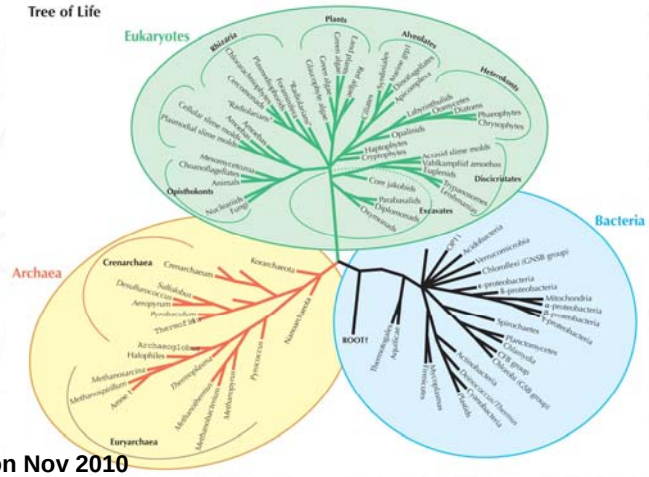
Molecular phylogeny

How to infer phylogenetic trees using molecular sequences



Tore Samuelsson Nov 2010

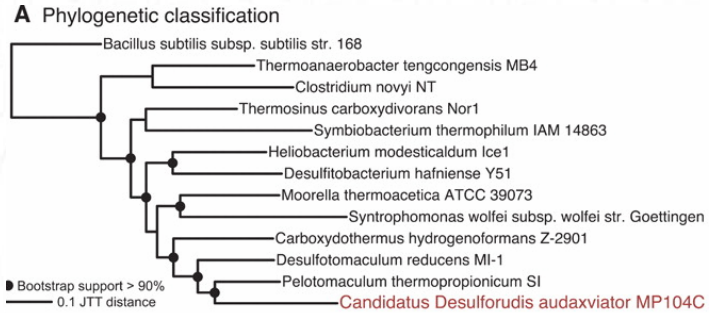
Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

Candidatus Desulforudis audaxviator

Life is Lonely at the Center of the Earth

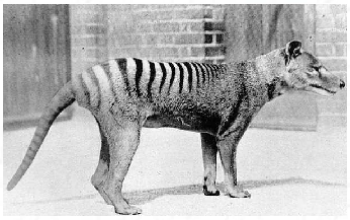
Environmental genomics reveals a single-species ecosystem deep within earth. Chivian et al. Science 2008.

