Evolutionary algorithm for inverse pseudoknotted RNA folding inspired by Lévy flights

Computational Approaches to RNA structure and function

Cyrille Merleau Nono Saha and Matteo Smerlak

August 11, 2022

Max Planck Institute for Mathematics in the sciences
Structure of Evolution Group.
Nono S.C. Merleau et al. BMC Bioinformatic 20022

Presentation plan

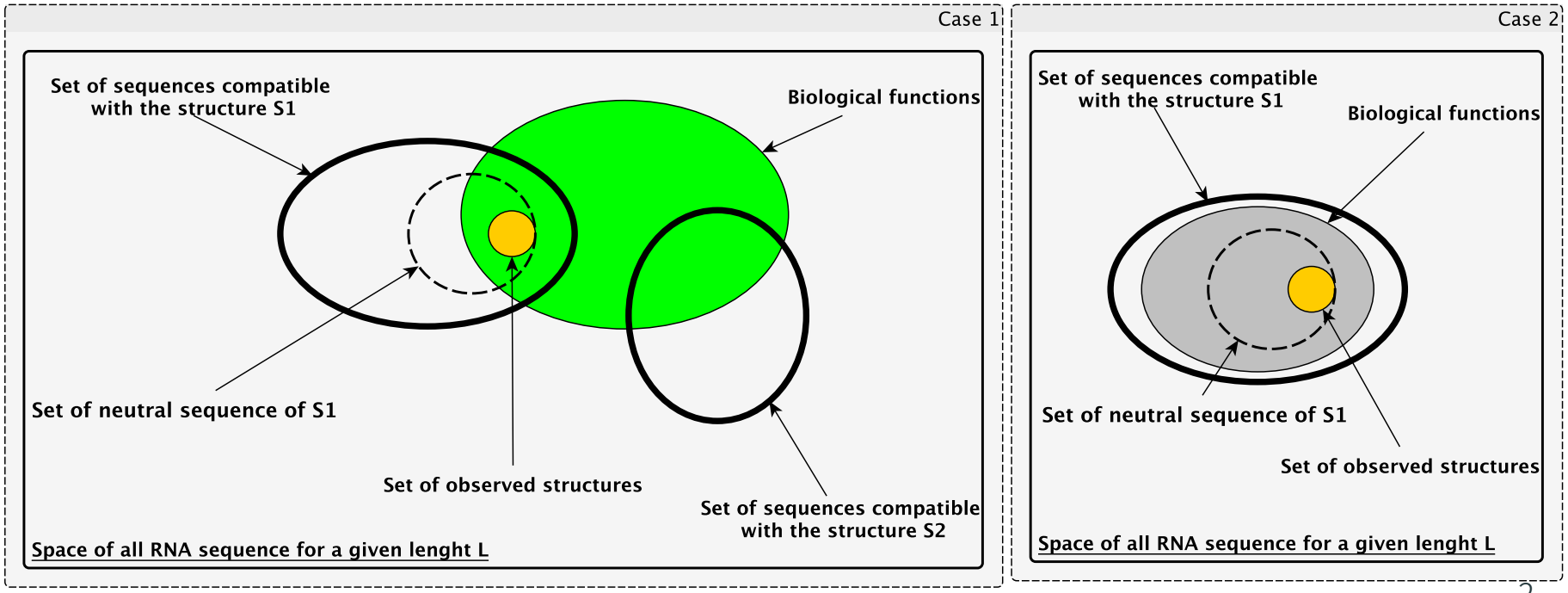
1. RNA design paradigm
2. Optimization problem
3. Evolutionary algorithm (EA)
4. Mutation operator
5. Experimental results
6. Discussion and conclusion

RNA design paradigm

RNA design paradigm

RNA Sequence ⟵ Secondary structure ⟵ Function

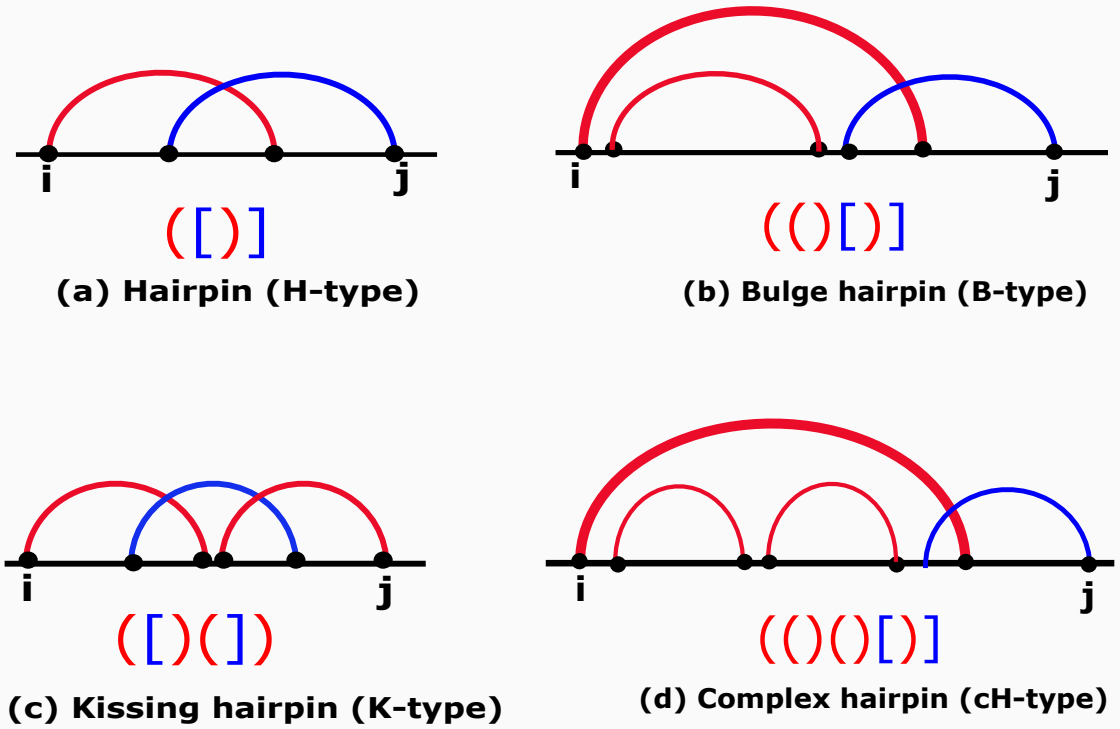
- Only a small fraction of structural function are observable.
- Different structures may imply the same functions.
- Not all sequences with the same structure share the same function.
- The perfect imaginary case 2.



