Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 21

1

TODAY'S TOPICS

➢ **Distance estimation**

- Synonymous & nonsynonymous distances

➢ **Phylogenetic reconstruction**

- Introduction (terminologies, rooting, *etc.*)
- Distance methods (UPGMA, FM, ME)

2

Universal Genetic Code

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Stop	TGA	Stop
	TTG		TCG		TAG		TGG	Trp
C	CTT		CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

- Synonymous (silent) substitutions DOES NOT change amino acids
- Nonsynonymous (replacement) substitutions DOES change amino acids

3

Synonymous/nonsynonymous distance methods

➢ **Nei-Gojobori method** (Nei and Gojobori, 1986)

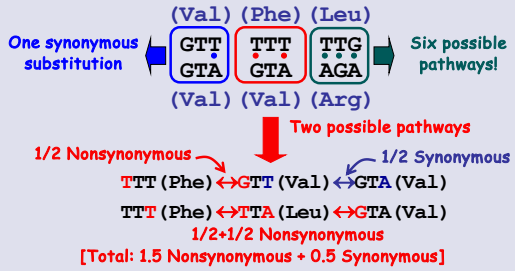
- Number of synonymous differences: S_d
- Number of nonsynonymous differences: N_d
- Proportion of synonymous differences: p_S
- Proportion of nonsynonymous differences: p_N
  - → $p_S = S_d/S$, $p_N = N_d/N$
    - S: Number of synonymous sites
    - N: Number of nonsynonymous sites
- Jukes-Cantor correction for multiple-hits
  - → $d_S = -3/4\ln(1-4p_S/3)$
  - → $d_N = -3/4\ln(1-4p_N/3)$

K2P or Tajima-Nei (1-parameter+base freq.) correction is also used in modified versions

4

Synonymous/nonsynonymous distance methods

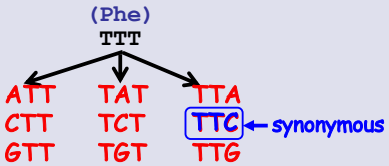
➢ How to count synonymous/nonsynonymous differences



5

Synonymous/nonsynonymous distance methods

➢ How to count synonymous/nonsynonymous sites



Synonymous sites (S):	0	+	0	+	1/3 = 1/3
Nonsynonymous sites (N):	3/3	+	3/3	+	2/3 = 8/3

- Count the number of sites from each codon and sum up for each sequence. Take the average from two sequences.

6

Synonymous/nonsynonymous distance methods

- **Nei-Gojobori method** (Nei and Gojobori, 1986)
  - Number of synonymous differences: S_d
  - Number of nonsynonymous differences: N_d
  - Proportion of synonymous differences: p_S
  - Proportion of nonsynonymous differences: p_N
    - → $p_S = S_d/S$, $p_N = N_d/N$
      - S: Number of synonymous sites
      - N: Number of nonsynonymous sites
  - Jukes-Cantor correction for multiple-hits
    - → $d_S = -3/4\ln(1-4p_S/3)$
    - → $d_N = -3/4\ln(1-4p_N/3)$

K2P or Tajima-Nei (1-parameter+base freq.) correction is also used in modified versions

7

Nucleotide substitution patterns

Table 2. Comparisons among the methods for estimating synonymous and nonsynonymous substitution numbers per site

	Gene: length (species compared with *D. melanogaster*)					
	~~GC:60%~~ *Adhr*: 816 bp (*D. teissieri*)		~~GC:80%~~ *Adh* 762 bp (*D. ps. bogotana*)		~~GC:4%~~ *Col*: 1497 bp (*D. yakuba*)	
Method	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous
NG	0.402 ± 0.060	0.009 ± 0.004	0.604 ± 0.080	0.053 ± 0.010	0.380 ± 0.041	0.007 ± 0.003
LWL	0.394 ± 0.058	0.009 ± 0.004	0.599 ± 0.080	0.054 ± 0.010	0.364 ± 0.040	0.007 ± 0.002
PBL ←	0.328 ± 0.052	0.009 ± 0.004	0.561 ± 0.078	0.054 ± 0.010	0.401 ± 0.051	0.007 ± 0.003

