I529 HW2-Section1

Chia-Hsuan Chou

Feb. 19, 2016

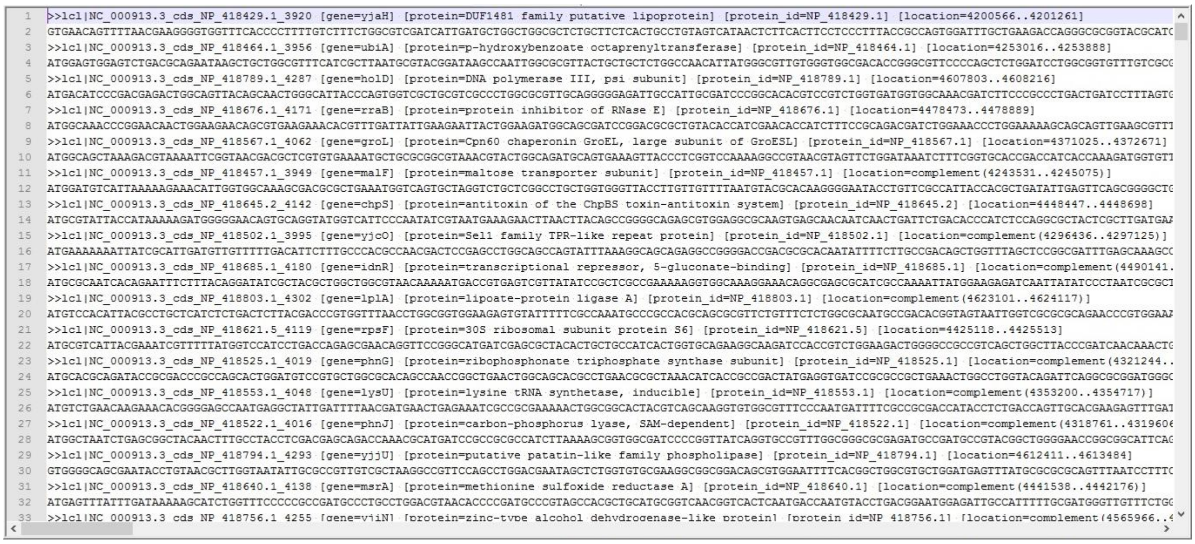
I. Goal

The goal is to test the prediction accuracy of possible *E. coli* genes based on a first order markov chain which generated from 1000 *E coli* genes.

II. Procedure

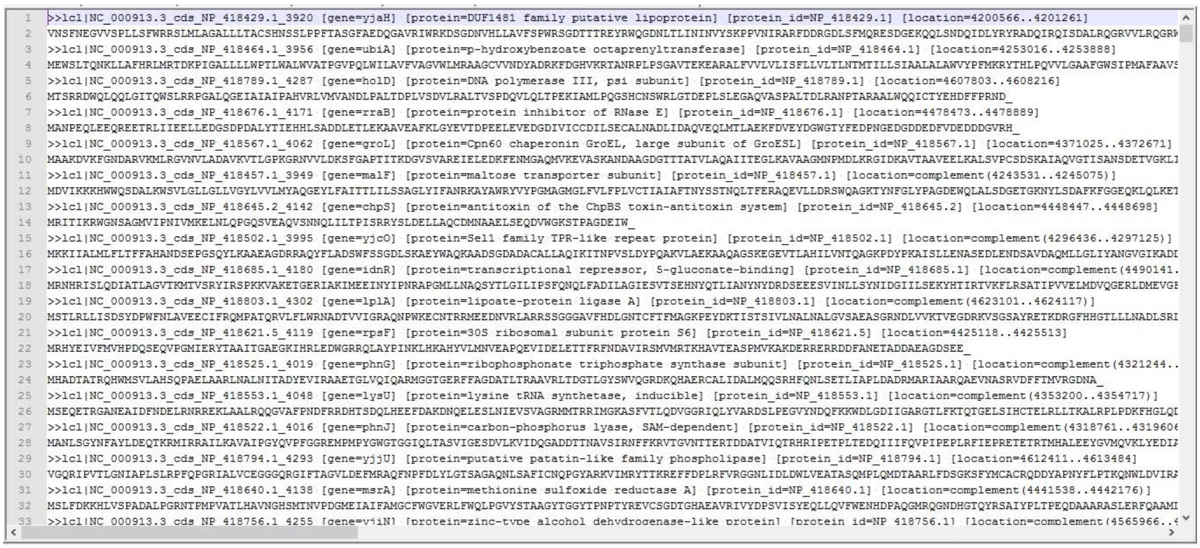
First, I collected the entire *E.coli* genome from NCBI and selected 1000 gene sequences from the genome as my training data (attached in the folder which called "gene_1000.fasta"). Next, I used these gene sequences to build a first order markov chain in order to reference it later. Then, I input a test data (attached in the folder which called "500_gene.fasta") in order to test the accuracy of my first order markov chain. In order to test my prediction, I calculated the log-likelihood ratio of $P_i = p_i / p_0$ for my test data. Here notices that I assume every nucleotide is independent to its adjacent codons so that we can multiply every probability of nucleotide based on the markov chain dictionary I made and the outcome will be p_i. Also, for the random model of coding DNA, I assume the probability of every nucleotide is 1/4 = 0.25, so the multiplication of every nucleotide in predicted sequence will be p_0. Moreover, I assume that the initial probability for every nucleotide is 0.25. Finally, the result will be the predicted genes and their likelihood and the program gives us two FASTA files with predicted genes and the translated protein respectively.

