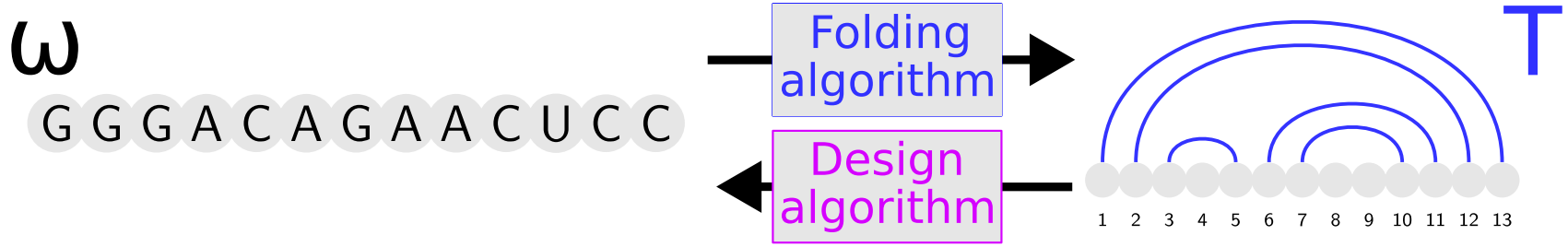
(NP-hard?) RNA Design in linear time and space! (most of the time)

$\rightarrow$ e.g. for structures without isolated stacks or base pairs

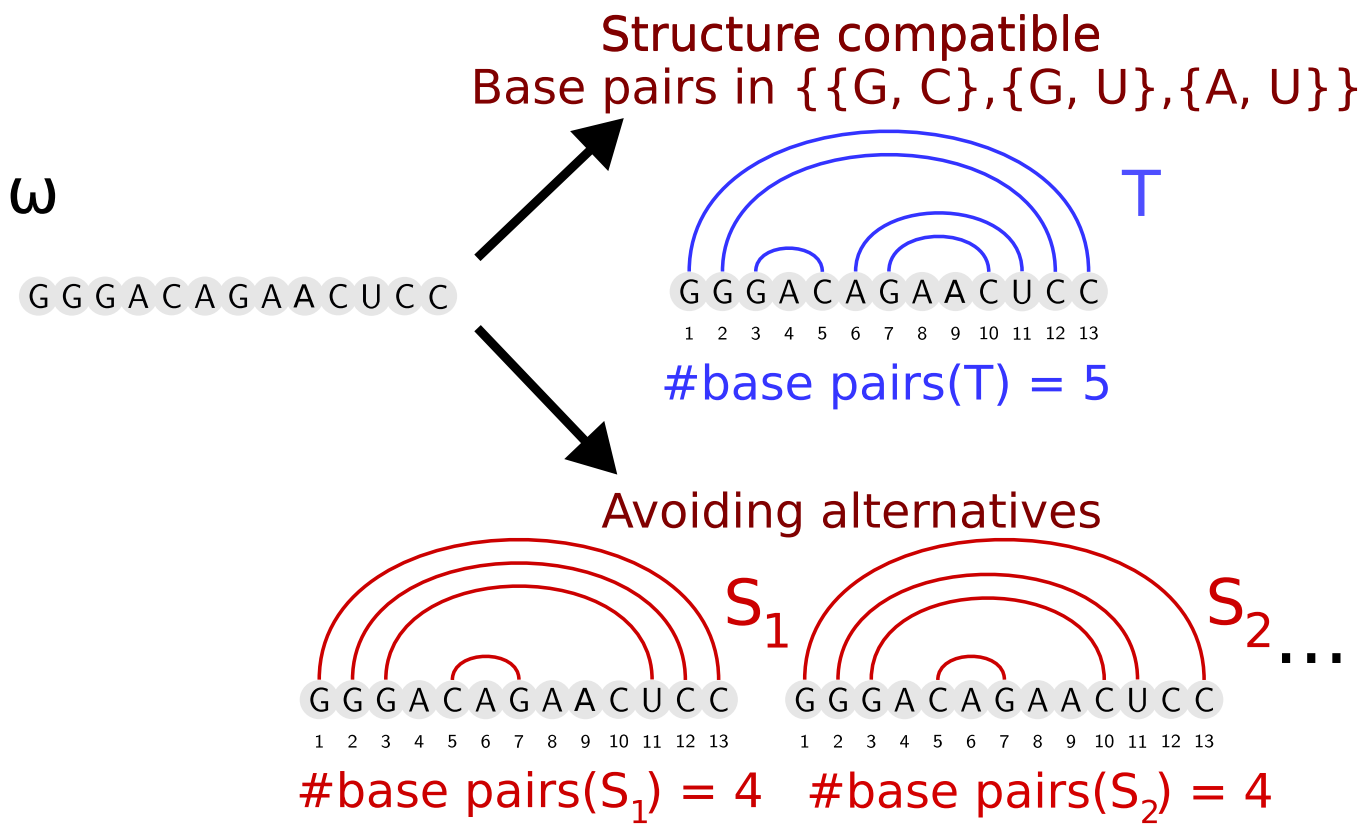
Théo Boury[1], Laurent Bulteau[2], Yann Ponty[1]

