Advanced bioinformatics for next generation sequencing

Handin 1, part 2

Stefano Pellegrini (mlq211)

September 18, 2020

1 Model

1.1 Likelihood model

The likelihood model for each position is the following:

$$\begin{aligned}\log(L(\theta_j)) &= \sum_i \log\left(\sum_j p\left(X_i, G_j \mid \theta_j\right)\right) \\ &= \sum_i \log\left(\sum_j p(X_i \mid G_j, \theta_j)\; p(G_j \mid \theta j)\right) \\ &= \sum_i \log\left(\sum_j p(X_i \mid G_j)\; \theta j\right)\end{aligned} \tag{1}$$

where $\sum_j \theta_j = 1$, and $p(X_i \mid G_j, \theta) = p(X_i \mid G_j)$, and $p\left(G_j \mid \theta_j\right) = \theta_j$.

$G \in \{A, C, G, T\}$ is the true unknown genotype, therefore it is the latent variable. θ is the allele frequency that we want to estimate, and lastly, X represents the observed base of the read i.

The genotype likelihood is defined as:

$$P(X \mid G_j, \theta_j) = P(X \mid G_j) \propto \prod_{i=0}^{n} P\left(X_i \mid G_j\right) \tag{2}$$

where $P\left(X_i \mid G_j\right) = \begin{cases} \frac{\epsilon_i}{3} & X_i \neq G_j \\ 1 - \epsilon_r & X_i = G_j \end{cases}$.

In equation (2), $X_i \in \{A, C, G, T\}$ is the nucleotide of the read i. ϵ is the probability of having a wrong base in the read i, which can be computed as $\epsilon = 10^{\frac{-Q}{10}}$, where Q is the quality score of that base.

1.2 Q and M step

The Q (estimation) step of the EM algorithm is defined as

$$\begin{aligned} q_i\left(G_j\right) &= p\left(G_j \mid X_i, \theta_j^{(n)}\right) \\ &= \frac{p\left(X_i \mid G_j\right)\; \theta_j^{(n)}}{p(X \mid \theta_j^{(n)})} \end{aligned} \tag{3}$$

. q_i is the helping function or the posterior probability of the genotype G_j, given the base (X_i) at position i and the $\theta^{(n)}$, which is the best estimation of the frequency after n steps of the EM algorithm.

The M (maximization) step of the EM algorithm is defined as

$$\theta_j^{(n+1)} = \frac{\sum_i q_i\left(G_j\right)}{\sum_i \sum_j q_i\left(G_j\right)} \tag{4}$$

2 Implementation

Number of sites with most common allele frequency less than 0.9 are 9. Allele frequency for all sites are: A = 0.3089898, C = 0.3128351, G = 0.1312083, T= 0.2469668