Abstract: Given a collection of rooted phylogenetic trees with overlapping sets of leaves, a supertree $S$ is a single tree whose set of leaves is the union of the input sets of leaves and such that $S$ agrees with each input tree when restricted to the leaves of the input tree. Typically with trees from real data, no supertree exists, and various methods may be utilized to reconcile the incompatibilities in the input trees. This talk focuses on a measure of robustness of a supertree method called its "radius" $R$. For example, if $R = 1/10$, then whenever $T$ is a candidate binary tree and for all rooted triples $abc$ in $T$ we have that $\{a,b,c\}$ occur together in some input tree and that more than 90% of the input triples involving $\{a,b,c\}$ are in fact $abc$ (a strong assumption), then the method outputs $T$ as the supertree; but this might fail if 90% is replaced by 89%. It is shown that the maximal possible radius for a method is $R = 1/2$. Many familiar methods, both for supertrees and consensus trees, have $R = 0$, indicating that they need not output a tree $T$ that would seem to be the natural correct answer. Some methods with the maximal possible $R = 1/2$ are indicated. Extensions may be presented concerning supertree methods that rely on input distance information as well as the topology of the input trees.
Robustness of supertree methods for reconciling dense incompatible data

by Stephen J. Willson
Department of Mathematics
Iowa State University
Ames, Iowa 50011
USA

swillson@iastate.edu

23 October 2007
The Supertree Problem

Different researchers use different methods and different genes and different collections of taxa to produce their phylogenetic trees.

But we expect a single "tree of life."

When can we combine the different trees into a single tree containing all the taxa?
A purely mathematical version:

Here are 3 rooted trees.

Do they fit together into a single rooted tree (a supertree)?
In this case there are several possible supertrees:
With biological data it is more likely that no true supertree exists.

There are various ways to try to find an approximate supertree that fits as well as possible. Incompatible data need to be reconciled.
Supertree methods have two aspects:

(a) Extrapolation
(b) Reconciliation of incompatible data

Problem. Study quantitatively how well a supertree method reconciles incompatible data.
Some Supertree Methods


(2) MinCutSupertree. (Semple and Steel 2000)


(4) Matrix Representation using Flipping (MRF). (Burleigh et al. 2004)

(5) Most Similar Supertree (MSS).  (Creevey et al. 2004)

(6) MaxCutTriplets. (Snir and Rao 2006)

Book: Phylogenetic Supertrees: Combining information to reveal the Tree of Life (ed. Olaf Bininda-Emonds, 2004)
A rooted phylogenetic tree with leaf set $X$ can be described by its clusters.

The cluster of a vertex consists of all the leaf descendants of that vertex.

$X = \{1,2,3,4,5,6,7\}$.

The cluster of 9 is $\{1,2,3,4\}$.
The cluster of 10 is $\{5,6,7\}$.
The children of $\{1,2,3,4\}$ are $\{1,2,3\}$ and $\{4\}$.
In fact, given such rooted input trees, if there exists a supertree, then one can be found fast by the procedure BUILD (Aho et al. 1981).

A common situation is as follows: One seeks the children of the cluster U.
1. Form a graph $A(U)$ in which the vertices are the members of U.
2. Find edges in $A(U)$ in some manner.
3. If the graph is disconnected, the children of U are the members of the components of $A(U)$. Otherwise, do something to disconnect $A(U)$. 
A common situation occurs when the edges in $A(U)$ have a numerical weight obtained in some manner.

$U = \{a,b,c,d,e\}$

One method to deal with this is to use a **minimum cut**:
A cut for a decomposition of $U$ into $Y$ and $Z$ is the sum of the weights of edges broken by the decomposition.
Cut $\{a\}, \{b,c,d,e\}$ is 2.
Cut $\{a,b\}, \{c,d,e\}$ is 3.
Cut $\{a,b,c\}, \{d,e\}$ is 4.
There is a unique minimum cut $\{a\}, \{b,c,d,e\} = 2$.
Make the children of $U = \{a,b,c,d,e\}$ be $\{a\}$ and $\{b,c,d,e\}$.
Another method is to use a **minimum threshold**:

For a threshold \( \tau \), let \( A_\tau(U) \) include only edges with weight \( > \tau \).

If \( \tau = 2 \), \( A_2(\{a,b,c,d,e\}) \) is

The graph decomposes into \( \{a\}, \{b\}, \{c\}, \{d,e\} \).
If $\tau = 1$,

the graph decomposes into \{a,b\}, \{c,d,e\}.
If \( \tau = 0 \),

the graph does not decompose.
The method of **minimum threshold** utilizes the smallest threshold $\tau$ such that $A_\tau(U)$ decomposes into at least two components. In this case, the minimum threshold is $\tau = 1$ and $A_1(U)$ is

The children of $\{a,b,c,d,e\}$ using minimum threshold are $\{a,b\}$ and $\{c,d,e\}$.
There is always a minimum threshold, possibly 0. It can be computed rapidly. The **minimum threshold tree** is the tree found by always utilizing the minimum threshold criterion.

The issue is how to choose the weights.
Let $T$ be a rooted tree. Given distinct leaves $a, b, c$, say $ab \mid c$ in $T$ if $T$ has a cluster that contains $a$ and $b$ but not $c$, and say $T$ contains the **rooted triplet** $ab \mid c$. 

