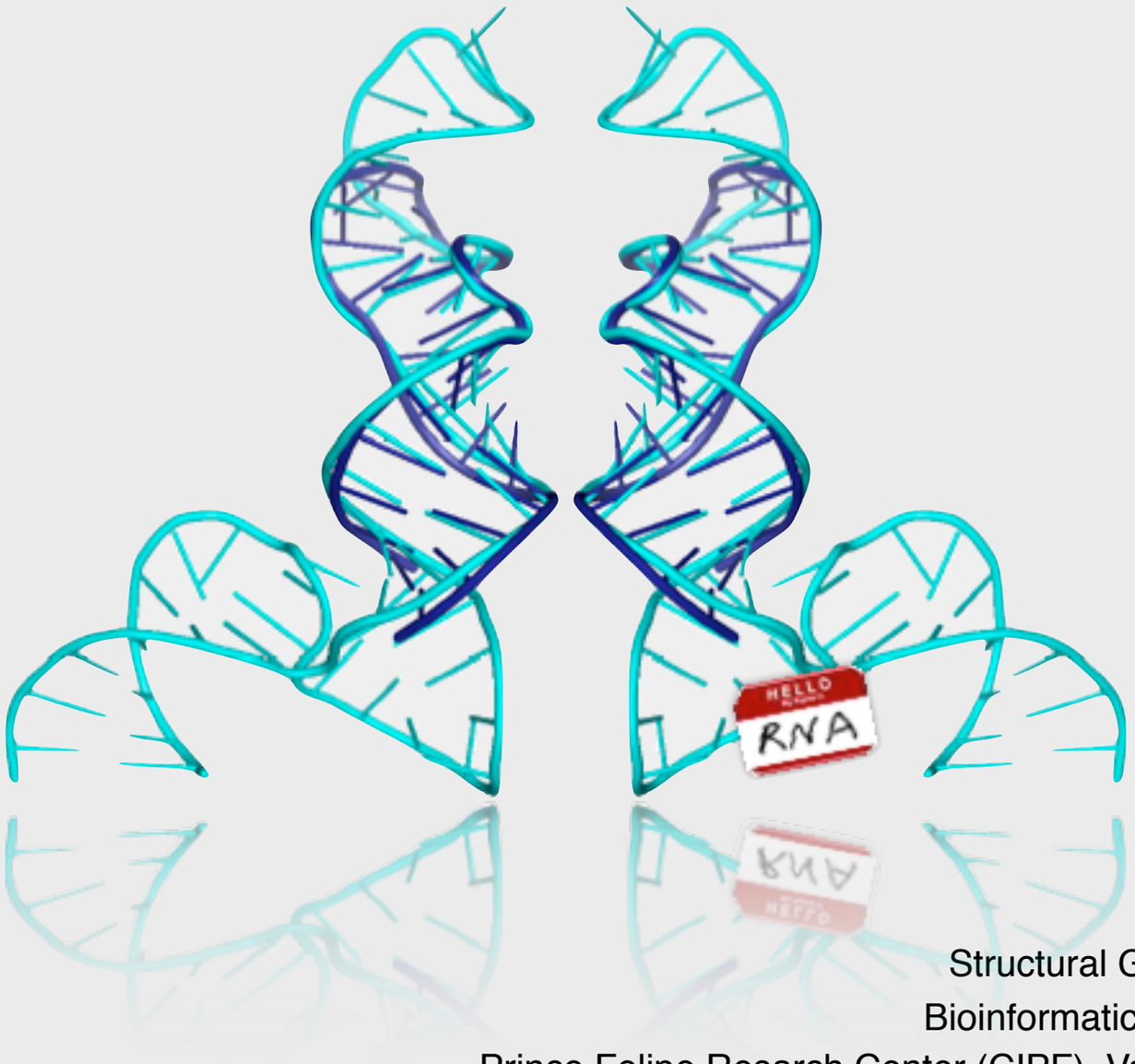
Quantifying the relationship between sequence and three-dimensional structure conservation in RNA



Emidio Capriotti

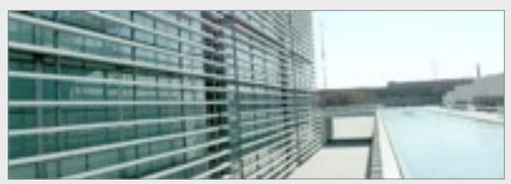
Marc A. Marti-Renom

http://sgu.bioinfo.cipf.es

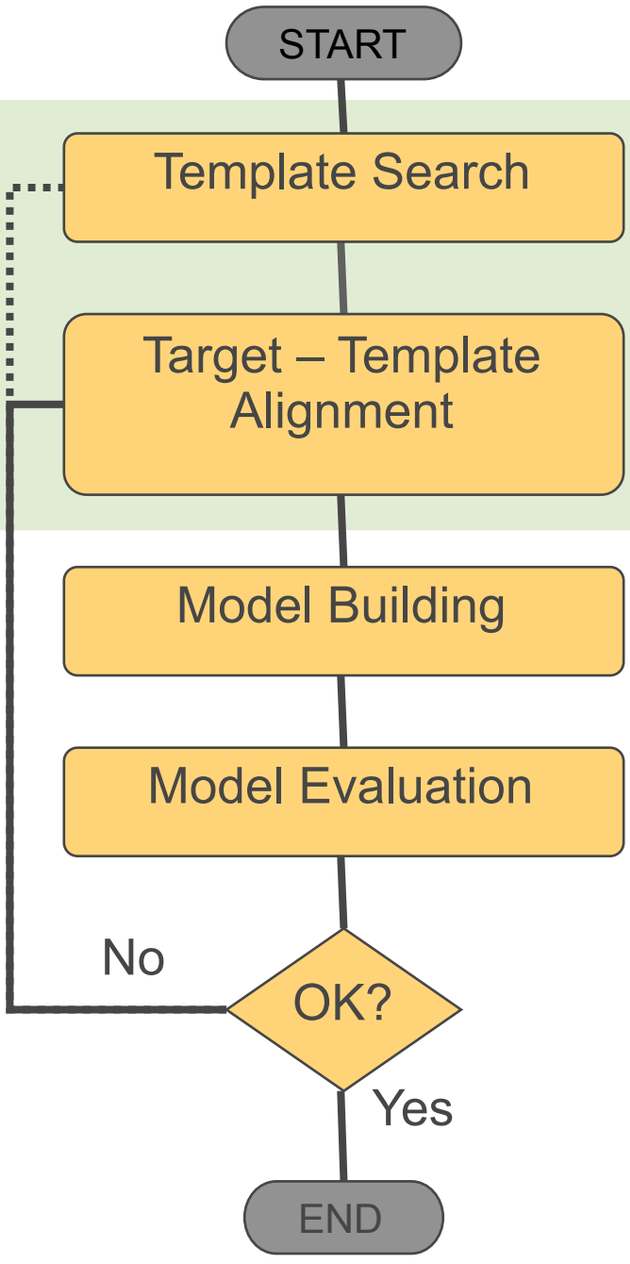
Structural Genomics Unit

Bioinformatics Department

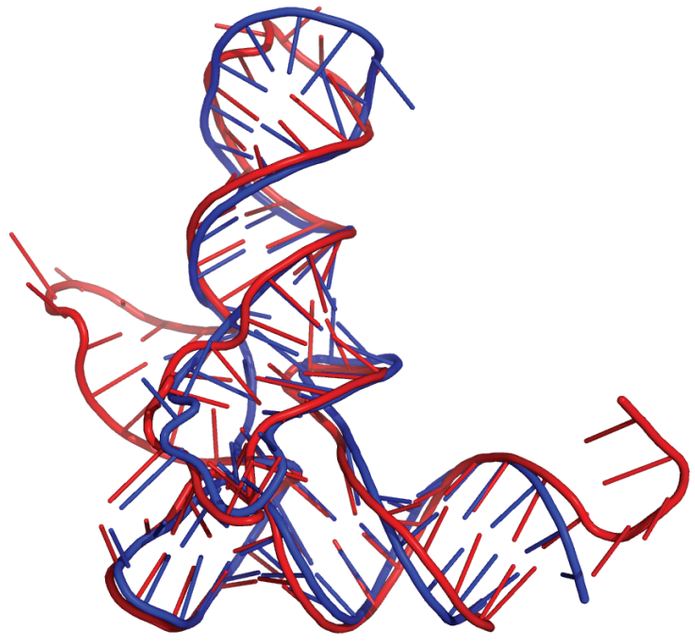
Prince Felipe Resarch Center (CIPF), Valencia, Spain



