Phylogeny inference

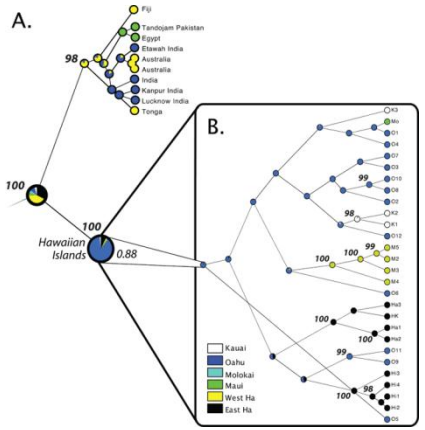
- Studying evolutionary relatedness among various groups of organisms (species, populations), through molecular sequence data (and also through morphological data).

- The course *Phylogenetic data analysis* (period IV, 2010) is an in-depth course on this topic

- The link: http://evolution.genetics.washington.edu/phylip/software.html shows that this field of science is an exceptionally popular one, 385 software packages at the moment.

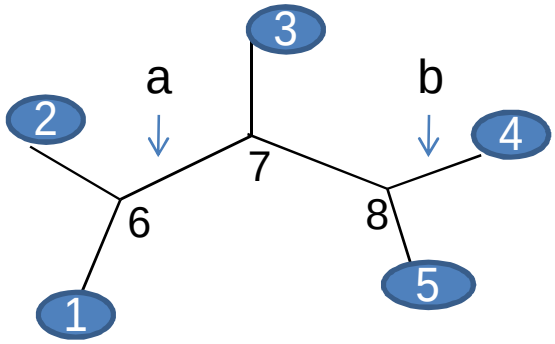
- The most widely used are PHYLIP, PAUP, MEGA, MrBAYES

- This lecture starts by a demo with MEGA4 (http://www.megasoftware.net/) and exercise session 5 relates to getting started with (easy) phylogeny reconstructions

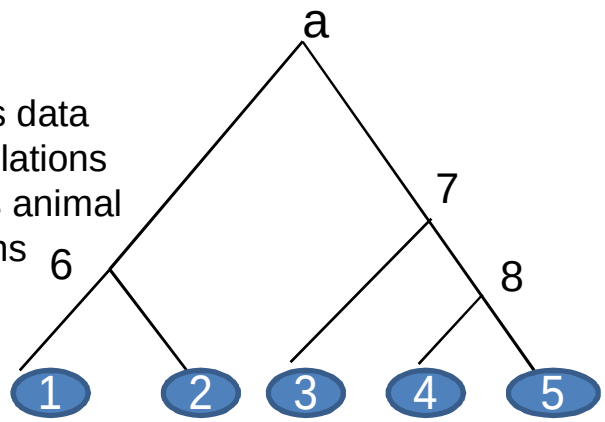


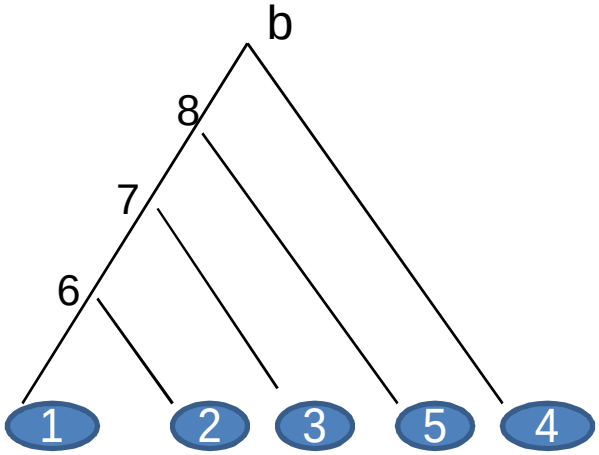
Basic concepts and terms

- Leaves, external nodes 1,2,3,4,5 are observations which may be, depending on the situation, sequences from different species, populations etc. They are often called OTUs = Operational Taxonomic Units. Internal nodes 6,7,8 are hypothetical sequences in ancestral units
- The tree is unrooted.



- In case evidence exists for depicting the root (for example, a or b), a rooted three can be constructed.





- For example, is there is data from different human populations and from chimpanzee, this animal is an outgroup and a means for rooting a tree
- Rooting requires external evidence and cannot be done on the basis of the data which is under a given study.

Number of possible rooted and unrooted trees

The number of unrooted trees

$b_n = (2(n-1)-3)b_{n-1} = (2n-5)b_{n-1} = (2n-5) * (2n-7) * \ldots * 3 * 1 = (2n-5)! / ((n-3)!2^{n-3}), n > 2$

Number of rooted trees b'_n is

$b'_n = (2n-3)b_n = (2n-3)! / ((n-2)!2^{n-2}), \quad n > 2$

that is, the number of unrooted trees times the number of branches in the trees

n	B_n	b'_n
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
20	2.22E+020	8.20E+021
30	8.69E+036	4.95E+038

Maximum parsimony

Entitia non sunt multiplicanda praeter necessitatem

William of Ockham (1280-1350)

- Discrete character states, the shortest pathway leading to these is chosen as the best tree. A model-free method.

- Parsimony, Occam´s razor, a philosophical concept, “the best hypothesis is the one requiring the smallest number of assumptions”.

- Parsimony analyses have been used from the early 1970s.

- The principle of maximum parsimony: identification of a tree topology/topologies that require the smallest number of evolutionary changes, i.e. transformations of one character state into another, to explain the differences among OTUs (operational taxonomic units).

- The goal of minimizing evolutionary change is often defended on philosophical grounds: when two hypotheses provide equally valid explanations for a phenomenon, the simpler one should always be preferred.

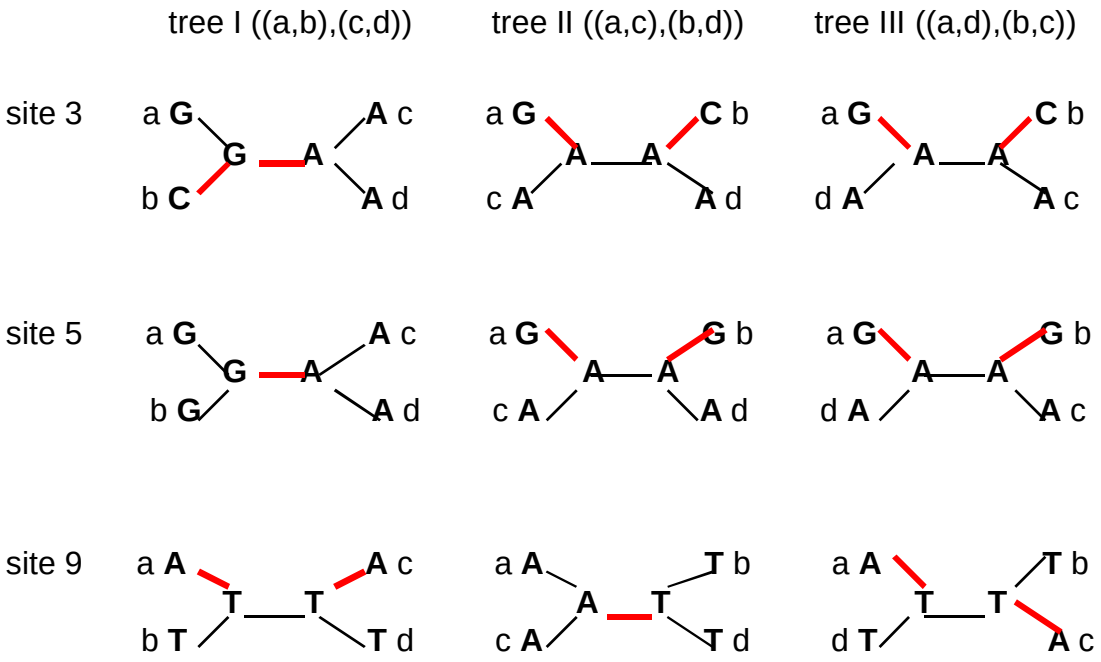
- Two subproblems:

(1) determining the amount of character change, or tree length, required by any given tree
(2) searching over all possible tree topologies for the trees that minimize this length.

Informative and uninformative sites

	1	2	3	4	5	6	7	8	9
OTU a	**A**	**A**	**G**	**A**	**G**	**T**	**T**	**C**	**A**
OTU b	**A**	**G**	**C**	**C**	**G**	**T**	**T**	**C**	**T**
OTU c	**A**	**G**	**A**	**T**	**A**	**T**	**C**	**C**	**A**
OTU d	**A**	**G**	**A**	**G**	**A**	**T**	**C**	**C**	**T**
					+		+		+

Four OTUs, three possible unrooted trees: I, II, III



A site is informative only when there are at least two different kinds of nucleotides at the site, each of which is represented in at least two OTUs

A nucleotide site is *informative* only if it favors a subset of trees over the other possible trees. *Invariant* (1, 6, 8 in the picture) and *uninformative* sites are not considered.

Variable sites:

Site 2 is uninformative because all three possible trees require 1 evolutionary change, G ->A.

Site 3 is uninformative because all trees require 2 changes.

Site 4 is uninformative because all trees require 3 changes.

Site 5 is informative because tree I requires one change, trees II and III require three changes
Site 7 is informative, like site 5

Site 9 is informative because tree II requires one change, trees I and III require two.

Inferring the maximum parsimony tree

■ Identification of all informative sites and for each possible tree the minimum number of substitutions at each informative site is calculated.

■ In the example for sites 5, 7 and 9:

tree I requires 1, 1, and 2 changes
tree II requires 2, 2, and 1 changes
tree III requires 2, 2, and 2 changes.

■ Summing the number of changes over all the informative sites for each possible tree and chosing the tree associated with the smallest number of changes:

Tree I is chosen because it requires 4 changes, II and III require 5 and 6 changes.

■ In the case of 4 OTUs an informative site can favor only one of the three possible alternative trees. For example, site 5 favors tree I over trees II and III, and is thus said to support tree I. The tree supported by the largest number of informative sites is the most parsimonious tree. In the cases where more than 4 OTUs are involved, *an informative site may favor more than one tree and the maximum parsimony tree may not necessarily be the one supported by the largest number of informative sites.*

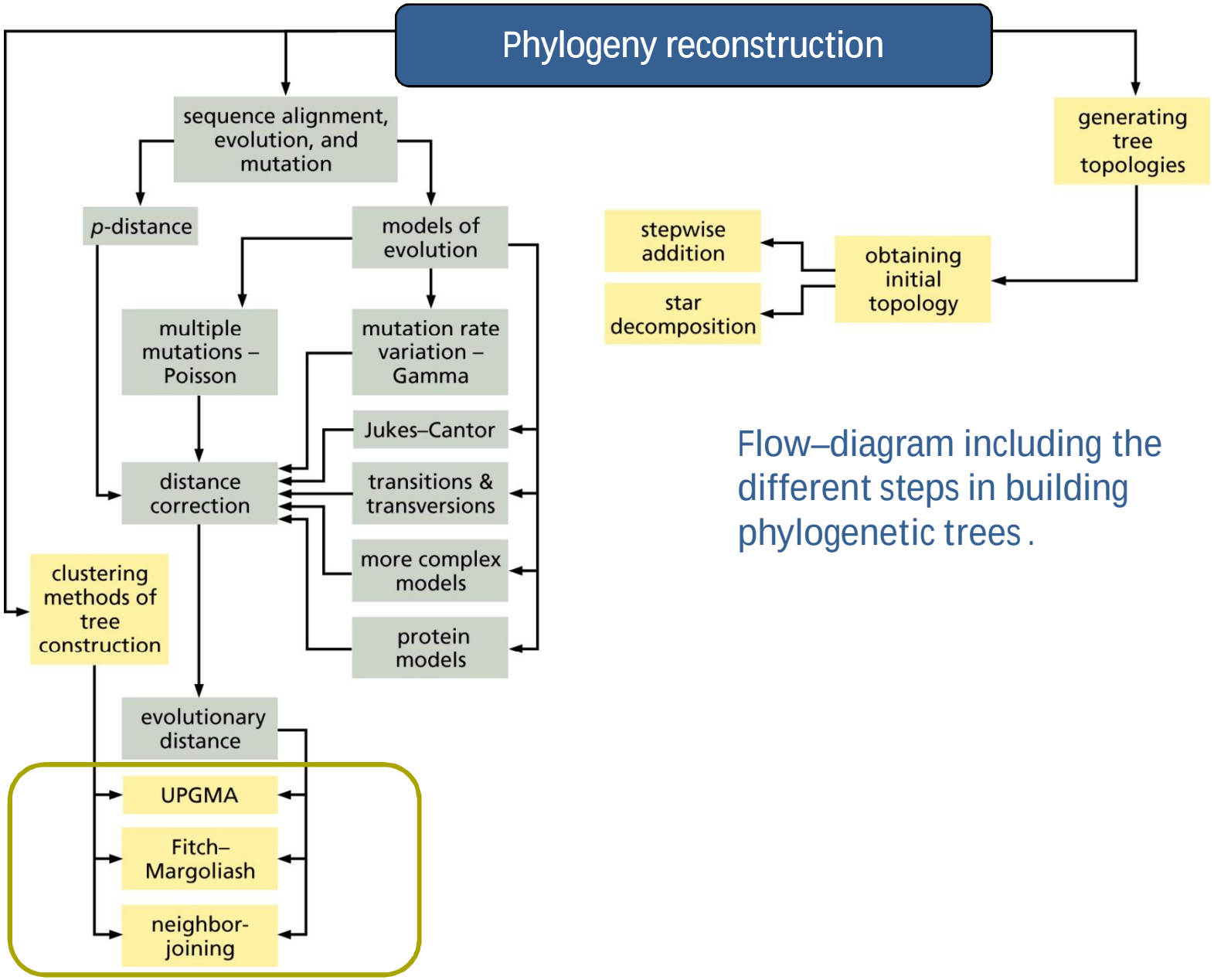
Exhaustive and heuristic searching for the maximum parsimony tree

- The total number of substitutions at both informative and uninformative sites in a particular tree is called the tree length.
- When the number of OTUs is small, it is possible to look at *all possible trees*, determine their length, and choose among them the shortest one(s) = *exhaustive search*.

- Large number of sequences (more than about 12) makes exhaustive searches impossible.

- Short-cut algorithms, for example ´branch-and-bound´: First an arbitrary tree is considered (or a tree obtained by another methods, for example some distance method), and compute the minimum number of substitutions for the this tree, which is considered as the "upper bound" to which the length of any other tree is compared. The rationale is that the maximum parsimony tree must be either equal in length to this tree *or shorter.*

- Above 20 sequences heuristic searches are needed: only a manageable subset of all the possible trees is examined. Branch swapping (rearrangement) is used to generate topologically similar trees from a initial one. Subtree pruning and regrafting is one method.



