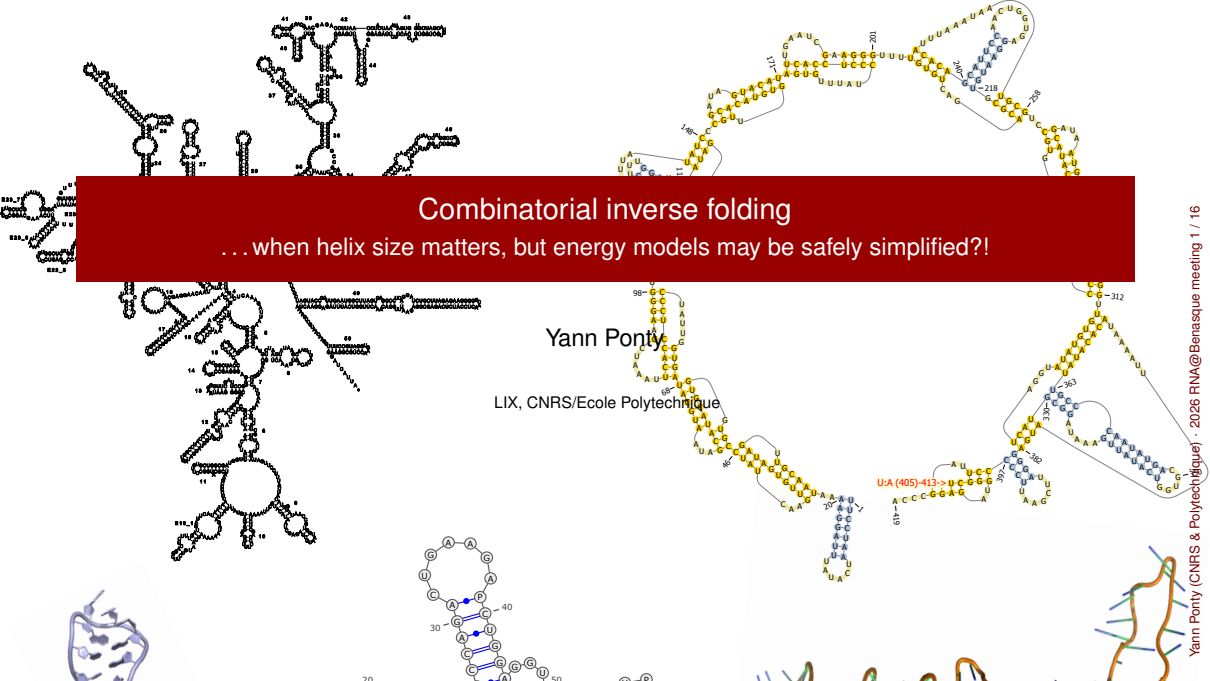


Combinatorial inverse folding

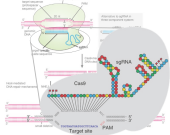
. . . when helix size matters, but energy models may be safely simplified?!

Yann Ponty

LIX, CNRS/Ecole Polytechnique

RiboNucleic Acids (RNA) in Human biology/health: Friend **and** Foe!

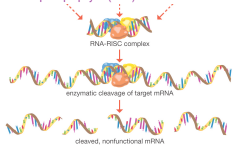
Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



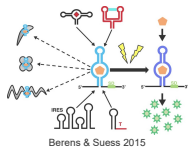
Hendel *et al*, 2015; Agrotis & Ketteler, 2015

Rational design

Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



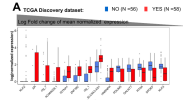
Encyclopaedia Brittannica, Inc 2013



Berens & Suess 2015

Sensor of metabolites
Riboswitches

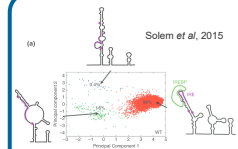
Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...



[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)

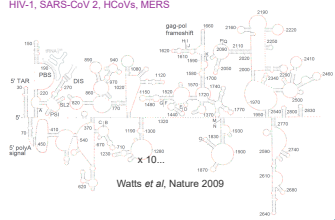
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)



Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy,
Cystic fibrosis, Rett syndrome...

