Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

Phylogenetics

naydin@yildiz.edu.tr
http://www3.yildiz.edu.tr/~naydin

What is Molecular Phylogenetics…

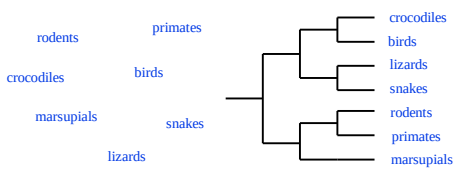
- Phylogenetics
  - the study of evolutionary relationships in organisms,
  - one part of the larger field of systematics, which also includes taxonomy.
    - The term taxonomy connotes the process and methodology for the naming and classification of organisms.
- The systematics
  - the branch of biology that deals with classification and nomenclature; taxonomy

…What is Molecular Phylogenetics…

- The context of evolutionary biology is phylogeny,
  - the connections between all groups of organisms as understood by ancestor/descendant relationships.
- The molecular mechanisms of organisms studied strongly suggests that all organisms on earth have a common ancestor.
  - Thus, the species are related to each other by the virtue of having evolved from the same common ancestor.
- Such a relationship of species is called phylogeny and it's graphical representation is called a phylogenetic tree.

…What is Molecular Phylogenetics

- Example:
  - relationship among species



Etymology

- phyl-(or phylo-)
  - Latin, from Greek, from phylē, phylon.
    - tribes, races or phyla
- phyla
  - a direct line of descent within a group
- genetics (genetic + -ics)
  - From Greek genetikos from genesis (origin)
    - laws of origination (1872)
    - study of heredity (from 1891)
- phylogeny
  - the evolutionary history of a kind of organism
  - the evolution of a genetically related group of organisms
- phylogenetics
  - a branch of science that deals with phylogeny

A Brief History of Molecular Phylogenetics

- 1900s
  - Immunochemical studies
    - cross-reactions stronger for closely related organisms
      - Nuttall (1902) - apes are closest relatives to humans!
- 1960s - 1970s
  - Protein sequencing methods, electrophoresis, DNA hybridization and Polymerase Chain Reaction (PCR) contributed to a boom in molecular phylogeny
- late 1970s to present
  - Discoveries using molecular phylogeny
    - Endosymbiosis - Margulis, 1978
    - Divergence of phyla and kingdom - Woese, 1987
    - Many Tree of Life projects completed or underway

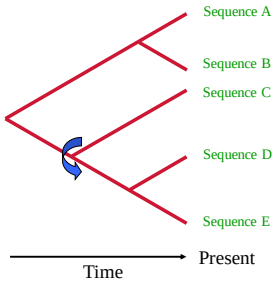
Molecular data vs. Morphology/Physiology

- Strictly heritable entities
- Data is unambiguous
- Regular & predictable evolution
- Quantitative analyses
- Ease of homology assessment
- Relationship of distantly related organisms can be inferred
- Abundant and easily generated with PCR and sequencing

- Can be influenced by environmental factors
- Ambiguous modifiers: "reduced", "slightly elongated", "somewhat flattened"
- Unpredictable evolution
- Qualitative argumentation
- Homology difficult to assess
- Only close relationships can be confidently inferred
- Problems when working with micro-organisms and where visible morphology is lacking

7

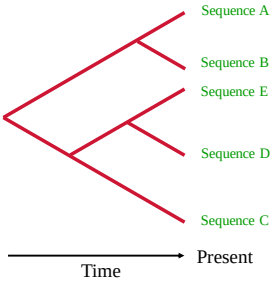
Phylogenetic concepts: Interpreting a Phylogeny



- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

8

Phylogenetic concepts: Interpreting a Phylogeny



- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

9

Tree Terminology

- Relationships are illustrated by a phylogenetic tree / dendrogram
  - Combination of Greek dendro/tree and gramma/drawing
  - A dendrogram is a tree diagram
    - frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.
  - Dendrograms are often used in computational biology
    - to illustrate the clustering of genes or samples, sometimes on top of heatmaps.

