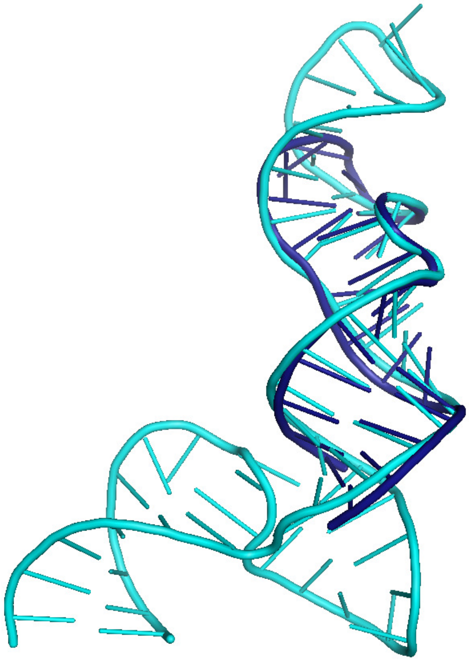
SARA: a method for RNA structural alignment and function annotation



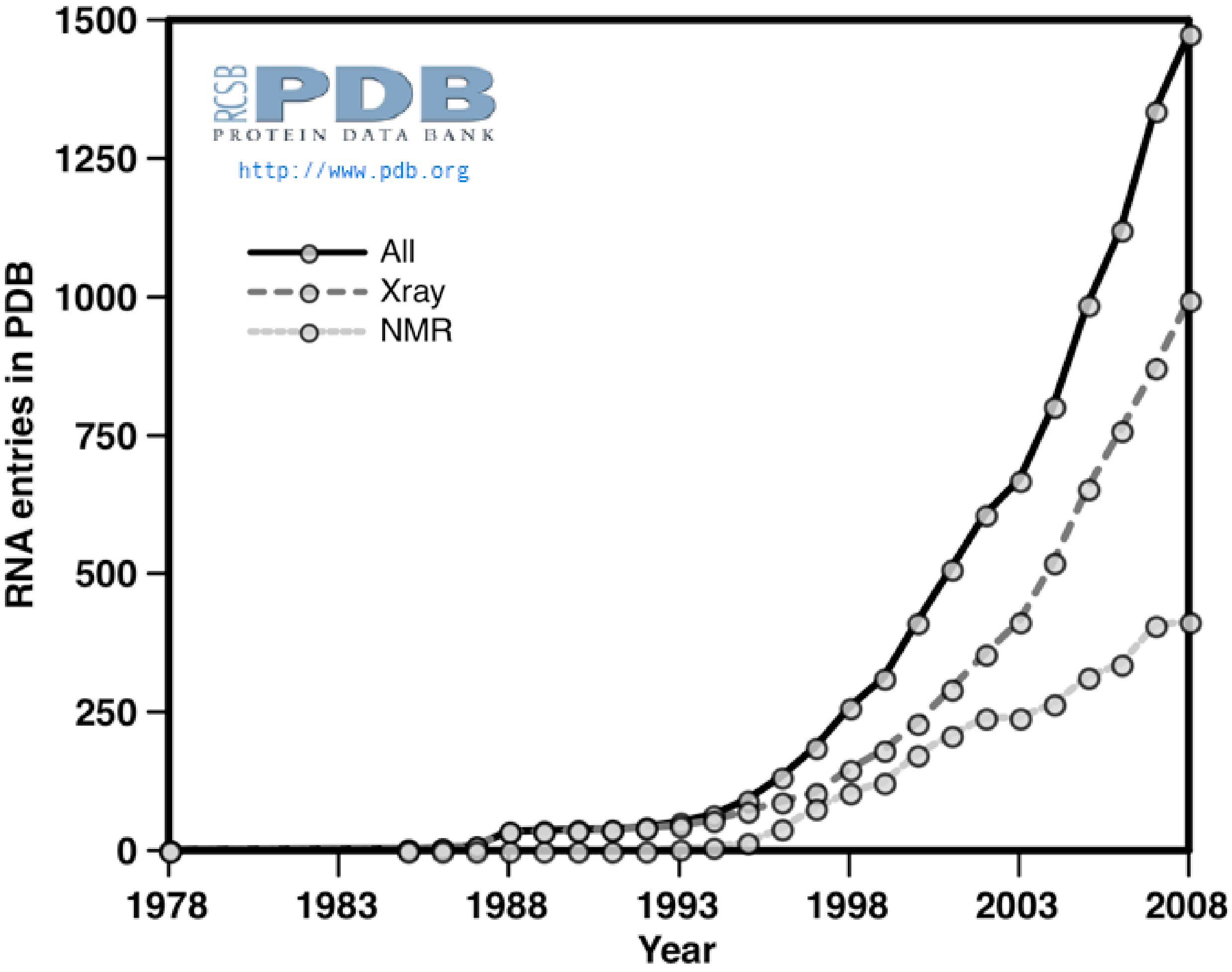
Emidio Capriotti
and Marc Marti-Renom

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Research Center (CIPF), Valencia (Spain)
http://sgu.bioinfo.cipf.es/

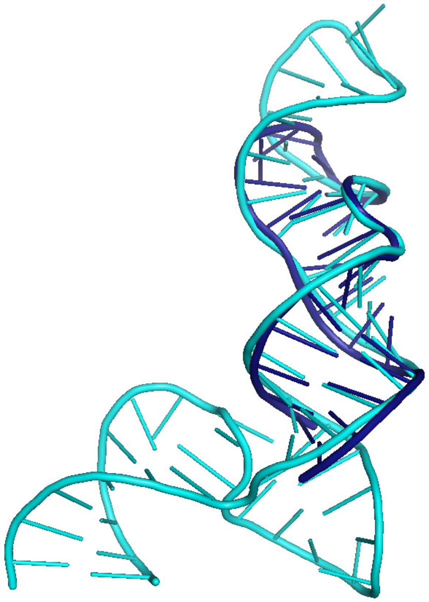
Lipari (ME) Italy 18/06/2009

RNA Structure

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



