Machine learning approaches for assessing predicted structures of RNAs and proteins

François GASTRIN

M2 Bio-informatique - Université Paris Cité
Laboratoire IBISC
Supervisor : Guillaume POSTIC

June 28, 2022

université PARIS-SACLAY | GRADUATE SCHOOL Informatique et Sciences du Numérique

Table of Contents

The protein folding problem

- A major scientific question [1]
- **NP-hardness** [2]
- For an average size protein of 100 residues
  $\rightarrow$ find the structure with the lowest Gibbs free energy
  $\Rightarrow$ on a total of $\sim 10^{30}$ conformations

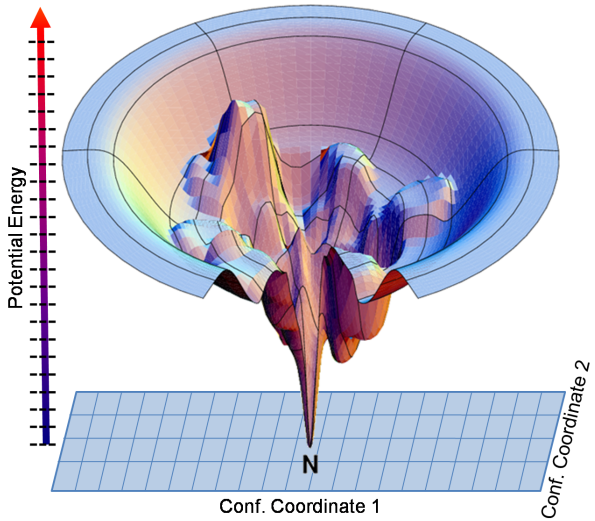
Heuristic

- Non-exhaustive algorithmic search of the conformation space
- Optimization of an objective function

1. "What Don't We Know ?", *Science*, 2005
2. Unger and Moult, 1993

Objective function : "knowledge-based"

Approximates Gibbs free energy → minimization



- Physical potentials → Fail
- Statistical "potentials" → AlphaFold [3]
  ⇒ Build on native (experimental) structures

3. Senior *et al.*, 2020

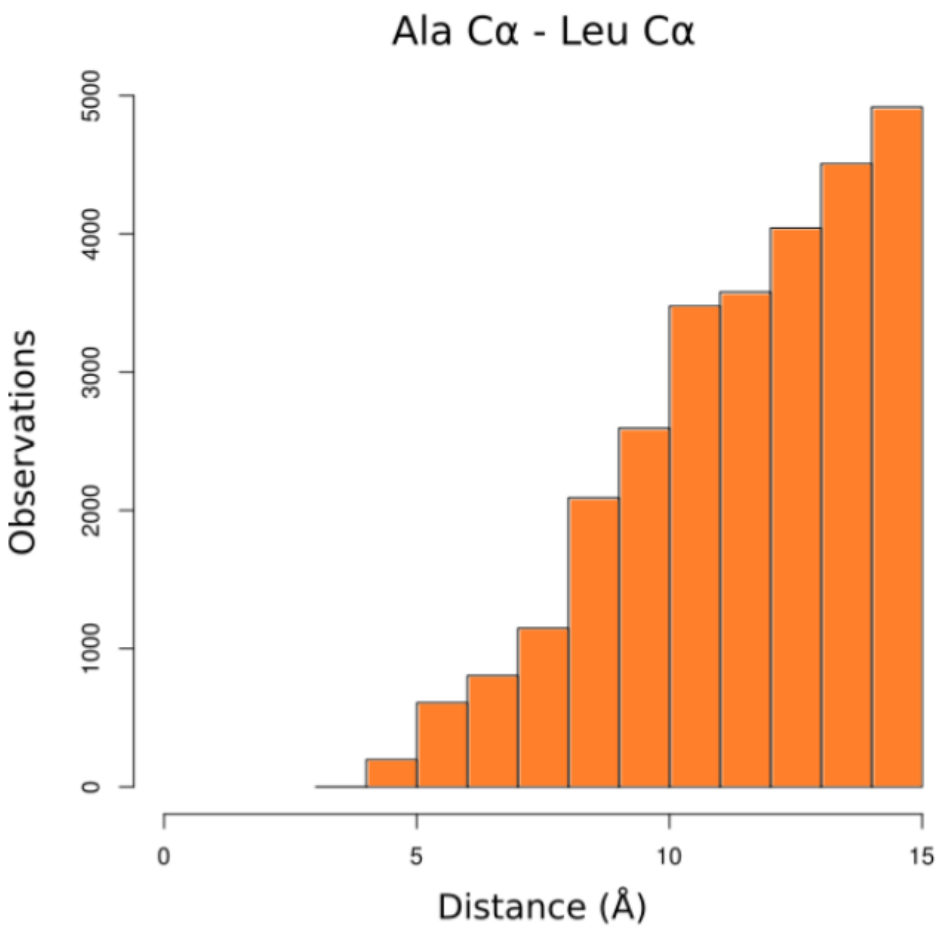
Objective function : "knowledge-based"

