Markov Chains and Hidden Markov Models

= stochastic, generative models

(Drawing heavily from Durbin et al., Biological Sequence Analysis)

BCH364C/391L Systems Biology / Bioinformatics – Spring 2015
Edward Marcotte, Univ of Texas at Austin

Markov Chains and Hidden Markov Models are important probabilistic models in computational biology

Some of their applications include:
- Finding genes in genomes
- Mapping introns, exons, and splice sites
- Identifying protein domain families
- Detecting distant sequence homology
- Identifying secondary structures in proteins
- Identifying transmembrane segments in proteins
- Aligning sequences

& outside biology, they have many uses, including:
- Speech, handwriting, and gesture recognition
- Tagging parts-of-speech
- Language translation
- Cryptanalysis

and so on....
The key idea of both of these types of models is that:

**Biological sequences can be modeled as series of stochastic (i.e., random) events.**

It’s easy to see how a random process might model stretches of DNA between genes and other important regions.

BUT, the idea of modeling something as structured and meaningful as a gene or protein sequence by a similar process might seem odd.

It’s important to realize exactly what we’re modeling.

The idea behind hidden Markov models is not that the sequence is random, but that the sequence we observe is one of many possible instances of some underlying process or object.

E.g., actin differs slightly from organism to organism.

Imagine an “ideal”, but unobservable, actin, defined by certain underlying specific physico-chemical properties important for its function. What we see in nature is not this ideal gene, but numerous instances of observed sequences, all just a bit different.

In the hidden Markov model, the underlying process or structure is represented as hidden, unobservable states and the observed sequences represent possible sequences compatible with these states.

We would say that the observed sequence is emitted from the hidden state.
Let’s start with an easier case: **Markov chains**

We’ll explore a simple non-biological model: a coin-toss

Flip a coin a bunch of times and observe the results, e.g.

TTTHTTHHTTHTHTHHHHHTTTTTTTHTHHTTTTHHHTHH

We could model this process as two states:
- H for heads,
- T for tails,
and the probability of switching between them:

A sequence is a walk along this graph:

H T H H T H H ...
With a fair coin:

The chance of seeing heads or tails is equal, and the chance of seeing heads following tails and vice versa is equal.

Therefore, the transition probabilities (corresponding to the arrows above) are:

<table>
<thead>
<tr>
<th>Position i:</th>
<th>Position i+1:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head</td>
</tr>
<tr>
<td>Head</td>
<td>0.5</td>
</tr>
<tr>
<td>Tail</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Important: All probabilities leading out of a state add up to 1!

With a biased coin (e.g. tails comes up 90% of the time):

The chance of seeing heads or tails is not equal, nor is the chance of seeing heads following tails and vice versa.

We might have the same model, but with skewed transition probabilities:

<table>
<thead>
<tr>
<th>Position i:</th>
<th>Position i+1:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head</td>
</tr>
<tr>
<td>Head</td>
<td>0.1</td>
</tr>
<tr>
<td>Tail</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Important: All probabilities leading out of a state add up to 1!
Now, imagine a scenario where the observed sequence of coin flips was actually generated by 2 coins, one fair and one biased.

To decide whether we are looking at a sequence of coin flips from the biased or fair coin, we could evaluate the ratio of the probabilities of observing the sequence by each model:

$$\frac{P(X \mid \text{fair coin})}{P(X \mid \text{biased coin})}$$

Does this remind you of something we’ve seen before? How might we test where the fair & biased coins were swapped along a long stretch of coin flips?

How might we test where the fair & biased coins were swapped along a long stretch of coin flips?

One way using our current Markov chain model is to calculate the ratio of probabilities (e.g. log odds ratio) in a *sliding window* along the sequence:

![Log odds ratio graph](image-url)
How about a biological application? A classic example is **CpG islands**

In animal genomes, the dinucleotide CG is strongly underrepresented (note: NOT the base pair C:G, but rather 5’-CG-3’)

Why? C’s are often methylated, and methylated C’s mutate at higher rates into T’s. So, over time, CG’s convert to TG’s EXCEPT around promoters, which tend not to be methylated.

