Selected combinatorial problems in RNA Bioinformatics

...and some solutions

Yann Ponty*,•,†

+ Many collaborators

* Recently back from Simon Fraser University/PIMS, Vancouver, Canada

• LIX, CNRS/Ecole Polytechnique

† Amib project-team, Inria Saclay
Fundamental *dogma* of molecular biology

DNA
{A, C, G, T}*

![DNA structure and sequence](image)
Fundamental *dogma* of molecular biology

DNA
\{A, C, G, T\}*

Pol

RNAs
\{A, C, G, U\}*

THE CODE (genes)
THE MACHINE (enzymes)

\[
\begin{array}{cccccccc}
A & T & G & G & T & T & A & C \\
T & A & C & C & A & A & T & G \\
\end{array}
\]
Fundamental *dogma* of molecular biology

DNA
\{A, C, G, T\}*

RNA
\{A, C, G, U\}*

THE CODE (genes)
THE MACHINE (enzymes)

---

Yann Ponty (CNRS/Polytechnique)
Combinatorial problems in RNA Bioinformatics
14/03/2016 – LAMSADE Seminar
Fundamental *dogma* of molecular biology

- **DNA** \{A, C, G, T\}*
- **RNAs** \{A, C, G, U\}*

\[
\begin{align*}
\text{DNA} & \rightarrow \text{RNAs} \\
\{A, C, G, T\}^* & \rightarrow \{A, C, G, U\}^*
\end{align*}
\]
Fundamental *dogma* of molecular biology

- DNA
  \{A, C, G, T\}^*

- RNAs
  \{A, C, G, U\}^*

---

**THE CODE** (genes)

**THE MACHINE** (enzymes)

---

Pol

---

Yann Ponty (CNRS/Polytechnique)

Combinatorial problems in RNA Bioinformatics

14/03/2016 – LAMSADE Seminar
Fundamental *dogma* of molecular biology

DNA
\{A, C, G, T\}^* 

RNAs
\{A, C, G, U\}^*
**Fundamental *dogma* of molecular biology**

DNA
\{A, C, G, T\}*

RNAs
\{A, C, G, U\}*

ATGGTTACCCAT

TACCAATTGGGTATA

AUGGUUACCCAU
Fundamental *dogma* of molecular biology

- **DNA** \{A, C, G, T\}*
- **RNAs** \{A, C, G, U\}*
- **Proteins** \{Ala, Arg, . . . , Val\}*

\[20^+\] Amino acids
Fundamental *dogma* of molecular biology

**DNA**
\{A, C, G, T\}*

**RNAs**
\{A, C, G, U\}*

**Proteins**
\{Ala, Arg, . . . , Val\}*

20+ Amino acids
Fundamental *dogma* of molecular biology

**DNA**
{A, C, G, T}*

**RNAs**
{A, C, G, U}*

**Proteins**
{Ala, Arg, . . ., Val}*

20+ Amino acids

20+ Amino acids

ATGGTTATACCCCAT

TACCAATGGGTA

AUGGUUAACCCCAU

Ribosome

Met Val
**Fundamental dogma of molecular biology**

DNA
\{A, C, G, T\}*

RNAs
\{A, C, G, U\}*

Proteins
\{Ala, Arg, . . . , Val\}*

20+ Amino acids

**THE CODE** (genes)

**THE MACHINE** (enzymes)

Ribosome

\begin{align*}
  &\text{Met} \\
  &\text{Val} \\
  &\text{Thr}
\end{align*}
**Fundamental dogma of molecular biology**

- **DNA** \{A, C, G, T\}*
- **RNAs** \{A, C, G, U\}*
- **Proteins** \{Ala, Arg, . . . , Val\}*

20+ Amino acids

\[\text{ATGGTTATCCCAT} \]
\[\text{TAACCAATGGGTA} \]

Ribosome

\[\text{AUGGUUAACCCAU} \]

Met Val Thr His

Yann Ponty (CNRS/Polytechnique)  Combinatorial problems in RNA Bioinformatics  14/03/2016 – LAMSADE Seminar
Fundamental *dogma* of molecular biology

- **DNA**
  \{A, C, G, T\}*

- **RNAs**
  \{A, C, G, U\}*

- **Proteins**
  \{Ala, Arg, . . . , Val\}*

20+ Amino acids
Fundamental *dogma* of molecular biology

**THE CODE**
- (genes)
- DNA
  \{A, C, G, T\}*

**RNAs**
\{A, C, G, U\}*

**Proteins**
\{Ala, Arg, . . . , Val\}*

20+ Amino acids

### THE CODE
ATGGTTACCCAT

### DNA
TACCAATGGGTA

### RNAs
AUGGUUAACCCAU

### Proteins
Met Val Thr His Ile Leu His Asn
Fundamental *dogma* of molecular biology

**THE CODE** (genes)

DNA
\{A, C, G, T\}*

**THE MACHINE** (enzymes)

Proteins
\{Ala, Arg, ..., Val\}*

20+ Amino acids

**RNAs**

\{A, C, G, U\}*

\[\text{ATGGTTACCCCAT}\]

\[\text{TACCAATGGGTA}\]

\[\text{AUGGUAACCCAU}\]

\[\text{Met Val Thr His Ile Leu His Asn}\]
Fundamental *dogma* of molecular biology

**THE CODE** *(genes)*
DNA
{A, C, G, T}*

**MEH...**
RNAs
{A, C, G, U}*

**THE MACHINE** *(enzymes)*
Proteins
{Ala, Arg, ..., Val}*
20+ Amino acids

The code:
- DNA: {A, C, G, T}*
- RNAs: {A, C, G, U}*
- Proteins: {Ala, Arg, ..., Val}*

The machine:

- **THE CODE** 
- **MEH...** 
- **THE MACHINE**

Amino acids:
- Met
- Val
- Thr
- His
- Ile
- Leu
- His
- Asn
Fundamental *dogma* of molecular biology
Fundamental *dogma* of molecular biology
Fundamental *dogma* of molecular biology

RNA functions
- Messenger
- Translation
- Regulation
- Enzyme
- Catalytic
- …
RNA world: Resolving the *chicken vs egg* paradox at the origin of life...

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

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

**R. Dawkins.** *The Ancestor’s Tale: A Pilgrimage to the Dawn of Evolution*
RNA world: Resolving the *chicken vs egg* paradox at the origin of life.

