AMIB(io): **A**lgorithms and **M**odels for **I**ntegrative **Bio**logy

Resp.: Yann Ponty

AMIBio team
Ecole Polytechnique
CNRS UMR 7161 LIX

June 15, 2020

Members

Yann Ponty CRCN
(Head of AMIBio)

Philippe Chassignet MCF

Cédric Chauve PR
(Visiting prof.)

Mireille Régnier DR

Jean-Marc Steyaert PR
(Aemeritus)

Julie Bernauer CR
2014–2015 → nVidia

Postdoc: Christelle Rovetta 2017, Afaf Saaidi 2018

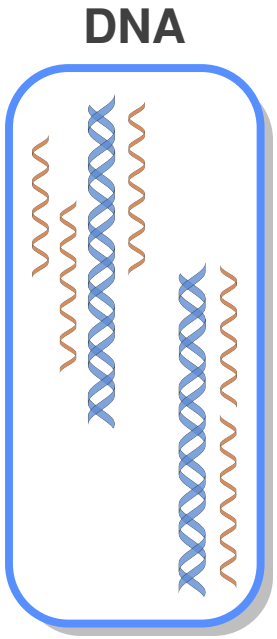
PhD Students:

Jorgelindo Da Veiga – 2016
Juraj Michalik – 2016
Hua-Ting Yao – 2018

Ha Nguyen Ngoc – 2017
Pauline Pommeret – 2017

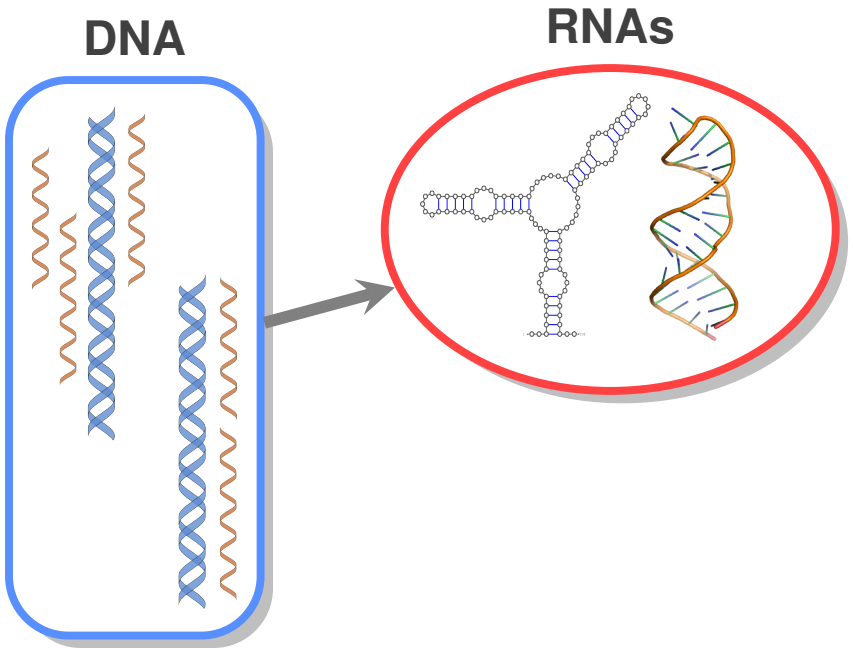
9 defended PhDs over evaluation period

Biological context



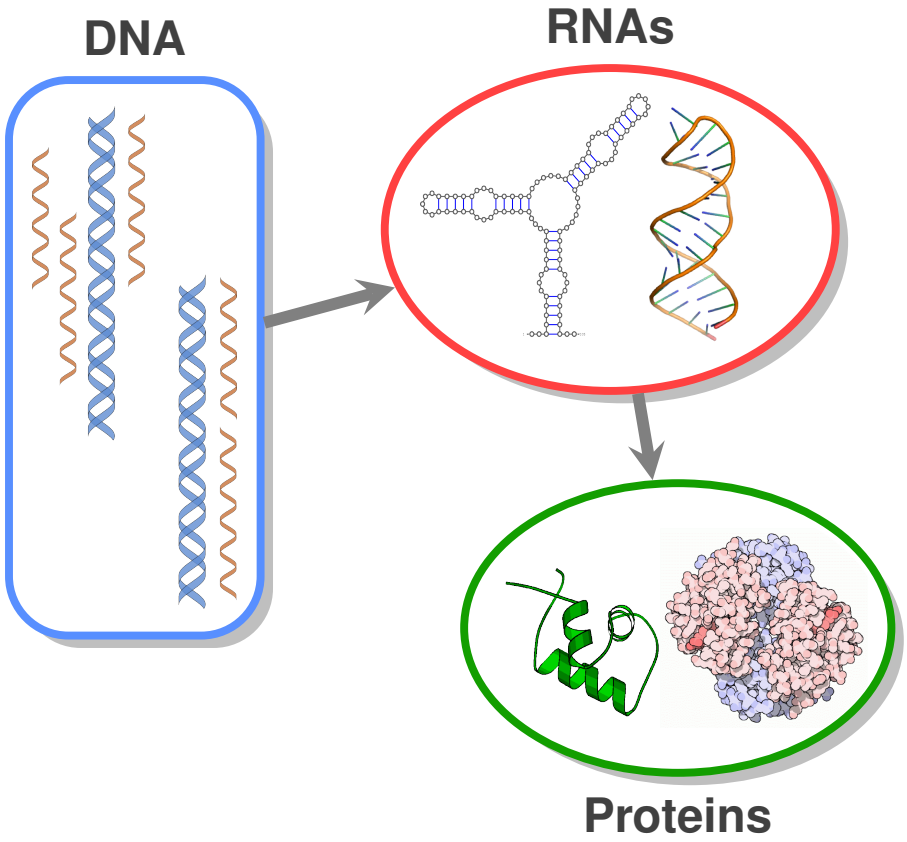
Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



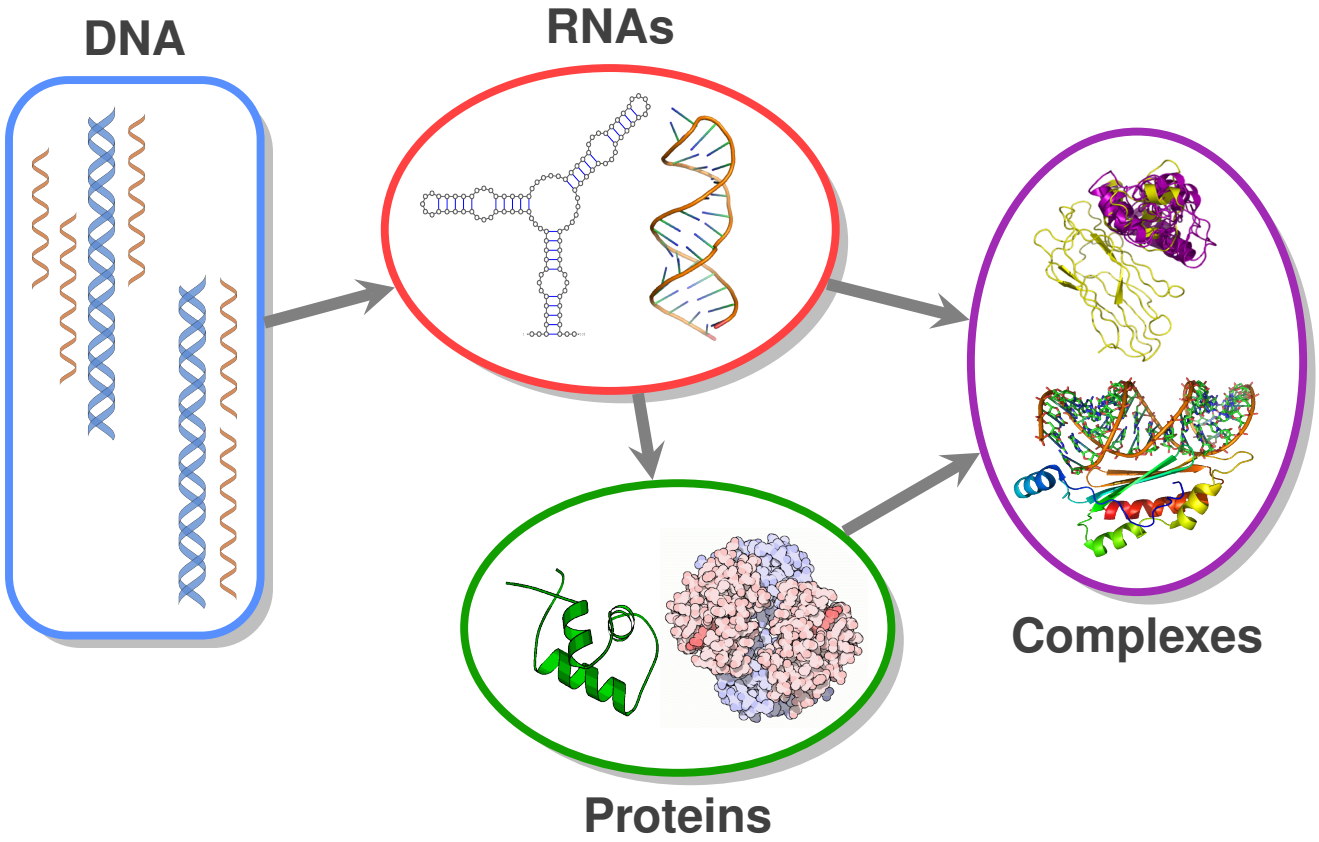
Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



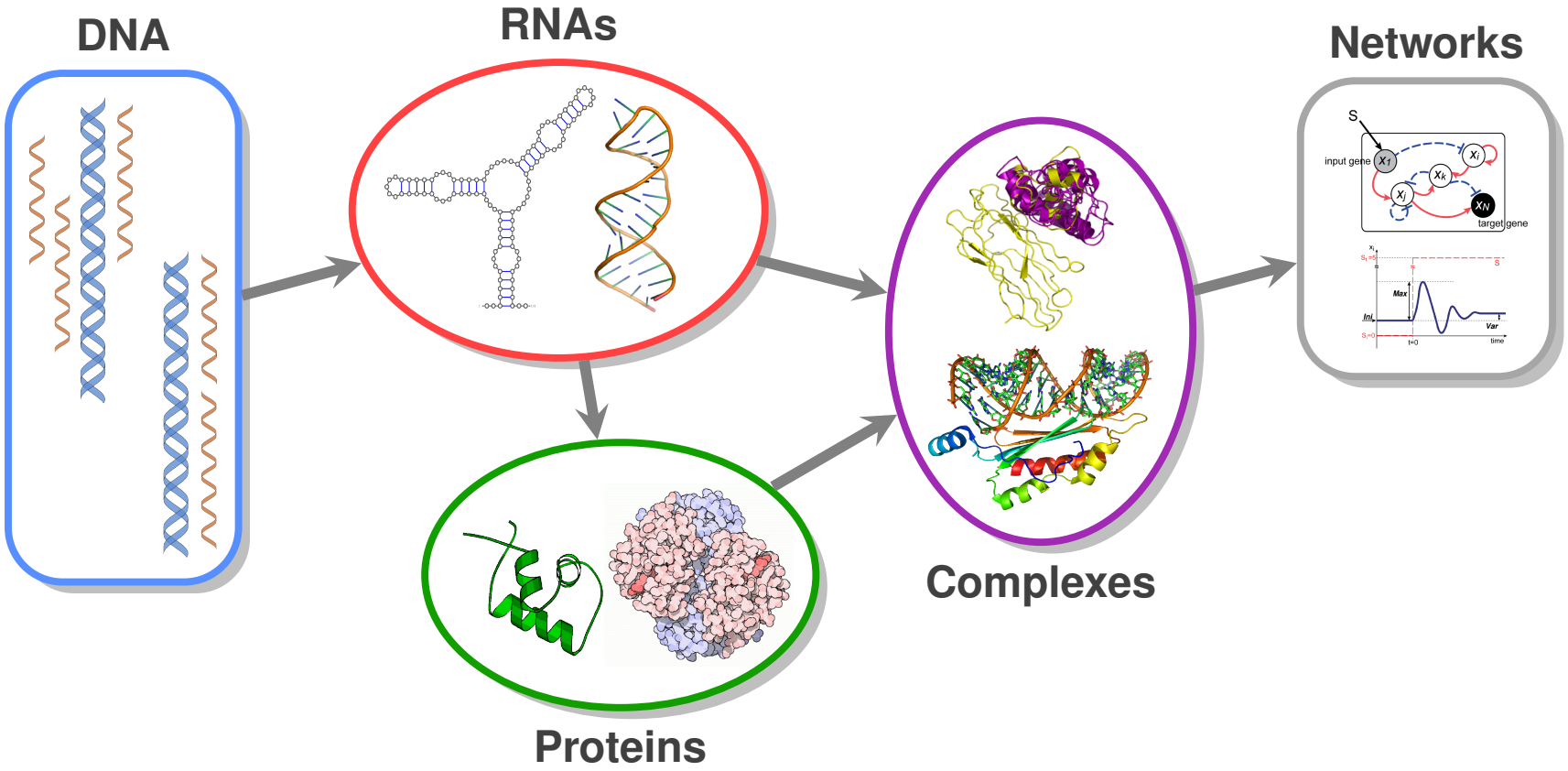
Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



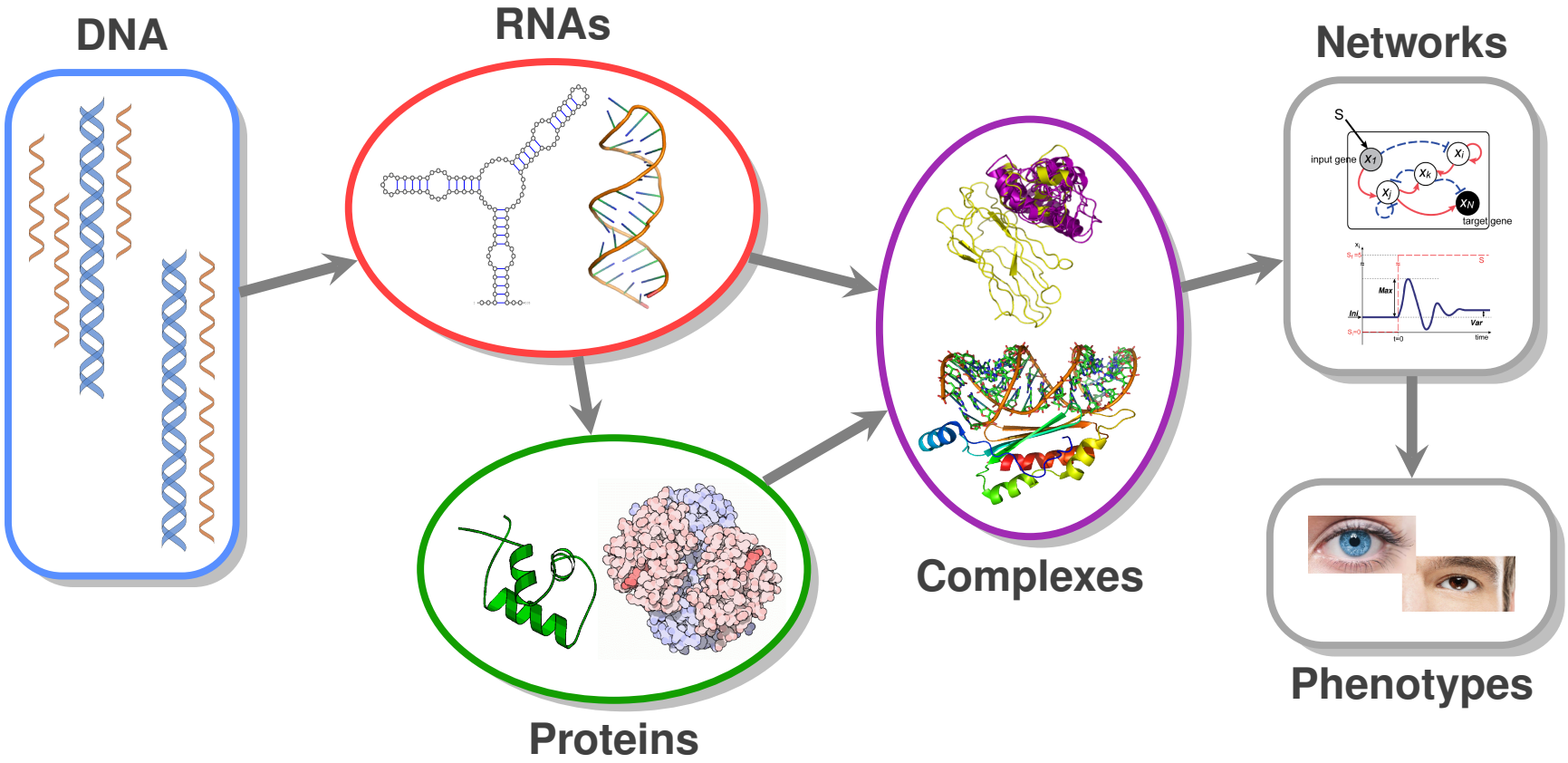
Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



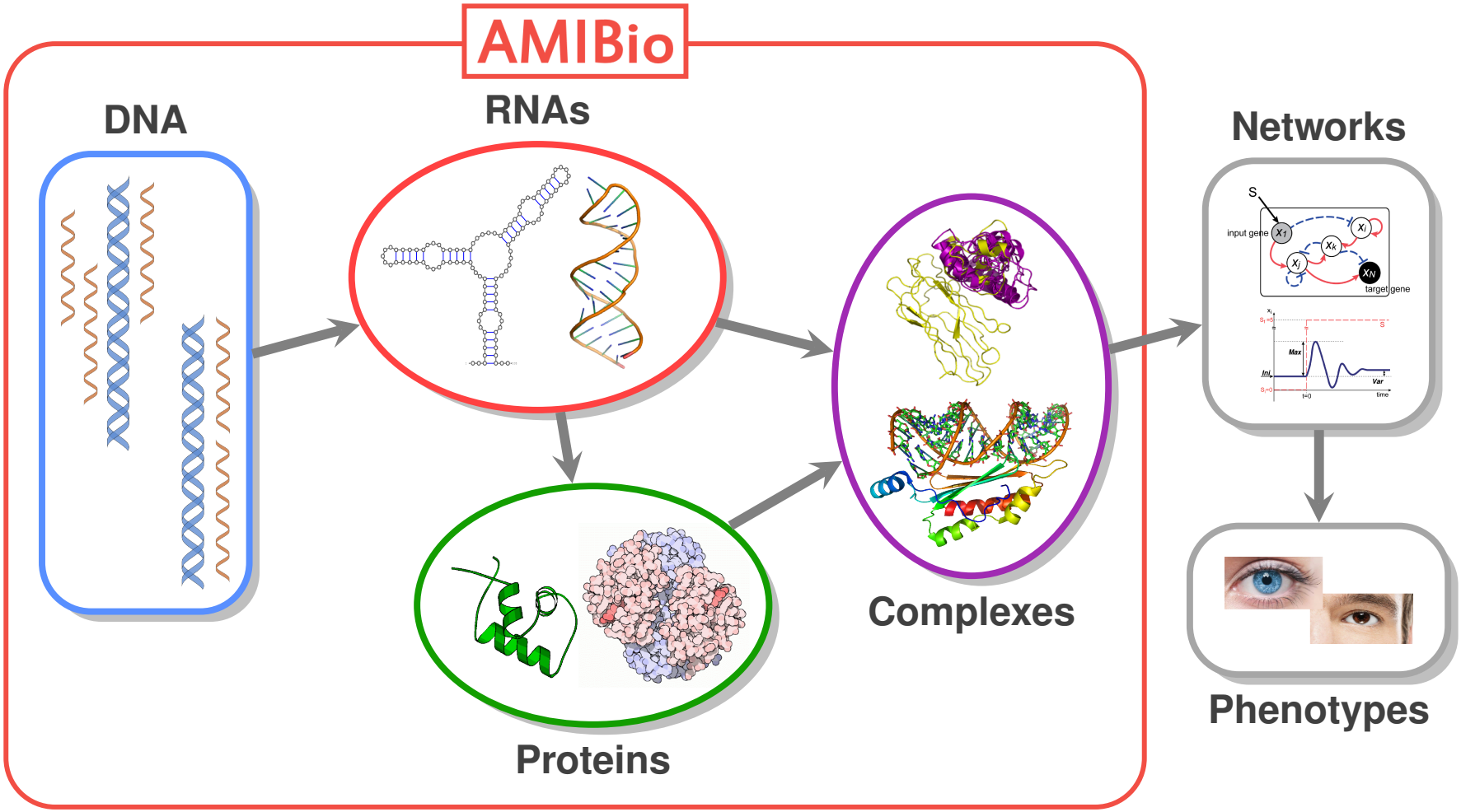
Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



Algorithms and **M**odels for **I**ntegrative **B**iology

Biological context



Algorithms and **M**odels for **I**ntegrative **B**iology

Predictive Bioinformatics

- Regulatory motifs in genomic datasets

···CGUCAGCUAGCGCAUCG···ACGCAAGCUAGCGCUCGU···

···AAUAUUUAAAUAUACGA···AUUAAAUAUAGAUAUUUUAAA···

- Find repeated motifs
- $\mathbb{P}(|\text{Longest repetition}| \geq 9 \mid H_0)$???

