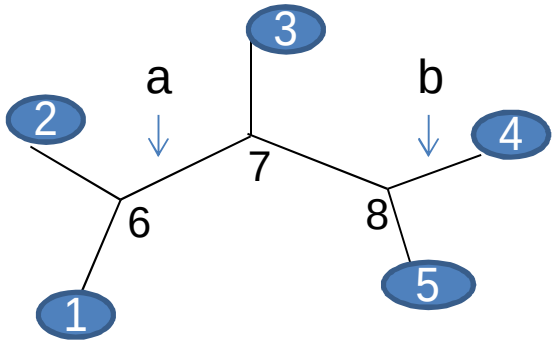
.

SEQUENCE DATA-ANALYSIS

Clustering sequences – phylogenetic trees
Basics of molecular evolution

BASIC TERMINOLOGY

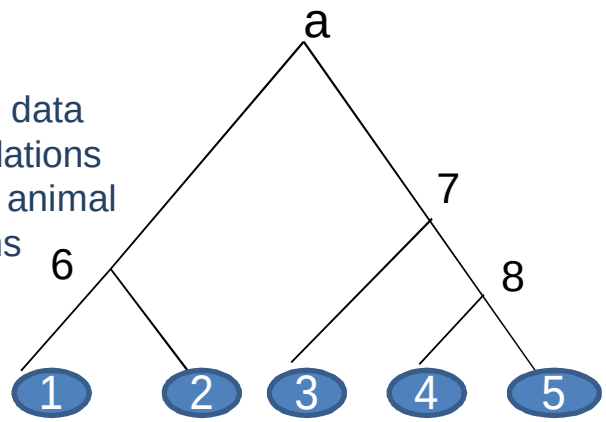
- **Leaves, external nodes** 1,2,3,4,5 are observations which may be, depending on the situation, sequences from different species, populations etc. They are often called **OTUs = Operational Taxonomic Units**. Internal nodes 6,7,8 are hypothetical sequences in ancestral units

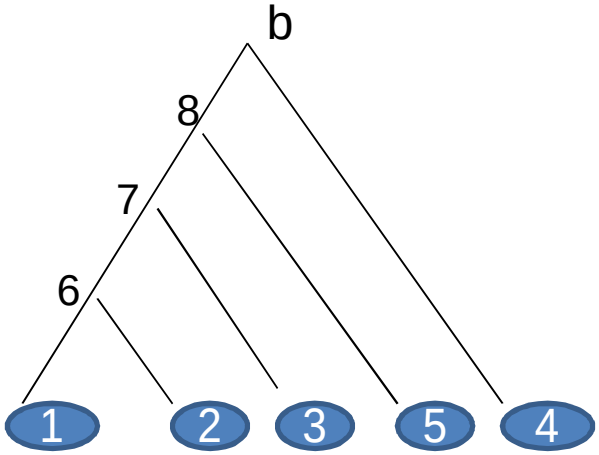


- The tree is **unrooted**.

- In case evidence exists for depicting the root (for example, a or b), a **rooted tree** 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 TOPOLOGIES

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

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

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 IN PHYLOGENY INFERENCE

- Parsimony, **Occams razor**, a philosophical concept.

  Monk William of Ockham (1280-1350):

  "*Entitia non sunt multiplicanda praeter necessitate", entities should not be multiplied more than necessary,*

  "The best hypothesis is the one requiring the smallest number of assumptions"

- The principle of *maximum parsimony* (MP) in phylogeny inference involves the identification of a tree topology that requires the *smallest number of changes* to explain the observed differences. The shortest pathway leading to these is chosen as the best tree.

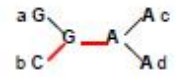
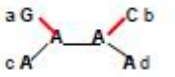
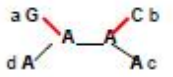
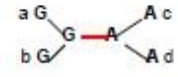
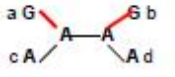
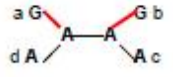
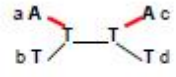
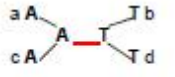
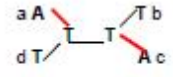
- Two subproblems:
  - Determining the amount of character change, or tree length, required by any given tree.
  - Searching over all possible tree topologies to find the tree that minimize this length.

INFORMATIVE AND UNINFORMATIVE SITES FOR PARSIMONY ANALYSIS

● An example, four OTUs (operational taxonomic units), nine 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 can form three possible unrooted trees, I, II, III

	tree I ((a,b),(c,d))	tree II ((a,c),(b,d))	tree III ((a,d), (b,c))	**NEWICK-formats**
site 3				
site 5				
site 9				

A nucleotide site is informative only if it favors a subset of trees over the other possible trees. *Invariant* (1, 6, 8 in the example) 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 two 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

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

● 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 choosing 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.

FITCH´S PARSIMONY

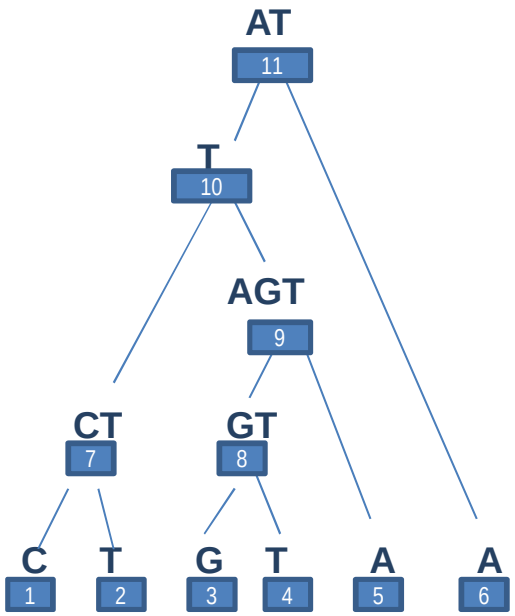
- *The rule*:
  - The set at an interior node is the intersection of its two immediately descendant sets if the intersection is not empty.
  - Otherwise it is the union of the descendant sets.
  - For every occasion that a union is required to form the nodal set, a nucleotide substitution at this position must have occurred at some point during the evolution for this position. Thus, counting the number of unions gives the minimum number of substitutions required to account for descendant nucleotides from a common ancestor, given the phylogeny assumed at the outset.

- The example next page (taken from textbook W-H Li, *Molecular evolution,* 1997) considers the case of six OTUs, and one particular *site*, at which the nucleotides are

	....*site*.....
OTU 1	C
OTU 2	T
OTU 3	G
OTU 4	T
OTU 5	A
OTU 6	A

- The six OTU´s have five (unknown, to be inferred) ancestors: 7, 8, 9, 10, 11.

