Fitness Functions for RNA Secondary Structure Design

Max Ward, University of Western Australia
Eliot Courtney, University of Western Australia
Elena Rivas, Harvard University

August 11, 2022

Secondary Structure Design

- We restrict ourselves to secondary structure design
- Design (or inverse folding) is basically inverting the folding function
- Folding: `GCGGAUGACCGC` $\longrightarrow$ `((((....))))`
- Design: `GCGGAUGACCGC` $\longleftarrow$ `((((....))))`

Algorithm Components

- Many published algorithms. Generally, three components:
- A *model* with associated *folding* algorithm (this is almost always just running RNAfold)
- A *fitness function* (E.g. minimizing distance to the target structure)
- A *search algorithm* (E.g. genetic algorithm)

Search Algorithms

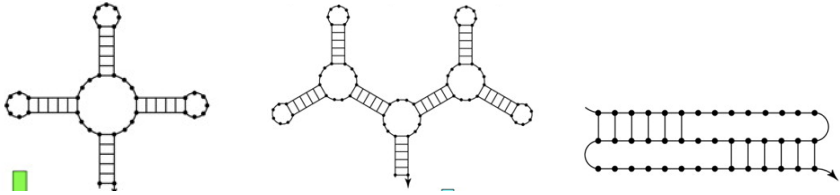
- Adaptive random walk, genetic algorithms, simulated annealing, ant colony optimization, Monte Carlo methods, constraint programming, hierarchical decomposition, crowd sourced by humans, and more

Fitness Functions

- We found four major types of fitness function
- Minimizing structure distance
- Minimizing free energy
- Maximizing ensemble probability
- Minimizing ensemble defect
- We wanted to know the strengths and weaknesses of each
- Previous work by Dirks *et al.* suggested ensemble defect was the best, but with major caveats!

Dirks *et al.*

- They found probability and ensemble defect both solved all puzzles using an adaptive random walk
- Concluded ensemble defect was better due to convergence time
- They used an "easy" testing set. 11 structures. 9 are tRNA variations, 1 larger multi-loop, 1 minimal pseudoknot



- Robert M Dirks et al. "Paradigms for computational nucleic acid design". In: *Nucleic acids research* 32.4 (2004), pp. 1392–1403

Structure Distance

- We have a target structure and a structure distance function. Minimize the distance between the predicted structure and the target structure
- Hamming distance and base pair distance appear to be the most commonly used
- How do we deal with ties? (average or minimum)
- Is MFE folding too chaotic?
- The most widely used
- RNAinverse, RNA-SSD, MODENA, SIMARD, MCTS-RNA, ...

Free Energy

- Make the free energy of the target structure given your sequence as low as possible
- Can we compare energy values from different ensembles?
- This can be solved exactly using dynamic programming. Too easy?
- INFO-RNA, “Fourier Representations for Black-Box Optimization over Categorical Variables”, ...

Probability

- Make the probability of the target structure as large as possible in the ensemble for your sequence
- Computed using McCaskill's algorithm (or Inside-outside algorithm)
- Can we compare probabilities from different ensembles?
- RNAinverse, Frankenstein, ...

Ensemble Defect

- Assume we are using a kind of Hamming distance.

$$d(s,t) = \sum_{1 \leq i \leq |s|} \begin{cases} 0 & \text{if } s_i = t_i \\ 1 & \text{otherwise} \end{cases}$$

- Ensemble defect is the ensemble probability weighted sum of these distances

$$\mathcal{D}(p,t) = \sum_{s \in S(p)} \mathbf{P}(s \mid p) \times d(s,t)$$

- Proposed by Dirks *et al.*

Ensemble Defect

- Can be computed efficiently using the base pairing probability table
- Seems to combine the best of structure distance and probability
- RNAstructure, NUPACK, ...

Synthetic Structures

- We generated 3200 structures for testing
- Create a sequence, generate all suboptimal structures within a window, sample uniformly
- "Easy" 1600 of length 40 with a 1 *kcal/mol* window
- "Hard" 1600 of length 80 with a 5 *kcal/mol* window
- All fitness functions were tested using an Adaptive Random Walk for 1000 steps

Adaptive Random Walk

function WALK(t, f, steps) ▷ t is a target structure, f is a fitness function
  $p \leftarrow$ a random sequence where $t \in S(p)$ ▷ A valid initial sequence
  loop steps iterations
    $p' \leftarrow$ MUTATE(p, t) ▷ Do a minimal random mutation that ensures $t \in S(p')$
    if $f(p') \geq f(p)$ **then** ▷ Accept any mutation that is not worse
      $p \leftarrow p'$
    end if
  end loop
  return p
end function

function MUTATE(p, t)
  $i \leftarrow$ a random index in the range $[1, |p|]$ ▷ Mutate a single nucleotide
  $p' \leftarrow p$
  $p'[i] \leftarrow$ a random nucleotide in $\{\texttt{A}, \texttt{U}, \texttt{G}, \texttt{C}\}$
  if i is paired in t **then** ▷ Fix the corresponding paired index if there is one
    $j \leftarrow$ the index paired to i in t
    $p'[j] \leftarrow$ a random valid paired nucleotide with $p'[i]$
  end if
  return p'
end function

Easy Result

Results on synthetic structures of length 40. The "# Correct" is the number of correct solutions out of 1600. The "Correct Rate" is the ratio of the number of correct solutions and 1600. The "Unique Solver" column contains the number of structures for which a fitness function was the only fitness function to find a correct solution.

