BIO 285/CSCI 285/MATH 285
Bioinformatics
Programming Lecture 12
Phylogenetic Tree - UPGMA

Instructor: Lei Qian

Fisk University

Phylogenetic inference

1. Selection of sequences for analysis

- **DNA:**
  - Higher phylogenetic signal:
    - Synonymous vs nonsynonymous substitutions

- **Protein:**
  - Phylogenetic signal less predominant than in DNA
  - Better to construct a tree for evolutionary distant species or genes

- **RNA:** rRNA often used for constructing species trees

Phylogenetic inference

2. Multiple sequence alignment

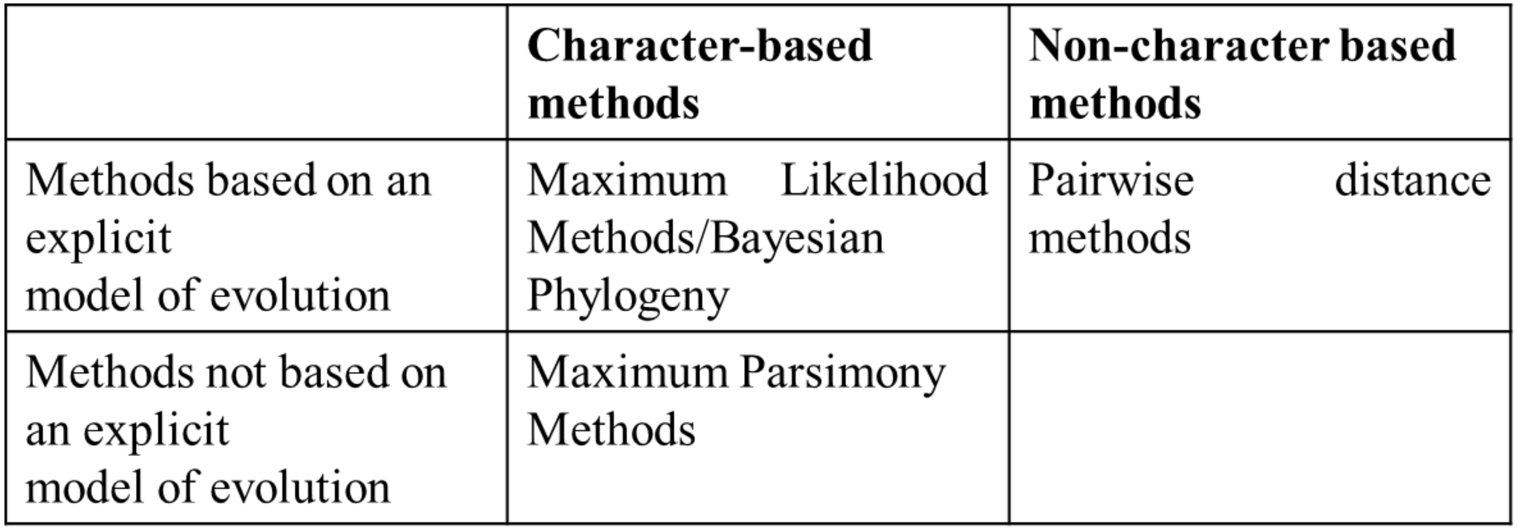
- This is a critical step in the analysis as in many cases the alignment of amino acids or nucleotides in a column implies that they share a common ancestor
- If you misalign a group of sequences you will still be able to produce a tree. However, it is not likely to be biologically meaningful.

Crap in is crap out!

- Inspect the alignment to be sure that all sequences are homologous
- Some times with ClustalW distantly related sequences are not well aligned. Try different gap and extension parameters to improve the alignment
- Only use these columns of the multiple alignment for which you have data for all organisms or sequences. Delete the columns for which this is not the case.
- Delete columns with gaps

Phylogenetic inference

3. Tree building



	Character-based methods	**Non-character based methods**
Methods based on an explicit model of evolution	Maximum Likelihood Methods/Bayesian Phylogeny	Pairwise distance methods
Methods not based on an explicit model of evolution	Maximum Parsimony Methods	

Phylogenetic Tree

Definition of the "best"

- **Distance-based**
  a distance is a measure of the overall differences/similarities between two objects
- **Character-based**
  a character is a characteristic that has well-defined, limited number of states
- **Maximum likelihood**
  Finds a tree such that the likelihood of the data given the tree structure is maximum

