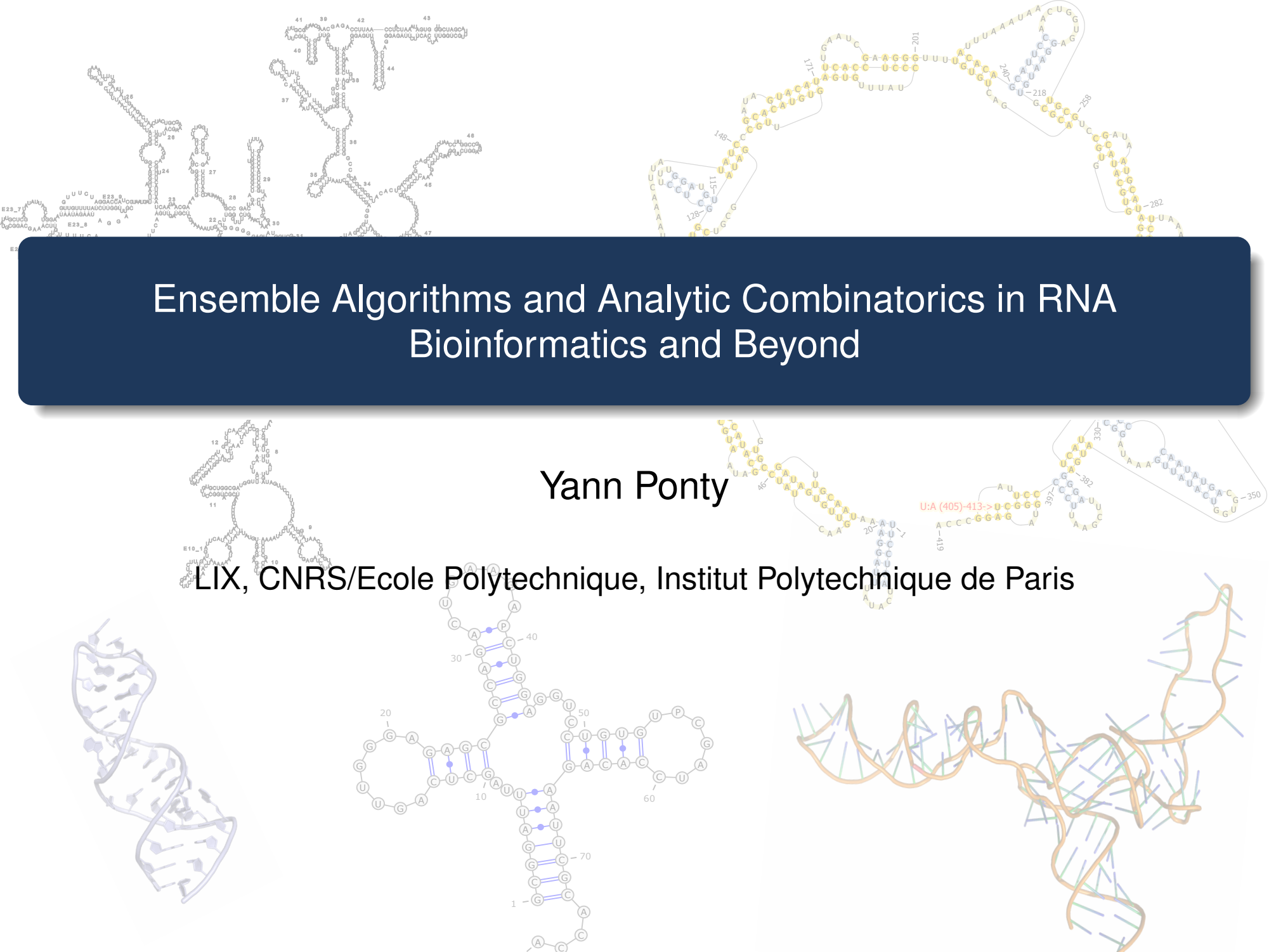
Ensemble Algorithms and Analytic Combinatorics in RNA Bioinformatics and Beyond

Yann Ponty

LIX, CNRS/Ecole Polytechnique, Institut Polytechnique de Paris



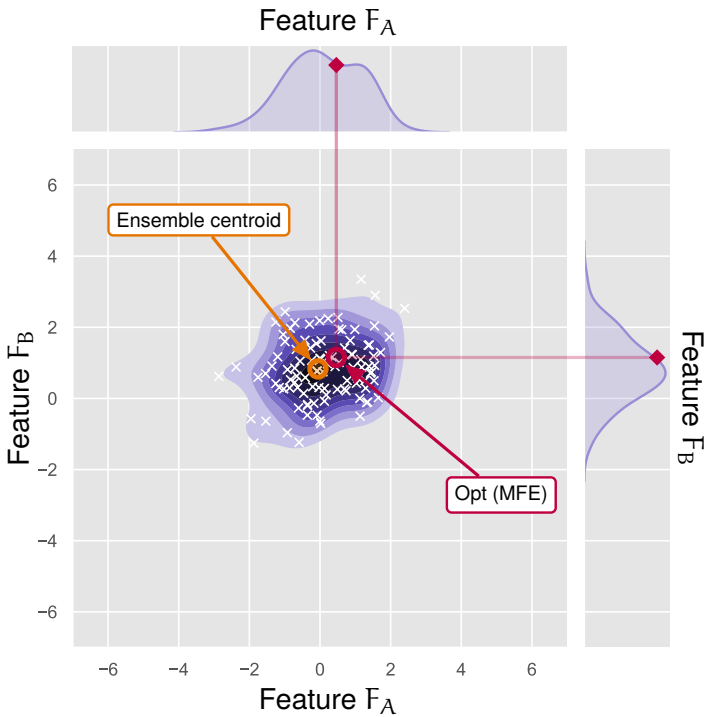
Introduction

It's my party, and I'll cry if I want to. *Lesley Gore*

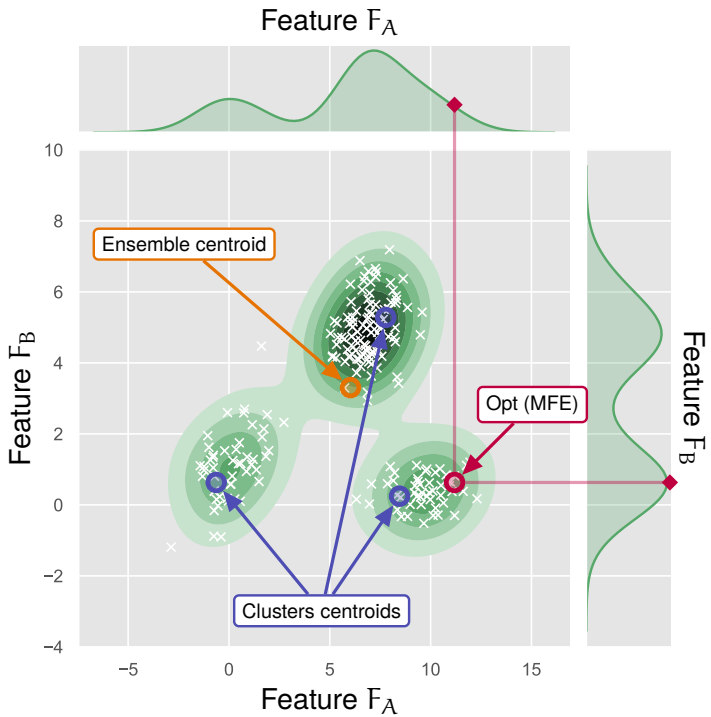
Beyond optimality

Why do we use optimization?

- For fun: after all, who doesn't just love algorithms?
- For money: operations research, network design...
- For love of exhaustivity: exact negative results
- To predict the unobservable: probabilistic modeling + Occam's razor



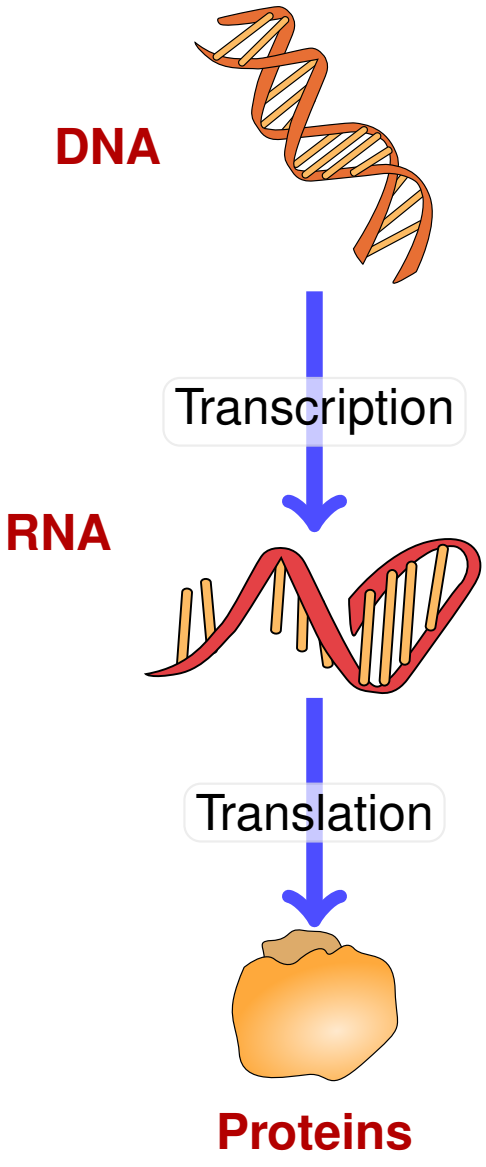
A – Concentrated ensemble



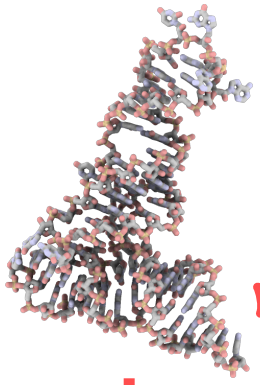
B – Fragmented ensemble

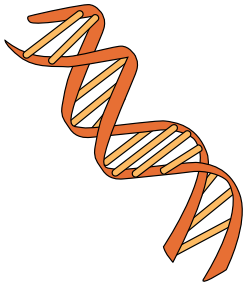
→ Ensemble analyses and algorithms

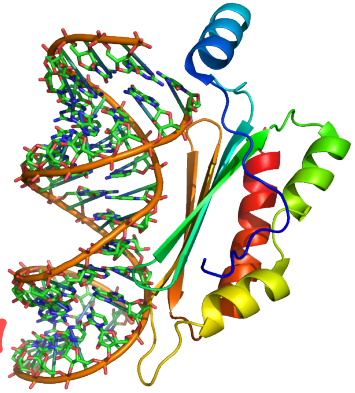
Fundamental dogma of molecular biology

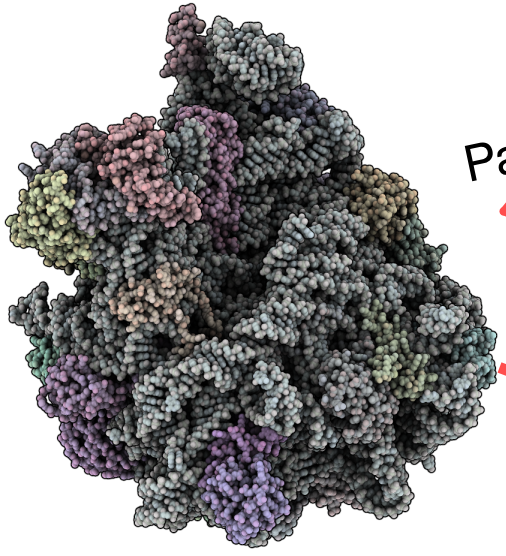


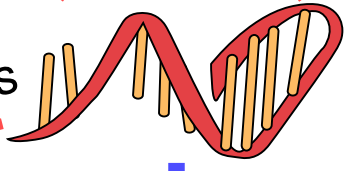
Fundamental dogma of molecular biology (v2.0)

