

Impact Of The Energy Model On The Complexity Of RNA Folding With Pseudoknots

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

⊙ University of Florida, Gainesville, USA

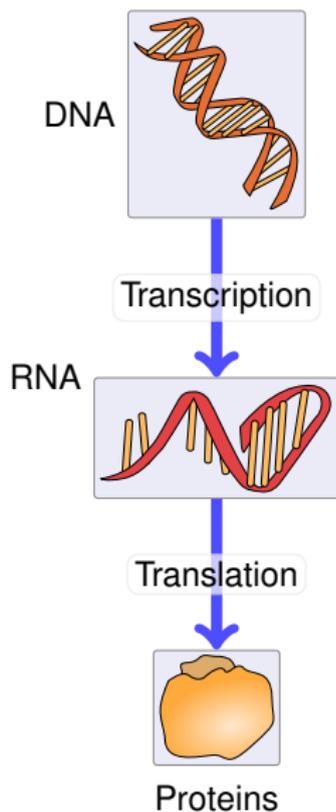
♣ Albert Ludwigs University, Freiburg, Germany

• LIX, CNRS/Ecole Polytechnique, France

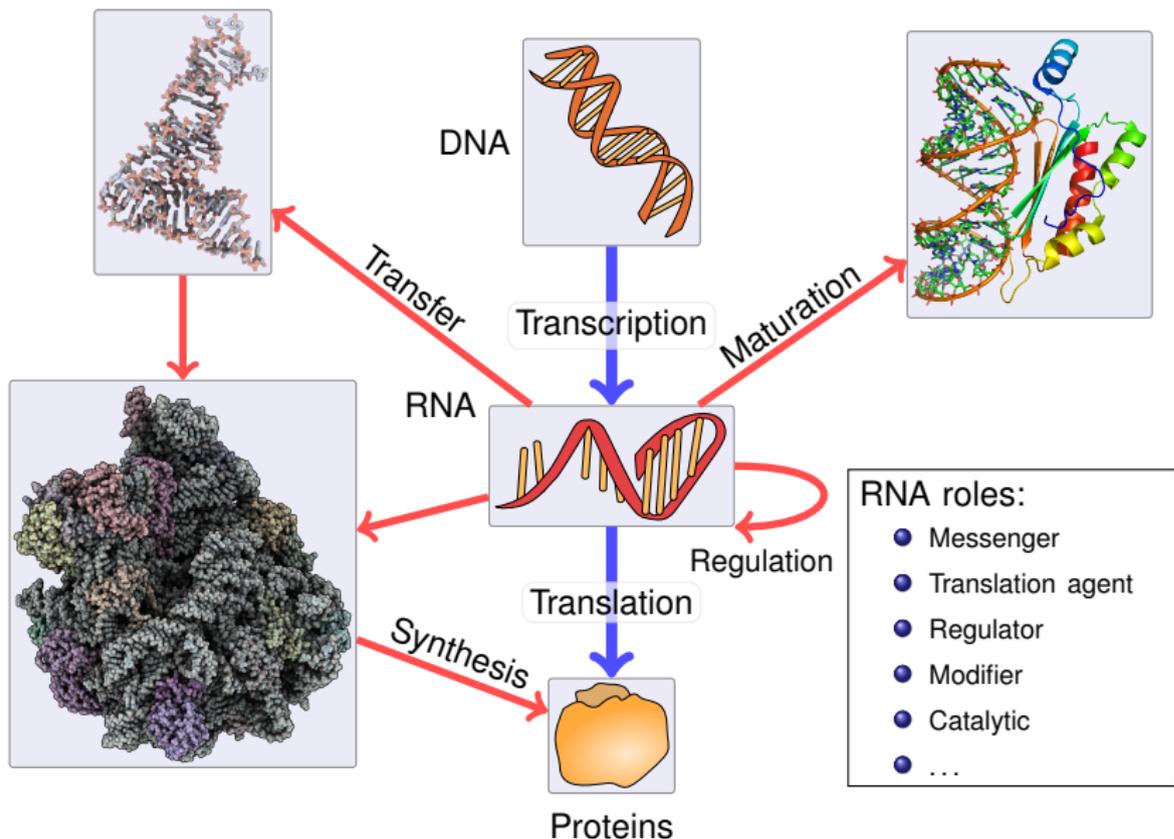
◇ AMIB Team-Project, INRIA, Saclay, France