Predictive Bioinformatics

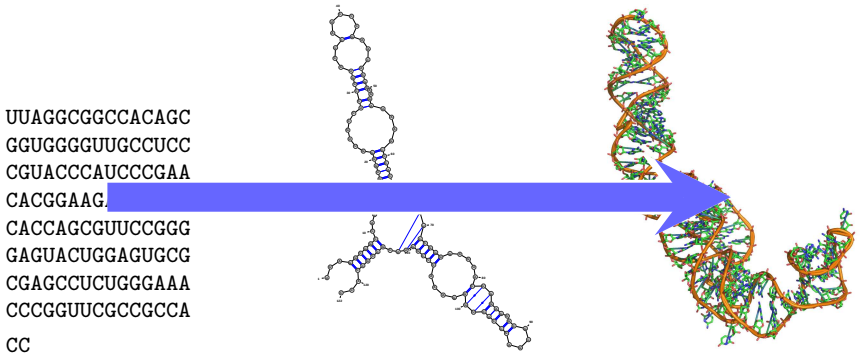
- Regulatory motifs in genomic datasets

···CGUCAGCUAGCGCAUCG···ACGCAAGCUAGCGCUCGU···

···AAUAUUUAAAUAUACGA···AUUAAAUAUAGAUAUUUUAAA···

  - Find repeated motifs
  - $\mathbb{P}(|\text{Longest repetition}| \geq 9 \mid H_0)$???

- RNA folding



  - Energy minimization
  - Thermodynamic equilibrium
  - Comparative modeling
  - Folding kinetics

Predictive Bioinformatics

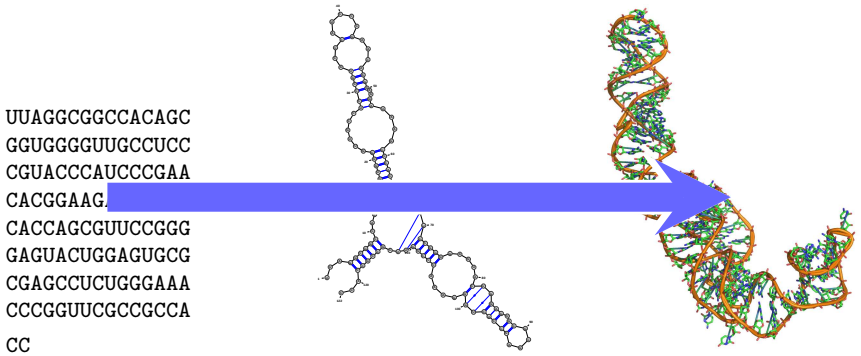
- Regulatory motifs in genomic datasets

···CGUCAGCUAGCGCAUCG···ACGCAAGCUAGCGCUCGU···

···AAUAUUUAAAUAUACGA···AUUAAAUAUAGAUAUUUUAAA···

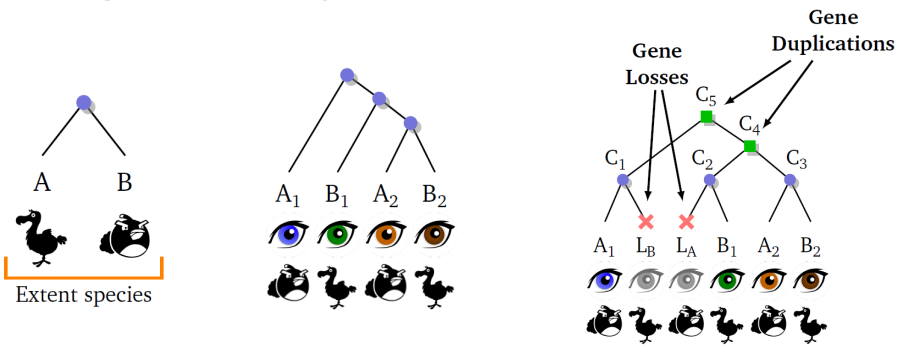
  - Find repeated motifs
  - $\mathbb{P}(|\text{Longest repetition}| \geq 9 \mid H_0)$???

- RNA folding



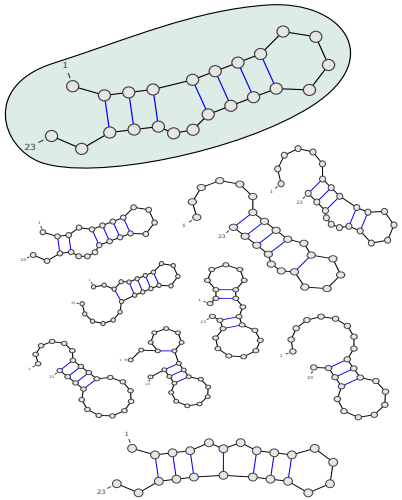
  - Energy minimization
  - Thermodynamic equilibrium
  - Comparative modeling
  - Folding kinetics

- Comparative genomics



  - Parsimonious reconciliation
  - Ancestral features
  - Probabilistic support

Ensemble analysis



Algorithms

Dynamic Programming

Sampling
Random Generation

Clustering

Best Solution
(Parsimonious/Optimal)

Probabilistic Support

Maximum Expected Accuracy

Robustness to errors

Centroid Structures

Combinatorial Specifications

Models

Analytic Combinatorics

Generating Functions

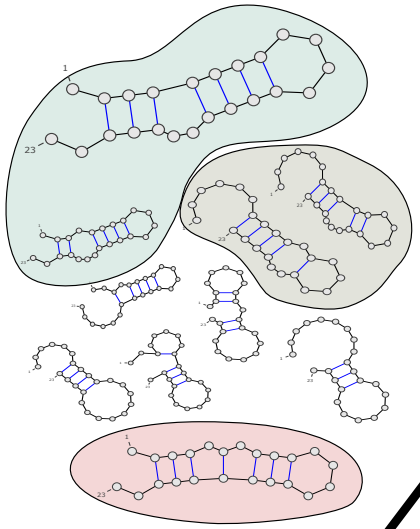
Transfer Theorems

Significance scores

Asymptotic properties

- [TCS'14'16] [Bioinformatics'17'17'17] [BMC Bioinfo'14] [RECOMB'18]...

Ensemble analysis



Algorithms

Dynamic Programming

Sampling
Random Generation

Clustering

Best Solution
(Parsimonious/Optimal)

Probabilistic Support

Maximum Expected Accuracy

Robustness to errors

Centroid Structures

Combinatorial Specifications

Models

Analytic Combinatorics

Generating Functions

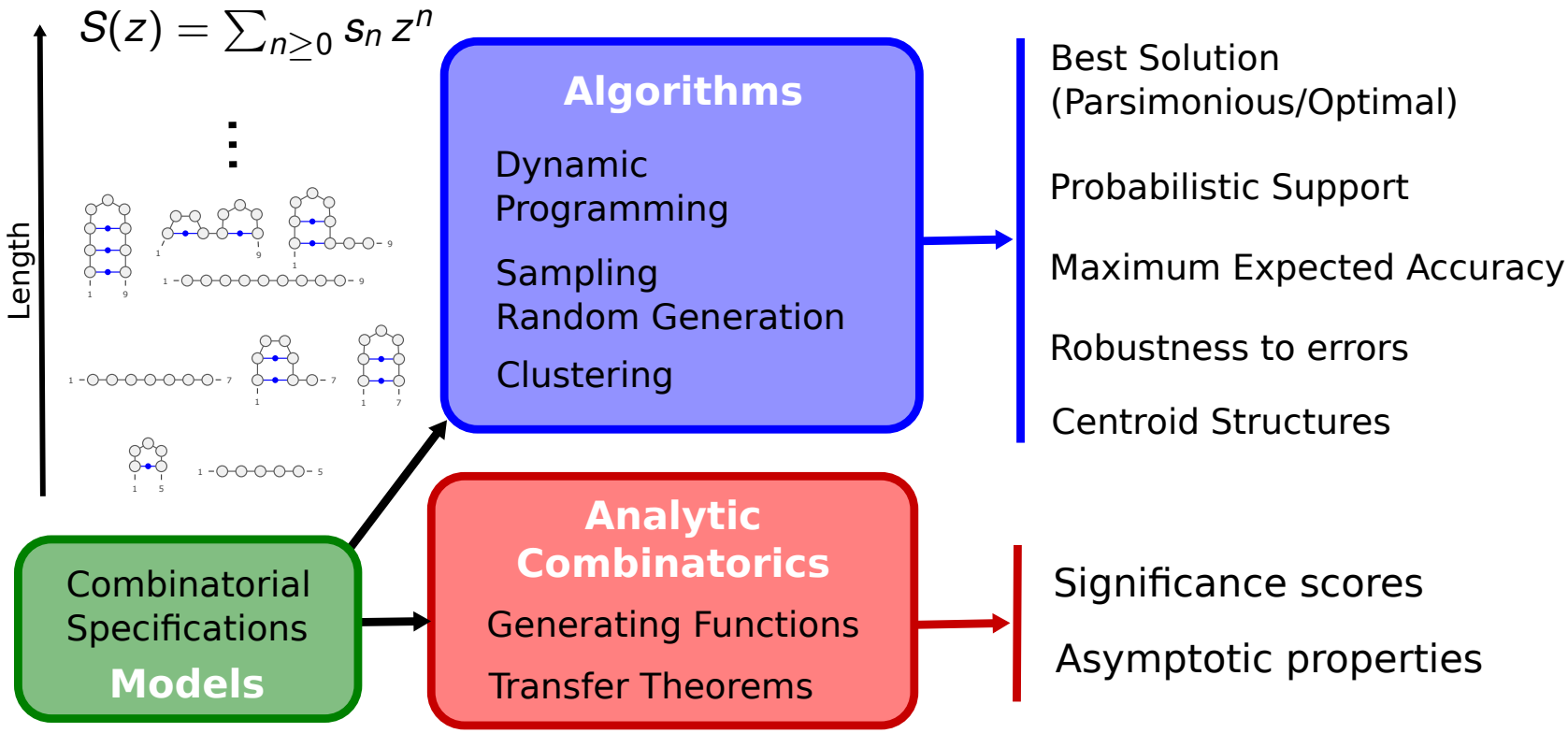
Transfer Theorems

Significance scores

Asymptotic properties

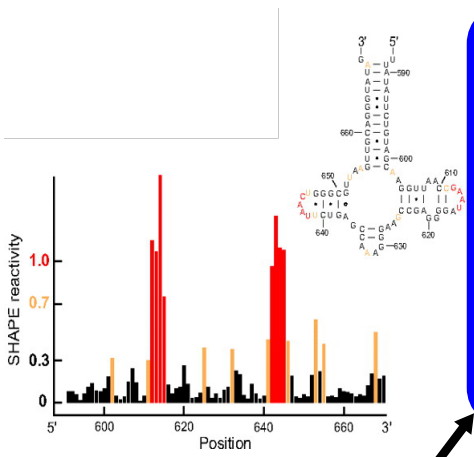
- [TCS'14'16] [Bioinformatics'17'17'17] [BMC Bioinfo'14] [RECOMB'18]...
- [Nucleic Acids Res*'13*'14'16'*'16*'18] [Bioinformatics'13'16'16] [J Comp Bio'13'18] [Algorithmica'17] [RECOMB'13'13'15'18] [ISMB/ECCB'13'17] [ACM/BCB'13] [CPM'15]...

Ensemble analysis



- [TCS'14'16] [Bioinformatics'17'17'17] [BMC Bioinfo'14] [RECOMB'18]...
- [Nucleic Acids Res*'13*'14'16'*'16*'18] [Bioinformatics'13'16'16] [J Comp Bio'13'18] [Algorithmica'17] [RECOMB'13'13'15'18] [ISMB/ECCB'13'17] [ACM/BCB'13] [CPM'15]...
- [TCS'13] [J Discrete Algo'13] [Algo Mol Biol'14] [Frontiers'16] [ANALCO'14]...

Ensemble analysis



Algorithms

Dynamic Programming

Sampling
Random Generation

Clustering

Experimental Data

Best Solution
(Parsimonious/Optimal)

Probabilistic Support

Maximum Expected Accuracy

Robustness to errors

Centroid Structures

Combinatorial Specifications

Models

Analytic Combinatorics

Generating Functions

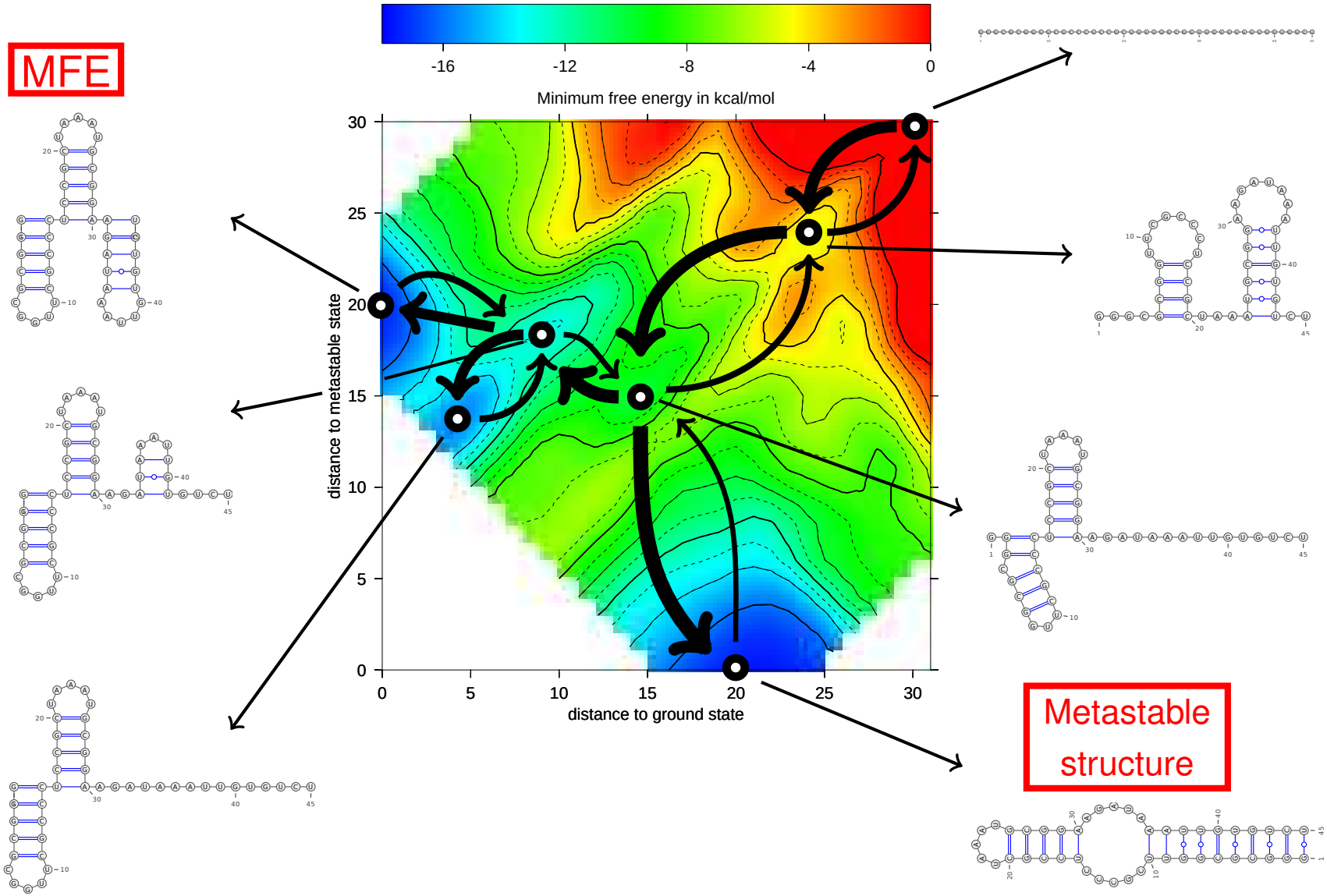
Transfer Theorems

Significance scores

Asymptotic properties

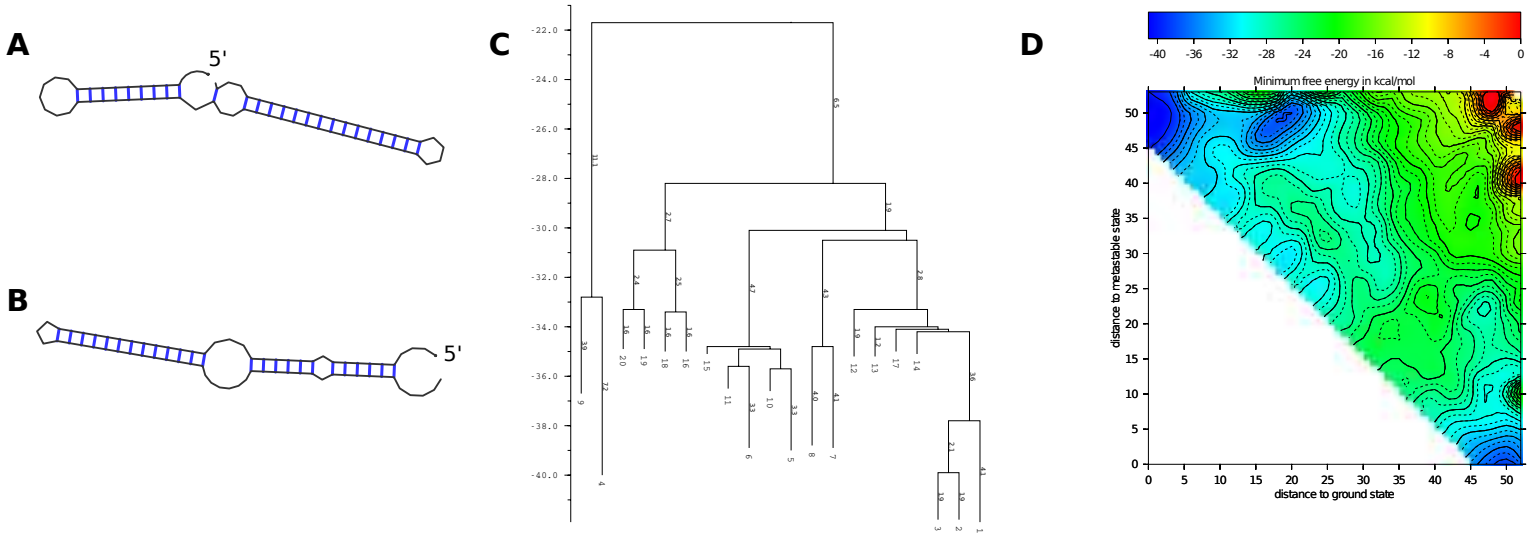
- [TCS'14'16] [Bioinformatics'17'17'17] [BMC Bioinfo'14] [RECOMB'18]...
- [Nucleic Acids Res*'13*'14'16'*'16*'18] [Bioinformatics'13'16'16] [J Comp Bio'13'18] [Algorithmica'17] [RECOMB'13'13'15'18] [ISMB/ECCB'13'17] [ACM/BCB'13] [CPM'15]...
- [TCS'13] [J Discrete Algo'13] [Algo Mol Biol'14] [Frontiers'16] [ANALCO'14]...
- [Plos CB'13'15] [RNA'14] [Nucleic Acids Res'14'16'17]...

