BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1112–1119
doi:10.1093/bioinformatics/btt096

Sequence analysis

Advance Access publication February 28, 2013

Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package

Carsten Kemena[1,2,†], Giovanni Bussotti[1,2,†], Emidio Capriotti[3], Marc A. Marti-Renom[4,5] and Cedric Notredame[1,2,*]

[1]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain, [3]Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 35249 Birmingham, AL, USA, [4]Structural Genomics Team, Genome Biology Group, Centre Nacional d'Anàlisis Genòmic (CNAG), 08028 Barcelona, Spain and [5]Gene Regulation, Stem Cells and Cancer program, Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Aligning RNAs is useful to search for homologous genes, study evolutionary relationships, detect conserved regions and identify any patterns that may be of biological relevance. Poor levels of conservation among homologs, however, make it difficult to compare RNA sequences, even when considering closely evolutionary related sequences.

Results: We describe SARA-Coffee, a tertiary structure-based multiple RNA aligner, which has been validated using BRAliDARTS, a new benchmark framework designed for evaluating tertiary structure–based multiple RNA aligners. We provide two methods to measure the capacity of alignments to match corresponding secondary and tertiary structure features. On this benchmark, SARA-Coffee outperforms both regular aligners and those using secondary structure information. Furthermore, we show that on sequences in which <60% of the nucleotides form base pairs, primary sequence methods usually perform better than secondary-structure aware aligners.

Availability and implementation: The package and the datasets are available from http://www.tcoffee.org/Projects/saracoffee and http://structure.biofold.org/sara/.

Contact: cedric.notredame@crg.es

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

Received on October 1, 2012; revised on January 25, 2013; accepted on February 18, 2013

1 INTRODUCTION

Recent reports of a large number of previously unknown RNA genes (Guttman *et al.*, 2009) have prompted a renewed interest in the field of non-coding RNA analysis. This shows well in the growing number of scientific reports uncovering a rapidly expanding range of new functions, and it now appears that non-coding RNAs are involved in most parts of the cell machinery, including X inactivation [Xists (Brown *et al.*, 1992)], genome integrity maintenance [piwi-interacting RNA (Farazi *et al.*, 2008)], transcript knockdown and cell differentiation [miRNA (Lee *et al.*, 1993)] as well as nuclear trafficking [NRON (Willingham *et al.*, 2005)], among others. From a functional standpoint, the main consequence of high-throughput sequencing has certainly been the discovery of a large number of long non-coding RNAs (lncRNAs), simultaneously identified as un-reported non-coding ENCODE transcripts (Orom *et al.*, 2010) and as conserved genomic regions with active promoter chromatin signatures (Guttman *et al.*, 2009). The exact function of this new class remains a matter of debate, although mounting bodies of evidence suggest their involvement in gene regulation, either through trans- (Rinn *et al.*, 2007) or cis-acting (Orom *et al.*, 2010) mechanisms. Other reports are also suggesting the potential usage of lncRNAs as biomarkers (Romanuik *et al.*, 2009). In humans only, the latest ENCODE catalog lists >17 000 lncRNA genes, and probably more have to come as a wider range of tissues get deep-sequenced. It remains a matter of debate whether these lncRNAs have evolutionary conserved secondary structure. A difficulty when looking for such structures is the fast evolutionary pace of these molecules, a property that makes it hard to produce structurally informative sequence alignments.

This limitation is rather serious, as our capacity to make sense of so much new information will significantly depend on our ability to build accurate homology-based models (Capriotti and Marti-Renom, 2008a). In the present work, we borrow some concepts developed for protein sequence comparison and show that RNA structural information can be used to derive more informative multiple sequence alignment (MSA) models. This approach amounts to defining a perfect RNA alignment as the one maximizing the matching of structurally equivalent elements. Such accurate alignments are critical for various modeling applications, including evolutionary reconstruction, database search using improved context-free stochastic grammar model and fine-grain structural modeling of novel family members. We show that even a small amount of three-dimensional (3D) or two-dimensional (2D) experimental structure can help improve these models.

Alignment methods rely on the notion that key features are usually preserved by evolution through purifying selection. Multiple comparison models can therefore reveal functional

