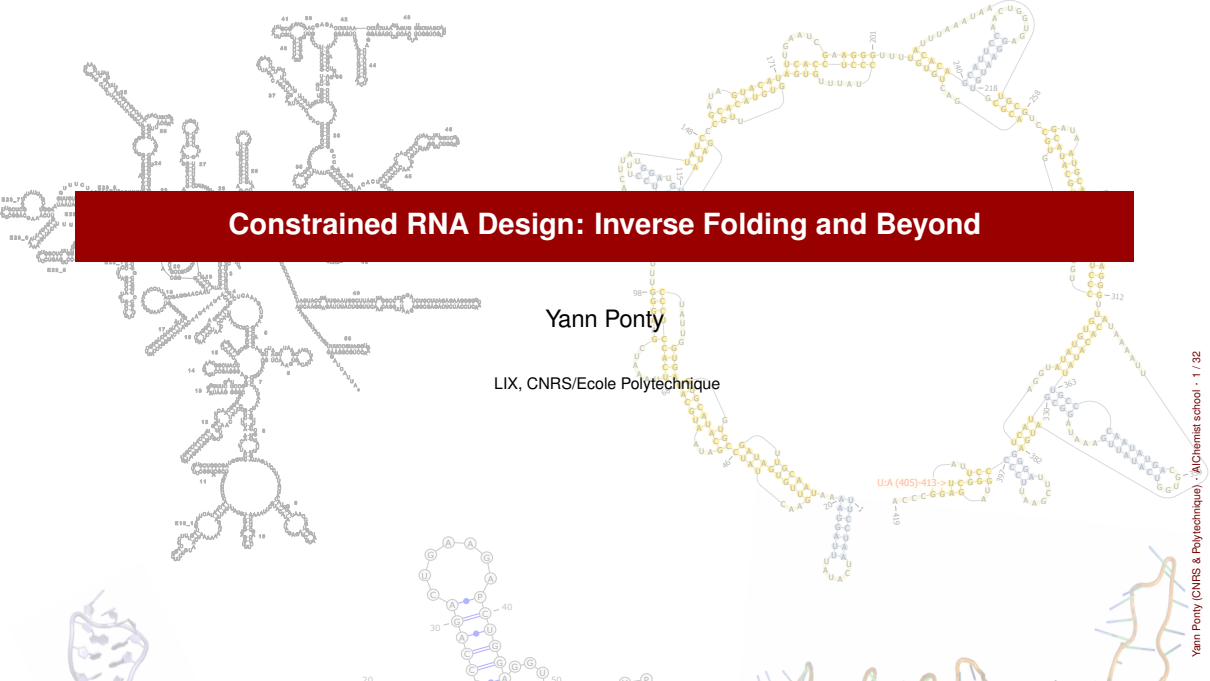
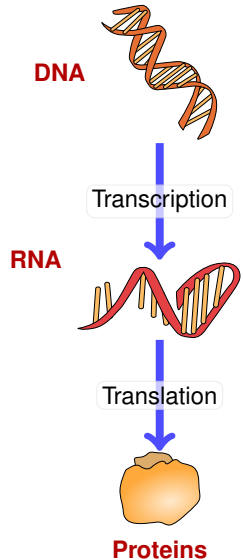
Constrained RNA Design: Inverse Folding and Beyond

Yann Ponty

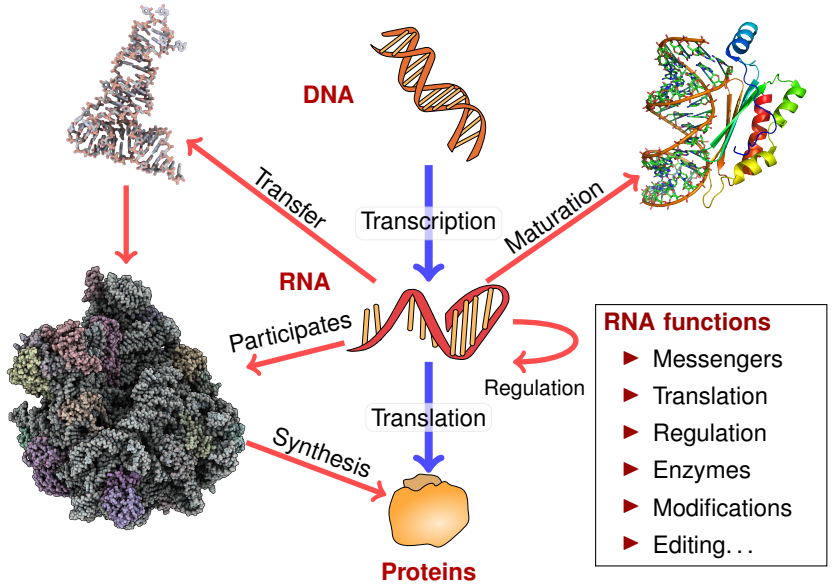
LIX, CNRS/Ecole Polytechnique



Fundamental dogma of molecular biology



Fundamental dogma of molecular biology (v2.0)



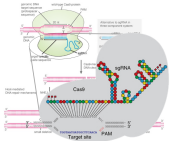
RiboNucleic Acids (RNA) in Human biology/health: Friend **and** Foe!

RiboNucleic Acids (RNAs)

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

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...



