Finding Optimal Phylogenetic Trees

Katherine St. John

City University of New York
American Museum of Natural History

23 June 2015
Outline

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

Charles Darwin, 1837
Analogy: Find the Highest Point
Analogy: Find the Highest Point

Sampling:

- Choose 1000 random points.
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.
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.
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
Sampling only on the island misses peaks elsewhere.
Goal: Find the point with the optimal score
**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,...
Local Search Techniques

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

Many variations on the theme: branch-and-bound, MCMC, genetic algorithms,...
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)
**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
**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

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

**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?

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

\[
\text{# of trees} = 1 \cdot 3 \cdot 5 \cdots (2^n - 5) = (2^n - 5)!! \\
\sim \frac{1}{\sqrt{\pi}} \cdot 2^n \cdot 2^n \cdot 2^n - 2^n \cdot 2^n - 5 \cdot 2^n - 5 \cdot 2^n - 5 \cdot 2^n - 5 \\
\sim \frac{1}{2} \cdot 2^\frac{1}{2} \cdot (2e)^n.
\]

(For \( n \geq 50 \), \( \exists \) more possible tree topologies than there are atoms in the universe.)
Schröder, 1870 (see Semple & Steel, 2003):

\[
\text{# of trees} = 1 \cdot 3 \cdot 5 \cdots \cdots (2n - 5) \\
= (2n - 5)!! \\
\sim \frac{1}{\sqrt{\Pi}} 2^{n-2}n!n^{-\frac{5}{2}} \\
\sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^nn^{n-2}
\]
Schröder, 1870 (see Semple & Steel, 2003):

\[
\# \text{ of trees} = 1 \cdot 3 \cdot 5 \cdots (2n - 5) \\
= (2n - 5)!! \\
\sim \frac{1}{\sqrt{\pi}} 2^{n-2} n! n^{-\frac{5}{2}} \\
\sim \frac{1}{2\sqrt{2}} (\frac{2}{e})^n n^{n-2}
\]

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

How many taxa?

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

David Hillis, 2002
How Many Trees?

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

How many taxa?

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

David Hillis, 2002
(For \( n \geq 50 \), \( \exists \) more possible tree topologies than there are atoms in the universe.)

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

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

David Hillis, 2002
Goal: Find the tree with the optimal score
Goal: Find the tree with the optimal score

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

Local search techniques prevail:
  ▶ Begin with a tree
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)
**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
**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,...
Given a set of organisms, 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
  - Maximum Likelihood: find most likely tree (with respect to a model of evolution)
Given a set of organisms, which tree is optimal?

- Two standard criteria for optimality:
  - **Maximum Parsimony**: find tree with fewest changes
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)
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)
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).
Find the tree that can explain the observed sequences with a minimal number of substitutions.
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.
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.

![Diagram of a phylogenetic tree with sequences ACCCT, AACGA, GACGT, AACGT, GACGT, GGCCTA]
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.

```
ACCCCT

GACGT

AACGT

AACGA

G{A,G}CG?

GACGT

GGCGA
```
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.

![Diagram of phylogenetic tree with sequences ACCCT and AACGA at the root, and G{A,G}CG? at the terminal nodes. The internal nodes have sequences GACGT and AACGT.]
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.

![Diagram of a phylogenetic tree with sequences and labels]
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.
Given sequences for leaves, find tree with minimal parsimony score:

- ACCCT
- AACGA

- GACGT
- AACGT
- GACGT
- GGCGA

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

NASA Blue Marble Bathymetry
Analogy: Parsimony

Find the lowest point.

NASA Blue Marble Bathymetry
For every $n$, treespace is the space of all phylogenetic trees on a $n$ taxa, under a fixed distance metric.
A treespace with assigned scores is often called a landscape.
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?

The phylogeny problem is that of finding those trees which optimise some function of the input data. We may end up with several trees optimising some optimality criterion (say parsimony) and a completely different set of trees optimising another criterion. There are hundreds of phylogenetic optimality criteria! This is a problem, but it's not the one in which I'm interested today.

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). If very smooth, 'hill climbing' will work well. If very rugged, need more sophisticated searches that use the underlying structure of the space.
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).

If very smooth, ‘hill climbing’ will work well.

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

If very smooth, ‘hill climbing’ will work well.

If very rugged, need more sophisticated searches that use the underlying structure of the space.
Different metrics yield different neighbors: places you can reach in 10 minutes from Grand Central Station walking versus transit
Adjusting Search Space

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

The same data, organized by different tree metrics.
Attraction Basins

resalliance.org
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.
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

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

Charles Darwin, 1837
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)
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.
**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.”
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.
$(100)$: Does shrinking common subchains in trees preserve SPR distance?

