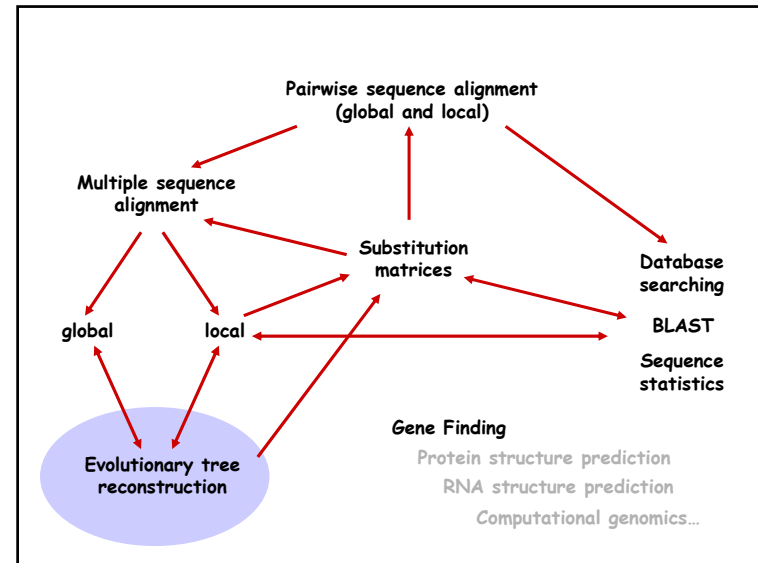


- Today:
 - Introduction to projects
 - Phylogeny reconstruction: Maximum parsimony
- Tuesday
 - PS2 is due
 - Phylogeny reconstruction: Distance-based methods

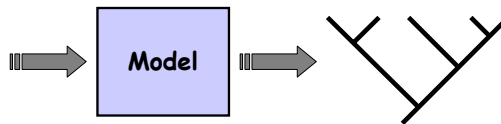


Phylogeny reconstruction

```

...atgcaaggagtcgcagagc...
...atgcaaggctctcgtagtgt...
...atggaggctctcccagtg...
...atgcgacgtcagtatagg...
...atgtgtggtctggcagtg...
...atgcgacctctcggagaat...

```



Given

- Multiple sequence alignment
- Model of sequence evolution

find the (binary) tree that is the best explains the data with respect to the model.

Finding the optimal tree

Given k taxa,

- Consider all trees with k leaves
- Score each tree with respect to chosen evolutionary model.
- Select highest scoring tree(s)

Phylogeny reconstruction is *NP*-complete:

Except in special cases when the data obeys specific constraints, the only way to find the best tree is to consider all trees.

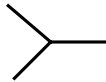
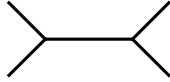
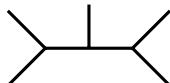
Finding the optimal tree

Given k taxa,

- Consider all trees with k leaves
- Score each tree with respect to chosen evolutionary model.
- Select highest scoring tree(s)

How many trees are there?

How many unrooted trees with k leaves?

		k	$E(k)$	$T(k)$
• Three taxa		3	3	1
• Four Taxa		4	5	3
• Five taxa		5	7	15
			

Number of unrooted trees for k taxa

$$E(k) = E(k-1) + 2 = 2k - 3$$

$$T(k) = E(k-1)T(k-1) = \prod_{i=3}^{k-1} (2i-3)$$

$$T(k) = \frac{(2k-5)!}{2^{k-3}(k-3)!}$$

The number of trees gets big fast

Number of leaves	Number of unrooted binary trees
3	1
4	3
5	15
6	105
10	2,027,025
20	2.2×10^{20}
50	2.8×10^{74}
500	1×10^{1074}

How big is that?

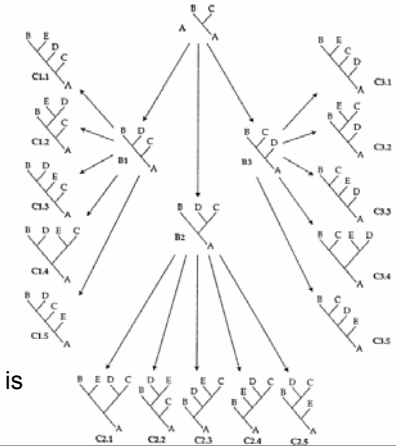
Number of leaves Number of unrooted binary trees

...	...
20	2.2×10^{20}
50	2.8×10^{74}
500	1×10^{1074}

Age of the universe (seconds):	4.42×10^{17}
Diameter of the universe:	2.70×10^{10}
Number of stars in the universe:	10^{22}

How do you find the optimal tree?

