Genome 540 Discussion

January 23rd, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

- Assignment 3
- JASPAR

Assignment 3

Overview

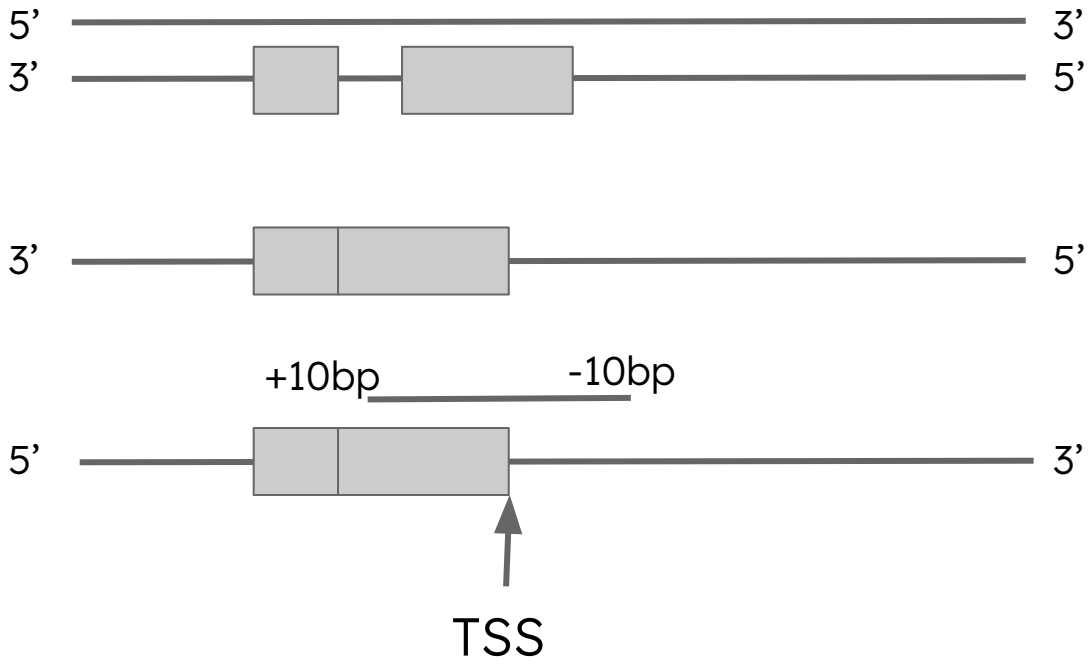
1. Parse a genbank file (.gbff) and…
   a. Extract all CDS features
   b. Read in the sequence
2. Build a site model for translation start sites (TSS)
   a. Use CDS features to get nucleotide frequencies +/- 10bp around all TSS (21bp total including TSS)
   b. Use sequence to get nucleotide frequencies throughout the genome *on both strands*
   c. Compute the weights using the log2 ratios of the frequencies
3. Use the site model to compute scores at
   a. Every annotated TSS
   b. The entire genome (21bp window) on both strands

complement(join(...)) example

Example: complement(join(15..20,25..35))

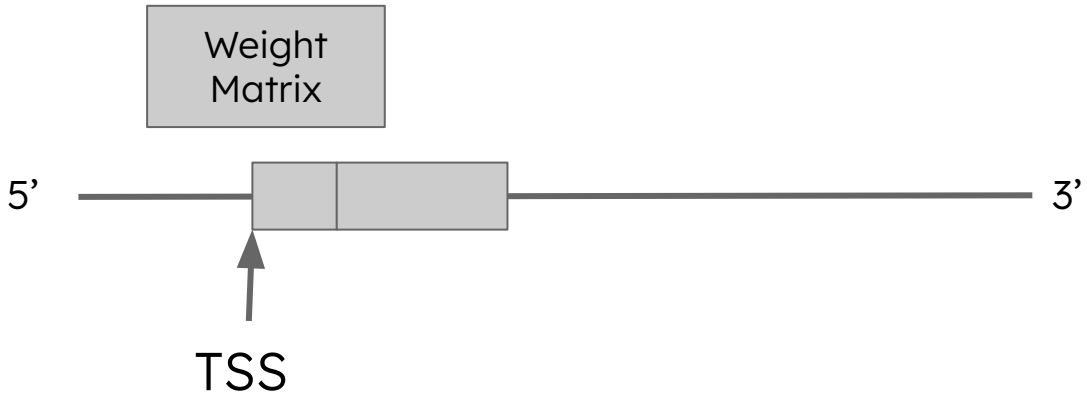
15..20,25..35
- Coordinates on + strand
- But take sequence on reverse complement

join(15..20,25..35)



Question

Does the site model represent genomic DNA or the processed mRNA?

