# Local Search for RNA Design

Denny C. Dai, Herbert H. Tsang and Kay C. Wiese

Bioinformatics Research Lab
School of Computing Science, Simon Fraser University
Canada

CIBCB, Mar. 30th 2009

# Background

- RNA primary sequence
  - Ribonucleic Acid molecule consisting of a long chain of nucleotide units
  - string of length $n$ over alphabet set $\{A, C, G, U\}$
- RNA secondary structure
  - coarse-grained representation of RNA spatial shape
  - formed through base pairing among nucleotides
  - stable in minimum free energy (MFE) ground state

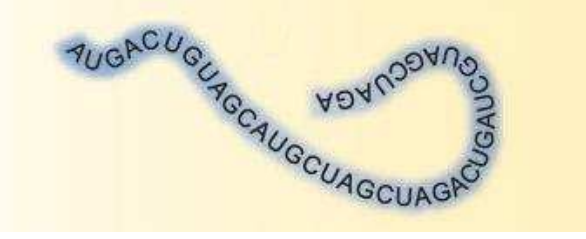



Figure: RNA Primary Sequence

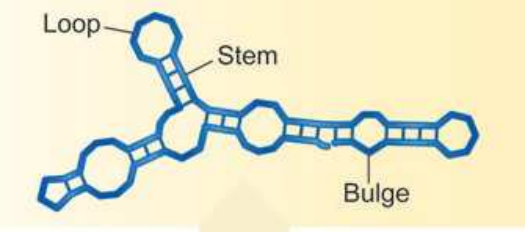
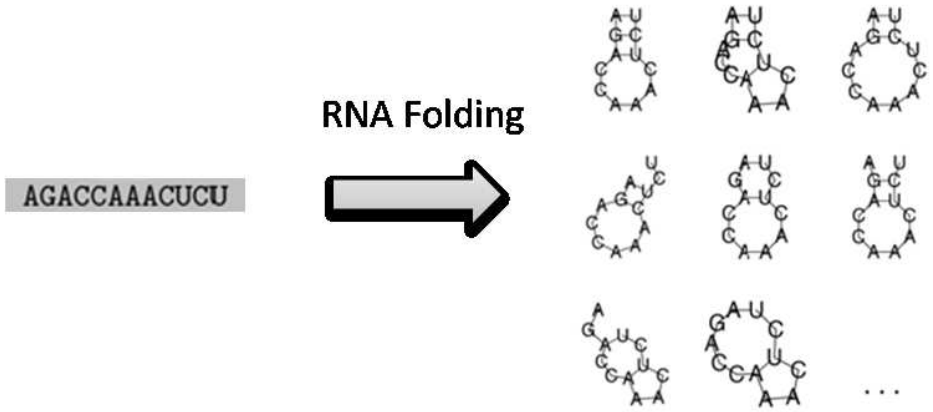



Figure: RNA Secondary Structure

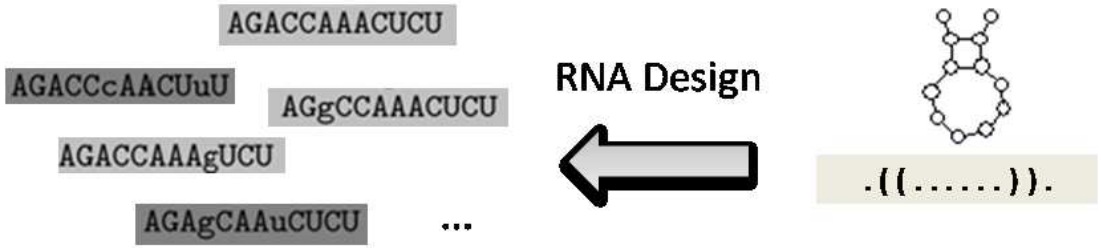
# Background

- RNA Folding Problem
  - finding the minimum free energy (MFE) secondary structure
  - efficient dynamic programming algorithm
    - $\mathbf{O}(n^3)$ under simplified thermodynamic model



# Background

- RNA Design Problem
  - finding RNA primary sequence folded into target shape
  - NP hard
    - heuristic search algorithms



## General design criteria

RNA design shall exhibit both high affinity and specificity.

- Sequence affinity
  - Target folding energy $e(r, S^*)$ measures thermodynamic stability. [1]
- Structure specificity
  - Soft constraint: structural distance $d(S^o, S^*)$ is small
  - Hard constraint: $d(S^o, S^*) = 0$

[1] $r$ is RNA sequence, $S^o$ is the MFE structure of $r$, $S^*$ is the target structure for design

## General design criteria

RNA design shall exhibit both high affinity and specificity.

