Advanced Algorithms

Rearrangement distance of phylogenetic trees

Kernelisation, fpt and approximation algorithm

Jonathan Klawitter · WS20
Phylogenetic trees

... represent the evolutionary history of a set of taxa.

- Leaves are labelled with taxa.
- Each taxon represents a species, population, individual organism, gene, chromosome, ... .
- Edge lengths represent amount of time passed or genetic distance.
Phylogenetic trees

... represent the evolutionary history of a set of taxa.

- Leaves are labelled with taxa.
- Each taxon represents a species, population, individual organism, gene, chromosome, ...
- Edge lengths represent amount of time passed or genetic distance.
- Inference methods compute a phylogenetic tree based on some model and data.

by Jenna McCullough 2016
Phylogenetic trees

Let $X = \{1, 2, 3, \ldots n\}$.

A \textbf{(rooted, binary) phylogenetic tree} $T$ is a rooted tree with the following properties:

- The unique \textbf{root} is labeled $\rho$ and has outdegree 1.
- The leaves are bijectively labeled by $X$.
- All other vertices have indegree 1 and outdegree 2.

\textbf{Remarks.}

Here, in our definition

- vertices have \textbf{no heights} and
- the order of leaves does not matter.
Problem

For the same taxa, we may infer different phylogenetic trees because of the use of

- different inference methods,
- different models, or
- different data.

We want to be able to compare different phylogenetic trees. How?

Goal.
Define a metric on phylogenetic trees on $X$ and devise algorithms to compute it.

Idea.
Count the number of rearrangement operations that are necessary to transform $T$ into $T'$. 
**Subtree Prune & Regraft (SPR)**

An **SPR** operation transforms one phylogenetic tree into another one.

- **Subtree**
  - An SPR operation transforms one phylogenetic tree into another one.
  - Note that an SPR operation is reversible.

- **Prune**
  - Subtree
  - Prune
  - Regraft

- **Regraft**
  - Spruce
  - Replant

- **Note**
  - An SPR operation is reversible.
**SPR-graph**

SPR induces the **SPR-graph** $G = (V, E)$:

- $V = \{ T \mid T \text{ is a phylogenetic tree on } X \}$
- $\{ T, T' \} \in E$ if $T$ can be transformed into $T'$ with a single SPR operation
SPR-distance

The **SPR-distance** $d_{SPR}(T, T')$ of $T$ and $T'$ is defined as the distance of $T$ and $T'$ in the SPR-graph $G$.

**Lemma 1.**
The SPR-graph $G$ is connected.

**Proof** as exercise or in discussion.

**Lemma 2.**
The SPR-distance is a metric.

**Proof.** $G$ is connected and undirected.

**Goal.**
Compute the SPR-distance $d_{SPR}(T, T')$.

... but $G$ is huge!

$$|V(G)| = (2n - 3)!! = (2n - 3) \cdot (2n - 5) \cdot \ldots \cdot 5 \cdot 3$$
The **SPR-distance** $d_{SPR}(T, T')$ of $T$ and $T'$ is defined as the distance of $T$ and $T'$ in the SPR-graph $G$.

**Lemma 1.** The SPR-graph $G$ is connected.

**Proof** as exercise or in discussion.

**Lemma 2.** The SPR-distance is a metric.

**Proof.** $G$ is connected and undirected.

**Goal.** Compute the SPR-distance $d_{SPR}(T, T')$.

...but $G$ is huge!

$$|V(G)| = (2n - 3)!! = (2n - 3) \cdot (2n - 5) \cdot \ldots \cdot 5 \cdot 3$$

- Can we rephrase the problem?
Maximum agreement forests

\[ T \xrightarrow{\text{SPR}} T' \xrightarrow{\text{SPR}} T'' \]

\[ F \text{ into } T \]

\[ F \]

\[ F \text{ into } T'' \]
Maximum agreement forests

An agreement forest \( F \) of \( T \) and \( T'' \) is a forest \( \{T_\rho, T_1, T_2, \ldots, T_k\} \) such that

- the label sets of the \( T_i \) partition \( X \cup \{\rho\} \),
- \( \rho \) is in the label set of \( T_\rho \), and
- there exist edge-disjoint embeddings of subdivisions of the \( T_i \)'s into \( T \) and \( T'' \) that cover all edges.

If \( k \) is minimal, \( F \) is a maximum agreement forest (MAF).
Characterisation

Let $T$ and $T'$ be two phylogenetic trees on $X$. Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$. Define

$$m(T, T') = k = |F| - 1.$$
Characterisation

Let $T$ and $T'$ be two phylogenetic trees on $X$.

Let $F = \{T_{\rho}, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$.

