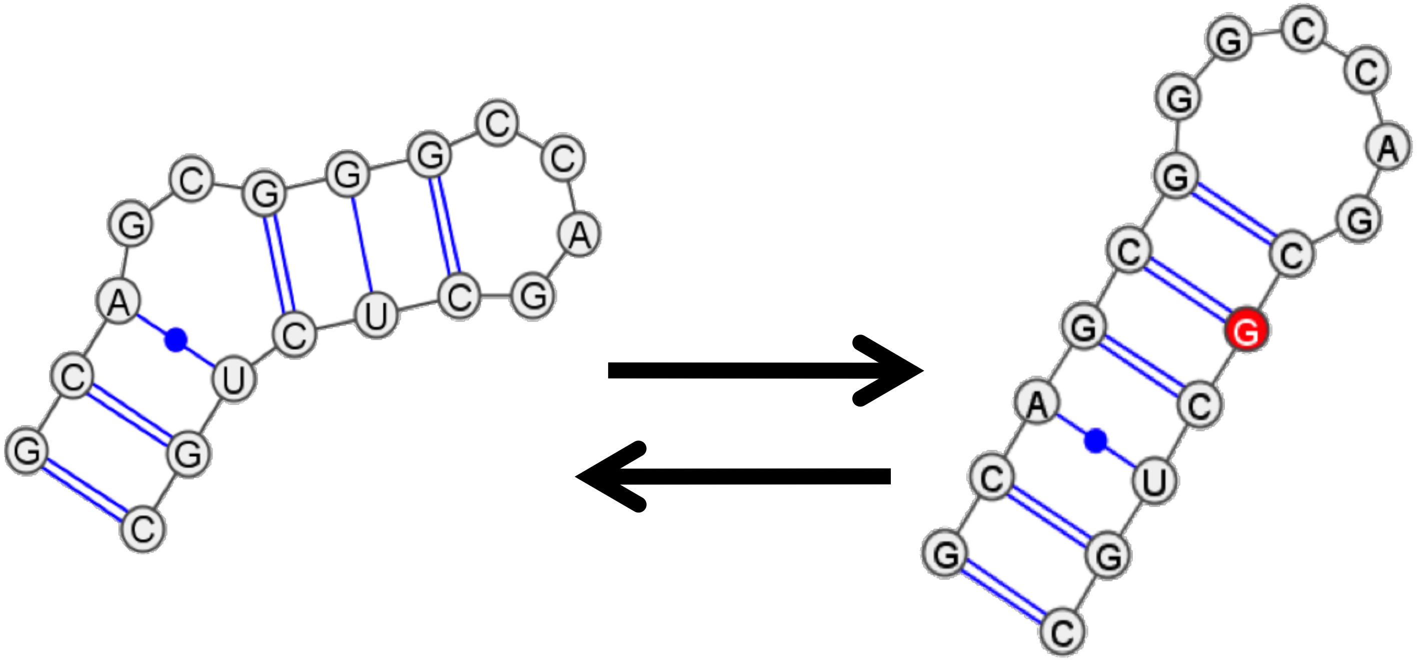
COMP598: Advanced Computational Biology Methods & Research

Exploring the RNA mutational Landscape: Algorithms & Applications

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

Includes slides from V. Reinharz

Overview

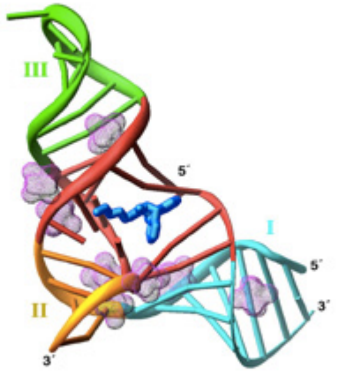


How mutations affect structures… and vice versa!

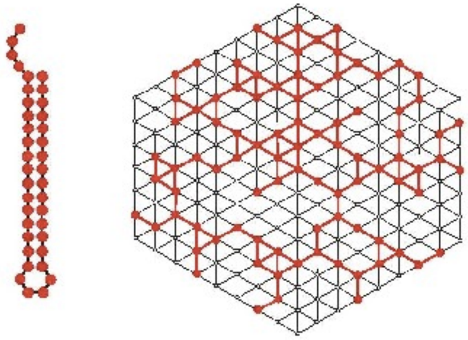
- Brute force approach: Slow & not scalable.
- **Our Approach:** Fast, scalable… & elegant!

Motivations

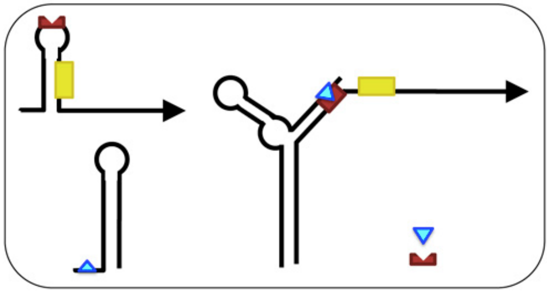
- Analysis of molecular Functions



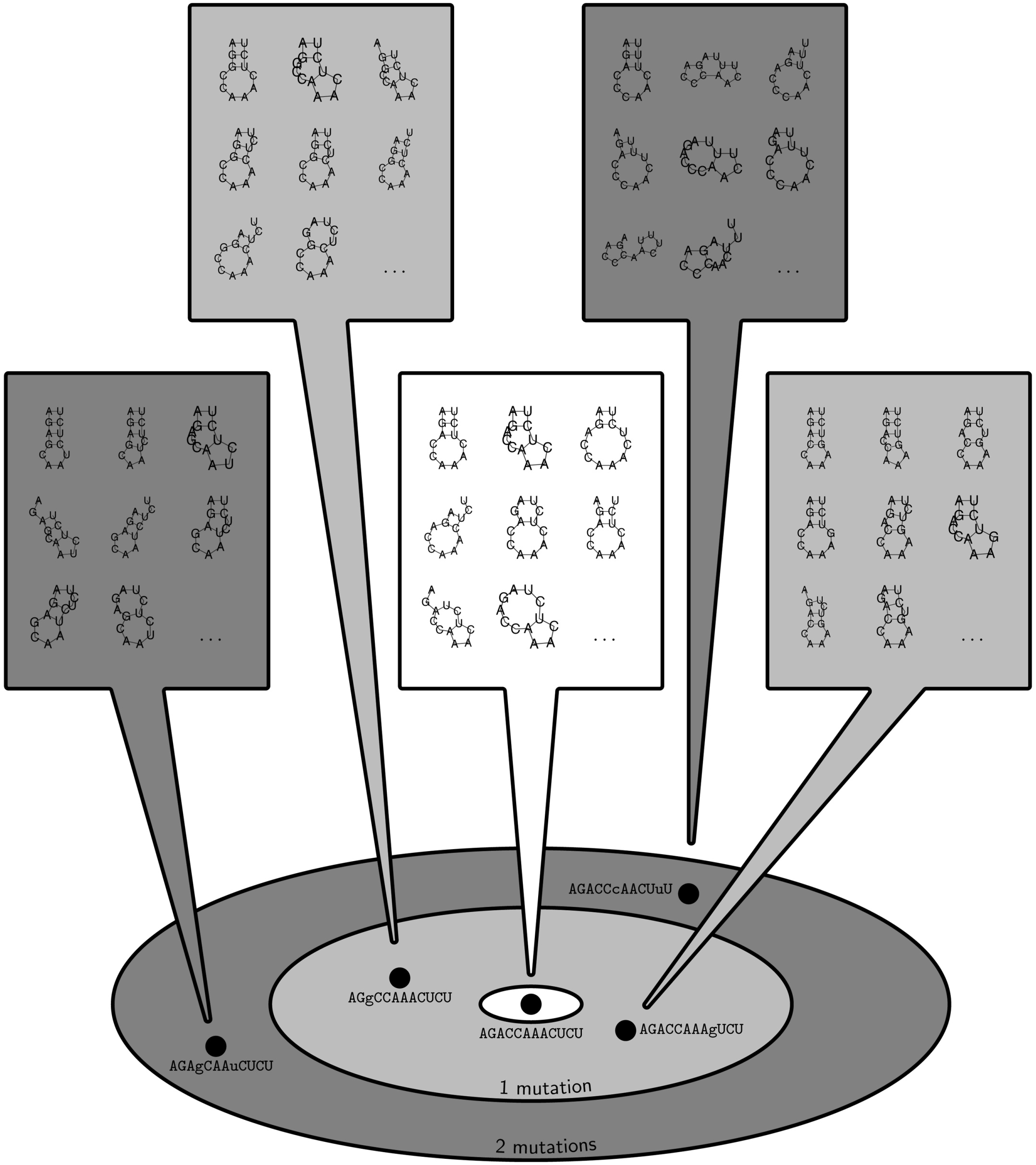
- Evolutionary studies



- Synthetic biology systems



RNAmutants


AGACCcAACUuU
AGgCCAAACUCU
AGACCAAACUCU
AGACCAAAgUCU
AGAgCAAuCUCU
1 mutation
2 mutations

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

Outline

- **Computing the Mutational Landscape**

  (Waldispühl *et al.*, 2008)

- **Controlling the nucleotide distribution**

  (Waldispühl & Ponty, 2011)

- **Applications**

  (Lam *et al.*, 2011; Levin *et al.*, 2012; Reinharz *et al.*, 2013)

RNA sequence-structure maps

CCUCAACGAAGC

UUUACGGCUAGC

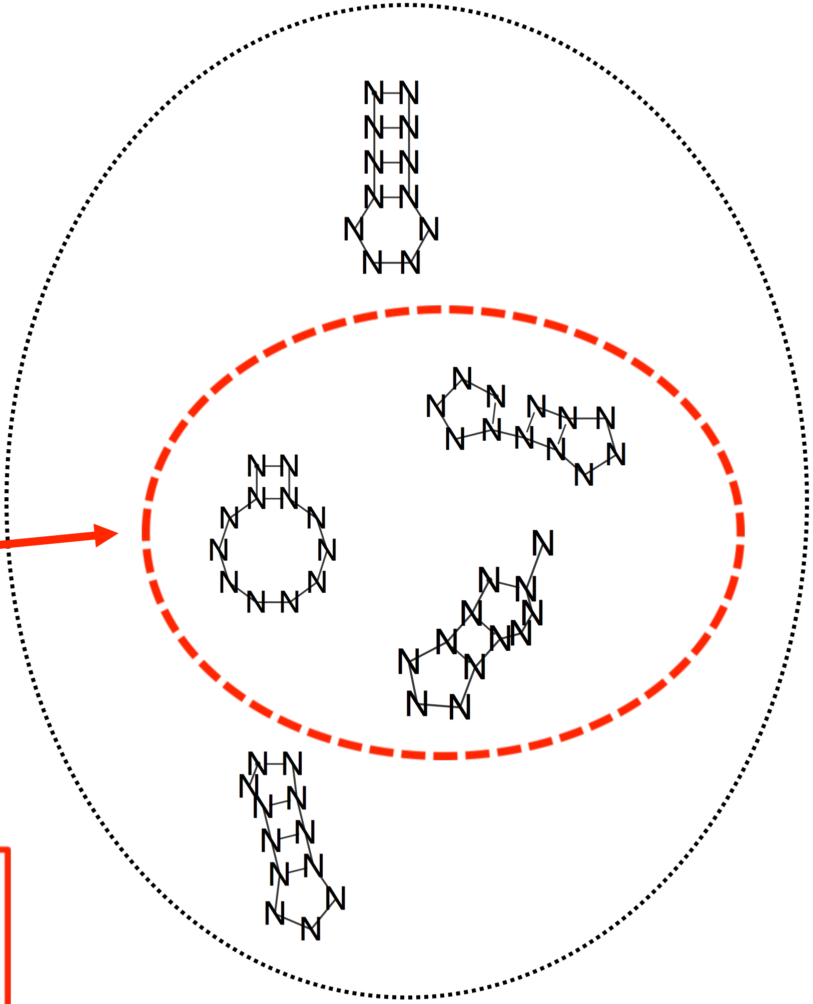
UAUACGGCCAGC

UUUAAGGCCAGC

UUUAGGGCCAGC

UCUCAAACGGCU

Z



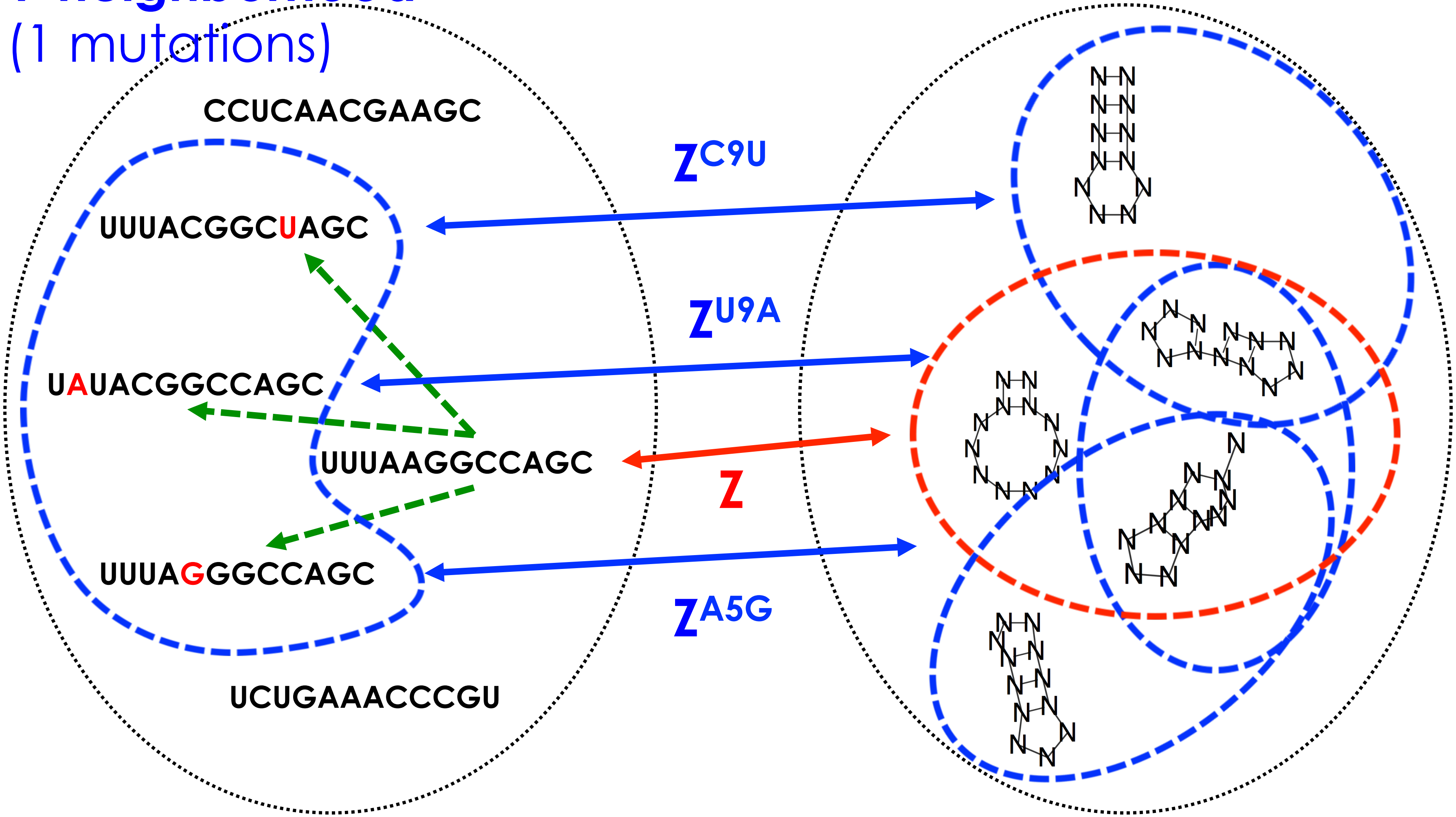
Structure ensemble

$$Z(s) = \sum_{S} \exp(-\beta \cdot E(s,S))$$

Boltzmann partition function

Parameterization of the mutational landscape

1-neighborhood
(1 mutations)