RNA comparative modeling



RNA structural space

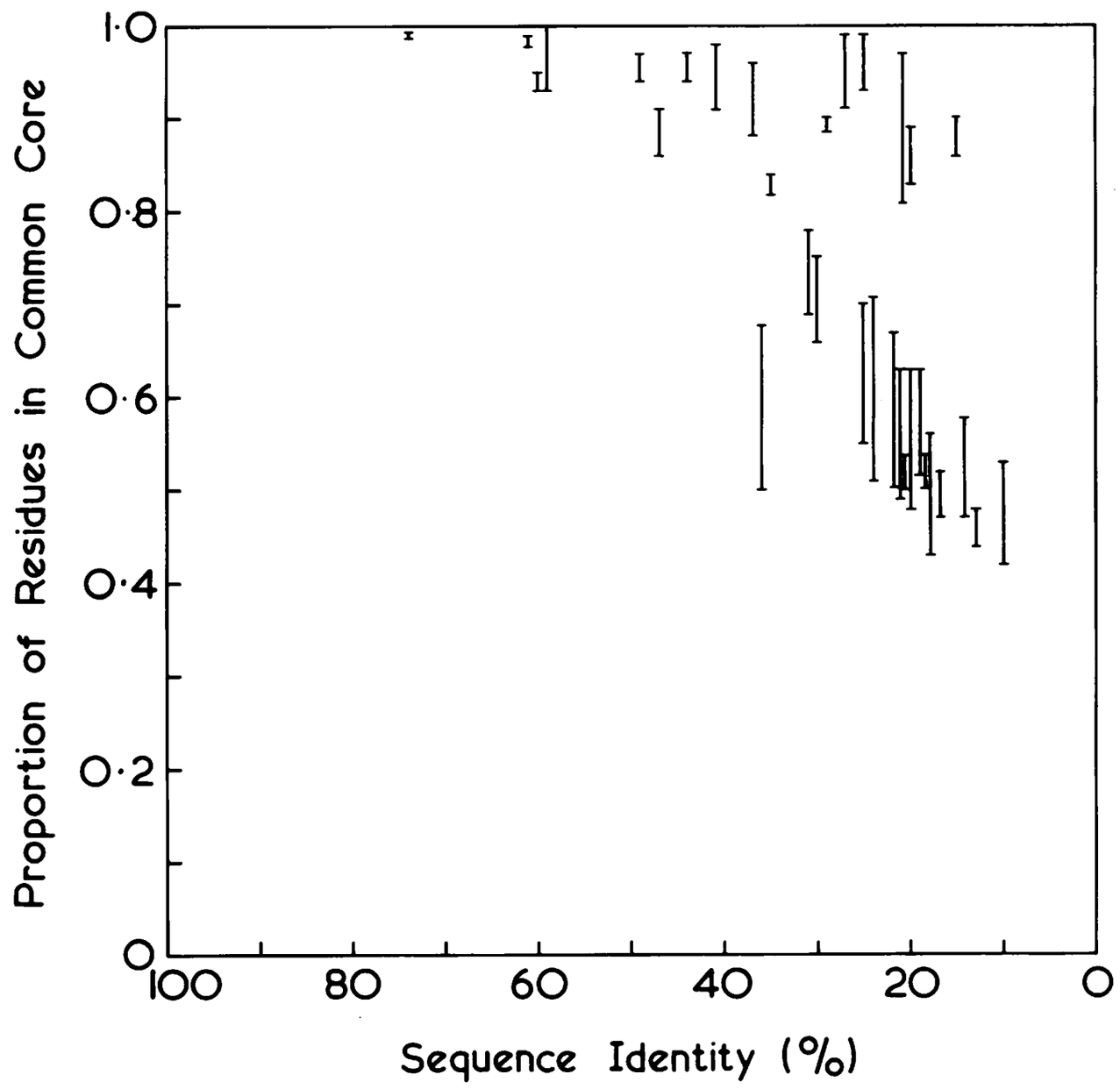


Seq-Str conservation in PROTEINS

The EMBO Journal vol.5 no.4 pp.823–826, 1986

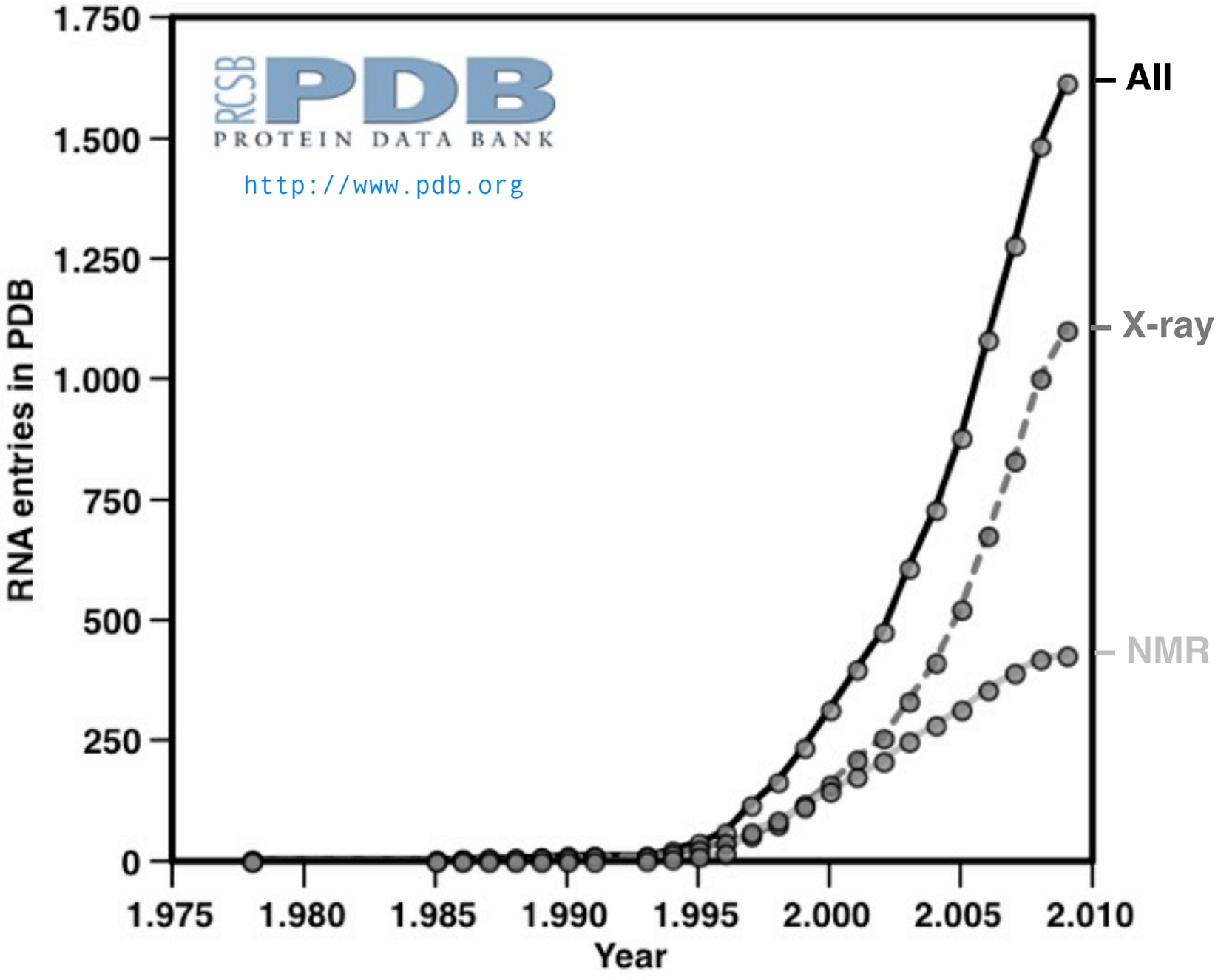
The relation between the divergence of sequence and structure in proteins

Cyrus Chothia[1] and Arthur M.Lesk[2]



