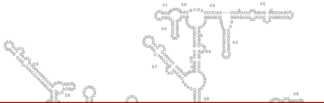
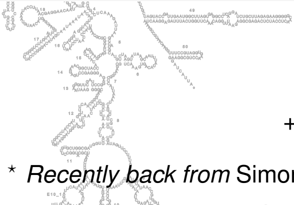
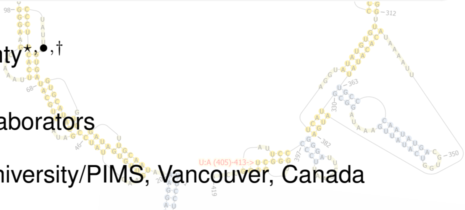


Selected combinatorial problems in RNA Bioinformatics

. . . and some solutions





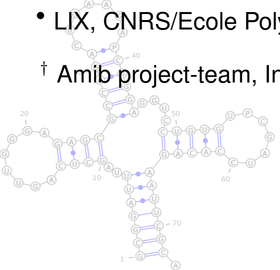
Yann Ponty[*,•,†]

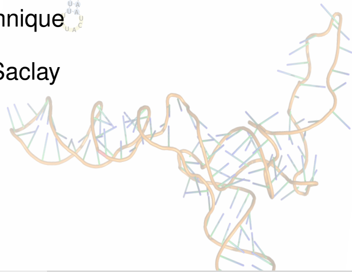
+ Many collaborators

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

[•] LIX, CNRS/Ecole Polytechnique

[†] Amib project-team, Inria Saclay

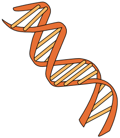


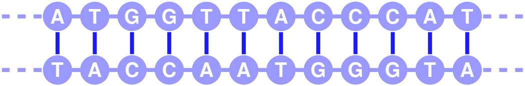


Fundamental *dogma* of molecular biology

DNA

$\{A, C, G, T\}^\star$

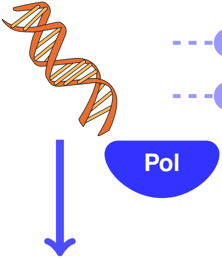


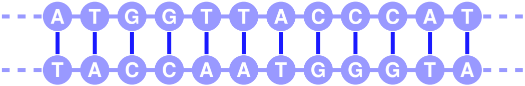


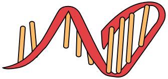
Fundamental *dogma* of molecular biology

DNA

$\{A, C, G, T\}^\star$







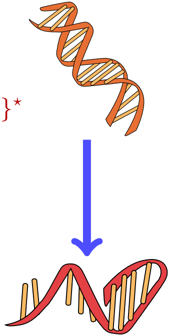
RNAs

$\{A, C, G, U\}^\star$

Fundamental *dogma* of molecular biology

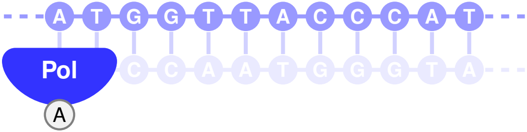
DNA

$\{A, C, G, T\}^\star$



RNAs

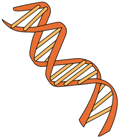
$\{A, C, G, U\}^\star$

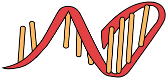


Fundamental *dogma* of molecular biology

DNA

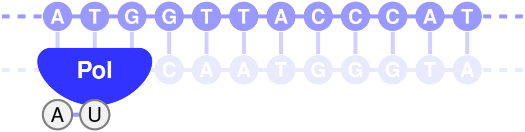
$\{A, C, G, T\}^\star$





RNAs

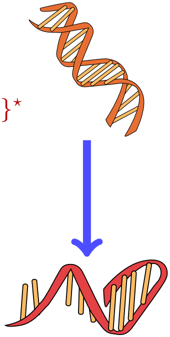
$\{A, C, G, U\}^\star$



Fundamental *dogma* of molecular biology

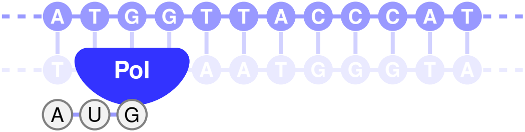
DNA

$\{A, C, G, T\}^\star$

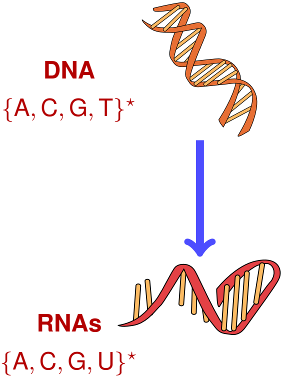


RNAs

$\{A, C, G, U\}^\star$



Fundamental *dogma* of molecular biology

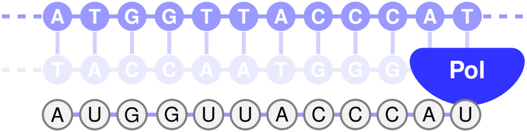


DNA

$\{A, C, G, T\}^\star$

RNAs

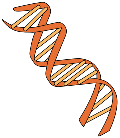
$\{A, C, G, U\}^\star$

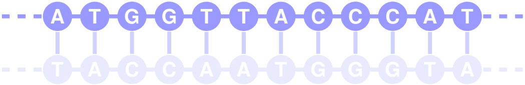


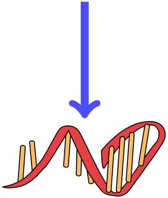
Fundamental *dogma* of molecular biology

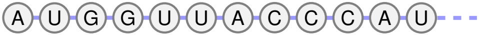
DNA

$\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^\star$





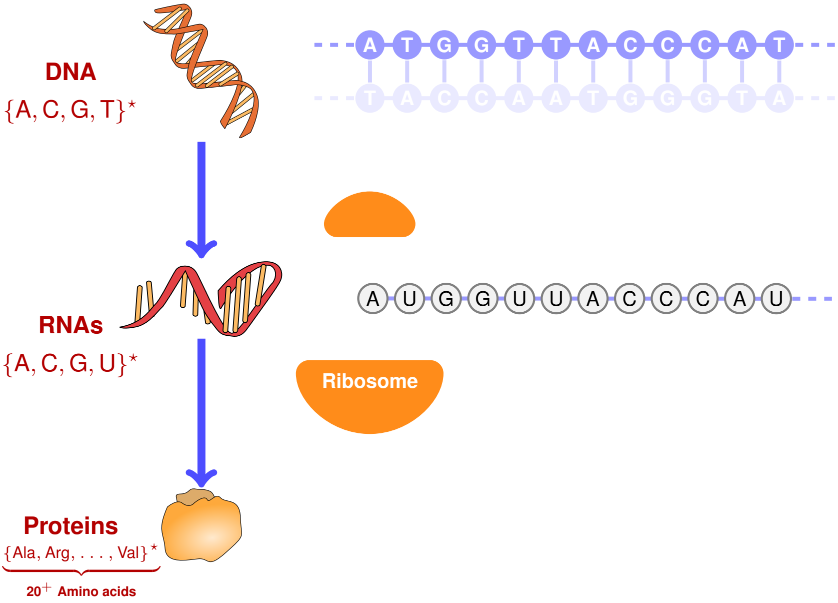




RNAs

$\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^\star$

Fundamental *dogma* of molecular biology



DNA

$\{A, C, G, T\}^\star$

A T G G T T A C C C A T

T A C C A A T G G G T A

RNAs

$\{A, C, G, U\}^\star$

A U G G U U A C C C A U

Ribosome

Proteins

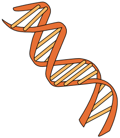
$\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star$

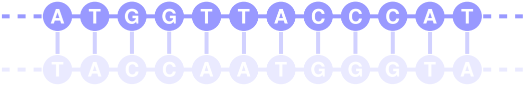
20^+ Amino acids

Fundamental *dogma* of molecular biology

DNA

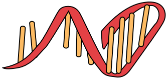
$\{A, C, G, T\}^\star$

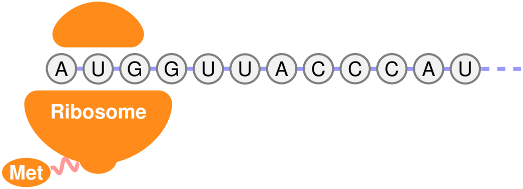




RNAs

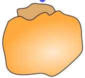
$\{A, C, G, U\}^\star$



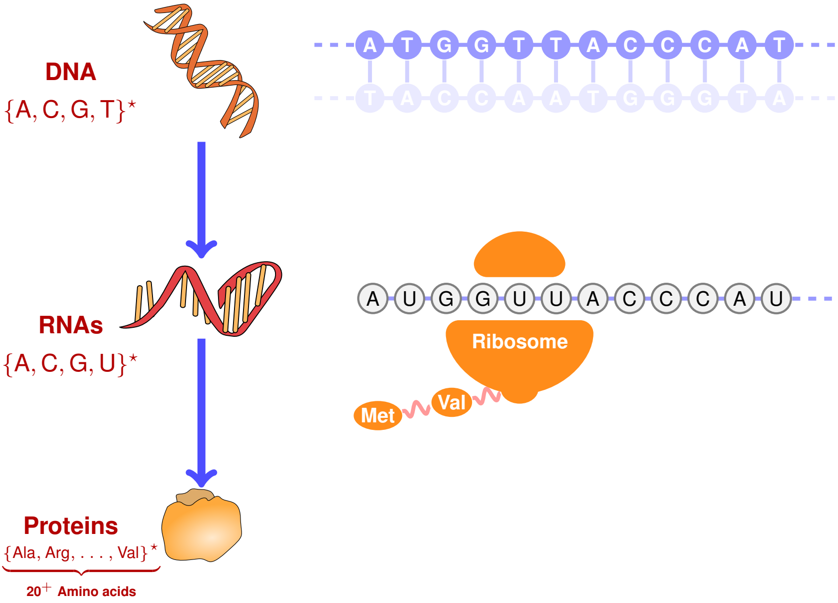


Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$



Fundamental *dogma* of molecular biology



DNA

$\{A, C, G, T\}^\star$

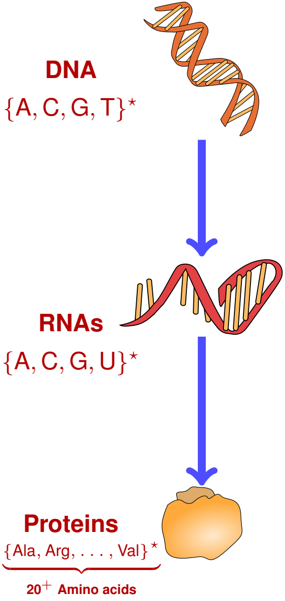
RNAs

$\{A, C, G, U\}^\star$

Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \textbf{ Amino acids}}$

Fundamental *dogma* of molecular biology



DNA

$\{A, C, G, T\}^\star$

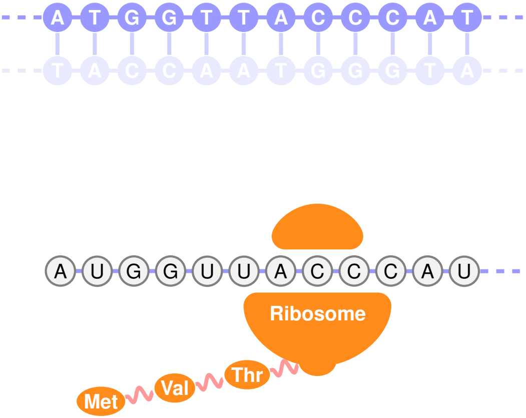
RNAs

$\{A, C, G, U\}^\star$

Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{}$

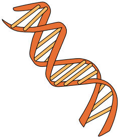
20^+ **Amino acids**

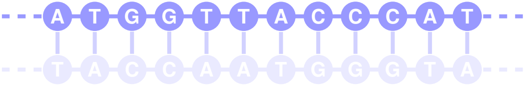


Fundamental *dogma* of molecular biology

DNA

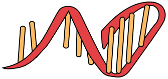
$\{A, C, G, T\}^\star$

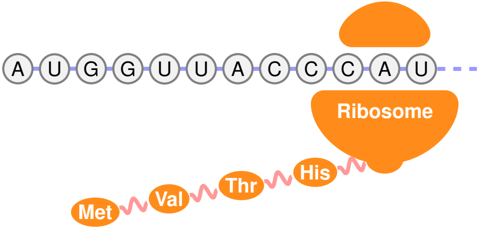




RNAs

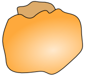
$\{A, C, G, U\}^\star$





Proteins

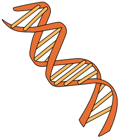
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{\textbf{20}^+\ \textbf{Amino acids}}$

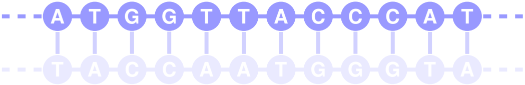


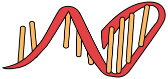
Fundamental *dogma* of molecular biology

DNA

$\{A, C, G, T\}^\star$

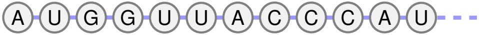


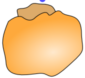




RNAs

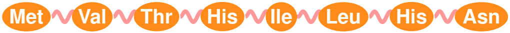
$\{A, C, G, U\}^\star$



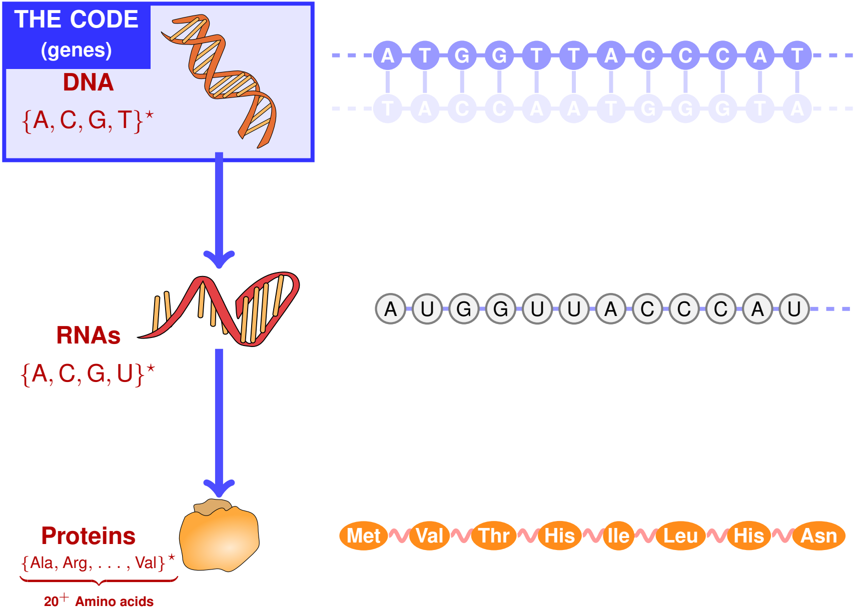


Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \dots, \text{Val}\}^\star}_{\textbf{20}^+ \textbf{ Amino acids}}$