CCUCAACGAAGC

UUUACGGCUAGC

UAUACGGCCAGC

UUUAAGGCCAGC

UUUAGGGCCAGC

UCUGAACCCGU

Z^{C9U}

Z^{U9A}

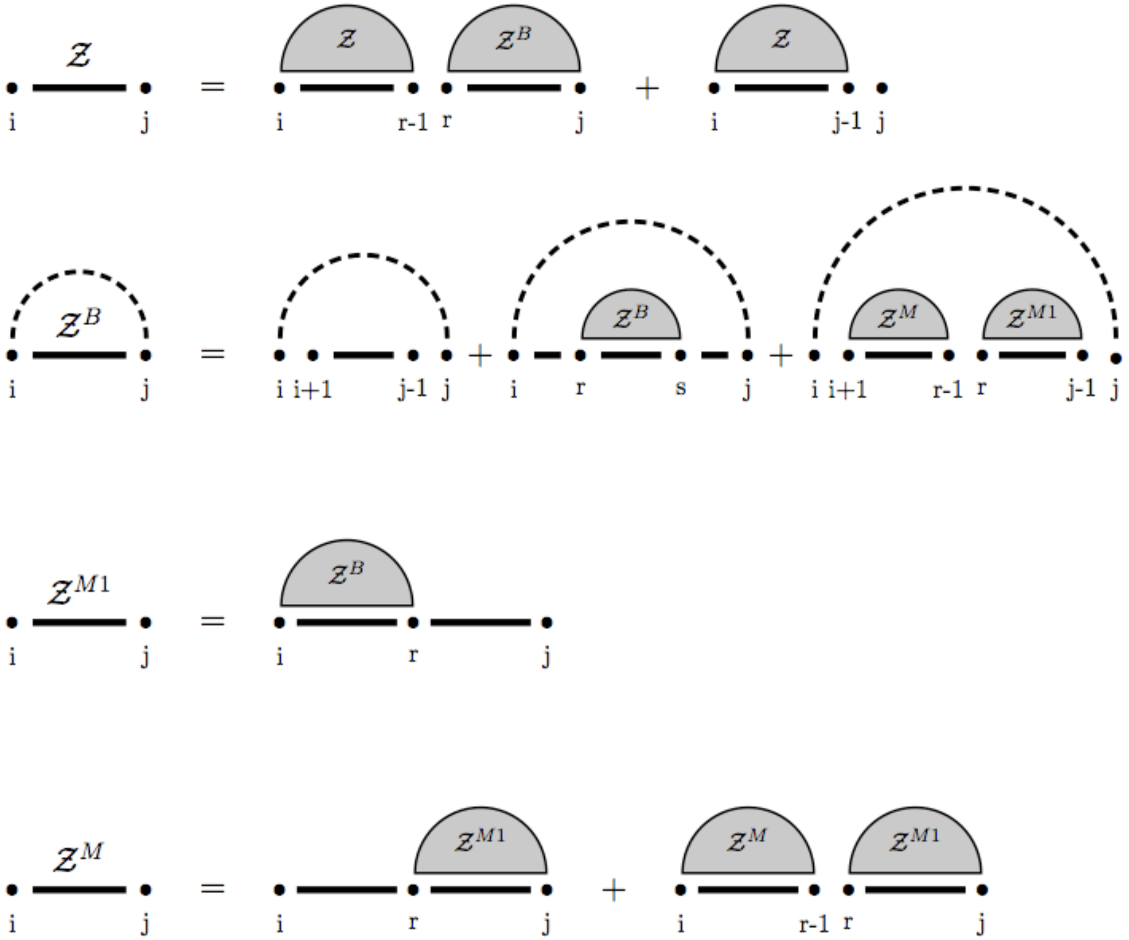
Z

Z^{A5G}

Sequence ensemble

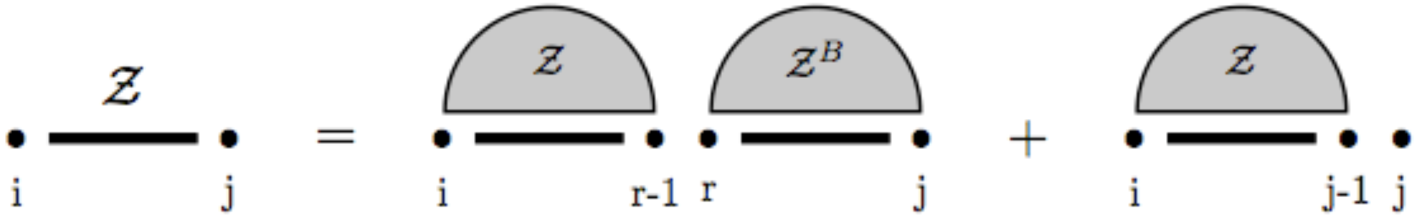
Structure ensemble

Classical Recursions (Zuker & Stiegler, McCaskill)



Enumerate all secondary structures

Classical Recursions (Zuker & Stiegler, McCaskill)

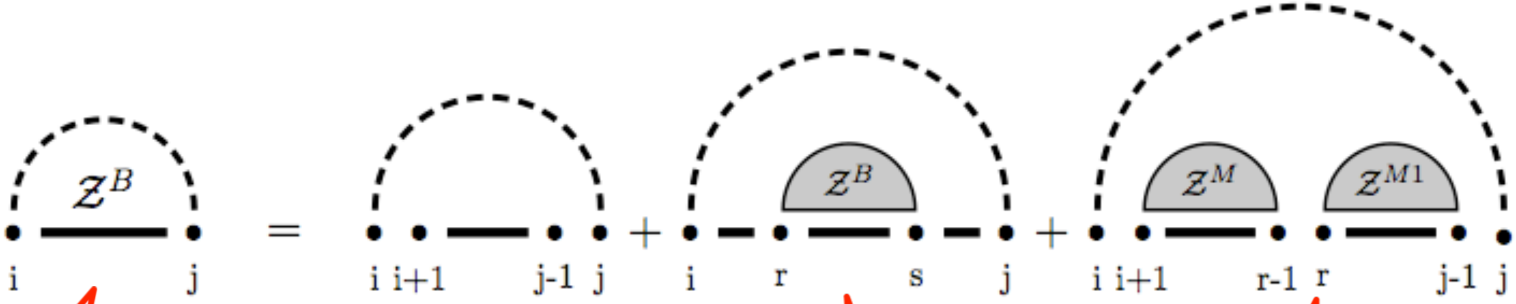


Any Secondary Structure on $S_{i,j}$

Index j base pair with r (i≤r(j)

Index j does NOT base pair

Classical Recursions (Zuker & Stiegler, McCaskill)



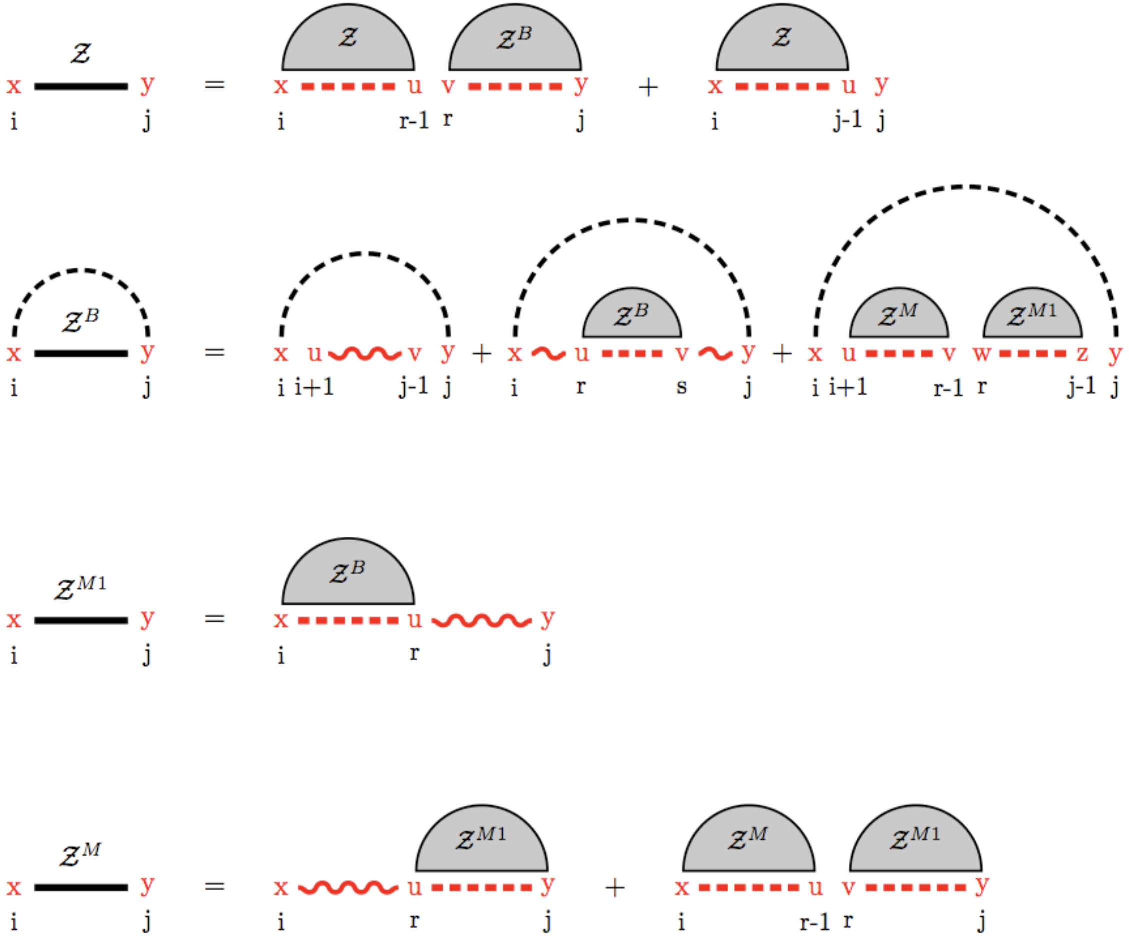
Secondary Structures on $S_{i,j}$ s.t. (i,j) base pair

Hairpin

Internal loop. (r,s) base pair

Multi-loop

RNAmutants Generalize Classical Algorithms



Enumerate all secondary structures over all mutants

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

Our approach

RNAmutants

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

RNAmutants:

$$Z = \sum_{s} \sum_{S} \exp(-\beta \cdot E(s,S))$$

Single sequence:

$$Z(s) = \sum_{S} \exp(-\beta \cdot E(s,S))$$

- Backtrack to sample mutants & secondary structures.

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

Sampling k-mutants

Seed

Classic: 0 mutation

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

RNAmutants: 1 mutation

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

RNAmutants: 10 mutations

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

C+G content of samples increases.

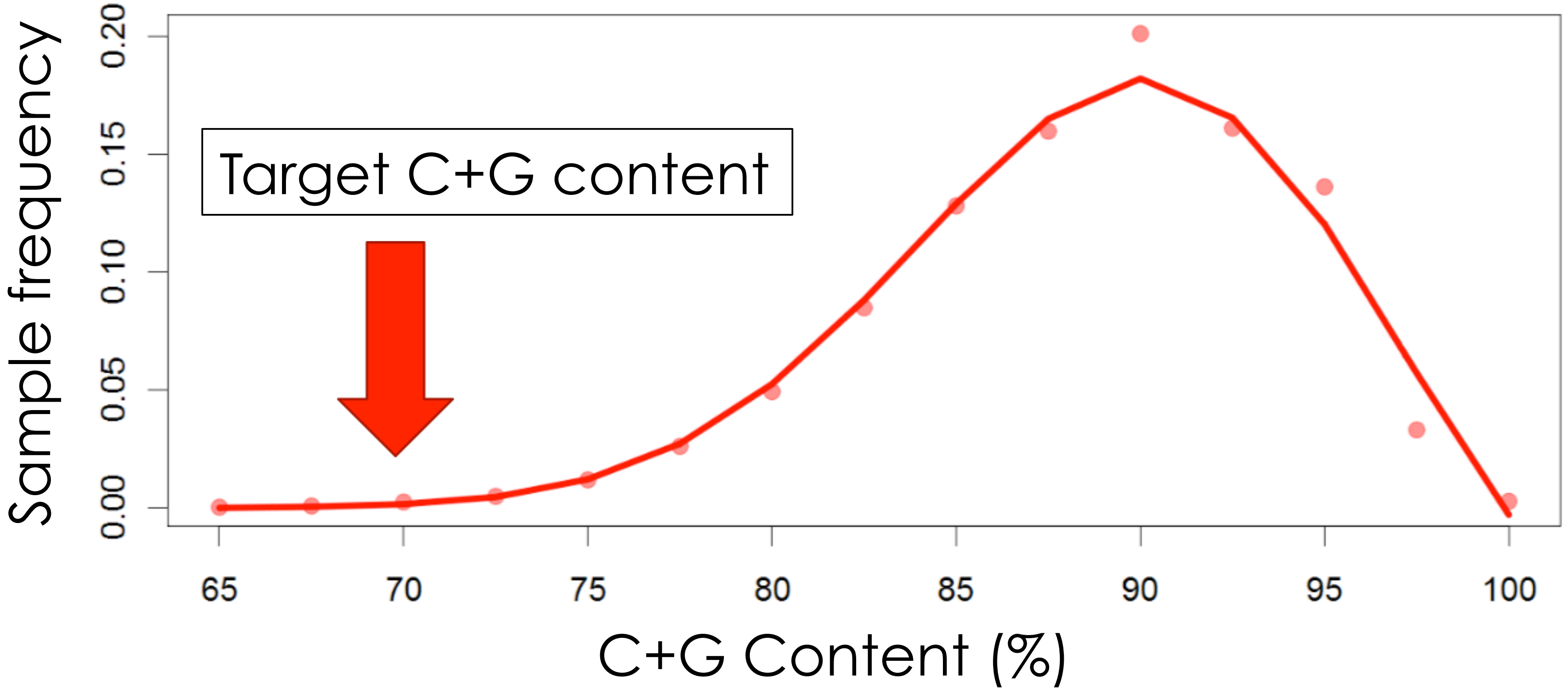
Outline

- **Computing the Mutational Landscape**
  (Waldispühl *et al.*, 2008)

