COMP 598: RNA inverse folding

Jérôme Waldispühl
School of Computer Science
McGill University
RNA folding

Sequence ensemble

Structure ensemble

MFE

UUUAAGGCCAGC

UUUACGGCUAGC

UUUAACGGCCAGC

UUUAGGGCCAGC

UAUACGGCCAGC

UUUAAGGCCAGC

CCUCAACGAAAGC

UCUGAAACCCGU

CCUCAACGAAAGC

UUUAAAGGCCAGC

UUUAACGGCCAGC

UUUAAGGCCAGC

UCUGAAACCCGU
RNA inverse folding

Sequence ensemble:
- CCUCAACGAAGC
- UUUACGGCUAGC
- UAUACGGCCAGC
- UUUAGGGCCAGC
- UUUAAGGCCAGC
- UCUGAAACCCGU

Structure ensemble:
- MFE

Diagram showing the relationship between the sequence ensemble and the structure ensemble.
Applications

(Chappell et al., 2015)

(Grabow & Jaeger, 2014)
Complexity

Folding: $\Theta(n^3)$

Inverse Folding: NP-hard?

- (Schnall-Levin et al., 2008): Design NP-hard for HMM & SCFG
- (Manuch et al., 2015): Rules for RNA designability
Local Search

Target

Design

Compute MFE structure
Target

Design

Local Search

Mutate sequence to get closer to target
Local Search

Target

Design

Fold and mutate again
Local Search

Target

Design
Local Search

2 Step process:
1. Seed selection
2. Adaptive walk
2 Step process:

1. Seed selection
2. Adaptive walk

Require:

- Folding algorithm (MFE or ensemble prediction)
- Metric to estimate the similarity between the target and predicted structure.
Objective function

- Random: Satisfy base pair complementarity
- Energy minimization: lowest free energy
- MFE: Guarantee MFE criterion
- Sequence symmetry minimization (SSM): Prohibit repeated subsequences
- Energy minimization + SSM
- Probability: optimize frequency in ensemble
- Average incorrect nucleotides $n(s^*)$

\[
p(s^*) = \frac{e^{-\beta\cdot E(s^*)}}{Q} \]

\[
n(s^*) = N - \sum_{1 \leq i \leq N} \sum_{1 \leq j \leq N+1} P_{i,j}S_{i,j}^* \]

(Dirks et al., 2004)
Objective function

(Dirks et al., 2004)
Timeline

1994  RNAinverse       Original program
2006  INFO-RNA         Optimized seed selection
2007  RNA-SSD          Improved search strategy
2011  NUPACK           Optimized objective function
2012  Frnakenstein     Multiple target structures
2013  IncaRNAation     Improved sampling of seeds
2014  RNAiFold         Exhaustive search
Connected sequence space
Disconnected sequence space under constraints

Forbidden motifs: \{ AU, UA\}
IncaRNAtion

Our approach: Inside-outside algorithm

Features:
• Linear time and space complexity
• No seed
• Sequence constraints (GC-content, Profile, etc.)
• Can be used to generate better seeds for local search algorithms.

(Reinharz et al., 2013)
Computing the sequence landscape

(Reinharz et al., 2013)
Extracting good candidates

(Reinharz et al., 2013)
Optimizing seeds

Secondary structure fixed

(Reinharz et al., 2013)
IncarRNAtion Algorithm

\[ Z = \text{weight of } nt^1 \]
IncaRNAtion Algorithm

\[ Z = \text{weight of } nt^2 \land nt^{17} \]
IncaRNAtion Algorithm

$$Z = \text{weight of } nt^3 \land nt^{16}$$

![Diagram showing RNA structure with probabilities and weights]

**Probability**

- Green: 0.0
- Orange: 0.5
- Blue: 1.0
IncaRNAtion Algorithm

\[ Z = \text{weight of } nt^4 \land nt^8 \]
IncaRNAtion Algorithm

\[ Z = \text{weight of } nt^{10} \land nt^{14} \]
Technical details

• After rejection, the weights only impact the performance, not the probability (i.e. unbiased).

• Complexity \( \mathcal{O}(n^3 \cdot k^2 + m \cdot k \cdot n\sqrt{n} \cdot \log(n)) \)
  where \( n \) size, \( k \) #mutations, \( m \) #samples.

• Partition function can be written as a polynomial:

\[
Z = \sum_{i=0}^{n} a_i \cdot w^i
\]

After \( n \) iterations we can calculate all \( a_i \)'s and exactly solve the weight/C+G% relationship.

Remark: In practice, less iterations are necessary.
Results: Success rate

IncaRNAtion only:

IncaRNAtion + RNAinverse (global + local):

(Reinharz et al., 2013)
Results: Running time

![Graph showing running time vs. length for Incarnation + RNAinverse and RNA-SSD. The graph peaks at approximately 80 length units.](graph.png)

(Reinharz et al., 2013)