Massachusetts Institute of Technology
6.877 Computational Evolutionary Biology, Fall, 2005
Laboratory 2, Part I: Evolution, polymorphism, and the coalescent

Handed out: October 18, 2005 **Due: October 31, 2005**

This portion of the laboratory is PART I of 2 parts. PART II will be handed out next Monday. PART I consists of a series of 'problem set' type exercises followed by some warmup computer exercises.

Exercises with the basic theory.
Preamable: some review and 'cheat sheet' summary.

0. Definitions and descriptive statistics for DNA sequences

Heterozygosity (also known as 'gene diversity') is the probability that two random sequences are different. To calculate it, one can straightforwardly examine all sequence pairs and coun the fraction of the pairs in which the two sequences are different from each other. It is often faster to start by counting the number of copies of each type in the data. Let n_i denote the number of copies of type *i* and let *n* be the sum of all these types. Then the heterozyosity is estimated by:

$$H = 1 - \sum_i \left(\frac{n_i}{n}\right)\left(\frac{n_i - 1}{n-1}\right)$$

Number of segregating sites. A 'segregating site' is a site that is polymorphic in the data. The number of sites is usually denoted S (or S_n if referring to n samples).

Mean pairwise difference. Let k_{ij} denote the number of nucleotide site differences between sequences of type *i* and sequences of type *j*. The mean pairwise difference is:

$$\prod = \sum_{i<j} k_{ij} \left(\frac{n_i}{n}\right)\left(\frac{n_j}{n-1}\right)$$

Mean pairwise difference per nucleotide. If the sequences are *L* bases long, we can standardize the above value by the length:

$$\pi = \Pi / L$$

Mismatch distribution. A histogram whose i^{th} 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 i^{th} entry is the number of polymorphic sites at which the mutant allele is present in *i* copies within the sample. Here, the indexing *i* ranges from 1 to $n\text{–}1$.

Folded site frequency spectrum. Often one cannot tell which allele is the mutant and which is ancestral. In that case, we combine the entries for *i* and $n\text{–}i$ so the new *i* ranges from 1 through $n/2$.

Neutral substitutions. The rate at which neutral mutants are substituted into the population depends on several factors.

The number of new mutant genes introduced per generation. In a population of $2N$ genes, this number is $2Nu$.

The fraction of those mutants that eventually become fixed. If drift continues long enough, all but one of the genes in the current population will be lost, and the one that survives will be fixed. Each gene has a probability of $1/(2N)$ of increasing to fixation (its initial frequency).

The rate of substitution is the number of new mutants per generation ($2Nu$) times the probability that a given mutant will be fixed, $1/(2N)$. So this product is just u: remarkably, the substitution rate is just equal to the rate of mutation of neutral alleles. It does not depend on the size of the population or the extent of subdivision within it.

We now turn to the problem set questions.

Question 1. Effective population size.

We have seen that the effective population size, N_e, plays a crucial role in all the theoretical accounts of evolution. Since we multiply this (typically large) number by another (typically very small) factor, *s, u, m,* or *r* (recombination rate), to get the crucial value of theta, it's important to get this value as correct as we can make it, because this will have profound impact on the outcome of polymorphism patterns. Thus, it's actually extremely important to figure out how to correct for different variations that are actually seen biologically. In class (and the slides), we saw how to correct for fluctuating population size by 'back calculating' what the population 'should have been' in order to get the same change in variance as would be produced by a Wright-Fisher model. This idea, viewed from another perspective, is also covered in the Rice book, pages 111–112. Here's another common correction that is often encountered in biological situations.

Elephant seals possess an interesting system of mating. One alpha-male lies in the center of a harem of females and attempts to mate with as many of these females as possible throughout the course of the mating season. The harem is also surrounded by 5 beta-males, usually younger and smaller, who lie around the harem and protect it from invasion by other males. When the alpha-male starts to mate, the beta-males use his distraction to do some mating of their own with the females at the outer edges of the harem. Other males, not alphas or betas, stay in the water and are not allowed to mate at all. One would think that the alpha male sires the most offspring in this situation, however, recent genetic studies have found that beta-males as a group actually have at least as much mating success as the alpha-male. From a typical harem 50% of the children are sired from beta males and 50% are sired from the alpha. Given that there are 40 alpha-males, and 2000 females in the population, assuming that each of the harems is exactly the same size, assuming that all beta-males have exactly the same probability of having an offspring, and assuming that all females are fertile and available for mating, **calculate the effective population size** for this population of elephant seals. NOTE: assume non-overlapping generations for simplicity.

