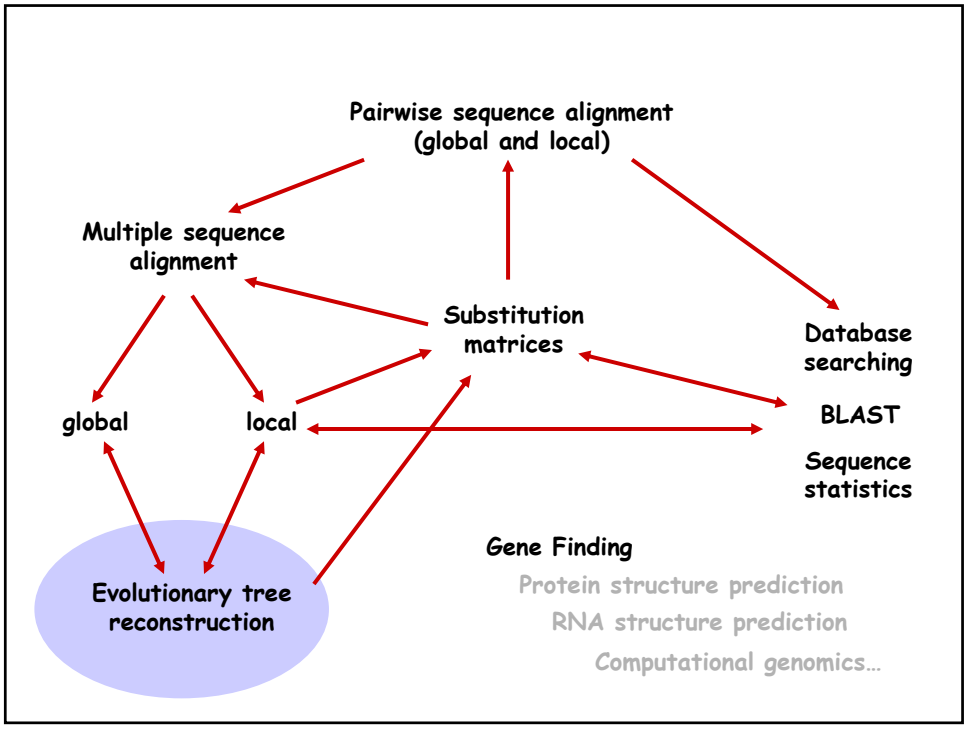
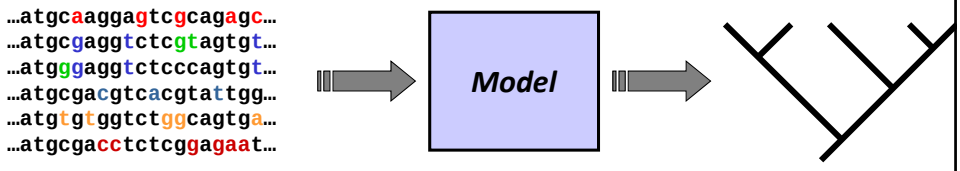


Phylogeny reconstruction



Given

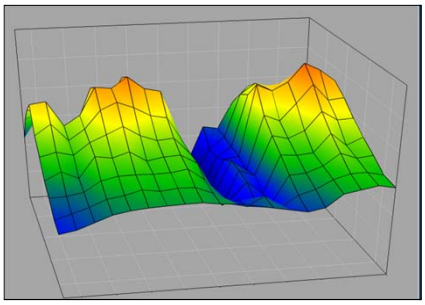
- Multiple sequence alignment
- Model of sequence evolution

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 optimization criterion.
- Select the optimal tree(s)



Tree 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.

Categories of tree reconstruction methods

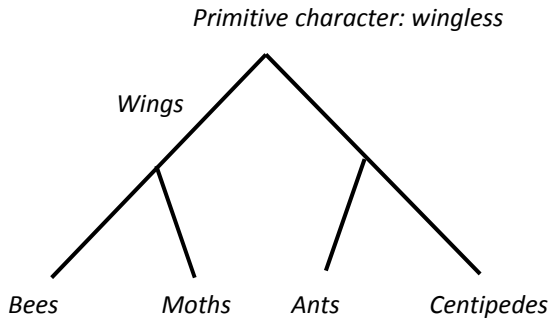
	Parsimony	Distance	Maximum likelihood estimation	Bayesian methods
Character data	**x**		**x**	**x**
Pairwise distances		**x**		

