Population Genetics, Recombination Histories & Global Pedigrees

Population Genetics and Genealogies

Finding Minimal Recombination Histories

Global Pedigrees
Wright-Fisher Model of Population Reproduction

**Haploid Model**

i. Individuals are made by sampling with replacement in the previous generation.

ii. The probability that 2 alleles have same ancestor in previous generation is $1/2N$

**Assumptions**

1. Constant population size
2. No geography
3. No Selection
4. No recombination

**Diploid Model**

Individuals are made by sampling a chromosome from the female and one from the male previous generation with replacement.
10 Alleles’ Ancestry for 15 generations

Generate new generations according to the haploid model

Sort generations forward in time, such that individual 1’s children in the next generation comes first, then the children of 2...

Pick individuals in the present and label edges to their ancestors back in time
Waiting for most recent common ancestor - MRCA

Distribution of time until 2 alleles had a common ancestor, $X_2$?:

$$P(X_2 > 1) = \frac{2N-1}{2N} = 1 - \frac{1}{2N}$$

$$P(X_2 > j) = \left(1 - \frac{1}{2N}\right)^j$$

$$P(X_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \cdot \frac{1}{2N}$$

Mean, $E(X_2) = 2N$.

Ex.: $2N = 20,000$, Generation time 30 years, $E(X_2) = 600,000$ years.
**P(k):=P{k alleles had k distinct parents}**

\[ P(k) = \frac{2N^k}{(2N)^k} \approx (k^2 < 2N) \ 1 - \frac{k}{2N} \approx e^{-\frac{k^2}{2N}} \]

**Ancestor choices:**

<table>
<thead>
<tr>
<th>Ancestor choices:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>k -&gt; any</td>
<td>k -&gt; k</td>
<td>k -&gt; k-1</td>
<td>k -&gt; j</td>
<td></td>
</tr>
<tr>
<td>(2N)^k</td>
<td>2N * (2N-1) * ... * (2N-(k-1)) (\Rightarrow) ((2N)_{[k]})</td>
<td>(\begin{pmatrix} k \ 2 \end{pmatrix}) ((2N)_{[k-1]})</td>
<td>(S_{k,j}) ((2N)_{[j]})</td>
<td></td>
</tr>
</tbody>
</table>

\(S_{k,j}\) - the number of ways to group \(k\) labelled objects into \(j\) groups. (Stirling Numbers of second kind.)
Expected Total height of tree: \( H_k = 2(1-1/k) + \ldots + 1/(k-1) \) ca= 2*ln(k-1)

i. Infinitely many alleles finds 1 allele in finite time.
ii. It takes less than twice as long for \( k \) alleles to find 1 ancestor as it does for 2 alleles.

Expected Total branch length in tree, \( L_k: 2*(1 + 1/2 + 1/3 + \ldots + 1/(k-1)) \) ca= 2*ln(k-1)
6 Realisations with 25 leaves

Observations:
Variation great close to root.
Trees are unbalanced.
The probability that the ancestor of the sample of size $n$ is in a sub-sample of size $k$ is
\[
\frac{(n + 1)(k - 1)}{(n - 1)(k + 1)}
\]
Letting $n$ go to infinity gives $(k-1)/(k+1)$, i.e. even for quite small samples it is quite large.
Three Models of Alleles and Mutations.

**Infinite Allele**
- i. Only identity, non-identity is determinable
- ii. A mutation creates a new type.

**Infinite Site**
- i. Allele is represented by a line.
- ii. A mutation always hits a new position.

**Finite Site**
- i. Allele is represented by a sequence.
- ii. A mutation changes nucleotide at chosen position.
Infinite Site Model

Final Aligned Data Set:
Recombination-Coalescence Illustration  

*Intensities*

<table>
<thead>
<tr>
<th>Coales.</th>
<th>Recomb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\rho$</td>
</tr>
<tr>
<td>1</td>
<td>$(1+b)\rho$</td>
</tr>
<tr>
<td>3</td>
<td>$(2+b)\rho$</td>
</tr>
<tr>
<td>6</td>
<td>$2\rho$</td>
</tr>
<tr>
<td>3</td>
<td>$2\rho$</td>
</tr>
<tr>
<td>1</td>
<td>$2\rho$</td>
</tr>
</tbody>
</table>