(2D) Structure Modeling

Genomic material for Human pathogens
HIV-1, SARS-CoV 2, HCoVs, MERS



Watts *et al*, Nature 2009

RNA Inverse Folding

Definition (INVERSE-FOLDING(E) problem)

Input: Secondary structure R + Energy distance $\Delta > 0$

Output: RNA sequence $\omega \in \Sigma^\star$ such that:

$$\forall R' \neq R \text{ compatible with } \omega : \; E(\omega, R') \geq E(\omega, R) + \Delta$$

or $\varnothing$ if no such sequence exists.

Difficult problem:

- Introduced in the 90s by Vienna/Leipzig wunderkids [Hofacker *et al* Monatshefte für Chemie 1994]
- NP-hardness in maximalist setting (Energy model as input [Schnall-Levin *et al*, ICML'08])
- NP-hardness for BP maximization with partial assignment [Bonnet *et al*, RECOMB'18]
- Hard to study: Non local, no theoretical framework, too many parameters...

Foreword and caveats

- *"All models are wrong, but some are useful"* George Box

  Reductive models enable analysis. . . . . while hopefully retaining relevance

- In ViennaRNA we trust: RNAFold, not Nature (yet), scores

  Ideally, uniform sampling from the preimage

- Is inverse RNA design actually hard? Worthy of further attention?

Molecular biology, Structural biology Comparative genomics, Computer Science. . .	→	YES
DNA computing	→	Meh...

Foreword and caveats

- *"All models are wrong, but some are useful"* George Box

  Reductive models enable analysis. . . . . while hopefully retaining relevance

- In ViennaRNA we trust: RNAFold, not Nature (yet), scores

  Ideally, uniform sampling from the preimage

- Is inverse RNA design actually hard? Worthy of further attention?

  Molecular biology, Structural biology
  Comparative genomics, Computer Science. . . → YES
  DNA computing → Meh...

Foreword and caveats

- *"All models are wrong, but some are useful"* George Box

  Reductive models enable analysis. . . . . while hopefully retaining relevance

- In ViennaRNA we trust: RNAFold, not Nature (yet), scores

  Ideally, uniform sampling from the preimage

- Is inverse RNA design actually hard? Worthy of further attention?

  Molecular biology, Structural biology
  Comparative genomics, Computer Science. . . $\rightarrow$ YES
  DNA computing $\rightarrow$ Meh...

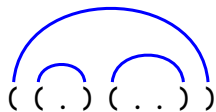
Inverse Folding in Base Pair maximization (maxBP) model

Definition (Inverse folding in maxBP)

Input: Target secondary structure R, *i.e.* set of pairwise non-crossing BPs

Output: RNA sequence $\omega \in \Sigma^\star$ such that:

- Target R compatible with ω;
- $\forall R' \neq R$, compatible with ω: $|R'| < |R|$.



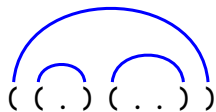
Inverse Folding in Base Pair maximization (maxBP) model

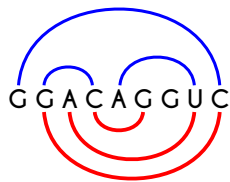
Definition (Inverse folding in maxBP)

Input: Target secondary structure R, *i.e.* set of pairwise non-crossing BPs

Output: RNA sequence $\omega \in \Sigma^\star$ such that:

- Target R compatible with ω;
- $\forall R' \neq R$, compatible with ω: $|R'| < |R|$.





Designability in simple BP-based energy models

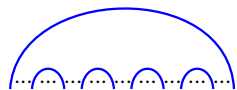
Partial characterization of designable structures [Hales *et al*, CPM'15 & Algorithmica'17]

- Saturated structures (all positions paired): Designable $\Leftrightarrow$ Multiloops degrees ≤ 4 (+ $\Theta(n)$ algo.)
- Designable $\Rightarrow$ Avoid multiloops with *degree* 5^+ (m_5), or *degree* 3^+ *with* 1^+ *unpaired* ($m_{3\circ}$).

