Clades and clans probability in Yule trees

Sha (Joe) Zhu, Dr James Degnan, and Prof Mike Steel
University of Canterbury

November 9, 2011
The Yule-Harding-Kingman (YHK) process

Clades probability

Clans probability

To sum up
**Definition**
A clade of a rooted tree is a subset of $X$ that corresponds to the set of leaves that are descended from any internal vertex.
**Definition**

A clade of a rooted tree is a subset of $X$ that corresponds to the set of leaves that are descended from any internal vertex.

**Example:**

Set $\{a, b, c\}$ is a clade.

Set $\{a, b, c, d\}$ is not a clade.
The Yule-Harding-Kingman (YHK) process

Clades probability

Clans probability

To sum up
Yule Tree

Definition
Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.
Yule Tree

Definition
Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.

Example:
Yule Tree

Definition

Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.

Example:
Yule Tree

Definition
Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.

Example:
Yule Tree

Definition
Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.

Example:
Definition
Yule tree construction starts from a $X_2$ tree, choose a pendant edge randomly, and add a node onto this pendant edge. Repeat this process on a $X_k$ tree, choose a pendent edge at random, and add the $(k + 1)$th node onto it to form a $X_{k+1}$ tree.

Example:
Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one.
Coalescent Tree

Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one.

Example:

\[ a \quad b \quad c \quad d \]
Coalescent Tree

**Definition**
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one.

Example:
Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one.

Example:

```
   a
  /|
 / |\n/  v  \
/    \
/      \
 b  c  d
```

Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one.

Example:
Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one. Example:

```
    a   b
    c   d
```
Definition
The coalescent process starts from the bottom of a tree. Randomly choose two lineage to coalesce into one, and repeat this process on the remaining lineages until all lineages coalesce into one. Example:

```
     /
    / 
   /   
  a    b
     
     /
    / 
   /   
  c    d
```
Coalescent Tree
The Yule-Harding-Kingman (YHK) process

Clades probability

Clans probability

To sum up
**Definition**
A clade of a rooted tree is a subset of $X$ that corresponds to the set of leaves that are descended from any internal vertex.

**Example:**

Set $\{a, b, c\}$ is a clade.

Set $\{a, b, c, d\}$ is not a clade.
Yule process properties (I) (Aldous, 1995)

\textbf{(EP)} If $T'$ is obtained from $T$ by permuting its leaves, then

$$P(T = T') = P(T = T).$$
(EP) If \( T' \) is obtained from \( T \) by permuting its leaves, then

\[
P(T = T') = P(T = T).
\]

Example:
Let \( p_n(A) \) be the probability that \( A \) is a clade of \( T \), and \( a = |A| \), then

\[
p_n(A) = p_n(a).
\]

From Rosenberg (2003) we have:

\[
p_n(a) = \begin{cases} 
  \frac{2n}{a(a + 1)} \left( \frac{n}{a} \right)^{-1}, & \text{if } 1 \leq a \leq n - 1; \\
  0, & \text{otherwise}.
\end{cases}
\]
Probability of $A$ is a clade of $T$ (Rosenberg, 2003; Brown, 1994)
Clades probabilities

- Probability of $A$ is a clade of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of $A$ and $X - A$ are clades of $T$ (Rosenberg, 2003; Brown, 1994)
Clades probabilities

- Probability of $A$ is a clade of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of $A$ and $X - A$ are clades of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of three clades, and four clades (Brown, 1994)
Clades probabilities

- Probability of $A$ is a clade of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of $A$ and $X - A$ are clades of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of three clades, and four clades (Brown, 1994)
Clades probabilities

- Probability of $A$ is a clade of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of $A$ and $X - A$ are clades of $T$ (Rosenberg, 2003; Brown, 1994)
- Probability of three clades, and four clades (Brown, 1994)
- $k$ clades
(GE) For any proper (and non-empty) subset $A$ of $X$, and any rooted binary phylogenetic tree $T$ with leaf set $X - A$:

$$\mathbb{P}(T_{X|X-A} = T | A \in c(T)) = \mathbb{P}(T_{(X-A)} = T).$$

(SC) $$\mathbb{P}(T_{X|A} = T) = \mathbb{P}(T_A = T).$$
Clades probability of A and B, $p_n(a, b)$

<table>
<thead>
<tr>
<th>Case</th>
<th>Diagram</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A = B$</td>
<td></td>
<td>$A \subsetneq B$</td>
<td>$B \subsetneq A$</td>
</tr>
<tr>
<td></td>
<td>$A \cap B = \emptyset$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$A \cup B = X$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$A \cup B \subsetneq X$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>otherwise</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More than $k$ clades,