RNA design and optimization problem

Input: An RNA secondary structure with pseudoknot pattern

Output: One or many RNA sequences with a certain optimum property



Objective functions:

- Distance to the target structure (PD).
- Energy distance
- Ensemble defect (ND)
- probability defect
- suboptimal defect
- Multi-objective function

aRNAque's algorithm approach

EA for RNA Inverse folding: general principle

Initial Population of RNA sequences $I \subset \mathcal{C}$

Genetic Operations:

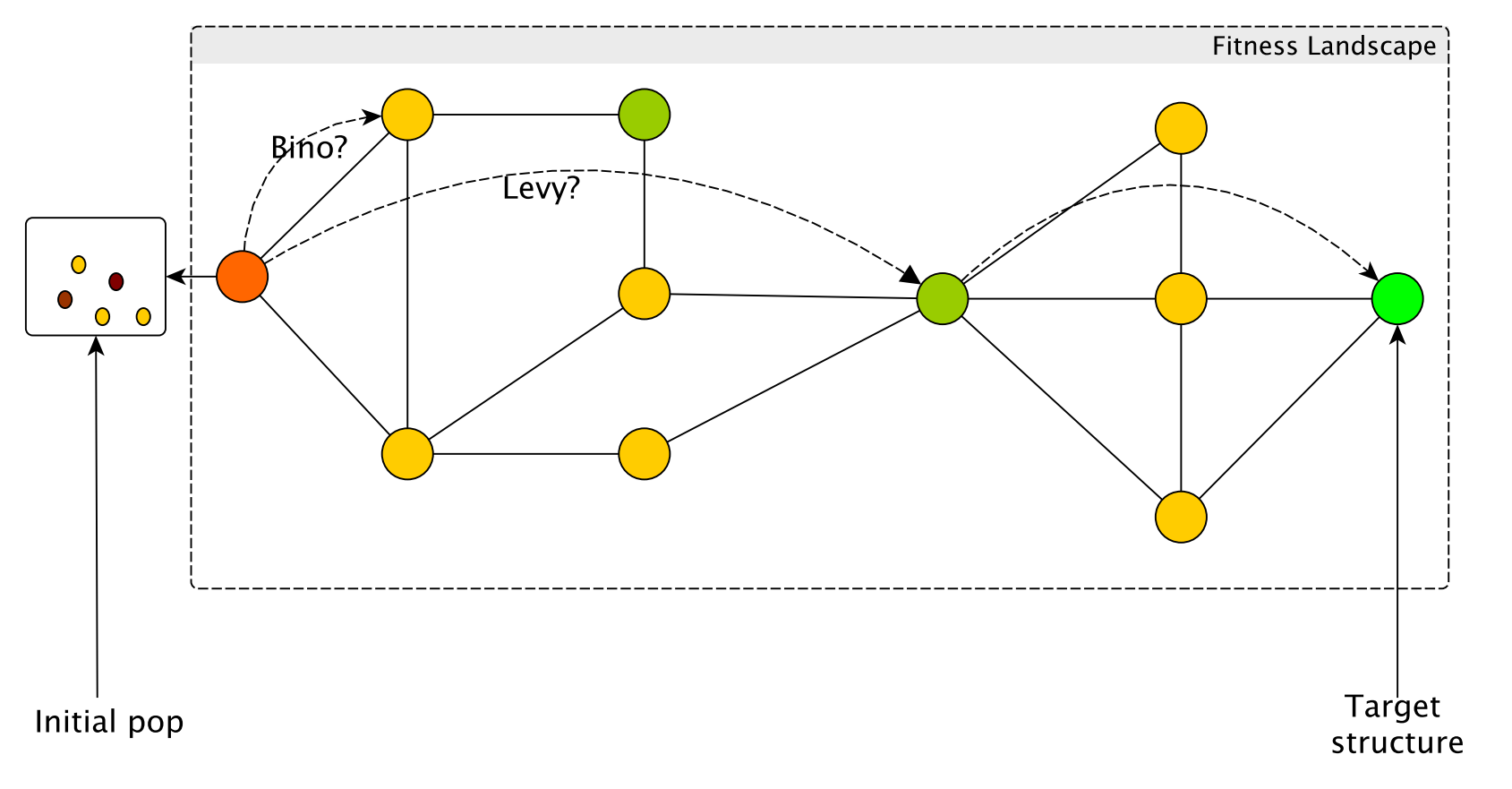
1. Selection
2. Mutation
3. Crossover
4. Elitism

