Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 22

1

TODAY'S TOPICS

- Phylogenetic reconstruction
  - Distance methods (FM, ME, NJ)
  - Character-based methods (maximum parsimony)
- Assignment 10

2

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.

3

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):

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

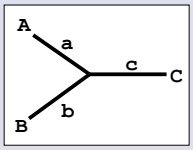
n: number of taxa in the tree
D_{ij}: observed distances between i and j
E_{ij}: estimated distances between i and j (calculated by branch lengths)

4

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

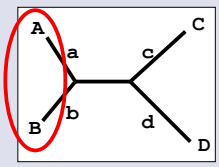
$$\begin{cases} d_{AB} = a + b \\ d_{AC} = a + c \\ d_{BC} = b + c \end{cases} \Rightarrow \begin{aligned} a &= (d_{AB} + d_{AC} - d_{BC}) / 2 \\ b &= (d_{AB} + d_{BC} - d_{AC}) / 2 \\ c &= (d_{AC} + d_{BC} - d_{AB}) / 2 \end{aligned}$$

5

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}

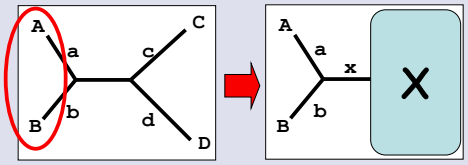
- Choose two taxa (*e.g.,* with the smallest distance)
  → A and B

6

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



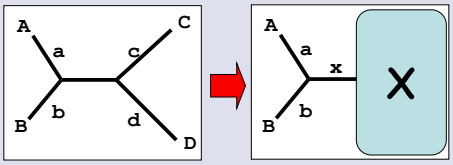
- Choose two taxa (*e.g.*, with the smallest distance)
- Remaining taxa are combined into a single composite taxon ➜ X

7

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



Recalculate the distance matrix
(as shown in UPGMA)

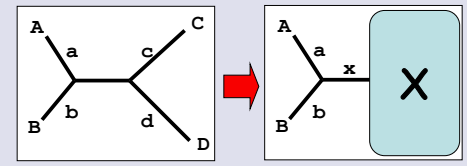
	B	X
A	d_{AB}	d_{AX}
B		d_{BX}

8

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}

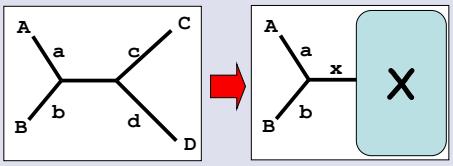
➜ a and b can be calculated as before

9

Phylogenetic methods (Distance)

➢ Estimation of branch lengths: more than 3 taxa

(Fitch and Margoliash, 1967)



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

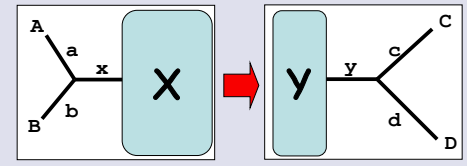
	B	X
A	d_{AB}	d_{AX}
B		d_{BX}

10

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)

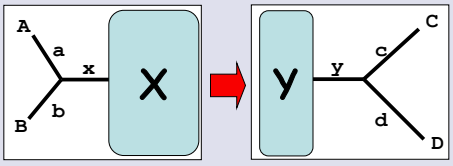
	C	D
Y	d_{YC}	d_{YD}
C		d_{CD}

11

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}

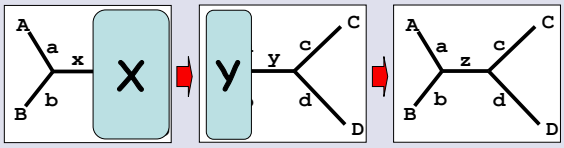
➜ c and d can be calculated as before

12

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

13

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)

14

Phylogenetic methods

➢ Data types and tree-building methods

[Data types]

[Tree-building methods]	Distances	Characters
Clustering	UPGMA Neighbor joining ← Approximates minimum evolution tree	
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.