Moriyama and Powell (1997) J Mol Evol 45:378-391

NG: Nei-Gojobori method (Nei & Gojobori 1986): based on JC model
LWL: Li-Wu-Luo method (Li et al. 1985): based on K2P model
PBL or Li93: Pamilo-Bianchi-Li method (Pamilo and Bianchi 1993; Li 1993)
Kumar method (available in MEGA; modification to PBL)

NG method underestimates the number of synonymous sites: S
LWL method overestimates the number of synonymous sub.: S_d
PBL method corrected problems found in both NG and LWL methods

8

Available distance method programs

- MEGA X http://www.megasoftware.net/
  - → Includes synonymous & nonsynonymous distances
- PAML http://abacus.gene.ucl.ac.uk/software/paml.html
  - → Includes Yang and Nielsen (2000) method [yn00]
- SNAP https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html
  - → Synonymous & nonsynonymous (Nei-Gojobori) distance only
- Ape (R package for Analysis of Phylogenetics and Evolution)
  - → Includes many distance methods https://emmanuelparadis.github.io/index.html https://cran.r-project.org/web/packages/ape/index.html
- Phylip3.698 http://evolution.genetics.washington.edu/phylip.html
  - → JC, K2P, F84 (HKY85), LogDet, gamma distances
  - → Dayhoff's PAM, JTT, PMB (Probability Matrix from Blocks), Kimura's PAM approximation, gamma distances
  - → On the Web: http://phylemon.bioinfo.cipf.es (ver. 3.68)
  - → In EMBOSS: http://emboss.toulouse.inra.fr/cgi-bin/emboss/ (found in Phylogeny sections)
  - → See "How to use Phylip" on Canvas

⚠ ClustalW2 (ClustalX2) → ~~K2P for DNA, hybrid between Kimura and PAM for protein!~~

$p \le 0.75$	Use Kimura's PAM distance approximation method
$0.75 < p \le 0.93$	Use a conversion table with 0.01 interval (.75, .751, ...)
~~$0.93 < p$~~	~~$k = 10.0$ [arbitral constant]~~

9

Introduction to phylogeny

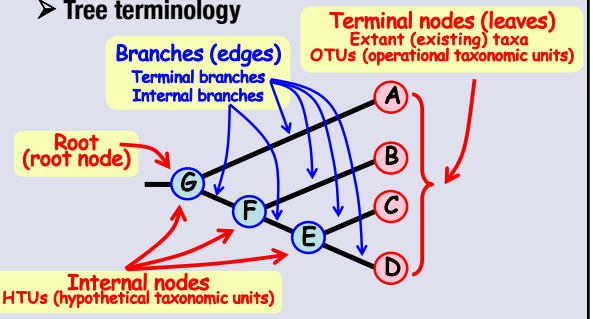
- **Phylogeny (phylogenetic tree)**
  - → a graphic representation of evolutionary relationships among genes or organisms
  - True phylogeny cannot be known
    - We cannot actually observe the long-term evolution!
  - Phylogenetic relationships can be only inferred
  - Phylogenetic relationships are reconstructed based on the information available (*e.g.*, sequences)
  - → represents a hypothesis of evolutionary relationships among gene or protein sequences: gene tree
  - → Organismal relationships are inferred based on phylogenetic analysis: species tree

Note: Gene trees do not always represent species trees!

