INFO 529

**HW1-Section 3**

Authors:

Chia-Hsuan Chou

Venkata Prudhvi Raj Indana

Jing Wang

January 29, 2016

**Contents**

1. Goal……………………………………………………………………………3
2. Procedure………………………………………………………………………3
3. Result…………………………………………………………………………..4
4. Appendix………………………………………………………………………7
5. Author Contributions………………………………………………………….9
6. **Goal**

Our goal is to find potential E.coli genes from a given DNA sequence file based on E.coli codon-usage.

1. **Procedure**

First, we collected the entire *E.coli* genome from NCBI and selected 1000 gene sequences from the genome as our input (attach in the folder which called “gene\_1000.fasta”). Next, we used these gene sequences to compute the codon usage table and build a probabilistic model (see Appendix A) in order to reference it later. Then, we input a given DNA sequence and the program predicted all possible open reading frames (ORFs) in that given DNA sequence. Moreover, it calculated the log-likelihood ratio of Pi = pi / p0 for every predicted ORF. Here notices that we assume every codon is independent to its adjacent codons so that we can multiply every probability of codon based on the codon-usage dictionary we made and the outcome will be pi. Also, for the random model of coding DNA, we assume the probability of every codon is 1/64 = 0.0156, so the multiplication of every codon in every predicted ORF will be p0. Finally, the result will be the likelihood of every predicted ORF and the program gives us two FASTA files with predicted genes and the translated protein respectively (see Appendix B and Appendix C).

1. **Result**

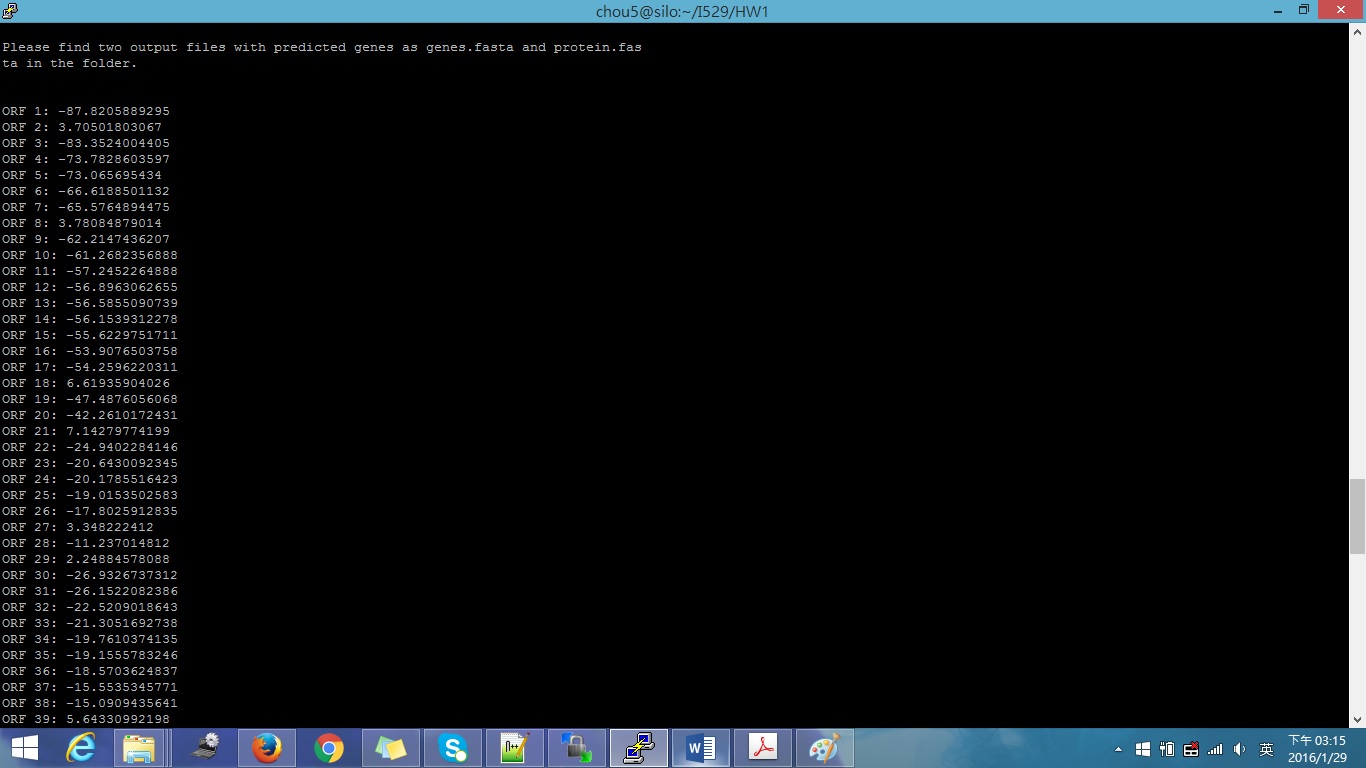
****

Figure 1. Total of 130 predicted ORFs and their likelihood (cont.)