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

elements that would otherwise be difficult to identify on a single sequence. This is especially true for structured RNA molecules where compensated mutations are frequent signatures for evolutionarily maintained stem loops. This strategy has been extensively used for the successful elucidation of ribosomal RNA secondary structures (Gutell and Fox, 1988). Unfortunately, producing alignments accurate enough to be used for secondary-structure prediction is a challenging task, especially when dealing with distantly related sequences. Two main obstacles exist that prevent the computation of informative homology-based models. First of all, RNA sequences are composed of a four-letter alphabet, with no higher-order meta-alphabet (like proteins' amino acid code) that would help powering statistical analysis. As a consequence, structure similarity becomes hard to infer when sequences have <60% identity (Abraham *et al.*, 2008; Capriotti and Marti-Renom, 2010). Second, for many RNAs, sequence evolution is mostly constrained by the maintenance of secondary structure elements stabilized through a combination of canonical and non-canonical base-pairings. Under such constraints, it has been shown that sequences can evolve rapidly while exploring so-called neutral networks (Huynen *et al.*, 1996). The combination of a small alphabet with rapid evolution makes it difficult to use standard alignment tools like Basic Local Alignment Search Tool-based approaches (Altschul *et al.*, 1990). To address these limitations, one can tap into the evolutionary signal contained in di-nucleotides that results from the co-evolution of adjacent bases. This approach has been recently shown to be effective enough for the improvement of database search accuracy (Bussotti *et al.*, 2011). Unfortunately, the signal thus uncovered is modest and unlikely to result in significantly improved alignments. A more suitable solution involves the simultaneous estimation of sequence and structural conservation using Sankoff algorithm (Sankoff, 1985). As effective as it may be in theory, this approach is hampered by prohibitive memory and CPU requirements, a limitation that has prompted the development of a large number of faster approximate heuristics for the inclusion of secondary structure information when aligning RNA. Some of the most popular tools include R-Coffee (Wilm *et al.*, 2008), LocARNA (Will *et al.*, 2007) and Consan (Dowell and Eddy, 2006). Consan combines expectation maximization with a sophisticated banded dynamic programming strategy, which results in a heuristic approximation of Sankoff algorithm. The Consan algorithm that only aligns two sequences at a time can easily be combined with a consistency-based multiple sequence aligner like T-Coffee (Notredame *et al.*, 2000) or R-Coffee (Wilm *et al.*, 2008) to assemble highly accurate RNA MSAs.

Consistency-based aligners (Do *et al.*, 2005; Notredame *et al.*, 2000; Roshan and Livesay, 2006; Wilm *et al.*, 2008) rely on the compilation of an exhaustive library of all-against-all pairwise alignments. This library is extended to derive a position-specific scoring scheme, used to compute a standard progressive alignment. The main strength of multiple aligners like T-Coffee is to allow any third-party pairwise aligner to be used for the library generation. This property was previously used to generate structure-based protein alignments (O'Sullivan *et al.*, 2004) by combining structural pairwise aligners like SAP (Taylor and Orengo, 1989). We show here how this approach, originally developed for proteins, can easily be extended to RNA sequence alignments, provided suitable pairwise tools are used to build the pairwise library. Structure-based RNA alignment algorithms include SARA (Capriotti and Marti-Renom, 2008b, 2009), DIAL (Ferre *et al.*, 2007), ARTS (Dror *et al.*, 2005), LaJolla (Bauer *et al.*, 2009), R3D Align (Rahrig *et al.*, 2010) and SARSA (Chang *et al.*, 2008). These tools belong to a recently described class of aligners that make use of experimentally derived 3D structures. In this study, we chose the SARA structural aligner that estimates series of unit vectors between consecutive C3′ atoms (shown by the authors to be the most suitable for this task) and aligns them using dynamic programming, to minimize the root mean square deviation between superimposed atoms. As a stand-alone pairwise structural aligner, SARA is directly usable within the T-Coffee framework, and we describe in this article, a framework suitable for validating the effectiveness of combining these two tools for the generation of 3D structure-based RNA MSAs.

The benchmarking of an RNA 3D structure-based method like SARA-Coffee is not an easy task. First of all, one needs reference datasets of sequences with known 3D structures. When it comes to benchmarking RNA alignments, BRAliBase (Gardner *et al.*, 2005) is usually referred to as the reference collection of choice. However, it cannot be used in the context of this work, as <7.2% of the sequences that make up the datasets match a known 3D structure (≥95% identity), thus making it an impractical reference dataset for the benchmark of tertiary-structure aligners. This limitation prompted us to assemble BRAliDARTS, a new reference dataset that only contains DARTS clusters (Abraham *et al.*, 2008) of structurally homologous sequences, further filtered for their suitability (see Section 2). The second issue relates to the nature of the reference. Our goal being the evaluation of a structural aligner, any reliance on a structure-based reference alignment would have turned our approach into the de-facto comparison of two alternative structural alignment strategies (ours and the one used for the reference). We therefore decided to do an alignment-free assessment of our method by evaluating SARA-Coffee's ability to match structurally equivalent features such as pairs of paired residues, or internal structural distances. This comparison of internal structures was carried out by adapting the NiRMSD (Armougom *et al.*, 2006), a method designed to estimate the structural accuracy of protein MSAs (see Section 2), to use distances between the C3′ atoms of the aligned RNA sequences.

SARA-Coffee is heavily dependent on available RNA 3D structure, an information source only available in small quantities. For instance, the latest PDB release (November 2012) contains <4100 chains longer than eight nucleotides (a minimum for the SARA algorithm), with a mere 2325 mapping onto the 6 million or so sequences in Rfam. This shortage severely restricts the scope of a method like SARA-Coffee, and we therefore decided to broaden the scope of this work by going beyond a mere pure structure-based validation. We also tried to estimate the usefulness of 3D structural information when computing MSAs so as to provide the community with guidelines on how these data may be used as efficiently as possible, and also to determine when these data are critically needed. We were especially interested in determining the usefulness of 3D data when doing 2D modeling. This question is relevant in a context where

it may soon be relatively easy to use next-generation sequencing to do massive secondary structure estimation at minimal cost (Kertesz *et al.*, 2010; Wan *et al.*, 2012), even though the reliability of such techniques remains to be established.