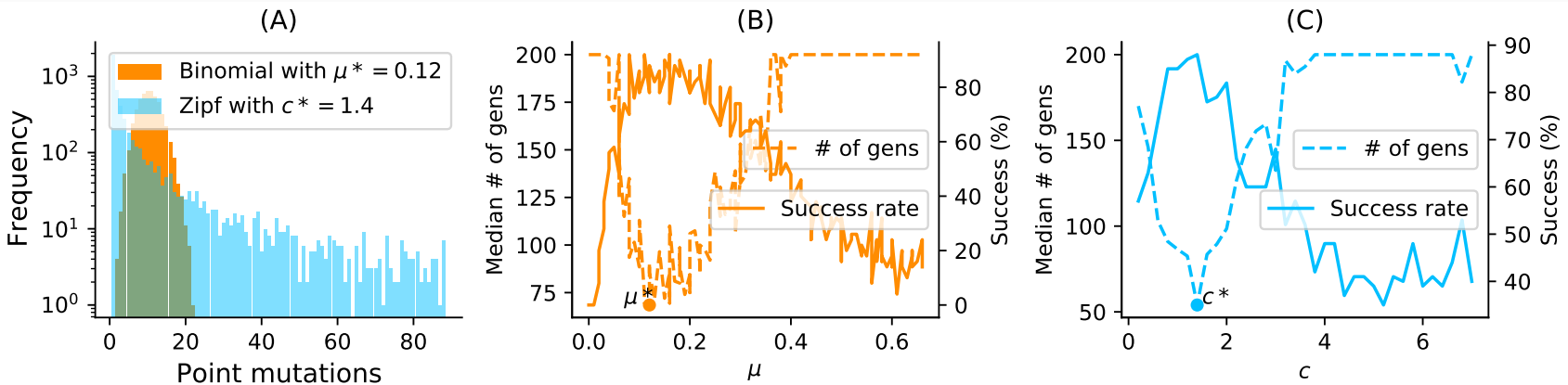
RNA sequences that fold to the target structure $\{s \in \mathcal{C} \mid \Sigma(s) = \sigma*\}$

Target RNA secondary structure $\sigma* \in \Sigma_L$

- MODENA [*Akito Taneda*, 2011]
- ERD [*Ali Esmaili-Taheri et al.*, 2015]

The proposed algorithm: Lévy flights and EA





The proposed algorithm: Lévy mutation scheme

- $P = \{S_1 \ldots S_n\}$
- $P_C = \{w_{AU}, w_{GU}, w_{GC}\}$
- $P_N = \{w_A, w_U, w_C, w_G\}$
- $\mathcal{D}(c, L) = \frac{1/n^c}{\sum_{k=1}^{L} 1/k^c}$
- P_C, P_N: tunable parameters
- B_i and $U_i \sim \mathcal{D}$

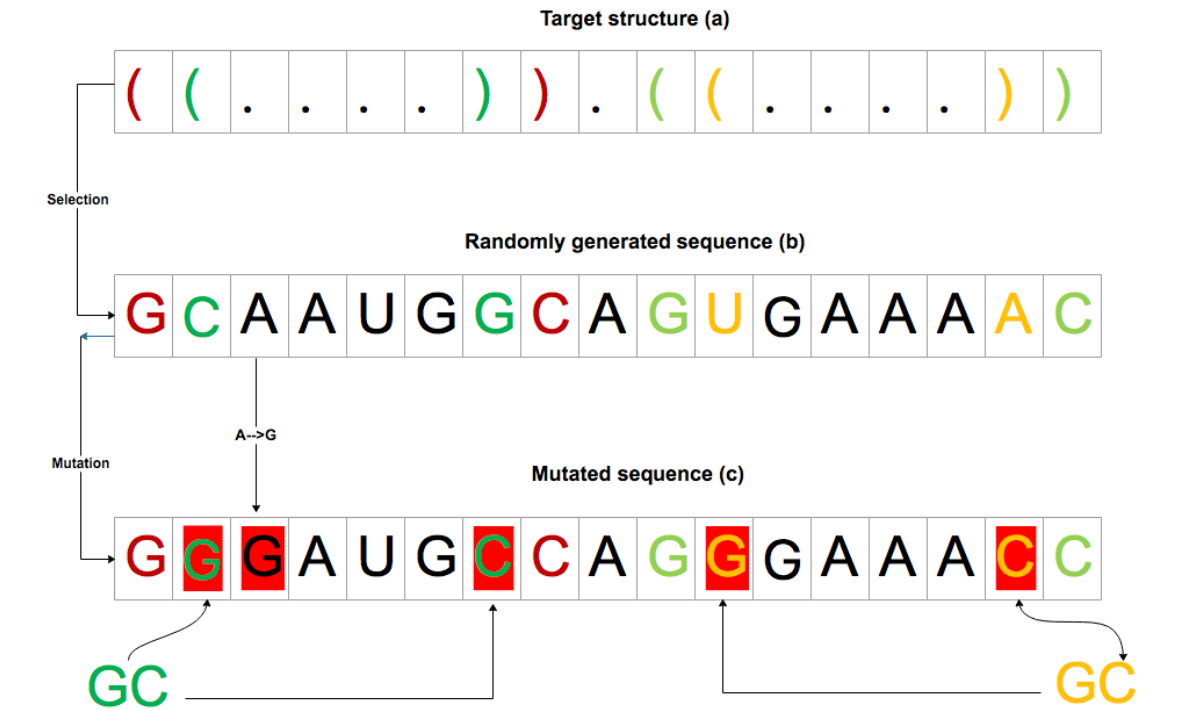


Figure 2: Illustration of the mutation algorithm

Experimental results

Experimental results: benchmark protocols

Benchmark dataset

- The `PseudoBase++`: a set of 265 pseudoknotted RNA structures.

