Methods and Algorithms

RNA STRUCTURE PREDICTION

Overview

- Nucleic Acid Structures
- Secondary Structure Prediction
- Tertiary Structure Prediction
- Quaternary Structure Prediction
- A Dynamic Programming Approach
- CyloFold: Secondary Structure Prediction
- Vienna RNA Package
- References
Nucleic Acid Structures

- Primary Structure
  - the sequence of RNA bases
- Secondary Structure
  - a two dimensional folding containing an annotation of which base pairs are formed.
- Tertiary Structure
  - a three dimensional folding containing a base sequence with base pair annotation and we describe the spatial location of every atom
- Quaternary Structure
  - Higher-level organization: DNA into chromatin; RNA interactions between RNA units in the ribosome or spliceosome, etc.
Primary Structures

- DNA
- RNA
  - mRNA: Messenger RNA is encoding for the primary protein sequence
  - tRNA: Transfer RNA is active in the translation process at the ribosome, binds an amino acid with its anti-sense codon.
  - rRNA: Binds with proteins to form the ribosome.
  - snRNA: Short RNA performing splicing, regulation, etc. in the nucleus.
  - miRNA: Short RNA inhibiting translation of mRNA, or increasing RNA structure degradation.

RNA Base Pairing

DNA
- Thymine (T), Adenine (A), Cytosine (C), Guanine (G)
- Sugar-phosphate backbone with deoxyribose
- Usually double stranded
- Bonds T-A, C-G

RNA
- Uracil (U) (instead of T), Adenine (A), Cytosine (C), Guanine (G)
- Sugar-phosphate backbone with ribose
- Usually double stranded
- Bonds A-U, G-C, G-U
Sugar-phosphate backbone with ribose

RNA Base Pairs

G-U Wobble pair highly unstable
Nucleic Acid Pairs

Wobble Pairs
Stable RNA Pairs

Secondary Structure

- Different parts of the strands can be connected to each other.
- DNA
  - C – A
  - T – G
- RNA
  - A – U (2 H-bonds)
  - G – C (2 H-bonds)
  - G – U (‘wobble’ pair: very unstable, 3 H-bonds)
RNA Secondary Structure

- Single Strand
  - unpaired bases

- Stem
  - Pairing of palindromic DNA or RNA structures
  - DNA: ---CCTTAGXXXXXXXCTAAGG---
  - RNA: ---CCUUAGXXXXXXXCUAAGG---

Hairpin Loop
- Loop at the end of a stem, not containing any other structures
- DNA: ---CCTTAGXXXXXXXCTAAGG---
- RNA: ---CCUUAGXXXXXXXCUAAGG---
- Empirical: loop size $|XXX...XXX| > 3$

Bulge Loop
- Loop at want side of a stem, because of non matching insert in one side of the palindromic sequence
- DNA: ---GGGCCCTTAGXXXXXXXCTAAGGNNNNNNCC---
- RNA: ---GGGCCUUAGXXXXXXXCUAAGGCC---

Interior Loop
- Two non matching parts of the sequence between 2 stems.
- DNA: ---GGGCCCTTAGXXXXXXXCTAAGGNNNNNNCC---
- RNA: ---GGGCCUUAGXXXXXXXCUAAGNNNNNNCCC---
RNA Secondary Structure

- Junction (multiloop)
  - Part of the secondary structure where multiple ends of stems meet.
- Pseudoknot
  - Pairing between bases on hairpinloops.
  - “Common belief that pseudoknots contribute very little to the energy balance of the RNA molecule. For that reason, and for another computational reason, it is a common practice to ignore their locations when predicting RNA folding.” (?)
Tertiary Structure

Image from: http://www.tutorvista.com/biology/trna
Quaternary Structure

http://rna.ucsc.edu/rnacenter/ribosome_images.html
Interface views of the 50S (left) and 30S (right) ribosomal subunits. with labels

70S Ribosome (open book view)
**Formal definition of RNA secondary structure**

**Definition:**
Given a single stranded RNA sequence of length $N$, $R = r_1, \ldots, r_N$, a secondary structure of $R$ is defined to be a set $S$ of disjoint base pair indices, $(i, j)$ such that $1 \leq i < j \leq N$ and $r_i$ and $r_j$ form a valid (matching) base pair.

**Restriction:**
adjacent and very close bases cannot be paired, i.e., we have the following restriction over all pairs $(i, j): j - i > 3$

Empirical evidence suggests that the minimal hairpin loop size is greater than 3.