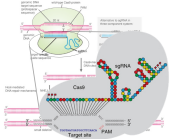
Hendel et al, 2015; Agrotis & Ketteler, 2015

RiboNucleic Acids (RNAs)

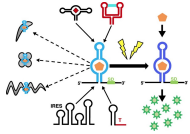
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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



Berens & Suess 2015

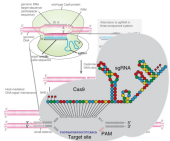
Sensor of metabolites
Riboswitches

RiboNucleic Acids (RNAs)

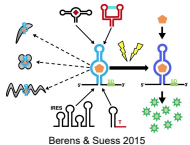
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



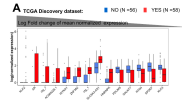
Hendel et al, 2015; Agrotis & Ketteler, 2015



Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



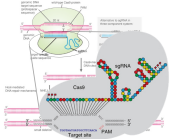
[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)

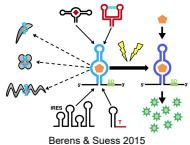
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



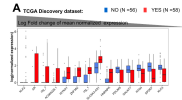
Hendel et al, 2015; Agrotis & Ketteler, 2015



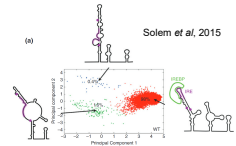
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



[NGuyen *et al*, 2021]



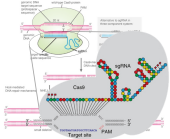
Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy, Cystic fibrosis, Rett syndrome...

RiboNucleic Acids (RNAs)

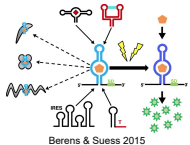
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



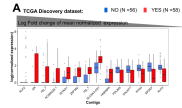
Hendel et al, 2015; Agrotis & Ketteler, 2015



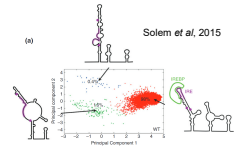
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



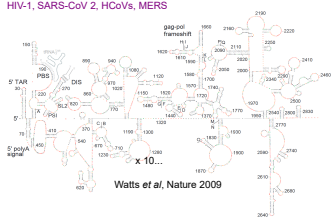
[NGuyen *et al*, 2021]



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

RiboNucleic Acids (RNAs)

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

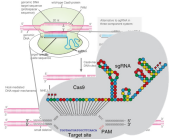


Watts *et al*, Nature 2009

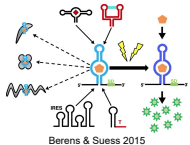
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



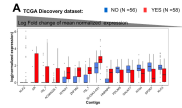
Hendel et al, 2015; Agrotis & Ketteler, 2015



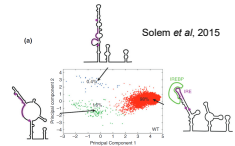
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



[NGuyen *et al*, 2021]

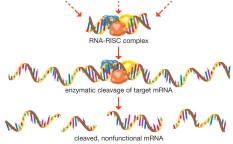


Solem *et al*, 2015

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

RiboNucleic Acids (RNAs)

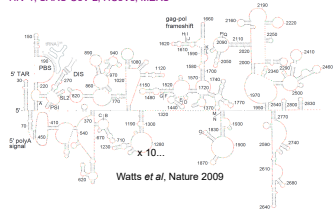
Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1), Hereditary transthyretin amyloidosis (ATTRv), Acute hepatic porphyria (AHP)



Encyclopaedia Britannica, Inc 2013

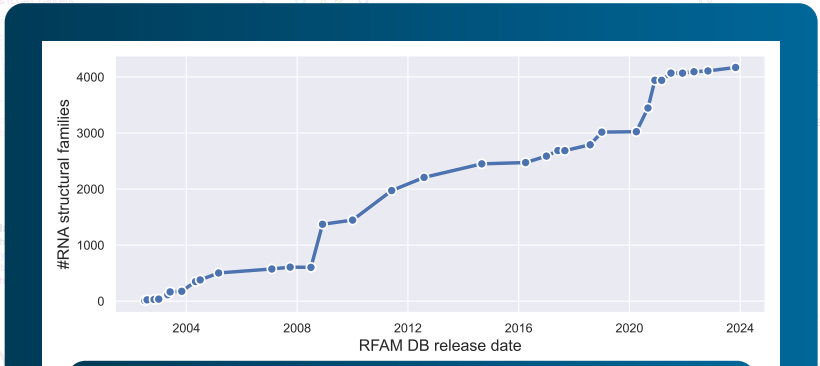
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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



Watts *et al*, Nature 2009

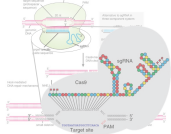
RiboNucleic Acids (RNA) in Human biology/health: Friend **and** Foe!



RNA functional diversity is (largely) enabled by deep structural diversity

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

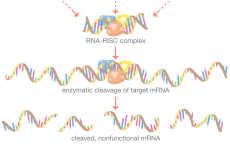
Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



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)

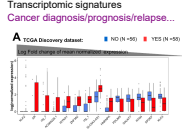


Encyclopaedia Britannica, Inc 2013

Berens & Suess 2015

Sensor of metabolites
Riboswitches

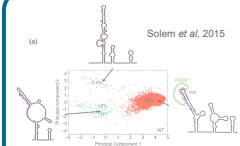
Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...



