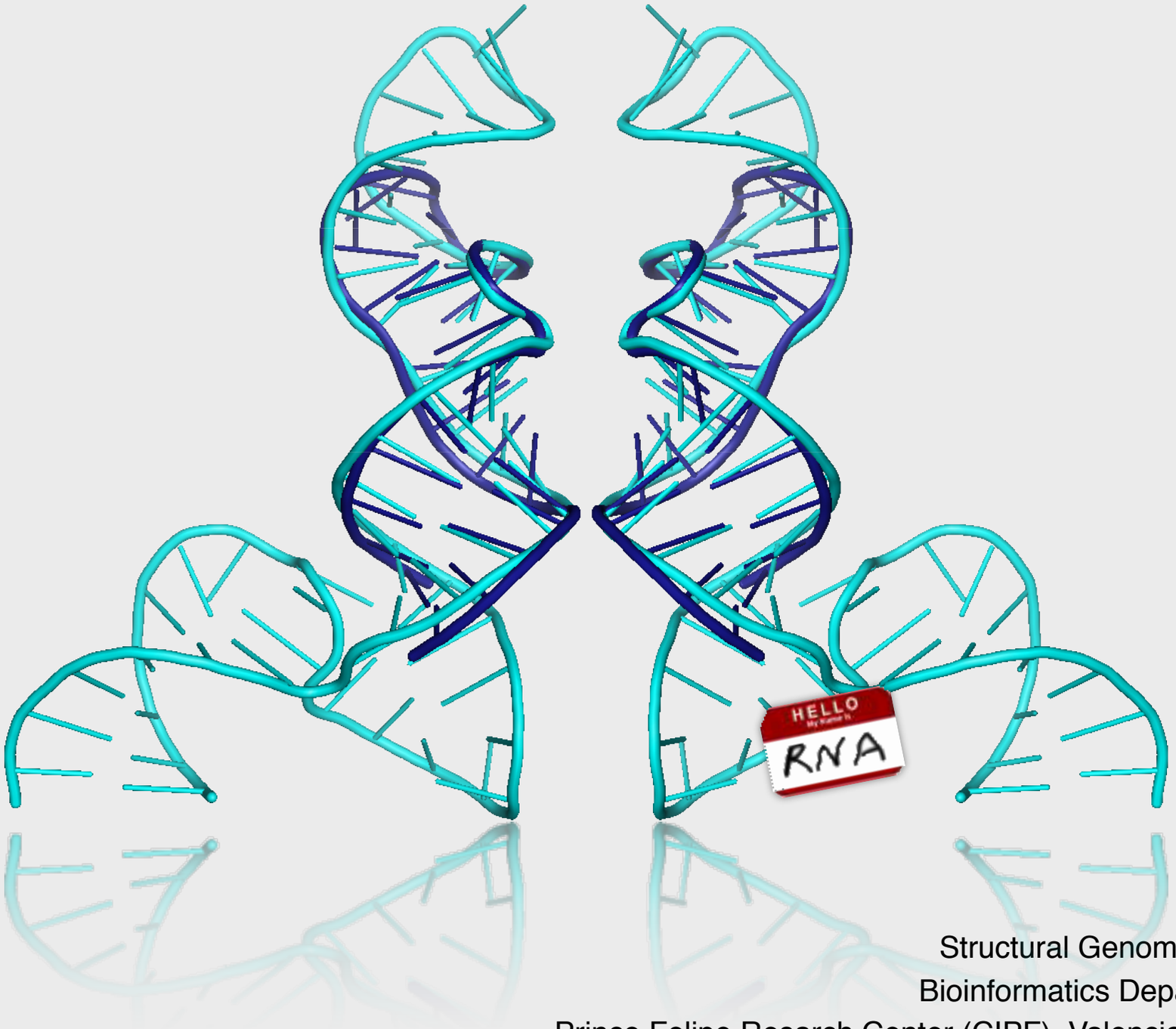
SARA: a tool for RNA structure alignment



Emidio Capriotti

Marc A. Marti-Renom

http://sgu.bioinfo.cipf.es

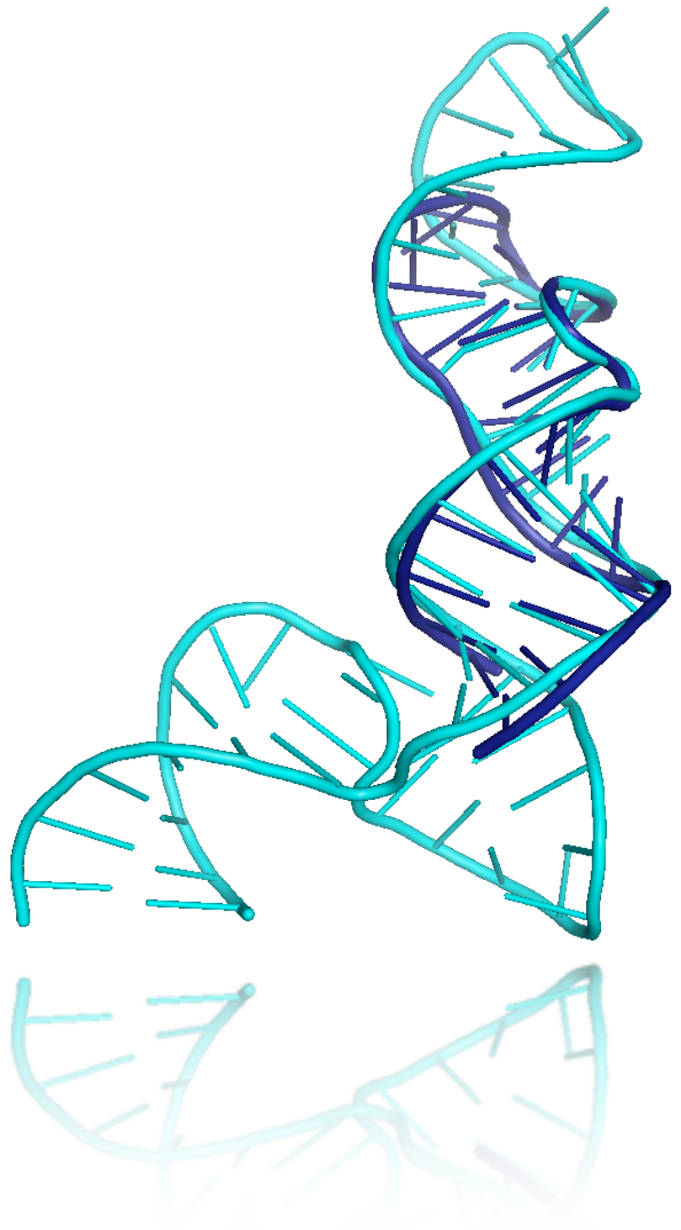
Structural Genomics Unit

Bioinformatics Department

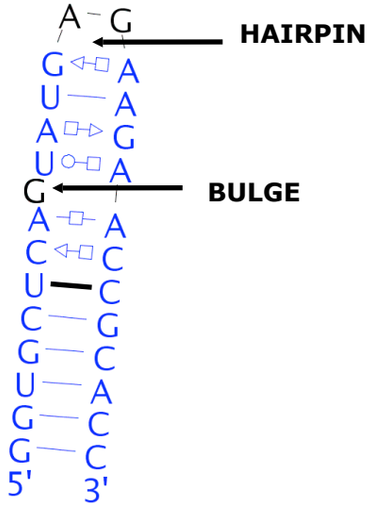
Prince Felipe Resarch Center (CIPF), Valencia, Spain

Summary

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



RNA structure

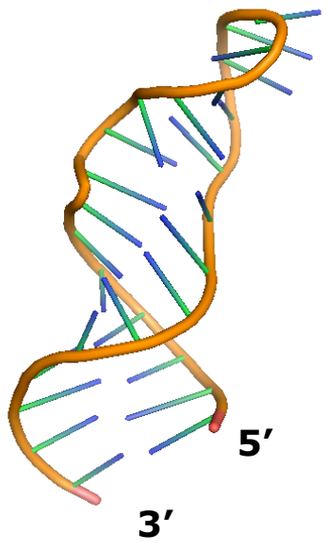


Primary Structure

>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC

Secondary Structure

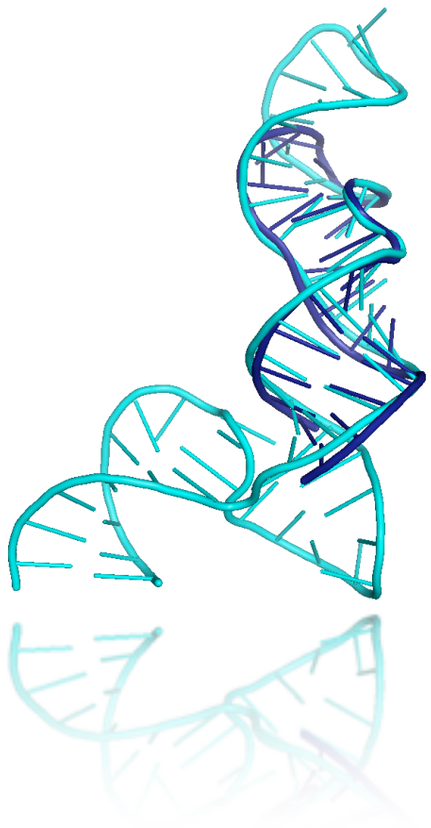
>Mutant Rat 28S rRNA sarcin/ricin domain
GGUGCUCAGUAUGAGAAGAACCGCACC
((((((((.((((..))))))))))))