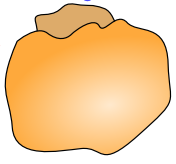












DNA

Transfer

Transcription

Maturation

RNA

Participates

Regulation

Translation

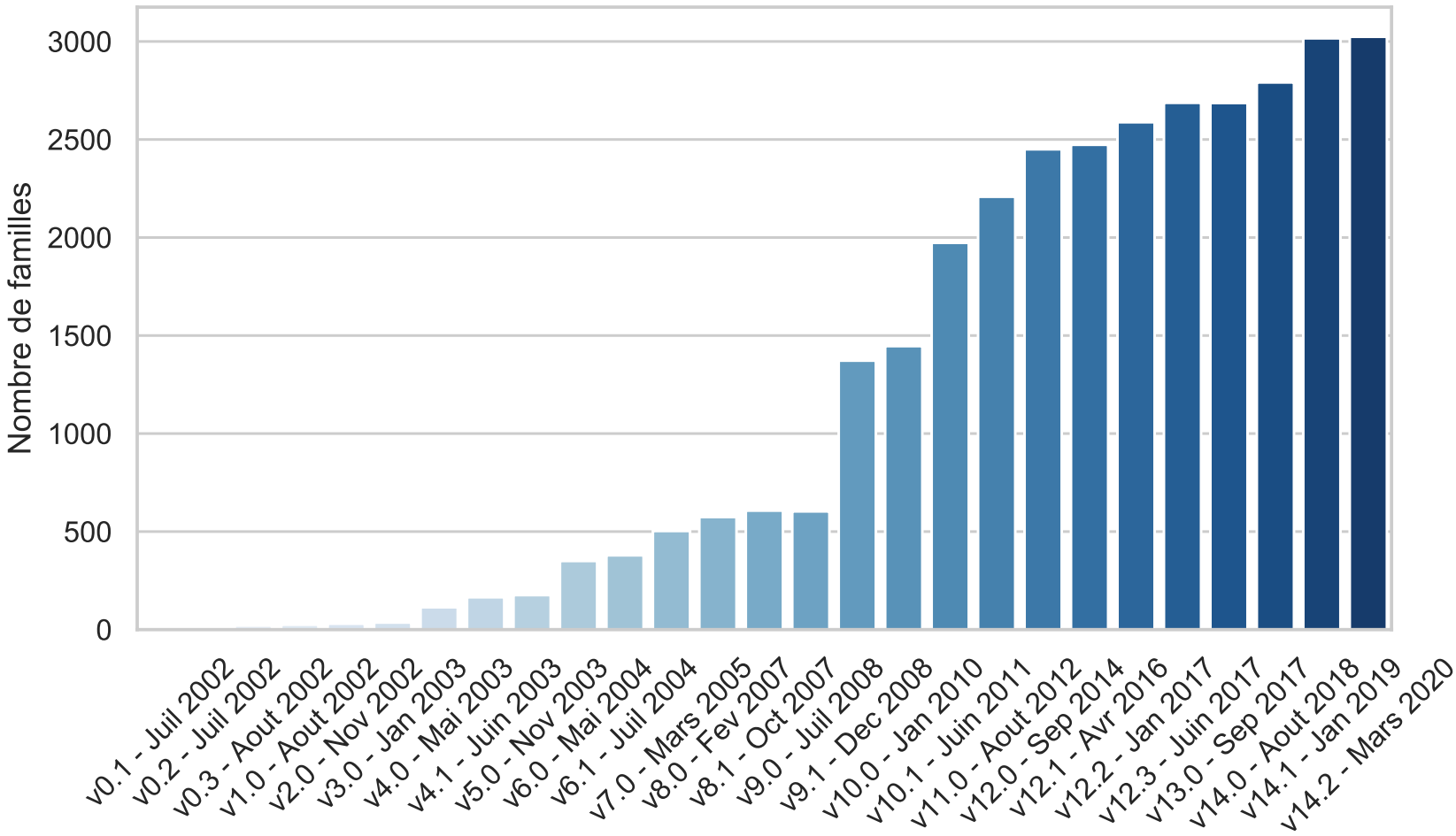
Synthesis

Proteins

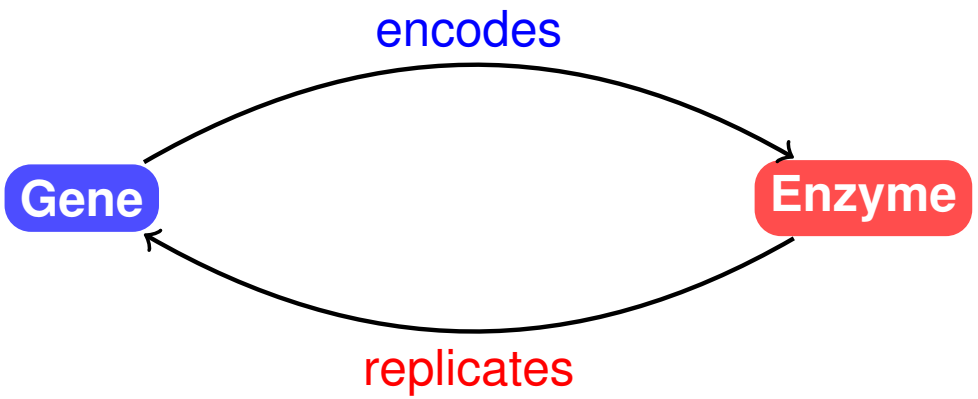
RNA functions

- Messengers
- Translation
- Regulation
- Enzymes
- Modifications
- Editing. . .

Fundamental dogma of molecular biology (v2.0)



A chicken and eggs paradigm at the origin of life

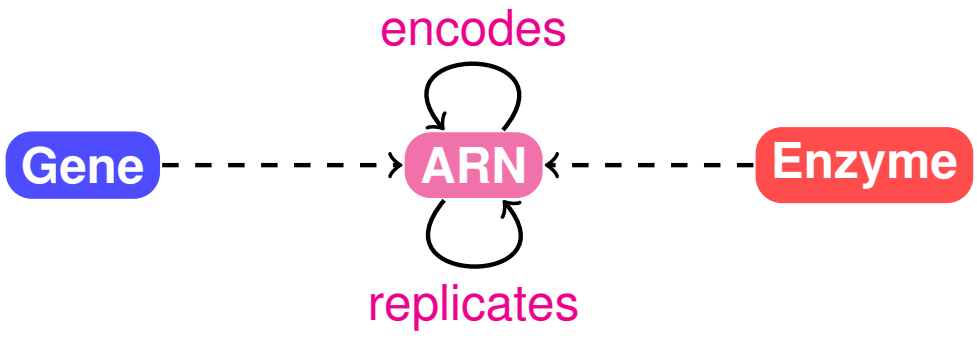


A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

[. . .] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

R. Dawkins. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

A chicken and eggs paradigm at the origin of life



A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

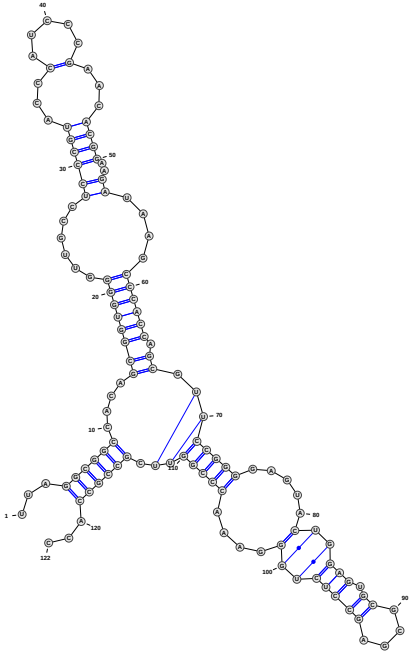
[. . .] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

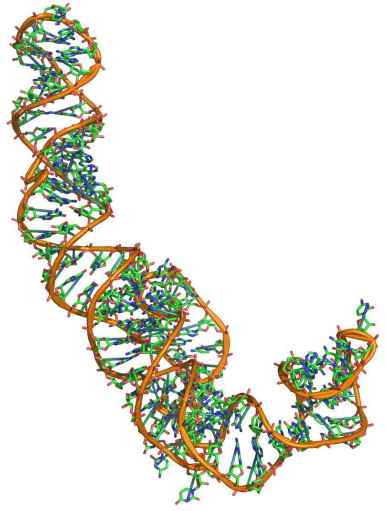
R. Dawkins. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

RNA structure(s)

RNA = Linear Polymer = Nucleotides sequence $w \in \{A, C, G, U\}^\star$

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```



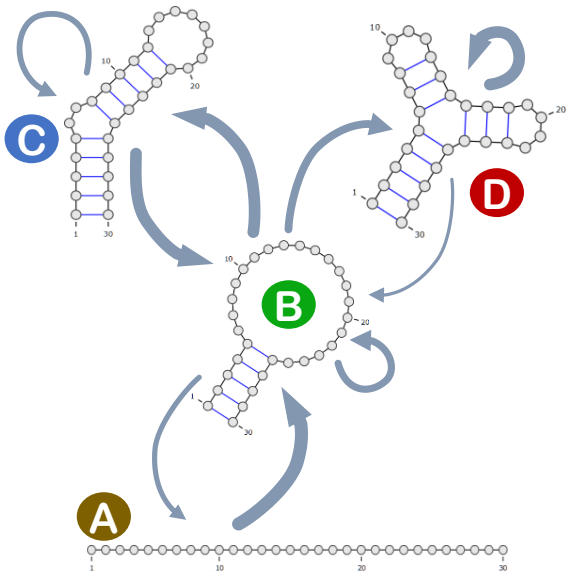


Primary struct. Secondary (2D) struct. Tertiary ($\approx$ 3D) struct.

Source: 5s rRNA (PDBID: 1K73:B)

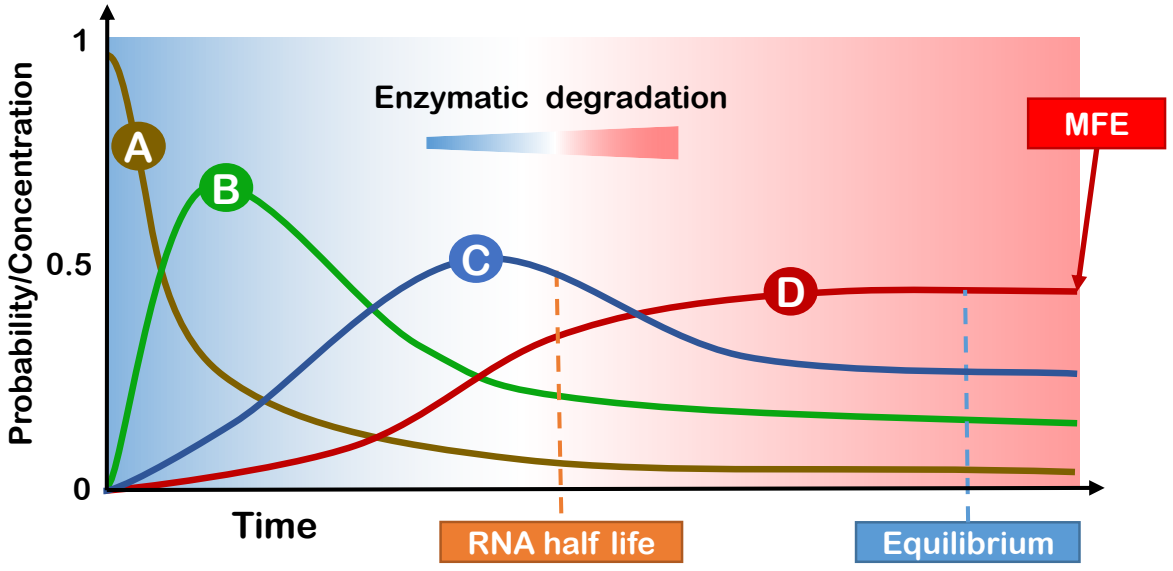
Secondary structure S = Set of base-pairs $(i, j) \in [1, n]^2$ such that:

- Monogamy: Each position $x \in [1, n]$ involved in at most one base-pair
- No crossing base-pairs: $\forall (i, j) \in S, \nexists (k, l) \in S$ such that $i < k < j < l$
- Steric constraints: $\forall (i, j) \in S, |i - j| > \theta$ ($\theta = 1$ or 3)
- Valid base pairs: $\forall (i, j) \in S$, $\{w_i, w_j\}$ is either {G, C},{A, U}, or {G, U}



A – Kinetic Landscape

Continuous-time Markov chain



B – Evolution of concentrations

Given free-energy $E : \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^\star \times \mathcal{S} \to \mathbb{R}$, at the Boltzmann equilibrium one has:

$$\mathbb{P}(S \mid w) \propto e^{-E(w,S)/RT}$$

- Minimum Free-Energy (MFE): Relevant structure = Most stable/probable
- Partition function: Equilibrium properties of Boltzmann ensemble
- Kinetics: Finite-time evolution of concentrations/probabilities

Part 1. Applied Analytic Combinatorics

All models are wrong, but some are useful. *George Box*

$\mathcal{S}$: Combinatorial class, *i.e.* (possibly infinite) set of *things*

s_n: Number of things of size/length n

$$S(z) = \sum_{n \geq 0} s_n \, z^n$$

In enumerative combinatorics:

- Only size matters...
- $S(z)$ easily obtained from principled decomposition of $\mathcal{S}$ $\rightarrow$ Symbolic method
- Asymptotic equivalent of s_n through singularity analysis $\rightarrow$ Analytic combinatorics
- Bivariate form $S(z, u) = \sum_{n,k} s_{n,k} \, z^n \, u^k$

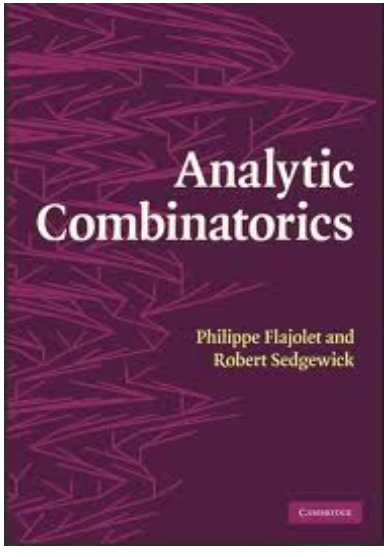
Enumerative combinatorics can be used to characterize the precise (asymptotic) behavior of abstract models for RNA sequence/structure.