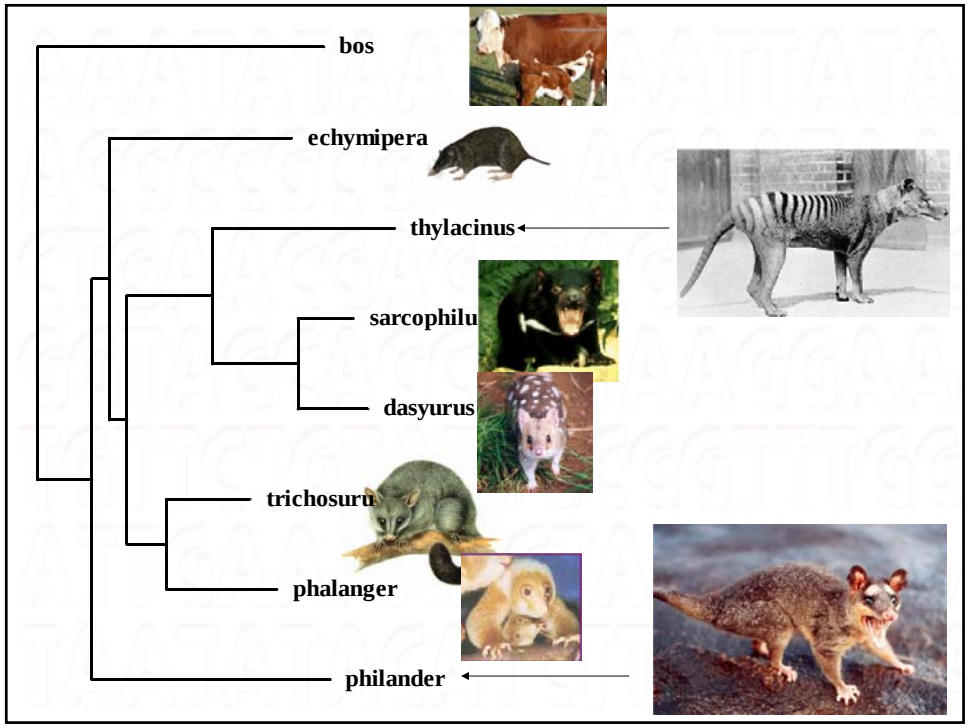


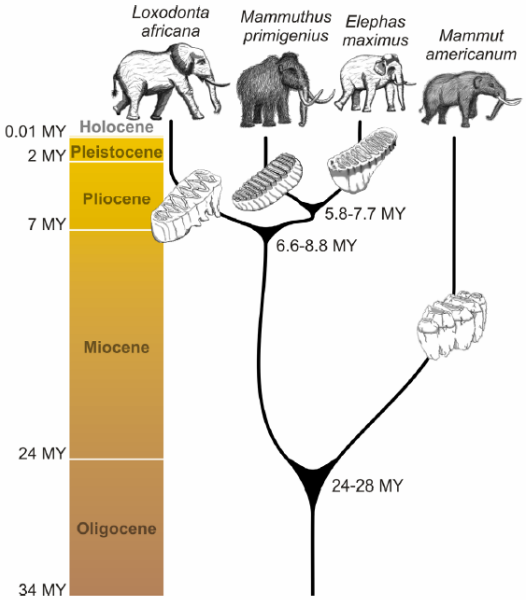
Molecular phylogeny to examine extinct species - I

Is the south american opossum

evolutionary related
to the australian ‘marsupial wolf’ ?







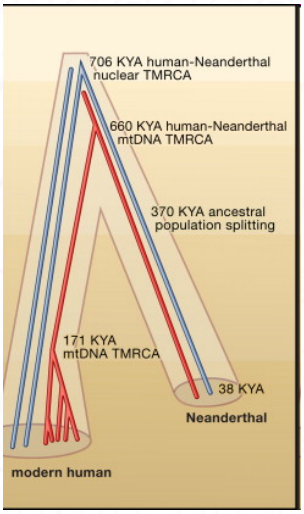
Molecular phylogeny to examine extinct species - II

"Sequencing the nuclear genome of the extinct woolly mammoth". Miller et al. Nature Nov. 2008

Molecular phylogeny to examine extinct species - III

Phylogeny of Neanderthal individuals

Svante Pääbo



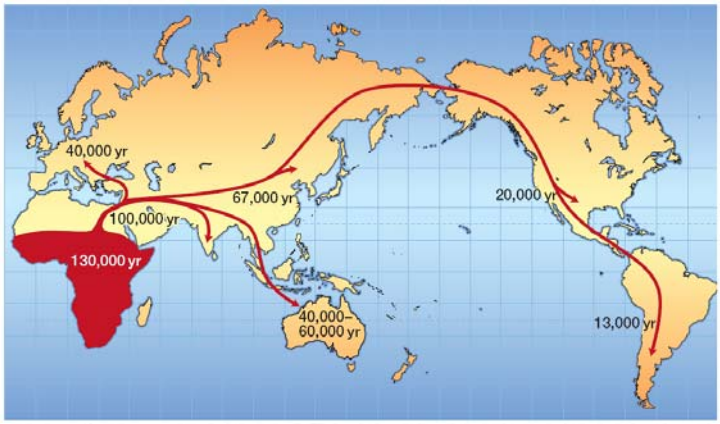
A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing

Richard E. Green,[1,*] Anna-Sapfo Malaspinas,[2] Johannes Krause,[1] Adrian W. Briggs,[1] Philip L.F. Johnson,[3] Caroline Uhler,[4] Matthias Meyer,[1] Jeffrey M. Good,[1] Tomislav Maricic,[1] Udo Stenzel,[1] Kay Prüfer,[1] Michael Siebauer,[1] Hernán A. Burbano,[1] Michael Ronan,[5] Jonathan M. Rothberg,[6] Michael Egholm,[5] Pavao Rudan,[7] Dejana Brajković,[8] Željko Kućan,[7] Ivan Gušić,[7] Mårten Wikström,[9] Liisa Laakkonen,[10] Janet Kelso,[1] Montgomery Slatkin,[2] and Svante Pääbo[1]
[1]Max-Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany

Mitochondrial genome variation and the origin of modern humans

Max Ingman*, Henrik Kaessmann†, Svante Pääbo† & Ulf Gyllensten*

** Department of Genetics and Pathology, Section of Medical Genetics, Rudbeck Laboratory, University of Uppsala, S-751 85 Uppsala, Sweden*
† Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany



"Out of Africa" hypothesis
Modern humans evolved from archaic forms only in Africa. Archaic humans living in Asia and Europe (like the Neanderthal) were replaced by modern humans migrating out of Africa.

