Genome 540 Discussion

March 7th, 2024

Clifford Rostomily

Genome Sciences
UNIVERSITY OF WASHINGTON

Assignment 9

Overview

- Calculate emission probabilities from 2 files containing:
  - Alignment column counts from a large set of ancient repeat sequences
  - Conserved alignment column counts from putative functional sites
- Using these emission probabilities and the given transition and initiation probabilities find “conserved” and “not conserved” regions in an alignment of human, dog, and mouse

Calculating Emission Probabilities

Neutral State: Ancient Repeat Sequences

AAA	10222095
AAC	481243
AAT	420185
AAG	1415675
AA-	273456
ACA	852624
ACC	179459
ACT	99493
ACG	167810
AC-	29636
ATA	874547
ATC	113150
ATT	220714
ATG	185789

etc …

1st base: human
2nd base: dog
3rd base: mouse

Conserved State: Putative Functional Sites

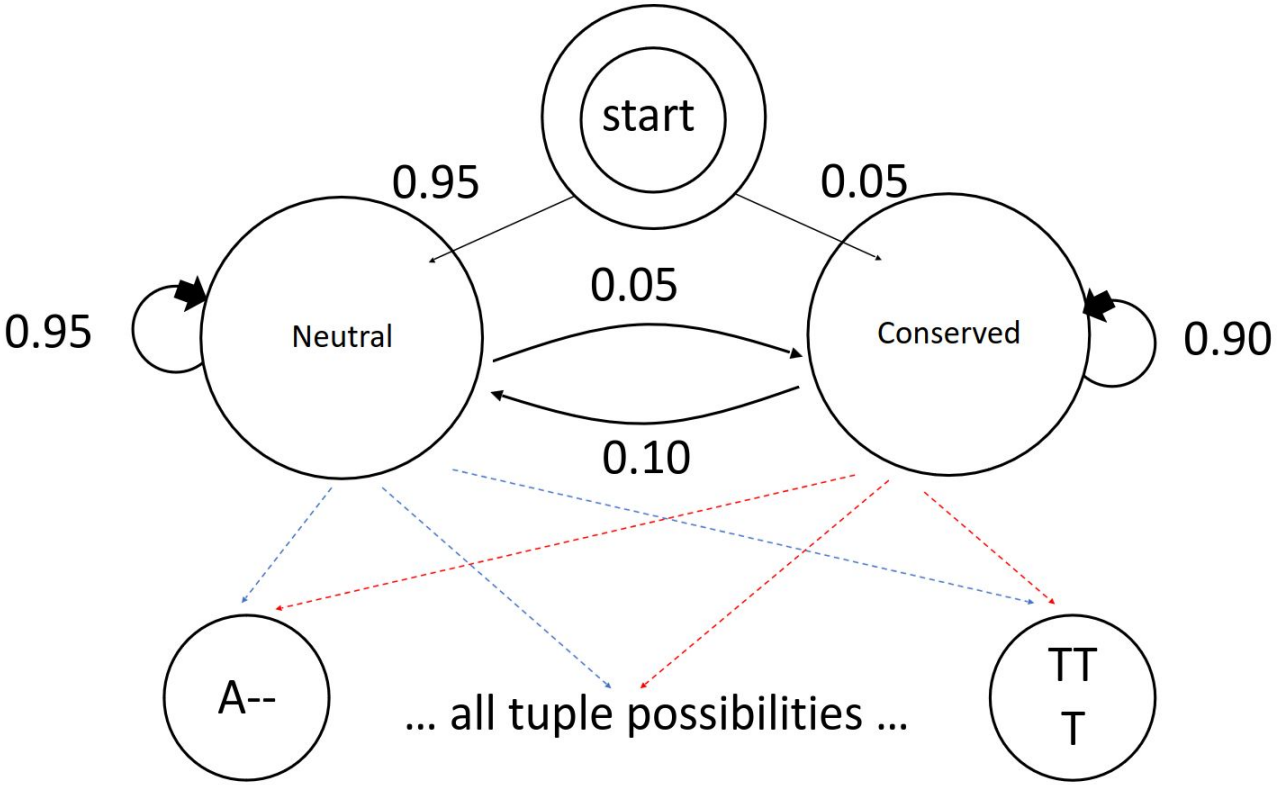
AAA	2375583
AAC	21337
AAT	10886
AAG	56328
AA-	3205
ACA	33210
ACC	12122
ACT	2270
ACG	5187
AC-	374
ATA	21805
ATC	2871
ATT	7426
ATG	4369

etc …

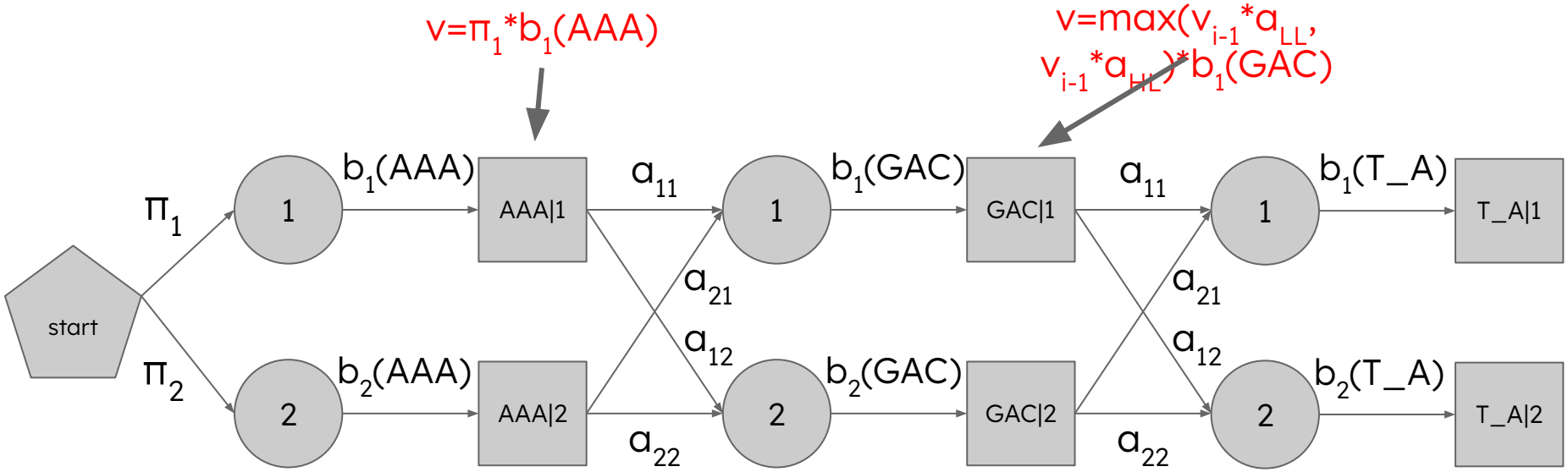
Input data

```
# chr7:26924045-26924056
hg18    TGCTCACATTTT
canFam2 --CTCACAGTTT
mm9     ------CGCTT-

# chr7:26924057-26924120
hg18    CTAGAAGGATTAATGTTCTGTAGATCTATTGATCTTCTACATTCTTCTTAAAGTATCCAGGGTA
canFam2 TCAGAGGGATTAGTGTTCTGTGGATCTATTGATCTTCTGCACTCTTCCTAAAGTATCTGGGGGA
mm9     CCAGAGGGAGTGGTGTTCTGTAGATCTATCGACCTTC--CACGCAGCTAAAAGTACCCTGGGTG

# chr7:26924121-26924289
hg18    ATCATTAACAATACTTTGTTTTGATTTACTTGCCTGGTGTCTGAGGCTCTTCCAGCTCTCTACAATACATTTGCGCTTTATTCATGATGCTTATTCTGTAGATAAAGACAGCACATTACTGGCATTTGTAACTGGGAGGCTTAAAATTTTTAAACATAAAATTAGAGAT
canFam2 ATCATTAGCAACACTTTGTTCTGATCTACTTGCCTGTCATCCAAGGCTATTCCAGCTCTCTAAAATACATTTGTGCTTTATTCATCATGCTTATTCTATA-ATAAAGACCTTACCTTACTGGCATTTATAACTGGGAGGCATAAGACTTTTAAAAATTAGATTATATGT
mm9     ------------ACTTCGCTCTGCTCCACTTGCCTGACATCCAAGGCTCTCCCAGTTCTGTATAATGTCTTCGTGTTTCATTCACTATTCTTATGTTATA----AAGACTGAGTGTTCTTGGCACCTTCAATTGGAAAGCTTAA----TCAAAAAAGTAGAT-------

# chr7:26924290-26924313
hg18    AATCTAATGTTTAGATTAGGGTTA
canFam2 ------------------------
mm9     ----------TTAGA-------TA

# chr7:26924314-26924339
hg18    GATTTTTAAATAGGGTATAGAACTTC
canFam2 GCCTTTTAAGTAGGGTGTAGATTTTC
mm9     -CCTTTTAAATGAGACACAGATCTTC
```

HMM For HW9

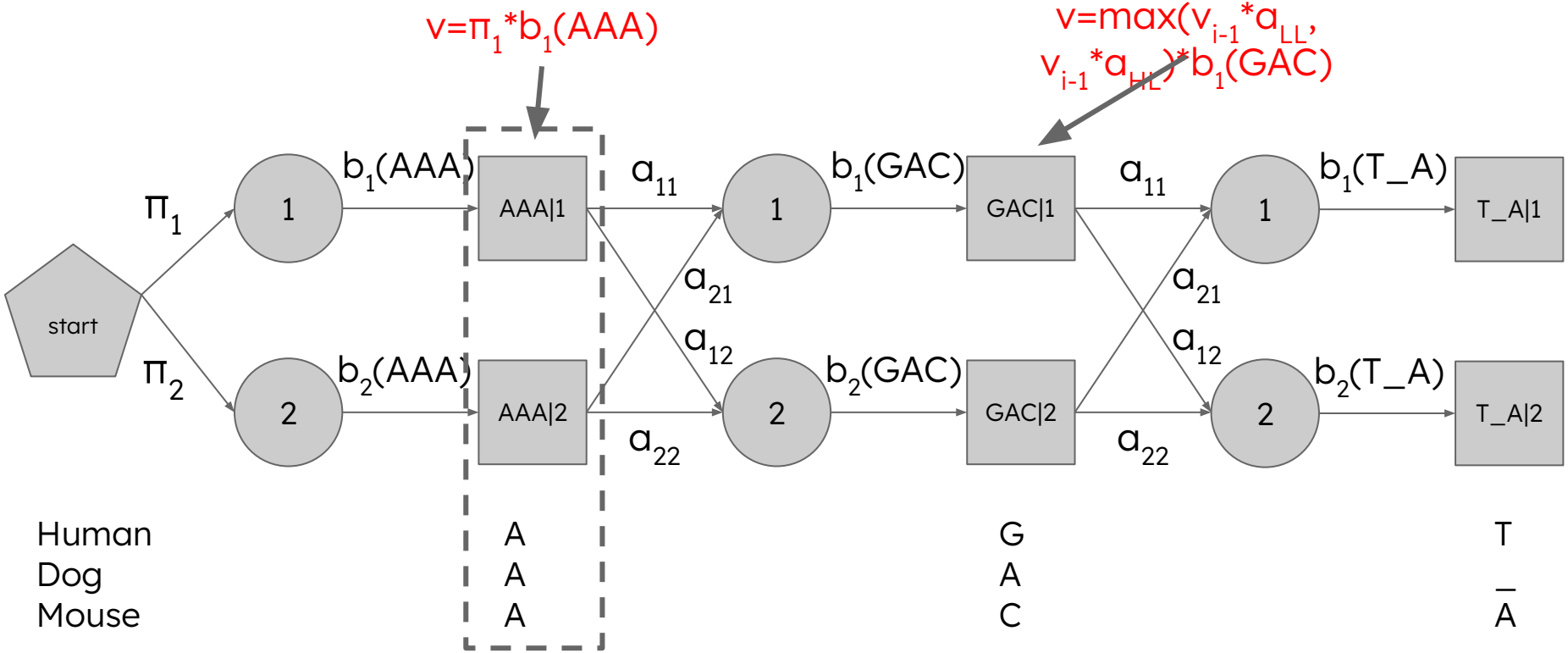

start
0.95
0.05
0.95
Neutral
0.05
Conserved
0.90
0.10
A--
… all tuple possibilities …
TT
T

Viterbi - Most probable sequence of states

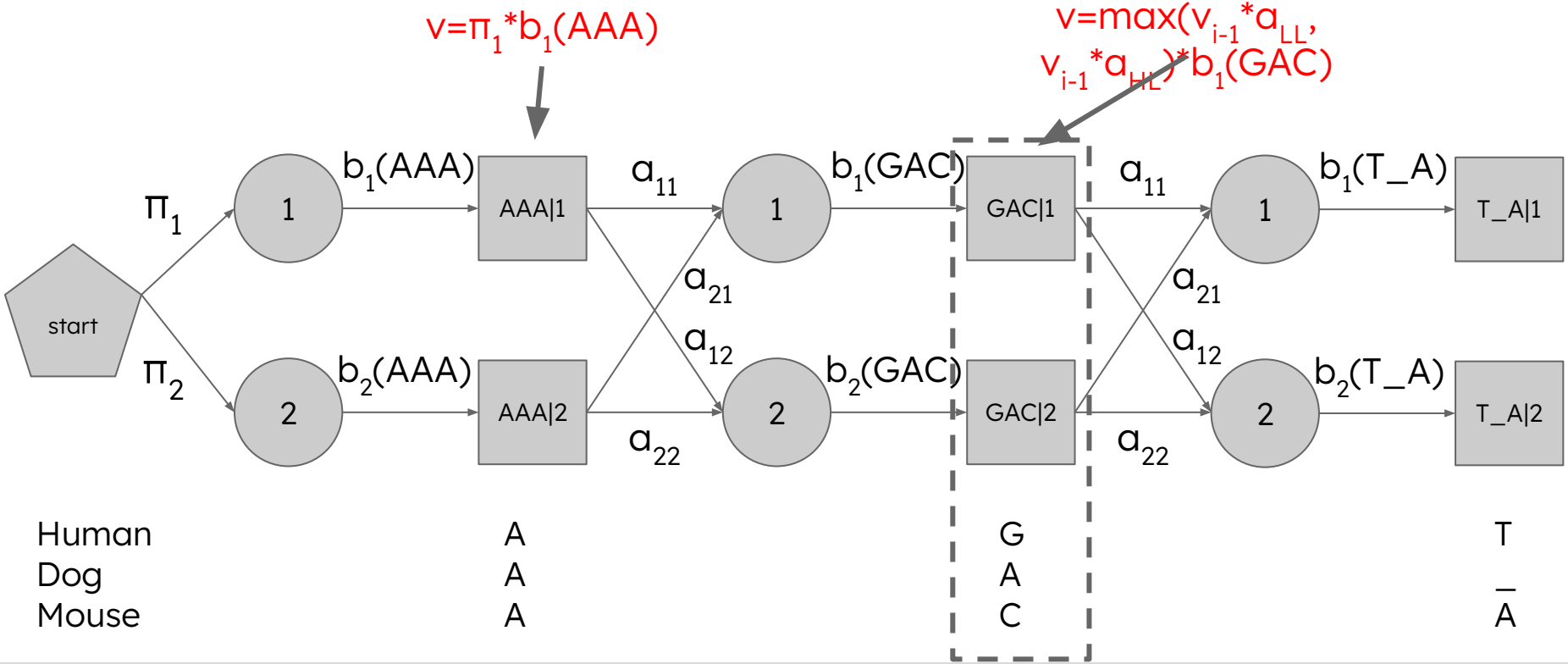


Human	A	G	T
Dog	A	A	_
Mouse	A	C	A

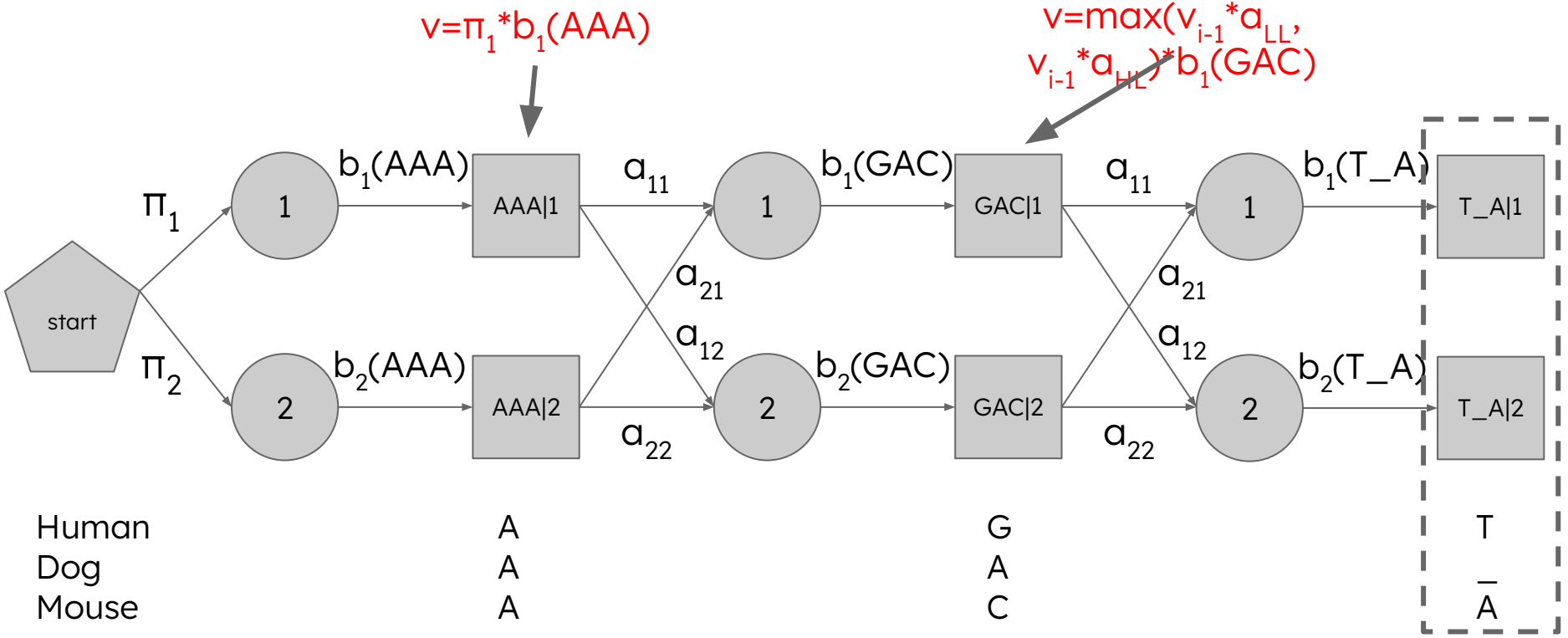
Process as a sliding window



Process as a sliding window



Process as a sliding window



Human	A	G	T
Dog	A	A	_
Mouse	A	C	A

Output

- State and segment histograms
- Parameter values
  - Initiation/transition probabilities you were given in the assignment
  - Emission probabilities you calculated from neutral and conserved data sets
- Coordinates of 10 longest conserved segments (report positions relative to the start of the chromosome)
- Brief annotations for the 5 longest conserved segments (look at UCSC genome browser, and make sure using the correct genome version, e.g. hg18)

```
State Histogram:
1=5
2=3

Segment Histogram:
1=2
2=1
```

```
Initial State Probabilities:
1=0.90000
2=0.10000

Transition Probabilities:
1,1=0.99000
1,2=0.01000
2,1=0.20000
2,2=0.80000

Emission Probabilities:
1,A--=0.20000
1,A-A=0.20000
1,A-C=0.20000
1,A-G=0.20000
1,A-T=0.20000
.
.
.
2,A--=0.10000
2,A-A=0.20000
2,A-C=0.25000
2,A-G=0.25000
2,A-T=0.20000
etc..
```

```
Longest Segment List:

116741000 116752000
116745000 116756000
etc.. (give 10 longest from state 2)

Annotations:

Start: 116741000
End: 116752000
Overlaps with exon3 of the protein coding gene cMyc

Start: 116745000
End: 116756000
Overlaps with exon4 of the protein coding gene cMyc

etc.. (give 5 longest)
```

You’re almost there!

- HW9 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template
- Thanks for a great quarter!