Flow–diagram including the different steps in building phylogenetic trees .

This picture is modified from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 277.

Distance matrix methods

Neighbor-joining phylogeny by MEGA-software

- Introduction to getting started with phylogenies in practice: exercise session 5 with MEGA-software

- Introduced in 1960´s, based on clustering algorithms of Sokal and Sneath (1963, Principles of numerical taxonomy).

- Calculatation of a measure of the distance between each pair of units (for example, species) and then finding a tree that predicts the observed set of distances as closely as possible.

- The data is reduced to a matrix table of pairwise distances, which can be considered as estimates of the branch length separating a given pair of units compared.

- Each distance infers the best unrooted tree for a given pair of units. In effect, there is a number of estimated two-unit trees, and finding the *n*-unit tree that is implied is the task.

- The individual distances are not exactly the path lengths in the full *n*-unit tree between two units: finding the full tree that does the best job of approximating the individual two-unit trees.

Distances in a phylogenetic tree

- Distance matrix D = (d_{ij}) gives pairwise distances for *leaves* of the phylogenetic tree

- In addition, the phylogenetic tree will now specify distances between leaves and internal nodes

- Distances d_{ij} in evolutionary context satisfy the following conditions:

  Symmetry: $d_{ij} = d_{ji}$ for each i, j
  Distinguishability: $d_{ij} \neq 0$ if and only if $i \neq j$
  Triangle inequality: $d_{ij} \leq d_{ik} + d_{kj}$ for each i, j, k
  Distances satisfying these conditions are called metric
  In addition, evolutionary mechanisms may impose additional constraints on the distances: *additive* and *ultrametric* distances

- A tree is called *additive*, if the distance between any pair of leaves (i, j) is the sum of the distances between the leaves and a node k on the shortest path from i to j in the tree

  $d_{ij} = d_{ik} + d_{jk}$

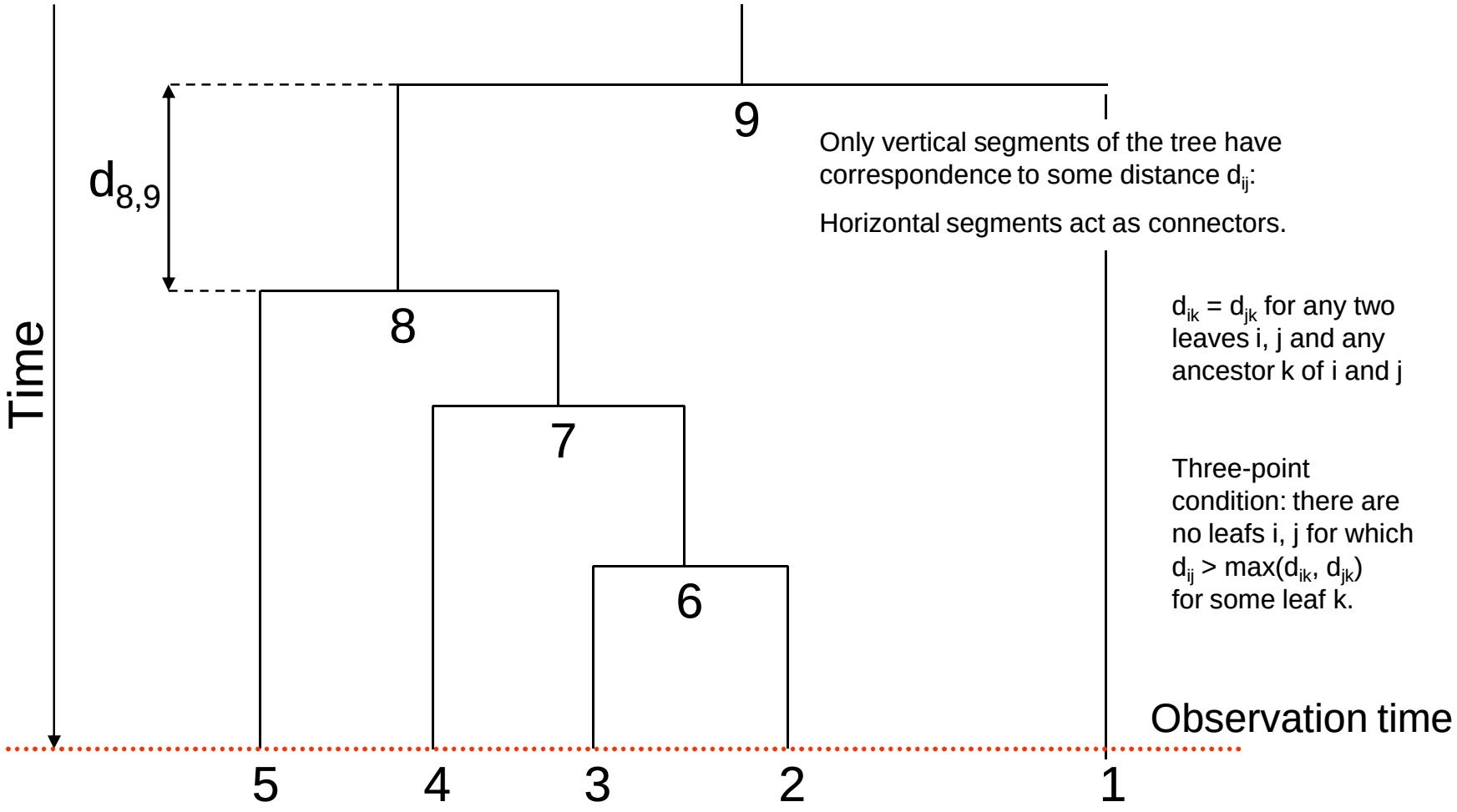
- A rooted additive tree is called an *ultrametric tree*, if the distances between any two leaves i and j, and their common ancestor k are equal

  $d_{ik} = d_{jk}$

- Edge length d_{ij} corresponds to the time elapsed since divergence of i and j from the common parent ,i.e. edge lengths are measured by a "*molecular clock*" with a constant rate

Ultrametric tree

Distances to be ultrametric can be found by the three-point condition:
D corresponds to an ultrametric tree if and only if for any three species (OTUs) i, j and k, the distances satisfy $d_{ij} \leq \max(d_{ik}, d_{kj})$



Only vertical segments of the tree have correspondence to some distance d_{ij}:

Horizontal segments act as connectors.

$d_{ik} = d_{jk}$ for any two leaves i, j and any ancestor k of i and j

Three-point condition: there are no leafs i, j for which $d_{ij} > \max(d_{ik}, d_{jk})$ for some leaf k.

UPGMA algorithm

- Unweighted Pair Group Method using arithmetic Averages, constructs an ultrametric phylogenetic tree via clustering

- The algorithm works by at the same time merging two clusters and creating a new node on the tree. The tree is built from leaves towards the root.

Neighbor-joining algorithm

- Neighbor joining has similarities to UPGMA, Differences in the choice of function f(C_1, C_2) and how to assign the distances

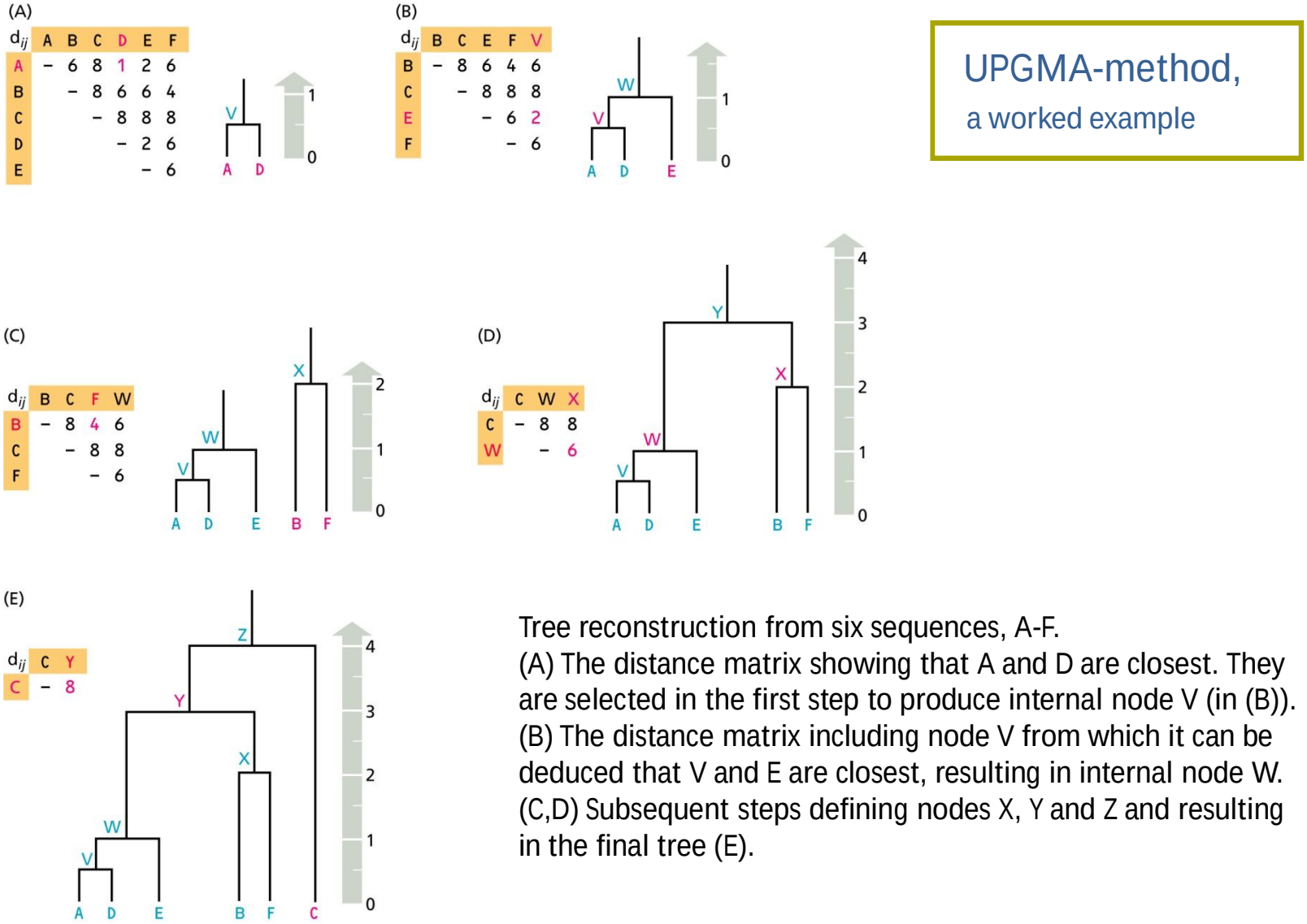
  Find clusters C_1 and C_2 that minimise a function f(C_1, C_2)
  Join the two clusters C_1 and C_2 into a new cluster C
  Add a node to the tree corresponding to C
  Assign distances to the new branch

  - The distance d_{ij} for clusters C_i and C_j is $$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$$
  - Let u(C_i) be the separation of cluster C_i from other clusters defined by $u(C_i) = \frac{1}{n-2} \sum_{C_j} d_{ij}$ where n is the number of clusters.
  - Instead of trying to choose the clusters C_i and C_j closest to each other, neighbor joining at the same time

    Minimises the distance between clusters C_i and C_j and
    Maximises the separation of both C_i and C_j from other clusters

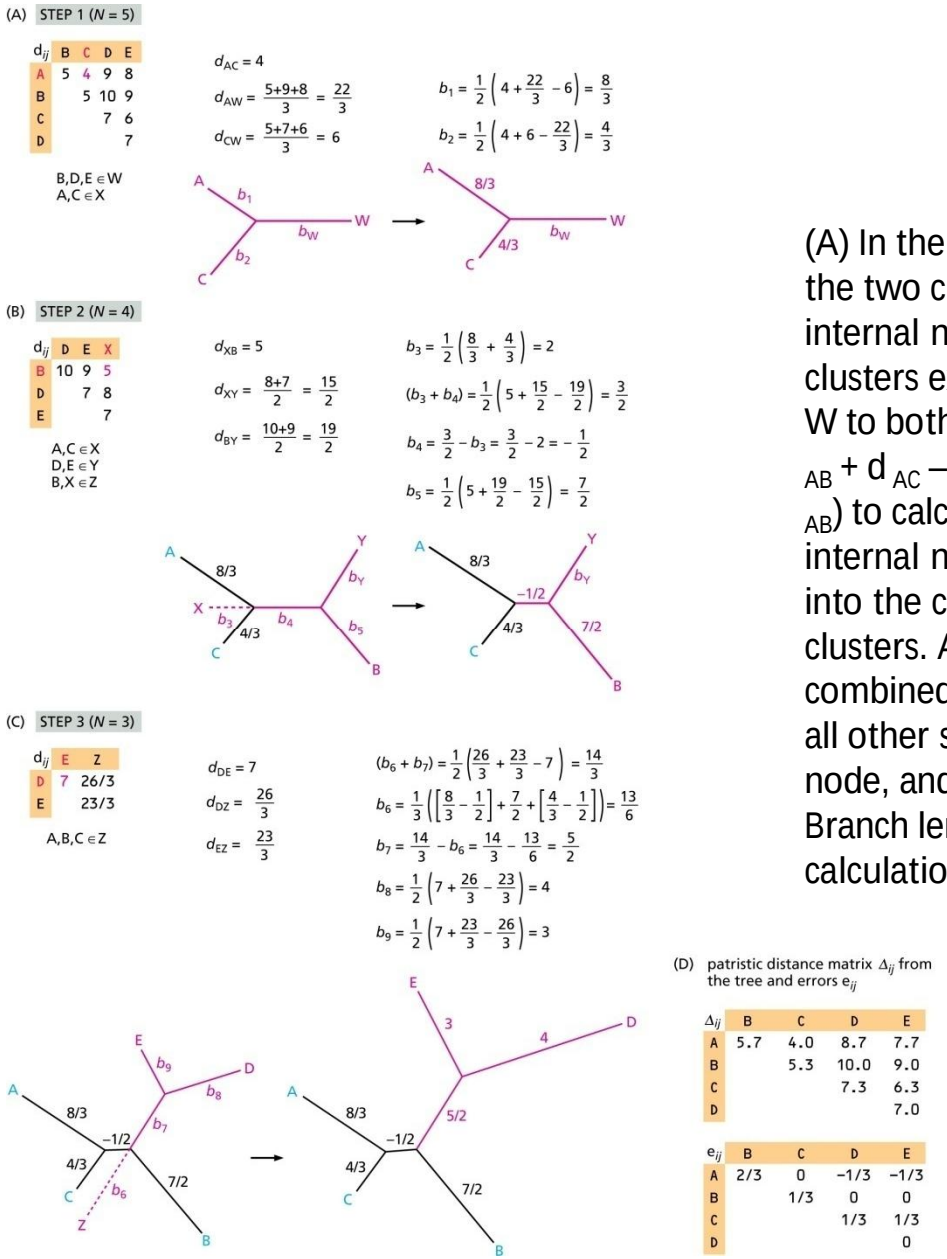
UPGMA-method,

a worked example



Tree reconstruction from six sequences, A-F.
(A) The distance matrix showing that A and D are closest. They are selected in the first step to produce internal node V (in (B)).
(B) The distance matrix including node V from which it can be deduced that V and E are closest, resulting in internal node W.
(C,D) Subsequent steps defining nodes X, Y and Z and resulting in the final tree (E).