1 2 | 4
1 4 | 5
5 7 | 3
not 2 4 | 1
not 5 6 | 7
Suppose we are given as data a collection $\mathcal{D} = \{T_i : i \in \Lambda\}$ of rooted trees.

For each distinct $a,b,c$ define $spt_D(ab|c)$

= the **normalized triplet support** of $ab \mid c$ as follows:

Let $\text{den}(a,b,c) = \# \{i \in \Lambda: \{a,b,c\} \subseteq L(T_i)\}$

$\text{num}(ab|c) = \# \{i \in \Lambda: ab \mid c \text{ in } T_i\}$.

$spt_D(ab|c) = \begin{cases} 
\frac{\text{num}(ab|c)}{\text{den}(a,b,c)} & \text{if } \text{den}(a,b,c) > 0 \\
0 & \text{if } \text{den}(a,b,c) = 0.
\end{cases}$
\[
\begin{align*}
\text{num}(45|2) &= 1 \\
\text{den}(4,5,2) &= 2 \\
\text{spt}_D(45|2) &= 1/2 \\
\text{num}(45|3) &= 2 \\
\text{num}(34|5) &= 1 \\
\text{den}(4,5,3) &= 3 \\
\text{den}(3,4,5) &= 3 \\
\text{spt}_D(45|3) &= 2/3 \\
\text{spt}_D(34|5) &= 1/3 \\
\text{num}(35|4) &= 0 \\
\text{den}(3,4,5) &= 3 \\
\text{spt}_D(35|4) &= 0
\end{align*}
\]
It is immediate that for all distinct $a$, $b$, $c$,
\[0 \leq \text{spt}\mathcal{D}(abl_c) \leq 1\]
\[\text{spt}\mathcal{D}(abl_c) = \text{spt}\mathcal{D}(bal_c)\]
\[\text{spt}\mathcal{D}(abl_c) + \text{spt}\mathcal{D}(acl_b) + \text{spt}\mathcal{D}(bla_c) \leq 1.\]
Given $a, b$ in $U$, define
\[ \text{mspt}_U(a,b) = \max \{\text{spt}_D(ab|c) : a,b,c \text{ are distinct elements of } U\}. \]

\[ \text{num}(45|2) = 1 \]
\[ \text{den}(4,5,2) = 2 \]
\[ \text{spt}_D(45|2) = 1/2 \]

\[ \text{num}(45|3) = 2 \]
\[ \text{den}(4,5,3) = 3 \]
\[ \text{spt}_D(45|3) = 2/3 \]

\[ \text{num}(45|1) = 1 \]
\[ \text{den}(4,5,1) = 2 \]
\[ \text{spt}_D(45|1) = 1/2 \]

\[ \text{mspt}_U(4,5) = \max \{\text{spt}_D(45|1), \text{spt}_D(45|2), \text{spt}_D(45|3)\} = 2/3. \]

\[ \text{mspt}_U(4,5) \text{ is the } \text{normalized triplet} \text{ support for edge } \{4,5\}. \]
The normalized triplet supertree

Construct a tree $S$ as follows:

(1) $X$ is a cluster of $S$.

Repeat until done:
(2) (i) If $U$ is a cluster of $S$ with two members, $U = \{a, b\}$, then the children of $U$ are $\{a\}$ and $\{b\}$.

(ii) If $U$ is a cluster of $S$ with more than two members, form $A(U)$ in which the edge $\{a, b\}$ has weight $mspt_U(a, b)$. Find the minimal threshold $\tau$ such that $A_\tau(U)$ is disconnected. The components of $A_\tau(U)$ are the children of $U$.

The resulting tree $S$ is the normalized triplet supertree or NTS.
The minimum threshold is $\tau = 1/2$.

The children of $\{1,2,3,4,5\}$ in the minimal threshold tree are $\{1\}$ and $\{2,3,4,5\}$. 
A({2,3,4,5})

The minimum threshold is $\tau = 1/2$. 
$A(\{2,3,4,5\})$

The minimum threshold is $\tau = 1/2$.
The children of $\{2,3,4,5\}$ are $\{2\}$ and $\{3,4,5\}$. 
A({3,4,5})

The minimum threshold is $\tau = 1/3$.
The children of \{3,4,5\} are \{3\} and \{4,5\}.
The normalized triplet supertree NTS has clusters 
\{1,2,3,4,5\}, \{2,3,4,5\}, \{1\}, \{2\}, \{3,4,5\}, \{3\}, \{4,5\}, \{4\}, \{5\}.
Problem. Study how well a supertree method reconciles incompatible data.

Let $D = \{T_i : i \in \Lambda\}$ be a collection of input trees, where $T_i$ has leaf set $L(T_i)$. We seek a supertree for $D$.

We say that collection $D$ is dense if for every three distinct taxa $a, b, c$ in $X$, there exists an input tree $T_i$ such that $\{a, b, c\} \subseteq L(T_i)$.

A special case of dense data arises when each $T_i$ has leafset $X$. In this case, the problem is the same as finding a consensus tree.