15

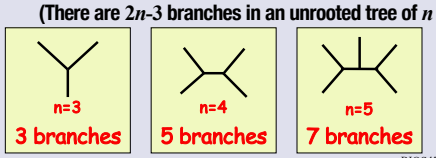
Phylogenetic methods (Distance)

➢ Minimum evolution (ME)

- The tree that minimizes the tree length is regarded as the best estimate of the phylogeny
  - ➜ Tree length = Sum of the branch lengths

$$S = \sum_{i=1}^{2n-3} e_i$$

n: number of taxa in the tree, e_i: length of the branch i
(There are $2n$-3 branches in an unrooted tree of n taxa)



16

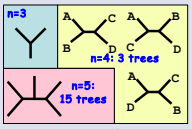
Phylogenetic methods (Distance)

➢ Minimum evolution (ME)

- The tree that minimizes the tree length is regarded as the best estimate of the phylogeny
  - ➜ Tree length = Sum of the branch lengths

$$S = \sum_{i=1}^{2n-3} e_i$$

n: number of taxa in the tree
e_i: length of the branch i
(There are $2n$-3 branches in an unrooted tree of n taxa)



- reconstructs additive distance trees
- reconstructs unrooted trees
- If n is large, impossible to examine all possible topologies

17

Phylogenetic methods

➢ Data types and tree-building methods

[Data types]

[Tree-building methods]	Distances	Characters
Clustering	UPGMA Neighbor joining ← Approximates minimum evolution tree	
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.

18

Phylogenetic methods (Distance)

➢ **Neighbor joining (NJ)**

Saito and Nei (1987) and correction by Studier and Keppler (1988)

- Clustering method (fast computation)
- A good heuristic method for estimating the minimum evolution tree
  - ➜ No guarantee to find the ME tree
  - ➜ In practice, the NJ tree is often the same or very similar to the ME tree
- No assumption for the constant evolutionary rate
  - ➜ Evolutionary rate can vary among lineages
- reconstructs unrooted trees

ME tree: has the topology with the shortest tree length

$$S = \sum_{i=1}^{2n-3} e_i$$

19

Phylogenetic methods (Distance)

➢ **Neighbor joining (NJ)**

Example: a distance matrix for 5 OTUs

	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					

1) Start with a star phylogeny

2) S_0: the sum of all branch lengths

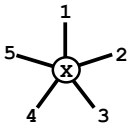
$$S_0 = \sum_{i=1}^{m} L_{iX} = \sum_{i<j}^{m} d_{ij} \Big/ (m-1)$$

L_{iX}: branch length between OTU i and node X

d_{ij}: distance between OTUs i and j

m: number of OTUs

$S_0 = (d_{12}+d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}+d_{34}+d_{35}+d_{45})/4$
$= (.53+.99+.80+1.02+.93+.65+.82+.73+.81+.94)/4$
$= 2.055$



For 3 OTUs: $d_{12}+d_{13}+d_{23}=(a+b)+(a+c)+(b+c)=2(a+b+c)=2S_0$
S_0=(Sum of 3 distances)/(3-1)

20

Phylogenetic methods (Distance)

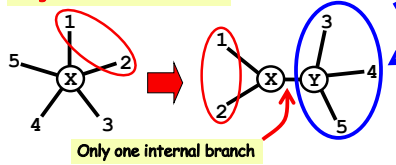
➢ **Neighbor joining (NJ)**

3) Take 2 OTUs (a and b) as a pair (neighbors) and calculate the tree length (S_{ab}) from this topology

4) Find the OTU pair that makes the shortest S_{ab}

Neighbors: OTU1 and OTU2

Put all other OTUs in one cluster



Calculate $S_{12}, S_{13}, S_{14}, \ldots, S_{45}$

Find the shortest S_{ab}

Only one internal branch

21

Phylogenetic methods (Distance)

➢ **Neighbor joining (NJ)**

3') Calculate the sum of all branch length (S_{ab}) when OTUs a and b are neighbors.

