Journées MAGNUM – Partie 1

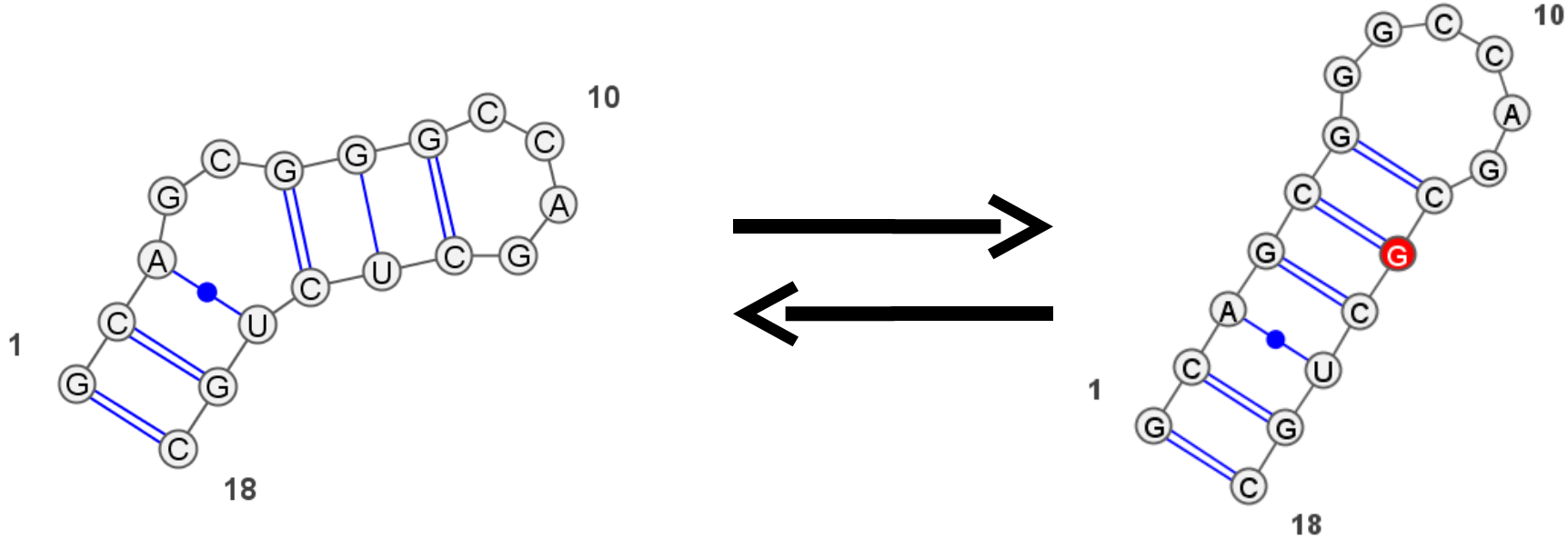
An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure

Jérôme Waldispühl, PhD
School of Computer Science,
McGill Centre for Bioinformatics,
McGill University, Canada

Yann Ponty, PhD
Laboratoire d'informatique (LIX),
École Polytechnique, France

Overview

RNAmutants: Algorithms to explore the RNA mutational landscape



Understanding how mutations influence RNA secondary structures **AND** how structures influence mutations (Waldispühl et al., PLoS Comp Bio, 2008).

Sampling k-mutants

Seed

```
CAGUGAUUGCAGUGCGAUGC   (-1.20)
..((.(((((...)))))))
```

Classic: 0 mutation

```
CAGUGAUUGCAGUGCGAUcC   (-3.40)
..(.((((((...))))))
CAGUGAUUGCAGUGCGgUGC   (-0.30)
((.((....)).))......
CAGUGAUcGCAGUGCGAUGC   (-3.10)
.....(((((...)))))..
```

RNAmutants: 1 mutation

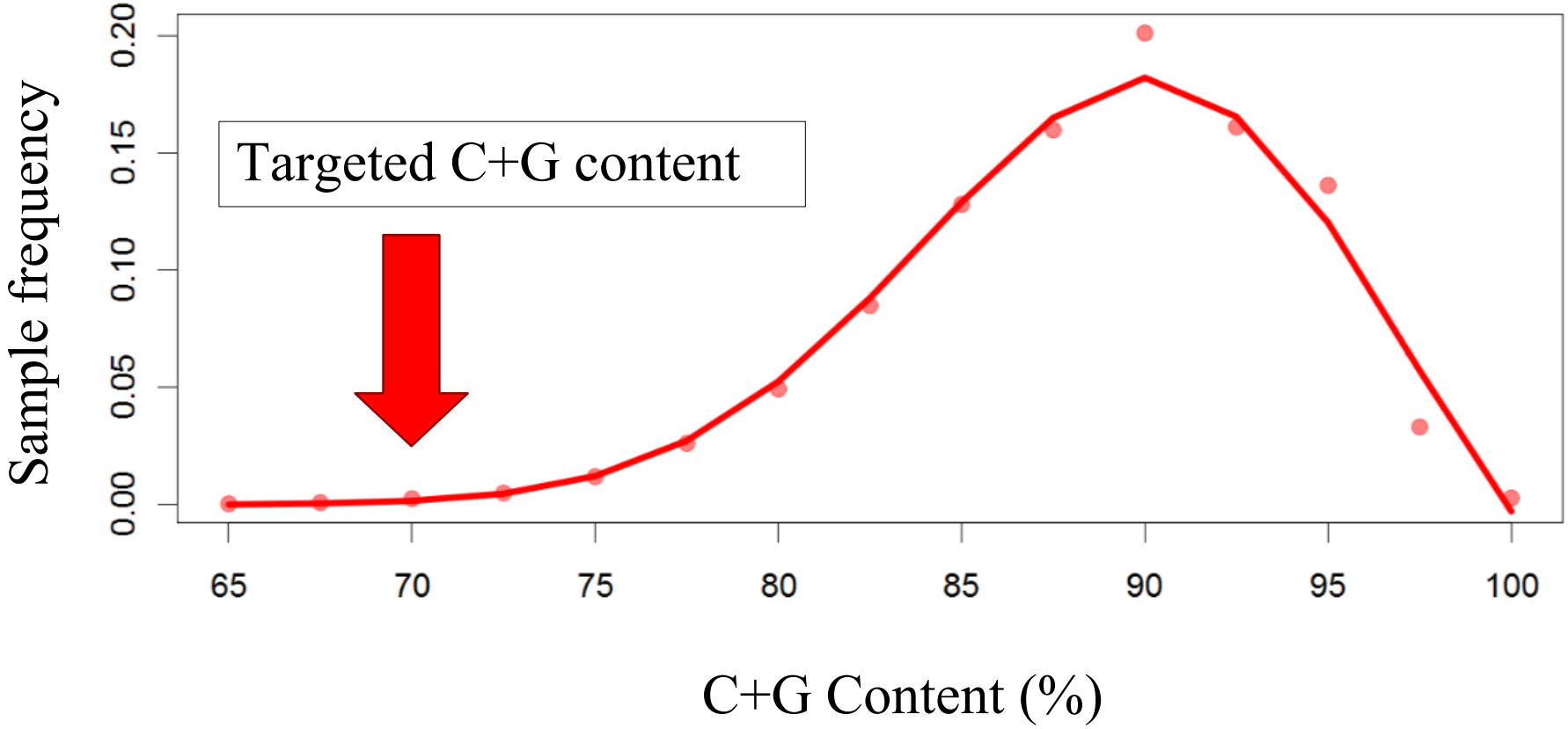
```
uAGcGccgGgAGacCGgcGC   (-18.00)
..(((((((....)))))))
CccUGgccGCAagGCcAgGg   (-20.40)
(((((((((....)))))))
CcGUGgccGCgagGCcAcGg   (-19.10)
(((((((((....)))))))
```

RNAmutants: 10 mutations

Sample k mutations increasing the folding energy

Consequence: Increase in C+G content

Objectives



How to efficiently sample sequences at arbitrary C+G contents … without introducing a bias!