FITCH´S PARSIMONY, EXAMPLE



● One possible tree topology for the example site (previous page). The nucleotide at nodes 7, 8 and 9 cannot be determined uniquely under the parsimony rule. At node 10 T is chosen as it is shared by the sets at the two descendant nodes, 7 and 9. The nucleotide at node 11 cannot be determined uniquely. Parsimony requires it to be either A or T.

● At nodes 7, 8 and 10 nucleotide A could be included as a possible ancestral nucleotide because A is a possible common ancestral nucleotide (node 11) of all the six OTUs.

● **NEWICK-format,** the commonly agreed format for phylogeny topologies (not only parsimony), of the tree is **(((1,2) ((3,4) 5)) 6)**

● Consider other possible topologies for the example site. For example:

(((2,4) 1) (3 (5,6)))

Inferred nucleotides at nodes 7, 8, 9, 10 and 11 ?

FITCH´S PARSIMONY

- In the example tree (previous page), the nucleotide at node 10 is the intersection of the sets at nodes 7 and 9. The set at node 9 is the union of the sets at nodes 8 and 5.

- Counting the number of unions gives the minimum number of substitutions required to account for descendant nucleotides from a common ancestor, given the phylogeny assumed at the outset. In the example this number is 4.

    - There are many other alternative trees, each of which requires 3 substitutions. Thus, unlike the case of four OTUs, an informative site may favor many alternative trees.

PARSIMONY ANALYSES

● 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.

In the course Biometry and bioinformatics II we go further with parsimony analyses, as well as other methods, maximum likelihood and bayesian phylogenetics.

In BB_I assignments 1, 2 and 3 we use only the distance matrix based methods UPGMA and neighbor joining, and parsimony for assignment 1, by using the software MEGA5.

PHYLOGENY METHODS BASED ON DISTANCE MATRICES

● Distances are computed for all pairs of OTUs and a phylogenetic tree is constructed by considering the relationships among these distance values. Distances are numbers of nucleotide substitutions between sequences. Distances are simple p-distances or based on some nucleotide substitution model.

● The **unweighted pair-group method with arithmetic mean, UPGMA,** is the simplest method for tree construction. It was originally developed in 1950´s for constructing taxonomic **phenograms**, i.e. trees that reflect the phenotypic similarities between OTUs.

● The most widely used method is the **neighbor-joining, NJ,** algorithm, developed in 1987. NJ tree is usually the first tree constructed for a given research problem, followed by other methods, parsimony, maximum likelihood and bayesian trees.

DISTANCES

- 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

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

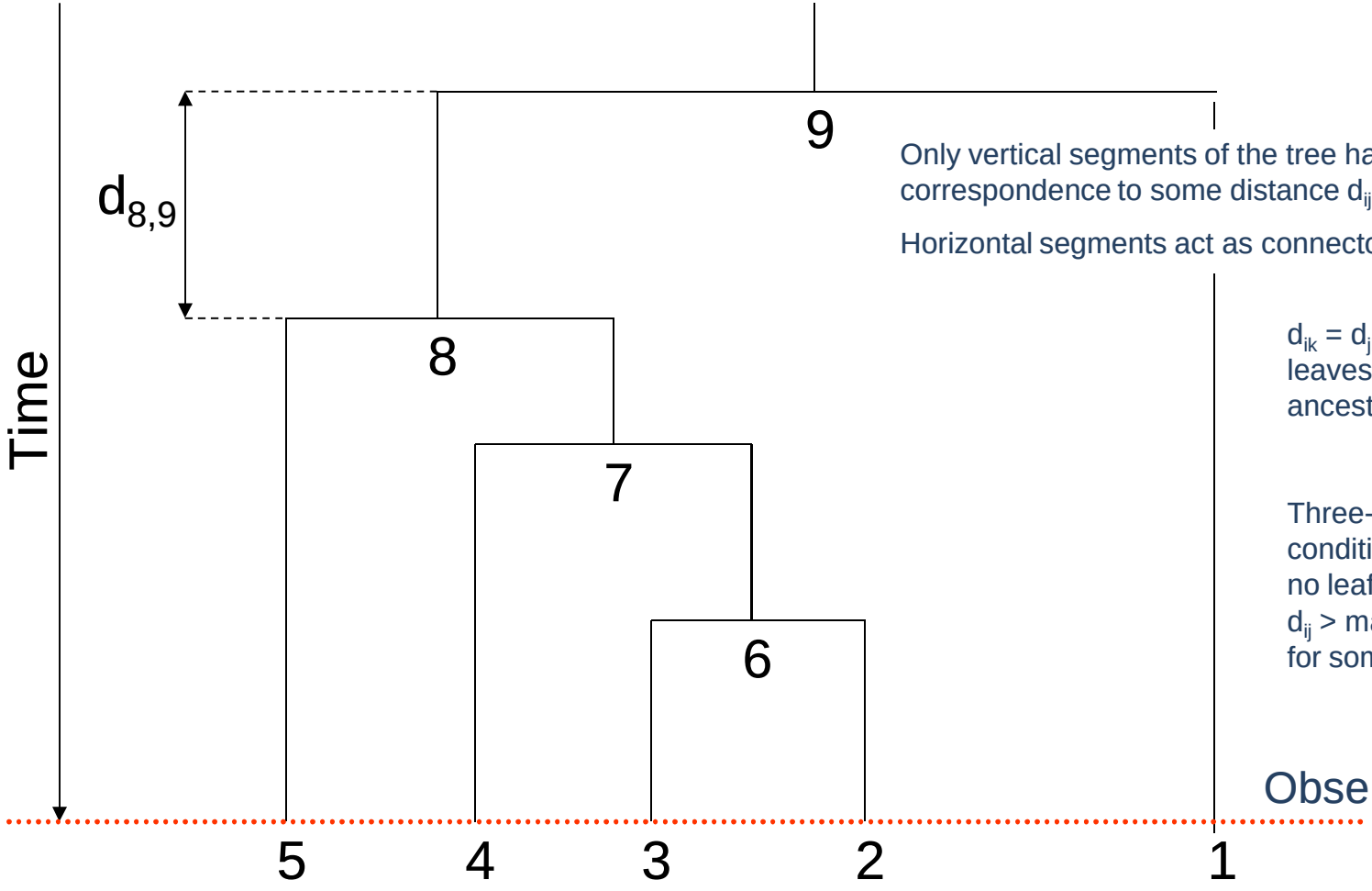
- A rooted additive tree is called an *ultrametric tree*, if the distances between any two leaves i and j, and their common ancestor k are equal

  $$d_{ik} = d_{jk}$$

- 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 -method

- The UPGMA method employs a sequential clustering algorithm, in which local topological relationships are inferred in order of decreasing similarity and a phylogenetic tree is built in a stepwise manner.
  - The two OTUs that are most similar to each other, i.e. have the shortest distance, are first identified.
  - The two OTUs are treated as a new single OUT, a **composite OTU**
  - Then, from among the new group of OTUs, the pair with highest similarity is identified, and so on, until only two OTUs are left.