July 5th – CPM'12

Fundamental dogma of molecular biology



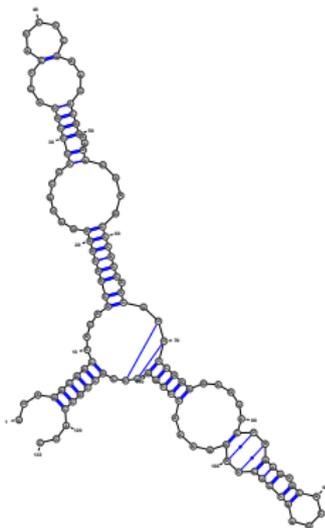
Fundamental dogma of molecular biology



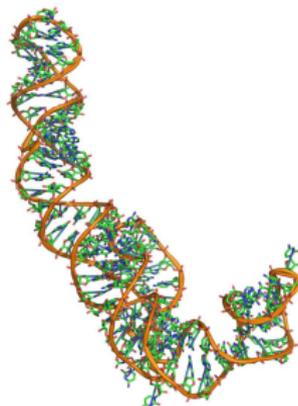
RNA structure

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

Primary structure



Secondary structure
(Matching)



Tertiary structure

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

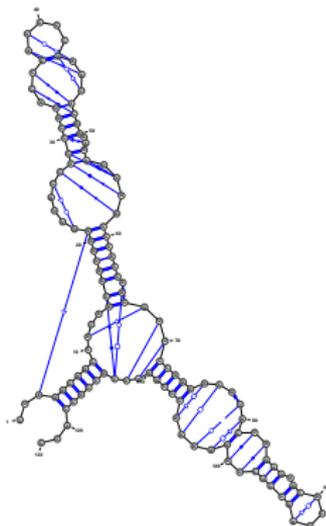
Bottom-up approach to molecular biology

Understand and predict how RNA folds to decipher its function(s).

RNA structure

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

Primary structure



Secondary⁺ structure
(Matching)



Tertiary structure

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

Bottom-up approach to molecular biology

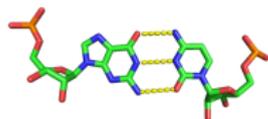
Understand and predict how RNA folds to decypher its function(s).

Crossing interactions

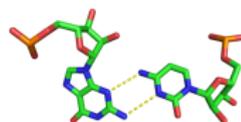
- **Non-canonical base-pairs:**

Any base-pair **other than** {(A-U), (C-G), (G-U)}

OR interacting in a non-standard way (WC/WC-Cis) [Leontis 01].

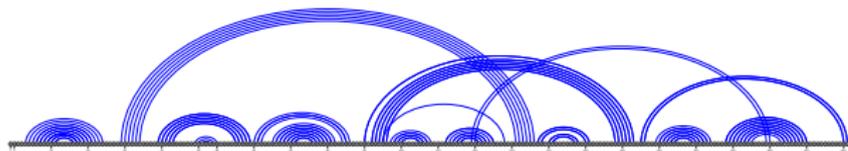


Canonical CG base-pair (WC/WC-Cis)



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

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



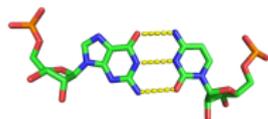
Group I Ribozyme (PDBID: 1Y0Q:A)

Crossing interactions

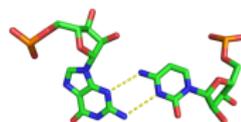
- **Non-canonical base-pairs:**

Any base-pair **other than** {(A-U), (C-G), (G-U)}

OR interacting in a non-standard way (WC/WC-Cis) [Leontis 01].

