Algorithms for Bioinformatics (Autumn 2015)

Exercise 2 (Tue 15.9., 10-12, B222)

If you cannot make it to the exercise session, please e-mail your solutions and the reason why you cannot attend to daniel.valenzuela@cs.helsinki.fi before the exercise session to get credit.

Some of the problems below are programming exercises on the Rosalind platform at `http://rosalind.info/problems/list-view/?location=bioinformatics-textbook-track`

1. Solve Rosalind problem BA2B: *Find a Median String*.
   Hint: Problem BA2H is a subproblem of Median String and you might try solving it first.

2. (a) Take a Rosalind input dataset for the Median String problem (with string length $n > 20$), increment the value of k, and use your Median String implementation to solve it. Keep incrementing k until the running time exceeds 5 minutes. For example, if the original input had $k = 6$, try $k = 7$, $k = 8$, and so on. At what value of k did the time exceed 5 minutes?

   (b) What is the number of character comparisons performed by your Median String program? Express the answer as a formula based on the parameters k, n and t. Verify the formula by modifying your program to compute and output this number.

   (c) How many character comparisons per second did your program do for the input in part (a) that took over 5 minutes. Based on this and the formula in part (b), estimate how large value of k your program could handle in one day? How large in one year?

3. Solve Rosalind problem BA2E: *Implement GreedyMotifSearch with Pseudocounts*
   Hint: Problem BA2C is a subproblem.

4. Design a dataset for Motif Finding for which GreedyMotifSearch (with pseudocounts) fails to find the best set of motifs.
   Hint: Force GreedyMotifSearch to make wrong choices in the early rounds.

5. Solve Rosalind problem BA2F: *Implement RandomizedMotifSearch*

6. Solve Rosalind problem BA2G: *Implement GibbsSampler*