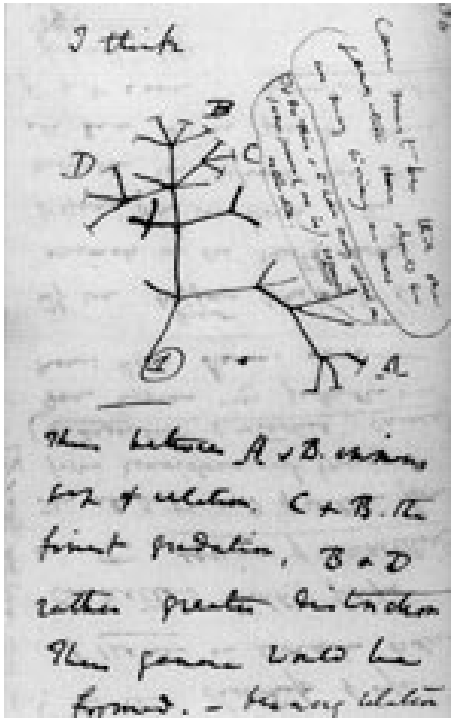
Finding Optimal Phylogenetic Trees

Katherine St. John

City University of New York
American Museum of Natural History

23 June 2015

Outline



Charles Darwin, 1837

- Treespaces and Landscapes
- Metrics & Search
- Preprocessing to Improve Search
- Maximum Likelihood & Continuous Treespace
- When Trees are Not Enough....

Analogy: Find the Highest Point

polymaps.org

Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.

Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.
- Find height at each point.

Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.
- Find height at each point.
- Output the sampled point with largest height.

Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.
- Find height at each point.
- Output the sampled point with largest height.
- Will you reach the highest point?

Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.
- Find height at each point.
- Output the sampled point with largest height.
- Will you reach the highest point?
- Only if very lucky or a very dense sample.

Analogy: Find the Highest Point

Hill Climbing:

- Start at the harbor.

Analogy: Find the Highest Point

Hill Climbing:

- Start at the harbor.
- Can see 25 meters in all directions.

Analogy: Find the Highest Point

Hill Climbing:

- Start at the harbor.
- Can see 25 meters in all directions.
- Walk upwards, repeat.

Analogy: Find the Highest Point

Hill Climbing:

- Start at the harbor.
- Can see 25 meters in all directions.
- Walk upwards, repeat.
- Will you reach the highest point?

Analogy: Find the Highest Point

Hill Climbing:

- Start at the harbor.
- Can see 25 meters in all directions.
- Walk upwards, repeat.
- Will you reach the highest point?
- Maybe, but maybe not.

Analogy: Find the Highest Point

Hill Climbing:

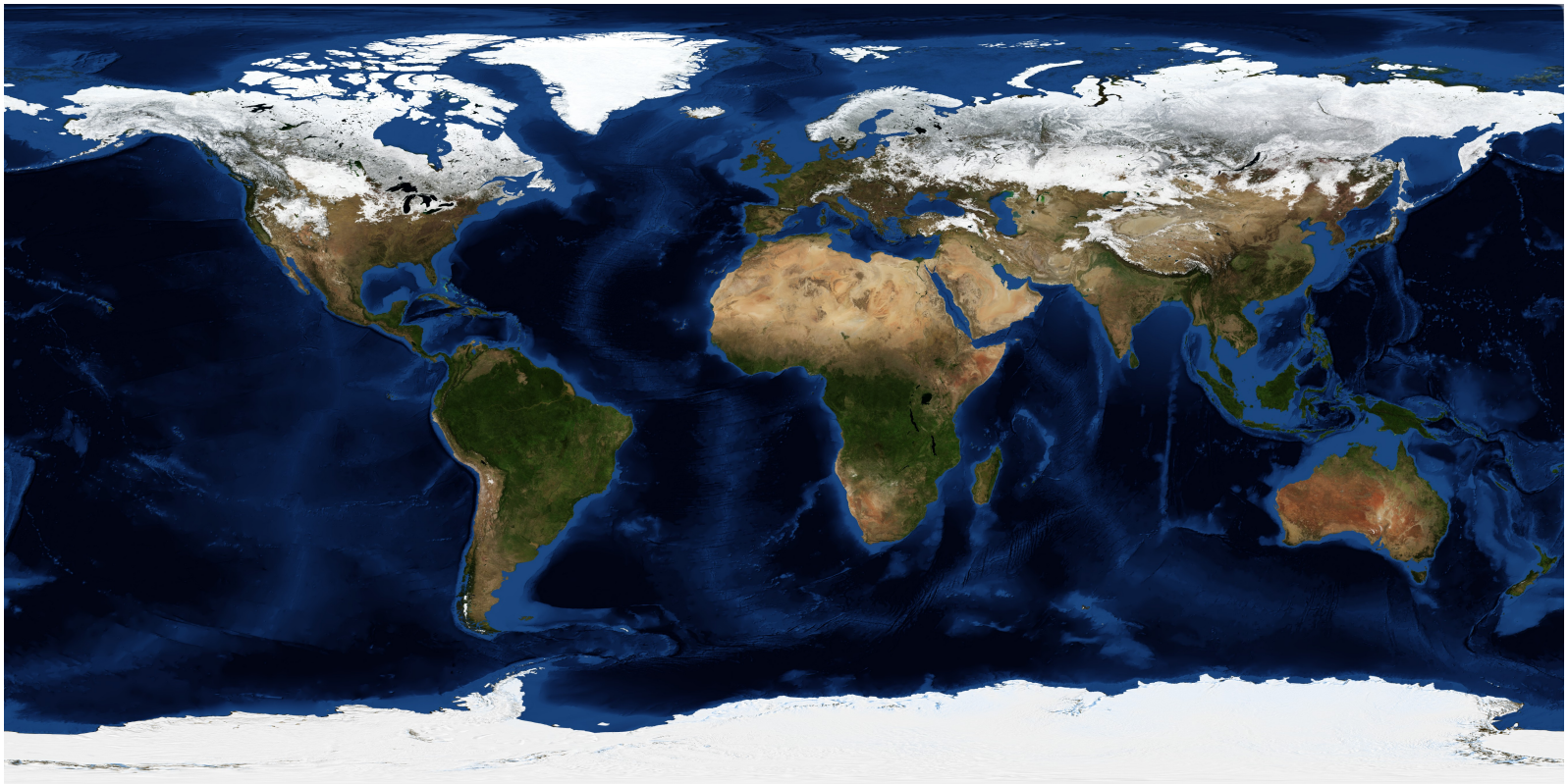
- Start at the harbor.
- Can see 25 meters in all directions.
- Walk upwards, repeat.
- Will you reach the highest point?
- Maybe, but maybe not.
  - Could reach small peaks, but miss the larger ones.

Analogy: Find the Highest Point

Hill Climbing:

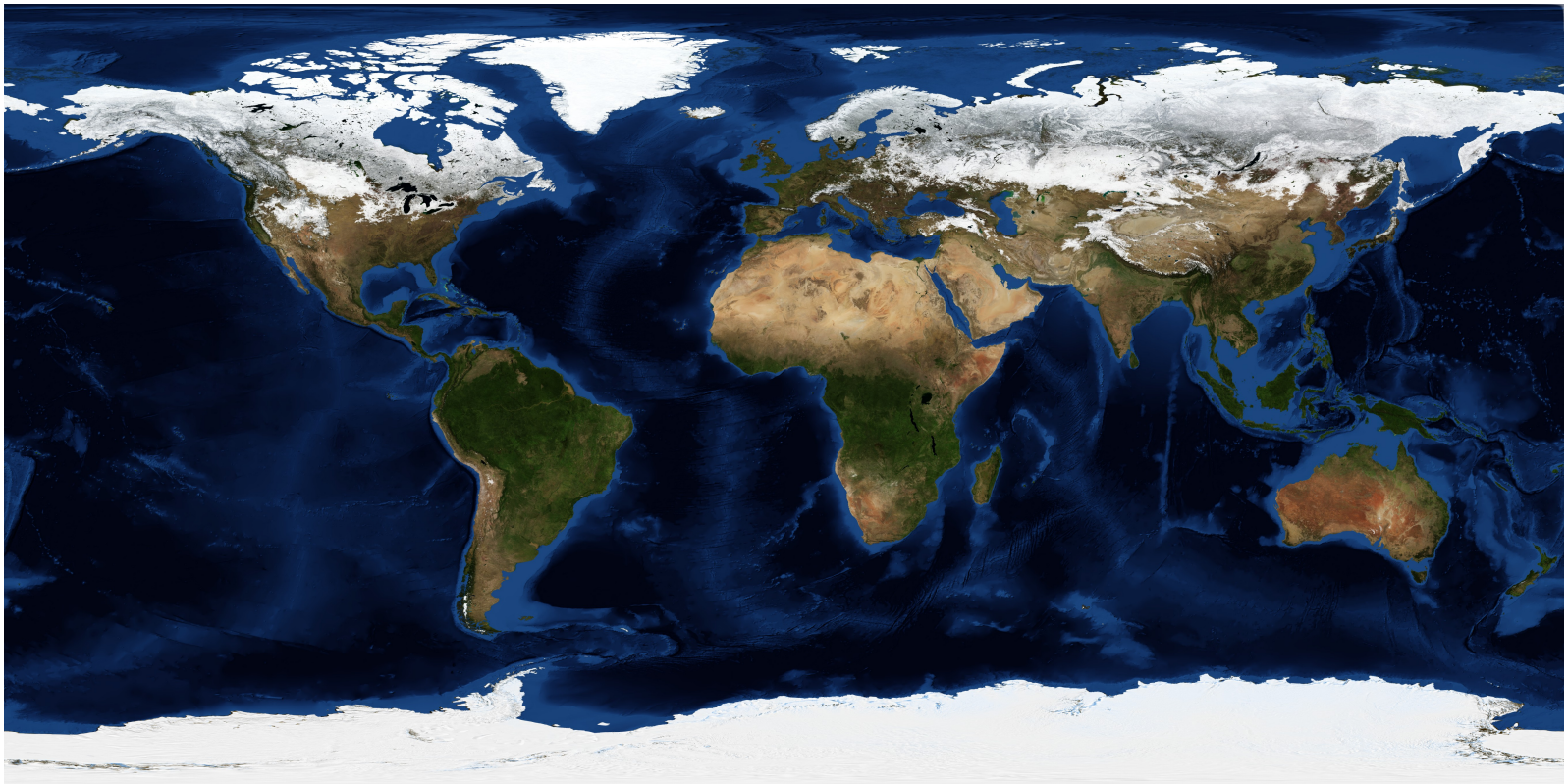
- Start at the harbor.
- Can see 25 meters in all directions.
- Walk upwards, repeat.
- Will you reach the highest point?
- Maybe, but maybe not.
  - Could reach small peaks, but miss the larger ones.
  - Start in multiple places to see more.

Analogy: Find the Highest Point



NASA Blue Marble

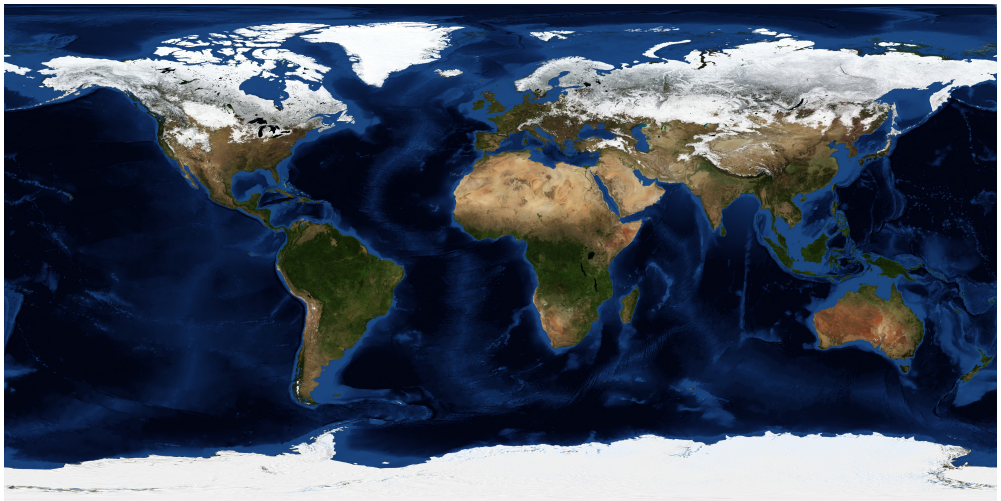
Analogy: Find the Highest Point



NASA Blue Marble

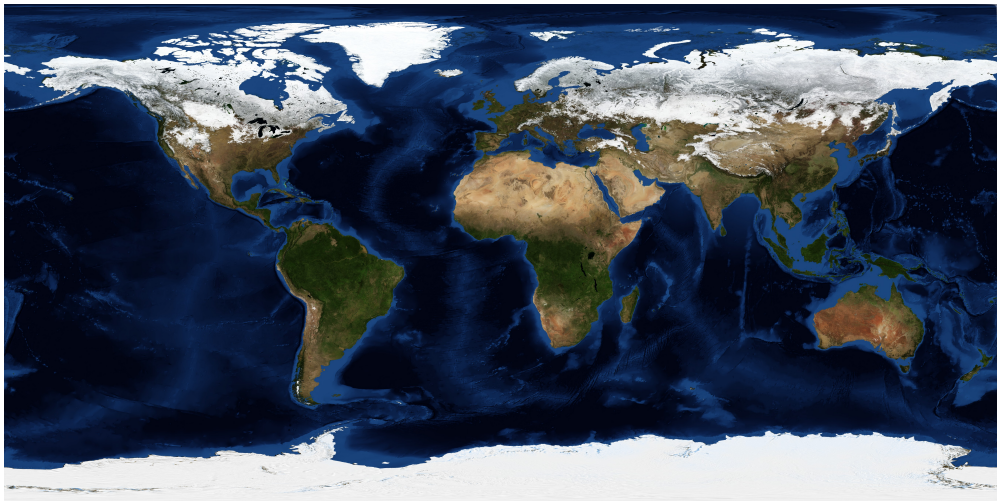
Sampling only on the island misses peaks elsewhere.

Local Search Techniques



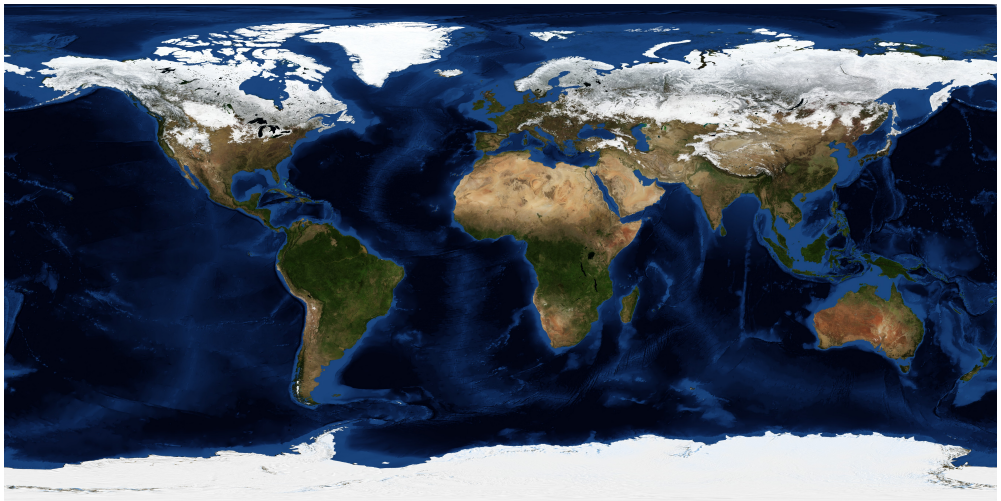
- **Goal:** Find the point with the optimal score

Local Search Techniques



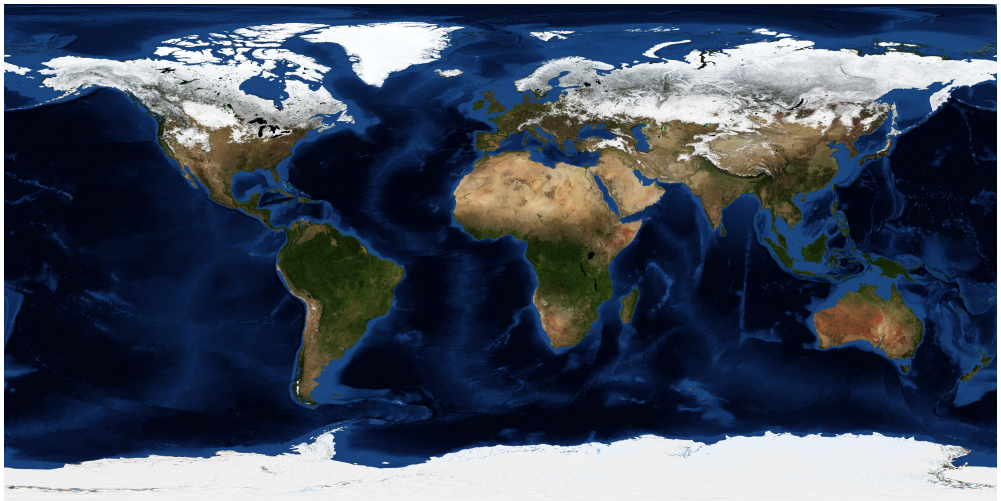
- **Goal:** Find the point with the optimal score
- Local search techniques prevail:

Local Search Techniques



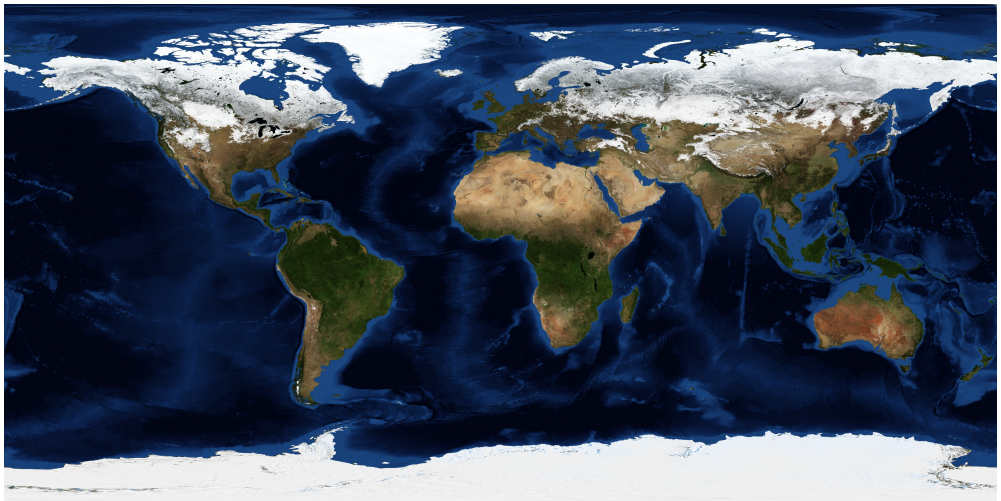
- **Goal:** Find the point with the optimal score
- Local search techniques prevail:
  - Begin with a point

Local Search Techniques



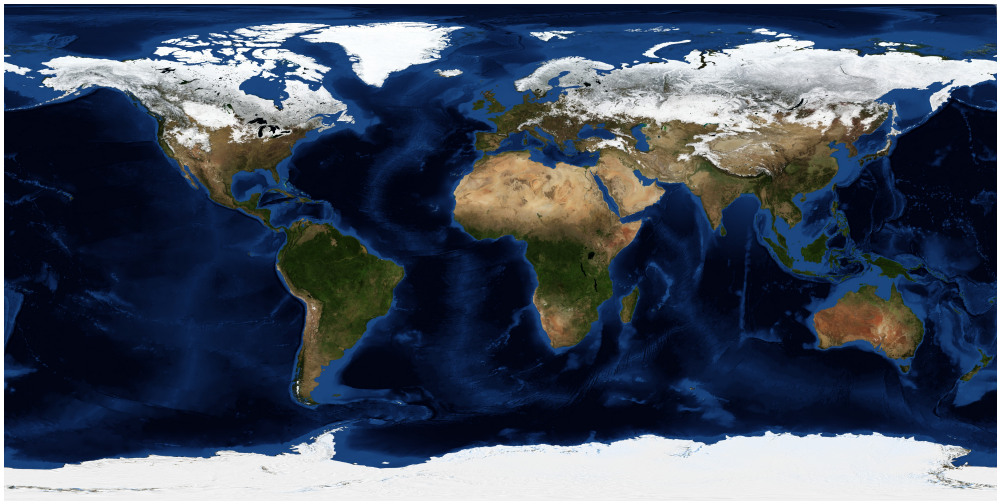
- **Goal:** Find the point with the optimal score
- Local search techniques prevail:
  - Begin with a point
  - Choose the next point from its neighbors (e.g. best scoring)

Local Search Techniques



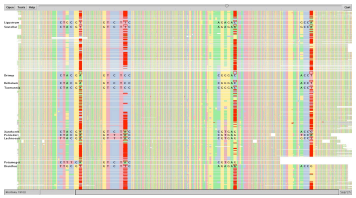
- **Goal:** Find the point with the optimal score
- Local search techniques prevail:
  - Begin with a point
  - Choose the next point from its neighbors (e.g. best scoring)
  - Repeat

Local Search Techniques

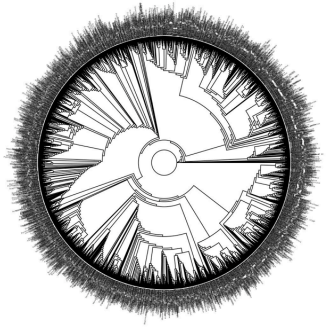


- **Goal:** Find the point with the optimal score
- Local search techniques prevail:
  - Begin with a point
  - Choose the next point from its neighbors (e.g. best scoring)
  - Repeat
- Many variations on the theme: branch-and-bound, MCMC, genetic algorithms,...

Goal: Find Optimal Evolutionary History



rBCL sequences



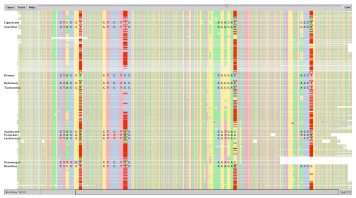
Hillis Lab

Input: Sequences of k Characters on n taxa

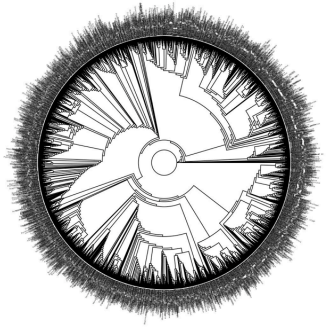
Output: Evolutionary History (Tree) on n leaves

Optimality Criteria : Two popular ones, both NP-hard.

Goal: Find Optimal Evolutionary History



rBCL sequences



Hillis Lab

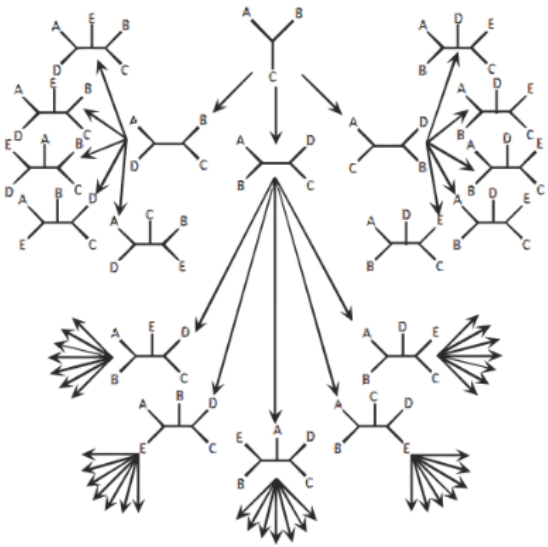
Input: Sequences of k Characters on n taxa

Output: Evolutionary History (Tree) on n leaves

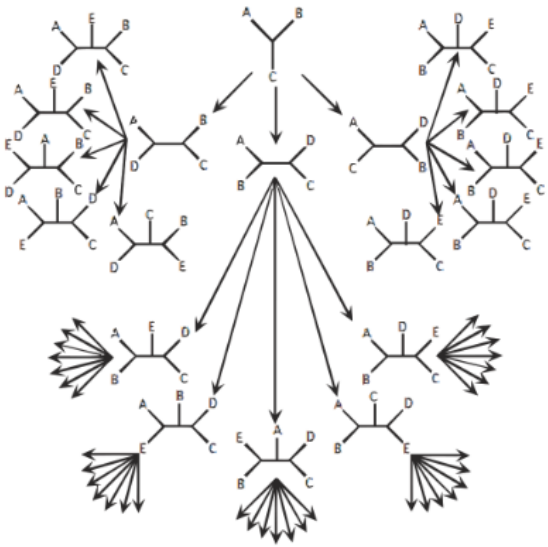
Optimality Criteria : Two popular ones, both NP-hard.

Underlying assumption: *Evolution is tree-like.*

How Many Phylogenetic Trees?



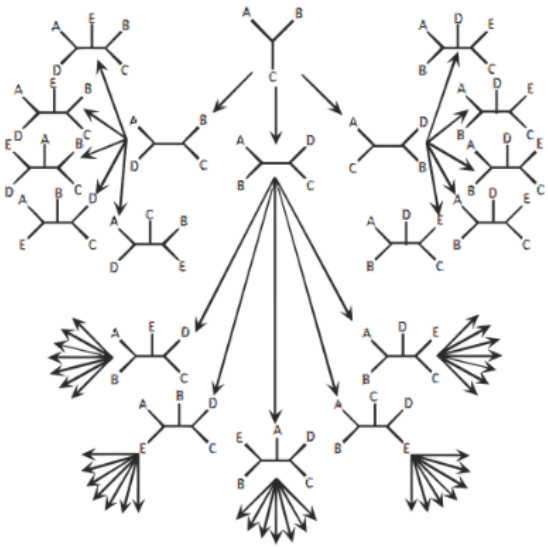
How Many Phylogenetic Trees?



Schröder, 1870 (see Semple & Steel, 2003):

$$
\begin{aligned}
\#\text{ of trees} &= 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-5) \\
&= (2n-5)!! \\
&\sim \frac{1}{\sqrt{\Pi}} 2^{n-2} n! n^{\frac{-5}{2}} \\
&\sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-2}
\end{aligned}
$$

How Many Phylogenetic Trees?



Schröder, 1870 (see Semple & Steel, 2003):

$$\begin{aligned} \#\text{ of trees} &= 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-5) \\ &= (2n-5)!! \\ &\sim \frac{1}{\sqrt{\Pi}} 2^{n-2} n! n^{\frac{-5}{2}} \\ &\sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-2} \end{aligned}$$

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)

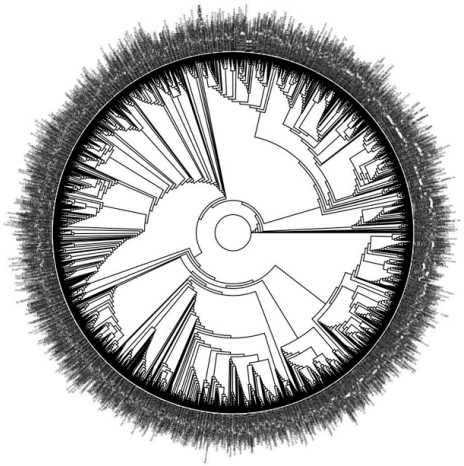
How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)

How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)

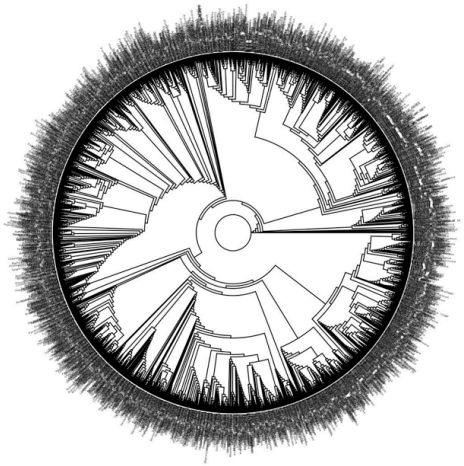
How many taxa?



David Hillis, 2002

How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)



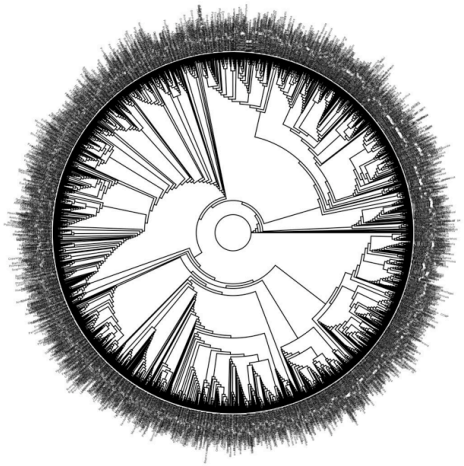
David Hillis, 2002

How many taxa?

- classifying species: *n ranges from dozens to thousands (think beetles!)*

How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)



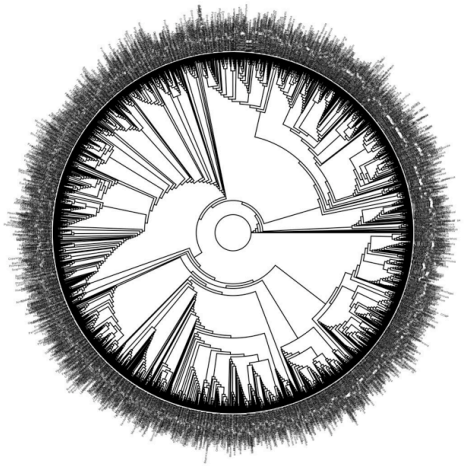
David Hillis, 2002

How many taxa?

- classifying species: *n ranges from dozens to thousands (think beetles!)*
- building the "Tree of Life": *$n \sim$ million species*

How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)



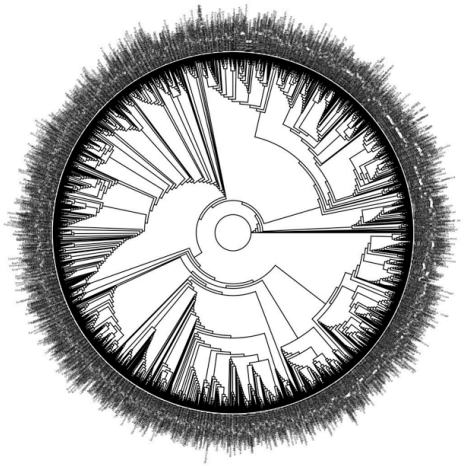
David Hillis, 2002

How many taxa?

- classifying species: *n ranges from dozens to thousands (think beetles!)*
- building the "Tree of Life": *$n \sim$ million species*
- designing the flu vaccine and other drugs: *$n \sim$ hundreds of isolates*

How Many Trees?

(For $n \geq 50$, $\exists$ more possible tree topologies than there are atoms in the universe.)

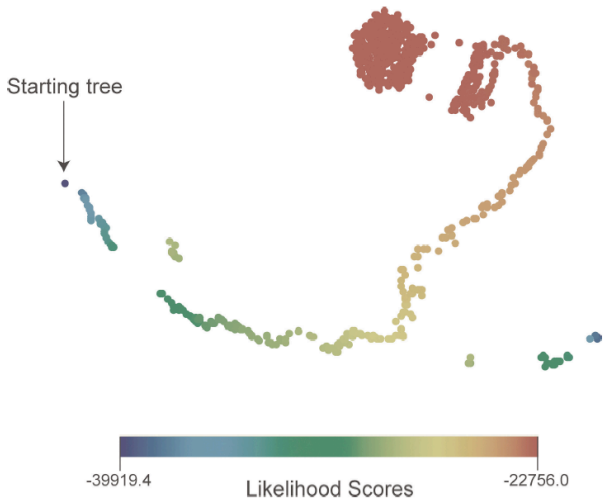


David Hillis, 2002

How many taxa?