10

Tree Terminology

- A cladogram is a type of phylogenetic tree that only shows tree topology
  - the shape indicating relatedness.
    - It shows that, say, humans are more closely related to chimpanzees than to gorillas,
      - but not the time or genetic distance between the species.
  - Combination of Greek clados/branch and gramma/drawing
    - topology
      - the branching structure of the tree

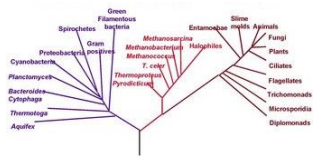
11

Tree Terminology

- The branching pattern is called the tree's topology
- Trees can be represented in several forms:

Rectangular cladogram

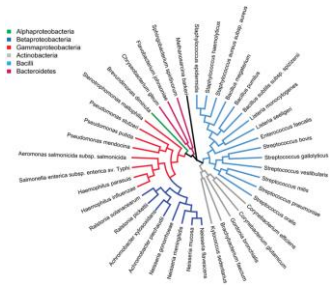
Slanted cladogram



12

Tree Terminology

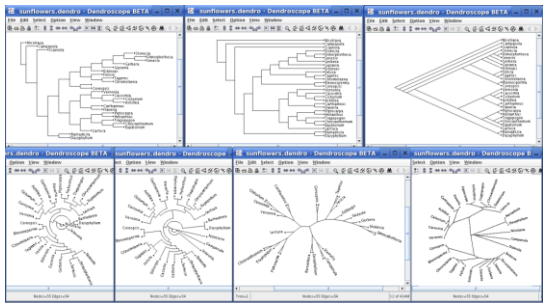
- Circular cladogram



13

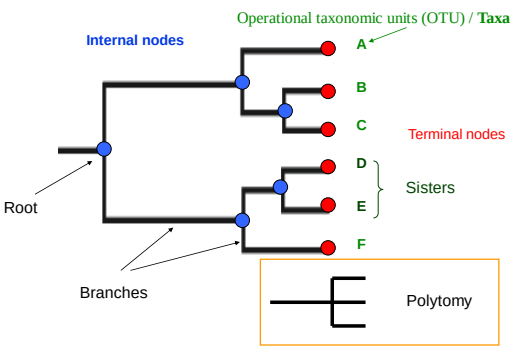
Same tree - seven different views

Rectangular Phylogram, Rectangular Cladogram, Slanted Cladogram, Circular Phylogram, Circular Cladogram, Radial Phylogram and Radial Cladogram



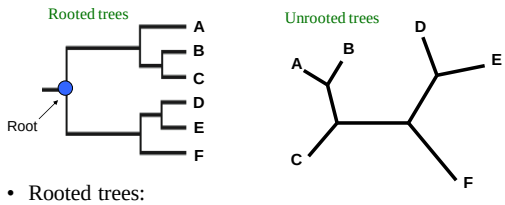
14

Tree Terminology



15

Tree Terminology



- Rooted trees:
  - has a root that denotes common ancestry
- Unrooted trees:
  - Only specifies the degree of kinship among taxa but not the evolutionary path

Taxon, plural taxa. (taxonomy): Any group or rank in a biological classification into which related organisms are classified.

16

Number of trees

- The number of rooted trees for n species:

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

- The number of unrooted trees for n species:

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}$$

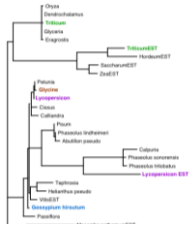
- Number of possible rooted and unrooted trees that can describe the possible relationships among fairly small numbers of data sets.

Number of Data Sets	Number of Rooted Trees	Number of Unrooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
10	34,459,425	2,027,025
15	213,458,046,767,875	7,905,853,580,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

17

Tree Terminology

- Scaled trees:

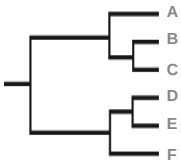


  - Branch lengths are proportional to the number of nucleotide/amino acid changes that occurred on that branch (usually a scale is included).
- In the best of cases, scaled trees are also additive, meaning that
  - the physical length of the branches connecting any two nodes is an accurate representation of their accumulated differences.

18

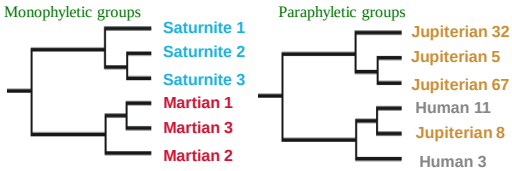
Tree Terminology

- Unscaled trees:



  - Branch lengths are not proportional to the number of nucleotide/amino acid changes
    - usually used to illustrate evolutionary relationships only.
  - line up all terminal nodes and convey only their relative kinship without making any representation regarding the number of changes that separate

Tree Terminology



- Monophyletic groups:
  - All taxa within the group are derived from a single common ancestor and members form a natural clade.
- Paraphyletic groups:
  - The common ancestor is shared by other taxon in the group and members do not form a natural clade.

Gene vs. Species Trees

- Gene tree
  - a phylogenetic tree based on the divergence observed within a single homologous gene.
    - Such trees may represent the evolutionary history of a gene but not necessarily that of the species in which it is found.
- Species trees
  - best obtained from analyses that use data from multiple genes.
    - For example, a study on the evolution of plant species used more than 100 different genes to generate a species tree for plants.

