Introductory Python Programming for Genomics (BIOS274)

Problem Set #1 (due Monday (Nov 25) before 5pm, submit your script and output via canvas)

Given a list of restriction enzyme recognition sites and the cutting site offsets, create components of a typical restriction map.

The sequences are contained in the fasta file: `rosalind_dna.fsa`

The two dictionaries with restriction enzyme information are:

```
enzyme_sites = {'EcoRI': 'GAATTC', 'HindIII': 'AAGCTT',
                'BamHI': 'GGATCC', 'HpaI': 'GTTAAC',
                'HaeIII': 'GGCC'}
cutsite_offset = {'EcoRI': 1, 'HindIII': 1, 'BamHI': 1,
                  'HpaI': 3, 'HaeIII': 2}
```

The goal is to see what you can do, be challenged. There are three parts to this problem set. Start with part 1, and then of part 2. Part 3 is optional, provided if you feel inspired. We'll go over the solutions in class.

An important part of this problem is to output a clear readable format, below is a suggestion. Add appropriate commands to your scripts and use informative variable names.

Please submit your script and output files via canvas

1. Login to canvas.stanford.edu, and select the BIOS274.
2. Go to the Assignments tab and submit your work to Problem Set 1.
3. Upload your .ipynb file.
4. Create the html file, in Jupyter select the 'Download as' option in the File tab. Select the 'html (.html)' option. A .html file will be downloaded.
5. Upload an the .html version of your script.
6. Upload your output files as .txt files.

Part #1:

The information you need to complete this assignment are provided in the enzyme_sites and cut_offset dictionaries above. The DNA sequences to be processed are contained within the fasta file, `rosalind_dna.fsa`, available on canvas in the Files>Problem Sets>Problem Set 1 folder. Include mention of sequence(s) that aren't cut by any of the enzymes. For those that do have recognition sites, list the cut site(s) in 1-based ordering (the first base of the DNA sequence is position 1). Below is a suggested format for the sequences Rosalind_6820 and Rosalind_3684.

Sequence: Rosalind_6820 (cut sites)
HpaI 118
HaeIII 596

Sequence: Rosalind_3684 (cut sites)
HaeIII 106, 121, 263, 408, 800, 916

Part #2.
Extend the output above to include the length of the DNA fragments produced per enzyme.

Sequence: Rosalind_6820 (fragment sizes)
HpaI 118, 877
HaeIII 596, 399

Sequence: Rosalind_3684 (fragment sizes)
HaeIII 106, 15, 142, 145, 392, 116, 48

Part #3. [optional]
Print the DNA sequences, and its complement, with the cut sites annotated. A possible presentation is below.

```
Sequence: Rosalind_2711

     1  CTACCGAGAGGTGCCGTCAAATTCTGCCTTTAACCCCCACATGTAGCTCAGTAACTGAGC
        GATGGCTCTCCACGGCAGTTTAAGACGGAAATTGGGGGTGTACATCGAGTCATTGACTCG
                HaeIII
                  |
    61  GGCTTGACGGCCCAGGTGCAGAGACGTACGATGCGTGAGCCTGCACAATAACCCACCATT
        CCGAACTGCCGGGTCCACGTCTCTGCATGCTACGCACTCGGACGTGTTATTGGGTGGTAA
                    HindIII
                     |
   121  TAGACCATTCAAAAGCTTCCAGACAGTCTAGCTCGAAGAAATTTTTACTCGCTACTAGAC
        ATCTGGTAAGTTTTCGAAGGTCTGTCAGATCGAGCTTCTTTAAAAATGAGCGATGATCTG
                                                      HaeIII
                                                        |
   181  CCGGGTTTCGGAACAAATTGACCAAGAGGACAGTTGTCCCAGCGGCCTCGCCGACGTCTT
        GGCCCAAAGCCTTGTTTAACTGGTTCTCCTGTCAACAGGGTCGCCGGAGCGGCTGCAGAA


   241  AGTGCATCTAGTTCTTGAGTATACTTACTATACTTTACGCCGCTATCTAAAACCCACCAC
        TCACGTAGATCAAGAACTCATATGAATGATATGAAATGCGGCGATAGATTTTGGGTGGTG
```