10

Introduction to phylogeny

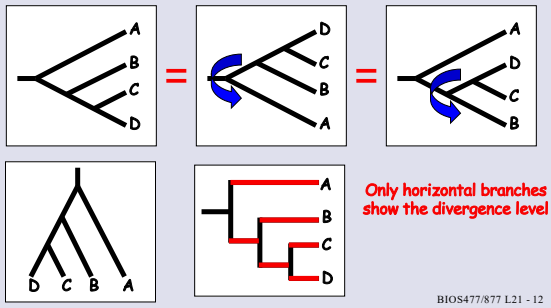
- Tree terminology



11

Introduction to phylogeny

- Many ways of drawing trees

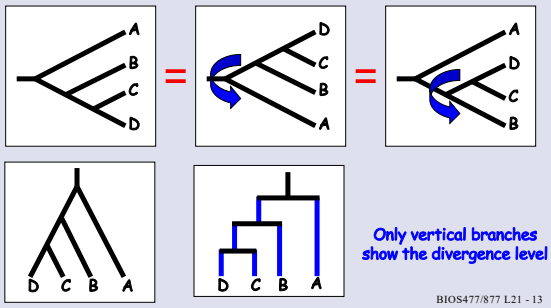


Only horizontal branches show the divergence level

12

Introduction to phylogeny

➢ Many ways of drawing trees

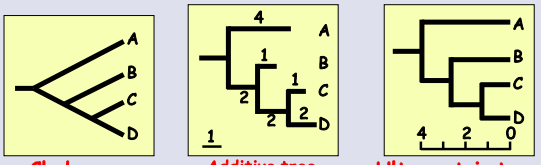


Only vertical branches show the divergence level

13

Introduction to phylogeny

➢ Three different types of trees



Cladogram
Relative recency of common ancestry (or branching order)
No quantitative information

Additive tree (phenogram)
Branch lengths show the amount of evolutionary changes

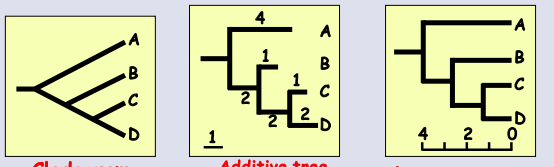
Ultrametric tree
shows evolutionary time

In ultrametric trees, end nodes are all equidistant from the root of the tree
➜ possible only assuming molecular clock (constant evolutionary rate)

14

Introduction to phylogeny

➢ Three different types of trees



Cladogram

Additive tree (phenogram)

Ultrametric tree

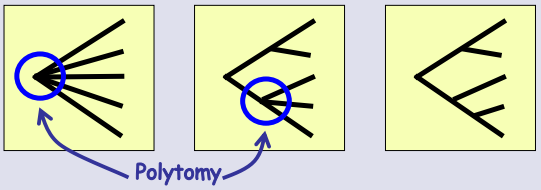
Branch length has no information

Branch length shows the amount of divergence

15

Introduction to phylogeny

➢ Resolution of trees



Star tree
No resolution

Partially resolved

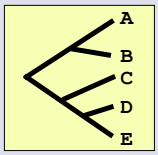
Fully resolved
(bifurcating tree)

16

Introduction to phylogeny

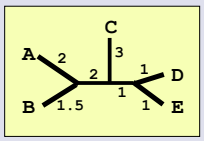
➢ Nested parentheses format: Newick format

Rooted



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

Unrooted



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

((A:2, B:1.5):2,C:3, (D:1, E:1):1);

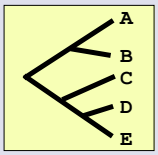
Branch lengths

17

Introduction to phylogeny

➢ Nested parentheses format: Newick format

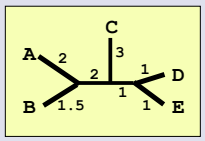
Rooted



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

2 clusters divided by the root

Unrooted

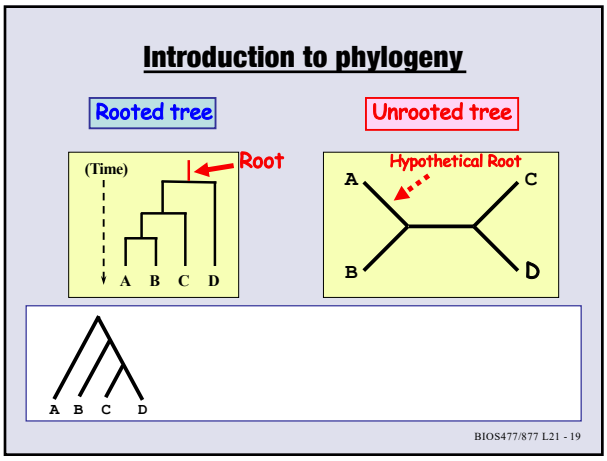


((A,B), C, (D, E));

