Our haplotype phasing algorithm uses expectation maximization to determine the appropriate pair of haplotypes corresponding to a genotype. 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. The algorithm splits the genome into several small pieces that are more easily manageable. For each piece the algorithm is performed and the pieces are stitched back together at the end.

The file em.py contains a function called em that 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 algorithm is run for a specified number of runs. The expectation step first finds the frequency of each pair. .................................... not sure what you changed

After the expectation algorithm has run for its allotted number of runs the maximization step finds the best haplotypes.

The rest is stuff you added...