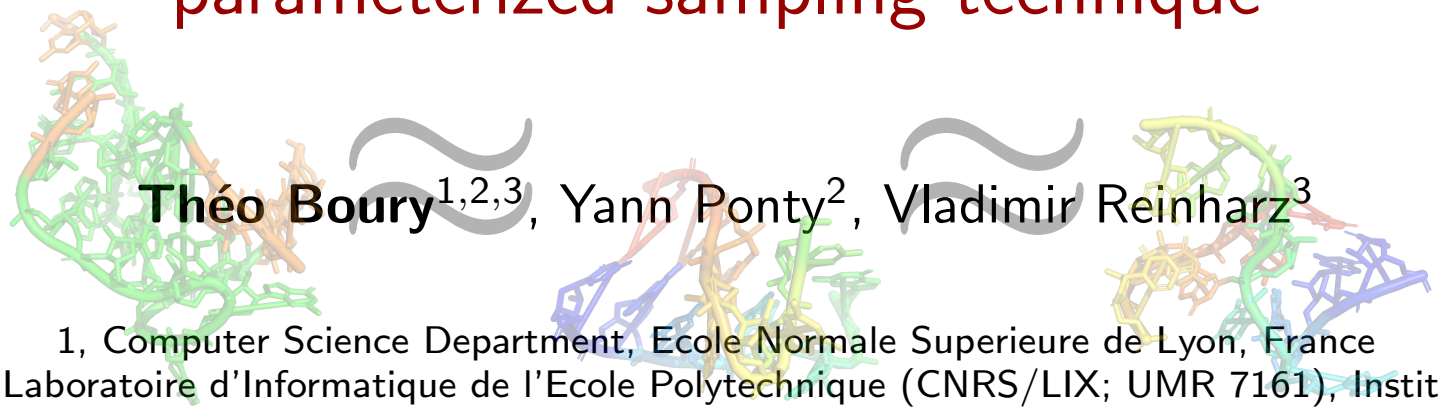
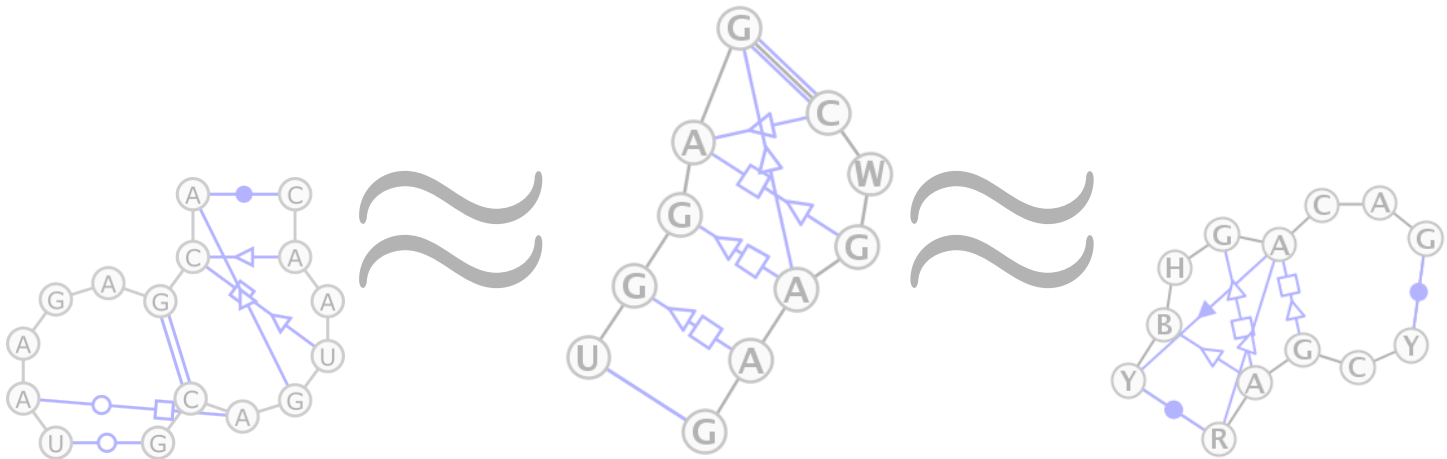
Automatic exploration of the natural variability of RNA non-canonical geometric patterns with a parameterized sampling technique



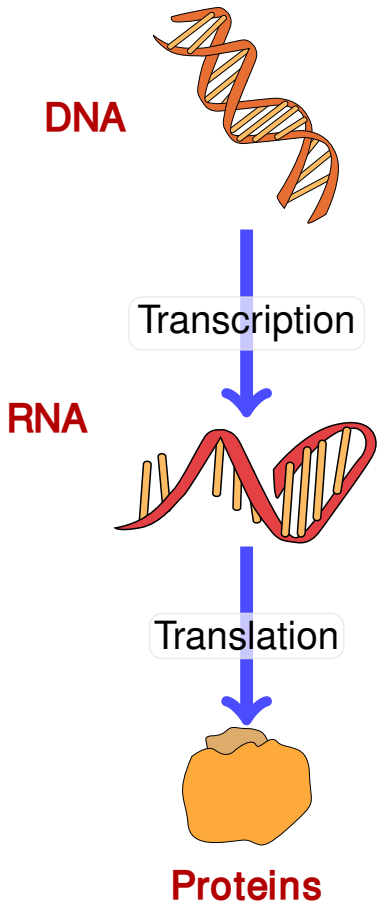
Théo Boury[1,2,3], Yann Ponty[2], Vladimir Reinharz[3]

1, Computer Science Department, Ecole Normale Superieure de Lyon, France
2, Laboratoire d'Informatique de l'Ecole Polytechnique (CNRS/LIX; UMR 7161), Institut Polytechnique de Paris, France
3, Department of Computer Science, Université du Québec à Montréal, Canada



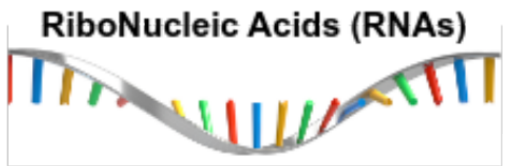
Outline

The RiboNucleic Acid (RNA) molecule... a simple intermediate ?

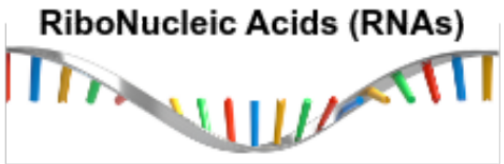


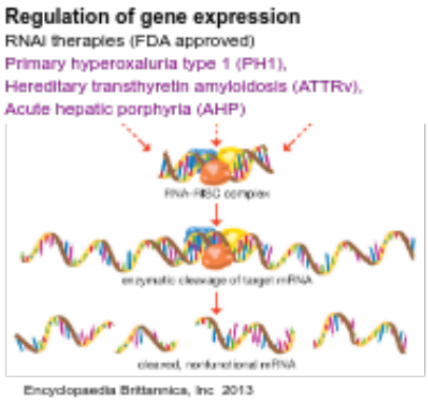
- RNA is a molecule of interest in itself

RNA "functions"



A few RNA "functions"?



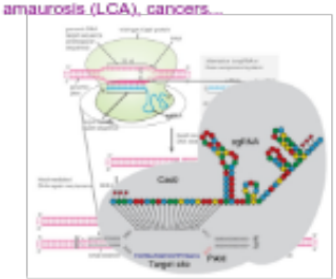


A few RNA "functions"?

Targeting system for DNA Editing

CRISPR therapies

Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...

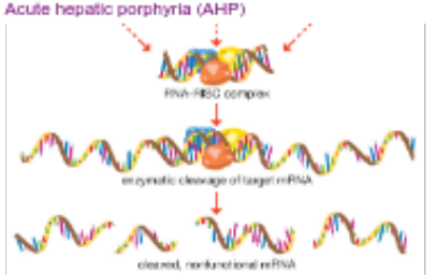


Hendel et al. 2015; Agrotis & Ketteler, 2015

Regulation of gene expression

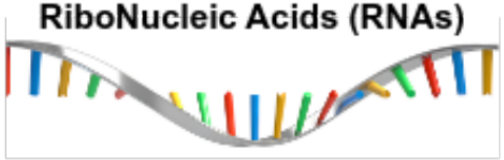
RNAi therapies (FDA approved)

Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



Encyclopaedia Britannica, Inc. 2013

RiboNucleic Acids (RNAs)

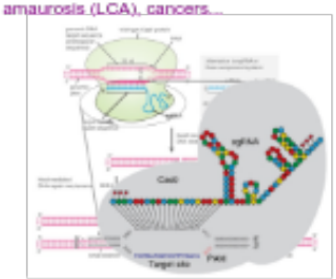


Some RNA "functions"

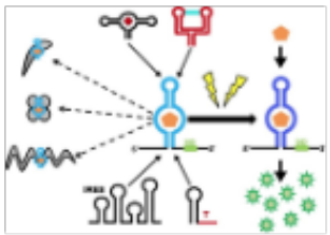
Targeting system for DNA Editing

CRISPR therapies

Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



Hendel et al. 2015; Agrotis & Ketteler, 2015

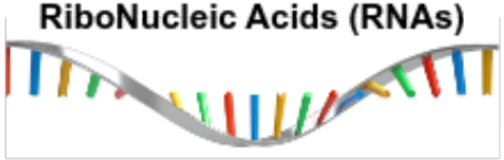


Berens & Suess 2015

Sensor of metabolites

Riboswitches

RiboNucleic Acids (RNAs)



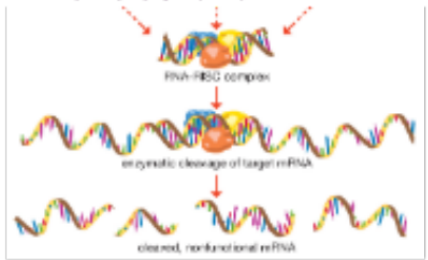
Regulation of gene expression

RNAi therapies (FDA approved)

Primary hyperoxaluria type 1 (PH1),

Hereditary transthyretin amyloidosis (ATTRv),

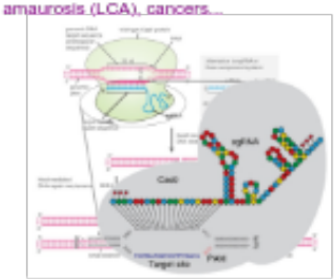
Acute hepatic porphyria (AHP)



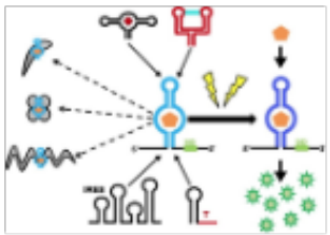
Encyclopaedia Britannica, Inc. 2013

Some RNA "functions"

Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



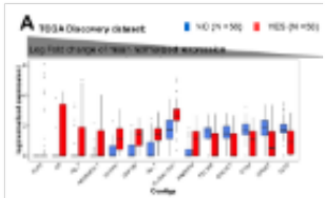
Hendel et al. 2015; Agrotis & Ketteler, 2015



Berens & Suess 2015

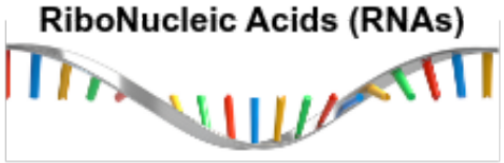
Sensor of metabolites
Riboswitches

Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...

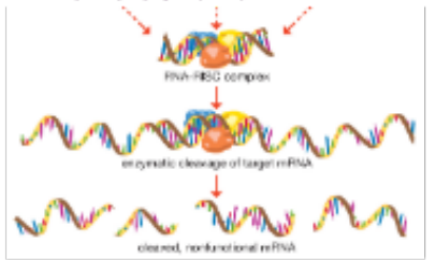


[NGuyen *et al*, 2021]

RiboNucleic Acids (RNAs)



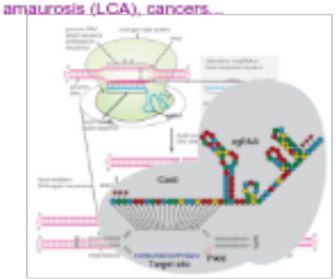
Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



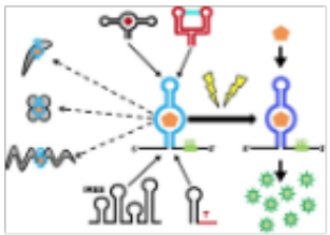
Encyclopaedia Britannica, Inc. 2013

A lot of RNA "functions"!

Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



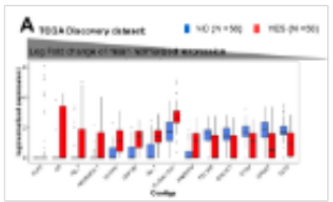
Hendel et al. 2015; Agrotis & Ketteler, 2015



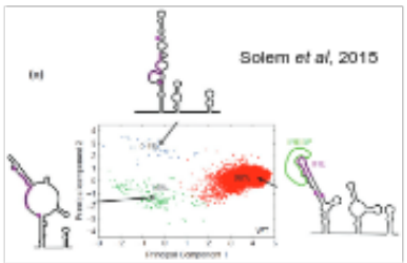
Berens & Suess 2015

Sensor of metabolites
Riboswitches

Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...

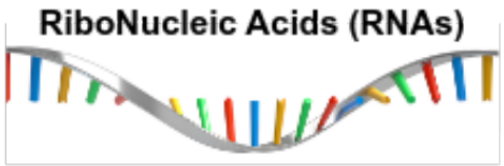


[NGuyen *et al*, 2021]

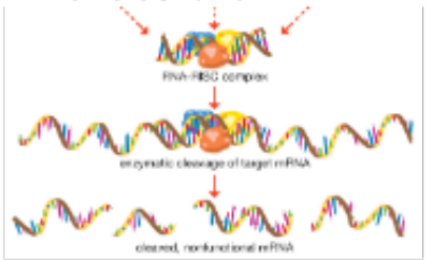


Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy,
Cystic fibrosis, Rett syndrome...

RiboNucleic Acids (RNAs)



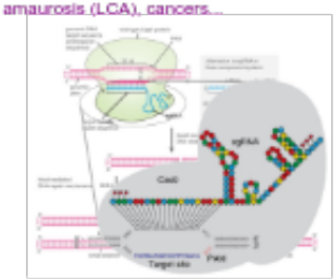
Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



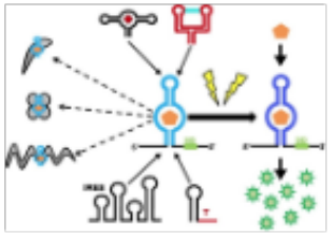
Encyclopaedia Britannica, Inc. 2013

A lot of RNA "functions"!

Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



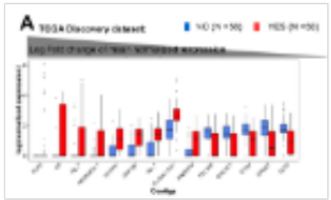
Hendel et al. 2015; Agrotis & Ketteler, 2015



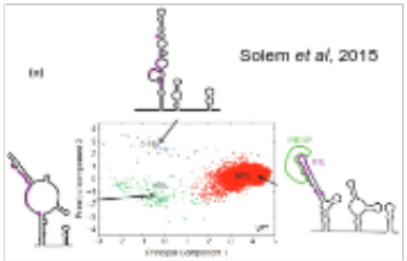
Berens & Suess 2015

Sensor of metabolites
Riboswitches

Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...