III. Result



```
>>lcl|NC_000913.3_cds_NP_418429.1_3920 [gene=yjaH] [protein=DUF1481 family putative lipoprotein] [protein_id=NP_418429.1] [location=4200566..4201261]
GTGAACAGTTTTAACGAAGGGGTGGTTTCACCCCTTTTGTCTTTCTGGCGTCGATCATTGATGCTGGCTGGCGCTCTGCTTCTCACTGCCTGTAGTCATAACTCTTCACTTCCTCCCTTTACCGCCAGTGGATTTGCTGAAGACCAGGGCGCGGTACGCATC
>>lcl|NC_000913.3_cds_NP_418464.1_3956 [gene=ubiA] [protein=p-hydroxybenzoate octaprenyltransferase] [protein_id=NP_418464.1] [location=4253016..4253888]
ATGGAGTGGAGTCTGACGCAGAATAAGCTGCTGGCGTTTCATCGCTTAATGCGTACGGATAAGCCAATTGGCGCGTTACTGCTGCTCTGGCCAACATTATGGGCGTTGTGGGTGGCGACACCGGGCGTTCCCCAGCTCTGGATCCTGGCGGTGTTTGTCGCC
>>lcl|NC_000913.3_cds_NP_418789.1_4287 [gene=holD] [protein=DNA polymerase III, psi subunit] [protein_id=NP_418789.1] [location=4607803..4608216]
ATGACATCCCGACGAGACTGGCAGTTACAGCAACTGGGCATTACCCAGTGGTCGCTGCGTCGCCCTGGCGCGTTGCAGGGGGAGATTGCCATTGCGATCCCGGCACACGTCCGTCTGGTGATGGTGGCAAACGATCTTCCCGCCCTGACTGATCCTTTAGTG
>>lcl|NC_000913.3_cds_NP_418676.1_4171 [gene=rraB] [protein=protein inhibitor of RNase E] [protein_id=NP_418676.1] [location=4478473..4478889]
ATGGCAAACCCGGAACAACTGGAAGAACAGCGTGAAGAAACACGTTTGATTATTGAAGAATTACTGGAAGATGGCAGCGATCCGGACGCGCTGTACACCATCGAACACCATCTTTCCGCAGACGATCTGGAAACCCTGGAAAAAGCAGCAGTTGAAGCGTTT
>>lcl|NC_000913.3_cds_NP_418567.1_4062 [gene=groL] [protein=Cpn60 chaperonin GroEL, large subunit of GroESL] [protein_id=NP_418567.1] [location=4371025..4372671]
ATGGCAGCTAAAGACGTAAAATTCGGTAACGACGCTCGTGTGAAAATGCTGCGCGGCGTAAACGTACTGGCAGATGCAGTGAAAGTTACCCTCGGTCCAAAAGGCCGTAACGTAGTTCTGGATAAATCTTTCGGTGCACCGACCATCACCAAAGATGGTGTT
>>lcl|NC_000913.3_cds_NP_418457.1_3949 [gene=malF] [protein=maltose transporter subunit] [protein_id=NP_418457.1] [location=complement(4243531..4245075)]
ATGGATGTCATTAAAAAGAAACATTGGTGGCAAAGCGACGCGCTGAAATGGTCAGTGCTAGGTCTGCTCGGCCTGCTGGTGGGTTACCTTGTTGTTTTAATGTACGCACAAGGGGAATACCTGTTCGCCATTACCACGCTGATATTGAGTTCAGCGGGGCTG
>>lcl|NC_000913.3_cds_NP_418645.2_4142 [gene=chpS] [protein=antitoxin of the ChpBS toxin-antitoxin system] [protein_id=NP_418645.2] [location=4448447..4448698]
ATGCGTATTACCATAAAAAGATGGGGAACAGTGCAGGTATGGTCATTCCCAATATCGTAATGAAAGAACTTAACTTACAGCCGGGGCAGAGCGTGGAGGCGCAAGTGAGCAACAATCAACTGATTCTGACACCCATCTCCAGGCGCTACTCGCTTGATGAA
>>lcl|NC_000913.3_cds_NP_418502.1_3995 [gene=yjcO] [protein=Sel1 family TPR-like repeat protein] [protein_id=NP_418502.1] [location=complement(4296436..4297125)]
ATGAAAAAAATTATCGCATTGATGTTGTTTTTGACATTCTTTGCCCACGCCAACGACTCCGAGCCTGGCAGCCAGTATTTAAAGGCAGCAGAGGCCGGGGACCGACGCGCACAATATTTTCTTGCCGACAGCTGGTTTAGCTCCGGCGATTTGAGCAAAGCC
>>lcl|NC_000913.3_cds_NP_418685.1_4180 [gene=idnR] [protein=transcriptional repressor, 5-gluconate-binding] [protein_id=NP_418685.1] [location=complement(4490141.
