Molecular Clock

- If rate of allele substitutions is constant for molecular variants, then this rate can be used as a **molecular clock**, and these predictions can be used to infer time of divergence between sequences
- We can now predict the amount of ***d***ivergence or ***d*** (*genetic **d**istance in the molecular population genetics literature*) as the number of different nucleotide sites between two nucleotide sequences over ***t*** generations, if the neutral mutation rate ***u*** is known:

$$d = 2ut$$

- It can be solved for $t = d/(2u)$
- ***d*** can be measured from sequence data, and if ***u*** is known, then ***t*** can be also estimated: for example, if $d = 0.02$ (2% divergence) and $u = 10^{-8}$, then ***t***, the time since divergence of the two sequences is $0.02/(2\times10^{-8}) = 10^6$ generations
- Similar, it can be solved for ***u***: $u = d/2t$
- Assuming regular replacement, the difference in amino acid or nucleotide sequence between two species may serve as a **molecular clock**, indicating the time since two species diverged from a common ancestor

Molecular Clock

- How to calculate the amount of protein ***d***ivergence ***d*** (aka d_p, d_{PC} or K_{aa}) the mean number of amino acid substitutions per site?
- Assuming independence at different sites we can estimate ***d*** using the Poisson distribution for number of ***s***ubstitutions ***s***: $Pr_S = e^{-d}d^s/s!$
- The probability of no substitution at any site ($s = 0$) is $Pr_0 = e^{-d}$
- The probability of one or more substitutions is $Pr_{>1} = 1 - e^{-d}$
- The observed ***p***roportion of sites ***p*** at which the sequences are different can be considered as a proxy for the "realized" probability $Pr_{>1}$:

$$p = 1 - e^{-d}$$

- It can be solved for ***d***: $e^{-d} = 1-p$

$$\ln(e^{-d}) = \ln(1-p)$$

d or $d_{PC} = -\ln(1-p)$ (**P**oisson **C**orrection ***d***istance)

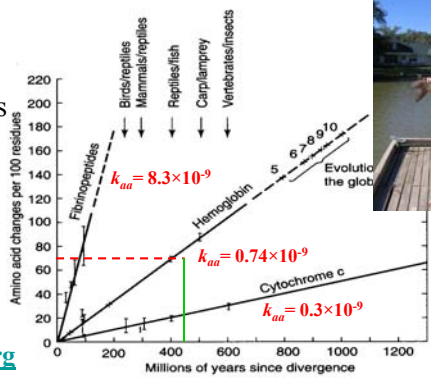
$$V_d = p/[(1-p)N] \qquad \text{s.e.} = \sqrt{V_d}$$

Molecular Clock

- The amount of amino acid divergence is $d = 2Tk_{aa}$
- The rate of substitution per amino acid per year is $k_{aa} = d/2T$, where T is the number of years since the divergence of two species from their common ancestor
- $d = -\ln(1-p)$

Exar 6.2, 8.2. The amino acid sequences of protein molecules in a number of organisms were compared soon after the concept of the molecular clock was proposed. For example, the α chain of the hemoglobin molecule was sequenced, and the sequence in humans and carp differed at 68 out of 140 sites. $p = 68/140 = 0.486$. Therefore, $d = 0.665 \pm 0.082$, and because the time since divergence of humans and carp is about 450 million years (Kumar and Hedges, 1998), $k_{aa} = 0.74 \times 10^{-9}$.

FIGURE 6.3 (8.3) The amount of amino acid substitution for three proteins having different rates of substitution. The horizontal axis gives the times since divergence of various organisms in millions of years (Dickerson 1971).



http://www.timetree.org

3

Molecular Clock

Amino acid probability substitution table

http://www.proteinstructures.com/Sequence/Sequence/amino-acid-substitution.html

A number of refinements in the theory used in estimating the mean number of substitutions per site have been proposed (see Nei and Kumar, 2000, for comparisons of these approaches). In addition, Dayhoff *et al.* (1978) empirically determined, from a survey of comparisons between a number of proteins, the probability of any given amino acid being replaced by any of the other 19. From this information, they constructed a 20 × 20 matrix that gives the probabilities for all possible transitions, which can then be used to predict changes in amino acid sequence (Nei and Kumar, 2000). These probabilities reflect the observation that changes between amino acids similar in biochemical properties are much more likely than are changes between greatly different amino acids.

https://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff

Additional reading:

Nei, M. and S. Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

http://www.timetree.org

4

Amino acid probability substitution table

BLOSUM62 matrix (S_{ij}):
(**BLO**cks of amino acid **SU**bstitution **M**atrix)

$$S_{ij} = \frac{1}{\lambda} \cdot \log_2 \frac{p_{ij}}{q_i q_j}$$

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Here, p_{ij} is the probability of two amino acids ***i*** and ***j*** replacing each other in a homologous sequence, and q_i and q_j are the background probabilities of finding the amino acids ***i*** and ***j*** in any protein sequence at random. The factor **λ** is a scaling factor, set such that the matrix contains easily computable integer values.

5

Phylogenomics: Choice of Scoring Matrix

PAM-1 — PAM-250

BLOSUM100 — BLOSUM30

→

Small evolutionary distance
Strong similarity for short sequence

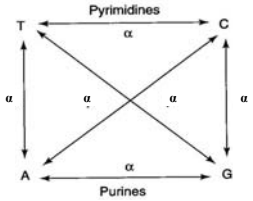
Large evolutionary distance
Weak similarity over stretched length

- **PAM** - **P**oint **A**ccepted **M**utations (% Accepted Mutations)
- **BLOSUM** - Blocks Substitution Matrix (**BLO**cks of amino acid **SU**bstitution **M**atrix)

6

Molecular Clock

- How to calculate the amount of nucleotide divergence ***d***, the mean number of nucleotide substitutions per site?
- If we assume that **α** is a substitution rate between any pair of nucleotides (Jukes & Cantor 1969):



- Then, the probability of a nucleotide to change is **3α** (total mutation rate) and remain unchanged at time ***t*** is $\mathbf{Pr}_t = 1 - 3\alpha$
- The probability of this nucleotide remain unchanged at time ***t*+1** is $\mathbf{Pr}_{t+1} = (1 - 3\alpha)\mathbf{Pr}_t + \alpha(1 - \mathbf{Pr}_t)$, where

  $(1 - 3\alpha)\mathbf{Pr}_t$ is the probability that the nucleotide remains unchanged from time ***t*** to time ***t*+1** , and

  $\alpha(1 - \mathbf{Pr}_t)$ is the probability of back mutation, that is at time ***t*** the nucleotide was different from the original one $(1 - \mathbf{Pr}_t)$, but with a probability **α**, it changed back at time ***t*+1** to the original nucleotide

Molecular Clock

- $\mathbf{Pr}_{t+1} = (1 - 3\alpha)\mathbf{Pr}_t + \alpha(1 - \mathbf{Pr}_t)$ can be written as the amount of change:

$$\mathbf{Pr}_{t+1} - \mathbf{Pr}_t = -4\alpha\mathbf{Pr}_t + \alpha$$

- or assuming continuous time can be written as differential equation:

$$\frac{d\mathbf{Pr}}{dt} = -4\alpha\mathbf{Pr}_t + \alpha$$

- it can be solved for $\mathbf{Pr}_t$:

$$\mathbf{Pr}_t = \tfrac{1}{4} + \tfrac{3}{4}e^{-4\alpha t}$$

- the probability that a nucleotide remains the same in **2** sequences at time ***t***:

$$\mathbf{Pr}_t = \tfrac{1}{4} + \tfrac{3}{4}e^{-8\alpha t}$$

Molecular Clock

- the probability that a nucleotide remains the same in 2 sequences at time ***t***:

$$\mathbf{Pr}_t = ¼ + ¾e^{-8\alpha t}$$

- the probability that sites are different $p = 1 - \mathbf{Pr}_t$ and $\mathbf{Pr}_t = 1 - p$

$$1 - p = ¼ + ¾e^{-8\alpha t}$$

$$p = ¾\,(1 - e^{-8\alpha t})$$

$$e^{-8\alpha t} = 1 - 4p/3$$

$$\ln e^{-8\alpha t} = \ln(1 - 4p/3)$$

$$-8\alpha t = \ln(1 - 4p/3)$$

- the expected number of substitution per site in a lineage is $3\alpha t$, for two lineages $6\alpha t$ that can be considered as a genetic distance $d = 6\alpha t$
  - $8\alpha t = -\ln(1 - 4p/3)$ or $^4/_3 6\alpha t = -\ln(1 - 4p/3)$ or $^4/_3 d = -\ln(1 - 4p/3)$

$$d = -¾\ln(1 - 4p/3) = d_{JC}$$

$$V_d = p(1 - p)/[N(1 - 4p/3)^2]$$

Molecular Clock

$$d_{JC} = -\frac{3}{4}\ln\left(1 - \frac{4p}{3}\right)$$

- where d_{JC} is the **J**ukes-**C**antor's genetic ***d***istance, and ***p*** is an observed ***p***roportion of nucleotide substitutions between two sequences
- an estimate of the rate of substitution per nucleotide site per year is then

$$k = \frac{d_{JC}}{2T}$$

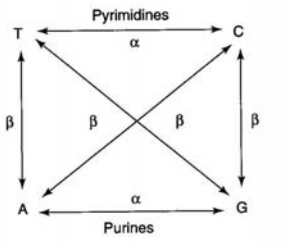


Figure 8.4. The rates of nucleotide substitutions for transitions (α) and transversions (β) under the two-parameter model of Kimura (1980).

