Table 1: Data from Buri’s (1956) table 17, p. 395. $V$ is variance of population allele frequencies and $H$ is heterozygosity.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Series I</th>
<th>Series II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H$</td>
<td>$V$</td>
</tr>
<tr>
<td>0</td>
<td>0.514</td>
<td>0.006</td>
</tr>
<tr>
<td>1</td>
<td>0.464</td>
<td>0.026</td>
</tr>
<tr>
<td>2</td>
<td>0.504</td>
<td>0.031</td>
</tr>
<tr>
<td>3</td>
<td>0.456</td>
<td>0.042</td>
</tr>
<tr>
<td>4</td>
<td>0.448</td>
<td>0.050</td>
</tr>
<tr>
<td>5</td>
<td>0.428</td>
<td>0.055</td>
</tr>
<tr>
<td>6</td>
<td>0.403</td>
<td>0.062</td>
</tr>
<tr>
<td>7</td>
<td>0.402</td>
<td>0.072</td>
</tr>
<tr>
<td>8</td>
<td>0.358</td>
<td>0.083</td>
</tr>
<tr>
<td>9</td>
<td>0.348</td>
<td>0.090</td>
</tr>
<tr>
<td>10</td>
<td>0.325</td>
<td>0.105</td>
</tr>
<tr>
<td>11</td>
<td>0.305</td>
<td>0.112</td>
</tr>
<tr>
<td>12</td>
<td>0.263</td>
<td>0.123</td>
</tr>
<tr>
<td>13</td>
<td>0.255</td>
<td>0.136</td>
</tr>
<tr>
<td>14</td>
<td>0.216</td>
<td>0.140</td>
</tr>
<tr>
<td>15</td>
<td>0.202</td>
<td>0.155</td>
</tr>
<tr>
<td>16</td>
<td>0.210</td>
<td>0.160</td>
</tr>
<tr>
<td>17</td>
<td>0.197</td>
<td>0.165</td>
</tr>
<tr>
<td>18</td>
<td>0.183</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Practice Exercises for the 2nd Exam
November 7, 2007

1 Genetic Drift

Table 1 presents some of the data from Buri’s (1956) experiment. We have also put these data onto the class web site so that you will not have to type them in by hand. This homework assignment would be hard work by hand, but it will be easy if you use a spreadsheet or write your own program.

Parameters The allele frequency in generation 0 is $p = 1/2$; in series I, $2N_e = 18$. You will not need the data for series II.

*EXERCISE 1 Using the $H_0$ and $2N_e$ values for series I (table 1) and the formula for pure genetic drift

$$H_t = H_0(1 - 1/2N)^t$$

Gillespie’s eqn 2.3, p. 27 calculate the expected heterozygosity in generations 1–18. To display the results, make a graph with generation on the horizontal axis and heterozygosity on the vertical axis. Plot observed heterozygosity using one symbol (or color) and predicted heterozygosity using another.

The next few questions involve several gene diversity estimates from the African human population:

- Protein polymorphisms: 0.179
- Restriction site polymorphisms: 0.322
- Tetra-nucleotide repeat polymorphisms: 0.769
- Di-nucleotide repeat polymorphisms: 0.807

For the sake of argument, let us assume that (1) all these systems are at equilibrium between mutation and genetic drift under the model of infinite alleles, and (2) the mutation rate for protein polymorphisms is $10^{-6}$ per generation.

*EXERCISE 2 What do the protein polymorphism data imply about the effective size of the African population? (Hint: use Gillespie’s equation 2.10.)

*EXERCISE 3 What would you estimate the mutation rates to be for

1. restriction site polymorphisms?
2. tetra-nucleotide repeats?
3. di-nucleotide repeats?

(Hint: use Gillespie 2.10 again.)

*EXERCISE 4 It is often assumed that $u \approx 10^{-6}$ for nuclear protein polymorphisms, and that $u \approx 10^{-3}$ for STRs such as the tetra-nucleotide and di-nucleotide repeats in our data. Are these values consistent with your results? If not, speculate about the discrepancy.

