# CSEP 590A Assignment 5

# Protein Coding Sequence Gene Prediction Algorithm

# Name: Dipak C Boyed

## 1. Introduction

In this project I have implemented a protein coding gene prediction algorithm in a console executable that traverses a given DNA sequence to find real genes and outputs the following as the result to the console output:

ORF length histogram with data on matching genes predicted using:

(a) Matching stop codon locations in a gene bank (this is for reference/comparison only).

(b) 3rd order Markov Model with training data for a trusted/background model coming from ORFs of length greater/lesser than given thresholds.

### 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

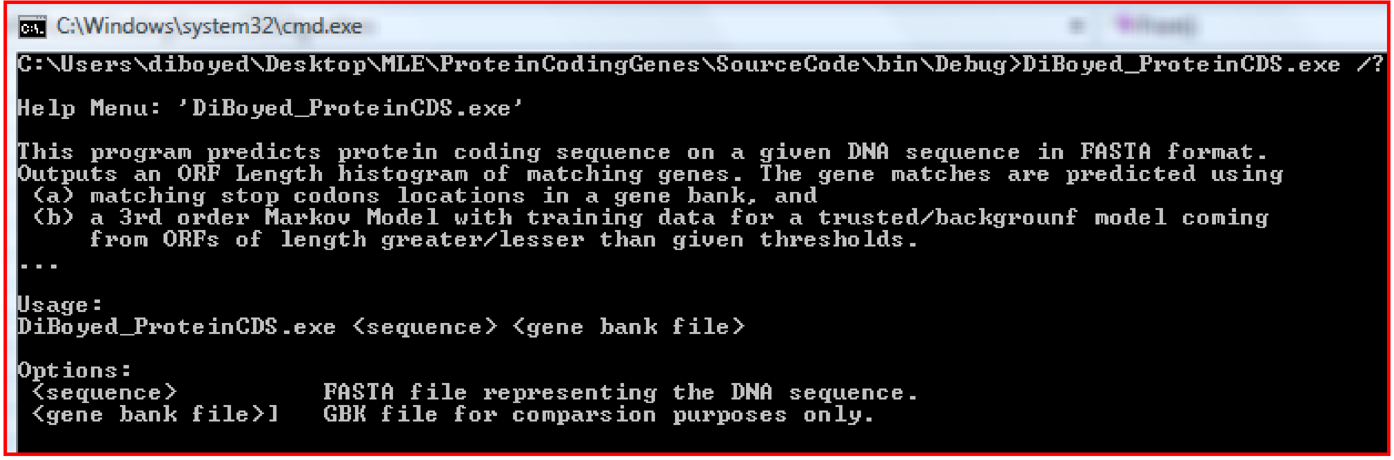
The following major classes define the source code for my sequence alignment application:

| File Name | Description |
|---|---|
| Program.cs | Container class for the console application |
| CDSAlgorithm.cs | Class representing the protein coding sequence prediction algorithm to perform. |

