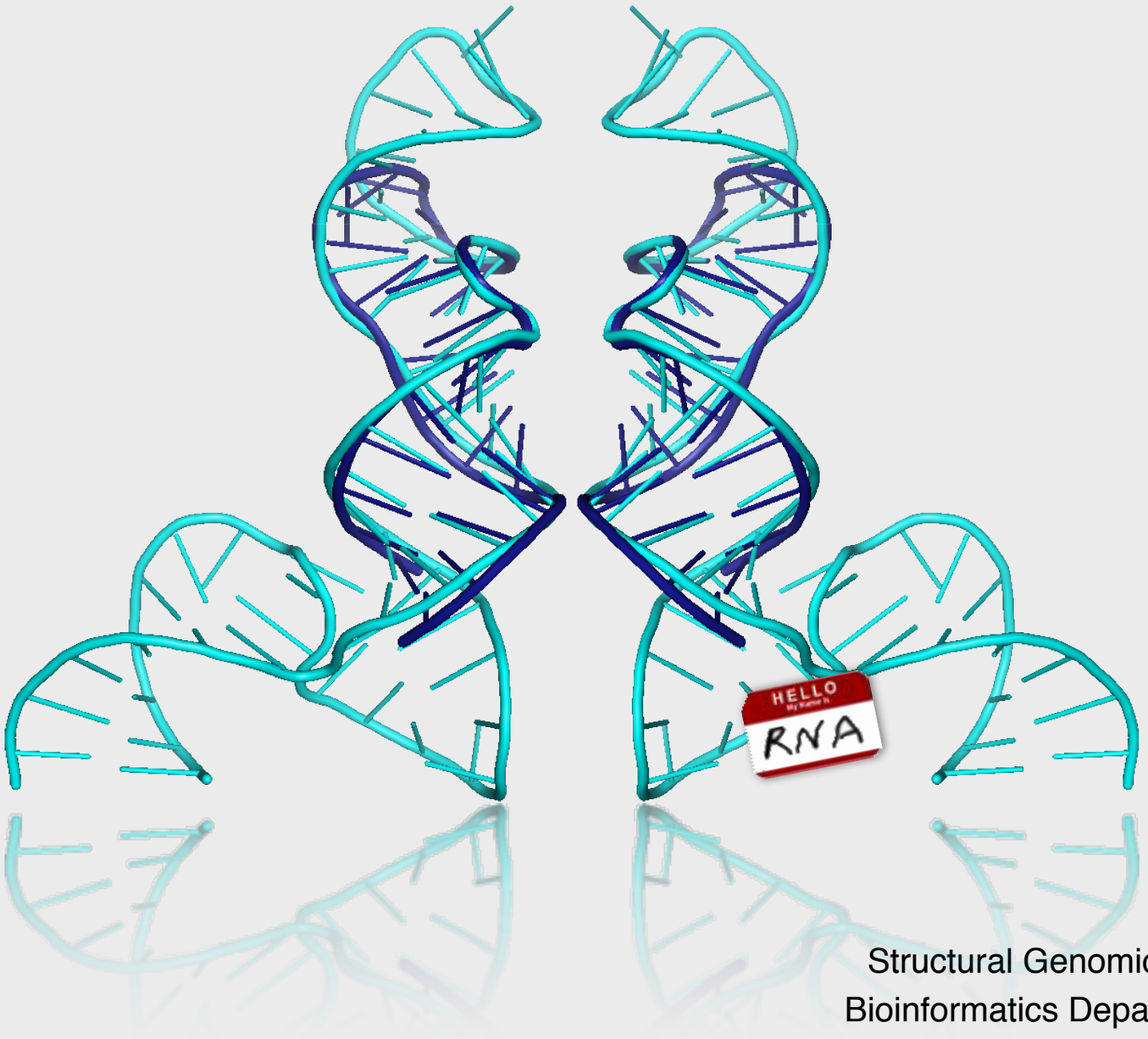
RNA structure alignment by a unit-vector approach



Emidio Capriotti

Marc A. Marti-Renom

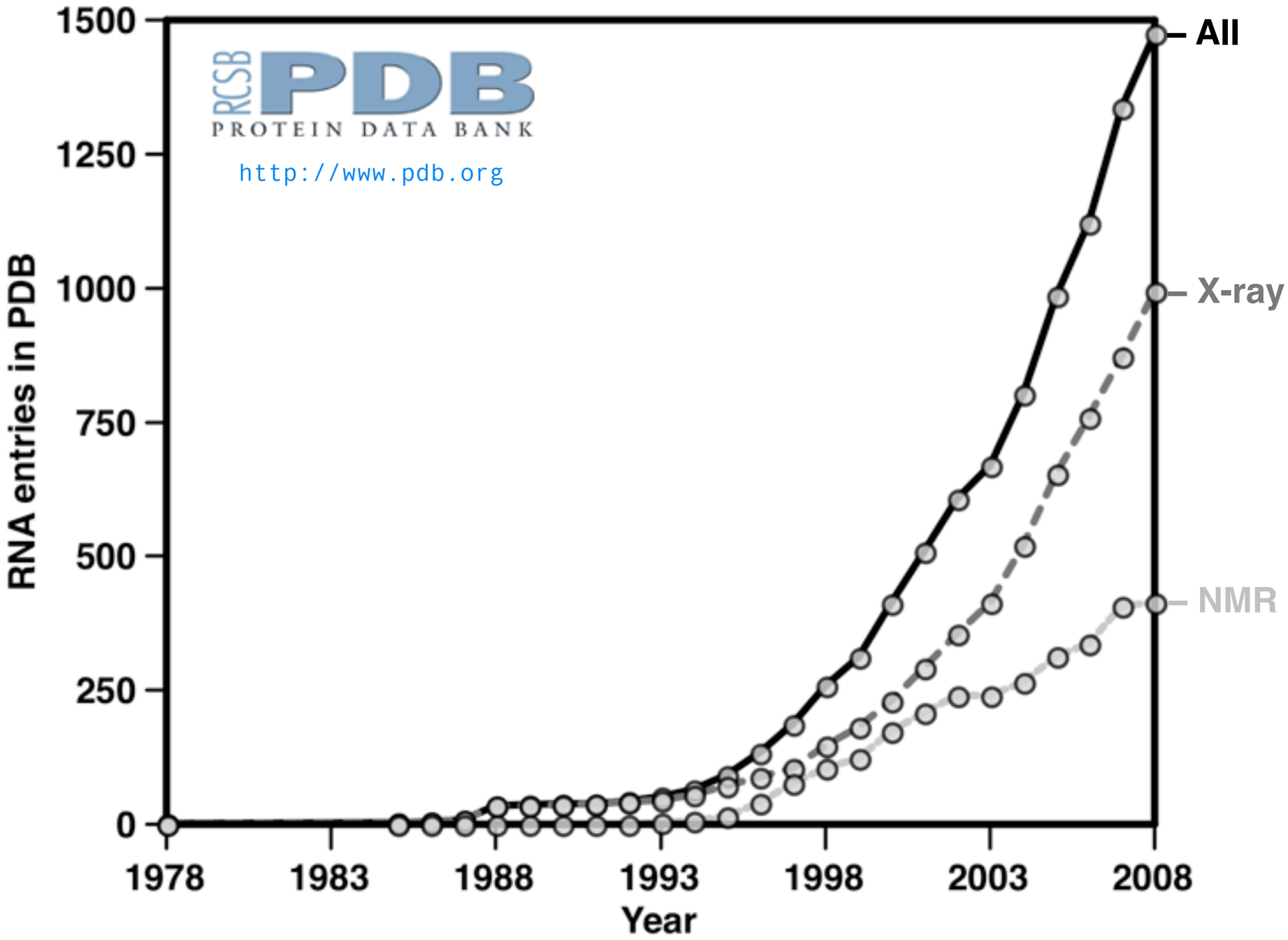
http://sgu.bioinfo.cipf.es

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain

ECCB08
Cagliari (Italy)
22-26 September 2008

RNA structure

The PDB database contains ~1,500 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