Folding tools

- `IPknot`: no energy prediction.
- `Hotknots`: free energy predictions

Termination criteria

- 0 distance to the target structure
- Number of generations (Max 200 *gen.*)

Performed tasks

- Levy mutation *vs.* Local search (One-point mutation)
- `antaRNA` *vs.* `aRNAque` (EA+Lévy mutation)

Experimental results: Lévy search *vs* Local search in EA

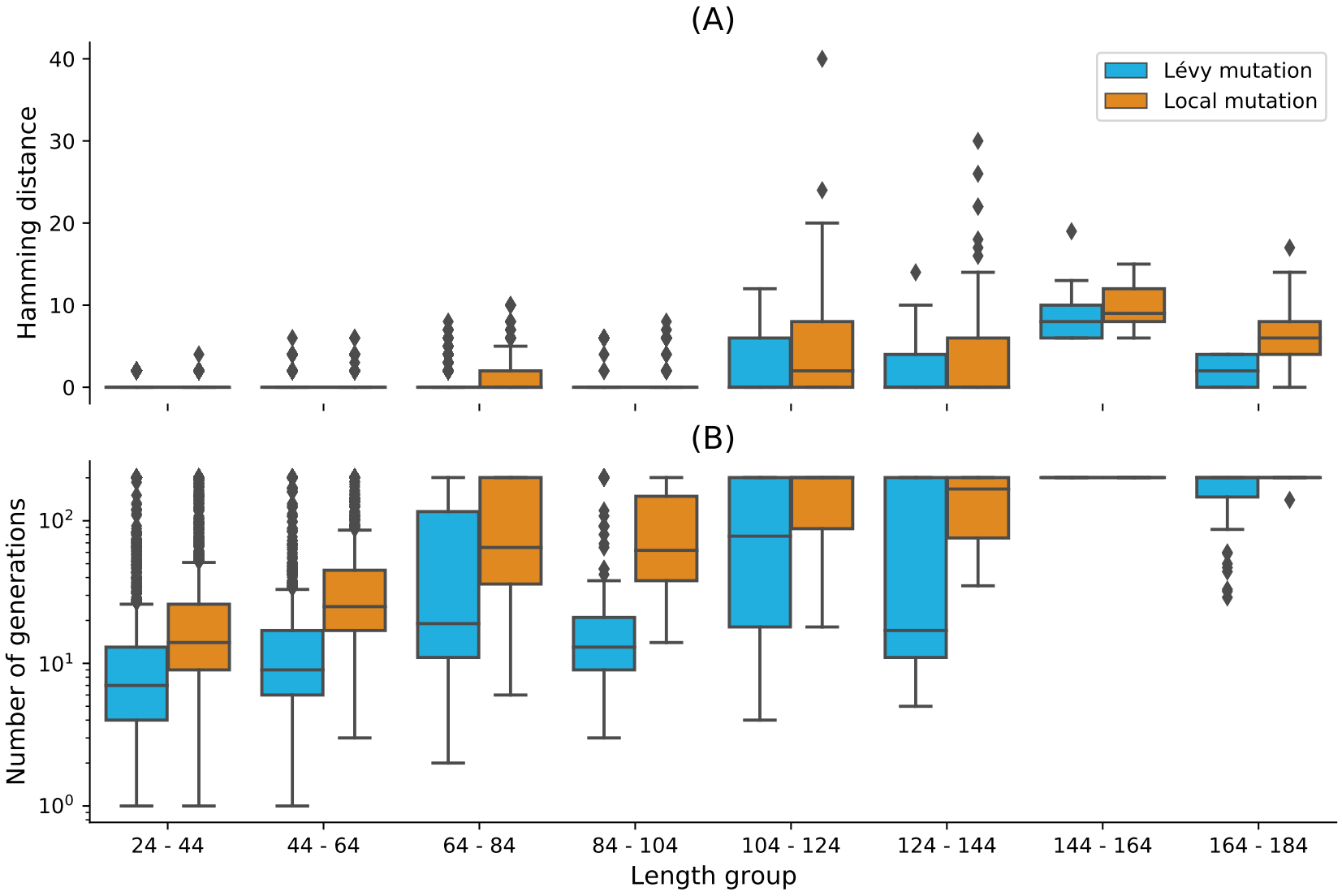


Figure 3: aRNAque: Lévy mutation *vs.* One-point mutation ($p-$value ≈ 0.16, $p-$value≈ 0.0004)

