

Empirical performances, ML/AI an advanced dynamic programming

Yann Ponty

AMIBio Team
École Polytechnique/CNRS

Historical paradigms towards 2D prediction

Definition (Ab initio folding)

Starting from sequence, find conformation that minimizes free-energy.

Advantages:

- ▶ Mechanical nature allows the (in)validation of models
- ▶ Reasonable complexity
 $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/space
- ▶ *Exhaustive* nature

Limitations:

- ▶ Hard to include PKs
- ▶ Highly dependent on energy model
- ▶ No cooperativity
- ▶ Limited performances

Definition (Comparative approach)

Starting from homologous sequences, postulate common structure and find best possible tradeoff between folding & alignment.

Avantages :

- ▶ Better performances
- ▶ (Limited) cooperativity
- ▶ Self-improving

Limitations

- ▶ Easily unreasonable complexity
- ▶ Non exhaustive search
- ▶ Captures *transient* structures

Historical paradigms towards 2D prediction

Definition (Ab initio folding)

Starting from sequence, find conformation that minimizes free-energy.

Advantages:

- ▶ Mechanical nature allows the (in)validation of models
- ▶ Reasonable complexity
 $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/space
- ▶ *Exhaustive* nature

Limitations:

- ▶ Hard to include PKs
- ▶ Highly dependent on energy model
- ▶ No cooperativity
- ▶ Limited performances

Definition (Comparative approach)

Starting from homologous sequences, postulate common structure and find best possible tradeoff between folding & alignment.

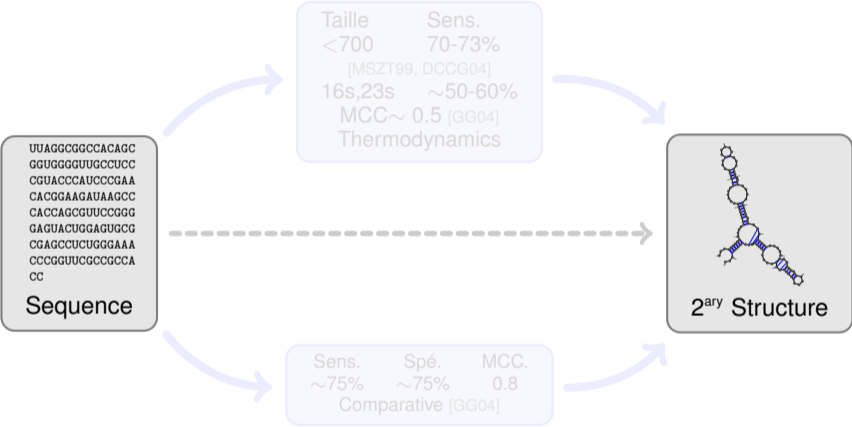
Avantages :

- ▶ Better performances
- ▶ (Limited) cooperativity
- ▶ Self-improving

Limitations

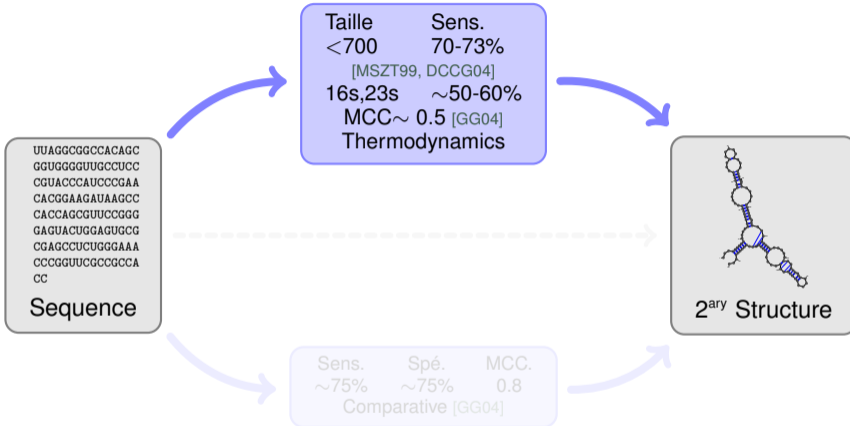
- ▶ Easily unreasonable complexity
- ▶ Non exhaustive search
- ▶ Captures *transient* structures

Typical performances



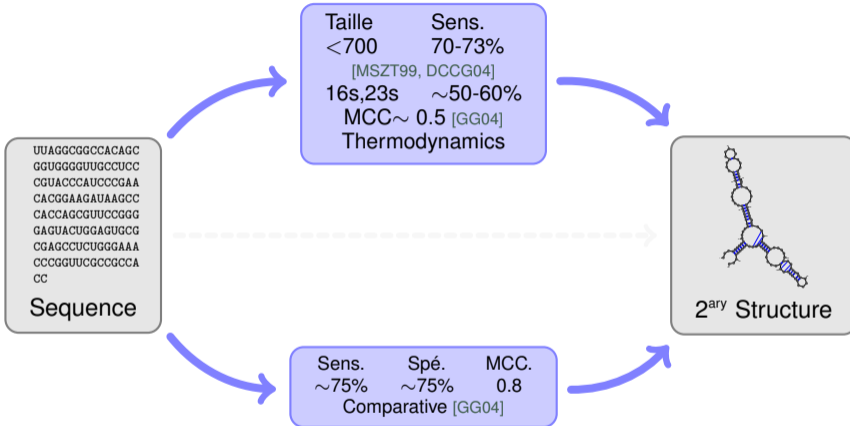
Reminder:
$$MCC = \frac{t^+t^- - f^+f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$$

Typical performances



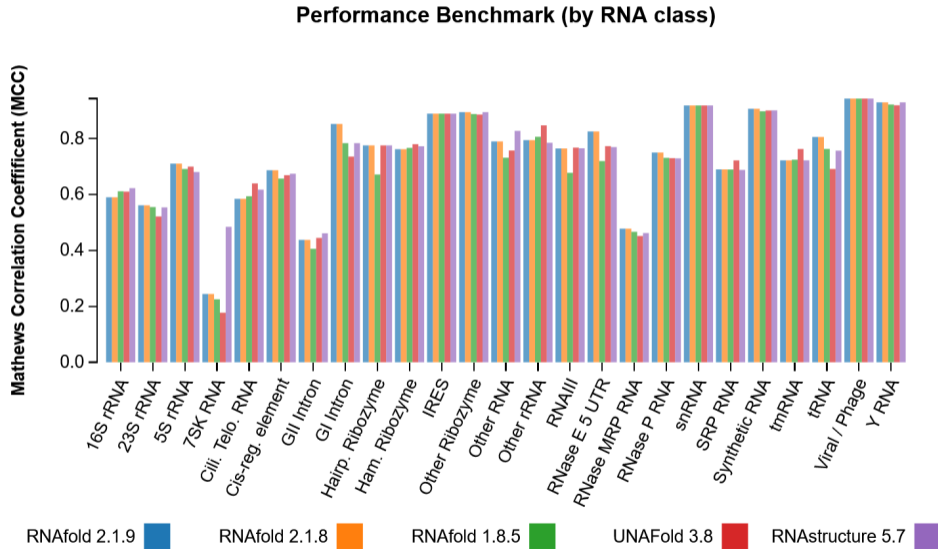
Reminder:
$$MCC = \frac{t^+t^- - f^+f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$$

