Topological Concordance of Gene Trees and Species Tree

jatin narula
Gene Trees and Species Tree

- Gene Loss and Duplication
- Horizontal Gene Transfer and Recombination
- Stochastic Factors
Topological Concordance

Takahata Congruence => Topological Concordance
Topological Concordance for Multiple Lineages

Collapsed gene tree is both topologically concordant and Takahata congruent.
Monophyletically Concordant

Monophyletically Concordant

All species are Monophyletic but gene tree is not topologically concordant

Monophyletic + Topologically Concordant = Monophyletically Concordant
Speciodendric Genes

Orthology: Genes whose homology was the result of speciation and subsequent descent, with no duplication.

Speciodendricity: Gene Tree constructed from all copies of the gene in all species is topologically concordant.

Speciodendric: Orthologous

Xenologous: Paralogous
The Problem

“conditioned on the species tree topology and assuming no gene exchange between species, what is the probability that a tree of orthologous genes is topologically concordant with a species tree?”
Takahata Concordance Probability

\[ P(\text{Takahata Congruence}) = \sum_{m=1}^{m} \sum_{n=1}^{s} \sum_{k=1}^{m+n} g_{rm}(T_3) * g_{sn}(T_3) * g_{m+n,k}(T_2) * F_{k,A,B}(m,n,0) \]

\( P(\text{r lineages derive from m lineages at time } T_3) * P(\text{s lineages derive from n lineages at time } T_3) \)

\* \( P(\text{m+n lineages at } T_3 \text{ derive from } k \text{ lineages at time } T_3+T_2) \)

\* \( P(\text{at least one interspecific coalescence occurs during this process, and that the most recent interspecific coalescence joins a lineage from species A and a lineage from species B}) \)
Topological Concordance Probability

\[
P(\text{Topological Concordance}) =
\sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \left[ g_{rm}(T_3) g_{sn}(T_3) g_{m+n,k}(T_2) \times \left[ F_{k}^{A,B}(m, n, 0) + [1 - F_{k}^{A,B}(m, n, 0)] \times \sum_{x=1}^{k-1} \left[ W_{(m, n), (x, k-x)}(T_2) \right] \right] \right] \times \sum_{l=1}^{q} \left[ g_{yl}(T_3 + T_2) F_{l}^{A,B}(x, k-x, l) \right].
\]
Key Determinants of Topological Concordance:

$$T_2$$

Concordance Highly Likely

Small $$T_2$$

Large $$T_2$$

Trifurcation

$$(r,s,q)=(10,10,10), T_3=0.05$$

$$(r,s,q)=(10,10,10), T_3=2$$

$$(r,s,q)=(1,1,1), T_3=any$$

$$(r,s,q)=(1,1,10), T_3=0.05$$
Key Determinants of Topological Concordance: $T_3$

- Small $T_3$
- Large $T_3$
- Concordance Unlikely
Key Determinants of Topological Concordance:
Sample Sizes

Large T3

Small T3
Probability of Speciodendricity

\[ P(\text{Speciodendricity}) = P(\text{Topological Concordance} \mid \text{Sample sizes} = \text{no. of copies of gene in respective species}) \]

\[ \approx P(\text{Topological Concordance} \mid \text{Sample sizes} = \infty) \]
Maximal Useful Sample Sizes

\[
|E[A_{T_3} | A_0 = r] - E[A_{T_3} | A_0 = \infty]| < \varepsilon,
\]

| $T_3$ | Large-sample limiting mean number of ancestral lineages at time $T_3$ | $\log_2(\varepsilon)$ | Lower bound $R$ | Mean number of ancestral lineages at time $T_3$ with a sample size of $R$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.020</td>
<td>−3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1.418</td>
<td>−1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−2</td>
<td>3</td>
<td>1.204</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−3</td>
<td>6</td>
<td>1.294</td>
</tr>
<tr>
<td>1</td>
<td>2.370</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>3</td>
<td>1.577</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−1</td>
<td>6</td>
<td>1.879</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−2</td>
<td>14</td>
<td>2.126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−3</td>
<td>30</td>
<td>2.248</td>
</tr>
<tr>
<td>0.05</td>
<td>40.335</td>
<td>4</td>
<td>61</td>
<td>24.474</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>161</td>
<td>32.373</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>361</td>
<td>36.345</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>760</td>
<td>38.335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1560</td>
<td>39.335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−1</td>
<td>3160</td>
<td>39.835</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−2</td>
<td>6360</td>
<td>40.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−3</td>
<td>12759</td>
<td>40.210</td>
</tr>
</tbody>
</table>
Estimation of Parameters $T_2$, $T_3$ and $N$

- Known Species Tree Topology
- Choose multiple independent loci
- Construct Gene trees assuming values of $T_2$, $T_3$
- Estimate Likelihood of Parameter Values

\[
\text{Lik}(T_3, T_2) \propto P_C(r, s, q, T_3, T_2)^x Q_{(AC)B}(r, s, q, T_3, T_2)^y \\
\times Q_{(BC)A}(r, s, q, T_3, T_2)^z.
\]  

(20)
### Extension to Four or More Species

#### Balanced

<table>
<thead>
<tr>
<th>Topology Is (((AB)(CD)))</th>
<th>Gene tree topology</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(((AB)(CD)))</td>
<td>(g_{21}(T_3 + T_2) g_{21}(T_2))  + (g_{22}(T_3 + T_2) g_{22}(T_2) \frac{1}{3})  + (g_{22}(T_3 + T_2) g_{22}(T_2) \frac{1}{5})  + (g_{22}(T_3 + T_2) g_{22}(T_2) \frac{1}{5})</td>
<td>0.6867</td>
</tr>
</tbody>
</table>