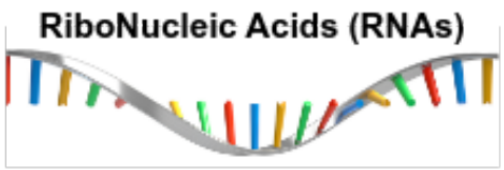


[NGuyen *et al*, 2021]

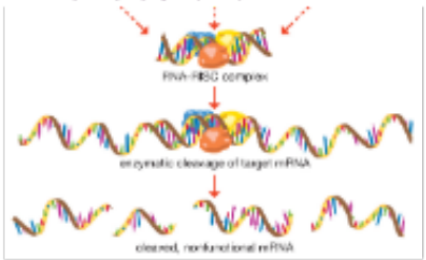


Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy,
Cystic fibrosis, Rett syndrome...

RiboNucleic Acids (RNAs)

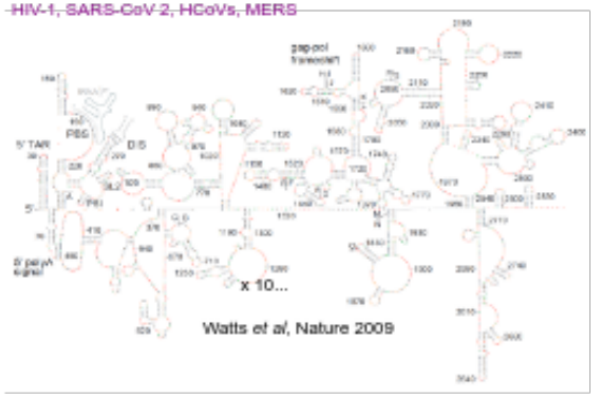


Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



Encyclopaedia Britannica, Inc. 2013

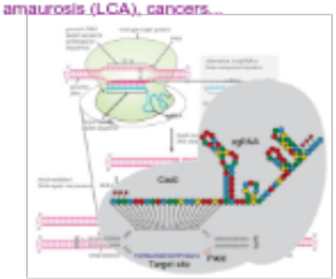
Genomic material for Human pathogens
HIV-1, SARS-CoV-2, HCoVs, MERS



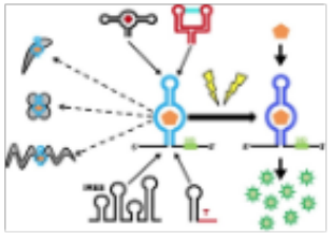
Watts *et al*, Nature 2009

A lot of RNA "functions"!

Targeting system for DNA Editing
CRISPR therapies
Sickle-cell anemia, β-thalassamia, Leber congenital amaurosis (LCA), cancers...



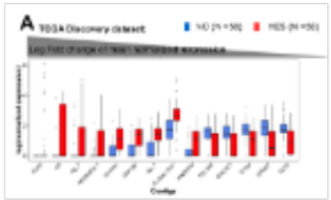
Hendel et al. 2015; Agrotis & Ketteler, 2015



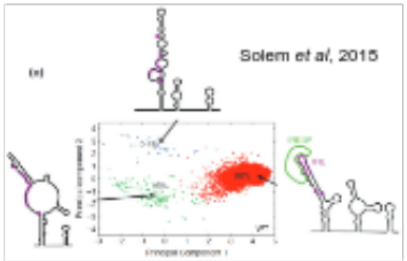
Berens & Suess 2015

Sensor of metabolites
Riboswitches

Quantitative expression
Transcriptomic signatures
Cancer diagnosis/prognosis/relapse...

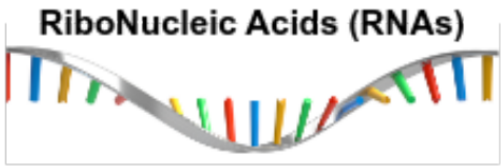


[NGuyen *et al*, 2021]

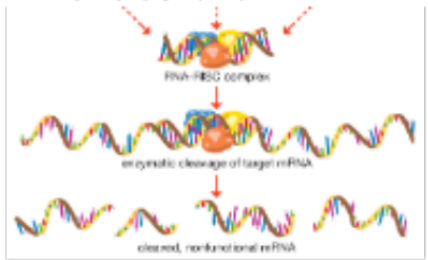


Non-coding mutations
lncRNAs, miRNAs, structure-associated (RiboSnitches)
β-thalassemia, duchenne muscular dystrophy,
Cystic fibrosis, Rett syndrome...

RiboNucleic Acids (RNAs)



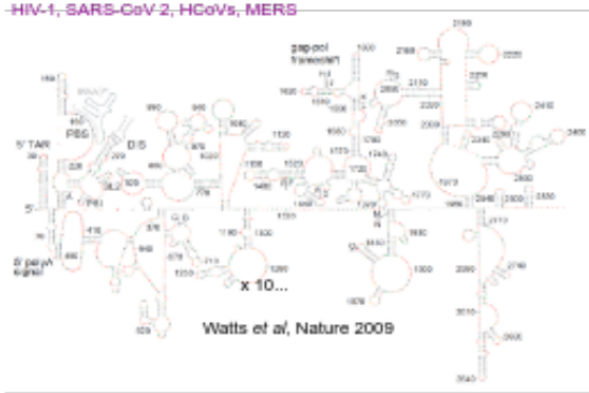
Regulation of gene expression
RNAi therapies (FDA approved)
Primary hyperoxaluria type 1 (PH1),
Hereditary transthyretin amyloidosis (ATTRv),
Acute hepatic porphyria (AHP)



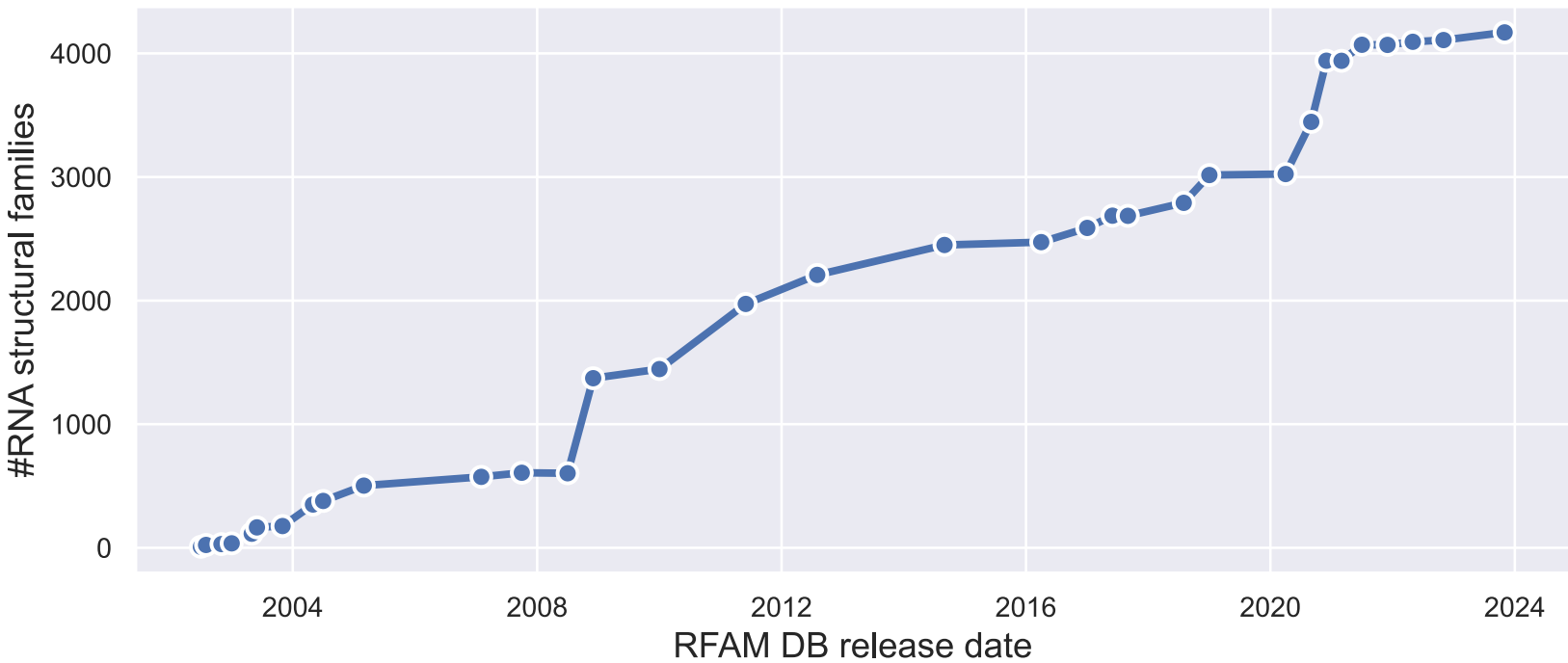
Encyclopaedia Britannica, Inc. 2013

Encodes proteins
mRNA Vaccines
COVID-19, Malaria (Zika, CMV, Cancers?)

Genomic material for Human pathogens
HIV-1, SARS-CoV-2, HCoVs, MERS

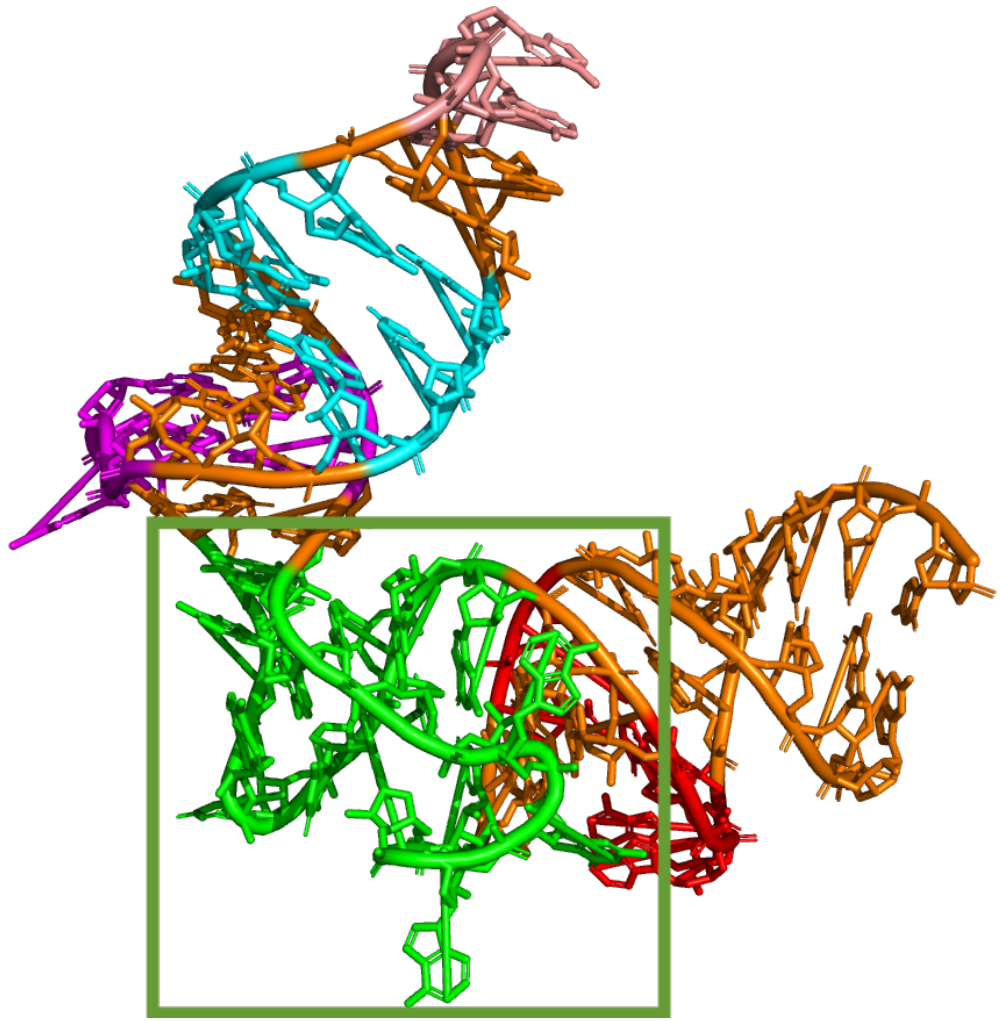


How many functional RNAs are there?

