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

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 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-paring 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.

Table 1. Composition of the different sets used in this work

	RNA chains	Alignments
PDBNov06	2179	–
NR95	277	38 226
NR95-HR	51	1 275
NR95-SCOR	60	1 770
OPT	141	300
RAND	300	44 850

We begin this article by describing the benchmark sets used and the algorithm behind the SARA program in Section 2. Then we assess the accuracy of the SARA alignments in Section 3. Finally in Section 4, we discuss its applicability and usage for RNA structure pairwise alignment.

2 METHODS

2.1 RNA structure and alignment sets

A total of 1101 structure files (November 2006), which contained 2179 RNA chains, were downloaded from the PDB database (Table 1). The NR95 set was obtained by removing: (i) sequence redundancy at 95% identity from the initial set of downloaded structures (PDBNov06) using the 'cd_hit' program (Li and Godzik, 2006), (ii) all RNA structures shorter than 20 nucleotides and larger than 320 nucleotides (i.e. 407 and 23 chains, respectively) and (iii) all RNA structures with only P trace atoms. The redundancy filter was applied to reduce the CPU time for calculating all-against-all comparisons within each set. Moreover, very short RNA structures (e.g. <20 nucleotides) invariably result in random pairwise alignments (i.e. *P*-value <3.0) and very large RNA structures significantly increased the needed CPU time for computing our alignment sets. The NR95 set results in 277 chains and 38 226 pairwise alignments. To generate the NR95-HR set, an additional filter was applied to remove all RNA structures with crystal resolution greater than 4.0 Å and with missing backbone atoms, resulting in 51 chains (1275 pairwise alignments). To measure the ability of SARA in detecting functional similarities between two RNA structures, we generated the NR95-SCOR set that contained pairs of structures in the NR95 that were functionally annotated in the same SCOR class (Tamura *et al.*, 2004). The NR95-SCOR set resulted in 60 chains (1770 pairwise alignments) covering 18 SCOR functional classes. Finally, to calculate a background distribution of alignments between two unrelated RNA molecules, 300 RNA structures were generated by connecting randomly selected nucleotide backbone conformations over 42 possible rotamers (Murray *et al.*, 2003). The RAND set contained 300 structures (44 850 pairwise alignments) uniformly distributed over lengths between 20 and 320 nucleotides.

To optimize the independent parameters of SARA, an OPT set was generated by randomly selecting 300 pairwise structure alignments with a *P*-value higher than 5 from those obtained from the all-against-all comparison of the structures in the NR95 set. The 1275 pairwise alignments from the NR95-HR set were used to benchmark the accuracy of the SARA alignments compared to those obtained with the other only available stand-alone program at the time (i.e. the ARTS program). The 1770 pairwise alignments in NR95-SCOR set were used to benchmark the accuracy of SARA for automatically annotated RNA structures. Finally, the 44 850 pairwise alignments from the RAND set were used to calculate the background distribution of random alignments.

The entire sets of RNA structures are available for download at http://sgu.bioinfo.cipf.es/datasets/.

2.2 SCOR database

The structural classification of RNA database contains 579 PDB entries with 5350 internal loops and 2920 hairpin loops (SCOR 2.0.3, October 2004). SCOR provides a classification of RNA structural motifs, function, tertiary interactions as well as their relationships. Structural elements in the SCOR database are organized in directed acyclic graph architectures, allowing multiple parent classes for a motif. The SCOR database was used to generate the NR95-SCOR set with annotated classification mentioned earlier.

2.3 Alignment evaluation

The choice of metric to evaluate a structure alignment is difficult because two different measures are needed (i.e. accuracy and coverage). In this work, we have used two scores that calculate the percentage of superposed nucleotides or base pairs within a given distance cut-off. Thus, such scores quantify at the same time both accuracy and coverage of a given alignment. First, the percentage structural identity (PSI) is

$$\mathrm{PSI} = 100\frac{n_{\mathrm{al}}}{N} \tag{1}$$

where n_{al} is the number of aligned nucleotides within a threshold of 4.0 Å and N is the length of the shorter of the two RNA structures. Second, the percentage of aligned secondary structure (PSS) is

$$\mathrm{PSS} = 100\frac{p_{\mathrm{al}}}{NP} \tag{2}$$

where p_{al} is the number of aligned base pairs within a threshold of 4.0 Å and NP is the smallest number of base pairs of the two aligned RNA structures.