2 METHODS

2.1 Benchmarking dataset

Our benchmark is a collection of 41 datasets, each made of several unaligned homologous structures. These datasets were compiled from the DARTS database (Abraham *et al.*, 2008). DARTS stores 1333 RNA structures that can be clustered in 94 structurally homogenous subgroups using ARTS (Dror *et al.*, 2005). Not all DARTS sequences are suitable for the approach described here, and some filtering was needed to define a usable subset. The initial dataset was filtered by: (i) removing all sequences tagged as fragments by DARTS, (ii) converting all non-canonical residue symbols into an N, (iii) updating outdated PDB structures with their newer versions, (iv) removing RNA–DNA hybrids and structures including heteroatoms, (v) removing structures containing less than nine nucleotides, (vi) removing clusters in which X3DNA (Lu and Olson, 2003) failed to extracting at least one secondary structure, (vii) removing structures with discrepancies between the ATOM and the SEQRES PDB fields and (viii) removing clusters with less than three sequences. The final dataset resulted in a total of 41 distinct sequence sets containing 486 structures (see Supplementary Materials). The most common RNA types in this BRAliDARTS are rRNA and tRNA, a more detailed description is shown in Supplementary Table S2. We named this dataset collection BRAliDARTS, by reference to BRAliBase (Gardner *et al.*, 2005), a popular reference dataset used for RNA aligner benchmarks. BRAliDARTS can be downloaded from http://www.tcoffee.org/Projects/saracoffee. We used these same data to define a high-quality subset named BRAliDARTS-HQ. We did so by (i) removing all non X-ray structures, (ii) keeping only structures with a resolution lower than 2.85 Å and (iii) removing all structures in which RNA sequences have a fraction of base-pair residues lower than 48%. This high-quality dataset resulted in a set of 10 clusters with a total of 79 sequences.

2.2 Benchmark

In BRAliDARTS, reference datasets do not come along with reference alignments, and are merely defined as sets of homologous sequences each with an associated 3D structure. It is important to stress that the benchmark strategy described here relies on all the considered sequences having an experimentally known 3D structure. We used two MSA method-independent metrics. The first one is adapted from the NiRMSD (Armougom *et al.*, 2006), a measure originally defined to evaluate protein MSAs by comparing the variation in intra-molecular distances (as inferred from the evaluated MSA itself and the 3D structure of the considered sequences). The NiRMSD can be described as a normalized form of the distance RMSD. In this work, the original package was adapted to evaluate intra-molecular distances using the RNA ribose C3′ instead of the peptidic alpha carbons. We choose the C3′ atom having this most conserved inter-distance of all RNA backbone atoms (Capriotti and Marti-Renom, 2008b) The principle of a distance RMSD is to compare variations of distances between pairs of aligned residues. Its main advantage over a standard RMSD is its non-reliance on a structural superposition. The alignment columns declare equivalent residues, and intra-molecular distances are directly estimated within the non-superposed 3D structures.

The second metric is named Secondary Structure Sum of Pairs (3SP). 3SP is a simple measure estimating the number of base pairs where each side of the pair is aligned with equally contacting residues. We used the m3 implementation of this measure originally described in Notredame *et al.* (1997), and formalized as follows:

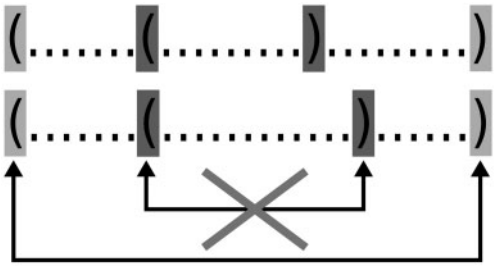

Fig. 1. Schema of the 3SP score computation. The outer base pairs match while the inner one do not thus are not considered for the 3SP computation

$$3SP = \frac{\sum_{i,j} P_{i,j}}{\sum_{i,j} \min(p_i, p_j)}$$

where $P_{i,j}$ is the number of residue pairs found to be in agreement when considering the pairwise alignment of sequences i and j. Residue pairs were estimated from the original PDB structures using the X3DNA. Figure 1 displays a schematic overview of these metrics and shows how when using a Newick-like representation of RNA secondary structures, the 3SP metrics amount to estimating the number of matching parenthesis aligned with equally matching parenthesis, and normalizing this value by its theoretical maximum. Figure 2 shows an example of two colored alignments using this metric.

2.3 SARA-Coffee

Our new method is called SARA-Coffee and is based on R-Coffee. R-Coffee is a consistency aligner for RNA. It can be described as a modified version of T-Coffee able to incorporate predicted secondary structure [RNAPlfold (Bernhart *et al.*, 2006)]. The main advantage of consistency-based aligners is their capacity to better integrate sequence and structural information across large datasets. They may be described as improved progressive alignment heuristics and have been shown to outperform simpler algorithms on many occasions. The default T-Coffee algorithm operates as follows: given a collection of sequences to align, it starts making all the pairwise comparisons and stores them in a data structure named primary library. This primary library can be described as collection of all aligned residue pairs collected from the primary alignments. The library is then processed by combining all possible residue pairs containing one common residue (as defined by indexes and sequences) into residue triplets. The library of triplets is named an extended library. To create a multiple alignment, the sequences are then clustered into a guide tree and incorporated one-by-one into the final MSA while following the guide tree order. At that stage, and in contrast to other aligners, like ClustalW, T-Coffee does not use a substitution matrix but a scoring scheme derived from the extended library. This scoring scheme assigns to the matching of any residue pair, a score equal to the number of triplets (in the extended library) that link this same pair through a third intermediate residue.