Fundamental *dogma* of molecular biology



THE CODE (genes)

DNA

$\{A, C, G, T\}^\star$

A T G G T T A C C C A T

T A C C A A T G G G T A

RNAs

$\{A, C, G, U\}^\star$

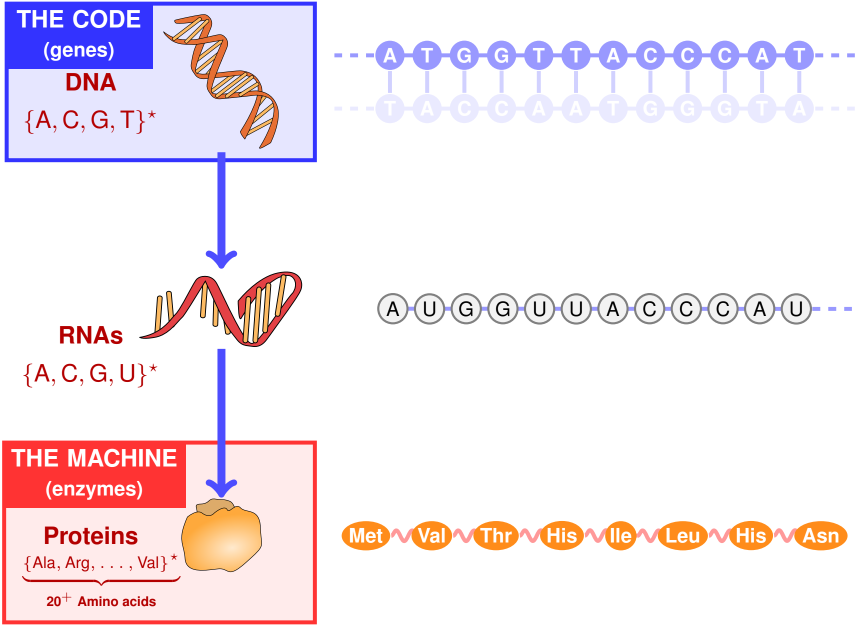
A U G G U U A C C C A U

Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

Met Val Thr His Ile Leu His Asn

Fundamental *dogma* of molecular biology



THE CODE
(genes)

DNA

$\{A, C, G, T\}^\star$

- - A T G G T T A C C C A T - - -

- - T A C C A A T G G G T A - - -

RNAs

$\{A, C, G, U\}^\star$

A U G G U U A C C C A U - - -

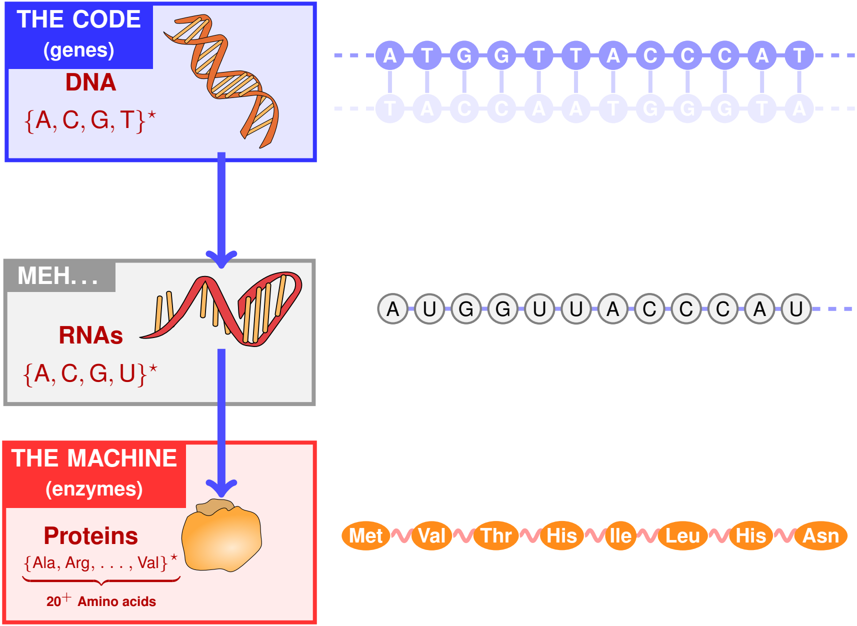
THE MACHINE
(enzymes)

Proteins

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

Met Val Thr His Ile Leu His Asn

Fundamental *dogma* of molecular biology



THE CODE (genes)

DNA

$\{A, C, G, T\}^\star$

- - - A T G G T T A C C C A T - - -

- - - T A C C A A T G G G T A - - -

MEH. . .

RNAs

$\{A, C, G, U\}^\star$

A U G G U U A C C C A U - - -

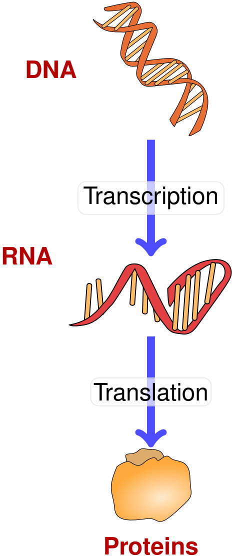
THE MACHINE (enzymes)

Proteins

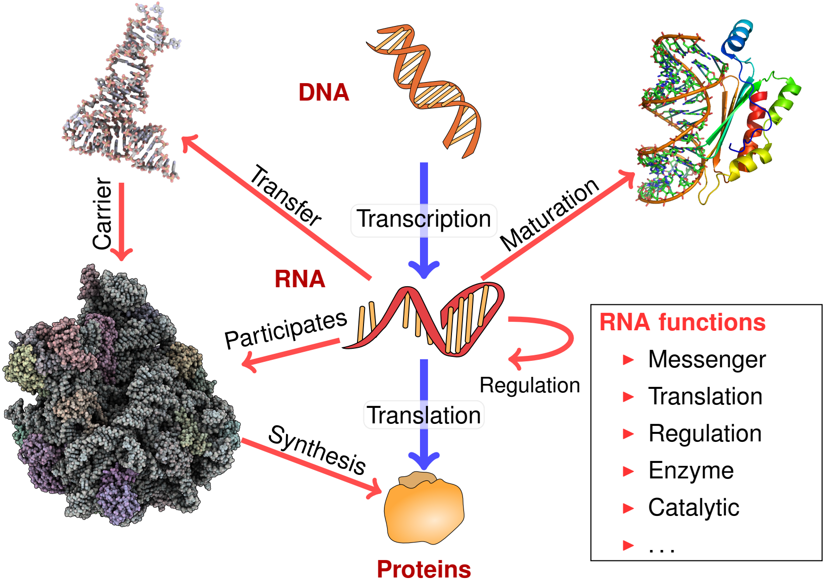
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^{+}\ \text{Amino acids}}$

Met Val Thr His Ile Leu His Asn

Fundamental *dogma* of molecular biology



Fundamental *dogma* of molecular biology



DNA

Transcription

RNA

Transfer

Maturation

Carrier

Participates

Regulation

Translation

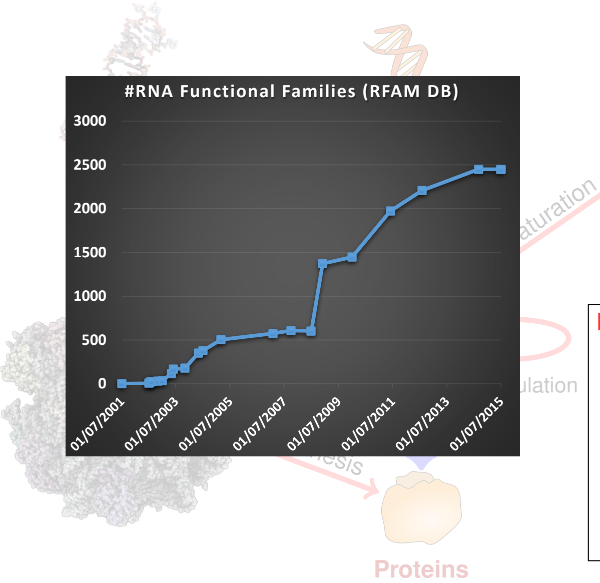
Synthesis

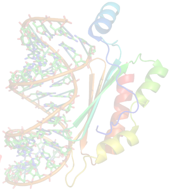
Proteins

RNA functions

- Messenger
- Translation
- Regulation
- Enzyme
- Catalytic
- ...

Fundamental *dogma* of molecular biology

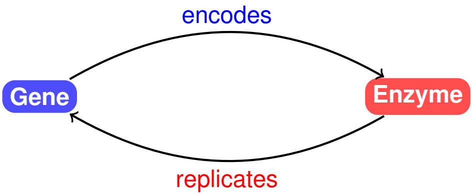




RNA functions

- Messenger
- Translation
- Regulation
- Enzyme
- Catalytic
- . . .

RNA world: Resolving the *chicken vs egg* paradox at the origin of life...

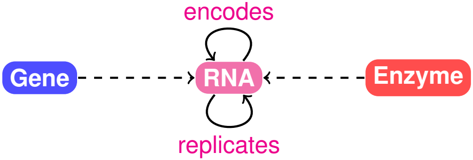


A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

[...] 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...



encodes

Gene

RNA

Enzyme

replicates

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*.

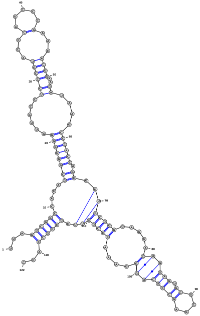
[. . .] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

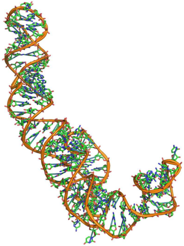
R. Dawkins. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

RNA structure(s)

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

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





Primary structure | Secondary structure | Tertiary structure

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

Definition (Secondary Structure)

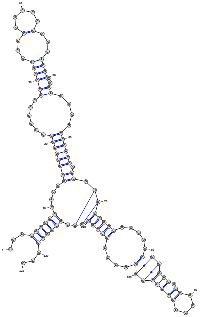
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

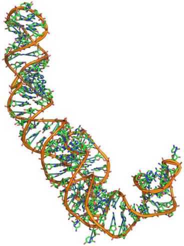
- **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints:** $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).

RNA structure(s)

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

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





Primary structure   Secondary structure   Tertiary structure

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

Definition (Secondary Structure)

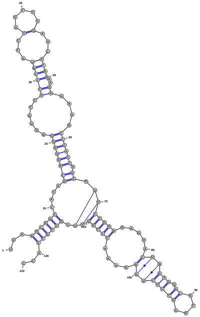
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

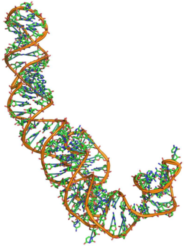
- **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints:** $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).

RNA structure(s)

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

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





Primary structure Secondary structure Tertiary structure

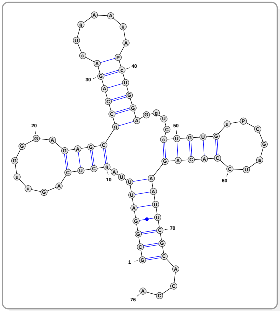
Source: 5s rRNA (PDBID: 1K73:B)

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:** $\nexists(i, j), (k, l) \in S$ 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*

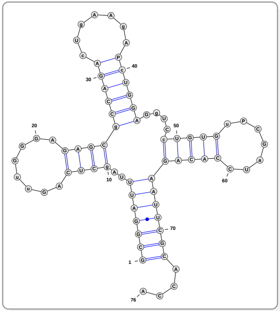
Supporting intuitions

Different representations

Common combinatorial structure

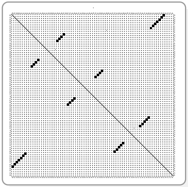
*Additional steric constraints

Various representations for a versatile biomolecule



Outer-planar graphs

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



Dot plots

Adjacency matrices*

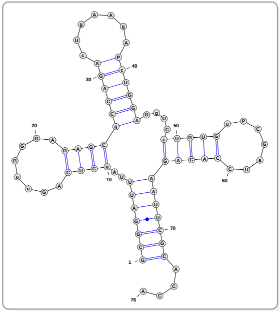
Supporting intuitions

Different representations

Common combinatorial structure

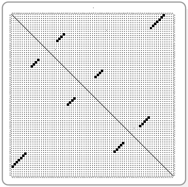
*Additional steric constraints

Various representations for a versatile biomolecule



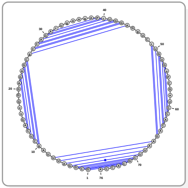
Outer-planar graphs

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



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

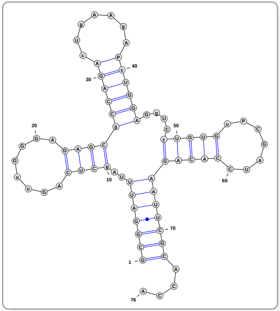
Supporting intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

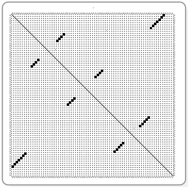


Outer-planar graphs

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

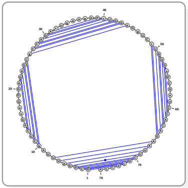
```
(((((((..((((........))))((((((.......))))))....(((((.......))))))))))))....
```

Motzkin words*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

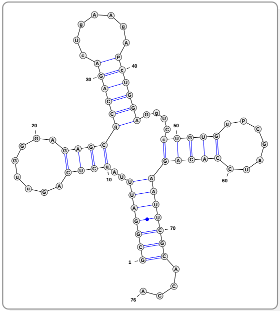
Supporting intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

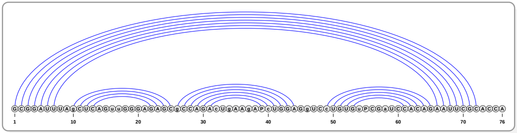


Outer-planar graphs

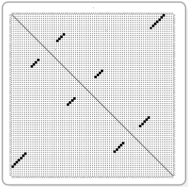
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

Motzkin words*

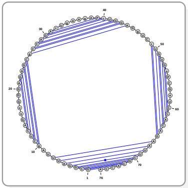


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

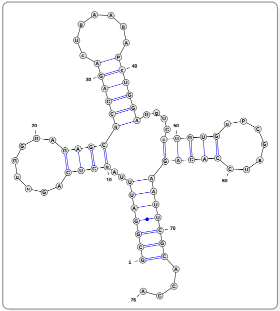
Supporting intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

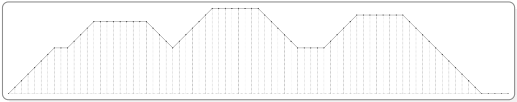


Outer-planar graphs

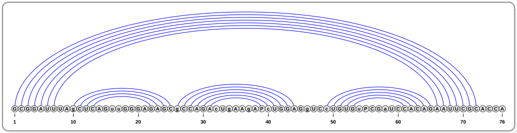
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

