

# Non-redundant sampling for locally optimal\* structures

Juraj Michalik   Yann Ponty   Helene Touzet

Equipe AMIB  
Laboratoire LIX - Inria Saclay

Ecole Doctorale Interfaces

February 14, 2017

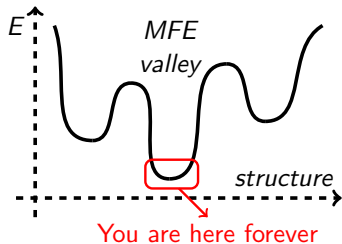
The Inria logo is written in a red, cursive script.The logo for the University of Paris-Saclay, featuring the text "université" in a purple serif font and "PARIS-SACLAY" in a purple sans-serif font below it, with a small purple dot above the "é".The AMIB logo consists of the letters "AMIB" in a large, bold, blue sans-serif font with a slight 3D effect.


# Overview

- 1 Introduction
- 2 Concepts
- 3 Decomposition of Nussinov local minima
- 4 Non-redundant sampling
- 5 Results
- 6 Conclusion

# Introduction

- RNAs - structural diversity, usually important at a functional level
- Thermodynamic equilibrium (McCaskill, 1990): partition function  
→ base-pairing probabilities within Boltzmann ensemble



-  Equilibrium assumption not always valid:
  - ▶ **Riboswitches**: 2 conformations with significant  $\Delta G$ , **both** active **yet** difference unmitigated by sole presence/absence of ligand.
  - ▶ **Co-transcriptional folding** would **not** happen at equilibrium!
- Also, RNA degrades quickly - MFE frequently not achieved




**Importance of kinetic effects in formation of RNA structure**



**Study RNA folding kinetics**

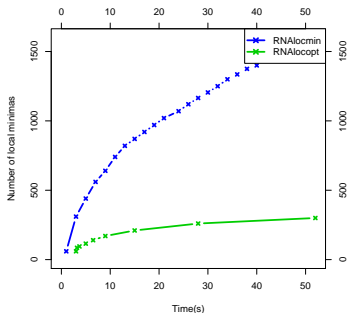
# RNA kinetics study

## RNA kinetics analysis methods - 2 classes:

- *Simulation methods (statistical)* - simulates RNA folding base by base/helix by helix  
→ #trajectories required for reproducibility increases fast
- *4-step plan (approximative)*:
  - ▶ **Sampling** of representative set of structures
  - ▶ **Assembling** of representation of RNA folding landscape from samples
  - ▶ **Estimation** of transition rates between different parts of folding landscape representation
  - ▶ **Investigation**, notably evolution of concentrations during time
-  sampling quality is essential, following steps depend on it:
  - ▶ Missing **functional** structure  Losing part of RNA folding space
  - ▶ Missing **transitive** structure  Energy barrier overestimation

# Diversity is problematic

- Suboptimal structures (Wuchty *et al.*, 1999)  
➔ Combinatorial explosion
- Stochastic Sampling (Ding and Lawrence, 2003): Saturation  
➔ High redundancy



(Kucharik *et al.*, 2014)

- Most of sampling strategies:  $P(\text{sample}) \propto e^{-\frac{E}{RT}}$
- Problem: oversampling of structures close to MFE

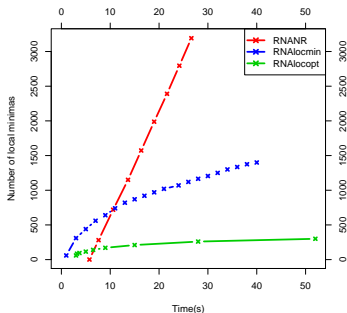
To overcome this problem:



**Non-redundant sampling**

# Diversity is problematic

- Suboptimal structures (Wuchty *et al.*, 1999)  
➔ Combinatorial explosion
- Stochastic Sampling (Ding and Lawrence, 2003): Saturation  
➔ High redundancy



(Kucharik *et al.*, 2014)

