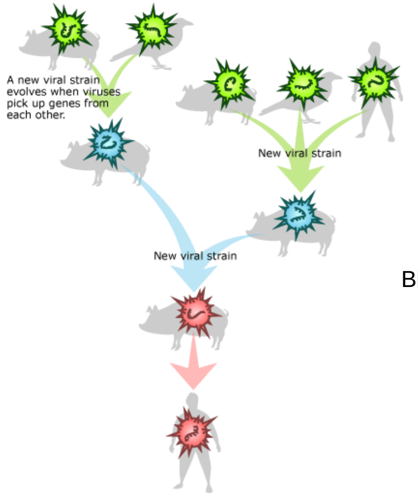
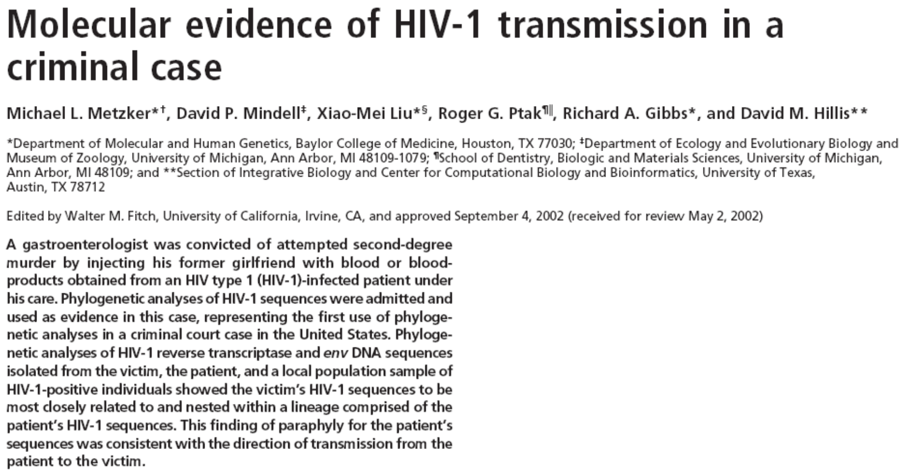
Molecular phylogeny - Using molecular sequences to infer evolutionary relationships

Tore Samuelsson Feb 2015

Molecular phylogeny is being used in the identification and characterization of new pathogens, like viruses and bacteria


A new viral strain evolves when viruses pick up genes from each other.
New viral strain
New viral strain

Bird flu virus evolution



Molecular evidence of HIV-1 transmission in a criminal case

Michael L. Metzker*[†], David P. Mindell[‡], Xiao-Mei Liu*[§], Roger G. Ptak[¶‖], Richard A. Gibbs*, and David M. Hillis**

*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; [‡]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079; [¶]School of Dentistry, Biologic and Materials Sciences, University of Michigan, Ann Arbor, MI 48109; and **Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

Edited by Walter M. Fitch, University of California, Irvine, CA, and approved September 4, 2002 (received for review May 2, 2002)

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States. Phylogenetic analyses of HIV-1 reverse transcriptase and *env* DNA sequences isolated from the victim, the patient, and a local population sample of HIV-1-positive individuals showed the victim's HIV-1 sequences to be most closely related to and nested within a lineage comprised of the patient's HIV-1 sequences. This finding of paraphyly for the patient's sequences was consistent with the direction of transmission from the patient to the victim.

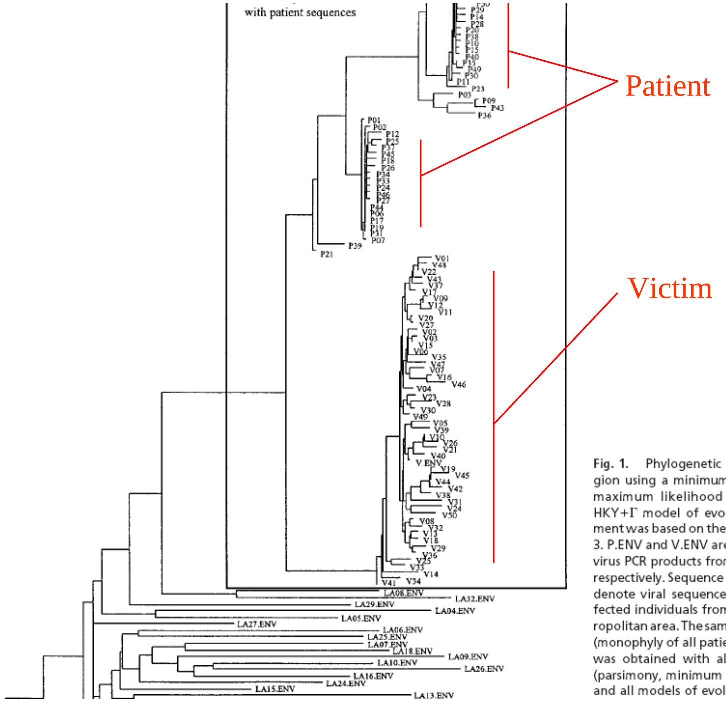
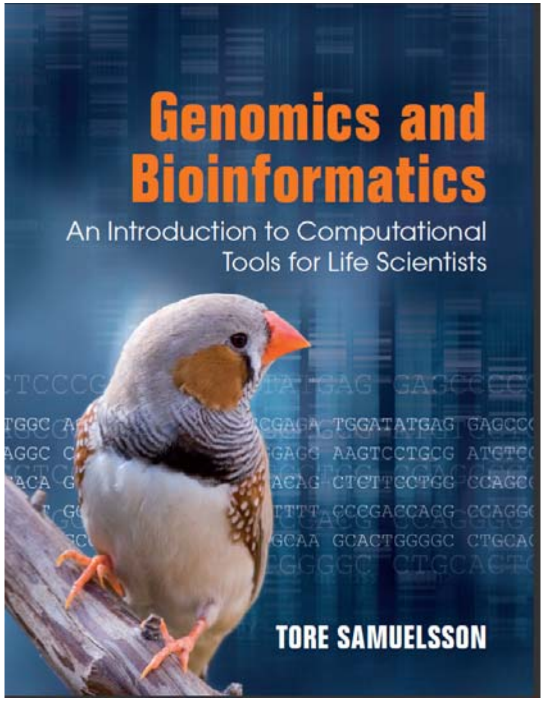