Typical problems

- How many secondary structures on n nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between 3'/5' ends? [Clote, Ponty, and Steyaert, 2012]

The symbolic method, a generic framework for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics

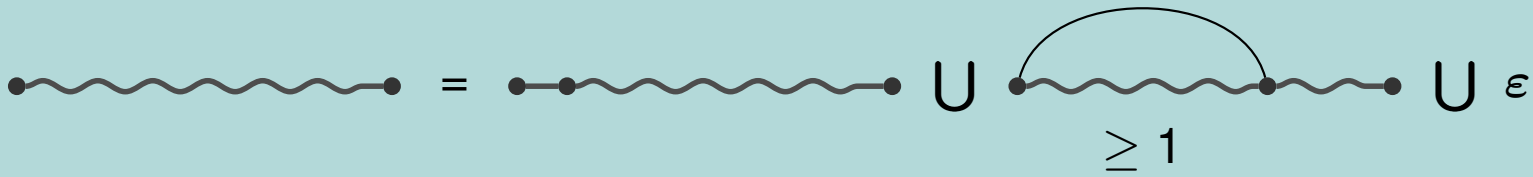


[Flajolet and Sedgewick, 2009]

Goal: Generating function $S(z) = \sum_{n\geq 0} s_n z^n$

where s_n : #Secondary structures of length n

1



2

$$\begin{aligned} S &\rightarrow \bullet\, S \mid (\, S_{>0}\,)\, S \mid \varepsilon \\ \mathbf{T} &\rightarrow \bullet\, S \mid (\, \mathbf{T}\,)\, S \end{aligned}$$

3

$$S(z) = \frac{1-z+z^2-\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2}$$

4

$$\rho = \frac{3-\sqrt{5}}{2} = 1-\phi$$

$$s_n = \sqrt{\frac{15+7\sqrt{5}}{8\pi}} \cdot \frac{\left(\frac{3+\sqrt{5}}{2}\right)^n}{n\sqrt{n}}(1+\mathcal{O}(1/n)) \sim 1.1 \cdot \frac{2.6^n}{n\sqrt{n}}$$

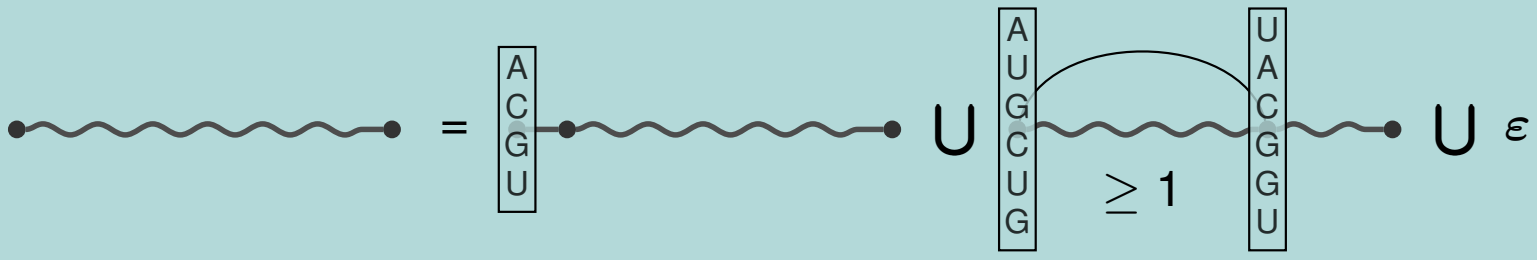
[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

Expected #2D struct. compatible with random RNA

Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$

where s_n = #Compatible (Sequence/Sec. struct.) pairs of length n

1



2

$$S \to \begin{matrix}(_A\ T)_U\ S\\ (_U\ T)_A\ S\\ (_G\ T)_C\ S\\ (_C\ T)_G\ S\\ (_U\ T)_G\ S\\ (_G\ T)_U\ S\end{matrix} \mid \begin{matrix}\bullet_A\ S\\ \bullet_U\ S\\ \bullet_G\ S\\ \bullet_C\ S\end{matrix} \mid \varepsilon \qquad T \to \begin{matrix}(_A\ T)_U\ S\\ (_U\ T)_A\ S\\ (_G\ T)_C\ S\\ (_C\ T)_G\ S\\ (_U\ T)_G\ S\\ (_G\ T)_U\ S\end{matrix} \mid \begin{matrix}\bullet_A\ S\\ \bullet_U\ S\\ \bullet_G\ S\\ \bullet_C\ S\end{matrix}.$$

3

$$S(z) = \frac{1-4z+6z^2-\sqrt{1-8z+4z^2-48z^3+36z^4}}{12z^2}$$

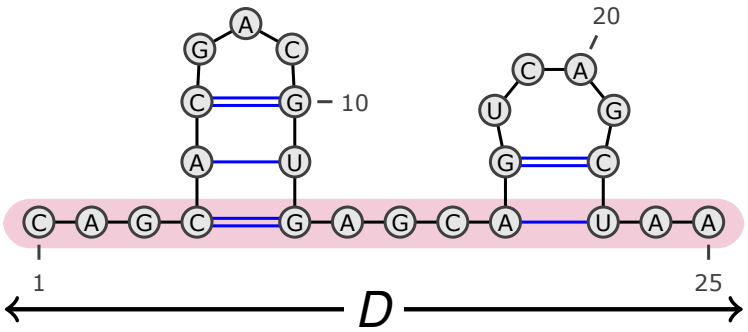
4

$$\rho = \text{InfSing}(1-8z-4z^2-48z^3+36z^4) \quad 1/\rho \approx 8.164$$

$$s_n \in \Theta\left(\frac{\rho^{-n}}{n\sqrt{n}}\right) \quad \to \quad \text{Expected\#Sec.Str.} = s_n/4^n \in \Theta(2.04^n/n\sqrt{n})$$

[Zuker and Sankoff, 1984]

5'–3' Distance: Expectation through derivation



Goal: Bivariate Generating Function

$$S(z,u) = \sum_{n\geq 0}\sum_{d\geq 0} s_{\theta,n,d}\, z^n\, u^d$$

$s_{\theta,n,d}$ = #2D structures of length n having 5'-3' distance d

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

3

$$E_\theta(z) = \left.\frac{\partial T(z,u)}{\partial u}\right|_{u=1} = \frac{\begin{pmatrix} 2-9z+14z^2-8z^3+2z^5 \\ +z^{\theta+2}(-4+10z-10z^2+2z^3)+z^{2\theta+4}(2-z) \\ -(2-5z+4z^2-2z^{\theta+2}+z^{\theta+3})\sqrt{\Delta_\theta}\end{pmatrix}}{2(1-z)^2z^4}$$

$$\Delta_\theta := 1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}$$

4

$$D_n \sim \frac{2-5\rho+4\rho^2-2\rho^{\theta+2}+\rho^{\theta+3}}{(1-\rho)\rho^2} - 1,\ \rho \text{ smallest root of } \Delta_\theta = 0$$

[Clote, Ponty, and Steyaert, 2012]

Homopolymer model = All positions can form base pairs

Random RNAs are almost circularized [Clote, Ponty, and Steyaert, 2012]

Expected 5'–3' dist. is asymptotically constant at the Boltzmann equilibrium

Efficient Boltzmann sampling [Ponty, 2008]

Average/worst case complexities of classic Boltzmann sampling in $\Theta(n\sqrt{n})/\Theta(n^2)$, and can be improved to $\Theta(n \log n)$ worst-case

Connectivity of RNA networks [Surujon, Ponty, and Clote, 2019]

RNA networks induced by BP addition, removal and shifts are not small world

Scarcity of RNA phenotypes [Yao, Chauve, Regnier, and Ponty, 2019]

The proportion of designable RNA 2D structures is exponentially decreasing on the sequence length

Hua-Ting Yao

Part 2. Random generation

Everything we care about lies somewhere in the middle, where pattern and randomness interlace. *James Gleick*

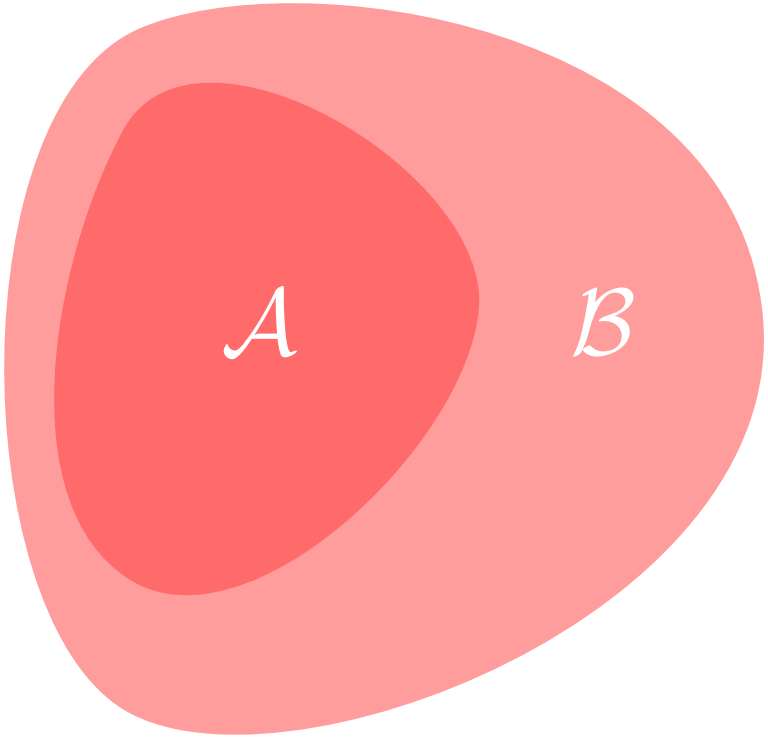
The power of rejection

Goal

Generate (pseudo-)random, uniformly-distributed, objects from a combinatorial class $\mathcal{A}$

If $\mathcal{A}$ can be specified (grammar):
$\rightarrow$ Precompute/use derivations probs

But many classes cannot be specified!



Rejection method:
- Find efficiently generated $\mathcal{B}$, $\mathcal{A} \subset \mathcal{B}$
- Draw random objects uniformly from $\mathcal{B}$
- Reject objects in $\mathcal{B} \setminus \mathcal{A}$

$\rightarrow$ Random uniform generator for $\mathcal{A}$

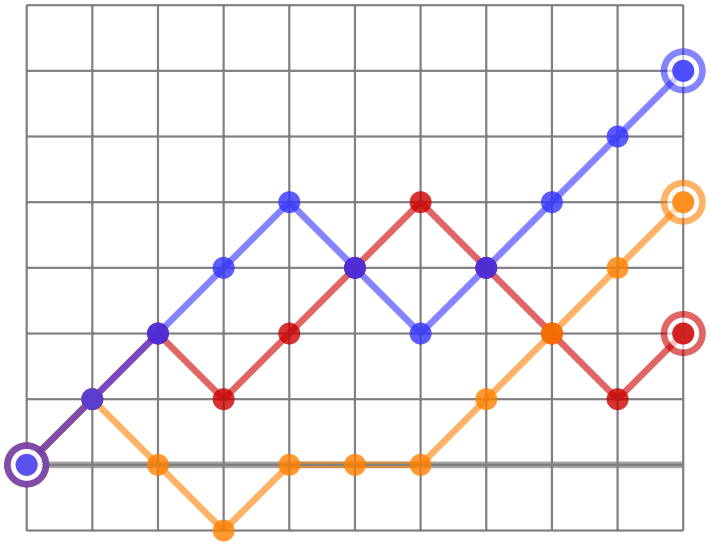
Expected complexity $\mathfrak{C}_{\mathcal{A}}$: $\dfrac{\mathfrak{C}_{\mathcal{B}} \times |\mathcal{B}|}{|\mathcal{A}|}$

Extensions: Non-uniform distr for $\mathcal{B}$ and/or $\mathcal{A}$, anticipated rejection...

Culminating paths

Culminating paths $\mathcal{C}$: Walks over $\{\searrow, \nearrow\}$ that start at $(0, 0)$, finish at (n, h) (length n), such that:

1. Always stay positive
2. Achieve highest height on their last step



Goal

Generate culminating path of length n uniformly at random

Asymptotics: $c_n = \kappa \frac{2^n}{n}(1 + \mathcal{O}(1/P(n)))$

Specification unattainable (transcendental, non D-finite, gen. fun.)!