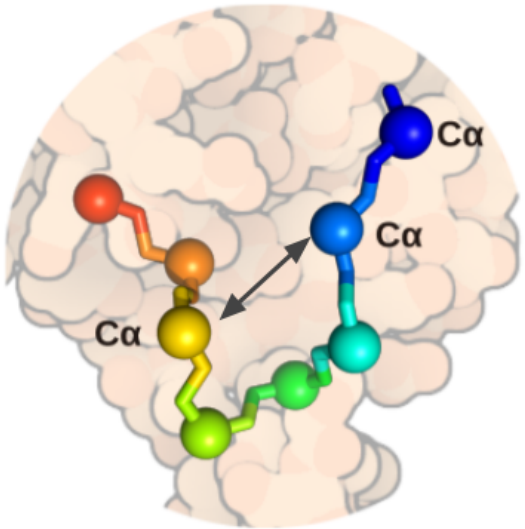
How do they work ?

Can we improve them ?
(+RNA structures ?)

Table of Contents

Statistical "potentials" : How do they work ?





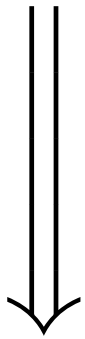
Interatomic distances between residues

The most frequent an interaction (observed in native structures), the most favorable the score (i.e. predicted quality)

Statistical "potentials" : How do they work ?

- "Statistical potentials of mean force" [4] :

$$score = -\sum_{i=1}^{n} log\left[\frac{p(i|native)}{p(i|non-native)}\right]$$



- Multinomial naive Bayes (MNB) classifier :

$$p(native|\vec{X}) = \frac{p(native)\prod_{i=1}^{n} p(i|native)^{x_i}}{p(\vec{X})}$$

