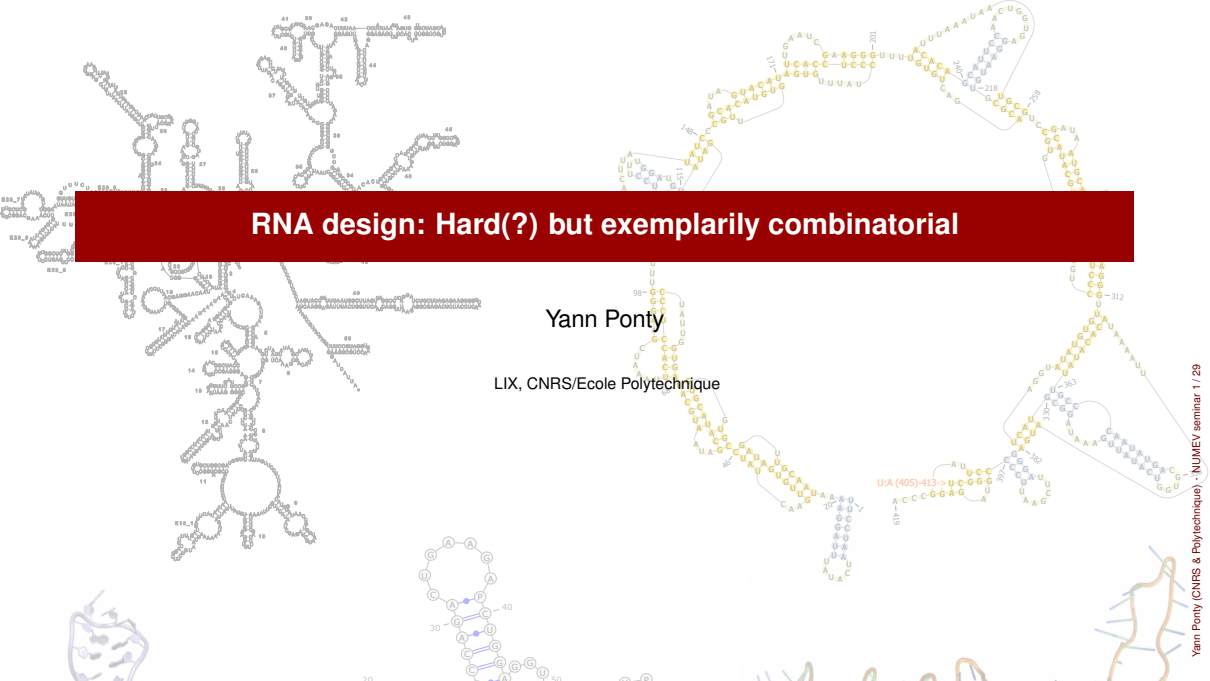
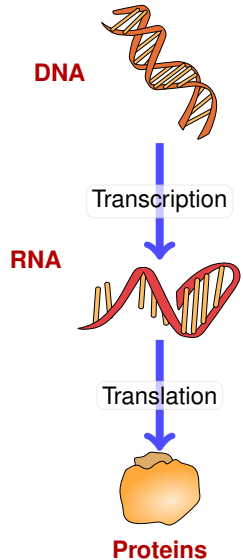
RNA design: Hard(?) but exemplarily combinatorial

Yann Ponty

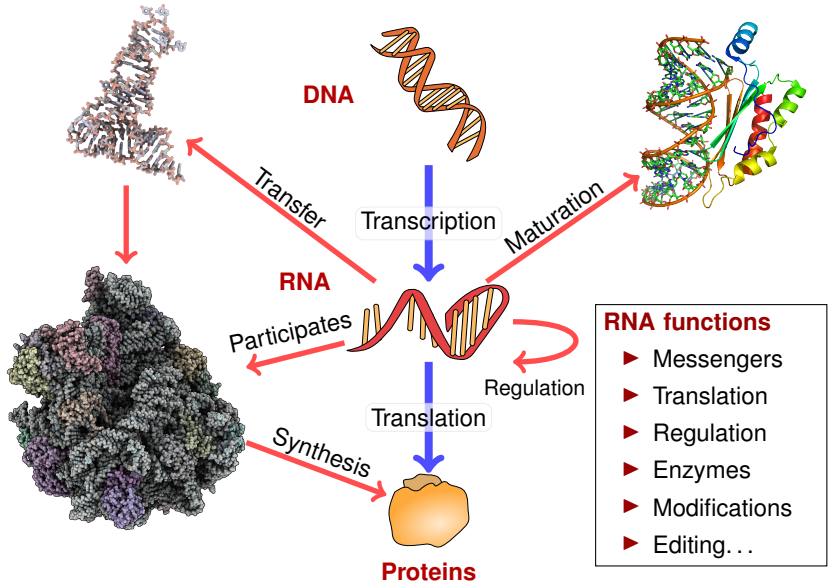
LIX, CNRS/Ecole Polytechnique



Fundamental dogma of molecular biology



Fundamental dogma of molecular biology (v2.0)



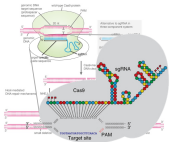
RiboNucleic Acids (RNA) in Human biology/health: Friend **and** Foe!

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...



