(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13

Peter Güntert

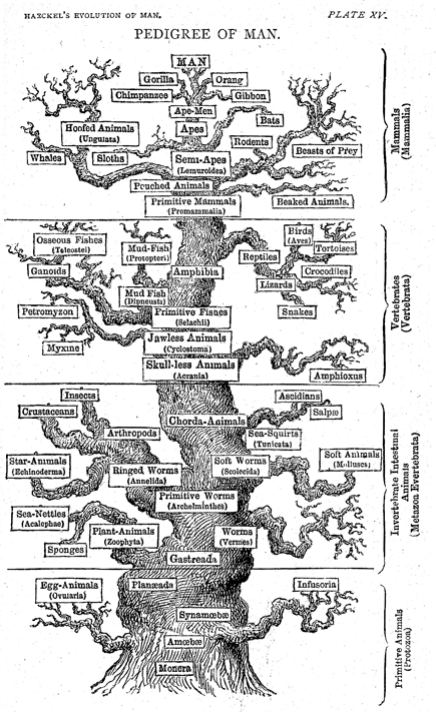
Phylogeny

Outline

- Introduction to evolution and phylogeny
- Nomenclature of trees
- Five stages of molecular phylogeny:
    1. selecting sequences
    2. multiple sequence alignment
    3. models of substitution
    4. tree-building
    5. tree evaluation
- Practical approaches to making trees

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.

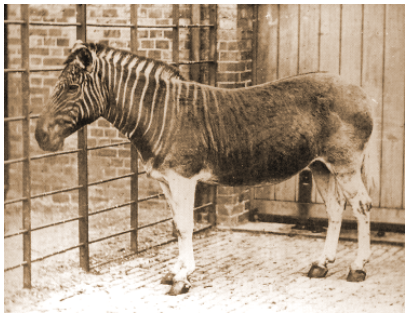
Introduction

- 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:

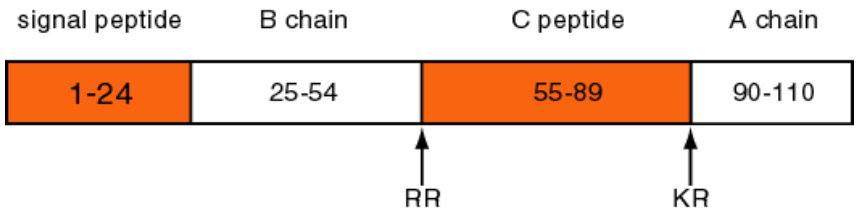
- How many genes are related to my favorite gene?
- How related are whales, dolphins & porpoises to cows?
- Where and when did HIV or other viruses originate?
- What is the history of life on earth?
- Was the extinct quagga more like a zebra or a horse?



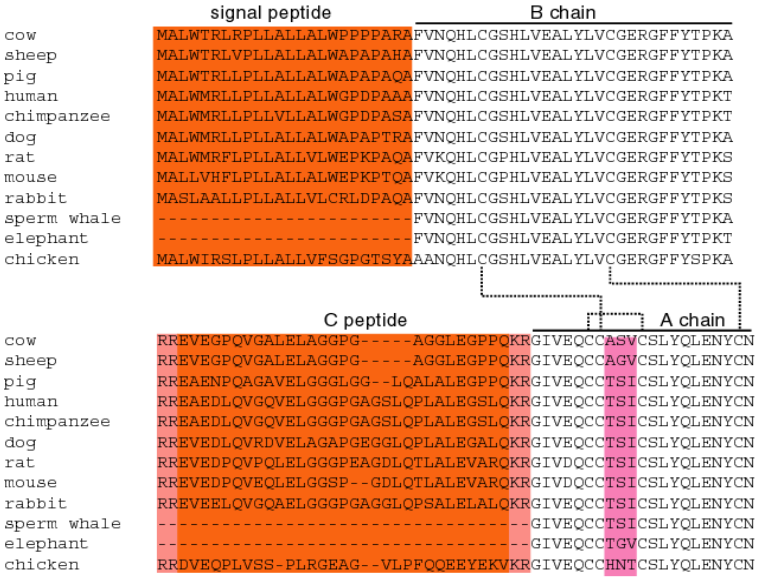
Historical background

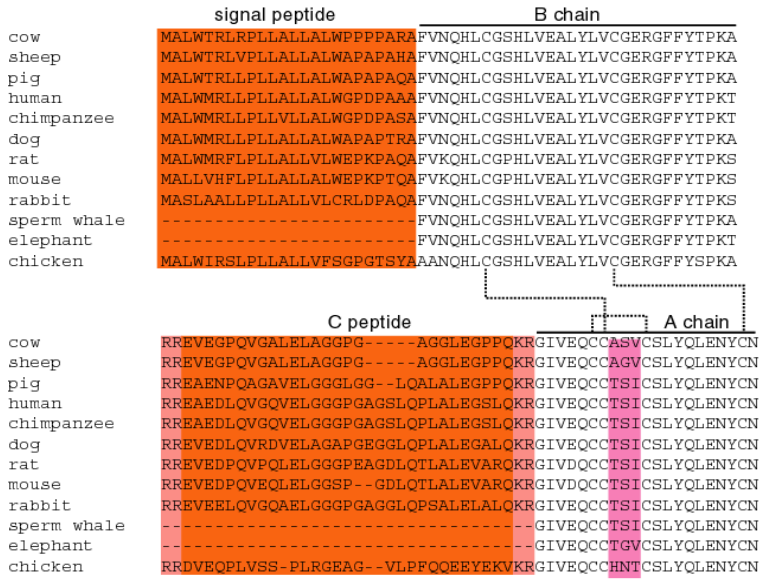
- Studies of molecular evolution began with the first sequencing of proteins, beginning in the 1950s.
- In 1953 Frederick Sanger and colleagues determined the primary amino acid sequence of insulin.

Mature insulin consists of an A chain and B chain heterodimer connected by disulphide bridges



The signal peptide and C peptide are cleaved, and their sequences display fewer functional constraints.