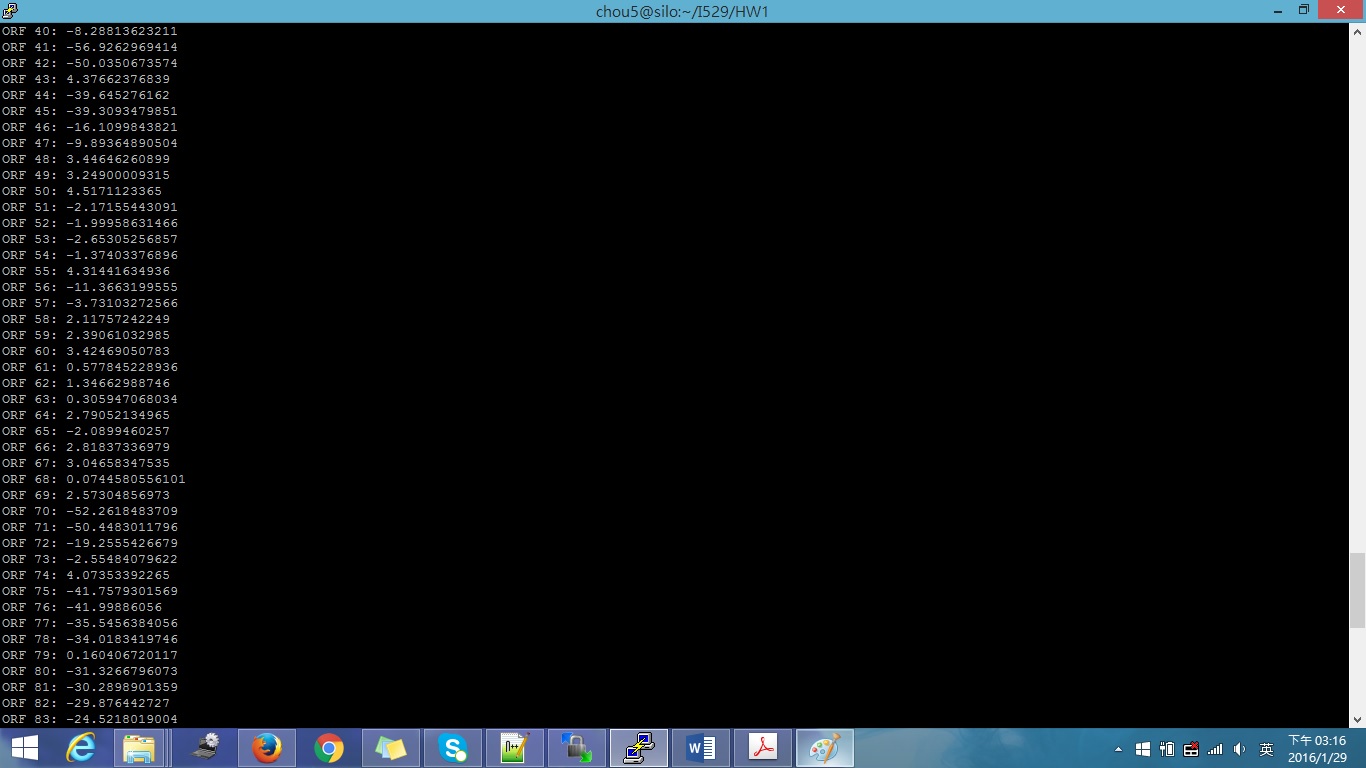


Figure 2. Total of 130 predicted ORFs and their likelihood (cont.)

