CpG Educate User's Guide

Aaron Garrett
Jacksonville State University
agarrett@jsu.edu

Last Updated:
December 5, 2012

Contents

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.

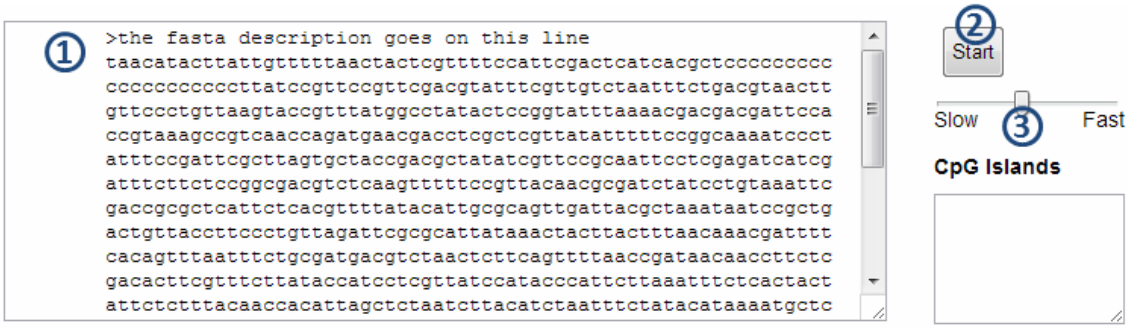


Figure 1: Sliding Windows Interface Inputs

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.

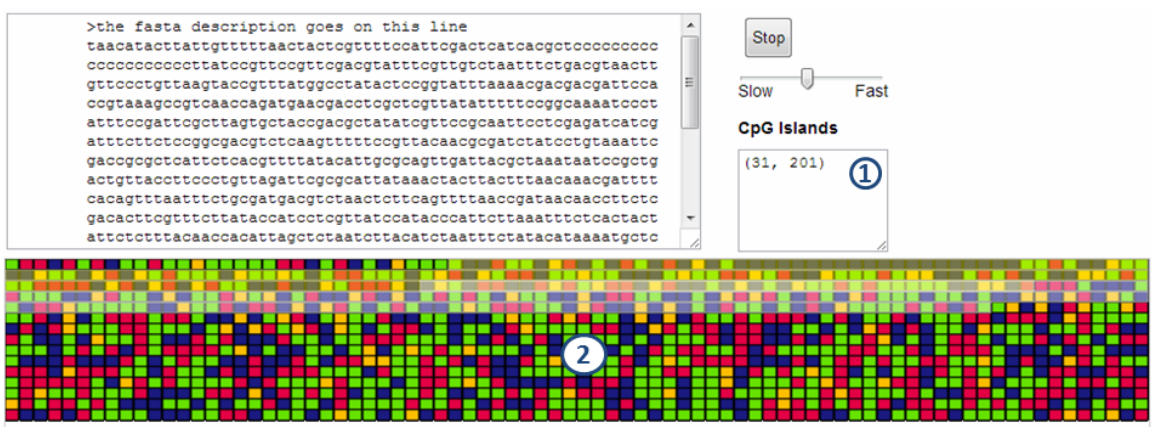


Figure 2: Sliding Windows Interface Outputs

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.

