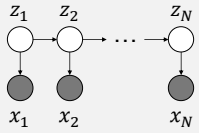
1

BTRY 4840/6840, CS 4775
Computational Genetics and Genomics



September 18, 2018

2

Announcements

- Problem set 2 due in one week
- Fill out survey on my office hours:
  - Currently 83% say they can attend 4:30-5:30pm, ≤50% for other times

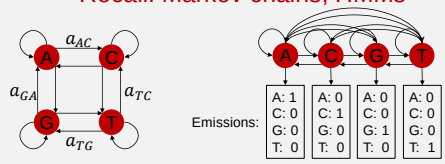
3

Today's lecture

- Continuing Hidden Markov models (HMMs)
  - Viterbi algorithm
  - Forward algorithm
  - Forward/backward algorithm
  - Posterior state probabilities and posterior decoding

4

Recall: Markov chains, HMMs



- Difference between Markov chain, Hidden Markov model (HMM):
  - In Markov chain, states are observable, states unobservable in HMM
  - Both have: states, initial probabilities, transition probabilities
  - HMM has emission probabilities

5

Hidden Markov models
CpG island example

6

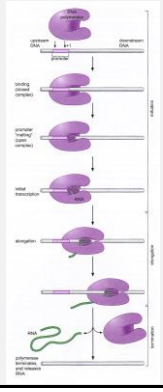
Biology background: CpG dinucleotide

- 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

7

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)



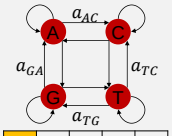
8

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?

9

Can train Markov chains for CpG island / not



+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

-	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

- 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
- Approach: use two Markov chains
  In CpG island: + model; not island: - model
- How would you assign transition probabilities in these DNA Markov chains?

Is MLE

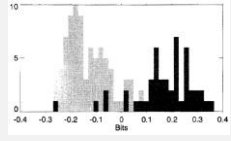
$a_{kl}^{+} = \frac{c_{kl}^{+}}{\sum_{l'} c_{kl'}^{+}}, c_{kl}^{+}$ dinucleotide counts

10

Have two opposite models: want to classify

- Can compute likelihood of a given input sequence under both models: $P(\boldsymbol{x}|\text{model}+)$, $P(\boldsymbol{x}|\text{model}-)$
- How should we discriminate between these?
- Use log-odds ratio:

$$S(x) = \log\frac{P(\boldsymbol{x}|\text{model}+)}{P(\boldsymbol{x}|\text{model}-)} = \sum_{i=1}^{N} \log\frac{a_{x_{i-1}x_i}^{+}}{a_{x_{i-1}x_i}^{-}}$$



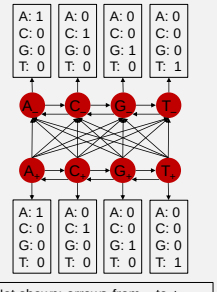
11

Problem 2: locate CpG islands in sequence

- Given (long) genome sequence, locate CpG islands
- How?
  - Could divide up sequence into 100 bp windows, compute log odds scores, call windows with positive scores CpG islands
  - Not ideal:
    - Why 100 bp? CpG islands vary in length
    - In general won't start/end at window boundary
- Instead: use hidden Markov model
  - Observed DNA is derived from some underlying hidden biological process that impacts the DNA sequence we do see
  - Can incorporate Markov models that give expected frequency of dinucleotides between bases within CpG islands/non-island

12

CpG island hidden Markov model

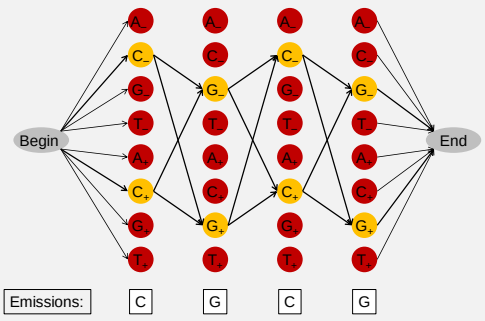


Not shown: arrows from – to + some arrows between +/– states

- Can merge the two Markov models, make states hidden
  - States get relabeled, with +, –
  - Both "emit" A,C,G,T
- Transition probabilities:
  - Based on those in Markov chains, but need probability of switching: + to – and – to +
- Why such a large number of states?
  - Why not use only two: +/ – ?
  - Need memory of previous nucleotide to model dinucleotides

13

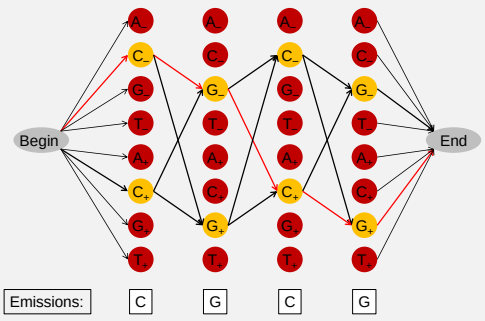
Numerous possible paths through states



Want to know most likely state path for given observed sequence

14

What if the path was given to us?