**Nested Edges Graph**

The RNA secondary structure can be described as a graph $G = (V, E)$ where $V$ is the set of bases in $R$ and $E$ is defined as:

$E = \{(v_i, v_j) | i = j - 1\} \cup \{(i, j) \text{ in } S\}$,

where $S$ is the set of disjoint base pair indices defining the RNA secondary structure.

This defines a graph with two types of edges:
- edges between subsequent bases defining the primary structure
- edges between base pairs defined by the secondary structure.

If we restrict the secondary structure, so that it cannot contain Pseudoknots, the graph representation of the secondary structure would be obeying the Nested Edges Graph definition:

*for all $(v_i, v_j), (v_k, v_l) \in E, \text{k in [i, j] if and only if } l \text{ in [i, j]}*
Optimal RNA Secondary Structure

The optimal secondary structure of RNA is commonly viewed as the most energetically stable RNA secondary structure.

Idea

- Define a stability score for a given secondary structure.
- A greater number of valid pairs (A-U and C-G) corresponds to a more stable secondary structure.

More refined

- allow also the loose pair G-U
- take into account different weights for different motifs in the secondary structure

Algorithm for maximizing the number of base pairs

- The problem of finding the optimal base pair matching of a given primary RNA sequence is an NP-Complete problem (see also [3]).

- Base pair maximization though not optimal, is a good heuristic algorithmic approach for predicting the RNA secondary structure, using sequence alignment.

- Base pair maximization is done using a dynamic programming algorithm, that dynamically builds the optimal secondary structure of all subsequences of the given RNA.

Base Pair Maximization a Dynamic Programming Approach

$S(i, j)$ is defined to be the number of base pairs in the folding of the subsequence $r_i...r_j$ of $R$ which results in the highest number of base pairs.

Base condition:

$S(i, j) = 0$ for each $i, j$ such that $i = j, i = j + 1$

The four recurrence options:

1. Adding another base pair: $r_i r_j$ are a valid base pair, and we add it to the best folding of subsequence $r_{i+1}...r_{j-1}$
2. Adding another base at the beginning of the sequence $r_{i+1}...r_j$
3. Adding another base to the end of the sequence $r_i...r_{j-1}$
4. Finding the best cut point $k$ of the sequence $r_i...r_j$ (a bifurcation)

Initial matrix $S$ for the RNA Sequence GGGAAAAUCC
Base Pair Maximization a Dynamic Programming Approach

$S(i, j)$ is defined to be the number of base pairs in the folding of the subsequence $r_i...r_j$ of $R$ which results in the highest number of base pairs.

The four recurrence options:

1. Adding another base pair: $r_i, r_j$ are a valid base pair, and we add it to the best folding of subsequence $r_{i+1}...r_{j-1}$

2. Adding another base at the beginning of the sequence $r_{i+1}...r_j$
Base Pair Maximization a Dynamic Programming Approach

$S(i, j)$ is defined to be the number of base pairs in the folding of the subsequence $r_i...r_j$ of $R$ which results in the highest number of base pairs.

3. Adding another base to the end of the sequence $r_i...r_{j-1}$

4. Finding the best cut point $k$ of the sequence $r_i...r_j$ (a bifurcation)
Base Pair Maximization a Dynamic Programming Approach


Base Pair Maximization a Dynamic Programming Approach

Recurrence relation:

\[ S(i, j) = \max \begin{cases} S(i + 1, j - 1) + 1 & \text{(if i,j is a base pair)} \\ S(i + 1, j) \\ S(i, j - 1) \\ \max_{i < k < j} S(i, k) + S(k + 1, j) \end{cases} \]

Optimal secondary structure score \( S(i,j) \) is in upper-right corner.
Pseudoknots

Note that \( S(i, j) \) is independent of the overall structure, and relies only on the local structure of \( r_i \ldots r_j \).

Pseudoknots
- Are base pairs between distant hairpin loop motifs.
- Two distant hairpin loops are found in the middle of two different subsequences
- They do not change the local score of the two subsequences
- A pseudoknot increases the overall score of the fold.
- Enumerating all potential structures is exponential in the number of bases.

Solution: do not allow pseudoknots [3,4]. Reasonable assumption [2]:
1. Close pseudoknots contribute very little to the overall stability of the fold.
2. Distant pseudoknots are very rare.

Note: For three-dimensional modeling of the RNA structure, Pseudoknots cannot be ignored, since they actually affect the spatial position of atoms.

