Genome 540 Discussion

January 15th, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

- Assignment 1 comments/common issues
- Assignment 2

Some common issues on HW1

- Match length histogram logic is incorrect
  - Does a match length of 1 make sense for two 10Mb sequences?
- The position of the longest match is shifted by 1bp
- The description of the longest match is not included/incorrect
- The number of non-alphabetic characters is incorrect
  - Only count the sequence position numbers!

Some more comments on HW1

- **Match the template!!!**
- gzip your homework
  - **gzip** lastname_firstname_hw1.txt
- Include your name in the homework
  - lastname_firstname_hw1.txt.gz
- You only need to submit on the “real data”

Comparing your result to the template

- Write your program
- Run it on the test data
- Run a diff between your program and the template
    - If your program is correct the answers should be the same
    - If your program is formatted correctly it should be EXACTLY the same (up to the Program line, and excluding manually written responses and the header)
    - Diff your_name_hw1.txt template.txt
- If using VSCode you can use the select for compare tool:
    - https://semanticdiff.com/blog/visual-studio-code-compare-files/

Assignment #2

Part 1 - Write a program

- The program should:
    - Read in a fasta file
    - Determine the frequencies of the nucleotides and dinucleotides (based on the forward strand) and the length of the sequence
    - Generate 3 sequences of the same length as the input file using:
        - the length (equal frequency assumption)
        - nucleotide frequency (order 0-Markov)
        - dinucleotide frequency (order 1-Markov)
    - Save these sequences as fasta files

Equal Frequency Model

A: 0.25
T: 0.25
G: 0.25
C: 0.25

Order 0 Markov Model

seq: ACTGA
length = 5

A: 2 T: 1 G: 1 C: 1	÷ 5 =	A: 0.4 T: 0.2 G: 0.2 C: 0.2
Number of times each base occurs		Probability of observing each base

Order 1 Markov Model

seq: ACTGATGATGGTACA

Length = 15, Number of dinucleotides = 14

	A	T	G	C
A	0	2	0	2
T	1	0	3	0
G	2	1	1	0
C	1	1	0	0

Dinucleotide Frequencies

e.g. # AT = 2

	A	T	G	C
A	0	.143	0	.143
T	.071	0	.214	0
G	.143	.071	.071	0
C	.071	.071	0	0

Dinucleotide Probabilities

e.g. P(AT) = 0.143

	A	T	G	C
A	0	.5	0	.5
T	.25	0	.75	0
G	.5	.25	.25	0
C	.5	.5	0	0

Nucleotide Conditional Probabilities

e.g. P(T|A) = 0.5

Part 2 - Simulate Sequences

- Using your program simulate 3 sequences from the mouse genomic region in HW1 using:
  - An equal frequency assumption
  - An order-0 Markov model
  - An order-1 Markov model
- Output sequences should be the same length as the input
- Store the sequences as fasta files

Part 3 - Run your HW1 on those seqs.

- Run your program from HW1 on each of those sequences
    - Sequence 1 should always be the 10Mb human region from HW1,
    - Sequence 2 should be your simulated sequence

Reminders

- HW2 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template