# Algorithms for Bioinformatics (Autumn 2014)

## Course Exam 21.10.2014 — Solutions and grading

See the course material for more detailed solutions.

1. **Branch-and-bound and motif finding.** (12 points)

   Define the Motif Finding and Median String problems and explain why they are actually the same problem. Describe briefly the idea of the branch-and-bound solution for solving the problem.

   **Grading:**

   - For both problems:
     - +1 for input
     - +1 for output
     - +1 for score/distance
   - +3 for describing the connection between the score for Motif Finding and the distance for Median String.
   - +3 for explaining the main idea of branch-and-bound.

2. **Edit distance and approximate string matching.** (3+3+3+3 points)

   Compare the problem of computing the edit distance and the $k$-errors problem (i.e., approximate string matching with edit distance):

   (a) What is the difference in the definition of the problems?

   (b) What is the difference in their solutions using dynamic programming?

   (c) Solve the $k$-errors problem for pattern $P =$ `TGAA`, text $T =$ `GTAATGTA` and $k = 1$ by filling the dynamic programming matrix. List all approximate occurrences of $P$ in $T$.

   (d) Choose one of the approximate occurrences of $P$ in $T$ and compute the edit distance between $P$ and that occurrence by filling the dynamic programming matrix. Give also the corresponding alignment.

   **Grading:**

   (a)
   - +1 for string-string vs. string-substring comparison
   - +1 for no threshold vs. threshold
   - +1 for reporting distance vs. reporting positions

   (b)
   - +1 for difference in first row initialization
   - +1 for difference in reporting result
   - +1 for mentioning that the rest of the computation is the same

   (c) and (d)
   - +2 for correctly filled matrix
   - +1 for correct occurrences (TAA and TGTA) or alignment

3. **Reductions and sequencing.** (4+4+4 points)

   Suppose that a sequencing by hybridization experiment produces the following (multi)set $S = \{\texttt{AAA, AAG, AGA, AGT, GAT, TAA, TAG}\}$.

   (a) Use the Eulerian path approach to show that $S$ is *not* a 3-mer spectrum of a string.

   (b) Show that if one nucleotide in one of the strings in $S$ is substituted for another nucleotide, the new $S$ becomes a 3-mer spectrum of a string.

   (c) The new $S$ is the 3-mer spectrum of multiple strings. How many? Which strings?

   **Grading:**

   (a)
   - +2 for drawing the correct graph with 2-mers as vertices and 3-mers as edges
   - +2 for a condition that makes the graph non-Eulerian

   (b)
   - +2 for coming up with the correct substitution AAA → ATA
   - +1 for drawing the correct graph
   - +1 for explaining why the graph is now Eulerian

   (c)
   - +2 for each string (TAAGATAGT and TAGATAAGT)

4. **Your choice.** (12 points)

   Choose one of the (non-trivial) problems studied during the course (in study groups, lectures, or/and exercises) not related to the assignments above. Define the problem (input, output), explain how the problem is motivated by molecular biology, and describe an algorithm for the problem by either simulating an example or by giving its pseudocode.

   **Grading**:

   - 4p – Correct definition
   - 4p – Correct motivation
   - 4p – Correct simulation of pseudo-code