We study how various methods work on dense data.
Suppose $T$ is a rooted tree and $a,b,c$ are leaves.

\[
spt_T(ab|c) = \begin{cases} 
1 & \text{if } ab|c \text{ in } T \\ 
0 & \text{otherwise} 
\end{cases}
\]

$T$

\[
spt_T(34|2) = 1 \\
spt_T(12|4) = 0 \\
spt_T(24|1) = 1
\]
Main Theorem. Suppose $\mathcal{D}$ is dense and $T$ is a binary rooted tree.

Assume that for all distinct $x, y, z$ in $X$ we have
$$\left| spt_{\mathcal{D}}(xylz) - spt_T(xylz) \right| < 1/2.$$

Then the normalized triplet supertree $S$ is $T$. 
The condition
\[ | \text{spt}(xyz) - \text{spt}_T(xyz) | < 1/2 \]
for all \( x, y, z \) means

(1) If \( xyz \) in \( T \), then more than half the input \( \{x,y,z\} \) triplets are \( xyz \).
(2) If \( xyz \) is not in \( T \), then fewer than half the input \( \{x,y,z\} \) triplets are \( xyz \).

In this situation, the NTS is \( T \).
If T is not binary, there is a more complicated statement.

**Theorem.** Suppose $\mathcal{D}$ is dense.
Assume that for all distinct x, y, z in X we have
$$|\text{spt}_{\mathcal{D}}(xyz) - \text{spt}_T(xyz)| < 1/2.$$  
Then the normalized triplet supertree S is a refinement of T.
Following Atteson 1999, say that the (dense $l_\infty$) radius $R$ of a method of supertree reconstruction is $\alpha$ provided that,

(1) Whenever $D$ is a dense collection of input trees and $T$ is a binary tree such that for all distinct $a, b, c$

$$|spt_D(ab|c) - spt_T(ab|c)| < \alpha$$

then the method constructs $T$; and

(2) For every $\beta > \alpha$ there exists a binary tree $T$ and a dense collection $D$ of input trees such that for all distinct $a, b, c$

$$|spt_D(ab|c) - spt_T(ab|c)| < \beta$$

but the method does not reconstruct $T$.

Alternatively, call such $\alpha$ the dense robustness radius $R$.

**Theorem.** The normalized triplet supertree method has $R \geq 1/2$. 
**Theorem.** $R = 1/2$ is best possible for any method.

Suppose that we have as input $n$ copies of $T$ and $m$ copies of $W$.

- $\text{spt}_T(ab|c) = 1$
  \[
  |\text{spt}_T(ab|c) - \text{spt}_D(ab|c)| = m/(n+m)
  \]
- $\text{spt}_T(ba|c) = 0$
  \[
  |\text{spt}_T(ba|c) - \text{spt}_D(ba|c)| = m/(n+m)
  \]
- $\text{spt}_W(ac|b) = 0$
  \[
  |\text{spt}_W(ac|b) - \text{spt}_D(ac|b)| = 0
  \]

\[
\begin{align*}
\text{spt}_D(ablc) &= n/(n+m) = \text{mspt}(a,b) \\
\text{spt}_D(bcla) &= m/(n+m) = \text{mspt}(b,c) \\
\text{spt}_D(aclb) &= 0
\end{align*}
\]
Suppose a method yields a tree $Y$ when for all distinct $x$, $y$, $z$ in $X$
\[ |\text{spt}_D(xyz) - \text{spt}_Y(xyz)| < \beta \quad \text{with} \quad \beta > 1/2.\]

We could choose $m > n$ with $m/(n+m) < \beta$.
Then also $n/(n+m) < \beta$.
So the method would have to output both $T$ and $W$. 
Theorem. The normalized triplet supertree method has dense robustness radius $R = 1/2$. 
Complexity of NTS

**Theorem.** Let $X$ be a finite set, $|X| = n$, and let $D$ be a collection of $m$ input rooted trees with leaves from $X$.

The computation of the normalized triplet supertree $S$ takes time

$$O(n^4) + O(n^3 m).$$
Some Supertree Methods


(2) MinCutSupertree. (Semple and Steel 2000)

(3) PageMinCutSupertree. (modified by Page 2003)

(4) Matrix Representation using Flipping (MRF). (Burleigh et al 2004)

(5) Normalized Triplet Supertree (NTS).
<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTS</td>
<td>polynomial</td>
</tr>
<tr>
<td>MRP</td>
<td>NP-hard</td>
</tr>
<tr>
<td>MinCutSupertree (orig)</td>
<td>polynomial</td>
</tr>
<tr>
<td>MinCutSupertree (Page)</td>
<td>polynomial</td>
</tr>
<tr>
<td>triplet MRP</td>
<td>NP-hard</td>
</tr>
<tr>
<td>MRF</td>
<td>NP-hard</td>
</tr>
<tr>
<td>Method</td>
<td>Dense robustness radius</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>NTS</td>
<td>1/2</td>
</tr>
<tr>
<td>MRP</td>
<td>0 [R ≤ 0.05]</td>
</tr>
<tr>
<td>MinCutSupertree (orig)</td>
<td>0</td>
</tr>
<tr>
<td>MinCutSupertree (Page)</td>
<td>0</td>
</tr>
<tr>
<td>triplet MRP</td>
<td>1/2</td>
</tr>
<tr>
<td>triplet MRF</td>
<td>1/2</td>
</tr>
</tbody>
</table>
Interpretation:

There exists a binary tree $T$ and dense input data $\mathcal{D}$ such that
(1) whenever $xylz$ in $T$, then at least 99% of the input $\{x,y,z\}$ triplets are $xylz$; but
(2) the MinCutSupertree is not $T$. 
A special case is consensus trees, when all input trees have the same leaf set.
Majority consensus tree $M$

although if you include clusters in order of occurrence until you fail to get a tree, then we obtain
Strict Consensus

Adams Consensus

Normalized Triplet Consensus = MinCutSupertree Consensus
Consensus tree computation (all input trees have leaf set X)

<table>
<thead>
<tr>
<th>Method</th>
<th>dense robustness radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTS</td>
<td>1/2</td>
</tr>
<tr>
<td>MRP</td>
<td>0</td>
</tr>
<tr>
<td>strict consensus</td>
<td>0</td>
</tr>
<tr>
<td>triplet MRP</td>
<td>1/2</td>
</tr>
<tr>
<td>triplet MRF</td>
<td>1/2</td>
</tr>
<tr>
<td>majority consensus</td>
<td>0</td>
</tr>
<tr>
<td>Adams consensus</td>
<td>0</td>
</tr>
<tr>
<td>MinCutSupertree (orig)</td>
<td>0</td>
</tr>
<tr>
<td>MinCutSupertree (Page)</td>
<td>0</td>
</tr>
</tbody>
</table>

NTS is much slower than strict consensus and majority consensus which take time $O(n \cdot m)$. And NTS applies only to rooted trees.
There are many variants on the method, also with radius 1/2, which may possibly extrapolate better than NTS.

**Theorem.** Suppose a supertree method utilizes minimal threshold trees with the edge support $spt_U(a,b)$ for the edge \{a,b\} in $A(U)$. Suppose
(1) if there exists $c \in U$ with $spt_D(ab|c) > 1/2$, then $spt_U(a,b) > 1/2$; and
(2) if for every $c \in U$ we have $spt_D(ab|c) < 1/2$, then $spt_U(a,b) < 1/2$.
Then the dense robustness radius is 1/2.
Lemma. Suppose that $\mathcal{D}$ is dense and $T$ is a tree with leaf set $X = \bigcup L(T_i)$. Assume that for all distinct $x, y, z$ in $X$ we have
\[ |\text{spt}_\mathcal{D}(xyz) - \text{spt}_T(xyz)| < 1/2. \]
Suppose that the tree $T$ has a cluster $U$ with exactly 2 children $A$ and $B$. Suppose $S$ is the NTS and $U$ in $S$. Then $A^{1/2}(U)$ is disconnected and its components are exactly $A$ and $B$. The children of $U$ in $S$ are $A$ and $B$. 
Lemma. Suppose that $\mathcal{D}$ is dense and $T$ is a tree with leaf set $X = \bigcup L(T_i)$. Assume that for all distinct $x, y, z$ in $X$ we have

$$|\text{spt}_{\mathcal{D}}(xyz) - \text{spt}_T(xyz)| < \frac{1}{2}.$$ 

Suppose that the tree $T$ has a cluster $U$ with exactly 2 children $A$ and $B$. Suppose $S$ is the NTS and $U$ in $S$. Then $A_{1/2}(U)$ is disconnected and its components are exactly $A$ and $B$. The children of $U$ in $S$ are $A$ and $B$.

Consider $a_1$ and $a_2$ in $A$. Choose $b_1$ in $B$. Then $a_1 \ a_2 \ b_1$ in $T$. Hence $\text{spt}_T(a_1 \ a_2 \ b_1) = 1$. Hence $\text{spt}_{\mathcal{D}}(a_1 \ a_2 \ b_1) > \frac{1}{2}$. Hence $\text{mspt}(a_1, a_2) > \frac{1}{2}$. Hence there is an edge $\{a_1, a_2\}$. 
Similarly there is an edge \{a_i, a_j\} in $A_{1/2}(U)$.
And there is an edge \{b_i, b_j\} in A_{1/2}(U).
Is there an edge \{a_1, b_1\}? Consider any \(c \neq a_1, c \neq b_1\). Say \(c = a_2\).
Then \(a_1 \ c \mid b_1\) in \(T\).
\[\text{spt}_T(a_1 \ c \mid b_1) = 1.\]
\[\text{spt}_T(a_1 \ b_1 \mid c) = 0.\]
\[\text{spt}_D(a_1 \ b_1 \mid c) < 1/2\]

Indeed \(\text{spt}_D(a_1 \ b_1 \mid c) < 1/2\) for all \(c \neq a_1, c \neq b_1\).
Hence \(\text{mspt}(a_1, b_1) < 1/2\).
There is no edge \(\{a_1, b_1\}\) in \(A_{1/2}(U)\).
A_{1/2}(U)

The children of U in the NTS are A and B.
General Properties of NTS

Wilkinson et al 2004 propose some "desiderata for liberal supertrees."

(1) **Uniqueness.** The method should give a unique answer.

(2) **Plenary.** The resulting supertree should contain all the leaves of the input trees.

(3) **Order invariance.** The method should not be influenced by the order in which members of $\mathcal{D}$ are introduced, or the order of leaves in the adjacency matrix of an input tree.