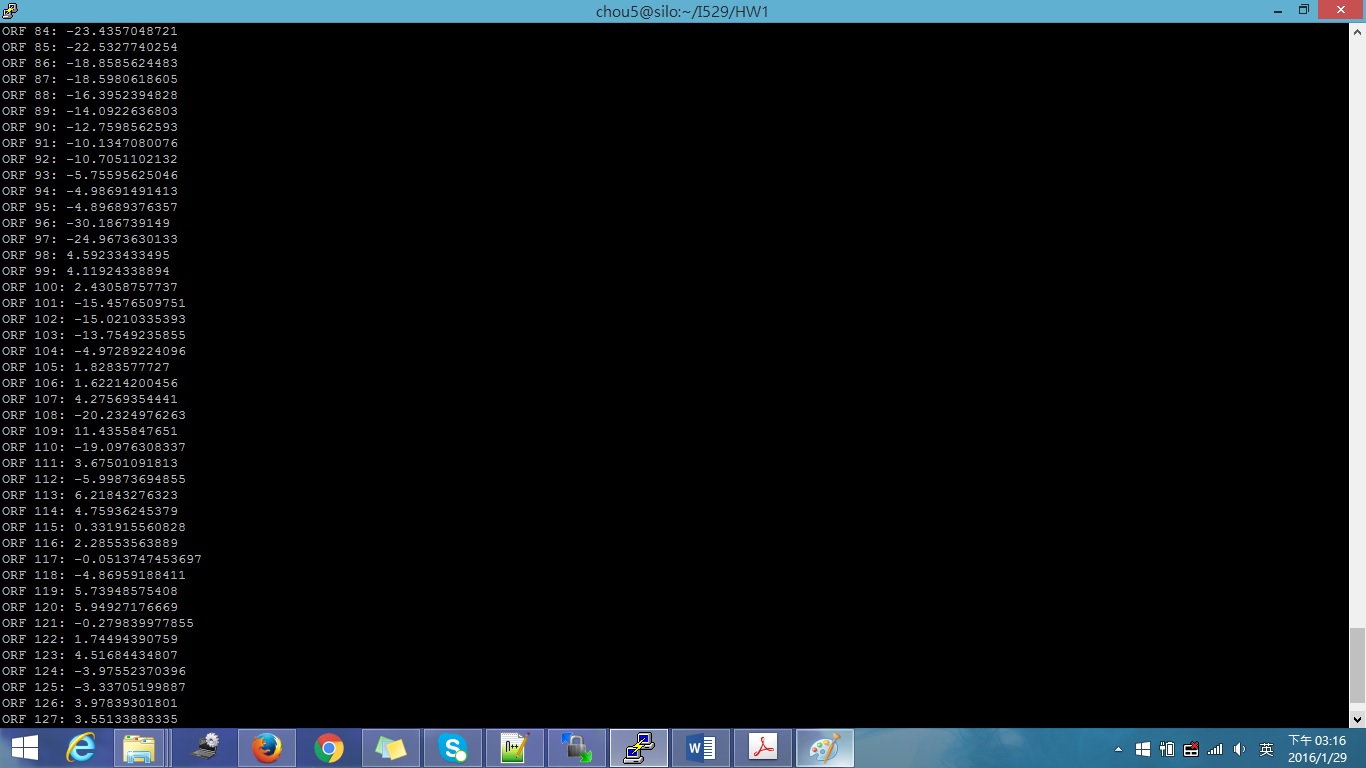


Figure 3. Total of 130 predicted ORFs and their likelihood (cont.)

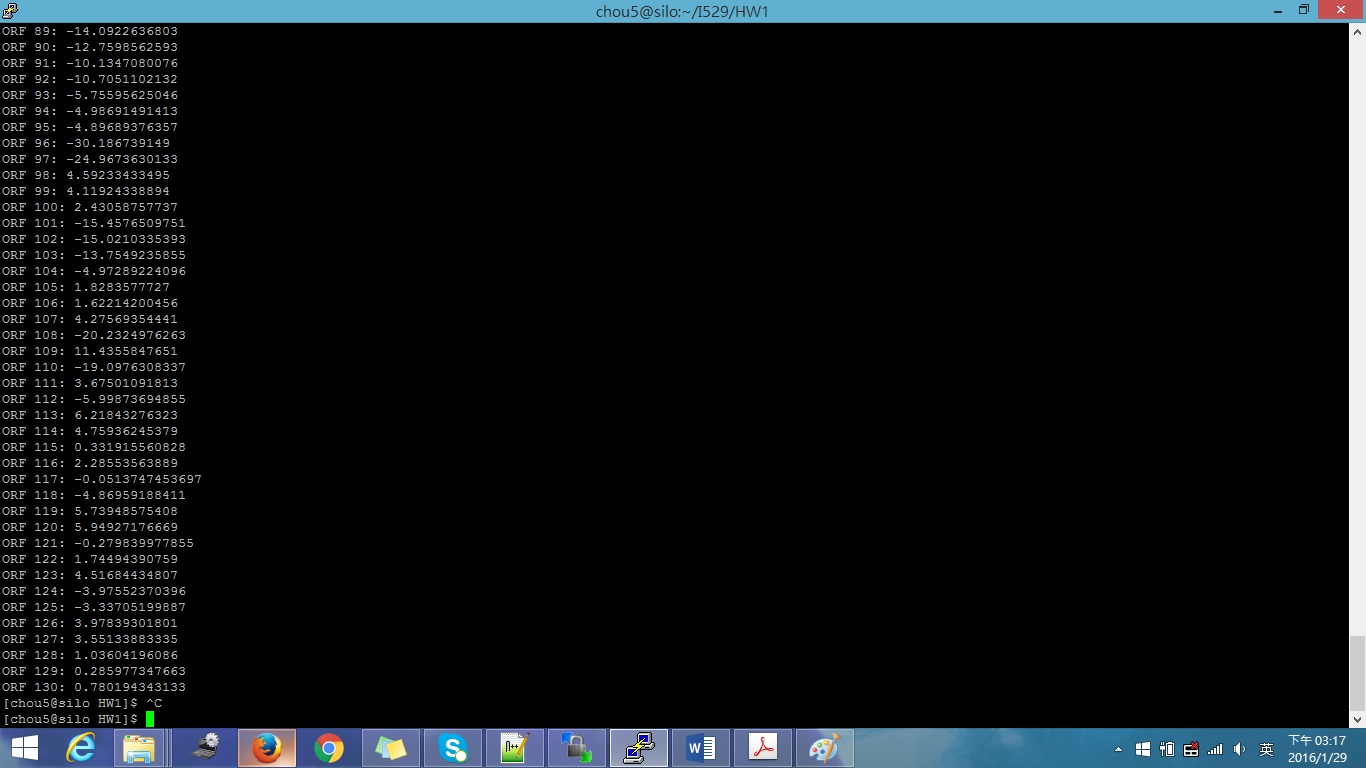


Figure 4. Total of 130 predicted ORFs and their likelihood

For our program, there are 130 predicted ORFs (Figure 1, 2, 3, 4) and 46 predicted genes (see Appendix B) which belong to *E.coli*. The two output files contain with the predicted genes and proteins which belongs to *E.coli* (see Appendix B and Appendix C).