1. Exhaustive search (<12 taxa)



(Phylogeny reconstruction is NP-complete.)

How do you find the optimal tree?

Method	Result	Time	Typical k
Exhaustive search	Optimal solution	$T(k)$	12

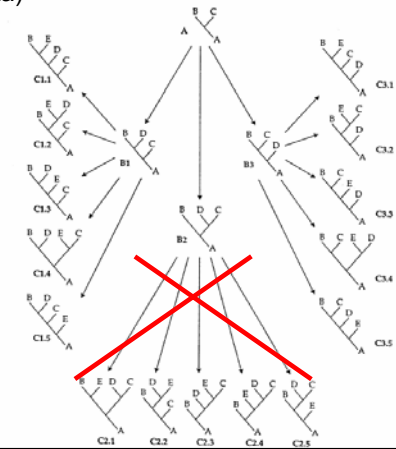
How do you find the optimal tree?

2. Branch-and-bound (<18 taxa)

Score is non-decreasing as you add edges

```

L = {T3}, C = infinity
For i = 3 to k {
  For each tree, t, in L{
    If Score(t) > C, prune search
    If Score(t) < C, C = Score(t).
    For every edge, e, in t {
      t' = t plus a new edge at e
      NewL = NewL U {t'}
    }
  }
  L = NewL
}
    
```



How do you find the optimal tree?

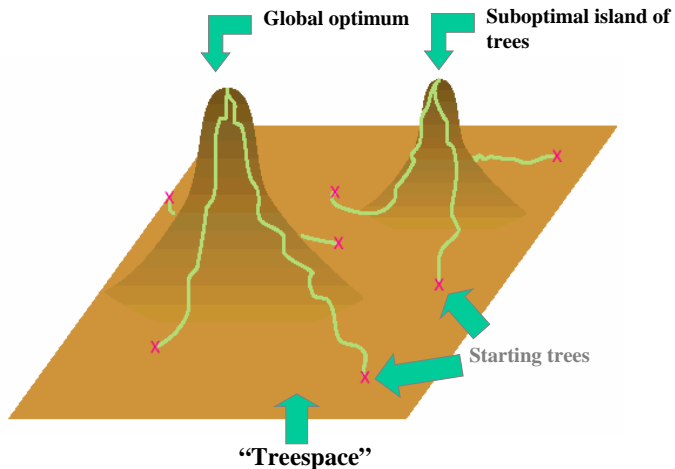
Method	Result	Time	Typical k
Exhaustive search	Optimal solution	$T(k)$	12
Branch and bound	Optimal solution	$\leq T(k)$	18

How do you find a pretty good tree?

3. Heuristic search

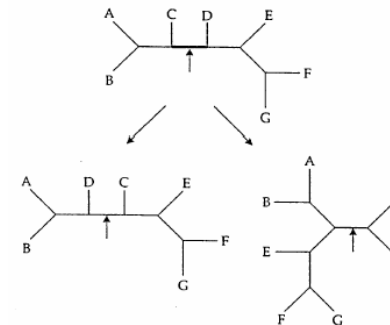
Search for optimal trees by finding good trees and then rearranging them in the hopes of finding an even better tree

Heuristic search



Branch swapping

Nearest-neighbor interchange (NNI)



How do you find the optimal tree?

Method	Result	Time	Typical k
Exhaustive search	Optimal solution	$T(k)$	12
Branch and bound	Optimal solution	$\leq T(k)$	18
Heuristic search	Suboptimal solution	<i>You choose</i>	<i>You choose</i>

Finding the optimal tree

Given k taxa,

- Consider all trees with k leaves
- Score each tree with respect to chosen evolutionary model.
- Select highest scoring tree(s)

Criteria for evaluating which tree best fits the data:

- Maximum parsimony (character data)
- Minimum evolution (distance data)
- Maximum Likelihood (character data)

Multiple Sequence Alignment as Character Data

Each column (or site) is one character.

```

~~~~ALTEKQEALLKQSWEVLKQNIPAHSRLFALIIEAA...
~~~~MALTEKQEALLKQSWEVLKQNIPAHSRLFALIIEAA...
~~~~MALTERQEALLKQSWEVLKQNIGHSRLFALIIEAA...
~~~~~EALLKQSWEVLKQNIGHSCLFALIIEAA...
    
```

