RNAconTest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency

ERIK S. WRIGHT

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania 15219, USA

ABSTRACT

The importance of noncoding RNA sequences has become increasingly clear over the past decade. New RNA families are often detected and analyzed using comparative methods based on multiple sequence alignments. Accordingly, a number of programs have been developed for aligning and deriving secondary structures from sets of RNA sequences. Yet, the best tools for these tasks remain unclear because existing benchmarks contain too few sequences belonging to only a small number of RNA families. RNAconTest (RNA consistency test) is a new benchmarking approach relying on the observation that secondary structure is often conserved across highly divergent RNA sequences from the same family. RNAconTest scores multiple sequence alignments based on the level of consistency among known secondary structures belonging to reference sequences in their output alignment. Similarly, consensus secondary structure predictions are scored according to their agreement with one or more known structures in a family. Comparing the performance of 10 popular alignment programs using RNAconTest revealed that DAFS, DECIPHER, LocARNA, and MAFFT created the most structurally consistent alignments. The best consensus secondary structure predictions were generated by DAFS and LocARNA (via RNAalifold). Many of the methods specific to noncoding RNAs exhibited poor scalability as the number or length of input sequences increased, and several programs displayed substantial declines in score as more sequences were aligned. Overall, RNAconTest provides a means of testing and improving tools for comparative RNA analysis, as well as highlighting the best available approaches. RNAconTest is available from the DECIPHER website (<http://DECIPHER.codes/Downloads.html>).

Keywords: multiple sequence alignment; secondary structure prediction; benchmark; noncoding RNA; consensus secondary structure

INTRODUCTION

Multiple sequence alignment forms the basis of many comparative analyses of noncoding RNA sequences such as the prediction of secondary structure. The alignment of RNA sequences remains an unsolved challenge in computational biology (Fallmann et al. 2017). While protein sequences can be accurately aligned down to 30% similarity, RNA sequences become difficult to align below 60% similarity. This discrepancy can be attributed to the much smaller alphabet of RNA sequences (i.e., four nucleotides versus 20 amino acids) and the fact that many different primary sequences can fold into similar secondary structures through complementary base-pairing (Bussotti et al. 2013). For this reason, most RNA-specific tools take advantage of structural conservation to improve alignment performance. Despite the wide variety of tools that have been published, it remains unclear which tools are the best for RNA multiple sequence alignment. However, it is generally believed that programs performing simultane-

ous folding and alignment outperform programs that do not consider structure during alignment.

Several benchmarks for RNA multiple sequence alignment have been published. The popular BRaliBase II (Gardner et al. 2005) and 2.1 (Wilm et al. 2006) benchmarks were constructed from Rfam (Kalvari et al. 2018) seed alignments, which are often derived from alignments created by automated programs. Since the ground truth is unknown in these cases, it is foreseeable that accuracy is partly based on the ability to match other programs' outputs. Furthermore, the BRaliBase benchmarks contain only two to 15 sequences per set and many test sets were generated by repeated resampling of the same few RNA families. The paucity of families with low levels of sequence similarity resulted in the infamous "BRaliBase dent," where programs appear to perform worse at an intermediary similarity range than they do for highly

Corresponding author: eswright@pitt.edu

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.073015.119>.

© 2020 Wright This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://majournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

dissimilar sequences (Löwes et al. 2017). More recently, a collection of manually curated multiple alignments was generated for RNA families with a known structure (Widmann et al. 2012). Although structures can be used to adjust alignments by hand, there is no guarantee that the resulting alignment is optimal (Morrison 2009). Thus, existing benchmarks for RNA multiple sequence alignment leave room for improvement, especially as the number of RNA families and solved 3D structures have continued to increase.

In contrast to RNA, there is a much wider variety of benchmarks for protein multiple sequence alignment. Many benchmarks rely on the superposition of multiple 3D crystal structures to identify colocated residues. Structural superposition is more challenging for RNA sequences because RNAs have modular 3D structures that behave as a collection of semi-independent rigid bodies (Rahrig et al. 2013; Čech et al. 2015; Piątkowski et al. 2017). Even though many more benchmarks are available for proteins than RNAs, there remains little consensus as to which benchmarking approach is optimal (Aniba et al. 2010; Edgar 2010; Iantorno et al. 2014). Recently, a new protein benchmark was developed to overcome some flaws in previous approaches by indirectly comparing protein multiple sequence alignments based on their utility for predicting correct secondary structure (Fox et al. 2015). This approach was shown to strongly correlate with the results of 3D superposition-based protein benchmarks (Sievers and Higgins 2019) but has yet to be applied for benchmarking RNA alignment programs.

The goal of this study was to develop a new benchmark, RNAconTest (RNA consistency test), for assessing RNA multiple sequence alignments based on their degree of secondary structure consistency. Ten different alignment programs were compared using 29 Rfam families with two or more solved structures. This approach is based on the assumption that better secondary structure agreement results from a better alignment, that is, the alignment adequately represents a biologically conserved structure. The same concept was applied to assess consensus secondary structure predictions according to their agreement with empirically determined structures, although many alternative benchmarks already exist for this purpose (Puton et al. 2013; Miao et al. 2017). Collectively, the results reveal differences among programs for RNA alignment and could assist with improving future programs.

