

# RNA Comparative Structure Modeling...

## three steps ahead



**Marc A. Marti-Renom**

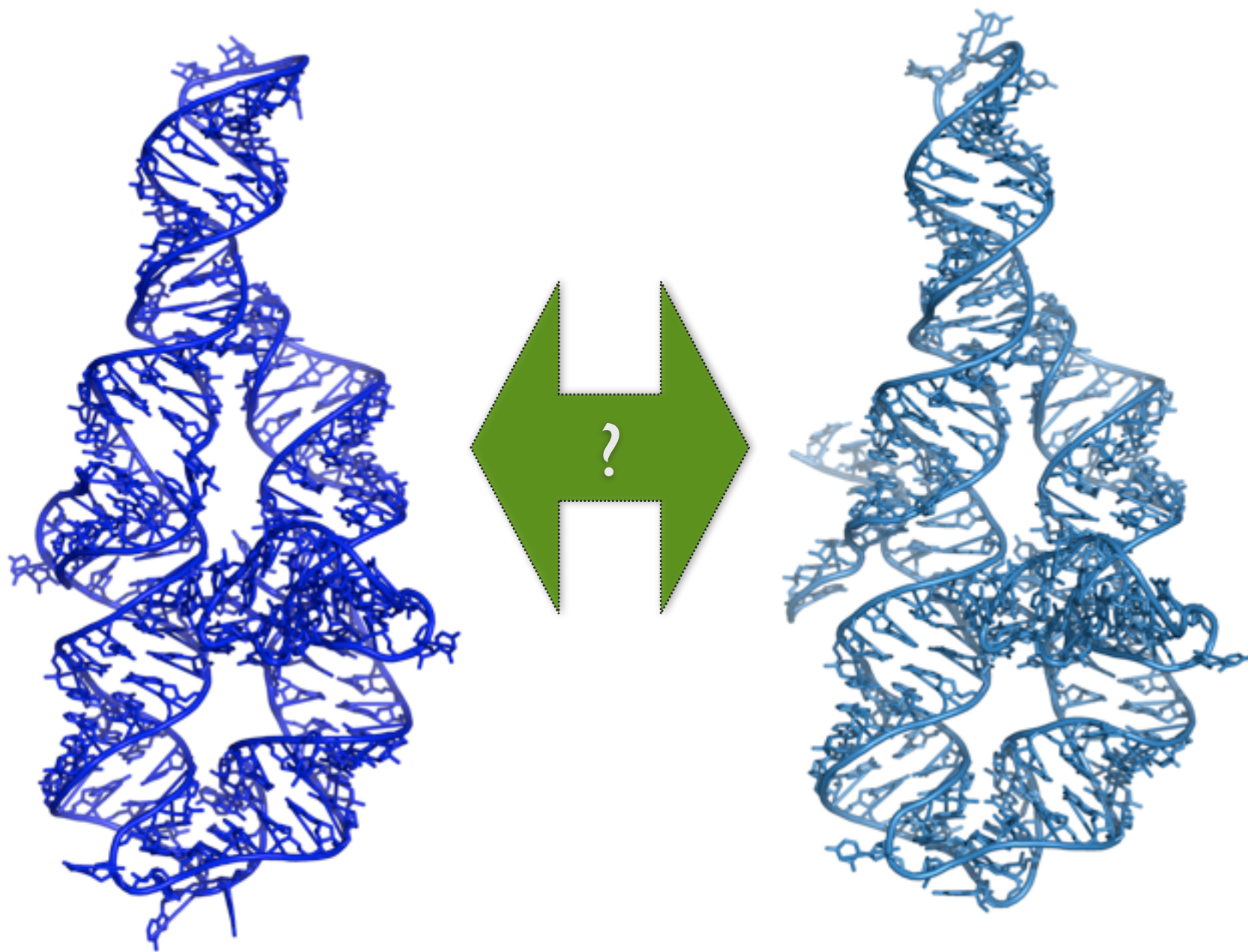
<http://sgu.bioinfo.cipf.es>

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



# First step

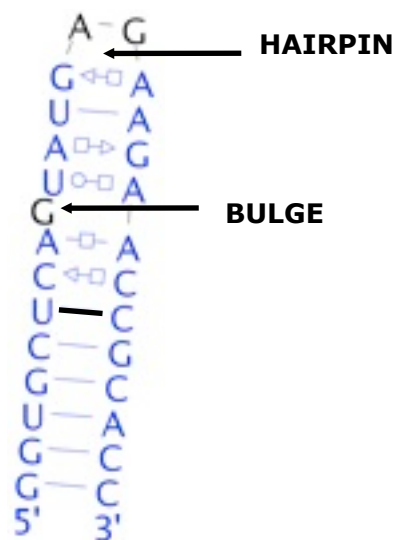
Can we reliably compare RNA structures?



# RNA structure

## Primary Structure

>Mutant Rat 28S rRNA sarcin/ricin domain  
GGUGCUCAGUAUGAGAAGAACCGCACC



## Secondary Structure

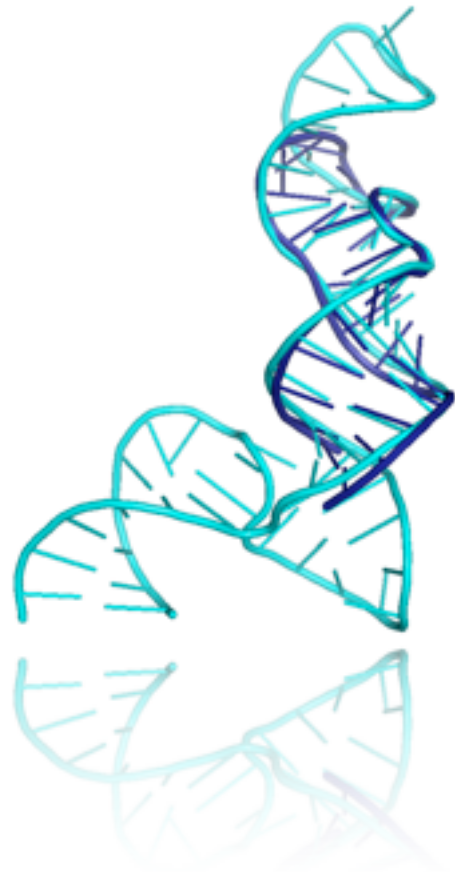
>Mutant Rat 28S rRNA sarcin/ricin domain  
GGUGCUCAGUAUGAGAAGAACCGCACC  
(((((((((. ((((. . ))))))) )))))))



## 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.

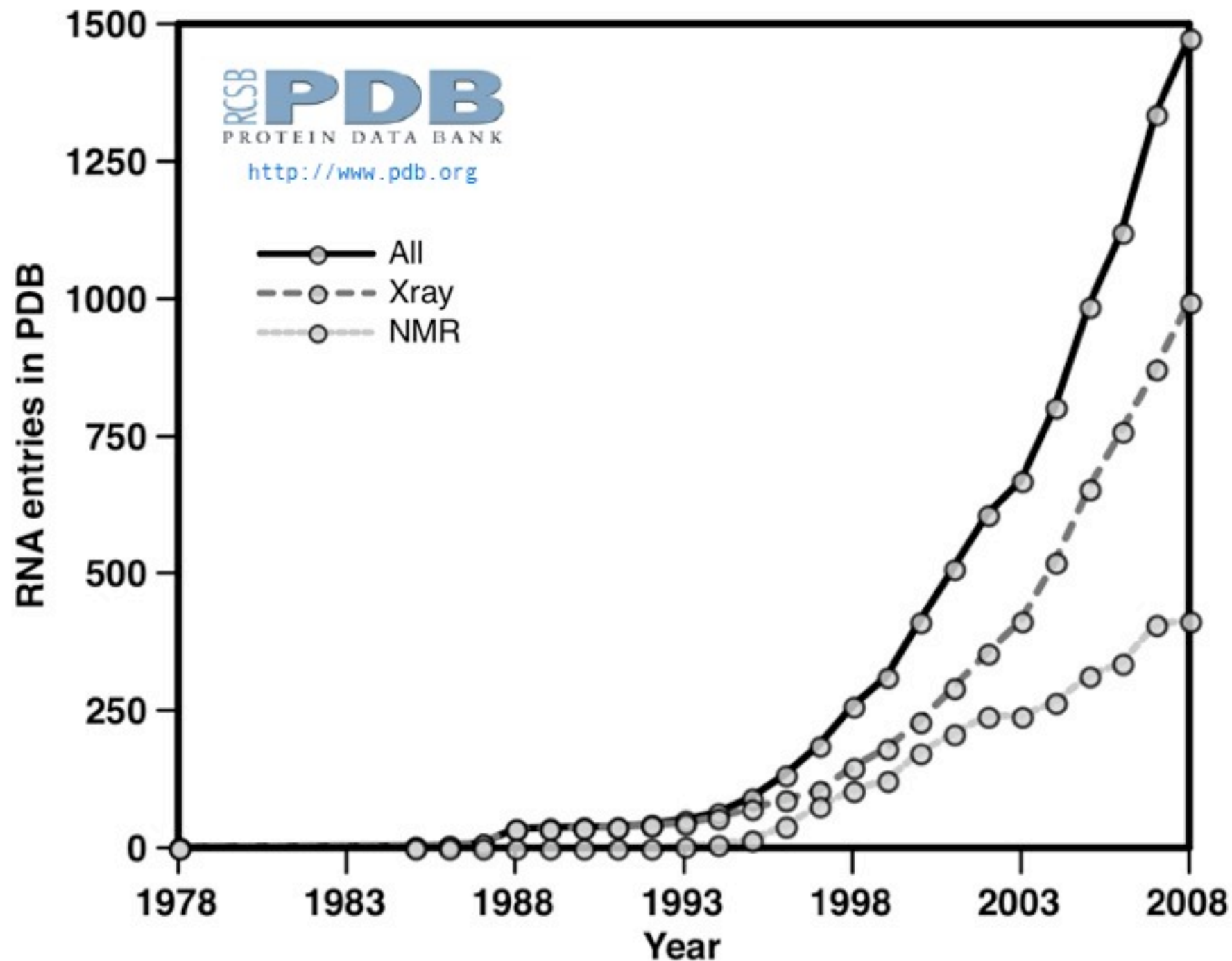
In contrast to simple structural superposition, where at least some equivalent residues of the two structures are known, structural alignment **does not require prior knowledge of the equivalent positions**.

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

Currently **more than 1500 RNA structures** are deposited in the PDB (Mar 09)



# RNA structure datasets

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

NR95

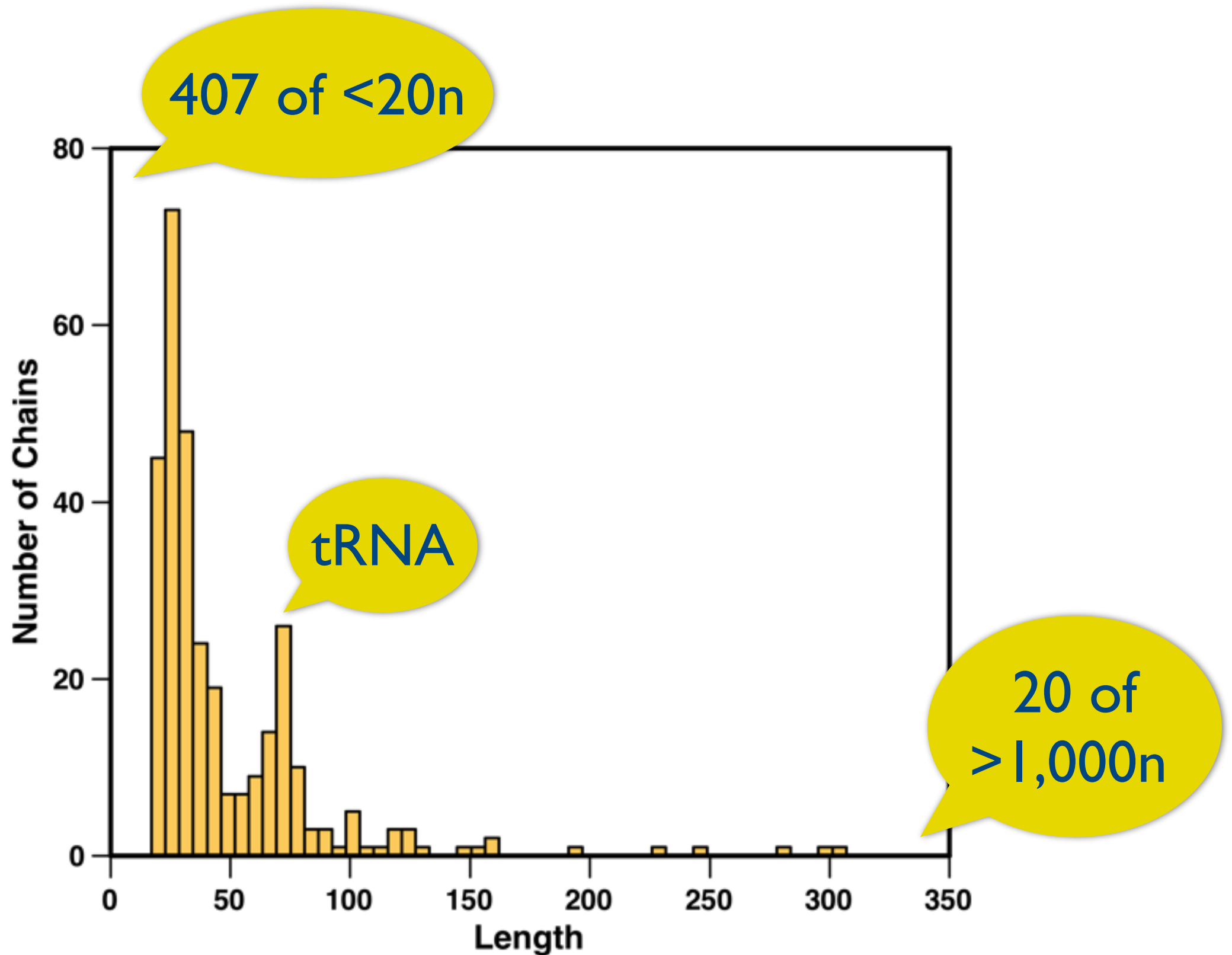
HR

\* from PDB November 06.

