

Selected combinatorial problems in RNA Bioinformatics

...and some solutions

AMIB^{*,†}

+ Many collaborators

• LIX, CNRS/Ecole Polytechnique

† Amib project-team, Inria Saclay

* Université Paris-Saclay

► Staff members

- Philippe Chassignet (MCF Ecole Polytechnique)
- Laurent Mouchard (MCF Université de Rouen – Associé)
- Yann Ponty (CR CNRS – Resp. équipe)
- Mireille Régnier (DR Inria/CNRS – DU LIX)
- Jean-Marc Steyaert (Pr Ecole Polytechnique – Aemeritus)

► PhD Students

- Alice Héliou (Ecole Polytechnique)
- Amélie Héliou (Ecole Polytechnique)
- Juraj Michalik (ANR/FWF – Inria)
- Jorgelindo Moreira Da Veiga (CIFRE Soredab)
- Afaf Saaidi (FRM – CNRS)
- Antoine Soulé (Ecole Polytechnique/Univ. McGill)
- Wei Wang (Université Paris-Sud/LRI)

+ Pauline Pommeret (Inria Engineer) and Evelyne Rayssac (Asst/Adm)

Research:

- **Enumerative combinatorics**, **combinatorial optimization** and **stringology** ...
- ... for **structural biology** and **(comparative) genomics** ...
- ... with a strong taste for **RiboNucleic Acids (RNAs)**.

► Staff members

- Philippe Chassignet (MCF Ecole Polytechnique)
- Laurent Mouchard (MCF Université de Rouen – Associé)
- Yann Ponty (CR CNRS – Resp. équipe)
- Mireille Régnier (DR Inria/CNRS – DU LIX)
- Jean-Marc Steyaert (Pr Ecole Polytechnique – Aemeritus)

► PhD Students

- Alice Héliou (Ecole Polytechnique)
- Amélie Héliou (Ecole Polytechnique)
- Juraj Michalik (ANR/FWF – Inria)
- Jorgelindo Moreira Da Veiga (CIFRE Soredab)
- Afaf Saaidi (FRM – CNRS)
- Antoine Soulé (Ecole Polytechnique/Univ. McGill)
- Wei Wang (Université Paris-Sud/LRI)

+ Pauline Pommeret (Inria Engineer) and Evelyne Rayssac (Asst/Adm)

Research:

- **Enumerative combinatorics**, **combinatorial optimization** and **stringology** ...
- ... for **structural biology** and **(comparative) genomics** ...
- ... with a strong taste for **RiboNucleic Acids (RNAs)**.

► Staff members

- Philippe Chassignet (MCF Ecole Polytechnique)
- Laurent Mouchard (MCF Université de Rouen – Associé)
- Yann Ponty (CR CNRS – Resp. équipe)
- Mireille Régnier (DR Inria/CNRS – DU LIX)
- Jean-Marc Steyaert (Pr Ecole Polytechnique – Aemeritus)

► PhD Students

- Alice Héliou (Ecole Polytechnique)
- Amélie Héliou (Ecole Polytechnique)
- Juraj Michalik (ANR/FWF – Inria)
- Jorgelindo Moreira Da Veiga (CIFRE Soredab)
- Afaf Saaidi (FRM – CNRS)
- Antoine Soulé (Ecole Polytechnique/Univ. McGill)
- Wei Wang (Université Paris-Sud/LRI)

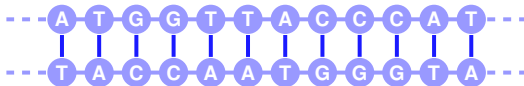
+ Pauline Pommeret (Inria Engineer) and Evelyne Rayssac (Asst/Adm)

Research:

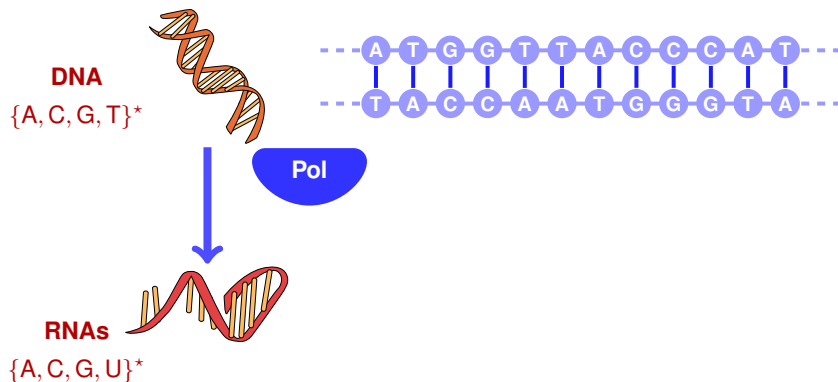
- **Enumerative combinatorics**, **combinatorial optimization** and **stringology** ...
- ... for **structural biology** and **(comparative) genomics** ...
- ... with a strong taste for **RiboNucleic Acids (RNAs)**.

Fundamental *dogma* of molecular biology

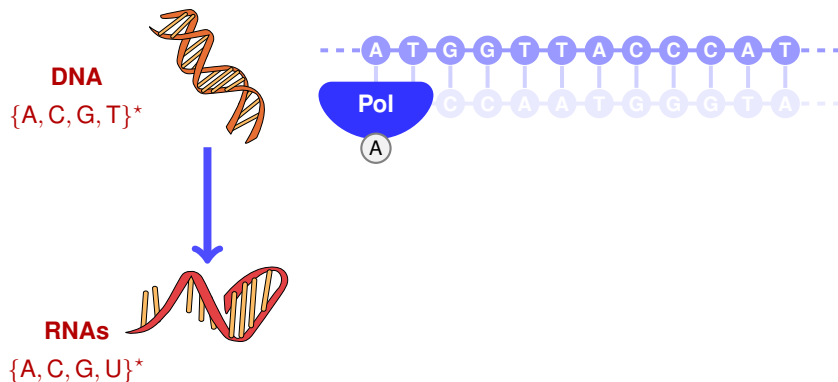
DNA
 $\{A, C, G, T\}^*$



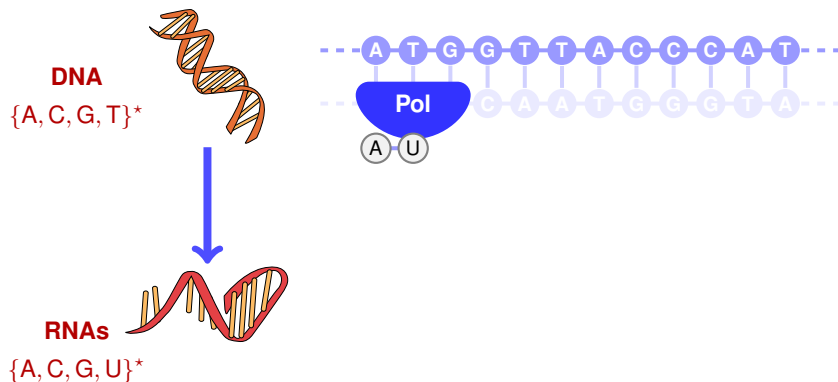
Fundamental *dogma* of molecular biology



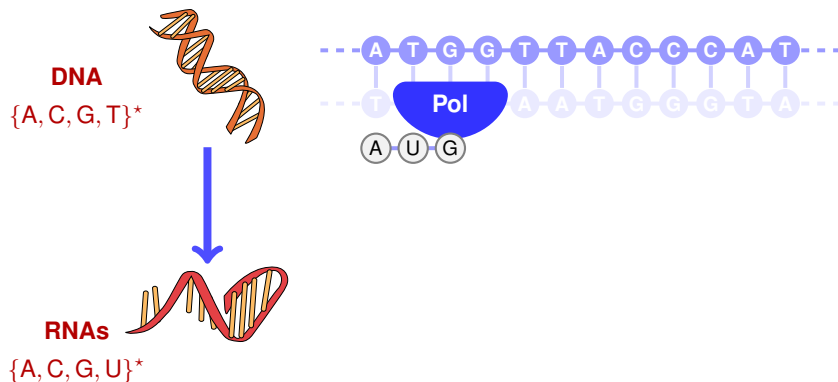
Fundamental *dogma* of molecular biology



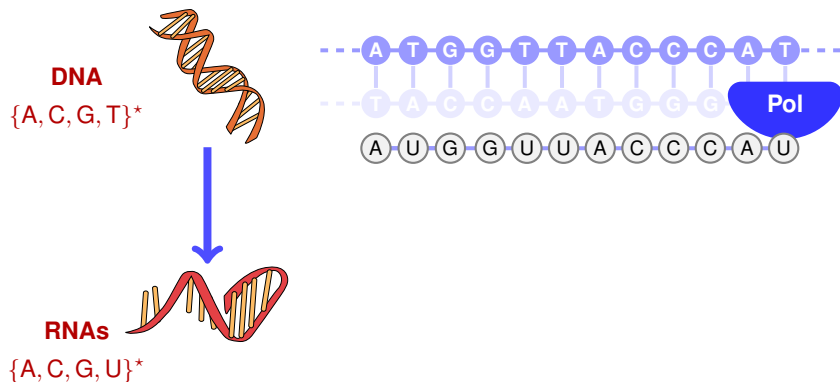
Fundamental *dogma* of molecular biology



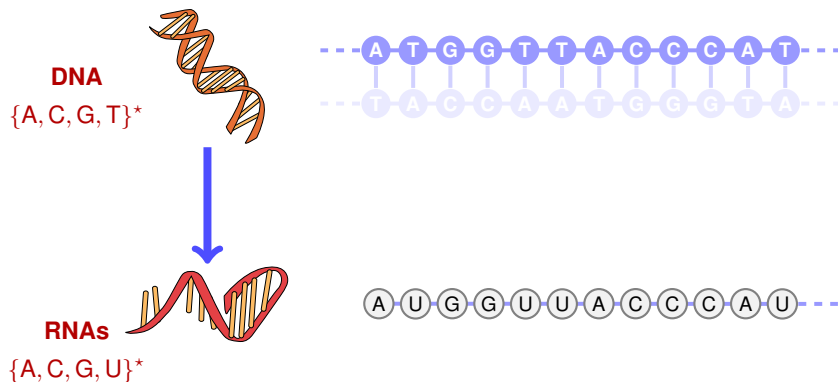
Fundamental *dogma* of molecular biology



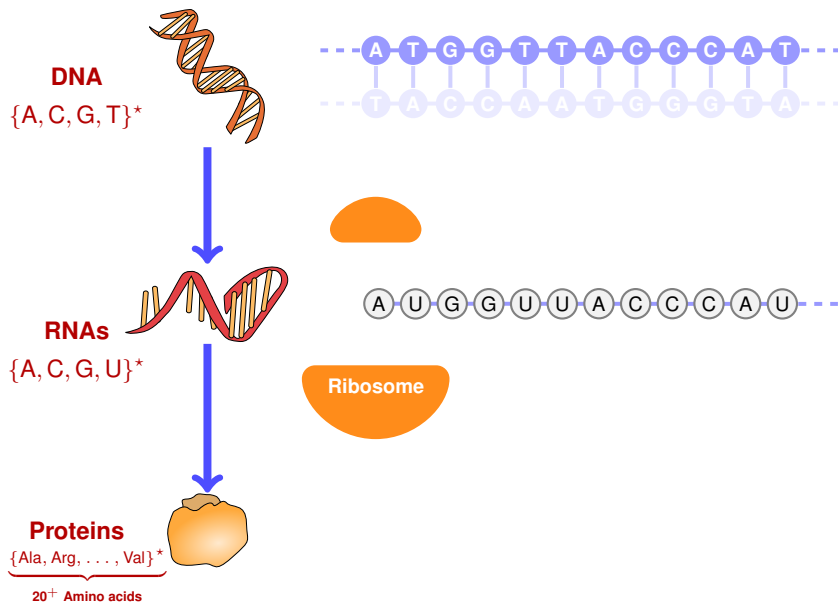
Fundamental *dogma* of molecular biology