Outline

- Background: RNAmutants in a nutshell
  Algorithms to sample RNA secondary structures and mutations.

- Our approach: Adaptive sampling
  Uniformly shifting the distribution of samples.

- Results: Evolutionary studies
  Insights on the evolutionary pressure stemming from an optimization of the thermodynamical stability.

Outline

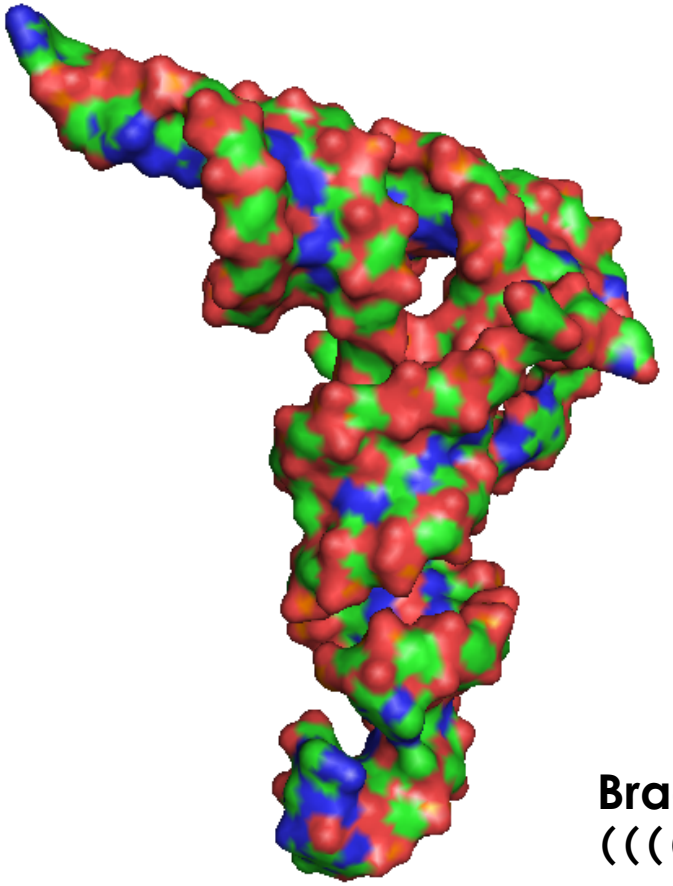
- **Background: RNAmutants in a nutshell**
  Algorithms to sample RNA secondary structures and mutations.

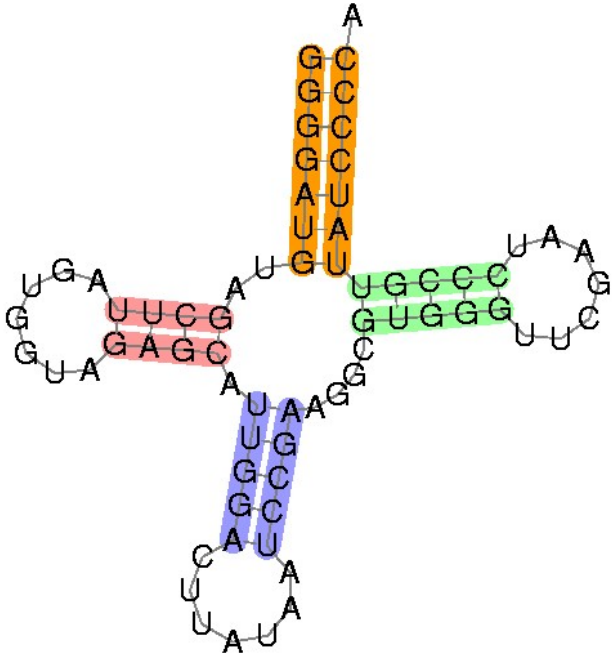
- Our approach: Adaptive sampling
  Uniformly shifting the distribution of samples.

- Results: Evolutionary studies
  Insights on the evolutionary pressure stemming from an optimization of the thermodynamical stability.

RNA secondary structure

The **secondary structure** is the ensemble of base-pairs in the structure.

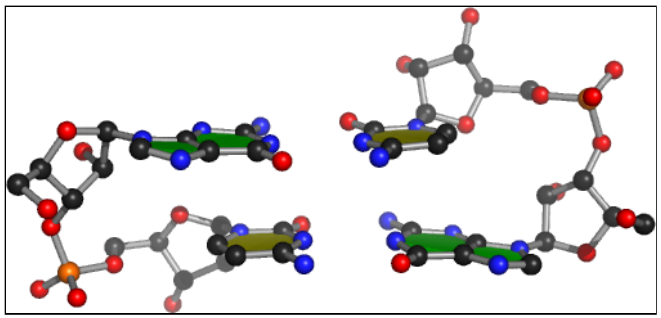




Bracket notation:

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

Loop decomposition



Stacking pairs

Hairpin loop

Bulge loop

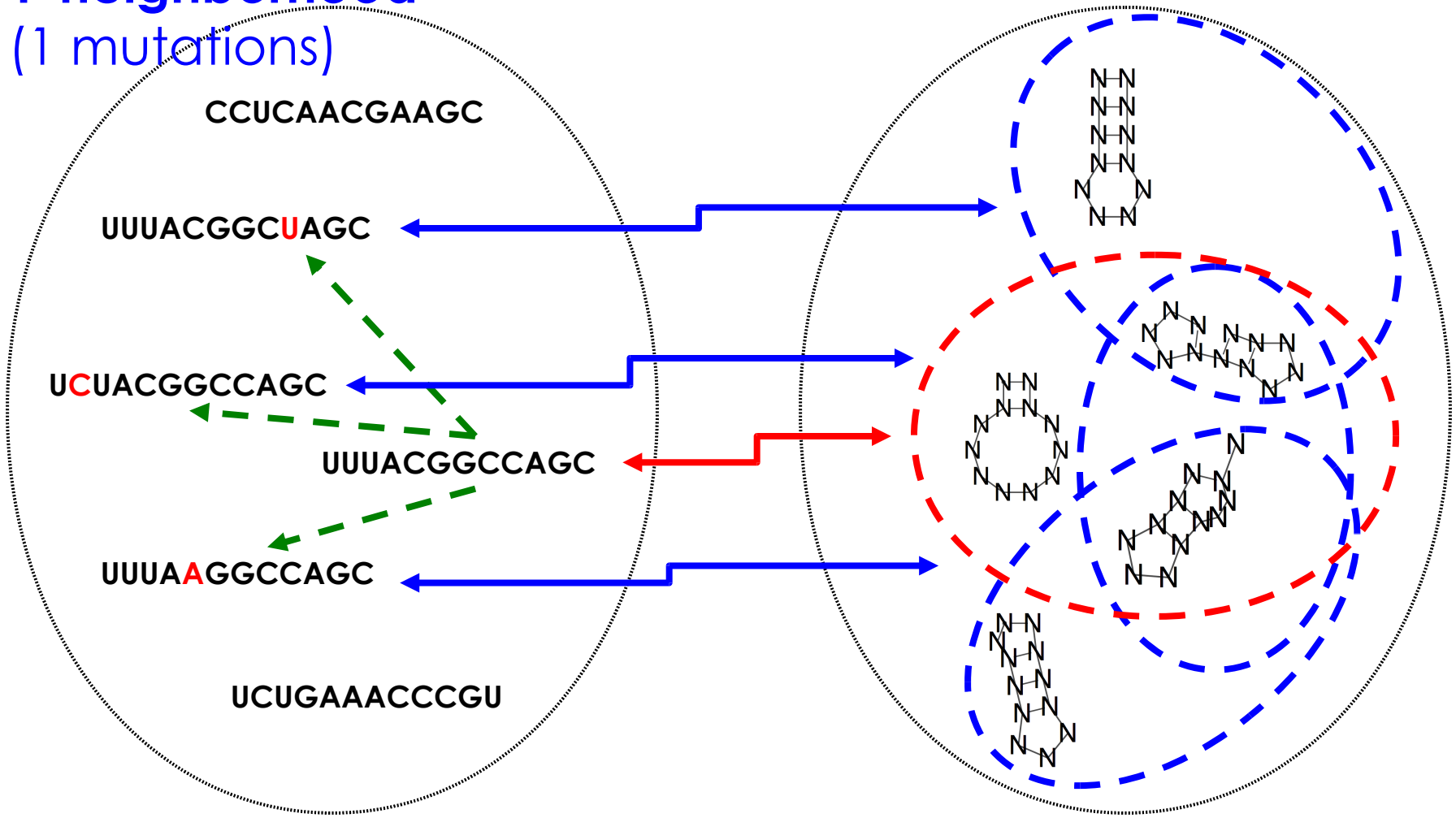
Interior loop

Multibranched loop

Unconstrained single-stranded region

Parameterization of the mutational landscape

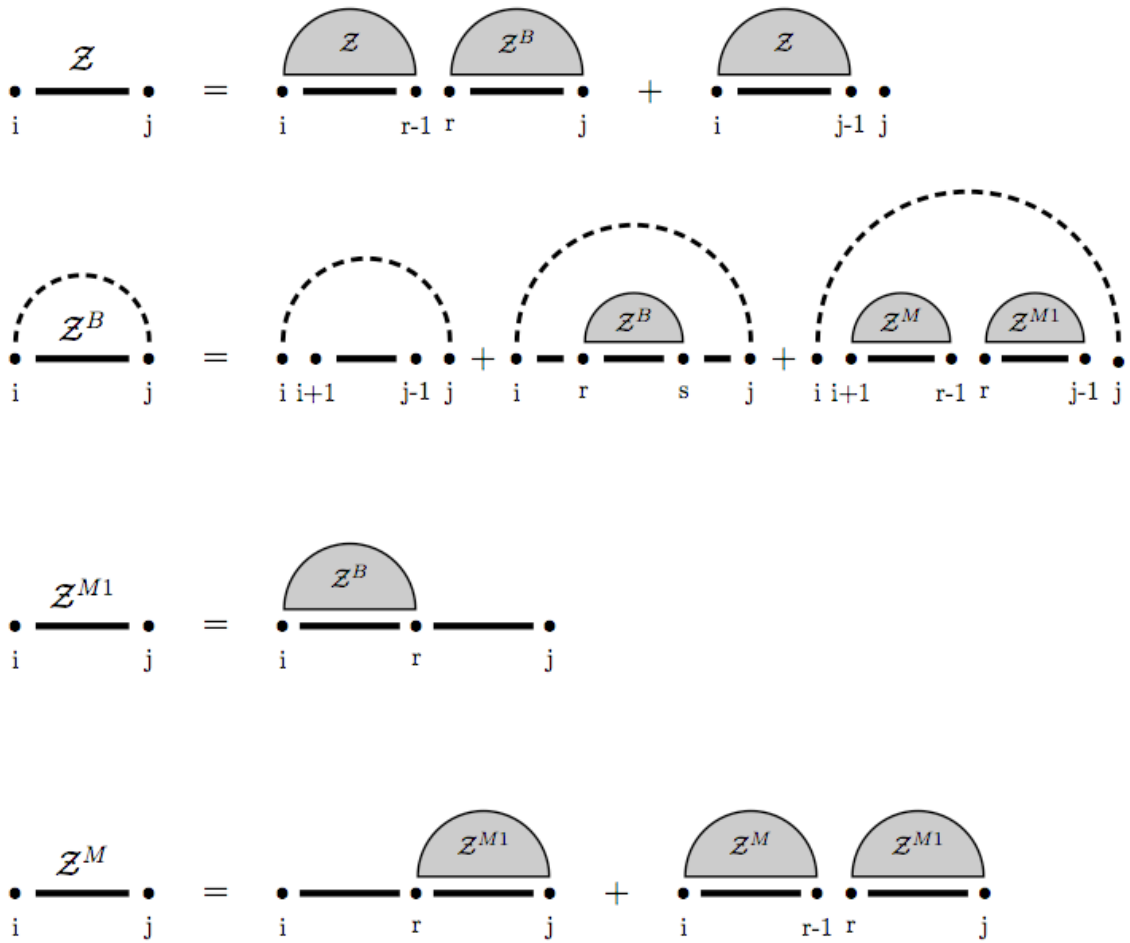
1-neighborhood
(1 mutations)



Sequence ensemble

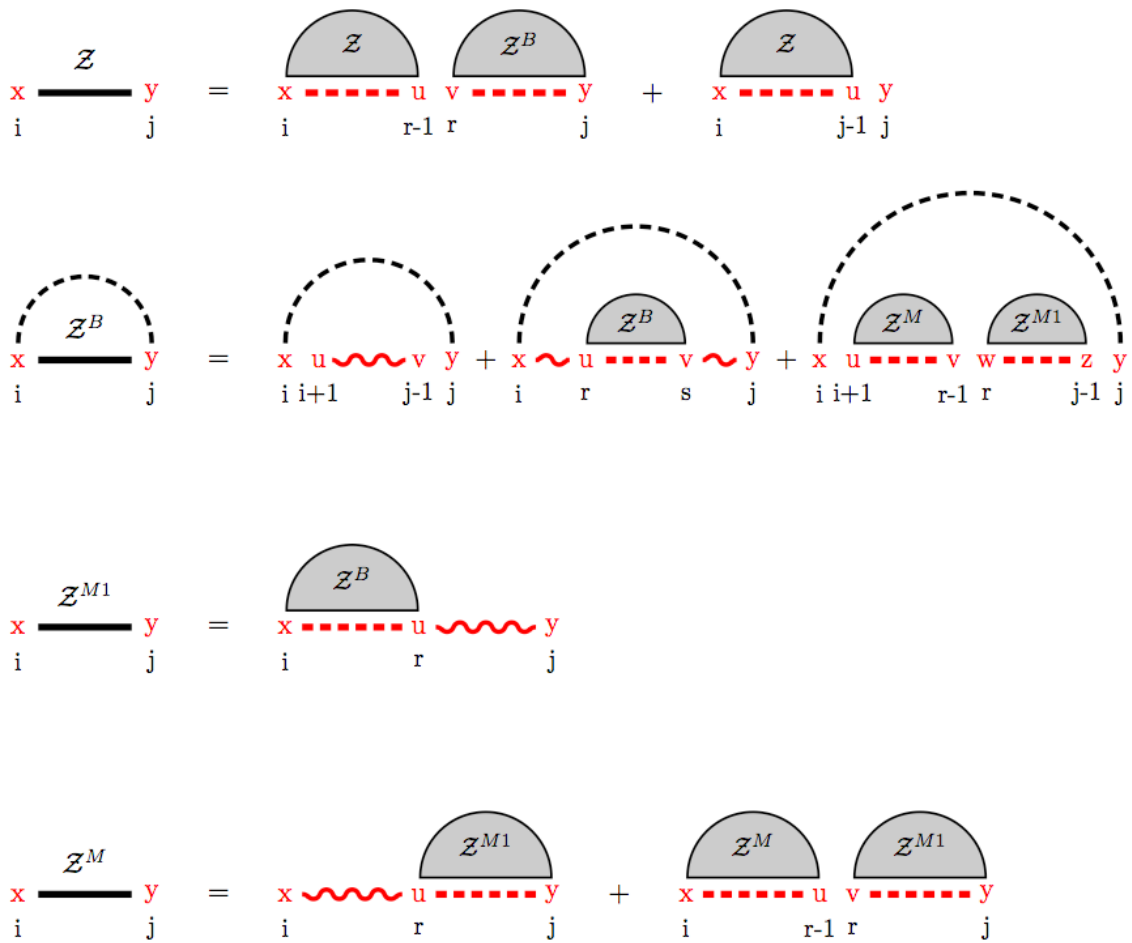
Structure ensemble

Classical Recursions (Zuker & Stiegler, McCaskill)



