CSCI2951 HW1 | Ben Siranosian | 2014-09-30

Problem 1: Linkage Disequilibrium

One major drawback of measures of LD is that they can be biased when comparing two regions on the same genome or regions between genomes. LD can be applied to solve many problems in biology and medicine, but the biases can make interpretation of the results difficult. For example, LD can be used to estimate the evolutionary age of an allele by comparing with nearby regions. However, comparing regions invokes the biases described earlier. This could lead to an incorrect estimate of the age of the allele.

A second application of LD is to map disease causing loci to a certain region on a chromosome. This requires comparing the LD of regions surrounding the probable disease loci. Once again, the biases indicated earlier could lead to an incorrect estimate of the allele location.

Problem 2: Haplotype Phasing

2A: Warmup

A total of 8 haplotypes are needed to explain the genotypes. They are:

h1	10011	h5	10110
h2	00111	h6	01000
h3	01111	h7	11001
h4	10100	h8	11101

In this case, the haplotype phasing is unique. Changing the order of the input does not change the result of the phasing. Every genotype is explained by a unique set of two haplotypes. In more complicated cases, though, this might not be true. One input order may produce different results than another input order. Additionally, multiple resolved haplotypes could be used to explain a given genotype. This creates another type of ambiguity that is not seen in the simple example.

2B: Coding the Clark Method

I implemented the Clark method of phasing in a python script (clark_phasing.py). This script can be used in the following way:

```
USAGE: python clark_phasing.py genotype_file
Takes as input a list of genotypes (defined by 0,1,2), one per line.
Computes the clark phasing of the input with rules defined in the
accompanying document.By default runs 1000 iterations of the algorithm
and picks the best solution (fewest orphans and fewest explaining
haplotypes).
```

I made some decisions to make the phasing problem simpler. First, I remove duplicate genotypes from the input data. Genotype frequencies could be used by a more advanced algorithm to make better decisions about haplotypes, possibly by recording the number of times a haplotype explanation is used and making frequent haplotypes more likely. Second, I remove duplicates from the list of resolved haplotypes, for the same reason as above.

The general outline of my algorithm is as follows:

1) Read input data from file. Remove duplicates from the set of genotypes. Randomize genotypes remaining.
2) Resolve homozygote genotypes (those with 0 ambiguous sites)
   a) Add the resolved haplotype to a set of possible explanations
3) Resolve heterozygote genotypes (those with 1 ambiguous site)
   a) Add the resolved haplotypes to a set of possible explanations
4) Look at the next genotype in the list. If it can be explained by a haplotype in the resolved set, use the first possible explanation to phase it.
   a) Add the complementary haplotype to the set of possible explanations
   b) If the genotype cannot be explained, pass on to the next genotype and hope it can be explained in a future iteration.
5) Iterate the above step until the number of genotypes explained does no longer decrease. The remaining genotypes are orphans for this iteration.

I repeat steps 4-6 several (1000) times and record the result from each iteration. I then only consider the results in which the smallest number of orphans are left behind. Within this subset, I choose the result with the smallest number of explanations as my solution. If multiple solutions exist with the smallest number of explanations, one of them is chosen at random. This already answers the third part of this question, Parsimony Phasing. By running the algorithm many times and picking a solution with the fewest number of explaining haplotypes, we are picking the most parsimonious solution. While the most parsimonious solution to the provided genotypes contained 10 haplotypes, I occasionally observed solutions with more haplotypes (up to 14).

The result from running my algorithm on the provided data follows:

```
Number of orphans remaining: 0
Number of haplotypes explaining resolved genotypes: 10
Haplotypes explaining genotypes
        [0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1]
        [1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0]
        [1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1]
        [1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1]
        [1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1]
        [1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1]
        [1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1]
        [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
        [1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1]
        [1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

Problem 3: The Lady Tasting Tea

An excellent way to solve Sir Ronald Fisher's problem would be to use the test named after the statistician himself: Fisher's exact test. This test can provide an exact answer to the question, "what is the probability that the lady's arrangement of tea into two categories is the result to random chance?" We can then fix a threshold for this probability, below which we will believe in the lady's abilities, and interpret the result of the experiment in light of these statistics. Finding *P* to be below this threshold will

be equivalent to rejecting the null hypothesis that the lady grouped the tea in such an ordering by random chance alone. Fisher's exact test is appropriate for this experiment because it can be modeled as a 2x2 contingency table where the margins are fixed. If values in the table are listed as *a, b, c, d* and the sum of the values in the table is *n*, the probability of obtaining the given values under the hypergeometric distribution is:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

If we set a threshold of $P < 0.05$ for believing in the lady's abilities, she must get every cup correct. See the contingency table below for an explanation.

All guesses are correct

	Guess 1	Guess 2	Totals
True 1	4	0	4
True 2	0	4	4
Totals	4	4	8

$$P = \frac{\binom{4+0}{4}\binom{0+4}{0}}{\binom{8}{4+0}} = 0.014$$

If the lady guesses one cup incorrect, we will not believe in her abilities:

3 guesses correct

	Guess 1	Guess 2	Totals
True 1	3	1	4
True 2	1	3	4
Totals	4	4	8

$$P = \frac{\binom{3+1}{3}\binom{1+3}{3}}{\binom{8}{3+1}} = 0.224$$

If more statistical power is required (*P* must be smaller for us to believe in her abilities) a larger sample size could be used.