Genome 540 Discussion

March 7th, 2024
Clifford Rostomily
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

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Emission Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>10222095</td>
</tr>
<tr>
<td>AAC</td>
<td>481243</td>
</tr>
<tr>
<td>AAT</td>
<td>420185</td>
</tr>
<tr>
<td>AAG</td>
<td>1415675</td>
</tr>
<tr>
<td>AA−</td>
<td>273456</td>
</tr>
<tr>
<td>ACA</td>
<td>852624</td>
</tr>
<tr>
<td>ACC</td>
<td>179459</td>
</tr>
<tr>
<td>ACT</td>
<td>99493</td>
</tr>
<tr>
<td>ACG</td>
<td>167810</td>
</tr>
<tr>
<td>AC−</td>
<td>29636</td>
</tr>
<tr>
<td>ATA</td>
<td>874547</td>
</tr>
<tr>
<td>ATC</td>
<td>113150</td>
</tr>
<tr>
<td>ATT</td>
<td>220714</td>
</tr>
<tr>
<td>ATG</td>
<td>185789</td>
</tr>
<tr>
<td>etc ...</td>
<td></td>
</tr>
</tbody>
</table>

**Conserved State:** Putative Functional Sites

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Emission Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>2375583</td>
</tr>
<tr>
<td>AAC</td>
<td>21337</td>
</tr>
<tr>
<td>AAT</td>
<td>10886</td>
</tr>
<tr>
<td>AAG</td>
<td>56328</td>
</tr>
<tr>
<td>AA−</td>
<td>3205</td>
</tr>
<tr>
<td>ACA</td>
<td>33210</td>
</tr>
<tr>
<td>ACC</td>
<td>12122</td>
</tr>
<tr>
<td>ACT</td>
<td>2270</td>
</tr>
<tr>
<td>ACG</td>
<td>5187</td>
</tr>
<tr>
<td>AC−</td>
<td>374</td>
</tr>
<tr>
<td>ATA</td>
<td>21805</td>
</tr>
<tr>
<td>ATC</td>
<td>2871</td>
</tr>
<tr>
<td>ATT</td>
<td>7426</td>
</tr>
<tr>
<td>ATG</td>
<td>4369</td>
</tr>
<tr>
<td>etc ...</td>
<td></td>
</tr>
</tbody>
</table>

1\textsuperscript{st} base: human  
2\textsuperscript{nd} base: dog  
3\textsuperscript{rd} base: mouse
Input data

# chr7:26924045-26924056
hg18     TGTCACATTTT
CanFam2   --CTCACAGTTT
mm9       ------GCTTT-

# chr7:26924057-26924120
hg18     CTAGAAAGGATTATATGTGTTGCTGATAGTCATGATCTTCTAGATTCTGCTCCATTAAAGTATCCAGGTA
CanFam2   TCAAGGGATTATATGTGTTGCTGATAGTCATGATCTTCTAGATTCTGCTCCATTAAAGTATCCAGGTA
mm9       CCAAGGGATTATATGTGTTGCTGATAGTCATGATCTTCTAGATTCTGCTCCATTAAAGTATCCAGGTA

# chr7:26924121-26924289
hg18     AATGACAATACATTCTTTGATTACCTACCTGCTGCCCTCAACCTGCTACATAGATTTTCTGTTTCTGTATCTGATCTGAGATATAGACACCTTTATTATTTATTATTATGAGAT
CanFam2   AATGACAATACATTCTTTGATTACCTACCTGCTGCCCTCAACCTGCTACATAGATTTTCTGTTTCTGTATCTGATCTGAGATATAGACACCTTTATTATTTATTATTATGAGAT
mm9       AATGACAATACATTCTTTGATTACCTACCTGCTGCCCTCAACCTGCTACATAGATTTTCTGTTTCTGTATCTGATCTGAGATATAGACACCTTTATTATTTATTATTATGAGAT

# chr7:26924230-26924313
hg18     AACTTAATGTGGAAGTTAGTTGGGTA
CanFam2   AACTTAATGTGGAAGTTAGTTGGGTA
mm9       AACTTAATGTGGAAGTTAGTTGGGTA

# chr7:26924314-26924339
hg18     GATTTTTAATAGGTATAGAATACCTC
CanFam2   GATTTTTAATAGGTATAGAATACCTC
mm9       GATTTTTAATAGGTATAGAATACCTC
HMM For HW9

... all tuple possibilities ...
Viterbi - Most probable sequence of states

\[ v = \pi_1 b_1(AAA) \]

\[ v = \max(v_{i-1} a_{LL}, v_{i-1} a_{HL}) b_1(GAC) \]

**Diagram:**
- **States:** 1, 2
- **Transitions:**
  - \( a_{11} \) from 1 to 1
  - \( a_{21} \) from 2 to 1
  - \( a_{12} \) from 1 to 2
  - \( a_{22} \) from 2 to 2
- **Initial State:** \( \pi_1 \)
- **Final States:**
  - T.A.1
  - T.A.2

**Alphabets:**
- A
- G
- T

**Human:**
- A
- G
- T

**Dog:**
- A
- A
- A

**Mouse:**
- A
- A
- A
Process as a sliding window

\[ v = \pi_1 b_1(AAA) \]

\[ v = \max(v_{i-1} a_{LL}, v_{i-1} a_{HL}) b_1(GAC) \]
Process as a sliding window

\[ v = \pi_1 b_1(\text{AAA}) \]

\[ v = \max(v_{i-1} a_{LL}, v_{i-1} a_{HL}) b_1(\text{GAC}) \]

- **Human**
  - A
  - G
  - T

- **Dog**
  - A
  - A
  - T

- **Mouse**
  - A
  - A
  - A

**Diagram:**

1. Start state
2. Transition 1 (\( b_1(\text{AAA}) \))
3. Transition 2 (\( b_2(\text{AAA}) \))
4. Transition 1 (\( b_1(\text{GAC}) \))
5. Transition 2 (\( b_2(\text{GAC}) \))
6. Transition 1 (\( b_1(T_A) \))
7. Transition 2 (\( b_2(T_A) \))
Process as a sliding window

\[ v = \pi_1 b_1(\text{AAA}) \]

\[ v = \max(v_{i-1} \cdot a_{\text{LL}'}, v_{i-1} \cdot a_{\text{HL}}) \cdot b_1(\text{GAC}) \]

**Human**
- A
- A
- A

**Dog**
- A
- A
- A

**Mouse**
- A
- A
- 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!