Easy to calculate probability of path/observations (e.g., loaded/fair die earlier)

15

How do we find the most likely path?

- Can compute probability of a given path, how can we find the best one?
- Dynamic programming!
  - Highest probability path to a given state is product of:
    a) Highest probability path to all previous states (subproblem)
    b) Transition probability to from previous state to given state
    c) Emission probability
- Viterbi algorithm / Viterbi decoding
  - Terminology: HMM decoding: path of states that produce some sequence (note: not necessarily *correct* path)

16

Hidden Markov models

Viterbi decoding: maximum likelihood path

17

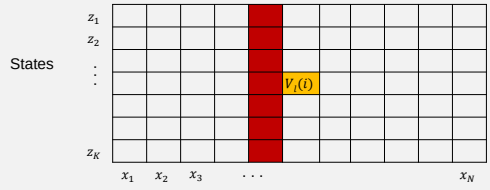
Goal: maximum likelihood decoding

- Want to maximize joint probability $P(\boldsymbol{x}, \boldsymbol{z})$ over $\boldsymbol{z}$
  - Recall: $\boldsymbol{x}$ – sequence of observations; $\boldsymbol{z}$ – sequence of states
- For $\boldsymbol{x}, \boldsymbol{z}$ given, we have:

$$P(\boldsymbol{x}, \boldsymbol{z}) = a_{0z_1} \cdot e_{z_1}(x_1) \cdot \prod_{i=2}^{N} e_{z_i}(x_i) \cdot a_{z_{i-1}z_i}$$

18

Viterbi algorithm summary

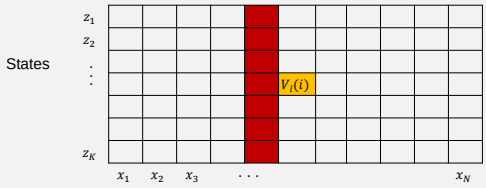


- Let $V_k(i-1)$ be the probability of the highest probability path to state k, up to observation x_{i-1}
- Assume we know $V_k(i-1)$, for all k
  How do we get $V_l(i)$ for a given l?

$$V_l(i) = e_l(x_i) \cdot \max_k (V_k(i-1) \cdot a_{kl})$$

19

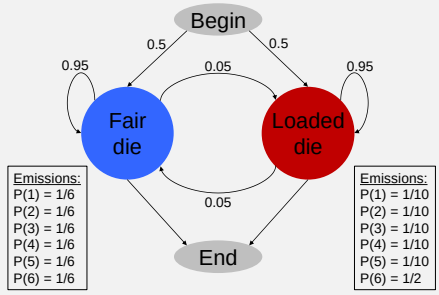
Viterbi algorithm



- Initialization: $V_k(1) = a_{0k}e_k(x_1) = P(z_1 = k)P(x_1|z_1 = k)$
- Iteration: $V_i(i) = e_i(x_i) \cdot \max_k(V_k(i-1) \cdot a_{kl})$
- Final value: $\max_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z}) = \max_k V_k(N)$
- How do we get path? Trace back: keep back pointers to max
- Runtime, Space complexity: $O(K^2N)$, $O(KN)$, respectively

20

Example: occasionally dishonest casino



21

Example: Viterbi path for dishonest casino

- 300 rolls generated by sampling Markov chain
- Viterbi decoding shown – what do you notice?

22

Desired uses of HMMs (highlighted done)

- 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 $\boldsymbol{x}$?

23

Hidden Markov models

What is the likelihood of the data $P(\boldsymbol{x})$ based on a given parameterization of an HMM?