Kinetics of RNA molecules



[**Michalik**, Touzet, **Ponty**, ECCB/ISMB'17 and Bioinformatics 2017]

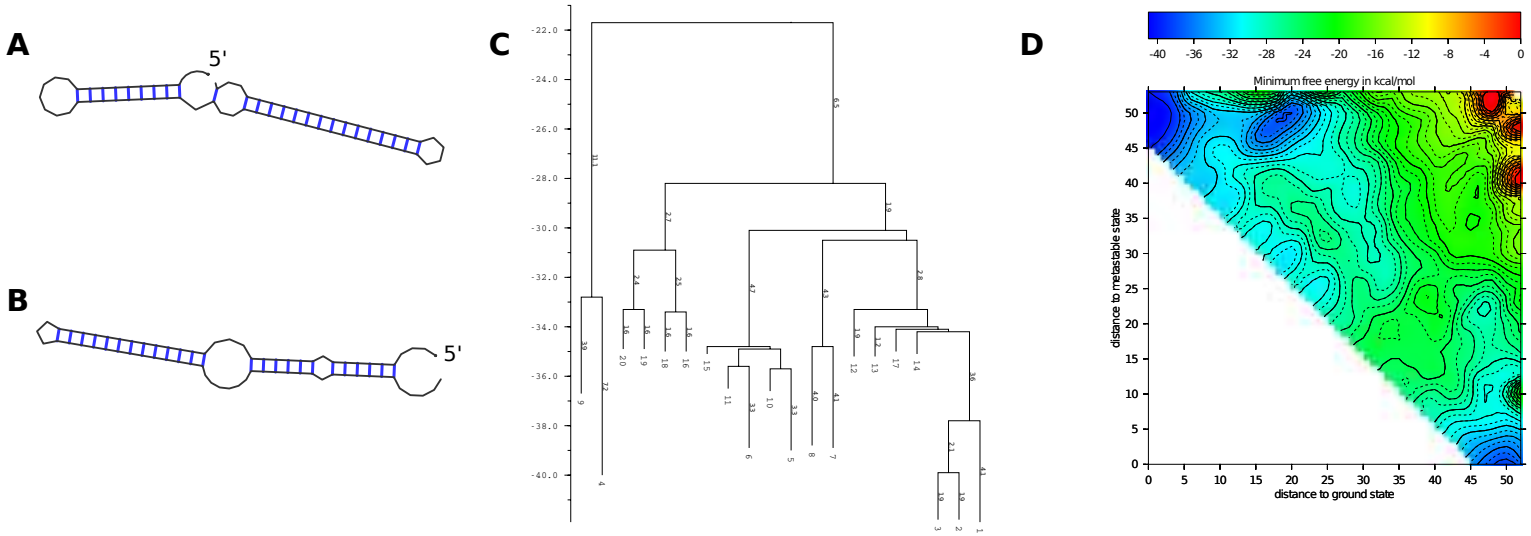
Non-redundant sampling for RNA 2D kinetics



- RNA 2D studies reveal dynamic behaviors (Markov process);
- Bottleneck: Build kinetic landscape (combinatorial explosion);

[**Michalik**, Touzet, **Ponty**, ECCB/ISMB'17 and Bioinformatics 2017]

Non-redundant sampling for RNA 2D kinetics



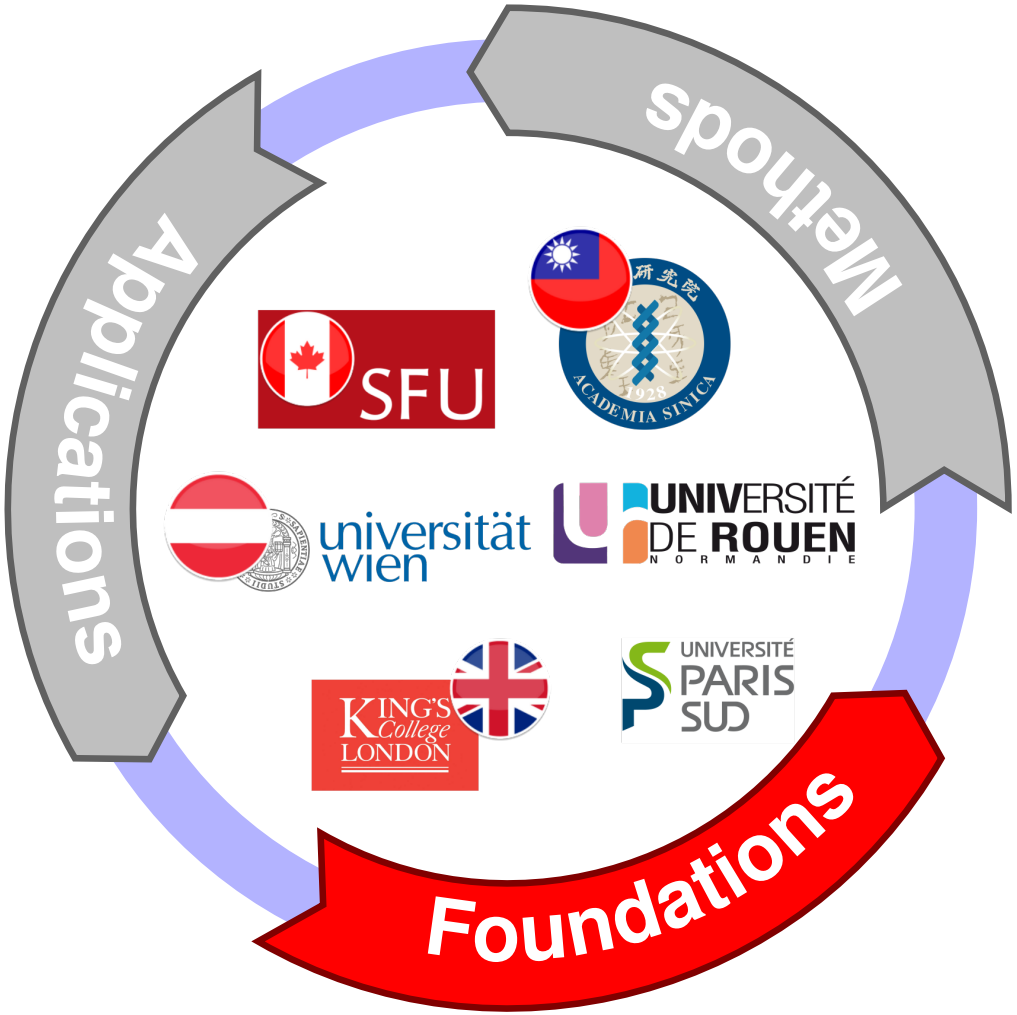
- RNA 2D studies reveal dynamic behaviors (Markov process);
- Bottleneck: Build kinetic landscape (combinatorial explosion);
- Our project: Advanced sampling methods to approximate landscapes, enabling kinetics studies beyond 1k NTs.

$\Rightarrow$ ANR/FWF-funded RNALands project (2015 – 2019)

Partners: TBI Vienna, EPI Bonsai (Inria Lille), Paris-Saclay

[**Michalik**, Touzet, **Ponty**, ECCB/ISMB'17 and Bioinformatics 2017]

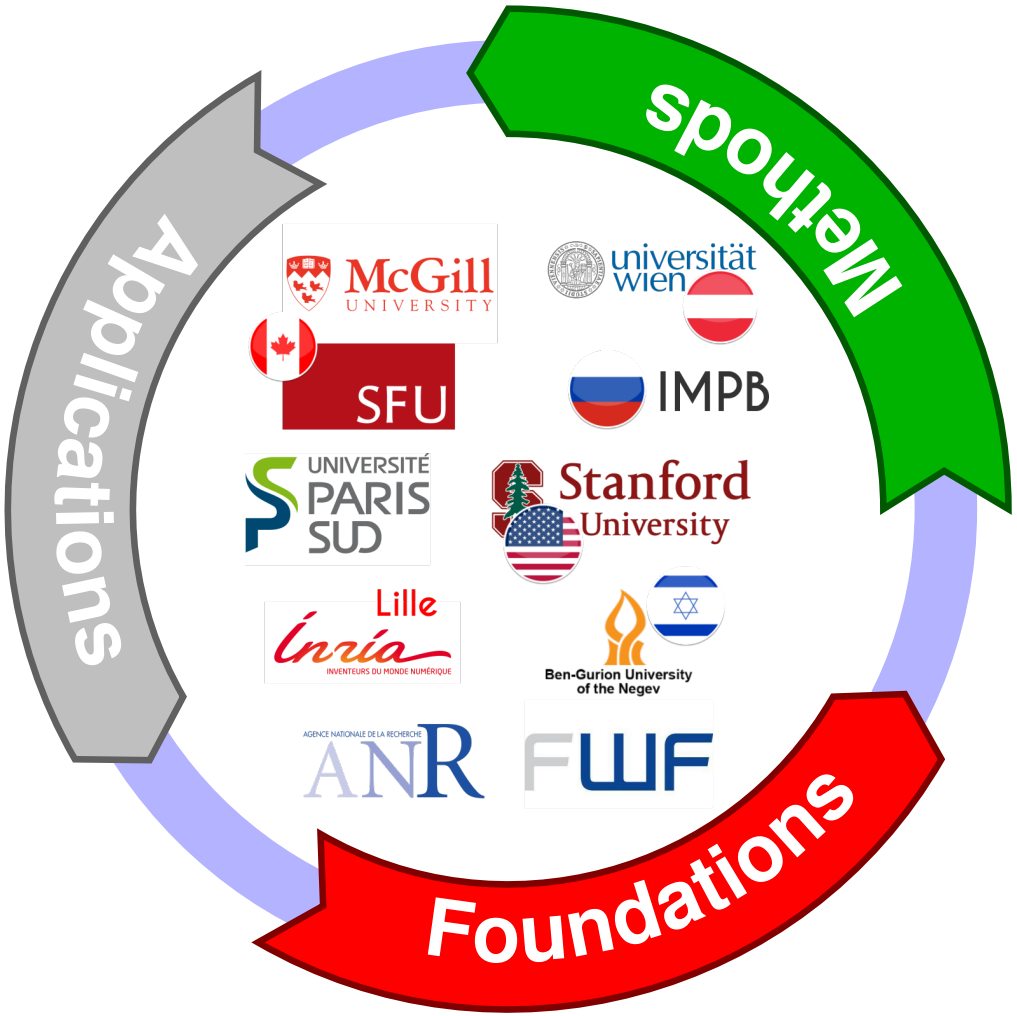
Vision



- Enumerative combinatorics
- Stringology
- Algebraic Dynamic Programming
- FPT algorithms

- Comparative genomics
- Statistical genomics
- Folding prediction
- RNA Design [RECOMB'18]
- 3D modeling [Plos CB 2013/2015]
- Circular RNAs [RNA Biology 2017]
- HIV modeling [NAR 2017]
- Structural basis of myristoylation [Nature Chemistry 2018]

Vision

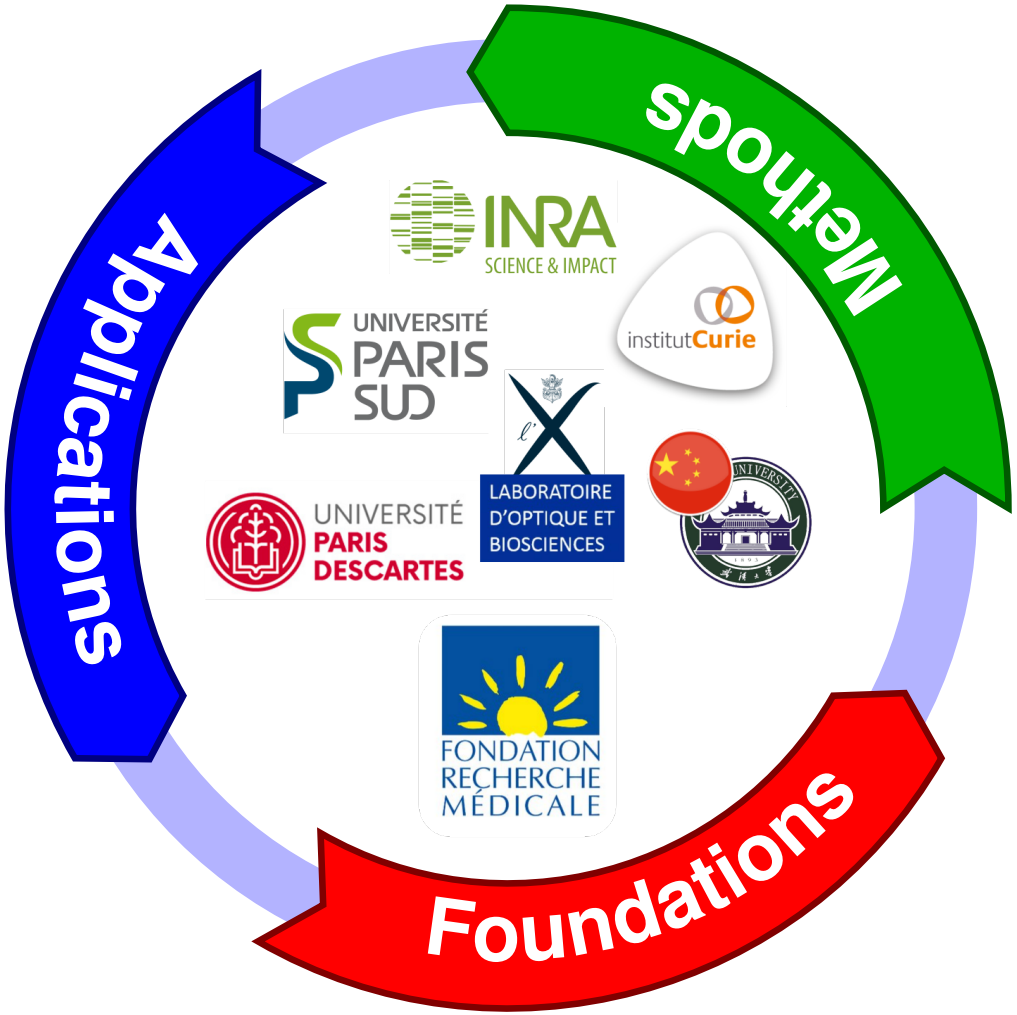


- Enumerative combinatorics
- Stringology
- Algebraic Dynamic Programming
- FPT algorithms

- Comparative genomics
- Statistical genomics
- Folding prediction
- RNA Design [RECOMB'18]

- 3D modeling [Plos CB 2013/2015]
- Circular RNAs [RNA Biology 2017]
- HIV modeling [NAR 2017]
- Structural basis of myristoylation [Nature Chemistry 2018]

Vision

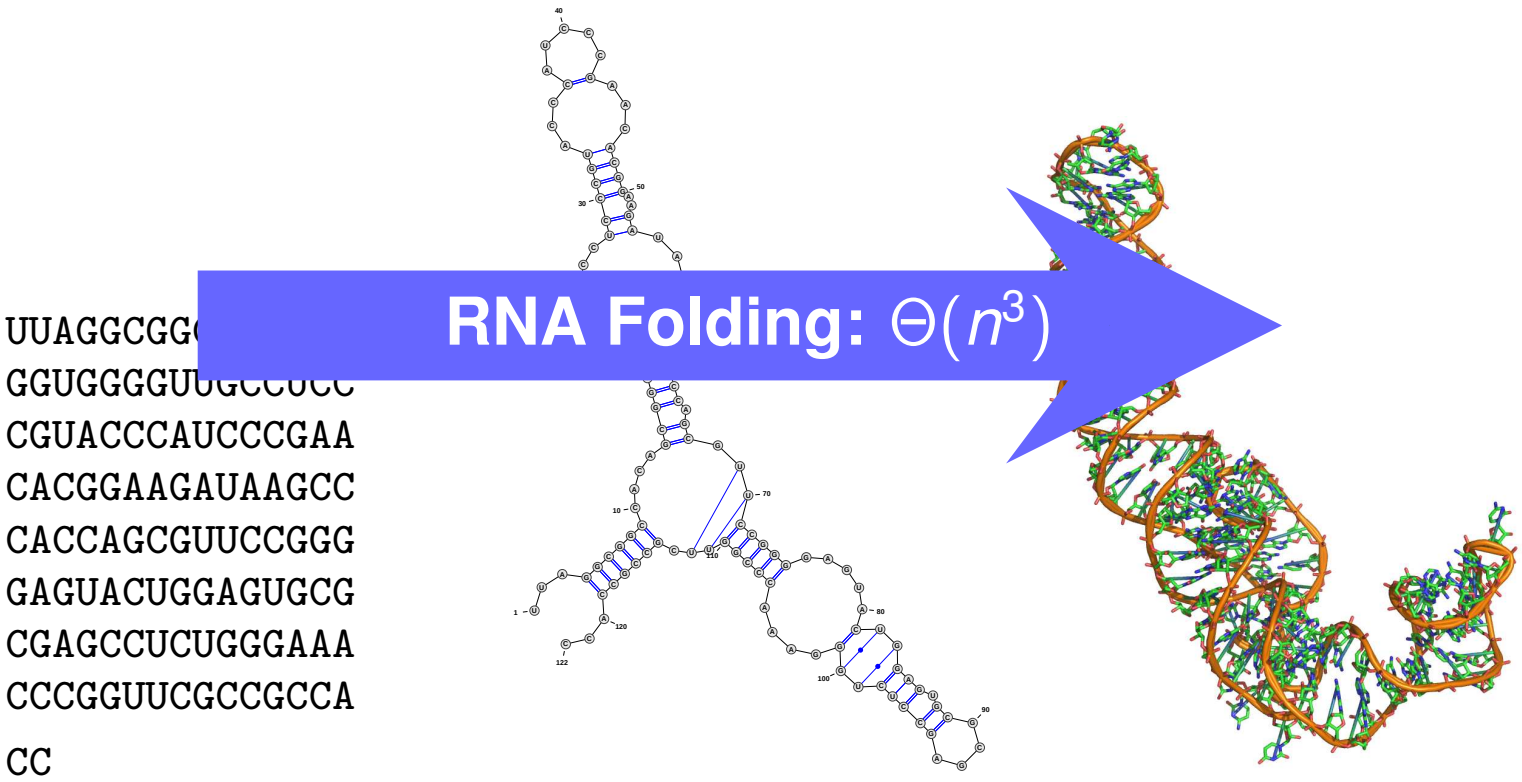


- Enumerative combinatorics
- Stringology
- Algebraic Dynamic Programming
- FPT algorithms

- Comparative genomics
- Statistical genomics
- Folding prediction
- RNA Design [RECOMB'18]

- 3D modeling [Plos CB 2013/2015]
- Circular RNAs [RNA Biology 2017]
- HIV modeling [NAR 2017]
- Structural basis of myristoylation [Nature Chemistry 2018]