Home assignment - Phylogeny of
Neanderthal - modern humans - monkeys

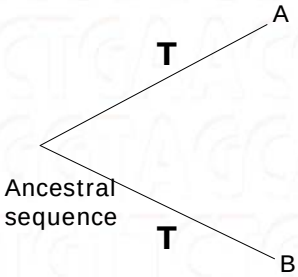
Starting point is multiple alignment of complete mitochondrial genomes from

* 8 modern humans of different origin, including 2 African sequences
* One Neanderthal genome (2008)
* Gorilla, Bonobo, Chimpanzee

* Relationship of Neanderthal to modern humans?
* Modern humans and "Out of Africa" hypothesis?
* What primate is most closely related to humans?

Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer



A molecular clock may be used in the estimation of *time of divergence* between two species

r = K / 2T or
T = K/2r

where

r = rate of nucleotide substitution (estimated from fossil records)

K = number of substitutions K between the two homologous sequences

T = ***Time of divergence between the two species***

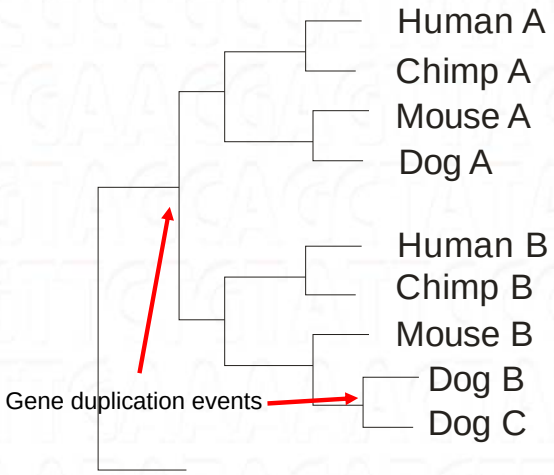
Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

Analysis of orthology and paralogy



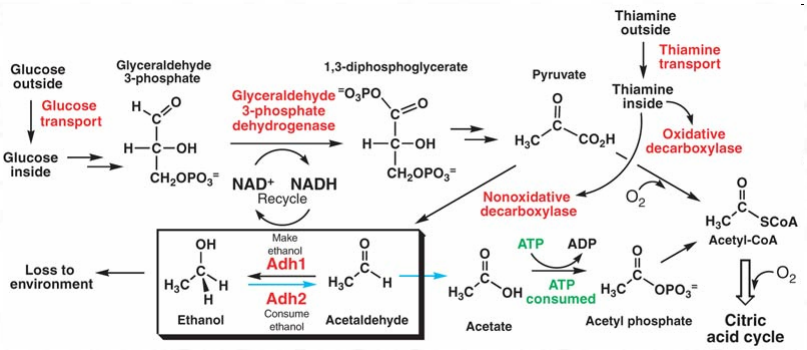
Compare Zvelebil & Baum p. 242

Applications of phylogenetic methods

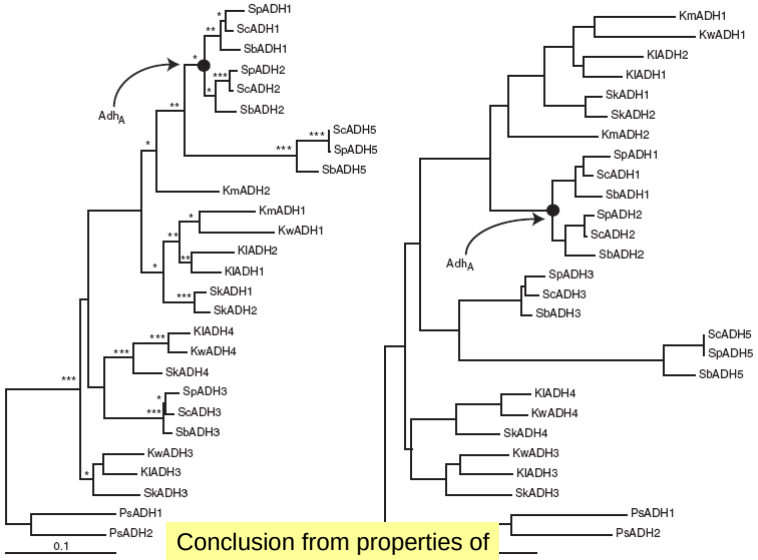
- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