(4) **Sizeless.** The method should not be biased by input trees with large size. Note that if $\mathcal{D} = \{T_i: i \in \Lambda\}$, then NTS is precisely the same as the multiset $\mathcal{D}'$ where each $T_i$ is replaced by the set of its rooted triples and star triples.

(5) **Shapeless.** The method should not be biased by the shape of input trees.
(6) **Pareto.** Relationships in all input trees should appear in the output.

**Theorem.** Suppose that $D$ contains at least one input tree containing all the taxa $X$. Suppose $\text{spt}_D(ab|c) = 1$. Then $ab|c$ in the NTS.

(7) **Weightable.** The method should allow different trees to be weighted.

(8) **Speed.** NTS has polynomial time-complexity.

(9) **Assessable.** The method should allow a measure of the amount of support of the output supertree. The smallest number $\alpha$ such that for all $x, y, z$,

$$|\text{spt}_D(xy|z) - \text{spt}_T(xy|z)| \leq \alpha$$

gives a measure of the overall support for a tree $T$. A smaller $\alpha$ means a better fit.
Real data

Philip, Creevey, and McInerney (Mol. Biol. Evol. 2005) considered 780 single-gene trees for a set of ten eukaryotes. The numbers of taxa in a given tree ranged widely. They performed a Most Similar Supertree analysis using the software package CLANN. They published and analyzed the resulting tree.
et al. (2003). The data were then analyzed so that the process of substitution from one category to another (general time-reversible) was optimized, as well as the tree topology using an MCMC approach with the p4 software (http://www.nhm.ac.uk/zoology/external/p4.htm). The resulting support values for every internal node in figure 1 was unity (analysis kindly carried out by Peter Foster). This analysis effectively only analyses substitutions across, not within, Dayhoff categories and is similar to a transversion analysis of DNA sequences.

Finally, we carried out a variety of tests of alternative hypotheses using the most conservative alignment (the alignment with site categories 4, 5, 6, 7, and 8 removed). The results are to be found in table 3. All alternatives were rejected at least once at the 5% level by one test, and most alternatives were rejected at least twice. In other words, changing the topology is a significantly worse interpretation of the evolution of these sequences.

**Discussion**

We have addressed a phylogenetic analysis problem where there are five known nodes and only two unknown nodes. We have only used provably single-copy genes from completed genomes, and therefore this analysis is different to most other analyses in this respect. In the results presented here, support for the Coelomata hypothesis was

![Diagram](image-url)

Fig. 2.—Relative support for competing hypotheses. Data set 8 indicates that category 8 has been removed, while data set 7 + 8 indicates that categories 7 and 8 have been removed, and so forth. (a) The bootstrap support values for the three possible topologies for the interrelationships of the three kingdoms: animals (A), plants (P), and fungi (F) where the data are not controlled. (b) The bootstrap support values for the three kingdoms using input data sets screened for sequence length and ability to recover “known” nodes. (c) Bootstrap proportions for the three possible relationships of the animal phylogeny using uncontrolled data. (d) Bootstrap support for the animal phylogeny using controlled data. The relationships in (a) and (b) are hypothesized to have diverged 950–1,259 MYA, while the relationships in (c) and (d) are hypothesized to have diverged 642–761 MYA (Douzery et al. 2004).

---

**Table 1**

Statistics for the Data Sets Used in This Study

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Length</th>
<th>Uncontrolled</th>
<th>Controlled</th>
<th>Total Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Animal-Plant</td>
<td>Animal-Plant</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InvG</td>
<td>436,450</td>
<td>100</td>
<td>91</td>
<td>480</td>
</tr>
<tr>
<td>8</td>
<td>304,759</td>
<td>98</td>
<td>76</td>
<td>350</td>
</tr>
<tr>
<td>7 + 8</td>
<td>270,733</td>
<td>94</td>
<td>75</td>
<td>299</td>
</tr>
<tr>
<td>6 + 7 + 8</td>
<td>241,555</td>
<td>98</td>
<td>60</td>
<td>255</td>
</tr>
<tr>
<td>5 + 6 + 7 + 8</td>
<td>211,929</td>
<td>90</td>
<td>67</td>
<td>203</td>
</tr>
<tr>
<td>4 + 5 + 6 + 7 + 8</td>
<td>180,869</td>
<td>86</td>
<td>50</td>
<td>153</td>
</tr>
<tr>
<td>1 + 8</td>
<td>170,128</td>
<td>76</td>
<td>68</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,452</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>870</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>747</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>620</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>516</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>725</td>
</tr>
</tbody>
</table>

**Note.**—The rows refer to the particular treatment of the data, whether the data were analyzed using a method that deals effectively with rate variation across sites or whether the data were stripped of different categories of the fastest-evolving sites. The numbers in the various columns refer to support values for the uncontrolled or the controlled single-gene trees or the concatenated alignment. NA means not applicable.
I used the 664 trees (out of 780 obtained at the web site http://bioinf.nuim.ie/supplementary/eukaryotes/ ) that contained humans.

The NTS is topologically the same tree as the one they published.

