Hidden Markov Models and BioInformatics, Part I

Steven R. Dunbar

March 9, 2017

Outline

Background

- Prof E. Moriyama (SBS) has a BioInformatics Seminar
- SBS, Math, Computer Science, Statistics
- Extensive use of program "HMMer"
- Britney (Hinds) Keel thesis (2015)
- Yixiang Zhang thesis (2016)

Tentative Titles

- HMMer for Mathematicians
- HMM and HMMer
- HMMer for Dummies
- HMM and DMMer
- Hidden Markov Models and BioInformatics

Brief History

- Hidden Markov Models developed in the 1960s and 1970s for satellite communication (Baum, Viterbi, 1967).
- HMMs adapted for general communications engineering in 1960s, 1970s.
- HMMs adapted for speech recognition in the 1970s and 1980s (Bell Labs, IBM)
- HMMs in computational biology (Krogh et. al., 1994)
- HMMer 1.0 - 3.1 (Eddy et al. 1998, 2015)

The problem

"When explaining a command, or language feature, or hardware widget, first describe the problem it is designed to solve."

Software engineering principle from David Martin, quoted in "Bumper-sticker computer science" by John Bentley, *Programming Pearls*, Communications of the ACM, March 1986

Setting a Simple Stage: Facts about DNA

- DNA/Genome analysis (4 nucleotides, A, T, C, G)
- Reverse complementarity (A-T, C-G) across 2 strands
- Because of reverse complementarity, measure length in "base-pairs" bp
- DNA has a start end ($5'$) and a finish end ($3'$) (only one way to read the "tape")
- "previous", "next" base make sense, gives a *temporal orientation* based on the direction

Scale of the Simple Stage: DNA

- viruses have several thousand bp
- bacteria have a few million bp
- mosquito has 300 million bp
- mouse has 2.4 billion bp
- human DNA has 3.3 billion nucleotides (base-pairs) in each strand
- wheat is much longer than human DNA, 17 billion bp

CpG Islands 1

In the human genome the dinucleotide CpG (written CpG to distinguish it from the CG base-pair across two strands) is rarer than would be expected from the independent probabilities of C and G, for reasons of chemistry.

CpG Islands 2

For example, in the human genome, which has a 42% GC content, a pair of nucleotides consisting of cytosine followed by guanine would be expected to occur

$$0.21 * 0.21 = 4.41\%$$

of the time. The frequency of CpG dinucleotides in human genomes is 1% — less than one-quarter of the expected frequency.

The total number of CpG sites in humans is 28 million.

CpG Islands 3

For biologically important reasons, the chemistry is suppressed in short regions of the genome, such as around the promoters or start regions of many genes. In these regions, we see many more `CpG` dinucleotides than elsewhere.

Such regions are called CpG islands. They are typically a few hundred to a few thousand bases long.

CpG Islands 4: Problems

Given a short stretch of genomic sequence, how would we decide if it comes from a CpG island or not? (This is the *scoring* or *evaluation* problem defined later.)

Second, given a long sequence, how would we find the CpG islands in it, if there are any? (This is the *segmentation* or *estimation* problem defined later.)

Simplifying Assumptions

- "Just" DNA/Genome analysis with 4 nucleotides, A, T, C, G (and not protein analysis with 20 amino acids)
- 2 States along the DNA strand: AT-rich, GC-rich (and not more detailed functional units like exons, introns, promoters, etc.)

Segmentation Problem 1

This is a common problem in bioinformatics.

New DNA sequence: Which state (GC-rich or AT-rich) is the most likely to have generated each nucleotide position in that DNA sequence?

This is the problem of finding the most probable state path, assigning the most likely state to each position in the DNA sequence.

The problem of finding the most probable state path is also called *segmentation*.

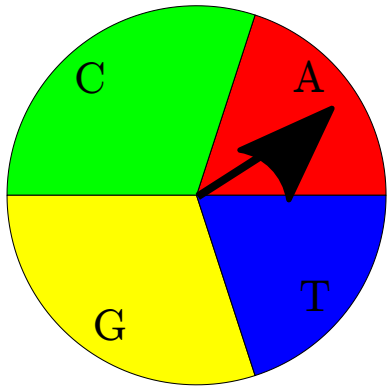
Segmentation Problem 2

For example, give a DNA sequence of 1000 nucleotides, you may wish to use your HMM to *segment* the sequence into blocks that were probably generated by the "GC-rich" state or by the "AT-rich" state.

