


OPTIMALITY SOLUTIONS TO INFERRING TREES: PARSIMONY

PARSIMONY

[HTTP://EN.WIKTIONARY.ORG/WIKI/PARSIMONY](http://en.wiktionary.org/wiki/parimony)

Noun
parimony (uncountable)

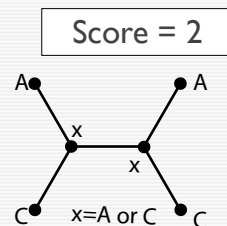
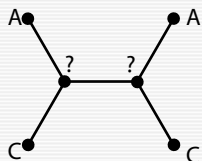
[edit]  [Wikipedia has an article on: Parimony](#)

- Great reluctance to spend money unnecessarily.
 - 1776, Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*: Parsimony, and not industry, is the immediate cause of the increase of capital. Industry, indeed, provides the subject which parsimony accumulates. But whatever industry might acquire, if parsimony did not save and store up, the capital would never be the greater.
- By analogy from (1), principle of using the least resources or explanations to solve a problem.

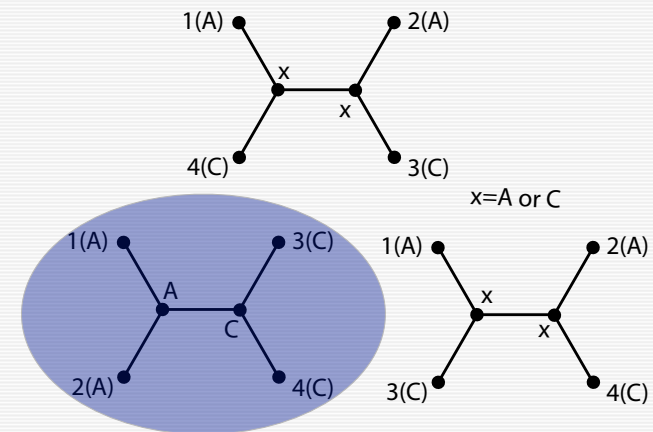
- find a tree (or group of trees) that minimizes the amount of mutations needed to describe the data
- Two subproblems:
 - Given the topology, and leaf labels, find the minimum cost of the tree
 - Find the topology that minimizes the cost

PARSIMONY EXAMPLES

- Let each leaf be labeled with a letter from some alphabet
 - Nucleotides, amino-acid residues, presence or absence of a trait
- Define the cost of changing one letter to another (a substitution), $c(x,y)$
 - The simplest case $c(x,y) = 1$, if $x \neq y$, and $c(x,x) = 0$.
- How would you assign interior node labels to minimize the total cost of the tree below?



TOPOLOGY SEARCH EXAMPLE



- Which topology is the best?

PARSIMONY ON MOLECULAR SEQUENCES

- Consider an alignment of nucleotide sequences

	20	30	40	50	60	70	80	90								
Human	GTC	ATTCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Chimpanzee	ATT	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAT	GAAC	GGA
Gorilla	GTT	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Orangutan	ACC	AGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Gibbon	ACC	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA

- We seek the topology that minimizes the cumulative parsimony score across all sites of the alignment

- L = length of tree, N = number of sites, l_j = length at site j

$$L(t) = \sum_{j=1}^N l_j$$

- To score each site, we need to solve a parsimony problem (assign interior labels optimally at that site)

INFORMATIVE SITES

- For standard parsimony (score = 0 or 1 for match or mismatch) some alignment columns will have the same score for all topologies – these sites are called uninformative

	20	30	40	50	60	70	80	90								
Human	GTC	ATTCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Chimpanzee	ATT	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAT	GAAC	GGA
Gorilla	GTT	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Orangutan	ACC	AGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA
Gibbon	ACC	TGCT	CATAA	TGCC	CACGG	ACTT	ACATCC	TACTT	ATTTG	CGCT	AGC	AAACT	CAAACT	TAC	GAAC	GGA

Invariable sites. Score 0 for all

Single difference. Score 1 for all topologies

- An informative sites must have at least two different characters with at least two instances of each character.