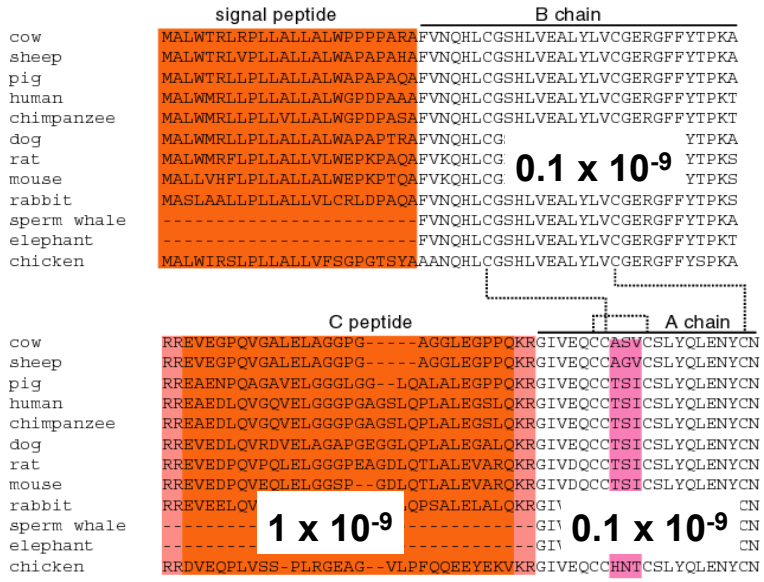




Note the sequence divergence in the disulfide loop region of the A chain.

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 the rate of nucleotide (and of amino acid) substitution is about 6–10 fold higher in the C peptide, relative to the A and B chains.



Number of nucleotide substitutions/site/year

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 insulin have undergone an extremely rapid rate of evolutionary change

Arrows indicate positions at which guinea pig insulin (A chain and B chain) differs from both human and mouse.

Molecular clock hypothesis

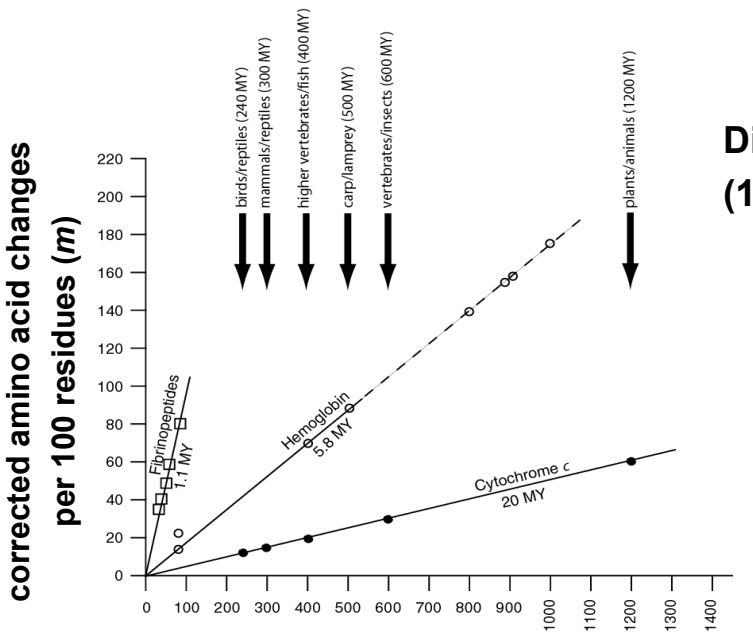
- In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes *c*, and fibrinopeptides. Some proteins appeared to evolve slowly, while others evolved rapidly.
- Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:

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)

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 for proteins: rate of substitutions per aa site per 10^9 years

Fibrinopeptides	9.0
Kappa casein	3.3
Lactalbumin	2.7
Serum albumin	1.9
Lysozyme	0.98
Trypsin	0.59
Insulin	0.44
Cytochrome c	0.22
Histone H2B	0.09
Ubiquitin	0.010
Histone H4	0.010

Molecular clock hypothesis: implications

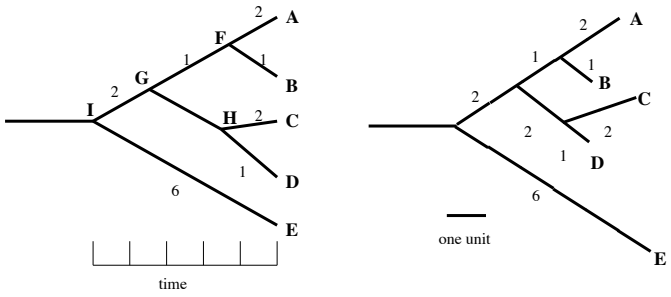
- If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.

Molecular phylogeny: nomenclature of trees

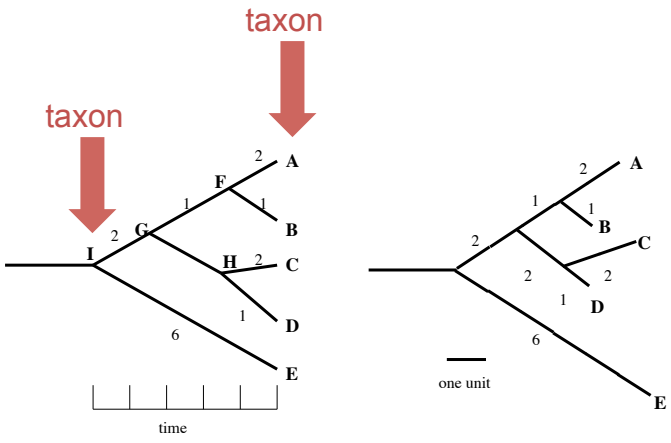
There are two main kinds of information inherent to any tree: topology and branch lengths.

We will now describe the parts of a tree.

Molecular phylogeny uses trees to depict evolutionary relationships among organisms. These trees are based upon DNA and protein sequence data.



Tree nomenclature

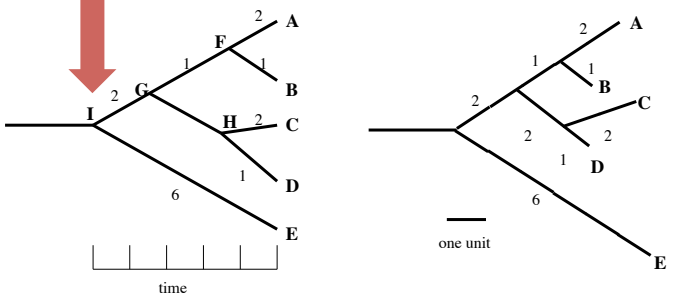


Tree nomenclature

operational taxonomic unit (OTU)

such as a protein sequence

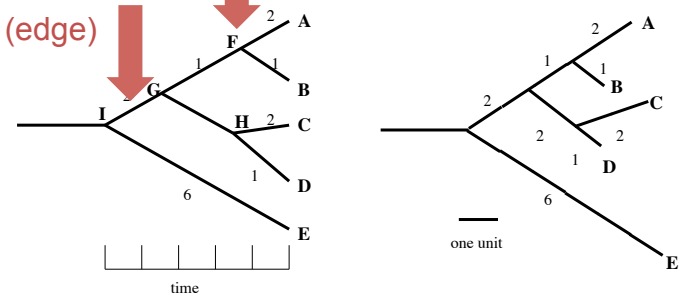
taxon



Tree nomenclature

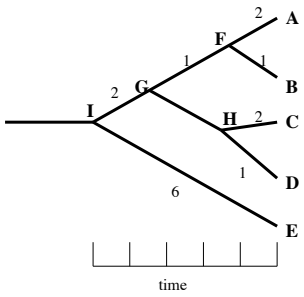
Node (intersection or terminating point of two or more branches)

branch (edge)



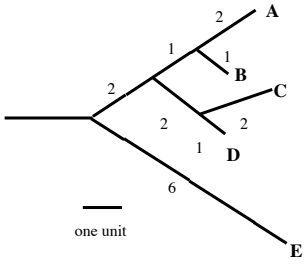
Tree nomenclature

Branches are unscaled...