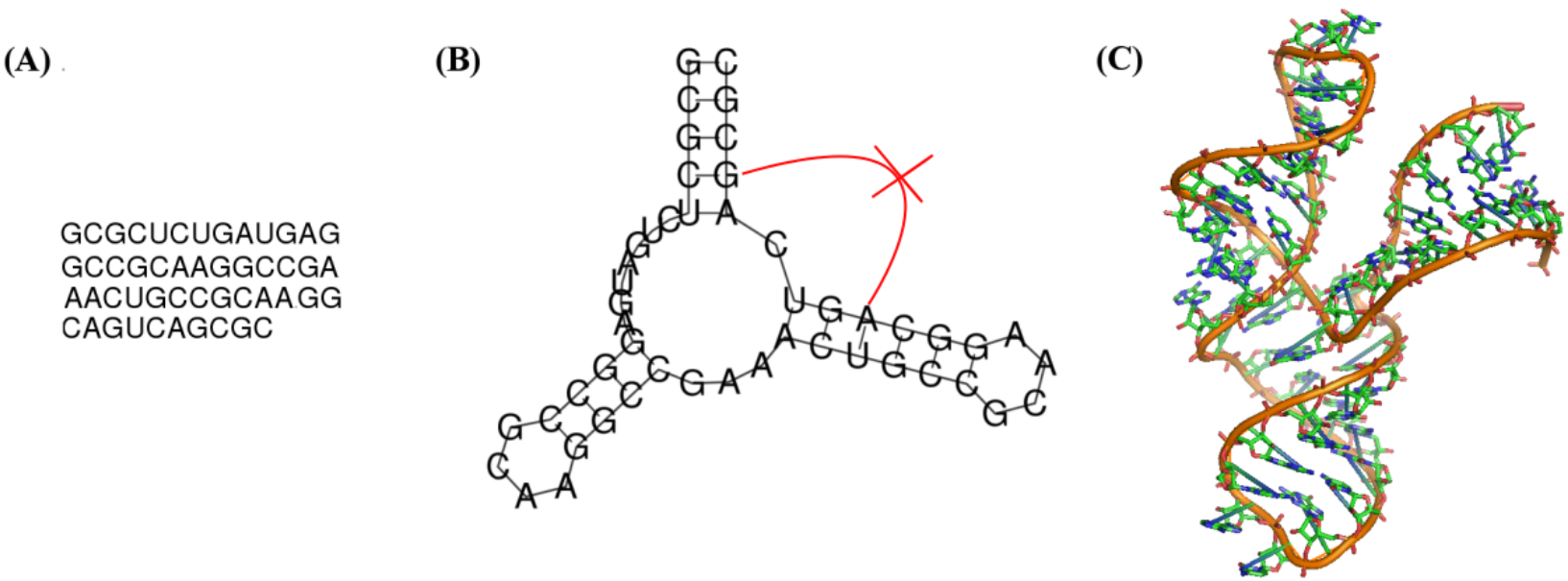


The 3D RNA structure

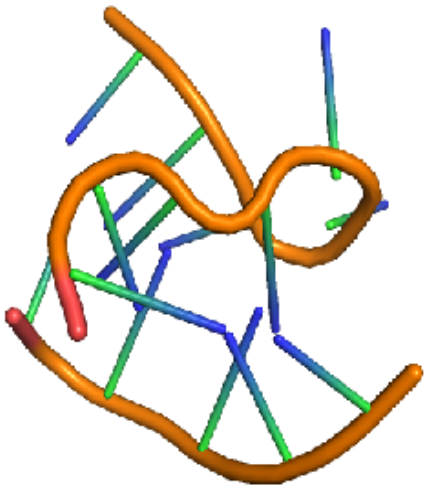
The 3D RNA structure



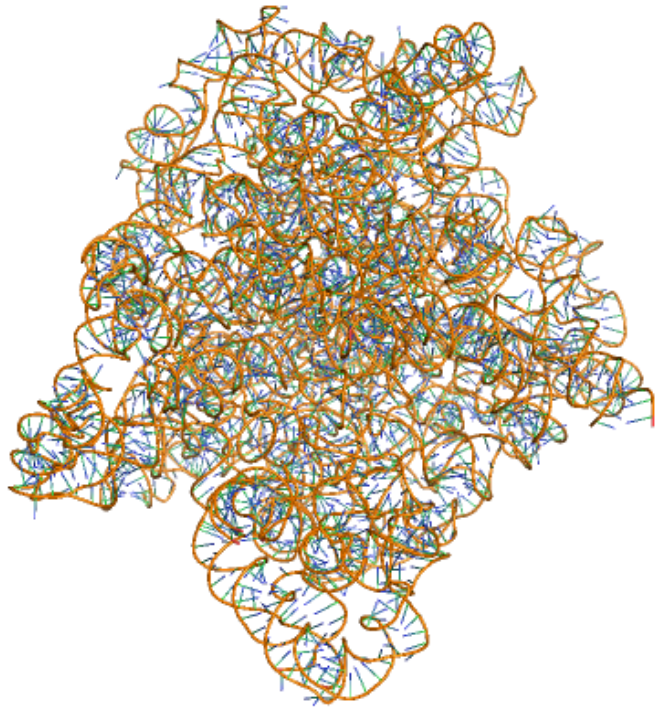
Different level of abstraction for RNA



Where is Waldo?

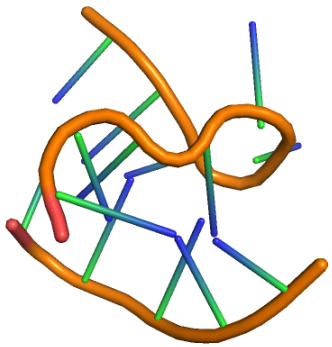


Motif

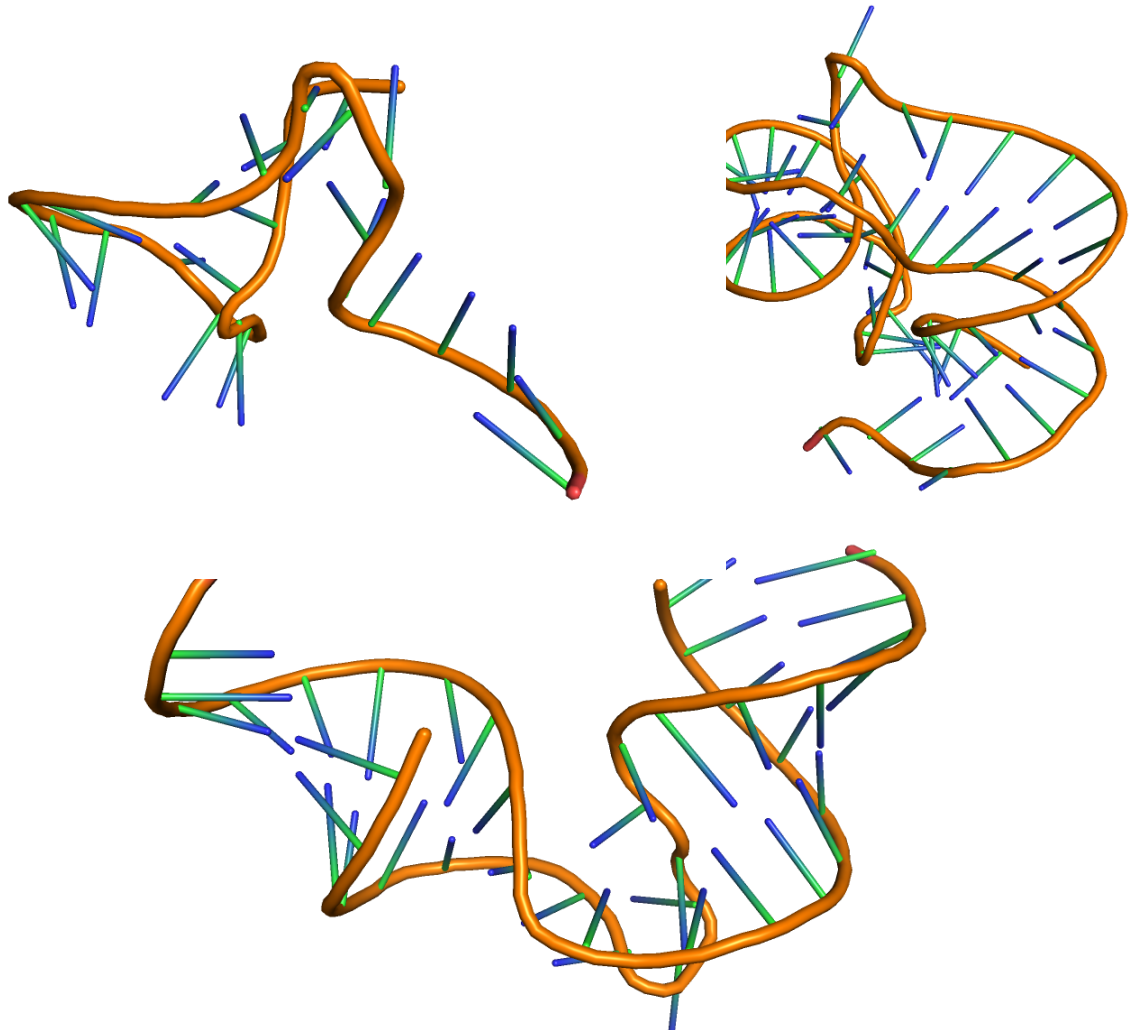


RNA 4V9F

Where is Waldo?

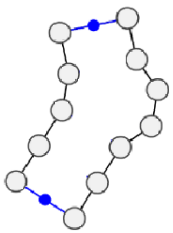


Motif

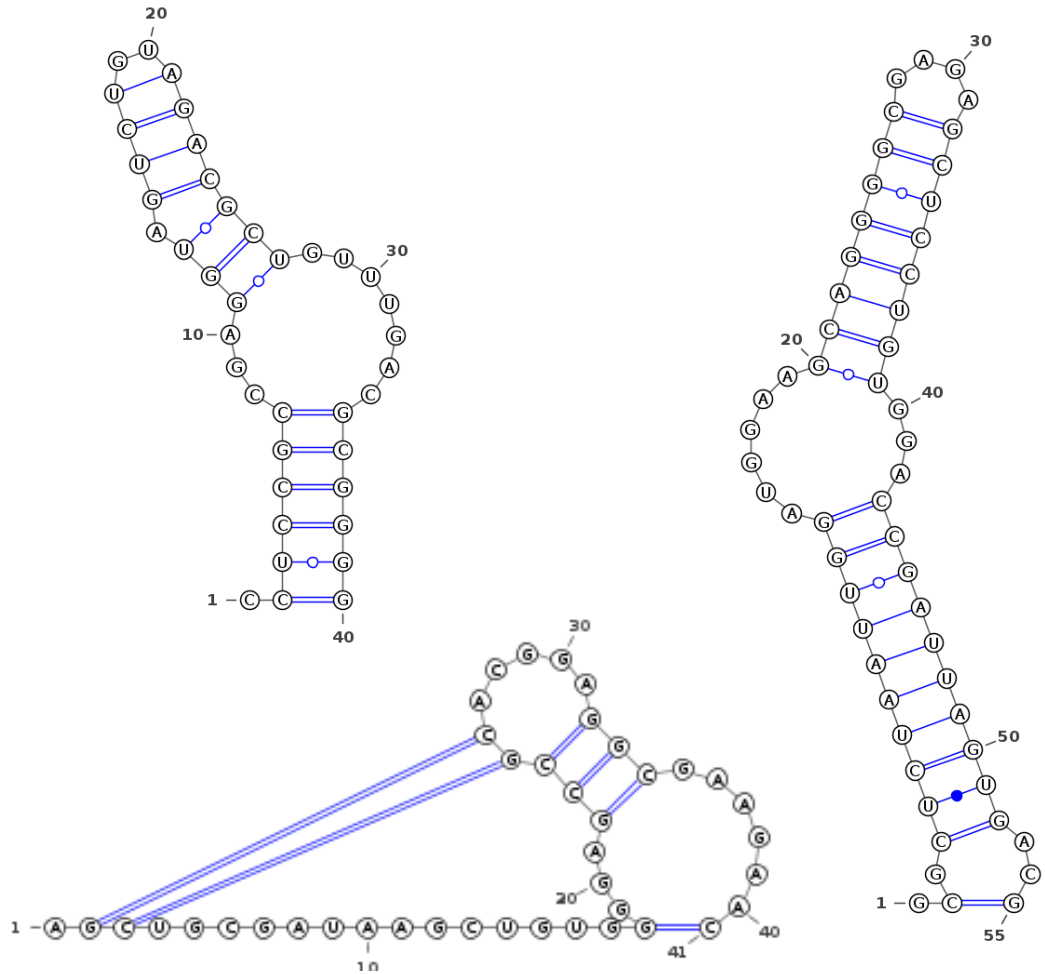


Subparts of RNA 4V9F

Where is Waldo?

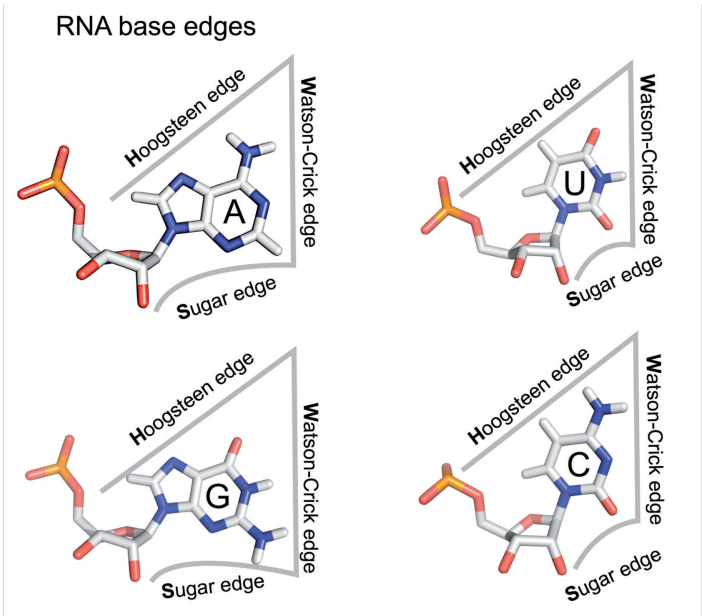


Motif



Subparts of RNA 4V9F

Non canonical annotations (Leontis-Westhof) to the rescue!



