Phylogenetics:
Parsimony

COMP 571
Luay Nakhleh, Rice University

1

The Problem

- Input: Multiple alignment of a set S of sequences
- Output: Tree T leaf-labeled with S

2

Assumptions

- Characters are mutually independent
- Following a speciation event, characters continue to evolve independently

3

4

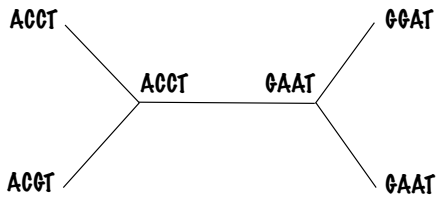
* In parsimony-based methods, the inferred tree is fully labeled.

5-1

ACCT GGAT

ACGT GAAT

5-2



A Simple Solution: Try All Trees

6

- Problem:
  - (2n-3)!! rooted trees
  - (2m-5)!! unrooted trees

A Simple Solution: Try All Trees

7

Number of Taxa	Number of unrooted trees	Number of rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
20	2.22E+20	8.20E+21
30	8.69E+36	4.95E+38
40	1.31E+55	1.01E+57
50	2.84E+74	2.75E+76
60	5.01E+94	5.86E+96
70	5.00E+115	6.85E+117
80	2.18E+137	3.43E+139

Solution

8

- Define an optimization criterion
- Find the tree (or, set of trees) that optimizes the criterion
- Two common criteria: parsimony and likelihood

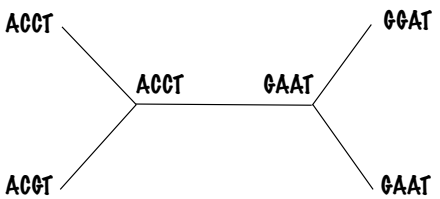
Parsimony

9

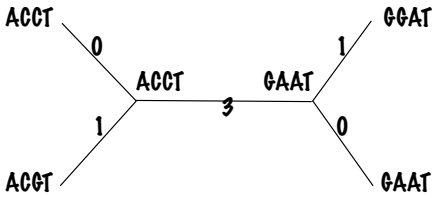
10

- The parsimony of a fully-labeled unrooted tree *T*, is the sum of lengths of all the edges in *T*
- Length of an edge is the Hamming distance between the sequences at its two endpoints
- PS(T)

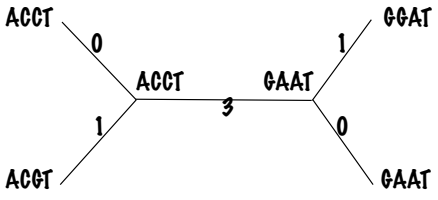
11-1



11-2



11-3



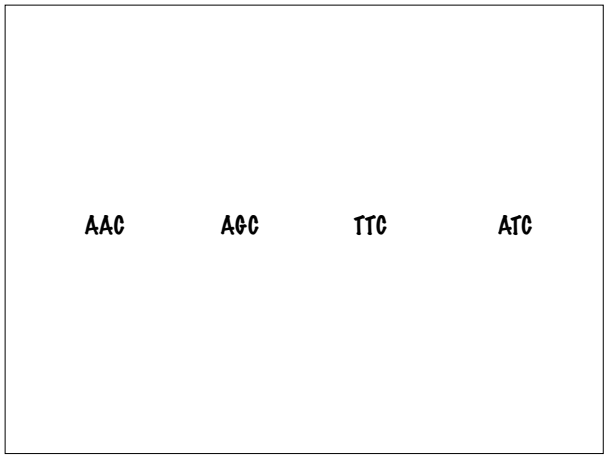
Parsimony score = 5

12

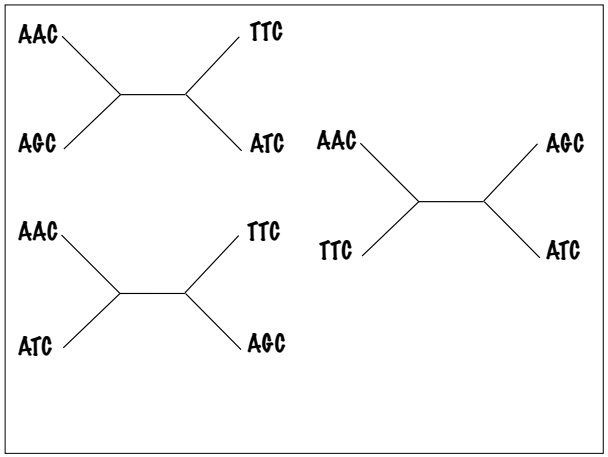
Maximum Parsimony (MP)

