Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 23

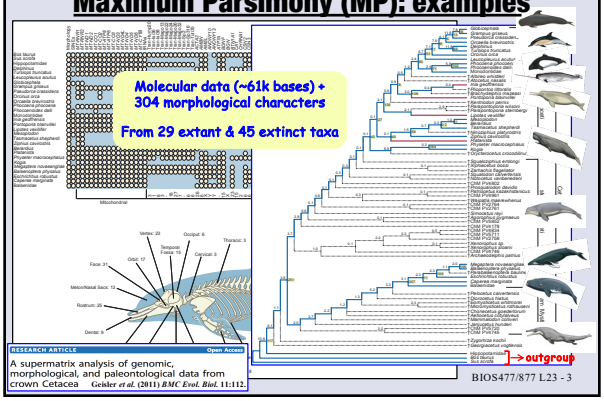
1

TODAY'S TOPICS

- ➢ Phylogenetic reconstruction
- Maximum parsimony method: use examples
- Maximum likelihood method
- Tree searching
  (Exhaustive, Branch-and-Bound, Heuristic)
- Branch support (bootstrap analysis, etc.)

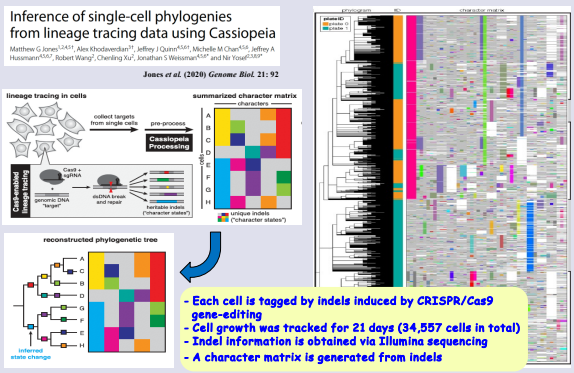
2

Maximum Parsimony (MP): examples



3

Maximum Parsimony (MP): examples



4

Phylogenetic methods

- ➢ Data types and tree-building methods

[Tree-building methods]	[Data types] 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.

5

Phylogenetic methods (Character-based)

- ➢ Maximum Likelihood (ML)
  - Chooses the tree that makes the observed data the most probable evolutionary outcome
  - Likelihood = Conditional probability of obtaining the observed sequences given a hypothesis (substitution model and tree)

$L(\tau, \theta)$ = Prob(Data | τ, θ)
= Prob(Aligned sequences | tree, model of evolution)

τ: a tree (including its topology, branch lengths)
θ: a set of parameters for a substitution model

➜ The likelihood calculated is not the probability of tree or evolutionary model! It is the probability of the data.

6

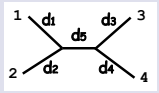
Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

$L(\tau, \theta)$ = Prob(Data | τ, θ)

Find the topology that gives the maximum $L(\tau, \theta)$, and simultaneously estimate all required parameters

- To compute the likelihood of a given tree,
  - τ: the topology and the maximum likelihood estimates for the tree's branch lengths (d_1, d_2, ...) need to be found
  - θ: the best values for the parameters for the evolutionary model need to be found, too

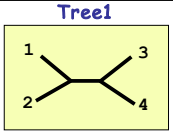


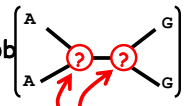
$$\begin{cases} r_t = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \\ s_t = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \end{cases}$$

7

Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

Data	Tree1
1 ATATT	
2 ATCGT	
3 GCAGT	
4 GCCGT	

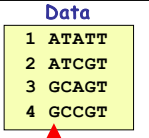
L(AAGG|Tree)=Prob 

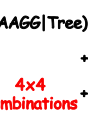
All possible combinations

8

Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

Data	Tree1
1 ATATT	
2 ATCGT	
3 GCAGT	
4 GCCGT	

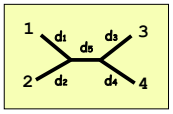
L(AAGG|Tree)=Prob() + Prob( ) + Prob( ) + Prob( )
+ Prob( ) + Prob( ) + Prob( ) + Prob( )
+ Prob( ) + Prob( ) + Prob( ) + Prob( )
+ Prob( ) + Prob( ) + Prob( ) + Prob( )

4x4 combinations

9

Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

Data	Tree1
1 ATATT	
2 ATCGT	
3 GCAGT	
4 GCCGT	

L(AAGG|Tree)=Prob() + ...

