Properties of the TBR metric

Peter Humphries

pjh96@student.canterbury.ac.nz

Department of Mathematics and Statistics

University of Canterbury

Tree rearrangement operations

Tree rearrangement operations

- used to change one phylogenetic tree to another by a sequence of localised manipulations

Tree rearrangement operations

- used to change one phylogenetic tree to another by a sequence of localised manipulations
- introduced by Robinson (1969) for unrooted trees, and has since been extended in a number of ways

Tree rearrangement operations

- used to change one phylogenetic tree to another by a sequence of localised manipulations
- introduced by Robinson (1969) for unrooted trees, and has since been extended in a number of ways
- measure the degree of similarity between two phylogenetic trees on the same set of taxa

Tree rearrangement operations

- used to change one phylogenetic tree to another by a sequence of localised manipulations
- introduced by Robinson (1969) for unrooted trees, and has since been extended in a number of ways
- measure the degree of similarity between two phylogenetic trees on the same set of taxa
- quantify (roughly) the number of hybridisation events required to resolve the incompatibility of two trees

Tree rearrangement operations

- used to change one phylogenetic tree to another by a sequence of localised manipulations
- introduced by Robinson (1969) for unrooted trees, and has since been extended in a number of ways
- measure the degree of similarity between two phylogenetic trees on the same set of taxa
- quantify (roughly) the number of hybridisation events required to resolve the incompatibility of two trees
- have algorithmic applications for optimisation searches through a tree space

Operations on unrooted trees

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees
  - Nearest Neighbour Interchange (NNI)

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees
  - Nearest Neighbour Interchange (NNI)
  - Subtree Prune and Regraft (SPR)

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees
  - Nearest Neighbour Interchange (NNI)
  - Subtree Prune and Regraft (SPR)
  - Tree Bisection and Reconnection (TBR)

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees
  - Nearest Neighbour Interchange (NNI)
  - Subtree Prune and Regraft (SPR)
  - Tree Bisection and Reconnection (TBR)
- each of these operations induces a metric on $\mathscr{T}_n$

Operations on unrooted trees

- denote by $\mathscr{T}_n$ the set of all binary unrooted phylogenetic trees labelled by $\{1, \ldots, n\}$
- three main operations on unrooted trees
  - Nearest Neighbour Interchange (NNI)
  - Subtree Prune and Regraft (SPR)
  - Tree Bisection and Reconnection (TBR)
- each of these operations induces a metric on $\mathscr{T}_n$

GIVEN A TREE $\mathcal{T}$, HOW MANY OTHER TREES CAN BE REACHED BY APPLYING A SINGLE TREE REARRANGEMENT OPERATION?

Unit neighbourhoods

Unit neighbourhoods

- for all $\mathcal{T} \in \mathscr{T}_n$, we define the TBR unit neighbourhood of $\mathcal{T}$ by

$$N_{\mathrm{TBR}}(\mathcal{T}) = \{\mathcal{T}' \in \mathscr{T}_n : d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = 1\}$$

Unit neighbourhoods

- for all $\mathcal{T} \in \mathscr{T}_n$, we define the TBR unit neighbourhood of $\mathcal{T}$ by

$$N_{\mathrm{TBR}}(\mathcal{T}) = \{\mathcal{T}' \in \mathscr{T}_n : d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = 1\}$$

- for both NNI and SPR, the exact size of the unit neighbourhood is known and is independent of the shape of $\mathcal{T}$

Unit neighbourhoods

- for all $\mathcal{T} \in \mathscr{T}_n$, we define the TBR unit neighbourhood of $\mathcal{T}$ by

$$N_{\mathrm{TBR}}(\mathcal{T}) = \{\mathcal{T}' \in \mathscr{T}_n : d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = 1\}$$

- for both NNI and SPR, the exact size of the unit neighbourhood is known and is independent of the shape of $\mathcal{T}$
- for TBR, the size of the unit neighbourhood does depend on tree shape

Unit neighbourhoods

- for all $\mathcal{T} \in \mathscr{T}_n$, we define the TBR unit neighbourhood of $\mathcal{T}$ by

$$N_{\mathrm{TBR}}(\mathcal{T}) = \{\mathcal{T}' \in \mathscr{T}_n : d_{\mathrm{TBR}}(\mathcal{T}, \mathcal{T}') = 1\}$$