* Input: a multiple alignment S of n sequences
* Output: tree T with n leaves, each leaf labeled by a unique sequence from S, internal nodes labeled by sequences, and PS(T) is minimized

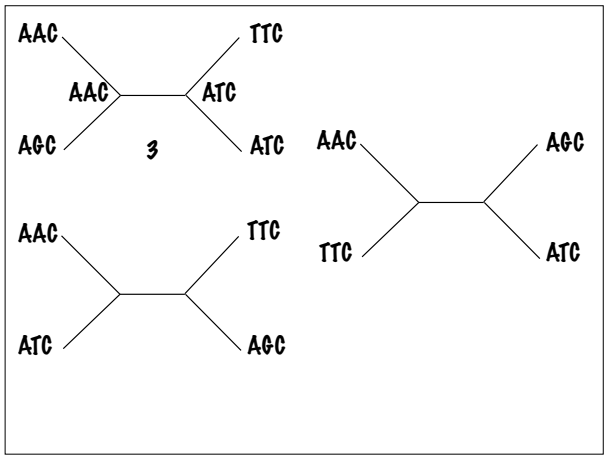
13

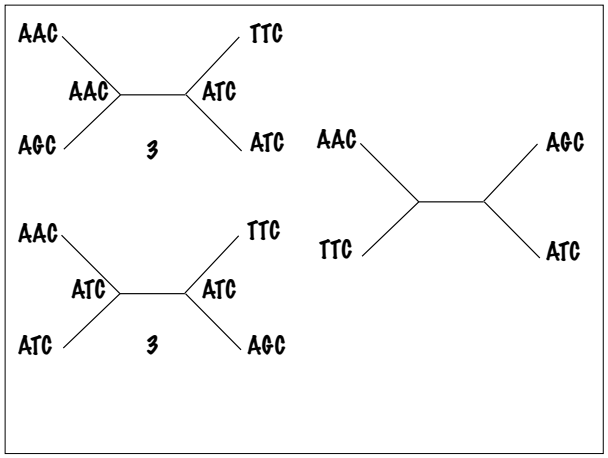

AAC
AGC
TTC
ATC

14-1

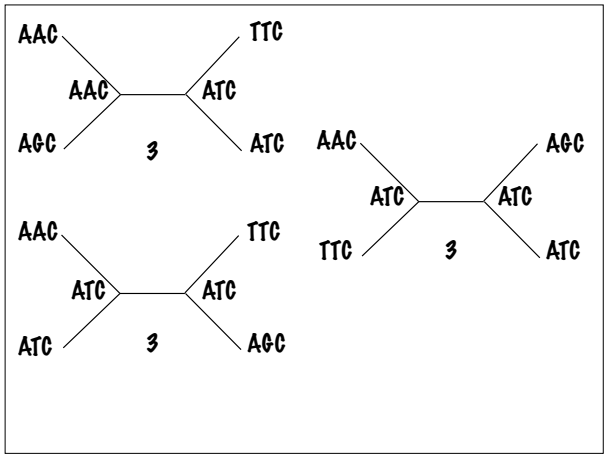

AAC
TTC
AGC
ATC
AAC
AGC
TTC
ATC
AAC
TTC
ATC
AGC

14-2

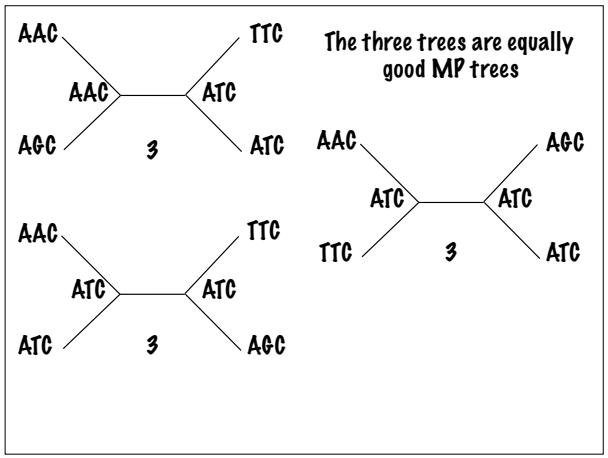