$P_{AA}(d_1) \times P_{AA}(d_2) \times P_{AA}(d_5) \times P_{AG}(d_3) \times P_{AG}(d_4)$

$P_{ij}(d_k)$: the probability that two sequences separated by d_k would have character states i and j

Transition probability matrix for JC model (see Lec. 20 & "Derivation of JC equation")

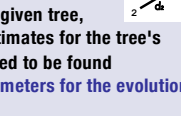
$$P(t) = \begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix} \quad \begin{cases} r_t = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \\ s_t = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \end{cases}$$

$d_k = \alpha t_k$, and $P_{ij}(d_k) = \begin{cases} r_{tk} \text{ if } i=j \\ s_{tk} \text{ if } i \neq j \end{cases}$

10

Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

Data	Tree1
1 ATATT	
2 ATCGT	
3 GCAGT	
4 GCCGT	

L_{Tree1} = L(AAGG|Tree1) x L(TTCC|Tree1) x L(ACAC|Tree1) x L(TGGG|Tree1) x L(TTTT|Tree1)
= $\prod^5 L_i$ where L_i = L(site$_i$|Tree1)

$\ln(L_{Tree1})$ = ln[L(AAGG|Tree1)] + ln[L(TTCC|Tree1)] + ln[L(ACAC|Tree1)] + ln[L(TGGG|Tree1)] + ln[L(TTTT|Tree1)]
= $\sum^5$ ln[L(site$_i$|Tree1)]

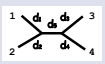
Likelihood of the data (multiple alignment) given Tree1

11

Phylogenetic methods (Character-based)

➢ **Maximum Likelihood (ML)**

$L(\tau, \theta)$ = Prob(Data | τ, θ)



- To compute the likelihood of a given tree,
  - τ: the maximum likelihood estimates for the tree's branch lengths (d_1, d_2, ...) need to be found
  - θ: the best values for the parameters for the evolutionary model need to be found, too
    - → [JC model] α (single evolutionary rate)
      [2-paramter model] transition/transversion ratio (TS/TV)
      base composition, gamma shape parameter (a)
      etc. etc.
    - → Number of the parameters depends on the model

$$\begin{cases} r_t = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \\ s_t = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \end{cases}$$

Find the topology that gives the maximum $L(\tau, \theta)$, and simultaneously estimate all required parameters

12

Phylogenetic methods

	Distances	Characters
Clustering	UPGMA NJ	
Optimality criterion	ME FM	MP ML

- Clustering *vs.* search methods
  - Clustering methods (UPGMA, Neighbor-joining)
    - → Do not search all possible topologies
    - → Very fast
    - → Produce only one tree
  - Search methods
    - → Use optimality criterion (minimum evolution, maximum parsimony, maximum likelihood)
    - → Exhaustive search for all possible topologies is not possible for a large number of taxa
    - → A heuristic search algorithm needs to be used

How can we search trees?

13

Phylogenetic methods: tree searching

- Number of possible tree topologies

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1

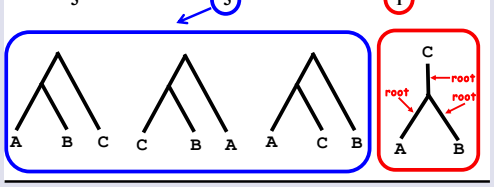
A —— B

14

Phylogenetic methods: tree searching

- Number of possible tree topologies

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1

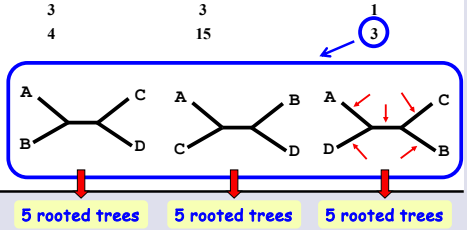


15

Phylogenetic methods: tree searching

- Number of possible tree topologies

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3



16

Phylogenetic methods: tree searching

- Number of possible tree topologies

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

Impossible to examine all of the possible trees!