Structural alignment

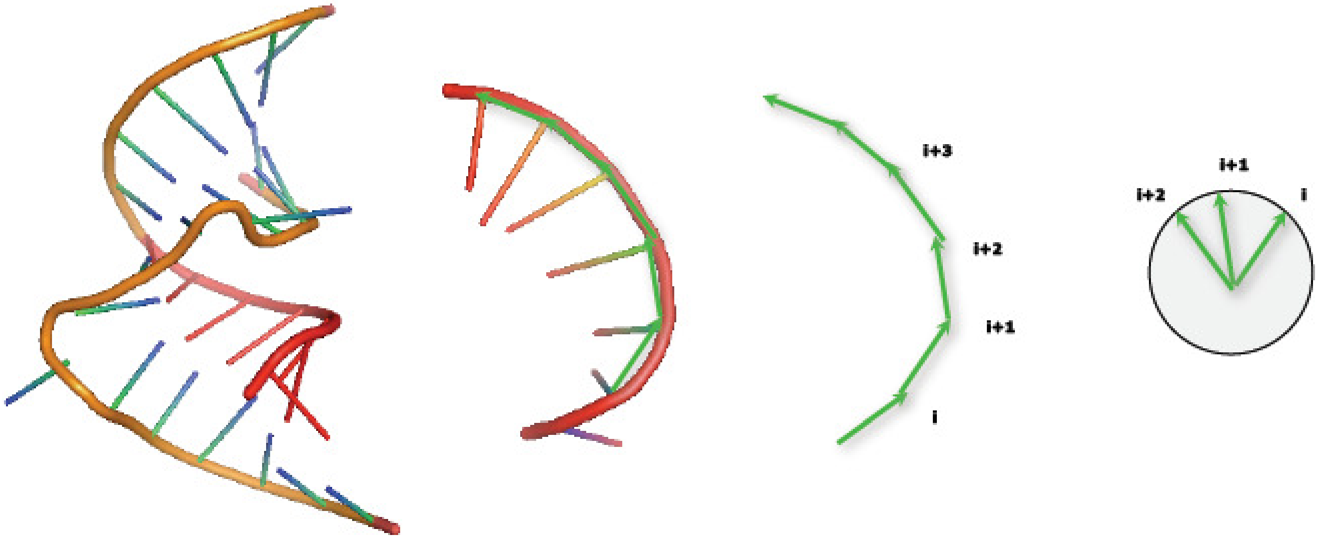


Structural alignment attempts to estabish equivalences between two or more polymer structures based on their shape and their **three-dimensional conformations.**

In contrast to the structural superimposition, where at least some equivalences are known, structural alignments **does not require any a priori knowledge of the equivalents positions**.