AAC
TTC
AAC
ATC
AGC
ATC
3
AAC
AGC
TTC
ATC
AAC
TTC
ATC
AGC


AAC
TTC
AAC
ATC
AGC
3
ATC
AAC
AGC
TTC
ATC
AAC
TTC
ATC
ATC
ATC
3
AGC

14-3

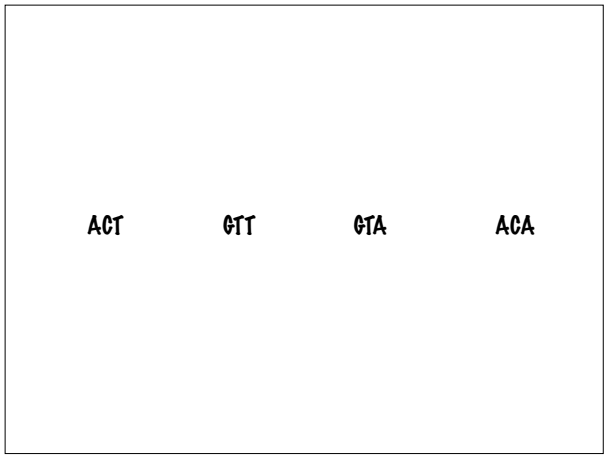

AAC
TTC
AAC
ATC
AGC
3
ATC
AAC
AGC
ATC
ATC
TTC
3
ATC
AAC
TTC
ATC
ATC
ATC
3
AGC

14-4

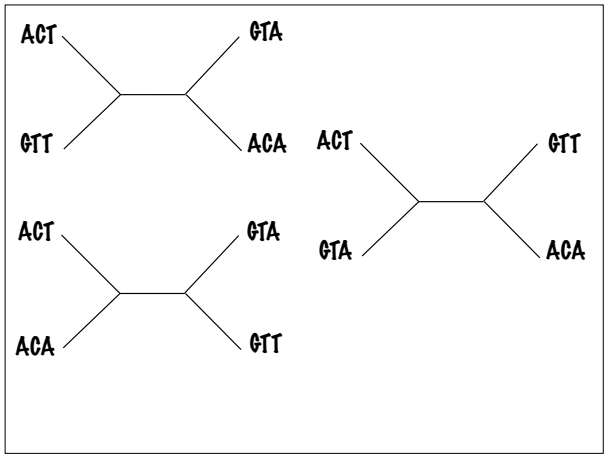

AAC
TTC
AAC
ATC
AGC
3
ATC
The three trees are equally good MP trees
AAC
AGC
ATC
ATC
TTC
3
ATC
AAC
TTC
ATC
ATC
ATC
3
AGC