antaRNA *vs.* aRNAque: performance per pseudoknot types

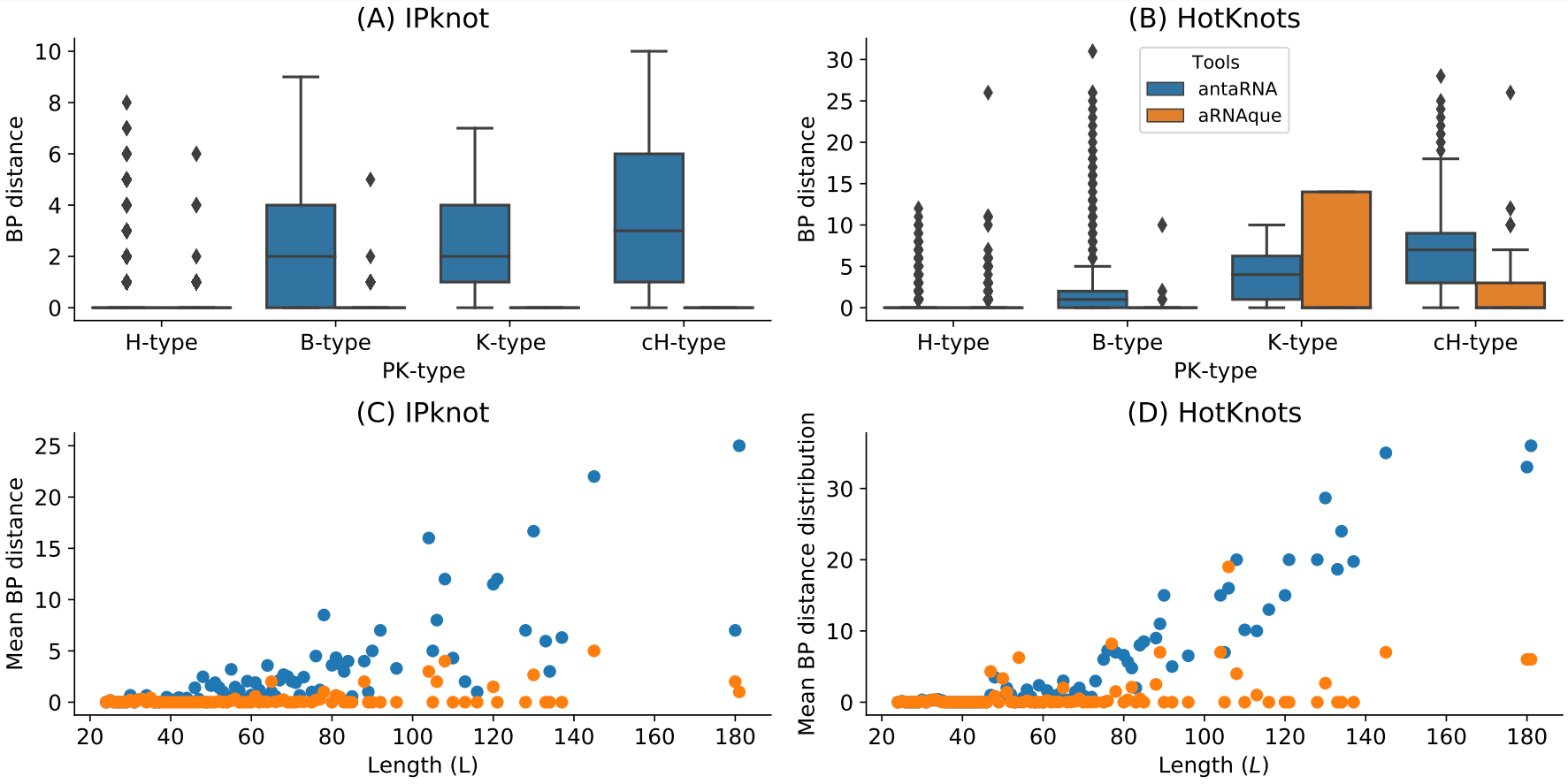


Figure 4: aRNAque *vs.* antaRNA

antaRNA *vs.* aRNAque: GC-content control

- Measures the concentration of G-C nucleotide in S and influences its stability and biological function [Wang et *al.*, 2014, Science].
- $tGC = 0.25$; $P_C = \{0.125, 0.125, 0.3, 0.3, 0.075, 0.075\}$; $P_N = \{0.125, 0.125, 0.375, 0.375\}$