Structural alignment has been used as valuable tool for the comparison of proteins including the **inference of evolutionary relationship** between proteins with low level of sequence similatity.

Unit Vector representation

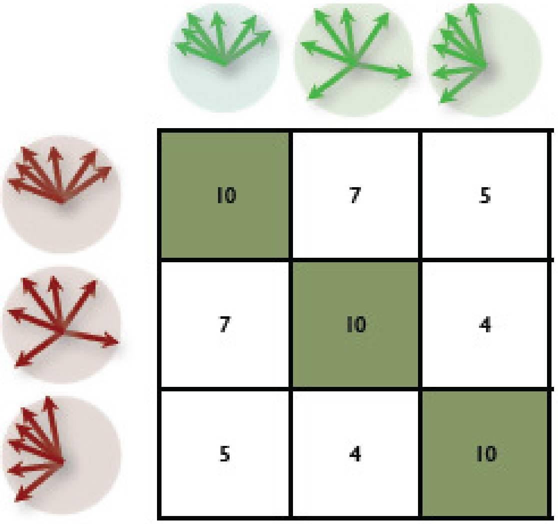


A **Unit Vector** is the **normalized vector between two successive atoms of the same type.**

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

Ortiz et al. Proteins 2002

Unit Vector scoring



10	7	5
7	10	4
5	4	10

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

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

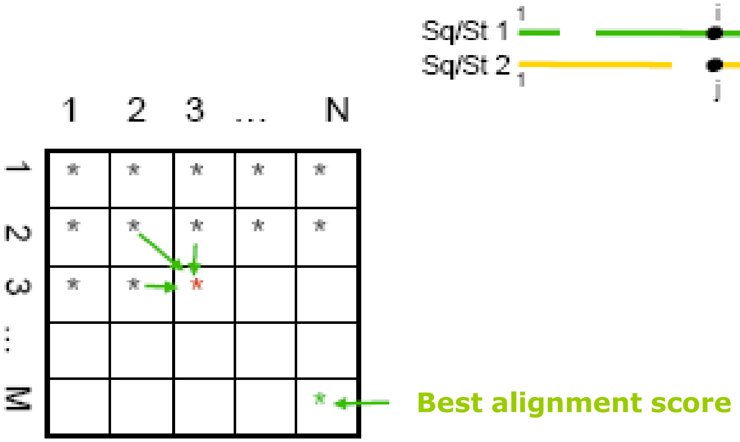
$$\Delta(URMS^R, URMS^{ij}) = 10 \Rightarrow URMS^R > URMS^{ij}$$
$$\Delta(URMS^R, URMS^{ij}) = 0 \Rightarrow URMS^R \le URMS^{ij}$$

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



$$D_{ij}=\max\begin{cases} D_{i-1,j-1} & + s(x_i,y_j)\\ D_{i-1,j} & + g \\ D_{i,j-1} & + g\end{cases}$$

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 by a variant of the MaxSub algorithm.

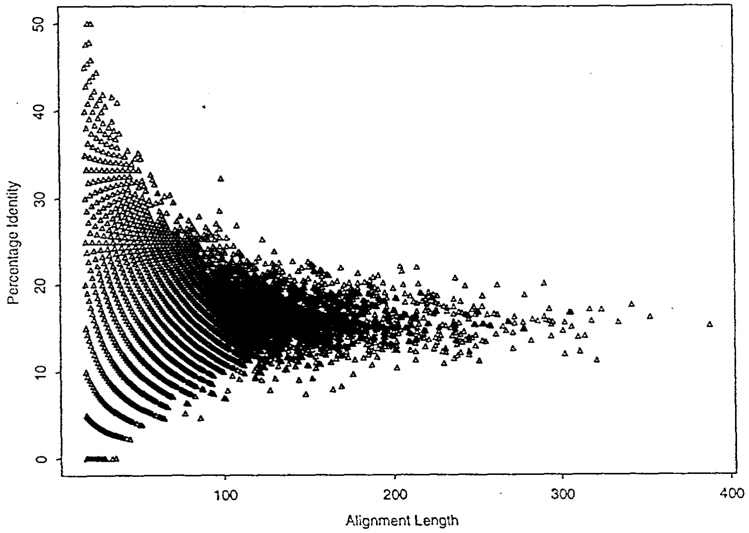
For each alignment the number of close atoms is used to **evaluate the percentage of sequence, secondary structure and tertiary structural identities.**