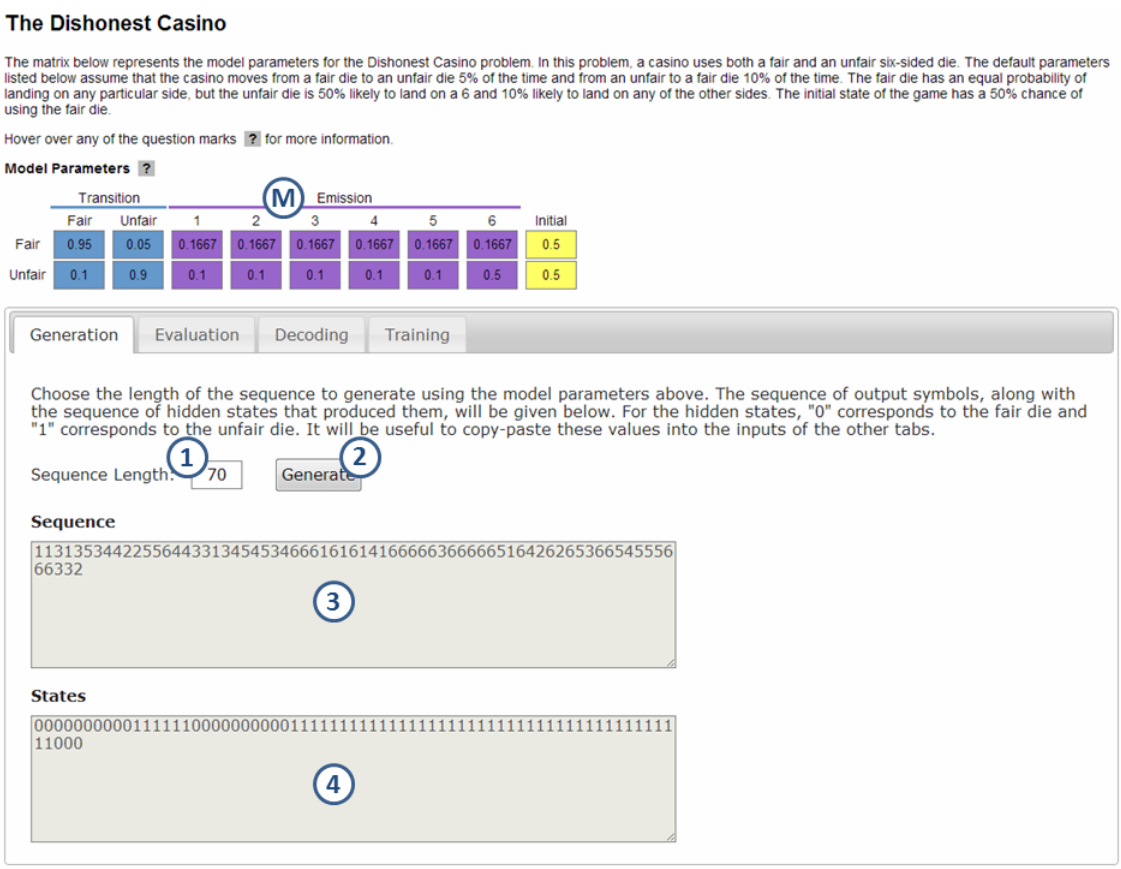


Figure 3: Dishonest Casino Generation Tab

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.