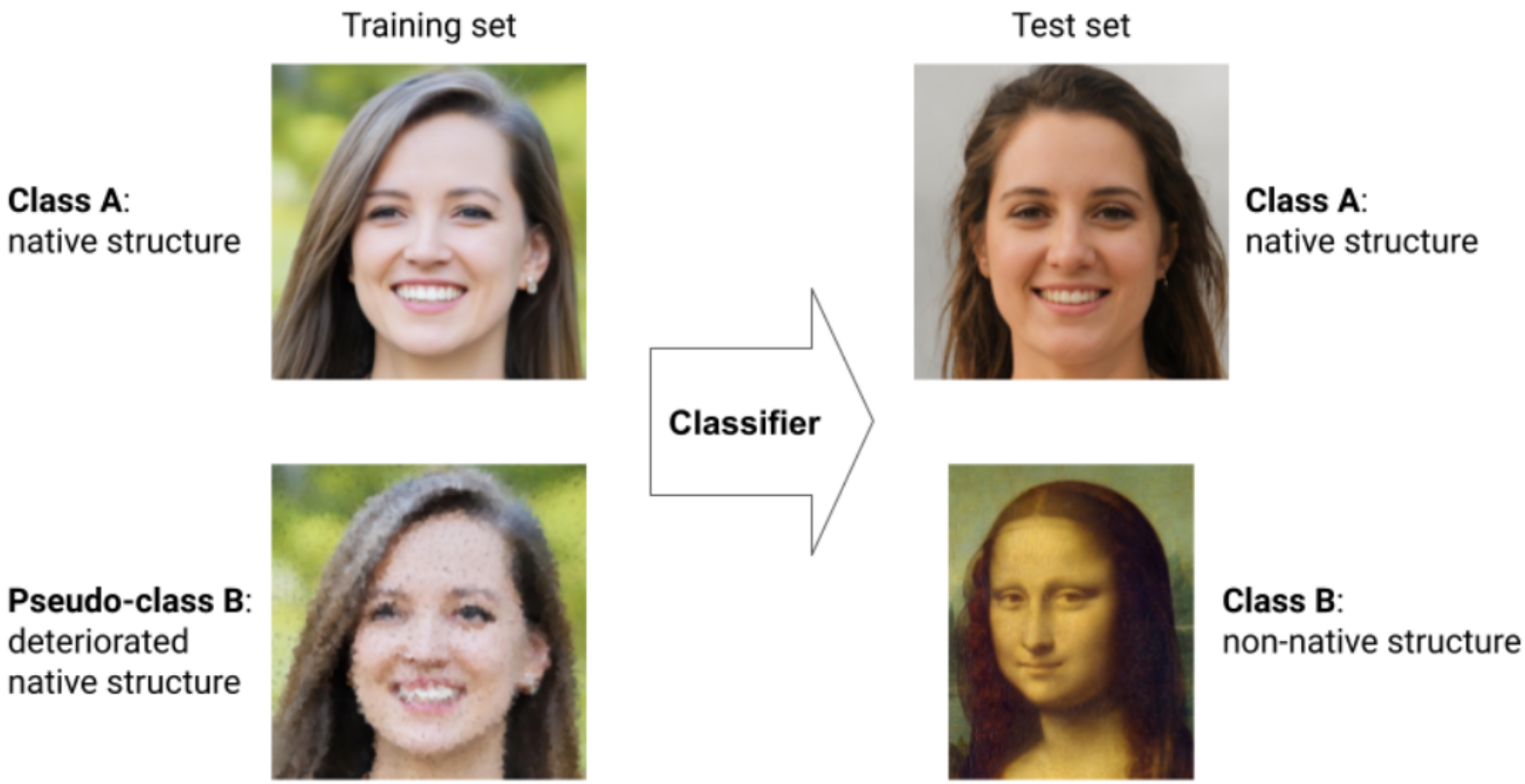
i : interaction type = distances bins + residue pair types
$\vec{X}$: feature vector

4. Sippl, 1990

Statistical "potentials" : How do they work ?

Generalized principle :

- Binary classifier $\rightarrow$ trained using a single class



Statistical "potentials" : Can we improve them ?

- Including RNAs ?
- Using the same data organization :

Data instance	Features							Class
	AA [0.0, 0.5[	AC [0.0, 0.5[	…	UU [0.0, 0.5[	…	GU [14.5, 15.0[	UU [14.5, 15.0[	
Native 1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,10}$	…	$x_{1,299}$	$x_{1,300}$	Native
Native 2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,10}$	…	$x_{2,299}$	$x_{2,300}$	Native
…	…	…	…	…	…	…	…	…
Native n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,10}$	…	$x_{n,299}$	$x_{n,300}$	Native
Averaged 1	$\text{avg}(x_{1,1}: x_{1,10})$	$\text{avg}(x_{1,1}: x_{1,10})$	…	$\text{avg}(x_{1,1}: x_{1,10})$	…	$\text{avg}(x_{1,291}: x_{1,300})$	$\text{avg}(x_{1,291}: x_{1,300})$	Non-native
Averaged 2	$\text{avg}(x_{2,1}: x_{2,10})$	$\text{avg}(x_{2,1}: x_{2,10})$	…	$\text{avg}(x_{2,1}: x_{2,10})$	…	$\text{avg}(x_{2,291}: x_{2,300})$	$\text{avg}(x_{2,291}: x_{2,300})$	Non-native
…	…	…	…	…	…	…	…	…
Averaged n	$\text{avg}(x_{n,1}: x_{n,10})$	$\text{avg}(x_{n,1}: x_{n,10})$	…	$\text{avg}(x_{n,1}: x_{n,10})$	…	$\text{avg}(x_{n,291}: x_{n,300})$	$\text{avg}(x_{n,291}: x_{n,300})$	Non-native