#### Unbalanced

<table>
<thead>
<tr>
<th>Topology Is (((AB)\ C)\ D))</th>
<th>Gene tree topology</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(((AB)\ C)\ D))</td>
<td>(g_{21}(T_3)\left[ g_{22}(T_2) + g_{22}(T_2) \frac{1}{3}\right] + g_{22}(T_3))  \times \left[ g_{31}(T_3) \frac{1}{3} + g_{32}(T_3) \frac{1}{3} + g_{33}(T_3) \frac{1}{3}\right]</td>
<td>0.5556</td>
</tr>
</tbody>
</table>
Summary

• Likelihood functions for observed gene tree conditioned on proposed species tree
• Inference of most likely Species History
• Estimation of optimal sample sizes that maximize concordance probability
• Estimation of divergence times and ancestral population sizes
• Identification of Speciation Genes
• Assumes equality and stability of population sizes
• Ignores gene exchange, mistaken orthology and other stochastic effects that cause discordance
References

probability that two lineages coalesce in the immediately preceding generation = probability that they share a parent = $1/N$

$P_c(t \text{ generations}) = (1 - 1/N)^{t-1}(1/N) \approx \exp(-t/N)/N$ (for large $N$)

Let $T = t/N$

Probability that $p$ lineages coalesce to $p-1$ at $T$

$f_{p-1}(T) = p\ast (p-1)\ast \exp(-T)/2$

$T_{p-1}$ is the waiting time for $p$ lineages to coalesce to $p-1$ with distribution $f_{p-1}$

Define $S_{ij} = \sum_{p=j}^{p=i-1} T_p$

Then $P(i \text{ lineages converge to } j \text{ in } T) = P(S_{ij} = T) = g_{ij}(T)$

(Hudson 1983, Takahata and Nei 1985)

$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{(k)}}{j! (k-j)! i_{(k)}}$$

where $a_{(k)} = a(a+1) \cdots (a+k-1)$ for $k \geq 1$ with $a_{(0)} = 1$; and $a_{(k)} = a(a-1) \cdots (a-k+1)$ for $k \geq 1$ with $a_{(0)} = 1$.

$$g_{11}(T) = 1, \quad g_{21}(T) = 1 - e^{-T}, \quad g_{31}(T) = 1 - \frac{3}{2} e^{-T} + \frac{1}{2} e^{-3T}$$

$$g_{22}(T) = e^{-T}, \quad g_{32}(T) = \frac{3}{2} e^{-T} - \frac{3}{2} e^{-3T}, \quad g_{33}(T) = e^{-3T}.$$
probability that during the coalescence of \(m+n+l\) lineages from A, B and C, at least one interspecific coalescence occurs during this process, and that the most recent interspecific coalescence joins a lineage from species A and a lineage from species B.

\[
F^A_B(m,n,l) = \frac{ab}{\binom{a+b+c}{2}} + F^A_B(a-1, b, c) \frac{\binom{a}{2}}{\binom{a+b+c}{2}} \\
+ F^A_B(a, b-1, c) \frac{\binom{b}{2}}{\binom{a+b+c}{2}} \\
+ F^A_B(a, b, c-1) \frac{\binom{c}{2}}{\binom{a+b+c}{2}}. \tag{21}
\]

\(F^A_B(a, b, c) = 0\) if \(a+b+c \leq k\).
Probability of Monophyletic Concordance

\[ P_{M3}(r, s, q, T_3, T_2) \]

\[ = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \sum_{l=1}^{q} \left[ g_{rm}(T_3) g_{sn}(T_3) g_{m+n,k}(T_2) \right. \]

\[ \times g_{ql}(T_3+T_2) \left[ \delta_{k,1} \left[ 1 - F_2^{A,B}(m, n, 0) \right] \frac{2}{l(l+1)} \right. \]

\[ + \left( 1 - \delta_{k,1} \right) \left[ 1 - F_2^{A,B}(m, n, 0) \right] \]

\[ \times \sum_{x=1}^{k-1} \left[ W_{(m,n),(x,k-x)}(T_2) \left[ 1 - F_3^{A,B}(x, k-x, l) \right] \right. \]

\[ - F_3^{A,C}(x, k-x, l) - F_3^{B,C}(x, k-x, l) \left] \frac{1}{3} \right] \left] \right] \right]. \] (11)

\[ P_{monophyly}(r, s, q, T_3, T_2) \]

\[ = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \sum_{l=1}^{q} \left[ g_{rm}(T_3) g_{sn}(T_3) g_{m+n,k}(T_2) \right. \]

\[ \times g_{ql}(T_3+T_2) \left[ \delta_{k,1} \left[ 1 - F_2^{A,B}(m, n, 0) \right] \frac{2}{l(l+1)} \right. \]

\[ + \left( 1 - \delta_{k,1} \right) \left[ 1 - F_2^{A,B}(m, n, 0) \right] \]

\[ \times \sum_{x=1}^{k-1} \left[ W_{(m,n),(x,k-x)}(T_2) \left[ 1 - F_3^{A,B}(x, k-x, l) \right] \right. \]

\[ - F_3^{A,C}(x, k-x, l) - F_3^{B,C}(x, k-x, l) \left] \frac{1}{3} \right] \left] \right] \right]. \] (12)