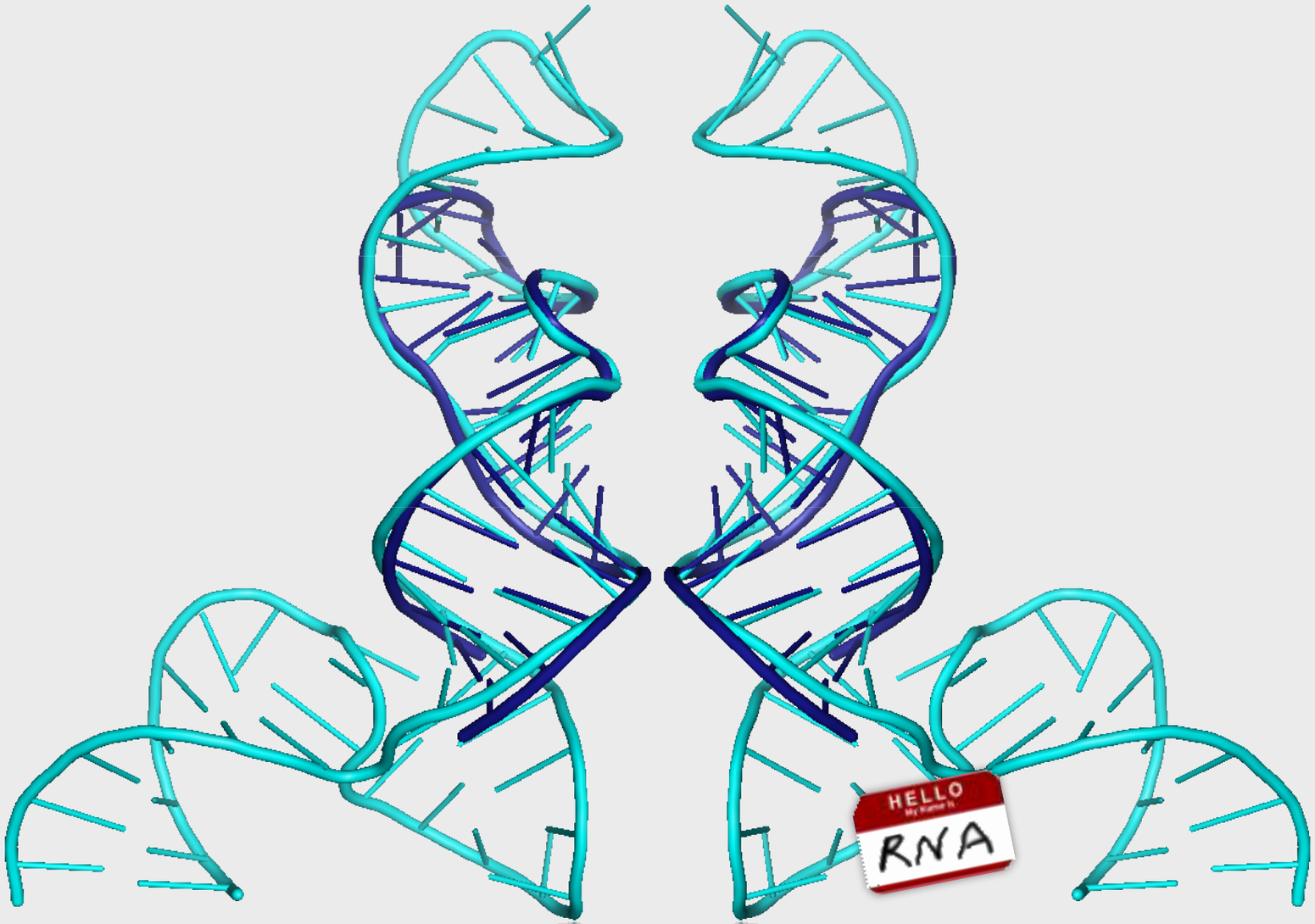
SARA: a tool for RNA structural alignment



Emidio Capriotti

Marc A. Marti-Renom

http://sgu.bioinfo.cipf.es

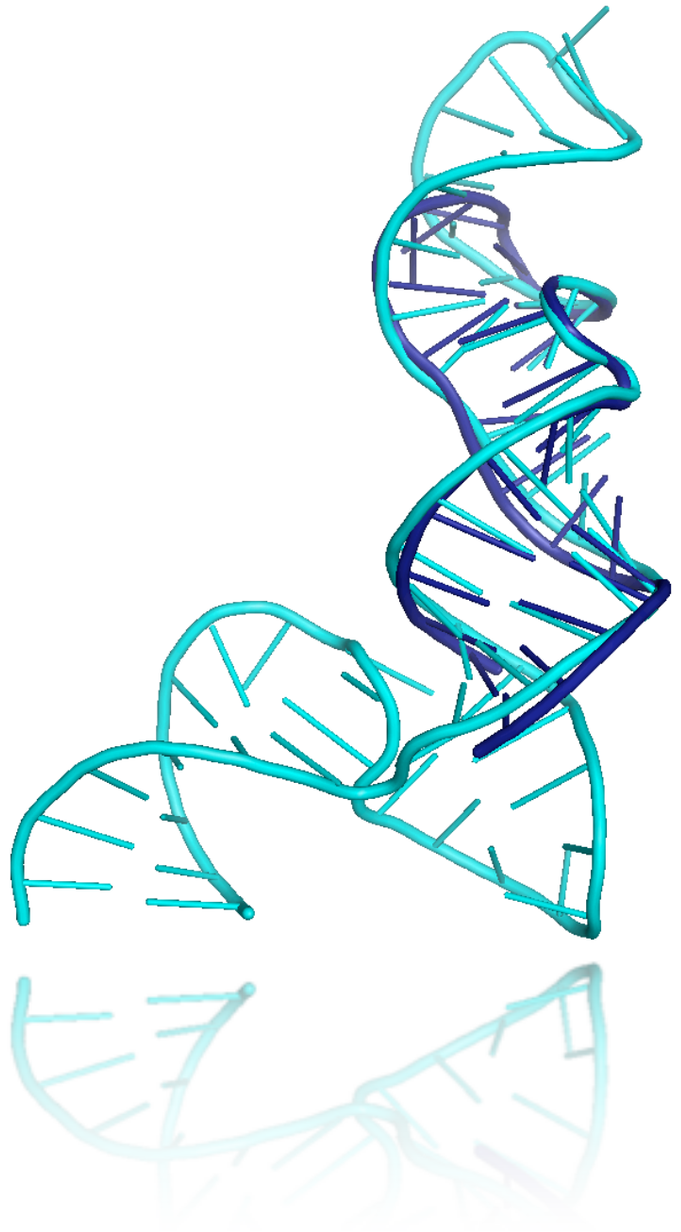
Structural Genomics Unit

Bioinformatics Department

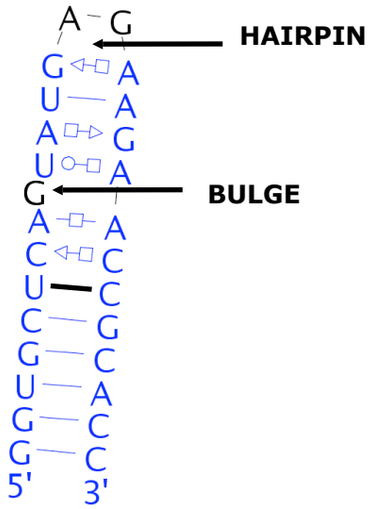
Prince Felipe Resarch Center (CIPF), Valencia, Spain

Summary

- Introduction
- RNA Structural Alignment
  - Problem definition
  - Datasets
  - Structure representation
  - Alignment method
  - Statistical evaluation
- Method
  - Method optimization
  - Results
  - Comparison with ARTS
- Conclusion