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

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

**R. Dawkins.** *The Ancestor’s Tale: A Pilgrimage to the Dawn of Evolution*
**RNA structure(s)**

**RNA** = Linear Polymer = Sequence in \{A, C, G, U\}*

```
UUAGGCAGCAGC
GGUGGGGUUGCCUCC
CGUACCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

**Primary structure**  **Secondary structure**  **Tertiary structure**

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

---

**Definition (Secondary Structure)**

A secondary structure \( S \) for an RNA \( w \) is a set of base-pairs \((i, j) \in [1, n]^2\) such that:

- **Monogamy:** Each position \( x \in [1, n] \) involved in at most one base-pair;
- **Non-crossing base-pairs:** \( \forall (i, j), (k, l) \in S \) such that \( i < k < j < l \);
- **Steric constraints:** \( \forall (i, j), \) one has \( i < j \) and \( j - i > \theta \) (where \( \theta := 1 \) typically).
**RNA structure(s)**

**RNA** = Linear Polymer = Sequence in \{A, C, G, U\}*

UUAGGGCCGCCAACAAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAAGUAAGCC
CACCAGCGUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Primary structure  Secondary structure  Tertiary structure

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

**Definition (Secondary Structure)**

A **secondary structure** $S$ for an RNA $w$ is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- **Monogamy**: Each position $x \in [1, n]$ involved in at most one base-pair;
- **Non-crossing base-pairs**: $(i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints**: $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).
RNA structure(s)

**RNA** = Linear Polymer = Sequence in \{A, C, G, U\}*

- UUAGGCGGCCAAGC
- GGUGGGGUUGCCUCC
- CGUACCCAUCCCGAA
- CACGGAAGAUAAGCC
- CACCAGCGUUCCGGG
- GAGUACUGGAGUGCG
- CGAGCCUCUGGGAAA
- CCCCGGUUCGCCGCCA
- CC

**Primary structure**

**Secondary structure**

**Tertiary structure**

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

**Definition (Secondary Structure)**

A secondary structure \(S\) for an RNA \(w\) is a set of base-pairs \((i, j) \in [1, n]^2\) such that:

- **Monogamy:** Each position \(x \in [1, n]\) involved in at most one base-pair;
- **Non-crossing base-pairs:** \(\not\exists (i, j), (k, l) \in S\) such that \(i < k < j < l\);
- **Steric constraints:** \(\forall (i, j), \) one has \(i < j\) and \(j - i > \theta\) (where \(\theta := 1\) typically).
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected

Supporting intuitions

Different representations
Common combinatorial structure

* Additional steric constraints
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, \( \Delta(G) \leq 3 \), 2-connected*

Supporting intuitions

Different representations
Common combinatorial structure

* Additional steric constraints

Dot plots
Adjacency matrices*
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

Supporting intuitions

Different representations
Common combinatorial structure

* Additional steric constraints
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected

Motzkin words*

Supporting intuitions
Different representations
Common combinatorial structure
* Additional steric constraints
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

Motzkin words*

Non-crossing arc-annotated sequences*

Supporting intuitions

Different representations
Common combinatorial structure

* Additional steric constraints
Various representations for a versatile biomolecule

Outer-planar graphs
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

Motzkin words*

Positive 1D meanders* over $S = \{+1, -1, 0\}$

Non-crossing arc-annotated sequences*

Supporting intuitions
Different representations
Common combinatorial structure

* Additional steric constraints

Dot plots
Adjacency matrices*

Non-crossing arc diagrams*
Part. I: Predicting how RNA folds
**Thermodynamics view**

At the **nanoscale**, RNA folding can be adequately viewed as a **Markov process**, whose **stationary distribution** is the **Boltzmann distribution**.

**Definition (Thermodynamic equilibrium)**

Each structure $S$ **compatible** with an RNA $w$ observed with probability:

$$ P(S \mid w) = \frac{e^{-E_w(S) / kT}}{\mathcal{Z}_w} $$

and

$$ \mathcal{Z}_w \equiv \sum_{S'} e^{-E_w(S') / RT} \{\text{Partition function}\} $$

$E_w(S)$: **free-energy** of $S$ over $w$; $R$: Boltzmann constant; and $T$: temperature.
Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) representative of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

mRNA half-life: $\sim 7h$ (Mouse [Sharova2009])

$T \rightarrow \infty$
Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) representative of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

mRNA half-life: $\sim 7h$ (Mouse [Sharova2009])
**Paradigms for RNA structure prediction**

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

\[ T = 0h \]

mRNA half-life: \(~7h\)  
(Mouse [Sharova2009])
Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

$mRNA$ half-life: $\sim 7h$

(Mouse [Sharova2009])
Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

mRNA half-life: $\sim 7h$

(Mouse [Sharova2009])

$T = 2h$
Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) representative of the Boltzmann ensemble
- **2010s–?????** Embracing kinetics

mRNA half-life: $\sim 7$ h (Mouse [Sharova2009])
Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–????** Embracing kinetics

mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

$T = 10h$
Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–????** Embracing kinetics

$mRNA$ half-life: $\sim 7h$ (Mouse [Sharova2009])
Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) representative of the Boltzmann ensemble
- **2010s–????** Embracing kinetics

mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

$T = 10h$
Problem statement

▶ RNA structure $S$: (Partial) matching of positions in sequence $w$

▶ Motifs: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)

▶ Energy model:
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$
Problem statement

- **RNA structure** $S$: (Partial) matching of positions in sequence $w$
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops . . .)
- **Energy model**:
  - Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  - Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$
Problem statement

- **RNA structure** $S$: (Partial) matching of positions in sequence $w$
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model**:
  - Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  - Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$
Problem statement

- **RNA structure** $S$: (Partial) matching of positions in sequence $w$
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model**:
  - Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  - Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$
Problem statement

- **RNA structure** $S$: (Partial) matching of positions in sequence $w$
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model**:
  - **Motif** → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  - **Free-Energy** $E_w(S)$: Sum over (independently contributing) motifs in $S$

\[
E_S = 2 \cdot \Delta \begin{pmatrix} U \\ G \end{pmatrix} + 4 \cdot \Delta \begin{pmatrix} G \\ C \end{pmatrix} + 2 \cdot \Delta \begin{pmatrix} C \\ G \end{pmatrix}
\]
Problem statement

▸ RNA structure $S$: (Partial) matching of positions in sequence $w$
▸ Motifs: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops . . .)
▸ Energy model:
  • Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  • Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$

$$E_S = \Delta\begin{pmatrix}C & G \\ G & C \end{pmatrix} + \Delta\begin{pmatrix}G & G \\ C & C \end{pmatrix} + \Delta\begin{pmatrix}U & G \\ G & C \end{pmatrix} + \Delta\begin{pmatrix}U & G \\ G & C \end{pmatrix} + \Delta\begin{pmatrix}U & G \\ G & C \end{pmatrix}$$
Problem statement

- **RNA structure** $S$: (Partial) matching of positions in sequence $w$
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model**:  
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$  
  Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in $S$

$$E_S = \Delta \begin{pmatrix} C \ G \ C \ G \\ G \ C \ C \ G \end{pmatrix} + \Delta \begin{pmatrix} G \ C \ G \ C \\ C \ G \ C \ C \end{pmatrix} + \Delta \begin{pmatrix} U \ G \ C \ C \\ G \ C \ C \ G \end{pmatrix} + \Delta \begin{pmatrix} C \ U \ A \ C \\ G \ G \ U \ G \end{pmatrix} + \Delta \begin{pmatrix} C \ G \ A \ G \\ A \ C \ G \ C \end{pmatrix}$$
Problem statement

- **RNA structure** \( S \): (Partial) matching of positions in sequence \( w \)
- **Motifs**: Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model**:
  - Motif \( \rightarrow \) Free-energy contribution \( \Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\} \)
  - Free-Energy \( E_w(S) \): Sum over (independently contributing) motifs in \( S \)

**Definition (MFE-PREDICT(\( E \)) problem)**

- **Input**: RNA sequence \( w \in \{A, C, G, U\}^* \).
- **Output**: (Constrained) matching \( S^* \) of Minimal Free-Energy \( E_w(S^*) \).
RNA folding: non-crossing matchings

**RNA** = Linear Polymer = Sequence in \{A, C, G, U\}*

**Structure** = Non-crossing matching

```
UUAGCGGCCACAGC
GGUGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
```

**MFE folding prediction:** $\mathcal{O}(n^3)$