Character data

- A **character** is a well-defined feature that in a taxonomic unit can assume one out of **two or more** mutually exclusive **character states**.
- Character: variable
  - e.g., Height, Weight, Color
- State: values that character can take on
  - e.g., 2.39m, 14.7kg, red…

Character data

Bees	Wings
Moths	Wings
Ants	No wings
Centipedes	No wings



Multiple Sequence Alignment as Character Data

```
~~~~ALTEKQEALSWEVLKQNIPAHSRLFALIIEAA…
~~~MALTEKQEALSWEVLKQNIPAHSRLFALILEAA…
~~~MALTERQEALSWEVLKQNIPGHSRLFALIIEAA…
~~~~~~~~~~EALSWEVLKQNIPGHSCLFALIIEAA…
```

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

Each column (or site) is one character.

- DNA: 4 states
- Amino acids: 20 states

Other molecular features

- e.g., gain and loss of introns

Distance data

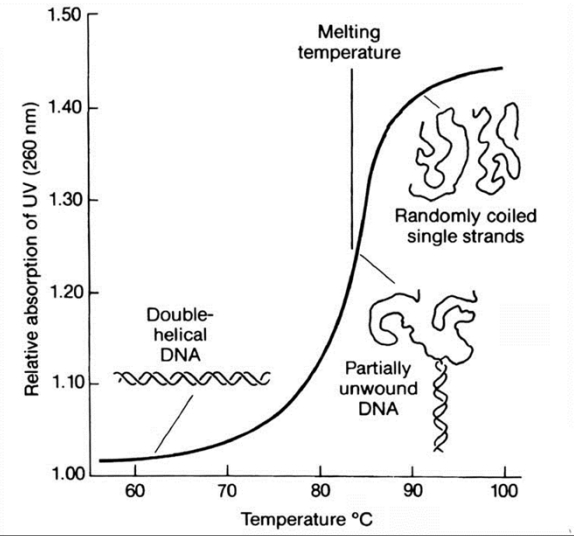
- Pairwise distances between taxa with the usual geometric properties
- Most molecular data yield character states that are subsequently converted into distances.

CHARACTERS ⟶ DISTANCE

DISTANCE ⟶/ CHARACTERS

10

Some molecular data can only be expressed as distances.



Calculating distances from MSAs

~~~~ALTEKQEALLKQSWEVLKQNIPAHSLRLFALIIEAA…
~~~MALTEKQEALLKQSWEVLKQNIPAHSLRLFALILEAA…
~~~MALTERQEALLKQSWEVLKQNIPGHSLRLFALIIEAA…
~~~~~~~~~~EALLKQSWEVLKQNIPGHSLCLFALIIEAA…

- Use the pairwise alignment between *taxon i* and *taxon j induced by the MSA.*
- Assess the amino acid changes
- Correct for multiple substitutions

	Rabbit	Pig	Chicken
Human	4	5	8
Rabbit	0	5	11
Pig		0	11

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)

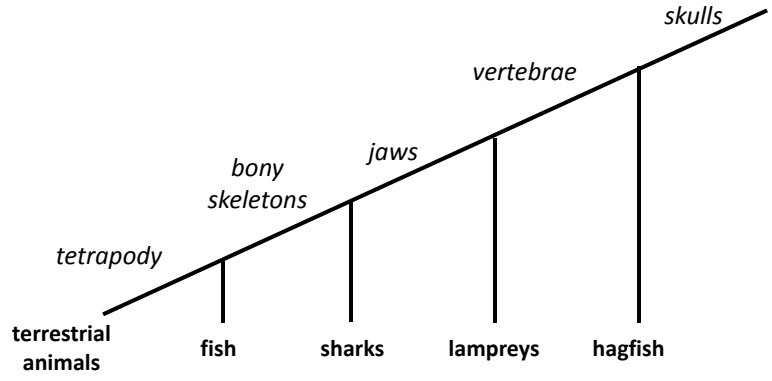
Categories of tree reconstruction methods

	Parsimony	Distance	Maximum likelihood estimation	Bayesian methods
Character data	x		x	x
Pairwise distances		x		

Maximum Parsimony: Nature is thrifty

The best tree requires the fewest mutations.

e.g., jaws were only "invented" once



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

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 character, c /* 1≤c≤n */

Find the optimal labeling of internal nodes

score(t) = score(t) + count_mutations(c)

Return the tree(s) with minimum score.