Typical performances



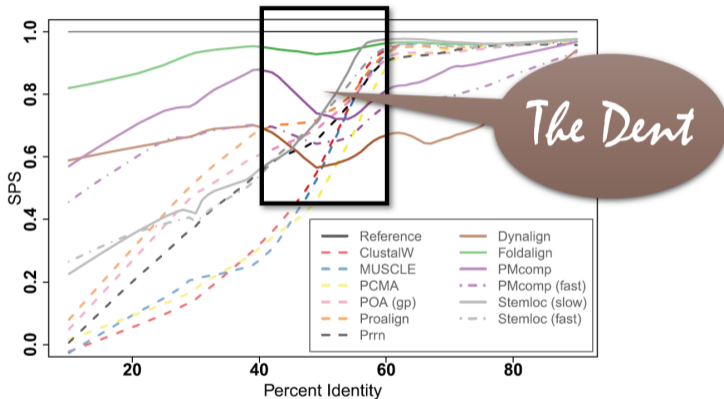
Reminder:
$$MCC = \frac{t^+t^- - f^+f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$$

Detailed performances of 2D folding algorithms



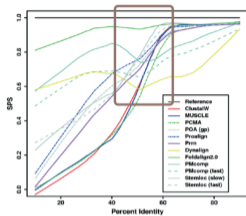
Biased benchmarks: precedent in comparative folding/alignment

Bralibase: Benchmark for comp. RNA folding [Gardner,Wilm & Washietl, NAR 2005]

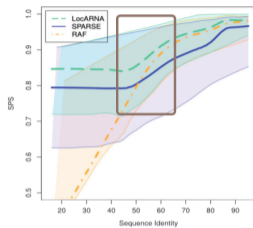


[Löwes *et al*, Briefings in Bioinfo 2016]

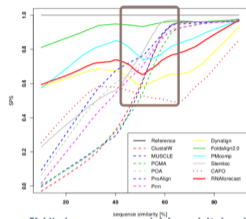
Biased benchmarks: precedent in comparative folding/alignment



[Gardner *et al*, NAR 2005]



[Will *et al*, Bioinformatics 2015]



[Höschmann *et al*, Unpublished]

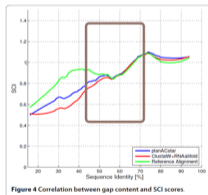
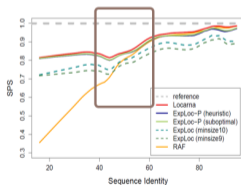
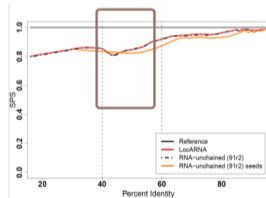


Figure 4 Correlation between gap content and SCI scores.

[Bremges *et al*, BMC Bioinfo, 2010]



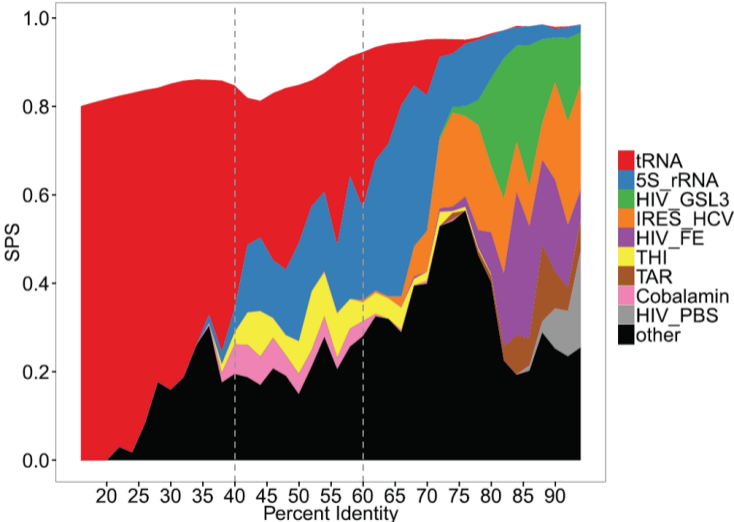
[Schmiedl *et al*, RECOMB 2012]



[Bourgeade *et al*,] Comp Biol, 2015]

[Löwes *et al*, Briefings in Bioinfo 2016]

Biased benchmarks: precedent in comparative folding/alignment

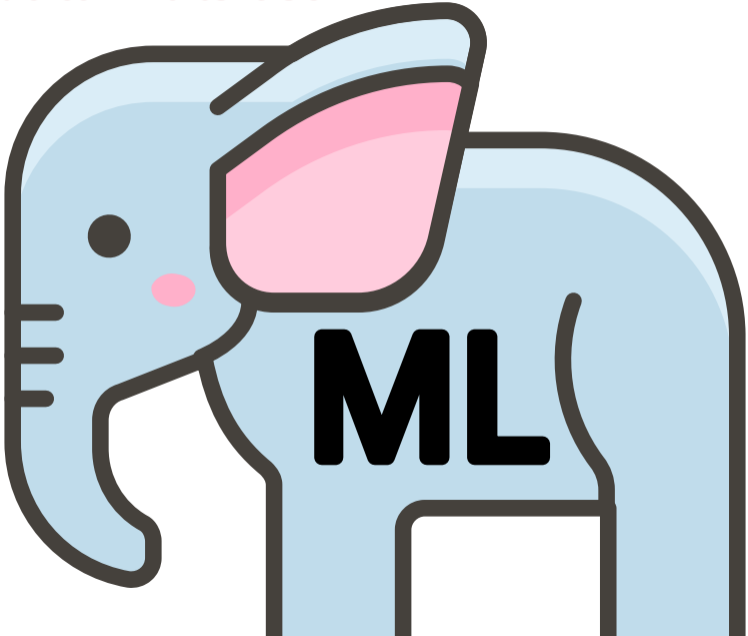


[Löwes *et al*, Briefings in Bioinfo 2016]

The elephant in the room – **2010s version**



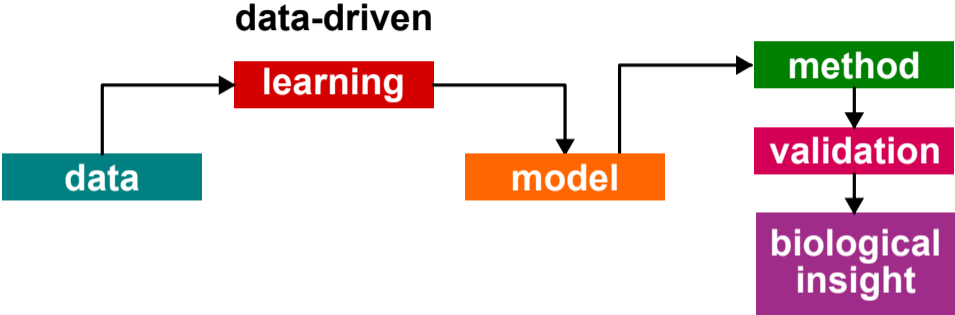
The elephant in the room – **2020s version**



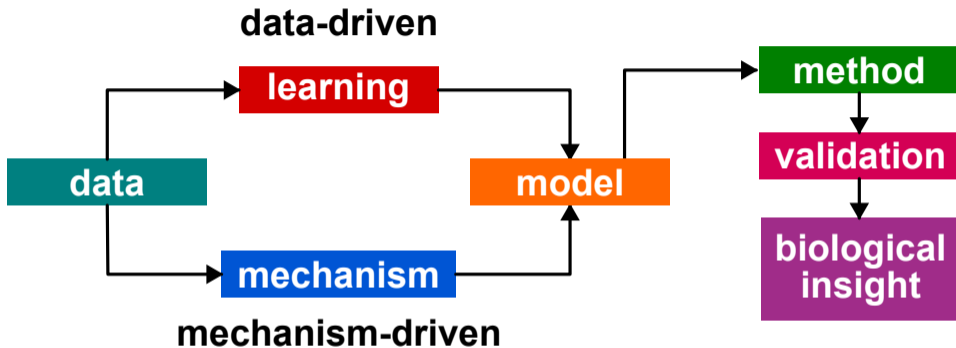
A personal take on predictive Bioinformatics



A personal take on predictive Bioinformatics



A personal take on predictive Bioinformatics



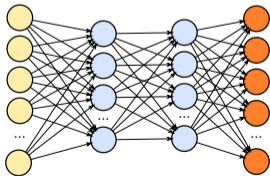
Method dev. as a modeling discipline:

Mechanism-driven model + Exact/deterministic algorithms
→ Performance as (in)validation of model

Machine Learning (ML): The beauty...

