Genome 540 Discussion

Conor Camplisson

March 7th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

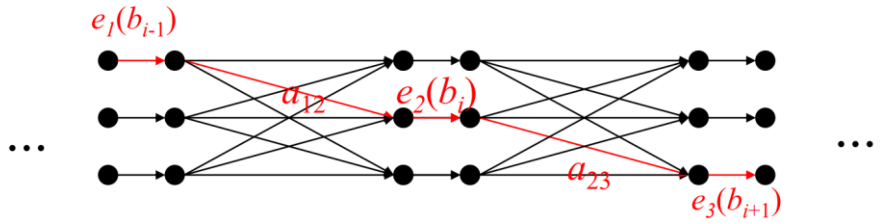
Outline

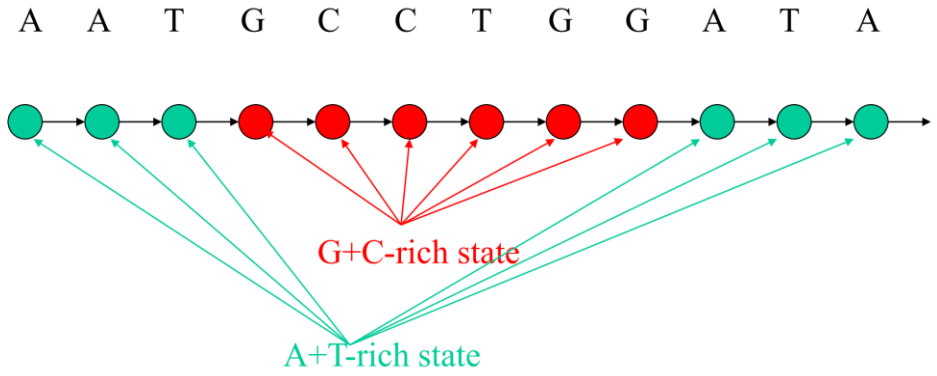
- Homework 8 Wrap-up
- Homework 9 Overview

Outline

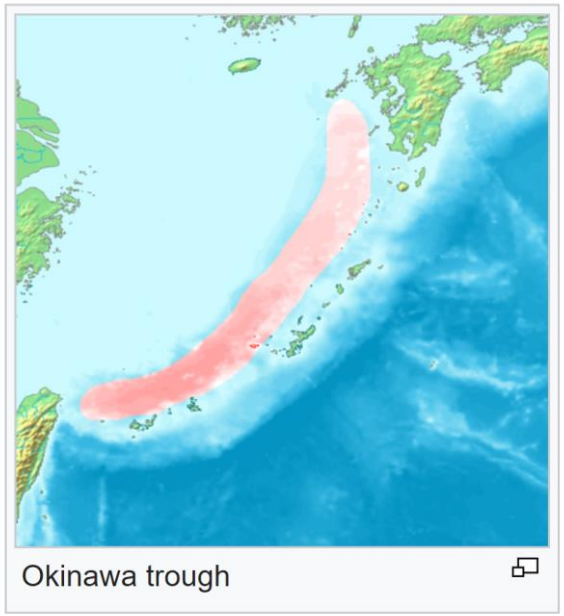
- Homework 8 Wrap-up
- Homework 9 Overview

Homework 8 Overview





Pyrococcus horikoshii



ACGGCTTTGACCAC

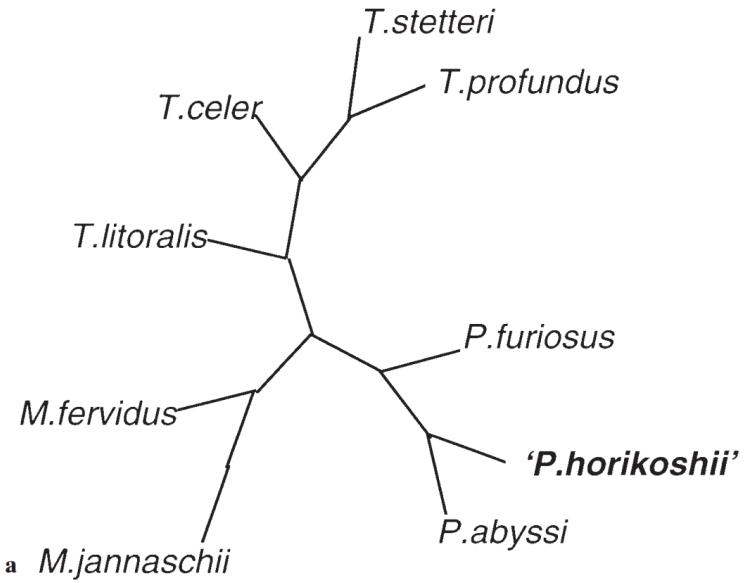
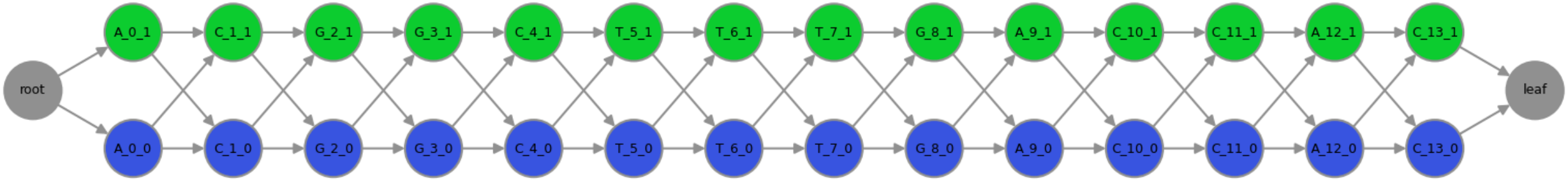