- Consider a case of four OTUs, A, B, C and D. The pairwise distances are given by the following matrix

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

- Let us assume that d_{AB} has the smallest value. Then, A and B are the first to be clustered, and the branching point is positioned at a distance of $d_{AB}/2$.

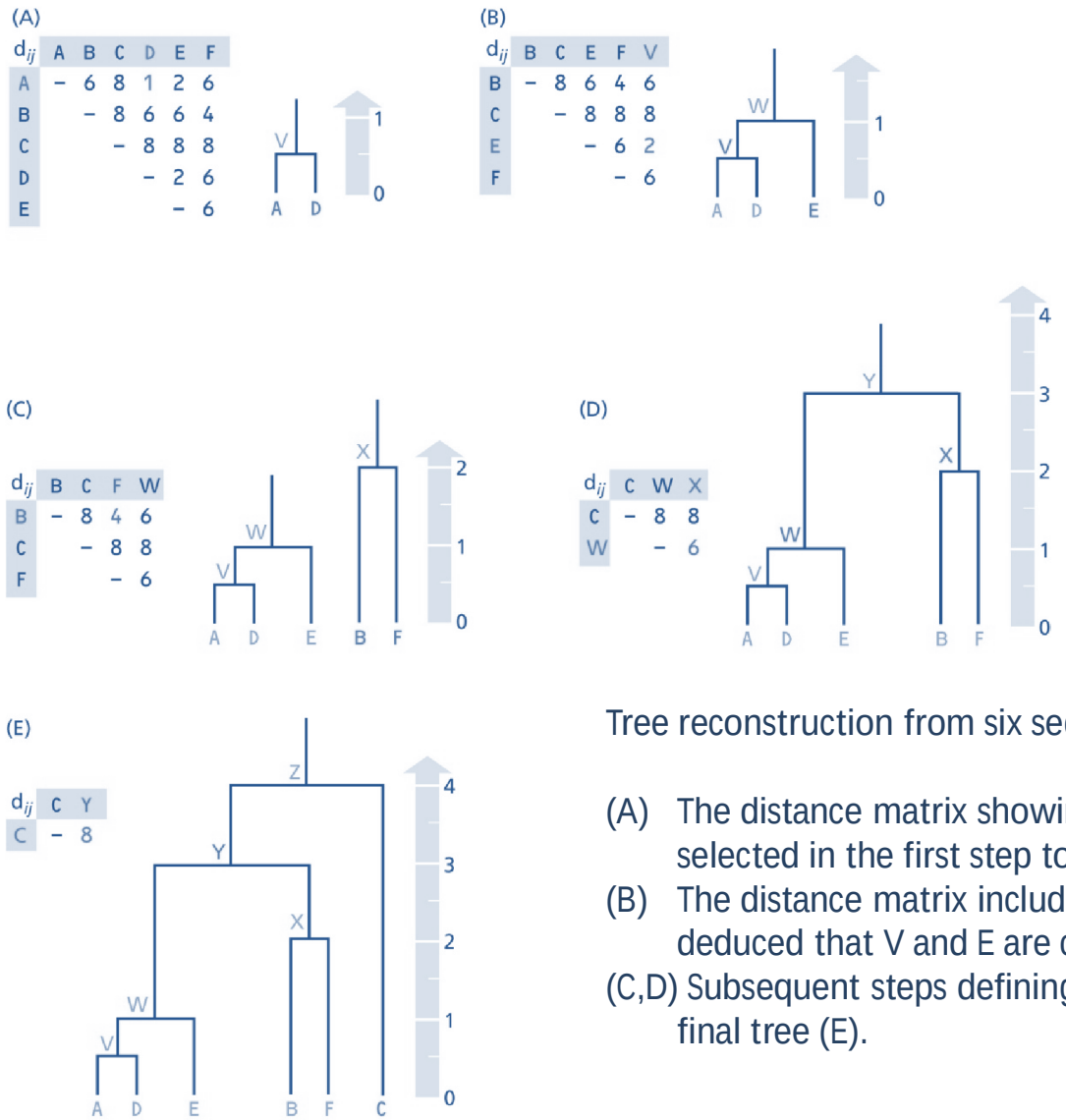
	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

- A new distance matrix is computed by using AB composite OTU.