RNA structure

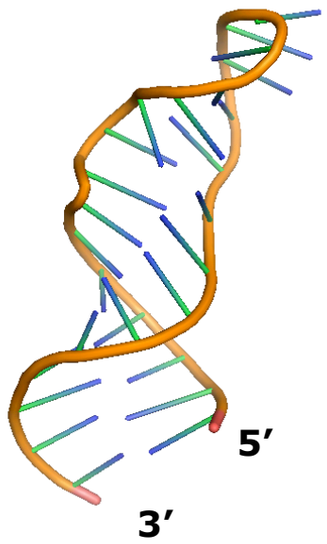


Primary Structure

>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC

Secondary Structure

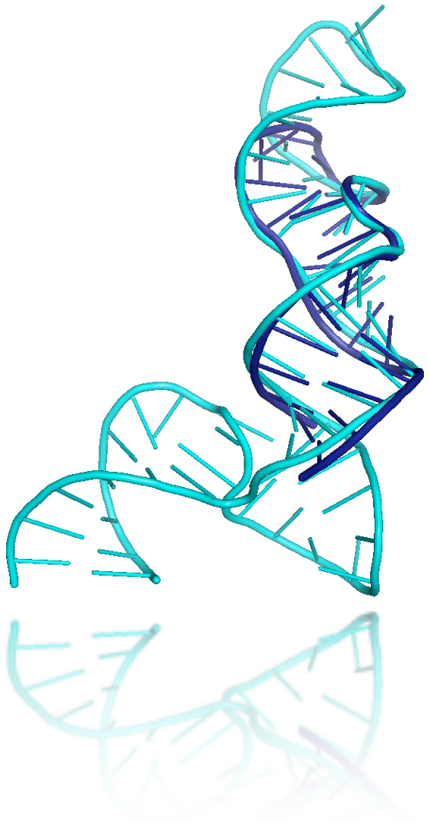
>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC
((((((((.((((..))))))))))))



Tertiary Structure

Secondary Structure interactions and other interactions like pseudoknots, hairpin-hairpin interactions etc.

Structural alignment



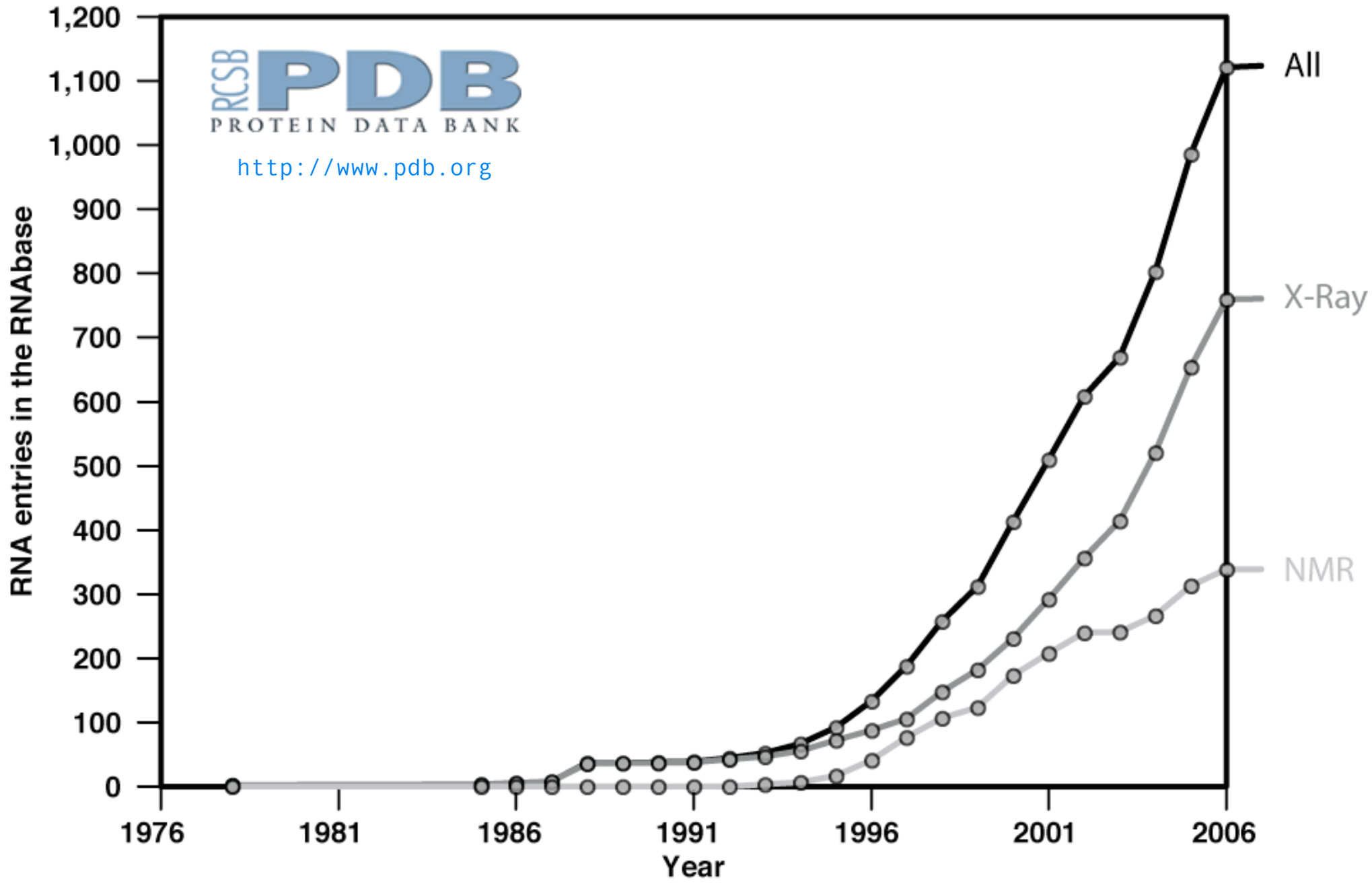
Structural alignment attempts to establish equivalences between two or more polymer structures based on their shape and three-dimensional conformation.

In contrast to simple structural superposition, where at least some equivalent residues of the two structures are known, structural alignment does not require prior knowledge of the equivalent positions.

Structural alignment has been used as a valuable tool for the comparison of proteins, including the inference of evolutionary relationships between proteins of remote sequence similarity.

RNA structure

Today, the PDB database contains more than 1,300 RNA structures.



RNA structure datasets

RNA STRUCTURE*	**1,101**
RNA CHAINS	**2,179**
Non-Redundant RNA CHAINS**	**744**
RNA CHAINS (20≤ Length ≤310)	**313**
HIGH RESOLUTION RNA SET***	**54**