RESULTS

Scores on the RNAconTest benchmark are based on the consistency of known secondary structures across multiple sequences in an alignment or when compared to a predicted consensus structure (Fig. 1). Alignment and structure scores were defined using secondary structures in a manner that is analogous to weighted sum of pairs scores that are

commonly used to compare sequence alignments (see Materials and Methods). Each set of reference sequences was derived from Rfam families with one or more solved 3D structures (Table 1). Only the unique sequences in each reference set were kept because many solved structures are redundant. Since reference sets contained a small number of unique sequences (1 to 93) with empirically supported structures, supplemental RNA sequences ($N = 10$ to 5120) were added from each Rfam family's full alignment. This resulted in a set of 29 different Rfam families with two or more solved structures, and 22 additional Rfam families with a single solved structure (Table 1). At least two reference structures are required to calculate an alignment score, whereas only one structure is needed to compare with a consensus secondary structure prediction (Fig. 1).

Ten different alignment programs were tested for their ability to generate structurally consistent alignments. The programs vary in the way in which they handle folding and alignment. Two generic alignment programs, Clustal Omega (Sievers et al. 2011) and MUSCLE (Edgar 2004), do not make use of secondary structure predictions. Three programs, DECIPHER (Wright 2015), MAFFT (Katoh and Standley 2013), and R-Coffee (Wilm et al. 2008), are extensions of programs originally designed for protein alignment but handle RNA sequences differently by incorporating secondary structure predictions. The remaining five programs were specifically designed for the alignment of noncoding RNAs: DAFS (Sato et al. 2012), MASTR (Lindgreen et al. 2007), MXSCARNA (Tabei et al. 2008), LocARNA, and SPARSE (Will et al. 2015). Seven out of ten programs provide a consensus secondary structure prediction with their output (i.e., all except Clustal Omega, MUSCLE, and MAFFT).

Alignment consistency

As expected, all programs generated alignments following a monotonic pattern of lower consistency scores with increasing distance (Fig. 2A). Unlike previous benchmarks, there was no sudden transition to lower scores beyond a specific similarity threshold, suggesting that the concept of an alignment "twilight zone" may be less applicable to RNA sequences than proteins. However, a component of the steady decline in scores may be partly due to true biological variability among secondary structures as the sequences evolve. For example, consistency scores varied ~10% between different reference sets with similar average sequence identities (Fig. 3). Therefore, it is not feasible to determine an upper bound on consistency scores, and RNAconTest benchmark results should only be considered on a relative basis.

On average, LocARNA (0.758) was the highest scoring program (Table 2), followed by MAFFT (0.745), DECIPHER (0.744), and DAFS (0.739). However, the difference among the top scoring programs was not statistically

FIGURE 1. Example scoring for an Rfam family. (A) Five selenocysteine transfer RNA (RF01852) reference sequences were supplemented with 1280 randomly selected sequences from the same Rfam family to generate an alignment using DECIPHER. (B) After discarding the supplemental sequences, alignment scores are calculated for each column (site) in the alignment according to the degree of consistency among their known secondary structures. Structures are shown in dot bracket notation (i.e., base-pairings "(" and ")", pseudoknots "[" and "]"", and unpaired positions "." with chain breaks masked by "+" symbols. (C) These per-site scores are averaged, after weighting by column occupancy, to compute the final alignment score (0.72). (D) In a similar manner, DECIPHER's consensus structure prediction is compared to each reference sequence's known structure to calculate the fraction of matching paired and unpaired bases. These scores are averaged across sequences to obtain the final structure score (0.69). (E) An arc diagram connecting all empirically supported base-pairings in the alignment shows that only a subset of the predicted base-pairings was consistent among sequences. (F) The predicted consensus structure represented most of the consistent base-pairings in the alignment.

significant ($P \geq 0.02$), except in single cases where LocARNA had higher scores than DAFS ($N = 10$) and MAFFT ($N = 20$). For larger sets of sequences, the lack of statistical significance could be attributable to the relatively few alignments completed by LocARNA within the 12 h time limit (Table 2). The statistically significant scores were used to rank programs (Fig. 4), revealing that the highest scoring programs all held positions at the top of the hierarchy. MASTR (0.545) was consistently the lowest scoring program placed at the bottom of the hierarchy. Notably, MUSCLE (0.722) performed reasonably well given that it does not take into account secondary structure. SPARSE (0.662), which is an extension of LocARNA, performed considerably worse than LocARNA. Furthermore, SPARSE displayed considerable score variability within families, suggesting that the heuristics it uses to speed up LocARNA are sometimes detrimental to accuracy. Overall, all alignment programs performed better than expected by random chance, except on the reference sets with lowest average similarity (Fig. 3).

