Complexity aspects of RNA folding on complex conformation spaces

Saad Sheikh[⊙,◇] Rolf Backofen[♣] Yann Ponty[•,◇]

[⊙] Bloomberg RD, New York, USA

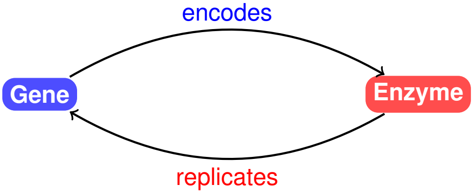
[♣] Albert Ludwigs University, Freiburg, Germany

[•] LIX, CNRS/Ecole Polytechnique, France

[◇] AMIB Team-Project, INRIA, Saclay, France

Sep 30th – MBI workshop’15

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

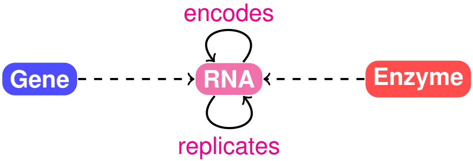


A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

[...] 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…



The RNA World
THIRD EDITION
PROTEIN
mRNA
RNA
RNA
DNA
RNA
Edited by Raymond F. Gesteland
Thomas R. Cech
John F. Atkins

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*.

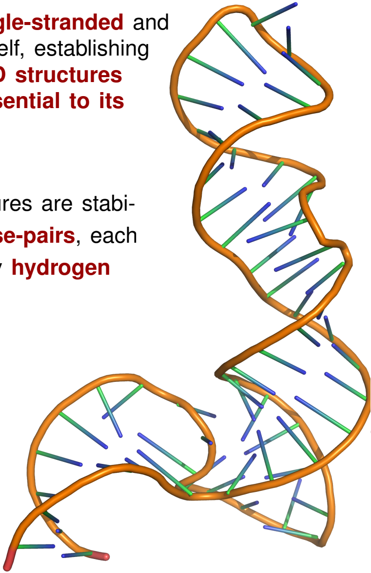
[. . .] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

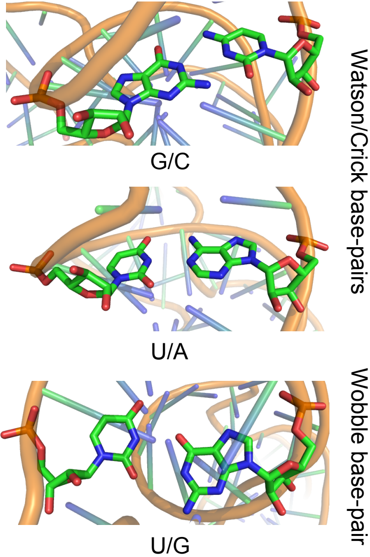
R. Dawkins. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

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**.

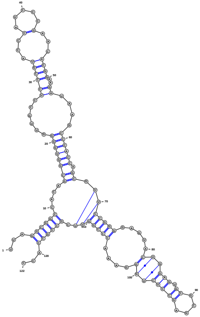


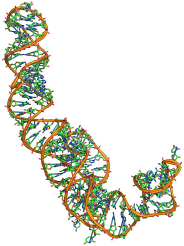


Canonical base-pairs

RNA structure(s)

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





Primary structure

Secondary structure

Tertiary structure

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

Definition (Secondary Structure)

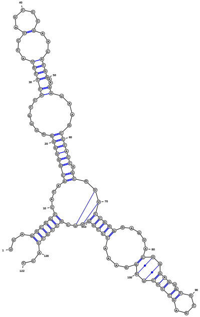
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i,j) \in [1,n]^2$ such that:

- **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).

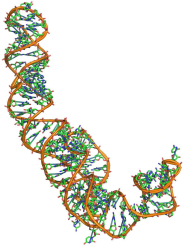
RNA structure(s)

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Primary structure



Secondary structure



Tertiary structure

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

Definition (Secondary Structure)

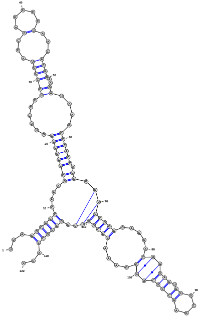
A **secondary structure** S for an RNA w is a set of **base-pairs** $(i,j) \in [1,n]^2$ such that:

- **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).

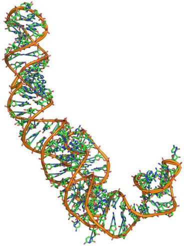
RNA structure(s)

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

Primary structure



Secondary structure



Tertiary structure

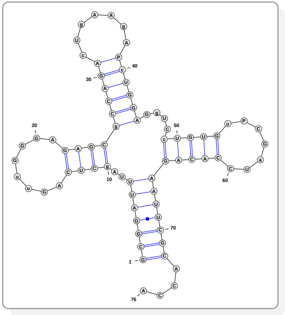
Source: 5s rRNA (PDBID: 1K73:B)

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*

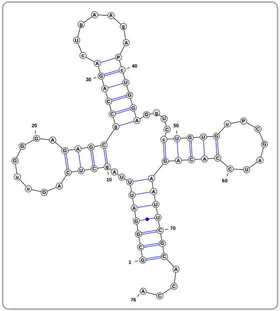
Diversity supports intuitions

Different representations

Common combinatorial structure

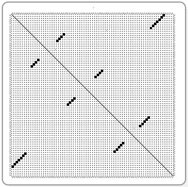
*Additional steric constraints

Various representations for a versatile biomolecule



Outer-planar graphs

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



Dot plots

Adjacency matrices*

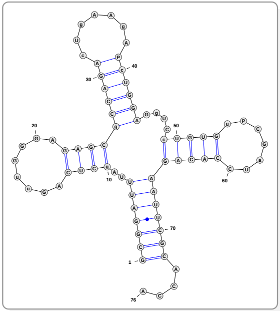
Diversity supports intuitions

Different representations

Common combinatorial structure

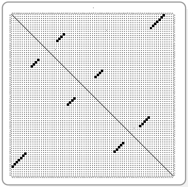
*Additional steric constraints

Various representations for a versatile biomolecule



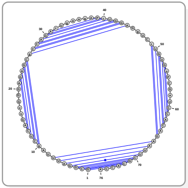
Outer-planar graphs

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



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

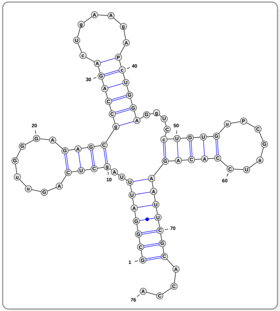
Diversity supports intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

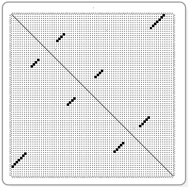


Outer-planar graphs

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

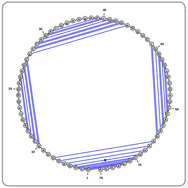
```
(((((((..((((........))))((((((.......))))))....(((((.......))))))))))))....
```

Motzkin words*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

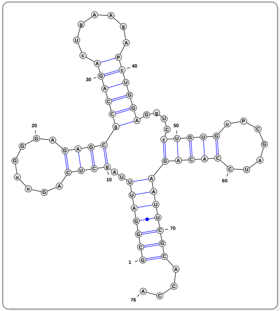
Diversity supports intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