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

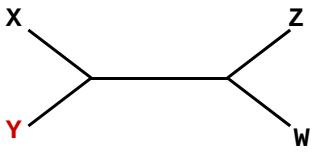
For each character, c /* 1≤c≤n */

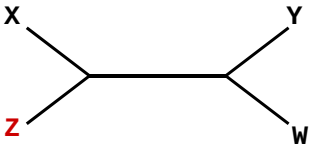
Find the optimal labeling of internal nodes

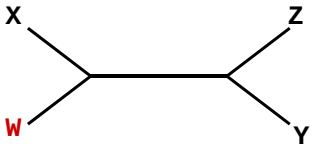
score(t) = score(t) + count_mutations(c)

Return the tree(s) with minimum score.

Trees with four leaves







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 character, c /* 1≤c≤n */

Find the optimal labeling of internal nodes

score(t) = score(t) + count_mutations(c)

Return the tree(s) with minimum score.

Determining the parsimony score of a *given* tree

Input:

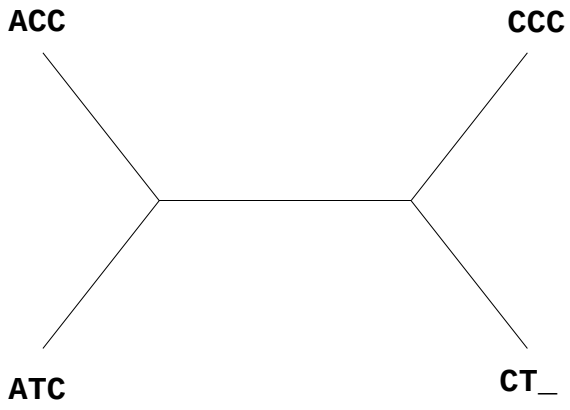
- MSA: *k* taxa, *n* columns, aka characters or "sites".
- Tree: *T.*
- An assignment of the sequences in the MSA to the leaves of *T.*

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.)

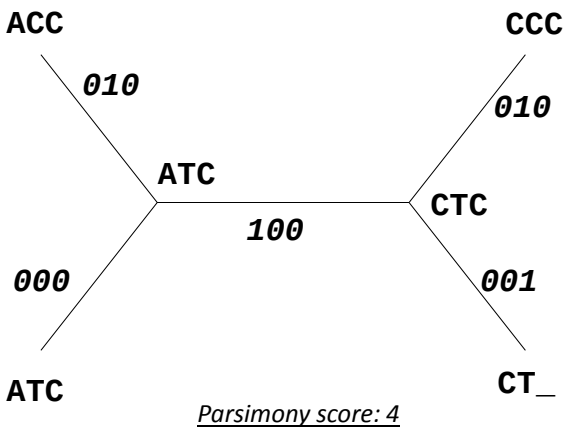
Inferring ancestral sequences and computing the parsimony score

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



Inferring ancestral sequences and computing the parsimony score

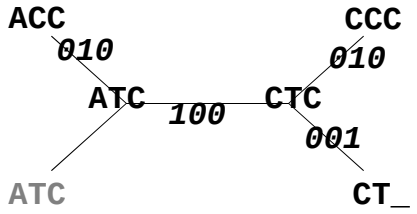
(1) **ACC**
(2) **ATC**
(3) **CCC**
(4) **CT_**

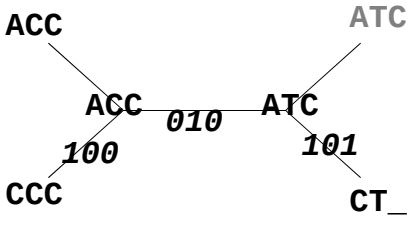


Parsimony score: 4

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)* = SCORE (left(*i*))
  - *R(r)* = SCORE (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

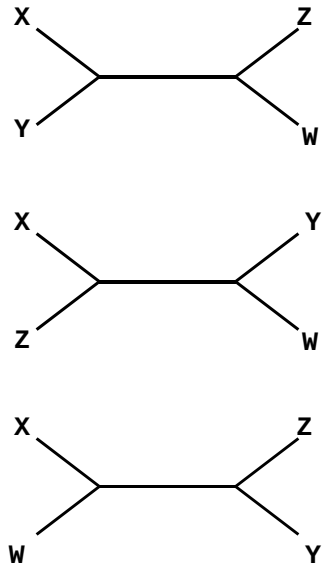
Some problems with parsimony

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

Informative sites:

Columns that distinguish alternate trees

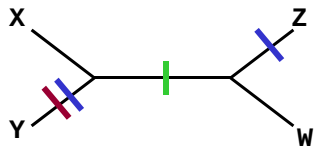
X	**C**	**A**	**G**
Y	**T**	**G**	**G**
Z	**C**	**C**	**T**
W	**C**	**A**	**T**



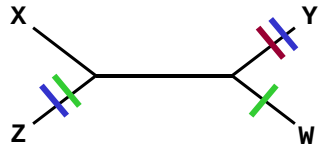
Informative sites

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

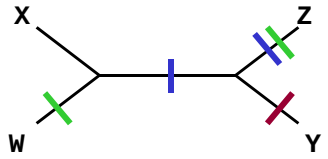
4



5



5



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

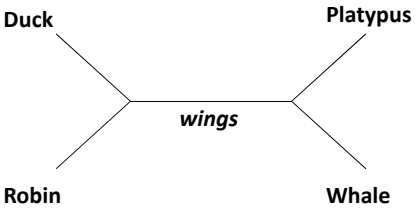
Not all columns are informative!

Some problems with parsimony

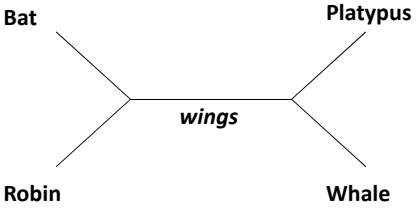
- Not all characters are informative
- Data may not be parsimonious
- There may be more than one parsimonious tree

Problem: Not all characters are parsimonious

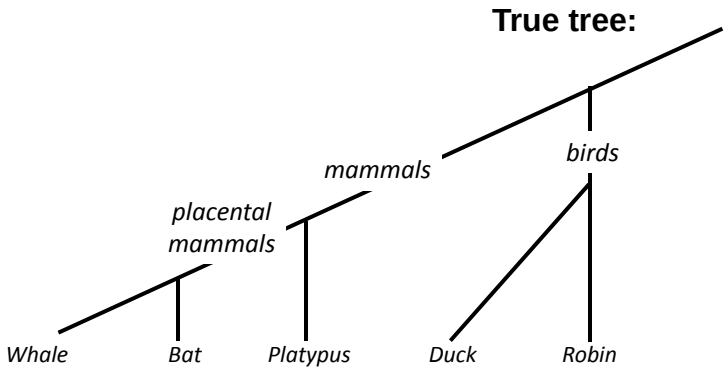
	Wings
Duck	x
Robin	x
Platypus	
Whale	



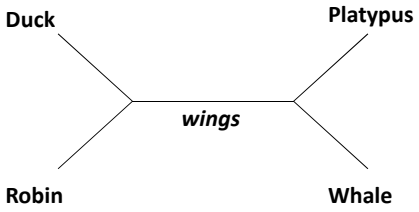
	Wings
Bat	x
Robin	x
Platypus	
Whale	



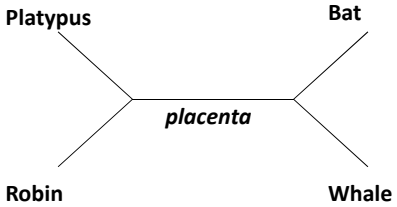
Problem: Not all characters are parsimonious



	Wings
Duck	x
Robin	x
Platypus	
Whale	

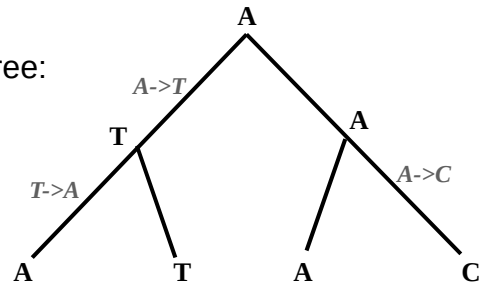


	Placenta
Bat	x
Robin	
Platypus	
Whale	x

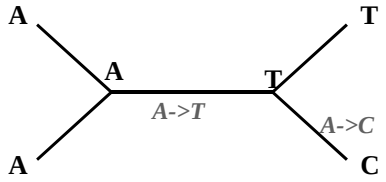


If the mutation rate is high,
sequence data is not parsimonious

True tree:



Most parsimonious, but false, tree:

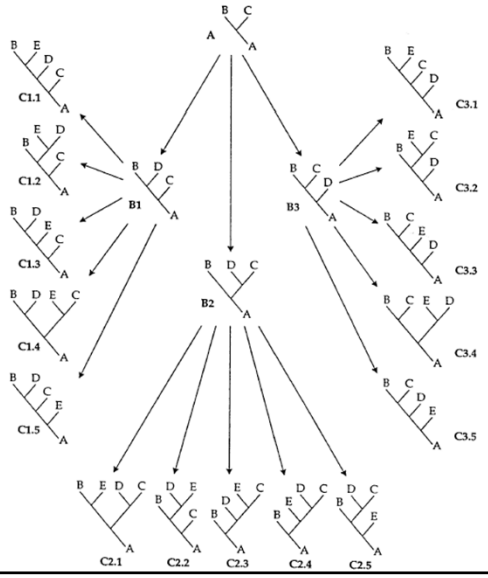


Some problems with parsimony

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

How do you find the optimal tree?

1. Exhaustive search (<12 taxa)



(Phylogeny reconstruction is NP-complete.)

How do you find the optimal tree?

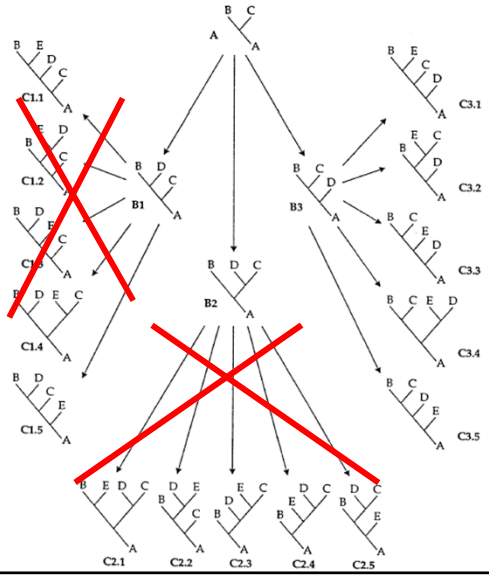
Method	Result	Time	Typical k
Exhaustive search	Optimal solution	*T(k)*	12

How do you find the optimal tree?

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

   Score is non-decreasing as you add edges

L = {T_3}, *C = infinity*
For i = 3 to k {
  For each tree, *t,* in L{
    If *Score(t) > C, skip*
    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?

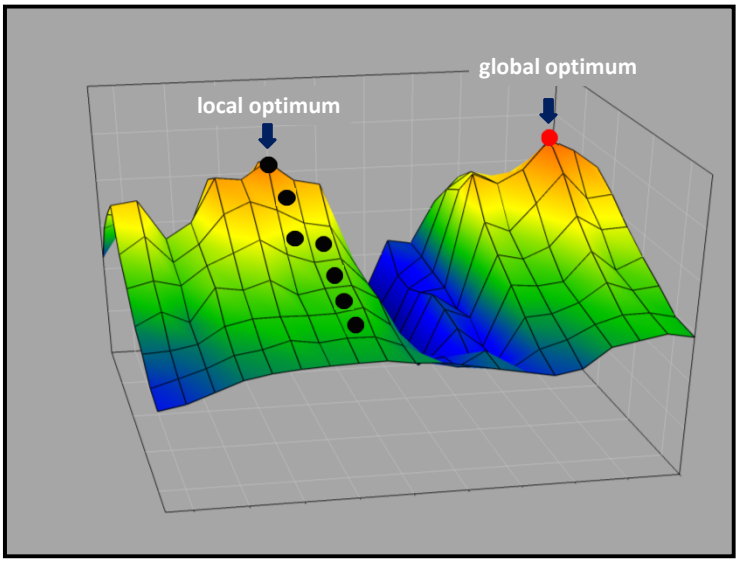
Method	Result	Time	Typical k
Exhaustive search	Optimal solution	*T(k)*	12
Branch and bound	Optimal solution	*≤T(k)*	18

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

Phylogeny reconstruction uses heuristic search



How do you find the optimal tree?

Method	Result	Time	Typical k
Exhaustive search	Optimal solution	*T(k)*	12
Branch and bound	Optimal solution	*≤T(k)*	18
Heuristic search	Suboptimal solution	*You choose*	*You choose*