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

RNA 2D folding vs RNA structural design



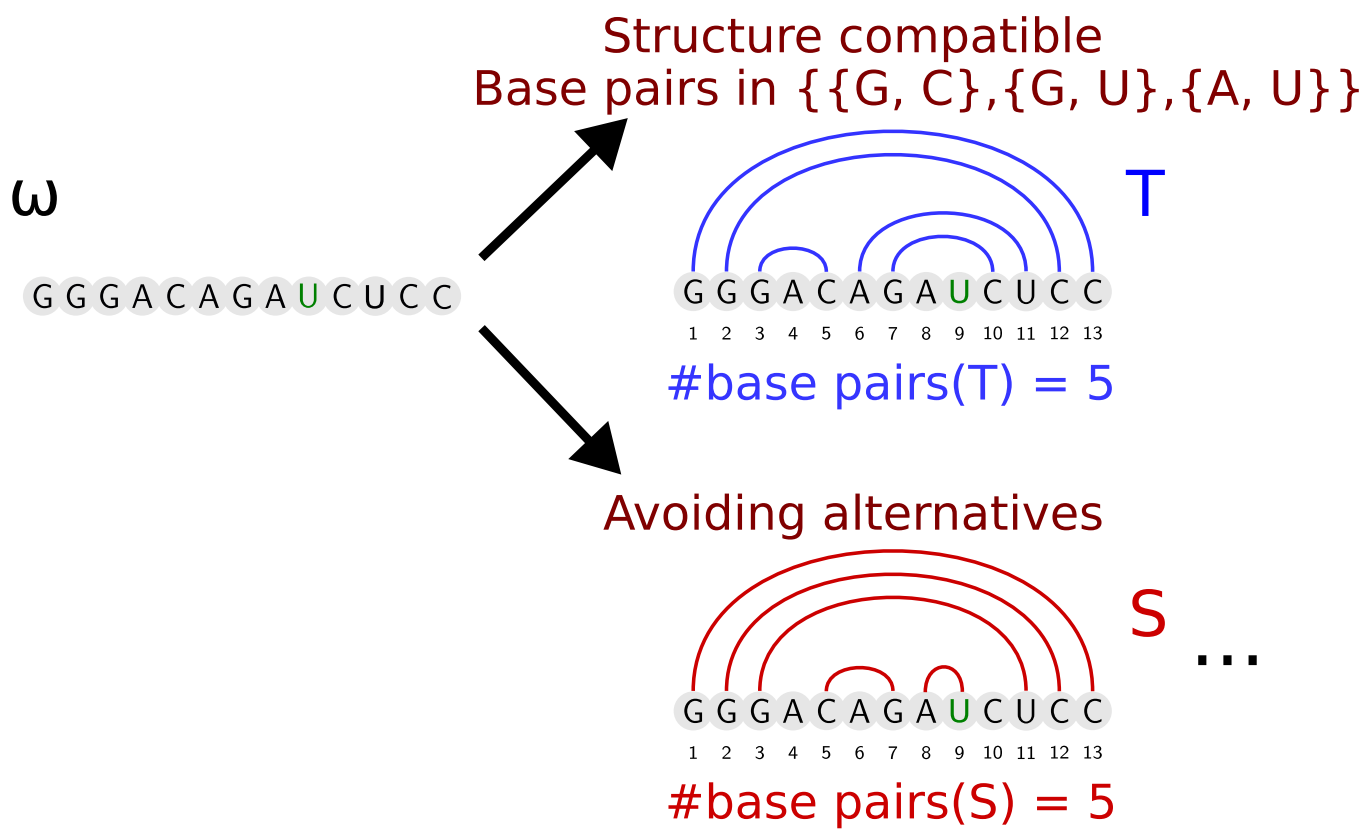
Inverse folding (IF): Formal definition



Goal: Find ω such that T is unique+optimal+valid fold for ω

$$\forall S \neq T, S \text{ comp. with } \omega, \#\text{BasePairs}(S) < \#\text{BasePairs}(T)$$

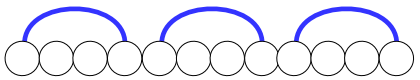
Inverse folding (IF): Formal definition



Goal: Find ω such that T is unique+optimal+valid fold for ω

$$\forall S \neq T, S \text{ comp. with } \omega, \#\text{BasePairs}(S) < \#\text{BasePairs}(T)$$

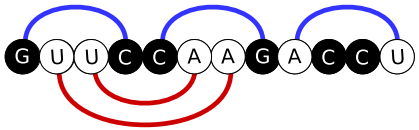
Minimal undesignable structures [Halès et al, 2017]

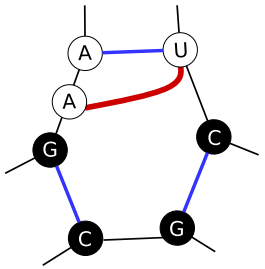


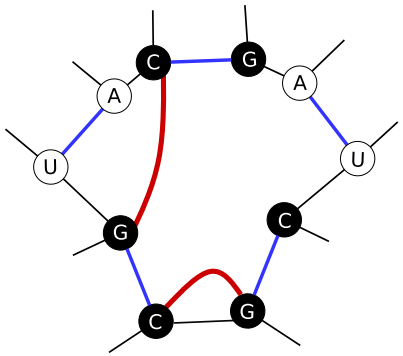
Undesignable structure

m3o motif

m5 motif

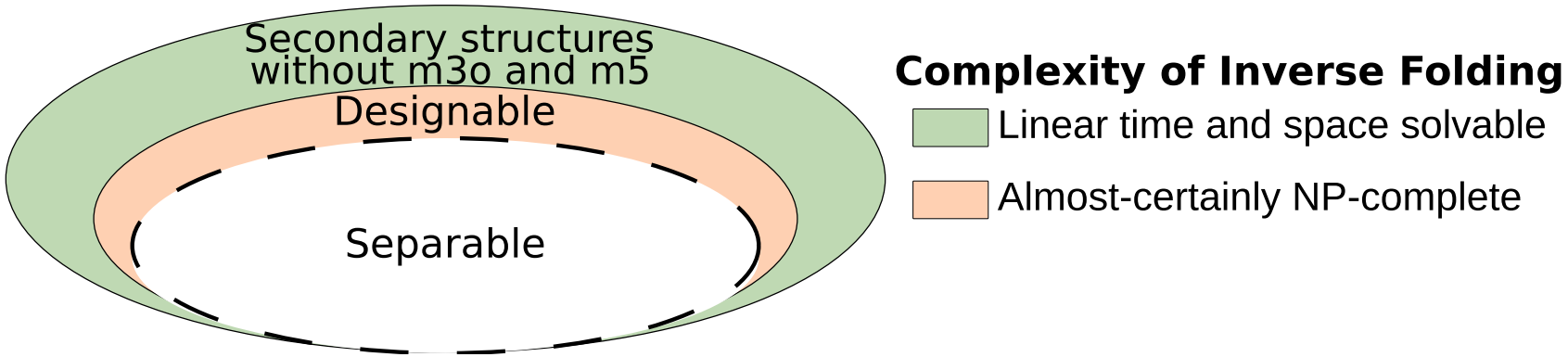






- Any target featuring occurrence of m3o or m5 is not designable
- Inverse folding (+ minor constraints) is NP-hard [Bonnet et al, 2018]
  $\rightarrow$ Infinite (+ exp. growth) list of min undesignable motifs
  (unless P=NP)
- Decision version of Inverse Folding not reducible to pattern matching

Design complexity

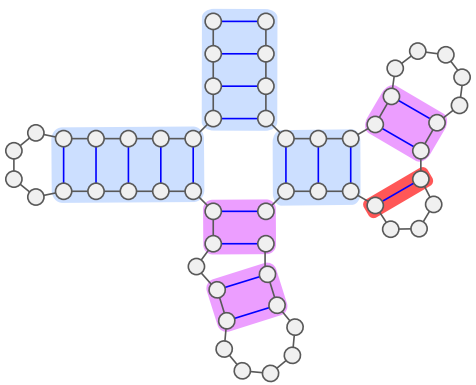


Inverse Folding efficiently solved for non-bonsai structures

$h_{min} =$ Minimum #base pairs in an helix

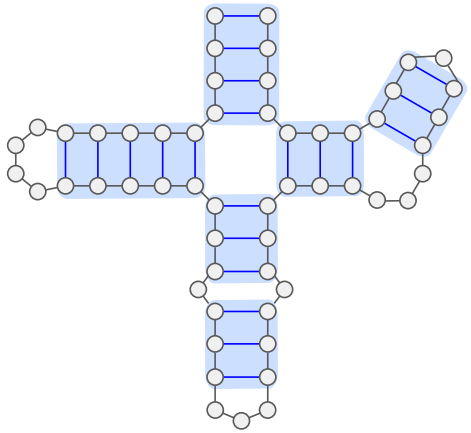
Theorem: Inverse folding (IF) is solved in linear time and space for secondary structures with $h_{min} = 3$. [This talk]

Structure with $h_{min} = 1$



Presumably NP-complete to design...