14-5

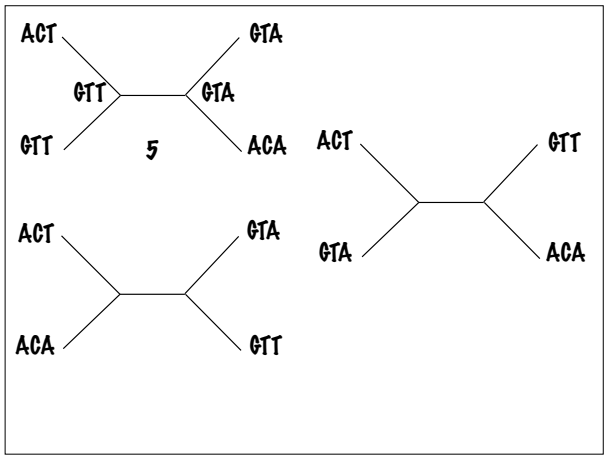
15

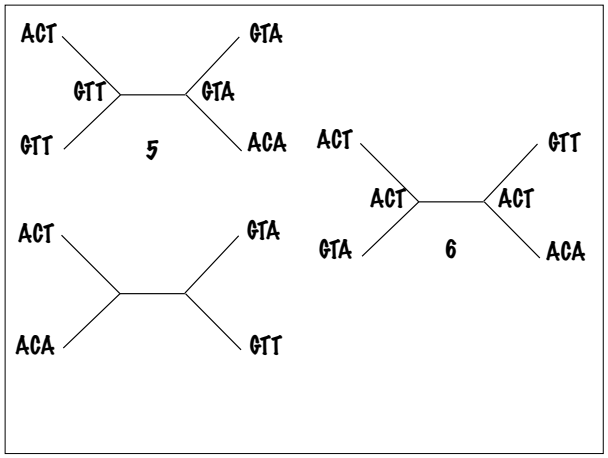

ACT
GTT
GTA
ACA

16-1

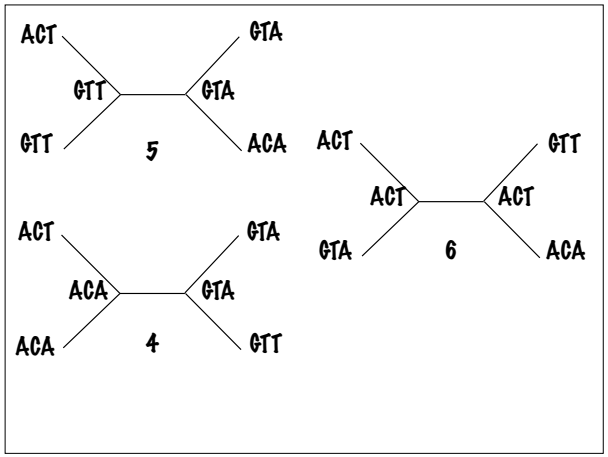

ACT
GTA
GTT
ACA
ACT
GTT
GTA
ACA
ACT
GTA
ACA
GTT

16-2

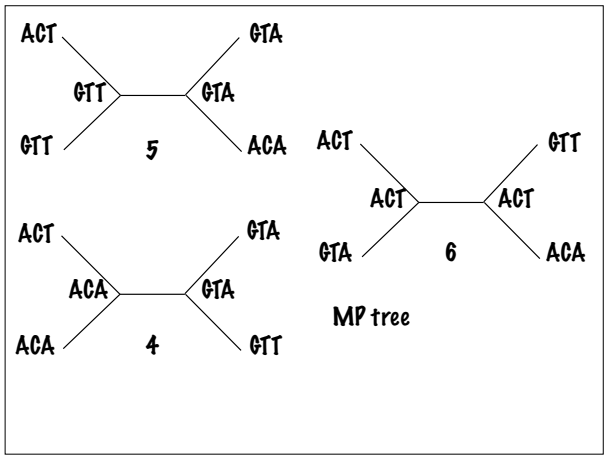

ACT
GTA
GTT
GTA
GTT
ACA
5
ACT
GTT
GTA
ACA
ACT
GTA
ACA
GTT


ACT
GTA
GTT
GTA
GTT
ACA
5
ACT
GTT
ACT
ACT
GTA
ACA
6
ACT
GTA
ACA
GTT

16-3


ACT
GTA
GTT
GTA
GTT
ACA
5
ACT
GTT
ACT
ACT
GTA
ACA
6
ACT
GTA
ACA
GTT
4

16-4


ACT
GTA
GTT
GTA
GTT
ACA
5
ACT
GTT
ACT
ACT
GTA
ACA
6
ACT
GTA
ACA
GTT
4
MP tree

16-5

Weighted Parsimony

17

- Each transition from one character state to another is given a weight
- Each character is given a weight
- See a tree that minimizes the weighted parsimony

18

- Both the MP and weighted MP problems are NP-hard

A Heuristic For Solving the MP Problem

19

- Starting with a random tree *T*, move through the tree space while computing the parsimony of trees, and keeping those with optimal score (among the ones encountered)
- Usually, the search time is the stopping factor

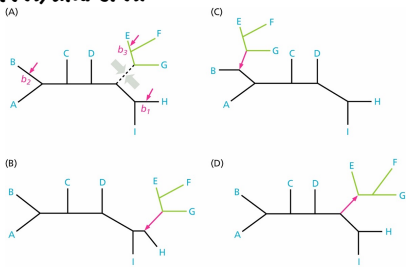
Two Issues

- How do we move through the tree search space?
- Can we compute the parsimony of a given leaf-labeled tree efficiently?

20

Searching Through the Tree Space

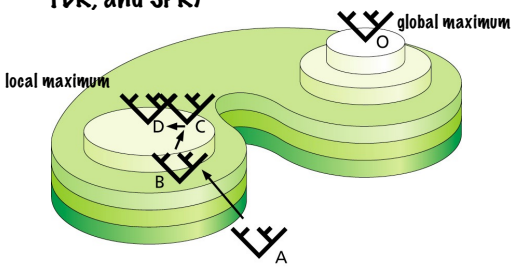
- Use tree transformation operations (NNI, TBR, and SPR)



21

Searching Through the Tree Space

- Use tree transformation operations (NNI, TBR, and SPR)



22

Computing the Parsimony Length of a Given Tree

23

- Fitch's algorithm
- Computes the parsimony score of a given leaf-labeled rooted tree
- Polynomial time