…OTUs are neatly aligned, and nodes reflect time

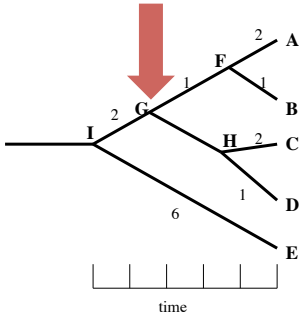
Branches are scaled...



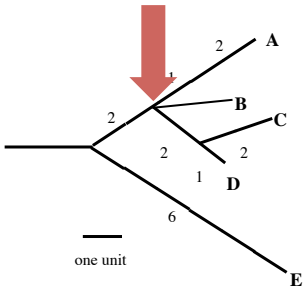
…branch lengths are proportional to number of amino acid changes

Tree nomenclature

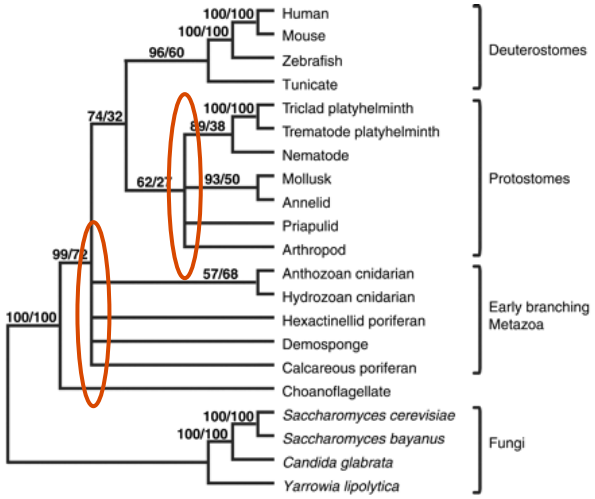
Bifurcating internal node



multifurcating internal node



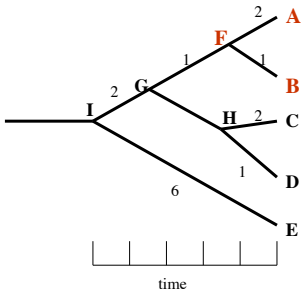
Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes



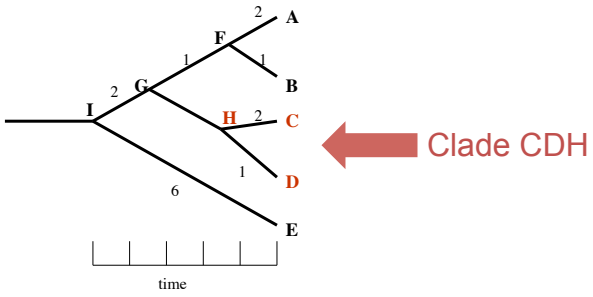
Rokas, A. et al. Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933 (2005), Fig. 1.

Tree nomenclature: clades

Clade ABF (monophyletic group)



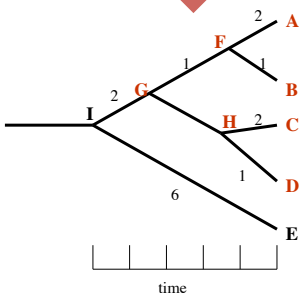
Tree nomenclature

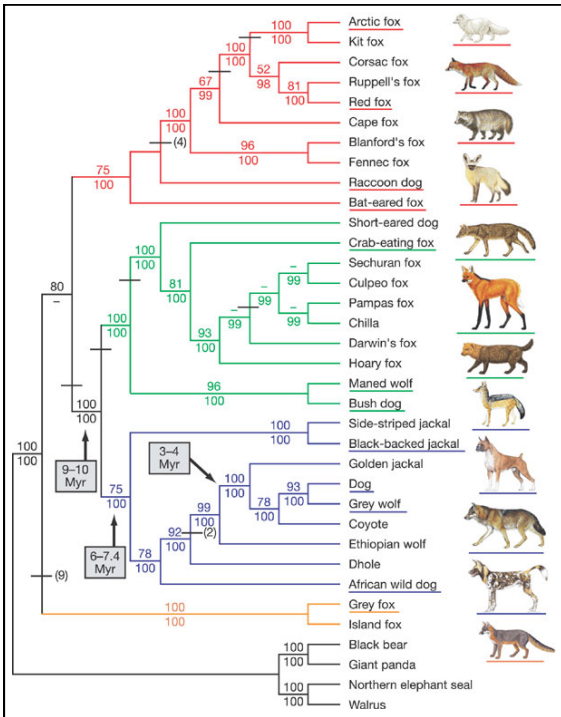


Clade CDH

Tree nomenclature

Clade ABF/CDH/G





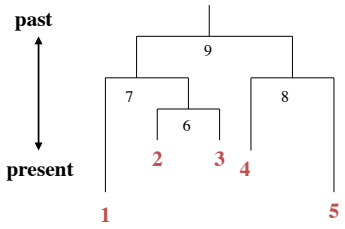
Examples of clades

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

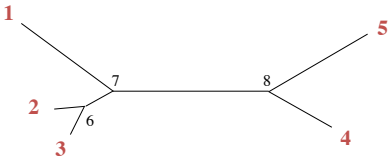
Tree roots

- 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

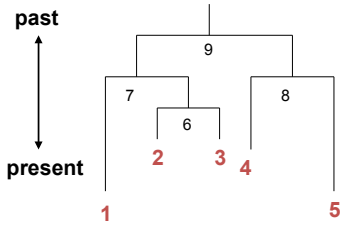


Rooted tree
(specifies evolutionary path)

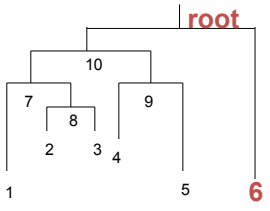


Unrooted tree

Tree nomenclature: outgroup rooting



Rooted tree



Outgroup
(used to place the root)

Enumerating trees

Cavalii-Sforza and Edwards (1967) derived the number of possible unrooted trees (N_U) for n OTUs ($n \geq 3$):

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

The number of bifurcating rooted trees (N_R):

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

For 10 OTUs (e.g. 10 DNA or protein sequences), the number of possible rooted trees is ≈ 34 million, and the number of unrooted trees is ≈ 2 million. Many tree-making algorithms can exhaustively examine every possible tree for up to ten to twelve sequences.