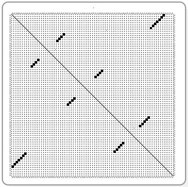
Motzkin words*



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

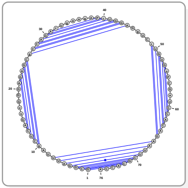


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

Supporting intuitions

Different representations

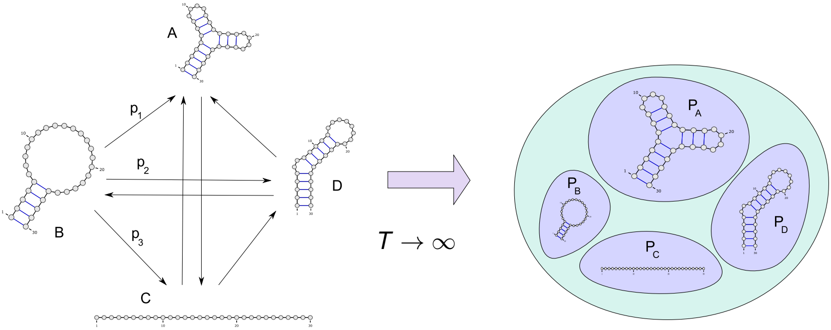
Common combinatorial structure

*Additional steric constraints

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:

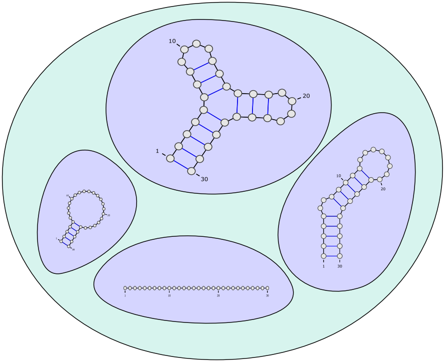
$$\mathbb{P}(S \mid w) = \frac{e^{\frac{-E_w(S)}{kT}}}{\mathcal{Z}_w} \qquad \text{and} \qquad \mathcal{Z}_w \equiv \sum_{S'} e^{\frac{-E_w(S')}{RT}} \quad \{\text{Partition function}\}$$

$E_w(S)$: **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



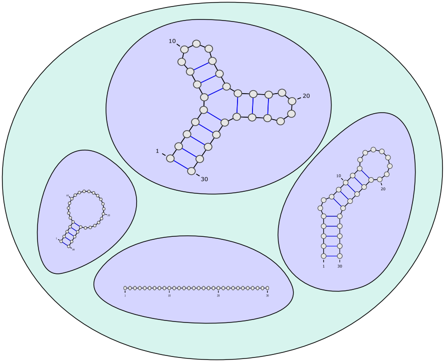
$T \to \infty$

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



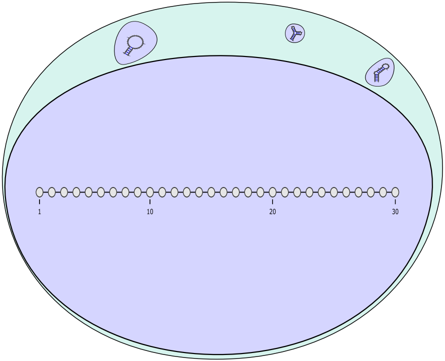
$T \rightarrow \infty$

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



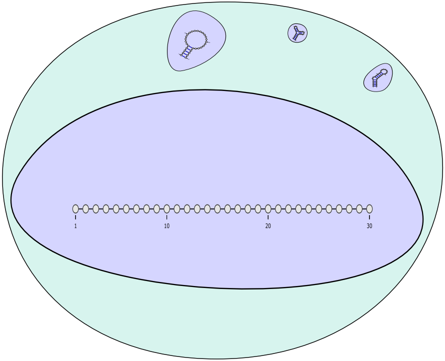
$T = 0$h

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



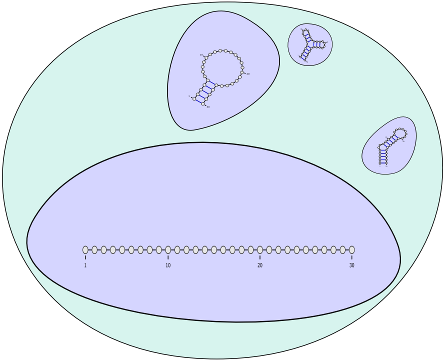
$T = 1\text{h}$

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



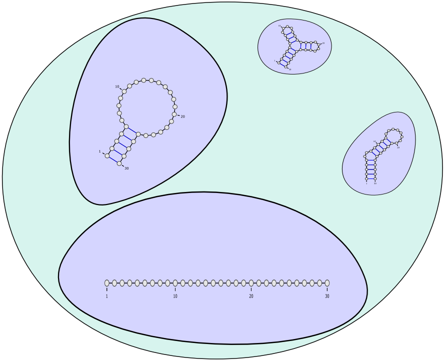
$T = 2\text{h}$

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



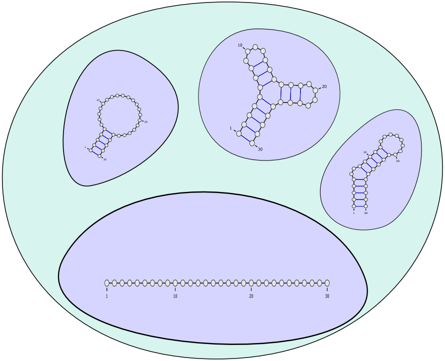
$T = 5\text{h}$

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



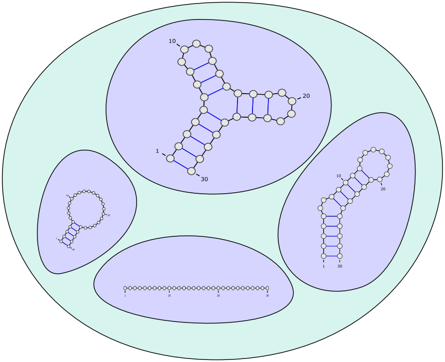
$T = 10\text{h}$

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



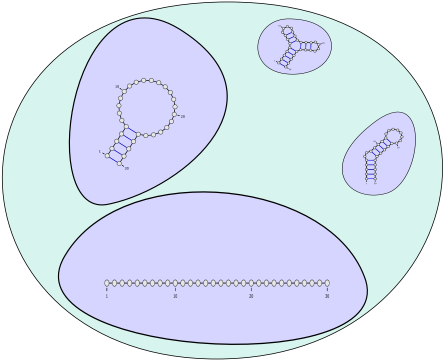
$T \to \infty$

mRNA half-life: ∼7h
(Mouse [Sharova2009])

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

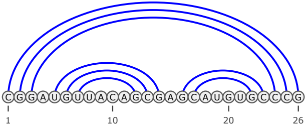
- **1978–1990s** Functional structure = Minimal Free-Energy
- **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- **2010s–????** Embracing kinetics

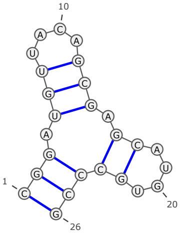


$T = 10\text{h}$

mRNA half-life: ∼7h
(Mouse [Sharova2009])

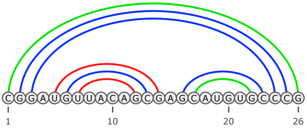
Problem statement

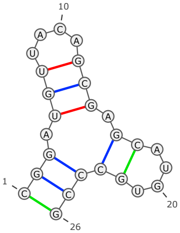




- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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

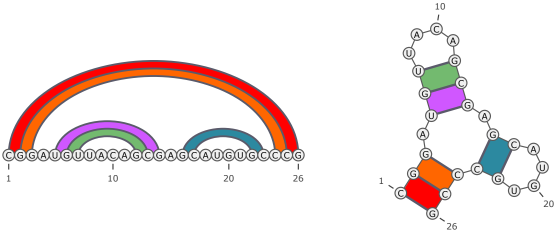
Problem statement





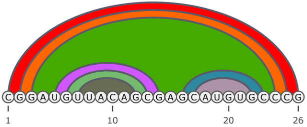
- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: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

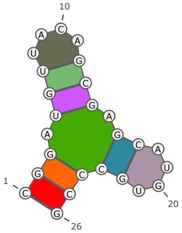
Problem statement



- **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

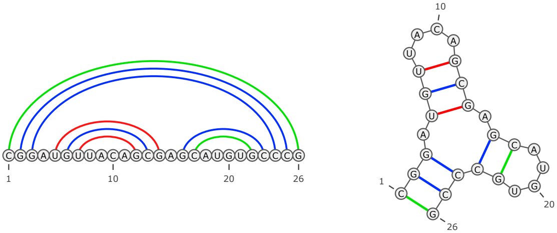
Problem statement





- **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

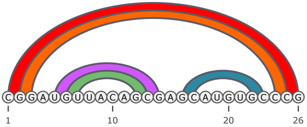
Problem statement

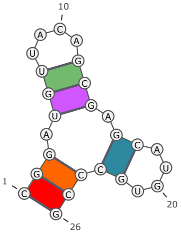


- **RNA structure** ***S*****:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$**:**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

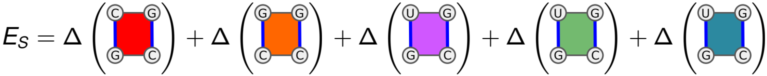
$$E_S = 2 \cdot \Delta \begin{pmatrix} \text{U} \\ | \\ \text{G} \end{pmatrix} + 4 \cdot \Delta \begin{pmatrix} \text{G} \\ | \\ \text{C} \end{pmatrix} + 2 \cdot \Delta \begin{pmatrix} \text{C} \\ | \\ \text{G} \end{pmatrix}$$

Problem statement

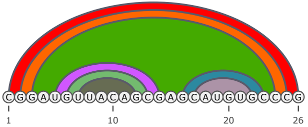


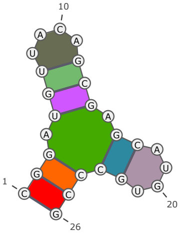


- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: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

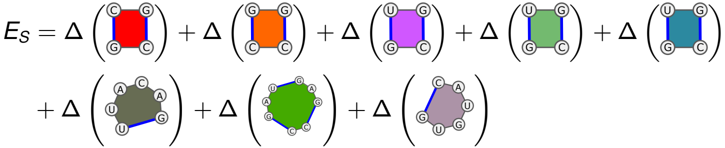


Problem statement

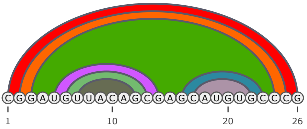


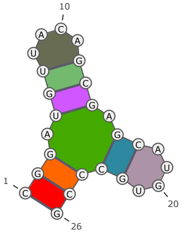


- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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



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

Definition (MFE-PREDICT(E) problem)