	C1	C2	C3	C4
Bees	A	H	S	R
Moths	A	H	S	R
Ants	G	H	S	R
Centipedes	G	H	S	C

Multiple Sequence Alignment as Distance Data

```

Human ~~~~ALTEKQEALLKQSWEVLKQNIPAHSRLFALIIEAA...
Rabbit ~~~~MALTEKQEALLKQSWEVLKQNIPAHSRLFALIIEAA...
Pig ~~~~MALTERQEALLKQSWEVLKQNIGHSRLFALIIEAA...
Chicken ~~~~~EALLKQSWEVLKQNIGHSCLFALIIEAA...
    
```

	Human	Rabbit	Pig	Chicken
Human	0	3	7	9
Rabbit		0	6	8
Pig			0	6
Chicken				0

Finding the optimal tree

Given k taxa,

- Consider all trees with k leaves
- Score each tree with respect to chosen evolutionary model.
- Select highest scoring tree(s)

Criteria for evaluating which tree best fits the data:

- Maximum parsimony (character data)
 - Minimum evolution (distance data)
 - Maximum Likelihood (character data)

Maximum Parsimony

- Parsimony score = the minimum number of changes (mutations) needed to explain data.
- Assumptions
 - Purifying selection dominates
 - Changes are rare
 - No multiple substitutions
 - Sites are independent

Inferring ancestral sequences and computing the parsimony score

Given a tree topology

- Associate characters with leaves of tree
- Find the optimal labeling of internal nodes
- Count mutations

Finding the most parsimonious tree

Given k taxa and n characters (e.g., columns in an MSA),

For each topology, t , with k leaves

$$\text{score}(t) = 0$$

For each character, c /* $1 \leq c \leq n$ */

Find the optimal labeling of internal nodes

$$\text{score}(t) = \text{score}(t) + \text{count_mutations}(c)$$

Return the tree(s) with minimum score.

Determining the parsimony score of a given tree

Input:

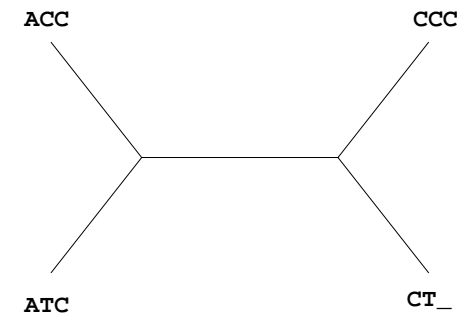
- MSA: k taxa, n columns, aka characters or "sites".
- Tree: T .
- An assignment of the sequences in the MSA to the leaves of T .

Output:

- Score: The minimum number of mutations, over all possible ancestral sequences, required to explain the data
- The ancestral sequences that minimize the score (sometimes.)

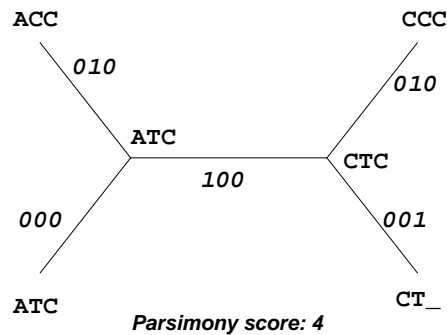
Inferring ancestral sequences and computing the parsimony score

- (1) ACC
- (2) ATC
- (3) CCC
- (4) CT_



Inferring ancestral sequences and computing the parsimony score

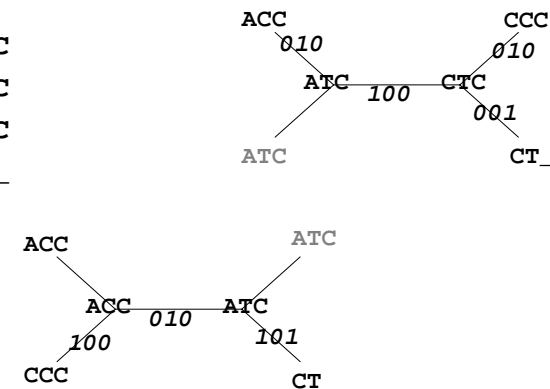
- (1) ACC
- (2) ATC
- (3) CCC
- (4) CT_



Note:

there can be more than one most parsimonious tree

- (1) ACC
- (2) ATC
- (3) CCC
- (4) CT_



Determining the parsimony score of a tree

Fitch's algorithm

- Input: tree, leaf labels
- Output: minimum number of mutations required to explain leaf labels
- *Does not determine the ancestral sequences!*
- Durbin *et al.*, p 175.

Fitch's algorithm

Root tree arbitrarily; Global $C = 0$.

