NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

School of Science

Information Technologies in Medicine and Biology

Direction: *Bioinformatics*

Algorithms in Structural Bioinformatics

Postgraduate Student: *Begetis Nikolaos*

Professor: *Emiris Ioannis*

Deadline Date: *05/03/2013*

Assignment 2

1(a).

We are assigned to find all optimal secondary structures of RNA sequence CACGGUUAG by Zuker's energy minimization, using initialization

$$j+3 > i \rightarrow V(i,j) = W(i,j) = \infty$$

hairpin energy h(i,j) = j-i+5, and stem function s(i,j)=-3,0 and 3, respectively, for Watson-Crick pairs, GU pairs, and all other pairs.

Energy Minimization W(i,j)	C	A	C	G	G	U	U	A	G
C	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 8	8 8	8 8	5 5	5 5
A	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 8	8 8	5 5	5 5
C	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 8	8 8	8 8	6 8
G	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 ∞	9 8	10 8
G	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 ∞	9 8
U	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	8 ∞
U	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞
A	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞	∞ ∞
G	∞	∞	∞	∞	∞	∞	∞	∞	∞

Figure 1: Zuker's energy minimization and backtrack path, using initialization j+3 > I ➔ V(i,j) = W(i,j) = ∞

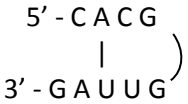
In addition to all the above, we could also ignore multiloops and bulges. Figure 1 shows the final results of the Zuker's energy minimization. It is worthy to mention that on the **bottomleft** corner of every table cell there are set the three out of the four possible scores given from the four possible ancestors of each cell, respectively. The best of the scores are highlighted in red color (we keep in mind that we need the minimum possible score), when they come from an ancestor cell, e.g. for the cell W(I,j), the ancestor cell are W(i-1,j) and W(i,j+1). When the best score comes from the bottomleft diagonal cell, it means that the minimum score comes from the refined energy function V(i,j) [1.dynpr.pdf, pp.30-32, I.Z. Emiris].

As we earlier mentioned we can ignore multiloops and interior bulges, and as a result the only two structure elements that affect in the energy function V(i,j) are hairpins and Watson-Crick stems. Actually, as shown in Figure 1, all the left and bottom scores come from ancestor cell, while all the diagonal scores come from the refined energy function V(i,j). So, in respect to that, we highlighted in blue color the best scores when they come from a hairpin and in green color the best scores when they come from a match, using the formulas given from the assignment.

The backtrack path is shown by the pale red cells beginning from the topright corner W(1,9) to the cell W(4,5). To figure out which path to follow we consulted [Sebastian Will[1,2]] lectures from Freiburg University. By using the referenced Zuker's algorithm we concluded that the bases C-G (W(1,9)), A-G (W(2,9)) and A-A (W(2,8)) do not match in pairs, while on the contrary A-U (W(2,7)) match to a Watson-Crick pair, making a stem. Finally, a pyr-pyr pair C-U (W(3,6)), and a pur-pur pair G-G (W(4,5)) makes a hairpin.

So, the optimal secondary structure of RNA is that shown in Figure 1, containing 1 matched pairs and 7 unmatched nucleotides, from which the two make a hairpin. This secondary structure may come from two different paths with the only difference between them, the base from which the initial base was predicted to begin.

Given this backtrack path, the RNA secondary structure is as follows:



Finally, an example of the table filling as shown in Figure 1 is illustrated below (the 1st step):

$$W(i,j) = \min \begin{cases} W(i+1,j) = W(2,4) = \infty \\ W(i+1,j) = W(1,3) = \infty \\ V(i,j) = V(1,4) \rightarrow h(1,4) = 4\text{-}1\text{+}5 = 8 \\ \qquad\qquad S(i,j) + V(i+1,j-1) = S(1,4) + V(2,3) = -3 + \infty = \infty \\ W(i,k) + W(k+1,j) : i+1 < k < j \rightarrow k = 3: W(1,3) + W(4,4) = \infty \end{cases}$$

[1] http://www.bioinf.uni-freiburg.de/Lehre/Courses/2008_WS/V_RNA/
[2] http://www.bioinf.uni-freiburg.de/Lehre/Courses/2008_WS/V_RNA/Slides/vorl3.pdf