HMM Applications

Announcements

- Homework 5 on the CourseWeb site.
- Due Monday Nov. 7 (a week from today).
- A shorter assignment than usual, because of your midterm.
- Lecture next Monday will be presented by Eun Kang (your TA), introducing sequence alignment algorithms. (I have to be at NIH that day).
- We decided not to include the Viterbi algorithm in the material covered by the midterm.
- Videos of my lectures from the 2009 course are collected on **thinking.bioinformatics.ucla.edu/teaching**, in case that's helpful.

An Epsilon Test

A researcher interested in a hypothesis h performs an experimental observation X, and finds that $p(X_1|h) < 0.05$ for the observation X_1 from his first experiment. He now wishes to define a rigorous test for rejecting the hypothesis: if for any confidence level ε, no matter how small, he can obtain $p(X_1, X_2, \ldots X_n|h) < \varepsilon$ by simply repeating the experiment some finite number of times n, the hypothesis is rejected. Another researcher insists that this test is invalid due to possible bias. Who is right? Justify your answer mathematically.

Epsilon Test Lessons

- Many of you had the right idea (that the likelihood would go down exponentially with increasing sample size n), but didn't consider how that invalidates this proposed test.
- Many others focused on the question of whether the observations were independent, which is a valid concern (since the question did not say they were).
- A few of you reversed the conditional probability, i.e. interpreted $p(X_1, X_2, \cdots, X_n|h)$ as $p(h|X_1, X_2, \cdots, X_n)$.

All those responses missed the big picture here, namely that the whole idea of this test is wrong, because the likelihood cannot be interpreted on an absolute scale, only *relative* to another likelihood model for the same observations.

Use Posterior Odds Ratios

Take-home lesson: to avoid this problem, compare two different models using their *odds ratio*.

- based on a p-value test:

$$\frac{p(h|t^+)}{p(h_0|t^+)} = \frac{p(t^+|h)p(h)}{p(t^+|h_0)p(h_0)}$$

- based on regular likelihood:

$$\frac{p(h|obs)}{p(h_0|obs)} = \frac{p(obs|h)p(h)}{p(obs|h_0)p(h_0)}$$

This is just Bayes' Law expressed in ratio form.

Basic SNP Scoring P-value

You are scoring a candidate SNP under the same assumptions as before: N reads from a pool of multiple people (assume each read is untagged and comes from a different person), with a fixed sequencing error rate ε. Any site where a "mutant" basecall b' is observed (in addition to the "reference" basecall b) is considered a candidate SNP. Propose an extreme value test t^+ that ensures a false positive rate $p(t^+|h^-) = \alpha$, where h^- means "there is no SNP at this position".

P-value Lessons

- Most of you realized that the null hypothesis should consider a binomial event with probability ε, i.e. the sequencing error rate.
- Poisson model? No, that applies to events that occur at a uniform density over continuous time intervals. Here sequencing error is a binomial event (error / no error) that occurs in discrete trials (i.e. the individual basecall).
- Many of you didn't consider what the observable metric for the extreme value test should be, to distinguish real SNPs from pure sequencing error. To make an extreme value test, you need such a metric: in this case, the total number of b' observations.
- Many of you did not think of this in terms of an explicit likelihood model $p(k|\varepsilon, N)$. As soon as you do so, the p-value test becomes dead easy: $p(K \geq k|\varepsilon, N) \leq \alpha$.

Why an Extreme Value Test?

Because *likelihood* cannot be interpreted on an absolute scale, we want to convert it to a form that *can*.

- $p(k|\varepsilon, N)$: Likelihood! Could be "big" (e.g. 0.1) or "small" (e.g. 10^{-25}) but those values are not meaningful in themselves.
- $p(K \geq k|\varepsilon, N)$: p-value! No longer subject to the arbitrariness of the likelihood of a single value of k. Instead we are measuring the probability of *all* values of k above some threshold.
- $p(K \leq k|\varepsilon, N)$: also a p-value. Useful if you're looking for values of k much *smaller* than expected under your null hypothesis.

Make sure you get the direction right!