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

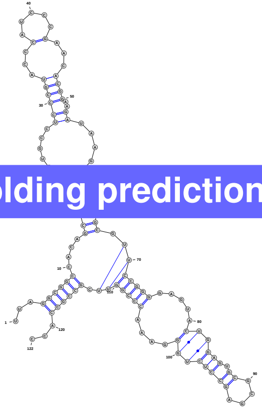
RNA folding: non-crossing matchings

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

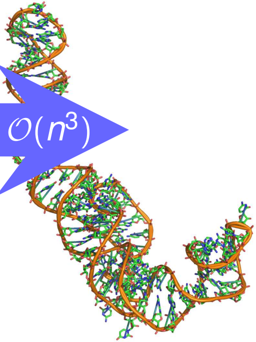
Structure = **Non-crossing** matching

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA

CC
```



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



Primary Structure

Secondary Structure

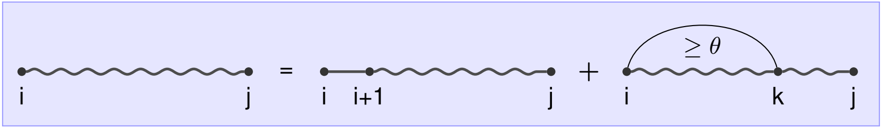
Tertiary Structure

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 $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ *time/memory using dynamic programming*



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

$\boldsymbol{N}_{i,j}$: Max #base-pairs over interval $[i, j]$

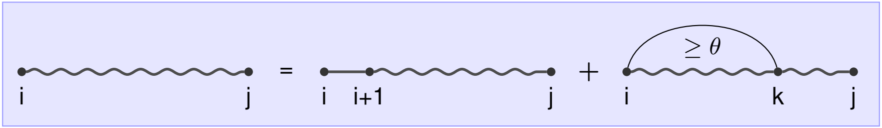
$$\boldsymbol{N}_{i,t} = 0, \quad \forall t \in [i, i+\theta]$$

$$\boldsymbol{N}_{i,j} = \min \begin{cases} \boldsymbol{N}_{i+1,j} & \{i \text{ unpaired}\} \\ \min\limits_{k=i+\theta+1}^{j} E_{i,k} + \boldsymbol{N}_{i+1,k-1} + \boldsymbol{N}_{k+1,j} & \{i \text{ paired to } k\} \end{cases}$$

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



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

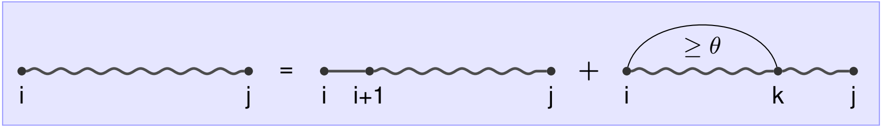
$\boldsymbol{C}_{i,j}$: Number of secondary structures compatible with interval $[i,j]$

$$\begin{aligned} \boldsymbol{C}_{i,t} &= \mathbf{1}, \quad \forall t \in [i, i+\theta] \\ \boldsymbol{C}_{i,j} &= \sum \left\{ \begin{array}{lr} \boldsymbol{C}_{i+1,j} & \{i \text{ unpaired}\} \\ \sum_{k=i+\theta+1}^{j} \mathbb{1}_{\text{comp.}(i,k)} \times \boldsymbol{C}_{i+1,k-1} \times \boldsymbol{C}_{k+1,j} & \{i \text{ paired to } k\} \end{array} \right. \end{aligned}$$

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



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

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

$$\begin{aligned} \mathcal{Z}_{i,t} &= \mathbf{1}, \quad \forall t \in [i, i+\theta] \\ \mathcal{Z}_{i,j} &= \sum \left\{ \begin{array}{cr} \mathcal{Z}_{i+1,j} & \{i \text{ unpaired}\} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{i,k}}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} & \{i \text{ paired to } k\} \end{array} \right. \end{aligned}$$

Dynamic programming (DP) for RNA folding

Many extensions:

- Comparative folding [Sankoff1985]
- Equilibrium base-pairing probabilities [McCaskill1990]
- Moments of additive features [Miklos2005,Ponty2011]
- Δ 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]

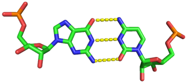
Made possible by:

- **Completeness**/**Unambiguity** of decomposition
  $\exists$ energy-preserving bijection between **derivations of DP scheme** and **search space**
- Objective function **additive** with respect to DP scheme

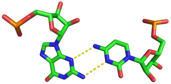
$\Rightarrow$ **Combinatorial Dynamic Programming**

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]**.

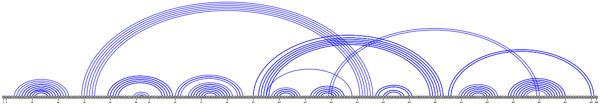


Canonical CG base-pair (WC/WC-Cis)



Non-canonical base-pair (Sugar/WC-Trans)

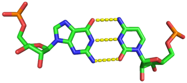
- **Pseudoknots:** Crossing sets of nested stable base-pairs



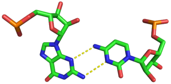
Group I Ribozyme (PDBID: 1Y0Q:A)

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]**.

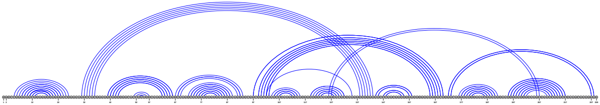


Canonical CG base-pair (WC/WC-Cis)



Non-canonical base-pair (Sugar/WC-Trans)

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



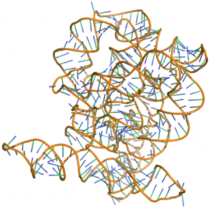
Group I Ribozyme (PDBID: 1Y0Q:A)

Including crossing interactions

- **Non-canonical base-pairs:** Lead to **local crossings** and **promiscuity**
  Any ba se-pair (A-U), (C-G), (G-U)}
  OR int dard wa (WC/WC-Cis) [Leontis 2001]

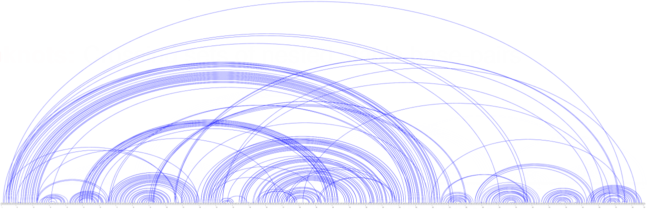
Canonical (WC/WC-Cis)   Non-canonical base-pair (Sugar/WC-Trans)

- **Pseudoknots**



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

Example: Group II Intron (PDB ID: 3IGI)

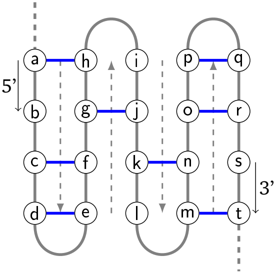


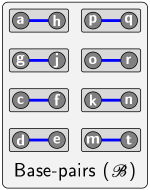
Group I Ribozyme (PDBID: 1Y0Q:A)

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



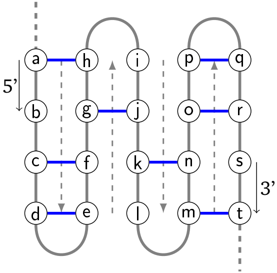


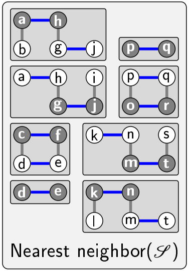
Solved in $\mathcal{O}(n^3)$ [Tabaska1998]
(Max-weighted matching)
Unrealistic!

Energy models

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

- **Base-pair model $\mathcal{B}$**: Nucleotides (w_i, w_j) at (i, j)
  $\rightarrow \Delta_{\mathcal{B}}(w_i, w_j)$
- **Nearest-neighbor model $\mathcal{N}$:** Nucl. at (i, j) and $(i+1, j\text{-}1)$ + partners (or $\varnothing$)
  $\rightarrow \Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}})$
- **Stacking pairs model $\mathcal{S}$:** Nucl. at (i, j) and $(i+1, j\text{-}1)$ **only if** latter paired
  $\rightarrow \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$





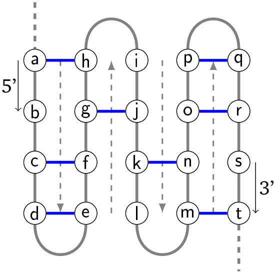
NP-hard [Lyngso2000,Akutsu2000]

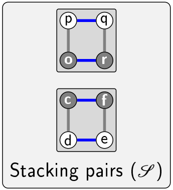
Too expressive?

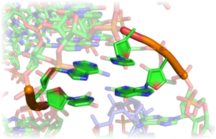
Energy models

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

- **Base-pair model $\mathcal{B}$**: Nucleotides (w_i, w_j) at (i, j)
  $\rightarrow \Delta_{\mathcal{B}}(w_i, w_j)$
- **Nearest-neighbor model $\mathcal{N}$:** Nucl. at (i, j) and $(i+1, j\text{-}1)$ + partners (or $\varnothing$)
  $\rightarrow \Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}})$
- **Stacking pairs model $\mathcal{S}$:** Nucl. at (i, j) and $(i+1, j\text{-}1)$ **only if** latter paired
  $\rightarrow \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$





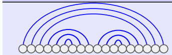
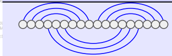
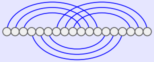


Captures stablest motifs

Still NP-hard [Lyngso2004]

. . . but PTAS [Lyngso2004]

The full monty

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov1980]	P [Ieong2003]	P [Zuker1981]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong2003]	NP-Hard [Ieong2003]
Planar	Approx.	2-approx. ≈[Ieong2003]	2-approx. [Ieong2003]	???
	Comp.	P [Tabaska1998]	NP-Hard [Lyngso2004] (any* Δ model) [Sheikh2012]	NP-Hard [Lyngso2000] [Akutsu2000]
General	Approx.	Duh...	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngso2004] 1/5 (any Δ model) [Sheikh2012]	APX-Hard [Sheikh2012]

Missing:

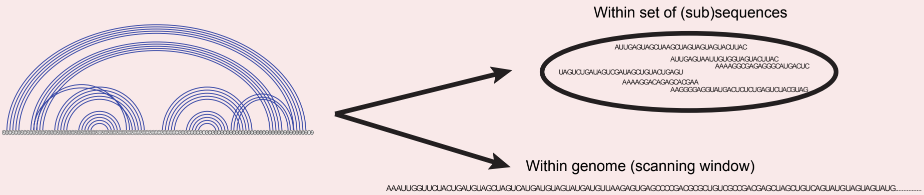
- Base-pair maximization in planar model (probably NP-hard)
- Partition function (mostly in P cases), Boltzmann-Gibbs sampling
- **Relevance** of approximation???

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



Within set of (sub)sequences

AUUGAGUAGCUAAGCUAGUAGUAGUACUUAC

AUUGAGUAAUUGUGGUAGUACUUAC

AAAAGGCGAGAGGGCAUGACUC

UAGUCUGAUAGUCGAUAGCUGUACUGAGU

AAAAGGACAGAGCACGAA

AAGGGGAGGUAUGACUCUCUGAGUCUACGUAG

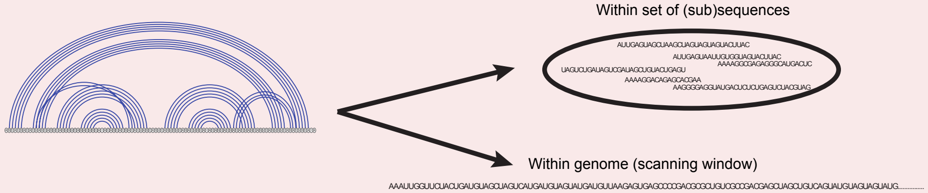
Within genome (scanning window)

AAAUUGGUUCUACUGAUGUAGCUAGUCAUGAUGUAGUAUGAUGUUAAGAGUGAGCCCCGACGCGCUGUCGCCGACGAGCUAGCUGUCAGUAUGUAGUAGUAUG...............