Distance based Methods

Distance based methods:

- calculate the distances between molecular sequences using some distance metric
- A clustering method (UPGMA, neighbour joining) is used to infer the tree from the pairwise distance matrix
- treat the sequence from a horizontal perspective, by calculating a single distance between entire sequences

- Advantage:
  - Fast
  - Allow using evolutionary models
- Disadvantage:
  - sequences reduced to one number

Distance based Methods

Character based methods:

- treat the sequences from a vertical perspective
- they search for each column of the alignment, the simplest explanation for how the characters evolved.
- For instance, MP involves a search for a tree with the fewest number of amino acid (or nucleotide character) changes that account for the observed differences between the protein (gene) sequences.

Distance Calculation

Approach:

- align pairs of sequences and count the number of differences (Hamming distance).
- For an alignment of length N with n sites at which there are differences: D= (n/N*100).

Problem:

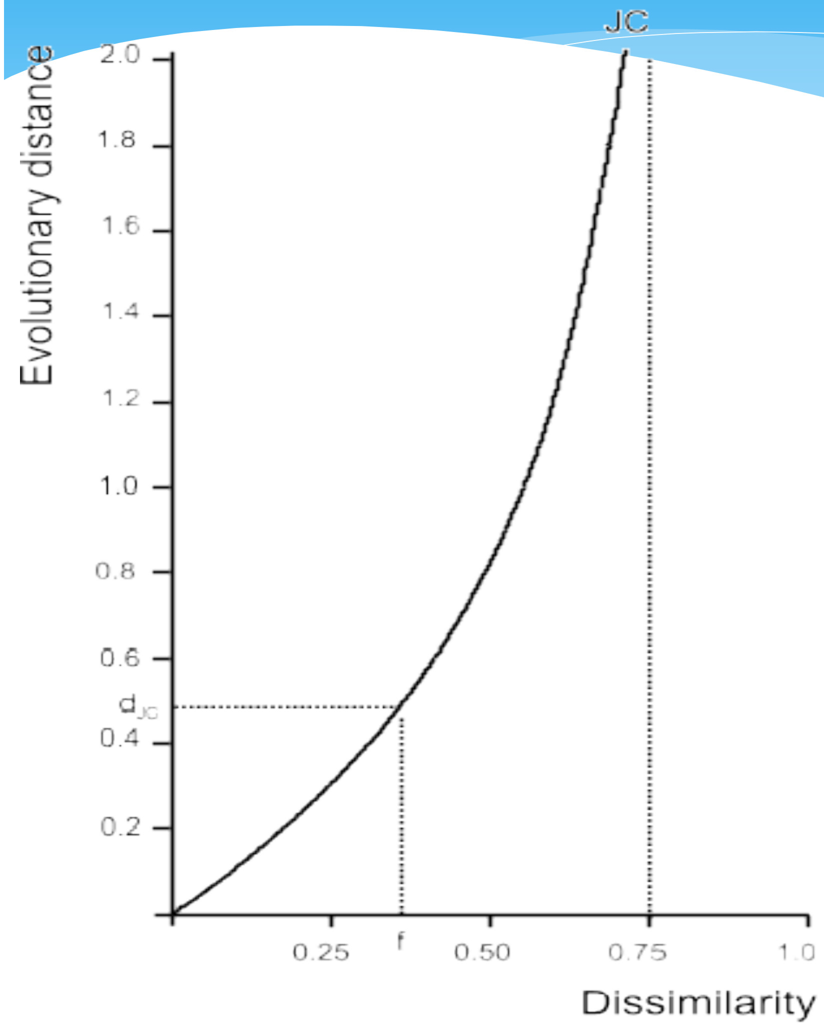
- observed differences <> actual genetic distances between the sequences.

=> dissimilarity is an underestimation of the true evolutionary distance, because of the fact that some of the sequence positions are the result of multiple events

Solution:

- Use an evolutionary model that corrects for multiple mutations

Distance Calculation



A model of evolution is based on certain presumptions!