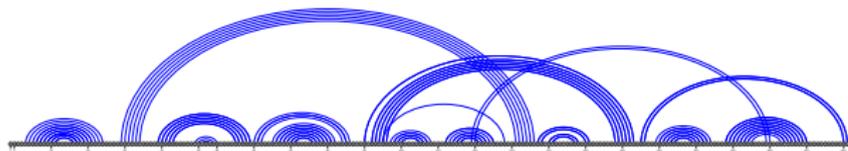


Canonical CG base-pair (WC/WC-Cis)



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

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



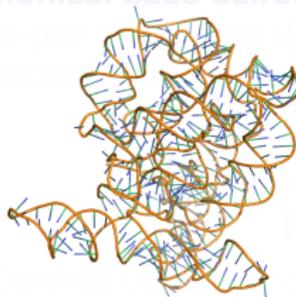
Group I Ribozyme (PDBID: 1Y0Q:A)

Crossing interactions

- **Non-canonical base-pairs:**

Any base

OR inter



(A-U), (C-G), (G-U)

Standard Watson-Crick

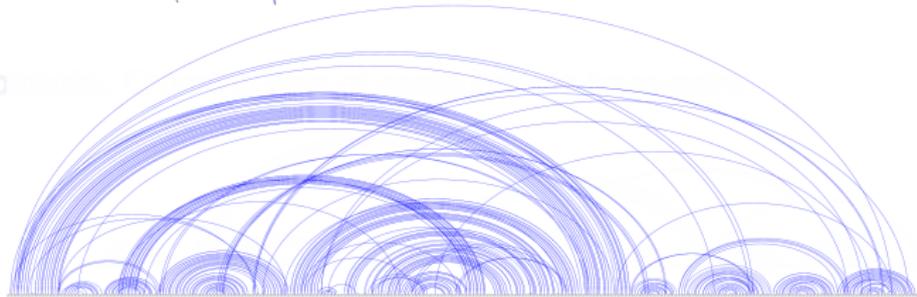
Crossing interactions, once ignored, are now **ubiquitous!**

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

VC/WC-Cis

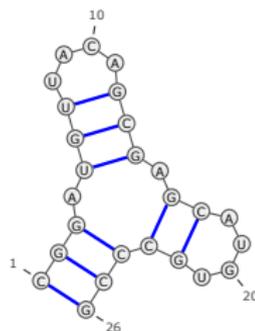
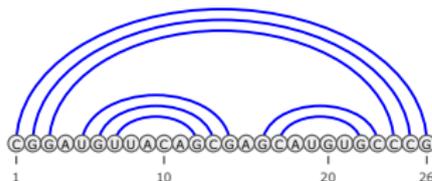
Non-canonical base-pair (Sugar/WC-Trans)

- **Pseudo**



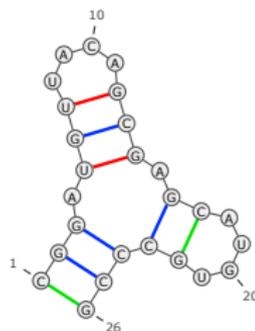
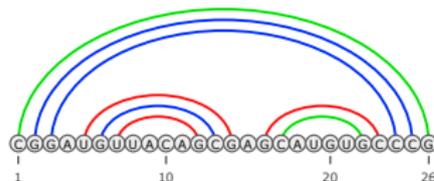
Group I ribozyme (PDB ID: 1RUC)

Problem statement



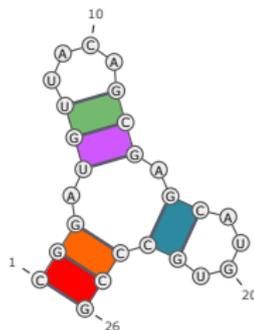
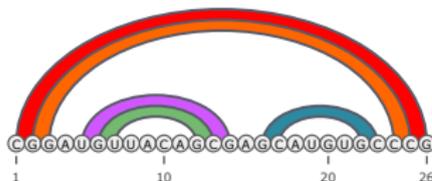
- **RNA structure S :** (Partial) matching of positions in sequence w
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model:**
 - Motif** \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$:**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

Problem statement



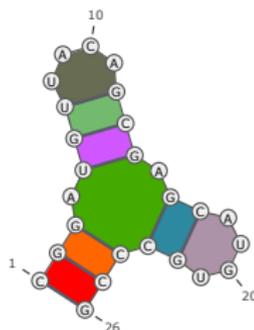
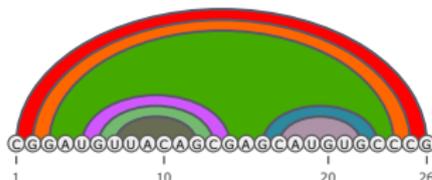
- **RNA structure S :** (Partial) matching of positions in sequence w
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model:**
 - Motif \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$: Sum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