```
Primary Structure       Secondary Structure       Tertiary Structure
```

5s rRNA (PDBID: 1K73:B)
Dynamic programming (DP) for RNA folding

Theorem (NussinovJacobson1980 + ZukerStiegler80)

Max #base-pairs/min weight/minimum free-energy structure can be solved in $O(n^3)/O(n^2)$ time/memory using dynamic programming

$$E_{i,k}: \text{Free-energy contribution of base-pair (i, k).}$$

$$N_{i,j}: \text{Max #base-pairs over interval [i, j]}$$

$$N_{i,t} = 0, \quad \forall t \in [i, i + \theta]$$

$$N_{i,j} = \min \left\{ \begin{array}{ll}
N_{i+1,j} & \{i \text{ unpaired}\} \\
\min_{k=i+\theta+1}^{j} E_{i,k} + N_{i+1,k-1} + N_{k+1,j} & \{i \text{ paired to } k\}
\end{array} \right.$$
Dynamic programming (DP) for RNA folding

**Theorem (NussinovJacobson1980 + ZukerStiegler80)**

Max #base-pairs/min weight/minimum free-energy structure can be solved in $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/memory using dynamic programming

\[
E_{i,k} : \text{Free-energy contribution of base-pair } (i, k). \quad (-1/ + \infty \text{ or } \Delta G(s_i \equiv s_k))
\]

\[
C_{i,t} = 1, \quad \forall t \in [i, i + \theta]
\]

\[
C_{i,j} = \sum \left\{ \sum_{k=i+\theta+1}^{j} 1_{\text{comp.}(i,k)} \times C_{i+1,k-1} \times C_{k+1,j} \right\}
\]

\{i unpaired\}

\{i paired to k\}

\[
C_{i,j} = \sum \left\{ \sum_{k=i+\theta+1}^{j} 1_{\text{comp.}(i,k)} \times C_{i+1,k-1} \times C_{k+1,j} \right\}
\]

\{i unpaired\}

\{i paired to k\}
Dynamic programming (DP) for RNA folding

**Theorem (Nussinov-Jacobson1980 + Zuker-Stiegler80)**

Max #base-pairs/min weight/minimum free-energy structure can be solved in $O(n^3) / O(n^2)$ time/memory using dynamic programming

\[ i \quad \quad \quad \quad \quad \quad j = \quad \quad \quad \quad \quad \quad i \quad i+1 \quad \quad \quad \quad \quad \quad j + \quad \quad \quad \quad \quad \quad i \quad k \quad \quad \quad \quad \quad \quad j \geq \theta \]

$E_{i,k}$: Free-energy contribution of base-pair $(i, k)$. $(-1/ + \infty$ or $\Delta G(s_i \equiv s_k)$)

\[ Z_{i,j} = \sum_{S \text{ comp. with } w_{[i,j]}} e^{\frac{-E_W(S)}{RT}} = \text{Partition function of structures compatible with interval } [i,j] \]

\[ Z_{i,t} = 1, \quad \forall t \in [i, i+\theta] \]

\[ Z_{i,j} = \sum \left\{ \begin{array}{ll}
\sum_{k=i+\theta+1}^{j} e^{\frac{-E_{i,k}}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} & \{i \text{ unpaired}\} \\
Z_{i+1,j} & \{i \text{ paired to } k\} 
\end{array} \right. \]
Dynamic programming (DP) for RNA folding

Many extensions:

- Comparative folding [Sankoff1985]
- Equilibrium base-pairing probabilities [McCaskill1990]
- Moments of additive features [Miklos2005, Ponty2011]
- \( \Delta \text{kcal.mol}^{-1} \) suboptimal structures of MFE [Wuchty1999]
- Basic crossing structures [Rivas1999]
- Exact sampling in Boltzmann distr. [Ding2003, Ponty2008]
- Moments of additive features [Miklos2005, Ponty2011]
- Maximum expected accuracy structure [Do2006]
- Distance-classified partitioning of Boltzmann ens. [E.Freyhult2007a]

Made possible by:

- Completeness/Unambiguity of decomposition
  - \( \exists \) energy-preserving bijection between derivations of DP scheme and search space
- Objective function additive with respect to DP scheme
  \( \Rightarrow \) Combinatorial Dynamic Programming

\( E_{i,k} \): Free-energy contribution of base-pair \((i,k)\).

\( Z_{i,j} = \sum_{s_{\text{comp}}} e^{-\frac{E_{(s)}}{RT}} \): Partition function of structures compatible with interval \([i,j]\).
Including crossing interactions

- **Non-canonical base-pairs**: Lead to local crossings and promiscuity
  Any base-pair **other than** \{(A-U), (C-G), (G-U)\}
  OR interacting in a non-standard way (WC/WC-Cis) [Leontis2001].

- **Pseudoknots**: Crossing sets of nested stable base-pairs

![Canonical CG base-pair (WC/WC-Cis)](image1)
![Non-canonical base-pair (Sugar/WC-Trans)](image2)

![Group I Ribozyme (PDBID: 1Y0Q:A)](image3)
Including crossing interactions

- **Non-canonical base-pairs:** Lead to local crossings and promiscuity
  - Any base-pair other than \{(A-U), (C-G), (G-U)\}
  - OR interacting in a non-standard way (WC/WC-Cis) [Leontis2001].

- **Pseudoknots:** Crossing sets of nested stable base-pairs
Including crossing interactions

- **Non-canonical base-pairs:** Lead to local crossings and promiscuity
- Any base-pair other than \{(A-U), (C-G), (G-U)\} OR interacting in a non-standard way (WC/WC-Cis) \[\text{Leontis2001}\]

Crossing interactions, once ignored, are now **ubiquitous**!

**Example:** Group II Intron (PDB ID: 3IGI)
Energy models

Three models, based on interacting positions \((i, j)\):

- **Base-pair model** \(\mathcal{B}\): Nucleotides \((w_i, w_j)\) at \((i, j)\)
  \[ \rightarrow \Delta_{\mathcal{B}}(w_i, w_j) \]

- **Nearest-neighbor model** \(\mathcal{N}\): Nucl. at \((i, j)\) and \((i+1, j-1)\) + partners (or \(\emptyset\))
  \[ \rightarrow \Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}}) \]

- **Stacking pairs model** \(\mathcal{S}\): Nucl. at \((i, j)\) and \((i+1, j-1)\) **only if** latter paired
  \[ \rightarrow \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1}) \]

Solved in \(O(n^3)\) [Tabaska1998] (Max-weighted matching)
**Unrealistic!**
Energy models

Three models, based on interacting positions \((i,j)\):

- **Base-pair model** \(\mathcal{B}**: Nucleotides \((w_i, w_j)\) at \((i,j)\)
  \[\Delta_{\mathcal{B}}(w_i, w_j)\]

- **Nearest-neighbor model** \(\mathcal{N}**: Nucl. at \((i,j)\) and \((i+1,j-1)\) + partners (or \(\emptyset\))
  \[\Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}})\]

- **Stacking pairs model** \(\mathcal{S}**: Nucl. at \((i,j)\) and \((i+1,j-1)\) **only** if latter paired
  \[\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})\]

Too expressive?
Energy models

Three models, based on interacting positions \((i, j)\):

- **Base-pair model** \(\mathcal{B}\): Nucleotides \((w_i, w_j)\) at \((i, j)\)
  \[ \Delta_{\mathcal{B}}(w_i, w_j) \]

- **Nearest-neighbor model** \(\mathcal{N}\): Nucl. at \((i, j)\) and \((i+1, j-1)\) + partners (or \(\emptyset\))
  \[ \Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}}) \]

- **Stacking pairs model** \(\mathcal{S}\): Nucl. at \((i, j)\) and \((i+1, j-1)\) **only if** latter paired
  \[ \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1}) \]

Captures stablest motifs
Still NP-hard [Lyngso2004]
… but PTAS [Lyngso2004]
## The full monty

<table>
<thead>
<tr>
<th></th>
<th>Base-pairs</th>
<th>Stacking-Pairs</th>
<th>Nearest-Neighbor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comp.</strong></td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>[Nussinov1980]</td>
<td>[leong2003]</td>
<td>[Zuker1981]</td>
<td></td>
</tr>
<tr>
<td><strong>Non-crossing</strong></td>
<td>Approx.</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Comp.</strong></td>
<td>???</td>
<td>NP-Hard</td>
<td>NP-Hard</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[leong2003]</td>
<td>[leong2003]</td>
</tr>
<tr>
<td><strong>Planar</strong></td>
<td>Approx.</td>
<td>2-approx.</td>
<td>2-approx.</td>
</tr>
<tr>
<td></td>
<td>≈[leong2003]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comp.</strong></td>
<td>P</td>
<td>NP-Hard</td>
<td>NP-Hard</td>
</tr>
<tr>
<td>[Tabaska1998]</td>
<td>[Lyngso2004]</td>
<td>[Lyngso2000]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(any* Δ model)</td>
<td>[Akutsu2000]</td>
<td></td>
</tr>
<tr>
<td><strong>General</strong></td>
<td>Approx.</td>
<td>Duh...</td>
<td>APX-Hard</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Sheikh2012]</td>
</tr>
</tbody>
</table>

### Missing:
- Base-pair maximization in planar model (probably NP-hard)
- Partition function (mostly in P cases), Boltzmann-Gibbs sampling
- **Relevance** of approximation???

**Rem.:** Exact polynomial algorithms for restricted Pseudoknots [PontySaule2011]
Part. II: Finding RNAs in genomes
Sequence structure alignment for ncRNA search and homology-modeling

Search for novel ncRNA instances

Structure prediction by homology modelling
Sequence structure alignment for ncRNA search and homology-modeling