Tertiary Structure

Secondary structure interactions and other interactions such as pseudoknots, hairpin-hairpin interactions, etc.

Structural alignment



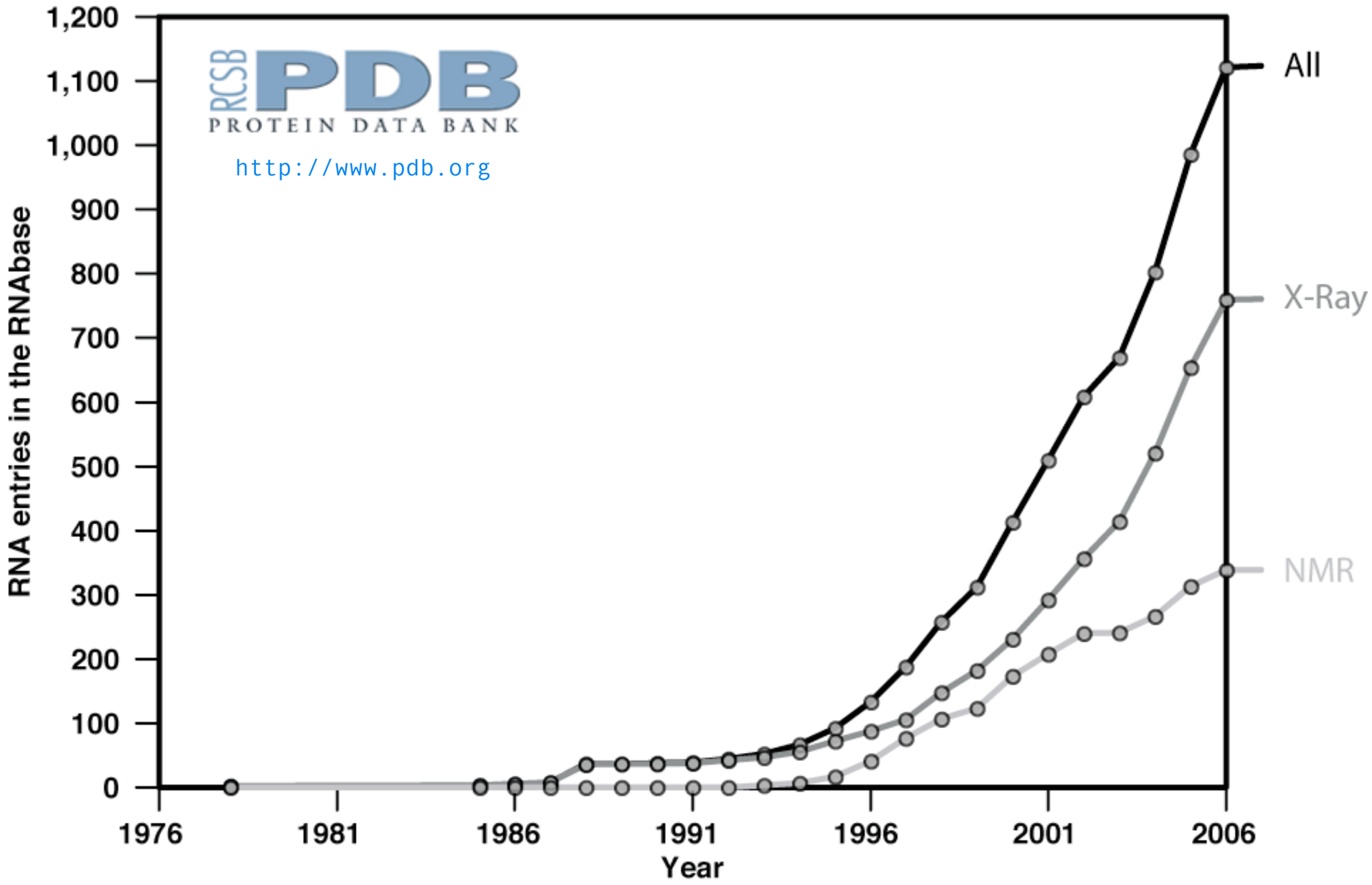
Structural alignment attempts to establish equivalences between two or more polymer structures based on their shape and three-dimensional conformation.

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**	**708**
RNA CHAINS (20≤ Length ≤310)	**277**
SCOR SET***	**60**
HIGH RESOLUTION RNA SET****	**51**

