Genome 540 Discussion

Conor Camplisson

February 28th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

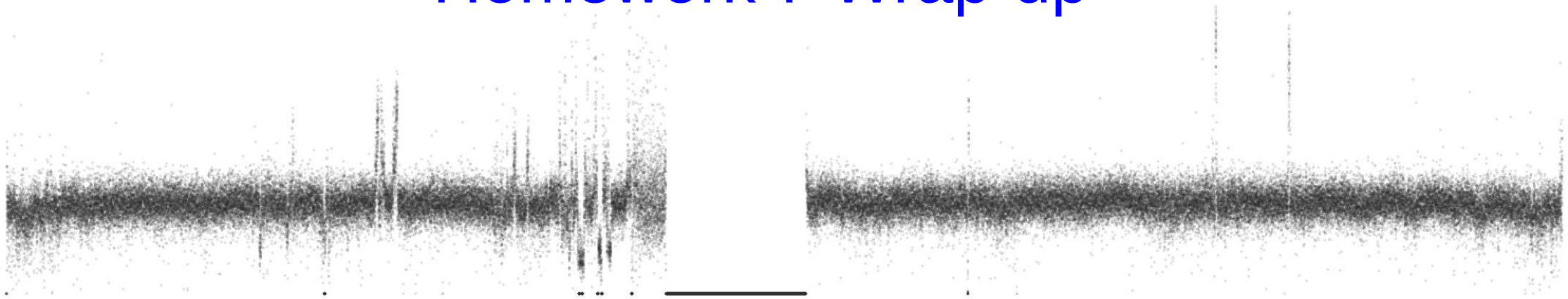
Outline

- Homework 7 wrap-up
- Homework 8 overview & questions

Outline

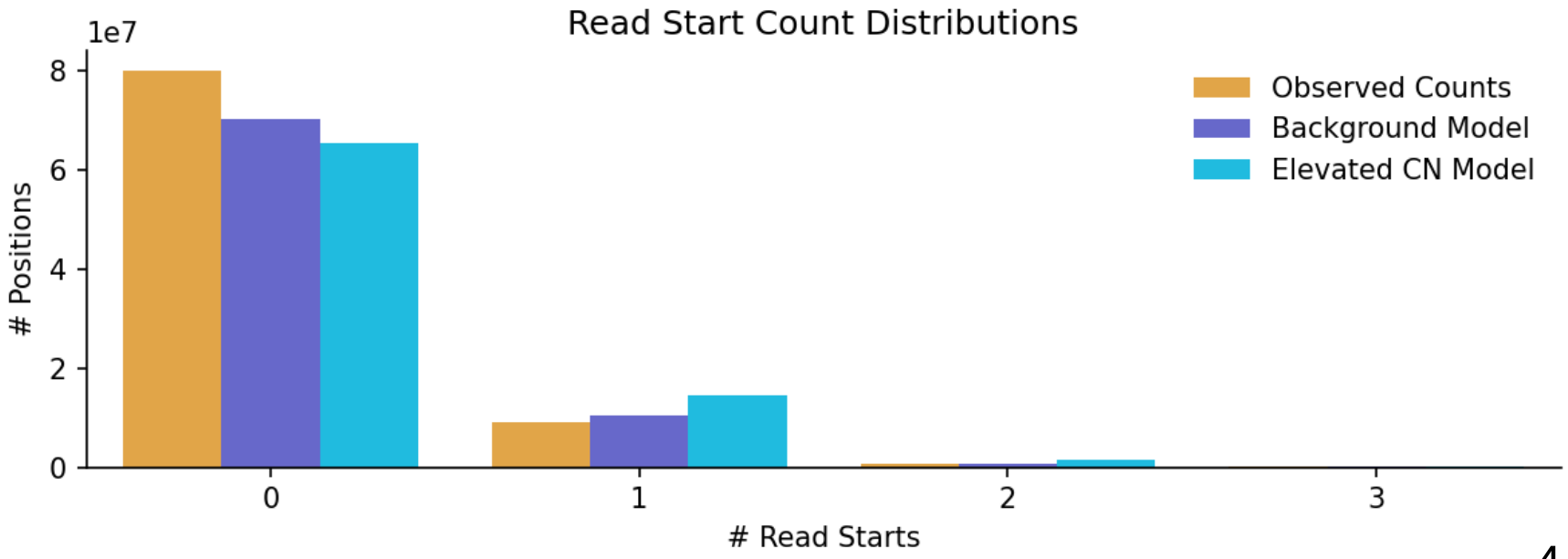
- Homework 7 wrap-up
- Homework 8 overview & questions

Homework 7 Wrap-up



chm13.chr16.txt

```
16      1       0
16      2       0
16      3       0
16      4       0
16      5       0
      [ ... ]
16      14793   0
16      14794   1
16      14795   3
16      14796   0
      [ ... ]
```



Homework 7 Wrap-up

```
Assignment: GS540 HW7
Name: {YOURNAME}
Email: {YOUREMAIL}
Language: {YOURLANGUAGE}
Running time: {YOURRUNTIME}

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

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

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

```
Simulated data:
5 {# of segments with score >= 5}
6 {# of segments with score >= 6}
7 {# of segments with score >= 7}
```

```
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
(only first/last 3 shown here)
.
.
.
28 {# of segments with score >= 28}
29 {# of segments with score >= 29}
30 {# of segments with score >= 30}
```

for scores between 5 and 30

```
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}
```

As discussed in lecture, Karlin-Altschul theory predicts that, for LLR scores using logarithmic base b, the number of D-segments with scores >= s should be proportional to b^-s (b to the power -s; this is the reciprocal of the corresponding LR). Since your scores used logarithmic base 2, if N_seg(s1) is the number of D-segments found with score value >= s1, and N_seg(s2) is the number of D-segments found with score value >= s2, then the ratio N_seg(s1)/N_seg(s2) should be approximately equal to 2^(s2 - s1). Consider the following questions:

- Does this relationship appear to be true for the simulated data?
- Is it true for the real data?
- Would you expect it to be true for the real data?
- What score threshold is a reasonable one to use for the real data, to ensure a very low false positive rate?

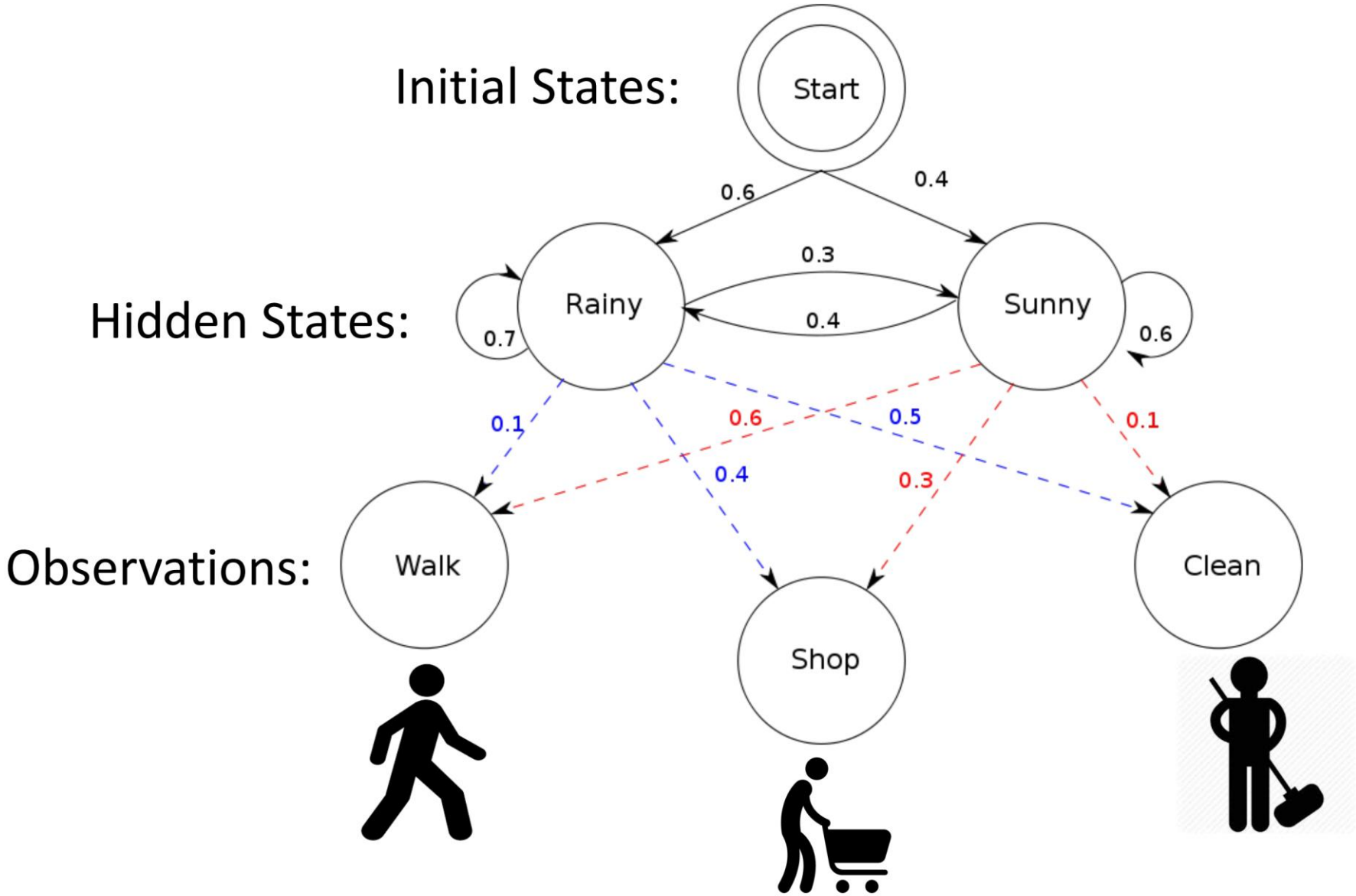
http://bozeman.mbt.washington.edu/compbio/mbt599/assignments/hw7.html

Outline

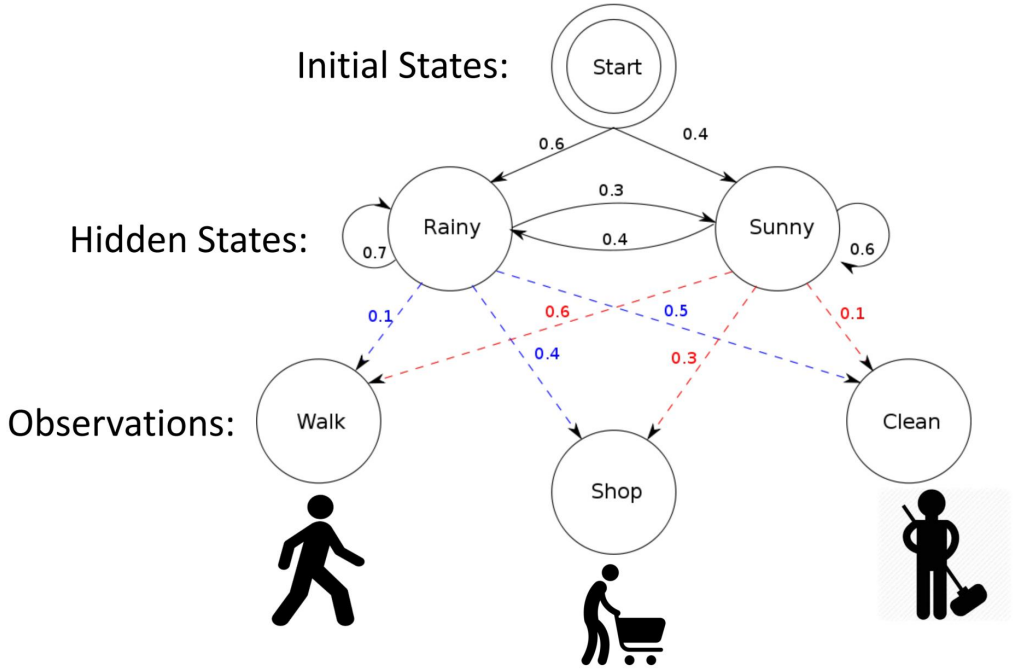
- Homework 7 wrap-up

- Homework 8 overview & questions

HMM Overview

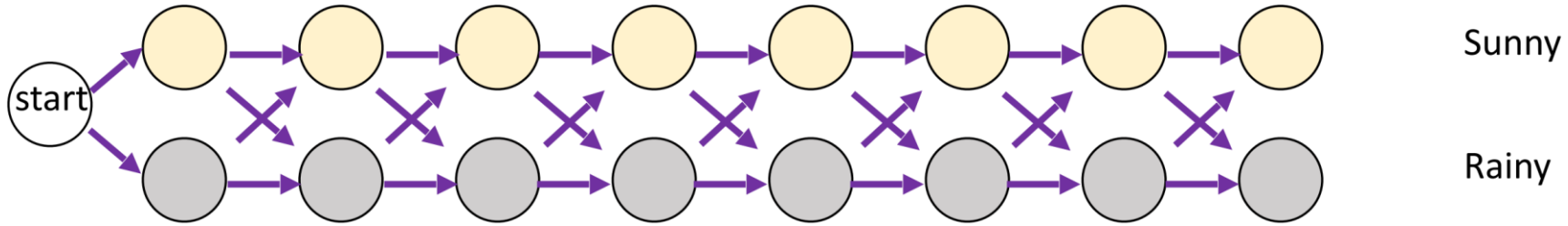


HMM Overview

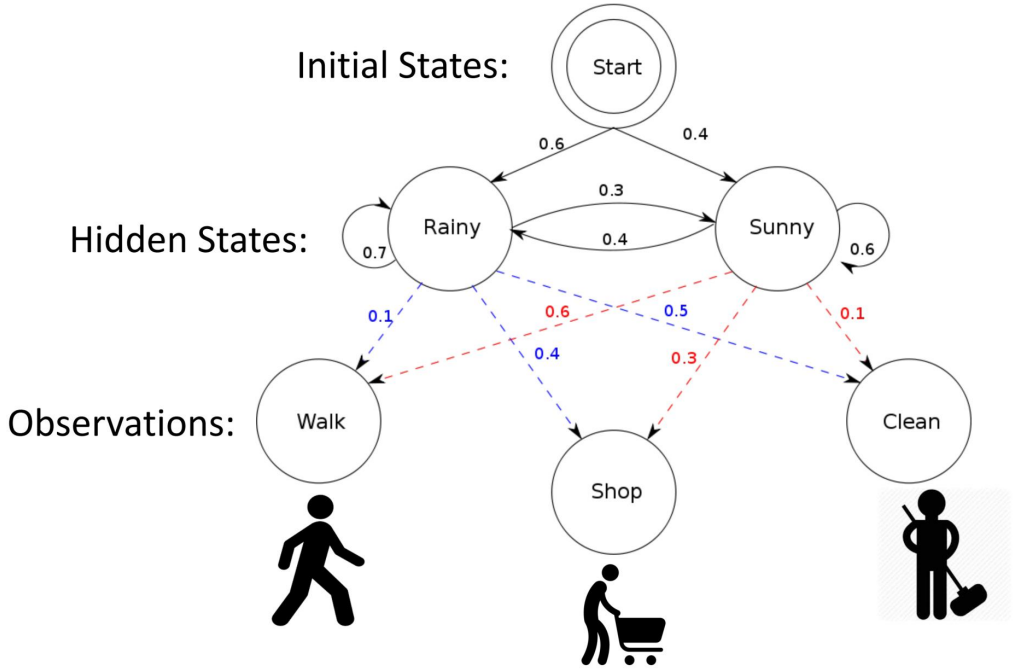


Observation List:

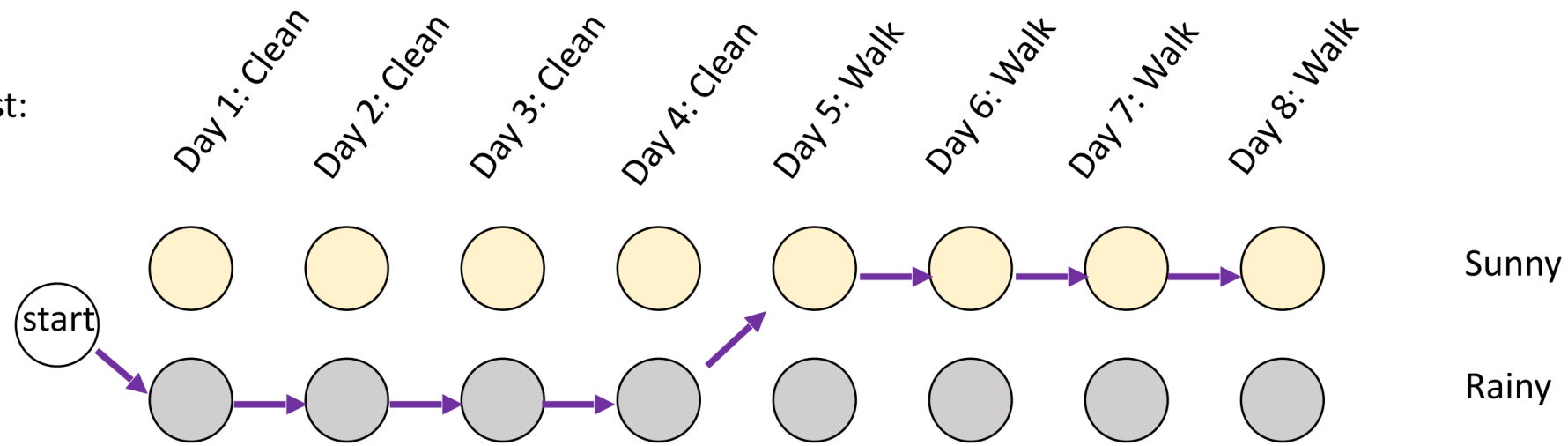
All possible paths



HMM Overview



Observation List:



HMM Overview

A general definition of HMM

$$HMM = (S, V, B, A, \Pi)$$

N states

$$S = \{s_1, \ldots, s_N\}$$

M symbols

$$V = \{v_1, \ldots, v_M\}$$

Initial state probability:

$$\Pi = \{\pi_1, \ldots, \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1$$

π_i : *prob of starting at state* s_i

State transition probability:

$$A = \{a_{ij}\} \quad 1 \le i, j \le N \quad \sum_{j=1}^{N} a_{ij} = 1$$

a_{ij} : *prob of going* $s_i \rightarrow s_j$

Output probability:

$$B = \{b_i(v_k)\} \quad 1 \le i \le N, 1 \le k \le M \quad \sum_{k=1}^{M} b_i(v_k) = 1$$

$b_i(v_k)$: *prob of "generating"* v_k *at* s_i

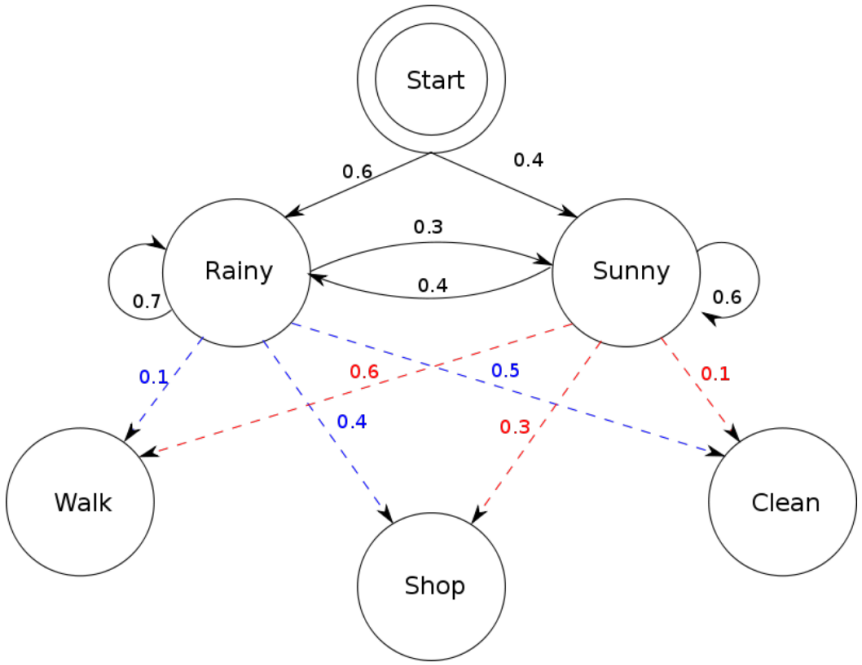
Three Basic Problems in HMMs

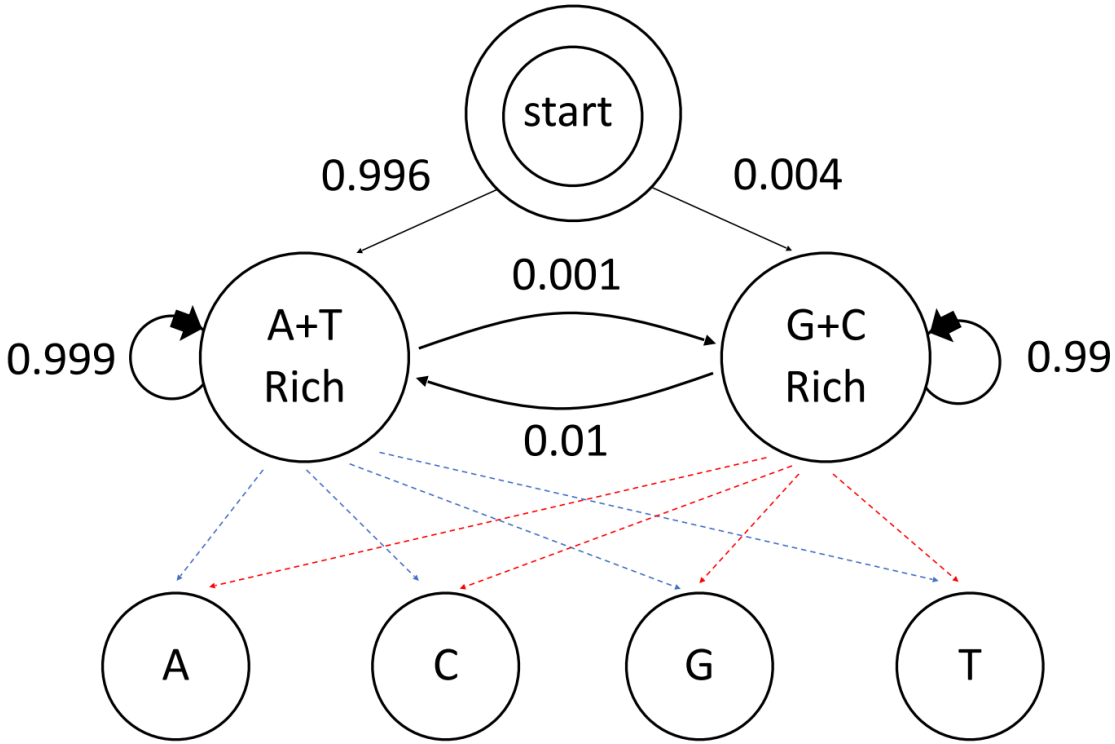
Given a set of observation sequences $O = O_1 O_2 \cdots O_T$ and the HMM parameters $\lambda = (A, B, \pi)$, computing the probability $P(O|\lambda)$

Given a set of observation sequences $O = O_1 O_2 \cdots O_T$ and the HMM parameters $\lambda = (A, B, \pi)$, computing the optimal state sequences

Given a set of observation sequences $O = O_1 O_2 \cdots O_T$ adjusting the HMM parameters $\lambda = (A, B, \pi)$ to maximize the probability $P(O|\lambda)$

Applying this Concept To Genomics/HW





A = 0.291
C = 0.209
G = 0.209
T = 0.291

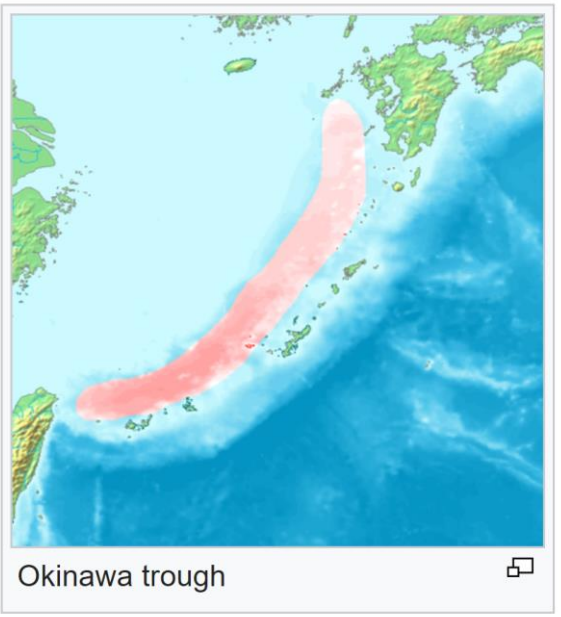
A = 0.169
C = 0.331
G = 0.331
T = 0.169

Homework 8 Overview

Pyrococcus horikoshii	
Scientific classification	
Domain:	Archaea
Kingdom:	Euryarchaeota
Phylum:	Euryarchaeota
Class:	Thermococci
Order:	Thermococcales
Family:	Thermococcaceae
Genus:	*Pyrococcus*
Species:	***P. horikoshii***
Binomial name	
Pyrococcus horikoshii Erauso *et al.* 1993	

Pyrococcus horikoshii

- Hyperthermophile, 98 C !
- Anaerobic archaeon
- Isolated from Okinawa Trough
- Growth enhanced by Sulfur
- 32 min doubling time (growth rate)


Okinawa trough

package, 1996). The phylogenetic tree diagrams were generated by the PHYLIP suite of programs (Kuhner and Felsenstein 1994).

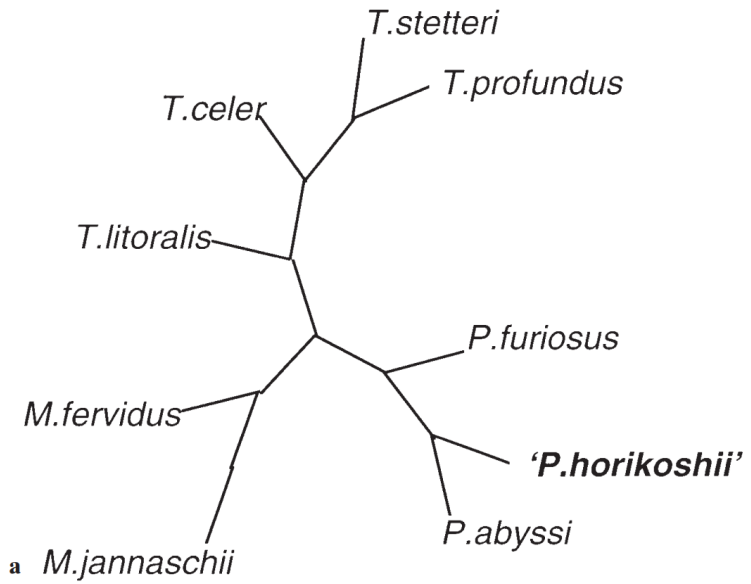


Fig. 3. Phylogenetic trees representing the relatedness of *P. horikoshii*

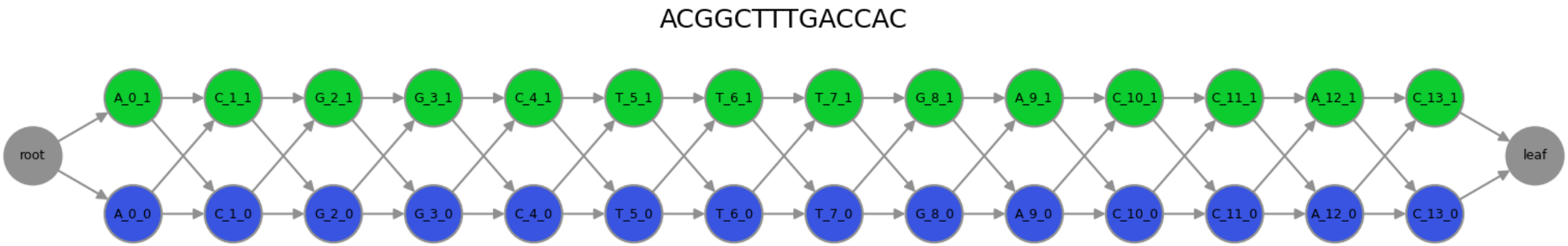
González, J., Masuchi, Y., Robb, F. et al. *Extremophiles* **2**, 123–130 (1998). https://doi.org/10.1007/s007920050051

Homework 8 Overview

- 2-state HMM for detecting GC-rich regions in *Pyrococcus horikoshii* genome
  - state 1: AT-rich, state 2: GC-rich
  - starting parameters are given: initiation, transition, emission probabilities

- Use Baum-Welch training to find improved parameter estimates
  - each iteration: compute log-likelihood of the sequence, new probabilities

- Run until parameter estimates converge
  - stop when log-likelihood increase < 0.1



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

Homework 8 Overview

Baum-Welch Algorithm

Baum–Welch is an expectation-maximization algorithm that uses the forward–backward algorithm.

1. Use the **forward algorithm** to calculate the forward probabilities for the HMM.

2. Use the **backward algorithm** to calculate the backward probabilities for the HMM.

3. Re-estimate transition, emission, and initial probabilities by calculating the expected number of each edge type

4. Calculate the new log likelihood of the model (the likelihood of our observations given our re-tuned model)

5. Repeat until the change in log likelihood is smaller than a given threshold or when a maximum number of iterations is passed.

Homework 8 Overview

Re-estimation of parameters

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time } (t = 1) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{\substack{t=1 \\ \text{s.t. } O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

Homework 8 Overview

Underflow – Important!

COURSE-RELATED MATERIALS:

- Math Notation
- Biological Review Slides: Gene and genome structure in prokaryotes and eukaryotes and characteristics of sequence data; Genbank and other sequence databases.
- Nature paper on human genome sequence
- Nature paper on mouse genome sequence
- Siepel et al. paper on PhyloHMMs & sequence conservation
- Rabiner tutorial on HMMs
- HMM scaling tutorial (Tobias Mann)

Problem: numbers too small to be stored in a variable

Solutions:

- Scale weights to be close to 1
  - affects all paths by same constant factor, which can be multiplied back later

- Use log weights, so can add instead of multiplying
  - Ex: Instead of 0.0001 * 0.0002, you can do: log(0.0001) + log(0.0002)

What about when you need to sum probabilities in logspace?

- See this blogpost for a solution or Tobias Mann
- https://gasstationwithoutpumps.wordpress.com/2014/05/06/sum-of-probabilities-in-log-prob-space/

Homework 8 Overview

Notes for debugging

1. Try calculating some simple forward and backward probabilities by hand to check your algorithm

2. **The likelihood at each iteration should increase; if it decreases, then you have a bug**

3. Have a print statement in your program to keep track of iterations as your program is running. The assignment will provide an estimate on the number of iterations to converge.

Homework 8 Questions ?

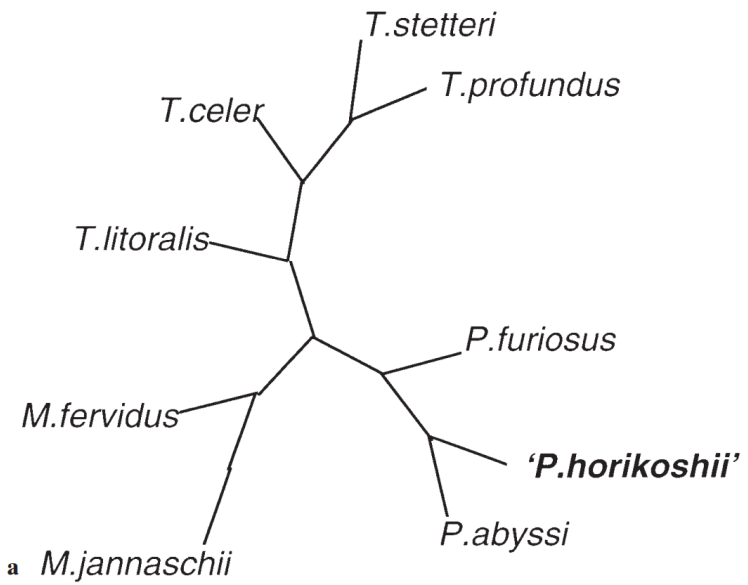
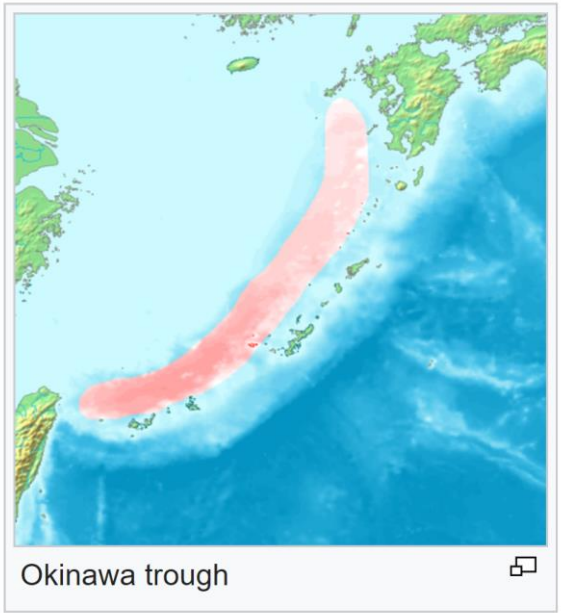
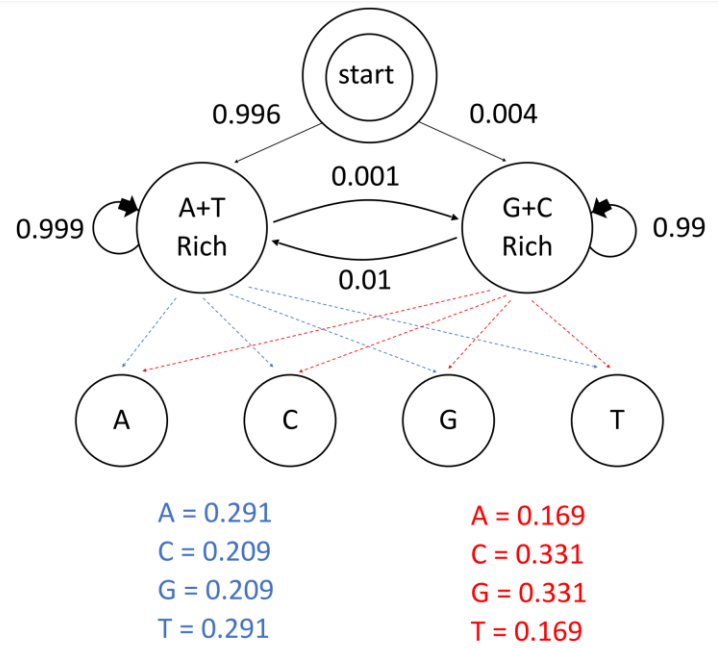
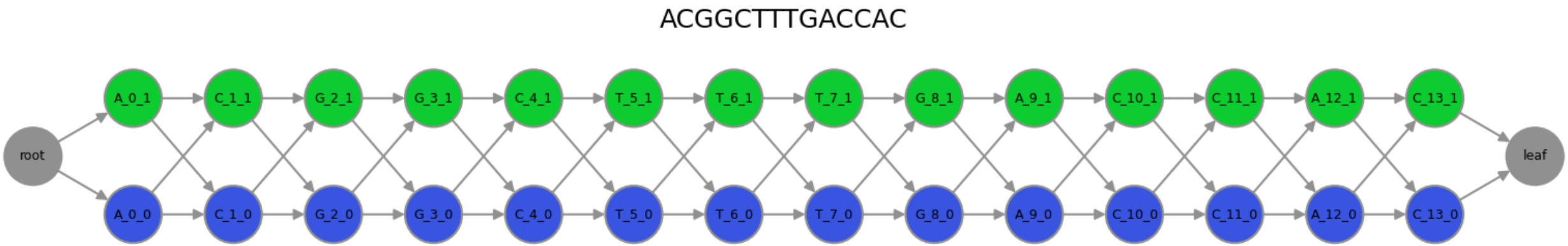


Fig. 3. Phylogenetic trees representing the relatedness of *P. horikoshii*







Reminders

- Homework 8 due this Sunday Mar. 5, 11:59 pm
- Homework 9 will be posted by tomorrow

. . .