- for both NNI and SPR, the exact size of the unit neighbourhood is known and is independent of the shape of $\mathcal{T}$
- for TBR, the size of the unit neighbourhood does depend on tree shape
  - for all $\mathcal{T} \in \mathscr{T}_n$ we have

$$2(n-3)(2n-7) \leq |N_{\mathrm{TBR}}(\mathcal{T})| \leq (2n-3)(n-3)^2$$

Recursion equation for $\mathcal{T}$

Recursion equation for $\mathcal{T}$

- if we let $\mathcal{T}$ be a tree in $\mathscr{T}_n$, then there are three ways to apply a TBR operation to $\mathcal{T}$

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

Recursion equation for $\mathcal{T}$

- if we let $\mathcal{T}$ be a tree in $\mathscr{T}_n$, then there are three ways to apply a TBR operation to $\mathcal{T}$
  - retain a cherry $\{x, y\}$ of $\mathcal{T}$

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

Recursion equation for $\mathcal{T}$

- if we let $\mathcal{T}$ be a tree in $\mathscr{T}_n$, then there are three ways to apply a TBR operation to $\mathcal{T}$
  - retain a cherry $\{x, y\}$ of $\mathcal{T}$
  - cut either x or y from $\mathcal{T}$ and reattach so that $\{x, y\}$ is no longer a cherry

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

Recursion equation for $\mathcal{T}$

- if we let $\mathcal{T}$ be a tree in $\mathscr{T}_n$, then there are three ways to apply a TBR operation to $\mathcal{T}$
  - retain a cherry $\{x, y\}$ of $\mathcal{T}$
  - cut either x or y from $\mathcal{T}$ and reattach so that $\{x, y\}$ is no longer a cherry
  - cut a proper subtree of $\mathcal{R}_1$ or $\mathcal{R}_2$ and attach adjacent to either x or y

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

Recursion equation for $\mathcal{T}$

- if we let $\mathcal{T}$ be a tree in $\mathscr{T}_n$, then there are three ways to apply a TBR operation to $\mathcal{T}$
  - retain a cherry $\{x, y\}$ of $\mathcal{T}$
  - cut either x or y from $\mathcal{T}$ and reattach so that $\{x, y\}$ is no longer a cherry
  - cut a proper subtree of $\mathcal{R}_1$ or $\mathcal{R}_2$ and attach adjacent to either x or y

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

- $\xi(\mathcal{R})$ counts the number of ways to prune and re-root a proper subtree of a rooted tree $\mathcal{R}$

Calculating $\xi(\mathcal{R})$

Calculating $\xi(\mathcal{R})$

- delete the root of $\mathcal{R}$ to give two rooted trees $\mathcal{R}_1$ and $\mathcal{R}_2$

$$\xi(\mathcal{R}) = |E(\mathcal{R}_1)| + |E(\mathcal{R}_2)| - 2 + \xi(\mathcal{R}_1) + \xi(\mathcal{R}_2)$$

Calculating $\xi(\mathcal{R})$

- delete the root of $\mathcal{R}$ to give two rooted trees $\mathcal{R}_1$ and $\mathcal{R}_2$
  - ► we can either cut one of $\mathcal{R}_1$ and $\mathcal{R}_2$ and root it on some edge ...

$$\xi(\mathcal{R}) = |E(\mathcal{R}_1)| + |E(\mathcal{R}_2)| - 2 + \xi(\mathcal{R}_1) + \xi(\mathcal{R}_2)$$

Calculating $\xi(\mathcal{R})$

- delete the root of $\mathcal{R}$ to give two rooted trees $\mathcal{R}_1$ and $\mathcal{R}_2$
  - we can either cut one of $\mathcal{R}_1$ and $\mathcal{R}_2$ and root it on some edge ...
  - ... or cut a proper subtree of $\mathcal{R}_1$ or $\mathcal{R}_2$

$$\xi(\mathcal{R}) = |E(\mathcal{R}_1)| + |E(\mathcal{R}_2)| - 2 + \xi(\mathcal{R}_1) + \xi(\mathcal{R}_2)$$

Calculating $\xi(\mathcal{R})$

- delete the root of $\mathcal{R}$ to give two rooted trees $\mathcal{R}_1$ and $\mathcal{R}_2$
  - ► we can either cut one of $\mathcal{R}_1$ and $\mathcal{R}_2$ and root it on some edge ...
  - ► ... or cut a proper subtree of $\mathcal{R}_1$ or $\mathcal{R}_2$

$$\xi(\mathcal{R}) = |E(\mathcal{R}_1)| + |E(\mathcal{R}_2)| - 2 + \xi(\mathcal{R}_1) + \xi(\mathcal{R}_2)$$