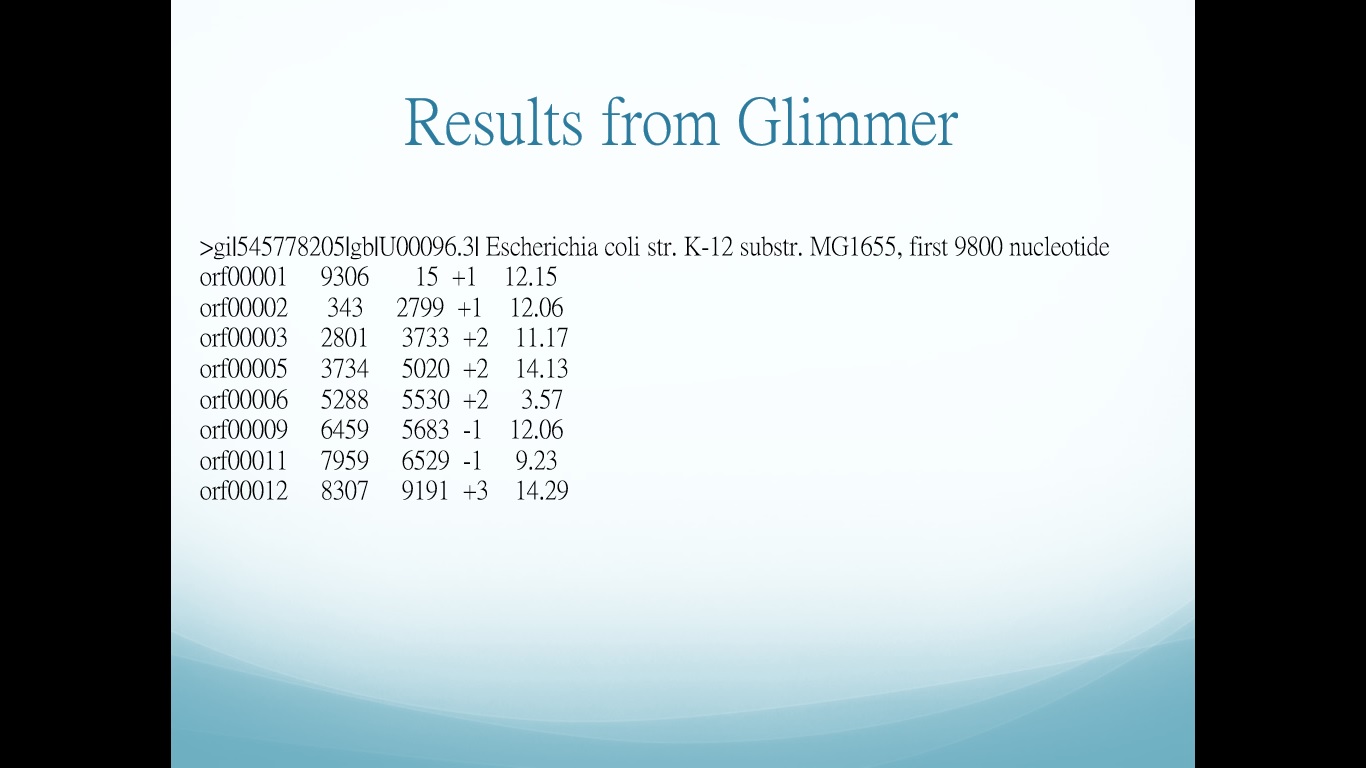


Figure 5. The 8 predicted genes from Glimmer

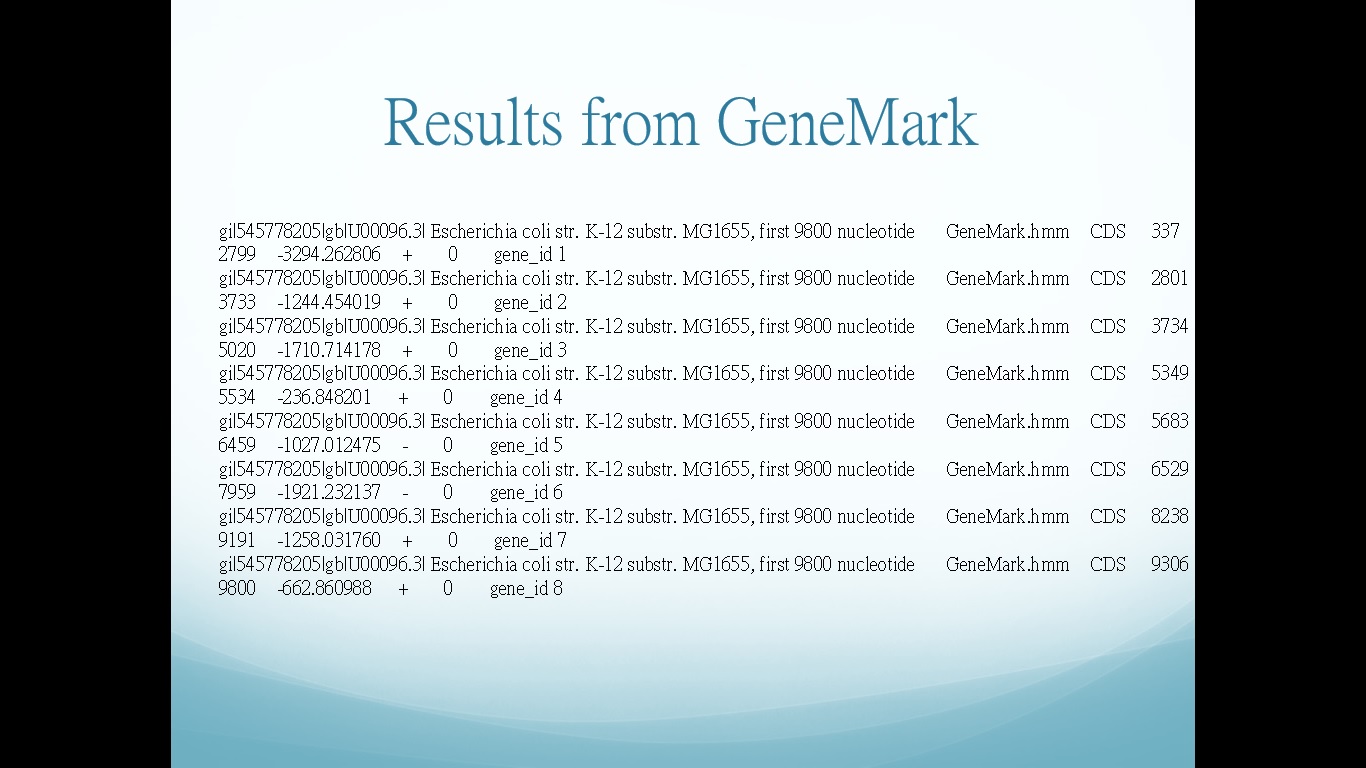