Thus, **CpG ‘islands’ often indicate nearby genes.** Finding them was a classic method for annotating genes.

How could we make a CpG island finding model analogous to the fair/biased coin model?

**A CpG island model might look like:**

- **CpG island model:**
  - $p(C \rightarrow G)$ is higher
  - Could calculate $P(X | \text{CpG island})$

- **Not CpG island model:**
  - $p(C \rightarrow G)$ is lower
  - Could calculate $P(X | \text{not CpG island})$

( of course, need the parameters, but maybe these are the most important....)

(or log ratio) along a sliding window, just like the fair/biased coin test
In these simple models, called *Markov chains*, we don’t have hidden states.

BUT, we could have used a *hidden Markov model*:

Now, the underlying state (the choice of coin) is hidden. Each state *emits* H or T with different probabilities.

These are the *emission probabilities* for each state

The *transition probabilities* might be something like:

<table>
<thead>
<tr>
<th>Position i:</th>
<th>Position i+1:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fair</td>
</tr>
<tr>
<td>Fair</td>
<td>0.9</td>
</tr>
<tr>
<td>Biased</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Important questions we might like to ask:

1. Given an observed sequence and a model, what is the most likely sequence of hidden states?
   
i.e., what is the path through the HMM that maximizes $P(p,X|l)$, where $p$ is the sequence of states?)?

In our coin example, we might be given an observed sequence:

HTHTHTHTTTTTTTTTTTTTTTTTTTHTHTHTHTHT

and want to identify when the biased coin was used:

FFFFFFFFFFFFBBBFFFFBBBBBBBBBBBBBBBBBBFFFFFFFFFF

Answer: Use the Viterbi algorithm.
We’ll see this shortly.

---

Important questions we might like to ask:

2. Given a sequence of observations, can we calculate the probability that the sequence was derived from our model?
   
i.e., can we calculate $P(X|l)$, where $X$ is our observed sequence, and $l$ represents our HMM?

For example, we might want to know if a given protein sequence is a member of a certain protein family.

---

**Graphic Summary**
- Positive conserved domains have been detected, click on the image below for detailed results.
- These results are for globin proteins, specifically for globin-like superfamily.
- Distribution of 187 Blast Hits on the Query Sequence:
  - Query hits: 1-30, 31-60, 61-90, 91-120, 121-150, 151-180
  - Color key for alignment scores:
    - Red: low scores
    - Blue: high scores

---

E.g. as we saw before (although calculated a bit differently)
Important questions we might like to ask:

2. **Given a sequence of observations**, can we calculate the probability that the sequence was derived from our model? 
   *i.e.*, can we calculate \( P(X|I) \), where \( X \) is our observed sequence, and \( I \) represents our HMM?

For example, we might want to know if a given protein sequence is a member of a certain protein family.

**Answer:** Yes. Use the *forward algorithm*. We’ll see this shortly.

---

Important questions we might like to ask:

3. **Given a model**, what is the most likely sequence of observations?

For example, after having trained an HMM to recognize a type of protein domain, what amino acid sequence best embodies that domain?

**Answer:** Follow the maximum transition and emission probability at each state in the model. This will give the most likely state sequence and observed sequence.
Important questions we might like to ask:

4. How do we train our HMM?
   *i.e.*, given some training observations, how do we set the emission and transition probabilities to maximize $P(X|l)$?

Answer: If the state sequence is known for your training set, just directly calculate the transition and emission frequencies. With sufficient data, these can be used as the probabilities.

This is what you will do in Problem Set #2.

With insufficient data, probabilities can be estimated from these (e.g., by adding pseudo-counts).

If the state path is unknown, use the forward-backward algorithm (also known as the Baum-Welch algorithm).

Important questions we might like to ask:

5. How do we choose the best HMM topology from the many possible choices?

Answer: Good question. No great answer.

Often trial-and-error, and understanding the essential features of the system that you are modeling.
Each of these algorithms (the Viterbi, forward, and forward-backward) uses dynamic programming to find an optimal solution.