Modeling CpG Islands as a Markov Chain

CpG islands show very different conditional probabilities $p(X_{t+1}|X_t)$ than non-CpG island sequence. This suggests we can use a Markov chain model to detect them in any sequence.

- Given a 4 x 4 table of the conditional probabilities $p(X_{t+1}|X_t)$ measured in CpG islands, describe how you would construct a Markov chain model of CpG island sequences. Specifically, describe the state graph structure you would use:
  - what are its nodes?
  - what edges exist?
  - what will you use as the transition probabilities associated with the edges?

CpG Markov Model Lessons

Most of you got this right or very close...

- One mistake: confused the state graph with the information graph. Do not mix these up!
- Some of you made extra assumptions that were not in the question, e.g. that CpG island can only contain C and G, or that the question was asking you to model non-CpG island sequences as well. Make sure to read the question carefully.
- Some of you proposed dinucleotide states (e.g. CC, CG). That is not needed, because the conditional probability $p(X_{t+1}|X_t)$ directly captures these dinucleotide relationships. In other words, the dinucleotide aspect of this problem is captured by the *edges* not the *nodes*.

Deriving the Transition Matrix

Given a table of the 16 dinucleotide frequencies $p(X_t, X_{t+1})$ measured for CpG islands, indicate how you would derive the transition matrix for your Markov chain model.

Deriving the Transition Matrix Lessons

Again, most of you got this right.

- a few of you did not distinguish between joint probability i.e. $p(X_t, X_{t+1})$ vs. conditional probability i.e. $p(X_{t+1}|X_t)$. You were given the former and needed to convert it to the latter.
- Some of you started by writing the right principle but didn't finalize it.

Remember the chain rule and the summation rule (eliminate a variable from a joint probability). They'll help you solve problem after problem.

Uniform Probabilities?

Say we wish our model of the Occasionally Dishonest Casino to have the same hidden state probability $p(X_t = F)$ at all times t. Assuming the transition probabilities are $\tau_{FL} = p(L|F) = 0.05$ and $\tau_{LF} = p(F|L) = 0.1$, how can we achieve this goal? (If this is not possible with the given data, just say "Insufficient data").

Uniform Probabilities Lessons

- about half of you understood that this Markov chain will tend to spend about a 2:1 ratio of its time in the F vs L states.
- One thing that may be confusing is that the stationary distribution applies to two quite different ways of computing the distribution of a Markov chain:
  - *time average*: e.g. if we were to generate a single sequence of states, as the sequence length $L \to \infty$, the ratio $n_i/L \to \pi_i$ (where n_i is the count of occurences of state s_i in the sequence). This is the average of a single sequence over time.
  - *population average*: if we were to generate N such sequences, $n_{ti}/N \to \pi_i$ (where n_{ti} is the count of occurences of state s_i at time t), for large sample size $N \to \infty$ and long times $t \to \infty$. (This true regardless of what the initial distribution at time $t = 1$ was.) This is the average of the whole population at a single time point.

Note that *both* of these converge to the stationary distribution.

Population Averaged Distribution

- Whenever someone talks about "the distribution at time t" $p(X_t)$, they are talking about a population-averaged distribution.
- E.g. the balance equation is easiest to interpret in terms of $p(X_t), p(X_{t+1})$

$$\sum_i \pi_i \tau_{ij} = \sum_i \pi_j \tau_{ji}$$

- Note that we can set any distribution we want as the starting distribution $p(X_1)$.
- At later times t it will converge to the stationary distribution, i.e. $p(X_t) \to \vec{\pi}$ as $t \to \infty$.
- The distributions $p(X_t)$ at different times t will be different! (unless $p(X_1) = \vec{\pi}$).

Viterbi subpaths Lessons

- About 2/3 of you got this right!
- **Take-home Lesson**: remember that Viterbi optimality is proved by induction, i.e. we can only prove that the viterbi path $\vec{V}^t$ is optimal if we assume that we knew the optimal path $\vec{V}^{t-1}$ for *each* possible state of X_{t-1}.

