BIMS 8601 Assignment 4: Scan a DNA sequence for motif matching

Due April 8, 2022

1. Implement motif matching scan: (6 pt)

Given a DNA sequence of length N and the position weight matrix (PWM) of a motif of length w, the goal is to scan the DNA sequence and its *reverse complement* to calculate a motif matching score at every position. The output will be two lists of scores, which can generate two curves representing the motif matching pattern for the sequence and its reverse complement, respectively. For simplicity, the DNA sequence contains A, C, G, T only, i.e., no N.

The motif matching score can be calculated as a log likelihood ratio:

$$S = \log_2 \frac{\Pr(x \text{ from } \theta_m)}{\Pr(x \text{ from } \theta_0)}$$

where

$$\Pr(x \text{ from } \theta) = \prod_{i=1}^{w} p(X_i|\theta)$$

θ_m is the motif PWM; θ_0 is the genome background:
$p_0(\mathrm{A},\ \mathrm{C},\ \mathrm{G},\ \mathrm{T}) = [0.28,\ 0.22,\ 0.22,\ 0.28]$

In practice, as we are only interested in a positive matching score, the final output score can be adjusted as

$$S_{\mathrm{OUTPUT}} = \max(S, 0)$$

2. Run motif matching scan on the 11 provided sequences for the RBPJ motif. Generate plots for the RBPJ motif matching patterns for each sequence. Use different colors for the two curves for original sequence and its reverse complement sequence. (3 pt)
The PWM can be found at:
https://github.com/cphg/compgen/blob/master/resources/motif_assignment/rbpj_pwm.txt
The sequences can be found at:
https://github.com/cphg/compgen/blob/master/resources/motif_assignment/notch1_binding_sequences.fa

3. Discuss any interesting observations you find from the results. (1 pt)