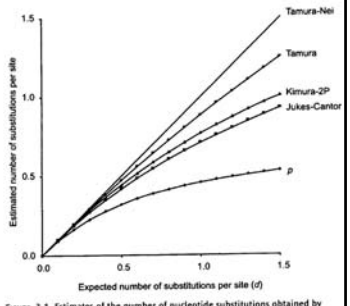
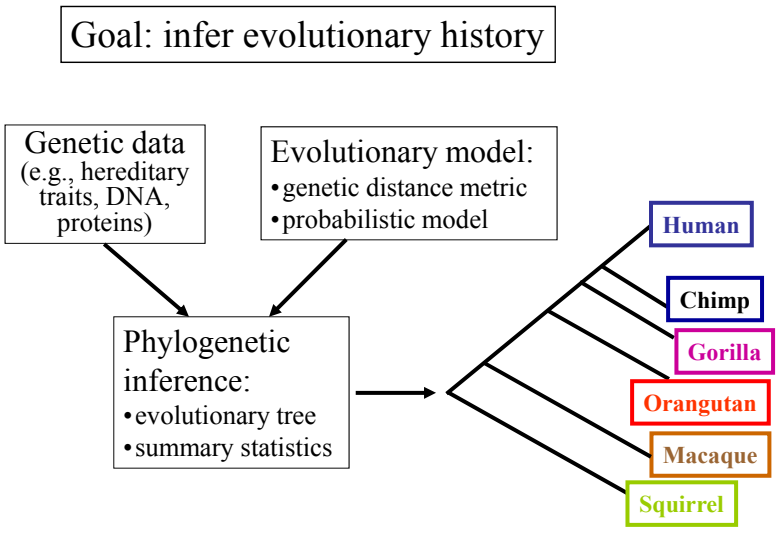


FIGURE 3.1. Estimates of the number of nucleotide substitutions obtained by different distance measures when actual nucleotide substitution follows the Tamura-Nei model. The nucleotide frequencies assumed are $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$; and the two transition/transversion rate ratios assumed are $\alpha_1/\beta = 4$ and $\alpha_2/\beta = 8$.

Molecular Phylogenetics

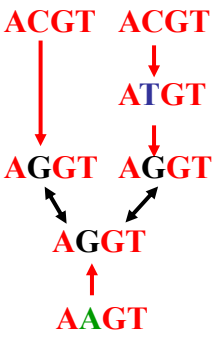
- Molecular Phylogenetics **is aimed at the reconstruction of** Evolutionary History **and relationships between genes, populations, species and other taxa based on molecular data (allele and haplotype frequencies, nucleotide and amino acid sequences)**
- **It assumes that the similarity and differences in DNA or inherited traits reflects evolutionary relationships**

Phylogenetic Analysis: Overview



Finding a Phylogeny

- no guarantees of correctness
  - **based on evidence, but there is often more than one way to arrive at the same answer**
  - **all we can *observe* is distance or differences from which we *infer* relatedness; homology vs. homoplasy: similarity due to a common ancestor vs. similarity due to convergent evolution**
- reconstructs the "most likely" history
  - **minimum evolution (ME) and parsimony find the shortest tree or evolutionary tree that explains the observations with the fewest possible changes**



(Occam's Razor philosophical principle)

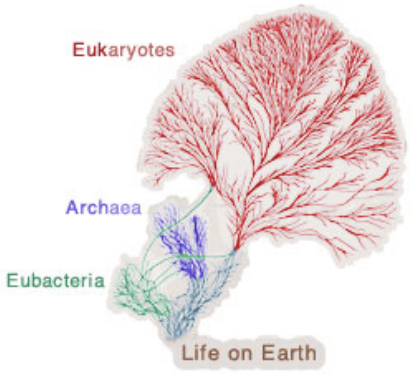
'To every complex problem there is a solution that is simple, straightforward, … and wrong.' Henry Louis Mencken (1880–1956)

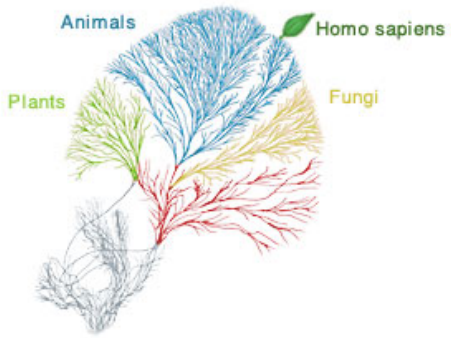
A brief history of phylogeny

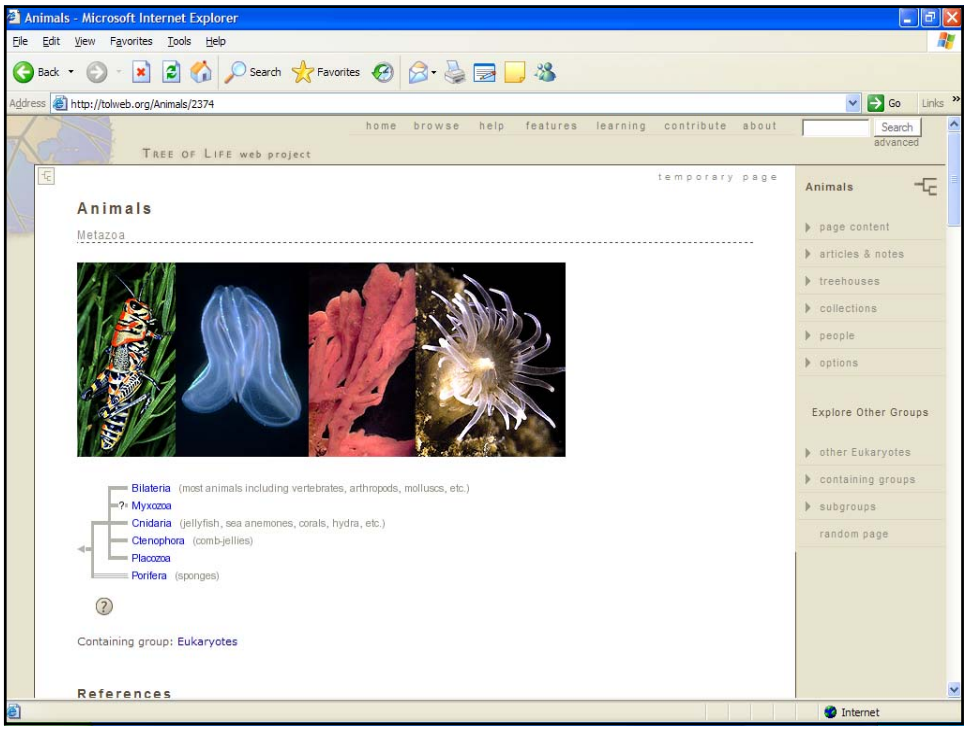
- The ancient Greeks divided life into classes or "forms" based upon their morphology and created early taxonomy
- Carl Linnaeus (1707–1778), a Swedish botanist, physician and zoologist laid the foundations for the modern scheme of Binomial nomenclature. He is considered as the father of modern taxonomy and modern ecology
- Darwin's Origin of Species (1859) propelled efforts to find the links between species
- Alfred Hershey and Martha Chase showed that DNA is the molecule of inheritance (1952)
- Emile Zuckerkandl and Linus Pauling used hemoglobin amino acid sequences to produce a primate phylogeny (1962)
- NSF begins funding efforts to construct the "Tree of Life" (2005)

The Tree of Life

http://tolweb.org

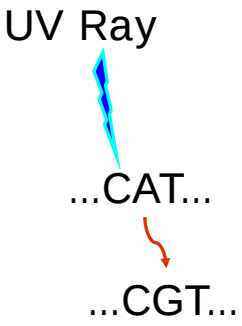






Molecular phylogeny: What makes it possible?

- Point mutations (substitution): one base is replaced with another



With only point mutations, easy to tell how close two genomes are - just count the different bases

Mutations caused by copying errors (enzymes slipping, etc.)

- Insertions: one or more bases are inserted

...GCATG... → ...GCA**CA**TG...

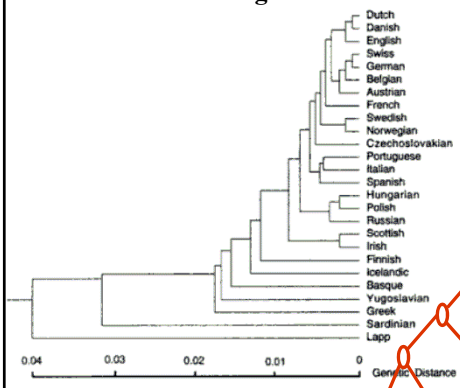
- Deletions: one or more bases are removed

...GCATCATG... → ...GCATG...

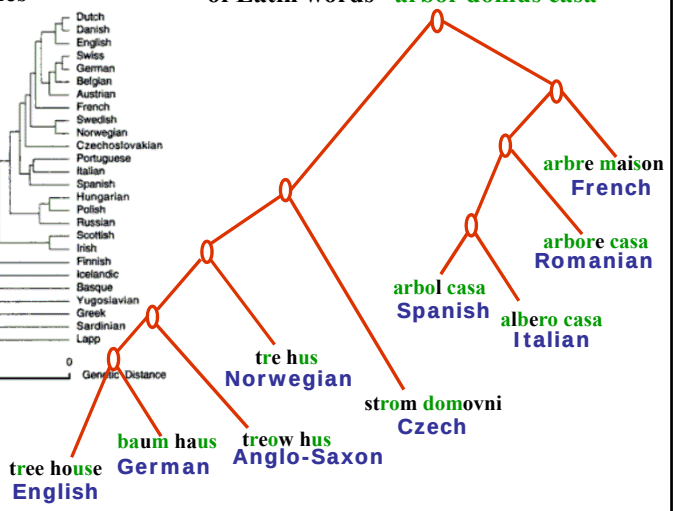
- and in sequence alignment, we can easily quantify differences

Phylogenetic trees based on genetic and social information

Phylogenetic tree based on evolution of 88 genes



Phylogenetic tree based on language evolution of Latin words "arbor domus casa"



From Cavalli-Sforza *et al.* "The History and Geography of Human Genes", page 268, Figure 5.5.1. Genetic tree of 26 European populations. F_{ST} distances are based on an average of 88 genes

Phylogeny: Standard Assumptions

1) Sequences diverge by bifurcation events

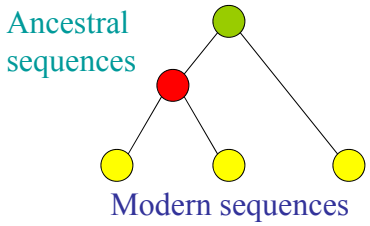
2) Sequences are essentially independent once they diverge from their common ancestor