2.4 Algorithm overview

SARA implements a unit-vector representation of RNA structures that calculates a set of vectors between consecutive atoms of a user-selected type (Kedem *et al.*, 1999). A similar approach has been previously used for protein pairwise structure alignment by the MAMMOTH program (Ortiz *et al.*, 2002). Such a simplified representation is a key for finding structurally equivalent atoms between two rigid body structures within seconds of CPU time. SARA calculates an alignment by the following procedure: (i) for each input RNA structure, its atom trace is calculated by selecting all contiguous atoms of a user-defined type; (ii) the resulting atom trace is used to calculate all unit-vectors between consecutive atoms; (iii) a set of k unit-vectors are mapped into a unit-sphere for each nucleotide, where k is a user-defined parameter; (iv) an all-against-all score matrix is calculated with the unit-vector root mean square (URMS) distances between all pairs of unit-spheres from each structure (Chew *et al.*, 1999); (v) a dynamic programming procedure (Needleman and Wunsch, 1970) using zero end gap penalties is applied to the scoring matrix to identify the global alignment between the two structures; (vi) a variant of the MaxSub algorithm (Ortiz *et al.*, 2002; Siew *et al.*, 2000) is used to maximize the number of atoms within 3.5 Å distance between the two structures and (vii) a *P*-value and its minus logarithm are calculated to assess the statistical significance of the resulting alignment score.

Differently from protein structure alignment, where maximizing the number of aligned residues or minimizing the RMSD are usually the main goals, for RNA structures it is also important that base pairs are correctly aligned. Therefore, steps (i) and (vi) of the SARA algorithm also include the possibility of using secondary structure information calculated by the 3DNA program (Lu and Olson, 2003). If such option is selected by the user, SARA will calculate an atom trace using only contiguous atoms involved in base-pairing (step i) and will maximize the number of base-pair atoms within 3.5 Å distance cut-off (step vi).

2.5 Alignment score and significance

The score corresponding to the match of two unit-spheres of k unit-vectors is calculated by their URMS (i.e. $\mathrm{URMS}^{i,j}$ for pairs of sets i and j). Such score

($S_{i,j}$) is obtained by finding the rotation of the two unit-spheres that minimizes the distances between the two sets of k unit-vectors:

$$S_{i,j} = \frac{(\text{URMS}^R - \text{URMS}^{i,j})}{\text{URMS}^R} f(\text{URMS}^{i,j}, \text{URMS}^R) \quad (3)$$

where $f(\text{URMS}^{i,j}, \text{URMS}^R)$ is equal to 10 for $\text{URMS}^{i,j} > \text{URMS}^R$ otherwise 0 and URMS^R is the minimum distance between two random sets (Chew *et al.*, 1999):

$$\text{URMS}^R = \sqrt{2.0 - \frac{2.84}{\sqrt{k}}} \quad (4)$$

where k is the number of unit-vectors in a unit-sphere. The final score of the pairwise alignment is the sum of individual scores ($S_{i,j}$) for the optimal path between two compared structures minus the affine gap penalties after dynamic programming (Needleman and Wunsch, 1970). The optimal gap initiation and extension penalties for SARA with and without secondary structure information were identified by maximizing the number of aligned nucleotides for the training set of RNA pairs. The grid-like search for optimal parameters explored all combinations for initiation gap penalty from −9 to 0 in steps of 1, extension gap penalty from −0.8 to 0 in steps of 0.2, and number of unit-vectors in a unit-sphere from 3 to 7.

After an alignment is produced, SARA calculates its PSI as well as P-value for estimating the probability of obtaining an equal- or better-scored alignment by chance. The distribution of the PSI scores for a set of random RNA structures follows an extreme value distribution and the probability for a given alignment to obtain a score x larger than z is calculated by integrating the Gumbel distribution:

$$P(x > z) = 1 - \exp\left(-\exp\left(-\frac{\pi}{\sqrt{6}} z - \gamma\right)\right) \quad (5)$$

where $\gamma = 0.5772$, $\pi = 3.1416$ and z is calculated using μ, σ that better fits to the extreme value distribution.

$$z = \frac{x - \mu}{\sigma} \quad (6)$$

2.6 Optimal parameters

The parameters in the SARA program have been optimized using the OPT set of pairwise alignments. The parameters that were optimized for the SARA program include number of atoms to calculate a unit-sphere, open and extension gap penalties and use of secondary structure information (Table 2).