Structure prediction by homology modelling

Sequence structure alignment for ncRNA search and homology-modeling

Search for novel ncRNA instances



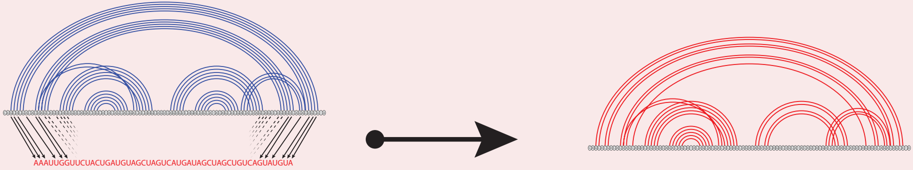
Within set of (sub)sequences

AUUGAGUAGCUAAGCUAGUAGUAGUACUUAC
AUUGAGUAAUUGUGGUAGUACUUAC
AAAAGGCGAGAGGGCAUGACUC
UAGUCUGAUAGUCGAUAGCUGUACUGAGU
AAAAGGACAGAGCACGAA
AAGGGGAGGUAUGACUCUCUGAGUCUACGUAG

Within genome (scanning window)

AAAUUGGUUCUACUGAUGUAGCUAGUCAUGAUGUAGUAUGAUGUUAAGAGUGAGCCCCGACGCGCUGUCGCCGACGAGCUAGCUGUCAGUAUGUAGUAGUAUG..............

Structure prediction by homology modelling

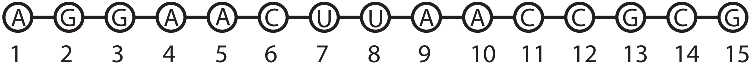


AAAUUGGUUCUACUGAUGUAGCUAGUCAUGAUAGCUAGCUGUCAGUAUGUA

Context: Multiple Structural levels

Primary Structure

- Represents nucleotides sequence
- No interaction **Boring. . .**

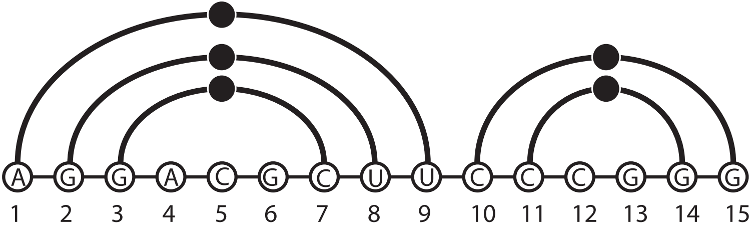


Context: Multiple Structural levels

Secondary Structure

- Scaffold/blueprint for 3D
- Only includes non-crossing canonical interactions (WC/WC cis, GC/AU/GU)
- Any nucleotide has ≤ 1 partner

Better…



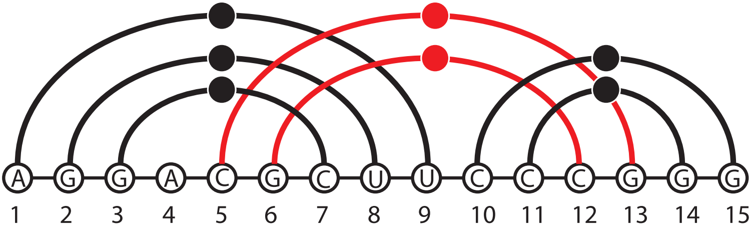
Context: Multiple Structural levels

Secondary Structure with Pseudoknots

- Includes all canonical crossing interactions
- Any nucleotide has ≤ 1 partner

Wow...

Pseudoknots play a major part in the architecture of some RNAs
Yet they are hard to handle algorithmically!

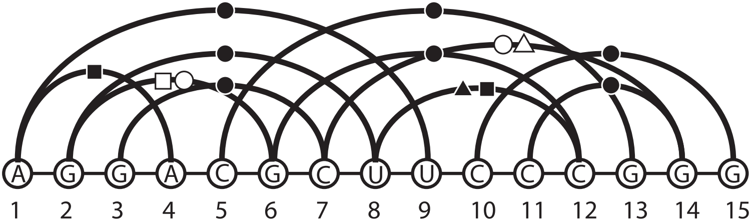


Context: Multiple Structural levels

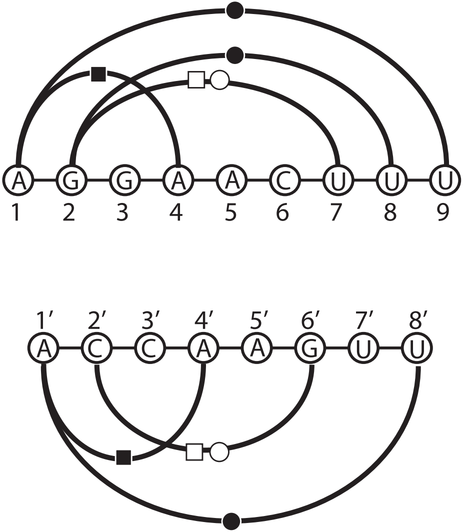
Extended secondary structure

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

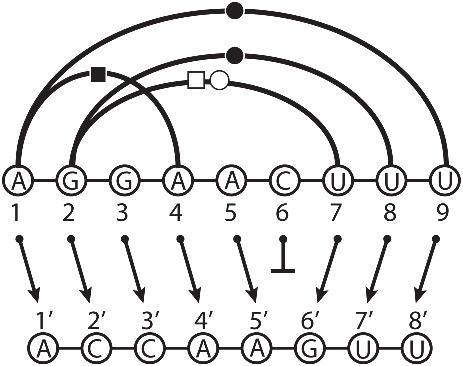
Now we're talking!



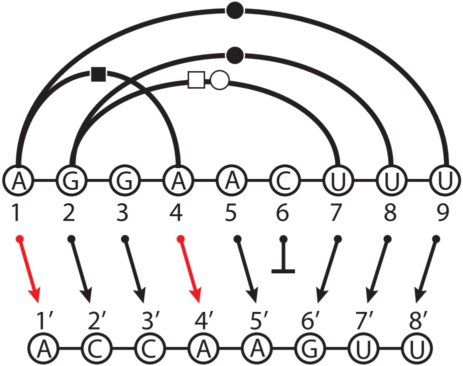
Sequence-structure alignment



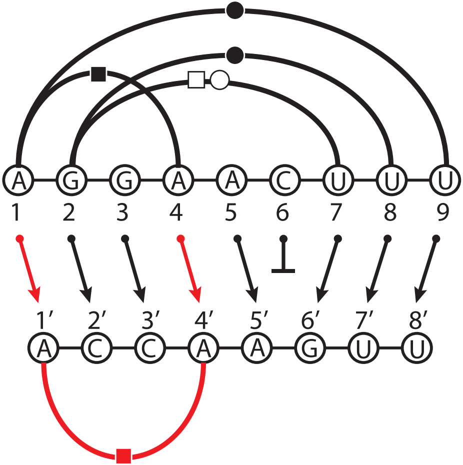
Sequence-structure alignment



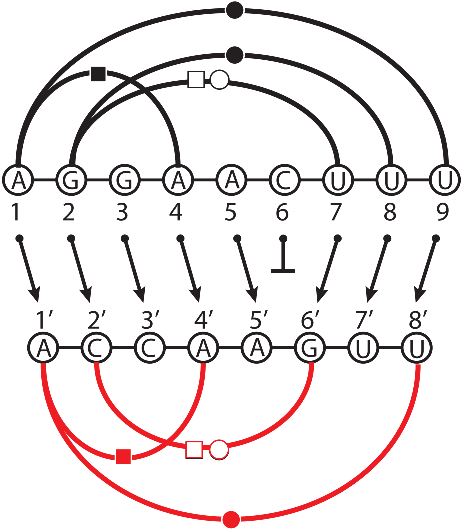
Sequence-structure alignment



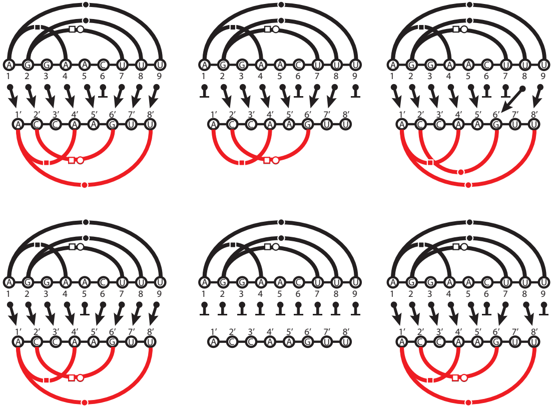
Sequence-structure alignment



Sequence-structure alignment



Sequence-structure alignment

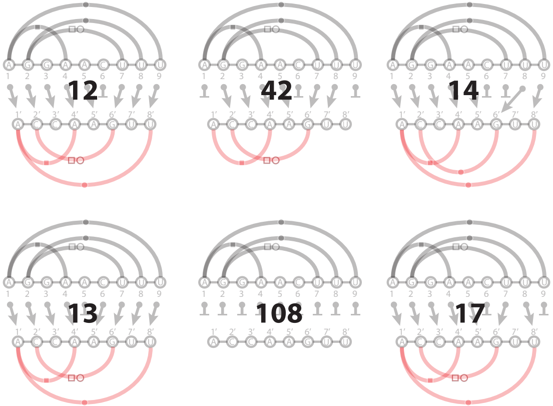


Sequence-structure alignment Problem

Input: (Extended) Secondary structure S + Sequence ω
Output: Minimal-cost alignment (mapping subject to constraints)

Variant: Affine gap cost model

Sequence-structure alignment



Sequence-structure alignment Problem

Input: (Extended) Secondary structure S + Sequence ω

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

Variant: Affine gap cost model

Complexity of structure-sequence alignment

$n =$ Structure Length, $m =$ Sequence Length

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

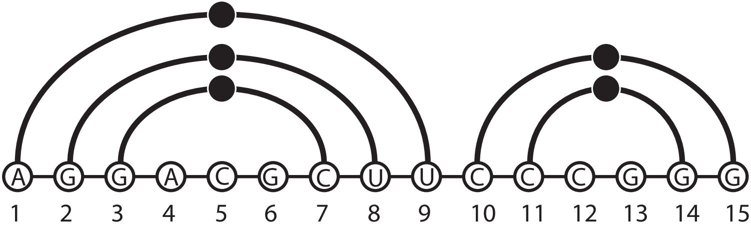
Jiang *et al.* 2001

Complexity of structure-sequence alignment

$n =$ Structure Length, $m =$ 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

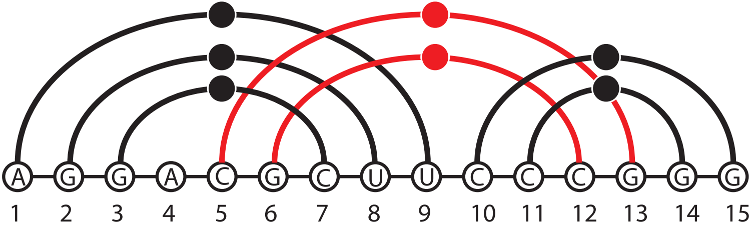


Complexity of structure-sequence alignment

$n =$ Structure Length, $m =$ 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

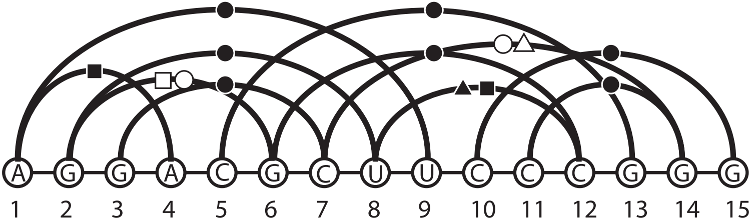


Complexity of structure-sequence alignment

$n =$ Structure Length, $m =$ Sequence Length

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

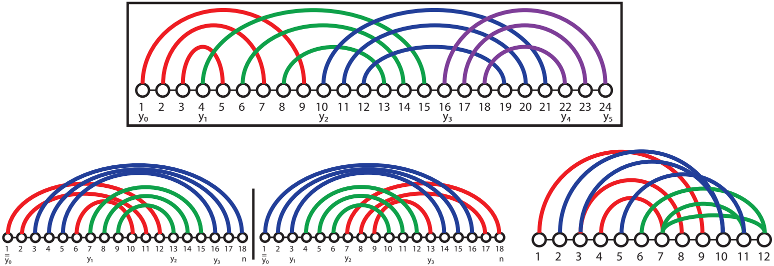
Jiang *et al.* 2001



Complexity of struct.-seq. alignment: Polynomial classes

$n =$ Structure Length, $m =$ Sequence Length, $b =$ #Bands

Standard Pseudoknots	$O(n \cdot m^b)$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^3)$

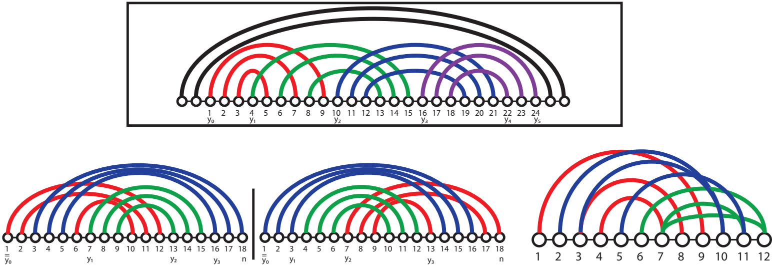


Han *et al.* 2008

Complexity of struct.-seq. alignment: Polynomial classes

$n =$ Structure Length, $m =$ Sequence Length, $b =$ #Bands

Standard Pseudoknots	$O(n \cdot m^{b})$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^{3})$

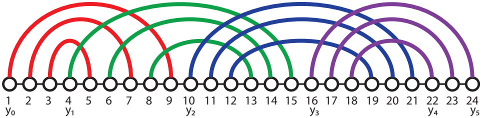


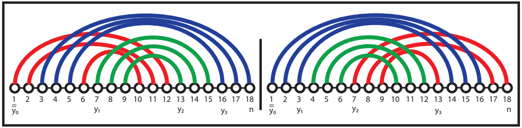
Han *et al.* 2008

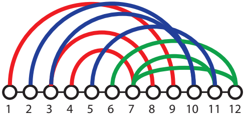
Complexity of struct.-seq. alignment: Polynomial classes

$n =$ Structure Length, $m =$ Sequence Length, $b =$ #Bands

Standard Pseudoknots	$O(n \cdot m^b)$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^3)$





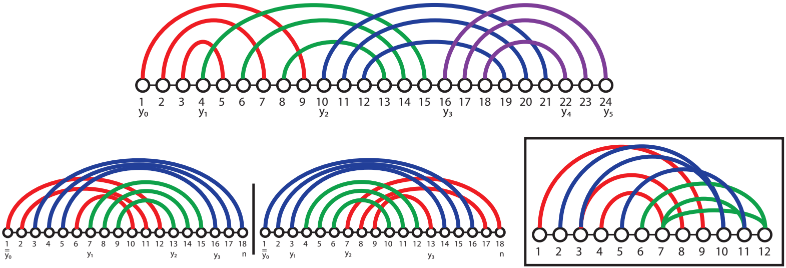


Wong *et al.* 2011

Complexity of struct.-seq. alignment: Polynomial classes

$n =$ Structure Length, $m =$ Sequence Length, $b =$ #Bands

Standard Pseudoknots	$O(n \cdot m^{b})$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^{3})$

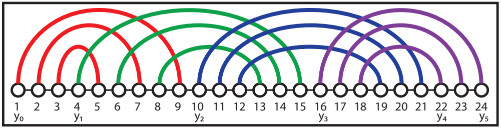


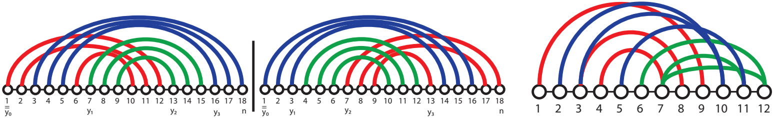
Wong *et al.* 2012

Complexity of struct.-seq. alignment: Polynomial classes

$n =$ Structure Length, $m =$ Sequence Length, $b =$ #Bands

Standard Pseudoknots	$O(n \cdot m^b)$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^3)$

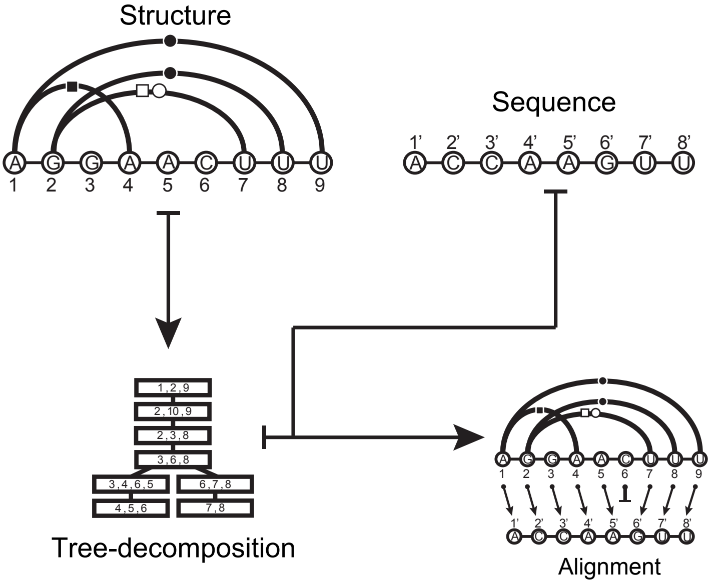




+ Other $O(n.m^4)/O(n.m^6)$ classes based on folding DP schemes

[Möhl/Will/Backofen 2009]

Outline of general parameterized approach



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

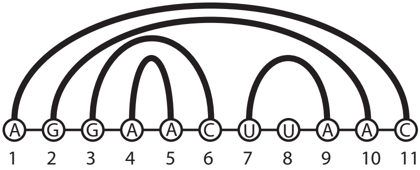
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

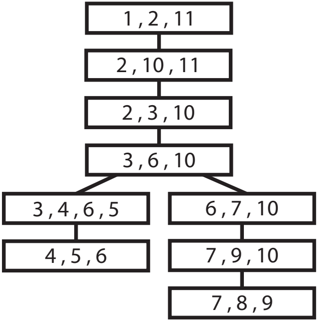
Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

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

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





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

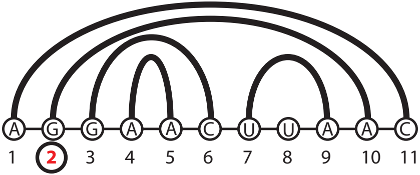
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

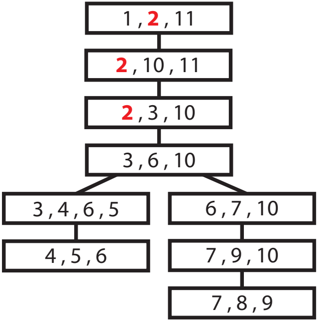
Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

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

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





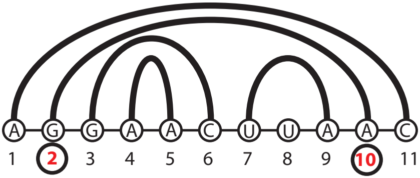
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

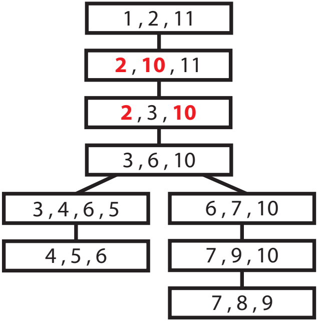
Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

Sets of structure-side positions (**bags** $\{\mathcal{B}_i\}$), 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 $x \in \mathcal{B} \cap \mathcal{B}'$, than x is in **every bag** $\mathcal{B}''$ on the path from $\mathcal{B}$ to $\mathcal{B}'$




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

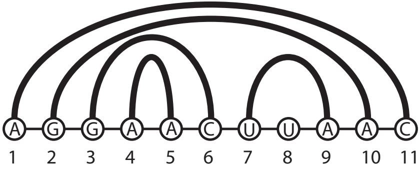
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

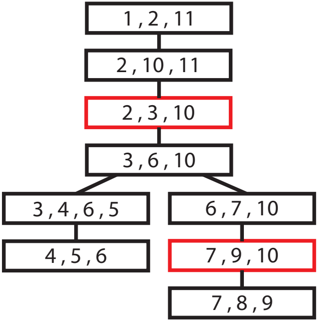
Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

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

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





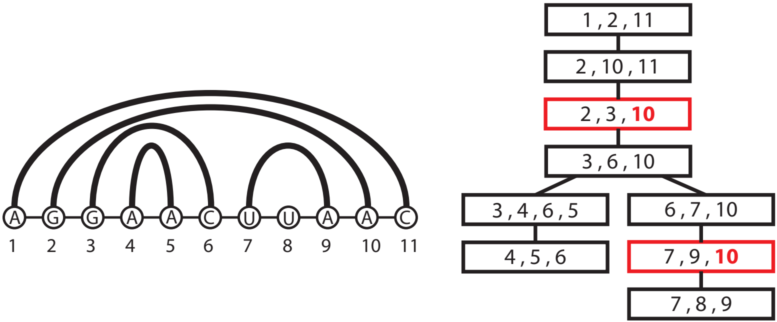
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

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

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



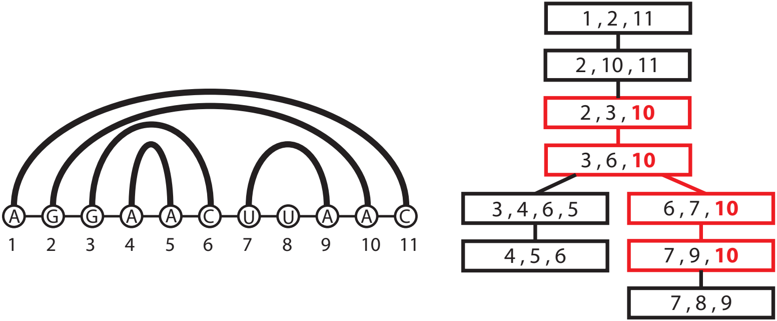
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

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

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



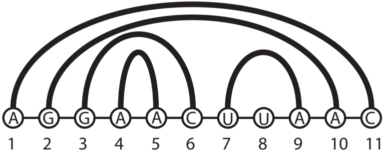
Tree decomposition of RNA structure [*Rinaudo et al. 2012*]