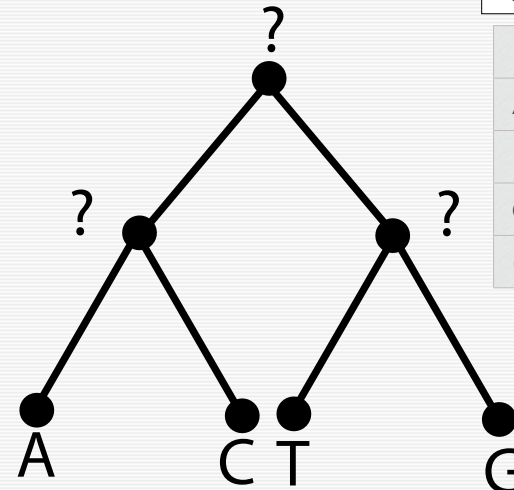
SANKOFF'S ALGORITHM

- Permits, for a fixed topology, to compute the optimal interior node label assignment and the parsimony score for a user specified cost function $c(x,y)$
- Uses the fact that the score of the subtree rooted at some interior node n is independent of the rest of the tree given the label of n 's parent.
- For each node n in the tree, the algorithm populates two arrays (of dimension equal to the size of the alphabet) – leaf and internal – in the topology (except the arbitrarily chosen interior node designated as root):
 - $\alpha_n(i)$ - the optimal score of the subtree rooted at n , given that the label of n 's parent is i .
 - $\beta_n(i)$ - the label at n that achieves score $\alpha_n(i)$
- The arrays can be computed recursively from the leaves up to the tree root
- The second pass from the root down to the leaves assigns the optimal labels

STEP 1: Assign 0 to observed leaf states

Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0



parent	A	T	C	G
α	0	∞	∞	∞
β	A	A	A	A

parent	A	T	C	G
α	∞	∞	0	∞
β	C	C	C	C

parent	A	T	C	G
α	∞	0	∞	∞
β	T	T	T	T

parent	A	T	C	G
α	∞	∞	∞	0
β	G	G	G	G

STEP 2: Traverse the tree from the leaves up (postorder) and populate cost/label arrays

Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

TRY A: $C(A_{n-1}, A_n) + 3 + 4 = 7$
 TRY T: $C(A_{n-1}, T_n) + 0 + 2 = 5$
 TRY C: $C(A_{n-1}, C_n) + 4 + 4 = 17$
 TRY G: $C(A_{n-1}, G_n) + 2 + 0 = 6$

parent	A	T	C	G
α	9	7	9	8
β	A	T	C	G

parent	A	T	C	G
α	5	2	2	3
β	T	T	T	G

parent	A	T	C	G
α	0	∞	∞	∞
β	A	A	A	A

parent	A	T	C	G
α	∞	∞	0	∞
β	C	C	C	C

parent	A	T	C	G
α	∞	0	∞	∞
β	T	T	T	T

parent	A	T	C	G
α	∞	∞	∞	0
β	G	G	G	G

STEP 3: Label the root

Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

FOR A:

$$\begin{aligned} & \text{MIN}(C(A,A)+5, C(A,T)+2, C(A,C)+2, C(A,G)+3) \\ & + \text{MIN}(C(A,A)+9, C(A,T)+7, C(A,C)+9, C(A,G)+8) \\ & = 5 + 9 = 14 \end{aligned}$$