**Table 1 List of source files**

### 1.3 Program Input

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

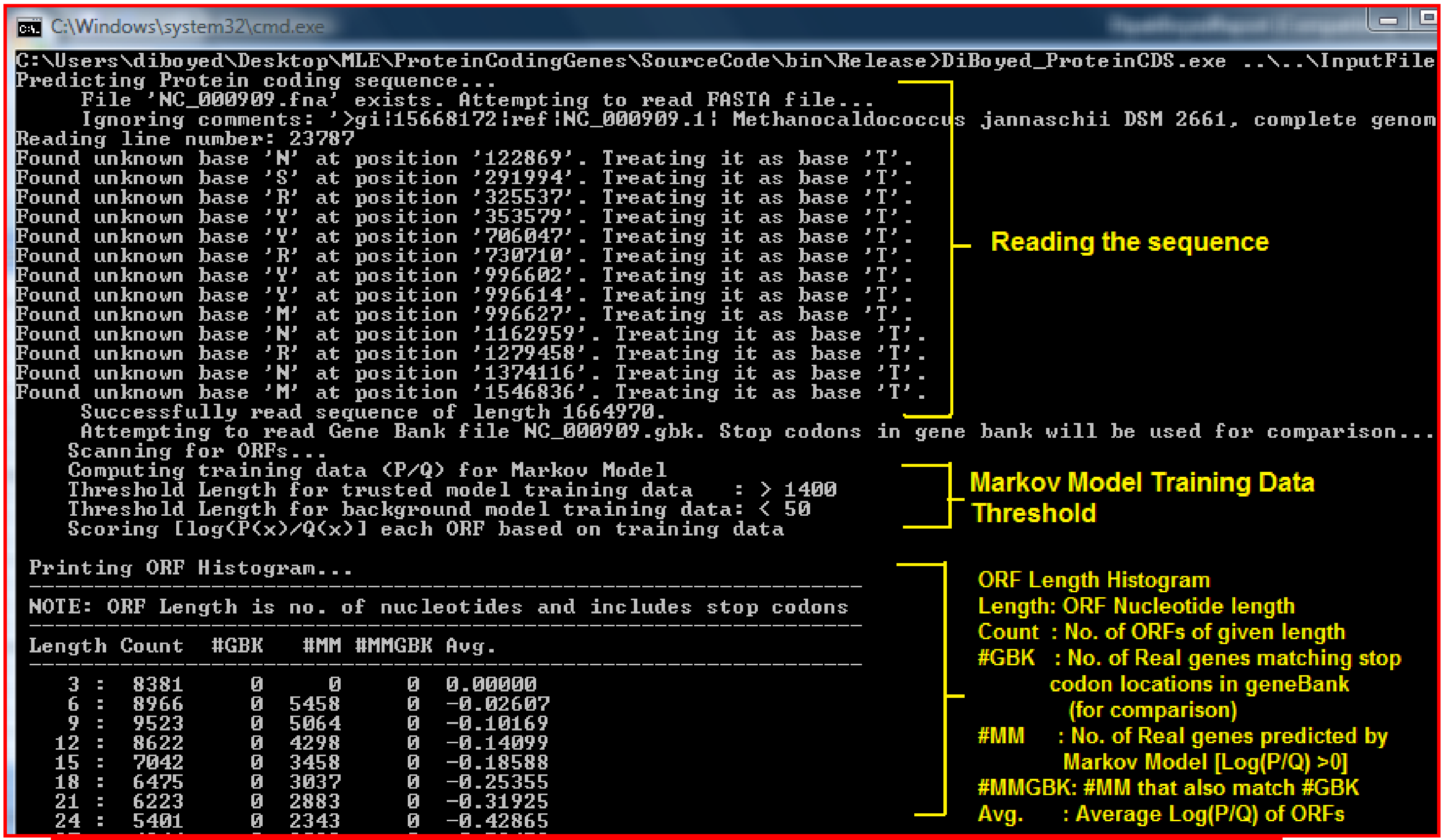


```
C:\Users\diboyed\Desktop\MLE\ProteinCodingGenes\SourceCode\bin\Debug>DiBoyed_ProteinCDS.exe /?

Help Menu: 'DiBoyed_ProteinCDS.exe'

This program predicts protein coding sequence on a given DNA sequence in FASTA format.
Outputs an ORF Length histogram of matching genes. The gene matches are predicted using
 (a) matching stop codons locations in a gene bank, and
 (b) a 3rd order Markov Model with training data for a trusted/backgrounf model coming
     from ORFs of length greater/lesser than given thresholds.
...

Usage:
DiBoyed_ProteinCDS.exe <sequence> <gene bank file>

Options:
 <sequence>           FASTA file representing the DNA sequence.
 <gene bank file>]    GBK file for comparsion purposes only.
```

**Figure 1 Program Help Menu**

### 1.4 Program Output

The following figures shows an example of the results/output displayed by the application (only parts of ORF length histogram are shown in the figure below):

