Ancestral Inference in Molecular Biology

NTHU #2

Equations used in the overheads are given the numbers they have in the accompanying text.
The story so far . . .

We were looking at the Wright-Fisher model from a genealogical perspective.

Next we analyze what happens to the tree linking the $n$ chromosomal segments in a large population of size $N$. 
A sample of size $n = 2$

Since individuals choose their parents at random,

$$\mathbb{P}(\text{2 individuals have 2 distinct parents}) = \left(1 - \frac{1}{N}\right).$$

Since those parents are themselves a random sample from their generation, we may iterate this argument:

$$\mathbb{P}(\text{First common ancestor } > r \text{ generations ago})$$

$$= \left(1 - \frac{1}{N}\right)^r.$$
By rescaling time in units of $N$ generations, so that $r = Nt$, and letting $N \to \infty$ we see that this probability is

$$\left(1 - \frac{1}{N}\right)^{Nt} \to e^{-t}.$$ 

Thus the time until the first common ancestor of the sample of two genes has approximately an exponential distribution with mean 1.
More generally, the number of distinct parents of a sample of size \( k \) individuals can be thought of as the number of occupied cells after \( k \) balls have been dropped (uniformly and independently) into \( N \) cells.

\[
g_{k,j} \equiv \mathbb{P}(k \text{ individuals have } j \text{ distinct parents})
\]

\[
= N(N-1) \cdots (N-j+1) S_k^{(j)} N^{-k} \quad j = 1, 2, \ldots, k \quad (3.2)
\]

where \( S_k^{(j)} \) is a Stirling number of the second kind; \( S_k^{(j)} \) is the number of ways of partitioning a set of \( k \) elements into \( j \) nonempty subsets.
Now let

\[ A^N_n(l) \equiv \text{number of distinct ancestors in generation } l \text{ of a sample of size } n \text{ at time 0.} \]

\[ A^N_n(\cdot) \text{ is Markovian with transition probabilities given by} \]

\[ \mathbb{P}(A^N_n(l + 1) = j | A^N_n(l) = k) = g_{kj}, 1 \leq j, k \leq n \]

Fix the sample size \( n \), and let \( N \to \infty \).
Then for $k \leq n$,

$$g_{k,k-1} = \mathcal{S}^{(k-1)}_k \frac{N(N-1) \cdots (N-k+2)}{N^k}$$

$$= \binom{k}{2} \frac{1}{N} + O(N^{-2}),$$

since $\mathcal{S}^{(k-1)}_k = \binom{k}{2}$. For $j < k - 1$, we have

$$g_{k,j} = \mathcal{S}^{(j)}_k \frac{N(N-1) \cdots (N-j+1)}{N^k} = O(N^{-2})$$
and

\[ g_{k,k} = N^{-k} N(N-1) \cdots (N-k+1) \]

\[ = 1 - \binom{k}{2} \frac{1}{N} + O(N^{-2}). \]

The transition matrix \( P_N \) of the chain satisfies

\[ P_N = I + \frac{1}{N} Q + O(N^{-2}), \]

where the non-zero elements of \( Q \) are given by

\[ q_{k,k} = -\binom{k}{2}, \quad q_{k,k-1} = \binom{k}{2}. \]
Hence in $g = Nt$ generations,

$$P^g_N = \left(I + \frac{1}{N}Q + O(N^{-2})\right)^{Nt} \to \exp(Qt)$$

This suggests that the ancestral process in a large population can be approximated by a very simple *pure death* process.
The approximating ancestral process $A_n(\cdot)$ is a pure death process that starts from $A_n(0) = n$, and decreases by jumps of size one only.

The waiting time $T_k$ in state $k$ is exponential with parameter $\left(\frac{k}{2}\right)$, the $T_k$ being independent for different $k$.

See Page 21 for details.
Chapter 4. Time to the most recent common ancestor

The TMRCA of a sample of \( n \) is

\[
W_n = T_n + T_{n-1} + \cdots + T_2
\]  \hspace{1cm} (4.3)

where \( T_k \) are independent exponential random variables with parameter \((k^2)\). It follows that

\[
\mathbb{E}W_n = \sum_{k=2}^{n} \mathbb{E}T_k = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2 \left(1 - \frac{1}{n}\right).
\]

Therefore

\[
1 = \mathbb{E}W_2 \leq \mathbb{E}W_n \leq \mathbb{E}W_N < 2.
\]
Note that $T_2$ makes a substantial contribution to the sum (4.3) for $W_n$. For example, on average for over half the time since its MRCA, the sample will have exactly two ancestors.