Numbers of possible trees is extremely large for >10 sequences

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}

Five stages of phylogenetic analysis

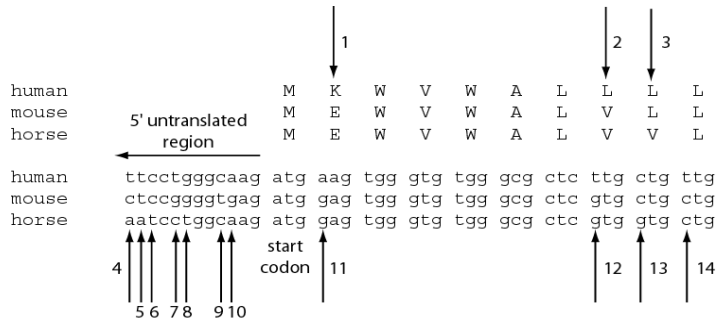
1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Selection of a substitution model
4. Tree building
5. Tree evaluation

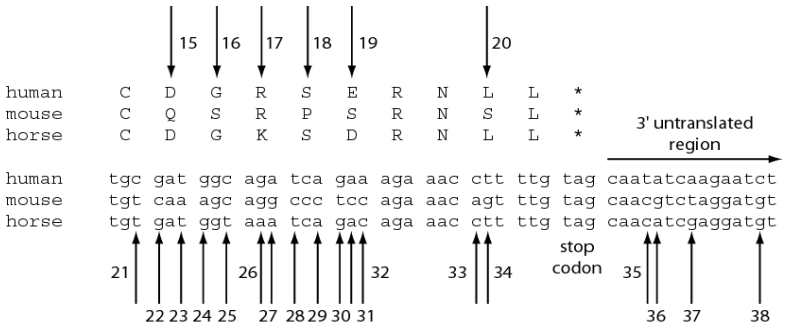
Stage 1: Use of DNA, RNA, or protein

For some phylogenetic studies, it may be preferable to use protein instead of DNA sequences. We saw that in pairwise alignment and in BLAST searching, protein is often more informative than DNA.

Stage 1: Use of DNA, RNA, or protein

- For phylogeny, DNA can be more informative.
- The protein-coding portion of DNA has synonymous and nonsynonymous substitutions. Thus, some DNA changes do not have corresponding protein changes.





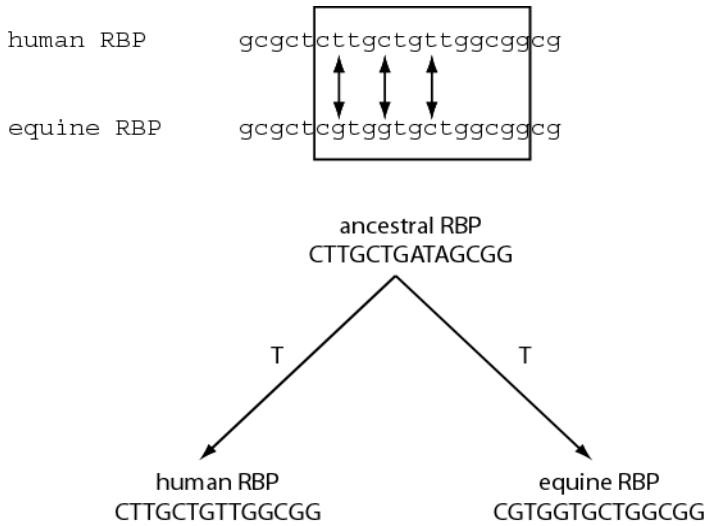
Stage 1: Use of DNA, RNA, or protein

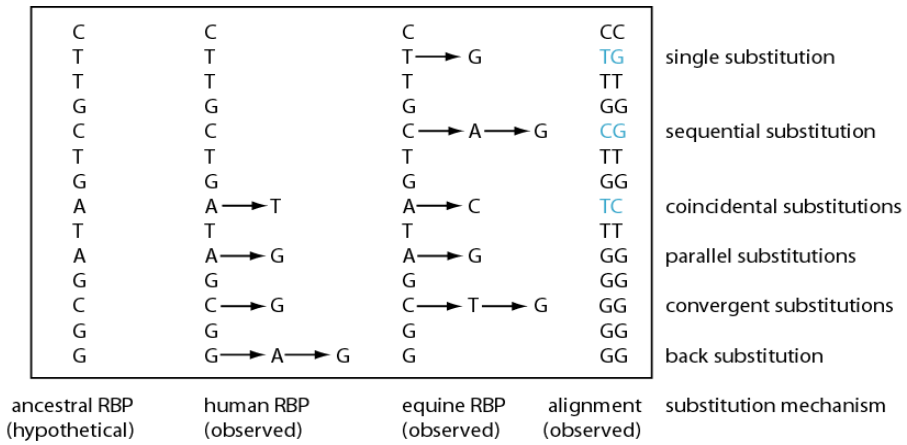
- For phylogeny, DNA can be more informative.
- The protein-coding portion of DNA has synonymous and non-synonymous substitutions. Thus, some DNA changes do not have corresponding protein changes.
- If the synonymous substitution rate (d_S) is greater than the non-synonymous substitution rate (d_N), 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.

Substitutions in a DNA sequence alignment can be directly observed, or inferred





Stage 1: Use of DNA, RNA, or protein

- For phylogeny, DNA can be more informative.
- Noncoding regions (such as 5' and 3' untranslated regions) may be analyzed using molecular phylogeny.
- Pseudogenes (nonfunctional genes) are studied by molecular phylogeny
- Rates of transitions and transversions can be measured.
- Transitions: purine (A⟷ G) or pyrimidine (C ⟷T) substitutions
- Transversion: purine ⟷ pyrimidine

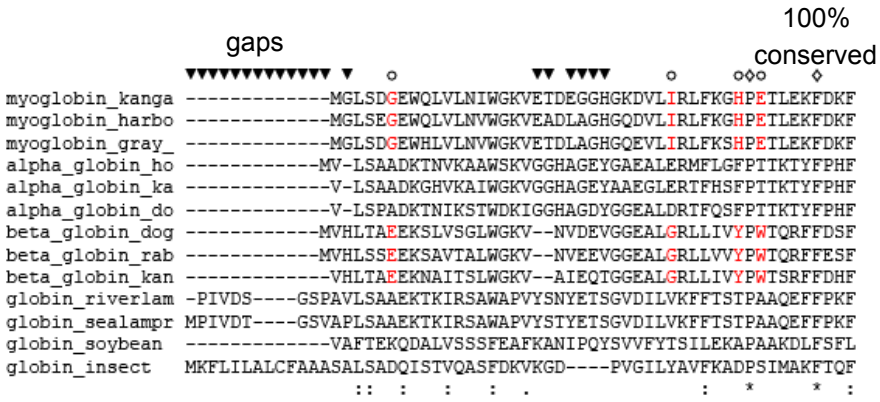
Five stages of phylogenetic analysis

1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Selection of a substitution model
4. Tree building
5. Tree evaluation

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.)