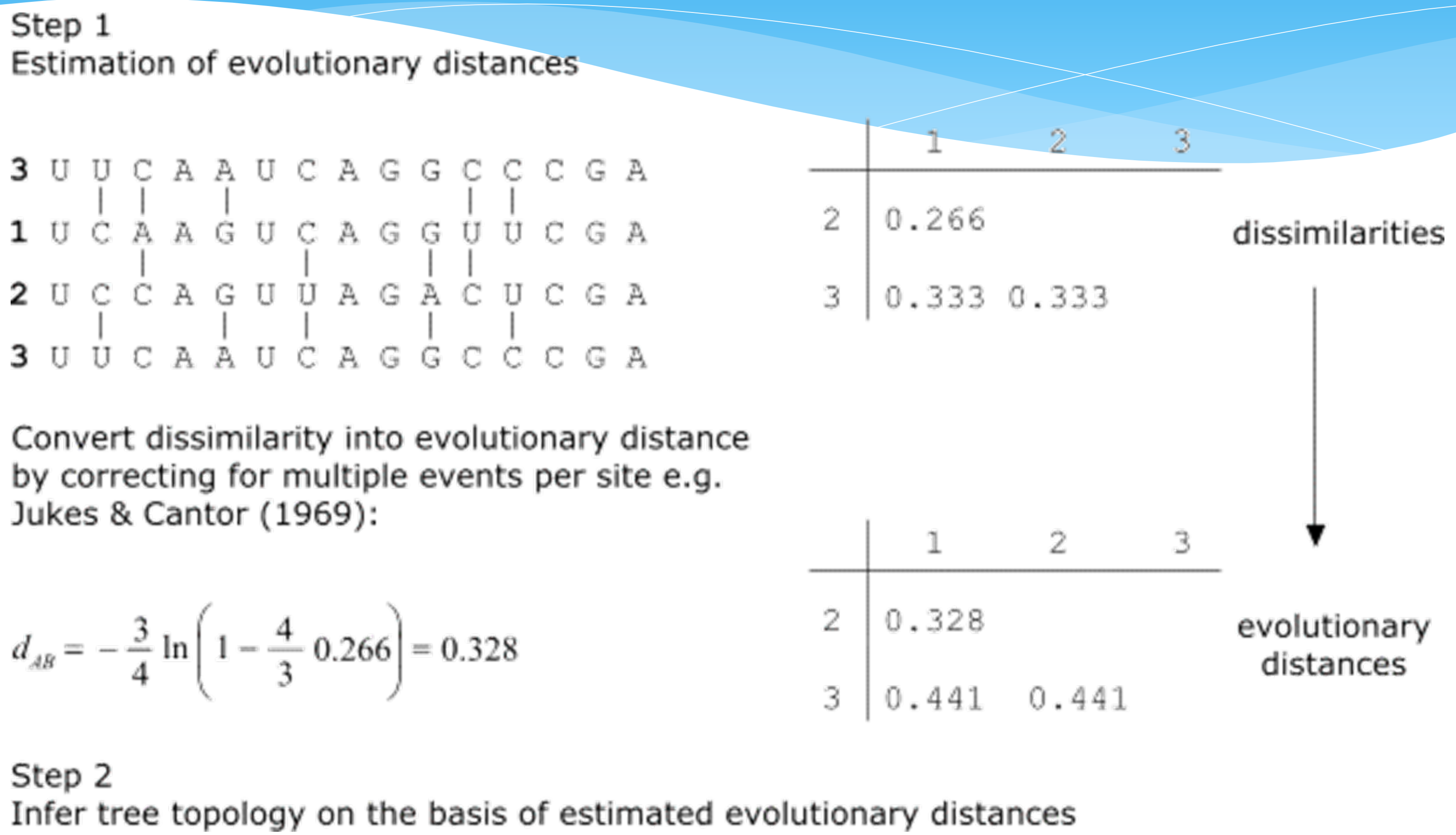
E.g. substitution model of Jukes & Cantor (1969)

- all substitutions are independent
- all sequence positions are equally likely to change
- substitutions occur randomly among the four types of nucleotides: there is no bias in the direction of change
- no insertions or deletions have occurred