[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)

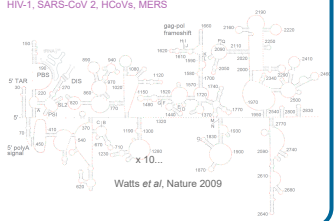
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)



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

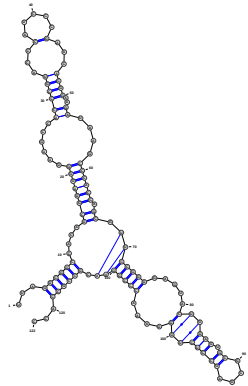


Watts *et al*, Nature 2009

RNA structure(s)

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

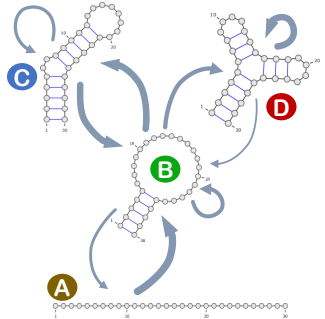
```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

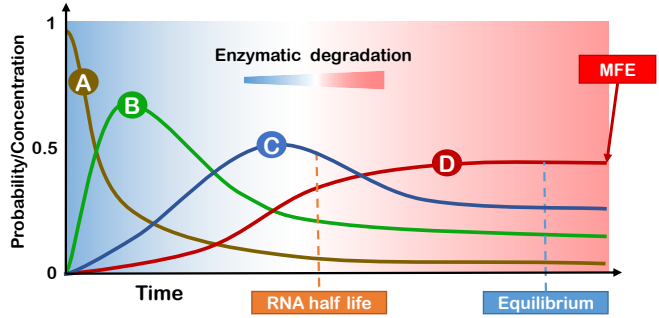


Primary struct. Secondary (2D) struct. Tertiary ($\approx$ 3D) struct.

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

Paradigms in RNA structural bioinformatics





A – Kinetic Landscape

Continuous-time Markov chain

B – Evolution of concentrations

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

$$\mathbb{P}(S \mid w) = e^{-E(w,S)/RT} / \mathcal{Z}(S) \quad (\mathcal{Z} \text{ partition function})$$

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

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

$\mathcal{O}(99)$ reasons to perform a rational design of structural RNAs

1. To stress test our understanding of how RNA folds
   Misfolded RNAs reveal gaps in our energy models and conformational descriptions

2. To create building blocks for synthetic systems
   Rationally-designed RNAs increase orthogonality

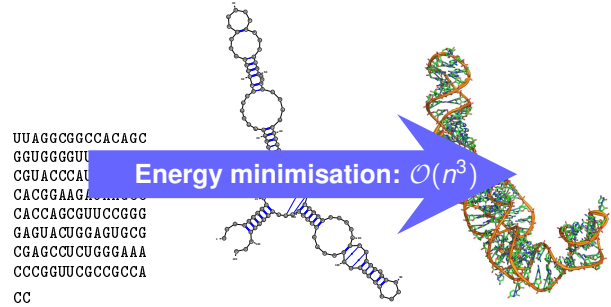
3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

4. To help search for homologous sequences (remote homology)
   Include designed/unseen homologs in multiple sequence alignments (e.g. cov. models)

5. To perform controlled experiments
   Test statistical support of theories (w/o confirmation bias)

6. To fuel RNA-based therapeutics
   Sequence-based (siRNA, synthetic genes), but structure also matters

Minimum Free-Energy (MFE) folding



Nussinov's [PNAS, 1980] $\Theta(n^3)$ algorithm finds Min. Free-Energy structure (base-pairs)

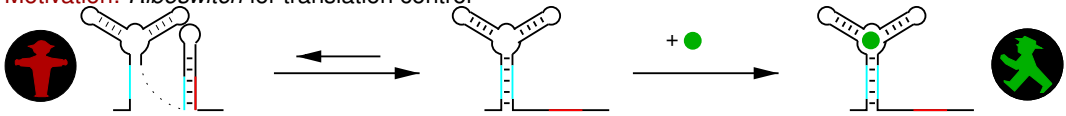
$$\mathrm{MFE}_{i,j} = \min \left\{ \mathrm{MFE}_{i+1,j}; \sum_k E(i,k) + \mathrm{MFE}_{i+1,k-1} + \mathrm{MFE}_{k+1,j} \right\}$$

Trivially adapted into joint OPT of Energy and Codon Adaptation Index for given protein sequence
→ Yield-optimized mRNA vaccines [Zhang *et al*, Nature 2023]

Positive multiple design

Positive design for multiple RNA structures

Motivation: *Riboswitch* for translation control



Multiple target structures

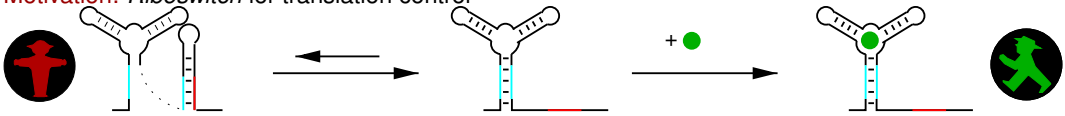
```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...)))...
```

Objective: To randomly generate RNA sequences under constraints

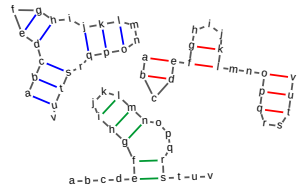
1. Validity for targeted structures wrt base pairing nucleotides
2. Stability (low free-energy, comparable across structures...) of target structures
3. Constrained composition: (prescribed G+C content), ± motifs...

Positive design for multiple RNA structures

Motivation: *Riboswitch* for translation control



Multiple target structures



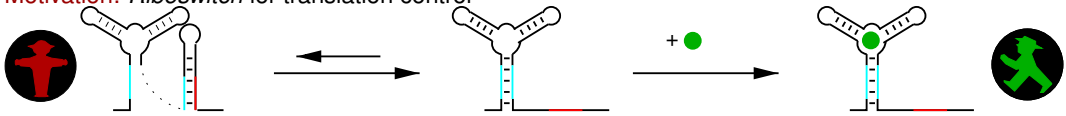

