# CSE 280a Final Exam

March 21, 2019

## Notes

*You must work independently on the exam. No collaboration with any other person is allowed. Please cite any published resource that you use. The answers must be type-written. Points will be given for clarity, and concise exposition. The goal of the class is to learn algorithmic techniques applied to the analysis of genetic data, and the exam is meant to reflect that. Please let me know if you have feedback.*

## Answer all of the following:

1. **Equilibria.** Suppose we have sampled the two chromosomes of a diploid individual and find a run of homozygosity involving $k$ consecutive sites, each of which has a population minor allele frequency of at least 0.1. How large does $k$ need to be so that the p-value of observing this data under Hardy Weinberg equilibrium becomes lower than $10^{-4}$.

2. **Recombination.** We did not study recombination in class in detail. This problem does not assume that you know the biological underpinnings, but is focused on the combinatorial analysis. Consider a SNP matrix $M$ with $n$ distinct haplotypes, and $m$ distinct SNPs. Assume that the infinite sites assumption holds, but we do have recombination. *As the recombination events do happen at specific locations in between sites, we cannot represent the genealogy by a tree, nor can we reorder columns of the matrix.* Also, the Matrix is unrooted so that 0 does not necessarily represent the ancestral mutation. Consider the matrix $M$ restricted to two columns $i$ and $j$. We say that $i, j$ satisfy the 4-gamete rule if the rows of the restricted matrix contain all combinations $00, 01, 10, 11$.

   (a) *True or False?* If $i, j$ satisfy 4-gamete rule, then at least one recombination event happened between $i$ and $j$. Give a concise reason for your answer.

   (b) Consider 4 columns $j_1 \leq j_2 \leq j_3 \leq j_4$. (a) If $(j_1, j_2)$ satisfy the 4-gamete rule, and $(j_3, j_4)$ satisfy the 4-gamete rule, can you lower bound the number of historical recombination events in the matrix $M$? (In other words, provide a number $n$ saying that at least $n$ recombination events had to have occurred). (b) If $(j_1, j_3)$ satisfy the 4-gamete rule, and $(j_2, j_4)$ satisfy the 4-gamete rule, can you lower bound the number of historical recombination events in the matrix $M$?

   (c) Using the ideas in steps 2a and 2b, describe an algorithm that computes a lower bound on the number of historical recombination events in the entire SNP matrix. Credit will be given for efficiency *and* for the quality of the lower bound.

3. **Coalescent theory.** Consider a haploid population evolving according to the WF model from a diploid population size $N$, and assume that the infinite-sites assumption holds. Sample $n$ individuals from the current generation in a region evolving neutrally with scaled mutation rate $\theta$ and no recombination. The sample is in the form of a binary SNP matrix. Compute the expected number of 1's in the matrix.

4. **Haplotype assembly and Optimization.** Consider $n$ heterozygous sites on a single individual's chromosomal pair, where the alleles at any location are given by $0, 1$. Then, the phase resolved chromosomes will be a pair of complementary strings. For example, consider $n = 3$. In this case, the possible haplotype phase resolution is one of

$$\{(000, 111), (001, 110), (010, 101), (011, 100)\}.$$

As the strings are complementary, we can describe the *haplotype* simply by one of the two (e.g., 001 is the same as $(001, 110)$. By convention, we use 0 to represent the first bit, making possible haplotypes $000, 001, 010, 011$.

To resolve the haplotypes, we use short read sequencing, and map it back to the genome. Consider a haplotype string $h$ with $n$ bits. Each mapped read $r$ is also a string of length $n$, with $r_j \in \{0, 1, -\}$, with extra '-' symbols at locations not covered by the read. For example, let $n = 3$, and read $r = (01\text{-})$. Then, $r$ is *consistent* with haplotypes 010 (or the complement 101) and 011, but not with 001 or 000. It is clear that a collection of such reads could phase the entire chromosome. However, if some if the reads have erroneous calls, that makes the problem difficult. In this example, if 000 was the true haplotype, we would have to flip one bit in the read $r$.

Given a collection of reads $R$ and a haplotype $h$, the *edit-error* $E(R, h)$ is the number of bits that need to be flipped in the read collection $R$, so that all reads are consistent with $h$ or its complement. Describe an integer linear program to compute $\min_h E(R, h)$.

5. **MCMC.** Define an odd-permutation $\pi_n$ as a permutation over $1, \ldots, n$ such that all odd numbers are in odd positions, and even numbers are in even positions. For example, let $n = 4$. Then, $[1, 2, 3, 4]$ and $[3, 4, 1, 2]$ are examples of odd permutations, while $[1, 3, 2, 4]$ is not. You are given a score $\mathcal{S}(\pi)$ for each odd-permutation $\pi$. Describe an MCMC algorithm that samples each odd-permutation with probability proportional to its score. Recall that in an MCMC method, whenever we are in state $i$, we pick a state $j$ according to a proposal distribution $A[i, j]$, and move to to state $j$ (or stay in $i$) with a specific transition probability $T[i, j]$. Therefore, your solution should clarify the state space of the markov chain, the proposal distribution, and the transition probabilities, and you should provide an argument showing that you are indeed sampling from the target distribution when the markov chain converges.