Character and Distance Data

- The molecular data used to generate phylogenetic trees fall into one of two categories:
  - Characters
    - a well-defined feature that can exist in a limited number of different states
  - Distances
    - a measure of the overall, pairwise difference between two data sets
- Both DNA and protein sequences are examples of data that describe a set of discrete character states.

Methods in Phylogenetic Reconstruction

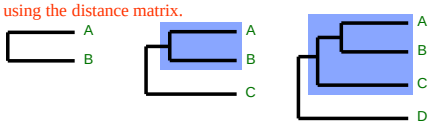
- Distance Based Methods
  - calculate pairwise distances between sequences, and group sequences that are most similar.
  - This approach has potential for computational simplicity and therefore speed
- Character Based Methods (Maximum parsimony)
  - assumes that shared characters in different entities result from common descent.
  - Groups are built on the basis of such shared characters, and the simplest explanation for the evolution of characters is taken to be the correct, or most parsimonious one.
- Probabilistic Methods (Maximum likelihood)
  - compute the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution.

Comparison of Methods

Distance	**Maximum parsimony**	**Maximum likelihood**
• Uses only pairwise distances	• Uses only shared derived characters	• Uses all data
• Minimizes distance between nearest neighbors	• Minimizes total distance	• 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 in Phylogenetic Reconstruction

- Distance Based Methods
  - Using a sequence alignment, pairwise distances/dissimilarities are calculated
  - Creates a distance/dissimilarity matrix
  - A phylogenetic tree is calculated with clustering algorithms, using the distance matrix.



  - Examples of clustering algorithms include
    - Unweighted Pair Group Method using Arithmetic averages (UPGMA)
    - Neighbor Joining clustering.

UPGMA

- Unweighted-Pair-Group Method with Arithmetic mean (UPGMA)
  - Oldest and simplest distance matrix method
  - Originally proposed in the early 1960s to help with the evolutionary analysis of morphological characters,
  - requires data that can be condensed to a measure of genetic distance between all pairs of taxa being considered.
  - requires a distance matrix such as one that might be created for a group of 4 taxa called A, B, C, and D.

UPGMA

- Assume that the pairwise distances between each of the taxa are given in the following matrix:

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

  - d_{AB} : distance between species A and B
  - d_{AC} : distance between species A and C
  - ...

UPGMA

- UPGMA begins by clustering the two species with the smallest distance separating them into a single, composite group.
  - Assume that the smallest value in the distance matrix corresponds to d_{AB} in which case species A and B are the first to be grouped (AB).
    - After the first clustering, a new distance matrix is computed with the distance between the new group (AB) and species C and D being calculated as

$$d_{(AB)C} = \frac{d_{AC}+d_{BC}}{2} \text{ and } d_{(AB)D} = \frac{d_{AD}+d_{BD}}{2}$$

UPGMA

- The species separated by the smallest distance in the new matrix are then clustered to make another new composite species.
- The process is repeated until all species have been grouped.
  - If scaled branch lengths are to be used on the tree to represent the evolutionary distance between species, branch points are positioned at a distance halfway between each of the species being grouped
    - i.e., at $d_{AB}/2$ for the first clustering