RNA structure

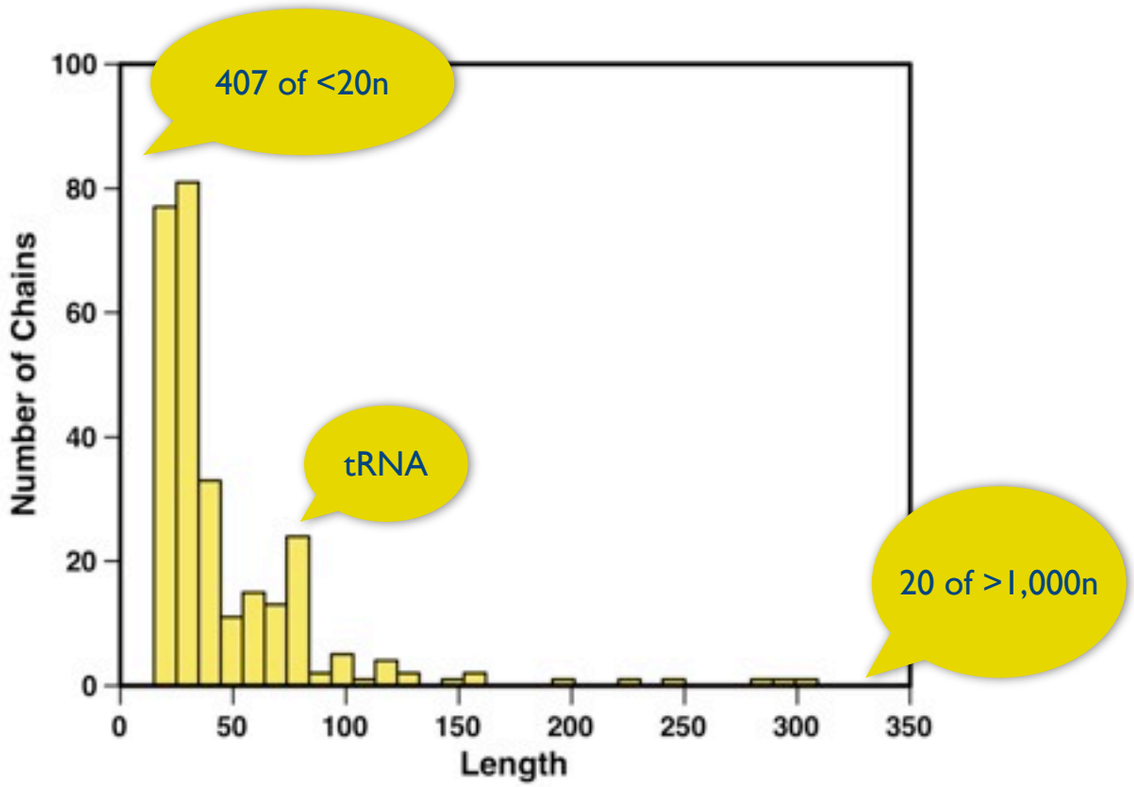
The PDB database contains ~1,600 RNA structures.



RNA structure datasets

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

RNA STRUCTURE*	**1,101**
RNA CHAINS	**2,179**
Non-Redundant RNA CHAINS**	**708**
RNA CHAINS (20≤ Length ≤310) [NR95]	**277**
SCOR SET* [SCOR]**	**60**
HIGH RESOLUTION RNA SET** [HR]**	**51**



* *from PDB November 06.*

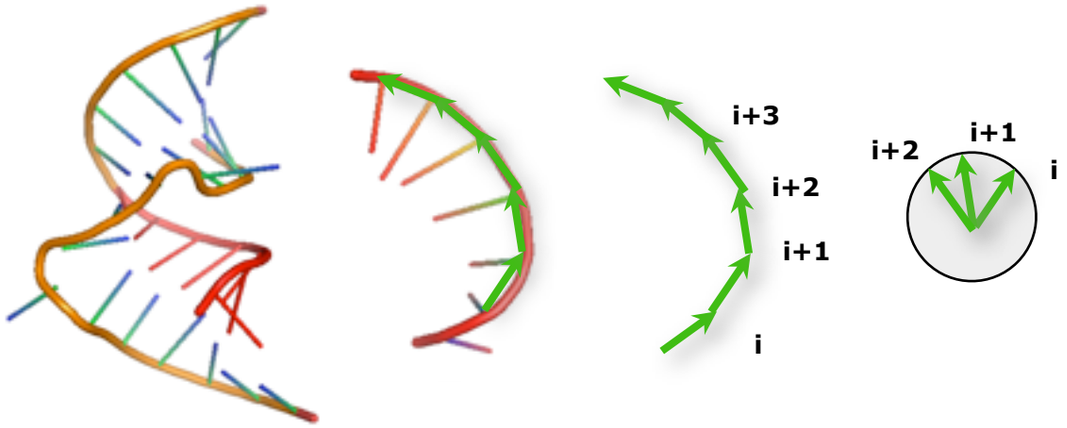
** *non-redundant 95% sequence identity*

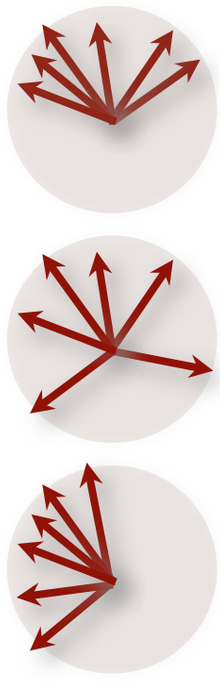
*** *SCOR functions with at least two chains*

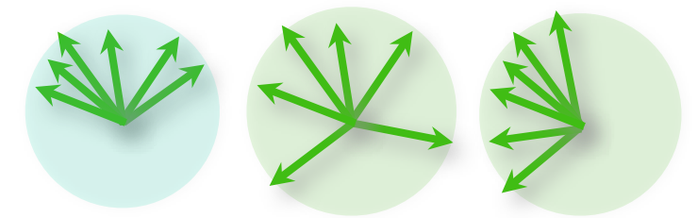
**** *resolution below 4.0 Å and with no missing backbone atoms.*