Designability in simple BP-based energy models

Partial characterization of designable structures [Hales *et al*, CPM'15 & Algorithmica'17]

- Saturated structures (all positions paired): Designable $\Leftrightarrow$ Multiloops degrees ≤ 4 $(+\ \Theta(n)$ algo.)
- Designable $\Rightarrow$ Avoid multiloops with *degree* 5^+ (m_5), or *degree* 3^+ *with* 1^+ *unpaired* ($m_{3\circ}$).



m_5

$m_{3\circ}$

Remark: Non-designable motifs (obstructions) exist in all energy model/design criterion

Corollary [Yao *et al*, ACM-BCB'19&J Math Biol 2026]]

The fraction of designable structures is exponentially decreasing with length n

. . . but (fortunately) $\exists$ infinite family of unsaturated designable 2D structures [Jedwab *et al*, TCS 2020]

Separated coloring of RNA tree representations

Proper coloring of base pairs/nodes:

- Each base pair $\rightarrow$ one out of 3 colors: ● $\rightarrow$ G · C; ○ $\rightarrow$ C · G; ◍ $\rightarrow$ A · U or U · A.
- Parent node colored ● (resp. ○) cannot have ○ (resp. ●) children;
- Children satisfy $\#● \leq 1$, $\#○ \leq 1$, $\#◍ \leq 2$ and $\#● + \#○ < 2$

Definitions:

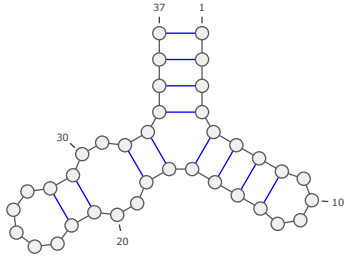
- Level of a base pair = $\#● - \#○$ on path to root
- Coloring separated if proper and ◍ base pairs and unpaired positions at different levels

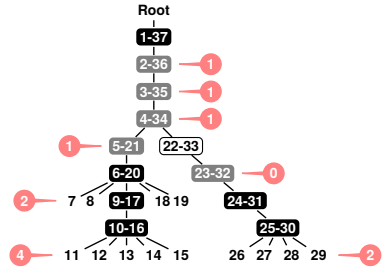
Theorem ([Hales *et al.* Algorithmica 2017])

$\exists$ Separated (proper) coloring for structure $\Rightarrow$ Structure is designable in $\Theta(n)$ time

Separated Coloring

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





GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC

Intuition: All unpaired → A, so alternative structures need to pair every G, C, and A:

- If ⬤ base pairs conform to target, then ⬤/◯ behave as saturated structure;
- If some ⬤ misbehaves, then some ⬤ must interact with leaf, delimiting regions with $\#G \neq \#C$, so we lose at least one G − C pair.

Theorem ([Boury *et al.* Alg Mol Biol 2025])

However, deciding if a structure is separable is NP-hard

Modulo m-separated coloring

Reminder: Separated coloring = induced levels of ⬤ and leaves do not overlap.

Idea: Prescribe allowed values for the levels mod m of ⬤ (set ξ) and leaves (set $\bar{\xi} := [0, m-1] \setminus \xi$).

Definition: Given m and $\xi \subseteq [0, m-1]$, a $(\xi, \bar{\xi})$-separated coloring χ is such that, for each node v:

$$\blacktriangleright \chi(v) = \bigcirc \rightarrow \text{Level}_\chi(v) \bmod m \in \xi \quad \text{and} \quad \blacktriangleright v \text{ is leaf } \rightarrow \text{Level}_\chi(v) \bmod m \in \bar{\xi}.$$

Proposition (Boury *et al*, Alg Mol Biol 2026)

Given m and $(\xi, \bar{\xi})$, a $(\xi, \bar{\xi})$-separated coloring can be found in $\Theta(m.n)$-time using dyn. prog.