(Implies fixed parameter tractability.)
($100)$: Does shrinking common subchains in trees preserve SPR distance?

(Implies fixed parameter tractability.)

Bordewich & Semple ‘05: Yes, for rooted trees.
($100)$: 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?

(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?

(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?

(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 is the minimal number of moves that transforms one tree into the other.
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 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)
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).
An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of un-rooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.
Bryant’s Challenge: Walking Through Trees

An **NNI-walk** is a sequence $T_1, T_2, \ldots, T_k$ of un-rooted 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

An **NNI-walk** is a sequence $T_1, T_2, \ldots, T_k$ of un-rooted 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 $T$?
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)
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 $t + \Theta(n^2)$. 
For every $n$, treespace is the space of all phylogenetic trees on a $n$ taxa, under a fixed metric.
For every $n$, `treespace` is the space of all phylogenetic trees on $n$ taxa, under a fixed metric.

The diameter of neighborhoods varies by metric:

- NNI: Li, Tromp & Zhang '96
- SPR & TBR: Atkins & McDiarmid '15
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:

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>$n = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNI</td>
<td>$\Theta(n \log n)$</td>
<td>4</td>
</tr>
<tr>
<td>SPR</td>
<td>$n - \Theta(\sqrt{n})$</td>
<td>2</td>
</tr>
<tr>
<td>TBR</td>
<td>$n - \Theta(\sqrt{n})$</td>
<td>2</td>
</tr>
</tbody>
</table>

NNI: Li, Tromp & Zhang ‘96

SPR & TBR: Atkins & McDiarmid ‘15
For every $n$, treespace is the space of all phylogenetic trees on a $n$ taxa, under a fixed metric.
For every \( n \), treespace is the space of all phylogenetic trees on \( n \) taxa, under a fixed metric.

The size of neighborhoods varies by metric (Allen & Steel ‘01):

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):

<table>
<thead>
<tr>
<th>Metric</th>
<th>General Expression</th>
<th>Value for $n = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNI</td>
<td>$2n - 6$</td>
<td>4</td>
</tr>
<tr>
<td>SPR</td>
<td>$2(n - 3)(2n - 7)$</td>
<td>12</td>
</tr>
<tr>
<td>TBR</td>
<td>$&lt; (2n - 3)(n - 3)^2$</td>
<td>12</td>
</tr>
</tbody>
</table>
Outline

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

Charles Darwin, 1837
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.

K. St. John (CUNY & AMNH)
Preprocessing the Data: Finding Easy Instances

Identical Sequences

A  GTTAGAAGCGGCGCCAGCGAC...
B  GTTAGAAGCGGCGCCAGCGAC...
C  GTTAGAAGCGGCGCCAGCGAC...
D  GTTAGAAGCGGCGCCAGCGAC...
E  GTTAGAAGCGGCGCCAGCGAC...
F  GTTAGAAGCGGCGCCAGCGAC...

945 Rooted Trees on 6 Leaves
Preprocessing the Data: Finding Easy Instances

Identical Sequences

A  GTTAGAAGCGCGCCAGCGAC...
B  GTTAGAAGCGCGCCAGCGAC...
C  GTTAGAAGCGCGCCAGCGAC...
D  GTTAGAAGCGCGCCAGCGAC...
E  GTTAGAAGCGCGCCAGCGAC...
F  GTTAGAAGCGCGCCAGCGAC...

945 Rooted Trees on 6 Leaves

Easy Instance: all trees have same score.
If all characters were constant (i.e. all ‘G’), then parsimony score is the same for all trees.
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.
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 \( r = \text{number of states} \).

**Worst case:** like taxa are scattered across the tree and many changes occur across the edges.
Compatible Characters

- Theoretical Min
- Current Min
- New Min
- Parsimony score
- 2-dimensional mapping of distance between trees

Simple observation: when the characters are compatible:
- The minimal scoring tree is the 'perfect phylogeny.'
- The score grows by at least $i$ for each "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.
Compatible Characters

Simple observation: when the characters are compatible:

- The minimal scoring tree is the ‘perfect phylogeny.’
Compatible Characters

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).

Ford, S, & Wheeler ‘14
Compatible Characters

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.