A grid-like search procedure was performed to identify the set of parameters that resulted in the maximum number of superposed nucleotides (n_{al}) for the whole set of alignments. The optimal parameters for SARA using secondary structure information were set to three consecutive unit-vectors between base pairs for each unit-sphere, an opening gap penalty of −7.0, and extension gap penalty of −0.6. If the secondary structure of an RNA molecule is not used, the optimal parameters are seven consecutive unit-vectors, −8.0 opening gap penalty and −0.2 extension gap penalty.

2.7 SARA and ARTS comparison

The default implementations of the SARA and ARTS programs were compared to each other by their accuracy in aligning pairs of RNA structures from the NR95-HR set (i.e. 1275 RNA structural alignments from an all-against-all comparison of 51 chains). The ARTS program is able to align a pair of RNA structures if at least two base pairs can be calculated from each structure. Therefore, 16 chains of the NR95-HR were joined to co-crystallized RNA chains that were base pairing with the query structure. To fairly compare the results from the SARA and ARTS programs, we implemented an algorithm similar to ProSup (Lackner *et al.*, 2000), which calculates a distance matrix from the coordinates of a pair of aligned structures. A distance between two atoms i and j (one from each structure) was used to calculate a distance score ($\text{ds}_{i,j}$):

$$\text{ds}_{i,j} = \max(0, t - d_{i,j}) \quad (7)$$

where t is the distance threshold set to 4.0 Å. The resulting distance matrix is then used in a dynamic programming algorithm with 0 gap penalties, which results in the identification of all equivalent positions that are within 4.0 Å distance between the two aligned structures. The final equivalences (i.e. n_{al}) are used to calculate the PSI as well as the PSS of the alignment.

Table 2. Optimal parameters for the SARA program

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

3 RESULTS

3.1 Structure representation

The accuracy of SARA depends on the atom representing the RNA structure. For proteins, most of the available structure alignment methods consider Cα atom as the best descriptor because consecutive Cα atoms have a conserved distance of ∼3.5 Å. However, the RNA backbone is more flexible than protein backbone and consecutive atoms of the same type have much variable distances (Capriotti and Marti-Renom, 2008). Previous methods developed for RNA structure alignment used different representations. For example, the PRIMOS and COMPADRES programs use specific η and θ pseudo angles between P and C4′ atoms, the ARTS program uses vector distances between P atoms forming a base-pair interaction, and the DIAL program uses an all-atom representation from which several torsion angles are calculated.

To select the atom type representing an RNA structure in SARA, we have used the structures in the NR95-HR set for calculating the distribution of distances between consecutive C3′, C4′, O3′ and P atoms (Fig. 1). The average distances between the backbone atoms were 5.8, 6.1, 6.2 and 6.1 Å for the C3′, C4′, O3′ and P atom types, respectively. The most variable distance corresponded to the P atom type with 0.6 Å of mean standard deviation of the distribution. The C3′ atom type was the most conserved distance among the tested backbone atom types (i.e. 0.4 Å of mean standard deviation). Therefore, the SARA method has been optimized for the use of C3′ atom to represent an RNA structure. Alternatively, SARA can use the P atom trace when no other atoms are present in the crystallographic conformation of the structure.

3.2 Statistical significance of the alignment score

The statistical significance of a pairwise structure alignment score depends on the length of the shorter of the two aligned structures (N). To calculate such statistic, it is necessary to have a background set of pairwise structure alignment scores representing all possible comparisons between two structures. For proteins, such distribution has been typically obtained by comparing randomly selected pairs of known structures. However, for RNA the current coverage of the structural space is limited (i.e. contains only about 1300 entries in the PDB) and is sparsely distributed as indicated by its length distribution (Fig. 2). Three major regions can be detected in the distribution of lengths for known RNA structures: (i) a large number of molecules with lengths shorter than 50 nucleotides, (ii) a peak between 50 and 100 nucleotide mostly populated by tRNA molecules and (iii) a poorly populated region with lengths over

