Searching for exact solutions for the inverse folding problem using graphs and parameterization.

Théo Boury[1], Laurent Bulteau[2], Yann Ponty[1]

1, Laboratoire d'Informatique de l'Ecole Polytechnique (CNRS/LIX; UMR 7161), Institut Polytechnique de Paris, France
2, Laboratoire d'Informatique Gaspard Monge (CNRS/LIGM; UMR 8049), Université Gustave Eiffel, France

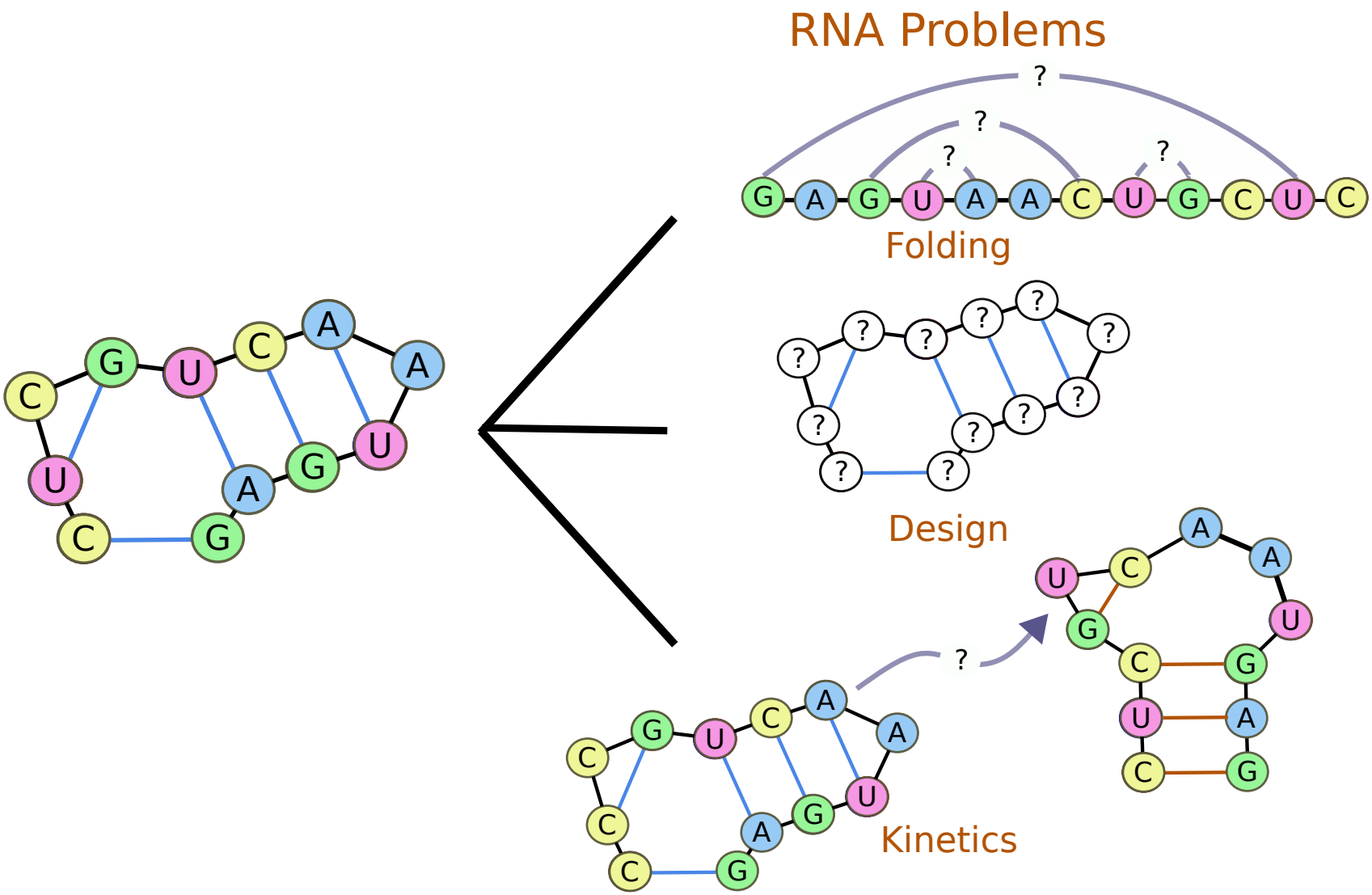
A vague definition of design

Rational design targets a desired biological function

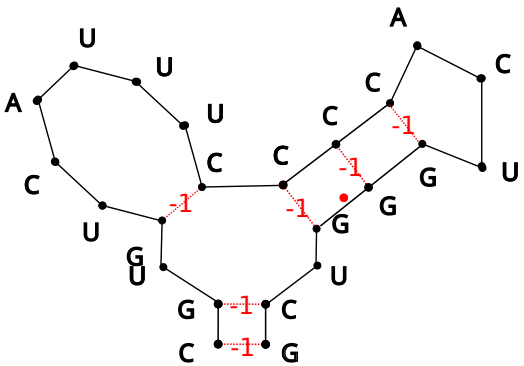
Criteria typically split:

- Positive design ($\approx$ Affinity)
- Negative design ($\approx$ Specificity)

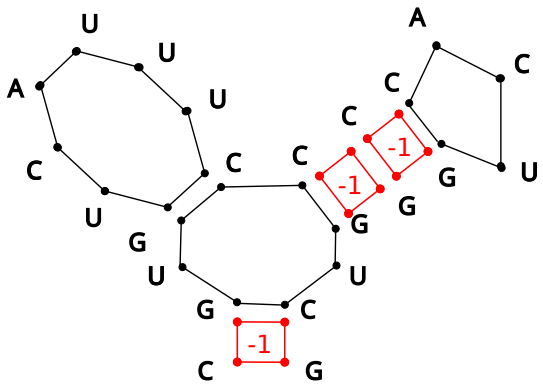
The RNA molecule: 2D abstraction and problems



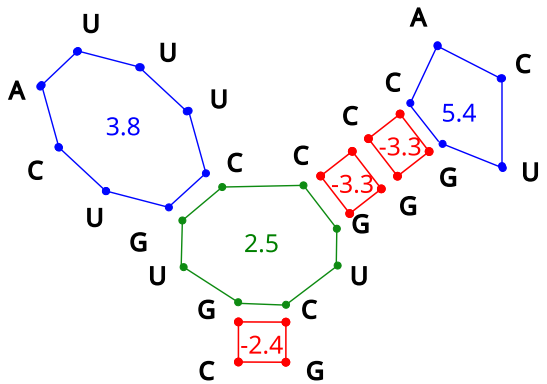
Difficulties of this problem depend heavily on the underlying energy model



