Phylogenetic Tree Construction- Introduction, Distance Based Methods

Manu Madhavan

Lecture 8

Recap

- Sequence alignment
- Pairwise- Global and Local
- Needleman-Wunsch, Smith-Waterman
- BLAST
- Multiple Sequence Alignment and CLUSTALW
- Scoring matrices- PAM

Outline

- Phylogenetic Tree
- Methods for reconstructing Phylogenetic Trees
- Distance based methods and Character based methods
- UPGMA algorithm

Molecular Phylogeny

- Searching for the evidences of important events in evolutionary history
- Evolutionary events- substitution, insertion, deletition,...
- **Phylogenetic Tree** showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor

Molecular Phylogeny

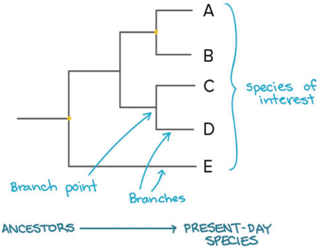
Molecular phylogeny is a relatively new scientific discipline that involves the comparative analysis of the nucleotide sequences of genes and the amino acid sequences and structural features of proteins from which evolutionary histories and relationships, and in some cases also functions, can be inferred.

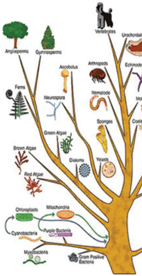
Molecular Phylogeny

- Earlier analysis were Phynotype to Genetype (top-down)
- Taxonomists were forced to rely on comparisons of phenotypes (how organisms looked) to infer their genotypes (the genes that gave rise to their physical appearance).
- This analysis have limitations: Sometimes similar phenotypes can evolve in organisms that are distantly related in a process called **convergent evolution**
- Molecular phylogenies are more reliable because the effects of natural selection are generally less pronounced at the sequence level.

Phylogenetic Tree

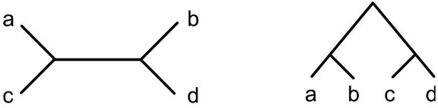
- A graphical representation of the evolutionary relationship among three or more genes or organisms
- Nodes and branches
- Terminal nodes represents genes/organisms
- Internal nodes represents common ancestors
- **Newick Format:** (((A,B),(C,D)),E)





Phylogenetic Tree

- Internal nodes either **bifurcating or multifurcating**
- **Scaled Trees:** ones in which branch lengths are proportional to the differences between pairs of neighboring nodes.
- Unsealed trees line up all terminal nodes and convey only their relative kinship without making any representation regarding the number of changes that separate them.
- Rooted and unrooted trees



(a)

(b)

Phylogenetic Tree

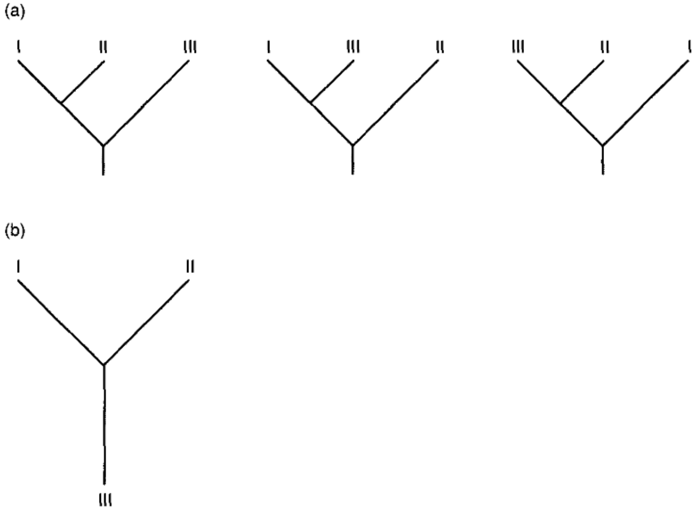
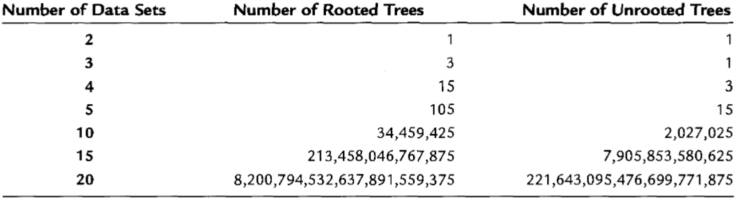