To solve this problem, you should read the relevant section in the Rice textbook, pages 113–114; please show the details of your work and how you arrive at your answer (not just a final number!) (And, Yes, one can just go through the Rice book and follow that derivation, but it's more interesting to work out the details as applied to this particular case.)

Question 2. Neutral evolution and drift. How would an increase in population size affect the rate of neutral substitutions in the generations immediately following the increase?

Question 3. Neutral evolution and drift. Consider a stretch of noncoding, nonfunctional DNA that is 1, 000 nucleotides long. Assume that the mutation rate is high – 2 changes per site per million generations.

Question 3(i) In a population of 10,000 individuals, how many new alleles will be created in the population each generation?

Question 3(ii) What fraction of these new alleles will ultimately be fixed?

Question 3(iii) What is the rate of molecular evolution? (substitutions per site per generation)

Question 3(iv) In a population of 50 individuals with the same mutation rate, how many new alleles will be created per generation? What fraction of these will ultimately be fixed? What is the rate of molecular evolution?

Question 3(v) Which population will have greater heterozygosity at any one time? Why?

Question 4. The neutral theory of evolution

Question 4(i) In a population of 10,000 individuals with a neutral mutation rate of 1 nucleotide change per site per 10 million generations, what is the expected heterozygosity per site?

Question 4(ii) What is the expected heterozygosity for a protein that is 500 amino acids long? Express the meaning of this result in a sentence (i.e., express it in words).

Question 4(iii) What is the expected heterozygosity if there are 12 individuals in the population?

Question 5. Segregating sites, *S*.
What is the ratio between the expected value of *S* in a sample of 100 DNA sequences and the expected value in a sample of 200? (You should do the calculation, and, if you want to go a little further, try to derive the ratio mathematically).

Question 5. Nucleotide polymorphisms and the number of segregating sites, *S*
Here is a set of 10 (real) DNA sequences, each with 40 sites, taken from a mitochondrial DNA sample:

```
seq1          AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA□
seq2          .......T.. .C......TG C......C.. ..........□
seq3          ..C....... .......... .......... ..........□
seq4          .......T.. .C......TG C......... G.........□
seq5          .......... .......... .......... ..........□
seq6          .....A.... ........T. C......... G....C....□
seq7          ..C....T.. .C......TG C......... G.........□
seq8          .....A.T.. TC......TG C......... G.........□
seq9          .......... .......... C......... ..........□
seq10         .G...A.... ........T. C......C.. .T....C..G□
```

Each column corresponds to a different site. Periods indicate sites that are identical to the site in sequence 1. To save typing, the file is available for download in the labs section: Question 5 sequence data. Use these data to answer the following questions.

Question 5(i). Calculate the mean pairwise difference among the sequences, the nucleotide diversity π.

Question 5(ii). Use this value to estimate θ.

Question 5(iii). Calculate the number of segregating sites, *S*, given this sample (more properly denoted S_{10}).

Question 5(iv). Use this value to provide a second estimate of θ.

Question 5(v). What can you infer from the similarity (or dissimilarity) of these two estimates?

To ease your computation, I have posted a very simple `awk` script `piS` that computes the diversity statistic π and an estimate of θ based on the Waterston segregating sites statistic (along with some other things, viz, Tajima's *D* statistic), for a set of aligned sequences like this. The script is available in the labs section. and ought to run (i.e., it's been tested on) any linux box; Mac OS X, etc. (you can download it as a text file and run it locally – I will check re getting it going on the MIT server).
It is invoked this way:

```
awk –f piS.awk <name of sequence file> > <myhomedir> myoutputfile.txt
```

where you should make sure `myoutputfile.txt` is a file in your home directory. Running this program will give you output like the following,

```
                         pi: 53.626812
          Segregating sites: 184/1878
theta_hat[estimated from S]: 49.273068
                 Tajima's D: 0.354335
a1=3.734292 a2=1.602387 b1=0.362319 b2=0.242754
c1=0.094530 c2=0.067558 e1=0.025314 e2=0.004345
```

The a1, a2, etc. coefficients are those needed to compute Tajima's *D,* as part of the variance calculation devised by Tajima.

Part 2: Analyzing some real polymorphism data

The same program can be used to analyze full-size data sets. As an example, consider the paper by Hey *et al.* (1993) in *Genetics* that examined a 1.9kb stretch of one locus, the *per* locus, in *Drosophila melanogaster* and three sister species. They were interested in details about speciation history, and common ancestors, and so were eventually drawn to a coalescent analysis. Here, we'll start to get our hands dirty by taking a look at the whole dataset.