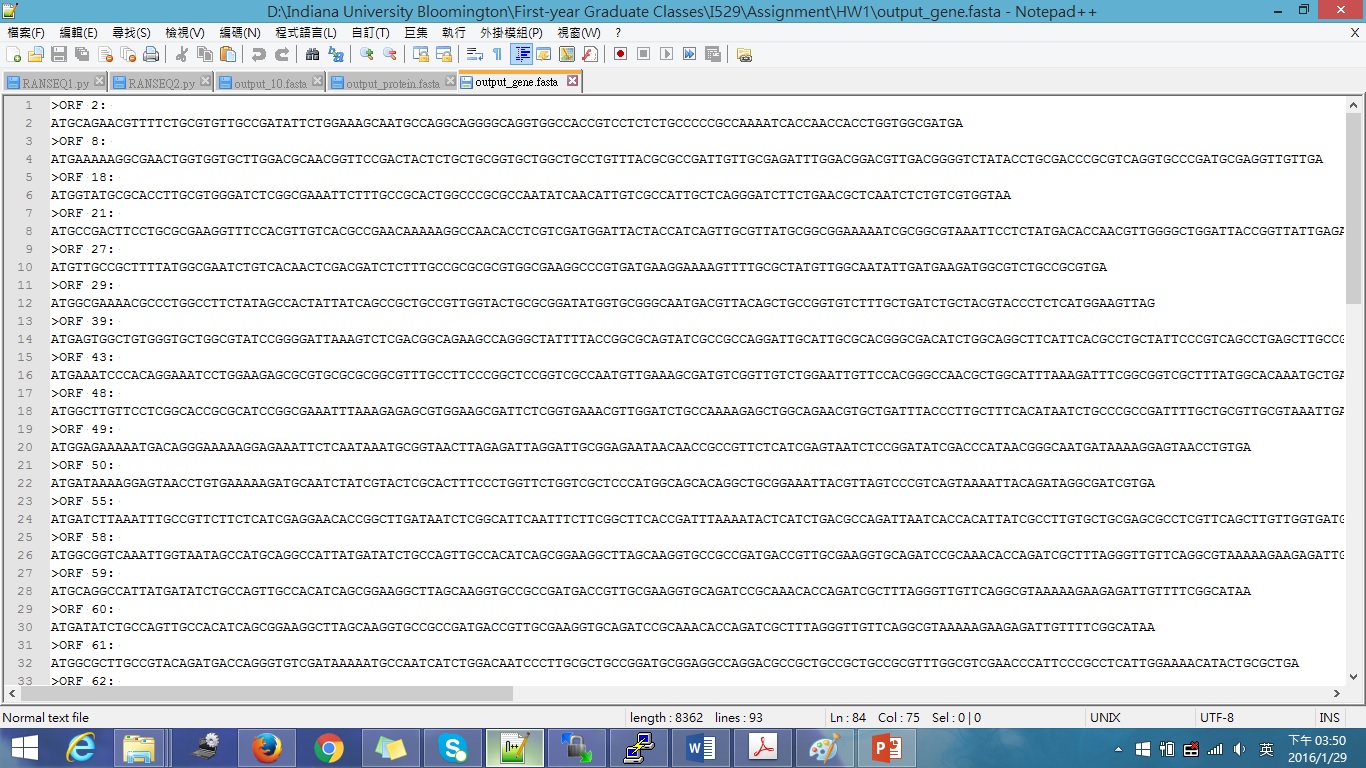
Figure 6. The 8 predicted genes from GeneMark

