CSEP 590A Assignment 4
Hidden Markov Model Viterbi Algorithm
Name: Dipak C Boyed

1. Introduction

In this project I have implemented the Hidden Markov Model (HMM) Viterbi algorithm in a console executable that traverses a given DNA sequence to find high GC content patches and output the following as the result to the console output:
Number and size of each high GC content patch in the DNA sequence and log probability (ln) of the Viterbi path.
In addition, I also implemented an E-M based Viterbi training model that iterates 10 times over the data to converge on the log probability of the Viterbi path and the GC content patches.

1.1 Supported Environment

NOTE: All development was done on the Windows platform (OS: Vista) in Visual Studio 2005 and the programs were compiled and linked against .NET 2.0 framework in C#.
I was able to successfully build and use the program in the Windows platform.

1.2 Source Code

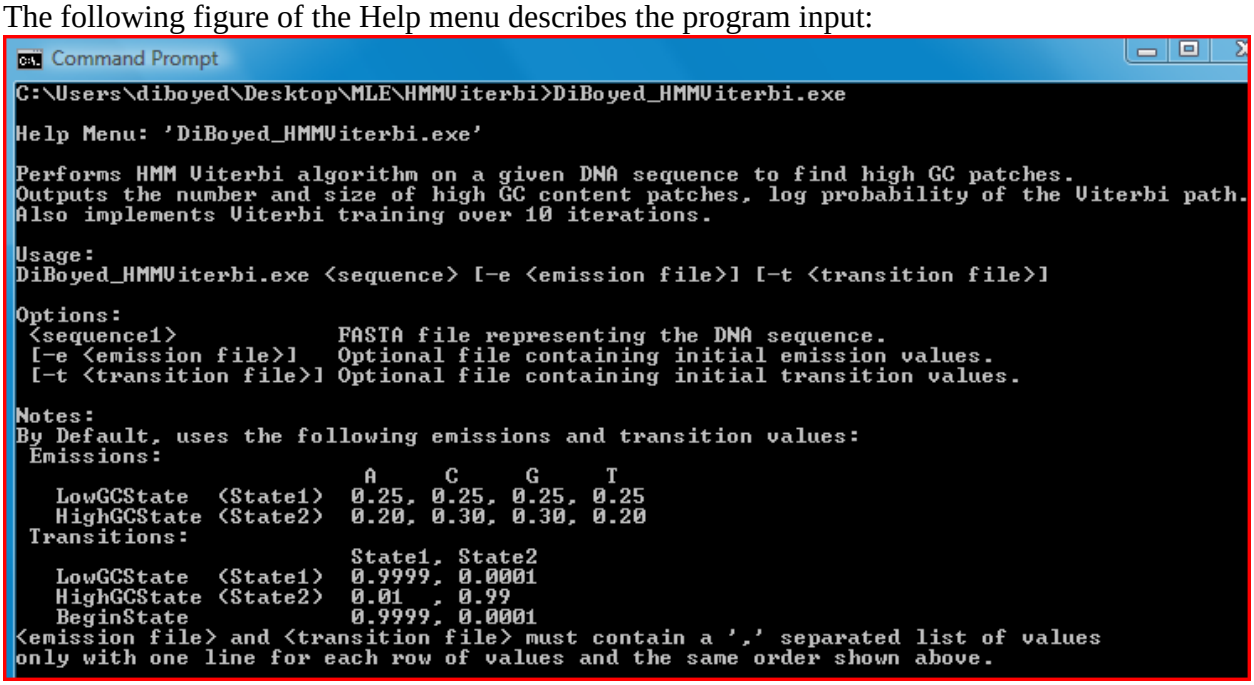
The following major classes define the source code for my sequence alignment application:

File Name	Description
Program.cs	Container class for the console application
HMMViterbiAlgorithm.cs	Class representing the Viterbi algorithm to perform and common functionality.

Table 1 List of source files

1.3 Program Input

The following figure of the Help menu describes the program input:



```
Command Prompt
C:\Users\diboyed\Desktop\MLE\HMMViterbi>DiBoyed_HMMViterbi.exe

Help Menu: 'DiBoyed_HMMViterbi.exe'

Performs HMM Viterbi algorithm on a given DNA sequence to find high GC patches.
Outputs the number and size of high GC content patches, log probability of the Viterbi path.
Also implements Viterbi training over 10 iterations.

Usage:
DiBoyed_HMMViterbi.exe <sequence> [-e <emission file>] [-t <transition file>]

Options:
 <sequence1>            FASTA file representing the DNA sequence.
 [-e <emission file>]   Optional file containing initial emission values.
 [-t <transition file>] Optional file containing initial transition values.

Notes:
By Default, uses the following emissions and transition values:
 Emissions:
                          A     C     G     T
   LowGCState  (State1)  0.25, 0.25, 0.25, 0.25
   HighGCState (State2)  0.20, 0.30, 0.30, 0.20
 Transitions:
                         State1, State2
   LowGCState  (State1)  0.9999, 0.0001
   HighGCState (State2)  0.01   , 0.99
   BeginState            0.9999, 0.0001
<emission file> and <transition file> must contain a ',' separated list of values
only with one line for each row of values and the same order shown above.
```

