Assignment 1

Ethan Holleman

October 18, 2021

1 Part 1

1.1 Definitions

- i: Index over nucleotide sequences.
- P: Length of kmers to consider in the model.
- m: Indexes over all kmers of a given nucleotide sequence.
- j: Indexes over nucleotides of a given kmer m.
- k: Indexes over nucleotides (A, T, G, C).
- l: Indexes over models where 0 indicates model cooresponding to transcription factor bindind motif (foreground) and 1 indicates non-binding region (background).
- $X_{i,m,j,k}$: indicator variable that equals 1 if nucleotide j of kmer m of sequence i is equal to base k and equals 0 otherwise.
- $C_{i,m,l}$: Indicator variable that will equal 1 if kmer m of sequence i was drawn from model l and otherwise equals 0.
- $\boldsymbol{X}$: The set of all $X_{i,m,j,k}$.
- $\boldsymbol{C}$: The set of all $C_{i,m,l}$.
- $P(C_{i,m,l} = 1) = \lambda_l$ where l is either 0 or 1 indicating foreground or background.
- $P(X_{i,m,j,k} = 1 | C_{i,m,l} = 1) = \psi_{j,k}^{(l)}$, where $\sum_k \psi_{j,k}^{(l)} = 1$

  This should define two 4 by P matries where each row sums to 1 and represents the probibility of observing each nucleotide at a given position j of a kmer given kmer was selected from model l.
- The model parameters are denoted as $\theta = \{\lambda_l, \psi_{j,k}^{(l)}\}$

1.2 Complete log likelihood

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

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

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

$$= \log \prod_i \prod_m (\prod_j \prod_k \prod_l [\lambda_l \psi_{j,k}^{(l)}]) \tag{4}$$

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

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