If OTUs 1 and 2 are the neighbors:

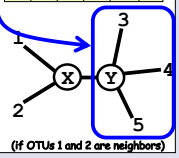
	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					

$$S_{12} = L_{1X} + L_{2X} + L_{XY} + \sum_{i=3}^{m} L_{iY}$$

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

$$\sum_{i=3}^{m} L_{iY} = \sum_{3 \le i<j}^{m} d_{ij} \Big/ (m-3)$$

$$L_{XY} = \left[\sum_{i=3}^{m} (d_{1i} + d_{2i}) - (m-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{m} L_{iY} \right] \Big/ 2(m-2)$$



(if OTUs 1 and 2 are neighbors)

22

Phylogenetic methods (Distance)

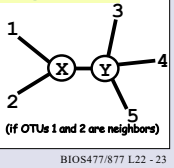
➢ **Neighbor joining (NJ)**

$$L_{XY} = \left[\sum_{i=3}^{m} (d_{1i} + d_{2i}) - (m-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{m} L_{iY} \right] \Big/ 2(m-2)$$

Sum of all distances that include L_{XY}

Sum of all irrelevant branch lengths

$\sum_{i=3}^{5}(d_{1i} + d_{2i})$
$= d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}$
$= L_{1X}+ L_{XY}+L_{Y3}+L_{1X}+ L_{XY}+L_{Y4}+L_{1X}+ L_{XY}+L_{Y5}$
$+L_{2X}+ L_{XY}+L_{Y3}+L_{2X}+ L_{XY}+L_{Y4}+L_{2X}+ L_{XY}+L_{Y4}$
$=3(L_{1X}+ L_{2X}) + 6L_{XY} + 2(L_{Y3}+L_{Y4}+L_{Y5})$
Thus, $L_{XY} = |(d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25})$
$-3(L_{1X}+ L_{2X})-2(L_{Y3}+L_{Y4}+L_{Y5})|/6$



(if OTUs 1 and 2 are neighbors)

23

Phylogenetic methods (Distance)

➢ **Neighbor joining (NJ)**

3') Calculate the sum of all branch length (S_{ab}) for all OTU pairs

	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					

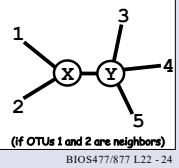
$$S_{12} = L_{1X} + L_{2X} + L_{XY} + \sum_{i=3}^{m} L_{iY}$$

$$= d_{12} + \left[\sum_{i=3}^{m} (d_{1i} + d_{2i}) - (m-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{m} L_{iY} \right] \Big/ 2(m-2)$$

$$+ \sum_{3 \le i<j}^{m} d_{ij} \Big/ (m-3)$$

$$= \sum_{i=3}^{m} (d_{1i} + d_{2i}) \Big/ 2(m-2) + d_{12}/2 + \sum_{3 \le i<j}^{m} d_{ij} \Big/ (m-2)$$

$S_{12} = (d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25})/6+d_{12}/2+(d_{34}+d_{35}+d_{45})/3$
$= (.99+1.02+.82+.80+.93+.73)/6+.53/2+(.65+.81+.94)/3$
$= 1.95 < S_0 = 2.055$



(if OTUs 1 and 2 are neighbors)

24

Phylogenetic methods (Distance)

➢ Neighbor joining (NJ)

3) Sum of all branch lengths (S_{ab}) is calculated for all pairs of OTUs

4) Find the shortest S_{ab}

5) Create a new node (A) that connects OTUs a and b.