\*\* non-redundant 95% sequence identity

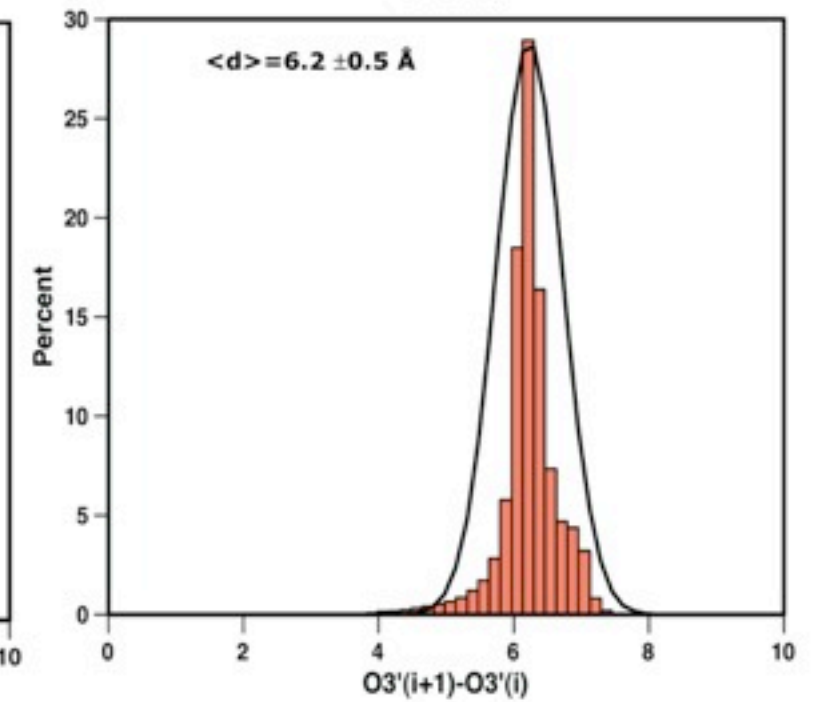
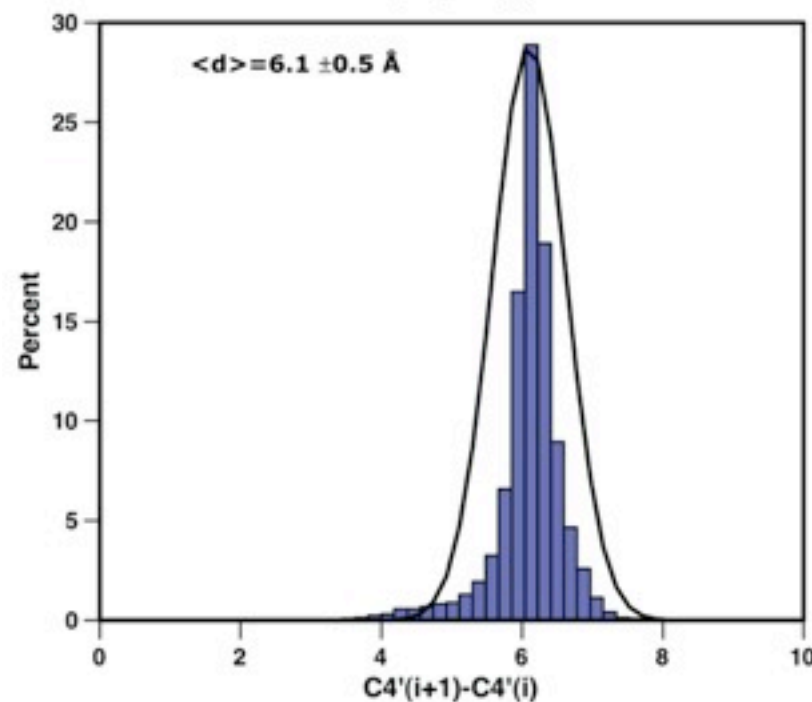
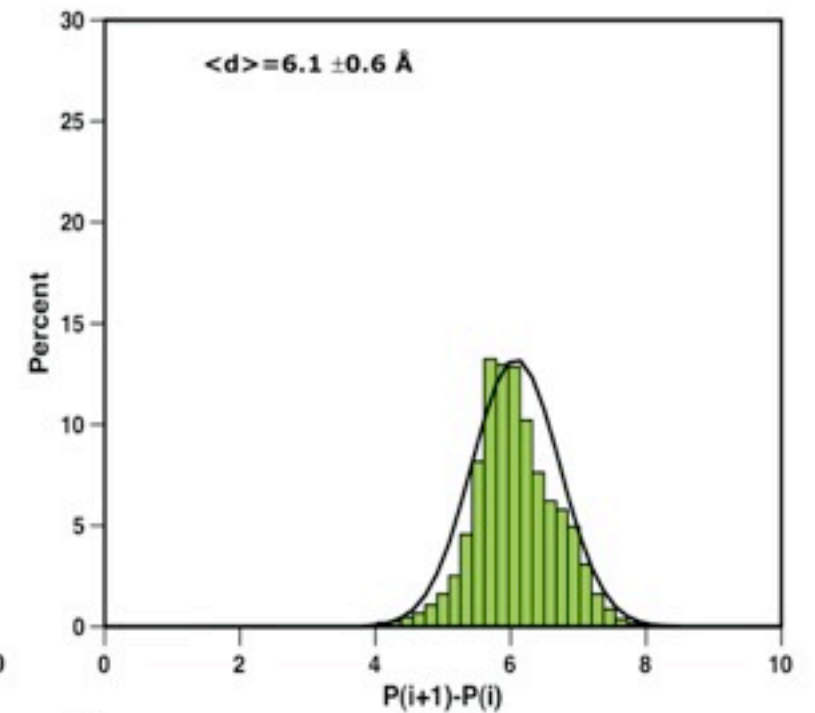
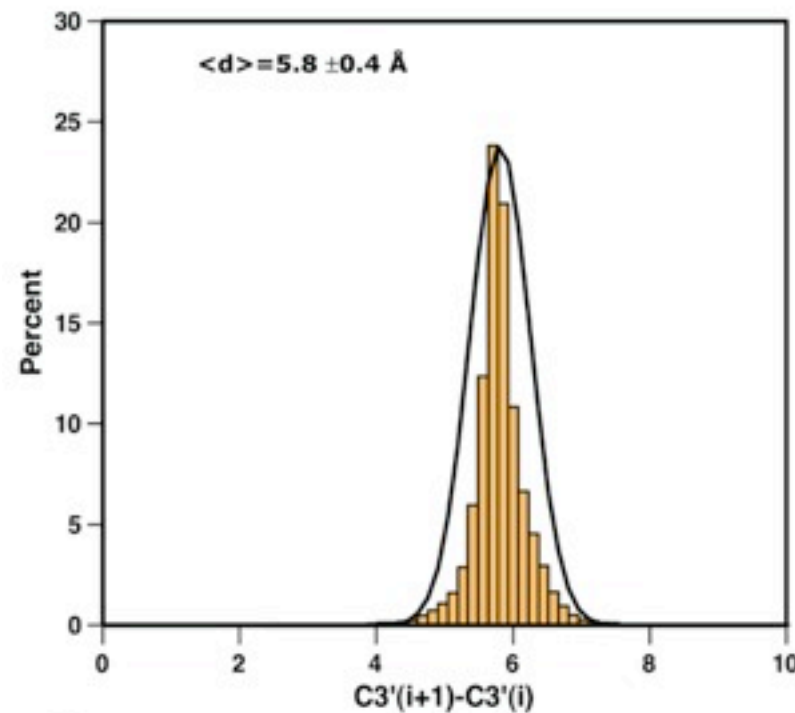
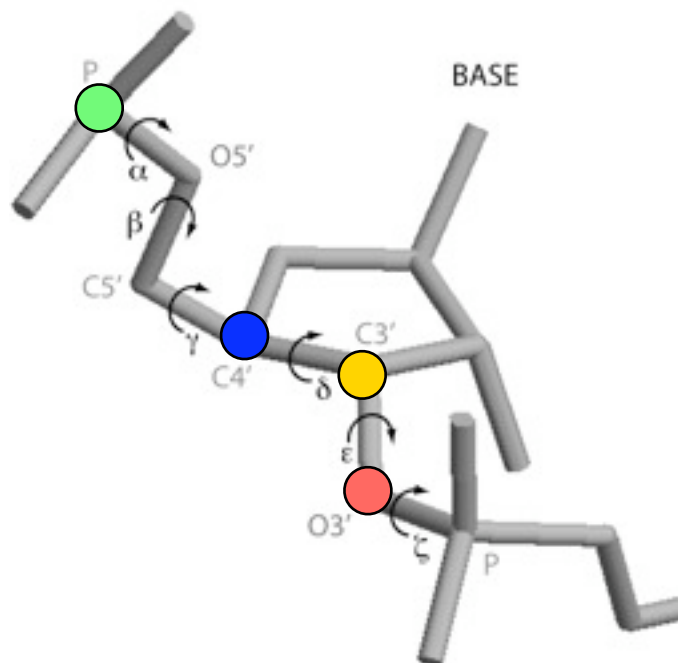
\*\*\* Resolution below 4.0 Å and with no missing backbone atoms.

# Dataset distribution



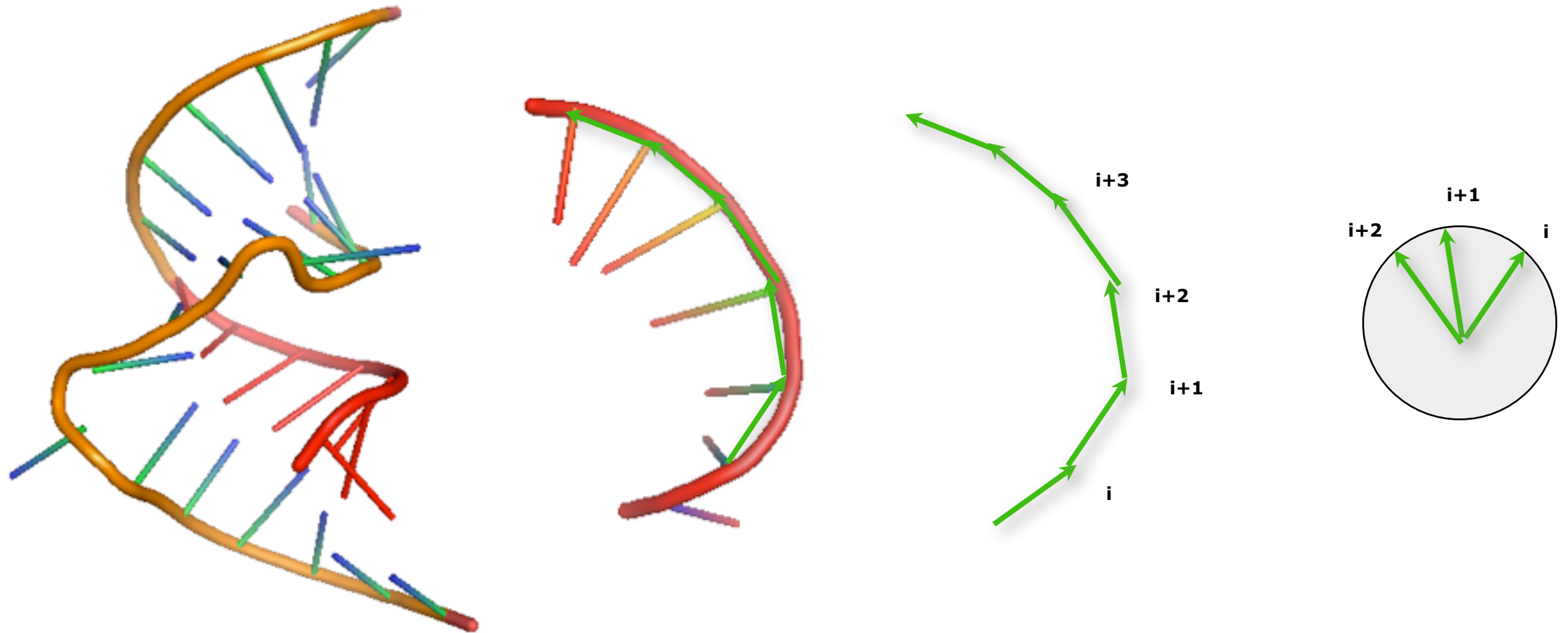
# Atom selection

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.

