CpG Educate User's Guide

Aaron Garrett Jacksonville State University agarrett@jsu.edu

> Last Updated: December 5, 2012

Contents

1	Introduction	1
2	Sliding Windows	2
3	Dishonest Casino	3
	3.1 Generation	3
	3.2 Evaluation	4
	3.3 Decoding	5
	3.4 Training	6
4	CpG Islands	7
	4.1 Generation	8
	4.2 Evaluation	9
	4.3 Decoding	10
	4.4 Training	11

1 Introduction

The CpG Educate online software allows users to explore the mathematics of finding CpG islands in genomic data. The software was designed as a companion to the book "Mathematical Concepts and Methods in Modern Biology," edited by Raina Robeva and Terrell Hodge. This document serves as a brief user's guide to the various features of the software. Specific questions not covered in this guide may be sent to the author via email (listed above). The CpG Educate website can be found at http://inspired.jsu.edu/~agarrett/cpg/index.htm. The software has been most fully tested using Google's Chrome browser, and users are recommended to access the suite through Chrome. The online tools are broken into three groups, each focusing on a different problem or approach. These three groups are described in more detail in the following sections.

2 Sliding Windows

The Sliding Windows application allows the user to watch the progress of the sliding windows algorithm as it process a user-defined nucleotide sequence. The initial Sliding Windows interface can be seen in Figure 1. This figure has numbered the three inputs to the Sliding Windows interface, and each is detailed below.

- 1. Nucleotide Sequence: This is the sequence in which CpG islands should be found. It may contain an initial Fasta description line (beginning with a ">") if desired.
- 2. Start/Stop Button: This button begins and ends the search process.
- 3. **Speed Control**: This slider can be used to control how slowly or quickly the animation moves. However, due to browser processing overhead, even at the fastest setting it will not be particularly fast.



Figure 1: Sliding Windows Interface Inputs

The two outputs of the Sliding Windows system are numbered in Figure 2 and are detailed below.

- 1. Island Coordinates: This box lists the starting and ending locations on the nucleotide sequence of the CpG island.
- 2. Sequence Visualization: This is a visual representation of the nucleotide sequence (A, C, G, and T are represented as red, blue, yellow, and green, respectively). It also shows a white-transparent overlay on the current window being processed. This animation will move backward and forward as the algorithm proceeds. Any islands that are found are highlighted with a yellow-transparent overlay.



Figure 2: Sliding Windows Interface Outputs

3 Dishonest Casino

The Dishonest Casino application allows the user to specify the parameters to a hidden Markov model to solve the Dishonest Casino problem, in which a fair and unfair die are used with specified probabilities. The initial interface for the Dishonest Casino is shown in Figure 3. The model parameters are labeled "M" in the figure, and, while they have the typical default values for the Dishonest Casino problem, they may instead be specified by the user. The interface also has four tabs—Generation, Evaluation, Decoding, and Training. The model parameters apply to all four tabs.

The Dishonest Casino

The matrix below represents the model parameters for the Dishonest Casino problem. In this problem, a casino uses both a fair and an unfair six-sided die. The default parameters listed below assume that the casino moves from a fair die to an unfair die 5% of the time and from an unfair to a fair die 10% of the time. The fair die has an equal probability of landing on any particular side, but the unfair die 50% likely to land on a 6 and 10% likely to land on any of the other sides. The initial state of the game has a 50% chance of using the fair die.											
Hover over any of the question marks ? for more information.											
Model Parameters ?											
Transition (M) Emission											
Fair Unfair 1 2 3 4 5 6 Initial											
Fair 0.95 0.05 0.1667 0.1667 0.1667 0.1667 0.1667 0.1667 0.1667 0.1667											
Unfair 0.1 0.9 0.1 0.1 0.1 0.1 0.1 0.5 0.5											
Generation Evaluation Decoding Training											
Choose the length of the sequence to generate using the model parameters above. The sequence of output symbols, along with the sequence of hidden states that produced them, will be given below. For the hidden states, "0" corresponds to the fair die and "1" corresponds to the unfair die. It will be useful to copy-paste these values into the inputs of the other tabs. Sequence Length: 70 Generate Sequence 11313534422556443313454534666161614166666366666516426265366545556 66332											
States 0000000001111110000000000111111111111											

Figure 3: Dishonest Casino Generation Tab

3.1 Generation

The Generation tab can also be seen in Figure 3. Its user inputs and outputs are enumerated below.