Alignment of orthologous globins



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

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 or gaps).

Five stages of phylogenetic analysis

1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Selection of a substitution model
4. Tree building
5. Tree evaluation

Stage 4: Tree-building methods: distance

- 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.

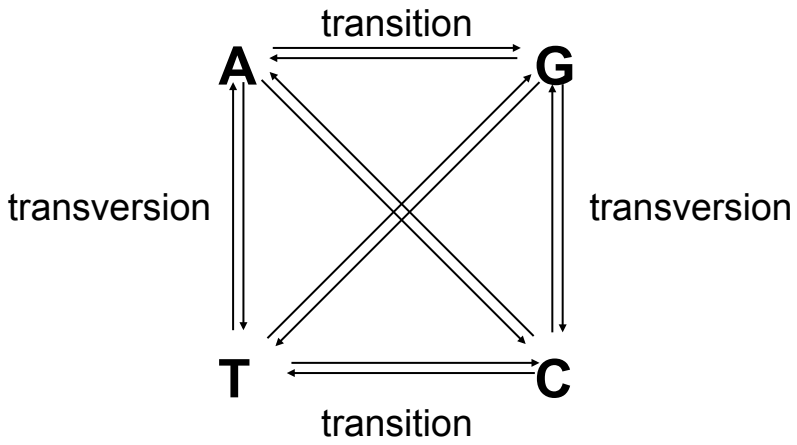
Stage 4: Tree-building methods: distance

- Jukes and Cantor (1969) proposed a corrective formula ($p = n/N$):

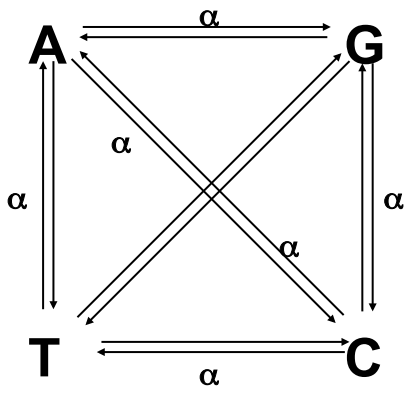
$$D = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

- 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.

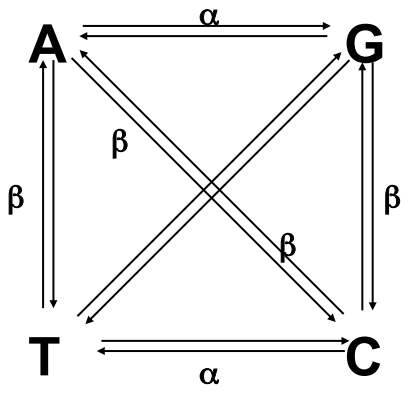
Models of nucleotide substitution



Jukes and Cantor one-parameter model of nucleotide substitution ($\alpha = \beta$)



Kimura model of nucleotide substitution (assumes $\alpha \neq \beta$)



Stage 4: Tree-building methods: distance

Jukes and Cantor (1969) proposed a corrective formula:

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

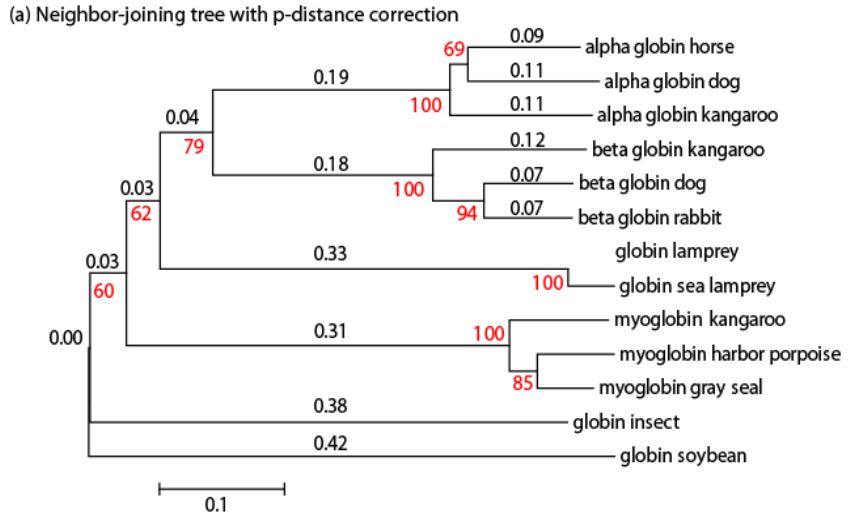
Consider an alignment where 3/60 aligned residues differ.
The normalized Hamming distance is 3/60 = 0.05.
The Jukes-Cantor correction is

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

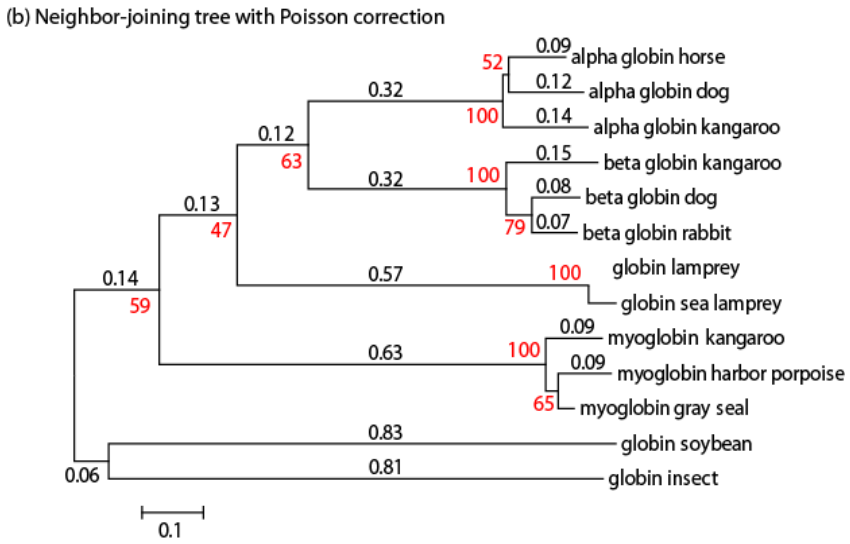
When 30/60 aligned residues differ, the Jukes-Cantor correction is more substantial:

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

Each model can affect the topology and branch lengths of the tree



p-distance correction



Poisson correction

Five stages of phylogenetic analysis

1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Selection of a substitution model
4. Tree building
5. Tree evaluation

Stage 4: Tree-building methods

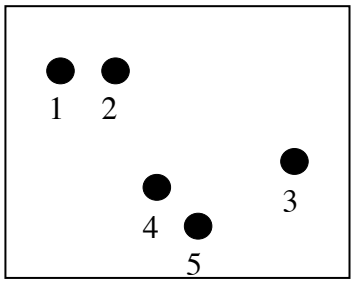
- Different tree-building methods: distance-based and character-based.
- Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score.
- Examples of distance-based algorithms are UPGMA and neighbor-joining.

Stage 4: Tree-building methods

1. distance-based
2. character-based: maximum parsimony
3. character- and model-based: maximum likelihood
4. character- and model-based: Bayesian

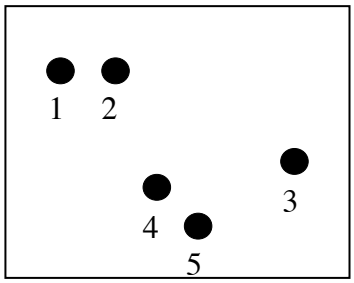
Tree-building methods: UPGMA

UPGMA is
unweighted pair group method
using arithmetic mean



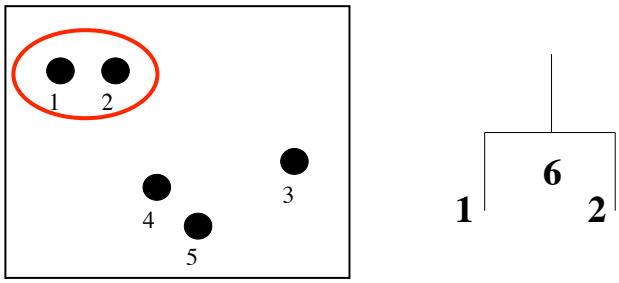
Tree-building methods: UPGMA

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.



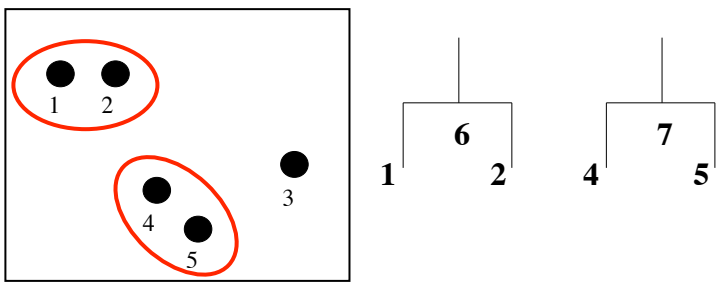
Tree-building methods: UPGMA

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



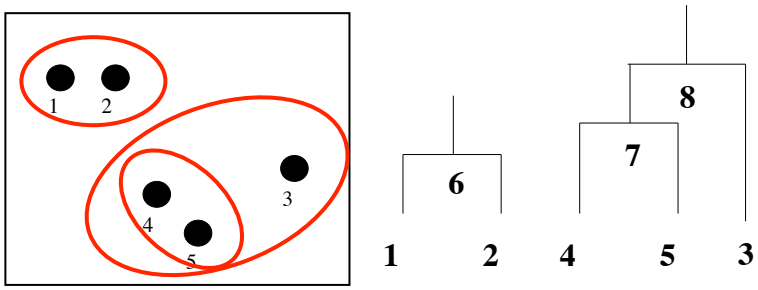
Tree-building methods: UPGMA

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



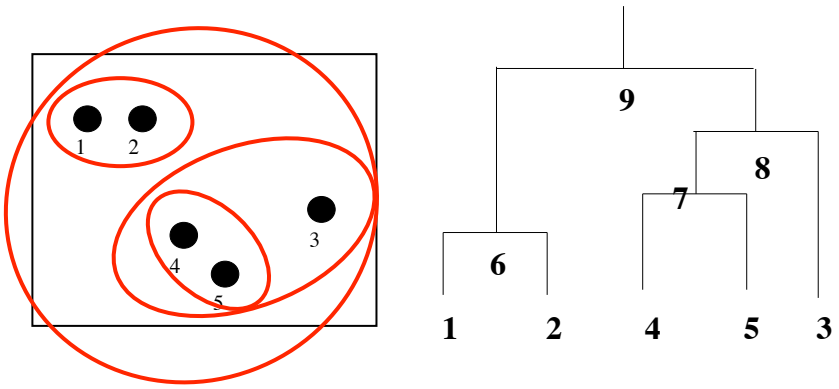
Tree-building methods: UPGMA

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 other approaches, e.g. the neighbor-joining approach.

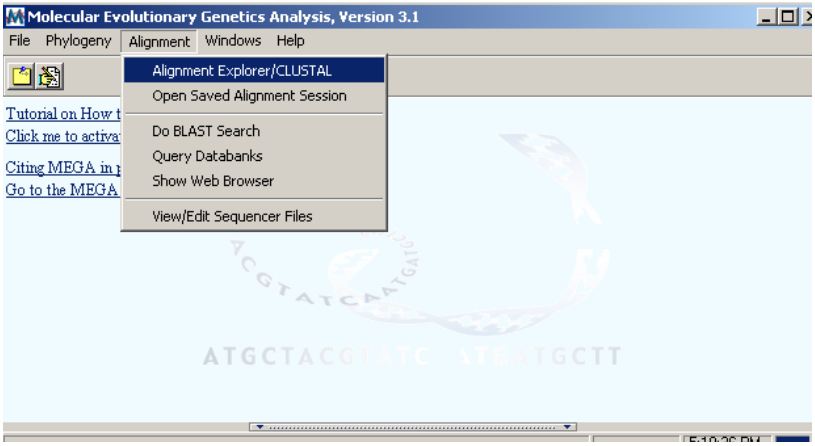
MEGA software

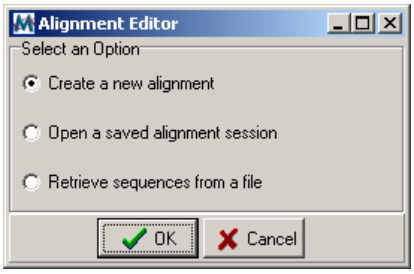
Molecular Evolutionary Genetics Analysis

http://www.megasoftware.net/

MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses.