Needleman and Wunsch J.MB 1970
Siew et al. Bioinformatics 2000

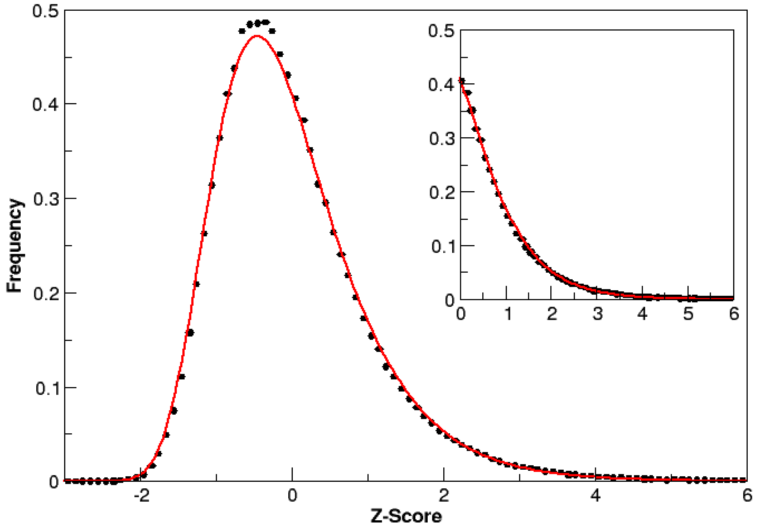
Background distribution

Considering **RNA sequences with identity below 25%** we have produced a **set of pairwise alignments** that has been used **to calculate the empirical background distribution.**
From such distribution we can then evaluate μ and σ needed to calculated the p-value for P(s≥x).

Empiric



Analytic



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

Karlin and Altschul PNAS 1990

RNA function annotation

The proposed algorithm to align RNA structures called **SARA** has been **applied** in the **development** of an automatic **method for RNA function annotation** based on structural alignment.

The adopted strategy consists in the **comparison of a query RNA** structure **with a set o representative RNA structures** with **function annotation reported in the SCOR** database.

For each comparison a score is returned and the **predicted fuction** will be assigned sorting the list pairwise alignments and **considering the fuction of the RNA with highest score.**

The **score** is the **mean logarithm of the negative P-values** obtaining comparing the **percentage of sequence, secondary structure and tertiary stucture identities** of the best alignment with their relative background distributions.

Capriotti and Marti-Renom Bioinformatics 2008
Tamura et al. NAR 2004

Datasets composition

We test our method selecting a set **no identical RNA structures with more than 20 nucleotides, 3 base pairs**.

A **set of RNA structure** has been collected considering only RNAs with **fuction annotation in SCOR** dataset.

R-FSCOR: **representative set of 192 RNA structures** with SCOR annotated functions. Cluster according **to deepest SCOR function and reclustered** considering a **percentage of structural identity threshold of 90%.**

Datasets	Number of chains	Number of alignments	Number of different SCOR functions
RNA08	451	101,475	-
BgALI	451	50,995	-
FSCOR	419	-	168
R-FSCOR	192	-	168
T-FSCOR	227	-	88

