Using an Expectation Maximization Algorithm to Estimate Allele Frequencies of Cystic Fibrosis

By Rachel Ferina

Mentor: Seth Temple

Cystic fibrosis is a recessive genetic disease that affects the lungs and digestive system. This quarter I implemented an Expectation Maximization (EM) Algorithm to estimate population allele frequencies based on a random sample of 200 subjects with 12 cases diagnosed with cystic fibrosis. In this sample, the genotypes of the subjects diagnosed with cystic fibrosis are known because they required both recessive alleles. However, the genotypes of the unaffected subjects are unknown, as they could be homozygous or heterozygous dominant. EM algorithms are useful for calculations like this with missing or incomplete data. The algorithm consists of two steps. The E step stands for expectation and involves computing the expected genotype based on the observed phenotype. The M step is the maximization step which determines the allele frequencies that maximize a likelihood equation given the expected phenotypes. The EM algorithm switches between the E step and M step until convergence of the likelihood occurs. Likelihood is a product of probability and for the EM algorithm it increases with each iteration. To compute the log likelihood, I derived an equation where the counts are multiplied by log frequencies. The Hardy-Weinberg equilibrium allele frequencies are assumed for the allele frequencies, as the equilibrium principle assumes the allele frequencies are not changing and the population is not evolving, which is ideal for the sample. To account for the number of dominant alleles, I combined the Hardy-Weinberg allele frequencies for both the heterozygous and homozygous dominant cases in one log. I coded the log likelihood equation in RStudio along with a while loop to update the likelihoods.

Incomplete Log Likelihood: $n_A\log(p^2 + 2pq) + n_a\log(q^2)$

The E step of the algorithm enables the calculation of the current allele frequency estimates. It also splits up the unaffected group and predicted the number of heterozygous and homozygous dominant subjects out of the 188 subjects that do not have Cystic Fibrosis. The M step estimates new allele frequencies that are conditional on the current allele frequency estimates and updates frequencies via gene counting, which is essentially counting the copies of a gene present in a population. The EM algorithm converged after 15 iterations, and I found the frequency of the recessive allele in the sample to be 0.245, and the frequency of the dominant allele to be 0.755.

In addition to this project, this quarter I learned about applications of statistics and heredity. I expanded my understanding of pedigrees by computing kinship and inbreeding coefficients based on Wright's path counting formula. I also learned about identical by descent states, which measure if alleles are copies of the same ancestral gene and indicate relatedness. I gained programming experience with RStudio and visualized segments of identity by descent using the command line hap-ibd program. Overall, I learned how statistics facilitates research in biology and how biology informs statistical models.