17

Tree-searching methods (unrooted)

- Exhaustive search
  - Possible only for a few taxa (11 or fewer)
  - Number of possible unrooted tree:

$$B(t) = \prod_{i=3}^{t} (2i-5)$$

t : number of taxa

e.g., $B(7) = 1 \times 3 \times 5 \times 7 \times 9 = 945$
$B(10) = B(7) \times 11 \times 13 \times 15 > 2 \times 10^6$
$B(20) > 2 \times 10^{20}$

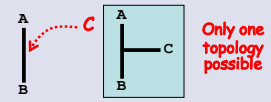
  - An algorithm is required to guarantee generation of all possible trees

18

Tree-searching methods (unrooted)

- **Exhaustive search**

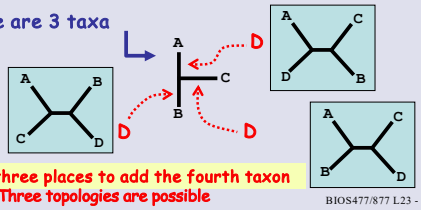
If there are 2 taxa
Only one possible topology



Only one topology possible

There is only one place to add the third taxon

If there are 3 taxa



There are three places to add the fourth taxon
Three topologies are possible

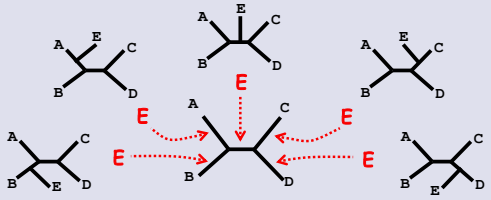
BIOS477/877 L23 - 19

19

Tree-searching methods (unrooted)

- **Exhaustive search**

If there are 4 taxa



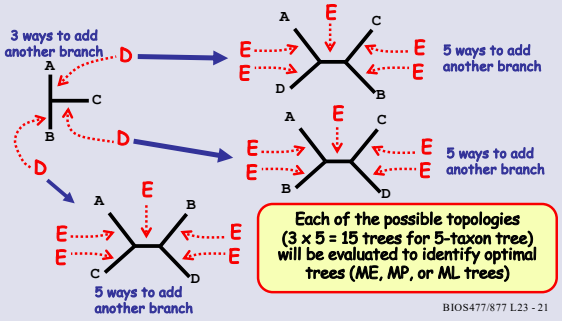
There are five places to add the fifth taxon
Five topologies are possible

BIOS477/877 L23 - 20

20

Tree-searching methods (unrooted)

- **Exhaustive search for 5-taxon trees**



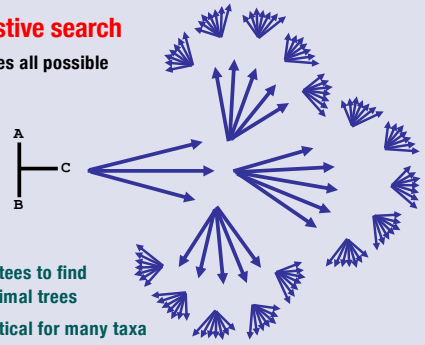
Each of the possible topologies (3 x 5 = 15 trees for 5-taxon tree) will be evaluated to identify optimal trees (ME, MP, or ML trees)

BIOS477/877 L23 - 21

21

Tree-searching methods (unrooted)

- **Exhaustive search**
  - searches all possible trees
  - guarantees to find the optimal trees
  - Impractical for many taxa



BIOS477/877 L23 - 22

22

Tree-searching methods (unrooted)

- **Branch-and-bound method**
  - Alternative exact method
  - Useful for 25 or so taxa
  - Implicitly evaluates all possible trees
  - But cuts off search paths that do not lead to optimal trees
    - ➜ Reduces the number of trees to evaluate
    - ➜ Will find the optimal tree(s)

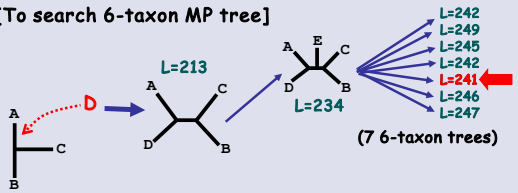
BIOS477/877 L23 - 23

23

Tree-searching methods (unrooted)

- **Branch-and-bound method**

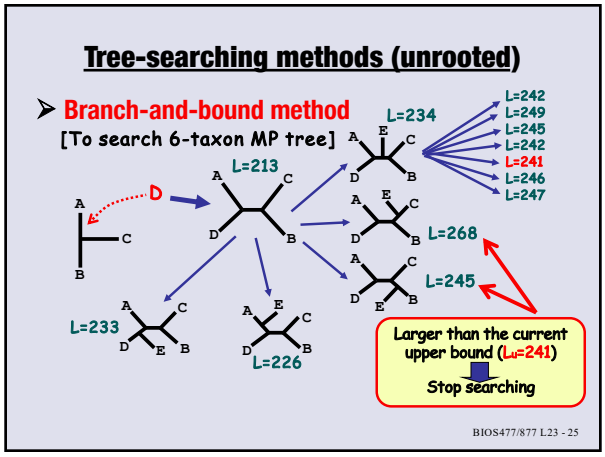
[To search 6-taxon MP tree]