```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...)))...
```

Objective: To randomly generate RNA sequences under constraints

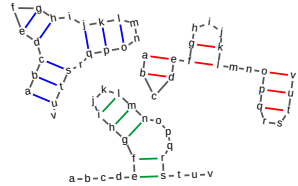
1. Validity for targeted structures wrt base pairing nucleotides
2. Stability (low free-energy, comparable across structures...) of target structures
3. Constrained composition: (prescribed G+C content), ± motifs...

Positive design for multiple RNA structures

Motivation: *Riboswitch* for translation control



Multiple target structures

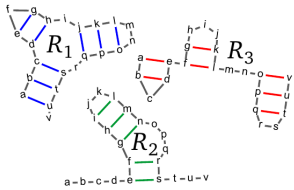



```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...))...
```

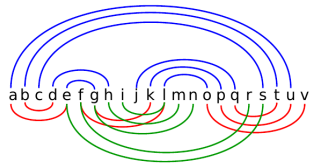
Objective: To randomly generate RNA sequences under constraints

1. Validity for targeted structures wrt base pairing nucleotides
2. Stability (low free-energy, comparable across structures...) of target structures
3. Constrained composition: (prescribed G+C content), $\pm$ motifs...

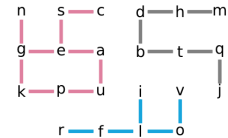
Counting the number of valid sequences



i) Input Structures



ii) Merged Base-Pairs



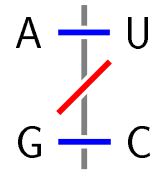
iii) Compatibility Graph

Question: How many valid sequences over $\Sigma^n := \{A, C, G, U\}^n$?

Problem (#ValidSequences)

Input: Secondary structures $\mathcal{R} = \{R_1, \ldots, R_k\}$ of length n
Output: Number of valid sequences

$$\#Designs = \left|\left\{S \in \Sigma^n \mid \forall (i,j) \in R_\ell, (S_i, S_j) \text{ forms a valid base pair}\right\}\right|$$



Valid base pairs

Valid sequences and independent sets [Hammer, P, Wang, Will, RECOMB 2018]

Theorem (Designs ≈ Independent sets)

Let G be a bipartite and connected dependency graph:

$$\#\text{Designs}(G) = 2 \times \#\text{Designs}^{\star}(G) = 2 \times \#\text{IndSets}(G)$$

$$\Rightarrow \text{For general graphs: } \#\mathit{Designs}(G) = \prod_{cc \in CC(G)} 2 \times \#\mathit{IndSets}(cc) = 2^{|CC(G)|} \times \#\mathit{IndSets}(G)$$

But $\#\mathit{IndSets}(G)$ is #P-hard on bipartite graphs ($\#\mathit{BIS}$) [Dyer & Greenhill'00]

Theorem (Classic counting complexity)

Counting #Designs is #P-hard.

No Poly-Time algorithm for #Designs(G) unless $\#P = FP$ ($\Rightarrow P = NP$)

Theorem (Parameterized complexity for treewidth)

#Designs is Fixed-Parameter Tractable ($O(f(tw).P(n))$) for the Treewidth parameter tw

Valid sequences and independent sets [Hammer, P, Wang, Will, RECOMB 2018]

Theorem (Designs ≈ Independent sets)

Let G be a bipartite and connected dependency graph:

$$\#\text{Designs}(G) = 2 \times \#\text{Designs}^{\star}(G) = 2 \times \#\text{IndSets}(G)$$

$$\Rightarrow \text{For general graphs: } \#\mathit{Designs}(G) = \prod_{cc \in CC(G)} 2 \times \#\mathit{IndSets}(cc) = 2^{|CC(G)|} \times \#\mathit{IndSets}(G)$$

But $\#\mathit{IndSets}(G)$ is #P-hard on bipartite graphs ($\#\mathit{BIS}$) [Dyer & Greenhill'00]

Theorem (Classic counting complexity)

Counting #Designs is #P-hard.

No Poly-Time algorithm for #Designs(G) unless $\#P = FP$ ($\Rightarrow P = NP$)

Theorem (Parameterized complexity for treewidth)

#Designs is Fixed-Parameter Tractable ($O(f(tw).P(n))$) for the Treewidth parameter tw

Valid sequences and independent sets [Hammer, P, Wang, Will, RECOMB 2018]

Theorem (Designs ≈ Independent sets)

Let G be a bipartite and connected dependency graph:

$$\#\text{Designs}(G) = 2 \times \#\text{Designs}^{\star}(G) = 2 \times \#\text{IndSets}(G)$$

$$\Rightarrow \text{For general graphs: } \#\mathit{Designs}(G) = \prod_{cc \in CC(G)} 2 \times \#\mathit{IndSets}(cc) = 2^{|CC(G)|} \times \#\mathit{IndSets}(G)$$

But $\#\mathit{IndSets}(G)$ is #P-hard on bipartite graphs ($\#\mathit{BIS}$) [Dyer & Greenhill'00]

Theorem (Classic counting complexity)

Counting #Designs is #P-hard.

No Poly-Time algorithm for #Designs(G) unless $\#P = FP$ ($\Rightarrow P = NP$)

Theorem (Parameterized complexity for treewidth)

#Designs is Fixed-Parameter Tractable ($O(f(tw).P(n))$) for the Treewidth parameter tw

Tree decomposition and treewidth

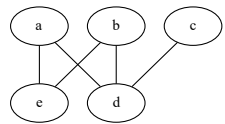
A tree decomposition T for graph $G = (V, E)$:

1. Nodes of T = Bags, *i.e.* subsets of V;
2. Every vertex must be found in ≥ 1 bag;
3. Each edge must be represented in ≥ 1 bag;
4. Nodes featuring any $v \in V$ form a connected subtree of T

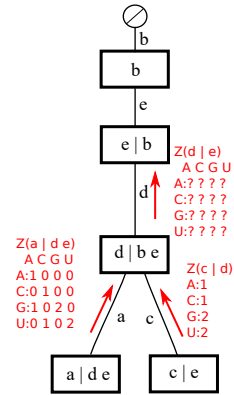
```
a   b   c   d   e