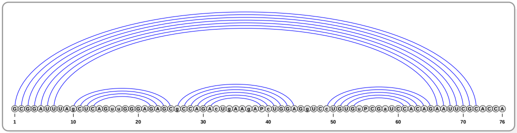


Outer-planar graphs

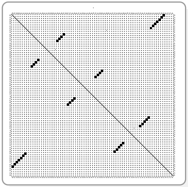
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

Motzkin words*

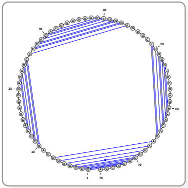


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

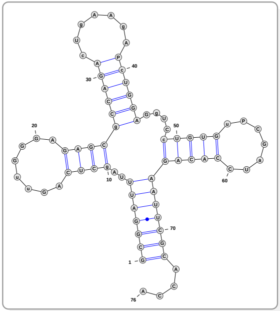
Diversity supports intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

Various representations for a versatile biomolecule

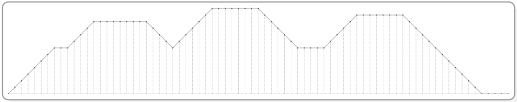


Outer-planar graphs

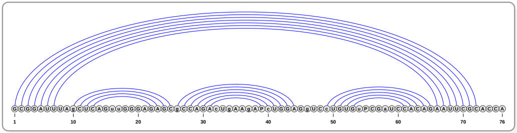
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

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

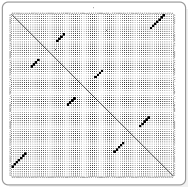
Motzkin words*



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

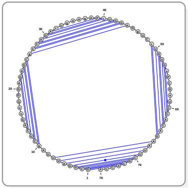


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

Diversity supports intuitions

Different representations

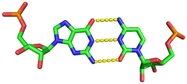
Common combinatorial structure

*Additional steric constraints

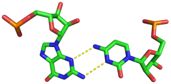
Crossing interactions

MFE structure, part. func.... computed exactly in $\Theta(n^3)$ time in the absence of:

- **Non-canonical base-pairs:**
  Any base-pair **other than** {(A-U), (C-G), (G-U)}
  OR interacting in a non-standard way (WC/WC-Cis) [Leontis 01].

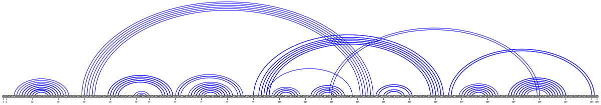


Canonical CG base-pair (WC/WC-Cis)



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

- **(Pseudo?)knots:** Crossing sets of nested stable base-pairs

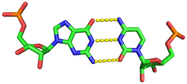


Group I Ribozyme (PDBID: 1Y0Q:A)

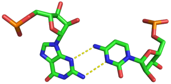
Crossing interactions

MFE structure, part. func.. . . computed exactly in $\Theta(n^3)$ time in the absence of:

- **Non-canonical base-pairs:**
  Any base-pair **other than** {(A-U), (C-G), (G-U)}
  OR interacting in a non-standard way (WC/WC-Cis) [Leontis 01].

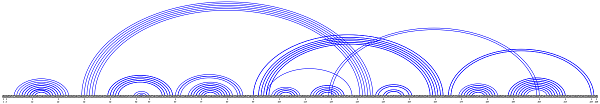


Canonical CG base-pair (WC/WC-Cis)



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

- **(Pseudo?)knots:** Crossing sets of nested stable base-pairs

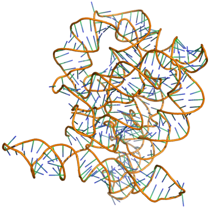


Group I Ribozyme (PDBID: 1Y0Q:A)

Crossing interactions

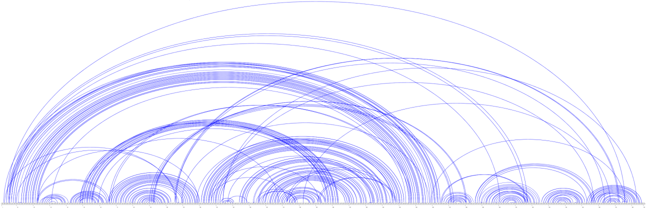
MFE structure, part. func.. . . computed exactly in $\Theta(n^3)$ time in the absence of:

- **Non-c**
  Any ba
  OR int



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

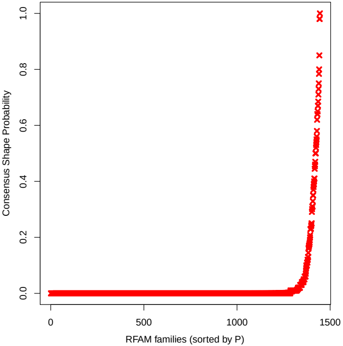
Example: Group II Intron (PDB ID: 3IGI)

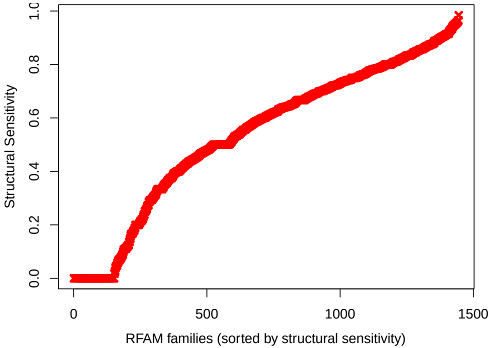


- **(Pseu**

(Pseudo)-knots: They Walk Amongst Us…

…and are responsible for (some of) the shortcomings of predictive tools.





Software: RNAFold **Database:** RFAM Release 10.1

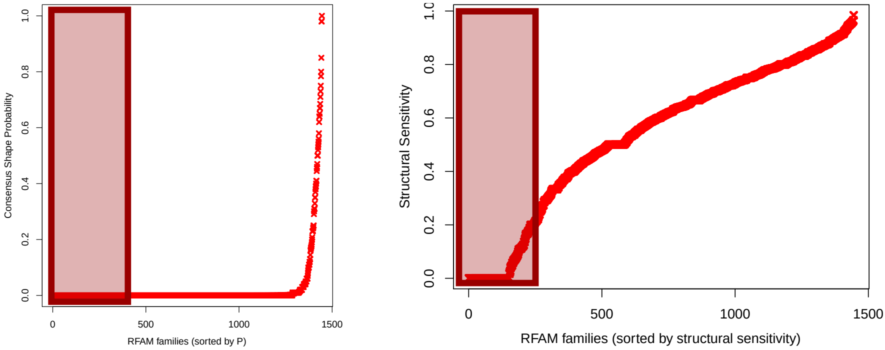
Many of **those** families feature pseudoknots (RFAM consensus or predictions)

⇒ **Include pseudoknots to folding space of structure prediction algorithms**

Looks tough (this talk) but restricting the search space helps (Orland's talk)

(Pseudo)-knots: They Walk Amongst Us…

…and are responsible for (some of) the shortcomings of predictive tools.



