Genome 373: Hidden Markov Models III

Doug Fowler
We talked about two decoding algorithms last time.

What is meant by decoding?
We talked about two decoding algorithms last time.

What were the two algorithms?
Viterbi vs F-B Algorithms

Viterbi

start: at the beginning of the sequence of symbols $x$

algorithm:

results:
Viterbi vs F-B Algorithms

Viterbi

**start:** at the beginning of the sequence of symbols $x$

**algorithm:** moving forward, find the likeliest path for each state in each position and discard all the rest

**results:**
Viterbi vs F-B Algorithms

Viterbi

**start:** at the beginning of the sequence of symbols $x$

**algorithm:** moving forward, find the likeliest path for each state in each position and discard all the rest

**results:** the most likely state path
Viterbi vs F-B Algorithms

Viterbi

**start:** at the beginning of the sequence of symbols $x$

**algorithm:** moving forward, find the likeliest path for each state in each position and discard all the rest

**results:** the most likely state path

This process is often referred to as **decoding** an HMM, because it reveals the most likely sequence of (hidden, encoded) states
Viterbi vs F-B Algorithms

F-B Algorithms

**start:** at the $i^{th}$ position

**algorithm:**

**results:**
Viterbi vs F-B Algorithms

F-B Algorithms

**start:** at the $i^{th}$ position

**algorithm:** moving forward or backward, sum the probabilities of paths leading to a particular state $\pi_i = k$

**results:**
F-B Algorithms

**start**: at the $i^{th}$ position

**algorithm**: moving forward or backward, sum the probabilities of paths leading to a particular state $\pi_i = k$

**results**: the probability that the model was in state $k$ at position $i$

The F-B algorithms can be used to solve many other **decoding** problems (e.g. find the most probable state at position $i$)
Outline

• A splice site finding HMM

• An ORF finding HMM

• A gene finding HMM
A HMM For Finding 5’ Splice Sites

• We want to draw a model that will identify the G in a 5’ splice site in a DNA sequence

• What are the states that will be important for the model (i.e. what is on either side of a 5’ intron splice site)?

• How should we arrange them?

• Where should our transition arrows go?

• What will each state emit?

• Assort into groups of 4 or 5 and develop a solution in ~5m

• Nominate someone to come up and draw/explain your solution in ~1m
OK, now that we have a model (and were given a sequence and transition/emission probabilities), what is our next step?
A HMM For Finding 5’ Splice Sites

Find the most likely state path, which enables us to label the sequence with the most likely location of exon, splice site and intron!
Outline

• A splice site finding HMM

• An ORF finding HMM

• A gene finding HMM
A HMM For Finding ORFs

- We want to draw a model that will identify the ORFs in a sequence of arbitrary length. Don’t worry about introns for now.

- What starts an ORF? What ends one?

- What should be the order of the features we care about?

- Where should our transition arrows go?

- What will each state emit?

- Permute groups (nobody stays together), develop a solution in ~5m

- Nominate someone to come up and draw/explain your solution in ~1m
A HMM For Finding ORFs

I came up with one that explicitly models all three nucleotides of the start and stop codons.
Outline

• A splice site finding HMM
• An ORF finding HMM
• A gene finding HMM
A Gene Finding HMM

• Draw a model of an HMM that can find genes

• The model will work on DNA sequence and should have states that describe genes.

• Try to think of all the states you will need as a first task (e.g. exon, intron, promoter, etc, etc).

• Then, try to organize/connect the states in such a way that transitions between them are biologically meaningful

• Finally, think about how you would find transition/emission probabilities for your model!

• Permute groups again. Take ~10m and come up with a model. Nominate someone to come up and draw/explain your solution in ~1m
A Gene Finding HMM

This is from Burge and Karlin’s GENSCAN paper and is a good solution to the problem.
This is from Burge and Karlin’s GENSCAN paper and is a good solution to the problem

Note that their model takes into account:

- the fact that genes can occur on both DNA strands
- The fact that introns/exons have phase based on where codons are interrupted
Where Do We Get Transition Probabilities?

One of the most common ways to get transition probabilities for a HMM is by learning them from training sequences.

Training sequences are sequences where we know the states (for gene finding, a set of known genes and their sequences).

In essence, we just use the emission frequencies from each state and the transition frequencies between states.

This is an example of machine learning, which is what we will cover next!
Three ways to think about HMMs

Evaluation

Given a HMM $M$ and a sequence of symbols $x$: Find $P(x | M)$

Or given a model and a sequence, calculate the probability of that particular sequence arising from the model.

Decoding

Given a HMM $M$ and a sequence of symbols $x$: Find the sequence of states $\pi$ that maximizes $P(x, \pi | M)$

Learning

Given a HMM $M$ with unspecified transition/emission probabilities and sequence(s) $x$

Find the likeliest transition/emission probabilities
Markov Model References/Reading

Biological Sequence Analysis, Durbin et al. 2010 Chapter 3

An Introduction to Bioinformatics Algorithms, Jones and Pevzner 2004


What is a Hidden Markov Model?, Eddy, 2004, 22, pg 1315-1317