This initial T-Coffee algorithm has been modified to align RNA, by adding to the primary library all pairs that would result from the systematic matching of residues predicted to do Watson and Crick (WC) base pairs. For instance, if a residue x of sequence A forms a WC base pair with residue y of that same sequence and is found aligned with residue w of sequence B that forms a WC base pair with residue z of sequence B, then the primary library will contain both the pair xw and the pair yz, even if xz does not appear in any alignment. This approach, similar to the

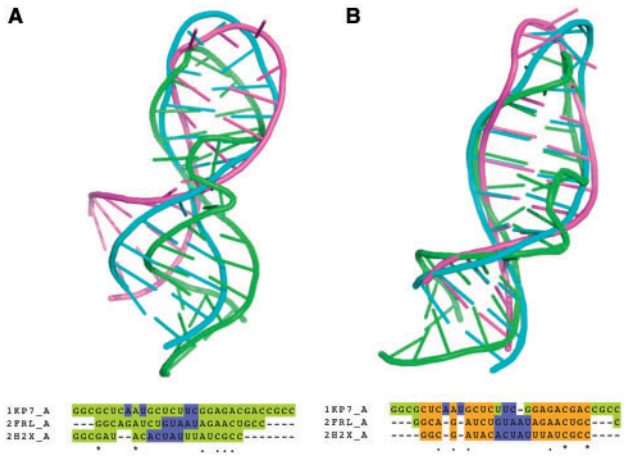

Fig. 2. Example of two alternative RNA MSAs and the associated tertiary structure superposition (top). The dataset is the DARTs cluster 45 (telomerase templates). The MSA in the left panel was produced by ClustalW (3SP score: 0.08, NiRMSD: 8.85 Å, average RMSD: 10.4 Å); the right panel MSA was produced by SARA-Coffee (3SP: 0.96, NiRMSD: 2.74 Å, average RMSD: 2.4 Å. The color gradient on the MSA, from dark-green to orange, indicates the level of agreement of the secondary structures (lowest to highest). Nucleotides colored in violet are not involved in a base-pairing

four-ways consistency described in the RNA alignment flavor of Mafft, amounts to enriching the primary library with structural (predicted or experimental) secondary structure information. The rest of the process is entirely identical to T-Coffee, including the library extension.

The main strength of T-Coffee is its primary library and the possibility it gives to populate this library with any method able to produce high-quality pairwise alignments. In the present work, SARA-Coffee uses the SARA RNA pairwise structural alignment method (Capriotti and Marti-Renom, 2008b) and relies on the assumption that all considered sequences have a known 3D structure. SARA is then used to produce all the structure-based pairwise sequence alignment to populate the library. The 3D structures are also used to extract the correct secondary structures (X3DNA) with which the library is further extended by projecting the base pairs.

To determine the influence of 2D/3D structure information, we also designed two additional T-Coffee implementations: R-CoffeeReal and BestPairs. R-CoffeeReal is the default R-Coffee running with experimental secondary structures rather than predicted. BestPair is a mixture of SARA-Coffee and R-Coffee that runs SARA on a single pair of sequences only (the most closely related) and ignores true structural information for all the other pairs of sequences.

2.4 Alignment comparison

We compared SARA-Coffee with both generic and structure aware aligners. Generic aligners include: ClustalW 1.82 (Larkin *et al.*, 2007), MAFFT (default) 6.624b (Katoh *et al.*, 2005), Probalign 1.4 (Roshan and Livesay, 2006), ProbconsRNA 1.1 (Do *et al.*, 2005) and T-Coffee 8.28 (Notredame *et al.*, 2000). Structure aware aligners use structural information while assembling an MSA. Structural information can either be predicted from single sequences, as in LocARNA 1.6.2 (Will *et al.*, 2007), MAFFT-qinisi 6.864b, MXSCARNA 2.1 (Tabei *et al.*, 2008) and R-Coffee 8.28 (Wilm *et al.*, 2008), or using compensated mutations, as in Consan-Coffee 8.28 (Dowell and Eddy, 2006).

2.5 Implementation/distribution

SARA-Coffee is part of the standard T-Coffee distribution, an open-source freeware available from http://www.tcoffee.org. It requires the SARA program as a plug-in, which is available from http://structure.biofold.org/sara/. The benchmark dataset, including the evaluation procedure, is available from http://www.tcoffee.org/Projects/saracoffee.

3 RESULTS

Our main goal was to estimate the effectiveness of structural information incorporation when assembling RNA MSAs. We were especially interested in quantifying the usefulness of 3D information and its relative merits in comparison with inferred secondary structures. To address this problem, we focused a large part of this work on the design of BRAliDARTS, a structure-based benchmark framework. Aside from its full reliance on 3D information, BRAliDARTS' main strength is its total independence from any reference MSA. This independence makes it possible to avoid any bias toward specific alignment methods.

Having assembled BRAliDARTS, we tested five generic aligners, five secondary-structure aware aligners and the three new methods described here on the 41 datasets of BRAliDARTS. We then calculated 3SP and NiRMSD, the two metrics developed for BRAliDARTS. The 3SP estimates the fraction of base pairs aligned with a potentially homologous pair. This metric merely requires knowing the secondary structure of the considered sequences. The NiRMSD is used to estimate the variation of intra-molecular distances across homologous pairs of residue pairs (as defined by the MSA one evaluates). It relies on the notion that in a correct alignment, the distance between two residues in a structure should be as similar as possible to the distance between homologous residues in another structure. This measure is made local by only considering, for any given residue, the difference of distances between a residue and its neighborhood across a structure. For proteins, a sphere of radius 20 Å was reported to be an optimal size. In the context of this work, we tested three values: 20, 50 and 99 (Supplementary Data). We found the 50 Å limit to yield the most informative correlations, even though no strong differences were observed between 20 and 50 Å cutoffs.