Backward in time
Local Inference of Recombinations

**Recoding**
- At most 1 mutation per column
- 0 ancestral state, 1 derived state

**Incompatibility:**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Four combinations**

- 00
- 10
- 01
- 11

**Myers-Griffiths (2002):** Number of Recombinations in a sample, \( N_R \), number of types, \( N_T \), number of mutations, \( N_M \) obeys:

\[
N_R \geq N_T - N_M - 1
\]
"Observing" Recombinations: **Hudson & Kaplan’s $R_M$**

If you equate $R_M$ with expected number of recombinations, this could be used as an estimator. Unfortunately, $R_M$ is a gross underestimate of the real number of recombinations.
Minimal Number of Recombinations

The Kreitman data (1983): 11 sequences, 3200bp, 43(28) recoded, 9 different

Last Local Tree Algorithm:

Data

1 2 i-1 i L

1
2
n

Trees

How many local trees?

- Unrooted
- Coalescent

How many neighbors?

\[ \frac{(2n - 2)!}{2^{n-1}(n-1)!} \]

\[ 3n^2 - 13n + 14 \]

\[ \frac{n!(n-1)!}{2^{n-1}} \]

\[ \sim n^3 \]
Metrics on Trees based on subtree transfers.

Trees including branch lengths

Unrooted tree topologies

Rooted tree topologies

Tree topologies with age ordered internal nodes

Pretending the easy problem (unrooted) is the real problem (age ordered), causes violation of the triangle inequality:
Tree Combinatorics and Neighborhoods

Observe that the size of the unit-neighbourhood of a tree does not grow nearly as fast as the number of trees

\[ \delta(T) := \text{number of trees one SPR operation away from a given tree } T. \]

<table>
<thead>
<tr>
<th></th>
<th>Unrooted</th>
<th>Rooted</th>
<th>Dendrograms</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td># of trees</td>
<td>( \delta )</td>
<td># of trees</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>12</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>30</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>56</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
<td>90</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
<td>132</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>182</td>
<td>34,459,425</td>
</tr>
</tbody>
</table>

\( (2n - 3)! = \frac{(2n-2)!}{2^{n-1}(n-1)!} \)

\( 2(n-3)(2n-7) \)

\( 3n^2 - 13n + 14 \)

\( 4(n-2)^2 - 2 \sum_{m=1}^{n-2} \log_2(m+1) \)

\( \frac{n! (n-1)!}{2^{n-1}} \)

\( \frac{1}{3} (2n^3 - 3n^2 - 20n + 39) \)

\( \frac{1}{6} \left\{ 4n^3 - 9n^2 - 13n + 42 - 3(2n+3) \left[ \frac{n-1}{2} \right] + 9 \left( \left\lfloor \frac{n-1}{2} \right\rfloor \right)^2 \right\} \)

Allen & Steel (2001)

Song (2003+)
**minARGs: Recombination Events & Local Trees**

**Minimal ARG**

**True ARG**

**Reconstructed ARG**

**Mutation information on both sides**

**Mutation information on only one side**

\[
E_{70}(R) = 132 \\
E_{70}(R_{visible}) = 46
\]

\[\rho \sum_{j=1}^{n-1} j^{-1} \]

\[ \text{n=8, } \Theta=40 \]

\[ \text{n=8, } \Theta=15 \]

\[ \text{n=7, } \rho=10, \Theta=75 \]

**Hudson-Kaplan**

**Myers-Griiths**

**Song-Hein**
Counting + Branch and Bound Algorithm

<table>
<thead>
<tr>
<th>$n$</th>
<th>$2$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>30</td>
<td>573</td>
<td>16 875</td>
<td>689 175</td>
</tr>
<tr>
<td>3</td>
<td>108</td>
<td>6 286</td>
<td>743 387</td>
<td>149 861 079</td>
</tr>
<tr>
<td>4</td>
<td>330</td>
<td>62 589</td>
<td>32 482 009</td>
<td>35 523 729 489</td>
</tr>
<tr>
<td>5</td>
<td>866</td>
<td>445 137</td>
<td>893 479 326</td>
<td>4 938 627 635 669</td>
</tr>
<tr>
<td>6</td>
<td>2 143</td>
<td>3 302 506</td>
<td>29 521 615 942</td>
<td>962 962 451 049 968</td>
</tr>
<tr>
<td>7</td>
<td>4 611</td>
<td>17 409 443</td>
<td>568 860 072 916</td>
<td>91 812 561 254 804 105</td>
</tr>
<tr>
<td>8</td>
<td>9 728</td>
<td>98 432 218</td>
<td>13 273 296 248 617</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>18 378</td>
<td>420 106 717</td>
<td>195 515 335 378 914</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>34 552</td>
<td>1 917 604 869</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>59 577</td>
<td>6 985 275 356</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BB & Heuristic minimal ancestral recombination graphs

