

INFO 529

HW1-Section 3

Authors:
Chia-Hsuan Chou
Venkata Prudhvi Raj Indana
Jing Wang

January 29, 2016

Contents

I. Goal.....	3
II. Procedure.....	3
III. Result.....	4
IV. Appendix.....	7
V. Author Contributions.....	9

I. Goal

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

II. 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 $P_i = p_i / p_0$ 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 p_i . 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 p_0 . 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).

III. Result

```
chou5@
Please find two output files with predicted genes as genes.fasta and protein.fasta in the folder.

ORF 1: -87.8205889295
ORF 2: 3.70501803067
ORF 3: -83.3524004405
ORF 4: -73.7828603597
ORF 5: -73.065695434
ORF 6: -66.6188501132
ORF 7: -65.5764894475
ORF 8: 3.78084879014
ORF 9: -62.2147436207
ORF 10: -61.2682356888
ORF 11: -57.2452264888
ORF 12: -56.8963062655
ORF 13: -56.5855090739
ORF 14: -56.1539312278
ORF 15: -55.6229751711
ORF 16: -53.9076503758
ORF 17: -54.2596220311
ORF 18: 6.61935904026
ORF 19: -47.4876056068
ORF 20: -42.2610172431
ORF 21: 7.14279774199
ORF 22: -24.9402284146
ORF 23: -20.6430092345
ORF 24: -20.1785516423
ORF 25: -19.0153502583
ORF 26: -17.8025912835
ORF 27: 3.348222412
ORF 28: -11.237014812
ORF 29: 2.24884578088
ORF 30: -26.9326737312
ORF 31: -26.1522082386
ORF 32: -22.5209018643
ORF 33: -21.3051692738
ORF 34: -19.7610374135
ORF 35: -19.1555783246
ORF 36: -18.5703624837
ORF 37: -15.5535345771
ORF 38: -15.0909435641
ORF 39: 5.64330992198
```

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

```
chou5@
ORF 40: -8.28813623211
ORF 41: -56.9262969414
ORF 42: -50.0350673574
ORF 43: 4.37662376839
ORF 44: -39.645276162
ORF 45: -39.3093479851
ORF 46: -16.1099843821
ORF 47: -9.89364890504
ORF 48: 3.44646260899
ORF 49: 3.24900009315
ORF 50: 4.5171123365
ORF 51: -2.17155443091
ORF 52: -1.99958631466
ORF 53: -2.65305256857
ORF 54: -1.37403376896
ORF 55: 4.31441634936
ORF 56: -11.3663199555
ORF 57: -3.73103272566
ORF 58: 2.11757242249
ORF 59: 2.39061032985
ORF 60: 3.42469050783
ORF 61: 0.577845228936
ORF 62: 1.34662988746
ORF 63: 0.305947068034
ORF 64: 2.79052134965
ORF 65: -2.0899460257
ORF 66: 2.81837336979
ORF 67: 3.04658347535
ORF 68: 0.0744580556101
ORF 69: 2.57304856973
ORF 70: -52.2618483709
ORF 71: -50.4483011796
ORF 72: -19.2555426679
ORF 73: -2.55484079622
ORF 74: 4.07353392265
ORF 75: -41.7579301569
ORF 76: -41.99886056
ORF 77: -35.5456384056
ORF 78: -34.0183419746
ORF 79: 0.160406720117
ORF 80: -31.3266796073
ORF 81: -30.2898901359
ORF 82: -29.876442727
ORF 83: -24.5218019004
```