Machine Learning as a tool for scientific discovery

- ▶ Great promises
- ▶ Self-improving methods
- ▶ Generates/prioritizes hypotheses
- ▶ Available workforce (ubiquitous in curriculums)
- ▶ Highly promoted/funded by research institutions and glamorous journals. . .



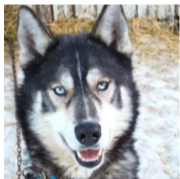
**Shut up and
take my money**



Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- ▶ **Tricky evaluation (data leakage) → Extrapolation/generalization???**
- ▶ Reproducibility issues (code/datasets availability, stability, retraining)
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ Educational deadend?
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)



(a) Husky classified as wolf



(b) Explanation

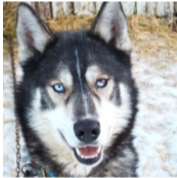
[Ribeiro et al, KDD'16]



Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- ▶ **Tricky evaluation (data leakage) → Extrapolation/generalization???**
- ▶ Reproducibility issues (code/datasets availability, stability, retraining)
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ Educational deadend?
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)

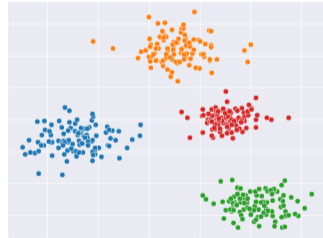


(a) Husky classified as wolf



(b) Explanation

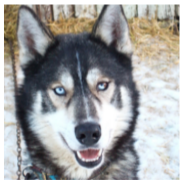
[Ribeiro et al, KDD'16]



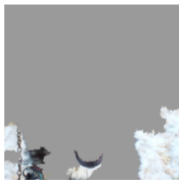
Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- ▶ **Tricky evaluation (data leakage) → Extrapolation/generalization???**
- ▶ Reproducibility issues (code/datasets availability, stability, retraining)
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ Educational deadend?
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)

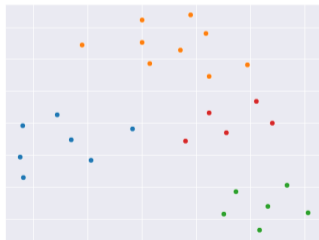


(a) Husky classified as wolf



(b) Explanation

[Ribeiro et al, KDD'16]



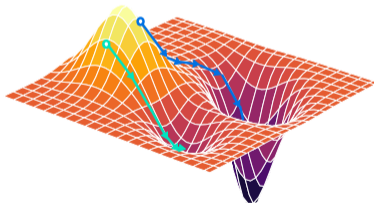
Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio* :

- ▶ Tricky evaluation (data leakage) → Extrapolation/generalization???
- ▶ **Reproducibility issues (code/datasets availability, stability, retraining)**
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ Educational deadend?
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)

Available upon request

***aka iff I'm in a good mood,
PhD/postdoc still in lab, HDDs haven't burned,
pharma hasn't expressed interest in data...***



Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

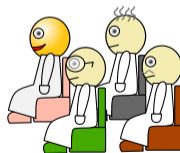
- ▶ Tricky evaluation (data leakage) → Extrapolation/generalization???
- ▶ Reproducibility issues (code/datasets availability, stability, retraining)
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ **Educational deadend?**
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)



Fifth law of thermodynamics (continued)

```
...  
-0.31622776601683794 0.31622776601683794  
-0.3157663248679193 0.3160839282916222  
0.006806069733149146 0.17777128902976705  
0.4472135954999579 1.433348584081719  
-1.5736761136523203 1.433348584081719  
-0.0002340648727882 0.4522609460629265  
...
```

24235/1020400



Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio* :

- ▶ Tricky evaluation (data leakage) → Extrapolation/generalization???
- ▶ Reproducibility issues (code/datasets availability, stability, retraining)
- ▶ Fishing expeditions/storytelling, selective reporting
- ▶ Educational deadend?
- ▶ Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble...)



A crowded ML field for RNA 2D prediction



Method	Output	PKs?	Architecture	Availability
CONTRAFold	Pairwise contacts	No	CLLM	Code+weights+webserver
EternaFold	Pairwise contacts	No	CLLM	Code+weights+webserver
DMfold	DBN	Yes	bi-LSTM	Code only
RNA-state-inf	Binary paired/unpaired	N/A	bi-LSTM	Code only
SPOT-RNA2	Pairwise contacts	Yes	CNN	Code+weights+webserver
CROSS	Binary paired/unpaired	N/A	CNN-like	Webserver
RPreS	Binary paired/unpaired	N/A	bi-LSTM+CNN	Code only
2dRNA	Pairwise contacts	Yes	bi-LSTM+CNN	Webserver
2dRNA-LD	Pairwise contacts	Yes	bi-LSTM+CNN	Webserver
SPOT-RNA	Pairwise contacts	Yes	CNN+bi-LSTM	Code+weights+webserver
MXfold2	Pseudo-dG	No	CNN+bi-LSTM	Code+weights+webserver
CNNFold	Pairwise contacts	Yes	CNN(NxN input)	Code+weights
UFold	Pairwise contacts	Yes	CNN(NxN input)	Code+weights+webserver
CDPfold	DBN	No	CNN(NxNinput)	Code
E2Efold	Pairwise contacts	Yes	Transformer+CNN	Code+weights
ATTFold	Pairwise contacts	Yes	Transformer+CNN	No

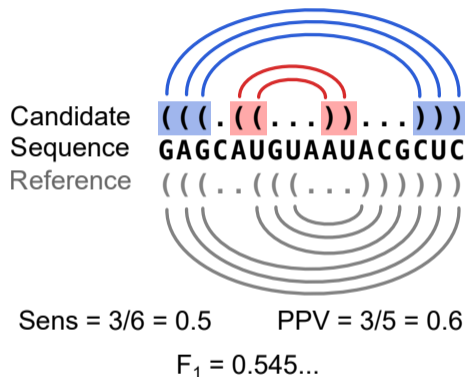
[Wu *et al*, Briefings in Bioinfo 2023]

Performances of 2D structure prediction

RNAstrand benchmark

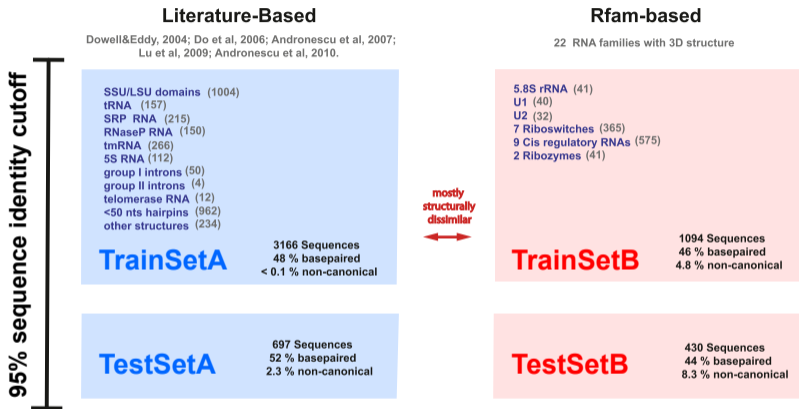
[Adronescu *et al*, BMC Bioinf 2008]

Method	F ₁
RNAfold 1.8.5	0.737
UNAFold 3.8	0.725
RNAstructure 5.7	0.744



$$F_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

The TORNADO dataset



[Rivas *et al*, RNA 2012]