Define

$$m(T, T') = k = |F| - 1.$$ 

**Theorem 3.** $m(T, T') = d_{\text{SPR}}(T, T')$

**Proof** of “$\leq$” by induction on $d = d_{\text{SPR}}(T, T')$.

- Case $d = 1$ is easy. ✓
- Assume $m(T, T') \leq d_{\text{SPR}}(T, T')$ holds for all $d \leq \ell$. 

![Diagram showing phylogenetic trees and MAF](image-url)
Characterisation

Let $T$ and $T'$ be two phylogenetic trees on $X$.
Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$.
Define
\[ m(T, T') = k = |F| - 1. \]

**Theorem 3.** \( m(T, T') = d_{\text{SPR}}(T, T') \)

**Proof** of “≤” by induction on \( d = d_{\text{SPR}}(T, T') \).

- If \( d = \ell + 1 \), then there exists $T''$ with $d_{\text{SPR}}(T, T'') = \ell$ and $d_{\text{SPR}}(T'', T') = 1$.
- There exists MAF $F'$ for $T$ and $T''$ and MAF $F''$ for $T''$ and $T'$. 

\[ \begin{array}{c}
T \\
\rho \\
\ell \text{ SPR} \\
\rho \\
F' \\
\rho \\
\rho_1 \rho_2 \\
\rho \\
T'' \\
\rho \\
\rho_1 \rho_2 \\
\rho \\
T' \\
\rho \\
\rho_1 \rho_2 \\
\rho \\
F'' \\
\rho \\
\rho_1 \rho_2 \\
\rho \\
\end{array} \]
Characterisation

Let $T$ and $T'$ be two phylogenetic trees on $X$.
Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$.
Define
\[
m(T, T') = k = |F| - 1.
\]

**Theorem 3.** \(m(T, T') = d_{\text{SPR}}(T, T')\)

**Proof** of “\(\leq\)” by induction on \(d = d_{\text{SPR}}(T, T')\).
- If \(d = \ell + 1\), then there exists $T''$ with $d_{\text{SPR}}(T, T'') = \ell$ and $d_{\text{SPR}}(T'', T') = 1$.
- There exists MAF $F'$ for $T$ and $T''$ and MAF $F''$ for $T''$ and $T'$. 

\[\begin{array}{cccccc}
T & T'' & F' & T'' & T' & F'' & F \\
\rho & \rho & \rho & \rho & \rho & \rho & \rho \\
\ell_{\text{SPR}} & & & & & & \\
\end{array}\]
Characterisation

Let $T$ and $T'$ be two phylogenetic trees on $X$. Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$. Define

$$m(T, T') = k = |F| - 1.$$ 

**Theorem 3.** $m(T, T') = d_{SPR}(T, T')$

**Proof** of "$\geq"$ by induction on $d = m(T, T')$.

- Case $d = 1$ is easy. ✓
- Assume $m(T, T') \geq d_{SPR}(T, T')$ holds for all $d \leq \ell$. 

$$F$$

$$T$$

$$T'$$

$\rho$

$\rho$

$\rho$

$\rho$

$\rho$

$\rho$

SPR
Let $T$ and $T'$ be two phylogenetic trees on $X$. Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of $T$ and $T'$. Define

$$m(T, T') = k = |F| - 1.$$ 

**Theorem 3.** $m(T, T') = d_{\text{SPR}}(T, T')$

**Proof** of “$\geq$” by induction on $d = m(T, T')$.

- Let $F$ be a MAF of $T$ and $T'$ of size $\ell + 2$.
- There exists a $T_i$ that can be pruned in $T$.
- Regraft $T_i$ according to the embedding of $F$ into $T' \Rightarrow T''$ & $F'$
- $F'$ is an AF for $T'$ and $T''$
- $\Rightarrow d_{\text{SPR}}(T'', T') \leq \ell$
- $d_{\text{SPR}}(T, T'') = 1$
- $d_{\text{SPR}}(T, T') \leq \ell + 1 = m(T, T')$
Theorem 4. [HJWZ ’96, BS ’05]
Computing $d_{\text{SPR}}(T, T')$ is NP-hard.

Proof is by reduction from Exact Cover by 3-Sets.

Plan.
- Construct kernel of the problem.
  - Replace $T$ and $T'$ with smaller $S$ and $S'$.
  - We should be able to get $d_{\text{SPR}}(T, T')$ from $d_{\text{SPR}}(S, S')$.
- Show that size of the kernel depends on $d_{\text{SPR}}(T, T')$.
- Devise an fpt algorithm by computing $d_{\text{SPR}}$ for kernel.
- Devise an approximation algorithm.
Kernelisation – Subtrees

**Common subtree reduction.**
- Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label.

\[
T \quad T' \quad S \quad S'
\]

**Lemma 5.** Applying the common subtree reduction is safe; i.e. \(d_{SPR}(T, T') = d_{SPR}(S, S')\).

**Proof.**
- Suppose is covered by two trees of MAF
- then there is alternative MAF
Kernelisation – Chains

Chain reduction.
- Replace any chain of leaves that occurs identically in both trees by three new leaves.