Fitch's Algorithm

24

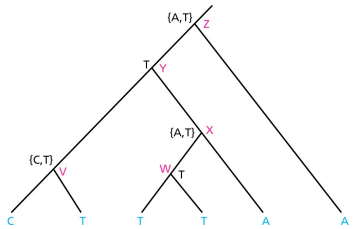
- Alphabet Σ
- Character c takes states from Σ
- v_c denotes the state of character c at node v

Fitch's Algorithm

25

- Bottom-up phase:
- For each node v and each character c, compute the set $S_{c,v}$ as follows:
  - If v is a leaf, then $S_{c,v}=\{v_c\}$
  - If v is an internal node whose two children are x and y, then

$$S_{c,v} = \begin{cases} S_{c,x} \cap S_{c,y} & S_{c,x} \cap S_{c,y} \neq \emptyset \\ S_{c,x} \cup S_{c,y} & \text{otherwise} \end{cases}$$



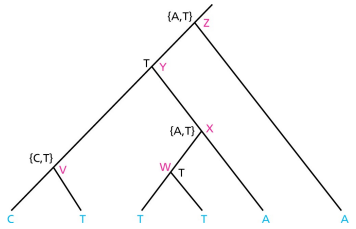
26

Fitch's Algorithm

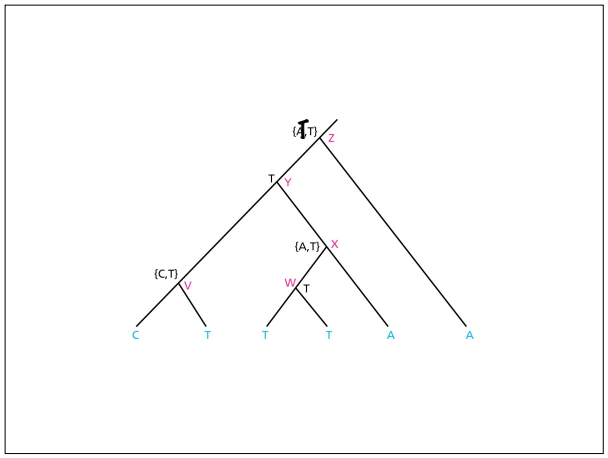
- Top-down phase:
- For the root r, let r_c=a for some arbitrary a in the set $S_{c,r}$
- For internal node v whose parent is u,

$$v_c = \begin{cases} u_c & u_c \in S_{c,v} \\ \text{arbitrary } \alpha \in S_{c,v} & \text{otherwise} \end{cases}$$

