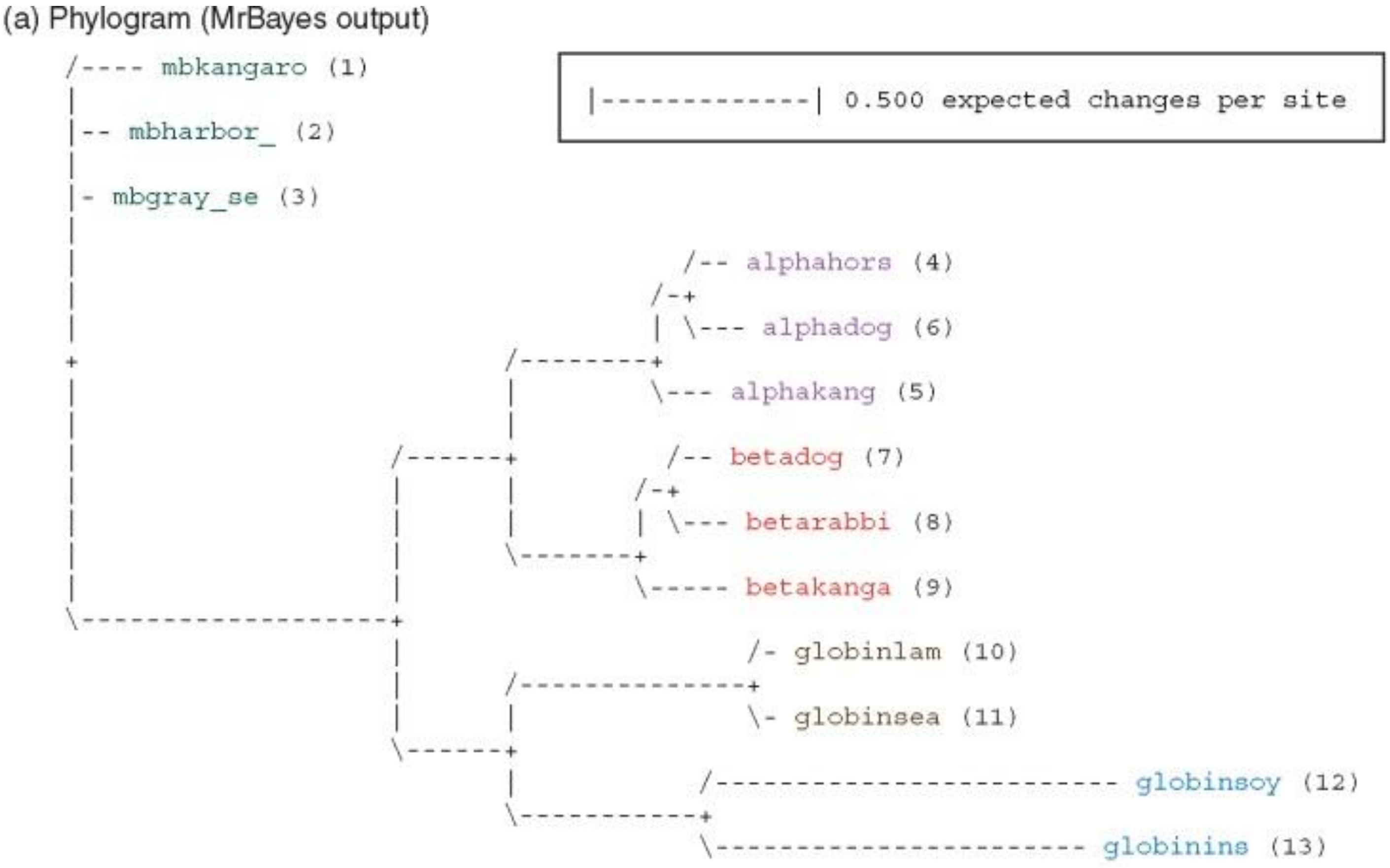


(a) Phylogram (MrBayes output)

```
/---- mbkangaro (1)                        |------------| 0.500 expected changes per site
|
|-- mbharbor_ (2)
|
|- mbgray_se (3)
|
|                                      /-- alphahors (4)
|                                    /-+
|                                    | \--- alphadog (6)
|                          /---------+
|                          |         \--- alphakang (5)
+                          |
|                /-------+           /-- betadog (7)
|                |         |       /-+
|                |         |       | \--- betarabbi (8)
|                |         \-------+
|                |                 \----- betakanga (9)
\----------------+
                 |                         /- globinlam (10)
                 |         /--------------+
                 |         |               \- globinsea (11)
                 \-------+
                           |          /------------------------ globinsoy (12)
                           \----------+
                                      \---------------------- globinins (13)
```

(b) Radial tree with clade credibility values (MrBayes output)