Figure 1 Program Help Menu

1.4 Program Output

The following figure shows an example of the results/output displayed by the application:

```
C:\Users\diboyed\Desktop\MLE\HMMViterbi>DiBoyed_HMMViterbi.exe HMMViterbi\InputFiles\NC_000909.fna
Performing HMM Viterbi algorithm...
     File 'NC_000909.fna' exists. Attempting to read FASTA file...
     Ignoring comments: '>gi|15668172|ref|NC_000909.1| Methanocaldococcus jannaschii DSM 2661, com
Reading line number: 23787
Found unknown base 'N' at position '122869'. Treating it as base 'T'.
Found unknown base 'S' at position '291994'. Treating it as base 'T'.
Found unknown base 'R' at position '325537'. Treating it as base 'T'.
Found unknown base 'Y' at position '353579'. Treating it as base 'T'.
Found unknown base 'Y' at position '706047'. Treating it as base 'T'.
Found unknown base 'R' at position '730710'. Treating it as base 'T'.
Found unknown base 'Y' at position '996602'. Treating it as base 'T'.
Found unknown base 'Y' at position '996614'. Treating it as base 'T'.
Found unknown base 'M' at position '996627'. Treating it as base 'T'.
Found unknown base 'N' at position '1162959'. Treating it as base 'T'.
Found unknown base 'R' at position '1279458'. Treating it as base 'T'.
Found unknown base 'N' at position '1374116'. Treating it as base 'T'.
Found unknown base 'M' at position '1546836'. Treating it as base 'T'.
     Successfully read sequence of length 1664970.
************************************
Displaying Results of iteration 1:

 Printing Emission values...
 ---------------------------------------------
                      |  A     C     G     T
 ---------------------------------------------
 LowGCState  (State1) | 0.25, 0.25, 0.25, 0.25,
 HighGCState (State2) | 0.20, 0.30, 0.30, 0.20,
 ---------------------------------------------

 Printing Transition values...
 ---------------------------------------
                      |  State1, State2
 ---------------------------------------
 LowGCState  (State1) | 0.9999 , 0.0001
 HighGCState (State2) | 0.01 , 0.99
 BeginState           | 0.9999 , 0.0001
 ---------------------------------------


 No. of hit areas: '2'
 Location,  Length
 154651...159579, 4929
 638464...643447, 4984
 Log Probability of Viterbi Path : -2.308117e+006
************************************

************************************
Displaying Results of iteration 2:
...
```

Reading the sequence

Results of pass #1

Emission

HMM Parameters

Transition

Number and size of hits

Ln P(Viterbi Path)

Results of pass #2

...

Figure 2 Sample program output

Besides writing current operations to console, the program writes a Results section at the end of each iteration (1 through 10):

(i) The HMM parameters (Emission and Transition values recalculated using Viterbi training).

NOTE: in order to make indexing easier, I list the BeginState transition values in the end of my transition matrix.

(ii) The number of hits (patches with high GC content) and size of each such patch.

(iii) The log probability of the Viterbi path.

2. Test Results

For all Dice examples I tested, I assumed the following:

- Rolls of 1-5 were treated as A
- Roll of 6 was treated as C
- State1 => Fair dice, State2 => Loaded dice.
- Hit area represents rolls where Viterbi predicts a '**Loaded** dice' was used.
- Emissions and Transitions were supplied using the –e/-t optional switches in my application.

2.1 Simple Test Case: Casino Dice rolling '316664':
First, I ran the simple casino dice example presented in the lec06-casino-hmm.xls spreadsheet.

Input	DiBoyed_HMMViterbi.exe HMMViterbi\InputFiles\hw4diceSimple.fasta -e HMMViterbi\InputFiles\hw4DiceEmissions.txt -t HMMViterbi\InputFiles\hw4DiceSimpleTransitions.txt
Output	OutputLogs\hw4DiceSimple.output.txt

The results (number and size of hit areas as well as log probability of Viterbi path) matched the results provided in the spreadsheet (http://www.cs.washington.edu/education/courses/csep590a/08au/slides/lec06-casino-hmm.xls)

2.2 Test Case: Casino Dice 300 rolls (Durbin fig 3.5)
Next, I ran the casino dice example of 300 rolls from Durbin fig 3.5.

