Genome 540 Discussion

January 23rd, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

- Assignment 3
- Assignment 4

Assignment 3

Any questions?

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

Assignment 4

Overview

Part 1: Write a program to find the highest-weight path in a directed acyclic graph using dynamic programming

Part 2: Run your program on a linked list created from DNA sequence

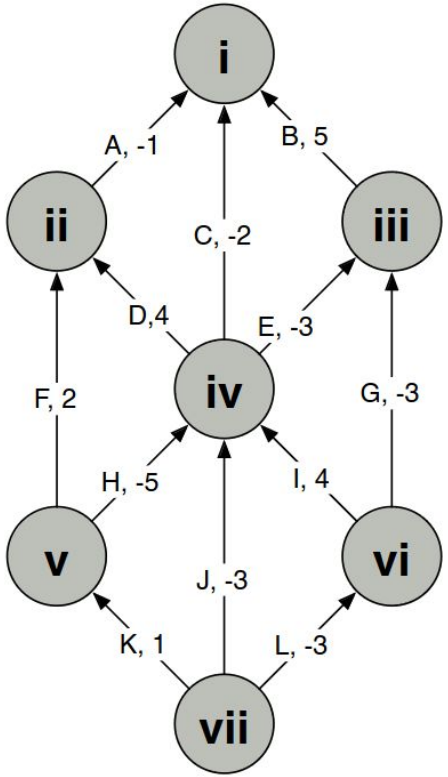
Program 1: Highest weight path

1. Convert graph to text file of **vertices** and **edges** by hand
2. Use dynamic programming to find the max weight path through the graph (Lectures 7/8)
   a. Overall
   b. With constraints (START/END)
3. Output
   a. Path Score
   b. The start/end vertex on the path
   c. Labels for all the edges on path (in order)

Example:
V vii START
V vi
V v
…
E A ii i -1
E B iii i 5

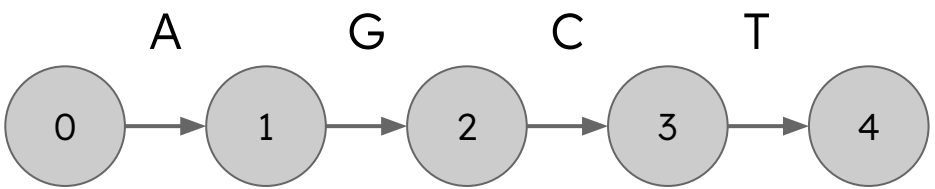
Part 1
Score: 8.0
Begin: vi
End: ii
Path: ID

Part 2
Score: 4.0
Begin: vii
End: i
Path: LIDA



Program 2: DNA Linked List

1. Create a linked list from a DNA sequence and a scoring scheme
   a. Positions are vertices
   b. Bases are edges
2. Run your program from part 1 on the graph



Example:

Scores

A = -1.49

T = -1.49

G = .74

C = .74

Sequence: AGCT

Graph:

0
1
2
3
4
A -1.49
G .74
C .74
T -1.49

Reminders

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