- **Controlling the nucleotide distribution**
  (Waldispühl & Ponty, 2011)

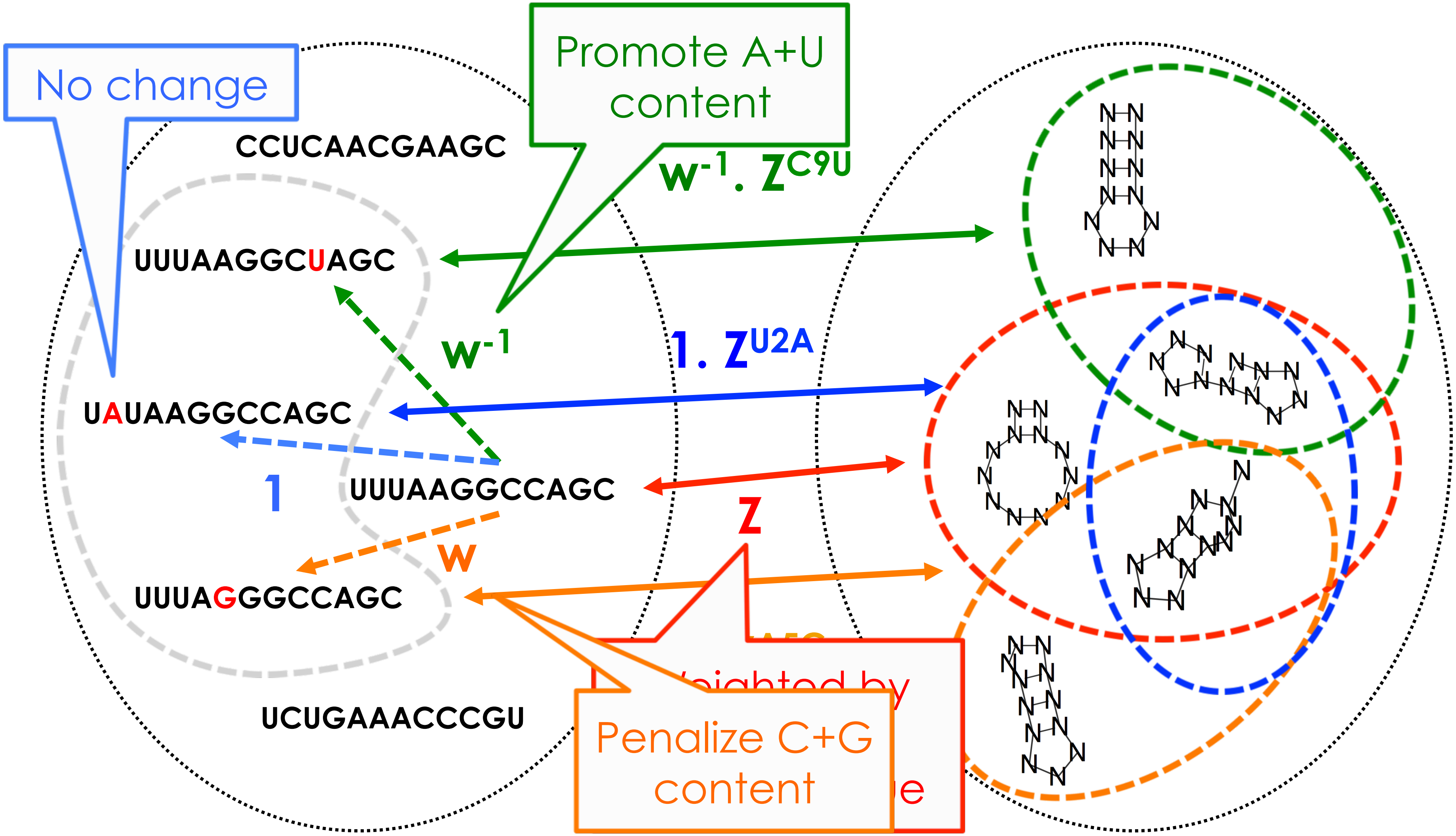
- **Applications**
  (Lam *et al.*, 2011; Levin *et al.*, 2012; Reinharz *et al.*, 2013)

Objectives



- Sampling at targeted CG% decreases exponentially with the length.
- How to efficiently sample sequences at arbitrary CG% contents … without bias!

Our approach: Weighting mutations



No change

Promote A+U content

Penalize C+G content

...ighted by

...e

CCUCAACGAAGC

UUUAAGGCUAGC

UAUAAGGCCAGC

UUUAAGGCCAGC

UUUAGGGCCAGC

UCUGAACCCGU

w^{-1}

$w^{-1} . Z^{C9U}$

$1 . Z^{U2A}$

1

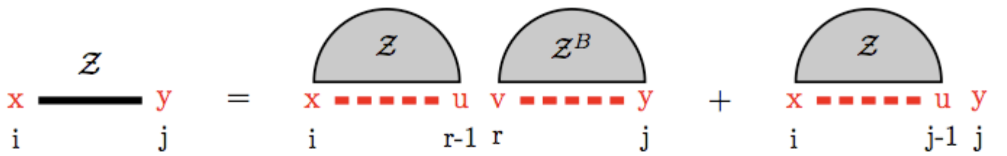
Z

w

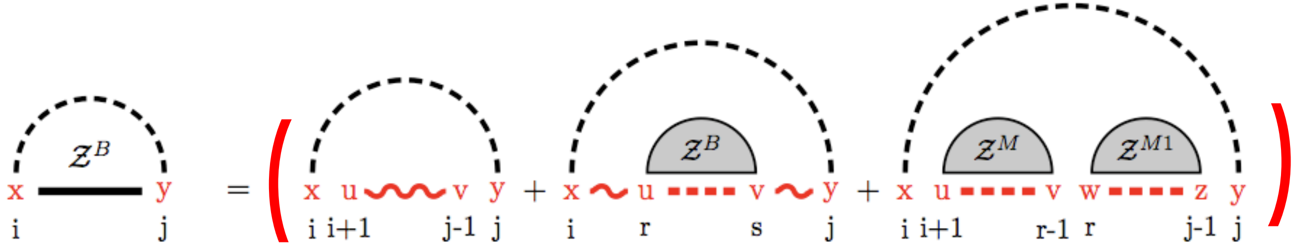
Sequence ensemble

Structure ensemble

Weighting recursive equations

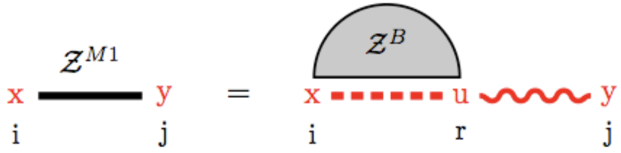


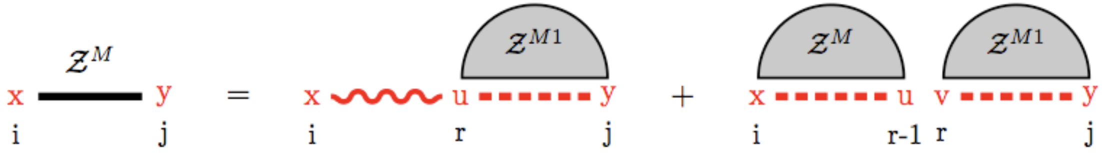
× W(j,y)



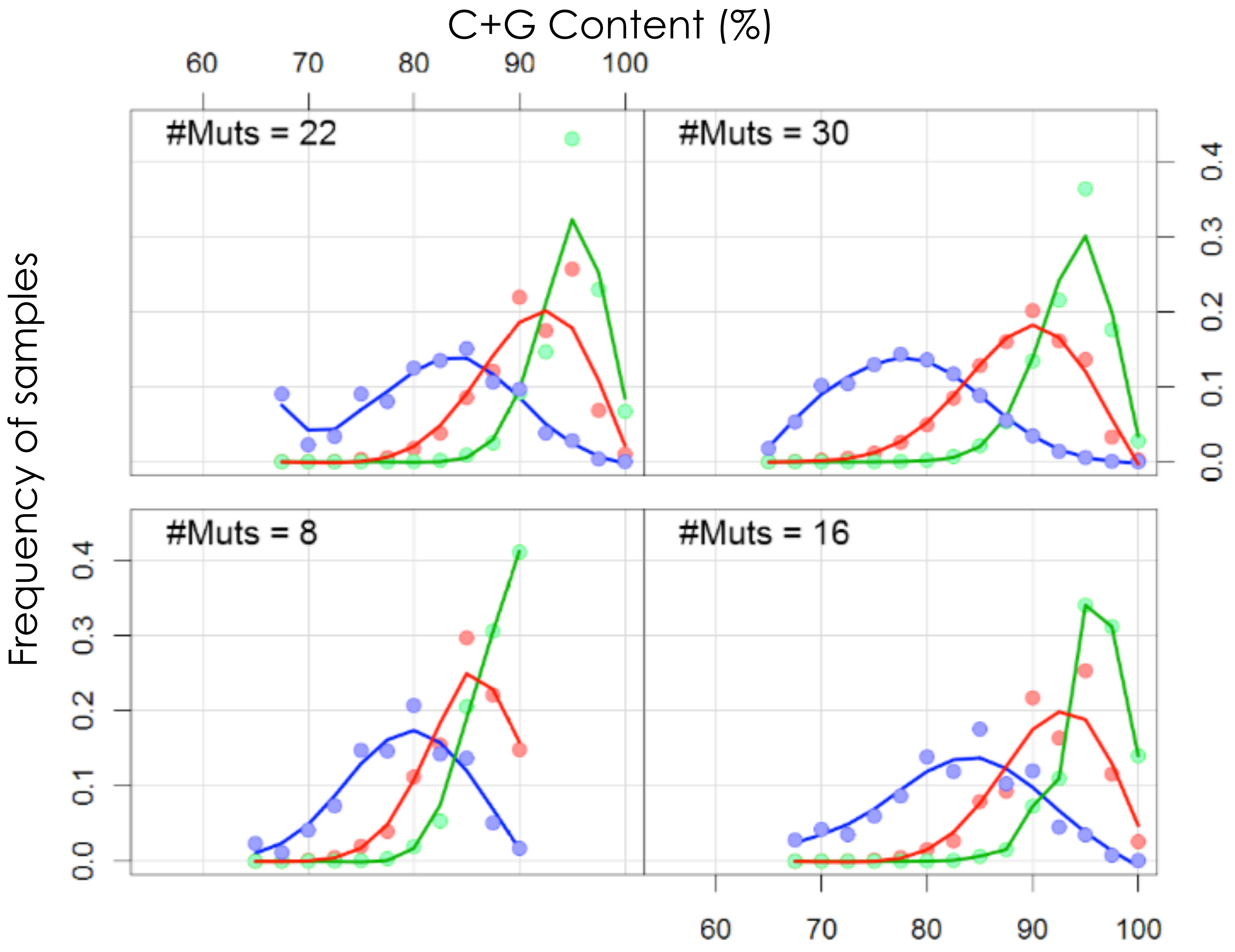
× W(i,x) × W(j,y)