- Sequence affinity
  - Target folding energy $e(r, S^*)$ measures thermodynamic stability. [1]
- Structure specificity
  - Soft constraint: structural distance $d(S^o, S^*)$ is small
  - Hard constraint: $d(S^o, S^*) = 0$

[1] $r$ is RNA sequence, $S^o$ is the MFE structure of $r$, $S^*$ is the target structure for design

## Design of local search heuristics

### Effective exploration of high-dimensional sequence space

- Structure distance

  - minimize $d(S^o, S^*)$
  - explicit specificity design paradigm
  - $d = 0$ leads to exact design

## Design of local search heuristics

### Effective exploration of high-dimensional sequence space

- Structure distance
- Folding probability

- maximize probability of folding $r$ into target structure $S^*$

$$p(r, S^*) = \frac{1}{Q} e^{-E(r,S^*)/RT}$$

where

$$Q = \sum_{S \in \Omega} e^{-E(r,S)/RT}$$

- implicit specificity & affinity design $p \in [0, 1]$

## Design of local search heuristics

### Effective exploration of high-dimensional sequence space

- Structure distance
- Folding probability
- Nucleotide error

- minimize incorrect base content that are not properly paired

$$n(S^*) = N - \sum_{i,j} P_{ij} S^*_{ij}$$

  - $P_{ij}$: probability of forming base pair between position $i, j$ (over all structure assembles at sequence $r$).

## Design of local search heuristics

### Effective exploration of high-dimensional sequence space

- Structure distance
- Folding probability
- Nucleotide error

- minimize incorrect base content that are not properly paired

$$n(S^*) = N - \sum_{i,j} P_{ij} S^*_{ij}$$

- Sequences adopting structure close to $S^*$ has small $n(S^*)$ error.
- $n(S^*) \approx 0$ is equivalent to $p(S^*) \approx 1$

## Design of local search heuristics

Effective exploration of high-dimensional sequence space

- Structure distance
- Folding probability
- Nucleotide error
- Target fold energy

- minimize target fold energy

$$e(r, S^*)$$

- measuring thermodynamic stability while adopting $S^*$.
- explicit affinity design heuristics

# Comparison of Algorithms

| | RNAinverse | RNA-SSD | INFO-RNA | rnaDesign |
|---|---|---|---|---|
| struct dis' | √ | √ | √ | √ |
| folding prob | √ | | | |
| nt errors | | | | √ |
| folding energy | | | √ | √ |
| initial phase | random | sampling | DP | random |

RNAinverse

'i' adaptive local search

'ii' sensitive to initial search position (random restart)

'iii' easily trapped in local optima for exact design

'iv' less scalability

# Comparison of Algorithms

| | RNAinverse | RNA-SSD | INFO-RNA | rnaDesign |
|---|---|---|---|---|
| struct dis' | √ | √ | √ | √ |
| folding prob | √ | | | |
| nt errors | | | | √ |
| folding energy | | | √ | √ |
| initial phase | random | sampling | DP | random |

RNA-SSD

'i' Hierarchical decomposition/reassemble & Stochastic local search

'ii' runtime grows polynomially with $N$

'iii' heuristic initial sequence sampling

# Comparison of Algorithms

| | RNAinverse | RNA-SSD | INFO-RNA | rnaDesign |
|---|---|---|---|---|
| struct dis' | √ | √ | √ | √ |
| folding prob | √ | | | |
| nt errors | | | | √ |
| folding energy | | | √ | √ |
| initial phase | random | sampling | DP | random |

INFO-RNA

'i' Dynamic programming & Stochastic local search

'ii' DP for constructing initial sequence (target fold energy)

'iii' fastest runtime

# Comparison of Algorithms

| | RNAinverse | RNA-SSD | INFO-RNA | rnaDesign |
|---|---|---|---|---|
| struct dis' | √ | √ | √ | √ |
| folding prob | √ | | | |
| nt errors | | | | √ |
| folding energy | | | √ | √ |
| initial phase | random | sampling | DP | random |

rnaDesign

'i' stochastic local search

'ii' adaptive random walk in the sequence space

# rnaDesign Local Search

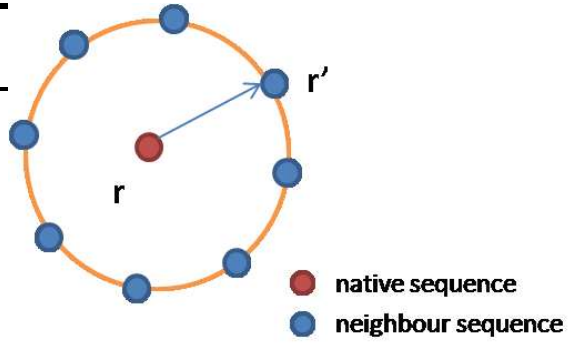
**Algorithm 1** rnaDesign

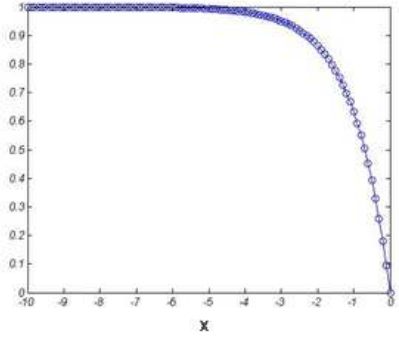
**Require:** initial sequence $r_0$, target structure $S^*$