Forward algorithm

24

Law of total probability gives $P(\boldsymbol{x})$

- Want to know the probability that the data was generated by the model
- From law of total probability, have:

$$P(\boldsymbol{x}) = \sum_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z})$$

- We know how to compute:
  - $P(\boldsymbol{x}, \boldsymbol{z})$ for a given $\boldsymbol{z}$: multiply transition, emission probabilities
  - $\max_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z})$: Viterbi algorithm
- How do we compute $P(\boldsymbol{x})$?
  - Exponential number of paths
  - Could just use Viterbi to approximate, but that's only one path
  - ➢ Dynamic programming still works

25

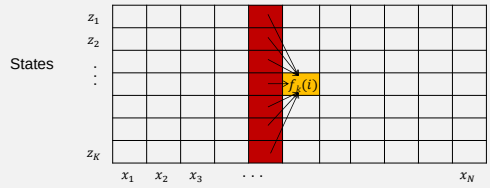
Forward probability

- Will compute iteratively
- Let $f_k(i) = P(x_1, \ldots, x_i, z_i = k)$: forward probability
- Given $f_j(i-1)$ at some step, the following holds:

$$f_k(i) = e_k(x_i) \sum_{j=1}^{K} f_j(i-1) \cdot a_{jk}$$

26

Forward probability calculation



- Initialization: $f_k(1) = a_{0k} e_k(x_1) = P(z_1 = k) P(x_1 | z_1 = k)$
- Iteration: $f_k(i) = e_k(x_i) \cdot \sum_{j=1}^{K} f_j(i-1) \cdot a_{jk}$
- Final value: $P(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(N)$
- Runtime, Space complexity: $O(K^2 N)$, $O(KN)$, respectively

No trace back: not about one path

27

Applications of forward algorithm

- Forward probabilities are used in:
  - Forward-backward algorithm to compute posterior probability of each state
  - Posterior decoding
  - Unsupervised learning via Baum-Welch
  - Sampling a state path from the distribution $P(\boldsymbol{z}|\boldsymbol{x})$
- In principle can use $P(\boldsymbol{x})$ to compare two different HMMs, but in practice this is uncommon
  - Instead, one chooses hidden states and sets parameters (transition/emission probabilities) from theory or via learning

28

Desired uses of HMMs (highlighted done)

- 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 $\boldsymbol{x}$?

29

Hidden Markov models

What is $P(z_i = k|\boldsymbol{x})$?

Forward-backward algorithm

30

Computing $P(z_i = k|\boldsymbol{x})$

- Want $P(z_i = k|\boldsymbol{x})$
- By definition of conditional probability, we have

$$P(z_i = k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, z_i = k)}{P(\boldsymbol{x})}$$

  - Forward probability gives $P(\boldsymbol{x})$
  - How do we compute the numerator?

$$\begin{aligned} P(\boldsymbol{x}, z_i = k) &= P(x_1, \ldots, x_i, z_i = k) P(x_{i+1}, \ldots, x_N | x_1, \ldots, x_i, z_i = k) \\ &= P(x_1, \ldots, x_i, z_i = k) P(x_{i+1}, \ldots, x_N | z_i = k) \\ &= f_k(i) \cdot \underbrace{P(x_{i+1}, \ldots, x_N | z_i = k)}_{b_k(i)} \end{aligned}$$

31

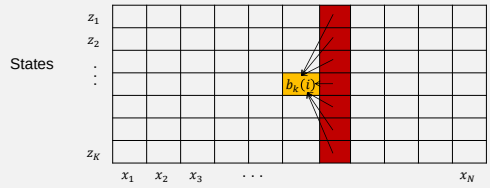
Backward probability

- Similar to forward probability, but calculated using observed data *after* a given step i
- Let $b_k(i) = P(x_{i+1}, \ldots, x_N | z_i = k)$: backward probability
- Given $b_k(i+1)$ at some step, the following holds:

$$b_k(i) = \sum_{l=1}^{K} a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$$

32

Backward probability calculation



- Initialization: $b_k(N) = a_{k0}$ (typically 1)
- Iteration: $b_k(i) = \sum_{l=1}^{K} a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$
- Note: $b_k(i)$ does not include emission probability for step i
- Final value: $P(\boldsymbol{x}) = \sum_{k=1}^{K} a_{0k} e_k(x_1) b_k(1)$

Analogous to forward
- Computes $P(\boldsymbol{x})$
- No trace back

33

Final notes

- Summary:
  - Hidden Markov models: very general framework for analyzing data from a given model
    - Viterbi algorithm
    - Forward algorithm