Assignment-2

2018101075

1. Simulation is done in q1.py.
The output is $P(N \geq 240) = 0.7721$

Using normal approximation to binomial distribution

$p = 0.25$

$q = 1 - p = 0.75$

$n = 1000$

$\mu = np = 250$

By central limit theorem

$$\sigma = \sqrt{npq} = 13.69$$

$$z = \frac{240 - \mu}{\sigma} = -0.7305$$

Using the z scores, we get

$$P(x \geq 240) = 0.5 (1 - (-0.5552))$$
$$= 0.77776$$

On comparison, we see that both the values almost match. (with a small error). If we increase the number of simulations, we see that the obtained value becomes closed to the normal approximation value.

2. $p = 0.3$

$q = 1 - p = 0.7$

$n = 10$

$P(X = n) = {}^{n}C_n \, p^n q^{n-n}$

$P(X = 0) = {}^{10}C_0 (0.3)^0 (0.7)^{10} = 0.0282$

$P(X > 2) = {}^{10}C_x (0.3)^2 (0.7)^8 = 0.2335$

$E(X) = np = 10 \times 0.3 = 3$

$Var(X) = npq = 10 \times 0.3 \times 0.7 = 2.1$

3. K-mers ($K \geq 2$) are used to identify regions having aberrant base compositions that indicate genome segments acquired by lateral transfer.

Based on K-mer analysis, the different parametric methods at the gene level are :-

- Codon Usage Bias - i.e. unequal usage of synonymous codons.
- Amino acid Usage Bias :- deviation in the frequency of usage of individual amino acids over the average usage of all 20 AA.
- GC content at codon positions - frequency of occurrence of G & C at 3 codon positions, GC1, GC2 & GC3 for the set of genes compared to the whole genome and set of genes.

The frequencies of K-tuples have a number of applications:

- For eukaryotes, gene regions, in general, have a different base composition than non-genic regions. Different gene classes have different codon usage frequencies, eg. highly expressed genes, that differ from organism to organism - useful in identifying horizontally transferred genes.
- Sometimes the observed frequencies of K-words can be used to make inferences about DNA sequences.
- K-tuple ($K > 3$) frequencies can also assist in predicting whether an unannotated sequence is coding or non-coding.
- Predict gene expression: using $K = 3$, compute CAI within ORFs to identify highly expressed genes. (CAI ~ 1)
- K-tuple frequencies and other content-based measures such as the presence of particular signals are among the [redacted] statistical properties employed by computational gene finding tools.

K-mer distributions are well-preserved among related strains/species. As a result, bacterial genomes can be clustered into natural groups according to K-mer distribution similarities.

4. To get the plot, run q4.py. 4.png contains the image of the output
The diagonal lines of 'x' from top-left to bottom-right represents conserved region.

5. Run q5.py to get the plot. 5.png contains the image of the output with diagonals coloured. Repeat regions can be identified by the top left - bottom right diagonal
The main sequence is in yellow color, while sequences with 4 elements repeat are highlighted using purple and those with 5 elements are highlighted with other colours. (Sequences with lower element repeats are ignored / not highlighted).

6. Run q6.py to get the plot. 6.png contains the image of the output with diagonals coloured.
Complementary Repeat regions can be identified by the top left - bottom right diagonals
The longest self-complementary region found here is marked here in yellow (length 6) and is 'UUGCAUGCCCA'. Other smaller regions are marked in purple.