- classifying species: *n ranges from dozens to thousands (think beetles!)*
- building the "Tree of Life": *$n \sim$ million species*
- designing the flu vaccine and other drugs: *$n \sim$ hundreds of isolates*
- determining the origins of HIV infection: *$n \sim$ thousands of strains*

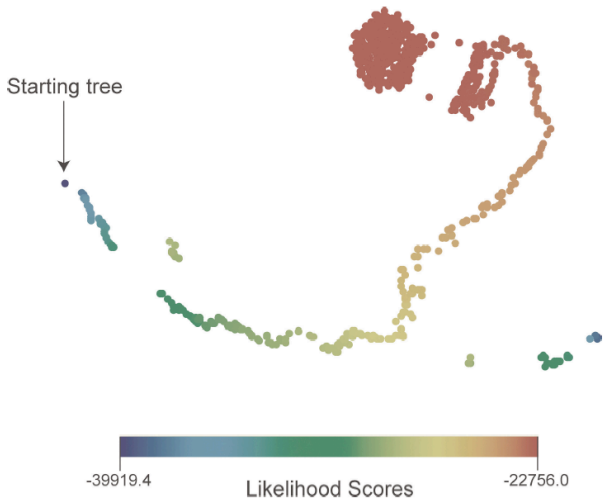
Searching for Optimal Trees



Hillis, Heath, S, 2005

- **Goal:** Find the tree with the optimal score

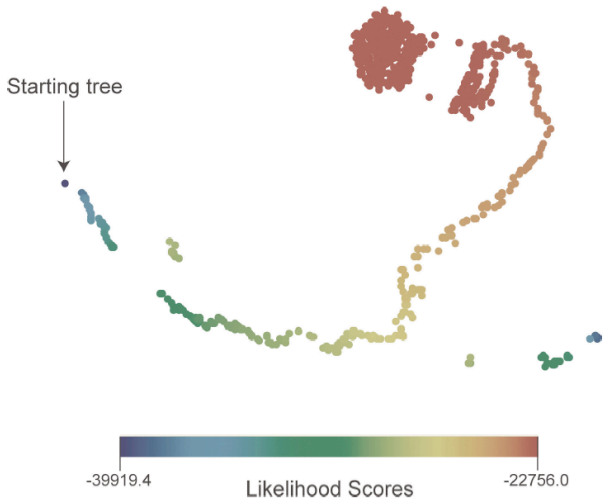
Searching for Optimal Trees



Hillis, Heath, S, 2005

- **Goal:** Find the tree with the optimal score
- Local search techniques prevail:

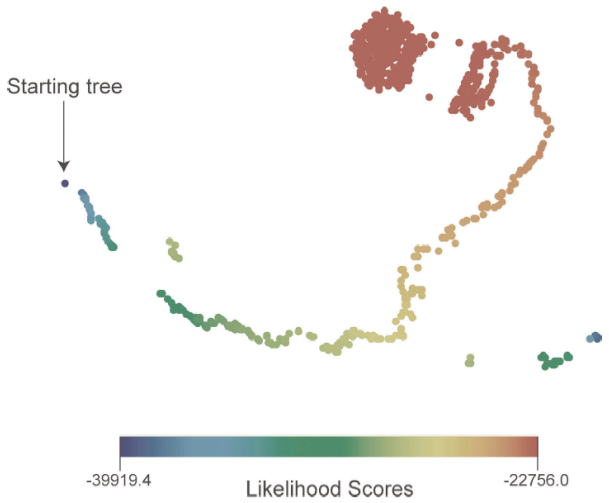
Searching for Optimal Trees



Hillis, Heath, S, 2005

- **Goal:** Find the tree with the optimal score
- Local search techniques prevail:
  - Begin with a tree

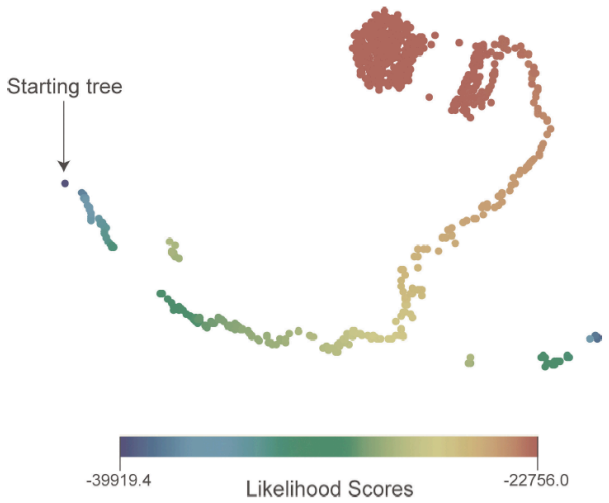
Searching for Optimal Trees



Hillis, Heath, S, 2005

- **Goal:** Find the tree with the optimal score
- Local search techniques prevail:
  - Begin with a tree
  - Choose the next tree from its neighbor (e.g. best scoring)

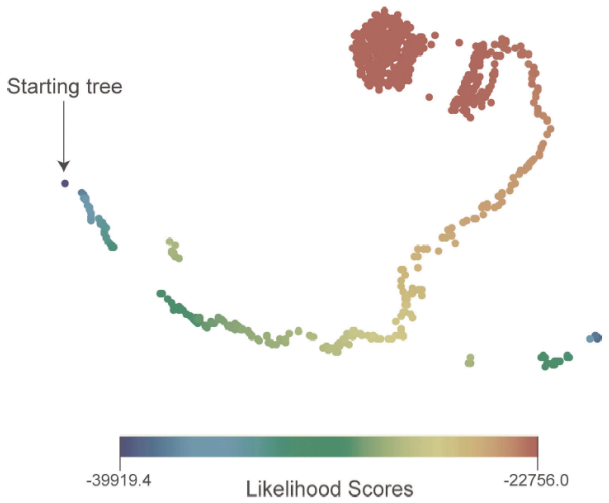
Searching for Optimal Trees



Hillis, Heath, S, 2005

- **Goal:** Find the tree with the optimal score
- Local search techniques prevail:
  - Begin with a tree
  - Choose the next tree from its neighbor (e.g. best scoring)
  - Repeat

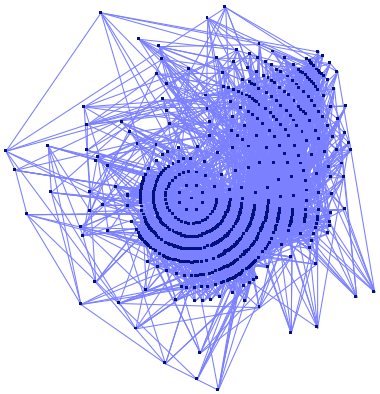
Searching for Optimal Trees



Hillis, Heath, S, 2005

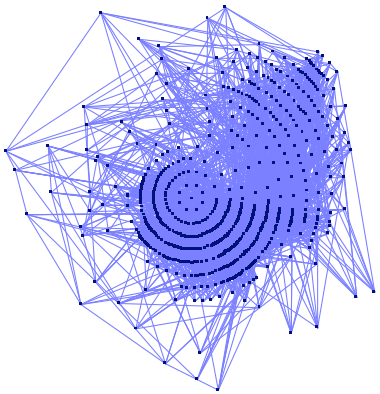
- **Goal:** Find the tree with the optimal score
- Local search techniques prevail:
  - Begin with a tree
  - Choose the next tree from its neighbor (e.g. best scoring)
  - Repeat
- Many variations on the theme: branch-and-bound, MCMC, genetic algorithms,...

Optimality Criteria



Given a set of organisms, which tree is optimal?

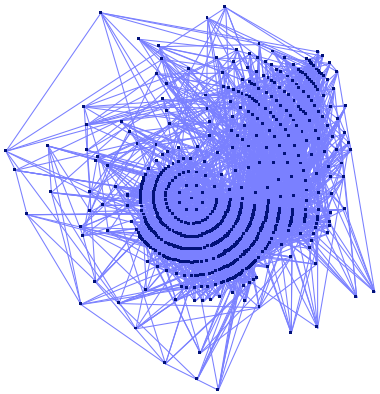
Optimality Criteria



Given a set of organisms, which tree is optimal?

- Two standard criteria for optimality:

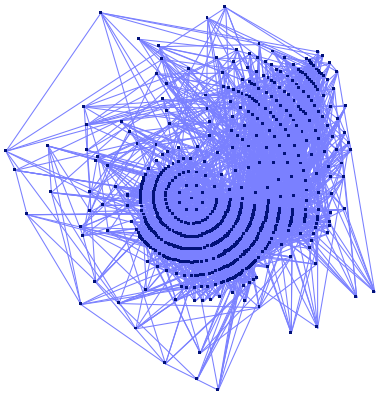
Optimality Criteria



Given a set of organisms, which tree is optimal?

- Two standard criteria for optimality:
  - Maximum Parsimony: find tree with fewest changes

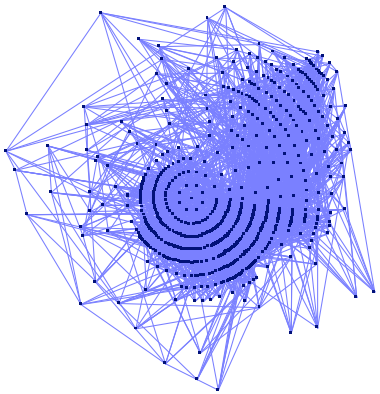
Optimality Criteria



Given a set of organisms, which tree is optimal?

- Two standard criteria for optimality:
  - Maximum Parsimony: find tree with fewest changes
  - Maximum Likelihood: find most likely tree (with respect to a model of evolution)

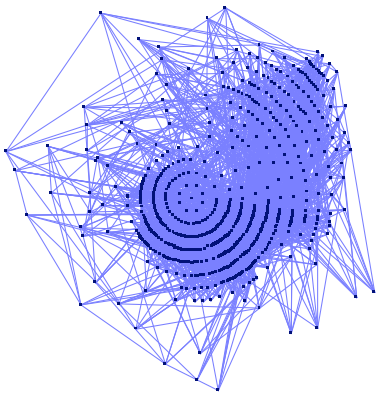
Optimality Criteria



Given a set of organisms, which tree is optimal?

- Two standard criteria for optimality:
  - Maximum Parsimony: find tree with fewest changes
  - Maximum Likelihood: find most likely tree (with respect to a model of evolution)

Which Tree is Optimal?



Given a set of organisms, which tree is optimal?

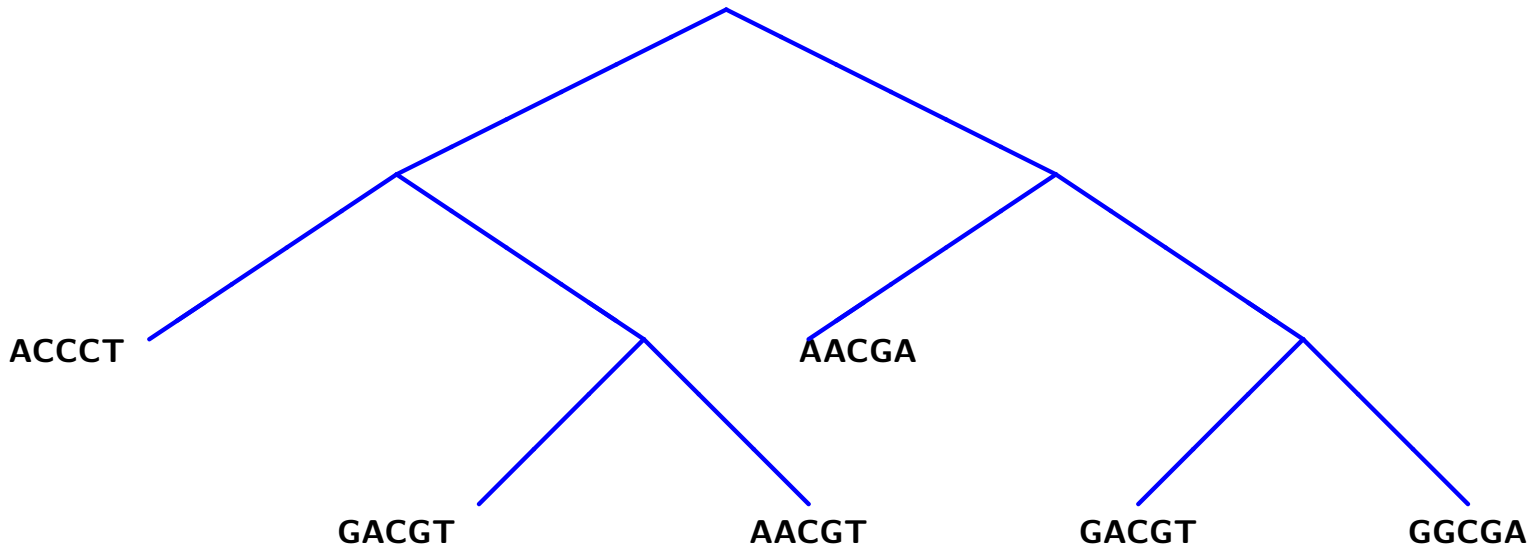
- Two standard criteria for optimality:
  - Maximum Parsimony: find tree with fewest changes. (NP-hard, Foulds & Graham, 1982).
  - Maximum Likelihood: find most likely tree (with respect to a model of evolution) (NP-hard, Roch, 2008).

Maximum Parsimony

- Find the tree that can explain the observed sequences with a minimal number of substitutions.

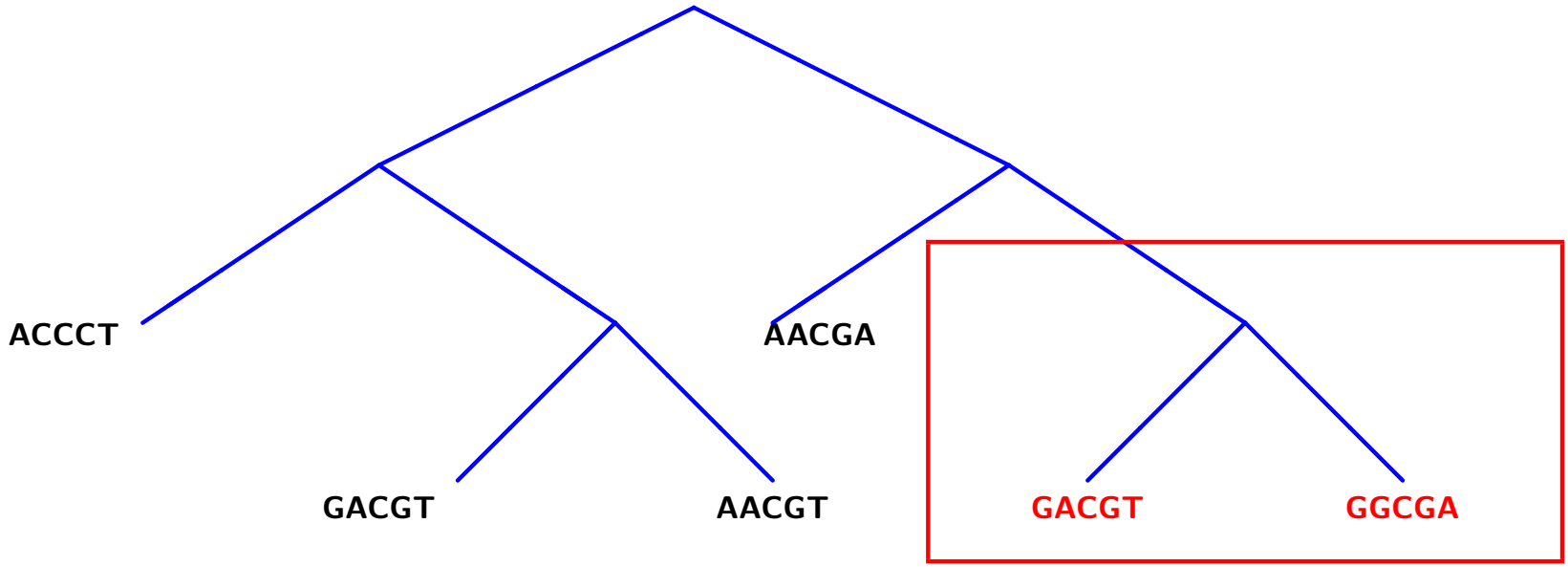
Maximum Parsimony

- Find the tree that can explain the observed sequences with a minimal number of substitutions.
- Given sequences for leaves and a tree, first measure "minimal number of substitutions."



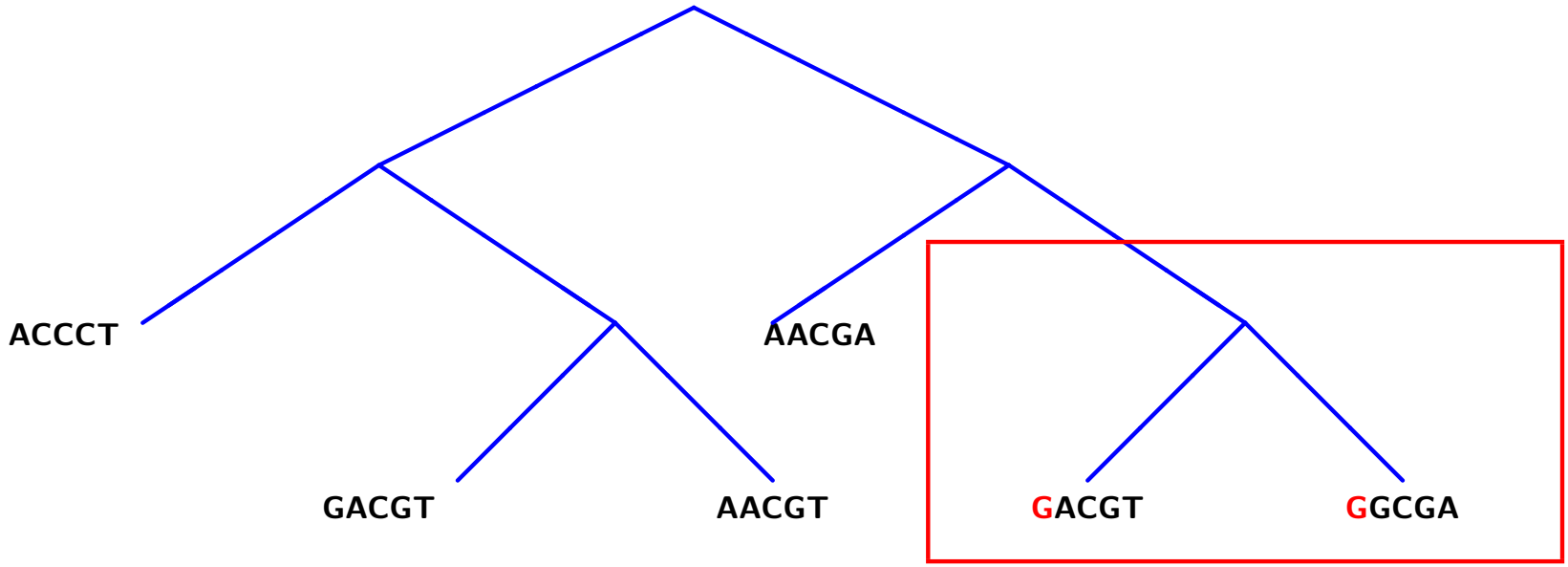
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



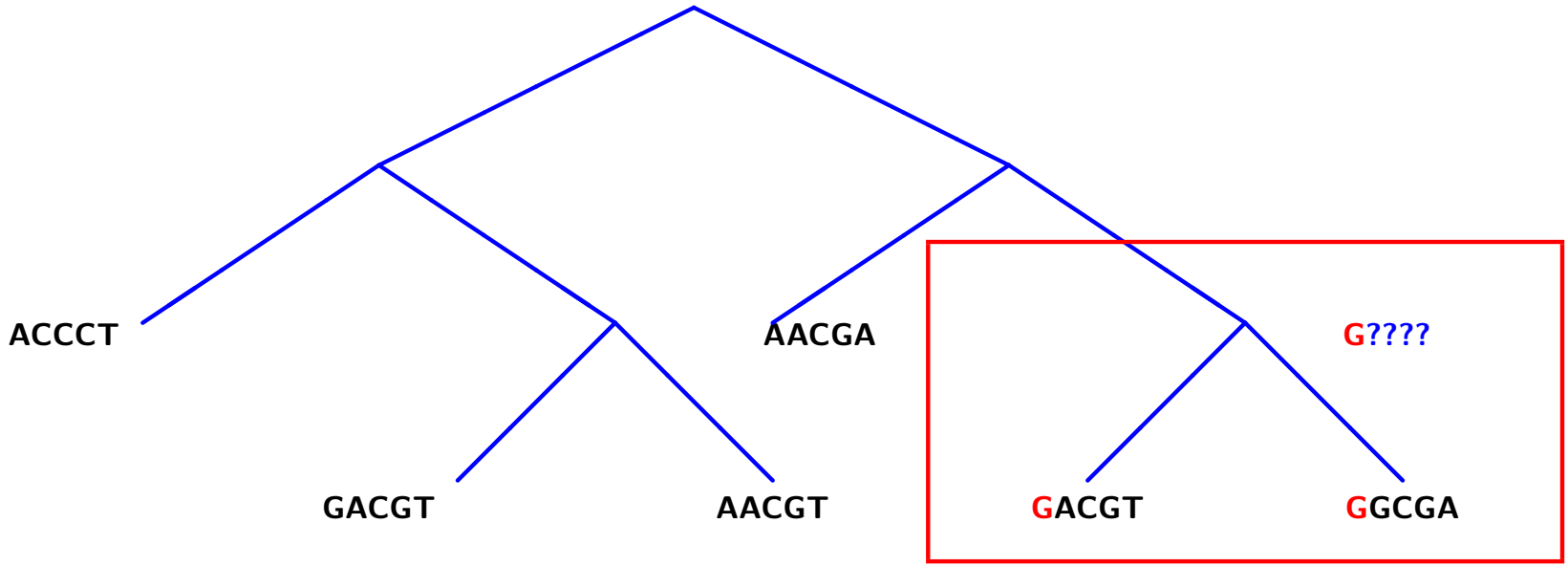
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



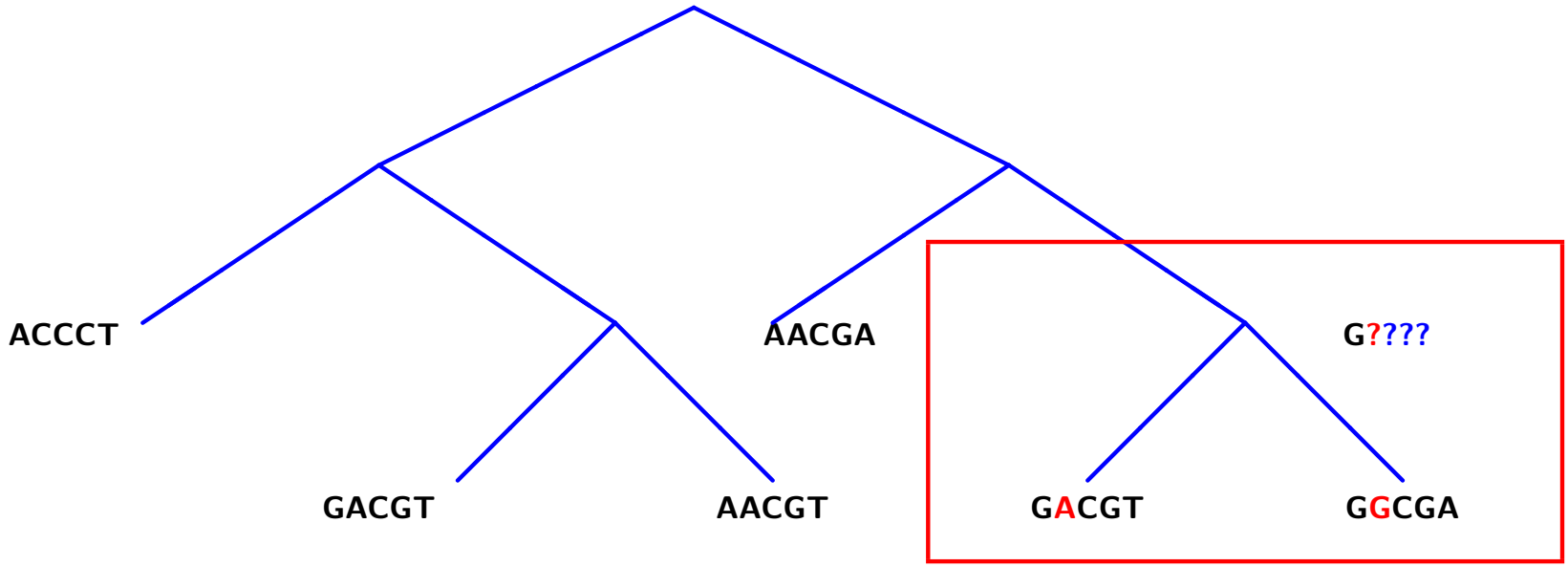
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



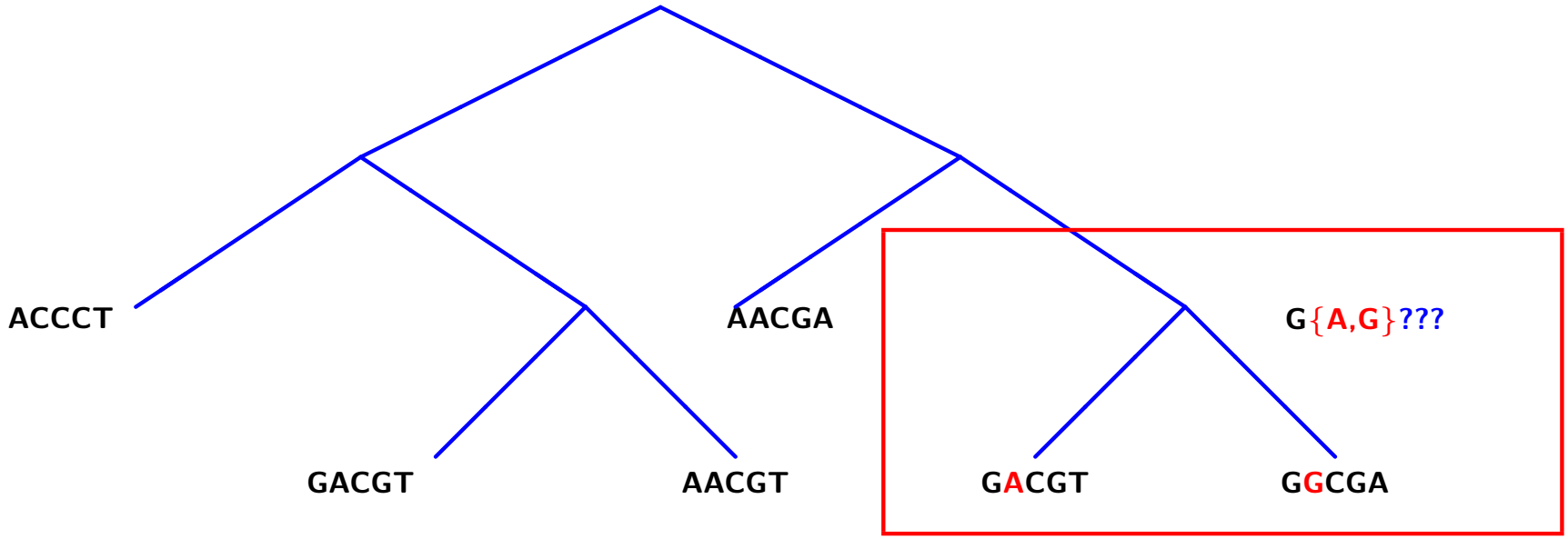
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



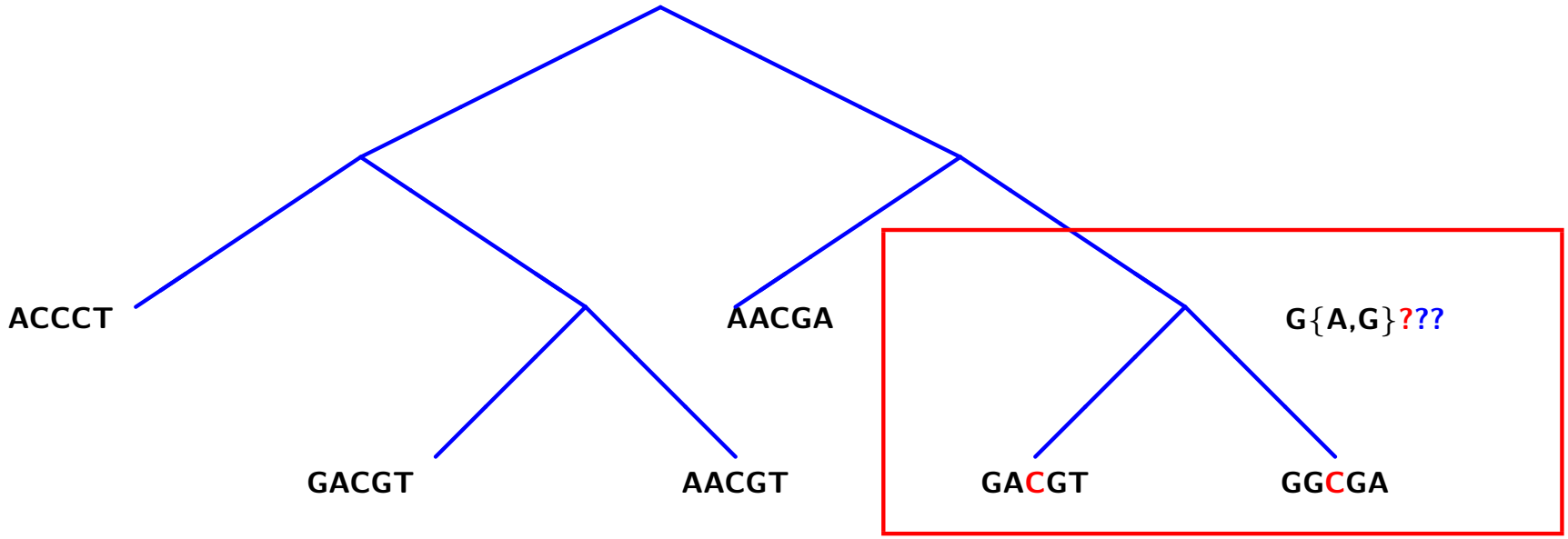
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



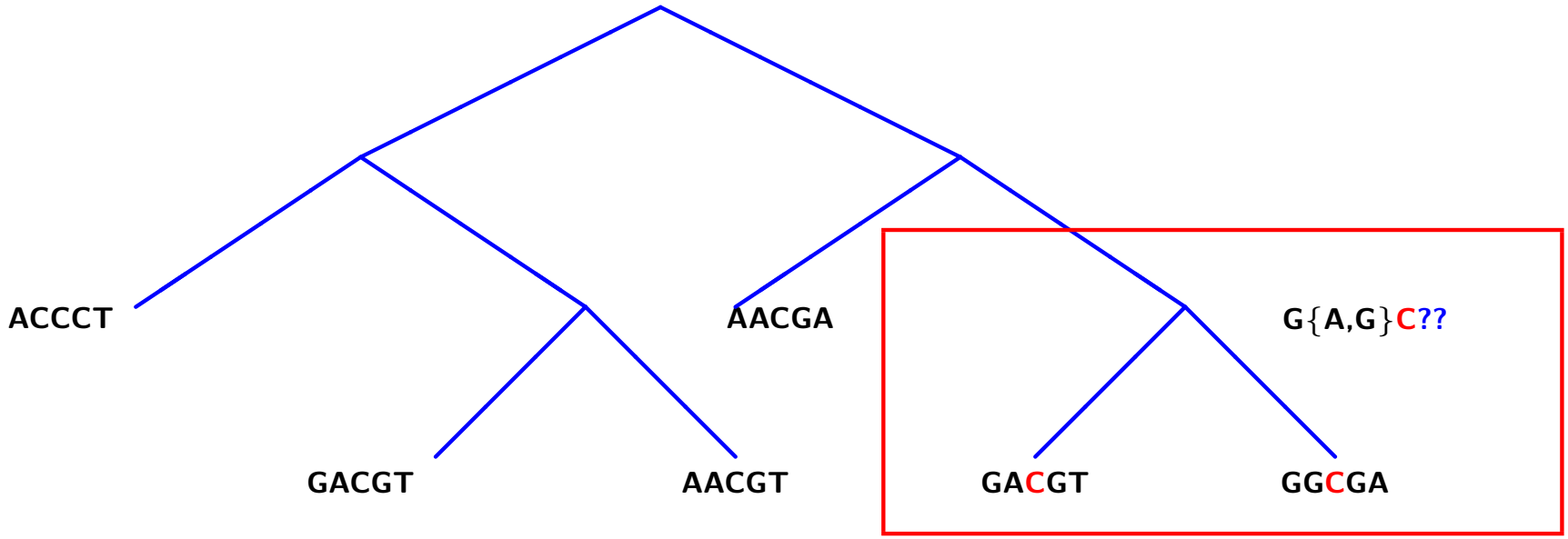
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