Remark: Conditions 1 and 2 play symmetric roles up to 180° rotation
$\rightarrow$ Candidate superset: $\mathcal{B} := \{w_1.\overline{w_2} \mid w_1 \text{ and } w_2 \text{ positive} \vee |w_1| = |w_2|\} \supset \mathcal{C}$

Linear time algorithm for objects in $\mathcal{B}_n + |\mathcal{B}_n|/c_n \in \Theta(1)$ [Barcucci et al., 1994]

$\rightarrow$ Linear rejection algorithm for highly complex objects [Bousquet-Mélou and Ponty, 2008]

Rejection to control additive features

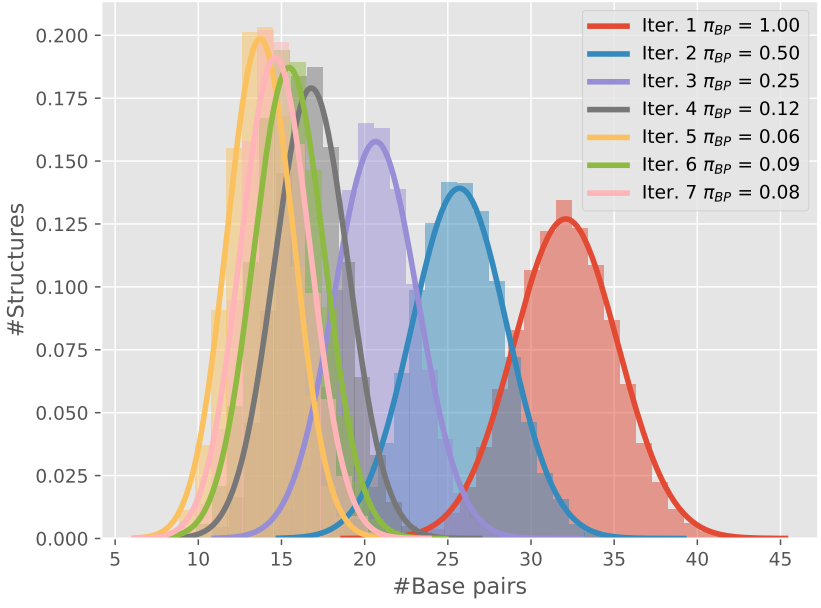
Specifiable class $\mathcal{A}$ + Additive feature $F : \mathcal{A} \to \mathbb{N}$

Goal: Uniform generation in $\mathcal{A}^k = \{w \in \mathcal{A} \mid F(w) = k\}$, typically not specifiable

Fortunately, efficient generation often possible for $\mathcal{A}$ in π-weighted distribution:

$$\mathbb{P}_\pi(w \in \mathcal{A}_n) = \frac{\pi^{F(w)}}{\sum_{w' \in \mathcal{A}} \pi^{F(w')}}$$

[Denise, Ponty, and Termier, 2010]



Multidimensionnal Boltzmann sampling [Bodini and Ponty, 2010]

Idea: Use π to control expectation + Exploit (provable) concentration
$\to \Theta(n^{1+d/2})$ rejection generation with d features in connected grammars

Circumventing the cost of rejections

Redundancy is uninformative and wasteful for the statistical analysis of comb. classes

Weighted coupons collector [Du Boisberranger, Gardy, and Ponty, 2012]

In weighted distributions, at the full collection, avg #copies is exponential on n.

\+ Recursive algorithms adapted to sample without replacement [Lorenz and Ponty, 2013]

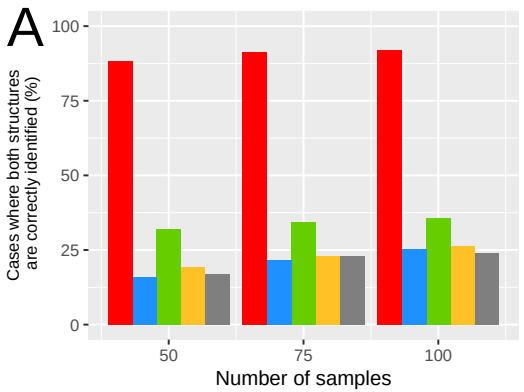
Estimating from a collection $\mathbf{t}$ of m distinct objects, sampled without replacement:

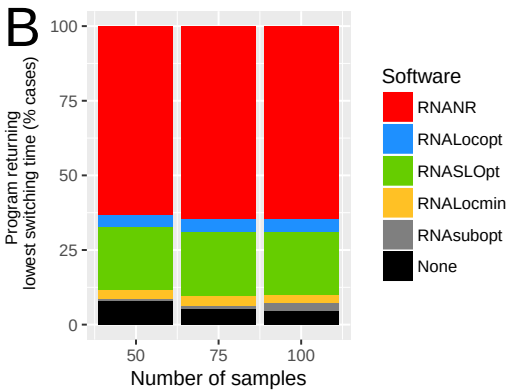
$$\tilde{F}(\mathbf{t}) = \frac{1}{m}\sum_{i=1}^{m} F(t_i)\Big(1 - \sum_{t\in\Theta_{i-1}} \mathbb{P}(t) + (m-i)\times\mathbb{P}(t_i)\Big) \text{ with } \Theta_i := (t_1,\ldots,t_i)$$

Non-redundant estimator [Rovetta, Michálik, Lorenz, Tanzer, and Ponty, 2020]

$\tilde{F}$ is unbiased, consistent, and has lower variance than naive estimator.

Applications to RNA kinetics [Michálik, Touzet, and Ponty, 2017]



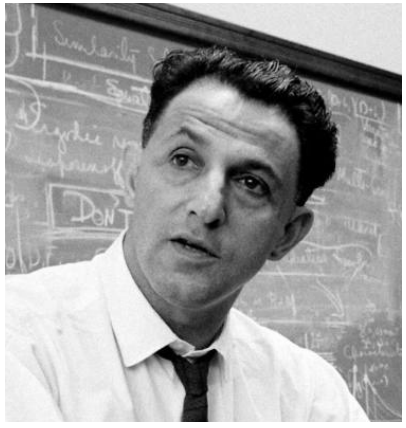


Juraj Michalik

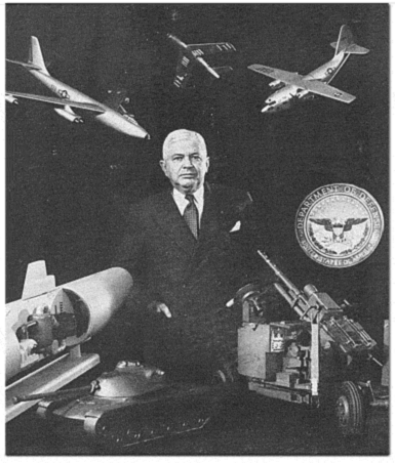
Part 3. Dynamic programming

It was not even something a Congressman could object to. . .
Richard Bellmann

Dynamic programming



R. Bellman



C. E. Wilson

Dynamic Programming = Generic algorithmic technique for optimization

Principle: Given objective function E, express $\max_{S \in \Omega_I} E(S)$ for a problem I, as a monotonous function of its opt. values over subproblems

- Memoization: Efficiently compute optimal value of (sub)problems
- Backtrack: Recompose element of search space achieving optimal value

Benefits: exponential speedup over brute-force search, and robustness to modifications of objective function parameters

$\Rightarrow$ Ubiquitous in Bioinformatics: Alignment, Reconciliations, RNA folding...

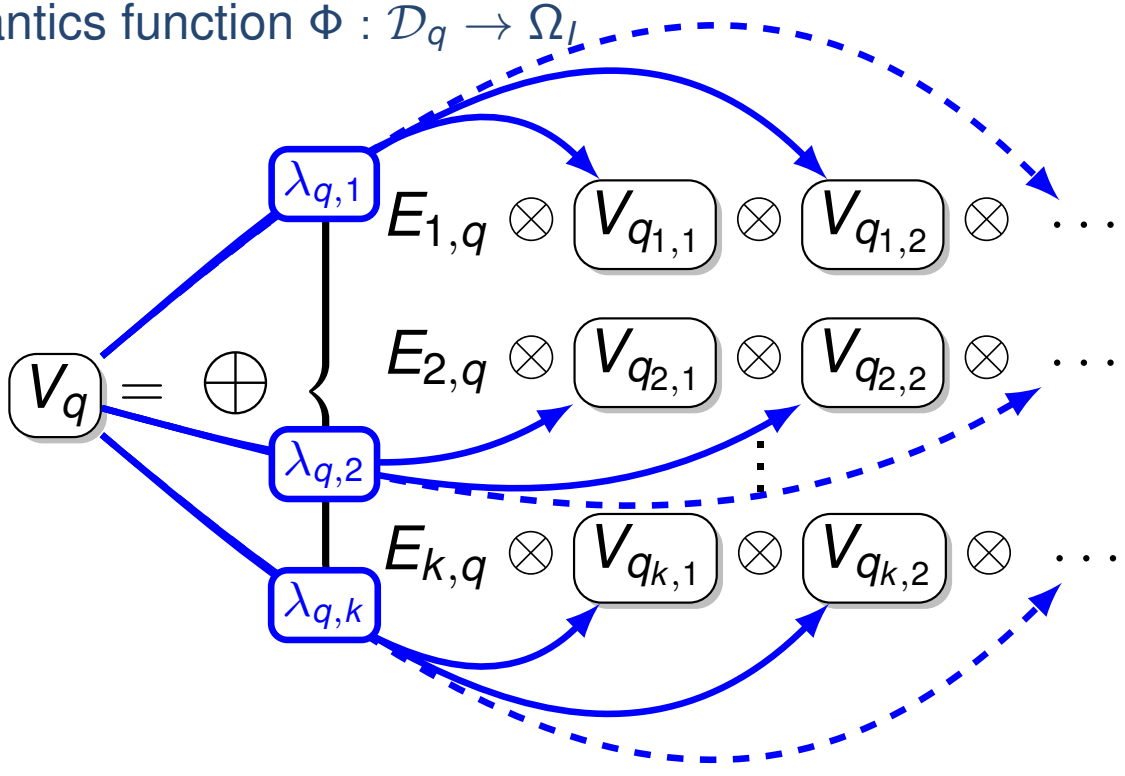
Reasoning structurally about dynamic programming

- I : Instance (a.k.a. problem)
- Q : State space for dyn. prog. scheme (LHS terms, I initial state)
- Ω_q : Search space associated with state q
- $\mathcal{T}_q$: Terms generated by DP scheme from state $q \in Q_i$
- Semantics function $\Phi : \mathcal{D}_q \to \Omega_I$

$$V_q = \bigoplus \begin{cases} E_{1,q} \otimes V_{q_{1,1}} \otimes V_{q_{1,2}} \otimes \cdots \\ E_{2,q} \otimes V_{q_{2,1}} \otimes V_{q_{2,2}} \otimes \cdots \\ \vdots \\ E_{k,q} \otimes V_{q_{k,1}} \otimes V_{q_{k,2}} \otimes \cdots \end{cases}$$

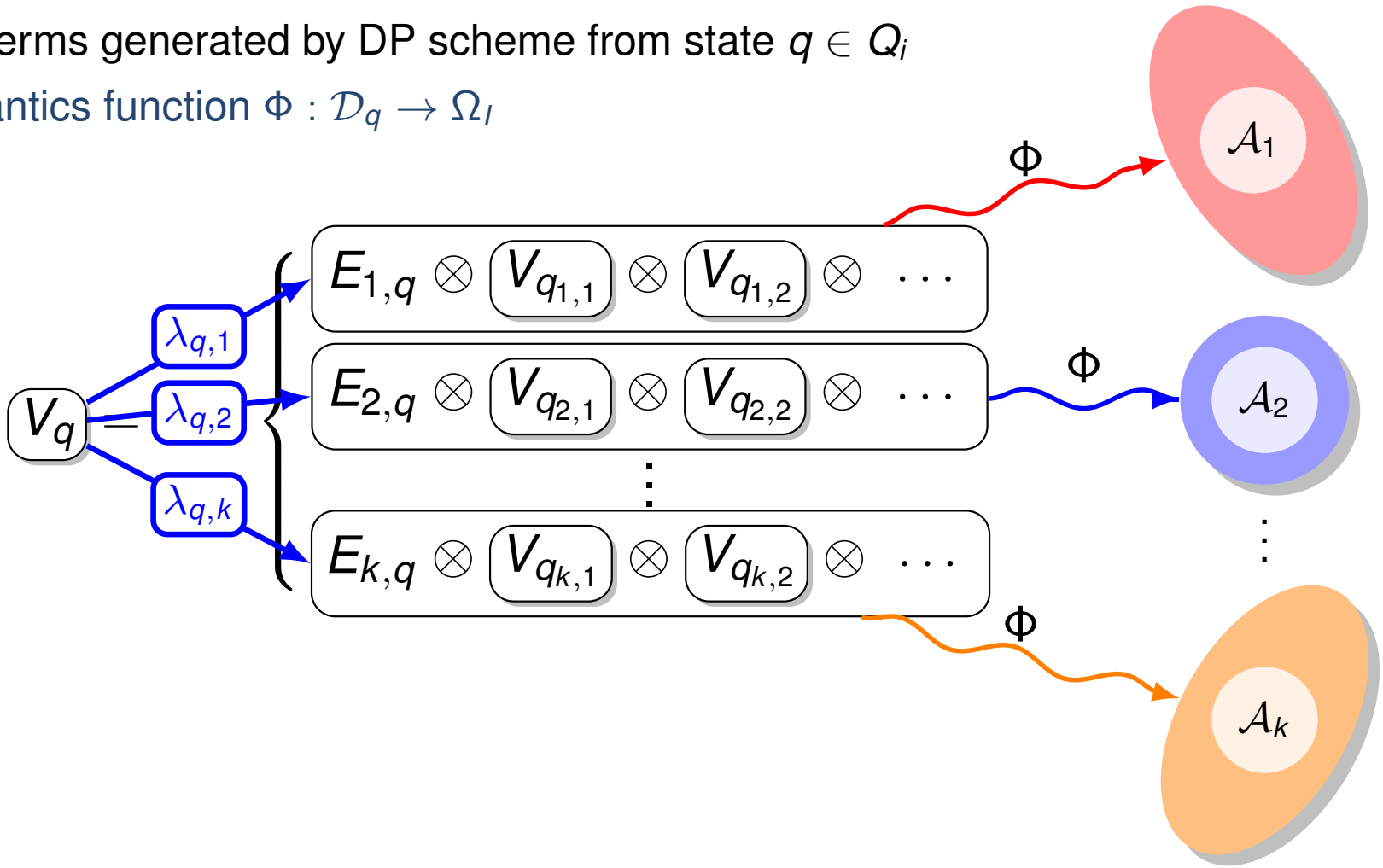
Reasoning structurally about dynamic programming

