Alignment Algorithms

Midterm Results

- Big improvement over scores from the previous two years.
- Since this class' grade is based on the previous years' curve, that means this class will get higher grades than the previous years.
- Your reward for the "extra pain" of the new in-class exercises:
  - you've learned more than previous classes did;
  - you're getting better grades than previous classes did.

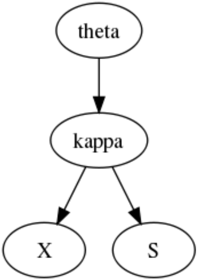
A Few Lessons From the Midterm

The main problem I see is students *trying to remember the "right answer"* rather than *thinking about what the question is asking*.

- Example: $\sum_Y p(X|Y) = ?$
- If you believe this sums to 1 (or $p(X)$), why not at least check an example? e.g. p(rain|day-of-week). People who think things through for themselves get in the habit of looking for easy ways to check their answer, because they've learned this is easy and valuable.
- You could easily have checked those possible answers against the equations you were provided on the equation sheet.

Example: Two Sisters Model

Many of you proposed the following information graph for the model that assumes the two women are sisters:



But this assumes they are the same person!

- Even before taking this class, everyone here knew that sisters are not genetically identical. The only way I can interpret this answer is that you were trying to *remember* the match model from the forensic test problem you saw before, rather than *thinking* about what this question was asking.

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 Probability Lessons

- Hopefully, homework 5 helped you learn the forward-backward algorithm by implementing it.
- Please review the following to distinguish them clearly in your minds:
- $p(\theta_t|\vec{X}^n)$ (posterior): **this is what we want to know.** The forward-backward algorithm computes this in two halves:
- $p(\theta_t, \vec{X}^t)$ (forward probability): the first half.
- $p(\vec{X}^n_{t+1}|\theta_t)$ (backward probability): the second half.
- Many of you mixed up forward vs. posterior, or left out the observable completely.

$$p(\theta_t, \vec{X}^n) = p(\theta_t, \vec{X}^t) p(\vec{X}^n_{t+1}|\theta_t)$$

Forward Probability for t=1?

- Some of you ignored the specific question about the t=1 case. Sounds like regurgitating what you read rather than thinking about the question you were asked.
- Some of you referred to the stationary distribution rather than the prior. No. $p(\theta_1)$ is the prior. Only equals $\vec{\pi}$ if you specifically set that to be true.
- Some of you referred to $p(\theta_1|\vec{X}_1)$ as "the likelihood that observation X1 was emitted by state theta1", or defined the forward probability as a *posterior* $p(\theta_1|X_1)$. You're reversing the conditional probability; watch out!
- One of you defined the forward probability as $p(\theta_1, \theta_2, ..., \theta_t)$.

Forward-backward is easy to remember (and understand) if you just bear in mind what the **goal** is: $p(\theta_t|\vec{X}^n)$: just *one* hidden variable, and *all* the observables.

Forward-Backward Probabilities

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

Backward Probability Lessons

Two things confused you:

- the START state;
  - think of this as "before time t=1". i.e. $p(\theta_0 = \text{START}) = 1$.
  - lets us handle priors just like a regular transition, i.e. $\tau_{START,i} = p(\theta_1 = s_i | \theta_0 = \text{START}) = p(\theta_1 = s_i)$.
  - Some of you gave $p(\vec{X}_{n+1}^{n} | \theta_n) = 1$ as your answer. Sounds like the end rather than START.
- the backwards probability.
  - Some of you seem to associate the backwards probability with θ_n... but the question specifically asked about the backwards probability of the START.
  - as before, some of you left the observables out of the backward probability. *Remember the goal*!

Backwards to START?

- One of you wondered how it could even be possible to go “backward” to the START state, because I said “the HMM never goes back to the START state.”
- Does $b_{0,START} > 0$ mean there are directed edges that point to the START state?
- No. The word “backward” is confusing you here. The “backward probability” does not mean that there are directed edges pointing “backward” (e.g. to START), but rather that $p(\vec{X}^n_{t+1}|\theta_t)$ must be computed in *reverse order* (last edge first) in order to be computationally efficient.
- (You could compute it in forward order, but the computational complexity would be horrible. Try it yourself to see why.)
- Think of this as analogous to the Viterbi “backtracking” stage: i.e. we are not following edges that *point backwards*, but rather we are traversing in *reverse order* edges that point forwards.

Good Nomenclature can save your @#$!

How are you going to keep the concepts clear in your mind, if you don't have clear, consistent ways of naming them?