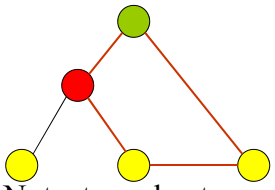
3) The probability of observing nucleotide ***k*** at site ***j*** in the future depends only on the current nucleotide at site ***j*** (Markov Chain assumption)

4) Different sites (characters) within a sequence evolve independently

Phylogenies represented with trees

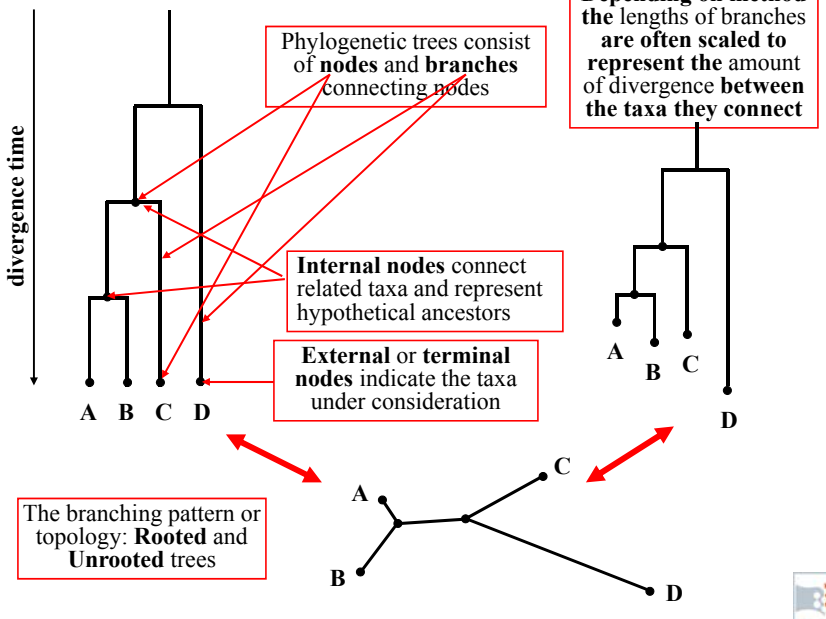
- **A tree is a *directed acyclic graph* consisting of *nodes* (sequences) and *edges* (relationships).**
- **There exists a single unique path between any pair of nodes.**



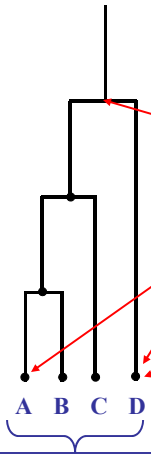


Not a tree, due to **cycle, but can be used in the network analysis**

Construction of phylogenetic trees from molecular data



Construction of phylogenetic trees from molecular data



The branching pattern or topology:

Rooted and **Unrooted** trees

There are two major methods to define a root for a tree:

1) to find the **middle point** between the **two most distant terminal nodes**

2) to use an **outgroup**, the most ancestral taxon that from paleontological or other data is assumed to have separated earlier than the other taxa under study

Definition: **OTUs – Operational Taxonomic Units** that could be genes, haplotypes, individuals, populations, species or other biological groups and taxa

Construction of phylogenetic trees from molecular data

- **It is important to differentiate between gene trees and species trees**

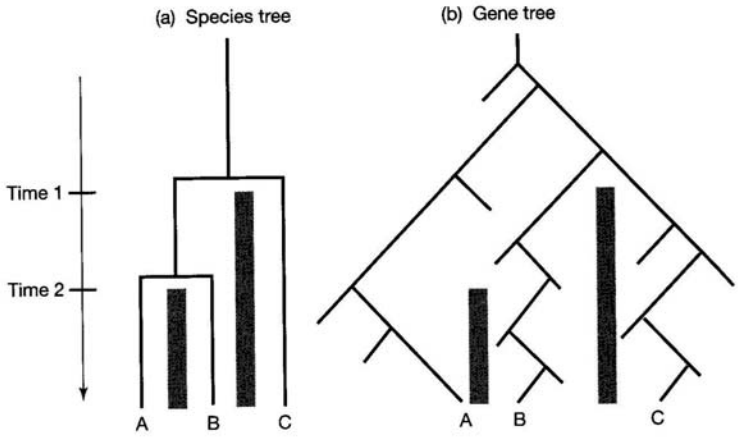
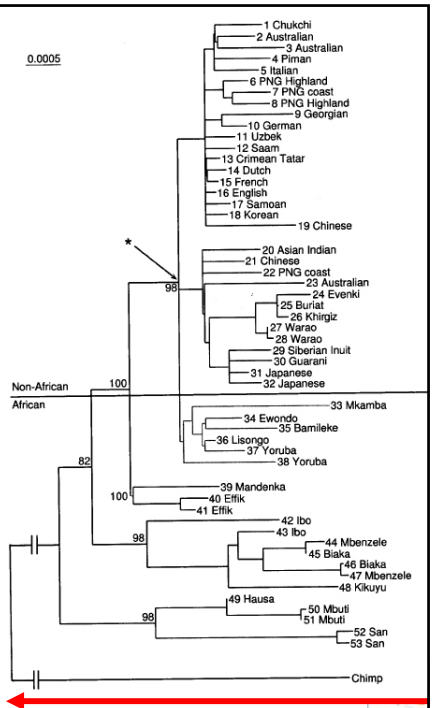


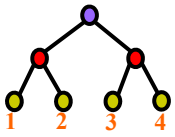
Figure 11.1. (a) A species tree that results from one barrier (shaded bar) splitting the ancestral species at time 1 and another barrier splitting part of the species at time 2. As a result, species A and B are most closely related. (b) A gene tree for a gene in the same species in which the ancestral species is polymorphic for several sequences. By chance, the surviving sequences in species B and C are more closely related than the surviving sequences in species A and B (after Kocher, 2003).

Human mtDNA Phylogeny

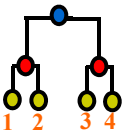
- *Vertical layout* is relatively meaningless (e.g. swapping any 2-way branch has no effect on tree meaning)
- **Evolutionary distance** (horizontal scale, in this diagram) is the most important information in the phylogeny, and is reflected in the tree structure (grouping).
- Phylogeny is *not* classification, but ***distance***, i.e. there is no answer to "how many clusters?"

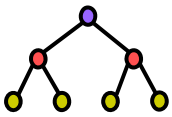


Phylogenetic tree types

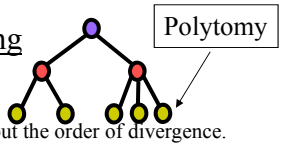


Forked vs. Hierarchical



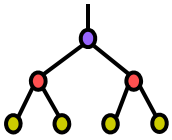


Bifurcating vs. Multifurcating

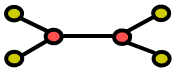


Polytomies: **Soft** vs. **Hard**

- **Soft**: designate a lack of information about the order of divergence.
- **Hard**: the hypothesis that multiple divergences occurred simultaneously



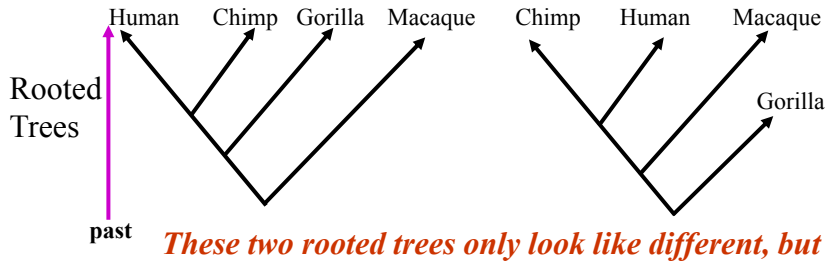
Rooted vs. Unrooted



Evolutionary vs. Phylograms

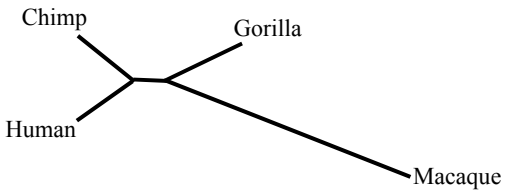
Ultrametric vs. Additive

Rooted vs. Unrooted Trees



These two rooted trees only look like different, but both can be presented as the same unrooted tree:

Unrooted Tree

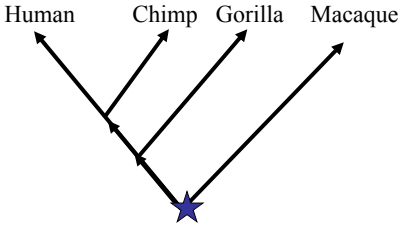


The direction of evolution is specified in a rooted tree but not in an unrooted tree!

Rooted Trees Are Directed Graphs

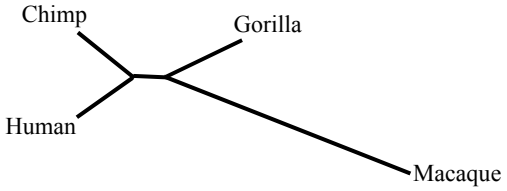
In a rooted tree each edge is directed, pointing from root to "leaves"

Rooted Tree

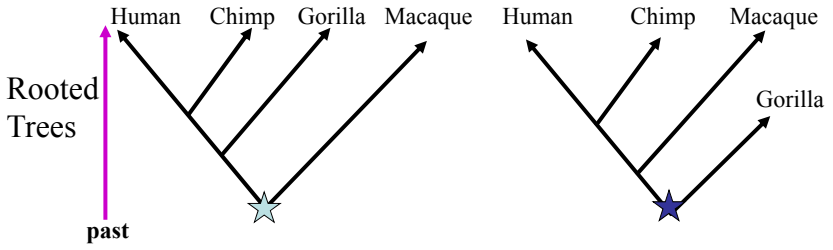


In an unrooted tree, the edges are "undirected"

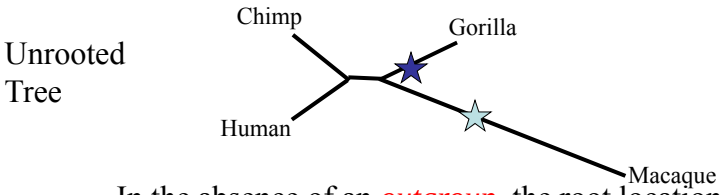
Unrooted Tree



Choosing a Root Location?



