Algorithms in Bioinformatics

Andrey Prjibelski (andrewprzh@gmail.com),
Ira Vasilinetc (vasilinetc.ira@gmail.com)

April 16, 2014

Task 13

Implement Nussinov algorithm for predicting secondary RNA structure with the following constraints:

- Complimentary base pairs should be at least 4 non-complimentary nucleotides apart (see fig. 1)

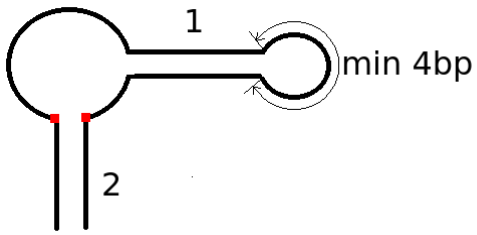

Figure 1: Hairpin loop must be at least 4 bases long. Automatically this condition makes all complementary base-pairs at least 4 nucleotides apart (including ones marked with red).

- The prediction structure should prefer stacked base pairs over individual base pairs. This can be accomplished by allowing every two stacked base pairs to have an additional score contribution d. You may try to find an appropriate value for d by varying it from 0 to some small positive number. To incorporate stacked base pair contributions, the DP algorithm may need to be revised to consider "state transition" because the first base pair and other base pairs are scored differently due to the stacked base pair score contribution. (see fig. 2)

Input:
RNA sequence in FASTA format.
Output:

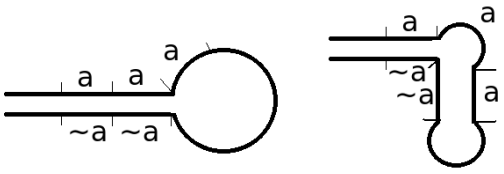

Figure 2: The structure on the left picture is preferred over the right one.

RNA secondary structure. Output all pairs of complimentary sequences with their positions in the RNA sequence.
Example:
Input:
>seq
CCCCCCACGUUUUUCCCUUUUGGGUUUUUACGUCCCCCCC
Output:
6,ACGU-29,ACGU
14,CCC-21,GGG

Task 14

Implement Zuker algorithm for predicting secondary RNA structure.
Input:
RNA sequence in FASTA format.
Output: RNA secondary structure. Output all pairs of complimentary sequences with their positions in the RNA sequence.
Example:
Input:
>seq
CCCCACGUUUUUACGUCCCC
Output:
4,ACGU-12,ACGU

Task 15

We define an exact tandem repeat as a subsequence that appears consecutively at least twice. For example, AATT is a tandem repeat in CCAATTAATTAATTCC. Design an efficient algorithm for finding the longest exact tandem repeat within a given sequence.
Input:
Sequence in FASTA format.
Output:
Longest tandem repeat in FASTA format.
Example:
Input:

>seq
AAAAAACGTACGTACGTAAAAAA
Output:
>repeat
ACGT

Task 16

Design an efficient algorithm for finding the longest exact repeat with at most one mismatch in a given sequence.

Input:
Sequence in FASTA format.

Output:
Two instances of the longest repeat with maximum 1 mismatch in FASTA format.

Example:

Input:
>seq
ACGTAAAAAAAAACCAAAATAAAACGG
Output:
>repeat1
AAAATAAAAC
>repeat2
AAAAAAAAAC