Fig. 1. Phylogenetic analysis of the gp120 region using a minimum evolution criterion and maximum likelihood distances assuming an HKY+Γ model of evolution. Nucleotide alignment was based on the protein alignment in Fig. 3. P.ENV and V.ENV are DNA sequences for provirus PCR products from the patient and victim, respectively. Sequence names beginning with LA denote viral sequences from control HIV-1 infected individuals from the Lafayette, LA, metropolitan area. The same pattern of relationships (monophyly of all patient and victim sequences) was obtained with all phylogenetic methods (parsimony, minimum evolution, and Bayesian) and all models of evolution examined. In addi-



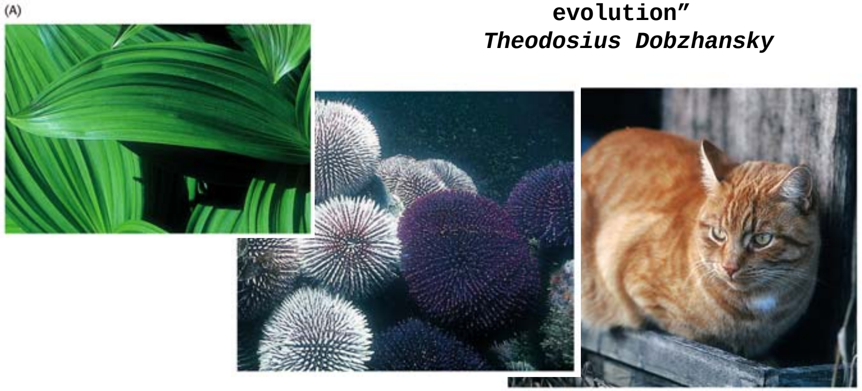
Genomics and Bioinformatics

An Introduction to Computational Tools for Life Scientists

TORE SAMUELSSON

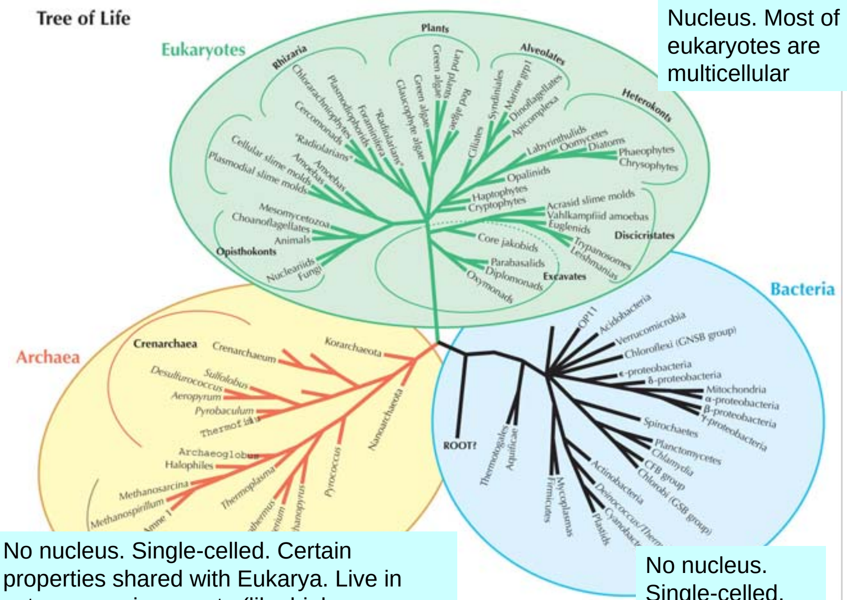
Evolution

"Nothing in biology makes sense
except in the light of
evolution"
Theodosius Dobzhansky



Organisms are remarkably uniform at the molecular level

This uniformity reveals that organisms on Earth have arisen from a common ancestor



Nucleus. Most of eukaryotes are multicellular

No nucleus. Single-celled. Certain properties shared with Eukarya. Live in extreme environments (like high temperature or high salt)

No nucleus. Single-celled. Ubiquitous

Principles of evolution

At the molecular level evolution is a process of mutation with selection

* Reproduction
* Variation
* Competition/selective pressure

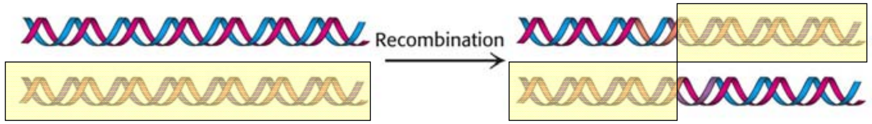
Mutations : changes in base sequence of DNA

1) single nucleotide change (point mutation)
transition (purine to purine or pyrimidine
to pyrimidine
C->T , T->C, A->G, G->A)
transversion (purine to pyrimidine
or py to pu
A->T, T->A, C->G, G->C etc

2) insertion / deletion of one or several
nucleotides

Such mutations are the result of
* ***Replication errors***
* ***Chemicals & irradiation***

Mutations : Homologous recombination cause large rearrangements in the genome



New gene families arise by
gene duplication and divergence

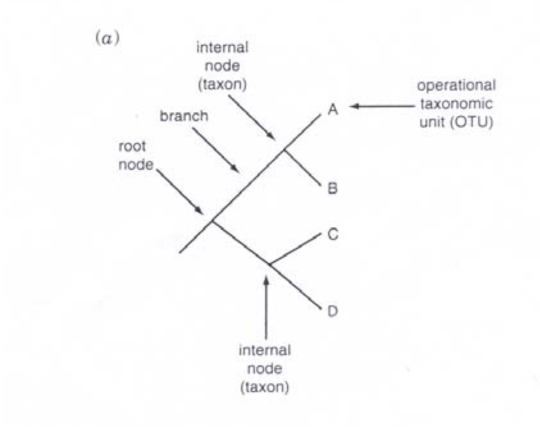
Molecular phylogeny

Phylogeny
Inference of evolutionary relationships

Molecular phylogeny
uses sequence information
(as opposed to other characteristics
frequently used in the past such as
morphological features)

Goals
* Deduce trees to show how
species/populations/invividuals/molecular sequences
are related

Nomenclature of trees



(More ancient) (More recent)