- Most of sampling strategies:  $P(\text{sample}) \propto e^{-\frac{E}{RT}}$
- Problem: oversampling of structures close to MFE

To overcome this problem:



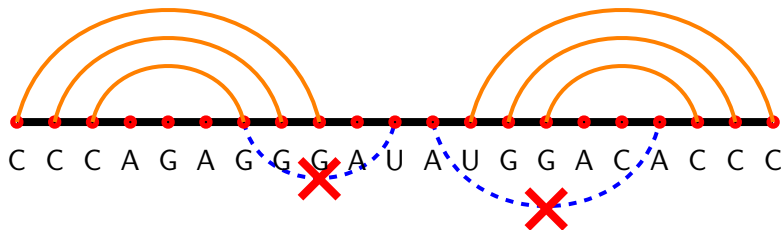
**Non-redundant sampling**

# Concepts

## Secondary structure (in this context):

Set of base pairs within an RNA sequence with following restrictions

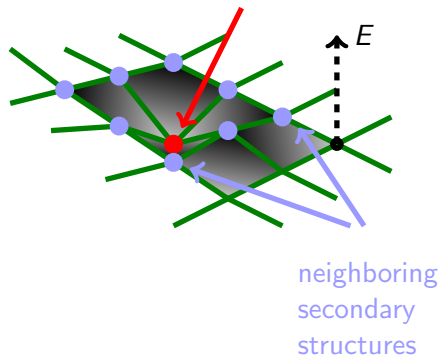
- Only pairs  $\in \{\{C, G\}, \{A, U\}, \{G, U\}\}$  permitted
- No base triplets
- No pseudoknots



Orange and blue paths cannot coexist within the same structure

# Locally optimal secondary structures

Local Minimum (LM) in RNA folding space



## Local Minima (LM)

- Minimal free energy **within neighborhood**
- **Neighbors of structure:** All structures obtained by single base pair addition/removal

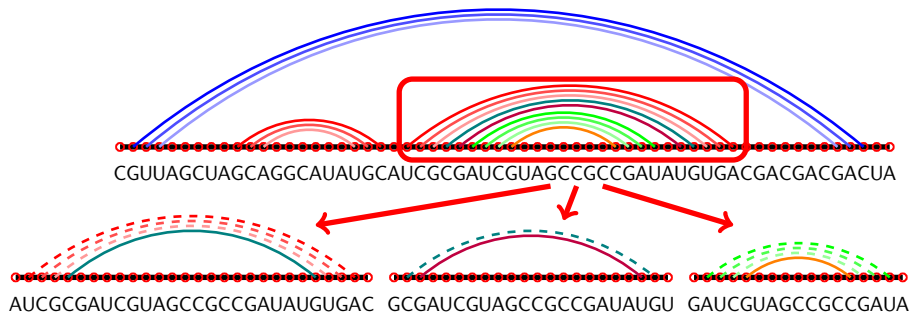
Energy model: Base pair maximization: RNANR  $\rightarrow$  Nussinov LMs...  
... **but also** w.r.t. Turner model: RNAlocopt (WA. Lorenz *et al.*, 2011),  
RNAlocmin (Kucharik *et al.*, 2014)  $\rightarrow$  Turner LMs



# Flat structures

 Beyond this point: min. helix length = 3 & stems of length 3 considered together

Nussinov model: Decomposition of secondary structures into **flat structures**, i.e. maximal by juxtaposition (Saffarian *et al.*, 2012):



# Decomposition of local minima

## Central idea:

- Generate **all flat structures** for RNA sequence (Saffarian *et al.*, 2012)
- Find free energy  $E_f$  of each flat structure  $f$  ( $\approx$  loops in Turner model) ... based on new interface to Vienna RNA package <sup>1</sup>
- **Combine** flat structures in any possible ways to obtain complete **Nussinov local minima** (while keeping track of free-energy)