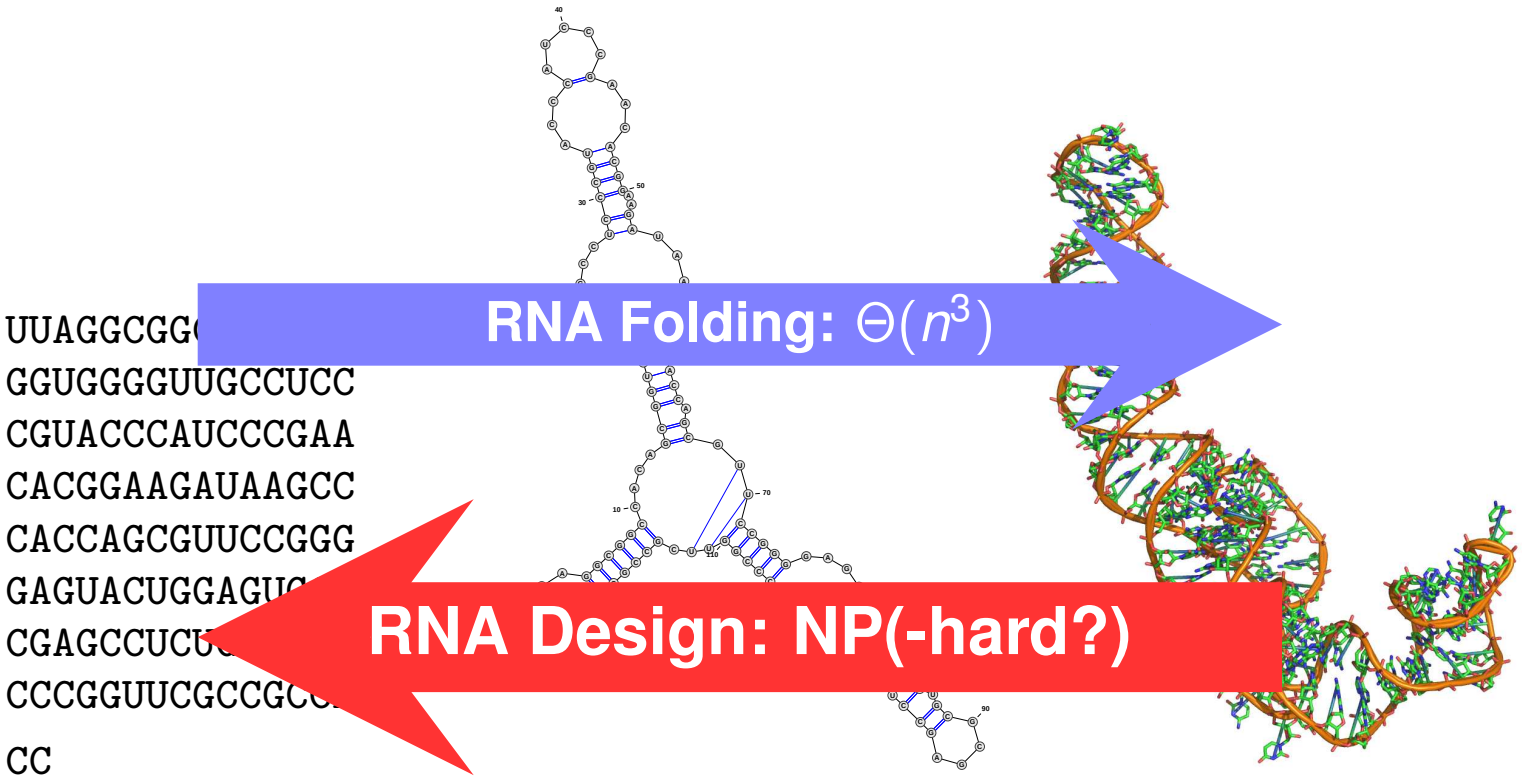
Random Generation for RNA Design 1/3



Primary structure Secondary structure Tertiary structure

5s rRNA (PDBID: 1K73:B)

Random Generation for RNA Design 1/3



Primary structure Secondary structure Tertiary structure

5s rRNA (PDBID: 1K73:B)

Random Generation for RNA Design 2/3

Input: Set of constraints

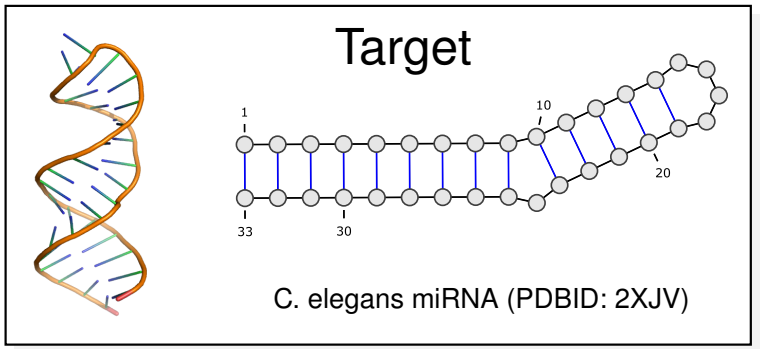
- Secondary structure
- Pattern avoid-ance/occurrence
- Energy/robustness
- Putative interactions
- ...

Goal:

Design of active RNAs

Method:

Generation + Selection



Generation

Candidates

Sequence	GC	Ener.	Prob.	mfe?
CUGUUCUGUACGUUGGCGAACGUGGCGGAGCAG	61	-15.7	0.38	Y
GGUCGGGCUGUAUUACAUGAUACAAGCCUGACU	48	-17.6	0.27	Y
AUACUGGGUGCGGGUGCCCCUGUGACUUAGUAU	55	-13.9	0.35	Y
GGGUGCCGUAAUGGUCACUCAUUCGUGGCAUUC	55	-9.8	0.11	Y
GGAGUACCUCAGCGUCCAUGCUGCGGGUGCUCC	67	-20.3	0.63	Y
GAAAUGGCCUGGAGUGAGUUCUGUGGUCAUUUU	45	-11	0.45	Y
CGAUAGGUGGCGAAGUGCUUUGUACAUUUAUCG	45	-11.2	0.53	Y
UAAACUAGGUGAUACUAGUGUCAACCUAGUUUA	33	-14.8	0.29	N
GAGGAGAUGUUACCCAGGGGUAGUUGUUUCCUU	48	-11.2	0.05	N
AAAUUAUUUCUUUGAUAUUAAAGAAGGUGGUUU	21	-4.9	0.07	N

Selection	$\leq$60	$\leq$-13	$\geq$0.3	=Y

Random Generation for RNA Design 3/3

Fact #1: Selection is expensive

$\Rightarrow$ Capture constraints during generation stage

Fact #2: Goals of synthetic biology are evolving

$\Rightarrow$ Need for modular approaches

Our approach: Non-uniform *Boltzmann* random generation

References: [1] TCS 2010, [2] AOFA'10, RECOMB'11, [4] NAR'12, [5] TCS'13
[6] CPM/Algorithmica'17, [7] RECOMB'2018

Random Generation for RNA Design 3/3

Fact #1: Selection is expensive

$\Rightarrow$ Capture constraints during generation stage

Fact #2: Goals of synthetic biology are evolving

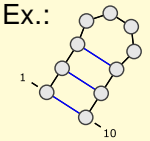
$\Rightarrow$ Need for modular approaches

Our approach: Non-uniform *Boltzmann* random generation

Easy targets:
Language models

Sec. str. compatibility [4]

Ex.:



$$
\begin{aligned}
S_0 &\to AS_1U \mid US_1A \mid GS_1C \mid CS_1G \\
S_1 &\to AS_2U \mid US_2A \mid GS_2C \mid CS_2G \\
S_2 &\to AS_3U \mid US_3A \mid GS_3C \mid CS_3G \\
S_3 &\to AS_4 \mid US_4 \mid GS_4 \mid CS_4 \\
S_4 &\to AS_5 \mid US_5 \mid GS_5 \mid CS_5 \\
S_5 &\to AS_6 \mid US_6 \mid GS_6 \mid CS_6 \\
S_7 &\to A \mid U \mid G \mid C
\end{aligned}
$$

Pattern avoidance/occurrence
(FDA$\times$CFG $\subseteq$ CFG)

$\Rightarrow$ Context-free grammars

Hard constraints

References: [1] TCS 2010, [2] AOFA'10, RECOMB'11, [4] NAR'12, [5] TCS'13
[6] CPM/Algorithmica'17, [7] RECOMB'2018

Random Generation for RNA Design 3/3

Fact #1: Selection is expensive

$\Rightarrow$ Capture constraints during generation stage

Fact #2: Goals of synthetic biology are evolving

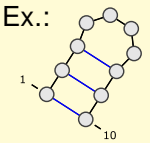
$\Rightarrow$ Need for modular approaches

Our approach: Non-uniform *Boltzmann* random generation

Easy targets:
Language models

Sec. str. compatibility [4]

Ex.:



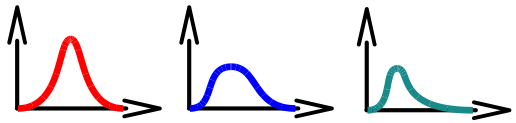
$$
\begin{aligned}
S_0 &\to AS_1U \mid US_1A \mid GS_1C \mid CS_1G \\
S_1 &\to AS_2U \mid US_2A \mid GS_2C \mid CS_2G \\
S_2 &\to AS_3U \mid US_3A \mid GS_3C \mid CS_3G \\
S_3 &\to AS_4 \mid US_4 \mid GS_4 \mid CS_4 \\
S_4 &\to AS_5 \mid US_5 \mid GS_5 \mid CS_5 \\
S_5 &\to AS_6 \mid US_6 \mid GS_6 \mid CS_6 \\
S_7 &\to A \mid U \mid G \mid C
\end{aligned}
$$

Pattern avoidance/occurrence
(FDA$\times$CFG $\subseteq$ CFG)

$\Rightarrow$ Context-free grammars

Hard constraints

Additional feature distributions
Typically Gaussian



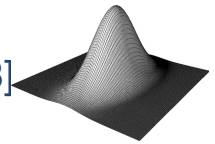
$\Rightarrow$ Weighted random generation [1]

$$\mathbb{P}(w) = \frac{\pi_a^{n_a}\pi_b^{n_b}\cdots}{\mathcal{Z}_\pi} \qquad \mathcal{Z}_\pi := \sum_{w\in S} \pi_a^{n_a}\pi_b^{n_b}\cdots$$

\+ Low variances

$\Rightarrow$ **Efficient rejection**

Multidim. Boltzmann [2,3]



Soft constraints

References: [1] TCS 2010, [2] AOFA'10, RECOMB'11, [4] NAR'12, [5] TCS'13
[6] CPM/Algorithmica'17, [7] RECOMB'2018

Random Generation for RNA Design 3/3

Fact #1: Selection is expensive

⇒ Capture constraints during generation stage

Fact #2: Goals of synthetic biology are evolving

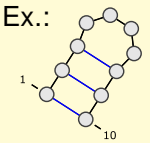
⇒ Need for modular approaches

Our approach: Non-uniform *Boltzmann* random generation

Easy targets:
Language models

Sec. str. compatibility [4]

Ex.:



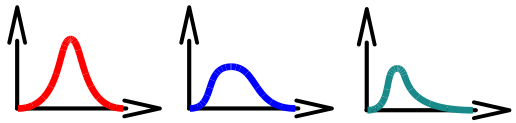
$S_0 \to AS_1U \mid US_1A \mid GS_1C \mid CS_1G$
$S_1 \to AS_2U \mid US_2A \mid GS_2C \mid CS_2G$
$S_2 \to AS_3U \mid US_3A \mid GS_3C \mid CS_3G$
$S_3 \to AS_4 \mid US_4 \mid GS_4 \mid CS_4$
$S_4 \to AS_5 \mid US_5 \mid GS_5 \mid CS_5$
$S_5 \to AS_6 \mid US_6 \mid GS_6 \mid CS_6$
$S_7 \to A \mid U \mid G \mid C$

Pattern avoidance/occurrence
(FDA$\times$CFG $\subseteq$ CFG)

⇒ Context-free grammars

Hard constraints

Additional feature distributions
Typically Gaussian



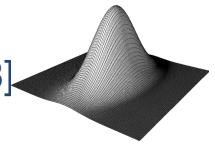
⇒ Weighted random generation [1]

$$\mathbb{P}(w) = \frac{\pi_a^{n_a}\pi_b^{n_b}\cdots}{\mathcal{Z}_\pi} \qquad \mathcal{Z}_\pi := \sum_{w\in S} \pi_a^{n_a}\pi_b^{n_b}\cdots$$

\+ Low variances

⇒ **Efficient rejection**

Multidim. Boltzmann [2,3]



Soft constraints

Complex features

Robustness

Predicted folding
(2D/MFold, 3D/MCFold...)

Stability
(Molecular dynamics)

Interactions
(RNACofold, Docking)

Non-redundant generation [5]

Multiple Targets [7]

Posterior filtering

References: [1] TCS 2010, [2] AOFA'10, RECOMB'11, [4] NAR'12, [5] TCS'13
[6] CPM/Algorithmica'17, [7] RECOMB'2018

Service and visibility

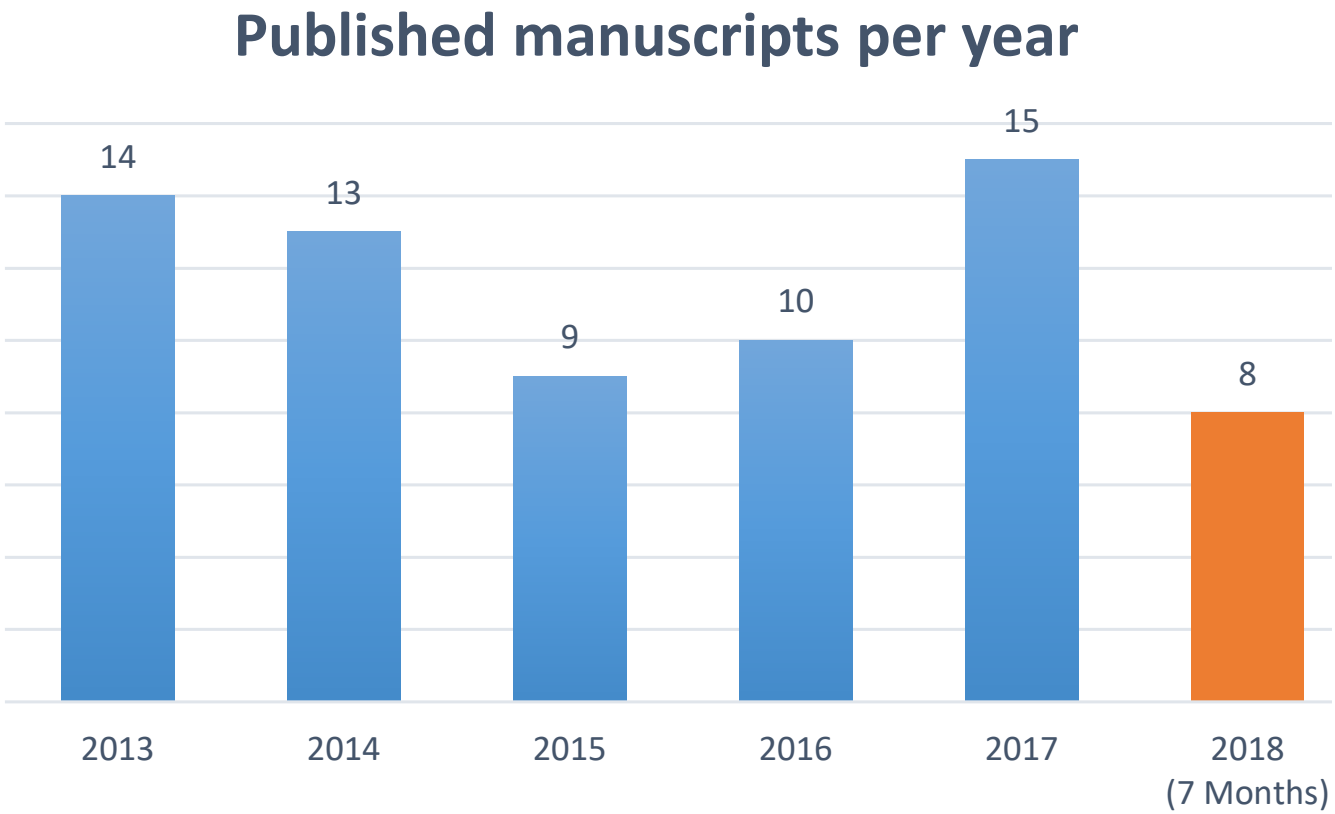
Program Committees

Local chairs

Editorial duties

+ M. Regnier head of LIX
+ Y. Ponty@coNRS 2013-17
+ Involvement in
  GdR BIM (GT MASIM)
  GdR IM (co-orga SeqBio'15)
+ AlsoSB'19 Winter School@CIRM

Publications



5 $\rightarrow$ 4 permanents (+ M. Regnier's tenure@LIX & J.-M. Steyaert's aemeritus)

Yet sustained research activity in selective venues!

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB , Paris V , INRA , I2BC , Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...
- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB , Paris V , INRA , I2BC , Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...
- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB , Paris V , INRA , I2BC , Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...
- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB, Paris V, INRAE, I2BC, Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...
- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB, Paris V, INRA, I2BC, Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...

EUKARYS
Advanced life science technologies

- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB, Paris V, INRA, I2BC, Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...

EUKARYS
Advanced life science technologies

- Nurture our roots in Discrete Maths and Algorithms

Future developments and challenges

- Maintain critical mass ($\rightarrow$ 2019 X/CNRS competitions)
- Extend our scope of methods (3D, ML, Disc. Algo.) through LIX collabs (COMBI, DaSciM, STREAM, MAX) and hirings...
- ...while keeping our focus on RNA
- Consolidate our software developments (web-servers)
- Strengthen our experimental collaborations:
  LOB, Paris V, INRA, I2BC, Wuhan
- Broaden our scope of applications: Livestock genetics, synthetic biology, therapeutic RNAs (Eukarys)...

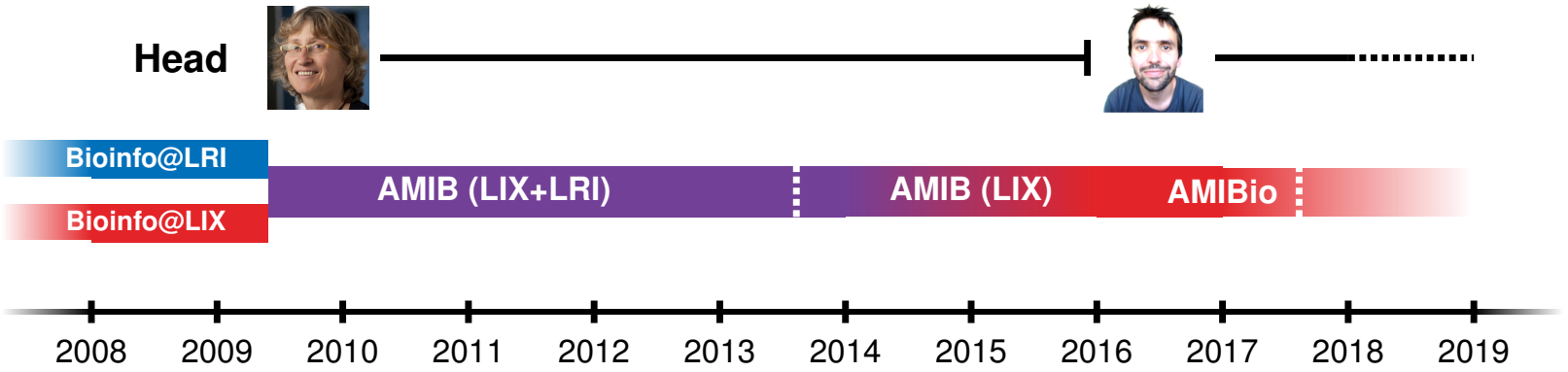
EUKARYS
Advanced life science technologies

- Nurture our roots in Discrete Maths and Algorithms

MERCI

Annexes

History


Head
Bioinfo@LRI
Bioinfo@LIX
AMIB (LIX+LRI)
AMIB (LIX)
AMIBio
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Local interactions

- Joint seminar with Bioinfo Team@LRI (Paris-Sud)
- Joint DIGICOSME project with LRI (L. Paulevé)

Teaching:

- *Programme d'approfondissement* in Bioinformatics@l'X
- Paris-Area Masters in Bioinformatics
  - Paris-Saclay AMI2B
  - Sorbonne Universités BIM

Main national collaborations

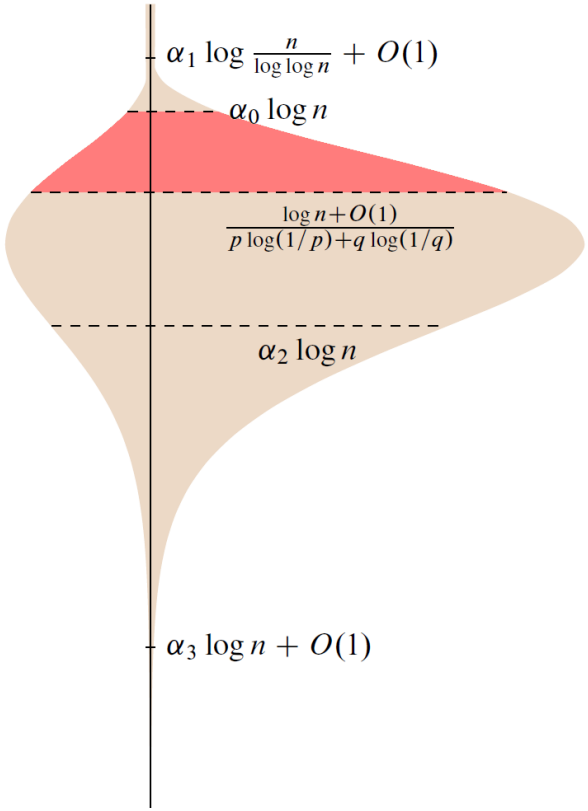
- Univ. Paris-Sud – RNA Bioinformatics, machine learning and algorithmic game theory for molecular modeling – `VARNA` 💾 & `GenRGenS` 💾
- LOB🧪, Ecole Polytechnique – Circular RNAs
- Univ. Paris V🧪 – SHAPE probing analysis
  Funding: FRM DiffProbing
- Univ. Lille I/Inria Lille Nord Europe – RNA kinetics – `RNANR` 💾
  Funding: ANR/FWF RNALands grant
- Univ. de Rouen – Dynamic range queries and minimal absent words – `MAW` 💾
- Institut Curie🧪 – Variant detection in Cancer Genomes – `SV-Bay` 💾