SARA, a unit vector approach

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





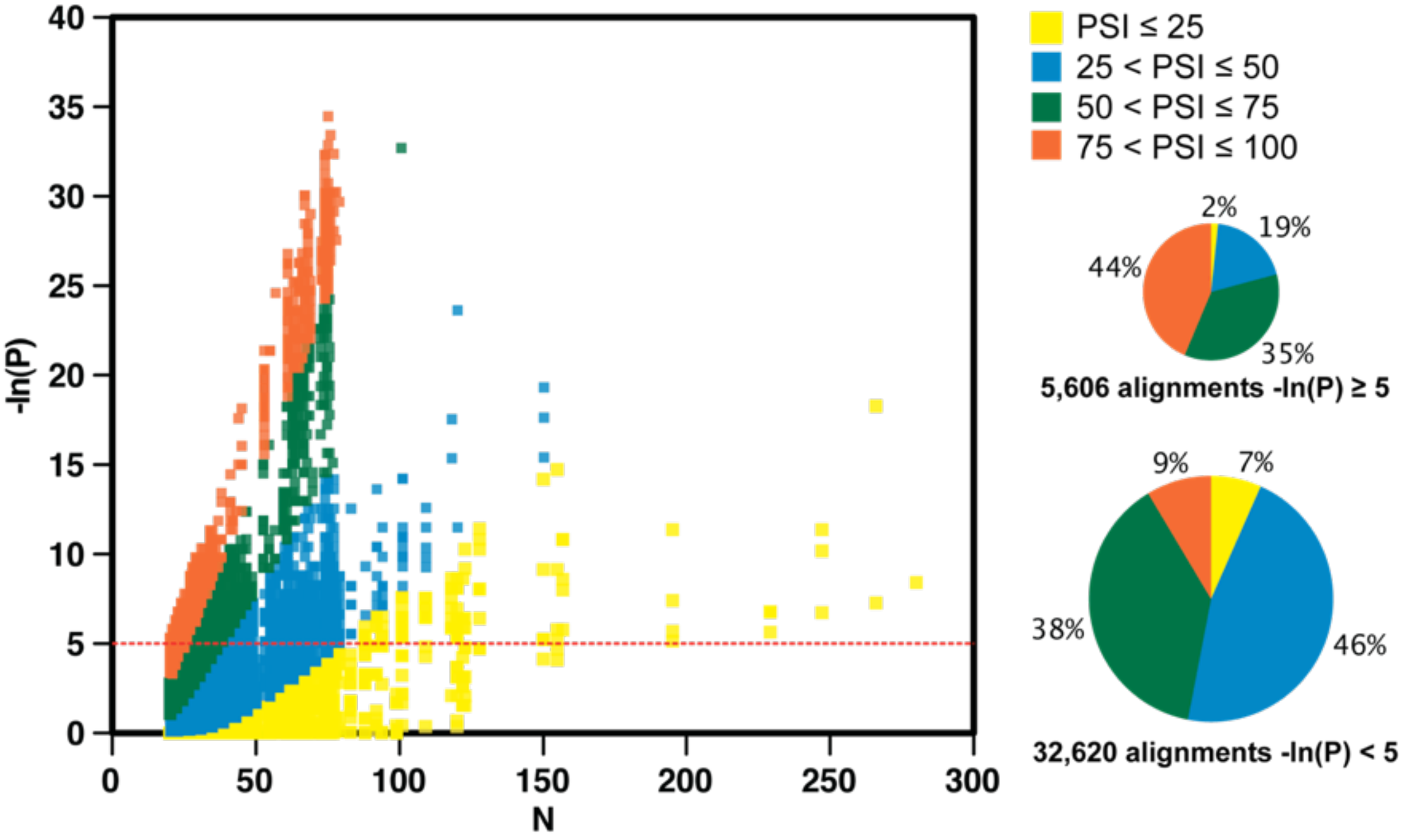


10	7	5
7	10	4
5	4	10

Ortiz et al. *Proteins* 2002

Structural alignments

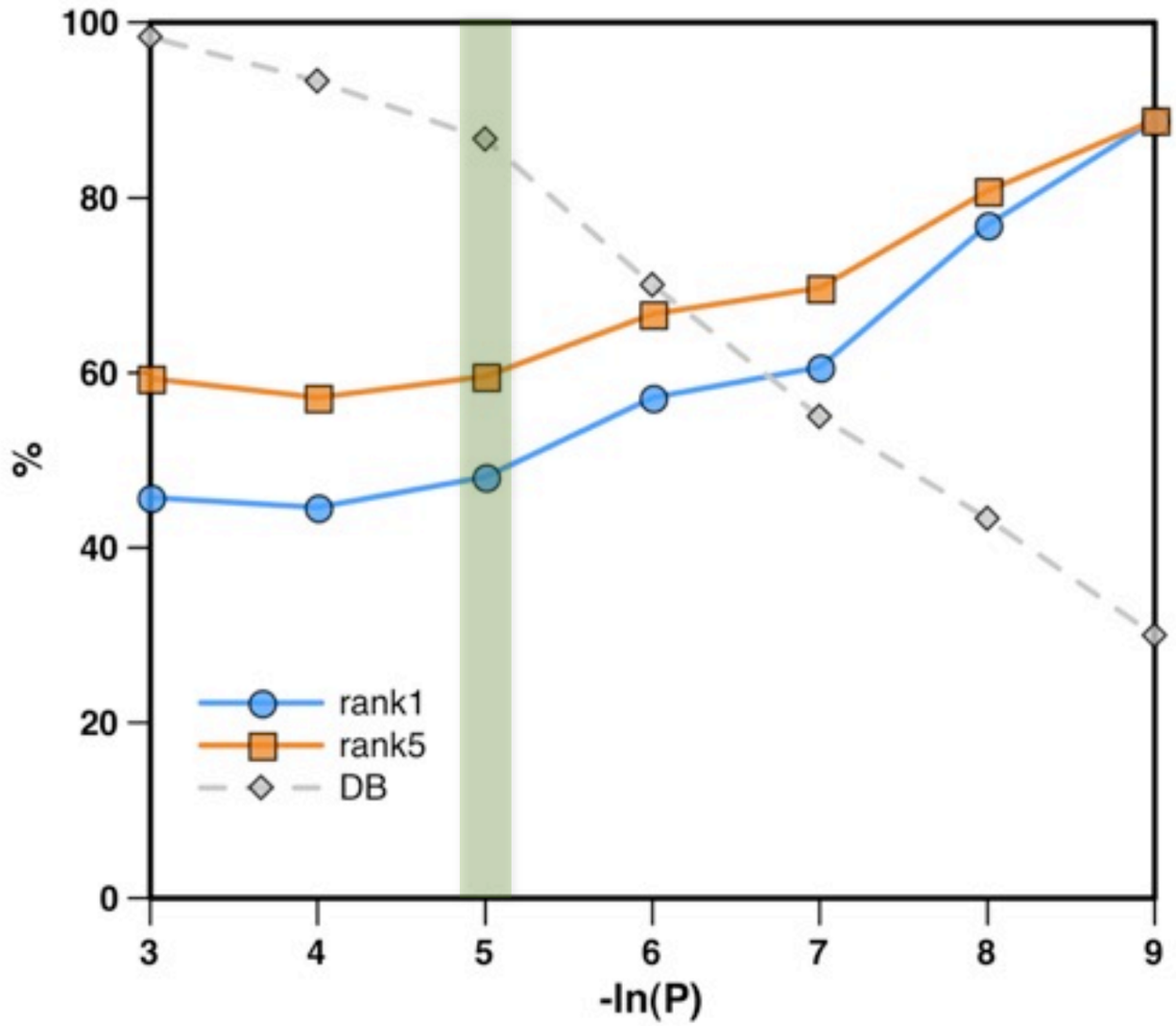
all-against-all comparison of structures in the NR95 set