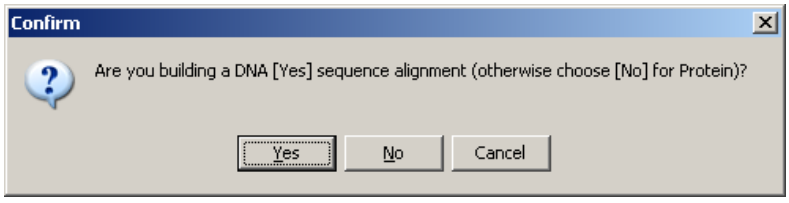
Use MEGA to make phylogenetic trees

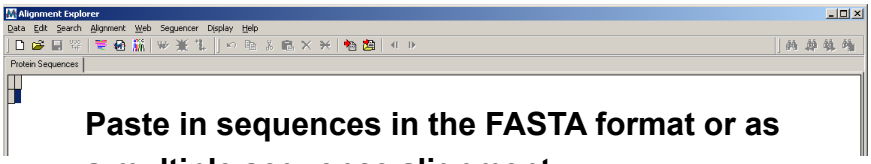




Open the alignment editor…

Choose DNA or protein…





Paste in sequences in the FASTA format or as a multiple sequence alignment…

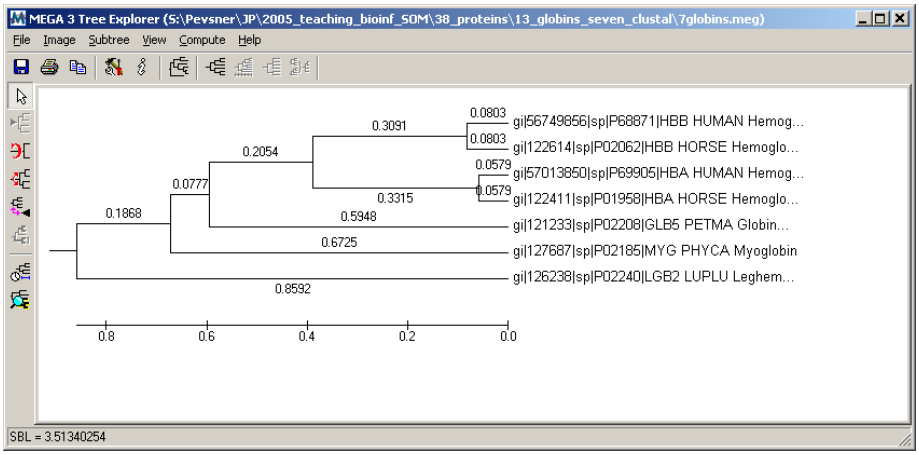
You can use a set of protein or DNA sequences in the FASTA format obtained from HomoloGene

homologene.txt - Notepad

File Edit Format View Help

```
>gi|7661534|ref|NP_054862.1| CD274 molecule [Homo sapiens]
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHG
EEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNAPYNKINQR
ILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVTSTLRINTTTNEIFYCT
FRRLDPEENHTAELVIPELPLAHPPNERTHLVILGAILLCLGVALTFIFRLRKGRMMDVKKCGIQDTNSK
KQSDTHLEET
>gi|114623690|ref|XP_001140705.1| PREDICTED: CD274 antigen isoform 2 [Pan troglodytes]
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHG
EEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNAPYNKINQR
ILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVTSTLRINTTTNEIFYCT
FRRLDPEENHTAELVIPELPLAHPPNERTHLVILGAILLCLGVALTFIFCLRKGRMMDVKKCGIQDTNSK
KQSDTHLEET
>gi|73946918|ref|XP_541302.2| PREDICTED: similar to CD274 antigen [Canis familiaris]
MRMFSVFTFMAYCHLLKAFTITVSKDLYVVEYGGNVTMECKFPVEKQLNLFALIVYWEMEDKKIIQFVNG
KEDLKVQHSSYSQRAQLLKDQLFLGKAALQITDVRLQDAGVYCCLIGYGGADYKRITLKVHAPYRNISQR
ISVDPVTSEHELMCQAEGYPEAEVIWTSSDHRVLSGKTTITNSNREEKLFNVTSTLNINATANEIFYCTF
QRSGPEENNTAELVIPEVSHHSSRSLAGNFL
>gi|119900350|ref|XP_613366.3| PREDICTED: similar to programmed death ligand 1 [Bos taurus]
MECQAFTITVSKDLYVVEYGSNVTLECRFPVDKQLNLLVLVVYWEMEDKKIIQFVNGKEDPNVQHSSYHG
RAQLLKDQLFLGKAALQITDVKLQDAGVYCCLISYGGADYKRITLKVNAPYRKIYHTISVDPVTSEHELT
CQAEGYPEADVIWTSSDHQVLSGKTSITSSKREEKLFNVTSTLRINTTADKIFYCTFRRLGHEENNTAEL
VIPEPYLDPAKKRNHLVTLGALFLCLSVTLAVIFCLKRDVRMMDVEKCDTRDMNSKQQNDQRYAVGQGAA
DDGELKKPKLRKQKLRKKRRTKEEGIKVPWKETLVLPNGGRLINTCEKEKGHF
>gi|11230798|ref|NP_068693.1| CD274 antigen [Mus musculus]
MRIFAGIIFTACCHLLRAFTITAPKDLYVVEYGSNVTMECRFPVERELDLLALVVYWEKEDEQVIQFVAG
EEDLKPQHSNFRGRASLPKDQLLKGNAALQITDVKLQDAGVYCCIISYGGADYKRITLKVNAPYRKINQR
ISVDPATSEHELICQAEGYPEAEVIWTNSDHQPVSGKRSVTTSRTEGMLLNVTSSLRVNATANDVFYCTF
WRSQPGQNHTAELIIPELPATHPPQNRTHWVLLGSILLFLIVVSTVLLFLRKQVRMLDVEKCGVEDTSSK
NRNDTQFEET
>gi|109460012|ref|XP_574652.2| PREDICTED: similar to CD274 antigen [Rattus norvegicus]
MRIFAVLIVTACSHVLAAFTITAPKDLYVVEYGSNVTMECRFPVEQKLDLLALVVYWEKEDKEVIQFVEG
EEDLKPQHSSFRGRAFLPKDQLLKGNAVLQITDVKLQDAGVYCCMISYGGADYKRITLKVNAPYRKINQR
ISMDPATSEHELMCQAEGYPEAEVIWTNSDHQSLSGETTVTTSQTEEKLLNVTSVLRVNATANDVFHCTF
WRVHSGENHTAELIIPELPVPRLPHNRTHWVLLGSVLLFLIVGFTVFFCLRKQVRMLDVEKCGFEDRNSK
NRNVRGDVSSVEPSEPRGGIGSLWSVKERARGTWVQGLKNGTGEEKRTKKVLEEEPGTKDISTGDTAKQV
THQSSRAIS
>gi|118103980|ref|XP_424811.2| PREDICTED: similar to B7-H1 [Gallus gallus]
MMEKLLLLHIFLFCWRSLNALFTVEAPKSLYTAELGSNVTMECVFPVNGKLKFRDLSVIWEKKDEVRKDV
YILLKGKEDSGSQHSDFQGRIKLLKENLDFGQSLLQISNVKLRDAGLYHCLIEYGGADYKTINLKVQAPY
RTITQEVVSTGDKEWKLTCQSEGYPKAEVMWQNGECQDLTDKANTSYETGSDQLYRVTSTLTVKNRTCEN
FRCIFWNKEIQENTSANLYILDSADDVLWTESRRFVWPVLIVSALVGSVPITVCIRKARASKDCRTRMAK
SSIHITKLSKDKGAHDCRGPSFEDAELKYIQIETT
```