Based on a model of evolution, we can derive an equation that expresses the relationship between dissimilarity (fraction of observed differences) and evolutionary distance (fraction of expected differences)

$$D_{AB} = -\frac{3}{4}\ln\left(1-\frac{4}{3}f_{AB}\right)$$

Distance Calculation



Step 1
Estimation of evolutionary distances

```
3 U U C A A U C A G G C C C G A
    | |   |           | |
1 U C A A G U C A G G U U C G A
      |       |     | |
2 U C C A G U U A G A C U C G A
    |     |   |     |   |
3 U U C A A U C A G G C C C G A
```

	1	2	3
2	0.266		
3	0.333	0.333	

dissimilarities

Convert dissimilarity into evolutionary distance by correcting for multiple events per site e.g. Jukes & Cantor (1969):

$$d_{AB} = -\frac{3}{4}\ln\left(1-\frac{4}{3}0.266\right) = 0.328$$

	1	2	3
2	0.328		
3	0.441	0.441	

evolutionary distances

Step 2
Infer tree topology on the basis of estimated evolutionary distances

UPGMA

UPGMA (Unweight Pair Group Method using Arithmetic averages)

- One of the most popular phylogenetic tree algorithms.
- Produce a rooted tree (unlike MP method).
- UPGMA produces **ultrametic** trees. The distance from any internal node (including the root) to its descendant leaves is identical!
- Assume a constant rate of evolution rate (molecular clock hypothesis).

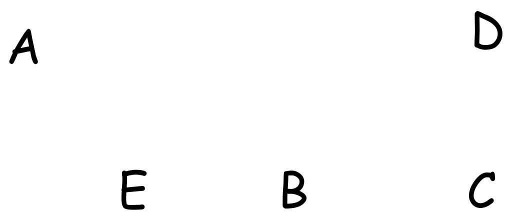
UPGMA

- Create a distance matrix between all pairs of taxa
- Iteratively do following until all taxa are merged
  - Merge the pair (x, y) with smallest distance d(x, y) and form xy
  - Set distance
    - $d(z, xy) = \frac{d(z,x)+d(z,y)}{2}$ for all z
  - For clusters,
    - $d(C_i, C_j) = \frac{1}{|C_i||C_j|} \Sigma_{p\ in\ C_i, q\ in\ C_j} d(p, q)$

      ($|C_i|$ is the number of taxa in C_i)
    - $d(z, C_iC_j) = \frac{d(z,C_i)|C_i|+d(z,C_j)|C_j|}{|C_i|+|C_j|}$

UPGMA

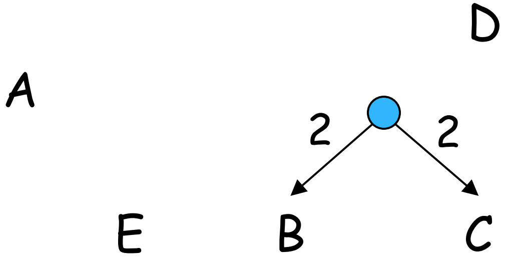
Choose two clusters with minimum distance and combine them



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

UPGMA

$$d(A,BC) = \frac{d(A,B)+d(B,C)}{2} = \frac{10+12}{2} = 11$$



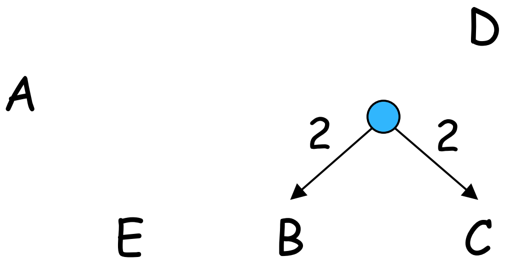
Distance of new cluster to nodes in original clusters is half of original distance