This picture is from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 279.



(D) patristic distance matrix Δ_{ij} from the tree and errors e_{ij}

Δ_{ij}	B	C	D	E
A	5.7	4.0	8.7	7.7
B		5.3	10.0	9.0
C			7.3	6.3
D				7.0

e_{ij}	B	C	D	E
A	2/3	0	−1/3	−1/3
B		1/3	0	0
C			1/3	1/3
D				0

This picture is from : 277. Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 281.

Fitch-Margoliash method, NOTE: This distance-method not in MEGA-software

(A) In the first step the shortest distance is used to identify the two clusters (A,C) which are combined to create the next internal node. A temporary cluster (W) is defined as all clusters except these two, and the distances calculated from W to both A and C. The method then uses equations $b_1 = ½(d_{AB} + d_{AC} – d_{BC}$, $b_2 = ½(d_{AB} + d_{BC} – d_{AC})$, $b_3 = ½(d_{AC} + d_{BC} – D_{AB})$ to calculate the branch lengths from A and C to the internal node that connects them. (B) A and C are combined into the cluster X and the distances calculated from the other clusters. After identifying B and X as the next clusters to be combined to create cluster Z, the temporary cluster Y contains all other sequences. X is the distance b_3 from the new internal node, and the distance between the internal nodes is b_4. Branch length b_4 is negative (not realistic); in future calculations this branch is treated like all others. (C) Combining sequences A,B and C into cluster Z, the sequences D and E are added to the tree in the final step. (D) The final tree has a negative branch length. The tables give the patristic distances (those measurer on the tree itself) and the errors (e_{ij}). The tree has a wrong topology,

as becomes clear with the neighbor-joining tree from the same data.

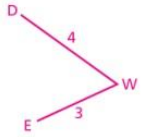
Neighbor-joining method,

a worked example

(A) STEP 1 ($N = 5$)

d_{ij}	B	C	D	E	U_i	$3\delta_{ij}$	B	C	D	E	
A	5	4	9	8	26		−40	−36	−32	−32	A
B		5	10	9	29			−36	−32	−32	B
C			7	6	22				−34	−34	C
D				7	33					−42	D
E					30						E

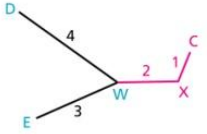
D and E are neighbors through internal node W with $d_{DW} = \frac{1}{2}\left(7 + \frac{33-30}{3}\right) = 4$ and $d_{EW} = 7 - 4 = 3$.



(B) STEP 2 ($N = 4$)

d_{ij}	B	C	W	U_i	$2\delta_{ij}$	B	C	W	
A	5	4	5	14		−20	−18	−18	A
B		5	6	16			−18	−18	B
C			3	12				−20	C
W				14					W

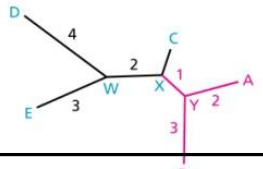
C and W are neighbors through internal node X with $d_{CX} = \frac{1}{2}\left(3 + \frac{12-14}{2}\right) = 1$ and $d_{WX} = 3 - 1 = 2$.



(C) STEP 3 ($N = 3$)

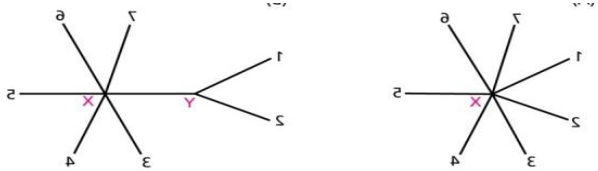
d_{ij}	B	X	U_i	δ_{ij}	B	X	
A	5	3	8		−12	−12	A
B		4	9			−12	B
X			7				X

Three alternatives (of which here we choose one of the two with an internal node):
A and X are neighbors through internal node Y with $d_{AY} = 2$ and $d_{XY} = 1$ or
B and X are neighbors through internal node Y with $d_{BY} = 3$ and $d_{XY} = 1$.
Whichever is chosen, the remaining distance d_{AY} or d_{BY} will be found in the next d_{ij} matrix.



The distance matrix is the same as in the Fitch-Margoliash example.
At each step the distances are converted by using the algorithm which minimizes the total tree distance (the minimum evolution principle).

The first step:



(A) Star-tree in which all sequences are joined directly to a single internal node X with no internal branches.

(B) After sequences 1 and 2 have been identified as the first pair of nearest-neighbors, they are separated from node X by and internal node Y. The method calculates the brabch lengths from sequences 1 and 2 to node Y to complete the step.

This picture is from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 284.

Parameters of nucleotide change

One-parameter model, the ´Jukes-Cantor model´

- Assumption: nucleotide substitutions occur with equal probabilites, α
- The rate of substitution for each nucleotide is 3α per unit time

	A	T	C	G
A		**α**	**α**	**α**
T	**α**		**α**	**α**
C	**α**	**α**		**α**
G	**α**	**α**	**α**	

- At time 0: A at a certain nucleotide site, $P_{A(0)} = 1$
- Question: probability that this site is occupied by A at time *t* , $P_{A(t)}$?
- At time 1, probability of still having A at this site is

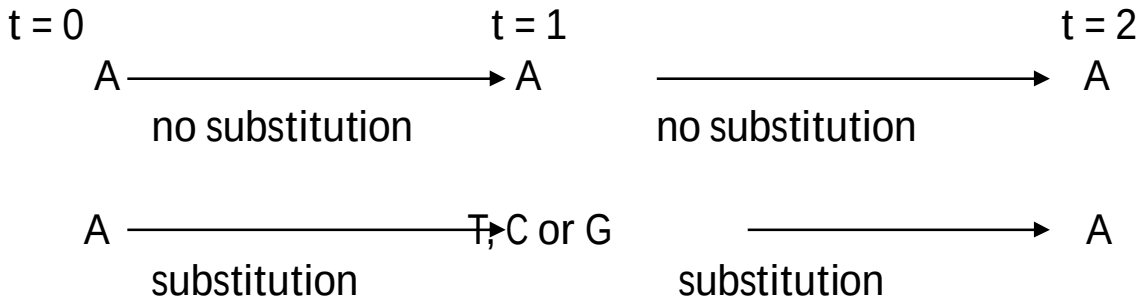
$$P_{A(1)} = 1 - 3\alpha \qquad (1)$$

- 3α is the probability of A changing to T, C, or G

- The probability of the site having A at time 2 is

$$P_{A(2)} = (1 - 3\alpha)P_{(A1)} + \alpha\,[1 - P_{A(1)}] \qquad (2)$$

- This includes two possible course of events:



- The following recurrence equation applies to any t

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha[1 - P_{A(t)}] \qquad (3)$$

Note that this holds also for t = 0, because $P_{A(0)} = 1$ and thus
$P_{A(0+1)} = (1 - 3\alpha)\,P_{A(0)} + \alpha\,[1 - P_{A(0)}] = 1 - 3\alpha$
which is identical with equation (1).

- The amount of change per unit time, rewriting equation (3):

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3\alpha P_{A(t)} + \alpha[1 - P_{A(t)}] = -4\alpha P_{A(t)} + \alpha \qquad (4)$$

- Approximating the previous discrete-time model by a continuous-time model, by regarding $\Delta P_{A(t)}$ as the rate of change at time t

$$dP_{A(t)} / dt = -4\alpha P_{A(t)} + \alpha \qquad (5)$$

- The solution of this first-order linear differential equation is

$$P_{A(t)} = \tfrac{1}{4} + (P_{A(0)} - \tfrac{1}{4})e^{-4\alpha t} \qquad (6)$$

- The starting condition was A at the given site, $P_{A(0)} = 1$, consequently

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{3}{4} e^{-4\alpha t} \qquad (7)$$

- Equation holds regardless of the initial conditions, for example if the initial nucleotide is not A, then $P_{A(0)} = 0$, and the probability of having A at time t

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{1}{4} e^{-4\alpha t} \qquad (8)$$

- Equations (7) and (8) describe the substitution process. If the initial nucleotide is A, then $P_{A(t)}$ decreases exponentially from 1 to ¼ . If the initial nucleotide is not A, then $P_{A(t)}$ will increase monotonically from 0 to ¼ .

- Under this simple model, after reaching equilibrium, $P_{A(t)}=P_{T(t)}=P_{C(t)}=P_{G(t)}$ for all subsequent times.

- Equation (7) can be rewritten in a more explicit form to take into account that the initial nucleotide is A and the nucleotide at time t is also A

$$P_{AA(t)} = ¼ + ¾ e^{-4\alpha t} \qquad (9)$$

- If the initial nucleotide is G instead of A, from equation (8)

$$P_{GA(t)} = ¼ + ¼ e^{-4\alpha t} \qquad (10)$$

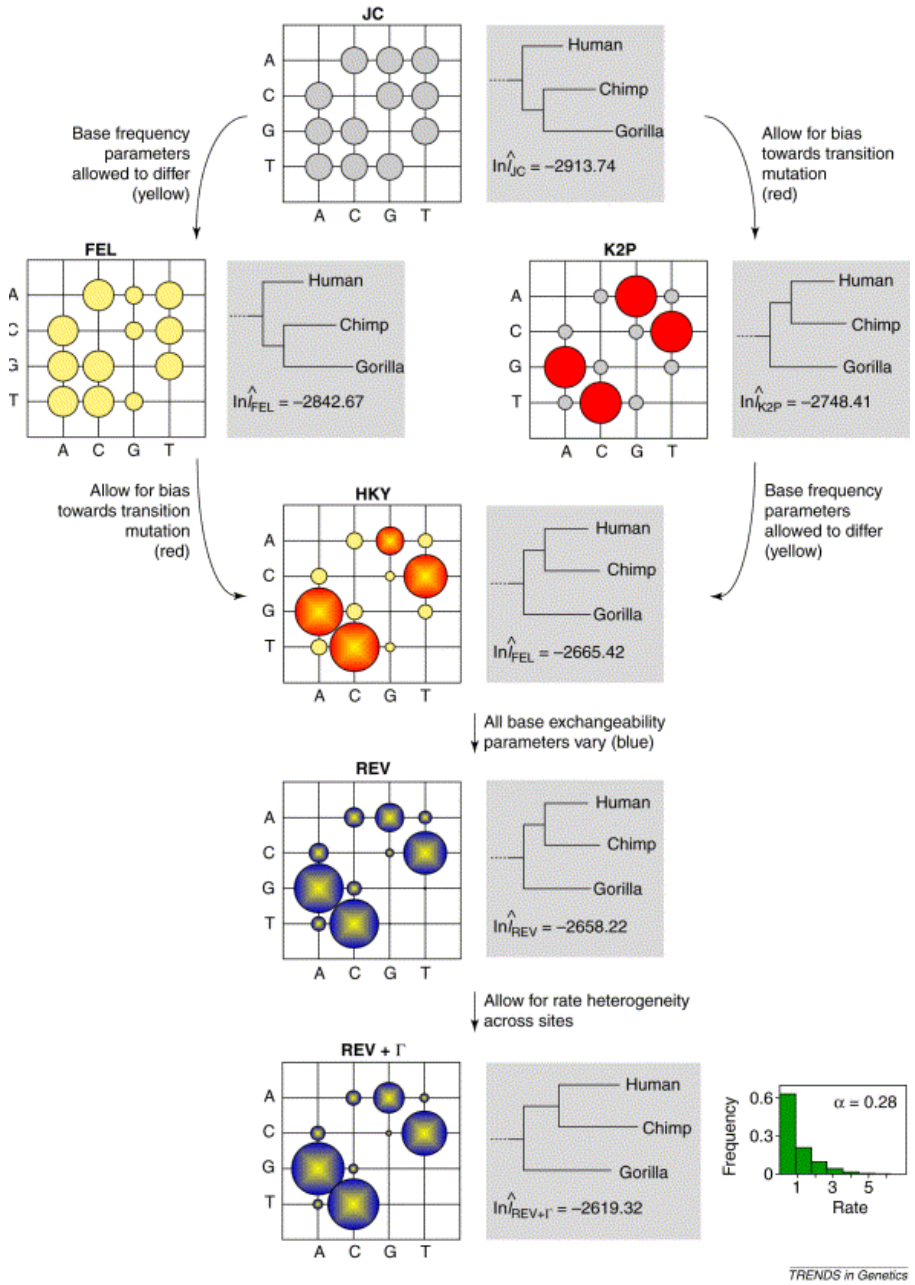
- Since all the nucleotides are equivalent under the Jukes-Cantor model, the general probability, $P_{ij(t)}$, that a nucleotide will become *j* at time *t*, given that it was *i* at time 0, equations (9) and (10) give the general probabilities $P_{ii(t)}$ and $P_{ij(t),}$ where $i \neq j$.

Kimura´s two- parameter -model

- The Jukes-Cantor –model was introduced in 1969 when virtually nothing was known about nucleotide substitution
- In 1980 Kimura proposed different parameters for transitions and transversions.
- Transition is a nucleotide change between purines, A and G, and pyrimidines, T and C. Transversion is a purine – pyrimide change.
- The rate of transition change is α and transversion change is β per unit time

	A	T	C	G
A		**β**	**β**	**α**
T	**β**		**α**	**β**
C	**β**	**α**		**β**
G	**α**	**β**	**β**	

- The development of models of sequence evolution is a very active field, profuse amount of publications.
- Two approaches:
    - Empirically using properties calculated through comparisons of large numbers of observed sequences. For example simply counting apparent replacements between many closely related sequences.
    - Empirical models result in fixed parameter values which are estimated only once and the n assumed to be applicable to all datasets => computationally easy to use
    - Parametrically on the basis of the chemical or biological properties of DNA and amino acids. For example, incorporating a parameter to describe the relative frequency of transition (purine- purine, pyrimidine-pyrimidine) and transversion (purine –pyrimidine).
    - Parameter values are derived from the dataset in each particular analysis.
- Both methods result in Markov process models, defined by matrices containing the relative rates (=the relative numbers, on average, and per unit time). From these are calculated the probabilities of change from any nucleotide to any other nucleotide, including the probability of remaining the same, over any period of evolutionary time at any site



Relationships among six standard models of nucleotide evolution.