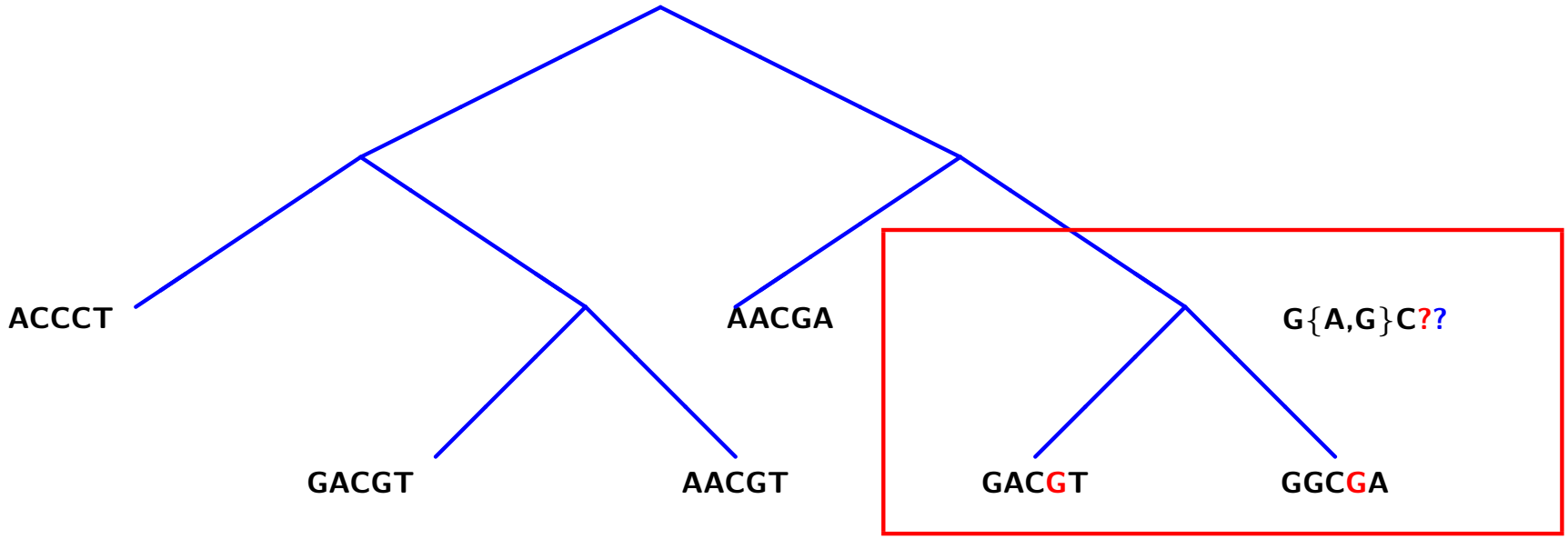
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



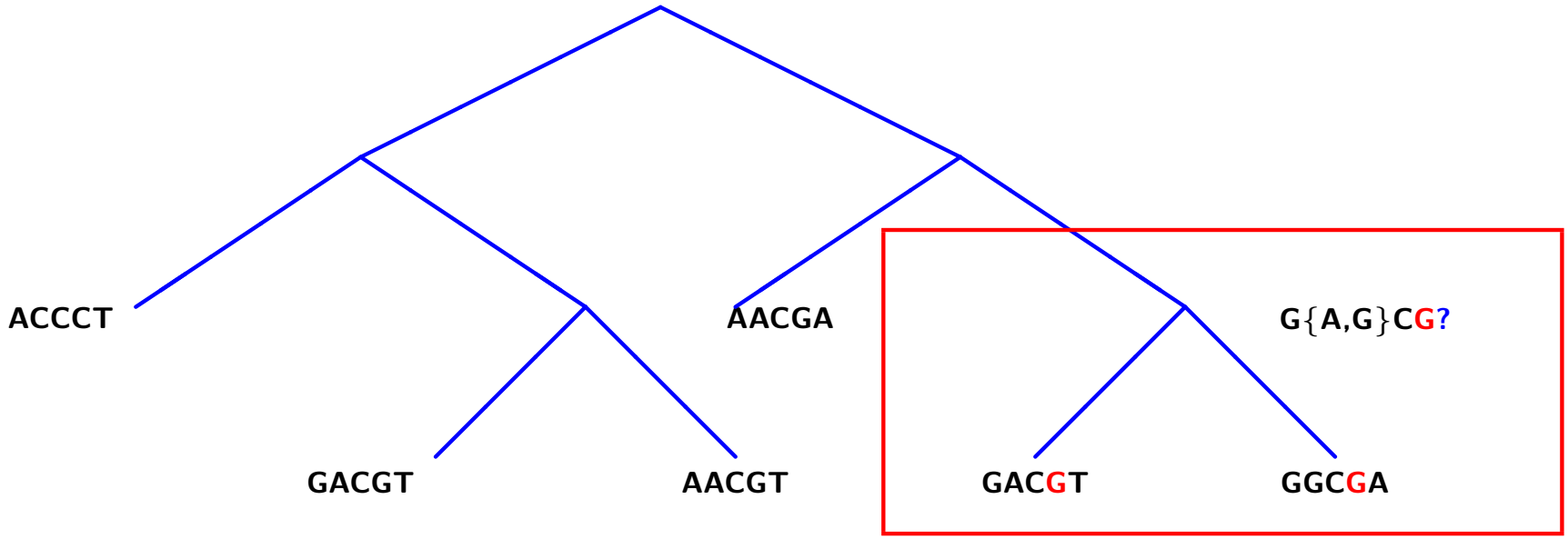
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



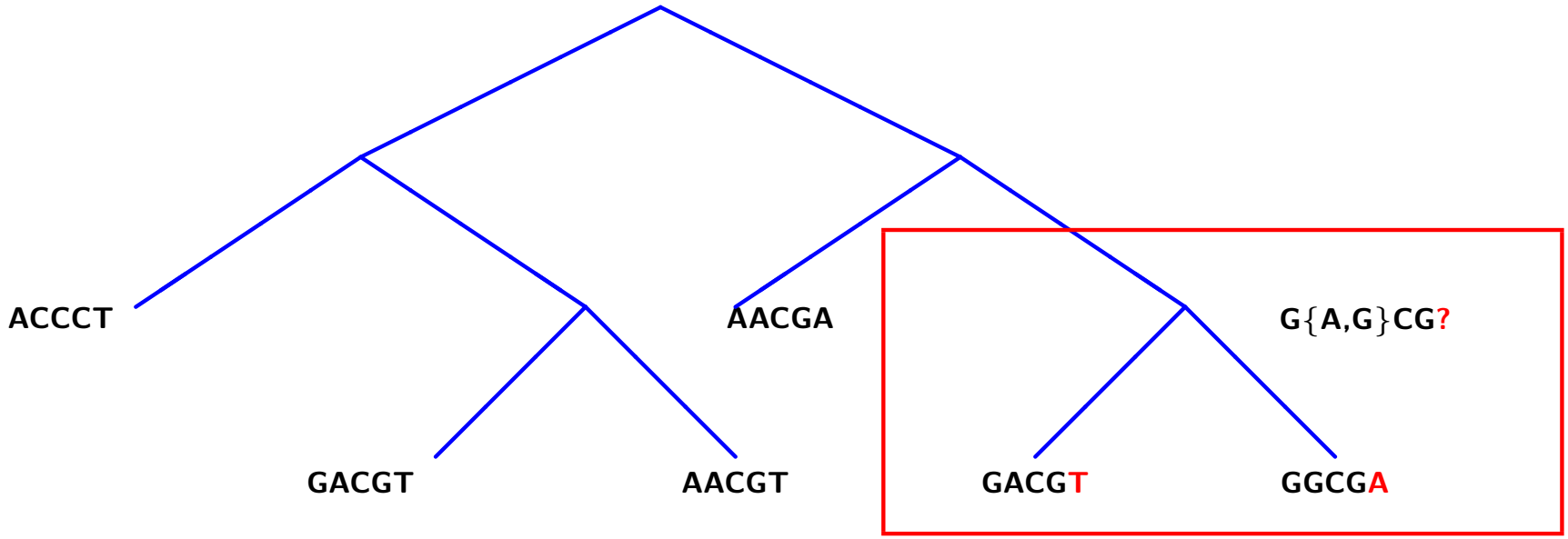
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



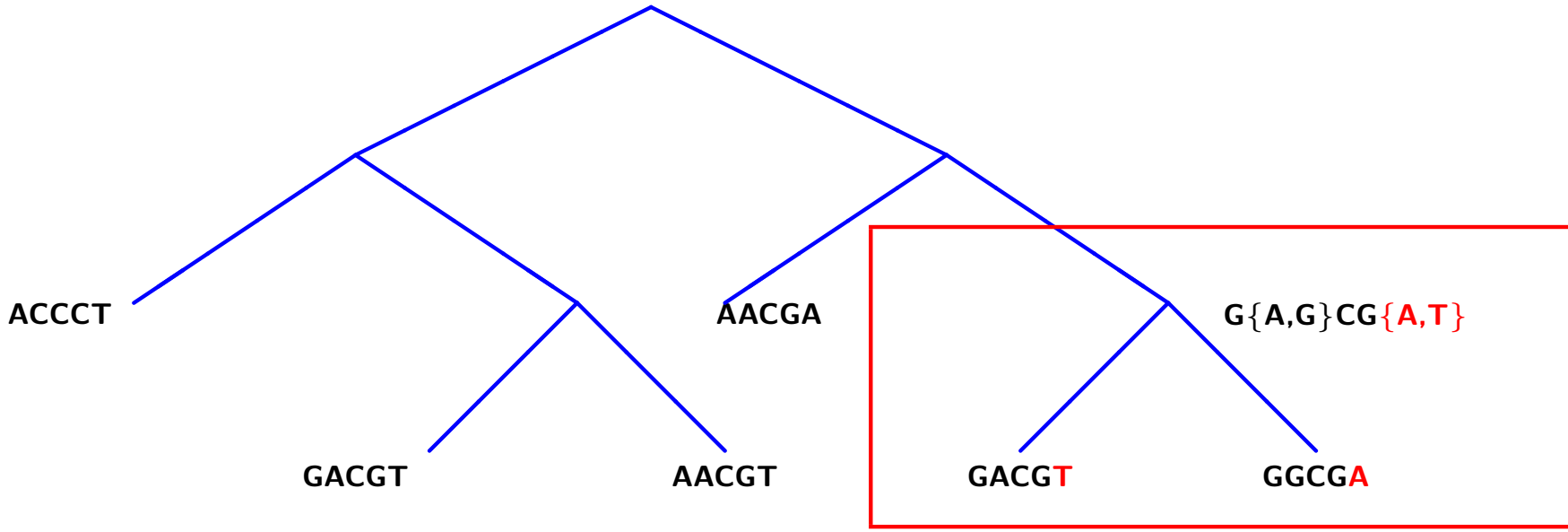
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



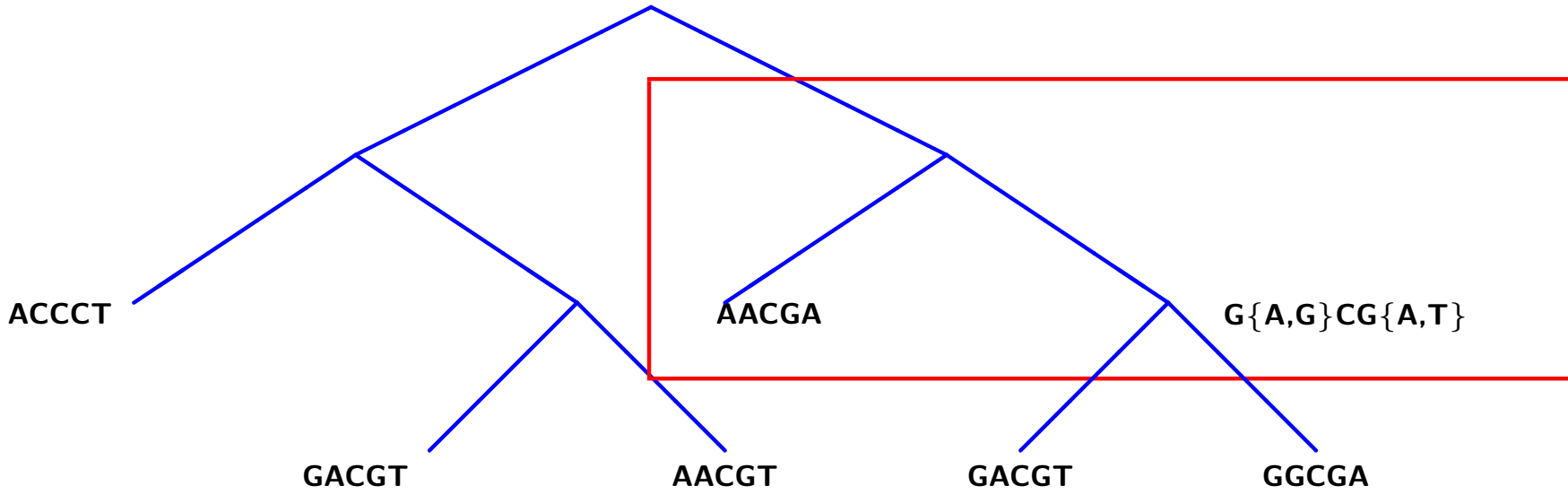
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



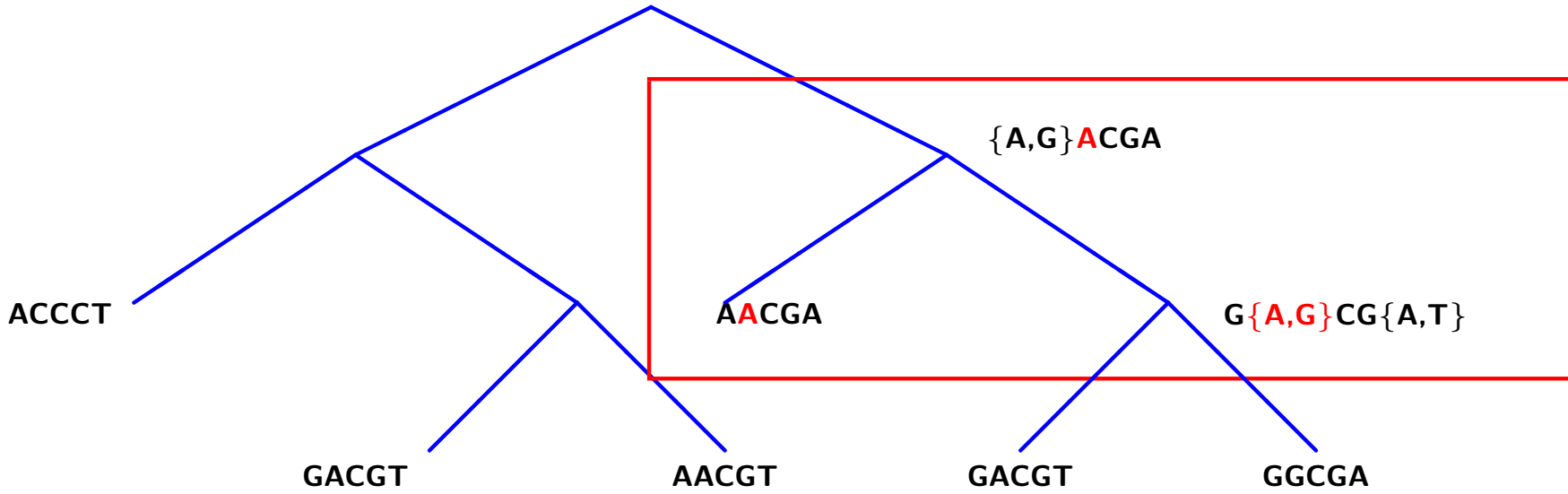
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



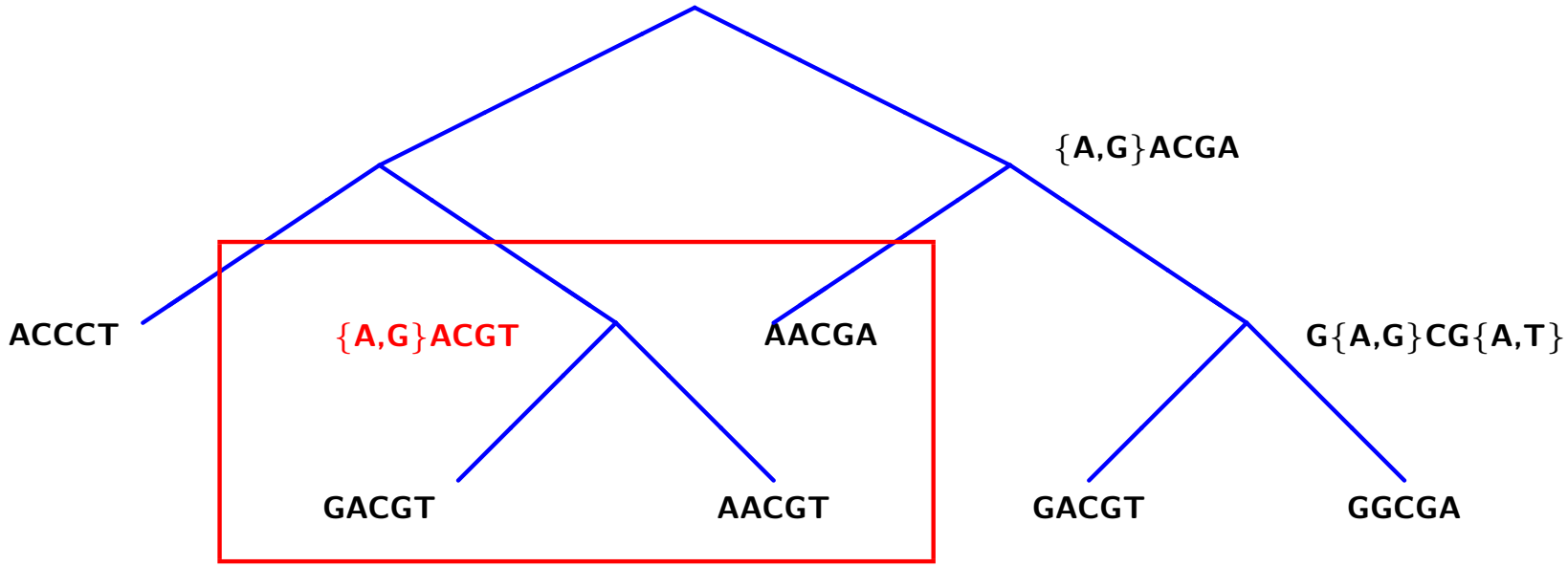
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



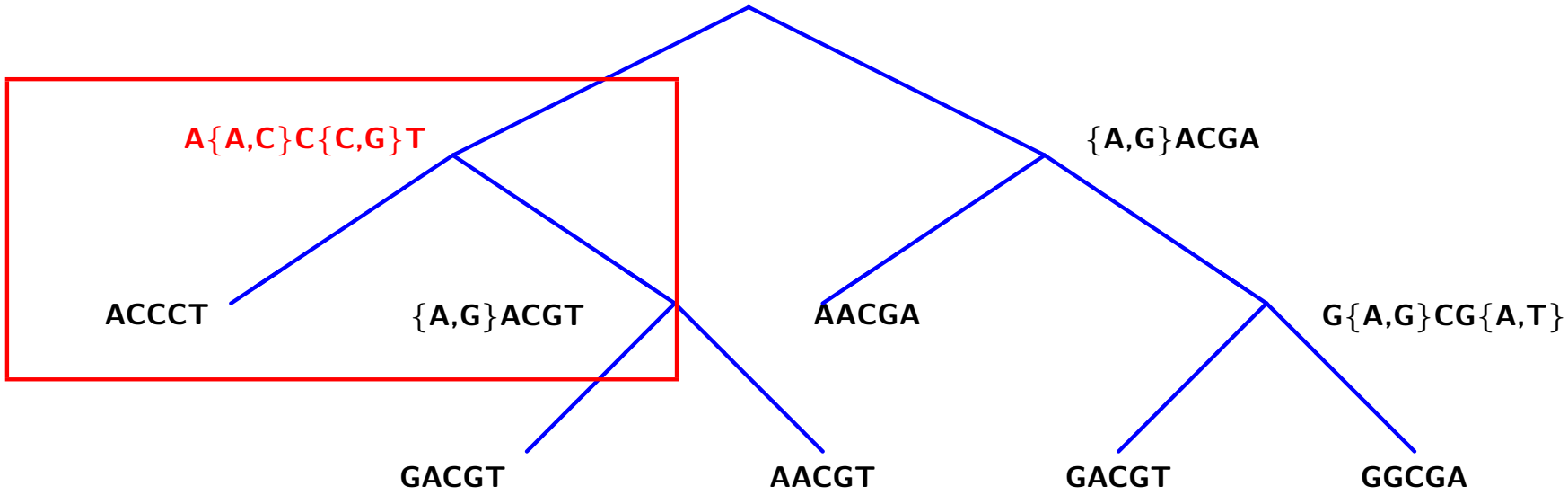
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



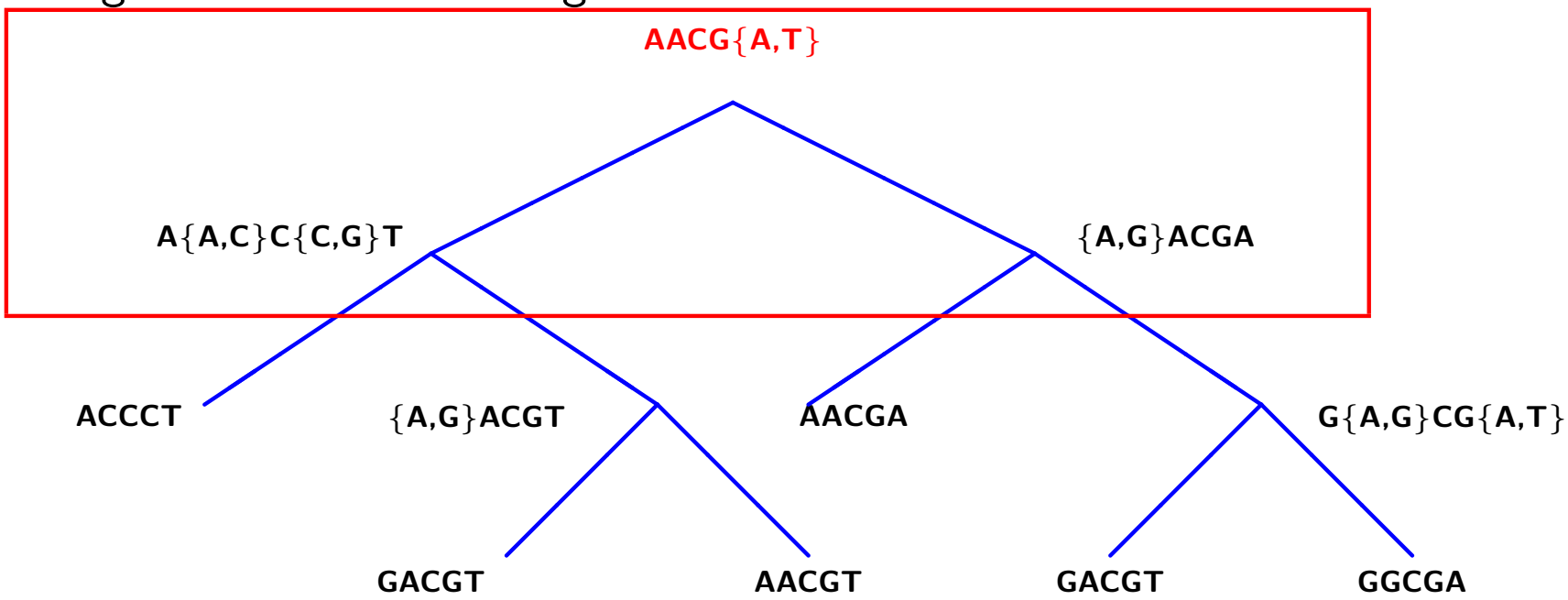
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



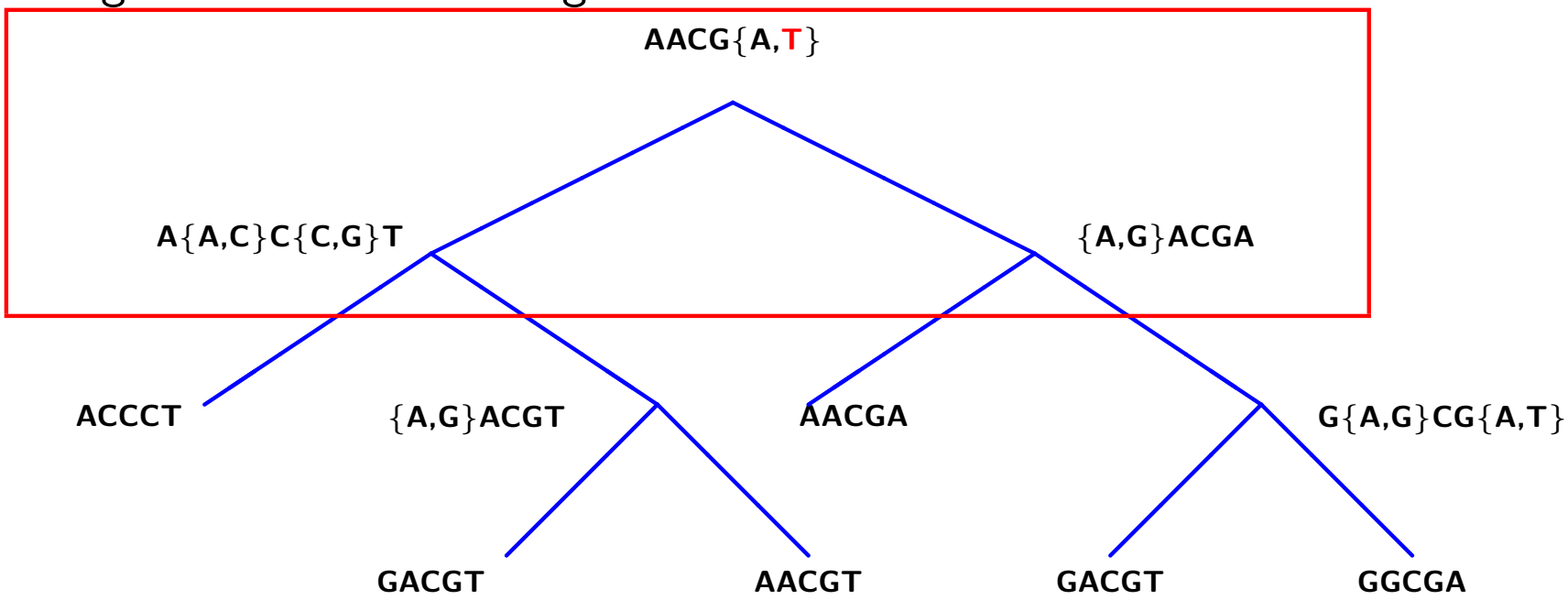
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



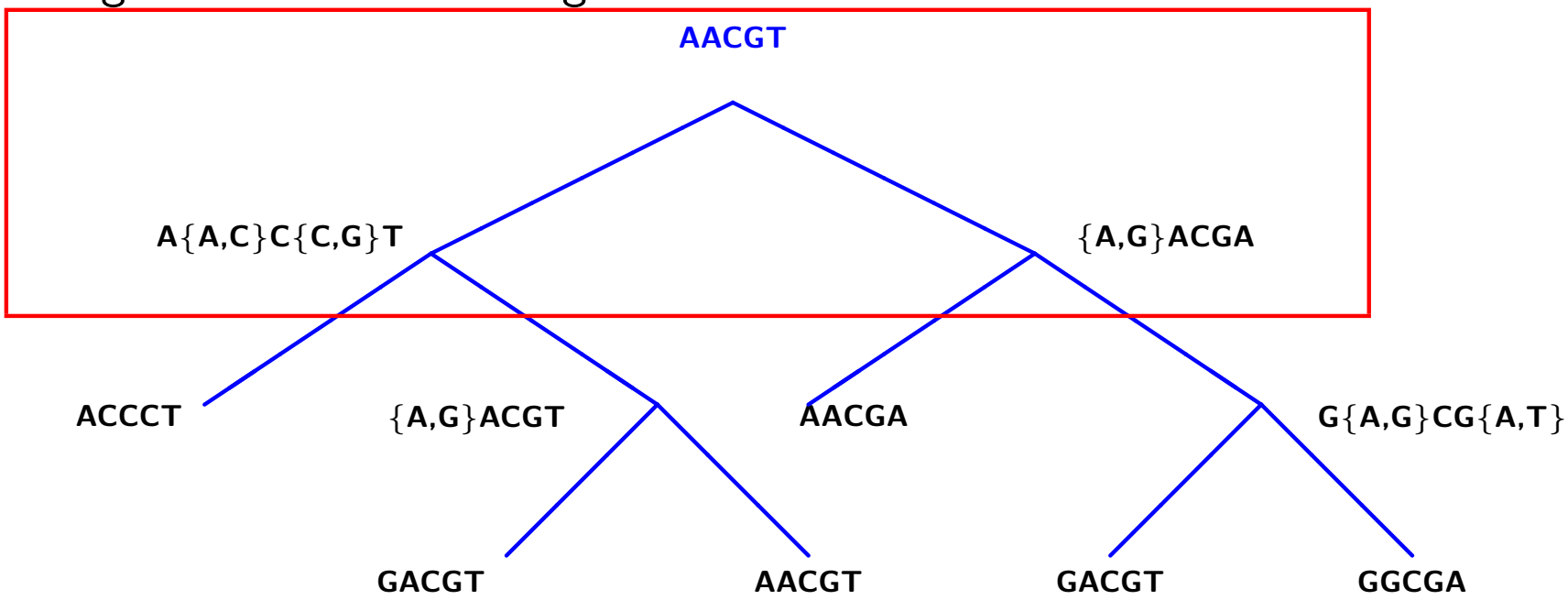
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