Enumerate all secondary structures

RNAmutants Generalize Classical Algorithms



Enumerate all secondary structures over all mutants

(Waldispuhl et al., *ECCB*, 2002)

Our approach

RNAmutants

- Explore the complete mutation landscape.
- *Polynomial* time and space algorithm.
- Compute the partition function for all sequences:

RNAmutants:

$$Z = \sum_{s} \sum_{S} \exp\left(-\frac{E(s,S)}{RT}\right)$$

Single sequence:

$$Z(s) = \sum_{S} \exp(b \times E(s,S))$$

- Sample by backtracking in dynamic prog. Tables.

  => Weighted random generation (recursive method)

(Waldispuhl et al., *PLoS Comp Bio*, 2008)

Sampling k-mutants

Seed

```
CAGUGAUUGCAGUGCGAUGC   (-1.20)
..((.(((((...)))))))
```

Classic: 0 mutation

```
CAGUGAUUGCAGUGCGAUcC   (-3.40)
..(.((((((...))))))
CAGUGAUUGCAGUGCGgUGC   (-0.30)
((.((....)).))......
CAGUGAUcGCAGUGCGAUGC   (-3.10)
.....(((((...)))))..
```

RNAmutants: 1 mutation

```
uAGcGccgGgAGacCGgcGC   (-18.00)
..(((((((....)))))))
CccUGgccGCAagGCcAgGg   (-20.40)
(((((((((....)))))))
CcGUGgccGCgagGCcAcGg   (-19.10)
(((((((((....)))))))
```

RNAmutants: 10 mutations

Sample k mutations increasing the folding energy

Consequence: Increase in C+G content

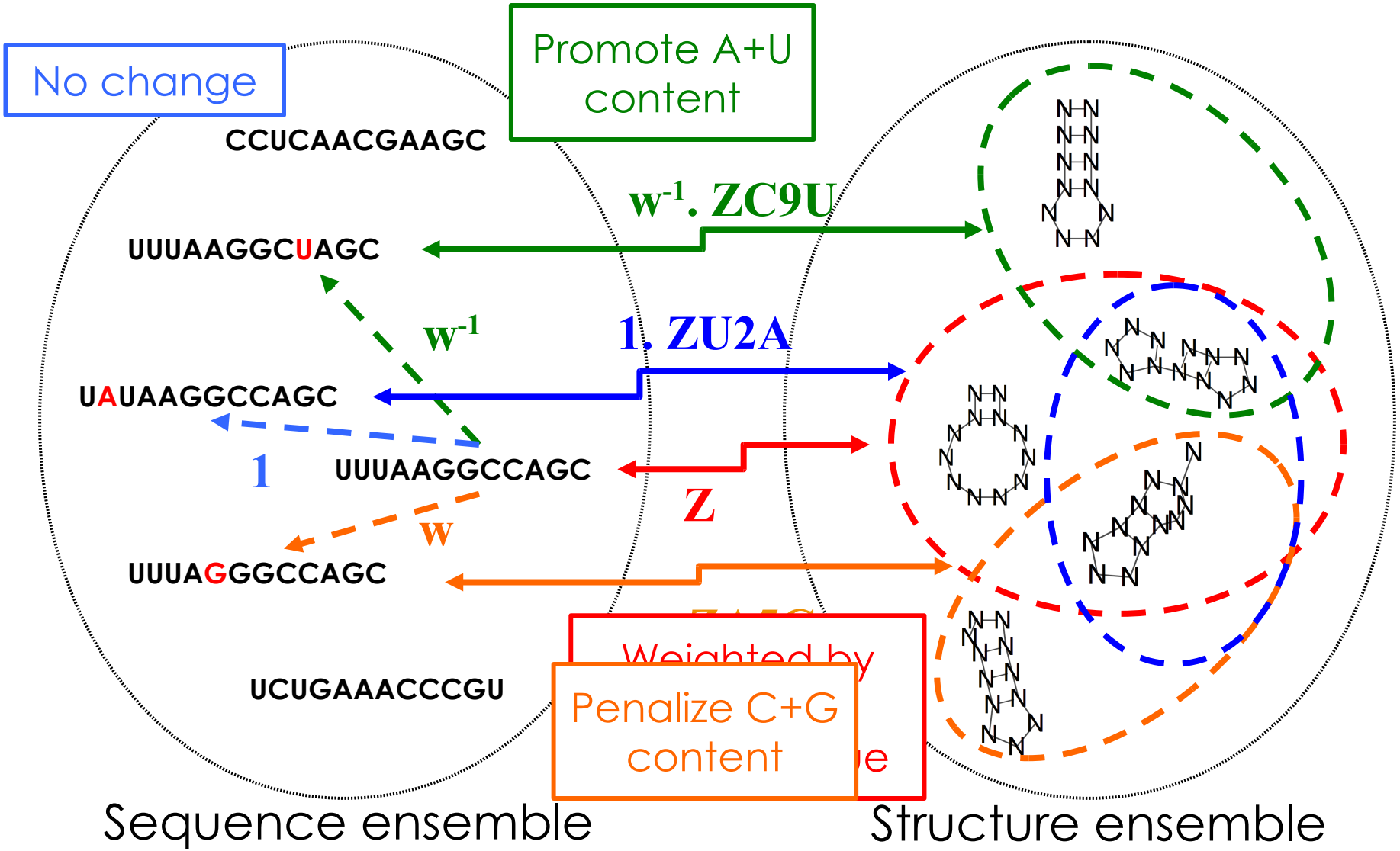
Outline

- Background: RNAmutants in a nutshell
  Algorithms to sample RNA secondary structures and mutations.