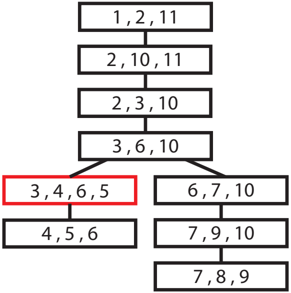
Structure-centric alignment ⇒ Constraints

- Adjacent positions in structure → **Precedence**
- Paired positions → **Both partners needed to assign score**

Sets of structure-side positions (**bags** $\{\mathcal{B}_i\}$), 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 $x \in \mathcal{B} \cap \mathcal{B}'$, than x is in **every bag** $\mathcal{B}''$ on the path from $\mathcal{B}$ to $\mathcal{B}'$





1 , 2 , 11

2 , 10 , 11

2 , 3 , 10

3 , 6 , 10

3 , 4 , 6 , 5

4 , 5 , 6

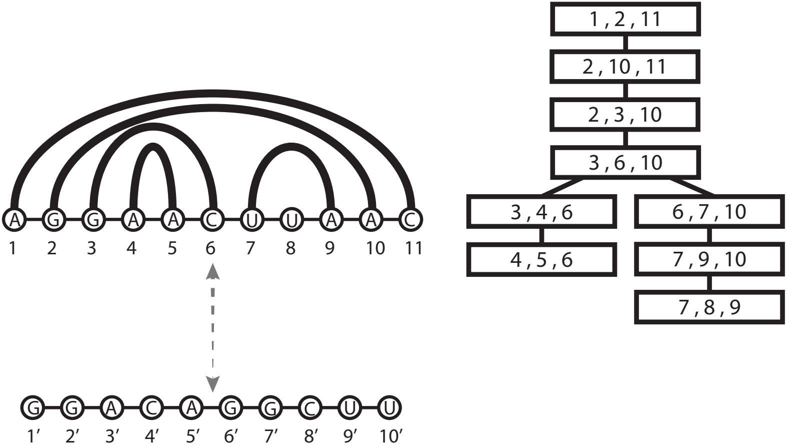
6 , 7 , 10

7 , 9 , 10

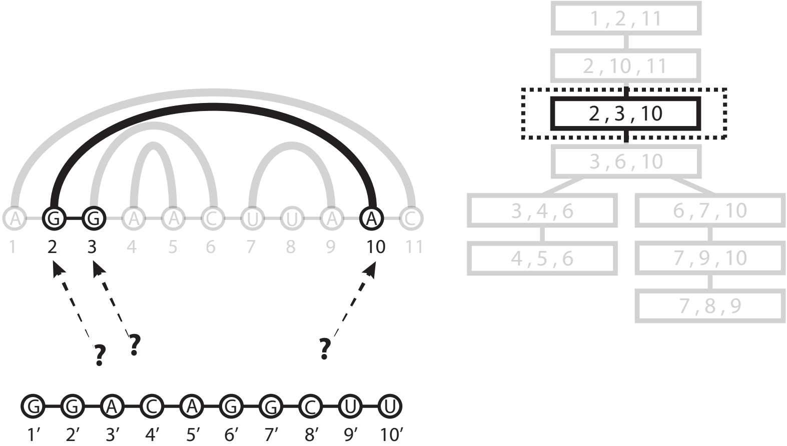
7 , 8 , 9

Width ***k*** **=** Size of biggest bag minus one.

Tree-Decomposition-based Alignment



Tree-Decomposition-based Alignment



(Fixed-parameter tractable??) algorithm [*Rinaudo et al. 2012*]

Theorem

Input: Structure S of length n; Sequence w of length m $\rightarrow$ Tree dec. of S, width k
Best alignment computed in $\mathcal{O}\left(n.m^{k+1}\right)/\mathcal{O}\left(n.m^{k}\right)$ time/space $\rightarrow$ **not FPT!**

Dynamic programming equation:

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

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)

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.

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

n → Structure length
m → Sequence length
k → Class-specific structural parameter

Tree Decomposition vs The World [*Rinaudo et al. 2012*]

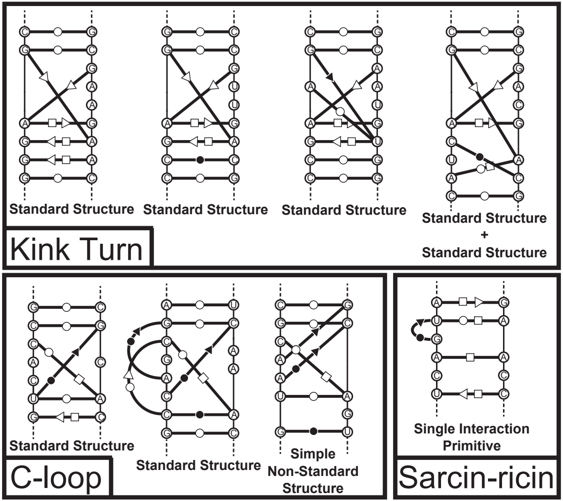
Specialized complexities

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

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

n → Structure length
m → Sequence length
k → Class-specific structural parameter

New classes of structures *[Rinaudo et al. 2012]*


Standard Structure

Standard Structure

Standard Structure

Standard Structure
+
Standard Structure

Kink Turn

Standard Structure

Standard Structure

Simple
Non-Standard
Structure

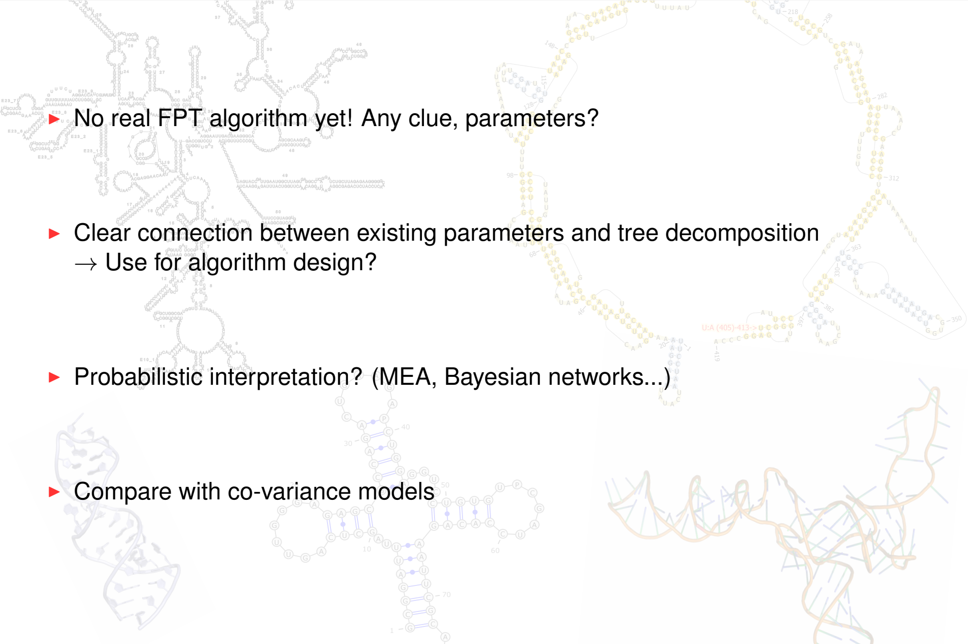
C-loop

Single Interaction
Primitive

Sarcin-ricin

Recursive Classical Structures	$O(n \cdot m^{k+2})$
Standard Structures	$O(n \cdot m^{k})$
Simple Non-Standard Structures	$O(n \cdot m^{k+1})$
Extended Triple Helices	$O(n \cdot m^{3})$

Half-time summary

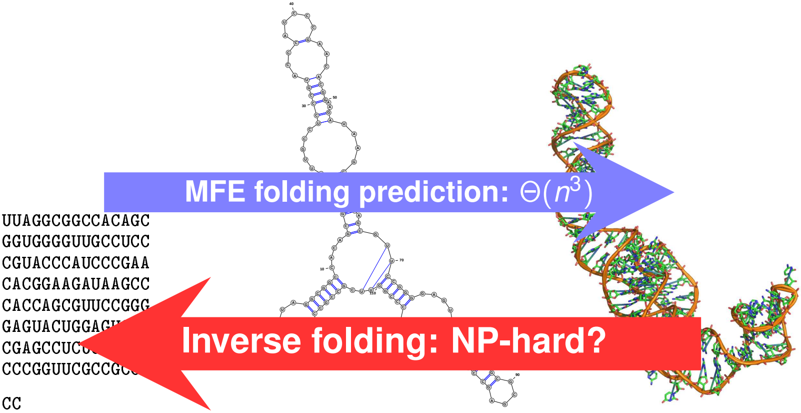


- No real FPT algorithm yet! Any clue, parameters?
- Clear connection between existing parameters and tree decomposition
  → Use for algorithm design?
- Probabilistic interpretation? (MEA, Bayesian networks...)
- Compare with co-variance models

Part. III: Designing RNAs

RNA inverse folding

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



UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAG
CGAGCCUC
CCCGGUUCGCCGC

CC

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

Inverse folding: NP-hard?

Primary Structure

Secondary Structure

Structure Tertiaire

5s rRNA (PDBID: 1K73:B)

RNA Inverse Folding

$\mathcal{M} =$ energy model

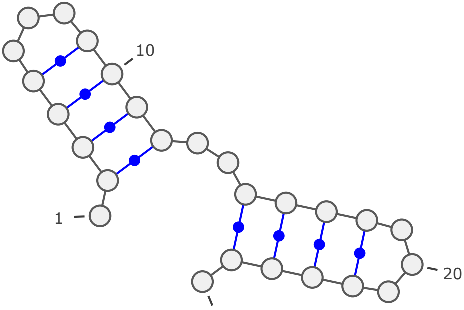
Definition (INVERSE-FOLDING(E) problem)

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

$$\forall S' \in \mathcal{S}|w| \setminus \{S\} : \; E_{w,S'} \geq E w, S + \Delta$$

or $\varnothing$ if no such sequence exists.

No (obvious?) optimal substructure property:



RNA Inverse Folding

$\mathcal{M} =$ energy model

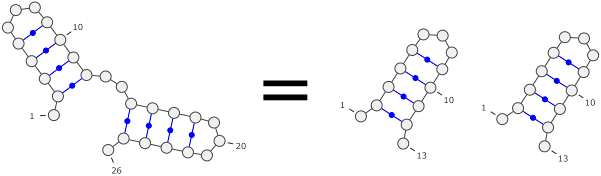
Definition (Inverse-folding(E) problem)

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

$$\forall S' \in \mathcal{S}|w| \setminus \{S\} : \; E_{w,S'} \geq E w, S + \Delta$$

or $\varnothing$ if no such sequence exists.

No (obvious?) optimal substructure property:



RNA Inverse Folding

$\mathcal{M}$ = energy model

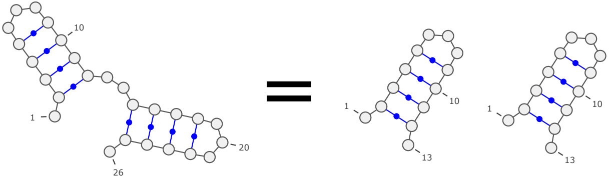
Definition (INVERSE-FOLDING(E) problem)

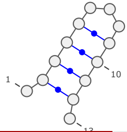
Input: Secondary structure S + Energy distance $\Delta > 0$.
Output: RNA sequence $w \in \Sigma^\star$ such that:

$$\forall S' \in \mathcal{S}|w| \setminus \{S\} :\ E_{w,S'} \geq E w, S + \Delta$$

or $\varnothing$ if no such sequence exists.

No (obvious?) optimal substructure property:





folds ←

AAGAGUCGCUCUC

RNA Inverse Folding

$\mathcal{M}$ = energy model

Definition (INVERSE-FOLDING(E) problem)

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

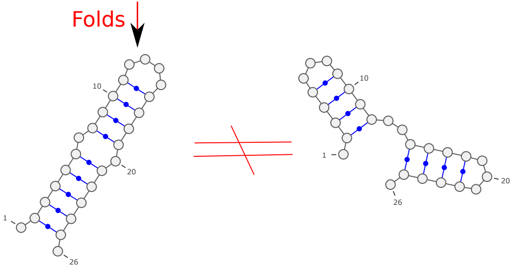
$$\forall S' \in \mathcal{S}|w| \setminus \{S\} : \; E_{w,S'} \geq E w, S + \Delta$$

or $\varnothing$ if no such sequence exists.

No (obvious?) optimal substructure property:

AAGAGUCGCUCUCAAGAGUCGCUCUC

Folds



RNA Design Problem

$\mathcal{M}$ = energy model

Definition (INVERSE-FOLDING(E) problem)

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

$$\forall S' \in \mathcal{S}|w| \setminus \{S\} : \; E_{w,S'} \geq Ew, S + \Delta$$

or $\varnothing$ 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**!... **CO-NP**???
- **Reason:** Non locality, no theoretical frameworks, too many parameters...

$\Rightarrow$ **Stick to a simplified model!**

RNA Design Problem (simplified)

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

Problem (Inverse-Folding(Σ) problem)

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

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

or $\varnothing$ if no such sequence exists.

Designable(Σ): All designable structures

RNA Design Problem (simplified)

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

Problem (INVERSE-FOLDING(Σ) problem)

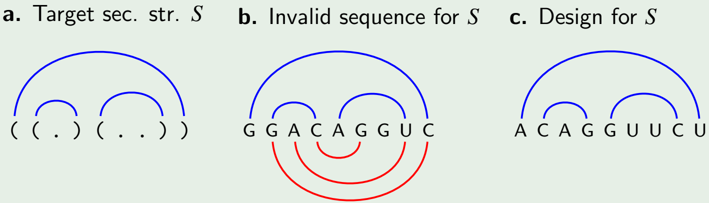
Input: *Secondary structure S*
Output: *RNA sequence $w \in \Sigma^\star$ — called a design for S — such that:*

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

or $\varnothing$ if no such sequence exists.

Designable(Σ)**:** All designable structures

Example

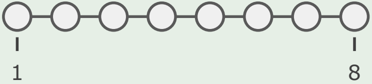


Our Results: Definitions and notations

Given a secondary structure S:

- Unpaired$_S$ = Set of all unpaired positions of S.
- S is **saturated** $\Leftrightarrow$ Unpaired$_S = \varnothing$.
  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



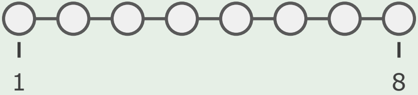
Unpaired$_S = \{4, 8\}$

Our Results: Definitions and notations

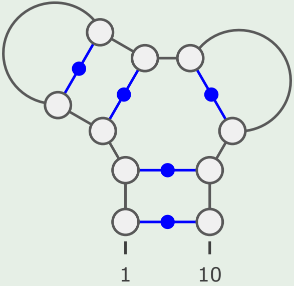
Given a secondary structure S:

- Unpaired_S = Set of all unpaired positions of S.
- S is **saturated** ⇔ $\text{Unpaired}_S = \varnothing$.
  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



Unsaturated



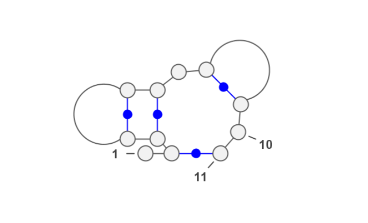
Saturated

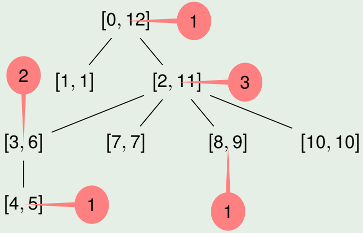
Our Results: Definitions and notations

Given a secondary structure S:

- Unpaired_S = Set of all unpaired positions of S.
- S is **saturated** $\Leftrightarrow \text{Unpaired}_S = \varnothing$.
  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





$D(S) = 3$

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≤ 2 + empty structures ;

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

Our Results: Designability over Restricted Alphabets

