Genome 540 Discussion

February 20th, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Assignment 7

Overview

- Part 1: Use your predicted D-segments from hw6 to
  - Generate a new scoring scheme
  - Simulate background sequence
- Part 2: Run your D-segment program on the background and compare to the real data
- Part 3: Answer some questions

Part 1: New scoring scheme

Read start histogram for non-elevated copy-number segments:
0=331908 **- 8422401 (# Ns, don’t forget this)**
1=19439
2=4272
>=3=1332

Background frequencies:
0={#.####}
1={#.####}
2={#.####}
>=3={#.####}

Read start histogram for elevated copy-number segments:
0=1656
1=542
2=352
>=3=499

Target frequencies:
0={#.####}
1={#.####}
2={#.####}
>=3={#.####}

log2(target freq./background freq.)

Scoring scheme:
0={#.####}
1={#.####}
2={#.####}
>=3={#.####}

Part 1: Simulate new background sequence

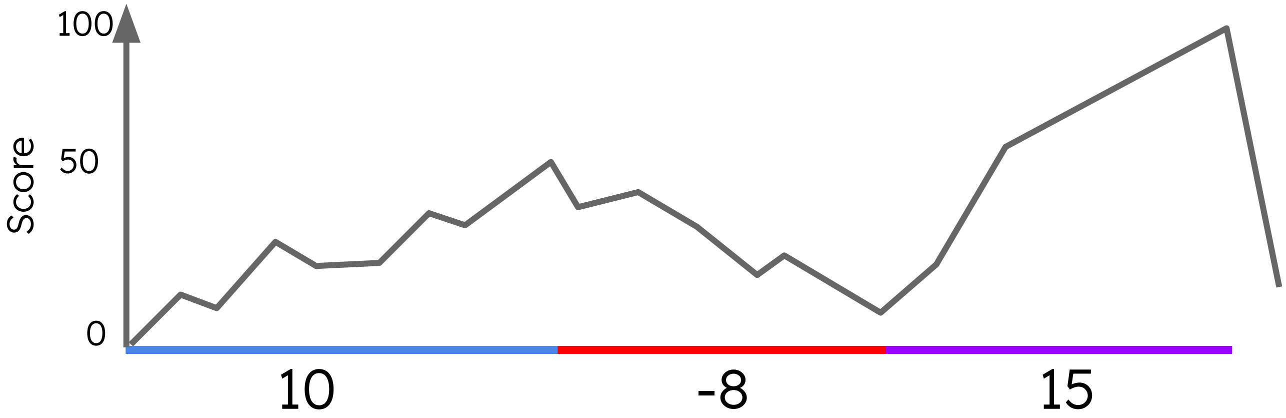
```
N = length of sequence to be simulated (length of seq. In HW6 - 8,422,401)
bkgd[r] = frequency of background sites with r read starts (r = 0, 1, 2, 3)
for each i = 1...N
    x = random number between 0 and 1 (uniform distribution)
    if x < bkgd[0]
        sim_seq[i] = 0
    else if x < bkgd[0] + bkgd[1]
        sim_seq[i] = 1
    else if x < bkgd[0] + bkgd[1] + bkgd[2]
        sim_seq[i] = 2
    else
        sim_seq[i] = 3
```

Part 2: Run D-seg and compare

Real data:
5 {# of segments with score >= 5}
6 {# of segments with score >= 6}
7 {# of segments with score >= 7}
.
.
.
list all the segment score counts for scores between 5 and 30
(only first/last 3 shown here)
.
.
.
28 {# of segments with score >= 28}
29 {# of segments with score >= 29}
30 {# of segments with score >= 30}

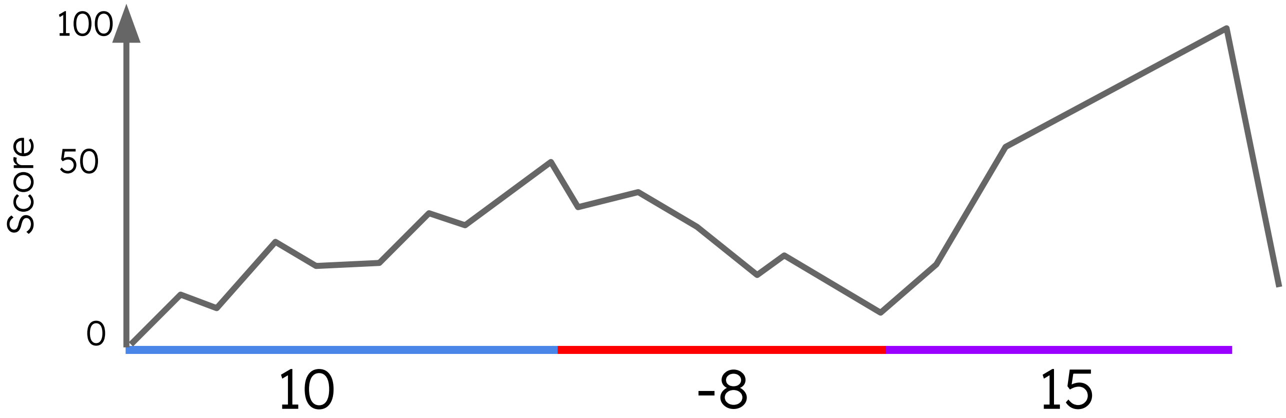
Simulated data:
5 {# of segments with score >= 5}
6 {# of segments with score >= 6}
7 {# of segments with score >= 7}
.
.
.
list all the segment score counts for scores between 5 and 30
(only first/last 3 shown here)
.
.
.
28 {# of segments with score >= 28}
29 {# of segments with score >= 29}
30 {# of segments with score >= 30}