Structure with $h_{min} = 3$



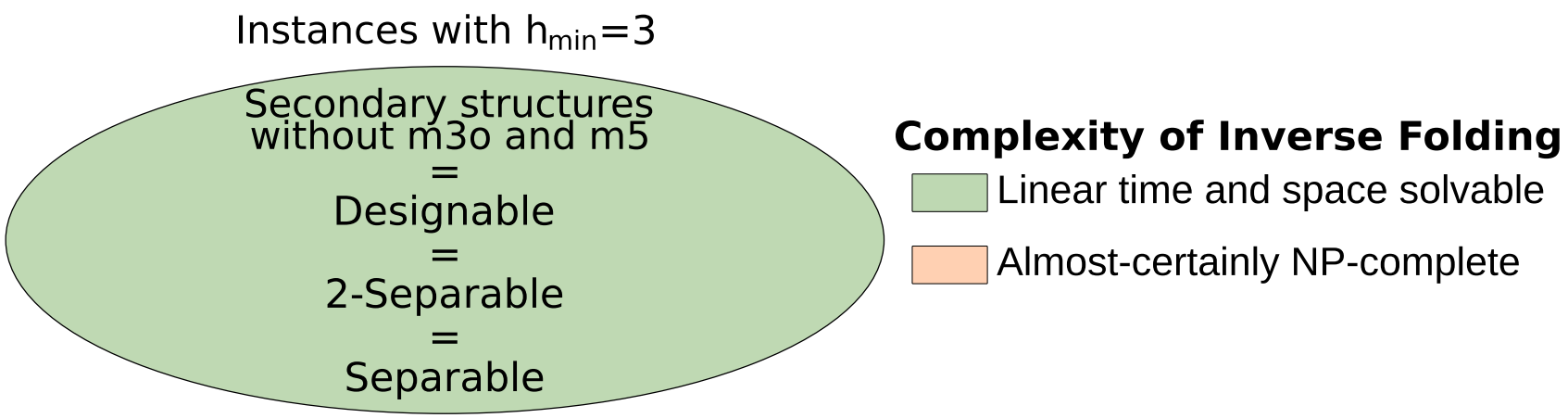
Designed in linear time!

Helices of size 3+

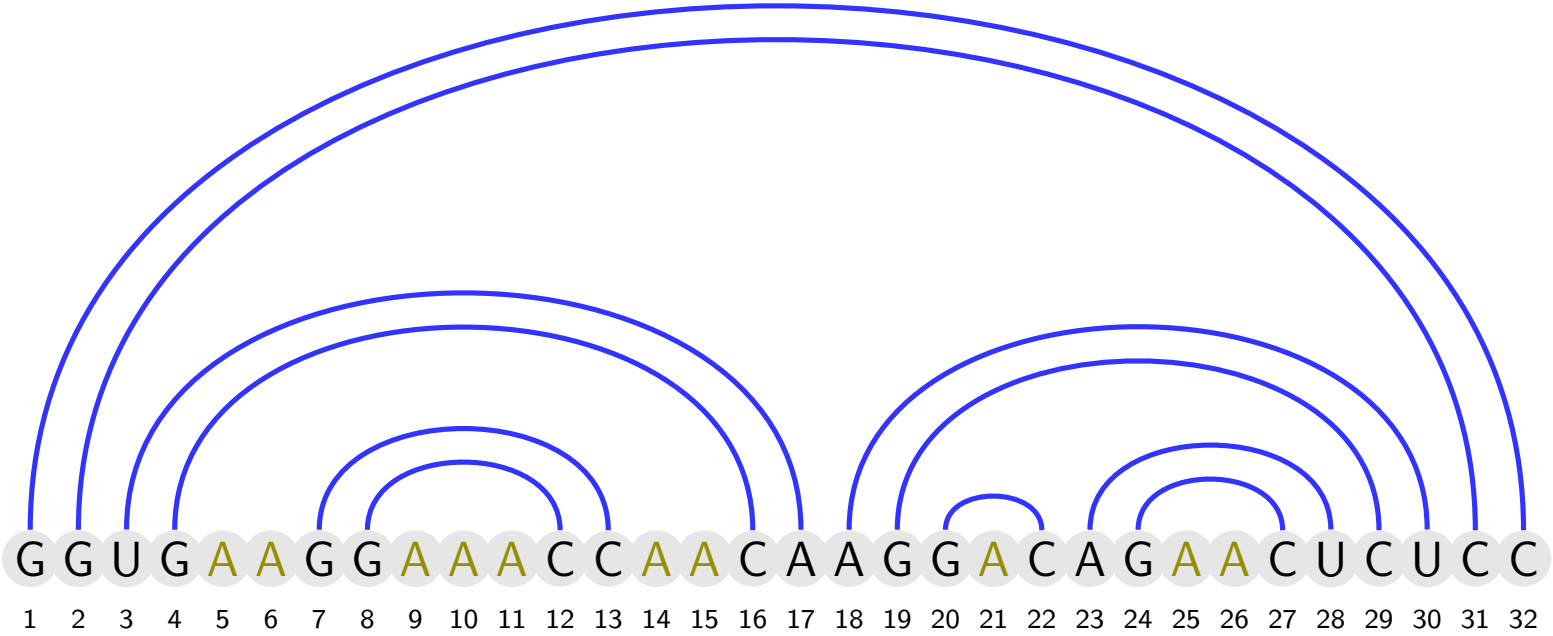
Isolated stacks (e.g. helices of size 2)

Isolated base pairs (e.g. helices of size 1)

Design complexity when $h_{min} = 3$



Technical point: Separable structure [Halès et al, 2017]

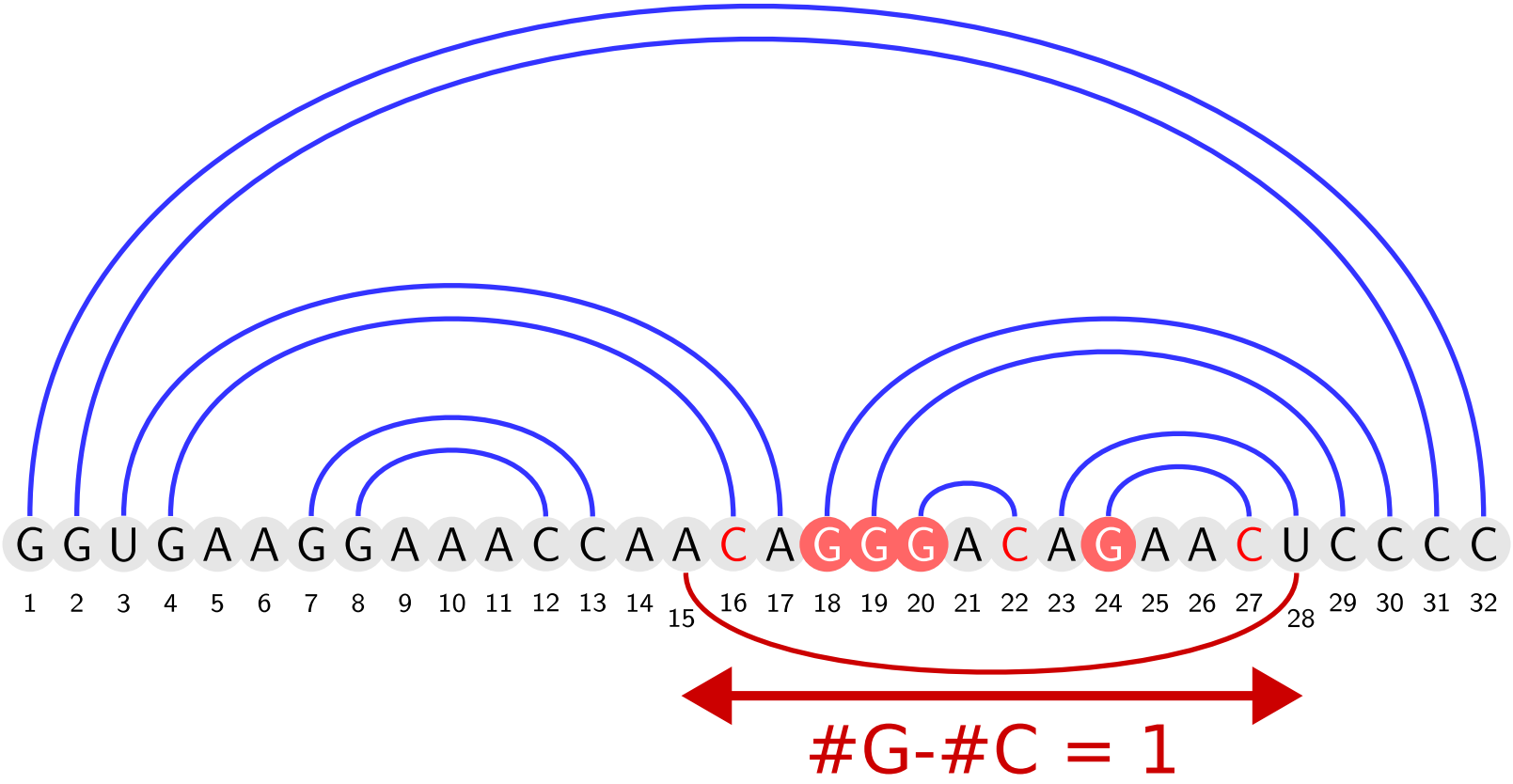


- Unpaired: A, Paired: G, U, C
- Separated sequence $\leftrightarrow$ $\forall$ alternative A-U, $\#G - \#C \neq 0$
- Separable structure $\leftrightarrow$ $\exists\ \omega$, separated sequence $\rightarrow \omega$, solution for IF