Fitness Function	# Correct	Correct Rate	GC-percent	Unique Solver
Structure Distance (BPD; arbitrary tie breaking)	1269	0.79	0.50	1
Structure Distance (BPD; average tie breaking)	1419	0.89	0.50	1
Structure Distance (BPD; minimum tie breaking)	1135	0.71	0.50	0
Structure Distance (HD; arbitrary tie breaking)	1286	0.80	0.50	2
Structure Distance (HD; average tie breaking)	1436	0.90	0.50	0
Structure Distance (HD; minimum tie breaking)	1111	0.69	0.50	1
Free Energy	250	0.16	0.74	0
Probability	1554	0.97	0.52	4
Ensemble Defect	1527	0.95	0.51	2

Hard Results

Results on synthetic structures of length 80. The "# Correct" is the number of correct solutions out of 1600. The "Correct Rate" is the ratio of the number of correct solutions and 1600. The "Unique Solver" column contains the number of structures for which a fitness function was the only fitness function to find a correct solution.

Fitness Function	# Correct	Correct Rate	GC-percent	Unique Solver
Structure Distance (BPD; arbitrary tie breaking)	351	0.22	0.50	0
Structure Distance (BPD; average tie breaking)	454	0.28	0.50	1
Structure Distance (BPD; minimum tie breaking)	273	0.17	0.49	0
Structure Distance (HD; arbitrary tie breaking)	370	0.23	0.50	0
Structure Distance (HD; average tie breaking)	482	0.30	0.50	0
Structure Distance (HD; minimum tie breaking)	267	0.17	0.50	2
Free Energy	17	0.01	0.77	0
Probability	1201	0.75	0.59	112
Ensemble Defect	945	0.59	0.58	10

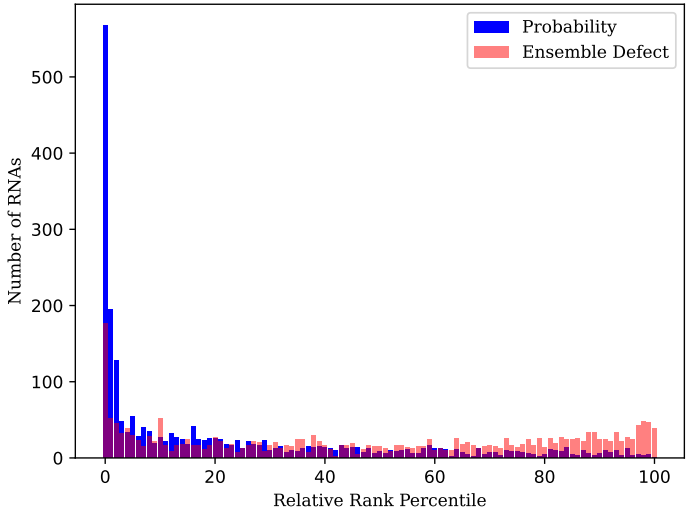
Real RNAs

- We also tested on natural sequences with known, conserved secondary structures
- The ArchiveII data set containing 3948 sequence structures
- First, we generated at least 200 000 suboptimal structures for each sequence (surprisingly difficult to do)
- If the true structure is not in the 200 000 we pick the closest by base pair distance
- If the distance is more than 5%, we remove the sequence
- 1719 sequences were removed

Real RNAs

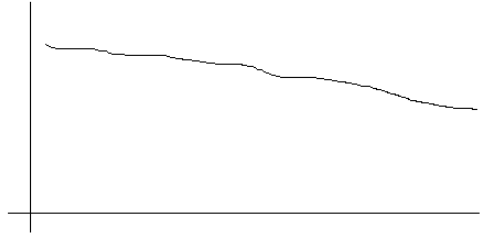
- Each of our 200 000 structures are from the same sequence/ensemble
- We can compute the probability for each, and rank them by this
- Similarly, we can rank by ensemble defect
- If a fitness function is good, the true structure should be ranked highly
- This is a bit weird. Instead of comparing different sequences we're comparing different structures

Real RNA


Probability
Ensemble Defect
Number of RNAs
0
100
200
300
400
500
0
20
40
60
80
100
Relative Rank Percentile

Why is ED Failing?

- The distribution of ensemble defects is flat



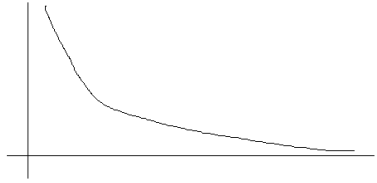
- The min ensemble defect structure is often a compromise

Why is ED Failing?

```
((((((((((...(((.......)))...))))......((((..((((....))))...))))...))))))....
    Rank=0,     Probability=0.10, Ensemble Defect=0.42
(((((((......(((.......))).((((((.....))).))))...(((((.......)))))))))))....
    Rank=1,     Probability=0.04, Ensemble Defect=0.33
(((((((..((((.........)))).(((((.......)))))).....(((((.......)))))))))))....
    Rank=14,    Probability=0.008, Ensemble Defect=0.40
(((((((......(...........).((((((.....))).))))....((((.......)))).)))))))....
    Rank=57934, Probability=2.8e-7, Ensemble Defect=0.32
```

Why is Probability Good?

- I only have guesses
- You can compare probabilities from different distributions because almost all probability distributions have roughly the same shape



- It's hard to increase the probability of a structure without increasing the probability of similar structures
- It's hard to increase the probability of a structure without decreasing the probability of dissimilar structures

Paper

Max Ward, Eliot Courtney, and Elena Rivas. “Fitness Functions for RNA Structure Design”. In: *bioRxiv* (Under Revision). DOI: 10.1101/2022.06.16.496369