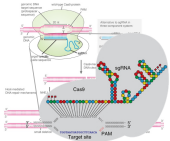
Hendel et al, 2015; Agrotis & Ketteler, 2015

RiboNucleic Acids (RNAs)

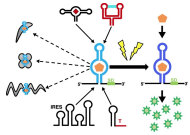
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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

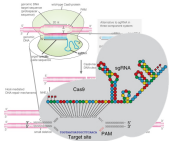
Sensor of metabolites
Riboswitches

RiboNucleic Acids (RNAs)

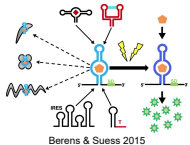
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



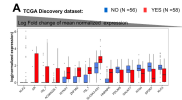
Hendel et al, 2015; Agrotis & Ketteler, 2015



Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



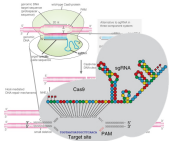
[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)

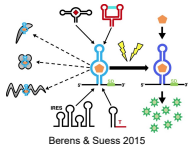
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



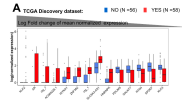
Hendel et al, 2015; Agrotis & Ketteler, 2015



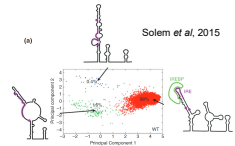
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



[NGuyen *et al*, 2021]



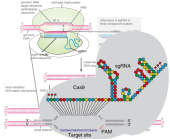
Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy, Cystic fibrosis, Rett syndrome...

RiboNucleic Acids (RNAs)

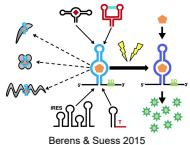
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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...



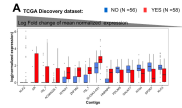
Hendel et al, 2015; Agrotis & Ketteler, 2015



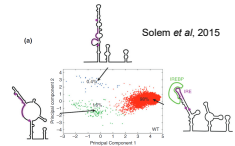
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



[NGuyen *et al*, 2021]



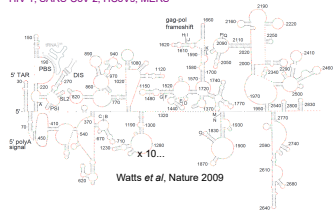
Solem *et al*, 2015

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?)

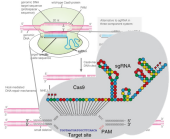
Genomic material for Human pathogens
HIV-1, SARS-CoV 2, HCoVs, MERS



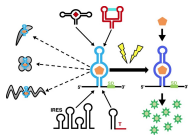
Watts *et al*, Nature 2009

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

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



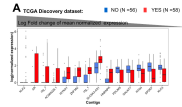
Hendel et al, 2015; Agrotis & Ketteler, 2015



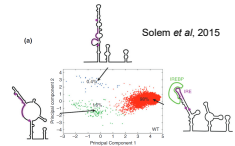
Berens & Suess 2015

Sensor of metabolites
Riboswitches

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



[NGuyen *et al*, 2021]

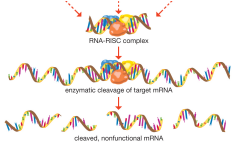


Solem *et al*, 2015

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

RiboNucleic Acids (RNAs)

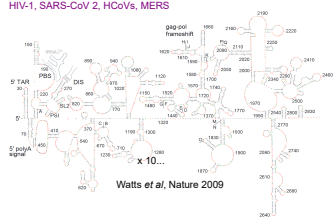
Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1), Hereditary transthyretin amyloidosis (ATTRv), Acute hepatic porphyria (AHP)



Encyclopaedia Britannica, Inc 2013

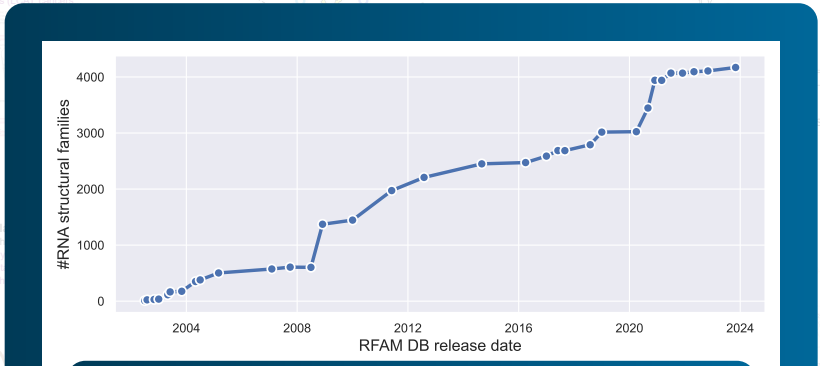
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

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



Watts *et al*, Nature 2009

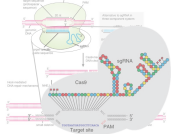
RiboNucleic Acids (RNA) in Human biology/health: Friend **and** Foe!



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

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

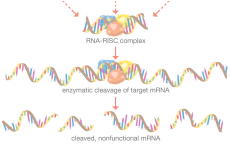
Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



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)

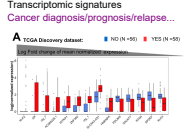


Encyclopaedia Britannica, Inc 2013

Berens & Suess 2015

Sensor of metabolites
Riboswitches

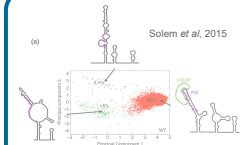
Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...