NR95

SCOR

HR

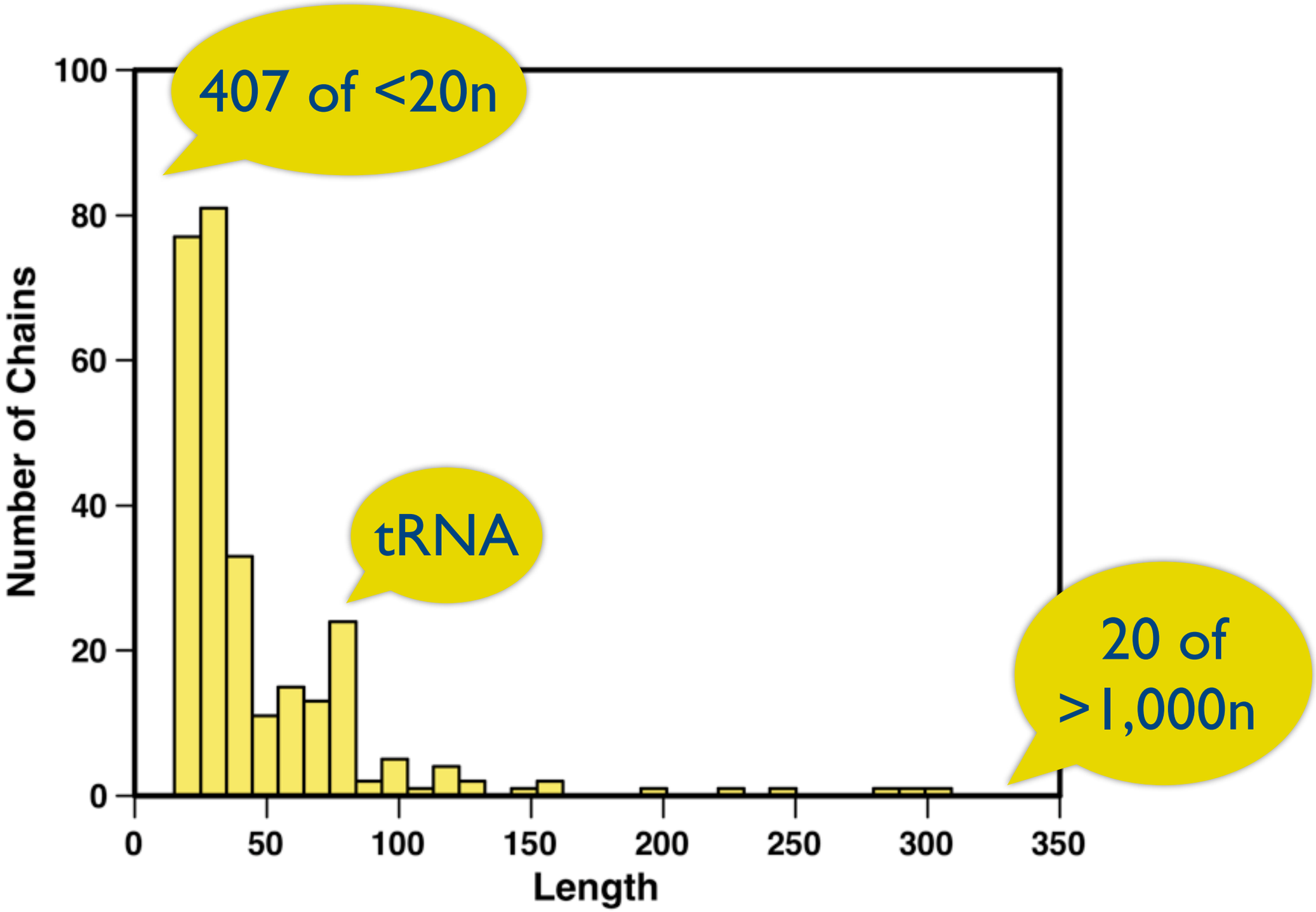
* *from PDB November 06.*

** *non-redundant 95% sequence identity*

*** *SCOR functions with at least two chains*

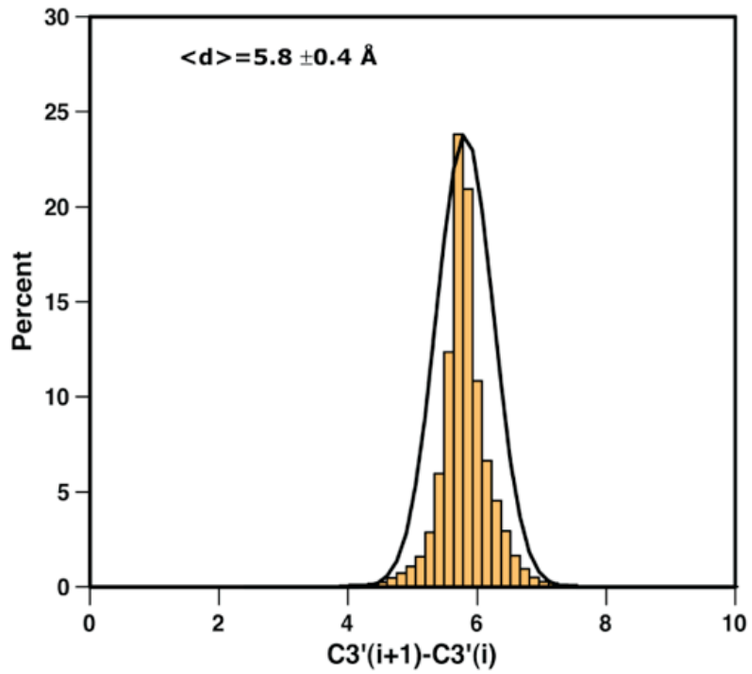
**** *resolution below 4.0 Å and with no missing backbone atoms.*

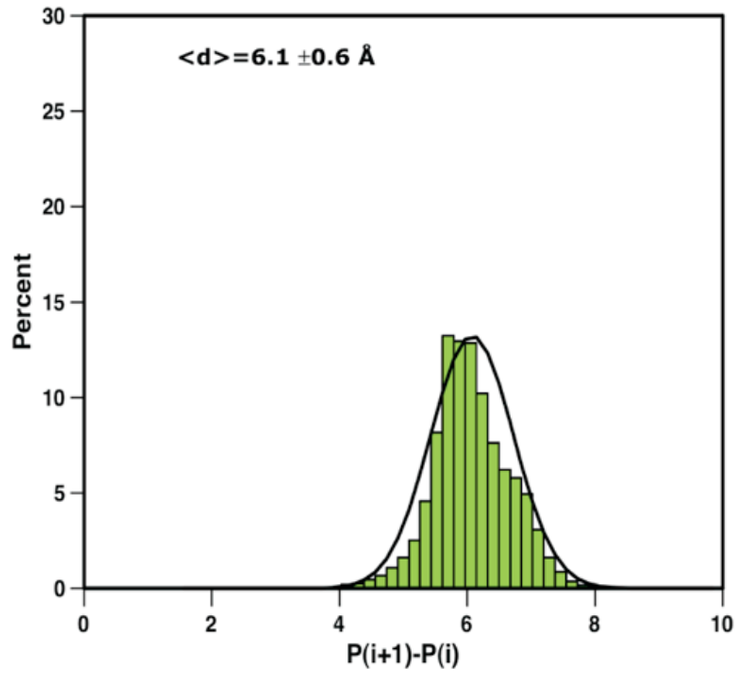
Dataset distribution

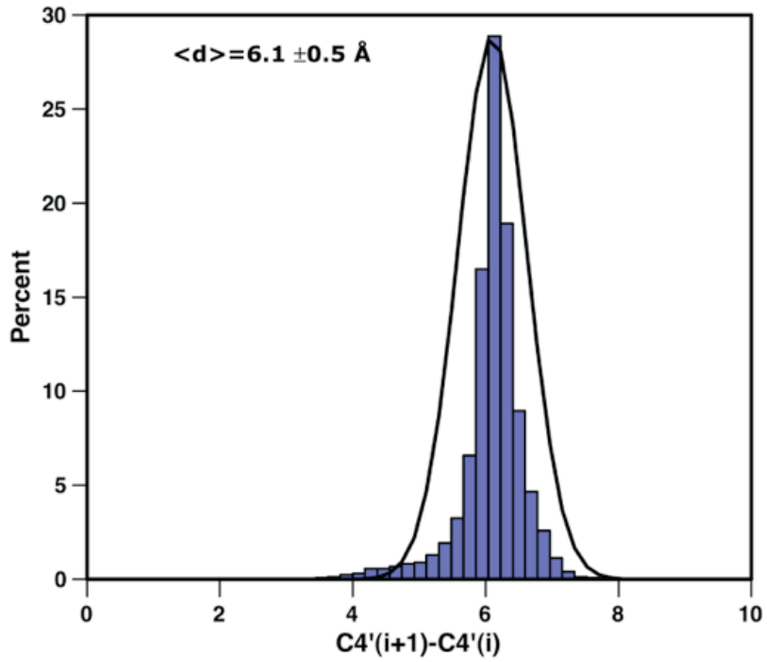


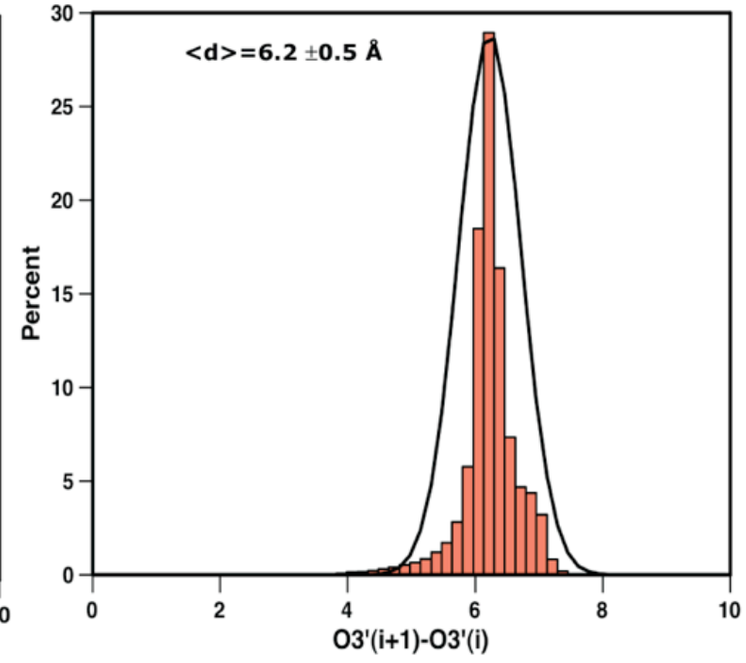
Atom selection

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.