$\rightarrow$ Replace the MNB by other classifiers RF, SVM, or CNN
$\Rightarrow$ **It should work...**

Training procedure

Protein dataset

- 1917 native structures of non-redundant proteins (sequence identity < 20.0%) [5]
- Chain lengths → from 50 to 250 amino acids
- Resolution → 2.5 Å maximum

5. Postic et al., 2020

RNA dataset

- 174 structures with no RNA-protein or RNA-DNA complexes [6]
- Chain lengths → 10 nucleotide minimum
- Resolution → less than 3.5 Å

6. Tan et al., 2022

Benchmarking procedure

Benchmarking datasets

Protein-specific : 3DRobot [7] → 200 non-redundant proteins (native) + 300 decoys generated (for each native) ⇒ 60,200 structures

RNA-specific : Melo's lab dataset [8] → 85 RNA chains (native) + 500 decoys generated (for each native) ⇒ 42,585 structures

7. Deng et al., 2016
8. Capriotti et al., 2011

Accuracy measures

- Pearson correlation coefficient : true model quality (TM-score [9] for proteins / DI [10] for RNA) VS predicted model quality (scoring function)
- Enrichment score (ES) : (Predicted top10% ∩ True top10%)÷N
- Native structures found → ranked first or in the top five

9. Zhang and Skolnick, 2004
10. Parisien et al., 2009

