Hidden Markov Model User Manual v1.0

Anton E. Weisstein, Truman State University

Jan. 5, 2013

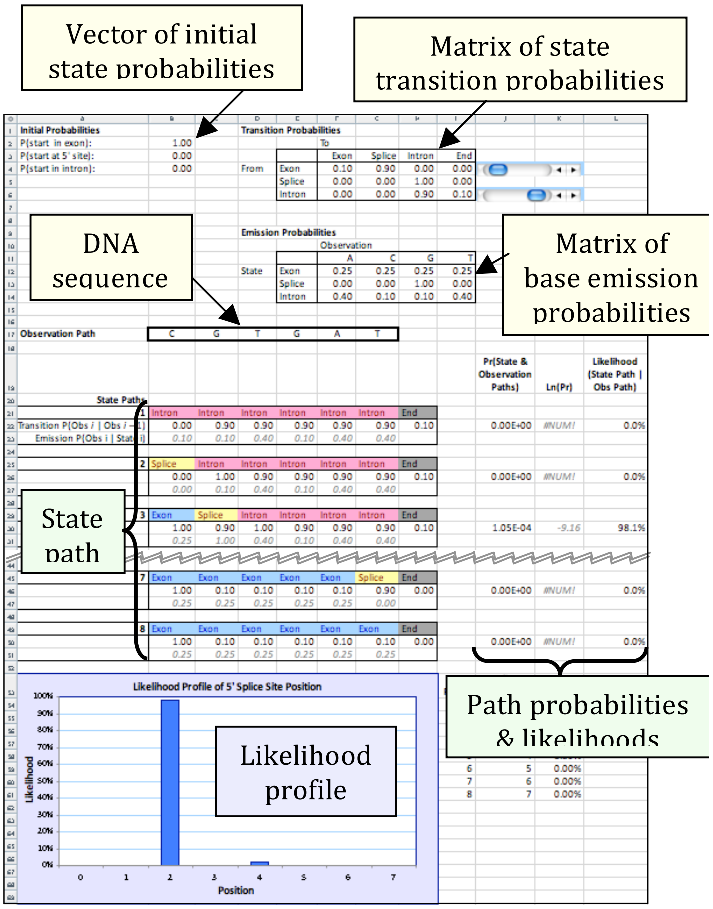
Hidden Markov models (HMMs) are a set of mathematical tools that can be used to draw inferences about genomic and evolutionary processes. In general, an HMM calculates the probability of each scenario that could have resulted in an observed data set, then uses statistical methods to determine which scenarios are most likely to have occurred. For example, HMMs can be applied to DNA sequence data from different species to infer the likely pattern of relationships among those species.

The *Excel* workbook "Hidden Markov Model" illustrates the mathematical workings of an HMM, using Eddy's (2004) example of locating the 5' splice site within a DNA sequence. By manipulating model parameters such as the average length and base composition of introns vs. exons, the user can study how such changes influence the model's decisions about where the splice site is likeliest to occur. For ease of demonstration, the workbook begins by analyzing a very short sequence (6 bp) before re-creating Eddy's full, 26-bp model.

1. Simple Model

The first sheet in the workbook, appropriately named "Simple Model", demonstrates the calculations involved in building an HMM to determine the exact location of an exon/intron boundary. The user enters a short DNA sequence and sets model parameters. The workbook then uses these values to compute the likelihood of each potential 5' splice site location.

Cells B2–B4 contain the probabilities that the DNA sequence begins in different regions of a gene. In HMM parlance, these regions are described as ***states*** of the model. Cells F4–I6 give the probability per nucleotide of moving from one state to another. For example, if G4 = 0.60, then a nucleotide in the gene's exon has a 60% chance of being followed by a nucleotide in the splice site. This model focuses only on 5' splice sites, so states will always occur in the order Exon – Splice – Intron – End.



Vector of initial state probabilities
Matrix of state transition probabilities
DNA sequence
Matrix of base emission probabilities
State path
Path probabilities & likelihoods
Likelihood profile
Likelihood Profile of 5' Splice Site Position
Likelihood
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1
2
3
4
5
6
7
Position

Initial Probabilities	
P(start in exon):	1.00
P(start at 5' site):	0.00
P(start in intron):	0.00

Transition Probabilities

		To			
		Exon	Splice	Intron	End
From	Exon	0.10	0.90	0.00	0.00
	Splice	0.00	0.00	1.00	0.00
	Intron	0.00	0.00	0.90	0.10

Emission Probabilities

		Observation			
		A	C	G	T
State	Exon	0.25	0.25	0.25	0.25
	Splice	0.00	0.00	1.00	0.00
	Intron	0.40	0.10	0.10	0.40