For each model the matrix of rates of substitutions between nucleotides is shown (represented by a bubble plot where the area of each bubble indicates the corresponding rate), a partial representation of a hominoid phylogeny as inferred by that model from a mitochondrial sequence dataset, and the maximum log-likelihood value obtained. For the REV+Γ model also the gamma distribution of rates among sites described by the inferred parameter value α=0.28 is shown. The reverse-J shape of the graph indicates that the majority of sites have low rates of evolution, with some sites having high rates of evolution. The JC model assumes that all nucleotide substitutions occur at equal rates. The models become more advanced moving down the figure, as illustrated in the bubble plots by their increasing flexibility in estimating relative replacement rates and as reflected by increasing log-likelihoods. Note how the inferred maximum likelihood phylogeny changes significantly as the models become more advanced (compare JC with K2P); inferred branch lengths also tend to increase (compare REV to REV+Γ). Arrows show where models are nested within each other; that is, where the first model is a simpler form of the next. For example, the JC model is nested within the K2P model (it is a special case arising when κ is fixed equal to 1), but the K2P model is not nested with the FEL model.

Model comparisons

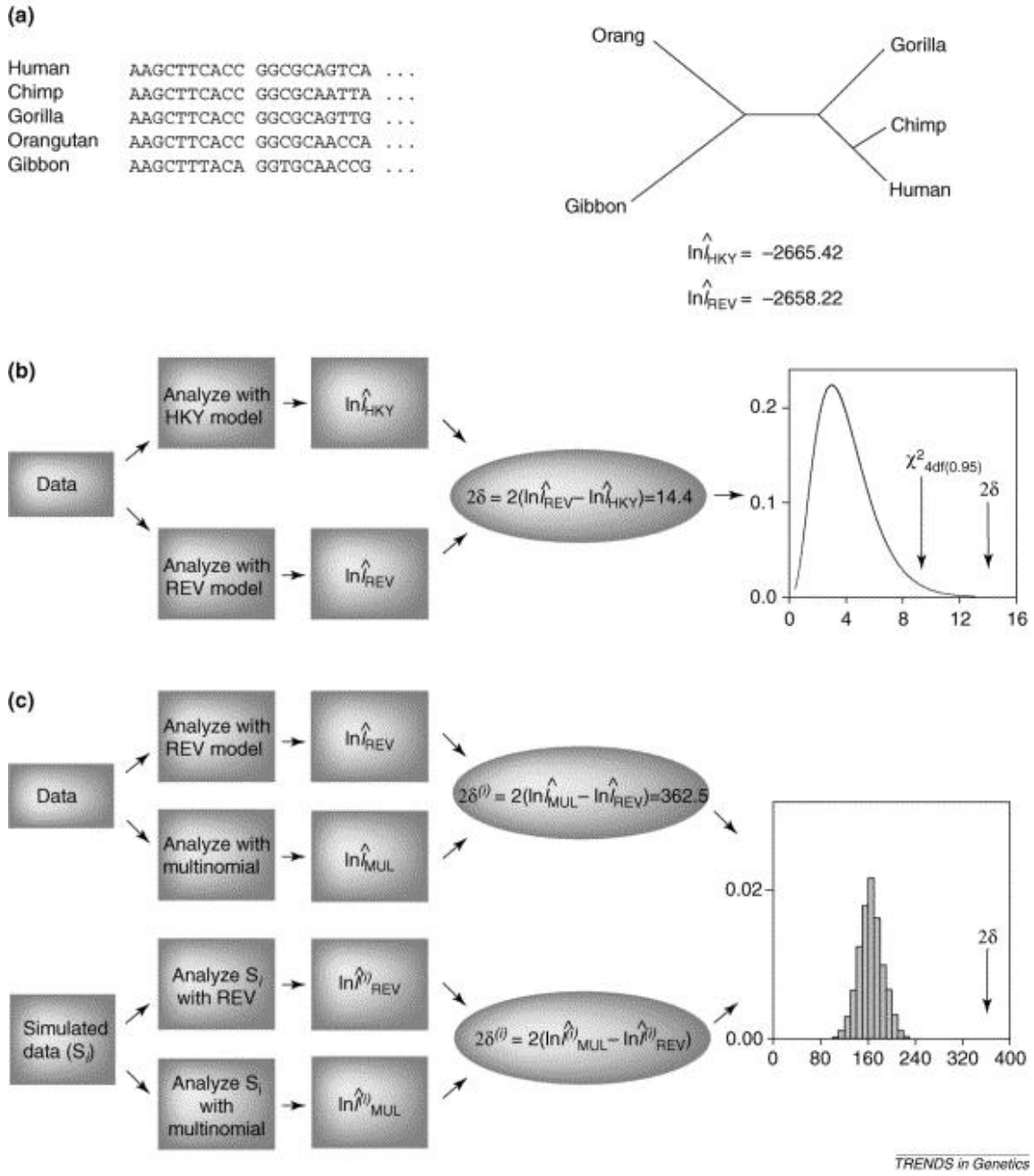
- The likelihood framework permits estimation of parameter values and their standard errors from the observed data (with no need for any a priori knowledge).

- For example, a transition / transversion bias estimated as $\kappa = 2.3 +/- 0.16$ effectively excludes the possibility that there is no such bias ($\kappa = 1$), whereas $\kappa = 2.3 +/- 1.6$ does not.

- Likelihood ratio tests compare two competing models, using their maximized likelihoods with a statistic, 2δ, that measures how much better an explanation of the data the alternative model gives. To perform a significance test, the distribution of values of 2δ expected under the simpler hypothesis is required. If the observed value of 2δ is too great to be consistent with this distribution (P-values), the simpler model is rejected in favour of the more complex model.

- When two models being compared are nested, the simpler model being a special case of the more complex model obtained by constraining certain free parameters to take particular values, then the required distribution for 2δ is usually a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models.

- When the models are not nested (the usual situation with complex models) the required distribution of 2δ can be estimated by Monte Carlo simulation or parametric bootstrapping.
- Figure in the next page illustrates a test for assessing whether one particular model is a statistically adequate description of the evolution of a set of sequences.
- This test almost invariably indicates that current models of sequence evolution are not explaining the evolutionary patterns in the data fully (sequences have been evolving by natural selection



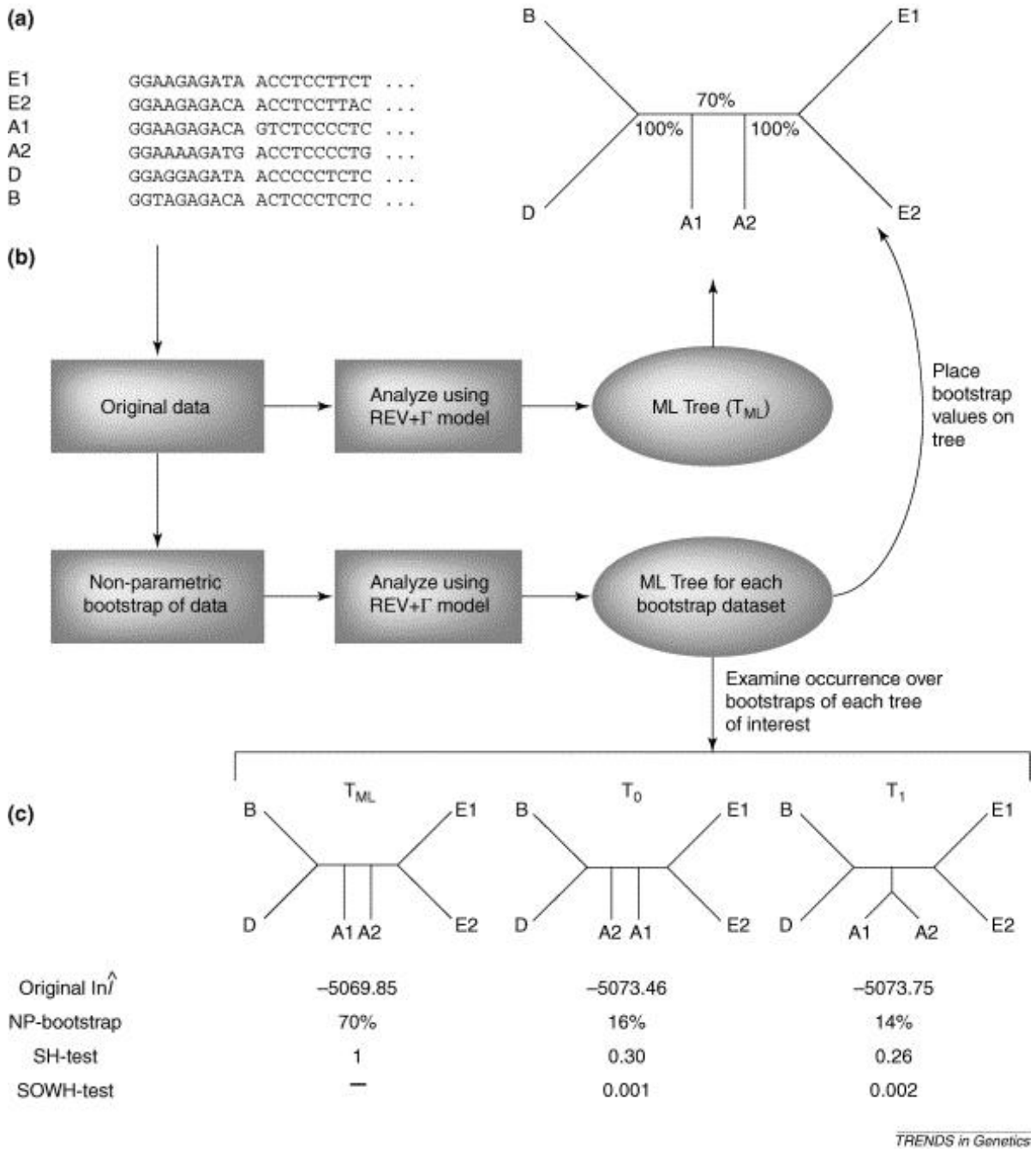
Statistical tests of models.

(a) Part of the mitochondrial sequence dataset used in previous figure, and the maximum likelihood phylogenetic tree and likelihood values from the HKY) and REV models.

(b) The statistical test to compare these models of nucleotide substitution, in which the likelihood ratio statistic 2δ is compared with a $\chi 42$ distribution. The observed value of 2δ, 14.4, has a *P*-value considerably less than 0.05, and the HKY model is rejected in favour of the REV model.

(c) The test of the adequacy of the REV model. The test statistic is derived from a comparison of the REV model and a multinomial model that identifies the maximum possible likelihood attainable under any model. The test distribution is estimated by parametric bootstrapping, in which simulated datasets S_i (generated using the maximum likelihood phylogeny and substitution model parameters estimated with the REV model) and are subjected to the same analysis as the original data. Comparison of the test statistic and the distribution of values obtained from simulated data indicates that the observed value 2δ is far in excess of what is expected if the REV model were accurate, and we can conclude that a more complex evolutionary model is necessary to describe the patterns of evolution of these sequences fully.

- Considering the primary interest in the topology of the inferred evolutionary tree:
  - As with estimates of model parameters, a single point-estimate is of little value without some measure of the confidence.
  - Non-parametric bootstrapping: comparisons of an inferred tree with a set of bootstrap replicate trees, typically in the form of tabulation of the proportion of the bootstrap replicates in which each branch from the inferred tree occurs.
  - Difficulty in the precise interpretation of what these values represent.

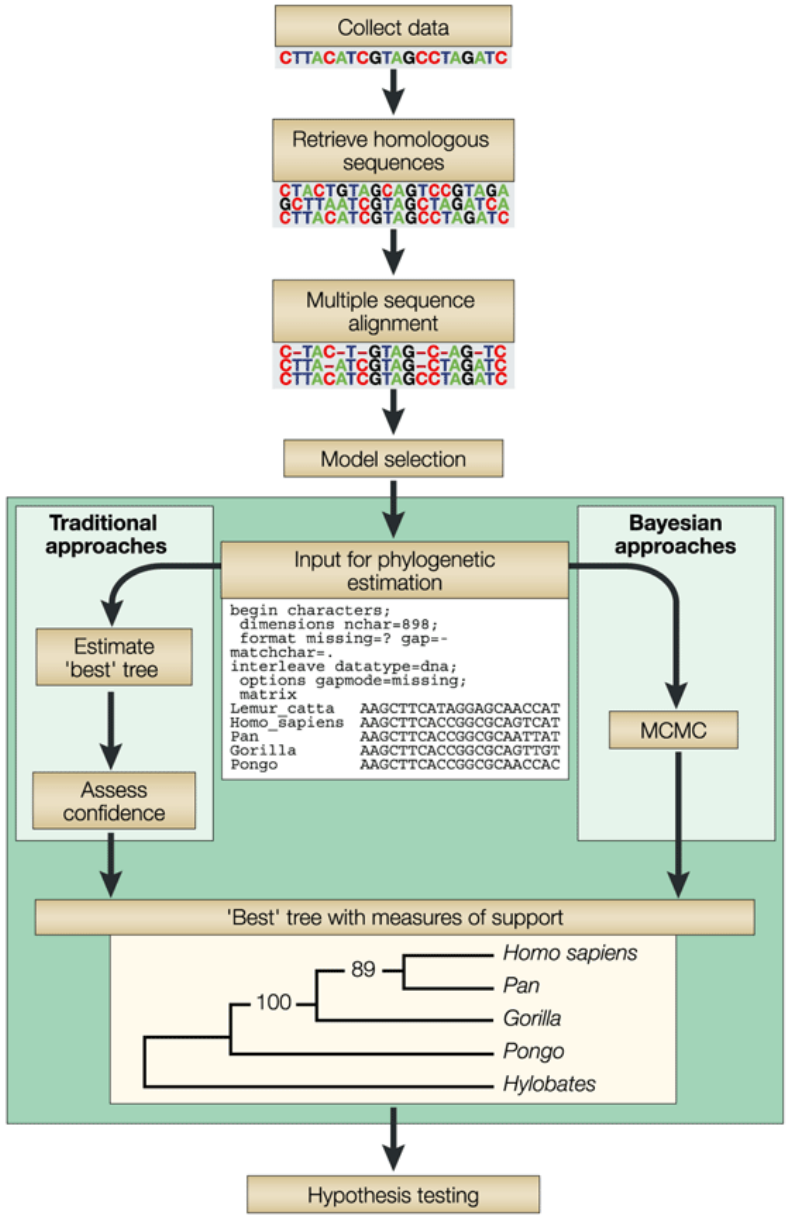


Statistical tests of tree topologies.