TrainSetA vs **TestSetA**: 95% sim. cutoff → Learn k -mer to template association

(May happen even for extreme cutoffs)

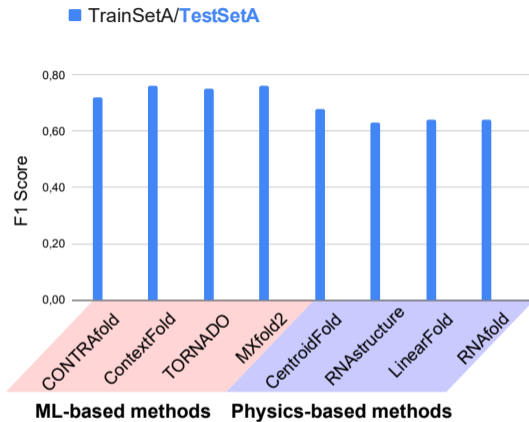
TrainSetA vs **TestSetB**: Rewards learning structurally generalizable models

Performances of 2D structure prediction

RNAstrand benchmark

[Adronescu *et al*, BMC Bioinf 2008]

Method	F ₁
RNAfold 1.8.5	0.737
UNAFold 3.8	0.725
RNAstructure 5.7	0.744



[Sato *et al*, Nature Comm 2021]

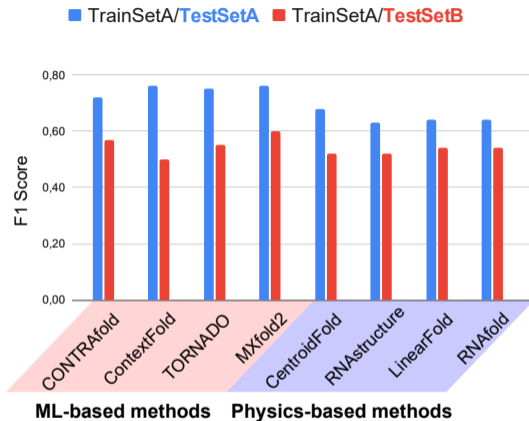
$$F_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

Performances of 2D structure prediction

RNAstrand benchmark

[Adronescu *et al*, BMC Bioinf 2008]

Method	F ₁
RNAfold 1.8.5	0.737
UNAFold 3.8	0.725
RNAstructure 5.7	0.744



[Sato *et al*, Nature Comm 2021]

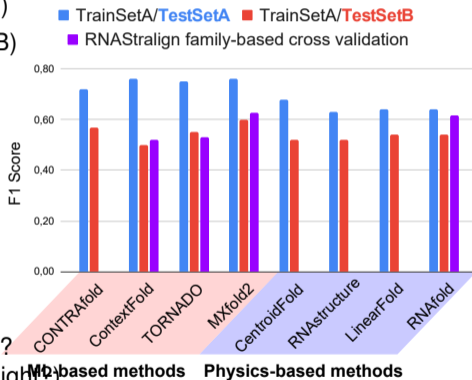
$$F_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

The (nc)RNA datasphere

- ▶ 34M sequences, inc 22M presumably structured (RNACentral)
- ▶ 4000+ functional ncRNA families (RFAM)
- ▶ 250-300 non-redundant 3D models (PDB)

Existing methods trained on datasets:

- ▶ highly-redundant sequence-wise
- ▶ low-diversity structure-wise



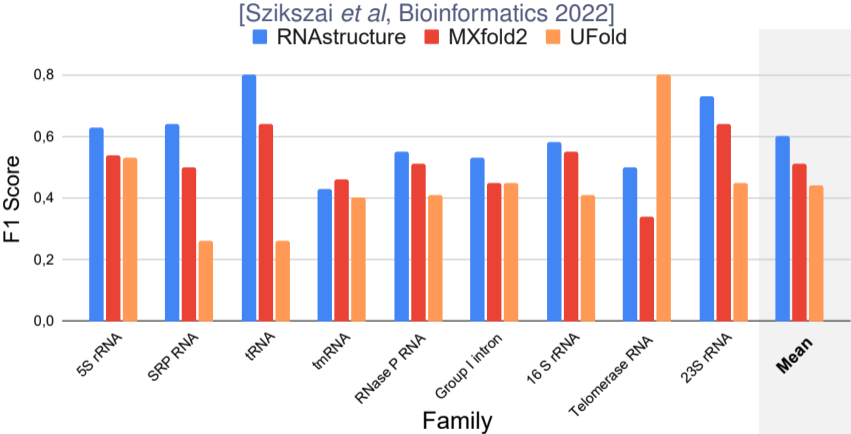
Do ML methods generalize to new structures?
(Do ML perfs translate into *new* biological insight?)

ML-based methods

Physics-based methods

[Sato *et al*, Nature Comm 2021]

Generalization to new families/structures remains problematic

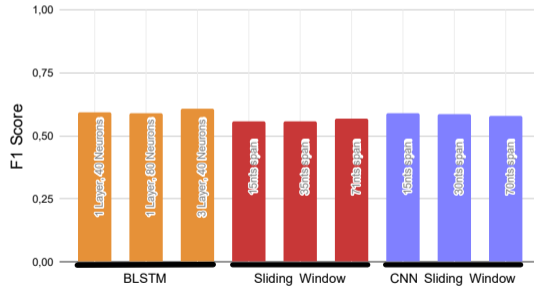


Family-fold cross-validation on **ArchiveII** dataset [Sloma & Mathews, RNA 2016]
3974 RNAs of length 77-438 (large rRNAs split into smaller domains)

What if you had access to (unbounded) additional data?

Idea: Assess NN models' capacity to emulate RNAfold on random sequences

[Flamm *et al*, Frontiers in Bioinfo 2022]

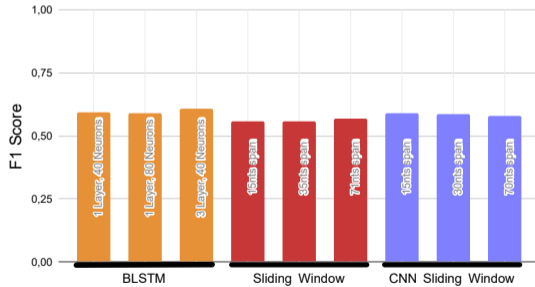


Perfs *plateau* at 80k seq/structs (70nts)

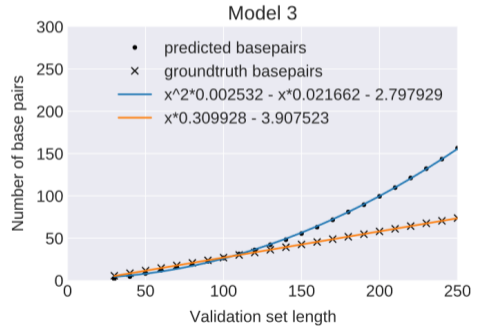
What if you had access to (unbounded) additional data?