It is known that alignment programs tend to decrease in accuracy as more sequences are incorporated into the alignment (Sievers et al. 2013). Across the range of supplemental sequences added (10 to 5120), Clustal Omega, MASTR, MUSCLE, and R-Coffee exhibited steady falloffs in score as more sequences were aligned (Fig. 2B). In contrast, DAFS, DECIPHER, MAFFT, and MXSCARNA all displayed relatively constant or increasing scores.

Consistency of secondary structure predictions

The scores for consensus secondary structure predictions told a somewhat different story than alignment scores. Structure scores tended to decrease with greater distance among the aligned sequences. However, some programs displayed a local peak in structure scores around 50% sequence identity (Fig. 2C), suggesting that intermediate levels of identity balance the benefits of covariation information against the loss of alignment quality. The highest structure scores were achieved by DAFS (0.690) and LocARNA (0.680), with neither program being statistically significantly better than the other (Table 2). Interestingly, LocARNA uses RNAalifold for their structure predictions, yet R-Coffee (0.524) had considerably lower structure scores despite also relying on RNAalifold. DECIPHER's predictions, which are purely based on mutual information (MI) rather than free energy, improved substantially ($>10\%$ on average) by increasing the number of sequences in the alignment (Fig. 2D). In contrast, MASTR, MXSCARNA, R-Coffee, and SPARSE all exhibited declines in structure scores on larger alignments.

Empirical scalability

Scalability was split into two components: number of sequences being aligned (n) and average length of sequences (l). In terms of elapsed time, Clustal Omega was the

TABLE 1. Reference sets and characteristics

Accession	Structures	Average width	Sequences ^a	Description
RF00001	30	117	140,644	5S ribosomal RNA
RF00002	16	154	4777	5.8S ribosomal RNA
RF00004	1	143	16,770	U2 spliceosomal RNA
RF00005	93	74	1,429,931	tRNA
RF00008	2	52	3180	Hammerhead ribozyme (type III)
RF00009	2	346	1353	Nuclear RNase P
RF00010	2	342	6481	Bacterial RNase P class A
RF00011	1	248	790	Bacterial RNase P class B
RF00013	1	122	3725	6S/SsrS RNA
RF00015	1	92	7692	U4 spliceosomal RNA
RF00017	2	299	42,477	Metazoan signal recognition particle RNA
RF00020	1	113	7704	U5 spliceosomal RNA
RF00023	3	365	6693	Transfer-messenger RNA
RF00026	3	84	47,589	U6 spliceosomal RNA
RF00028	3	228	2611	Group I catalytic intron
RF00029	2	86	15,728	Group II catalytic intron
RF00032	1	26	30,503	Histone 3'-UTR stem-loop
RF00037	1	30	1881	Iron response element I
RF00059	4	80	12,593	TPP riboswitch (THI element)
RF00100	1	57	15,277	7SK RNA
RF00162	2	93	6026	SAM riboswitch (S box leader)
RF00166	1	72	139	PrrB/RsmZ RNA family
RF00167	4	68	2703	Purine riboswitch
RF00168	2	165	2261	Lysine riboswitch
RF00169	2	85	5208	Bacterial small signal recognition particle RNA
RF00174	1	164	14,211	Cobalamin riboswitch
RF00177	35	1476	22,047	Bacterial small subunit ribosomal RNA
RF00234	5	132	943	glmS glucosamine-6-phosphate activated ribozyme
RF00374	2	83	165	Gammaretrovirus core encapsidation signal
RF00380	1	158	1128	ykoK leader
RF00442	1	94	901	Guanidine-I riboswitch
RF00504	3	88	4598	Glycine riboswitch
RF01073	1	59	3211	Gag/pol translational readthrough site
RF01734	2	49	2138	Fluoride riboswitch
RF01739	1	61	1125	Glutamine riboswitch
RF01750	3	71	1673	ZMP/ZTP riboswitch
RF01763	2	41	665	Guanidine-III riboswitch
RF01831	3	92	662	THF riboswitch
RF01846	1	270	643	Fungal small nucleolar RNA U3
RF01852	5	89	2359	Selenocysteine transfer RNA
RF01854	1	264	1143	Bacterial large signal recognition particle RNA
RF01857	1	136	344	Archaeal signal recognition particle RNA
RF01959	1	1495	640	Archaeal small subunit ribosomal RNA
RF01960	29	1817	21,434	Eukaryotic small subunit ribosomal RNA
RF02095	1	71	368	mir-2985-2 microRNA precursor
RF02253	1	29	1202	Iron response element II
RF02540	18	2920	1539	Archaeal large subunit ribosomal RNA
RF02541	46	2904	25,583	Bacterial large subunit ribosomal RNA
RF02543	21	3515	32,268	Eukaryotic large subunit ribosomal RNA
RF02553	1	80	194	Y RNA-like
RF02683	1	86	252	NiCo riboswitch

^aTotal number of sequences in the Rfam family's full alignment used to randomly select supplemental sequences.