A rooted tree is just an unrooted tree with a ***root location*** chosen on one branch

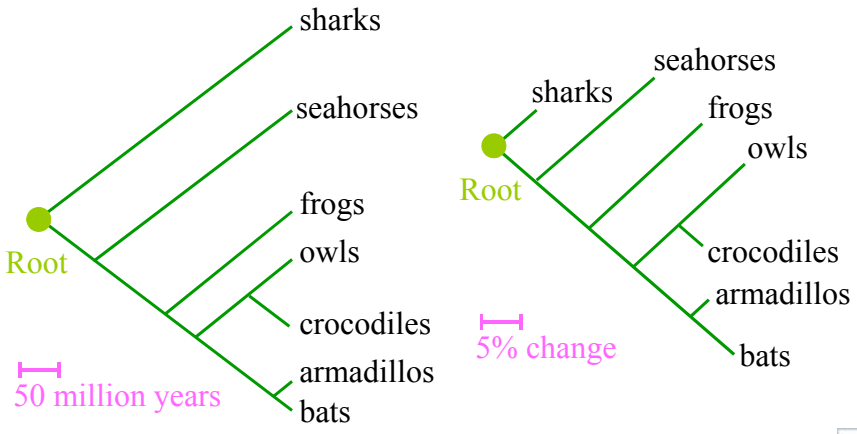


In the absence of an *outgroup*, the root location is often unclear and therefore arbitrary.

Tree Types

Evolutionary trees measure ***time***

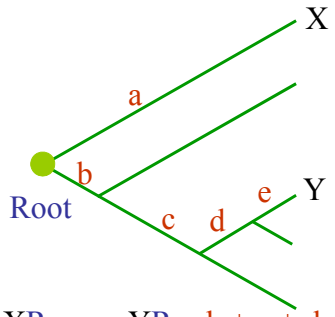
Phylograms measure ***change***



Tree Properties

Ultrametric

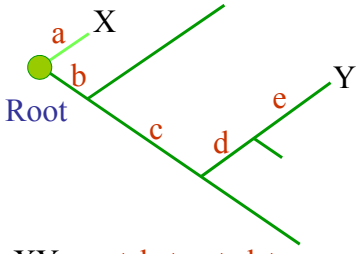
All tips are an equal distance from the root.



XR = a = YR = b + c + d + e

Additive

Distance between any two tips equals the total branch length between them.



XY = a + b + c + d + e

In simple scenarios, evolutionary trees are ultrametric, and phylograms are additive.

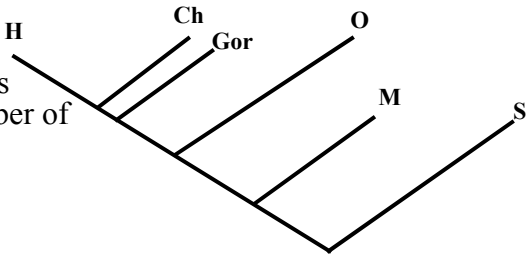
Character Differences → Branch Lengths and Order

Characters

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Phylogenetic Tree

- Length of a branch is proportional to number of differences
- Tree shows order of divergences



Branch length represents the extent of change - expected number of substitutions per nucleotide site. The longer the branch the more change

Character Differences → Sequence Distances

Characters

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Distances

	Chimp	Gor	Hum	Mac	Orang	Sq
Chimp	-	5	2	8	8	9
Gor		-	5	7	6	8
Hum			-	9	8	9
Mac				-	2	4
Orang					-	5
Sq						-

Molecular Phylogenetics Methods

Classification of the major approaches and methods of phylogenetic analysis and constructing phylogenetic trees from molecular data:

1) Cluster analysis based on the measures of pairwise similarity and dissimilarity (e.g., genetic distance) or so-called distance matrix methods (phenetics); it finds the tree that best fits the pairwise distances between taxa

2) Maximum Parsimony analysis based on discrete characters (cladistics); it finds the tree that requires the fewest changes to explain the data

3) Maximum Likelihood analysis (probabilistic methods including Bayesian methods); it finds the most likely tree

4) Multivariate analysis (PCA, PCoordA, Multidimensional Scaling, etc.)

Molecular Phylogenetics Methods

- When there is good statistical confidence in the topology of phylogenetic trees, then the major methods all work well and are reasonably comparable
- However, different methods may use different assumptions, and if some of the assumptions are violated, then some methods may work better and produce more realistic trees than others
- The produced trees should be biologically meaningful!

Construction of phylogenetic trees from molecular data: Distance matrix methods (aka "*algorithmic methods*")

- Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
- Neighbor-Joining (NJ) method

Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

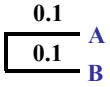
- The UPGMA is the simplest method of tree construction.
- It was originally developed for constructing taxonomic phenograms, i.e. trees that reflect the phenotypic similarities or dissimilarities (phenetic distances) between OTUs (sometimes called “phenetics”).
- It can also be used to construct phylogenetic trees, but it assumes that the rates of evolution are approximately constant among the different lineages.
- UPGMA needs the matrix of pairwise genetic distances that can be based on practically any data - allele frequencies, nucleotide or amino-acid substitutions, restriction sites, etc.
- UPGMA employs a sequential clustering algorithm, in which local topological relationships are identified in order of similarity, and the phylogenetic tree is build in a stepwise manner.

Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

- **Let us assume that we have the following matrix of pairwise genetic distances for 6 OTUs:**

OTUs	A	B	C	D	E
B	0.2				
C	**0.4**	**0.4**			
D	**0.6**	**0.6**	**0.6**		
E	**0.6**	**0.6**	**0.6**	**0.4**	
F	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**

- We first identify among all OTUs two OTUs that have the smallest genetic distance and, therefore, are the most similar to each other (**A** and **B**).
- We now cluster these pair of OTUs that are separated by a distance of **0.2**. The branching point is positioned at a distance of 0.2/2 = 0.1. We thus construct a subtree as follows:



Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

- Following the first clustering A and B are considered as a new single composite OTU_{AB}
- We can now calculate the new distance matrix as follows:

$D_{(AB)C} = (D_{AC} + D_{BC})/2 = 0.4$

$D_{(AB)D} = (D_{AD} + D_{BD})/2 = 0.6$

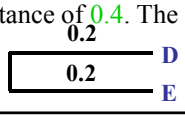
$D_{(AB)E} = (D_{AE} + D_{BE})/2 = 0.6$

$D_{(AB)F} = (D_{AF} + D_{BF})/2 = 0.8$

OTUs	AB	C	D	E
C	**0.4**			
D	**0.6**	**0.6**		
E	**0.6**	**0.6**	0.4	
F	**0.8**	**0.8**	**0.8**	**0.8**

In other words the distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU. Then a new distance matrix is recalculated using the newly calculated distances and the whole cycle is being repeated

- In the new matrix we identify the next pair that have the smallest genetic distance now (D and E).
- We now cluster this pair of OTUs that are separated by a distance of 0.4. The branching point is positioned at a distance of 0.4/2 = 0.2.
- We thus construct the second subtree as follows:



Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

- Following the second clustering D and E are considered as a new single composite OTU_{DE}
- We can now calculate the new distance matrix as follows:

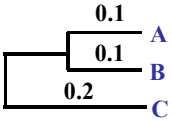
$D_{(DE)(AB)} = (D_{D(AB)} + D_{E(AB)})/2 = 0.6$

$D_{(DE)C} = (D_{DC} + D_{EC})/2 = 0.6$

$D_{(DE)F} = (D_{DF} + D_{EF})/2 = 0.8$

OTUs	AB	C	DE
C	0.4		
DE	**0.6**	**0.6**	
F	**0.8**	**0.8**	**0.8**

- In the new matrix we identify the next pair that have the smallest genetic distance now (AB and C).
- We now cluster this pair of OTUs that are separated by a distance of 0.4. The branching point is positioned at a distance of 0.4/2 = 0.2.
- We thus construct the next subtree as follows:



Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

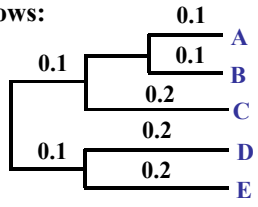
- **Following the third clustering** AB **and** C **are considered as a new single composite** OTU_{ABC}
- **We can now calculate the new distance as follows:**

$D_{(ABC)(DE)} = (D_{(AB)(DE)} + D_{C(DE)})/2 = 0.6$

$D_{(ABC)F} = (D_{(AB)E} + D_{CF})/2 = 0.6$

OTUs	ABC	DE
DE	0.6	
F	0.8	0.8

- **In the new matrix we identify the next pair that have the smallest genetic distance now (**ABC **and** DE**).**
- **We now cluster this pair of OTUs that are separated by a distance of** 0.6**. The branching point is positioned at a distance of 0.6/2 = 0.3.**
- **We thus construct the next subtree as follows:**



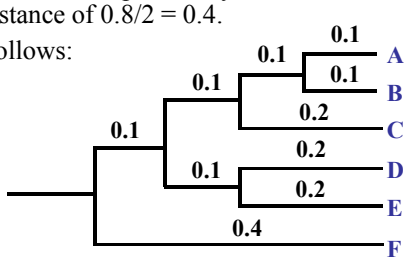
Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

- Following the forth clustering ABC and DE are considered as a new single composite OTU_{ABCDE}
- We can now calculate the final distance as follows:

$D_{(ABCDE)F} = (D_{(ABC)F} + D_{(DE)F})/2 = 0.8$

OTUs	ABCDE
F	0.8

- Finally, we cluster this pair of OTUs that are separated by a distance of 0.8. The branching point is positioned at a distance of 0.8/2 = 0.4.
- We thus construct the final tree as follows:



- Although this method leads essentially to an unrooted tree, UPGMA assumes equal rates of mutation along all the branches. The theoretical root, therefore, must be equidistant from all OTUs. We can here thus apply the method of mid-point rooting.
- The root of the entire tree is then positioned at dist (ABCDE)F/2 = 0.4.

Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

However, there are some pitfalls:

- The UPGMA clustering method is *very sensitive* to unequal evolutionary rates. This means that when one of the OTUs has incorporated more mutations over time than the other OTU, one may end up with a tree that has the wrong topology.
- Clustering works only if the data are *ultrametric*:
  - ultrametric distances are defined by the satisfaction of the '*three-point condition*'.
  - what is the three-point condition?
  - for any three taxa {A,B,C}:
    - $AB \leq \max(AC,BC)$
    - $AC \leq \max(AB,BC)$
    - $BC \leq \max(AC,BC)$

    or in other words: the two greatest distances are equal, or all terminal nodes are equidistant from the root, which follows from the UPGMA assumption that the evolutionary rate is the same for all branches
- If the assumption of constant evolutionary rate among lineages does not hold, then UPGMA may give an erroneous topology.
- It should only be used for closely related OTUs, or when there is constancy of evolutionary rate.

Neighbor-Joining (NJ) method

- **Neighbor-joining (Saitou and Nei, 1987) is also** a distance based cluster method**, but unlike UPGMA** does not require data to be ultrametric**.**
- **Therefore, it has become the method of choice for many types of molecular data because it does not require that all lineages have diverged by equal amounts, and it** can incorporate different rates of evolution in different lineages
- **Like UPGMA it also** proceeds in a stepwise fashion **through minimizing the sum of the branch lengths at each step in the total tree (**the minimum evolution (ME) principle).
- NJ uses the minimum evolution criterion in each step (but not as an overall criterion) and hence is a good way to produce a tree in the first step of an heuristic search strategy when using ME as optimality criterion.

(Saitou & Nei, 1987)

Neighbor-Joining (NJ) method

- The neighbor-joining method is a special case of the star decomposition method:
- The raw data are provided as a distance matrix, and the initial tree is a star tree.
- Then a modified distance matrix is constructed, in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes.
- The tree is constructed by linking the least-distant pair of nodes in this modified matrix.
- When two nodes are linked, their common ancestral node is added to the tree, and the terminal nodes with their respective branches are removed from the tree.
- This pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size.
- At each stage in the process two terminal nodes are replaced by one new node.
- The process is complete when two nodes remain, separated by a single branch.

(Saitou & Nei, 1987)

Neighbor-Joining (NJ) method

Example of the method

- Suppose we have the following tree:
- Since B and D have accumulated mutations at a higher rate than A, the three-point criterion is violated, and the UPGMA method cannot be used since this would group together A and C rather than A and B. In such a case the neighbor-joining method is one of the recommended methods. The raw data of the tree are represented by the following distance (D_{ij}) matrix:
- We have in total 6 OTUs (N=6).
- Step 1: We calculate the net divergence R_i for each OTU from all other OTUs:

$R_A = 5+4+7+6+8=30$

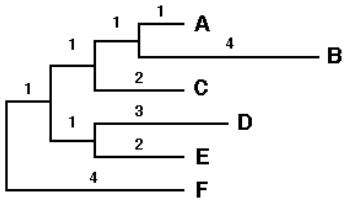
$R_B = 42$

$R_C = 32$

$R_D = 38$

$R_E = 34$

$R_F = 44$



OTUs	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Neighbor-Joining (NJ) method

Step 2: Then, to determine the nearest neighbors we calculate a new distance matrix (S_{ij}) for each pair of OUTs using the formula:

$$S_{ij} = (2T - R_i - R_j)/2(N-2) + D_{ij}/2 \qquad (11.2a),$$

where $T = \sum D_{ij}$

Furthermore, since this matrix is used only to determine the nearest neighbors, and since T is the same for all pairs of i and j, to facilitate computing we can transform it into:

$$2S_{ij} = (2T - R_i - R_j)/(N-2) + D_{ij}$$

$$2S_{ij} = 2T/(N-2) - (R_i + R_j)/(N-2) + D_{ij}$$

$$2S_{ij} - 2T/(N-2) = D_{ij} - (R_i + R_j)/(N-2)$$

$2S_{ij} - 2T/(N-2)$ can be replaced by Q_{ij} :

$$Q_{ij} = D_{ij} - (R_i + R_j)/(N-2)$$

(for simplicity I keep using S_{ij} instead of Q_{ij} in further slides)

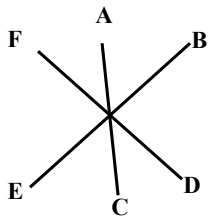
Neighbor-Joining (NJ) method

Step 2: Then, we calculate a new distance matrix (S_{ij}) for each pair of OUTs using the formula:

$S_{ij} = D_{ij} - (R_i + R_j)/(N-2)$ or in the case of the pair **A**, **B**:

$S_{AB} = D_{AB} - (R_A + R_B)/(N-2) =$ **-13**

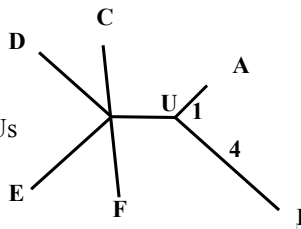
- Then, we start with a star tree:



OTUs	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

Step 3: We choose as neighbors those two OTUs for which S_{ij} is the smallest. These are **A** and **B**, and **D** and **E**. Let's take **A** and **B** as neighbors and form a new node called **U**. Then, we calculate the branch length from the internal node **U** to the external **OTU**s **A** and **B**.

- $S_{AU} = D_{AB}/2 + (R_A - R_B)/2(N-2) = 1$
- $S_{BU} = D_{AB}/2 + (R_B - R_A)/2(N-2) = 4$
  $= D_{AB} - S_{AU} = 4$



Neighbor-Joining (NJ) method

Step 4: Now we define new distances from U to each other terminal node:

$D_{CU} = (D_{AC} + D_{BC} - D_{AB})/2 = 3$
$D_{DU} = (D_{AD} + D_{BD} - D_{AB})/2 = 6$
$D_{EU} = (D_{AE} + D_{BE} - D_{AB})/2 = 5$
$D_{FU} = (D_{AF} + D_{BF} - D_{AB})/2 = 7$

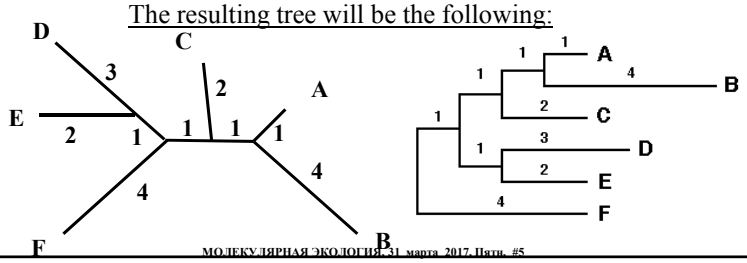
and we create a new matrix:

OTUs	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

$N = N - 1 = 5$

The entire procedure is repeated starting at **step 1** (calculate the net divergence R_i); **step 2** (calculate a new distance matrix, S_{ij}); **step 3**: (choose two new neighbors with smallest S_{ij}, create a new node and calculate the branch lengths)

The resulting tree will be the following:



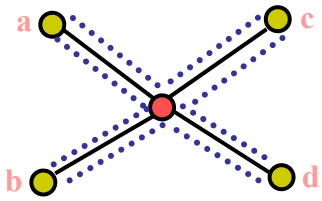
Neighbor-Joining (NJ) method

Advantages and disadvantages of the neighbor-joining method

- Advantages
  - **is fast and thus suited for large datasets and for bootstrap analysis**
  - **permits lineages with largely different branch lengths**
  - **permits correction for multiple substitutions**
- Disadvantages
  - **sequence information is reduced**
  - **gives only one possible tree**
  - **strongly dependent on the model of the minimum evolution used.**
- **Note:** ***its suitability to handle large datasets has led to the fact that the method is widely used by molecular evolutionists. With the rapid growth of sequence databases it is still one of the few methods that allows the rapid inclusion of all homologous sequences present in the database in a single tree. A good example can be found in the Ribosomal Database Project (http://rdp.cme.msu.edu/) that maintains a tree of life based on all available ribosomal RNA sequences.***

Neighbor Joining:
An algorithm for finding the shortest tree

Start with a star (no hierarchical structure)



$$S_O = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum D_{ij},$$

The length of the tree

Number of OTUs

Pair-wise distances

Neighbor Joining

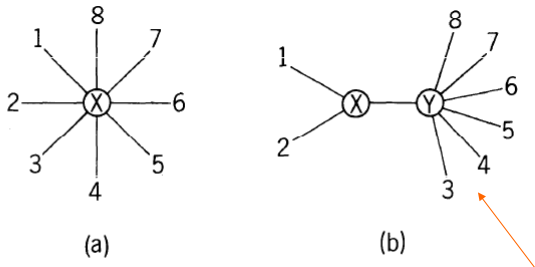


FIG. 2.—(a), A starlike tree with no hierarchical structure; and (b), a tree in which OTUs 1 clustered.

The following can be used to calculate the length of this tree:

$$L_{XY} = \frac{1}{2(N-2)}\left(\sum_{k=3}^{N}(D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{N} L_{iY}\right)$$

$$L_{1X} + L_{2X} = D_{12}$$

$$\sum_{i=3}^{N} D_{iY} = \frac{1}{(N-3)-1}\sum_{i=3}^{N}\sum_{i<j}^{N} D_{ij}$$

Neighbor Joining

At each step, each pair of possible neighbors are considered, and the one producing the shortest tree is chosen (minimal evolution criteria).