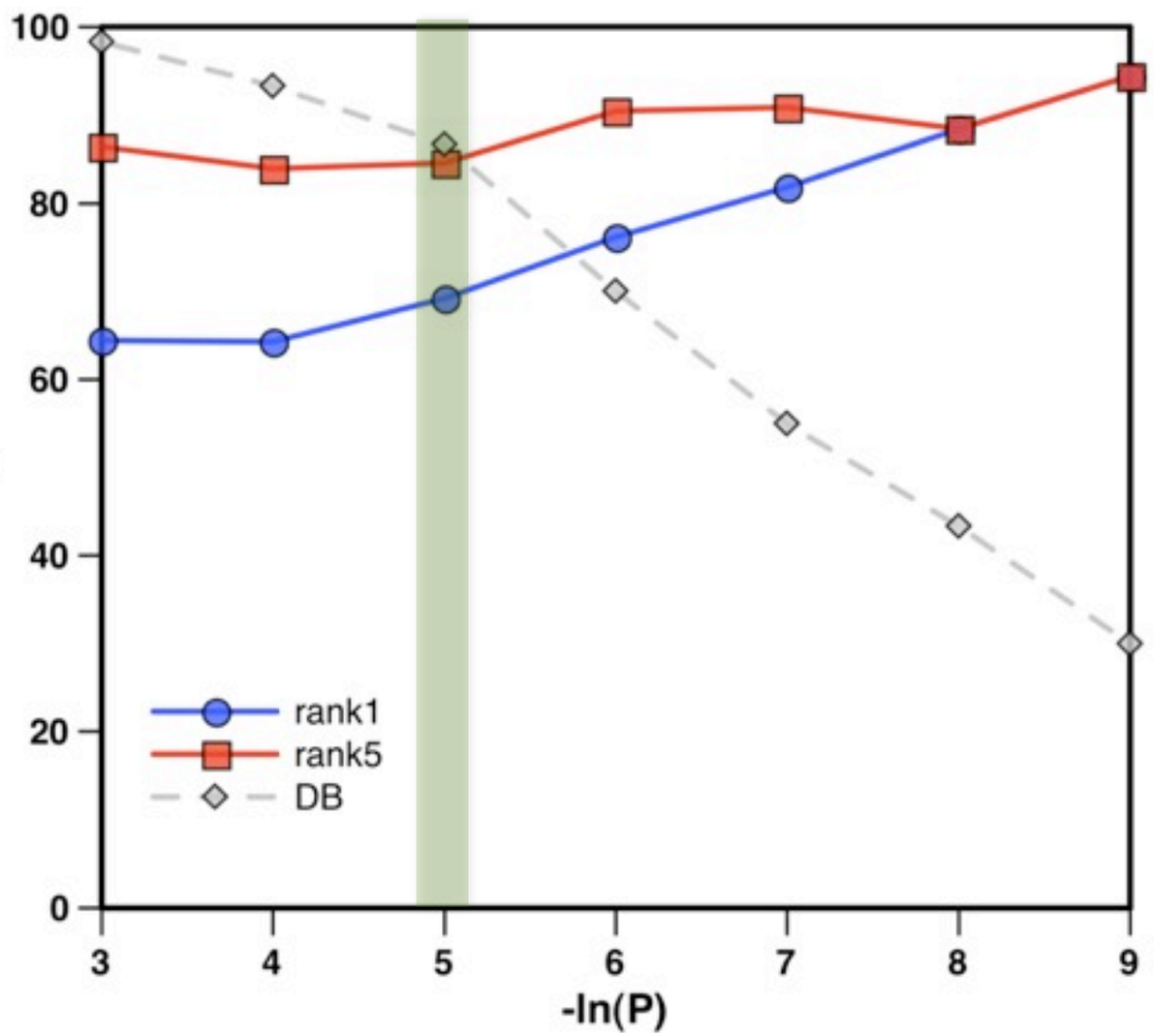
Function assignment

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

Rank of **deepest SCOR function**

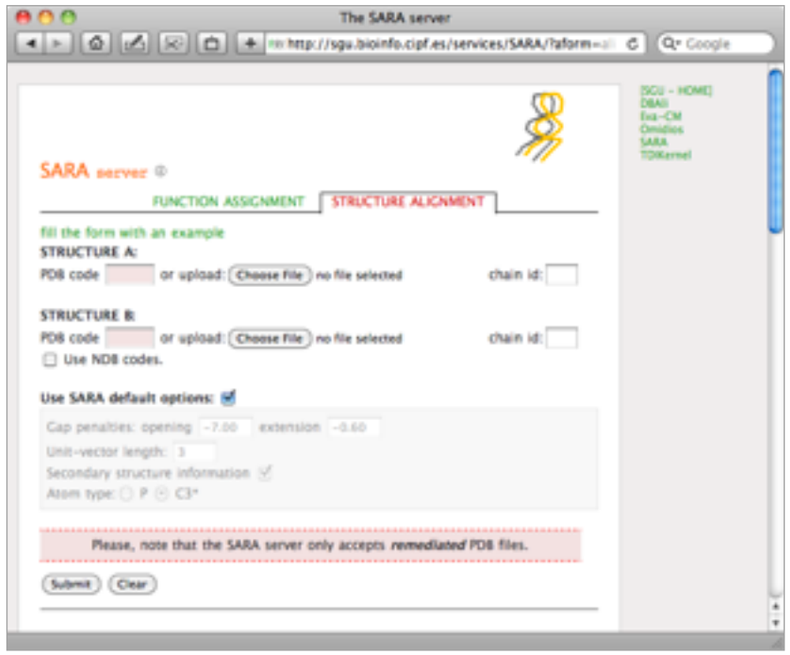


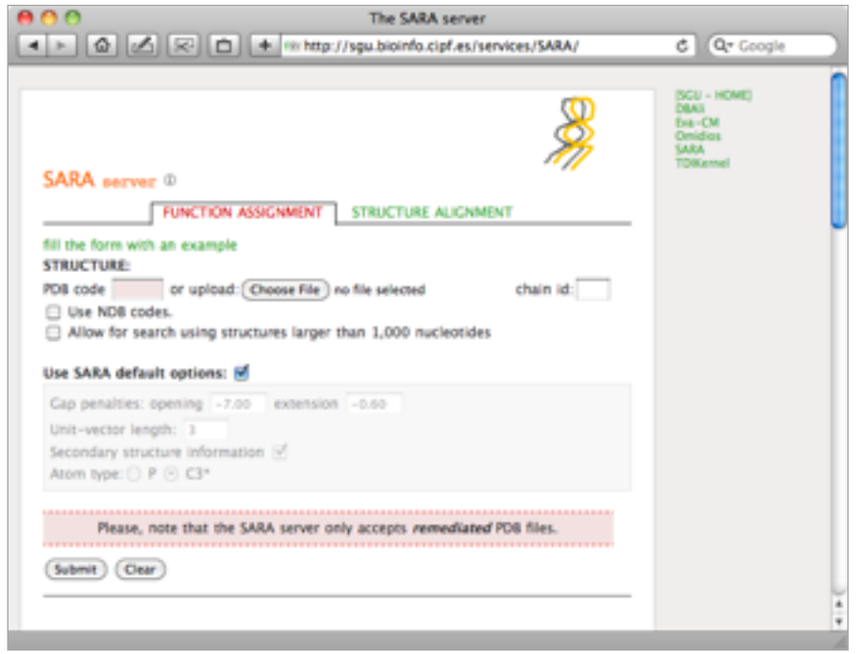
Rank of **related SCOR function**



SARA server

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





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 μRNAs 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 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 η and θ pseudo angles as well as the sugar packer phase. The FR3D programs 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.

*To whom correspondence should be addressed.

Capriotti, E. & Marti-Renom, M.A.
Bioinformatics (2008) **24**:i112-i118

W260–W265 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 29 May 2009*
doi:10.1093/nar/gkp433

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

Capriotti, E. & Marti-Renom, M.A.
Nucleic Acids Research (2009) **37**:W260-5

Seq-Str datasets

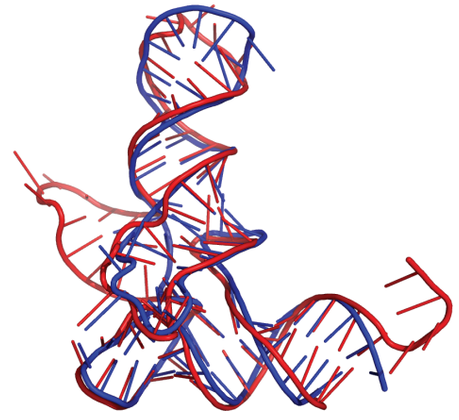
http://sgu.bioinfo.cipf.es/datasets/

Dataset	Number of structures	Number of alignments
NR-RNA09	451	50,995
HA-RNA09	114	589

Staphylococcus phage group I ribozyme (1y0q:A)
Synthetic I Intron fragment (1u6b:B)

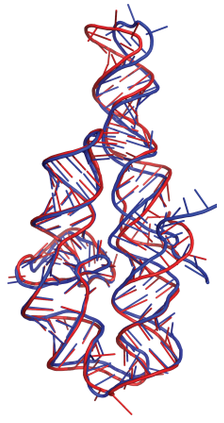
Aligned nucleotides:	120
RMSD:	1.8 Å
Sequence Identity:	34.0 %
Secondary Structure Identity:	52.1 %
Structure Identity:	60.9 %
Sequence -ln(p-value):	18.2
Secondary structure -ln(p-value):	10.3
Structure -ln(p-value):	15.6
Mean -ln(p-value):	**14.7**

Pyrococcus horikoshii tRNA(Leu) (1wz2:C)
Acuifex aeolicus tRNA(Met) (2ct8:C)



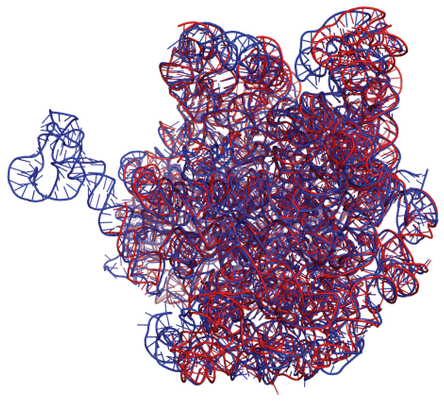
Aligned nucleotides:	65
RMSD:	1.9 Å
Sequence Identity:	56.8 %
Secondary Structure Identity:	88.5 %
Structure Identity:	87.8 %
Sequence -ln(p-value):	10.2
Secondary structure -ln(p-value):	5.2
Structure -ln(p-value):	7.2
Mean -ln(p-value):	**7.5**

Synthetic P4-P6 RNA ribozyme (1l8v:A)
Synthetic P4-P6 RNA ribozyme (2r8s:R)



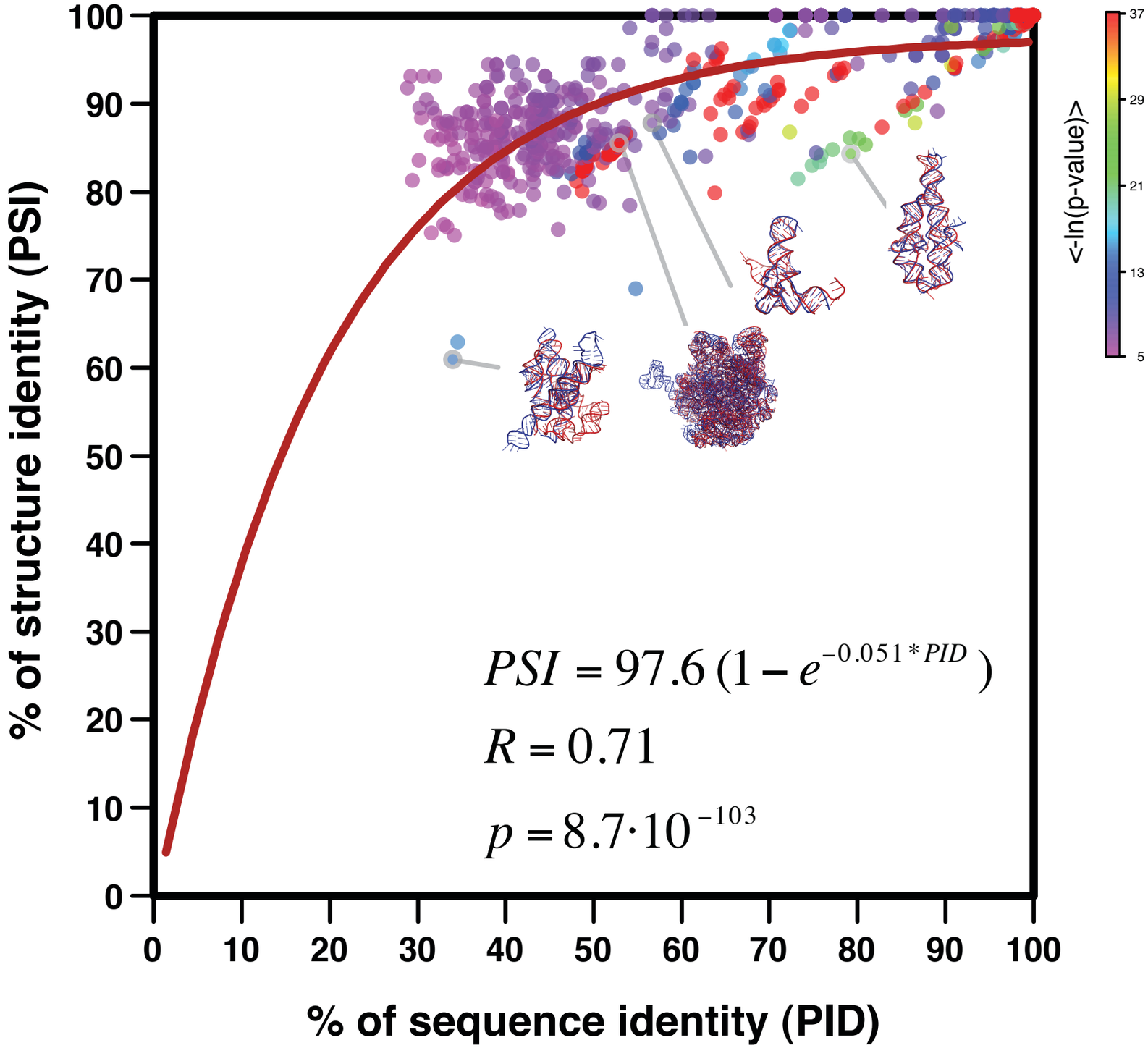
Aligned nucleotides:	134
RMSD:	1.8 Å
Sequence Identity:	80.9 %
Secondary Structure Identity:	81.0 %
Structure Identity:	85.4 %
Sequence -ln(p-value):	37.0
Secondary structure -ln(p-value):	17.1
Structure -ln(p-value):	19.4
Mean -ln(p-value):	**24.5**