$d_{(AB)C} = (d_{AC} + d_{BC}) / 2$

$d_{(AB)D} = (d_{AD} + d_{BD}) / 2$

UPGMA –method - a worked example

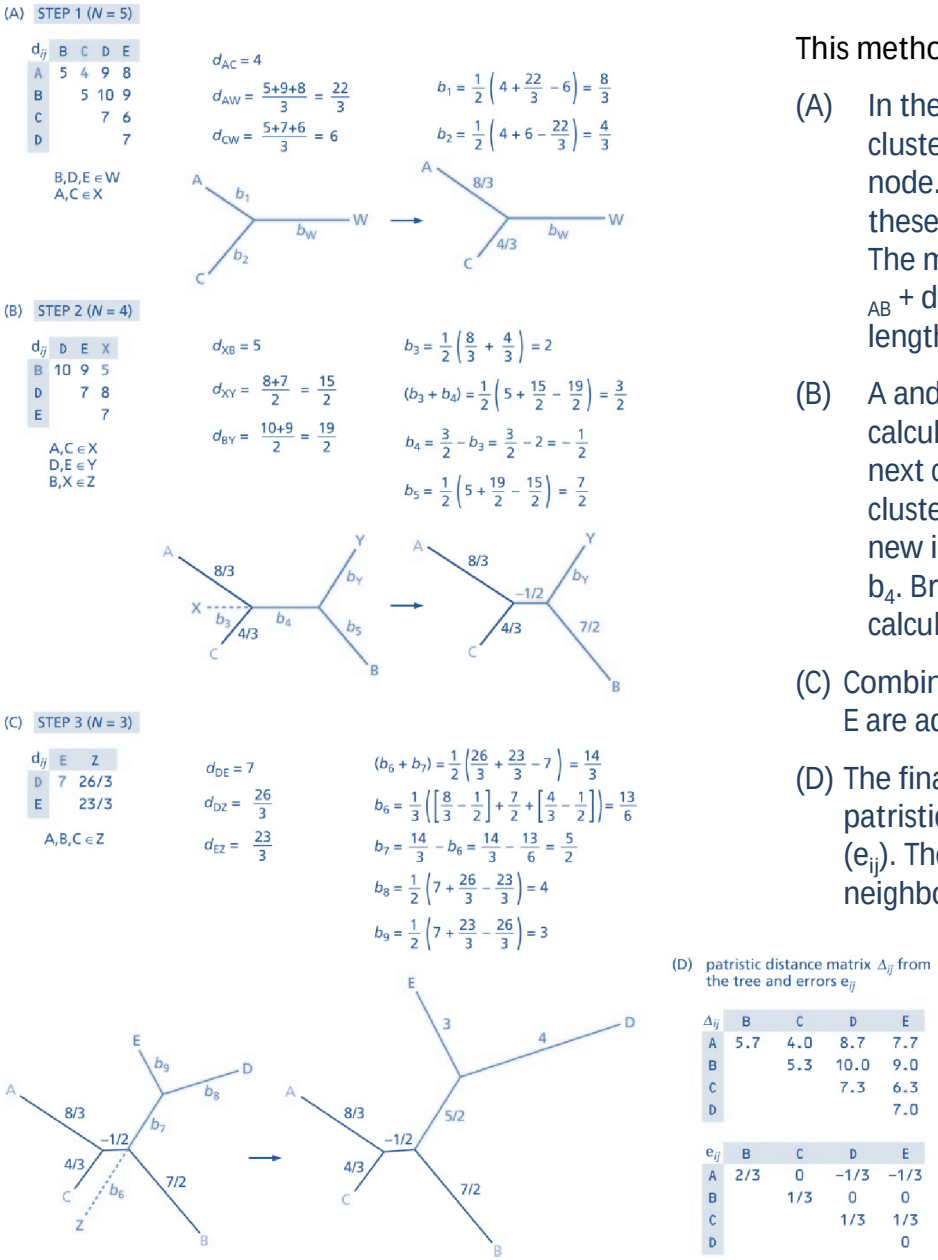


Tree reconstruction from six sequences, A-F.

(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.

FITCH-MARGOLIASH METHOD - a worked example



This method is not in MEGA5-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.

(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.

NEIGHBOR JOINING, NJ, 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

- *NJ is easy to use, and understand the results. However, the algorithm is not easy stuff.*

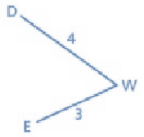
In case you want to learn more about NJ, see the attached original paper.

NJ-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

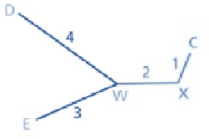
D and E are neighbors through internal node W with $d_{DW} = \frac{1}{2}\left(7 + \frac{33-30}{3}\right) = 4$ and $d_{EW} = 7 - 4 = 3$.



(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

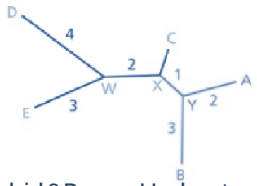
C and W are neighbors through internal node X with $d_{CX} = \frac{1}{2}\left(3 + \frac{12-14}{2}\right) = 1$ and $d_{WX} = 3 - 1 = 2$.



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

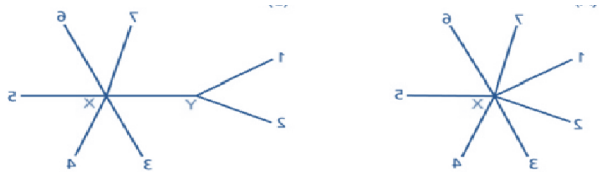
	d_{ij} B	X	U_i	δ_{ij} B	X	
A	5	3	8	−12	−12	A
B		4	9		−12	B
X			7			X

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).

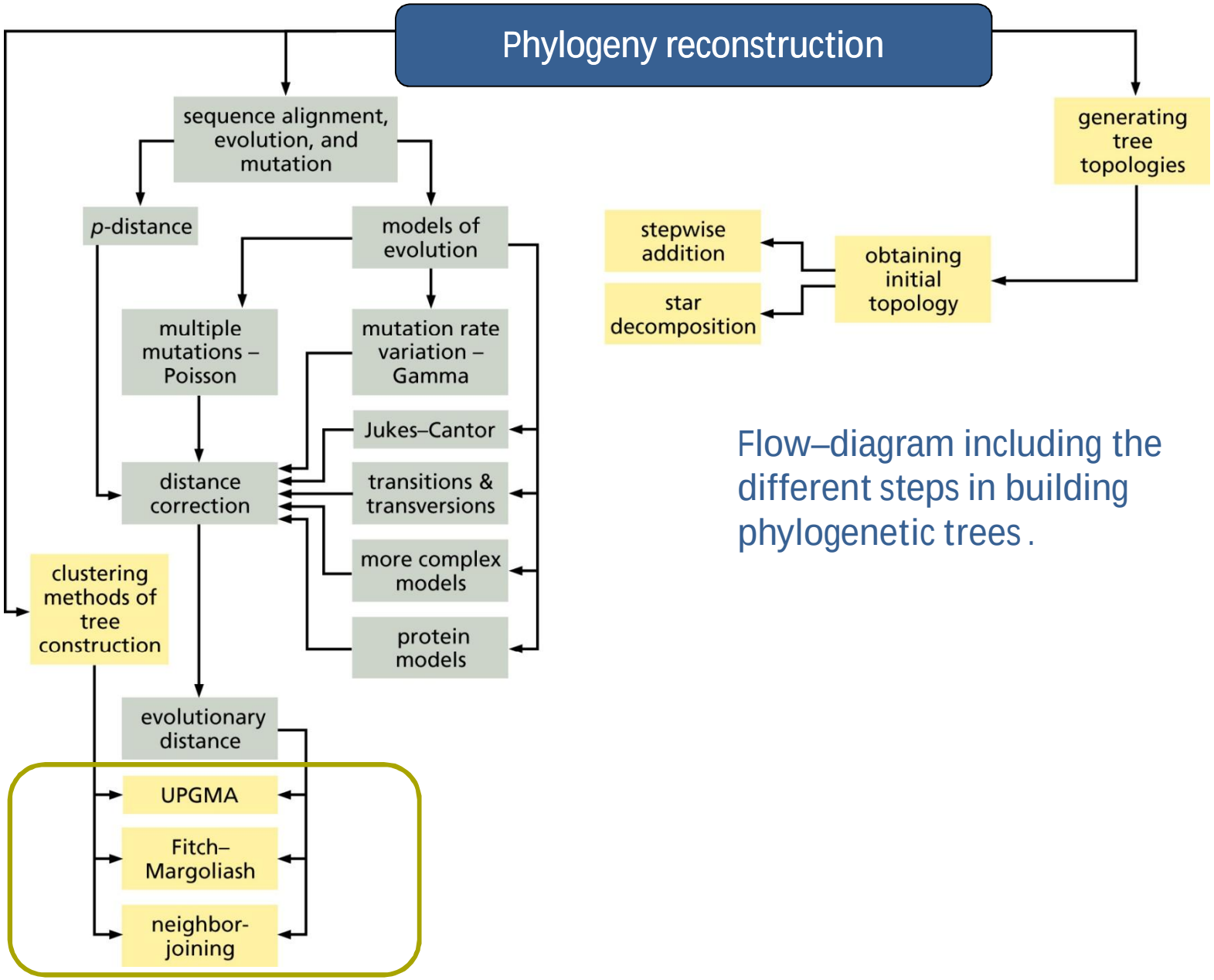
The first step:



(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 branch 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

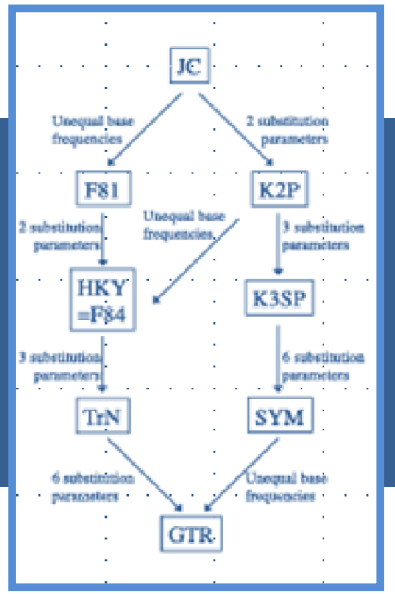


Flow–diagram including the different steps in building phylogenetic trees .

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

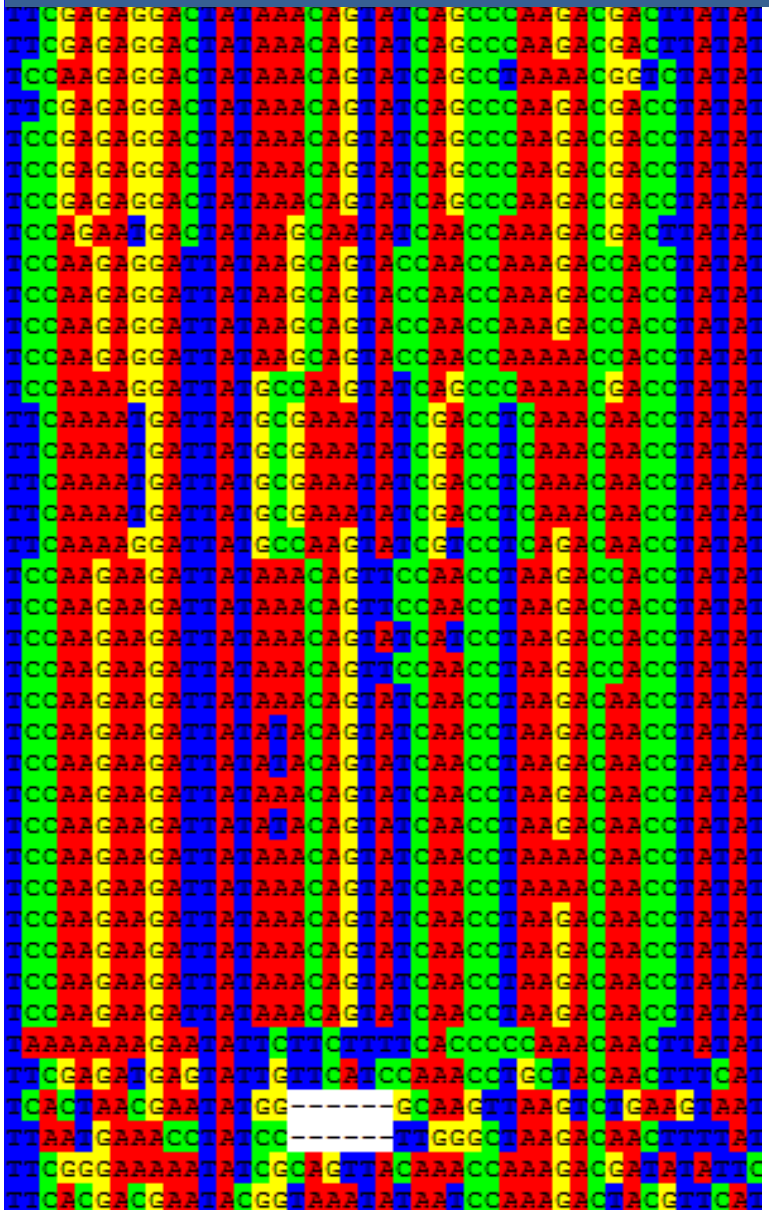
http://evolution.genetics.washington.edu/phylip/software.html