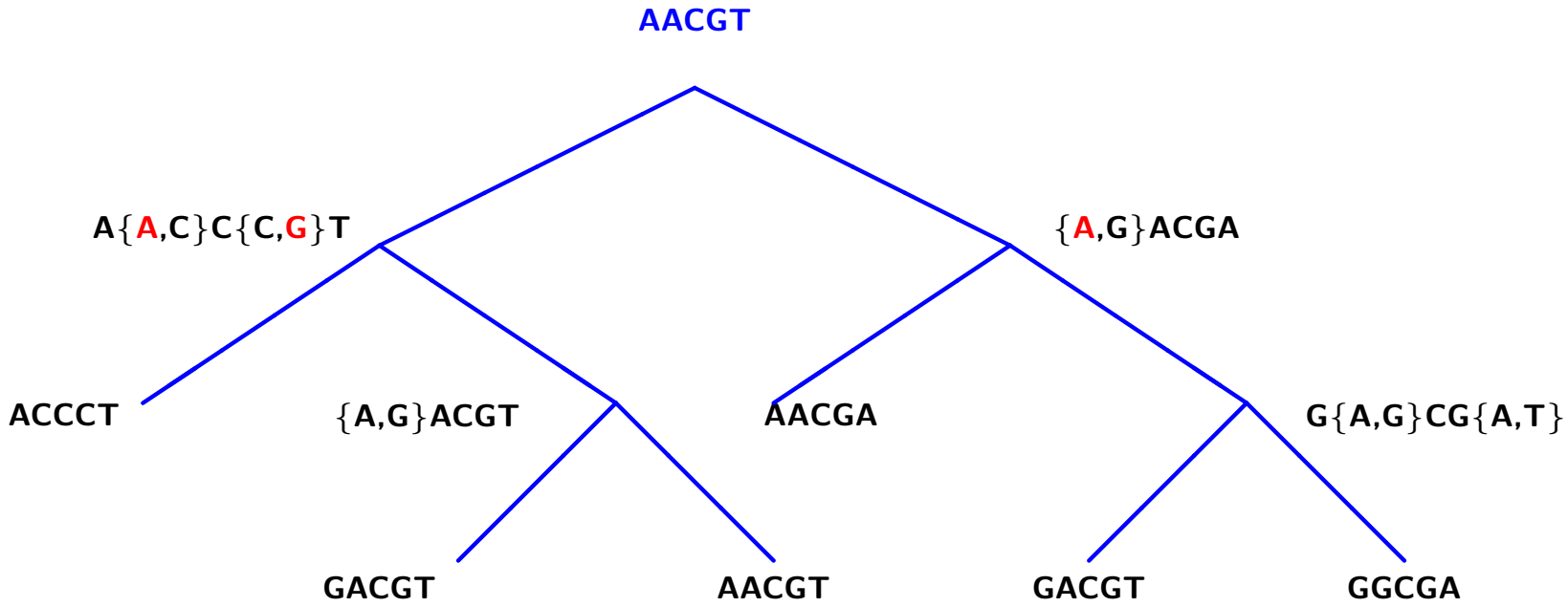
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



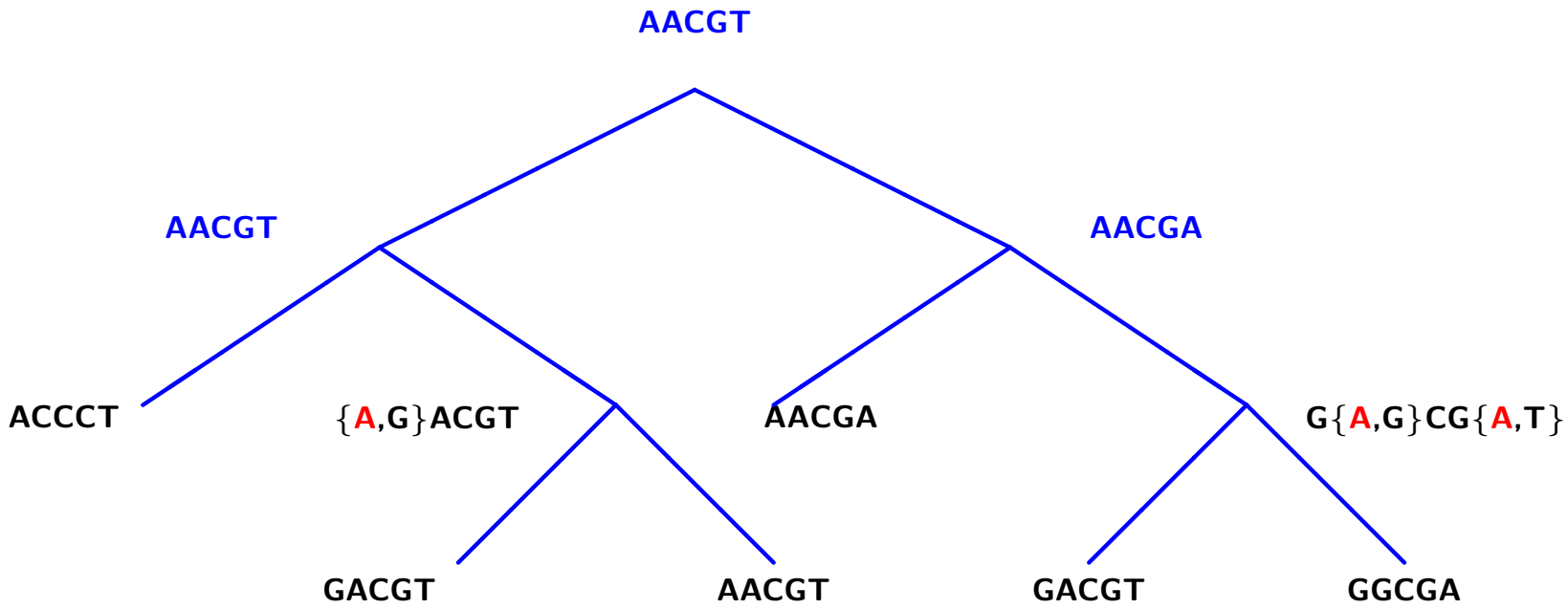
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



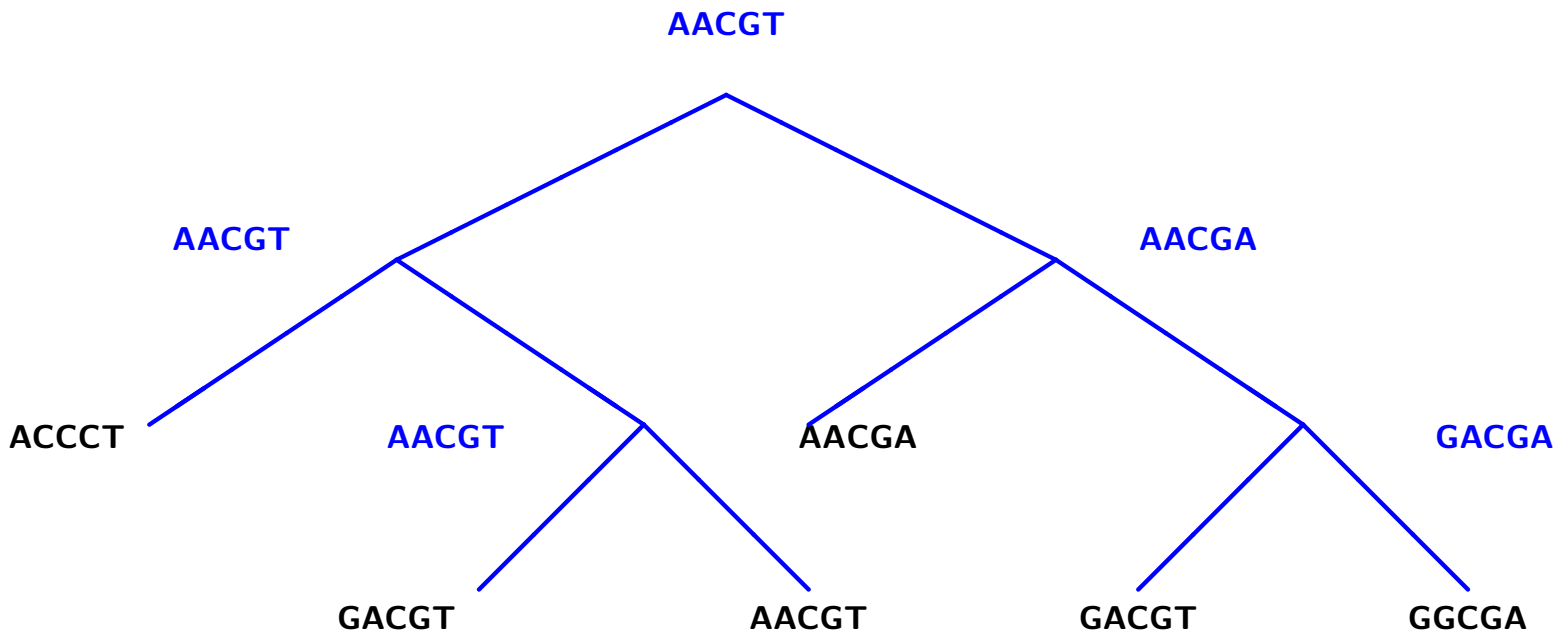
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



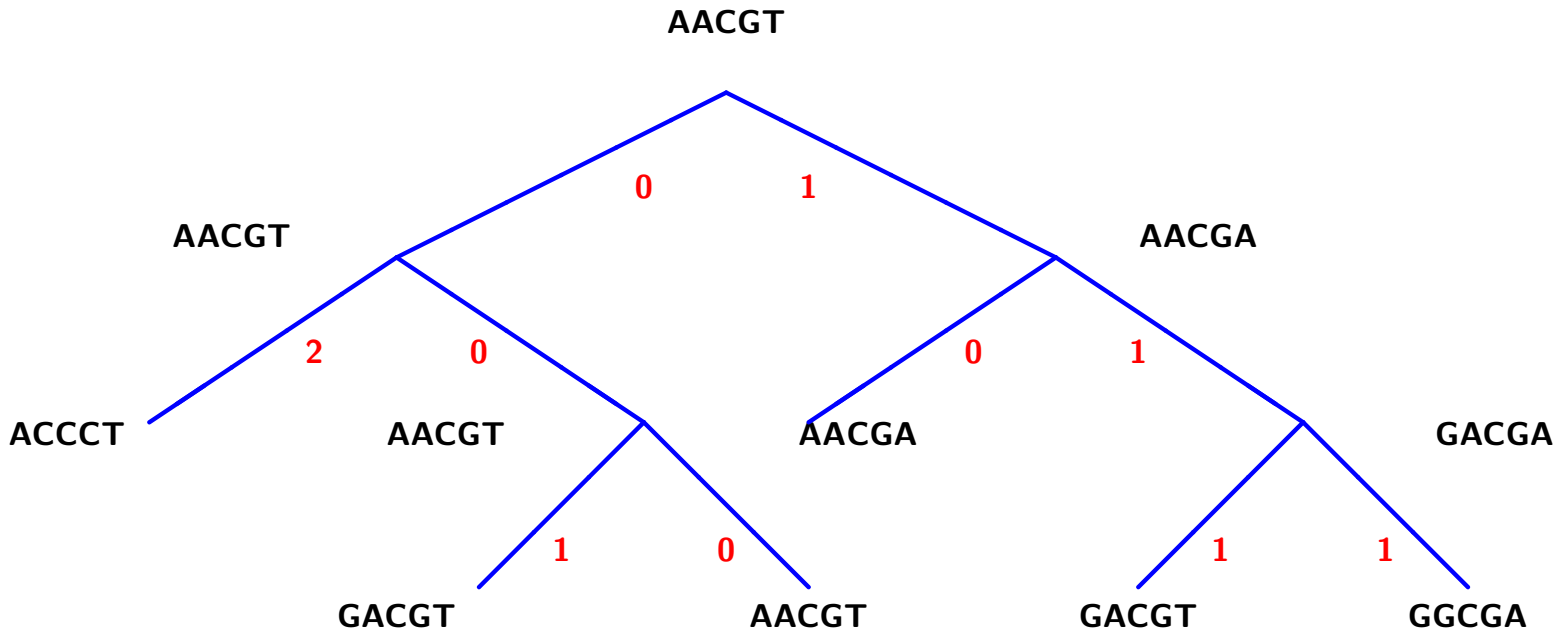
Maximum Parsimony

- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



Maximum Parsimony

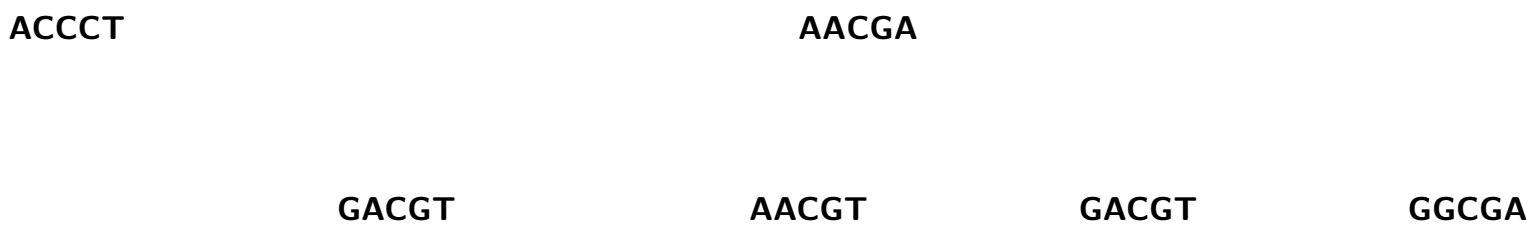
- Given sequences for leaves and a tree, first measure "minimal number of substitutions."
- Label the internal nodes with sequences that have minimal number of changes. Then count changes.



Total change, called the parsimony score is 7.

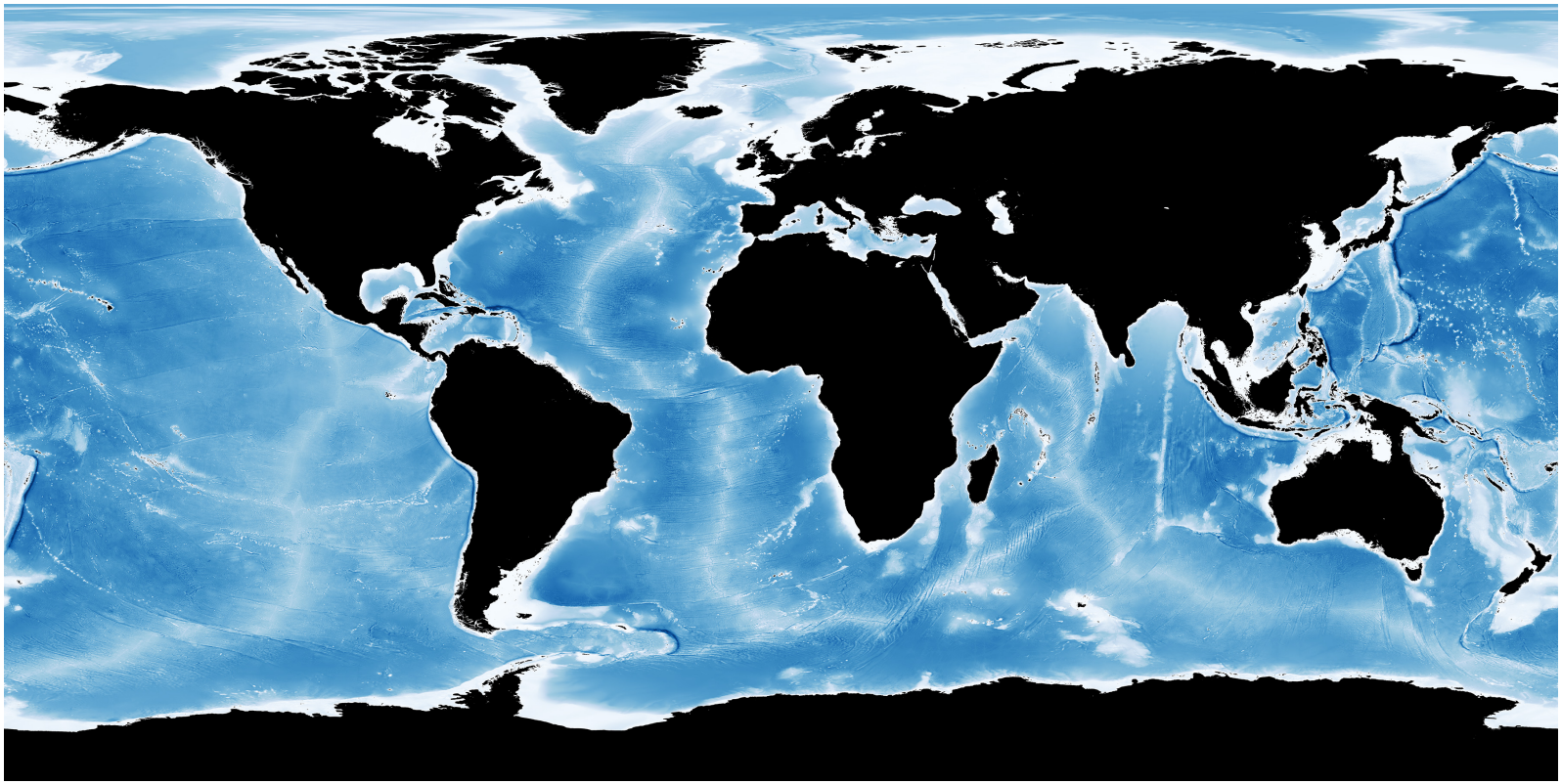
Maximum Parsimony

- Given sequences for leaves, find tree with minimal parsimony score:



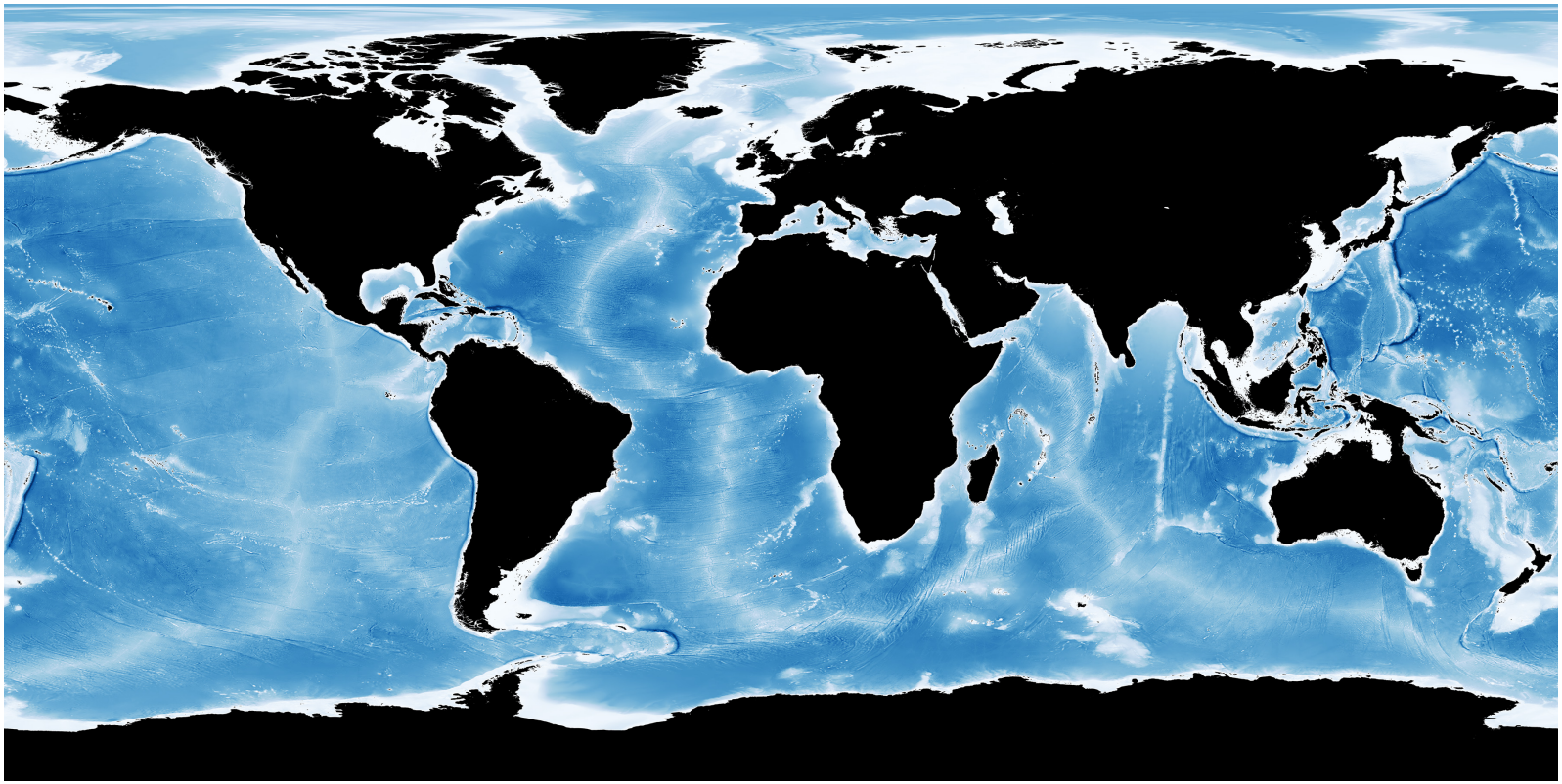
(Can you find a tree with a score better than 7?)

Analogy: Parsimony



NASA Blue Marble Bathymetry

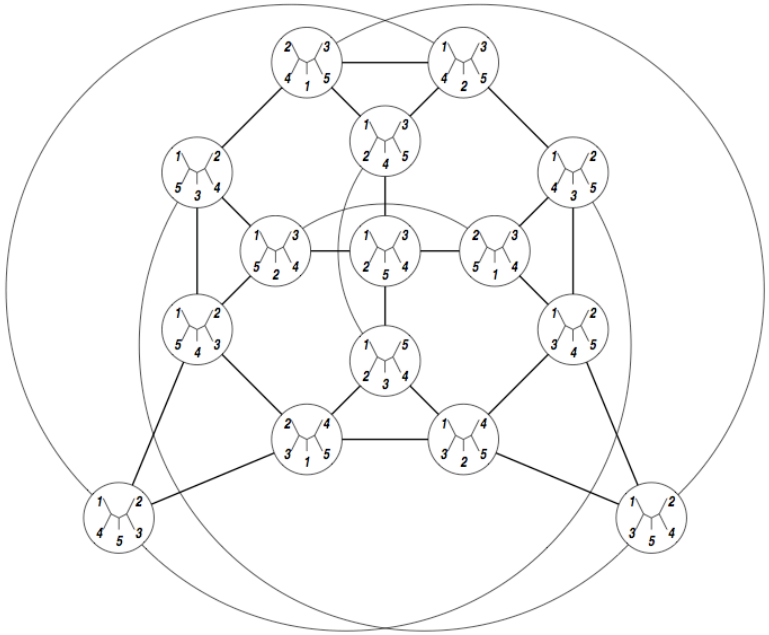
Analogy: Parsimony



NASA Blue Marble Bathymetry

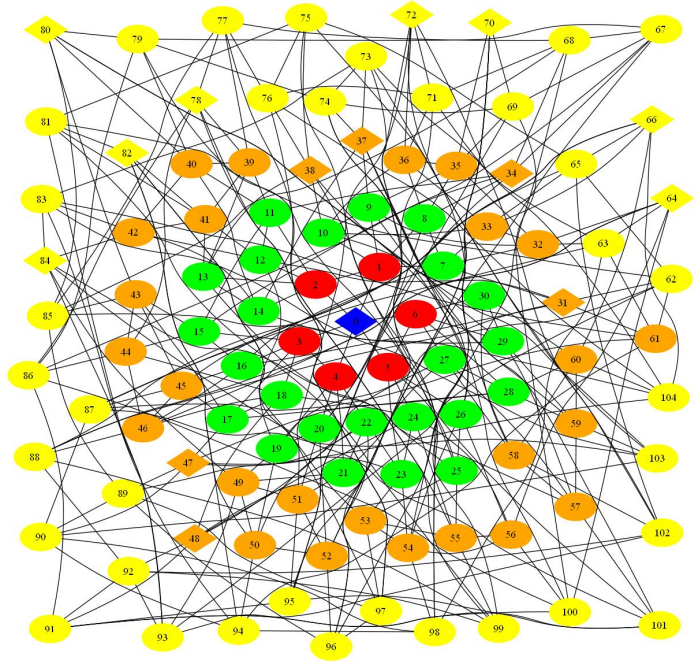
Find the lowest point.

Treespace



Treespace for $n = 5$ under NNI

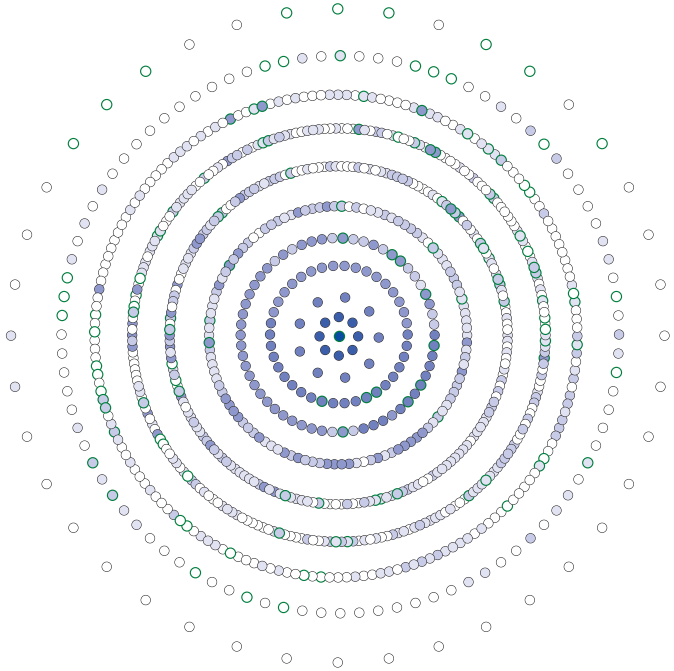
Bastert *et al.*, 2002



Treespace for $n = 6$

For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed distance metric.

Landscapes



Parsimony score for compatible characters for $n = 7$ (Urheim, Ford, & S, submitted)

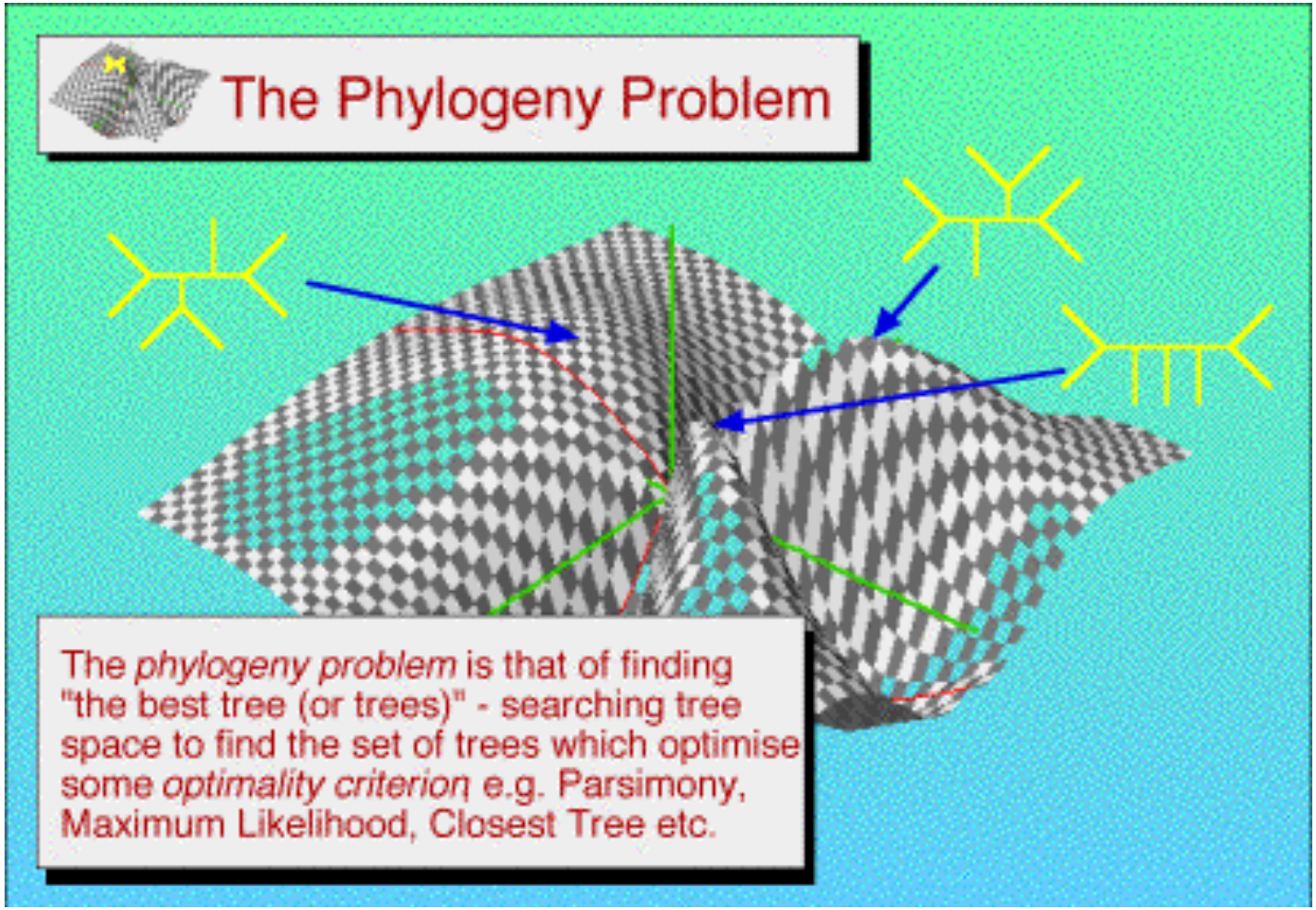
A treespace with assigned scores is often called a landscape.

Hillis' Helicopter

wiki commons

David Hillis: Representing tree scores as height, he wanted a visualization with a 'helicopter' to fly over the space of trees.

What does the landscape look like?



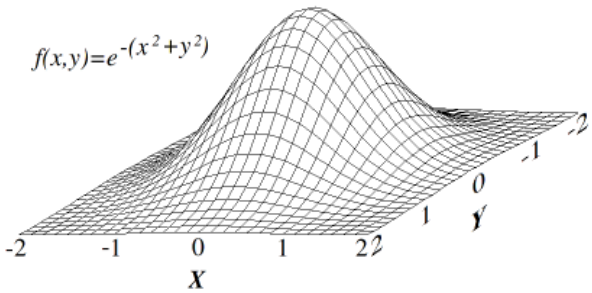
Mike Charleston, 1996

What does the landscape look like?

Each landscape depends on the number of taxa and the score of each tree (usually derived from the inputted character sequences).

What does the landscape look like?

Each landscape depends on the number of taxa and the score of each tree (usually derived from the inputted character sequences).

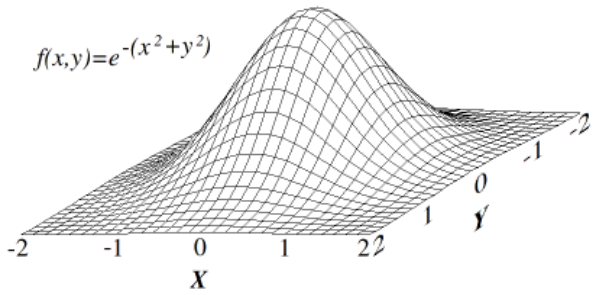


(from wikipedia)

If very smooth, 'hill climbing' will work well.

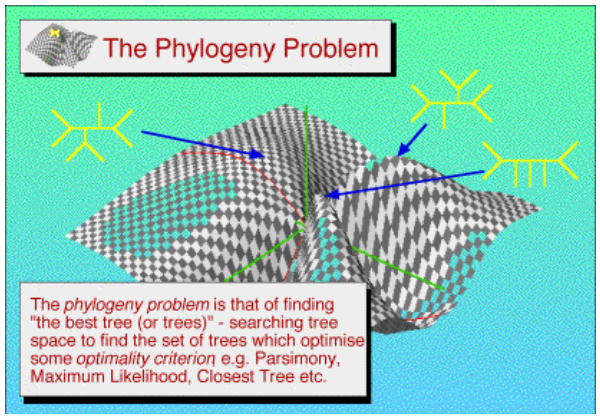
What does the landscape look like?

Each landscape depends on the number of taxa and the score of each tree (usually derived from the inputted character sequences).