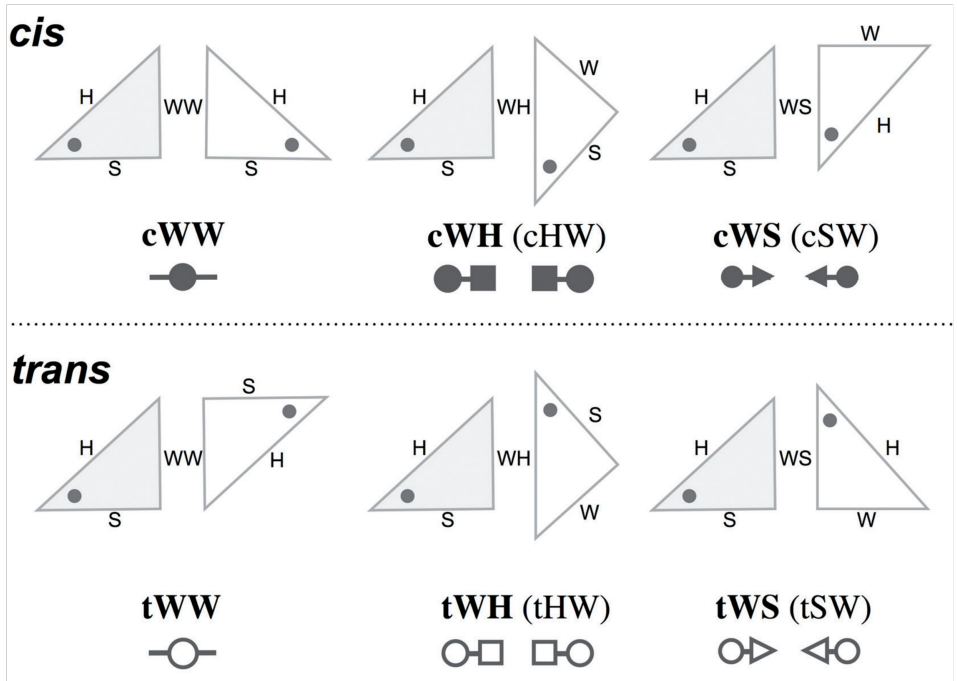
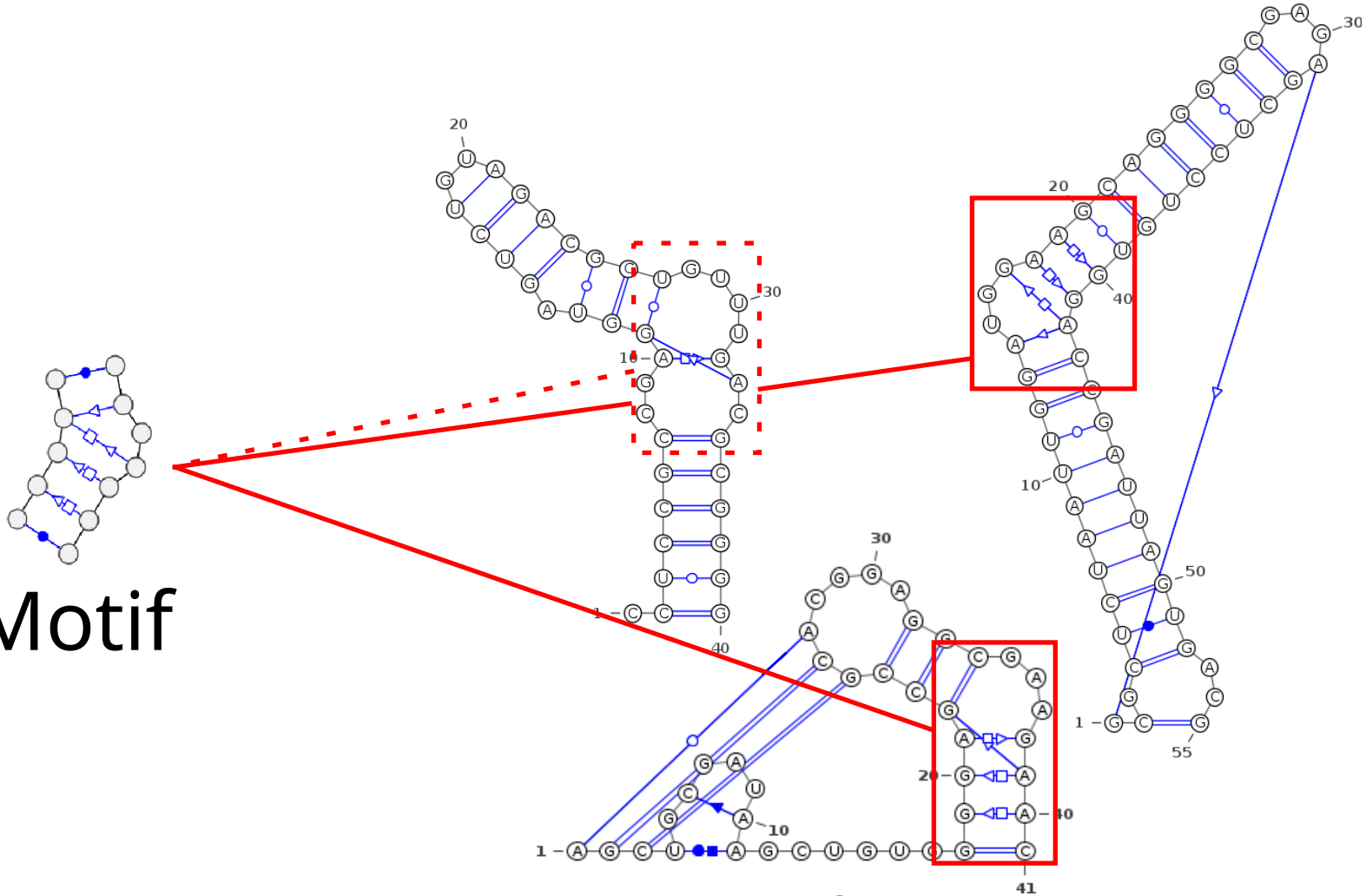
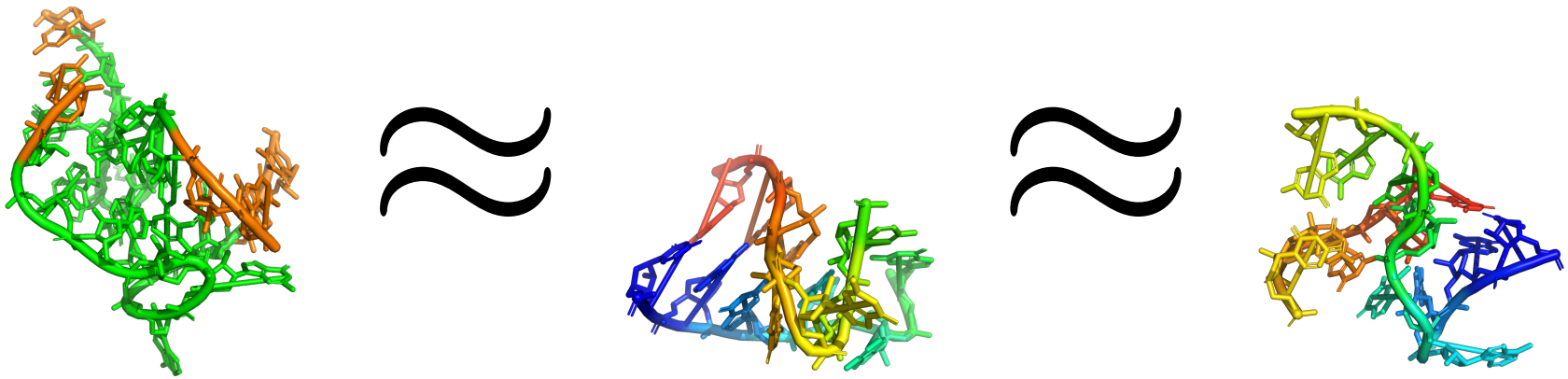


Figure adapted from Almakarem et al, 2011

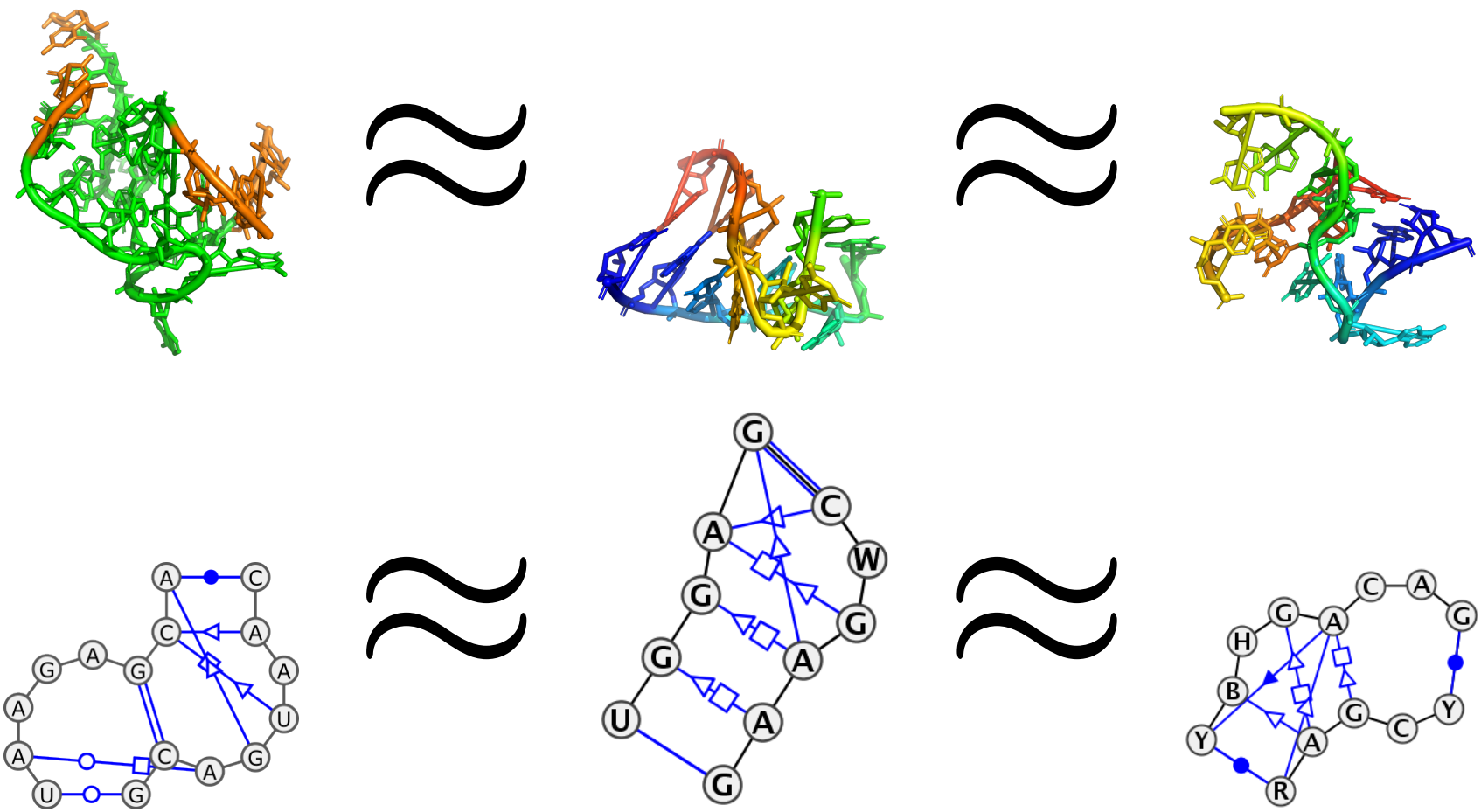
Where is Waldo (again)?



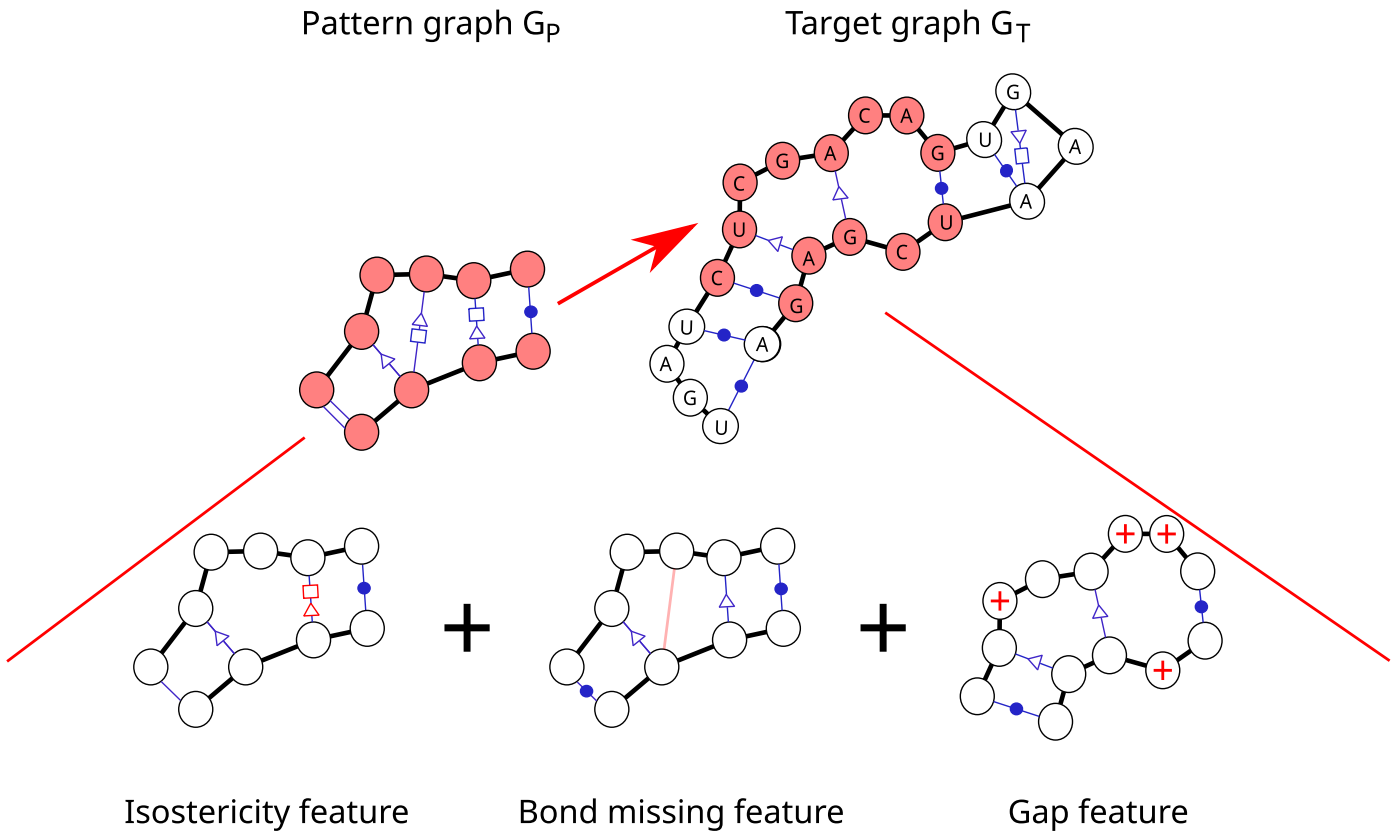
3D homology...



...Is not always obvious even with the non-canonical structure



Achieved result



New method : FuzzTree

- Sample RNA subgraphs in a neighborhood of G_P.
- Used neighborhoods: isostericity, missing bonds and gaps.
- Complexity: XP in G_P treewidth.
- Other state-of-the-art methods: only exact matches.

Outline

Problem formalism

Input: Pattern graph $G_P = \left(V_P, E_P = B_P \sqcup \overline{B}_P\right)$ ($\prec$ −Hamiltonian) , target graph $G_T = \left(V_T, E_T = B_T \sqcup \overline{B}_T\right)$ and neighborhood thresholds (T^L, T^E, T^G, D_{edge}, D_{gap})

Output: Mapping $M : V_P \to V_T$ such that:

1. $\forall (u, v) \in {V_P}^2, u \prec v \Rightarrow M(u) \prec M(v)$ **(monotonicity)**
2. $\sum_{(u,v)\in\overline{B}_P} \text{ISO}\left(L\left(u, v\right), L\left(M(u), M(v)\right)\right) \leqslant T^L$ **(label compatibility)**
3. $\sum_{(u,v)\in\overline{B}_P} 1 - \mathbb{1}_{(M(u),M(v))\in\overline{B}_T} \leqslant T^E$ **(few missing edges)**
4. $\forall (u, v) \in \overline{B}_P, (M(u), M(v)) \notin \overline{B}_T, \text{GEO}\left(M(u), M(v)\right) \leqslant D_{\text{edge}}$ **(edge distance limit)**
5. $\sum_{(p_0,\ldots,p_k)\in P, k\geqslant 3} \text{GEO}(p_0, p_k) \leqslant T^G$ **(path size limitation)**
6. $\forall (u, v) \in B_P, \exists (p_0, p_1, p_2, ..., p_k) \in P$ such that **(no missing backbone path)**
   - $p_0 = M\left(u\right), p_k = M\left(v\right)$ **(*)**
   - $\text{GEO}\left(p_0, p_k\right) \leqslant D_{\text{gap}}$ **(**)**

or $\varnothing$ if no such mapping exists.

Corresponding NP-complete Problem

Our problem specializes in Hamiltonian Subgraph Isomorphism Problem, known to be NP-complete:

Input: Pattern graph ($\prec$ −Hamiltonian) $G_P = (V_P, E_P)$; Target graph $G_T = (V_T, E_T)$

Output: Mapping $M : V_P \to V_T$ such that

- $\forall (u, v) \in {V_P}^2, u \prec v \Rightarrow M(u) \prec M(v)$ **(monotonicity)**
- $\forall (u, v) \in E_P, (M(u), M(v)) \in E_T \Rightarrow L((u, v)) = L((M(u), M(v)))$ **(label comp.)**
- $\forall (u, v) \in E_P, (M(u), M(v)) \in E_T$ **(no missing edge)**

or $\varnothing$ if no such mapping exists.

State of the art methods for Subgraph Isomorphism Problem

1976 Exact method: Ullmann's method

1995 Exact method: Color-Coding

2004 Heuristic: VF2

2018 Heuristic improvements: VF2++ and VF3

2021 Fuzzy method: VeRNAI

Parametrized complexity: Complexity uses a parameter p that depicts "property" of the input. Two classes of complexity:

Fixed-Parameter Tractable FPT: $O\left(f(p)n^{O(1)}\right)$

Slicewise polynomial XP: $O\left(f(p)n^{g(p)}\right)$

A sampling method: why and how?

Why sampling?

- Serve to obtain only a subset of all mappings in the RNA.
- To filter automatically mappings that are "too far" from G_P.

How to sample?