Local optimality ensured by **saturation** of all flat structures



Cannot add any new base pair without creating a conflict

## How to sample Nussinov Local Minima?

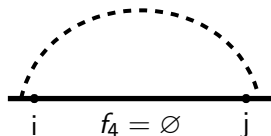
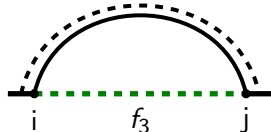
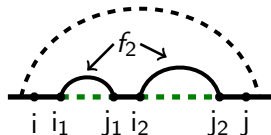
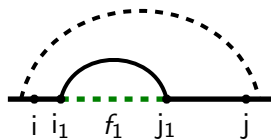
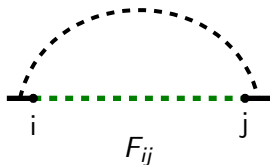
---

<sup>1</sup>Thanks Ronny!

# Dynamic programming scheme for flat structure assembly

$F$  = Set of flat structures

$$Z_{i,j} = \sum_{f \in F_{ij}} e^{-E_f/RT} \begin{cases} \prod_{[k,l] \in f} Z_{k,l} & \text{if } f \neq \emptyset \\ 1 & \text{if } f = \emptyset \end{cases}$$



Memory complexity:  $\mathcal{O}(n^2)$   
 Time complexity:  $\mathcal{O}(|F| \cdot n)$   
 $\in \mathcal{O}(|F| \cdot \max(\#MLbranches))$

[...]

# Statistical Sampling of Nussinov LMs

## Partition function

$$\mathcal{Z} = \sum_{s \in \mathcal{S}} e^{\frac{-E_s}{k_B T}}$$

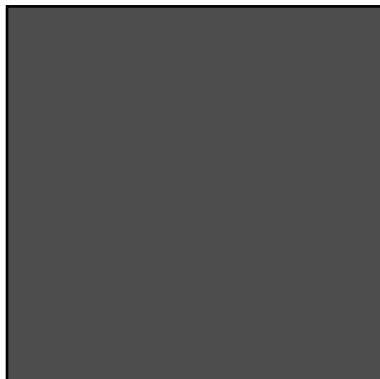
$\mathcal{S}$  = space of secondary structures  $s$

$E_s$  = energy of specific state  $s$

$k_B$  = Arbitrary constant

$T$  = Absolute temperature

Here,  $\mathcal{S}$  = Nussinov LMs secondary structures under structural restrictions



Space  $\mathcal{S}$  of secondary structures of interest  $s$

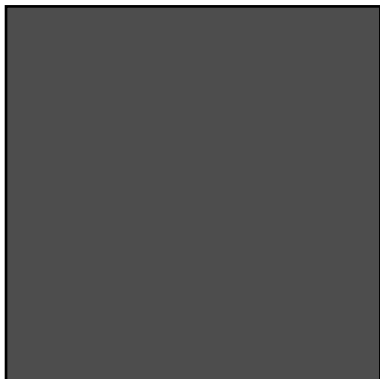
# Statistical Sampling of Nussinov LMs



Secondary structure  $s$  of RNA sequence

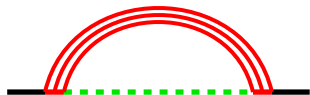
$$\mathcal{Z}_s = \mathcal{Z}$$

$$P(s) = 1$$

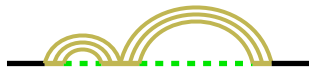


Space  $\mathcal{S}$  of secondary structures of interest  $s$

# Statistical Sampling of Nussinov LMs



Secondary structure  $a$  of RNA sequence



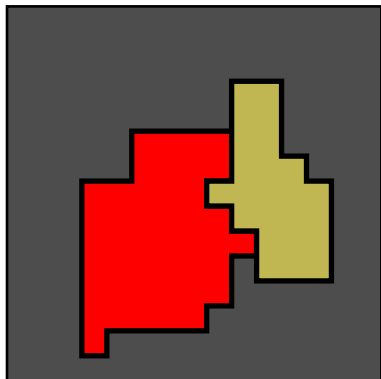
Secondary structure  $a_1$  of RNA sequence