Search for novel ncRNA instances

Structure prediction by homology modelling
Context: Multiple Structural levels

Primary Structure

- Represents nucleotides sequence
- No interaction

Boring...
Secondary Structure

- Scaffold/blueprint for 3D
- Only includes non-crossing canonical interactions (WC/WC cis, GC/AU/GU)
- Any nucleotide has \( \leq 1 \) partner
Pseudoknots play a major part in the architecture of some RNAs. Yet they are hard to handle algorithmically!
Context: Multiple Structural levels

Extended secondary structure

- Captures any interaction (canonical and non-canonical)
- Possibly, multiple partners per position

Now we’re talking!
Sequence-structure alignment

A UUCAAGG U
21 3 4 5 6 7 8 9
A UGAACC U
2'1' 3' 4' 5' 6' 7' 8'

Yann Ponty (CNRS/Polytechnique)
Combinatorial problems in RNA Bioinformatics
14/03/2016 – LAMSADE Seminar
Sequence-structure alignment
Sequence-structure alignment
Sequence-structure alignment

Yann Ponty (CNRS/Polytechnique)
Sequence-structure alignment

A UUCAAGG U
21 3 4 5 6 7 8 9
A UGAACC U
2'1' 3' 4' 5' 6' 7' 8'

Yann Ponty (CNRS/Polytechnique)
Combinatorial problems in RNA Bioinformatics
14/03/2016 – LAMSADE Seminar
Sequence-structure alignment Problem

**Input:** (Extended) Secondary structure $S$ + Sequence $\omega$

**Output:** Minimal-cost alignment (mapping subject to constraints)

**Variant:** Affine gap cost model
Sequence-structure alignment Problem

**Input:** (Extended) Secondary structure $S$ + Sequence $\omega$

**Output:** Minimal-cost alignment (mapping subject to constraints)

**Variant:** Affine gap cost model
### Complexity of structure-sequence alignment

\[ n = \text{Structure Length}, \; m = \text{Sequence Length} \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary Structure – Sequence</td>
<td>( O(n \cdot m^3) )</td>
</tr>
<tr>
<td>Pseudoknots – Sequence</td>
<td>MAX-SNP-Hard</td>
</tr>
<tr>
<td>Extended Secondary Structure – Sequence</td>
<td>MAX-SNP-Hard</td>
</tr>
</tbody>
</table>

Jiang et al. 2001
Complexity of structure-sequence alignment

\[ n = \text{Structure Length}, \ m = \text{Sequence Length} \]

<table>
<thead>
<tr>
<th>Secondary structure – Sequence</th>
<th>( O(n \cdot m^3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudoknots – Sequence</td>
<td>MAX-SNP-Hard</td>
</tr>
<tr>
<td>Extended Secondary Structure – Sequence</td>
<td>MAX-SNP-Hard</td>
</tr>
</tbody>
</table>

Jiang et al. 2001
Complexity of structure-sequence alignment

\[ n = \text{Structure Length}, \ m = \text{Sequence Length} \]

<table>
<thead>
<tr>
<th></th>
<th>(O(n \cdot m^3))</th>
<th>(\text{MAX-SNP-Hard})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary Structure – Sequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pseudoknots – Sequence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended Secondary Structure – Sequence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Jiang *et al.* 2001
Complexity of structure-sequence alignment

\[ n = \text{Structure Length}, \ m = \text{Sequence Length} \]

| Secondary Structure – Sequence | \( O(n \cdot m^3) \) |
| Pseudoknots – Sequence          | MAX-SNP-Hard          |
| **Extended Secondary Structure – Sequence** | MAX-SNP-Hard          |

Jiang et al. 2001
Complexity of struct.-seq. alignment: Polynomial classes

\[ n = \text{Structure Length}, \quad m = \text{Sequence Length}, \quad b = \#\text{Bands} \]

<table>
<thead>
<tr>
<th>Standard Pseudoknots</th>
<th>( O(n \cdot m^b) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Embedded Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Simple Non-standard Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Standard Triple Helices</td>
<td>( O(n \cdot m^3) )</td>
</tr>
</tbody>
</table>

Han et al. 2008
Complexity of struct.-seq. alignment: Polynomial classes

\[ n = \text{Structure Length}, \ m = \text{Sequence Length}, \ b = \text{#Bands} \]

<table>
<thead>
<tr>
<th>Pseudoknot Type</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Pseudoknots</td>
<td>( O(n \cdot m^b) )</td>
</tr>
<tr>
<td><strong>Standard Embedded Pseudoknots</strong></td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Simple Non-standard Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Standard Triple Helices</td>
<td>( O(n \cdot m^3) )</td>
</tr>
</tbody>
</table>

Graphical representations of pseudoknots and triple helices are shown in the image.

Han et al. 2008
Complexity of struct.-seq. alignment: Polynomial classes

\[ n = \text{Structure Length}, \ m = \text{Sequence Length}, \ b = \#\text{Bands} \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard Pseudoknots</strong></td>
<td>(O(n \cdot m^b))</td>
</tr>
<tr>
<td><strong>Standard Embedded Pseudoknots</strong></td>
<td>(O(n \cdot m^{b+1}))</td>
</tr>
<tr>
<td><strong>Simple Non-standard Pseudoknots</strong></td>
<td>(O(n \cdot m^{b+1}))</td>
</tr>
<tr>
<td><strong>Standard Triple Helices</strong></td>
<td>(O(n \cdot m^3))</td>
</tr>
</tbody>
</table>

Wong et al. 2011
Complexity of struct.-seq. alignment: Polynomial classes

\[ n = \text{Structure Length}, \ m = \text{Sequence Length}, \ b = \#\text{Bands} \]

<table>
<thead>
<tr>
<th></th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Pseudoknots</td>
<td>( O(n \cdot m^b) )</td>
</tr>
<tr>
<td>Standard Embedded Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Simple Non-standard Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td><strong>Standard Triple Helices</strong></td>
<td>( O(n \cdot m^3) )</td>
</tr>
</tbody>
</table>

Wong et al. 2012
Complexity of struct.-seq. alignment: Polynomial classes

\[ n = \text{Structure Length, } m = \text{Sequence Length, } b = \#\text{Bands} \]

<table>
<thead>
<tr>
<th>Type of Pseudoknots</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Pseudoknots</td>
<td>( O(n \cdot m^b) )</td>
</tr>
<tr>
<td>Standard Embedded Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Simple Non-standard Pseudoknots</td>
<td>( O(n \cdot m^{b+1}) )</td>
</tr>
<tr>
<td>Standard Triple Helices</td>
<td>( O(n \cdot m^3) )</td>
</tr>
</tbody>
</table>

+ Other \( O(n.m^4)/O(n.m^6) \) classes based on folding DP schemes

[Möhl/Will/Backofen 2009]
Outline of general parameterized approach

Structure

Sequence

Tree-decomposition

Alignment

[Rinaudo, Ponty, Barth, Denise, WABI 2012]
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment $\Rightarrow$ Constraints

- Adjacent positions in structure $\Rightarrow$ Precedence
- Paired positions $\Rightarrow$ Both partners needed to assign score

Sets of structure-side positions (bags $\{B_i\}$), in a tree such that:

- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in $\geq 1$ bag
- If $x \in B \cap B'$, than $x$ is in every bag $B''$ on the path from $B$ to $B'$. 

![Diagram of RNA structure and tree decomposition](Image)
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment $\Rightarrow$ Constraints

- Adjacent positions in structure $\Rightarrow$ Precedence
- Paired positions $\Rightarrow$ Both partners needed to assign score

Sets of structure-side positions (bags $\{B_i\}$), in a tree such that:
- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in $\geq 1$ bag
- If $x \in B \cap B'$, then $x$ is in every bag $B''$ on the path from $B$ to $B'$
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment $\Rightarrow$ Constraints

- Adjacent positions in structure $\Rightarrow$ Precedence
- Paired positions $\Rightarrow$ Both partners needed to assign score

Sets of structure-side positions (bags $\{B_i\}$), in a tree such that:
- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in $\geq 1$ bag
- If $x \in B \cap B'$, than $x$ is in every bag $B''$ on the path from $B$ to $B'$
Structure-centric alignment \(\Rightarrow\) Constraints

- Adjacent positions in structure \(\Rightarrow\) Precedence
- Paired positions \(\Rightarrow\) Both partners needed to assign score

Sets of structure-side positions (bags \(\{B_i\}\)), in a tree such that:

- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in \(\geq 1\) bag
- If \(x \in B \cap B'\), then \(x\) is in every bag \(B''\) on the path from \(B\) to \(B'\)