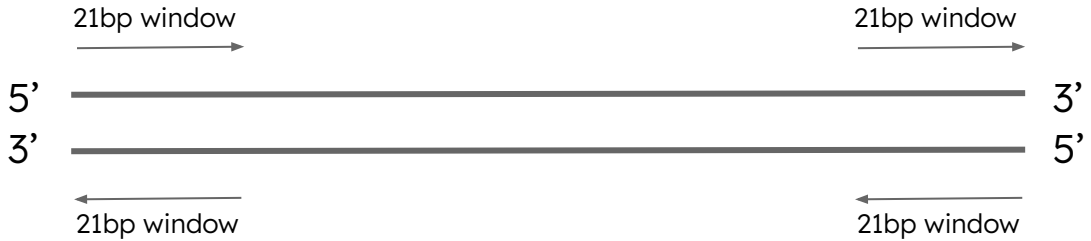


Building the weight matrix

Steps:

1. Compute the background nucleotide frequencies
    a. Forward and reverse strands
2. Count matrix
    a. Compute the nucleotide counts around every TSS
3. Frequency Matrix
    a. Compute the proportion of times a nucleotide occurs at each position
4. Weight matrix
    a. Weight = log2([nt freq at motif position] / [background nt freq])
    b. If a nt has a frequency = 0, assign it a weight of -99.0

Computing site scores

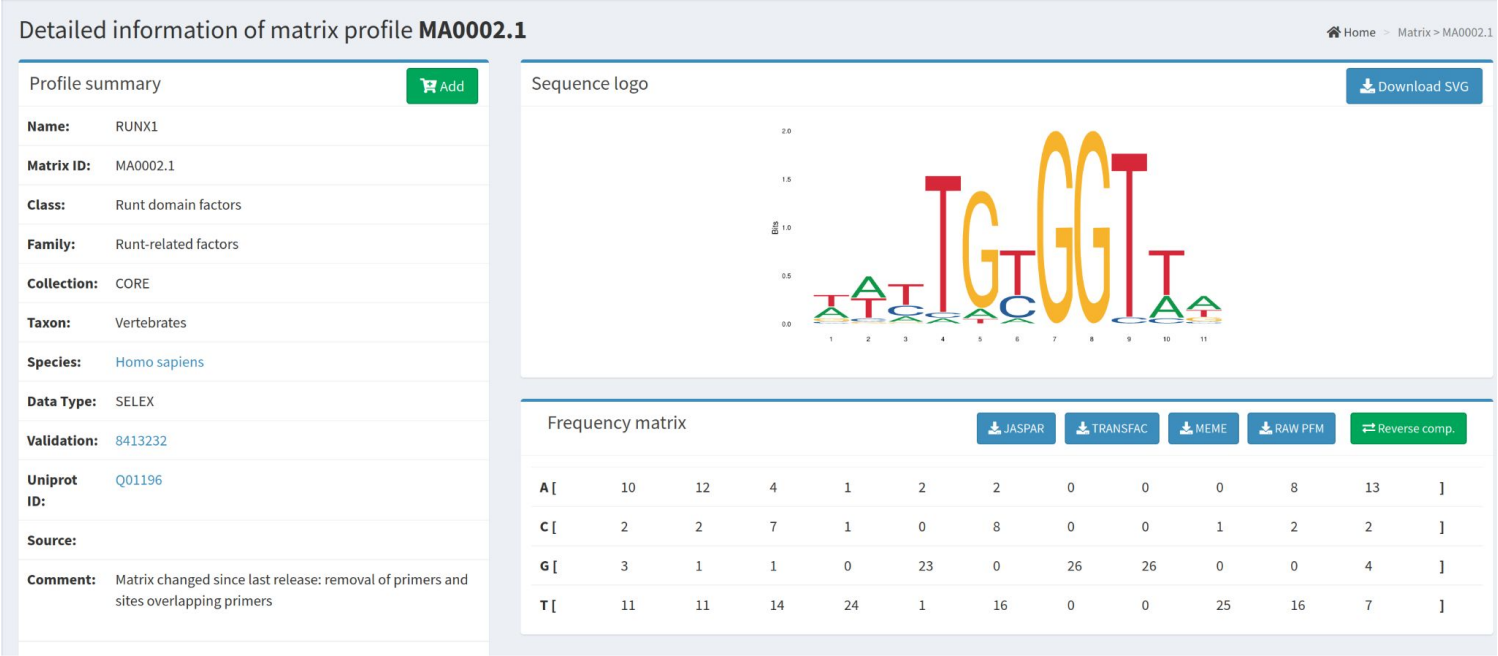


- Use weight matrix to compute site scores at **all** positions in the genome
  - Score = sum of weights for nucleotide present at each position
  - Scores should be associated with motif **centered** on that position
  - Don’t extend window beyond the genome
  - Run on forward and reverse strands

Other things…

- Positions are inclusive (5..10) is 6bp starting at 5 and ending at 10
- Use **double** precision numbers
- Ignore duplicates
- Ignore CDS sequences with “>” and “<” characters
  - If a CDS contains these the position is uncertain and you can skip that CDS

JASPAR



https://jaspar2020.genereg.net/

Reminders

- HW3 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template