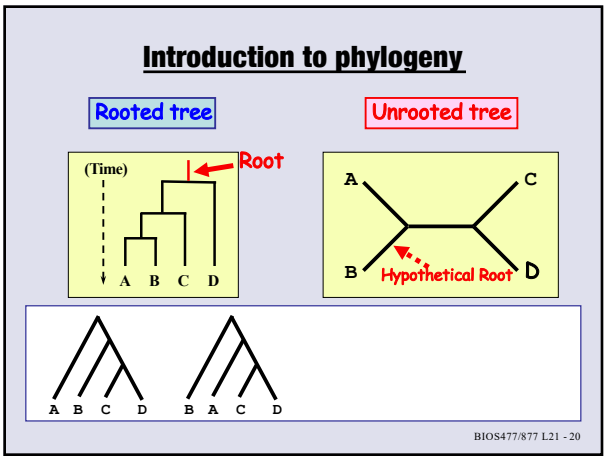
((A:2, B:1.5):2,C:3, (D:1, E:1):1);

3 clusters and no root

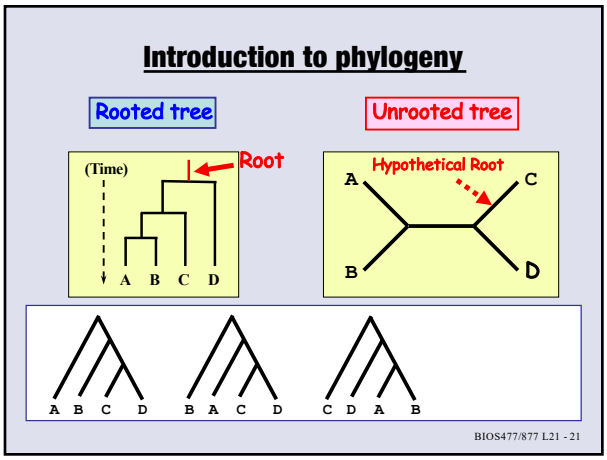
18



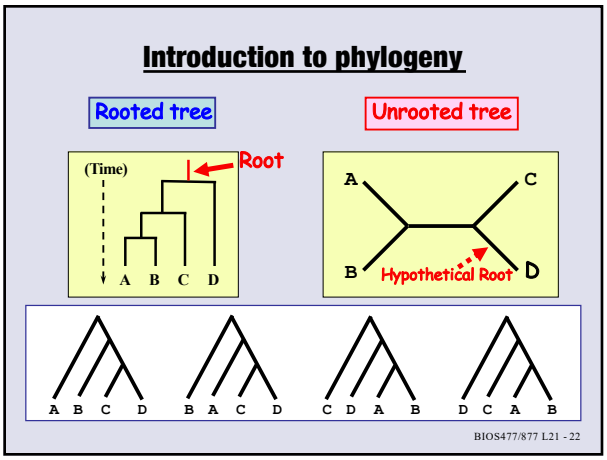
19



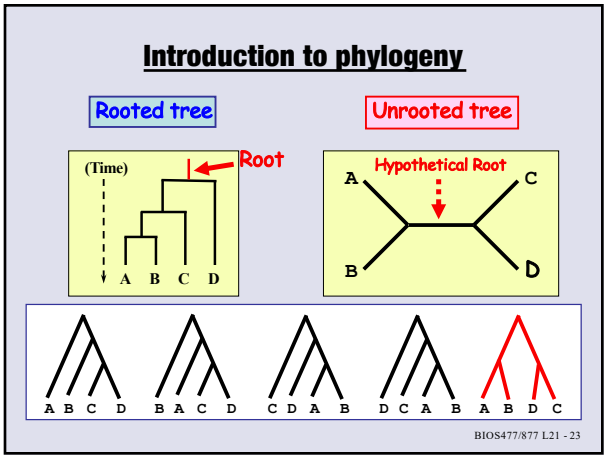
20



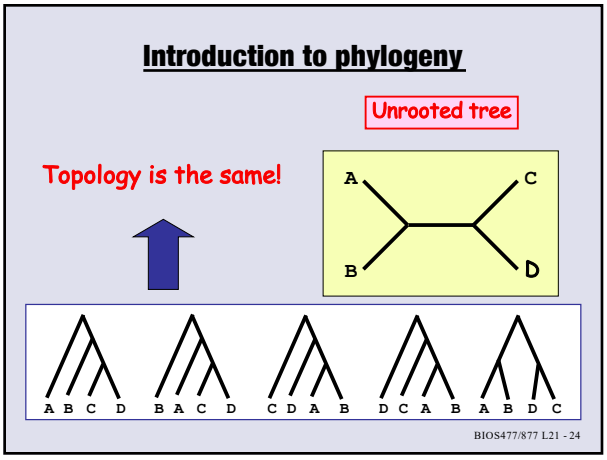
21



22



23



24

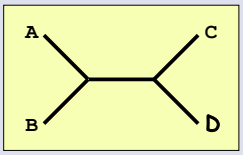
Introduction to phylogeny

➢ **Outgroup**

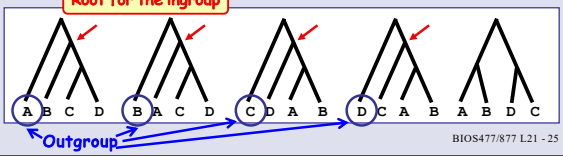
➜ used to "root" unrooted tree

• Biological information is required to choose appropriate outgroup

Unrooted tree



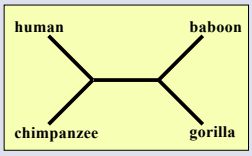
Root for the ingroup



25

Introduction to phylogeny

➢ **Outgroup**



Which species can be used as the outgroup?

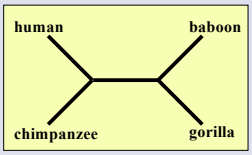
26

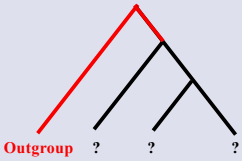
Introduction to phylogeny

➢ **Outgroup**

➜ used to "root" unrooted tree

• Biological information is required to choose appropriate outgroup





Most appropriate outgroup?

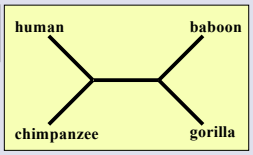
27

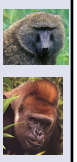
Introduction to phylogeny

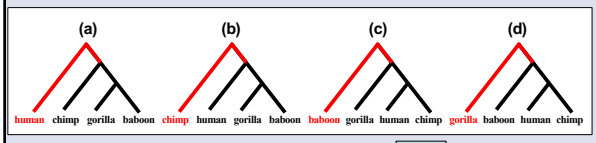
➢ **Outgroup**

➜ used to "root" unrooted tree

• Biological information is required to choose appropriate outgroup







Which tree has the most appropriate outgroup?

28

Phylogenetic methods

➢ Data types and tree-building methods

[Data types]

[Tree-building methods]	Distances	Characters
Clustering	UPGMA Neighbor joining	
Optimality criterion	Minimum evolution Fitch-Margoliash	Maximum parsimony Maximum likelihood

(Bayesian inference)

Examine all possible topologies based on a certain criterion

Yang and Rannala (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13: 303-314.