```
AGGAACCUUAAAC
1 2 3 4 5 6 7 8 9 10 11
1, 2, 11
2, 10, 11
2, 3, 10
3, 6, 10
3, 4, 6, 5
4, 5, 6
6, 7, 10
7, 9, 10
7, 8, 9
```
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment $\Rightarrow$ Constraints

- Adjacent positions in structure $\Rightarrow$ Precedence
- Paired positions $\Rightarrow$ Both partners needed to assign score

Sets of structure-side positions (bags $\{B_i\}$), in a tree such that:

- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in $\geq 1$ bag
- If $x \in B \cap B'$, than $x$ is in every bag $B''$ on the path from $B$ to $B'$

Yann Ponty (CNRS/Polytechnique) Combinatorial problems in RNA Bioinformatics 14/03/2016 – LAMSADE Seminar
Structure-centric alignment \implies Constraints

- Adjacent positions in structure
  \implies Precedence
- Paired positions
  \implies Both partners needed to assign score

Sets of structure-side positions (bags \{B_i\}), in a tree such that:
- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in \( \geq 1 \) bag
- If \( x \in B \cap B' \), then \( x \) is in every bag \( B'' \) on the path from \( B \) to \( B' \)
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment $\Rightarrow$ Constraints

- Adjacent positions in structure $\Rightarrow$ Precedence
- Paired positions $\Rightarrow$ Both partners needed to assign score

Sets of structure-side positions (bags $\{B_i\}$), in a tree such that:

- Every position in the structure appears at least once
- Each interacting pair of positions simultaneously appear in $\geq 1$ bag
- If $x \in B \cap B'$, than $x$ is in every bag $B''$ on the path from $B$ to $B'$

![Diagram showing tree decomposition]

Width $k =$ Size of biggest bag minus one.
Tree-Decomposition-based Alignment

Yann Ponty (CNRS/Polytechnique)
Tree-Decomposition-based Alignment

\[ \begin{align*}
1, 2, 11 \\
2, 10, 11 \\
2, 3, 10 \\
3, 6, 10 \\
3, 4, 6 \\
4, 5, 6 \\
6, 7, 10 \\
7, 9, 10 \\
7, 8, 9
\end{align*} \]

\[ \begin{align*}
\text{G UUGGACAG C} \\
\text{2' 1' 3' 4' 5' 6' 7' 8' 9' 10'}
\end{align*} \]
(Fixed-parameter tractable?) algorithm [Rinaudo et al. 2012]

**Theorem**

**Input:** Structure $S$ of length $n$; Sequence $w$ of length $m$ → Tree dec. of $S$, width $k$

Best alignment computed in $O(n.m^{k+1})/O(n.m^k)$ time/space → not FPT!

**Dynamic programming equation:**

$$
\text{Cost}(l, f) = \min_{f'=(\mu', \delta') \in \mathcal{F}|X_l}
\left\{ \phi(X_l, f') + \sum_{s \text{ child of } l} \text{Cost}(s, f'|X_s, l) \right\},
$$

where $\phi(X_l, f')$ : local cost contribution of alignment $f'$ to a bag $X_l$

**Algorithm:** Depth-first order, Compute/Memorize Cost (+Best assignment)

**Bonus:**

- Free extension to affine gaps cost models;
- Time complexity reduced to $\Theta(n.m^k)$ for smooth tree-decompositions. (Smooth = Proper index of a bag replaces a neighboring index in the parent bag)
Specialized complexities

For previous classes of biologically-relevant structures, our algorithm has **equal or better** complexities than *ad hoc* algorithms.

<table>
<thead>
<tr>
<th>Class of Structures</th>
<th>Time comp.</th>
<th>Multiple interactions</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recursive Classical Structures</td>
<td>$O(n \cdot m^{k+2})$</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Secondary Structures (Pseudoknot-free)</td>
<td>$O(n \cdot m^3)$</td>
<td></td>
<td>[Jiang et al 02]</td>
</tr>
<tr>
<td>Embedded Standard Pseudoknots</td>
<td>$O(n \cdot m^{k+1})$</td>
<td></td>
<td>[Han et al 08]</td>
</tr>
<tr>
<td>Standard Structures</td>
<td>$O(n \cdot m^k)$</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Standard Pseudoknots</td>
<td>$O(n \cdot m^k)$</td>
<td></td>
<td>[Han et al 08]</td>
</tr>
<tr>
<td>2-Level Recursive Simple Non-Standard PKs</td>
<td>$O(n \cdot m^{k+2})$</td>
<td></td>
<td>[Wong et al 11]</td>
</tr>
<tr>
<td>Simple Non-Standard Structures</td>
<td>$O(n \cdot m^{k+1})$</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Simple Non-Standard Pseudoknots</td>
<td>$O(n \cdot m^{k+1})$</td>
<td></td>
<td>[Wong et al 11]</td>
</tr>
<tr>
<td>Extended Triple Helices</td>
<td>$O(n \cdot m^3)$</td>
<td>√</td>
<td>[Wong et al 12]</td>
</tr>
<tr>
<td>Triple Helices</td>
<td>$O(n \cdot m^3)$</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

$n \rightarrow$ Structure length  
$m \rightarrow$ Sequence length  
$k \rightarrow$ Class-specific structural parameter
Tree Decomposition vs The World [Rinaudo et al. 2012]

Specialized complexities

For previous classes of biologically-relevant structures, our algorithm has equal or better complexities than ad hoc algorithms.

<table>
<thead>
<tr>
<th>Class of Structures</th>
<th>Time comp.</th>
<th>Multiple interactions</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recursive Classical Structures</td>
<td>$O(n \cdot m^{k+2})$</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Secondary Structures (Pseudoknot-free)</td>
<td>$O(n \cdot m^3)$</td>
<td>✓</td>
<td>[Jiang et al 02]</td>
</tr>
<tr>
<td>Embedded Standard Pseudoknots</td>
<td>$O(n \cdot m^{k+1})$</td>
<td>✓</td>
<td>[Han et al 08]</td>
</tr>
<tr>
<td>Standard Structures</td>
<td>$O(n \cdot m^k)$</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Standard Pseudoknots</td>
<td>$O(n \cdot m^k)$</td>
<td>✓</td>
<td>[Han et al 08]</td>
</tr>
<tr>
<td>2-Level Recursive Simple Non-Standard PKs</td>
<td>$O(n \cdot m^{k+2})$</td>
<td>✓</td>
<td>[Wong et al 11]</td>
</tr>
<tr>
<td>Simple Non-Standard Structures</td>
<td>$O(n \cdot m^{k+1})$</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Simple Non-Standard Pseudoknots</td>
<td>$O(n \cdot m^{k+1})$</td>
<td>✓</td>
<td>[Wong et al 11]</td>
</tr>
<tr>
<td>Extended Triple Helices</td>
<td>$O(n \cdot m^3)$</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Triple Helices</td>
<td>$O(n \cdot m^3)$</td>
<td>✓</td>
<td>[Wong et al 12]</td>
</tr>
</tbody>
</table>

$n \rightarrow$ Structure length  
$m \rightarrow$ Sequence length  
$k \rightarrow$ Class-specific structural parameter
New classes of structures [Rinaudo et al. 2012]

Recursive Classical Structures
- Standard Structures
- Simple Non-Standard Structures
- Extended Triple Helices

Kink Turn

C-loop

Sarcin-ricin

\[O(n \cdot m^k + 2)\]
\[O(n \cdot m^k)\]
\[O(n \cdot m^{k+1})\]
\[O(n \cdot m^3)\]
Half-time summary

- No real FPT algorithm yet! Any clue, parameters?

- Clear connection between existing parameters and tree decomposition → Use for algorithm design?

- Probabilistic interpretation? (MEA, Bayesian networks...)

- Compare with co-variance models
Part. III: Designing RNAs
RNA inverse folding

**RNA** = Linear Polymer = Sequence in \{A, C, G, U\}*

**MFE folding prediction:** $\Theta(n^3)$

**Inverse folding:** NP-hard?

Primary Structure  Secondary Structure  Structure Tertiaire

UUAGCGGCCACAGC  GGUGGGUUGCUC  CGUACCCAUCCCGA  CAGGGAAGAUAAGCC  CACCAGCGUUCCGGG  GAGUACUGGAGUGCG  CGAGCCUCUGGGAAA  CCCGGUUCGCCGCAC  CCUC

5s rRNA (PDBID: 1K73:B)
RNA Inverse Folding

\[ M = \text{energy model} \]

**Definition (\textsc{inverse-folding}(E) problem)**

**Input:** Secondary structure \( S \) + Energy distance \( \Delta > 0 \).

**Output:** RNA sequence \( w \in \Sigma^* \) such that:

\[ \forall S' \in S | w | \setminus \{ S \} : E_{w,S'} \geq E_w, S + \Delta \]

or \( \emptyset \) if no such sequence exists.

No (obvious?) optimal substructure property:
RNA Inverse Folding

\[ M = \text{energy model} \]

**Definition (Inverse-Folding}(E)\text{ problem)\)**

**Input:** Secondary structure \( S \) + Energy distance \( \Delta > 0 \).

**Output:** RNA sequence \( w \in \Sigma^* \) such that:

\[ \forall S' \in S \setminus \{S\} : E_{w,S'} \geq E_{w,S} + \Delta \]

or \( \emptyset \) if no such sequence exists.

No (obvious?) optimal substructure property:
RNA Inverse Folding

\[ M = \text{energy model} \]

**Definition (INVERSE-FOLDING(\(E\)) problem)**

**Input:** Secondary structure \( S \) + Energy distance \( \Delta > 0 \).

**Output:** RNA sequence \( w \in \Sigma^* \) such that:

\[ \forall S' \in S \mid w \setminus \{S\} : E_{w,S'} \geq E_w, S + \Delta \]

or \( \emptyset \) if no such sequence exists.

No (obvious?) optimal substructure property:

\[
\begin{align*}
1 & 10 \\
10 & \\
20 & \\
26 & \\
1 & 10 \\
13 & 10 \\
1 & 10 \\
13 & 10 \\
\end{align*}
\]

folds

\[
\begin{align*}
AAGAGUCGCUCUC
\end{align*}
\]
RNA Inverse Folding

\[ \mathcal{M} = \text{energy model} \]

**Definition (INVERSE-FOLDING}(E) \text{ problem)**

**Input:** Secondary structure \( S \) + Energy distance \( \Delta > 0 \).

**Output:** RNA sequence \( w \in \Sigma^* \) such that:

\[ \forall S' \in S \setminus \{ S \} : E_{w,S'} \geq E_w, S + \Delta \]

or \( \emptyset \) if no such sequence exists.

No (obvious?) optimal substructure property:

AAGAGUCGCUCUC AAGAGUCGCUCUC

Folds

AAGAGUCGCUCUC

Folds

Yann Ponty (CNRS/Polytechnique)

Combinatorial problems in RNA Bioinformatics

14/03/2016 – LAMSADE Seminar
RNA Design Problem

\[ M = \text{energy model} \]

**Definition (INVERSE-FOLDING(\(E\)) problem)**

**Input:** Secondary structure \( S \) + Energy distance \( \Delta > 0 \).
**Output:** RNA sequence \( w \in \Sigma^* \) such that:

\[ \forall S' \in S \setminus \{ S \} : E_{w,S'} \geq E_w, S + \Delta \]

or \( \emptyset \) if no such sequence exists.

**Difficult problem:** No (obvious??) substructure property

- **Existing algorithms/software (20+):** Heuristics or Exponential-time

- **Complexity of problem unknown (despite [Schnall Levin et al (2008)])**
  Clearly in \( P! \ldots \text{CO-NP???} \)

- **Reason:** Non locality, no theoretical frameworks, too many parameters…

⇒ Stick to a simplified model!
RNA Design Problem (simplified)

Simplified formulation for Watson-Crick model $\mathcal{W}$ and $\Delta = 1$:

**Problem ([INVERSE-FOLDING](\Sigma) problem)**

*Input:* Secondary structure $S$

*Output:* RNA sequence $w \in \Sigma^*$ — called a design for $S$ — such that:

$$\text{RNA-FOLD}_\mathcal{W}(w) = \{S\}$$

or $\emptyset$ if no such sequence exists.

**Designable($\Sigma$):** All designable structures
RNA Design Problem (simplified)

Simplified formulation for Watson-Crick model $\mathcal{W}$ and $\Delta = 1$:

**Problem (INVERSE-FOLDING($\Sigma$) problem)**

*Input*: Secondary structure $S$

*Output*: RNA sequence $w \in \Sigma^*$ — called a design for $S$ — such that:

$$\text{RNA-FOLD}_\mathcal{W}(w) = \{S\}$$

or $\emptyset$ if no such sequence exists.

**Designable($\Sigma$)**: All designable structures

**Example**

- **a.** Target sec. str. $S$
- **b.** Invalid sequence for $S$
- **c.** Design for $S$

---

Yann Ponty (CNRS/Polytechnique)

Combinatorial problems in RNA Bioinformatics

14/03/2016 – LAMSADÉ Seminar
Our Results: Definitions and notations

Given a secondary structure $S$:

- $\text{Unpaired}_S = \text{Set of all unpaired positions of } S$.
- $S$ is saturated $\iff \text{Unpaired}_S = \emptyset$.
- $\text{Saturated} = \text{Set of all saturated structures}$.
- $\text{Paired degree of base-pair} = \text{#Helices on the loop}$.
- $D(S) = \text{Maximal } \text{paired degree} \text{ of nodes in the tree representation of } S$.

Example

1  8

Unpaired$_S = \{4, 8\}$
Our Results: Definitions and notations

Given a secondary structure $S$:

- **Unpaired$_S$** = Set of all unpaired positions of $S$.
- **$S$ is saturated** $\iff$ **Unpaired$_S$** = $\emptyset$.
- Saturated = Set of all saturated structures.
- **Paired degree of base-pair** = #Helices on the loop.
- $D(S)$ = Maximal *paired degree* of nodes in the tree representation of $S$.

Example