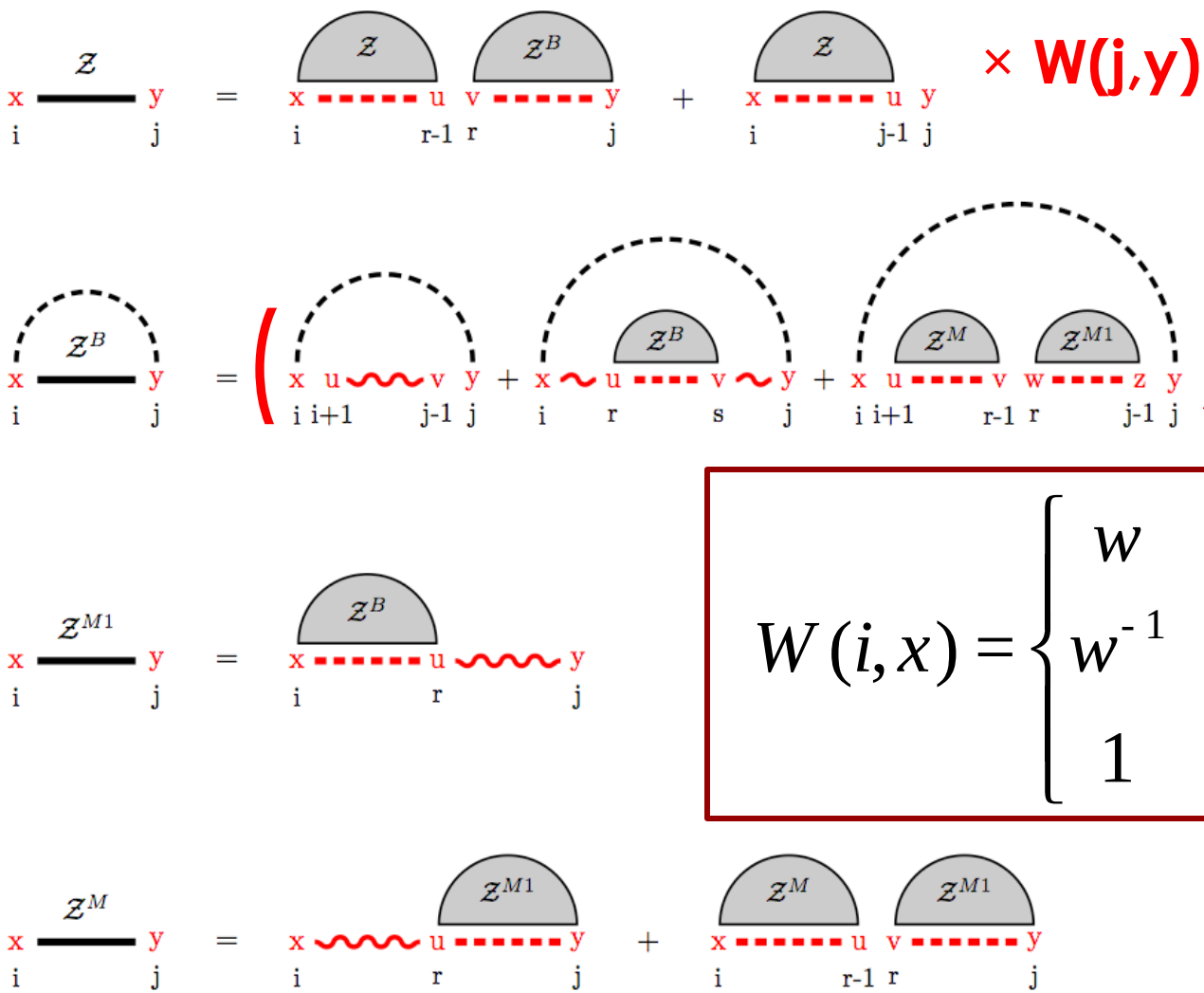
- **Our approach: Adaptive sampling**
  Uniformly shifting the distribution of samples.

- Results: Evolutionary studies
  Insights on the evolutionary pressure stemming from an optimization of the thermodynamical stability.

Our approach: Weighting mutations

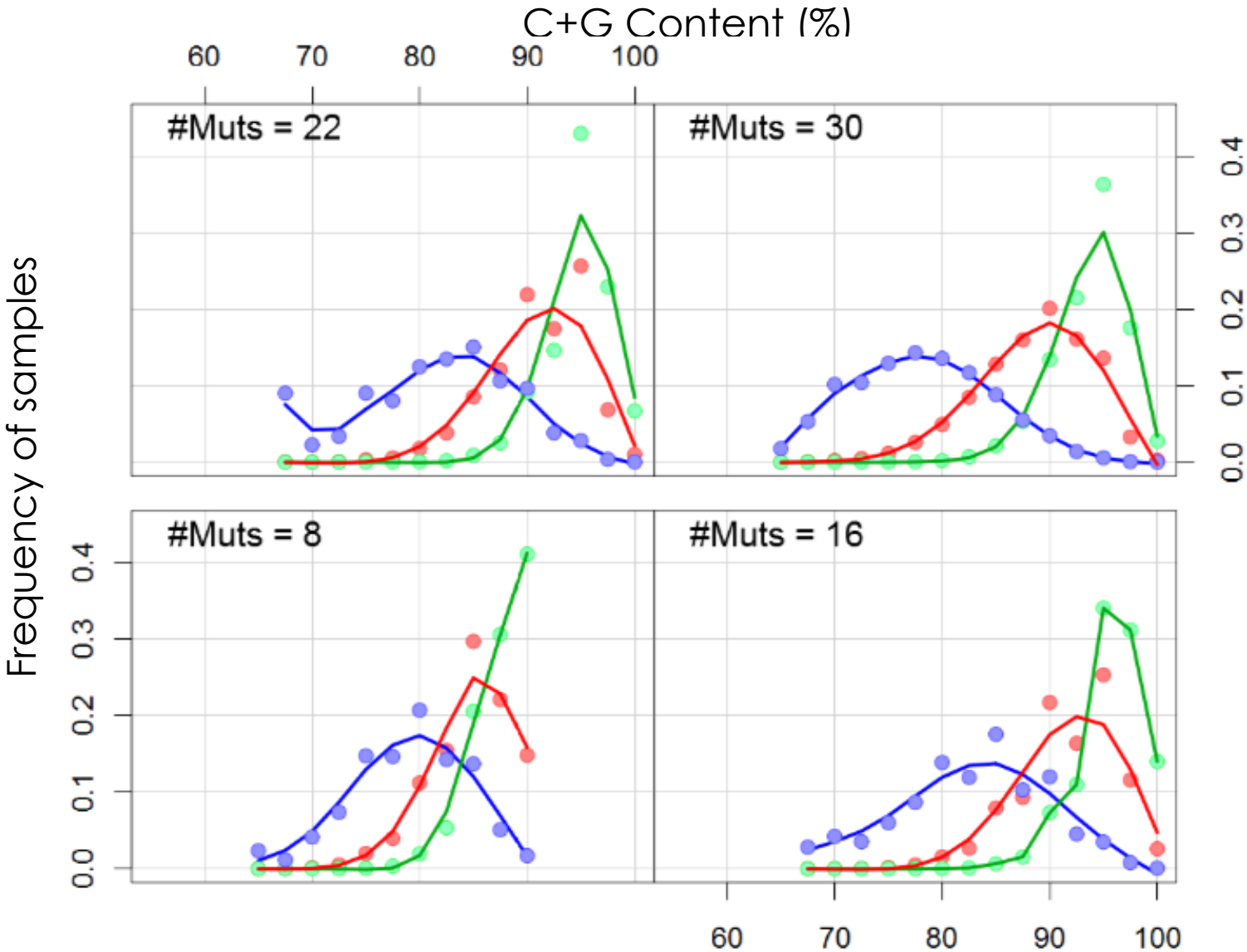


Weighting recursive equations



$$W(i,x) = \begin{cases} w & \text{If } A,U \rightarrow C,G \\ w^{-1} & \text{If } C,G \rightarrow A,U \\ 1 & \text{Otherwise} \end{cases}$$

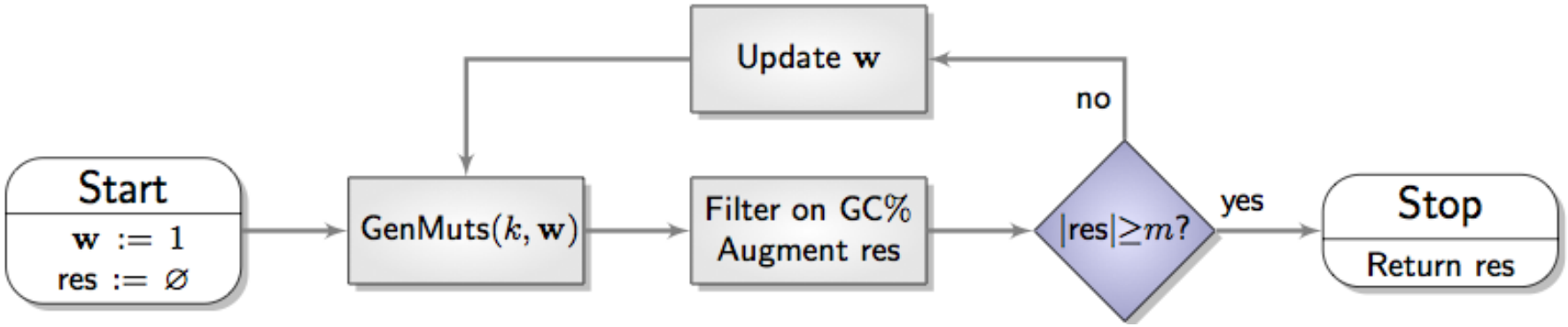
Effect of weighted sampling



■ Unweighted sampling ■ weighted (**w**=1/2) ■ weighted (**w**=2)

An Adaptive Sampling approach

Idea: Figure out a suitable weight *on the fly*.