SCORE (i)

- If i is a leaf, return $\{label(i)\}$
- Else
 - $R(l) = \text{SCORE}(\text{left}(i))$
 - $R(r) = \text{SCORE}(\text{right}(i))$
 - If $R(r) \cap R(l) = \emptyset$
 - $R(i) = R(r) \cup R(l)$ // No label avoids mutation
 - $C = C + 1$ // Pass all labels up tree
 - Else
 - $R(i) = R(r) \cap R(l)$ // Choose label that avoids mutation

Final score = C

Some problems with parsimony

- Not all characters are informative
- Data may not be parsimonious
- There may be more than one parsimonious tree

Finding the most parsimonious tree

Given k taxa and n characters (e.g., columns in an MSA),

For each topology, t , with k leaves

$score(t) = 0$

For each of the n characters

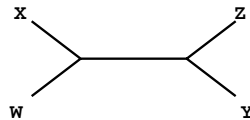
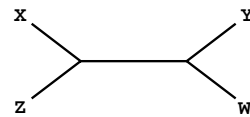
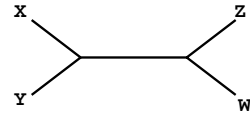
Find the optimal labeling of internal nodes

$score(t) = score(t) + count_mutations$

Not all columns are informative!

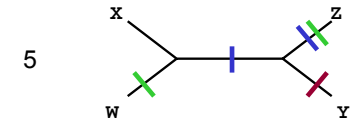
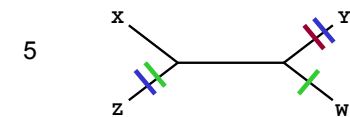
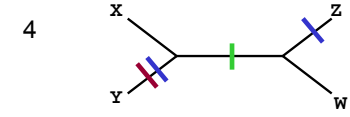
Informative sites:
Columns that distinguish alternate trees

X C A G
Y T G G
Z C C T
W C A T



Informative sites

X C A G
Y T G G
Z C C T
W C A T
1 2 I

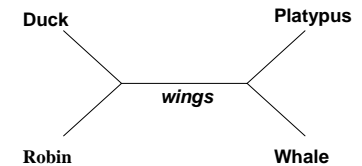


Some problems with parsimony

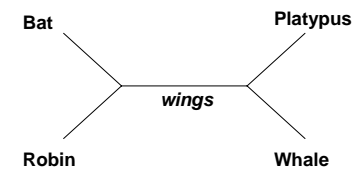
- Not all characters are informative
- Data may not be parsimonious
- There may be more than one parsimonious tree

Problem: Not all characters are parsimonious

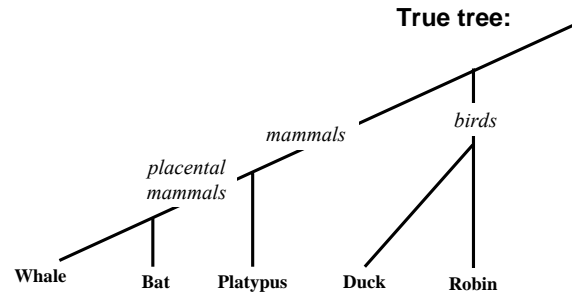
	Wings
Duck	x
Robin	x
Platypus	
Whale	



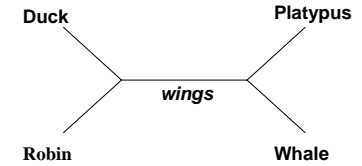
	Wings
Bat	x
Robin	x
Platypus	
Whale	



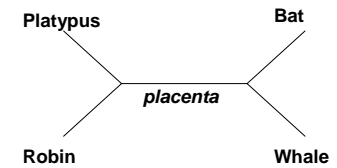
Problem: Not all characters are parsimonious



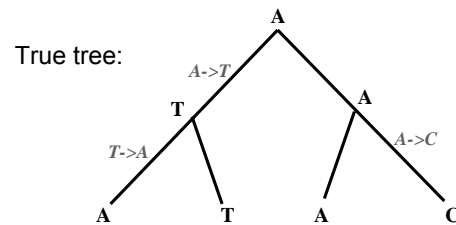
	Wings
Duck	x
Robin	x
Platypos	
Whale	



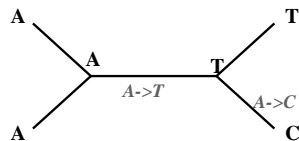
	Placenta
Bat	x
Robin	
Platypos	
Whale	x



If the mutation rate is high,
sequence data is not parsimonious



Most parsimonious, but false, tree:



Some problems with parsimony

- Not all characters are informative
- Data may not be parsimonious
- There may be more than one parsimonious tree

Note:

there can be more than one most parsimonious tree

- (1) ACC
- (2) ATC
- (3) CCC
- (4) CT_