$$P(a) = \frac{z_a}{z_s} = \frac{z_a}{z}$$

$$P(a_1) = \frac{z_{a_1}}{z_s} = \frac{z_{a_1}}{z}$$

$$\mathcal{A} \cap \mathcal{A}_1 = \emptyset$$

$$P(a) + P(a_1) < 1$$



Space  $\mathcal{A}$  of secondary structures  $a$   
Space  $\mathcal{A}_1$  of secondary structures  $a_1$   
 $\mathcal{A} \subset \mathcal{S}, \mathcal{A}_1 \subset \mathcal{S}$

# Statistical Sampling of Nussinov LMs

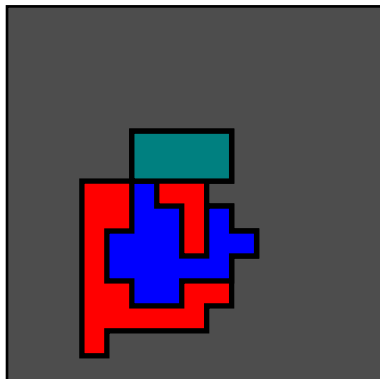


Secondary structure  $b$  of RNA sequence



Secondary structure  $b_1$  of RNA sequence

$$P(b|a) = \frac{z_b}{z_a}, P(b_1|a) = \frac{z_{b_1}}{z_a}$$
$$B \cap B_1 = \emptyset, P(b) + P(b_1) < P(a)$$
$$P(b) = P(b|a) \cdot P(a) = \frac{z_b}{z_a} \cdot \frac{z_a}{z} = \frac{z_b}{z}$$



Space  $B$  of secondary structures  $b$   
Space  $B_1$  of secondary structures  $b_1$   
 $B \subset \mathcal{A} \subset \mathcal{S}, B_1 \subset \mathcal{A} \subset \mathcal{S}$

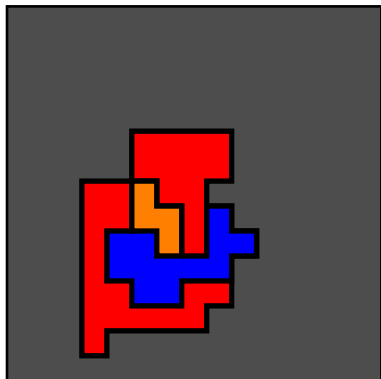
# Statistical Sampling of Nussinov LMs



Secondary structure  $c$  of RNA sequence

$$P(c|b) = \frac{Z_b}{Z_c}$$


$$P(c) = P(c|b).P(b|a).P(a) = \frac{Z_c}{Z}$$

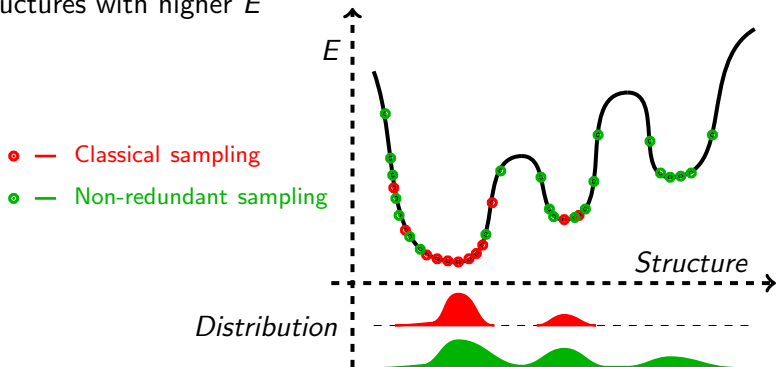


Space  $\mathcal{C}$  of secondary structures  $c$   
 $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A} \subset \mathcal{S}$



# Non-redundant sampling

- Each structure picked up **at most once**  Faster access to structures with higher  $E$