(a) Part of a dataset of six HIV-1 nucleotide sequences, from HIV-1 subtypes A (two examples), B, D and E (two examples), and the maximum likelihood (ML) phylogenetic inference under the REV+Γ model. The ML phylogeny (TML) differs from the conventional tree (T1), which would group the two subtype A sequences together.

(b) A non-parametric bootstrap analysis of confidence in TML: analysis of many bootstrapped datasets allows calculation of the proportion of replicates in which branches appearing in TML also arise in the bootstrap trees – these values are indicated in (a). Note that the central branch does not receive a statistically significant bootstrap proportion, indicating that there is some uncertainty about the position of the subtype A sequences.

(c) The likelihoods assigned to TML, T1 and another plausible tree T0, and the proportion of the time these trees are inferred from non-parametric bootstrap datasets. Note that T0 and T1 are each recovered a considerable proportion of the time.


Collect data
CTTACATCGTAGCCTAGATC
Retrieve homologous sequences
CTACTGTAGCAGTCCGTAGA
GCTTAATCGTAGCTAGATCA
CTTACATCGTAGCCTAGATC
Multiple sequence alignment
C-TAC-T-GTAG-C-AG-TC
CTTA-ATCGTAG-CTAGATC
CTTACATCGTAGCCTAGATC
Model selection
Traditional approaches
Input for phylogenetic estimation
Bayesian approaches
begin characters;
dimensions nchar=898;
format missing=? gap=-
matchchar=.
interleave datatype=dna;
options gapmode=missing;
matrix
Lemur_catta AAGCTTCATAGGAGCAACCAT
Homo_sapiens AAGCTTCACCGGCGCAGTCAT
Pan AAGCTTCACCGGCGCAATTAT
Gorilla AAGCTTCACCGGCGCAGTTGT
Pongo AAGCTTCACCGGCGCAACCAC
Estimate 'best' tree
MCMC
Assess confidence
'Best' tree with measures of support
89
100
Homo sapiens
Pan
Gorilla
Pongo
Hylobates
Hypothesis testing
Nature Reviews | Genetics

Frequentist and Bayesian tree confidence and credibility

- Quantifying the uncertainty of a phylogenetic estimate is at least as important a goal as obtaining the phylogenetic estimate itself.

- Measures of phylogenetic reliability point out what parts of a tree can be trusted when interpreting the evolution of a group and guide future efforts in data collection that can help resolve remaining uncertainties.

- Bootstrapping (in distance methods and in maximum likelihood phylogenies) and posterior probability (in Bayesian phylogenies)

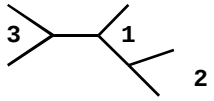
Bootstrapping proceduce

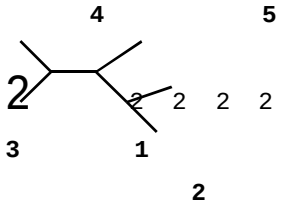
- Sample

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

- Pseudosample 1

1 1 1 1 2 6 6 6 8 8 10 13 13 13 13 15 16 17 17 19

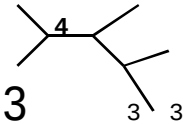




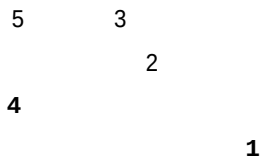
- Pseudosample 2

2 2 2 2 5 7 8 8 9 10 11 12 12 14 14 17 17 18 20 20

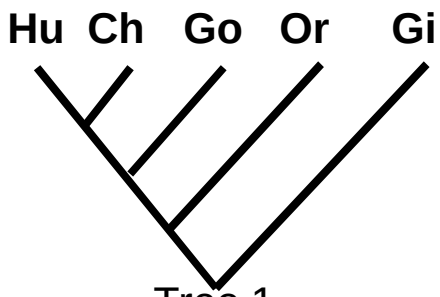
- Pseudosample 3



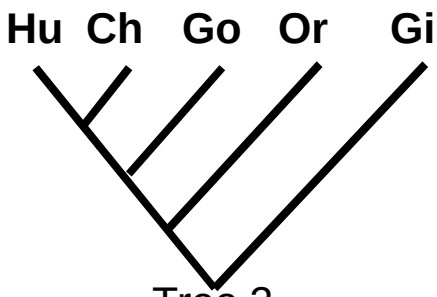
3 3 3 5 5 6 7 7 9 9 11 11 11 11 11 12 12 18 18 18



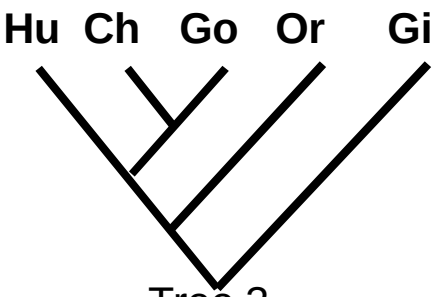
Strict Consensus Tree



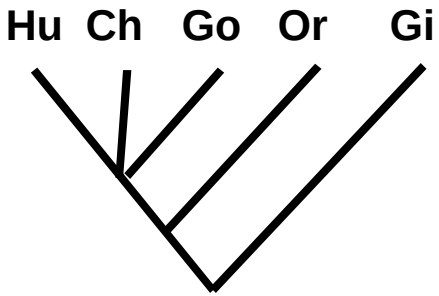
Tree 1



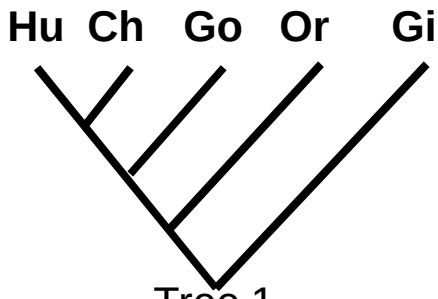
Tree 2



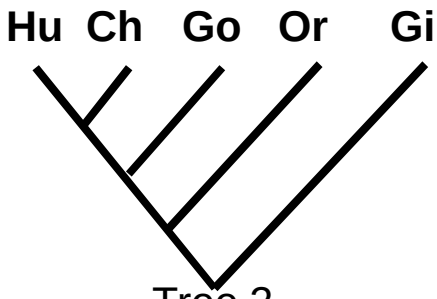
Tree 3



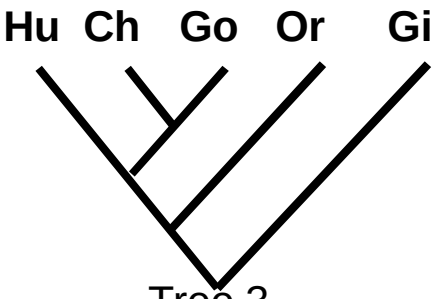
Majority Rule Consensus Tree



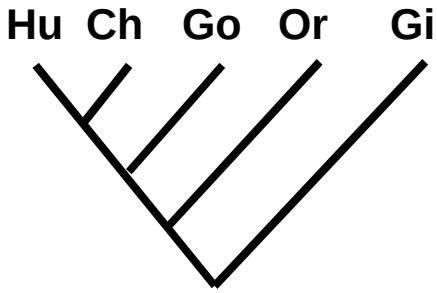
Tree 1

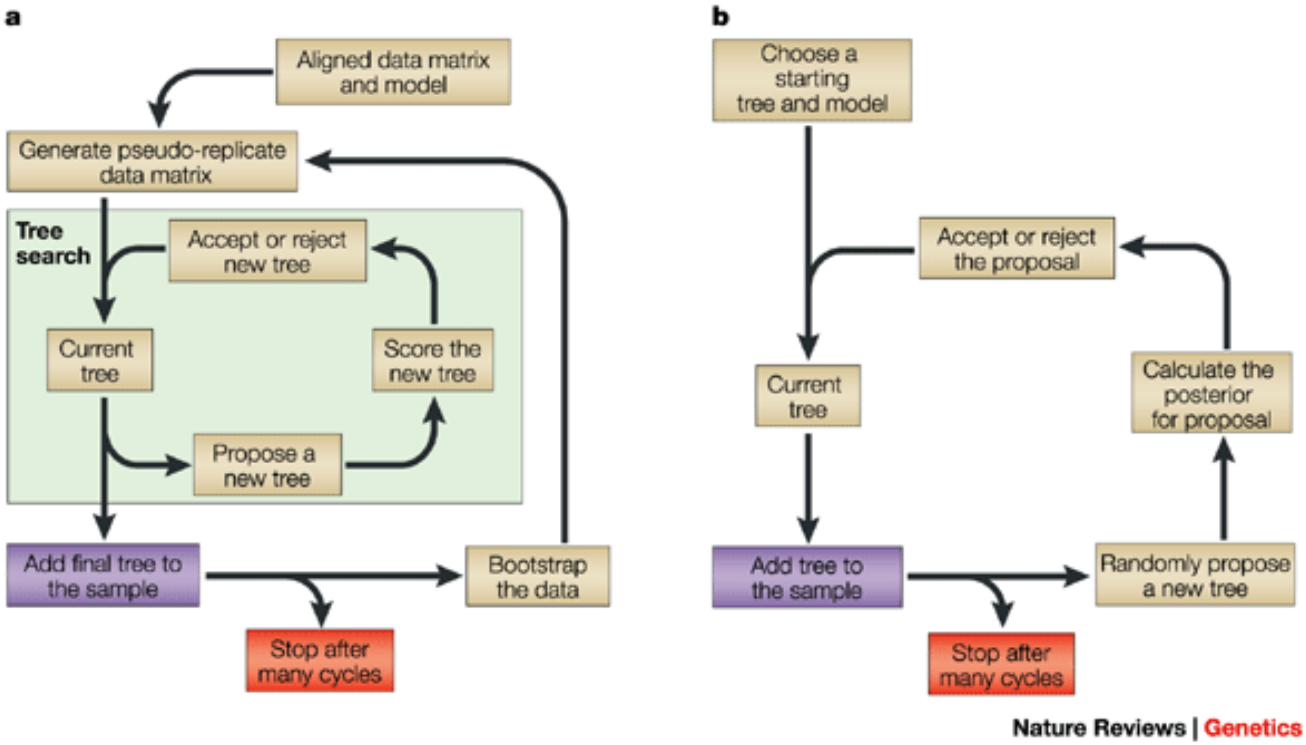


Tree 2



Tree 3





The bootstrapping approach .When optimality-criterion methods are used, a tree search (green box) is performed for each data set, and the resulting tree is added to the final collection of trees. A wide variety of tree-search strategies have been developed, but most are variants of the same basic strategy. An initial tree is chosen, either randomly or as the result of an algorithm — such as neighbour joining . Changes to this tree are proposed; the type of move can be selected randomly or the search can involve trying every possible variant of a particular type of move .The new tree is scored and possibly accepted. Some search strategies are strict hill-climbers — they never accept moves that result in lower scores; others (genetic algorithms or simulated annealing) occasionally accept worse trees in an attempt to explore the tree space more fully.

The Markov chain Monte Carlo (MCMC) methodology is similar to the tree-searching algorithm, but the rules are stricter. From an initial tree, a new tree is proposed. The moves that change the tree must involve a random choice that satisfies several conditions. The MCMC algorithm also specifies the rules for when to accept or reject a tree. MCMC yields a much larger sample of trees in the same computational time, because it produces one tree for every proposal cycle versus one tree per tree search (which assesses numerous alternative trees) in the traditional approach. However, the sample of trees produced by MCMC is highly auto-correlated. As a result, millions of cycles through MCMC are usually required, whereas many fewer (of the order of 1,000) bootstrap replicates are sufficient for most problems.

Examples of phylogeny inference in practice

1) A classical example: phylogenetics in court

Molecular evidence of HIV-1 transmission in a criminal case
By: Metzker *et al., PNAS October 29, 2002, 99: 14292-14297.* doi: 10.1073/pnas.222522599

- 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.

■ Phylogenetic analyses of the gp120 and RT sequences (two genes of the virus) to examine relationships among the patient, victim, and LA (geographical reference area) control viral DNA sequences. The analyses that formed the basis of the results we presented in court were conducted by using the optimality criteria of parsimony and minimum evolution (neighbor joining algorithm).

■ These approaches were used because they **were accepted by the court** in a pre-trial hearing as meeting the criteria for admissibility of evidence. Analyses based on direct likelihood evaluations of the sequence data were not computationally feasible at the time of the pretrial hearing or court case.

■ Recent developments of Markov-chain Monte Carlo (MCMC) approaches have, however, made a Bayesian analysis under a likelihood model feasible. Therefore, additional post-trial analyses were conducted with MCMC Bayesian analysis by using the Metropolis-coupled MCMC algorithm implemented in the program mrBayes.

■ Bayesian analysis was based on a General-Time-Reversible model of sequence evolution, with γ-distributed rate heterogeneity among sites and a calculated proportion of invariable sites (GTR+Γ+I).

■ For each gene, 5,000,000 MCMC generations, and sampled solutions once every 100 generations. After 2,500,000 generations, it was determined that the searches had reached equilibrium by plotting the values for the likelihood scores and the various parameters of the model. We therefore used the samples from the final 2,500,000 generations to compute 95% confidence intervals for the model parameters shown in Tables (next page), and to assess the posterior probabilities of the relationships between the victim and patient sequences.

■ Note that the two genes are very different (see the parameter estimates). This serves here as an example of the importance of models (cf. above, the simple Jukes-Cantor model would not be adequate).

Means and 95% confidence intervals for parameters of the GTR + Γ + I model for gp120 sequences (one gene from the virus)

Parameter	Mean	95% confidence interval
C–T substitution rate	5.03	3.60–7.03
C–G substitution rate	0.97	0.57–1.54
A–T substitution rate	.75	0.52–1.07
A–G substitution rate	3.87	91–5.10
A–C substitution rate	2.34	1.60–3.34
Frequency of A	0.40	0.37–0.43
Frequency of C	0.15	0.13–0.17
Frequency of G	0.23	0.21–0.25
Frequency of T	0.22	0.20–0.25
α (shape of Γ distribution)	0.53	0.43–0.68
Proportion of invariable sites	0.08	0.01–0.18

Means and 95% confidence intervals for parameters of the GTR + Γ + I model for the RT sequences (another gene from the virus)

Parameter	Mean	95% Confidence interval
C–T substitution rate	110.36	23.04–195.53
C–G substitution rate	17.59	2.82–42.02
A–T substitution rate	7.62	1.34–17.32
A–G substitution rate	83.01	16.29–171.17
A–C substitution rate	16.60	3.41–35.62
Frequency of A	0.40	0.36–0.43
Frequency of C	0.17	0.14–0.19
Frequency of G	0.20	0.17–0.23
Frequency of T	0.23	0.20–0.26
α (shape of Γ distribution)	0.94	0.38–1.94
Proportion of invariable sites	0.50	0.29–0.63

■ For the parsimony and minimum evolution analyses, nonparametric bootstrapping was used to test the *a priori* hypothesis of a relationship between the victim and patient sequences. The generally accepted standard for rejecting a null hypothesis (in this case, the null hypothesis is that the sequences obtained from the victim are not most closely related to sequences obtained from the patient) is $P < 0.05$.