Beagle
Try each in turn until shortest route is determined

Margarita
Just follow road seeming to lead in the right direction

Kwarg
Choice based on location of next crossroads
Time versus Spatial 1: Coalescent-Recombination

i. The process is non-Markovian

ii. The trees cannot be reduced to Topologies
**Time versus Spatial 2: Pedigrees**

Elston-Stewart (1971) - Temporal Peeling Algorithm:

Condition on parental states
Recombination and mutation are Markovian

Lander-Green (1987) - Genotype Scanning Algorithm:

Condition on paternal/maternal inheritance
Recombination and mutation are Markovian
**Time versus Spatial 3: Phylogenetic Alignment**

- **Optimisation Algorithms**
  - Indels of length 1 (David Sankoff, 1973) **Spatial**
  - Indels of length k (Bjarne Knudsen, 2003) **Temporal**
- **Statistical Alignment**

**Spatial:**

**Temporal:**
The Griffiths-Ethier-Tavare Recursions

<table>
<thead>
<tr>
<th>No recombination:</th>
<th>Infinite Site Assumption</th>
<th>History Graph:</th>
<th>Recursions Exists</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ancestral State Known</td>
</tr>
</tbody>
</table>

Possible Histories without Recombination for simple data example

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

- recombination 27 ACs

+ recombination \(9 \times 10^8\) ACs
Ancestral configurations to 2 sequences with 2 segregating sites
<table>
<thead>
<tr>
<th></th>
<th>0-ARG</th>
<th>1-ARG</th>
<th>2-ARG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>States</strong></td>
<td>5</td>
<td>+15</td>
<td>+10</td>
</tr>
<tr>
<td>(\Theta, \rho = 2, 0)</td>
<td>0.148</td>
<td>0.148</td>
<td>0.148</td>
</tr>
<tr>
<td>(\Theta, \rho = 1, 1)</td>
<td>0.037</td>
<td>0.082</td>
<td>0.090</td>
</tr>
<tr>
<td>(\Theta, \rho = 2, 2)</td>
<td>0.032</td>
<td>0.074</td>
<td>0.085</td>
</tr>
</tbody>
</table>

---

**Diagram**

2nd

1st

---
Likelihood Calculations on the ε-ARG

Example:

010
010
101
101
110

0-ARG

1-ARG

2-ARG
• For sites=7, sequences=6, then $10^{-7}$ th of states visited

• Cut-off $\varepsilon$-ARG, $\varepsilon=2$
Let $T$ be the time, when somebody was everybody’s ancestor.

Chang’s result: $\lim T^*/\log_2(N) = 1$  prob. 1

Unify the two processes:

I. Sample more individuals   II. Let each have 2 parents with probability $p$.

Result: A discontinuity at 1.

For $p<1$ change $\log_2 \rightarrow \log_p$

Comment: Genetic Ancestors is a vanishing set within Genealogical Ancestors.
Reconstructing global pedigrees: Superpedigrees

Steel and Hein, 2006

The gender-labeled pedigrees for all pairs defines global pedigree

Gender-unlabeled pedigrees don’t!!

Benevolent Mutation and Recombination Process

Genomes with $\rho$ and $\mu/\rho \rightarrow \infty$

$\rho$ – recombination rate, $\mu$ – mutation rate

- All embedded phylogenies are observable
- Do they determine the pedigree?

Counter example:

Embedded phylogenies:
Infinite Sequences: From ARG to Pedigree

What can you observe from data (infinite sequences)?

A. The ARG?

B. Sequence of neighbor pairs of local trees with recombination points
   Going to neighbor triples, quadruples,.., be more restrictive than pairs?

C. Sequence of local trees?

D. Set of local trees?

E. Set/Sequences of local unrooted tree topologies?

F. Set/Sequences of local bipartitions? (neighbor pairs…

Given A/B/C/D/E above how much does that constrain set of pedigrees

How many pedigrees are compatible with A/B/C/D/E varying over data?
Infinite Sequences: From ARG to Pedigree

What can you observe from data (infinite sequences)?

A. The ARG?

B. Sequence of pairs of local trees with recombination points
   Going to triples, quadruples,.., be more restrictive than pairs?

C. Sequence of local trees?

D. Set of local trees?

E. Set/Sequences of local unrooted tree topologies?

F. Set/Sequences of bipartitions