Main international collaborations

- Simon Fraser University – RNA inverse folding, alignment, comparative genomics, combinatorics – `DeClone` 💾
- Univ. McGill – RNA design, SHAPE analysis – `IncaRNAtion` 💾
  Funding: ALARNA Inria/NSERC associate team
- Univ. Vienna – RNA kinetics, RNA Design – `RNANR` 💾
  Funding: ANR/FWF RNALands grant
- Stanford Univ. – RNA kinematics – `KGS` 💾
  Funding: ITSNAP associate team and France-Stanford program
- King's College, London – Minimal absent words – `MAW` 💾
- Univ. Wuhan 🧪 – Transcription speed from GROSeq experiments
  Funding: PHC Xu Guangqi (CampusFrance)
- Univ. Ben Gurion – RNA design
- IMPB Moscow – Clump and word combinatorics
  Funding: Inria/Russia CARNAGE associate team

Stringology/Analytic combinatorics 1/2

Objective: (Right maximal) Repetitions and Absent words of length k

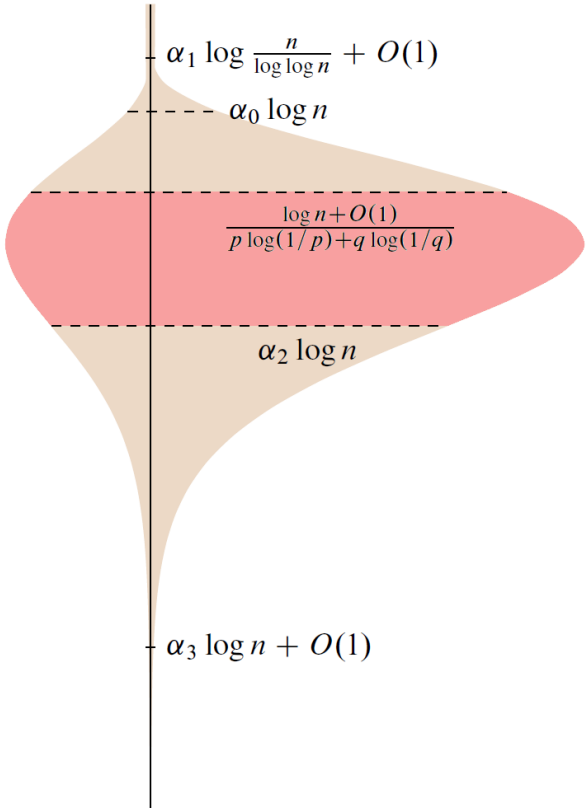


Bernoulli binary alphabet
[Hwang *et al*, 2009]

State of the Art

Stringology/Analytic combinatorics 1/2

Objective: (Right maximal) Repetitions and Absent words of length k

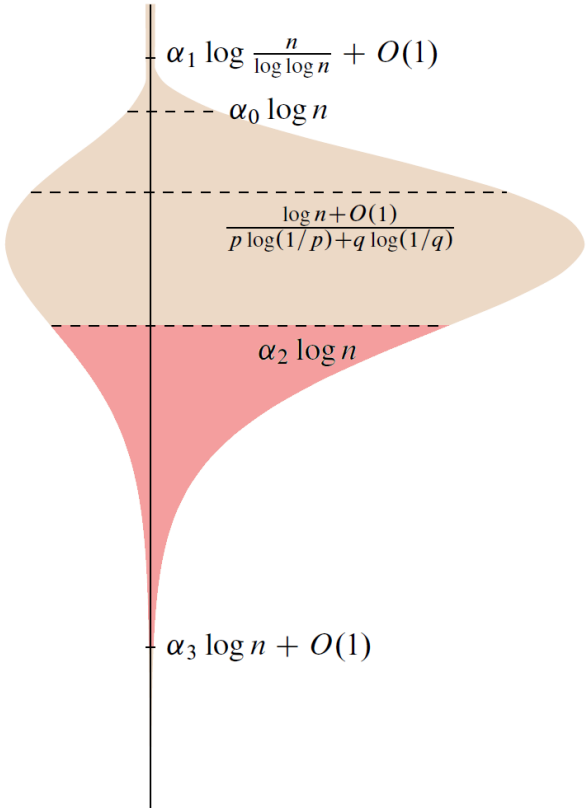


Bernoulli binary alphabet
[Hwang *et al*, 2009]

State of the Art

Stringology/Analytic combinatorics 1/2

Objective: (Right maximal) Repetitions and Absent words of length k

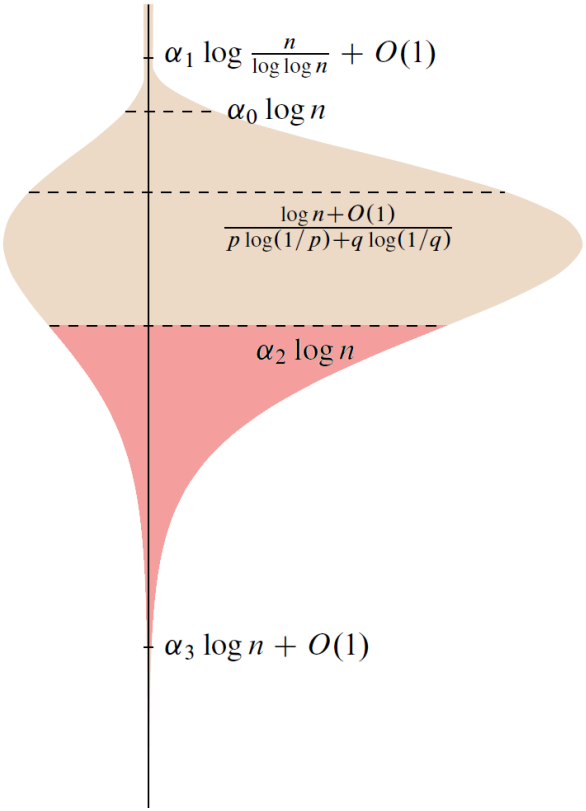


Bernoulli binary alphabet
[Hwang *et al*, 2009]

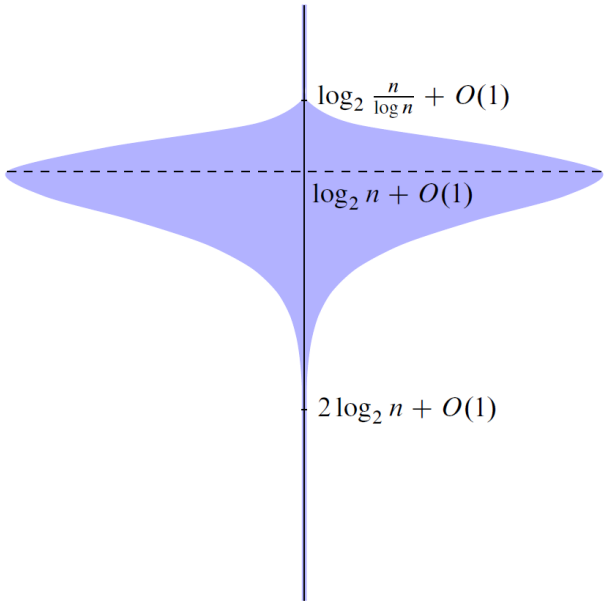
State of the Art

Stringology/Analytic combinatorics 1/2

Objective: (Right maximal) Repetitions and Absent words of length k



Bernoulli binary alphabet
[Hwang *et al*, 2009]



Symmetric Bernoulli binary
[Flajolet *et al*, 1983]

State of the Art

Stringology/Analytic combinatorics 2/2

Idea: At each level, distinguish Rare, Transient and Frequent nodes.
Transient nodes are most likely associated to maximal repetitions.

Analytic combinatorics

Objective function $\rho(k_1, \cdots, k_V)$

$\Downarrow$

3 kind of nodes $\leftrightarrow$ Large Deviation principle

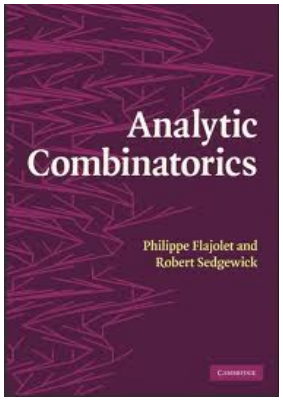
Results: [**Regnier-Chassignet**, Frontiers 2016]

- We recover the three previously identified domains
- Easily generalized to larger alphabets ($|\Sigma| > 2$)
- Information theory: $(k \cdot p_1, \cdots, k \cdot p_V)$

Stringology/Analytic combinatorics 2/2

Idea: At each level, distinguish Rare, Transient and Frequent nodes.
Transient nodes are most likely associated to maximal repetitions.

Analytic combinatorics



Objective function $\rho(k_1, \cdots, k_V)$

$\Downarrow$

3 kind of nodes $\leftrightarrow$ Large Deviation principle

Results: [**Regnier-Chassignet**, Frontiers 2016]

- We recover the three previously identified domains
- Easily generalized to larger alphabets ($|\Sigma| > 2$)
- Information theory: $(k \cdot p_1, \cdots, k \cdot p_V)$

Stringology/Analytic combinatorics 2/2

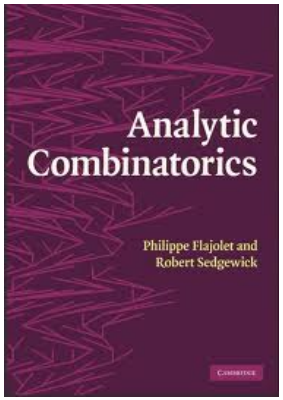
Idea: At each level, distinguish Rare, Transient and Frequent nodes.
Transient nodes are most likely associated to maximal repetitions.

Analytic combinatorics

Objective function $\rho(k_1, \cdots, k_V)$

$\Downarrow$

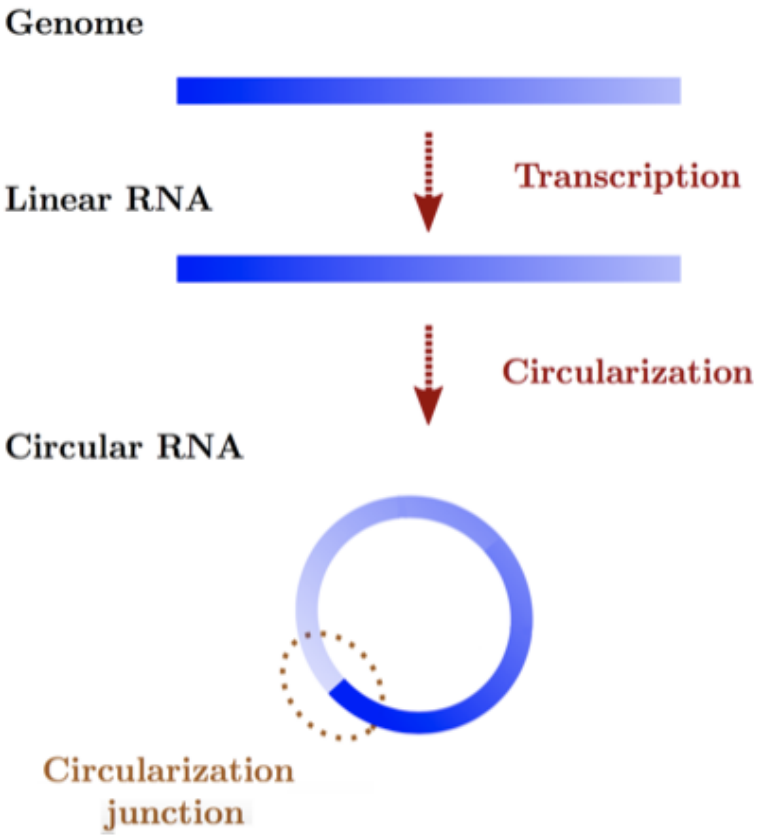
3 kind of nodes $\leftrightarrow$ Large Deviation principle



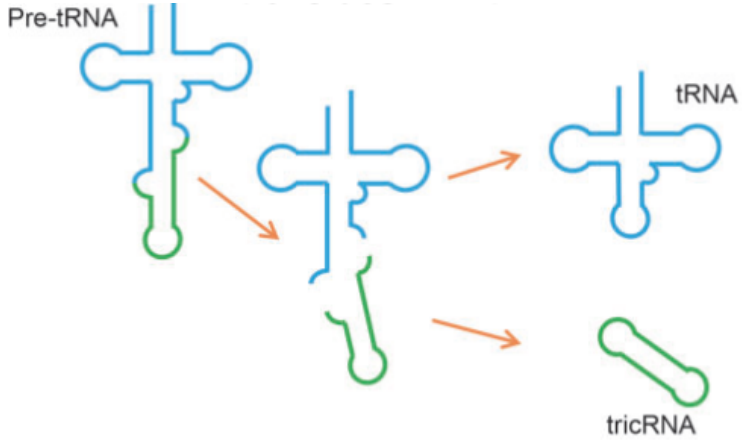
Results: [**Regnier-Chassignet**, Frontiers 2016]

- We recover the three previously identified domains
- Easily generalized to larger alphabets ($|\Sigma| > 2$)
- Information theory: $(k \cdot p_1, \cdots, k \cdot p_V)$

Circular RNAs 1/2

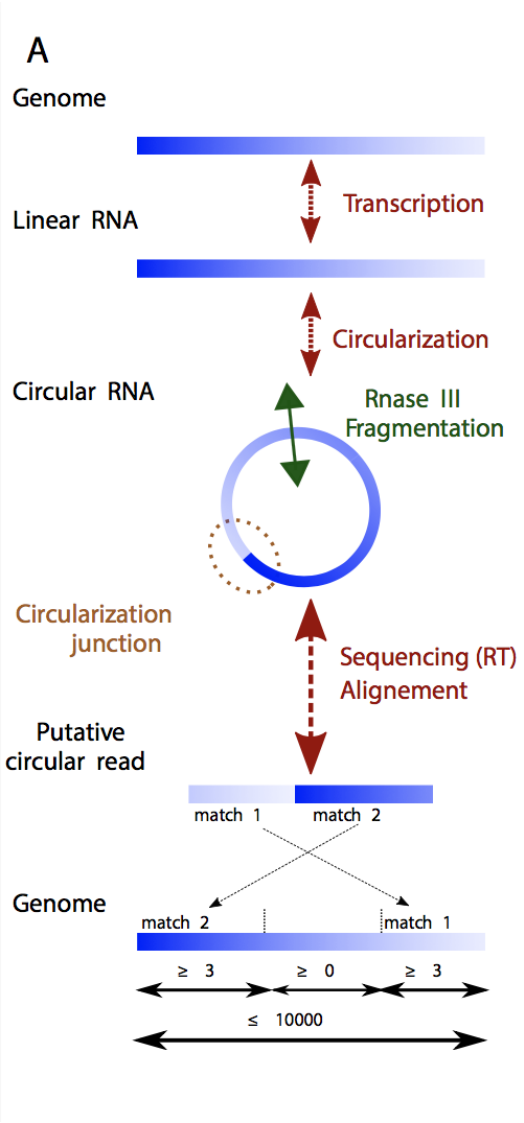


LOB partner : Pab1020 ligase



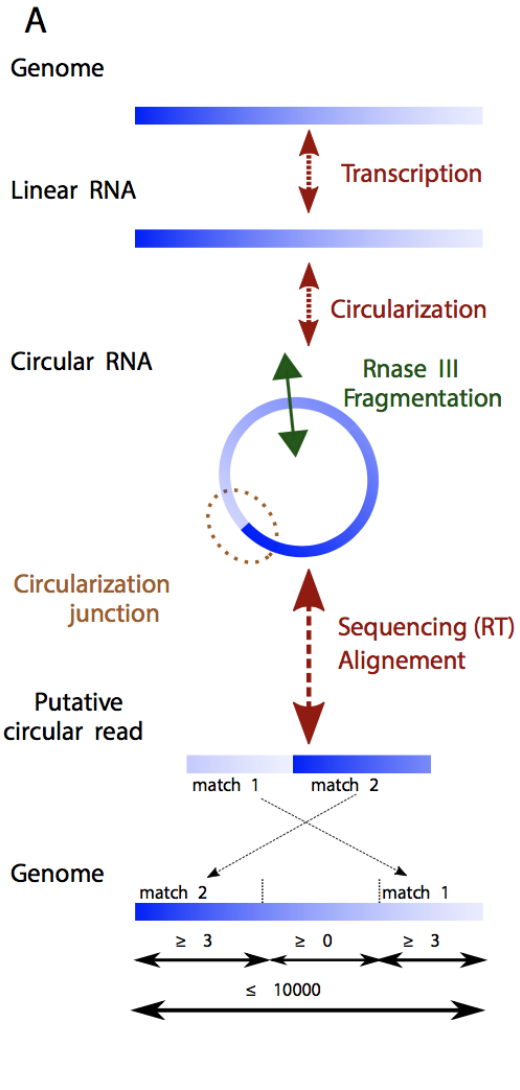
[Becker, **Heliou**, Djaout, Lestini, **Regnier**, Myllykallio, RNA Biol. 2017]

Circular RNAs 2/2



[Becker, **Heliou**, Djaout, Lestini, **Regnier**, Myllykallio, RNA Biol. 2017]

Circular RNAs 2/2



Our results:

- Confirm the role of Pab1020
- Exhibit circular RNAs
- Inactivate coding gene (IFREMER-Brest)
- Suggest other families

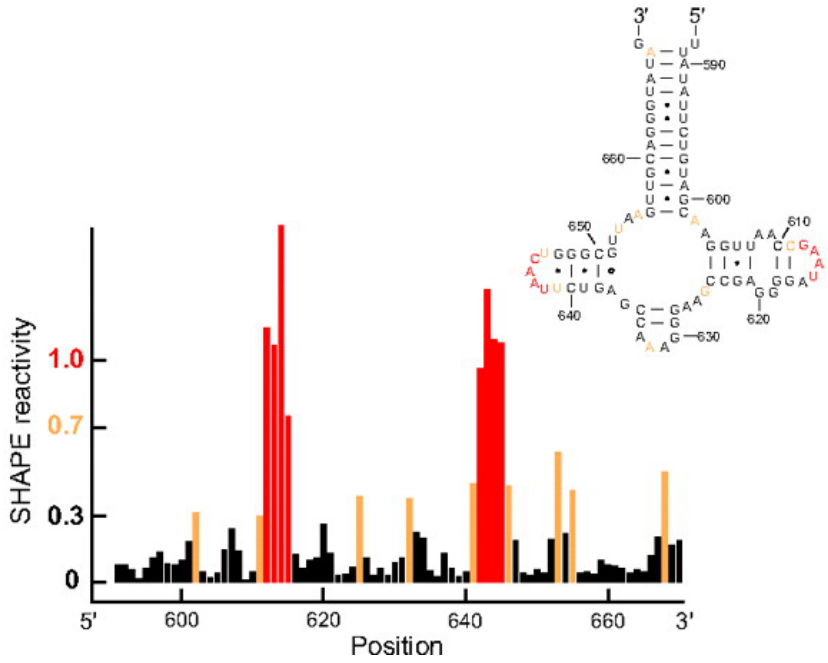
[Becker, **Heliou**, Djaout, Lestini, **Regnier**, Myllykallio, RNA Biol. 2017]

SHAPE probing

SHAPE = Selective 2'-Hydroxyl Acylation by Primer Extension

Produces accessibility profiles, i.e. *projections* of RNA structure.

in silico analysis of SHAPE data remains complex and misleading.