30

Phylogenetic methods (Distance)

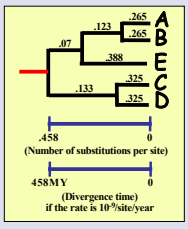
➢ **UPGMA**: unweighted pair-group method with arithmetic mean

• reconstructs ultrametric trees

➜ all terminal nodes are equidistant from the root of the tree

➜ equivalent to assuming a molecular clock

➜ assumes all sequences evolve at the same rate



• reconstructs a rooted tree

• extremely sensitive to unequal rates in different lineages ➜ could result in a wrong topology

31

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

Example: a distance matrix for 5 sequences

The smallest distance pair is grouped together

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

32

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

ABCDE → (AB) CDE

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

	C	D	E
A/B			
C			
D			

Distance matrix is recalculated with A and B as one group

33

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

ABCDE → (AB) CDE

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

	C	D	E
A/B	.895	.975	.775
C		.65	.81
D			.94

Distance matrix is recalculated with A and B as one group

$d(A/B,C) = [d(A,C)+d(B,C)]/2$
$d(A/B,D) = [d(A,D)+d(B,D)]/2$
$d(A/B,E) = [d(A,E)+d(B,E)]/2$

34

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

ABCDE → (AB) CDE

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

	C	D	E
A/B	.895	.975	.775
C		.65	.81
D			.94

The smallest distance pair

$d(A/B,C) = [d(A,C)+d(B,C)]/2$
$d(A/B,D) = [d(A,D)+d(B,D)]/2$
$d(A/B,E) = [d(A,E)+d(B,E)]/2$

35

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

ABCDE → (AB) CDE → (AB) (CD) E

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

	C	D	E
A/B	.895	.975	.775
C		.65	.81
D			.94

	C/D	E
A/B	.935	.775
C/D	.875	

If i/j is the new cluster:

$$d(i/j,k) = \frac{n_i}{n_i+n_j} d(i,k) + \frac{n_j}{n_i+n_j} d(j,k)$$

(n_i and n_j are the numbers of taxa in the groups)

$d(C/D, A/B) = ½d(C,A/B) + ½d(D,A/B)$
$d(C/D,E) = ½d(C,E) + ½d(D,E)$

Or from the original matrix, calculated also as
$d(C/D, A/B) = [d(C,A)+d(C,B)+d(D,A)+d(D,B)]/4$

$$d(X,Y) = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d(i,j)$$

N_X, N_Y: the number of taxa in group X and Y
i, j: all sequences in group X and Y

2-ways to calculate

36

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

ABCDE → (AB) CDE → (AB) (CD) E

	B	C	D	E
A	.53	.99	1.02	.82
B		.80	.93	.73
C			.65	.81
D				.94

	C	D	E
A/B	.895	.975	.775
C		.65	.81
D			.94

	C/D	E
A/B	.935	.775
C/D	.875	

The smallest distance pair

Or from the original matrix, calculated also as
$d(C/D, A/B) = [d(C,A)+d(C,B)+d(D,A)+d(D,B)]/4$

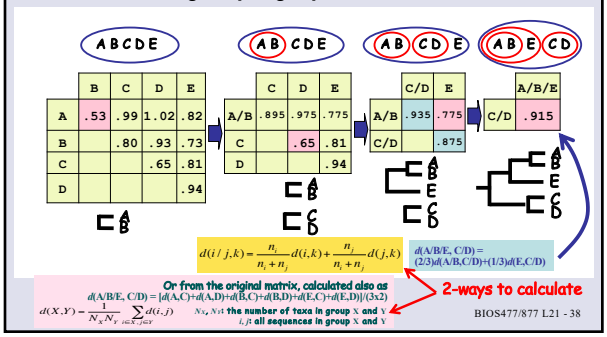
$d(C/D, A/B) = ½d(C,A/B) + ½d(D,A/B)$
$d(C/D,E) = ½d(C,E) + ½d(D,E)$

