Homework 2: Minichiello-Durbin Algorithm for Recombination & Fine Disease Mapping

CS1820 Spring 2018

Out: March 22, 2018
Due: 10:00 pm, April 4, 2018

Overview

In this assignment, you will gain experience with the Minichiello-Durbin Algorithm, which is used to construct Ancestral Recombination Graphs (ARGs) from large populations of haplotype data for recombination and disease mapping analysis. This assignment is worth a total of 50 points.

Suggested Reading

The full algorithm is detailed in the original paper by Mark J. Minichiello and Richard Durbin: Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs (2006)

An outline of this paper can be found on the course website (TODO), but we suggest reading the full paper for a deeper understanding of the methodology and applications of this algorithm.

Handin Instructions

You should hand in your answers to the following three parts (handwritten or typed) to the CS1820 bin. Do not include any identifying information other than your Banner ID.

Part 1: Conceptual Questions (10 points)

1. How many distinct haplotypes are possible in sampling m markers from the SNP alphabet \{0, 1, \_\} (where \_ denotes any unknown allele)? (1)

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 preferred to first set these breakpoints at the ends of longer shared tracts over shorter ones when working backwards in time to construct an ARG? (2)

3. How is the principle of maximum parsimony applied to the construction of an ARG? (2)

4. Discuss the medical applications of this algorithm. In other words, what are ARGs useful for? Once constructed, how are they used? (5)
Part 2: Constructing a Simple ARG (20 points)

Since implementing the full Minichiello-Durbin algorithm would be prohibitively time-intensive, the following example problem can be solved manually (using the principles of the actual algorithm).

Goal

Use the following haplotype data (each haplotype string represents SNPs at four different markers along a chromosome) to construct an ARG which sufficiently explains the genetic variation and pattern of disease presentation observed.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>0100</td>
<td>unaffected</td>
</tr>
<tr>
<td>0101</td>
<td>affected</td>
</tr>
<tr>
<td>0110</td>
<td>unaffected</td>
</tr>
<tr>
<td>1110</td>
<td>unaffected</td>
</tr>
<tr>
<td>1111</td>
<td>affected</td>
</tr>
</tbody>
</table>

Sketch an ARG which follows the methodology of the Minichiello-Durbin algorithm (see the ARG Inference Algorithm section of the paper linked above for details). You will probably find it easiest to begin with all five haplotypes on the lowest level of the ‘tree’ and gradually work upwards (backwards in time) to simplify, assigning coalescences, mutations and recombination events as necessary. Your final graph should be a series of branching events from a single common ancestor which results in all five of the haplotypes observed, with mutation and recombination events noted as in page 2 of the paper linked 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, and only if it can then coalesce with an existing haplotype*
- Attempt recombination events only when no coalescences or mutations are possible, and always seek to resolve shared tracts of maximal length between haplotypes

*This last condition is not part of the original algorithm, but is included here to simplify this application

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). For the sake of this application, consider mutation events to be more parsimonious than recombination events. If there are multiple minimal ARGs which satisfy these conditions, any one of them will receive full credit.

Submission

Turn in an ARG sketch as described above, along with answers to the following questions:

1. What is the haplotype of the common ancestor?
2. How many mutations occurred in this genealogical tree?
3. For all recombination events: between which markers did they occur, and what were the presumed parental haplotypes (recall that • denotes an unknown allele at a specific marker)?
4. Which marker appears to be responsible for this disease? Briefly explain why.
Part 3: Extracting Marginal Trees From an ARG (20 points)

For each locus (position) on the chromosome, there is a genealogical tree, called a marginal tree, embedded in the main ARG. This subtree only contains information about mutation and recombination events that are relevant to a specific locus or subsequence of SNPs. Figure 1 in the Minichiello-Durbin paper linked above has an example of a marginal tree extracted from an ARG.

Goal

Use the following haplotype data and completed ARG to create marginal trees, and attempt to identify disease-causing mutations and recombinations. In the ARG, the shaded sequences refer to affected individuals, the black circles reference mutations, and the numbers around a fork represent a recombination event between the left and right parental haplotypes between the markers at each of those loci.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>00101</td>
<td>affected</td>
</tr>
<tr>
<td>00110</td>
<td>unaffected</td>
</tr>
<tr>
<td>10000</td>
<td>unaffected</td>
</tr>
<tr>
<td>10001</td>
<td>affected</td>
</tr>
<tr>
<td>10011</td>
<td>affected</td>
</tr>
<tr>
<td>11111</td>
<td>unaffected</td>
</tr>
</tbody>
</table>

Submission

Create marginal trees (in a similar format to the paper, and the ARG above) for the following SNP subsequences (1-indexed). These may be hand-drawn or computer generated.

1. Positions 1-3
2. Position 4
3. Position 5

Given these trees, are you able to isolate a disease-causing mutation/recombination? If so, what is it? Answer this question in your submission.