L519: Bioinformatics: Theory & Application

HW6 (Due: **Dec. 5** Midnight)

http://darwin.informatics.indiana.edu/col/courses/L519

----------**Section 1** ----------

For the homework#6, you will have three exercises. There is no Perl script.

1. **Motif Finding**

You have already experienced how to use several motif finding programs like MEME and Gibbs sampler. Let's do one more round of such an exercise. You will be provided a set of sequences which are known to contain a TFBS (Transcription Factor Binding Site). These sequences are from yeast (*Saccharomyces cerevisiae*).

1. Use MATCH and P-MATCH programs to search for the possible binding site of the given sequences. You have to register in order to use the public TRANSFAC at the Gene Regulation webpage (http://www.gene-regulation.com/index.html). Have you identified the transcription factor binding sites? Then, **what is the transcription factor**? (Submit your search research (html file), too) and **what is the consensus for this TFBS?** (Note: the middle part of this TFBS is not well conserved. Only the beginning and ending portion are well conserved and important for TF recognition)
2. Now, let's see whether MEME and Gibbs whether they can discovery the hidden motif (TFBS). Try MEME and Gibbs. If you were able to find the length of the TFBS from the previous step, you can use it. If not, try a length range of [10,20].
    a. **Is MEME able to identify the TFBS, correctly?**
        i. **If not, do you see any similar motif? Then, what it is?**
    b. **Is Gibbs able to identify the TFBS, correctly?**
        i. **If not, do you see any similar motif? Then, what it is?**

    [Hint: For Gibbs, you better limit the max_sites per sequence [-E]]
3. **Comment what you have learned from this exercise.**

2. **RNA structure prediction**

As you learned from the class, RNA molecules can also have 2D and 3D structures like proteins. Typical 2D structures of RNA molecules are stem, hairpin loop, bulge loop, interior loop, junction (multi-loop), and pseudoknot. Refer to the class slides for detail. There have been many attempts to computationally predict the 2D / 3D structures of RNA molecules. The two mostly widely used programs are MFOLD and Vienna RNA package.

In this exercise, use any available RNA structure prediction programs including MFOLD and Vienna RNA package. Compare the results and report them with structure prediction results.

Seq1.
GTTAATGTAGCTTATAATAAAGCAAAGCACTGAAAATGCTTAGATGGATTCAAAAA
TCCCATAAACA

Seq2.
GCTTACGACCATATCACGTTGAATGCACGC
CATCCCGTCCGATCTGGCAAGTTAAGCAAC
GTTGAGTCCAGTTAGTACTTGGATCGGAGA
CGGCCTGGGAATCCTGGATGTTGTAAGCT

3. **dnSNP**

In the lab session 11, the public SNP database, NCBI's dbSNP, was introduced. Go to the dbSNP website, and find out how many SNPs have been identified for the following genes so far?

A. psen1 (all available species)?
B. psen1 Human?
C. BRCA1 Human?
D. BRCA1 Human and only from intron region?