	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

	A	BC	D	E
A	0	**11**	8	7
BC		0	5	15
D			0	12
E				0

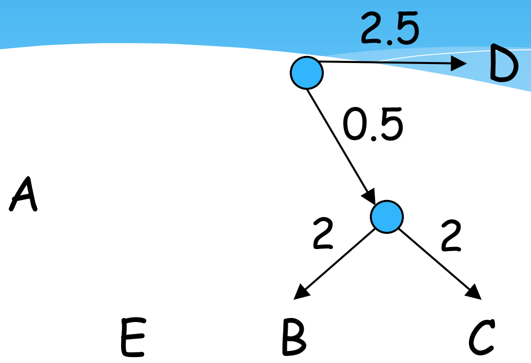
Update distance matrix

UPGMA



	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

UPGMA



	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

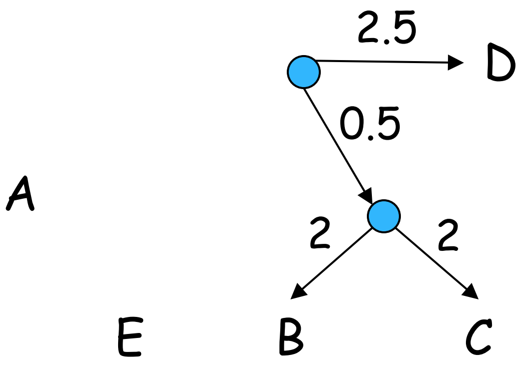
$$d(A,BCD) = \frac{2*d(A,BC)+d(A,D)}{2+1}$$

$$= \frac{2*11+8}{3} = 10$$

$$d(E,BCD) = \frac{2*15+12}{2+1} = 14$$

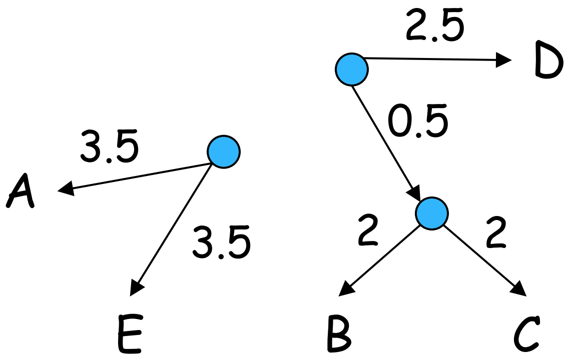
	A	BCD	E
A	0	10	7
BCD		0	14
E			0

UPGMA



	A	BCD	E
A	0	10	7
BCD		0	14
E			0

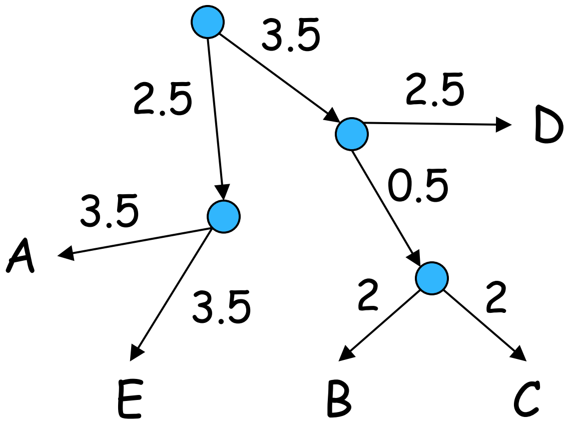
UPGMA



	A	BCD	E
A	0	10	7
BCD		0	14
E			0

	AE	BCD
AE	0	12
BCD		0

UPGMA

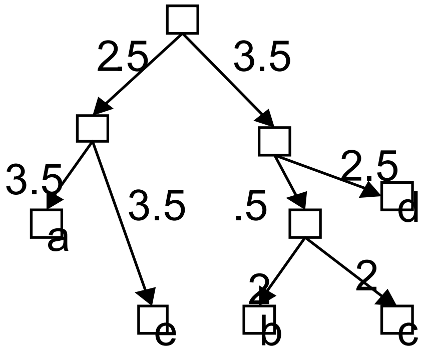


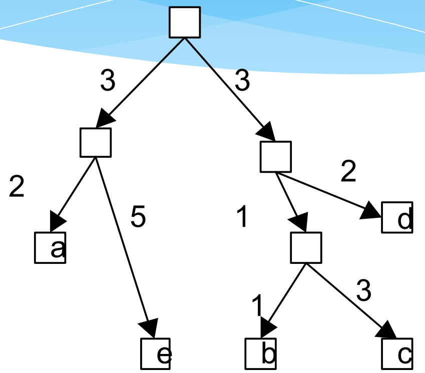
	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

produced tree

(((B, C), D), (A, E))

UPGMA Result





	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

	A	B	C	D	E
A	0	10	12	10	7
B		0	4	4	13
C			0	6	15
D				0	13
E					0

UPGMA

UIUC TeachEnG UPGMA game

http://teacheng.illinois.edu/PhylogeneticTree