The Combinatorial Interpretation of Formulas in Coalescent Theory

John L. Spouge
National Center for Biotechnology Information
NLM, NIH, DHHS
Motivation

- Evolution is fundamental to understanding biological networks.
- The infinite alleles model
  - was arguably the most important model of evolution until about 1990.
  - yields many simple solutions, mostly proved by induction.
  - still provides useful approximations (e.g., in LDHat, a program for estimating recombination rates along human chromosomes)
- If a difficult problem has a solution with a simple form, the solution had to be simple, and something is missing.
- In 1990, a simple solution was discovered, but not pursued.
- The solution
  - might indicate techniques useful in newer evolutionary realistic models.
  - provides people of a combinatorial bent with an instant understanding of a significant part of almost thirty years of evolutionary literature.

M Kimura & JF Crow (1964) Genetics 49:725
P Joyce & S Tavare (1990) Stoch Proc Appl 36:245
Overview

- Population Genetics Models – Forward Time
  - Wright-Fisher Model
  - Moran Model
  - Yule Model

- The Death Process for the Moran Model – Backward Time
  - The waiting times are independent of the combinatorics.
  - Invariance theorems: the ancestral tree of a population sample converges to a single limit for many genetics models (the coalescent process).

- The death process maps to random permutations, so counting permutations provides many results for the infinite alleles model.

- The mapping to random permutations also gives insight into deep results for the entire population.
Life

Reproduction
Mutation
Selection


Wright-Fisher Model Without Mutation

- The population is:
  - composed of haploid individuals; (No sex!)
  - of constant size $M$.
- Individuals reproduce simultaneously.
- Every offspring chooses a single parent uniformly at random.

RA Fisher (1930) The Genetical Theory of Natural Selection
S Wright (1931) Genetics 16:97
Wright-Fisher Model
Without Mutation
Some lineages die out by chance.

Some lineages expand by chance.

Some lineages do nothing much at all.
The population is:
- composed of haploid individuals;
- of constant size.

At each step
- Two individuals are chosen uniformly at random with replacement.
- One reproduces; the other (possibly the same individual) dies.
Moran Model Without Mutation
The population is:
- composed of haploid individuals;
- of constant size.

At each step
- Two individuals are chosen uniformly at random with replacement.
- One reproduces; the other (possibly the same individual) dies.
- With probability $u$, the new individual is a mutant type never seen before.
Infinite Alleles & Infinite Sites Models

Infinite alleles models

Every mutation is a type never seen before, but nothing is said about the structure of the mutation.

Infinite sites models

DNA

Every mutation is a type never seen before, occurring in a new place on the DNA.
Moran Model Without Mutation

1

2

3

4

5

6

7

8

Mt
Moran Model With Mutation

1

2

3

4

5

6

7

8

\[ M \]

\[ t \]

\[ 1 - u \]

\[ u \]
Moran Model With Mutation

Looking Backward – a Death Process

1. Let us temporarily ignore changes that do not affect the ancestry of the present population.

set of rooted trees ordered by height

$n_0 1 2 3 4 5 6 7 8$
Moran Model With Mutation
Looking Backward – a Death Process

"Ancestral Tree"
Moran Model With Mutation

Looking Backward – a Death Process

1. canonical representation of a permutation by its cycle structure

2. 8

3. 6

4. 5

5. 4

6. 7

7. 2

8. 1

M

n

0 1 2 3 4 5 6 7 8
Moran Model With Mutation
Looking Backward – a Death Process

M

n
0 1 2 3 4 5 6 7 8
Each permutation on $M$ objects corresponds to exactly one realization of the death process.
Moran Model With Mutation
Looking Backward – a Death Process

“Ancestral Permutation”
Moran Model With Mutation
Looking Backward

The variates $\Pi$ and $T$ are independent.

\[ P\{(\Pi, T) = (\pi, t)\} = \prod_{j=1}^{n} \left[ 1 - j \frac{u}{M} - j (j - 1) \frac{1 - u}{M^2} \right]^{t_{j-1} - t_{j+1} - 1} \]

\[ \left( \frac{u}{M} \right)^k \left( \frac{1 - u}{M^2} \right)^{n-k} \prod_{j=1}^{n} \left[ 1 - j \frac{u}{M} - j (j - 1) \frac{1 - u}{M^2} \right]^{t_{j-1} - t_{j+1} - 1} \]

The variates $\Pi$ and $T$ are independent.