Problem statement



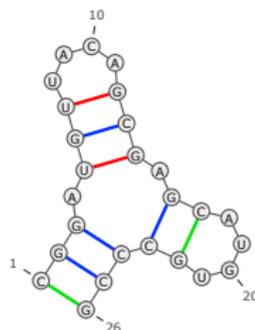
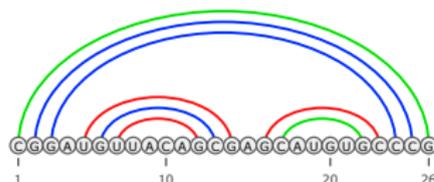
- **RNA structure S :** (Partial) matching of positions in sequence w
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model:**
 - Motif \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$
 - Free-Energy $E_w(S)$: Sum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

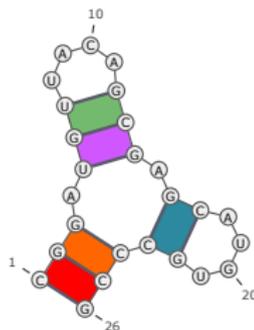
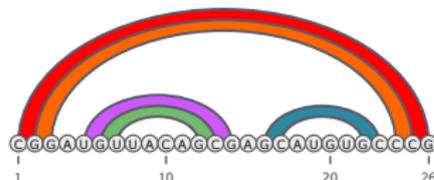
Problem statement



- **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

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

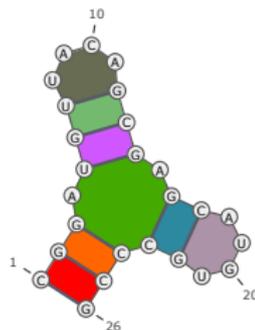
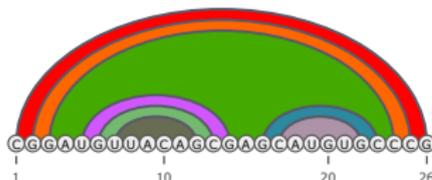
Problem statement



- **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



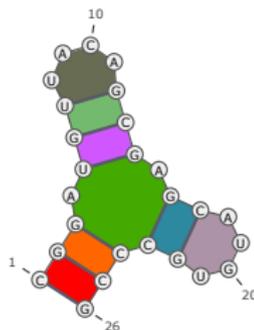
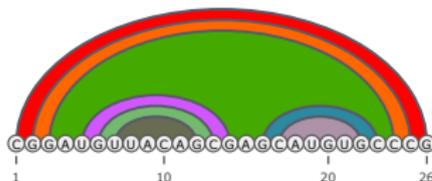
- **RNA structure S :** (Partial) matching of positions in sequence w
- **Motifs:** Sequence/structure features (e.g. Base-pairs, Stacking pairs, Loops...)
- **Energy model:**

Motif \rightarrow Free-energy contribution $\Delta(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

Free-Energy $E_w(S)$: Sum over (independently contributing) motifs in S

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

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 (RNA-PK-FOLD(E) problem)

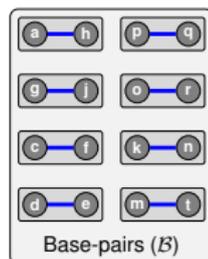
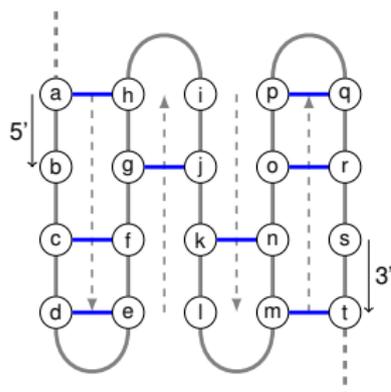
Input: RNA sequence $w \in \{A, C, G, U\}^*$.

Output: Matching S^* , having Minimal Free-Energy $E_w(S^*)$.