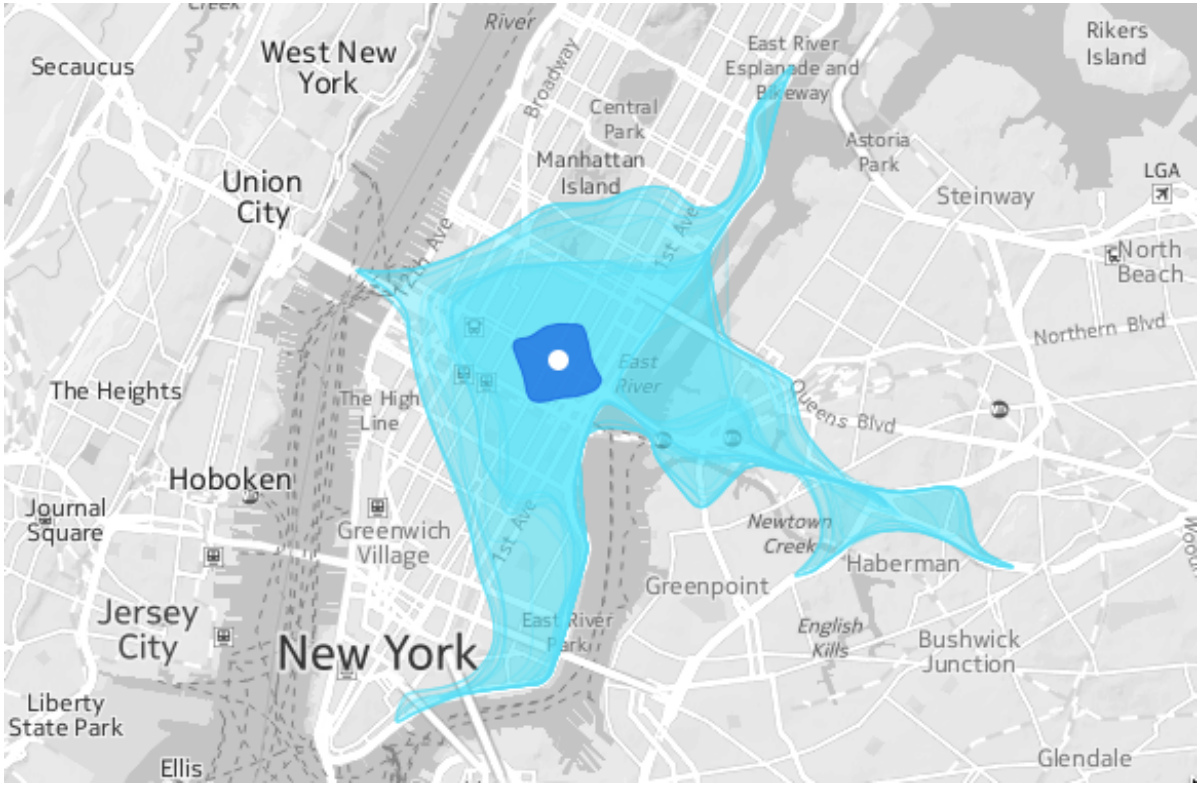
(from wikipedia)

If very smooth, 'hill climbing' will work well.



If very rugged, need more sophisticated searches that use the underlying structure of the space.

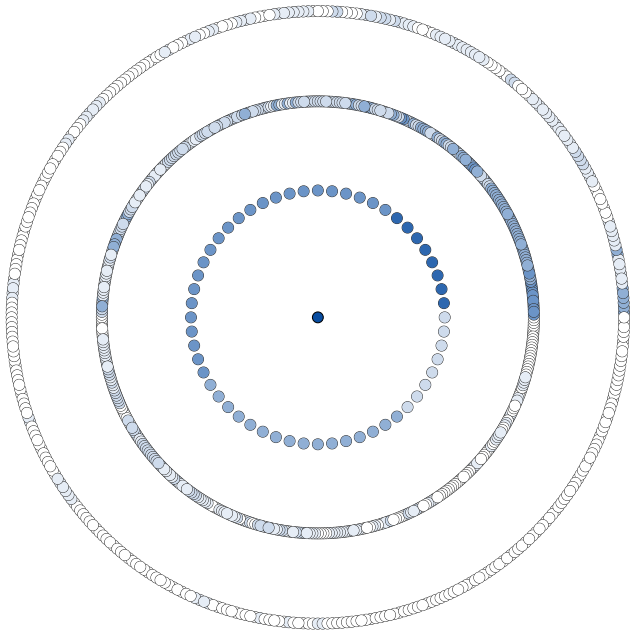
Analogy: Adjusting Search Space



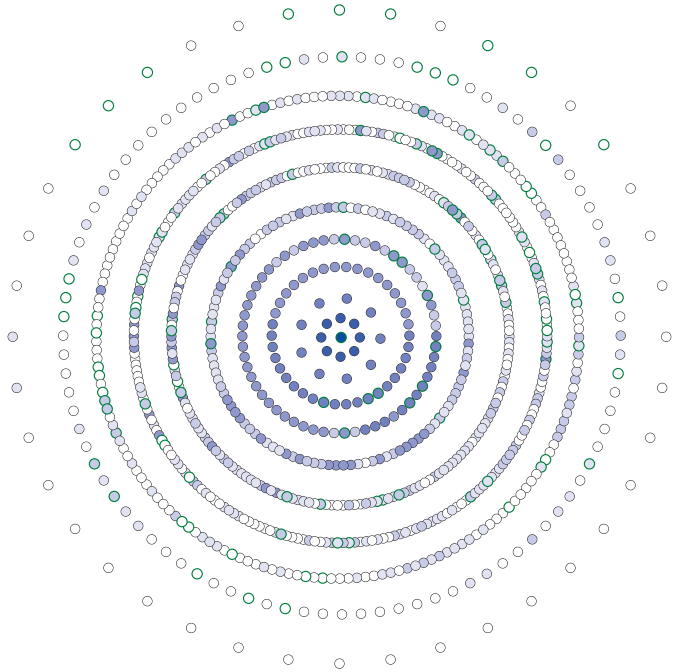
isoscope

Different metrics yield different neighbors: places you can reach in 10 minutes from Grand Central Station walking versus transit

Adjusting Search Space



SPR metric

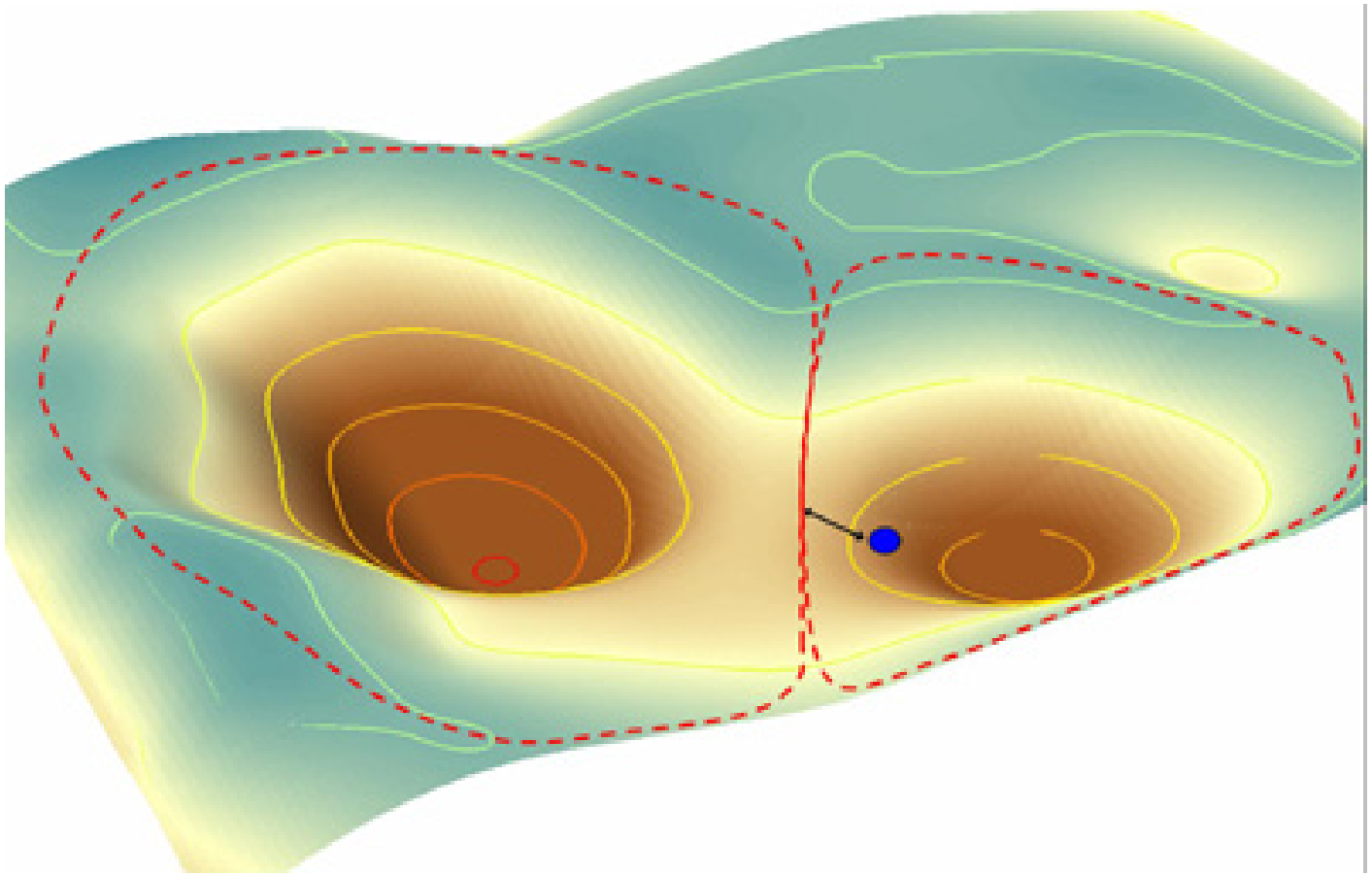


NNI metric

Parsimony score for compatible characters for $n = 7$ (Urheim, Ford, & S, submitted)

The same data, organized by different tree metrics.

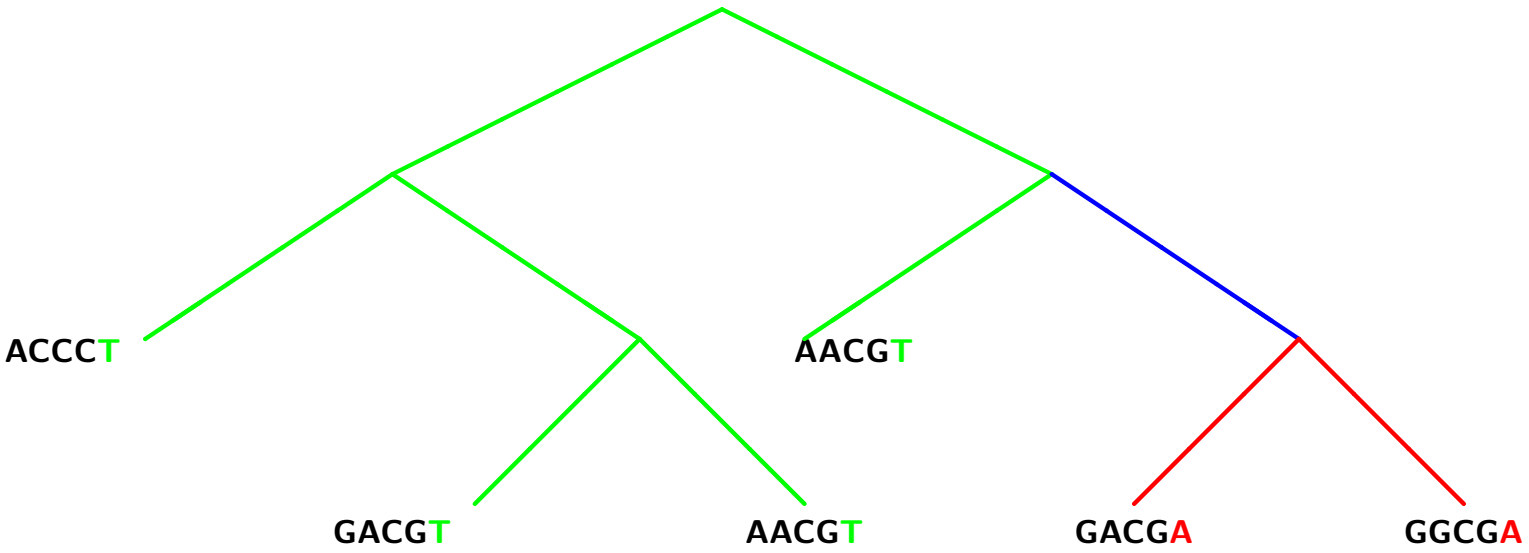
Attraction Basins



resalliance.org

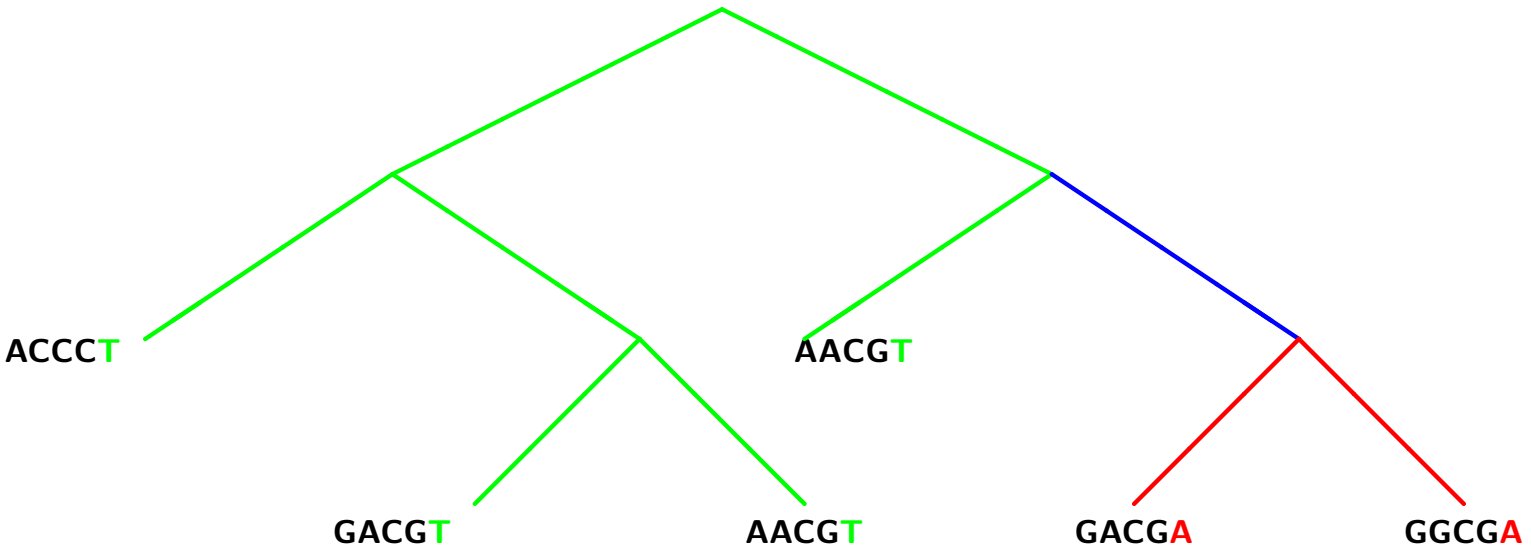
Compatible Characters

A character is compatible with a tree if each state induces a connected subtree:



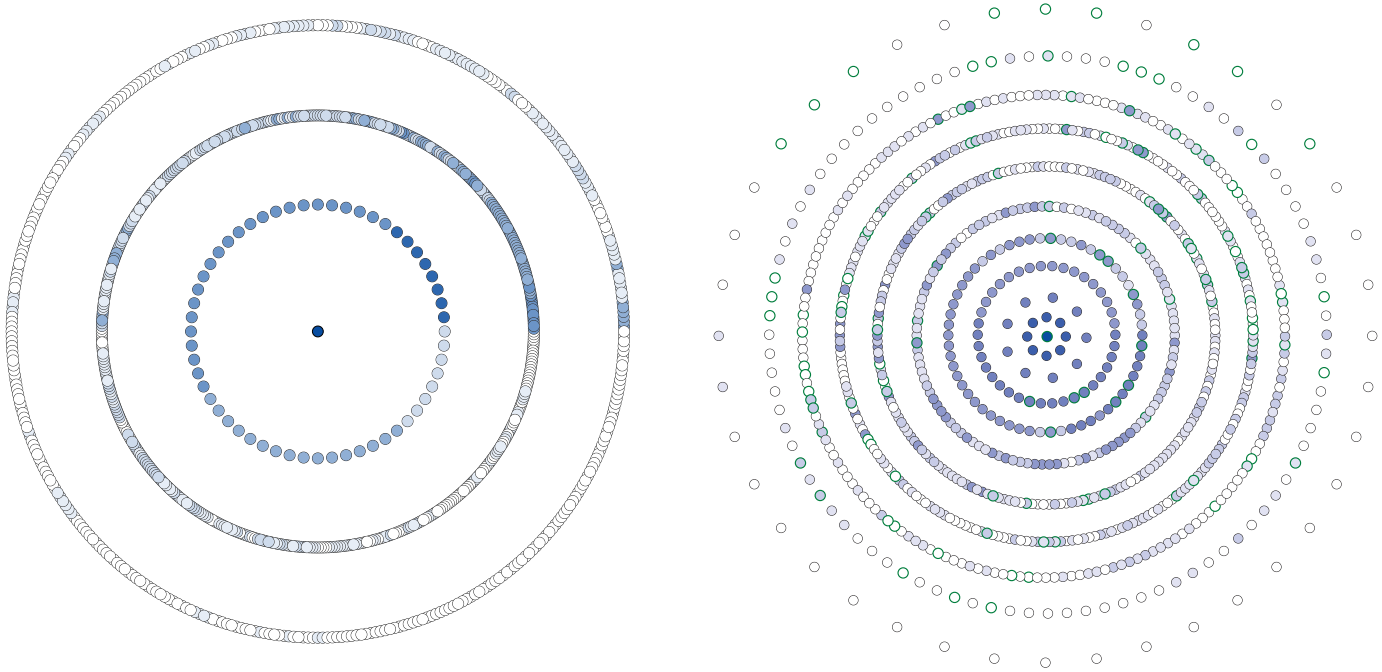
Compatible Characters

A character is compatible with a tree if each state induces a connected subtree:



A sequence of characters is compatible if there is at least one tree that all are compatible.

Adjusting Search Space



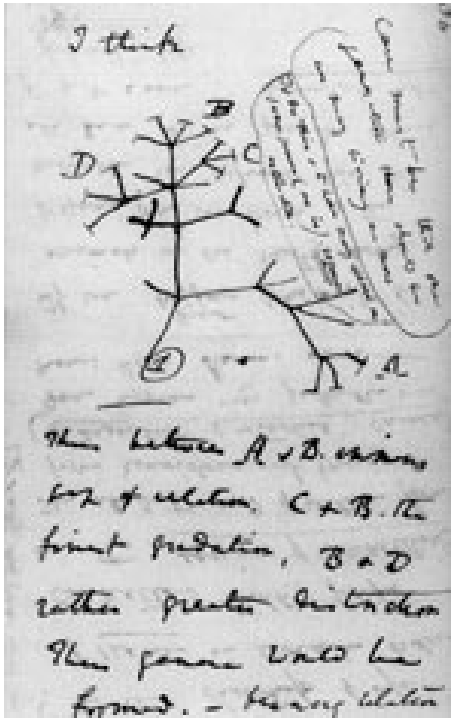
SPR metric NNI metric

Parsimony score for compatible characters for $n = 7$ (Urheim, Ford, & S, submitted)

Simplest Case: for compatible character sequences ('perfect data'):

- Under SPR, there is a single attraction basin.
- Under NNI, multiple attraction basins occur even for perfect data.

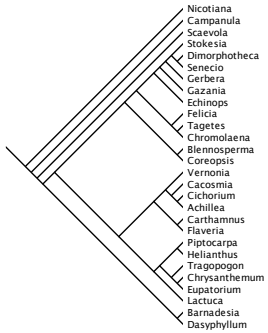
Outline

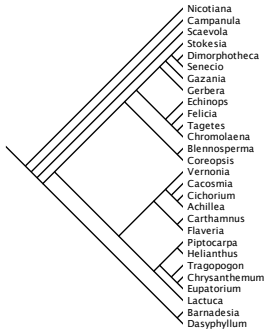


Charles Darwin, 1837

- Treespaces and Landscapes
- Metrics & Search
- Preprocessing to Improve Search
- Maximum Likelihood & Continuous Treespace
- When Trees are Not Enough....

Popular Tree Metrics

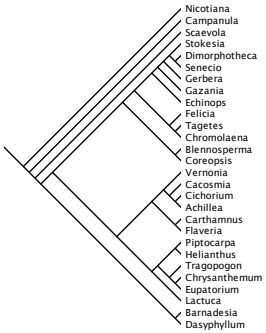


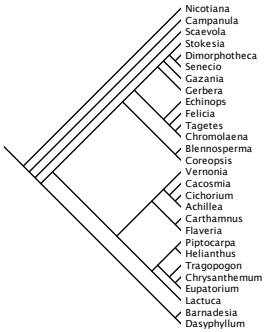


Those based on tree rearrangements:

- Subtree Prune and Regraft (SPR)
- Tree Bisection and Reconnection (TBR)
- Nearest Neighbor Interchange (NNI)

Popular Tree Metrics

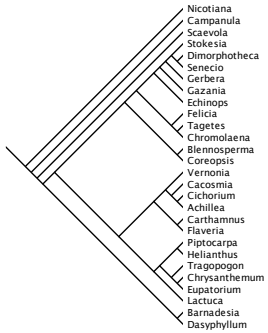




Those based on tree rearrangements:

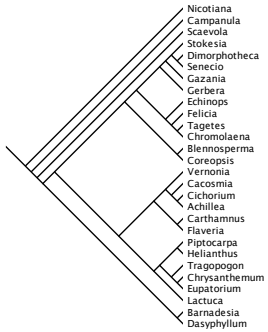
- Subtree Prune and Regraft (SPR)
- Tree Bisection and Reconnection (TBR)
- Nearest Neighbor Interchange (NNI)
- Used for Searching for Optimal Trees, NP-hard

Popular Tree Metrics



Those based on tree rearrangements:

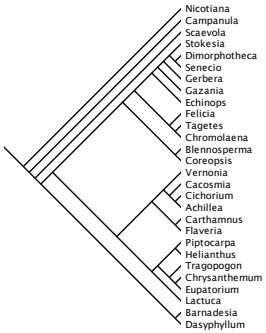
- Subtree Prune and Regraft (SPR)
- Tree Bisection and Reconnection (TBR)
- Nearest Neighbor Interchange (NNI)
- Used for Searching for Optimal Trees, NP-hard



Those based on comparing tree vectors:

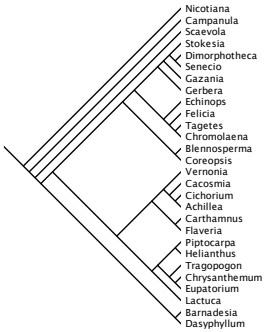
- Robinson-Foulds (RF)
- Rooted Triples (RT)
- Quartet Distance
- Billera-Holmes-Vogtmann (BHV or geodesic))

Popular Tree Metrics



Those based on tree rearrangements:

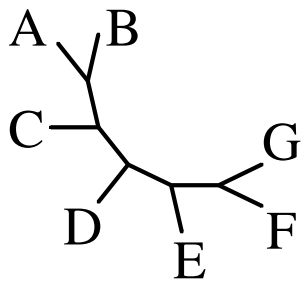
- Subtree Prune and Regraft (SPR)
- Tree Bisection and Reconnection (TBR)
- Nearest Neighbor Interchange (NNI)
- Used for Searching for Optimal Trees, NP-hard

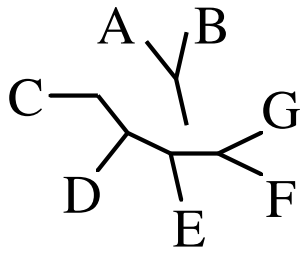


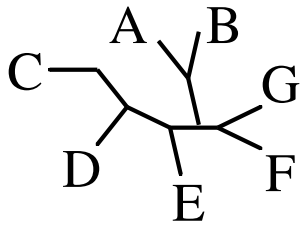
Those based on comparing tree vectors:

- Robinson-Foulds (RF)
- Rooted Triples (RT)
- Quartet Distance
- Billera-Holmes-Vogtmann (BHV or geodesic))
- Used for comparing trees, poly time

SPR Distance

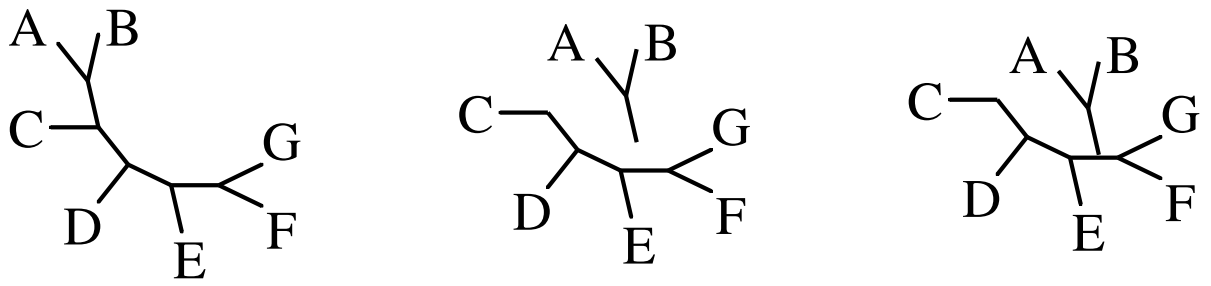






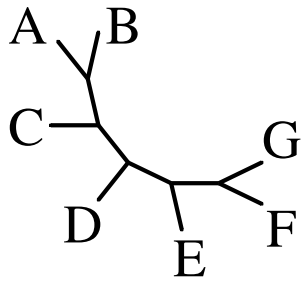
- SPR distance is the minimal number of moves that transforms one tree into the other.

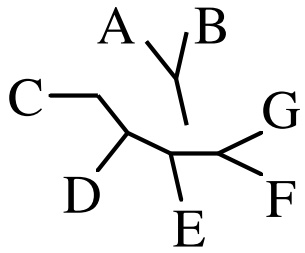
SPR Distance

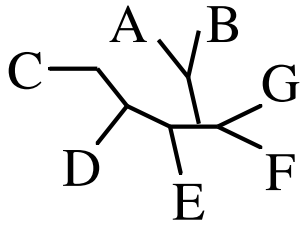


- SPR distance is the minimal number of moves that transforms one tree into the other.
- SPR for rooted trees is NP-hard (Bordewich & Semple '05).
- SPR for unrooted trees is NP-hard (Hickey *et al.* '08).
- SAT-based heuristic (Bonet & S '09).

Fixed Parameter Tractability for SPR







- Rooted: (Borderwich & Semple '05) Developed an agreement forest for SPR on rooted trees. Agreement forest gives NP-hardness and is used to show FPT.
- Unrooted: (Bonet & S '07) Used a variant of the reduction rules to get FPT.

Side Note: Phylogeny Problems

Steel's $100 Problems: "A choice of NZ$100 plus bottle of NZ wine, OR US$100, OR free registration and accommodation grant at the annual New Zealand phylogenetics meeting (value NZ$300 - flights not included!) for the first correct solution to any of these problems."

Side Note: Phylogeny Problems

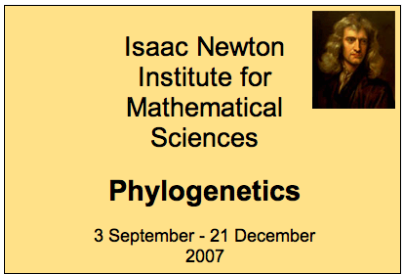
Steel's $100 Problems: "A choice of NZ$100 plus bottle of NZ wine, OR US$100, OR free registration and accommodation grant at the annual New Zealand phylogenetics meeting (value NZ$300 - flights not included!) for the first correct solution to any of these problems."

Penny Ante Problems: "A prize of your choice between $100 or a bottle of single malt whisky (for medicinal purposes only) . . . announced by the end of the NZ phylogenetics conference."

Side Note: Phylogeny Problems

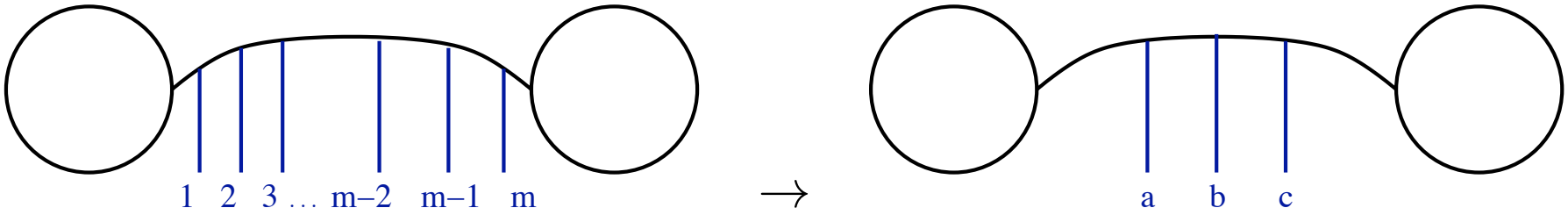
Steel's $100 Problems: "A choice of NZ$100 plus bottle of NZ wine, OR US$100, OR free registration and accommodation grant at the annual New Zealand phylogenetics meeting (value NZ$300 - flights not included!) for the first correct solution to any of these problems."

Penny Ante Problems: "A prize of your choice between $100 or a bottle of single malt whisky (for medicinal purposes only) . . . announced by the end of the NZ phylogenetics conference."



Isaac Newton Institute Challenges: A bottle of wine for those solved by the end of the 2007 INI program.

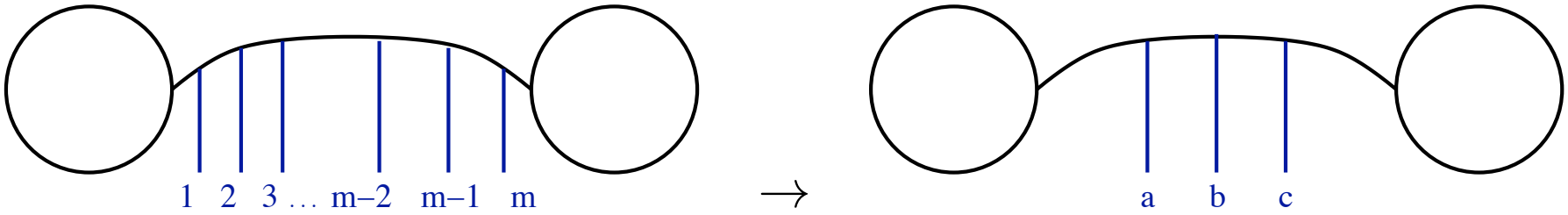
SPR Challenge



- ($100): Does shrinking common subchains in trees preserve SPR distance?

  (Implies fixed parameter tractability.)

