CSCI2951-N: Advanced Algorithms in Computational Biology Homework 2

Due: 11:59PM Tuesday October 21, 2014

Please handin your submission by emailing it to sorin@cs.brown.edu with subject "csci2951-N HW2 handin"

The Homework contains problems worth 120 points; i.e. there are 20 extra points.

Problem Reading

The following papers will help answering problems in this homework: *Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies* and *Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium*. These are the papers introducing the two SNP tagging algorithms presented in class: the LD-select algorithm and the Informativeness algorithm. Read the HapMap tutorial and visit the HapMap database, both links available on the Resources/class webpage.

Problem Tagging SNPs and the minimum informative set of SNPs. (20 points)

Definitions. Let M be an $m \times n$ haplotype matrix, of n individuals and m SNPs. Each individual (row) corresponds to one haplotype: $h_i \in \{0,1\}^n$, corresponds to the i^{th} haplotype, $1 \leq i \leq n$. Each column corresponds to a SNP: $s_j \in \{0,1\}^m$, corresponds to the j^{th} SNP, $1 \leq j \leq m$.

A biallelic SNP distinguishes two individuals namely the ones that have one allele from the ones having the other allele. Let $DIST(s)$, be the set of all pairs of haplotypes that have different alleles at SNP s. For example: Given the 3×4 haplotype matrix M

	s_1	s_2	s_3	s_4
h_1	0	0	1	0
h_2	1	0	0	1
h_3	1	1	0	1

$DIST(s_1) = \{(h_1, h_2), (h_1, h_3)\}$

Let S be the set of SNPs of M. For a set of SNPs, S' let $DIST(S_0)$ be the set of pairs of haplotypes of M distinguished by at least one SNP in S'. For example, $DIST(\{s_1, s_2\}) = \{(h_1, h_2), (h_1, h_3), (h_2, h_3)\}$. A set of SNPs S' *predicts* a SNP s, $s \notin S'$, if $DIST(s) \subseteq DIST(S')$.

A *minimum informative subset* of SNPs $S' \subseteq S$ is a set of SNPs such that every SNP of S is predicted by S' and S' has the minimum size among all such subsets. That is, $DIST(S') = DIST(S)$ and S'is of minimum size.

Problem. Given the following haplotype matrices, find a minimum informative subset of SNPs for each of them. M_3 is a subset of SNPs and CEU individuals from the HapMap HLA-DRA gene.

$$M_1 = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline h_1 & 0 & 0 & 1 & 0 \\ h_2 & 1 & 0 & 0 & 1 \\ h_3 & 1 & 1 & 0 & 1 \end{array}$$

$$M_2 = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline h_1 & 0 & 1 & 1 & 1 \\ h_2 & 1 & 0 & 0 & 1 \\ h_3 & 1 & 0 & 0 & 1 \end{array}$$

$$M_3 = \begin{array}{c|cccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline h_1 & 1 & 1 & 1 & 1 & 1 & 1 \\ h_2 & 0 & 1 & 1 & 1 & 0 & 1 \\ h_3 & 0 & 1 & 0 & 1 & 0 & 1 \\ h_4 & 0 & 1 & 0 & 1 & 0 & 1 \\ h_5 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Problem Expectation Maximization Haplotype Frequencies and Haplotype Phasing Algorithm (60 points)

Implement the Expectation Maximization algorithm presented in class. The algorithm should take as input a text file with a genotype on each line and output a text file with two haplotypes for each input genotype. Line i of the input should correspond to lines $(i * 2) - 1$ and $i * 2$ of the output. Your implementation should also take as input the number of iterations and output the haplotype frequencies. You may set the initial haplotype frequencies or make it a configurable parameter.

You may implement the algorithm in any language as long as it can be run from departmental machines. Include a README with instruction on how to run the program and also include the source code in the submission.

Problem EM Phasing (20 points)

Phase the following genotypes using the EM method. Start with the initial guess that all haplotype frequencies are equally likely. Show the intermediate haplotype frequencies and counts.

$$\begin{array}{cccc} g_1 & 0 & 2 & 0 \\ g_2 & 1 & 2 & 0 \\ g_3 & 1 & 2 & 2 \end{array}$$

Problem Fermat Urns Problem (20 extra points)

Two urns contain the same total number of balls, some blacks and some whites in each. From each urn are drawn n ($n \geq 3$) balls with replacement. Find the number of drawings and the composition of the two urns so that the probability that all white balls are drawn from the first urn is equal to the probabilty that the drawing from the second is either all whites or all blacks.