Pipeline for training and benchmarking

- Cross-platform
- User-friendly
- Open-source

https://github.com/FranGASTRIN/my_RNAAssessment

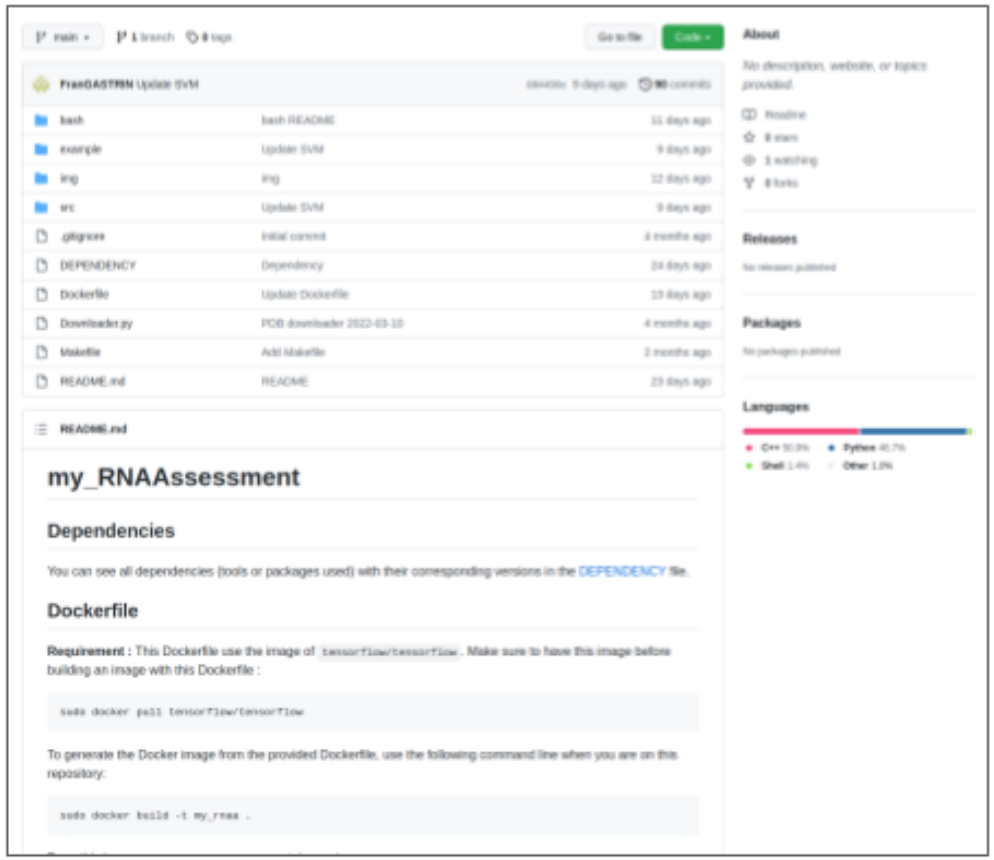


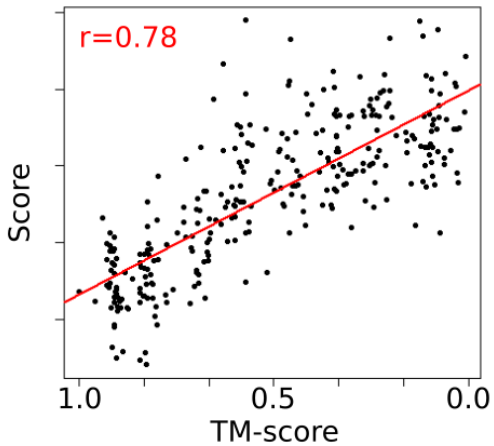
Table of Contents

Protein-specific

The higher the values, the more accurate the method :

	MNB	RF	SVM	CNN
Native structures found	44 (46)	44 (46)	50 (58)	50 (60)
Enrichment score	8.4 ±0.8	8.3 ±0.9	8.6 ±0.8	8.7 ±0.7
Pearson correlation coefficient	0.83 ± 0.07	0.82 ± 0.08	0.89 ± 0.06	0.90 ± 0.06

- SVM > RF and MNB $\Rightarrow$ SVM intrinsically binary
- CNN $\simeq$ SVM : theoretical limit reached for performance ?
- RF $\simeq$ MNB : not the best decision function ?

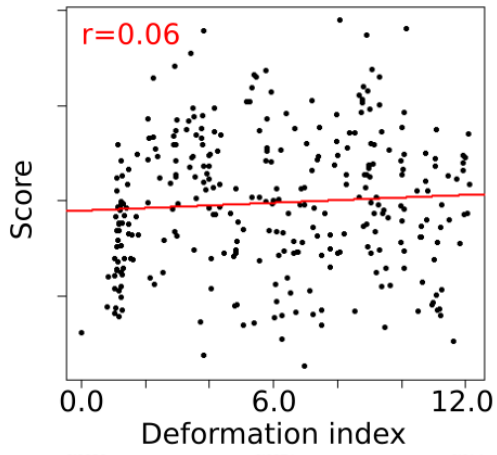


RNA-specific

The higher the values, the more accurate the method :

	MNB	RF	SVM	CNN
Native structures found	29 (45)	28 (47)	38 (50)	38 (52)
Enrichment score	3.2 ± 0.7	3.3 ± 0.7	3.8 ± 0.6	3.8 ± 0.5
Pearson correlation coefficient	0.56 ± 0.16	0.56 ± 0.15	0.61 ± 0.15	0.62 ± 0.14

- Low resolution representation (single atom)
  $\Rightarrow$ C4' or C1' are not equivalent to Cβ in proteins
- MNB test on all-atom $\rightarrow$ good results