Also

$$\mathbb{E}(W_N - W_n) = 2 \left( \frac{1}{n} - \frac{1}{N} \right) < \frac{2}{n}$$

so the mean difference between the time for a sample to reach its MRCA and the time for the whole population to reach its MRCA is small.
Using the independence of the $T_k$, 

$$\text{Var}(W_n) = \sum_{k=2}^{n} \text{Var}(T_k) = \sum_{k=2}^{n} \binom{k}{2}^{-2}$$

It follows that

$$1 = \text{Var}(W_2) \leq \text{Var}(W_n) \leq \lim_{n \to \infty} \text{Var}(W_n) \approx 1.16.$$ 

The distribution of $W_n$ is given in (4.4) and (4.5), and Exercise 4.5 shows that

$$e^{-t} \leq \mathbb{P}(W_n > t) \leq 3e^{-t}$$

for all $n$ and $t$. 
The coalescent

Which individuals are descended from which ancestors?

• For some fixed time $t$, label the individuals in the sample from the set $\{1, \ldots, n\}$

• Define a (random) equivalence relation $\sim$ on $[n] \equiv \{1, 2, \ldots, n\}$ by

$$i \sim j \quad \text{if and only if individuals } i \text{ and } j \text{ share a common ancestor at time } t.$$
We can also represent this information as a genealogical tree.

- The lengths of the various branches are proportional to the times between the various events.
- It is convenient to think of the $n$-coalescent as a random, rooted, binary tree, with lengths attached to the edges.
- The structure of the genealogical process translates easily to the random tree.
Simulated trees
Robustness in the coalescent

Three examples:

• WF model.

\[ \mathbb{E} \nu_1 = 1, \text{Var}(\nu_1) = 1 - 1/N \]

• \( \nu \equiv (1, 1, \ldots, 1) \).

\[ \mathbb{E} \nu_1 = 1, \text{Var}(\nu_1) = 0 \]

• \( \nu = (0, \ldots, N, 0, \ldots, 0) \) for some \( i = 1, \ldots, N \)

\[ \mathbb{E}(\nu_1) = 1, \text{Var}(\nu_1) = N - 1. \]
Cannings’ model

The following assumptions about the random variables $(\nu_1, \ldots, \nu_N)$ are natural:

(i) Constant population size requires that $\nu_1 + \cdots + \nu_N = N$.

(ii) The collection of random variables $\nu_1, \ldots, \nu_N$ is exchangeable. That is, the distribution of offspring numbers does not depend on the way in which the individuals are labeled.

(iii) The distribution of $(\nu_1, \ldots, \nu_N)$ is the same in each generation. This is time stationarity.
(iv) The joint distribution of \((\nu_1, \ldots, \nu_N)\) is independent of family sizes in other generations.

This is *neutrality*: offspring numbers for particular individuals do not depend on ancestral offspring numbers.
Theorem 4.3 [Kingman, 1982] Assume that

(i) \( \sigma_N^2 \equiv \text{Var}(\nu_1) \to \sigma^2 \in (0, \infty) \) as \( N \to \infty \);

(ii) \( \sup_N \mathbb{E}(\nu_1^k) < \infty \quad k = 3, 4, \ldots \).

Then

\[ A_N^n \left( \lfloor \sigma^{-2} N \cdot \rfloor \right) \Rightarrow A_n(\cdot) \text{ as } N \to \infty. \]

Note: this is NOT the only interesting limiting regime. Multiple coalescent events also occur in some limits.
The parameter $\sigma^2$ matters!

If time is scaled in units of $N$ generations, then

$$A^N_n ([N \cdot ]) \Rightarrow A_n (\sigma^2 \cdot )$$

On this time scale, the waiting time $T_j$ while the sample has $j$ distinct ancestors has an exponential distribution with mean

$$\mathbb{E}T_j = \frac{2}{\sigma^2 j(j - 1)}$$

in coalescent units, or

$$\frac{2N}{\sigma^2 j(j - 1)}$$

in units of generations.
Variable population size

For the WF model, measure time in units of the present population size $N = N(0)$. Suppose population time $t$ earlier had relative size $f(t)$.

