CSCI2820: Medical Bioinformatics Homework 4

Problems 1-2 are due on November 27, 2013. Problem 3 is due on December 4, 2013.

Please handin your submission of Problems 1 and 2 by emailing it to Derek_Aguiar@brown.edu with subject “csci2820 hw4 handin” or leaving it in my mailbox in the mailroom on the fourth floor by November 27. Problem 3 is due by email with subject “csci2820 final 1pager handin” or in my mailbox by December 4.

Problem 1: Minichiello-Durbin Algorithm and Ancestral Recombination Graph Reconstruction

Three individuals, changing objectives (20%)

Your task is to find the Ancestral Recombination Graph for the graph with leaf node haplotypes 101010, 110100, and 000101 and ancestral root node 000000 (Figure 1). You can create a new node by using either a mutation operation or a recombination operation and combine two or more nodes with a coalescence operation. The cost of a mutation event is 1; the cost of a recombination is 1; the cost of a coalescence is 0. Our objective is to minimize the sum of mutations and recombinations:

\[ \text{minimize} (\text{mutations} + \text{recombinations}) \]  (1)

You will be graded based on how close your tree objective values are to the true minimum.

How does the tree change if the cost of mutation is 2 and the cost of recombination is 1? What is the minimum object value in this case?

Marginal Trees (20%)

Compute the ARG for Figure 2 using the Minichiello-Durbin algorithm. Compute the marginal trees for the ARG and the mutation that maximally segregates cases (red nodes) from controls (blue nodes).

Problem 2: Lost Alleles (20%)

Given a population of 100 diploid humans that follows the Wright-Fisher assumptions we presented in class (non-overlapping generations, random mating, sampling with replacement, constant population size, etc.) what is the probability that an allele present in one copy will not be present in the next generation? What is the probability if the allele is present on two chromosomes? Give a general formula given \( k \) copies of the allele.
Figure 1: Three haplotypes and an ancestral 000000 haplotype.

Figure 2: Red nodes denote cases and blue nodes denote controls.
Problem 3: Final project 1-pager (40%) DUE DATE: December 4

Having covered most of the GWAS topics, we will start to focus on the final project (which counts for the final exam). This project should be interesting to you (ideally related to your own research) and also be related to the course material. A list of sample project topics related to what was covered in class is given in [http://cs.brown.edu/courses/csci2820/projects.html](http://cs.brown.edu/courses/csci2820/projects.html) but you are welcome to create your own project.

The Final Project for computational students will consist of three parts

1. ∼10 minute presentation

2. a paper describing your project including an abstract, introduction, body (this may include methods, algorithms, results, discussion), and references; this paper should be at most 5 pages

3. software/algorithms and documentation; as always, extra credit (and Pastiche pie!) will be given for Mathematica implementations

The Final Project for biological students will consist of a presentation and paper with no computational assignment. The workload will be comparable to the computational Final Project – please discuss your project directly with Sorin.

Write a project summary (1 page or less). In this summary you should

1. describe the main ideas

2. describe the hypothesis you are testing and/or algorithm you are implementing

3. commit to a deliverable that you will hand-in prior to the final class due date which is December 21