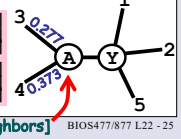
6) Branch lengths are calculated using Fitch-Margoliash method.

$$L_{Aa} = \frac{1}{2(m-2)}\left[(m-2)d_{ab} + \sum_{i=1}^{m} d_{ai} - \sum_{i=1}^{m} d_{bi}\right]$$

$$L_{Ab} = \frac{1}{2(m-2)}\left[(m-2)d_{ab} - \sum_{i=1}^{m} d_{ai} + \sum_{i=1}^{m} d_{bi}\right]$$

	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					

S_{34} is the shortest: (a=3, b=4, m=5)



[S_{34}=1.93 is the shortest ➔ OTUs 3 and 4 are the neighbors]

25

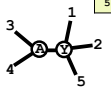
Phylogenetic methods (Distance)

➢ Neighbor joining (NJ)

7) The new distance matrix is calculated using the combined OTU A (for a and b).

Distance between the new OTU A and the remaining OTUs k ($1 \le k \le m$ where $k \ne a$ and $k \ne b$):

$$d_{Ak} = (d_{ak} + d_{bk} - d_{ab}) / 2$$



For a=3, b=4, and m=5, k=1, 2, and 5

$d_{A1} = (d_{31}+d_{41}-d_{34})/2 = (0.99+1.02-0.65)/2=0.68$

$d_{A2} = (d_{32}+d_{42}-d_{34})/2 = (0.80+0.93-0.65)/2=0.54$

$d_{A5} = (d_{35}+d_{45}-d_{34})/2 = (0.81+0.94-0.65)/2=0.55$

	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					

	1	2	A	5
1		.53	.68	.82
2			.54	.73
A				.55
5				

26

Phylogenetic methods (Distance)

➢ Neighbor joining (NJ)

3-7) Repeat the steps:

- For each OTU pair (neighbors), calculate S_{ab},
- Find the smallest S_{ab}, and
- Calculate the next distance matrix.

... (S_{12} is the shortest: OTUs 1 and 2 are the next neighbors)

A new distance matrix for 4 OTUs (m=4)

	1	2	A	5
1		.53	.68	.82
2			.54	.73
A				.55
5				

8) Continue until all OTUs are clustered.

For the last 3 OTUs (see Lecture 21, slide 43):

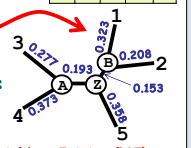
$L_{AZ}=(d_{A5}+d_{AB}-d_{B5})/2=0.193$

$L_{BZ}=(d_{BA}+d_{B5}-d_{A5})/2=0.153$

$L_{5Z}=(d_{5A}+d_{5B}-d_{AB})/2=0.358$

(or $d_{5A}-L_{AZ}$ or $d_{5B}-L_{BZ}$)

	B	A	5
B		.345	.51
A			.55
5			

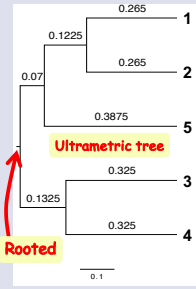


Neighbor Joining (NJ) tree

27

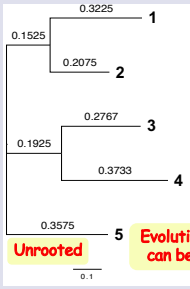
Phylogenetic methods (Distance)

UPGMA



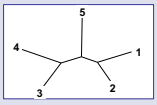
Rooted

NJ



Evolutionary rates can be different

	1	2	3	4	5
1		.53	.99	1.02	.82
2			.80	.93	.73
3				.65	.81
4					.94
5					



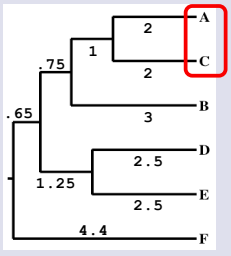
((1:0.3225, 2:0.2075):0.1525, (3:0.2767, 4:0.3733):0.1925, 5:0.3575);

(((1:0.265, 2:0.265):0.1225, 5:0.3875):0.07, (3:0.325, 4:0.325):0.1325);