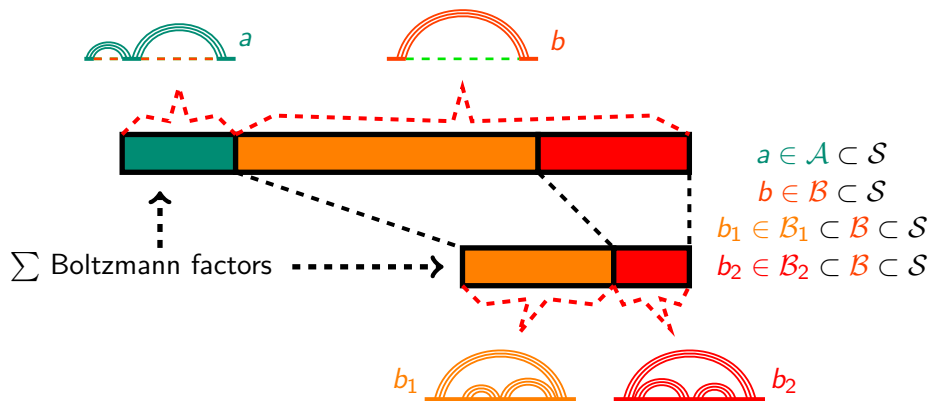


- Problem:** Avoid choosing sample after first selection

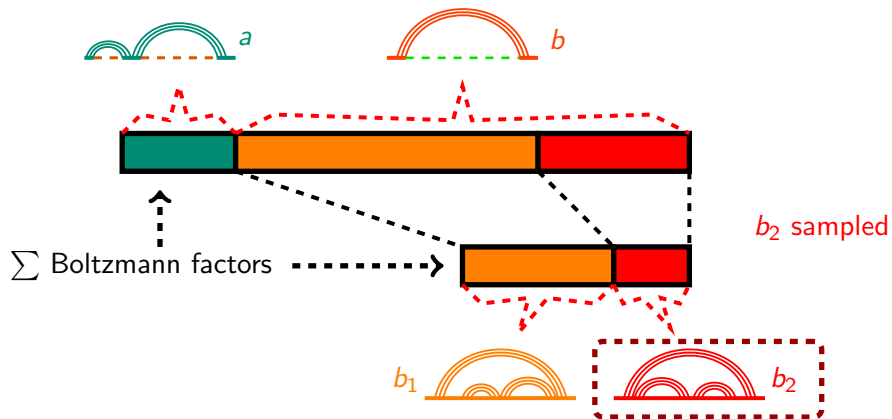
**Solution :** After generation of a given structure  $S$ , adjust probabilities of flat structures depending on their capacity to generate  $S$  again.

