582670 Algorithms for Bioinformatics, 4 cr — Exam 19.10.2011 — Solutions/grading

1,2 graded by Niko and 3,4 by Veli

1. **Branch-and-bound and motif finding.**

   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. **Greedy algorithms and genome rearrangements.**

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

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

   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 4 reversals:

   ```
   5 4 3 1 2 8 7 (9 6)      6bp
   5 4 3 1 2 (8 7 6) 9      4bp
   5 4 3 (1 2) 6 7 8 9      3bp
   (5 4 3 2 1) 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 at most one more reversal than any optimal solution could achieve.

   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.

3. **Reductions and sequencing.**

   Simulate the 4-approximation algorithm for the shortest common superstring problem on the set $S = \{\texttt{GAAT,ACTA,ATGA,CTAC}\}$. (Recall that this is the algorithm using prefix graph and reduction to minimum weight cycle cover.) Visualize also the minimum weight perfect matching corresponding to the minimum weight cycle cover.

   Solution. This was the same input as in the exercises, but each string complemented.

   Grading:

   - +8 points for correct prefix graph, cycle cover, and solution.
   - +4 points when showing also the perfect matching.
   - Some points reduced when mistakes e.g. in prefix graph affecting the solution.

4. **Your choice.**

   Choose one of the (non-trivial) problems studied during the course (in study groups, lectures, or/and exercises) not related to the three 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. Most popular were partial digest, small parsinomy, and UPGMA.

   Grading:

   - +8 points from correct simulation or pseudocode.
   - +4 points from correct motivation.
   - Some points reduced when the description was not clear enough, or mistakes in simulation.

Grading: 12+12+12+12=48