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



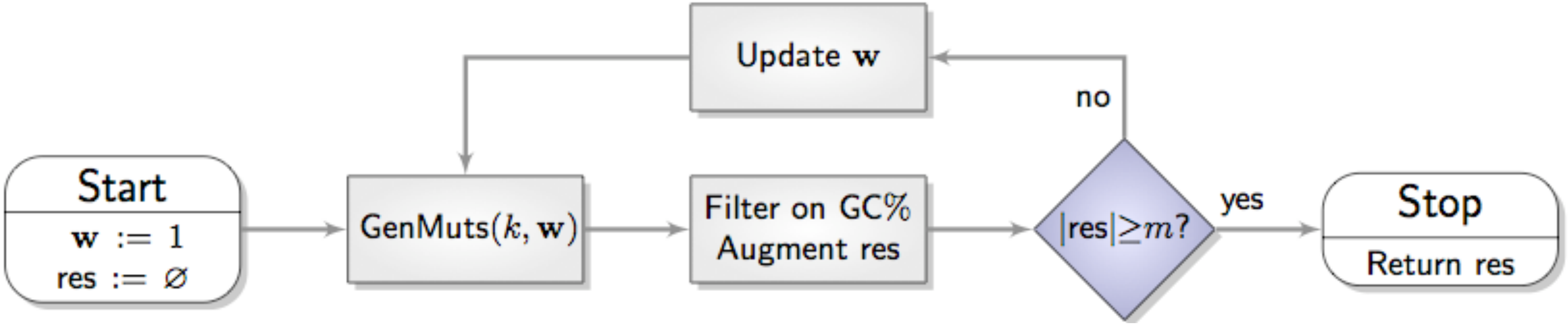


Effect of weighted sampling



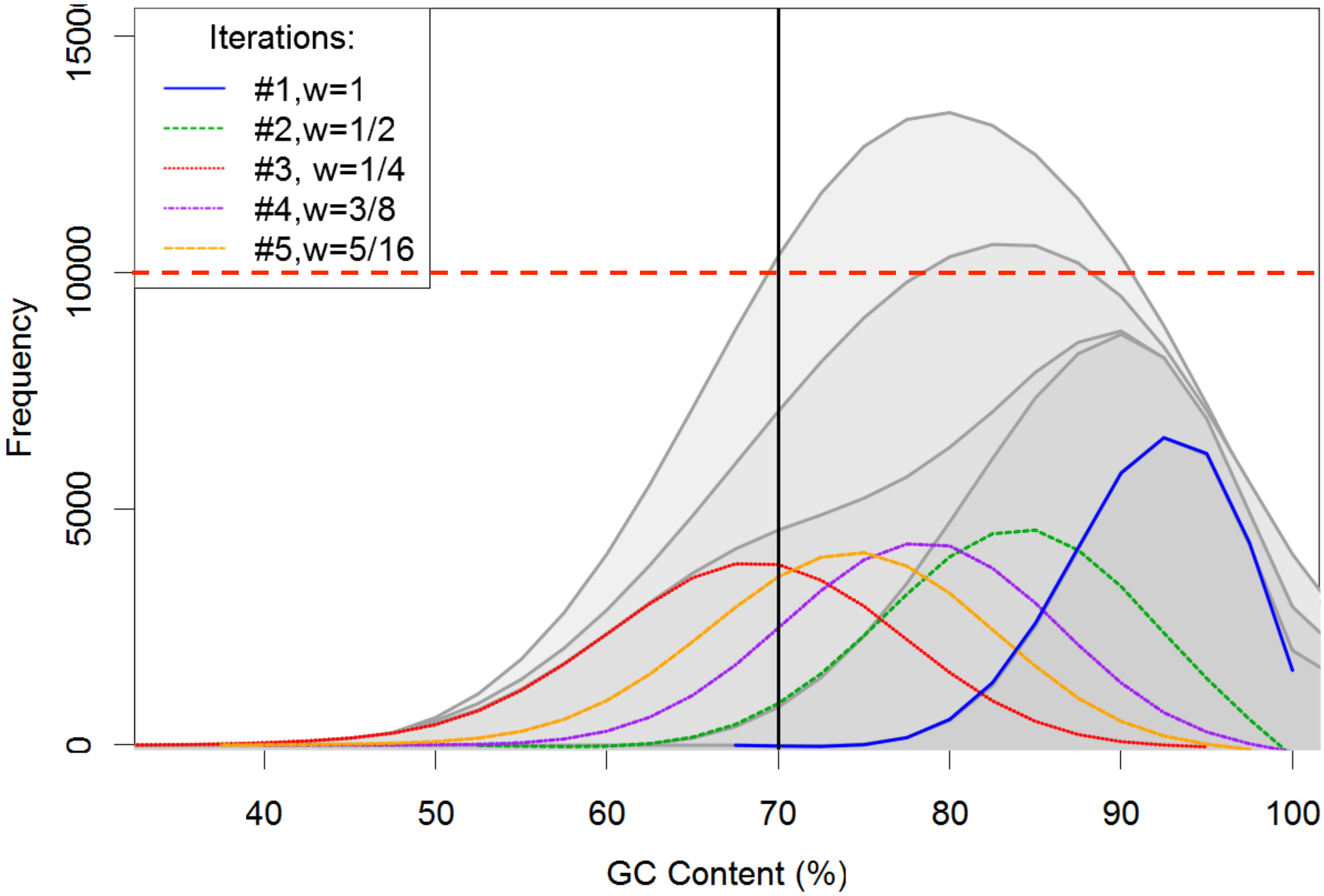
🟥 Unweighted sampling 🟦 weighted (**w**=1/2) 🟩 weighted (**w**=2)

Sampling pipe-line



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

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



■ Cumulative distribution

Technical details

- After rejection, the weights only impact the performance, not the probability (i.e. unbiased).
- Complexity $O(n^3 \cdot k^2 + m \cdot k \cdot n\sqrt{n} \cdot \log(n))$

  where *n* size, *k* #mutations, *m* #samples.
- Partition function can be written as a polynomial:

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

After *n* iterations we can calculate all a_i's and exactly solve the weight/C+G% relationship.

Remark: In practice, less iterations are necessary.

Outline

- **Computing the Mutational Landscape**
  (Waldispühl *et al.*, 2008)

- **Controlling the nucleotide distribution**
  (Waldispühl & Ponty, 2011)

- **Applications**
  (Lam *et al.*, 2011; Levin *et al.*, 2012; Reinharz *et al.*, 2013)

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

Applications

- Signature of evolutionary pressure - `RNAmutants`

(Waldispuhl *et al.*, 2008; Waldispühl & Ponty, 2011)

- Prediction of deleterious mutation - `corRna`

(Lam *et al.*, 2011)

- Design of RNA with target structure - `RNAensign`

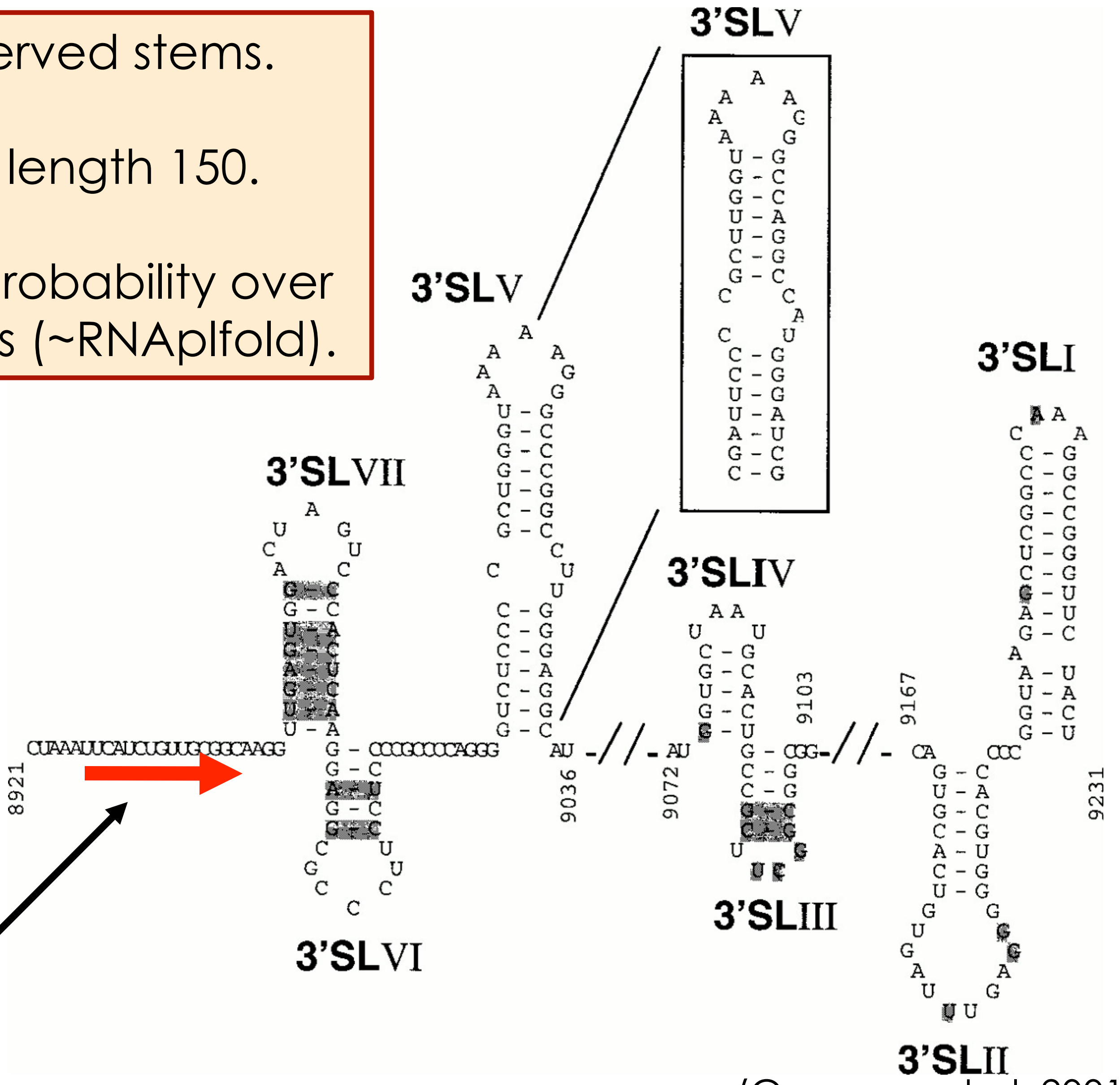
(Levin *et al.*, 2012)

- Error correction in NGS data - `RNApyro`

(Reinharz *et al.*, 2013)

Scan of GB virus C

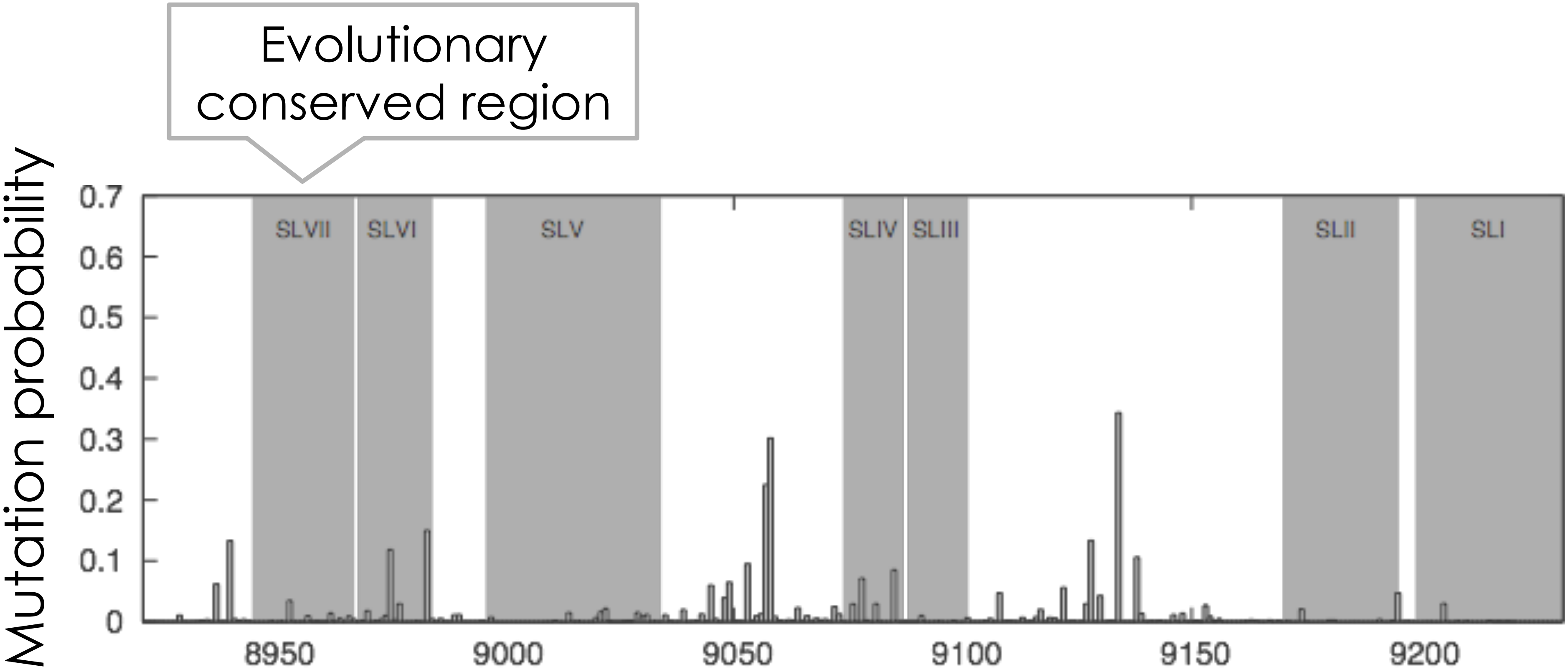
- 7 evolutionary conserved stems.
- Scan using frame of length 150.
- Average mutation probability over all overlapping frames (~RNAplfold).



Open frame

(Cucenau et al.,2001)

Scan of GB virus C

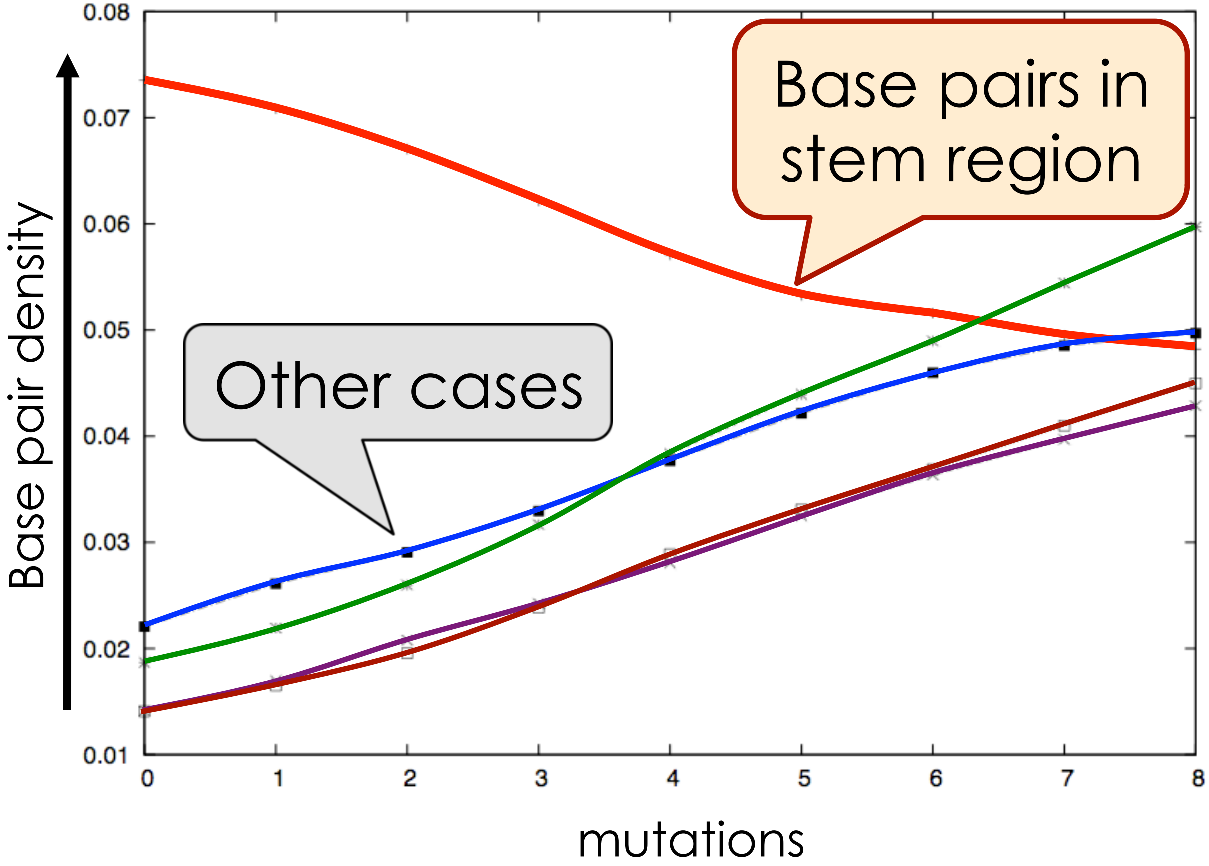


Results: Energetically favorable mutations are distributed outside the evolutionary conserved regions.