■ In forensic studies, however, there is no widely accepted standard for the meaning of *beyond a reasonable doubt*. Under a wide range of conditions, bootstrap proportions (BP) have been shown to represent a conservative estimate of phylogenetic confidence, and 1-BP was used as a conservative estimate of *p* (the probability of type I error) in a test of the *a priori* hypothesis . Because of the importance of estimating the strength of the results, as many bootstrap replications as were computationally feasible for each analysis were constructed.

■ For parsimony analyses, 100,000 bootstrap replicates, whereas for the more computationally intense maximum-likelihood distance analyses (in which large numbers of pairwise distances had to be recalculated for each replicate), 1,000 (gp120) to 10,000 (RT) replicates.

■ In the parsimony analyses, all 100,000 bootstrap replicates of the gp120 gene data supported the victim and patient sequences as the most closely related within the analysis ($P < 0.00001$), and 95,826 bootstrap replicates of the RT gene data supported the victim sequences as embedded within a group of patient sequences ($P < 0.04174$).

■ In the maximum-likelihood distance analyses, all 1,000 bootstrap replicates of the gp120 gene data ($P < 0.001$) supported the closer relationship between the patient and victim viral sequences compared with any of the LA controls, and all 10,000 bootstrap replicates of the RT gene data ($P < 0.0001$) supported the victim sequences as embedded within a group of patient sequences. All 25,000 sampled trees from the MCMC analyses also supported these relationships ($P < 0.00004$). The relationships of the patient and victim RT sequences were virtually identical based on both the originally sampled sequences (sequenced at BCM) and those subsequently sequenced at an other laboratory. (**NOTE:** Maximum likelihood and Bayesian phylogenies have not been introduced during this Introduction to bioinformatics course.)

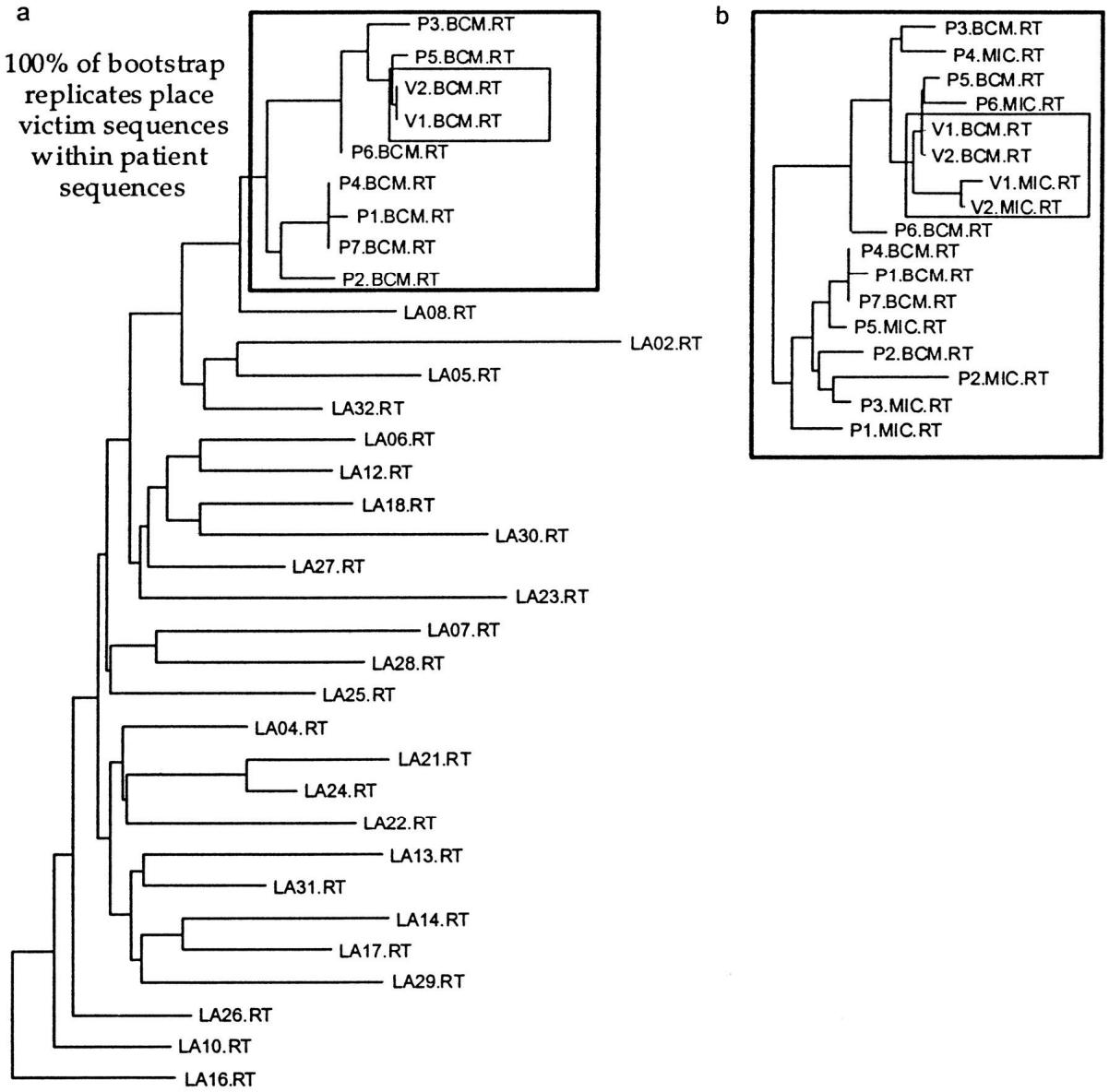
■ **The close relationship between the victim and patient samples was thus supported by both of the genes that we examined, using all major methods of phylogenetic analysis (parsimony, minimum evolution, and likelihood), and a broad range of evolutionary models**.

Phylogenetic analysis of the gp120 region using a minimum evolution criterion assuming HKY+gamma model of evolution.

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 addition to the 100% bootstrap support of this relationship for the minimum evolution analyses, the parsimony bootstrap support and the Bayesian posterior support were also 100%.

Metzker M. L. et.al. PNAS 2002;99:14292-14297

PNAS



Metzker M. L. et.al. PNAS 2002;99:14292-14297

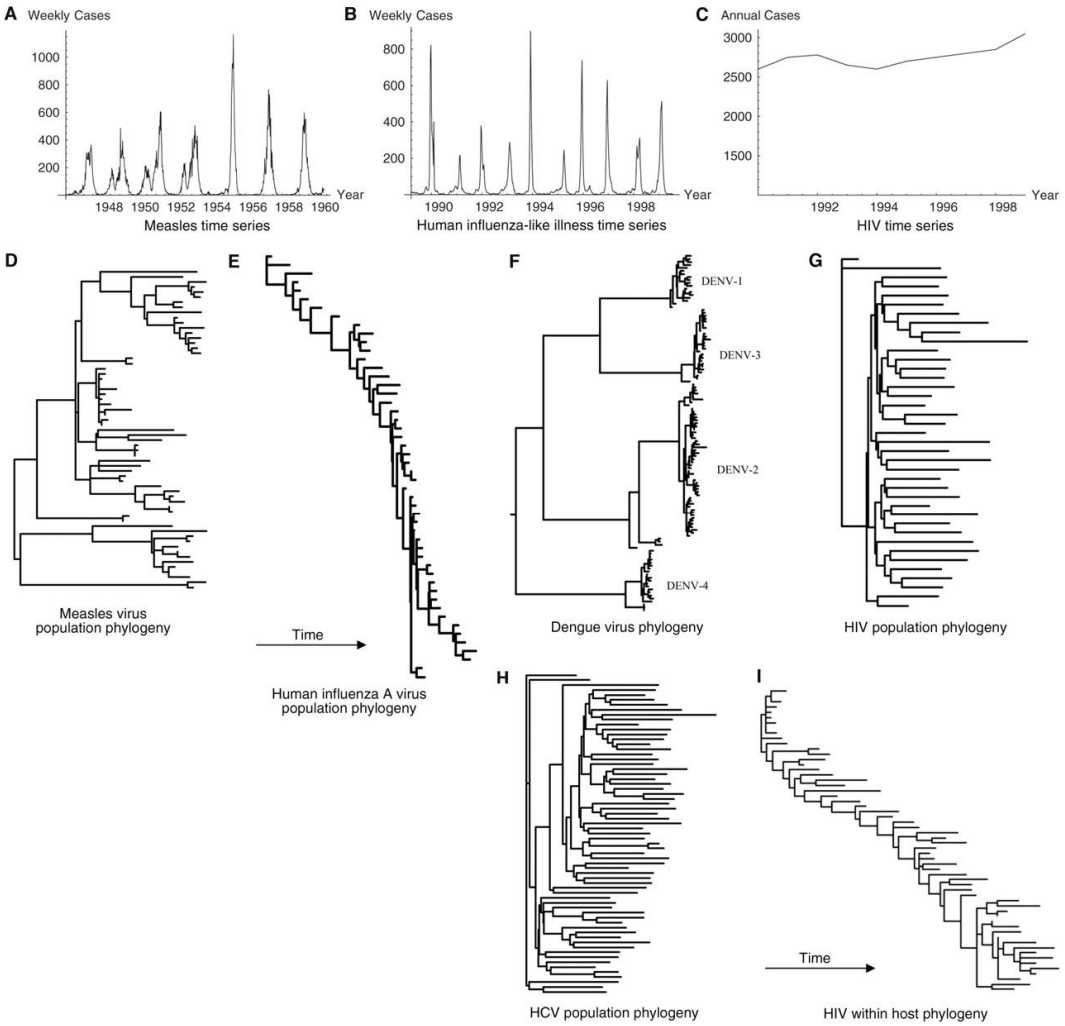
Phylogenetic analysis of the RT region; details of the analysis are the same as for previous page.

(a) Tree based on sequences from BCM .

(b) Subtree of patient and victim sequences, including those added by MIC. In both *a* and *b*, the smaller set of boxed sequences represents the sequences from the victim, and the larger set of boxed sequences represents the patient plus victim sequences. The victim sequences were found to be embedded within the patient sequences in all analyses and for all models of evolution examined. In addition to the 100% bootstrap support of this relationship for the minimum evolution analyses, the parsimony bootstrap support was 96% and the Bayesian posterior support was 100%.

PNAS

2) Phylodynamics: understanding the behaviour of viruses, what happens to their gene sequences during time epochs



(A) Prevaccination measles dynamics: weekly case reports for Leeds, UK.
(B) Weekly reports of influenza-like illness for France.
(C) Annual diagnosed cases of HIV in the United Kingdom.
(D) Measles phylogeny: the measles virus nucleocapsid gene [63 sequences, 1575 base pairs (bp)].
(E) Influenza phylogeny: the human influenza A virus (subtype H3N2) hemagglutinin (*HA1*) gene longitudinally sampled over a period of 32 years (50 sequences, 1080 bp).
(F) Dengue phylogeny: the dengue virus envelope gene from all four serotypes (DENV-1 to DENV-4, 120 sequences, 1485 bp).
(G) HIV-1 population phylogeny: the subtype B envelope (E) gene sampled from different patients (39 sequences, 2979 bp).
(H) HCV population phylogeny: the virus genotype 1b E1E2 gene sampled from different patients (65 sequences, 1677 bp).
(I) HIV-1 within-host phylogeny: the partial envelope (E) gene longitudinally sampled from a single patient over 5.8 years.

3) Resolving the ”tree of life”, the dream of Darwin

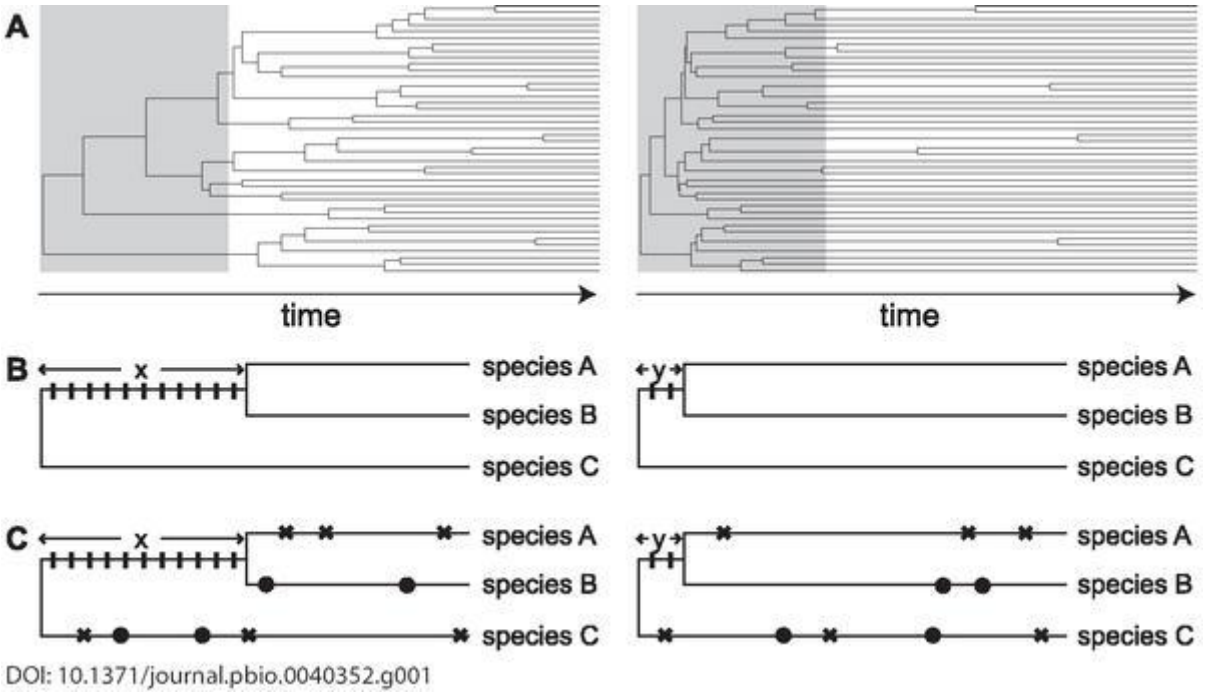
■ Genome analyses are delivering unprecedented amounts of data from an abundance of organisms, raising expectations that in the near future, resolving the tree of life (TOL) will simply be a matter of data collection. However, recent analyses of some key clades in life's history have produced “bushes” and not resolved trees. The patterns observed in these clades are both important signals of biological history and symptoms of fundamental challenges that must be confronted.