Idea: Assess NN models' capacity to emulate RNAfold on random sequences

[Flamm *et al*, Frontiers in Bioinfo 2022]

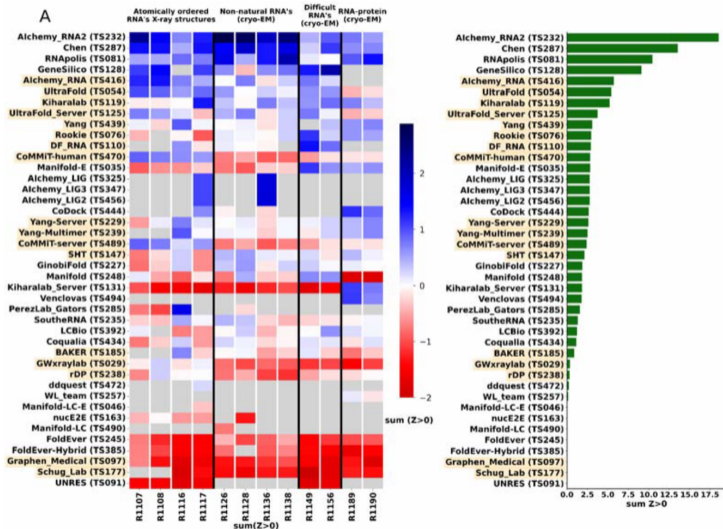


Perfs *plateau* at 80k seq/structs (70nts)



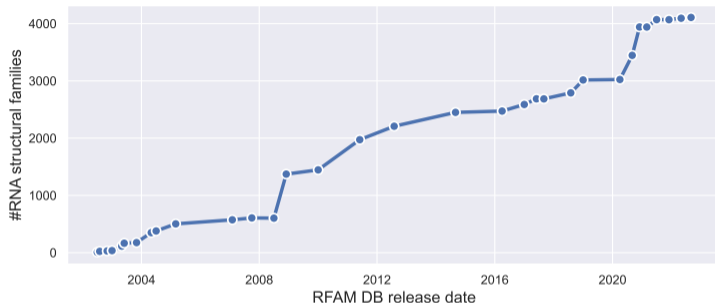
Popular CNN predicts $\Theta(n^2)$ BPs!

RNA 3D structure: No AlphaFold moment at CASP15



[Das *et al*, under review]

Conclusions and musings



- ▶ Still a need for improved RNA prediction (possibly ML-based)
- ▶ Purely combinatorial methods still \pm state-of-the-art for new families. . .
- ▶ Hybrid approaches *à la* MxFold2: Best of both worlds?
- ▶ Assessing intrinsic limits of architectures: RNAFold as surrogate model

Conclusions and musings

So what's special about RNA?

- ▶ Modular but combinatorial structure
- ▶ New folds being routinely discovered (+ can be designed)
- ▶ Relatively scarce 3D data
- ▶ Opportunity: Tons of probing data (ML)
- ▶ Potential of LLMs/transformers (incoming)
- ▶ Pseudoknots-ready algorithms

Conclusions and musings

RNA/Bioinfo community needs to enforce stricter standards for ML publications:

- ▶ Enforce datasets and source code availability
[Szikszai *et al*, Bioinfo'22] found 4/8 recent DL methods non-functional
- ▶ Realistic retraining mandatory
Precondition for self-improvement, benchmarking of novel methods
- ▶ Consider mechanistic and ML methods as largely incomparable
- ▶ Better datasets/benchmarks needed, but perhaps not sufficient
- ▶ Sequence-based leakage should be systematically investigated

Suboptimal structures

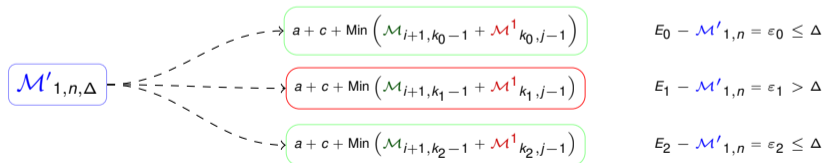
Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate suboptimal structures (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

- ▶ Compute minimum free-energy matrices
- ▶ **Backtrack on any contribution within Δ of MFE;**
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (brute-force ou Sort)



Suboptimal structures

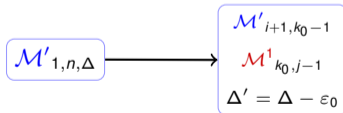
Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate suboptimal structures (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

- ▶ Compute minimum free-energy matrices
- ▶ Backtrack on any contribution within Δ of MFE;
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (brute-force ou Sort)



Suboptimal structures

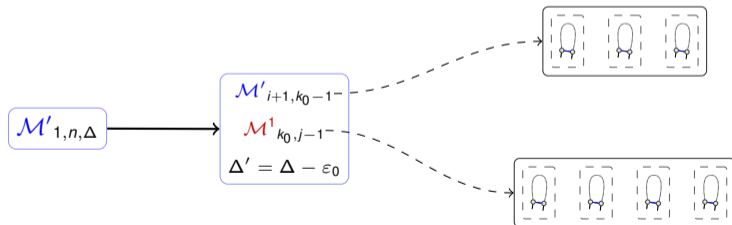
Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate suboptimal structures (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

- ▶ Compute minimum free-energy matrices
- ▶ Backtrack on any contribution within Δ of MFE;
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (brute-force ou Sort)



Suboptimal structures

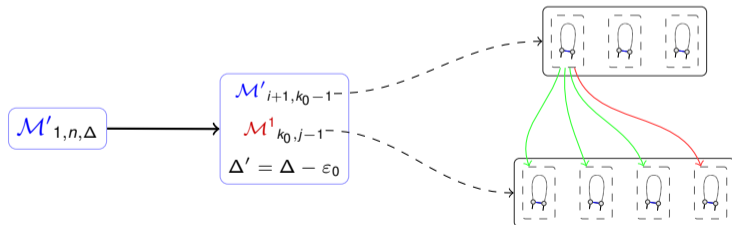
Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate suboptimal structures (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

- ▶ Compute minimum free-energy matrices
- ▶ Backtrack on any contribution within Δ of MFE;
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (brute-force ou Sort)



Suboptimal structures

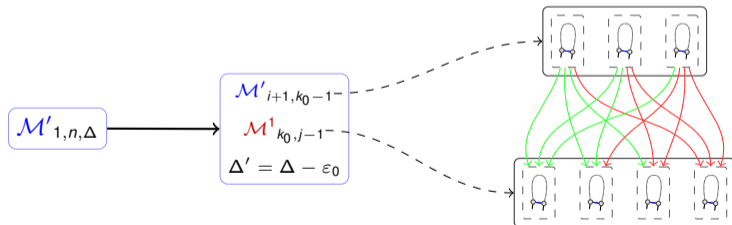
Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate suboptimal structures (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

- ▶ Compute minimum free-energy matrices
- ▶ Backtrack on any contribution within Δ of MFE;
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (brute-force ou Sort)



Suboptimal structures

Prob.: Simplified energy model (no pseudoknots, only canonical BPs)

⇒ **Native** structure (functional) could be **overthrown**.

⇒ Investigate **suboptimal structures** (RNASubopt [WFHS99]),

i.e. build all structures within Δ KCal.mol⁻¹ of MFE:

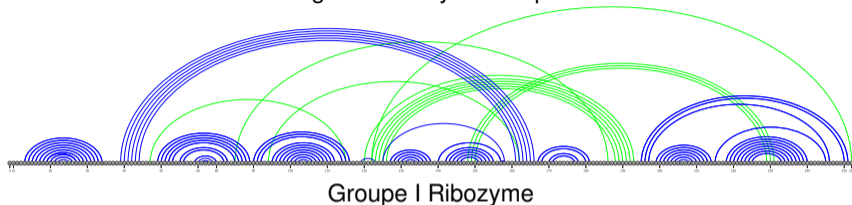
- ▶ Compute minimum free-energy matrices
- ▶ Backtrack on any contribution within Δ of MFE;
- ▶ Update Δ such that future backtracks create ≥ 1 struct.
- ▶ Recursively generate subopts and combine (**brute-force** ou **Sort**)

⇒ Time complexity (**Sort**) : $\mathcal{O}(n^3 + n \cdot k \log(k))$

(k grows exponentially fast with Δ !)

Predicting pseudoknotted structures

Pseudoknots are essential to the folding and activity of multiple RNA families.



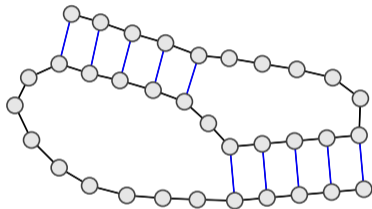
Their disregard within current folding algorithms stems both from **algorithmic** and **energetic** intricacies.