As our aim was to quantify the importance of structural information when assembling an RNA MSA, we first calculated the BP-index (base pair index) for each sequence, representing the fraction of paired residues as determined by X3DNA. Such score varies significantly across datasets and ranges from 6 to 90%, with a median close to 70%. We split the BRAliDARTS accordingly in two subsets, one containing datasets made of low-density secondary structures (21 datasets, BP-index $\leq$70%) and a second one containing high-density structures (20 datasets, BP-index >70%) (Supplementary Table S1). We then averaged readouts for each alignment method in both the high- and the low-density bin (Table 1). On the low-density dataset, we found the 3SP readouts to behave roughly according to expectations, with primary methods delivering results ~10% points lower than their secondary counterparts (0.51 versus 0.60). Tertiary methods like SARA-Coffee were among the best. These observations are in stark disagreements with NiRMSDs readouts that show

primary structure-based methods outperforming those based on secondary structure (6.73 versus 7.27 Å, the lowest values being the best ones). SARA-Coffee is one of the only method consistently delivering the best (NiRMSD) or close to the best (3SP) readouts. The different behavior of the two metrics is reflected well in Figure 3a, where no correlation appears to exist between the two measures. The lack of secondary structure in this dataset explains why the 3SP score fails to properly estimate the 3D alignment accuracy.

By contrast to this first series of results, our measures on the other subset of BRAliDARTS, the one with highly connected structures, gave different results. On this dataset, the 3SP and the NiRMSD measures are strongly correlated (−0.79, Fig. 3b). The differences between primary and secondary or tertiary methods are also much more pronounced. We found SARA-Coffee to be 20 points more accurate (3SP) and 1 Å (NiRMSD) better than sequence-based methods. On these two metrics, five of six secondary structure-based methods outperform all the sequence-based ones. This result confirms that on densely structured sequences, one can improve MSA accuracy by using secondary or tertiary structure information, with the best results being achieved with 3D information. A possible confounding factor when observing this correlation might be the effect of low-resolution structures, in which the BP-index could have been underestimated. In that case, the correlation might have to do more with structural data quality than with base-pairing density. To rule out this possibility, we used BRAliDARTS-HQ, a third reference dataset made of a small number of carefully selected high-quality and high-density structures, and found the correlation to be even stronger (−0.92, Fig. 3c). We also tried to rule out the possibility that the SARA-Coffee improvement might have resulted from the use of experimental (as opposed to predicted) secondary structures rather than tertiary structure information. For that purpose, we designed R-CoffeeReal, an adaptation of R-Coffee that explicitly uses experimental secondary structures. Results (Table 1) show that on the high structural density dataset, R-CoffeeReal manages to improve significantly over R-Coffee on most datasets, regardless of the considered metrics (3SP or NiRMSD).

The burden of requiring an experimental tertiary structure for each RNA sequence one wants to align dramatically limits the scope of SARA-Coffee. Unfortunately, such data are scarce. For instance, of the 4100 RNA PDB chains longer than eight nucleotides available from the PDB, only 2325 have a close homolog (>95% identity) to some of the 6 million sequences reported in Rfam 11.0. To evaluate the effectiveness of SARA-Coffee in a more realistic context, we asked whether using only a handful of structures might be enough to significantly improve MSA modeling (BestPair method). The results are not strongly conclusive,

Table 1. Comparative accuracy

Structural information	Method	BP/nuc < 0.70		BP/nuc ≥ 0.70		Time (s)
		3SP	NiRMSD	3SP	NiRMSD	
Primary	ClustalW	0.46	6.18	0.55	5.74	1
	T-Coffee	0.54	7.08	0.66	5.42	37
	Mafft	0.54	6.30	0.66	5.43	4
	Probalign	0.48	6.51	0.60	5.76	12
	ProbconsRNA	0.55	7.56	0.69	5.54	14
	Average	*0.51*	*6.73*	*0.63*	*5.58*	
Secondary	Mafft-qinsi	0.58	8.11	0.77	5.32	20
	LocARNA	0.64[a]	6.55	0.79	5.11	601
	MXSCARNA	0.61	7.53	0.80	5.45	15
	R-Coffee	0.57	7.36	0.77	5.32	229
	R-CoffeeReal	0.61	7.15	0.80	5.23	511
	Consan-Coffee	0.61	6.91	0.80	5.11	1 168 135
	Average	*0.60*	*7.27*	*0.79*	*5.26*	
Tertiary or secondary	BestPair	0.59	7.14	0.76	5.31	551
Tertiary	SARA-Coffee	0.62	5.69[a]	0.83[a]	4.53[a]	19 324

Structural information indicates the type of information used by the considered method. Method: multiple sequence alignment method. BP: base-pairing threshold of the considered subset. 3SP: average 3SP score; NiRMSD: average NiRMSD in Å; Time: total time on the full set. [a]Best readouts in each column.