- 1. Sequence Length: This input determines the number of die rolls to generate.
- 2. Generate Button: Clicking this button initiates the sequence generation.
- 3. Sequence Area: This output area holds the generated sequence. It is uneditable, but its contents may be copied.
- 4. **States Area**: This output area holds the hidden states that generated the sequence above. It is uneditable, but its contents may be copied.

3.2 Evaluation

The Evaluation tab can be seen in Figure 4. Its user inputs and outputs are enumerated below.

- 1. **Sequence Area**: This input should be the sequence of die rolls to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. States Area: This optional input should be the hidden states that generated the sequence above. Once again, this could be copied from the Generation tab.
- 3. Evaluate Button: Clicking this button initiates the evaluation.
- 4. **Graph**: This graph plots the probability that a particular element of the sequence was generated by an unfair hidden state. If the States Area input is used, the "unfair" bars show where the actual unfair die was used. Otherwise, they will not be displayed.

Generation	Evaluation	Decoding	Training											
Enter the sequence to evaluate (usually copied from the Generate tab) and, optionally, the hidden states that produced the sequence (from the Generate tab). For the hidden states, "0" corresponds to the fair die and "1" corresponds to the unfair die. The posterior probabilities for being in a state generated by the unfair die will then be calculated and presented as a graph below. If the hidden states are given, those with a state of "1" will appear as areas of dark gray background on the graph.														
Sequence	Sequence ?													
11313534422556443313454534666161614166666366666516426265366545556 66332														
00352	66332 1													
States (opti	onal) ?													
0000000000	111111000000	000011111111	111111111											
	2													
				Evaluate										
1.2			\sim	unfair I probability										
0.8		\sim	(han .										
0.6		(4)		\sim										
0.4				Ŭ V										
0.2	\sim	/												
1	10 20	30	40	50 60 70										

Figure 4: Dishonest Casino Evaluation Tab

3.3 Decoding

The Decoding tab can be seen in Figure 5. Its user inputs and outputs are enumerated below.

- 1. **Sequence Area**: This input should be the sequence of die rolls to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. States Area: This optional input should be the hidden states that generated the sequence above. Once again, this could be copied from the Generation tab.
- 3. Decode Button: Clicking this button initiates the decoding.
- 4. **Decoding Output**: This output shows the color-coded input sequence. Here, symbols in red were likely generated by an unfair die, and symbols with gray backgrounds were actually generated by an unfair die.
- 5. $\log_2(\text{probability})$: This output is the logarithm of the probability that the hidden state output above actually generated the input sequence.

Constation Fundamentary Training
Generation Evaluation Decoung Training
Enter the sequence to decode (usually copied from the Generate tab) and, optionally, the hidden states that produced the sequence (from the Generate tab). For the hidden states, "0" corresponds to the fair die and "1" corresponds to the unfair die. The Viterbi algorithm will then be used to decode the sequence. In the output, the red symbols are the predicted "1" states and the gray background symbols are actual "1" states (if supplied).
Sequence ?
11313534422556443313454534666161614166666366666516426265366545556 66332
States (optional) ?
11000
Decoding Output 11313334422336644333134545334 Inog_(probability):-120.378933 5

Figure 5: Dishonest Casino Decoding Tab

3.4 Training

The Training tab can be seen in Figure 6. Its user inputs and outputs are enumerated below.