- I : Instance (a.k.a. problem)
- Q : State space for dyn. prog. scheme (LHS terms, I initial state)
- Ω_q : Search space associated with state q
- $\mathcal{T}_q$: Terms generated by DP scheme from state $q \in Q_i$
- Semantics function $\Phi : \mathcal{D}_q \to \Omega_I$



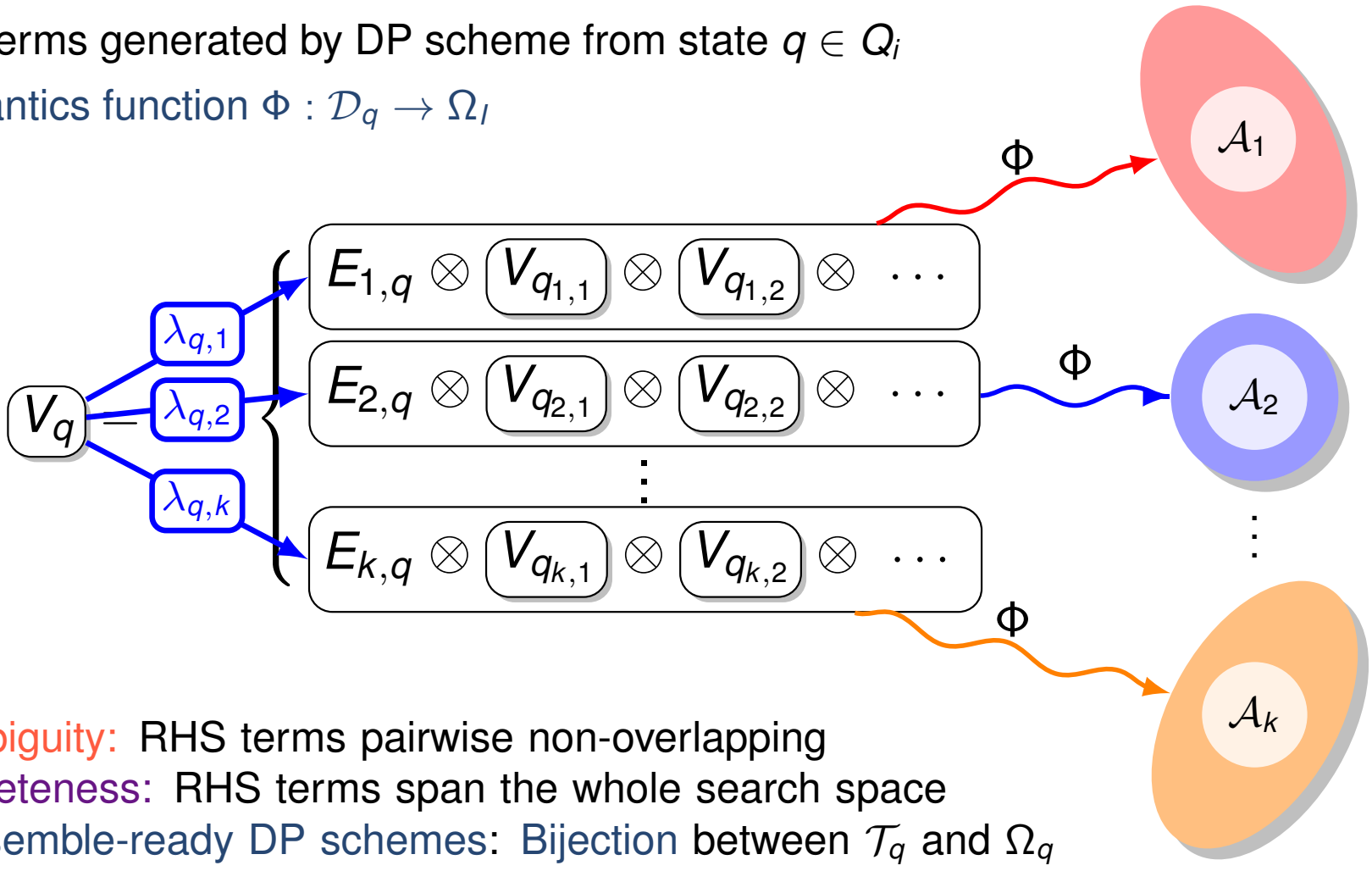
Reasoning structurally about dynamic programming

- I : Instance (a.k.a. problem)
- Q : State space for dyn. prog. scheme (LHS terms, I initial state)
- Ω_q : Search space associated with state q
- $\mathcal{T}_q$: Terms generated by DP scheme from state $q \in Q_i$
- Semantics function $\Phi : \mathcal{D}_q \to \Omega_I$



Reasoning structurally about dynamic programming

- I : Instance (a.k.a. problem)
- Q : State space for dyn. prog. scheme (LHS terms, I initial state)
- Ω_q : Search space associated with state q
- $\mathcal{T}_q$: Terms generated by DP scheme from state $q \in Q_i$
- Semantics function $\Phi : \mathcal{D}_q \to \Omega_I$



Unambiguity: RHS terms pairwise non-overlapping
Completeness: RHS terms span the whole search space
$\to$ Ensemble-ready DP schemes: Bijection between $\mathcal{T}_q$ and Ω_q

A toolbox of ensemble applications

Objective function: $E : \Omega \to \mathbb{R} \cup \{+\infty\}$

Optimization paradigm:

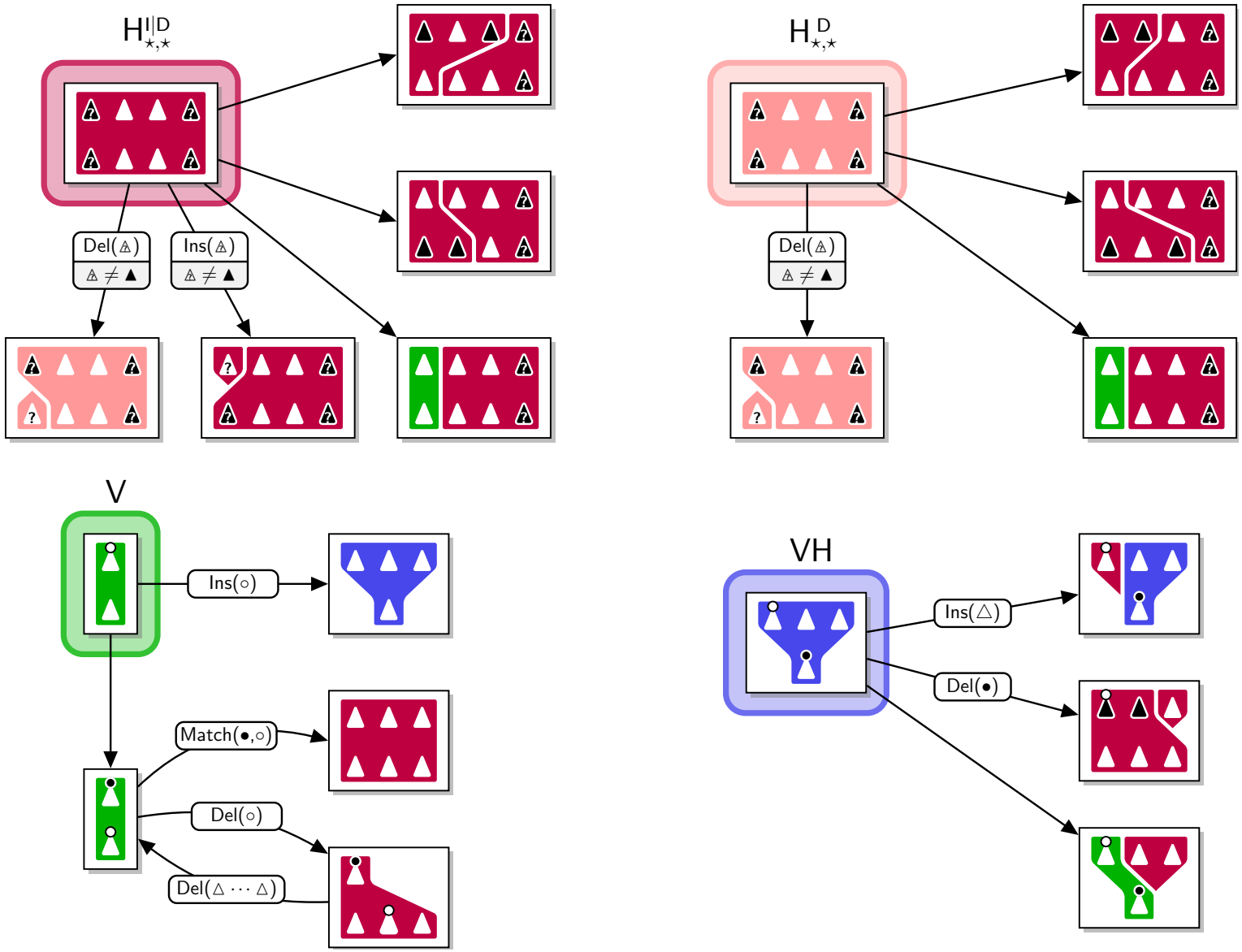
- Optimal solution $S^\star = \operatorname{argmin}_{S \in \Omega} E(S)$ [Bellman, 1954]
- k-best, or Δ-suboptimals [Waterman and Byers, 1985; Wuchty et al., 1999; Huang and Chiang, 2005]
- Parametric optimization [Gusfield et al., 1994; Pachter and Sturmfels, 2004]

Ensemble applications in Boltzmann-Gibbs distribution:

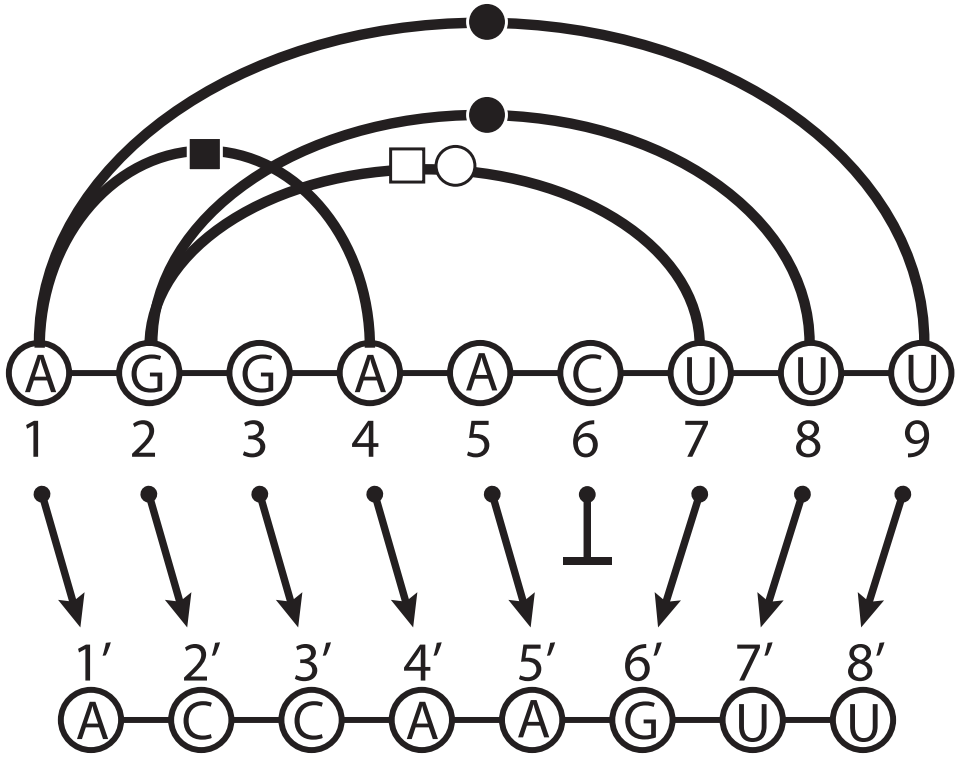
$$\text{Probability of } S \in \Omega \to \mathbb{P}(S) = \frac{e^{-E(S)/kT}}{\mathcal{Z}} \text{ where } \mathcal{Z} := \sum_{S \in \Omega} e^{-E(S)/kT}$$