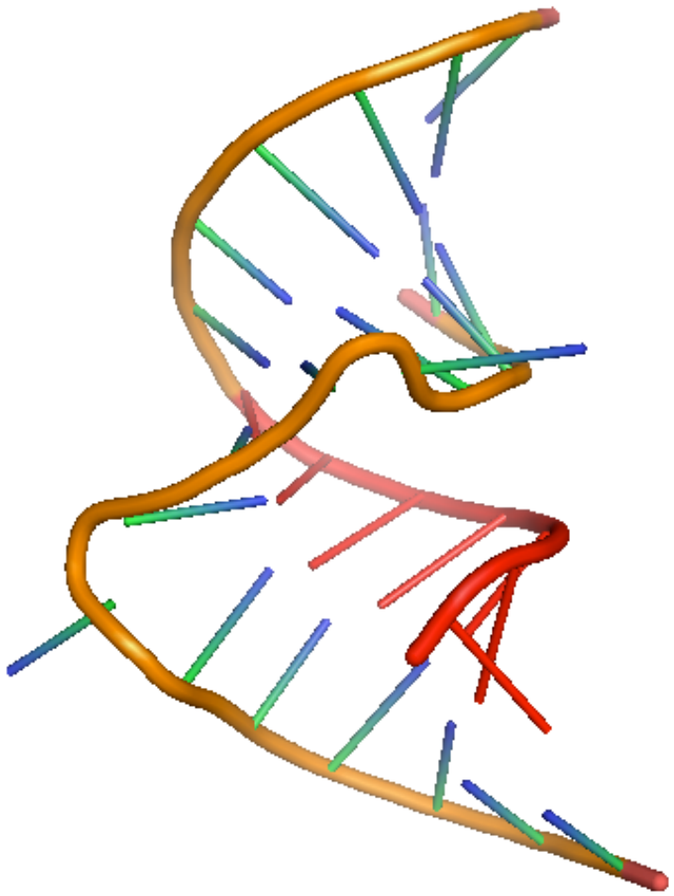


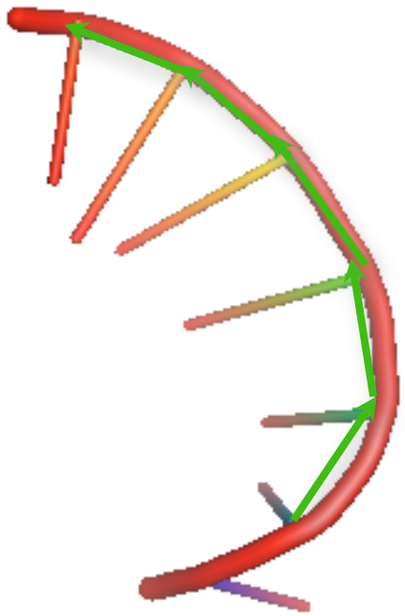


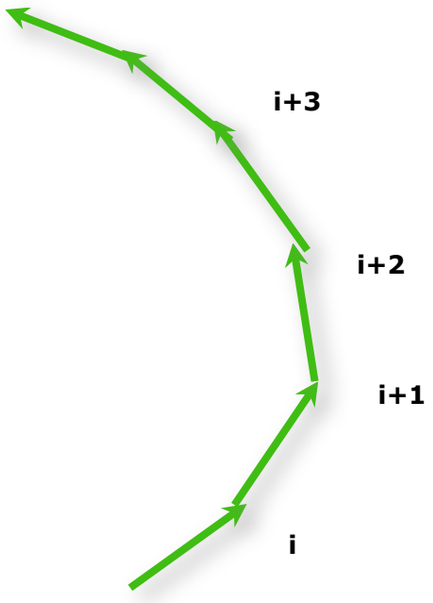


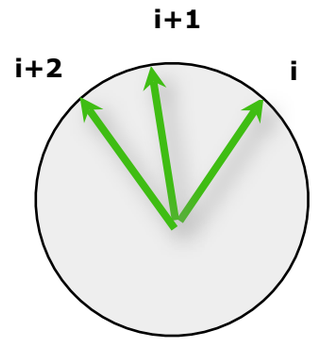
Unit Vector I

Representation









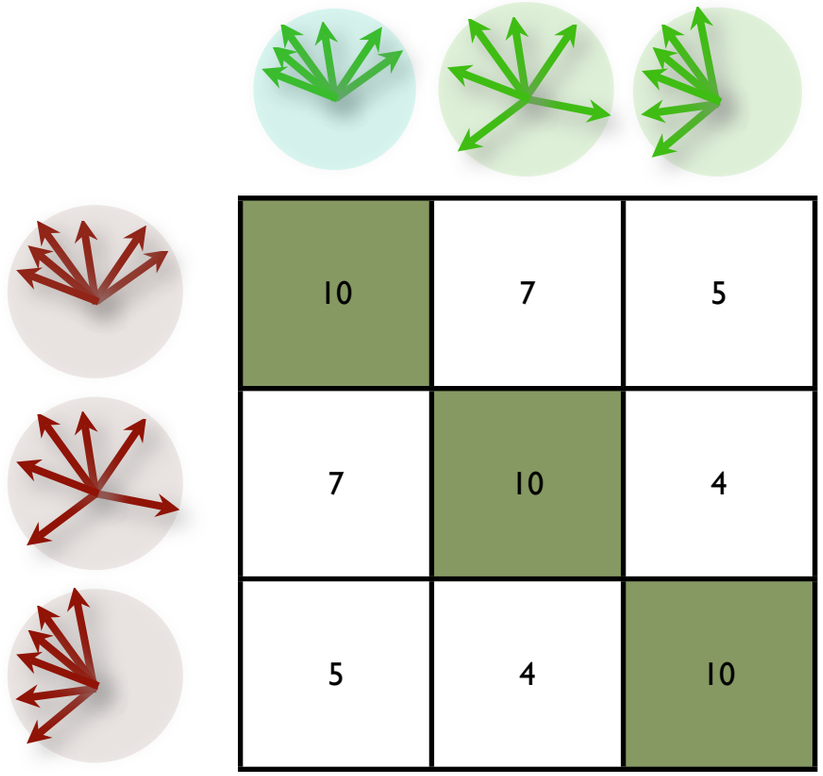
A Unit Vector is the normalized vector between two successive atoms of the same type.

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



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

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

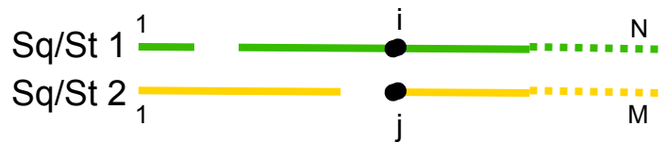
$$\Delta(URMS^{R}, URMS^{ij}) = 10 \Rightarrow URMS^{R} > URMS^{ij}$$
$$\Delta(URMS^{R}, URMS^{ij}) = 0 \Rightarrow URMS^{R} \leq URMS^{ij}$$

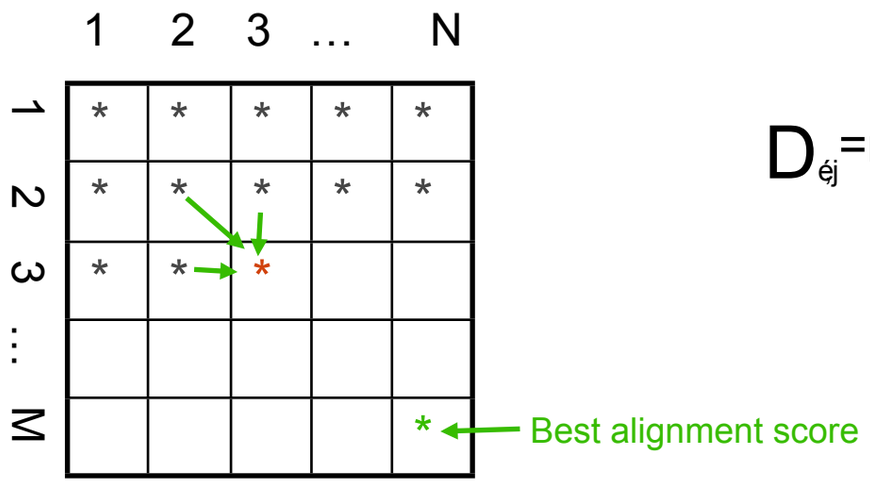
For each position *i*, the *k* consecutive unit vectors are grouped and aligned to the *j* set of unit vectors. Each pair of aligned unit vectors will be evaluated by calculating Unit Root Mean Square distance ($URMS^{ij}$).

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 then 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 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 equivalent selected atoms within 4.0 Å in the space are calculated using a variant of the MaxSub algorithm. For each alignment the number of close atoms is used to evaluate the percentage of structural identity (PSI).