$k$ cycles
$k = 4$
Moran Model With Mutation
Looking Backward

\[
\left( \frac{u}{M} \right)^k \left( \frac{1-u}{M^2} \right)^{n-k} \prod_{j=1}^{n} \left[ 1 - j \frac{u}{M} - j(j-1) \frac{1-u}{M^2} \right]^{t_j-t_{j+1}-1}
\]

The variates \( \Pi \) and \( T \) are independent.

\[
P\{(\Pi, T) = (\pi, t)\}
\]

\[
P\{\Pi = \pi\}
\]

\[
P\{\Pi = \pi \in S_n\} \propto \theta^k
\]

\( k \) cycles

\( k = 4 \)
Moran Model With Mutation
Looking Backward from a Population

\[ n \]

1 2 3 4 5 6 7 8

\[ M \]

0 1 2 3 4 5 6 7 8
Each permutation on \( n \) objects corresponds to exactly one realization of the death process.
Moran Model With Mutation
Looking Backward from a Sample

\[
\left( \frac{u}{M} \right)^{k} \left( \frac{1-u}{M^2} \right)^{n-k} \prod_{j=1}^{n} \left[ 1 - j \frac{u}{M} - j(j-1) \frac{1-u}{M^2} \right]^{t_j-t_{j+1}-1}
\]

\[
P\{ (\Pi, T) = (\pi, t) \}
\]

\[
P\{ T = t \}
\]

Fix

\[
M \to \infty
\]

\[
u = \frac{\theta}{M + \theta}
\]

\[
\theta = \frac{Mu}{1-u}
\]

By suitably scaling time, the waiting times between “deaths” are independent exponential variates. There are “invariance theorems” showing that when their times are suitably scaled, the Wright-Fisher Model, the Moran Model, and many other genetics models converge to limits, so-called “coalescent processes”.

The distribution of the death process for the population at any time has the same form as for a sample.
Sample Path of the Death Process & the Cycle Structure of Permutations

\[ \pi = (1 3)(2 7 4)(5)(6 8) \]

The Fundamental Theorem of Infinite Alleles Mutation Models

To calculate the probability of any particular ancestral tree, enumerate the corresponding ancestral permutations with \( \theta \) marking the cycles and then normalize to a probability. In a coalescent limit (infinite population with suitable scaling of time), the waiting times between events are independent exponential variates.

P Joyce & S Tavare (1990) Stoch Proc Appl 36:245
The Ancestral Permutation and the Sampling Permutation

In practice, individuals are sampled from the population before they are placed into the ancestral tree. Without loss of generality, let the sampling be sequential, and without replacement. The sampling order of the $n$ individuals requires a second, “sampling permutation”, to put the individuals into the unique order required for the correspondence between the ancestral graph and the ancestral permutation.

$$\pi = (1 \ 3)(2 \ 7 \ 4)(5)(6 \ 8)$$

$$\sigma = \begin{pmatrix} 6 & 3 & 4 & 7 & 8 & 5 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}$$

$$\mathbb{P}\{\Sigma = \sigma \in S_n\} = 1 / n!$$

The variates $\Pi$, $\Sigma$, and $T$ are independent.
A Light Warm-up
Moran Model Without Mutation ($\theta=0$)
Looking Backward – a Death Process
Yule Process Without Mutation ($\theta=0$) Looking Forward

What is the distribution of the number of leaves on either side of the first split?
Each cyclic permutation of $n$ objects corresponds to exactly one realization of the death process (or the Yule process).

The distribution of the number of leaves on either side of the first split is uniform.
Combinatorial Notations
Combinatorial Notations

Rising factorial

Stirling number of the first kind

The number of permutations of $n$ objects with $k$ cycles

The rising factorial is the ordinary generating function of permutations, with $\theta$ marking the number of cycles.

Falling factorial

Something a Little Heavier

J Hein et al. (2005) Gene Genealogies, Variation and Evolution
Probability of $k$ Types in the Sample With Mutation

$$\theta^n P\{k \text{ types}\}$$

The number of permutations of $n$ objects with $k$ cycles, marked with $\theta^k$

Stirling number of the first kind
Stirling cycle number

$k$ cycles
$k = 4$

Ewens’ Sampling Formula

\[ \theta^n P \{ 1^{a_1}, 2^{a_2}, \ldots, n^{a_n} \} \]

\[ k = \sum_{i=1}^{n} a_i \]

\[ n = \sum_{i=1}^{n} ia_i \]

\[ k \text{ cycles} \]

\[ k = 4 \]

Most Recent Common Ancestor (MRCA) Without Mutation

What is the probably that in a sample of $n$ individuals, a random subset of $m$ coalesces to its MRCA before coalescing with any other individuals?

In the (circular) sample permutation, the individuals are consecutive.

In the (circular) ancestral permutation, relative to the $m$ individuals in the sample permutation, both end numbers are smaller than the internal numbers.