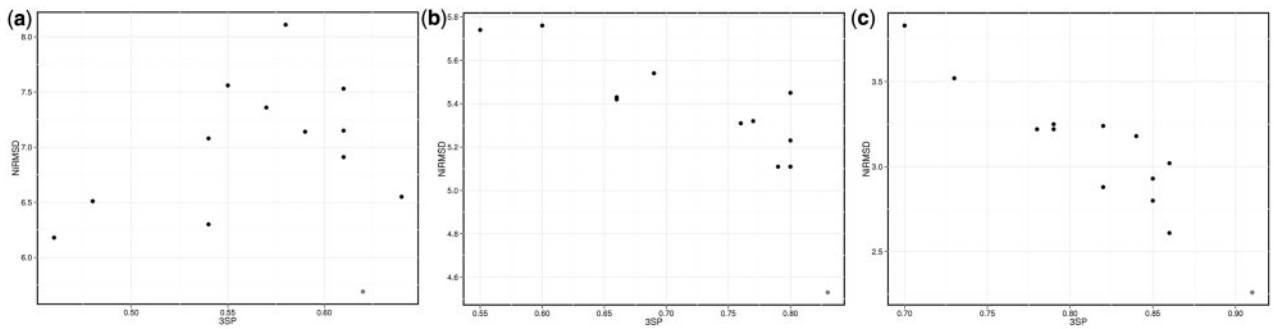

Fig. 3. Correlations between the NiRMSD and the 3SP measure. In all three figures SARA-Coffee is represented by the point with the lowest NiRMSD value. (**a**) Low-density structures: each point represents one of the 13 MSA methods tested here. The horizontal axis corresponds to the average 3SP, expressed as the fraction of base pairs aligned with equally matching pairs (high values correspond to the best readouts). The average was estimated on the 21 datasets in which ≤70% of the nucleotides are involved in a base pair (low-density structures). On this axis, the highest values correspond to the best readouts. The vertical axis corresponds to a similar average made on the NiRMSD readouts. The Pearson correlation between the two variables is 0.19. (**b**) High-density structures: similar graph estimated on the 20 datasets in which >70% of the nucleotides are involved in a base pair. The Pearson correlation is −0.79. (**c**) Same measures on the BRAliDARTS-HQ

with BestPair being only slightly more accurate than other structure-based methods (NiRMSD), albeit significantly less than SARA-Coffee and not significantly more accurate with respect to the 3SP measure than any structure-informed methods. This result, which is rather consistent with similar protein analysis, suggests a strong dependence between the final model accuracy and the overall amount of available structural information.

All together, the results measured on the low- and the high-density BRAliDARTS subsets suggest some heterogeneity and a strong sensitivity to datasets structural composition. When dealing with highly structured sequences, secondary and tertiary structure-based methods perform better, and result in highly correlated secondary (3SP) and tertiary (NiRMSD) readouts. This correlation seems to disappear when analyzing low-density structures. We tested this hypothesis a bit further by taking advantage of the availability of 13 alternative MSAs for each single dataset. This variety allowed us to estimate a Pearson correlation coefficient between the 3SP and the NiRMSD of each single dataset and plot the resulting values against the BP-index (Fig. 4). Despite a rather weak correlation, the trend shows an increasing correlation above a BP-index of 60%, with most datasets with >80% BP-index having strong correlations. This result suggests secondary structure-based methods to be best suited for datasets having a BP-index >80%. We tested this hypothesis by measuring on each dataset, the difference in NiRMSD readouts between primary and secondary structure-based methods (Fig. 5a). As expected, we found that below a BP-index of 60%, primary methods tend to give better results, whereas above this value, secondary-structure aware methods often result in an improvement. A similar analysis carried out by comparing primary- and tertiary-based approaches (Fig. 5b) shows that SARA-Coffee yields its most significant improvements on datasets with a BP-index >60%, but rarely degrades the MSAs below this value.

We completed our analysis by doing a pairwise comparison of all the methods considered here and by counting, for each metric, the number of times any method outperforms any other method (Supplementary Figs S2 and S3). Such a comparison is important, as it makes it possible to estimate the statistical support for the observed differences. We found most differences to be statistically significant on the 3SP method, whereas on the NiRMSD, SARA-Coffee is the only aligner whose behavior appears to be statistically different from most alternatives on most datasets. These comparisons, which reflect individual dataset readouts, also support the notion of secondary structure information being more useful when dealing with highly structured sequences.

The CPU requirements of SARA-Coffee are significantly higher than those of sequence-based methods, and we found our method to be ~100 times slower than LocARNA, but also ~100 times faster than Consan-Coffee, with the number of sequences being the main source of CPU cost (Supplementary Fig. S1). Considering the number of available PDB structures, this makes SARA-Coffee a realistic option for the computation of all currently available datasets on a standard desktop machine.