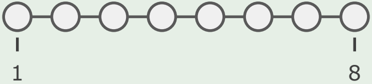
$\Sigma_{c,u} =$ Alphabet with c pairs of complementary bases and u unpairable bases.

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.

Example



Our Results: Designability over Restricted Alphabets

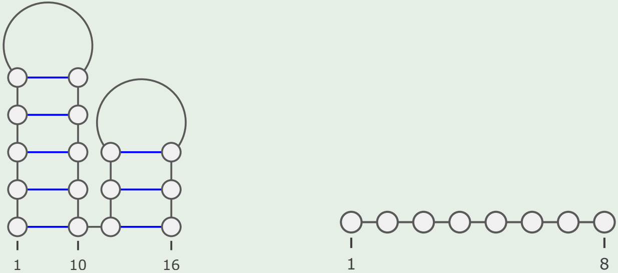
$\Sigma_{c,u}$ = Alphabet with *c* pairs of complementary bases and *u* unpairable bases.

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

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

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

Example



Our Results: Designability over Restricted Alphabets

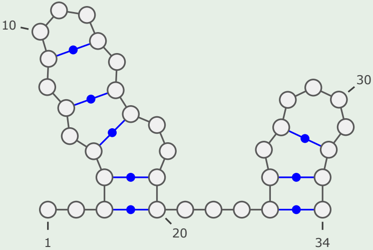
$\Sigma_{c,u} =$ Alphabet with c pairs of complementary bases and u unpairable bases.

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

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

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

Example



+ 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

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

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

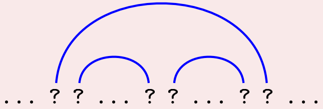
R2 $\Sigma_{1,0} \Rightarrow$ Designable = Saturated with degree$\leq$ 2 + empty structures ;

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

Question: Why not degree 3?

Proof.

Within an internal node:



□

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;

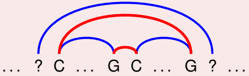
R2 $\Sigma_{1,0} \Rightarrow$ Designable $=$ Saturated with degree$\leq$ 2 + empty structures ;

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

Question: Why not degree 3?

Proof.

Within an internal node:



Either we get a repeat…

□

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;

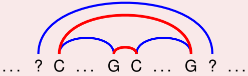
R2 $\Sigma_{1,0} \Rightarrow$ Designable = Saturated with degree≤ 2 + empty structures ;

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

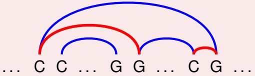
Question: Why not degree 3?

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

$\Sigma_{c,u} =$ 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≤ 2 + empty structures ;

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

This can be easily generalized to:

Lemma

For any structure S in Designable$(\Sigma_{c,u})$, $D(S) \leq 2c$.

Our Results: Designability over the Complete Alphabet

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

Without unpaired position → complete characterization:

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

With unpaired positions → 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: Designability over the Complete Alphabet

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

Without unpaired position → complete characterization:

R4 $\Sigma_{2,0} \Rightarrow$ Saturated Designable = Degree≤ 4.

With unpaired positions → partial characterization:

R5 (Necessary) Designable structure cannot contain "*a multiloop of degree* ≥ 5" (motif m_5) or "*a multiloop with unpaired position of degree* ≥ 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: Designability over the Complete Alphabet

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

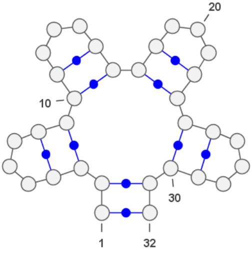
Without unpaired position → complete characterization:

R4 $\Sigma_{2,0} \Rightarrow$ Saturated Designable = Degree≤ 4.

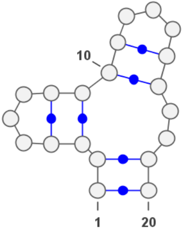
With unpaired positions → partial characterization:

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

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



m5



m3o

Our Results: Designability over the Complete Alphabet

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

Without unpaired position → complete characterization:

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

With unpaired positions → partial characterization:

R5 (Necessary) Designable structure cannot contain "*a multiloop of degree* ≥ 5" (motif m_5) or "*a multiloop with unpaired position of degree* ≥ 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 → G · C;
- white → C · G;
- grey → 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 $= \varnothing$

Our Results: Separated Coloring

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

- black → G · C;
- white → C · G;
- grey → 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 $= \varnothing$

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:

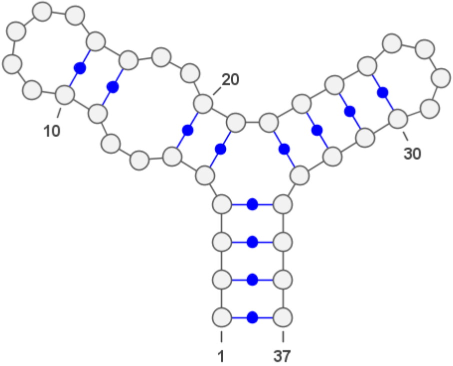
1. each internal node has at most one black, one white and two grey children;
2. a grey node has at most one grey child;
3. a black node does not have a white child; and
4. a white node does not have a black child.

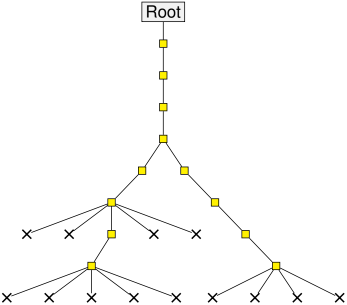
Level of a node = #black nodes $-$ #white nodes on the path to the root.

Separated coloring: Levels of grey nodes $\cap$ Levels of unpaired nodes $= \varnothing$

Our Results: Separated Coloring (example)

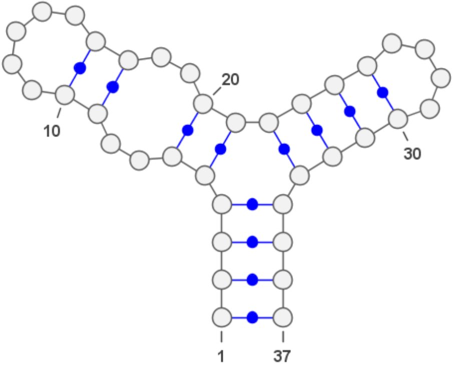
Descendant restrictions: Any node → ≤ 1 black & ≤ 1 White & ≤ 2 Grey;
Grey → 0/1 Grey; Black → 0 White; White → 0 Black.
(● → GC ○ → CG ● → AU|UA × → U)

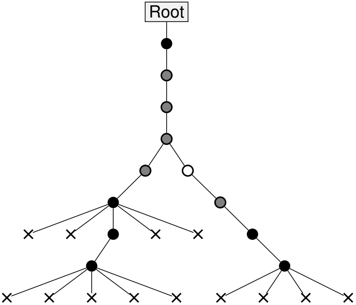




Our Results: Separated Coloring (example)

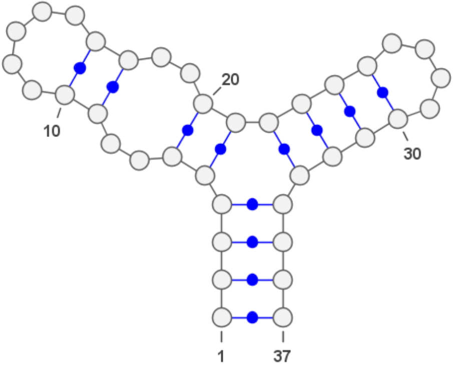
Descendant restrictions: Any node → ≤ 1 black & ≤ 1 White & ≤ 2 Grey;
Grey → 0/1 Grey; Black → 0 White; White → 0 Black.
(● → GC ○ → CG ● → AU|UA × → U)

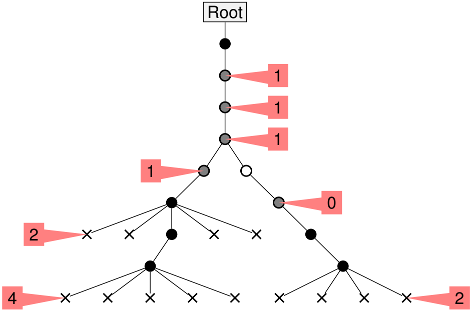




Our Results: Separated Coloring (example)

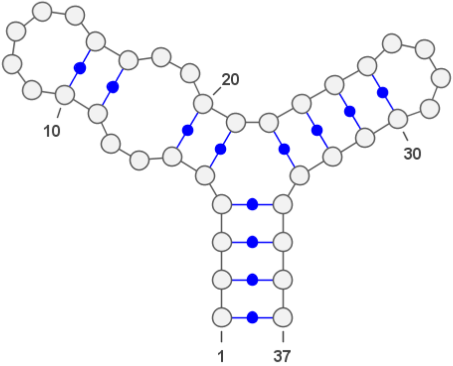
Descendant restrictions: Any node → ≤ 1 black & ≤ 1 White & ≤ 2 Grey;
Grey → 0/1 Grey; Black → 0 White; White → 0 Black.
(● → GC ○ → CG ◉ → AU|UA × → U)

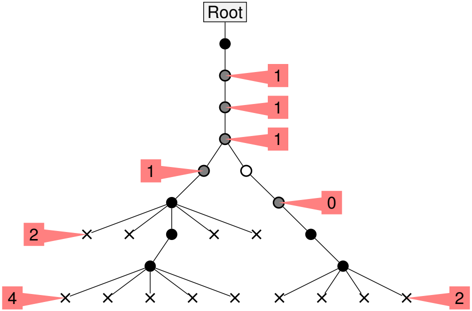




Our Results: Separated Coloring (example)

Descendant restrictions: Any node → ≤ 1 black & ≤ 1 White & ≤ 2 Grey;
Grey → 0/1 Grey; Black → 0 White; White → 0 Black.
(● → GC ○ → CG ● → AU|UA × → U)

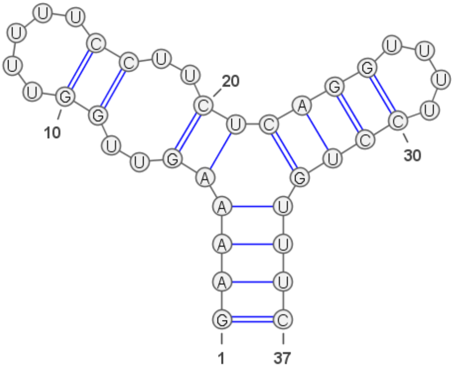


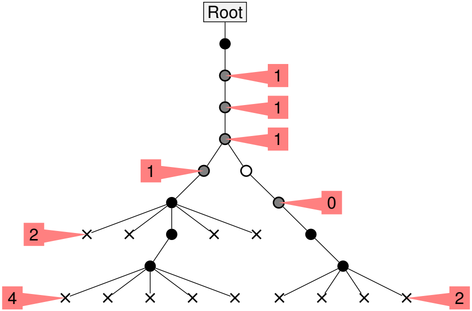


Levels of grey nodes: 0,1
Levels of leaves: 2,4
Separated coloring

Our Results: Separated Coloring (example)

Descendant restrictions: Any node $\to \leq 1$ black & ≤ 1 White & ≤ 2 Grey;
Grey $\to 0/1$ Grey; Black $\to 0$ White; White $\to 0$ Black.
(● → GC ○ → CG ◍ → AU|UA × → U)





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

⇒ **Design:** GAAAAGUUGGUUUUUCCUUCUCAGGUUUUCCUGUUUC

Our Results: Designability over the complete alphabet

$\Sigma_{2,0} = \{\mathrm{A, U, C, G}\}$ + $\{\mathrm{G - C, A - U}\}$ base pairs.

Without unpaired position → complete characterization:

R4 $\Sigma_{2,0} \Rightarrow$ Saturated Designable $=$ Degree≤ 4.

With unpaired positions → partial characterization:

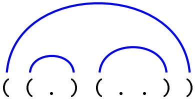
R5 (Necessary) Designable structure cannot contain "*a multiloop of degree* ≥ 5" (motif m_5) or "*a multiloop with unpaired position of degree* ≥ 3" (motif $m_{3\circ}$).