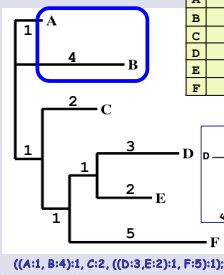
28

Phylogenetic methods (Distance)

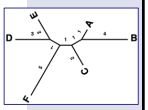
UPGMA



NJ



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



((A:1, B:4):1, C:2, ((D:3,E:2):1, F:5):1);

(((((A:2, C:2):1, B:3):0.75, ((D:2.5,E:2.5):1.25):0.65, F:4.4);

Assuming constant molecular clock

29

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 (Character-based)

➢ **Maximum Parsimony (MP)**

- Chooses the tree(s) that require(s) the fewest evolutionary changes = the shortest tree(s)
- Data: individual sites
- For each site (column), reconstruct the evolution of that site based on a given tree topology and with fewest possible evolutionary changes
- Tree length = Sum of the minimum numbers of character changes over all sites

$$L(\tau) = \sum_{i=1}^{N} l_i$$

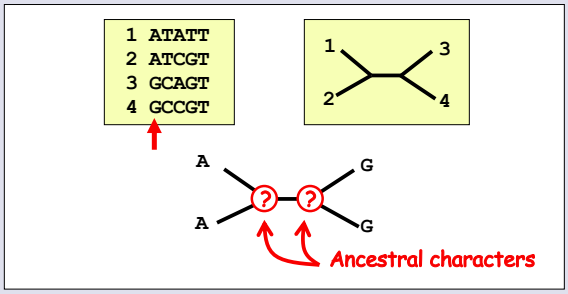
τ: tree topology
N: number of sites (characters)
l_i: tree length for a single site i (amount of character change)

BIOS477/877 L22 - 31

31

Phylogenetic methods (Character-based)

➢ **Maximum Parsimony (MP)**

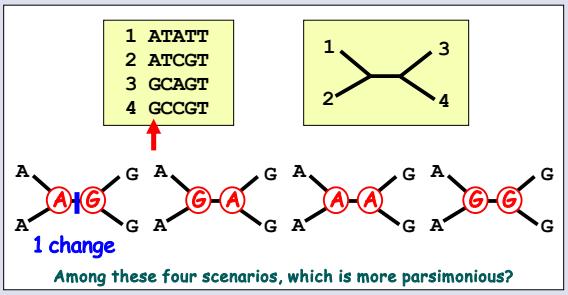


BIOS477/877 L22 - 32

32

Phylogenetic methods (Character-based)

➢ **Maximum Parsimony (MP)**

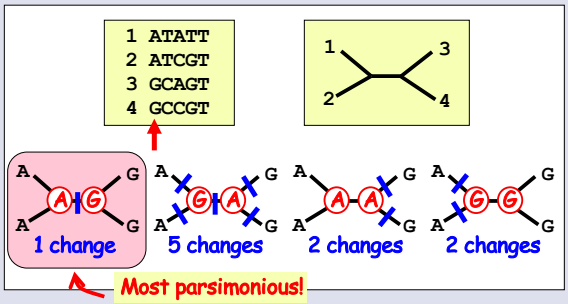


BIOS477/877 L22 - 33

33

Phylogenetic methods (Character-based)

➢ **Maximum Parsimony (MP)**

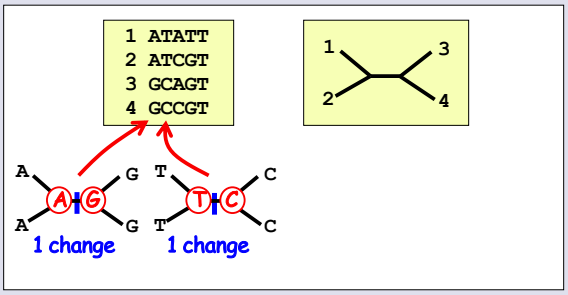


BIOS477/877 L22 - 34

34

Phylogenetic methods (Character-based)