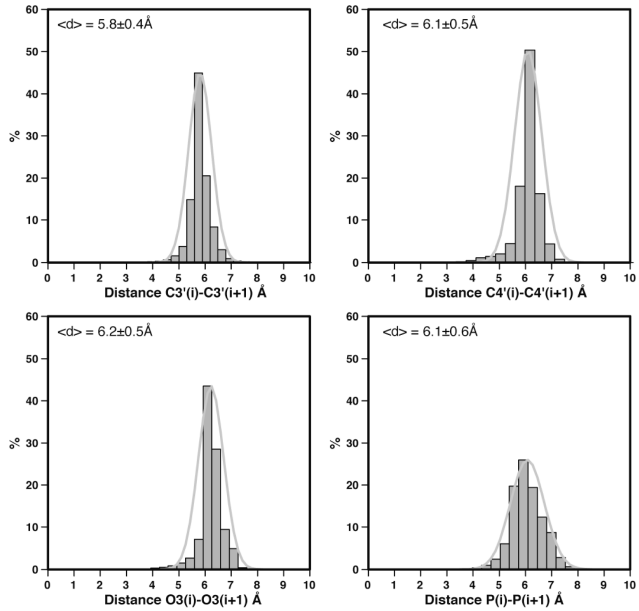

Fig. 1. Distance distributions for consecutive backbone atoms C3′, C4′, O3′ and P calculated from the 51 chains in the NR95-HR set. The average and standard deviation distances are provided within each plot.

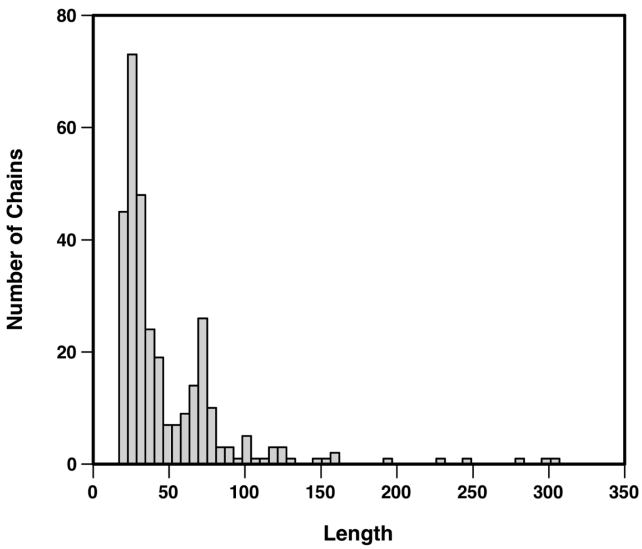

Fig. 2. RNA length distribution for the 277 structure chains in the NR95 set.

100 nucleotides. Such unequal length distribution precludes the use of known RNA structures to calculate a background distribution of pairwise scores. Indeed, the scores resulting from an NR95 all-against-all comparison did not follow an extreme value distribution needed to calculate a statistical significance of a score (data not shown). Thus, we had generated a set of random structures (i.e. RAND set), which was used to calculate a background distribution of alignment scores. The final set of alignments was then divided in 30 bins spamming from 20 to 320 nucleotides of length. The

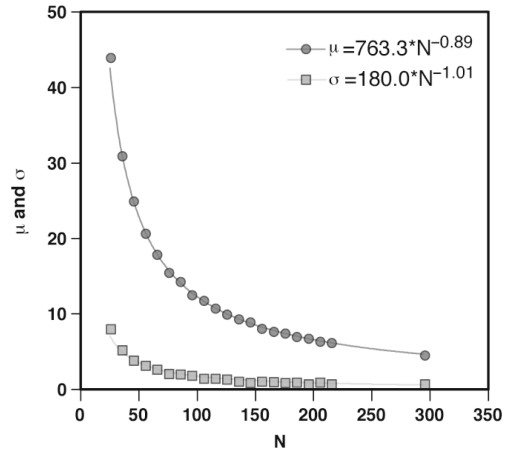

Fig. 3. Fitting of the μ and σ values as a function of the shorter of the two RNA aligned structures (N). The score distribution was obtained from 44 850 pairwise alignments calculated using the RAND set of RNA structures.