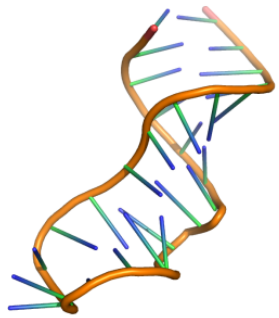


Table of Contents

Conclusion and Perspectives

Generalization of the statistical “potentials” formalism
$\Rightarrow$ Performances increased, but there is still room for improvement

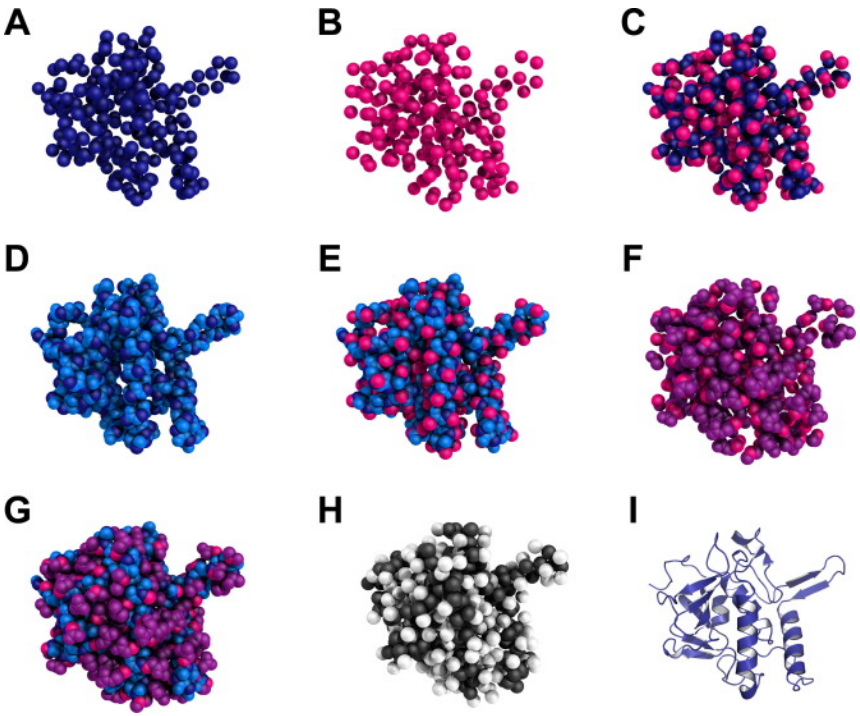
1. Structure representation
2. RNA training set
3. Dihedral angles added to features

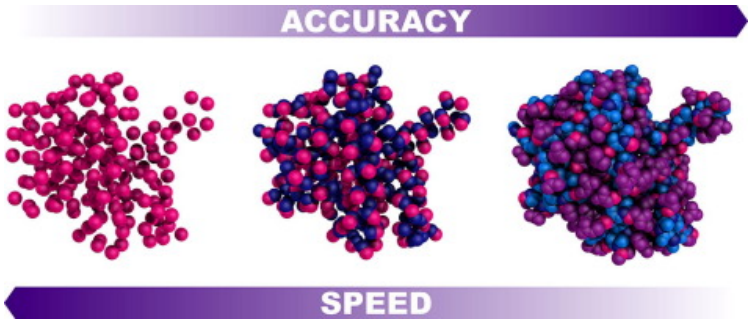
Possibility of improvement #1

Representation

Try low- and high-resolution representations for RNA [11]

