Genome 540 discussion

February 25th, 2025

Joe Min

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

Homework 7

HMM intuition

Homework 7

Homework 7

1. Using the D-segments identified in HW6,
   a. Generate a new scoring scheme
   b. Simulate a background sequence
2. Using S = -D = 5,
   a. Run your D-segment algorithm on the simulated sequence
   b. Re-run your D-segment algorithm on the HW6 data
3. Answer four short answer questions

Example: generate a new scoring scheme

1. Generate background and target frequencies
   a. Background = (elevated + non-elevated (- #Ns if 0)) / sum(all)
   b. Target = elevated / sum(elevated)

read start counts

non-elevated
0=18422401 - 8422401
1=200000
2=30000
>=3=4000

total = 10,234,000

Hardcoded #Ns for chromosome 16

elevated
0=10000
1=200
2=30
>=3=4

total = 10,234

totals

background
0=10000000 + 10000
1=200000 + 200
2=30000 + 30
>=3=4000 + 4

total = 10,244,234

freqs

background
0=10010000 / 10244234
1=200200 / 10244234
2=30030 / 10244234
>=3=4004 / 10244234

Example: generate a new scoring scheme

1. Generate background and target frequencies
   a. Background = (elevated + non-elevated (- #Ns if 0)) / sum(all)
   b. Target = elevated / sum(elevated)

```
# freqs as ratios

## background
0=10010000 / 10244234
1=200200 / 10244234
2=30030 / 10244234
>=3=4004 / 10244234
```

```
# freqs format

## background
0={#.####}
1={#.####}
2={#.####}
>=3={#.####}
```

```
# final freqs

## background
0=0.9771
1=0.0195
2=0.0029
>=3=0.0004
```

```
## target
0=10000 / 10234 = 0.9971
1=200 / 10234 = 0.0195
2=30 / 10234 = 0.0029
>=3=4 / 10234 = 0.0004
```

Example: generate a new scoring scheme

2. Using LLRs with base 2, create a new scoring scheme

```
# final freqs                          # scoring scheme                  # scheme format

## background      ## target
0=0.9771           0=0.9971            0=log2(0.9971 / 0.9771)         0={#.####}
1=0.0195           1=0.0195            1=log2(0.0195 / 0.0195)         1={#.####}
2=0.0029           2=0.0029            2=log2(0.0029 / 0.0029)         2={#.####}
>=3=0.0004         >=3=0.0004          >=3=log2(.0004 / .0004)         >=3={#.####}
```

Example: simulate a sequence of read start counts

Using the calculated background frequencies, simulate a new sequence that is the same length as chromosome 16 (minus Ns)

# totals	# final freqs
## background	## background
0=10010000	0=0.9771
1=200200	1=0.0195
2=30030	2=0.0029
>=3=4004	>=3=0.0004
total = 10,244,234	

10,244,234

```
N = length of sequence to be simulated
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
```

This is iteratively sampling from the background distribution

D-segment output format

Generate a histogram of segments >= to a given score

scoring scheme

0=log2(0.9971 / 0.9771)
1=log2(0.0195 / 0.0195)
2=log2(0.0029 / 0.0029)
>=3=log2(.0004 / .0004)

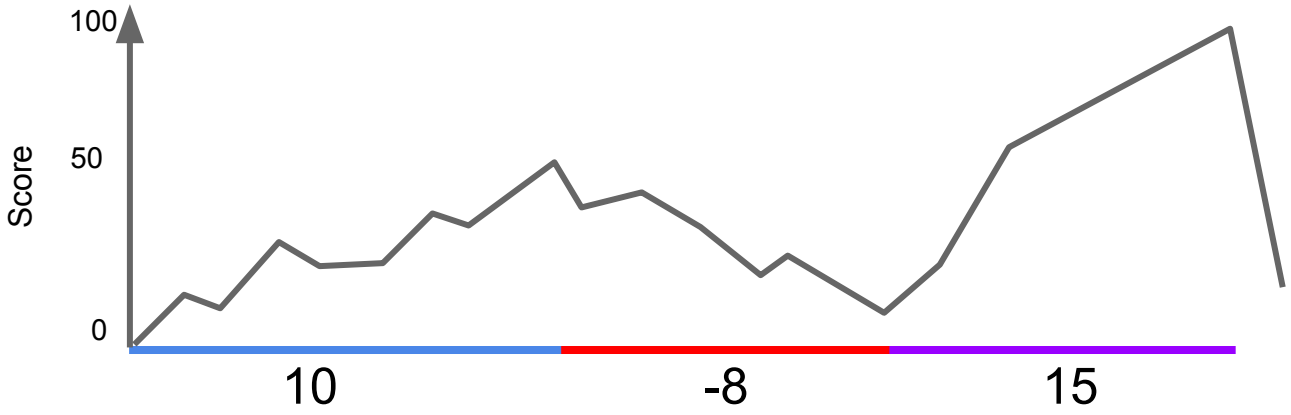
S = 5
D = -5

real sequence:
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 sequence:
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}

D-segment output example

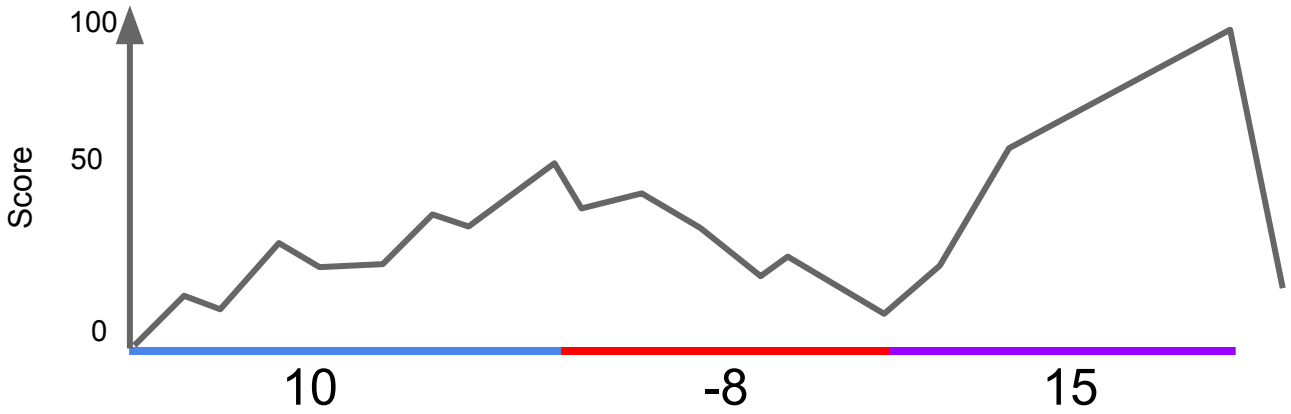
After identifying and scoring segments, you can add them to the histogram



5: ?
6: ?
7: ?
.
.
.
28: ?
29: ?
30: ?

D-segment output example

After identifying and scoring segments, you can add them to the histogram



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
.
.
.
30: 0

Comparing simulated data to real data

Finally, output the ratios of adjacent histogram values to understand the distribution shape of segment scores

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 intuition

HMM conceptual example

If we wanted to utilize an HMM that represent winter weather temperatures, let’s consider one implemented as a two state HMM:

- State 1: Normal cold winter weather
- State 2: In a warm spell

HMM conceptual example

These two states might determine a day's temperature max with the following known emission probabilities:

State 1 (normal weather)

- >=40F: 10%
- <40F: 90%

State 2 (warm spell)

- >=40F: 60%
- <40F: 40%

HMM conceptual example

So, if we measure the max temperature for 5 consecutive days, we might see:

Day 1	Day 2	Day 3	Day 4	Day 5
>= 40F	< 40F	>= 40F	>= 40F	>= 40F

And just by eye, this looks like these all could have been emitted by staying constantly in state 2, so we might take this as evidence that we are in a warm spell

HMM conceptual example

More formally, we can calculate the most probable set of 5 states that could have produced the observed data. For simplicity, consider staying in one state the whole time:

	Day 1	Day 2	Day 3	Day 4	Day 5
	>= 40F	< 40F	>= 40F	>= 40F	>= 40F
State 1 probs	0.1	0.9	0.1	0.1	0.1
State 2 probs	0.6	0.4	0.6	0.6	0.6

HMM conceptual example

If we had been in State 1 the entire time, the probability of the observed data would be (0.1)^4 * 0.9 = 0.00009

State 2 the whole time: (0.6)^4 * 0.4 = 0.0518

	Day 1	Day 2	Day 3	Day 4	Day 5
	>= 40F	< 40F	>= 40F	>= 40F	>= 40F
State 1 probs	0.1	0.9	0.1	0.1	0.1
State 2 probs	0.6	0.4	0.6	0.6	0.6

HMM conceptual example

Given transition probabilities between states (1→2 and 2→1) we could also compute the probability of any sequence of hidden states

	Day 1	Day 2	Day 3	Day 4	Day 5
	>= 40F	< 40F	>= 40F	>= 40F	>= 40F
State 1 probs	0.1	0.9	0.1	0.1	0.1
State 2 probs	0.6	0.4	0.6	0.6	0.6

HMM conceptual example

We did not consider any cases where there would be transitions, but just wanted to drive home:

- HMMs are tools for finding the most probable set of underlying states (e.g., “warm spell”) for a given set of observed data (max temp from 5 consecutive days)

Office hours

Reminder:

Homework 7 is due Sunday, March 2nd at 11:59pm!