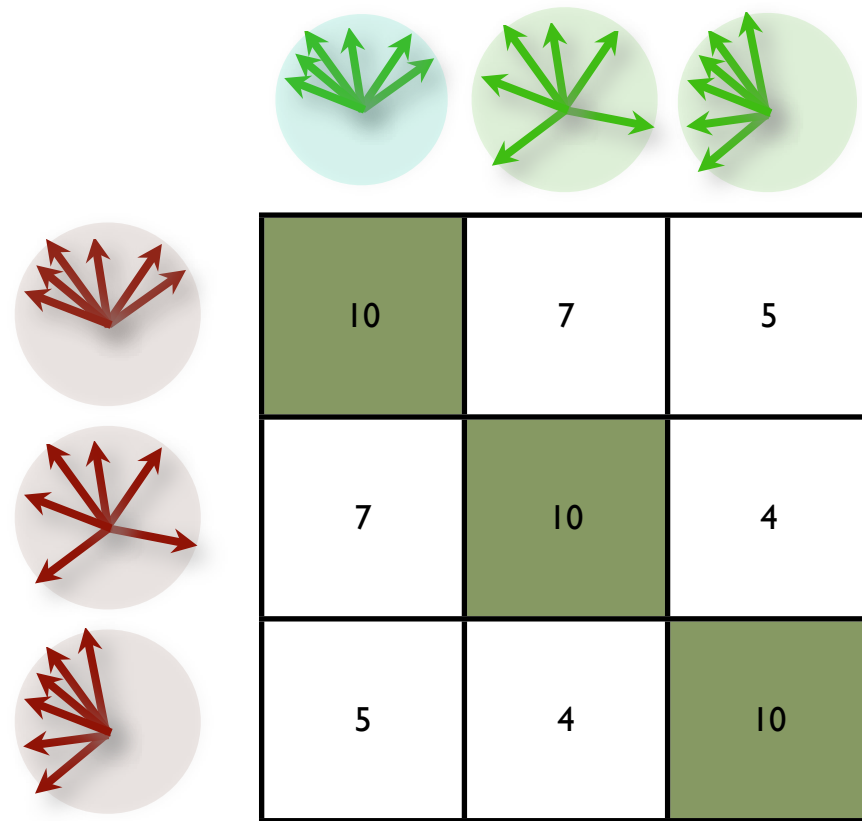
For each position  $i$  consider the  **$k$  consecutive vectors**, which will be mapped into a **unit sphere** representing the local structure of  $k$  residues.

Ortiz et al. *Proteins* 2002



# Unit Vector II

## Scoring



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

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

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

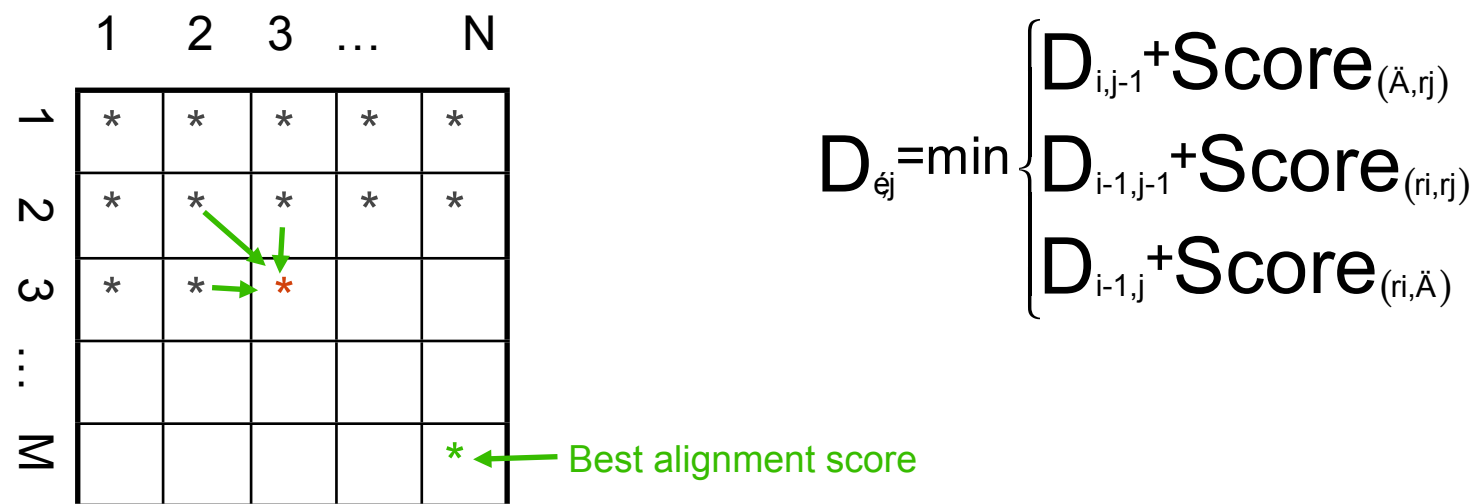
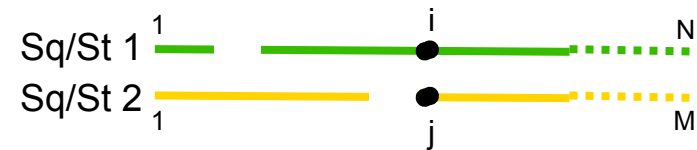
$$\Delta(U RMS^R, URMS^{ij}) = 0 \Rightarrow URMS^R \leq URMS^{ij}$$

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



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**.

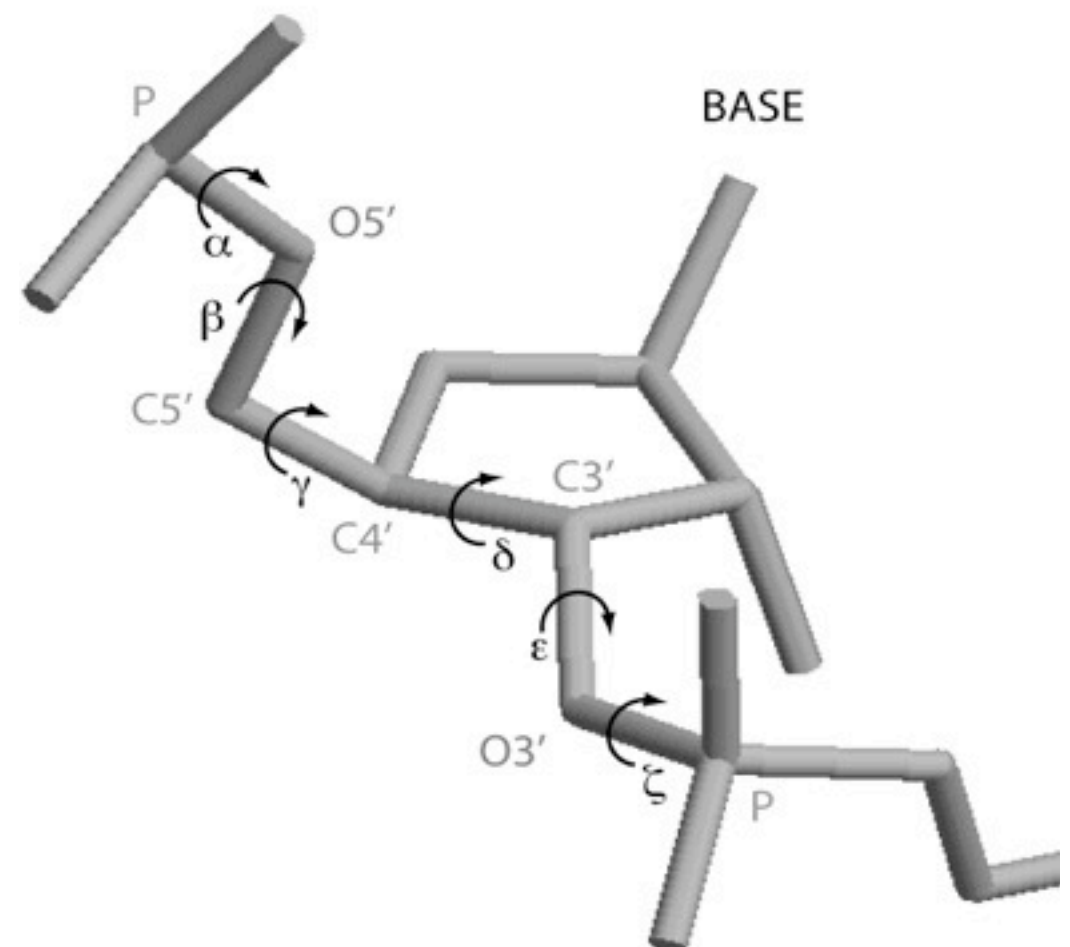
# Random RNA

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** ( $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ ) for each nucleotide.

The **RNA backbone is rotameric** and only 42 conformations have been described from a set of high resolution structures .

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

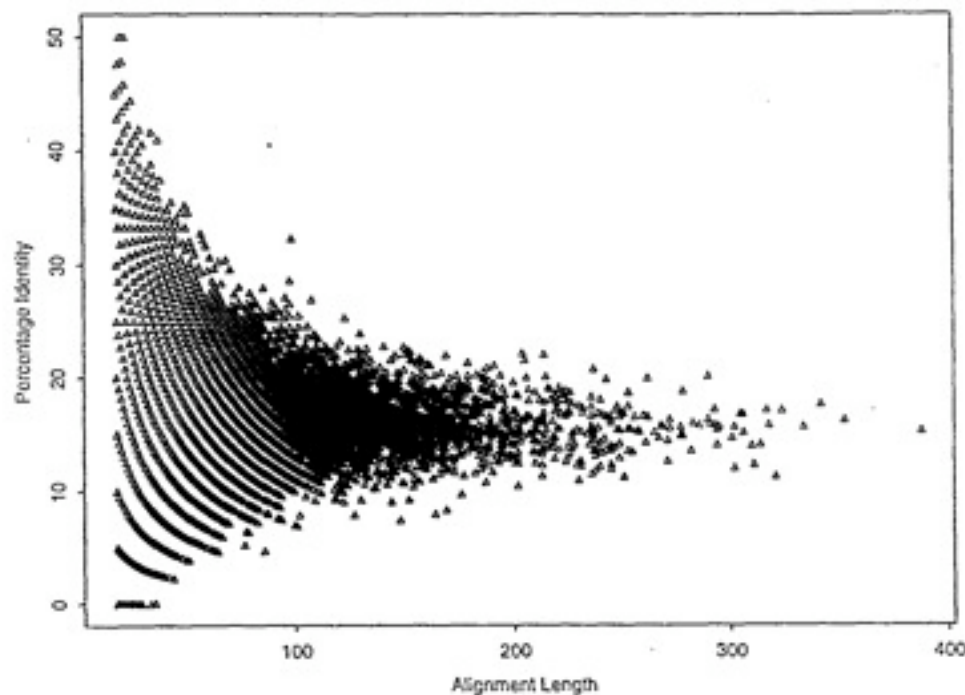


Murray et al PNAS 2003

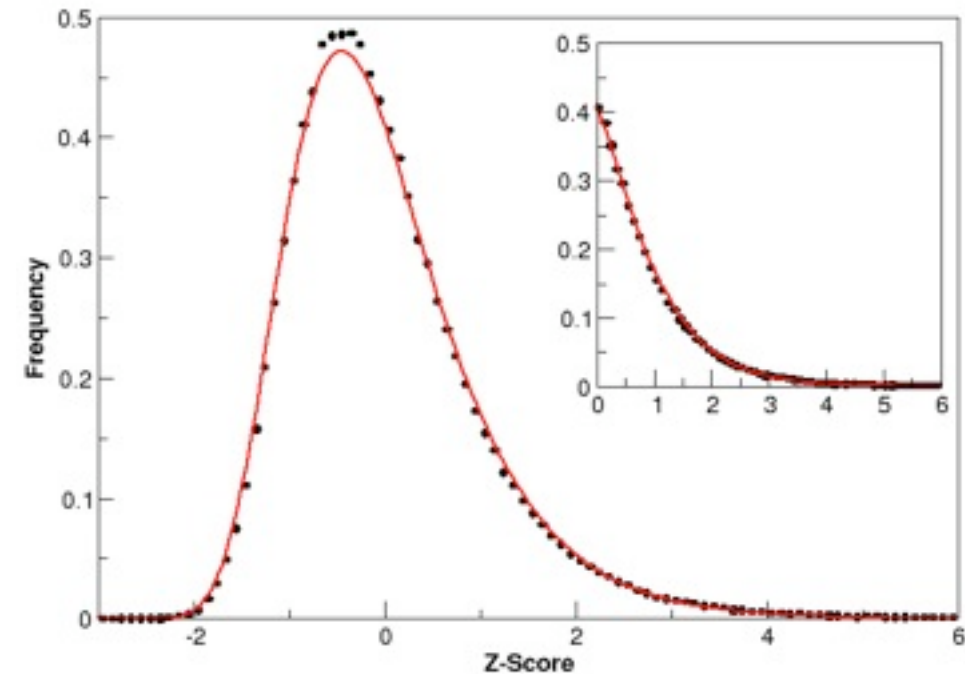
# Background distribution

Considering a dataset of 300 random RNA structures, we have produced ~45,000 pairwise alignments that resulted in an empirical distribution. From such distribution we can then evaluate  $\mu$  and  $\sigma$  needed to calculate the p-value for  $P(s \geq x)$ .