- Partition function $\mathcal{Z}$ [McCaskill, 1990]
- Transition probabilities $\to$ Inside/outside algorithms [Baker, 1979; McCaskill, 1990]
- Stochastic sampling [Ding and Lawrence, 2003; Ponty, 2008]
- Non-redundant sampling [Lorenz and Ponty, 2013; Michálik, Touzet, and Ponty, 2017]
- Multidim./constrained sampling [Bodini and Ponty, 2010; Waldispühl and Ponty, 2011; Reinharz et al., 2013]
- Moments of additive features [Miklós et al., 2005; Ponty and Saule, 2011]
- Exact distribution of features (DFT) [Senter, Sheikh, Dotu, Ponty, and Clote, 2012]

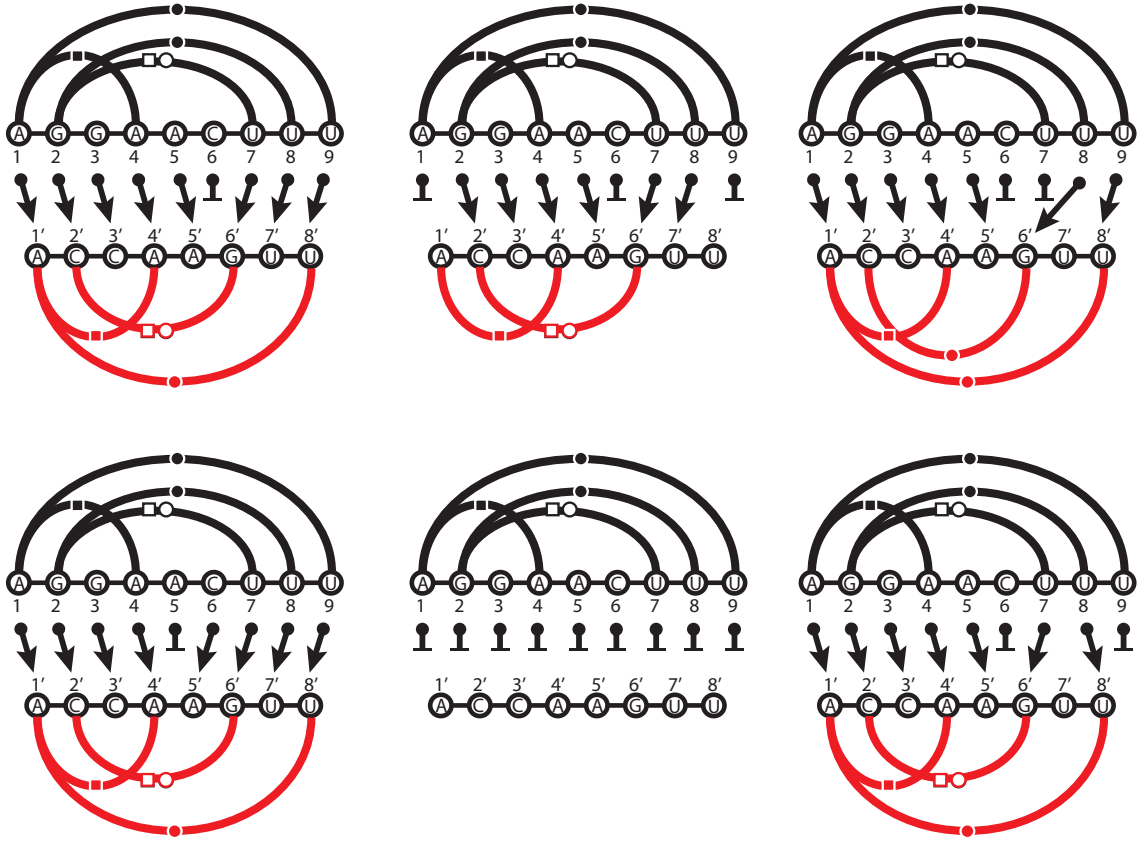
Unambiguous DP scheme for tree alignments



[Chauve, Courtiel, and Ponty, 2018]


A G G A A C U U U
1 2 3 4 5 6 7 8 9
1′ 2′ 3′ 4′ 5′ 6′ 7′ 8′
A C C A A G U U

Sequence-structure alignment



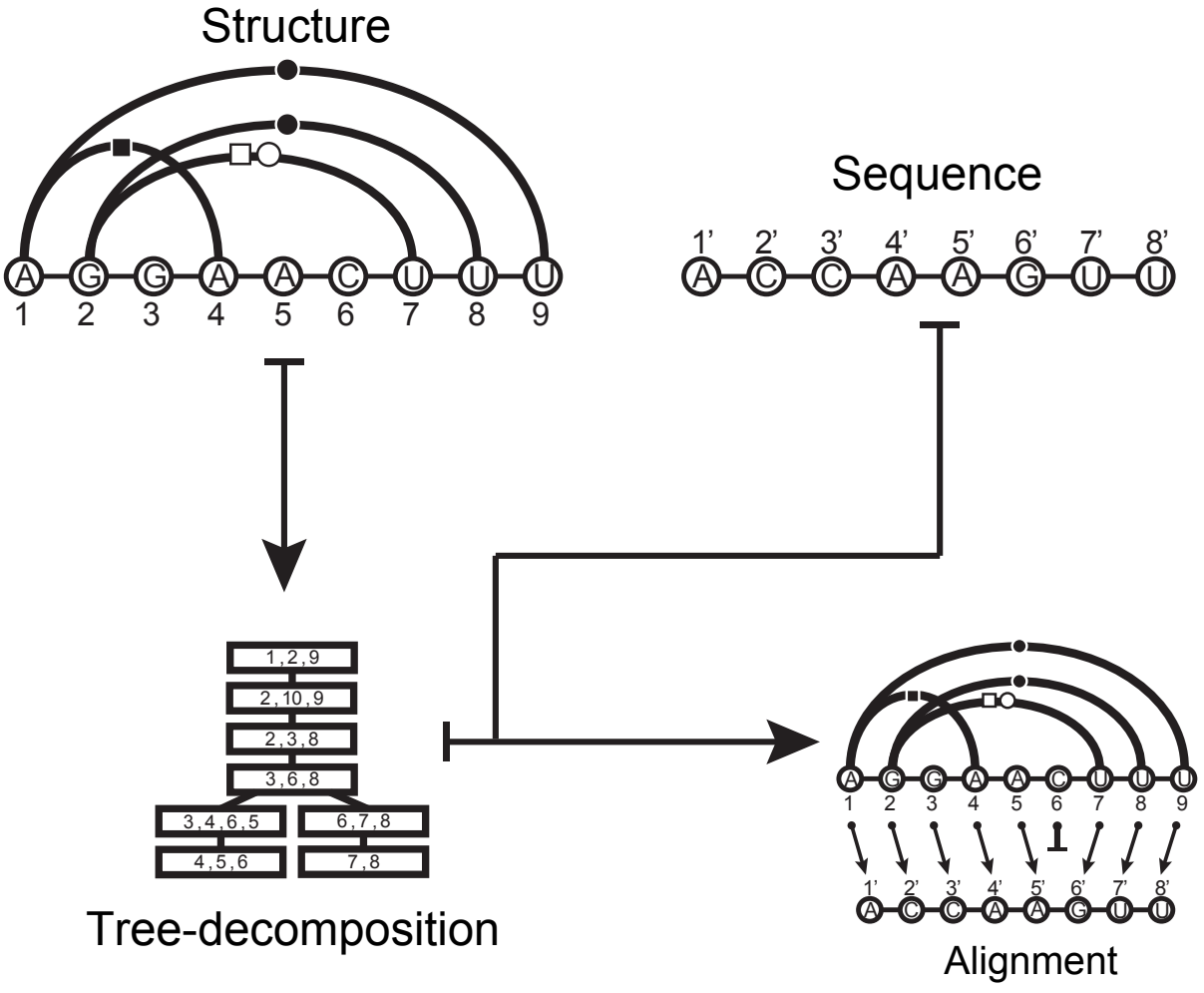
Sequence-structure alignment Problem

Input: (Extended) Secondary structure S ($|S| = n$) + Sequence ω ($|\omega| = m$)
Output: Minimal-cost alignment (mapping subject to constraints)

NP-hard problem + hard to approximate (max-SNP-hard, not PTAS unless P=NP)

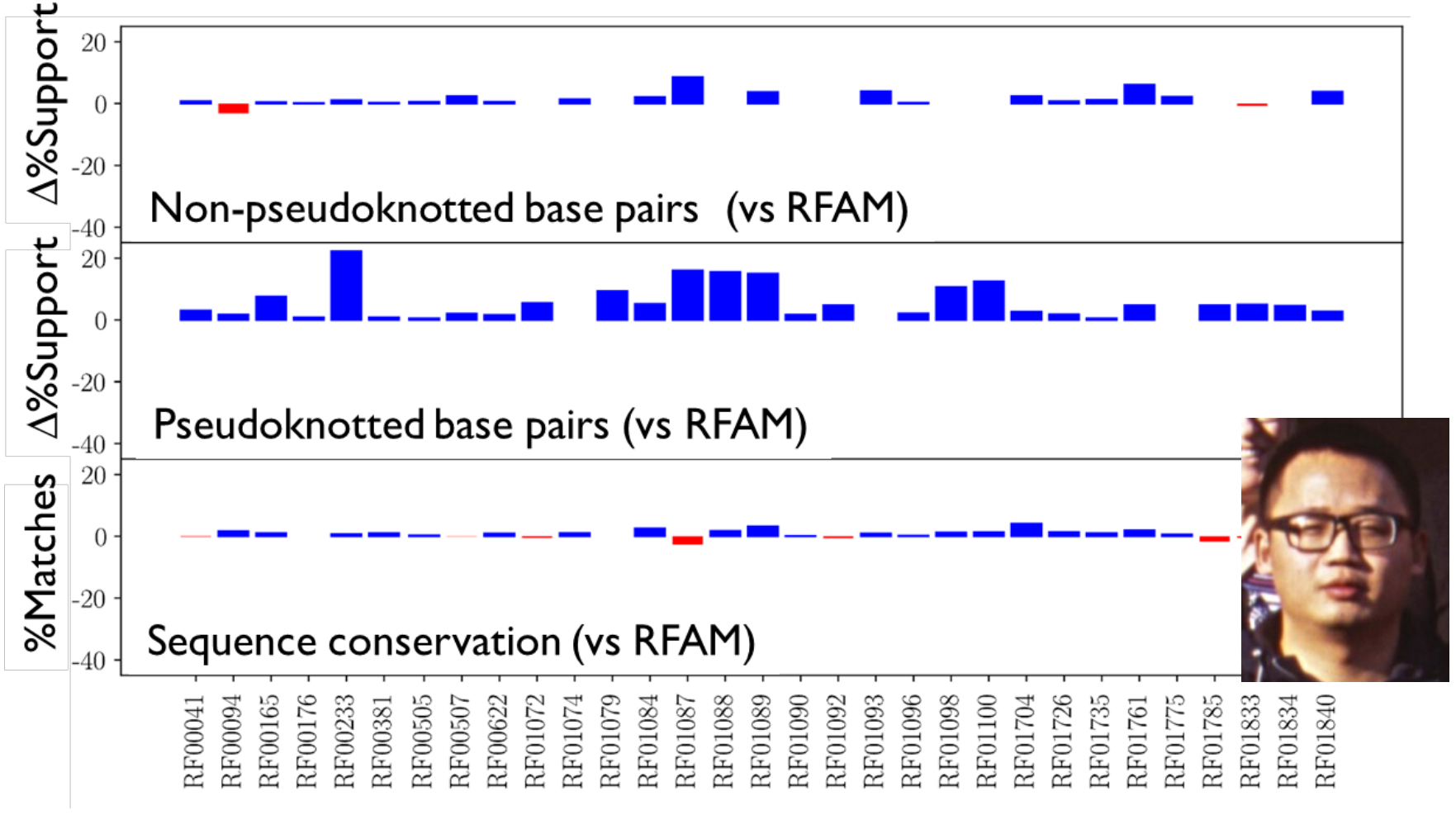
An XPT algorithm for sequence-structure alignment

