Assignment 1: Estimating Variants in Two Yeast Genomes

Abdullah Faqih Al Mubarok - vpx267

September 22, 2023

1 Model

1. Write the likelihood model that uses both the observed bases and the quality scores. The 16 fractions of allele configurations are the parameters.

   Answer:

   The likelihood model can be written as follow:

   $$L(\theta) = p(X|\theta) = \prod_i p(X_i|\theta) \tag{1}$$

   Where $\theta = \theta_{\{A,C,G,T\}^2}$ are all 16 allele combination frequencies between two yeasts and it sums to 1 ($\sum_{\{A,C,G,T\}^2} \theta = 1$). X is a data where X_i is the observed bases of site $i \in \{1, 2, ..., 5000\}$ for the two yeast individual. Here, we assume that the sites are independent of each other.

   Next, we introduce the latent variable B_i which is the 16 possible allele combinations at site i:

   $$\begin{aligned} p(X_i|\theta) &= \sum_{b_1 \in \{A,C,G,T\}} \sum_{b_2 \in \{A,C,G,T\}} p(X_i|B_i = b_1, b_2) p(B_i = b_1, b_2|\theta) \\ &= \sum_{b_1 \in \{A,C,G,T\}} \sum_{b_2 \in \{A,C,G,T\}} p(X_i|B_i = b_1, b_2) \theta_{b_1,b_2} \end{aligned}$$

   where b_j is base for individual $j \in \{1, 2\}$. Furthermore, we assume that the two yeast individuals are independent,

   $$p(X_i|\theta) = \sum_{b_1 \in \{A,C,G,T\}} \sum_{b_2 \in \{A,C,G,T\}} p(X_{i1}|B_{i1} = b_1) p(X_{i2}|B_{i2} = b_2) \theta_{b_1,b_2} \tag{2}$$

   Since the yeast are in haploid, the genotype likelihood of site i of an individual j is defined as follow:

   $$p(X_{ij}|b) = \prod_{d=1}^{D_{ij}} p(b_d|b) \tag{3}$$

where

$$p(b_d|z) = \begin{cases} \frac{\epsilon_d}{3} & b_d \neq b \\ 1-\epsilon_d & b_d = b \end{cases}$$

D_{ij} = number of reads of site j on an individual i

ϵ_d = probability of wrong base at depth d which can be obtained from $10^{\frac{-Q}{10}}$

2. Report the E step) (Q function) and M step of the EM algorithm that you will need for the optimization in order to get the maximum likelihood estimates.

Answer:

The Q function can be written as follow:

$$\begin{aligned} q_i(B = b_1, b_2) &= p(B = b_1, b_{b2}|X_i, \theta^{(n)}) \\ &= \frac{p(X_i|B = b_1, b_2, \theta^n)p(B = b_1, b_2|\theta^{(n)})}{\sum_{b_1'} \sum_{b_2'} p(X_i|B = b_1', b_2', \theta^n)p(B = b_1', b_2'|\theta^{(n)})} \\ &= \frac{p(X_i|B = b_1, b_2)\theta^{(n)}_{b_1,b_2}}{\sum_{b_1'} \sum_{b_2'} p(X_i|B = b_1', b_2')\theta^{(n)}_{b_1',b_2'}} \\ &= \frac{p(X_{i1}|B_{i1} = b_1)p(X_{i2}|B_{i2} = b_2)\theta^{(n)}_{b_1,b_2}}{\sum_{b_1'} \sum_{b_2'} p(X_{i1}|B_{i1} = b_1')p(X_{i2}|B_{i2} = b_2')\theta^{(n)}_{b_1',b_2'}} \end{aligned}$$

Where q_i is the helper function of the allele combination of yeast individual B_j and θ^n is the estimation of the allele combination fraction at step n.

The, the M step can be written as:

$$\theta^{(n+1)}_{b_1,b_2} = \frac{\sum_i q_i(B = b_1, b_2)}{\sum_i \sum_j q_i(B = b_1, b_2)} \tag{4}$$

2 Implementation

1. Estimate 16 fractions of allele configurations and make a barplot of the results.

   Answer:

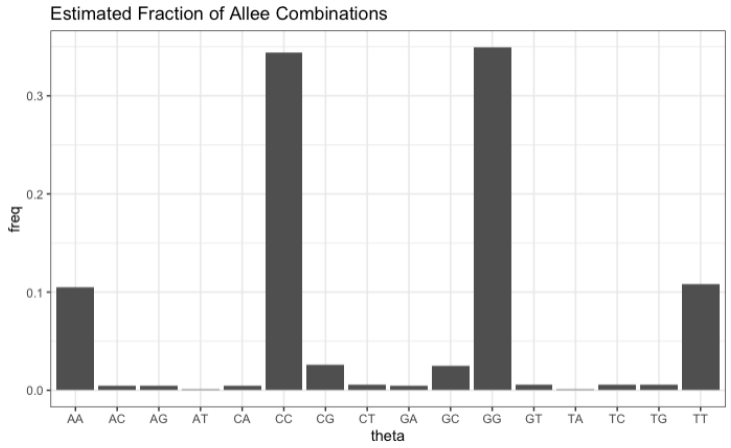

Figure 1: Estimated fractions of allele configurations

2. Based on your results what is the estimated number of sites that are variable (sites when individuals 1 and 2 have different alleles)?

   Answer:

   The Estimated total number of site that are variable is 4662