Assignment 1

Ethan Holleman

October 19, 2021

1 Part 1

1.1 Definitions

- i: Index over nucleotide (input) sequences.
- P: Length of motif to consider in the model.
- j: Indexes over start positions of all possible motifs in a given sequence i of length L assuming all sequences have length L. $0 \leq j \leq L - P + 1$
- m: Indexes of each nucleotide of motif j of sequence i. $0 \leq m \leq P$.
- l: Indexes over models where 1 indicates model corresponding to transcription factor binding motif (foreground) and 0 indicates non-binding region (background).
- k: Indexes over nucleotides (A, T, G, C).
- $X_{i,j,m,k}$: Indicator variable in which $X_{i,j,m,k} = 1$ if base m of the motif beginning at position j of sequence i is equal to nucleotide k and is 0 otherwise.
- C_i: Vector of motif start positions for each sequence i. If $C_i = j$ then then motif j is the transcription factor binding site.

  We use the indicator variable $C_{i,j} = 1$ to represent when $C_i = j$.
- $P(C_{i,j} = l) = \lambda_j$ and $0 \leq \lambda_j \leq 1$.
- $P(X_{i,j,m,k} = l | C_{i,j} = l) = \psi_{k,m}^{(l)}$ where $\sum_k \psi_{k,m}^{(l)} = 1$
- $\boldsymbol{X}$: The set of all $X_{i,j,m,k}$.
- $\boldsymbol{C}$: The set of all C_i.
- The set of all model parameters is denoted as $\boldsymbol{\theta} = \{\lambda_j, \psi_{k,m}^{(l)}\}$

1.2 Complete log likelihood

$$Q(\theta|\boldsymbol{X},\boldsymbol{C}) = \log(P(\boldsymbol{X},\boldsymbol{C}|\theta)) \tag{1}$$

$$= \log \prod_i \prod_j P(X_{i,j}, C_{i,j}|\theta) \tag{2}$$

$$= \log \prod_i \prod_j (\prod_m \prod_k \prod_l [P(X_{i,j,m,k} = 1|C_{i,j} = l, \theta) P(C_{i,j} = l|\theta)]^{X_{i,j,m,k} C_{i,j}}) \tag{3}$$

$$= \log \prod_i \prod_j (\prod_m \prod_k \prod_l [\lambda_j \psi_{k,m}^{(l)}])^{X_{i,j,m,k} C_{i,j}} \tag{4}$$

$$= \sum_i \sum_j \sum_m \sum_k \sum_l X_{i,j,m,k} C_{i,j} \log[\psi_{k,m}^{(l)}] \tag{5}$$

$$= \sum_i \sum_j \sum_m \sum_k \sum_l X_{i,j,m,k} C_{i,j} \log[\psi_{k,m}^{(l)}] \tag{6}$$

$$= \sum_i \sum_j \sum_m \sum_k \sum_l X_{i,j,m,k} C_{i,j} \log \lambda_j + \sum_i \sum_j \sum_m \sum_k \sum_l X_{i,j,m,k} C_{i,j} \log \psi_{k,m}^{(l)} \tag{7}$$