NR95

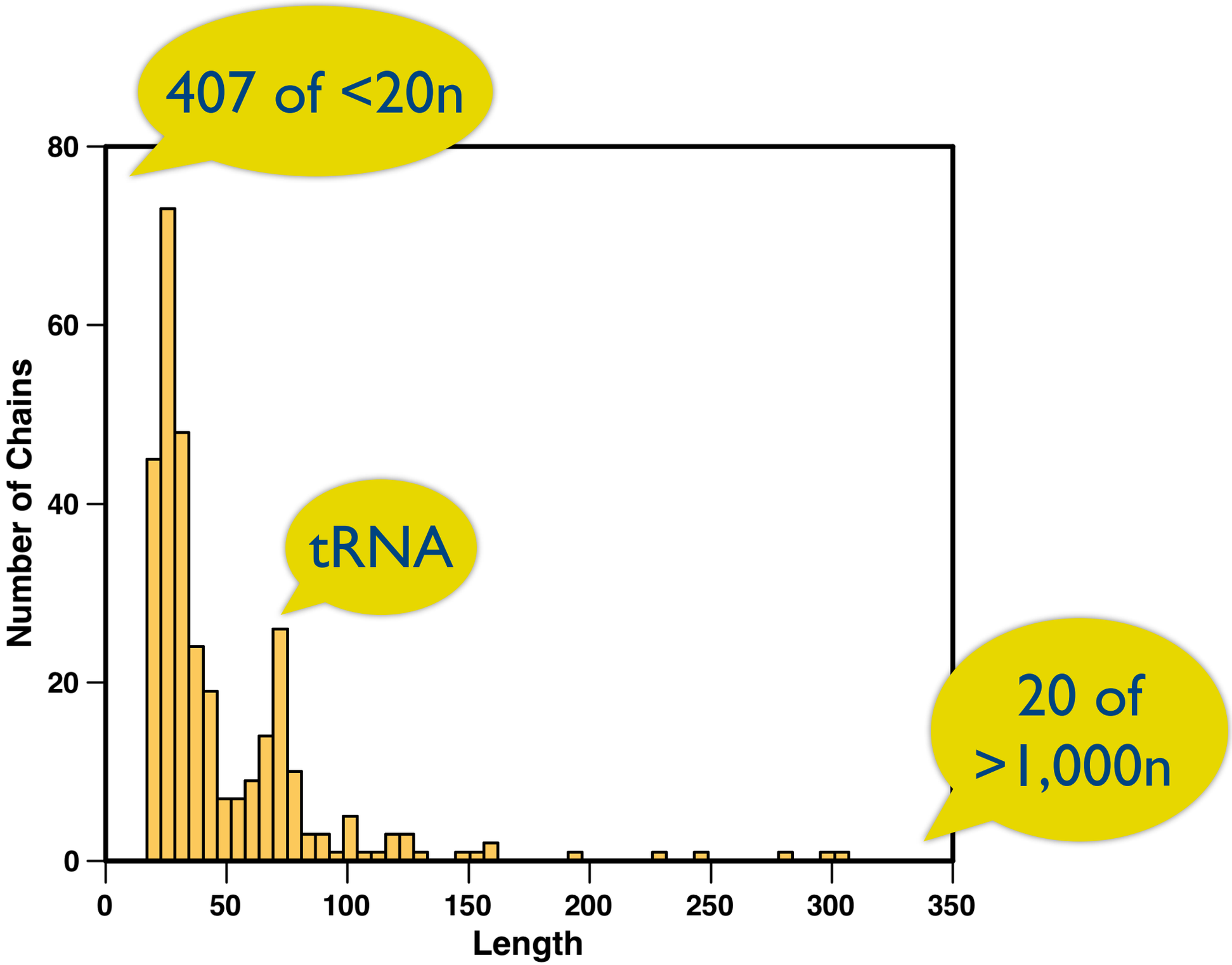
HR

* *from PDB November 06.*

** *non-redundant 95% sequence identity*

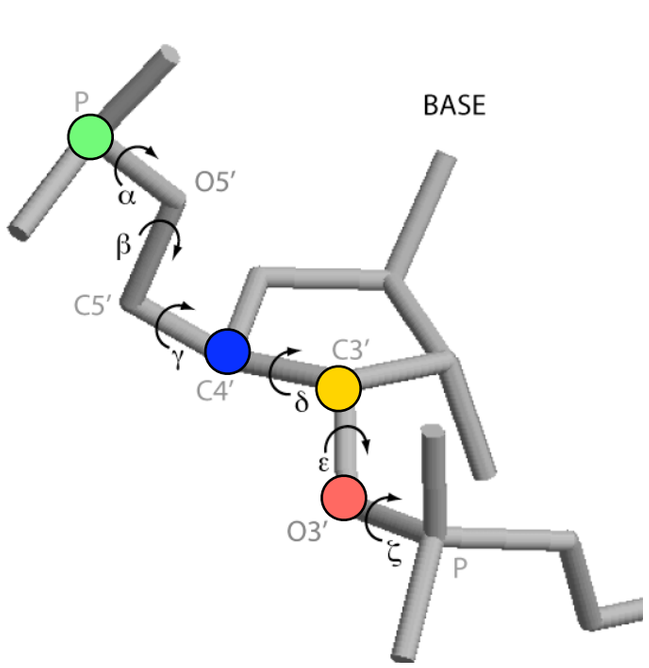
*** *Resolution below 4.0 Å and with no missing backbone atoms.*

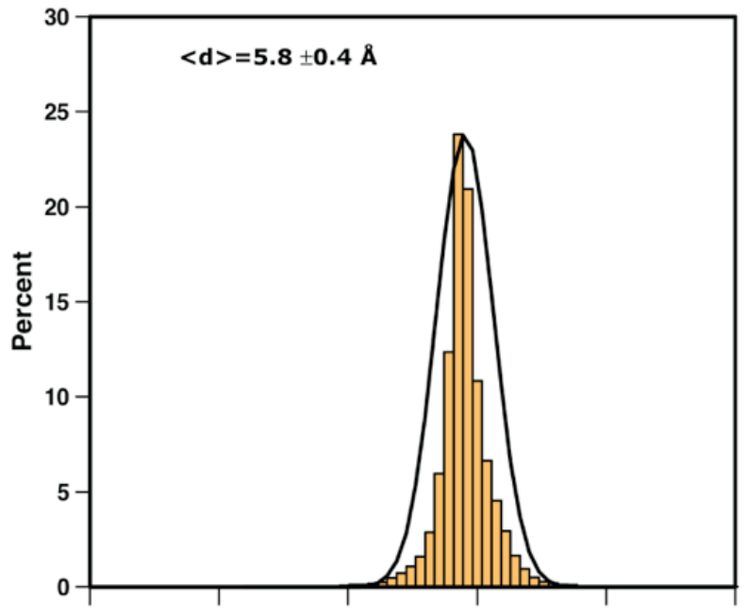
Dataset distribution

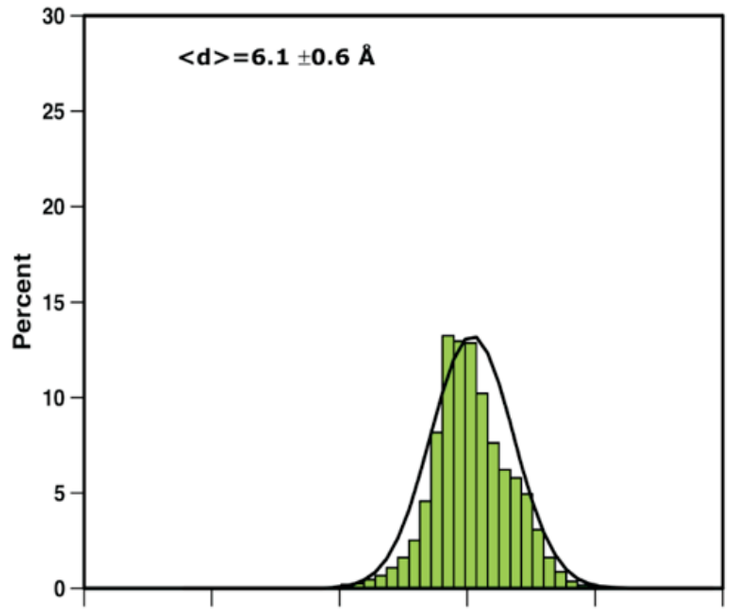


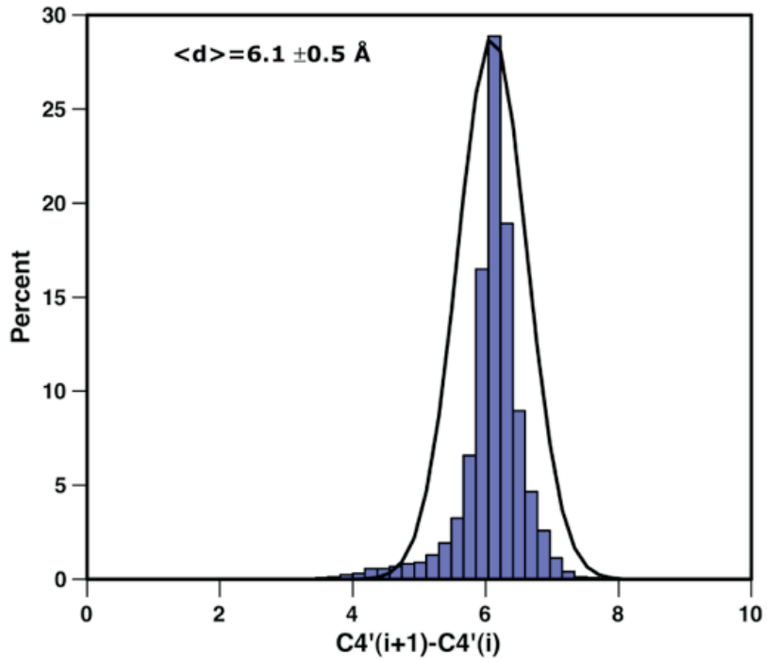
Atom selection

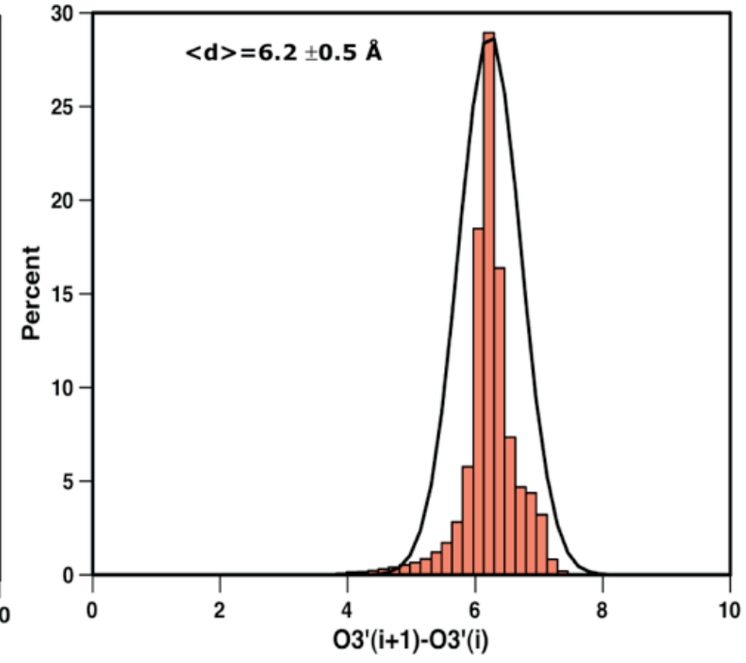
The best backbone atom that represents the RNA structure has been selected by evaluating the distribution of the distances between consecutive atoms in structures from the NR95 set.