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



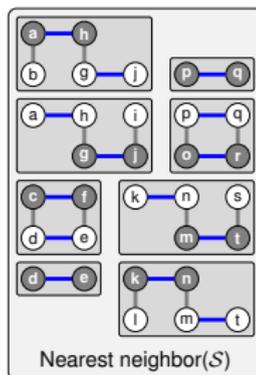
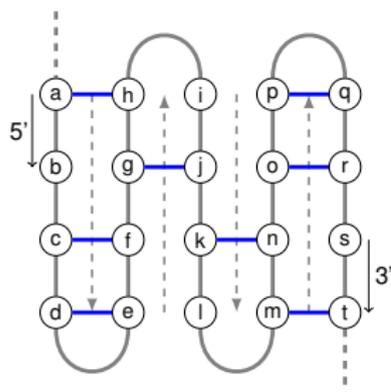
Solved in $\mathcal{O}(n^3)$ [Tabaska 98]
(Max-weighted matching)

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

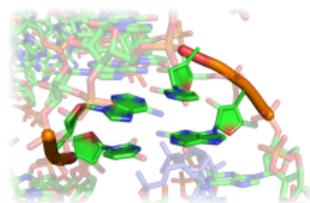
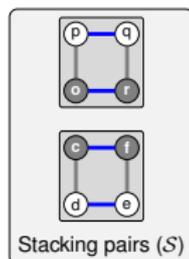
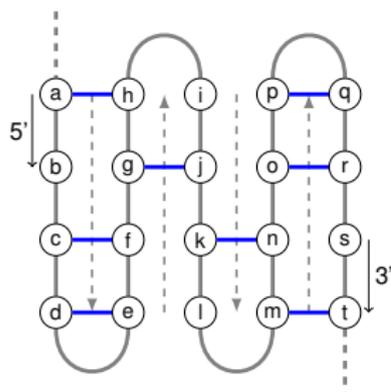


NP-hard [Lyngsø 00, Akutsu 00]
Too expressive?

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

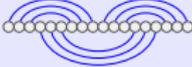
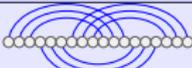


Captures stablest motifs

Still NP-hard [Lyngsø 04]

... but PTAS [Lyngsø 04]

State of the art

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

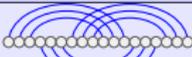
→ Dishonest

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 [leong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [leong 03]	NP-Hard [leong 03]
Planar	Approx.	2-approx. \approx [leong 03]	2-approx. [leong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ϵ -approx. $\in \mathcal{O}(n^{4^{1/\epsilon}})$ [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 ($\exists \mathcal{M}$)?

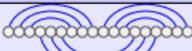
→ Dishonest

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 [leong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [leong 03]	NP-Hard [leong 03]
Planar	Approx.	2-approx. \approx [leong 03]	2-approx. [leong 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 ($\exists \mathcal{M}$)?

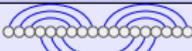
→ **Dishonest**

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 [leong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [leong 03]	NP-Hard [leong 03]
Planar	Approx.	2-approx. \approx [leong 03]	2-approx. [leong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ϵ -approx. $\in \mathcal{O}(n^{4^{1/\epsilon}})$ [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 ($\exists \mathcal{M}$)?

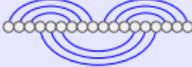
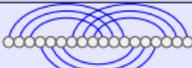
→ **Dishonest**

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 [leong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [leong 03]	NP-Hard [leong 03]
Planar	Approx.	2-approx. \approx [leong 03]	2-approx. [leong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04]	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ϵ -approx. $\in \mathcal{O}(n^{4^{1/\epsilon}})$ [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**

(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] \rightarrow 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.

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 $w_X := C^{x_1} A C^{x_2} A C^{x_3} A \dots A C^{x_n} \underbrace{A G^K A G^K A \dots A G^K}_{m \text{ times}}$ and $\delta := \Delta_S(C, G, C, G)$
- 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 (C^a, C^b, C^c).
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then $\text{RNA-PK-FOLD}(S) \in P \Rightarrow 3\text{-PARTITION} \in P$.

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

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 $w_X := C^{x_1} A C^{x_2} A C^{x_3} A \dots A C^{x_n} \underbrace{A G^K A G^K A \dots A G^K}_{m \text{ times}}$ and $\delta := \Delta_S(C, G, C, G)$
- 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 (C^a, C^b, C^c) .
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then $\text{RNA-PK-FOLD}(S) \in P \Rightarrow 3\text{-PARTITION} \in P$.