Table 3. Statistical significance of a SARA alignment score

$-\ln(P)$ threshold	False positives PSI $\leqslant$ 25%	False negatives PSI $\geqslant$ 75%
5.0	91 (1.6%)	2797 (8.6%)
4.0	135 (1.5%)	1067 (3.7%)
3.0	251 (1.8%)	141 (0.5%)

frequency of the resulting PSI values in each bin was fitted to an extreme value distribution. Finally, a power law function was used for the analytic estimation of μ and σ as a function of N (Fig. 3). The μ and σ parameters were then used to calculate the z-score (z) and the P-value ($P(x>z)$), which estimates the probability to obtain by chance an alignment of a score x larger than z [Equations (5) and (6)].

The ability of the P-value for selecting statistically significant alignments was tested using the alignments from an all-against-all comparison of the NR95 set. The relationship between N, the minus logarithm of the P-value ($-\ln(P)$) and PSI allows us to determine a significance threshold for the selection of accurate alignments (Table 3 and Fig. 4). Setting a conservative threshold of 5.0 $-\ln(P)$, the percentage of false positive alignments [i.e. alignments with $-\ln(P) \geqslant 5.0$ and PSI $\leqslant$ 25%] is 1.6 and the percentage of false negatives [i.e. alignments with $-\ln(P) < 5.0$ and PSI $\geqslant$ 75%] is 8.6. Similar results were obtained for alignments using the secondary structure information.

In summary, for typical RNA structures (i.e. $N < 50$), a $-\ln(P)$ value higher than 5 would most of the time result in an alignment of more than 25 nucleotides within 4.0 Å. Therefore, the results show that the background distribution obtained by comparing randomly generated RNA structures can be used for estimating a score, which correlates with the significance of the alignment at the same time that it is independent of N.

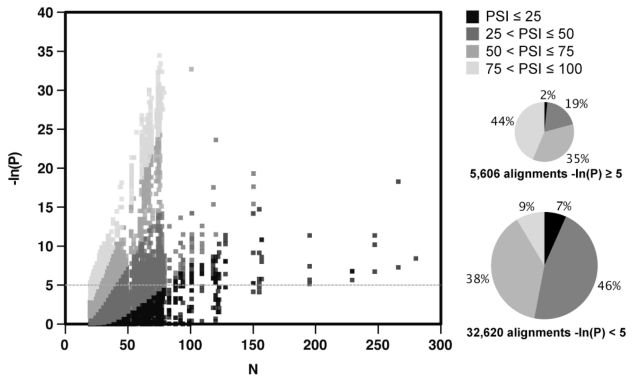

Fig. 4. $-\ln(P)$ scores for the 36 226 pairwise alignments from the all-against-all comparison in the NR95 set are plotted as function of the shorter of the two RNA aligned structures (N). The pie charts, which volumes are proportional to the number of alignments, show the percentage distribution of alignments depending on PSI.

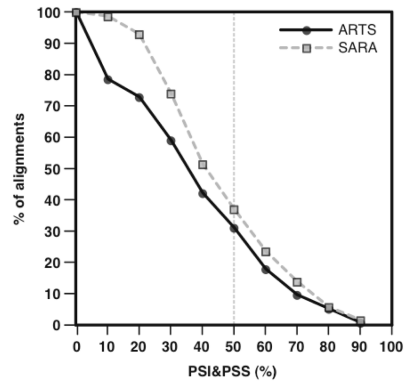

Fig. 5. ARTS and SARA comparison. The percentage of alignments for the all-against-all comparison of the chains in the NR95-HR set is plotted as a function of the minimal percentage of PSI and PSS.

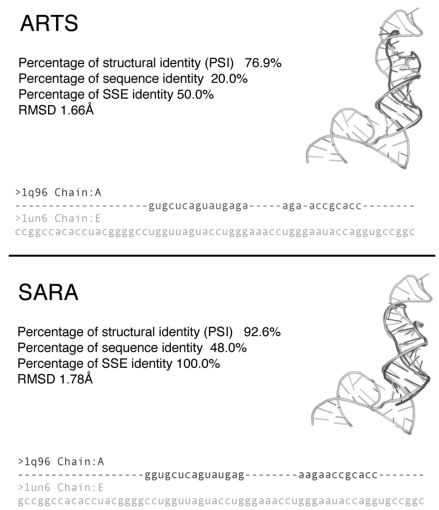

Fig. 6. ARTS and SARA comparison. Structural alignment of a sarcin/ricin domain 28S rRNA (PDB code 1q96 chain A) with a 5S Ribosomal RNA (PDB code 1un6 chain E). The PSI, PSS and RMSD scores were calculated from the superposed coordinates using a local implementation of the ProSup algorithm (see Section 2).