Haloarcula marismortui 23S RNA (3cce:0)
Thermus thermophilus 23S RNA (3d5b:A)

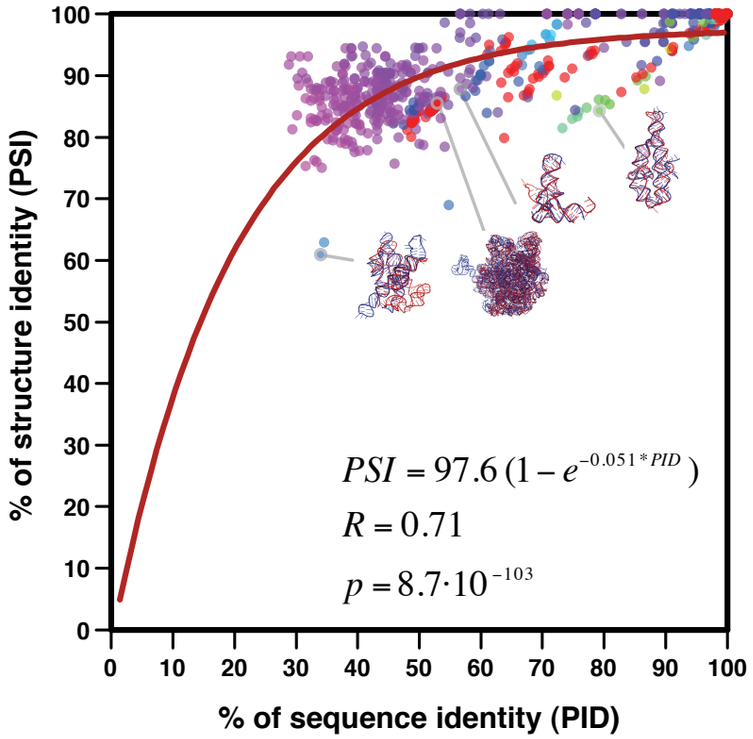


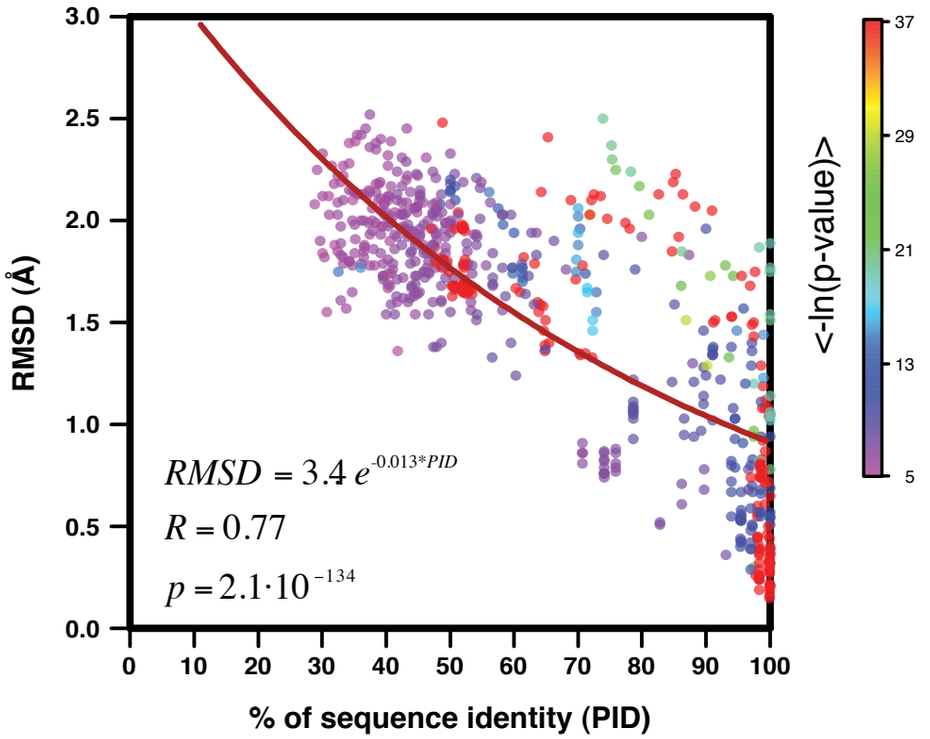
Aligned nucleotides:	2,347
RMSD:	1.7 Å
Sequence Identity:	52.7 %
Secondary Structure Identity:	75.7 %
Structure Identity:	85.2 %
Sequence -ln(p-value):	37.0
Secondary structure -ln(p-value):	37.0
Structure -ln(p-value):	37.0
Mean -ln(p-value):	**37.0**