Empirical



Analytic



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

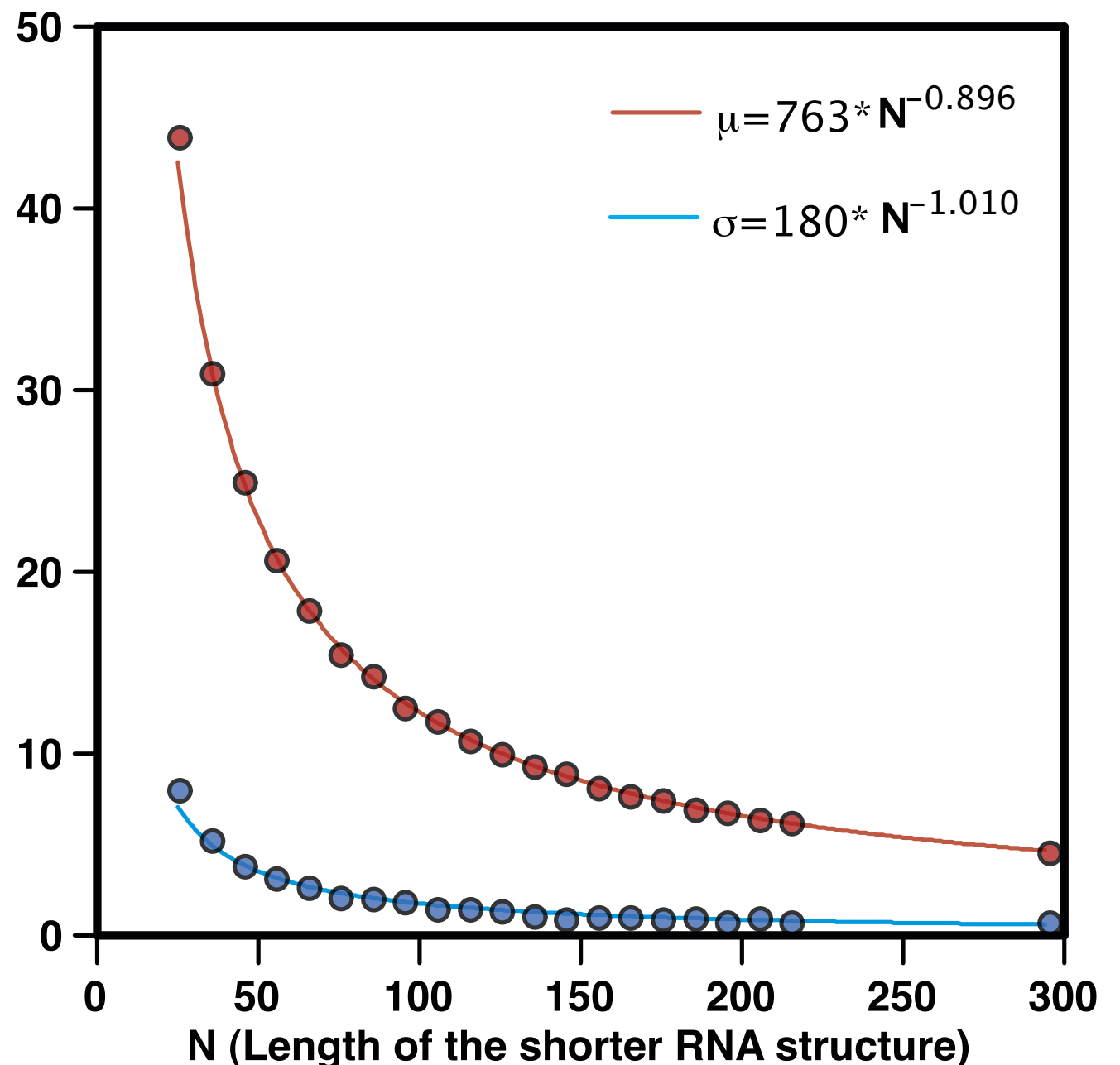
# Mean and sigma

The score distribution depends on the length of the molecule.

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

For each bin the  $\mu$  and  $\sigma$  values are evaluated fitting the data to an EVD.

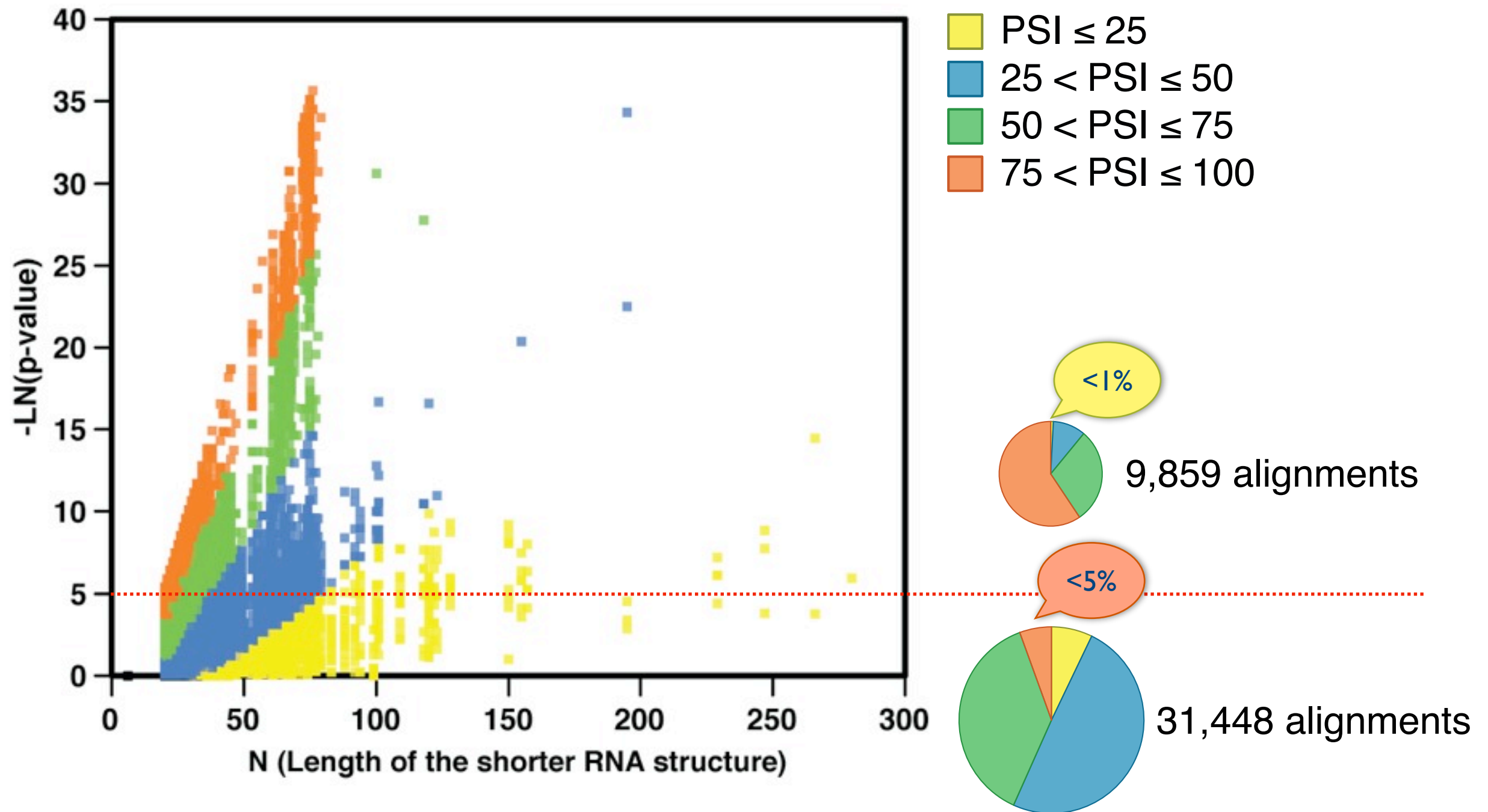
The relations between N and  $\mu$ ,  $\sigma$  values are extrapolate fitting them to a power low function ( $r \approx 0.99$ ).