1: **while** $MFE(S)$ not equal to target $S^*$ **do**
2: pick neighbor $r'$ of $r$
3: accept $r'$ with probability $p$

$$p = 1 - e^{\{wd \cdot \Delta D + we \cdot \Delta E + wt \cdot \Delta T\}}$$

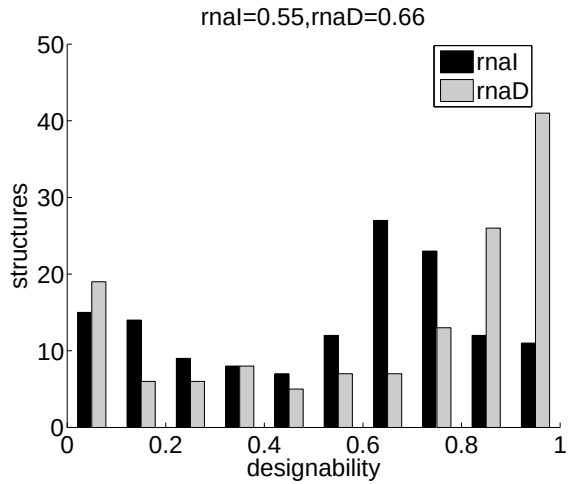
4: update sequence $r$
5: **end while**
6: **return** $r$