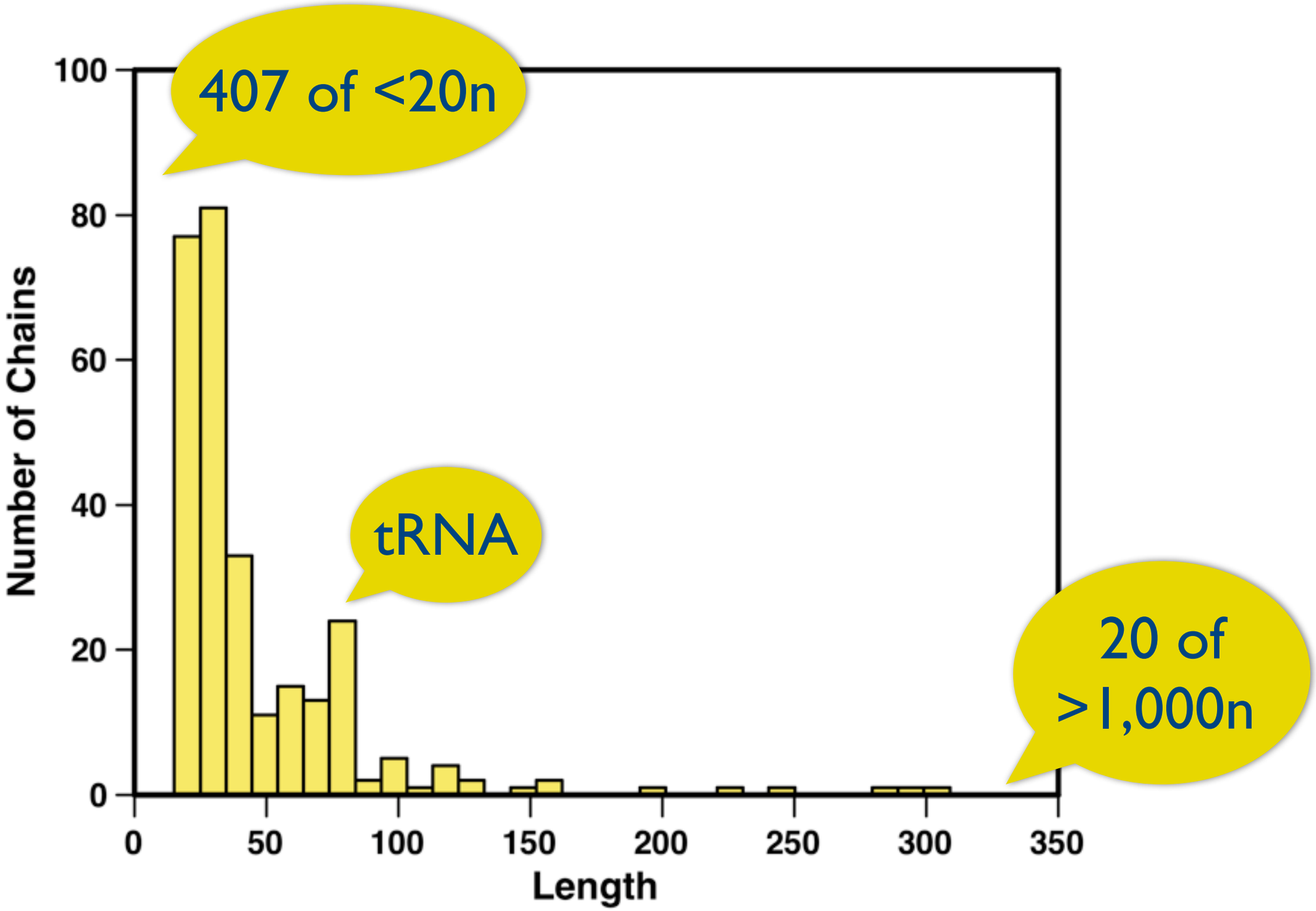
* *from PDB November 06.*

** *non-redundant 95% sequence identity*

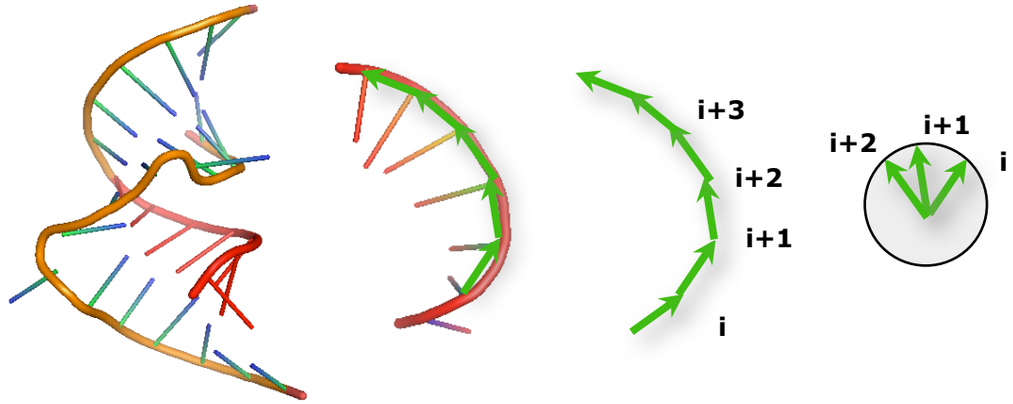
*** *SCOR functions with at least two chains*

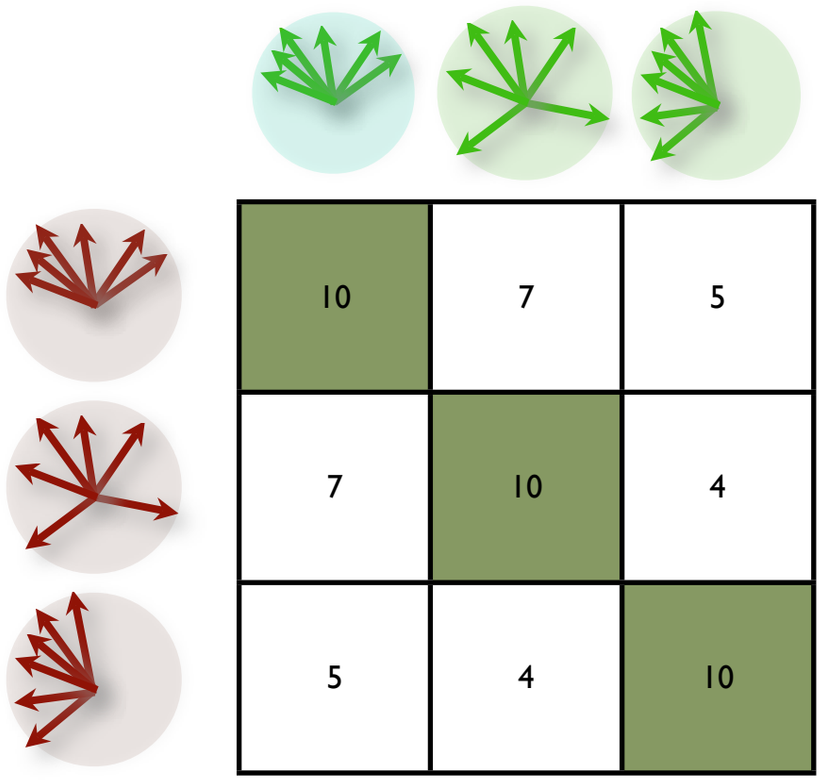
**** *resolution below 4.0 Å and with no missing backbone atoms.*

Dataset distribution



Unit Vector





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

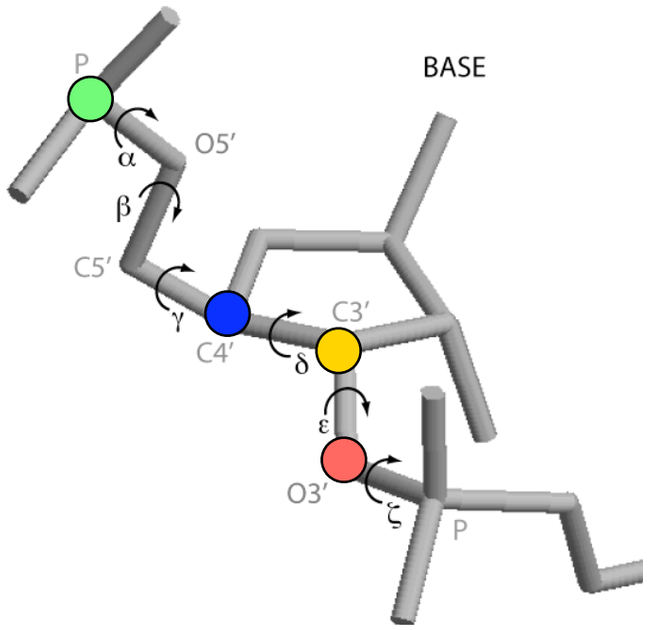
$$S_{ij} = \frac{(URMS^R - URMS^{ij})}{URMS^R} \Delta(URMS^R, URMS^{ij})$$

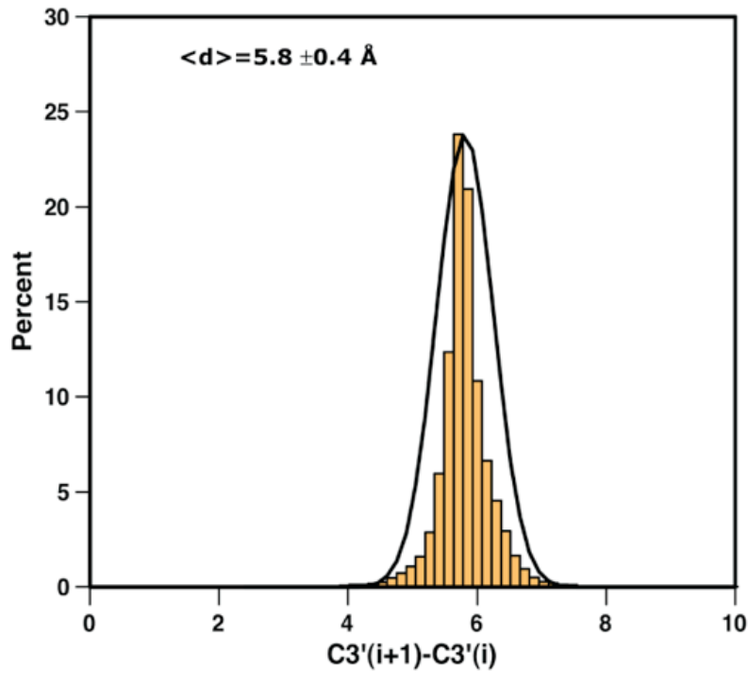
$$\Delta(URMS^R, URMS^{ij}) = 10 \Rightarrow URMS^R > URMS^{ij}$$
$$\Delta(URMS^R, URMS^{ij}) = 0 \Rightarrow URMS^R \leq URMS^{ij}$$