UPGMA - example

- Consider the following alignment between five different DNA sequences

```
          10        20        30        40        50
A:  GTGCTGCACGG CTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
B:  ACGCTGCACGG CTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
C:  GTGCTGCACGG CTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
D:  GTATCACACGA CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
E:  GTATCACATAG CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

- Pairwise distance matrix

Species	A	B	C	D
B	9	-	-	-
C	8	11	-	-
D	12	15	10	-
E	15	18	13	5

  - Smallest distance:
    - d_{DE} , so
      - Species D and species E are grouped

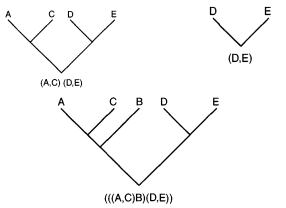
UPGMA - example

Species	A	B	C	D
B	9	-	-	-
C	8	11	-	-
D	12	15	10	-
E	15	18	13	5

Species	A	B	C
B	9	-	-
C	8	11	-
DE	$\frac{12+15}{2}=13.5$	$\frac{15+18}{2}=16.5$	$\frac{10+13}{2}=11.5$

Species	B	AC
AC	$\frac{9+11}{2}=10$	-
DE	$\frac{15+18}{2}=16.5$	$\frac{13.5+11.5}{2}=12.5$

Species	(AC)B
(AC)B	–
DE	$\frac{16.5+12.5}{2}=19.5$



31

Estimation of Branch Lengths

- Tree describes the relatedness of sequences
- It is possible for the topology of phylogenetic trees to convey information about
  - the relative degree to which sequences have diverged.
  - Scaled trees that convey that information, often referred to as cladograms,
    - the length of branches correspond to the inferred amount of time that the sequences have been accumulating substitutions independently.

32

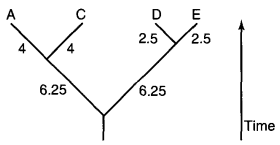
Estimation of Branch Lengths

- The relative length of each branch in a cladogram can be calculated using the information in a distance matrix.
  - In the example, the d_{DE} is 5,
    - the pair of branches connecting each of those species to their common ancestor should each be $d_{DE}/2$ or 2.5 units long on a tree with scaled branch lengths.
  - A and C should be connected to their common ancestor by branches that are $d_{AC}/2$ or 4 units long,
  - The branch point between (AC) and (DE) should be connected to (AC) and (DE) by branches that are both $d_{(AC)(DE)}/2$ or 6.25 units long,

33

Estimation of Branch Lengths

- A scaled tree showing the branch lengths separating four of the species depicted in slide 30.

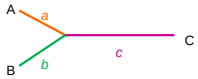


  - Branch lengths are shown next to each branch.
  - Branches are also drawn to scale to reflect the amount of differences between all species.
- This very simple approach to estimating branch lengths actually allows UPGMA to intrinsically generate rooted phylogenetic trees.

34

Estimation of Branch Lengths

- Determining branch lengths for a scaled tree is only slightly more complicated
  - when it cannot be assumed that evolutionary rates are the same for all lineages.



  - The simplest tree whose branch lengths might have some meaningful information is one with just three species (A, B, C) and one branch point, such as the one shown.
- On such a tree, the length of each of the three branches can be represented by a single letter (a, b, and c) for which we know the following must be true:

$d_{AB} = a + b$; $d_{BC} = b + c$; $d_{AC} = a + c$

35

Estimation of Branch Lengths

- Phylogeny reconstruction for 3 sequences
  - There is a single tree topology
  - The branch lengths (a, b, c) :

$a + b = d_{AB}$

$b + c = d_{BC}$

$a + c = d_{AC}$

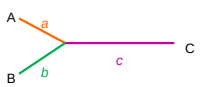
	A	B	C
A	-	$a+b$	$a+c$
B	-	-	$b+c$
C	-	-	-

- Input:
  - d_{AB}, d_{BC} and d_{AC} (pairwise distances)
- Output:

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

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

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

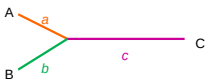


36

Estimation of Branch Lengths - example

- Distance matrix of 3 sequences and unrooted tree

	A	B	C
A	--	22	39
B	--	--	41
C	--	--	--



  - distance from A to B = $a + b = 22$ (1)
  - distance from A to C = $a + c = 39$ (2)
  - distance from B to C = $b + c = 41$ (3)
- subtracting (3) from (2) yields:

  $b + c - (a + c) = b - a = 41 - 39 = 2$ (4)

Estimation of Branch Lengths - example

- adding (1) and (4) yields

  $a + b + b - a = 2b = 22 + 2 = 24$

  $2b = 24$

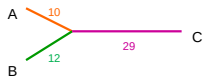
  $b = 24 / 2 = 12$
- so

  $a + b = a + 12 = 22;$

  $a = 22 - 12 = 10$
- finally

  $a + c = 10 + c = 39;$

  $c = 39 - 10 = 29$

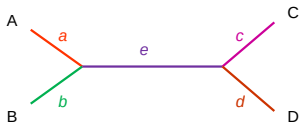


