Algorithms for Bioinformatics (Autumn 2015)

Course Exam, Tuesday 20.10.2015 — Solutions and grading

See the course material for more detailed solutions.

1. **String reconstruction using Eulerian path approach.** (6+6 points)

   A sequencing by hybridization experiment produces the following (multi)set $S = \{\texttt{ACC},\texttt{ACT},\texttt{CAC},\texttt{CCA},\texttt{CTA},\texttt{GAC},\texttt{TAG}\}$.

   (a) Use the Eulerian path approach to deduce the unique string whose 3-mer composition is S. Show the intermediate steps.

   (b) Show that there is a 3-mer such that when added to S there are multiple strings whose 3-mer composition is the new S. How many such 3-mers there are? Justify your answer.

   Grading:

   (a)
   - +3 for correct deBruijn graph
   - +3 for correct string/path

   (b)
   - +3 for correct addition (AGA)
   - +2 for stating that there are no other such 3-mers
   - +1 for good justification

---

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

   Compare the problem of computing the edit distance (i.e., global alignment) and the k-errors problem (i.e., approximate string matching or fitting alignment 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 $P = \texttt{ACTT}$, text $T = \texttt{CATTACAT}$ 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 (ATT and ACAT) or alignment

3. **2-breaks and rearrangements.** (6+6 points)

   2-breaks on *cyclic* synteny block graphs correspond to three types of rearrangements on *circular* genomes.

   (a) Name the three types of rearrangements, and give an example of each using signed permutations.

   (b) For each type of rearrangement, give an example of a corresponding 2-break showing the synteny block graph before and after the 2-break.

   Grading:

   - +1 for each correct name (reversal, fission, fusion) or correct description
   - +1 for each correct example with signed permutations
   - +2 for each correct example of a 2-break

---

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 or pseudo-code