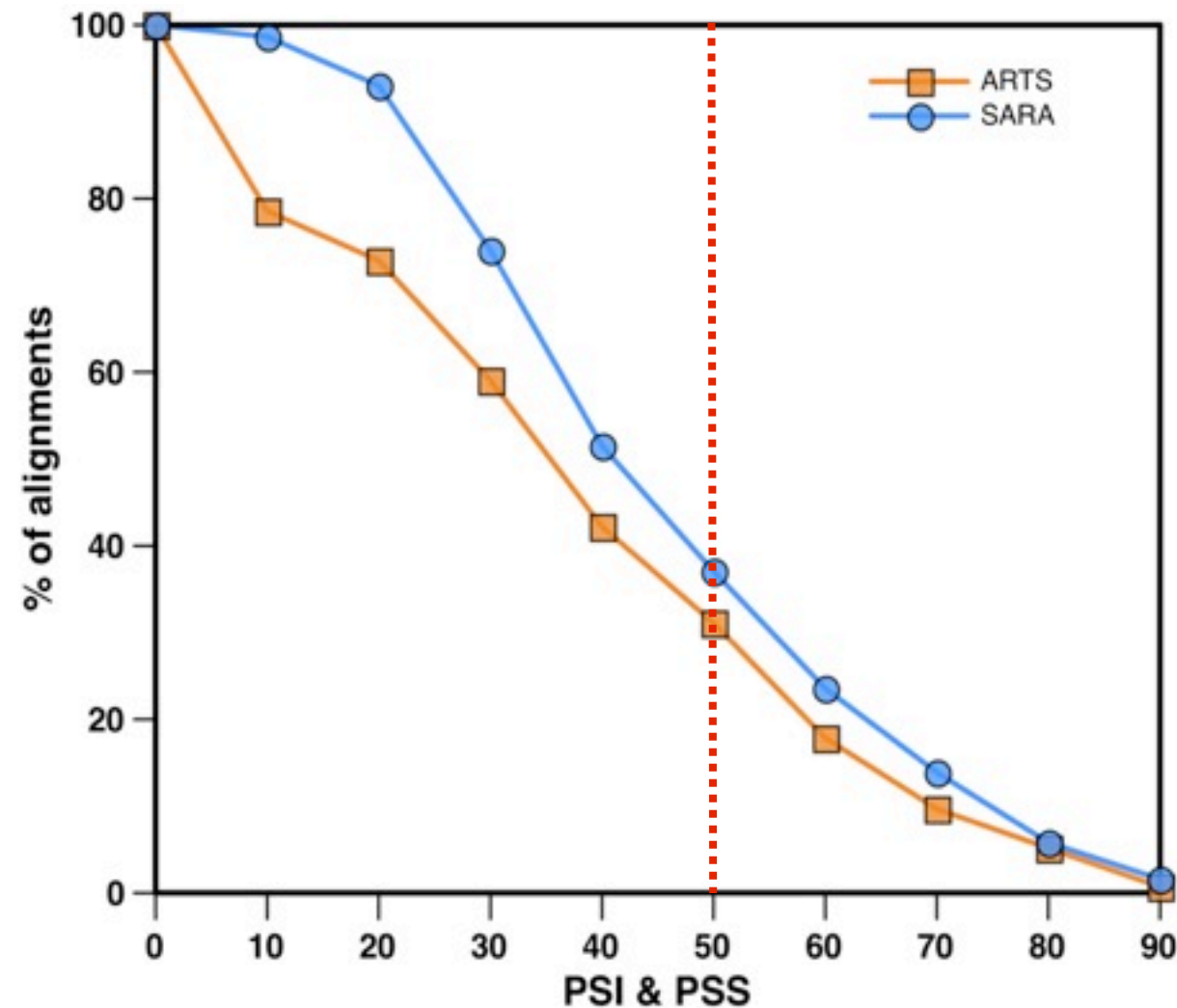


# Statistical significance

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



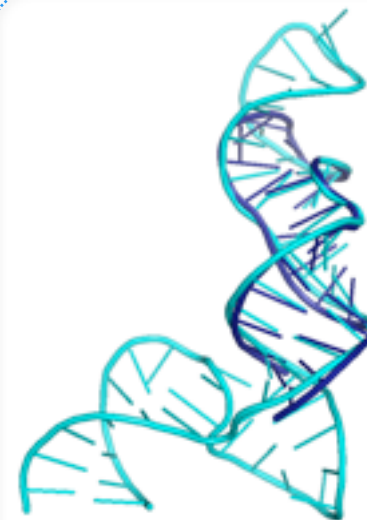
# SARA .vs. ARTS



**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

gccggccacaccuacggggccugguuaguaccugggaaaccugggaaauaccaggugccggc



## ARTS

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

>1q96 Chain:A

-----gugcucaguaugaga-----aga-accgcacc-----

>1un6 Chain:E

ccggccacaccuacggggccugguuaguaccugggaaaccugggaaauaccaggugccggc

# SARA Alignments

A) *Staphylococcus* phage group I ribozyme (1y8qA)  
Human group I Intron fragment (1u6bB)



Aligned nucleotides:	120
RMSD:	1.8 Å
Sequence Identity:	34.0 %
Secondary Structure Identity:	52.1 %
Structure Identity:	60.9 %
Sequence -ln(p-value):	18.2
Secondary structure -ln(p-value):	10.3
Structure -ln(p-value):	15.6
Mean -ln(p-value):	14.7

B) *Pyrococcus horikoshii* tRNA(Leu) (1wz2C)  
*Acuifex aeolicus* tRNA(Met) (2ct8C)



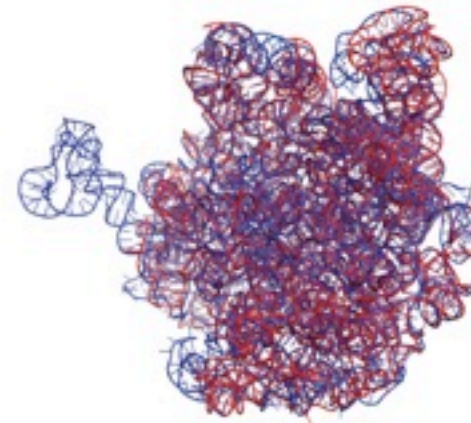
Aligned nucleotides:	65
RMSD:	1.9 Å
Sequence Identity:	56.8 %
Secondary Structure Identity:	88.5 %
Structure Identity:	87.8 %
Sequence -ln(p-value):	10.2
Secondary structure -ln(p-value):	5.2
Structure -ln(p-value):	7.2
Mean -ln(p-value):	7.5

C) Synthetic P4-P6 RNA ribozyme (118vA)  
*Mus musculus* P4-P6 RNA ribozyme (2r8sR)



Aligned nucleotides:	134
RMSD:	1.8 Å
Sequence Identity:	80.9 %
Secondary Structure Identity:	81.0 %
Structure Identity:	85.4 %
Sequence -ln(p-value):	37.0
Secondary structure -ln(p-value):	17.1
Structure -ln(p-value):	19.4
Mean -ln(p-value):	24.5

D) *Haloarcula marismortui* 23S RNA (3cce0)  
*Thermus thermophilus* 23S RNA (3d5bA)



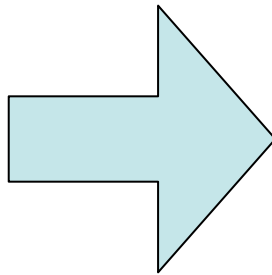
Aligned nucleotides:	2,347
RMSD:	1.7 Å
Sequence Identity:	52.7 %
Secondary Structure Identity:	75.7 %
Structure Identity:	85.2 %
Sequence -ln(p-value):	37.0
Secondary structure -ln(p-value):	37.0
Structure -ln(p-value):	37.0
Mean -ln(p-value):	37.0



# Second step...

Can we reliably predict RNA function from structure?





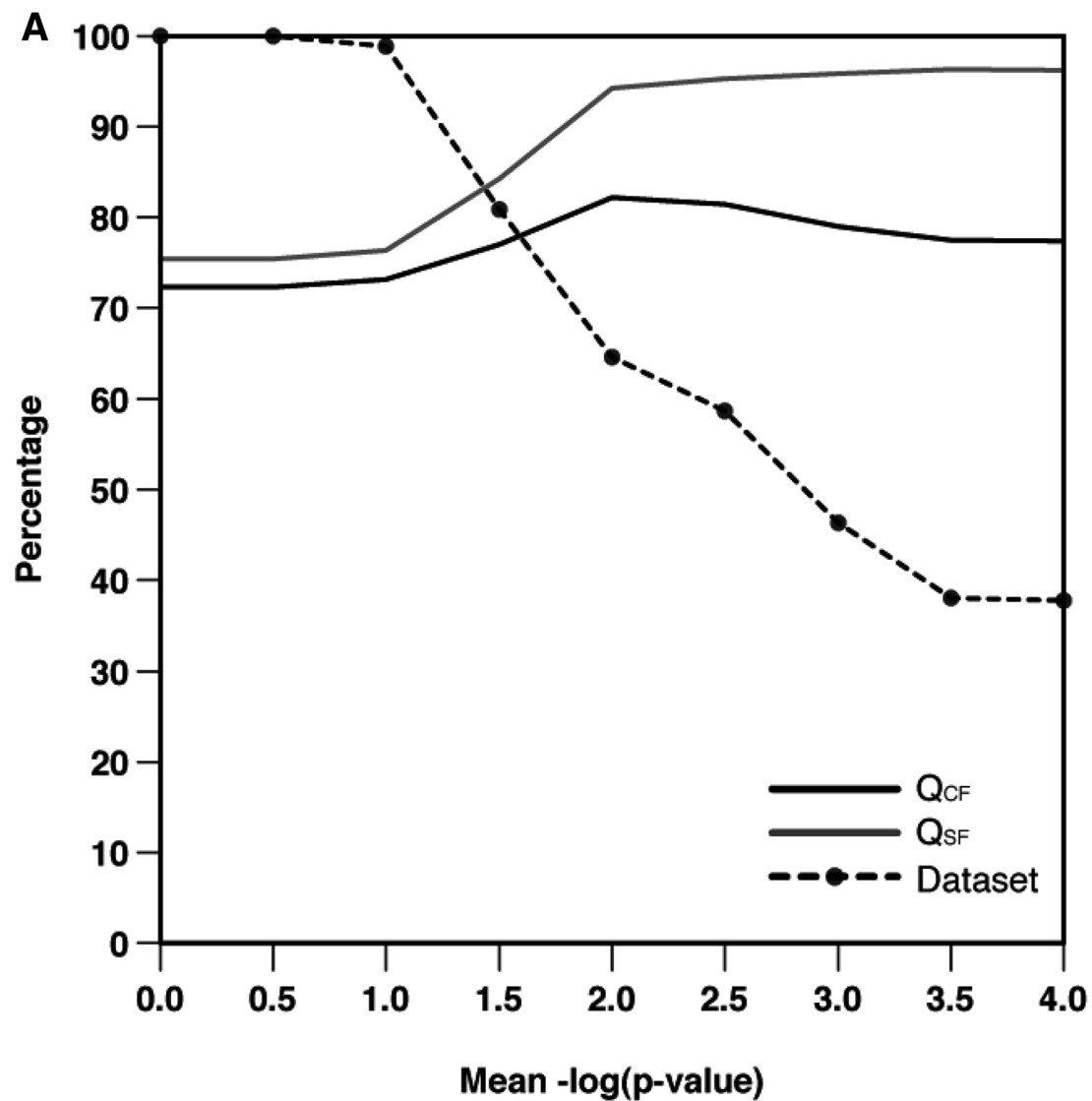
19



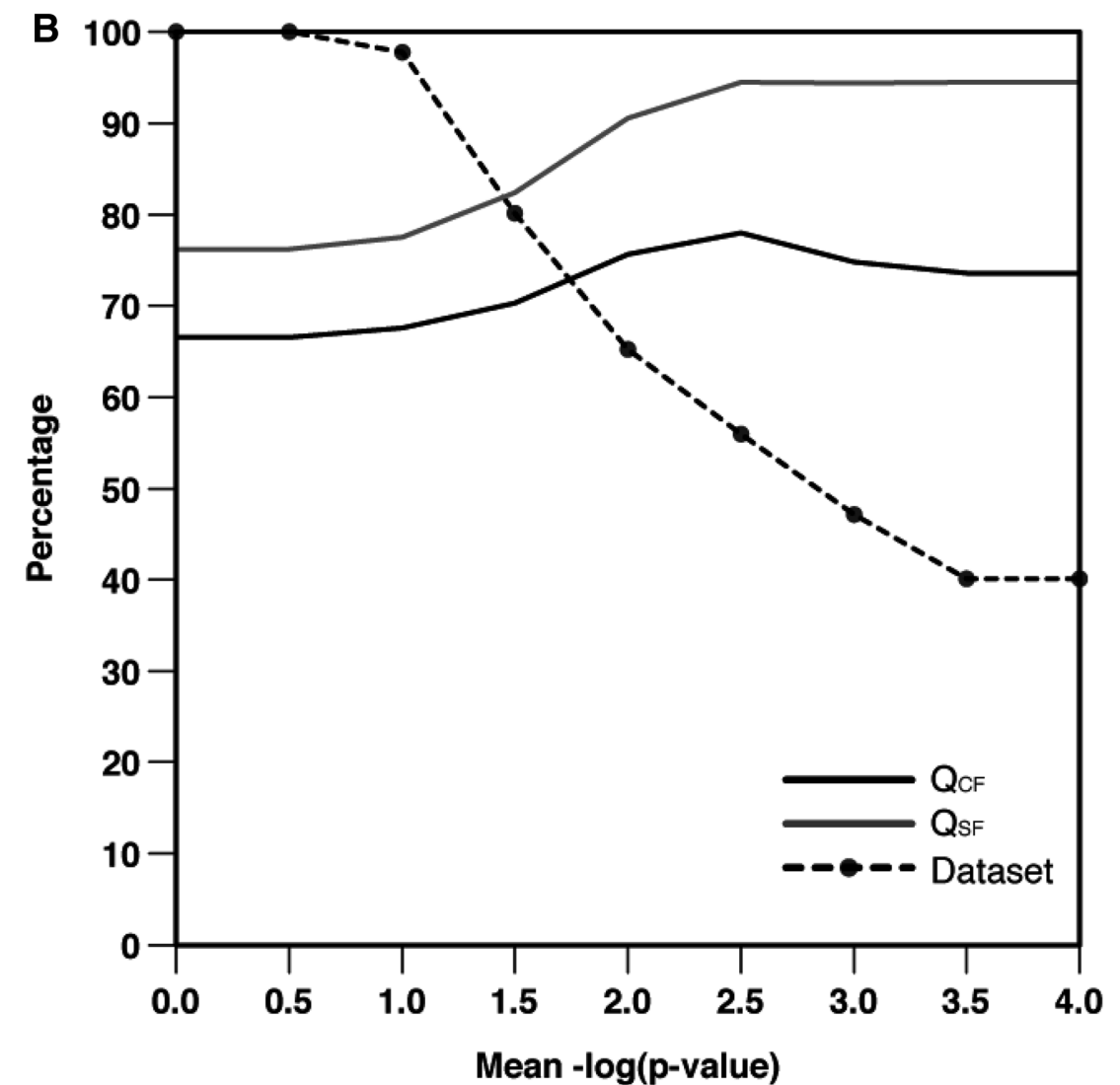
# Results

Datasets	Number of chains	Number of alignments	Number of different SCOR functions
RNA09	451	101 475	
BgALI	451	50 995	
FSCOR	419		168
R-FSCOR	192		168
T-FSCOR	227		88