Observation Path	C	G	T	G	A	T				
State Paths								Pr(State & Observation Paths)	Ln(Pr)	Likelihood (State Path | Obs Path)
1	Intron	Intron	Intron	Intron	Intron	Intron	End			
Transition P(Obs *i* | Obs *i* – 1)	0.00	0.90	0.90	0.90	0.90	0.90	0.10	0.00E+00	#NUM!	0.0%
Emission P(Obs i | State i)	0.10	0.10	0.40	0.10	0.40	0.40				
2	Splice	Intron	Intron	Intron	Intron	Intron	End			
	0.00	1.00	0.90	0.90	0.90	0.90	0.10	0.00E+00	#NUM!	0.0%
	0.00	0.10	0.40	0.10	0.40	0.40				
3	Exon	Splice	Intron	Intron	Intron	Intron	End			
	1.00	0.90	1.00	0.90	0.90	0.90	0.10	1.05E-04	-9.16	98.1%
	0.25	1.00	0.40	0.10	0.40	0.40				
7	Exon	Exon	Exon	Exon	Exon	Splice	End			
	1.00	0.10	0.10	0.10	0.10	0.90	0.00	0.00E+00	#NUM!	0.0%
	0.25	0.25	0.25	0.25	0.25	0.00				
8	Exon	Exon	Exon	Exon	Exon	Exon	End			
	1.00	0.10	0.10	0.10	0.10	0.10	0.00	0.00E+00	#NUM!	0.0%
	0.25	0.25	0.25	0.25	0.25	0.25				

6	5	0.00%
7	6	0.00%
8	7	0.00%

Cells F12–I14 give the state-specific probability of producing, or ***emitting***, a particular nucleotide. In the initial model, all four nucleotides are equally likely to occur in an exon, but introns are GC-poor. For the sake of simplicity, the splice site is considered to be only one nucleotide long: in the initial model, that nucleotide is always a guanine. Finally, Cells B17–G17 contain the nucleotide sequence to be analyzed.

Cells B21–L51 show the model's calculations. For each potential ***state path***, the model computes (i) the probability that the DNA sequence starts in the appropriate state, (ii) the probability of precisely following the state path, and (iii) the probability of emitting the observed DNA sequence if the state path is followed. For example, state path 1 (Cells B21–H21) posits that the first nucleotide, a cytosine, occurs in an intron. The initial model gives a 0% probability of starting in an intron (Cell B4), and a 10% frequency of cytosines within an intron (Cell G14), so these values are reported. The second nucleotide, a guanine, is also proposed to occur in an intron. The initial model therefore reports the 90% probability of remaining within an intron (Cell H6) and the 10% frequency of guanines within an intron (Cell H14). This process continues through all six nucleotides, then reports the 10% probability of transitioning from the last nucleotide's state — an intron — to the end of the sequence (Cell I6).

HMMs assume that each nucleotide is independent of the others, much as the result of a coin flip does not depend on previous or future flips. This assumption of independence is biologically unrealistic, but it greatly simplifies the math: each nucleotide's individual probabilities can be simply multiplied together to calculate the total probability of the overall sequence. For state path 1, this combined probability is reported in Cell J22. Unsurprisingly, given that one of the numbers being multiplied was zero, the resulting product is also zero, meaning that this state path is not consistent with the observed sequence. In fact, under the initial model, state paths 3 and 5 are the only ones possible, because only they correctly place the splice site at a G.

Computationally, it is faster to add numbers together than to multiply them. This can make a huge difference when running a large HMM that may contain many millions of state paths. For this reason, most HMMs calculate the combined probability of a state path not by multiplying the individual nucleotide probabilities, but by adding the logarithms of those probabilities, using the identity $\ln(ab) = \ln(a) + \ln(b)$. The resulting ***log-probabilities*** are reported in Column K. Note that $\ln(0)$ is undefined, so state paths with probability zero yield an error message in this column.

If the model correctly includes all possible state paths, adding up all of the probabilities in Column J should give the total probability p of emitting the observed sequence. But by definition, that sequence has indeed been observed, so the individual state path probabilities should be calibrated to sum to one. (In much the same way, if I have prior knowledge that a poker player is holding either a flush or a full house, each of those two hands is much likelier than if I lacked such knowledge.) Each state path's probability is therefore divided by the total probability *p* to compute the path's likelihood given observed DNA sequence. The results are displayed in Column L and graphed in the likelihood profile at the bottom of the worksheet.

A few brief explorations may help clarify how the model works while also providing insight into the logic of HMMs:

1. Change the observed DNA sequence. Alterations in GC richness and in the position of any G nucleotides may influence the likelihood of the 5' splice site occurring at a specific position. Note that the sequence must consist only of standard DNA nucleotides (A, C, G, and T).