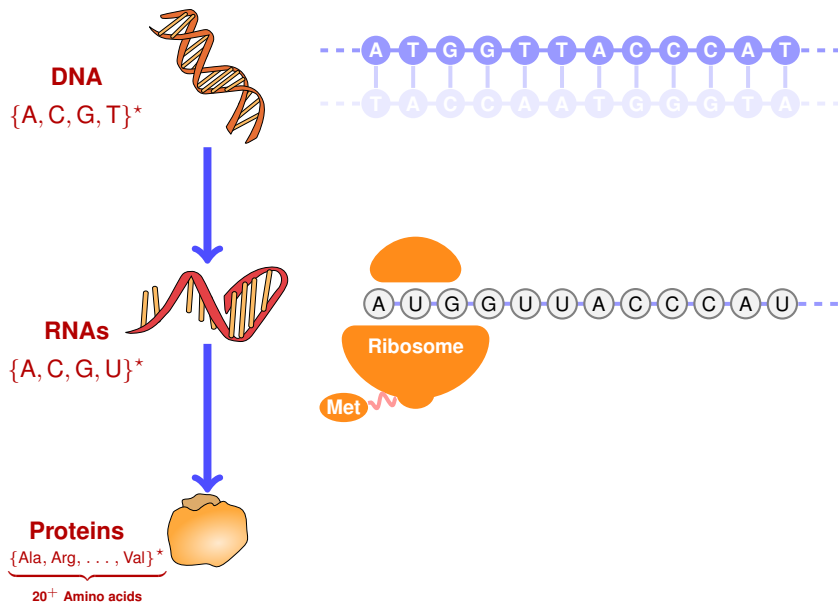
Fundamental *dogma* of molecular biology



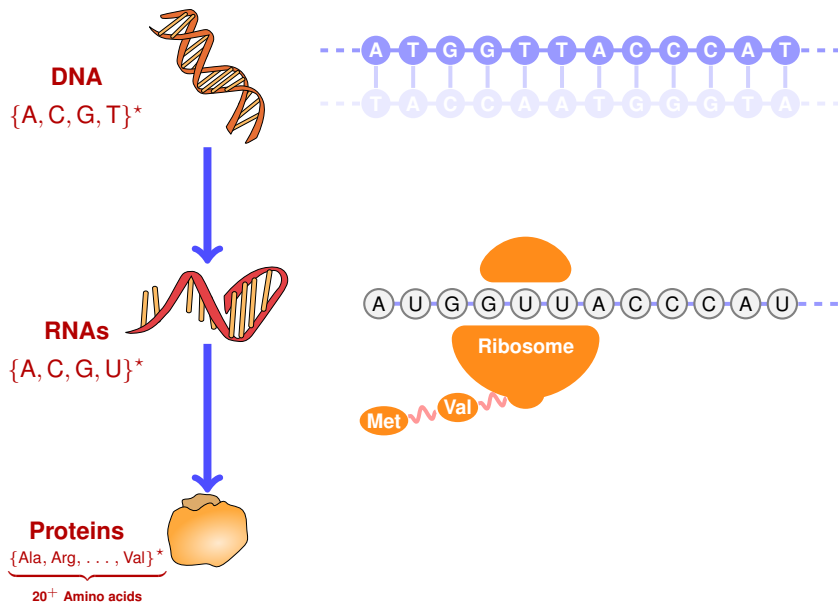
Fundamental *dogma* of molecular biology



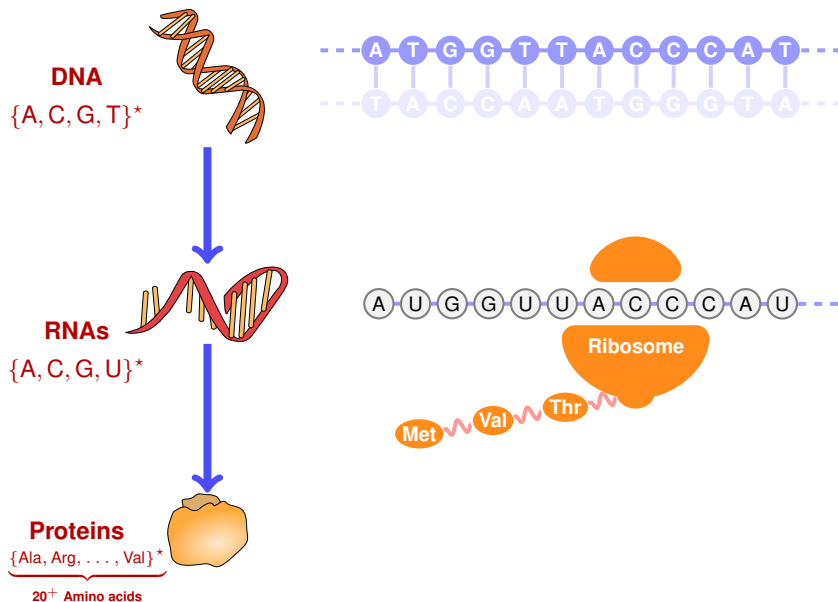
Fundamental *dogma* of molecular biology



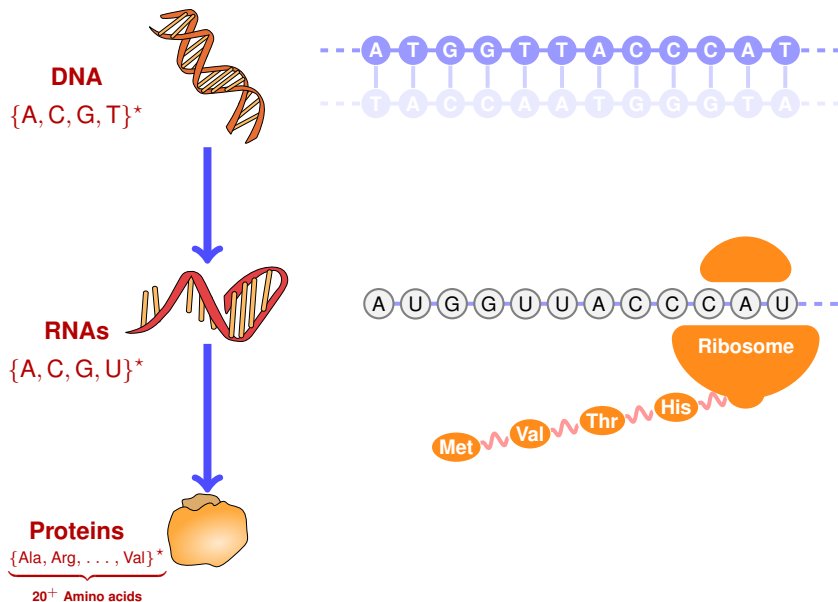
Fundamental *dogma* of molecular biology



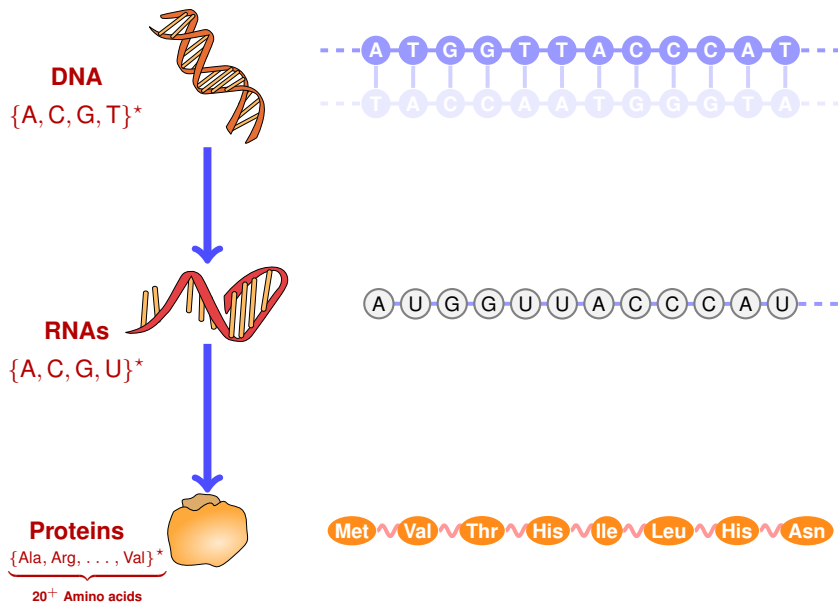
Fundamental *dogma* of molecular biology



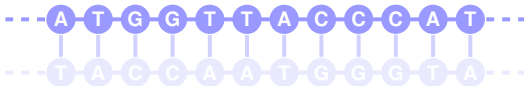
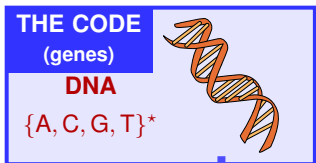
Fundamental *dogma* of molecular biology



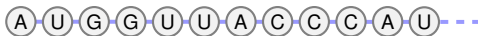
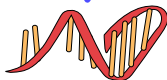
Fundamental *dogma* of molecular biology



Fundamental *dogma* of molecular biology



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

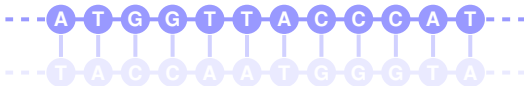
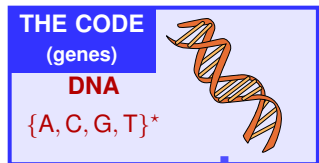


Proteins
 $\{Ala, Arg, \dots, Val\}^*$

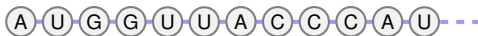
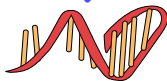
20⁺ Amino acids



Fundamental *dogma* of molecular biology



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

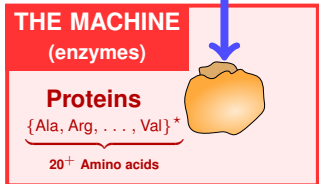
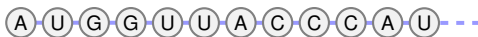
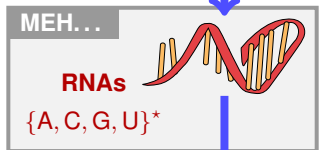
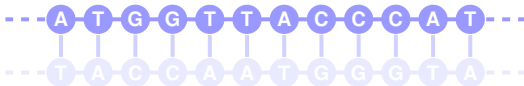
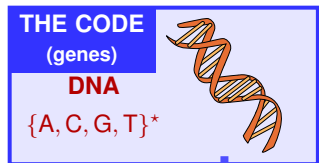


THE MACHINE
(enzymes)

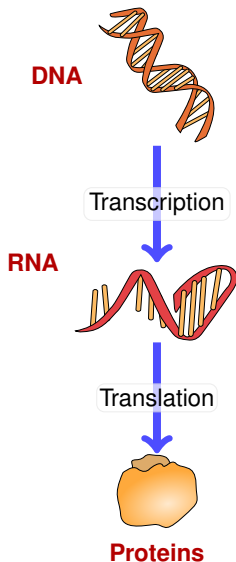
Proteins
{Ala, Arg, . . . , Val}^{*}
20⁺ Amino acids