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

```

chou5@
ORF 84: -23.4357048721
ORF 85: -22.5327740254
ORF 86: -18.8585624483
ORF 87: -18.5980618605
ORF 88: -16.3952394828
ORF 89: -14.0922636803
ORF 90: -12.7598562593
ORF 91: -10.1347080076
ORF 92: -10.7051102132
ORF 93: -5.75595625046
ORF 94: -4.98691491413
ORF 95: -4.89689376357
ORF 96: -30.186739149
ORF 97: -24.9673630133
ORF 98: 4.59233433495
ORF 99: 4.11924338894
ORF 100: 2.43058757737
ORF 101: -15.4576509751
ORF 102: -15.0210335393
ORF 103: -13.7549235855
ORF 104: -4.97289224096
ORF 105: 1.8283577727
ORF 106: 1.62214200456
ORF 107: 4.27569354441
ORF 108: -20.2324976263
ORF 109: 11.4355847651
ORF 110: -19.0976308337
ORF 111: 3.67501091813
ORF 112: -5.99873694855
ORF 113: 6.21843276323
ORF 114: 4.75936245379
ORF 115: 0.331915560828
ORF 116: 2.28553563889
ORF 117: -0.0513747453697
ORF 118: -4.86959188411
ORF 119: 5.73948575408
ORF 120: 5.94927176669
ORF 121: -0.279839977855
ORF 122: 1.74494390759
ORF 123: 4.51684434807
ORF 124: -3.97552370396
ORF 125: -3.33705199887
ORF 126: 3.97839301801
ORF 127: 3.55133883335

```

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

```

ORF 128: 1.03604196086
ORF 129: 0.285977347663
ORF 130: 0.780194343133

```

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).

```

>gil545778205|gb|U00096.3| Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide
orf00001  9306   15 +1  12.15
orf00002   343  2799 +1  12.06
orf00003  2801  3733 +2  11.17
orf00005  3734  5020 +2  14.13
orf00006  5288  5530 +2   3.57
orf00009  6459  5683 -1  12.06
orf00011  7959  6529 -1   9.23
orf00012  8307  9191 +3  14.29

```

Figure 5. The 8 predicted genes from Glimmer

gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	337
2799 -3294.262806 + 0 gene_id 1			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	2801
3733 -1244.454019 + 0 gene_id 2			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	3734
5020 -1710.714178 + 0 gene_id 3			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	5349
5534 -236.848201 + 0 gene_id 4			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	5683
6459 -1027.012475 - 0 gene_id 5			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	6529
7959 -1921.232137 - 0 gene_id 6			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	8238
9191 -1258.031760 + 0 gene_id 7			
gil545778205 gb U000096.3 Escherichia coli str. K-12 substr. MG1655, first 9800 nucleotide	GeneMark.hmm	CDS	9306
9800 -662.860988 + 0 gene_id 8			

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.

IV. Appendix

A. Codon-Usage Table (frequency per 1000 bp)

CTT: 10.30967530779820, 'ATG': 27.736331578153628, 'ACA': 6.282709073348543, 'ACG': 14.471408539735407, 'ATC': 25.06987027072123, 'AAC': 21.873967187444848, 'ATA': 3.6643788364474137, 'AGG': 1.026796171333776, 'CCT': 6.741558612413324, 'ACT': 8.4614421993974000, 'AGC': 16.621263023465502, 'AAG': 10.07222869317726, 'AGA': 1.761597181444510, 'CAT': 12.616757955263774, 'AAT': 16.220170769038244, 'ATT': 29.55568604423566, 'CTG': 55.20312916133214, 'CTA': 3.757432239474537, 'CTC': 10.887248154173445, 'CAC': 9.972758710479301, 'AAA': 32.73554543733495, 'CCG': 24.453792567920964, 'AGT': 8.028262564615963, 'CCA': 8.272126655307735, 'CAA': 15.16449595538571, 'CCC': 5.6088740859107520, 'TAT': 16.034063962984000, 'GGT': 24.51796732862932, 'TGT': 4.935039098472962, 'CGA': 3.3627574611181170, 'CAG': 30.139676366681748, 'TCT': 8.025053826580544, 'GAT': 31.61890460100947, 'CGG': 5.5992478718044980, 'TTT': 21.931724472082376, 'TGC': 6.484859569579879, 'GGG': 11.34930643127365, 'TAG': 0.2085679723021733, 'GGA': 7.556578073409509, 'TAA': 1.9380777733925025, 'GGC': 30.61777833395903, 'TAC': 12.764359904893006, 'TTC': 16.425530003305000, 'TCG': 9.055058735949737, 'TTA': 12.908753116486817, 'TTG': 13.239253134134877, 'TCC': 8.660383957593318, 'ACC': 24.367156640964673, 'TCA': 6.43672849048609, 'GCA': 19.656729204970976, 'GTA': 10.707558824190034, 'GCC': 26.093457704019585, 'GTC': 15.254340620377413, 'GCG': 35.48543393368822, 'GTG': 27.46679758317851, 'GAG': 18.078030091545298, 'GTT': 17.57425821998466, 'GCT': 15.129199836996108, 'TGA': 0.9080728640233083, 'GAC': 19.84925348709606, 'CGT': 21.44078755266342, 'TGG': 15.745277539796373, 'GAA': 39.233239959056505, 'CGC': 22.63122936380351