Definition 2.1 (Multidimensional Boltzmann distribution): Given a motif G_P, the probability $\mathbb{P}$ to sample a (mapped) graph M in an RNA depends on its (pseudo-)**energy** E:

$$\mathbb{P}(M) = \frac{e^{-E(M)}}{\mathcal{Z}} \text{ where } \mathcal{Z} = \sum_{M'} e^{-E(M')} \quad (1)$$

Accounting the features to get a subset of solutions

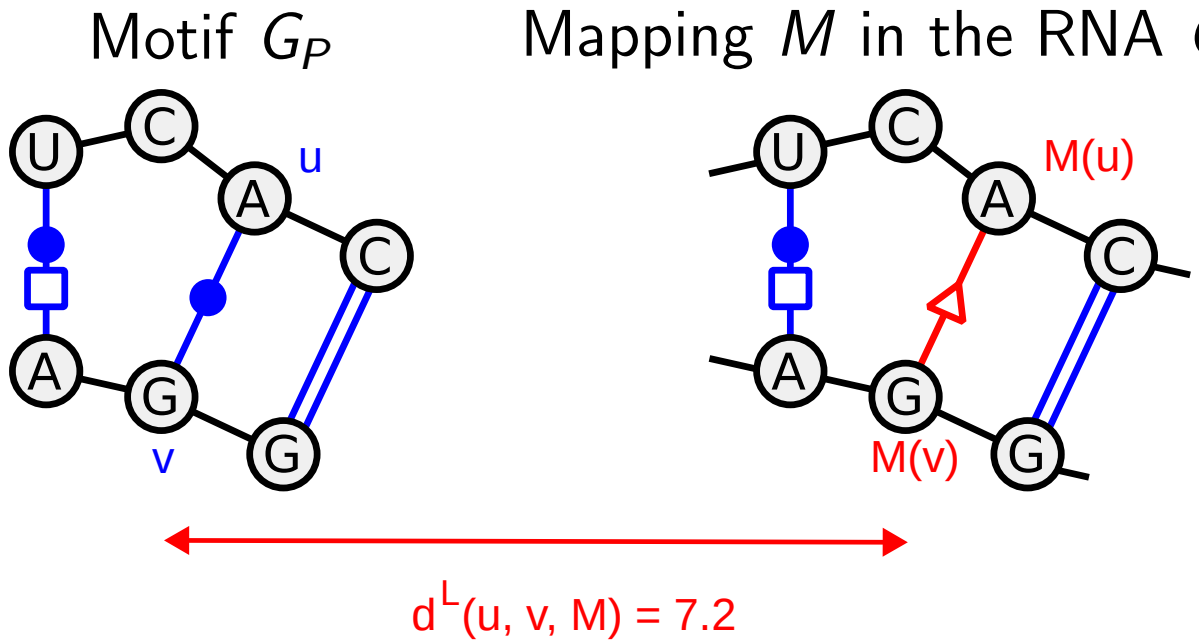
Features of a mapping M are taken into account through an additive (pseudo-)energy function:

$$E(M) = \sum_{(u,v)\in E_P} w_L \times d^L(u,v,M) + w_E \times d^E(u,v,M) + w_G \times d^G(u,v,M)$$

Where w_L, w_E, w_G are real positive valued weights.

- Exact mapping corresponds to $E(M) = 0$.
- $E(M) \neq 0$ and $E(M) << \infty$ corresponds to fuzzy matches.

The label compatibility feature d^L



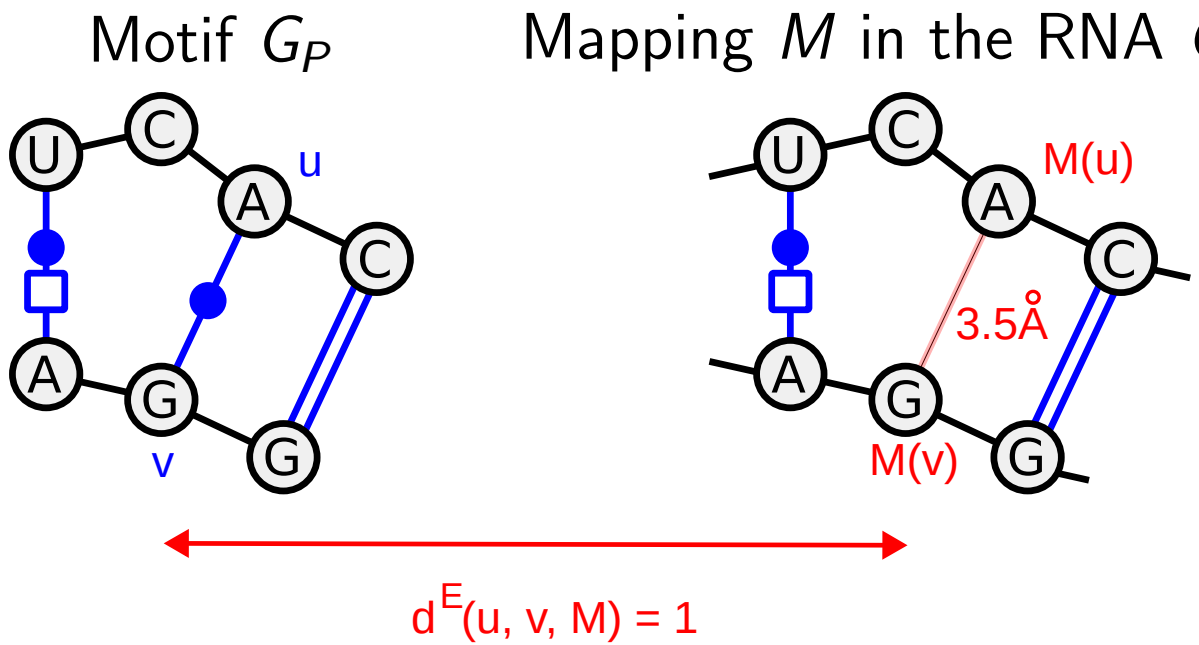
Definition 2.2 (Label compatibility feature d^L):

$$d^L(u, v, M) = ISO(\text{Label}(u, v), \text{Label}(M(u), M(v)))$$

- Isostericity ISO[1] compares both the 12 canonical and non-canonical base pairing families.

[1] Stombaugh et al, 2009, Nucleic Acids Research

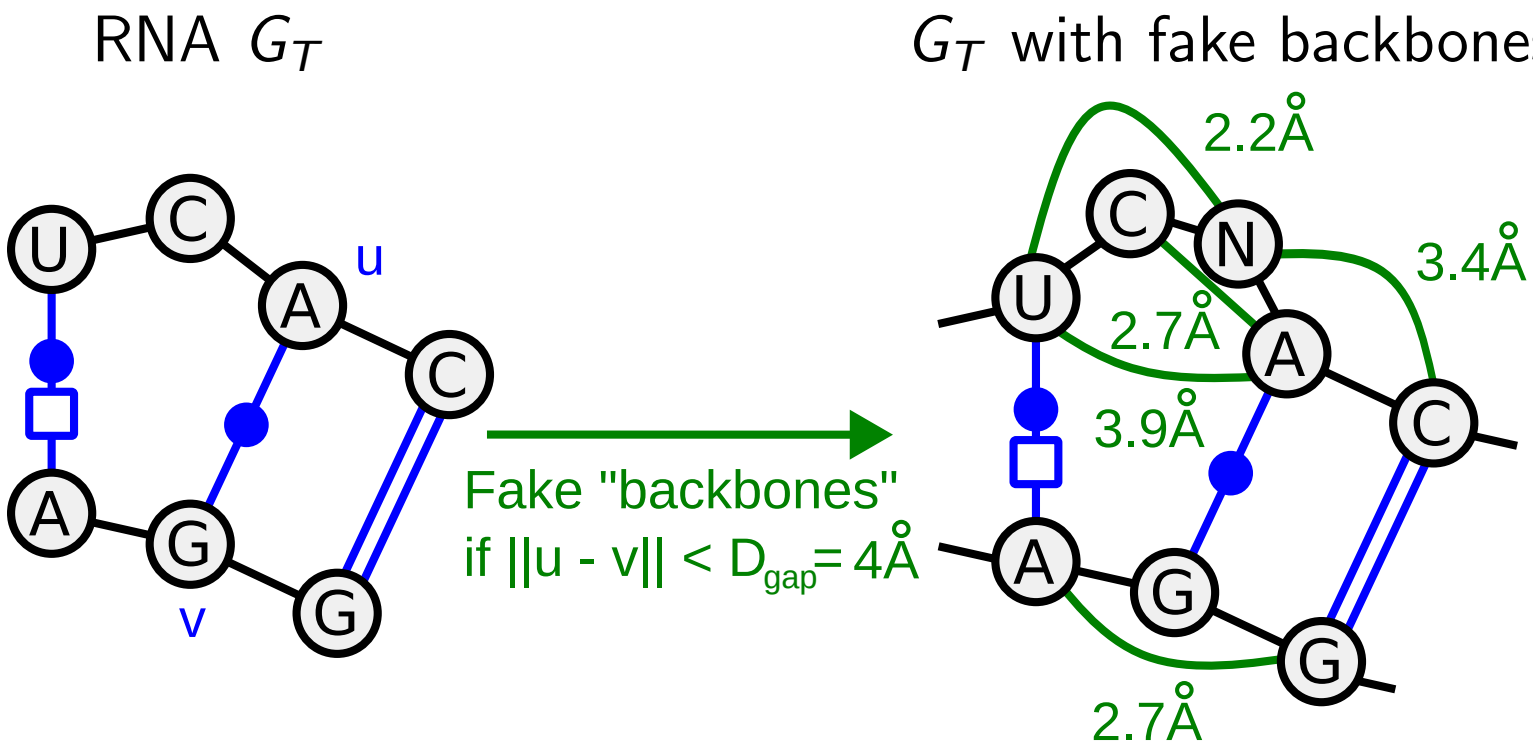
The bond missing feature d^E



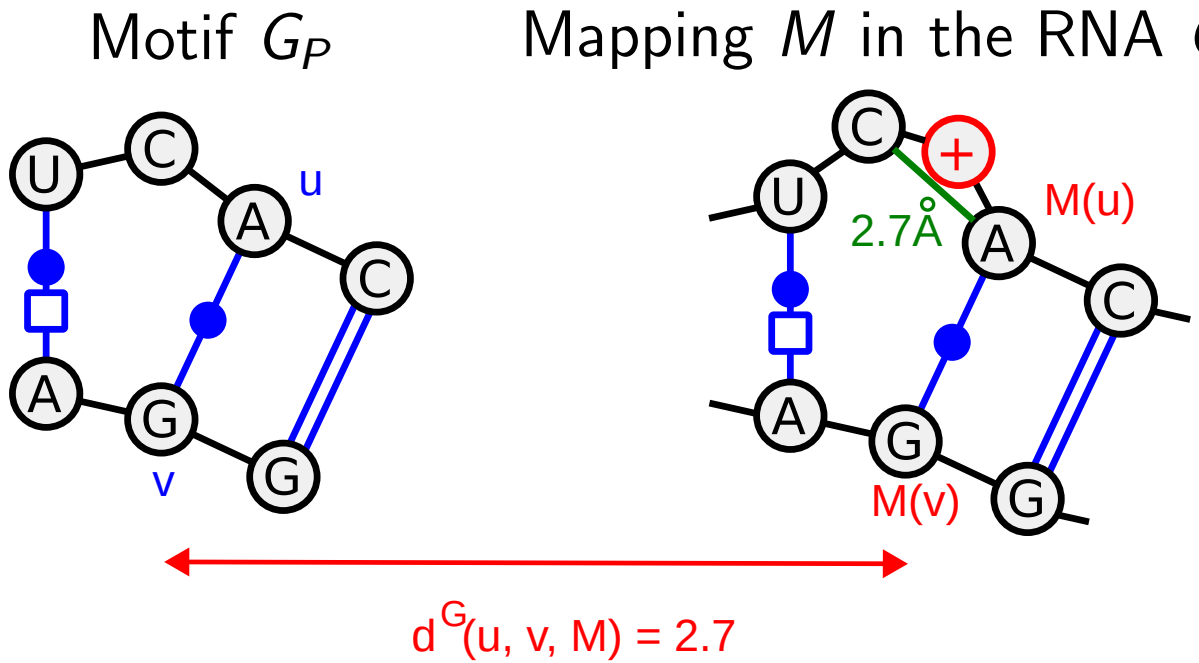
Definition 2.3 (Bond missing missing feature d^E):

$$d^E(u, v, M) = \begin{cases} 0 & \text{if } (u, v) \in B_P \cap (M(u), M(v)) \in B_T \\ & \text{or } (u, v) \in \overline{B}_P \cap (M(u), M(v)) \in \overline{B}_T \\ 1 & \text{if } (u, v) \in \overline{B}_P \cap (M(u), M(v)) \notin \overline{B}_T \\ & \text{and } \mathrm{GEO}(M(u), M(v)) \leqslant D_{\text{edge}} \\ \infty & \text{otherwise} \end{cases}$$

Fake backbones creation for gaps



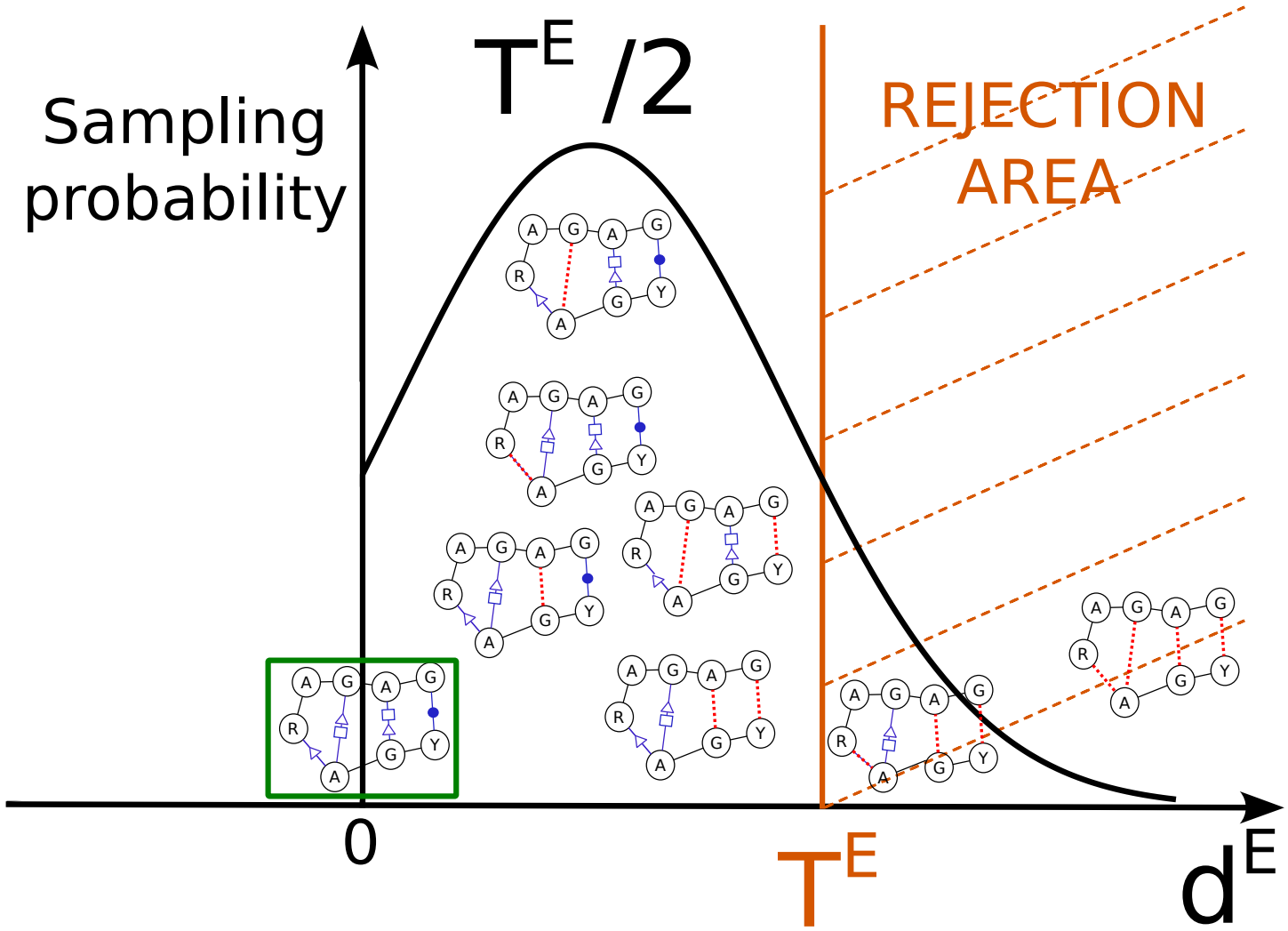
Gap feature d^G



Definition 2.4 (Gap difference d^G):

$$d^G(u, v, M) = \begin{cases} \text{GEO}(M(u), M(v)) & \text{if } (M(u), M(v)) \text{ is a "Fake Edge" in } E_T \\ 0 & \text{otherwise} \end{cases}$$

