COMP 561 Assignment 3

Jingyuan Wang 260860682

November 18, 2019

Question 1

a. Implemented in **predict.py**, the required fields are calculated and stored in **config.txt**.
Average length of intergenic region: 1030.5586107091171
Average length of genic region: 975.9979188345474
Nucleotide frequency table for intergenic: {’A’: 567669, ’C’: 519985, ’G’: 482119, ’T’: 566485}
Codon frequency table for genic regions: {’ATG’: 16262, ’AAT’: 11835, ’GCG’: 18997, ’TTC’: 8546, ’AAA’: 22315, ’TTA’: 12031, ’GTA’: 6898, ’CAT’: 7950, ’TCT’: 6951, ’AAG’: 8506, ’ATT’: 19281, ’AAC’: 12606, ’TTT’: 15960, ’TTG’: 14132, ’TCG’: 5769, ’TCC’: 3684, ’ATC’: 15707, ’GGT’: 16772, ’TGG’: 8207, ’GAC’: 9008, ’CAA’: 20682, ’GCC’: 13867, ’GTC’: 9024, ’CCC’: 3654, ’AGT’: 7212, ’GGC’: 15684, ’GCT’: 13111, ’GAA’: 24383, ’CGC’: 10996, ’GAG’: 15143, ’CTT’: 7950, ’GTG’: 17793, ’CTA’: 5502, ’ACC’: 12832, ’CAG’: 11506, ’CCA’: 7976, ’CTG’: 18147, ’GAT’: 23182, ’GCA’: 12053, ’ACG’: 7032, ’CTC’: 9067, ’GTT’: 10314, ’CCT’: 7009, ’TGT’: 3587, ’CAC’: 6628, ’CGT’: 12610, ’AGC’: 8757, ’ACA’: 4717, ’ACT’: 8097, ’GGG’: 5382, ’TAC’: 8865, ’TAT’: 9771, ’CCG’: 6706, ’CGA’: 3279, ’TCA’: 6547, ’CGG’: 1789, ’ATA’: 2239, ’TAA’: 1244, ’GGA’: 4649, ’TGC’: 2613, ’AGA’: 1600, ’AGG’: 570, ’TGA’: 407, ’TAG’: 355} (Stop codon is detected, however, is removed when calculating probabilities in later procedure.)

b. See file in **predict.py**. Usage is
python predict.py Vibrio_vulnificus.ASM74310v1.dna.nonchromosomal.fa config.txt
, where the first argument is fasta file to find gene and the second argument is config file which stores pre-computed parameters from question 1.a.

c. See gene prediction in **result.gff3**.

d. My prediction is compared with the given annotation of *Vibrio vulnificus* with an accuracy of 40.2974%. See implementation of this question in **evaluate.py** (Usage: *python evaluate.py*). See fraction of annotated genes to be reported in comparison to my prediction in **fraction_anno.gff3** while my predicted gene annotations compared to given annotation is in **my_fraction_anno.gff3**. Those gene fractions were computed with program, yet I will give an example of each category. Only sequence names and gene indices are shown here and another script was used to attain the real sequence with nucleotides for further evaluation.