state	A	T	C	G
α	14	9	15	10

parent	A	T	C	G
α	9	7	9	8
β	A	T	C	G

parent	A	T	C	G
α	5	2	2	3
β	T	T	T	G

parent	A	T	C	G
α	0	∞	∞	∞
β	A	A	A	A

parent	A	T	C	G
α	∞	∞	0	∞
β	C	C	C	C

parent	A	T	C	G
α	∞	0	∞	∞
β	T	T	T	T

parent	A	T	C	G
α	∞	∞	∞	0
β	G	G	G	G

STEP 4: Label the rest of tree

state	A	T	C	G
α	14	9	15	10

parent	A	T	C	G
α	9	7	9	8
β	A	T	C	G

parent	A	T	C	G
α	5	2	2	3
β	T	T	T	G

parent	A	T	C	G
α	0	∞	∞	∞
β	A	A	A	A

parent	A	T	C	G
α	∞	∞	0	∞
β	C	C	C	C

parent	A	T	C	G
α	∞	0	∞	∞
β	T	T	T	T

parent	A	T	C	G
α	∞	∞	∞	0
β	G	G	G	G

STEP 4: Label the rest of tree

Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

state	A	T	C	G
α	14	9	15	10

parent	A	T	C	G
α	9	7	9	8
β	A	T	C	G

parent	A	T	C	G
α	5	2	2	3
β	T	T	T	G

parent	A	T	C	G
α	0	∞	∞	∞
β	A	A	A	A

parent	A	T	C	G
α	∞	∞	0	∞
β	C	C	C	C

parent	A	T	C	G
α	∞	0	∞	∞
β	T	T	T	T

parent	A	T	C	G
α	∞	∞	∞	0
β	G	G	G	G

SANKOFF'S ALGORITHM

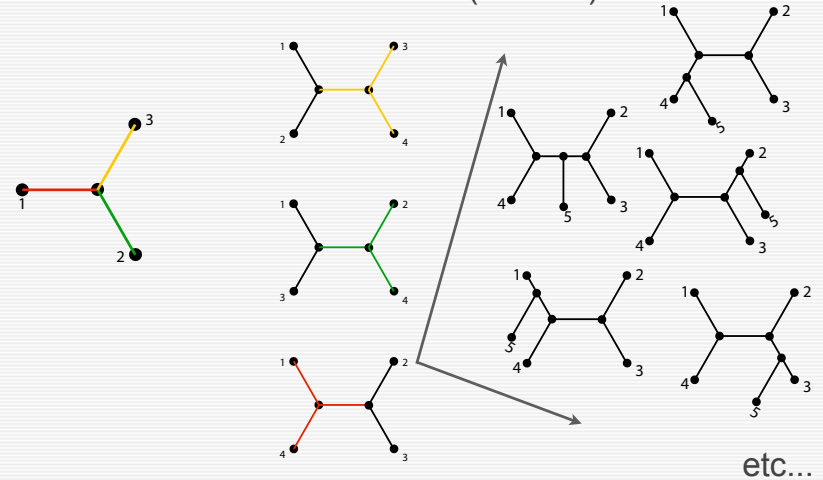
- we have the leaf and interior node labels that gives the minimum cost of a provided topology
- find the topology with the overall minimum cost = Tree Searching

N	T _u (N)
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025

N	T _u (N)
20	2.22E+20
50	2.84E+74
100	1.70E+182

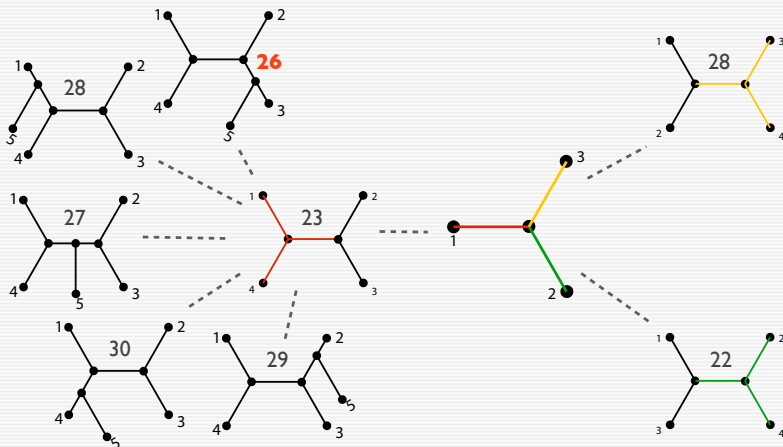
TREE SEARCHING

Exhaustive Search (n ≤ 11)



TREE SEARCHING

Branch & Bound (n = 12-25)