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

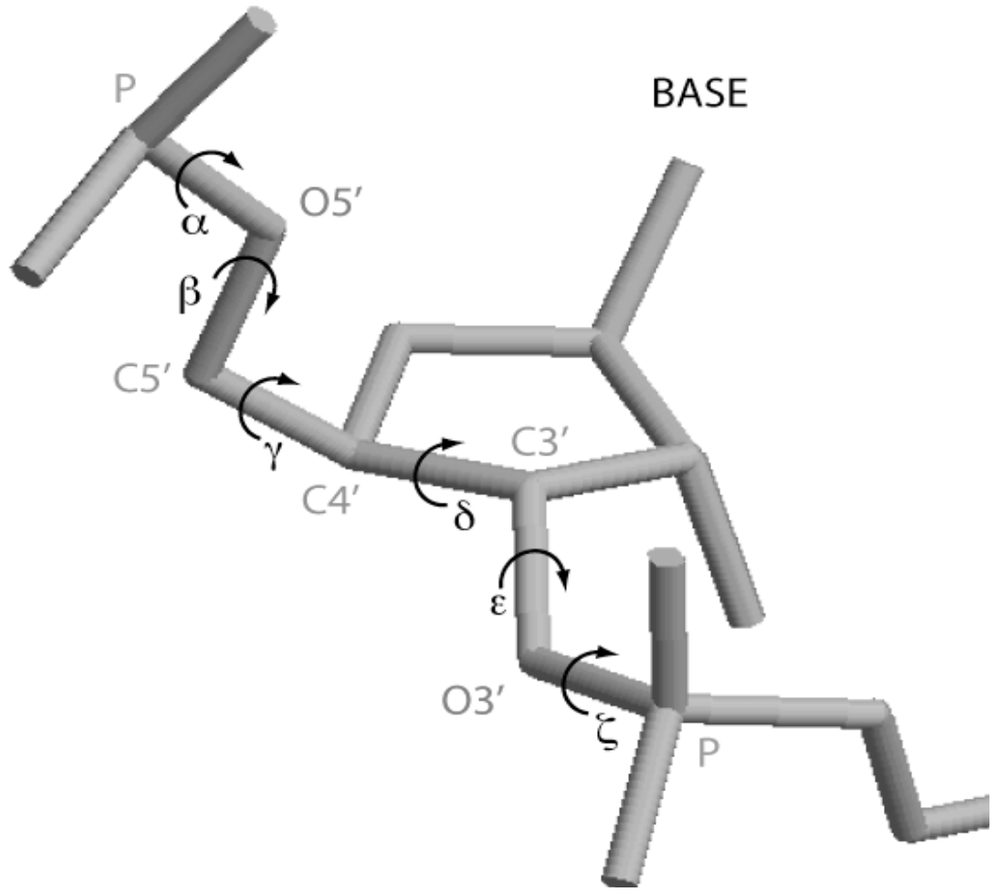
Random RNA structures

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.