(   .   .   )   .
.   (   (   )   )
(   (   .   )   )
```

Target structures



Dependency graph



Tree decomposition

w : Width of tree decomposition T ($=\max_{b \in B} |b| - 1$)

Let $b = (v; v_1 \ldots) \subseteq V$ a bag of T, and T_b be the subtree rooted at b

$$\#\text{Designs}(T_b \mid b_2 \leftarrow v_2 \ldots) = \sum_{\substack{b_1 \leftarrow v_1 \\ v_1 \in \{A,C,G,U\}}} \prod_{c \text{ child of } b} \#\text{Designs}(T_c \mid b_1 \leftarrow v_1, b_2 \leftarrow v_2 \ldots)$$

$\rightarrow$ #Designs (resp. **partition function**) computable in $\Theta(n\, k\, 2^w)$ time for k struct. of length n

Tree decomposition and treewidth

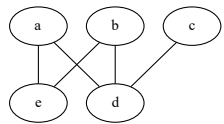
A tree decomposition T for graph $G = (V, E)$:

1. Nodes of T = Bags, *i.e.* subsets of V;
2. Every vertex must be found in ≥ 1 bag;
3. Each edge must be represented in ≥ 1 bag;
4. Nodes featuring any $v \in V$ form a connected subtree of T

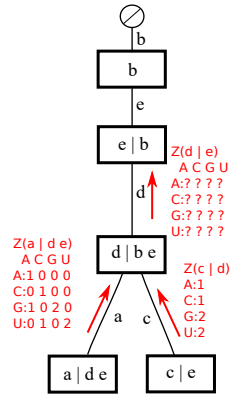
```
a   b   c   d   e