Base pairs energy model



Stackings energy model



Turner energy model

Accuracy

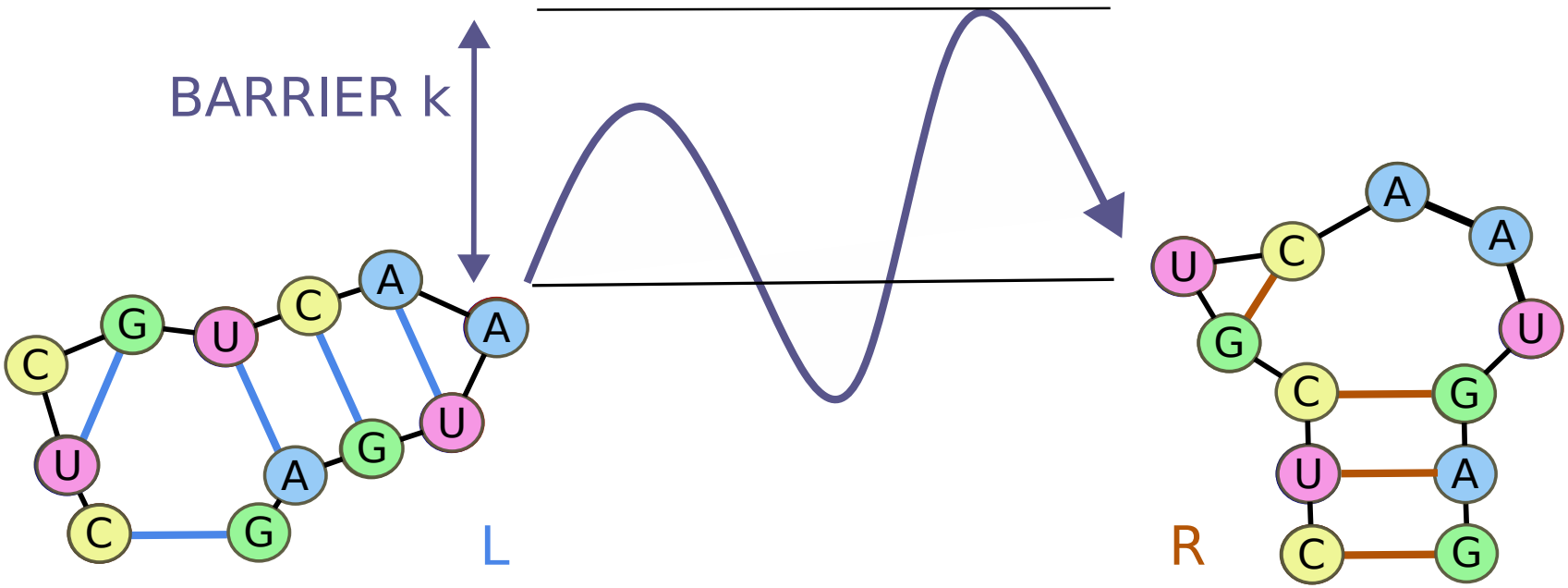
Simplicity

[1]

- Stacking model: a reasonable compromise and starting point

[1] Adapted from Ronny Lorenz's PhD

Design along a kinetics reconfiguration pathway



- **Compatible** over the pathway

The energy barrier problem

Problem 1 (RNA Energy-Barrier):

Input: Sequence ω; Secondary structures L and R; Energy barrier $k \in \mathbb{N}^+$

Output: True if there exists a sequence $S_0 \cdots S_\ell$ of secondary structures such that

- $S_0 = L$ and $S_\ell = R$;
- $E_{\mathcal{M}}(\omega, S_i) - E_{\mathcal{M}}(\omega, L) \leq k, \forall i \in [0, \ell]$;
- $|S_i \triangle S_{i+1}| = 1, \forall i \in [0, \ell - 1]$.

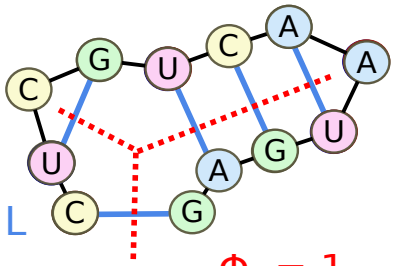
False otherwise.

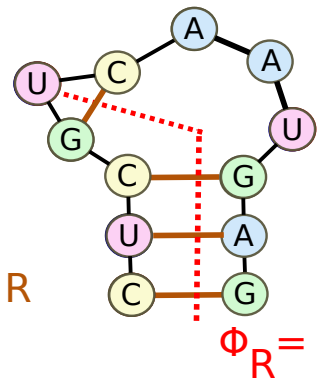
- $\cdots$ An NP-hard problem[2] (even in base pairs model)!
- Heuristically solved: solely an upper bound[3]

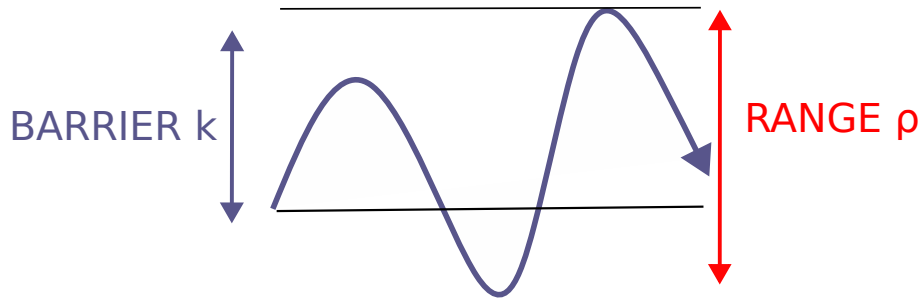
[2] Manuch et al, Nature Computing, 2009

[3] Dotu et al, NAR, 2010

Solving exactly the energy barrier problem







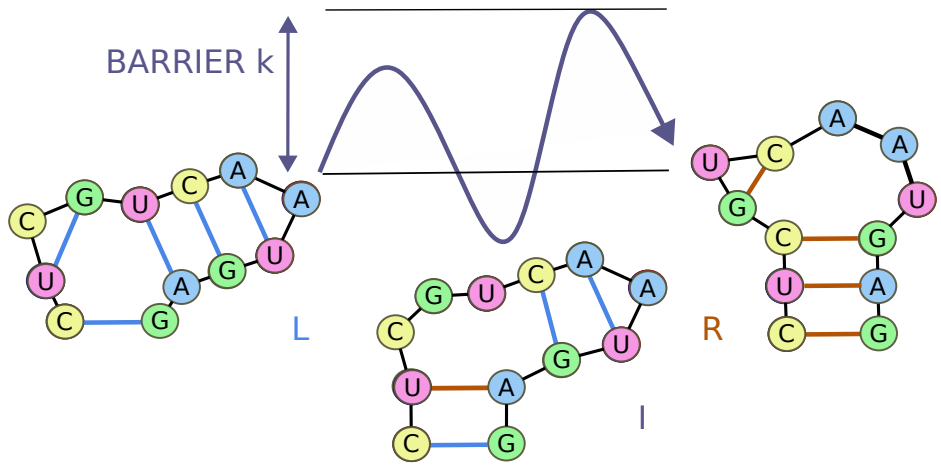
We proposed in the base pairs model: [Boury, Bulteau, Marchand, Ponty, 2023]

