HW3: Ancestral Recombination Graphs

CS 182/282 Spring 2024

Released: Thursday, April 4th, 2024

Due: 11:59pm on Thursday, April 11th, 2024

Overview

Computational biology leverages the statistical power of large-scale datasets to obtain valuable insights regarding genetic determiners of specific traits. In this HW, you will explore the development, theory and inference of ARGs, which can reveal subtle patterns of inheritance across genetically-diverse populations.

This assignment is worth a total of 50 points.

Reading

- Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs (Minichiello and Durbin, 2006)

Handin

Submit your answers to the following problems as a PDF on Gradescope. You may include images of hand-drawn diagrams if necessary, but all written responses must be typed up. Do not include any identifying information on your handin.

P1: Reading Questions (14 points)

For all reading questions, answers must be given in your own words. Any answer which relies too heavily on the direct wording from the readings will receive point deductions.

Read through Minichiello and Durbin linked above and answer the following questions:

1. Refer to heuristic 1 described on the 3rd page of the paper. Why do you think favoring mutation or coalescence events over recombination events help us to construct more plausible ARGs?
2. Recombination breakpoints are set at the ends of *shared tracts* between pairs of haplotypes in order to explain their derivation from a common ancestor. Why is it generally preferable to first set these breakpoints at the ends of longer shared tracts over shorter ones when working backward in time to construct an ARG?
3. Another related problem of interest is *haplotype phasing*, which involves attributing complementary SNPs to specific parental chromosomes within an individual. How can this algorithm be extended to handle "unphased" data from diploid chromosomes?
4. Fine-scale disease mapping is made possible through powerful tools which can simulate many possible ARGs to explain a given dataset. How is statistical inference applied to determine whether a given SNP is likely to be associated with a particular trait?
5. Discuss the medical applications of this algorithm. What kinds of insights can analysis of ARGs yield?

P2: ARG Reconstruction (20 points)

Consider the following haplotype dataset, in which each haplotype string represents SNPs at the same three markers along a chromosome and disease status is indicated:

Haplotype	Status
010	affected
100	affected
011	unaffected
101	unaffected
111	unaffected

Construct an ARG which follows the methodology of the Minichiello-Durbin algorithm (see the *ARG Inference Algorithm* section of their paper for details). You may find it helpful to begin with all five haplotypes on the lowest level of the 'tree' and gradually work upwards (backward in time) to simplify, assigning coalescences, mutations and recombination events as necessary. Your final graph should present a series of branching events from a single common ancestor which results in all five of the haplotypes observed, with mutation and recombination events notated as in page 2 of the paper above.

As you construct your ARG, note the following assumptions made by the original algorithm:

- Attempt **coalescences** only when two haplotypes are *equal*; that is, at every marker, they either match (both are 0 or both are 1), or one allele is • (unknown)
- Attempt **mutations** only to resolve a *single* haplotype which differs from *all other* haplotypes at the marker of interest.
- Attempt **recombination events** only when no coalescences or mutations are possible, and always seek to resolve shared tracts of *maximal length* between haplotypes

You should seek the **most parsimonious** (minimal) ARG you can construct; that is, a tree which fully explains the observed haplotype data with the fewest recombination and mutation events (this also implies the fewest total events along a single lineage). **Your ARG should contain only one recombination event**. Also, in this assignment, any parsimonious ARG should have a common ancestor which is represented in the contemporary population. If there are multiple minimal ARGs which satisfy these conditions, any one of them will receive full credit.

Submit your maximally-parsimonious ARG, along with answers to the following questions:

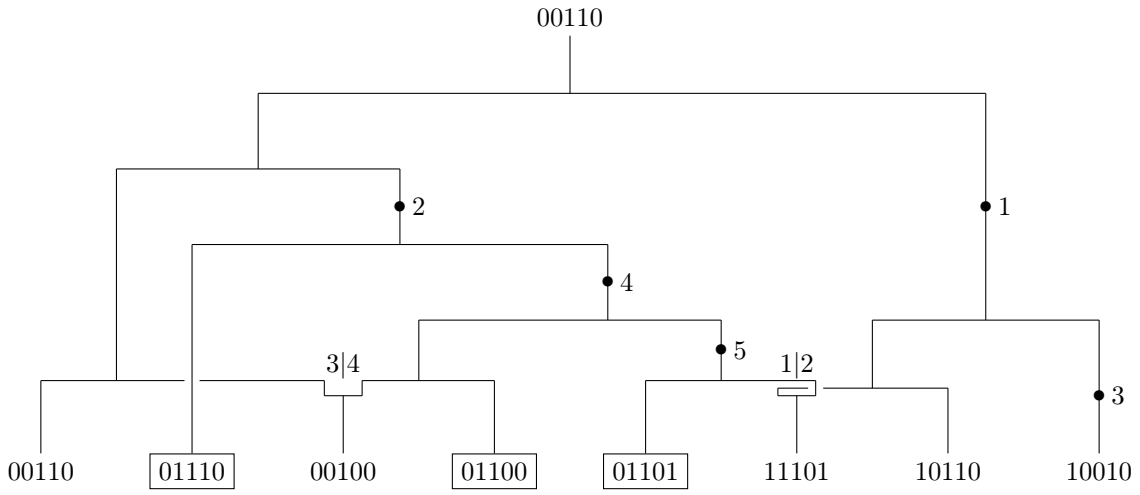
1. What is the haplotype of the common ancestor?
2. How many mutations occurred in this geneological tree?
3. Between which markers did each recombination event occur? What were the implied parental haplotypes for each such event (recall that • denotes an unknown allele at a specific marker)?
4. Which marker appears to be responsible for this disease? Briefly justify your answer.

Bonus (5 points): Sketch all possible minimal ARGs which could explain the haplotypes observed. Provide an example of additional data which, if known, could support the validity of a particular (arbitrary) ARG over others. Describe how this information could aid in selecting the most plausible ARG.

P3: Marginal Trees and Fine Mapping Disease (16 points)

For each locus (SNP) or set of loci within a haplotype, a *marginal tree* can be extracted from the full ARG. This subtree only contains information about mutation and recombination events that are relevant to the locus or loci of interest. See Figure 1 in Minichiello & Durbin's paper above for examples of marginal trees extracted from an ARG.

Consider the full ARG below, representing the geneological history of eight haplotypes. Boxed haplotypes indicate the presence of a disease trait; fork notation indicates the location of breakpoints during a recombination event.



Construct marginal trees for the following SNPs, adhering to the example format given in the paper:

1. Locus 1
2. Locus 2
3. Loci 4 and 5

Now we will utilize the marginal trees to map the disease trait along the chromosome. Assume that this disease is *monogenic*; that is, it can be attributed to a single gene (locus).

4. Which locus or loci are most likely responsible for this disease? Explain your inference.
5. Notice that the affected/unaffected haplotypes do not segregate perfectly based on a single allele. Provide a possible reason for this observation.

   Hint: You may find the concept of epistasis helpful.