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

R7 If $S \in$ Designable($\Sigma_{2,0}$), then k-stutter $S^{[k]} \in$ Designable($\Sigma_{2,0}$).

Our Results: *k*-Stutter

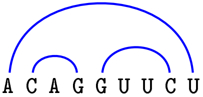
Designable structure:



Then 2-stutter is designable as well:

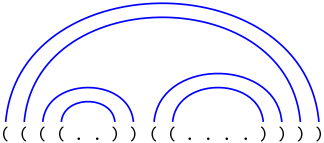
Our Results: *k*-Stutter

Designable structure:



A C A G G U U C U

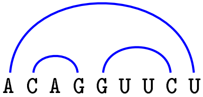
Then 2-stutter is designable as well:



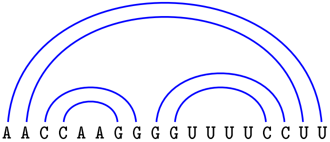
((((. .)) ((. . . .))))

Our Results: *k*-Stutter

Designable structure:

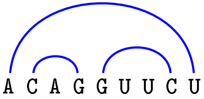


Then 2-stutter is designable as well:

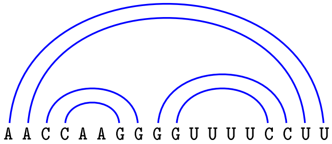


Our Results: *k*-Stutter

Designable structure:



Then 2-stutter is designable as well:



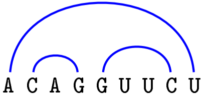
Proof idea: w: Design for S; $S' \neq S^{[k]}$: Alternative folding for k-stutter $w^{[k]}$:

- 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 = 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 $\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})$

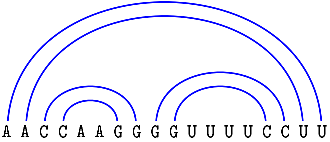
$\Rightarrow$ $w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $\mathcal{M}$)

Our Results: *k*-Stutter

Designable structure:



Then 2-stutter is designable as well:



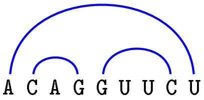
Proof idea: w: Design for S; $S' \neq S^{[k]}$: Alternative folding for k-stutter $w^{[k]}$:

- 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 = 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 $\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})$

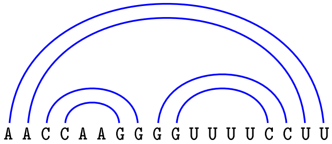
$\Rightarrow$ $w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $\mathcal{M}$)

Our Results: *k*-Stutter

Designable structure:



Then 2-stutter is designable as well:



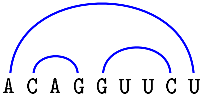
Proof idea: w: Design for S; $S' \neq S^{[k]}$: Alternative folding for k-stutter $w^{[k]}$:

- Compact k consecutive positions → 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 $E_{\mathcal{M}}(S_c) = E_{\mathcal{M}}(S)$ ($\Rightarrow S_c = S$), or $E_{\mathcal{M}}(S_c) > E_{\mathcal{M}}(S)$ (holds for some c)
- Thus $\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})$

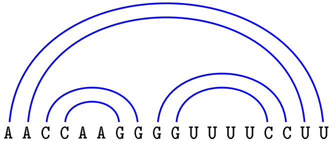
$\Rightarrow$ $w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $\mathcal{M}$)

Our Results: *k*-Stutter

Designable structure:



Then 2-stutter is designable as well:



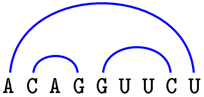
Proof idea: w: Design for S; $S' \neq S^{[k]}$: Alternative folding for k-stutter $w^{[k]}$:

- 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 = 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 $\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})$

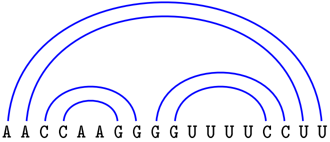
$\Rightarrow$ $w^{[k]}$ is design for $S^{[k]}$ (holds for any base-pair additive $\mathcal{M}$)

Our Results: *k*-Stutter

Designable structure:



Then 2-stutter is designable as well:



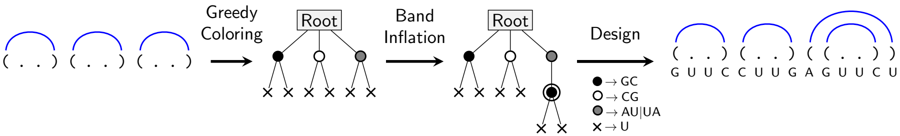
Proof idea: w: Design for S; $S' \neq S^{[k]}$: Alternative folding for k-stutter $w^{[k]}$:

- 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 = 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 $\sum_c E_{\mathcal{M}}(S_c) > k \cdot E(S) = E(S^{[k]})$

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

Our Results: Structure-Approximating Algorithm

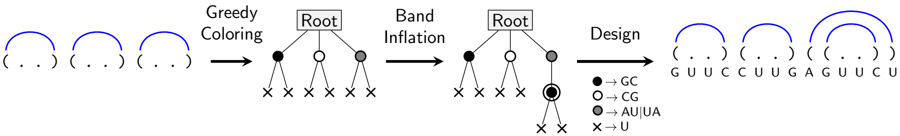
R8 Any structure S without m_5 and $m_{3\circ}$ can be transformed in $\Theta(n)$ time into a designable structure S', by adding at most a single base-pair to its helices.



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

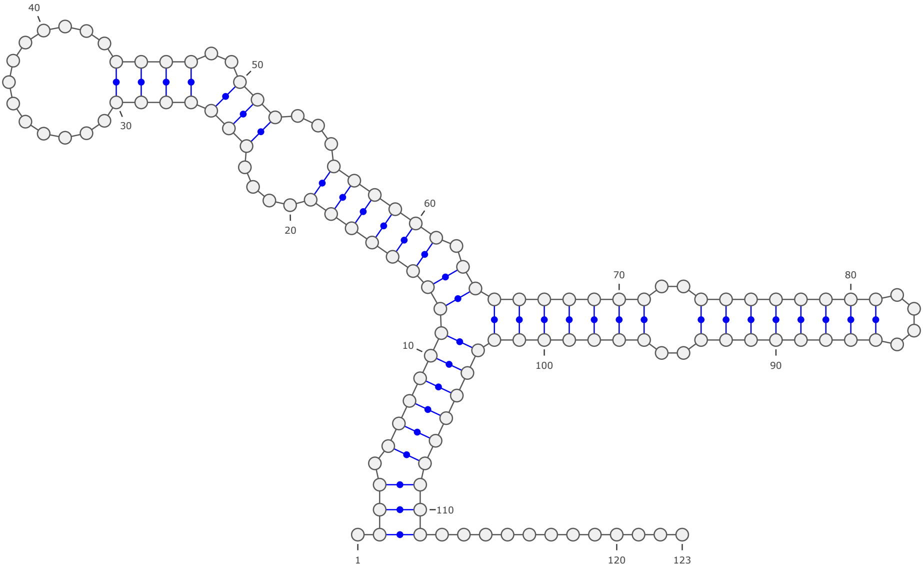
Our Results: Structure-Approximating Algorithm

R8 Any structure S without m_5 and $m_{3\circ}$ can be transformed in $\Theta(n)$ time into a designable structure S', by adding at most a single base-pair to its helices.

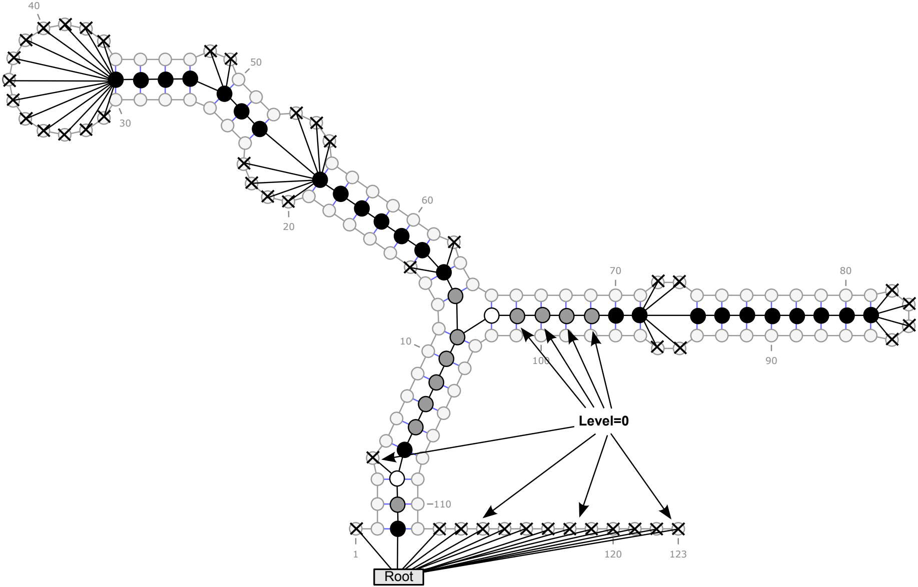


Main idea: Offset grey vertices and leaves to odd/even levels
→ Coloring is now **separated**

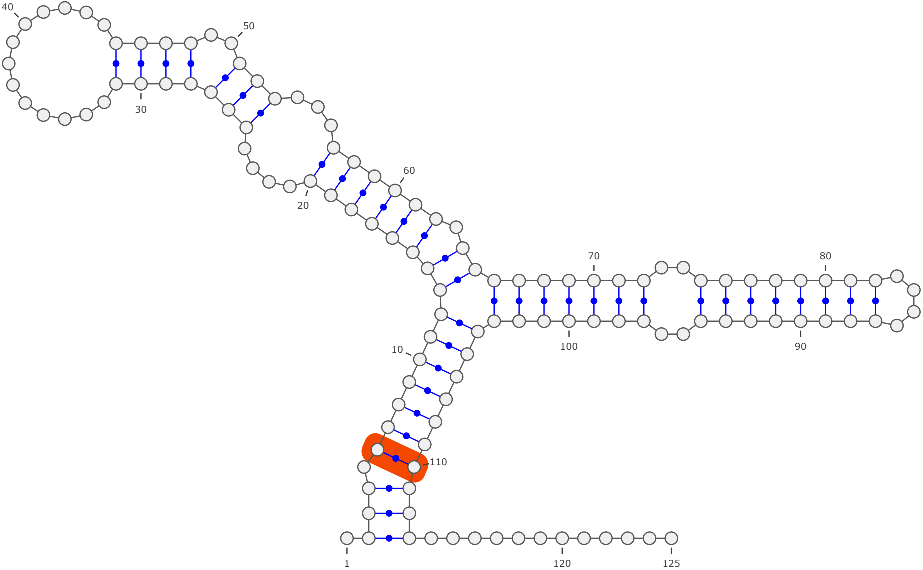
Example



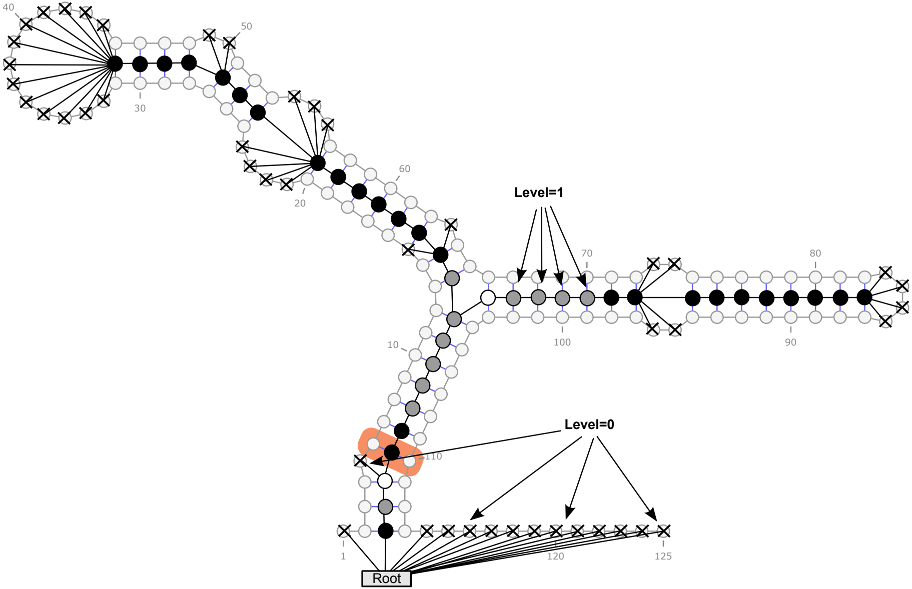
Example



Example



Example



Level=1

Level=0

Generalization

Theorem

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

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

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 $\mathrm{G} - \mathrm{U}$ base pair in optimal fold, since $\alpha > \gamma$.
Numbers of $\mathrm{G} - \mathrm{C}$ and $\mathrm{A} - \mathrm{U}$ 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 ($\mathrm{A}-\mathrm{U}, \mathrm{G}-\mathrm{C}, \mathrm{G}-\mathrm{U}$ + 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 $\in \Theta\left(\frac{\alpha^n}{n\sqrt{n}}\right)$ $(\theta = 0 \rightarrow \alpha = 3)$
  - #Designable structures $\in \mathcal{O}\left(\frac{\beta^n}{n\sqrt{n}}\right), \beta < \alpha$

  Proportion of designable structures: $\left(\frac{\beta}{\alpha}\right)^n$, **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 ($\mathrm{A}-\mathrm{U}, \mathrm{G}-\mathrm{C}, \mathrm{G}-\mathrm{U}$ + 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 & $m_{3\circ}$) can be found in any energy model
  ⇒ **Designable structures** ⊂ **Tree-like** objects with **forbidden motifs**
  + **Basic analytic combinatorics** (*à la* Philippe Flajolet):
  - #Secondary structures $\in \Theta\left(\frac{\alpha^n}{n\sqrt{n}}\right)$ $(\theta = 0 \to \alpha = 3)$
  - #Designable structures $\in \mathcal{O}\left(\frac{\beta^n}{n\sqrt{n}}\right), \beta < \alpha$

  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 ($\mathsf{A}-\mathsf{U},\mathsf{G}-\mathsf{C},\mathsf{G}-\mathsf{U}$ + weights)
  $\Rightarrow$**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 & $m_{3\circ}$) 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 $\in \Theta\left(\frac{\alpha^n}{n\sqrt{n}}\right)$ $(\theta = 0 \to \alpha = 3)$
  - #Designable structures $\in \mathcal{O}\left(\frac{\beta^n}{n\sqrt{n}}\right), \beta < \alpha$

  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 ($\mathrm{A}-\mathrm{U}, \mathrm{G}-\mathrm{C}, \mathrm{G}-\mathrm{U}$ + weights)
  $\Rightarrow$**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 & $m_{3\,\circ}$) can be found in any energy model

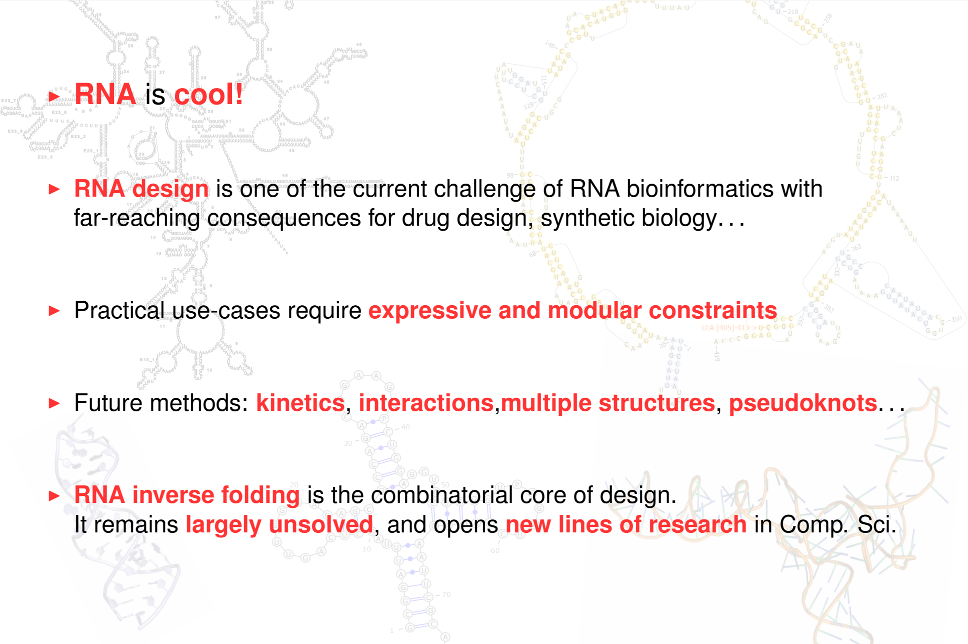
  $\Rightarrow$ **Designable structures** $\subset$ **Tree-like** objects with **forbidden motifs**

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

  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**

Conclusion (Design)



- **RNA** is **cool!**
- **RNA design** is one of the current challenge 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! [Thachuk2010]
- **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

Agence Nationale de la Recherche ANR

Inria INVENTEURS DU MONDE NUMÉRIQUE

cnrs dépasser les frontières