2 Two Loci

Consider the following data set:

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>$AB$</th>
<th>$Ab$</th>
<th>$aB$</th>
<th>$a$</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>70</td>
<td>50</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

*EXERCISE 5 What are the four haplotype frequencies (Gillespie’s $x_1$, $x_2$, $x_3$, and $x_4$)?

*EXERCISE 6 What are the frequencies $p_A$ and $p_B$ of alleles $A$ and $B$?
\* EXERCISE 7 What is \( D \)?

\* EXERCISE 8 What \( r^2 \)? (This \( r \) has nothing to do with the recombination rate, which is also called \( r \). See Gillespie’s p. 105.)

\* EXERCISE 9 Suppose the recombination rate is \( r = 1/2 \). What would \( D \) be after 10 generations?

\* EXERCISE 10 Repeat question 9, but this time assume that \( r = 0.001 \).

If two sites are 1 centimorgan (cM) apart on a chromosome, then the rate of recombination between them is \( r = 0.01 \). (This is the definition.) How does this relate to physical distance along the chromosome? In general, 1 cM is about a million base pairs—a “megabase.” In the exercise below, you will use this generalization to calculate the probability \( \lambda \) of recombination between two adjacent nucleotides.

As a first approximation, you might suppose that

\[
r = \lambda k
\]

where \( k \) is the physical distance between the two sites—the number of nucleotides between them. Our rule of thumb says that \( r = 0.01 \) when \( k = 10^6 \). Thus,

\[
\lambda = 0.01/10^6 = 10^{-8}
\]

The rate of recombination between adjacent nucleotides is about \( 10^{-8} \).

But this calculation is only approximate. To see why, consider sites A and B below.

---A1-------------------------B1

---A2-------------------------B2

---A1-----\ /--------\ /--------B1

---A2-----/ \--------/ \--------B2

In the “Before” picture, we see two chromosomes one the \( A_1B_1 \) haplotype, the other with \( A_2B_2 \). The haplotypes, however are unchanged. There were two cross-overs, but no recombination. In general, a recombination event results from an odd number of cross-overs. When this number is even, the haplotypes stay the same.

In 1919, JBS Haldane showed that the probability of an odd number of crossovers is

\[
r = (1 - e^{-2\lambda k})/2
\]

This is called Haldane’s mapping function.\(^1\)

\* EXERCISE 11 Use the rule of thumb (a centimorgan is a megabase) to solve for \( \lambda \). How does your result compare to the approximate calculation above?

\* EXERCISE 12 Make a graph. The horizontal axis represents \( \lambda k \) and should run from 0 to 0.5. On this graph, plot two curves: (1) the approximate formula \( r \approx \lambda k \), and (2) Haldane’s mapping function, as defined above. The vertical difference between the two lines shows the error involved in the approximate formula. Over what range of \( \lambda k \) is the approximation satisfactory? What is this range, assuming that \( \lambda = 10^{-8} \).

\* EXERCISE 13 On p. 110, Gillespie shows that a selective sweep removes variation at linked sites provided (roughly) that \( r/s < 0.1 \). How large is the affected region of the chromosome if (a) \( s = 0.001 \), (b) \( s = 0.01 \), and (c) \( s = 0.1 \)?

Answers to Exercises

\* EXERCISE 1 First find the value of \( H_0 \), the heterozygosity in the initial generation. There are two equally valid ways to do this. You could take \( H_0 = 2p(1 - p) = 0.5 \), where \( p = 0.5 \) is the initial allele frequency, or you could take \( H_0 = 0.514 \), since that was the observed heterozygosity in the initial generation.

The answers in table 1 assume that \( H_0 = 0.5 \).

\* EXERCISE 2 Equation 2.9 says that

\[
\hat{H} = \frac{4Nu}{4Nu + 1}
\]

In evaluating this equation, there is more than one way to proceed. The approach I use here is one you may not have thought of. The first step in my approach is to notice that the equation above can be re-expressed as

\[
4Nu = \frac{\hat{H}}{1 - \hat{H}}
\]  

(1)

The quantity \( \hat{H}/(1 - \hat{H}) \) is interesting in its own right because, at equilibrium between mutation and drift, it is proportional to \( N \) and is also proportional to \( u \).

