# CSE 280a Final exam

March 13, 2008

## Notes

*You are required to work independently on the exam. No collaboration with any other person is allowed, except for Long question 1, where you can ask outside sources, but not someone in your class. Please cite any published resource that you use.*

*The answers must be type-written. Points will be given for clarity, and concise exposition. Email your final exam to me by midnight, Thursday, March 20.*

## Short Questions. Answer all of the following:

1. Given the SNP matrix below, provide the minimum number of recombinations that would explain a history assuming infinite-sites (no site mutates more than once). You can do it either by hand computations, or by writing scripts. However, your final answer must have a concise written explanation of the basis for the bound, and must not simply cite the output of a program as an answer.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| E | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| G | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| H | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| I | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |

2. Consider a fixed population of size $N$ evolving with mutation rate $\mu$ changes per generation per individual, and recombination rate $\rho = 0$. Compute the expected number of polymorphic sites in a population of $n$ individuals. How does this change for $\rho > 0$?

3. You would like to detect inversion polymorphism in a population sample. Note that if some haplotypes are inverted w.r.t others, the inverted and non-inverted haplotypes will not recombine, but evolve independently. In other words, you will not see both alleles in the inverted, and the non-inverted population. Given a binary SNP matrix sampling the population, you want to partition the individuals into two sub-populations such that a large number of SNPs are non-segregating (have the same value) in at least one of the two populations. The table below shows two SNPs ($s_1, s_3$) that have this property, and one ($s_2$) that does not.

| $s_1$ | $s_2$ | | $s_3$ | |
|---|---|---|---|---|
| 0 | 0 | | 1 | |
| 0 | 1 | | 0 | $R'$ |
| ⋮ | | | ⋮ | Inverted |
| 0 | 0 | ... | 1 | |
| 1 | 1 | | 1 | |
| 0 | 0 | | 1 | $R - R'$ |
| 1 | 1 | | 1 | |
| ⋮ | | | ⋮ | Non-inverted |
| 0 | 1 | | 1 | |

Formally, the data can be thought of as a set $R$ of haplotypes. A partition is a subset $R' \subseteq R$ of rows. The score of the partition is $\mathcal{S}(R')$ is the number of SNPs that are identical in either $R'$, or $R - R'$.

Design an integer linear program that takes as input a binary SNP matrix (set $R$), and returns $R'$ with the maximum score. Can you formulate it as a Linear Program (no integer constraints)?

# Long Questions. Answer only one of the following.

1. Data collection: The supplementary data in the following review by Sabeti lists many regions of DNA under positive selection. Extract as many of the data-sets as you can.

   ```
   http://www.sciencemag.org/cgi/content/full/312/5780/1614/DC1
   Positive Natural Selection in the Human Lineage
   P. C. Sabeti
   ```

   Note that in this problem, you are not asked to do any analysis, but simply to extract data. You can either go to the primary source, and get their data, or go the HapMap project, and extract data consistent with these regions. If you extract from HapMap, you need to explain your procedure. If you go to the primary source, you should cite it. Finally, each data-set must be numbered as in the Sabeti paper. Ex: Table S1, Number 1. Credit will be given for the number of data-sets, and the confidence in your assessment of these regions.

2. Summarize the arguments in the following 2 papers:

   ```
   http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0030104
   http://www.springerlink.com/content/e803426681664g43/
   ```

   Specifically, for the second paper, describe the (a) data, (b) methods used, and (c) conclusions. Comment out the differences between the Y-chromosome, and mtDNA phylogeny. Finally, describe your own approach to detecting, and placing recurrent mutations.

3. A "structural polymorphism" is a large scale genetic event (insertion, deletion, or inversion) that is 'polymorphic in a population. Review the following papers on structural polymorphisms. Your review must summarize computational techniques used to detect such polymorphisms. Review additional literature and/or do some computational tests to answer the following: Is there a genomic signal (low copy repeats/motifs) that can cause such polymorphisms to occur?

   ```
   1. Nature Genetics 38, 86 - 92 (2006).

   Common deletion polymorphisms in the human genome

   Steven A McCarroll et al.

   2. A high-resolution survey of deletion polymorphism in the human genome
   Nature Genetics 38, pp75 - 81 (2006)

   -Donald F Conrad, T Daniel Andrews, Nigel P Carter, Matthew
   ```

E Hurles and Jonathan K Pritchard

3. Fine-scale structural variation of the human genome.
E Tuzun, AJ Sharp, JA Bailey, R Kaul, VA Morrison,
Nature Genetics 37, 727 - 732 (2005)