ATGCGCAATCACAGAATTTCTTTACAGGATATCGCTACGCTGGCTGGCGTAACAAAAATGACCGTGAGTCGTTATATCCGCTCGCCGAAAAAGGTGGCAAAGGAAACAGGCGAGCGCATCGCCAAAATTATGGAAGAGATCAATTATATCCCTAATCGCGCT
>>lcl|NC_000913.3_cds_NP_418803.1_4302 [gene=lplA] [protein=lipoate-protein ligase A] [protein_id=NP_418803.1] [location=complement(4623101..4624117)]
ATGTCCACATTACGCCTGCTCATCTCTGACTCTTACGACCCGTGGTTTAACCTGGCGGTGGAAGAGTGTATTTTTCGCCAAATGCCCGCCACGCAGCGCGTTCTGTTTCTCTGGCGCAATGCCGACACGGTAGTAATTGGTCGCGCGCAGAACCCGTGGAAA
>>lcl|NC_000913.3_cds_NP_418621.5_4119 [gene=rpsF] [protein=30S ribosomal subunit protein S6] [protein_id=NP_418621.5] [location=4425118..4425513]
ATGCGTCATTACGAAATCGTTTTTATGGTCCATCCTGACCAGAGCGAACAGGTTCCGGGCATGATCGAGCGCTACACTGCTGCCATCACTGGTGCAGAAGGCAAGATCCACCGTCTGGAAGACTGGGGCCGCCGTCAGCTGGCTTACCCGATCAACAAACTG
>>lcl|NC_000913.3_cds_NP_418525.1_4019 [gene=phnG] [protein=ribophosphonate triphosphate synthase subunit] [protein_id=NP_418525.1] [location=complement(4321244..
ATGCACGCAGATACCGCGACCCGCCAGCACTGGATGTCCGTGCTGGCGCACAGCCAACCGGCTGAACTGGCAGCACGCCTGAACGCGCTAAACATCACCGCCGACTATGAGGTGATCCGCGCCGCTGAAACTGGCCTGGTACAGATTCAGGCGCGGATGGGC
>>lcl|NC_000913.3_cds_NP_418553.1_4048 [gene=lysU] [protein=lysine tRNA synthetase, inducible] [protein_id=NP_418553.1] [location=complement(4353200..4354717)]
ATGTCTGAACAAGAAACACGGGGAGCCAATGAGGCTATTGATTTTAACGATGAACTGAGAAATCGCCGCGAAAAACTGGCGGCACTACGTCAGCAAGGTGTGGCGTTTCCCAATGATTTTCGCCGCGACCATACCTCTGACCAGTTGCACGAAGAGTTTGAT
>>lcl|NC_000913.3_cds_NP_418522.1_4016 [gene=phnJ] [protein=carbon-phosphorus lyase, SAM-dependent] [protein_id=NP_418522.1] [location=complement(4318761..4319606
ATGGCTAATCTGAGCGGCTACAACTTTGCCTACCTCGACGAGCAGACCAAACGCATGATCCGCCGCGCCATCTTAAAAGCGGTGGCGATCCCCGGTTATCAGGTGCCGTTTGGCGGGCGCGAGATGCCGATGCCGTACGGCTGGGGAACCGGCGGCATTCAG
>>lcl|NC_000913.3_cds_NP_418794.1_4293 [gene=yjjU] [protein=putative patatin-like family phospholipase] [protein_id=NP_418794.1] [location=4612411..4613484]
GTGGGGCAGCGAATACCTGTAACGCTTGGTAATATTGCGCCGTTGTCGCTAAGGCCGTTCCAGCCTGGACGAATAGCTCTGGTGTGCGAAGGCGGCGGACAGCGTGGAATTTTCACGGCTGGCGTGCTGGATGAGTTTATGCGCGCGCAGTTTAATCCTTTC
>>lcl|NC_000913.3_cds_NP_418640.1_4138 [gene=msrA] [protein=methionine sulfoxide reductase A] [protein_id=NP_418640.1] [location=complement(4441538..4442176)]
ATGAGTTTATTTGATAAAAAGCATCTGGTTTCCCCCGCCGATGCCCTGCCTGGACGTAACACCCCGATGCCCGTAGCCACGCTGCATGCGGTCAACGGTCACTCAATGACCAATGTACCTGACGGAATGGAGATTGCCATTTTTGCGATGGTTGTTTCTGG
>>lcl|NC_000913.3_cds_NP_418756.1_4255 [gene=yjjN] [protein=zinc-type alcohol dehydrogenase-like protein] [protein_id=NP_418756.1] [location=complement(4565966..4
```

