Introduction to Bioinformatics (autumn 2005)

Excersice 2

Group	time	place
Riikka Kaven	Tuesday 11.10 at 12.15–14.00	BK106

1. Simulate the greedy breakpoint-based reversal sorting algorithm (`ImprovedBreakPointReversalSort` on page 135 in J&P), on input π = 10 8 6 4 2 9 7 5 3 1. How much better can the optimal algorithm be on this input?

2. Show that the not improved breakpoint-based reversal sorting algorithm (`BreakPointReversalSort` on page 133) terminates on any input and is at least as good as `ImprovedBreakPointReversalSort`. (Hint: it suffices to show that in any permutation that only has increasing strips, the best reversal produces a decreasing strip.)

3. You are given the following DNA sequences: `DNA` = {`aacggt,ccgtaa,taaggt`}. The task is to find the optimal 4-mer motif (see Sect. 4.5 page 93- in J & P for definitions).

   a) Draw the search tree representing the 4-mer alignments of these three sequences (all possible 4-mer starting point configurations). Label all nodes with their starting positions. In addition, label the interior nodes of the search tree with their optimistic maximal score values, and the leaf nodes with their consensus scores. What is the maximum consensus score? What is the consensus motif string?

   b) Mark the subtrees that would be pruned in a branch-and-bound strategy.

   c) What is $TotalDistance(v, DNA)$, where v is the consensus motif?

4-5. Before the DNA sequencing technologies were developed, biologists used *Restriction Mapping* to reveal genetic markers: Rectriction enzymes were used to cut the DNA into fragments, whose lengths could be measured by gel electrophoresis. However, knowing only a few fragment lengths gives no clue about where the actual cutting took place. To get more information, one can measure the fragment lengths between *all* pairs of different restriction sites. Such measurement is called *partial digest*. The goal is to reveal the exact positions of restriction sites based on the full set of fragment lengths. More formally, in *partial digest problem* one is asked to find a set $X = \{x_1, x_2, \ldots, x_n\}$ of points on a line such that the multiset $\Delta X = \{x_j - x_i : 1 \leq i < j \leq n\}$ corresponds to the measured fragment lengths.

   a) Try to reveal X corresponding to $\Delta X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 10, 11, 12\}$ with a greedy algorithm: Fix the two outermost points based on the largest distance in ΔX, and remove the largest distance from it. Then repeat the following: Fix a point as near to the leftmost point as possible so that the largest distance in ΔX is satisfied, and remove the newly created inter-point distances from ΔX.

b) The greedy approach above correctly reveals all but one point in the example. Let us improve the strategy so that when fixing point to the left does not work, we try fixing it as near to the rightmost point as possible. Verify that this new strategy reveals also the last missing point in the example.

c) Verify that even the improved greedy strategy fails in revealing $X = \{0, 1, 4, 5, 8, 10\}$, given the corresponding inter-point distances.

Notice that a small modification of the above algorithm leads to a correct algorithm; instead of greedily choosing left/right, one examines both possibilities (see page 90 in J & P). This may lead to exponential branching, but usually (with high probability on random data) only one of the possibilities is feasible.