Assignment-3 Instructions

1. Last date of submission: **12-April-2019**

2. You have to submit **program files** (.m or .py file) **only**.

3. **Display** all the results.

4. We evaluates **only working code**. **Syntax errors** leads to zero marking .

5. You have to **clearly comment** on the program that, code is for which question and what are the results it will generates.

6. Upload all the codes as a single zip file with your **name and roll number** in the file name (e.g. Rahul_16126.zip)

7. Late submission will leads to **complete loss** of your marks.

Assignment-3: There are Three questions that carries overall 10 M.

Note: Use the DNA sequences given below to do all the following problems.

S0 =
CGGCCTGAAGCGACGTCGTATCATATCAATCGCATGTCATCGCCGTCTACGCCCCGGAGACTAAACCCTGCCGCATGATAATGTGGTCTACTGAGTTCTTCATGG
GGCAGGGGATCATGAATCGTGCAAGACCCAAGCCCCTACCAAAGAGACCACGAGGTCATTAGTCTTCCTAGGCGACTAGTTCTGTCGCGCTCTCACCATTTCTTC
TCATGGGGAACTCAGAACTGGATGAATGTCCCTTAGACCCTGTTTTCCTCGCGTGAAAAAGTACCTTTTAGAGCATTCAAATATGTCGACCGAAGAACCTGTAGT
TAAATCCGTCGCATTAACTCTTAGAGGGCCGGAGCTAAGACCAAGTCTATCACGCGCGCTCAAACATGAGGGAGATTGGTCCATTTGTGGGAGATTAGCCAAGC
ATCATGGAACTACCTCTTTCCATACAATTTCGGCCTTGCCATATTCCATTTAATGAAAGCTACGCCTCGAGCCGTTAAGCCCGTCAATAGAACTGGTTACCTAAGG
CCAGTACCAACGGAATGGCTGGAGGTCGCGCCACGAATATGGTGCCTTTTTCCTGTAGCTCGTGTCGGCCGAAGA

S1 =
AGGCGTCAAGTGTCGGCGGGGCATATTAATGGCGTGTTGCTAAGCTGGACAGTCAAAGTGCCCAACTCAGCTGCGCCGCAGCGTATTCCGACGGCTTCTCCATG
AGGGAAAAGATCGCAAAACGGGTAAGTTTTAAATTTGTAATAATAAGACGATTTGCCAACTGGTCCCGAAAGGGGAATGAGTTTGCCACAGACCCCCTGTCTGT
TCGTCCCAAAAATCAGGGTCCAGATGAGTTGTACCTGAGGGTCCATTTCTTCTTTTAGCTGATTGATTCCCGGATGACCCCTACGTGTCGCTCAGAAAGACAGTA
CGTCGACGCGTCACCTTAACATAGGGGTTGCCCAGGCCCGGCCCTAACCGAATTGGCATCCACAAACATAGGAAAGATTGATCCAATAAAAAGAAATCAGCCGC
GTACCATTATGTTAGCTATATCTGGGCATTGGCGTCCGTGCCGTCCTTTGACTAATAACGGTTACTCCCCAAGCAGTTATACCGGTGGGCAAAACTGGTCGATGG
ACTCGGCGGTGAATAGTCCGATCGGCGCACACGCCATGAGCAGGATGCATTCTTCCTGTAACCTGTGACAACTGCGGG

Q1. Write the code in MATLAB/PYTHON to

(a) Compute frequencies of various pairs of bases appearing at same site in two DNA sequences. Construct 4x4 array sequence table, with bases in seq1 (So) along top, seq2 (S1) along side, **in order A,G,C,T**.

(b) display the fraction of sites with each base in S0 from sequence table in (a).

(c) display the fraction of sites with each base in S1.

(d) display the conditional probability from the sequence table (a) **(4M)**

Q2. Write a MATLAB/PYTHON code to compute the following distances to 10 decimal digits for the above sequences (a) Jukes-Cantor (b) Kimura 2-parameter (c) Paralinear. Which of these is likely to be a better estimate of the number of substitutions per site that actually occurred? Explain**.** **(3M)**

Q3. Construct a makeup 4×4 Markov matrix M with all positive entries and an initial p_0. Make sure the diagonal entries of M are the largest. (a) Use a computer to observe that, after many time steps, $p_t = M^t p_0$ appears to approach some equilibrium. Estimate the equilibrium vector as accurately as you can. Do you get an eigenvector of M with eigenvalue 1? (b) Use a computer to compute the eigenvectors and eigenvalues of M. Is 1 an eigenvalue? Is your estimate of the equilibrium close to its eigenvector? (C) Are your computations in part (b) consistent with the two theorems about Markov matrices that is discussed in the class? (**3M)**