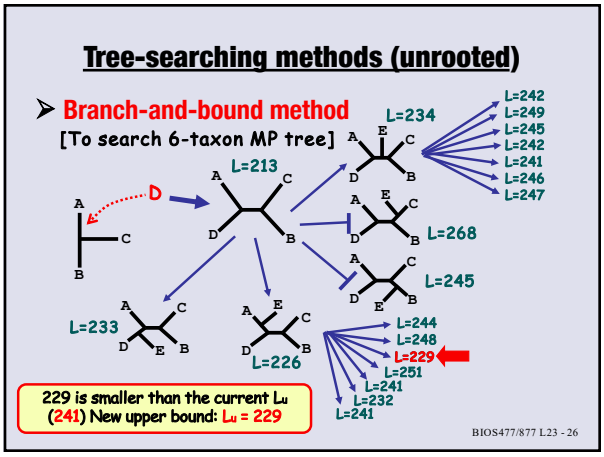
241 becomes the upper bound (L_u=241) for searching 6-taxon trees
(MP example: searching the shortest tree length, L)

BIOS477/877 L23 - 24

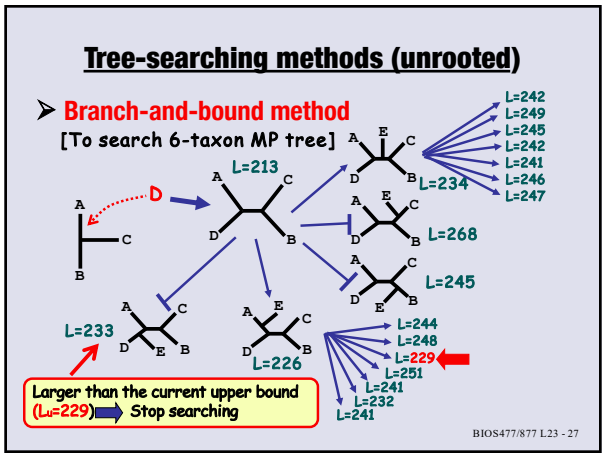
24



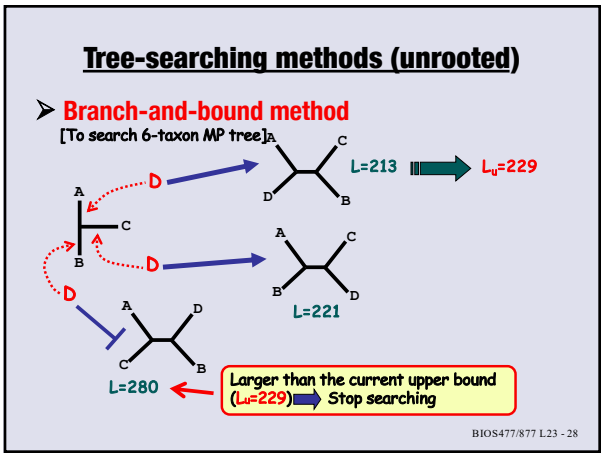
25



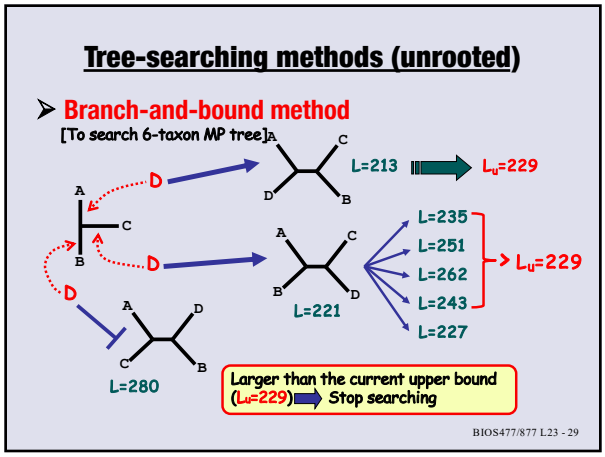
26



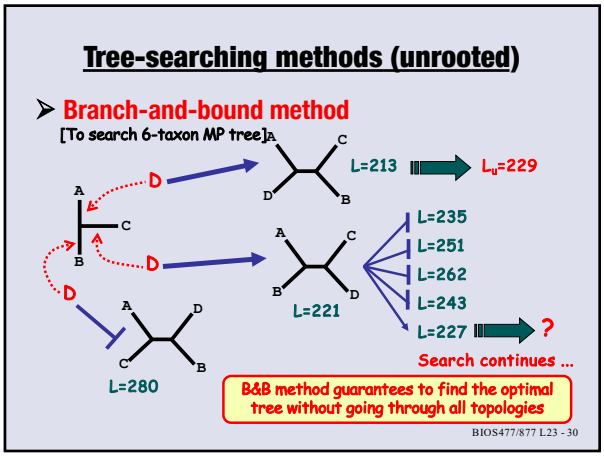
27



28



29

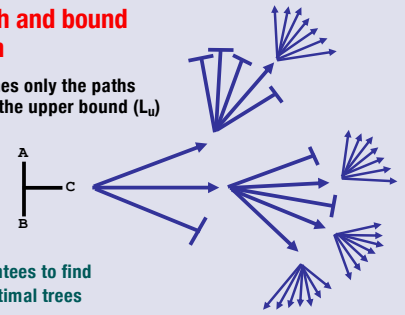


30

Tree-searching methods (unrooted)

➢ **Branch and bound search**

- searches only the paths under the upper bound (L_u)



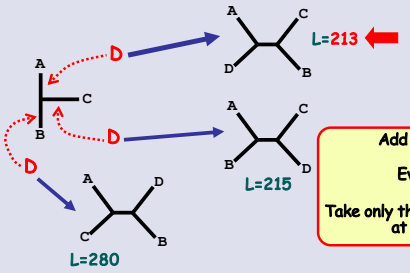
- guarantees to find the optimal trees

31

Tree-searching methods (unrooted)

➢ **Heuristic (approximation) method**

- Stepwise addition: greedy algorithm

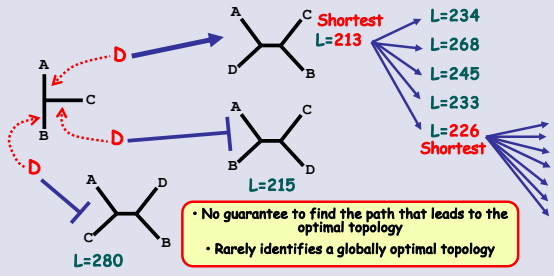


32

Tree-searching methods (unrooted)

➢ **Heuristic (approximation) method**

- Stepwise addition: greedy algorithm