- XP in Range ($O(n^{2\rho}\sqrt{n}m)$-time ($m = |E|$), $O(n^2)$-space)
- XP in Arboricity $\Phi = \min(\Phi_B, \Phi_R)$ ($O(n^{\phi+1})$-time, $O(n^{\phi})$-space)

Open question 1:

- Extension to a stacking model?

Design along a "direct" pathway



Definition (Direct pathway): A pathway $S_0 \cdots S_\ell$ is dais to be direct iff it contains only base pairs from L and R.

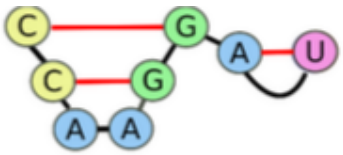
- Barrier does not depend on sequence
- Positive design always possible for two structures [Flamm et al, GCB, 2003]

$\Rightarrow$ Random generation of RNAs achieving barrier less than k from L to R (if possible) can be performed in linear time

Open question 2:

- Indirect pathways?

Inverse folding (positive/negative structural design)

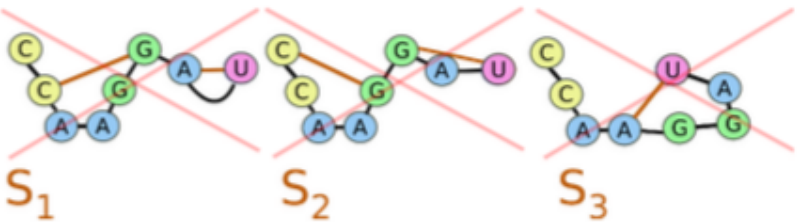


Positive Design (in P)

- Optimally stable structure
- Base pairs chosen from {{A,U},{C,G},{G,U}}

Negative Design: NP-hard [Bonnet et al, JCB 2020]