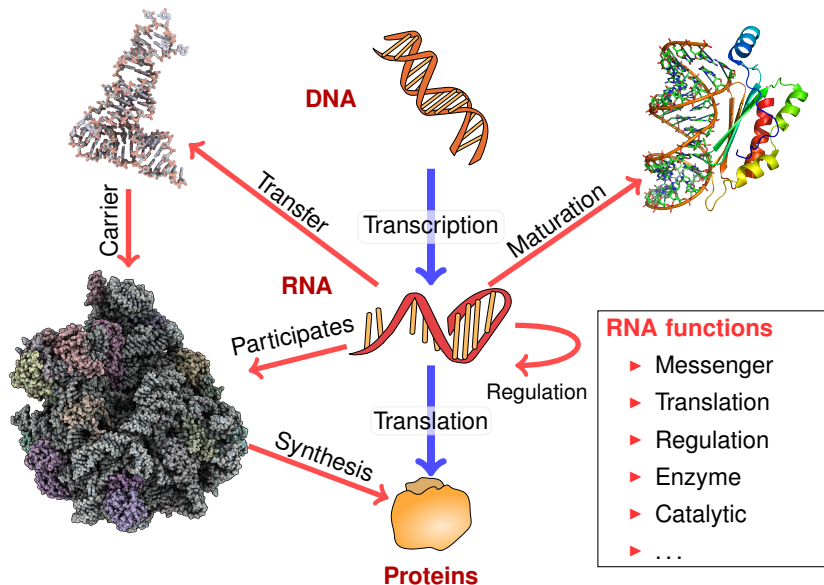
Fundamental *dogma* of molecular biology



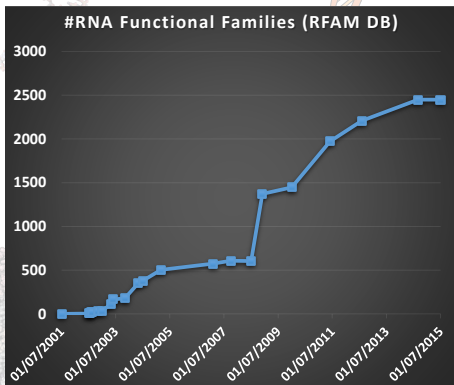
Fundamental *dogma* of molecular biology



Fundamental *dogma* of molecular biology



Fundamental *dogma* of molecular biology

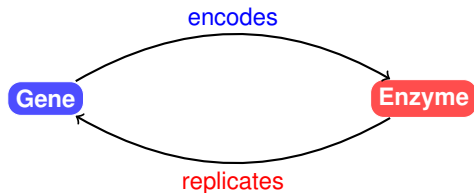


RNA functions

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

Proteins

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

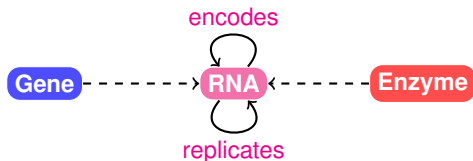


A **gene** big enough to specify **an enzyme** would be too big to replicate accurately without the aid of **an enzyme** of the very kind that it is trying to specify. So the system *apparently cannot get started*.

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

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

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



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

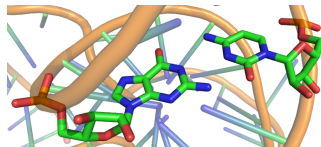
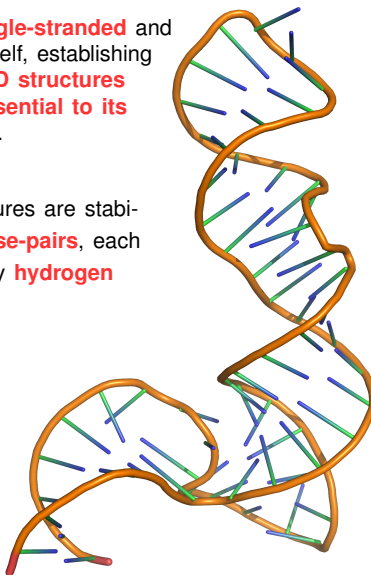
[...] This is the **RNA World**. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why **RNA might just be good enough at both roles to break out of the Catch-22**.

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

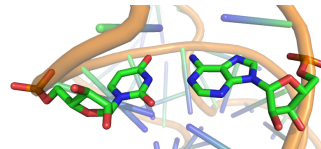
RNA folding

RNA is **single-stranded** and **folds** on itself, establishing **complex 3D structures** that are **essential to its function(s)**.

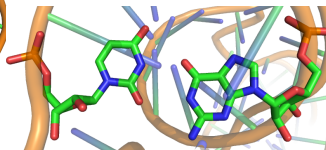
RNA structures are stabilized by **base-pairs**, each mediated by **hydrogen bonds**.



G/C



U/A



U/G

Watson/Crick base-pairs

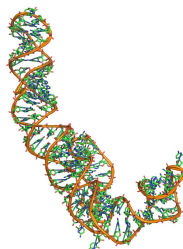
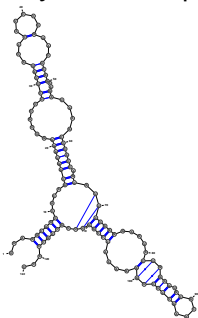
Wobble base-pair

Canonical base-pairs

RNA structure(s)

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

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCAUCCCGAA
CACGGAAGUAAGCC
CACCAGCGUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```



Primary structure

Secondary structure

Tertiary structure

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

Definition (Secondary Structure)

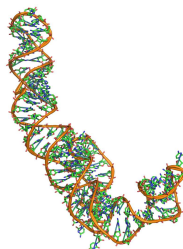
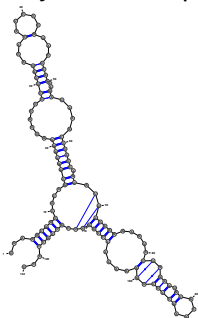
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- ▶ **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- ▶ **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ 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 $\{A, C, G, U\}^*$

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAGCC
CACCAGCGUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```



Primary structure

Secondary structure

Tertiary structure

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

Definition (Secondary Structure)

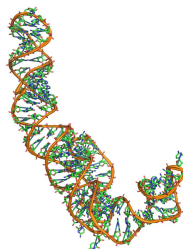
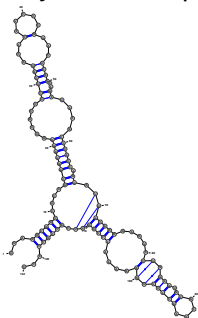
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- ▶ **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- ▶ **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ 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 $\{A, C, G, U\}^*$

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAGCC
CACCAGCGUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```



Primary structure

Secondary structure

Tertiary structure

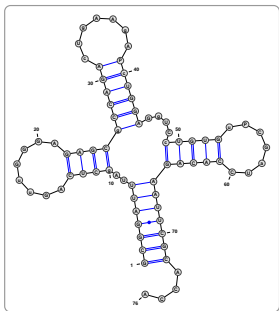
Source: 5s rRNA (PDBID: 1K73:B)

Definition (Secondary Structure)

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

- ▶ **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- ▶ **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ 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*

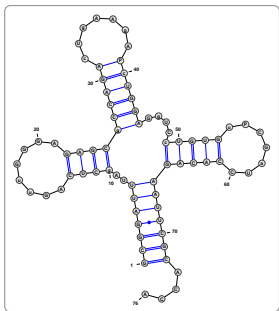
Supporting intuitions

Different representations

Common combinatorial structure

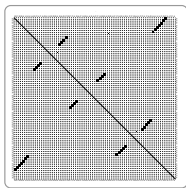
* Additional steric constraints

Various representations for a versatile biomolecule



Outer-planar graphs

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



Dot plots

Adjacency matrices*

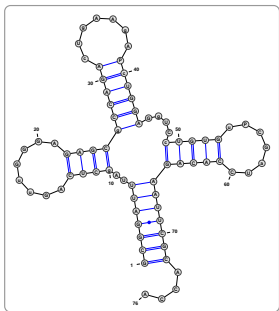
Supporting intuitions

Different representations

Common combinatorial structure

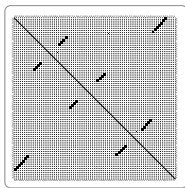
* Additional steric constraints

Various representations for a versatile biomolecule

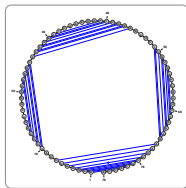


Outer-planar graphs

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



Dot plots
Adjacency matrices*



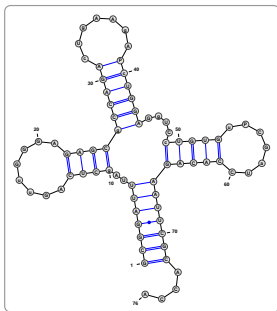
Non-crossing arc diagrams*

Supporting intuitions

Different representations
Common combinatorial structure

* Additional steric constraints

Various representations for a versatile biomolecule

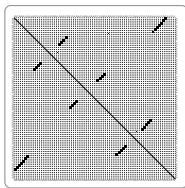


Outer-planar graphs

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

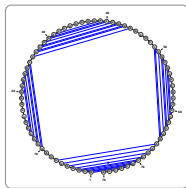
(((((...(((...))))))(((...))))...(((...))))))...))

Motzkin words*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

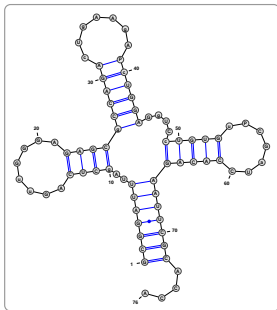
Supporting intuitions

Different representations

Common combinatorial structure

* Additional steric constraints

Various representations for a versatile biomolecule

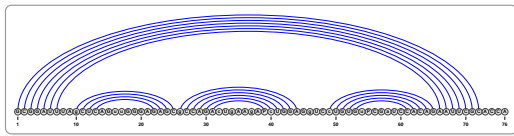


Outer-planar graphs

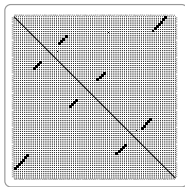
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

Motzkin words*

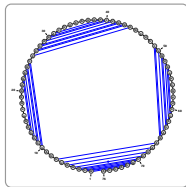


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

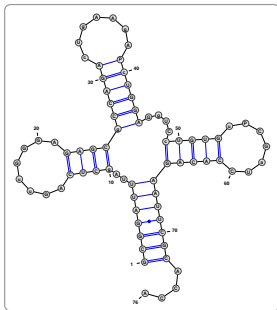
Supporting intuitions

Different representations

Common combinatorial structure

* Additional steric constraints

Various representations for a versatile biomolecule

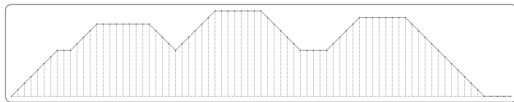


Outer-planar graphs

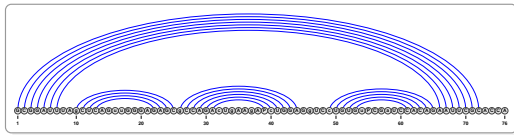
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

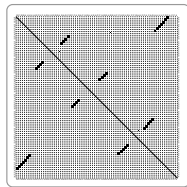
Motzkin words*



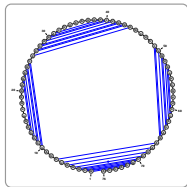
Positive 1D meanders* over $S = \{+1, -1, 0\}$



Non-crossing arc-annotated sequences*



Dot plots
Adjacency matrices*



Non-crossing arc diagrams*

Supporting intuitions

Different representations

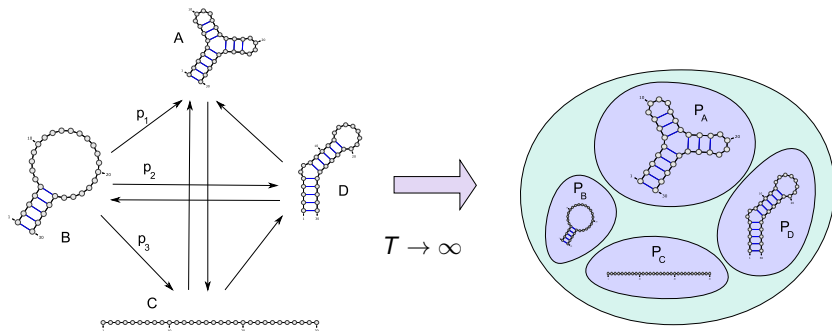
Common combinatorial structure

* 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:

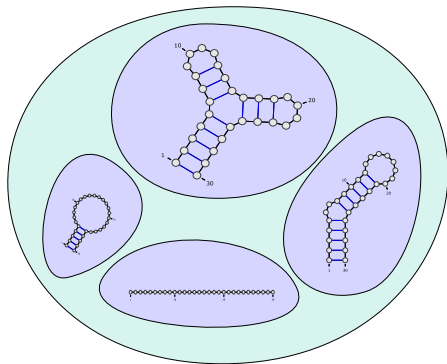
$$\mathbb{P}(S \mid w) = \frac{e^{-\frac{E_w(S)}{kT}}}{\mathcal{Z}_w} \quad \text{and} \quad \mathcal{Z}_w \equiv \sum_{S'} e^{-\frac{E_w(S')}{RT}} \quad \{\text{Partition function}\}$$

$E_w(S)$: **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



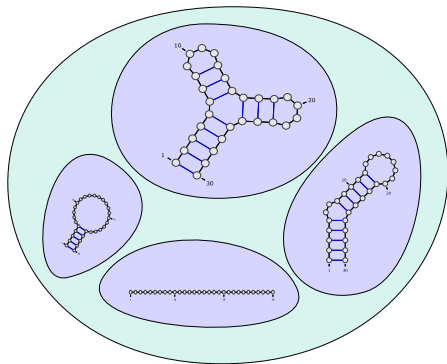
mRNA half-life: $\sim 7\text{h}$
(Mouse [\[Sharova2009\]](#))

$$T \rightarrow \infty$$

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

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



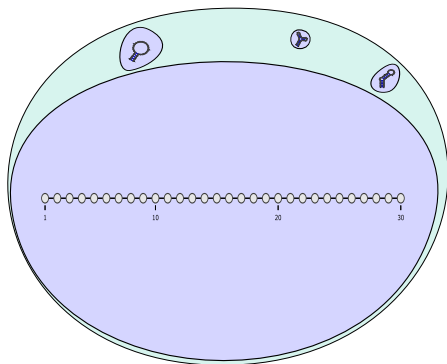
mRNA half-life: $\sim 7\text{h}$
(Mouse [Sharova2009])

$$T \rightarrow \infty$$

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

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



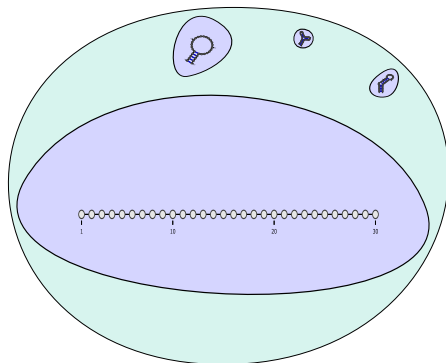
mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

$T = 0h$

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



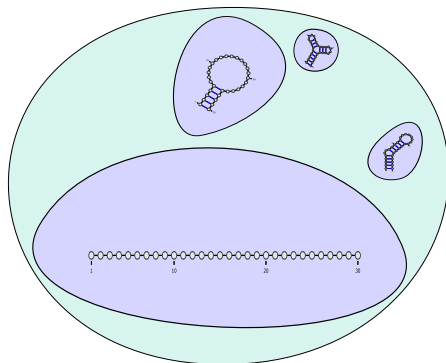
$T = 1\text{h}$

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

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

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

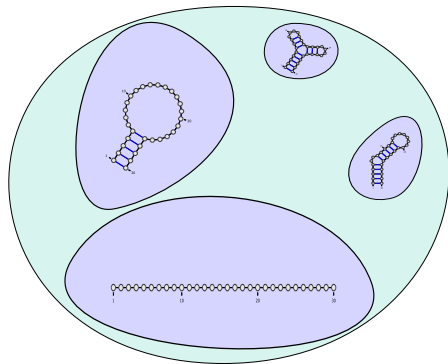


mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

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

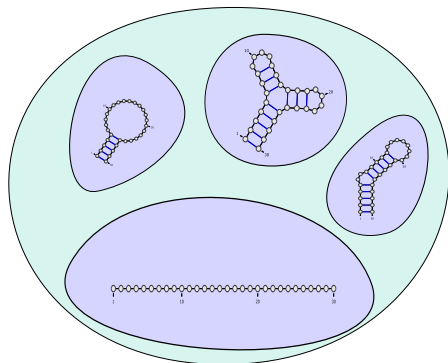


mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

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



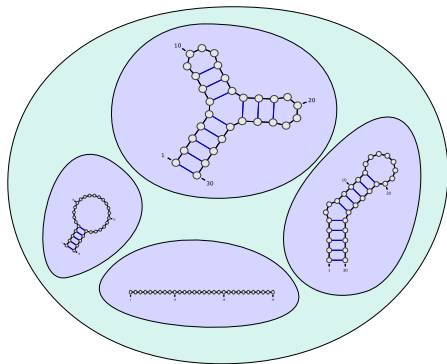
$T = 10h$

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

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

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



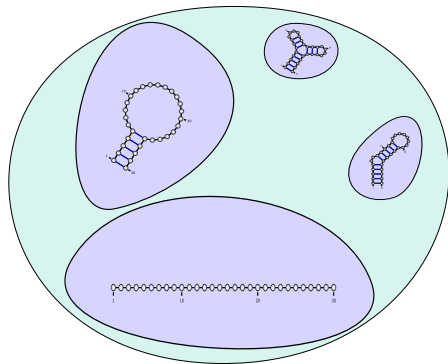
mRNA half-life: $\sim 7\text{h}$
(Mouse [Sharova2009])

$$T \rightarrow \infty$$

Thermodynamics vs Kinetics

Paradigms for RNA structure prediction

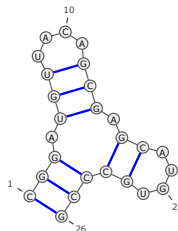
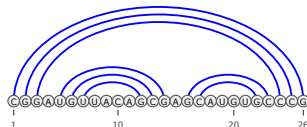
- ▶ **1978–1990s** Functional structure = Minimal Free-Energy
- ▶ **1990s–2010s** Functional structure(s) **representative** of the Boltzmann ensemble
- ▶ **2010s–????** Embracing kinetics



$T = 10h$

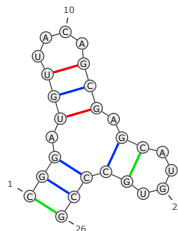
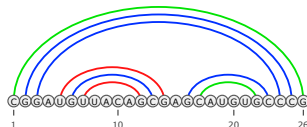
mRNA half-life: $\sim 7h$
(Mouse [Sharova2009])

Problem statement



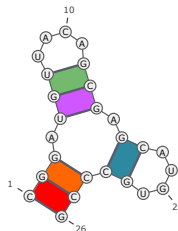
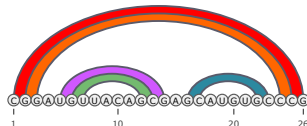
- ▶ **RNA structure S :** (Partial) matching of positions in sequence w
- ▶ **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- ▶ **Energy model:**
 - Motif** \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$:**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

Problem statement



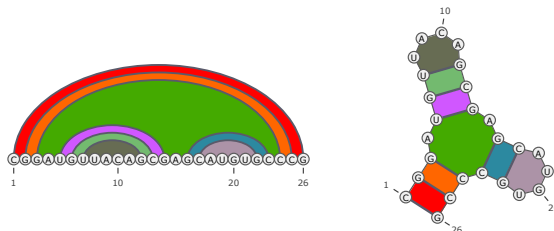
- ▶ **RNA structure S :** (Partial) matching of positions in sequence w
- ▶ **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- ▶ **Energy model:**
 - Motif** \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$:**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

Problem statement



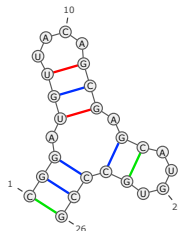
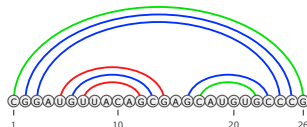
- ▶ **RNA structure S :** (Partial) matching of positions in sequence w
- ▶ **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- ▶ **Energy model:**
 - Motif** \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$:**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

Problem statement



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

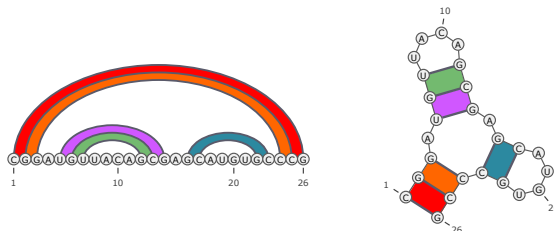
Problem statement



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

$$E_S = 2 \cdot \Delta \left(\begin{array}{c} \textcircled{\text{U}} \\ \textcolor{red}{|} \\ \textcircled{\text{G}} \end{array} \right) + 4 \cdot \Delta \left(\begin{array}{c} \textcircled{\text{G}} \\ \textcolor{blue}{|} \\ \textcircled{\text{C}} \end{array} \right) + 2 \cdot \Delta \left(\begin{array}{c} \textcircled{\text{G}} \\ \textcolor{green}{|} \\ \textcircled{\text{G}} \end{array} \right)$$

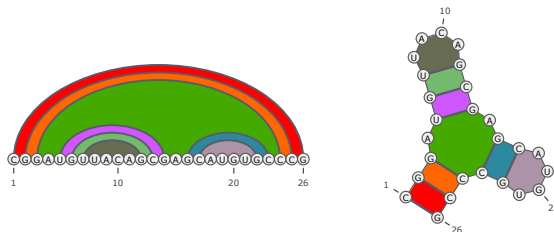
Problem statement



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

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

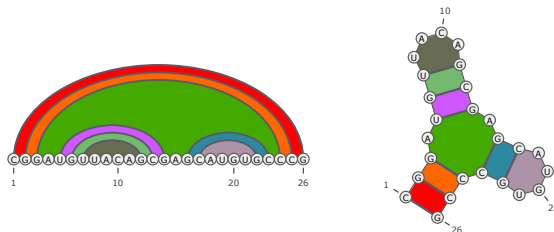
Problem statement



- ▶ **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

$$\begin{aligned}
 E_S = & \Delta \left(\begin{array}{c} \text{C} \quad \text{G} \\ | \quad | \\ \text{G} \quad \text{C} \end{array} \right) + \Delta \left(\begin{array}{c} \text{G} \quad \text{G} \\ | \quad | \\ \text{C} \quad \text{C} \end{array} \right) + \Delta \left(\begin{array}{c} \text{U} \quad \text{G} \\ | \quad | \\ \text{G} \quad \text{C} \end{array} \right) + \Delta \left(\begin{array}{c} \text{U} \quad \text{G} \\ | \quad | \\ \text{G} \quad \text{C} \end{array} \right) + \Delta \left(\begin{array}{c} \text{U} \quad \text{G} \\ | \quad | \\ \text{G} \quad \text{C} \end{array} \right) \\
 & + \Delta \left(\begin{array}{c} \text{A} \quad \text{C} \\ | \quad | \\ \text{U} \quad \text{G} \end{array} \right) + \Delta \left(\begin{array}{c} \text{A} \quad \text{C} \\ | \quad | \\ \text{U} \quad \text{G} \end{array} \right) + \Delta \left(\begin{array}{c} \text{C} \quad \text{A} \\ | \quad | \\ \text{G} \quad \text{U} \end{array} \right)
 \end{aligned}$$

Problem statement



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

Definition (MFE-PREDICT(E) problem)

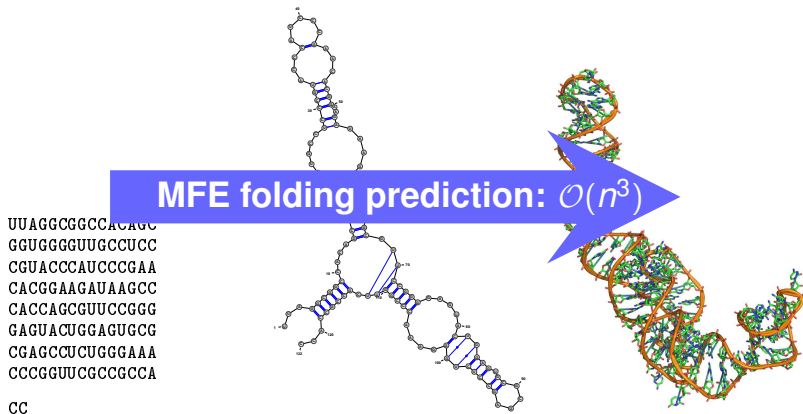
Input: RNA sequence $w \in \{A, C, G, U\}^*$.