The NTS satisfied that for all $x, y, z$

\[ |\text{spt}_D(xyz) - \text{spt}_T(xyz) | \leq 0.61111 \]
Geometric interpretation

Consider the collection of rooted X-trees, where \(|X| = n\). Every rooted X-tree \(T\) is uniquely determined by its set of rooted triples \(abc\). The total number of possible rooted triples is

\[
m = \binom{n}{3} = \frac{n(n-1)(n-2)}{2}
\]

Let \(H_X = [0,1]^m\) be the \(m\)-dimensional hypercube in which each coordinate corresponds to a rooted triple \(abc\).

\(H_X\). There are \(m\) dimensions.
$H_X$ is a kind of rooted-tree-landscape space.

To any rooted $X$-tree $T$ there corresponds $spt_T \in H_X$ given by $(spt_T)_{ab|c} = spt_T(ab|c)$.

Each rooted $X$-tree corresponds to a distinct corner of $H_X$. No two distinct trees correspond to the same corner. Not all corners correspond to trees.
To every collection $D$ of rooted trees with leaf-sets contained in $X$, there is a point $\text{spt} D \in H_X$ given by $(\text{spt} D)_{ab|c} = \text{spt} D(ab|c)$.

The $\ell_\infty$ norm on $H_X$ is defined by

$$\|u-v\|_\infty = \max \{|u_{ab|c} - v_{ab|c}|\}$$

If \text{spt} $T$ is the upper front left corner, then the red box is

$$\{u \in H_X : \|u - \text{spt} T\|_\infty < 1/2\}.$$
**Theorem.** Suppose $T$ is binary and
\[ \| \text{spt } \mathcal{D} - \text{spt } T \|_\infty < 1/2. \]
Then the NTS is $T$.

If $\text{spt } T$ is the upper front left corner, and $\text{spt } \mathcal{D}$ is in the red box, then the NTS is $T$. 
The picture is deceptive since \( m \) is very large.

If \( X \) contains 7 taxa, then \( m = 105 \).

The volume of \( H_X \) is 1.

The volume of the red box is \( \left( \frac{1}{2} \right)^{105} \approx 2.465 \times 10^{-32} \).

The total volume of all red boxes for all 10,395 binary rooted trees with 7 taxa is \( 2.563 \times 10^{-28} \).

But this ignores the 35 constraints that for each \( a, b, c \)

\[ \text{spt}_D(ablc) + \text{spt}_D(aclb) + \text{spt}_D(bcla) \leq 1. \]
For the Philip et al. dataset with 10 eukaryotes, if $T = NTS$ then

$$\| \text{spt} \mathcal{D} - \text{spt} T \|_{\infty} = 0.61111$$

Here $\dim(H_X) = 360$. 

Here $\dim(H_X) = 360$. 
**Closest Tree Problem.**

Given the set of input trees $D$, find a rooted $X$-tree $T$ that minimizes

$$\|spt_D - spt_T\|_\infty$$

The NTS solves this problem if the solution $T$ is binary and satisfies

$$\|spt_D - spt_T\|_\infty < 1/2.$$
**Closest Tree Problem.** Given the set of input rooted trees $\mathcal{D}$, find a rooted $X$-tree $T$ that minimizes

$$\| spt_{\mathcal{D}} - spt_{T} \|_{\infty}$$

Note

$$\| spt_{\mathcal{D}} - spt_{T} \|_{\infty} = \max \{ | spt_{\mathcal{D}}(xy|z) - spt_{T}(xy|z) | \}$$

$$= \max \{ \max \{ spt_{\mathcal{D}}(xy|z) : xy|z \text{ not in } T \}, \max \{ (1 - spt_{\mathcal{D}}(xy|z)) : xy|z \text{ in } T \} \}$$

Let $D_{\text{weak}}(T) := \max \{ spt_{\mathcal{D}}(xy|z) : xy|z \text{ not in } T \}$. 
**Weak Closest Tree Problem.** Given the set of input rooted trees $\mathcal{D}$, find a rooted X-tree $T$ that minimizes

$$D_{\text{weak}}(T) := \max \{ \text{spt}_{\mathcal{D}}(xyz) : xyz \text{ not in } T \}.$$
**Weak Closest Tree Problem.** Given the set of input trees $\mathcal{D}$, find an X-tree $T$ that minimizes

$$D_{\text{weak }}(T) := \max \{ \text{spt } \mathcal{D}(xyz) : xyz \text{ not in } T \}.$$

**Theorem.** The Normalized Triplet Supertree $S$ solves the Weak Closest Tree Problem.

There is typically not a unique solution.
**Theorem.** Suppose every input tree in $\mathcal{D}$ is binary. Suppose $\mathcal{D}$ is dense. Suppose $T$ solves the Closest Tree Problem. Let $S$ be the Normalized Triplet Supertree. Then

$$\text{Dweak}_{\mathcal{D}}(S) \leq \|spt_T - spt_{\mathcal{D}}\|_\infty \leq 2 \text{Dweak}_{\mathcal{D}}(S).$$
Comparison with MinCutSupertree

Suppose $m$ is a positive integer. If $m = 5$, we have the input

$\begin{array}{cccccc}
T & a & b & 1 & 2 & 3 \\
& & & 4 & 5 & \\
Y & a & b & 1 & 2 & 3 \\
& & & 4 & 5 & \\
\end{array}$

with $t$ copies of $T$ and $y$ copies of $Y$
Input also one copy of $Z_1, ..., Z_m$. 
The data are close to $T$.