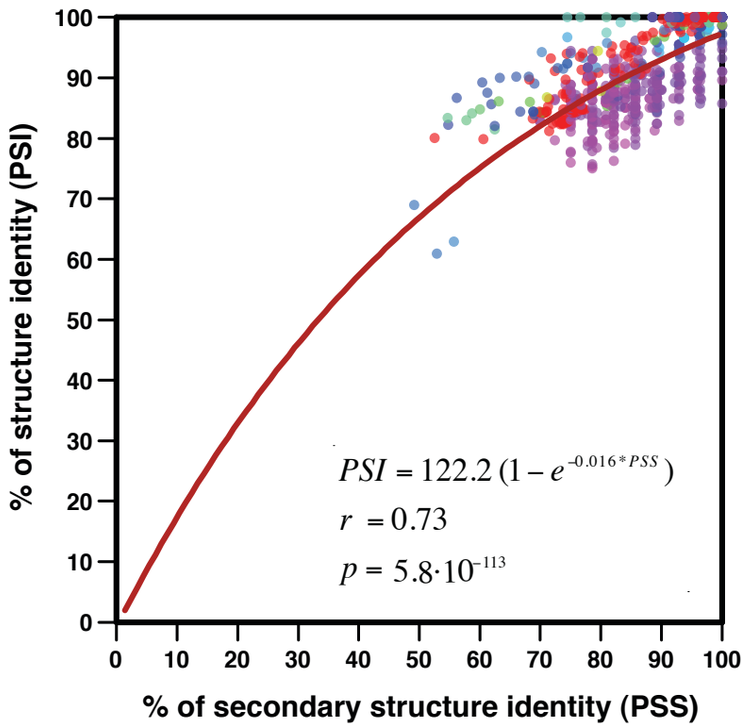
Seq-Str conservation

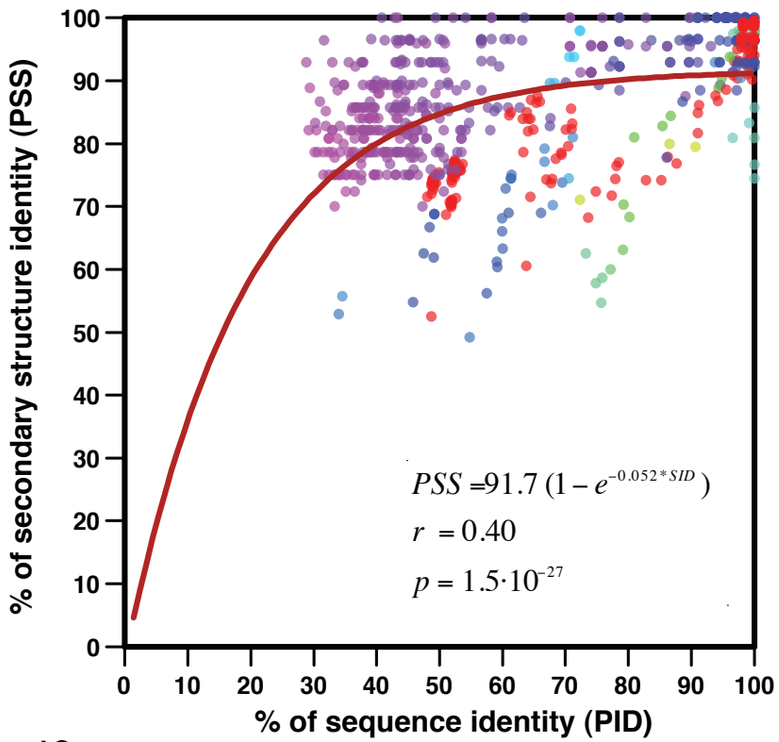


Seq-Str conservation



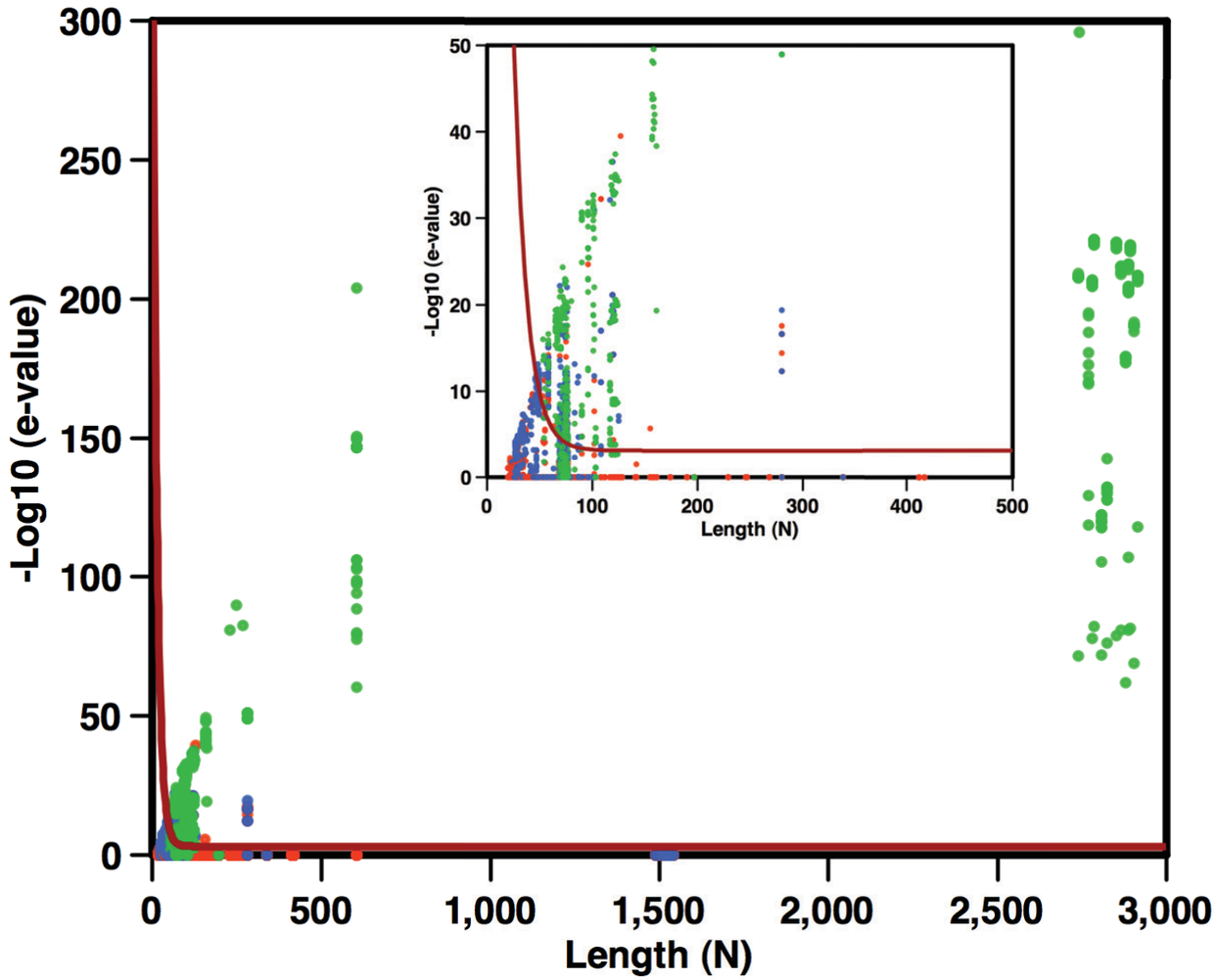






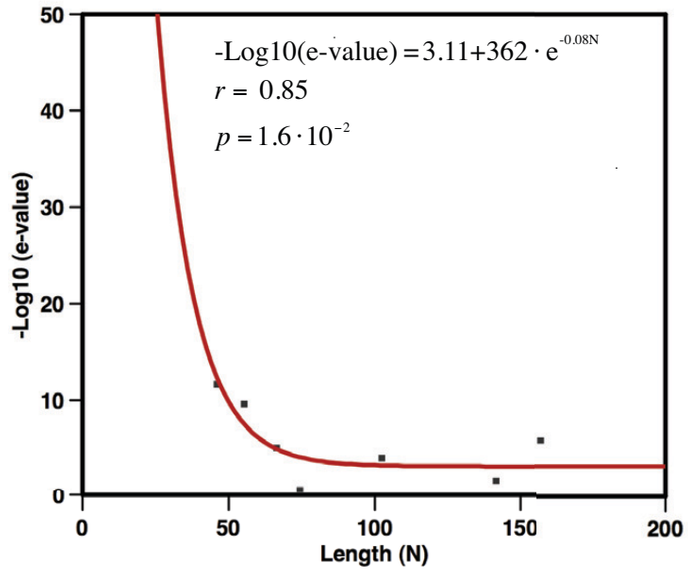
Twilight zone

Alignments by Infernal from Sean Eddy's Lab



- All -log(p-values) ≤ 4.5
- At least one -log(p-value) ≤ 4.5
- All -log(p-values) > 4.5

B

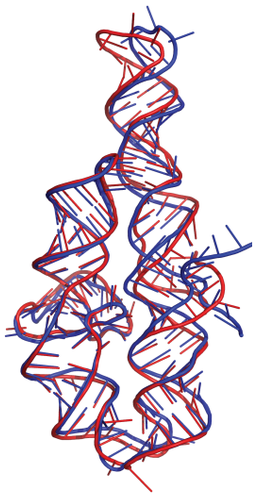


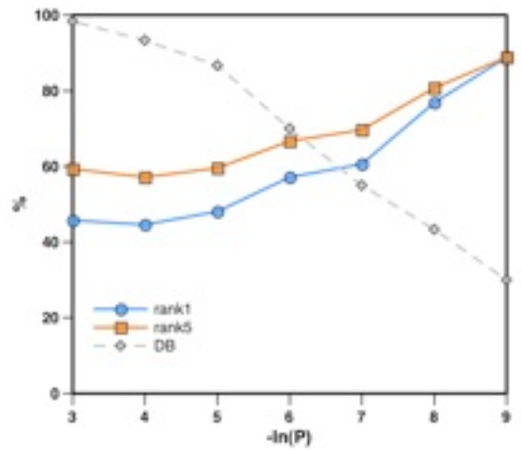
CM would result in accurate models for **1/4 of the RFam** database

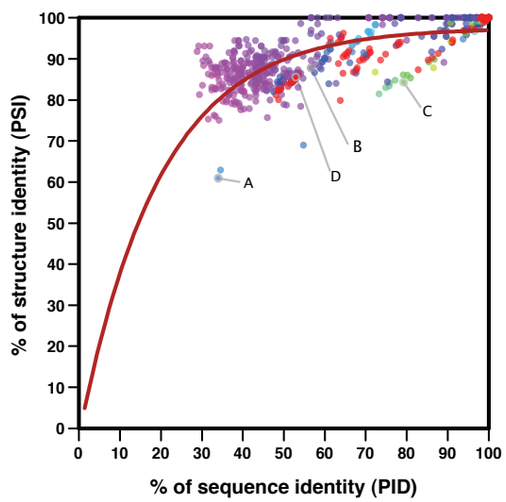
Infernal e-value < 10^{-4} to a known structure

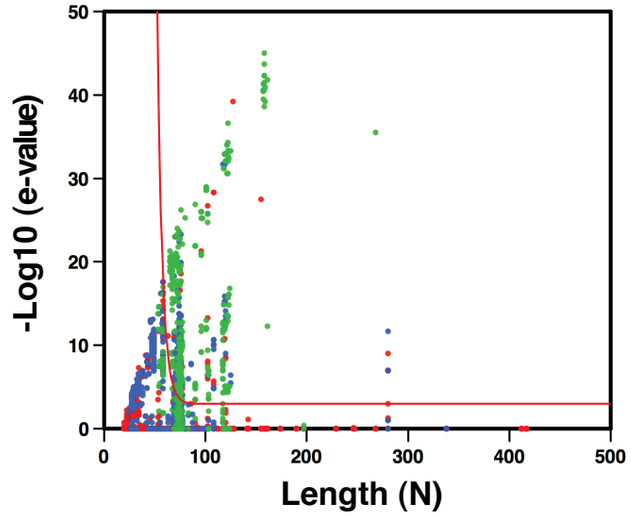
Summary

Capriotti, E. & Marti-Renom, M.A. *BMC Bioinformatics* (2010) **11**:322
Capriotti, E. & Marti-Renom, M.A. *Nucleic Acids Research* (2009) **37**:W260-5
Capriotti, E. & Marti-Renom, M.A.. *Bioinformatics* (2008) **24**:i112-i118









Capriotti and Marti-Renom *BMC Bioinformatics* 2010, **11**:322
http://www.biomedcentral.com/1471-2105/11/322

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Quantifying the relationship between sequence and three-dimensional structure conservation in RNA

Emidio Capriotti[1,2] and Marc A Marti-Renom*[1]

Abstract

Background: In recent years, the number of available RNA structures has rapidly grown reflecting the increased interest on RNA biology. Similarly to the studies carried out two decades ago for proteins, which gave the fundamental grounds for developing comparative protein structure prediction methods, we are now able to quantify the relationship between sequence and structure conservation in RNA.