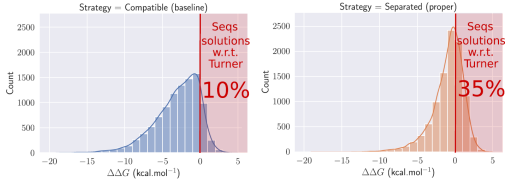
$\rightarrow$ (modulo) m-separated colorings can be found in $\Theta(2^m.n)$ time/$\Theta(m.n)$ memory (FPT)

Also, any target structure with 3^+ BP/helix (+ many with helix length 2) is (modulo) 2-separated

Theorem (Boury *et al*, Alg Mol Biol 2026)

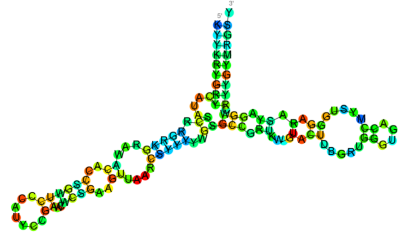
For targets w/o isolated stack/BP, maxBP inverse folding solvable in $\Theta(n)$ time

The good, the bad and the ugly



- Separated designs: promising first step towards inverse folding in Turner model
- Amenable to uniform random generation
  $\rightarrow$ Cardinality estimates for neutral network
- Good seeding strategy for local search
  [Boury, $\cdots$, Yao, RECOMB'25]

- maxBP considers unrealistic competing structures (*e.g.* isolated BPs)...
- ...and overlooks real Turner competitors
- Simple BP model artificially restricts loops
  $\rightarrow$ Crucial RNA structures are maxBP undesignable



Inverse folding in the stacks maximization (maxStacks) model [Boury et al, WABI 2026]

Energy: $\#Stacks(R) = \sum_{\text{Helix } H \in R} |H| - 1$

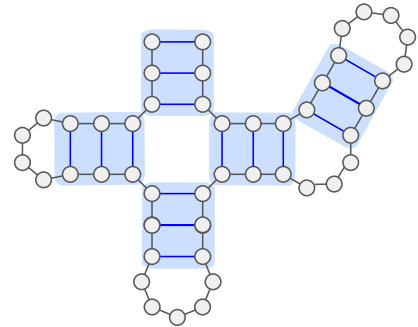
Definition (maxStacks inverse folding)

Input: Target secondary structure R

Output: RNA sequence $\omega \in \Sigma^\star$ such that:

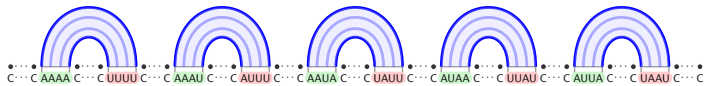
- Target R comp. with ω;
- $\forall R' \neq R$, compatible with ω:

$$\#Stacks(R') < \#Stacks(R).$$



The maxStacks model differs from maxBP:

- Multiloops of arbitrary degree Δ can be designed



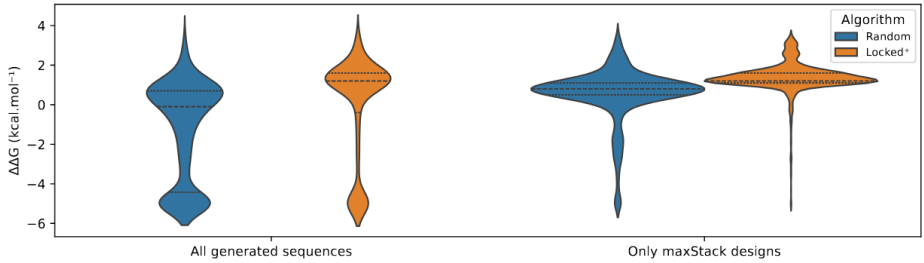
- Result: Exact linear algo (Locked$^+$) for targets with helices greater than $\lceil \log_{3.56}(\Delta) \rceil + 6.2$ BPs

Stacks-maximizing designs are frequently designs for the Turner energy model

Definition: $\Delta\Delta G(\text{Candidate design } w, \text{Target } R) = \max_{R' \neq R} E(w; R') - E(w; R)$