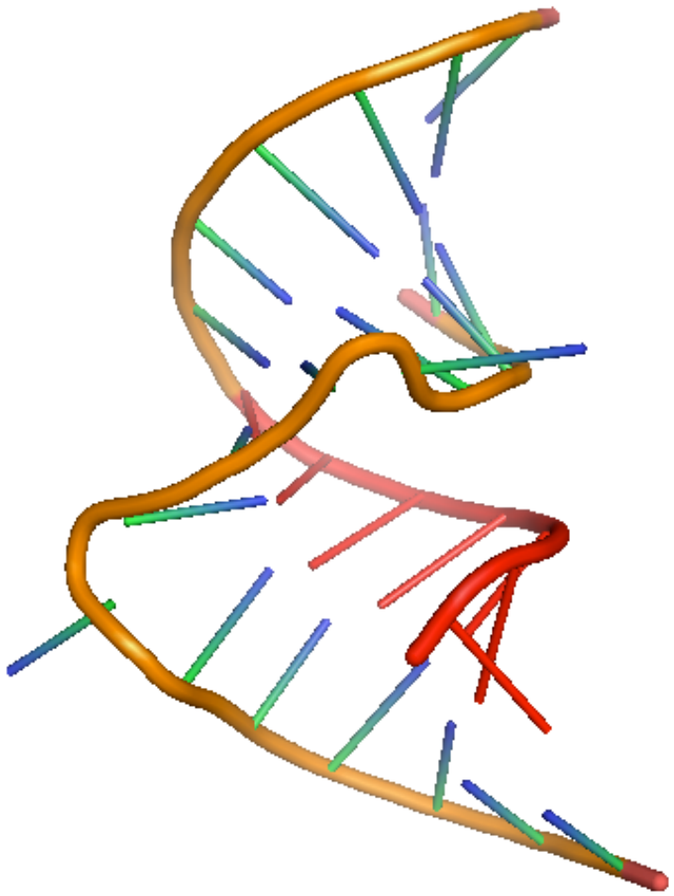


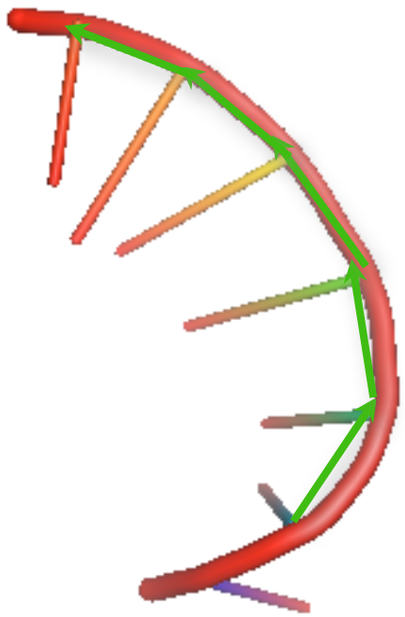


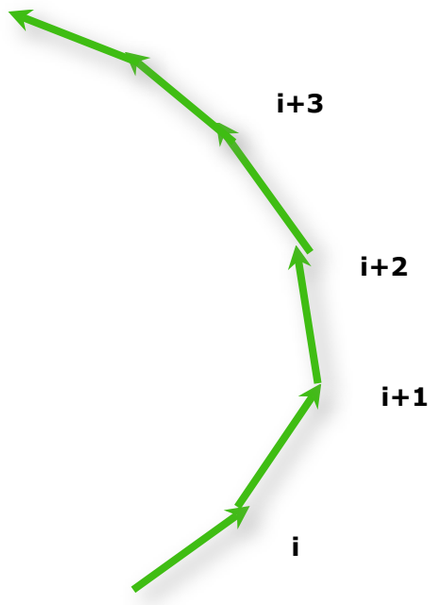


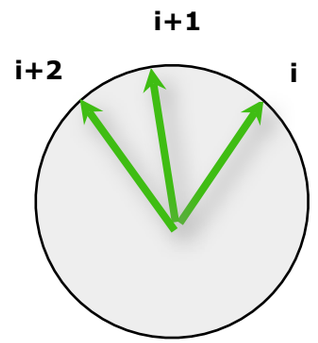
Unit Vector I

Representation









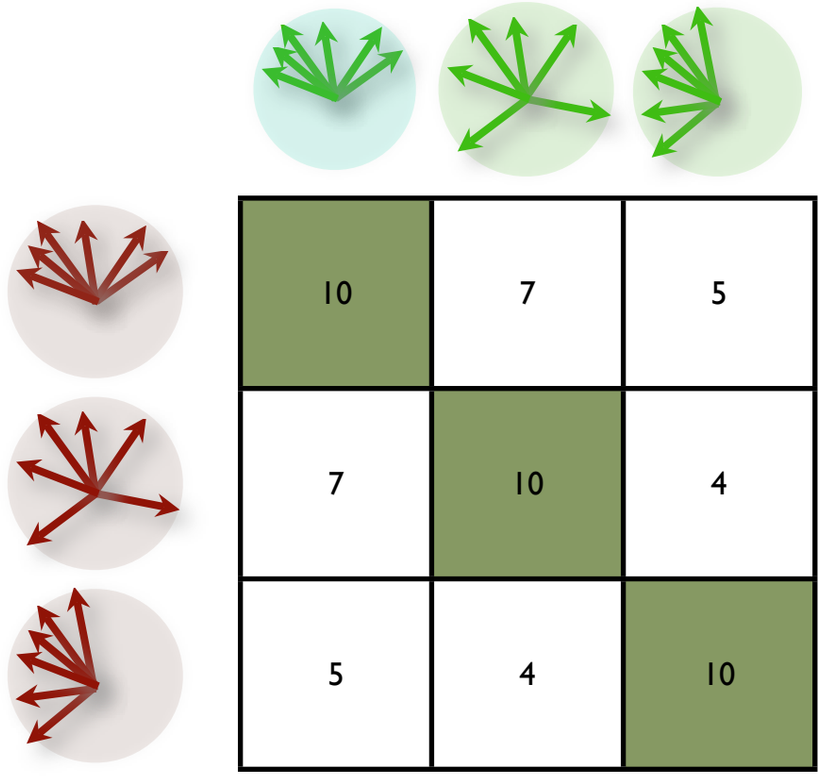
A Unit Vector is the normalized vector between two successive C3’ atoms.

For each position *i* consider the *k* consecutive vectors, which will be mapped into a unit sphere representing the local structure of k residues.