```
1 8
```

Unsaturated

```
1 10
```

Saturated
Our Results: Definitions and notations

Given a secondary structure $S$:

- $\text{Unpaired}_S = \text{Set of all unpaired positions of } S$.
- $S$ is saturated $\iff$ $\text{Unpaired}_S = \emptyset$.
- $\text{Saturated} = \text{Set of all saturated structures}$.
- **Paired degree of base-pair** $= \#\text{Helices on the loop}$.
- $D(S) = \text{Maximal paired degree of nodes in the tree representation of } S$.

**Example**

![Diagram showing a secondary structure and its tree representation with node labels and $D(S) = 3$]
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

- **R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures}; \)
- **R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures} ; \)
- **R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

**R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures;} \)

**R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures} ; \)

**R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

**Example**

```
    1  2  3  4  5  6  7  8

    1
```

Yann Ponty (CNRS/Polytechnique)  Combinatorial problems in RNA Bioinformatics  14/03/2016 – LAMSADÉ Seminar  35 / 47
Our Results: Designability over Restricted Alphabets

$$\Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.}$$

R1  $$\Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures;}$$

R2  $$\Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures ;}$$

R3  $$\Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2.$$

Example
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

- **R1**  \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures;} \)
- **R2**  \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures} ; \)
- **R3**  \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

**Example**

\[ + \text{miRNAs, some lncRNAs...} \]
Our Results: Designability over Restricted Alphabets

$\Sigma_{c,u} =$ Alphabet with $c$ pairs of complementary bases and $u$ unpairable bases.

**R1** $\Sigma_{0,u} \Rightarrow$ Designable $=$ Empty (single-stranded) structures;

**R2** $\Sigma_{1,0} \Rightarrow$ Designable $=$ Saturated with degree $\leq 2$ + empty structures;

**R3** $\Sigma_{1,1} \Rightarrow$ Designable $=$ Degree $\leq 2$.

**Question:** Why not degree 3?

**Proof.**
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

**R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures}; \)

**R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures}; \)

**R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

**Question:** Why not degree 3?

**Proof.**

Within an internal node:

\[ \ldots ? ? \ldots ? ? \ldots ? ? \ldots \]
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

**R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures;} \)

**R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures} ; \)

**R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

**Question:** Why not degree 3?

**Proof.**

Within an internal node:

... ? C ... G C ... G ? ...  

Either we get a repeat...
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.} \]

- **R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures;} \)
- **R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures;} \)
- **R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

**Question:** Why not degree 3?

**Proof.**

Within an internal node:

*Either we get a repeat...*

*... or some parent/child have complementary pairs.*

*+ Same principle at the root level.*
Our Results: Designability over Restricted Alphabets

\[ \Sigma_{c,u} = \text{Alphabet with } c \text{ pairs of complementary bases and } u \text{ unpairable bases.}\]

- **R1** \( \Sigma_{0,u} \Rightarrow \text{Designable} = \text{Empty (single-stranded) structures}; \)
- **R2** \( \Sigma_{1,0} \Rightarrow \text{Designable} = \text{Saturated with degree} \leq 2 + \text{empty structures}; \)
- **R3** \( \Sigma_{1,1} \Rightarrow \text{Designable} = \text{Degree} \leq 2. \)

This can be easily generalized to:

**Lemma**

*For any structure \( S \) in Designable(\( \Sigma_{c,u} \)), \( D(S) \leq 2c. \)*
\[ \Sigma_{2,0} = \{A, U, C, G\} + \{G - C, A - U\} \text{ base pairs.} \]

**Without unpaired position \(\rightarrow\) complete characterization:**

R4 \(\Sigma_{2,0} \Rightarrow\) Saturated Designable = Degree \(\leq 4\).

**With unpaired positions \(\rightarrow\) partial characterization:**

R5 (Necessary) Designable structure cannot contain “a multiloop of degree \(\geq 5\)” (motif \(m_5\)) or “a multiloop with unpaired position of degree \(\geq 3\)” (motif \(m_{3^o}\)).

R6 (Sufficient) Separated = Structure that admit a separated (proper) coloring. Then any Separated structure is Designable in \(\Sigma_{2,0}\).
Our Results: Designability over the Complete Alphabet

\[ \Sigma_{2,0} = \{A, U, C, G\} + \{G - C, A - U\} \text{ base pairs.} \]

Without unpaired position \(\rightarrow\) complete characterization:

**R4**  \(\Sigma_{2,0} \Rightarrow \text{Saturated Designable} = \text{Degree} \leq 4.\)

With unpaired positions \(\rightarrow\) partial characterization:

**R5** (Necessary) Designable structure cannot contain “\textit{a multiloop of degree } \geq 5\textit{”} (motif \(m_5\)) or “\textit{a multiloop with unpaired position of degree } \geq 3\textit{”} (motif \(m_3^\circ\)).

**R6** (Sufficient) \textit{Separated} = Structure that admit a separated (proper) coloring. Then any \textit{Separated structure is Designable} in \(\Sigma_{2,0}.\)
Our Results: Designability over the Complete Alphabet

\[ \Sigma_{2,0} = \{A, U, C, G\} + \{G − C, A − U\} \] base pairs.

Without unpaired position \(\Rightarrow\) complete characterization:

**R4** \(\Sigma_{2,0} \Rightarrow \text{Saturated Designable} = \text{Degree} \leq 4\).

With unpaired positions \(\Rightarrow\) partial characterization:

**R5** (Necessary) Designable structure cannot contain “a multiloop of degree \(\geq 5\)” (motif \(m_5\)) or “a multiloop with unpaired position of degree \(\geq 3\)” (motif \(m_3^o\)).

**R6** (Sufficient) Separated = Structure that admit a separated (proper) coloring. Then any Separated structure is Designable in \(\Sigma_{2,0}\).
Our Results: Designability over the Complete Alphabet

\[ \Sigma_{2,0} = \{A, U, C, G\} + \{G - C, A - U\} \text{ base pairs.} \]

Without unpaired position \(\rightarrow\) complete characterization:

R4 \(\Sigma_{2,0} \Rightarrow\) Saturated Designable = Degree \(\leq 4\).

With unpaired positions \(\rightarrow\) partial characterization:

R5 (Necessary) Designable structure cannot contain "a multiloop of degree \(\geq 5\)" (motif \(m_5\)) or "a multiloop with unpaired position of degree \(\geq 3\)" (motif \(m_{3\circ}\)).

R6 (Sufficient) Separated = Structure that admit a separated (proper) coloring. Then any Separated structure is Designable in \(\Sigma_{2,0}\).
Our Results: Separated Coloring

From the tree representation $T_S$ of structure $S$, color every paired node of $T_S$:

- black $\rightarrow G \cdot C$;
- white $\rightarrow C \cdot G$;
- grey $\rightarrow A \cdot U$ or $U \cdot A$.

Proper coloring:

1. each internal node has at most one black, one white and two grey children;
2. a grey node has at most one grey child;
3. a black node does not have a white child; and
4. a white node does not have a black child.

Level of a node = $\#$black nodes $-$ $\#$white nodes on the path to the root.

Separated coloring: Levels of grey nodes $\cap$ Levels of unpaired nodes $= \emptyset$
Our Results: Separated Coloring

From the tree representation $T_S$ of structure $S$, color every paired node of $T_S$:

- black $\rightarrow$ G · C;
- white $\rightarrow$ C · G;
- grey $\rightarrow$ A · U or U · A.

Proper coloring:

1. each internal node has at most one black, one white and two grey children;
2. a grey node has at most one grey child;
3. a black node does not have a white child; and
4. a white node does not have a black child.

Level of a node = #black nodes $-$ #white nodes on the path to the root.

Separated coloring: Levels of grey nodes $\cap$ Levels of unpaired nodes $= \emptyset$
Our Results: Separated Coloring

From the tree representation $T_S$ of structure $S$, color every paired node of $T_S$:

- black $\rightarrow$ G · C;
- white $\rightarrow$ C · G;
- grey $\rightarrow$ A · U or U · A.

Proper coloring:

1. each internal node has at most one black, one white and two grey children;
2. a grey node has at most one grey child;
3. a black node does not have a white child; and
4. a white node does not have a black child.

Level of a node $= \#\text{black nodes} - \#\text{white nodes}$ on the path to the root.

Separated coloring: Levels of grey nodes $\cap$ Levels of unpaired nodes $= \emptyset$
Our Results: Separated Coloring (example)

Descendant restrictions: Any node $\rightarrow \leq 1$ black & $\leq 1$ White & $\leq 2$ Grey;
Grey $\rightarrow 0/1$ Grey; Black $\rightarrow 0$ White; White $\rightarrow 0$ Black.
($\bullet \rightarrow GC$  $\circ \rightarrow CG$  $\bullet \rightarrow AU|UA$  $\times \rightarrow U)$

Levels of grey nodes: 0,1
Levels of leaves: 2,4
Separated coloring $\Rightarrow$ Design:

GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC

Yann Ponty (CNRS/Polytechnique)  Combinatorial problems in RNA Bioinformatics  14/03/2016 – LAMSADE Seminar
Our Results: Separated Coloring (example)

Descendant restrictions: Any node → ≤ 1 black & ≤ 1 White & ≤ 2 Grey; Grey → 0/1 Grey; Black → 0 White; White → 0 Black.

(● → GC  ○ → CG  ● → AU|UA  x → U)

Root

Levels of grey nodes: 0,1
Levels of leaves: 2,4

Separated coloring ⇒ Design:
GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC

Yann Ponty (CNRS/Polytechnique)
Combinatorial problems in RNA Bioinformatics
14/03/2016 – LAMSADE Seminar
Our Results: Separated Coloring (example)

Descendant restrictions: Any node $\rightarrow \leq 1$ black & $\leq 1$ White & $\leq 2$ Grey;
Grey $\rightarrow 0/1$ Grey; Black $\rightarrow 0$ White; White $\rightarrow 0$ Black.

($\bullet \rightarrow$ GC $\circ \rightarrow$ CG $\bullet \rightarrow$ AU|UA $\times \rightarrow$ U)
Our Results: Separated Coloring (example)

Descendant restrictions:

Any node $\rightarrow \leq 1$ black & $\leq 1$ White & $\leq 2$ Grey;
Grey $\rightarrow 0/1$ Grey; Black $\rightarrow 0$ White; White $\rightarrow 0$ Black.

($\bullet \rightarrow$ GC $\bigcirc \rightarrow$ CG $\blacklozenge \rightarrow$ AU|UA $\times \rightarrow$ U)

Levels of grey nodes: 0,1
Levels of leaves: 2,4
Separated coloring
Descendant restrictions: Any node $\rightarrow \leq 1$ black & $\leq 1$ White & $\leq 2$ Grey;
Grey $\rightarrow 0/1$ Grey; Black $\rightarrow 0$ White; White $\rightarrow 0$ Black.
($\bullet \rightarrow$ GC $\circ \rightarrow$ CG $\circ \rightarrow$ AU|UA $\times \rightarrow$ U)

Levels of grey nodes: 0,1
Levels of leaves: 2,4
Separated coloring

⇒ Design: GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC
Our Results: Designability over the complete alphabet

\[ \Sigma_{2,0} = \{ A, U, C, G \} + \{ G - C, A - U \} \text{ base pairs.} \]

Without unpaired position \( \rightarrow \) complete characterization:

**R4** \( \Sigma_{2,0} \Rightarrow \text{Saturated Designable} = \text{Degree} \leq 4. \)

With unpaired positions \( \rightarrow \) partial characterization:

**R5** (Necessary) Designable structure cannot contain “a multiloop of degree \( \geq 5 \)” (motif \( m_5 \)) or “a multiloop with unpaired position of degree \( \geq 3 \)” (motif \( m_3 \circ \)).

**R6** (Sufficient) Separated = Structure that admit a separated (proper) coloring. Then any Separated structure is Designable in \( \Sigma_{2,0} \).

**R7** If \( S \in \text{Designable}(\Sigma_{2,0}) \), then \( k \)-stutter \( S^{[k]} \in \text{Designable}(\Sigma_{2,0}) \).
Our Results: $k$-Stutter

Designable structure: 

Then 2-stutter is designable as well:
Our Results: \( k \)-Stutter

Designable structure:  

\[
A \quad C \quad A \quad G \quad G \quad U \quad U \quad C \quad U
\]

Then 2-stutter is designable as well:  

\[
( ( ( . . ) ) ) ( ( . . . . ) )
\]
Our Results: *k*-Stutter

Designable structure: \[ \text{A C A G G U U C U} \]

Then 2-stutter is designable as well: \[ \text{A A C C A A G G G U U U U C C U U} \]
Our Results: \( k \)-Stutter

Designable structure:

\[
\begin{array}{ccccccc}
A & C & A & G & G & U & U & C & U \\
\end{array}
\]

Then 2-stutter is designable as well:

\[
\begin{array}{cccccccccc}
\end{array}
\]

Proof idea: \( w \): Design for \( S \); \( S' \neq S^{[k]} \): Alternative folding for \( k \)-stutter \( w^{[k]} \):

- Compact \( k \) consecutive positions \( \rightarrow \) Multigraph \( G \) such that \( \Delta(G) = k \)
- Base-pair compatibility graph is bipartite \( \rightarrow \) \( G \) is also bipartite
- Therefore \( G \) is \( k \) edge-colorable
  - Any restriction of \( G \) to a given color \( c \) = Valid structure \( S_c \) for \( w \)
  - Either \( E_M(S_c) = E_M(S) (\Rightarrow S_c = S) \), or \( E_M(S_c) > E_M(S) \) (holds for some \( c \))
  - Thus \( \sum_c E_M(S_c) > k \cdot E(S) = E(S^{[k]}) \)
\( \Rightarrow \) \( w^{[k]} \) is design for \( S^{[k]} \) (holds for any base-pair additive \( M \))
Our Results: $k$-Stutter

**Proof idea:** $w$: Design for $S$; $S' \neq S^{[k]}$: Alternative folding for $k$-stutter $w^{[k]}$:

- Compact $k$ consecutive positions $\rightarrow$ Multigraph $G$ such that $\Delta(G) = k$
- Base-pair compatibility graph is bipartite $\rightarrow$ $G$ is also bipartite
- Therefore $G$ is $k$ edge-colorable
- Any restriction of $G$ to a given color $c = $ Valid structure $S_c$ for $w$
- Either $E_M(S_c) = E_M(S) (\Rightarrow S_c = S)$, or $E_M(S_c) > E_M(S)$ (holds for some $c$)
- Thus $\sum_c E_M(S_c) > k \cdot E(S) = E(S^{[k]})$

$w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $M$)
Our Results: $k$-Stutter

Designable structure: \[
\begin{array}{cccccccc}
A & C & A & G & G & U & U & C & U
\end{array}
\]

Then 2-stutter is designable as well:

\[
\begin{array}{cccccccc}
\end{array}
\]

Proof idea: \(w\): Design for \(S\); \(S' \neq S^{[k]}\): Alternative folding for \(k\)-stutter \(w^{[k]}\):

- Compact \(k\) consecutive positions \(\rightarrow\) Multigraph \(G\) such that \(\Delta(G) = k\)
- Base-pair compatibility graph is bipartite \(\rightarrow\) \(G\) is also bipartite
- Therefore \(G\) is \(k\) edge-colorable
- Any restriction of \(G\) to a given color \(c\) = Valid structure \(S_c\) for \(w\)
- Either \(E_{\mathcal{M}}(S_c) = E_{\mathcal{M}}(S) \Rightarrow S_c = S\), or \(E_{\mathcal{M}}(S_c) > E_{\mathcal{M}}(S)\) (holds for some \(c\))
  - Thus \(\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})\)
  - \(w^{[k]}\) is design for \(S^{[k]}\) (holds for any base-pair additive \(\mathcal{M}\))
Our Results: $k$-Stutter

Designable structure:

Then $2$-stutter is designable as well:

**Proof idea:** $w$: Design for $S$; $S' \neq S^{[k]}$: Alternative folding for $k$-stutter $w^{[k]}$:

- Compact $k$ consecutive positions $\rightarrow$ Multigraph $G$ such that $\Delta(G) = k$
- Base-pair compatibility graph is bipartite $\rightarrow$ $G$ is also bipartite
- Therefore $G$ is $k$ edge-colorable
- Any restriction of $G$ to a given color $c = \text{Valid structure } S_c$ for $w$
- Either $E_M(S_c) = E_M(S) (\Rightarrow S_c = S)$, or $E_M(S_c) > E_M(S)$ (holds for some $c$)
- Thus $\sum_c E_M(S_c) > k \cdot E(S) = E(S^{[k]})$

$\Rightarrow w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $M$)