$\forall S_i \neq S^*, E(S_i) > E(S^*)$



Formal definition

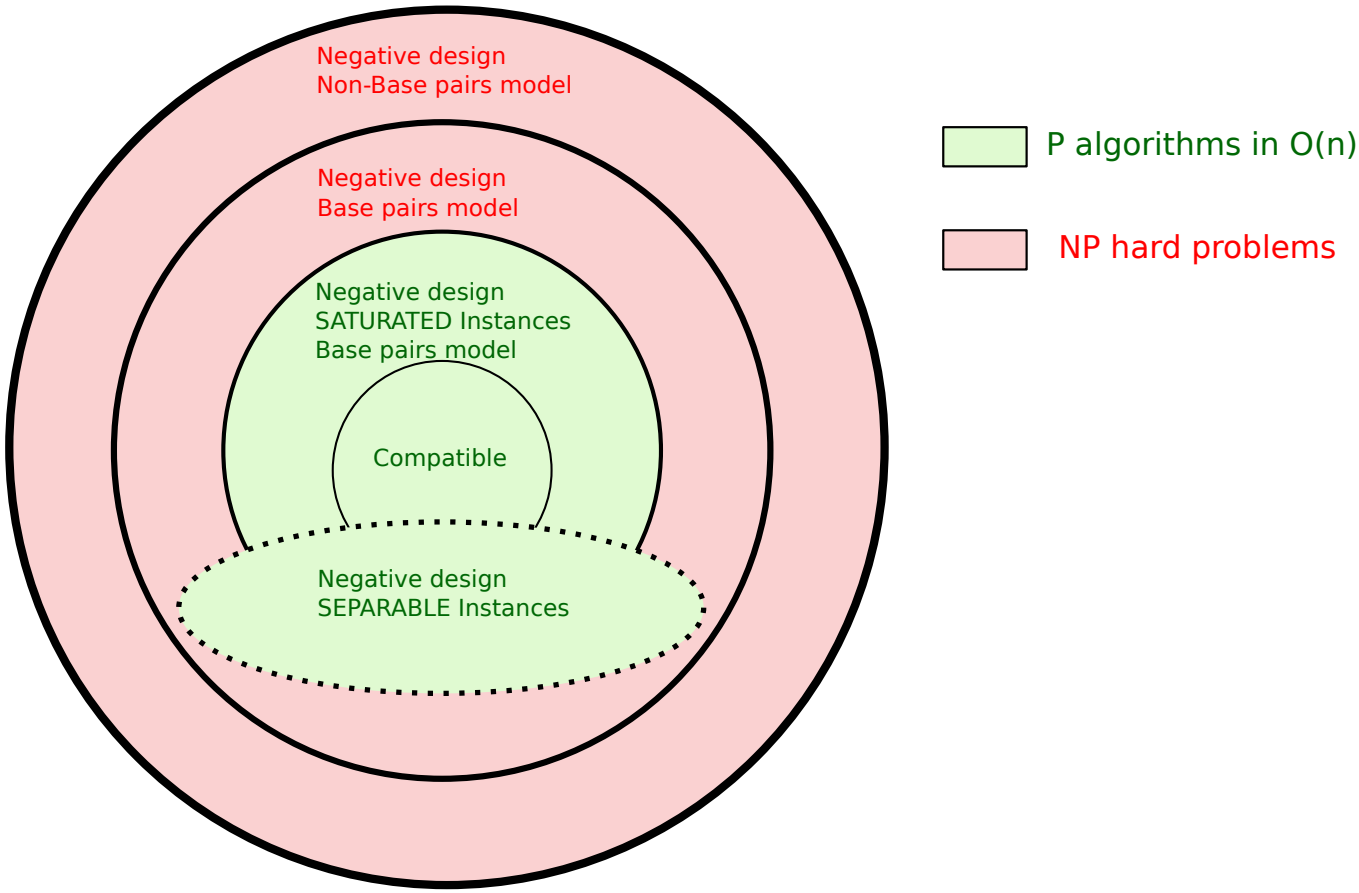
Definition (Design predicates assuming energy model $\mathcal{M}$): Given a target RNA secondary structure $S^\star$ and a length n, a sequence $\omega \in \{A, C, G, U\}^n$ can be called a design iff it respects some of the following predicates:

1. **Compatible**
   $\{\omega.(i), \omega.(j)\} \in \{\{A, U\}, \{G, U\}, \{G, C\}\} \forall (i, j) \in S$
2. **Positive Design**
   $E_{\mathcal{M}}(\omega', S^\star) \geqslant E_{\mathcal{M}}(\omega, S^\star), \forall \omega' \neq \omega$
3. **Negative Design**
   $E_{\mathcal{M}}(\omega, S) > E_{\mathcal{M}}(\omega, S^\star), \forall S \neq S^\star$

- In unweighted models: compatible $\rightarrow$ positive design

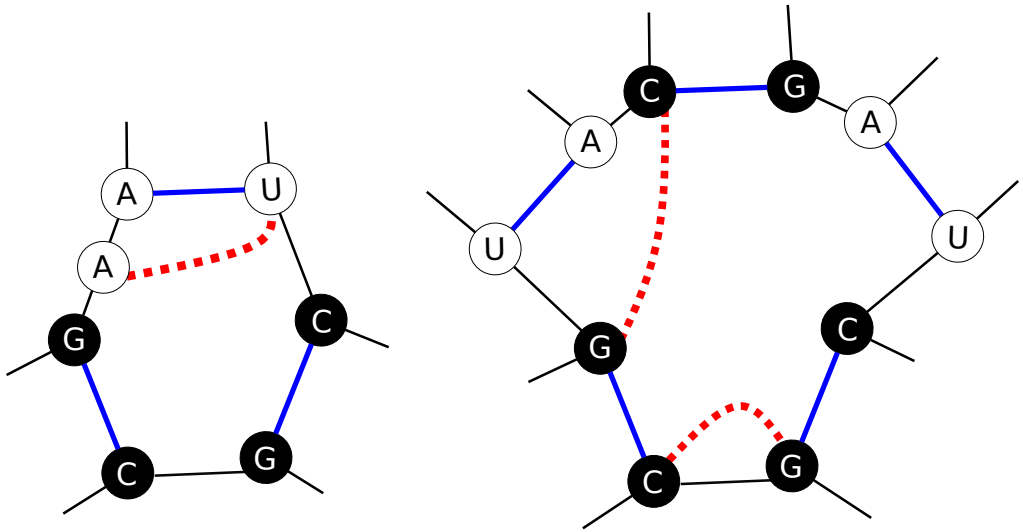
Halès et al[4]: what it brought to the table

How to find exact solutions that satisfies the negative design?