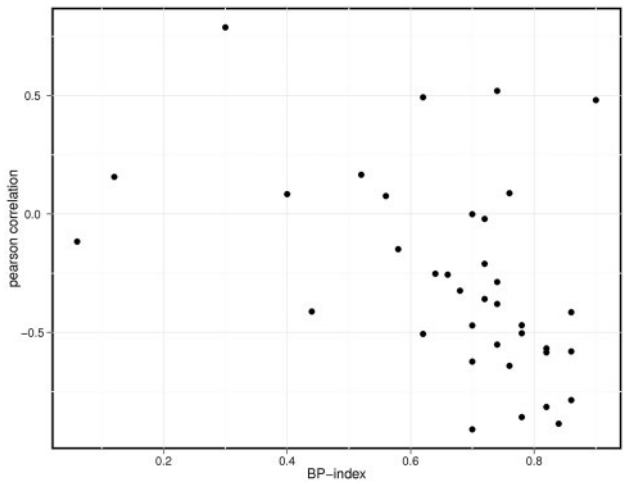

Fig. 4. Dependency of the NiRMSD/3SP correlation on structural density. Each point corresponds to one of the 41 datasets used for benchmark. The horizontal axis indicates the fraction of nucleotides involved in a base pair in the considered dataset. The vertical axis corresponds to the Pearson correlation coefficient between the NiRMSD and the 3SP readouts measured on the MSAs produced using the 13 alternative methods displayed on Table 1

4 DISCUSSION

In this work, we introduce SARA-Coffee, a new tool for generating multiple RNA alignments based on 3D structure. We show how the usage of tertiary information can result in significantly improved RNA multiple alignments. We quantified these improvements using a purpose-built benchmark framework named BRAliDARTS. BRAliDARTS is made of 41 collections of homologous RNA sequences with known 3D structures and two evaluation metrics independent from any reference alignment. The need to assemble our own reference dataset stems from the requirement of having references in which all sequences have a known 3D structure (PDB). This requirement prevented us from testing SARA-Coffee on more established reference datasets like BRAliBase. We nonetheless found that our BRAliDARTS benchmark on the 21 datasets with a high base-pair index shows a good agreement with the ranking of sequence-based alignment methods previously reported on BRAliBase (Wilm *et al.*, 2008). We also found this new dataset to confirm the observation made on BRAliBAse that Consan-Coffee is the most accurate aligner based on *ab-initio* predictions. These observations suggest that BRAliDARTS has, at least for this highly structured component, benchmarking properties similar to those of BRAliBase.

An important focus of our work has been the precise quantification of structural information usefulness when assembling an RNA MSA. The comparison of methods that use primary, secondary and/or tertiary structure information reveals that aligners using secondary structure information are rarely suitable when dealing with sequences in which <60% of the nucleotides are involved in a base pair. Below this figure, methods that rely on predictions appear to induce a degradation of MSA accuracy. By contrast, tertiary structure-based methods like SARA-Coffee almost always manage to improve MSA models accuracy,

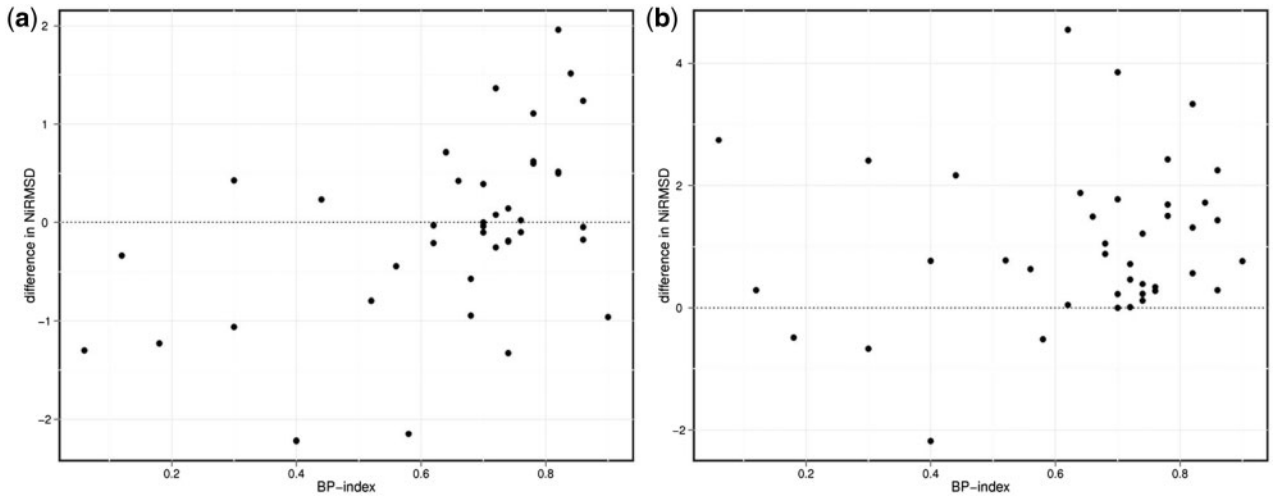

Fig. 5. Relative usefulness of structural information. (**a**) Primary versus secondary structure-based methods. Each point corresponds to one the 31 datasets used for benchmarking. The horizontal axis is the fraction of nucleotides involved in a base pair within the considered sequences (as estimated from the PDB structure). The vertical axis corresponds to the difference in NiRMSD average readouts when comparing the plain sequence alignment methods (ClustalW, T-Coffee, Mafft, Probalign and Probcons4RNA) and the secondary-structure aware methods (Mafft-qiinsi, LocARNA, MXSCARNA, R-Coffee and Consan-Coffee). Positive values indicate readouts where the secondary structure methods outperform their primary sequence counterparts. (**b**) Primary versus tertiary structure-based methods. Similar representation as in (**b**), with sequence methods being compared with SARA-Coffee, a tertiary structure method

regardless of the fraction of structured nucleotides. The easiest explanation for the lack of accuracy of aligners using predicted secondary structures is probably the tendency of over-predicting secondary structures on these sequences. Indeed, a closer inspection on the low-fraction structure shows that these datasets are enriched in heterodimer interactions (RNA/RNA or RNA/proteins), a finding that confirms the well-known issue of accurately predicting *ab-initio* structures without taking into account the folding context.