Method tests

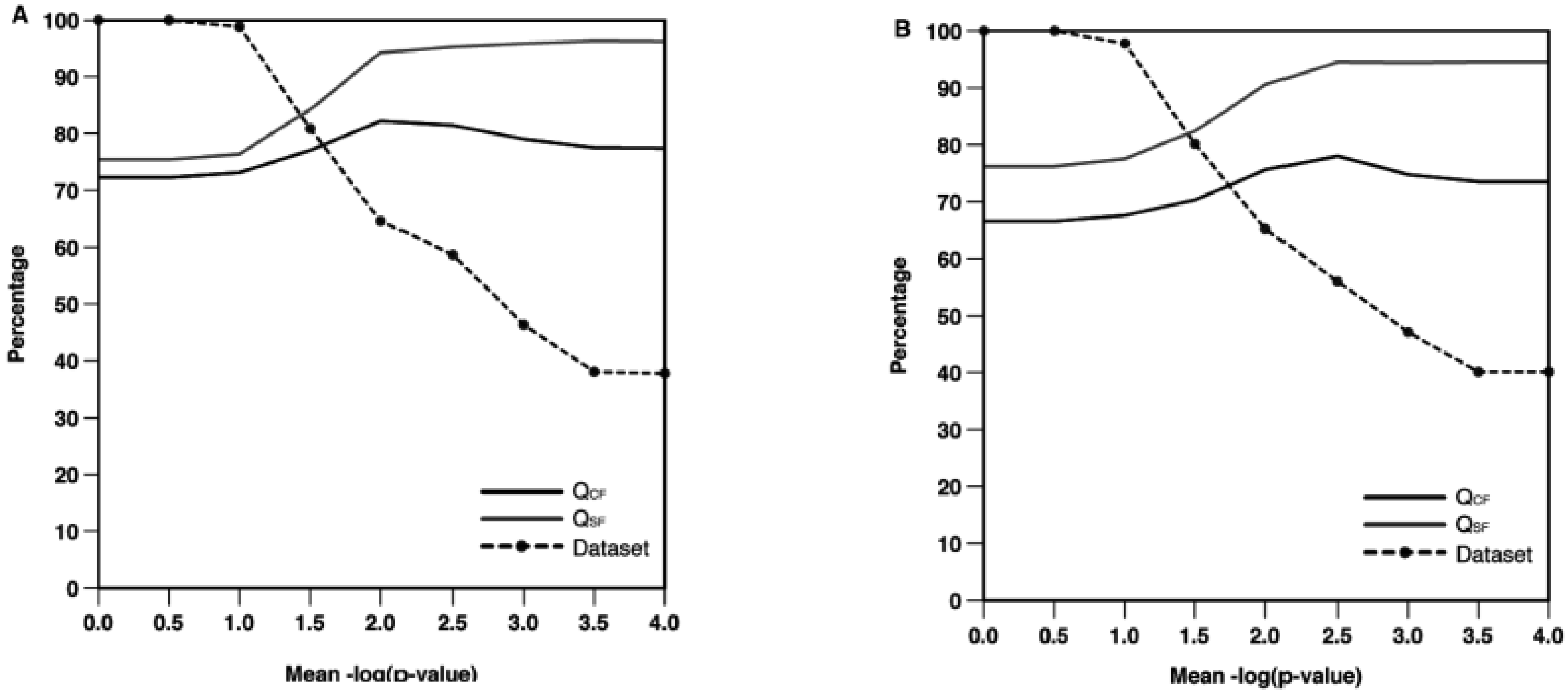
The **SARA method** for functional assignment has been **tested** calculating the **fraction of RNA for which a RNA with the same or similar function** is ranked **in the first position**. RNA with similar function differs only in the deepest SCOR classification term.

First test: Performance resulting from a **leave one out procedure over FSCOR** set.

Second test: Predicton accuracy on the **T-FSCOR** set composed by 227 **RNA structures in FSCOR and not in R-FSCOR.**