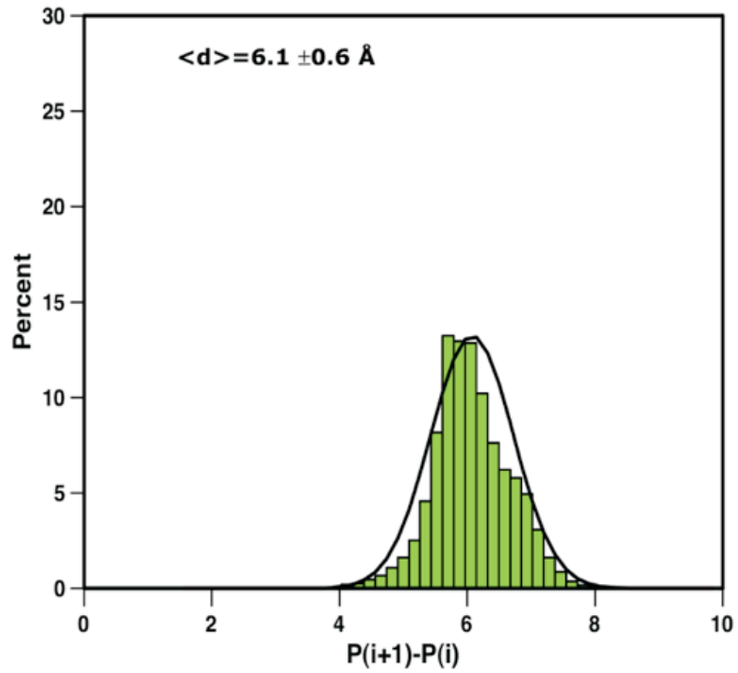
Ortiz et al. *Proteins* 2002

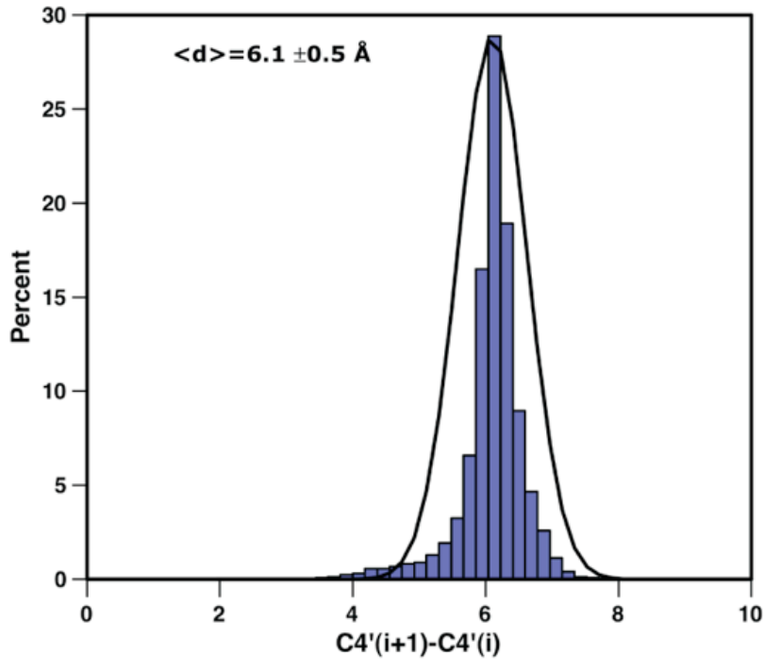
Atom selection

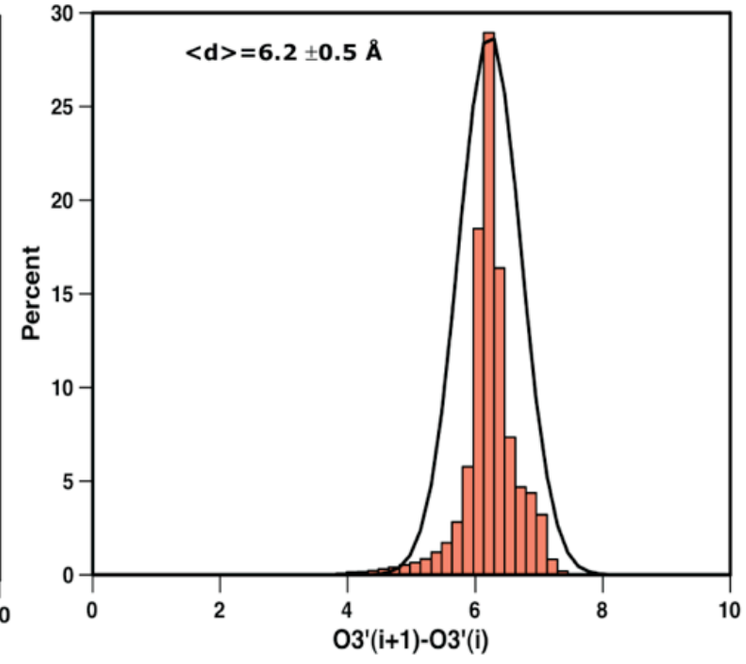
The best backbone atom that represents the RNA structure has been selected by evaluating the distribution of the distances between consecutive atoms in structures from the NR95 set.