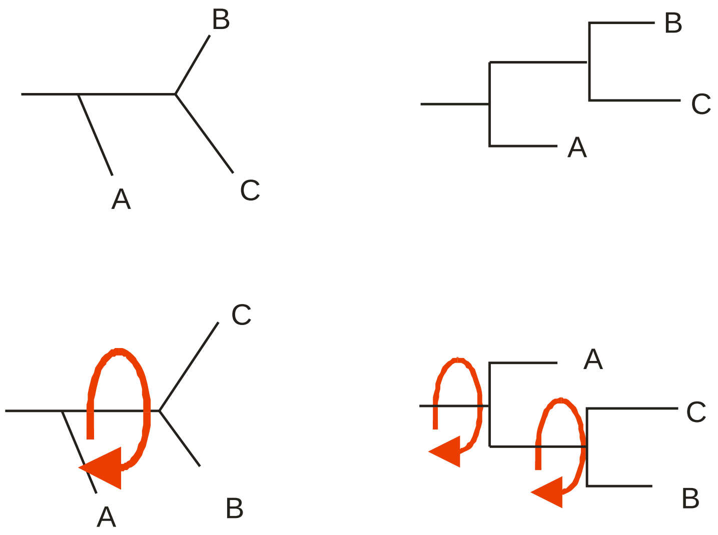
nodes
external (OTUs)
internal
root

branch
connects 2 nodes

OTUs are existing (observable) sequences / species / populations /individuals

an internal node is an inferred ancestor (not observed)

Different ways of showing the same tree



Nomenclature of trees

Rooted tree

Root - Common ancestor of all sequences in the tree

Unique path from the root to each of the other nodes

Direction of each path corresponds to evolutionary time

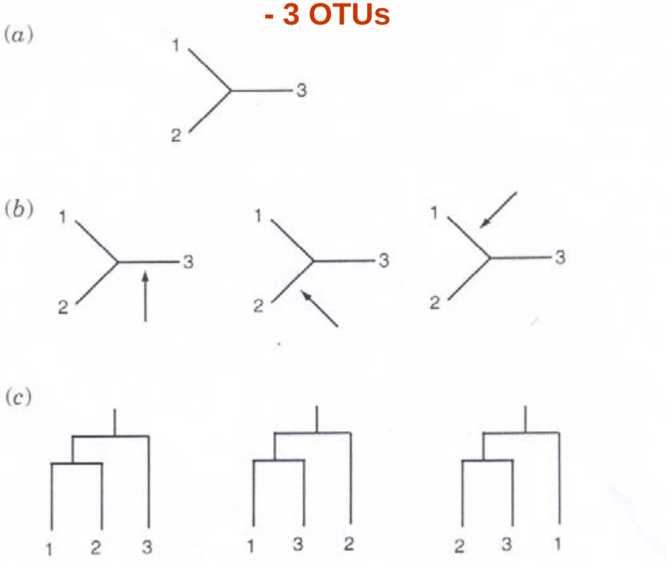
Unrooted tree

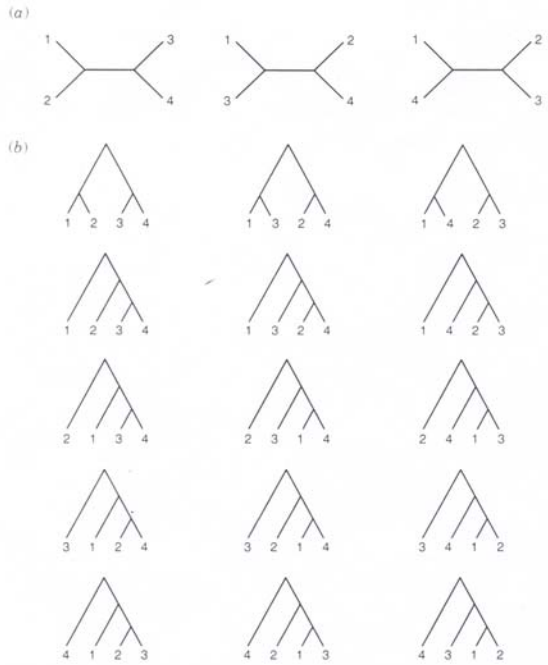
No root

No complete definition of evolutionary path

Direction of time is not determined

Comparing the numbers of rooted and unrooted trees - 3 OTUs





… and 4 OTUs

Goals of molecular phylogeny

Deduce the correct trees

* *Topology*
* *Branch lengths*

Phylogenetic analysis

- Selection of sequences for analysis

DNA?
RNA?
protein?

- Multiple sequence alignment

- Construction of tree

Slowly changing sequences

* Protein
* ribosomal RNA, for instance 16S rRNA

Useful for comparing widely divergent species.
Ribosomal RNA database (rdp.cme.msu.edu)
> 50,000 aligned sequences

More rapidly changing sequences

* DNA
* Mitochondrial DNA

Useful for comparing more closely related species or populations within a species.

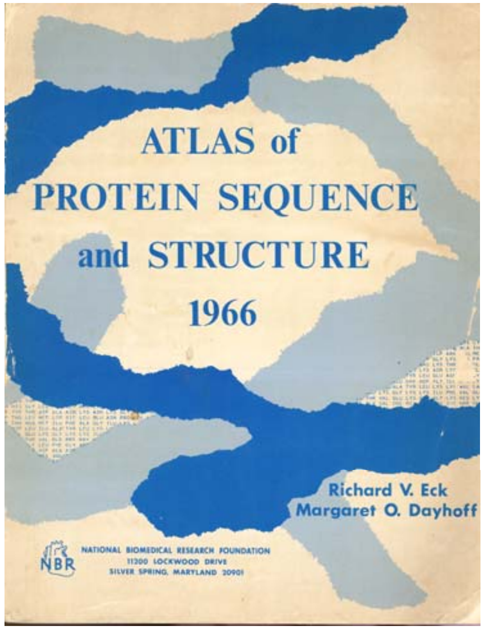
DNA sequences evolve more rapidly than protein sequences. This is to a large extent a result of the genetic code degeneracy.

```
Seq 1 GGC AAG CGA AGT
Seq 2 GGA AGA CGT TCA
Seq 1  G   R   R   S
Seq 2  G   K   R   S
```

Approximate rates of substitution
(number of substitutions per site & billion years)

rRNA	~ 0.1
protein	0.01 - 10
Hypervariable regions in mitochondria	10
HIV (RNA virus)	>1000