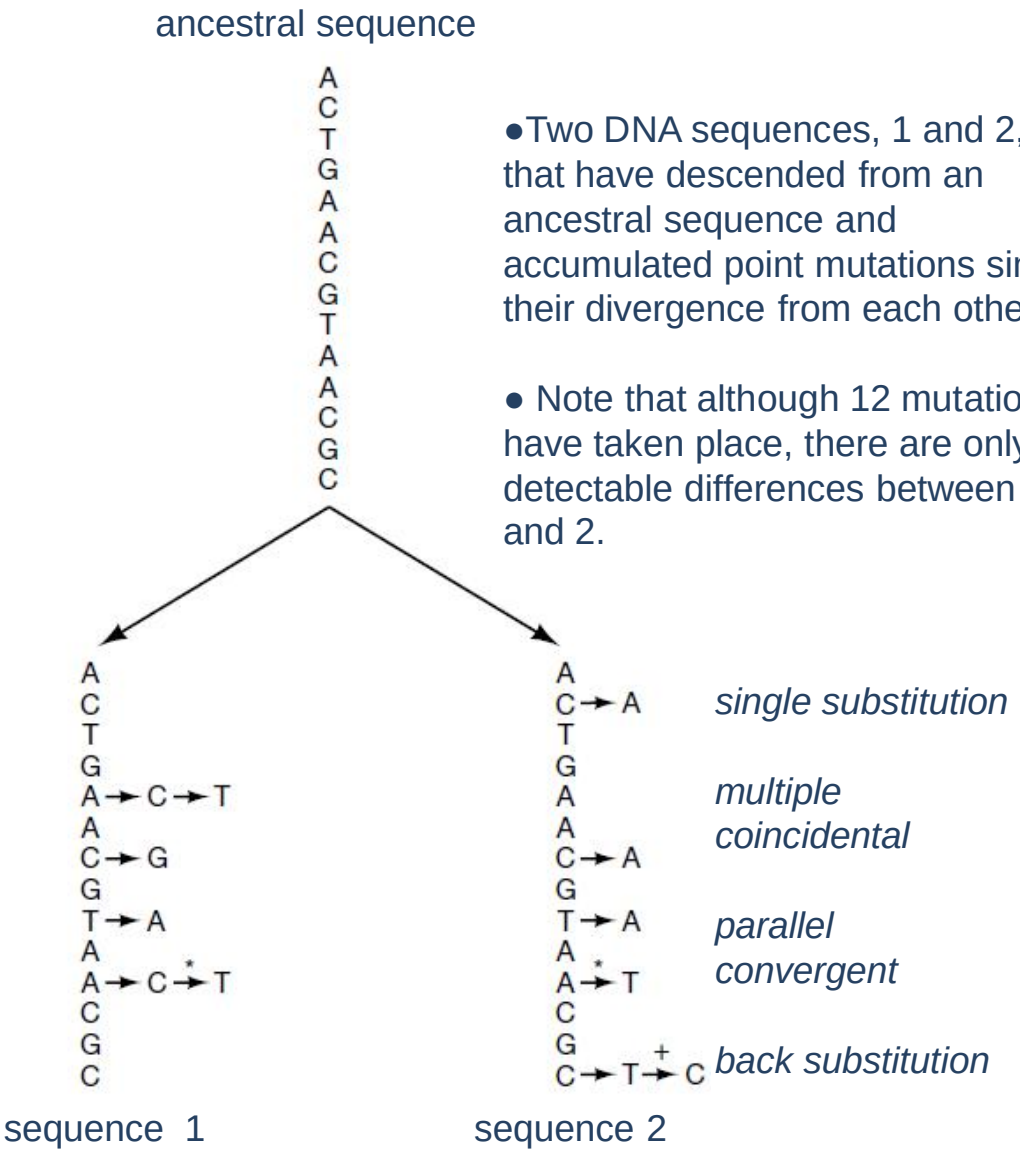
MODELLING SEQUENCE EVOLUTION



- Sequence differences that we observe now (among existing species or other items which are compared) are products of past mutation events, **substitutions**.
- Understanding the substitution process needs modelling.
- Historically, modelling started by the concept *molecular clock.*
- Parametrization: one parameter, *Jukes-Cantor model* → more realistic models

CURRENT DIFFERENCES ARE PRODUCTS OF PAST SUBSTITUTION EVENTS





- Two DNA sequences, 1 and 2, that have descended from an ancestral sequence and accumulated point mutations since their divergence from each other.
- Note that although 12 mutations have taken place, there are only 3 detectable differences between 1 and 2.

THE CONCEPT ´MOLECULAR CLOCK´

● In **1904 G.H.F. Nuttal** measured the amount of precipitate of normal blood serum from great apes, monkeys and some other mammals. His crude method, using rabbit antiserum directed against human serum, indicated that the amount of precipitate declined with the paleontological distance from humans

● The immunological method was later refined and played an important role in reconstructing primate phylogenies by **Morris Goodman** (**1962**, Hum Biol 34: 104-150, *Evolution of the immunologic species specificity of human serum proteins*).

● In **1962 Emile Zuckerkandl** and **Linus Pauling** worked on hemoglobin evolution and expressed the idea of *molecular anthropology* as a new discipline . They calibrated the amino acid substitution rate in mammalian hemoglobins and estimated the divergence times of hemoglobins (*Molecular disease, evolution and genic heterogeneity,* pp 189-225 in Horizons in Biochemistry).

● The historical conclusion was: the average rate of molecular evolution is constant and observations can be used for estimating time scales.

HISTORICAL RESULTS

Million Years

440

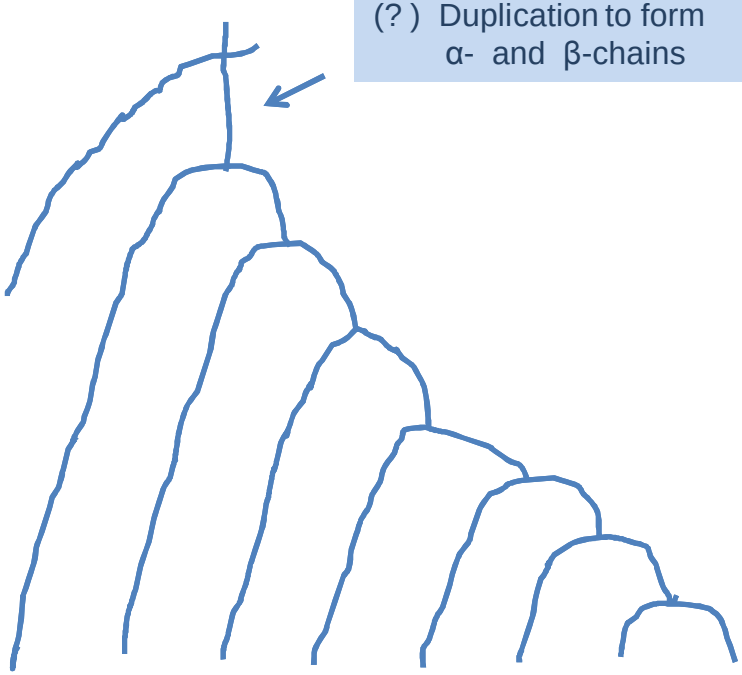
400

350

270

225
180
135
70

(?) Duplication to form α- and β-chains



■ Percent amino acid differences when the α-hemoglobin chains are compared among eight vertebrates together with their times of divergence (on the basis of paleontological views).

■ In mammals the α-chain consists of 141 amino acids.

■ Strong parallelism between divergence time and amino acid differences.

Hemoglobin α-chain
% amino acid differences

	Shark	Carp	Newt	Chick	Echi	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp			53.2	51.4	53.6	50.7	47.9	48.6
Newt				44.7	50.4	47.5	46.1	44.0
Chicken					34.0	29.1	31.2	24.8
Echidna						34.8	29.8	26.2
Kangaroo							23.4	19.1
Dog								16.3
Human								

HISTORICAL LOOK AT NUCLEOTIDE CHANGE MODELLING

● In 1965 Emile Zuckerkandl and Linus Pauling proposed the theory of *molecular clock: the rate of molecular evolution is approximately constant over time for all the proteins in all lineages.* This was the starting point for modelling approach to provide understanding for the first results concerning amino acid sequence differences between different animal species, at certain proteins.

● It has been known for a long time that the constancy is not true and that time of divergence between sequences *cannot* be dated simply by measuring the number of changes between sequences. Mutation rates vary among and within genomes, being affected by many factors such as chromosomal position, G+C (vs. A+T) content etc. Molecular clocks tick at different rates in different biological contexts. *We come back to this later during the course*

● **The proposition by Thomas Jukes and Charles Cantor** (in **1969**, a response to Zuckerkandl-Pauling theory), was important as regards subsequent achievement in modelling the DNA substitution process. Jukes and Cantor formulated **a stochastic model in which nucleotide substitutions occur at an equal rate.**

● Subsequently the model has been made more realistic by taking into account increased empirical knowledge about nucleotide substitutions.