[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)

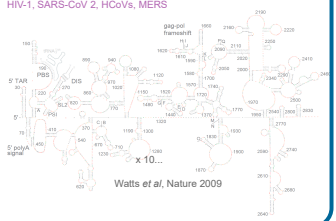
Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)



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

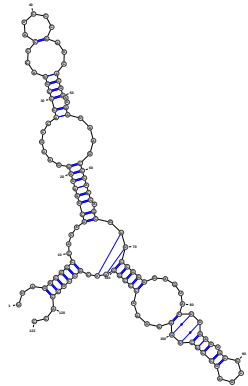


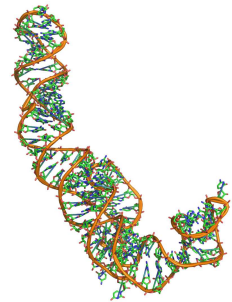
Watts *et al*, Nature 2009

RNA structure(s)

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

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

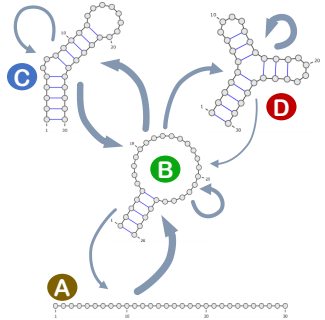


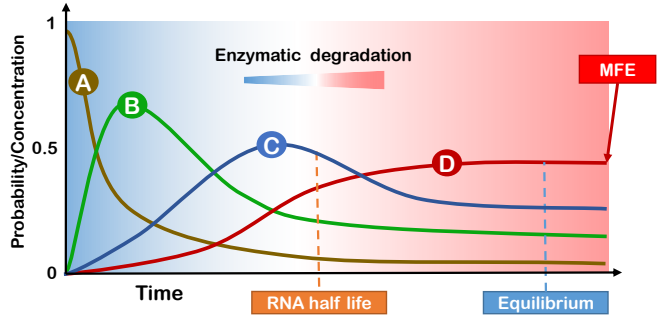


Primary struct. Secondary (2D) struct. Tertiary ($\approx$ 3D) struct.

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

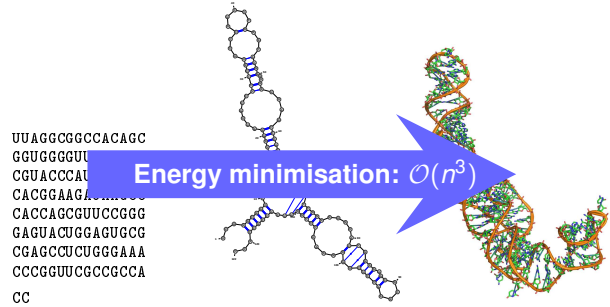
3. To assess the significance of observed phenomenon
   Random models should include all established traits, including adopting a well-defined structure

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)

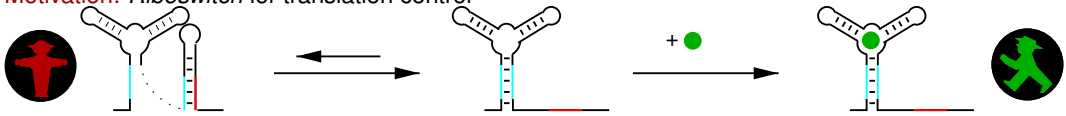
$$\mathrm{MFE}_{i,j} = \min \left\{ \mathrm{MFE}_{i+1,j}; \sum_k E(i,k) + \mathrm{MFE}_{i+1,k-1} + \mathrm{MFE}_{k+1,j} \right\}$$

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

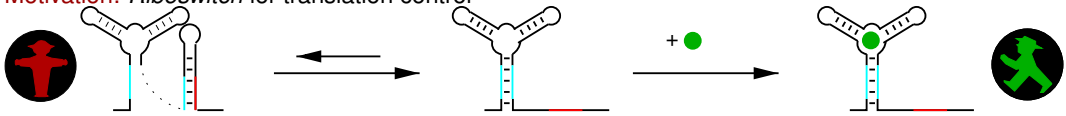
```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...)))...
```

Objective: To randomly generate RNA sequences under constraints

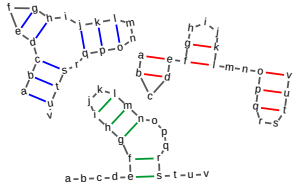
1. Validity for targeted structures wrt base pairing nucleotides
2. Stability (low free-energy, comparable across structures...) of target structures
3. Constrained composition: (prescribed G+C content), $\pm$ motifs...

Positive design for multiple RNA structures

Motivation: *Riboswitch* for translation control



Multiple target structures



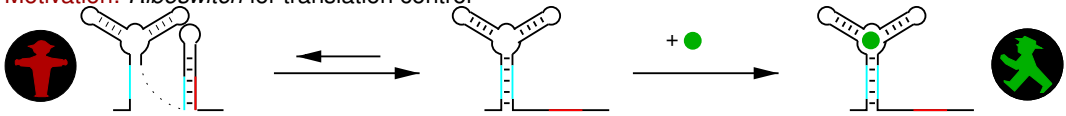

```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...))...
```

Objective: To randomly generate RNA sequences under constraints

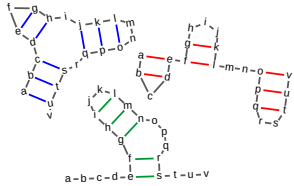
1. Validity for targeted structures wrt base pairing nucleotides
2. Stability (low free-energy, comparable across structures...) of target structures
3. Constrained composition: (prescribed G+C content), ± motifs...