Figure 1. The output file (“output_gene.fasta”) from ECgnfinder_mc.py



```
>>lcl|NC_000913.3_cds_NP_418429.1_3920 [gene=yjaH] [protein=DUF1481 family putative lipoprotein] [protein_id=NP_418429.1] [location=4200566..4201261]
VNSFNEGVVSPLLSFWRRSLMLAGALLLTACSHNSSLPPFTASGFAEDQGAVRIWRKDSGDNVHLLAVFSPWRSGDTTTREYRWQGDNLTLININVYSKPPVNIRARFDDRGDLSFMQRESDGEKQQLSNDQIDLYRYRADQIRQISDALRQGRVVLRQGRW
>>lcl|NC_000913.3_cds_NP_418464.1_3956 [gene=ubiA] [protein=p-hydroxybenzoate octaprenyltransferase] [protein_id=NP_418464.1] [location=4253016..4253888]
MEWSLTQNKLLAFHRLMRTDKPIGALLLLWPTLWALWVATPGVPQLWILAVFVAGVWLMRAAGCVVNDYADRKFDGHVKRTANRPLPSGAVTEKEARALFVVLVLISFLLVLTLNTMTILLSIAALALAWVYPFMKRYTHLPQVVLGAAFGWSIPMAFAAVS
>>lcl|NC_000913.3_cds_NP_418789.1_4287 [gene=holD] [protein=DNA polymerase III, psi subunit] [protein_id=NP_418789.1] [location=4607803..4608216]
MTSRRDWQLQQLGITQWSLRRPGALQGEIAIAIPAHVRLVMVANDLPALTDPLVSDVLRALTVSPDQVLQLTPEKIAMLPQGSHCNSWRLGTDEPLSLEGAQVASPALTDLRANPTARAALWQQICTYEHDFFPRND_
>>lcl|NC_000913.3_cds_NP_418676.1_4171 [gene=rraB] [protein=protein inhibitor of RNase E] [protein_id=NP_418676.1] [location=4478473..4478889]
MANPEQLEEQREETRLIIEELLEDGSDPDALYTIEHHLSADDLETLEKAAVEAFKLGYEVTDPEELEVEDGDIVICCDILSECALNADLIDAQVEQLMTLAEKFDVEYDGWGTYFEDPNGEDGDDEDFVDEDDDGVRH_
>>lcl|NC_000913.3_cds_NP_418567.1_4062 [gene=groL] [protein=Cpn60 chaperonin GroEL, large subunit of GroESL] [protein_id=NP_418567.1] [location=4371025..4372671]
MAAKDVKFGNDARVKMLRGVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGIDKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLI
>>lcl|NC_000913.3_cds_NP_418457.1_3949 [gene=malF] [protein=maltose transporter subunit] [protein_id=NP_418457.1] [location=complement(4243531..4245075)]
MDVIKKKHWWQSDALKWSVLGLLGLLVGYLVVLMYAQGEYLFAITTLILSSAGLYIFANRKAYAWRYVYPGMAGMGLFVLFPLVCTIAIAFTNYSSTNQLTFERAQEVLLDRSWQAGKTYNFGLYPAGDEWQLALSDGETGKNYLSDAFKFGGEQKLQLKET
>>lcl|NC_000913.3_cds_NP_418645.2_4142 [gene=chpS] [protein=antitoxin of the ChpBS toxin-antitoxin system] [protein_id=NP_418645.2] [location=4448447..4448698]
MRITIKRWGNSAGMVIPNIVMKELNLQPGQSVEAQVSNNQLILTPISRRYSLDELLAQCDMNAAELSEQDVWGKSTPAGDEIW_
>>lcl|NC_000913.3_cds_NP_418502.1_3995 [gene=yjcO] [protein=Sel1 family TPR-like repeat protein] [protein_id=NP_418502.1] [location=complement(4296436..4297125)]
MKKIIALMLFLTFFAHANDSEPGSQYLKAAEAGDRRAQYFLADSWFSSGDLSKAEYWAQKAADSGDADACALLAQIKITNPVSLDYPQAKVLAEKAAQAGSKEGEVTLAHILVNTQAGKPDYPKAISLLENASEDLENDSAVDAQMLLGLIYANGVGIKADI
>>lcl|NC_000913.3_cds_NP_418685.1_4180 [gene=idnR] [protein=transcriptional repressor, 5-gluconate-binding] [protein_id=NP_418685.1] [location=complement(4490141.
MRNHRISLQDIATLAGVTKMTVSRYIRSPKKVAKETGERIAKIMEEINYIPNRAPGMLLNAQSYTLGILIPSFQNQLFADILAGIESVTSEHNYQTLIANYNYDRDSEEESVINLLSYNIDGIILSEKYHTIRTVKFLRSATIPVVELMDVQGERLDMEVGE