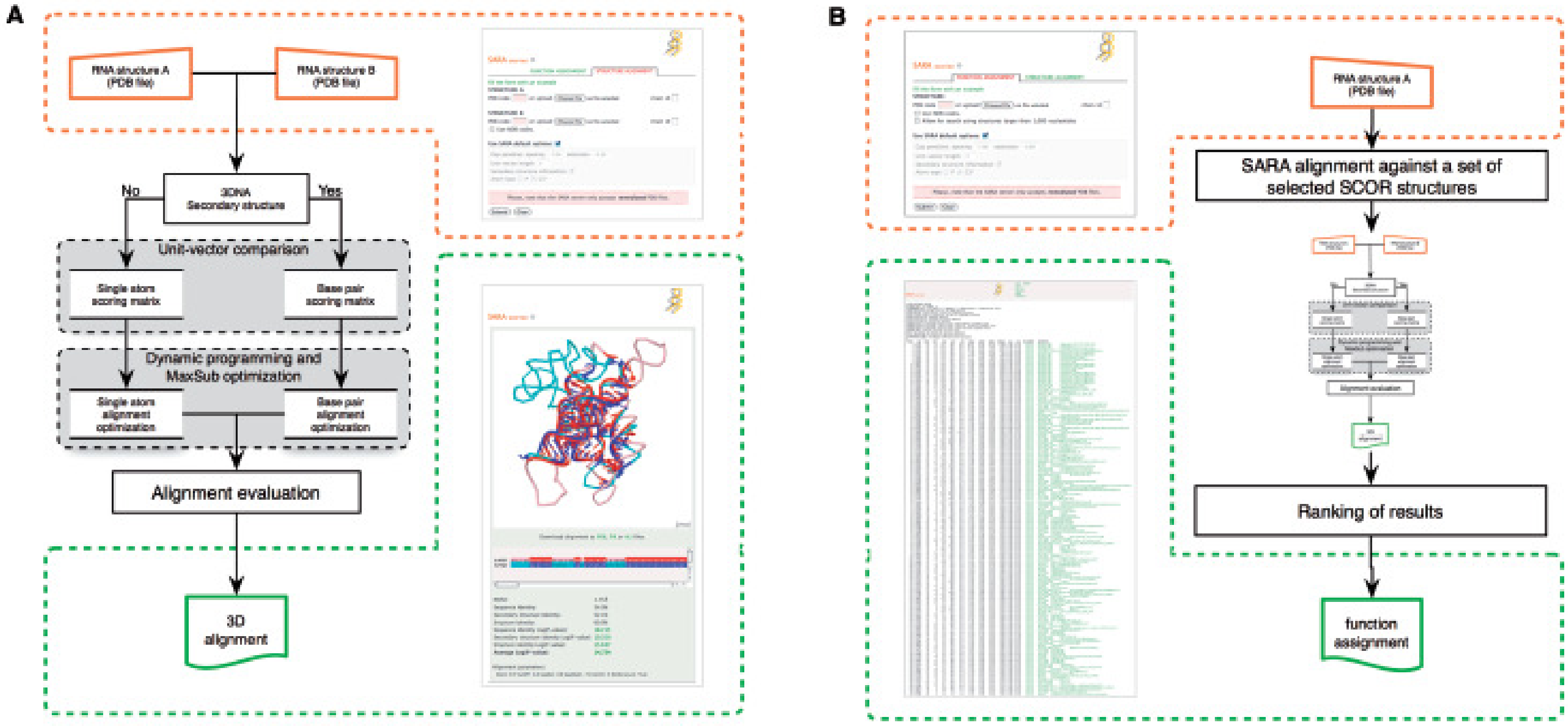
Results

The accuracy of **corrected function (Q_{CF})** and **similar function (Q_{SF})** assignment tasks has been plotted as a function of the mean negative logarithm of the P-values for the best alignment. In **(A)** the plot results from **leave one out on FSCOR** set and **(B) the performances on T-FSCOR** set



SARA server

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


A
RNA structure A (PDB file)
RNA structure B (PDB file)
No
3DNA Secondary structure
Yes
Unit-vector comparison
Single atom scoring matrix
Base pair scoring matrix
Dynamic programming and MaxSub optimization
Single atom alignment optimization
Base pair alignment optimization
Alignment evaluation
3D alignment
B
RNA structure A (PDB file)
SARA alignment against a set of selected SCOR structures
Ranking of results
function assignment

Capriotti and Marti-Renom NAR 2009

Acknowledgments

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

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

Functional Genomics Unit (CIPF)
Joaquin Dopazo
Fátima Al-Shahrour
José Carbonell
Ignacio Medina
David Montaner
Joaquin Tárraga
Ana Conesa
Toni Gabaldón
Eva Alloza
Lucia Conde
Stefan Goetz
Jaime Huerta Cepas
Marina Marcet
Pablo Minguez
Francisco García
Rafael Jiménez
Pablo Escobar

Bologna Biocomputing Group
Rita Casadio
Piero Fariselli
Remo Calabrese
University of Camerino
Mario Compiani

FUNDING
SPINNER Consortium
BIOSAPIENS Network of Excellence
Marco Polo Research Project
Marie Curie Reintegration Grant

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