**So, how to adjust the probabilities?**

# Non-redundant sampling

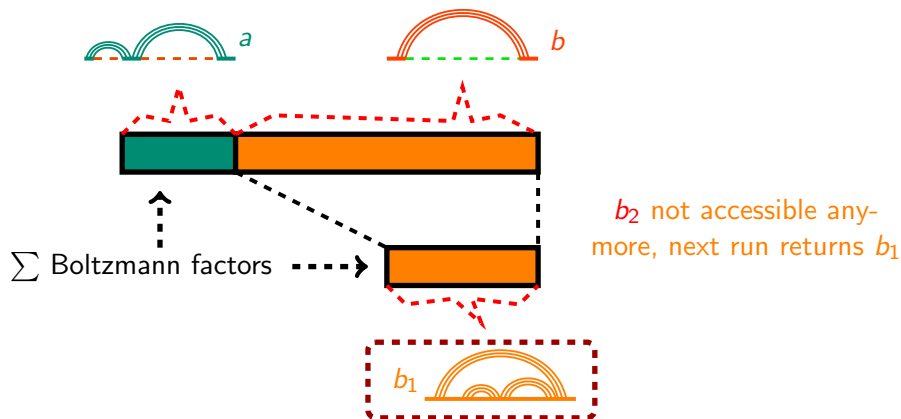


# Non-redundant sampling





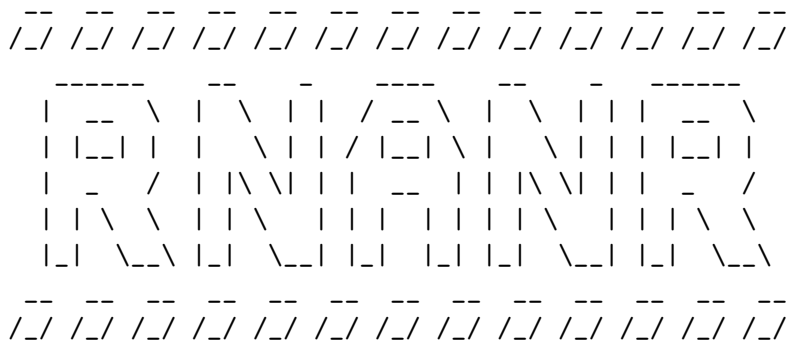
# Non-redundant sampling



Efficient access to the probabilities of generated LMs through dedicated data structure (no complexity overhead... details on demand)

# Results

## Implementation – RNANR



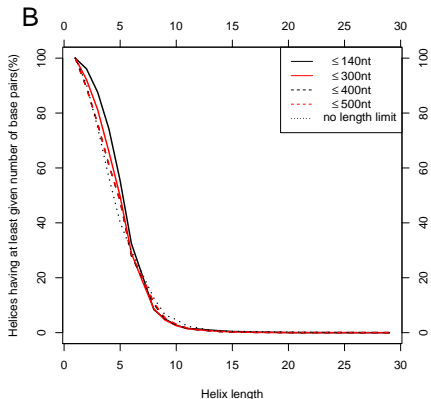
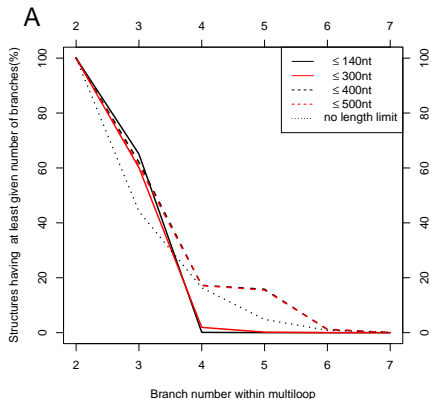
- **C implementation**, based on Vienna package's RNAlib
- Non-redundant sampling, exhaustive enumeration, counting, expressive structural restrictions
- **Availability:**

<https://project.inria.fr/rnaland/software/rnanr/>

# Results

## Structural Restrictions

- Space reduction using structural restrictions → complexity reduction!
- ⚠ Minimum helix length  $\alpha$ , max #branches within multiloop  $\gamma$
- ⚠ **Reminder:** min helix length = 3
- Statistics on **RNAstrand** (Andronescu *et al.*, 2006)



# Exhaustive LMs enumeration

Test on SV11

- SV11 has active metastable (MS) state at 28.5 kcal.mol<sup>-1</sup> of MFE
- MS-like conformations unreachable for sampling algorithms
- Currently, numerical precision issues with non-redundant sampling  
→ Exhaustive enumeration in restricted folding space
- Structural restrictions: min. helix length = 4, max #branches within multiloop = 4

## Results:

GGGCACCCCCUUCGGGGGUCACCCUCGCGUAGCUAGCUAGCGAGGGUUAAAGGGCCUUCUCCUCGCGUAGCUAACCAACGCGAGGUGACCCCCGAAAAGGGGGUUUCCCA

(((((.....((((((((((((((((((((.....((((((((((((((((.....)))))))))))))))).....)))))))))))))))).....)))))))).)))).



