Homework 4

Due Oct 30 before class starts.

1. A restriction enzyme cuts DNA at specific locations. Identify the number of cut sites in the genome of Bacillus subtilis of the EcoRI (GAATTC) motif. Your program should simply print out

```
Genome file: [MY FILE]
Number of EcoRI (GAATTC) cut sites: [Number of sites]
...
```

Here is a list of RE sites - you will need to re-write some of these to convert to a regular expression.

```
EcoRI   = "GAATTC"
Bsu15I  = "ATCGAT"
Bsu36I  = "CCTNAGG"
BsuRI   = "GGCC"
EcoRII  = "CCWGG"
```

The abbreviation for DNA ambiguity patterns can be found on this page.

Use the *B. subtilis* genome here and you can find a direct link to the strain 168 FTP folder of the assembly. You want the ".fna" file.

You can also use the E. coli K-12 genome from NCBI or what is available here.

The goal of this is to write generic code so you are welcome to run this on any genome really. There are many sequenced strains, it would be interesting to compare if the number of cut sites (or their size) varied among strains.

You can use this script as a starting point for reading a FASTA file

For the advanced programmers. Think about (or try) to make your program handle a folder of sequence files to read and provide a report.

2. Write a program to find which proteins have a Nuclear Localization Signal in yeast. The following paper Fries et al. Figure 8 demonstrates a sequence pattern. There is one part that is ambiguous (X)n where n varies some in the species. Feel free to write this targeted for the *Saccharomyces cerevisiae* search or just make it some arbitrarily long length based on all the data. Here is the sequence download site for Saccharomyces and here are the proteins in fasta format you can download.

- You can spot check how you tool does by looking at the Gene Ontology (GO) data for Saccharomyces. The 4th column indicates the ontology - a C there means 'cellular localization' so the rows that have a 'C' in the 4th column give you some indication of where the protein localizes to (if it is known). Typically a nuclear localization signal would indicate the protein localizes to the nucleus. So you could see how many of your hits have this prediction. Or you could test how many DON'T have this prediction since this signal sequence may not have been used to classify localization in GO

3. Task: You have been given a set of assembled RNA-Seq sequences that were assembled with Trinity. You would like to investigate how many of them have poly-A tails. Look for sequences that have AAAAAA or AATAAA, find the 3' most of these (e.g. the last one in the sequence) this paper talks about finding motifs related to poly-A positioning.

- Write a script to read in the data, and count which sequences have polyA sites – you can use this script as a starting point for reading a FASTA file

- Generate a distribution of polyA lengths (distance between the motif you found and the end of the contig). You will want to review how we capture a regular expression search match and then get the start or end of that match. E.g.

```
 5'<------------------------------------> 3'
                  AATAAAGAACAAAGTA
                     100       110
match ends at 100
sequence is 110 bp long
polyA tail is 10 bp long
```

The matching or sequence composition of is not always perfect for polyA (how to decide what is the LAST motif match is the problem). But give this a try - see if you can generate some summary statistics from this data.

- Compute summary statistics for these lengths (mean, median)

- Plot histogram of this distribution - using R. Here is a simple R script to plot histogram you can run like this - just make sure you file is called 'polyA_lengths.dat' or change the code in the R script.

```
$ R --no-save < histogram.R
```

You may want to revisit the regular expression lectures and see this page on Regular Expressions from Python.