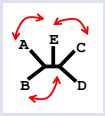


33

Tree-searching methods (unrooted)

➢ **Heuristic (approximation) method**

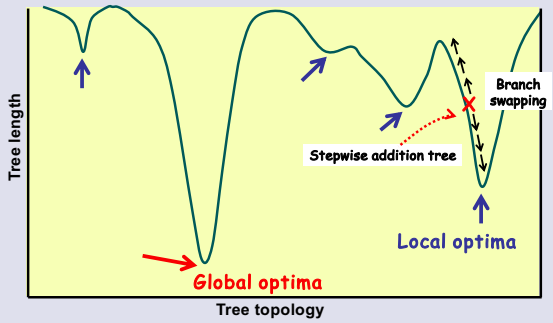
- Stepwise addition: greedy algorithm
  Rarely finds the global optima
  → Can be improved (refinement)



- Branch-swapping
  → Cutting off one or more subtrees and reassembling them to generate locally different trees
  → Repeat this process many times to find better topologies
  → Still can be entrapped in local optima
- Random addition + branch-swapping (many times)
  → Multiple optimal islands can be identified

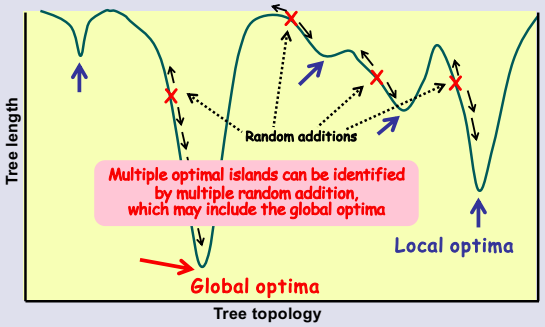
34

Tree-searching landscape



35

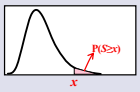
Tree-searching landscape



36

Reliability of inferred trees

- Bootstrap analysis
  - Evaluates the reliability of each cluster
  - A sampling technique to estimate the statistical errors when underlying sampling distribution is unknown