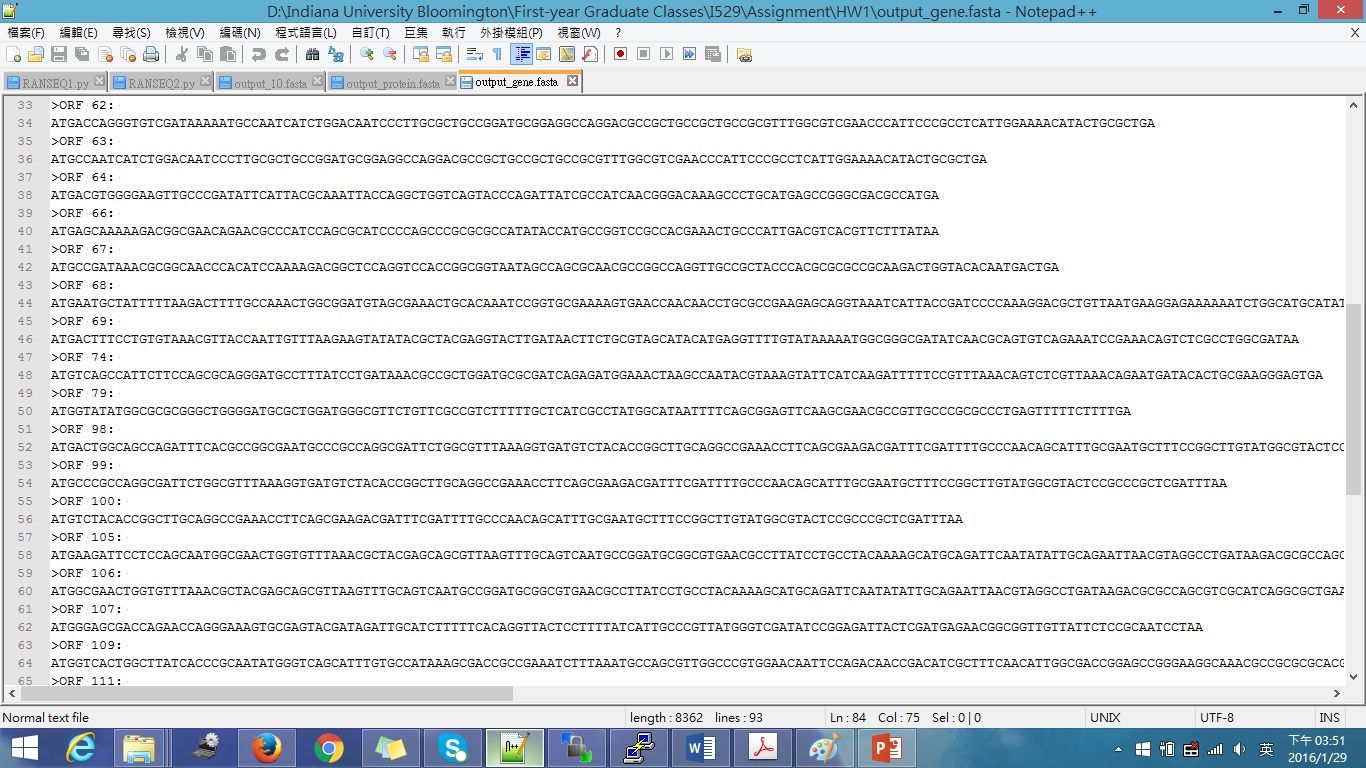
Moreover, we use Glimmer and GeneMark to predict our given DNA and they both predict 8 genes which are belong to *E.coli* (Figure 5, 6). By comparing our results with the two commercial available software, we found that our program contains many false positive results. Further optimization of the program to add more threshold to the score and prediction are required.

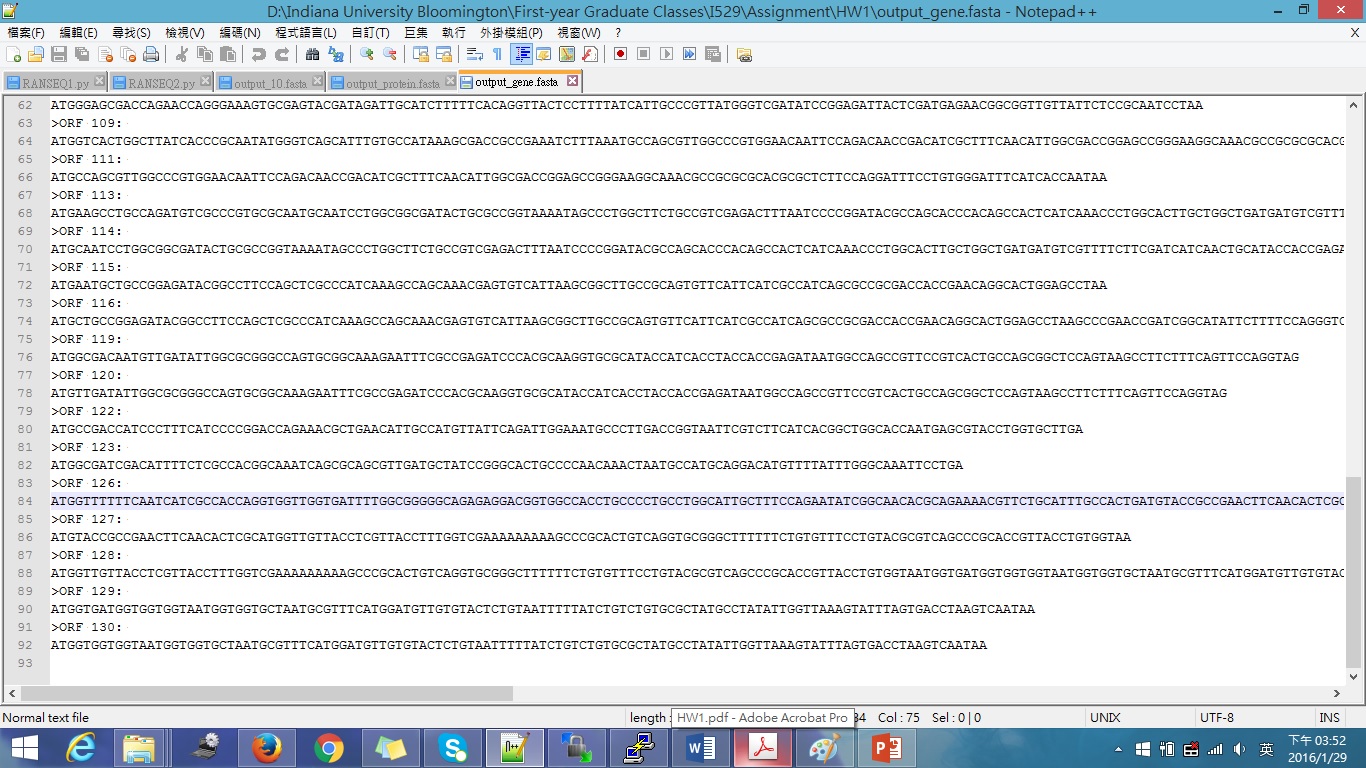
1. **Appendix**
2. Codon-Usage Table (frequency per 1000 bp)