Neighbor's Relation Method

- Popular variant of the UPGMA method
- emphasizes pairing species in such a way that
  - a tree is created with the smallest possible branch lengths overall.
- On any unrooted tree, pairs of species that are separated from each other by just one internal node are said to be neighbors.

Neighbor's Relation Method

- The topology of a phylogenetic tree such as the one shown below gives rise to some useful algebraic relationships between neighbors.



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

Species	A	B	C
B	$a+b$	-	-
C	$a+e+c$	$b+e+c$	-
D	$a+e+d$	$b+e+d$	$c+d$

- If the tree above is a true tree for which additivity holds, then the following should be true:

  $d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2e = d_{AB} + d_{CD} + 2e$
  - where a, b, c, and d are the lengths of the terminal branches and e is the length of the single central branch.

Neighbor's Relation Method

- The following conditions, known together as the four-point condition, will also be true:

  $d_{AB} + d_{CD} < d_{AC} + d_{BD}$; $d_{AB} + d_{CD} < d_{AD} + d_{BC}$
- It is in this way that a neighborliness approach considers all possible pairwise arrangements of four species and determines which arrangement satisfies the four-point condition.
  - An important assumption of the four-point condition is that branch lengths on a phylogenetic tree should be additive and, while it is not especially sensitive to departures from that assumption, data sets that are not additive can cause this method to generate a tree with an incorrect topology.

Neighbor's Relation Method - example

- Consider the alignment:

  A ACGCGTTGGGCGATGGCAAC

  B ACGCGTTGGGCGACGGTAAT

  C ACGCATTGAATGATGATAAT

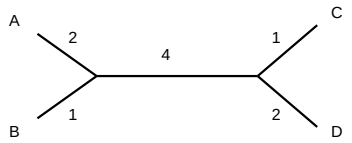
  D ACACATTGAGTGATAATAAT
- The distances between these sequences can be shown as a table:

	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-

Estimation of Branch Lengths - example

- Using this information, an unrooted tree showing the relationship between these sequences can be drawn:

	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-



43

Neighbor-Joining Methods

- A variant of neighborliness
- an agglomerative technique, and so operates using iteration,
  - building the tree from the bottom-up
- Start with a star-like tree with all species coming off a single central node regardless of their number.
- Neighbors are then sequentially found that minimize the total length of the branches on the tree.

44

Neighbor-Joining Methods

- The input is an $n \times n$ dissimilarity/distance matrix d.
- In the first iteration,
  - the n leaves are all in their own clusters;
- In subsequent iterations,
  - each cluster is a set of leaves,
    - but the clusters are disjoint.
- At the beginning of each iteration, the taxa are partitioned into clusters, and for each cluster we have a rooted tree that is leaf-labelled by the elements in the cluster.

45

Neighbor-Joining Methods

- During the iteration, a pair of clusters is selected to be made siblings;
  - this results in the trees for the two clusters being merged into a larger rooted tree by making their roots siblings.
- When there are only three subtrees, then the three subtrees are merged into a tree on all the taxa by adding a new node, r, and making the roots of the three subtrees adjacent to r.
- Note that this description suggests that neighbor joining produces a rooted tree.
  - However, after the tree is produced, the root is ignored, so that neighbor joining actually returns an unrooted tree.

46

Neighbor-Joining Methods

The neighbor joining algorithm. *Input:* $n \times n$ dissimilarity matrix $\mathbf{d}$ with $n \geq 4$
Output: Unrooted tree with n leaves labelled $1 \ldots n$

Initialization: Compute the $n \times n$ matrix $\mathbf{Q}$, defined by

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n}(d_{ik} + d_{jk}).$$

While $n > 3$, DO:

Find the pair i, j minimizing $Q_{i,j}$. Without loss of generality, we will call that pair a, b.
Make the rooted trees associated with taxa a and b siblings, and call the root of the tree you form u.
Update the distance matrix by deleting the rows and columns for a and b, and including a new row and column for u, and set $d_{u,k} = \frac{d_{ak} + d_{bk} - d_{ab}}{2}$ for all $k \neq u$. Decrement n by 1.

Now $n = 3$; return the star tree with a single internal node v where the roots of the three rooted trees are all adjacent to v.