- applying this gives sharp bounds on $\xi(\mathcal{R})$ of

$$2 + 2\sum_{i=2}^{n-1} \lfloor \log_2 i \rfloor \leq \xi(\mathcal{R}) \leq n^2 - 3n + 4$$

Upper bound

Upper bound

Theorem 1. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \leq \frac{2}{3}n^3 - 6n^2 + \frac{70}{3}n - 38$$

Moreover, for $n \geq 7$, equality holds if and only if $\mathcal{T}$ is a caterpillar.

Upper bound

Theorem 1. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \leq \frac{2}{3}n^3 - 6n^2 + \frac{70}{3}n - 38$$

Moreover, for $n \geq 7$, equality holds if and only if $\mathcal{T}$ is a caterpillar.

- contrasts directly with Song's (2003) result that rooted caterpillars minimise the size of the rooted SPR unit neighbourhood

Upper bound

Theorem 1. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \leq \frac{2}{3}n^3 - 6n^2 + \frac{70}{3}n - 38$$

Moreover, for $n \geq 7$, equality holds if and only if $\mathcal{T}$ is a caterpillar.

- contrasts directly with Song's (2003) result that rooted caterpillars minimise the size of the rooted SPR unit neighbourhood
- invites the conjecture that balanced unrooted trees will minimise the size of the TBR unit neighbourhood

Lower bound

Lower bound

Theorem 2. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \geq 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

Lower bound

Theorem 2. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \geq 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

- uses induction around a cherry at the end of a longest path

Lower bound

Theorem 2. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \geq 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

- uses induction around a cherry at the end of a longest path
- when $n \geq 9$, this gives a lower bound that is strictly larger than the size of $N_{\mathrm{SPR}}(\mathcal{T})$

Lower bound

Theorem 2. *For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have*

$$|N_{\mathrm{TBR}}(\mathcal{T})| \geq 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

- uses induction around a cherry at the end of a longest path
- when $n \geq 9$, this gives a lower bound that is strictly larger than the size of $N_{\mathrm{SPR}}(\mathcal{T})$
- does this bound grow faster than quadratically?

Lower bound (asymptotics)

Lower bound (asymptotics)

- define $f(n)$ by

$$f(n) = 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

Lower bound (asymptotics)

- define $f(n)$ by

$$f(n) = 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

  where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

- the generating function for the sequence $f(n)$ is

$$F(x) = \frac{2x^4}{(1-x)^3}\left(1 + 3x + 2x\sum_{k=2}^{\infty} x^{2^k}\right)$$

Lower bound (asymptotics)

- define $f(n)$ by

$$f(n) = 2n^2 - 8n + 2 + \sum_{i=2}^{n-4} l(i)$$

  where $l(i) = 2 + 2\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

- the generating function for the sequence $f(n)$ is

$$F(x) = \frac{2x^4}{(1-x)^3} \left(1 + 3x + 2x \sum_{k=2}^{\infty} x^{2^k} \right)$$

- that is, $[x^n]F(x) \leq |N_{\mathrm{TBR}}(\mathcal{T})|$ for all $\mathcal{T} \in \mathscr{T}_n$

Further work ...

Further work ...

- determine the asymptotic behaviour of the current lower bound

Further work ...

- determine the asymptotic behaviour of the current lower bound
- find a tight lower bound on the size of the TBR unit neighbourhood for all n, and characterise the trees that attain this bound

Further work ...

- determine the asymptotic behaviour of the current lower bound
- find a tight lower bound on the size of the TBR unit neighbourhood for all n, and characterise the trees that attain this bound
- find a closed form expression for $|N_{\mathrm{TBR}}(\mathcal{T})|$

Further work ...

- determine the asymptotic behaviour of the current lower bound
- find a tight lower bound on the size of the TBR unit neighbourhood for all n, and characterise the trees that attain this bound
- find a closed form expression for $|N_{\mathrm{TBR}}(\mathcal{T})|$

$$|N_{\mathrm{TBR}}(\mathcal{T})| = |N_{\mathrm{TBR}}(\mathcal{T} - x)| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

$$|N_{\mathrm{rSPR}}(\mathcal{T})| = 2(n-2)(2n-5) - 2\sum_{i=1}^{n-2}\gamma(v_i)$$