Resurrecting ancestral alcohol dehydrogenases from yeast

J Michael Thomson[1,4], Eric A Gaucher[2], Michelle F Burgan[3,4], Danny W De Kee[2], Tang Li[2], John P Aris[1] & Steven A Benner[1,3]



ADH1 produces ethanol
ADH2 consumes ethanol

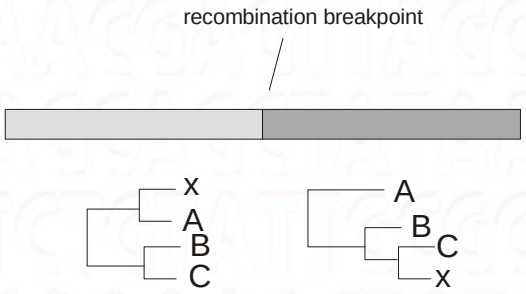


Conclusion from properties of ancestral protein:
ancestral Adh was mainly responsible for *making* ethanol

Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

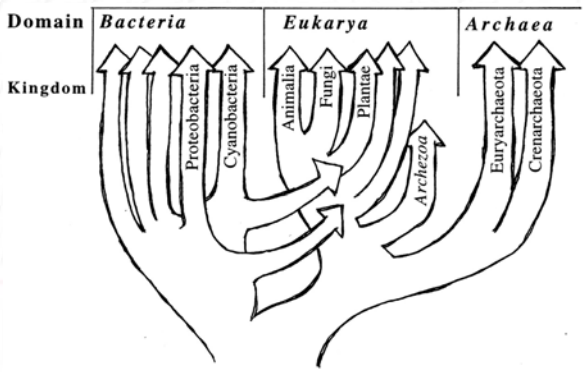
Detecting recombination break points - common in viral genomes