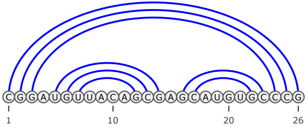
Software: RNAFold **Database:** RFAM Release 10.1

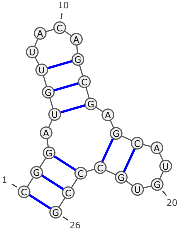
Many of **those** families feature pseudoknots (RFAM consensus or predictions)

⇒ **Include pseudoknots to folding space of structure prediction algorithms**

Looks tough (this talk) but restricting the search space helps (Orland's talk)

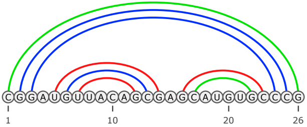
Problem statement

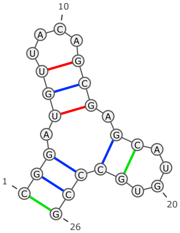




- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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

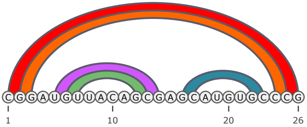
Problem statement

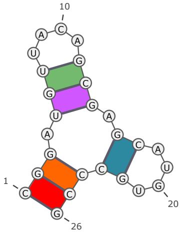




- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^{-} \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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

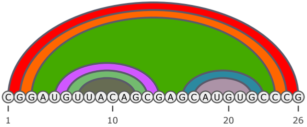
Problem statement

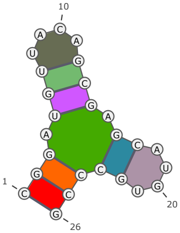




- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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

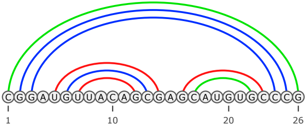
Problem statement

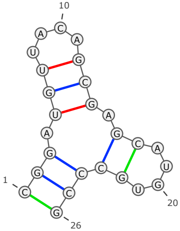




- **RNA structure S:** (Partial) matching of positions in sequence w
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif → Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^{-} \cup \{+\infty\}$
  Free-Energy $E_w(S)$: Sum over (independently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

Problem statement

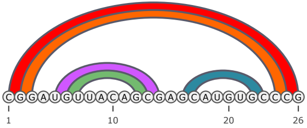


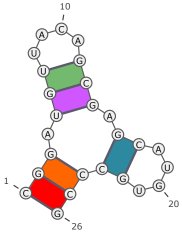


- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^{-} \cup \{+\infty\}$
  Free-Energy $E_w(S)$: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

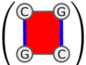
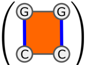
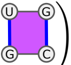
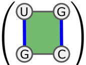
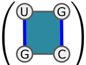
$$E_S = 2 \cdot \Delta\begin{pmatrix} U \\ | \\ G \end{pmatrix} + 4 \cdot \Delta\begin{pmatrix} G \\ | \\ C \end{pmatrix} + 2 \cdot \Delta\begin{pmatrix} C \\ | \\ G \end{pmatrix}$$

Problem statement

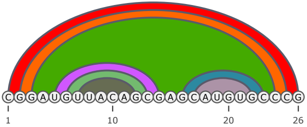


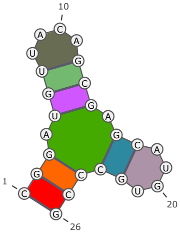


- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: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

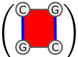
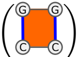
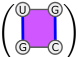
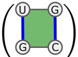
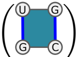
$E_S = \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg)$

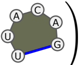
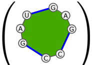
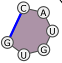
Problem statement



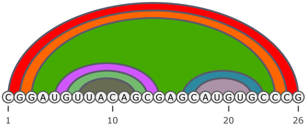


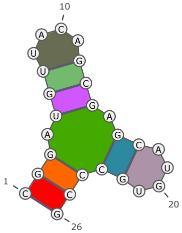
- **RNA structure *S*:** (Partial) matching of positions in sequence *w*
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops. . .)
- **Energy model:**
  Motif $\rightarrow$ Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
  Free-Energy $E_w(S)$: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

$E_S = \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg)$

$+ \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg) + \Delta\Bigg($  $\Bigg)$

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

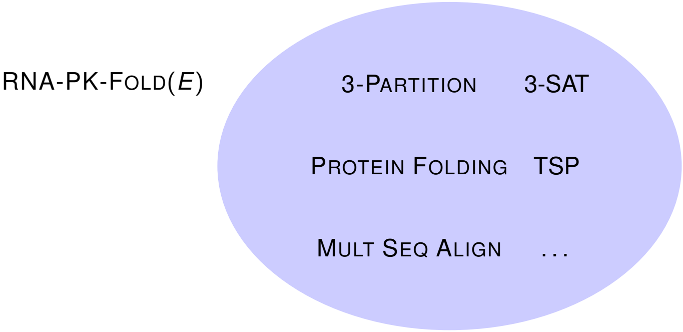
Definition (RNA-PK-FOLD(E) problem)

Input: RNA sequence $w \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^*$.
Output: Matching S^*, having Minimal Free-Energy (MFE) $E_w(S^*)$.

Are there **efficient** algorithms to predict MFE with **arbitrary** (pseudo)knots?
With **restricted** pseudoknots? On which **energy models**?

NP-hardness: Characterizing the difficulty of algorithmic problems

Question(s): Is RNA-PK-Fold(E) intrinsically difficult?



RNA-PK-Fold(E)

3-Partition

3-SAT

Protein Folding

TSP

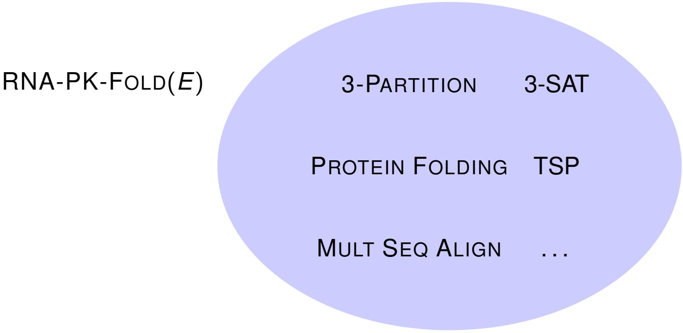
Mult Seq Align

...

NP-Complete problems

NP-hardness: Characterizing the difficulty of algorithmic problems

Question(s): ~~Is RNA-PK-FOLD(E) intrinsically difficult?~~
Is RNA-PK-FOLD(E) **as hard as** some reference **hard problem(s)**?



RNA-PK-FOLD(E)

3-PARTITION 3-SAT

PROTEIN FOLDING TSP

MULT SEQ ALIGN . . .

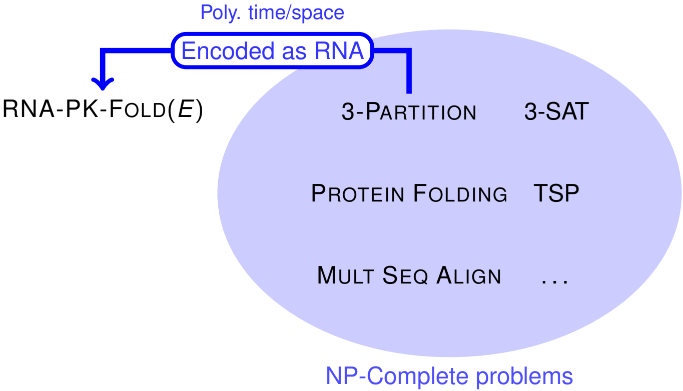
NP-Complete problems

NP-hardness: Characterizing the difficulty of algorithmic problems

Question(s): ~~Is RNA-PK-FOLD(E) intrinsically difficult?~~

Is RNA-PK-FOLD(E) **as hard as** some reference **hard problem(s)**?

⇔ Would solving RNA-PK-FOLD(E) **in polynomial time** (P) lead to a **polynomial-time algorithm** for other hard problems?