Our goal: To develop *hybrid* modeling approaches with experimentalists (Paris V) and bioinformaticians (McGill)

- Massively parallel derivation of profiles (PCR/NGS/EM);
- Clustering to model structure of viruses (HIV, Ebola);
- Joint analysis of multiple SHAPE profiles.

Funded by *Fondation pour la Recherche Médicale* (2015-2018)

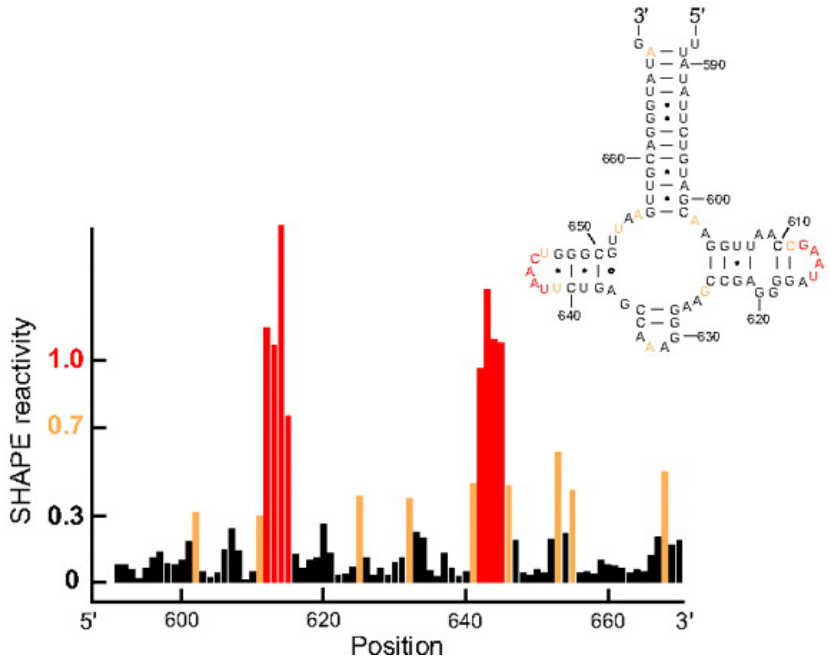
[Desforges, . . . , **Saaidi**, **Ponty**, Ohlmann, Sargueil, NAR 2017]

SHAPE probing

SHAPE = Selective 2'-Hydroxyl Acylation by Primer Extension

Produces accessibility profiles, i.e. *projections* of RNA structure.

in silico analysis of SHAPE data remains complex and misleading.



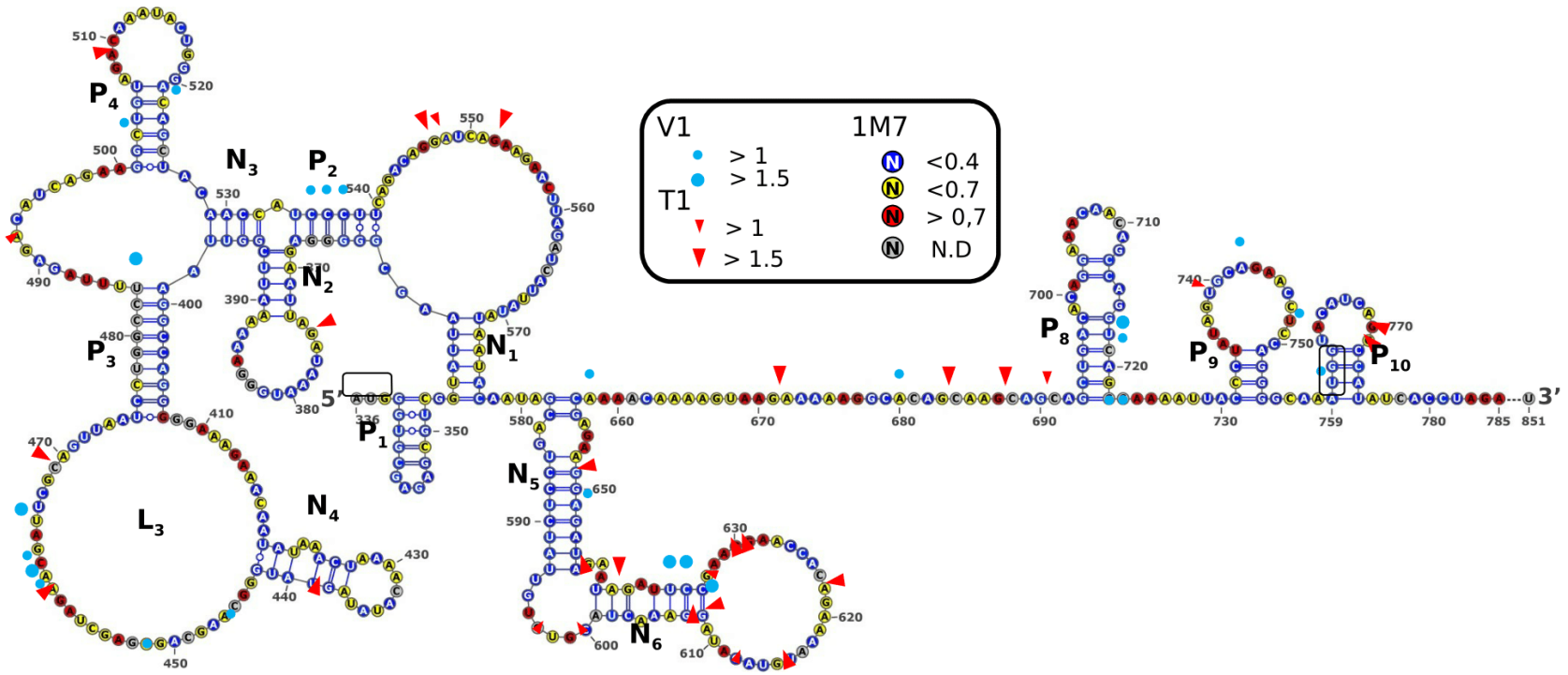
Our goal: To develop *hybrid* modeling approaches with experimentalists (Paris V) and bioinformaticians (McGill)

- Massively parallel derivation of profiles (PCR/NGS/EM);
- Clustering to model structure of viruses (HIV, Ebola);
- Joint analysis of multiple SHAPE profiles.

Funded by *Fondation pour la Recherche Médicale* (2015-2018)

[Desforges, . . . , **Saaidi**, **Ponty**, Ohlmann, Sargueil, NAR 2017]

HIV translational machinery

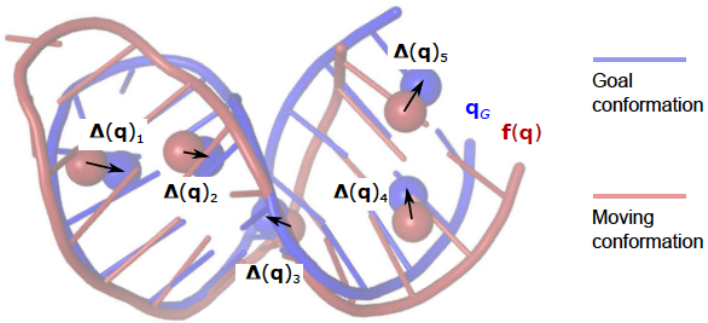


[Desforges, . . . , **Saaidi**, **Ponty**, Ohlmann, Sargueil, NAR 2017]

Kinematics 1/2

Given a full atoms initial conformation and a goal conformation with possibly only a few atoms positions

Find a feasible trajectory from initial to goals

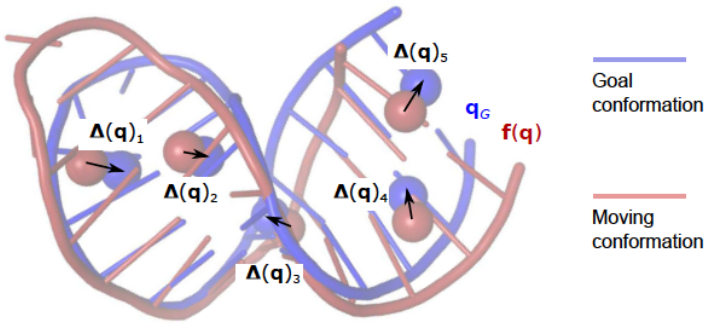


Developing an extension of KGS software (Kino-Geometric Sampling)

Kinematics 1/2

Given a full atoms initial conformation and a goal conformation with possibly only a few atoms positions

Find a feasible trajectory from initial to goals



Developing an extension of KGS software (Kino-Geometric Sampling)

Kinematics 2/2

We define a feasible path by two conditions:

- secondary structure preservation (WC hydrogen bonds)
- clash avoidance

The preservation of a WC interaction is guaranteed by 5 equations :

Deriving with respect to each DOF gives a Jacobian matrix denoted J. A perturbation δq is acceptable if $J.\delta q = 0$.

Clash avoidance using motion planning:

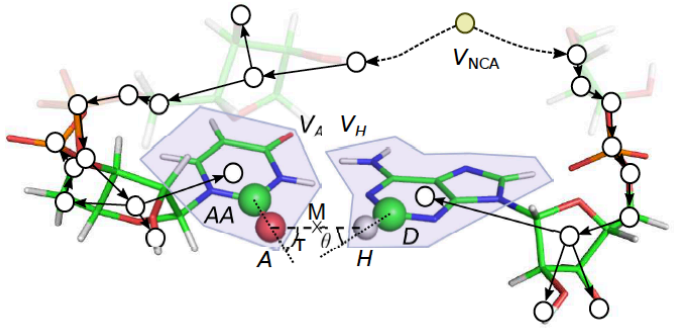
[**Heliou**, Budday, **Fonseca**, Van den Bedem, Bioinformatics 2017]

Kinematics 2/2

We define a feasible path by two conditions:

- secondary structure preservation (WC hydrogen bonds)
- clash avoidance

The preservation of a WC interaction is guaranteed by 5 equations :



$$M_{left} = M_{right}$$
$$\tau = \tau_{ini}$$
$$\theta = \theta_{ini}$$

Deriving with respect to each DOF gives a Jacobian matrix denoted J. A perturbation δq is acceptable if $J.\delta q = 0$.

Clash avoidance using motion planning:

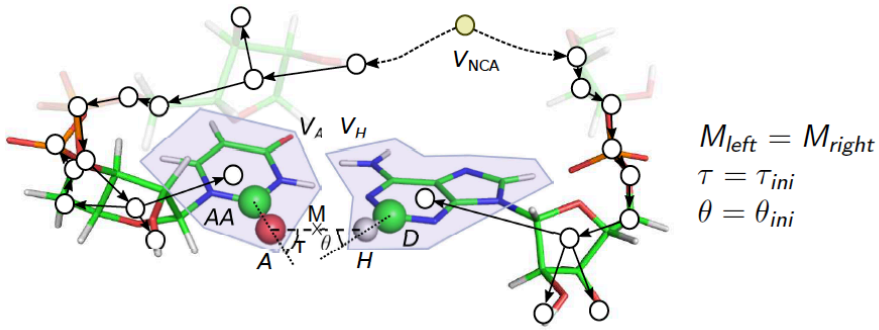
[**Heliou**, Budday, **Fonseca**, Van den Bedem, Bioinformatics 2017]

Kinematics 2/2

We define a feasible path by two conditions:

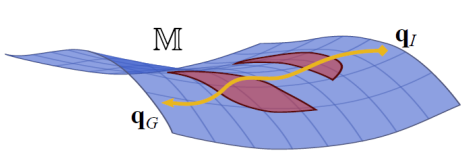
- secondary structure preservation (WC hydrogen bonds)
- clash avoidance

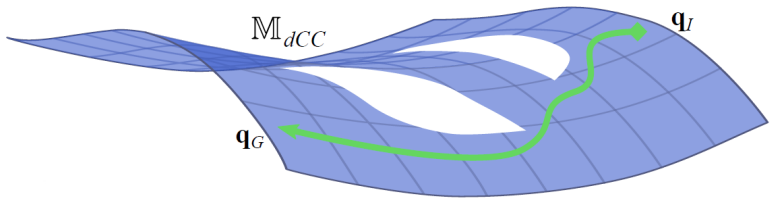
The preservation of a WC interaction is guaranteed by 5 equations :



Deriving with respect to each DOF gives a Jacobian matrix denoted J. A perturbation δq is acceptable if $J.\delta q = 0$.

Clash avoidance using motion planning:





[**Heliou**, Budday, **Fonseca**, Van den Bedem, Bioinformatics 2017]

Controlled experiments through RNA design

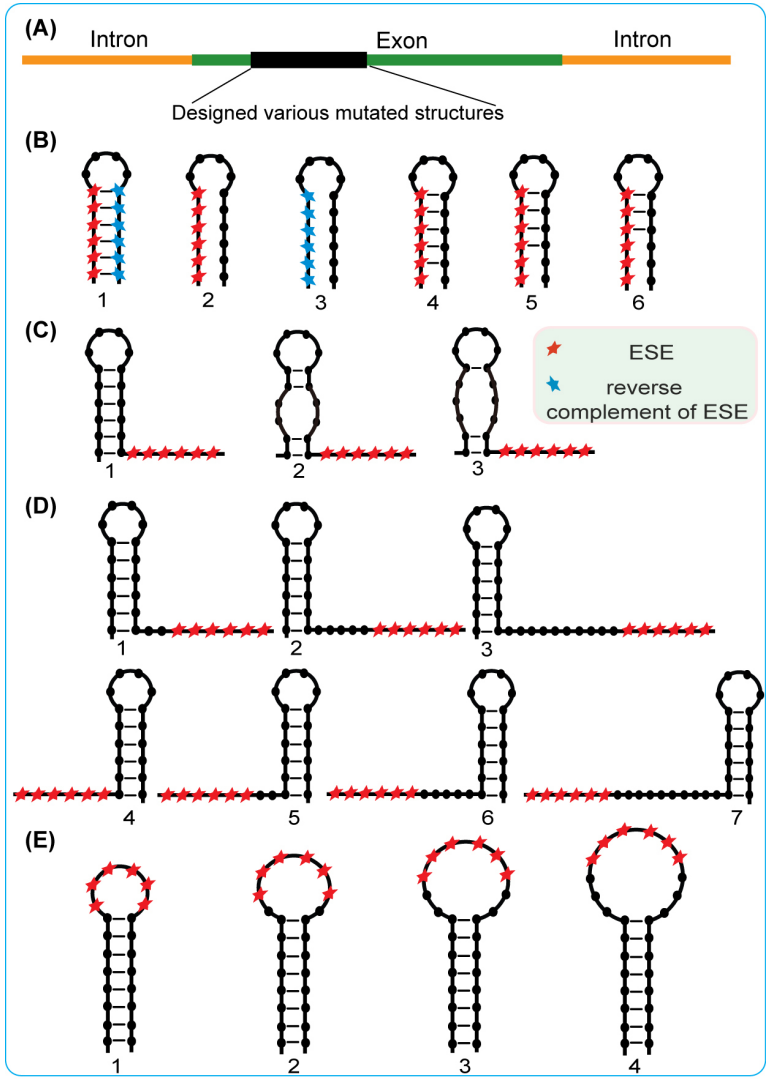
Motivation: Quantifying the impact of structure S on efficacy of a single Exon Splicing E Enhancers (ESE):

- Presence of given ESE motif E;
- Different structures S_1, S_2 . . . ;
- Avoid library of ($\sim$1500!) documented ESEs motifs.

Objectives. Design RNA which:

1. Folds into a given structure;
2. Features/avoids motifs.
3. Control GC%, Boltz. prob.. . . .

Structural context of ESE motif in transcript was shown to affect its functionality. [Liu *et al*, FEBS Lett. 2010]



RNA kinetics

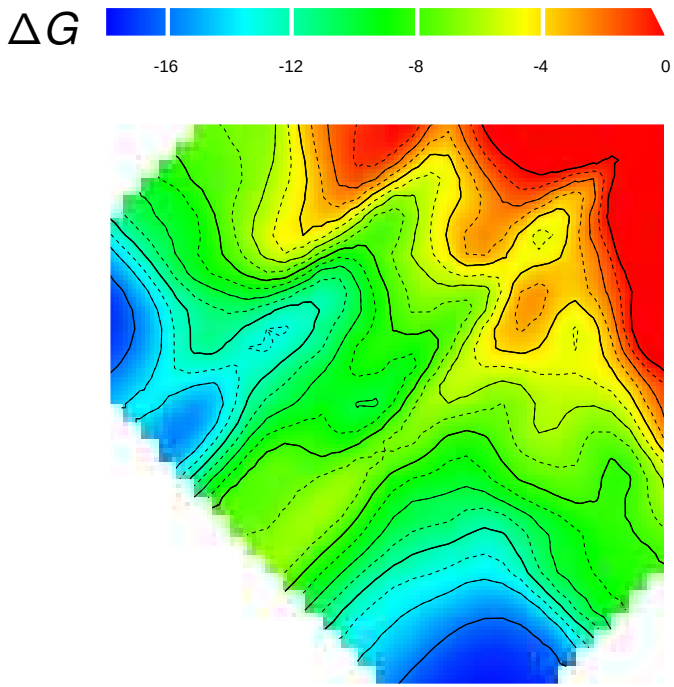
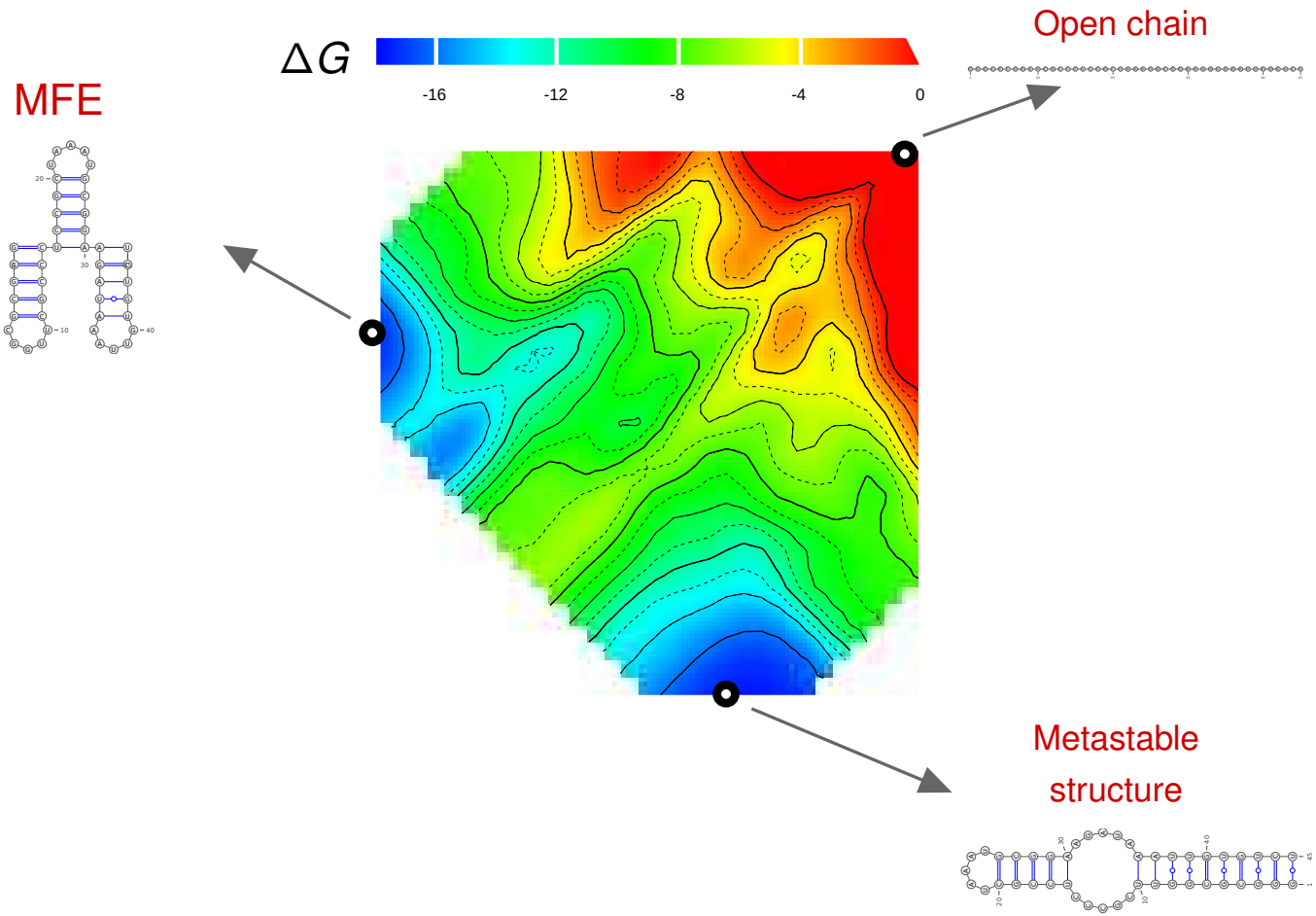