- 1. **Sequence Area**: This input should be the sequence of die rolls to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. Train Button: Clicking this button initiates the training.
- 3. **Training Output**: This table shows the model parameters after training. These should be compared to the model parameters marked "M" in Figure 3.

Generation Evaluation Decoding Training										
Enter the sequence to on which to train (usually copied from the Generate tab). The Baum-Welch algorithm will then be used to train the original model (given above) on this sequence. The resulting model parameters will be shown below.										
Sequence ?										
11313534422556443313454534666161614166666366666516426265366545556 66332										
Training Output										
Transition Emission										
Fair Loaded 1 2 3 4 5 6 Initial										
Loaded 0.0002 0.9998 0.1122 0.00 0.0930 0.0757 0.1369 0.5150 0.0000										

Figure 6: Dishonest Casino Training Tab

4 CpG Islands

The CpG Islands application allows the user to specify the parameters to a hidden Markov model to solve the task of finding CpG islands in nucleotide sequences. The model parameters are shown in Figure 7. The user may specify these parameters in two ways. First, the parameters may be loaded from a CSV file, formatted according to the example shown on the CpG Educate site. (Any spreadsheet application, such as Excel, can save to CSV format as one of the file types in the "Save" dialog box.) To do this, the user needs to click on the "Choose File" button labeled "1" in Figure 7. Second, the parameters may be entered manually in the area labeled "2" in Figure 7.

CpG Islands



Figure 7: CpG Islands Model Parameters

4.1 Generation

The Generation tab can be seen in Figure 8. Its user inputs and outputs are enumerated below.

- 1. Sequence Length: This input determines the number of nucleotides to generate.
- 2. Generate Button: Clicking this button initiates the sequence generation.
- 3. Sequence Area: This output area holds the generated sequence. It is uneditable, but its contents may be copied.
- 4. **States Area**: This output area holds the hidden states that generated the sequence above. It is uneditable, but its contents may be copied.

Generation	Evaluation	Decoding	Training								
Choose the length of the sequence to generate using the model parameters above. The sequence of output symbols, along with the sequence of hidden states that produced them, will be given below. The hidden states correspond to the columns in the transition matrix, where 0 represents the first column (A), 1 represents the second column (C), and so on. It will be useful to copy-paste these values into the inputs of the other tabs. Sequence Length: 70 Generate											
Sequence											
gctaacacggo	accctgcatacgac	aggacactgagt	gttgagtgcaacc	gagaaccctggctttgtgct							
		3									
States											
3126030133101112310221001003011122232323032110005667465112311222 32112											
		4									

Figure 8: CpG Islands Generation Tab

4.2 Evaluation

The Evaluation tab can be seen in Figure 9. Its user inputs and outputs are enumerated below.

- 1. **Sequence Area**: This input should be the sequence of nucleotides to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. States Area: This optional input should be the hidden states that generated the sequence above. Once again, this could be copied from the Generation tab.
- 3. Evaluate Button: Clicking this button initiates the evaluation.
- 4. **Graph**: This graph plots the probability that a particular element of the sequence was in a CpG island. If the States Area input is used, the "island" bars show where the actual islands are. Otherwise, they will not be displayed.

Generation	Evaluation	Decoding	Training										
Enter the sequence to evaluate (usually copied from the Generate tab) and, optionally, the hidden states that produced the sequence (from the Generate tab). States 0, 1, 2, 3 correspond to the island regions; states 4, 5, 6, 7 correspond to non-island regions. The posterior probabilities for being in an island state will then be calculated and presented as a graph below. If the hidden states are given, those representing islands (0, 1, 2, and 3) will appear as areas of dark gray background on the graph.													
Sequence ?													
gctaacacggca	gctaacacggcaccctgcatacgacaggacactgagtgttgagtgcaaccgagaaccctggctttgtgct												
	1												
States (optio	onal) ?												
31260301331	101112310221	001003011122	232322303	2110005667465112311222									
52112		~											
	2												
				Evaluate									
1.1				island									
1.0				probability -									
0.9													
0.7		4											
0.6	<u> </u>			~~~~									
0.5	~~~		~										
0.4		m	\sim	\sim									
0.3													
1	0 20	30	40	50 60 70									

Figure 9: CpG Islands Evaluation Tab

4.3 Decoding

The Decoding tab can be seen in Figure 10. Its user inputs and outputs are enumerated below.