Time and space complexity of the algorithm

- The algorithm requires the computation of a two-dimensional matrix of size \( N \times N \).
- At each recurrence step, calculating the first three options take a constant time, but going over at most \( N \) values of \( k \) at each cell of the matrix takes \( O(N) \).

Total time complexity is then \( O(N^3) \).
Space complexity, storing the matrix is \( O(N^3) \).
Drawbacks

Major drawbacks of the dynamic programming algorithm:
- Pseudoknots are ignored.
- The score incompletely reflects the stability model of the secondary structure obtained.

The first drawback is addressed in:
- CyloFold: secondary structure prediction including pseudoknots.

The second drawback was addressed by other algorithms that model the stability of the secondary structure better:
- The basic DP (Dynamic Programming) algorithm[9] only takes the number of base pairs into account,
- mFOLD[8] and Vienna[5] are based on a scoring system that takes the different motifs into account.

Open Problems

In [2] the score of each motif is drawn from laboratory experiments that measured the stability of the different motifs. (more stable => more likely).

Assumption:
- The stability of a given subsequence, is independent from the rest of the sequence, and only depends on the subsequence itself and its subsequences.

=> Dynamic Programming algorithms with more sophisticated recurrence rules follow.

But: in reality the equation “more stable => more likely” does not always hold!
- Algorithms based on thermodynamics predict only 50-70% of the base pairs correctly.
- Also a large number of RNA structures lie within 5-10% of the predicted global energy minimum[1].

=> RNA sequence folding and secondary structure prediction still an open and challenging problem.
References


CyloFold: secondary structure prediction including pseudoknots

by

Eckart Bindewald, Tanner Kluth and Bruce A. Shapiro

doi:10.1093/nar/gkq432
CyloFold: secondary structure prediction including pseudoknots

Computational RNA secondary structure prediction:
• approaches only allow a limited class of pseudoknot interactions or not at all.

CyloFold
• no restrictions in terms of pseudoknot complexity.
• based on simulating a folding process in a coarse-grained manner by choosing helices based on established energy rules.
• The steric feasibility of the chosen set of helices is checked during the folding process using a highly coarse-grained 3D model of the RNA structures.

The program is available as web server: http://cylofold.abcc.ncifcrf.gov.

Steric effects

• Steric effects arise from the fact that each atom within a molecule occupies a certain amount of space. If atoms are brought too close together, there is an associated cost in energy due to overlapping electron clouds (Pauli or Born repulsion), and this may affect the molecule’s preferred shape (conformation) and reactivity.
• The steric effect of tri-(tert-butyl)amine makes electrophilic reactions, like forming the tetraalkylammonium cation, difficult. It is difficult for electrophiles to get close enough to allow attack by the lone pair of the nitrogen (nitrogen is shown in blue)
Results

- The method is evaluated using two data sets of 26 and 241 RNA sequences.
- The approach is competitive compared to existing RNA secondary structure prediction programs:
  - pknotsRG
  - HotKnots and
  - UnaFold

Introduction

RNA and Structure

- Execute a huge variety of biochemical functions
- The secondary and tertiary structure is vital to the function of RNA such as ribosomal RNA, RNAase P or tRNA
- Experimentally determined structures are only available for a small fraction of RNAs.

=> Computational prediction of the base-pairing pattern (the secondary structure) of RNA is an important problem.
State of the Art

First breakthrough (1978) (1–5):
- dynamic programming algorithms that could predict the minimum free energy secondary structure of RNA sequence
- assumption that the structures are non-nested

More recently (6,7):
- Dynamic programming algorithms have been extended to allow certain classes of pseudoknots.

State of the Art

Problems to solve:
- RNA structures determined by X-ray crystallography or NMR showed that many RNAs contain non-nested base pairing interactions.
- Allowing all possible base pairing interactions leads to:
  - many more conformations to consider,
  - many conformations that are not sterically feasible.

RNA secondary structure prediction algorithms:
- iteratively adding substructures to an initially unfolded sequence (8,9).
- Genetic algorithms for exploring pseudoknotted structures and sub-optimal RNA structures (10–14). (A.P. Golt'yaev et al. 1995)

CyloFold:
- RNA secondary structure prediction without any restrictions in terms of pseudoknot complexity
- Check on the steric feasibility of the considered conformations.
CyloFold Algorithm

Underlying idea:
- maximizing matching helices in a secondary structure (10).