B. Output FASTA file for the predicted genes (Real file: output_gene.fasta)

```

2 >ORF 2:
3 ATGCAGACAGCTTTTCTGCGTGTTCGGATATTTCTGGAAGAACAAATGCCAGGCAGGGCAGAGTGGCCACCGCTCTCTCTGCCCCGCCAAAATCACCAACCATCGTGGCGATGA
4 >ORF 8:
5 ATGAAAGAGGCGAATCGTGGTCTTGGACGCAACGGTTCGCAGTACTCTGCTGCGGTGCTGGCTGCCGTTTAAOCGCGGATGTTGCGAGATTTGGACGGAAGCTTGACGGGGCTATACCTGCGACCCGCGTCAGGTGCCCGATCGAGAGTTGTGTGA
6 >ORF 18:
7 ATGTTATGCCGACCTTTCGTGGGATCTCGGCGAAATTTCTTTCGCGCATGGCCGCGGCGCAATATCACATTTCGCGCATTCGTCAGGATCTTCTGAACGCTCAATCTCTGTGCTGGTAA
8 >ORF 21:
9 ATGCCGATCTCTGCGCGAAGGTTTCCACGTTTGCACGCGCAACAAAAGGCCACACCTCTGTCGATGGAATTACTACACACAGTTGCGTATGCGCGCGAAATAATCGCGGCTAAATTCCTATGACACCAACGTTGGGGCTGGATTACCGGTTATTGAGT
10 >ORF 27:
11 ATGTTGCGCGCTTTTATGGCGAATCTGTCAACAATCGACGATCTCTTTGGCGCGCGCTGGCGAAGGCCGCTGATGAAGGAAAAGTTTGGCGTATGTTGGCAATATTGATGAAGATGGCGCTTGGCCGCTGA
12 >ORF 29:
13 ATGGCGAAAACGCCCTGGCTTCTATAGCCATATTATCAGCGCTGCGGTTGACTGTCGCGGATATGTCGCGGCAATGACGTTACACGTCGCGGTGCTTTTGCATGTCGTAAGTACCTCTCATGAAGTTAG
14 >ORF 39:
15 ATGAGTGGCTGTGGGTGCTGGCGATTCGGGGATTAAAGTCTCGACGCGAAGCGCAGGGCTATTTTACCGCGCAGTATCGCGCCAGGATGCAATTGCGCAGCGGACATCTGGCAGGCTTCATTACGCGCTGCTATTCGGCTCAGGCTGACGTTGCCG
16 >ORF 43:
17 ATGAATCCCAAGGAAATCTGGAAGAGCGCGTGGCGCGCGGTTTGGCTTCCCGGCTCGGTGCGCAATGTTGAAGCGATGTCGGTGTCTGGAATTGTTCCAGGGCGAACGCTGGCATTTTAAAGATTTCGGCGGTGCGCTTATGGCACAATGCTGT