(**Pseudoknots** = Crossings \Rightarrow foldings delimited by base-pair can no longer be assumed to be independent)

Type	Complexity	Reference
Secondary structures	$\mathcal{O}(n^3)$	[MSZT99]
L&P	$\mathcal{O}(n^5)$	[LP00]
D&P	$\mathcal{O}(n^5)$	[DP03]
A&U	$\mathcal{O}(n^5)$	[Aku00]
R&E	$\mathcal{O}(n^6)$	[RE99]
Unconstrained	NP-complete	[LP00]

Akutsu/Uemura Algorithm

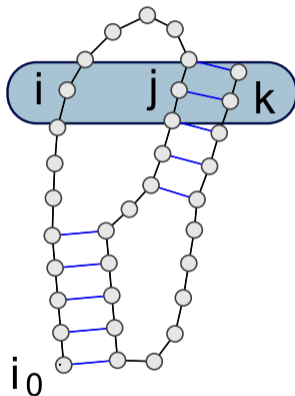
Goal: Capture a category of **simple, yet** recurrent, pseudoknots.



Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.

Akutsu/Uemura Algorithm

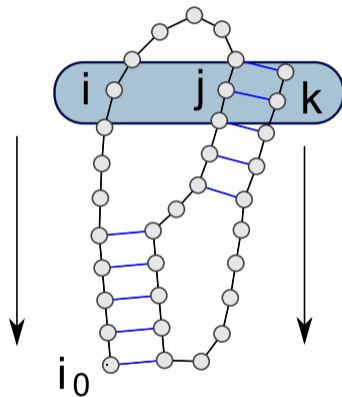
Goal: Capture a category of **simple, yet** recurrent, pseudoknots.



Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.

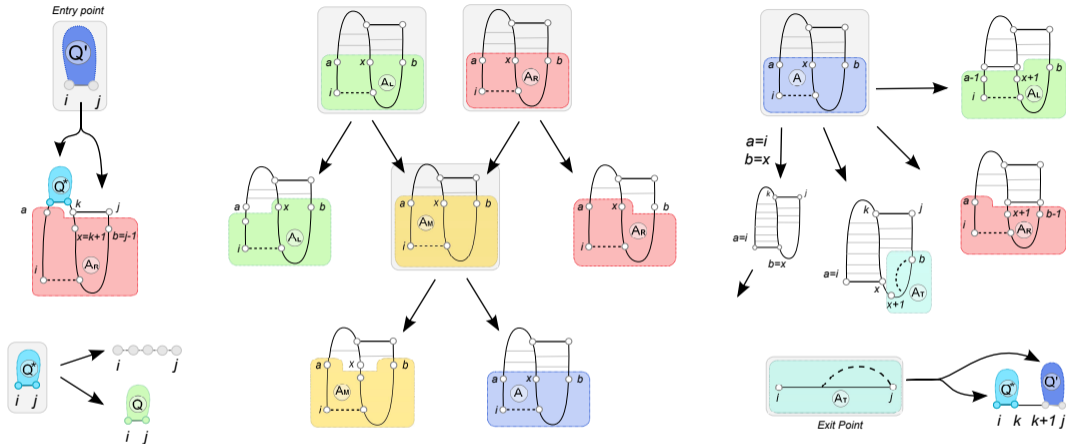
Akutsu/Uemura Algorithm

Goal: Capture a category of **simple, yet** recurrent, pseudoknots.



Idea: When such a PK motif is **rotated**, one can deduce the MFE of a triplet (i, j, k) from the MFE of triplets **directly below** it.

Akutsu/Uemura: Dynamic programming



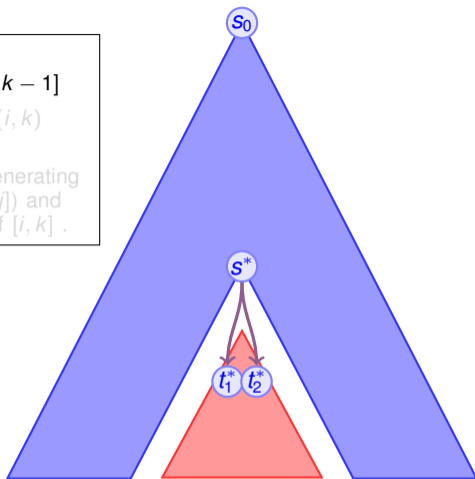
Application/Problem	Weight fun.	Time/Space	Ref.
Energy minimization	π_{bp}	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	[Aku00]
Partition function	$e^{-\frac{\pi_{bp}}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	$\Theta(n^5)$ [CC09]
BP probabilities	$e^{-\frac{\pi_{bp}}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	-
Sampling (k -struct.)	$e^{-\frac{\pi_{bp}}{RT}}$	$\mathcal{O}(n^4 + kn \log n)/\mathcal{O}(n^4)$	-

Exercise: Write DP equation for MFE computation, counting and partition function.

Inside/outside algorithm

Structure including base pair (i, k) :

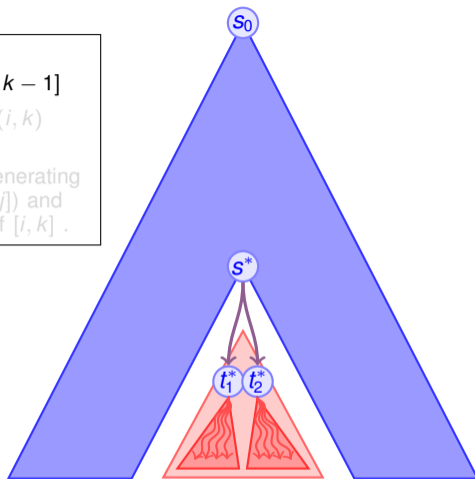
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ Complete structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

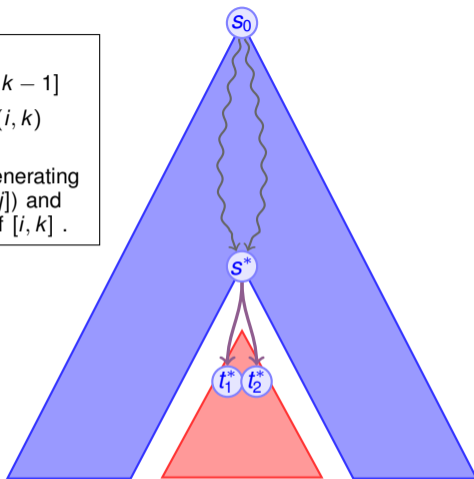
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ Complete structure by generating brother intervals $([k + 1, j])$ and recurse over the father of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

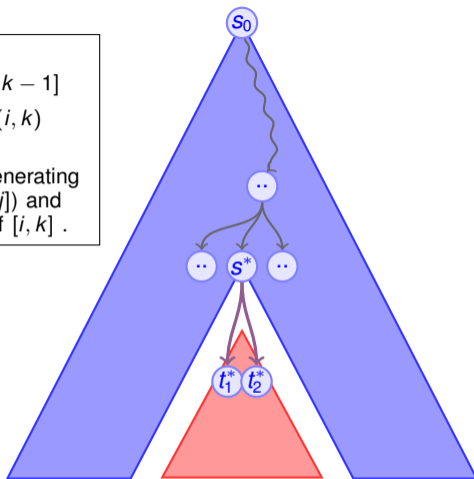
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ **Complete** structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

