CS284A: Algorithms for Molecular Biology Assignment #1

The yeast species *Saccharomyces cerevisiae* has been used in baking and brewing for thousands of years. It is one of most intensively studied eukaryotic model organisms in molecular and cell biology. *Saccharomyces cerevisiae* was the first eukaryotic genome that was completely sequenced. The genome is composed of about 13 million base pairs and ∼6,000 genes.

In this assignment, we will study the genome of *Saccharomyces cerevisiae*, in particular to identify regulatory motifs in promoter regions of the *S. cerevisiae* genes.

1. Download the following two files from:
   `http://www.ics.uci.edu/ xhx/courses/CS284A/assignments/PS1`

   (a) Promoter sequences: `yeast_orfs_promoter.fa`

   (b) Gene set G: `yeast_cellcycle_gene_cluster2.txt`

   The first file (a) contains promoter sequences of all *S. cerevisiae* genes. Each sequence in the file begins with a single-line description, followed by lines of sequences data. The description line is distinguished from the sequence data by a ">" symbol in the first column. The format is commonly referred to as the FASTA format. The word following ">" symbol is the ID for the sequence. The ID is typically the name of the gene for the sequence. Sometimes a promoter sequence is shared by two genes. In this case, The ID is the names of two genes connected by the symbol "_". For instance, ">YAL043C_YAL042W" represents that the sequence is the promoter of both gene YAL043C and gene YAL042W.

   The second file (b) contains a list of genes. These genes have been previously identified to share a similar expression pattern across different stages of *S. cerevisiae* cell cycle.

2. Motif discovery using enumeration-based method. This consists of the following several steps:

   (a) Write a program to enumerate all 6-mer motifs.

   (b) Write a program to return the reverse complement of any 6-mer. To get the reverse complement of a DNA sequence, first reverse the sequence and then convert letters `A,C,G,T` to `T,G,C,A` respectively. For instance, the reverse complement of `TGACCT` is `AGGTCA`.

   (c) For each 6-mer, count the number of genes (in `yeast_orfs_promoter.fa`) whose promoters contain the 6-mer or its reverse complement. Note that for each gene we only determine whether the gene contains the motif (or its reverse complement) or not. Even if the gene contains multiple sites of the motif, it will be counted as only once. Extra credit: any biological reasons to include the instances of the reverse complement?

(d) Repeat the same exercise in (c) using only genes contained in the gene set G.

(e) For each 6-mer, define and calculate a p-value to quantify the significance of its over-representation in the gene set G, using the entire gene set as a control.

(f) Rank all 6-mers with their p-values in ascending order. Output top 10 6-mers with the smallest p-values.

3. Experiment with the MEME algorithm (implementing the EM-algorithm, `http://meme.sdsc.edu/meme/meme.html`). Write down motifs discovered by MEME for gene set G.

4. Experiment with AlignACE algorithm (implementing Gibbs sampling algorithm, `http://atlas.med.harvard.edu/cgi-bin/alignace.pl`). Write down motifs discovered by AlignACE for gene set G.

5. Compare motifs discovered from the enumeration method, MEME and AlignACE. Comment on the differences.