➢ **Maximum Parsimony (MP)**

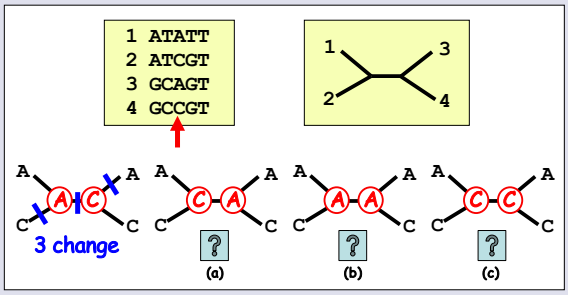


BIOS477/877 L22 - 35

35

Phylogenetic methods (Character-based)

➢ **Maximum Parsimony (MP)**

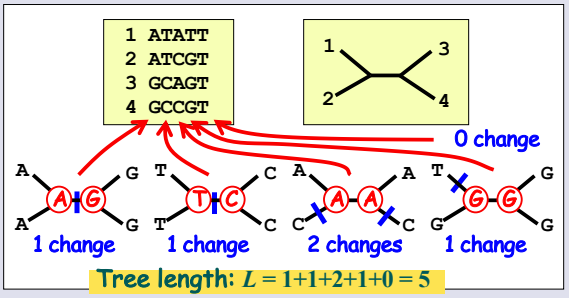


BIOS477/877 L22 - 36

36

Phylogenetic methods (Character-based)

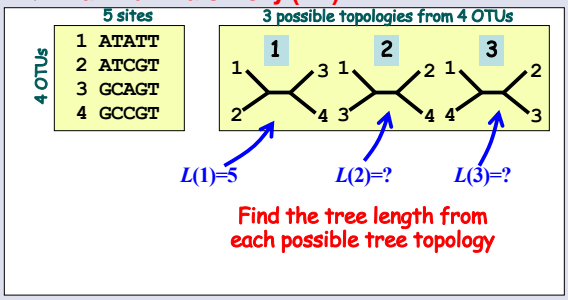
➢ Maximum Parsimony (MP)



40

Phylogenetic methods (Character-based)

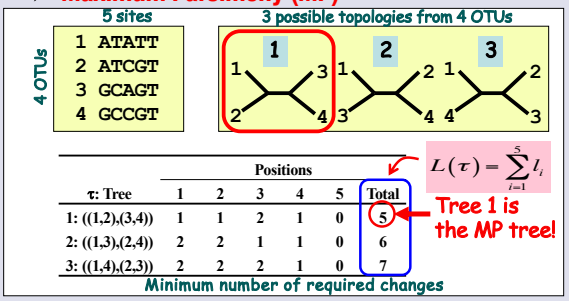
➢ Maximum Parsimony (MP)



41

Phylogenetic methods (Character-based)

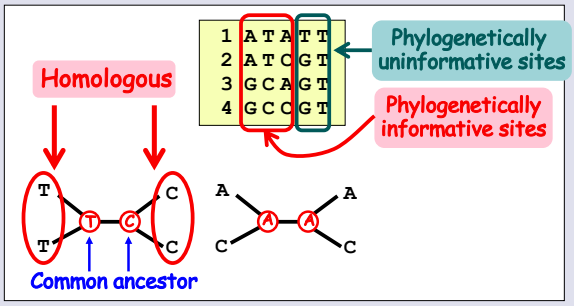
➢ Maximum Parsimony (MP)



42

Phylogenetic methods (Character-based)