Our Results: \( k \)-Stutter

Designable structure:

\[
\text{A C A G G U U C U}
\]

Then 2-stutter is designable as well:

\[
\text{A A C C A A G G G G U U U U C C U U}
\]

Proof idea: \( w \): Design for \( S \); \( S' \neq S^{[k]} \): Alternative folding for \( k \)-stutter \( w^{[k]} \):

- Compact \( k \) consecutive positions → Multigraph \( G \) such that \( \Delta(G) = k \)
- Base-pair compatibility graph is bipartite → \( G \) is also bipartite
- Therefore \( G \) is \( k \) edge-colorable
- Any restriction of \( G \) to a given color \( c \) = Valid structure \( S_c \) for \( w \)
- Either \( E_{\mathcal{M}}(S_c) = E_{\mathcal{M}}(S) \) (⇒ \( S_c = S \)), or \( E_{\mathcal{M}}(S_c) > E_{\mathcal{M}}(S) \) (holds for some \( c \))
- Thus \( \sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]}) \)

⇒ \( w^{[k]} \) is design for \( S^{[k]} \) (holds for any base-pair additive \( \mathcal{M} \))
Any structure $S$ without $m_5$ and $m_3^\circ$ can be transformed in $\Theta(n)$ time into a designable structure $S'$, by adding at most a single base-pair to its helices.

Main idea: Offset grey vertices and leaves to odd/even levels → Coloring is now separated
Our Results: Structure-Approximating Algorithm

**R8** Any structure $S$ without $m_5$ and $m_3^0$ can be transformed in $\Theta(n)$ time into a designable structure $S'$, by adding at most a single base-pair to its helices.

Main idea: Offset grey vertices and leaves to odd/even levels → Coloring is now separated
Example
Example
Example
Generalization

**Theorem**

All the above results hold in any energy models $\mathcal{M}$:

$$E_\mathcal{M}(X, Y) = \begin{cases} 
\alpha & \text{if } \{X, Y\} = \{G, C\} \\
\beta & \text{if } \{X, Y\} = \{A, U\} \\
\gamma & \text{if } \{X, Y\} = \{G, U\} \\
+\infty & \text{otherwise}
\end{cases}$$

such that $\alpha, \beta > \gamma$.

**Proof idea:** Stutter results holds for any base-pair additive model. Other results are based on $(G, C)$-saturated sequences. No G–U base pair in optimal fold, since $\alpha > \gamma$. Numbers of G–C and A–U base pairs are upper-bounded. $\Rightarrow$ Any alternative has same number of each base-pair as target structure.
Remarks

- Results also hold in **Nussinov** energy model \((A - U, G - C, G - U + \text{weights})\) ⇒ **Stacking** energy model? **Turner**?

- Characterized classes are mostly easy:
  - **Designable** classes → Linear time algorithms
  - **Non-designable** classes → Linear time membership tests

- Complexity of finding **separated coloring**?

- **Forbidden local motifs** (e.g. \(m_5\) & \(m_3\)) can be found in any energy model ⇒ **Designable structures** ⊂ **Tree-like** objects with **forbidden motifs**

  + Basic analytic combinatorics (à la Philippe Flajolet):
    - #Secondary structures ∈ \(\Theta \left( \frac{\alpha^n}{n\sqrt{n}} \right)\) \((\theta = 0 \rightarrow \alpha = 3)\)
    - #Designable structures ∈ \(\mathcal{O} \left( \frac{\beta^n}{n\sqrt{n}} \right)\), \(\beta < \alpha\)

Proportion of designable structures: \(\left( \frac{\beta}{\alpha} \right)^n\), exponentially decreasing with \(n\).

Possible consequences on **RNA neutral network** studies

+ motivation for identifying **new forbidden motifs**
Remarks

- Results also hold in Nussinov energy model \((A - U, G - C, G - U + \text{weights})\)
  
  \(\Rightarrow\) **Stacking** energy model? **Turner**?

- Characterized classes are mostly **easy**:
  - **Designable** classes \(\rightarrow\) Linear time **algorithms**
  - **Non-designable** classes \(\rightarrow\) Linear time **membership tests**

- Complexity of finding **separated coloring**?

- Forbidden local motifs (e.g. \(m_5\) & \(m_3\)) can be found in any energy model
  
  \(\Rightarrow\) **Designable structures** \(\subset\) **Tree-like** objects with **forbidden motifs**

  + **Basic analytic combinatorics** (à la Philippe Flajolet):
    - \#Secondary structures \(\in \Theta\left(\frac{\alpha^n}{n\sqrt{n}}\right)\) \((\theta = 0 \rightarrow \alpha = 3)\)
    - \#Designable structures \(\in \mathcal{O}\left(\frac{\beta^n}{n\sqrt{n}}\right)\), \(\beta < \alpha\)

  Proportion of designable structures: \((\frac{\beta}{\alpha})^n\), exponentially decreasing with \(n\).

  Possible consequences on **RNA neutral network** studies

  + motivation for identifying **new forbidden motifs**
Remarks

- Results also hold in Nussinov energy model (A – U, G – C, G – U + weights)
  ⇒ Stacking energy model? Turner?

- Characterized classes are mostly easy:
  - Designable classes → Linear time algorithms
  - Non-designable classes → Linear time membership tests

- Complexity of finding separated coloring?

- Forbidden local motifs (e.g. \(m_5\) & \(m_3\)) can be found in any energy model
  ⇒ Designable structures ⊂ Tree-like objects with forbidden motifs

  + Basic analytic combinatorics (à la Philippe Flajolet):
    - #Secondary structures ∈ \(\Theta\left(\frac{\alpha^n}{n\sqrt{n}}\right)\) (\(\theta = 0 \rightarrow \alpha = 3\))
    - #Designable structures ∈ \(O\left(\frac{\beta^n}{n\sqrt{n}}\right)\), \(\beta < \alpha\)

  Proportion of designable structures: \(\left(\frac{\beta}{\alpha}\right)^n\), exponentially decreasing with \(n\).

  Possible consequences on RNA neutral network studies

  + motivation for identifying new forbidden motifs
Remarks

- Results also hold in Nussinov energy model \((A - U, G - C, G - U + \text{weights})\) \(\Rightarrow\) Stacking energy model? Turner?

- Characterized classes are mostly easy:
  - Designable classes \(\rightarrow\) Linear time algorithms
  - Non-designable classes \(\rightarrow\) Linear time membership tests

- Complexity of finding separated coloring?

- Forbidden local motifs (e.g. \(m_5\) & \(m_3\circ\)) can be found in any energy model
  \(\Rightarrow\) Designable structures \(\subset\) Tree-like objects with forbidden motifs

  + Basic analytic combinatorics (à la Philippe Flajolet):
    - \#Secondary structures \(\in \Theta \left(\frac{\alpha^n}{n^{\theta}}\right)\) \((\theta = 0 \rightarrow \alpha = 3)\)
    - \#Designable structures \(\in O \left(\frac{\beta^n}{n^{\theta}}\right), \beta < \alpha\)

  Proportion of designable structures: \(\left(\frac{\beta}{\alpha}\right)^n\), exponentially decreasing with \(n\).
  Possible consequences on RNA neutral network studies
  + motivation for identifying new forbidden motifs
Conclusion (Design)

- **RNA is cool!**
  - **RNA design** is one of the current challenges of RNA bioinformatics with far-reaching consequences for drug design, synthetic biology.

- Practical use-cases require **expressive and modular constraints**

- Future methods: **kinetics, interactions, multiple structures, pseudoknots**

- **RNA inverse folding** is the combinatorial core of design. It remains **largely unsolved**, and opens **new lines of research** in Comp. Sci.
We need your help!

- **Crossing interactions (pseudoknots):** Finding the right parameter
- **RNA Kinetics:** Markov process... computing energy barrier is hard!
- **RNA Inverse folding/Design:** Complexity open! (missing theory?)
- **Beyond optimization:** Subopts, Boltzmann sampling...
Thanks

University McGill
  Vladimir Reinharz
  Jérôme Waldispühl

MIT
  Bonnie Berger
  Srinivas Devadas
  Alex Levin
  Mieszko Lis
  Charles O’Donnell

LRI – Univ. Paris Sud
  Alain Denise
  Philippe Rinaudo

Wuhan University
  Yi Zhang
  Yu Zhou

LIGM – Marne la Vallée
  Stéphane Vialette

LIX – Ecole Polytechnique
  Alice Héliou
  Saad Sheikh

Simon Fraser University
  Jozef Hales
  Jan Manuch (UBC)
  Ladislav Stacho
  Cédric Chauve
  Julien Courtiel

TBI Vienna
  Ronnie Lorenz
  Andrea Tanzer

Job offer: Postdoc on RNA kinetics@Inria Saclay+Lille