S_{ij} Matrices for Two Cycles of the NJ Method for the Data in Table 1

A. Cycle 1: Neighbors = [1, 2]

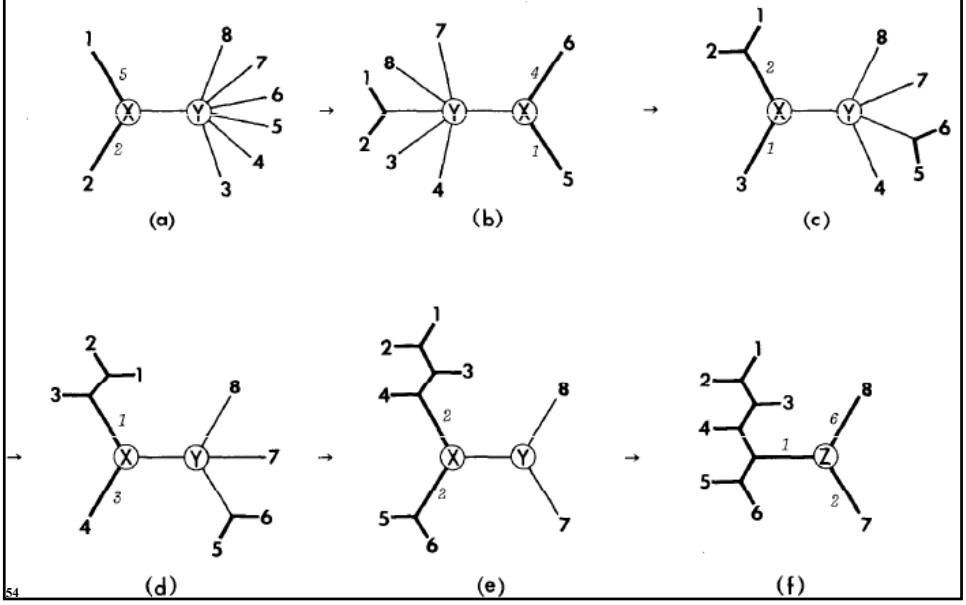
	OTU						
OTU	1	2	3	4	5	6	7
2 . .	36.67						
3 . .	38.33	38.33					
4 . .	39.00	39.00	38.67				
5 . .	40.33	40.33	40.00	39.67			
6 . .	40.33	40.33	40.00	39.67	37.00		
7 . .	40.17	40.17	39.83	39.50	38.83	38.83	
8 . .	40.17	40.17	39.83	39.50	38.83	38.83	37.67

B. Cycle 2: Neighbors = [5, 6]

	OTU					
OTU	1-2	3	4	5	6	7
3 . .	31.50					
4 . .	32.30	32.30				
5 . .	33.90	33.90	33.70			
6 . .	33.90	33.90	33.70	31.30		
7 . .	33.70	33.70	33.50	33.10	33.10	
8 . .	33.70	33.70	33.50	33.10	33.10	31.90

Neighbor Joining

As in UPGMA, a new internal branch is added at each step.



Maximum Parsimony (MP)

- Maximum Parsimony (MP) scores the number of changes between different character states that the minimum of changes are necessary to explain the observed data given the tree.
- The hypothesis is that the best is one that would require the fewest changes. The changes may be restricted in what kind of changes that are allowed.
- This score, often referred to as the tree's *length*, is the minimum number of changes for the tree. Thus, parsimony uses the observations without any attempt to use an explicit model to estimate the total number of changes.
- If we have ***k*** characters (sites) that each require ***l*** changes, the length of the tree, ***L***, is the sum
- $$L = \sum_{i=1}^{k} l_i$$
- Parsimony can be justified in several different ways; one is that it is a general principle in science known as *Occam's razor – there is no need to make more assumptions than necessary to explain the observations*.
- Hypotheses of homoplasy, that is when we have more steps (changes) than is minimally needed for the data, may be judged *ad hoc* in that they are attempts to explain the data that do not fit a particular hypothesis.
- The most parsimonious tree is the solution that requires the smallest number of such *ad hoc* hypothesis and is thus preferred.
- Parsimony is also appealing in that it does not require us to make any *explicit* assumptions about the process, and thus can be applied when we have data that can not easily be modeled.

Maximum Parsimony (MP)

The two most commonly used parsimony models:

- **Fitch parsimony**. All changes are assigned unit costs (or individual costs, i.e. transformational weighting); the cost of a change from 0 to 2 need not to be the sum of the changes from 0 to 1 and 1 to 2. However, the costs must obey the triangle inequality so the cost of going from 0 to 1 plus 1 to 2 can not be *less* costly than going from 0 to 2 in one step. Fitch parsimony is usually implied if one just uses the term "parsimony".
- **Wagner parsimony**. The character states are measured on an interval scale and thus are *ordered*. For a character having the states 0, 1, and 2 a change from state 0 to 2 on a tree would have the same cost as a change from 0 to 1 plus a change from 1 to 2.
- Wagner & Fitch parsimony have symmetric costs: cost (0→1) = cost (1→0).

Less frequently used:

- **Dollo parsimony**. Each character state is allowed to be gained only once on the tree, and if the distribution of character states on the tree does not fit, this must be explained by extra reversals (losses). This has been proposed as a way to analyze restriction site data, where *the probability of a loss is much higher than the probability of a gain*: cost (1→0) < cost (0→1).
- **Camin Sokal parsimony**. This was the first parsimony methods proposed, and is mentioned only for completeness. Here evolution are assumed to be irreversible and no reversals are allowed, but only multiple gains.

Maximum Parsimony (MP)

- All the above parsimony models can be regarded as special cases of **generalized parsimony**.
- **Generalized parsimony** assigns a cost for the transformation of any character state into each of all other character states.
- The cost for ***m*** different character states can be represented as a ***m*** × ***m*** matrix ***S***, where S_{ij} is the increase in tree length for a transformation from state ***i*** to state ***j***.
- If all costs are not equal, this is commonly referred to as ***transformational weighting*** or ***character state weighting***.
- Below are examples of cost matrices for Wagner, Fitch and Dollo parsimony as well as an example where transversions is weighted twice the cost of transitions.

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

Wagner parsimony

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

Fitch parsimony

	0	1	2	3
0	-	1M	2M	3M
1	1	-	1M	2M
2	2	1	-	1M
3	3	2	1	-

Dollo parsimony; M is an arbitrary, large number

	A	C	G	T
A	-	2	1	2
C	2	-	2	1
G	1	2	-	2
T	2	1	2	-

Transformational weighting - transversions twice as costly as transitions

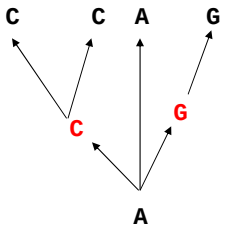
Maximum Parsimony (MP)

- Parsimony can be further elaborated by using another class of differential weighting. Instead of simply counting the minimum number of changes on a tree, certain characters may be given a higher weight (cost), implementing a *positional weighting* or *character weighting*. For example, the three positions in a codon can be given different weights. This is implemented by multiplying each characters length with a weight, ***w***. The tree length in this case is $L = \sum_{i=1}^{k} w_i l_i$
- The motivation for using differential weighting is to get a better approximation of the actual changes from the observed differences by giving less weight to characters or changes that are considered less informative (usually characters that seems to be changing a lot or changes that seem to occur more frequently). The actually numeric values of the weights applied are to some extent arbitrary, but the common practice is to let the data determine these weights according to some objective function.

Maximum likelihood (ML) methods

- Similar to maximum parsimony:
  - For each column of the alignment all possible trees are calculated
  - Trees with the least number of substitutions are more likely

(1) A G G C U C C A A
(2) A G G U U C G A A
(3) A G C C C A G A A
(4) A U U U C G G A A



P(change)=?

Maximum likelihood (ML)

- **Maximum likelihood (ML)** is a kind of estimate that is very common in statistics. For example, estimating the population mean with the average of a sample is a maximum likelihood (or *ML*) estimate. (Other common techniques in statistics are the methods of moments, which are used for pair-wise distances, and least squares).
- **ML** is different from parsimony in that an explicit model is used to calculate the score. The model in phylogenetic contexts consists of two parts:
  - a model of how the character state changes occur (probabilities of changes)
  - a tree with branch lengths
- The score used is the likelihood of a *model* (which includes the tree we want to evaluate), which is the conditional probability of the *data* (***D***) given the model (***H***). Or, phrased differently, it is the probability of getting the data we actually have got if the model (the tree ***T*** and the parameters ***Θ***) were true: $L_H \propto P(D|H) = P(D|T,\Theta)$

Maximum likelihood (ML)

- Note that the likelihood is the (conditional) probability of the *data* not a probability of the *model* (the tree), since the sum of the likelihood over all trees does not equal one.
- Normally, independence between characters is assumed and the likelihood of the tree is the product of the likelihoods of all characters. Since the likelihoods are very small numbers, the logarithm of the likelihood is normally used (with the log likelihood of the tree being the sum of the log likelihoods for the characters).

$$L = \prod_{i=1}^{k} L_i \qquad \ln L = \sum_{i=1}^{k} \ln(L_i)$$

- Calculating the likelihood for a tree is computationally very intense and takes considerably more time than calculating for example the tree length in parsimony, and increases for more complex models. Extra computational load is generated since all branch lengths are optimized numerically for each tree that is examined.

Maximum likelihood (ML)

- The models that are most commonly used for DNA sequences are sub models of the **GTR class of models** (General Time Reversible model, the mathematical expression of a substitution model presented as a table of rates at which each state is replaced with each alternative state), commonly modified to incorporate heterogeneity rate among the different sites. Some assumptions explicit or implicit when using these models in ML estimates are
- *Independence between sites* This may be violated by, for example, compensatory changes in rRNA genes
- *The conditional probability the same for all sites and not changing over time*. Different sites may evolve at different rates (violating the first part), e.g., positions in a codon; this can be handled by introducing the rate heterogeneity models.
- *Base composition at equilibrium (stationary).* The same base composition in all taxa, and along all edges. This is an assumption that is frequently violated, and there are more extensive models that try to handled this situation.
- *Constant rate (over time and in different lineages)*
- One advantage of maximum likelihood is that it will give a correct result in some cases where other methods fail (i.e., it is *consistent* in those cases, see below) – provided that the models used are correct... The need for explicit models are sometimes viewed as a weakness, but may also be a strength as the values for different parameters are visible and thus their validity can be assessed.