  - → Approximates the underlying distribution by resampling from the original dataset
  - First applied by Felsenstein (1985) for phylogenetic analysis

37

Reliability of inferred trees

- Bootstrap analysis
  - → Resampling is done on sequence data

	1	2	3	4	5	6	7	8	Site #
S1	G	C	A	G	T	A	C	T	
S2	G	T	A	G	T	A	C	T	
S3	A	C	A	A	T	A	C	C	
S4	A	C	A	A	C	A	C	T	
S5	G	C	G	G	C	A	T	T	

Multiple alignment

38

Reliability of inferred trees

- Bootstrap analysis
  - → Resampling is done on sequence data

1	2	3	4	5	6	7	8
G	C	A	G	T	A	C	T
G	T	A	G	T	A	C	T
A	C	A	A	T	A	C	C
A	C	A	A	C	A	C	T
G	C	G	G	C	A	T	T

8 independent samples

Configuration for each site is not changed

Each column of the multiple alignment is treated as an independent sample

39

Reliability of inferred trees

- Bootstrap analysis

1	2	3	4	5	6	7	8
G	C	A	G	T	A	C	T
G	T	A	G	T	A	C	T
A	C	A	A	T	A	C	C
A	C	A	A	C	A	C	T
G	C	G	G	C	A	T	T

	1
S1	G
S2	G
S3	A
S4	A
S5	G

Columns are randomly sampled

40

Reliability of inferred trees

- Bootstrap analysis

1	2	3	4	5	6	7	8
G	C	A	G	T	A	C	T
G	T	A	G	T	A	C	T
A	C	A	A	T	A	C	C
A	C	A	A	C	A	C	T
G	C	G	G	C	A	T	T

	14
S1	GG
S2	GG
S3	AA
S4	AA
S5	GG

Columns are randomly sampled

41

Reliability of inferred trees

- Bootstrap analysis

1	2	3	4	5	6	7	8
G	C	A	G	T	A	C	T
G	T	A	G	T	A	C	T
A	C	A	A	T	A	C	C
A	C	A	A	C	A	C	T
G	C	G	G	C	A	T	T

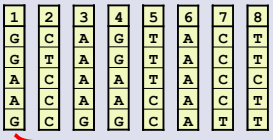
	146
S1	GGA
S2	GGA
S3	AAA
S4	AAA
S5	GGA

Columns are randomly sampled

42

Reliability of inferred trees

➢ Bootstrap analysis



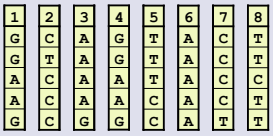

```
    1461
S1  GGAG
S2  GGAG
S3  AAAA
S4  AAAA
S5  GGAG
```

Multiple samplings are allowed from the same column

43

Reliability of inferred trees

➢ Bootstrap analysis



8 samples (sites)

A bootstrap replicate with the same length (the same number of sites) as the original alignment

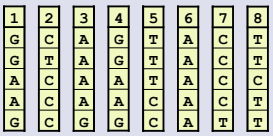
```
    14614853
S1  GGAGGTTA
S2  GGAGGTTA
S3  AAAAACTA
S4  AAAAATCA
S5  GGAGGTCG
```

8 samples (sites)

44

Reliability of inferred trees

➢ Bootstrap analysis




```
    14614853          74761232          85851124
S1  GGAGGTTA      S1  CGCAGCAC      S1  TTTTGGCG
S2  GGAGGTTA      S2  CGCAGTAT      S2  TTTTGGTG
S3  AAAAACTA      S3  CACAACAC      S3  CTCTAACA
S4  AAAAAT        S4  CACAACAC      S4  TCTCAACA
S5  GGAGGT        S5  TGTAGCGC      S5  TCTCGGCG
```

Many resamplings are done (~ 1000 times)