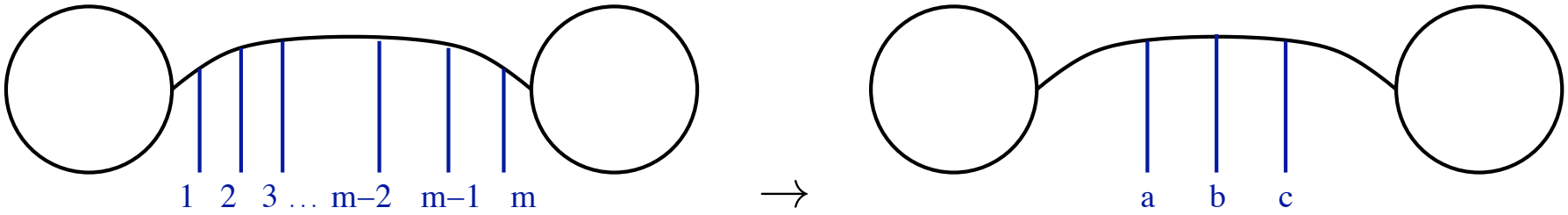
SPR Challenge



- ($100): Does shrinking common subchains in trees preserve SPR distance?

  (Implies fixed parameter tractability.)
- Bordewich & Semple '05: Yes, for rooted trees.

SPR Challenge

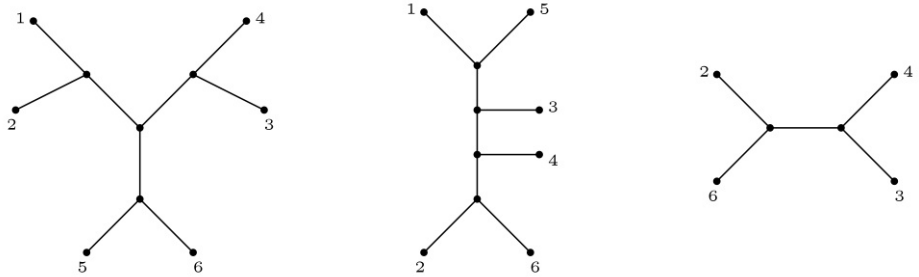


- ($100): Does shrinking common subchains in trees preserve SPR distance?

  (Implies fixed parameter tractability.)
- Bordewich & Semple '05: Yes, for rooted trees.
- Open for unrooted trees

  (uSPR is known to be FPT by other means, Bonet & S '09).

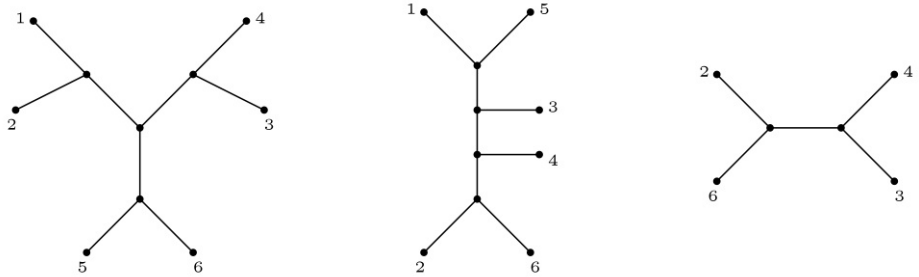
How little can two trees agree?



Steel & Székely, 2009

(INI): Given two unrooted binary phylogenetic trees T, T', an agreement set for T, T' is a subset Y of X for which $T|_Y = T'|_Y$.
Is there a constant c, so that for any two trees T, T' have an agreement subtree of size $c \log n$?

How little can two trees agree?

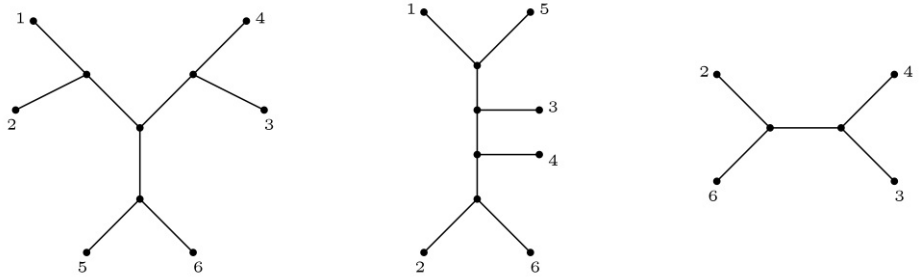


Steel & Székely, 2009

(INI): Given two unrooted binary phylogenetic trees T, T', an agreement set for T, T' is a subset Y of X for which $T|_Y = T'|_Y$.
Is there a constant c, so that for any two trees T, T' have an agreement subtree of size $c \log n$?

- Steel & Székely '09: The agreement subtree is of size $\Omega(\log(\log n))$.

How little can two trees agree?

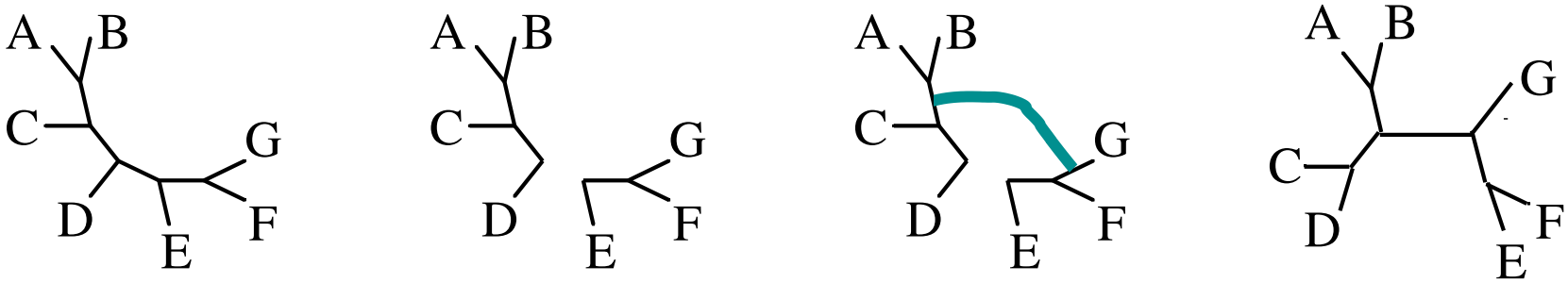


Steel & Székely, 2009

(INI): Given two unrooted binary phylogenetic trees T, T', an agreement set for T, T' is a subset Y of X for which $T|_Y = T'|_Y$.
Is there a constant c, so that for any two trees T, T' have an agreement subtree of size $c \log n$?

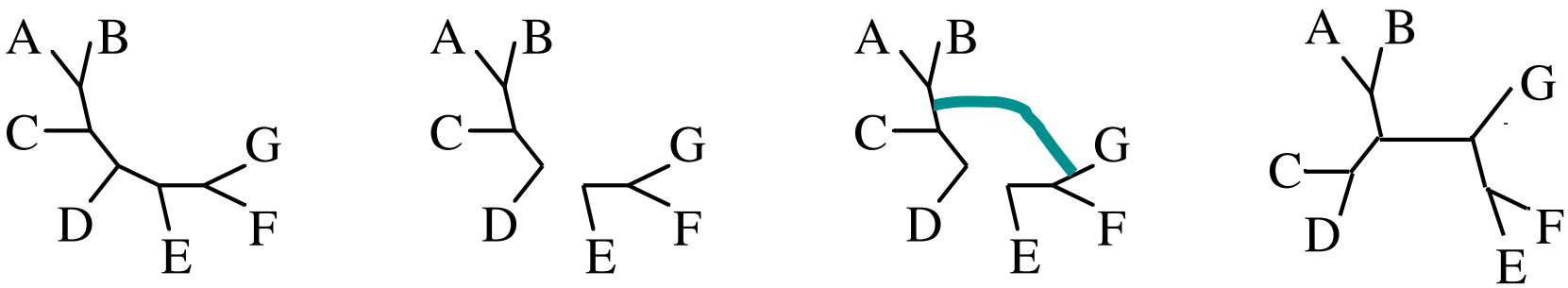
- Steel & Székely '09: The agreement subtree is of size $\Omega(\log(\log n))$.
- Martin & Thatte '12: The agreement subtree is of size $\Omega(\sqrt{\log n})$.

TBR Distance



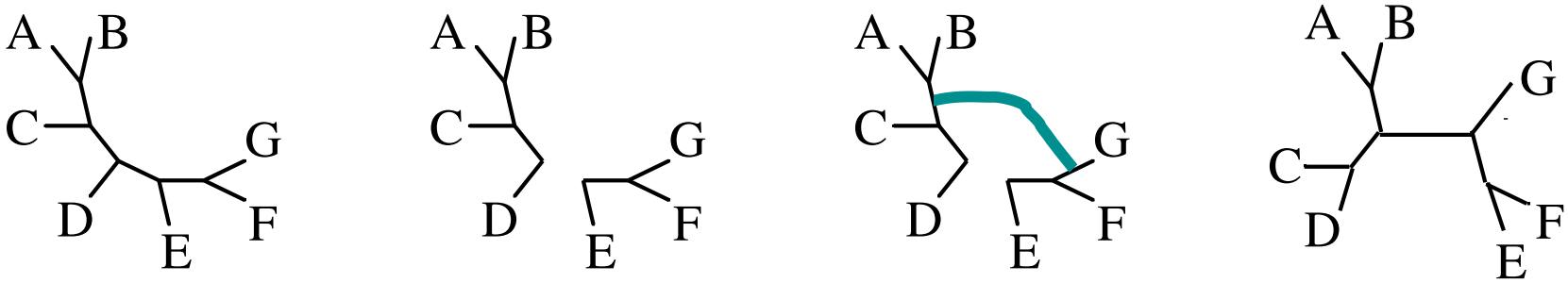
- TBR distance is the minimal number of moves that transforms one tree into the other.

TBR Distance



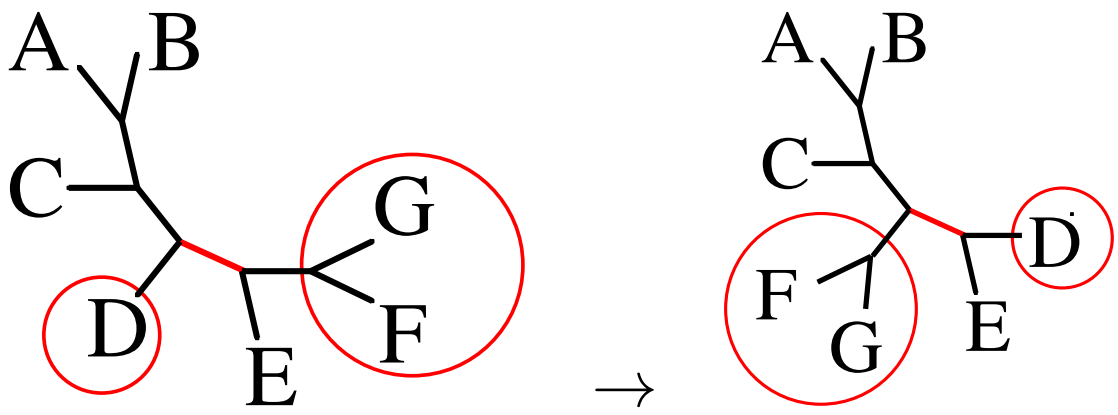
- TBR distance is the minimal number of moves that transforms one tree into the other.
- TBR is NP-hard and FPT. (Allen & Steel '01)

TBR Distance



- TBR distance is the minimal number of moves that transforms one tree into the other.
- TBR is NP-hard and FPT. (Allen & Steel '01)
- TBR has a linear time 5-approximation and a polynomial time 3-approximation (Amenta, Bonet, Mahindru, & S '06; Bordewich, McCartin, & Semple '08)

NNI Metric



The NNI distance between two trees is the minimal number of moves needed to transform one to the other (NP-hard, DasGupta *et al.* '97).

Bryant's Challenge: Walking Through Trees

David Bryant

An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.

Bryant's Challenge: Walking Through Trees

David Bryant

An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.

1. What is the shortest NNI walk that passes through all binary trees on n leaves?

Bryant's Challenge: Walking Through Trees

David Bryant

An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.

1. What is the shortest NNI walk that passes through all binary trees on n leaves?
2. Suppose we are given a tree T. What is the shortest NNI walk that passes through all the trees that lie at most one SPR (subtree prune and regraft) move from T?

Bryant's Challenge: Walking Through Trees

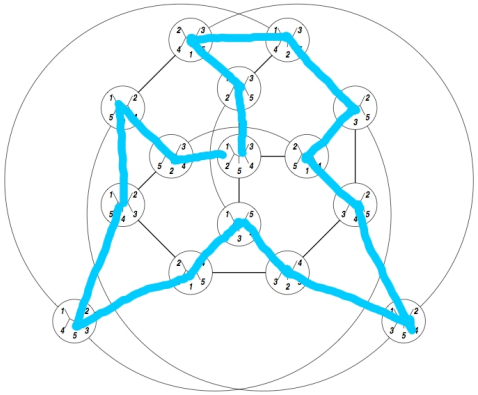
David Bryant

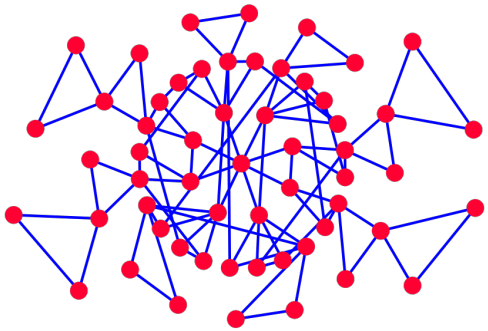
An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.

1. What is the shortest NNI walk that passes through all binary trees on n leaves?
2. Suppose we are given a tree T. What is the shortest NNI walk that passes through all the trees that lie at most one SPR (subtree prune and regraft) move from T?

NZ Penny Ante: $100 NZ or a bottle of fine whisky (Also appeared on the Isaac Newton Institute Phylogenetic Challenges, 2007)

Bryant's Challenges: Walking Through Trees

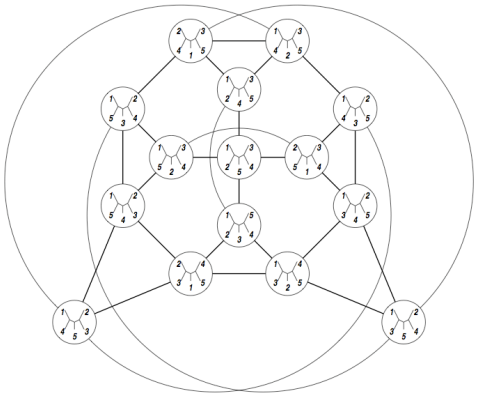




Penny Ante: An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.

1. What is the shortest NNI-walk that passes through all binary trees on n leaves?
   Gordon, Ford, & S '13:
   For all n, there exists a Hamiltonian path.
2. Suppose we are given a tree T. What is the shortest NNI-walk that passes through all the trees that lie at most one SPR (subtree prune and regraft) move from T?
   Caceres, Castillo, Lee, & S '13:
   For all n, it's $t + \Theta(n^2)$.

Treespace

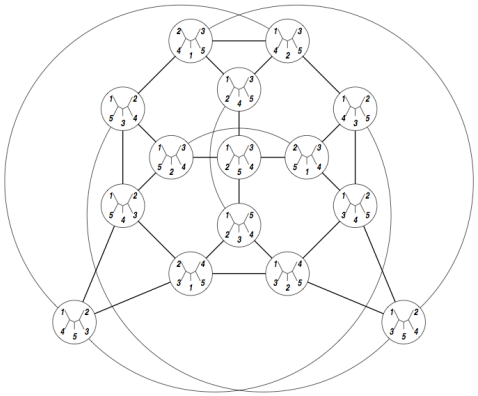


Treespace for $n = 5$ under NNI

Bastert *et al.*, 2002

- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.

Treespace

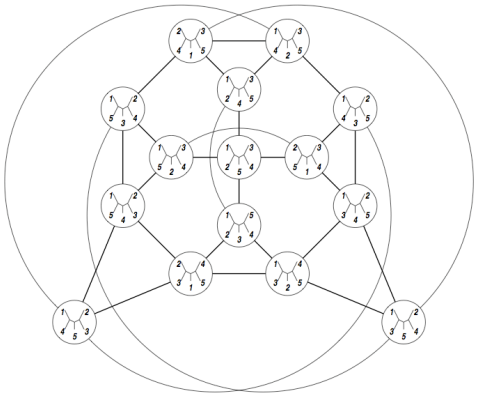


Treespace for $n = 5$ under NNI

Bastert *et al.*, 2002

- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.
- The diameter of neighborhoods varies by metric:

Treespace



Treespace for $n = 5$ under NNI

Bastert *et al.*, 2002

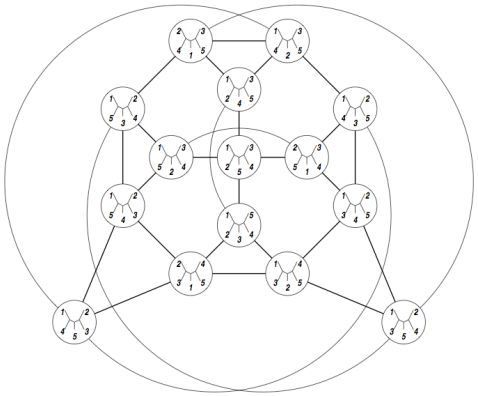
- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.
- The diameter of neighborhoods varies by metric:

	General	$n = 5$
NNI	$\Theta(n \log n)$	4
SPR	$n - \Theta(\sqrt{n})$	2
TBR	$n - \Theta(\sqrt{n})$	2

NNI: Li, Tromp & Zhang '96

SPR & TBR: Atkins & McDiarmid '15

Treespace

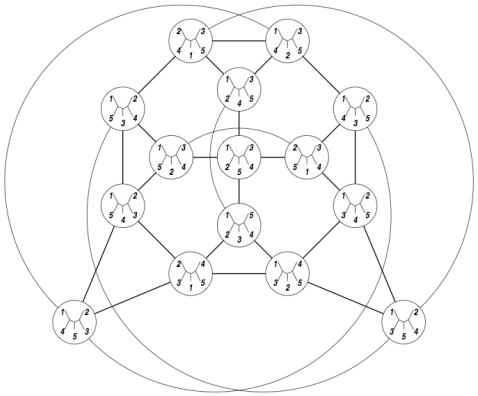


Treespace for $n = 5$ under NNI

Bastert *et al.* '02

- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.

Treespace

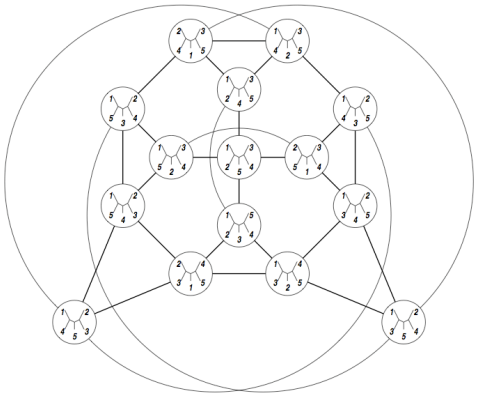


Treespace for $n = 5$ under NNI

Bastert *et al.* '02

- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.
- The size of neighborhoods varies by metric (Allen & Steel '01):

Treespace



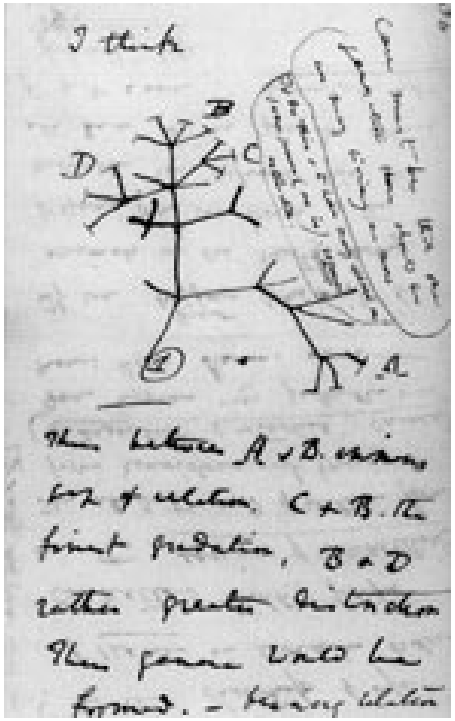
Treespace for $n = 5$ under NNI

Bastert *et al.* '02

- For every n, treespace is the space of all phylogenetic trees on a n taxa, under a fixed metric.
- The size of neighborhoods varies by metric (Allen & Steel '01):

	General	$n = 5$
NNI	$2n - 6$	4
SPR	$2(n-3)(2n-7)$	12
TBR	$< (2n-3)(n-3)^2$	12

Outline



Charles Darwin, 1837

- Treespaces and Landscapes
- Metrics & Search
- Preprocessing to Improve Search
- Maximum Likelihood & Continuous Treespace
- When Trees are Not Enough....

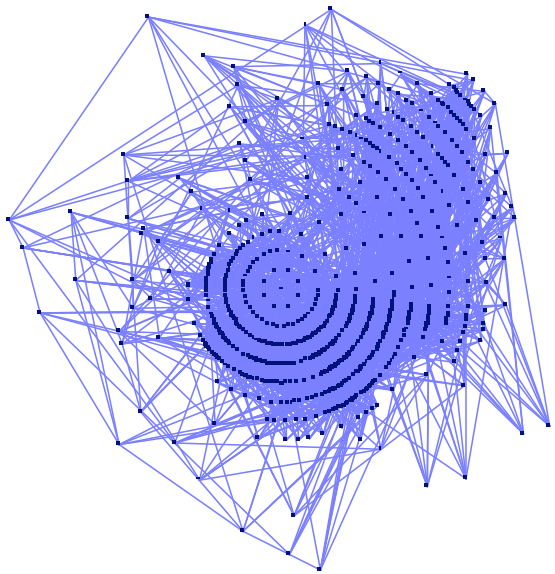
Preprocessing the Data: Finding Easy Instances

Preprocessing the Data: Finding Easy Instances

Identical Sequences

A	GTTAGAAGGCGGCCAGCGAC. . .
B	GTTAGAAGGCGGCCAGCGAC. . .
C	GTTAGAAGGCGGCCAGCGAC. . .
D	GTTAGAAGGCGGCCAGCGAC. . .
E	GTTAGAAGGCGGCCAGCGAC. . .
F	GTTAGAAGGCGGCCAGCGAC. . .

945 Rooted Trees on 6 Leaves

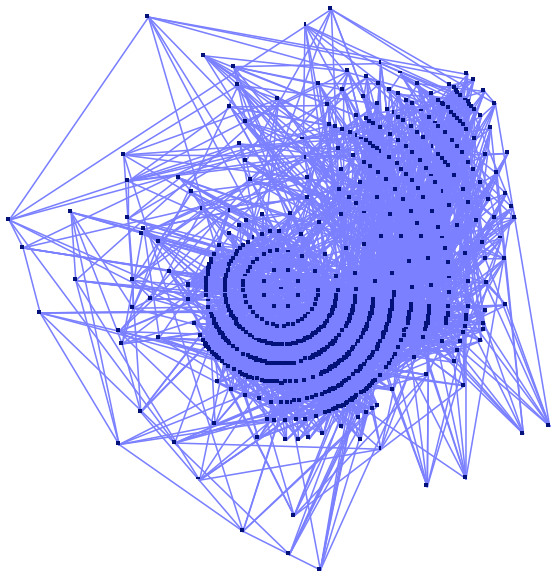


Preprocessing the Data: Finding Easy Instances

Identical Sequences

A	GTTAGAAGGCGGCCAGCGAC...
B	GTTAGAAGGCGGCCAGCGAC...
C	GTTAGAAGGCGGCCAGCGAC...
D	GTTAGAAGGCGGCCAGCGAC...
E	GTTAGAAGGCGGCCAGCGAC...
F	GTTAGAAGGCGGCCAGCGAC...

945 Rooted Trees on 6 Leaves



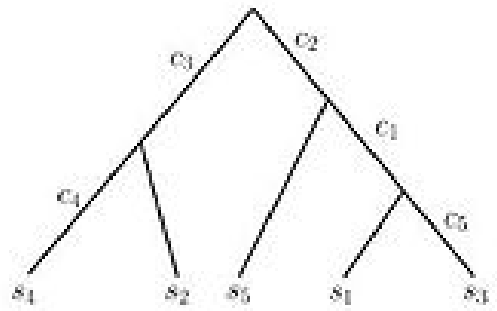
Easy Instance: all trees have same score.

Bounds on Parsimony Score

$M_1' =$

	c_2	c_1	c_3	c_5	c_4
s_1	1	1	0	0	0
s_2	0	0	1	0	0
s_3	1	1	0	1	0
s_4	0	0	1	0	1
s_5	1	0	0	0	0

$\longrightarrow \quad \mathcal{T}_1 =$



(from wikipedia.org)

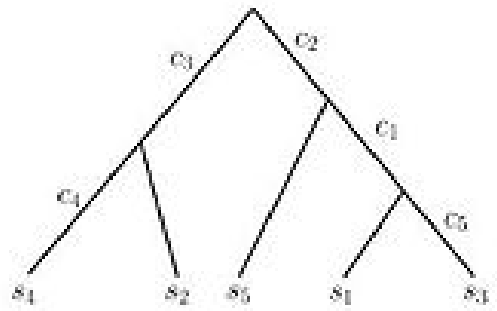
- If all characters were constant (i.e. all 'G'), than parsimony score is the same for all trees.

Bounds on Parsimony Score

$M_1' =$

	c_2	c_1	c_3	c_5	c_4
s_1	1	1	0	0	0
s_2	0	0	1	0	0
s_3	1	1	0	1	0
s_4	0	0	1	0	1
s_5	1	0	0	0	0

$\longrightarrow \quad \mathcal{T}_1 =$



(from wikipedia.org)

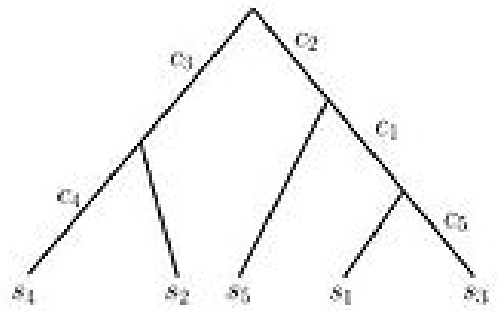
- If all characters were constant (i.e. all 'G'), than parsimony score is the same for all trees.
- Best (non-trivial) case: like taxa are 'grouped' together on the tree minimizing the number of changes to the $r-1$ where $r =$ `number of states`.

Bounds on Parsimony Score

$M_1' =$

	c_2	c_1	c_3	c_5	c_4
s_1	1	1	0	0	0
s_2	0	0	1	0	0
s_3	1	1	0	1	0
s_4	0	0	1	0	1
s_5	1	0	0	0	0

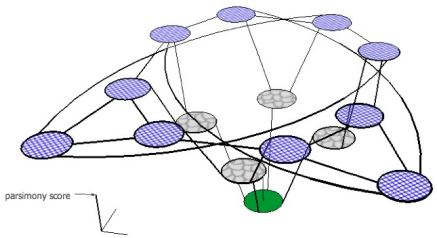
$\longrightarrow \quad \mathcal{T}_1 =$



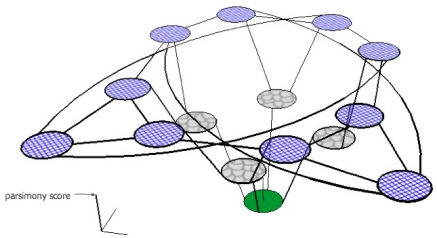
(from wikipedia.org)

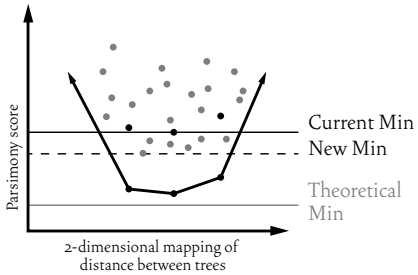
- If all characters were constant (i.e. all 'G'), than parsimony score is the same for all trees.
- Best (non-trivial) case: like taxa are 'grouped' together on the tree minimizing the number of changes to the $r-1$ where $r =$ `number of states`.
- Worst case: like taxa are scattered across the tree and many changes occur across the edges.

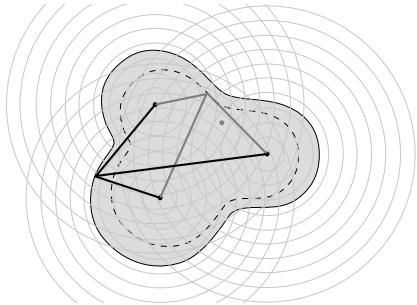
Compatible Characters



Compatible Characters





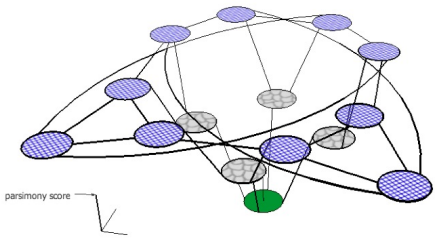


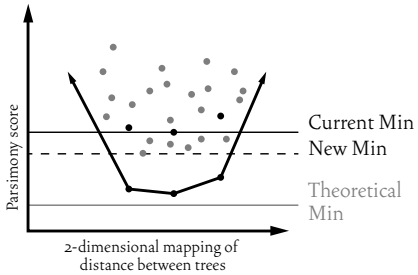
Ford, S, & Wheeler '14

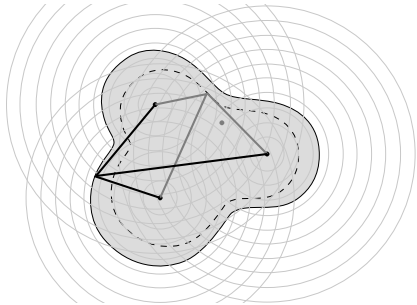
Simple observation: when the characters are compatible:

- The minimal scoring tree is the 'perfect phylogeny.'

Compatible Characters





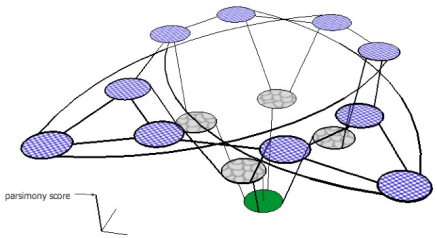


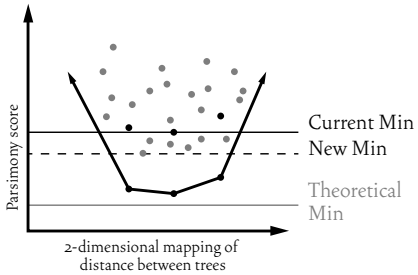
Ford, S, & Wheeler '14

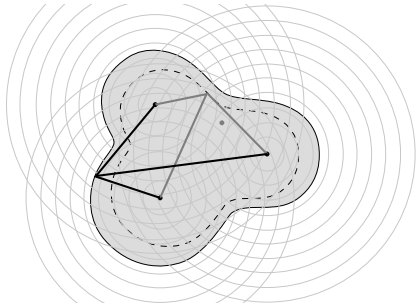
Simple observation: when the characters are compatible:

- The minimal scoring tree is the 'perfect phylogeny.'
- The score grows by at least i for each i "steps" taken (where steps are a relaxed Robinson-Foulds distance).

Compatible Characters







Ford, S, & Wheeler '14

Simple observation: when the characters are compatible:

- The minimal scoring tree is the 'perfect phylogeny.'
- The score grows by at least i for each i "steps" taken (where steps are a relaxed Robinson-Foulds distance).
- The sum of the bounds on compatible subsets of characters bounds the score across all the characters.

Metasiro

Metasiro americanus, Clouse & Wheeler '14

- *M. americanus* ('harvestmen') live in US south west.

Metasiro

Metasiro americanus, Clouse & Wheeler '14

- *M. americanus* ('harvestmen') live in US south west.
- Metasiro have poor dispersal and existed an exceptionally long time.

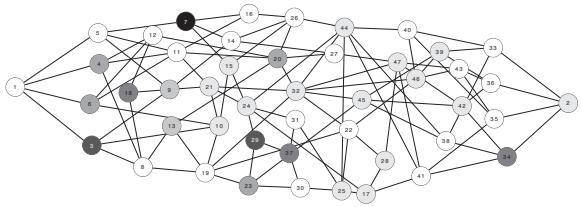
Metasiro

Metasiro americanus, Clouse & Wheeler '14

- *M. americanus* ('harvestmen') live in US south west.
- Metasiro have poor dispersal and existed an exceptionally long time.
- Used to test historical landmass movement hypothesis in phylogeography.