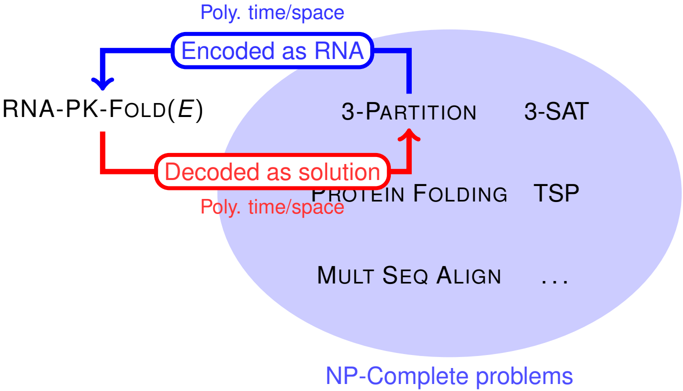


NP-hardness: Characterizing the difficulty of algorithmic problems

Question(s): ~~Is RNA-PK-FOLD(E) intrinsically difficult?~~

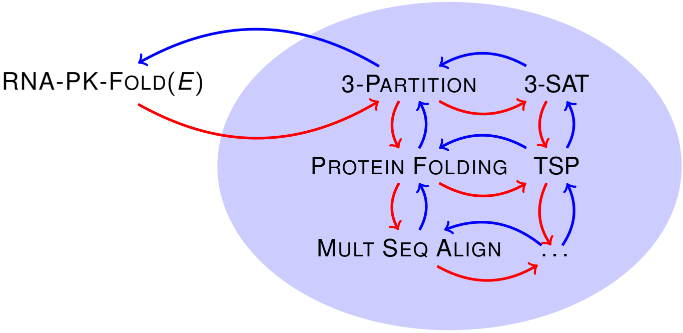
Is RNA-PK-FOLD(E) **as hard as** some reference **hard problem(s)**?

⇔ Would solving RNA-PK-FOLD(E) **in polynomial time** (P) lead to a **polynomial-time algorithm** for other hard problems?



NP-hardness: Characterizing the difficulty of algorithmic problems

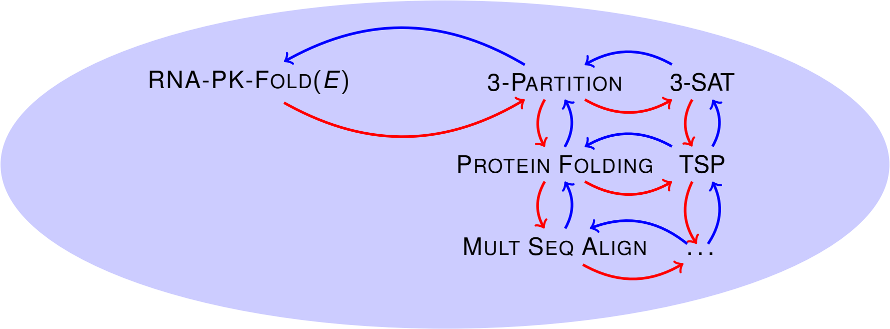
Question(s): ~~Is RNA-PK-FOLD(E) intrinsically difficult?~~
Is RNA-PK-FOLD(E) **as hard as** some reference **hard problem(s)**?
⇔ Would solving RNA-PK-FOLD(E) **in polynomial time** (P) lead to a **polynomial-time algorithm** for other hard problems?



NP-Complete problems

NP-hardness: Characterizing the difficulty of algorithmic problems

Question(s): ~~Is RNA-PK-FOLD(E) intrinsically difficult?~~
Is RNA-PK-FOLD(E) **as hard as** some reference **hard problem(s)**?
⇔ Would solving RNA-PK-FOLD(E) **in polynomial time** (P) lead to a **polynomial-time algorithm** for other hard problems?

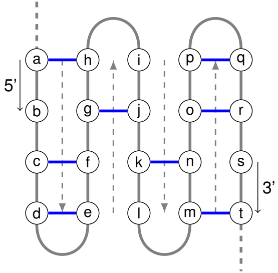


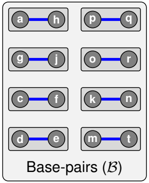
NP-Complete problems

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 $\rightarrow \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$



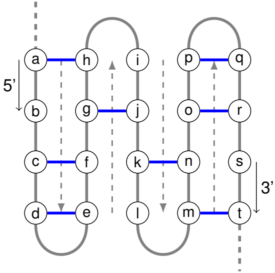


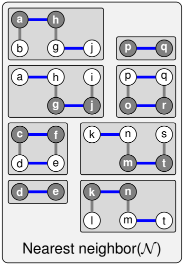
Solved in $\mathcal{O}(n^3)$ [Tabaska 98]
(Max-weighted matching)
Unrealistic!

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





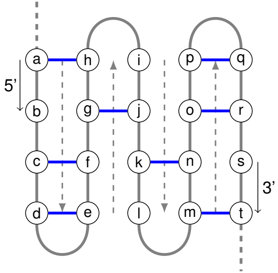
NP-hard [Lyngsø 00, Akutsu 00]

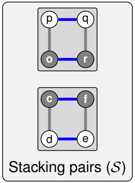
Too expressive?

Energy models

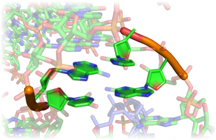
Three models, based on interacting positions (i,j):

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





Stacking pairs ($\mathcal{S}$)

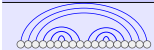
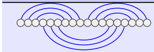
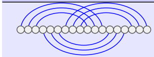


Captures stablest motifs

Still NP-hard [Lyngsø 04]

. . . but PTAS [Lyngsø 04]

State of the art

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04]	???

Missing:

- Qualitative difference between Stacking-pairs and Nearest-Neighbor models?
- Influence of $\mathcal{M}$ on hardness/approx. ratio (only unit-valued studied)

Biologists demand (Biology deserves) **honest hardness results**:

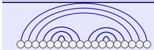
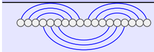
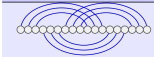
- Energy model as input: Pandora's box (e.g. RNA folding on infinite alphabet!)
- Model as parameter: Is problem hard…

  Sometimes ($\exists\mathcal{M}$)? → **Dishonest**

  Always ($\forall\mathcal{M}$)? Almost surely (w. p. 1)? → **Honest**

  Under reasonable assumptions + $\forall$ parameterization? → **Almost honest**

State of the art

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04]	???

Missing:

- Qualitative difference between Stacking-pairs and Nearest-Neighbor models?
- Influence of $\mathcal{M}$ on hardness/approx. ratio (only unit-valued studied)

Biologists demand (Biology deserves) **honest hardness results**:

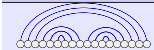
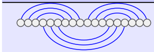
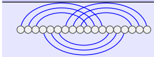
- Energy model as input: Pandora's box (e.g. RNA folding on infinite alphabet!)
- Model as parameter: Is problem hard. . .

  Sometimes ($\exists\mathcal{M}$)? → **Dishonest**

  Always ($\forall\mathcal{M}$)? Almost surely (w. p. 1)? → **Honest**

  Under reasonable assumptions + $\forall$ parameterization? → **Almost honest**

State of the art

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04]	???