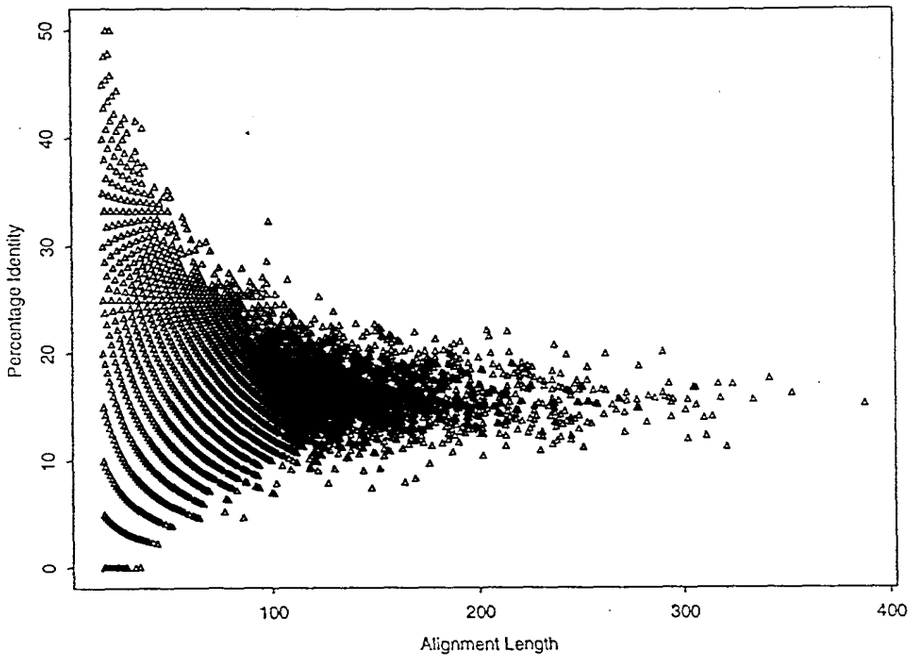




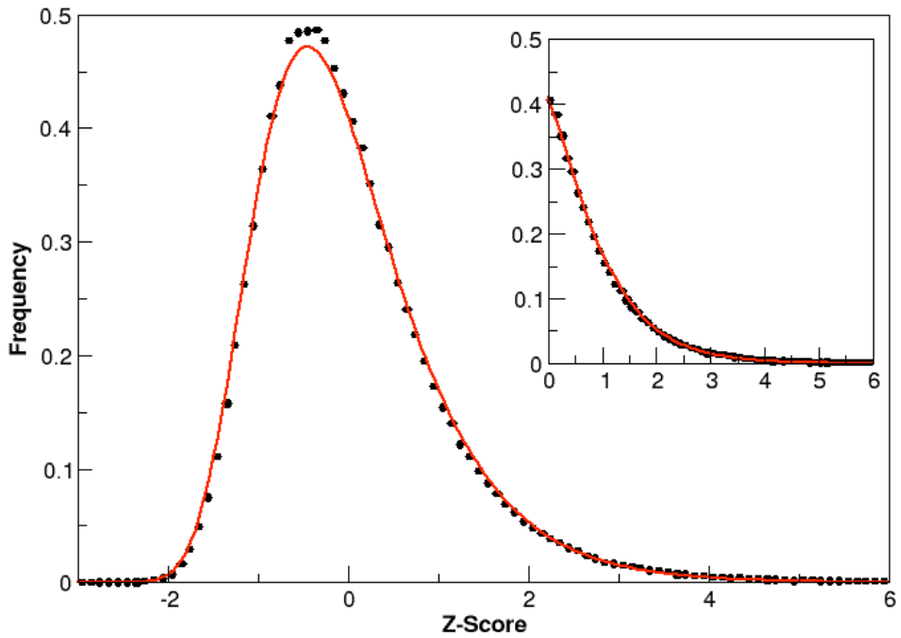
Background distribution

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



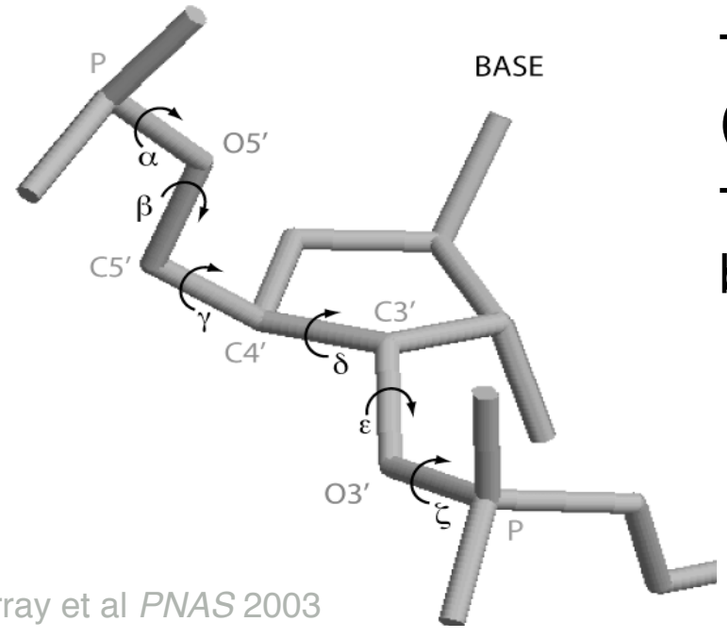
Analytic



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

Karlin and Altschul *PNAS* 1990

Random RNA



Murray et al *PNAS* 2003

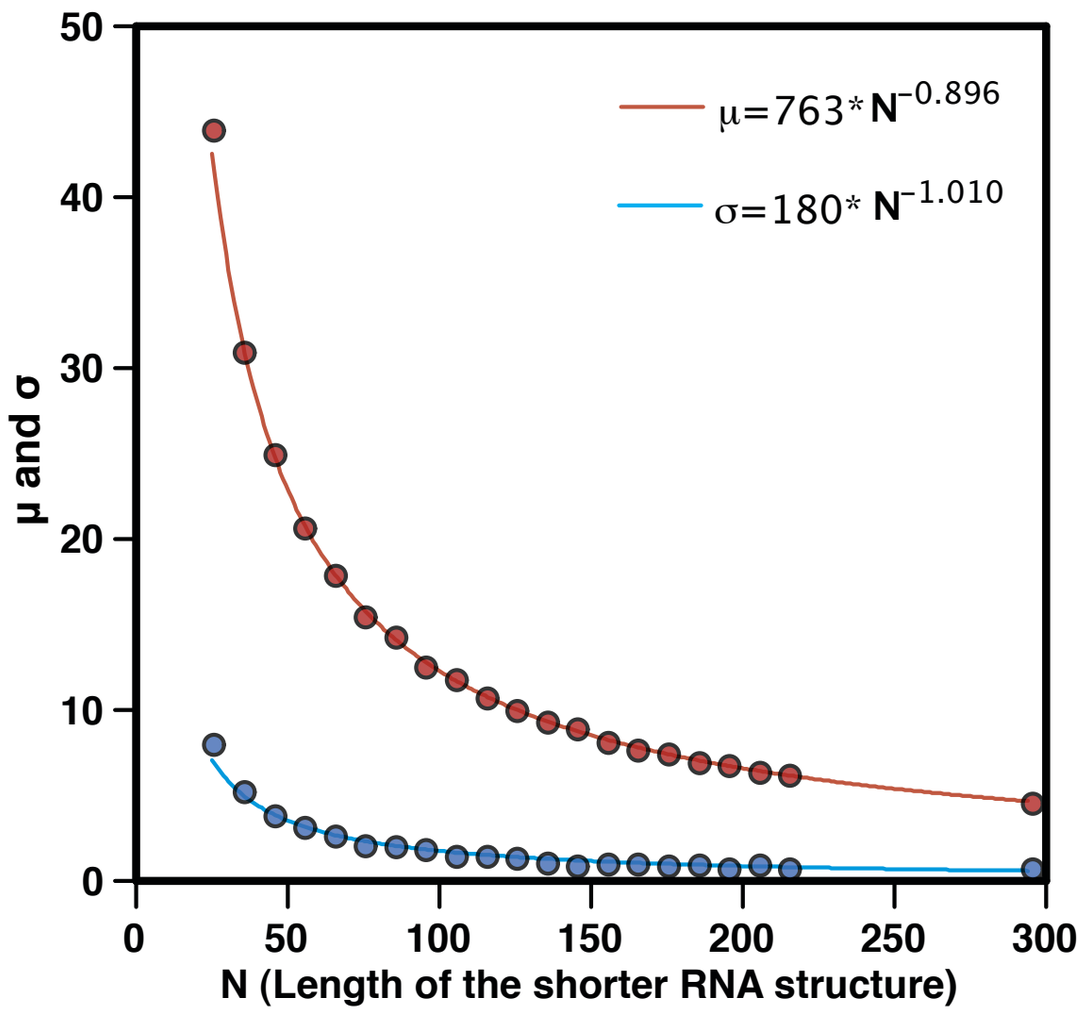
The RNA backbone can be described given the 6 torsion angle (α,β,γ,δ,ε,ζ) for each nucleotide.

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

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

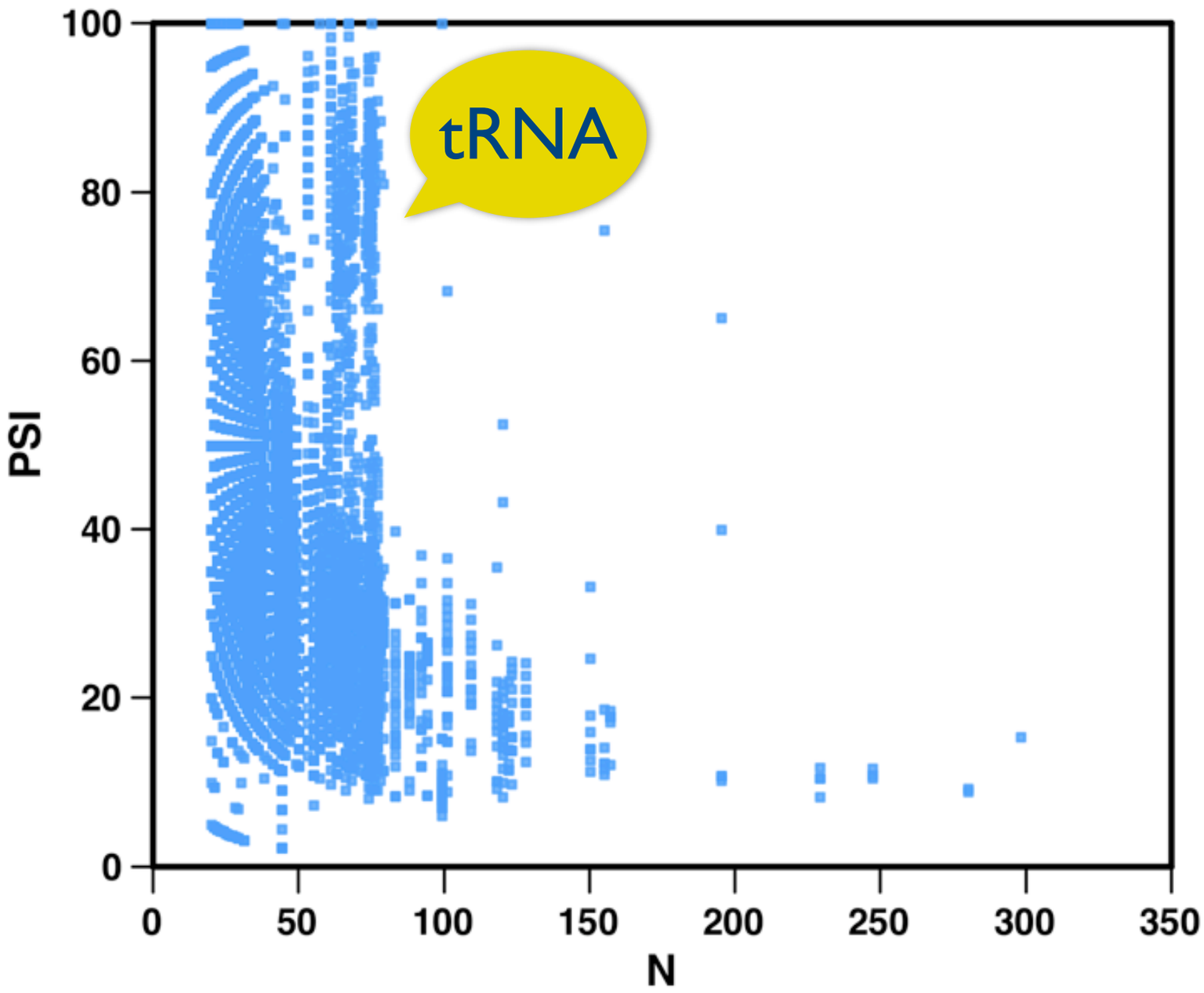
The accuracy of SARA method depends of a large number of parameters.