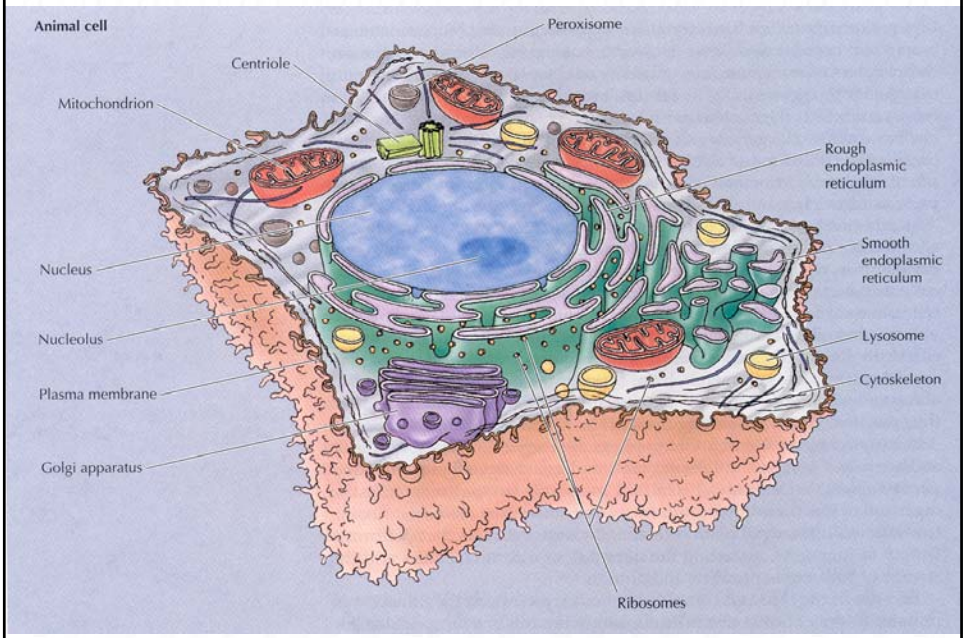
Applications of phylogenetic methods

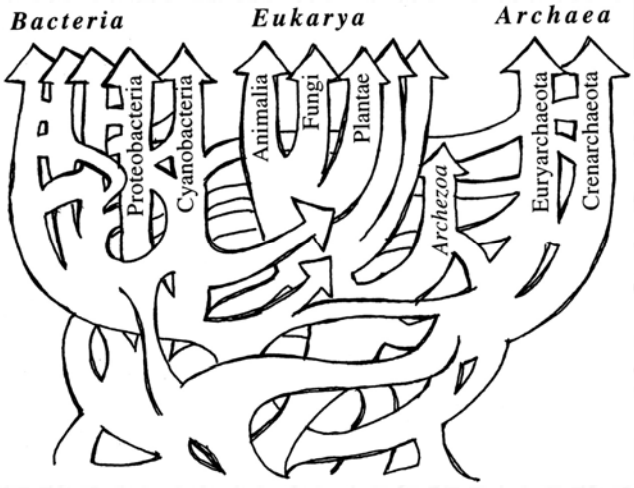
- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Determining the identity of new pathogens
- Detection of orthology and paralogy
- Reconstructing ancient proteins
- Detecting recombination break points
- Identification of horizontal gene transfer

Horizontal gene transfer - transfer of genes between species



Mitochondria and chloroplasts resulted from bacteria that lived in symbiosis with a primitive eukaryote. Eventually many genes were lost or transferred to the nuclear genome

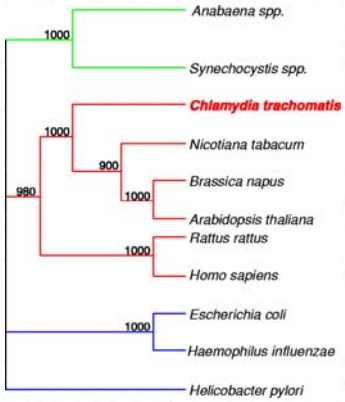




Many eubacterial genes have been transferred to archae and eukarya

Phylogenetic analysis may be used to identify horisontal gene transfer.

Some Chlamydia (Eubacteria kingdom) proteins group with plant homologs



Phylogeny of chlamydial enoyl-acyl carrier protein reductase as an example of horizontal transfer.

From: Stephens RS, et al Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science. 1998 Oct 23;282(5389):754-9.

Phylogenetic analysis

- Selection of sequences for analysis
- Multiple sequence alignment
- Construction of tree
- Evaluation of tree

Construction of the phylogenetic tree

Distance methods

Character methods
- Maximum parsimony
- Maximum likelihood

Distance methods

Simplest distance measure:

Consider every pair of sequences in the multiple alignment and count the number of differences.

Degree of divergence = Hamming distance (D)

$D = n/N$
where N = alignment length
n = number of sites with differences

Example:
AG**G**CTT**T**TCA
AG**C**CTT**C**TCA
D = 2/10 = 0.2

Character-based methods

* ***Maximum parsimony***
* Maximum likelihood

Maximum parsimony

parsimony - principle in science where the simplest answer is the preferred.

In phylogeny: The preferred phylogenetic tree is the one that requires the fewest evolutionary steps.

Maximum parsimony

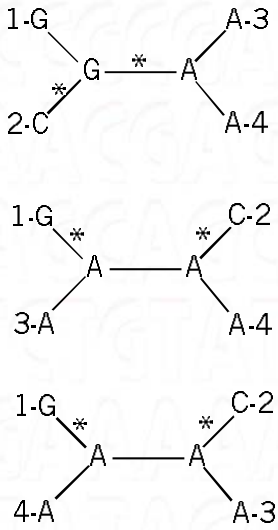
1. Identify all *informative sites* in the multiple alignment
2. For each possible tree, calculate the number of changes at each informative site.
3. Sum the number of changes for each possible tree.
4. Tree with the smallest number of changes is selected as the most likely tree.

Maximum parsimony

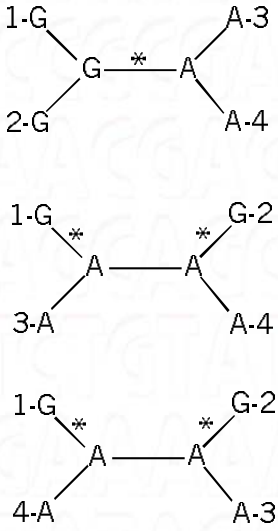
Identify informative sites

	Site								
	1	2	3	4	5	6	7	8	9
Sequence									
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

Site 3 - non - informative



Site 5 - informative



Summing changes:

	site 5	site 7	site 9	Sum
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III	2	2	2	6

⇒Tree I most likely.