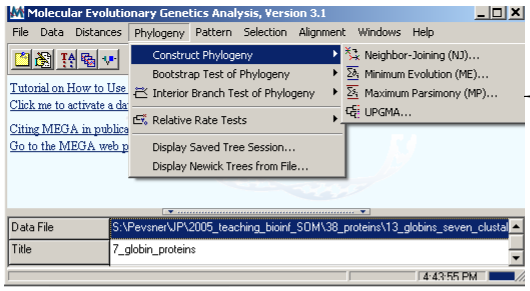
Use MEGA to make phylogenetic trees



Trees show the evolutionary relationships among proteins, or DNA sequences, or species…

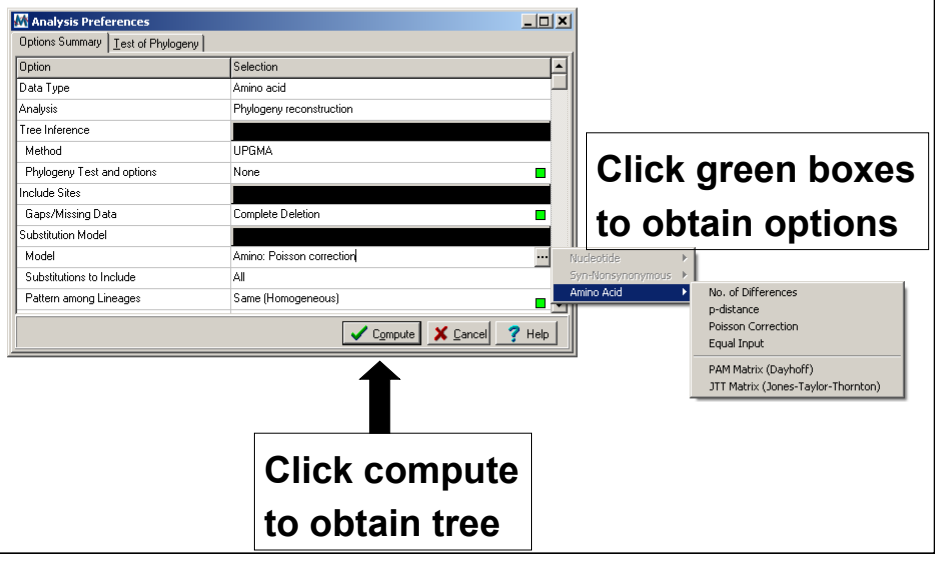
How to use MEGA to make a tree

1. Enter a multiple sequence alignment (.meg) file
2. Under the phylogeny menu, select one of these 4 methods…



Neighbor-Joining (NJ)
Minimum Evolution (ME)
Maximum Parsimony (MP)
UPGMA

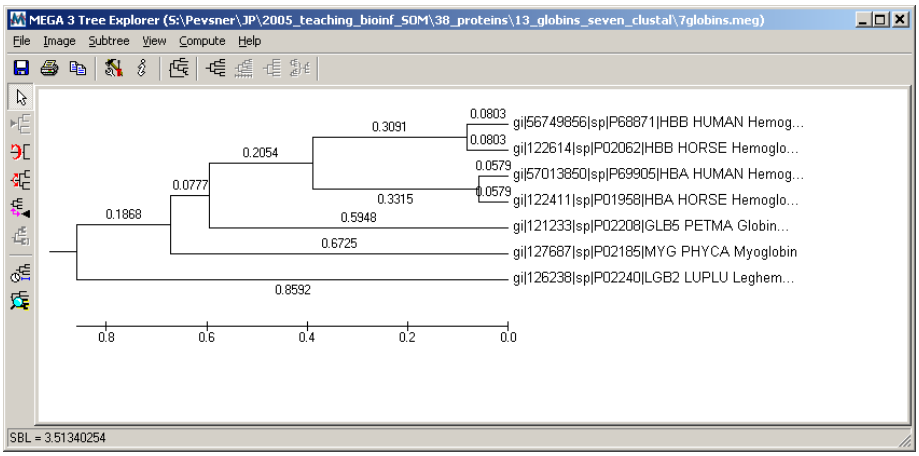
Use of MEGA for a distance-based tree: UPGMA



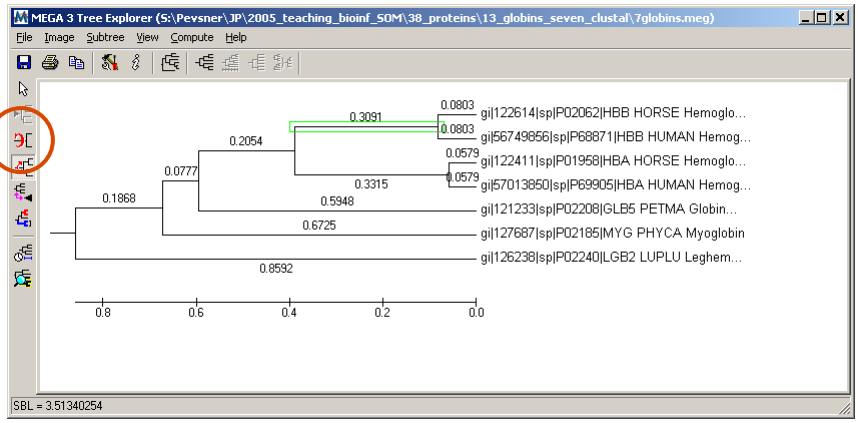
Click green boxes to obtain options

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.

Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching