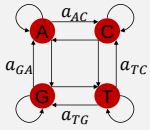
1

BTRY 4840/6840, CS 4775
Computational Genetics and Genomics



September 13, 2018

2

Announcements

- Problem set 2 out
  - Skeleton code that deals with input/output forthcoming

3

Today's lecture

- Finish sequence alignment
  - Affine gaps
  - Optimizations
- Hidden Markov models (HMMs)
  - Example
  - Definition of Markov chain, HMM
  - Viterbi algorithm

4

Affine gaps

5

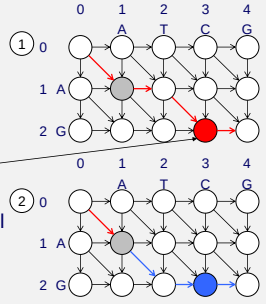
Affine gap penalties

- A bit more difficult than linear $\gamma(g) = -d - (g-1)e$
- Need to know whether a prior site is a gap or not
  - If previous alignment is match, gap opening penalty: $-d$
  - If previous alignment is gap in $\boldsymbol{x}$, gap extension penalty: $-e$
  - If previous alignment is gap in $\boldsymbol{y}$, gap extension penalty: $-e$
- Proposal: inspect trace back pointer of prior site and use e if so, d if not Does not work
  - Only have gap at prior site if gap length ≥1 better than match
  - Need to know best possible path that includes gap at prior site in order to decide on whether to extend or start new gap
- Solution: track best path with gap at given site
  - Instead of F, have M, I_x, I_y – latter two always gapped

6

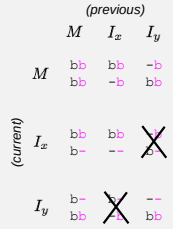
Example: one matrix insufficient for affine gaps

- Consider two paths
- Score for ① is $F(1,1) - d + s_{G,C} - d$
- Score for ② is $F(1,1) + s_{G,T} - d - e$
- If $s_{G,C} > s_{G,T}$, then $F(2,3) = F(1,1) - d + s_{G,C}$
- But if $s_{G,C} - d < s_{G,T} - e$, this assignment not optimal
- ➢ Does not "look forward" at $F(2,3)$. Need I_x, I_y



7

Affine gap recurrence relations



$$M(i,j) = \max \begin{cases} M(i-1,j-1) + s_{x_i,y_j} \\ I_x(i-1,j-1) + s_{x_i,y_j} \\ I_y(i-1,j-1) + s_{x_i,y_j} \end{cases}$$

$$I_x(i,j) = \max \begin{cases} M(i-1,j) - d \\ I_x(i-1,j) - e \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i,j-1) - d \\ I_y(i,j-1) - e \end{cases}$$

- Above is for Needleman-Wunsch
- Runtime complexity? $O(nm)$, but ~3× as long

8

Affine gap base cases

- Initialization: $M(0,0) = I_x(0,0) = I_y(0,0) = 0$
- Boundaries:

$$I_x(i,0) = I_x(i-1,0) - e, \quad 1 \le i \le n$$
$$I_y(0,j) = I_y(0,j-1) - e, \quad 1 \le j \le m$$
$$I_x(0,j) = -\infty, \quad 1 \le j \le m$$
$$I_y(i,0) = -\infty, \quad 1 \le i \le n$$
$$M(0,j) = M(i,0) = -\infty$$

- Final score:

$$\max_a S(x,y,a) = \max(M(n,m), I_x(n,m), I_y(n,m))$$

a an alignment

9

Optimizations

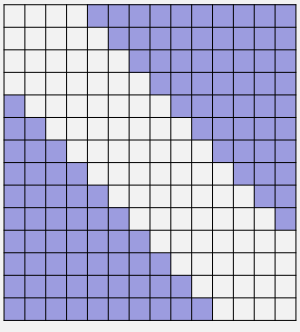
10

Efficiency in runtime, space usage

- Runtime of these alignment algorithms:
  - $O(nm)$
- Space usage:
  - $O(nm)$
- Can we do better?

11

Bounded dynamic programming



- Key idea: $F(i,j)$ defined only for bounded region
- Initialization:
  $F(i,0)$, $F(j,0)$ undefined outside bounded region
- Iteration for $|i-j| < k$:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s_{x_i,y_j} \\ F(i-1,j) - d \text{ if } j-(i-1)<k \\ F(i,j-1) - d \text{ if } i-(j-1)<k \end{cases}$$

- Runtime, space requirements?
  - $O(nk)$

12

Markov chain and Hidden Markov models

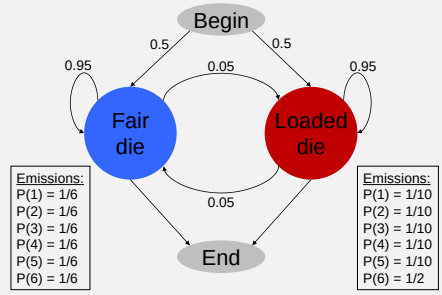
13

Example: dishonest casino

- Casino uses two dice:
  - Fair: P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6
  - Loaded: P(1) = P(2) = P(3) = P(4) = P(5) = 1/10
    P(6) = ½
- Casino switches dice on average every 20 rolls

14

Markov model of dice



15

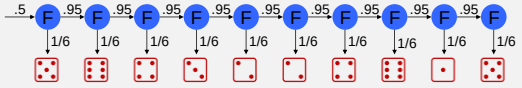
Probability of fair die for series of die rolls

- Suppose we observe 10 die rolls:

- Problem: assuming all rolls are from the same die, was it fair or loaded?
  - Will evaluate likelihood ratio

16

Calculating probabilities of hidden states

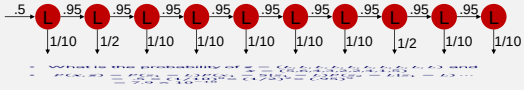


- What is the probability of $\boldsymbol{z} = (F, F, F, F, F, F, F, F, F, F)$ and $\boldsymbol{x} = (5,6,4,3,2,2,4,1,5)$
- $P(\boldsymbol{x}, \boldsymbol{z}) = P(z_1 = F)P(x_1 = 5|z_1 = F)P(z_2 = F|z_1 = F)\cdots$
  $= .5 \times (1/6)^{10} \times (.95)^9$
  $= 5.2 \times 10^{-9}$
- Why is this probability so small? Any particular sequence is rare
- What is the probability I have a dollar bill with serial number C 00273441 G?

➢ Likelihood ratio comparing two models still very informative

17

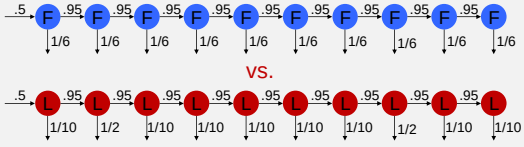
Calculating probabilities of hidden states



- What is the probability of $\boldsymbol{z} = (L, L, L, L, L, L, L, L, L, L)$ and $\boldsymbol{x} = (5,6,4,3,2,2,4,1,5)$
- $P(\boldsymbol{x}, \boldsymbol{z}) = P(z_1 = L)P(x_1 = 5|z_1 = L)P(z_2 = L|z_1 = L)\cdots$
  $= .5 \times (1/10)^8 \times (1/2)^2 \times (.95)^9$
  $= 7.9 \times 10^{-10}$

18

Compare the two possibilities



- Have $P(\boldsymbol{x}, \text{all fair}) = 5.2 \times 10^{-9}$
  $P(\boldsymbol{x}, \text{all loaded}) = 7.9 \times 10^{-10}$
- Likelihood ratio: $\frac{P(\boldsymbol{x}, \text{all fair})}{P(\boldsymbol{x}, \text{all loaded})} = 6.58$

➢ Nearly 7× more likely that the die was fair!

19

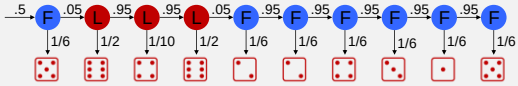
Comparison for different sequence

- For different sequence: $\boldsymbol{x} = (6,2,3,6,6,5,6,4,6,1)$
- Then $P(\boldsymbol{x}, \text{all fair}) = 5.2 \times 10^{-9}$ (same as before)
  $P(\boldsymbol{x}, \text{all loaded}) = .5 \times (1/10)^5 \times (1/2)^5 \times (.95)^9$
  $= 9.8 \times 10^{-8}$
- Likelihood ratio: $\frac{P(\boldsymbol{x}, \text{ all loaded})}{P(\boldsymbol{x}, \text{ all fair})} = 18.8$
  - Very likely to be loaded, unlikely to be fair

20

Likelihood of die being switched

- Now using different $\boldsymbol{z}$:



- Probability of $\boldsymbol{z} = (F, L, L, L, F, F, F, F, F, F)$
- $P(\boldsymbol{x}, \boldsymbol{z}) = P(z_1 = F)P(x_1 = 5|z_1 = F)P(z_2 = L|z_1 = F) \cdots$
  $= .5 \times (1/6)^7 \times (1/10)^1 \times (1/2)^2 \times (.95)^7 (.05)^2$
  $= 7.8 \times 10^{-11}$
- This state path unlikely, but
  - What about others state paths $\boldsymbol{z}$?
  - Other observed outcomes $\boldsymbol{x}$?

Want general way to analyze data given HMM specification

21

Desired uses of HMMs

- Evaluation:
  - **Given**: observed $\boldsymbol{x}$ and HMM specification
    Question: what is the joint probability of $\boldsymbol{x}$ and a given $\boldsymbol{z}$?
    Question: what is the likelihood of $\boldsymbol{x}$ based on the HMM?
- Decoding:
  - **Given**: observed $\boldsymbol{x}$ and HMM
    Question: what sequence of hidden states produced $\boldsymbol{x}$?
  - Viterbi decoding: most likely hidden state sequence
  - Posterior probability of hidden states: probability of each state z_i producing each x_i
    - Technically not a decoding: not path of states, but probabilities
- Learning:
  - **Given**: observed $\boldsymbol{x}$ and HMM without complete probabilities
    Question: what emission, transition probabilities produced $\boldsymbol{x}$?

22

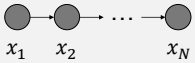
Markov chain (model) definition

- Markov chain: sequence of states at given times
  - In genomics, we often think of "time" as *position* on sequence
- Formally, Markov chain defined by
  - Set of states S: individual state at given time i denoted x_i
  - Matrix A of state transitions probabilities
    Element $A[k,l] = a_{kl} = P(x_i = l|x_{i-1} = k)$
  - Initial state probabilities $a_{0k} = P(x_1 = k)$
  - (Formulations vary: states sometimes states denoted π_i)
- Key feature: all states *observed* (not so in HMMs)
- Markov property: state x_i depends only on x_{i-1}, so
  - $P(\boldsymbol{x}) = P(x_1, x_2, \ldots, x_N)$
    $= P(x_N|x_{N-1})P(x_{N-1}|x_{N-2}) \cdots P(x_2|x_1)P(x_1)$

23

Graphical model representation of Markov chain

- Using graphical model notation:
  - Each node is a random variable
  - Shaded nodes are observed
  - White (unfilled) nodes are hidden
  - Arrows represent conditional dependence



  - Above implies
    $P(x_1, x_2, \ldots, x_N) = P(x_1)P(x_2|x_1) \cdots P(x_N|x_{N-1})$
    $= P(x_1) \prod_{i=2}^{N} P(x_i|x_{i-1})$
- This is 1st order Markov model: x_i depends on x_{i-1}
  - For K^{th} order Markov model, x_i depends on $x_{i-K}, \ldots x_{i-1}$

24

HMM notation (somewhat different in Durbin)

- States: $\boldsymbol{z} = (z_1, z_2, \ldots, z_L)$, for L observations
  - Examples: $\boldsymbol{z} = (F, F, F, F, F, F, F, F, F, F)$ – rolls from fair die
    $\boldsymbol{z} = (L, L, L, L, L, L, L, L, L, L)$ – rolls from loaded die
- Observations: $\boldsymbol{x} = (x_1, x_2, \ldots, x_L)$
  - Example: $\boldsymbol{x} = (5,6,4,3,2,2,4,1,5)$
- Transition probabilities $a_{kl} = P(z_i = l|z_{i-1} = k)$
  - Probability of moving to state l from previous state k
  - Example: $a_{FL} = P(z_i = L|z_{i-1} = F) = 0.05$
- Emission probabilities $e_k(b) = P(x_i = b|z_i = k)$
  - Probability of observing b given current state k
  - Example: $e_L(6) = P(x_i = 6|z_i = L) = 1/2$
- Initial probabilities $a_{0k} = P(z_1 = k)$

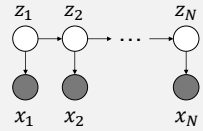
25

Hidden Markov model definition

- Formally, Hidden Markov model defined by
  - Set of states S: state at time/position i denoted z_i
  - Set of possible observations O: observation i denoted x_i
  - Matrix A of state transitions probabilities
    Element $A[k,l] = a_{kl} = P(z_i = l | z_{i-1} = k)$
  - Initial state probabilities $a_{0k} = P(z_1 = k)$
  - Emission probabilities $e_k(b) = P(z_i = b | z_i = k)$
- Key features:
  - States hidden (unobserved)
  - Markov property holds: state z_i depends only on z_{i-1}
    - (For first order HMM: most common type)
  - Observation x_i depends only on current state z_i

26

Graphical model representation of HMM



So:

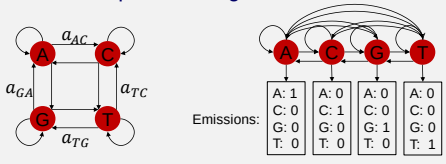
$$P(\boldsymbol{x}, \boldsymbol{z}) = P(x_1|z_1)P(z_1)\prod_{i=2}^{N} P(x_i|z_i)\,P(z_i|z_{i-1})$$

Note: in practice will want to evaluate
$P(\boldsymbol{x}) = \sum_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z})$ and/or $\max_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z})$
(i.e., where $\boldsymbol{z}$ unknown)

27

Genomic Markov chain, HMM

- Can model sequence using Markov chain, HMM



28

Hidden Markov models

CpG island example

29

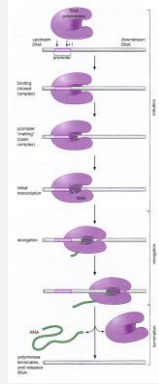
Biology background: CpG dinucleotide

- Dinucleotide sequence CG is typically written CpG
  - p for phosphate (between bases in DNA backbone): emphasizes this is dinucleotide, not base pairing
  - Can also talk about GC content: % G or C nucelotides in region, not dinucleotides
- CpG dinucleotides:
  - Cytosine in CpGs are often methylated
  - When methylated, have high rate of C→T mutations
    - Methylation: addition of methyl group (in this case to cytosine)
    - In mammals, 70-80% of CpG cytosines are methylated
  - Consequently CpGs are rarer in genome than expected

30

Gene promoter

- Gene promoter is:
  - Sequence where transcription is initiated
    - May or may not be transcribed, but near transcription start site (TSS)
  - Between ~100-1000 bp long
  - Subsequence bound by transcription factors:
    - A protein (i.e., product of a gene) that binds a specific DNA sequence
    - Recruits RNA polymerase, and thus controls the rate of transcription
      - In eukaryotes, often works in tandem with other elements (activators, repressors, others)



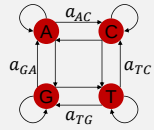
31

Biology background: CpG islands

- Methylation is suppressed in promoters, other regions of the genome
- Such regions have high rate of: CpG, GC content
  - Called CpG islands
- Problems:
  - Given short sequence, is it from a CpG island?
  - Given (long) genome sequence, locate CpG islands (regions with likely biological importance)
- How would you address the first question?

32

Can train Markov chains for CpG island / not



- Suppose we were given data labeled as *CpG island* or *not CpG island*
- Could use this to train a model
  - That is, set the parameters in the model based on the data

33

Final notes

- Summary – sequence alignment
  - Affine gaps possible, more work than linear
  - Optimize runtime via bounded dynamic programming
- Summary – hidden Markov models (HMMs):
  - Flexible tool to evaluate wide range of data (including genetic)
  - Can use to evaluate, decode, and learn parameters of HMMs
  - More detailed HMM example: CpG islands