Results: Here we introduce an all-against-all sequence- and three-dimensional (3D) structure-based comparison of a representative set of RNA structures, which have allowed us to quantitatively confirm that: (i) there is a measurable relationship between sequence and structure conservation that weakens for alignments resulting in below 60% sequence identity, (ii) evolution tends to conserve more RNA structure than sequence, and (iii) there is a twilight zone for RNA homology detection.

Discussion: The computational analysis here presented quantitatively describes the relationship between sequence and structure for RNA molecules and defines a twilight zone region for detecting RNA homology. Our work could represent the theoretical basis and limitations for future developments in comparative RNA 3D structure prediction.

Background

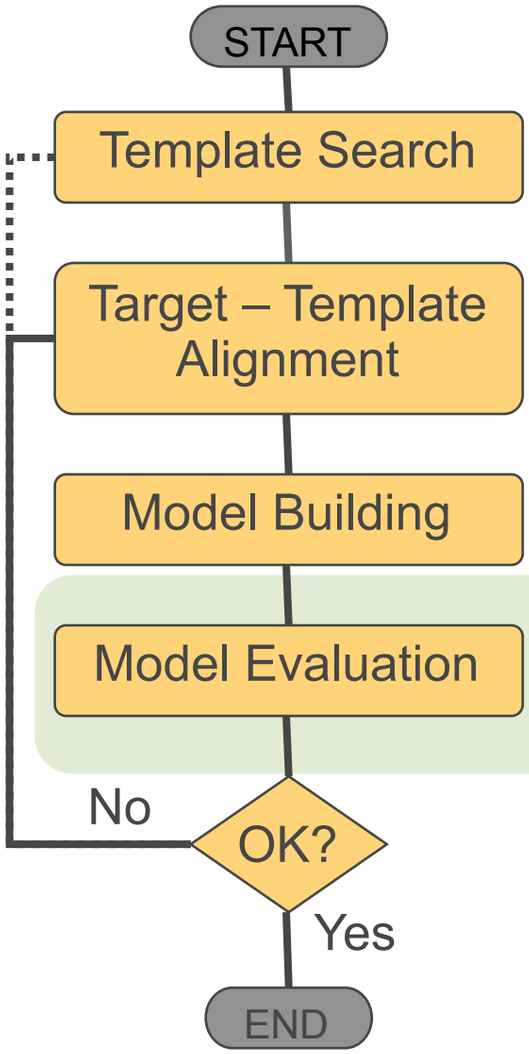
The view of RNA as a simple information transfer molecule has been challenged since the discovery of ribozymes, a class of RNA with enzyme-like functions [1-3]. RNA molecules are now known to carry a large repertory of biological functions such as transfer of information, enzymatic catalysis and regulation of cellular processes [4]. Similar to proteins, functional RNA molecules fold into specific three-dimensional conformations essential for performing their biological activity. Despite advances in characterizing the folding and unfolding of RNA molecules [5-8] and the significant increase of RNA structures deposited in the Protein Data Bank (PDB) [9], our knowledge of the atomic mechanism by which RNA molecules adopt their biological active structures is still limited [10]. Nonetheless, it is common knowledge that RNA 3D structure is more conserved than RNA sequence and that such principle could be used for comparative RNA structure prediction in a similar way it is done for proteins [11]. It was back in the eighties when Chothia and Lesk first quantified such evolutionary relationship for proteins [12-14]. Their seminal works on the relationship between protein sequence and structure conservation provided the theoretical grounds for many computational approaches in comparative protein structure and function prediction [11,15]. Their work concluded that the overall structural changes between two homologous proteins were proportional to their sequence differences. It was then estimated that homologous proteins aligning with less than 20% sequence identity could have large structural differences [14]. Such findings were later confirmed and expanded by several other studies [16-20].

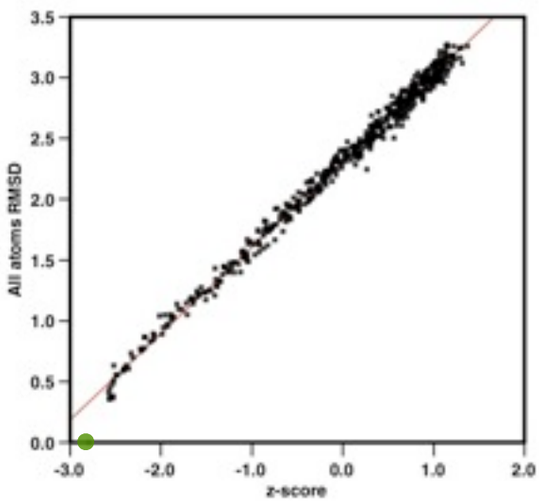
For RNA, the axiom of "function is more conserved than structure and structure is more conserved than sequence" has been adopted since the end of the sixties [21] and even reinforced with the analysis of newly determined large RNA containing complexes such as the ribosome [22-29]. The wealth of new structures has prompted the development of computational methods

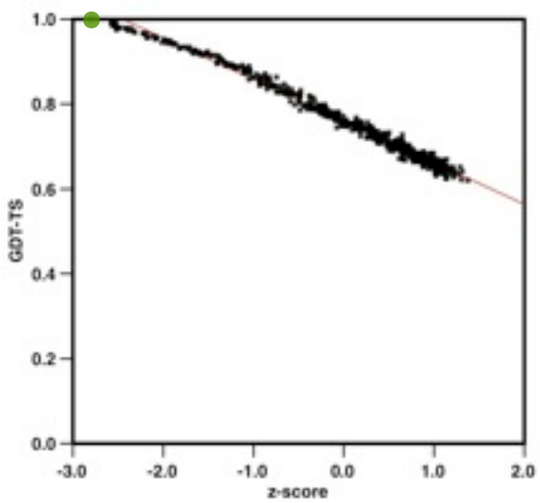
* Correspondence: mmarti@cipf.es
[1] Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain
Full list of author information is available at the end of the article

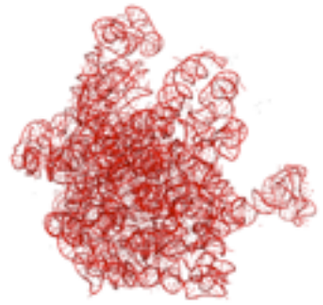
What's next...

Capriotti, E. *et al. Bioinformatics.* **Under revision.**







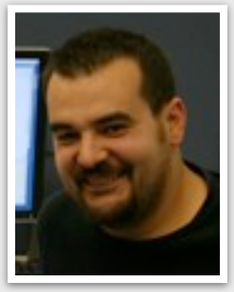
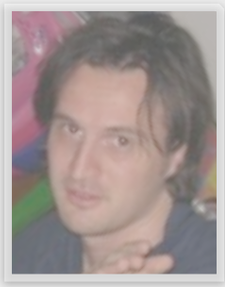


Ribosome 50S subunit

RNA model evaluation

Acknowledgments

http://sgu.bioinfo.cipf.es

COMPARATIVE MODELING
Andrej Sali (UCSF)
M. S. Madhusudhan (A*Star)
Narayanan Eswar (DUPON)
Min-Yi Shen (UCSF)
Ursula Pieper (UCSF)
Ben Webb (UCSF)
Maya Topf (Birbeck College)

MODEL ASSESSMENT
David Eramian (UCSF)
Min-Yi Shen (UCSF)
Damien Devos (EMBL)

FUNCTIONAL ANNOTATION
Andrea Rossi (Rinat-Pfizer)
Fred Davis (Janelia Fram)

FUNDING

Ministerio de Ciencia e Innovación
Marie Curie Reintegration Grant
STREP UE Grant

MODEL ASSESSMENT
Francisco Melo (CU of Chile)
Alejandro Panjkovich (CU of Chile)

NMR
Antonio Pineda-Lucena
Leticia Ortí
Rodrigo J. Carbajo

RNA STRUCTURE ASSESSMENT
Francisco Melo (CU of Chile)
Tomas Norambuena (CU of Chile)

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)