CS440
Assignment 3
Due March 13, 2019

Computer Science Department
Colorado State University

March 1, 2019

Preliminaries. The input is a string $S \in \Sigma^*$ in the alphabet $\Sigma = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{U}\}$. Let n be the length of S, and index characters of S from 1 to n. Character at location i is denoted by $S[i]$. S is the sequence of an RNA molecule. Characters in S are called *bases* or equivalently *nucleotides*.

Input file. The input to your program is called `input.txt`, which contains a representation of S. Make sure your program reads from that file in the current directory. The entry is put on one line. All characters are upper case. For instance,
line 1: `AACGAGUAUACGCGA`

Folding the input RNA. An RNA molecule folds into a 3D shape according to the laws of physics. Those laws can be simplified into the following constraints:

> **Constraint 1:** Certain pairs of nucleotides have tendency to form chemical bonds and make a pair. Particularly, `A - U`, `C - G`, and `G - U` are the only permissible pairings (Watson-Crick pairs).
>
> **Constraint 2:** Every base forms a pair with at most one other base, i.e. a base has two possible states: (i) unpaired, or (ii) paired with one other base.
>
> **Constraint 3:** There has to be at least 3 nucleotides in between the constituent ends of a pair, i.e., if $S[i]$-$S[j]$ is a base pair, then $j - i \geq 4$. RNA backbone cannot bend with high curvature hence this constraint.

In practice, the first step in RNA folding is secondary structure prediction, i.e., prediction of the list of base pairs. In this assignment, we simplify the laws of physics and assume RNA folds into a structure with maximum number of base pairs. A base pair, independent of its constituent bases, has unit score. We would like to find maximum score.

Origins of the problem. The problems of RNA folding and protein folding started in early 1970's upon discovery of biomolecule sequences. Knowing the sequence, the following question was spatial conformation. Those problems are still open due to complexity of atomic-level physics and computational complexity of search.

Output. Your program should print the maximum number of base pairs, subject to the three constraints above. Actual base pairs are not needed in the output. Your output is just one integer.

Grading. We will test your program on 10 different inputs and let it run for 1 minute each. Each test case is worth 10 points. If the output is correct, you get 10 points; otherwise, you get 0.

Implementation suggestions. The problem is in P. However, the polynomial time algorithm is beyond the scope of this course and takes time to implement. In this assignment, you are free to either implement the polynomial-time algorithm (if you happen to know it) or use any of the global optimization algorithms that we have discussed in class, namely simulated annealing, beam search, random restart, or genetic algorithms. Be prepared to deal with large input in 1 minute run time on the 120-unix-lab machines such as `denver`.

Upload your answer on Canvas in one zip file or tarball. Include a `README` file with running instructions and all the code/scripts you have written.