(just like aligning sequences)

Let’s revisit the CpG islands using an HMM:

- 8 states: one per nucleotide inside CpG islands (+) and one per nucleotide outside CpG islands (-)
- All possible transition probabilities are represented as arrows
- This is a particularly simple model: each state emits the nucleotide indicated with probability of 1 and has zero probability of emitting a different nucleotide.
Given a DNA sequence $X$ (e.g., CGATCGCG), how do we find the most probable sequence of states (e.g., ----++++)?

→ The Viterbi algorithm

We want to find the state path that maximizes the probability of observing that sequence from that HMM model.

Viterbi does this recursively using dynamic programming.

As with sequence alignment, we’ll construct a path matrix that captures the best score (i.e., highest probability) along a single path through the HMM up to each position. We’ll “grow” this matrix using a few simple recursion rule.

The rules (stated formally):

- $v$ is an entry in the Viterbi path matrix
- $e$ indicates an emission probability
- $a$ gives the transition probability between previous state $k$ and current state $l$
- $\text{pointer}(l) = \text{argmax}_k (v_k(l-1)a_{kl})$
- $\text{i.e., draw the pointer back to the entry that gave rise to the current best score}$

Initialization ($i=0$): $v_0(0) = 1$, $v_k(0)=0$ for $k > 0$

Recursion ($i=1$ to $L$): $v_i(l) = e(x_i) \max_k (v_k(i-1)a_{kl})$

Termination: $P(X,\rho^*) = \max_l (v_l(L)a_{l0})$

For each Viterbi matrix entry:
We try to maximize the product of prior score and transition from that state to this one. We then multiply that score times the emission probability for the current character.
Step 1: Initialize the path matrix.

Possible states

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>G</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A+</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C+</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G+</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>T+</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>A-</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For simplicity, let’s assume the transition probability from $B$ to each nucleotide is $1/8$. We’ll also ignore all transition probabilities except these for now:

Step 2: Calculate the elements of the $v_k$ matrix for $i = 1$.
Then keep going for $i = 2$, etc..

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>G</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A+</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C+</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>G+</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.13</td>
</tr>
<tr>
<td>T+</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>G-</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For simplicity, let’s assume the transition probability from $B$ to each nucleotide is $1/8$. We’ll also ignore all transition probabilities except these for now:

For example, the score $v_C(i=1) = 1 \cdot \max \{ 1^*1/8, 0^*a_{A,C}, 0^*a_{C,C}, ..., 0^*a_{T,C} \} = 1/8$
Step 3: Keep going for $i = 2$, etc..

The maximum scoring path scores 0.0032. The most likely state path is found by traceback from the 0.0032 to give C+G+C+G+.

In a longer sequence, the model would switch back & forth between CpG and non-CpG states appropriately.
Can this really work? Here’s a real example.

An HMM model of fair and loaded dice:

Reconstructing which was used when, using the Viterbi algorithm:

How do we calculate the probability of a sequence given our HMM model?

→ The forward algorithm

Subtle difference from Viterbi:
Viterbi gives the probability of the sequence being derived from the model given the optimal state path.

The forward algorithm takes into account all possible state paths.

Again, it does this recursively using dynamic programming.
The rules (stated formally):

Initialization ($i=0$):

\[ f_0(0) = 1, \quad f_k(0) = 0 \quad \text{for} \quad k > 0 \]

Recursion ($i=1$ to $L$):

\[ f_i(i) = e_i(x_i) \sum_k f_k(i-1)a_{ki} \]

Termination:

\[ P(x) = \sum_k f_k(L)a_{k0} \]

Note: No pointer! Just to calculate the probability of seeing this sequence from this model.

For each Viterbi matrix entry:

We try to maximize the product of prior score and transition from that state to this one. We then multiply that score times the emission probability for the current character.

---

A toy HMM for sequence alignment

Is this global or local alignment? How could you change the model to perform the other kind of alignment?

---

A toy HMM for 5′ splice site recognition (from Sean Eddy’s NBT primer linked on the course web page)

How might you design an HMM to recognize a given type of protein domain?
How might we design HMMs to recognize sequences of a given length?

What would this HMM produce?