Early days of molecular phylogeny



Margaret Dayhoff

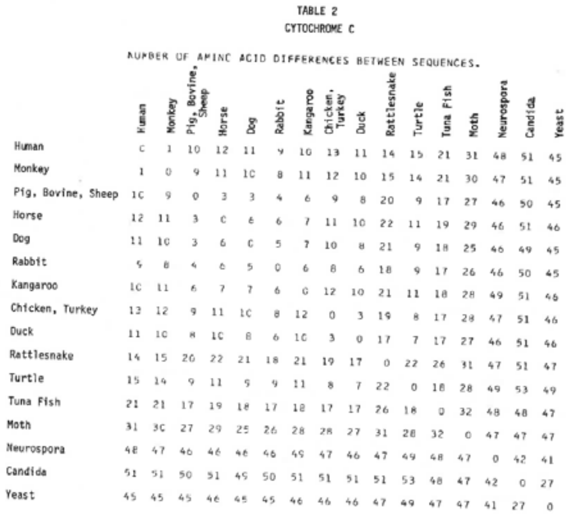
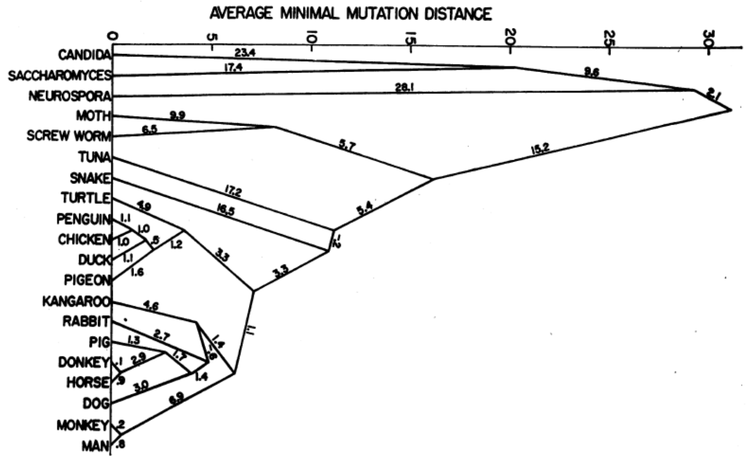


TABLE 2
CYTOCHROME C

NUPBER OF APINC ACID DIFFERENCES BETWEEN SEQUENCES.

	Human	Monkey	Pig, Bovine, Sheep	Horse	Dog	Rabbit	Kangaroo	Chicken, Turkey	Duck	Rattlesnake	Turtle	Tuna Fish	Moth	Neurospora	Candida	Yeast
Human	C	1	10	12	11	9	10	13	11	14	15	21	31	48	51	45
Monkey	1	0	9	11	1C	8	11	12	10	15	14	21	30	47	51	45
Pig, Bovine, Sheep	1C	9	0	3	3	4	6	9	8	20	9	17	27	46	50	45
Horse	12	11	3	C	6	6	7	11	10	22	11	19	29	46	51	46
Dog	11	1C	3	6	C	5	7	10	8	21	9	18	25	46	49	45
Rabbit	9	8	4	6	5	0	6	8	6	18	9	17	26	46	50	45
Kangaroo	1C	11	6	7	7	6	C	12	10	21	11	18	28	49	51	46
Chicken, Turkey	13	12	9	11	1C	8	12	0	3	19	8	17	28	47	51	46
Duck	11	1C	8	1C	8	6	1C	3	0	17	7	17	27	46	51	46
Rattlesnake	14	15	20	22	21	18	21	19	17	0	22	26	31	47	51	47
Turtle	15	14	9	11	9	9	11	8	7	22	0	18	28	49	53	49
Tuna Fish	21	21	17	19	18	17	18	17	17	26	18	0	32	48	48	47
Moth	31	3C	27	29	25	26	28	28	27	31	28	32	0	47	47	47
Neurospora	48	47	46	46	46	46	49	47	46	47	49	48	47	0	42	41
Candida	51	51	50	51	49	50	51	51	51	51	53	48	47	42	0	27
Yeast	45	45	45	46	45	45	46	46	46	47	49	47	47	41	27	0

Figure 4. Sequences of cytochrome c from 19 species. The amino acids common to all sequences and the allele groups at each position are shown. The sequences of the inferred common ancestors at each divergence point in the diagram are displayed below. Sites for which no single amino acid was most likely are left blank. The topology of the phylogenetic tree has been inferred from the sequences as explained in the text. The number of amino acid changes inferred between observed sequences and inferred ancestors are shown on the tree. The point of earliest time cannot be inferred directly from the sequences. We have placed it by assuming that on the average, species change at the same rate.



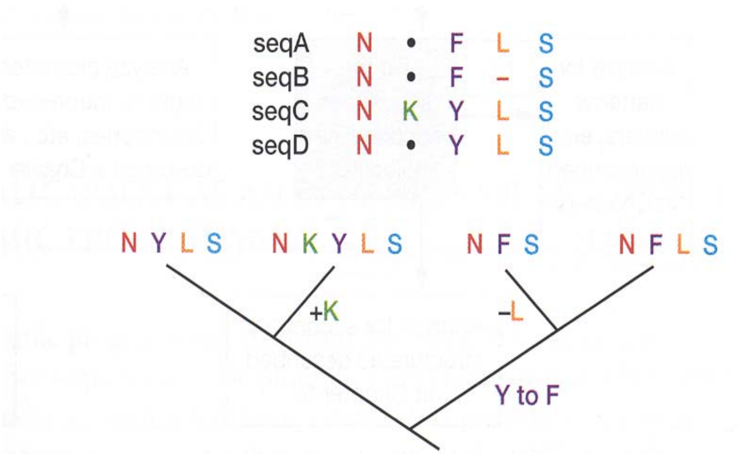
Phylogenetic analysis

- Selection of sequences for analysis

- ***Multiple sequence alignment***
  Alignment may be produced using methods such as CLUSTALW

- Construction of tree

Close relationship between multiple alignment and phylogenetic analysis



```
seqA   N  •  F  L  S
seqB   N  •  F  –  S
seqC   N  K  Y  L  S
seqD   N  •  Y  L  S
```

Inspecting the multiple alignment

Alignment should contain only homologous sequences. Overall identity should ideally be significant ensuring that the alignment is correct.

```
GGGCGGCGAGGCATTTATCGGGGGGTTGCAAAAT
GGGCGGTGAGGCATTTATCGGGGGGTTGCAAAAT
GGGCGGCGAAGCATAAATCGGGGAGTTGCAAAAT
GGGCGGCGAAGCATTTATCGGGGGGTTGCGAAAT
GGGCGGCGAGGCATTTATCGGGGGCTGCAAAAT
```

Phylogenetic analysis

- Selection of sequences for analysis

- Multiple sequence alignment

- ***Construction of 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:

AGGCTTTTCA
AGCCTTCTCA

D = 2/10 = 0.2

Generating a distance matrix

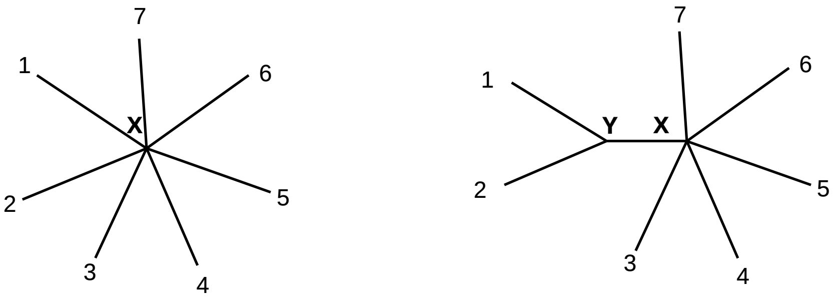
```
>A
GGACCACTACGAGCGCCTACGACGTA
>B
GGACCCCTACGAGCCCCTACGACGTA
>C
GGACCGCTGCGAGCTTCTACGACGTA
>D
GGACCTCTCCGGGCAGCTAGGACGTA
```

	A	B	C	D
A	0	2	4	6
B	2	0	4	6
C	4	4	0	6
D	6	6	6	0

OR

		B	C	D
A	-	2	4	6
B	-	-	4	6
C	-	-	-	6

Distance methods - Neighbor joining (1987)



Uses ***star decomposition*** method

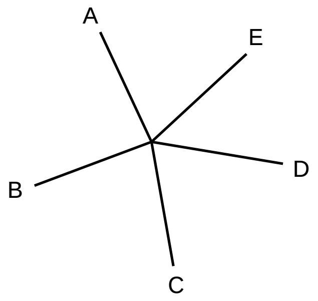
Neighbors: pair of nodes separated by one single node

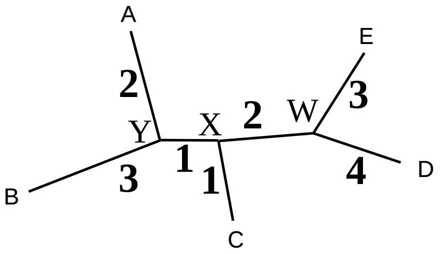
Minimal evolution: minimizing total branch length.

Generates ***unrooted tree***

Advantage: computationally ***fast***

	B	C	D	E
A	5	4	9	8
B		5	10	9
C			7	6
D				7





Character-based methods

- ***Maximum parsimony***
- Maximum likelihood
- Bayesian statistics

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.

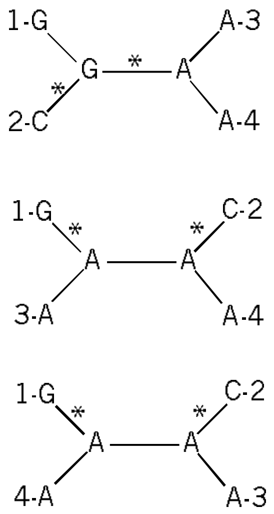
4. Tree with the smallest number of changes is selected as the most likely tree.

Maximum parsimony

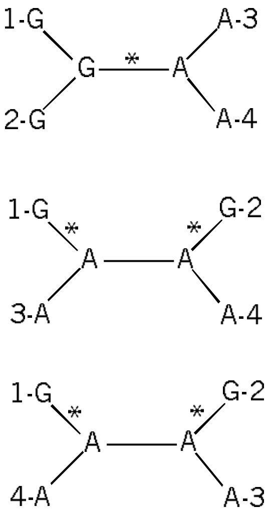
Identify informative sites

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

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)

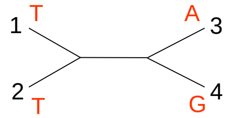
Character-based methods

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

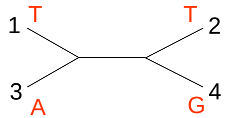
Consider the following multiple alignment

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

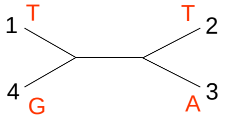
First , consider position 3 above (TTAG)
There are three possible unrooted trees for the OTUs 1-4:



Tree A

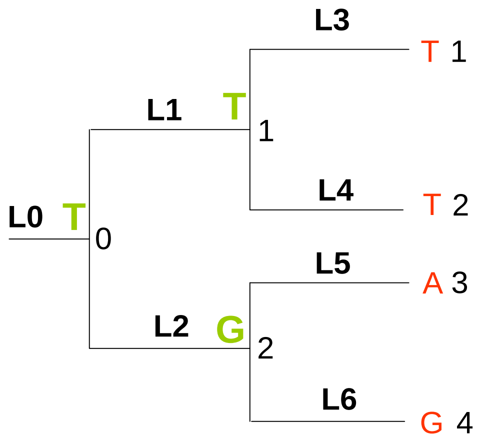


Tree B



Tree C

A rooted version of Tree A:



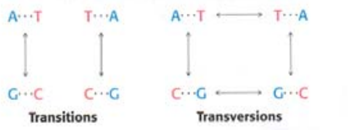
L(Tree1) = L0 * L1 * L2 * L3 * L4 * L5 * L6

Example of probability matrix for nucleotide substitutions

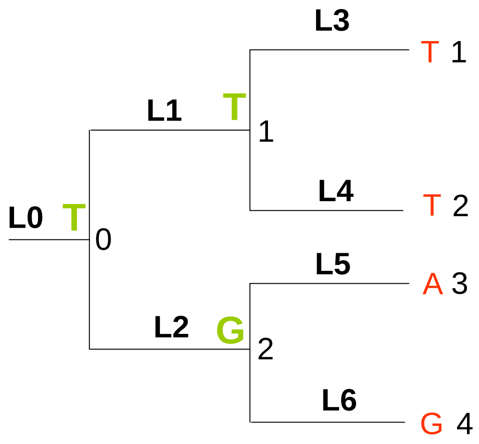
	A	C	T	G
A	~ 1	k	k	2k
C	k	~1	2k	k
T	k	2k	~1	k
G	2k	k	k	~1

where we here set k = 1E-6.