Result: [This talk]

Deciding if a structure is separable is NP-complete

Technical point: Separable structure

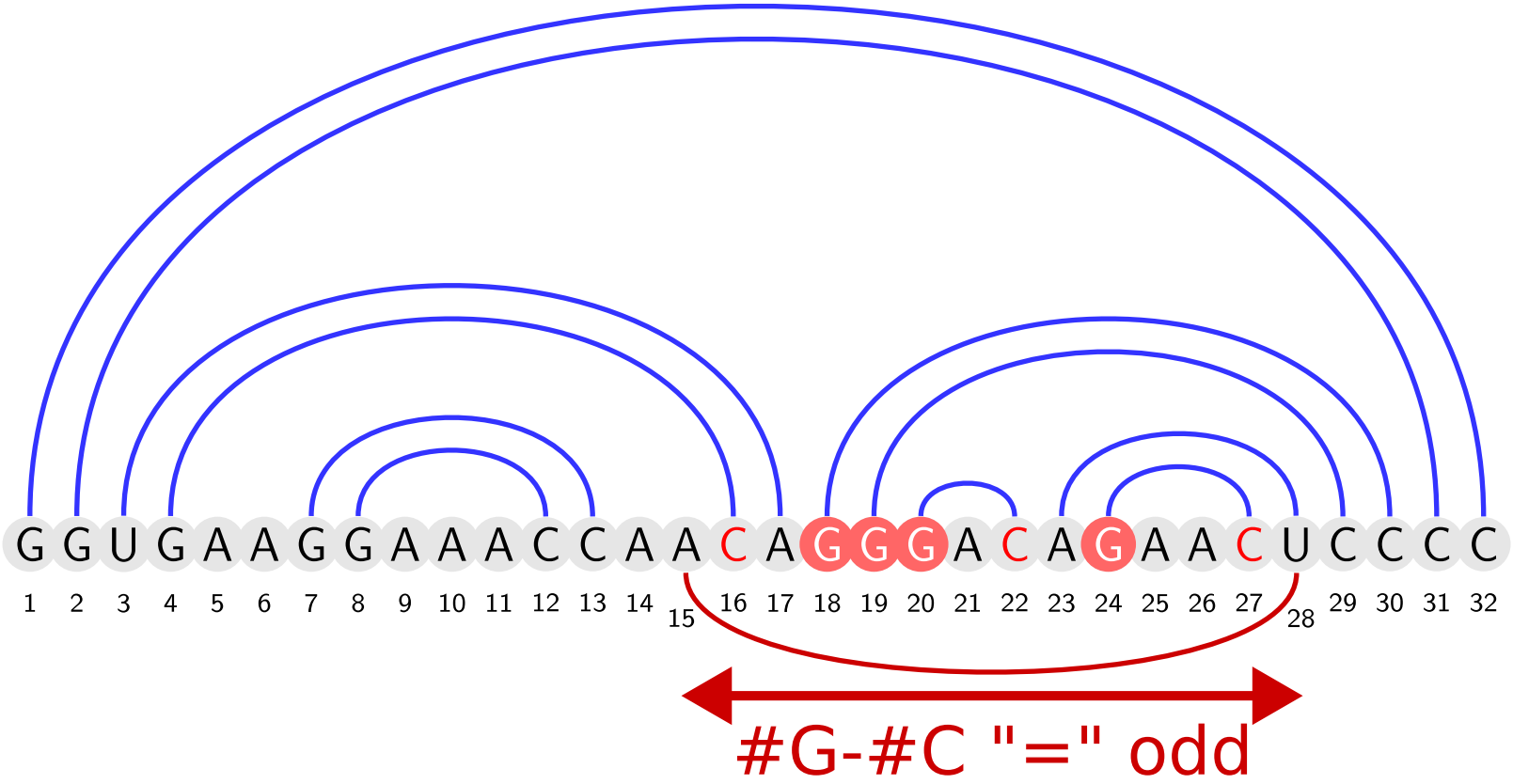


- Unpaired: A, Paired: G, U, C
- Separated sequence $\leftrightarrow$ $\forall$ alternative A-U, $\#G - \#C \neq 0$
- Separable structure $\leftrightarrow$ $\exists\ \omega$, separated sequence $\rightarrow \omega$, solution for IF

Result: [This talk]

Deciding if a structure is separable is NP-complete

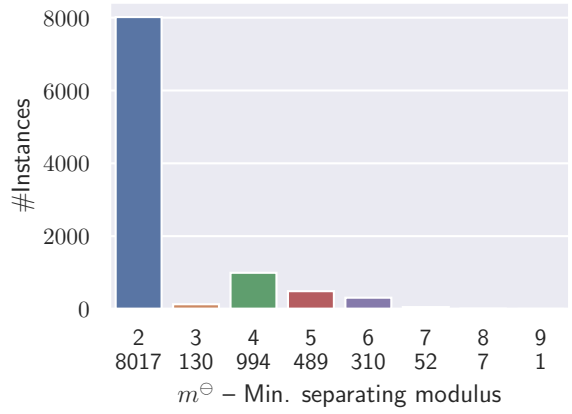
m-separable structure

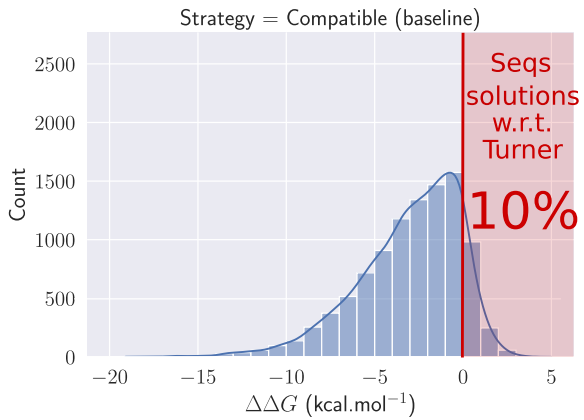


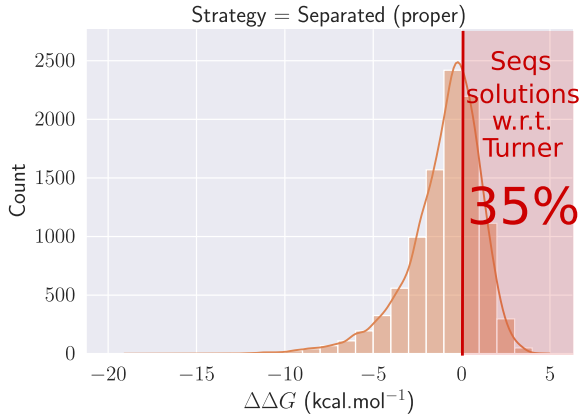
Result: [This talk]