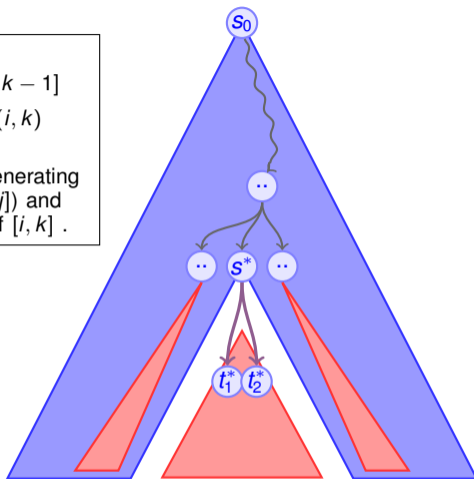
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ **Complete** structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

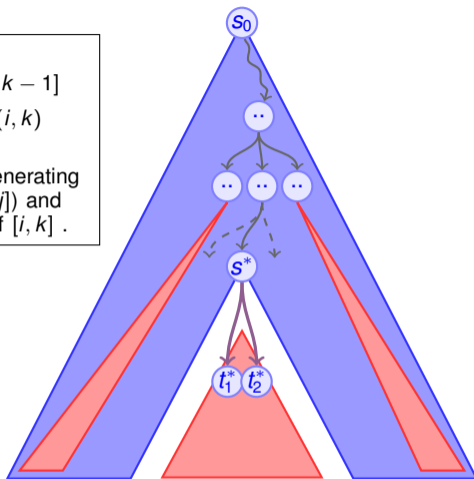
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ **Complete** structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

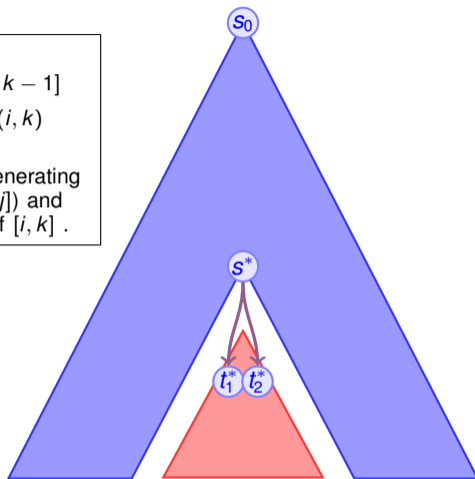
- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ **Complete** structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Inside/outside algorithm

Structure including base pair (i, k) :

- ▶ **Inside:** Structures over $[i + 1, k - 1]$
- ▶ **Outside:** Contexts of interval (i, k)
 - ▶ \forall interval $[i, j], i < j \leq k$
 - ▶ **Complete** structure by generating **brother intervals** $([k + 1, j])$ and recurse over the **father** of $[i, k]$.



Whenever some further **technical conditions** are satisfied, this decomposition is **complete** and **unambiguous**, and implies a **simple recurrence** for computing the base pair probability matrix in $\Theta(n^3)$.

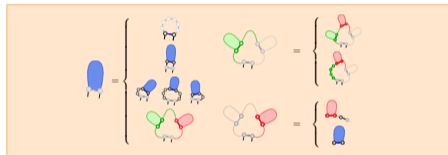
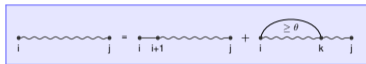
Alternatively: Duplicate sequence

What is a good dynamic programming scheme?

Correction of a (Ensemble) dynamic programming scheme:

Objective function **correctly** computed/inherited at **local level**

- + All the conformations can be obtained
- ⇒ Correct algorithm (Induction)



Enumerating search space helps **but** does not constitute a proof.

Need to **show equivalence** of DP schemes, e.g. use one to simulate the other and vice versa.

(Generating functions may help)

What is a good dynamic programming scheme?

Correction of a (Ensemble) dynamic programming scheme:

Objective function **correctly** computed/inherited at **local level**

+ **All the conformations can be obtained**

⇒ **Correct algorithm (Induction)**

$$C_{i,t} = 1, \quad \forall t \in [i, i + \theta]$$
$$C_{i,j} = \sum \left\{ \begin{array}{l} C_{i+1,j} \\ \sum_{k=i+\theta+1}^j 1 \times C_{i+1,k-1} \times C_{k+1,j} \end{array} \right.$$

Homopolymer (All pairs allowed) + $\theta = 1$
⇒ $C_{1,n} = 1, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \dots$



$$C'_{i,j} = \sum \left\{ \begin{array}{l} 1 \\ C'_{i+1,j-1} \\ \sum_{i',j'} C'_{i',j'} \\ \sum_k C_{i+1,k-1} \times C^1_{k,j-1} \end{array} \right.$$
$$C_{i,j} = \sum_k \left((C_{i,k-1} + 1) \times C^1_{k,j} \right)$$
$$C^1_{i,j} = C^1_{i,j-1} + C'_{i,j}$$

Homopolymer + $\theta = 1$
⇒ $C'_{1,n} = 0, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \dots$

Enumerating search space helps **but** does not constitute a proof.

Need to **show equivalence** of DP schemes, e.g. use one to simulate the other and vice versa.

(Generating functions may help)

Structural alignment: Why?

Hypothesis: Common evolutionary pressure = Common function .

Within certain RNA families (ex.: RNase-P), low sequence conservation **yet** high structural conservation.

Algorithmic problems:

- ▶ **Editing:** Compute *distance* between two secondary structures A and B .
Find minimal cost sequence of operations to turn A into B . Already NP-complete for two secondary structures [BFRS07].
- ▶ **Alignment:** Find minimal cost super-structure.
Generalizes sequence alignment. Polynomial ($\mathcal{O}(n^4)$) for secondary structures [BDD⁺08], NP-complete in 3D [SZS⁺08].
Alternatives: Local/global alignment, motifs search (aka small-in-large).
- ▶ **Superimposition:** Find solid-body geometric transform (Rotation, translation, zoom) to superimpose *as well as possible* the coordinates of two RNAs having **known matching**.
Polynomial in 3D [McL82].

Remark: Algorithmic hardness stems from finding the matching (i.e. combinatorial, not geometric).

FR3D: A geometric approach

When 3D models are available, the alignment problem can be tackled in a **purely geometric** setting.

Problem

Input: Motif m , target structure b (ordered set of 3D points).

Output: Matching of m versus a subset of b that minimizes a notion of geometric **discrepancy**.

Geometric discrepancy: In FR3D [SZS⁺08], a **discrepancy** function D combines two error functions L et A , respectively accounting for the **superimposability** (L) and **base orientation** (A) of m and b .

$$L = \sqrt{\min_{R,T} \sum_{i=1}^m \|b_i - R(T(m_i))\|^2} \quad A = \sqrt{\sum_{i=1}^m \alpha_i^2} \quad D = \frac{1}{m} \sqrt{L^2 + A^2}$$

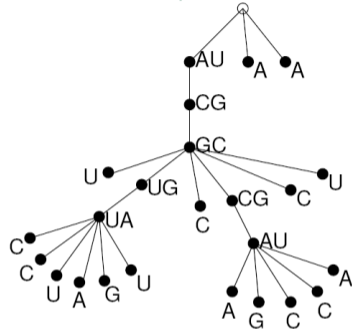
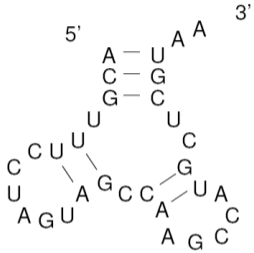
R, T : Rotation and translation. c_i : Center of mass (CM) of base m_i . α_i : Spread between orientation of CMs/bases in m_i et b_i .

Backtrack + Incremental pruning (Bounds on D) \Rightarrow **Combinatorial explosion!**

But exact search feasible for smaller motifs.