Positive design for multiple RNA structures

Motivation: *Riboswitch* for translation control



Multiple target structures

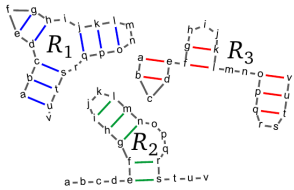



```
abcdefghijklmnopqrstuv
(((((.)).(((..))).))).
((.))((...))..(((..)))
....(((((..)))...))...
```

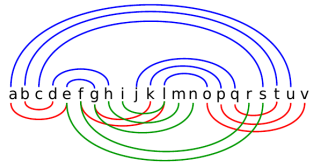
Objective: To randomly generate RNA sequences under constraints

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...

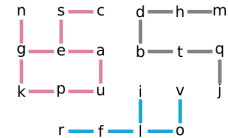
Counting the number of valid sequences



i) Input Structures



ii) Merged Base-Pairs



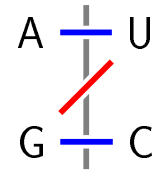
iii) Compatibility Graph

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

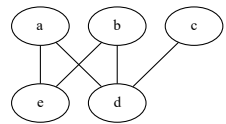
A tree decomposition T for graph $G = (V, E)$:

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

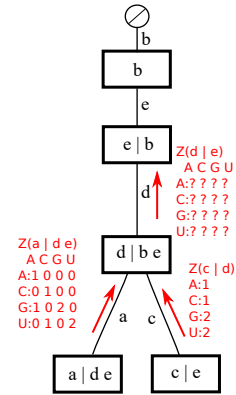
```
a   b   c   d   e

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

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

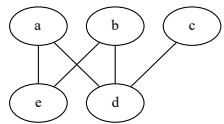
A tree decomposition T for graph $G = (V, E)$:

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

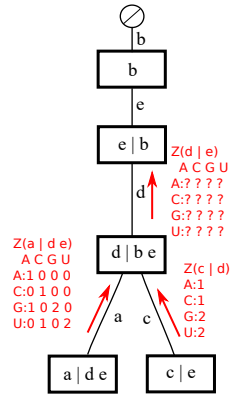
```
a   b   c   d   e

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

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

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

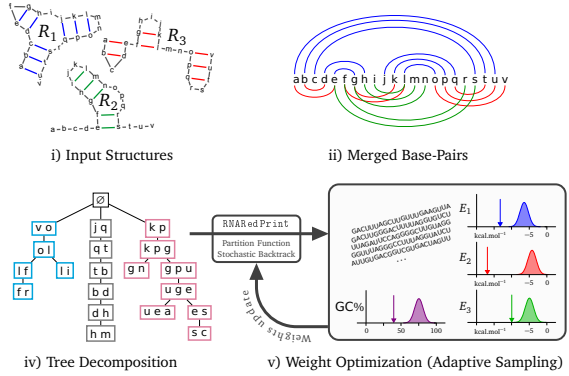
$\rightarrow$ #Designs (resp. **partition function**) computable in $\Theta(n\, k\, 2^w)$ time for k struct. of length n