Transitions are more likely than transversions

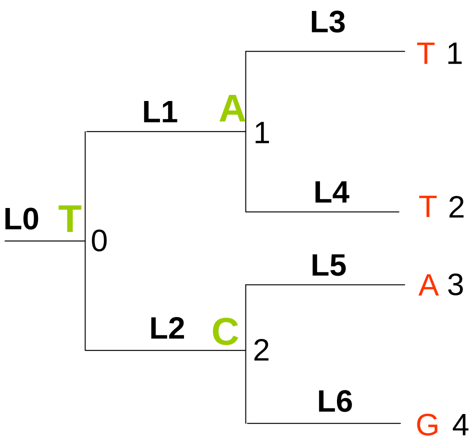


A rooted version of Tree A:



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)

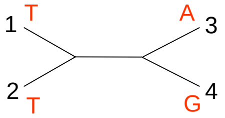
lnL = ln L(Tree pos1) + ln L(Tree pos2)
+ ln L(Tree pos3) +ln L(Tree pos4)

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

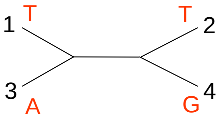
Consider the following multiple alignment

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

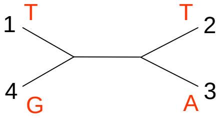
First , consider position 3 above (TTAG)
There are three possible unrooted trees for the OTUs 1-4:



Tree A



Tree B



Tree C

Character-based methods

- * Maximum parsimony
- * *Maximum likelihood*
  What is the probability of the data given the model?
- * ***Bayesian statistics***
  What is the probability of the model given the data?

Software for phylogenetic analysis

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

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

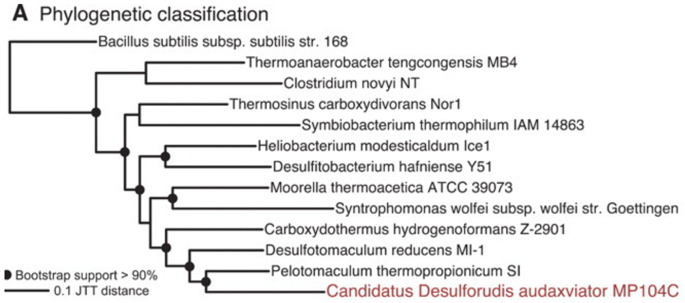
Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Identification of gene duplication events
- Reconstructing ancient proteins
- Identification of horizontal gene transfer

Candidatus Desulforudis audaxviator

Life is Lonely at the Center of the Earth

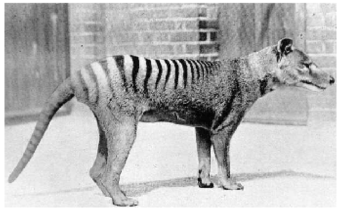
Environmental genomics reveals a single-species ecosystem deep within earth. Chivian et al. Science 2008.

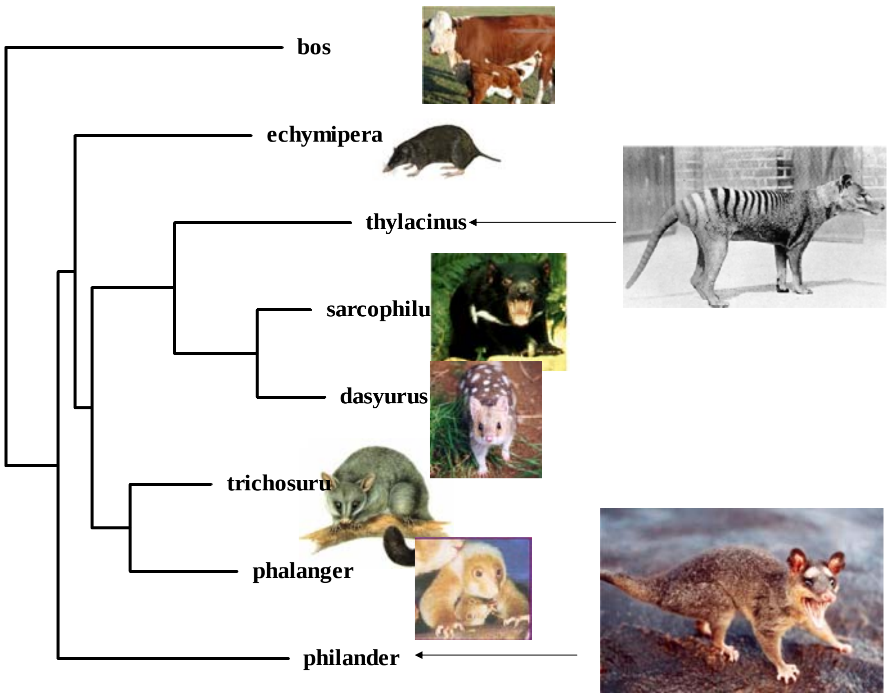

A Phylogenetic classification
Bacillus subtilis subsp. subtilis str. 168
Thermoanaerobacter tengcongensis MB4
Clostridium novyi NT
Thermosinus carboxydivorans Nor1
Symbiobacterium thermophilum IAM 14863
Heliobacterium modesticaldum Ice1
Desulfitobacterium hafniense Y51
Moorella thermoacetica ATCC 39073
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Carboxydothermus hydrogenoformans Z-2901
Desulfotomaculum reducens MI-1
Pelotomaculum thermopropionicum SI
Candidatus Desulforudis audaxviator MP104C
● Bootstrap support > 90%
0.1 JTT distance

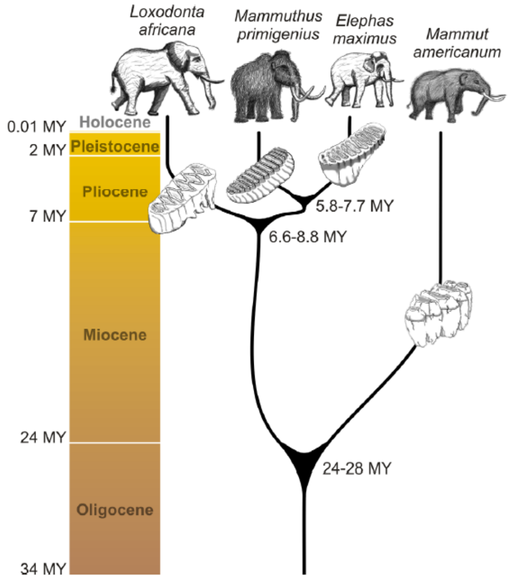
Molecular phylogeny to examine extinct species - I

Is the south american opossum

evolutionary related to the australian 'marsupial wolf' ?