Lemma 6. Applying chain reduction is safe; i.e. $d_{SPR}(T, T') = d_{SPR}(S, S')$.

Proof.
- Show there is a tree with abc-chain in a MAF of $S$ and $S'$.
- Swap abc-chain with original chain for MAF of $T$ and $T'$. 
Kernelisation – Chains

Chain reduction.
■ Replace any chain of leaves that occurs identically in both trees by three new leaves.

Lemma 6. Applying chain reduction is safe; i.e. $d_{\text{SPR}}(T, T') = d_{\text{SPR}}(S, S')$.

Proof.
■ Consider embedding of a MAF $F$ into $S$.  

Case 1
■ Consider embedding of a MAF $F$ into $S$.  

Proof.
Kernelisation – Chains

**Chain reduction.**
- Replace any chain of leaves that occurs identically in both trees by three new leaves.

\[ T \rightarrow T' \quad \text{and} \quad S \rightarrow S' \]

**Lemma 6.** Applying chain reduction is safe; i.e. \( d_{\text{SPR}}(T, T') = d_{\text{SPR}}(S, S') \).

**Proof.**
- Consider embedding of a MAF \( F \) into \( S \).

Case 2
Kernel size

**Theorem 7.**
Reduce $T$ and $T'$ to $S$ and $S'$ by exhaustively applying the reduction rules. Let $S$ and $S'$ be on $X'$. Then

$$|X'| \leq 28 \text{d}_{\text{SPR}}(T, T').$$

**Proof.** Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for $S$ and $S'$. Let $n(T_i)$ be $\#$ of $T_j$ that $T_i$ overlaps with in embedding of $F$ into $S$. Claim 1. $\sum_{i=\rho}^k (n(T_i) + n'(T_i)) \leq 4k = 4 \text{d}_{\text{SPR}}(T, T').$

\[|V(H)| = k + 1 = |E(H)| + 1\]

\[\sum n(T_i) = 2|E(H)| \leq 2k\]
Kernel size

**Theorem 7.**
Reduce $T$ and $T'$ to $S$ and $S'$ by exhaustively applying the reduction rules. Let $S$ and $S'$ be on $X'$. Then

$$|X'| \leq 28 d_{SPR}(T, T').$$

**Proof.** Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for $S$ and $S'$. Let $n(T_i)$ be the number of $T_j$ that $T_i$ overlaps with in the embedding of $F$ into $S$.

**Claim 1.** $\sum_{i=\rho}^{k} (n(T_i) + n'(T_i)) \leq 4k = 4 d_{SPR}(T, T').$

**Claim 2.** The number of leaves of $T_i \leq 7(n(T_i) + n'(T_i))$. 

$$\sum_{i=\rho}^{k} \text{# leaves of } T_i \leq \sum_{i=\rho}^{k} 7(n(T_i) + n'(T_i)) \leq 28k$$
FPT algorithm

Theorem 8.
Computing $d_{\text{SPR}}(T, T')$ is fixed-parameter tractable when parameterized by $d_{\text{SPR}}(T, T')$.

Proof.
- Reduce $T$ and $T'$ to $S$ and $S'$ by exhaustively applying the reduction rules.
- Let $S$ and $S'$ be on $X'$ and let $k = d_{\text{SPR}}(S, S')$.
- $S$ has at most $4|X'|^2$ neighbours.
  - $S$ has less than $2|X'|$ edges to cut and to attach to.  
    by Theorem 7
- Length-$k$ BFS from $S$ visits at most $O\left((4|X'|^2)^k\right) = O((56k)^{2k})$ trees.
- Since $k = d_{\text{SPR}}(S, S') = d_{\text{SPR}}(T, T')$, this yields an fpt algorithm.
Approximation algorithm

