CSCI2951-N: Advanced Algorithms in Computational Biology Homework 1

Due: 11:59PM Tuesday September 30, 2014

Please handin your submission by emailing it to sorin@cs.brown.edu with subject "csci2951-N HW1 handin"

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

Problem 1: Linkage Disequilibrium and a major drawback for all pairwise LD measures (20%)

All the LD measures used in the literature (see paper on comparing LD measures) are binary and depend on the allele frequencies of the two SNPs they are applied to. It is concluded, therefore, that measuring LD is a non-quantitiative task. Therefore, it is hard to compare the LD in two regions of a genome, or two regions on two different genomes, as each measurement could have a very specific bias that can confuse the analysis. The Slatkin survey (on the calss webpage) presents a variety of applications of LD to biological and medical problems. Present two such problems/applications where using LD measures with the above drawback would create analytical difficulties in solving those problems.

Problem 2: Clark Method for Haplotype Phasing (60%)

A warming up phasing problem

Phase the following set of genotypes using the Clark method. What can you say about different orders of applying chains of Clark rules and the corresponding phased solutions?

g_1	1	0	0	1	1
g_2	1	2	0	2	1
g_3	0	2	1	1	1
g_4	1	2	1	0	2
g_5	1	0	1	2	0
g_6	0	1	0	0	0

Coding the Clark Method

Code the Clark phasing algorithm by adding your rules to the parts of the algorithm method that are left undefined. You may modify this algorithm however you see fit, but be sure to document and describe the reasons for the decisions you made. For instance, if you want to always

produce a valid phasing, you may have to add new rules or routines to phase the orphans. E.g., use graph theory to model the phasing chains as graph theory constructs and use heuristics to get closer to the objective function of minimiing the number of genotype orphans (unphased at the end of the running of your algorithm). Run your phasing algorithm and compute the statistics of the inferred haplotypes in the sample. Use dataset1 for your algorithm and display your phased solution.

Design an algorithm for pure parsimony phasing

Clark's algorithm defines a sense of parsimony of "explanation". However, the pure parsimony objective function is simply to minimize the number of haplotypes that explain all genotypes. You can infer those haplotypes anyway you want, not necesarily by Clark rules. Implement the algorithm and compare results with your Clark Method algorithm.

Problem 3: Sir Ronald Fisher's "The Lady Tasting Tea" (40%)

As we will be studying a number of test statistics for statistical hypothesis testing and the caveats associated with them, a good warming up puzlle is the Fisher's puzzle. How would you attempt to solve this famous puzzle? The text of the puzzle is on the class webpage.