Maximum likelihood advantages

- Advantages of maximum likelihood over maximum parsimony:
  - Takes into account different rates of substitution between different amino acids and/or different sites
  - Statistically well-founded
  - Applicable to more diverse sequences

Finding the tree - Search methods

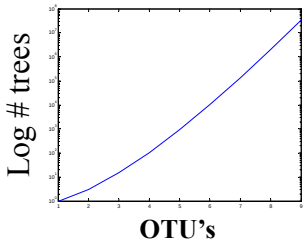
- The number of rooted trees (N_r) for n OTUs is given by Cavalli-Sforza & Edwards (1967):

$$N_r = (2n-3)!/(2^{n-2})(n-2)!$$

- The number of unrooted trees (N_u) for n OTUs is:

$$N_u = (2n-5)!/(2^{n-3})(n-3)!$$

Number of OTU's	Possible Number of	
	Rooted trees	Unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025
50	2.8×10^{76}	3×10^{74}



This rapid increase in number of trees to be analysed may make it impossible to apply some methods to very large datasets. In that case the MP and ML methods may become very time consuming, even on very fast computers.

(10^{79} protons in the universe)

Finding the tree - Search methods

Exact methods

Exhaustive enumeration

- If we evaluate each and every possible tree in turn, we will of course find the best tree. For data sets with few taxa (8-15 or so, depending on criterion used) exact methods - methods that are guaranteed to find the optimal tree - can be used. For larger data sets computing time will be prohibitively large and we have to use methods that are not guaranteed to find the optimal tree, heuristic approaches.
- It is not easy to get an appreciation for how big these numbers are, but the number of trees increases very rapidly. Four sequences have three different unrooted trees, seven sequences 945 trees, 10 sequences 2027025 trees. For not-too-impressive 53 sequences we have $2.75 \cdot 10^{80}$ unrooted trees, a number that is bigger than the estimated number of hydrogen atoms in the universe. So, this approach may only be feasible up to 11 taxa.

Branch and bound

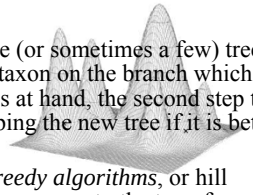
- We might do a little better if we can exclude some of the trees and just evaluate a subset of all possible trees. This can be achieved by first getting a fairly good (but not necessarily optimal) tree by some quick methods and than assembling a tree by adding one taxon at a time (in a sequence determined by some algorithm). The tree with the best score is selected for adding next taxa to the tree, and those trees with worse scores are discarded. Then, the next taxa is added, and again the tree with the best score is selected.
- The success of such a method, called *branch-and-bound*, will depend on the data. "Messy" data will decrease the efficiency and in the worst case, we will have to evaluate all possible trees and thus perform an exhaustive search. It may be worth trying branch-and-bound for up to 15-20 taxa or so.

Finding the tree - Search methods

Heuristic methods

- For most data sets we are forced to turn to *heuristic* or "quick-and-dirty" search methods, methods that try to find the best tree by reducing the set of trees examined and just calculating the score for some "likely" trees. However, these methods are not guaranteed to find the best tree(s) and one frequently will need to vary some of the parameters and try several times to get an adequate result.

Greedy algorithms or "hill-climbing"



- Heuristic methods *usually* proceed in two steps; first a single (or sometimes a few) tree(s) is (are) built by adding one taxon at a time, placing the added taxon on the branch which gives the best tree for the subset. When a tree containing all taxa is at hand, the second step tries to find a better tree by moving subtrees to other branches, keeping the new tree if it is better than the previous.
- The type of algorithms used is frequently of a kind called *greedy algorithms*, or hill climbing. This comes from the analogy of a method to find ones way to the top of a mountain (a peak of optimal score in tree search) when visibility is zero. It is quite simple: take one step in an arbitrary direction; if the ground is lower at the new position, go back and try another direction; if it is higher proceed with a new step. Eventually, one will end up in a spot where the ground is lower one step away in all directions – i.e., at a peak. However, we do not know if this is the summit of the mountain – it might be just a local peak. Of course we have the analogous problem when trying to find the best tree; the remedy is to start this "hill-climbing" from different points in the "tree landscape"
- There are variations on the procedures in both steps in the heuristic algorithms, and their performance will depend at the data; it takes some skills to make optimal use of these search methods. Remember – a tree with a better score is a better hypothesis according to the chosen criterion irrespective of how that tree is found.

Finding the tree - Search methods

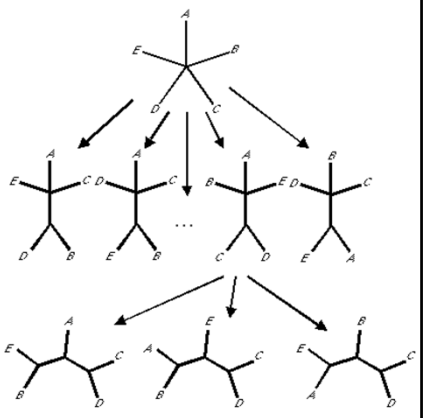
Step 1 - Obtaining the initial tree(s)

•There are several options to obtain the initial tree(s), and all are "correct" in some way but some are more efficient depending on the data at hand.

Star decomposition

Neighbor-joining described above is the most commonly used algorithm in this group. More generally, it is a divisive pair-wise clustering method. It can be used with any optimality criterion that can be evaluated on a polychotomous tree; neighbor-joining is an implementation using the minimum evolution criterion.

The algorithm start with all taxa connected in a star topology (all taxa connected to a single internal node). Next we evaluate all trees that can be constructed by joining two of the terminal nodes in a new group; the tree with the best score is kept to the next step. In each step when we form a new group, the number of branches connected to the central node is reduced by one. This continues until we have a dichotomous tree.



Finding the tree - Search methods

- Step 1 - Obtaining the initial tree(s)

Stepwise addition

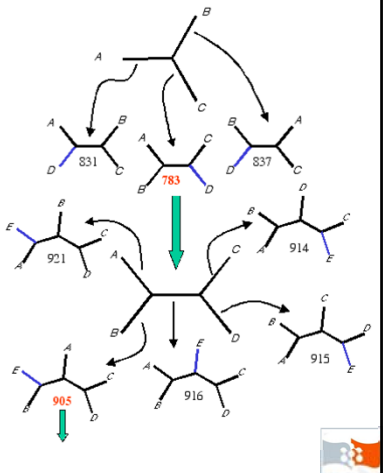
This is more extensive than start decomposition. When performing stepwise addition one starts with three taxa and then adds one at a time at the branch (trying each of the possible branches at the tree from the previous step) which gives the best tree according to the optimality criterion.

The order that the taxa are added can be varied in different ways

- randomly drawn from the remaining (not yet included) taxa
- the order in the data matrix
- a ranked list based on the average difference to the other taxa
- try each of the remaining taxa one at a time (at all possible branches) and pick the one that gives the best tree at each step

Random tree

- If we have a tree-scope with lots of peaks, a good strategy might be to simply generate a number of tree randomly and use those as starting points. Usually stepwise addition with random addition order performs better, though.



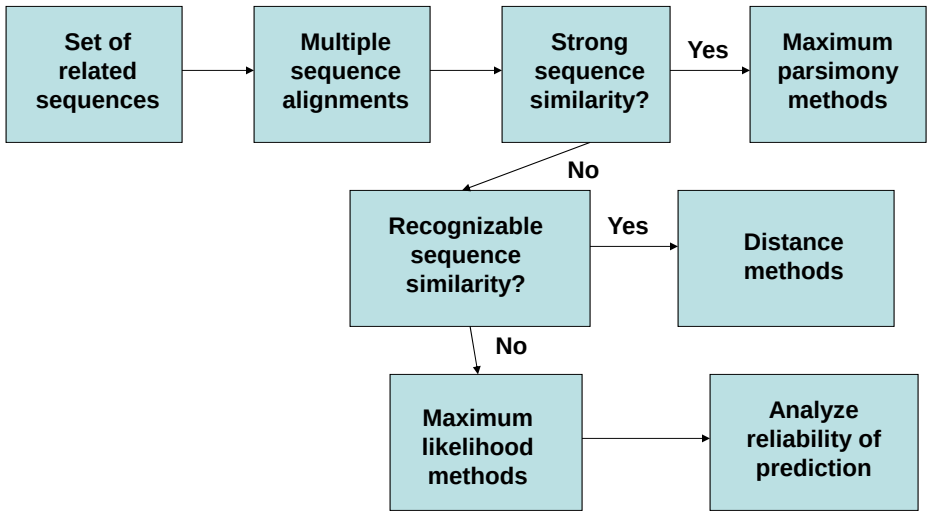
Finding the tree - Search methods

- Step 2 - Improving the initial tree(s)
- **This step is commonly referred to as *branch swapping*. The procedure is to move subtrees to other parts of the tree in order to increase the tree score. Depending on how extensive the swapping is we have (in order of increasing extensiveness):**
- Nearest Neighbour Interchange, NNI
- Subtree Pruning and Regrafting, SPR
- Tree Bisection and reconnection, TBR

Comparison of Methods

Neighbor-joining	Maximum parsimony	Maximum likelihood
Uses only pairwise distances	Uses only shared derived characters	Uses all data
Minimizes distance/length between nearest neighbors	Minimizes total distance/length	Maximizes tree likelihood given specific parameter values
Very fast	Slow	***Very*** slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, homoplasy rare)	Good for very small data sets and for testing trees built using other methods