\[
spt_T(ij \mid a) = 1 \\
spt_T(ij \mid b) = 1 \\
spt_T(ab \mid i) = 1 \\
spt_T(jk \mid i) = 1 \text{ if } i < j < k \\
spt_T(uv \mid w) = 0 \text{ otherwise}
\]

\[
spt_D(ij \mid a) = \frac{t + y + m - 2}{t+y+m} \\
spt_D(ij \mid b) = \frac{t + y + m - 2}{t+y+m} \\
spt_D(ab \mid i) = \frac{t + 1}{t + y + m} \\
spt_D(jk \mid i) = \frac{t+y + m - 2}{t + y + m} \text{ if } i < j < k
\]
Hence $|spt_D(uvw) - spt_T(uvw)| \leq \frac{(y + m - 1)}{(t + y + m)}$

We will choose $t = my + m^2 - 2m - 1$. For all $u, v, w$, as $y \to \infty$

$|spt_D(uvw) - spt_T(uvw)| \leq \frac{1}{(m+1)}$.

But a check shows that when $U = X$, $\{a,b\}$ is not a minimal cut and $\{a,b\}$ is not a child of the root in $\text{MinCutSupertree}$.

Hence for each $m$ we have $R \leq \frac{1}{(m+1)}$.

Hence $R = 0$. 
The same example applies to Page MinCutSupertree.
Distance Methods

Suppose we know the distances on the input trees. Then there are many more constraints.

Figure 1. A family $D$ of input rooted trees. Branch lengths in red.
The unique supertree with distances from Figure 1.
Figure 2. A different family of input trees with the same topology as in Figure 1 but different branchlengths.
The unique supertree with distances for Figure 2.
Figure 3. The supertree constructed by BUILD or NTS from the input trees in Figures 1 or 2, ignoring distances.

Moral: The use of distances can give extra resolution.
How do we use distances?

\( \lambda(a,b) \) is the length of the path in the graph of T from \( \{a\} \) to the most recent common ancestor of a and b.

Example:

\( \lambda(1,4) = 3 \)
\( \lambda(4,1) = 1 \)
\( \lambda(4,7) = 3 \)
\( \lambda(7,4) = 2 \)
d corresponds to the distance from mrca(1,2) to the root.

If \( \lambda \) comes from the tree, then
\[
\lambda(1,4) - \lambda(1,2) = d \\
\lambda(1,5) - \lambda(1,2) = d \\
\lambda(1,3) - \lambda(1,2) < d
\]

\[
d = \max\{\lambda(1,c) - \lambda(1,2) : c \in X, c \text{ distinct from 1 and 2}\}
\]
Given $\mathcal{D}$ in which each input tree has distances, in each input tree $T_i = (V_i, E_i, r, X_i)$ compute $\lambda_i(a,b)$ for $a$ and $b$ in $X_i$.

Let $\lambda_{\mathcal{D}}(a,b)$ be the average of the $\lambda_i(a,b)$.
$\lambda_{\mathcal{D}}$ is a partially defined function on $X \times X$. 
We wish to use a minimal threshold procedure to build a supertree from $\mathcal{D}$.

Suppose $U$ is a cluster. Form $A(U)$.

Define the support for the edge $\{a,b\}$ to be the estimated distance from $\text{mrca}(a,b)$ to the root for $U$. The bigger is this estimate, the more confident we are in the edge $\{a,b\}$.

Here $\text{spt}(1,2) = \lambda(1,4) - \lambda(1,2)$
Two possible formulas for support:

The **primary support** of \((a,b)\) is
\[
\text{pspt}(a,b) = \max \{ \lambda(a,c) - \lambda(a,b), \lambda(b,c) - \lambda(b,a) : c \in U \}
\]

The **confirmed support** of \((a,b)\) is
\[
\text{cspt}(a,b) = \max \{ \min \{ \lambda(a,c) - \lambda(a,b), \lambda(b,c) - \lambda(b,a) \} : c \in U \}.
\]

If \(\lambda\) is globally defined and is correct on the tree, these give the same numbers. If \(\lambda\) contains errors, these can be different numbers.
Algorithm BUILD-WITH-DISTANCES (W. 2004)

Suppose U is a nonempty set of vertices. Define A(U) as follows:

(i) The vertex set consists of the members of U.

(ii) The weight for edge \{a,b\} is pspt(a,b), provided it is positive.

Find the minimal $\tau$ such that $A_\tau(U)$ disconnects, and let the children of U be the components of $A_\tau(U)$.

The result is the **minimal threshold supertree** $S^P$ **with primary support**.