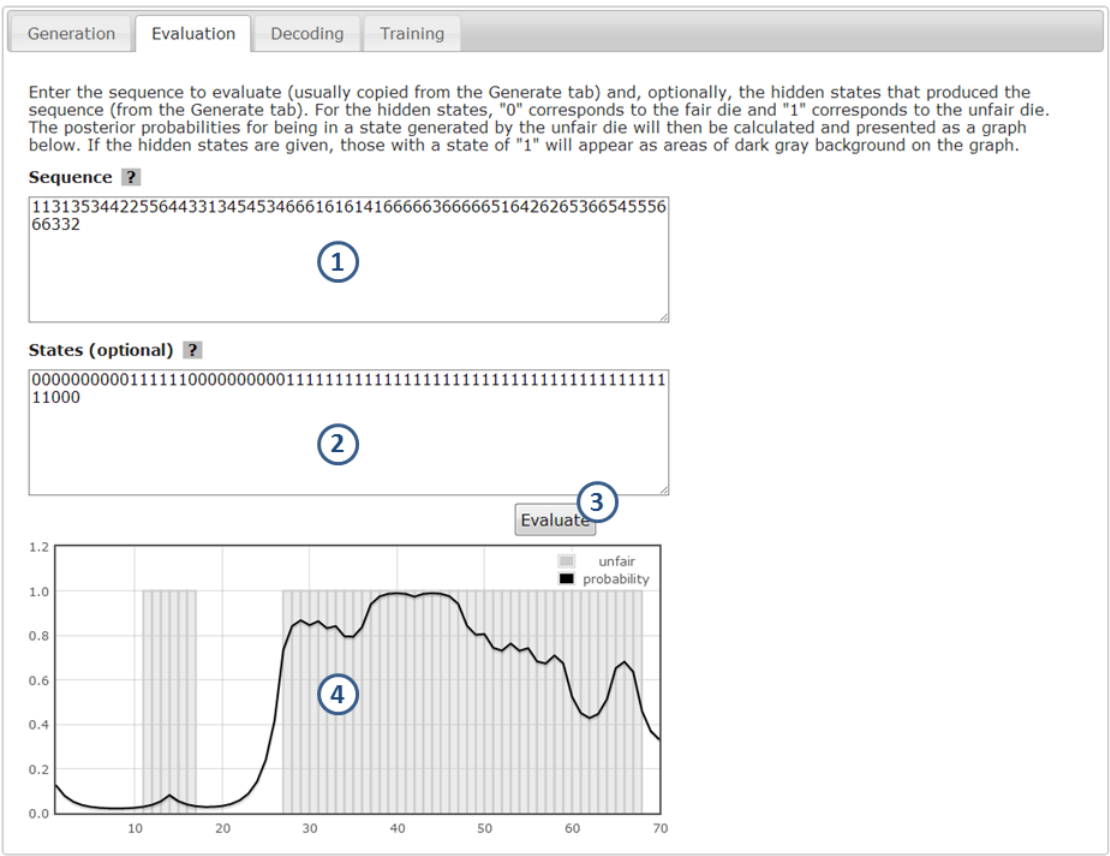


Figure 4: Dishonest Casino Evaluation Tab

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. **$\log_2$(probability)**: This output is the logarithm of the probability that the hidden state output above actually generated the input sequence.

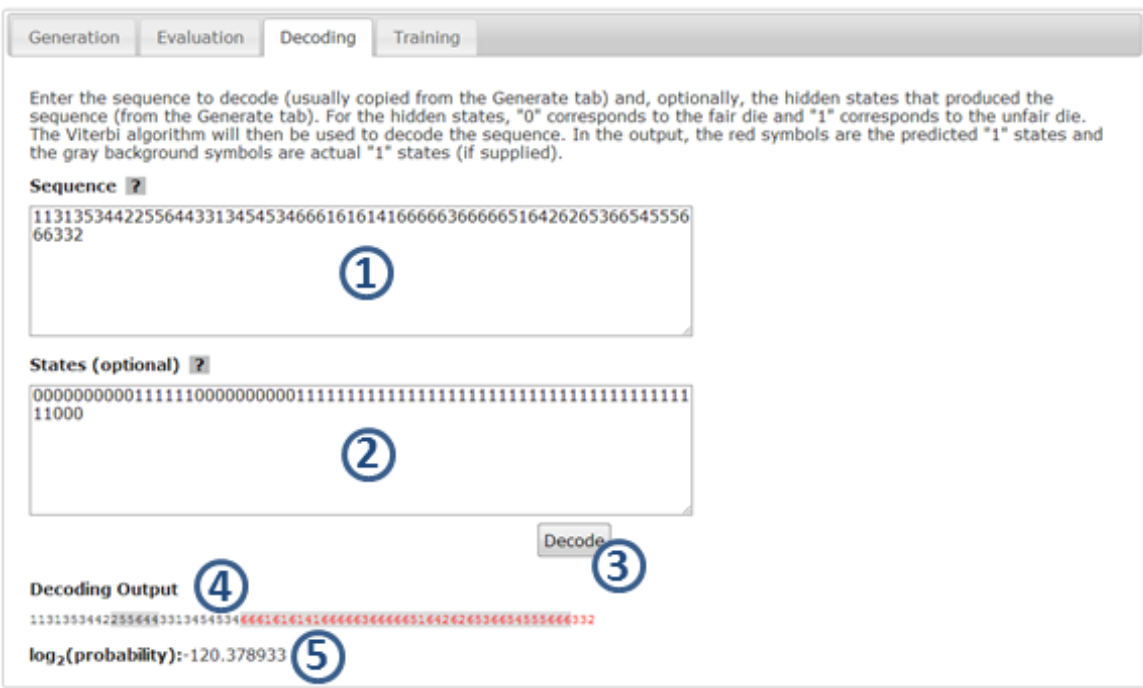


Figure 5: Dishonest Casino Decoding Tab

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.