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



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

Infrared, a declarative (weighted) constraint satisfaction framework

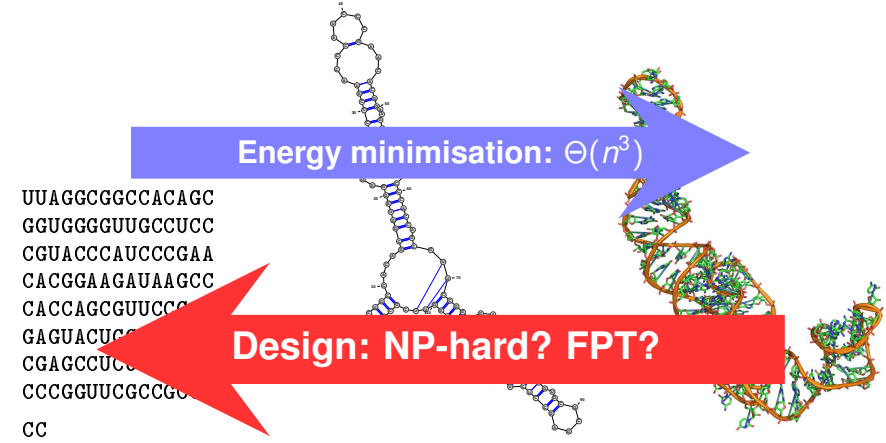
```python
import infrared as ir
import infrared.rna as rna
n, bps = len(target), rna.parse(target)
model = ir.Model(n, 4)
model.add_constraints(rna.BPComp(i, j) for (i, j) in bps)
model.add_functions([rna.GCCont(i) for i in range(n)], 'gc')
model.add_functions([rna.BPEnergy(i, j, (i-1, j+1) not in bps)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

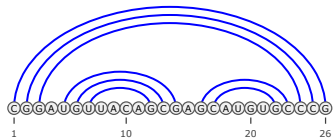
Inverse folding

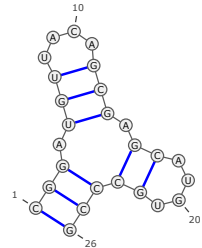
Minimum Free-Energy (MFE) folding



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

Energy model

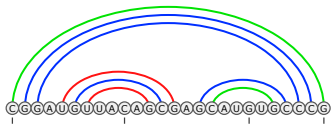


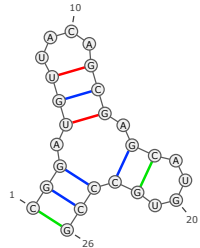


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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

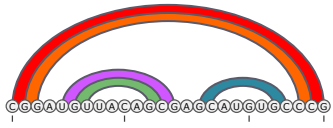


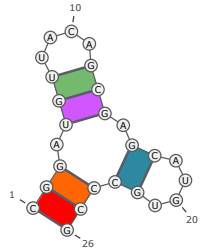


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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

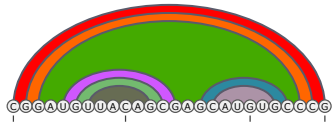


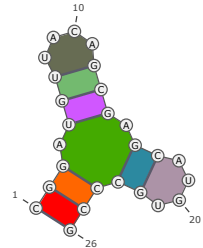


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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

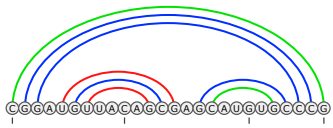


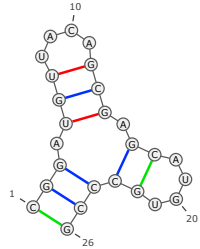


This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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



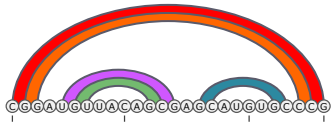


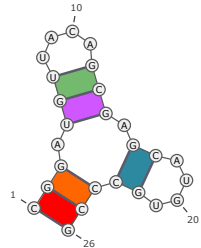
This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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



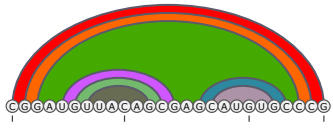


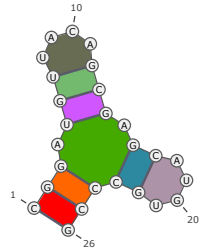
This section: Restriction to valid base-pairs = $\{(A,U),(G,C),(G,U)\}$

- 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

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

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

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

RNA Inverse Folding

Definition (INVERSE-FOLDING(E) problem)

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

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

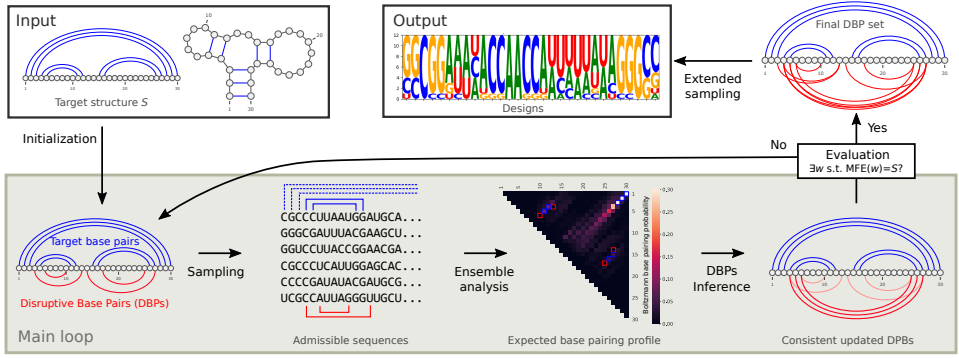
$$\forall R' \in \mathcal{S}_{|S|} \setminus \{R\} : \; E(S, R') \geq E(S, R) + \Delta$$

or $\varnothing$ if no such sequence exists.

Difficult problem: Probably no obvious DP decomposition

- Informally introduced by [Hofacker *et al* Monatshefte für Chemie/Chemical Monthly 1994]
- NP-hardness by including energy model in input [Schnall-Levin *et al*, ICML'08]
- NP-hardness for BP maximization with partial assignment [Bonnet *et al*, RECOMB'18]
- Reason(s): Non locality, no theoretical framework, too many parameters...
- Existing algorithms: Heuristics 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!

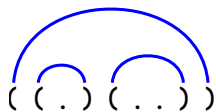
Inverse Folding in unitary Base Pair energy model (aka BP maximization)

Definition (INVERSE-FOLDING(BPmax) problem)

Input: Target secondary structure R, *i.e.* a set of base pairs

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

- Sequence S valid for structure R
- If S valid sequence for alt structure $R' \neq R$, then $|R'| < |R|$



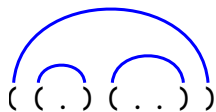
Inverse Folding in unitary Base Pair energy model (aka BP maximization)

Definition (INVERSE-FOLDING(BPmax) problem)

Input: Target secondary structure R, *i.e.* a set of base pairs

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

- Sequence S valid for structure R
- If S valid sequence for alt structure $R' \neq R$, then $|R'| < |R|$



Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

