# Assignment on Sequences

**Deadline: 27th Nov**

1. For a given sequence (File: *E. coli* K-12 MG 1655, GneBank id: U00096), find the frequency of occurrence of each of the nucleotides A, T, G, C’s, dinucleotides, and trinucleotides. For example for sequence ATGCCG dinucleotides would be AT, TG, GC etc and trinucleotides would be ATG, TGC, GCC etc.
2. Plot the density of nucleotides in a sequence (File: *E. coli* K-12 MG 1655). Graphically display n-mer (for n=2 and 3) in the given sequence. (You may use Excel sheet for obtaining the plots).
3. Consider the output generated by Program 1.
   1. What probability would you assign to an A being followed by a G? Under what assumptions does this seem appropriate?
   2. Given that there is an A at a given position along the E. coli genome, what probability would you assign to it being followed by a G, and why?
   3. Does the E. coli composition data suggest that the event we observe a G at one site is independent of the previous two bases? Explain fully with appropriate data.
4. For the given sequence (File: *E. coli* K-12 MG 1655), count number of purines (A,G) in each block (block size of 100,1000, and 10,000bp).
   1. Calculate the mean and standard deviation of the proportion of purines per block, and draw histograms of these numbers.
   2. Compare the results of (a) across the different block sizes and comment.
   3. For each block size calculate fraction of counts within 1, 2 and 3 standard deviations of the mean.