Assignment 1 Part 2

September 18, 2020

Shahriyar Mahdi Robbani

XQR418

0.0.1 Question 2

Write the likelihood model that uses both the observed bases and the quality scores for a single site. The frequencies of the four bases are the parameters. Remember that the true bases are not observed. NB! you have to explain all of the notation i.e. variables/parameters you use in your model.

The following likelihood model was used to estimate the frequencies:

$$\ell(f) = \sum_{i}^{N} \log \left(\sum_{j} p(b_i, G_j|f) \right)$$

where f is the base frequency to be estimated,
b_i is the ith base in N reads,
G_j is the true haploid Genotype j where $j \in \{A, C, G, T\}$,
$p(b_i, G_j|f)$ is the probability of a base b_i having the true genotype G_j given a certain frequency of j

Write the Q (estimation) and M step of the EM algorithm that you will need for the optimization. Use the same notation (variables/parameters) as you use to describe the likelihood.

Q Step:

$$Q_i(G_j) = p(G_j|b_i, f^{(n)}) = \frac{p(b_i|G_j, f^{(n)})p(G_j|f^{(n)})}{\sum_j p(b_i|G_j, f^{(n)})p(G_j|f^{(n)})}$$

where
$f^{(n)}$ is the estimated value of the frequency at the nth iteration fo the algorithm
$p(b_i|G_j, f^{(n)})$ is the likelihood of base i given the true genotype G_j and frequency at nth iteration
$p(G_j|f^{(n)})$ is the probability of thetrue genotype G_j occuring given the frequency at nth iteration

M step:

$$f_j^{n+1} = \frac{\sum_i^N p(G_j|b_i, f^{(n)})}{\sum_i^N \sum_j p(G_j|b_i, f^{(n)})}$$

0.0.2 Question 3

How many sites are there where the allele frequency of the most common allele is less than 0.9?

There are no sites where the allele frequency of the most common allele is less than 0.9.