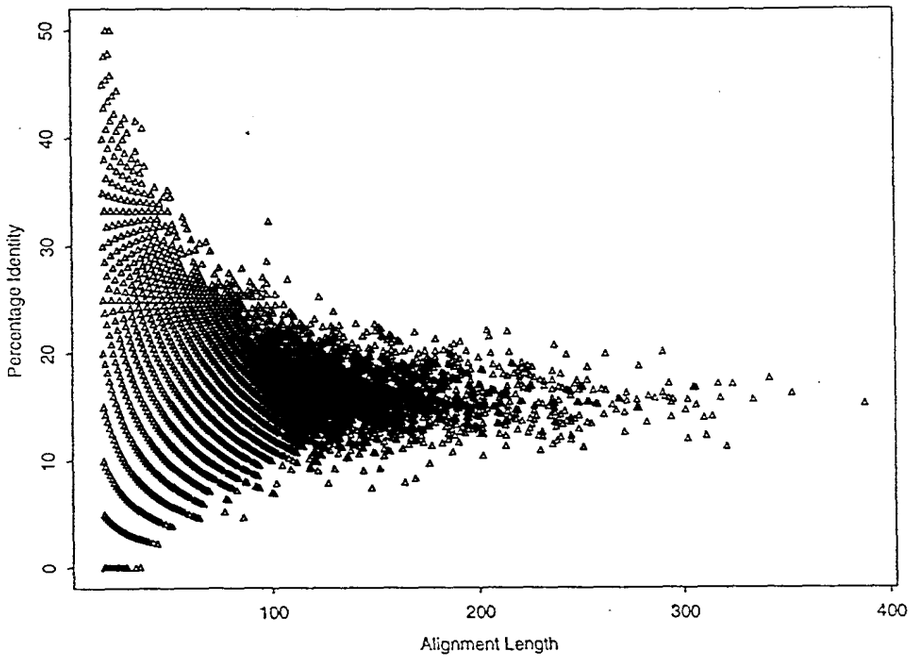


Murray et al PNAS 2003

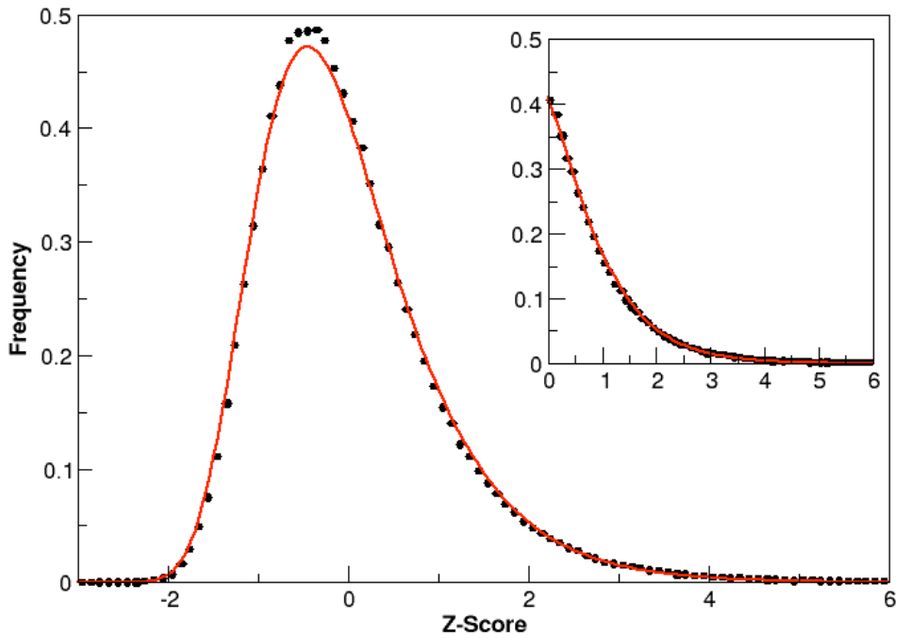
Background distribution

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

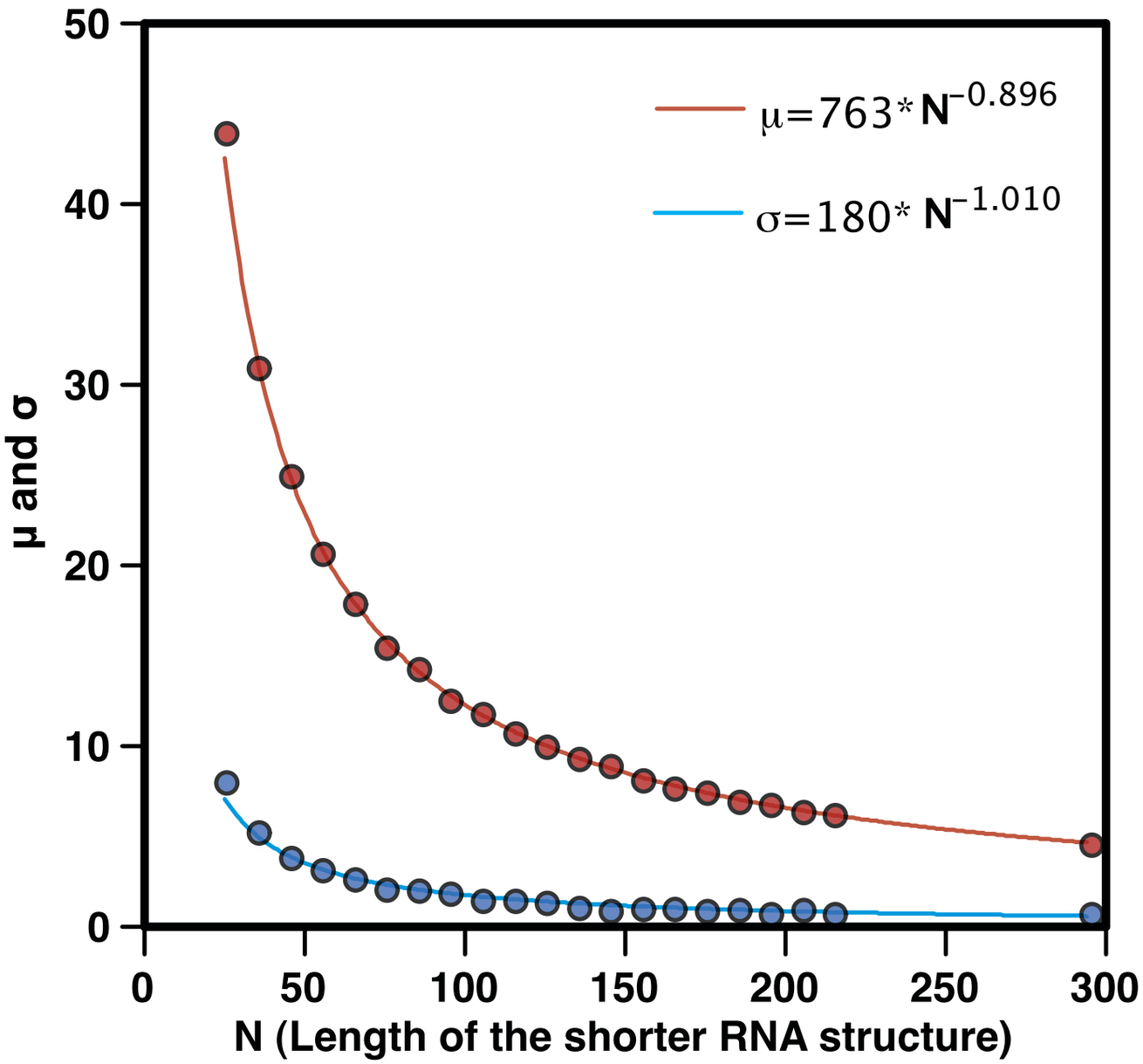
Mean and sigma

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

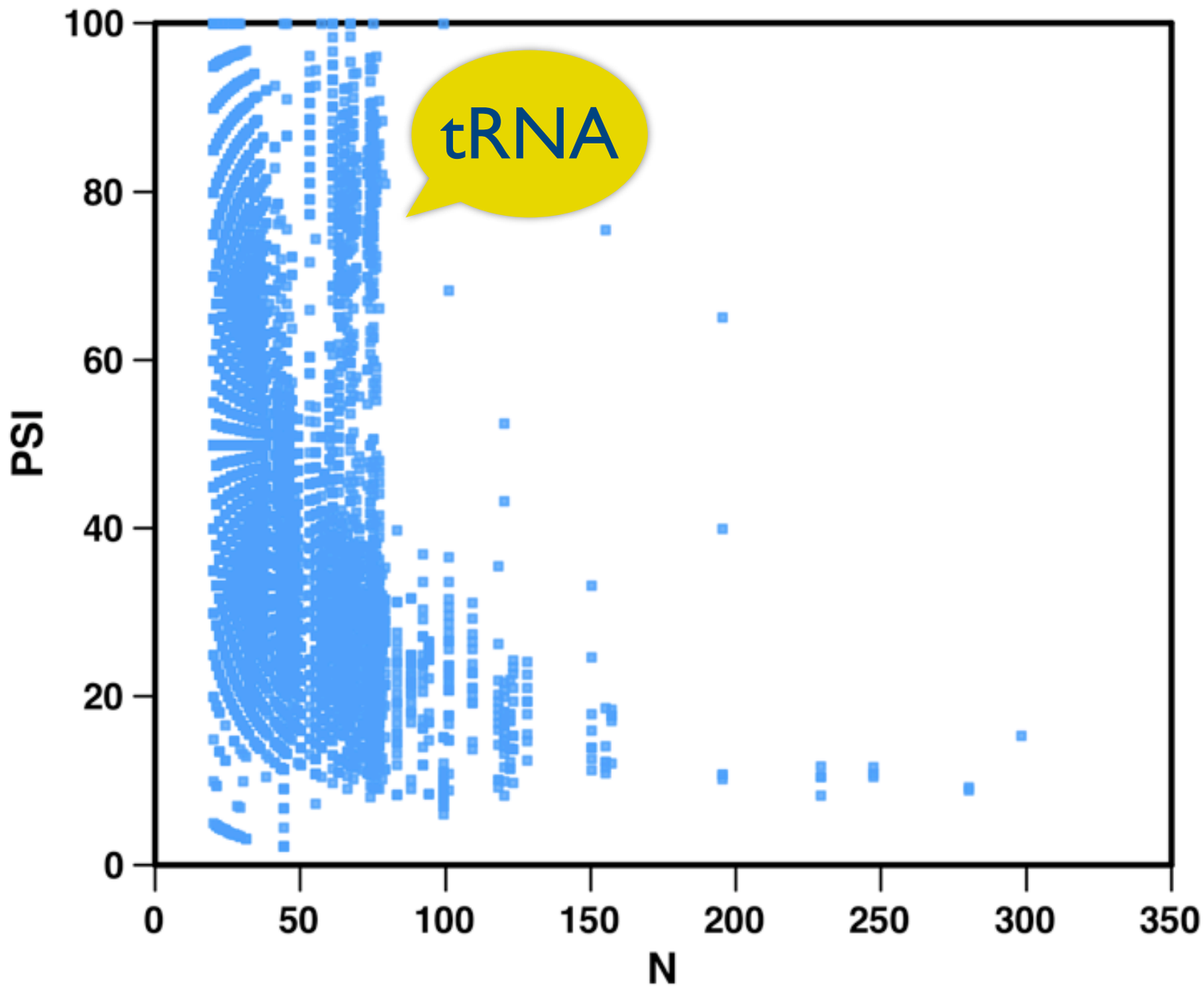
The accuracy of SARA method depends of a large number of parameters.

- 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 -9 to 0 and gap extension for -0.8 to 0
- Secondary structure information

	Gap opening	Gap extension	*k*
Secondary structure	-7.0	-0.6	3
No secondary structure	-8.0	-0.2	7