- 1. **Sequence Area**: This input should be the sequence of nucleotides to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. States Area: This optional input should be the hidden states that generated the sequence above. Once again, this could be copied from the Generation tab.
- 3. Decode Button: Clicking this button initiates the decoding.
- 4. **Decoding Output**: This output shows the color-coded input sequence. Here, symbols in red were likely a part of a CpG island, and symbols with gray backgrounds were actually in a CpG island.
- 5. $\log_2(\text{probability})$: This output is the logarithm of the probability that the hidden state output above actually generated the input sequence.

Generation Evaluation Decoding Training												
Enter the sequence to decode (usually copied from the Generate tab) and, optionally, the hidden states that produced the sequence (from the Generate tab). The Viterbi algorithm will then be used to decode the sequence. In the output, the red symbols are the predicted islands and the gray background symbols are actual islands (if supplied), corresponding to states 0, 1, 2, 3 from the simulation.												
Sequence ?												
gctaacacggcaccctgcatacgacaggacactgagtgttgagtgcaaccgagaaccctggctttgtgct												
1												
States (optional) ?												
31260301331011123102210010030111222323223032110005667465112311222 32112												
2												
Decode												
Decoding Output (4)												
gctaacacggcaccctgcatacgacaggacactgagtgttgagtgcaaccgagaaccctggctttgtgct												
log ₂ (probability):-112.877771 5												

Figure 10: CpG Islands Decoding Tab

4.4 Training

The Training tab can be seen in Figure 11. Its user inputs and outputs are enumerated below.

- 1. Sequence Area: This input should be the sequence of nucleotides to evaluate. For instance, this could be copy-pasted from the Generation tab.
- 2. Train Button: Clicking this button initiates the training.
- 3. **Training Output**: This table shows the model parameters after training. These should be compared to the model parameters shown in Figure 7.

Gei	Generation Evaluation Decoding Training															
Enter the sequence to on which to train (usually copied from the Generate tab). The Baum-Welch algorithm will then be used to train the original model (given above) on this sequence. The resulting model parameters will be shown below.																
Se	Sequence ?															
gc	gctaacacggcaccctgcatacgacaggacactgagtgttgagtgcaaccgagaaccctggctttgtgct															
	1															
										0		1				
									Г	rain						
Tra	aining	Output	3													
				Trans	sition					Emis	sion					
	Α	С	Т	G	а	С	t	g	а	С	t	g	Initial			
Α	0.222	0.666	0.111	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000			
С	0.235	0.294	0.353	0.000	0.000	0.000	0.000	0.117	0.000	1.000	0.000	0.000	0.000			
T	0.237	0.000	0.000	0.590	0.000	0.000	0.173	0.000	0.000	0.000	0.894	0.106	0.000			
G	0.000	0.857	0.000	0.143	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000			
a	a 0.125 0.000 0.000 0.000 0.000 0.375 0.000 0.500 1.000 0.000 0.000 0.000 0.000 0.000															
C	0.000	0.000	0.000	0.000	0.667	0.000	0.333	0.000	0.000	1.000	0.000	0.000	0.000			
t	0.000	0.000	0.000	0.000	0.000	0.000	0.277	0.723	0.000	0.000	1.000	0.000	0.000			
g	0.000	0.000	0.224	0.000	0.345	0.000	0.140	0.091	0.000	0.000	0.000	1.000	0.000			

Figure 11: CpG Islands Training Tab