The Algorithm
- Initially, a list (called a stem-list) of all possible helices with more than 3 bp is generated.
- Helices can contain (stable) Watson–Crick and GU–wobble base pairs.
- The score is set to be the sum of the free energy contribution of the already placed helices.
- Each folding simulation run is performed by picking helices from the stem list with a Boltzmann-weighted probability.
- Estimating the free energy contribution of an RNA double-helix is accomplished using the RNA Vienna package (2).

Vienna RNA Web Servers
The Vienna RNA Servers

- **RNAfold server** predicts minimum free energy structures and base pair probabilities from single RNA or DNA sequences.
- **RNAalifold server** predicts consensus secondary structures from an alignment of several related RNA or DNA sequences. You need to upload an alignment.
- **RNAinverse server** allows you to design RNA sequences for any desired target secondary structure.
- **RNACofold server** allows you to predict the secondary structure of a dimer.
- **RNAup server** allows you to predict the accessibility of a target region.
- **LocARNNA server** generates structural alignments from a set of sequences. In collaboration with the Bioinformatics Group Freiburg.

The Vienna RNA Servers

- **barriers server** allows you to get insights into RNA folding kinetics.
- **RNAz server** will assist you in detecting thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple sequence alignments.
- **Structure conservation analysis server** will assist you in detecting evolutionarily conserved RNA secondary structures in multiple sequence alignments.
- **RNAstrand server** allows you to predict the reading direction of evolutionarily conserved RNA secondary structures.
- **RNAxs server** assists you in siRNA design.
- **Bcheck** predicts rnpB genes
The Vienna RNA Servers

Get the Source code for:
- the Vienna RNA Package, our basic RNA secondary structure analysis software.
- The ALIDOT package for finding conserved structure motifs (add-on)
- The barriers program for analysis of RNA folding landscapes.

CyloFold Algorithm

<Begin CyloFold Algorithm Simulation>
- Each chosen helix is represented by a very coarse-grained 3D representation in a virtual 3D workspace, called a capsule.

A capsule is determined as follows:
An RNA double helix is represented by a cylinder (using a radius of 6.5 Å and a length of 2.7 Å times the number of base pairs) capped with a half-sphere on both ends.
CyloFold Algorithm

- Single stranded regions between helices constraints the maximum distance between the ends of the capped cylinders.
- New capsules are placed into the 3D simulation space at a random position obeying the distance-constraints.

The distance constraints
- The maximum distance between helix ends is $2.0 \text{ Å} + n \times 0.8 \text{ Å}$ with $n$ the length of the single strand.
- The minimum distance is $2.0 \text{ Å}$.

Restricted Capsules in 3D Space
CyloFold Algorithm

While (capsules collide and the number of attempts < 20) do
    placed capsule at a different random position;
    attempts += 1;
od

If (attempts == 20 and still capsule collision) then
    Optimize the positions of all Capsules thus far in order to minimize collisions and constraint violations.
    If (there is no collision-free position) then
        Discard the new helix and its capsule;
    else
        Store the new collision-free position;
    fi
fi

CyloFold Algorithm

- Helices that are part of the stem-list and that share bases with the newly placed helix are removed from the stem-list.
- In the next iteration the next helix is chosen until no more helices can be placed.
- Once no more helices can be placed, one simulation run is completed.

<End CyloFold Algorithm Simulation>

This total simulation process is run 50 times. The overall best-scoring structure is returned to the user.
CyloFold Algorithm

Implementation

- The CyloFold Algorithm is implemented in C++
- The CyloFold web (using Grails framework (18))
- A secondary structure prediction is computed using the cylofold binary on a Linux cluster.
- VARNA (19) is used to generate an image of the secondary structure prediction.
- The prediction results are stored in a relational database.
CyloFold Usage

- Enter a nucleotide sequence (as raw characters or in FASTA form, both ACGU and ACGT alphabets are accepted)
- The maximum sequence length: 300 nt.
- Returns a unique id, to access results at a later time.
- The server may take several minutes to compute a secondary structure prediction.