Output: (Constrained) matching S^* of Minimal Free-Energy $E_w(S^*)$.

RNA folding: non-crossing matchings

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

Secondary structure = **Non-crossing** matching



Primary Structure

Secondary Structure

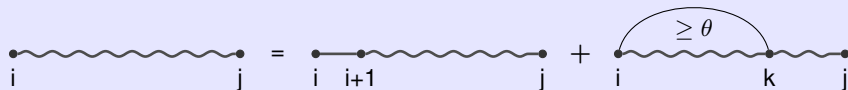
Tertiary Structure

5s rRNA (PDBID: 1K73:B)

Dynamic programming (DP) for RNA folding

Theorem (NussinovJacobson1980 + ZukerStiegler80)

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



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

$(-1 / +\infty$ or $\Delta G(s_i \equiv s_k))$

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

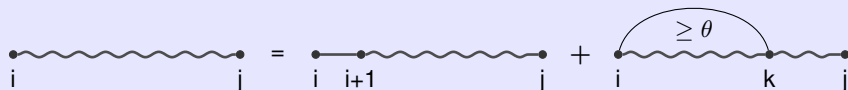
$$N_{i,t} = 0, \quad \forall t \in [i, i + \theta]$$

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

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 $\Delta G(s_i \stackrel{?}{=} s_k))$

$C_{i,j}$: Number of secondary structures compatible with interval $[i, j]$

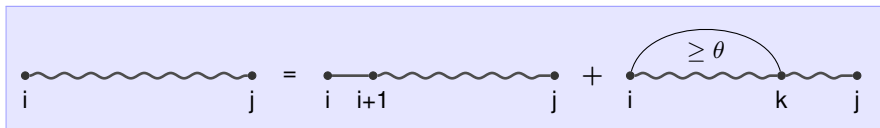
$$C_{i,t} = 1, \quad \forall t \in [i, i + \theta]$$

$$C_{i,j} = \sum \left\{ \begin{array}{ll} \sum_{k=i+\theta+1}^j \mathbb{1}_{\text{comp.}(i,k)} \times C_{i+1,k-1} \times C_{k+1,j} & \{i \text{ unpaired}\} \\ C_{i+1,j} & \{i \text{ paired to } k\} \end{array} \right.$$

Dynamic programming (DP) for RNA folding

Theorem (NussinovJacobson1980 + ZukerStiegler80)

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



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

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

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

$$Z_{i,j} = \sum \left\{ \begin{array}{ll} \sum_{k=i+\theta+1}^j e^{\frac{-E_{i+1,j}}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} & \{i \text{ unpaired}\} \\ e^{\frac{-E_{i,k}}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} & \{i \text{ paired to } k\} \end{array} \right.$$

Dynamic programming (DP) for RNA folding

Many extensions: (Jacobson1980 + ZukerStiegler80)

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

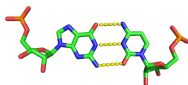
Made possible by:

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

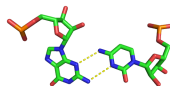
\Rightarrow **Combinatorial Dynamic Programming**

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].

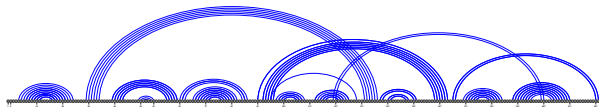


Canonical CG base-pair (WC/WC-Cis)



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

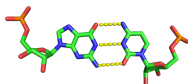
- **Pseudoknots:** Crossing sets of nested stable base-pairs



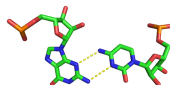
Group I Ribozyme (PDBID: 1Y0Q:A)

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].

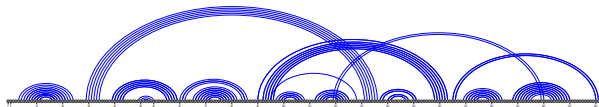


Canonical CG base-pair (WC/WC-Cis)



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

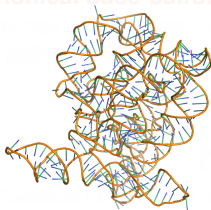
- ▶ **Pseudoknots:** Crossing sets of nested stable base-pairs



Group I Ribozyme (PDBID: 1Y0Q:A)

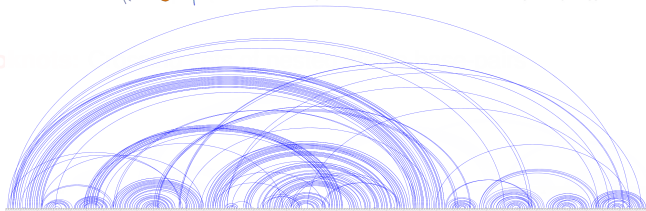
Including crossing interactions

- ▶ **Non-canonical base-pairs:** Lead to **local crossings** and **promiscuity**
Any base pair (A-U), (C-G), (G-U)
OR interactions (e.g. A-C, U-G, C-G, G-U)
Crossing interactions, once ignored, are now **ubiquitous!**



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

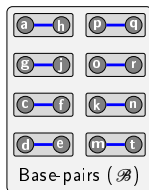
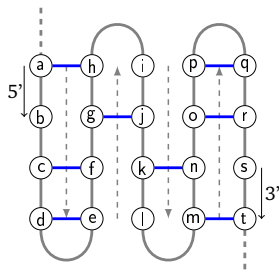
- ▶ **Pseudoknots**



Energy models

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

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

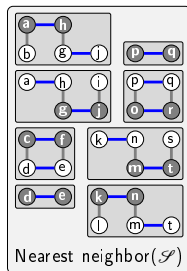
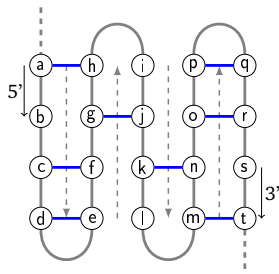


Solved in $\mathcal{O}(n^3)$ [Tabaska1998]
(Max-weighted matching)
Unrealistic!

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

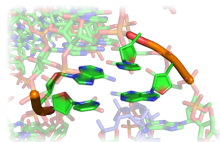
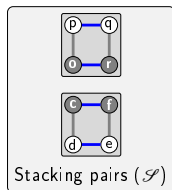
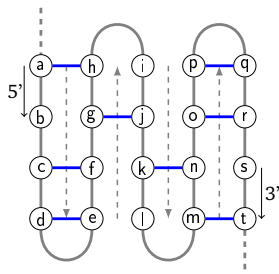


NP-
hard [Lyngso2000,Akutsu2000]
Too expressive?

Energy models



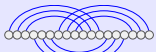
Three models, based on interacting positions (i, j) :

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



Captures stablest motifs
Still NP-hard [Lyngso2004]
... but PTAS [Lyngso2004]

The full monty

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Opt.	P [Nussinov1980]	P [leong2003]	P [Zuker1981]
Non-crossing	Approx.	—	—	—
	Opt.	???	NP-Hard [leong2003]	NP-Hard [leong2003]
Planar	Approx.	2-approx. ≈[leong2003]	2-approx. [leong2003]	???
	Opt.	P [Tabaska1998]	NP-Hard [Lyngso2004] (any* Δ model) [Sheikh2012]	NP-Hard [Lyngso2000] [Akutsu2000]
General	Approx.	Duh...	ϵ -approx. $\in \mathcal{O}(n^{4^{1/\epsilon}})$ [Lyngso2004] 1/5 (any Δ model) [Sheikh2012]	APX-Hard [Sheikh2012]

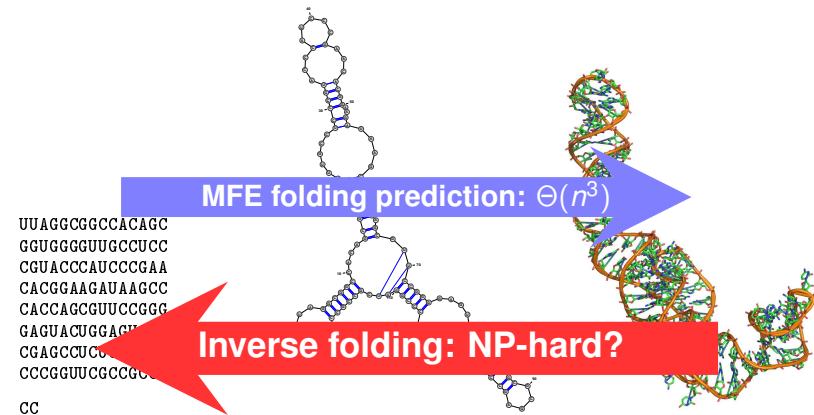
Missing:

- ▶ Base-pair maximization in planar model (probably NP-hard)
- ▶ **Relevance** of approximation???
- ▶ Partition function (Poly. cases), Boltzmann-Gibbs sampling
- ▶ FPT algorithms (Relevant parameters?)

Part. II: Designing RNAs

RNA inverse folding

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



Primary Structure

Secondary Structure

Structure Tertiaire

5s rRNA (PDBID: 1K73:B)

RNA Inverse Folding

\mathcal{M} = energy model

Definition (INVERSE-FOLDING(E) problem)

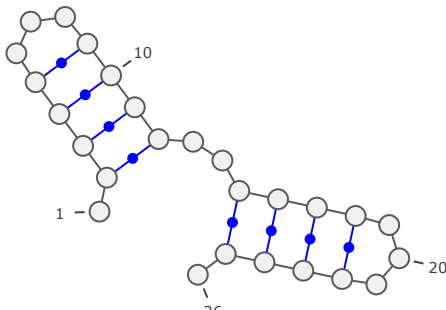
Input: Secondary structure S + Energy distance $\Delta > 0$.

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

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

or \emptyset if no such sequence exists.

No (obvious?) optimal substructure property:



RNA Inverse Folding

\mathcal{M} = energy model

Definition (INVERSE-FOLDING(E) problem)

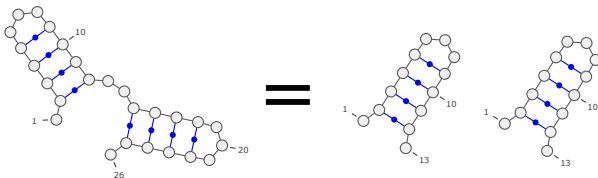
Input: Secondary structure S + Energy distance $\Delta > 0$.

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

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

or \emptyset if no such sequence exists.

