# BMI/CS 776 Spring 2007
# Homework #2

Prof. Colin Dewey

Due Friday, March 9th, 2007 by 11:59pm

The goal of this assignment is to become more familiar with the classic methods for discovering motifs within biological sequences.

To turn in your assignment, copy all relevant files to the directory:
`/u/medinfo/handin/bmi776/hw2/USERNAME`
where `USERNAME` is your account name for the BMI network. You must submit a file named `README` to this directory, which gives directions on how to compile (if necessary) and run your programs. For each question below, the `README` file should list the files relevant to that question (e.g., code, other files with written answers).

1. Write a program, `learnmotif`, that takes as input a set of DNA sequences and a width $W$, and learns an OOPS motif model for a motif of width $W$:

   ```
   learnmotif sequences.fasta width model_file positions_file
   ```

   where `sequences.fasta` is the input filename, `width` is the width of the motif model to learn, `model_file` is the name of a file to which you will output the learned motif model, and `positions_file` is the name of a file to which you will output the predicted location of the motif in each sequences. You may add other arguments as you see fit (e.g., maximum number of iterations, random number generator seed, etc.). Please document these extra arguments in your `README` file.

   The `positions_file` should simply contain a list of the best position for the motif in each sequence, one position per line. The `model_file` should contain a tab-delimited profile matrix, with the background frequencies in the first column.

   You may use either the EM algorithm (i.e., MEME OOPS), or the Gibbs sampling algorithm discussed in class for learning the motif model.

2. With `learnmotif`, discover the motif of width 14 hidden in sequences in the file: `http://www.biostat.wisc.edu/bmi776/hw/hw2_hidden_motif.fasta`.

3. Construct a *sequence logo* for the learned motif model from (2) by using the WebLogo Web form (`http://weblogo.berkeley.edu/`).

4. Search the JASPAR transcription factor binding profile database (`http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl`) using the profile matrix that you learned in (2). Which transcription factor binds to this motif?