CyloFold Output

The prediction result is presented as
1. An image of the predicted RNA secondary structure created by VARNA (19);
2. An extended bracket notation in which nested base pairs are denoted as pairs of nested parentheses and helices corresponding to pseudoknot interactions are denoted as letters;
3. The ‘CT’ file format (also used by mfold (5)): a list of the indices of the bases and their predicted base-pairing partners.
Results

- The method is evaluated using two data sets of 26 and 241 RNA sequences.
- The approach is competitive compared to existing RNA secondary structure prediction programs:
  - pknotsRG
  - HotKnots and
  - UnaFold

Results: Data Sets

<table>
<thead>
<tr>
<th>PDB</th>
<th>Description</th>
<th>L</th>
<th>PKF</th>
<th>CF</th>
<th>PK</th>
<th>HK</th>
<th>UF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA30</td>
<td>TYMY iRNA-like structure</td>
<td>44</td>
<td>13.6</td>
<td>0.74</td>
<td>0.77</td>
<td>0.71</td>
<td>0.96</td>
</tr>
<tr>
<td>IC20</td>
<td>HDV ribozyme</td>
<td>72</td>
<td>22.2</td>
<td>0.61</td>
<td>1.00</td>
<td>0.60</td>
<td>0.06</td>
</tr>
<tr>
<td>IH3</td>
<td>SRV-1 pseudoknot</td>
<td>56</td>
<td>33.3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>HHV</td>
<td>HHV RT head pseudoknot</td>
<td>20</td>
<td>26.6</td>
<td>0.95</td>
<td>1.00</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>JK9</td>
<td>MMTV RNA pseudoknot</td>
<td>32</td>
<td>25.0</td>
<td>0.85</td>
<td>1.00</td>
<td>0.73</td>
<td>0.85</td>
</tr>
<tr>
<td>IH16</td>
<td>HCV IRES domain</td>
<td>42</td>
<td>0.0</td>
<td>0.74</td>
<td>0.77</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>IPY</td>
<td>PEMV-1 PIP pseudoknot</td>
<td>27</td>
<td>22.2</td>
<td>0.89</td>
<td>1.00</td>
<td>0.80</td>
<td>0.94</td>
</tr>
<tr>
<td>I1K</td>
<td>Group I self-cleavage intron</td>
<td>70</td>
<td>0.0</td>
<td>0.91</td>
<td>0.87</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td>I1LN</td>
<td>Vical RNA pseudoknot</td>
<td>27</td>
<td>22.2</td>
<td>0.94</td>
<td>1.00</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>I09A</td>
<td>25S rRNA sarcin/ricin</td>
<td>27</td>
<td>0.0</td>
<td>0.91</td>
<td>0.83</td>
<td>1.00</td>
<td>0.77</td>
</tr>
<tr>
<td>I1UD</td>
<td>Guanine riboswitch</td>
<td>67</td>
<td>11.9</td>
<td>0.87</td>
<td>0.57</td>
<td>0.67</td>
<td>0.88</td>
</tr>
<tr>
<td>2A43</td>
<td>Luteoviral pseudoknot</td>
<td>26</td>
<td>23.0</td>
<td>0.93</td>
<td>1.00</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>5G1W</td>
<td>tmRNA pseudoknot</td>
<td>22</td>
<td>10.1</td>
<td>0.81</td>
<td>1.00</td>
<td>0.67</td>
<td>0.86</td>
</tr>
<tr>
<td>Z226</td>
<td>XMRV-1 riboswitch</td>
<td>34</td>
<td>8.5</td>
<td>0.80</td>
<td>0.75</td>
<td>0.95</td>
<td>0.89</td>
</tr>
</tbody>
</table>

L: Sequence length, PKF: fraction of pseudoknot interactions. For each of the four different prediction methods (CF, Cjfold, PK, pknotsRG, HK, HotKnots 2.0, UF, UnaFold) we report three different measures of prediction quality (SNS, sensitivity; PPV, positive predictive value).

SNS: average base pair prediction sensitivity.
PPV: the positive predictive value (how often are predicted base pairs part of the reference secondary structure),
Results

Table 2. Prediction results for a set of 241 RNA sequences that are part of PseudoBase for the programs CyloFold, pknotsRG (7), HotKnots 2.0 (8) and UNAFold (23)

<table>
<thead>
<tr>
<th></th>
<th>MCC</th>
<th>SNS</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CyloFold</td>
<td>0.752</td>
<td>0.763</td>
<td>0.747</td>
</tr>
<tr>
<td>pknotsRG</td>
<td>0.748</td>
<td>0.753</td>
<td>0.756</td>
</tr>
<tr>
<td>HotKnots 2.0</td>
<td>0.611</td>
<td>0.565</td>
<td>0.684</td>
</tr>
<tr>
<td>UNAFold</td>
<td>0.597</td>
<td>0.532</td>
<td>0.692</td>
</tr>
</tbody>
</table>

SNS, sensitivity of predicted base pairs; PPV, positive predictive value.

MCC: The average Matthews correlation coefficient is obtained by comparing the base pairing pattern of the predicted secondary structures with their respective reference secondary structure.

References