[4]Halès et al, Algorithmica, 2017

Obvious limits



m3o motif m5 motif

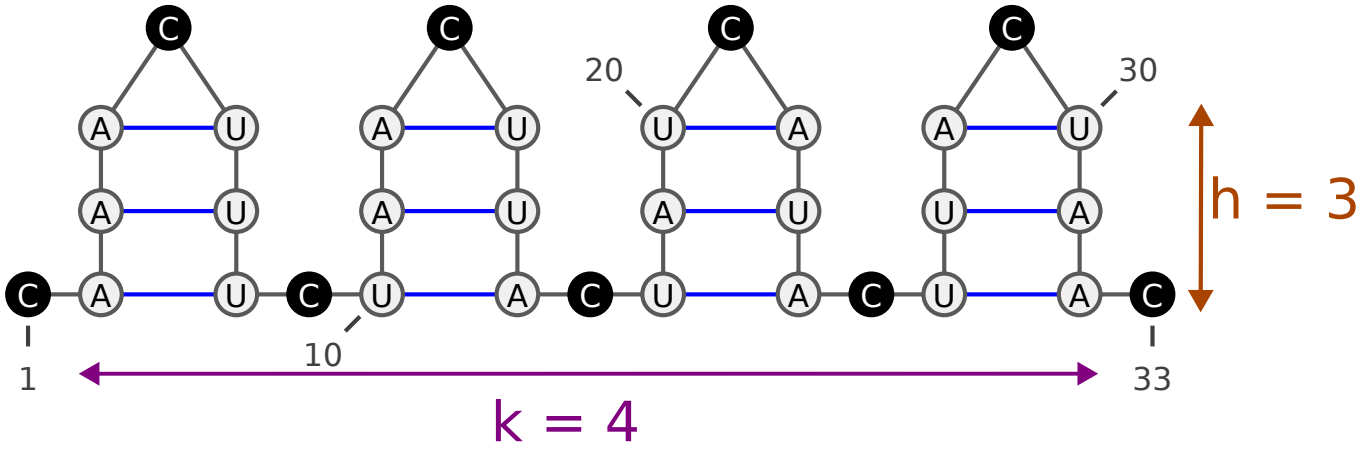
Limits

- m3o and m5 motifs do not yield a negative design
- $\cdots$ A direct consequence of the base pairs model!

Open question 3: Can we remove this restriction on the number of helices using a stacking energy model?

Possibility of the design with more helices

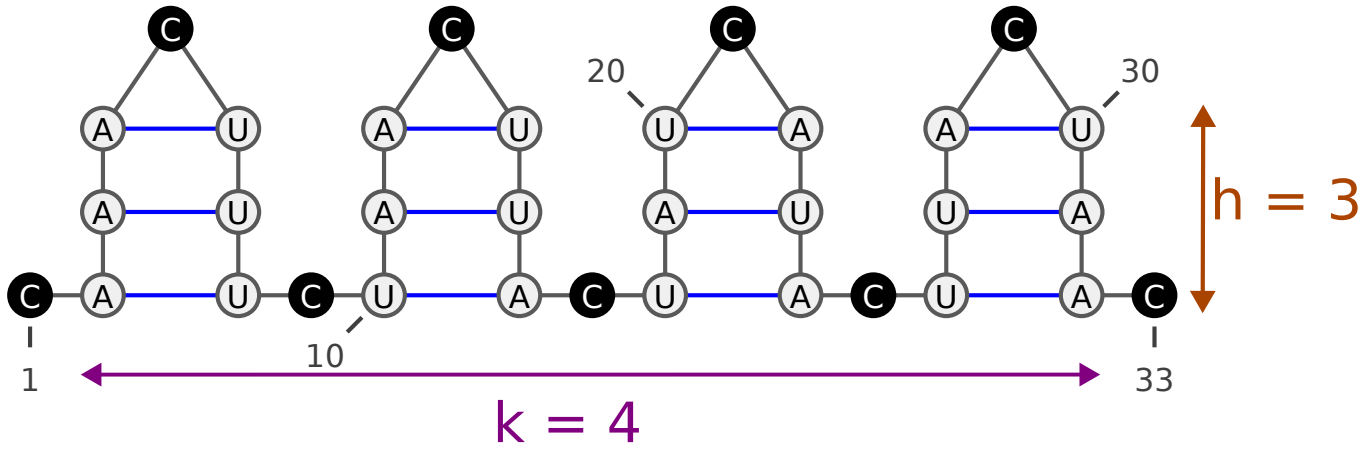
We work now using the stacking model:



Open question 3: Can we remove this restriction on the number of helices using a stacking energy model?