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

Scan of GB virus C

Base pair density in evolutionary conserved regions

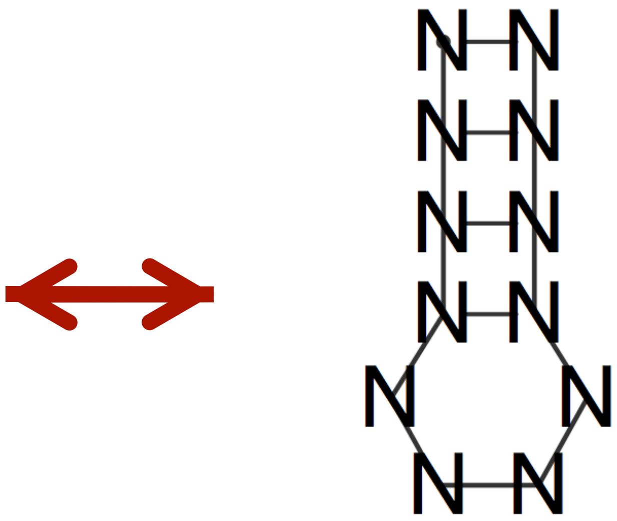


Results: Mutations decrease the base pair density in evolutionary conserved stem regions.

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

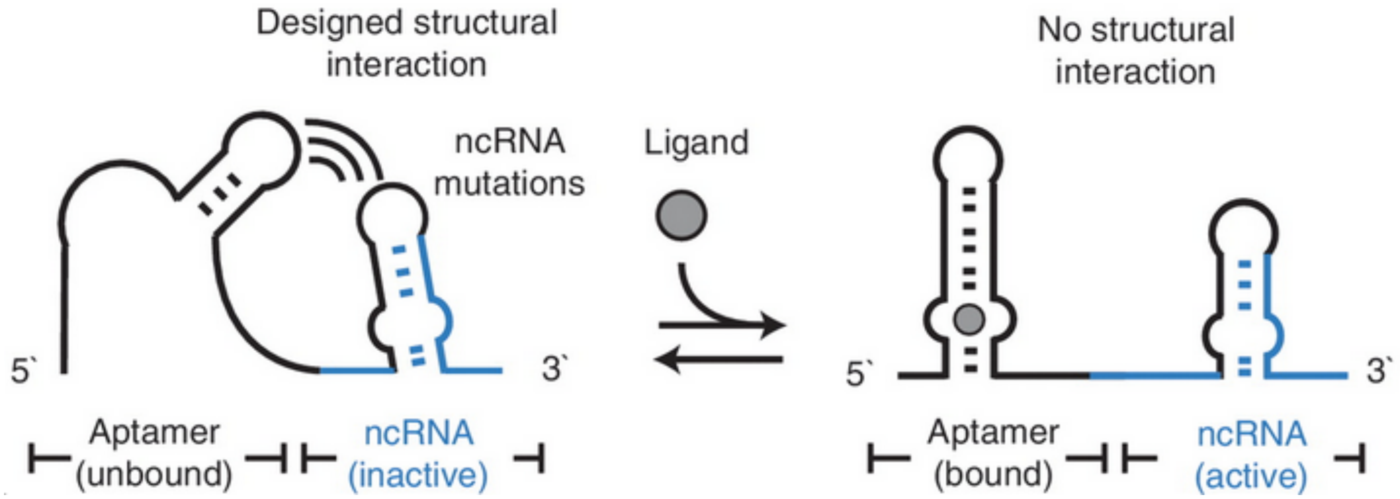
RNA secondary structure design

UCGGAG?GCCCGA



Heavily studied area: RNAinverse, RNA-SSD, INFO-RNA, …

Motivations



(Qi *et al.*, 2012)

- Designing new molecular functions
- Re-engineering existing RNAs
- RNA computing

Motivations

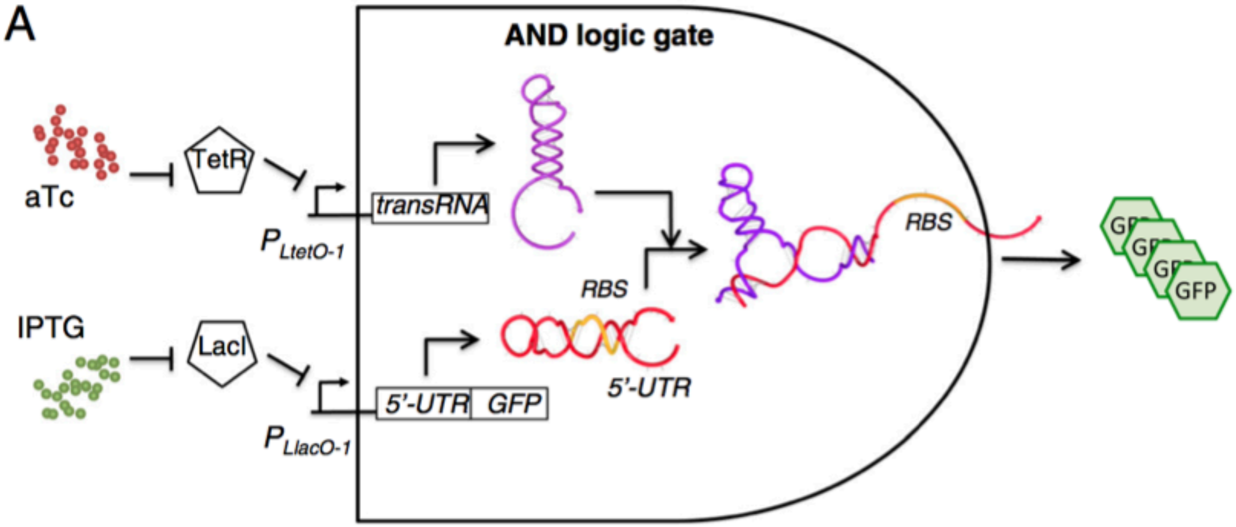
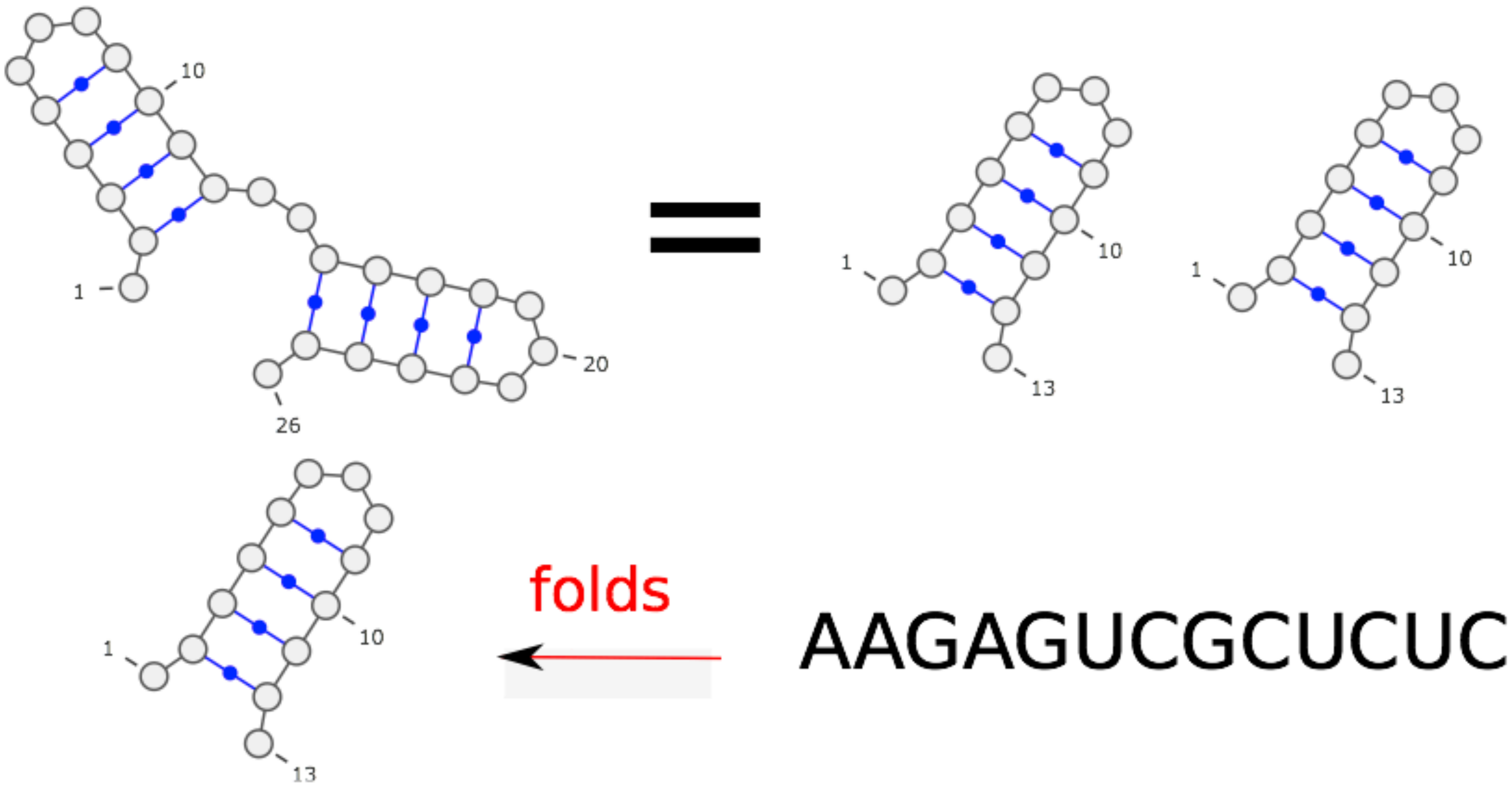


Figure : Rodrigo et al. 2012

- Designing new molecular functions
- Re-engineering existing RNAs
- RNA computing