leave one out on FSCOR

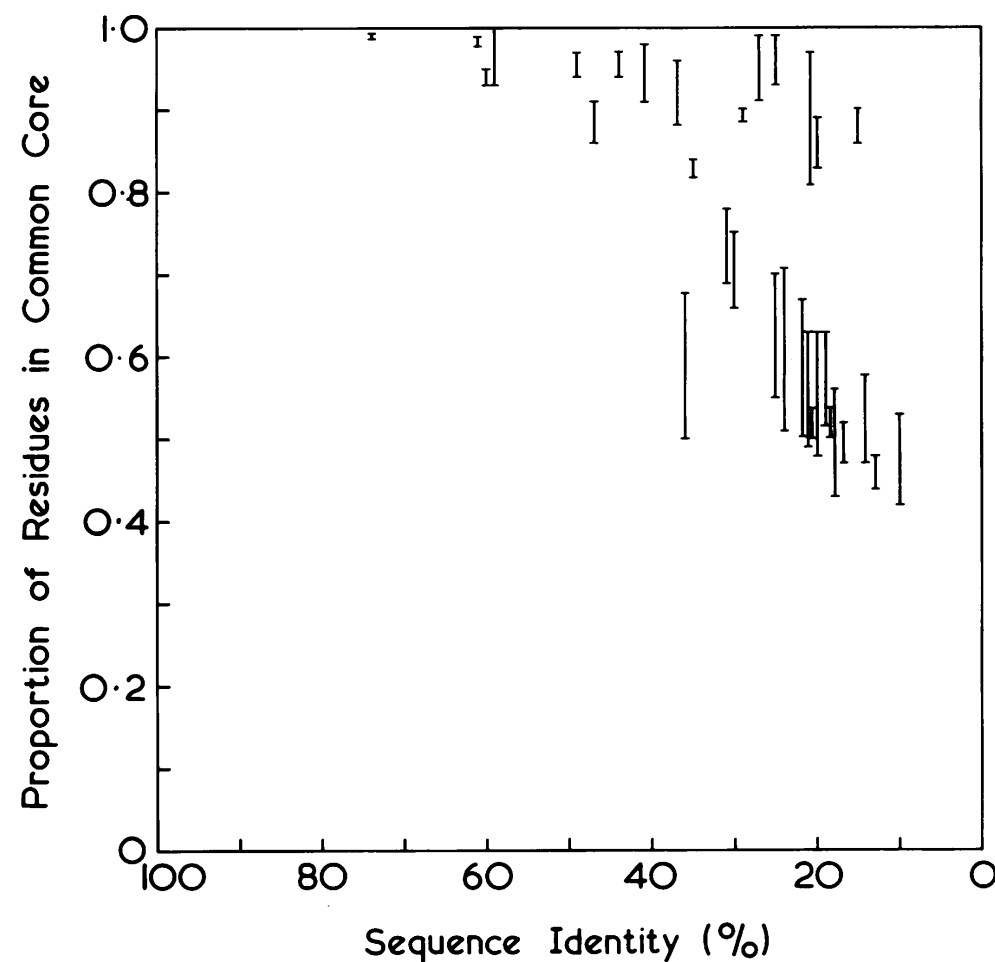


performances on T-FSCOR

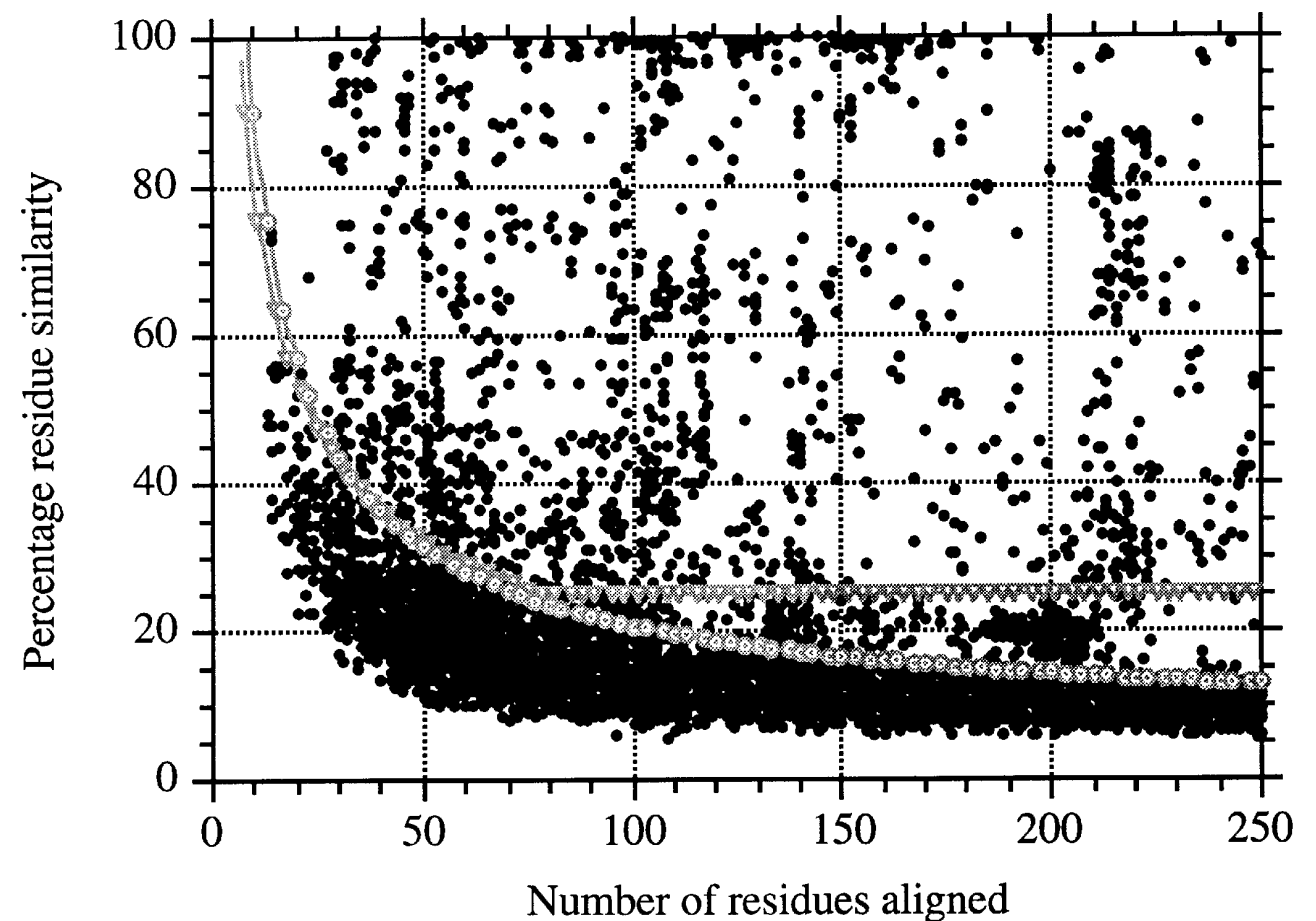


# Third step...

To what extent can we do comparative RNA structure prediction?

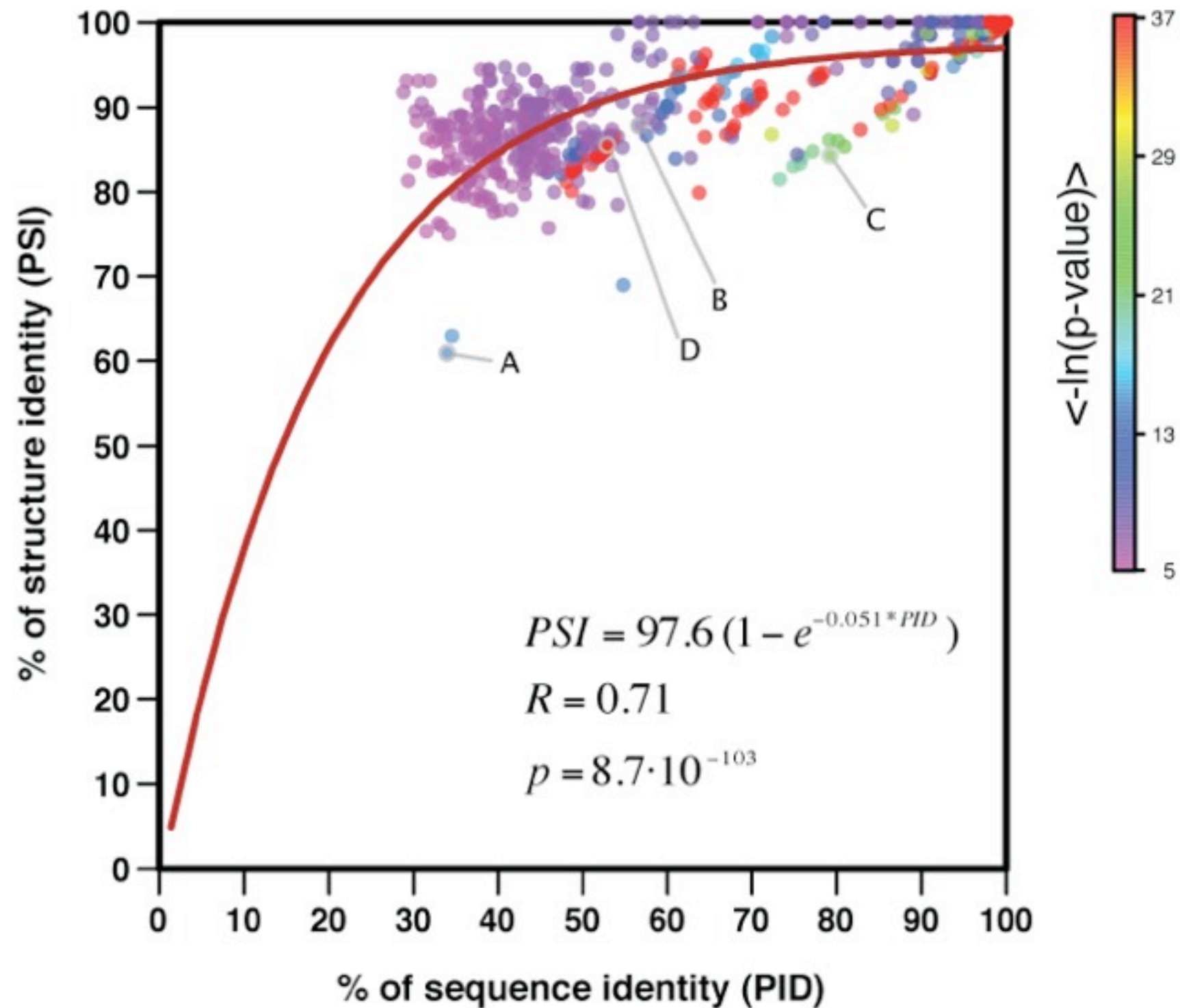
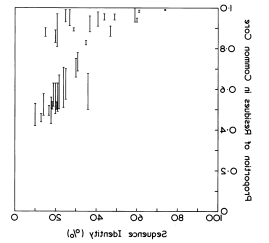