- 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

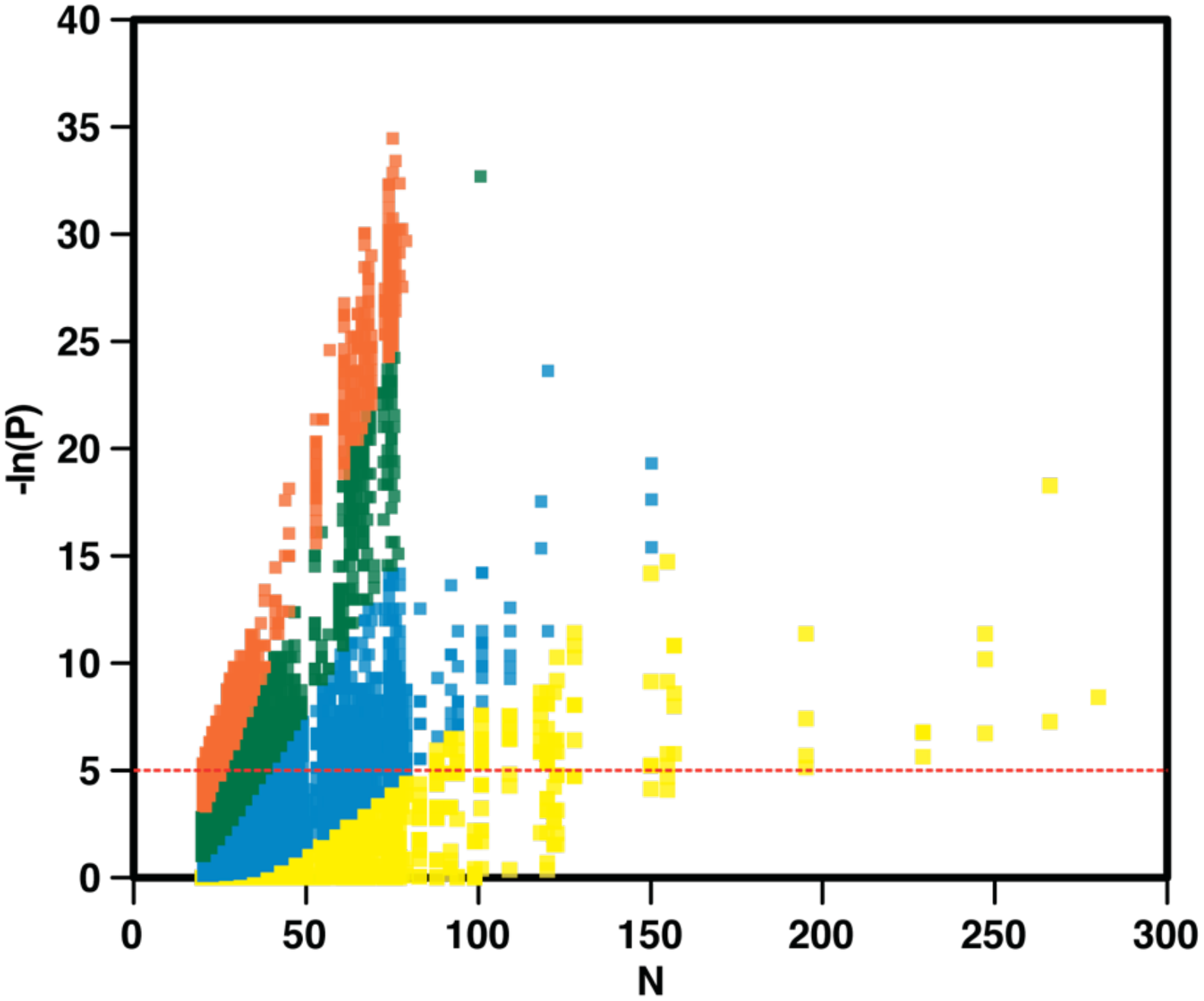
PSI distribution

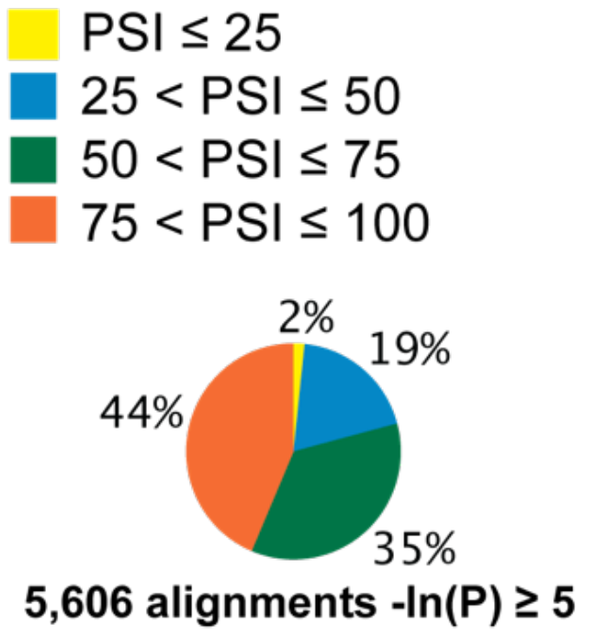
all-against-all comparison of structures in the NR95 set



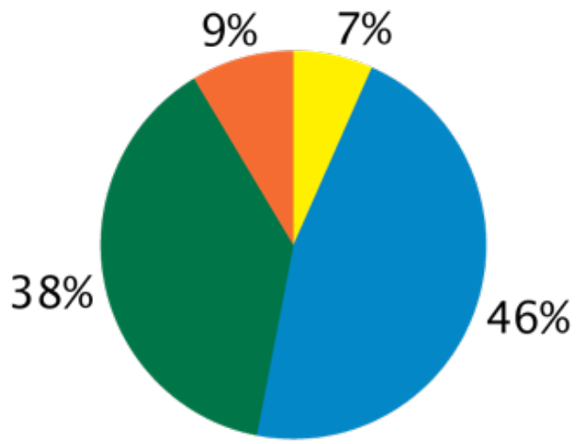
Statistical significance

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





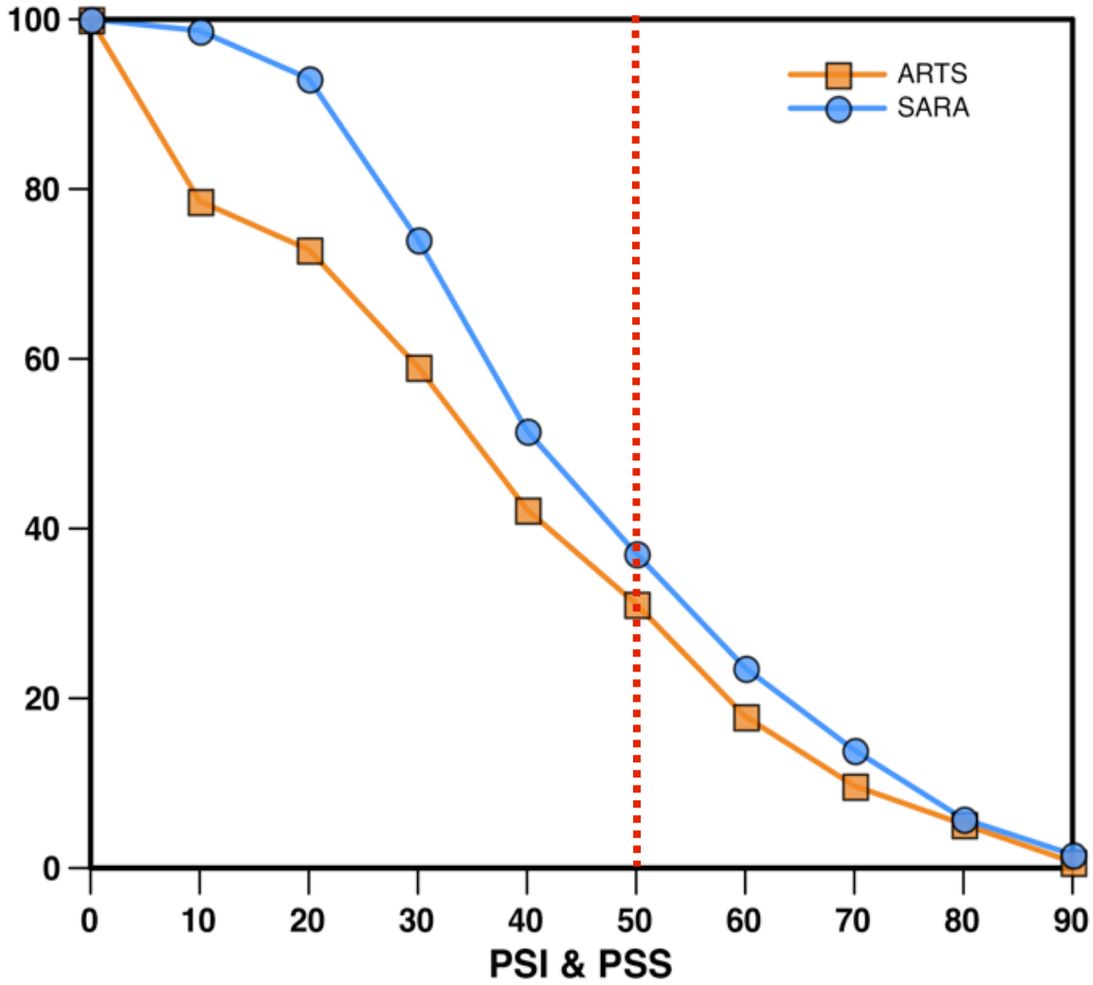
5,606 alignments -ln(P) ≥ 5



32,620 alignments -ln(P) < 5

Comparison with ARTS

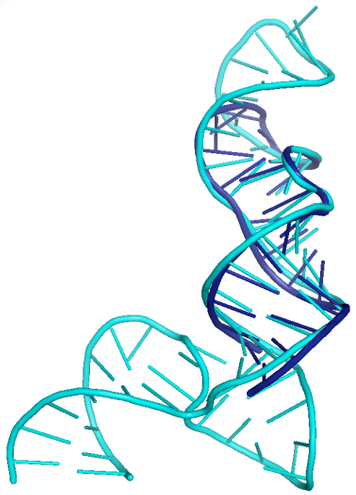
all-against-all comparison of structures in the HR set



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

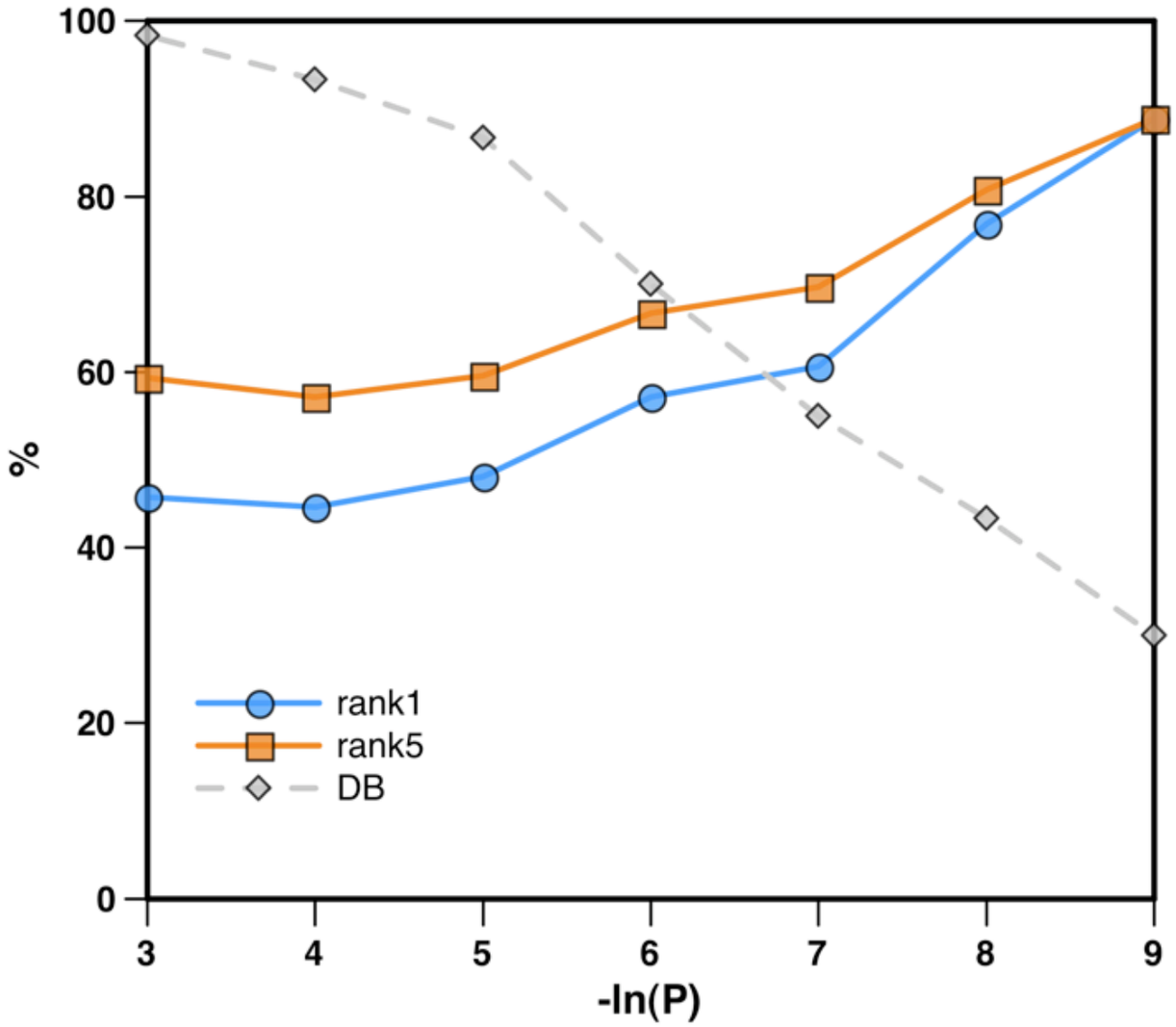
Percentage of structure identity (PSI) 76.9%
Percentage of sequence identity 20.0%
Percentage of SSE identity 79.2%
RMSD 1.66Å

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

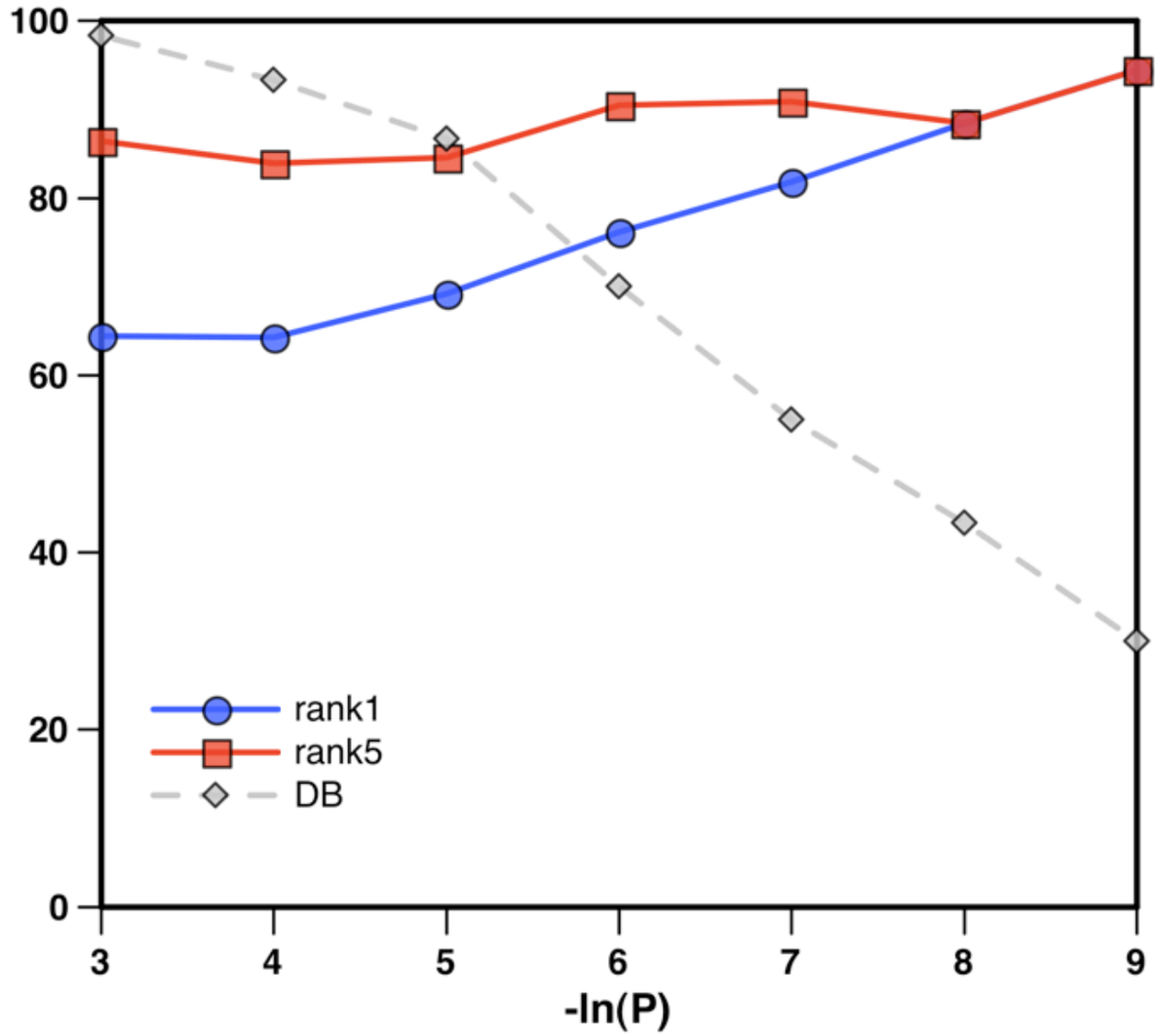
Function assignment

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

Rank of **deepest SCOR function**

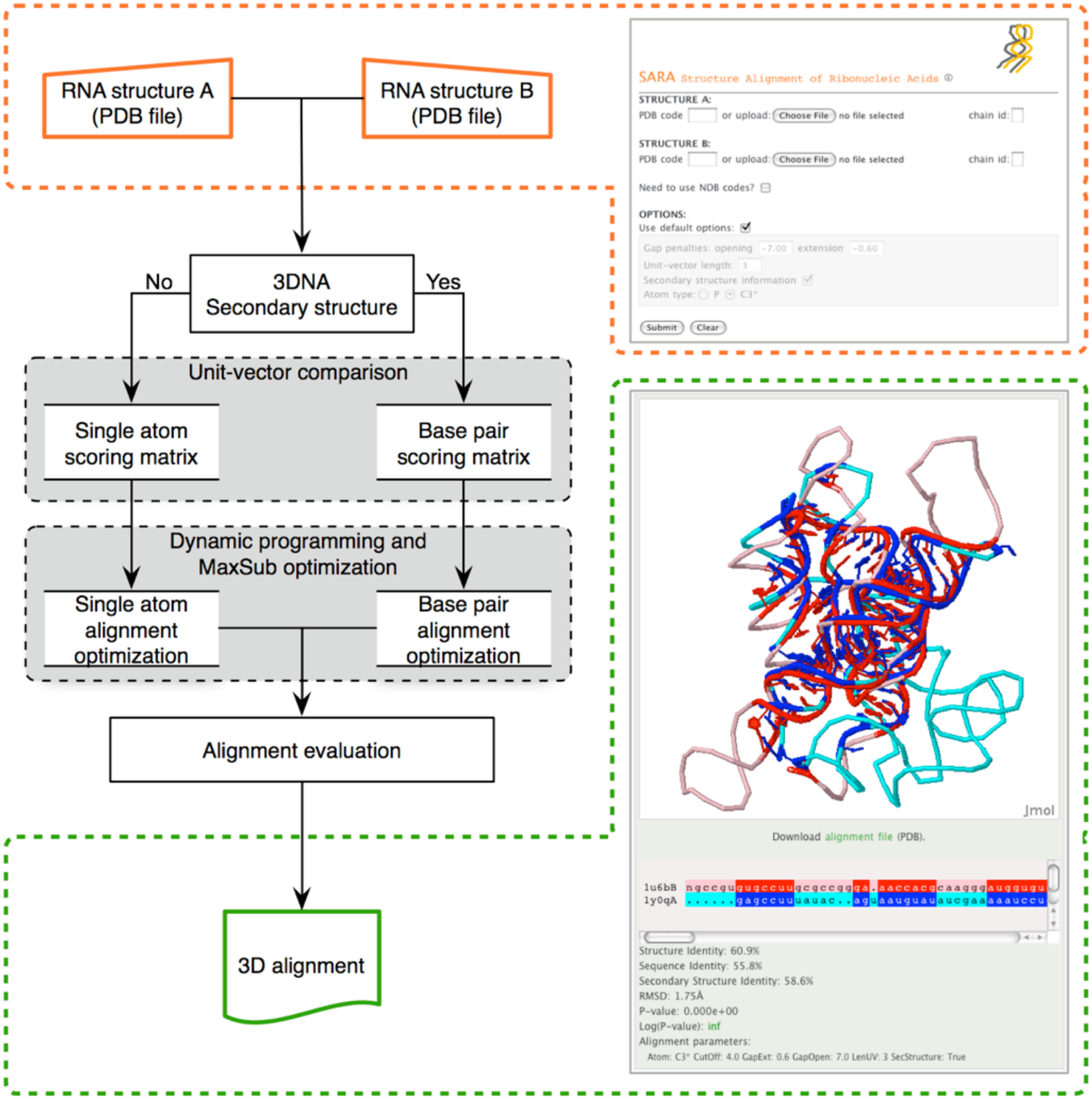


Rank of **related SCOR function**



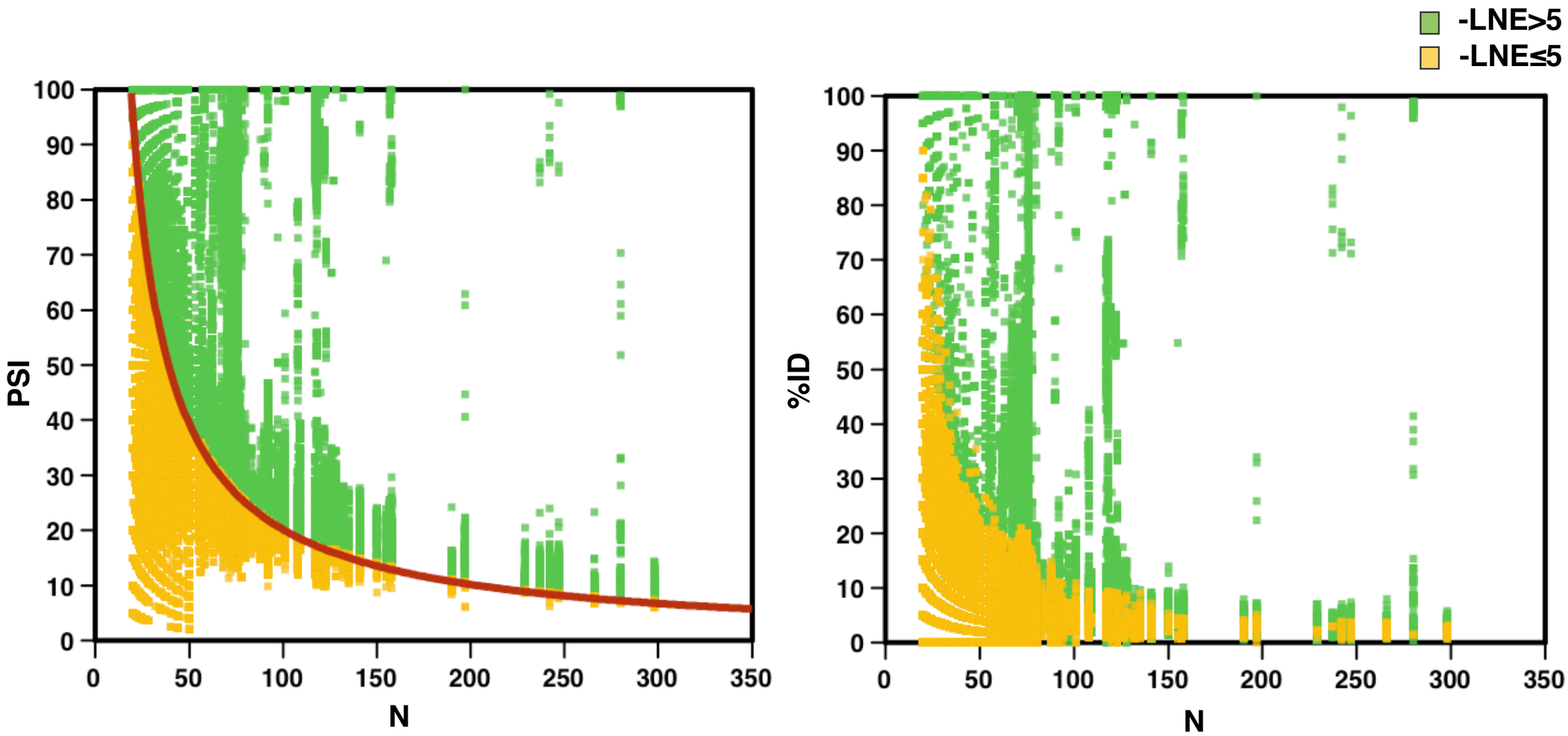
SARA server

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



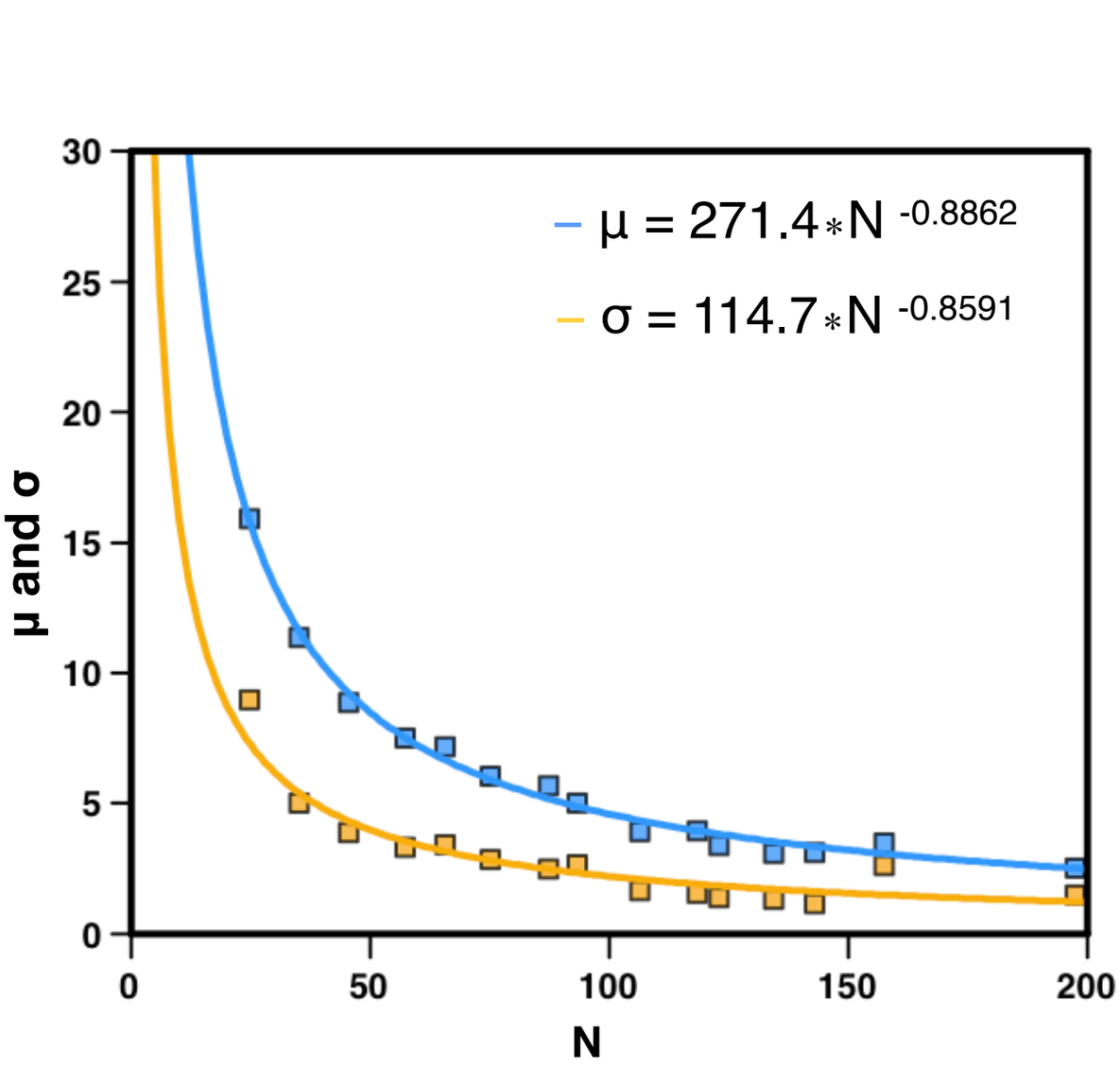
All against all alignments

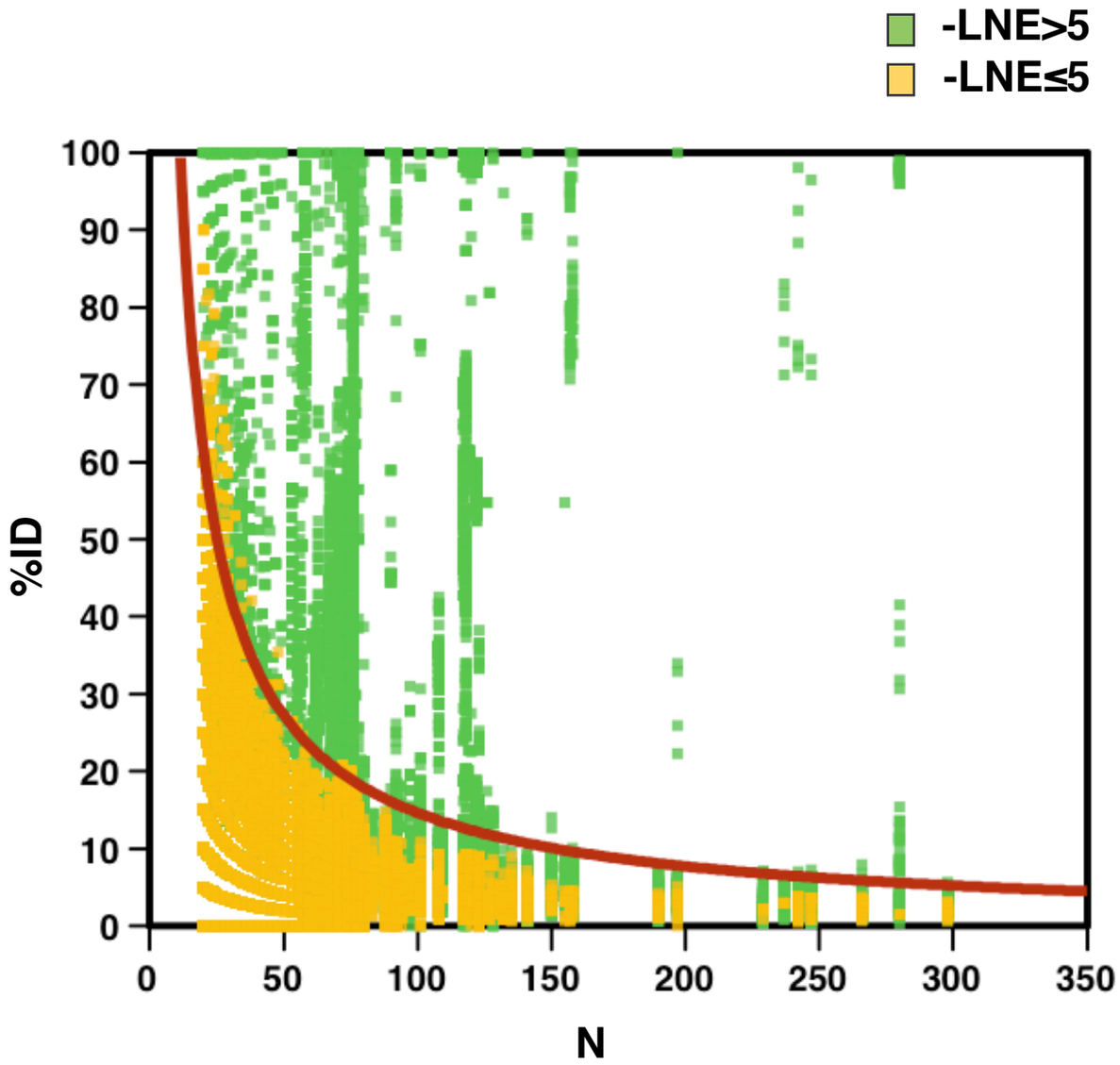
A set of 829 RNA chain structures from PDB (Jan 08) has been selected to study the relationship between sequence and structure similarity.



Sequence similarity distribution

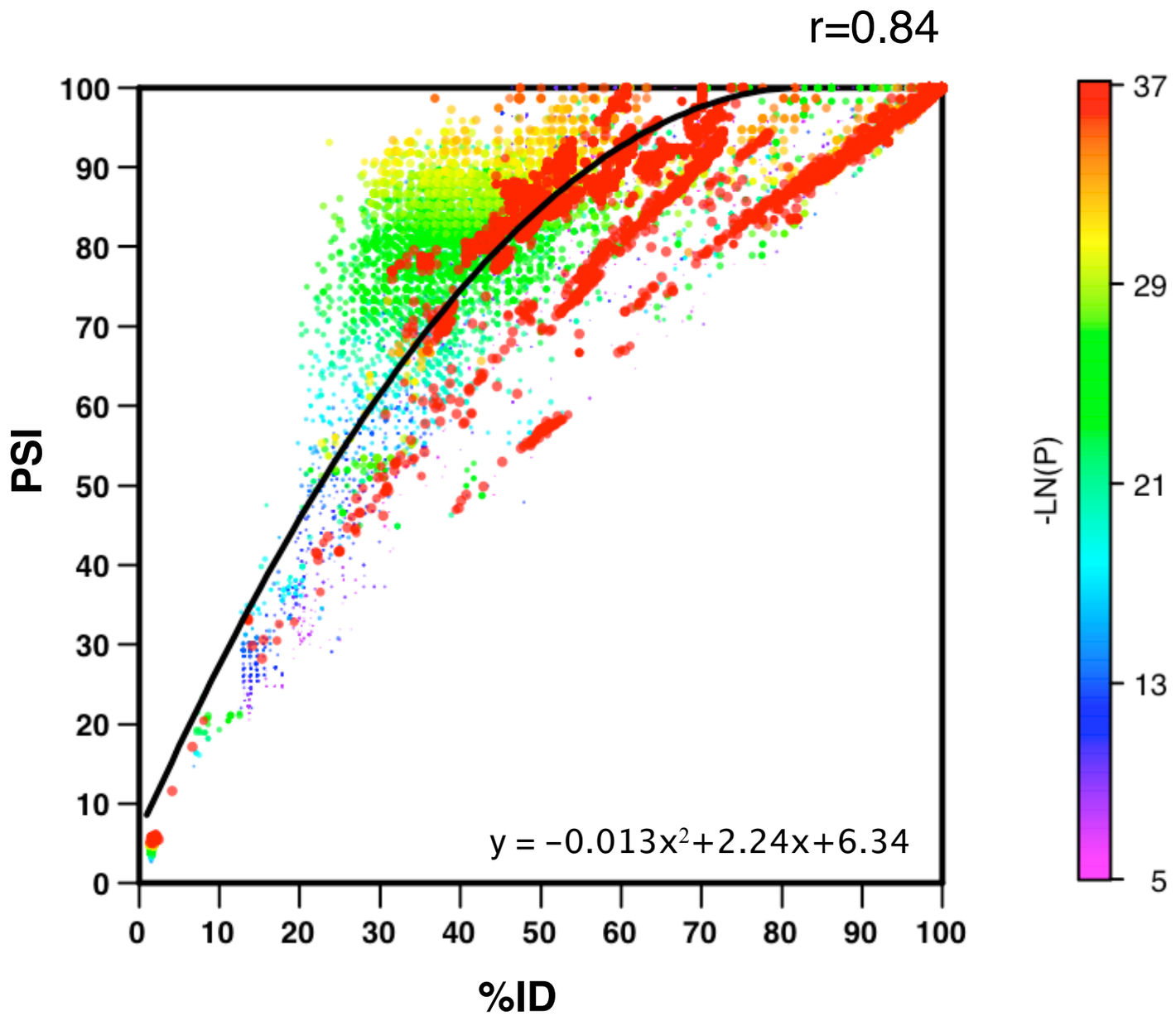
Using the subset of alignments with -LNE≤5 we evaluate the background distribution for the percentage of sequence identity (%ID)





RNA sequence and structure

The plot shows that tertiary structure is more conserved than sequence.



Conclusions and future directions

- The SARA method is a good alternative to other RNA structure alignment methods.
- The statistics obtained using the alignments between random generated structures have allowed to select high quality alignment.
- The subset of alignments with log(p-value) ≤ 5 has been used to evaluate the minimum level of sequence identity that corresponds to the conservation of the 3D structure.
- The RNA tertiary structure is more conserved than sequence.

- Develop new strategies to represent RNA secondary structure to improve the quality of the alignments
- A set of high quality alignments will be selected to derive the rules for the prediction of new RNA structures relying on sequence-structure alignment information.

Acknowledgments

Structural Genomics Unit (CIPF)
Marc A. Marti-Renom
Davide Bau
Emidio Capriotti

http://sgu.bioinfo.cipf.es

MAMMOTH ALGORITHM
Angel Ortiz
ARTS PROGRAM
Oranit Dror
Ruth Nussinov
Haim J. Wolfson

ECCB08 Travel Fellowship
granted by
BIOSAPIENS
Network of Excellence

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

Ángel Ramirez Ortiz,
June 30th 1966 - May 5th 2008