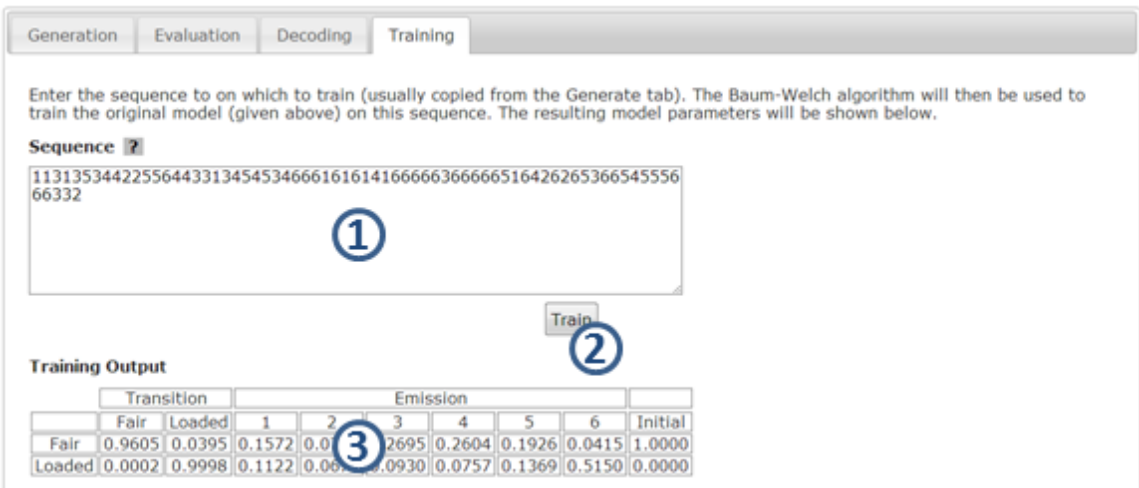
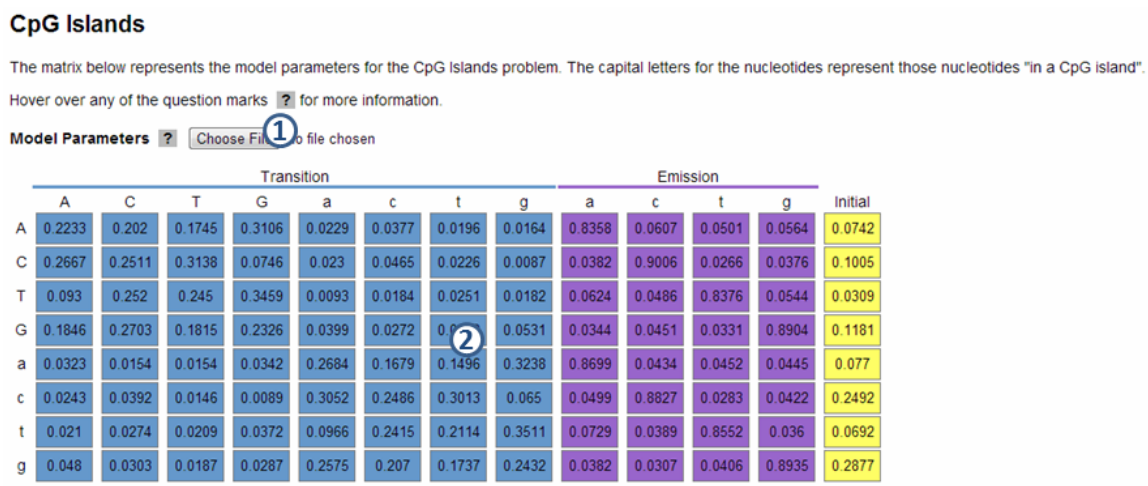


Figure 6: Dishonest Casino Training Tab

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

The matrix below represents the model parameters for the CpG Islands problem. The capital letters for the nucleotides represent those nucleotides "in a CpG island".

Hover over any of the question marks ? for more information.