FIGURE 2. Trends in score by sequence breadth and number. (A) All programs followed a monotonic trend of decreasing alignment score with increasing distance among aligned sequences. Each curve represents a smoothed cubic spline (with three degrees of freedom) fit to the set of all alignments for a given program. (B) Some programs declined in performance as more supplemental sequences were added to the reference alignment. Lines show the average alignment score relative to the baseline of 10 supplemental sequences. Lines end where fewer than five reference sets were aligned within the 12 h time limit per alignment. (C) Structure scores tended to decline with increasing distance among the aligned sequences. Each curve represents a smoothed cubic spline fit to the set of all alignments. Interestingly, some curves contain a peak at ~50% pairwise identity, particularly for covariation-based prediction methods (e.g., DECIPHER). (D) Average structure scores changed considerably for some programs as more supplemental sequences were included in the alignment. In particular, DECIPHER exhibited the largest gains in performance, whereas MASTR displayed the reverse trend.

fastest aligner tested, followed by DECIPHER. Table 2 shows that both programs scaled nearly linearly with the number of input sequences (n). DECIPHER was the only program to exhibit sublinear scalability with increasing length of sequences, likely due to heuristics it uses to constrain the dynamic programming matrix. Most other programs displayed quadratic or worse limiting behavior in n and/or l . In particular, SPARSE was similar in scalability to LocARNA, implying that its faster speed is not worth its lower accuracy. Clustal Omega (259) and DECIPHER (258) completed the most alignments within the 12 h time limit per alignment, while LocARNA (140) and SPARSE (142) completed the fewest.

DISCUSSION

The RNAconTest benchmark provides an alternative to existing benchmarks for RNA sequence alignments. The approach described here only tests for consistency among known secondary structures since the perfect alignment is unknown. Such a benchmarking approach would

have been impractical a decade ago when far fewer empirical RNA structures were available. The advantages of RNAconTest are that it does not rely on manual curation, encompasses a wide variety of RNA families without substantial redundancy, and more realistically reflects many user scenarios for RNA alignment (e.g., thousands of input sequences). Nevertheless, it relies on the assumption that greater structural consistency corresponds to greater underlying alignment accuracy. This assumption was shown to be reasonable for protein sequences (Sievers and Higgins 2019), but would be invalidated if secondary structure diverged substantially during RNA evolution.

It is known that RNA structures evolve in regions outside of their conserved core. For example, some tRNAs can have a different number of arms than the traditional four arm cloverleaf structure (Pons et al. 2019). Extensive structural variations also exist in RNase P (Brown 1999), group I introns (Jackson et al. 2002), and telomerase (Podlevsky et al. 2008) sequences. This raises the question of whether it is even feasible to align such divergent structures? The assumption made by most alignment software is that the

FIGURE 3. Alignment scores by Rfam family. Reference sets are ordered (left to right) by decreasing average similarity among the reference sequences (values along top horizontal axis). Alignment scores are ordered from fewest to most supplemental sequences within each family (left to right). Horizontal gray lines denote the scores of 100 random permutations of the reference structures belonging to each family, representing the range of results expected from incorrect alignments. Some programs exhibited instability (e.g., SPARSE), with clear failures to accurately align some reference sets belonging to the same family depending on which and/or how many supplemental sequences were added. The top scoring program differed from family to family, but DAFS, DECIPHER, LocARNA, and MAFFT were consistently high scoring.

input sequences are, at a minimum, homologous, if not orthologous (Morrison 2006). If true, it is reasonable to assume that greater structural consistency would result from a better alignment. However, the need for a large number of supplemental sequences likely introduces some false positive input sequences that score above Rfam's gathering cutoff (Nawrocki et al. 2015) for the full alignment but do

not belong in the RNA family. Such sequences would violate the assumption of homology made by most aligners, although they also likely reflect a realistic user scenario. As alignments grow in size, it is possible that not all input sequences are true homologs and alignment software is still expected to generate a high quality alignment of any homologous sequences.

A variety of different measures have been used to gauge the accuracy of secondary structure agreement (Gardner and Giegerich 2004; Xu et al. 2012). Here, the decision was made to focus on a single alignment score representing the agreement among paired and unpaired positions in known structures (see Materials and Methods). The alignment score takes into account the number of correctly/incorrectly aligned positions, as well as positions that were gapped that should have been aligned. The structure score incorporates correctly/incorrectly predicted base-pairings, as well as missing base-pairings. The advantage of these

scoring schemes is that they provide a unified measure of accuracy. Notably, this differs from a traditional benchmark in that the maximum score is unknown, that is, scores can only be compared on a relative (not absolute) basis. Notwithstanding this limitation, the ranking of programs on RNAconTest was generally consistent with previous benchmarks showing that LocARNA is currently the best

