

# Genome 540 Discussion

January 23rd, 2024

Clifford Rostomily

# 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

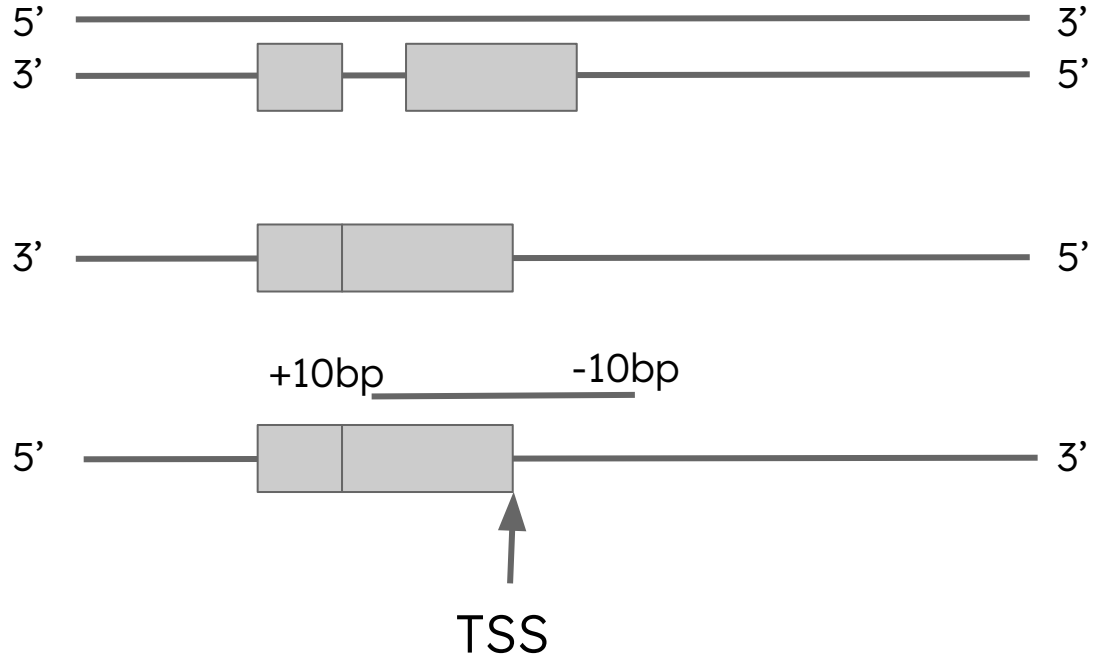
# complement(join(...)) example

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

15..20,25..35

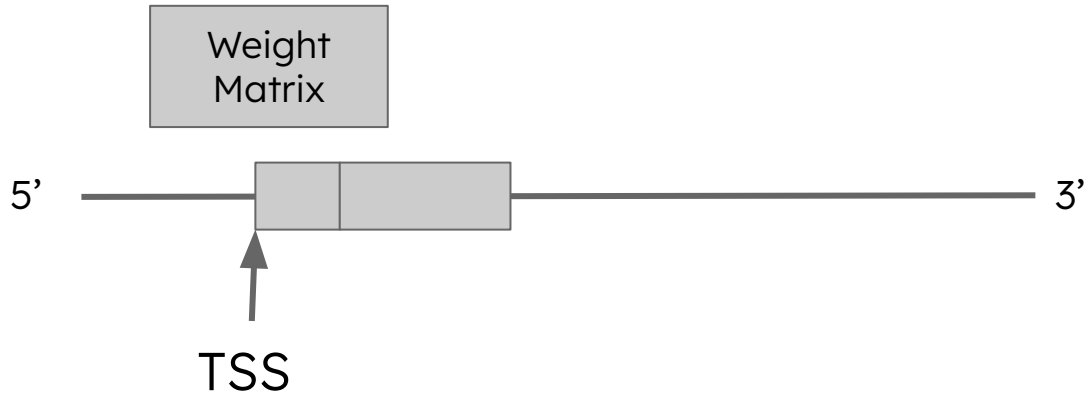
- Coordinates on + strand
- But take sequence on reverse complement

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.  $\text{Weight} = \log_2([\text{nt freq at motif position}] / [\text{background nt freq}])$
  - 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

## Detailed information of matrix profile **MA0002.1**

Home - Matrix - MA0002.1

### Profile summary

Add

**Name:** RUNX1

**Matrix ID:** MA0002.1

**Class:** Runt domain factors

**Family:** Runt-related factors

**Collection:** CORE

**Taxon:** Vertebrates

**Species:** [Homo sapiens](#)

**Data Type:** SELEX

**Validation:** [8413232](#)

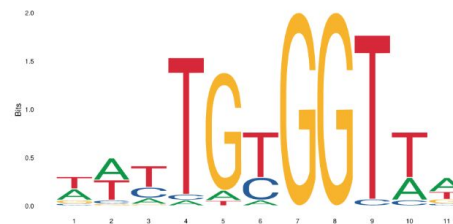
**Uniprot ID:** [Q01196](#)

**Source:**

**Comment:** Matrix changed since last release: removal of primers and sites overlapping primers

### Sequence logo

Download SVG



### Frequency matrix

JASPAR

TRANSFAC

MEME

RAW PFM

Reverse comp.

A	10	12	4	1	2	2	0	0	0	8	13	]
C	2	2	7	1	0	8	0	0	1	2	2	]
G	3	1	1	0	23	0	26	26	0	0	4	]
T	11	11	14	24	1	16	0	0	25	16	7	]

<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