A Local
Alignment
Approach to
RNA folding

Ben Chugg,
Coulter
Beeson,
Kenny
Drabble, Jeff
Jeyachandren

Background

Our Approach
Runtime
Analysis

Results

Conclusion

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.
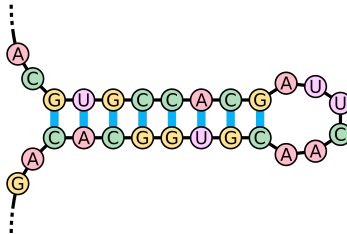

Figure: Source: http://rosalind.info/problems/pmch/

RNA Folding

A Local
Alignment
Approach to
RNA folding

Ben Chugg,
Coulter
Beeson,
Kenny
Drabble, Jeff
Jeyachandren

Background

Our Approach
Runtime
Analysis

Results

Conclusion

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 algorithm: Nussinov's algorithm which we will call NUSSINOV .

RNA Folding

A Local
Alignment
Approach to
RNA folding

Ben Chugg,
Coulter
Beeson,
Kenny
Drabble, Jeff
Jeyachandren

Background

Our Approach
Runtime
Analysis

Results

Conclusion

Let $r = r_1, \ldots, r_n$ be a strand of RNA, where $r_i \in \{A, C, G, U\}$ and let $S(i, j)$ denote the optimal score of folding the subsequence $s_i, s_{i+1} \ldots s_j \subset s$. Then,

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

It's clear that the runtime is $O(n^3)$.

k-Local Folding

Nussinov returns the mathematically optimal alignment under the energy minimization model. Therefore, any new approach cannot hope to beat the scores, but only improve the running time.

Idea: Subsequences of an RNA strand which are (near) palindromes of each other are likely to be a good match. Pair these regions and pass the leftover segments to Nussinov .

k-Local Folding

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 NUSSINOV to be folded as usual.

1: **procedure** K-LOCAL FOLDING(r, k)
2:    Initialize stack $S \leftarrow r$
3:    **while** The number of local alignments found is $\leq k$ and S not empty **do**
4:       $s \leftarrow \text{pop}(S)$
5:       Let $\bar{s}$ be the reverse of s
6:       Call Local Alignment on s and $\bar{s}$.
7:       **if** Local alignment found **then**
8:          Remove the aligned regions from s
9:          Push all unmatched regions of S back onto stack
10:       **end if**
11:    **end while**
12:    Call NUSSINOV on all unmatched regions of r.
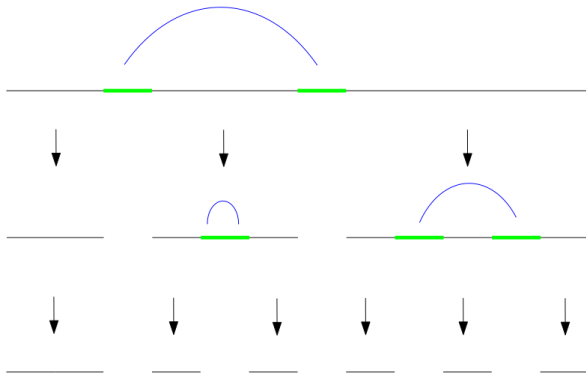13: **end procedure**

example

Figure: A strand of RNA undergoing multiple local alignments on successive subsequences. The unmatched regions at the bottom will be passed into Nussinov

k-Local Folding: Runtime Analysis

Fix an RNA strand r with length n.

Lemma

Let $\mathscr{A}$ be the set of local alignments found. Finding k disjoint local alignments of r takes time

$$O\left(\left(n - \sum_{A \in \mathscr{A}} \ell_A\right)^2 \log k\right),$$

where ℓ_A is the length of an alignment $A \in \mathscr{A}$.

Lemma

K-LOCAL FOLDING runs in time

$$O\left(\left(n - \sum_{A \in \mathscr{A}} \ell_A\right)^3 + \left(n - \sum_{A \in \mathscr{A}} \ell_A\right)^2 \log k\right).$$

A Local
Alignment
Approach to
RNA folding

Ben Chugg,
Coulter
Beeson,
Kenny
Drabble, Jeff
Jeyachandren

Background

Our Approach
Runtime
Analysis

Results

Conclusion

Proof of Lemma 1.
We can view the progress of the algorithm as a ternary tree:
amortized across each level we perform local alignment on a
sequence of size $n - \sum_{A \in \mathscr{A}} \ell_A$. Local alignment takes time
squared in the size of the input. Since the tree has depth
$\log(k)$, the result follows.   ◀

Proof of Lemma 2.
Running Nussinov takes cubic time in the size of the input.
K-Local Folding first finds k disjoint alignments, then runs
Nussinov on the remaining unmatched regions, which have
total length $n - \sum_{A \in \mathscr{A}} \ell_A$. The result follows by applying
lemma 1.   ◀

k-Local Folding: Runtime Analysis

We remark that since $\sum_{A \in \ell_A} \ell_A^3 \in O(n^3)$, $O((n - \sum_{A \in \mathscr{A}} \ell_A)^3) = O(n^3)$ so there is no asymptotic difference between K-LOCAL FOLDING and NUSSINOV. However, our hypothesis was that there may a difference in the run times in practice.

Run Time Results

A Local
Alignment
Approach to
RNA folding

Ben Chugg,
Coulter
Beeson,
Kenny
Drabble, Jeff
Jeyachandren

Extensions and Further Research

1. Modify K-LOCAL FOLDING to report possible pseudoknots
2. Compare K-LOCAL FOLDING to the Four Russians method of speeding up Nussinov.
3. Probabilistic Approach