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

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 $w_X := C^{x_1} A C^{x_2} A C^{x_3} A \dots A C^{x_n} \underbrace{A G^K A G^K A \dots A G^K}_{m \text{ times}}$ and $\delta := \Delta_S(C, G, C, G)$
- 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 (C^a, C^b, C^c).
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then $\text{RNA-PK-FOLD}(S) \in P \Rightarrow 3\text{-PARTITION} \in P$.

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

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 $w_X := C^{x_1} A C^{x_2} A C^{x_3} A \dots A C^{x_n} \underbrace{A G^K A G^K A \dots A G^K}_{m \text{ times}}$ and $\delta := \Delta_S(C, G, C, G)$
- 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 (C^a, C^b, C^c) .
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then $\text{RNA-PK-FOLD}(S) \in P \Rightarrow 3\text{-PARTITION} \in P$.

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

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 $w_X := C^{x_1} A C^{x_2} A C^{x_3} A \dots A C^{x_n} \underbrace{A G^K A G^K A \dots A G^K}_{m \text{ times}}$ and $\delta := \Delta_S(C, G, C, G)$
- 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 (C^a, C^b, C^c) .
- Since $|w_X| \in \mathcal{O}(n \cdot P(n))$, then $\text{RNA-PK-FOLD}(S) \in \mathcal{P} \Rightarrow 3\text{-PARTITION} \in \mathcal{P}$.

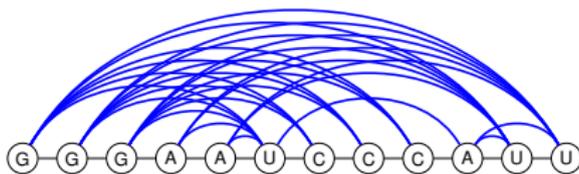
Reminder: 3-PARTITION is **strongly** NP-Hard [Garey 75], i.e. still hard if $x_i < P(n)$.

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

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\epsilon}})$) [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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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(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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

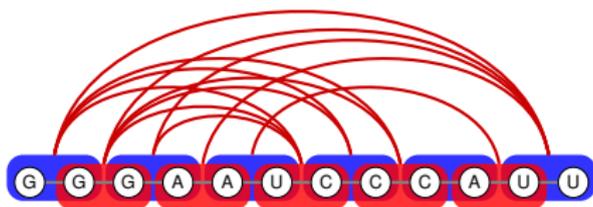


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- 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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

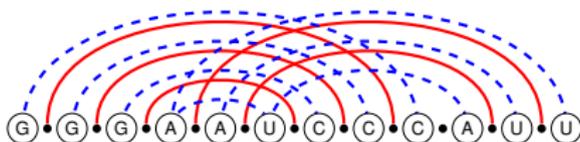


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\epsilon}})$) [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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

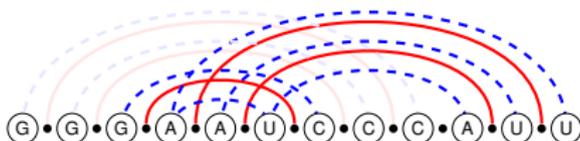


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- Existence of polynomial time approximation scheme (in $\mathcal{O}(n^{4^{1/\epsilon}})$) [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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

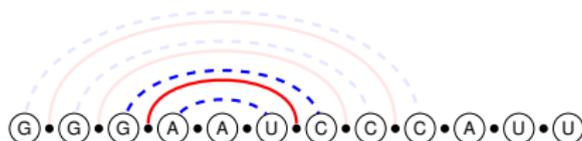


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- 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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

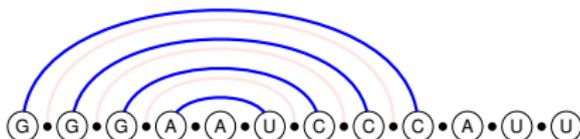


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- 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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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

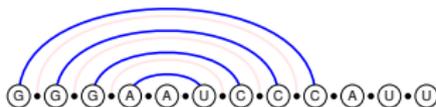


Honest $\mathcal{O}(n^3)$ 5-approximation for RNA-PK-FOLD(S)

- 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: $(i, i + 1) \rightarrow (j - 1, j)$ with weight $-\Delta_S(w_i, w_j, w_{i+1}, w_{j-1})$
- 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



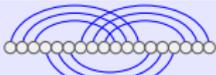
Complexity: At most $\mathcal{O}(n^3)$ (Max-weighted matching)

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 ($\forall S \rightarrow$ Honest)

Half-time summary

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

How hard is it to approximate the nearest neighbor model?