Local Design



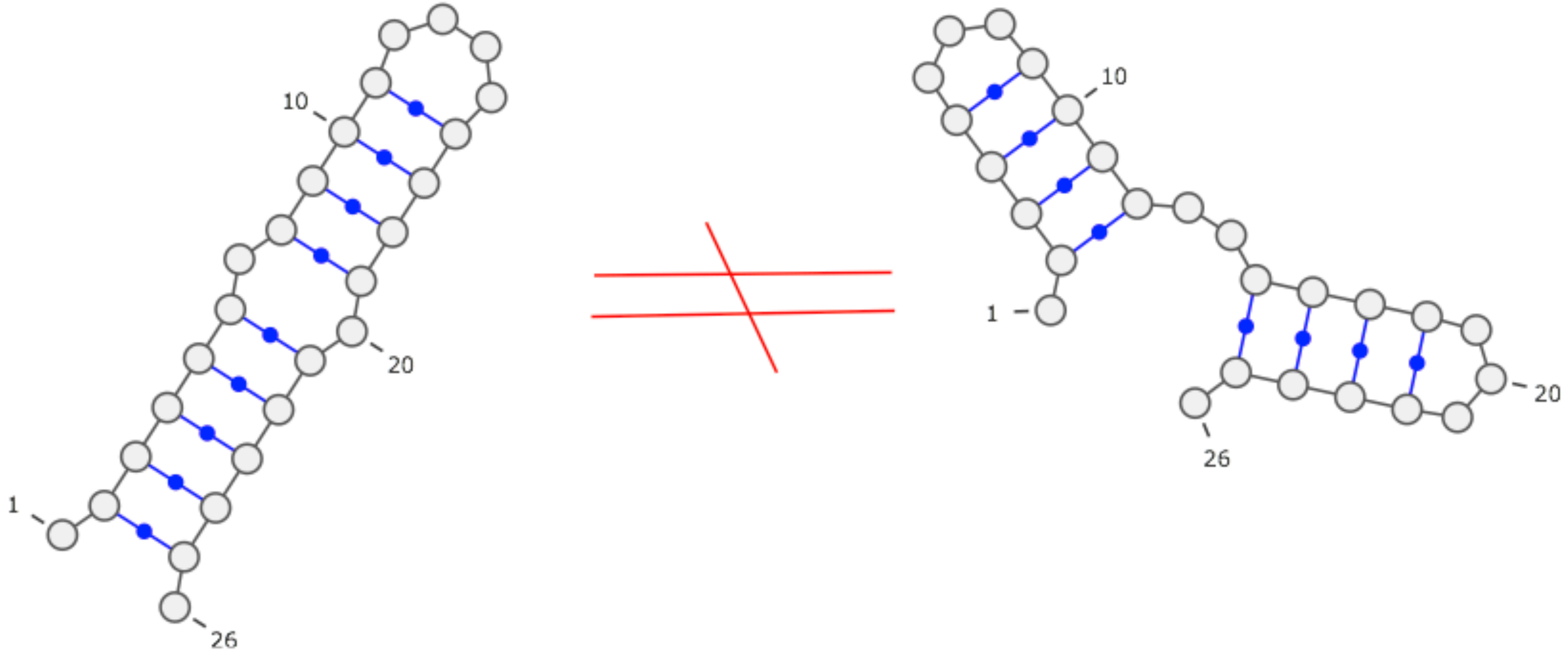
folds

AAGAGUCGCUCUC

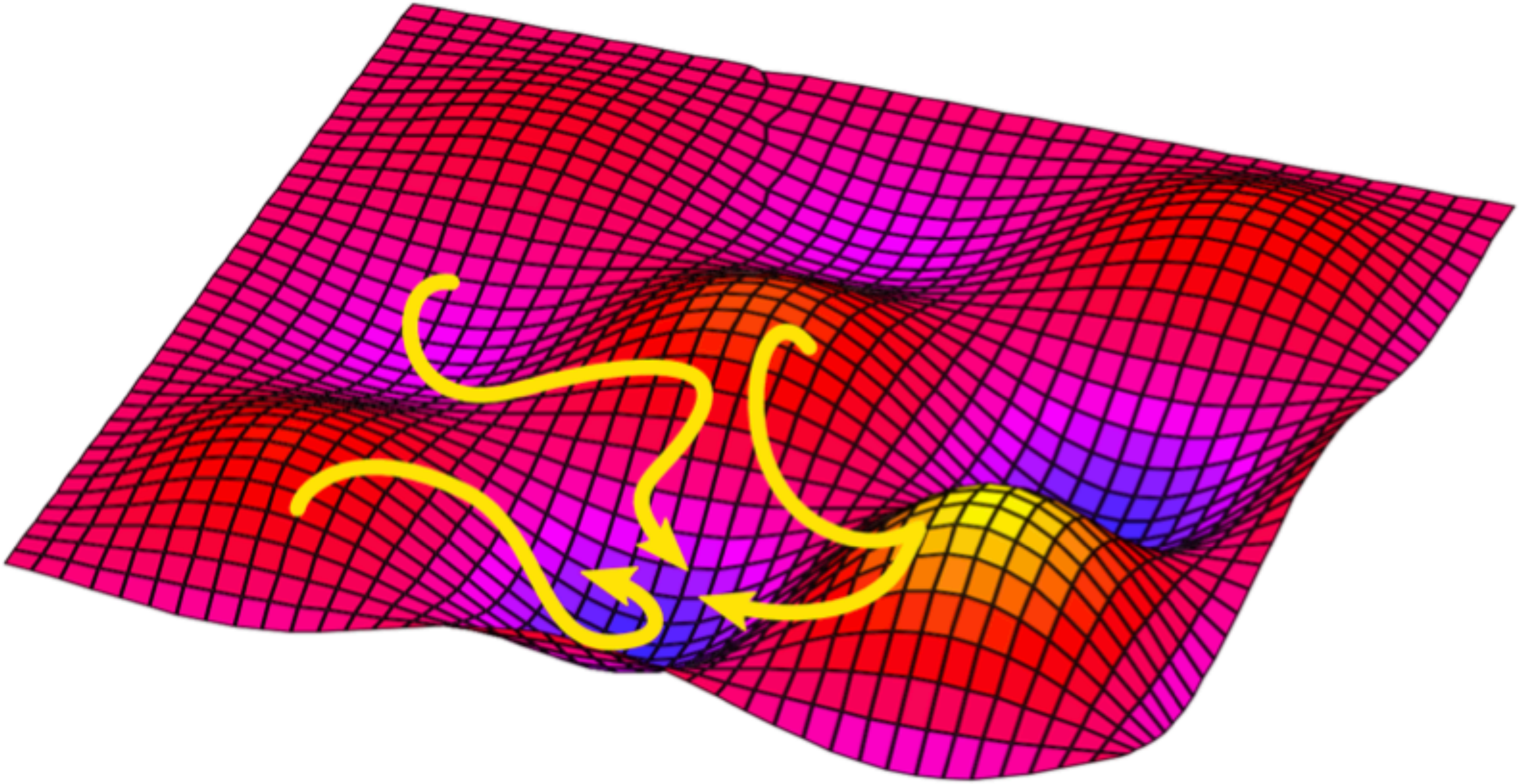
Local Design

AAGAGUCGCUCUCAAGAGUCGCUCUC

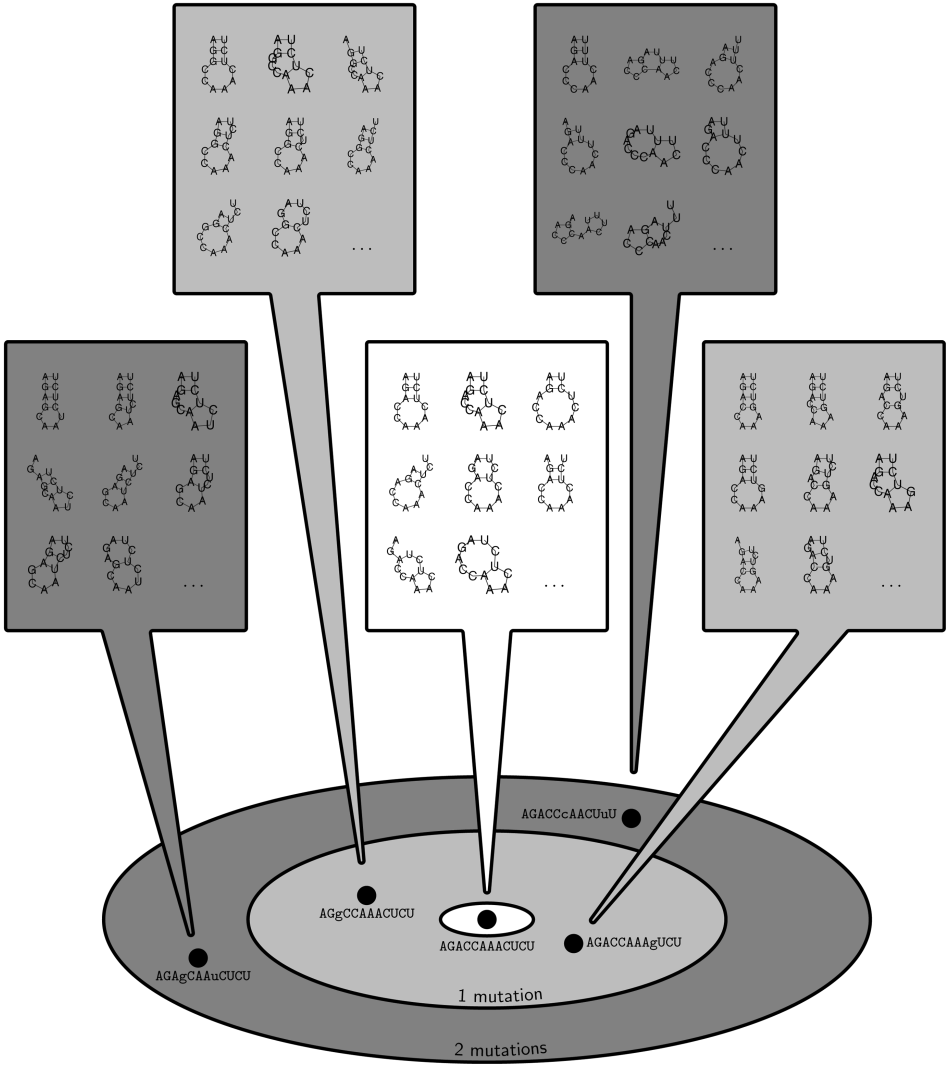
Folds



Local vs Global vs "Glocal"



RNA-ensign: Designing RNAs with RNAmutants



1. Select a random seed
2. Sample mutants from k-neighborhood with RNAmutants
3. Select sample with best fit to target

RNAensign

Our approach: global search strategy
(vs. local search heuristics)

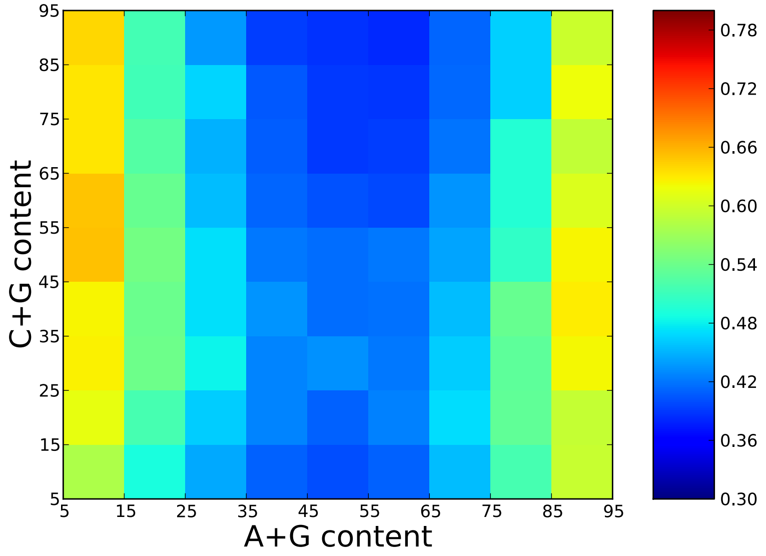
Objectives:

- How important is the choice of the seed ?
- Can we minimize the number of mutations ?
- Can we develop better design algorithm ?

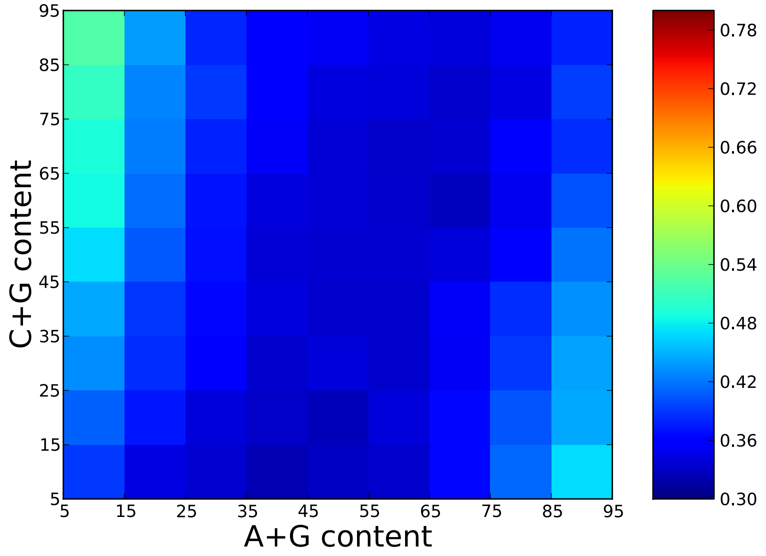
(Levin *et al.*, 2012)

Influence of the seed on the target stability

RNAmutants (global search)



RNAinverse (local search)

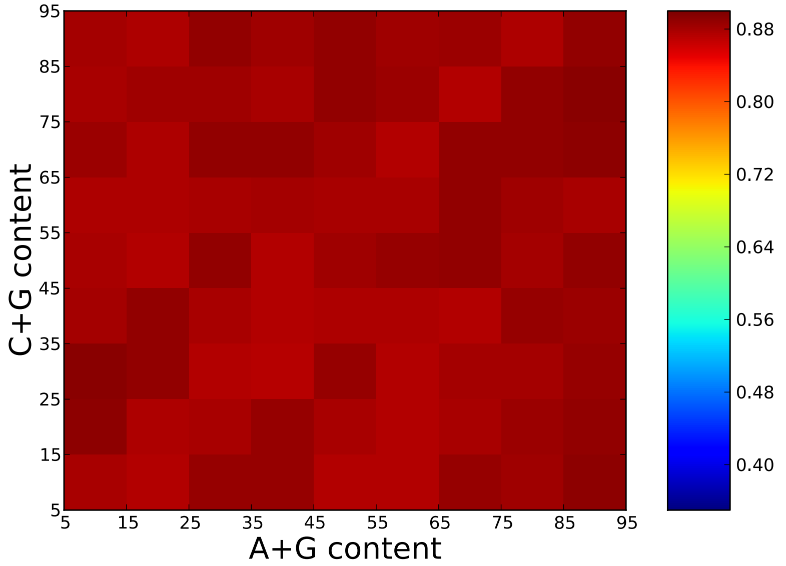


- 10 seeds with fized A+G and C+G content
- 100 structures generated using GenRGenS
- Average probability of the target structure on designed sequence.

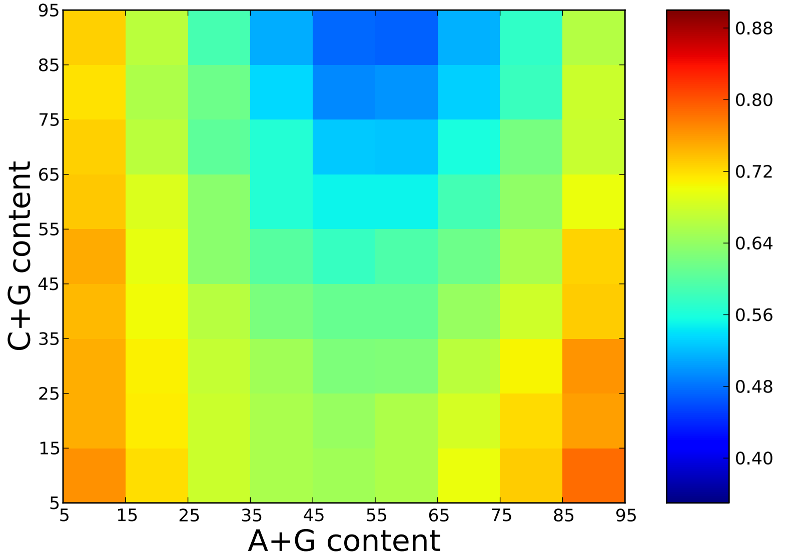
(Levin *et al.*, 2012)

Influence of the seed on the success rate

RNAmutants (global search)



RNAinverse (local search)



- 10 seeds with fized A+G and C+G content
- 100 structures generated using GenRGenS
- Average success rate.

BUT…

(Levin *et al.*, 2012)

Influence of the seed

	Probability			Entropy			Time		
Size	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
0-40	0.69	0.65	0.60	0.056	0.051	0.065	62	28	61
41-80	0.35	0.21	0.53	0.148	0.157	0.100	1883	742	711
81+	0.40	0.30	0.29	0.062	0.147	0.125	9332	2434	1269

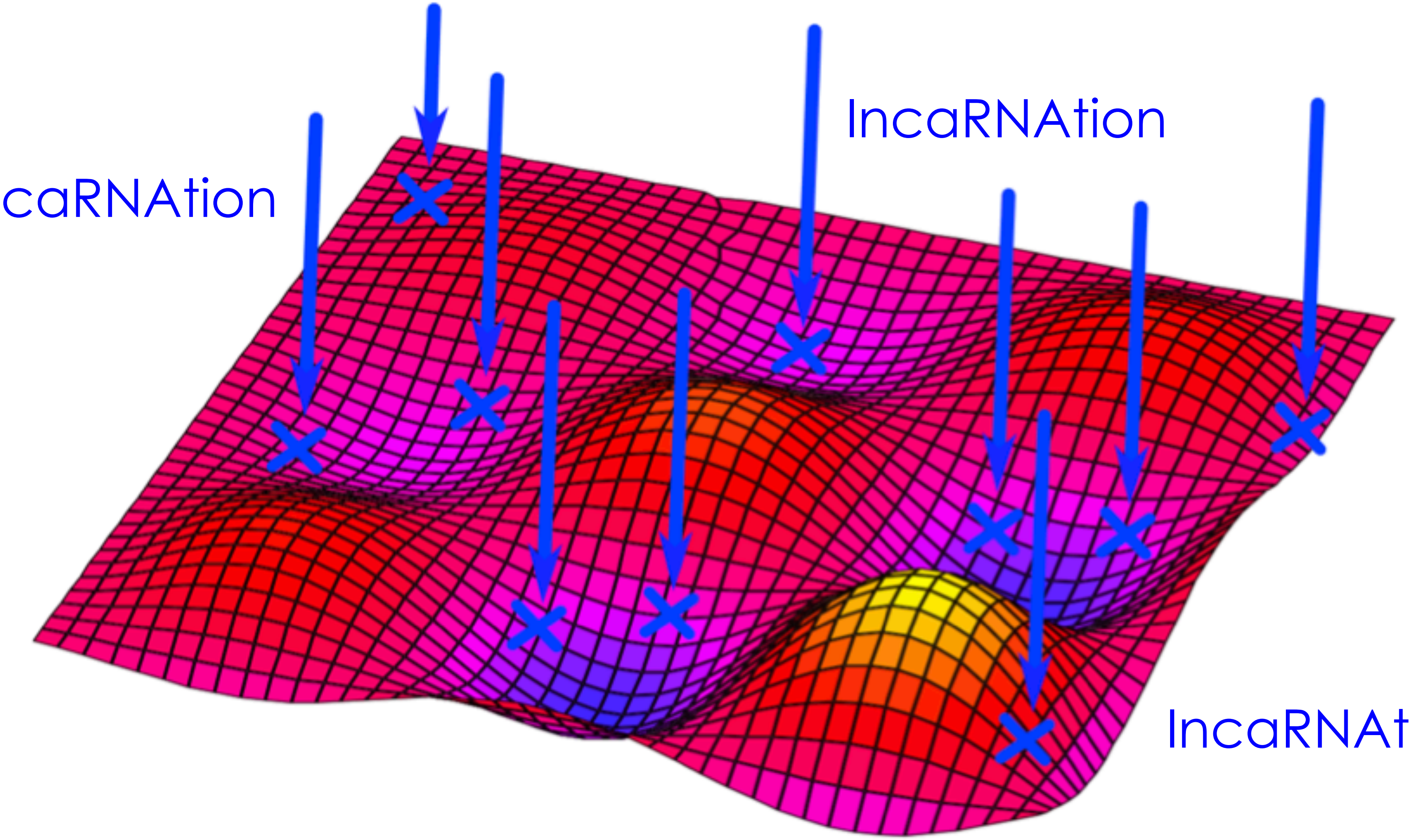
A: RNAmutants
B: RNAmutants with 50% of mutations
C: 10,000 runs of RNAinverse

Global search may has benefits for large structure **but** is computationally expensive.