11. Postic et al., 2021





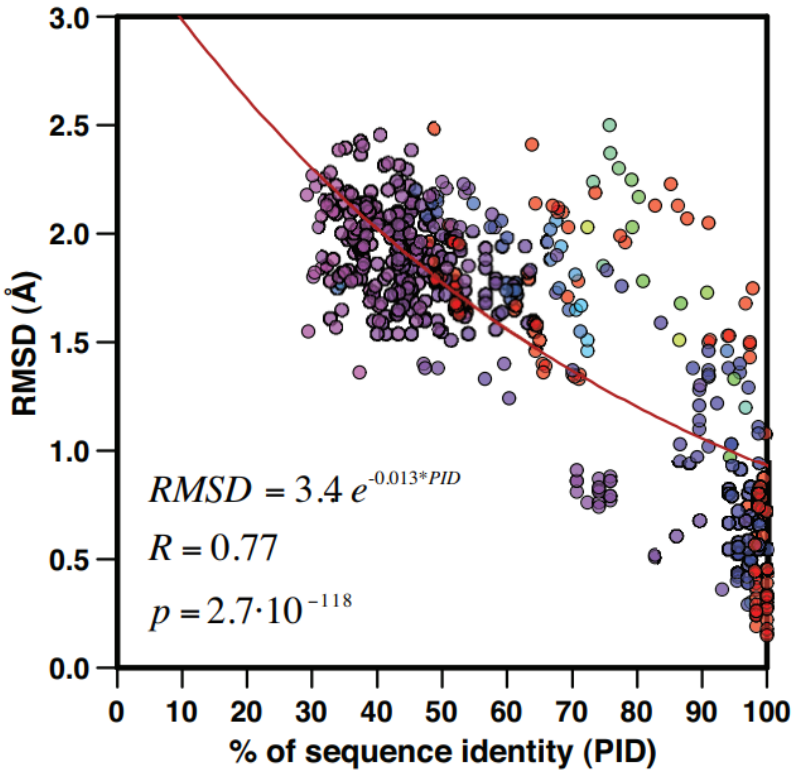
Possibility of improvement #2

RNA training set

Criteria for filtering redundancy of structures

$\Rightarrow$ Sequence/structure relation not as clear for RNA structures as in proteins [12]

12. Capriotti and Marti-Renom, 2010



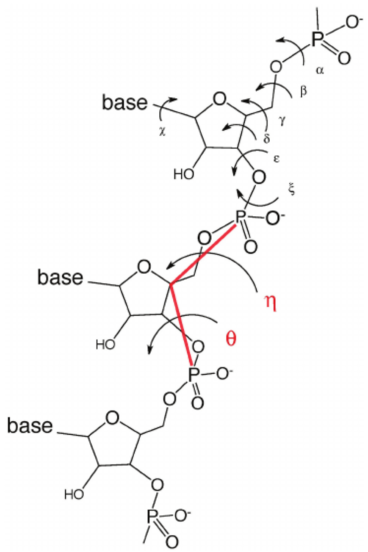
Possibility of improvement #3

Dihedral angles : additional features

Add to the feature vector the dihedral angles (i.e. ϕ and ψ for proteins). For RNA, we can use pseudotorsion angles η and θ, or η' and θ' :

- C++ program developed during this internship $\Rightarrow$ able to compute both of these torsion and pseudotorsion angles
- Faster than the only alternative available for RNA $\rightarrow$ AMIGOSIII from Pyle's lab [13]

13. Shine et al., 2022



1917 PDB files	CPU time (s)
default (Phi and Psi only)	7.0
default +options (file name,etc…)	7.1
default + Omega	9.2
default + Omega + options	9.3

RNANet : RNA 3D structures database (Becquey et al., 2021)