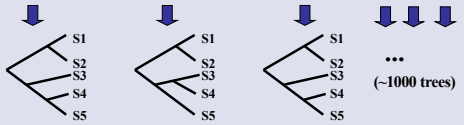
45

Reliability of inferred trees

➢ Bootstrap analysis

```
    14614853          74761232          85851124
S1  GGAGGTTA      S1  CGCAGCAC      S1  TTTTGGCG
S2  GGAGGTTA      S2  CGCAGTAT      S2  TTTTGGTG
S3  AAAAACTA      S3  CACAACAC      S3  CTCTAACA
S4  AAAAATCA      S4  CACAACAC      S4  TCTCAACA
S5  GGAGGTCG      S5  TGTAGCGC      S5  TCTCGGCG
```

... (~1000 alignments)

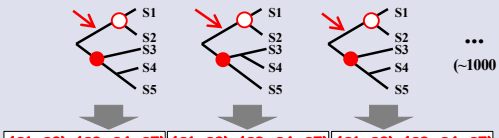


A phylogeny is reconstructed from each pseudoreplicate

46

Reliability of inferred trees

➢ Bootstrap analysis



(S1, S2), (S3, S4, S5) | (S1, S2), (S3, S4, S5) | (S1, S2), (S3, S4, S5)

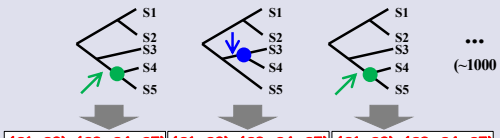
Count the number of each cluster

(S1, S2), (S3, S4, S5): 3/3 = 100%

47

Reliability of inferred trees

➢ Bootstrap analysis



(S1, S2), (S3, S4, S5) | (S1, S2), (S3, S4, S5) | (S1, S2), (S3, S4, S5)
(S1, S2, S3), (S4,S5) | (S3, S4), (S1, S2, S5) | (S1, S2, S3), (S4,S5)

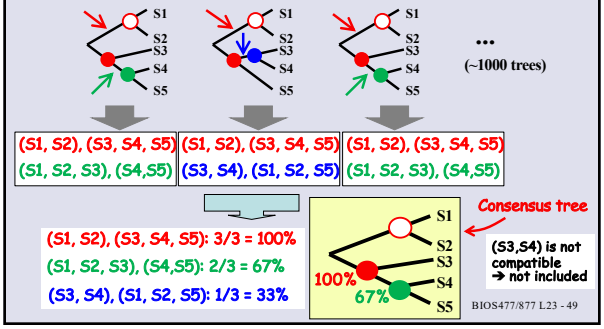
Count the number of each cluster

(S1, S2), (S3, S4, S5): 3/3 = 100%
(S1, S2, S3), (S4,S5): 2/3 = 67%
(S3, S4), (S1, S2, S5): 1/3 = 33%

48

Reliability of inferred trees

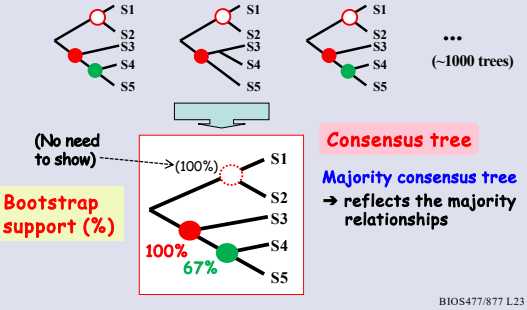
➢ Bootstrap analysis



49

Reliability of inferred trees

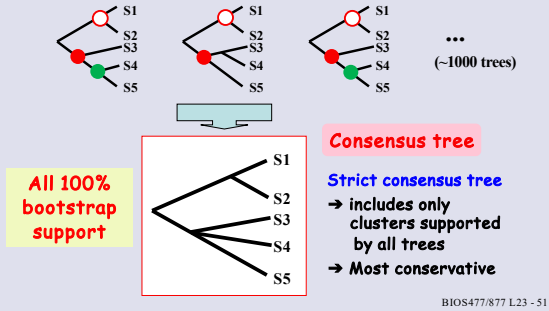
➢ Bootstrap analysis & consensus tree



50

Reliability of inferred trees

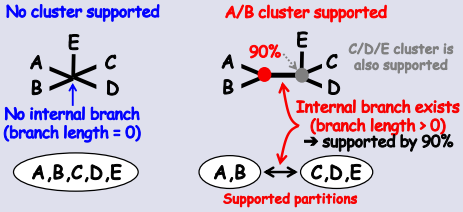
➢ Bootstrap analysis & consensus tree



51

Interpreting bootstrap support

➢ Bootstrap values can be considered to measure
➔ the probability of the branch length being greater than 0