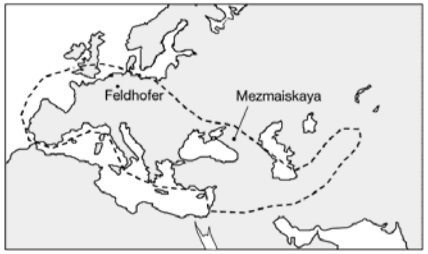


Molecular phylogeny to examine extinct species - II

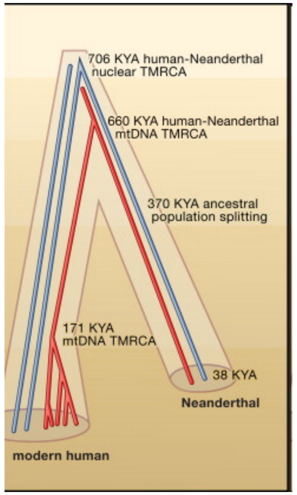
"Sequencing the nuclear genome of the extinct woolly mammoth". Miller et al. Nature Nov. 2008

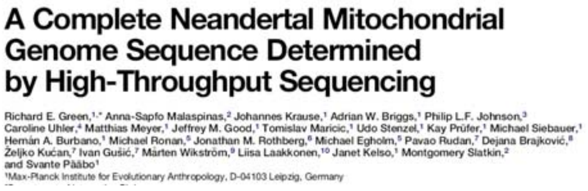
Molecular phylogeny to examine extinct species - III

Phylogeny of Neanderthal individuals



Svante Pääbo

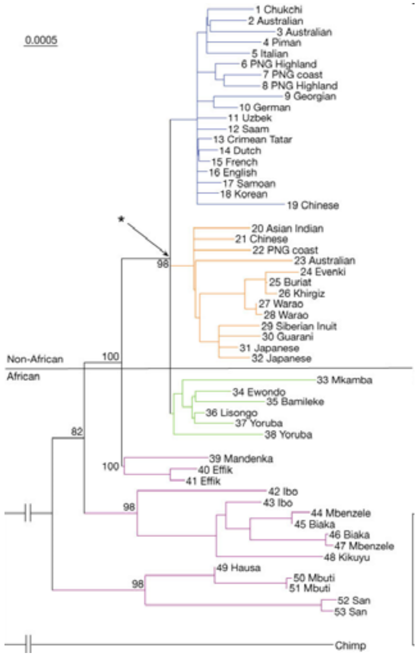

706 KYA human-Neanderthal nuclear TMRCA
660 KYA human-Neanderthal mtDNA TMRCA
370 KYA ancestral population splitting
171 KYA mtDNA TMRCA
38 KYA
Neanderthal
modern human

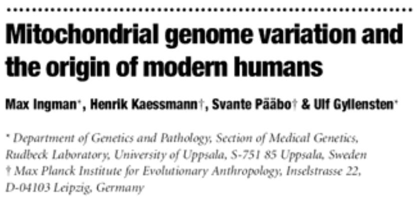


A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing

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

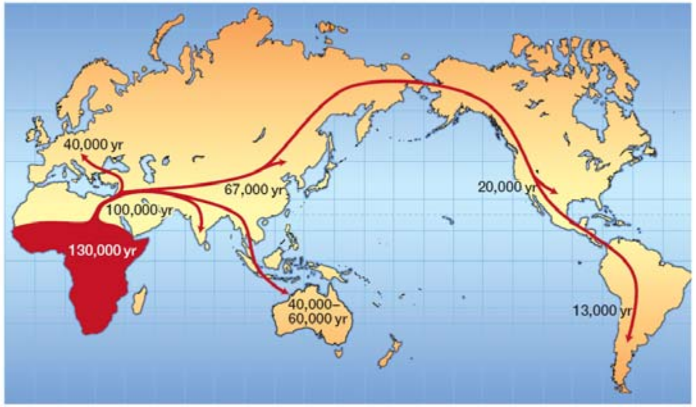




Mitochondrial genome variation and the origin of modern humans

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

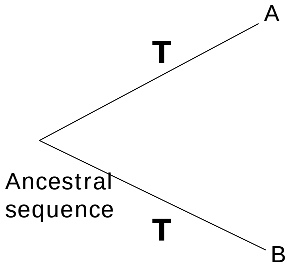


"Out of Africa" hypothesis

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.

Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Identification of gene duplication events
- Reconstructing ancient proteins
- 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)

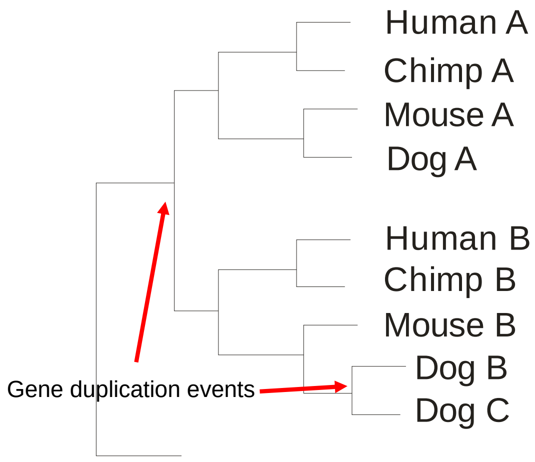
K = number of substitutions K between the two homologous sequences

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

Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Identification of gene duplication events
- Reconstructing ancient proteins
- Identification of horizontal gene transfer

Analysis of gene and protein evolution



Applications of phylogenetic methods

- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Identification of gene duplication events
- Reconstructing ancient proteins
- Identification of horizontal gene transfer

Resurrecting ancestral proteins responsible for ethanol digestion

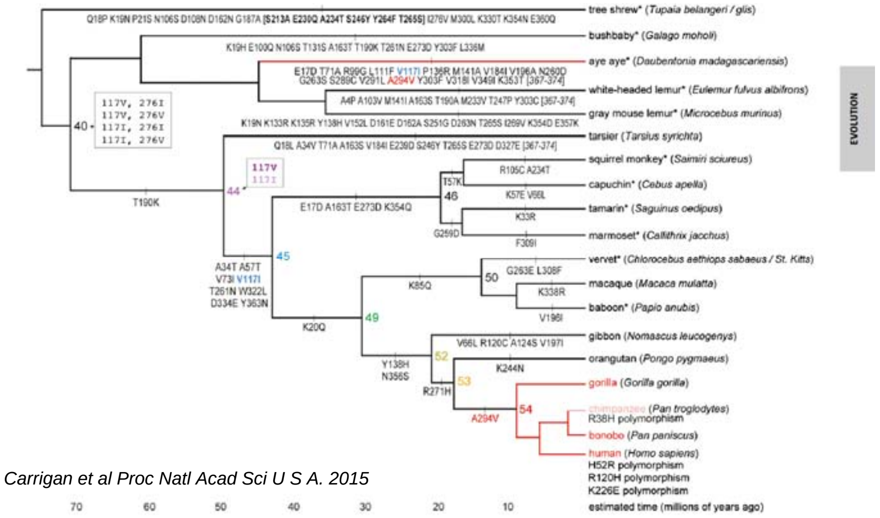
alcohol dehydrogenase

Ethanol => acetaldehyde

Modern humans can metabolize ethanol.
Many monkeys such as gibbon and orangutang cannot.
Was the ability to metabolize ethanol developed when humans started intentional fermentation of food?