Input	DiBoyed_HMMViterbi.exe HMMViterbi\InputFiles\hw4dice.fasta -e HMMViterbi\InputFiles\hw4DiceEmissions.txt -t HMMViterbi\InputFiles\hw4DiceTransitions.txt
Output	OutputLogs\hw4Dice.output.txt

The results (number and size of hit areas) exactly matched the results provided in the book and slides. The hit areas were found correctly in the very first iteration while the log probability of the Viterbi path converged after the 2nd iteration.

Results of 1st iteration:
No. of hit areas: '4' (Hit area is Loaded dice)

Location,	Length
49...66,	18
79...112,	34
180...192,	13
271...289,	19

Log Probability of Viterbi Path : -5.351774e+002

2.3 Test Results on NC_000909.fna

Finally, I ran the application on the NC_0009-9.fna (FASTA file) which is the genome sequence data for *Methanocaldococcus jannaschii*.

Input	DiBoyed_HMMViterbi.exe HMMViterbi\InputFiles\NC_000909.fna
Output	OutputLogs\NC_000909.output.txt

The two figures below summarize the results of the 1st iteration and the 10th iteration:

```
*************************************
Displaying Results of iteration 1:

 Printing Emission values...
 ---------------------------------------------
                        ¦  A     C     G     T
 ---------------------------------------------
 LowGCState  (State1) ¦ 0.25, 0.25, 0.25, 0.25,
 HighGCState (State2) ¦ 0.20, 0.30, 0.30, 0.20,
 ---------------------------------------------

 Printing Transition values...
 ---------------------------------------
                        ¦  State1, State2
 ---------------------------------------
 LowGCState  (State1) ¦ 0.9999 , 0.0001
 HighGCState (State2) ¦ 0.01 , 0.99
 BeginState           ¦ 0.9999 , 0.0001
 ---------------------------------------


 No. of hit areas: '2'
 Location,  Length
 154651...159579, 4929
 638464...643447, 4984
 Log Probability of Viterbi Path : -2.308117e+006
*************************************
```

Figure 3 Results of 1st iteration

As mentioned in the homework description, I saw 2 subsequences each of length around ~5000.

Viterbi training was used to re-compute the emission and transition values and the computation was re-run for 10 iterations. The below table displays the no. of hit areas and log probability of the Viterbi path in each iteration:

Iteration	No. of Hits	Ln P(Viterbi Path)
1	2	-2.308117E+06
2	23	-2.188057E+06
3	34	-2.187965E+06
4	35	-2.187960E+06
5	36	-2.187960E+06
6	36	-2.187960E+06
7	36	-2.187960E+06
8	36	-2.187960E+06
9	36	-2.187960E+06
10	36	-2.187960E+06

Table 2 Result summary for each iteration

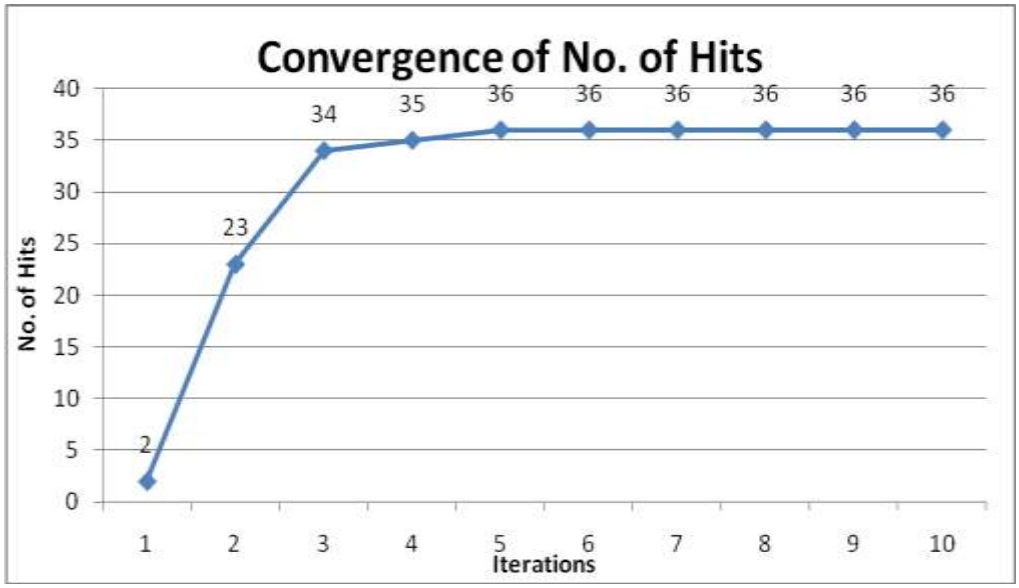
Finally, the figure below shows the result of the 10th iteration:

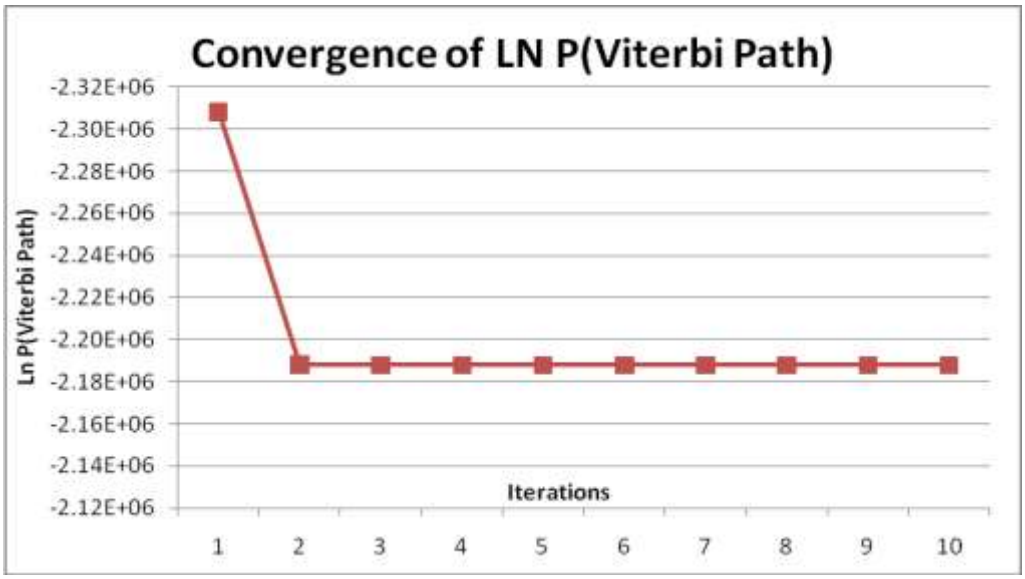
```
************************************
Displaying Results of iteration 10:


 Printing Emission values...
 ---------------------------------------------
                      |  A     C     G     T
 ---------------------------------------------
 LowGCState  (State1) | 0.35, 0.15, 0.16, 0.34,
 HighGCState (State2) | 0.18, 0.31, 0.32, 0.18,
 ---------------------------------------------

 Printing Transition values...
 ---------------------------------------
                      |  State1, State2
 ---------------------------------------
 LowGCState  (State1) | 0.999978185361408 , 2.18146385920347E-05
 HighGCState (State2) | 0.00244881300591796 , 0.997551186994082
 BeginState           | 0.9999 , 0.0001
 ---------------------------------------


 No. of hit areas: '36'
 Location,  Length
 97326...97541, 216
 97627...97823, 197
 111764...111856, 93
 118079...118179, 101
 138345...138419, 75
 154610...157697, 3088
 157782...159591, 1810
 186974...187067, 94
 190831...190907, 77
 215200...215296, 97
 227705...227782, 78
 291972...291997, 26
 303990...304080, 91
 358766...358942, 177
 359974...360046, 73
 402969...403057, 89
 412582...412635, 54
 552537...552862, 326
 619161...619236, 76
 637579...638153, 575
 638334...640132, 1799
 640217...643449, 3233
 643500...643767, 268
 763767...763845, 79
 764022...764095, 74
 774708...774788, 81
 863476...864151, 676
 873579...873778, 200
 883675...883755, 81
 951852...951968, 117
 1038544...1038622, 79
 1129124...1129194, 71
 1150142...1150402, 261
 1189943...1190054, 112
 1313165...1313251, 87
 1659451...1659520, 70
 Log Probability of Viterbi Path : -2.187960e+006
************************************
```

Figure 4 Results of 10th iteration

The results heavily overlapped the RNA location mentioned in the NC_000909.rnt file.

I also traced the below graphs to confirm the convergence of the hit areas and log probability over the iterations:


Convergence of No. of Hits
No. of Hits
Iterations
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
2
23
34
35
36
36
36
36
36
36


Convergence of LN P(Viterbi Path)
Ln P(Viterbi Path)
Iterations
-2.32E+06
-2.30E+06
-2.28E+06
-2.26E+06
-2.24E+06
-2.22E+06
-2.20E+06
-2.18E+06
-2.16E+06
-2.14E+06
-2.12E+06
1
2
3
4
5
6
7
8
9
10