Simplest Model: Multinomial Model

The *multinomial model* of DNA sequence evolution assumes that

- a random process to choose any of the four nucleotides at each position,
- predetermined, fixed probability distribution.



Failures of Multinomial Model

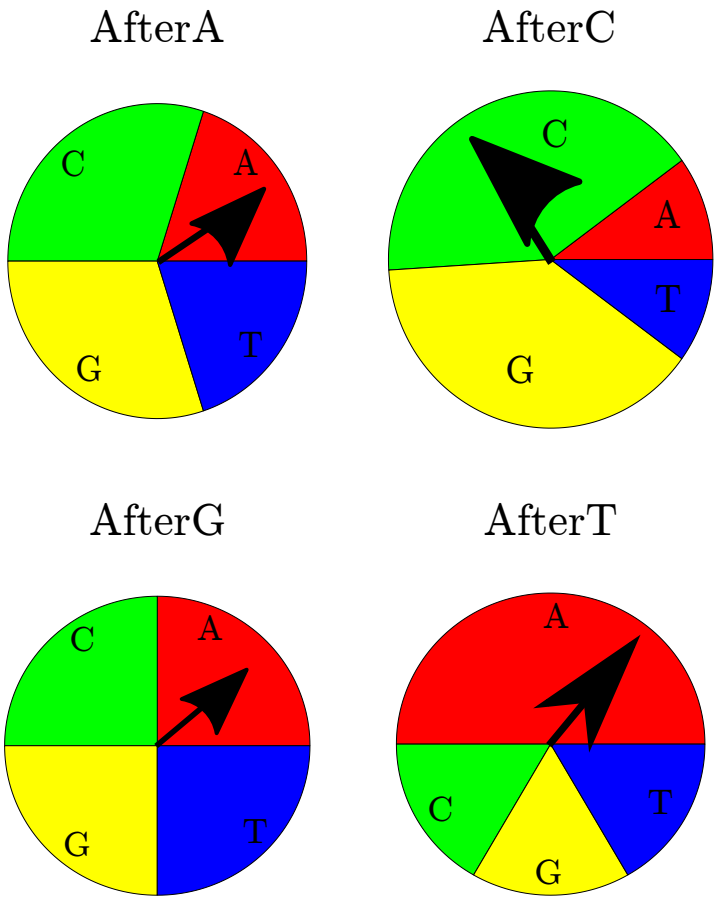
- Fixed probabilities may not hold if there are differences in nucleotide frequencies in different parts of the sequence.
- Probability does not depend on the nucleotides found at adjacent positions in the sequence. For some DNA sequences, this is not true,

Markov Models of DNA 1

A Markov sequence model assumes the sequence has been produced by a random process where the probability of choosing any one of the four nucleotides at a particular position depends on the nucleotide chosen for the previous position.

Markov Models of DNA 2

A Markov model: four different spinners: "afterA", "afterT", "afterG", and "afterC". Each spinner has four slices labeled "A", "T", "G", and "C", but in each spinner a different fraction of the wheel is taken up by the four slices.