(   .   .   )   .
.   (   (   )   )
(   (   .   )   )
```

Target structures



Dependency graph



Tree decomposition

w : Width of tree decomposition T (=$\max_{b \in B} |b| - 1$)
Let $b = (v; v_1 \ldots) \subseteq V$ a bag of T, and T_b be the subtree rooted at b

$$\#\text{Designs}(T_b \mid b_2 \leftarrow v_2 \ldots) = \sum_{\substack{b_1 \leftarrow v_1 \\ v_1 \in \{A,C,G,U\}}} \prod_{c \text{ child of } b} \#\text{Designs}(T_c \mid b_1 \leftarrow v_1, b_2 \leftarrow v_2 \ldots)$$

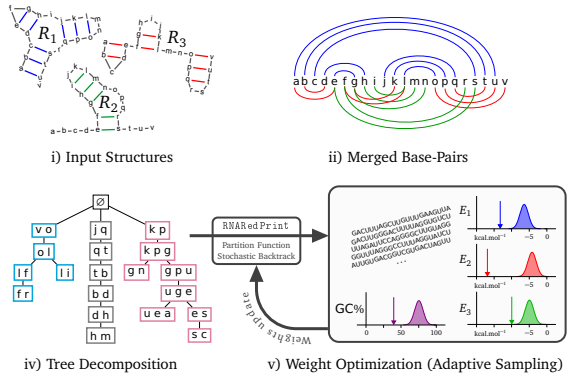
$\rightarrow$ #Designs (resp. **partition function**) computable in $\Theta(n\, k\, 2^w)$ time for k struct. of length n

Tree decomposition and Boltzmann sampling of sequences

First count (FPT), then stochastic backtrack $\Theta(n)$, based on Min width (tw) decomposition (FPT)

Theorem

Positive multiple design (unif./Boltzmann distr.) FPT for the treewidth parameter



i) Input Structures ii) Merged Base-Pairs iii) Compatibility Graph

iv) Tree Decomposition v) Weight Optimization (Adaptive Sampling) vi) Final Designs

RNARedPrint [Hammer, P, Wang, Will, RECOMB 2018 & BMC Bioinfo 2019]

Infrared, a declarative (weighted) constraint satisfaction framework

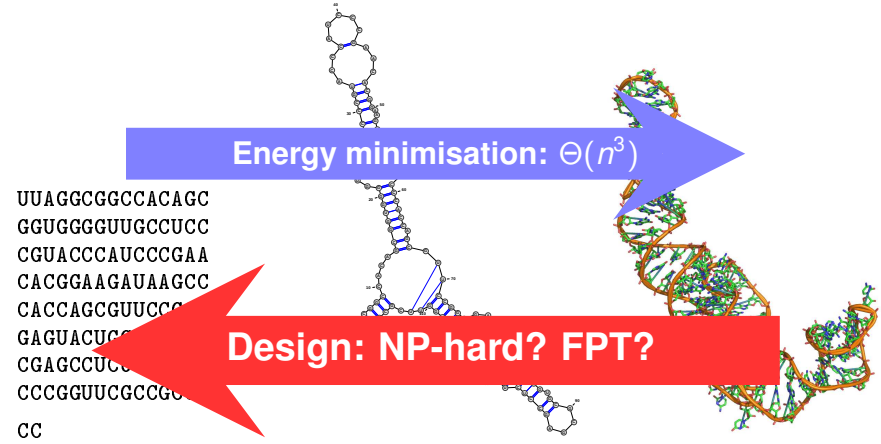
```python
import infrared as ir
import infrared.rna as rna
n, bps = len(target), rna.parse(target)
model = ir.Model(n, 4)
model.add_constraints(rna.BPComp(i, j) for (i, j) in bps)
model.add_functions([rna.GCCont(i) for i in range(n)], 'gc')
model.add_functions([rna.BPEnergy(i, j, (i-1, j+1) not in bps)
    for (i,j) in bps], 'energy')
model.set_feature_weight(-1.5, 'energy')
sampler = ir.Sampler(model)
samples = [sampler.sample() for _ in range(10)]
```

InfraRed [Yao *et al*, Algorithms Mol Biol 2024] generalizes RNARedPrint beyond RNA design tasks:

- Generic solver for sparse/weighted constraints networks, fueled by tree decomposition;
- Supports: optimization, exact sampling (unif./Boltzmann distr.), integers-value feature targets;
- Critical sections in `C`, conveniently interfaced in `Python`
- Illustrated on threading, network-based parsimony, alignment...

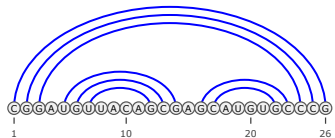
Inverse folding

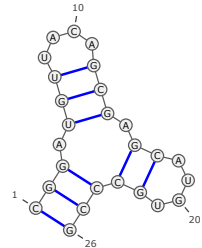
Minimum Free-Energy (MFE) folding



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

Energy model

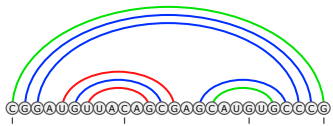


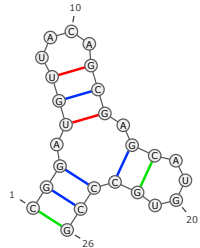


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops, ...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

Energy model

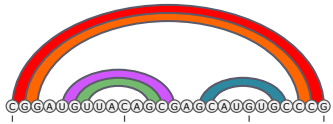


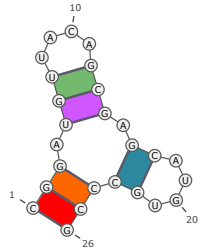


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

Energy model

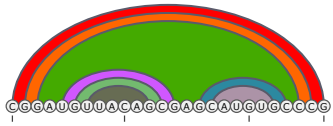


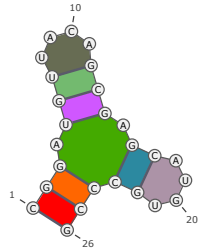


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

Energy model

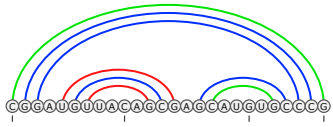


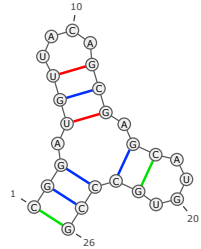


This section: Restriction to valid base-pairs = $\{(A, U), (G, C), (G, U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m, a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1, n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S, R)$: Sum of energies for motifs in R, given sequence S

Energy model



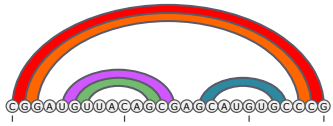


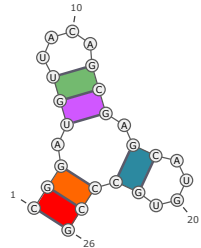
This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

$$E_R = 2 \cdot \Delta \left(\begin{matrix} U \\ | \\ G \end{matrix} \right) + 4 \cdot \Delta \left(\begin{matrix} G \\ | \\ C \end{matrix} \right) + 2 \cdot \Delta \left(\begin{matrix} C \\ | \\ G \end{matrix} \right)$$

Energy model



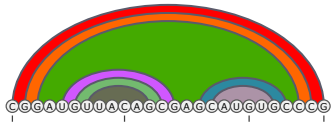


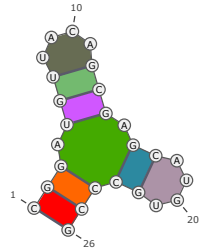
This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