TABLE 2. Programs tested and performance characteristics

Multiple alignment program	Structure prediction program	Version	Approximate empirical scalability ^a	Number of alignments completed ^b	Average alignment score ^c	Average structure score ^c
Clustal Omega	N/A	1.2.4	$O(n^{1.0}); O(l^{1.1})$	259	0.704	N/A
DAFS	DAFS	0.0.3	$O(n^{1.5}); O(l^{3.0})$	193	0.739	0.690
DECIPHER	DECIPHER	2.12.0	$O(n^{1.1}); O(l^{0.4})$	258	0.744	0.553
LocARNA	RNAalifold	1.9.2.1	$O(n^{2.0}); O(l^{2.6})$	140	0.758	0.680
MAFFT	N/A	7.427	$O(n^{2.0}); O(l^{1.8})$	194	0.745	N/A
MASTR	MASTR	1.0	$O(n^{1.2}); O(l^{1.8})$	188	0.545	0.475
MUSCLE	N/A	3.8.31	$O(n^{1.6}); O(l^{1.2})$	255	0.722	N/A
MXSCARNA	MXSCARNA	2.1	$O(n^{2.0}); O(l^{3.2})$	207	0.706	0.258
R-Coffee	RNAalifold	12.00	$O(n^{2.1}); O(l^{1.2})$	189	0.721	0.524
SPARSE	RNAalifold	1.9.2.1	$O(n^{2.0}); O(l^{2.6})$	142	0.662	0.593

^aScalability with the total number of sequences (n) and their average length (l).

^bAlignments (out of 259 possible) completed within the 12 h time limit per alignment.

^cAverage score for test sets completed by all programs, weighted equally by family.

FIGURE 4. Competitive hierarchies of programs. (A) Alignment programs were ranked based on whether their alignment scores were greater than other alignment programs, with higher nodes in the hierarchy indicating better performance (see Materials and Methods). Arrows point toward the program that had significantly ($P < 0.02$) lower scores for reference sets with a given number of supplemental sequences (N). DAFS, DECIPHER, LocARNA, and MAFFT were consistently the highest ranked programs. Only DECIPHER and LocARNA were never significantly outperformed by another program. (B) An analogous competitive hierarchy of programs that provided a consensus secondary structure prediction. DAFS and LocARNA dominated the hierarchy, while DECIPHER climbed in rank as more supplemental sequences were added.

available program for noncoding RNA alignment (Löwes et al. 2017).

Taken together, the results of this study suggest that there remains considerable room for improvement in the alignment of noncoding RNA sequences. This was evidenced by the fact that no program substantially outperformed all others across most Rfam families (Fig. 3). Users of alignment programs have many viable options for reasonably accurate RNA alignments. For a small number of input sequences ($n < 1000$), DAFS, LocARNA, and MAFFT all produced similarly accurate alignments. DECIPHER was the most scalable program that maintained high accuracy on alignments of a large number of sequences. DAFS and LocARNA (via RNAalifold) outperformed most other secondary structure prediction tools. Hopefully the availability of a new benchmark will reinvigorate developers' efforts to improve tools for analyzing noncoding RNAs.

MATERIALS AND METHODS

Construction of the RNAconTest benchmark

The benchmark was constructed from the set of 3016 families in Rfam (v14.1). This set was narrowed to 751 with at least 100 sequences available in their full alignment. Available 3D structures were downloaded from the Protein Data Bank (wwPDB consortium 2019) and used to determine an empirical secondary structure with DSSR (Lu and Olson 2003). Briefly, RNA chains were

extracted using PyMOL (v2.4.0a0) to prevent cross-chain bonds within multichain structures and any resulting fragments were mapped to their primary sequence. Masking symbols ("+") were used to denote chain breaks in the resulting reference secondary structures that were excluded from scoring. The final sets were reduced to the set of unique sequences with structures containing at least five paired bases and <10% imbalance between left and right pairings. Imbalance sometimes resulted from structures where the Rfam family only matched a small subsequence of the full-length chain (e.g., RF00029).

A set of up to 10 different input alignments were tested for each Rfam family. These input alignments were generated by supplementing the reference sequences with 10 to 5120 ($\{10 \cdot 2^k : 0 \leq k \leq 9\}$) randomly selected sequences from the same Rfam family (full alignment). These two sets of sequences were combined as the input to each of the alignment programs and then the supplemental sequences were discarded after alignment. This resulted in a set of up to 93 aligned reference sequences that were scored for their secondary structure consistency. A total of 29 Rfam families with two or more solved structures were used for calculating alignment scores (Fig. 3), and 51 Rfam families with one or more solved structures were used for determining consensus secondary structure scores (Fig. 5).