Bayesian inference of phylogeny

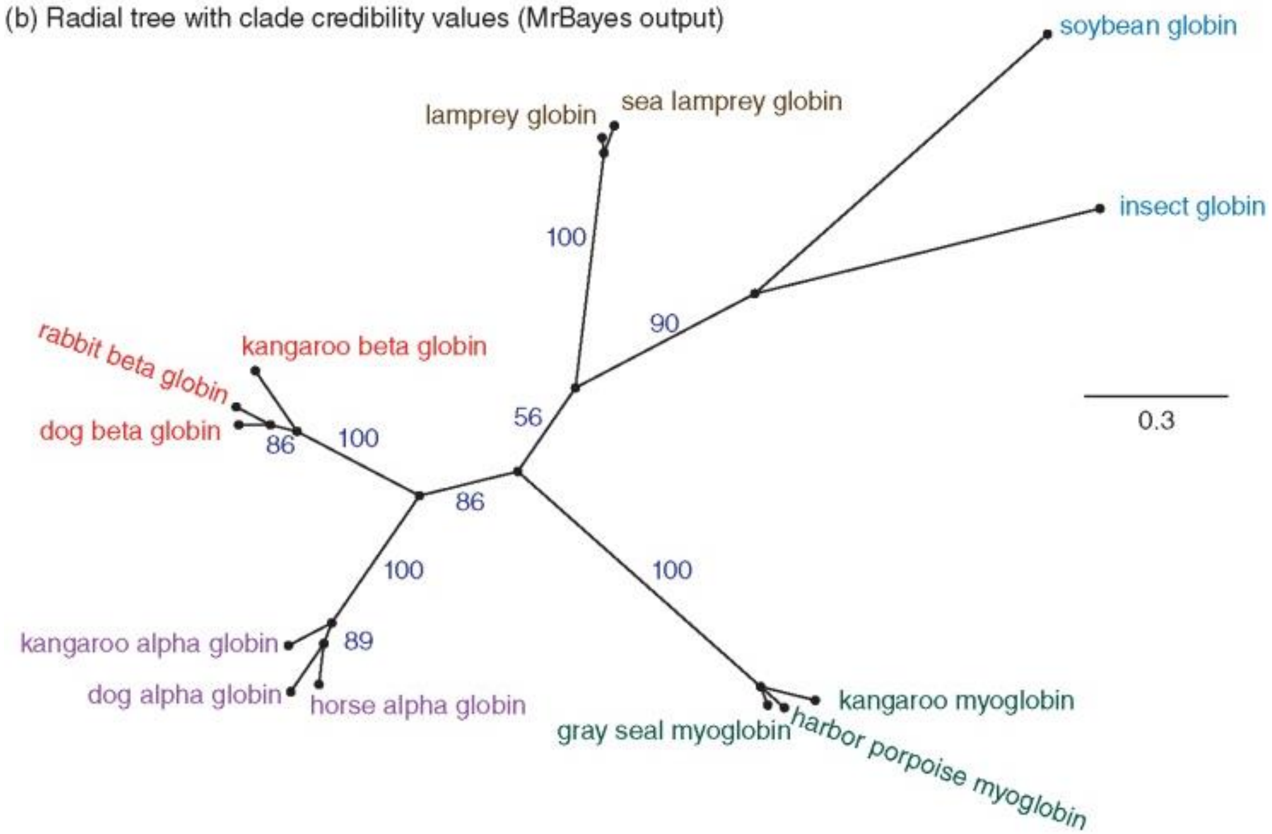


(a) Phylogram (MrBayes output)

```
|-------------| 0.500 expected changes per site

/---- mbkangaro (1)
|
|-- mbharbor_ (2)
|
|- mbgray_se (3)
|
|                                     /-- alphahors (4)
|                                   /-+
|                                   | \--- alphadog (6)
+                          /--------+
|                          |        \--- alphakang (5)
|                          |
|                   /------+         /-- betadog (7)
|                   |      |       /-+
|                   |      |       | \--- betarabbi (8)
|                   |      \-------+
|                   |              \----- betakanga (9)
\-------------------+
                    |                     /- globinlam (10)
                    |      /--------------+
                    |      |              \- globinsea (11)
                    \------+
                           |           /------------------------- globinsoy (12)
                           \-----------+
                                       \----------------------- globinins (13)
```

Phylogram shows clades (note myoglobins are unresolved).

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.