2-ways to calculate

37

Phylogenetic methods (Distance)

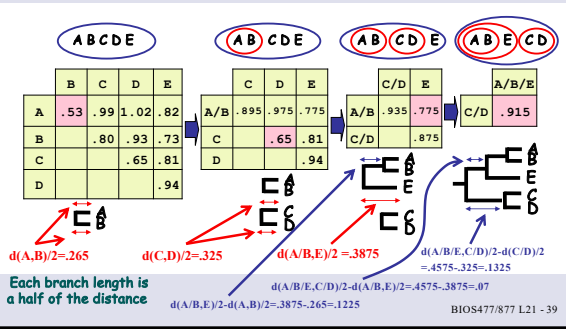
➢ **UPGMA**: unweighted pair-group method with arithmetic mean



38

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

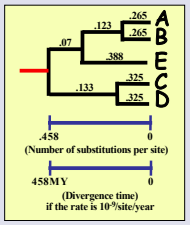


39

Phylogenetic methods (Distance)

➢ **UPGMA**: unweighted pair-group method with arithmetic mean

- reconstructs ultrametric trees
  - → all terminal nodes are equidistant from the root of the tree
  - → equivalent to assuming a molecular clock
  - → assumes all sequences evolve at the same rate
- reconstructs a rooted tree
- extremely sensitive to unequal rates in different lineages → could result in a wrong topology



40

Phylogenetic methods

➢ Data types and tree-building methods

[Data types]

[Tree-building methods]	Distances	Characters
Clustering	UPGMA Neighbor joining	
Optimality criterion	Minimum evolution Fitch-Margoliash	Maximum parsimony Maximum likelihood

(Bayesian inference)

Examine all possible topologies based on a certain criterion

Yang and Rannala (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13: 303-314.

41

Phylogenetic methods (Distance)

➢ Fitch-Margoliash method (weighted least-square)

(Fitch and Margoliash, 1967)

- Initial tree: constructed by clustering 2 OTUs with shortest distances
  - → similar to UPGMA
- No constant rate assumption (additive trees)
- Reconstructs unrooted trees
- Alternative trees are tested to identify the best tree based on the smallest percent standard deviation (PSD):

$$PSD = \sqrt{\frac{2\sum_{ij}\{(D_{ij} - E_{ij})/D_{ij}\}^2}{n(n-1)}} \times 100$$

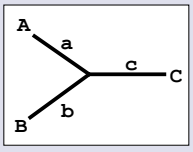
n: number of taxa in the tree
D_{ij}: observed distances between i and j
E_{ij}: estimated distances between i and j (calculated by branch lengths)

42

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: 3 taxa

(Fitch and Margoliash, 1967)



	B	C
A	d_{AB}	d_{AC}
B		d_{BC}

d_{AB}, d_{AC}, d_{BC}: distances between sequences A, B, and C
a, b, c: branch lengths

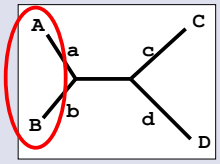
$$\begin{cases} d_{AB} = a + b \\ d_{AC} = a + c \\ d_{BC} = b + c \end{cases} \Rightarrow \begin{array}{l} a = (d_{AB} + d_{AC} - d_{BC})/2 \\ b = (d_{AB} + d_{BC} - d_{AC})/2 \\ c = (d_{AC} + d_{BC} - d_{AB})/2 \end{array}$$

43

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



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

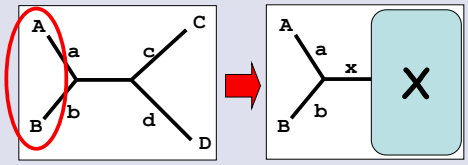
- Choose two taxa (*e.g.,* with the smallest distance) ➜ A and B