Scoring sequence alignments and consensus structure predictions

For each alignment that was generated, alignment scores were calculated by inserting gaps in the known secondary structures to match those in the (m) aligned reference sequences after

FIGURE 5. Structure scores by Rfam family. Reference sets are ordered (left to right) by decreasing average similarity among the aligned sequences (values along top horizontal axis). Structure scores are ordered from fewest to most supplemental sequences within each family (left to right). Horizontal gray lines denote the scores of 100 random permutations of the reference structures belonging to each family, representing the range of results expected from incorrect predictions. DAFS and LocARNA (via RNAalifold) outperformed the other programs for predicting consensus secondary structures.

removing any supplemental sequences (Fig. 1). Excluding terminal gaps and any masked positions, a score (s) was assigned to each pair of structures (x and y) at a site in the alignment based on whether they were consistent or inconsistent:

$$s(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

For example, if two structures were both unpaired (".") then the position was assigned a score of one, whereas if one was paired and the other was unpaired it was assigned a score of zero. Paired bases were required to pair with the same opposing site in the alignment in order to be considered consistent. At each site (i) in the alignment, the column score (cs) was defined as the number of consistent sequence pairs divided by the total number of possible sequence pairs (i.e., for all combinations of reference structures, j and k):

$$cs_i = \frac{\sum_{j=1}^{m-1} \sum_{k=j+1}^m s(x_j^i, y_k^i)}{m * (m - 1) / 2}$$

The alignment score for a set was calculated as the average of column scores across all (l) sites weighted by occupancy ($1 - f_G$):

$$\text{Alignment score} = \frac{\sum_{i=1}^l (cs_i * (1 - f_G^i))}{\sum_{i=1}^l (1 - f_G^i)}$$

In this way, sites including more nucleotides (i.e., those that have a lower fraction of gaps, f_G) had a greater influence over the alignment score. Thus, the alignment score can range from zero to one, where higher values correspond to greater secondary struc-

ture consistency in the aligned reference sequences.

A structure score was calculated in a similar manner but without requiring multiple reference sequences per Rfam family. Here, a consensus secondary structure prediction was compared to each reference structure individually. A reference's (ref) individual score (ss) was defined as the fraction of paired or unpaired sites in the reference structure that agreed with the consensus ($cons$) secondary structure:

$$ss = \frac{\sum_{i=1}^l s(cons_i, ref_i)}{l}$$

Agreement required that a paired position pair with the same opposing position in both the reference structure and predicted consensus structure. The structure score for a set was the average of scores across all (m) reference sequences in cases where there was more than one reference structure available:

$$\text{Structure score} = \frac{\sum_{j=1}^m ss_j}{m}$$

In this manner, structure scores range from zero to one, with higher values denoting greater agreement between the consensus prediction and the known structures

of reference sequences. Notably, perfect scores may be impossible to attain because of disagreement among known secondary structures or a suboptimal alignment. Nevertheless, higher relative scores are expected to result from better consensus structure predictions.

Structures were permuted randomly to estimate scores achievable by chance alone (i.e., horizontal gray lines in Figs. 3, 5). Here, the set of paired and unpaired positions were sampled without replacement (i.e., shuffled) in each reference structure before calculating an alignment and structure score. This process was repeated 100 times for each reference set to capture the distribution of scores expected from incorrect alignments or consensus secondary structure predictions. DECIPHER's alignments and consensus secondary structure predictions were used for this randomization procedure because it was relatively quick to generate replicate alignments and structure predictions. Structure scores lower than expected by random chance were often achieved by assigning many (incorrect) paired rather than unpaired (i.e., ".") positions in a consensus structure. Unpaired positions can match with any other unpaired position, which would be expected to result in higher scores by chance alone than incorrectly assigned base-pairings.

Alignment and structure prediction programs

Ten different alignment programs were compared, seven of which output a consensus secondary structure prediction (Table 2). Default invocations were used in all cases, with the exceptions of: Clustal Omega ("clustalo -seqtype=RNA -threads=1"),

MAFFT ("mafft-qinsi -thread 1"), and R-Coffee ("tcoffee -mode rcoffee -ncore=1"). Alignments were limited to 12 h using the `timeout` command. For this reason, average scores in Table 2 are only listed for sets completed by all programs. To avoid a potential conflict of interest (Boulesteix et al. 2013) as the developer of DECIPHER, an older version of DECIPHER was used that dated before the creation of RNAconTest. An attempt was made to incorporate other published alignment programs, but many could not be installed or were no longer available.

Unlike the other programs, DECIPHER is a package for the R programming language (R Core Team 2019). The DECIPHER algorithm for RNA sequences is analogous to that previously described for the alignment of protein sequences (Wright 2015). First, a k-mer-based guide tree is used to progressively construct an initial alignment. Then, secondary structure predictions for each sequence are derived from MI scores using an extension of a previously described method (Freyhult et al. 2005). In particular, the MI is modified by weights for each type of base-pairing and an average product correction (Buslje et al. 2009) is applied to the MI matrix. Secondary structures for each individual sequence are predicted from the consensus structure by removing paired positions that are inconsistent with base-pairing (e.g., A/A). Then a search is conducted between the remaining paired positions for stem-loops that are longer than expected by chance. The result is a predicted structure for each sequence, which is then incorporated during alignment with its own substitution matrix in the same way that base pairs are aligned. This process of constructing a guide tree, aligning, and calculating secondary structures is performed for two more iterations (by default) before producing the final alignment and consensus structure.