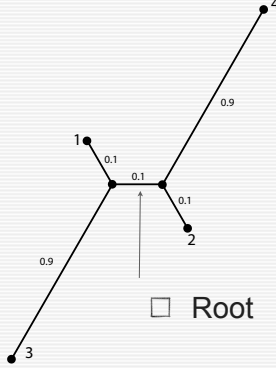


MORE ON PARSIMONY

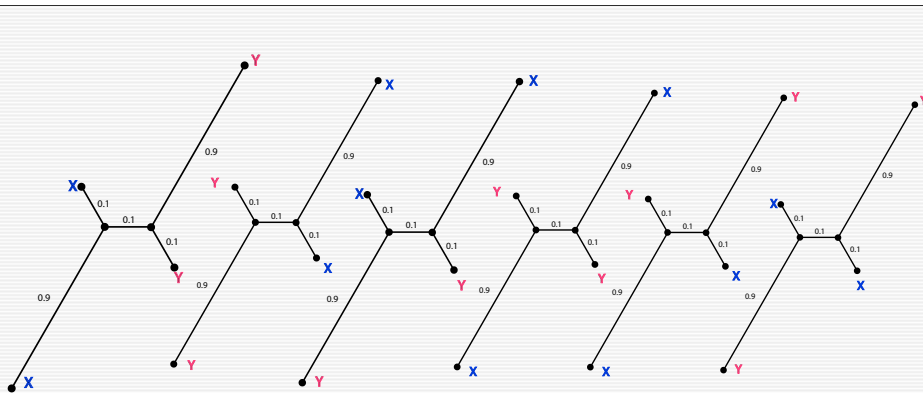
- Can be implemented very efficiently, permitting rapid screens of large sets of candidate trees
- Can be coupled with a branch and bound algorithm to exhaustively explore all topologies on ~20-30 taxa
- Works well if the assumptions of the method are not violated
 - The scoring matrix is reasonable
 - Branch lengths are short and not too different from one another

BUT...

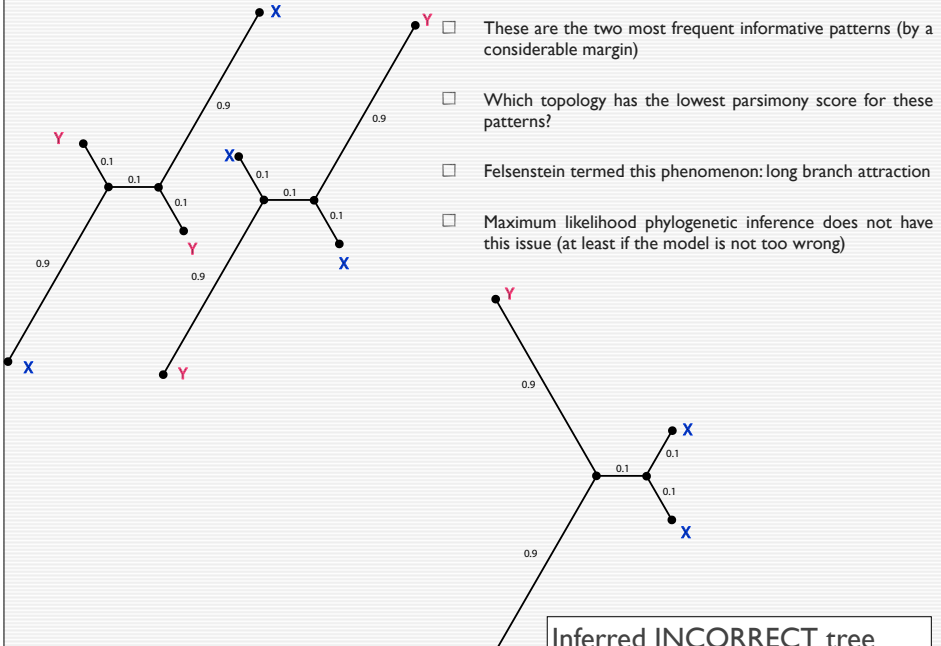
- Parsimony can also behave very poorly
- Under certain scenarios, the more data you give the method, the more certain it will be about inferring an incorrect tree
- This behavior is called positively misleading
- Example was given by Joe Felsenstein in a lead-up to his seminal work on using probabilistic models to reconstruct phylogenies.



- Consider the tree on the left.
- Treat each branch length as the probability that the sequence will mutate along this branch.
- Generate many sets of labels (alignment sites) using this model
- Reconstruct trees using parsimony (simple scoring function) from all sites.
- Which tree will parsimony tend to recover?



- What are the only 6 types of informative label patterns can be obtained?
- Which two have the highest probability of being generated?



- These are the two most frequent informative patterns (by a considerable margin)
- Which topology has the lowest parsimony score for these patterns?
- Felsenstein termed this phenomenon: long branch attraction
- Maximum likelihood phylogenetic inference does not have this issue (at least if the model is not too wrong)

Inferred INCORRECT tree