*Chothia et al.. EMBO J (1986) vol. 5 (4) pp. 823-6*

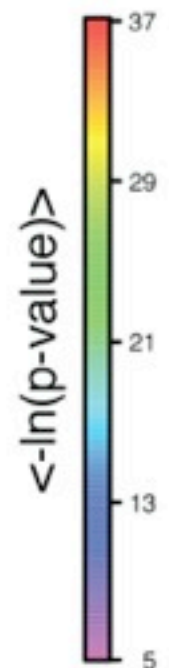
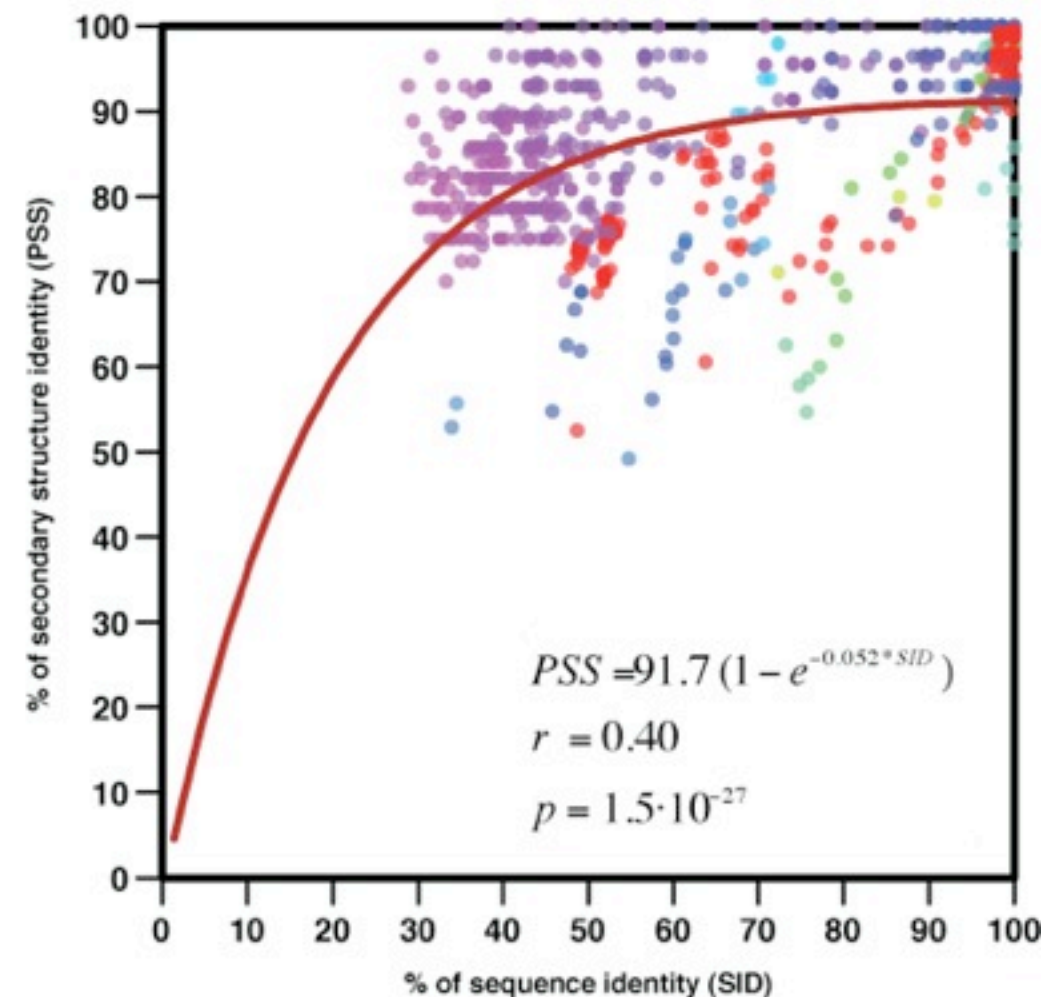
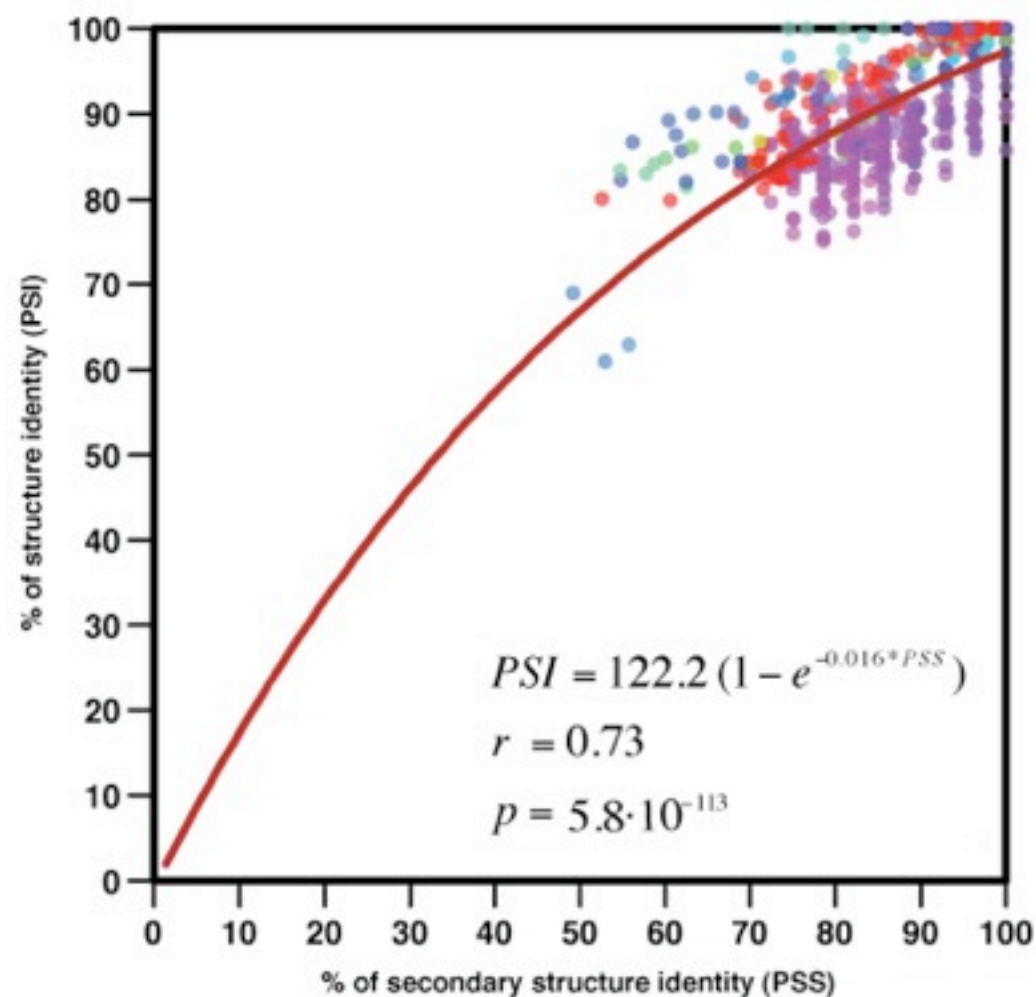
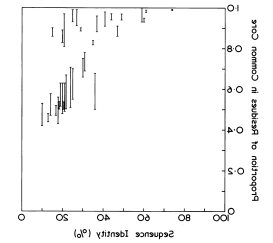


*Rost. Protein Eng (1999) vol. 12 (2) pp. 85-94*

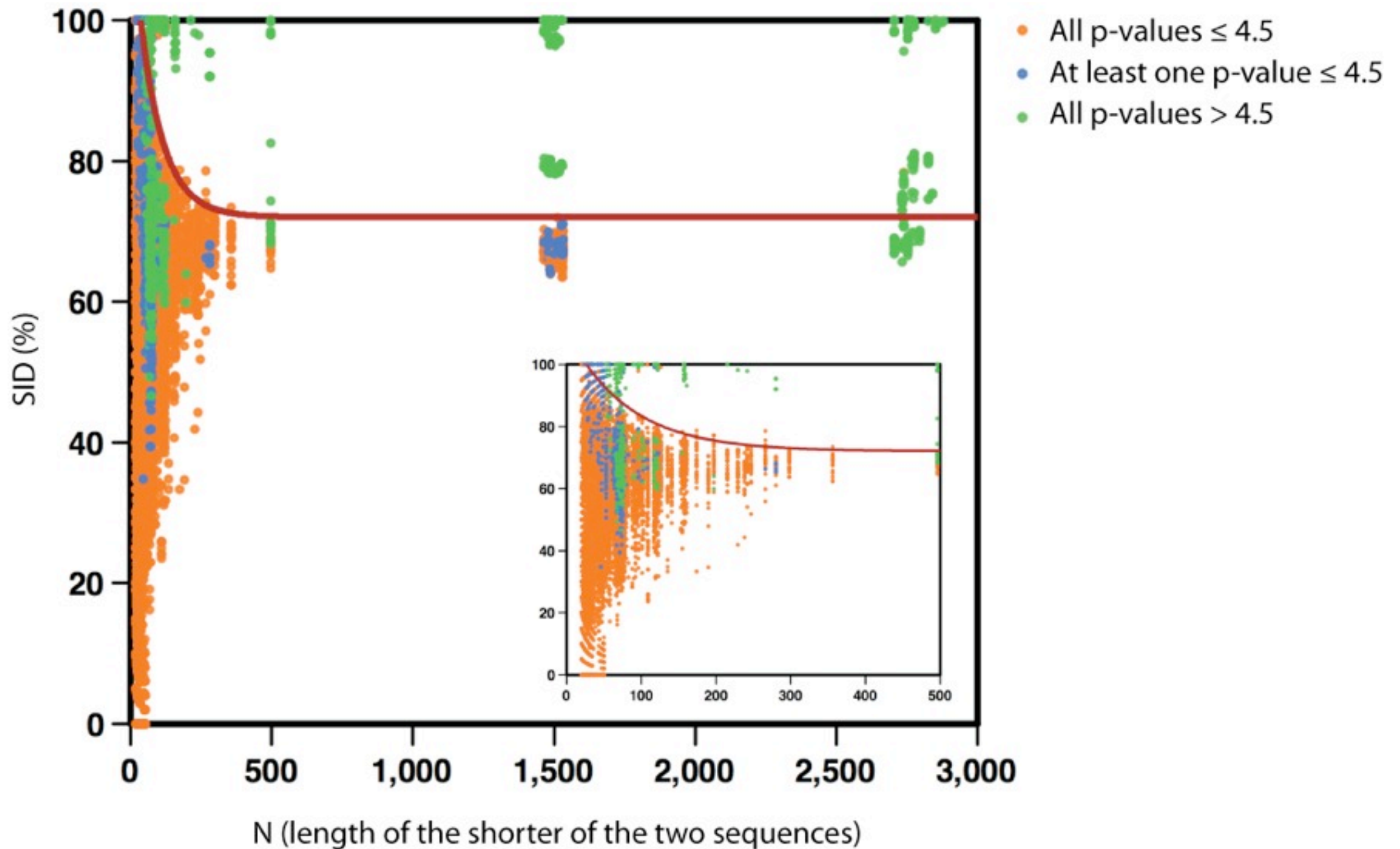
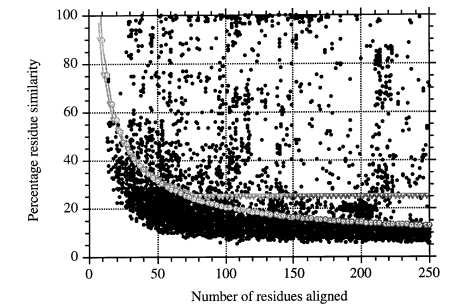
# Stx/Seq relationship



# SSE/Stx/Seq relationship



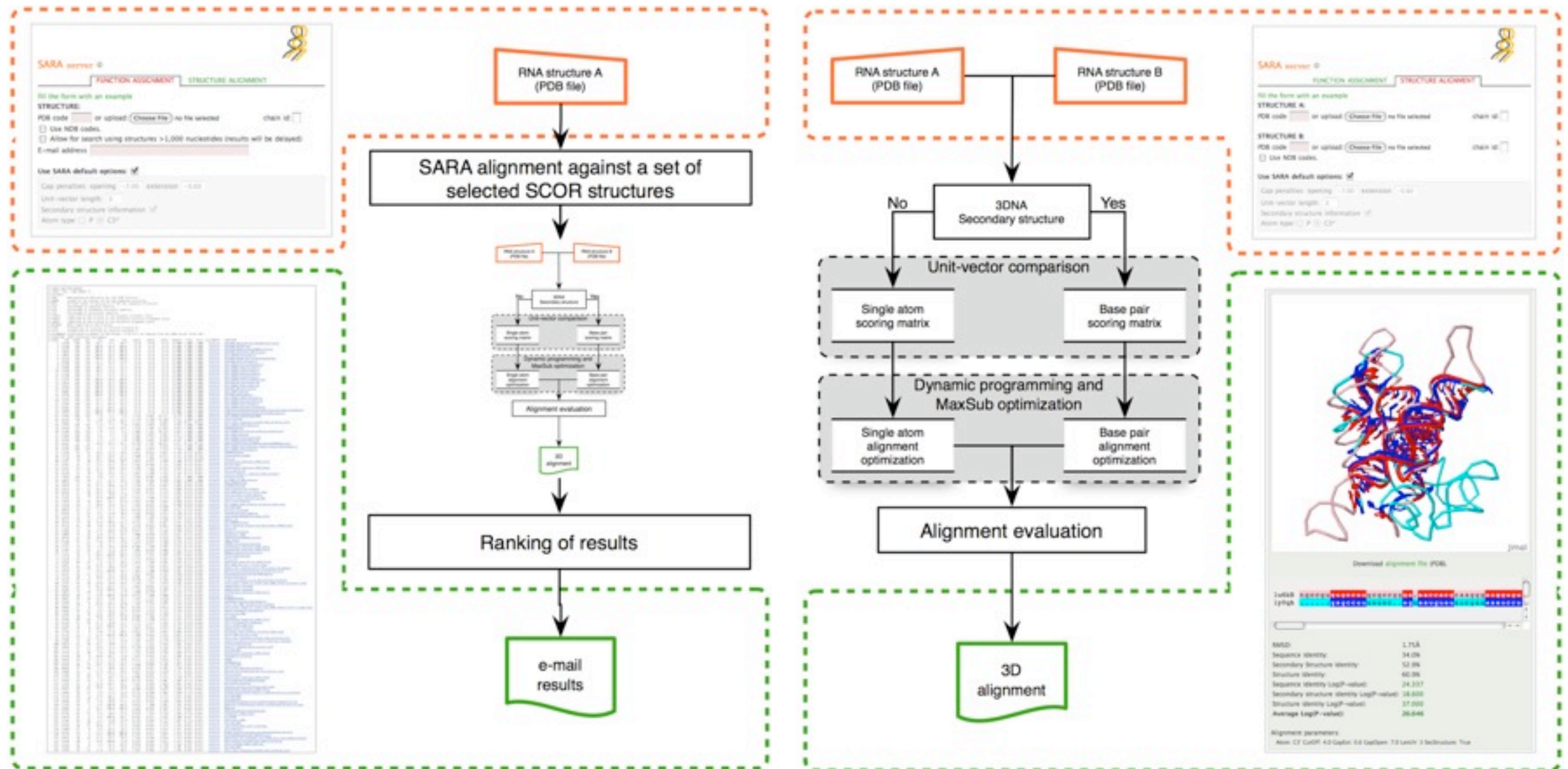
# Twilight Zone





# SARA server

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



Capriotti et al. Bioinformatics (2008) vol. 24 (16) pp. i112-i118

# SARA server

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

E. Capriotti, M. A. Marti-Renom (2008), *Bioinformatics* **24**:i112

E. Capriotti, M. A. Marti-Renom. (2009). *NAR* **37**:W260

## BIOINFORMATICS

Vol. 24 ECCB 2008, pages i112–i118  
doi:10.1093/bioinformatics/btn288

### RNA structure alignment by a unit-vector approach

Emidio Capriotti and Marc A. Marti-Renom\*

Bioinformatics and Genomics Department, Structural Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