$$E_R = \Delta\left(\begin{smallmatrix} C & G \\ G & C \end{smallmatrix}\right) + \Delta\left(\begin{smallmatrix} G & G \\ C & C \end{smallmatrix}\right) + \Delta\left(\begin{smallmatrix} U & G \\ G & C \end{smallmatrix}\right) + \Delta\left(\begin{smallmatrix} U & G \\ G & C \end{smallmatrix}\right) + \Delta\left(\begin{smallmatrix} U & G \\ G & C \end{smallmatrix}\right)$$

Energy model





This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- RNA structure R: Set of non-crossing base pairs (BPs)
- Motifs: Connected positions + content (e.g. Base Pairs, Stacking, Loops...)
- Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(m,a) \in \mathbb{R} \cup \{+\infty\}$, $m \subset [1,n]$, $a \in \Sigma^{|m|}$
  Free-energy $E(S,R)$: Sum of energies for motifs in R, given sequence S

$$E_R = \Delta\left(\begin{array}{c}C\ G\\G\ C\end{array}\right) + \Delta\left(\begin{array}{c}G\ G\\C\ C\end{array}\right) + \Delta\left(\begin{array}{c}U\ G\\G\ C\end{array}\right) + \Delta\left(\begin{array}{c}U\ G\\G\ C\end{array}\right) + \Delta\left(\begin{array}{c}U\ G\\G\ C\end{array}\right)$$
$$+ \Delta\left(\begin{array}{c}C\ A\\U\ G\end{array}\right) + \Delta\left(\begin{array}{c}\text{A U G A}\\\text{G C}\end{array}\right) + \Delta\left(\begin{array}{c}C\ A\\G\ U\end{array}\right)$$

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:

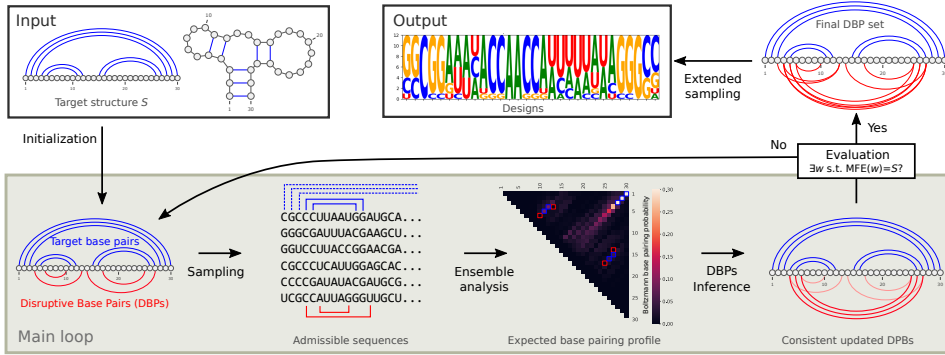
$$\forall R' \neq R \text{ compatible with } \omega : \; E(\omega, R') \geq E(\omega, R) + \Delta$$

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...
- Existing algorithms Heuristics, ML or Exponential time

RNA POsitive and Negative Design (RNAPOND) [Yao *et al*, RECOMB 2021]



RNAPond: Human-inspired heuristics based on identification of Disruptive Base Pairs (DBPs)

- Sample sequences compatible with target & avoiding DBPs ← Infrared (NP-hard, FPT on treewidth)
- Identify and forbid recurrent DPBs
- Iterate until solution found or treewidth threshold reached (def. $tw \leq 10$)

Close to state of the art heuristics in a crowded field

Existing approaches for negative design

Bio-inspired algorithms...

- FRNAKenstein - Hein@Oxford
- AntaRNA - Backofen@Freiburg
- ERD - Ganjtabesh@Tehran

...exact (exptime) approaches...

- RNAIFold - Clote@Boston College
- CO4 - Will@Leipzig

...based on local search...

- RNAInverse - TBI Vienna
- Info-RNA - Backofen@Freiburg
- RNA-SSD - Condon@UBC
- (Inca)RNAFBinv - Barash@BGU
- NUPack - Pierce@Caltech

...or ML/DL (Ribodiffusion...)

Typical issues:

- Single solution
- Strong impact of initialization strategy
- Synthesized sequences do not necessarily fold properly (kinetics)
- Overly GC-rich sequences
- Generative ML usually fails to generalize
- Few options to produce negative results

$\Rightarrow$ Establish combinatorial foundations!

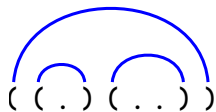
Inverse Folding in Base Pair maximization (maxBP) model

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: $|R'| < |R|$.



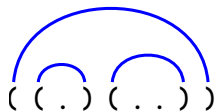
Inverse Folding in Base Pair maximization (maxBP) model

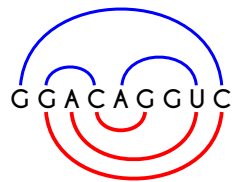
Definition (Inverse folding in maxBP)

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

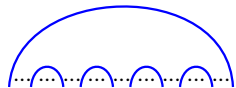
Partial characterization of designable structures [Hales *et al*, CPM'15 & Algorithmica'17]

- 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:

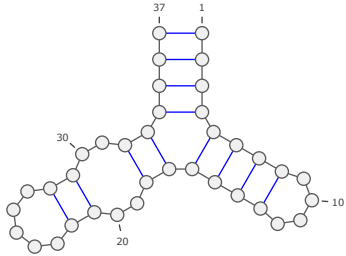
- Level of a base pair = $\#● - \#○$ on path to root
- Coloring separated if proper and ◍ base pairs and unpaired positions at different levels

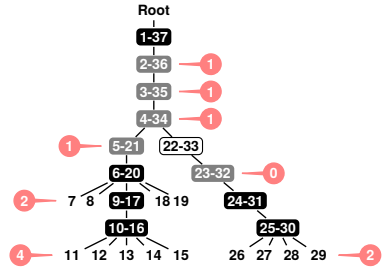
Theorem ([Hales *et al.* Algorithmica 2017])

$\exists$ 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

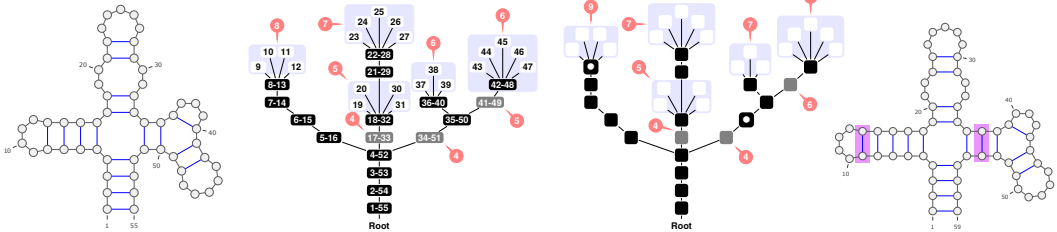
Structural approximate design

Artistic license is often acceptable (*aka* RNA function is typically robust to minor structural edits)

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