Missing:

- Qualitative difference between Stacking-pairs and Nearest-Neighbor models?
- Influence of $\mathcal{M}$ on hardness/approx. ratio (only unit-valued studied)

Biologists demand (Biology deserves) **honest hardness results**:

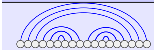
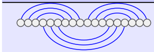
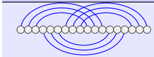
- Energy model as input: Pandora's box (e.g. RNA folding on infinite alphabet!)
- Model as parameter: Is problem hard. . .

  Sometimes ($\exists\mathcal{M}$)? → **Dishonest**

  Always ($\forall\mathcal{M}$)? Almost surely (w. p. 1)? → **Honest**

  Under reasonable assumptions + $\forall$ parameterization? → **Almost honest**

State of the art

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04]	???

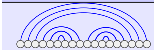
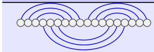
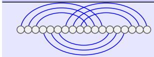
Missing:

- Qualitative difference between Stacking-pairs and Nearest-Neighbor models?
- Influence of $\mathcal{M}$ on hardness/approx. ratio (only unit-valued studied)

Biologists demand (Biology deserves) **honest hardness results**:

- Energy model as input: Pandora's box (e.g. RNA folding on infinite alphabet!)
- Model as parameter: Is problem hard...
  Sometimes ($\exists\mathcal{M}$)? → **Dishonest**
  Always ($\forall\mathcal{M}$)? Almost surely (w. p. 1)? → **Honest**
  Under reasonable assumptions + $\forall$ parameterization? → **Almost honest**

State of the art

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04]	???

Missing:

- Qualitative difference between Stacking-pairs and Nearest-Neighbor models?
- Influence of $\mathcal{M}$ on hardness/approx. ratio (only unit-valued studied)

Biologists demand (Biology deserves) **honest hardness results**:

- Energy model as input: Pandora's box (e.g. RNA folding on infinite alphabet!)
- Model as parameter: Is problem hard. . .
  Sometimes ($\exists\mathcal{M}$)? → **Dishonest**
  Always ($\forall\mathcal{M}$)? Almost surely (w. p. 1)? → **Honest**
  Under reasonable assumptions + ∀ parameterization? → **Almost honest**

(Almost!)-honest hardness of RNA-PK-FOLD($\mathcal{S}$)

For any **stacking energy model $\mathcal{S}$,** such that:

- Only G/C, A/U and G/U pairs are allowed
- Any other X/Y pair forbidden

$$\Rightarrow \Delta_{\mathcal{S}}(X, Y, *, *) = +\infty$$

(Such BPs are rarely observed [Stombaugh 09]$\longrightarrow$ Unstable)

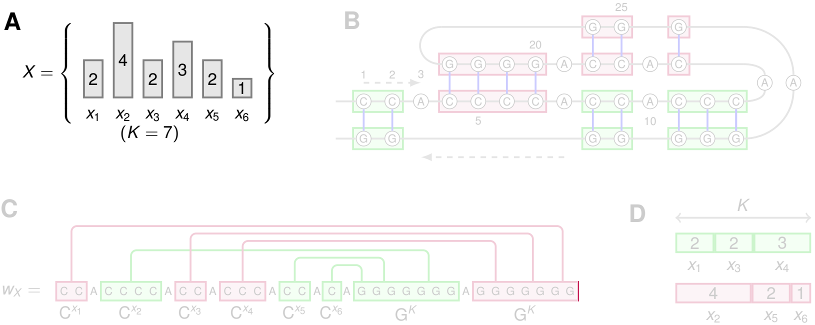
- Arbitrary energies associated with valid stackings

$$\Rightarrow \Delta_{\mathcal{S}}(X, Y, X', Y') < 0$$

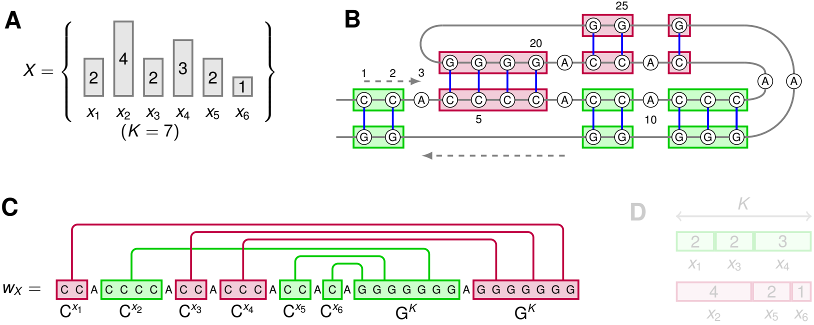
Theorem

RNA-PK-FOLD($\mathcal{S}$) *is* NP*-hard.*

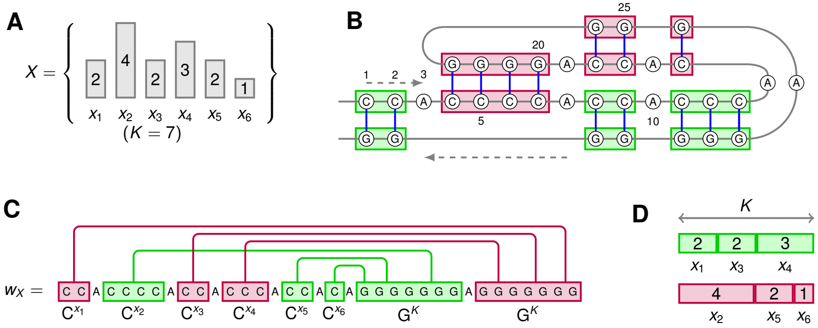
Example


A
X =
2
4
2
3
2
1
x1 x2 x3 x4 x5 x6
(K = 7)
B
25
20
1 2 3
5
10
C
wX =
C C A C C C C A C C A C C C A C C A C A G G G G G G G A G G G G G G G
C^x1 C^x2 C^x3 C^x4 C^x5 C^x6 G^K G^K
D
K
2 2 3
x1 x3 x4
4 2 1
x2 x5 x6

Example


A
X =
2
4
2
3
2
1
x1
x2
x3
x4
x5
x6
(K = 7)
B
25
20
1
2
3
5
10
G G G
G G G G A C C A C
C C A C C C C A C C A C C C
A A
G G G G G G G
C
w_X =
C C A C C C C A C C A C C C A C C A C A G G G G G G G A G G G G G G G
C^x1
C^x2
C^x3
C^x4
C^x5
C^x6
G^K
G^K
D
K
2
2
3
x1
x3
x4
4
2
1
x2
x5
x6

Example


A
$X = \{ 2, 4, 2, 3, 2, 1 \}$
x_1 x_2 x_3 x_4 x_5 x_6
$(K = 7)$
B
1 2 3
5
10
20
25
G G G G A C C A C
C C A C C C C A C C A C C C
G G G G G G G
G G G
A A
C
$w_X =$ C C A C C C C A C C A C C C A C C A C A G G G G G G G A G G G G G G G
C^{x_1} C^{x_2} C^{x_3} C^{x_4} C^{x_5} C^{x_6} G^K G^K
D
K
2 2 3
x_1 x_3 x_4
4 2 1
x_2 x_5 x_6

Proof

Definition (3-PARTITION problem)

Input: Sequence of integers $X = \{x_i\}_{i=1}^{n}$, summing to $n/3 \cdot K$, $K \in \mathbb{N}$.
Output: `True` iff X can be split into $m := n/3$ triplets $\{(x_{a_j}, x_{b_j}, x_{c_j})\}_{j=1}^{m}$ s. t.