The full *Drosophila* dataset as used in the original Hey *Genetics* paper and in this question is available for download: Sites Drosophila data set. (We provide this format for completeness because it can be used later on in Hey's own program, later on.) For now, though, we need the in the form that the awk program likes; that data set is in the labs section.

Question 6. Repeat Questions 5(i)–(v) for this dataset. Be sure to think about the fact that this is not data all drawn from a single species – in particular, it might not meet the requirements of the infinite site model either. (Please examine the data at least briefly for this possibility – whether this has an effect on the analysis is a different story, of course. In addition, as the Hey paper notes, there is the possibility of population growth and migration, none of which is taken into account by these summary statistics.)

In addition, at this point you should take a look at the description of Tajima's *D,* and go back and consider what it means in terms of the data for both Question 5 and Question 6. (We'll push on this same data more later on.) For a description of Tajima's *D,* see below.

Question 7. Tajima's *D*.

Fumio Tajima introduced a statistic that is widely used to test the null hypothesis of mutation-drift equilibrium and constant population size. Tajima considered two statistics: the mean pairwise difference π and the number *S* of segregating sites. Under the null hypothesis, the expected values of these statistics are:

$$E[\pi] = \theta$$

$$E[S] = a_1\theta$$

where θ=4*Nu*, 2*N* is the haploid population size, *u* is the mutation rate per generation, and a_1 is defined below, and should be familiar you from the coalescent process as the expected total length of time to coalescence, hence the time available for mutation:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

Under the null hypothesis, both π and S/a_1 estimate θ, so they should be roughly equal in value. If they are about equal in value, then we cannot reject the hypothesis. If they are very different, on the other hand, we reject the hypothesis. But how different is "very different"? The answer depends on how variable these two statistics are from sample to sample. Tajima obtained a formula for the sampling variance of these statistics, and so defined *D* this way: Let

$$V = Var[\pi - S / a_1]$$

denote the sampling variance of the difference between the two estimates. Then Tajima's *D* is:

$$D = \frac{\pi - S / a_1}{\sqrt{V}}$$

It expresses the difference between the two estimates relative to their standard error. If the difference between π and S/a_1 *were* normally distributed, then we could expect *D* to lie between negative 2 and +2 about 95% of the time. In fact, this difference is *not* normally distributed (technically, Tajima argued that it has what is called a *beta distribution*) but it is not too far off. We should be suspicious of values of *D* that are much outside the interval [–2*,* 2]. Later in the second part of this Lab, we shall see how to use coalescent programs to generate confidence intervals via simulation.

Although *D* is simple in concept, it is tedious to calculate. We need three pieces of data: *S*, and the number, *n*, of DNA sequences in the sample. Given these data, Tajima used the fact that the variance in the estimate of *S* derived from nucleotide diversity can be expressed this way:

$$Var[\hat{\theta}_T] = \frac{n+1}{3(n-1)}\theta + \frac{2}{9}\theta^2$$

If we go through a (fairly sophisticated) calculation of the difference in the variance of the two estimators of θ (which we omit here), as it turns out we need the following numbers:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2} \text{ (this is the expected total gene tree size)}$$

$$b_1 = \frac{n+1}{3(n-1)}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

$$c_1 = b_1 - 1/a_1$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^{\,2}}$$

$$e_1 = c_1 / a_1$$

$$e_2 = c_2 / (a_1^{\,2} + a_2)$$

And finally, the *D* value is computed as:

$$D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}}$$

Question 7 (at last). Lynn Jorde’s lab has published a large sample of DNA sequences from the D-loop of the human mitochondrial genome. There are 630 sites. For the Asian sample, we get:

```
77 sequences, 630 sites □
pi: 8.438483□
Segregating sites: 103/630□
theta_hat[estimated from S]: 20.958331 □
Tajima’s D: -2.021749 □
a1=4.914514 a2=1.631862 b1=0.342105 b2=0.228184 c1=0.138626 c2=0.086985□
e1=0.028208 e2=0.003374□
```

For the African sample:

```
72 sequences, 630 sites □
pi:15.339984□
Segregating sites: 88/630 □
theta_hat[estimated from S]: 18.155855 □
Tajima’s D: -0.525801 □
a1=4.846921 a2=1.630948 b1=0.342723 b2=0.228612 c1=0.136406 c2=0.085989□
e1=0.028143 e2=0.003423□
```

In one case, *D* is strongly negative and in the other case weakly negative. How would you interpret these results?

(This (almost) ends Part I of the laboratory.)