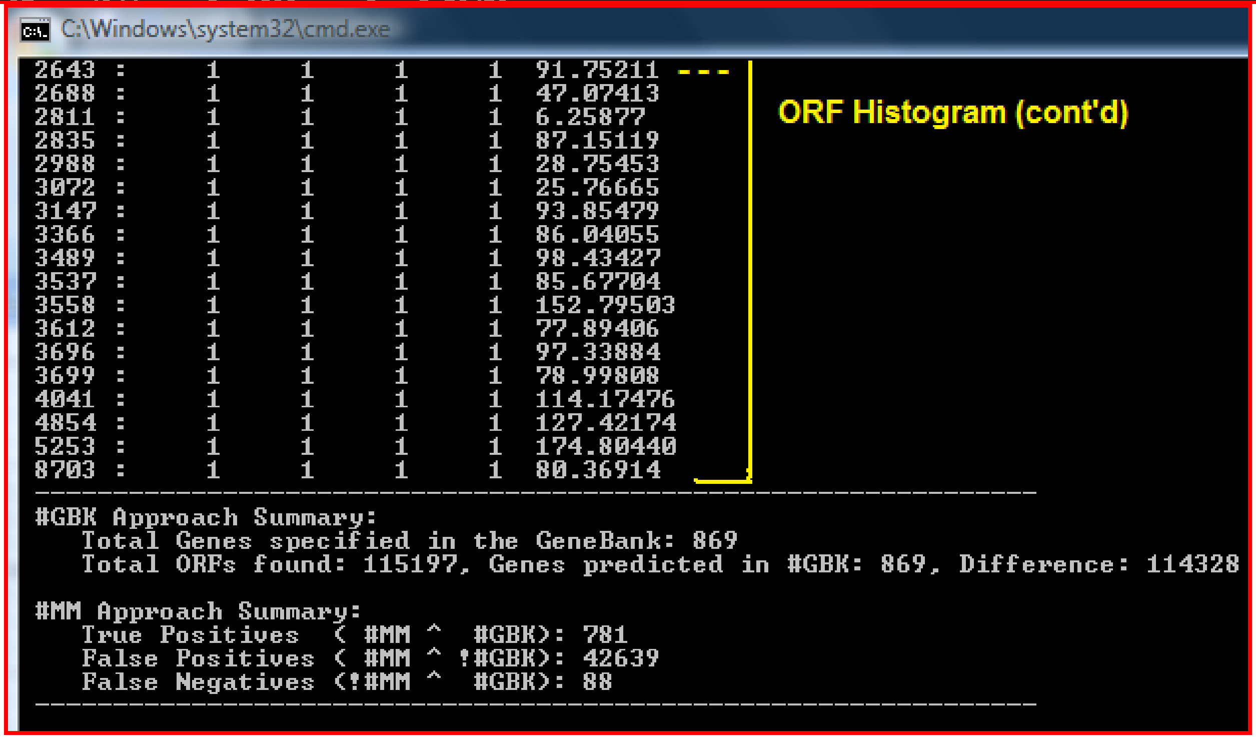


```
C:\Users\diboyed\Desktop\MLE\ProteinCodingGenes\SourceCode\bin\Release>DiBoyed_ProteinCDS.exe ..\..\InputFile
Predicting Protein coding sequence...
     File 'NC_000909.fna' exists. Attempting to read FASTA file...
     Ignoring comments: '>gi|15668172|ref|NC_000909.1| Methanocaldococcus jannaschii DSM 2661, complete genom
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.
     Attempting to read Gene Bank file NC_000909.gbk. Stop codons in gene bank will be used for comparison...
    Scanning for ORFs...
    Computing training data (P/Q) for Markov Model
    Threshold Length for trusted model training data    : > 1400
    Threshold Length for background model training data: < 50
    Scoring [log(P(x)/Q(x)] each ORF based on training data

 Printing ORF Histogram...
 -----------------------------------------------------------------
 NOTE: ORF Length is no. of nucleotides and includes stop codons
 -----------------------------------------------------------------
 Length Count   #GBK   #MM #MMGBK Avg.
 -----------------------------------------------------------------
     3 :   8381     0      0     0  0.00000
     6 :   8966     0   5458     0  -0.02607
     9 :   9523     0   5064     0  -0.10169
    12 :   8622     0   4298     0  -0.14099
    15 :   7042     0   3458     0  -0.18588
    18 :   6475     0   3037     0  -0.25355
    21 :   6223     0   2883     0  -0.31925
    24 :   5401     0   2343     0  -0.42865
```

Reading the sequence

Markov Model Training Data Threshold

ORF Length Histogram
Length: ORF Nucleotide length
Count : No. of ORFs of given length
#GBK : No. of Real genes matching stop codon locations in geneBank (for comparison)
#MM : No. of Real genes predicted by Markov Model [Log(P/Q) >0]
#MMGBK: #MM that also match #GBK
Avg. : Average Log(P/Q) of ORFs



```
2643 :       1      1      1      1   91.75211
2688 :       1      1      1      1   47.07413
2811 :       1      1      1      1   6.25877
2835 :       1      1      1      1   87.15119
2988 :       1      1      1      1   28.75453
3072 :       1      1      1      1   25.76665
3147 :       1      1      1      1   93.85479
3366 :       1      1      1      1   86.04055
3489 :       1      1      1      1   98.43427
3537 :       1      1      1      1   85.67704
3558 :       1      1      1      1   152.79503
3612 :       1      1      1      1   77.89406
3696 :       1      1      1      1   97.33884
3699 :       1      1      1      1   78.99808
4041 :       1      1      1      1   114.17476
4854 :       1      1      1      1   127.42174
5253 :       1      1      1      1   174.80440
8703 :       1      1      1      1   80.36914
-----------------------------------------------------------------
#GBK Approach Summary:
   Total Genes specified in the GeneBank: 869
   Total ORFs found: 115197, Genes predicted in #GBK: 869, Difference: 114328

#MM Approach Summary:
   True Positives  ( #MM ^  #GBK): 781
   False Positives ( #MM ^ !#GBK): 42639
   False Negatives (!#MM ^  #GBK): 88
-----------------------------------------------------------------
```

ORF Histogram (cont'd)

**Figure 2&3: Sample program output**

## 2. Results and Notes

- For all P/Q training data calculations and P/Q scoring, stop codons were not included.
- Stop codons were included in ORF length calculations.

### 2.1 Running the program on NC_000909.fna

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

| | |
|---|---|
| **Input** | DiBoyed_ProteinCDS.exe NC_000909.fna NC_000909.gbk |
| **Output** | OutputLogs\NC_000909.log.txt |

Shown below is the summary of the ORF Histogram output (Full details available in the Output log file):

```
----------------------------------------------------------------
 #GBK Approach Summary:
    Total Genes specified in the GeneBank: 869
    Total ORFs found: 115197, Genes predicted in #GBK: 869, Difference: 114328

 #MM Approach Summary:
    True Positives  ( #MM ^  #GBK): 781
    False Positives ( #MM ^ !#GBK): 42639
    False Negatives (!#MM ^  #GBK): 88
 ----------------------------------------------------------------
```