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m].$$

Proof. Reduction from 3-PARTITION:

- Let $w_X := \mathrm{C}^{x_1}\mathrm{A}\mathrm{C}^{x_2}\mathrm{A}\mathrm{C}^{x_3}\mathrm{A}\cdots\mathrm{A}\mathrm{C}^{x_n}\underbrace{\mathrm{A}\mathrm{G}^{K}\mathrm{A}\mathrm{G}^{K}\mathrm{A}\cdots\mathrm{A}\mathrm{G}^{K}}_{m\text{ times}}$ and $\delta := \Delta_{\mathcal{S}}(\mathrm{C},\mathrm{G},\mathrm{C},\mathrm{G})$
- Best matching S^* for w_X has free-energy $E(S^*)_{w_X} \leq E^* := \delta \cdot (K-3) \cdot m$.
- If X 3-partitionable, then matching induced by partition gives $E(S^*)_{w_X} = E^*$.
- If $E(S^*)_{w_X} = E^*$, then S^* *saturates* each G^K block, using three blocks $(\mathrm{C}^a, \mathrm{C}^b, \mathrm{C}^c)$.
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then RNA-PK-FOLD$(\mathcal{S}) \in \mathrm{P} \Rightarrow$ 3-PARTITION $\in \mathrm{P}$.

Reminder: 3-PARTITION is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

Proof

Definition (3-Partition problem)

Input: Sequence of integers $X = \{x_i\}_{i=1}^{n}$, summing to $n/3 \cdot K$, $K \in \mathbb{N}$.
Output: `True` iff X can be split into $m := n/3$ triplets $\{(x_{a_j}, x_{b_j}, x_{c_j})\}_{j=1}^{m}$ s. t.

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m].$$

Proof. Reduction from 3-Partition:

- Let $w_X := \mathrm{C}^{x_1}\mathrm{A}\mathrm{C}^{x_2}\mathrm{A}\mathrm{C}^{x_3}\mathrm{A}\cdots\mathrm{A}\mathrm{C}^{x_n}\underbrace{\mathrm{A}\mathrm{G}^{K}\mathrm{A}\mathrm{G}^{K}\mathrm{A}\cdots\mathrm{A}\mathrm{G}^{K}}_{m \text{ times}}$ and $\delta := \Delta_{\mathcal{S}}(\mathrm{C},\mathrm{G},\mathrm{C},\mathrm{G})$
- Best matching S^* for w_X has free-energy $E(S^*)_{w_X} \leq E^* := \delta \cdot (K-3) \cdot m$.
- If X 3-partitionable, then matching induced by partition gives $E(S^*)_{w_X} = E^*$.
- If $E(S^*)_{w_X} = E^*$, then S^* *saturates* each G^K block, using three blocks $(\mathrm{C}^a, \mathrm{C}^b, \mathrm{C}^c)$.
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then RNA-PK-Fold$(\mathcal{S}) \in \mathrm{P} \Rightarrow$ 3-Partition $\in \mathrm{P}$.

Reminder: 3-Partition is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

Proof

Definition (3-PARTITION problem)

Input: Sequence of integers $X = \{x_i\}_{i=1}^{n}$, summing to $n/3 \cdot K$, $K \in \mathbb{N}$.
Output: `True` iff X can be split into $m := n/3$ triplets $\{(x_{a_j}, x_{b_j}, x_{c_j})\}_{j=1}^{m}$ s. t.

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m].$$

Proof. Reduction from 3-PARTITION:

- Let $w_X := \mathrm{C}^{x_1}\mathrm{AC}^{x_2}\mathrm{AC}^{x_3}\mathrm{A}\cdots\mathrm{AC}^{x_n}\underbrace{\mathrm{AG}^K\mathrm{AG}^K\mathrm{A}\cdots\mathrm{AG}^K}_{m \text{ times}}$ and $\delta := \Delta_{\mathcal{S}}(\mathrm{C},\mathrm{G},\mathrm{C},\mathrm{G})$
- Best matching S^* for w_X has free-energy $E(S^*)_{w_X} \leq E^* := \delta \cdot (K-3) \cdot m$.
- If X 3-partitionable, then matching induced by partition gives $E(S^*)_{w_X} = E^*$.
- If $E(S^*)_{w_X} = E^*$, then S^* *saturates* each G^K block, using three blocks $(\mathrm{C}^a, \mathrm{C}^b, \mathrm{C}^c)$.
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then RNA-PK-FOLD$(\mathcal{S}) \in \mathrm{P} \Rightarrow$ 3-PARTITION $\in$ P.

Reminder: 3-PARTITION is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

Proof

Definition (3-PARTITION problem)

Input: Sequence of integers $X = \{x_i\}_{i=1}^{n}$, summing to $n/3 \cdot K$, $K \in \mathbb{N}$.
Output: `True` iff X can be split into $m := n/3$ triplets $\{(x_{a_j}, x_{b_j}, x_{c_j})\}_{j=1}^{m}$ s. t.

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m].$$

Proof. Reduction from 3-PARTITION:

- Let $w_X := \text{C}^{x_1}\text{AC}^{x_2}\text{AC}^{x_3}\text{A}\cdots\text{AC}^{x_n}\underbrace{\text{AG}^K\text{AG}^K\text{A}\cdots\text{AG}^K}_{m\text{ times}}$ and $\delta := \Delta_{\mathcal{S}}(\text{C},\text{G},\text{C},\text{G})$
- Best matching S^* for w_X has free-energy $E(S^*)_{w_X} \leq E^* := \delta \cdot (K-3) \cdot m$.
- If X 3-partitionable, then matching induced by partition gives $E(S^*)_{w_X} = E^*$.
- If $E(S^*)_{w_X} = E^*$, then S^* *saturates* each G^K block, using three blocks $(\text{C}^a, \text{C}^b, \text{C}^c)$.
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then RNA-PK-FOLD$(\mathcal{S}) \in \text{P} \Rightarrow$ 3-PARTITION $\in$ P.

Reminder: 3-PARTITION is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

Proof

Definition (3-PARTITION problem)

Input: Sequence of integers $X = \{x_i\}_{i=1}^{n}$, summing to $n/3 \cdot K$, $K \in \mathbb{N}$.
Output: `True` iff X can be split into $m := n/3$ triplets $\{(x_{a_j}, x_{b_j}, x_{c_j})\}_{j=1}^{m}$ s. t.

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m].$$

Proof. Reduction from 3-PARTITION:

- Let $w_X := \mathrm{C}^{x_1}\mathrm{A}\mathrm{C}^{x_2}\mathrm{A}\mathrm{C}^{x_3}\mathrm{A}\cdots\mathrm{A}\mathrm{C}^{x_n}\underbrace{\mathrm{A}\mathrm{G}^{K}\mathrm{A}\mathrm{G}^{K}\mathrm{A}\cdots\mathrm{A}\mathrm{G}^{K}}_{m\text{ times}}$ and $\delta := \Delta_{\mathcal{S}}(\mathrm{C}, \mathrm{G}, \mathrm{C}, \mathrm{G})$
- Best matching S^* for w_X has free-energy $E(S^*)_{w_X} \leq E^* := \delta \cdot (K - 3) \cdot m$.
- If X 3-partitionable, then matching induced by partition gives $E(S^*)_{w_X} = E^*$.
- If $E(S^*)_{w_X} = E^*$, then S^* *saturates* each G^K block, using three blocks $(\mathrm{C}^a, \mathrm{C}^b, \mathrm{C}^c)$.
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then RNA-PK-FOLD$(\mathcal{S}) \in \mathrm{P} \Rightarrow$ 3-PARTITION $\in \mathrm{P}$.