- Core: Structures with $h_{min} = 3 \rightarrow$ 2-separable
- In general: m-separated design is $\Theta(n.m.2^m)$ time and $\Theta(m.n)$ space (e.g. Fixed-Parameter Tractable in m)
- Corollary: Structures with $h_{min} = 3$ solved in $O(n)$!
- Bonus: Uniform sampling of sequences

Experimental results at a glance







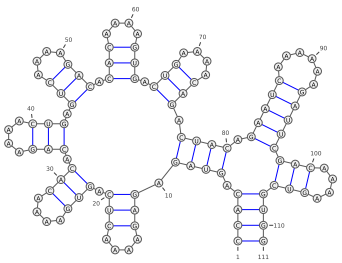
- Structures with $h_{min} \leqslant 2 \rightarrow$ Mainly designable in practice (around 100 nucleotides), but exponential decay of numbers of solutions
- Sampled solutions from 2-separable structures are more promising according to the Turner energy model than random compatible sequences

Conclusion

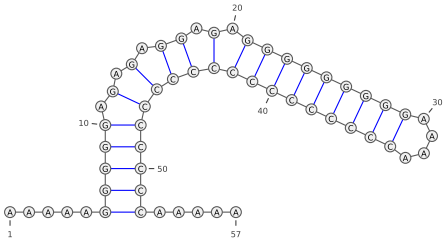
- $h_{min} = 3 \rightarrow$ IF solved in linear time with 2-separable
- $h_{min} \leqslant 2 \rightarrow$ IF partially solved, FPT in m with m-separable

Ongoing work (Return to the REAL world[1])

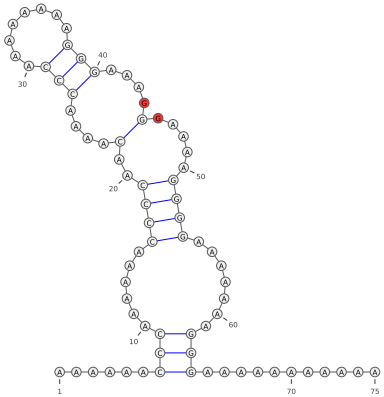
- Solutions as seed sequences of RNAinverse
- Test on Eterna v2 benchmark (100 artificial puzzles)
  $\rightarrow$ almost everyone is 2-separable
- One puzzle has a multiloop of degree 25
  $\rightarrow$ still linear, just with a large constant
- solved $\sim$80 puzzles, (almost) immediate solutions in half of them



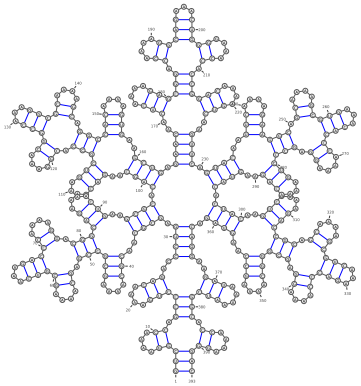
Puzzle #58



Puzzle #27



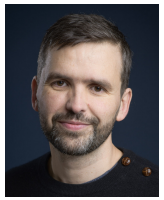
Puzzle #75



Puzzle #76

[1] where Hua-Ting tries to steal an authorship

Thanks to...



Yann Ponty

Laurent Bulteau

Collaborator: Hua Ting Yao (that did the last slide)

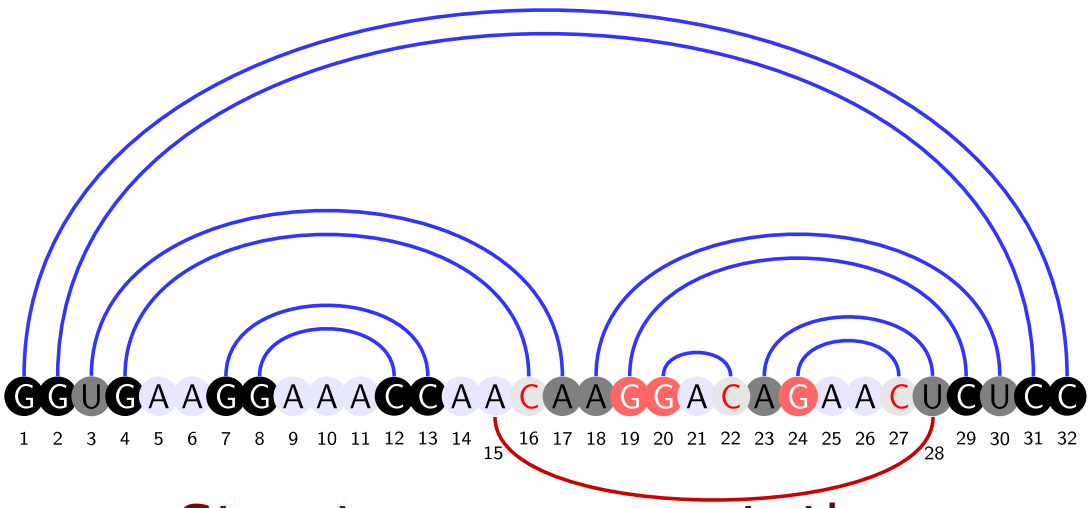
And to the other members of the AMIBio team:

Sarah Berkemer
Sebastian Will
Alan Azede
Nan Pan

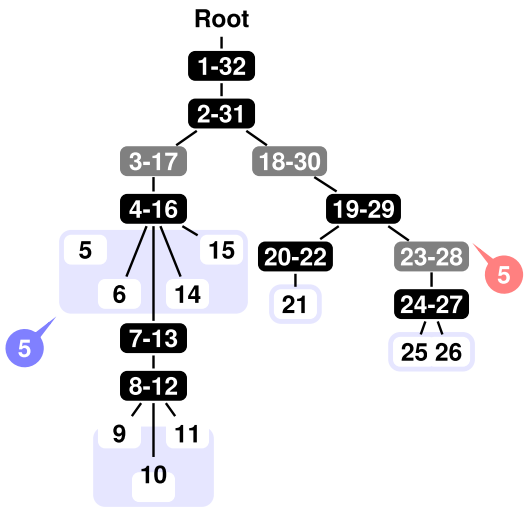
Link to the paper

You!

Tree formalism



Structure representation



Tree representation

Definition (Levels): Given a tree coloring, the level $L : V(T) \to \mathbb{Z}$ of a node v is $L(v) := |p|_{\bullet} - |p|_{\circ}$ where p denotes the color vector associated with the node sequence from parent(v) to Root.