Set $\lambda(t) = 1/f(t)$, $\Lambda(t) = \int_0^t \lambda(u)du$

Variable size ancestral process is a deterministic time change of regular one:

$$A^v_n(t) = A_n(\Lambda(t)), \ t > 0.$$ 

In an exponentially growing population most of the coalescence events occur near the root of the tree – it is harder to find common ancestors when the population is large. The resulting genealogy is ‘star-like’.
Mutations in the genealogical tree

Genetic variation observed in a sample of genes is a consequence of mutations arising on their coalescent.

- Genetic types are labelled by elements of the type space $E$.
- As mutations occur, labels of individuals move around according to a mutation process on $E$.
- An offspring of an individual of type $x \in E$ has a type in the set $B \subseteq E$ with probability $u(x, B)$. 
We suppose that conditional its parent’s type, the type of a particular offspring is independent of the types of other offspring, and of the demography of the population.

In particular, the offspring of different individuals mutate independently.
(i) Infinitely-many-alleles model (P54)

A mutation occurs in a given region with chance $u$. Mutants are labeled by independent $U(0,1)$ r.v.s. The example generates 5 types, $U_2, U_2, U_5, U_3, U_3$. 
(ii) Infinitely-many-sites model (P55)

Same structure as before, but sequences are labeled by sequence of mutations from root.

It results in a sample of five sequences: $(U_1, U_2)$, $(U_1, U_2)$, $(U_1, U_2, U_4, U_5)$, $(U_0, U_3)$, $(U_0, U_3)$

Think of the list as giving locations of new mutations that have arisen in those sequences. Each mutation introduces a new segregating site into the sample.
Can represent the sequences as a string of 0 and 1 (length 6 in this case).

1 denotes mutant, 0 ancestral type.

Get aligned sequences:

\[(U_1, U_2, U_4, U_5) = 011011\]
\[(U_1, U_2) = 011000\]
\[(U_0, U_3) = 100100\]
Finitely-many-sites: DNA sequences

A model allowing for base substitutions only.

- Probability $u$ that a mutation occurs somewhere in the sequence
- When mutation occurs, site $l$ is chosen to mutate with probability $h_l$
- Base at site $l$ changes according to a $4 \times 4$ matrix $P^{(l)}$

The $h_l$ may be markedly different (mutational hot spots)

The site-specific substitution probabilities $P^{(l)}$ may differ from site to site
Scaling the mutation rate

As before we maintain balance between drift and mutation by supposing that $u = O(1/N)$:

$$\theta = \lim_{N \to \infty} 2Nu \in (0, \infty)$$

Then mutations arise at points of Poisson processes of rate $\theta/2$, independently in different branches of the tree.

Other orders might be of interest – the approximating process need not then be a coalescent. Branching processes occur too.
Simulating a stationary sample

- Choose the whole genealogical tree according to the distribution of the coalescent.

- Choose the type of the MRCA of the sample according to the stationary distribution of the mutation process.

- Run the mutation process forward from the type of the MRCA, along the genealogical tree in such a way that on different branches of the tree these processes are independent.
Variable size

In a population of varying size, a ‘stationary sample’ means one that has been generated by simulating back to the MRCA, and then superimposing mutations down from that ancestor to the sample.

Can use other assignments of type at the MRCA
Branching Markov process
Relatives of the coalescent

The coalescent has been developed for many scenarios, such as:

- Population substructure, migration
- Selfing and other non-random mating systems
- Diploidy
- Selection
- Recombination

Reference: M. Nordborg (2001)
Why is the coalescent so useful?

The coalescent has revolutionized the way we interpret and think about polymorphism data.

- **Top down**

  A version of gene dropping

  Can simulate behavior of complex stochastic models for which ‘no theory’ can be found

  Do not need to simulate whole population at stationarity, then sample

  Very useful for inference, null distribution of test statistics
• **Bottom up**

Which trees are consistent with observed data?

Bayesian inference and likelihood methods
Chapter 5. Mitochondrial DNA
### Data [Table 5.1, p68]

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>allele</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>b</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>c</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>d</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>e</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>f</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>g</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>h</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>i</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>j</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>k</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>l</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>m</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>n</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

| allele freq. | 2 | 2 | 1 | 3 | 19 | 1 | 1 | 1 | 4 | 8 | 5 | 4 | 3 | 1 |