Example



Simulated data:
5: ?
6: ?
7: ?
8: ?
9: ?
10: ?
11: ?
12: ?
13: ?
14: ?
15: ?
16: ?
17: ?
18: ?
19: ?
20: ?
…

Example



Simulated data:
5: 2
6: 2
7: 2
8: 2
9: 2
10: 2
11: 1
12: 1
13: 1
14: 1
15: 1
16: 0
17: 0
18: 0
19: 0
20: 0
…

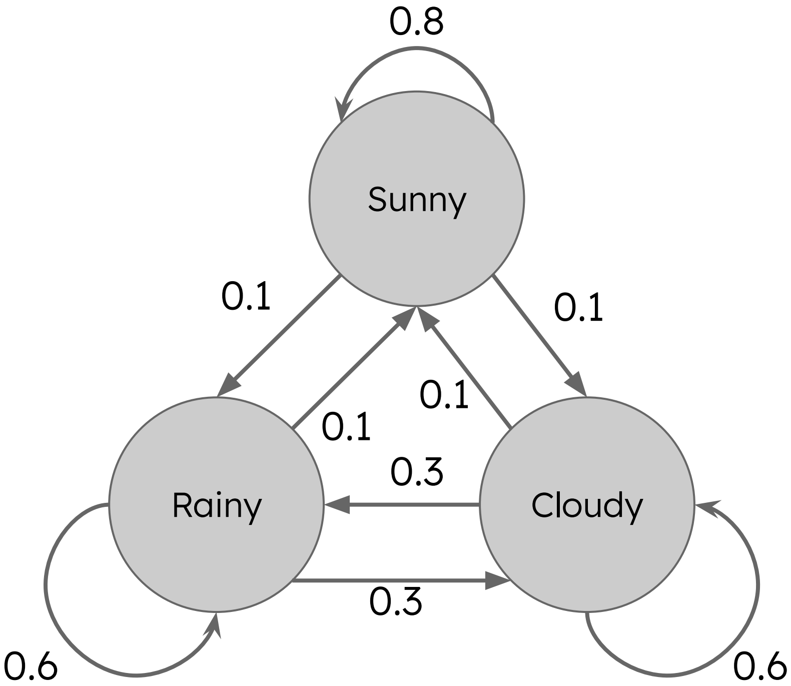
Part 2: Run D-seg and compare

Ratios of simulated data:
N_seg(5)/N_seg(6) {# of segments with score >= 5 / # of segments with score >= 6}
N_seg(6)/N_seg(7) {# of segments with score >= 6 / # of segments with score >= 7}
N_seg(7)/N_seg(8) {# of segments with score >= 7 / # of segments with score >= 8}
.
.
.
list all ratios
(only first/last 3 shown here)
.
.
.
N_seg(27)/N_seg(28) {# of segments with score >= 27 / # of segments with score >= 28}
N_seg(28)/N_seg(29) {# of segments with score >= 28 / # of segments with score >= 29}
N_seg(29)/N_seg(30) {# of segments with score >= 29 / # of segments with score >= 30}

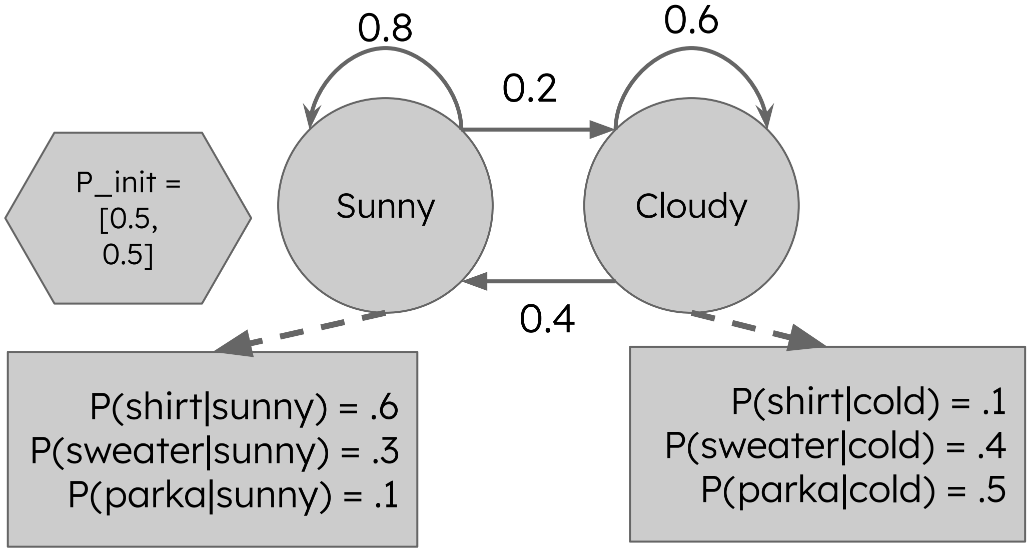
HMM Primer

Markov Chain vs. HMM

Markov Chain



HMM



P(shirt|sunny) = .6
P(sweater|sunny) = .3
P(parka|sunny) = .1

P(shirt|cold) = .1
P(sweater|cold) = .4
P(parka|cold) = .5

https://web.stanford.edu/~jurafsky/slp3/A.pdf

Markov Chain vs. HMM

Markov Chain

What is the probability of observing this sequence of states?

HMM

What are the most probable (unobserved) states given my observations?

e.g. I observed the sequence ATG, am I in a gene?

Reminders

- HW7 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template