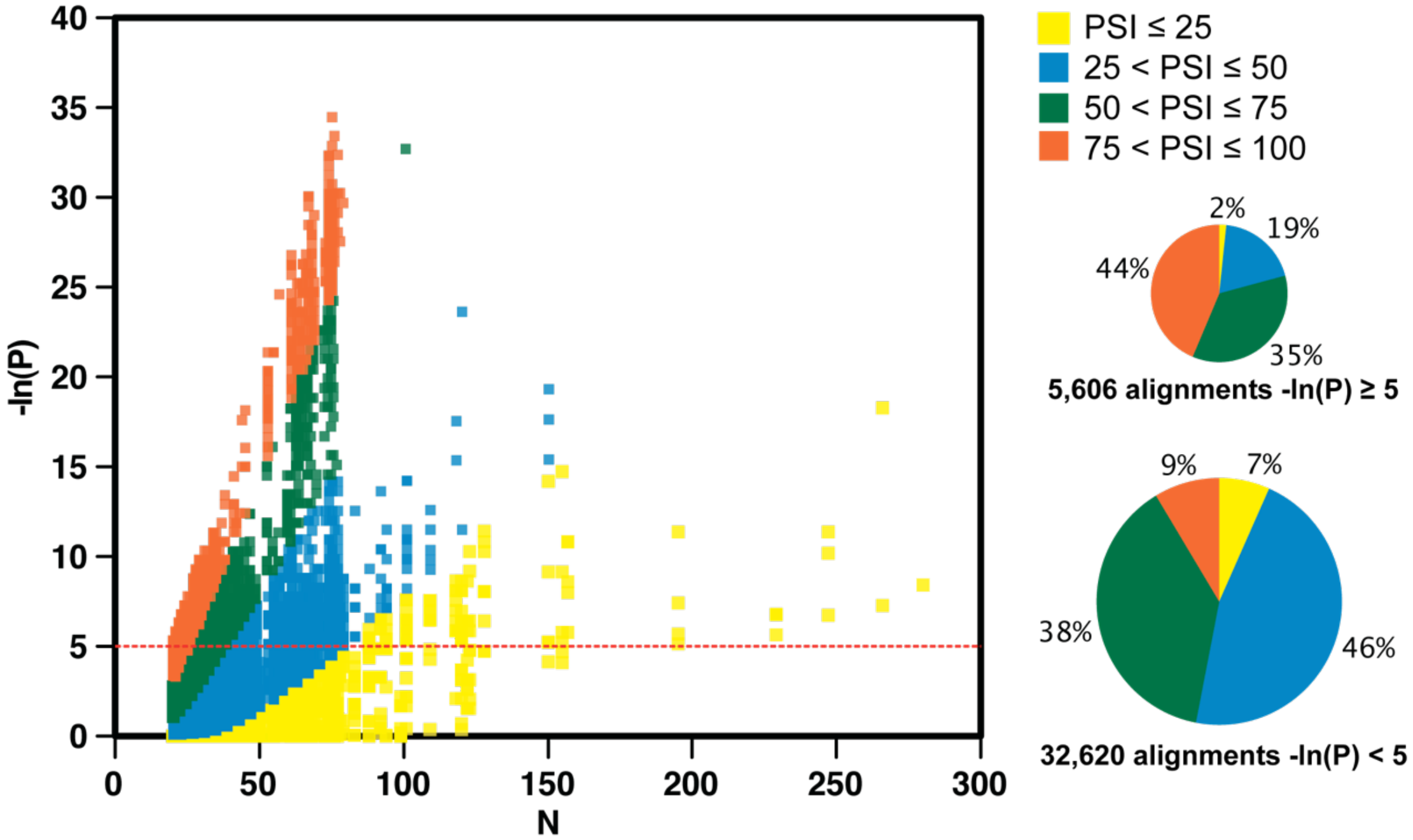
PSI distribution

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



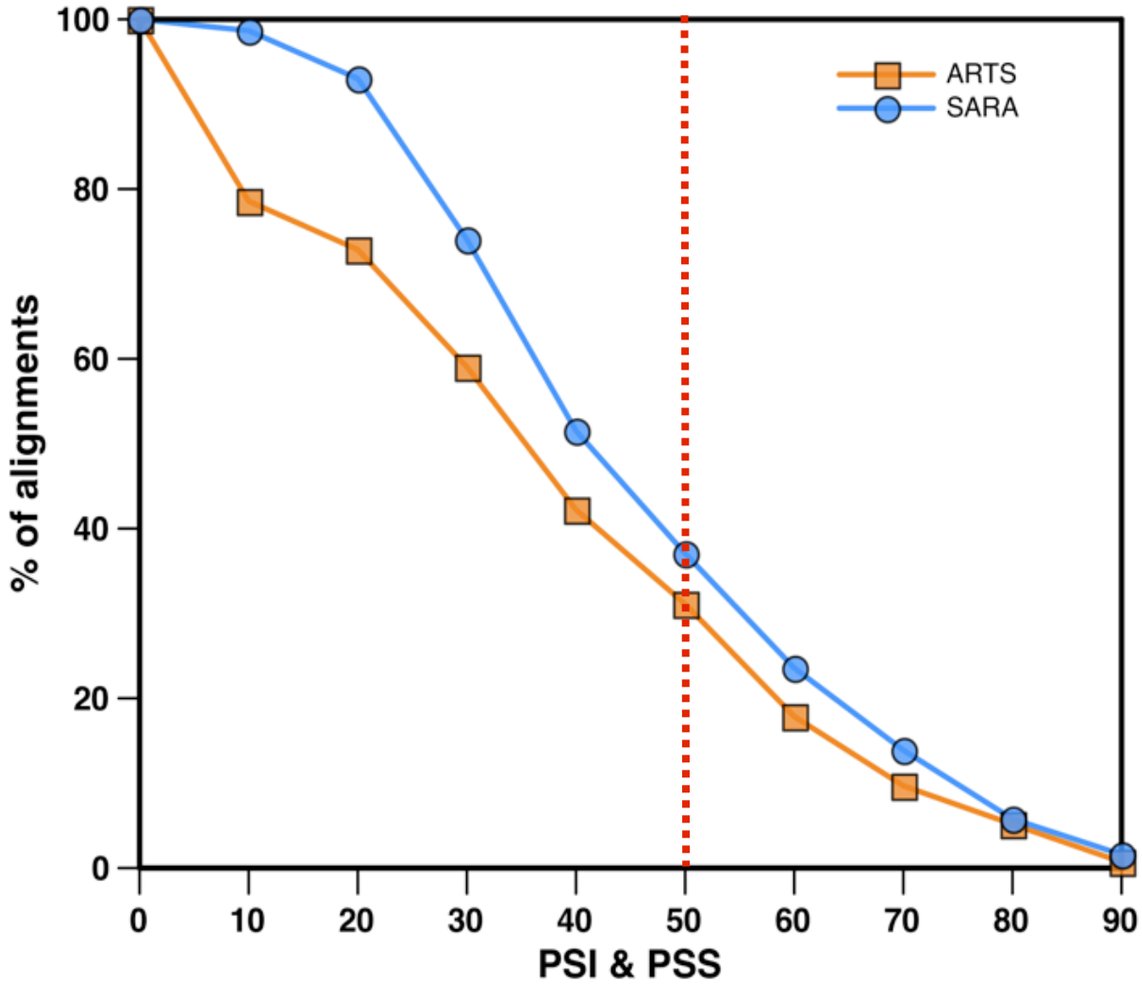
Statistical significance

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



Comparison with ARTS

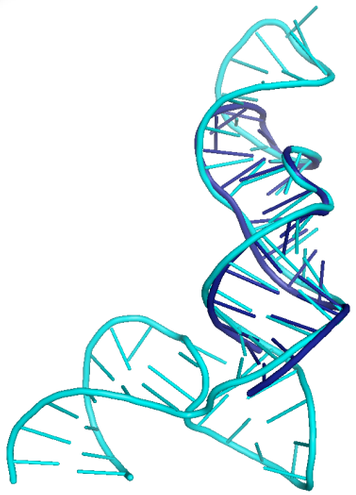
all-against-all comparison of structures in the HR set



PSI: % of structure identity

PSS: % of secondary structure identity

Cut-off distance: 4.0 Å



SARA

Percentage of structure identity (PSI) 92.6%
Percentage of sequence identity 48.0%
Percentage of SSE identity 100.0%
RMSD 1.78 Å

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

ARTS

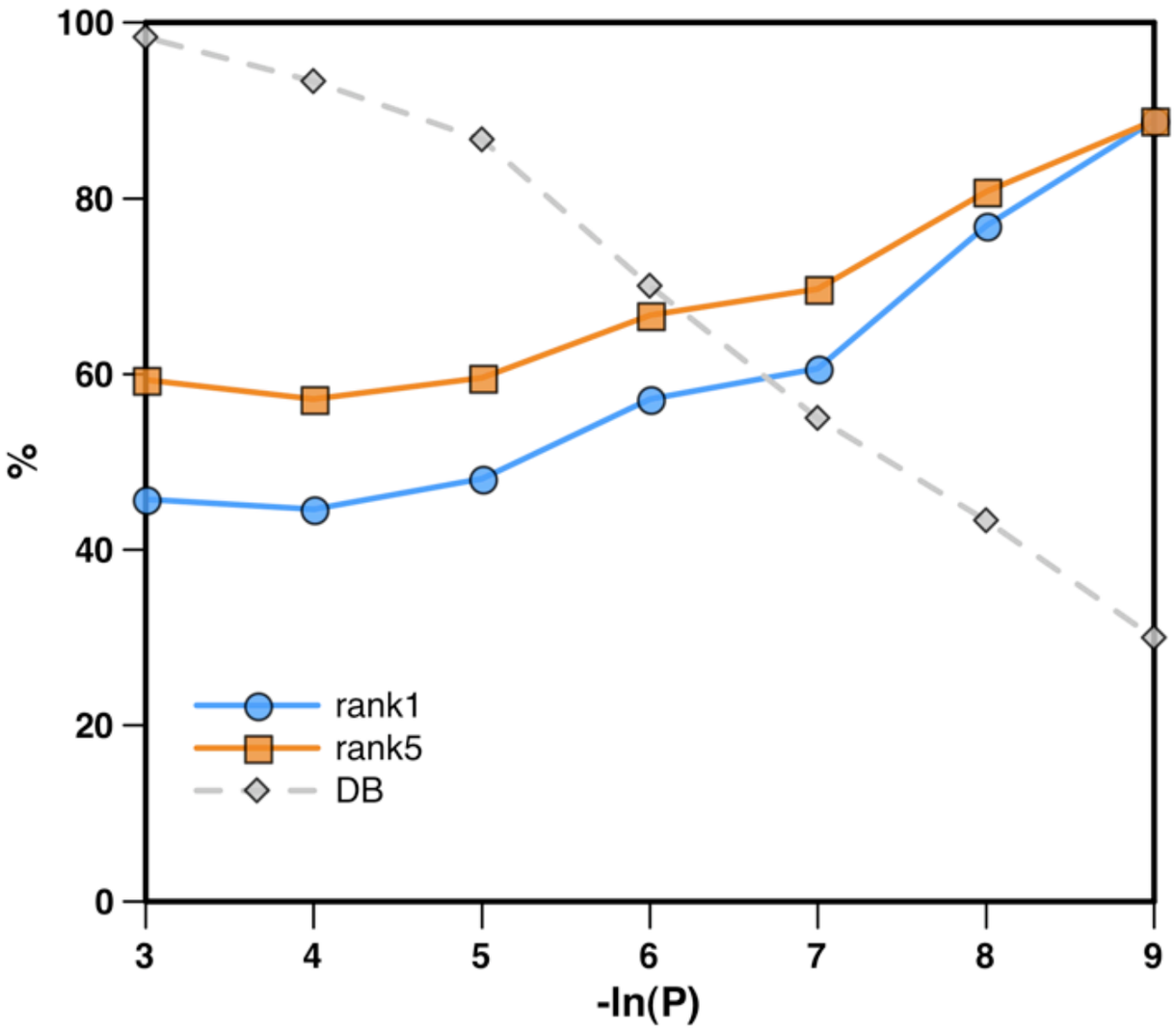
Percentage of structure identity (PSI) 76.9%
Percentage of sequence identity 20.0%
Percentage of SSE identity 79.2%
RMSD 1.66Å

```
>1q96 Chain:A
--------------------gugcucaguaugaga-----aga-accgcacc--------
>1un6 Chain:E
ccggccacaccuacggggccuggguuaguaccugggaaaccugggaauaccaggugccggc
```