(Levin *et al.*, 2012)

Local vs Global vs "Glocal"

Generate seed sequences with IncaRNAtion (Global search)



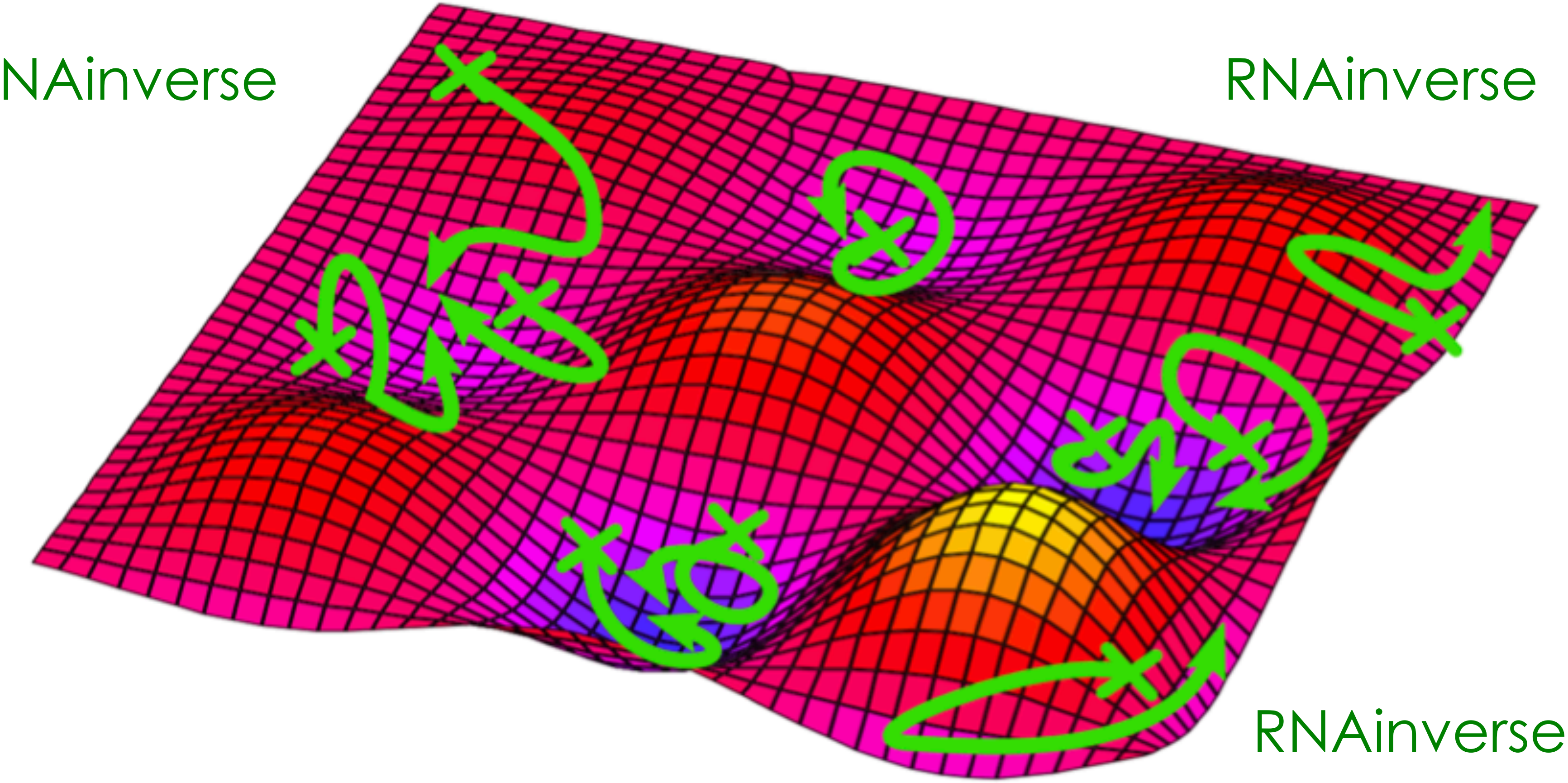
IncaRNAtion

IncaRNAtion

IncaRNAtion

Local vs Global vs "Glocal"

Optimize IncaRNAtion seeds with RNAinverse (local search)



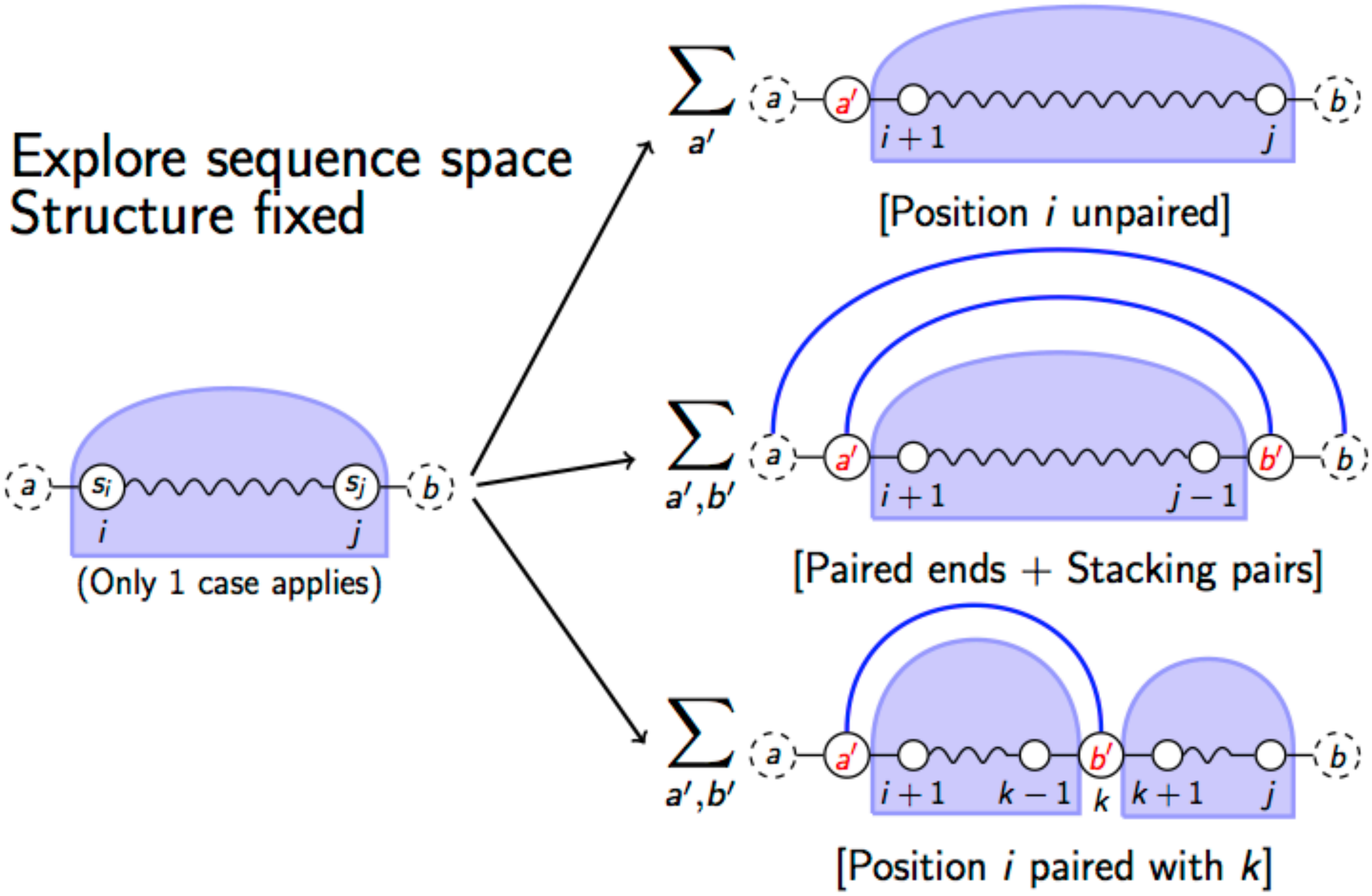
RNAinverse

RNAinverse

RNAinverse

DP Recursion
global

Explore sequence space
Structure fixed



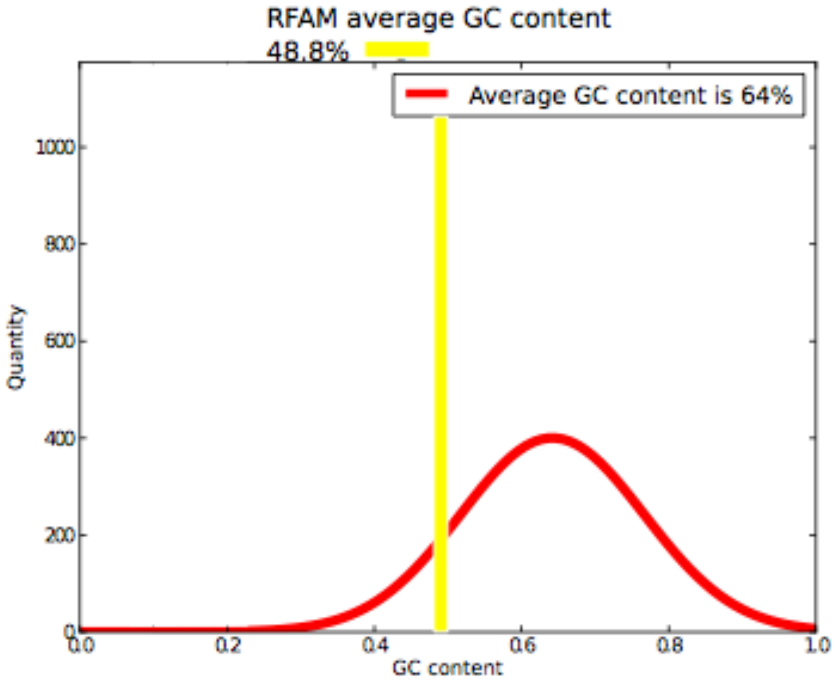
(Only 1 case applies)

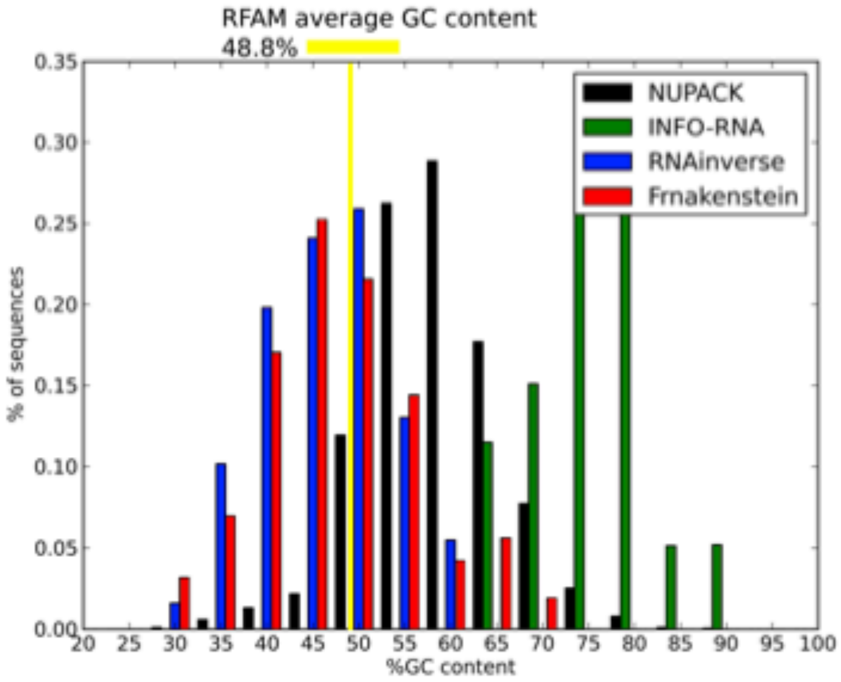
[Position i unpaired]

[Paired ends + Stacking pairs]

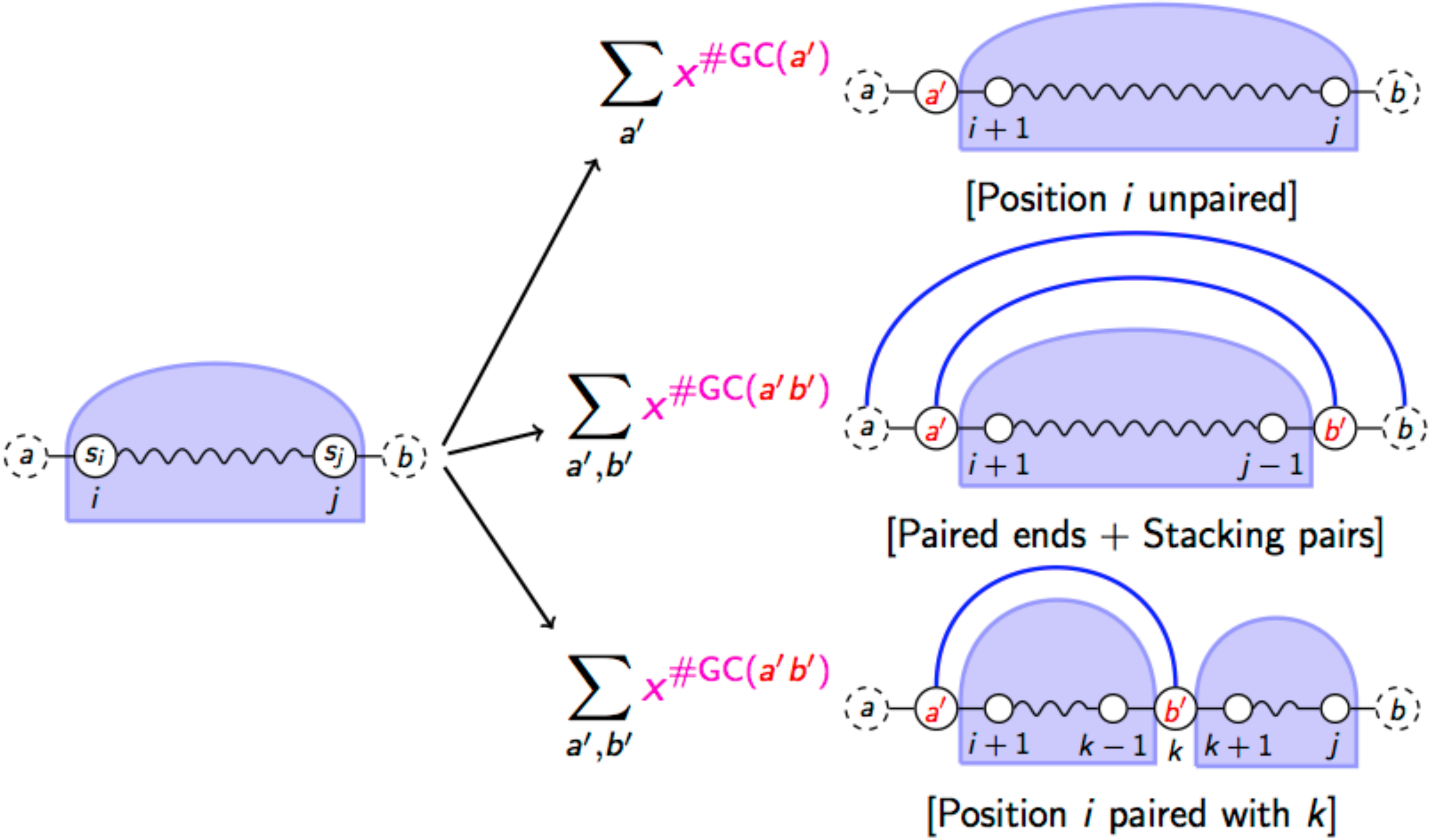
[Position i paired with k]

