Lecture 1

Descriptive Statistics for DNA Sequences

1.1 DNA sequence data

Not until the 1980s did population geneticists begin the study of DNA sequence data. Until then, our measures of genetic variation were incomplete. We worked only with a small fraction of the genetic variation in our samples. With DNA sequence data we were finally able to study it all.

But this opportunity posed an immediate challenge. How should we *measure* that variation? Population geneticists were used to summarizing variation with statistics such as the sample heterozygosity: the probability that two random gene copies are copies of different alleles. But if the DNA sequences are long enough, it is unlikely that any two of them will be identical. The heterozygosity, in other words, is always 1. Clearly, new measures of variation are needed.

Table 1.1: Ten DNA sequences, each consisting of 40 sites. The sites are numbered across the top. The dots represent sites that are identical to the *reference sequence* at the top.

```
              0000000001 1111111112 2222222223 3333333334
              1234567890 1234567890 1234567890 1234567890

Sequence01    AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA
Sequence02    .......T.. .C......TG C......C.. ..........
Sequence03    ..C....... .......... .......... ..........
Sequence04    .......T.. .C......TG C......... G.........
Sequence05    .......... .......... .......... ..........
Sequence06    .....A.... ........T. C......... G....C....
Sequence07    ..C....T.. .C......TG C......... G.........
Sequence08    .....A.T.. TC......TG C......... G.........
Sequence09    .......... .......... C......... ..........
Sequence10    .G...A.... ........T. C......C.. .T....C..G
Segregating:   ^^  ^ ^   ^^      ^^ ^      ^   ^^   ^^  ^
```

Table 1.1 presents 10 DNA sequences from some hypothetical species. Take a minute to study them. How many ways can you think of to summarize the variation in these data? This is precisely the problem that confronted population geneticists during the 1980s. The lecture that follows will summarize some of the ideas they came up with.

1.2 Statistics

Gene diversity (a.k.a. heterozygosity) is the probability that two random sequences are different. To calculate it, the straightforward approach is to examine all pairs and count the fraction of the pairs in which the two sequences are different from each other. It is often faster, however, to start by counting the number of copies of each type in the data. Let k_i denote the number of copies of type i, and $K = \sum k_i$ the number of gene copies in the sample. The the heterozygosity is estimated by

$$H = 1 - \sum_i \left(\frac{k_i}{K}\right)\left(\frac{k_i - 1}{K - 1}\right)$$

In the past, we have expressed heterozygosity as $2p(1-p)$ (for bi-allelic loci) or as $1 - \sum_i p_i^2$ (for loci with multiple alleles). These formulas are correct when p is the population allele frequency of the parents but contain a subtle bias when p is the allele frequency within a sample. The new formula corrects this bias.[1]

Number of segregating sites A "segregating site" is a site that is polymorphic in the data. The number of such sites is usually denoted by S.

Mean pairwise difference The average number Π of differences between pairs of sequences.

Mean pairwise difference per nucleotide If the sequences are L bases long, it is often useful to standardadize Π by dividing it by L. The resulting statistic is

$$\pi = \Pi / L$$

Mismatch distribution A histogram whose ith entry is the number of pairs of sequences that differ by i sites. Here, i ranges from 0 through the maximal difference between pairs in the sample.

Site frequency spectrum A histogram whose ith entry is the number of polymorphic sites at which the mutant allele is present in i copies within the sample. Here, i ranges from 1 to $K - 1$.

Folded site frequency spectrum It is often impossible to tell which allele is the mutant and which is ancestral. In that case, we combine the entries for i and $K - i$, so the new i ranges from 1 through $K/2$.

[1] Imagine drawing two gene copies without replacement from a sample of size K. The first is a copy of allele A_i with probability k_i/K. Given this, the second is a copy of A_i with probability $(k_i - 1)/(K - 1)$. Thus, the sum of these quantities is the homozygosity and 1 minus this sum is the heterozygosity.

1.3 Data analysis

1.3.1 The number (S) of segregating sites

On the last line of Table 1.1, segregating (i.e. polymorphic) sites are indicated with a caret (^). There are 15 such sites. Thus, the number of segregating sites is $S = 15$.

1.3.2 The mean pairwise difference (Π)

We want the average number of differences between pairs of individuals. There are two ways to do this calculation, the direct way and the easy way.

The direct way

Count the number of differences between each pair of sequences. For example, sequences 1 and 2 differ at 6 sites. Compare every pair of sequences, and write down the number of differences between each pair. If you do this (and I don't recommend it), you should end up with 45 numbers that sum to 248. The average is $\Pi = 248/45 = 5.511111$.

The easy way

The direct calculation involved two steps. Step 1 calculated the number (248) of pairwise differences, and then step 2 divided by the number (45) of pairs. The first of these numbers can be thought of as a sum over sites: the number of pairwise differences at site 1 plus that at site 2 and so on. The monomorphic sites make no contribution to this sum, so we need consider only the 15 polymorphic sites. And each site makes a contribution that is easy to calculate.

Suppose that at some site the sample contains only two nucleotides: x As and y Gs. Among pairs of sequences there will be some AA pairs, some AG pairs, and some GG pairs, but only the AG pairs will contribute a difference. The number of such pairs is $x \times y$, so this is the value that this particular site makes to the sum of pairwise differences.

For example, consider site 6 in the data above. There are 3 As and 7 Gs, so there are $3 \times 7 = 21$ AG pairs, and site 6 contributes 21 to the sum of pairwise differences. At site 2, on the other hand, there are 1 G and 9 As, so the site contributes $1 \times 9 = 9$ to the sum. Summing across the 15 polymorphic sites gives 248 as before.

There is also an easy way to find the number of pairs. In a sample of K sequences, there are $K(K-1)/2$ pairs. There are 10 sequences in the data above, so the formula gives $(10 \times 9)/2 = 45$ pairs.

1.3.3 Computer output

Here is the output of my seqstat program, which calculates descriptive statistics for DNA sequences:

```
%                              seqstat
%                (descriptive statistics from sequence data)
%                            by Alan R. Rogers
%                               Version 5-1
```

```
%                                30 Jan 2000
%                         Type `seqstat -- ' for help

% Cmd line: seqstat af10.seq

% Population 0
meanPairwiseDiff = 5.511111 ;
nsequences = 10 ;
nsites =  40 ;
mismatch = 1 5 3 2 2 6 8 8 5 2 2 1 ;
segregating = 15 ;
spectrum =  6 2 2 5 0 ;
% Count of minor allele at each polymorphic site:
%psite  site  count | psite  site  count
     1     2      1 |     9    21      3
     2     3      2 |    10    28      2
     3     6      3 |    11    31      4
     4     8      4 |    12    32      1
     5    11      1 |    13    36      1
     6    12      4 |    14    37      1
     7    19      4 |    15    40      1
     8    20      4 |
```

⋆ EXERCISE **1–1** Here is a set of 10 made-up DNA sequences, each with 10 nucleotide sites.

```
S01        AAACT GTCAT
S02        ..... A....
S03        ..G.. A....
S04        ..G.. A....
S05        ..... A....
S06        ..... AC...
S07        ..... A....
S08        ..... A....
S09        ..... A...C
S10        ..... A....
```

Calculate the mean pairwise difference, the number of segregating sites, the mismatch distribution and the site frequency spectrum.