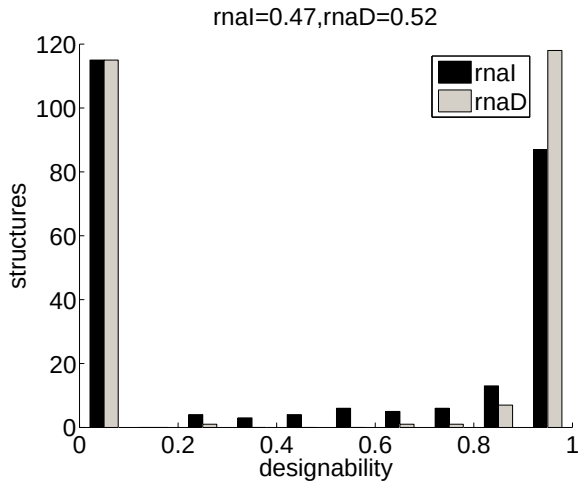




$p = 1 - e^x$

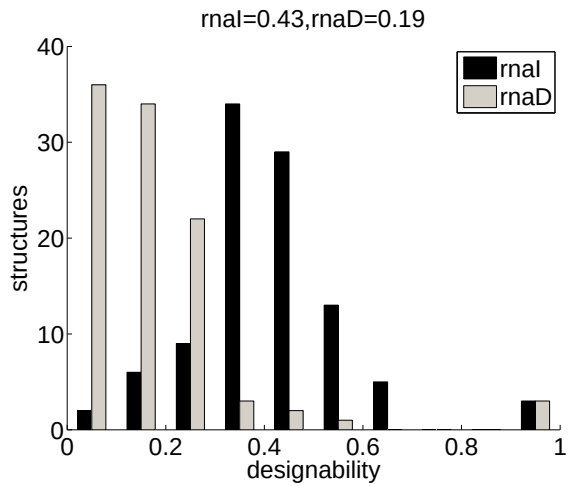
# Empirical Results



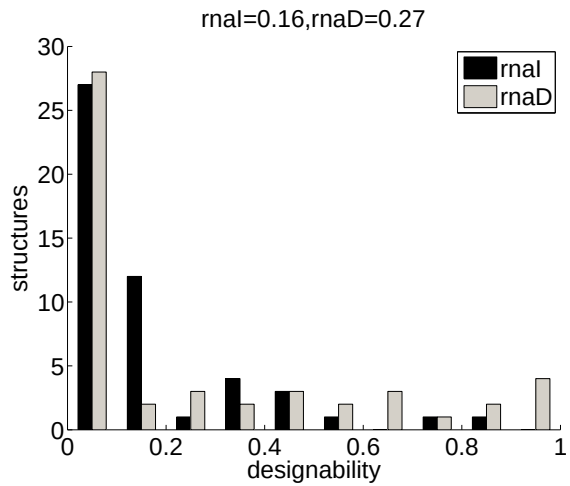

(a) TransferRNA




(b) SyntheticRNA

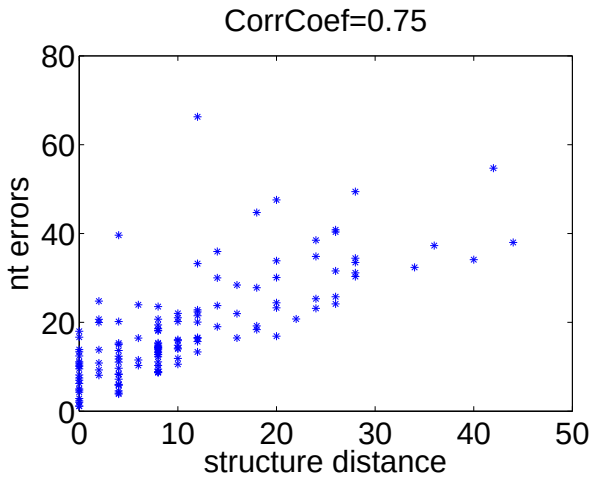



(c) RibosomalRNA

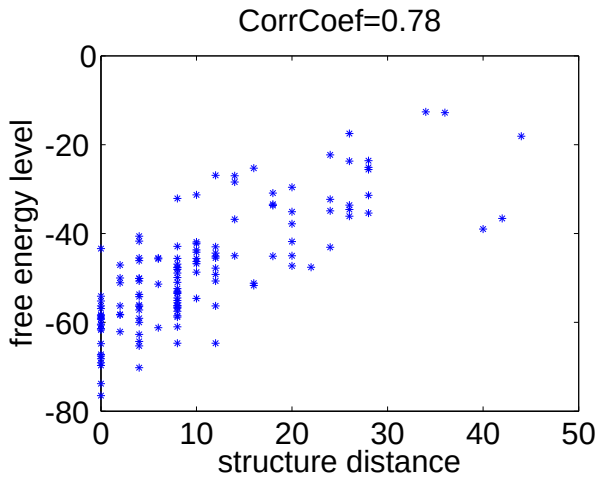



(d) RandomRNA

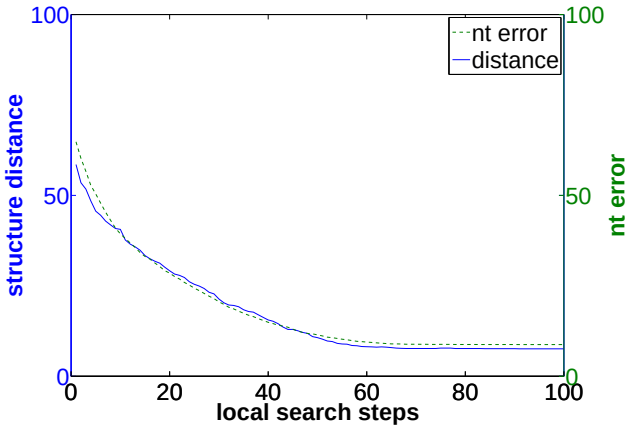
# Empirical Results



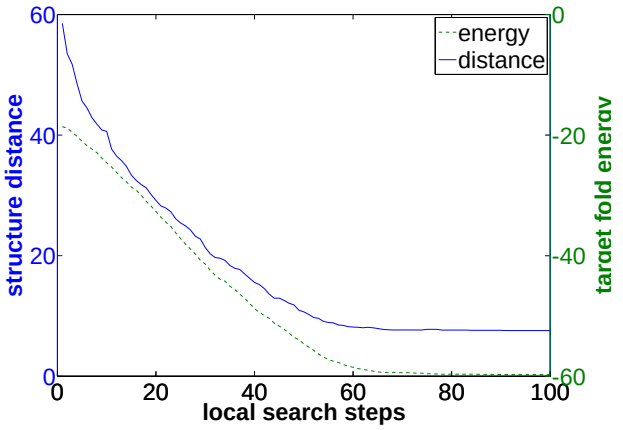

(e) nt errors & structure distance




(f) target fold energy & structure distance




(g) nt error runtime dist'




(h) target fold energy runtime dist'

# Conclusion and Discussion

- rnaDesign local search
  - outperforms RNAinverse in structure designability
  - capable of designing sequences having better thermodynamic stability
  - a combination of heuristic strategies leading to better design performance
    - smoother and less rugged combinatorial search landscape
    - performance is insensitive to initial search position (sequence)
    - less likely to be trapped in local optima
- performance issues
  - heavy-tail behavior
    - random mutation, restart scheme
  - repeated invocation of folding procedure
    - folding on partial structures
  - optimal parameter setting may vary across problem instances
    - algorithm robustness
    - online parameter tuning

# Question

- Thank you !