Possibility of the design with more helices

We work now using the stacking model:

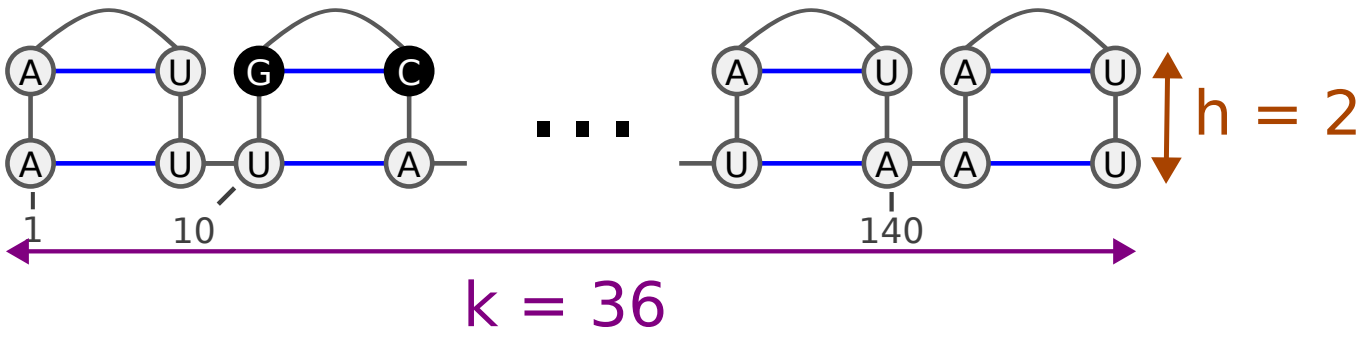


Open question 3: ~~Can we remove this restriction on the number of helices using a stacking energy model?~~ **Yes!**

Theorem (Helices of large enough size $\rightarrow$ Designable]): Given a saturated multiloop S of k helices of size h, with unpaired positions between and at extremities of each helix, if $log_2(k) < h$ then the structure is designable.

Open question 5: Motif generalization: removing terminal and in-between nodes?

· · · but we just push back the base pairs bounds

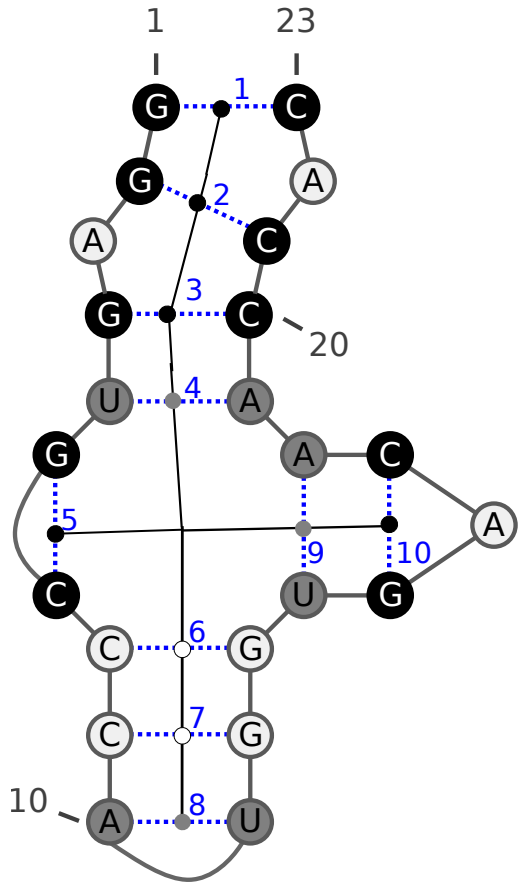


Theorem (Helices of small enough size → Non designable): Given a saturated multiloop S of k helices of size h, if $h < log_6(k)$ then the structure is non-designable.