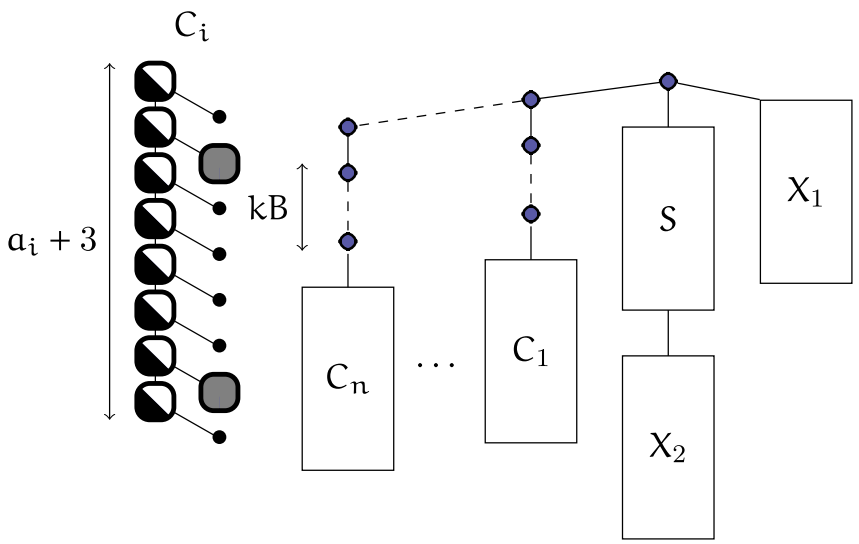
Decide separability is NP-complete

Problem 1 (Interval Packing):

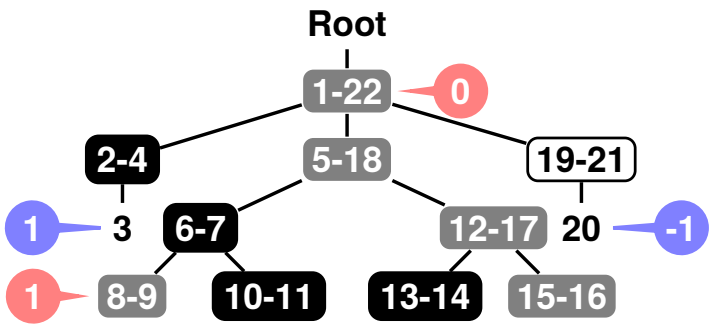
Input: set of distinct integers $A = \{a_1, \cdots, a_n\}$, integers k and B

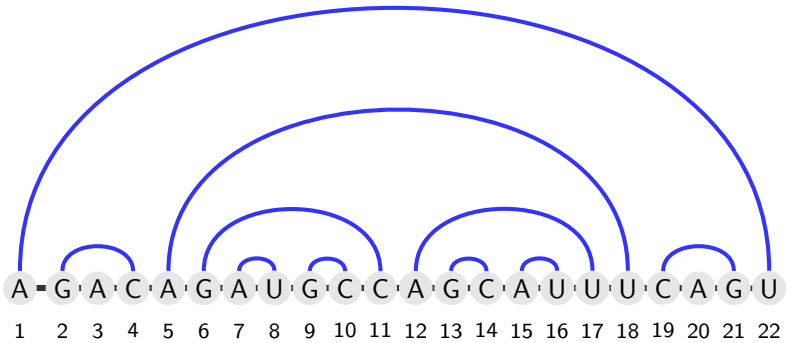
Output: function x from A to intervals of $[\![0, kB - 1]\!]$ such that:

- $x(a_i)$ is an interval of size a_i
- $x(a_i)$ and $x(a_j)$ are disjoint for $i \neq j$
- $x(a_i)$ does not contain both $jB - 1$ and jB for any i, j.



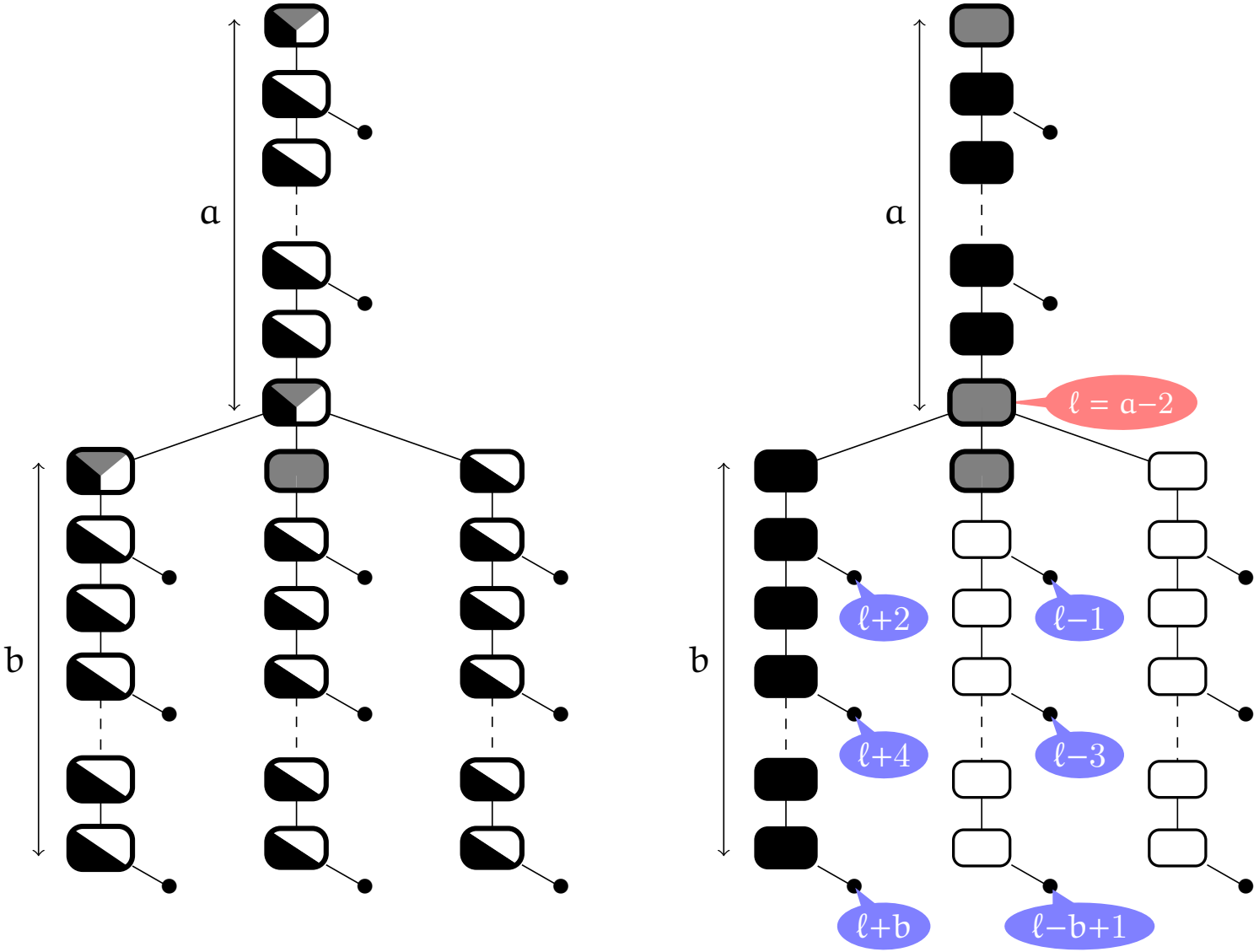
Which instances are non-separable but designable?





- Harder to find with helices of size 2 or more. (currently more than 1000 nucleotides long)

Core widget of the designable non-separable instances with helices of size 2



Modulo-separability

Definition ((Modulo) m-separability): Let m be an integer. A coloring *Color* is <u>m-separated</u> (or <u>separated with modulus m</u>) for a target secondary structure T, if an only if

$$\{Lv(v) \bmod m \mid Color(v) = \bullet\} \cap \{Lv(v) \bmod m \mid v \text{ is a leaf}\} = \varnothing$$

- Modulo separability coincides with separability with $m \geqslant \frac{n}{2}$

Problem 2 (Modulo Separability):
Input: A tree T (with no $m_{3\bullet}$ or m_5 motif), a modulus $m \in \mathbb{N}$
Output: A coloring of T that is m-separated, or $\perp$ if no such coloring exists.