Methods for phylogenetic trees construction: Overview



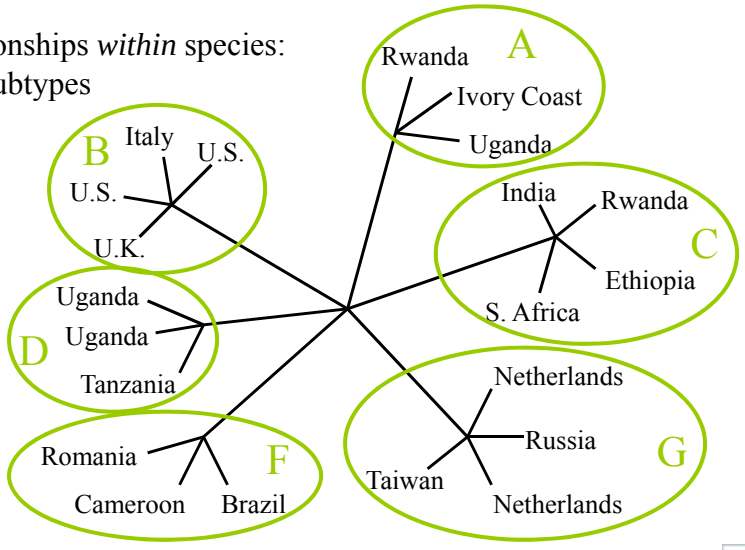
Phylogenetics applied: Forensics

Florida dentist case

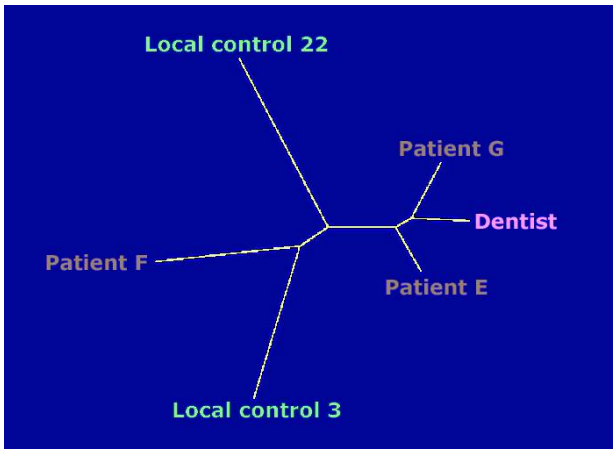
- 1990 case: Did a patient's HIV infection result from an invasive dental procedure performed by an HIV+ dentist?
- HIV evolves so fast that transmission patterns can be reconstructed from viral sequence (molecular forensics).
- Compared viral sequence from the dentist, three of his HIV+ patients, and two HIV+ local controls.

Phylogenetic trees of HIV subtypes

Relationships *within* species:
HIV subtypes



Florida dentist case results



- HIV sequences from HIV+ patients **G** and **E** were closer to the dentist than to the local controls
- HIV+ patient **F** contracted HIV likely from another source

Phylogenetics applied: Phylogeography

- Avise (2000):

  "field of study concerned with the principles and processes governing the geographical distributions of genealogical lineages…"

  "time and space are the jointly considered axes of phylogeography onto which (ideally) are mapped particular gene genealogies…"

- Spatial differentiation:

  Evolution vs. Dispersal or Vicariance

Phylogenetics applied

Evolution and Speciation

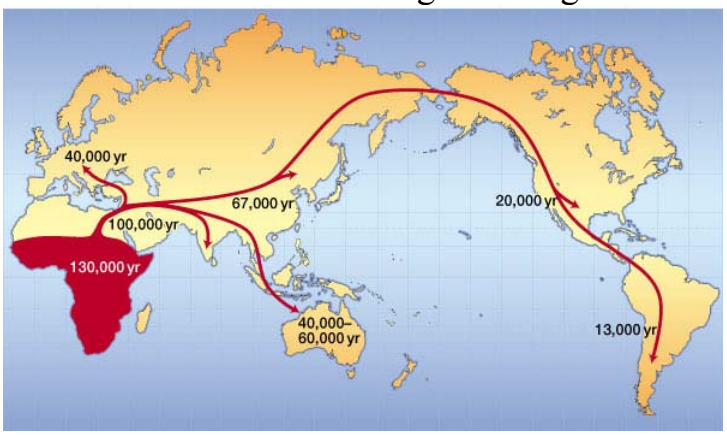
- Monophyletic vs. Polyphyletic or Paraphyletic origin

Conservation

- Evolutionary significant units (ESU)

Phylogenetics applied: Human origin

- Recent African vs. Multiregional origin



Evidence for a human mitochondrial origin in Africa: African sequence diversity is twice as large as that of non-African

Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. (2000) *Nature* 408: 708-713.

Question

Q: If we can accurately identify the evolutionary history of a species, can we extrapolate and predict future directions in its evolution?

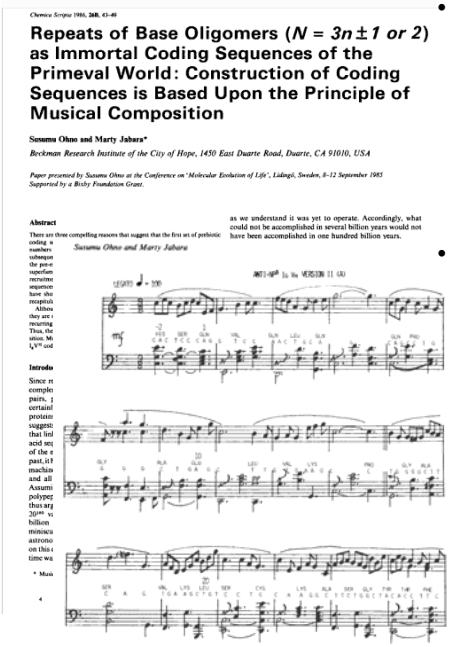
Summary

- Multiple sequence alignment gives us the opportunity to calculate evolutionary distances between sequences
- Phylogenetic trees have several different formats, some of which are stylistic and others, which convey information
- The optimal phylogenetic tree is hard to find, but there are several good ways of approximating it

Molecular Evolution and Music DNA

- Susumu Ohno (1928-2000) **suggested that repetition is a process that governs both Western music and DNA sequences: "*The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition*".**
- **Dan Graur, Ph.D.**
  John and Rebecca Moores Professor
  Department of Biology and Biochemistry University of Houston http://nsm.uh.edu/~dgraur/

Molecular Evolution and Music DNA



Repeats of Base Oligomers ($N = 3n \pm 1$ or 2) as Immortal Coding Sequences of the Primeval World: Construction of Coding Sequences is Based Upon the Principle of Musical Composition

Susumu Ohno and Marty Jabara*

Beckman Research Institute of the City of Hope, 1450 East Duarte Road, Duarte, CA 91010, USA

Paper presented by Susumu Ohno at the Conference on 'Molecular Evolution of Life', Lidingö, Sweden, 8–12 September 1985
Supported by a Bixby Foundation Grant.

Abstract

as we understand it was yet to operate. Accordingly, what could not be accomplished in several billion years would not have been accomplished in one hundred billion years.

Susumu Ohno and Marty Jabara

- The traditional musical composition embodied in sonata form consists of: (1) the exposition, in which the principal and secondary subjects are presented; (2) the development, in which one or both subjects are developed or worked out; (3) the recapitulation, in which both subjects are repeated by a coda (finale).
- Some nucleotide motifs (such as some decamers at the beginning and recapitulated near the end) can be considered as the principal subject, while the tandemly recurring motifs can be considered as the secondary subject.

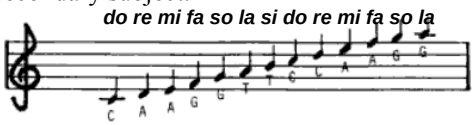
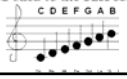


Fig. 1. Assignment of two alternative positions to four bases, A, G, T, and C in the ascending order, in the treble clef stave. This invariant rule permits the treble clef musical score to be transformed back to the base sequence with no ambiguity whatsoever.



Molecular Evolution and Music DNA

Here are some links to DNA/protein music:

- Algorithmic Arts: http://www.algoart.com, **by John Dunn. 1992. 1995. 1998. Protein Music:** http://whozoo.org/mac/Music/, **by Mary Ann Clark (Texas Wesleyan Univ). it contains a very nice** annotated source list of genetic music.
- NDB Musical Atlas : http://ndb-mirror-2.rutgers.edu/NDB/archives/MusicAtlas/index.html
- The Music of DNA: The Building Block of Life: http://www.healingmusic.org/SusanA/, **by Susan Alexjander in Partnership with Biologist David Deamer**
- Midi Music from DNA, Proteins and Math: http://education.llnl.gov/msds/music/, **Ron Rusay,Veikko Keranen, Tomi Laakso, Erik Jensen, Thomas Dunham. 1998.**
- DNA Music: http://www.dnamusic.com/, **Metamusic with Hemi-Sync® Sound Technology from The Monroe Institute featuring. by Barbara Bullard, Professor of Speech Communication Orange Coast College, Costa Mesa, CA. 1998-2002.**
- DNA Music Central - Human Genetic Code in Sound: http://www.dnamusiccentral.com/, **by Henry Alan Hargrove. 2001-2002.**
- Peter Gena's Home Page : http://www.petergena.com/
- Computational Biology - Applications - ProteinMusic: http://www.aber.ac.uk/~phiwww/pm/, **by Ross D King and A Karwath, Univ of Wales at Aberystwyth. 1996.**
- S2 Translation: http://www.nemeton.com/axis-mutatis/s2.html, **collaboration between Ross King and the band *Shamen* . "S2 is the receptor protein for 5-hydroxy tryptamine (Serotonin) and presumably for other tryptamines as well. It is thus one of the most important molecules in the mediation of both ordinary and non-ordinary (or "Shamanic") states of consciousness, which is why the molecule was chosen for this piece." - Colin Angus**
- DNA sequences - transposed into music : http://**www.mypage.bluewin.ch/molart/hugo.html**, **conversion by Daniel Schumperli (molecular biologist, clarinet), Lukas Frey (geographer, contrabass), and Rudolf von Steiger (space physicist, computer). 2001 (Switzerland).**
- Gene Music and Sangen Studio : http://www.toshima.ne.jp/%7Eedogiku/index.html, **by Nobuo Munakata, Kenshi Hayashi (Japan).**
- Genome Music : http://www.toddbarton.com/, **composer Todd Barton. 2001**.
- Molecular Music: http://www.molecularmusic.com, **by Dr. Linda Long at Exeter Univerity. mapping protein structure(!) to music. 2001.**
- AudioGenetics, Inc.: http://www.audiogenetics.com/, **founded by David Lane. 1998**.
- Sophia's Garden : http://www.sophiasgarden.org/music.html. **This piece of DNA music was created by Herb Moore for the** Sophia's Garden Foundation. **2003.**