- Keep all samples at targeted C+G% and reject others.
- Update **w** at each iteration using a bisection strategy.
- Stop when enough samples have been stored.

Algorithmic aspects

· After rejection, the weighted schema only impacts performance, not the probability distribution.

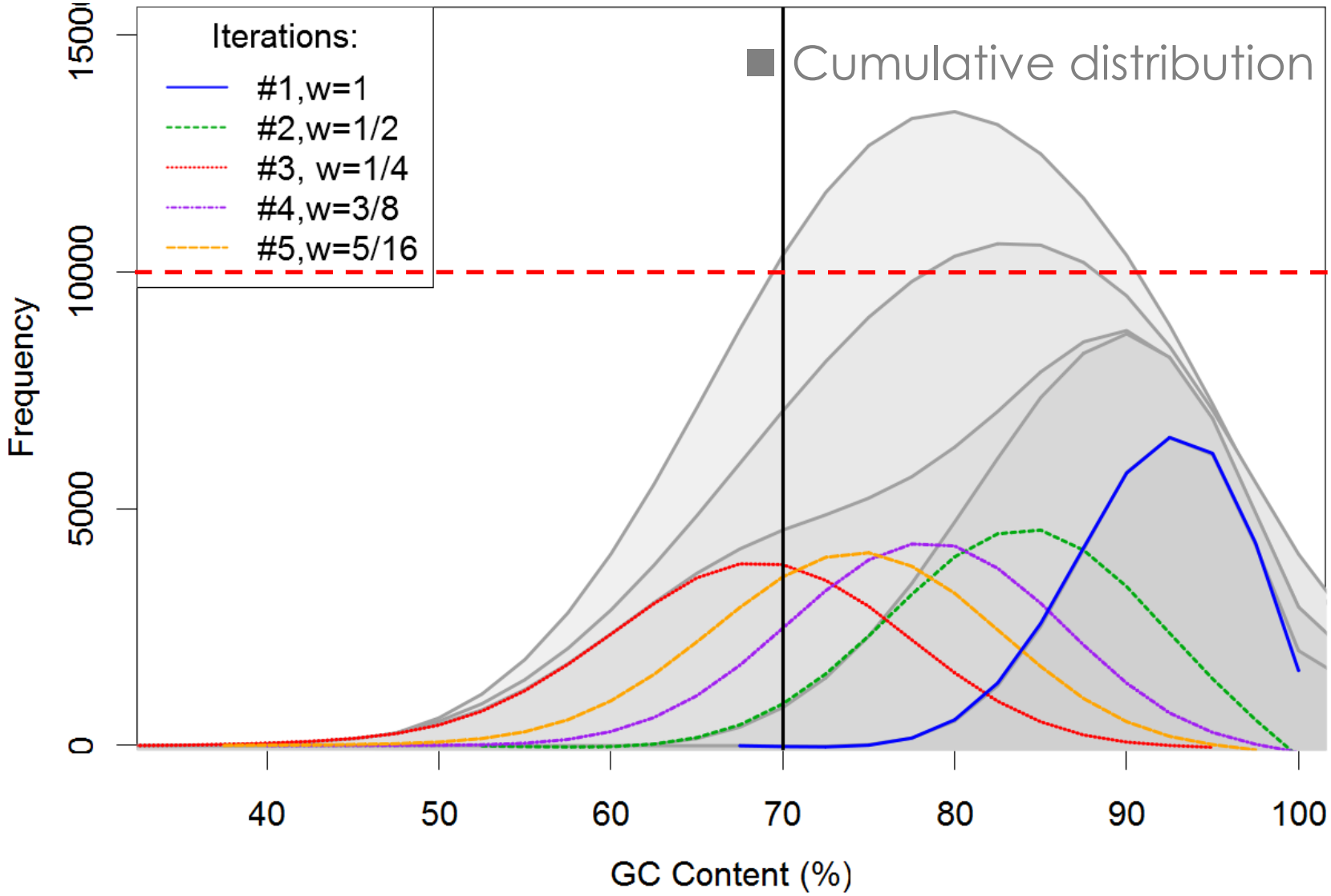
· Partition function can be written as a polynomial:

$$Z = \sum_{i=0}^{n} a_i \times w^i$$

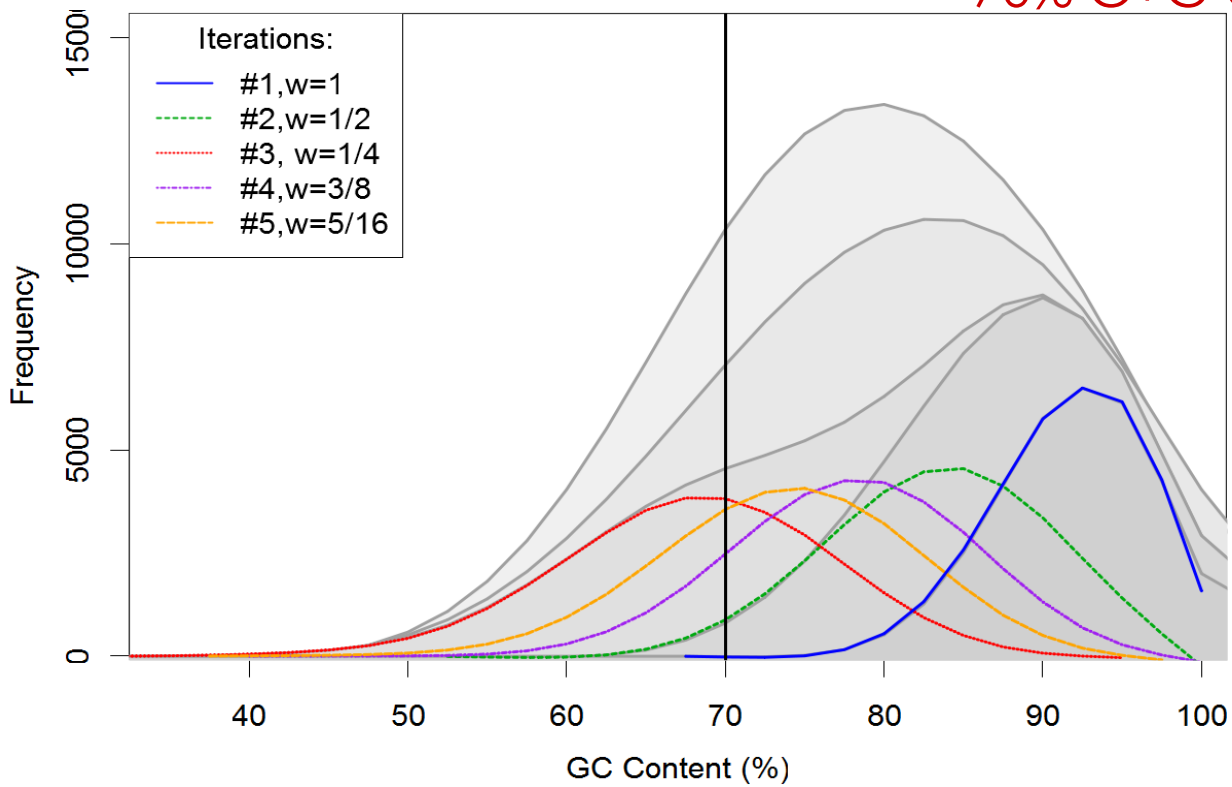
After *n* iterations one can compute all a_i through interpolation and *invert* the polynomial to compute optimal weight *w* (Gröbner base).