Our solution: Exact $\Theta(n.m^{t+1})$ DP scheme for RNA structure of tree-width t



[Rinaudo, Ponty, Barth, and Denise, 2012]

Supporting pseudoknotted could lead to improved alignments



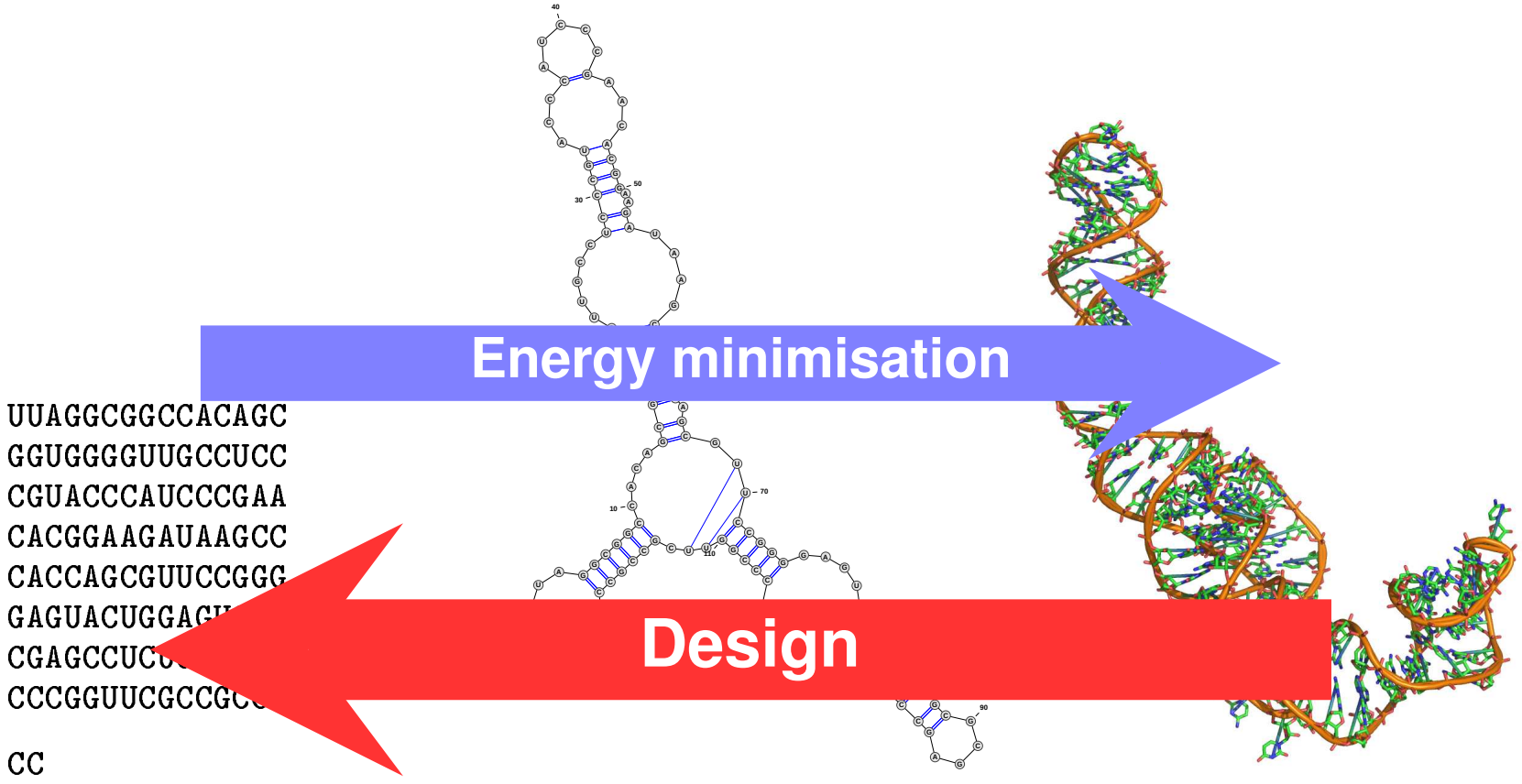
Wei Wang

Part 4. RNA design

Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself. *Francis Crick*

Negative design

RNA = Linear Polymer = Sequence over $\{A, C, G, U\}^\star$



Primary Structure Secondary Structure Tertiary Structure

5s rRNA 5s (PDBID: 1K73:B)

Why we design RNAs

- To fuel RNA-based therapeutics
  Sequence-based (siRNA, synthetic genes), but structure (or lack thereof) matters
- To perform controlled experiments
- To test/push our understanding of how RNA folds
  Misfolding RNAs reveal gaps in our energy models and descriptors for the conformational spaces
- To create building blocks for synthetic systems
  Rationally-designed RNAs increase orthogonality
- To assess the significance of observed phenomenon
  Background models should include established traits of RNA families. . .
  . . . including their adoption of a single structure
- To help search for homologous sequences
  Incomplete covariance models hindered by limited training sets
  Rational design can be used to enrich alignments (data augmentation)

Abstract goals and means of molecular design

But : To achieve a predefined biological function, as abstracted by a model.

Definition (Positive design)

To satisfy constraints induced by a model of function

In practice: To optimize affinity of interaction, to favor thermodynamic stability of a molecule, to respect sequence composition biases. . .

Definition (Negative design)

To avoid unwanted functions

In practice: To avoid off-target interactions, non-functional alternative foldings, kinetic traps. . . (inverse combinatorial problems)

In the context of RNA:

- Positive design: Seq/struct compatibility, composition, +/- motifs, energie(s)
  $\rightarrow$ Random generation, CSP
- Negative design: Target structure $\rightarrow$ Minimum Free-Energy + Boltzmann prob $\nearrow$
  $\rightarrow$ Local search, exp algorithms, black magic (heuristics, $\star$NN, crowdsourcing. . .)

Exact partial results on negative structural design

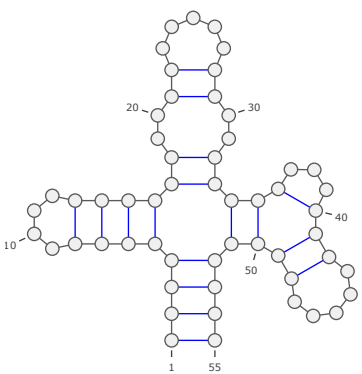
RNA negative structural design, aka inverse folding

Input: Target structure $S^\star$ and energy function E.
Output: RNA sequence w such that

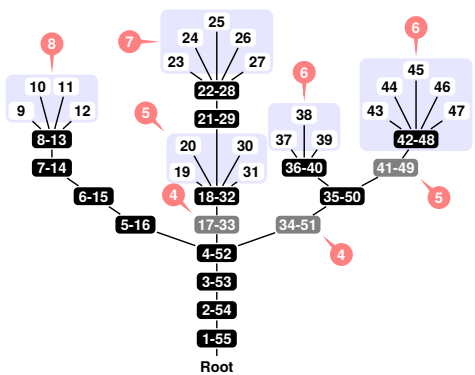
$$\{S \mid E(S; w) = \min_{S'} E(S'; w)\} = \{S^\star\}.$$

In base-pairs based energy models:

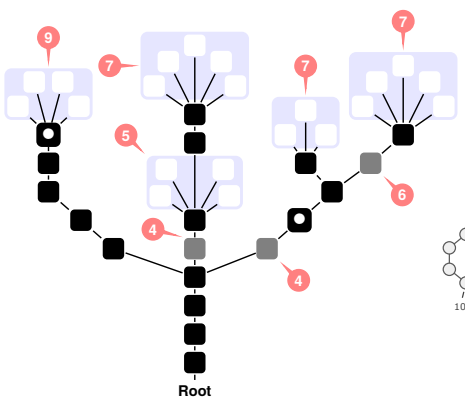
- NP-hard problem [Schnall-Levin et al., 2008; Bonnet et al., 2018]
- Polynomial for: [Hales, Héliou, Manuch, Ponty, and Stacho, 2017]
  - Saturated structures
  - Expansions (stutters) of designable structure (given design)
  - Structures admitting (given) constrained 3-coloring
  - Structures close to any structure avoiding two motifs



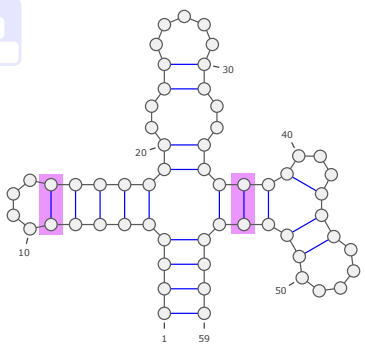
1) Target structure



2) Greedy proper coloring



3) Separated proper coloring



4) Designable structure

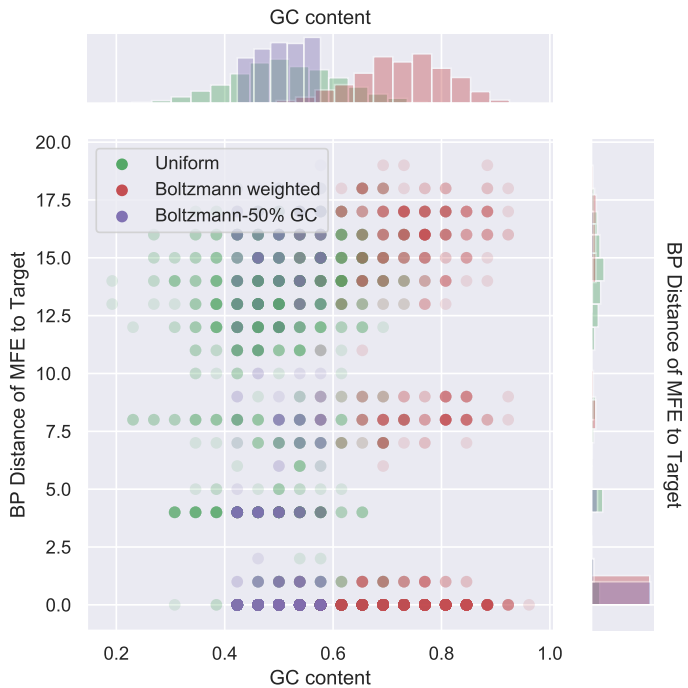
RNA positive structural design

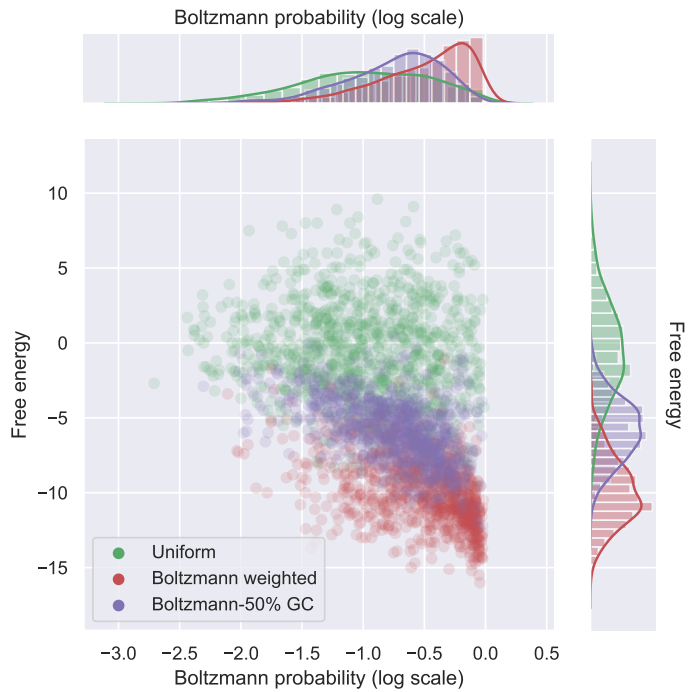
Input: Target structure S, and energy function E + Constraints
Output: RNA sequence w under constraints either minimizing E,
or generated with probability $\propto e^{-\beta E(S,w)}$

IncaRNAtion

[Reinharz, Ponty, and Waldispühl, 2013]

- Energy-weighted sampling (dual partition function)
- Rejection (multidim. Boltzmann) for GC% control

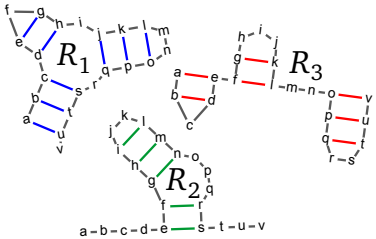




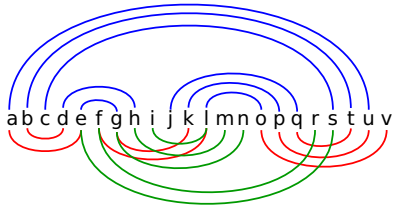
Considering multiple structures

Multiple targets: #P-hard counting/generation, but FPT

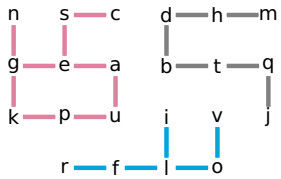
[Hammer, Wang, Will, and Ponty, 2019]