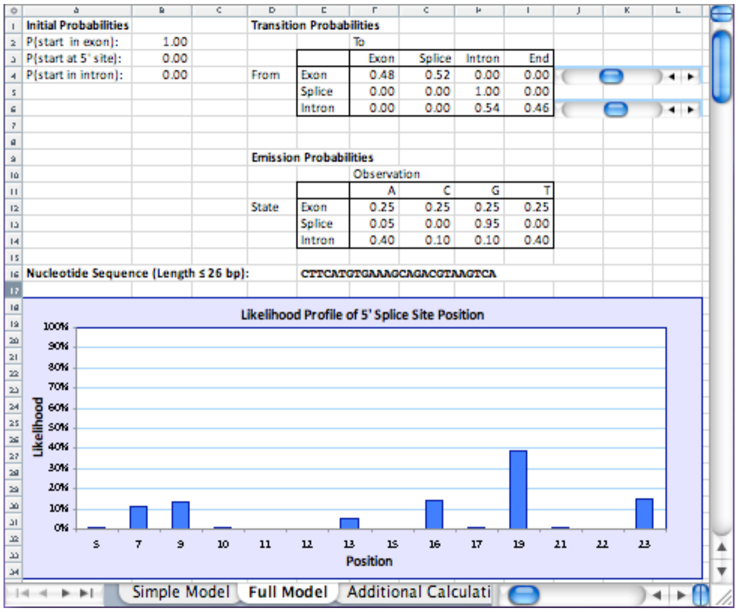
2. Change the base emission probabilities. Alterations in the base composition of exons vs. introns may affect the model's inferences about splice site position. In particular, similar base compositions in these two regions can severely compromise the model's ability to infer the splice site location with any degree of confidence. Note that each position within a given gene region must be occupied by one of the four nucleotides, so each <u>row</u> of the matrix must sum to one. If this condition is violated, the offending cells will be highlighted in red.

3. Change the state transition probabilities using the built-in scrollbars. Again, each <u>row</u> of the matrix must sum to one. Note that the exon-to-splice transition probability is the reciprocal of the average exon's length (e.g., if each nucleotide in an exon nucleotide has a 10% chance of being followed by a splice site, the average exon will be 10 bp long). Similarly, the intron-to-end transition probability is the reciprocal of the average intron's length. Adjusting these parameters will therefore shift the splice site's likely position to earlier or later in the sequence. The model assumes that the splice site is exactly 1 bp long, so the probabilities in the middle row cannot be changed.

4. Change the vector of initial state probabilities. The model is initially set up to assume that the DNA sequence starts in an exon; however, an actual read might start within an intron, or even at the splice site itself. A good model should therefore assign these possibilities an appropriate non-zero probability. By definition, the sequence must start in one of these regions, so these probabilities must sum to one.

2. Full Model

The second sheet of the “Hidden Markov Model” workbook contains an exact replica of the HMM described by Eddy (2004). This sheet can be reached using the navigation tabs at the bottom of the *Excel* workspace. Eddy’s HMM is designed to handle DNA sequences up to 26 bp in length; moreover, under the initial settings, the splice site can occur at either an A or a G, although the latter is far more likely.



Initial Probabilities

P(start in exon):	1.00
P(start at 5' site):	0.00
P(start in intron):	0.00

Transition Probabilities

		To			
		Exon	Splice	Intron	End
From	Exon	0.48	0.52	0.00	0.00
	Splice	0.00	0.00	1.00	0.00
	Intron	0.00	0.00	0.54	0.46

Emission Probabilities

		Observation			
		A	C	G	T
State	Exon	0.25	0.25	0.25	0.25
	Splice	0.05	0.00	0.95	0.00
	Intron	0.40	0.10	0.10	0.40

Nucleotide Sequence (Length ≤ 26 bp): CTTCATGTGAAAGCAGACGTAAGTCA

This sheet’s controls work almost exactly the same way as those for the preceding model. The only major difference is that the user may paste a DNA sequence directly into Cell E16: the worksheet itself will extract the individual nucleotides for analysis. The interface is also streamlined to focus attention on the model’s controls and graph rather than on the detailed calculations involved. However, the user can still view the relevant calculations by scrolling down to Rows 44–183 and/or by navigating to the “Additional Calculations” sheet.

Suggested explorations using the full model:

1. Study the effect of increasing or decreasing the base composition difference of exons vs. introns. How would you numerically measure the model’s statistical power to resolve the splice site’s location for a particular level of composition difference? Graph the model’s power as a function of composition difference, and determine the minimum difference needed to infer the splice site location of a 26-bp sequence with 95% confidence.

2. Using online resources, find the distribution of exon and intron lengths for a specific class of gene or a specific organism. How would you reduce this distribution to a single average length for use in the model? Enter the appropriate values in the transition probability matrix and calculate the likelihood profile for the splice site’s location. Compare your results to those of a neighbor who used the same DNA sequence but focused on a different gene or organism: how much do the results depend on the model’s transition matrix?

3. Terms and Conditions

History

Date: Jan. 5, 2013
Title: *Hidden Markov Model* Manual v1.0
Name: Anton E. Weisstein
Institutions: Washington University and Truman State University
Acknowledgements: This work was supported by a Freiburg Visiting Scholarship provided by Washington University in St. Louis.
Modifications: None (original version).