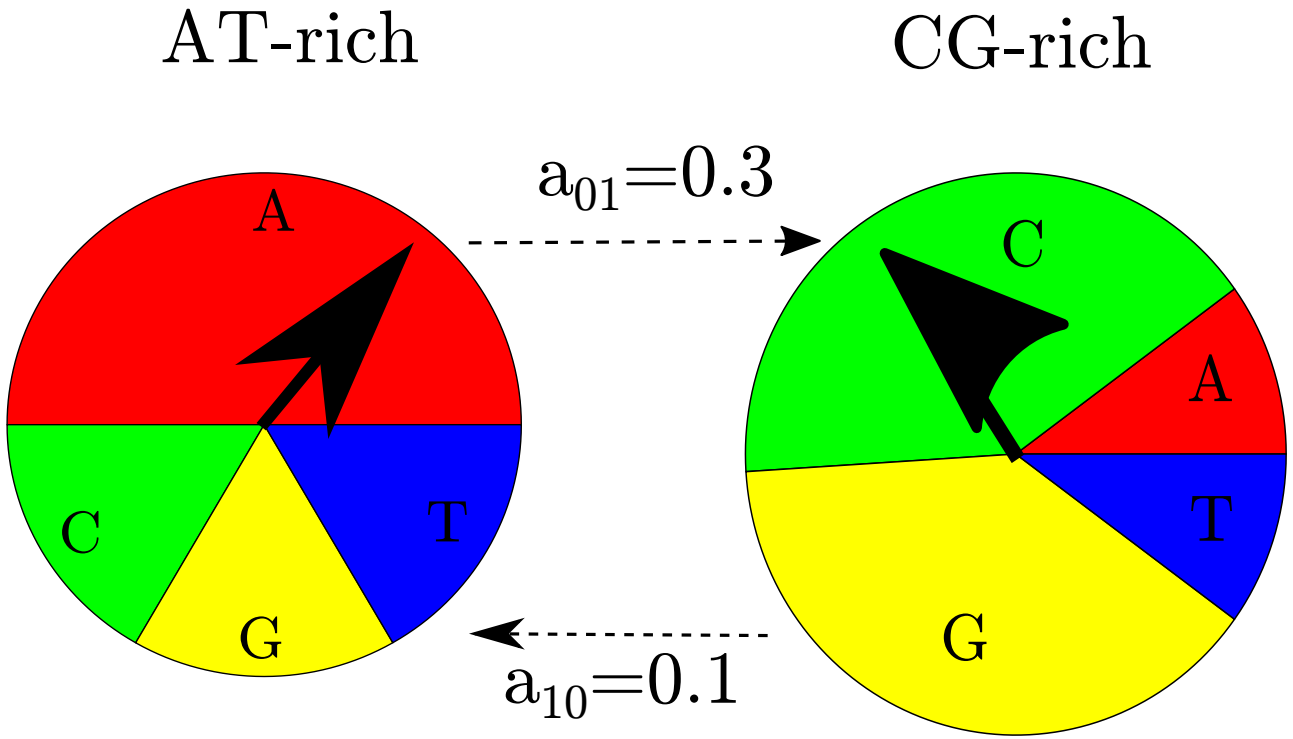
Hidden Markov Models

In a Hidden Markov model (HMM), the nucleotide found at a particular position in a sequence depends on the state at the previous nucleotide position in the sequence.

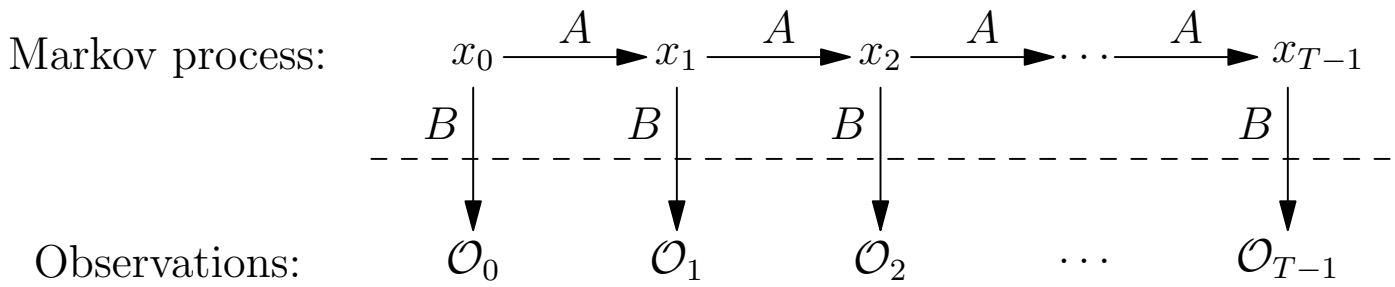
For example, a particular HMM may model the positions along a sequence as belonging to either one of two states, "GC-rich" or "AT-rich".

A more complex HMM may model the positions along a sequence as belonging to many different possible states, such as "promoter", "exon", "intron", and "intergenic DNA".