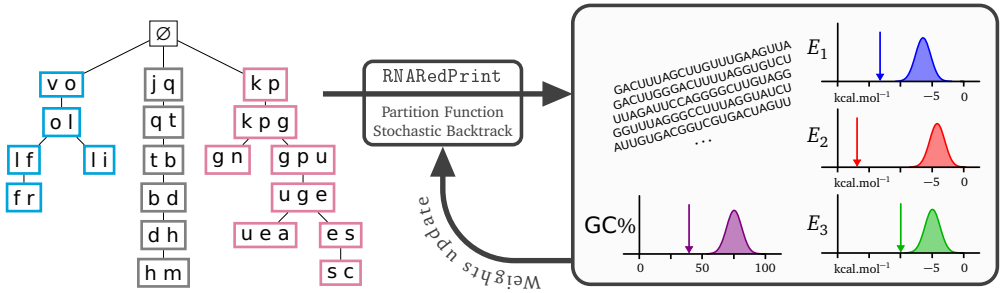
i) Input Structures



ii) Merged Base-Pairs

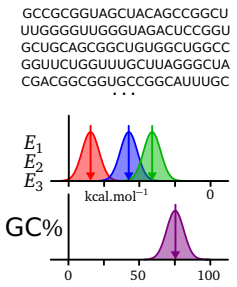


iii) Compatibility Graph



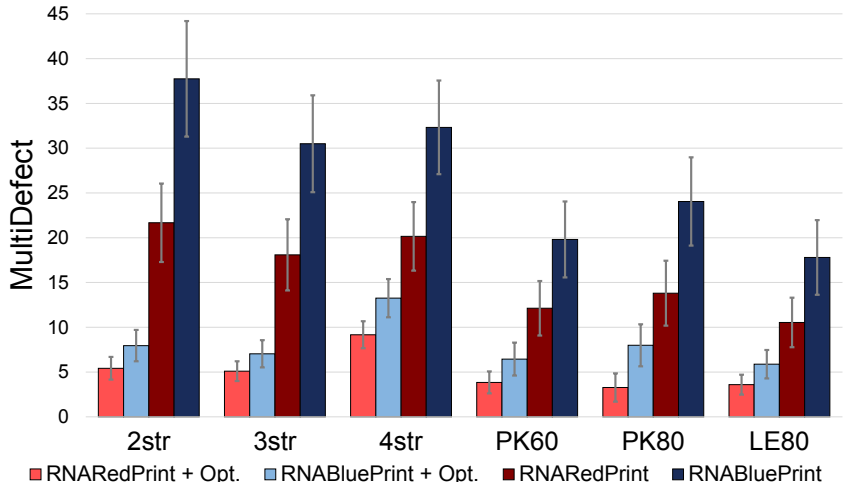
iv) Tree Decomposition

v) Weight Optimization (Adaptive Sampling)



vi) Final Designs

Energies controlled through multidimensionnal Boltzmann sampling



Conclusion

Oh, the places you'll go! *Dr Seuss*

Conclusions

- Enumerative contributions
- Algorithmic ensemble techniques
- Leading to Bioinformatics methods

What's next?

- **Predictive RNA structure from probing data**
  FRM → ANR/FWF PaRNAssus project
  with B. Sargueil (Paris Descartes) and R. Lorenz (TBI Vienna)
- **RNA design/evolution**
  ANR Decrypted project on Direct Coupling analysis
  with B. Sargueil (Paris Descartes) and S. Cocco (ENS Paris)

Afaf Saaidi

Powered by AMIBio founded at LIX by M. Régnier

Supported by

cnrs — Agence Nationale de la Recherche ANR — FRM Fondation Recherche Médicale — Institut Polytechnique de Paris — université Paris-Saclay — Inria Inventeurs du monde numérique

Merci – Thank you

Collaborators:

Ecole Polytechnique
- P. Chassignet, J.-M. Steyaert
- M. Régnier, S. Will

Université Paris-Saclay
- A. Denise, L. Paulevé, D. Barth
- D. Gautheret, F. Leclerc

Université Paris-Descarte
- N. Sargueil, N. Chamont

INRAE Jouy
- D. Rocha

Université Montpelier
- C. Scornavacca

Université Lorraine
- I. Chauvot de Beauchêne

Univ. Lille
- H. Touzet

McGill University
- J. Waldispühl, M. Blanchette

Simon Fraser University
- J. Hales, J. Manuch, L. Stacho
- C. Chauve, M. Mishna

Université du Québec à Montréal
- V. Reinharz

University of Vienna
- S. Will, S. Hammer
- A. Tanzer, R. Lorenz, I. Hofacker

Boston College
- P. Clote

Ben Gurion University
- D. Barash, A. Churkin

Wuhan University
- Y. Zhou

References I

J. K. Baker. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132, 1979. doi: 10.1121/1.2017061. URL https://doi.org/10.1121/1.2017061.

Elena Barcucci, Renzo Pinzani, and Renzo Sprugnoli. The random generation of directed animals. *Theoretical Computer Science*, 127(2):333–350, 1994.

Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

Olivier Bodini and Yann Ponty. Multi-dimensional Boltzmann Sampling of Languages. In Michael Drmota and Bernhard Gittenberger, editors, *AOFA - 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms - 2010*, volume AM of *DMTCS Proceedings*, pages 49–64, Vienna, Austria, June 2010. Discrete Mathematics and Theoretical Computer Science. URL https://hal.inria.fr/hal-00450763.

Édouard Bonnet, Paweł Rzążewski, and Florian Sikora. Designing RNA secondary structures is hard. In Benjamin J. Raphael, editor, *Research in Computational Molecular Biology - 22nd Annual International Conference, RECOMB 2018*, volume 10812 of *Lecture Notes in Computer Science*, pages 248–250, Paris, 2018. Springer.

Mireille Bousquet-Mélou and Yann Ponty. Culminating paths. *Discrete Mathematics and Theoretical Computer Science*, 10(2):125–152, 2008. URL https://hal.archives-ouvertes.fr/hal-00151979.

Cedric Chauve, Julien Courtiel, and Yann Ponty. Counting, generating, analyzing and sampling tree alignments. *International Journal of Foundations of Computer Science*, 29(5):741–767, 2018. URL https://hal.inria.fr/hal-01500116.

Peter Clote, Yann Ponty, and Jean-Marc Steyaert. Expected distance between terminal nucleotides of rna secondary structures. *Journal of Mathematical Biology*, 65(3):581–599, 2012.

Alain Denise, Yann Ponty, and Michel Termier. Controlled non uniform random generation of decomposable structures. *Theoretical Computer Science*, 411 (40-42):3527–3552, 2010. doi: 10.1016/j.tcs.2010.05.010. URL https://hal.archives-ouvertes.fr/hal-00483581.

Y. Ding and E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

Jérémie Du Boisberranger, Danièle Gardy, and Yann Ponty. The weighted words collector. In Nicolas, Broutin (INRIA, France), Luc, Devroye (McGill, and Cana, editors, *AOFA - 23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms - 2012*, volume AQ, pages 243–264, Montreal, Canada, June 2012. DMTCS. URL https://hal.inria.fr/hal-00666399.

Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, New York, NY, USA, 1 edition, 2009. ISBN 0521898064, 9780521898065.

D. Gusfield, K. Balasubramanian, and D. Naor. Parametric optimization of sequence alignment. *Algorithmica*, 12(4-5):312–326, nov 1994. doi: 10.1007/bf01185430.

Jozef Hales, Alice Héliou, Ján Manuch, Yann Ponty, and Ladislav Stacho. Combinatorial RNA Design: Designability and Structure-Approximating Algorithm in Watson-Crick and Nussinov-Jacobson Energy Models. *Algorithmica*, 79(3):835–856, 2017. doi: 10.1007/s00453-016-0196-x.

Stefan Hammer, Wei Wang, Sebastian Will, and Yann Ponty. Fixed-parameter tractable sampling for RNA design with multiple target structures. *BMC Bioinformatics*, 20(1):209, December 2019. doi: 10.1186/s12859-019-2784-7. URL https://hal.inria.fr/hal-02112888.

References II

Liang Huang and David Chiang. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 53–64. Association for Computational Linguistics, 2005.

Andy Lorenz and Yann Ponty. Non-redundant random generation algorithms for weighted context-free languages. *Theoretical Computer Science*, 502: 177–194, September 2013. doi: 10.1016/j.tcs.2013.01.006. URL https://hal.inria.fr/inria-00607745.

J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

Juraj Michálik, Hélène Touzet, and Yann Ponty. Efficient approximations of RNA kinetics landscape using non-redundant sampling. In *ISMB/ECCB - 25th Annual international conference on Intelligent Systems for Molecular Biology/16th European Conference on Computational Biology - 2017*, volume 33, pages i283 – i292, Prague, Czech Republic, 2017. doi: 10.1093/bioinformatics/btx269. URL https://hal.inria.fr/hal-01500115.

István Miklós, Irmtraud M. Meyer, and Borbála Nagy. Moments of the boltzmann distribution for RNA secondary structures. *Bull Math Biol*, 67(5):1031–1047, Sep 2005. doi: 10.1016/j.bulm.2004.12.003. URL http://dx.doi.org/10.1016/j.bulm.2004.12.003.

L. Pachter and B. Sturmfels. Parametric inference for biological sequence analysis. *Proceedings of the National Academy of Sciences*, 101(46): 16138–16143, nov 2004. doi: 10.1073/pnas.0406011101.

Yann Ponty. Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method. *Journal of Mathematical Biology*, 56(1-2):107–127, 2008. doi: 10.1007/s00285-007-0137-z. URL https://hal.inria.fr/inria-00548863.

Yann Ponty and Cédric Saule. A Combinatorial Framework for Designing (Pseudoknotted) RNA Algorithms. In *WABI - 11th Workshop on Algorithms in Bioinformatics - 2011*, Saarbrucken, Germany, 2011. URL https://hal.inria.fr/inria-00601060.

Vladimir Reinharz, Yann Ponty, and Jérôme Waldispühl. A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. *Bioinformatics*, 29(13):i308–15, July 2013. doi: 10.1093/bioinformatics/btt217. URL https://hal.inria.fr/hal-00840260. Extended version of ISMB/ECCB'13.

Philippe Rinaudo, Yann Ponty, Dominique Barth, and Alain Denise. Tree decomposition and parameterized algorithms for RNA structure-sequence alignment including tertiary interactions and pseudoknots. In Ben Raphael and Jijun Tang, editors, *WABI - 12th Workshop on Algorithms in Bioinformatics - 2012*, tba, Ljubljana, Slovenia, September 2012. University of Ljubljana. URL https://hal.inria.fr/hal-00708580.

Christelle Rovetta, Juraj Michálik, Ronny Lorenz, Andrea Tanzer, and Yann Ponty. Non-Redundant Sampling and Statistical Estimators for RNA Structural Properties at the Thermodynamic Equilibrium. working paper or preprint, September 2020. URL https://hal.inria.fr/hal-02288811.

Michael Schnall-Levin, Leonid Chindelevitch, and Bonnie Berger. Inverting the Viterbi algorithm: an abstract framework for structure design. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 904–911, 2008. doi: 10.1145/1390156.1390270.

Evan Senter, Saad Sheikh, Ivan Dotu, Yann Ponty, and Peter Clote. Using the fast fourier transform to accelerate the computational search for rna conformational switches. *PLoS One*, 7(12):e50506, 2012. doi: 10.1371/journal.pone.0050506. URL http://dx.doi.org/10.1371/journal.pone.0050506.

Defne Surujon, Yann Ponty, and Peter Clote. Small-world networks and RNA secondary structures. *Journal of computational biology : a journal of computational molecular cell biology*, 26(1):16–26, January 2019. doi: 10.1089/cmb.2018.0125. URL https://hal.inria.fr/hal-01424452.

References III

M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in Medecine and Biology*, volume 57 of *Lecture Notes in Biomathematics*, pages 360–365, 1985.

Jérôme Waldispühl and Yann Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Journal of Computational Biology*, 18(11):1465–79, November 2011. doi: 10.1089/cmb.2011.0181. URL https://hal.inria.fr/hal-00681928.

Michael Waterman. Secondary structure of single-stranded nucleic acids. *Advances in Mathematics: Supplementary Studies*, 1:167–212, 1978.

Michael S. Waterman and Thomas H. Byers. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Mathematical Biosciences*, 77(1):179 – 188, 1985. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(85)90096-3. URL http://www.sciencedirect.com/science/article/pii/0025556485900963.

S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49: 145–164, 1999.

Hua-Ting Yao, Cedric Chauve, Mireille Regnier, and Yann Ponty. Exponentially few RNA structures are designable. In *ACM-BCB - 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - 2019*, pages 289–298, Niagara-Falls, United States, 2019. ACM Press. doi: 10.1145/3307339.3342163. URL https://hal.inria.fr/hal-02141853.

Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bulletin of mathematical biology*, 46(4):591–621, 1984.