Results on Limiting the Search Space

Clouse & Wheeler ‘14



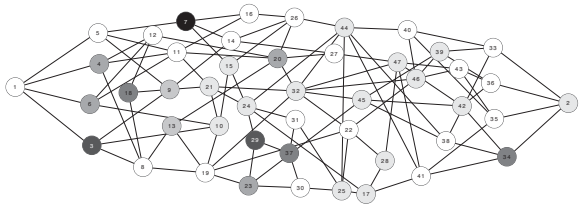
Ford, S, & Wheeler ‘14

Evaluated a *metasiro* data set from Clouse & Wheeler ‘14:

- Still an NP-hard problem but can reduce search space significantly

Results on Limiting the Search Space

Clouse & Wheeler ‘14



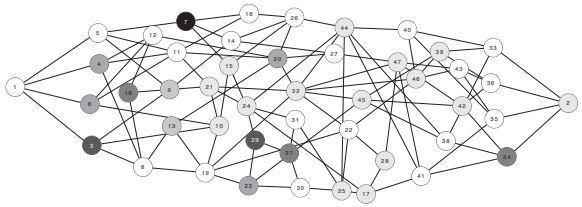
Ford, S, & Wheeler ‘14

Evaluated a *metasiro* data set from Clouse & Wheeler ‘14:

- Still an NP-hard problem but can reduce search space significantly
- 769 bp fragment of the mitochondrial gene cytochrome c oxidase subunit I (COI) sequenced.

Results on Limiting the Search Space

Clouse & Wheeler ‘14



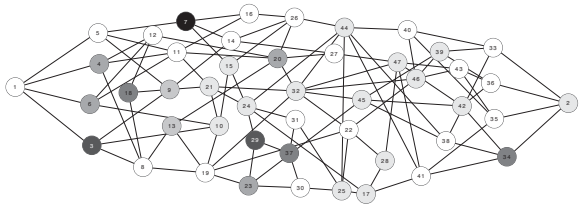
Ford, S, & Wheeler ‘14

Evaluated a *metasiro* data set from Clouse & Wheeler ‘14:

- Still an NP-hard problem but can reduce search space significantly
- 769 bp fragment of the mitochondrial gene cytochrome c oxidase subunit I (COI) sequenced.
- 62 taxa and 36 (out of 460) informative characters.

Results on Limiting the Search Space

Clouse & Wheeler ‘14



Ford, S, & Wheeler ‘14

Evaluated a *metasiro* data set from Clouse & Wheeler ‘14:

- Still an NP-hard problem but can reduce search space significantly
- 769 bp fragment of the mitochondrial gene cytochrome c oxidase subunit I (COI) sequenced.
- 62 taxa and 36 (out of 460) informative characters.
- Reduced to 57 unresolved trees, searched exhaustively.

Results on Limiting the Search Space

CI	number of taxa	actual size of tree space	PS_{t0} - M_t	number of anchor trees found	size of reduced search space	log reduction
0.97	14	7.91E+12	6	2	1.57E+06	-6.70
0.96	10	34459425	2	3	405	-4.93
0.95	18	6.33E+18	1	2	3.04E+04	-14.32
0.95	6	945	5	2	n/a	n/a
0.95	32	1.78E+42	2	2	4.98E+15	-26.55
0.93	39	1.31E+55	4	4	2.25E+39	-15.77
0.93	40	1.01E+57	18	2		
0.92	11	6.55E+08	24	2	n/a	n/a
0.92	16	6.19E+15	9	2	1.39E+13	-2.65
0.91	15	2.13E+14	16	2	1.28E+35	20.78
0.91	41	7.98E+58	6	2	7.42E+05	-53.03
0.89	13	3.16E+11	3	2	n/a	n/a
0.88	29	8.69E+36	60	2	n/a	n/a
0.88	15	2.13E+14	20	2	n/a	n/a
0.87	25	1.19E+30	3	3	9.36E+15	-14.11
0.87	19	2.22E+20	17	2	5.31E+17	-2.62
0.87	19	2.22E+20	8	2	4.57E+10	-9.69
0.86	20	8.20E+21	17	2	n/a	n/a
0.86	55	3.19E+86	28	2		
0.86	41	7.98E+58	8	2	9.99E+43	-14.90

Ford, S, & Wheeler '14

- Evaluated 600 datasets from TreeBase.

Results on Limiting the Search Space

CI	number of taxa	actual size of tree space	PS_{t0} - M_t	number of anchor trees found	size of reduced search space	log reduction
0.97	14	7.91E+12	6	2	1.57E+06	-6.70
0.96	10	34459425	2	3	405	-4.93
0.95	18	6.33E+18	1	2	3.04E+04	-14.32
0.95	6	945	5	2	n/a	n/a
0.95	32	1.78E+42	2	2	4.98E+15	-26.55
0.93	39	1.31E+55	4	4	2.25E+39	-15.77
0.93	40	1.01E+57	18	2		
0.92	11	6.55E+08	24	2	n/a	n/a
0.92	16	6.19E+15	9	2	1.39E+13	-2.65
0.91	15	2.13E+14	16	2	1.28E+35	20.78
0.91	41	7.98E+58	6	2	7.42E+05	-53.03
0.89	13	3.16E+11	3	2	n/a	n/a
0.88	29	8.69E+36	60	2	n/a	n/a
0.88	15	2.13E+14	20	2	n/a	n/a
0.87	25	1.19E+30	3	3	9.36E+15	-14.11
0.87	19	2.22E+20	17	2	5.31E+17	-2.62
0.87	19	2.22E+20	8	2	4.57E+10	-9.69
0.86	20	8.20E+21	17	2	n/a	n/a
0.86	55	3.19E+86	28	2		
0.86	41	7.98E+58	8	2	9.99E+43	-14.90

Ford, S, & Wheeler '14

- Evaluated 600 datasets from TreeBase.
- Still an NP-hard problem but can reduce search space significantly

Results on Limiting the Search Space

CI	number of taxa	actual size of tree space	PS_{t0} - M_t	number of anchor trees found	size of reduced search space	log reduction
0.97	14	7.91E+12	6	2	1.57E+06	-6.70
0.96	10	34459425	2	3	405	-4.93
0.95	18	6.33E+18	1	2	3.04E+04	-14.32
0.95	6	945	5	2	n/a	n/a
0.95	32	1.78E+42	2	2	4.98E+15	-26.55
0.93	39	1.31E+55	4	4	2.25E+39	-15.77
0.93	40	1.01E+57	18	2		
0.92	11	6.55E+08	24	2	n/a	n/a
0.92	16	6.19E+15	9	2	1.39E+13	-2.65
0.91	15	2.13E+14	16	2	1.28E+35	20.78
0.91	41	7.98E+58	6	2	7.42E+05	-53.03
0.89	13	3.16E+11	3	2	n/a	n/a
0.88	29	8.69E+36	60	2	n/a	n/a
0.88	15	2.13E+14	20	2	n/a	n/a
0.87	25	1.19E+30	3	3	9.36E+15	-14.11
0.87	19	2.22E+20	17	2	5.31E+17	-2.62
0.87	19	2.22E+20	8	2	4.57E+10	-9.69
0.86	20	8.20E+21	17	2	n/a	n/a
0.86	55	3.19E+86	28	2		
0.86	41	7.98E+58	8	2	9.99E+43	-14.90

Ford, S, & Wheeler '14

- Evaluated 600 datasets from TreeBase.
- Still an NP-hard problem but can reduce search space significantly
- Reduction highly dependent on number of anchor trees.

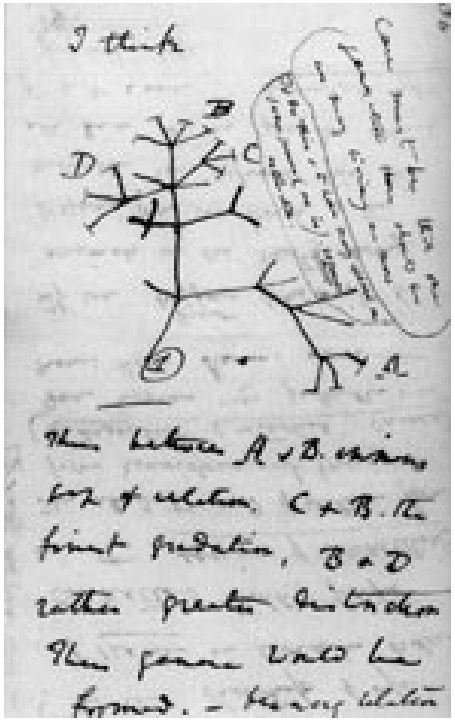
Results on Limiting the Search Space

CI	number of taxa	actual size of tree space	PS_{t0} - M_t	number of anchor trees found	size of reduced search space	log reduction
0.97	14	7.91E+12	6	2	1.57E+06	-6.70
0.96	10	34459425	2	3	405	-4.93
0.95	18	6.33E+18	1	2	3.04E+04	-14.32
0.95	6	945	5	2	n/a	n/a
0.95	32	1.78E+42	2	2	4.98E+15	-26.55
0.93	39	1.31E+55	4	4	2.25E+39	-15.77
0.93	40	1.01E+57	18	2		
0.92	11	6.55E+08	24	2	n/a	n/a
0.92	16	6.19E+15	9	2	1.39E+13	-2.65
0.91	15	2.13E+14	16	2	1.28E+35	20.78
0.91	41	7.98E+58	6	2	7.42E+05	-53.03
0.89	13	3.16E+11	3	2	n/a	n/a
0.88	29	8.69E+36	60	2	n/a	n/a
0.88	15	2.13E+14	20	2	n/a	n/a
0.87	25	1.19E+30	3	3	9.36E+15	-14.11
0.87	19	2.22E+20	17	2	5.31E+17	-2.62
0.87	19	2.22E+20	8	2	4.57E+10	-9.69
0.86	20	8.20E+21	17	2	n/a	n/a
0.86	55	3.19E+86	28	2		
0.86	41	7.98E+58	8	2	9.99E+43	-14.90

Ford, S, & Wheeler '14

- Evaluated 600 datasets from TreeBase.
- Still an NP-hard problem but can reduce search space significantly
- Reduction highly dependent on number of anchor trees.
- High consistency index (CI) empirically has the best reduction.

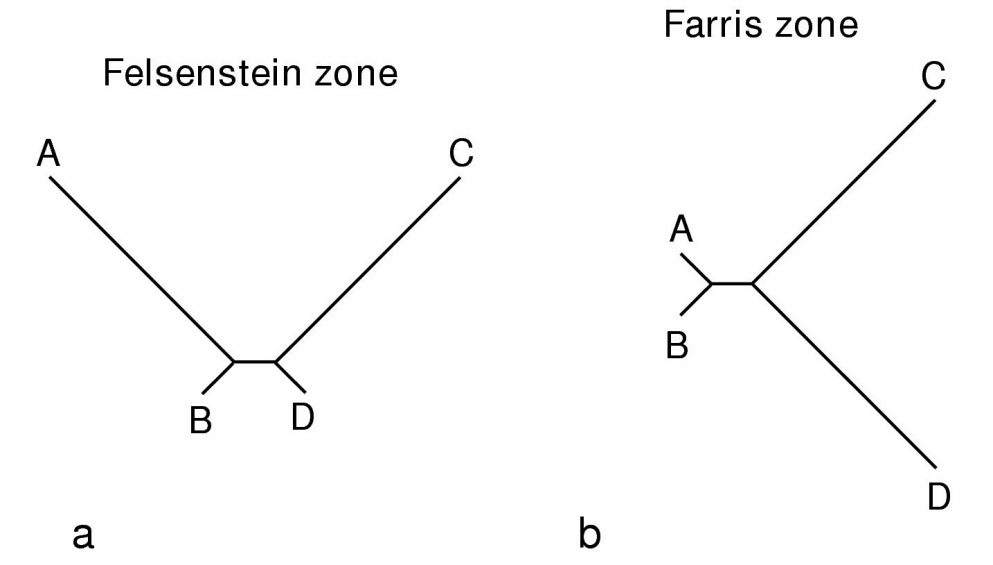
Outline



Charles Darwin, 1837

- Treespaces and Landscapes
- Metrics & Search
- Preprocessing to Improve Search
- Maximum Likelihood & Continuous Treespace
- When Trees are Not Enough....

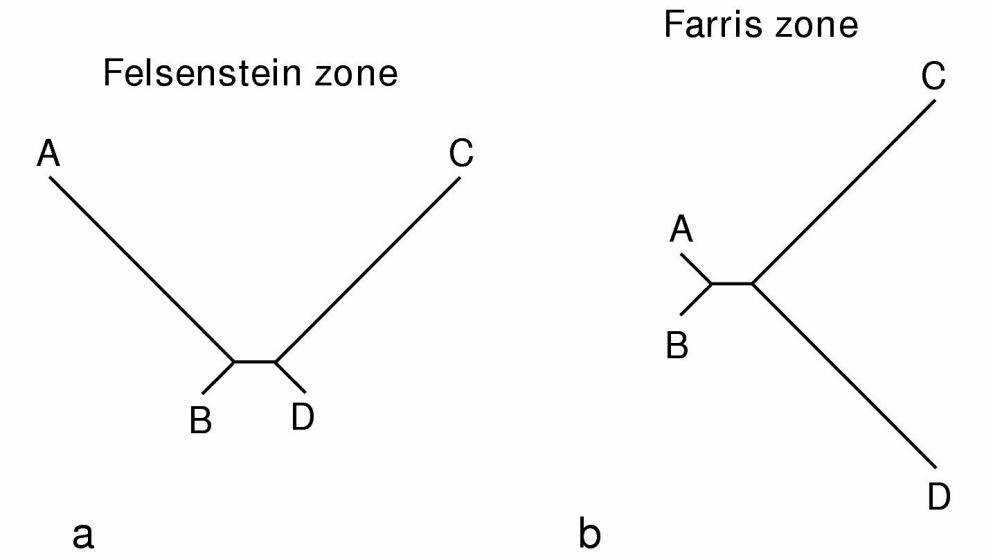
Maximum Likelihood Trees



Philippe *et al.*, '05

- Branch weights are part of the model.

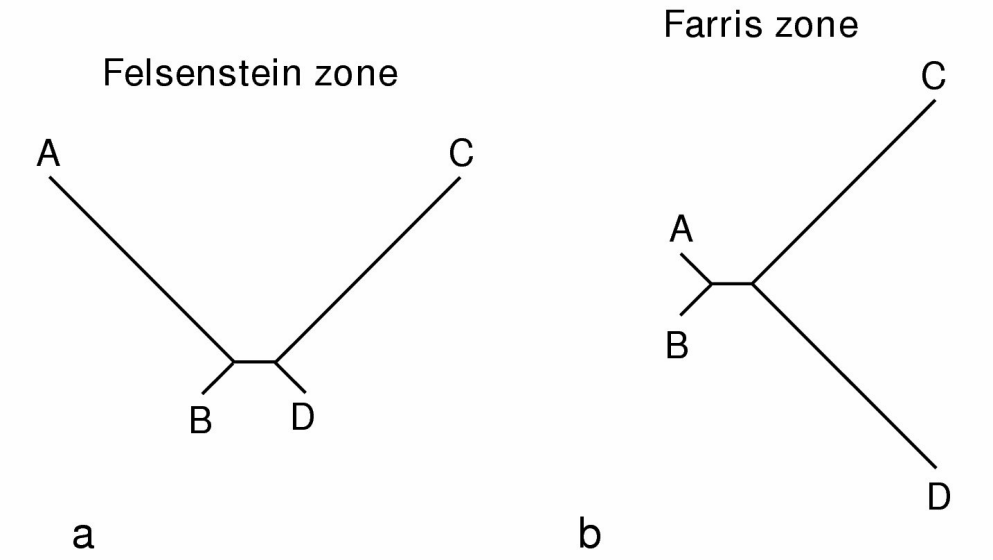
Maximum Likelihood Trees



Philippe *et al.*, '05

- Branch weights are part of the model.
- Indicated by length of edges in drawing.

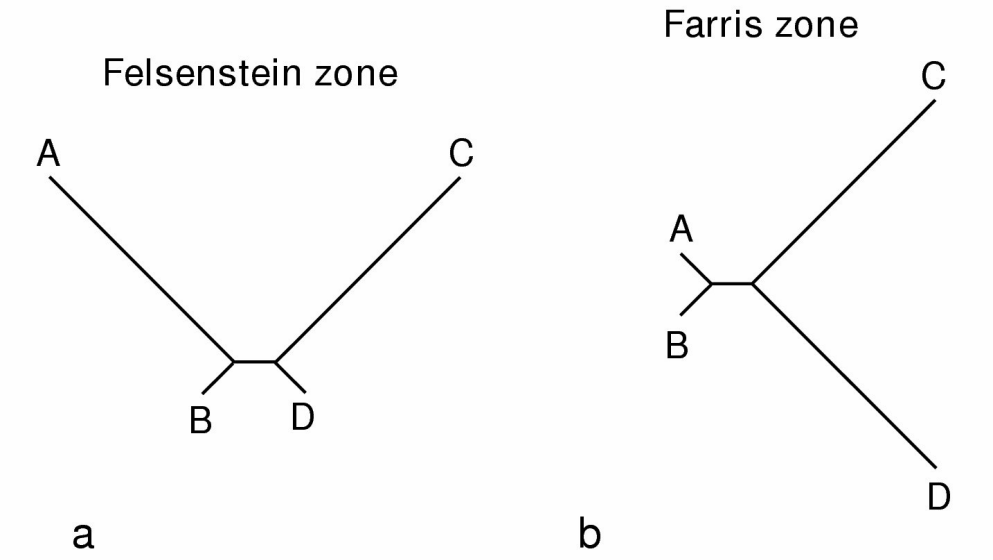
Maximum Likelihood Trees



Philippe *et al.*, '05

- Branch weights are part of the model.
- Indicated by length of edges in drawing.
- Two classic trees with same underlying topology.

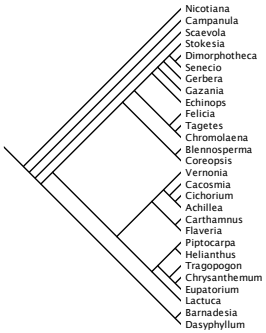
Maximum Likelihood Trees



Philippe *et al.*, '05

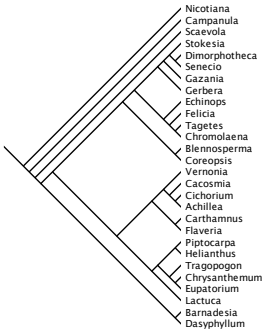
- Branch weights are part of the model.
- Indicated by length of edges in drawing.
- Two classic trees with same underlying topology.
- The metrics and search spaces above treat them as identical.

Popular Tree Metrics



Those based on tree rearrangements:

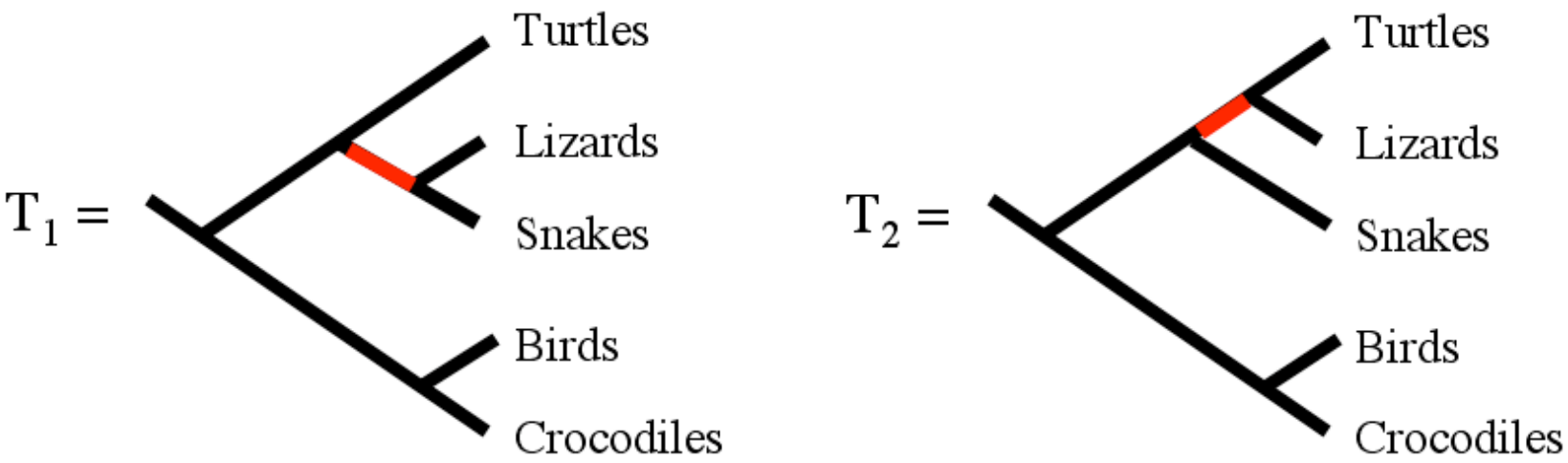
- Subtree Prune and Regraft (SPR)
- Tree Bisection and Reconnection (TBR)
- Nearest Neighbor Interchange (NNI)
- Used for Searching for Optimal Trees, NP-hard



Those based on comparing tree vectors:

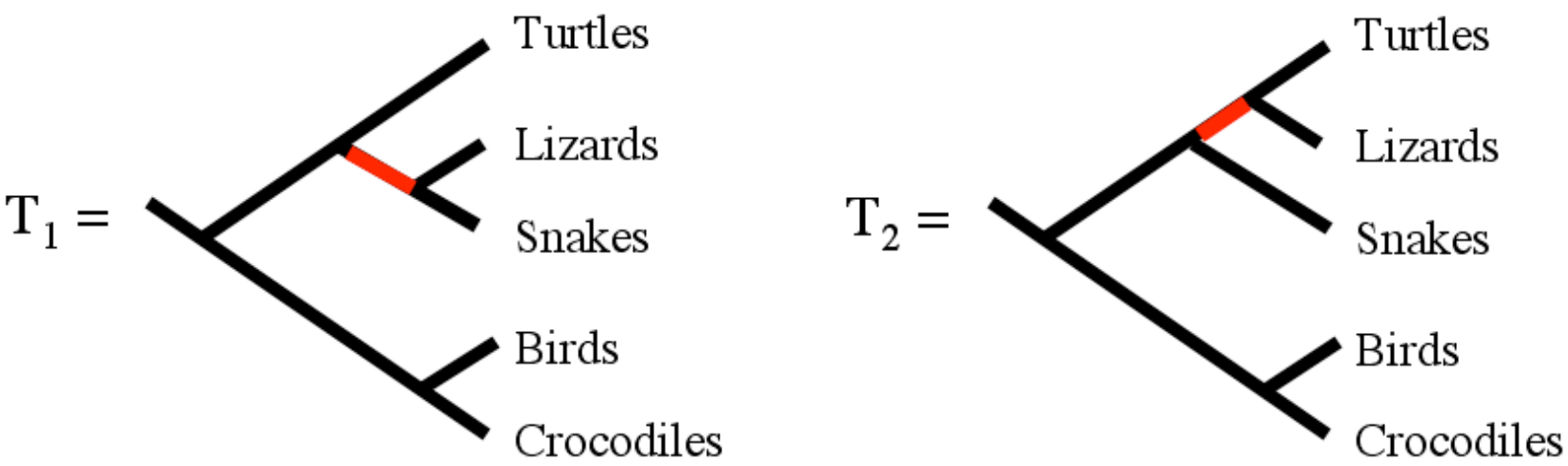
- Robinson-Foulds (RF)
- Rooted Triples (RT)
- Quartet Distance
- Billera-Holmes-Vogtmann (BHV or geodesic))
- Used for comparing trees, poly time

Robinson-Foulds Distance



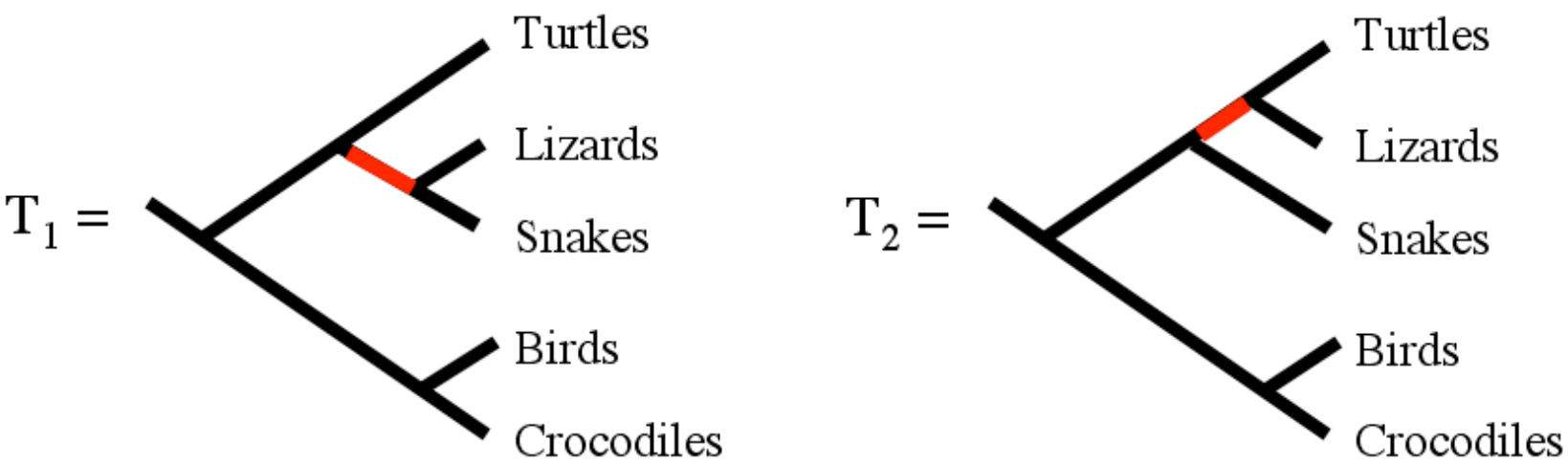
- The # of branches that occur in only one tree, or

Robinson-Foulds Distance



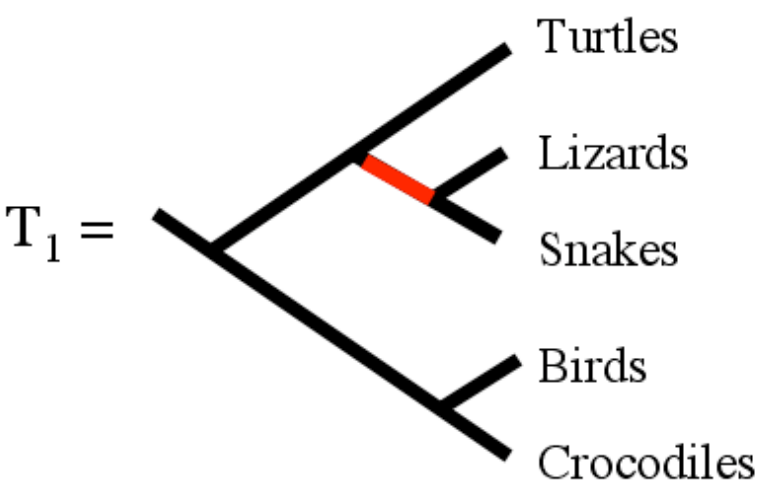
- The # of branches that occur in only one tree, or
- The size of the symmetric difference of the splits, or

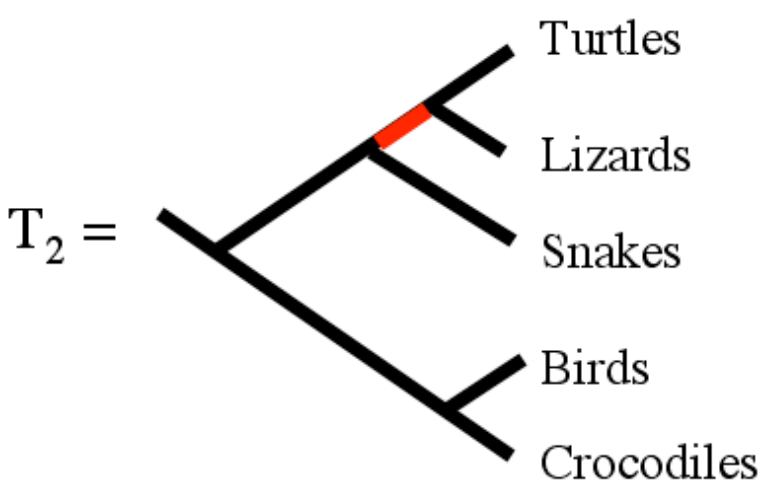
Robinson-Foulds Distance



- The # of branches that occur in only one tree, or
- The size of the symmetric difference of the splits, or
- The sum of the "false positives" and "false negatives."