In the real-world, RNA tertiary structure information is rather scarce, and we therefore had to ask whether alternative sources of information could be reasonable substitutes for tertiary structure data. For instance, it is now possible to do large-scale secondary structure predictions using high-throughput sequencing, and the two leading technologies for single-molecule sequencing techniques, PacBio and Nanopore, have been announcing kits dedicated to large-scale secondary structure determination. It is therefore realistic to consider that a wide amount of secondary structure information will soon be available, although its accuracy still needs to be verified. We tested the effect of using this information using a variation of the R-Coffee method named R-CoffeeReal. Our results are encouraging. They show that when dealing with highly structured RNAs (>70%), the use of experimental secondary structure results in MSAs significantly better than those obtained with alternative secondary methods, even though accuracy does not reach the level of pure tertiary structure-based alignments. This result was also supported by the high correlation (0.92) observed when measured on BRAliDARTS-HQ.

The main limitation of our work is probably its reliance on a rather small collection of datasets. Furthermore, sequences making up this dataset are also rather short (44 nucleotides on average, 213 at most). It is not entirely clear how the behavior of methods relying on secondary structure prediction can be extrapolated to longer sequences, as it is well-known that length tends to impact structure prediction accuracy (Ding *et al.*, 2008; Doshi *et al.*, 2004). By contrast, it is likely that the good performances measured on R-CoffeeReal and all methods using experimental data will hold reasonably well. Our approach is generic and could easily be extended to any pairwise RNA structure aligner. In practical terms, T-Coffee is an open-source package that can be adapted by anyone. It has been designed so as to allow the introduction of any third-party package with minimal effort, thus making the incorporation of new third-party RNA alignment package mostly a benchmarking exercise.

No ideal substitute seems to exist for experimental data when modeling RNA homology. A method like SARA-Coffee could therefore be useful for anyone requiring high-quality MSA modeling for sequences having known 3D structure, and it is probably a realistic expectation that its relevance will grow as RNA structural databases become more populated.

ACKNOWLEDGEMENTS

The authors thank Eric Westhof as well as the three anonymous reviewers for helpful discussions.

Funding: The research leading to these results has received funding from the Centre for Genomic Regulation (CRG) (to C.N. and C.K.); the Spanish Ministerio de Economía y Competitividad through the Plan Nacional (BFU2011-28575); the Euopean Union Seventh Framework Programme (FP7/2007-2013) under grant agreement Quantomics n° KBBE-2A-222664; the Fundació la Caixa within the La Caixa International PhD Fellowship Programme (to G.B.); and the

start-up funds from the Department of Pathology at the University of Alabama, Birmingham (to E.C.); Spanish 60 MINECO (BFU2010-19310 to M.A.M.-R.). It reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein.

Conflict of Interest: none declared.

REFERENCES

Abraham,M. *et al.* (2008) Analysis and classification of RNA tertiary structures. *RNA*, **14**, 2274–2289.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Armougom,F. *et al.* (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.

Bauer,R. *et al.* (2009) Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. *Algorithms*, **2**, 692–709.

Bernhart,S.H. *et al.* (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

Brown,C.J. *et al.* (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

Bussotti,G. *et al.* (2011) BlastR–fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.*, **39**, 6886–6895.

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

Capriotti,E. and Marti-Renom,M.A. (2008b) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.

Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–W265.

Capriotti,E. and Marti-Renom,M.A. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, **11**, 322.

Chang,Y.F. *et al.* (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res.*, **36**, W19–W24.

Ding,F. *et al.* (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.

Do,C.B. *et al.* (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.

Doshi,K.J. *et al.* (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.

Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400.

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

Farazi,T.A. *et al.* (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, **135**, 1201–1214.

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**, W659–W668.

Gardner,P.P. *et al.* (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.

Gutell,R.R. and Fox,G.E. (1988) A compilation of large subunit RNA sequences presented in a structural format. *Nucleic Acids Res.*, **16** (**Suppl.**), r175–r269.

Guttman,M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Huynen,M.A. *et al.* (1996) Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl Acad. Sci. USA*, **93**, 397–401.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lee,R.C. *et al.* (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

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.

Notredame,C. *et al.* (1997) RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **25**, 4570–4580.

Notredame,C. *et al.* (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol.Biol.*, **302**, 205–217.

Orom,U.A. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46–58.

O'Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Rahrig,R.R. *et al.* (2010) R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, **26**, 2689–2697.

Rinn,J.L. *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311–1323.

Romanuik,T.L. *et al.* (2009) Novel biomarkers for prostate cancer including noncoding transcripts. *Am. J. Pathol.*, **175**, 2264–2276.

Roshan,U. and Livesay,D.R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J.Appl. Math.*, **45**, 810–825.

Tabei,Y. *et al.* (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.

Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.

Wan,Y. *et al.* (2012) Genome-wide measurement of RNA folding energies. *Mol. Cell*, **48**, 169–181.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.

Willingham,A.T. *et al.* (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, **309**, 1570–1573.

Wilm,A. *et al.* (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.