Resurrecting ancestral proteins responsible for ethanol digestion

Ancestral sequences are inferred from present sequences and proteins are then produced in the lab to examine their properties.


tree shrew* (Tupaia belangeri / glis)
Q18P K19N P21S N106S D108N D162N G187A [S213A E230Q A234T S246Y Y264F T265S] I276V M300L K330T K354N E360Q
bushbaby* (Galago moholi)
K19H E100Q N106S T131S A163T T190K T261N E273D Y303F L336M
aye aye* (Daubentonia madagascariensis)
E17D T71A R99G L111F V117I P136R M141A V184I V196A N260D
G263S S289C V291L A294V Y303F V318I V349I K353T [367-374]
white-headed lemur* (Eulemur fulvus albifrons)
A4P A103V M141I A163S T190A M233V T247P Y303C [367-374]
gray mouse lemur* (Microcebus murinus)
K19N K133R K135R Y138H V152L D161E D162A S251G D263N T265S I269V K354D E357K
tarsier (Tarsius syrichta)
Q18L A34V T71A A163S V184I E239D S246Y T265S E273D D327E [367-374]
40
117V, 276I
117V, 276V
117I, 276I
117I, 276V
T190K
44
117V
117I
squirrel monkey* (Saimiri sciureus)
R105C A234T
capuchin* (Cebus apella)
T57K
46
K57E V66L
tamarin* (Saguinus oedipus)
E17D A163T E273D K354Q
K33R
G259D
marmoset* (Callithrix jacchus)
F309I
45
A34T A57T
V73I V117I
T261N W322L
D334E Y363N
vervet* (Chlorocebus aethiops sabaeus / St. Kitts)
50
G263E L308F
K85Q
macaque (Macaca mulatta)
K338R
baboon* (Papio anubis)
V196I
K20Q
49
gibbon (Nomascus leucogenys)
V66L R120C A124S V197I
52
orangutan (Pongo pygmaeus)
Y138H
N356S
K244N
53
R271H
gorilla (Gorilla gorilla)
54
A294V
chimpanzee (Pan troglodytes)
R38H polymorphism
bonobo (Pan paniscus)
human (Homo sapiens)
H52R polymorphism
R120H polymorphism
K226E polymorphism
70
60
50
40
30
20
10
estimated time (millions of years ago)
EVOLUTION

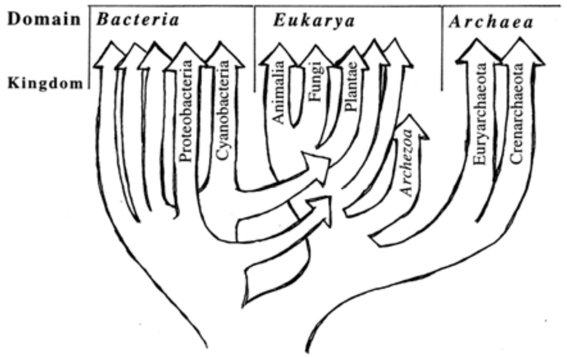
Carrigan et al Proc Natl Acad Sci U S A. 2015

Conclusion: The ability to digest ethanol appeared ~10 million years ago

Applications of phylogenetic methods

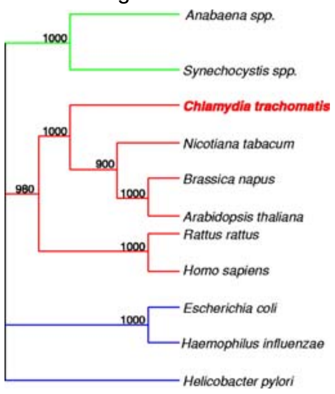
- Reconstruction of evolutionary history / Resolving taxonomy issues
- Estimating divergence times
- Identification of gene duplication events
- Reconstructing ancient proteins
- Identification of horizontal gene transfer

Horizontal gene transfer - transfer of genes between species



Phylogenetic analysis may be used to identify horisontal gene transfer.

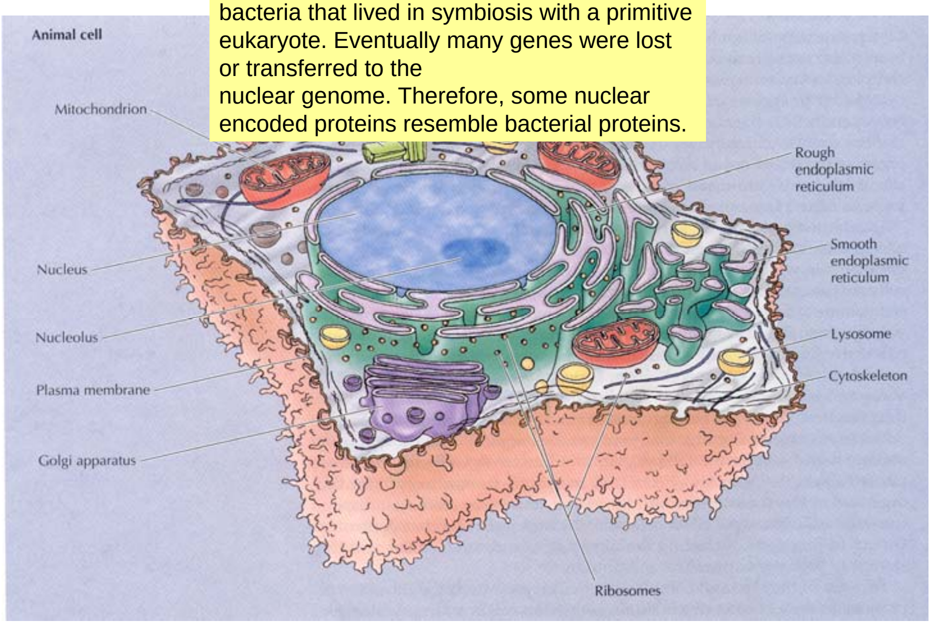
Some Chlamydia (Eubacteria kingdom) proteins group with plant homologs



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.

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. Therefore, some nuclear encoded proteins resemble bacterial proteins.


Animal cell
Mitochondrion
Rough endoplasmic reticulum
Smooth endoplasmic reticulum
Nucleus
Nucleolus
Lysosome
Cytoskeleton
Plasma membrane
Golgi apparatus
Ribosomes