- **to specify a variable**: "time index" t, e.g. θ_t.
- **to specify a state**: "state index" i, e.g. s_i.
- **to specify both**: e.g. $f_{ti} = p(\theta_t = s_i, \vec{X}^t)$
- **emission** (likelihood): $p(X_t|\theta_t)$.
- **transition**: $p(\theta_t = s_j|\theta_{t-1} = s_i) = \tau_{ij}$
- **posterior**: $p(\theta_t = s_j|\vec{X}^n)$

Sequence Alignment Matrix

- Pairwise alignment of two sequences $\vec{X}^m, \vec{Y}^n$
- Index letter positions in the two sequences with t, u, e.g. X_t, Y_u.
- Hypothesis: these two sequences are related to each other by some evolutionary process, e.g. mutation from a common ancestor: $\vec{A} \rightarrow \vec{X}$ and $\vec{A} \rightarrow \vec{Y}$.
- Want to be able to assess this hypothesis and gain insights from it, e.g.
  - calculate posterior odds vs. null hypothesis that they are unrelated;
  - predict which parts of the two sequences are "the same", i.e. aligned, and likely to perform same function. Easy when the two sequences are very similar, much harder when similarity weak.

Edit Operators

- In standard pairwise alignment, what are the allowed “edit operators” that transform one sequence into the other?
- Describe how each of these edit operations are represented on a sequence alignment matrix.

Alignment Paths

- Using the three edit operator "moves" $\leftarrow, \swarrow, \downarrow$, write the list of all possible paths representing global alignments of $\vec{X}^2$ to $\vec{Y}^1$. (For simplicity, you can write the moves as -, /, and |).
- If any of these paths correspond to the same alignment (i.e the same letter(s) of $\vec{X}^m, \vec{Y}^n$ are aligned to each other), group those paths together as one group.

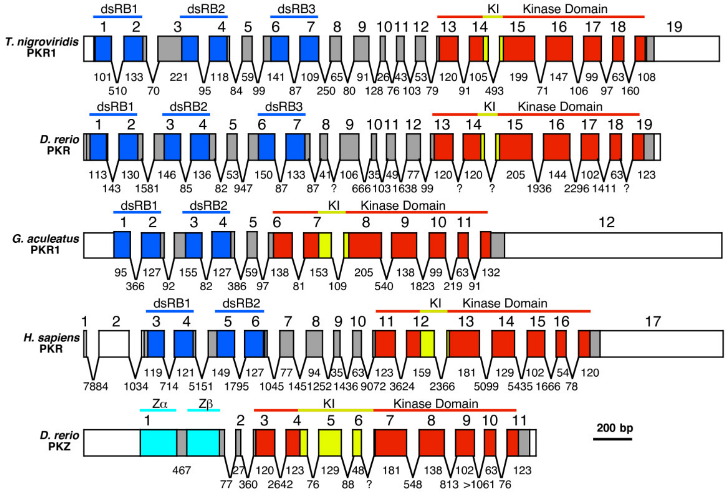
Viterbi Alignment

- Given the two sequences $\vec{X} = AGC\ldots$ and $\vec{Y} = ACC\ldots$
- Find the optimal global alignment of $\vec{X}^2$ to $\vec{Y}^1$ by filling out the corresponding Viterbi alignment matrix.
- Assume the following scores:
  - exact match: +5
  - substitution: -2
  - gap: -3

Alignment Scoring

- Say you are given the likelihood $p(X_t, Y_u|\text{aligned})$ derived from "true alignments" of sequences that are known to be evolutionarily related to each other.
- Based on the hypothesis testing principles we have learned, propose a scoring function for assessing whether X_t, Y_u should be aligned to each other or not.
- How does your scoring function indicate they *should* be aligned vs. *should not* be aligned?

Complex Gene Structure



Searching for regulatory sequences in introns

- You are studying introns in mammalian genes, looking for regulatory elements in orthologous introns.
- One important indicator of regulatory element function is *selection pressure against mutation*, i.e. finding regions of aligned introns from different mammals, that are mutated much less than the usual.
- This requires good intron alignments.
- You know that introns are poorly conserved and consequently hard to align accurately.

Intron Alignment Idea

- You have a database of matching *exons* in human and mouse (i.e. pairs of coding region segments in the human genome and mouse genome that align to each other optimally).
- A colleague suggests you use this matching exon information to find optimal alignments of matching introns as follows:
- take two consecutive human exons A, B (separated by an intron I) that are known to align to two consecutive mouse exons A', B' (separated by an intron I').
- Now find the top-scoring alignment of the two introns I, I' subject to the constraint that you already know the correct alignment of the exons on either side of the intron.

Intron Alignment

Assume that you know exactly the positions that are aligned at the end of exons A, A' and at the beginning of exons B, B'.

- What would be the easiest way to obtain the optimal intron alignment subject to the proposed constraint?
- State the proposed constraint in precise terms of the allowed paths on the alignment matrix.