Predicting CpG Islands Using an HMM

You wish to use your Markov chain model of CpG islands to predict which sequence regions in the human genome are actually CpG islands. You are given dinucleotide frequencies for non-CpG island sequences, and the probabilities of transitioning from CpG island to non-Cpg island at any nucleotide, and vice versa.

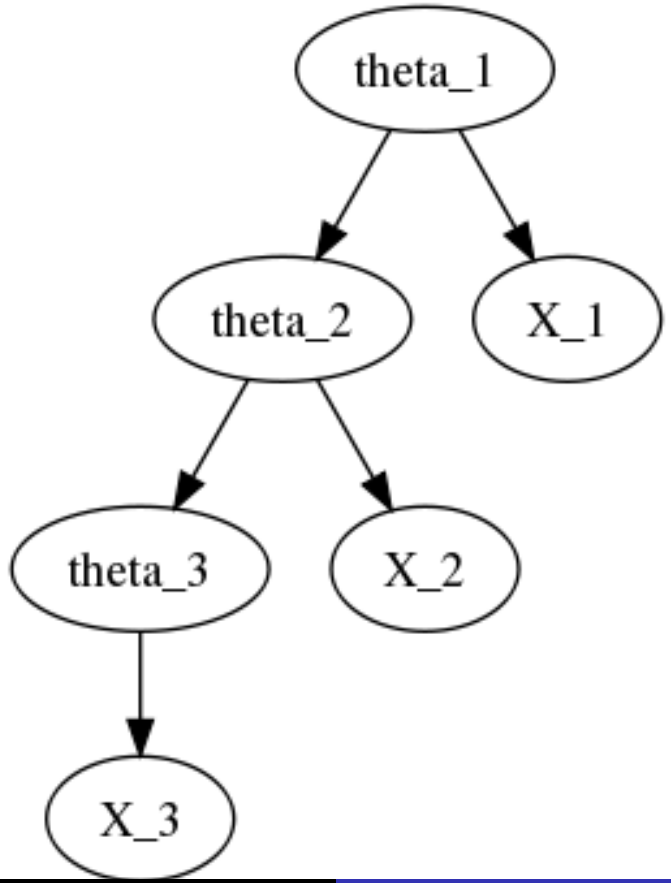
- describe the state graph of your HMM (you do not need to *draw* all its details; just state its structure unambiguously).
- draw an information graph of your HMM, including both the hidden and observable variables associated with three consecutive nucleotides in the genome sequence.

CpG Islands HMM Lessons

- a common error: working with state graphs has scrambled your understanding of information graphs. You are mixing up these two totally different things.
- Remember in the first week, we made a big fuss about distinguishing **variables** vs. **states**? It's a simple distinction, and I think all of you got it. But you may not be fully convinced of the *importance* of strictly observing this distinction.
- You are now seeing why it's important:
  - info graph = **variables**
  - state graph = **states**
- If you made these kinds of mistakes, please go back and review the variable vs. state definitions, and make sure this is 100% clear in your mind.

HMM Information Graph

- Remember I told you the information graph of any Markov chain is so simple, it's boring?
- It is always $X_1 \to X_2 \to X_3 \to \cdots$.
- An HMM is just the same thing decorated with associated observables:



Why not just two hidden states, CpG vs. not-CpG?

- we can create a CpG state (+), and a not-CpG state (-).
- Each state can have different emission probabilities, e.g. CpG could give much higher probability to emit C and G.
- **However**, such a model cannot capture the dinucleotide frequencies seen in CpG islands (e.g. of seeing G after C), because it does not include the nucleotide in the *state* (instead it is handled by the emission probability).
- Such a model is only capable of modeling nucleotide *composition* ("mononucleotide frequencies" if you wish).
- Because we need to model dinucleotide frequencies, we were forced to include the nucleotide in the states that we condition on e.g. A+.