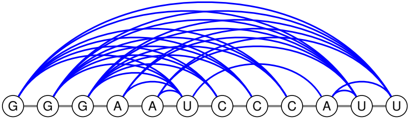
Reminder: 3-PARTITION is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

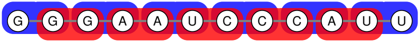


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

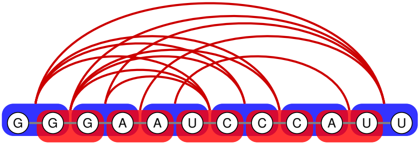


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

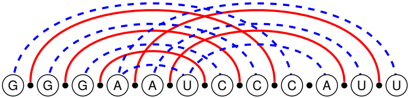


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

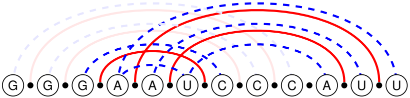


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \rightarrow (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

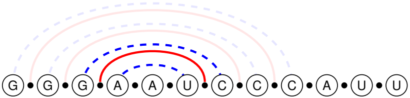


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

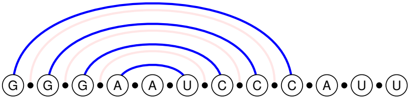


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) ⇒ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

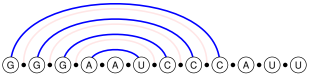


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD($\mathcal{S}$)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\varepsilon}})$) [Lyngsø 04]
- Base-pair maximization (unit cost) $\Rightarrow$ Arbitrary energies???

Algorithm:

1. Build weighted adjacency graph $G = (V, E)$
   - Vertices: Pairs of consecutive pos. $(i, i+1)$
   - Edges: $(i, i+1) \to (j-1, j)$ with weight $-\Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$
2. Compute maximal-weighted matching m'.
3. Loop over $p = (i, i+1), (j, j-1) \in m'$, ordered by decreasing weight:
   - Add result to output m, remove any $p' \in m'$ conflicting with p
4. Return m

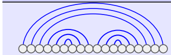
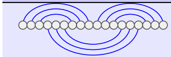
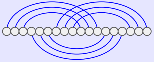


Complexity: At most $\mathcal{O}(n^3)$ (Max-weighted matching)

Approx. ratio: Initial matching m' has total energy smaller than OPT.
Loop 3: Each stacking pair p conflicts with ≤ 4 **pairs** in m', **having greater energy**.

$\Rightarrow$ Returned matching has free-energy $\leq 1/5$ of OPT ($\forall \mathcal{S} \to$ **Honest**)

Half-time summary

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04] **(any* Δ model)**	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04] **1/5 (any Δ model)**	**???**

How hard is it to approximate the nearest neighbor model?

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

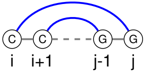
For some nearest-neighbor model $\mathcal{N}$, one has $\text{RNA-PK-F}\textsc{old}(\mathcal{N}) \notin APX$.

Proof. Consider the RNA seq. built from some 3-PARTITION instance X:

$$w_X = \text{C}^{x_1}\text{AC}^{x_2}\text{A}\cdots\text{AC}^{x_{3m}}\text{A}\,\underbrace{\text{G}^K\text{UG}^K\text{U}\cdots\text{G}^K\text{U}}_{m\text{ times}}\,\text{U}^{2m}$$

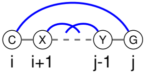
and the energy model:

$\Delta^*_{\mathcal{N}}$: (A)



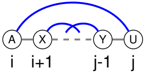
$\longrightarrow -1, \quad \forall i < j,$

(B)



$\longrightarrow -1, \quad \forall i < j, \forall X \neq C, \forall Y,$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

(C)



$\longrightarrow -1, \quad \forall i < j, \forall (X, Y),$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

(D) Any other motif $\longrightarrow +\infty, \quad \forall i < j,$

Claim: The energy of **any matching** of w_X is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

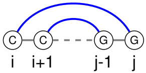
For some nearest-neighbor model $\mathcal{N}$, one has RNA-PK-FOLD($\mathcal{N}$) $\notin$ *APX*.

Proof. Consider the RNA seq. built from some 3-PARTITION instance X:

$$w_X = \mathrm{C}^{x_1}\mathrm{A}\mathrm{C}^{x_2}\mathrm{A}\cdots\mathrm{A}\mathrm{C}^{x_{3m}}\mathrm{A}\,\underbrace{\mathrm{G}^K\mathrm{U}\mathrm{G}^K\mathrm{U}\cdots\mathrm{G}^K\mathrm{U}}_{m\text{ times}}\,\mathrm{U}^{2m}$$

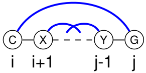
and the energy model:

$\Delta^*_{\mathcal{N}}$: (A)



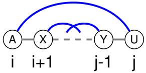
$\longrightarrow -1, \quad \forall i < j,$

(B)



$\longrightarrow -1, \quad \forall i < j, \forall X \neq C, \forall Y,$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

(C)



$\longrightarrow -1, \quad \forall i < j, \forall (X, Y),$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

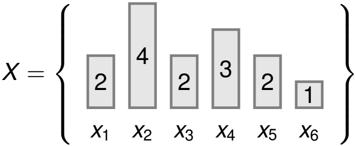
(D) Any other motif $\longrightarrow +\infty, \quad \forall i < j,$

Claim: The energy of **any matching** of w_X is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).

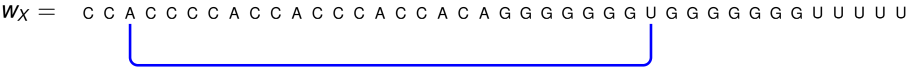
Three choices

Matching S^* is:

- **Empty** $\rightarrow \Delta_{\mathcal{N}}(S^*) = 0$
- **Invalid:** Some base-pair breaks some rule $\rightarrow \Delta_{\mathcal{N}}(S^*) = \infty$
- **Induces a 3-partition matching**



$(K = 7)$

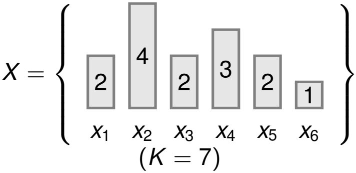


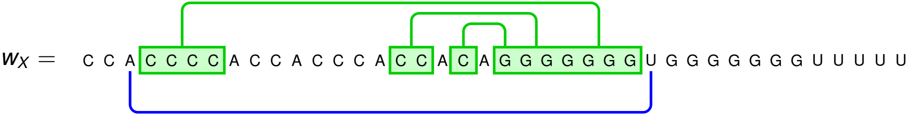
$w_X =$ C C A C C C C A C C A C C C A C C A C A G G G G G G G U G G G G G G G U U U U U

Three choices

Matching S^* is:

- **Empty** $\rightarrow \Delta_{\mathcal{N}}(S^*) = 0$
- **Invalid:** Some base-pair breaks some rule $\rightarrow \Delta_{\mathcal{N}}(S^*) = \infty$
- **Induces a 3-partition matching**