>>lcl|NC_000913.3_cds_NP_418803.1_4302 [gene=lplA] [protein=lipoate-protein ligase A] [protein_id=NP_418803.1] [location=complement(4623101..4624117)]
MSTLRLLISDSYDPWFNLAVEECIFRQMPATQRVLFLWRNADTVVIGRAQNPWKECNTRRMEEDNVRLARRSSGGGAVFHDLGNTCFTFMAGKPEYDKTISTSIVLNALNALGVSAEASGRNDLVVKTVEGDRKVSGSAYRETKDRGFHHGTLLLNADLSRI
>>lcl|NC_000913.3_cds_NP_418621.5_4119 [gene=rpsF] [protein=30S ribosomal subunit protein S6] [protein_id=NP_418621.5] [location=4425118..4425513]
MRHYEIVFMVHPDQSEQVPGMIERYTAAITGAEGKIHRLEDWGRRQLAYPINKLHKAHYVLMNVEAPQEVIDELETTFRFNDAVIRSMVMRTKHAVTEASPMVKAKDERRERRDDFANETADDAEAGDSEE_
>>lcl|NC_000913.3_cds_NP_418525.1_4019 [gene=phnG] [protein=ribophosphonate triphosphate synthase subunit] [protein_id=NP_418525.1] [location=complement(4321244..
MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLVQIQARMGGTGERFFAGDATLTRAAVRLTDGTLGYSWVQGRDKQHAERCALIDALMQQSRHFQNLSETLIAPLDADRMARIAARQAEVNASRVDFFTMVRGDNA_
>>lcl|NC_000913.3_cds_NP_418553.1_4048 [gene=lysU] [protein=lysine tRNA synthetase, inducible] [protein_id=NP_418553.1] [location=complement(4353200..4354717)]
MSEQETRGANEAIDFNDELRNRREKLAALRQQGVAFPNDFRRDHTSDQLHEEFDAKDNQELESLNIEVSVAGRMMTRRIMGKASFVTLQDVGGRIQLYVARDSLPEGVYNDQFKKWDLGDIIGARGTLFKTQTGELSIHCTELRLLTKALRPLPDKFHGLQD
>>lcl|NC_000913.3_cds_NP_418522.1_4016 [gene=phnJ] [protein=carbon-phosphorus lyase, SAM-dependent] [protein_id=NP_418522.1] [location=complement(4318761..4319606
MANLSGYNFAYLDEQTKRMIRRAILKAVAIPGYQVPFGGREMPMPYGWGTGGIQLTASVIGESDVLKVIDQGADDTTNAVSIRNFFKRVTGVNTTERTDDATVIQTRHRIPETPLTEDQIIIFQVPIPEPLRFIEPRETETRTMHALEEYGVMQVKLYEDIA
>>lcl|NC_000913.3_cds_NP_418794.1_4293 [gene=yjjU] [protein=putative patatin-like family phospholipase] [protein_id=NP_418794.1] [location=4612411..4613484]
VGQRIPVTLGNIAPLSLRPFQPGRIALVCEGGGQRGIFTAGVLDEFMRAQFNPFDLYLGTSAGAQNLSAFICNQPGYARKVIMRYTTKREFFDPLRFVRGGNLIDLDWLVEATASQMPLQMDTAARLFDSGKSFYMCACRQDDYAPNYFLPTKQNWLDVIRA
>>lcl|NC_000913.3_cds_NP_418640.1_4138 [gene=msrA] [protein=methionine sulfoxide reductase A] [protein_id=NP_418640.1] [location=complement(4441538..4442176)]
MSLFDKKHLVSPADALPGRNTPMPVATLHAVNGHSMTNVPDGMEIAIFAMGCFWGVERLFWQLPGVYSTAAGYTGGYTPNPTYREVCSGDTGHAEAVRIVYDPSVISYEQLLQVFWENHDPAQGMRQGNDHGTQYRSAIYPLTPEQDAAARASLERFQAAMI
>>lcl|NC_000913.3_cds_NP_418756.1_4255 [gene=yjjN] [protein=zinc-type alcohol dehydrogenase-like protein] [protein_id=NP_418756.1] [location=complement(4565966..4
```

Figure 2. The output file (“output_protein.fasta”) from ECgnfinder_mc.py

IV. Discussion

I also calculated the accuracy of my prediction and the accuracy is about 91 percent which I print it on the screen when running the program. Therefore, the accuracy is quite high that we use markov chain model. However, the accuracy is high might because we use all E. coli genes to test the program which means we do not need to consider that it might have non-coding genes inside the test data.