Ranking programs by score

Aggregate scores for the benchmark as a whole were unobtainable because each program completed a different number of alignments within the 12 h time limit. Instead, a test of statistical significance was performed to decide whether a program outperformed another on the set of reference alignments completed by both programs (for a given number of supplemental sequences). The Wilcoxon signed-rank test (one-sided) with a *P*-value threshold of 0.02 was used to qualify significance. To display these results, the programs were ranked according to a method for assigning hierarchy to directed networks (Gupte et al. 2011). This method relies on a scoring system wherein arrows pointing down the hierarchy are given a positive score (+1), and arrows pointing against the hierarchy are given a negative score (−1) in the same manner. The optimal ranking of programs was determined using the R package `rgenoud` for integer genetic optimization (Mebane and Sekhon 2011) by performing 100 replicates starting from different initial conditions. The most compact ranking was preferred in cases where multiple competitive hierarchies had equivalent scores.

Estimating scalability

Scalability in time (*t*) was approximated using a two-step procedure, first determining scalability with the number of input sequences (*n*) and then with their average length (*l*). In step 1, a line was fitted to the log of *n* versus the log of *t* for each reference

set. The median of the fitted slopes is reported in Table 2 in big-O notation. In step 2, each set's slope was used to remove the effect of *n* on *t*. This resulted in an average time per reference set that was used to determine the slope of the log of *l* versus the log of *t*. The limiting behavior of scalability as *l* becomes large is reported separately from *n* in Table 2. Notably, one processor was used for all programs but additional speedup might be achieved by specifying more than one processor where possible (i.e., Clustal Omega, DECIPHER, MAFFT, and R-Coffee). All tests were performed on a Dell PowerEdge T650 with an Intel Xeon processor (E5-2690 v4 2.6 GHz) and 256 GB of memory running CentOS 7.

Availability of the RNAconTest benchmark

The RNAconTest (v1.0) benchmark and associated results are available from the DECIPHER website (<http://DECIPHER.codes/Downloads.html>). The benchmark consists of reference sequences, structures, and supplemental sequences for each Rfam family, as well as output alignments for all programs tested. Two R functions are included for scoring multiple sequence alignments and consensus secondary structure predictions.

ACKNOWLEDGMENTS

I would like to thank David Hoksza for discussions about multiple structure alignment using MultiSETTER. This study was supported by the National Institute of Allergy and Infectious Diseases at National Institutes of Health (1DP2AI145058-01).

Received August 20, 2019; accepted January 28, 2020.

REFERENCES

- Aniba MR, Poch O, Thompson JD. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res* **38**: 7353–7363. doi:10.1093/nar/gkq625
- Boulesteix AL, Lauer S, Eugster MJ. 2013. A plea for neutral comparison studies in computational sciences. *PLoS One* **8**: e61562. doi:10.1371/journal.pone.0061562
- Brown JW. 1999. The ribonuclease P database. *Nucleic Acids Res* **27**: 314. doi:10.1093/nar/27.1.314
- Buslje CM, Santos J, Delfino JM, Nielsen M. 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* **25**: 1125–1131. doi:10.1093/bioinformatics/btp135
- Bussotti G, Notredame C, Enright AJ. 2013. Detecting and comparing non-coding RNAs in the high-throughput era. *Int J Mol Sci* **14**: 15423–15458. doi:10.3390/ijms140815423
- Čech P, Hoksza D, Svozil D. 2015. MultiSETTER: web server for multiple RNA structure comparison. *BMC Bioinformatics* **16**: 253. doi:10.1186/s12859-015-0696-8
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113. doi:10.1186/1471-2105-5-113
- Edgar RC. 2010. Quality measures for protein alignment benchmarks. *Nucleic Acids Res* **38**: 2145–2153. doi:10.1093/nar/gkp1196
- Fallmann J, Will S, Engelhardt J, Grüning B, Backofen R, Stadler PF. 2017. Recent advances in RNA folding. *J Biotechnol* **261**: 97–104. doi:10.1016/j.jbiotec.2017.07.007