GC Bias


RFAM average GC content
48.8%
Average GC content is 64%
Quantity
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
GC content


RFAM average GC content
48.8%
NUPACK
INFO-RNA
RNAinverse
Frnakenstein
% of sequences
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
%GC content

Weighted DP Recursion
global



$$\sum_{a'} x^{\#GC(a')}$$

[Position *i* unpaired]

$$\sum_{a',b'} x^{\#GC(a'b')}$$

[Paired ends + Stacking pairs]

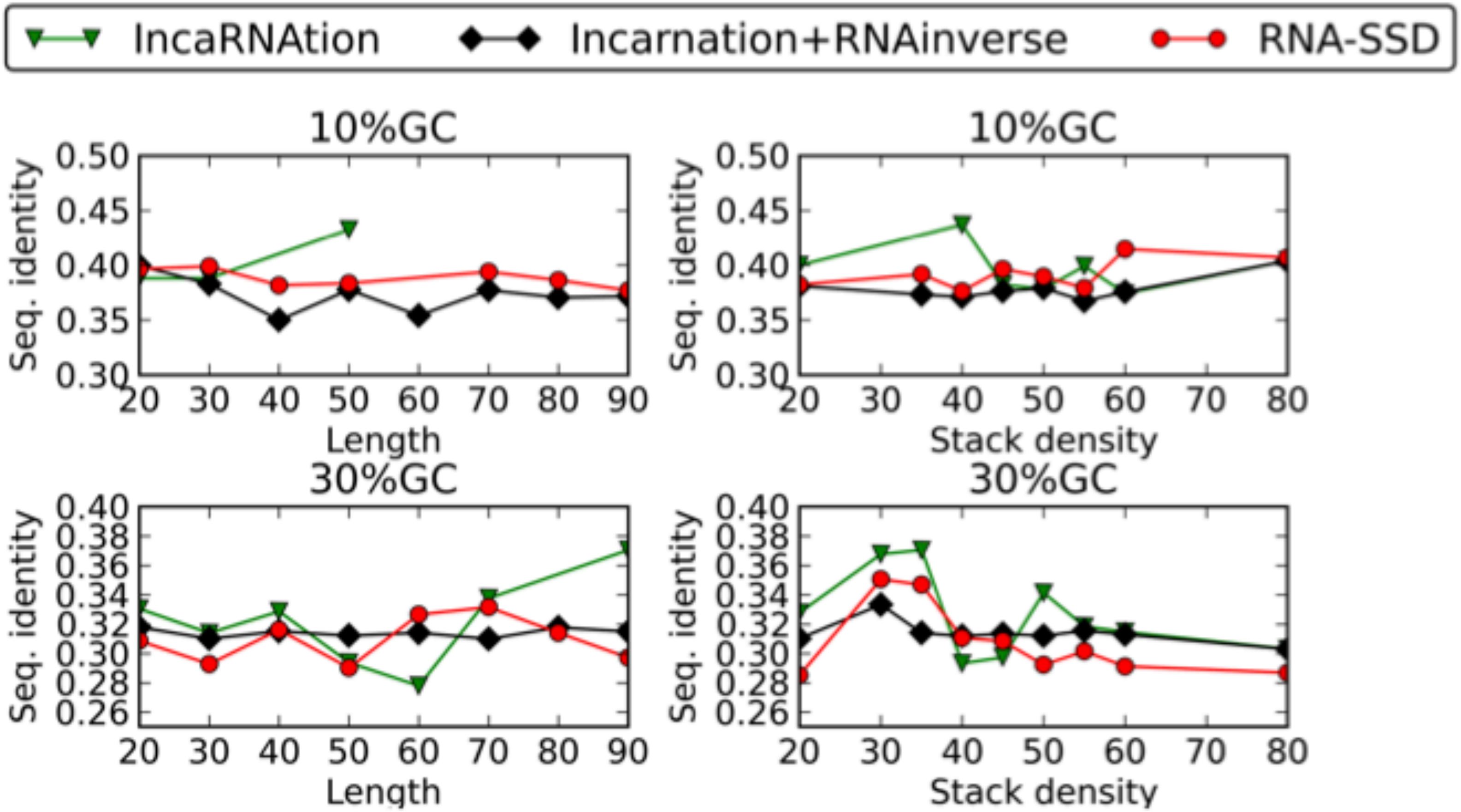
$$\sum_{a',b'} x^{\#GC(a'b')}$$

[Position *i* paired with *k*]

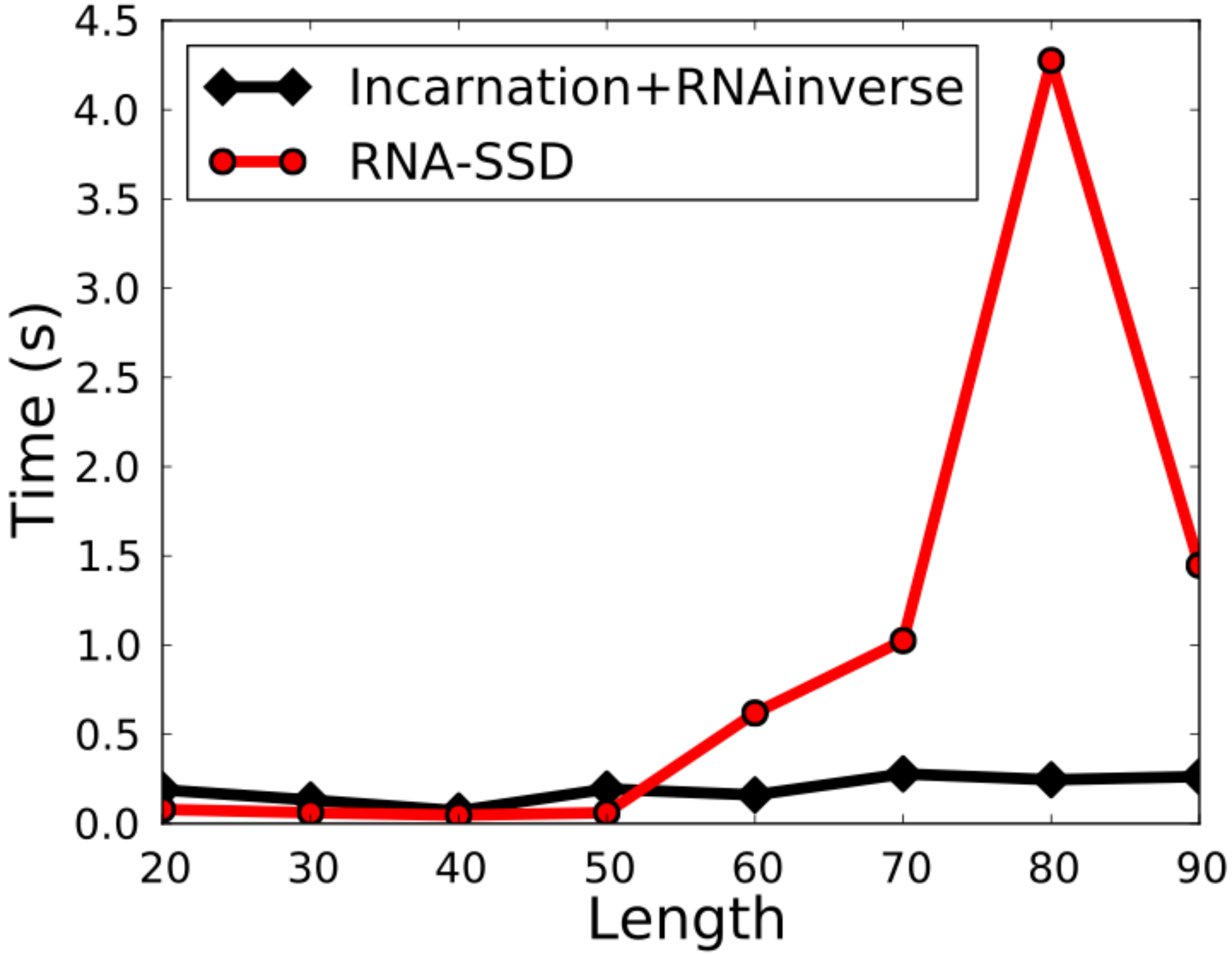
IncaRNAtion

RNA-ensign[1]	IncaRNAtion
Seeded	No Seed
Explore mutant space	Explore **full** sequence space
$\mathcal{O}(n^5)$	$\mathcal{O}(n)$
Complex energy model	Simple energy model
	Sequence constraints

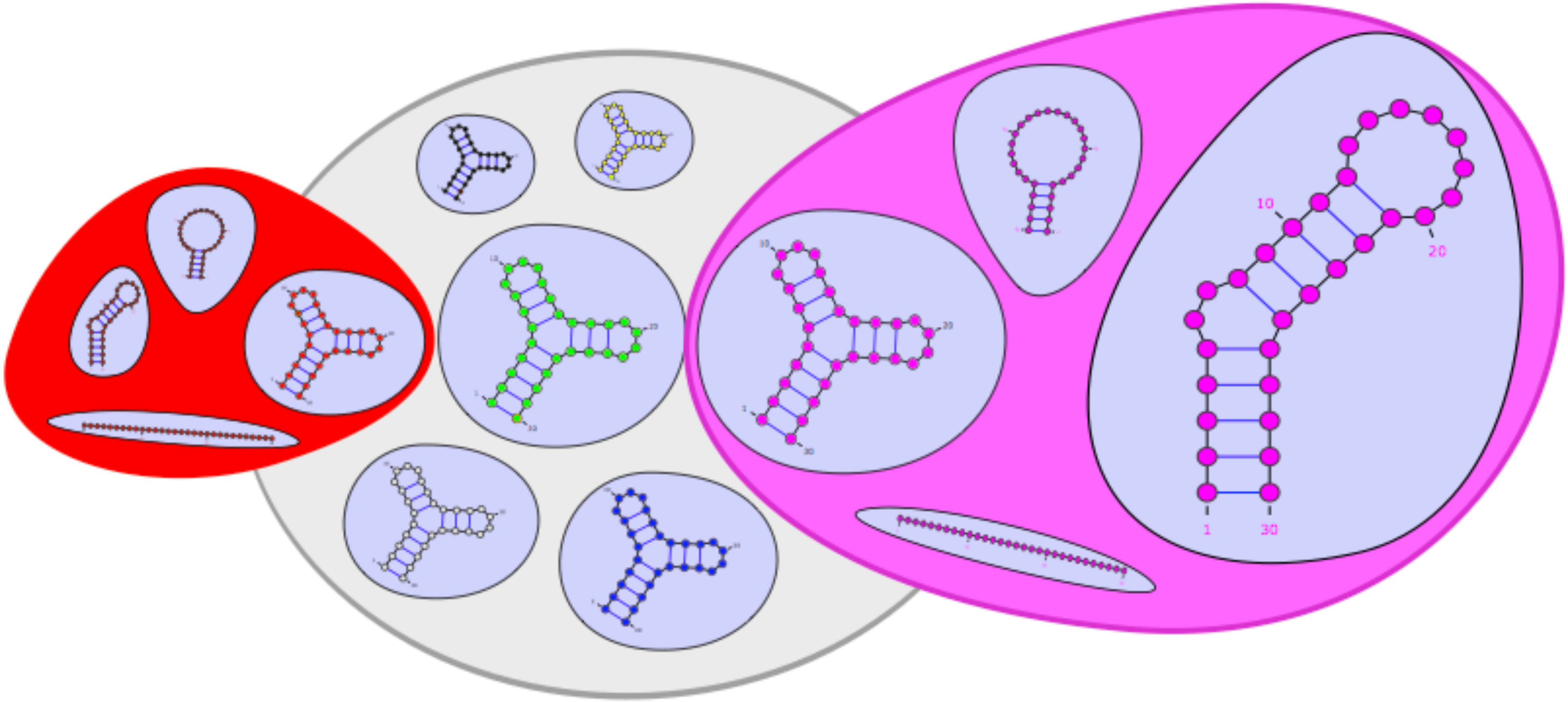
Incarnation + RNAinverse Results
Sequence identity



Incarnation + RNAinverse time



Affinity $\neq$ Specificity



Formally, a large affinity neither ensures preferential fold into a target, nor is it a necessary condition. . .

Acknowledgments

McGill
- Anwar Asbah
- David Becerra
- Carlos Gonzales
- Alfred Kam
- Edmund Lam
- Vladimir Reinharz

Ecole Polytechnique
- Yann Ponty
- Jean-Marc Steayert

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

Boston College
- Peter Clote

Google Inc.
- Behshad Behzadi

Fonds de recherche
sur la nature
et les technologies
Québec

NSERC
CRSNG

Would you like to know more?

- J. Waldispühl *et al.* (2008), **Efficient Algorithms for Probing the RNA Mutation Landscape**, Plos Comp. Bio.

- J. Waldispühl and Y. Ponty (2011), **An Unbiased Sampling Algorithm for the Exploration of RNA Mutational Landscape Under Evolutionary Pressure**, RECOMB.

- Levin *et al.* (2012), **A global sampling approach to designing and reengineering RNA secondary structures**, NAR.

- Reinharz *et al.* (2013), **A linear inside-outside algorithm for correcting sequencing errors in structured RNA sequences**, RECOMB.

- Reinharz *et al.* (2013), **A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotides distribution**, ISMB.

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