18 >ORF 48:
19 ATGGCTGTGCTTCGCGACCGCGCATTCGGCGGAAATTTAAGAGAGCGTGGAGCGATTCTCGGTGAAACGTTGGAATCTGCGAAAAGAGCTGGCAGAACGTCGTAATTACCTTCTTTACATAATCTGCCCGCGATTITGTCGCTTGGCTAAATTTG
20 >ORF 49:
21 ATGAGAAAAATGACAGGGAAGGAGGAAATTTCTCAATAATCGGTAACTTAGAGATTAGGATTGCGGAGATAACACCGCGGCTCTCATCGAGTAACTTCGAGATATCGACCCATAACGGGCAATGATAAAGAGGATAACCTGTGA
22 >ORF 50:
23 ATGATAAAGAGGATTAACCTGTGAAGAAGATGCAATCTATGCTACTGSCATTTTCCCTGGTTCTGGTCTGCCATGGCAGCAGAGGCTCGGAAATTAAGTTAGTCCGCTCAGTAAATACAGATAGGCGATGCTGA
24 >ORF 55:
25 ATGATCTTAAATTTGCCGTTCTCTCATCGAGGAACCGGGCTTGATAATCTCGGCACTCAATTTCTTGGGCTCACGATTTAAATAACTCATCTGAOCGAGATTAAACACCAATTATGCCCTGTGCTGCGAGCGCGCTGCTCAGCTTGTGGTGAATG
26 >ORF 58:
27 ATGGCGGTCAAAATGGTAAATGCAATGCAAGCCATTATGATATCTGCGAGTTGCCCATCAGCGGAAGGCTTAGCAAGGTGCGCGGATGACCGTTGCGAAGGTGCGAGATCGCGCAACACAGGATCGCTTTAGGTTGTTTCAGGCGTAAAGAGAGAGATTG
28 >ORF 59:
29 ATGACGCCATTATGATATCTGCGAGTTGCCCATCAGCGGAAGGCTTAGCAAGGTGCCCGGATGACCGTTGCGAAGGTGCGAGATCGCGCAACACAGGATCGCTTTAGGTTGTTTCAGGCGTAAAGAGAGAGATTGTTTTCGCGCATAA
30 >ORF 60:
31 ATGATATCTGCGAGTTGCCCATCAGCGGAAGGCTTAGCAAGGTGCCCGGATGACCGTTGCGAAGGTGCGAGATCGCGCAACACAGGATCGCTTTAGGTTGTTTCAGGCGTAAAGAGAGAGATTGTTTTCGCGCATAA
32 >ORF 61:
33 ATGGCGCTTGGCGTACAGATGACGAGGTTGCGATAAAATGCAATCATCTGGAACATCCCTTGGCTGCGGATGCGGAGGCGGAGCGCGCTGCGCGTGGCGGCTTGGCGTGAACCCATTCCCGGCTCAATTGGAAGAACATACTGCGCTGA