(Dishonest!) Inapproximability of Nearest-Neighbor model

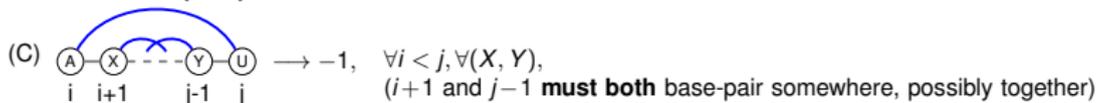
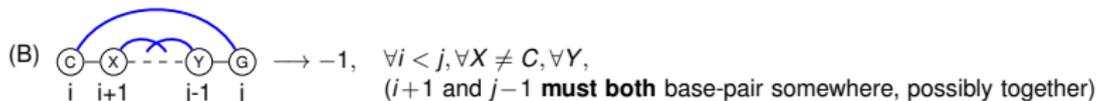
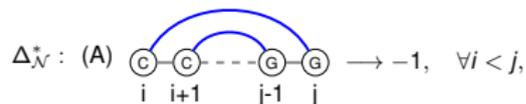
Theorem

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

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

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

and the energy model:



Lemma: 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

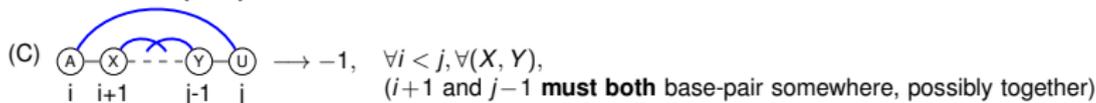
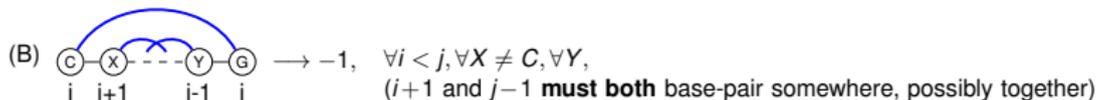
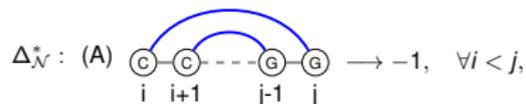
Theorem

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

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

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

and the energy model:



Lemma: 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 $\text{RNA-PK-FOLD}(\mathcal{N}) \notin \text{APX}$.

Proof (continued).

- The energy of **any matching** of w_X under \mathcal{N} is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).
 - **Reminder:** Polynomial-time $1/f(n)$ -approximation algorithm bound to produce solution having free-energy $\leq f(n) \cdot \text{OPT}$.
 - Any $1/f(n)$ -approx. algorithm, $f(n) > 0$, produces a matching of **negative free-energy** $\leq f(n) \cdot E^* < 0$ iff a matching of energy $E^* < 0$ exists...
 - ... i.e. iff X is 3-partitionable!
- \Rightarrow Unless $P = NP$, there is no **polynomial-time approximation algorithm of (non-necessarily constant) positive ratio** for $\text{RNA-PK-FOLD}(\mathcal{N})$.

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

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

Proof (continued).

- The energy of **any matching** of w_X under \mathcal{N} is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).
 - **Reminder:** Polynomial-time $1/f(n)$ -approximation algorithm bound to produce solution having free-energy $\leq f(n) \cdot \text{OPT}$.
 - Any $1/f(n)$ -approx. algorithm, $f(n) > 0$, produces a matching of **negative free-energy** $\leq f(n) \cdot E^* < 0$ iff a matching of energy $E^* < 0$ exists...
 - ... i.e. iff X is 3-partitionable!
- \Rightarrow Unless $P = NP$, there is no **polynomial-time approximation algorithm of (non-necessarily constant) positive ratio** for $\text{RNA-PK-FOLD}(\mathcal{N})$.

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

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

Proof (continued).