Model Parameters ? Choose File No file chosen

	Transition								Emission				
	A	C	T	G	a	c	t	g	a	c	t	g	Initial
A	0.2233	0.202	0.1745	0.3106	0.0229	0.0377	0.0196	0.0164	0.8358	0.0607	0.0501	0.0564	0.0742
C	0.2667	0.2511	0.3138	0.0746	0.023	0.0465	0.0226	0.0087	0.0382	0.9006	0.0266	0.0376	0.1005
T	0.093	0.252	0.245	0.3459	0.0093	0.0184	0.0251	0.0182	0.0624	0.0486	0.8376	0.0544	0.0309
G	0.1846	0.2703	0.1815	0.2326	0.0399	0.0272	[illegible]	0.0531	0.0344	0.0451	0.0331	0.8904	0.1181
a	0.0323	0.0154	0.0154	0.0342	0.2684	0.1679	0.1496	0.3238	0.8699	0.0434	0.0452	0.0445	0.077
c	0.0243	0.0392	0.0146	0.0089	0.3052	0.2486	0.3013	0.065	0.0499	0.8827	0.0283	0.0422	0.2492
t	0.021	0.0274	0.0209	0.0372	0.0966	0.2415	0.2114	0.3511	0.0729	0.0389	0.8552	0.036	0.0692
g	0.048	0.0303	0.0187	0.0287	0.2575	0.207	0.1737	0.2432	0.0382	0.0307	0.0406	0.8935	0.2877

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.

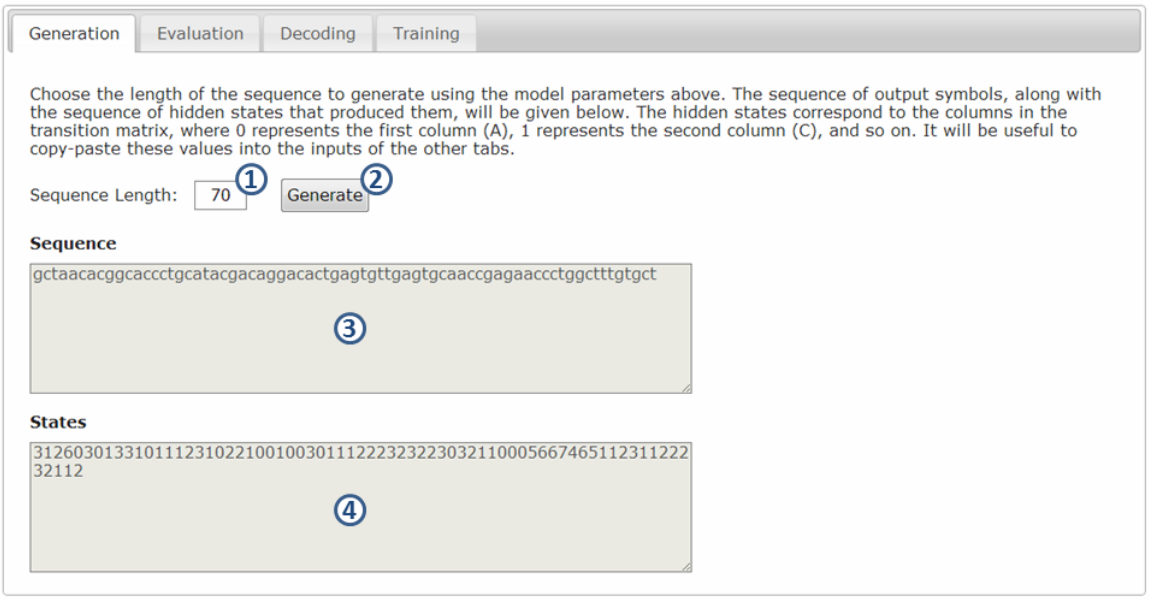


Figure 8: CpG Islands Generation Tab

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.