Ortiz et al. *Proteins* 2002

Unit Vector II

Scoring



$$URMS^{R} = \sqrt{2.0 - \frac{2.84}{\sqrt{k}}}$$

$$S_{ij} = \frac{(URMS^{R} - URMS^{ij})}{URMS^{R}} \Delta(URMS^{R}, URMS^{ij})$$

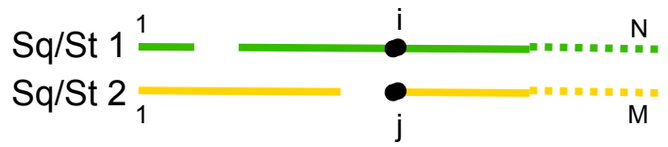
$$\Delta(URMS^{R}, URMS^{ij}) = 10 \Rightarrow URMS^{R} > URMS^{ij}$$
$$\Delta(URMS^{R}, URMS^{ij}) = 0 \Rightarrow URMS^{R} \leq URMS^{ij}$$

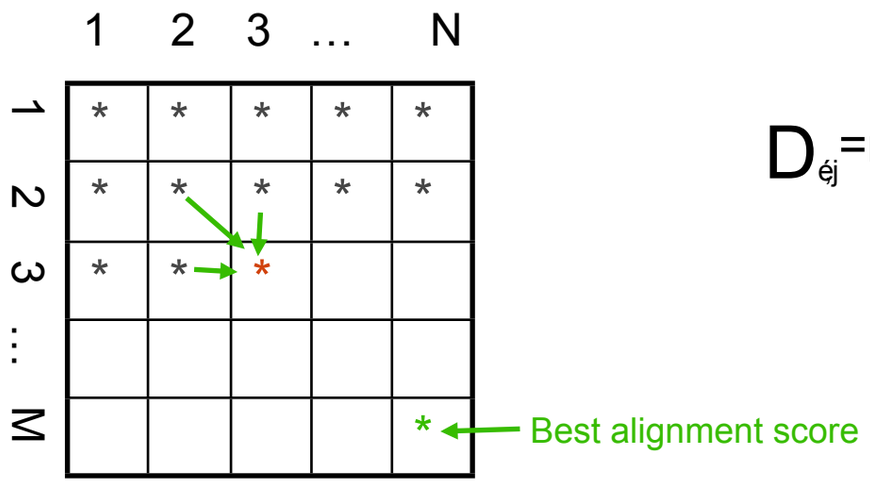
For each position i, the k consecutive unit vectors are grouped and aligned to the j set of unit vectors. Each pair of aligned unit vectors will be evaluated by calculating Unit Root Mean Square distance ($URMS^{ij}$).

The obtained URMS values are compared the minimum expected URMS distance between two random set of k unit vectors ($URMS^{R}$).

The alignment score is than calculated normalizing $URMS^{ij}$ to the $URMS^{R}$ value.

Alignment





$$D_{ij} = \min \begin{cases} D_{i,j-1} + Score_{(-,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,-)} \end{cases}$$

Backtracking to get the best alignment

A Dynamic Programming procedure is then applied to search for the optimal structural alignment using a global alignment with zero end gap penalties.

The maximum subset of local structures that have their corresponding C3' within 3.5 Å in the space are evaluated. The number of close atoms is used to evaluate the percentage of structural identity (PSI) using a variant of the MaxSub algorithm.

Needleman and Wunsch J. Mol.Biol 1970
Siew et al. Bioinformatics 2000

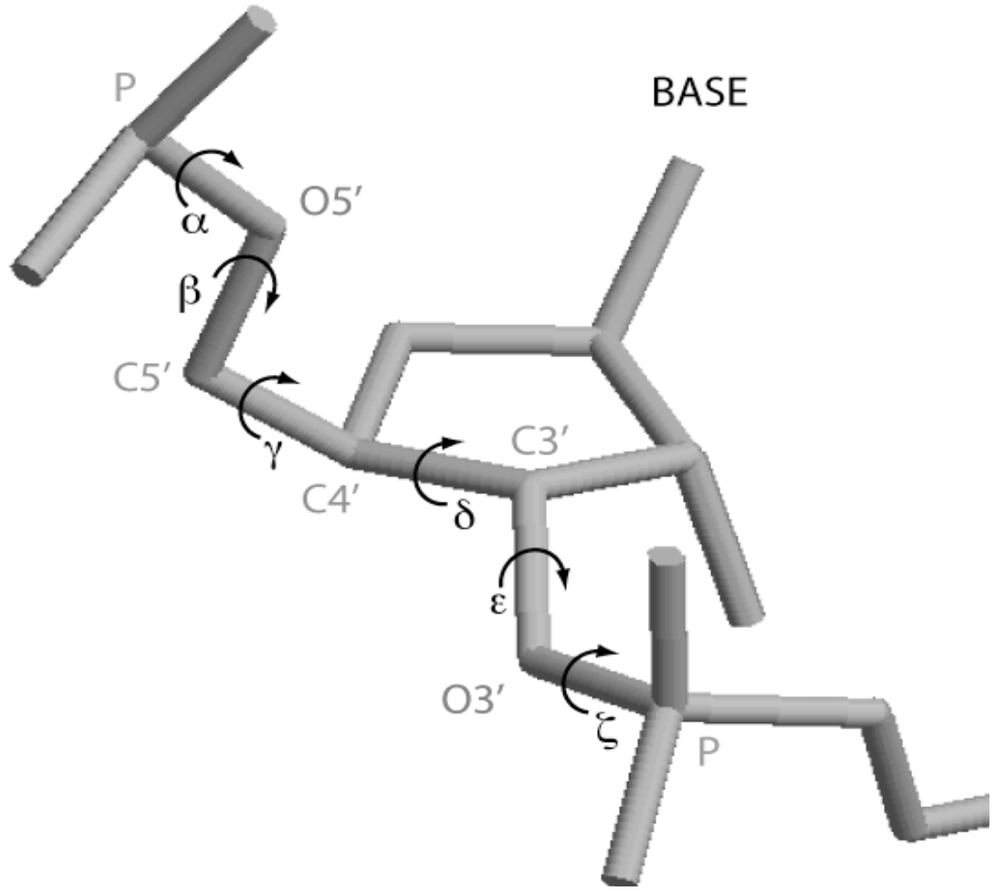
Random RNA structures

In order to build a background distribution that reproduce the scores given by the structural alignments of unrelated RNA sequences, we generated a set 300 random RNA sequences and structures with sequence length uniformly distributed between 20 and 320 nucleotides.

The RNA backbone can be described given the 6 torsion angle (α,β,γ,δ,ε,ζ) for each nucleotide.

The RNA backbone is rotameric and only 42 conformation have been described from a set o high resolution structures .

According to this observation we generated the 300 structures, randomly selecting the backbone angles among the 42 possible conformations.