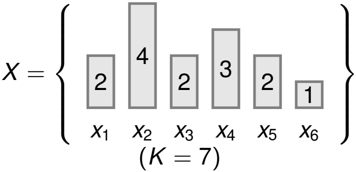



$w_X =$
C C A C C C C A C C A C C C A C C A C A G G G G G G G U G G G G G G G U U U U U

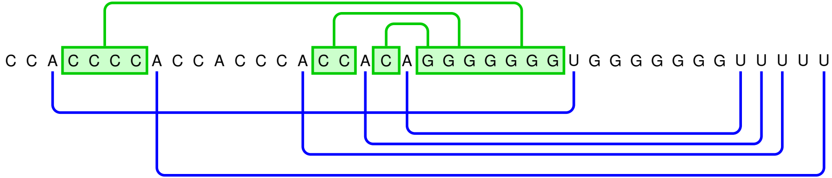
Three choices

Matching S^* is:

- **Empty** $\rightarrow \Delta_{\mathcal{N}}(S^*) = 0$
- **Invalid:** Some base-pair breaks some rule $\rightarrow \Delta_{\mathcal{N}}(S^*) = \infty$
- **Induces a 3-partition matching**



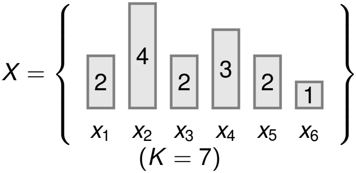
$w_X =$

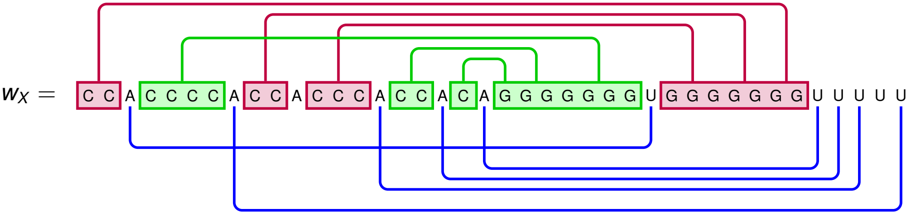

C C A C C C C A C C A C C C A C C A C A G G G G G G G U G G G G G G G U U U U U

Three choices

Matching S^* is:

- **Empty** $\rightarrow \Delta_{\mathcal{N}}(S^*) = 0$
- **Invalid:** Some base-pair breaks some rule $\rightarrow \Delta_{\mathcal{N}}(S^*) = \infty$
- **Induces a 3-partition matching**

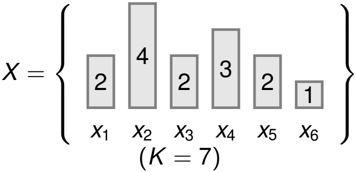


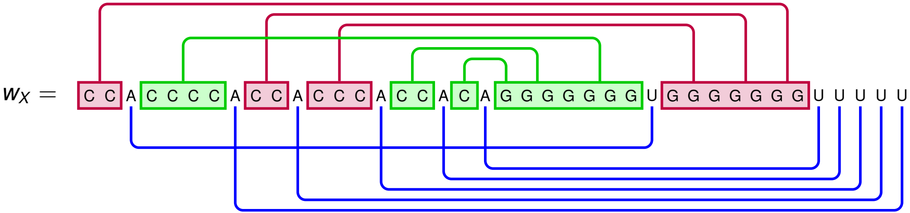

$w_X =$
C C A C C C C A C C A C C C A C C A C A G G G G G G G U G G G G G G G U U U U U

Three choices

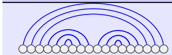
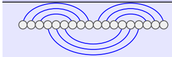
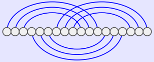
Matching S^* is:

- **Empty** $\rightarrow \Delta_{\mathcal{N}}(S^*) = 0$
- **Invalid:** Some base-pair breaks some rule $\rightarrow \Delta_{\mathcal{N}}(S^*) = \infty$
- **Induces a 3-partition matching**




w_X =
C C A C C C C A C C A C C C A C C A C A G G G G G G G U G G G G G G G U U U U U

Final summary

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04] **(any* Δ model)**	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04] **1/5 (any Δ model)**	**APX-Hard**

Conclusion

Incorporating pseudoknots is generally hard:

- **Dishonest** inapproximability result for nearest-neighbor model
- **Almost honest** general hardness result for stacking model
- **Honest** 5-approximation for stacking model

Still hope for **tractable exact algorithms** accounting for PKs:

- Parametrized approaches (*aka you get what you pay for. . .*)
- Topologically restricted sets of RNAs

Thanks for listening
Questions?

Job offers: PhD & Postdoc on RNA kinetics@Inria/Ecole Polytechnique

Funding

Agence Nationale de la Recherche ANR

Inria INVENTEURS DU MONDE NUMÉRIQUE

cnrs dépasser les frontières

Conclusion

Incorporating pseudoknots is generally hard:

- **Dishonest** inapproximability result for nearest-neighbor model
- **Almost honest** general hardness result for stacking model
- **Honest** 5-approximation for stacking model

Still hope for **tractable exact algorithms** accounting for PKs:

- Parametrized approaches (*aka you get what you pay for. . .*)
- Topologically restricted sets of RNAs

Thanks for listening
Questions?

Job offers: PhD & Postdoc on RNA kinetics@Inria/Ecole Polytechnique

Funding

Agence Nationale de la Recherche ANR

Inria INVENTEURS DU MONDE NUMÉRIQUE

cnrs dépasser les frontières

References I

Tatsuya Akutsu.
Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.
Discrete Appl. Math., vol. 104, no. 1-3, pages 45–62, 2000.

M. R. Garey & D. S. Johnson.
Complexity Results for Multiprocessor Scheduling under Resource Constraints.
SIAM Journal on Computing, vol. 4, no. 4, pages 397–411, 1975.

Samuel Ieong, Ming yang Kao, Tak wah Lam, Wing kin Sung & Siu ming Yiu.
Predicting RNA Secondary Structures with Arbitrary Pseudoknots by Maximizing the Number of Stacking Pairs.
Journal Of Computational Biology, vol. 10, no. 6, pages 981–995, 2003.

N. Leontis & E. Westhof.
Geometric nomenclature and classification of RNA base pairs.
RNA, vol. 7, pages 499–512, 2001.

R. B. Lyngsø& C. N. S. Pedersen.
RNA Pseudoknot Prediction in Energy-Based Models.
Journal of Computational Biology, vol. 7, no. 3-4, pages 409–427, 2000.

Rune Lyngsø.
Complexity of Pseudoknot Prediction in Simple Models.
In Proceedings of ICALP, 2004.

R. Nussinov & A.B. Jacobson.
Fast algorithm for predicting the secondary structure of single-stranded RNA.
Proc Natl Acad Sci U S A, vol. 77, pages 6903–13, 1980.

Jesse Stombaugh, Craig L. Zirbel, Eric Westhof & Neocles B. Leontis.
Frequency and isostericity of RNA base pairs.
Nucleic Acids Research, vol. 37, no. 7, pages 2294–2312, 2009.

References II

J. E. Tabaska, R. B. Cary, H. N. Gabow & G. D. Stormo.
An RNA folding method capable of identifying pseudoknots and base triples.
Bioinformatics, vol. 14, no. 8, pages 691–699, 1998.

M. Zuker & P. Stiegler.
Optimal computer folding of large RNA sequencesusing thermodynamics and auxiliary information.
Nucleic Acids Res., vol. 9, pages 133–148, 1981.