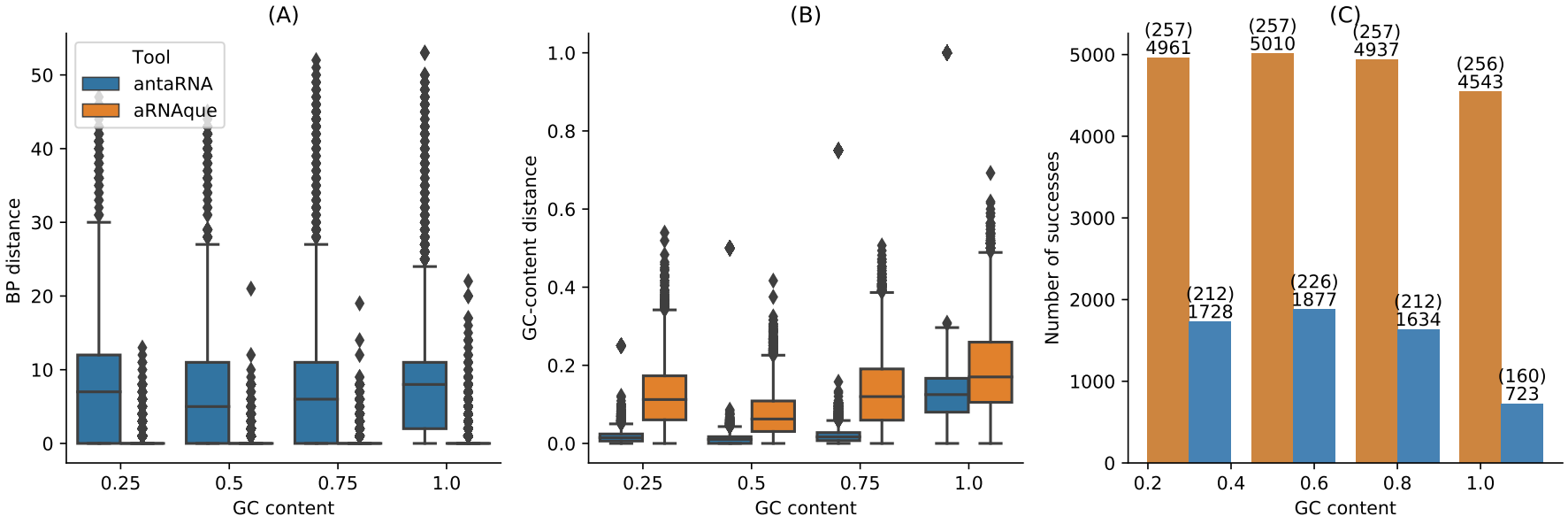


Figure 5: aRNAque *vs.* antaRNA

antaRNA *vs.* aRNAque: diversity of the designed sequences

- Less GC-content control
- High diversity

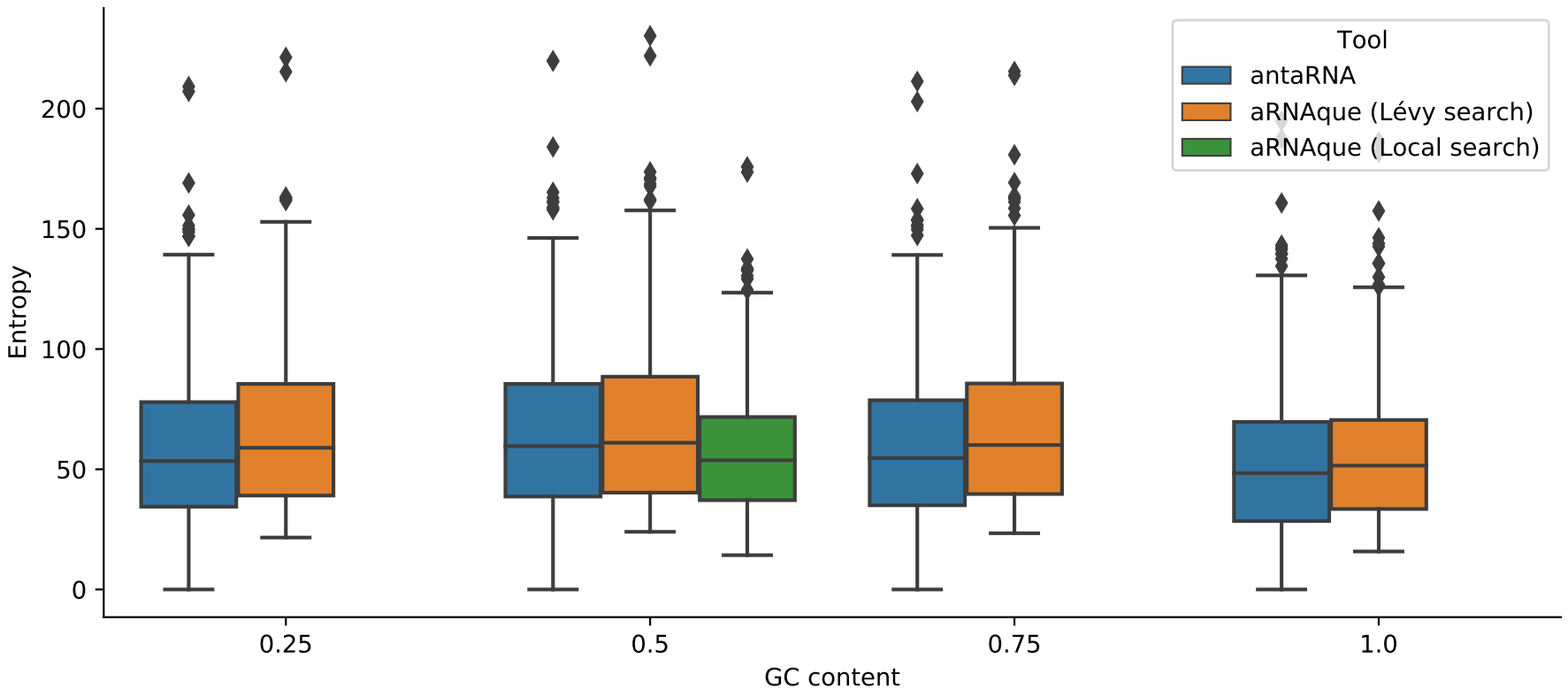


Figure 6: aRNAque *vs.* antaRNA

Conclusion and discussion

Why Lévy mutation could probably work better? (1)