Partial characterization of designable structures

- 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* ($m_{3\circ}$).
  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]

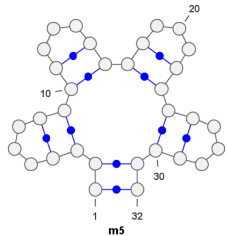
Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

Partial characterization of designable structures

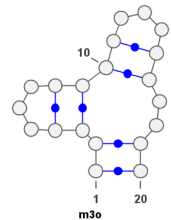
- Saturated structures (all positions paired): Designable $\Leftrightarrow$ Multiloops degrees ≤ 4 (+ $\Theta(n)$ algo.)
- Designable $\Rightarrow$ Avoid multiloops with *degree* ≥ 5 (m_5), or *degree* ≥ 3 *with* ≥ 1 *unpaired* ($m_{3\circ}$).

  Theorem: Similar motifs exist for any energy model and design criterion

  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]



m_5



$m_{3\circ}$

Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

Partial characterization of designable structures

- 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* ($m_{3\circ}$).
  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]
- $\exists$ Separated coloring for structure $\Rightarrow$ Designable (+ $\Theta(n)$ algo.)
  Each base pair $\rightarrow$ one out of 3 colors: ● $\rightarrow$ G · C; ○ $\rightarrow$ C · G; ◍ $\rightarrow$ A · U or U · A.
  Coloring rules: Within each loop, $\#● \leq 1$, $\#○ \leq 1$, $\#◍ \leq 2$ and $\#● + \#○ < 2$

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

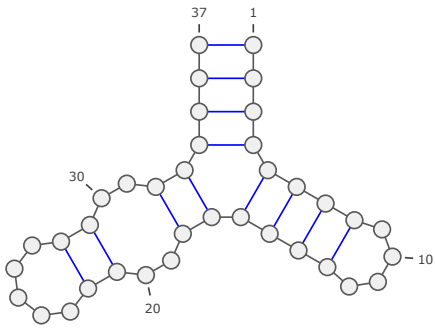
  Idea: Separated sequences (unpaired $\rightarrow$ A) uniquely fold since alt BPs segregate regions with $\#G \neq \#C$

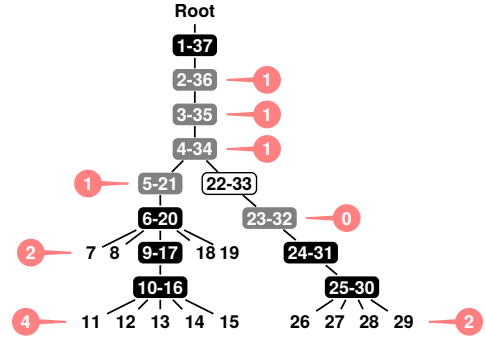
Separated Coloring (example)

Base pairs → 3 colors: ● → G · C; ○ → C · G; ◍ → A · U or U · A.

Coloring rules: Within each loop, #● ≤ 1, #○ ≤ 1, #◍ ≤ 2 and #● + #○ < 2

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





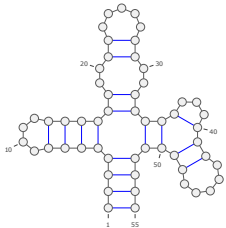
GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC

Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

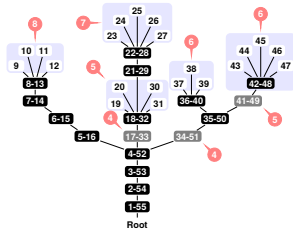
Partial characterization of designable structures

- 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* ($m_{3\circ}$).
  Theorem: Similar motifs exist for any energy model and design criterion
  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]
- $\exists$ Separated coloring for structure $\Rightarrow$ Designable (+ $\Theta(n)$ algo.)
  Corollary: 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

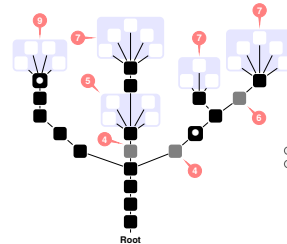
Example of structure-approximating design



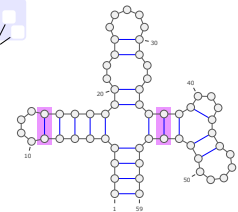
1) Target structure



2) Greedy proper coloring



3) Separated proper coloring



4) Designable structure

Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

Partial characterization of designable structures

- 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* ($m_{3\circ}$).
  Theorem: Similar motifs exist for any energy model and design criterion
  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]
- $\exists$ Separated coloring for structure $\Rightarrow$ Designable (+ $\Theta(n)$ algo.)
  Corollary: 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

  Remark: In structures with helices of length 3+ offsetting always possible

Theorem (Unpublished!)

Inverse folding solvable in Polytime ($\Theta(n)$) for target structures with 3+ BPs helices

Is inverse folding *really* NP hard?

Designability in simple BP-based energy models [Hales *et al*, CPM'15 & Algorithmica'17]

Partial characterization of designable structures

- 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* ($m_{3\circ}$).
  Theorem: Similar motifs exist for any energy model and design criterion
  Corollary: Fraction of designable structures decreases exponentially with n [Yao *et al*, ACM-BCB'19]
- $\exists$ Separated coloring for structure $\Rightarrow$ Designable (+ $\Theta(n)$ algo.)