Remark: In practice, few iterations are necessary

Example: 40 nt., 10000 samples, 30 mutations, 70% C+G content



Example: 40 nt., 10000 samples, 30 mutations, 70% C+G content



Overall, 150 000 samples required.
Naïve rejection: ~30 000 000 samples!
+ Ratio grows exponentially on length (gen. Fun.)

Note: Dynamic programming alternative in $O(n^7)$

Outline

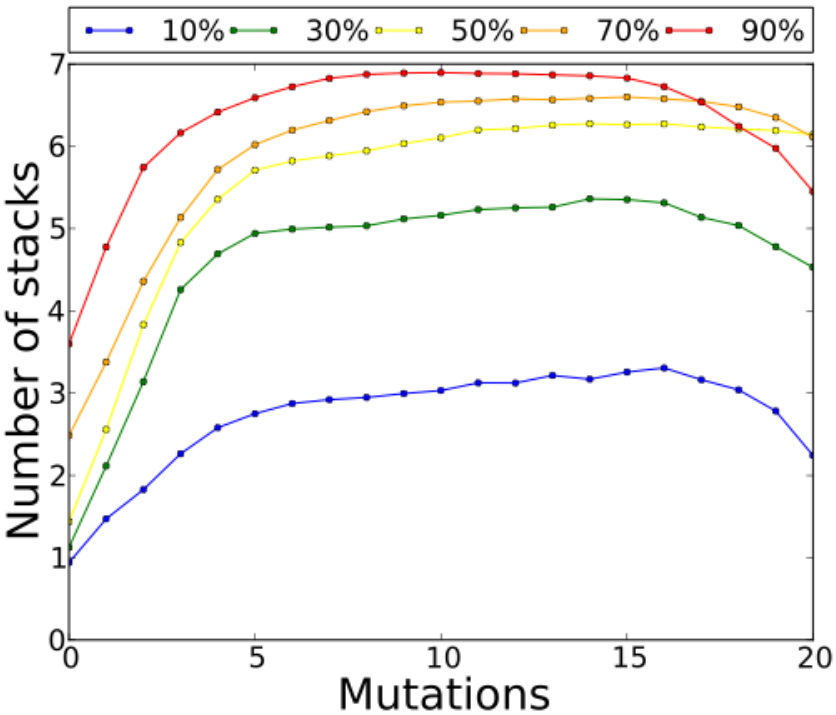
- Background: RNAmutants in a nutshell
  Algorithms to sample RNA secondary structures and mutations.

- Our approach: Adaptive sampling
  Uniformly shifting the distribution of samples.

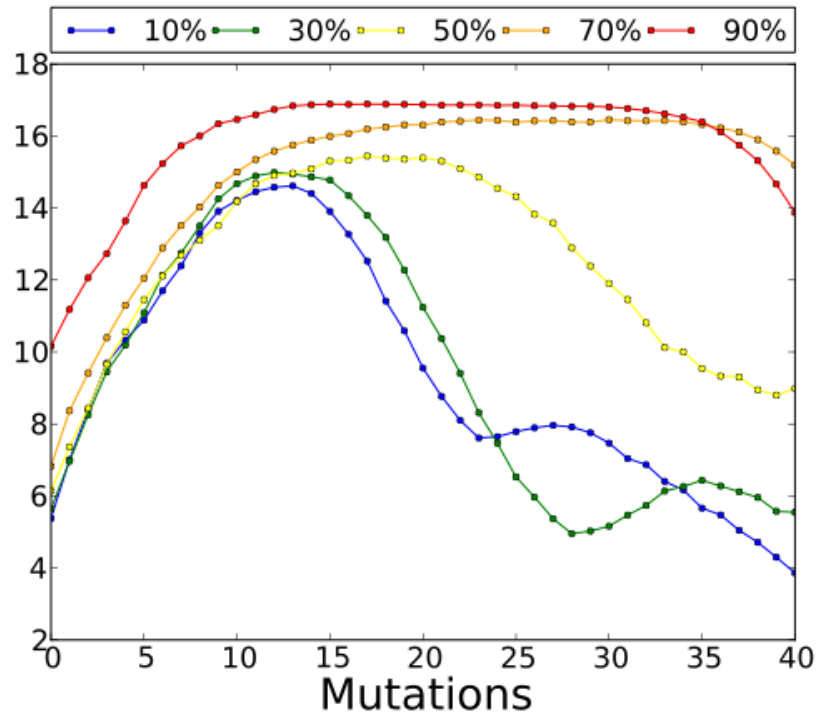
- **Results: Evolutionary studies**
  Insight into the evolutionary pressure stemming from an optimization of the thermodynamical stability.