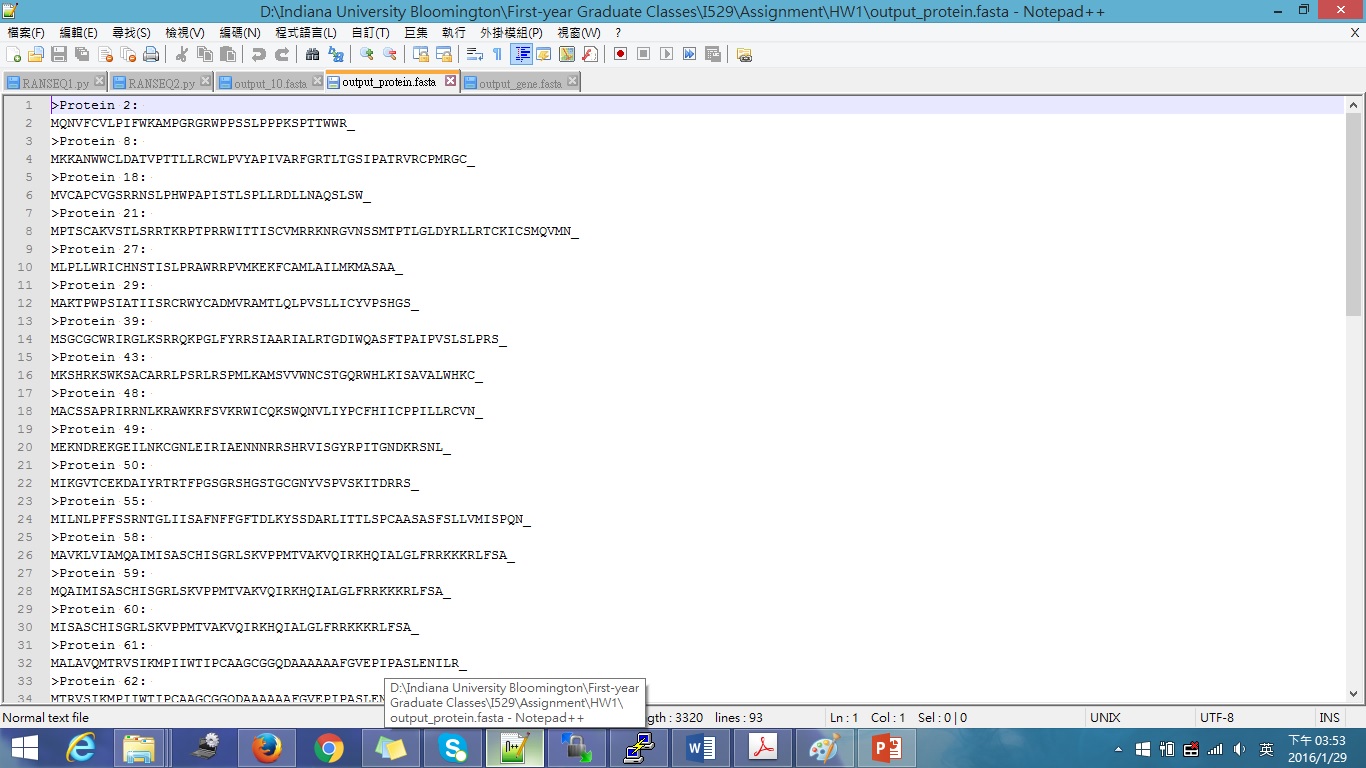
1. Output FASTA file for the predicted genes (Real file: output\_gene.fasta)