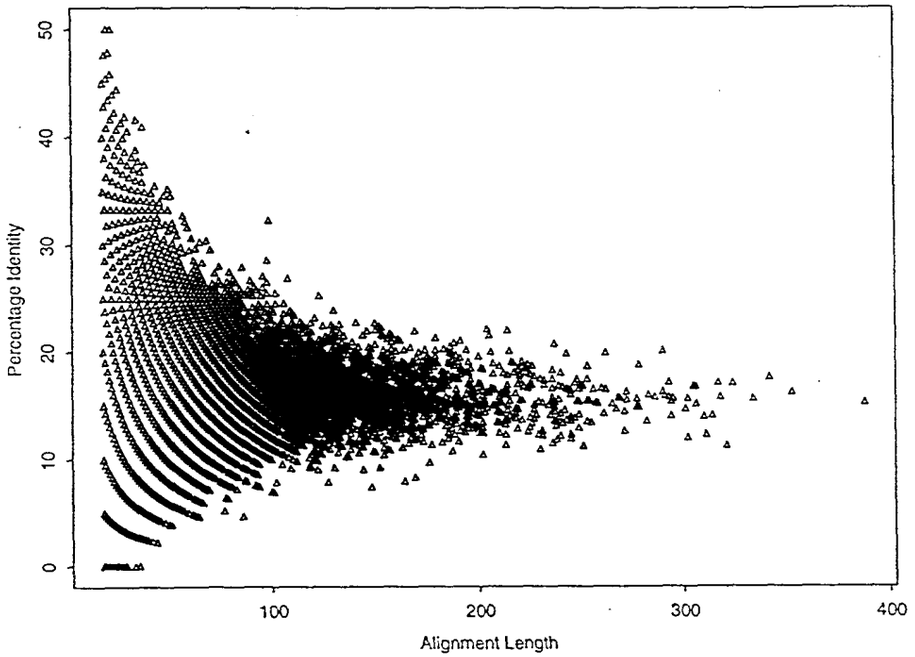


Murray et al PNAS 2003

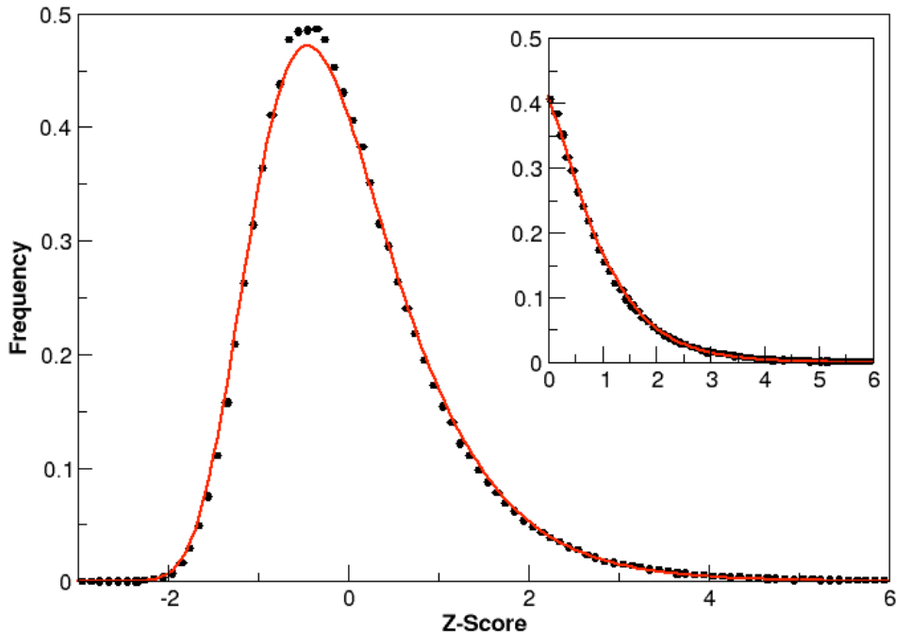
Background distribution

Considering a dataset of 300 random RNA structures, we have produced ~45,000 pairwise alignments that resulted in a empirical distribution. From such distribution we can then evaluate μ and σ needed to calculated the p-value for P(s>=x).

Empirical



Analytic



$$P(s \geq x) = 1 - \exp(-e^{-\lambda(s-\mu)})$$

Karlin and Altschul, 1990 *PNAS* **87**, pp2264

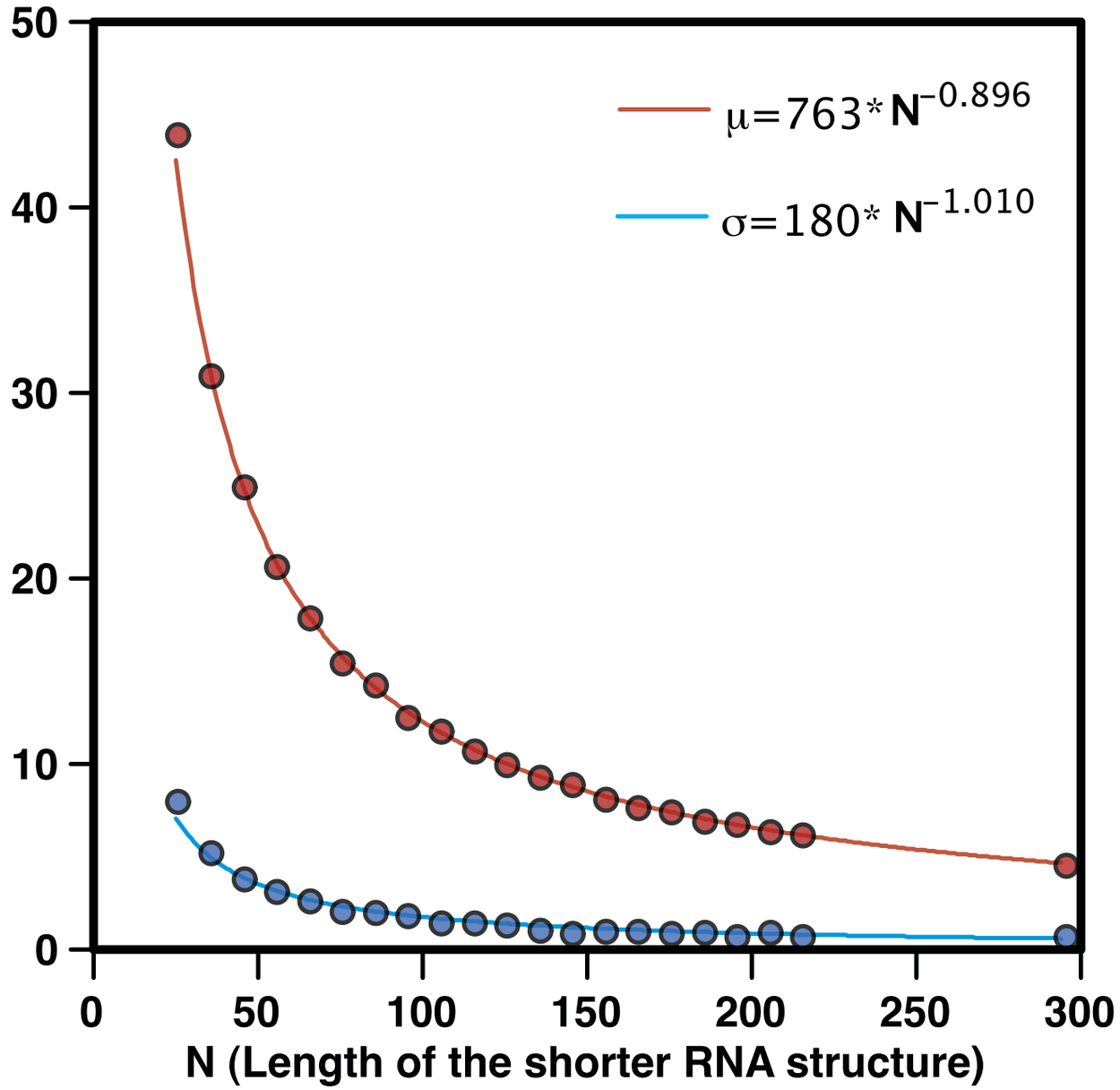
Mean and sigma

The score distribution depends on the length of the molecule.

We divided the resulting structural alignments (~45,000) in 30 bins according to the minimum sequence length of the two random structures (N).

For each bin the μ and σ values are evaluated fitting the data to an EVD.

The relations between N and μ, σ values are extrapolate fitting them to a power low function (r≈0.99).



Optimization

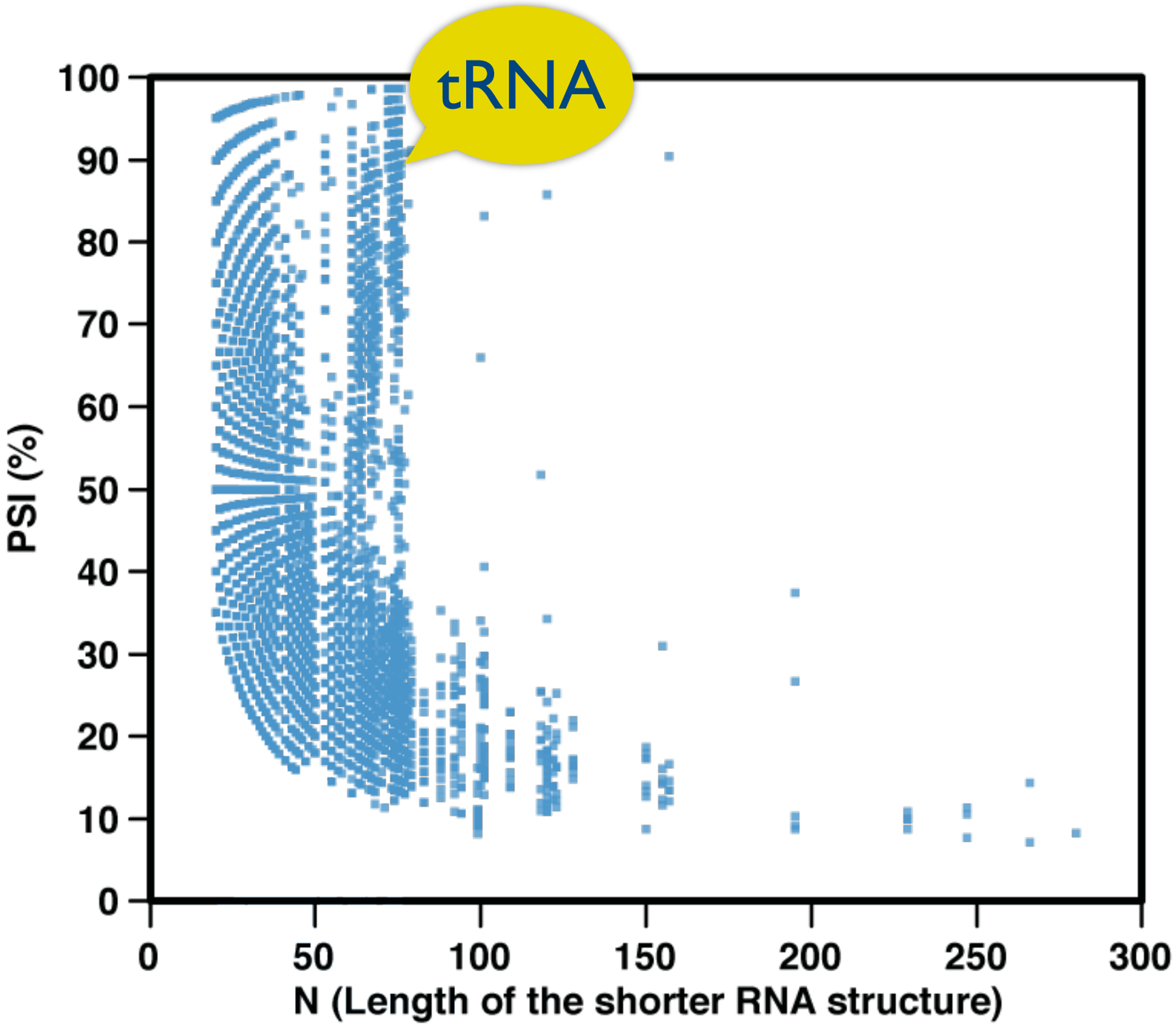
The accuracy of the method here presented depends of a large number of parameters. We optimized the method performing a grid-like search, over about 49,000 possible alignments between the chains in NR95 set, considering:

- C3’ and P backbone atoms for the unit vectors evaluation,
- k number of consecutive unit vectors, spamming from 3 to 9 and,
- values of gap opening from -8 to -6 and gap extension for -1.0 to –0.2

The best parameters corresponded to the use of 7 consecutive C3’ atoms using an opening gap penalty of -7.0 and extension gap penalty of -0.45.

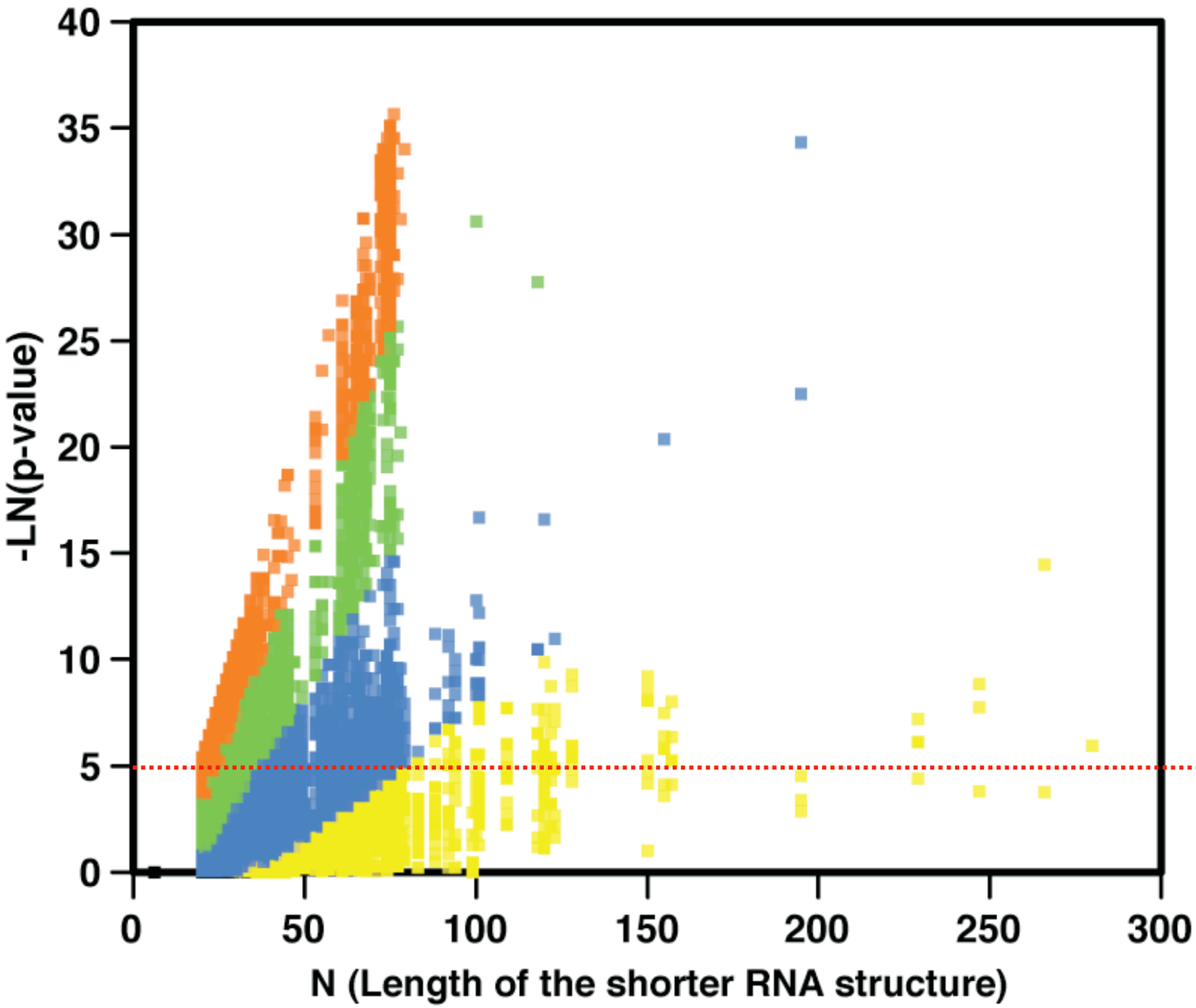
PSI distribution

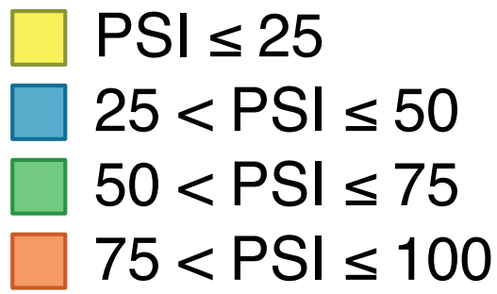
all-against-all comparison of structures in the NR95 set