■ The combination of the spacing of cladogenetic events and the high frequency of independently evolved characters (homoplasy) limit the resolution of ancient divergences. Because some histories may not be resolvable by even vast increases in amounts of conventional data, the identification of new molecular characters will be crucial to future progress.

■ The famous science writer, Richard Dawkins says: ... *“there is, after all, one true tree of life, the unique pattern of evolutionary branchings that actually happened. It exists. It is in principle knowable. We don't know it all yet. By 2050 we should – or if we do not, we shall have been defeated only at the terminal twigs, by the sheer number of species.”*

■ Examples of open questions: Who are tetrapods' (four-legged animals) closest living relatives? Which is the earliest-branching animal phylum? Answers to such fundamental questions would be easy if the historical connections among all living organisms in the TOL were known. Obtaining an accurate depiction of the evolutionary history of all living organisms has been and remains one of biology's great challenges. The discipline primarily responsible for assembling the TOL—molecular systematics—has produced many new insights by illuminating episodes in life's history, posing new hypotheses, as well as providing the evolutionary framework within which new discoveries can be interpreted.

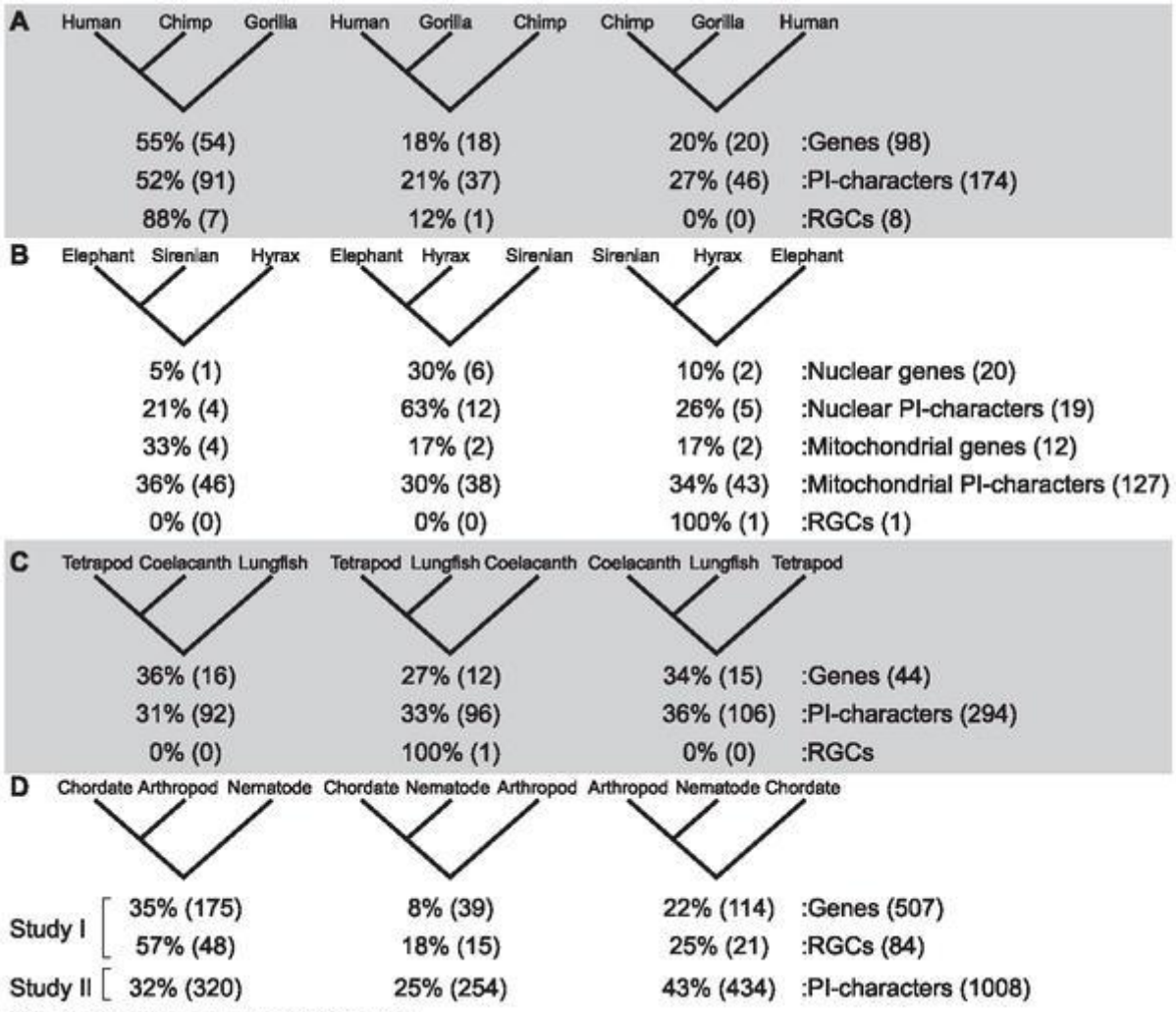
■ The TOL has been molded by cladogenesis and extinction. Starting from a single lineage that undergoes cladogenesis and splits into two, the rate at which the lineages arising from this cladogenetic event undergo further cladogenetic events determines the lengths of the nascent stems. Once these stems have been generated, the only process that can modify their lengths is extinction. At its core, the elucidation of evolutionary relationships is the identification, through statistical means, of the tree's stems



(A) Early in a clade's history (gray box), the number of cladogenetic events is smaller and the length of stems larger in tree-like (left) relative to bush-like clades (right).
(B) In the absence of homoplasy, the number of PICs (parsimony informatice characters) for a stem is proportional to its time span; many PICs (rectangles) accumulated on the long stem *x* (left), whereas few PICs accumulated on the short stem *y* (right).
(C) When the stem time span is long, the effect of homoplastic characters (crosses supporting a clade of species A and C and bullets supporting a clade of species B and C) is not sufficient to obscure the true signal (left). In contrast, the same number of homoplastic characters is sufficient to mislead reconstruction of short stems (right), because the number of homoplastic characters shared between species A and C (three crosses in each of the two species) is larger than the number of true PICs (two rectangles).

Homoplasy, homoplastic = the probability of several species acquiring the same nucleotide or amino acid independently.

From: Bushes in the Tree of Life, Rokas A, Carroll SB *PLoS Biology* Vol. 4, No. 11, e352 doi:10.1371/journal.pbio.0040352



DOI: 10.1371/journal.pbio.0040352.g002

(A) The human/chimpanzee/gorilla tree (5–8 million years ago).
(B) The elephant/sirenian/hyrax bush (57–65 million years ago).
(C) The tetrapod/coelacanth/lungfish bush (370–390 million years ago).
(D) The metazoan superbush (>550 million years ago).

In each panel, the three alternative topologies for each set of taxa are shown. Below each topology, the percentage and number (in parentheses) of genes, PICs, and RGCs (rare genomic changes) supporting that topology are shown (when available). Numbers of genes supporting each topology in (A), (C), and (D) are based on maximum likelihood analyses; numbers in (B) are based on parsimony. The observed conflicts are not dependent on the optimality criterion used; similar results have been obtained by analyses of the data under a variety of widely used optimality criteria. A fraction of genes in each panel is uninformative: (A), 6 of 98 genes; (B), 9 of 20 nuclear genes; (C), 1 of 44 genes; and (D), 179 of 507 genes.

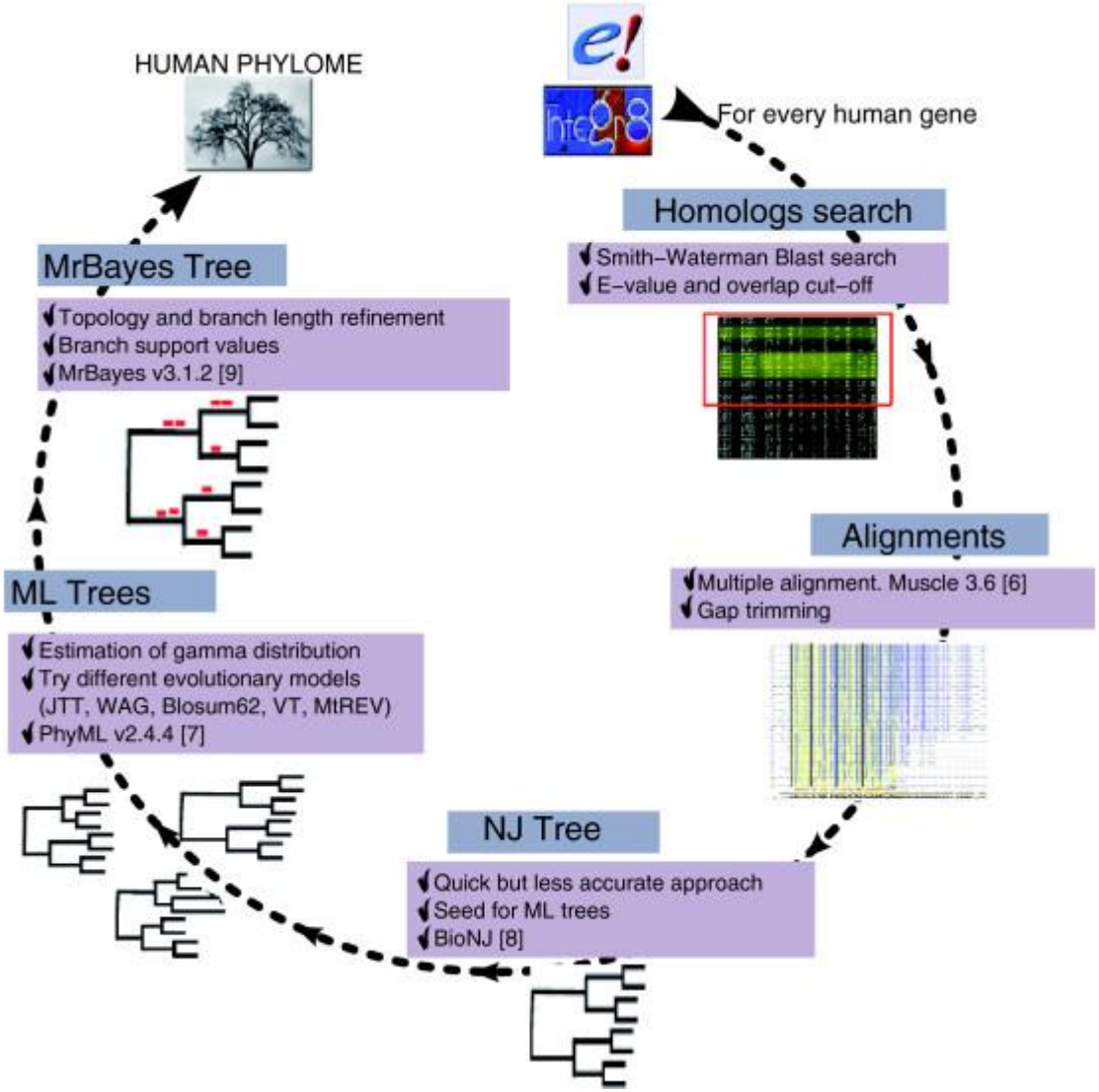
From: Bushes in the tree of life, Rokas A, Carroll SB *PLoS Biology* Vol. 4, No. 11, e352 doi:10.1371/journal.pbio.0040352

4) The phylome

- Complementing the concepts
  - transcriptome
  - proteome
  - interactome
  - metabolome

- Reconstruction of the evolutionary histories of all genes encoded in a genome

- The human phylome
  - *Genome Biology 8:R109 (2007)*
  - proteins encoded by 39 publicly available eukaryotic genomes



Schematic representation of the phylogenetic pipeline used to reconstruct the human phylome. Each protein sequence encoded in the human genome is compared against a database of proteins from 39 fully sequenced eukaryotic genomes to select putative homologous proteins. Groups of homologous sequences are aligned and subsequently trimmed to remove gap-rich regions. The refined alignment is used to build a NJ tree, which is then used as a seed tree to perform a ML likelihood analysis as implemented in PhyML, using four different evolutionary models (five in the case of mitochondrially encoded proteins). The ML tree with the maximum likelihood is further refined with a Bayesian analysis using MrBayes. Finally, different algorithms are used to search for specific topologies in the phylome or to define orthology and paralogy relationships.

From:
Huerta-Cepas *et al. Genome Biology* 2007 8:R109 doi:10.1186/gb-2007-8-6-r109

5) Phylogenetic inference may form a framework for understanding the evolution of characters, e.g. humanenness....