3.3 Alignment accuracy

The alignments produced by SARA were benchmarked for their accuracy compared to the ARTS program and for their use in automatic classification of RNA structures in the SCOR database. The benchmark against the ARTS program was performed using 1275 pairwise alignments from the NR95-HR set. At the time of the benchmark only two similar programs to SARA were available as web servers (Dror *et al.*, 2006; Ferre *et al.*, 2007) and only ARTS was published and available as a stand-alone program (Dror *et al.*, 2005). The benchmark for automatic assignment of RNA function was performed using 1770 pairwise alignments from the NR95-SCOR set. Although not updated since 2004, only one stable classification of RNA structures was available at the time of the benchmark (Tamura *et al.*, 2004).

To fairly compare the alignments produced by SARA and ARTS, all the structures in the benchmark set contained at least two base pairs. To obtain an alignment by ARTS, 16 of the 51 chains in the NR95-HR set had to be complemented by co-crystallized chains to which they were base pairing. Those co-crystallized chains were considered as a single chain for the benchmark exercise. Moreover, both the resulting superposed coordinates by SARA and ARTS were then used to calculate the PSI, and the PSS accuracy scores [Equations (1) and (2)]. It is important to note that, in average, when SARA uses secondary structure information the number of pairwise alignments with at least 50% PSI and PSS increases ∼13% for the NR95-HR set with respect to SARA not using such information.

SARA and ARTS cannot be statistically distinguished for their accuracy in aligning nucleotides of two RNA structures. However, in average, SARA tends to superpose ∼0.2 more nucleotides than ARTS for the whole set of alignments in the NR95-HR set. Moreover, SARA results in more superposed base pairs than ARTS. In particular, SARA has ∼12% more alignments that result in at least 50% PSS than ARTS. Such difference is significant as assessed by the parametric Student's t-test at the 95% confidence value (Marti-Renom *et al.*, 2002). Therefore, SARA results in ∼6% more alignments with both PSI and PSS at least 50% compared to ARTS (i.e. 37 and 31% of the alignments, respectively) (Fig. 5). By selecting only alignment pairs with at least 50% PSI and PSS in any of the two methods (i.e. 593 pairwise alignments), the differences between SARA and ARTS become significant at the 95% confidence value for both the PSI and PSS scores.

The present results quantify the increase in accuracy of RNA structure alignment compared to existing methods such as ARTS. Although such differences in accuracy may seem small, our benchmark indicates that in average the observed differences are statistically significant at the 95% confidence level. For example, the structural alignment of a sarcin/ricin domain 28S rRNA (PDB code 1q96 chain A) with a 5S Ribosomal RNA (PDB code 1un6 chain E) results in 50% more base pairs and 15.7% more nucleotides aligned by SARA compared to the ARTS alignment (Fig. 6). The difference in such alignment are due to a more accurate superposition of a loop region at the tip of the hairpin, which results in a slightly increase of the RMSD from 1.66 Å to 1.78 Å.

Predicting the function of newly discovered molecules is a very challenging task for modern computational biology. Such task is