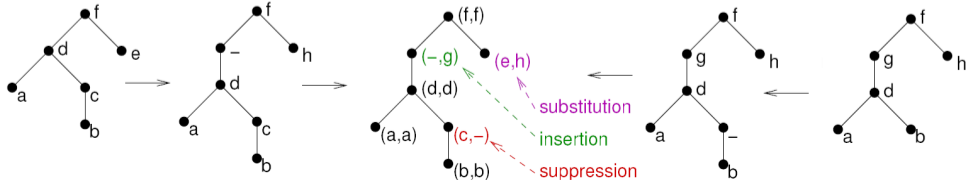
Structures to Trees

The alignment of two secondary structures is based on their **tree-like representations**¹.



Base pairs \Rightarrow internal nodes Unpaired bases \Rightarrow Leaves

Alignment = Complete matching having minimal cost.



Historic algorithm: Jiang, Wang & Zhang 95 [JWZ94]

Aligning Trees²

$$\delta(\text{Tree}_1, \text{Tree}_2) = \min \begin{cases} \delta(\text{Tree}_1, \text{Tree}_2) + \text{del}(\bullet) \\ \delta(\text{Tree}_1, \text{Tree}_2) + \text{ins}(\bullet) \\ \delta(\text{Tree}_1, \text{Tree}_2) + \text{subst}(\bullet, \bullet) \end{cases}$$

Aligning Forests

$$\delta(\text{Forest}_1, \text{Forest}_2) = \min \begin{cases} \min\{\delta(\text{Tree}_1, \text{Tree}_2) + \delta(\text{Forest}_1, \text{Forest}_2) \mid \text{Forest}_1 = \text{Tree}_1 \cup \text{Forest}_2\} + \text{del}(\bullet) \\ \min\{\delta(\text{Tree}_1, \text{Tree}_2) + \delta(\text{Forest}_1, \text{Forest}_2) \mid \text{Forest}_1 = \text{Tree}_1 \cup \text{Tree}_2\} + \text{ins}(\bullet) \\ \delta(\text{Tree}_1, \text{Tree}_2) + \delta(\text{Forest}_1, \text{Forest}_2) \end{cases}$$

Worst-case complexity in $\mathcal{O}(n^4)$ [JWZ94], on average in $\mathcal{O}(n^2)$ [HDD07].

But RNA-specific operations are lacking

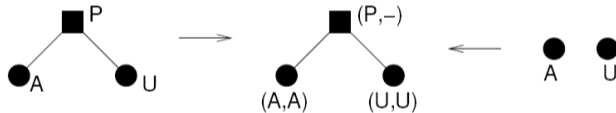
²Idem

Parametrization of operation costs, but some operations, atomic in a realistic model, must be composed from available ones.

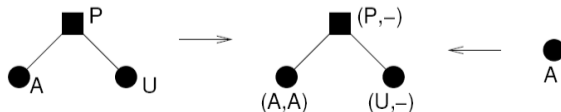
Example: To detach a base-pair, delete node (base-pair), and insert two leaves (bases).

RNAForester: Based on Jiang, Wang & Zhang algorithm
+ Integration of RNA-specific operations³.

arc-breaking



arc-altering








³Idem






$$\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) =$$

{	min	$\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \text{BDel}(\bullet)$	si \bullet base
		$\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \text{BIns}(\bullet)$	si \bullet base
		$\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \text{BSub}(\bullet, \bullet)$	si \bullet et \bullet bases
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{PDel}(\bullet)$	si \bullet paire
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{PIns}(\bullet)$	si \bullet paire
		$\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \text{PSub}(\bullet, \bullet)$	si \bullet et \bullet paires
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{Fus}(\bullet, \bullet, \bullet)$	si \bullet paire et \bullet base
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{Sci}(\bullet, \bullet, \bullet)$	si \bullet paire et \bullet base
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{GAlt}(\bullet, \bullet)$	si \bullet paire et \bullet base
		$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{DAlt}(\bullet, \bullet)$	si \bullet paire
$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{GComp}(\bullet, \bullet)$	si \bullet paire et \bullet base		
$\min\{\delta(\text{▲▲▲▲}, \text{▲▲▲▲}) + \delta(\text{▲▲▲▲}, \text{▲▲▲▲}) : \text{▲▲▲▲} = \text{▲▲▲▲}\} + \text{DComp}(\bullet, \bullet)$	si \bullet paire		






References I

-  **Tatsuya Akutsu.**
Dynamic programming algorithms for rna secondary structure prediction with pseudoknots.
Discrete Appl. Math., 104(1-3):45–62, 2000.
-  **G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet.**
Alignment of rna structures.
Transactions on Computational Biology and Bioinformatics, , 2008.
A paraître.
-  **Guillaume Blin, Guillaume Fertin, Irena Rusu, and Christine Sinoquet.**
Extending the Hardness of RNA Secondary Structure Comparison.
In Bo Chen, Mike Paterson, and Guochuan Zhang, editors, *ESCAPE'07*, volume 4614 of *LNCS*,
pages 140–151, Hangzhou, China, Apr 2007.
-  **S. Cao and S-J Chen.**
Predicting structured and stabilities for h-type pseudoknots with interhelix loop.
RNA, 15:696–706, 2009.
-  **K. Doshi, J. J. Cannone, C. Cobaugh, and R. R. Gutell.**
Evaluation of the suitability of free-energy minimization using nearest-neighbor energy
parameters for rna secondary structure prediction.
BMC Bioinformatics, 5(1):105, 2004.





References II

-  **Robert M Dirks and Niles A Pierce.**
A partition function algorithm for nucleic acid secondary structure including pseudoknots.
J Comput Chem, 24(13):1664–1677, Oct 2003.
-  **F. Ferrè, Y. Ponty, W. A. Lorenz, and Peter Clote.**
Dial: A web server for the pairwise alignment of two RNA 3-dimensional structures using nucleotide, dihedral angle and base pairing similarities.
Nucleic Acids Research, 35(Web server issue):W659–668, July 2007.
-  **P. Gardner and R. Giegerich.**
A comprehensive comparison of comparative rna structure prediction approaches.
BMC Bioinformatics, 5(1):140, 2004.
-  **Claire Herrbach, Alain Denise, and Serge Dulucq.**
Average complexity of the jiang-wang-zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm.
In Proceedings of MACIS 2007, Second International Conference on Mathematical Aspects of Computer and Information Sciences, 2007.
-  **M. Hochsmann, B. Voss, and R. Giegerich.**
Pure multiple RNA secondary structure alignments: A progressive profile approach.
01(1):53–62, 2004.

References III

-  Tao Jiang, Lusheng Wang, and Kaizhong Zhang.
Alignment of trees - an alternative to tree edit.
In *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 75–86, London, UK, 1994. Springer-Verlag.
-  R. B. Lyngsø and C. N. S. Pedersen.
RNA pseudoknot prediction in energy-based models.
Journal of Computational Biology, 7(3-4):409–427, 2000.
-  D. McLachlan.
Rapid comparison of protein structures.
Acta crystallographica A, 38(6):871–873, 1982.
-  D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner.
Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure.
Journal of Molecular Biology, 288(5):911–940, May 1999.
-  M. Parisien and F. Major.
The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.
Nature, 452(7183):51–55, 2008.

References IV

-  **E. Rivas and S.R. Eddy.**
A dynamic programming algorithm for RNA structure prediction including pseudoknots.
J Mol Biol, 285:2053–2068, 1999.
-  **B. A. Shapiro, Y. G. Yingling, W. Kasprzak, and E. Bindewald.**
Bridging the gap in rna structure prediction.
Curr Opin Struct Biol, 17(2):157–165, Apr 2007.
-  **M. Sarver, C. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis.**
FR3D: Finding local and composite recurrent structural motifs in RNA 3D.
Journal of Mathematical Biology, 56(1–2):215–252, January 2008.
-  **S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster.**
Complete suboptimal folding of RNA and the stability of secondary structures.
Biopolymers, 49:145–164, 1999.