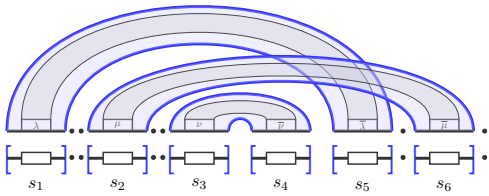
$\Delta\Delta G \nearrow \rightarrow$ Greater stability relative to competitors
$\Delta\Delta G \leq 0 \rightarrow$ Dominated/challenged by competitor structure
$\Delta\Delta G > 0 \rightarrow$ Solution to inverse folding

Random target structures: Length from 25 to 275nts, min helix length from 2 to 8 BPs



- Larger proportion of (better) Turner designs using Locked$^+$ (maxStack algorithm)
- Stacking maximization usable as filter to improve potential as design

Inverse folding including crossing BPs/PseudoKnots follows for (almost) free



General complexity open, yet probably NP-Hard

Definition (maxStacks PK inv. folding)

Input: Target crossing structure R

Output: RNA sequence $\omega \in \Sigma^\star$ such that:

- Target R comp. with ω;
- $\forall$ (crossing) $R' \neq R$, compatible with ω:

$$\#Stacks(R') < \#Stacks(R).$$

Remark: General pseudoknots allowed both for input and alternative structures

Theorem (Boury et al, WABI 2026)

Linear-time exact algorithm for target structures with H (crossing) helices with more than $\lceil \log_{3.56}(H) \rceil + 6.2$ base pairs

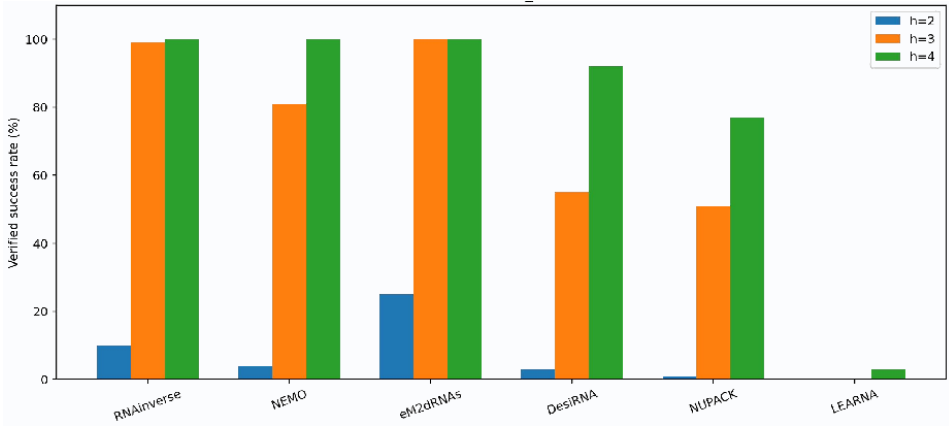
Surprisingly, as checking if given RNA sequence represents solution is NP-hard [Lyngsø, ICALP'04]

Conclusions

- Exact minimal solutions to inverse folding improve empirical design performances
- Longer helices really help structural design for deeper combinatorial reasons
- Interaction design with guaranteed specificity can be addressed in this framework
- Is RNA inverse folding still hard? Is method development still relevant?

Inverse folding remains hard anytime targets feature short helices

Random uniform target structure : Length 75 nts, min helix length 2 to 4



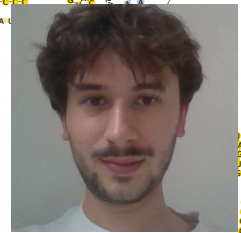
- Modest overall performances, especially in comparison to reported success rates
- Generalization issues of ML approaches?
- Old faithful – 3 decades old – RNAinverse still SOTA???

Thank you

Dr. Théo Bourry
(Open for postdoc)

Laurent Bulteau
CNRS@LIGM

Hua-Ting Yao
TBI Vienna



Samuel Gardelle
LIX@Polytechnique

- Jozef Hanč, Ján Mauch, and Ladislav Stacho (SFU, Canada)
- Ivo Hofacker and Leonhard Sidl (TBI Vienna)
- Sebastian Will and Sarah J. Berkemer (Ecole Polytechnique)
- Elena and Eric for (yet another!) awesome Benasque