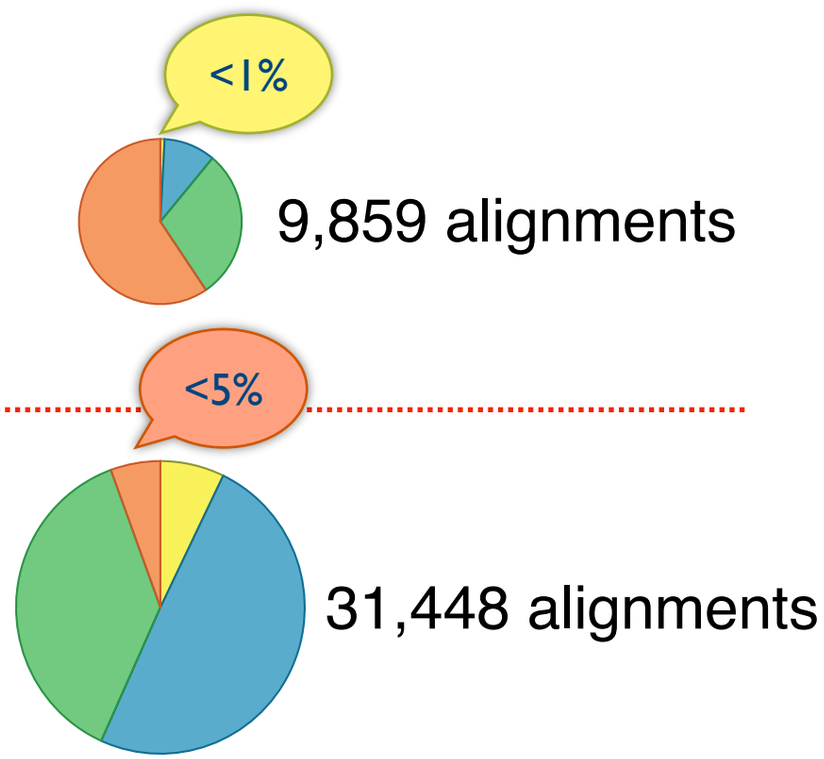


Statistical significance

all-against-all comparison of structures in the NR95 set

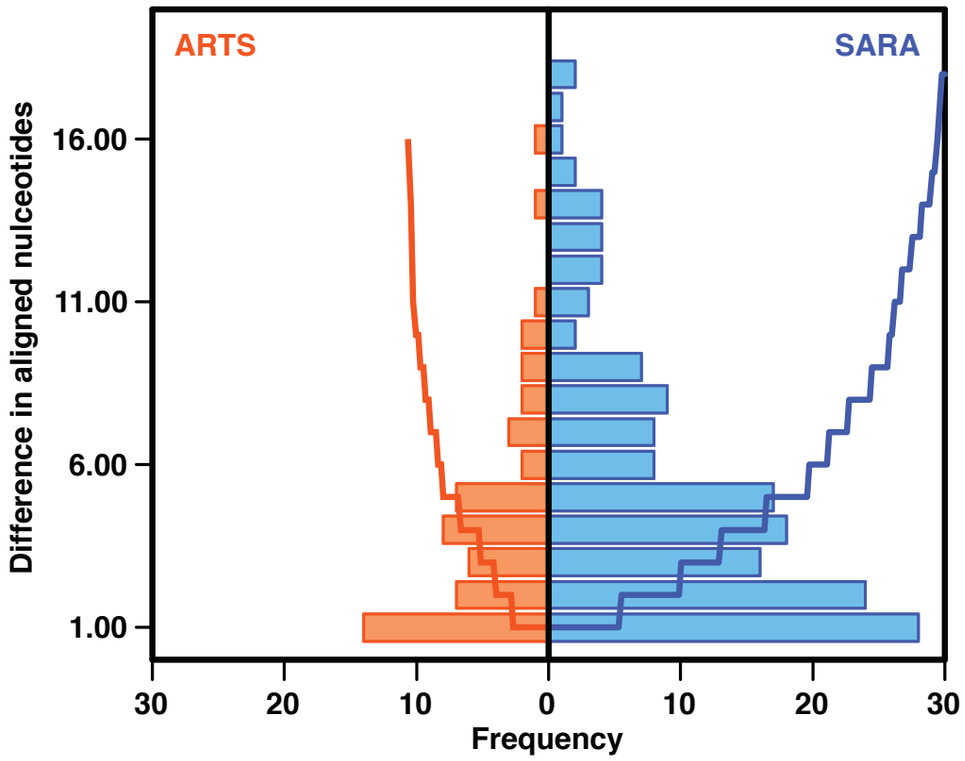


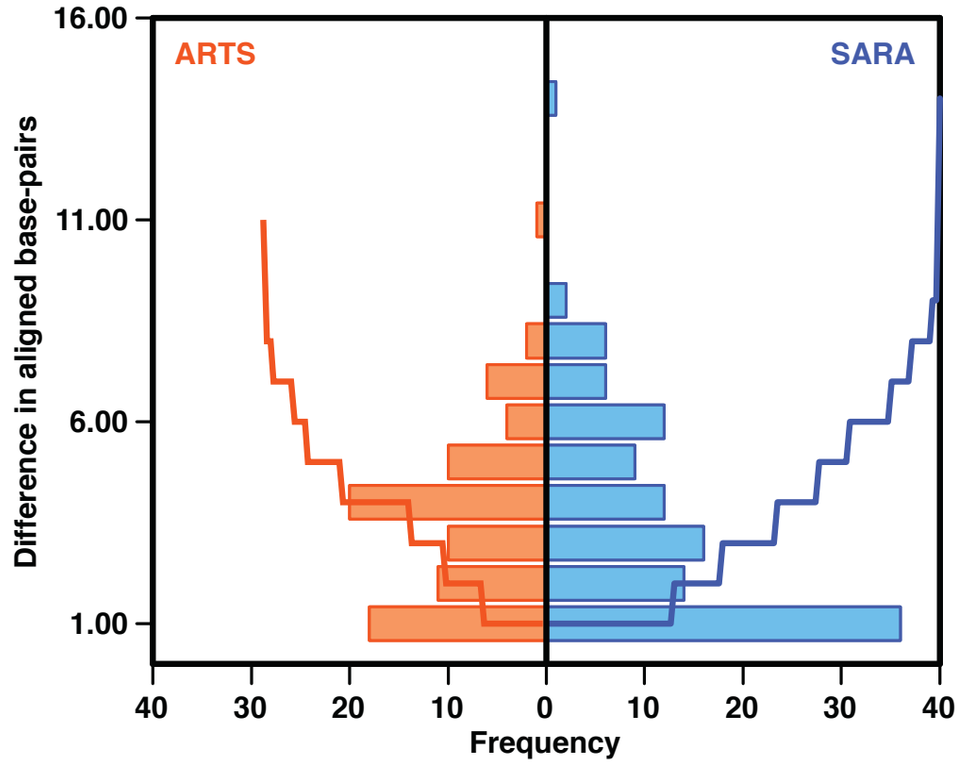




Comparison with ARTS

all-against-all comparison of structures in the HR set

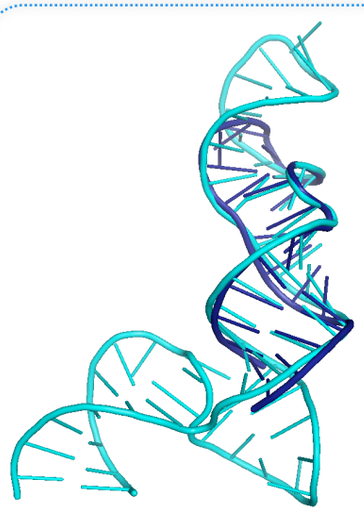




ARTS

Percentage of structural identity (PSI) 76.9%
Percentage of sequence identity 25.0%
Percentage of SSE identity 87.5%
RMSD 3.54Å

```
>1q96 Chain:A
--------------------gugcucag-uaugaga------aga--accgcacc--------
>1un6 Chain:E
ccggccacaccuacggggccugguua-guaccug-ggaaaccu-gggaauaccaggugccggc
```



SARA

Percentage of structural identity (PSI) 92.6%
Percentage of sequence identity 48.0%
Percentage of SSE identity 100.0%
RMSD 2.12Å

```
>1q96 Chain:A
--------------------ggugcucaguaugag---------aagaaccgcacc-------
>1un6 Chain:E
gccggccacaccuacggggccugguuaguacc-ugggaaaccugggaauaccaggugccggc
```

Conclusions

- The C3'–trace is a good representation of the RNA structure.

- The all-against-all alignments among the 300 random RNA structures provides a good set for generating a background distribution needed for calculating a p-value significance of the alignments. P-values larger than 5 are useful to detect reliable alignments.

- Our algorithm results in higher accuracy alignments than those produced by ARTS. For 226 pairs of structures that aligned with a -LN(p-value) > 5.0, SARA results in ~45% of alignments with higher number of aligned nucleotides and ~14% with higher number of aligned base-pairs than those by ARTS.

Acknowledgments

Structural Genomics Unit (CIPF)
Marc A. Marti-Renom
Emidio Capriotti
Peio Ziarsolo Areitioaurtena

Comparative Genomics Unit (CIPF)
Hernán Dopazo
Leo Arbiza
Francisco García

Functional Genomics Unit (CIPF)
Joaquín Dopazo
Fátima Al-Shahrour
José Carbonell
Ignacio Medina
David Montaner
Joaquin Tárraga
Ana Conesa
Toni Gabaldón
Eva Alloza
Lucía Conde
Stefan Goetz
Jaime Huerta Cepas
Marina Marcet
Pablo Minguez
Jordi Burguet Castell
Pablo Escobar

ARTS PROGRAM
Orinat Dror
Ruth Nussinov
Haim J. Wolfson

FUNDING
Prince Felipe Research Center
Marie Curie Reintegration Grant
STREP EU Grant
Generalitat Valenciana
MEC-BIO

http://bioinfo.cipf.es
http://sgu.bioinfo.cipf.es