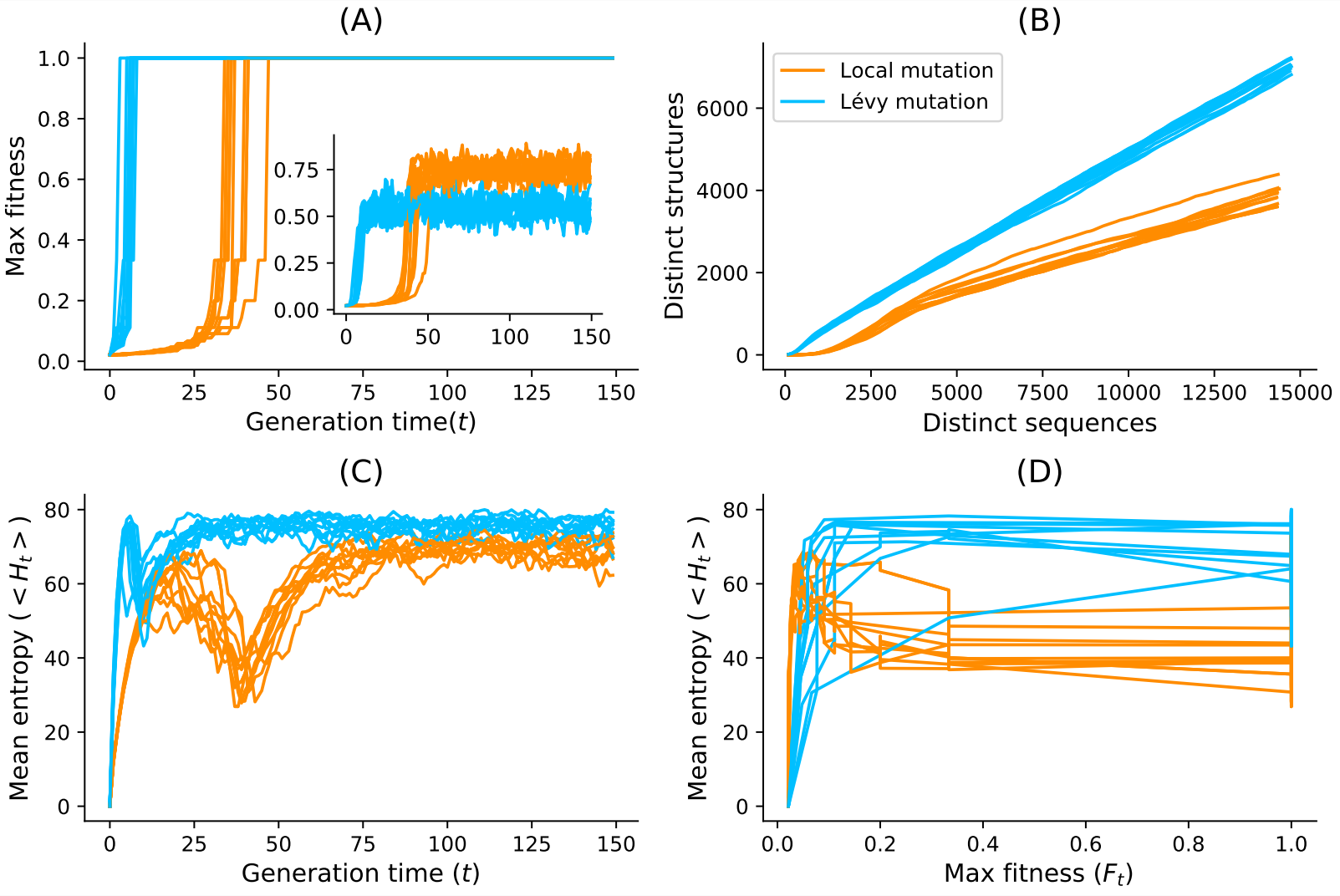
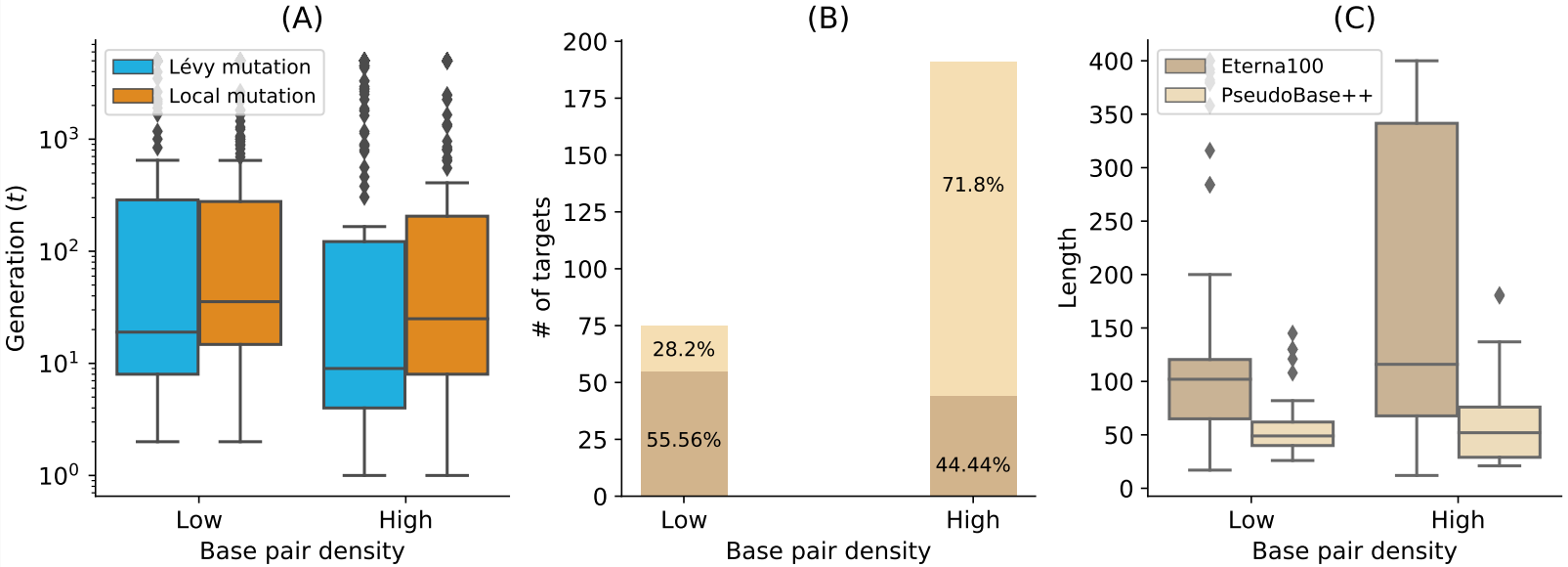
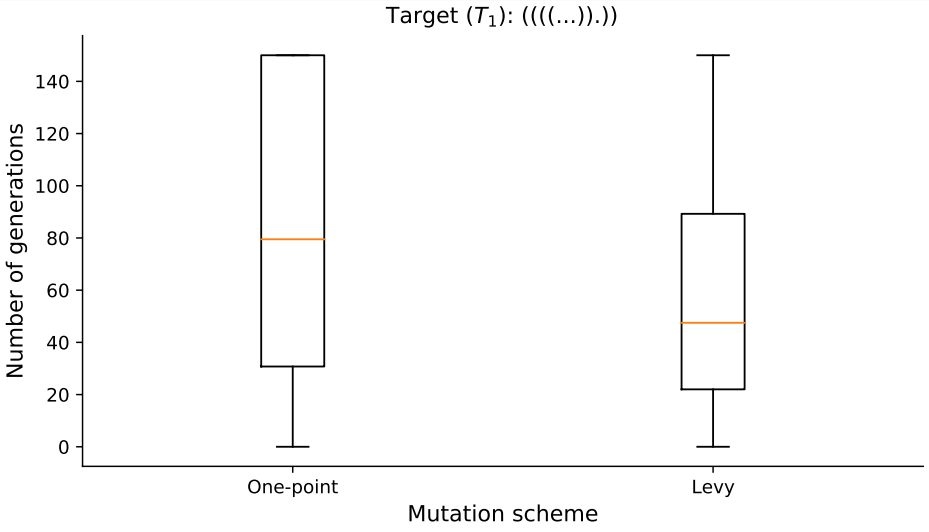