\(^1\)Haldane’s formula assumes that recombination occurs at the same rate throughout the chromosome. In reality, the rate varies greatly along the chromosome.
It is often more convenient to work with $H/(1 - H)$ than it is to work directly with $H$.

For protein polymorphisms, equation 1 becomes

$$4N \times 10^{-6} = \frac{0.179}{1 - 0.179} = 0.218$$

or

$$N = 10^6 \times 0.218/4 = 54,507$$

*EXERCISE 3* These problems are just like the previous one except that now it is $u$ that is unknown rather than $N$. To simplify the calculations, let us add another column to the data table:

<table>
<thead>
<tr>
<th>System</th>
<th>$\hat{H}$</th>
<th>$\hat{H}/(1 - \hat{H})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>0.179</td>
<td>0.218</td>
</tr>
<tr>
<td>Restriction site</td>
<td>0.322</td>
<td>0.475</td>
</tr>
<tr>
<td>Tetra-nucleotide repeats</td>
<td>0.769</td>
<td>3.329</td>
</tr>
<tr>
<td>Di-nucleotide repeats</td>
<td>0.807</td>
<td>4.181</td>
</tr>
</tbody>
</table>

According to equation 1, the right column is an estimate of $4Nu$. If the value of $N$ is the same for all these systems, then the differences in these numbers must be due to differences in $u$. (We're ignoring sampling error.) Dividing the second of these values by the first gives $0.475/0.218 = 2.17$. Thus, the mutation rate for restriction site polymorphisms must be 2.17 times as large as that for protein polymorphisms. Since we already know that the latter rate equals $10^{-6}$—that was assumed at the outset—the mutation rate for restriction site polymorphisms must be $2.17 \times 10^{-6}$.

Similarly, the rate for tetra-nucleotide repeat polymorphisms is

$$10^{-6} \times \frac{3.329}{0.218} = 1.53 \times 10^{-5}$$

and that for di-nucleotide repeats is

$$10^{-6} \times \frac{4.181}{0.218} = 1.92 \times 10^{-5}$$

*EXERCISE 4* The calculations above yield mutation rates for the repeat loci that are on the order of $10^{-5}$ rather than $10^{-3}$. This is two orders of magnitude smaller than the conventional view would predict.

*EXERCISE 5* $x_1 = 0.25$, $x_2 = 0.35$, $x_3 = 0.25$, $x_4 = 0.15$

*EXERCISE 6* $p_A = 0.6$, $p_B = 0.5$.

*EXERCISE 7* $D = -0.05$

*EXERCISE 8* $r^2 = 0.04$

*EXERCISE 9* $D_{10} = -0.000049$

*EXERCISE 10* $D_{10} = -0.0495$

*EXERCISE 11* Solving for $\lambda$ gives: $\lambda = -\ln(1 - 2r/2k)$. With $r = 0.01$ and $k = 10^6$, this is $\lambda = 1.01 \times 10^{-8}$. The approximation is thus very accurate here. (It is less accurate when $r$ is large.)

*EXERCISE 12* The approximation is within 10% of the truth provided that $\lambda k$ is less than about 0.1. For $\lambda = 10^{-8}$, this implies that $k < 10^7$. Thus, the approximate formula is good out to about 10 megabases.

*EXERCISE 13* The formula implies that variation will be removed at sites for which $r < 0.1s$. With $s = 0.001$, this is $r < 0.0001$. In view of our rule of thumb ($r = \lambda k$, discussed above), we have $\lambda k < 0.0001$. We estimated that $\lambda \approx 10^{-8}$, so the inequality becomes $k < 10000$. Thus, the sweep will remove most variation within about 10 kb of the selected site. In parts b and c, selection is stronger, and the answer becomes 100 kb (for part b) and 1000 kb (for c).