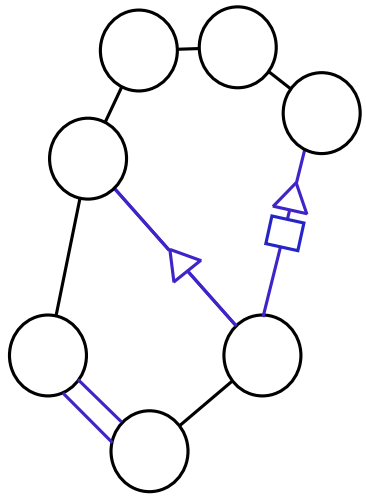
Sampling into a Boltzmann distribution

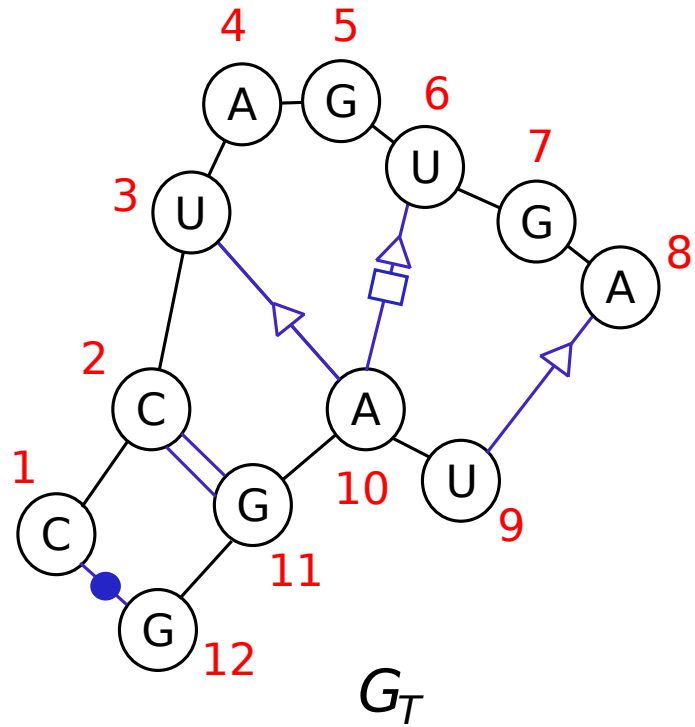


- We sample (given a pseudo-energy) instead of simply searching an optimal.

Decompose our instance into tree

Origin G_P



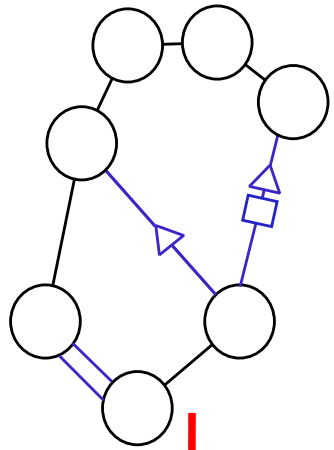


G_T

— *Backbone from 5' to 3'*
— *Labeled base pair*

Decompose our instance into tree

Origin G_P

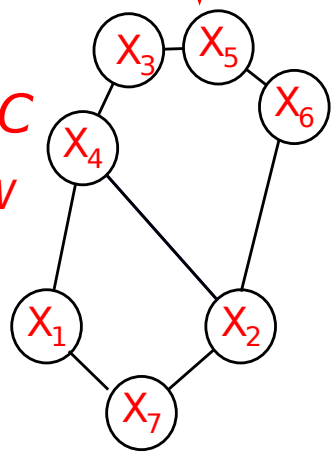


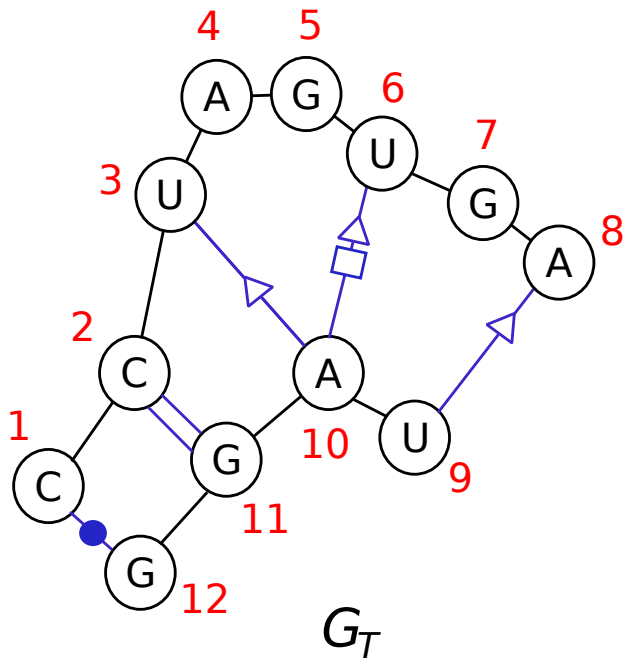
Constraints graph C

label(X_1, X_7) = cWW

label(X_2, X_4) = tSS

label(X_2, X_6) = tHS





G_T

— *Backbone from 5' to 3'*

— *Labeled base pair*

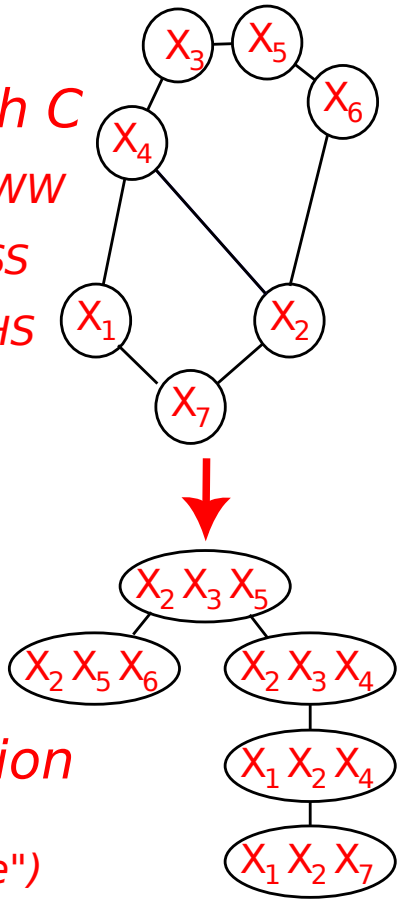
Decompose our instance into tree

Constraints graph C

label(X_1, X_7) = cWW

label(X_2, X_4) = tSS

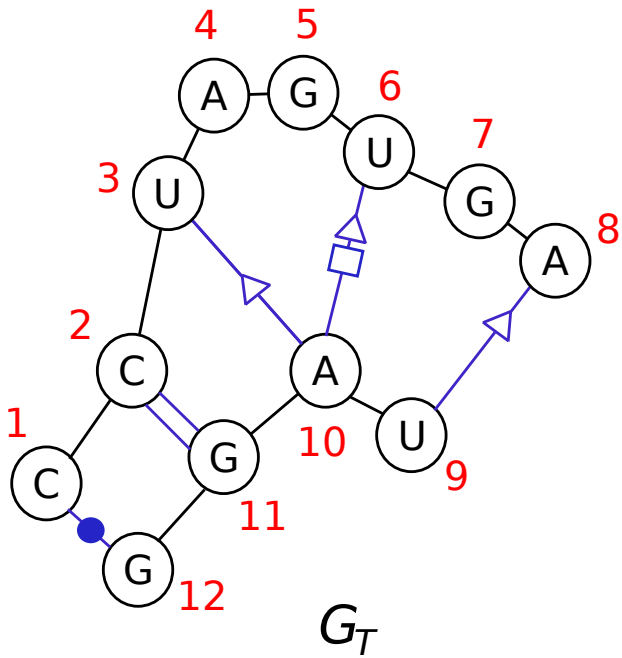
label(X_2, X_6) = tHS



Tree Decomposition

T(C)

(In reality, T(C) is``nice")



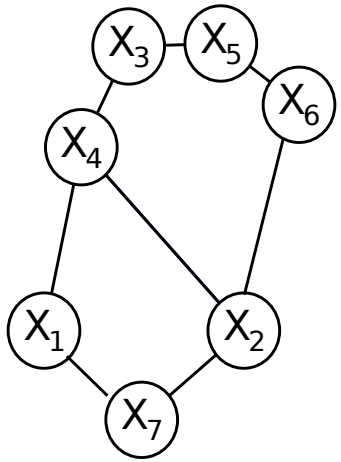
G_T

— *Backbone from 5' to 3'*

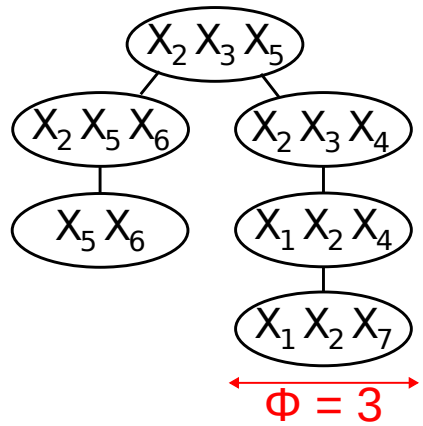
— *Labeled base pair*

Tree decomposition and tree width

Graph C



Tree Decomposition T



Given a graph $C = (V, E)$, a tree decomposition of G is a tree T composed of bags $B_1 ... B_t$ such as: [2]

1. $V \subset \bigcup_{i=1}^{t} B_i$
2. $\forall (X_i, X_j) \in E, \exists i \in [\![1, t]\!], (X_i \in B_k) \cap (X_j \in B_k)$
3. $\forall X_i \in V, \{B_k \mid X_i \in B_k\}$ is a subtree of T.

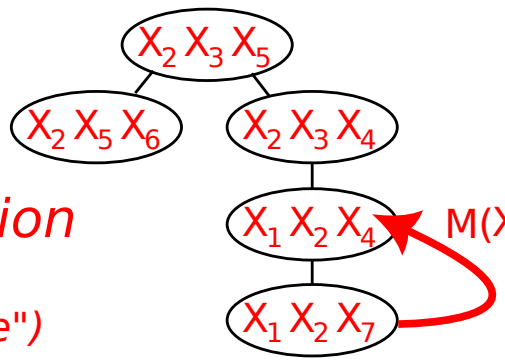
[2] Bodlaender et al, 2008

Decompose our instance into tree

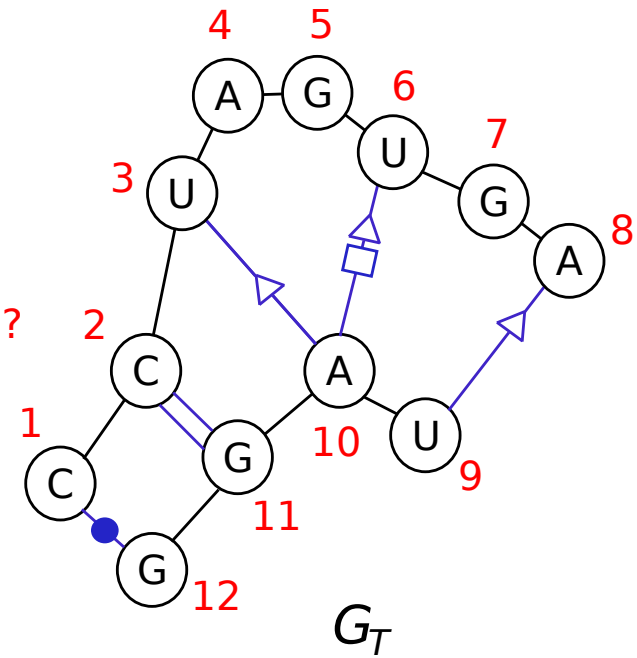
Tree Decomposition

T(C)

(In reality, T(C) is``nice")



$M(X_1) = 2$
$M(X_2) = 10$
$M(X_7) = 11$



— *Backbone from 5' to 3'*
— *Labeled base pair*

Dynamic programming on tree decomp. is automatized with Infrared [3]

[3] Hua-Ting et al, 2022, RNA Folding - Methods and Protocols

Complexity

- Complexity:
  - Partition function computation: $O\left(kn^{\phi+1}\right)$
  - Sampling: $O\left(knt\right)$