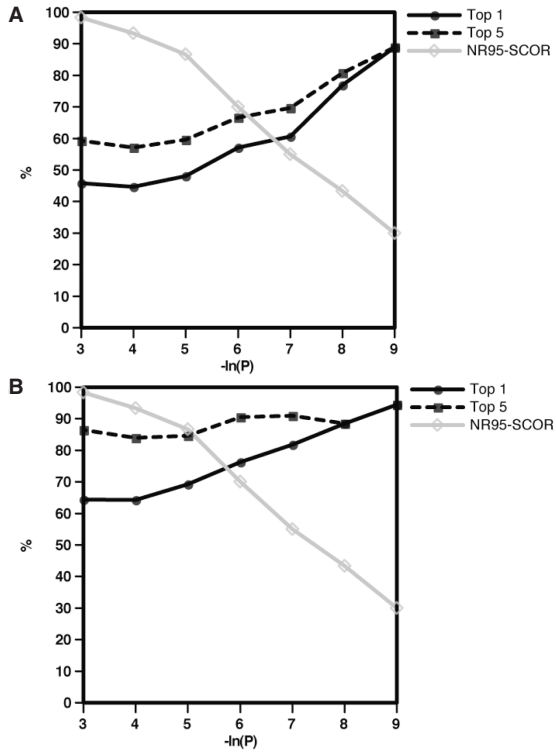

Fig. 7. Automatic assignment of RNA function by SARA. The percentages of correctly assigned SCOR classification (Top 1 and Top 5) as well as the coverage of the NR-SCOR database are plotted as a function of the $-\ln(P)$ alignment score. (**A**) Percentage of correctly assigned deepest SCOR class. (**B**) Percentage of correctly assigned parents in the direct acyclic graph of the SCOR classification.

usually facilitated when the molecular 3D conformation is known. For proteins, Structural Genomics initiatives are in need for faster and more accurate protein structure alignment methods to facilitate their function annotation (Friedberg, 2006; Godzik *et al.*, 2007). Therefore, one of the most appealing tasks for structure alignment methods is to produce sufficiently accurate alignments that would probe useful for transferring molecular function. The alignments produced by SARA were benchmarked for their accuracy for inferring RNA function from annotated structures in the SCOR database. The NR95-SCOR set with 60 chains and 1770 pairwise alignments was used to assess how many times SARA was able to identify as a top hit [i.e. with the highest $-\ln(P)$] and within the Top 5 hits a structure with the same SCOR classification as the query structure. For pairs of alignments with -ln(P) higher than 7, SARA was able to identify as top hit an identical classification for 61% of the query structure (Fig. 7A). Such accuracy increases to 70% if we allow a correct hit within the Top 5 alignments. Considering as correct a hit that has a common parent in the SCOR classification, SARA resulted in an accuracy of 72 and 91% for Top 1 and Top 5 hits, respectively (Fig. 7B).

4 DISCUSSION

The accelerating pace of RNA structures deposition is making RNA structure alignment methods a necessary tool for leveraging the wealth of data in the PDB. Despite such need, only two RNA structure alignment programs of general applicability have been published to date (Dror *et al.*, 2005, 2006; Ferre *et al.*, 2007). The DIAL and ARTS programs usually result in accurate structure alignments between related RNA structures or motifs. However, both programs have certain limitations, which we have tried to address by developing and implementing a new RNA structure alignment algorithm based on a unit-vector approach. The SARA program, introduced in this work, differentiates from the DIAL and ARTS programs in several aspects: (i) the unit-vector approach is of general applicability being possible to obtain a structure alignment even for structures with only a phosphate-trace, with no base pairs or with missing atoms; (ii) compared to DIAL, SARA reports the score of the final alignment as well as its statistical significance within seconds of computational time, which can be used to assess its relevance and (iii) the algorithm implemented in SARA is different to those in DIAL and ARTS making it an alternative choice for aligning two structures of remote similarity when different methods may result in very different pairwise alignments. Moreover, in this work we have also introduced a series of RNA structure sets, which constitute the first stable benchmark set for future development of RNA structure alignment methods. In particular, the RAND set of alignments may prove useful for generating random distributions of RNA structure alignment scores necessary for assessing the statistical significance of pairwise alignments.