Idea.
- Given reduced trees $T$ and $T'$ we compute an agreement forest $F$ by
- successively making “cuts” and “eliminations”.
- This shrink $T$ and $T'$ further and further.
- Show that $|F|$ is at most $3|F'|$, where $F'$ is a MAF of $T$ and $T'$. 
Approximation algorithm

\[
\text{APPROXDSPR}(T, T')
\]

\[
i \leftarrow 1
\]

\[
G_i \leftarrow T
\]

\[
H_i \leftarrow T'
\]

\[
\text{while } \exists \text{ pair of sibling leaves } a \text{ and } b \text{ in } G_i \text{ do}
\]

\[
\text{find the case that applies to } a \text{ and } b \text{ in } H_i
\]

\[
\text{apply the corresponding transaction}
\]

\[
\text{to obtain } G_{i+1} \text{ from } G_i \text{ and } H_{i+1} \text{ from } H_i
\]

\[
i++
\]

\[
\text{return } |H_i| - 1
\]
Approximation algorithm – example

$T = G_1$

$T' = H_1$

Case 2

- Should we cut of the leaves 1 or 2 or all in between them in $H_1$?
- Do parts of each!
Approximation algorithm – example

Case 1

If the same cherry occurs in \( H_i \), we can simply reduce it.
Case 4

- Leaf $b$ is the only leaf of a tree in $H_i$.
- Cut off $b$ in $G_i$. 

Leaf $b$ is the only leaf of a tree in $H_i$. 
Cut off $b$ in $G_i$. 

$G_3$ 

$H_3$
Approximation algorithm – example

\[ G_4 \]

\[ H_4 \]

Return 3.
Approximation algorithm – analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>$G_i$</th>
<th>$H_i$</th>
<th>$G_{i+1}$</th>
<th>$H_{i+1}$</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a b</td>
<td>a b</td>
<td>c</td>
<td>c</td>
<td>no mistake</td>
</tr>
<tr>
<td>2</td>
<td>a b</td>
<td>a b</td>
<td>a b</td>
<td>a b</td>
<td>3 cuts 1+ good</td>
</tr>
<tr>
<td>3</td>
<td>a b</td>
<td>a b</td>
<td>a b</td>
<td>a b</td>
<td>2 cuts 1+ good</td>
</tr>
<tr>
<td>4</td>
<td>a b</td>
<td>b</td>
<td>a b</td>
<td>b</td>
<td>1 cut 1 good</td>
</tr>
</tbody>
</table>
Approximation algorithm – analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>$G_i$</th>
<th>$H_i$</th>
<th>$G_{i+1}$</th>
<th>$H_{i+1}$</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
<td>no mistake</td>
</tr>
<tr>
<td>2</td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
<td><img src="image7" alt="Diagram" /></td>
<td><img src="image8" alt="Diagram" /></td>
<td>3 cuts 1+ good</td>
</tr>
<tr>
<td>3</td>
<td><img src="image9" alt="Diagram" /></td>
<td><img src="image10" alt="Diagram" /></td>
<td><img src="image11" alt="Diagram" /></td>
<td><img src="image12" alt="Diagram" /></td>
<td>2 cuts 1+ good</td>
</tr>
<tr>
<td>4</td>
<td><img src="image13" alt="Diagram" /></td>
<td><img src="image14" alt="Diagram" /></td>
<td><img src="image15" alt="Diagram" /></td>
<td><img src="image16" alt="Diagram" /></td>
<td>1 cut 1 good</td>
</tr>
</tbody>
</table>

**Theorem 9**

**APPROXDSPR** is a 3-approximation algorithm for $d_{SPR}(T, T')$ with an $O(|X|^2)$ running time.
Discussion

Phylogenetic trees.
- There are other classes of phylogenetic trees: unrooted, non-binary, ranked, ...
- Trees can be generalized to phylogenetic networks, which can also have indegree 2 outdegree 1 vertices.

Maximum Agreement Forests.
- Reframing (characterising) a problem in a different way, can sometimes make your life a lot easier.
- MAF can be generalized to Maximum Agreement Graphs, but these don't characterize the SPR-distance of networks anymore.
Discussion

**Kernelization.**
- Kernelization is an important technique to construct fpt algorithms.
- Result important since SPR-distance small in practice.
- Reduction rules actually give a kernel of size at most $15k - 9$.
- With further reduction rules can get size below $11k - 9$. [KL ’18]
- Divide & conquer algorithm can (in practice) reduce further reduce problem sizes. [LS ’11]

**Approximation algorithm.**
- There exist 2-approximation algorithms for the SPR-distance with a running time in $O(n^3)$. [CHW ’17]
Literature

Original papers:

- [BS '05] “On the computational complexity of the rooted subtree prune and regraft distance” for SPR, MAF, characterisation, fpt, divide & conquer
- [RSW '06] “The maximum agreement forest problem: Approximation algorithms and computational experiments”

Referenced papers:

- [HJWZ '96] “On the complexity of comparing evolutionary trees” for NP-hardness proof
- [KL '19] “New reduction rules for the tree bisection and reconnection distance”
- [CHW '17] “A New 2-Approximation Algorithm for rSPR Distance”
- [LS11] “A cluster reduction for computing the subtree distance between phylogenies”