HISTORY: JUKES TELLS HOW THE IDEA AROSE

http://www.garfield.library.upenn.edu/classics1990/A1990CZ67100002.pdf

How Many Nucleotide Substitutions *Actually* Took Place?

Thomas H. Jukes
Department of Biophysics and Medical Physics
University of California
Berkeley, CA 94720

In 1965 I met Charles R. Cantor, who was 23 years old and was a graduate student in chemistry at the University of California. We started talking about molecular evolution and about comparing the polypeptide chains of homologous proteins. Charles said that a computer program should be written for searching for evidence of this, but that he was frightfully busy working on his PhD thesis, and he had no time for this. The next day he had written the program, and in 1966 we wrote two notes on its use. We resolved to write a textbook on molecular evolution together, and Charles left for Columbia University as an assistant professor in 1966. I received a request from Hamish N. Munro for a chapter in his forthcoming volume III of *Mammalian Protein Metabolism*. He asked us to write on "Evolution of protein molecules," and we sent him the manuscript that was to have been incorporated in our book. It was published in Munro's book in 1969, and the article has 110 printed pages. Citations to our long article relate only to the following short passage in it, written by Charles.

It can be shown that the mean number of base differences at a single position on the mRNA, μ, is related to the observed fraction of residues with single base differences, p, by the expression

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

The equation (1) assumes that all single base changes (nucleotide substitutions) are equally probable and that the frequencies of all four bases in DNA are the same. This gives me the chance to point out that (1) should be called the Cantor equation, not Jukes and Cantor. The formula came into wide use when rapid DNA and RNA sequencing became available. From then on molecular biologists became interested in comparing sequences of homologous genes to study evolution. For example, a portion of the two sequences of human α and β hemoglobin genes is

α gene A C C A A C G T C A A G G C C G
C C T G G G G T A A G G T T

β gene T C T G C C G T T A C T G C C C
T G T G G G G G A A G G T G

showing 12 nucleotide substitutions (40 percent). The mean number of substitutions that has actually occurred is greater than 12, because of revertants, such as A to C to A, and multiple changes, such as A to C to G. Equation (1) corrects for these, and the probable total number of substitutions is 17 (57 percent), *not* 40 percent.

The two genes diverged from a common ancestor at least 4×10^8 years ago. Sharks go back in the fossil record for 400 million years and sharks have α and β hemoglobins (but lampreys do not). The equation tells us that the average rate of substitution per year per nucleotide site is about $0.57 \div (4 \times 10^8) = 1.4 \times 10^{-9}$. We carry with us in every red blood cell the evidence that we are in a line of descent from an ancestor who lived 400 million years ago!

An example of the use of equation (1) is in the article by C.L. Manske and D.J. Chapman.[1] These authors used the equation to correct their comparisons of 5S ribosomal RNA sequences for rev[illegible] and convergent mutations. See also references 2 and 3 for similar usage.

Charles returned to Berkeley in 1989 to direct the human genome project at the Lawrence Berkeley Laboratory.

HISTORICAL LOOK AT NUCLEOTIDE CHANGE MODELLING

● Since 1980´s it has been known that misincorporation errors (mutations) during DNA replication or repair are facilitated if a base is replaced by similar one and thus **transitions** (purine replaced by a purine, or pyrimidine replaced by pyrimidine) occur more frequently than **transversions** (purine replaced by a pyrimidine or vv). Differences in mutation rate tend to decrease TA and CG dimers and to produce an excess of CT and TG dimers, and many other kinds of biased processes (cf. the constancy in the Jukes-Cantor model).

● The development of models of sequence evolution is an active field and there is a large number of models.

● Two main approaches to building models of sequence evolution: An *empirical* one, using properties calculated through comparisons of large numbers of observed sequences (for example, counting apparent replacements between many closely related sequences). Empirical models result in fixed parameter values which are estimated only once and then assumed to be applicable to other datasets (=> easy to use computationally). The alternative approach is to build models *parametrically* on the basis of chemical or biological properties of DNA and amino acids. For example, incorporating a parameter to describe the relative frequency of transition to and transversion substitutions in the sequences studied. Both methods result in **Markov process models. (In Biometry and bioinformatics II we go further with this.)**

JUKES-CANTOR MODEL, ONE PARAMETER

- To study the dynamics of nucleotide substitution, assumptions on the probabilities of substitutions of one nucleotide by another are needed.
- **Assumption: all 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: Assumption that at a certain nucleotide site there is A, $P_{A(0)} = 1$
- 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

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

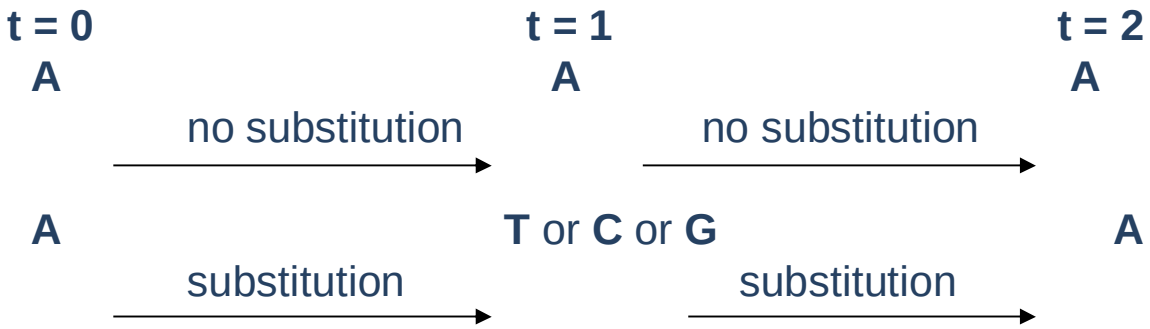
- 3α is the probability of A changing to T, C, or G

JUKES-CANTOR MODEL, ONE PARAMETER

- 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 courses of events:



- The following recurrence equation holds for 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).