44

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



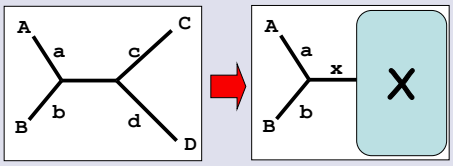
- Choose two taxa (*e.g.,* with the smallest distance)
- Remaining taxa are combined into a single composite taxon ➜ X

45

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



Recalculate the distance matrix
(as shown in UPGMA)

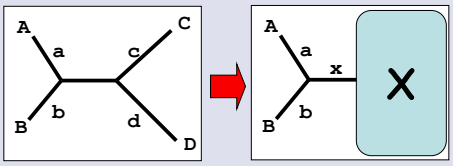
	B	X
A	d_{AB}	d_{AX}
B		d_{BX}

46

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



$$d_{AX} = (d_{AC} + d_{AD}) / 2$$
$$d_{BX} = (d_{BC} + d_{BD}) / 2$$

	B	X
A	d_{AB}	d_{AX}
B		d_{BX}

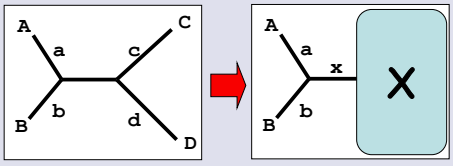
➜ a and b can be calculated as before

47

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



$$\begin{cases} d_{AB} = a + b \\ d_{AX} = a + x \\ d_{BX} = b + x \end{cases} \Rightarrow \begin{aligned} a &= (d_{AB} + d_{AX} - d_{BX}) / 2 \\ b &= (d_{AB} + d_{BX} - d_{AX}) / 2 \end{aligned}$$

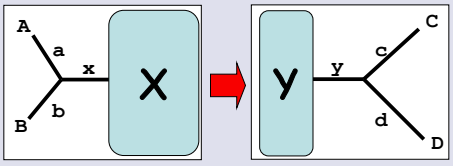
	B	X
A	d_{AB}	d_{AX}
B		d_{BX}

48

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



- The two taxa A and B are combined into a single composite taxon ➜ Y
- Recalculate the distances between Y and other taxa
  (as in UPGMA method)

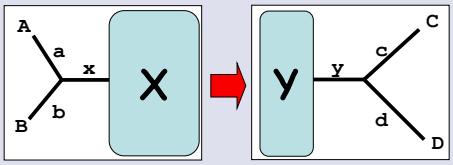
	C	D
Y	d_{YC}	d_{YD}
C		d_{CD}

49

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



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

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

	C	D
Y	d_{YC}	d_{YD}
C		d_{CD}

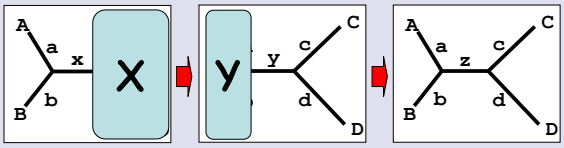
➜ *c* and *d* can be calculated as before

50

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



- If no other unresolved taxon remains,
  - ➜ *z* (internal branch length) can be calculated:
    e.g., $z = d_{AC} - a - c$
- If there are still more taxa,
  - ➜ choose two (*e.g.*, closest) taxa and repeat the above procedure

51

Phylogenetic methods (Distance)

➢ Fitch-Margoliash method (weighted least-square)

(Fitch and Margoliash, 1967)

- Initial tree: constructed by clustering 2 OTUs with shortest distances
  - ➜ similar to UPGMA
- No constant rate assumption (additive trees)
- Reconstructs unrooted trees
- Alternative trees are tested to identify the best tree based on the smallest percent standard deviation (PSD):

$$PSD = \sqrt{\frac{2\sum_{ij}\{(D_{ij} - E_{ij}) / D_{ij}\}^2}{n(n-1)}} \times 100$$

n: number of taxa in the tree
D_{ij}: observed distances between *i* and *j*
E_{ij}: estimated distances between *i* and *j* (calculated by branch lengths)

52