Ford, S, & Wheeler ‘14
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

*Metasisro 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

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

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

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.
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

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

Ford, S, & Wheeler '14

- Evaluated 600 datasets from TreeBase.
Results on Limiting the Search Space

<table>
<thead>
<tr>
<th>CI</th>
<th>number of taxa</th>
<th>actual size of tree space</th>
<th>$P_{S_0} - M_i$</th>
<th>number of anchor trees found</th>
<th>size of reduced search space</th>
<th>log reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97</td>
<td>14</td>
<td>7.91E+12</td>
<td>6</td>
<td>2</td>
<td>1.57E+06</td>
<td>-6.70</td>
</tr>
<tr>
<td>0.96</td>
<td>10</td>
<td>34459425</td>
<td>2</td>
<td>3</td>
<td>405</td>
<td>-4.93</td>
</tr>
<tr>
<td>0.95</td>
<td>18</td>
<td>6.33E+18</td>
<td>1</td>
<td>2</td>
<td>3.04E+04</td>
<td>-14.32</td>
</tr>
<tr>
<td>0.95</td>
<td>6</td>
<td>945</td>
<td>5</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.95</td>
<td>32</td>
<td>1.78E+42</td>
<td>2</td>
<td>2</td>
<td>4.98E+15</td>
<td>-26.55</td>
</tr>
<tr>
<td>0.93</td>
<td>39</td>
<td>1.31E+55</td>
<td>4</td>
<td>4</td>
<td>2.25E+39</td>
<td>-15.77</td>
</tr>
<tr>
<td>0.93</td>
<td>40</td>
<td>1.01E+57</td>
<td>18</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>11</td>
<td>6.55E+08</td>
<td>24</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>16</td>
<td>6.19E+15</td>
<td>9</td>
<td>2</td>
<td>1.39E+13</td>
<td>-2.65</td>
</tr>
<tr>
<td>0.91</td>
<td>15</td>
<td>2.13E+14</td>
<td>16</td>
<td>2</td>
<td>1.28E+35</td>
<td>20.78</td>
</tr>
<tr>
<td>0.91</td>
<td>41</td>
<td>7.98E+58</td>
<td>6</td>
<td>2</td>
<td>7.42E+05</td>
<td>-53.03</td>
</tr>
<tr>
<td>0.89</td>
<td>13</td>
<td>3.16E+11</td>
<td>3</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>29</td>
<td>8.69E+36</td>
<td>60</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>15</td>
<td>2.13E+14</td>
<td>20</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.87</td>
<td>25</td>
<td>1.19E+30</td>
<td>3</td>
<td>3</td>
<td>9.36E+15</td>
<td>-14.11</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>17</td>
<td>2</td>
<td>5.31E+17</td>
<td>-2.62</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>8</td>
<td>2</td>
<td>4.57E+10</td>
<td>-9.69</td>
</tr>
<tr>
<td>0.86</td>
<td>20</td>
<td>8.20E+21</td>
<td>17</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>55</td>
<td>3.19E+86</td>
<td>28</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>41</td>
<td>7.98E+58</td>
<td>8</td>
<td>2</td>
<td>9.99E+43</td>
<td>-14.90</td>
</tr>
</tbody>
</table>