Spinner Representation



States and Observations Representation



$A = (a_{ij}) = (\mathbb{P}[q_j \mid q_i]) =$ state transition probability matrix

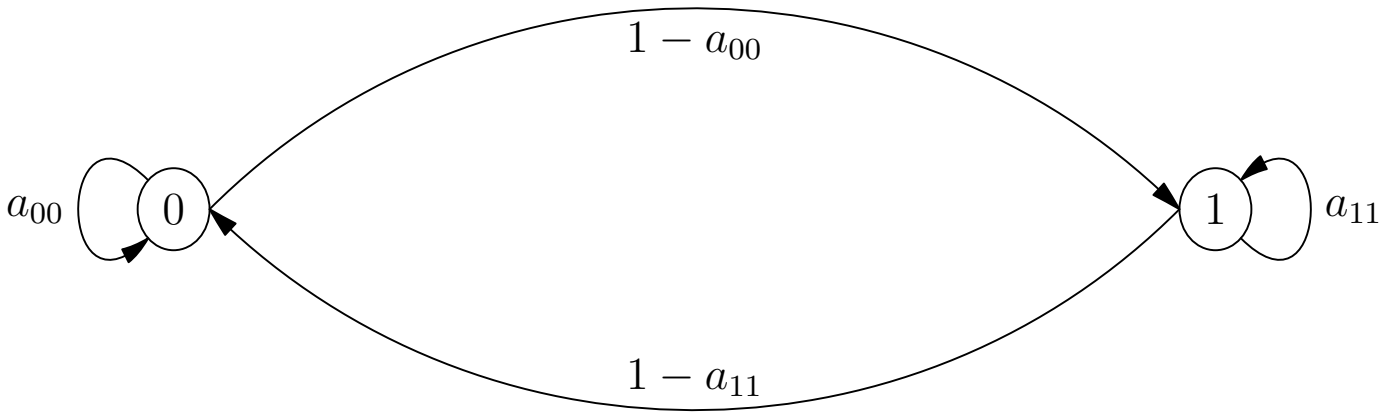
$B = (b_i(j)) = (b_{ij}) = \mathbb{P}[v_j \mid q_i]$ observation probability.

$\pi = \{\pi_j\} =$ initial state distribution at time 0.

$\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \ldots \mathcal{O}_{T-1}) =$ the observation sequence.

The HMM is denoted by $\lambda = (A, B, \pi)$.

HMM as State Graph



$$P[H \mid 0] = p_0 \qquad P[H \mid 1] = p_1$$

$$P[T \mid 0] = 1 - p_0 \qquad P[T \mid 1] = 1 - p_1$$

Evaluation Problem 1

Given the model $\lambda = (A, B, \pi)$ and a sequence of observations $\mathcal{O}$, find $\mathbb{P}[\mathcal{O} \mid \lambda]$. That is, *determine the likelihood of the observed sequence* $\mathcal{O}$, given the model.

Problem 1 is the **evaluation problem**: given a model and observations, how can we compute the probability that the model produced the observed sequence. We can also view the problem as: how we "score" or evaluate the model. If we think of the case in which we have several competing models, the solutions of problem 1 allows us to choose the model that best matches the observations.

Problem 2: Uncover the Hidden States

Given the model $\lambda = (A, B, \pi)$ and a sequence of observations $\mathcal{O}$, find an optimal state sequence. In other words, we want to uncover the hidden part of the Hidden Markov Model.

Problem 2 is the one in which we attempt to uncover the hidden part of the model, i.e. the state sequence. This is the *estimation problem*. Use an optimality criterion to discriminate which sequence best matches the observations. Two optimality criteria are common, and so the choice of criterion is a strong influence on the revealed state sequence.

Problem 3: Training and Parameter Fitting

Given an observation sequence $\mathcal{O}$ and the dimensions N and M, find the model $\lambda = (A, B, \pi)$ that maximizes the probability of $\mathcal{O}$. This can interpreted as training a model to best fit the observed data. We can also view this as search in the parameter space represented by A, B and π.

The solution of Problem 3 attempts to optimize the model parameters so as best to describe how the observed sequence comes about. The observed sequence used to solve Problem 3 is called a *training sequence* since it is used to train the model. This training problem is the crucial one for most applications of hidden Markov models since it creates best models for real phenomena.

Results about HMMs

There are efficient algorithms for

- *scoring*, the forward algorithm
- *segmentation* or *estimation*, the Viterbi algorithm
- *training*, the Baum-Welch algorithm

GENSCAN

- Gene finding on a genome
- Burge and Karlin, 1997, 1998
- Has 14 states
- States depend on whether a gene is in single contiguous region or in two or more separated pieces

Other HMM Models for Gene Finding

- Genemark (Lakashin, Borodvsky, 1993, 1995, 1998)
- BGF
- VEIL
- FGENESH
- GLIMMERHMM

HMMer

- uses extension of HMMs called profile-HMMs
- finds homologous (shared ancestry) protein or nucleotide sequences
- through multiple sequence alignment