Image : Lorenz *et al*, GCB'09

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

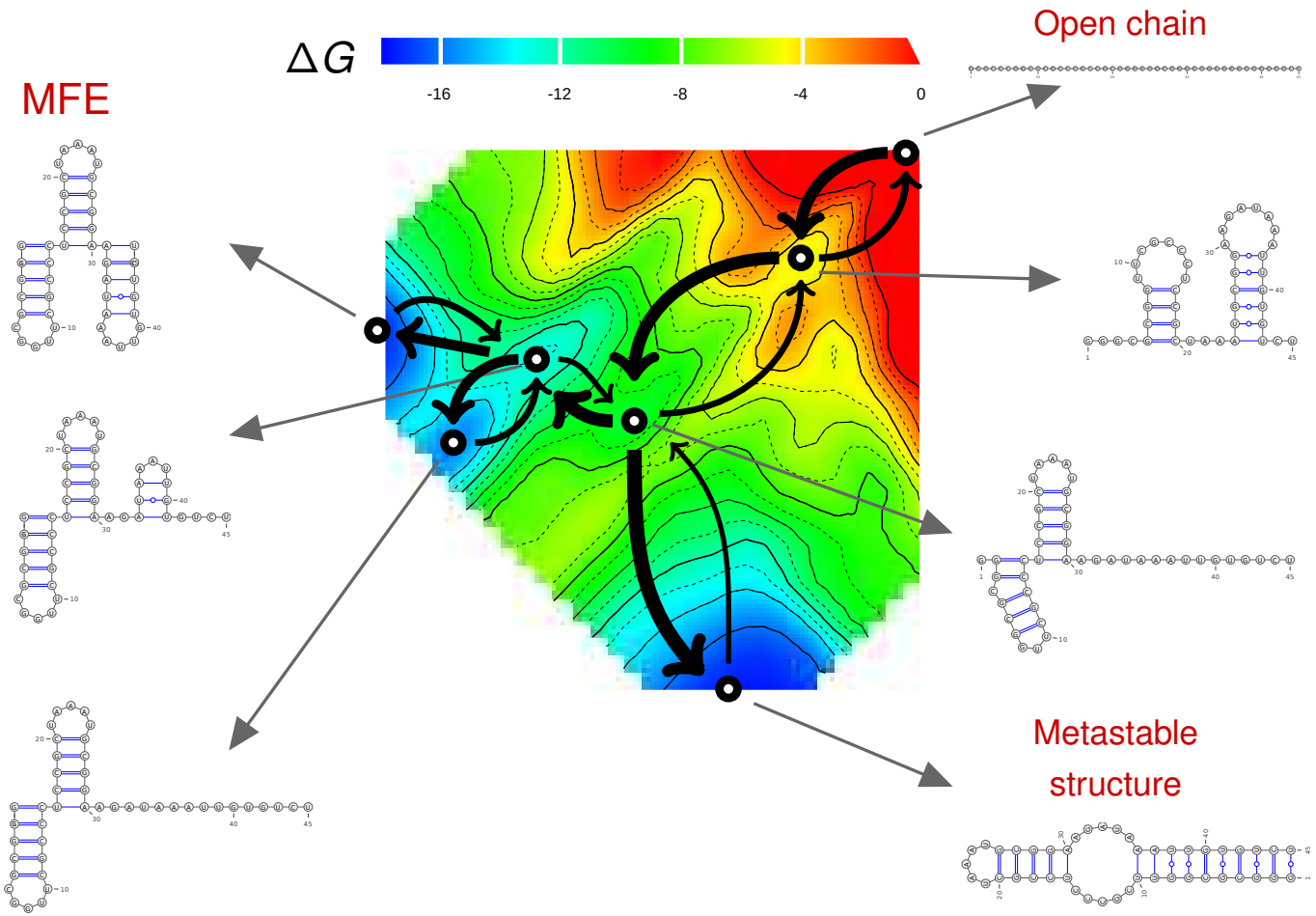
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

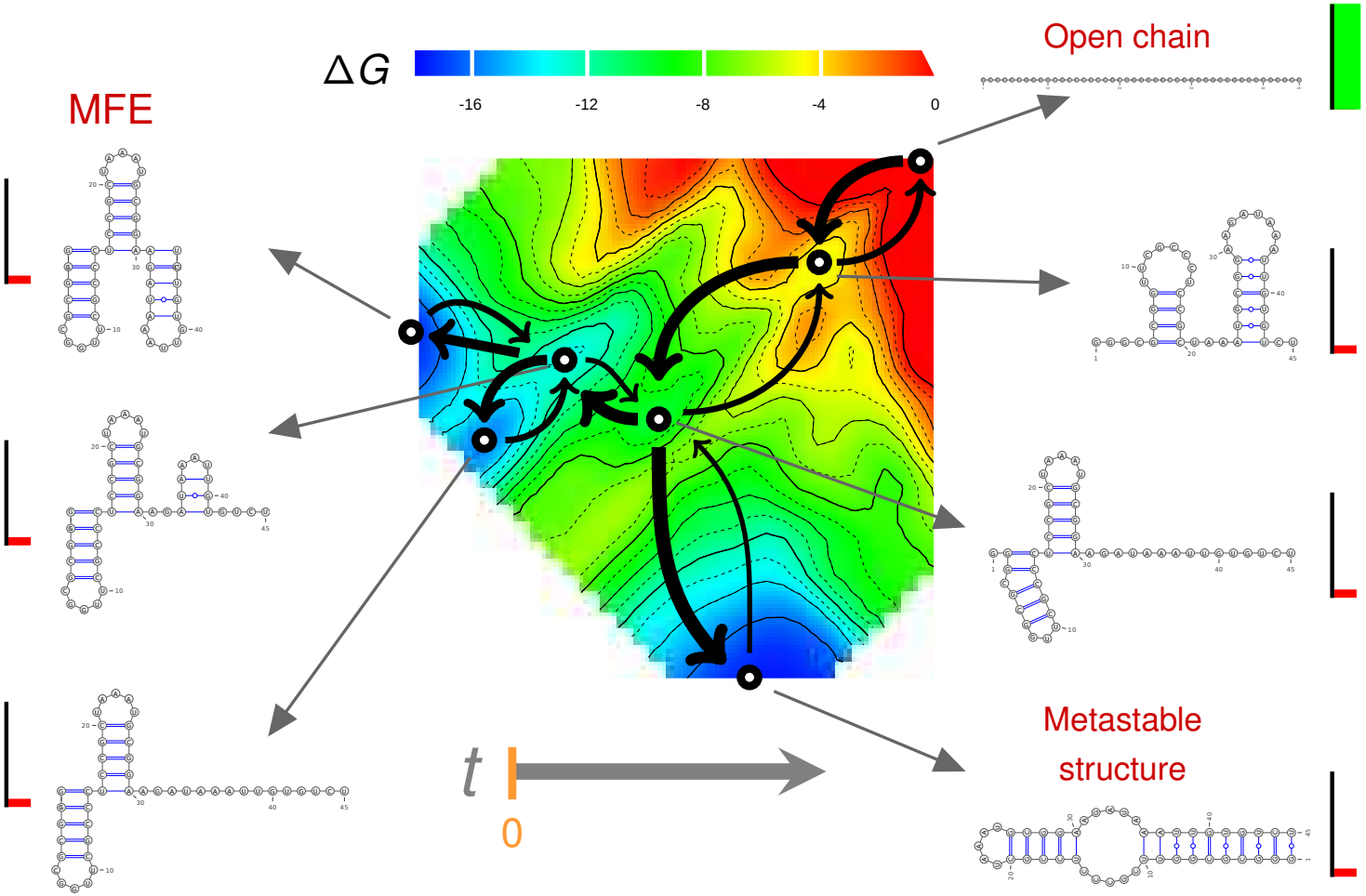
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

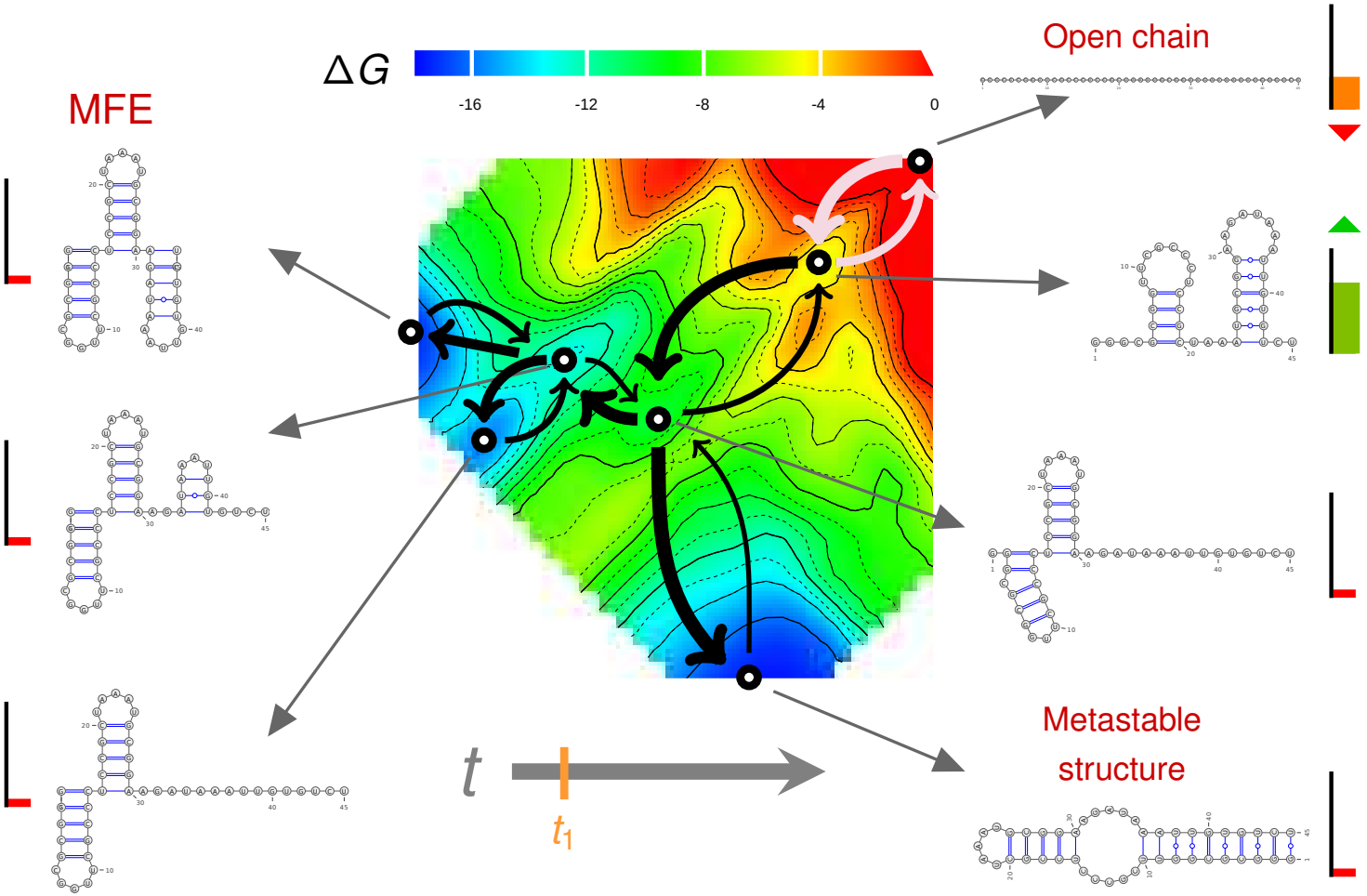
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

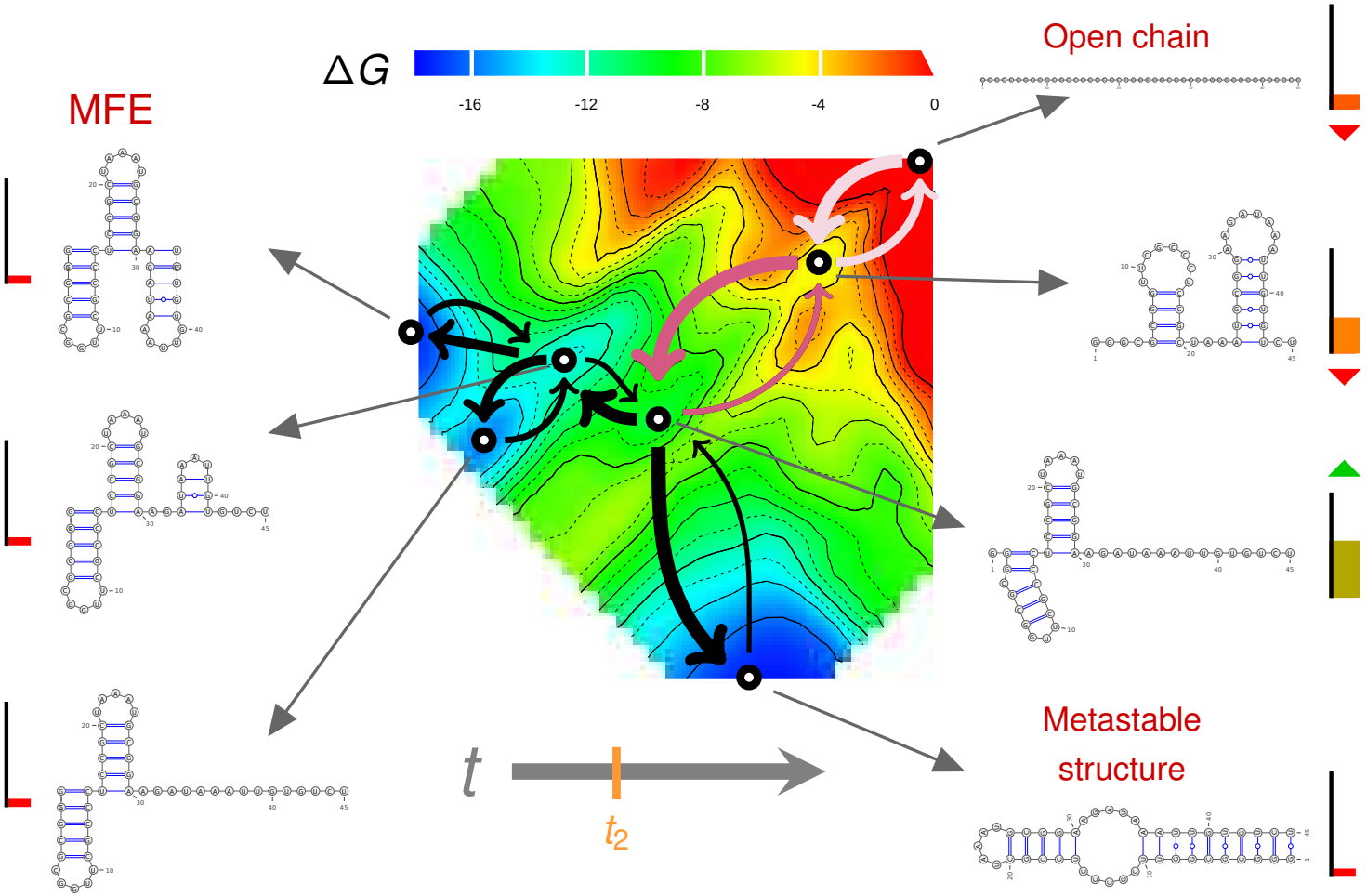
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

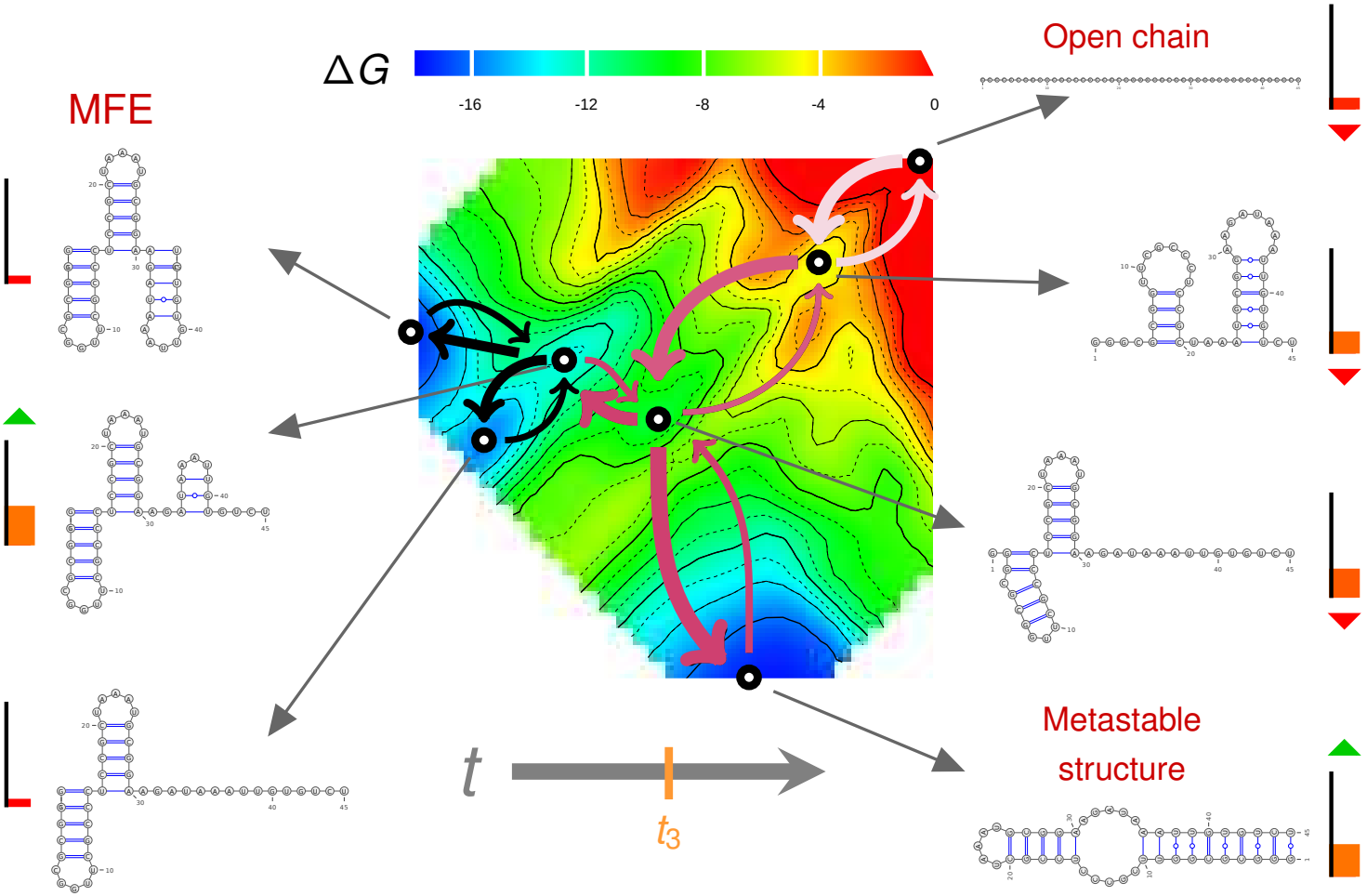
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