- The energy of **any matching** of w_X under \mathcal{N} is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).
 - **Reminder:** Polynomial-time $1/f(n)$ -approximation algorithm bound to produce solution having free-energy $\leq f(n) \cdot \text{OPT}$.
 - Any $1/f(n)$ -approx. algorithm, $f(n) > 0$, produces a matching of **negative free-energy** $\leq f(n) \cdot E^* < 0$ **iff a matching of energy** $E^* < 0$ **exists...**
 - ... i.e. iff X is 3-partitionable!
- \Rightarrow Unless $P = NP$, there is no **polynomial-time approximation algorithm of (non-necessarily constant) positive ratio** for $\text{RNA-PK-FOLD}(\mathcal{N})$.

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

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

Proof (continued).

- The energy of **any matching** of w_X under \mathcal{N} is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).
- **Reminder:** Polynomial-time $1/f(n)$ -approximation algorithm bound to produce solution having free-energy $\leq f(n) \cdot \text{OPT}$.
- Any $1/f(n)$ -approx. algorithm, $f(n) > 0$, produces a matching of **negative free-energy** $\leq f(n) \cdot E^* < 0$ **iff a matching of energy $E^* < 0$ exists...**
- ... i.e. iff X is 3-partitionable!

\Rightarrow Unless $P = NP$, there is no **polynomial-time approximation algorithm of (non-necessarily constant) positive ratio** for $\text{RNA-PK-FOLD}(\mathcal{N})$.

(Dishonest!) Inapproximability of Nearest-Neighbor model

Theorem

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

Proof (continued).

- The energy of **any matching** of w_X under \mathcal{N} is either 0 (no base-pair), $-|w_X| < 0$ ($\Rightarrow X$ is 3-partitionable) or $+\infty$ (any other case).
 - **Reminder:** Polynomial-time $1/f(n)$ -approximation algorithm bound to produce solution having free-energy $\leq f(n) \cdot \text{OPT}$.
 - Any $1/f(n)$ -approx. algorithm, $f(n) > 0$, produces a matching of **negative free-energy** $\leq f(n) \cdot E^* < 0$ **iff a matching of energy** $E^* < 0$ **exists...**
 - ... i.e. iff X is 3-partitionable!
- \Rightarrow Unless $\text{P} = \text{NP}$, there is no **polynomial-time approximation algorithm of (non-necessarily constant) positive ratio** for $\text{RNA-PK-FOLD}(\mathcal{N})$.

Conclusion

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

Nearest Neighbor model:

- **Dishonest** unapproximability → Hardness of approximating within ratio $f(r)$?
where r is largest ratio between contributions of motifs.

Stacking model:

- **Honest** + **efficient** polynomial-time approximation scheme
- **Approximations** do not guarantee **any** overlap with best solution.
→ Polynomial k -overlap algorithm? (Seems unlikely...)

Thanks for listening
Questions?

Thanks to



Conclusion

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

Nearest Neighbor model:

- **Dishonest** unapproximability → Hardness of approximating within ratio $f(r)$ where r is largest ratio between contributions of motifs.

Stacking model:

- **Honest** + **efficient** polynomial-time approximation scheme
- **Approximations** do not guarantee **any** overlap with best solution.
→ Polynomial k -overlap algorithm? (Seems unlikely...)

Thanks for listening
Questions?

Thanks to



Conclusion

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

Nearest Neighbor model:

- **Dishonest** unapproximability \rightarrow Hardness of approximating within ratio $f(r)$ where r is largest ratio between contributions of motifs.

Stacking model:

- **Honest** + **efficient** polynomial-time approximation scheme
- **Approximations** do not guarantee **any** overlap with best solution.
 \rightarrow Polynomial k -overlap algorithm? (Seems unlikely...)

Thanks for listening
Questions?

Thanks to



Conclusion

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

Nearest Neighbor model:

- **Dishonest** unapproximability → Hardness of approximating within ratio $f(r)$ where r is largest ratio between contributions of motifs.

Stacking model:

- **Honest** + **efficient** polynomial-time approximation scheme
- **Approximations** do not guarantee **any** overlap with best solution.
→ Polynomial k -overlap algorithm? (Seems unlikely. . .)

Thanks for listening
Questions?



Thanks to



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 leong, 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.



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 sequences using thermodynamics and auxiliary information.
Nucleic Acids Res., vol. 9, pages 133–148, 1981.