No (obvious?) optimal substructure property:



RNA Inverse Folding

\mathcal{M} = energy model

Definition (INVERSE-FOLDING(E) problem)

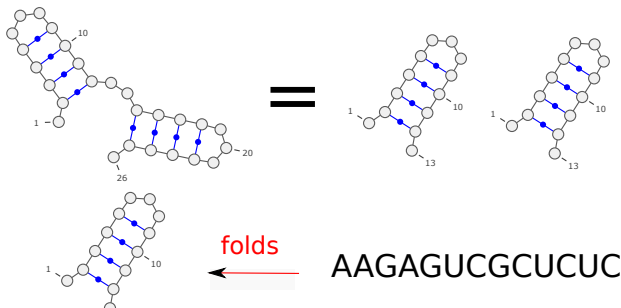
Input: Secondary structure S + Energy distance $\Delta > 0$.

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

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

or \emptyset if no such sequence exists.

No (obvious?) optimal substructure property:



RNA Inverse Folding

\mathcal{M} = energy model

Definition (INVERSE-FOLDING(E) problem)

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

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

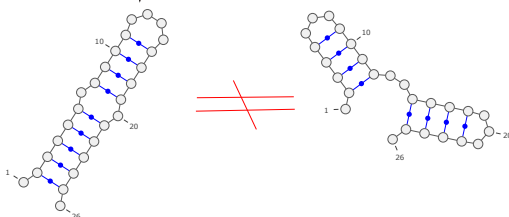
$$\forall S' \in \mathcal{S}|w| \setminus \{S\} : E_{w,S'} \geq E_{w,S} + \Delta$$

or \emptyset if no such sequence exists.

No (obvious?) optimal substructure property:

AAGAGUCGCUCUC AAGAGUCGCUCUC

Folds ↓



RNA Design Problem

\mathcal{M} = energy model

Definition (INVERSE-FOLDING(E) problem)

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

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

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

or \emptyset if no such sequence exists.

Difficult problem: No (obvious??) substructure property

- ▶ **Existing algorithms/software (20+):** Heuristics or Exponential-time
- ▶ Complexity of problem unknown (despite [Schnall Levin et al (2008)])
Clearly in **P**!... **CO-NP**???
- ▶ **Reason:** Non locality, no theoretical frameworks, too many parameters...

⇒ **Stick to a simplified model!**

RNA Design Problem (simplified)

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

Problem (INVERSE-FOLDING(Σ) problem)

Input: Secondary structure S

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

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

or \emptyset if no such sequence exists.

Designable(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 $w \in \Sigma^*$ — called a design for S — such that:

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

or \emptyset if no such sequence exists.

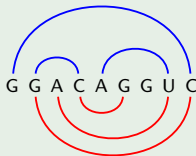
Designable(Σ): All designable structures

Example

a. Target sec. str. S



b. Invalid sequence for S



c. Design for S



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 ≤ 2 + empty structures ;

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

Example



Our Results: Designability over Restricted Alphabets

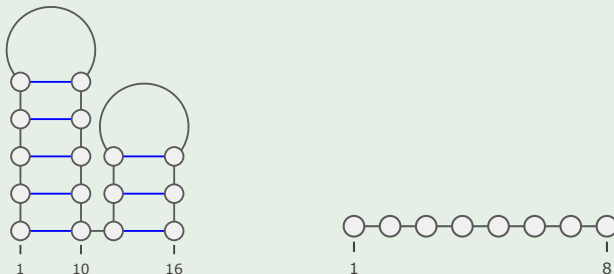
$\Sigma_{c,u}$ = Alphabet with c pairs of complementary bases and u unpairable bases.

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



Our Results: Designability over Restricted Alphabets

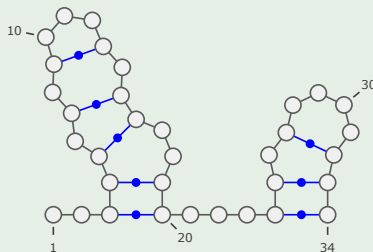
$\Sigma_{c,u}$ = Alphabet with c pairs of complementary bases and u unpairable bases.

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 the Complete Alphabet

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

Without unpaired position \rightarrow complete characterization:

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

With unpaired positions \rightarrow 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 \rightarrow complete characterization:

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

With unpaired positions \rightarrow partial characterization:

R5 (Necessary) Designable structure cannot contain “*a multiloop of degree ≥ 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.

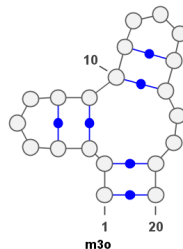
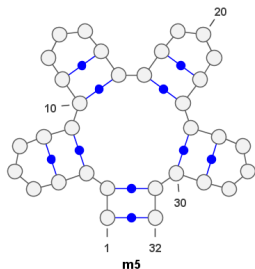
Without unpaired position \rightarrow complete characterization:

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

With unpaired positions \rightarrow 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 \rightarrow complete characterization:

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

With unpaired positions \rightarrow partial characterization:

R5 (Necessary) Designable structure cannot contain “*a multiloop of degree ≥ 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 \rightarrow complete characterization:

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

With unpaired positions \rightarrow partial characterization:

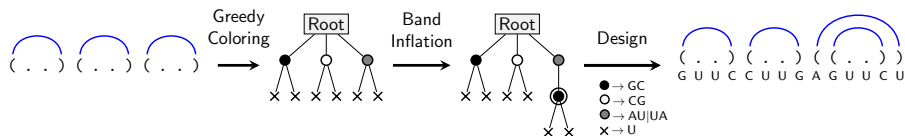
R5 (Necessary) Designable structure cannot contain “*a multiloop of degree ≥ 5* ” (motif m_5) or “*a multiloop with unpaired position of degree ≥ 3* ” (motif $m_{3\circ}$).

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

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

Our Results: Structure-Approximating Algorithm

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



Definition (μ structural approximation of design)

Input: Secondary structure S , Energy model E , Energy distance $\Delta > 0$.

Output: RNA sequence $w \in \Sigma^* + 2D$ structure S^* such that:

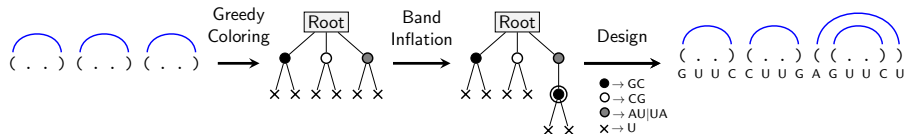
- 1 $\forall S' \in \mathcal{S}[w] \setminus \{S^*\} : E_{w,S'} \geq E_{w,S^*} + \Delta$
- 2 $\frac{Dist(S, S^*)}{|S|} \leq 1/\mu$

or \emptyset if no such sequence exists.

Remark: R8 is a 2 structural approximation wrt the tree-edit distance.

Our Results: Structure-Approximating Algorithm

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



Definition (μ structural approximation of design)

Input: Secondary structure S , Energy model E , Energy distance $\Delta > 0$.

Output: RNA sequence $w \in \Sigma^* + 2D$ structure S^* such that:

$$1 \quad \forall S' \in \mathcal{S}[w] \setminus \{S^*\} : E_{w,S'} \geq E_{w,S^*} + \Delta$$

$$2 \quad \frac{Dist(S, S^*)}{|S|} \leq 1/\mu$$

or \emptyset if no such sequence exists.

Remark: R8 is a 2 structural approximation wrt the tree-edit distance.

Theorem

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

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

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

Proof idea: Stutter results holds for any base-pair additive model.

Other results are based on (G, C)-saturated sequences

No G – U base pair in optimal fold, since $\alpha > \gamma$.

Numbers of G – C and A – U base pairs are upper-bounded.

⇒ Any alternative has same number of each base-pair as target structure.

Remarks

- ▶ Results also hold in **Nussinov** energy model (A – U, G – C, G – U + weights)
⇒ **Stacking** energy model? **Turner**?
- ▶ Characterized classes are mostly **easy**:
 - ▶ **Designable** classes → Linear time **algorithms**
 - ▶ **Non-designable** classes → Linear time **membership tests**
- ▶ **RNA Design**: P or NP? FPT?
- ▶ **Structural approximation** version of the problem (better ratios? NP-hard beyond some ratio?)

Remarks

- ▶ Results also hold in **Nussinov** energy model (A – U, G – C, G – U + weights)
⇒ **Stacking** energy model? **Turner**?
- ▶ Characterized classes are mostly **easy**:
 - ▶ **Designable** classes → Linear time **algorithms**
 - ▶ **Non-designable** classes → Linear time **membership tests**
- ▶ **RNA Design**: P or NP? FPT?
- ▶ **Structural approximation** version of the problem (better ratios? NP-hard beyond some ratio?)

Remarks

- ▶ Results also hold in **Nussinov** energy model (A – U, G – C, G – U + weights)
⇒ **Stacking** energy model? **Turner**?
- ▶ Characterized classes are mostly **easy**:
 - ▶ **Designable** classes → Linear time **algorithms**
 - ▶ **Non-designable** classes → Linear time **membership tests**
- ▶ **RNA Design**: P or NP? FPT?
- ▶ **Structural approximation** version of the problem (better ratios? NP-hard beyond some ratio?)

Remarks

- ▶ Results also hold in **Nussinov** energy model (A – U, G – C, G – U + weights)
⇒ **Stacking** energy model? **Turner**?
- ▶ Characterized classes are mostly **easy**:
 - ▶ **Designable** classes → Linear time **algorithms**
 - ▶ **Non-designable** classes → Linear time **membership tests**
- ▶ **RNA Design**: P or NP? FPT?
- ▶ **Structural approximation** version of the problem (better ratios? NP-hard beyond some ratio?)

Part. III: Finding RNAs in genomes

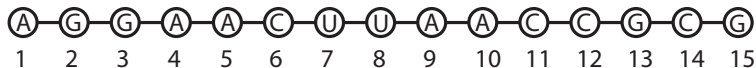
Wei Wang's PhD (collab. LRI@Paris Sud)

Context: Multiple Structural levels

Primary Structure

- ▶ Represents nucleotides sequence
- ▶ No interaction

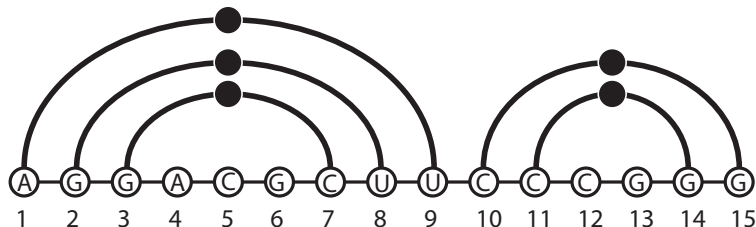
Boring...



Secondary Structure

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

Better...



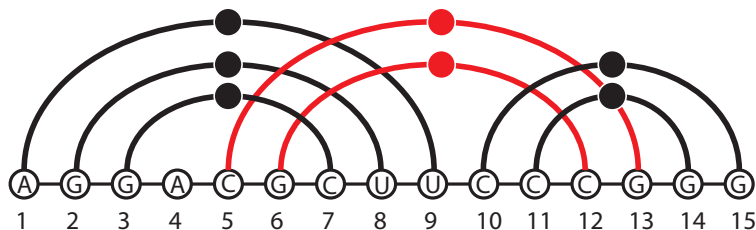
Context: Multiple Structural levels

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!

