Algorithm

Our haplotype phasing algorithm uses expectation maximization to determine the appropriate pair of haplotypes corresponding to a genotype. The algorithm considers the genomes of the individuals in a small window, so that the possible haplotypes are easier to manage. For each window, expectation-maximization is performed on the possible haplotypes corresponding to the window and a consensus is taken from the results of the different views of each SNP.

Implementation

There are two files, phasing.py and em.py. The phasing.py file handles manipulation of the input file, determination of all possible haplotypes for a given genotype, and all manipulations of this list for use in EM. The em.py file contains all parts relating to the expectation and maximization steps, and the main function for analysis.

The file em.py contains a function called em(), which takes the command-line arguments to determine which parameters and input file to use. The em() function first calls functions from the phasing.py file to create a list of haplotype pairs from the genome data. It loads the file into a list of lists and then converts this data into a list of genotypes. A list of all possible haplotypes from these genotypes is generated. Then this list is converted into a list of haplotype pairs. We end up with a list of individuals, each of which has a list of haplotype pairs. Then we make another list of haplotypes that removes all the duplicates. We use this to make a dictionary that keeps track of frequency of each haplotype.

The initial frequency of each haplotype is set to 1/(# number of haplotypes). Then the expectation-maximization is iterated for a specified number of runs. The em_step() first finds the probability of each pair, using the haplotype frequencies. Then, it uses those probabilities to find the overall frequency of each haplotype.

After the EM algorithm has run for its allotted number of runs, best() finds the best haplotypes. Finally, the consensus for each SNP is taken from the best haplotypes for each window, and is printed to file using output().