Fig. 3. Phylogenetic trees representing the relatedness of *P. horikoshii*



http://bozeman.mbt.washington.edu/compbio/mbt599/assignments/hw8.html

Outline

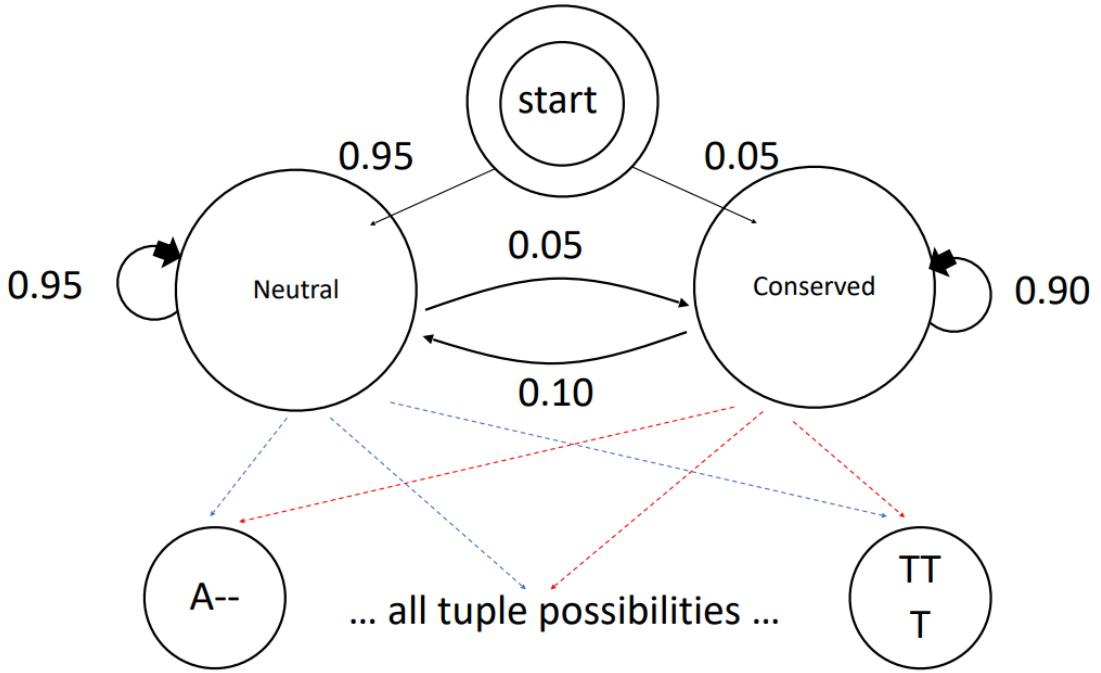
- Homework 8 Wrap-up

- Homework 9 Overview

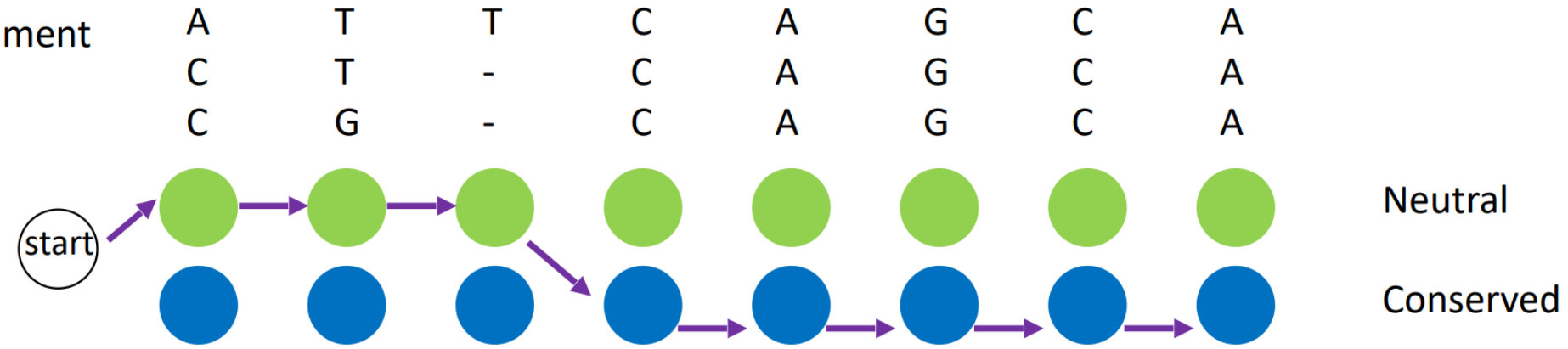
Homework 9 Overview

- ENCODE region 010 (chromosome 7)
- Multiple alignment of human, dog, and mouse
- 2 states:
  - neutral (fast-evolving)
  - conserved (slow-evolving)
- Emitted symbols are multiple alignment columns (e.g. ‘AAT’)
- Viterbi parse (no iteration)

HW9 – Model Structure



Observation: Alignment



HW9 – Model Parameters

Alignment Column Counts Provided

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
AT-	32253
AGA	2116012
AGC	139953
AGT	131553
AGG	881616
AG-	73372
A-A	760405
A-C	57350
A-T	56348
A-G	155911
A--	39186

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
AT-	294
AGA	81919
AGC	4455
AGT	2735
AGG	50413
AG-	796
A-A	6234
A-C	557
A-T	350
A-G	1349
A--	1282

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

Calculate Emission Probabilities

- For ‘neutral’ state emission probabilities, use observed frequencies in neutral data set (ancient repeat sequences)
- For ‘conserved’ state emission probabilities, use observed frequencies in functional data set

Initiation, Transition Probabilities

- Given in problem set description

HW9 – Input Data

Original maf format:

- Sequences broken into alignment blocks based on the species included
- Official file format specs

Homework file format:

- Only 3 species
- Gaps in human sequence were removed and ambiguous bases replaced with ‘A’ for simplicity

```
# chrX:152767699-152767743
hg18    ATAAAAACATTAAAAAAAATCAGCCACAGGACTTGGTCTTGGACC
canFam2 -----------------------------------------
mm9     -----------------------------------------

# chrX:152767744-152767853
hg18    CAAGTTAGAGCTAGGCCATGCTTGCTTAAAGGAGTGGCTGTAATTTTAAACAAGGCTAGTGGGAAAGT
canFam2 ----------------------------------------------------------------------
mm9     ----------------------------------------------------------------------
```

HW9 – Output

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)

HW9 – Output

```
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)
```

. . .