JUKES-CANTOR MODEL, ONE PARAMETER

- The amount of change in $P_{A(t)}$ 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.* With this approximation equation (4) is rewritten as

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

- 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 (6) 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

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

JUKES-CANTOR MODEL, ONE PARAMETER

- 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)

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

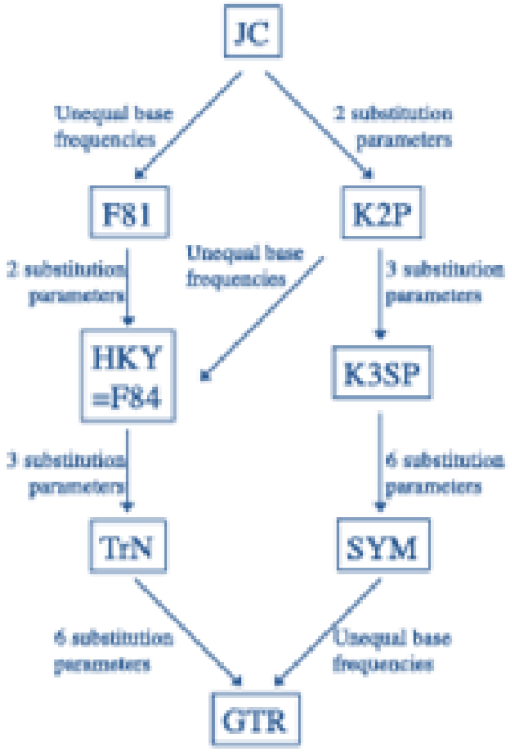
- Since all the nucleotides are equivalent under the Jukes-Cantor model, the general probability, $P_{ij(t)}$, that a nucleotide will become *j* at time *t*, given that it was *i* at time 0, equations (9) and (10) give the general probabilities $P_{ii(t)}$ and $P_{ij(t),}$ where $i \neq j$.

TWO PARAMETERS, KIMURA´s MODEL

- The Jukes-Cantor –model was introduced in 1969 when virtually nothing was known about nucleotide substitution
- **In 1980 Motoo 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 – pyrimidine 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**	

FLOW-DIAGRAM OF THE MOST WIDELY USED SUBSTITUTION MODELS



JC	"Jukes-Cantor"
F81	"Felsenstein 81"
K2P	"Kimura 2-Parameter"
K3SP	"Kimura 3-Parameter"
HKY	"Hasegawa-Kishino-Yano"
F84	"Felsenstein 84"
TrN	"Tamura-Nei"
SYM	"Symmetric"
GTR	"General Time Reversible"

- Starting with the simple Jukes-Cantor model, more general models are obtained by allowing unequal nucleotide frequencies and/or more than one substitution parameter. The most general model of this type is the GTR model that allows unequal base frequencies and prescribes a different substitution parameter for each of the six pairs of different nucleotides.

MODEL SELECTION

● Two approaches for model selection:

● **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.

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

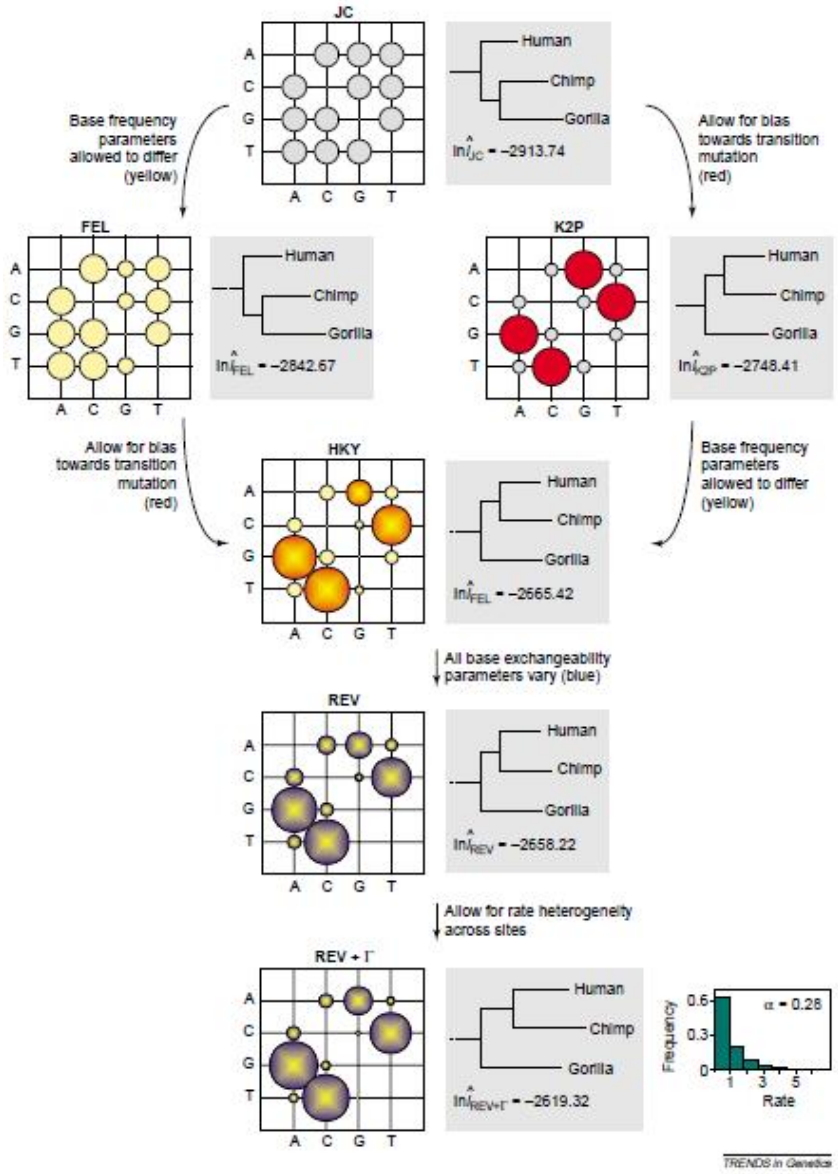
● For example, a transition / transversion bias estimated as $\kappa = 2.3 +/- 0.16$ effectively excludes the possibility that there is no such bias ($\kappa = 1$), whereas $\kappa = 2.3 +/- 1.6$ does not.

● **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.

Statistical model selection will be in the program of Biometry and bioinformatics II.

.RELATIONSHIPS AMONG SUBSTITUTION MODELS – AN EXAMPLE



- The sequence studied is a part of mitochondrial genome. Mitochondrial sequences are known to have highly biased transitions vs. transversions.

- The models JC , FEL, K2P, REV, REV +Γ (the inferred shape parameter value is α=0.28) are presented in a flowchart showing relationships between them. For each model, the matrix of rates of substitutions between nucleotides is represented by a bubble plot where the area of each bubble indicates the corresponding rate. 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.

- For the REV+Γ model 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.

- 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.

Assignment 1

By using MEGA5, construct UPGMA, neigbor joining and maximum parsimony trees. Compare the results.
Compare neighbor joining trees by using p-distance and Jukes-Cantor model.

Assignment 2

Construct only the neighbor joining tree by using the Jukes-Cantor model

Assignment 3

Construct only the neighbor joinig tree by using the Jukes-Cantor model

Before you start working with your own data, take the tutorial

"Walk through MEGA"

Read also the "*Short tutorial article*" in course webpage.