Dynamic programming scheme for modulo separability

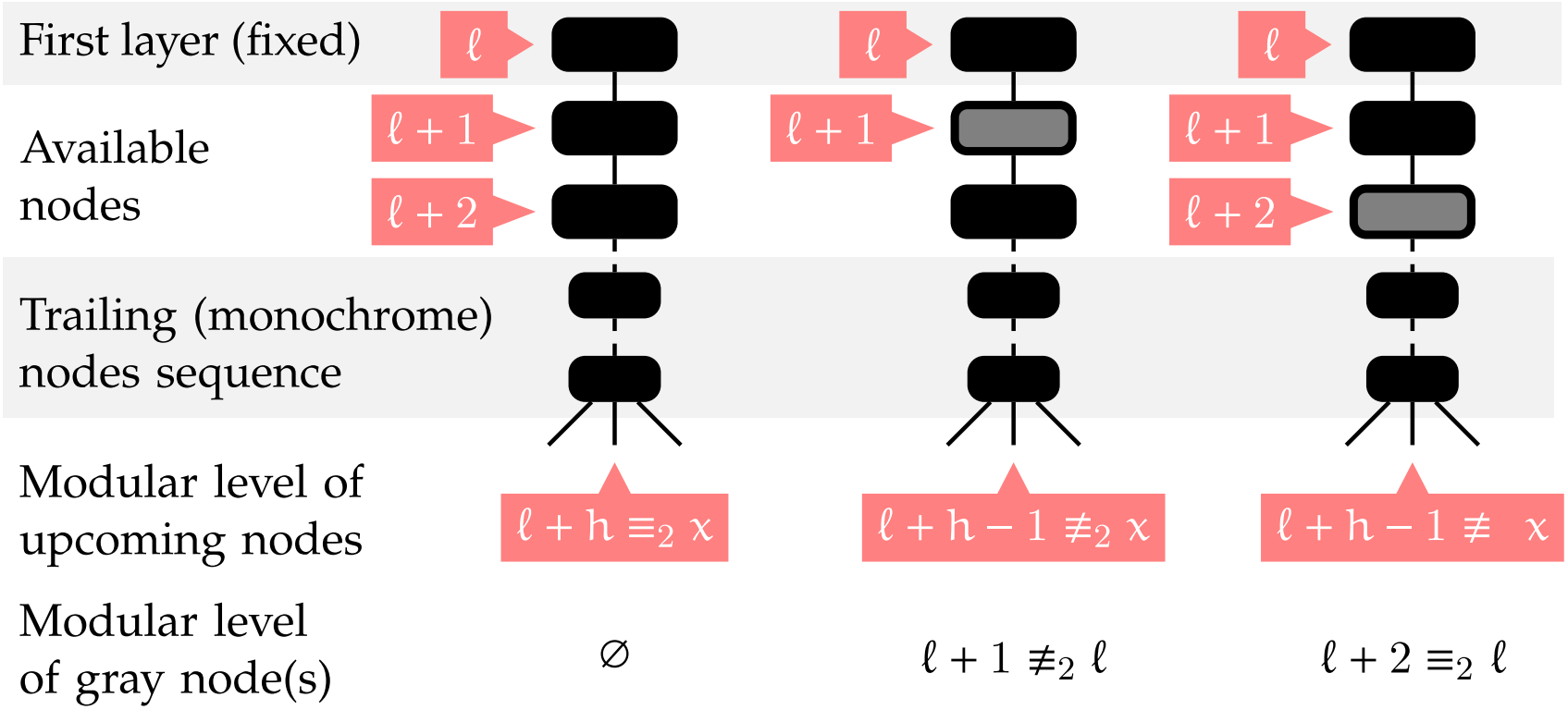
$$\mathrm{d}^{\xi_L}_{v\to c,l} = \begin{cases} \text{False} & \text{if } \ell \in \xi_L \wedge c = \bullet \\ & \text{or } \ell' \notin \xi_L \text{, and} \\ & \exists \text{ leaf in children}(v) \\ \text{True} & \text{if children}(v) = \emptyset \\ \displaystyle\bigvee_{\substack{c' \text{ "valid" coloring of} \\ \text{children}(v) \\ \text{given } v\to c}} \bigwedge_{v' \in \text{children}(v)} \mathrm{d}^{\xi_L}_{v'\to c'(v'),\ell'} & \text{otherwise.} \end{cases}$$

with $\ell' := \ell + \delta(c) \bmod m$

- $\mathrm{d}^{\xi_L}_{v\to c,\ell}$: existence of a valid assignment for a subtree of T rooted at internal node v, with v occurring at level ℓ, and being assigned a prior color c.
- ξ_L: Leaves levels (thus $[\![0, m[\![\setminus \xi_L$ are $\bullet$ levels.)
- δ: level increment induced by a color c

Instances with helices of size 3 or more are all separable

$$\boxed{x := \ell + h \bmod 2}$$



Theorem: Secondary structures with helices of size 3 or more are 2-separable (thus designable) in linear time

Inverse folding: Complexity Zoo

- NP-hard, 2008, Schnall-Levin et al $\cdots$
- Linear, 2017, Halès et al $\cdots$

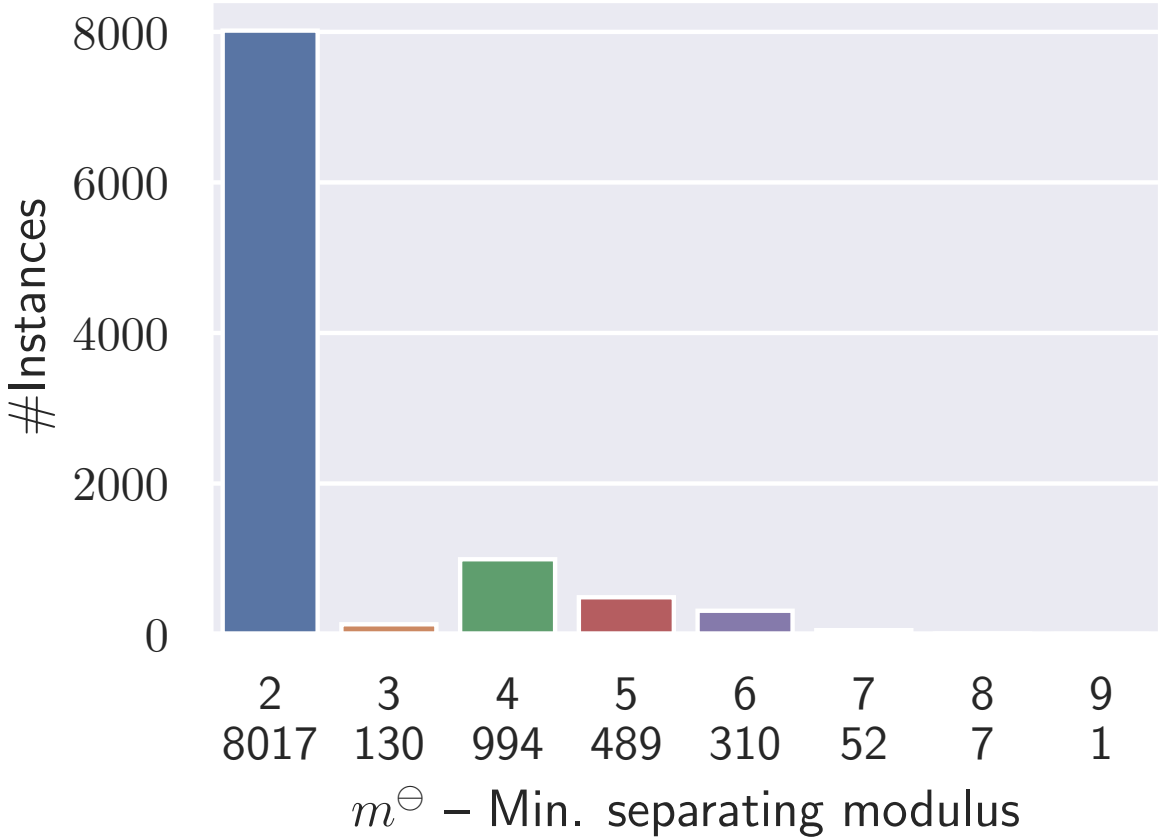
  But only on a subset called "separable instances".
- NP-hard, 2018, Bonnet et al $\cdots$

  But only an extension with constrained base pairs.
- Our contribution:

  Linear by avoiding isolated base pairs and stacks, 2024, Boury et al.