Real structure

Returned structure

MFE



# Comparison of Nussinov and Turner Local Minima

**Method:** Sampling Nussinov LMs + Gradient descent<sup>2</sup>

→ Final structure, ie Turner Local Minimum

	Samples% avg (std.dev)	$\Delta\Delta G$ avg (std.dev)	Base pair dist. avg (std.dev)
Within search space	59.57% (21.00)	0.071 (0.309)	0.129 (0.289)
Outside search space	40.42% (21.00)	1.248 (0.925)	1.550 (0.619)
Global average	100.00% (-)	<b>0.547</b> (0.817)	<b>0.703</b> (0.757)

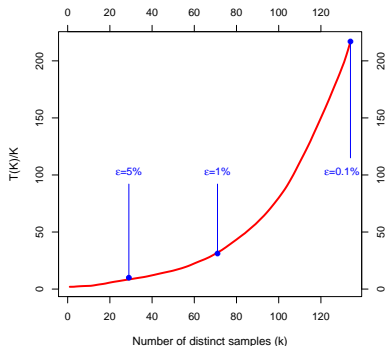
- More than half **Nussinov LMs** (52.4%) are also **Turner LMs**
- On average, a **Nussinov LM** is at  $\leq 0.55 \text{kcal.mol}^{-1}$  and 0.7 base pairs to its closest Turner
- When the final structure is in the search space, **Nussinov LMs** are almost always Turner LMs ( $\approx 90\%$ )

<sup>2</sup>Vienna package – Thanks Ronny and Gregor!

# Theoretical speedup

$T(K)$ : #redundant structures to obtain  $K$  #unique structures

Speed-up:  $T(K)/K = \text{Avg \#times a structure is (redundantly) sampled}$



*Expected number of duplicitious samples per unique structure (A)*

•  $S_s \subset \mathcal{S} = \text{set of obtained samples}$

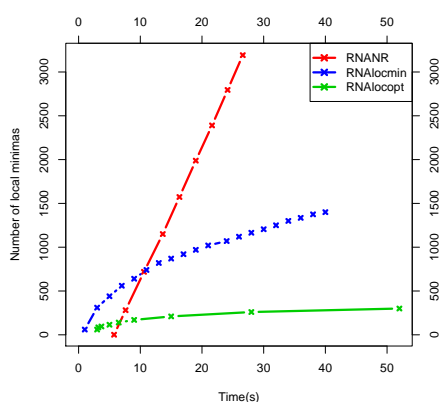
• Coverage = 
$$\frac{\sum_{s \in S_s} e^{-\frac{E_s}{kT}}}{Z}$$

• **Speedup**

$$= 1 + \frac{\sum_{i=0}^K \left( 1 - \sum_{j=1}^i \frac{e^{-\frac{E_j}{kT}}}{Z} \right)^{-1}}{K}$$

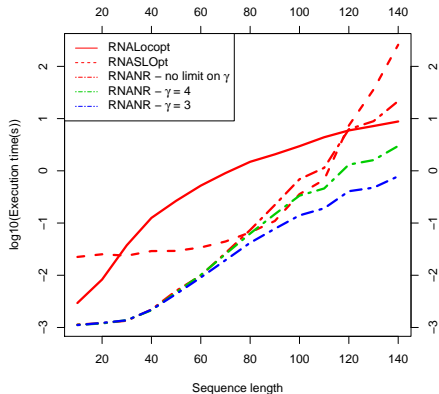
• Bigger speedup with higher coverage

# Practical speedup and complexity



*Number of LMs returned*

- No redundancy = faster coverage



*Comparison of speed of different software*

- Limiting #branches within multiloop reduces complexity

# Conclusion

## New features

- Considerable speed up for the exploration of RNA folding landscapes
- Expressive structural restriction without added cost

## Philosophical speedbump

- Exponential vs polynomial
- Non-redundant sampling can be easily implemented to any already existing sampling method
- Non-redundant sampling for statistical estimates: Does losing redundancy mean losing information?

## In progress ... **LOADING**

- Numerical stability issues when Boltzmann factors become too low
- Validation of our local minima for kinetics analysis
- Non-redundant sampling for Turner model,  $\chi$  scheduling. . .