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

Ford, S, & Wheeler '14
Results on Limiting the Search Space

<table>
<thead>
<tr>
<th>CI</th>
<th>number of taxa</th>
<th>actual size of tree space</th>
<th>PS\textsubscript{r0} - Mt</th>
<th>number of anchor trees found</th>
<th>size of reduced search space</th>
<th>log reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97</td>
<td>14</td>
<td>7.91E+12</td>
<td>6</td>
<td>2</td>
<td>1.57E+06</td>
<td>-6.70</td>
</tr>
<tr>
<td>0.96</td>
<td>10</td>
<td>34459425</td>
<td>2</td>
<td>3</td>
<td>405</td>
<td>-4.93</td>
</tr>
<tr>
<td>0.95</td>
<td>18</td>
<td>6.33E+18</td>
<td>1</td>
<td>2</td>
<td>3.04E+04</td>
<td>-14.32</td>
</tr>
<tr>
<td>0.95</td>
<td>6</td>
<td>945</td>
<td>5</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.95</td>
<td>32</td>
<td>1.78E+42</td>
<td>2</td>
<td>2</td>
<td>4.98E+15</td>
<td>-26.55</td>
</tr>
<tr>
<td>0.93</td>
<td>39</td>
<td>1.31E+55</td>
<td>4</td>
<td>4</td>
<td>2.25E+39</td>
<td>-15.77</td>
</tr>
<tr>
<td>0.93</td>
<td>40</td>
<td>1.01E+57</td>
<td>18</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>11</td>
<td>6.55E+08</td>
<td>24</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>16</td>
<td>6.19E+15</td>
<td>9</td>
<td>2</td>
<td>1.39E+13</td>
<td>-2.65</td>
</tr>
<tr>
<td>0.91</td>
<td>15</td>
<td>2.13E+14</td>
<td>16</td>
<td>2</td>
<td>1.28E+35</td>
<td>20.78</td>
</tr>
<tr>
<td>0.91</td>
<td>41</td>
<td>7.98E+58</td>
<td>6</td>
<td>2</td>
<td>7.42E+05</td>
<td>-53.03</td>
</tr>
<tr>
<td>0.89</td>
<td>13</td>
<td>3.16E+11</td>
<td>3</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>29</td>
<td>8.89E+36</td>
<td>60</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>15</td>
<td>2.13E+14</td>
<td>20</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.87</td>
<td>25</td>
<td>1.19E+30</td>
<td>3</td>
<td>3</td>
<td>9.36E+15</td>
<td>-14.11</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>17</td>
<td>2</td>
<td>5.31E+17</td>
<td>-2.62</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>8</td>
<td>2</td>
<td>4.57E+10</td>
<td>-9.69</td>
</tr>
<tr>
<td>0.86</td>
<td>20</td>
<td>8.20E+21</td>
<td>17</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>55</td>
<td>3.19E+86</td>
<td>28</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>41</td>
<td>7.98E+58</td>
<td>8</td>
<td>2</td>
<td>9.99E+43</td>
<td>-14.90</td>
</tr>
</tbody>
</table>

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

Ford, S, & Wheeler '14
Results on Limiting the Search Space

<table>
<thead>
<tr>
<th>CI</th>
<th>number of taxa</th>
<th>actual size of tree space</th>
<th>PS\textsubscript{t0} - M\textsubscript{t}</th>
<th>number of anchor trees found</th>
<th>size of reduced search space</th>
<th>log reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97</td>
<td>14</td>
<td>7.91E+12</td>
<td>6</td>
<td>2</td>
<td>1.57E+06</td>
<td>-6.70</td>
</tr>
<tr>
<td>0.96</td>
<td>10</td>
<td>34459425</td>
<td>2</td>
<td>3</td>
<td>405</td>
<td>-4.93</td>
</tr>
<tr>
<td>0.95</td>
<td>18</td>
<td>6.33E+18</td>
<td>1</td>
<td>2</td>
<td>3.04E+04</td>
<td>-14.32</td>
</tr>
<tr>
<td>0.95</td>
<td>6</td>
<td>945</td>
<td>5</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.95</td>
<td>32</td>
<td>1.78E+42</td>
<td>2</td>
<td>2</td>
<td>4.98E+15</td>
<td>-26.55</td>
</tr>
<tr>
<td>0.93</td>
<td>39</td>
<td>1.31E+55</td>
<td>4</td>
<td>4</td>
<td>2.25E+39</td>
<td>-15.77</td>
</tr>
<tr>
<td>0.93</td>
<td>40</td>
<td>0.1E+57</td>
<td>18</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>11</td>
<td>6.55E+08</td>
<td>24</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.92</td>
<td>16</td>
<td>6.19E+15</td>
<td>9</td>
<td>2</td>
<td>1.39E+13</td>
<td>-2.65</td>
</tr>
<tr>
<td>0.91</td>
<td>15</td>
<td>2.13E+14</td>
<td>16</td>
<td>2</td>
<td>1.28E+35</td>
<td>20.78</td>
</tr>
<tr>
<td>0.91</td>
<td>41</td>
<td>7.98E+58</td>
<td>6</td>
<td>2</td>
<td>7.42E+05</td>
<td>-53.03</td>
</tr>
<tr>
<td>0.89</td>
<td>13</td>
<td>3.16E+11</td>
<td>3</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>29</td>
<td>8.89E+36</td>
<td>60</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.88</td>
<td>15</td>
<td>2.13E+14</td>
<td>20</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.87</td>
<td>25</td>
<td>1.19E+30</td>
<td>3</td>
<td>3</td>
<td>9.36E+15</td>
<td>-14.11</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>17</td>
<td>2</td>
<td>5.31E+17</td>
<td>-2.62</td>
</tr>
<tr>
<td>0.87</td>
<td>19</td>
<td>2.22E+20</td>
<td>8</td>
<td>2</td>
<td>4.57E+10</td>
<td>-9.69</td>
</tr>
<tr>
<td>0.86</td>
<td>20</td>
<td>8.20E+21</td>
<td>17</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>55</td>
<td>3.19E+86</td>
<td>28</td>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>0.86</td>
<td>41</td>
<td>7.98E+58</td>
<td>8</td>
<td>2</td>
<td>9.99E+43</td>
<td>-14.90</td>
</tr>
</tbody>
</table>

- 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.