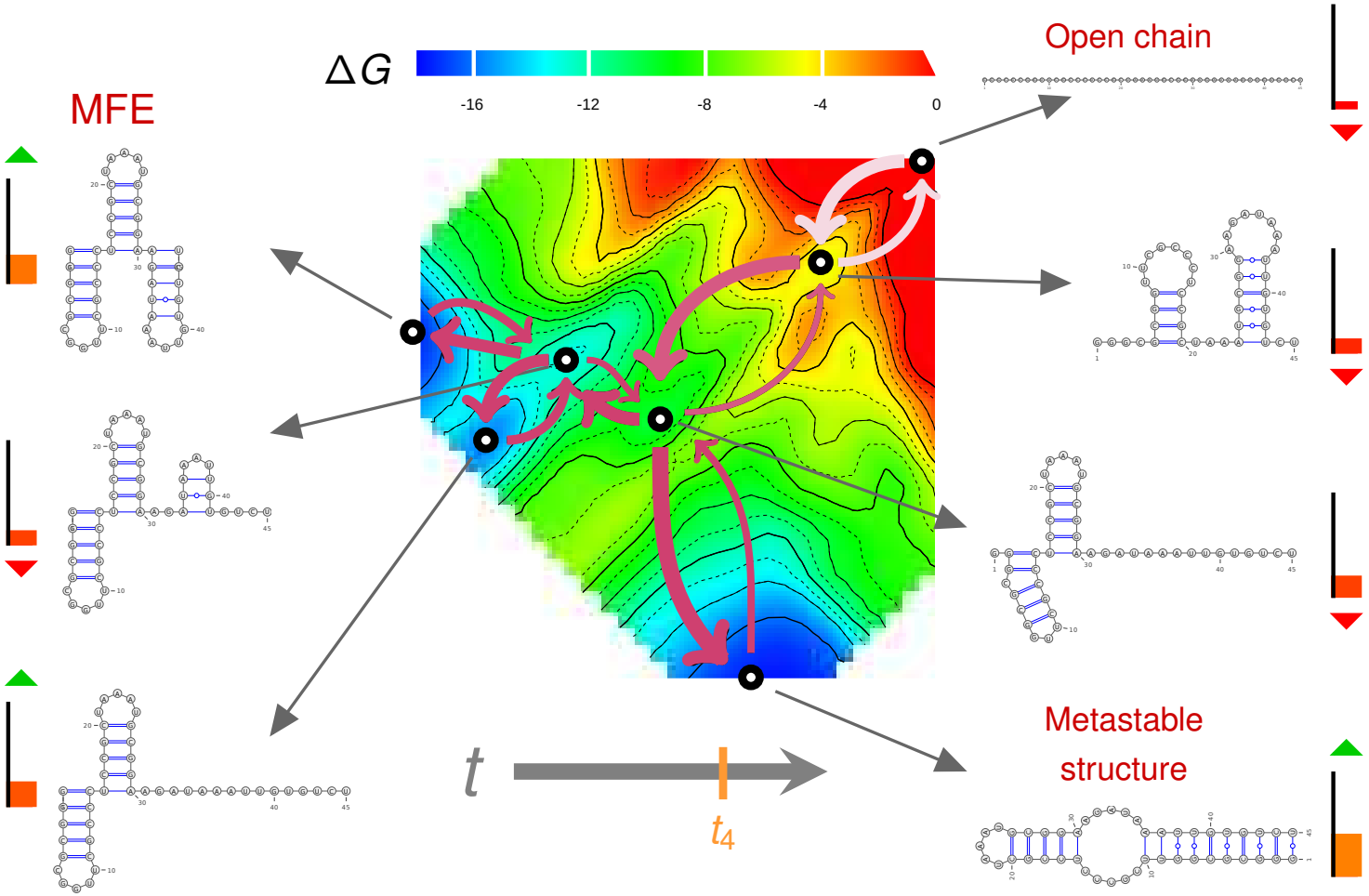
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

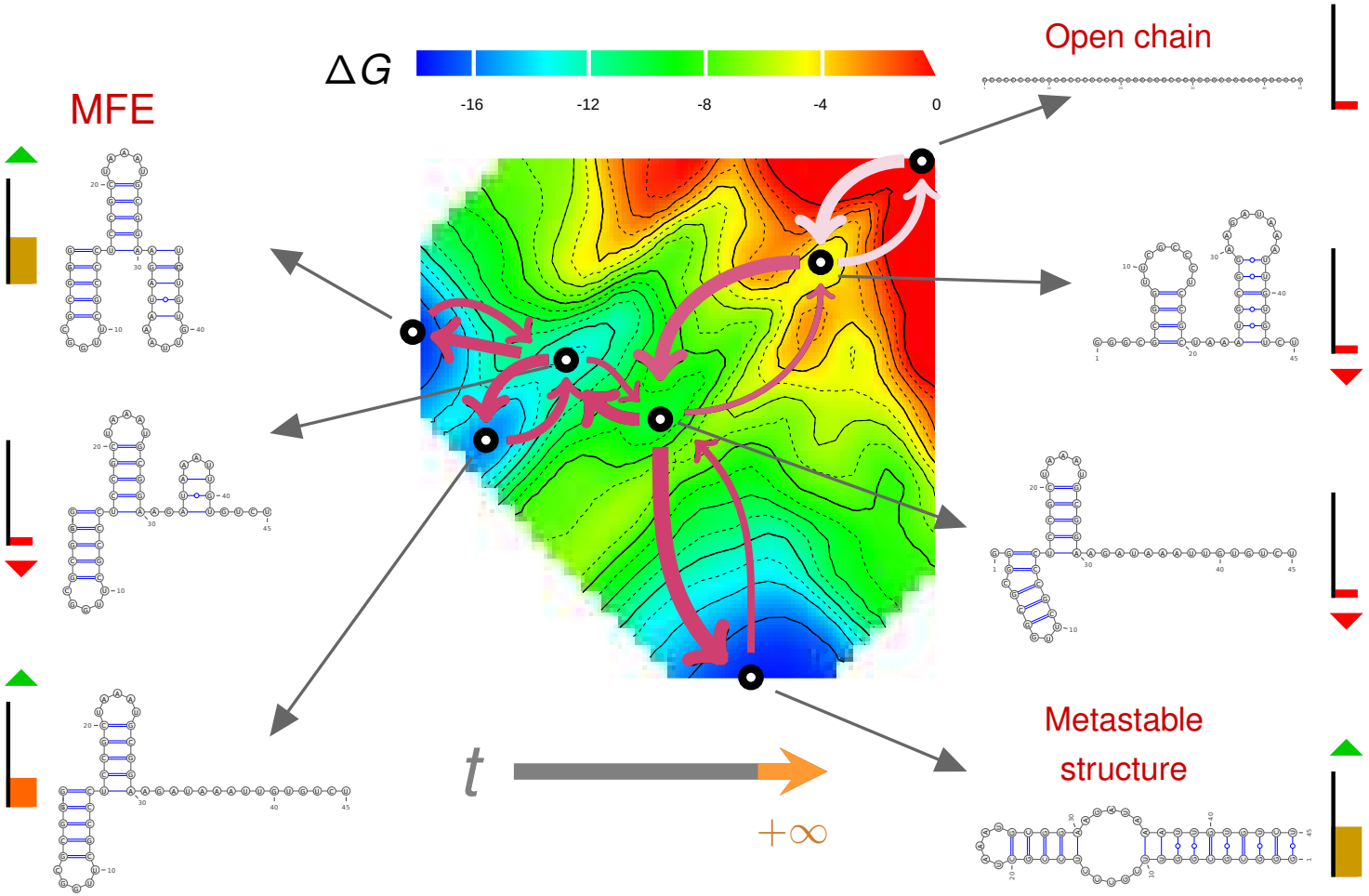
RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time $\rightarrow$ RNA kinetics!

RNA kinetics



Landscape : Lorenz *et al*, GCB'09, Structures : `Varna` - Darty *et al*, Bioinf. (2009)

Assuming a thermodynamic equilibrium sometimes misrepresents the reality of RNA folding in finite time → RNA kinetics!

ANR/FWF RNALands
Sampling 2D kinetics

Andrea Tanzer
PI, TBI Vienna

Yann Ponty
PI, Inria/Polytechnique

Alain Denise
LRI/IGM Paris-Sud

Ronny Lorenz
TBI Vienna

Loic Paulevé
CNRS/LRI Paris Sud

Mireille Regnier
Inria Saclay

Hélène Touzet
Inria/LIFL, Lille

Maria Waldl
TBI Vienna, Austria

+ Juraj Michalik (PhD) & Christelle Rovetta (Postdoc)

Game theory: convergence of no-regret algorithms

Collaboration with Johanne Cohen (GALAC, LRI) and Panayotis Mertikopoulos (POLARIS, LIG).

Discret N-players game, each player i has a finite set of strategies S_i. At each step, each player i chooses a strategy in S_i and receives a payoff $u_i : S_1 \times S_2 \times ... \times S_n \rightarrow \mathbb{R}$.

Definition: No-regret algorithms

The regret is sublinear: $Regret(T) = max_{s \in S_i}[\sum_{t=0}^{T} u_i(s, s_{-i}(t))] - \sum_{t=0}^{T} u_i(s_i(t), s_{-i}(t)) = o(T)$

We showed that some no-regret algorithms also converge to *Nash equilibrium*.

Definition: Nash Equilibrium

A Nash equilibrium is a state ,s^*, where no player has incentive to change its strategy alone: $u_i(s_i^*; s_{-i}^*) \geq u_i(s_i; s_{-i}^*), \forall i \in N, \forall s_i \in S_i$.

Game theory: convergence of no-regret algorithms

Collaboration with Johanne Cohen (GALAC, LRI) and Panayotis Mertikopoulos (POLARIS, LIG).

Discret N-players game, each player i has a finite set of strategies S_i. At each step, each player i chooses a strategy in S_i and receives a payoff $u_i : S_1 \times S_2 \times ... \times S_n \to \mathbb{R}$.

Definition: No-regret algorithms

The regret is sublinear: $Regret(T) = max_{s \in S_i}[\sum_{t=0}^{T} u_i(s, s_{-i}(t))] - \sum_{t=0}^{T} u_i(s_i(t), s_{-i}(t)) = o(T)$

We showed that some no-regret algorithms also converge to *Nash equilibrium*.

Definition: Nash Equilibrium

A Nash equilibrium is a state ,s^*, where no player has incentive to change its strategy alone: $u_i(s_i^*; s_{-i}^*) \geq u_i(s_i; s_{-i}^*), \forall i \in N, \forall s_i \in S_i$.

Game theory: convergence of no-regret algorithms

Collaboration with Johanne Cohen (GALAC, LRI) and Panayotis Mertikopoulos (POLARIS, LIG).

Discret N-players game, each player i has a finite set of strategies S_i. At each step, each player i chooses a strategy in S_i and receives a payoff $u_i : S_1 \times S_2 \times ... \times S_n \to \mathbb{R}$.

Definition: No-regret algorithms

The regret is sublinear: $Regret(T) = max_{s \in S_i}[\sum_{t=0}^{T} u_i(s, s_{-i}(t))] - \sum_{t=0}^{T} u_i(s_i(t), s_{-i}(t)) = o(T)$

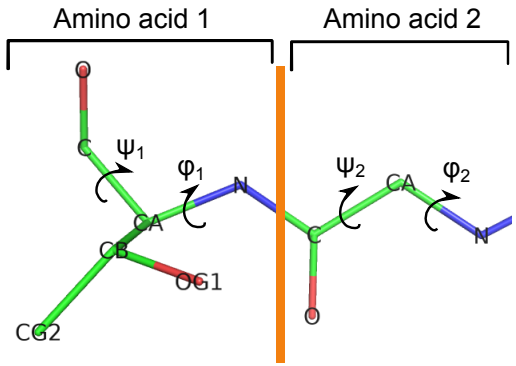
We showed that some no-regret algorithms also converge to *Nash equilibrium*.

Definition: Nash Equilibrium

A Nash equilibrium is a state ,s^*, where no player has incentive to change its strategy alone: $u_i(s_i^*; s_{-i}^*) \geq u_i(s_i; s_{-i}^*), \forall i \in N, \forall s_i \in S_i$.

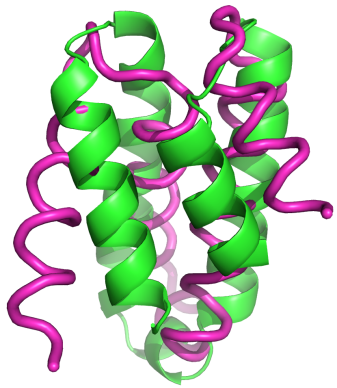
Protein folding with game theory

With advices of Raphaël Guerois and Jessica Andreani (CEA), and Frédéric Cazals (ABS, Nice).



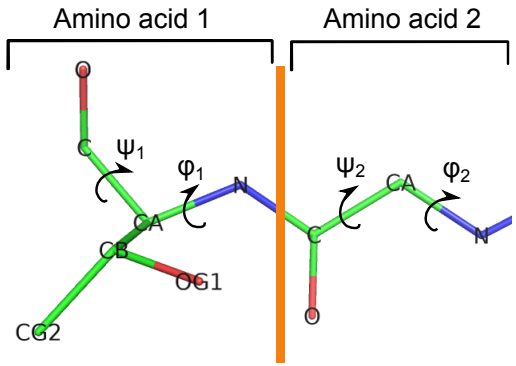
A protein is a sequence of amino acids. Each amino acid is a player. Strategies are couples of angles, the dihedral angles.
The payoffs correspond to the quality of the fold.

Encouraging results :



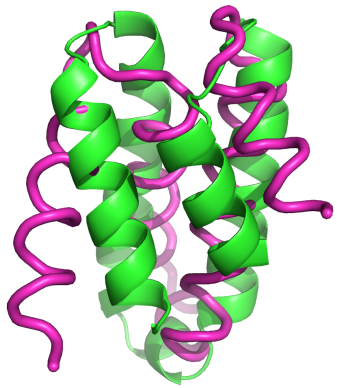
Protein folding with game theory

With advices of Raphaël Guerois and Jessica Andreani (CEA), and Frédéric Cazals (ABS, Nice).



A protein is a sequence of amino acids. Each amino acid is a player. Strategies are couples of angles, the dihedral angles.
The payoffs correspond to the quality of the fold.

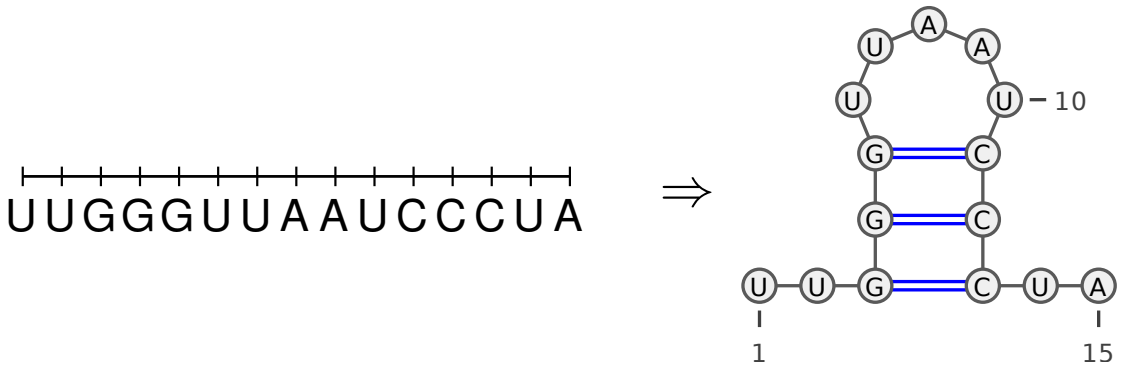
Encouraging results :



Introduction

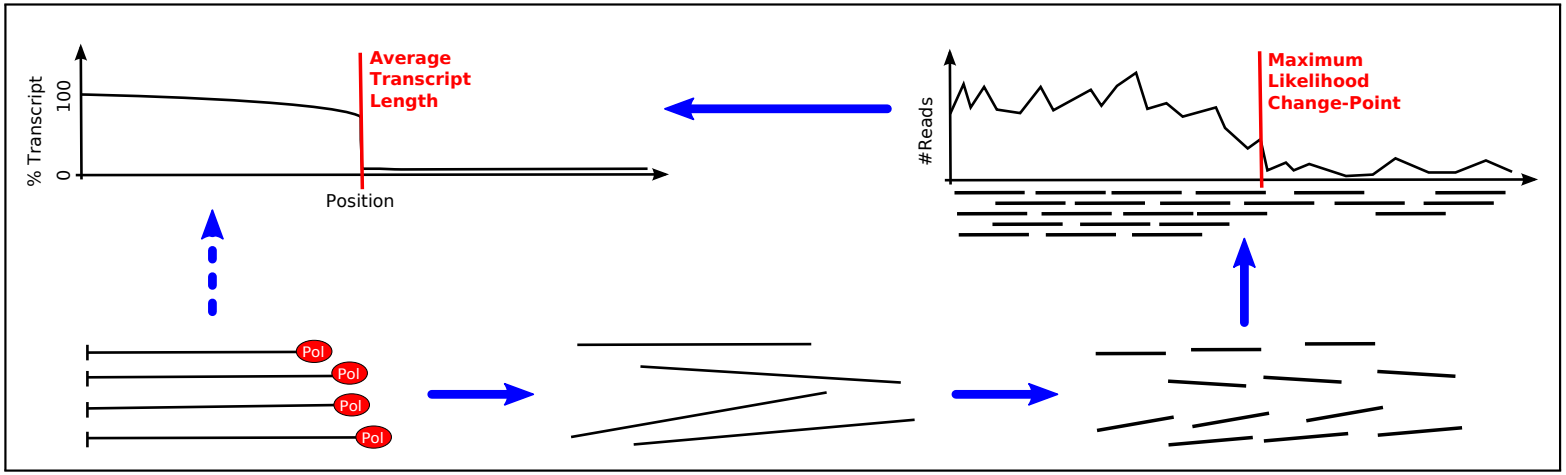
RNAs = **R**ibo**N**ucleic **A**cids

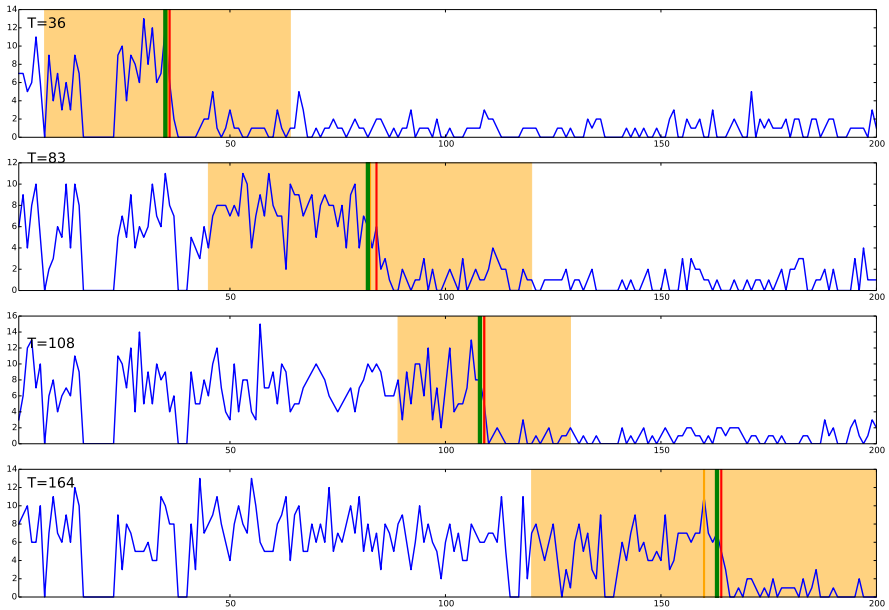
- Polymeric molecules composed of nucleotides A, C, G and U
- Diverse functions:
  - Coding and decoding $\rightarrow$ Coding RNAs
  - Regulatory $\rightarrow$ Non-coding RNAs
- Non-Coding sequences : function depends on structure
- Approximation : secondary structure : basepairing interactions (C-G, A-U, G-U)



Structure created with Varna(K. Darty *et al.*, 2009)

Segmentation for transcription rate estimation





- Goal: Estimate transcription rates for nascent RNAs
- Method: Conditional segmentation (prog. dyn.)
- Partners: Univ. Wuhan, Univ. Paris-Sud
- PHC XU GUANGQI project (2017-2018)