SARA takes as input the structures of two RNA molecules and calculates their global alignment within seconds of submission (e.g. <10 s for two RNA structures of $\sim$100 nucleotides). Moreover, SARA calculates a P-value to assess the statistical significance of an alignment score. The negative logarithm of the P-value has proven useful for selecting relevant structure alignment from a set of poorly significant ones. Pairwise structure alignments with a $-\ln(P)$ higher than 3.0 could be considered non-random with a $\sim$95% confidence threshold (i.e. $\sim$0.05 P-value). In average, alignments with $-\ln(P)$ higher than 3.0 have at least 16 superposed nucleotides within 4.0 Å. However, biologically relevant alignments may only be detected with $-\ln(P)$ threshold higher than 5.0. Similar to proteins, there is a decrease of structure identity as the sequence identity between two pairs of RNA molecules decreases. However, about 2% of the pairwise alignments from the NR95 dataset result in percentage of sequence identity <25% and percentage of structural identity >90%. Such alignments have a high percentage of conserved base-pairing ($\sim$85%) indicating that even in very low sequence identity, the secondary structure of RNA is very well conserved. At the threshold of $-\ln(P)$ of 5.0, SARA is able to correctly identify SCOR parents for about $\sim$85% of the NR95-SCOR set with accuracy of $\sim$70% and a $\sim$0.02 probability of false positive detection. Of the 60 chains in the NR95-SCOR set, SARA correctly recognized 27 RNA chains with the same deepest SCOR classification and 38 RNA chains with at least one father in common. The accuracy of SARA was also benchmarked against the ARTS program for its ability for producing pairwise RNA structure alignments with high PSI and PSS. SARA results in $\sim$6% extra pairwise alignments than ARTS with PSI and PSS higher than 50%. Moreover, in average, SARA results in $\sim$0.2 more nucleotides and $\sim$0.4 base pairs superposed for the whole NR95-HR set, which includes non-relevant pairs of alignments. Selecting only those pairwise alignments with PSI and PSS higher than 50% resulted in small but statistically significant differences between SARA and ARTS.

In summary, we have introduced a new algorithm for RNA pairwise structure alignment, which has been implemented in the SARA program. The algorithm was developed for general applicability even when incomplete RNA structures are available. Moreover, a new series of RNA structure and alignment sets have been introduced, which will allow further development of new and existing methods for RNA structure alignment. Despite the accuracy of SARA pairwise alignments and function assignments, the alignment of two molecular structures is always a difficult problem, which usually results in different answers depending on the method used. Therefore, SARA could be considered as a complementary method to those already developed such as DIAL and ARTS.

ACKNOWLEDGEMENTS

We acknowledge support from a Marie Curie International Reintegration Grant (FP6-039722) and Generalitat Valenciana (GV/2007/065). We also thank Oranit Dror, Ruth Nussinov and Haim J. Wolfson for making the latest version of the ARTS program available to us. Finally, this article is in memory of Angel R. Ortiz, whose work has inspired many others, including the one presented here.

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Berman,H.M. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Capriotti,E. and Marti-Renom,M.A. (2008) Computational RNA structure prediction. *Curr., Bioinformatics*, **3**, 32–45.

Chew,L. *et al*. (1999) Fast detection of common geometric substructure in proteins. *J. Comput. Biol.*, **6**, 313–325.

Dorsett,Y. and Tuschl,T. (2004) siRNAs: applications in functional genomics and potential as therapeutics. *Nat. Rev. Drug Discov.*, **3**, 318–329.

Doudna,J.A. (2000) Structural genomics of RNA. *Nat. Struct. Biol.*, **7** (Suppl), 954–956.

Dror,O. *et al*. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21** (Suppl. 2), ii47–ii53.

Dror,O. *et al*. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.*, **34**, W412–W415.

Duarte,C.M. *et al*. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.

Ferre,F. *et al*. (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res.*, **35** (Web Server issue), W659–W668.

Friedberg,I. (2006) Automated protein function prediction–the genomic challenge. *Brief Bioinform.*, **7**, 225–242.

Godzik,A. *et al*. (2007) Computational protein function prediction: are we making progress? *Cell Mol. Life Sci.*, **64**, 2505–2511.

Kedem,K. *et al*. (1999) Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, **37**, 554–564.

Lackner,P. *et al*. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.

Marti-Renom,M.A. *et al*. (2002) Reliability of assessment of protein structure prediction methods. *Structure (Camb.)*, **10**, 435–440.

Murray,L.J. *et al*. (2003) RNA backbone is rotameric. *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.

Ortiz,A.R. *et al*. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Sarver,M. *et al*. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.

Siew,N. *et al*. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*, **3**, e213.

Tamura,M. *et al*. (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.*, **32**, D182–D184.

Tinoco,I.,Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

Wadley,L.M. *et al*. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol.*, **372**, 942–957.

Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.

Zuker,M. and Sankoff,D. (1984) RNA secondary structure and their prediction. *Bull. Math. Biol.*, **46**, 591–621.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.