- The neighbor-joining method: a new method for reconstructing phylogenetic trees. (1987). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040454

47

Character-Based Methods of Phylogenetics

- The concept of **parsimony** is at the very heart of all character-based methods of phylogenetic reconstruction.
  - Parsimonious: someone who was especially careful with the spending of their money.
- In a biological sense, parsimony is used to describe
  - the process of attaching preference to one evolutionary pathway over another on the basis of which pathway requires the invocation of the smallest number of mutational events.

48

Character-Based Methods of Phylogenetics

- Phylogenetic trees represent theoretical models that depict the evolutionary history of 3 or more sequences.
- The two premises that underlie biological parsimony are quite simple:
  - Mutations are exceedingly rare events
  - The more unlikely events a model invokes, the less likely the model is to be correct.
- As a result, the relationship that requires the fewest number of mutations to explain the current state of the sequences being considered is the relationship that is most likely to be correct.

Character-Based Methods of Phylogenetics

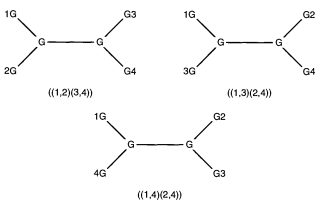
- Informative and Uninformative Sites

Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

- The short four-way multiple sequence alignment shown above contains positions that fall into two categories in terms of their information content for a parsimony analysis:
  - those that have information (are informative)
  - those that do not (are uninformative).
- The relationship between four sequences can be described by only three different unrooted trees

Character-Based Methods of Phylogenetics

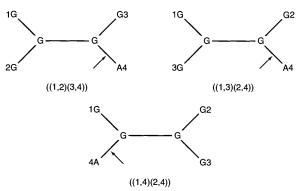
Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 1: All 4 sequences have the same character (a "G") and the position is therefore said to be invariant.
  - Invariant sites are clearly uninformative because each of the three possible trees that describe the relationship of the four sequences invokes exactly the same number of mutations (0)

Character-Based Methods of Phylogenetics

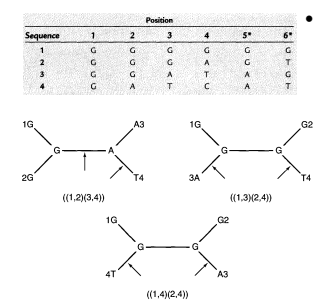
Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 2: Uninformative from a parsimony perspective because one mutation occurs in all three of the possible trees.

Character-Based Methods of Phylogenetics

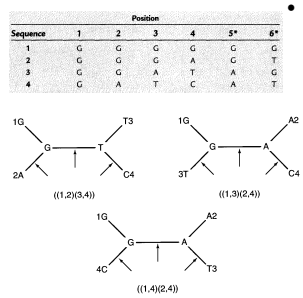
Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 3: Uninformative because all three trees require two mutations

Character-Based Methods of Phylogenetics

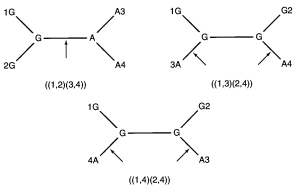
Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 4: Uninformative because all three trees require three mutations.

Character-Based Methods of Phylogenetics

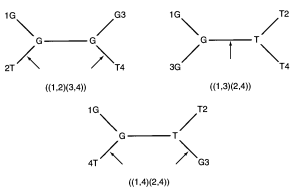
Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 5: Informative because one of the three trees invokes only one mutation and the other two alternative trees both require two mutations

Character-Based Methods of Phylogenetics

Sequence	Position 1	2	3	4	5*	6*
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



- Position 6: Informative because one of the three trees invokes only one mutation and the other two alternative trees both require two mutations

Character-Based Methods of Phylogenetics

- In general, for a position to be informative regardless of how many sequences are aligned,
  - it has to have at least two different nucleotides
  - each of these nucleotides has to be present at least twice.
- All parsimony programs begin by applying this fairly simple rule to the data set being analyzed.
- Notice that 4 of the 6 positions being considered in the alignment shown in slide 50 are simply discarded and not considered any further in a parsimony analysis.
  - All of those sites would have contributed to the pairwise similarity scores used by a distance-based approach, and this difference alone can generate substantial differences in the conclusions reached by both types of approaches.

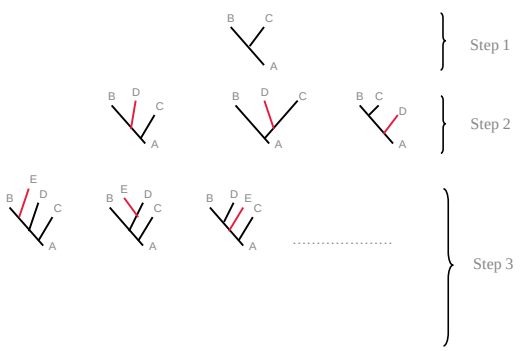
Character-Based Methods of Phylogenetics

- Once uninformative sites have been identified and discarded, implementation of the parsimony approach in its simplest form can be straightforward.
- Every possible tree is considered individually for each informative site.