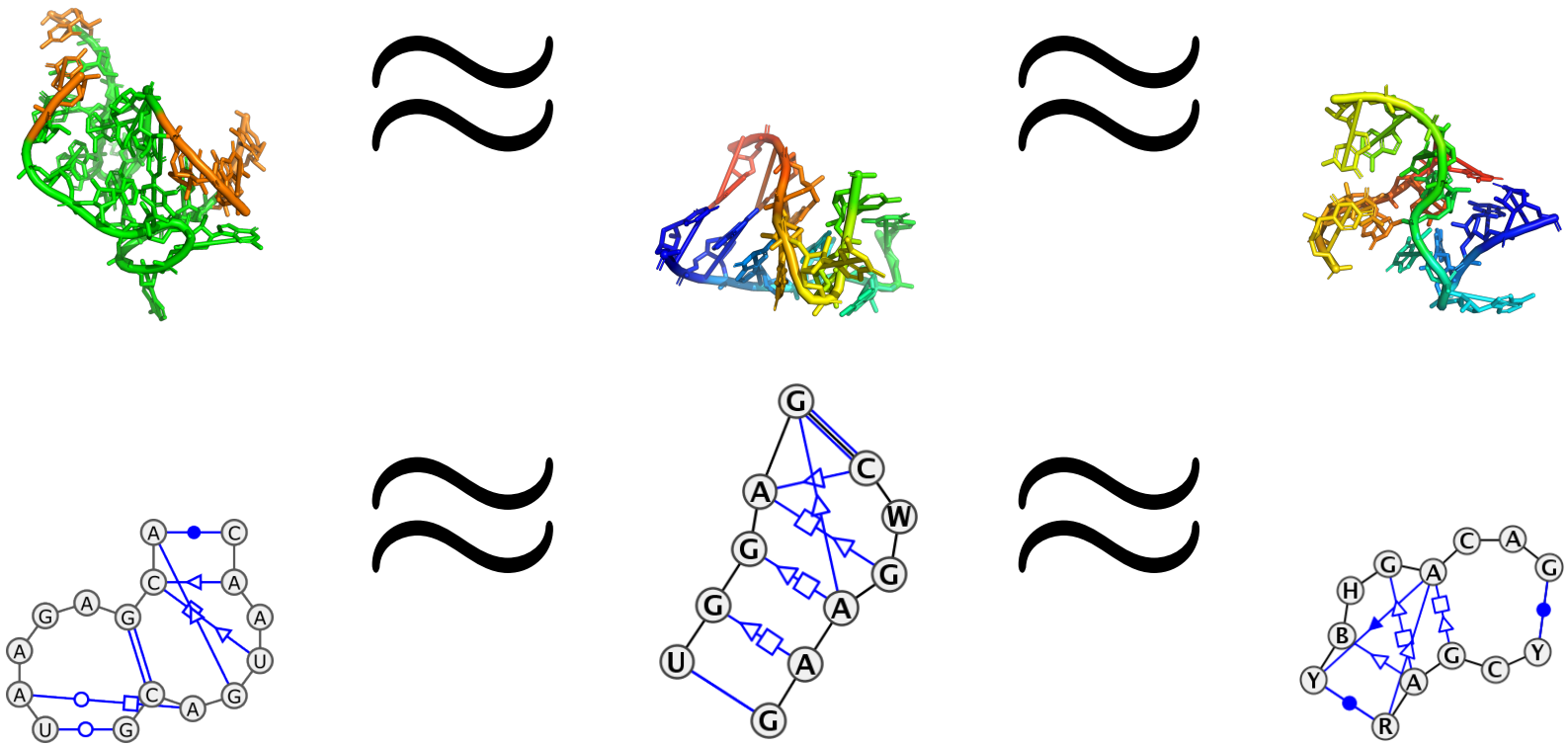
  With:
  - n: number of nodes in the RNA G_T
  - k: number of nodes in the motif G_P
  - ϕ: treewidth of G_P
  - t: number of samples

Outline

Validation dataset

The Kink-Turn family dataset:

- A biological family that contains 72 known motifs over more than 25 different RNAs.
- Kink-Turns are clustered in 18 different families according to atomic cristallography. [4]



[4] Petrov et al, 2013, RNA

Typical treewidth of Kink-Turn motifs

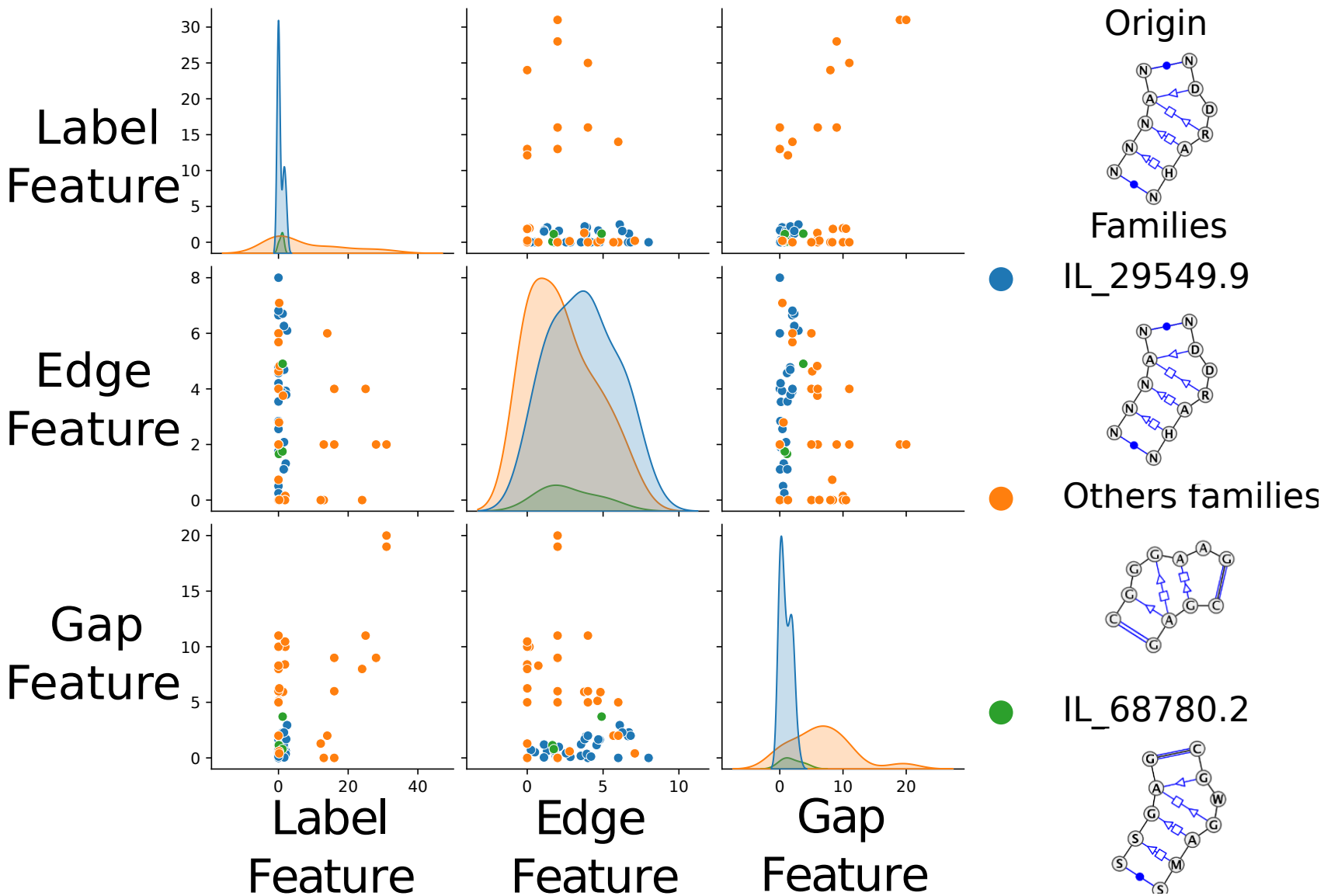
Number of Kink-Turn instances	Treewidth
50	2
21	3

- Complexity:

$$O\left(knt + kn^{\phi+1}\right)$$

- k, number of nodes in G_P.
- n, number of nodes in G_T.
- ϕ, treewidth of G_P.
- t, number of samples.

Kink-Turn Cartography

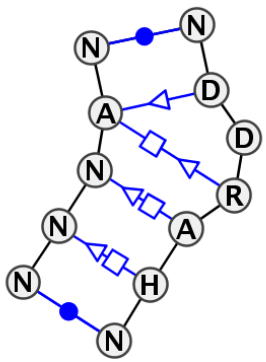


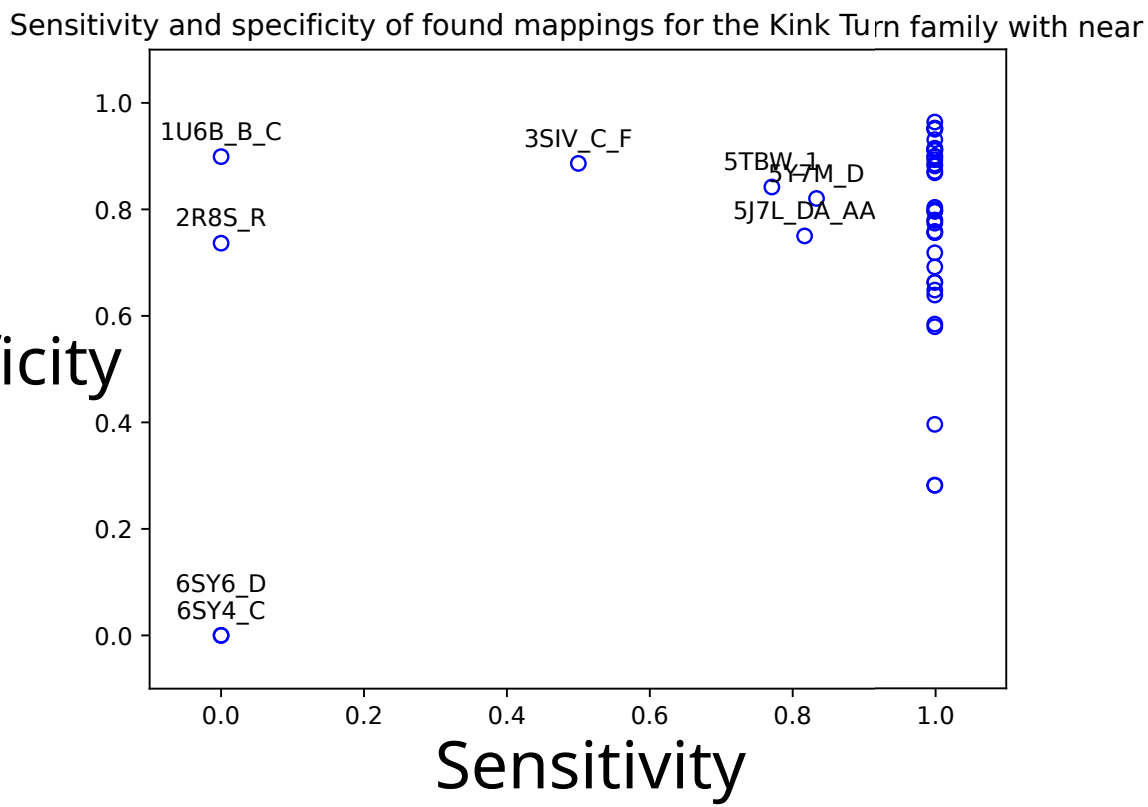
Retrieve the Kink-Turn family by requesting a single motif

We requested motif IL_5TBW_059 inside all RNA containing Kink-Turns.

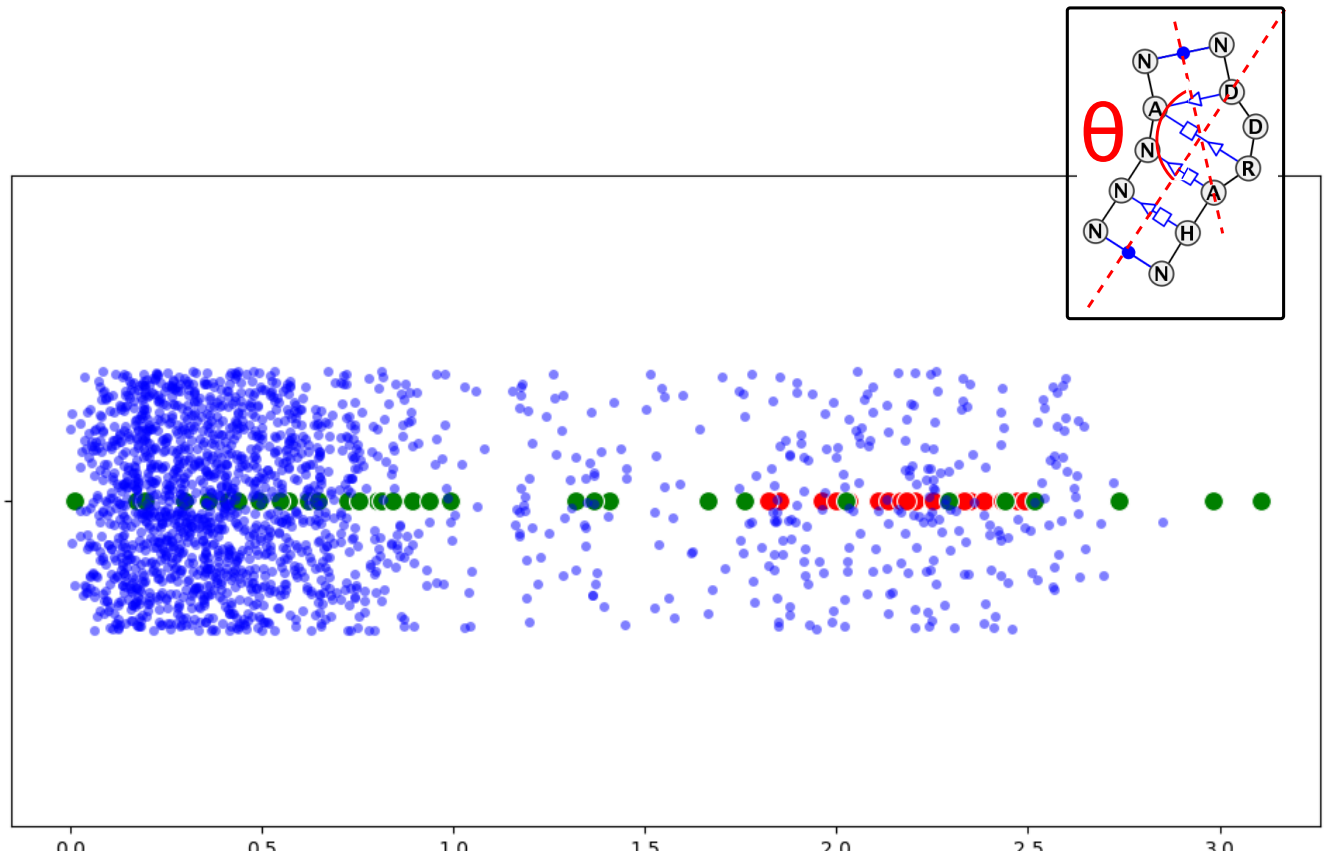
- Thresholds on neighborhoods: $T^L = 20$, $T^E = 4$ and $T^G = 20$.
- Used metrics: Sensitivity $= \frac{TP}{P}$ and Specificity $= \frac{TN}{N}$

Requested motif



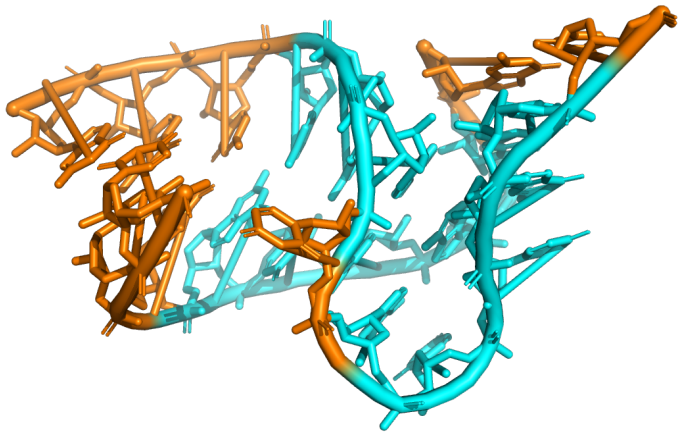


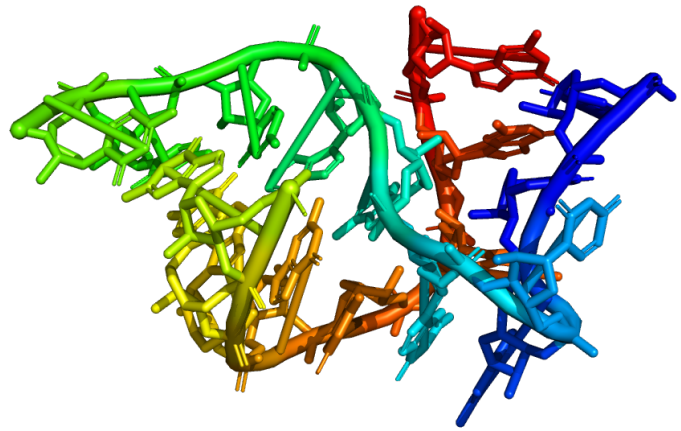
What about found motifs that are not labeled as Kink-Turns ?