Ford, S, & Wheeler '14
Outline

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

Charles Darwin, 1837
Branch weights are part of the model.

Philippe et al., ‘05
Branch weights are part of the model.
Indicated by length of edges in drawing.
Branch weights are part of the model.
Indicated by length of edges in drawing.
Two classic trees with same underlying topology.
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 number of branches that occur in only one tree, or

\[ T_1 = \]

\[
\text{Turtles} \\
\text{Lizards} \\
\text{Snakes} \\
\text{Birds} \\
\text{Crocodiles}
\]

\[ T_2 = \]

\[
\text{Turtles} \\
\text{Lizards} \\
\text{Snakes} \\
\text{Birds} \\
\text{Crocodiles}
\]

- The size of the symmetric difference of the splits, or

- The sum of the "false positives" and "false negatives."
Robinson-Foulds Distance

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

![Example Trees]

- $T_1 = \text{Turtles} \rightarrow \text{Lizards} \rightarrow \text{Snakes} \rightarrow \text{Birds} \rightarrow \text{Crocodiles}$
- $T_2 = \text{Turtles} \rightarrow \text{Lizards} \rightarrow \text{Snakes} \rightarrow \text{Birds} \rightarrow \text{Crocodiles}$
Robinson-Foulds Distance

- The number 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

T_1 = 

Turtles
Lizards
Snakes
Birds
Crocodiles

T_2 = 

Turtles
Lizards
Snakes
Birds
Crocodiles

Very popular
Robinson-Foulds Distance

- Very popular
- Calculated in linear time, using Day’s Algorithm (‘85)
Randomized $O(nt)$ for majority rule consensus (Amenta, Clarke, & S, WABI ‘03).

Linear time processing of tree reduction rules (Bonet, S, Amenta, & Mahindru ‘06).
Billera, Holmes, and Vogtmann ‘01 have a continuous metric space of trees.
Billera, Holmes, and Vogtmann ‘01 have a continuous metric space of trees.

View each split in a tree as a coordinate in the space.
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**

$T_0 = (1, 2, 3, 4, 5)$

$T_1 = ((1, 2), (3, (4, 5)))$

$T_2 = ((1, 2), (4, (3, 5)))$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>34</th>
<th>35</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_0$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$T_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$T_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$d_{RF}(T_0, T_1) = 2$

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

$d_{RF}(T_1, T_2) = 2$

$d_{BHV}(T_1, T_2) = 2$
Identify Edges of Orthants

\[ (0,1) \]

\[ (0,0) \]

\[ (1,1) \]

\[ (1,0) \]
Identify Edges of Orthants

All images from Billera, Holmes, Vogtmann '01

K. St. John (CUNY & AMNH)
Identify Edges of Orthants

(All images from Billera, Holmes, Vogtmann '01)
The geodesic is a shortest path on the surface between two points.
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)).
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).
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).
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

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

Charles Darwin, 1837
When Trees Are Not Enough

Underlying assumption above:

*Evolution is tree-like.*

Huson, Rupp, Scornavacca ‘10
Underlying assumption above: 

*Evolution is tree-like.*

In many cases, evolution produces a more tangled structure.

Huson, Rupp, Scornavacca ‘10
When Trees Are Not Enough

- 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

- **Nakhleh’s Enumeration Challenge I:** Given a phylogenetic $X$–network $N$

![Phylogenetic network diagram](image)

Leo van Iersel, 2013
Can’t see the trees for the ... network

- **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

- **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 $N$?

Leo van Iersel, 2013

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 $\gamma$ is 3 for general networks, and 5 for tree-child & normal networks.

K. St. John (CUNY & AMNH)
Can’t see the trees for the . . . network

- **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

- 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).

  McDiarmid, Semple, Welsh, 2015: $2 \gamma n \log n + O(n)$, where $\gamma$ is 3/2 for general networks, and 5/4 for tree-child & normal networks.
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 $\gamma$ is 3 for general networks, and 5 for tree-child & normal networks.
Can’t see the trees for the … network

- **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 \(\gamma = \frac{3}{2}\) for general networks, and \(\frac{5}{4}\) for tree-child & normal networks.
Interesting challenges in searching, comparing, analyzing, & visualizing sets of trees.

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.
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.
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.
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.
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.