BMI/CS 576 – Day 20

- Today
  - Training HMMs in the completely observed case
  - Applying HMMs to real data
- Next week
  - Clustering

Mid-semester survey

- 15 responses – Thank you to the responders!
- Diversity in responses for improvements
- Quizzes
  - primary purpose is to incentivize completing *all* reading/lecture prior to class period
  - not meant to comprehensively assess understanding
  - will increase time limits a bit to make it less stressful

HW4

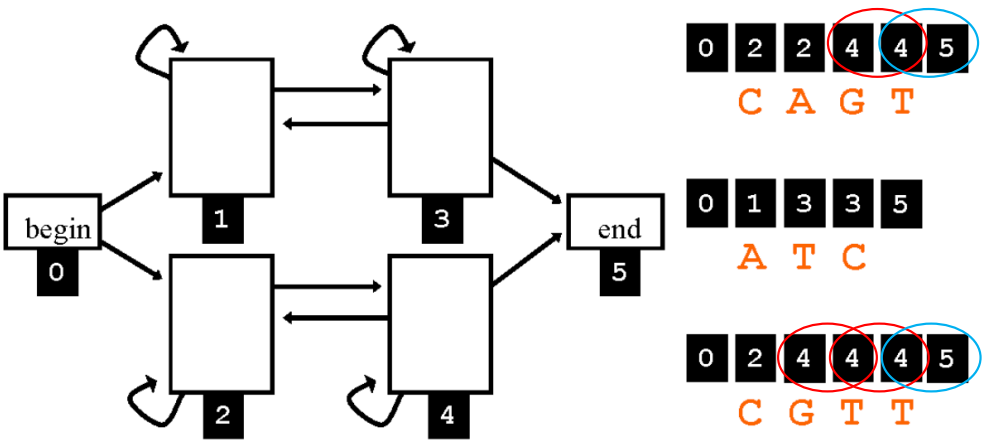
- Posted soon (likely tomorrow)
- Due before Thanksgiving
- HMMs
  - Use HMM class that we have been developing
  - Baum–Welch
  - Designing an HMM for a biological application

HMM applications lecture

- Video up shortly
- Meant to give you a sense of the diversity of applications of HMMs in molecular biology
- You will **not** be responsible for understanding the specifics of these applications
- **But:** thinking about these various HMMs should reinforce your understanding

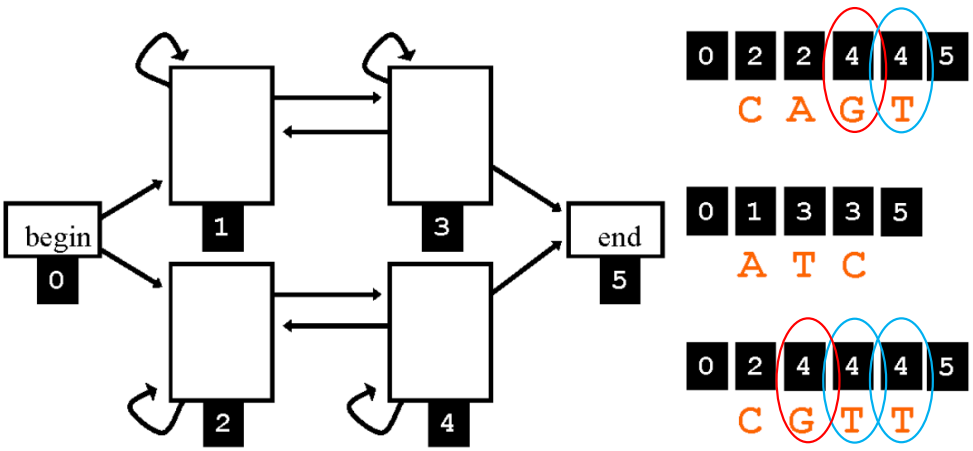
Quiz

Given the HMM below and three training sequences for which we have the true hidden state paths, what is the maximum likelihood estimate for the transition probability from state 4 to itself? Your answer should be a decimal number between 0 and 1.



$$\hat{a}_{44} = \frac{n_{4\to4}}{\sum_m n_{4\to m}} = \frac{n_{4\to4}}{n_{4\to4}+n_{4\to5}} \quad \frac{3}{3+2} = 0.6$$

Emission parameter estimate example



$$\hat{e}_4(G) = \frac{n_{4,G}}{\sum_c n_{4,c}} = \frac{n_{4,G}}{n_{4,A}+n_{4,C}+n_{4,G}+n_{4,T}} = \frac{2}{0+0+2+3} = 0.4$$

or with a pseudocount of 1 (Laplace estimates):

$$\hat{e}_4(G) = \frac{n_{4,G}+1}{\sum_c (n_{4,c}+1)} = \frac{n_{4,G}+1}{n_{4,A}+n_{4,C}+n_{4,G}+n_{4,T}+4} = \frac{3}{0+0+2+3+4} \approx 0.33$$

Baum–Welch

- Worked example of one iteration using quiz HMM
  - On Canvas (Day 20 section of Modules)
- Initialization
  - Can pick initial parameter values at random
  - If you have some knowledge of roughly what the parameter values should be (or at least which parameters are larger than others) use some informed guesses
    - May be more likely to reach a global optimum or at least a more plausible set of parameter values