#### ABSTRACT

**Motivation:** The recent discovery of tiny RNA molecules such as  $\mu$ RNAs and small interfering RNA are transforming the view of RNA as a simple information transfer molecule. Similar to proteins, the native three-dimensional structure of RNA determines its biological activity. Therefore, classifying the current structural space is paramount for functionally annotating RNA molecules. The increasing numbers of RNA structures deposited in the PDB requires more accurate, automatic and benchmarked methods for RNA structure comparison. In this article, we introduce a new algorithm for RNA structure alignment based on a unit-vector approach. The algorithm has been implemented in the SARA program, which results in RNA structure pairwise alignments and their statistical significance.

**Results:** The SARA program has been implemented to be of general applicability even when no secondary structure can be calculated from the RNA structures. A benchmark against the ARTS program using a set of 1275 non-redundant pairwise structure alignments results in ~6% extra alignments with at least 50% structurally superposed nucleotides and base pairs. A first attempt to perform RNA automatic functional annotation based on structure alignments indicates that SARA can correctly assign the deepest SCOR classification to >60% of the query structures.

**Availability:** The SARA program is freely available through a World Wide Web server <http://sgu.bioinfo.cipf.es/services/SARA/>

**Contact:** mmarti@cipf.es

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

#### 1 INTRODUCTION

Recent discoveries of new RNA functions are changing our view of RNA molecules and reinforcing the so-called 'RNA world' origin of life (Bartel, 2004; Dorsett and Tuschl, 2004; Doudna, 2000; Staple and Butcher, 2005). RNA is now known to play an important role in biological functions such as enzymatic activity (Staple and Butcher, 2005), gene transcriptional regulation (Bartel, 2004; Dorsett and Tuschl, 2004; Staple and Butcher, 2005) and protein biosynthesis regulation (Doudna, 2000). Therefore, much attention is lately being paid to the structural determination of RNA molecules. Such efforts have increased the pace of deposition of RNA structures in the Protein Data Bank (PDB) (Berman *et al.*, 2002). Currently (January 2008), the PDB database stores more than 1300 RNA structures. Such a wealth of data may allow, for first time, the analysis and characterization of the RNA structural space, which will help to characterize RNA function.

RNA folding is a hierarchical process by which base pairing formation affects the final three-dimensional (3D) conformation

\*To whom correspondence should be addressed.

of the RNA molecule (Tinoco and Bustamante, 1999). Hence, algorithms for RNA secondary structure prediction have classically been used for characterizing RNA structure and function. Although more than two decades have past since the development of the first algorithms for RNA secondary structure prediction (Nussinov and Jacobson, 1980; Zuker and Sankoff, 1984; Zuker and Stiegler, 1981), there has been limited development in RNA tertiary structure analysis and, in particular, in RNA structure comparison. Only recently, the PRIMOS/AMIGOS (Duarte *et al.*, 2003; Wadley *et al.*, 2007), FR3D (Sarver *et al.*, 2008), ARTS (Dror *et al.*, 2005, 2006) and DIAL (Ferre *et al.*, 2007) programs have been developed for structurally comparing two RNA molecules. The PRIMOS/AMIGOS programs search for structural similarities of consecutive RNA fragments with five or more nucleotides by comparing specific  $\eta$  and  $\theta$  pseudo angles as well as the sugar pucker phase. The FR3D program uses a base-centered approach for conducting a geometric search of local and composite RNA structures. The COMPADRES program, which implements the PRIMOS algorithm, has been applied for searching local structural motifs in known RNA structures (Wadley and Pyle, 2004). The ARTS program, which represents RNA structures by a set of contiguous four phosphate atoms or *quadrats*, detects similarities between *quadrats* after a rigid superimposition of two RNA structures followed by an optimization based on a bipartite graph strategy. Finally, the DIAL program, which implements a scoring function combining nucleotide, dihedral angles and base-pairing similarities, compares the two RNA structures using a dynamic programming algorithm.

Although the PRIMOS/AMIGOS, ARTS and DIAL programs, result in accurate RNA structure alignments, they have some limitations: (i) the PRIMOS/AMIGOS program have limited applicability to searching only for local motifs regardless of global similarities between two structures, (ii) the DIAL method, in its default version, only calculates an alignment score and requires substantial computational time to return a statistical evaluation of its significance and (iii) ARTS requires the existence of secondary structure elements in both structures to compute the final alignment. To overcome such limitations, we have developed a new RNA 3D alignment method (SARA), which does not require the assignment of base pairs from structure and provides a statistical assessment of the significance of the resulting alignment. The SARA algorithm uses a unit-vector approach inspired by the MAMMOTH program for protein structure alignment (Ortiz *et al.*, 2002). The SARA program has been benchmarked for its alignment accuracy against the ARTS program as well as for its use in RNA function prediction. Its general applicability will allow an all-against-all comparison of known RNA structures, which will help in characterizing the relationship between sequence, structure and function of RNA molecules.

W260–W265 Nucleic Acids Research, 2009, Vol. 37, Web Server issue  
doi:10.1093/nar/gkp433

Published online 29 May 2009

### SARA: a server for function annotation of RNA structures

Emidio Capriotti and Marc A. Marti-Renom\*

Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Received February 8, 2009; Revised May 5, 2009; Accepted May 11, 2009

#### ABSTRACT

**Recent interest in non-coding RNA transcripts has resulted in a rapid increase of deposited RNA structures in the Protein Data Bank. However, a characterization and functional classification of the RNA structure and function space have only been partially addressed. Here, we introduce the SARA program for pair-wise alignment of RNA structures as a web server for structure-based RNA function assignment. The SARA server relies on the SARA program, which aligns two RNA structures based on a unit-vector root-mean-square approach. The likely accuracy of the SARA alignments is assessed by three different P-values estimating the statistical significance of the sequence, secondary structure and tertiary structure identity scores, respectively. Our benchmarks, which relied on a set of 419 RNA structures with known SCOR structural class, indicate that at a negative logarithm of mean P-value higher or equal than 2.5, SARA can assign the correct or a similar SCOR class to 81.4% and 95.3% of the benchmark set, respectively. The SARA server is freely accessible via the World Wide Web at <http://sgu.bioinfo.cipf.es/services/SARA/>.**

#### INTRODUCTION

It is now known that RNA molecules are essential for a wide range of biological processes (1–6), which is changing the view of RNA as a simple vector of genetic information and reinforcing the hypothesis on the original 'RNA world' (7,8). Biosynthesis and transcription regulation (1–3,5), enzymatic action (5) and chromosome replication (4) are some of the functions that RNA molecules are now known to perform. RNA structure determination, which is accelerating its pace of deposition in the Nucleic Acid Database (NDB) (9) and the Protein Data Bank (PDB) (10), is thus becoming an essential and necessary tool for RNA function annotation. Although there are not

standard rules to infer function, at least for proteins (11–13), structure similarity is arguably one of the most reliable methods for comparative function annotation (14,15).

Several methods have already been developed for the alignment of two or more protein 3D structures (16). However, only few are available for RNA structure comparison (17–23). The PRIMOS and AMIGOS programs identify RNA structure motifs and compare RNA structures by describing them as a set of pseudo angles from the C4' and P atom trace (18,20). Both programs are limited to the comparison of RNA structures with the same number of nucleotides and only a newer version of AMIGOS can perform a comparison of a given structure against a set of RNA structures. The ARTS program was introduced as a general method for RNA structure alignment (17,24). ARTS describes RNA molecules with a set of 'quadrats' composed by four phosphate atoms of two consecutive base-pairs and uses a bipartite graph to find the maximum number of aligned 'quadrats' between two RNA structures. The DIAL program, developed to compare RNA structures using a dynamic programming algorithm (19), computes global, local and semi-global alignments by taking into account sequence similarity, dihedral angles and base-pair information from the two aligned structures. DIAL can also return the Boltzmann pair probabilities of the resulting alignments. However, such computation would double the runtime, hence the default in the DIAL server is not to calculate the pair probabilities. More recently, the SARSA server was developed to align two or more RNA structures using a structural alphabet of 23 nucleotide conformations (22). Both, the DIAL and SARSA servers were developed and benchmarked for their ability detecting short RNA motifs in a set of RNA structures. In contrast, the SARA program (21), which implementation for function assignment of RNA structures is here introduced, was recently developed to align two RNA structures based on a unit-vector alignment strategy (25). Given its implementation, an alignment by SARA shorter than 20 nt is likely to be indistinguishable from random structure alignments. The SARA program can be considered as an alternative

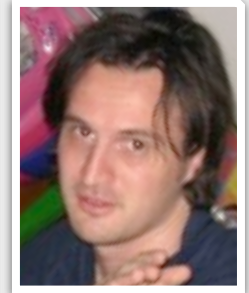
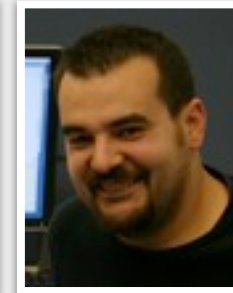
\*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: mmarti@cipf.es

© 2009 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Acknowledgments

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



## COMPARATIVE MODELING

**Andrej Sali**

M. S. Madhusudhan

**Narayanan Eswar**

Min-Yi Shen

**Ursula Pieper**

Ben Webb

Maya Topf (Birbeck College)

## RNA Alignments

Emidio Capriotti

## NMR

Antonio Pineda-Lucena

Leticia Ortí

Rodrigo J. Carbajo

## MAMMOTH

Angel R. Ortiz

## FUNCTIONAL ANNOTATION

Fatima Al-Shahrour

Joaquin Dopazo

## BIOLOGY

Jeff Friedman (RU)

James Hudsped (RU)

Partho Ghosh (UCSD)

Alvaro Monteiro (Cornell U)

Stephen Krilis (St. George H)

## Tropical Disease Initiative

Stephen Maurer (UC Berkeley)

Arti Rai (Duke U)

Andrej Sali (UCSF)

Ginger Taylor (TSL)

Matthew Todd (U Sydney)

## CCPR Functional Proteomics

Patsy Babbitt (UCSF)

Fred Cohen (UCSF)

Ken Dill (UCSF)

Tom Ferrin (UCSF)

John Irwin (UCSF)

Matt Jacobson (UCSF)

Tack Kuntz (UCSF)

Andrej Sali (UCSF)

Brian Shoichet (UCSF)

Chris Voigt (UCSF)

## EVA

Burkhard Rost (Columbia U)

Alfonso Valencia (CNB/UAM)

## CAMP

Xavier Aviles (UAB)

Hans-Peter Nester (SANOFI)

Ernst Meinjohanns (ARPIDA)

Boris Turk (IJS)

Markus Gruetter (UE)

Matthias Wilmanns (EMBL)

Wolfram Bode (MPG)

## MODEL ASSESSMENT

David Eramian

Min-Yi Shen

Damien Devos

## FUNCTIONAL ANNOTATION

Andrea Rossi (Rinat-Pfizer)

Fred Davis (Janelia Fram)

## FUNDING

Prince Felipe Research Center

Ministerio de Educación y Ciencia

STREP UE Grant

**Marie Curie Reintegration Grant**