```



```

33 >Protein 62:
34 MTRVSIKMPIIWTIPCAAGCGGQDAAAAAAFGVEPIPASLENILR_
35 >Protein 63:
36 MPIIWTIPCAAGCGGQDAAAAAAFGVEPIPASLENILR_
37 >Protein 64:
38 MTWGSPIFITQITRLVSTQIIAINGTKPCMSRATP_
39 >Protein 66:
40 MSKKTANRTPIQRIIPSPRAIYHAGPPFNCLTSRSL_
41 >Protein 67:
42 MPINAATHIQKTAAGPPAVIASATPARLELPTRAARLVHND_
43 >Protein 68:
44 MNAIFKTFAKLADVAKLHKSAGKSEPTTCAEEQVNHYSRSPKDAVNEGEKINHAYPSYCRSR_
45 >Protein 69:
46 MFLCKRYQLFKKYIRYEVLDNFCVATMRFCKIMAGDINAVSEIRNSLAWR_
47 >Protein 74:
48 MSAILPAQGCLYPDKRRWMDQRNKLSQLVVKYSSRFRLNSLVKQNDILARE_
49 >Protein 79:
50 MVYGARAGDALDGRSVRRLLFAHRLWHNFQRSSSERRCPRPEFFF_
51 >Protein 98:
52 MTGSQISRRRMPARRFWRLKVMSTFACRFKPSAKTISILPNSICECFACMAYSARSI_
53 >Protein 99:
54 MPARRFWRLKVMSTFACRFKPSAKTISILPNSICECFACMAYSARSI_
55 >Protein 100:
56 MSTFACRFKPSAKTISILPNSICECFACMAYSARSI_
57 >Protein 105:
58 MKIPFAMANWCLNATSSVKFAVNAAGGVNALSCQKHADSIYCRINVGLIRRASVASGAECRIVTRHLSFGDVCLDHRRDDHEELSYGGAAVAAVRVGASDCLAIHNVVTSRGRASDLHPSNSHDYHDLRSVILLTGLT_
59 >Protein 106:
60 MANWCLNATSSVKFAVNAAGGVNALSCQKHADSIYCRINVGLIRRASVASGAECRIVTRHLSFGDVCLDHRRDDHEELSYGGAAVAAVRVGASDCLAIHNVVTSRGRASDLHPSNSHDYHDLRSVILLTGLT_
61 >Protein 107:
62 MGATRTRESASTIDCIFFTGYSFYHCELMVDIRRLDENGCGCSPQS_
63 >Protein 109:
64 MVTGLSPAIVWSICAIRPPKSLNASVGFNNNSRQPTSLSTLATGAGKANAARTSSRISCGISSPINAERIFALRVTKSSFSISSISVRLNSGRSCGKNSP_

65 >Protein 111:
66 MPALARGTIPDNRRHFQHNREPEGRQTPRARALPGFFVGFHHQ_
67 >Protein 113:
68 MKPARCGRPCAMQSWRRYCAGKIALASAVETLIPGYASTHSHSSNPGTCWLMSFSSIINCIPPRKHGATLS_
69 >Protein 114:
70 MQSWRRYCAGKIALASAVETLIPGYASTHSHSSNPGTCWLMSFSSIINCIPPRKHGATLS_
71 >Protein 115:
72 MNAAGDTAFQLAHQSQTISVIKRLAAVFIHRHQRRDHRGTIGA_
73 >Protein 116:
74 MLPEIRPSSSPIKASKRVLSGLPQCSFIAISAATTEQALEPKPEPIGFIFFSVIATGICLPSSWQKRSHQ_
75 >Protein 119:
76 MATMLILARASAAKNFAEIPRKVRIPSPTTEIMASRSVTASGSSKPSFSSR_
77 >Protein 120:
78 MLILARASAAKNFAEIPRKVRIPSPTTEIMASRSVTASGSSKPSFSSR_
79 >Protein 122:
80 MPTIPIPGPETLNIAMLFREMLPTGNSSSSRLAPMSVPGA_
81 >Protein 123:
82 MAIDIFSFRQISAAMLSGHCENKLMPCRTCIFWANS_
83 >Protein 126:
84 MVFSIIATRNLVLILAGAERTVATCECLALLSRISATRKRKSAFATDVPPNFNTRMNVVLSLPLVEKKSPPHCQVRAFFCVSCTRQAPAPLVVMVMVVMVLMRFMDVVYSVIFICLCAMPILVKVFSDSLQ_
85 >Protein 127:
86 MYRRTSTLAWLLFRYLNSKKKARTVRCGLFSVFPVRVSPHRYLW_
87 >Protein 128:
88 MVVLSLPLVEKKSPPHCQVRAFFCVSCTRQAPAPLVVMVMVVMVLMRFMDVVYSVIFICLCAMPILVKVFSDSLQ_
89 >Protein 129:
90 MVVMVVMVLMRFMDVVYSVIFICLCAMPILVKVFSDSLQ_
91 >Protein 130:
92 MVVMVVMVLMRFMDVVYSVIFICLCAMPILVKVFSDSLQ_

```

V. 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.