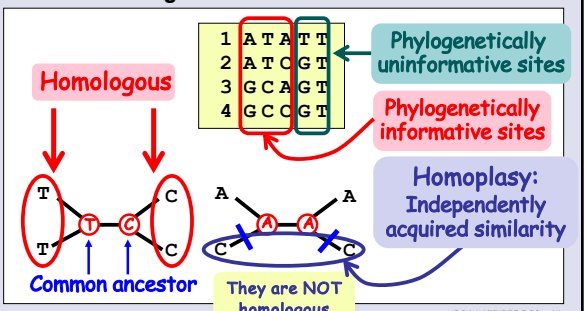
➢ Terminologies for character evolution



43

Phylogenetic methods (Character-based)

➢ Terminologies for character evolution

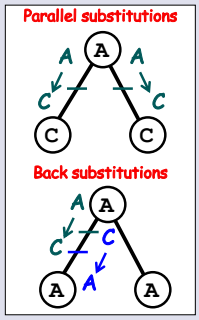


44

Phylogenetic methods (Character-based)

➢ Homoplasy

- Sharing of identical character states that cannot be explained by inheritance from the common ancestor of a group of taxa
- Caused by
  → parallel or back substitutions
- Homoplasy obscures the actual number of evolutionary events
- Fewer homoplasy is better



45

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

Not based on an explicit model of evolution

➜ What do we do if substitution patterns are biased? (e.g., saturation in transitional substitutions)

- Cost matrix (or weight matrix or step matrix)

  ➜ When counting the number of changes, use different weighting depending on the reliability of character change information

  Transversion weighting

  - Ts could be saturated and may not reflect the correct evolutionary relationships (less phylogenetic information) ➜ down-weight Ts
  - Phylogenetic information from Tv is more reliable ➜ up-weight Tv (more information from Tv is used)

46

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

$L(\tau) = \sum_{i=1}^{N} l_i$ τ: tree topology; N: number of sites (characters)

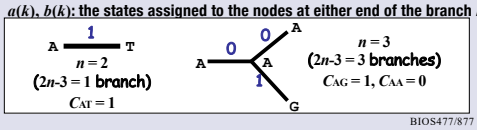
l_i: length for a single site i (amount of character change)

$l_i = \sum_{k=1}^{2n-3} C_{a(k)b(k)}$

n: number of OTUs ($2n$-3: number of branches)

C_{xy}: the cost associated with the change from state x to y

$a(k)$, $b(k)$: the states assigned to the nodes at either end of the branch k



47

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

- Cost matrix (or weight matrix or step matrix)

	A	T	C	G
A	0	1	1	1
T	1	0	1	1
C	1	1	0	1
G	1	1	1	0

A —1— T
$n = 2$
($2n$-3 = 1 branch)
$C_{AT} = 1$

To minimize the changes

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Identity substitution matrix used for alignment

To maximize the identity

48

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

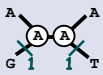
- Simple parsimony

	A	T	C	G
A	0	1	1	1
T	1	0	1	1
C	1	1	0	1
G	1	1	1	0

$C_{xy} = 0$ for $x = y$
$C_{xy} = 1$ for $x \neq y$

C_{xy}: the cost associated with the change from state x to y

$C_{AT} = 1$
$C_{AG} = 1$



$L = 1 + 1 = 2$

49

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

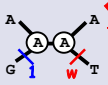
- Generalized parsimony (weighted parsimony)

	A	T	C	G
A	0	w	w	1
T	w	0	1	w
C	w	1	0	w
G	1	w	w	0

$C_{xy} = 0$ for $x = y$
$C_{xy} = 1$ for transition
$C_{xy} = w$ for transversion

C_{xy}: the cost associated with the change from state x to y

$C_{AT} = w$
$C_{AG} = 1$



Transversions are weighted more than transitions

$L = 1 + w$

50

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

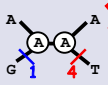
- Generalized parsimony (weighted parsimony)

	A	T	C	G
A	0	4	4	1
T	4	0	1	4
C	4	1	0	4
G	1	4	4	0

$C_{xy} = 0$ for $x = y$
$C_{xy} = 1$ for transition
$C_{xy} = 4$ for transversion

