582670 Algorithms for Bioinformatics, 4 cr — Exam 26.11.2013 — Solutions/grading

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.

   See course material for answers.

   Grading:

   - +6 points from defining $Score()$ and $TotalDistance()$.
   - +3 points for describing the connection between $Score()$ and $TotalDistance()$.
   - +3 points for explaining the main idea of branch-and-bound.
   - Some points given for sketches of the problem definitions and/or showing an example of a problem instance.

2. **Dynamic programming and sequence alignment.** (12 points)

   Give a small (but non-trivial) example of directed acyclic graph (DAG) with edge weights and lightest path computation on it. How is this computation related to sequence alignment?

   Solution:

   See the lecture slides for an example DAG and lightest path computation.

   The edit distance matrix can be thought of as a DAG and then the edit distance computation corresponds to finding a lightest path on that DAG.

   Grading:

   - +4 points for drawing a DAG
   - +4 points for lightest path computations on the DAG
   - +4 points for relating lightest path computation on the DAG to sequence alignment

3. **Greedy algorithms and genome rearrangements.** (12 points)

   Simulate the improved breakpoint reversal sort 4-approximation algorithm on the permutation

   ```
   3 9 8 2 1 5 6 7 4.
   ```

   Based on the properties of this problem instance, estimate how many more reversals does the algorithm make compared to the optimal solution?

   Solution:

Recall increasing and decreasing strips. One element strips are defined as decreasing, except for the special case of elements 1 and n: if they are located at their correct positions, then there is no breakpoint before / after, respectively, and they can be seen as increasing strips.

The improved breakpoint reversal sorting finds a reversal distance of 5 reversals:

```
3 (9 8 2 1 5 6 7 4)  6bp
3 4 (7 6 5 1 2) 8 9  4bp
(3 4 2 1) 5 6 7 8 9  3bp
1 2 (4 3) 5 6 7 8 9  2bp
1 2 3 4 5 6 7 8 9    0bp
```

Originally there were 6 breakpoints and, since each reversal can remove at most two breakpoints, we have $OPT \geq 3$ reversals. In this problem instance, the improved breakpoint reversal sort used four reversals and therefore it used at most one reversal more than the optimal solution.

Grading:

- +8 points for correct simulation of the improved breakpoint reversal sort.
- +4 points from relating the answer to the lower bound.
- Some points given for partially correct simulations.

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

Sequencing by hybridization experiment produces the following (multi)set $S =$ {AGC, AGG, CGA, GCA, GCG, GGC, GGG}.

(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?

(a) Draw the graph with $\ell - 1$-mers as nodes, and add an edge between two nodes is the corresponding ℓ-mer is in the spectrum (here $\ell = 3$). There is an Eulerian path in the graph if for one vertex $in(v) = out(v) - 1$, for one vertex $in(w) = out(w) + 1$ and for the rest of the vertices $in(u) = out(u)$. This is not the case for the given graph and thus there is no Eulerian path and the spectrum does not correspond to any one string.

Grading:

- +2 points for drawing the correct graph
- +2 points for the condition for having an Eulerian path

(b) GGG → GAG

Grading:

- +2 points for coming up with the correct substitution
- +2 points for drawing the correct graph

(c) Two strings: AGCGAGGCA and AGGCGAGCA

Grading:

- +2 points for each string

5. **Your choice.**

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

Solution:

Grading:

- +4 points for correct definition
- +4 points for correct simulation or pseudocode.
- +4 points for correct motivation.
- Some points reduced when the description was not clear enough, or mistakes in simulation.