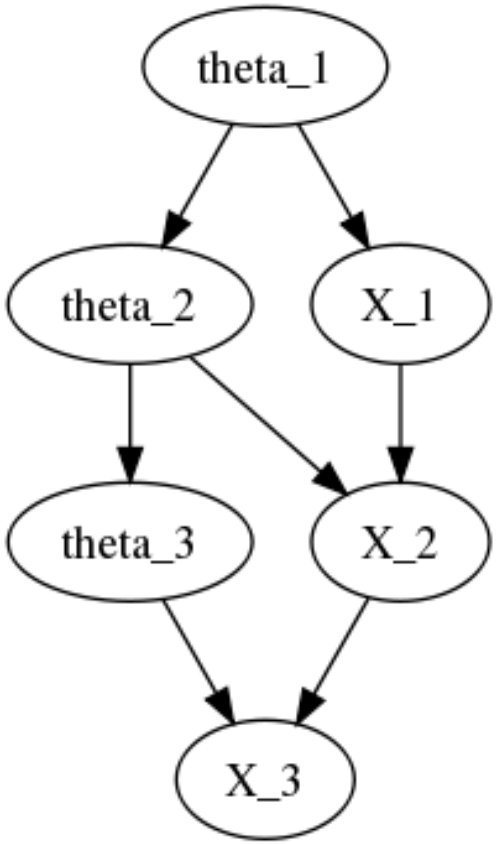
A Better Idea

- one of you suggested that this could be addressed by having each hidden variable θ_t emit *dinucleotide* probabilities e.g. $p(X_{t-1}, X_t|\theta_t)$.
- This is a good idea! But we must be careful not to have any variable X_t appear more than once as a subject variable.
- The right way to do this is by altering the emission probability to $p(X_t|\theta_t, X_{t-1})$.
- The payoff is that now we only need two states for our hidden variable: CpG (+) vs. not-CpG (-). Compare that with 8 states in the original model! That makes the computation 16 times faster!
- The *only* difference in the resulting probabilities is the handling of the first nucleotide in any CpG / not-CpG block. In the original model we used the stationary distribution; the new model uses the conditional probability as if the previous nucleotide were also part of the block. In practice, a negligible difference.