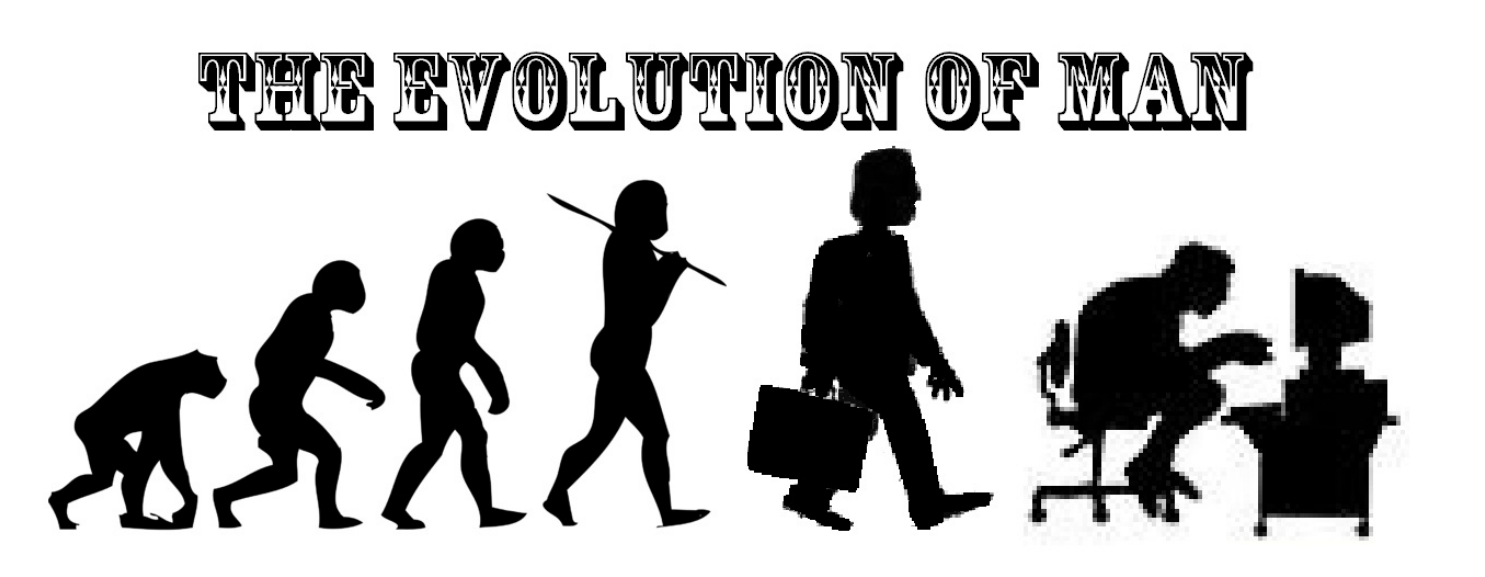
Figure 7: aRNAque: Lévy mutation *vs.* Local mutation

Why Lévy mutation could probably work better? (2)



- ((((...)))). : 243 sequences
- (((....))).. : 249 sequences
- ((((...)).)): 02 sequences
- ((.((...)))): 01 sequences





Thanks for your attention