- A running tally is maintained for each tree that keeps track of the minimum number of substitutions required for each position.
- After all informative sites have been considered, the tree (or trees) that needs to invoke the smallest total number of substitutions is labeled the most parsimonious.
  - https://paup.phylosolutions.com/

Character-Based Methods of Phylogenetics

- Maximum Parsimony
  - All possible trees are determined for each position of the sequence alignment
  - Each tree is given a score based on the number of evolutionary step needed to produce said tree
  - The most parsimonious tree is the one that has the fewest evolutionary changes for all sequences to be derived from a common ancestor
  - Usually several equally parsimonious trees result from a single run.

Maximum parsimony: exhaustive stepwise addition



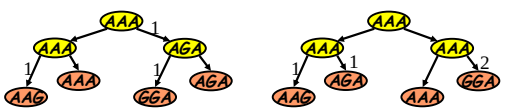
Maximum Parsimony - example

- Maximum parsimony methods predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences.
  - First, a multiple sequence alignment must first be obtained.
- For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes are identified.

Maximum Parsimony - example

- This continues for each position in the alignment.
- Those trees that produce the smallest number of changes overall for all sequence positions are identified.
  - This is a rather time consuming algorithm that only works well if the sequences have a strong sequence similarity.

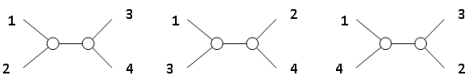
Maximum Parsimony - example



- The left tree is preferred over the right tree.

Maximum Parsimony - example

- Assuming we have 4 sequences
  - There are 3 possible trees:



- The optimal tree is obtained by adding the number of changes at each informative site for each tree, and picking the tree requiring the least total number of changes.
- For a large number of sequences the number of trees to examine becomes so large that it might not be possible to examine all possible trees.

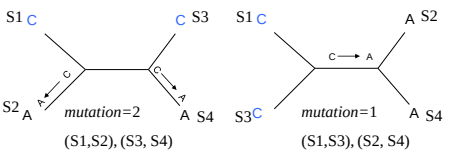
Maximum Parsimony - example

- Consider the following sequences

	S1	C	A	C	C	C	C	T	T	
	S2	A	A	C	C	C	C	A	T	
	S3	C	A	C	T	G	C	T	T	
	S4	A	A	C	T	G	C	T	A	
(S1,S2),(S3,S4)		2	0	0	1	1	0	1	1	**6** √
(S1,S3),(S2,S4)		1	0	0	2	2	0	1	1	**7**
(S1,S4),(S2,S3)		2	0	0	2	2	0	1	1	**7**

Maximum Parsimony - example

	S1	C	A	C	C	C	C	T	T	
	S2	A	A	C	C	C	C	A	T	
	S3	C	A	C	T	G	C	T	T	
	S4	A	A	C	T	G	C	T	A	
(S1,S2),(S3,S4)		2	0	0	1	1	0	1	1	**6** √
(S1,S3),(S2,S4)		1	0	0	2	2	0	1	1	**7**



Methods in Phylogenetic Reconstruction

- Maximum Likelihood
  - Creates all possible trees like Maximum Parsimony method
  - But instead of retaining trees with shortest evolutionary steps,
    - employs a model of evolution whereby different rates of transition/transversion ratios can be used
  - Each tree generated is calculated for the probability that it reflects each position of the sequence data
  - Calculation is repeated for all nucleotide sites
  - Finally, the tree with the best probability is shown as the maximum likelihood tree
    - usually only a single tree remains
  - It is a more realistic tree estimation because it does not assume equal transition-transversion ratio for all branches.

67

Tree Confidence

- All phylogenetic trees represent hypotheses regarding the evolutionary history of the sequences that make up a data set.
- Like any good hypothesis, it is reasonable to ask two questions about how well it describes the underlying data:
  - How much confidence can be attached to the overall tree and its component parts (branches)?
  - How much more likely is one tree to be correct than a particular or randomly chosen alternative tree?
- To address these two questions, a powerful resampling technique called bootstrapping has become the predominant favorite for addressing the first question, and a simple parametric comparison of two trees is typical of those used to address the second.

68

Bootstrapping

- Bootstrap analysis is a kind of statistical analysis to test the reliability of certain branches in the evolutionary tree.
  - In statistics, it is any test or metric that relies on random sampling with replacement.
  - It falls in the broader class of resampling methods.
- It involves resampling one's own data, with replacement, to create a series of bootstrap samples of the same size as the original data.