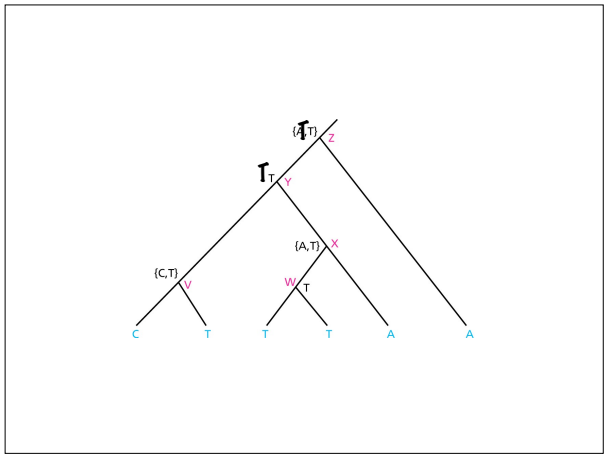
27



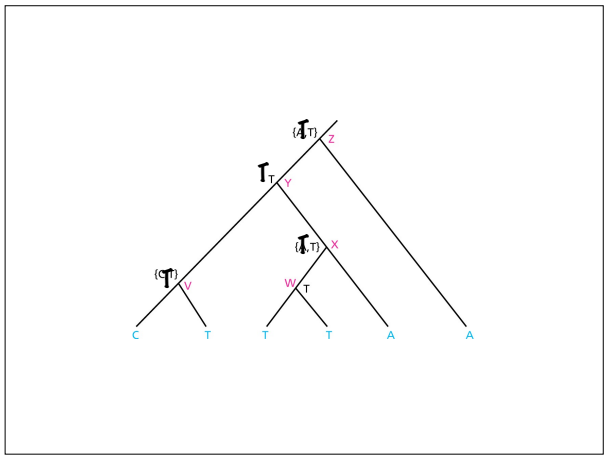
28-1


{A,T}
Z
T
Y
{A,T}
X
{C,T}
V
W
T
C
T
T
T
A
A

28-2

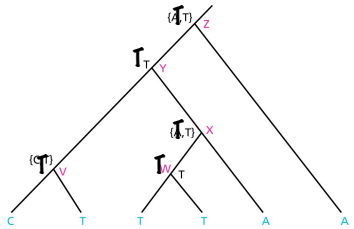

{A,T}
Z
T
Y
{A,T}
X
{C,T}
V
W
T
C
T
T
T
A
A

28-3

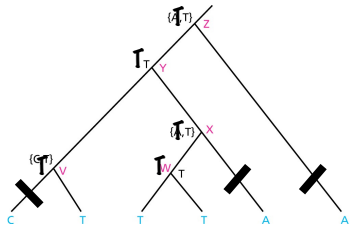

{A,T}
Z
T
Y
{A,T}
X
{C,T}
V
W
T
C
T
T
T
A
A

28-4

28-5



28-6



3 mutations

29

Fitch's Algorithm

* Takes time O(nkm), where n is the number of leaves in the tree, m is the number of sites, and k is the maximum number of states per site (for DNA, k=4)

Informative Sites and Homoplasy

30

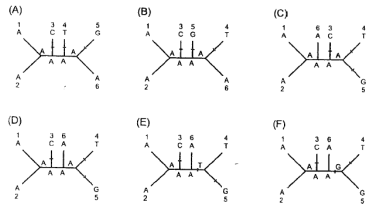
- Invariable sites: In the search for MP trees, sites that exhibit exactly one state for all taxa are eliminated from the analysis
- Only variable sites are used

Informative Sites and Homoplasy

31

- However, not all variable sites are useful for finding an MP tree topology
- Singleton sites: any nucleotide site at which only unique nucleotides (singletons) exist is not informative, because the nucleotide variation at the site can always be explained by the same number of substitutions in all topologies

32



C,T,G are three singleton substitutions ⇒non-informative site

All trees have parsimony score 3

Informative Sites and Homoplasy

33

- For a site to be informative for constructing an MP tree, it must exhibit at least two different states, each represented in at least two taxa
- These sites are called informative sites
- For constructing MP trees, it is sufficient to consider only informative sites

Informative Sites and Homoplasy

34

- Because only informative sites contribute to finding MP trees, it is important to have many informative sites to obtain reliable MP trees
- However, when the extent of homoplasy (backward and parallel substitutions) is high, MP trees would not be reliable even if there are many informative sites available

Measuring the Extent of Homoplasy

35

- The consistency index (Kluge and Farris, 1969) for a single nucleotide site (i-th site) is given by $c_i=m_i/s_i$, where
  - m_i is the minimum possible number of substitutions at the site for any conceivable topology (= one fewer than the number of different kinds of nucleotides at that site, assuming that one of the observed nucleotides is ancestral)
  - s_i is the minimum number of substitutions required for the topology under consideration

Measuring the Extent of Homoplasy

36

- The lower bound of the consistency index is not 0
- The consistency index varies with the topology
- Therefore, Farris (1989) proposed two more quantities: the retention index and the rescaled consistency index

The Retention Index

37

- The retention index, r_i, is given by $(g_i-s_i)/(g_i-m_i)$, where g_i is the maximum possible number of substitutions at the i-th site for any conceivable tree under the parsimony criterion and is equal to the number of substitutions required for a star topology when the most frequent nucleotide is placed at the central node

The Retention Index

38

- The retention index becomes 0 when the site is least informative for MP tree construction, that is, $s_i=g_i$

The Rescaled Consistency Index

$$rc_i = \frac{g_i - s_i}{g_i - m_i}\frac{m_i}{s_i}$$

39

Ensemble Indices

- The three values are often computed for all informative sites, and the ensemble or overall consistency index (CI), overall retention index (RI), and overall rescaled index (RC) for all sites are considered

40

Ensemble Indices

$$CI = \frac{\sum_i m_i}{\sum_i s_i}$$

$$RI = \frac{\sum_i g_i - \sum_i s_i}{\sum_i g_i - \sum_i m_i}$$

$$RC = CI \times RI$$

These indices should be computed only for informative sites, because for uninformative sites they are undefined

41

Homoplasy Index

42

- The homoplasy index is $HI = 1 - CI$
- When there are no backward or parallel substitutions, we have $HI = 0$. In this case, the topology is uniquely determined

A Major Caveat

43

- Maximum parsimony is not statistically consistent!

Questions?

44