Web interface development

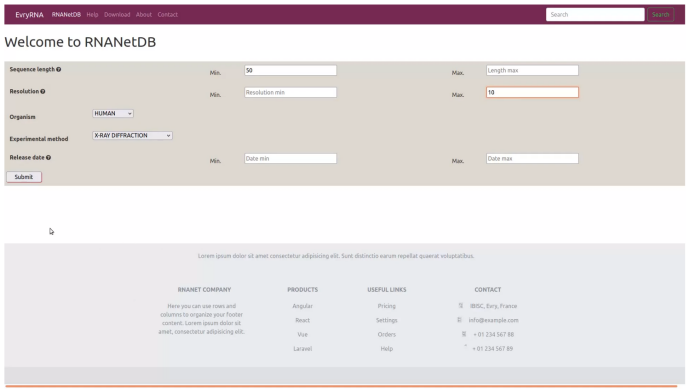
Back-end

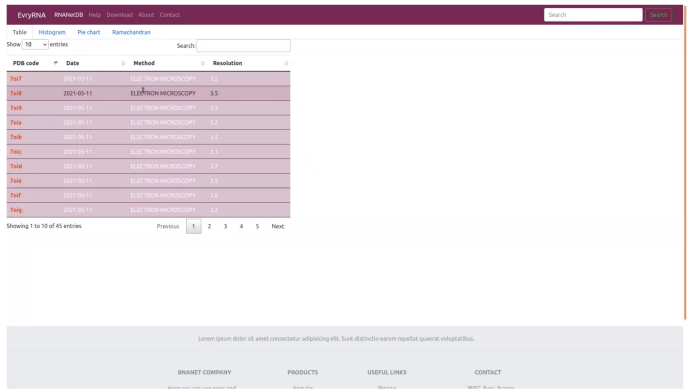
Server (LAMP)

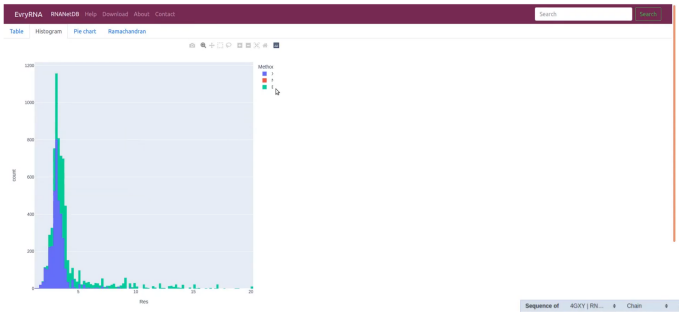
Client

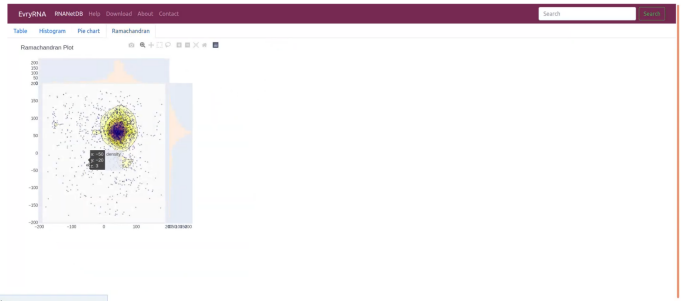
Front-end

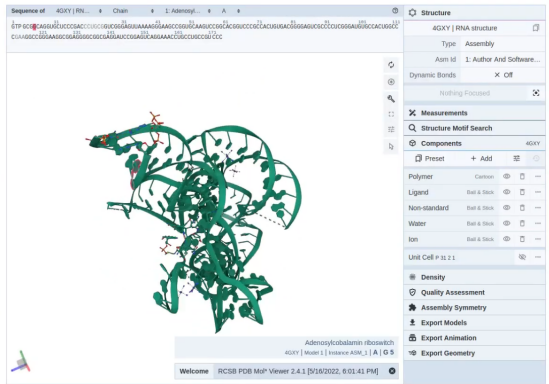
RNANet : RNA 3D structures database (Becquey et al., 2021)











Thank you for your attention

université PARIS-SACLAY | GRADUATE SCHOOL Informatique et Sciences du Numérique

Table of Contents

$$p(native \mid \vec{X}) = \frac{p(native) \prod_{i=1}^{n} p(i \mid native)^{x_i}}{p(\vec{X})} \quad (1)$$

$$score = log\left[\frac{\prod_{i=1}^{n} p(i \mid non-native)^{x_i}}{\prod_{i=1}^{n} p(i \mid native)^{x_i}}\right] + C \quad (2)$$

$$= \sum_{i=1}^{n} x_i \cdot log[p(i \mid non-native)] - \sum_{i=1}^{n} x_i \cdot log[p(i \mid native)] + C$$

$$= - \sum_{i=1}^{n} log\left[\frac{p(i \mid native)}{p(i \mid non-native)}\right] + C$$