Simpler-CpG Info Graph

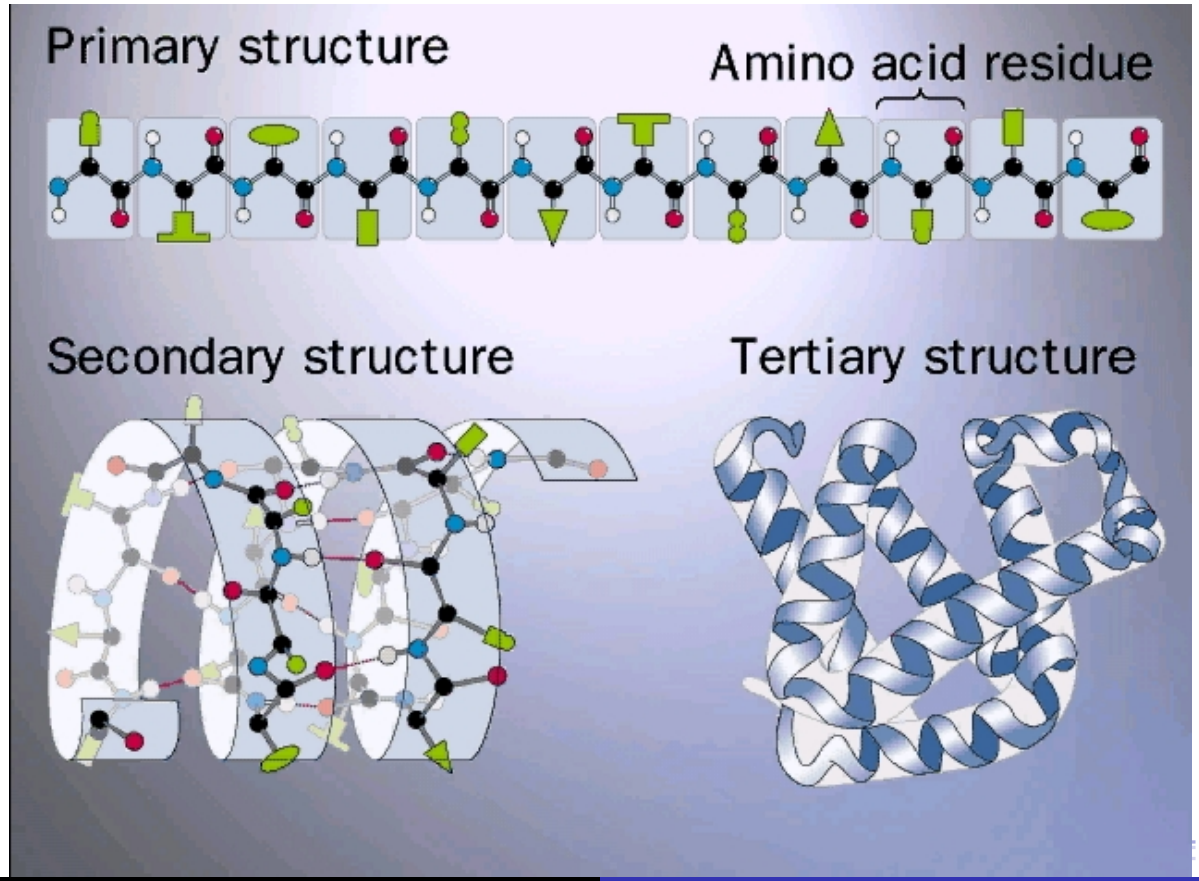
Then the joint probability looks like

$$p(\vec{X}^n, \vec{\theta}^n) = \cdots p(\theta_t|\theta_{t-1})p(X_t|\theta_t, X_{t-1})\cdots$$



Example: Secondary Structure Prediction

- Proteins are chains (strings) of a 20 letter alphabet (amino acids).
- The chain is flexible and automatically “folds” into three-dimensional structures based on its specific amino acid sequence.
- The simplest structural elements are α helix and β sheet.



Example: Secondary Structure Prediction

- When someone finds a new protein sequence, they often want to predict its secondary structure.
- Different secondary structures have different characteristic amino acid compositions.
- Simple prediction problem: given a protein sequence $\vec{X}^n$, predict H (helix) vs. S (sheet) vs. C ("random coil", which just means not helix or sheet) for each letter in the protein.
- What model structure? (Info graph; state graph)
- What data do we need to build the HMM?
- To make predictions and measure confidence, what do we need to compute?

Forward-Backward Probabilities

State the probabilistic meaning, for an HMM consisting of n hidden variables, of the forward probability of state $\Theta_1 = s_j$.

Forward-Backward Probabilities Answer

$$f_{1,j} = p(\vec{O}^1, \theta_1 = s_j) = p(\theta_1 = s_j)p(O_1|\theta_1 = s_j)$$

Forward-Backward Probabilities

State the probabilistic meaning, for an HMM consisting of n hidden variables, of the backward probability of the START state.

Forward-Backward Probabilities Answer

$$b_{0,START} = p(\vec{O}_1^n | \theta_0 = \text{START}) = p(\vec{O}^n)$$

Example: Profile Alignment HMM

- Being able to assign a new sequence to a known protein family is crucial for infering its function.
- Being able to align it accurately to other members of the family indicates which parts of the new sequence carry out different aspects of the function of that family.
- Tools for aligning to individual sequences in the database (one at a time, e.g. BLAST) are often less sensitive than a “profile” that captures the detailed patterns of conservation of a protein family (in other words, the many sequences in that family).
- These patterns show you what’s important vs. not, for recognizing a new member of the family.

Example: Profile Alignment HMM

- Say we want to "align" a new sequence $\vec{X}^n$ to a profile HMM.
- What is the info graph for a profile HMM?
- What is the state graph for a profile HMM?
- What parameters are required to build the model?
- What will the "variable index" t mean in this HMM?
- How does the HMM track which part of the profile emitted a given part of the new sequence?
- To identify which parts of the new sequence are confidently aligned to the profile, what must we compute?
- in big-O notation, how long will it take us to compute this for the whole sequence?

Profile HMM