Or if instead the weight for edge \{a,b\} is cspt(a,b), obtain the **minimal threshold supertree** $S^C$ **with confirmed support**.
Example. Use the primary support function pspt:

\[ U = \{1, 2, 3, 4, 5, 6\} \]

\[ \lambda(2,1) - \lambda(2,6) = 0 \]
\[ \lambda(2,3) - \lambda(2,6) = 2 \]
\[ \lambda(2,4) - \lambda(2,6) = 2 \]
\[ \lambda(2,5) - \lambda(2,6) = 0.5 \]

\[ \lambda(6,1) - \lambda(6,2) = 0.5 \]
\[ \lambda(6,3) - \lambda(6,2) = \text{undef} \]
\[ \lambda(6,4) - \lambda(6,2) = -1.5 \]
\[ \lambda(6,5) - \lambda(6,2) = 0.5 \]

\[ \text{pspt}(2,6) = 2 \]
Example. Use the confirmed support function cspt:

\[ U = \{1,2,3,4,5,6\}. \]

\[ \lambda(2,1)-\lambda(2,6) = 0 \quad \lambda(6,1)-\lambda(6,2) = 0.5 \]
\[ \lambda(2,3)-\lambda(2,6) = 2 \quad \lambda(6,3)-\lambda(6,2) = \text{undef} \]
\[ \lambda(2,4)-\lambda(2,6) = 2 \quad \lambda(6,4)-\lambda(6,2) = -1.5 \]
\[ \lambda(2,5)-\lambda(2,6) = 0.5 \quad \lambda(6,5)-\lambda(6,2) = 0.5 \]

\[ \text{cspt}(2,6) = 0.5 \]
NTS = MRP strict consensus of 18 MP trees
Suppose that the input trees $\mathcal{D} = \{T_i: i \in \Lambda\}$ have distances. Say that $\mathcal{D}$ is **distance-dense** if for each distinct $a$ and $b$ there exists $i \in \Lambda$ with $\{a, b\} \subseteq L(T_i)$.

**Theorem** Suppose that the tree $T$ is binary, the shortest branchlength is $\mathcal{K} > 0$, and the distance function on $T$ is $\lambda_T$.

Suppose that each tree in $\mathcal{D} = \{T_i: i \in \Lambda\}$ has distances and $\mathcal{D}$ is distance-dense. For each $a$ and $b$, let $\lambda_{\mathcal{D}}(a,b)$ be the estimated value

$$\lambda_{\mathcal{D}}(a,b) = \text{average}\{\lambda_i(a,b): a \text{ and } b \text{ are in } L(T_i)\}.$$  

(1) Assume for all $x$ and $y$

$$|\lambda_{\mathcal{D}}(x,y) - \lambda_T(x,y)| < \frac{1}{2} \mathcal{K}$$

Then $S^c = T$.

(2) Assume for all $x$ and $y$

$$|\lambda_{\mathcal{D}}(x,y) - \lambda_T(x,y)| < \frac{1}{4} \mathcal{K}$$

Then $S^p = T$.

(3) The numbers $(1/2)$ and $(1/4)$ above are the largest possible for the respective methods.
Following Atteson 1999, say that the (distance-dense $l_\infty$) radius $r$ of a method of supertree reconstruction from trees with distances is $\alpha$ provided that,

1. Suppose $\mathcal{D}$ is a distance-dense collection of input trees with distances yielding the function $\lambda$. Suppose $T$ is a binary tree with function $\lambda_T$ and minimal branchlength $K$. Suppose that for all distinct $x, y$
   \[ |\lambda_D(x,y) - \lambda_T(x,y)| < \alpha K. \]
   Then the method constructs $T$.

2. For every $\beta > \alpha$ there exists a binary tree $T$ with distances and shortest branchlength $K$, and there exists a distance-dense collection $\mathcal{D}$ of input trees such that for all distinct $x, y$
   \[ |\lambda_D(x,y) - \lambda_T(x,y)| < \beta K. \]
   but the method does not reconstruct $T$.

Alternatively, call such $\alpha$ the distance-dense robustness radius $r$. 
Theorem.
(1) The primary supertree method has $r = 1/4$.
(2) The confirmed supertree method has $r = 1/2$.
(3) $r = 1/2$ is best possible for any method.
Summary.

The confirmed supertree method is twice as robust as the primary supertree method on distance-dense data and has optimal distance-dense robustness radius.
Review

(1) Different methods for finding supertrees can differ significantly on how well they perform on dense data to reconcile incompatibilities.

(2) For topological input, Triplet MRP and NTS can reconcile incompatibilities close to a tree better than MRP or MinCutSupertree.

(3) For consensus methods, NTS reconciles incompatibilities close to a tree better than Majority Consensus and Strict Consensus.

(4) NTS solves an optimization problem in tree-space.

(5) NTS is polynomial-time and has other nice properties.

(6) Distance methods can add resolution if the input trees have distances.

(7) For distance-dense data, the confirmed supertree is more robust than the primary supertree.
Questions

(1) Is there an analogous quantitative measure for how well a supertree method extrapolates data?

(2) Which topological methods with radius 1/2 extrapolate well?

(3) What is the complexity of the Closest Tree Problem?

(4) What is the dense robustness radius of MRF and other methods?

(5) What happens if we use a different norm on $H_X$, e.g.

$$||u - v||^2_2 = \sqrt{\sum (u_{abc} - v_{abc})^2}$$
Acknowledgments

Thanks to the Isaac Newton Institute for Mathematical Sciences for its hospitality at an excellent facility.

Thanks to the e-Science Institute for hosting this Phyloinformatics Workshop.

Thanks to the organizers Vincent Moulton, Mike Steel, and Daniel Huson of the Phylogenetics Programme.

Thanks to Iowa State University for giving me sabbatical this semester.
Thank you for your attention.
References