Consider set $A_1, A_2, \ldots, A_k$ are clades, and $\bigcup_{i=1}^{k} A_i = X$

\[ p(a_1, \ldots, a_k; T_k) = \frac{2^{k-1} \prod_{i=1}^{k} a_i!}{n!} \prod_{v \in \mathcal{I}(T_k)} \left( \frac{1}{\sum_{i=1}^{k} a_i l_v(A_i) - 1} \right), \]

\[ p(a_1, \ldots, a_k) = \sum_{T_k} p(a_1, \ldots, a_k; T_k). \]
More than $k$ clades,

Consider set $A_1, A_2, \ldots, A_k$ are clades, and $\bigcup_{i=1}^{k} A_i = X$

$$p(a_1, \ldots, a_k; T_k) = \frac{2^{k-1} \prod_{i=1}^{k} a_i!}{n!} \prod_{v \in \mathcal{I}(T_k)} \left( \frac{1}{\sum_{i=1}^{k} a_i l_v(A_i) - 1} \right),$$

$$p(a_1, \ldots, a_k) = \sum_{T_k} p(a_1, \ldots, a_k; T_k).$$
More than $k$ clades,

Consider set $A_1, A_2, \ldots, A_k$ are clades, and $\bigcup_{i=1}^{k} A_i = X$

$$p(a_1, \ldots, a_k; T_k) = \frac{2^{k-1} \prod_{i=1}^{k} a_i!}{n!} \prod_{v \in \mathcal{I}(T_k)} \left( \frac{1}{\sum_{i=1}^{k} a_i l_v(A_i) - 1} \right),$$

$$p(a_1, \ldots, a_k) = \sum_{T_k} p(a_1, \ldots, a_k; T_k).$$
Corollary

For any two strict subsets $A, B$ of $X$, the correlation $\rho_n(A, B)$ is:

- **strictly negative**, if $A, B$ are not compatible, and undefined if $|A| = 1$ or $|B| = 1$.
- **strictly positive**, otherwise.
Corollary

For any two strict subsets $A, B$ of $X$, the correlation $\rho_n(A, B)$ is:

- strictly negative, if $A, B$ are not compatible, and undefined if $|A| = 1$ or $|B| = 1$.
- strictly positive, otherwise.

$$\rho_n(A, B) = \frac{p_n(A, B) - p_n(A)p_n(B)}{\sqrt{p_n(A)(1 - p_n(A))p_n(B)(1 - p_n(B))}}.$$
Corollary

For any two strict subsets $A, B$ of $X$, the correlation $\rho_n(A, B)$ is:

- strictly negative, if $A, B$ are not compatible, and undefined if $|A| = 1$ or $|B| = 1$.
- strictly positive, otherwise.
The Yule-Harding-Kingman (YHK) process

Clades probability

Clans probability

To sum up
Definition

We say that a subset $A$ of $X$ is a clan of an unrooted phylogenetic $X$-tree $T^{-\rho}$ if $A \mid (X - A)$ is a split of $T^{-\rho}$.
Definition
We say that a subset $A$ of $X$ is a clan of an unrooted phylogenetic $X-$tree $T^{-\rho}$ if $A \mid (X - A)$ is a split of $T^{-\rho}$.

Example:
Let $X = \{a, b, c, d, e, f\}$, $A = \{a, b, c, d\}$, and $X - A = \{e, f\}$. 

Set $A$ is not a clade.

Set $A$ is a clan.
Lemma

A set $A$ is a clan of a unrooted tree $T^-\rho$ if and only if either $A$ is a clade of $T$ or $X - A$ is a clade of $T$.

$$q_n(A) = p_n(A) + p_n(X - A) - p_n(A, X - A)$$

If $A \cap B = \emptyset$ and $A \cup B = X$,

$$q_n(A, B) = q_n(A)$$
If $A \cap B = \emptyset$ and $A \cup B \subseteq X$, 

\[ q_n(A, B) \]

Recall:

Case 3  Case 4  Case 5
The Yule-Harding-Kingman (YHK) process

Clades probability

Clans probability

To sum up
In this project,

- we derived exact probabilities for two compatible clades with any sizes;
- we have shown there is positive correlation between any two compatible clades;
- we have extended clades probabilities to clans probabilities.


Acknowledgments

- My supervisors: Mike and James.
- Marsden fund.
- Organizers: Barbara and Jeremy.