FIGURE 4.3 *All possible (a) rooted and (b) unrooted trees when only three species are considered.*

Phylogenetic Tree



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

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

Phylogenetic Tree

- Two molecular data used to construct :
  - Character: a well-defined feature that can exist in a limited number of different states
    - DNA/RNA/Protein sequence
    - Anatomically or behaviorally based features
  - Distance: a measure of the overall, pairwise difference between two data sets
    - Distance/similarity scores
    - Matching nucleotide
  - Pheneticists prefer distance based measures- which give light on evolutionary clock
  - Cladists prefer patterns in evolutionary pathways

UPGMA: Distance Based Methods

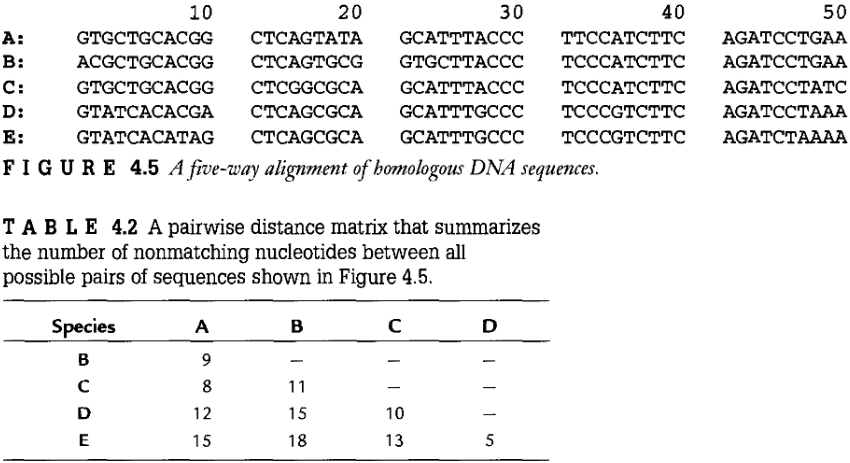
- Unweighted Pairwise group method with arithmetic mean
- Use genetic distance between taxa
- Input is a distance matrix

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

UPGMA

- UPGMA begins by clustering the two species with the smallest distance separating them into a single, composite group.
- 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 $d_{(AB)C} = 1/2(d_{AC} + d_{BC})$ and $d_{(AB)D} = 1/2(D_{AD} + d_{BD})$.
- 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

UPGMA



```
              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
```

FIGURE 4.5 *A five-way alignment of homologous DNA sequences.*

TABLE 4.2 A pairwise distance matrix that summarizes the number of nonmatching nucleotides between all possible pairs of sequences shown in Figure 4.5.

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

UPGMA

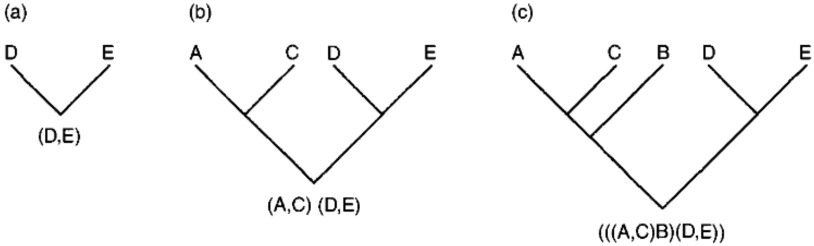


FIGURE 4.6 *A phylogenetic tree as it is constructed using the UPGMA method. (a) The first grouping (D,E) is between species D and E, which are connected by a single bifurcating branch. (b) The second grouping (A,C) is between species A and C, which are also connected to each other by a single bifurcating branch. (c) The last grouping ((A,C)B) unambiguously places the branching point for species B between that of the common ancestors of (A,C) and (D,E).*

Next

- Neighbor Joining Method
- Character based methods- Maximum Parsimony