C_{xy}: the cost associated with the change from state x to y

$C_{AT} = 4$
$C_{AG} = 1$



Transversions are weighted more than transitions

$L = 1 + 4 = 5$

51

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP)

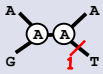
- Transversion parsimony

	A	T	C	G
A	0	1	1	0
T	1	0	0	1
C	1	0	0	1
G	0	1	1	0

$C_{xy} = 0$ for $x = y$
$C_{xy} = 0$ for transition
$C_{xy} = 1$ for transversion

C_{xy}: the cost associated with the change from state x to y

$C_{AT} = w$



Only transversions are counted

$L = 1$

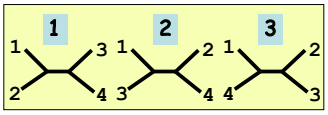
52

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP): simple, not weighted

	A	T	C	G
A	0	1	1	1
T	1	0	1	1
C	1	1	0	1
G	1	1	1	0

1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT



Tree	Sites 1	2	3	4	5	Total
1: ((1,2),(3,4))	1	1	2	1	0	5
2: ((1,3),(2,4))	2	2	1	1	0	6
3: ((1,4),(2,3))	2	2	2	1	0	7

$L(\tau) = \sum_{i=1}^{5} l_i$

MP tree!

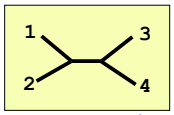
53

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP): weighted

	A	T	C	G
A	0	4	4	1
T	4	0	1	4
C	4	1	0	4
G	1	4	4	0

1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT



0 change

1 change | 1 change | 8 (4x2) changes | 4 changes

Tree length: $L = 1+1+8+4+0 = 14$

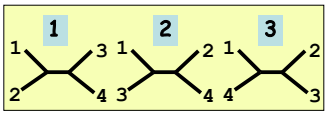
54

Phylogenetic methods (Character-based)

➢ Maximum Parsimony (MP): weighted

	A	T	C	G
A	0	4	4	1
T	4	0	1	4
C	4	1	0	4
G	1	4	4	0

1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT



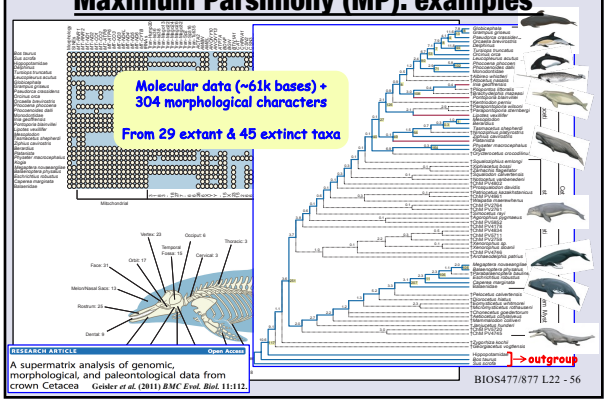
Tree	Sites 1	2	3	4	5	Total
1: ((1,2),(3,4))	1	1	8	4	0	14
2: ((1,3),(2,4))	2	2	4	4	0	12
3: ((1,4),(2,3))	2	2	8	4	0	16

$L(\tau) = \sum_{i=1}^{5} l_i$

MP tree!

55

Maximum Parsimony (MP): examples

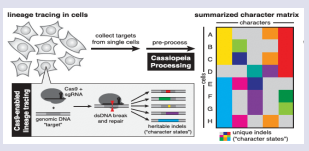


56

Maximum Parsimony (MP): examples

Inference of single-cell phylogenies from lineage tracing data using Cassiopeia

Jones et al. (2020) Genome Biol. 21: 92



- Each cell is tagged by indels induced by CRIPR/Cas9 gene-editing
- Cell growth was tracked for 21 days (34,557 cells in total)
- Indel information is obtained via Illumina sequencing
- A character matrix is generated from indels

57