The existence of such a bound means that there is more than a polynomial number of designs that we miss.[5]

[5] Consequence of Hua Ting's PhD

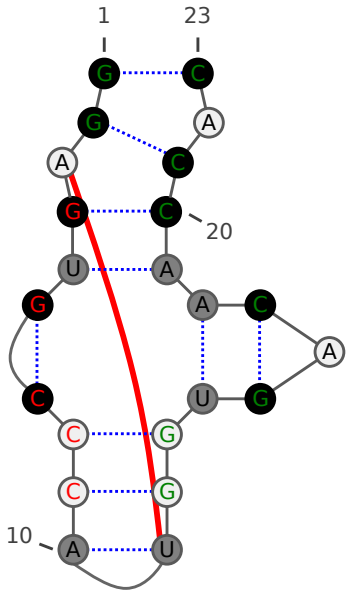
Halès et al approach[6]



- Base pairs are assigned greedily in a DFS manner
- Obtained sequence is a design but not necessarily negative!

[6] Halès et al, Algorithmica, 2017

The separability condition



Let ω be a sequence compatible with S with A on unpaired regions

Definition (Separability condition): ω is separated iff any alternative BP $(i, j) \notin S$ segregates different numbers of C and G.

Open question 6:

- Separability over a stacking energy model?

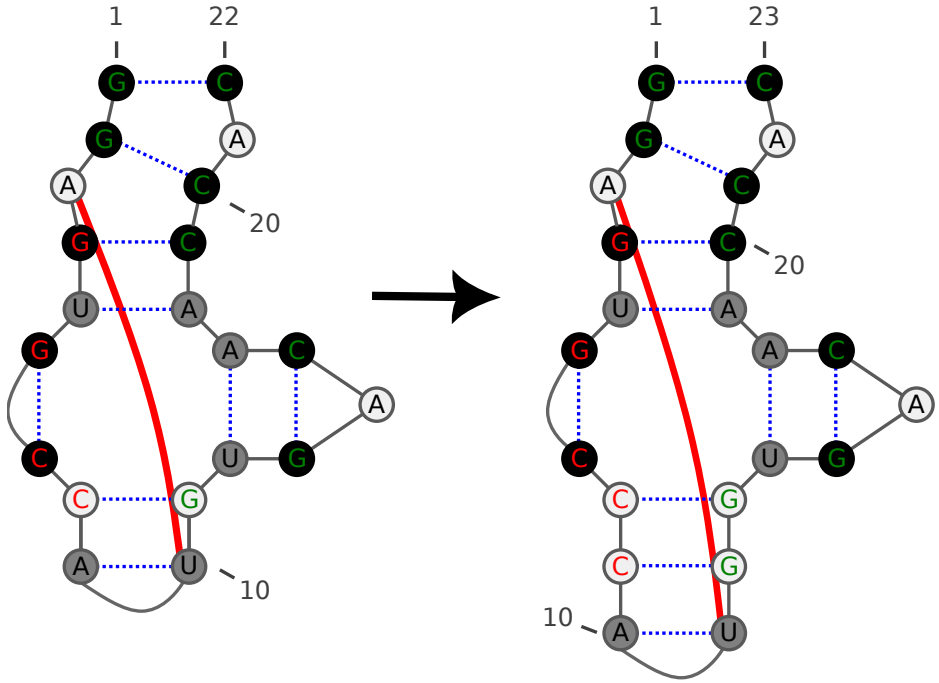
What about instances that are not separable?

One can solve the problem on a modified separable instance:

Definition (Disrupted design):
The sequence ω is a disrupted design at distance p iff we can add p nucleotides to ω and $p/2$ base pairs over them forming S' such as ω' is a design over S'.
We denote the smaller possible perturbation p_{min}.

- Linear algorithm achieving a disrupted design at distance $p \leqslant n$ (Halès et al)
- At most 1 added BP by helix

Disrupted design proposition



Another algorithm for disrupted negative design:

- XP algorithm finding a disrupted design at distance p_{min}
- Through step-by-step exploration of the possible disruptions

Open question 7:

- Find p_{min} value in polynomial complexity?

Sampling

Definition (Sampling): Given a set of design predicates P,we say that ω is a uniform design sample if ω satisfies P and $\mathbb{P}(\omega) = \frac{1}{|\{\omega' | \omega' \text{ satisfies } P\}|}$

Perspectives for sampling:

- Algorithms to sample given multiple structures is polynomial.
- Algorithms to sample given a sequence and some pairs constraints is FPT in treewidth. [7] [8]

Open question 8: What about negative design, in particular, how to enumerate exhaustively the alternatives?

[7] Hammer et al, BMC Bioinformatics, 2019

[8] Yao et al, RNA Folding - Methods and Protocols, 2022

Final word $\cdots$

As one may notice... I have a lot of "open" questions $\cdots$
$\ldots$ but I am also "open" to discussions!