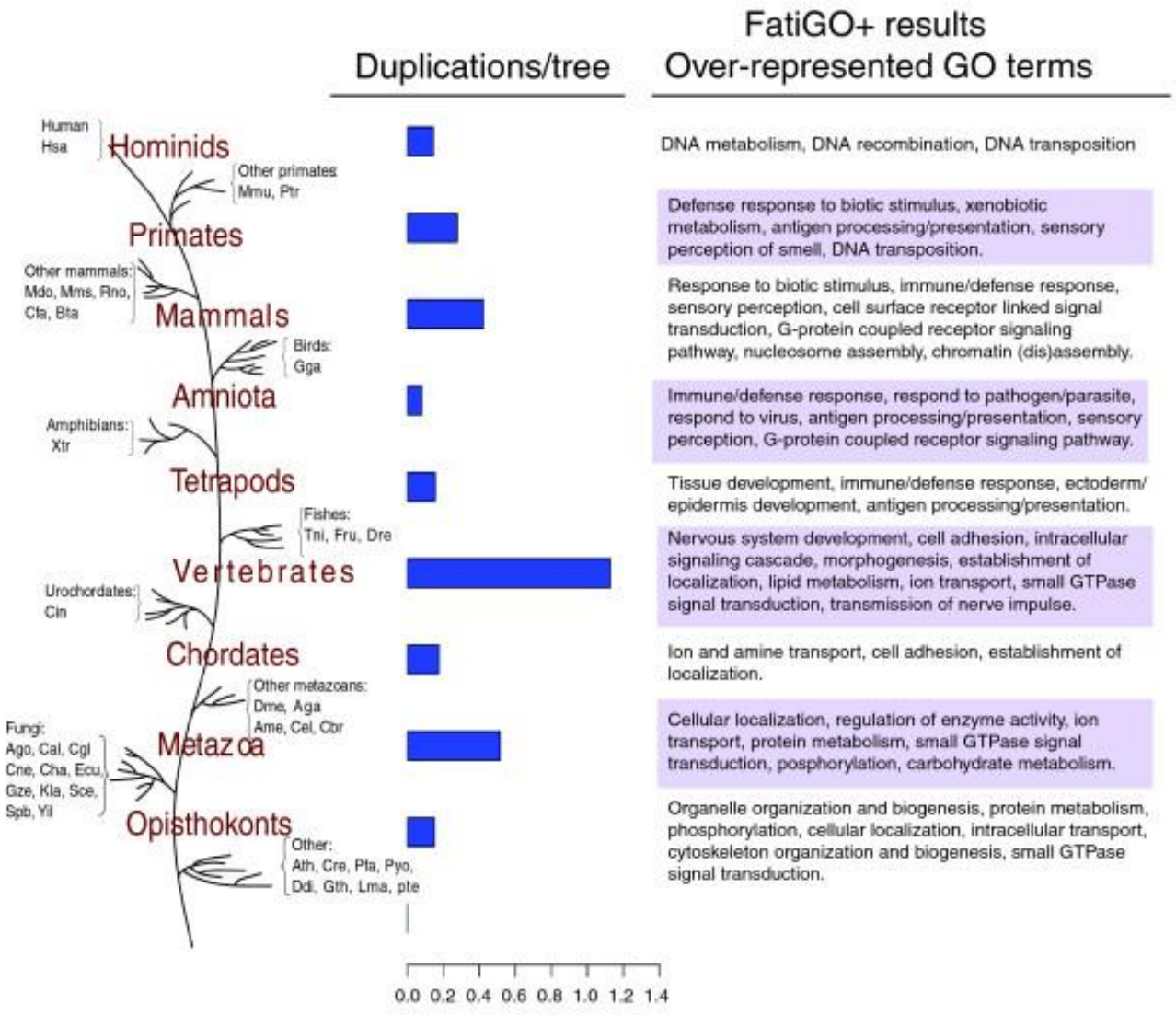
■ During the course of evolution, gene families have increase their size through events of gene duplication.

■ These events may correspond to massive duplications affecting many genes in the genome at the same time, such as in whole genome duplications (WGDs) or may be restricted to chromosomal segments or specific genes.

■ Not only recent genomics surveys have provided evidence for the abundance of duplicated genes in all organisms, but it has also been observed that gene duplication is often associated with processes of neo-functionalization and/or sub-functionalization.

■ To quantify the extent of gene duplication that has occurred in the lineages leading to human, the phylogenetic trees have been scanned to find duplication events, which have been mapped onto a species phylogeny that marks the major branching points in the lineage leading to hominids (next page).

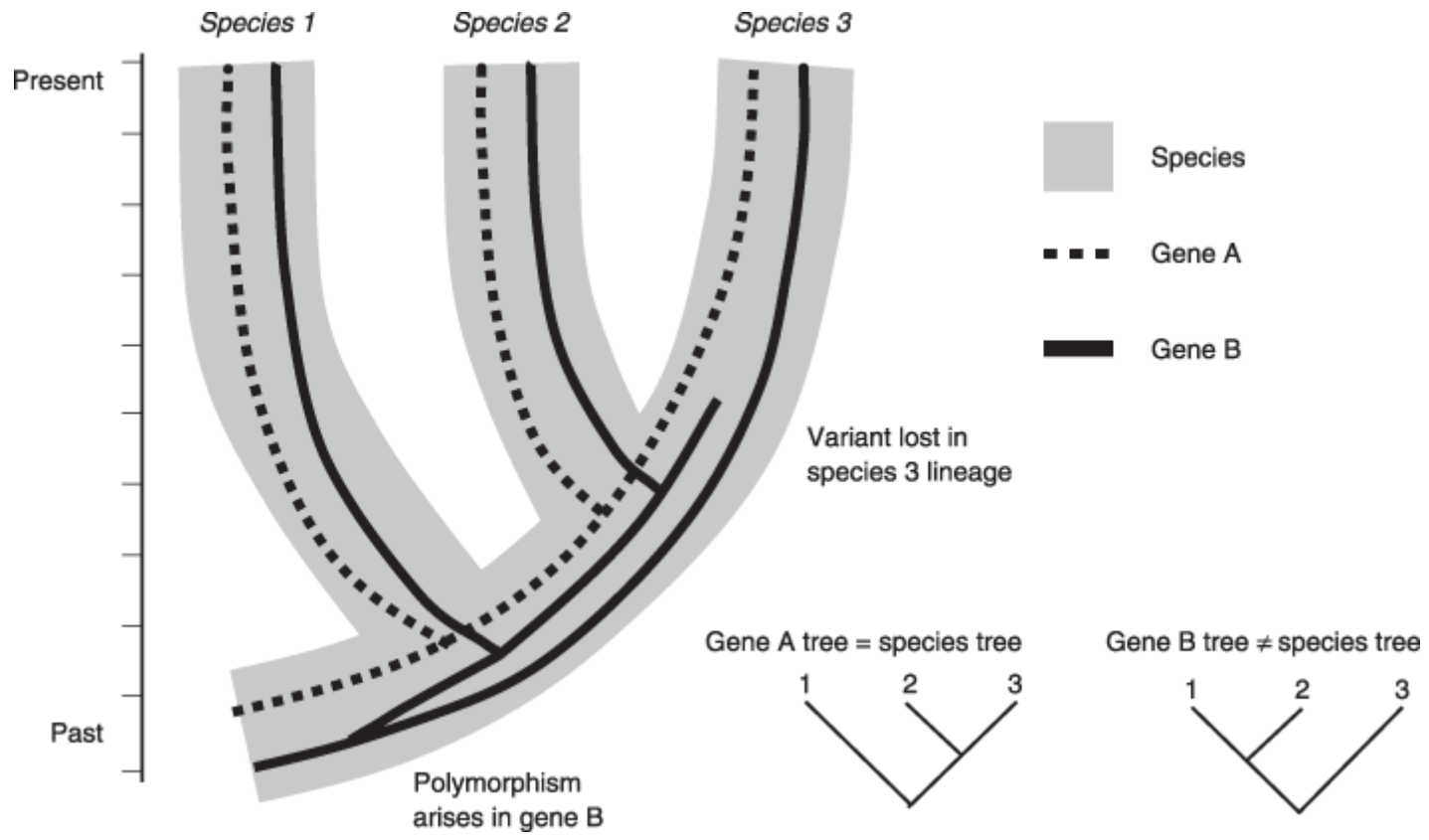
■ High duplication rates in the lineages leading to mammals, primates and hominids. This suggests that duplications have played a major role in the evolution of these groups.



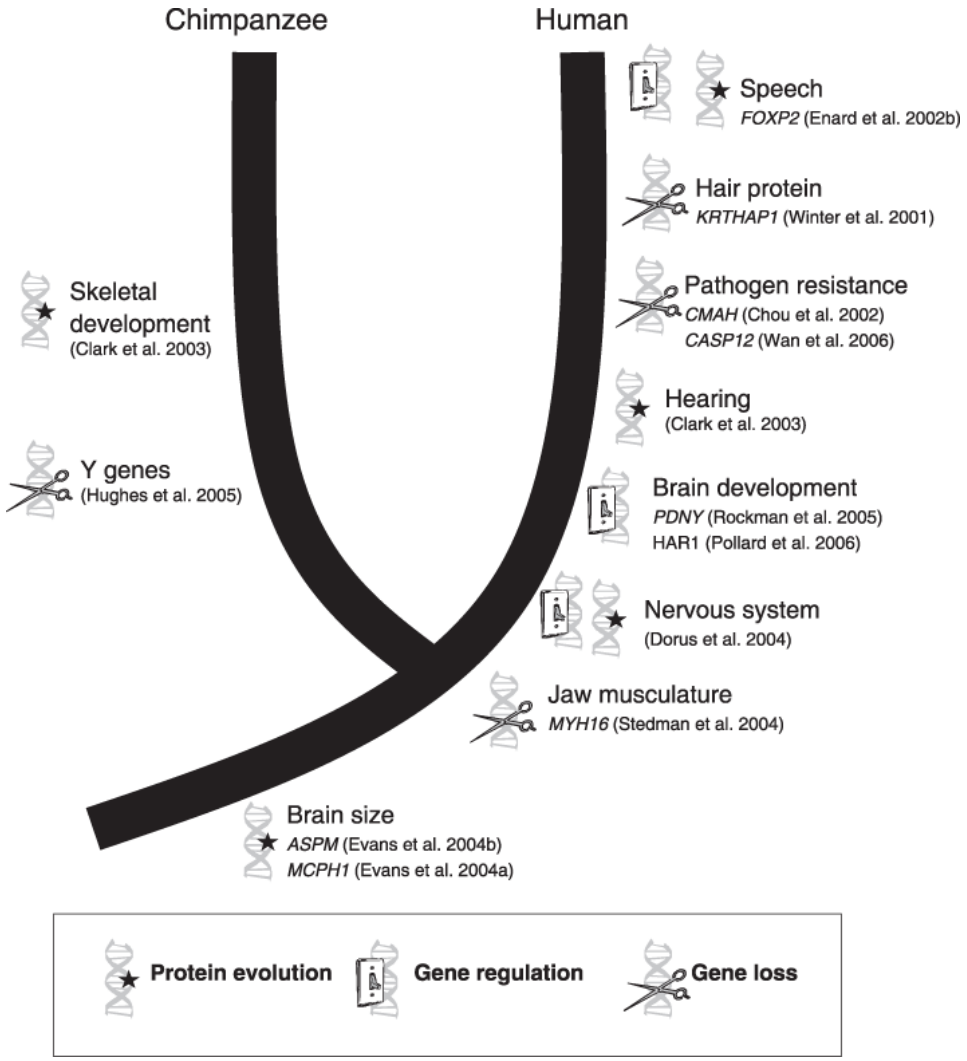
Estimates for the number of duplication events occurred at each major transition in the evolution of the eukaryotes.
Horizontal bars indicate the average number of duplications per gene. Boxes on the right list some of the GO terms of the biological process category that are significantly over-represented compared to the rest of the genome in the set of gene families duplicated at a certain stage.

GO = gene ontology

Gene trees and species trees may tell different stories

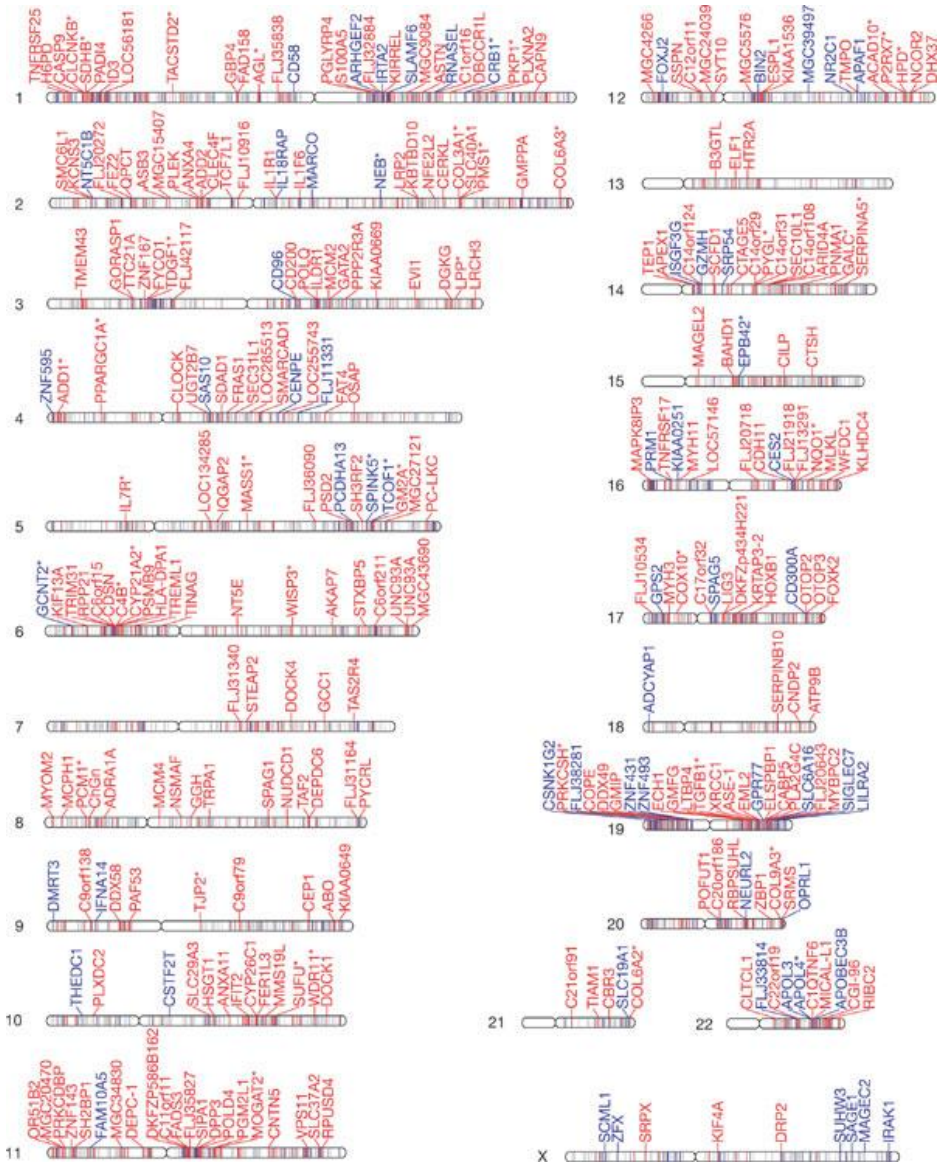


What is known about human-specific characters



- "Protein evolution" – related characters usually mean that accelerated evolution, "positive selection", has been noticed.

- This means that the phylogeny on the basis of such a gene is biased, i.e. has too long branches as compared to the phylogeny which is known to reflect evolution of e.g. "neutral" characters (markers which measure elapsed time).



- Human chromosomes, some genes shown.

- Genes with blue colour are those in which positive selection has been inferred.

- This field of science is very active and is one example of the utilization of phylogeny inference as a tool – the scientific question is not:

  ”what is the phylogeny”

but is:

” the species phylogeny should be .... however, it is not (too long branches, for example), why? Maybe this gene has experienced high rate of evolution and is thus one marker for something specific that has occurred in the history of human lineage?