- Annotated gene perfectly match predicted gene:
  contig_1 18767 19291

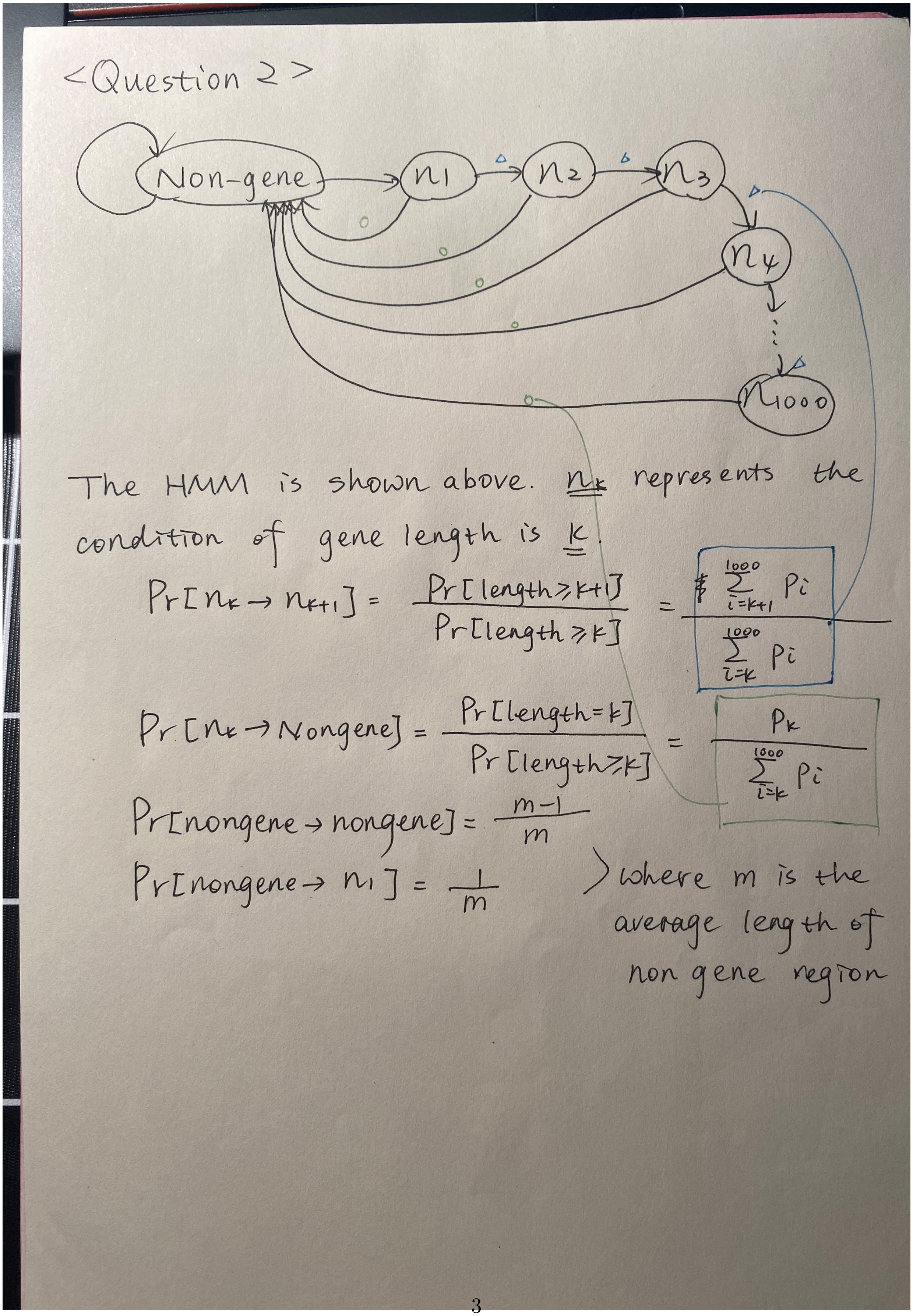
- Annotated gene matches only start of predicted gene:
  None

- Annotated gene matches only end of predicted gene:
  contig_1 6004 7083————predicted gene: contig_1 6634 7083
- Annotated gene match neither end of predicted gene:
  contig_1 1819 2256
- Predicted gene perfectly match annotated gene:
  contig_1 18767 19291
- Predicted gene matches only start of annotated gene:
  None
- Predicted gene matches only end of annotated gene:
  contig_1 6634 7083————contig_1 6604 7083
- Predicted gene match neither end of annotated gene:
  contig_1 5536 5712

e. It is shown that if the length of an annotated gene is too short, it is likely to be missed by the predictor, such as the following fragment from **contig_1 19327 19497**:
ATGGCCGTACAACAAAACCGTAAGACACGTTCTAAGCGTGGCATGCGTCGTTCACACGATGCG
CTAACTACAGCTGCACTATCTGTAGACGCGACTTCAGGTGAAACTCACCTACGTCACAACGTA
ACCGCTGAAGGTTACTACCGTGGCAAAAAGGTTATCAACAAGTAA
In our HMM model, an 'ATG' has a probability of approximately 0.001 to be marked as a start codon , whereas the probability of marking those as intergenic region is about 0.015. So it would need more reasonable codons to elevate the probability of current sequence as gene more than intergenic region.

On the other hand, my model would predict some extremely short sequences. e.g. **config_1 5536 5712**:
ATGCGTGCTGTGATAGTTATCGAAACTAATATCAAAGCCAGCGAAATCTTTTTGGTGCTCGATG
CTCACCGCAGCAATCATCTCTTCTGGAGACATACCCATCTGTTGCGCTTTAAGCATAATTGGCG
TGCCGTGAGCGTCGTCAGCACAGATGAAGTTTACAATGTTGCCGCGTAG
When using Viterbi algorithm, the average probability for a common codon is about 0.058, however, the probability for those three nucleotides to be predicted as intergenic region is $0.25^3 = 0.0156$. Consequently, when a start codon and end codon is detected within a small region, our model is very likely to predict them as genetic region. Moreover, as *Vibrio cholerae* has an average length of 1000 in genetic region, the short fragments we found are biologically not gene, although it resembles one.

Question 2


Non-gene
n_1
n_2
n_3
n_4
⋮
n_{1000}

<Question 2>

The HMM is shown above. $\underline{n_k}$ represents the condition of gene length is $\underline{k}$.

$$\Pr[n_k \rightarrow n_{k+1}] = \frac{\Pr[\text{length} \geq k+1]}{\Pr[\text{length} \geq k]} = \boxed{\frac{\cancel{\$}\sum_{i=k+1}^{1000} P_i}{\sum_{i=k}^{1000} P_i}}$$

$$\Pr[n_k \rightarrow \text{Nongene}] = \frac{\Pr[\text{length} = k]}{\Pr[\text{length} \geq k]} = \boxed{\frac{P_k}{\sum_{i=k}^{1000} P_i}}$$

$$\Pr[\text{nongene} \rightarrow \text{nongene}] = \frac{m-1}{m}$$

$$\Pr[\text{nongene} \rightarrow n_1] = \frac{1}{m}$$

} where m is the average length of non gene region