(In this case we are not considering branch lengths, only topology of tree is predicted)

Character-based methods

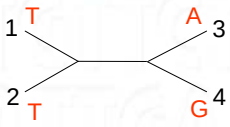
* Maximum parsimony
* ***Maximum likelihood***
  What is the probability that a particular tree generated the observed data under a specific model?

Maximum likelihood

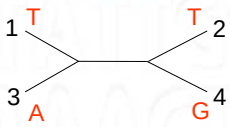
Consider the following multiple alignment

1	A	C	T	T
2	A	C	T	T
3	A	T	A	T
4	A	T	G	C

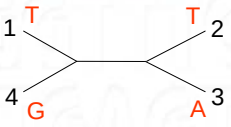
First , consider position 3 above (TTAG)
There are three possible unrooted trees for the OTUs 1-4:



Tree A



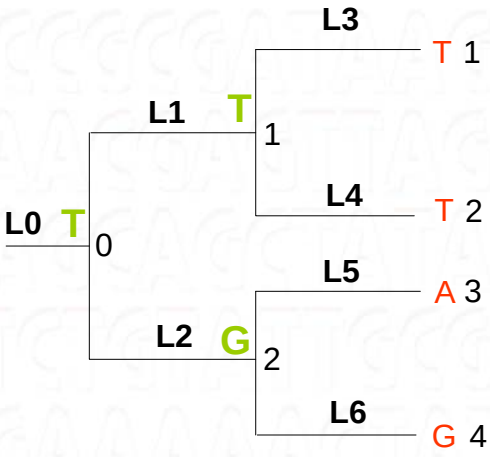
Tree B



Tree C

Maximum likelihood

A rooted version of Tree A:



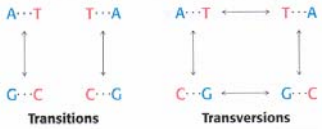
L(Tree1) = L0 * L1 * L2 * L3 * L4 * L5 * L6

Example of probability matrix for nucleotide substitutions

	A	C	T	G
A	~ 1	k	k	2k
C	k	~1	2k	k
T	k	2k	~1	k
G	2k	k	k	~1

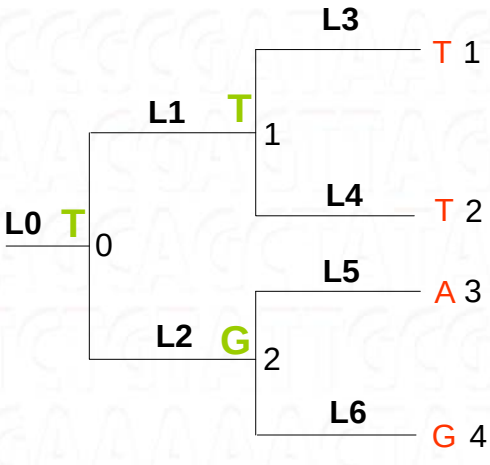
where we here set k = 1E-6.

Transitions are more likely than transversions



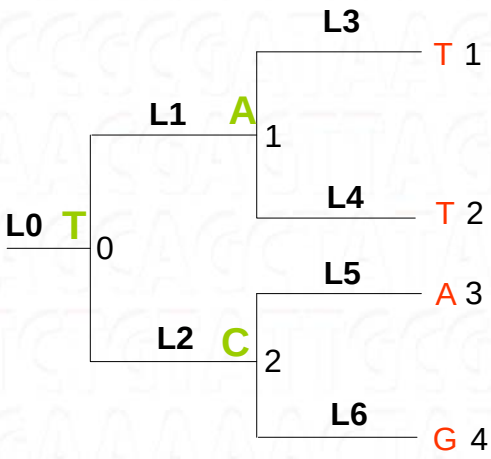
Maximum likelihood

A rooted version of Tree A:



L(Tree1) = L0 * L1 * L2 * L3 * L4 * L5 * L6 =
0.25 * 1 * 1E-6 * 1 * 1 * 2E-6 * 1 = 5E-13

A rooted version of Tree A:



L(Tree2) = L0 * L1 * L2 * L3 * L4 * L5 * L6 =
0.25 * 1E-6 * 2E-6 * 1E-6 * 1E-6 * 1E-6 * 1E-6 = 5E-37

L(Tree) = L(Tree1) + L(Tree2) + L(Tree3) .... L(Tree64)

Then we examine all positions of the alignment in the same way. Probability of tree is the product of probabilities for the different positions.