Beyond helices of size 3: instances with helices of size 2

There is no certainty that these instances are Modulo *m*-separable!

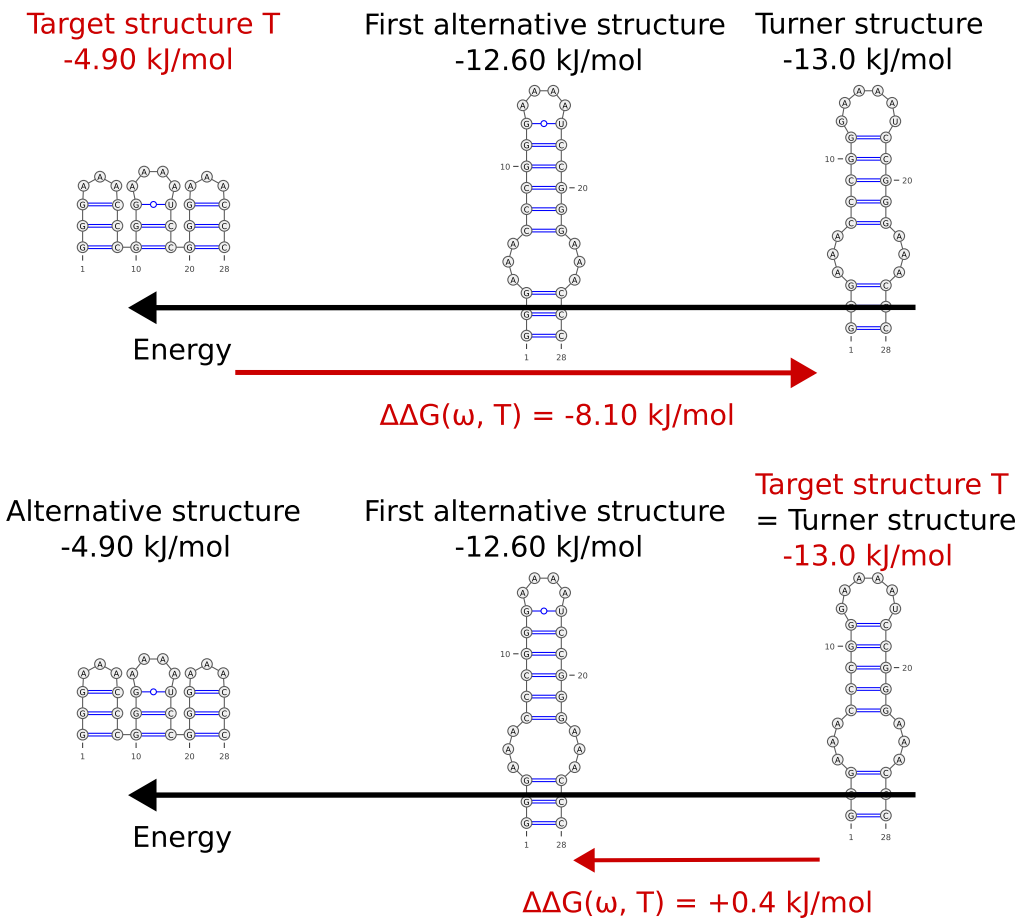


- Surprisingly enough, all instances containing helices of size 2 were found Modulo *m*-separable thus designable.

Turner energy of designed sequence with helices of size 3

$$\Delta\Delta G(\omega, T) := \Delta G(\omega, \alpha(w, T)) - \Delta G(\omega, T)$$

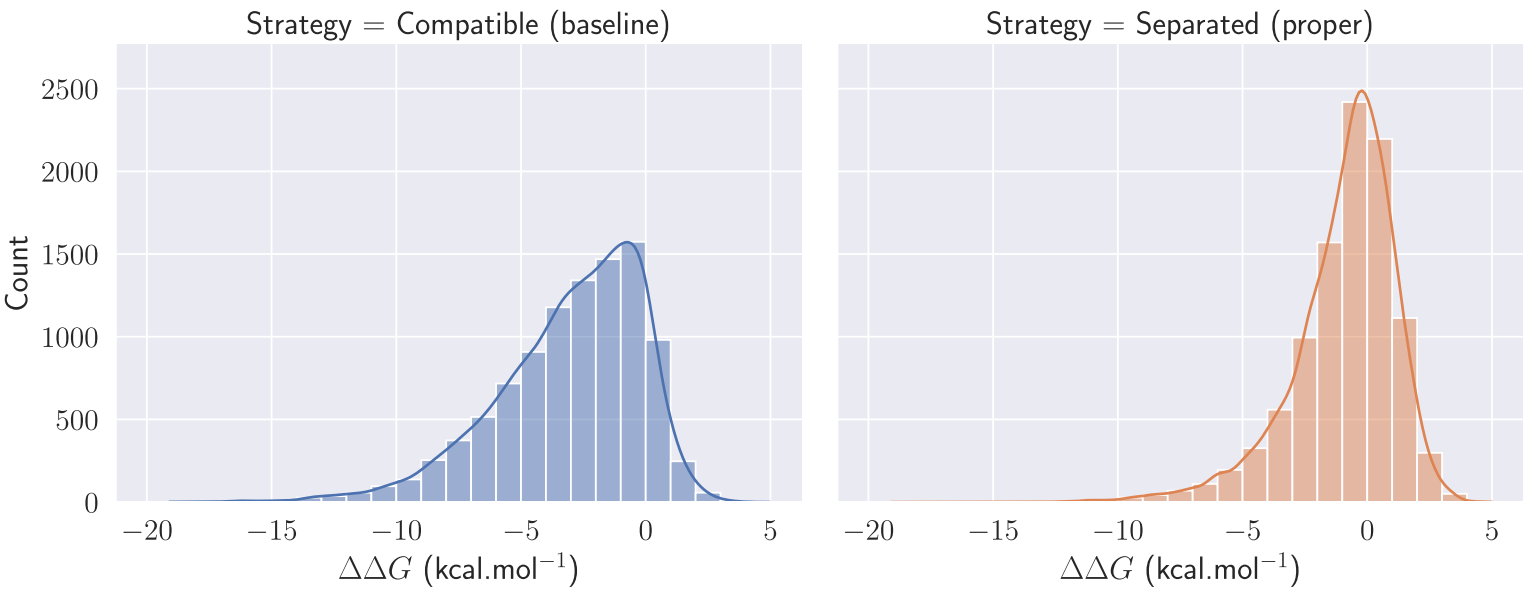
$$\alpha(\omega, T) := \min\{\Delta G(\omega, T') \mid |T' \triangle T| \geq 3\}$$



Turner energy of designed sequence with helices of size 3

$$\Delta\Delta G(\omega, T) := \Delta G(\omega, C(w, T)) - \Delta G(\omega, T)$$

$$C(\omega, T) := \min\{\Delta G(\omega, T') \mid |T' \triangle T| \geq 3\}$$



- Even if guarantied only in a base pairs model, our sequences represent better competitor in Turner energy model than simply compatible sequences