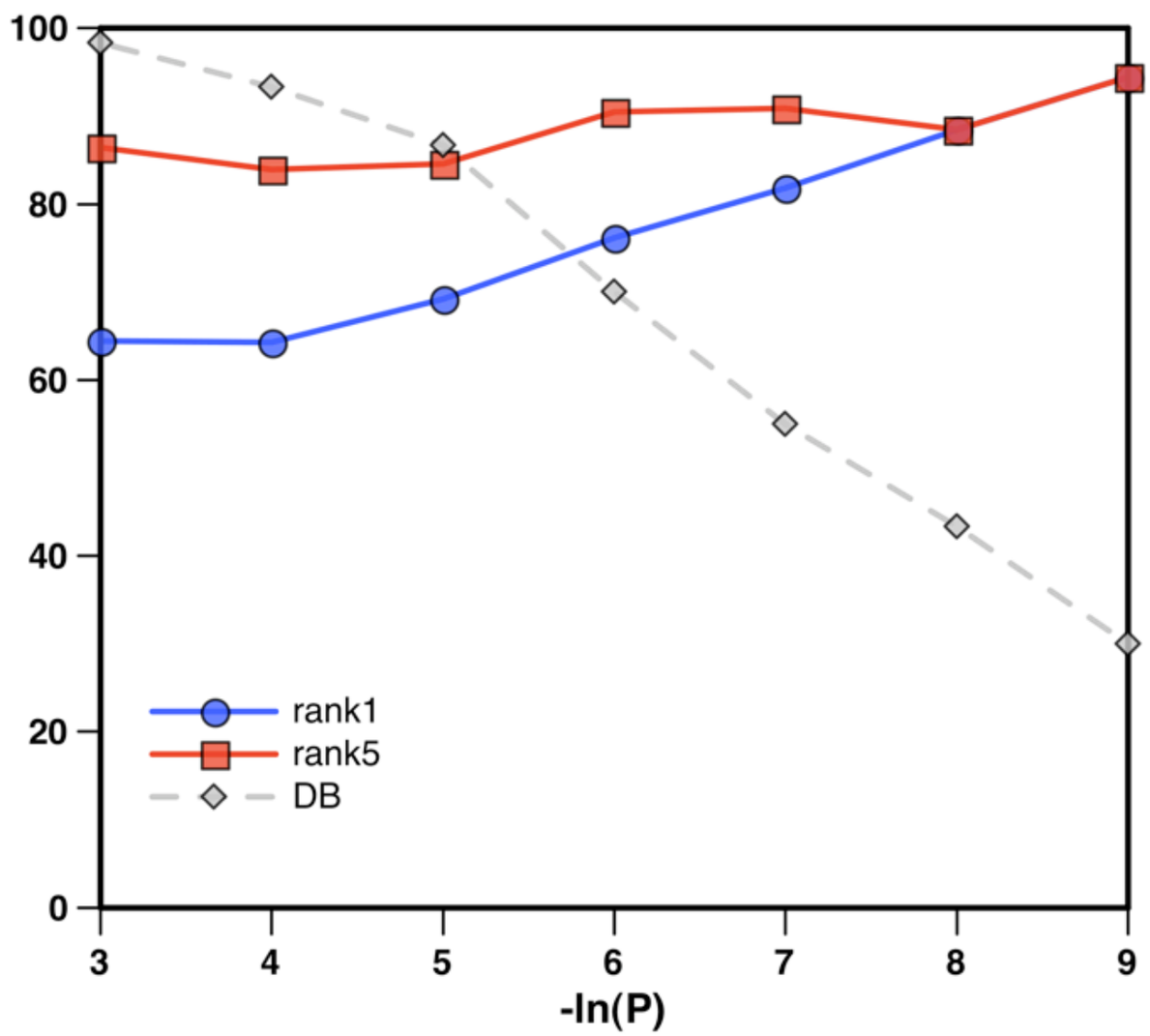
Function assignment

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

Rank of **deepest SCOR function**

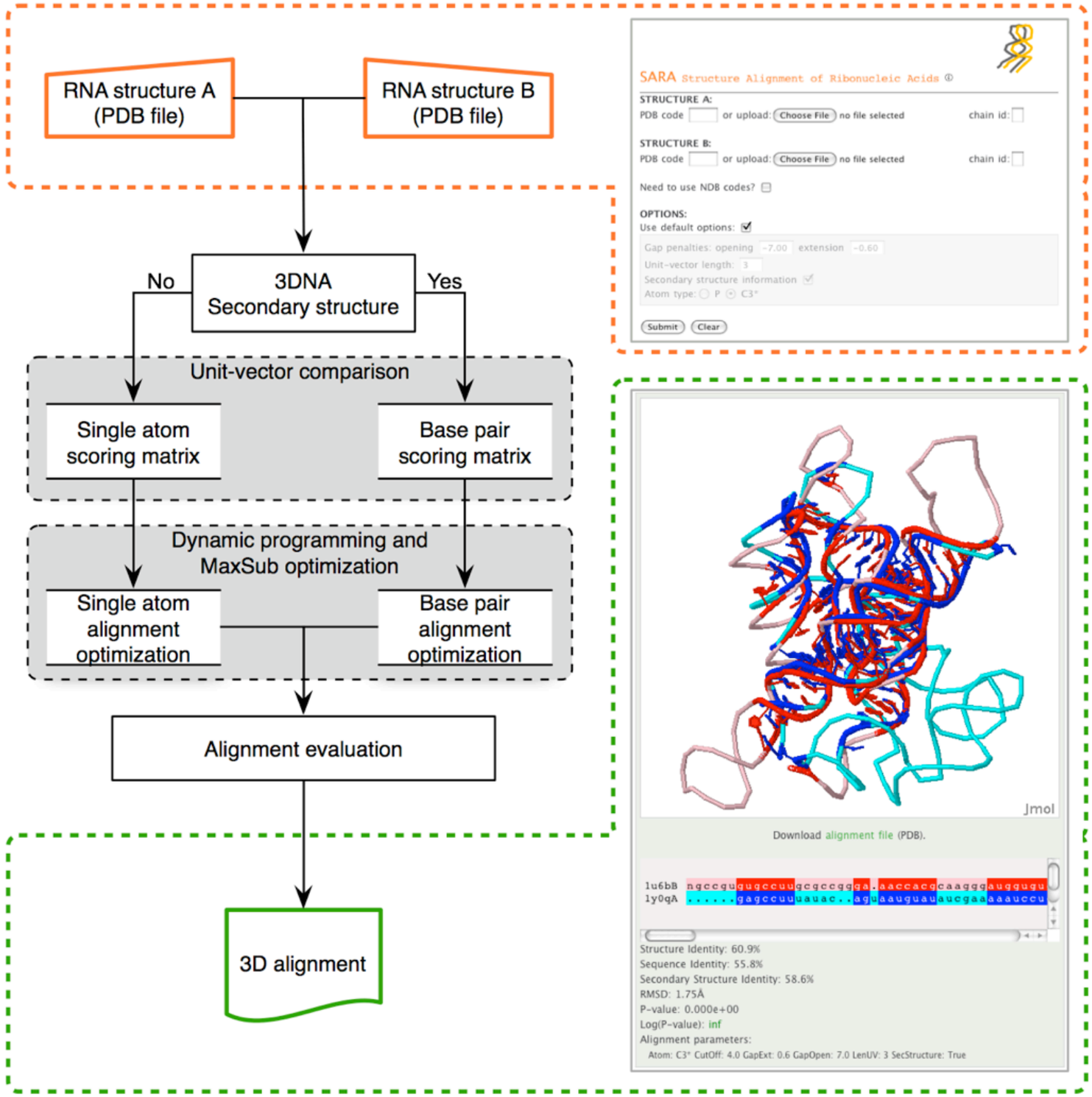


Rank of **related SCOR function**



SARA server

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



Conclusions

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

- An 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 and biologically relevant alignments.

- SARA results in higher accuracy alignments than those produced by ARTS, returning about 6% more alignment with PSI and PSS larger than 50% than ARTS.

- SARA algorithm can be used to automatic function assignment. When results with a $-ln(P)>5$ are selected, SARA correctly ranks, in the first position, 48% of RNA pairs with same deepest SCOR function (60% rank5) and 69% of RNA pairs with related SCOR function (85% rank5).

Acknowledgments

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

Comparative Genomics Unit (CIPF)
Hernán Dopazo
Leo Arbiza
François Serra

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
Francisco García
Rafael Jiménez
Pablo Escobar

MAMMOTH ALGORITHM
Angel Ortiz
ARTS PROGRAM
Oranit 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