- Fox G, Sievers F, Higgins DG. 2015. Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. *Bioinformatics* **32**: 814–820. doi:10.1093/bioinformatics/btv592
- Freyhult E, Moulton V, Gardner P. 2005. Predicting RNA structure using mutual information. *Appl Bioinformatics* **4**: 53–59. doi:10.2165/00822942-200504010-00006
- Gardner PP, Giegerich R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**: 140. doi:10.1186/1471-2105-5-140
- Gardner PP, Wilm A, Washietl S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* **33**: 2433–2439. doi:10.1093/nar/gki541
- Gupte M, Shankar P, Li J, Muthukrishnan S, Iftode L. 2011. Finding hierarchy in directed online social networks. WWW'11: Proceedings of the 20th international conference on world wide web, pp. 557–566. ACM, New York.
- Iantorno S, Gori K, Goldman N, Gil M, Dessimoz C. 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol Biol* **1079**: 59–73. doi:10.1007/978-1-62703-646-7_4
- Jackson SA, Cannone JJ, Lee JC, Gutell RR, Woodson SA. 2002. Distribution of rRNA introns in the three-dimensional structure of the ribosome. *J Mol Biol* **323**: 35–52. doi:10.1016/S0022-2836(02)00895-1
- Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**: D335–D342. doi:10.1093/nar/gkx1038
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780. doi:10.1093/molbev/mst010
- Lindgreen S, Gardner PP, Krogh A. 2007. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* **23**: 3304–3311. doi:10.1093/bioinformatics/btm525
- Löwes B, Chauve C, Ponty Y, Giegerich R. 2017. The BRaliBase dent—a tale of benchmark design and interpretation. *Brief Bioinformatics* **18**: 306–311.
- Lu XJ, Olson WK. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**: 5108–5121. doi:10.1093/nar/gkg680
- Mebane JWR, Sekhon JS. 2011. Genetic optimization using derivatives: the rgenoud package for R. *J Stat Softw* **42**. doi:10.18637/jss.v042.i11
- Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, Boniecki MJ, Bujnicki JM, Chen S-J, Cheng CY, et al. 2017. RNA-Puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23**: 655–672. doi:10.1261/rna.060368.116
- Morrison DA. 2006. Multiple sequence alignment for phylogenetic purposes. *Aust Syst Bot* **19**: 479–539. doi:10.1071/SB06020
- Morrison DA. 2009. Why would phylogeneticists ignore computerized sequence alignment? *Syst Biol* **58**: 150–158. doi:10.1093/sysbio/syp009
- Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43**: D130–D137. doi:10.1093/nar/gku1063
- Piątkowski P, Jabłońska J, Żyła A, Niedziałek D, Matelska D, Jankowska E, Waleń T, Dawson WK, Bujnicki JM. 2017. SuperNAlign: a new tool for flexible superposition of homologous RNA structures and inference of accurate structure-based sequence alignments. *Nucleic Acids Res* **45**: e150. doi:10.1093/nar/gkx631
- Podlevsky JD, Bley CJ, Omana RV, Qi X, Chen JJ. 2008. The telomerase database. *Nucleic Acids Res* **36**: D339–D343. doi:10.1093/nar/gkm700
- Pons J, Bover P, Bidegaray-Batista L, Arnedo MA. 2019. Arm-less mitochondrial tRNAs conserved for over 30 millions of years in spiders. *BMC Genomics* **20**: 665. doi:10.1186/s12864-019-6026-1
- Puton T, Kozłowski LP, Rother KM, Bujnicki JM. 2013. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res* **41**: 4307–4323. doi:10.1093/nar/gkt101
- R Core Team. 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rahrig RR, Petrov AI, Leontis NB, Zirbel CL. 2013. R3D Align web server for global nucleotide to nucleotide alignments of RNA 3D structures. *Nucleic Acids Res* **41**: W15–W21. doi:10.1093/nar/gkt417
- Sato K, Kato Y, Akutsu T, Asai K, Sakakibara Y. 2012. DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics* **28**: 3218–3224. doi:10.1093/bioinformatics/bts612
- Sievers F, Higgins DG. 2019. QuanTest2: benchmarking multiple sequence alignments using secondary structure prediction. *Bioinformatics* **36**: 90–95. doi:10.1093/bioinformatics/btz552
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539. doi:10.1038/msb.2011.75
- Sievers F, Dineen D, Wilm A, Higgins DG. 2013. Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics* **29**: 989–995. doi:10.1093/bioinformatics/btt093
- Tabai Y, Kiryu H, Kin T, Asai K. 2008. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* **9**: 33. doi:10.1186/1471-2105-9-33
- Widmann J, Stombaugh J, McDonald D, Chocholousova J, Gardner P, Iyer MK, Liu Z, Lozupone CA, Quinn J, Smit S, et al. 2012. RNASTAR: an RNA Structural Alignment Repository that provides insight into the evolution of natural and artificial RNAs. *RNA* **18**: 1319–1327. doi:10.1261/ma.032052.111
- Will S, Otto C, Miladi M, Möhl M, Backofen R. 2015. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics* **31**: 2489–2496. doi:10.1093/bioinformatics/btv185
- Wilm A, Mainz I, Steger G. 2006. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol* **1**: 19. doi:10.1186/1748-7188-1-19
- Wilm A, Higgins DG, Notredame C. 2008. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res* **36**: e52. doi:10.1093/nar/gkn174
- Wright ES. 2015. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**: 322. doi:10.1186/s12859-015-0749-z
- wwPDB consortium. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**: D520–D528. doi:10.1093/nar/gky949
- Xu Z, Almudevar A, Mathews DH. 2012. Statistical evaluation of improvement in RNA secondary structure prediction. *Nucleic Acids Res* **40**: e26. doi:10.1093/nar/gkr1081

RNA

A PUBLICATION OF THE RNA SOCIETY

RNAconTest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency

Erik S. Wright

RNA 2020 26: 531-540 originally published online January 31, 2020
Access the most recent version at doi:[10.1261/ma.073015.119](https://doi.org/10.1261/ma.073015.119)

References

This article cites 42 articles, 2 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/26/5/531.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>