Biostatistics 666
Problem Set 2
Due February 2, 2006

1. Consider a sample of $n = 2$ DNA sequences from a population of size $N_e = 250$ diploid individuals (i.e., $2N = 500$ sequences).

   a) What is the expected coalescence time for the two sequences?

   b) What is the probability that coalescence occurs at generation 500?

2. Consider a sample of $n = 10$ DNA sequences from a population of size $N_e = 2500$ diploid individuals.

   a) What is the expected time to the 1st, 2nd, 3rd ... 9th coalescence event?

   b) What is the expected coalescence time to the MRCA of all 10 sequences?

   c) What is the expected number of polymorphisms in the sample? Assume that the sequences are 1,000 base pairs long and that the mutation rate is 10^{-8} per base-pair per generation.

   d) What is the probability that all sequences are identical and no polymorphisms are observed in the sample?

3. Two 1,000 bp sequences were compared and 10 differences were observed. Assuming the mutation rate is 10^{-8} per base-pair, and the effective population size is $N_e = 10,000$, what is the probability of observing more than 3 differences? What assumptions did you make to complete the calculation?

   What might explain the large number of differences between the two sequences?