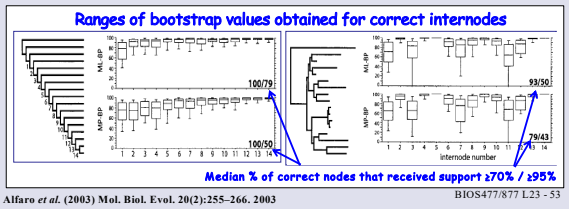
- Bootstrap values > 95% or 99% (often 70~80% used as a threshold)
  ➔ the internal branch length is considered to be significantly positive (>0); the two partitions are supported

52

Interpreting bootstrap support

➢ Bootstrap values can be considered to measure
➔ the probability of the branch length being greater than 0

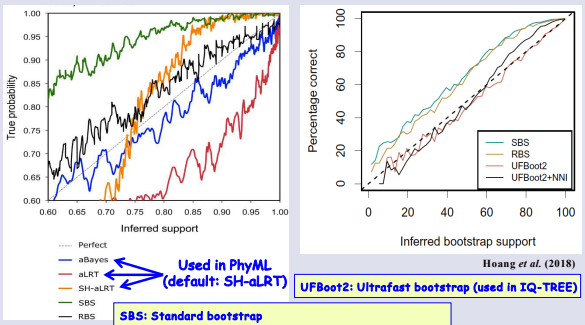
- Bootstrap values are also shown to give often conservative estimates of statistical confidence
- If the internal branches are very short or close to 0, both underestimation and overestimation could happen



Alfaro *et al.* (2003) Mol. Biol. Evol. 20(2):255–266. 2003

53

Interpreting branch supporting values



54

Renewing Felsenstein's phylogenetic bootstrap in the era of big data

F. Lemoine[1,2], J.-B. Domelevo Entfellner[3,4], E. Wilkinson[5], D. Correia[1], M. Dávila Felipe[1], T. De Oliveira[5,6] & O. Gascuel[1,7]*

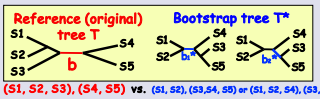
452 | NATURE | VOL 556 | 26 APRIL 2018

Transfer distance (or R distance):

The minimum number of elements to be transferred (or removed) to transform on partition into the other

(S1, S2), (S3, S4, S5) ⟺ Distance=0 ⟺ (S1, S2), (S3, S4, S5)

(S1, S2, S3), (S4, S5) ⟺ Distance=1 ⟺ (S1, S2), (S3, S4, S5)



Transfer Bootstrap Expectation (TBE) for branch *b*:

$$\mathrm{TBE}(b) = 1 - \frac{\overline{\phi(b, T^*)}}{p-1}$$

Average from all BS trees

p: number of taxa from the smaller of the two clusters

(S1, S2, S3), (S4, S5) vs. (S1, S2), (S3,S4, S5) or (S1, S2, S4), (S3,S5)

$\delta(b,b^*)$: Distance between branch b in tree T and branch b^* in BS tree T^*:

- Compare the two partitions defined by branch *b* in tree T and all branches in BS tree T*
- Find the b^* with the minimum $\delta(b,b^*) \Rightarrow \phi(b,T^*)$

$\delta(b,b1^*)=1$, $\delta(b,b2^*)=2$

BS: count only identical partitions

TBE: include most similar partitions (no need to be identical)

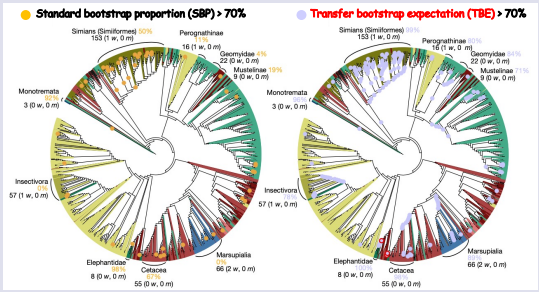
55

Transfer bootstrap for big phylogenies

Renewing Felsenstein's phylogenetic bootstrap in the era of big data

A fast and memory-efficient implementation of the transfer bootstrap

Sarah Lutteropp[1,*], Alexey M. Kozlov[1] and Alexandros Stamatakis[1,2]



- Deep branches in large phylogenies are often not supported by SBP
- TBE supports are higher without inducing falsely supported branches

56