A Local Alignment Approach to RNA folding

Ben Chugg, Coulter Beeson, Kenny Drabble, Jeff Jeyachandren

The University of British Columbia

April 6,2017

RNA Folding

RNA consists of the four base pairs Adenine (A), Guanine (G) , Cytosine (C) and Uracil (U). These base pairs of RNA pair in a complementary fashion: Adenine to Uracil (A – U) and Cytosine to Guanine (C – G).

Unlike DNA for which we are concerned with optimally aligning two strands, for RNA we are concerned with how the strand folds with itself.

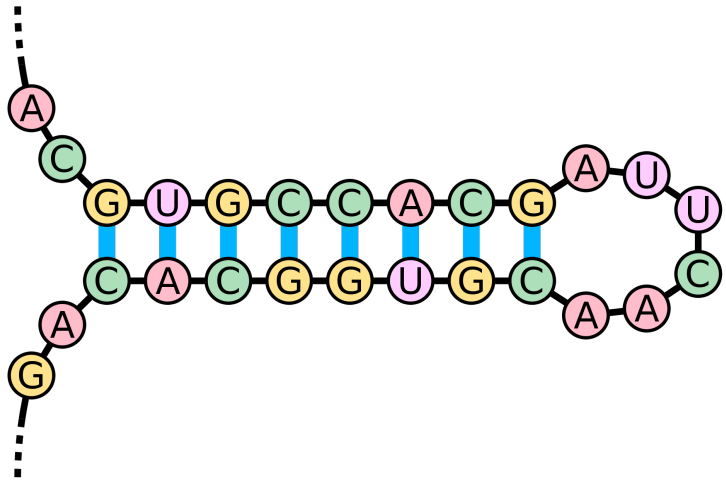

Figure: Source: http://rosalind.info/problems/pmch/

RNA Folding

There are several frameworks with which we can model RNA folding. We will use the Energy Minimization Model.
In this model, matches are scored as $+1$ and non-matches as 0. This lends itself to a dynamic programming approach. Let $r = r_1, \ldots, r_n$ be a strand of RNA, where $r_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ and let $S(i,j)$ denote the optimal score of folding the subsequence $s_i, s_{i+1} \ldots s_j \subset s$. Then,

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1, & \text{if } i,j \text{ base pair} \\ S(i+1, j), & \\ S(i, j-1), & \\ \max\limits_{i<k<j} \{S(i,k) + S(k+1, j)\}, & \text{bifurcation.} \end{cases}$$

RNA Folding

Problems with the Energy Minimization Model:

1. Simplicity
2. Does not capture pseudo-knots

Definition

A strand $r = r_1, \ldots, r_n$ is said to contain a pseudo knot if there exist indicies $i_1 < i_3 < i_2$ and $j_1 < j_3 < j_2$ such that $r_{i_1}, \ldots, r_{j_1}$ is aligned with $r_{i_2}, \ldots, r_{j_2}$ and $r_{i_3}, \ldots, r_{j_4}$ is paired with a subsequence outside the interval $r_{i_1}, \ldots, r_{j_2}$.

RNA Folding

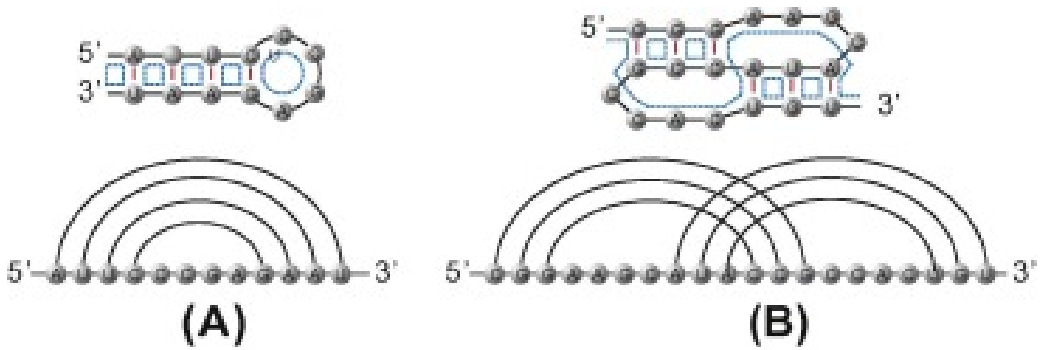

Figure: Source: `http://www.sciencedirect.com/science/article/pii/S0025556413001788`

We will call this typical RNA folding algoritm DP FOLDING .

k-Local Folding

Idea: Subsequences of an RNA strand which are (near) palindromes of each other are likely to be a good match. Formally, we define an algorithm k-Local Folding which takes as input an RNA strand r and a parameter k, runs a local alignment algorithm on the strand to find k high scoring — and disjoint — palindromic regions of r. It then passes the remaining unpaired regions to DP Folding to be folded as usual.

function k-Local Folding(r, k)
    Call Smith Waterman on r k times
end function