  Corollary: 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
  Remark: In structures with helices of length 3+ offsetting always possible

Theorem (Unpublished!)

Inverse folding solvable in Polytime ($\Theta(n)$) for target structures with 3+ BPs helices

Is inverse folding *really* NP hard?

Conclusions

- RNA design is a timely topic for Bio Maths/Computer Sciences with practical consequences
- Realistic setting: FPT algo for positive design and general heuristics for inverse folding
- Structure approximating design: a tractable alternative to a (possibly) NP hard model?
- Simple BP maximization: General $\Theta(n)$ algorithm for restricted inputs
- Forbidden motifs: Ubiquitous in DP-based inverse combinatorial optimization
- RNA structure seemingly harder in theory than in practice. Why?

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

Ecole Polytechnique
- H.T. Yao, B. Marchand, T. Boury
- 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

Support

cnrs Agence Nationale de la Recherche ANR FWF

Supplementary Slides

Consequences

Corollary (#Approximability for ≤ 5 structures) [Weitz'06]

For ≤ 5 structures (crossings allowed), #Design(G) can be approximated within any ratio in Poly-time (PTAS)

Corollary (#BIS-hardness for > 5 structures) [Cai, Galanis *et al* 16]

For more than 5 structures (crossings allowed), #Design is equally as hard to approximate as general #BIS.

Why crossings/Pseudokots? Because any bipartite graph of max degree Δ can be decomposed in Δ matchings in Poly-Time (Vizing theorem).

Connection between counting and sampling [Jerrum/Valiant/Vazirani'86].

Conjecture (#BIS-hardness of multiple positive design)

Quasi-uniform generation as hard as approximation of general #BIS

$\Rightarrow$ Sampling #P hard?

Consequences

Corollary (#Approximability for ≤ 5 structures) [Weitz'06]

For ≤ 5 structures (crossings allowed), #Design(G) can be approximated within any ratio in Poly-time (PTAS)

Corollary (#BIS-hardness for > 5 structures) [Cai, Galanis *et al*'16]

For more than 5 structures (crossings allowed), #Design is equally as hard to approximate as general #BIS.

Why crossings/Pseudokots? Because any bipartite graph of max degree Δ can be decomposed in Δ matchings in Poly-Time (Vizing theorem).

Connection between counting and sampling [Jerrum/Valiant/Vazirani'86].

Conjecture (#BIS-hardness of multiple positive design)

Quasi-uniform generation as hard as approximation of general #BIS

$\Rightarrow$ Sampling #P hard?

Consequences

Corollary (#Approximability for ≤ 5 structures) [Weitz'06]

For ≤ 5 structures (crossings allowed), #Design(G) can be approximated within any ratio in Poly-time (PTAS)

Corollary (#BIS-hardness for > 5 structures) [Cai, Galanis *et al*'16]

For more than 5 structures (crossings allowed), #Design is equally as hard to approximate as general #BIS.

Why crossings/Pseudokots? Because any bipartite graph of max degree Δ can be decomposed in Δ matchings in Poly-Time (Vizing theorem).

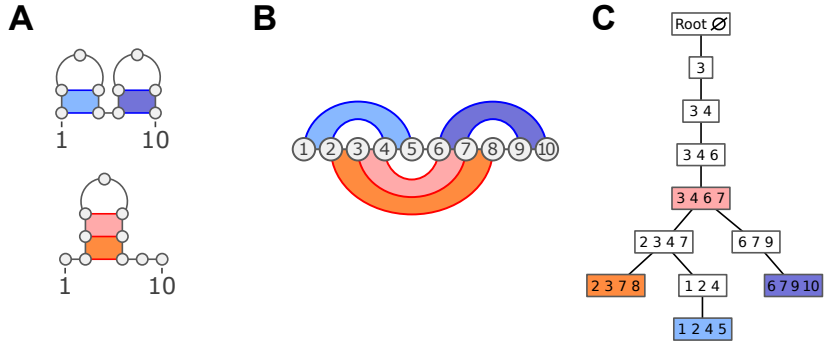
Connection between counting and sampling [Jerrum/Valiant/Vazirani'86].

Conjecture (#BIS-hardness of multiple positive design)

Quasi-uniform generation as hard as approximation of general #BIS

$\Rightarrow$ Sampling #P hard?

Our problem for general free-energy models



Question: Which partition function for valid sequences

Problem (PFDesigns)

Input: Structures $\mathcal{R} = \{R_1, \ldots, R_k\}$ of length n + Weight $(x_1, \ldots, x_k)$

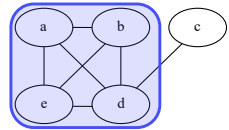
Output: Partition function

$$\mathcal{Z} = \sum_{\substack{S \in \Sigma^n \\ S \text{ valid for } \mathcal{R}}} \prod_{i=1}^{k} x_i^{E(S,R_i)}$$

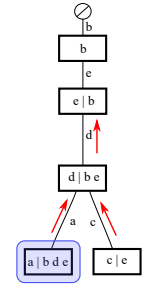
Counting/sampling, the Boltzmann-Gibbs way