L = L(Tree pos1) * L(Tree pos2) * L(Tree pos3) * L(Tree pos4)

lnL = ln L(Tree pos1) + ln L(Tree pos2)
+ ln L(Tree pos3) +ln L(Tree pos4)

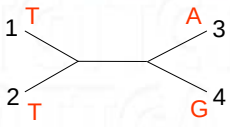
Finally , the Trees B and C are handled the same way. Tree with highest probability is preferred.

Maximum likelihood

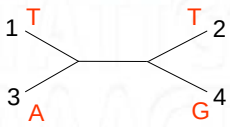
Consider the following multiple alignment

1	A	C	T	T
2	A	C	T	T
3	A	T	A	T
4	A	T	G	C

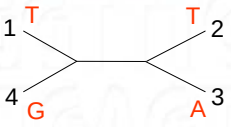
First , consider position 3 above (TTAG)
There are three possible unrooted trees for the OTUs 1-4:



Tree A



Tree B



Tree C

Phylogenetic analysis

- Selection of sequences for analysis
- Multiple sequence alignment
- Construction of tree
- ***Evaluation of tree***
  Bootstrapping

Evaluation of tree - Bootstrapping

(from www.icp.ucl.ac.be/~opperd/private/bootstrap.html)

Bootstrapping is a way of testing the reliability of the dataset and the tree, allows you to assess whether the distribution of characters has been influenced by stochastic effects.

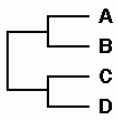
Bootstrapping in practice

Take a dataset consisting of in total n sequences with m sites each. A number of resampled datasets of the same size (n x m) as the original dataset is produced. However, each site is sampled at random and no more sites are sampled than there were original sites.

Sample 1

```
          0 1 2 0 3 0 1 2 0 1  (<- number of times each site is sampled)
          ---------------------

  A       A G G C U C C A A A       A     G G G U U U C A A A
  B       A G G U U C G A A A       B     G G G U U U G A A A
  C       A G C C C C G A A A       C     G C C C C C G A A A
  D       A U U U C C G A A C       D     U U U C C C G A A C
```

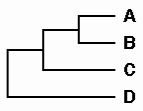
	A	B	C
B	1		
C	6	5	
D	8	7	4



Sample 2

```
          1 0 0 0 2 2 2 0 0 3
          ---------------------

  A       A G G C U C C A A A       A     A U U C C C C A A A
  B       A G G U U C G A A A       B     A U U C C G G A A A
  C       A G C C C C G A A A       C     A C C C C G G A A A
  D       A U U U C C G A A C       D     A C C C C G G C C C
```

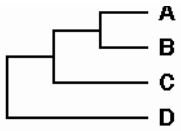
	A	B	C
B	2		
C	4	2	
D	7	5	3



Sample 3

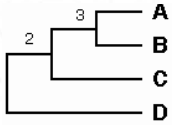
```
          2 2 3 0 0 0 1 0 0 2
          ___________________

A         A G G C U C C A A A       A     A A G G G G G C A A
B         A G G U U C G A A A       B     A A G G G G G G A A
C         A G C C C C G A A A       C     A A G G C C C G A A
D         A U U U C C G A A C       D     A A U U U U U G C C
```

	A	B	C
B	1		
C	4	3	
D	8	7	7



Consensus tree.

The number of times each branch point or node occurred (*bootstrap proportion*) is indicated at each node.



Bootstrapping typically involves 100-1000 datasets.

Bootstrap values > 70% are generally considered to provide support for the clade designation.

Software for phylogenetic analysis

PHYLIP (**Phyl**ogenetic **I**nference **P**ackage)
Joe Felsenstein
http://evolution.genetics.washington.edu/phylip.html

Examples in home assignment
- DNADIST = create a distance matrix
- NEIGHBOR = neighbor joining / UPGMA
- DNAPARS = maximum parsimony
- DNAML = maximum likelihood

PAUP (**P**hylogenetic **A**nalysis **U**sing **P**arsimony)

MrBayes