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

- Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

- Topologies and branch lengths of trees
- Tree roots
- Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

- Stage 1: sequence acquisition
- Stage 2: multiple sequence alignment
- Stage 3: models of DNA and amino acid substitution
- Stage 4: tree-building methods (distance-based; maximum parsimony; maximum likelihood; Bayesian methods)
- Stage 5: evaluating trees

Perspective

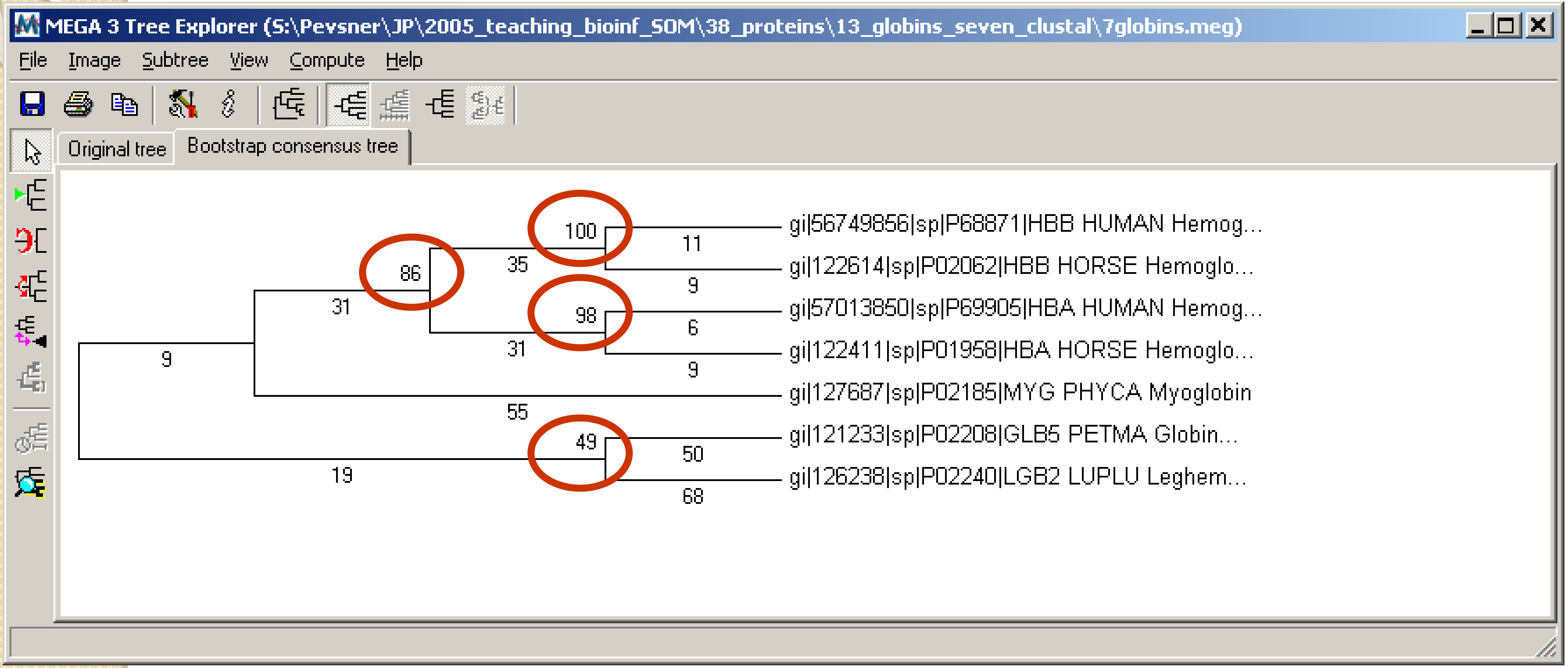
Stage 5: Evaluating trees

The main criteria by which the accuracy of a phylogentic 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


MEGA 3 Tree Explorer (S:\Pevsner\JP\2005_teaching_bioinf_SOM\38_proteins\13_globins_seven_clustal\7globins.meg)
File
Image
Subtree
View
Compute
Help
Original tree
Bootstrap consensus tree
100
11
gi|56749856|sp|P68871|HBB HUMAN Hemog...
86
35
gi|122614|sp|P02062|HBB HORSE Hemoglo...
9
31
98
gi|57013850|sp|P69905|HBA HUMAN Hemog...
6
31
gi|122411|sp|P01958|HBA HORSE Hemoglo...
9
9
gi|127687|sp|P02185|MYG PHYCA Myoglobin
55
49
gi|121233|sp|P02208|GLB5 PETMA Globin...
50
19
gi|126238|sp|P02240|LGB2 LUPLU Leghem...
68

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.

Outline

Introduction to molecular evolution

Principles of molecular phylogeny and evolution

Goals; historical background; molecular clock hypothesis; positive and negative selection; neutral theory of evolution

Molecular phylogeny: properties of trees

Topologies and branch lengths of trees

Tree roots

Enumerating trees and selecting search strategies

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

Five stages of phylogenetic analysis

Stage 1: sequence acquisition

Stage 2: multiple sequence alignment

Stage 3: models of DNA and amino acid substitution

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

Stage 5: evaluating trees

Perspective

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.