- In the case of nucleic acid (amino acid) sequences, the resampled data are the nucleotides (amino acids) of a sequence while the statistical significance of a specific cluster is given by the fraction of trees, based on the resampled data, containing that cluster.

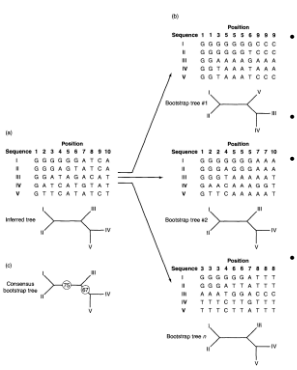
69

Bootstrapping

- Bootstrap tests allow for a rough quantification of those confidence levels.
- The basic approach of a bootstrap test is straightforward:
- A subset of the original data is drawn (with replacement) from the original data set and a tree is inferred from this new data set,
  - as illustrated in the next slide

70

Bootstrapping

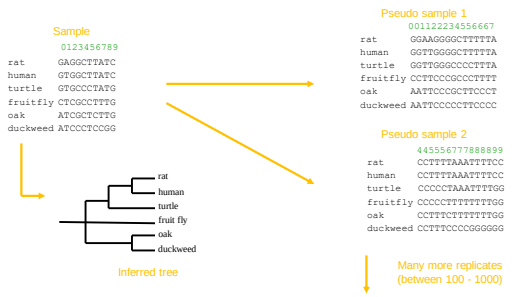


- Illustration of a boostrap test for a data set that is 10 characters long (positions: 1 through 10) and five sequences (I through V) deep.
- The original data set and its most parsimonious tree (a).
- A random number generator chooses 10 times from the original 10 positions and creates the three resampled data sets with their corresponding trees (b).
- A consensus tree (c) for the three resampled data sets is shown with circled values indicating the fraction of bootstrapped trees that support the clustering of sequences I and II and sequences IV and V.

71

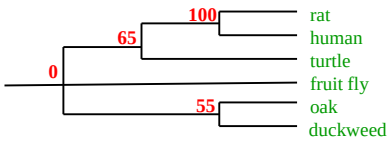
Bootstrapping

- Computational method to estimate the confidence level of a certain phylogenetic tree.



72

Bootstrap values



- Values are in percentages
- Conventional practice:
  - only values 60-100% are shown

73

Parametric Tests

- The underlying principle of parsimony suggests that the tree that invokes the smallest number of substitutions is the one that is most likely to depict the true relationship between the sequences.
- A common question is "How much more likely is the most parsimonious tree than a particular alternative that has previously been put forward to describe the relationship between these taxa?"

74

Parametric Tests

- One of the first parametric tests devised to answer such questions for parsimony analyses is that of H. Kishino and M. Hasegawa (1989).
  - Their test assumes that informative sites within an alignment are both independent and equivalent and uses the difference in the minimum number of substitutions invoked by two trees, D, as a test statistic.
- A value for the variance, V, of D is determined by considering each of the informative sites separately as follows:

$$V = \frac{n}{(n-1)} \sum_{k=1}^{n} \left[D_i - \frac{1}{n} \sum_{k=1}^{n} D_k \right]^2$$

where n is the number of informative sites.

75

Parametric Tests

- A paired t-test with n-1 degrees of freedom can then be used to test the null hypothesis that the two trees are not different from each other:

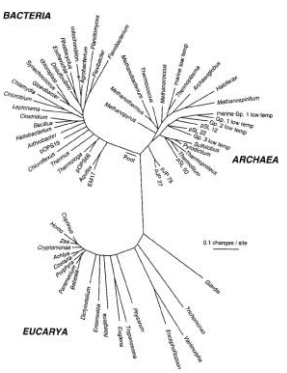
$$t = \frac{D}{n\sqrt{nV}}$$

- Alternative parametric tests are available not just for parsimony analyses but for distance matrix and maximum likelihood trees as well.

76

Some Discoveries Made Using Molecular Phylogenetics

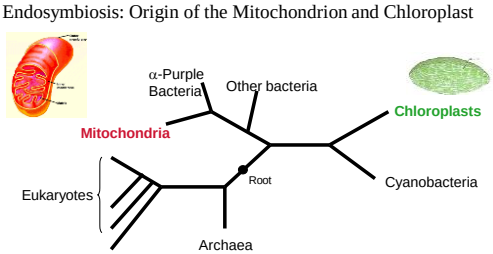
- Universal Tree of Life
  - Using rRNA sequences
  - Able to study the relationships of uncultivated organisms, obtained from a hot spring in Yellowstone National Park



77

Some Discoveries Made Using Molecular Phylogenetics

- Endosymbiosis: Origin of the Mitochondrion and Chloroplast

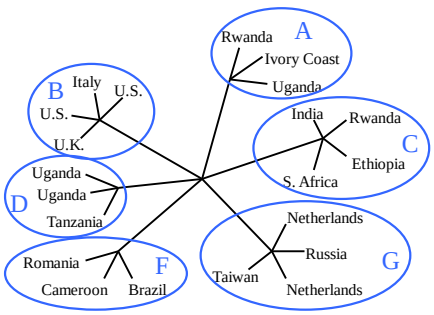


- **Mitochondria** and **chloroplasts** are derived from the α-purple bacteria and the cyanobacteria respectively, via separate endosymbiotic events.

78

Some Discoveries Made Using Molecular Phylogenetics

- Relationships within species: HIV subtypes



Problems and Errors in Phylogenetic Reconstruction

- Inherent strengths and weaknesses in different tree-making methodologies.
- More is better
  - Errors in inferred phylogeny may be caused by small data sets and/or limited sampling.
- Unsuitable sequences
  - those undergoing rapid nucleotide changes or slow to zero changes overtime may skew phylogenetic estimations

Problems and Errors in Phylogenetic Reconstruction

- Mutations:
  - Duplications, inversions, insertions, deletions etc. can give inaccurate signals
- Genomic hotspots:
  - small regions of rapid evolution are not easily detected
- Homoplasy:
  - nucleotide changes that are similar but occurred independently in separate lineages are mistakenly assumed as inherited changes
- Sample contamination / mislabeling:
  - always a possibility when working with large data sets