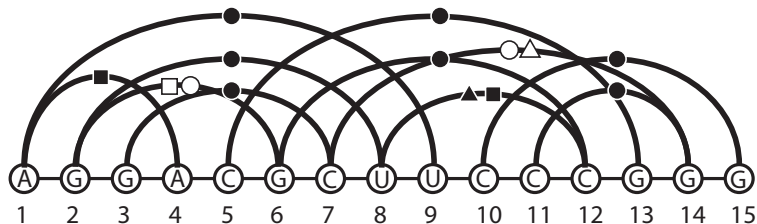


Context: Multiple Structural levels

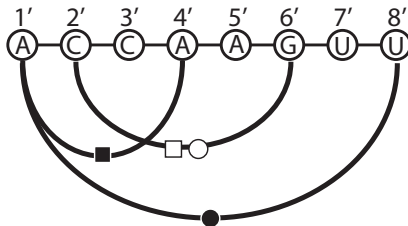
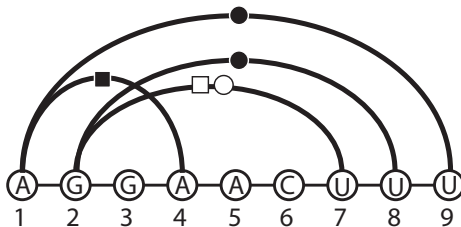
Extended secondary structure

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

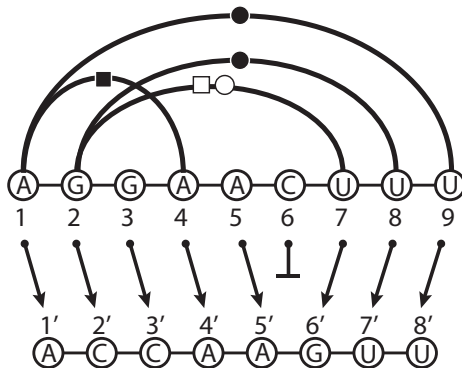
Now we're talking!



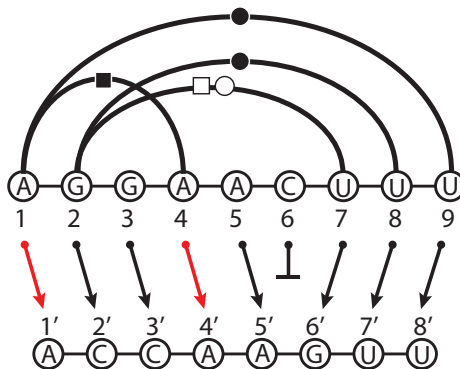
Sequence-structure alignment



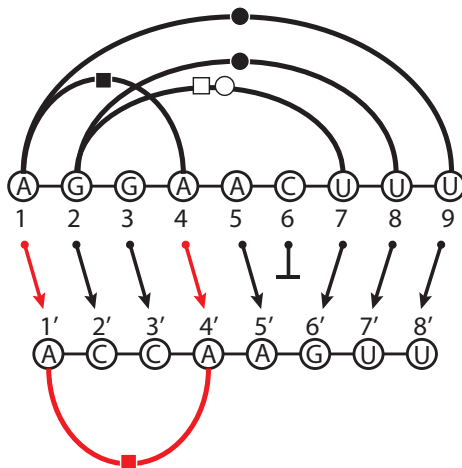
Sequence-structure alignment



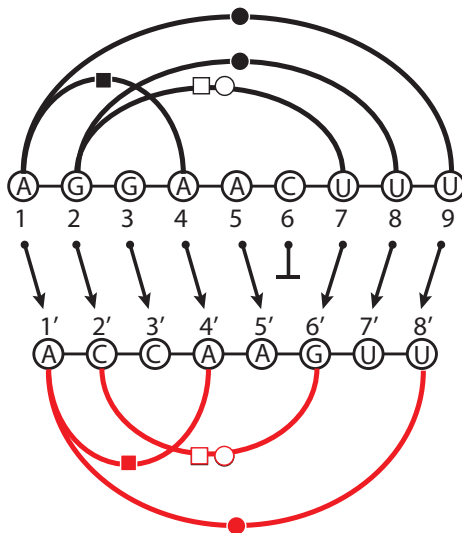
Sequence-structure alignment



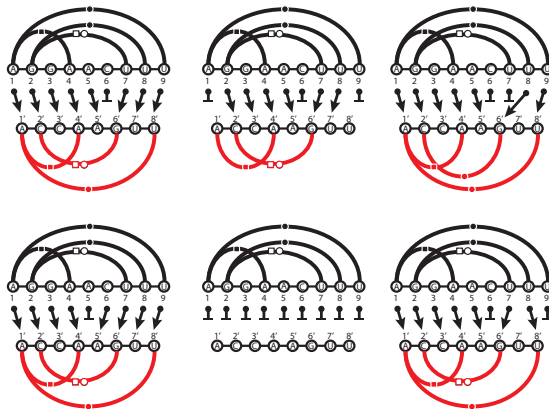
Sequence-structure alignment



Sequence-structure alignment



Sequence-structure alignment



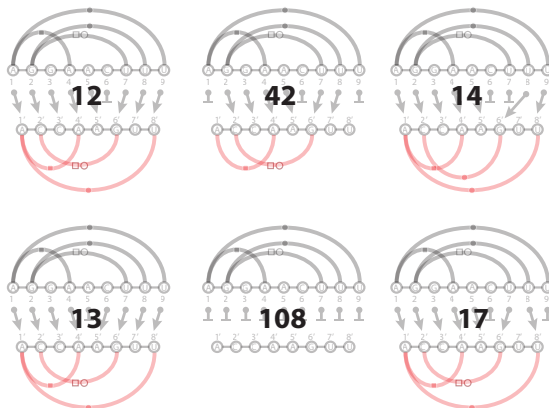
Sequence-structure alignment Problem

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

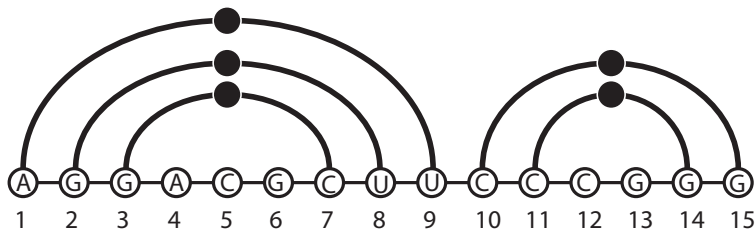
Jiang *et al.* 2001

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

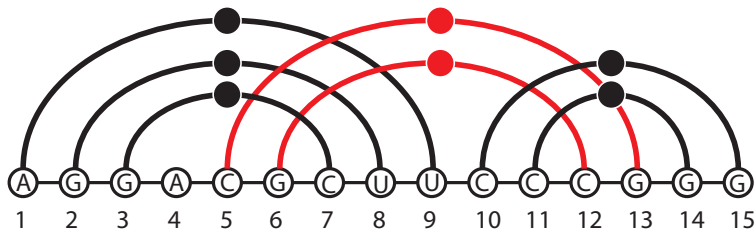


Complexity of structure-sequence alignment

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

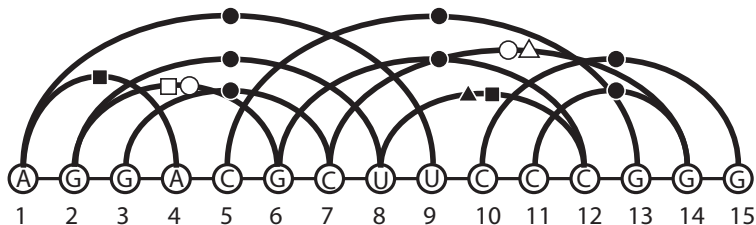


Complexity of structure-sequence alignment

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

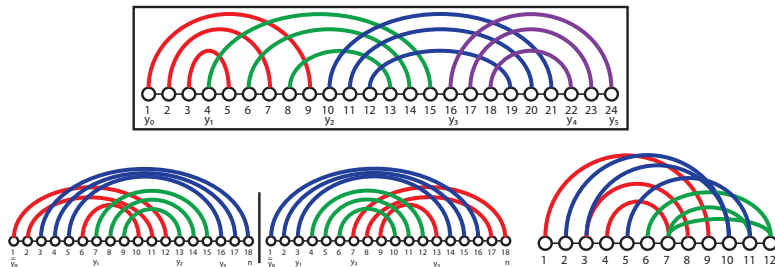
Jiang *et al.* 2001



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	$O(n \cdot m^3)$

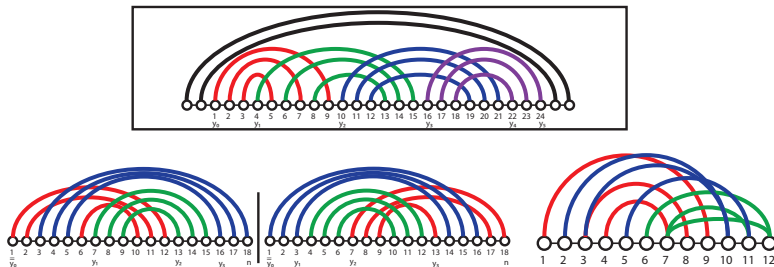


Han *et al.* 2008

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	$O(n \cdot m^3)$

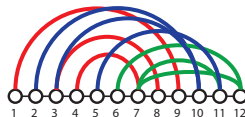
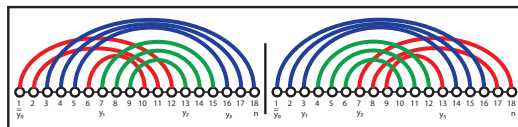
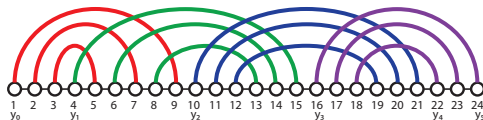


Han *et al.* 2008

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	$O(n \cdot m^3)$

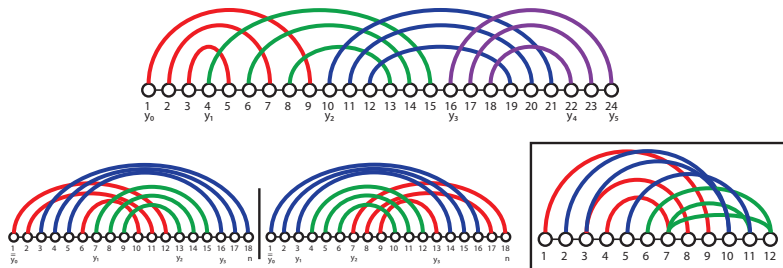


Wong *et al.* 2011

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	$O(n \cdot m^3)$

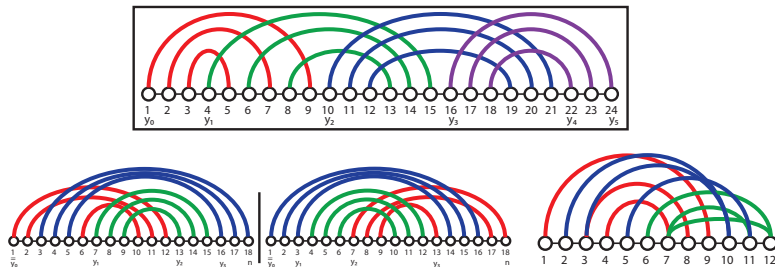


Wong *et al.* 2012

Complexity of struct.-seq. alignment: Polynomial classes

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

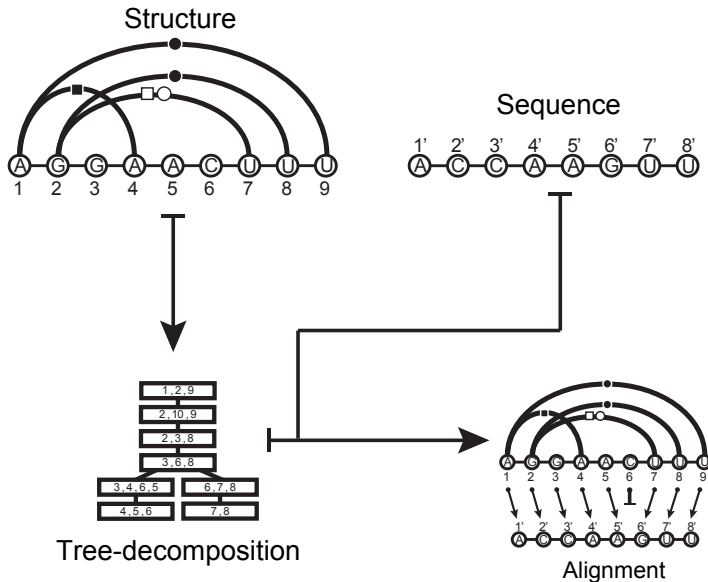
Standard Pseudoknots	$O(n \cdot m^b)$
Standard Embedded Pseudoknots	$O(n \cdot m^{b+1})$
Simple Non-standard Pseudoknots	$O(n \cdot m^{b+1})$
Standard Triple Helices	$O(n \cdot m^3)$