$\pi$

In Hein et al. (2005) Gene Genealogies, Variation and Evolution
PROBLEMS
No Matter how Great and Destructive your Problems May Seem Now, Remember, You’ve Probably only Seen the Tip of Them.

www.despair.com
Distribution of Ancestors Without Mutation

$k=4$ ancestors
Distribution of Ancestors
Without Mutation

\[ n!(n-1)! \mathbb{P}\{ A_k = \{\xi_1, \ldots, \xi_k \}\} \]

\#(realizations producing \( A_k \))

\[ |\xi_1|! \cdots |\xi_k|! \cdot k!(k-1)!(n-k)! \]

\( k \) = 4 ancestors

Distribution of Ancestors With Mutation

\[ n! \theta^n \mathbb{P} \left\{ A_k = \left\{ \xi_1, \ldots, \xi_k \right\}, \left( \eta_{(1)}, \ldots, \eta_{(j)} \right) \right\} \]

\#(realizations marked by \( \theta \) producing \( A_k \))

\[ \left[ \begin{array}{c|c|c|c|c} \xi_1 & \ldots & \xi_k & k! \\ \hline \eta_{(1)} & \ldots & \eta_{(j)} & ! \\ \hline \eta_{(1)} + \ldots + \eta_{(j)} & \ldots & \eta_{(j)} & ! \\ \hline \theta^j (n-k) & \theta^k & & & \end{array} \right] \]

\( k=4 \) ancestors

Distribution of Sample Types

\[ \theta^n P \left\{ A_0 = \left( \eta_1, \ldots, \eta_j \right) \right\} \text{ plus } k - j \text{ younger types} \]

\#(realizations marked by \( \theta \) producing \( A_0 \))

\[
\frac{n^{s_j}}{n(n-s_1) \cdots (n-s_{j-1})} \left[ \begin{matrix} n - s_j \\ k - j \end{matrix} \right] \theta^k
\]

oldest to youngest

total length of first \( i \) cycles

\[ s_i = \left| \eta_1 \right| + \ldots + \left| \eta_i \right| \]
Distribution of All Sample Types

\[ \theta^n \mathbb{P}\left\{ A_0 = \left( \eta_1, \ldots, \eta_k \right) \right\} \]

#(realizations producing \( A_0 \), marked by \( \theta \))

\[ \frac{n!}{n(n-s_1)\cdots(n-s_{k-1})} \theta^k \]

\( s_i = |\eta_1| + \ldots + |\eta_i| \)

\( (\eta_1, \ldots, \eta_k) = (2, 3, 1, 2) \)

oldest to youngest

Something Really Heavy

ADVERSITY
That which DOES NOT KILL ME POSTPONES the Inevitable.
www.despair.com
Embed a death process for a known sample of size $n$ in the death process for a “super-sample” of size $M$.

Use the correspondence with random permutations in $S_M$ to derive results for the super-sample, conditioned on the sample.

“It will be expected that various exact results hold for the Moran model…”

Let $M \to \infty$ to derive results relating the sample to an infinite population.

Because of invariance theorems, such results often pertain to several different genetics models.

Distribution of Sample Types Within an Infinite Population

\[
P\left\{ A_0 = \left( \eta_1, \ldots, \eta_k \right) \right\} = \frac{n!}{n \left( n - s_1 \right) \cdots \left( n - s_{k-1} \right)} \frac{\theta^k}{\theta^n}
\]

oldest to youngest types within sample

\[
P\left\{ A'_0 = \left( \eta_1, \ldots, \eta_k \right) \right\} = \frac{n!}{\left( \theta + n \right) \left( \theta + n - s_1 \right) \cdots \left( \theta + n - s_{k-1} \right)} \frac{\theta^k}{\theta^n}
\]

oldest to youngest types within sample are in fact the \( k \) oldest types in the population

\[
s_i = |\eta_1| + \ldots + |\eta_i|
\]

\( k = 1 \)


“Species Deletion Property”
PERSISTENCE

It’s Over, Man. Let Her Go.

www.despair.com
Summary

- Population Genetics Models – Forward Time
  - Models without mutation and infinite-alleles infinite-sites models
    - Wright-Fisher Model
    - Moran Model
    - Yule Model

- The Death Process for the Models – Backward Time
  - The waiting times are independent of the combinatorics.
  - Invariance theorems: the ancestral tree of a population sample converges to a single limit for many genetics models (the coalescent process).

- The death process maps to random permutations, so counting permutations provides many results for genetic samples.

- The mapping to random permutations also gives insight into results for the entire population
  - Species Deletion Property
Laszlo Szekely
Eva Czabarka

STRIFE
As Long as We Have Each Other, We'll Never Run Out of Problems.

www.despair.com