2-approximate design for any structure avoiding m_5 and $m_{3\circ}$ in $\Theta(n)$ time

Idea: Shift unpaired/leaves and ● to odd/even levels resp. by adding ≤ 1 BP in each helix



1) Target structure 2) Greedy proper coloring 3) Separated proper coloring 4) Designable structure

. . . but minimizing #Edits to make target separable is NP-hard. (Parameterized Complexity?)

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:

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

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)

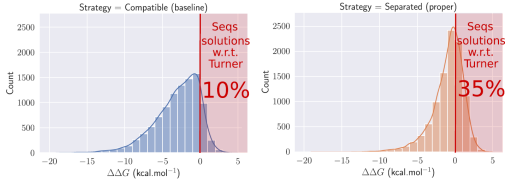
Also, any target structure with 3^+ BP/helix (+ many with helix length 2) is (modulo) 3-separated

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

For targets without isolated stacks and BPs, maxBP inverse folding solvable in $\Theta(n)$ time

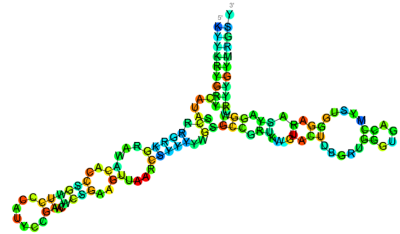
Is maxBP inverse folding really NP hard?

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 maxStacks model [incoming...]

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

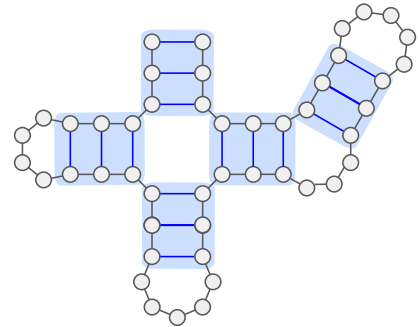
Definition (maxStacks inverse folding)

Input: Target secondary structure R

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

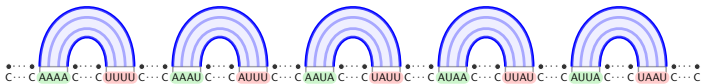
- Target R comp. with ω;
- $\forall R' \neq R$, compatible with ω:

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



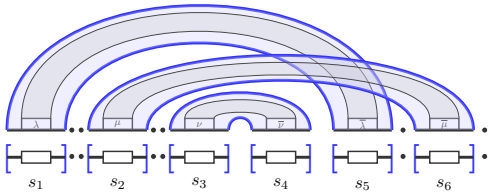
The maxStacks model differs from maxBP:

- Multiloops of arbitrary degree Δ can be designed



- For target structures have min helix length $\lceil \log_3(\Delta) \rceil + \mathcal{O}(1)$, inverse folding solved in $\Theta(n)$ time

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:

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

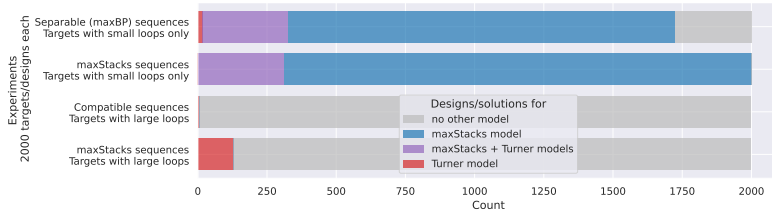
Remark: General pseudoknots allowed both for input and alternative structures

Theorem (Incoming...)

For target structures with H (crossing) helices + min length in $\lceil \log_3(H) \rceil + \mathcal{O}(1)$, maxStack inverse folding with general PK is solvable in $\Theta(n)$ time

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 are known to really help design, now we understand why
- Pseudoknot-aware inverse folding specializes into constrained interaction design

Open questions (help wanted):

- Parameterized complexity of finding separated coloring?
- Proportion of separable structures within Neutral Networks?
- Complexity of maxBP inverse folding w/o sequence constraints?
- Complexity of inverse folding in maxStack model?

Extensions and perspectives

- Onwards to the bench: Inspiration from exp. test of designs for SAM riboswitches
- More complex/realistic energy models (Stacks, Turner's Nearest Neighbors?)

  Extended conformational spaces (pseudoknots, non-canonical BPs)
- Kinetics-aware design (prescribed intermediates, energy barriers. . .)
- (Parameterized) complexity of general inverse folding?
- Potential/limitations of Machine Learning towards RNA design:

  Can ML learn negative design strategies from extent sequences? Novelty/orthogonality?
- NeuTral networks: Exponentially less designable structs (*aka* phenotypes) than initially thought

  $\rightarrow$ Refine phenotype/genotype studies?

Acknowledgments

Thanks to AIChemist organizers

- Ecole Polytechnique
  - H.T. Yao, B. Marchand
  - T. Boury, S. Gardelle
  - S. Will, S. Berkemer
  - M. Régnier, A. Héliou
- Univ Gustave Eiffel
  - L. Bulteau
- Faculté Pharmacie@Univ Paris Cité
  - B. Sargueil, P. Hardouin
- Stat. Physics@ENS Paris
  - J. Fernandez de Cossio Diaz
  - S. Cocco, R. Monasson, J.
- Simon Fraser University
  - J. Hales, J. Manuch, L. Stacho
  - C. Chauve
- McGill University
  - J. Waldispühl
- Université du Québec à Montréal
  - V. Reinharz
- University of Vienna
  - S. Hammer, R. Lorenz
- Ben Gurion University
  - D. Barash, M. Drory, A. Churkin