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

[Möhl/Will/Backofen 2009]

Outline of general parameterized approach



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

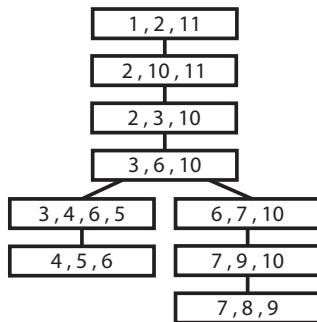
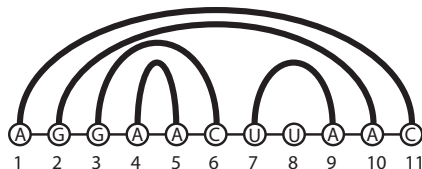
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **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 $x \in \mathcal{B} \cap \mathcal{B}'$, then x is in **every bag** \mathcal{B}'' on the path from \mathcal{B} to \mathcal{B}'



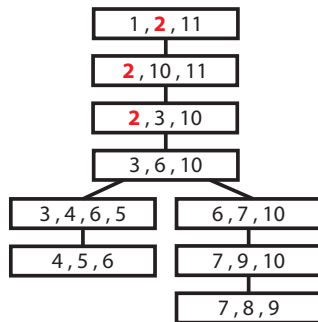
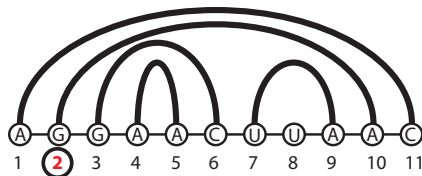
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **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 $x \in \mathcal{B} \cap \mathcal{B}'$, then x is in **every bag** \mathcal{B}'' on the path from \mathcal{B} to \mathcal{B}'



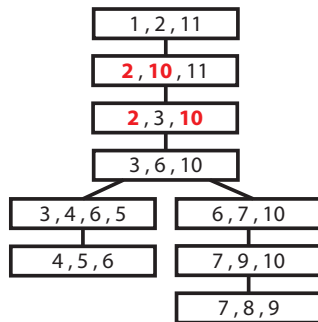
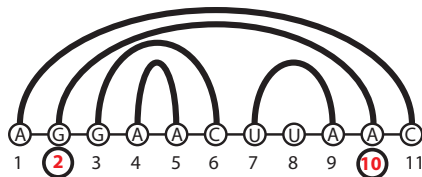
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **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 $x \in \mathcal{B} \cap \mathcal{B}'$, then x is in **every bag** \mathcal{B}'' on the path from \mathcal{B} to \mathcal{B}'



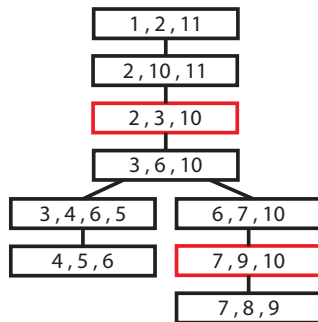
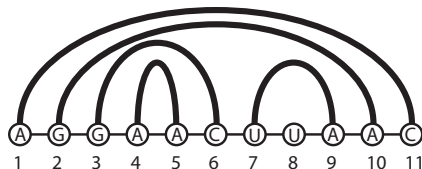
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **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 $x \in \mathcal{B} \cap \mathcal{B}'$, then x is in **every bag** \mathcal{B}'' on the path from \mathcal{B} to \mathcal{B}'



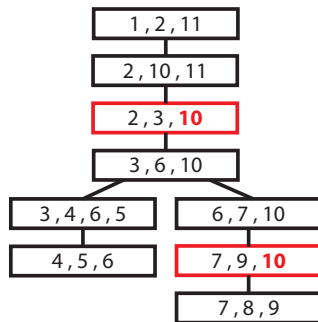
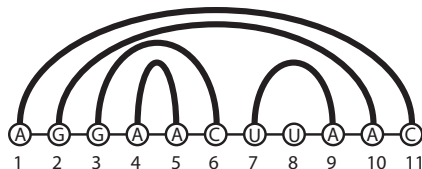
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **Both partners needed to assign score**

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

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



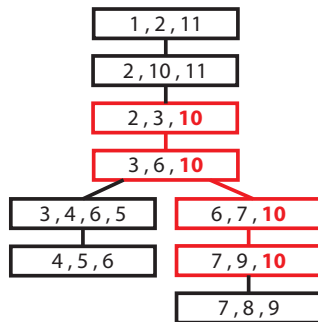
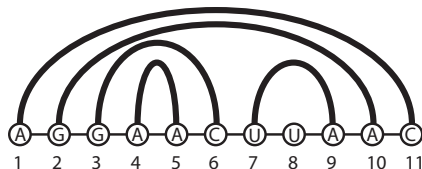
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **Both partners needed to assign score**

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



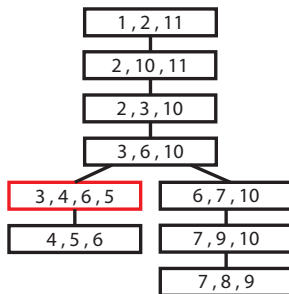
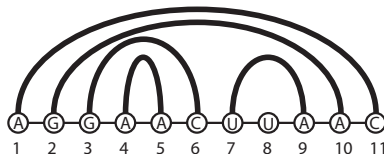
Tree decomposition of RNA structure [Rinaudo et al. 2012]

Structure-centric alignment \Rightarrow Constraints

- ▶ Adjacent positions in structure \rightarrow **Precedence**
- ▶ Paired positions \rightarrow **Both partners needed to assign score**

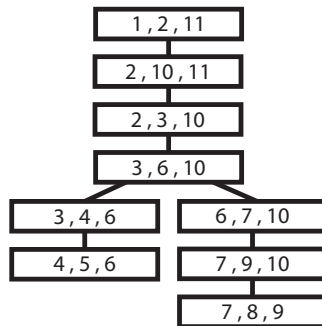
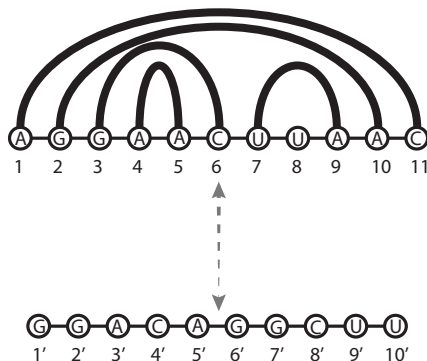
Sets of structure-side positions (**bags** $\{\mathcal{B}_i\}$), in a **tree** such that:

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

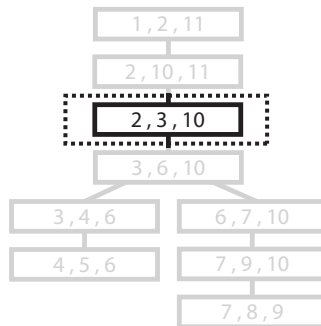
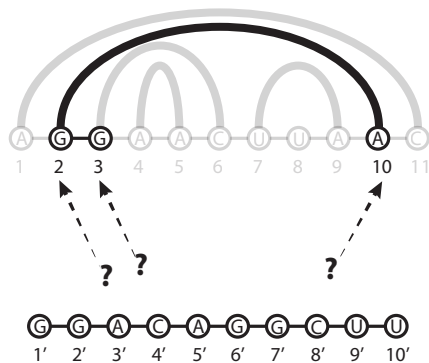


Width k = Size of biggest bag minus one.

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 $\mathcal{O}(n.m^{k+1})/\mathcal{O}(n.m^k)$ time/space \rightarrow **XPT, 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)

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

Specialized complexities

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

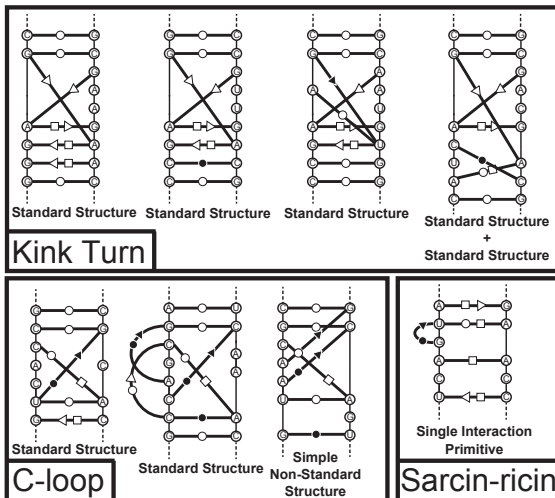
Class of Structures	Time comp.	Multiple interactions	Ref.
Recursive Classical Structures	$O(n \cdot m^{k+2})$	✓	–
└ Secondary Structures (Pseudoknot-free)	$O(n \cdot m^3)$		[Jiang et al 02]
└ Embedded Standard Pseudoknots	$O(n \cdot m^{k+1})$		[Han et al 08]
Standard Structures	$O(n \cdot m^k)$	✓	–
└ 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]



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

$$O(n \cdot m^{k+2})$$

$$O(n \cdot m^k)$$

$$O(n \cdot m^{k+1})$$

$$O(n \cdot m^3)$$

- ▶ Still not a real FPT algorithm! Clues, parameters?
- ▶ Probabilistic interpretation? (MEA, Bayesian networks...)
- ▶ Streaming version of structure/sequence alignment?

Part. IV: Minimal absent words

Definition: Minimal Absent Word

A minimal absent word of a sequence is an absent word whose proper factors (longest prefixes and suffixes) all occur in the sequence.

An upper bound on the number of minimal absent words is $\mathcal{O}(|\Sigma| \cdot n)$, with $|\Sigma|$ the size of the alphabet and n the size of the sequence.

Crochemore et al. 1998, Mignosi et al. 2002

$$S = \overset{0}{A} \overset{1}{A} \overset{2}{C} \overset{3}{A} \overset{4}{C} \overset{5}{A} \overset{6}{C} \overset{7}{C}$$

$MAWs(S) = \{AAA, AACACC, AACC, CAA, CACACA, CCA, CCC\}$

Application in genomic sequence analysis

- ▶ Alignment-free sequence comparison and local genome analysis
Journal of Theoretical Biology, 2012 and 2016, Yang et al.
- ▶ Linear-time sequence comparison using minimal absent words
LATIN 2016, Crochemore et al.

Computing minimal absent words in a sliding window.

Problem: Given a large sequence S of size n , and a smaller sequence W of size L , over a constant size alphabet (of size $|\Sigma|$).

Find the best alignment in terms of minimal absent words, the position x such that:

$$x = \min_{0 \leq i < n} (\text{Comp}(\text{MAWs}(S, i, L), \text{MAWs}(W)))$$

with $\text{MAWs}(S, i, L) = \text{MAWs}(S[i..i + L - 1])$

Goal: Solution in $\mathcal{O}(|\Sigma| \cdot n)$ time and $\mathcal{O}(|\Sigma| \cdot L)$ space .



Typical candidates for **Comp**:

- ▶ **Length weighted index** of the symmetric difference;
- ▶ **Jaccard distance of indices**

were found to perform well to compare sequences [Rahman et al 2016, BMC Research Notes].

We need your help!



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

[Thachuk2010]

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