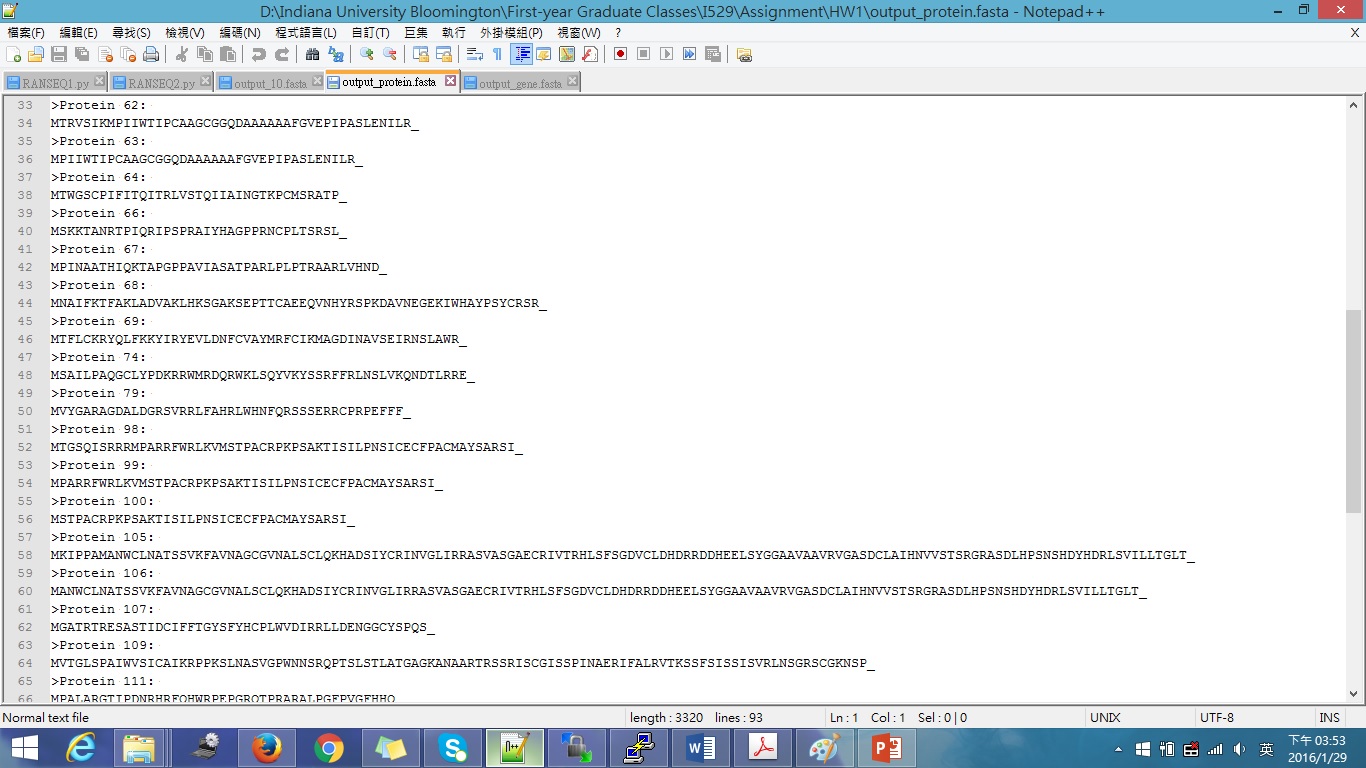


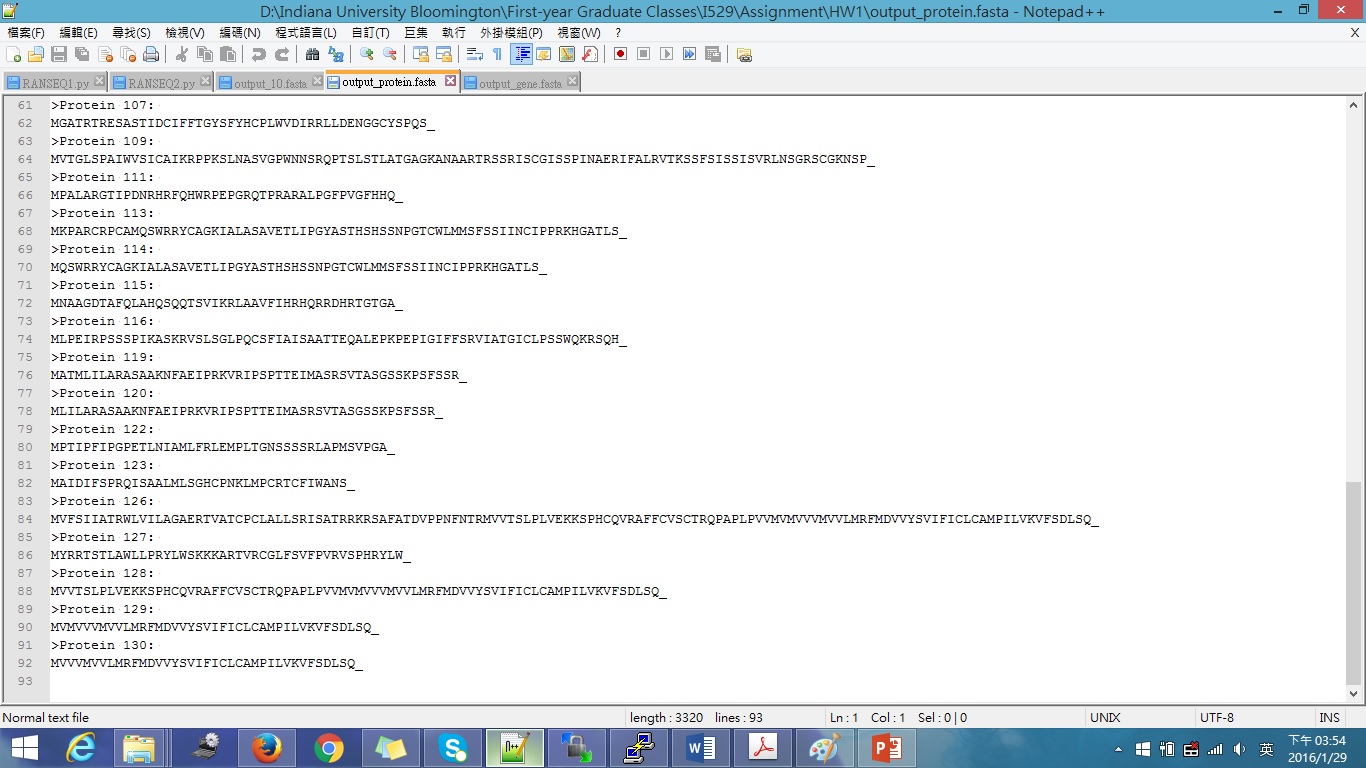




1. Output FASTA file for the predicted proteins (Real file: output\_protein.fasta)







1. **Author Contributions**

* Chia-Hsuan Chou: did coding for finding ORF and calculating likelihood and wrote report.
* Venkata Prudhvi Raj Indana: did coding for building codon usage table and debugging.
* Jing Wang: did coding for combining the above two codes, implemented Glimmer and GeneMark and made power point.
* Our group members are contributed evenly to the project.