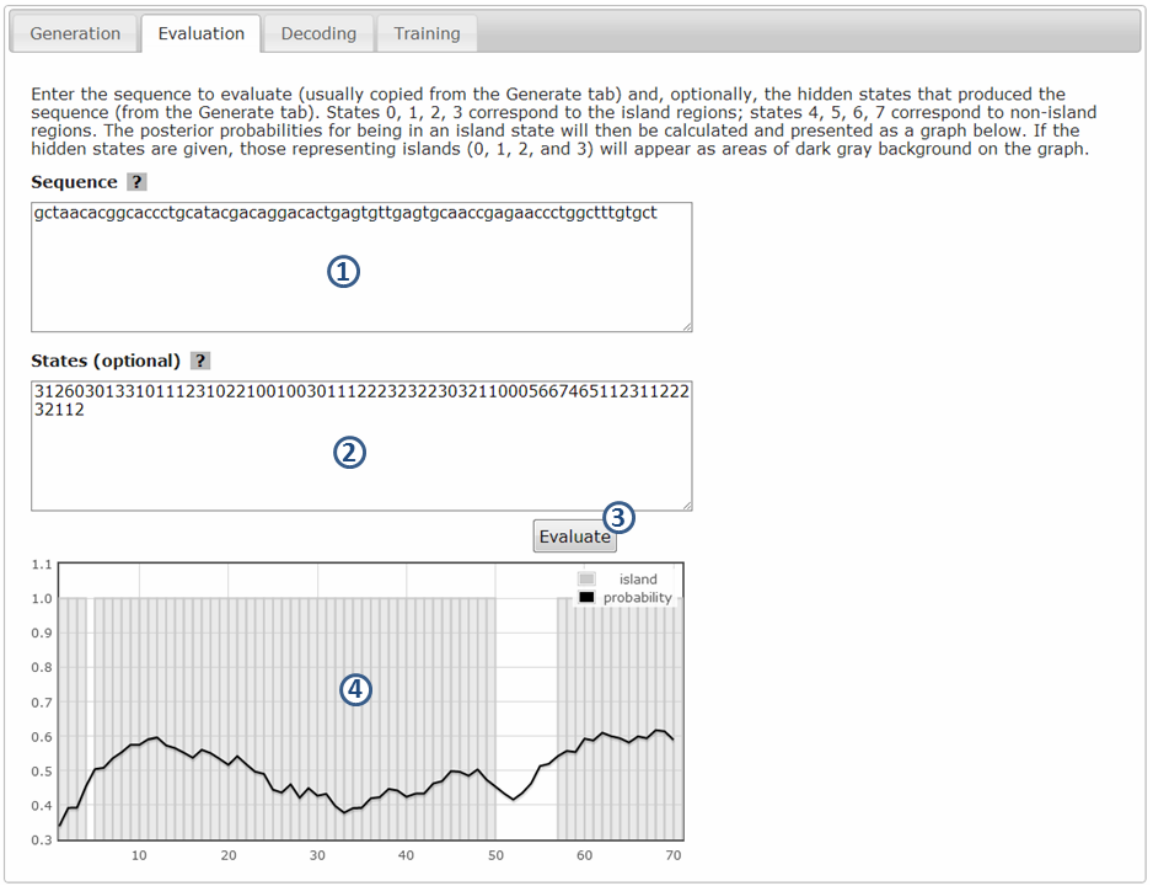


Figure 9: CpG Islands Evaluation Tab

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. **$\log_2$(probability)**: This output is the logarithm of the probability that the hidden state output above actually generated the input sequence.