```
a   b   c   d   e
(   .   .   )   .
.   (   (   )   )
(   (   .   )   )
```

Target Structures



Dependency Hypergraph



Tree Decomposition

$b = \{b_1, b_2 \ldots\}$: node of D
T_b : subtree rooted at b
w : Width of treedecomposition D

$$\mathcal{Z}(T_b \mid b_2 \leftarrow v_2 \ldots) = \sum_{\substack{b_1 \leftarrow v_1 \\ v_1 \in \{\mathrm{A,C,G,U}\}}} \prod_{i=1}^{k} x_i^{\sum_{E \in b} E(b, v_1, \ldots)} \prod_{c \text{ child of } b} \mathcal{Z}(T_c \mid b_1 \leftarrow v_1, \ldots)$$

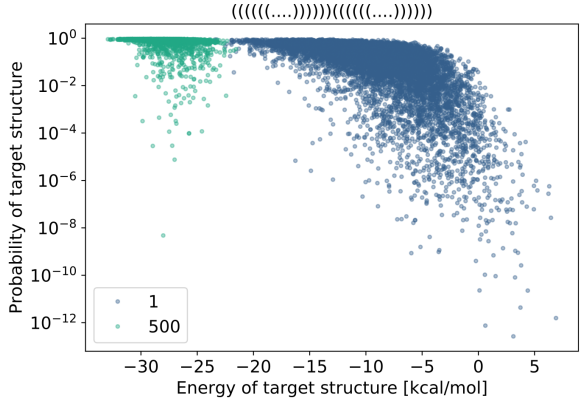
Complexity: $\Theta\left(n\,m\,k + n\,k\,2^{w+\#CC}\right)$ for sampling in Boltzmann-Gibbs distrib.

Practical impact of Boltzmann-Gibbs sampling

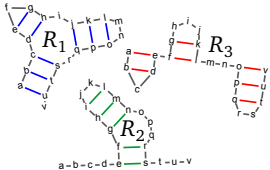
Boltzmann probability of structure R, pour une séquence S:

$$\mathbb{P}(R \mid S) = \frac{e^{-\frac{E(S,R)}{\beta T}}}{\mathcal{Z}_S} \qquad \mathcal{Z}_S := \sum_{R'} e^{-\frac{E(S,R')}{\beta T}}$$

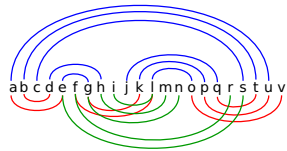
Objectif classique du design négatif ($\rightarrow$ spécificité)



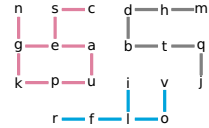
RNARedPrint: a flexible method for (positive) design



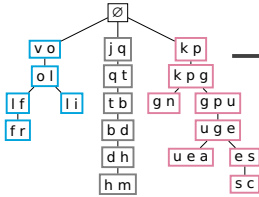
i) Input Structures



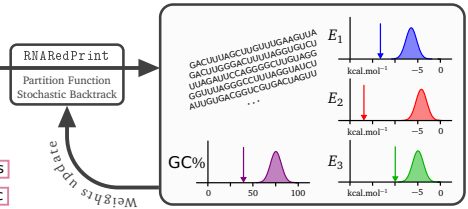
ii) Merged Base-Pairs



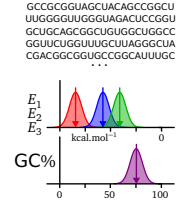
iii) Compatibility Graph



iv) Tree Decomposition



v) Weight Optimization (Adaptive Sampling)



vi) Final Designs

[Hammer/P/Wang/Will, RECOMB'18 + BMC Bioinfo 2019]

- Fixed Parameter Tractable algorithm based on tree width
- Uniform or Boltzmann-Gibbs sampling, to favor diversity and stability
- Multidimensional Boltzmann sampling for controlling free-energy, GC%...

https://github.com/yannponty/RNARedPrint

Multidimensional Boltzmann sampling

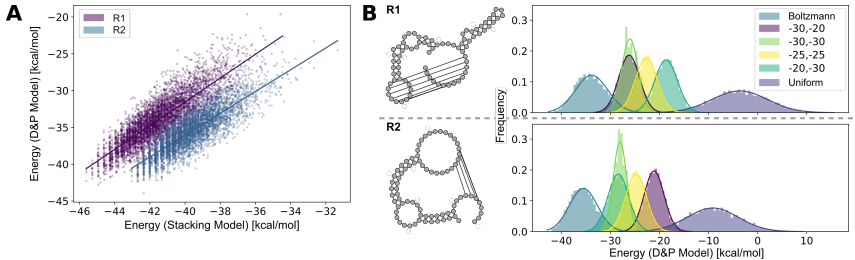
Multidimensional Boltzmann sampling [Bodini, P, DMTCS 2011]

Input: Targeted free-energies $(E^\star_\ell)^k_{\ell=1}$, weights $(x_\ell)^k_{\ell=1}$ such that $\mathbb{E}(E(w, S_\ell)) = E^\star_\ell, \forall \ell$:

$$\mathbb{P}(w \mid x_1 \cdots x_k) \sim \prod_{\ell=1}^{k} x_\ell^{E(w,S_\ell)} \text{ + Efficient rejection} \rightarrow \mathcal{O}(n^{k/2}) \text{ exact/} \mathcal{O}(\alpha^k) \text{ approx.}$$

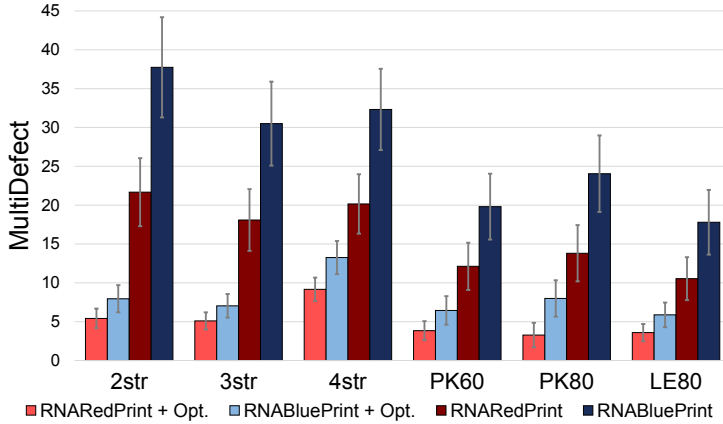
Empirical efficiency for additive *concentrated* constraints (GC%, dinucleotides ...)
$\rightarrow$ Partial functions $\rightarrow$ Hyper-edges, *aka* cliques[1]

To Do General framework for integer-valued constraints; Concentration tests.



[1] But tree width ↗

Strangely enough, it actually works!



$$\text{MultiDefect}(S, R_1 \cdots R_k) := \frac{\sum_{\ell=1}^{k} E(S, R_\ell) - \text{EFE}(S)}{k} + \frac{\sum_{1 \le \ell < j \le k} |E(S, R_\ell) - E(S, R_j)|}{2\binom{k}{2}}$$