Low CG-contents favor structural diversity

Simulation at **fixed G+C content** from random seeds



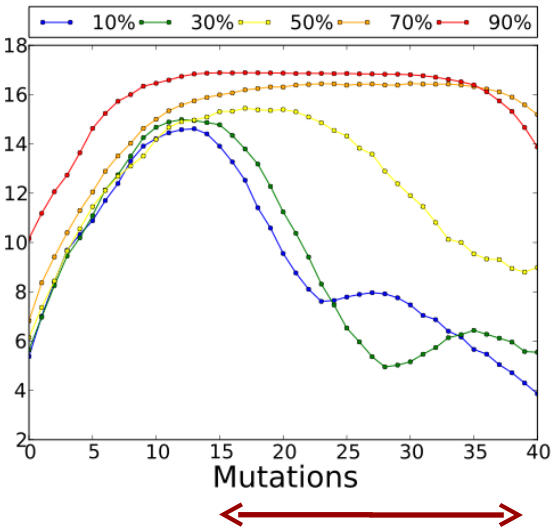
20 nucleotides

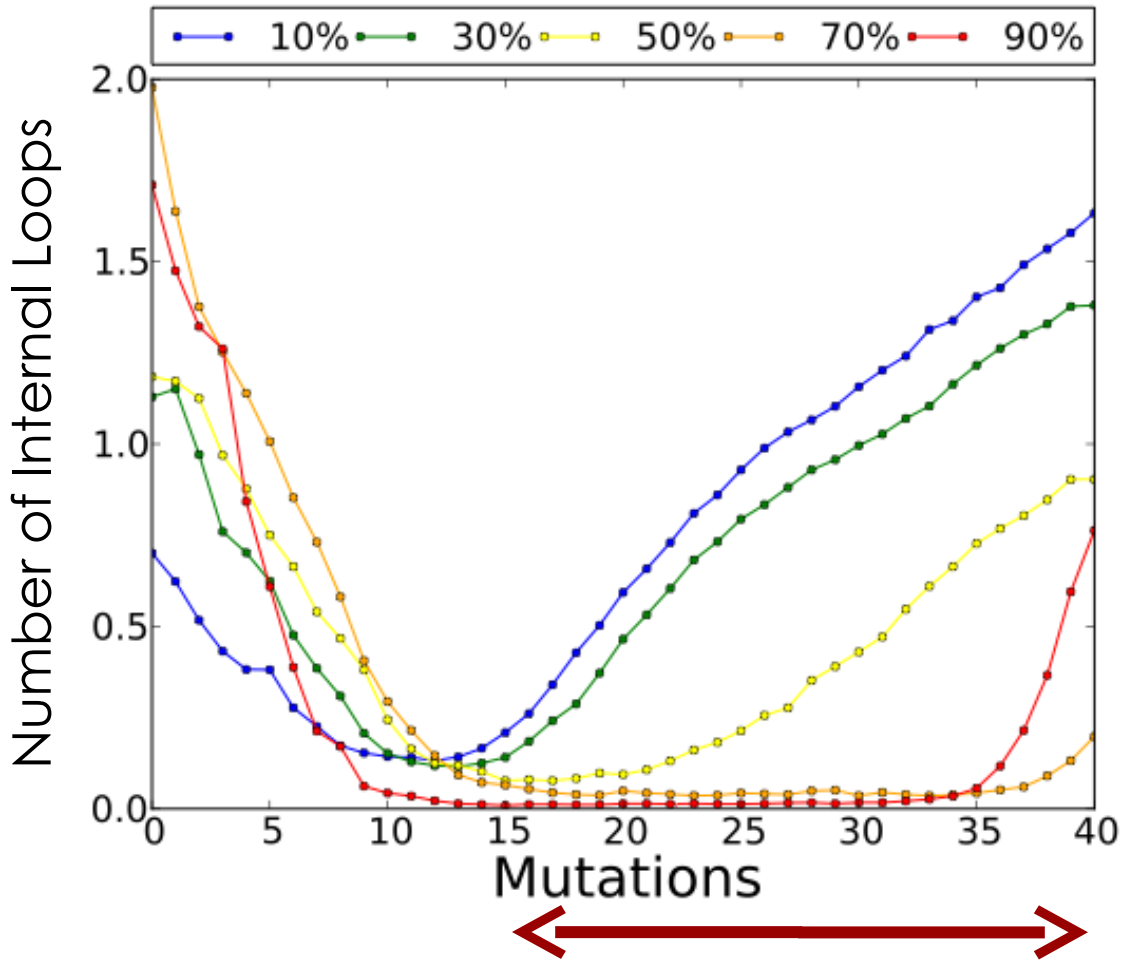


40 nucleotides

10% 30% 50% 70% 90%

Low C+G contents favor internal loop insertion

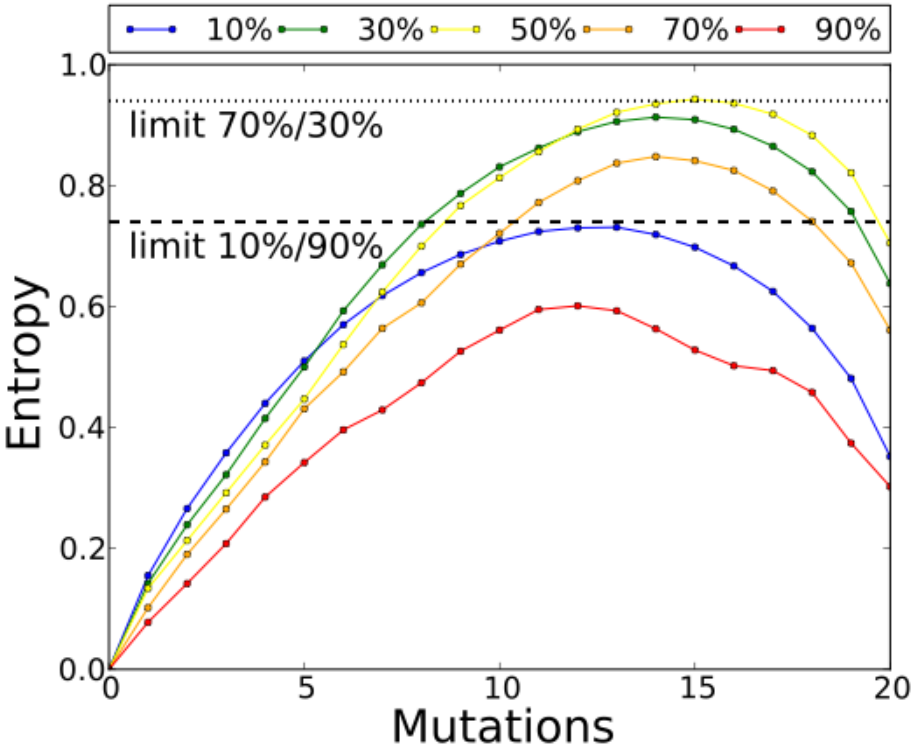




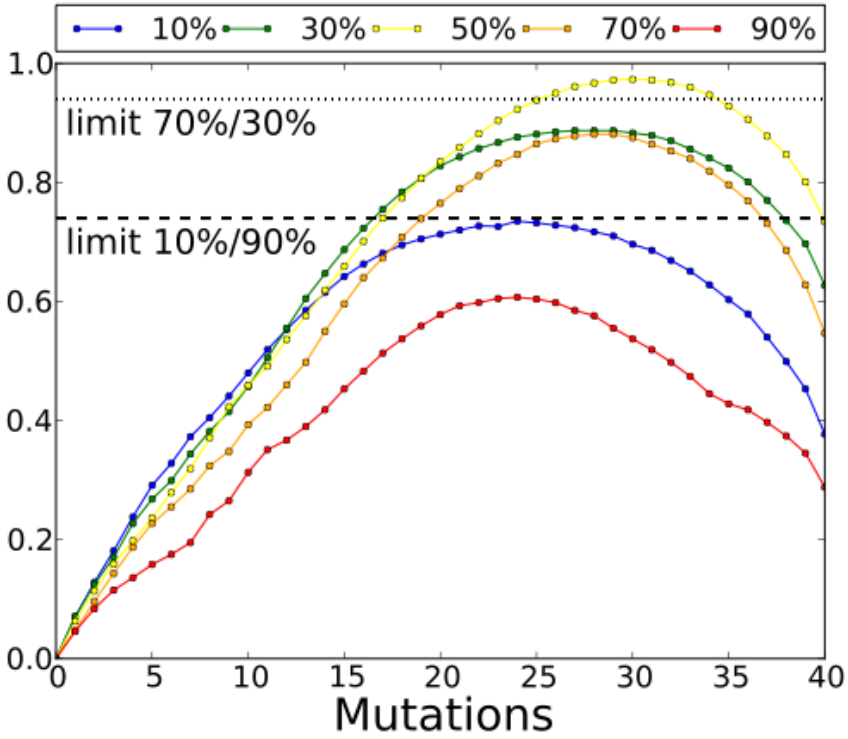
■ 10% ■ 30% ■ 50% ■ 70% ■ 90%

High G+C-contents reduce evolutionary accessibility

Simulation at **fixed G+C content** from random seeds



20 nucleotides



40 nucleotides

■ 10% ■ 30% ■ 50% ■ 70% ■ 90%

Perspectives

- More studies of Sequence-Structure maps.
- Applications to RNA design.
- Same techniques can be applied to other parameters (e.g. number of base pairs).
- Can be generalized to multiple parameters.

Acknowledgments

Jérôme Waldispühl

Assistant Prof. McGill, Montreal Canada.

INRIA
· Philippe Flajolet

MIT
· Bonnie Berger
· Srinivas Devadas
· Mieszko Lis
· Alex Levin
· Charles W. O'Donnell

Google Inc.
· Behshad Behzadi

Ecole Polytechnique
· Jean-Marc Steyaert

Boston College
· Peter Clote

University of Paris 6
· Olivier Bodini

University of Paris 11
· Alain Denise

ANR MAGNUM

NSERC CRSNG

Would you like to know more?

- O. Bodini, and Y. Ponty
Multi-dimensional Boltzmann Sampling of Languages,
Proceedings of AOFA'10, 49--64, 2010

- J. Waldispühl, S. Devadas, B. Berger and P. Clote,
Efficient Algorithms for Probing the RNA Mutation Landscape,
Plos Computational Biology, 4(8):e1000124, 2008.

http://csb.cs.mcgill.ca/RNAmutants