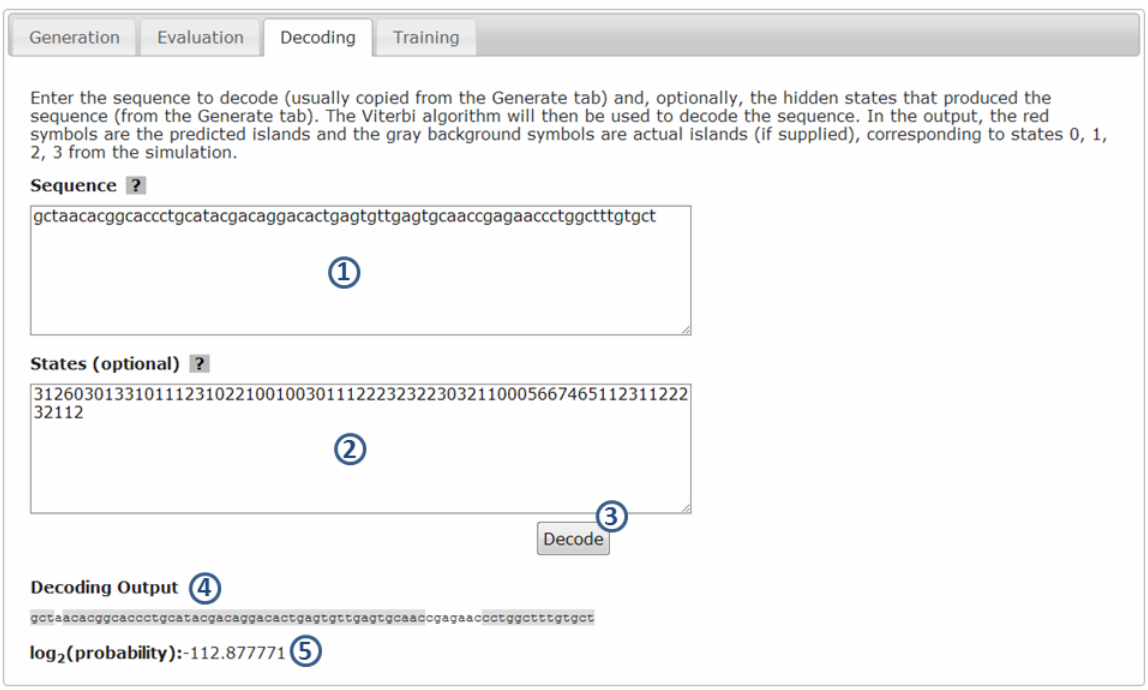


Figure 10: CpG Islands Decoding Tab

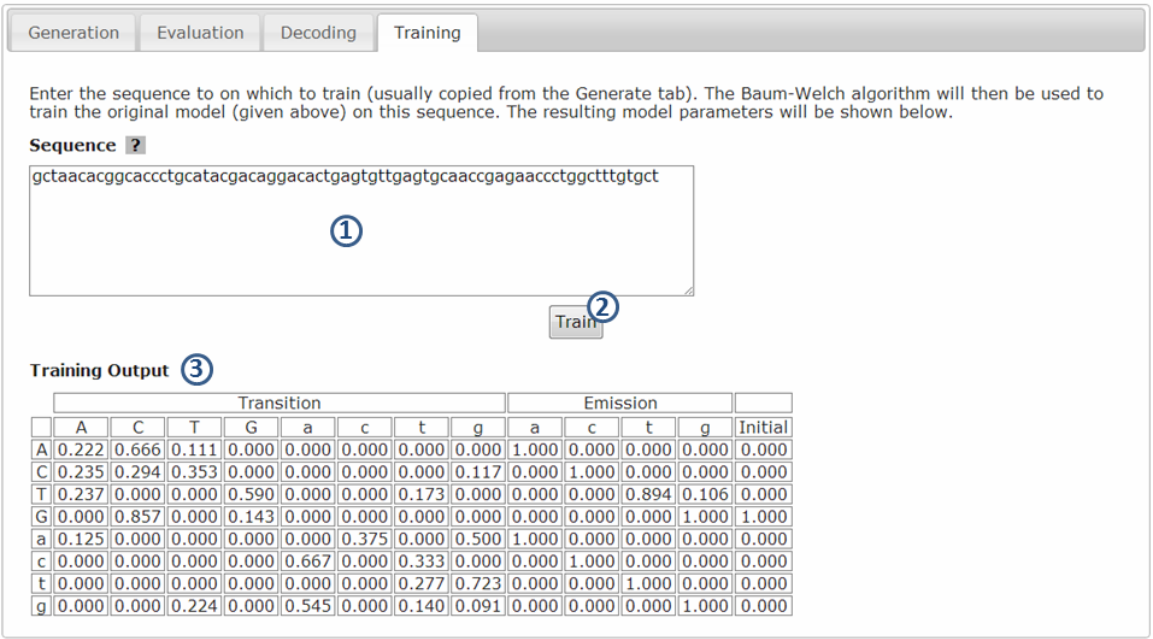
4.4 Training

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.



	Transition								Emission				
	A	C	T	G	a	c	t	g	a	c	t	g	Initial
A	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
C	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	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
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	0.545	0.000	0.140	0.091	0.000	0.000	0.000	1.000	0.000

Figure 11: CpG Islands Training Tab