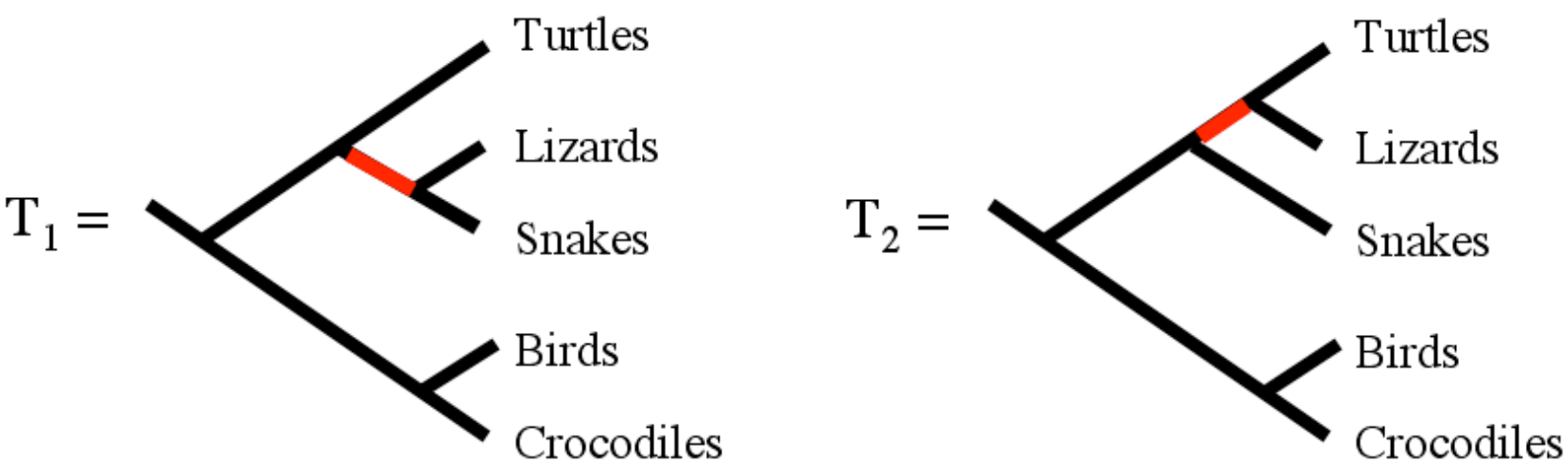
Robinson-Foulds Distance





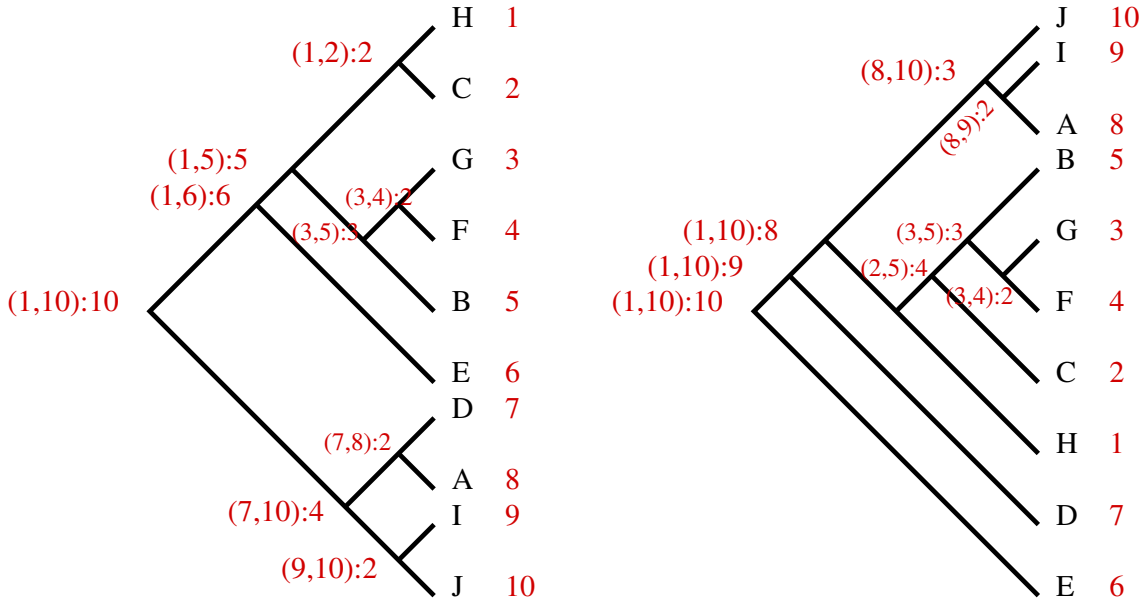
- Very popular

Robinson-Foulds Distance



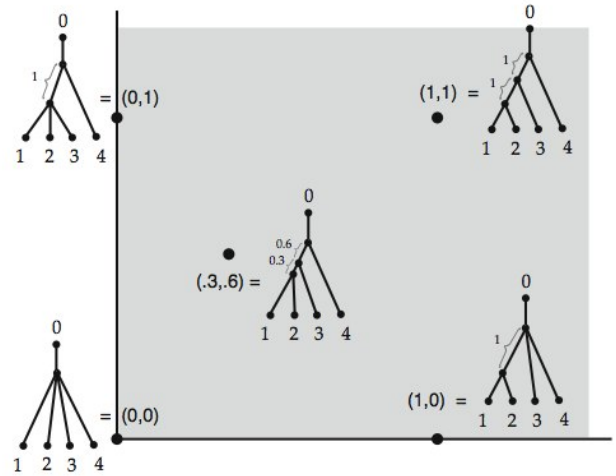
- Very popular
- Calculated in linear time, using Day's Algorithm ('85)

Applications



- Randomized $O(nt)$ for majority rule consensus (Amenta, Clarke, & S, WABI '03).
- Linear time processing of tree reduction rules (Bonet, S, Amenta, & Mahindru '06).

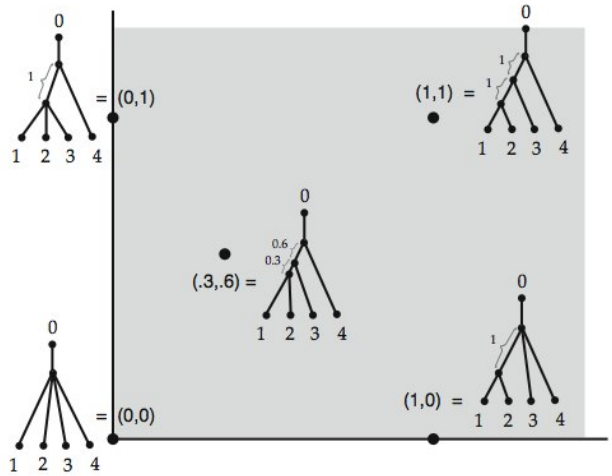
BHV Distance



Billera, Holmes, Vogtmann '01

- Billera, Holmes, and Vogtmann '01 have a continuous metric space of trees.

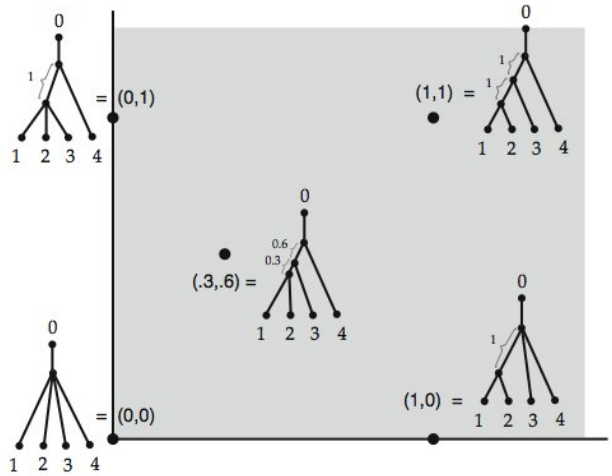
BHV Distance



Billera, Holmes, Vogtmann '01

- Billera, Holmes, and Vogtmann '01 have a continuous metric space of trees.
- View each split in a tree as a coordinate in the space.

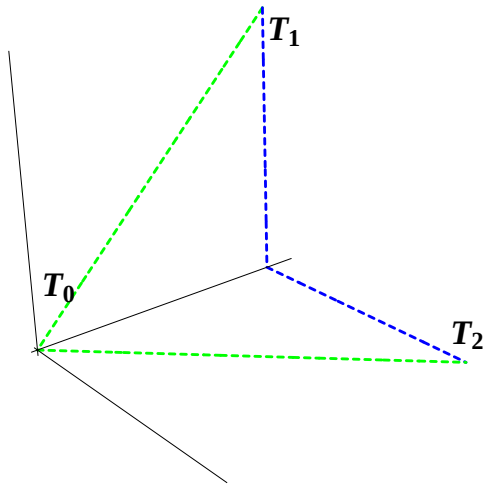
BHV Distance



Billera, Holmes, Vogtmann '01

- Billera, Holmes, and Vogtmann '01 have a continuous metric space of trees.
- View each split in a tree as a coordinate in the space.
- Identify edges of orthants to form space

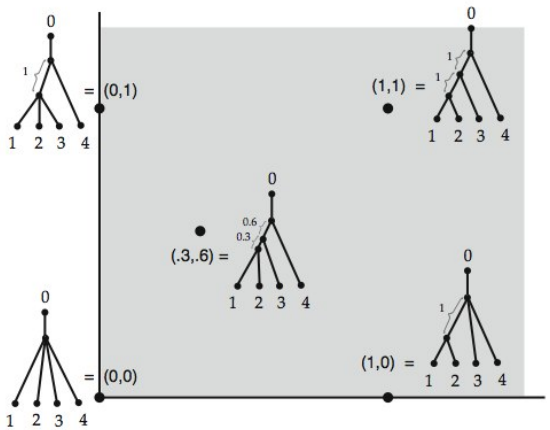
Tree Vectors



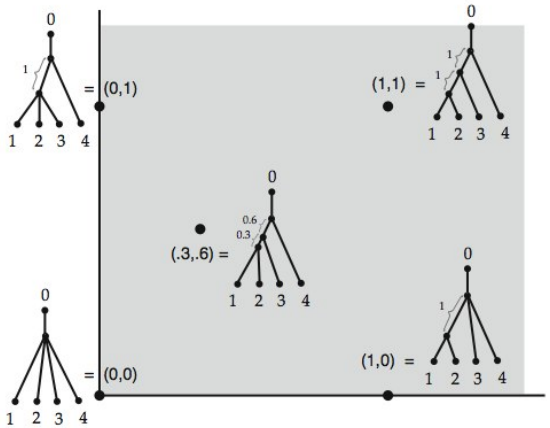
	$\frac{1}{2345}$	$\frac{2}{1345}$	$\frac{3}{1245}$	$\frac{4}{1235}$	$\frac{5}{1234}$	$\frac{12}{345}$	$\frac{13}{245}$	$\frac{14}{235}$	$\frac{15}{234}$	$\frac{23}{145}$	$\frac{24}{135}$	$\frac{25}{134}$	$\frac{34}{125}$	$\frac{35}{124}$	$\frac{45}{123}$
$T_0 = (1,2,3,4,5)$	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
$T_1 = ((1,2),(3,(4,5))$	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
$T_2 = ((1,2),(4,(3,5))$	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0

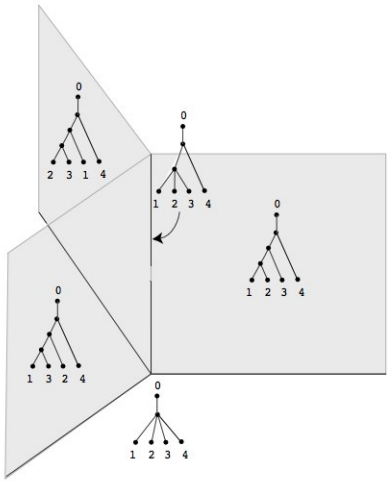
$d_{RF}(T_0, T_1) = 2$
$d_{BHV}(T_0, T_1) = \sqrt{2}$
$d_{RF}(T_1, T_2) = 2$
$d_{BHV}(T_1, T_2) = 2$

Identify Edges of Orthants

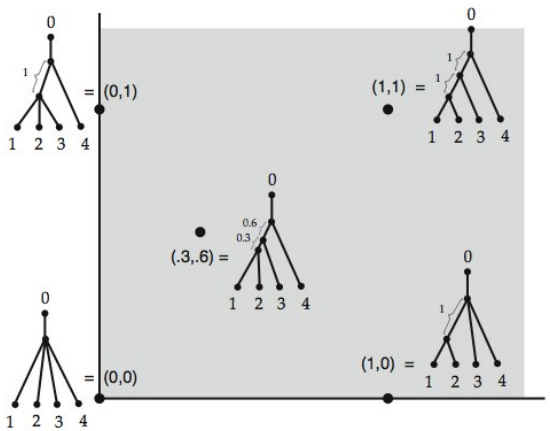


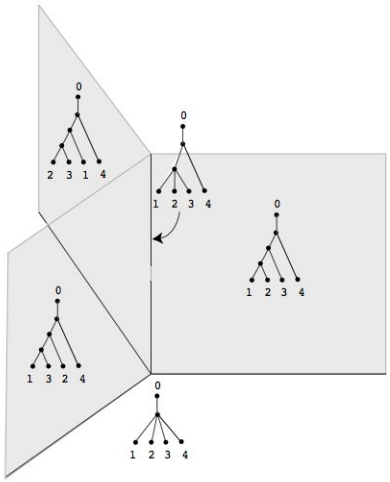
Identify Edges of Orthants

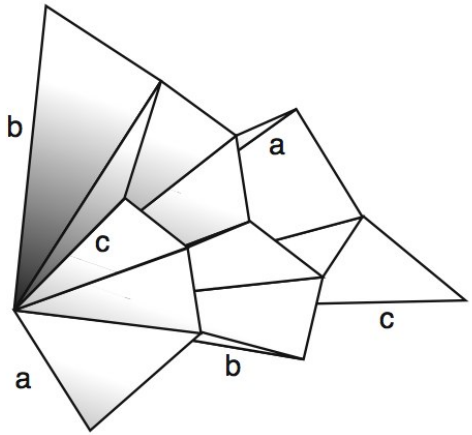




Identify Edges of Orthants

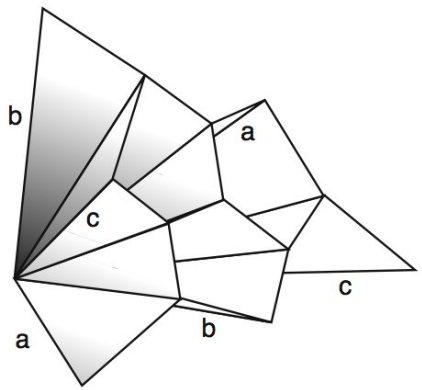






(All images from Billera, Holmes, Vogtmann '01)

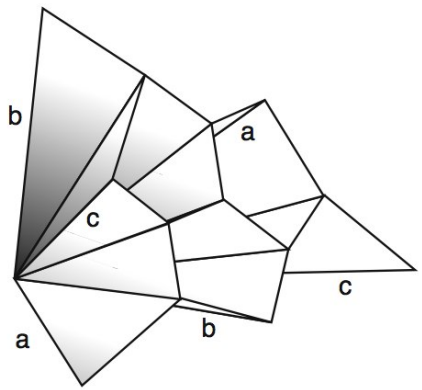
Geodesic Distance



Billera, Holmes, Vogtmann '01

- The geodesic is a shortest path on the surface between two points.

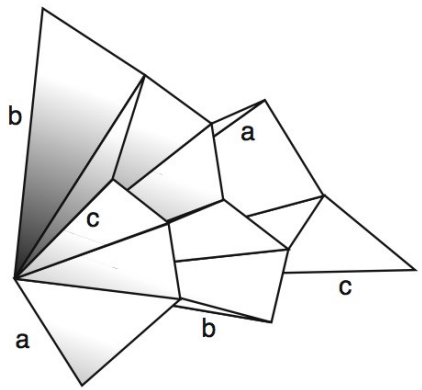
Geodesic Distance



Billera, Holmes, Vogtmann '01

- The geodesic is a shortest path on the surface between two points.
- Deep mathematics used to show the geodesic is a distance. (this negatively curved space is CAT(0)).

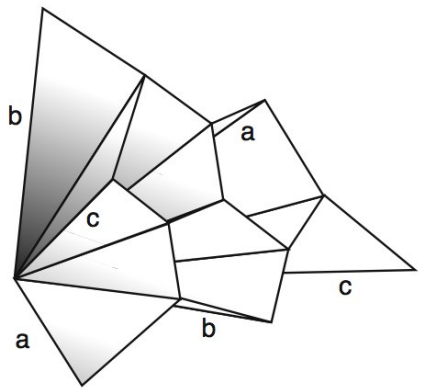
Geodesic Distance



Billera, Holmes, Vogtmann '01

- The geodesic is a shortest path on the surface between two points.
- Deep mathematics used to show the geodesic is a distance. (this negatively curved space is CAT(0)).
- Polynomial time ($O(n^4)$) to compute (Owen & Provon, 2011).

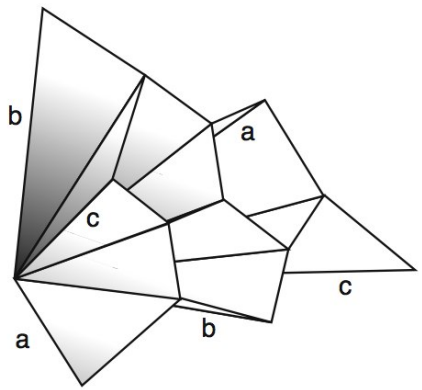
Geodesic Distance



Billera, Holmes, Vogtmann '01

- The geodesic is a shortest path on the surface between two points.
- Deep mathematics used to show the geodesic is a distance. (this negatively curved space is CAT(0)).
- Polynomial time ($O(n^4)$) to compute (Owen & Provon, 2011).
- Linear time approximation (Amenta, Godwin, Postarnakevich, and S '07).

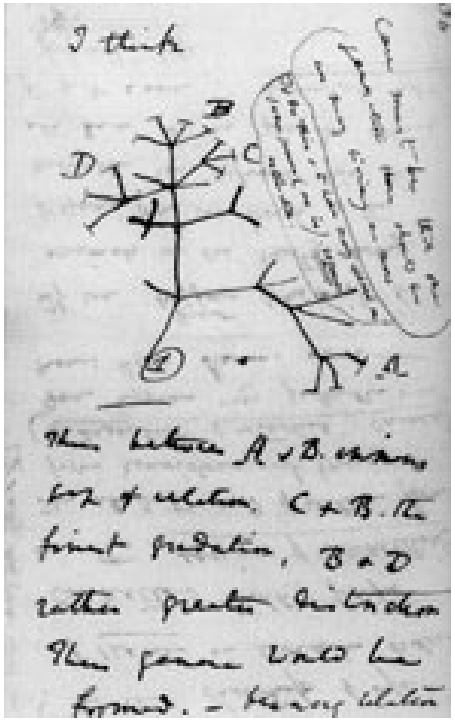
Geodesic Distance



Billera, Holmes, Vogtmann '01

- The geodesic is a shortest path on the surface between two points.
- Deep mathematics used to show the geodesic is a distance. (this negatively curved space is CAT(0)).
- Polynomial time ($O(n^4)$) to compute (Owen & Provon, 2011).
- Linear time approximation (Amenta, Godwin, Postarnakevich, and S '07).
- Averages computed via Freéchet means (Miller, Owen, & Proven '12, Bacák '12)

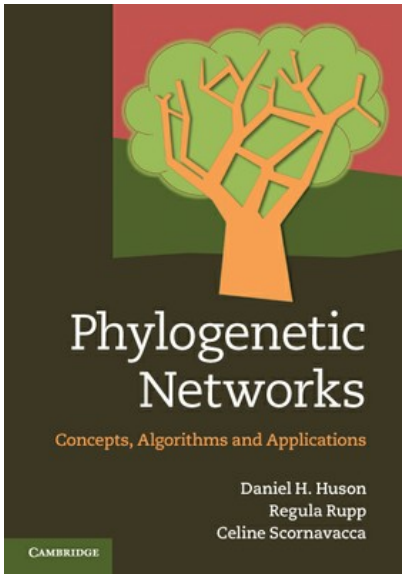
Outline



Charles Darwin, 1837

- Treespaces and Landscapes
- Metrics & Search
- Preprocessing to Improve Search
- Maximum Likelihood & Continuous Treespace
- When Trees are Not Enough....

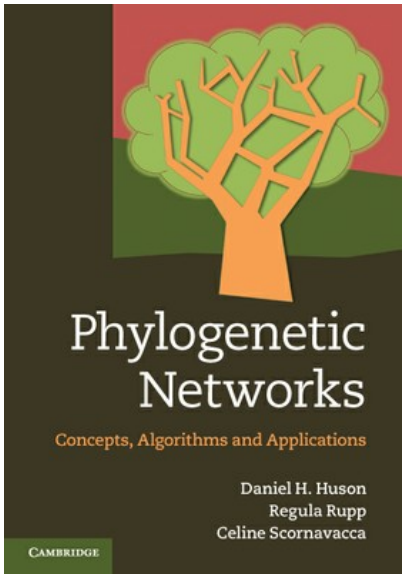
When Trees Are Not Enough



Huson, Rupp, Scornavacca '10

- Underlying assumption above:
  Evolution is tree-like.

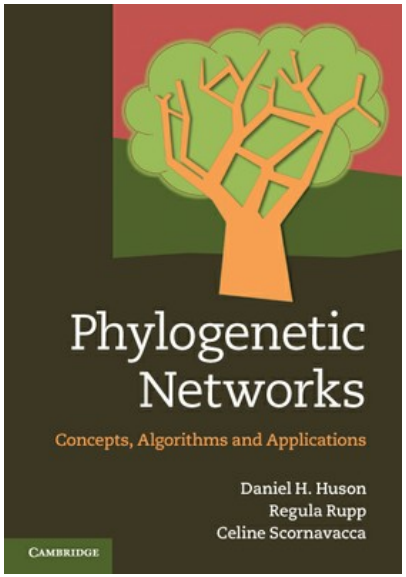
When Trees Are Not Enough



Huson, Rupp, Scornavacca '10

- Underlying assumption above: *Evolution is tree-like.*
- In many cases, evolution produces a more tangled structure.

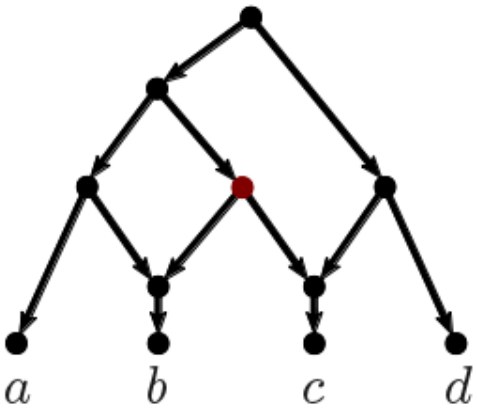
When Trees Are Not Enough



Huson, Rupp, Scornavacca '10

- Underlying assumption above: *Evolution is tree-like.*
- In many cases, evolution produces a more tangled structure.
- Networks (leaf-labeled, directed acyclic graphs) are used to model reticulate evolution.

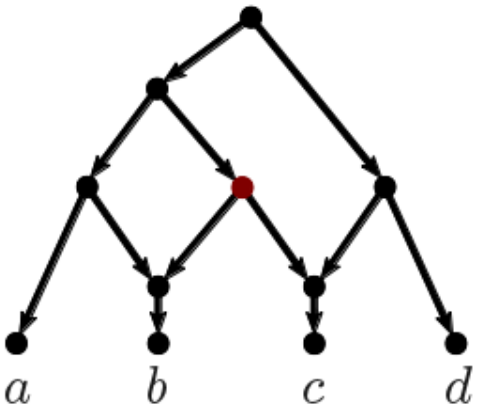
Can't see the trees for the . . . network



Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N

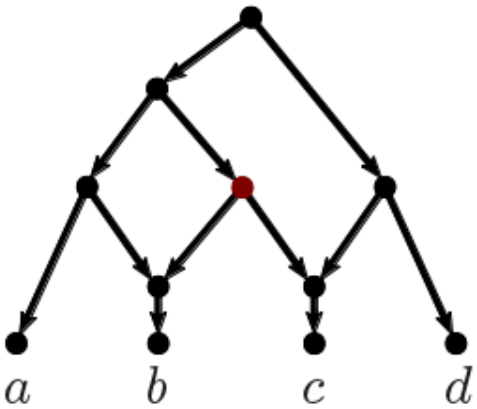
Can't see the trees for the . . . network



Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X),

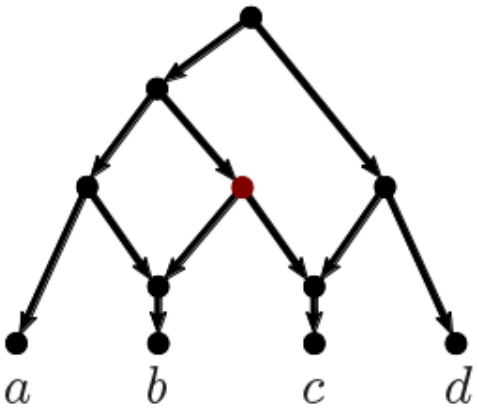
Can't see the trees for the . . . network



Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X), how many unique trees are displayed by N?

Can't see the trees for the . . . network

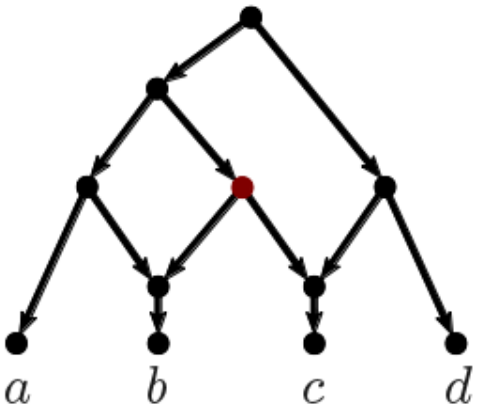


Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X), how many unique trees are displayed by N?

  Linz, S, Semple, 2013: It's #P-complete.

Can't see the trees for the . . . network



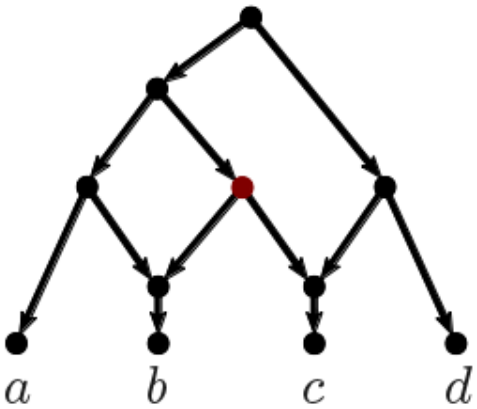
Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X), how many unique trees are displayed by N?

  Linz, S, Semple, 2013: It's #P-complete.

- Nakhleh's Enumeration Challenge II: Counting nets (up to isomorphism or other equivalence).

Can't see the trees for the . . . network



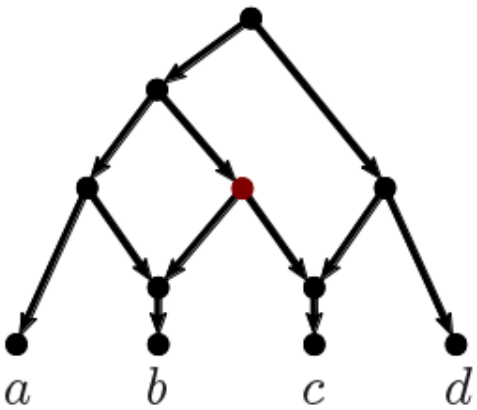
Leo van Iersel, 2013

- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X), how many unique trees are displayed by N?

  Linz, S, Semple, 2013: It's #P-complete.

- Nakhleh's Enumeration Challenge II: Counting nets (up to isomorphism or other equivalence). What is the number of unique (up to digraph isomorphism) rooted phylogenetic networks on n taxa and with h reticulation nodes?

Can't see the trees for the . . . network



Leo van Iersel, 2013

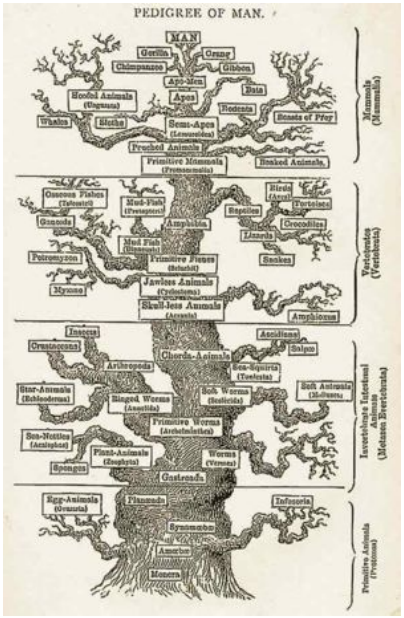
- Nakhleh's Enumeration Challenge I: Given a phylogenetic X–network N (rooted binary DAG leaf labeled bijectively by set X), how many unique trees are displayed by N?

  Linz, S, Semple, 2013: It's #P-complete.

- Nakhleh's Enumeration Challenge II: Counting nets (up to isomorphism or other equivalence). What is the number of unique (up to digraph isomorphism) rooted phylogenetic networks on n taxa and with h reticulation nodes?

  McDiarmid, Semple, Welsh, 2015: $2^{\gamma n \log n + O(n)}$, where γ is $\frac{3}{2}$ for general networks, and $\frac{5}{4}$ for tree-child & normal networks.

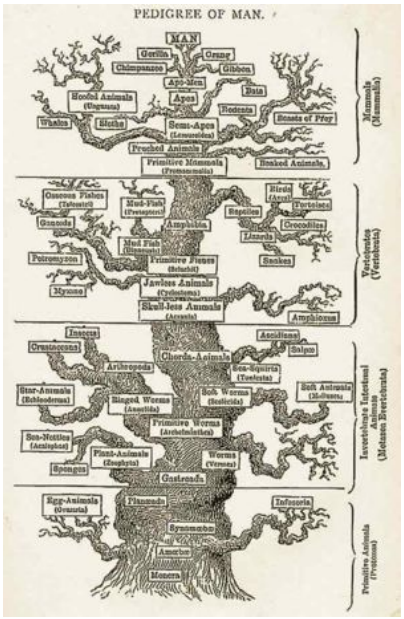
Summary



Haeckel's Tree of Life, 1879

- Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.

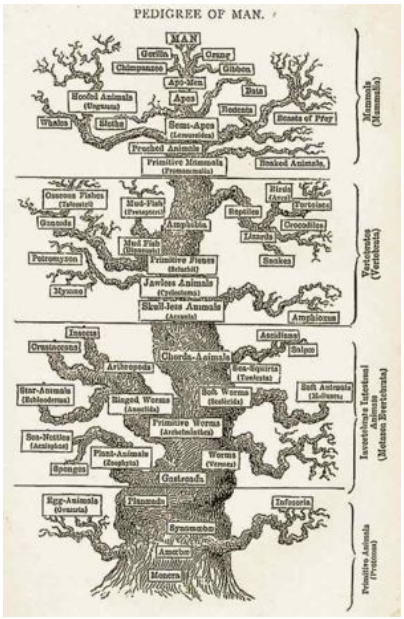
Summary



Haeckel's Tree of Life, 1879

- Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.
- Explosion of data overwhelms search techniques: Sampling 10 million trees is insignificant when 10^{200} trees.

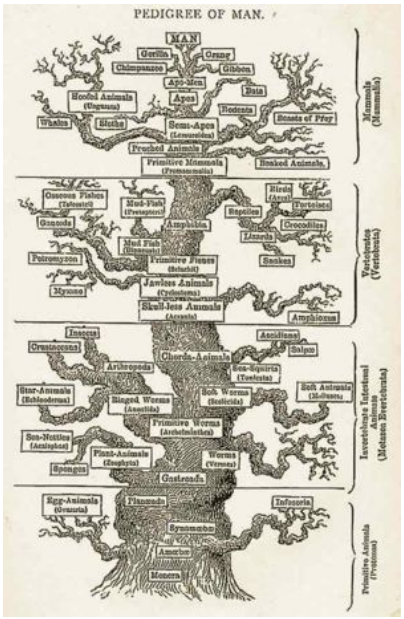
Summary



Haeckel's Tree of Life, 1879

- Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.
- Explosion of data overwhelms search techniques: Sampling 10 million trees is insignificant when 10^{200} trees.
- Optimality criteria's NP-hardness comes from seemingly random data.

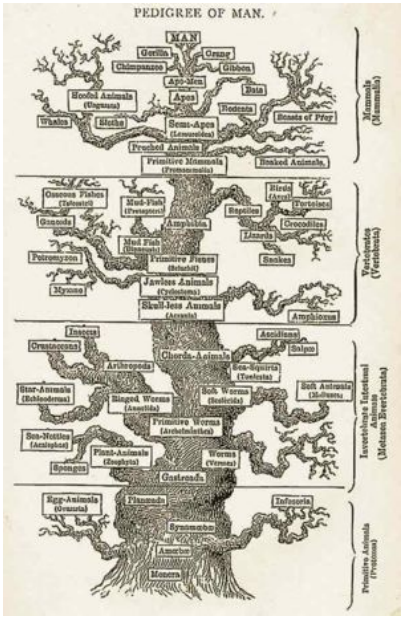
Summary



Haeckel's Tree of Life, 1879

- Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.
- Explosion of data overwhelms search techniques: Sampling 10 million trees is insignificant when 10^{200} trees.
- Optimality criteria's NP-hardness comes from seemingly random data.
- But biology is not random. The processes create identifiable patterns and easy instances.

Summary



Haeckel's Tree of Life, 1879

- Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.
- Explosion of data overwhelms search techniques: Sampling 10 million trees is insignificant when 10^{200} trees.
- Optimality criteria's NP-hardness comes from seemingly random data.
- But biology is not random. The processes create identifiable patterns and easy instances.
- Better understanding of the underlying structure of treespace can improve the search for optima.

Treespace Working Group

A team of students in mathematics, computer science, and biology contributed to this work:

> *Ann Marie Alcocer, Kadian Brown, Alan Caceres, Juan Castillo, Efrain Colon, Samantha Daley, John De Jesus, Eric Ford, Kevaughn Gordon, Kaitlin Hansen, Michael Hintze, Daniele Ippolito, Jinnie Lee, Ling Li, Joan Marc, Oliver Mendez, Diquan Moore, Daniel Packer, and Rachel Spratt.*

Acknowledgments

- The Simons Foundation for collaboration & travel funding,
- The US National Science Foundation for their generous support, and
- The New York Louis Stokes Alliance for Minority Participation in Research for student funding.