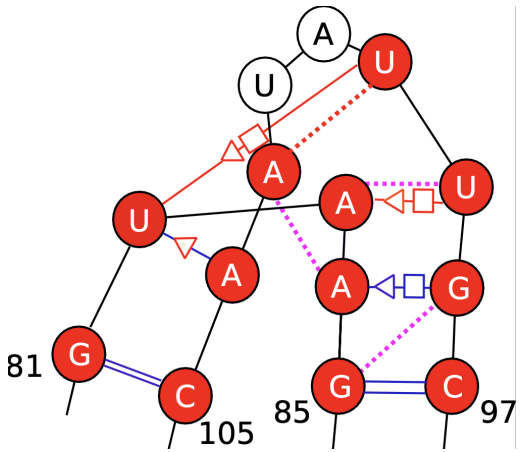


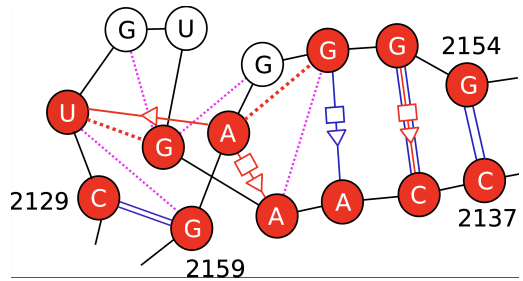
- Some of our motifs angles in green are in the range of the Kink-Turn angles in red.

Suggestions of new motifs using our methods



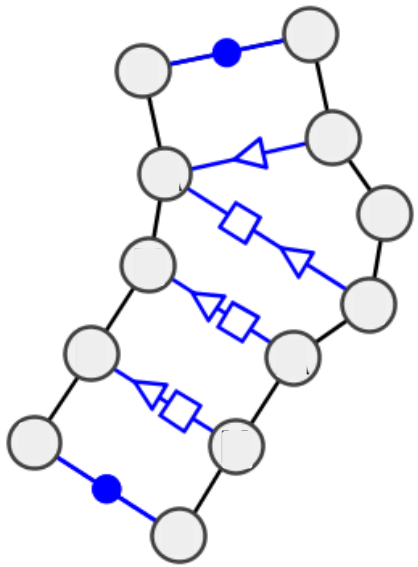




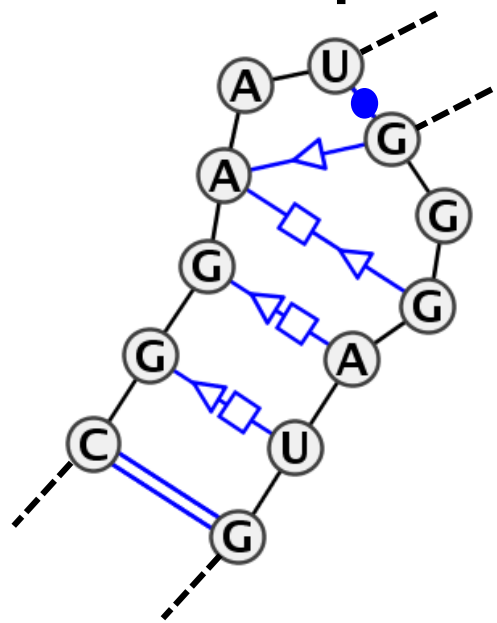


Case study: A Kink-Turn that we missed

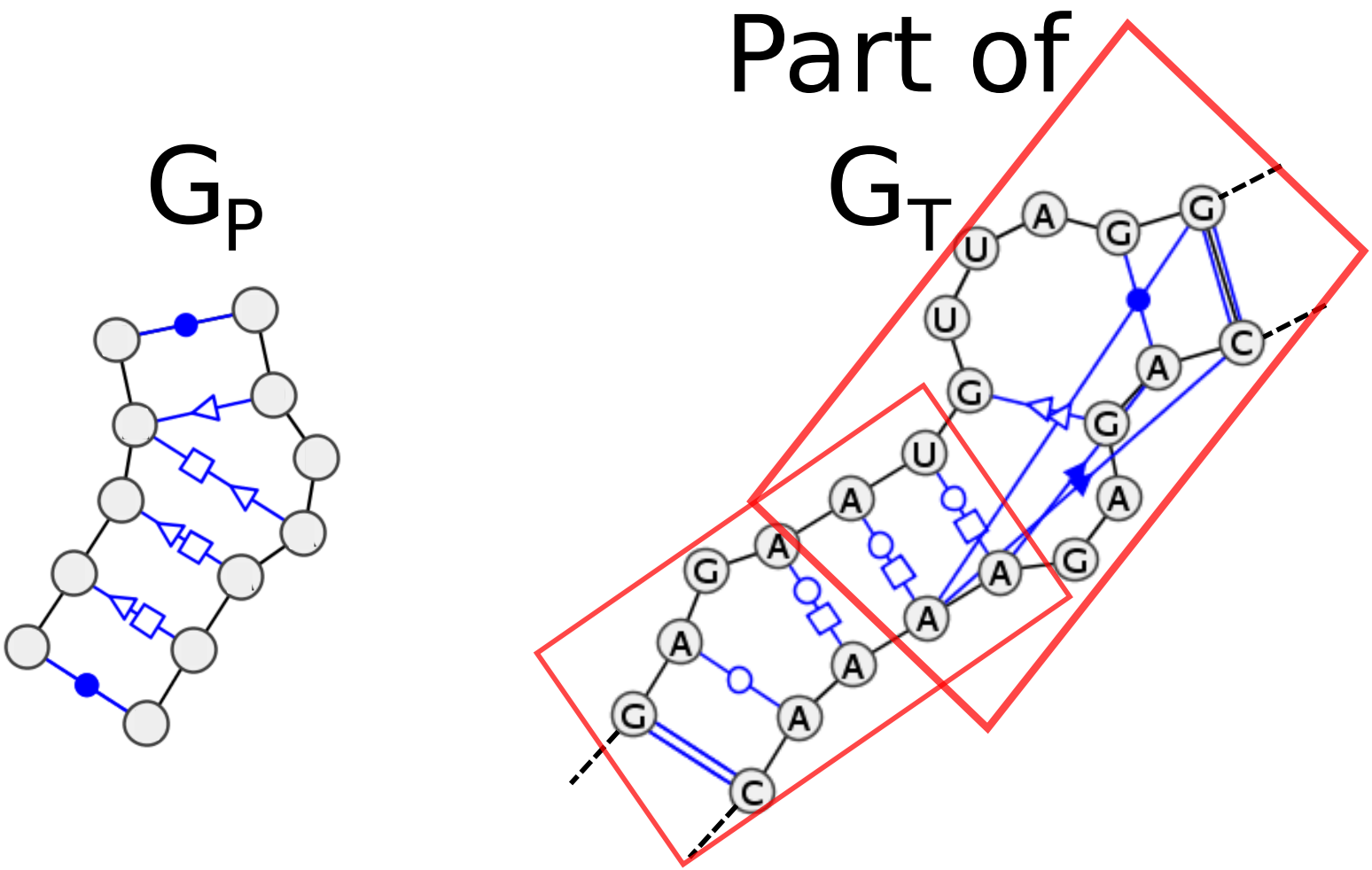
G_P



Part of G_T



Case study: Kink-Turn that we missed but that we can hope to cover with multiple mappings



Conclusion and future work

- We proposed an exact sampling solution for the Fuzzy Monotonous Subgraph Isomorphism Problem.
- Complexity is rooted in the treewidth of the requested motif:

$$O\left(knt + kn^{\phi+1}\right)$$

- We used isostericity, missing bonds and gaps to catch a wide variety of RNA motifs as observed on the Kink-Turn.

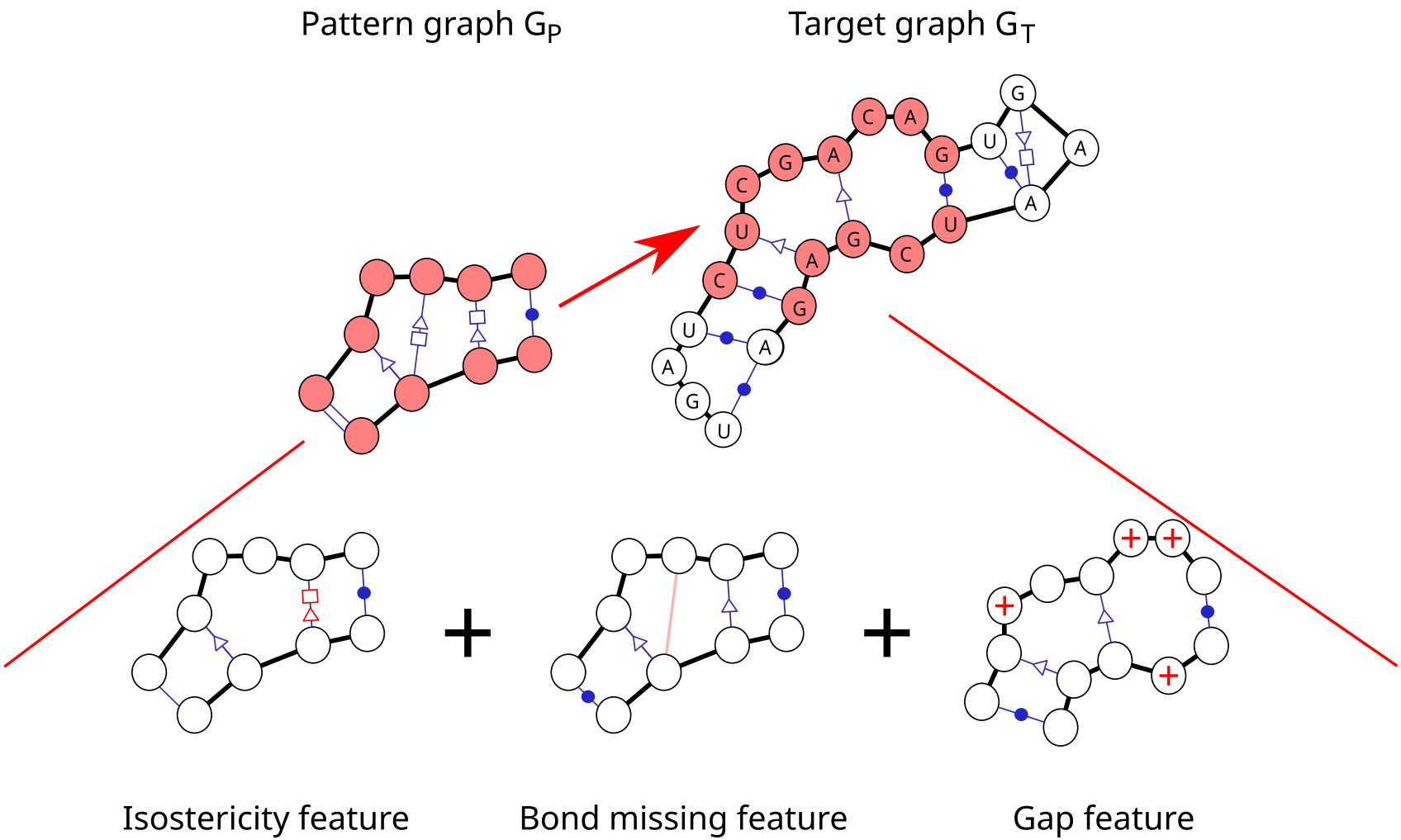
Future work:

- Further evaluate the efficiency of FuzzTree on diverse RNA modules
- Possibility to introduce new metrics without additional work
- Discover unknown RNA motifs unlisted until now thanks to our neighborhoods

Acknowledgements

You!

FuzzTree


Pattern graph G_P
Target graph G_T
Isostericity feature
Bond missing feature
Gap feature

Isostericity computation

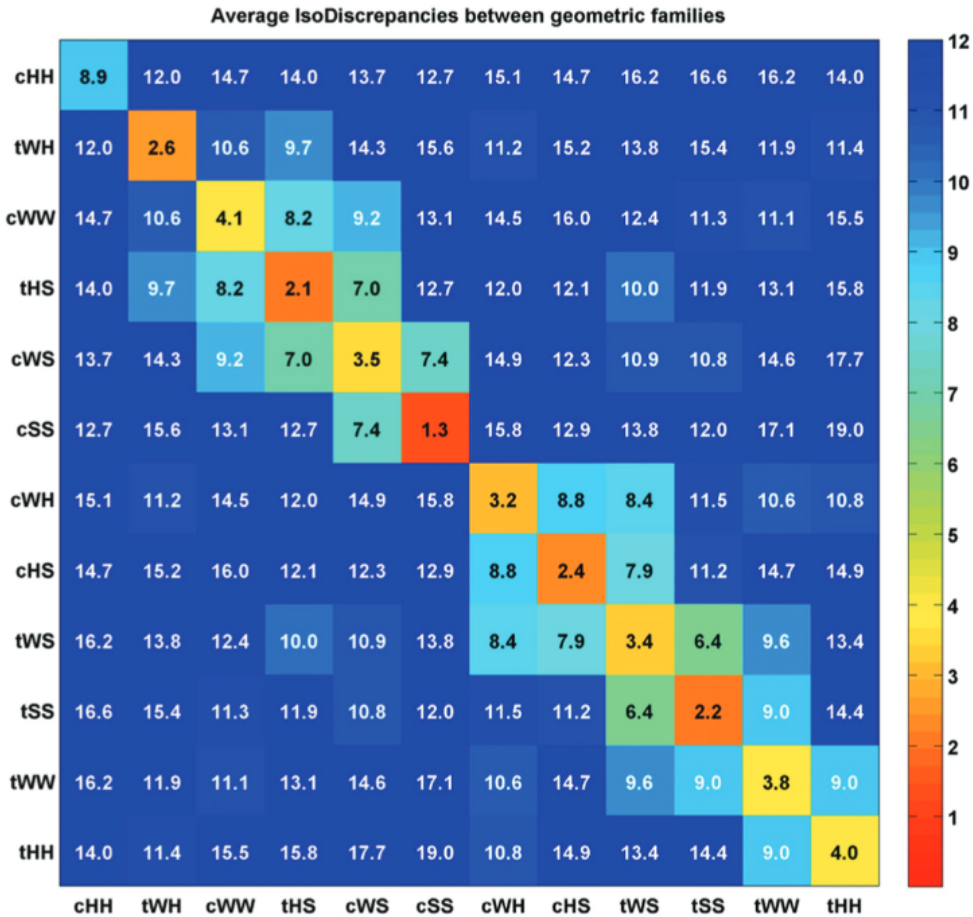


Figure adapted from
Zirben al, 2009