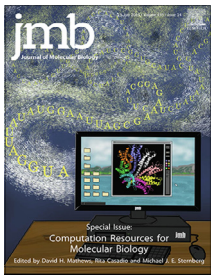

rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling

Chengxin Zhang [1,2,3], **Yang Zhang** [1,4*] **and Anna Marie Pyle** [2,3,5*]

*1 - **Department of Computational Medicine and Bioinformatics,** University of Michigan, Ann Arbor, MI 48109, USA*
*2 - **Department of Molecular,** Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA*
*3 - **Howard Hughes Medical Institute,** Chevy Chase, MD 20815, USA*
*4 - **Department of Biological Chemistry,** University of Michigan, Ann Arbor, MI 48109, USA*
*5 - **Department of Chemistry,** Yale University, New Haven, CT 06511, USA*

Correspondence to Yang ZhangAnna Marie Pyle: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. (Y. Zhang). Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA (A.M. Pyle). *zhang@zhanggroup.org (Y. Zhang), anna.pyle@yale.edu (A.M. Pyle)@pylelab (A.M. Pyle)*
https://doi.org/10.1016/j.jmb.2022.167904
Edited by David Mathews

Abstract

The multiple sequence alignment (MSA) is the entry point of many RNA structure modeling tasks, such as prediction of RNA secondary structure (rSS) and contacts. However, there are few automated programs for generating high quality MSAs of target RNA molecules. We have developed rMSA, a hierarchical pipeline for sensitive search and accurate alignment of RNA homologs for a target RNA. On a diverse set of 365 non-redundant RNA structures, rMSA significantly outperforms an existing MSA generation method (RNAcmap) by approximately 20% and 5% higher F1-scores for rSS and long-range contact prediction, respectively. rMSA is available at https://zhanggroup.org/rMSA/ and https://github.com/pylelab/rMSA.

Introduction

Detecting homologous sequences and aligning them into a multiple sequence alignment (MSA) is the first step for many bioinformatics analyses, such as structure prediction and functional annotation. Due to the importance of MSA in protein biology, especially in protein structure prediction, many sophisticated MSA generation approaches have been proposed.[1–3] It was found that improved MSAs alone can immediately result in more accurate protein structure prediction.

By contrast, in RNA biology, there are few studies on how to construct a high quality MSA for a target sequence, although several studies[4–5] have shown that MSAs are as important to RNA structure prediction as they are for protein structure prediction. To our knowledge, the only two automated pipeline for constructing MSA for a target RNA are RNAcmap[6] and RNAlien.[7] RNAcmap performs a relatively straightforward process of converting blastn[8] hits into a sequence profile (in the form of a covariance model, CM), which is then used by the cmsearch program of Infernal[9] to search the NCBI nucleotide (nt) database. As shown later, this simple approach can lead to unnecessarily large MSAs that are not guaranteed to be optimal for RNA structure prediction. On the other hand, RNAlien performs iterative blastn search, MSA construction and CM building using increasingly permissive taxonomy restraints. Since RNAlien only aligns blastn hits rather than hits from the more sensitive Infernal method, its homolog

detection has limited sensitivities. Consequently, RNAlien alignments tend to be shallow, as shown later.

It should be clarified that our work discusses the generation of an MSA for a target sequence by simultaneous sequence database search and homolog alignment. This is different from the "RNA MSA problem" to generate an MSA for given set of homologs; this problem is addressed by many existing programs such as LocARNA,[10] TurboFold II,[11] and RAF.[12] These algorithms do not perform database searches for an RNA sequence.

Perhaps due to the lack of standard pipelines for RNA homology search and MSA generation, it is a common practice for RNA structure modeling studies[13–16] to use MSAs from the Rfam[17] database for benchmarks. Rfam alignments are usually of exceptionally high quality for three reasons: first, their initial alignments are constructed semi-manually by a human expert and cannot be easily replicated by automated programs; second, most Rfam entries are for relatively well characterized RNA families with a substantial number of sequences; third, 5′ and 3′ termini of RNAs that are not well aligned with other homologs are excluded from the Rfam alignment. Due to the semi-manual curation and biased selection nature of Rfam, aforementioned methodological studies using Rfam MSAs as benchmarks cannot be generalized to a less studied target RNA that is not covered by Rfam.

An area of RNA biology that is in dire need of an automated MSA generator is covariance analysis, also called coevolution analysis, of RNA secondary structure (rSS).[13–14,18] Although RNA rSS prediction is a classical problem that can be addressed by single-sequence-based predictions derived from thermodynamics[19–20] or, more recently, supervised machine learning,[4] MSA-based rSS modeling by covariance analysis still has its unique applications. In a covariance analysis, co-mutation patterns across different positions of the MSA are computed, where pairs of nucleotides with statistically greater amount of co-mutation events are considered more likely to physically interact, e.g., through canonical base pairing. This is a classical approach in RNA bioinformatics and had led to the successful prediction of conserved rSS in 5S rRNAs[21] and introns.[22] Thus, covariance analysis is often used as the statistical evidence for evolutionarily conserved rSS, while alternative methods using thermodynamics or supervised machine learning cannot readily consider the evolutionary aspect of rSS. This is especially true for large RNAs such as long non-coding RNAs and single-stranded RNA genomes, which are known to have well-defined rSS[23–26] but the evolutionary significance of their rSS is elusive.[14,27] Since covariance analysis depends completely on statistical features calculated from the MSA, the correctness of this analysis is contingent on the quality of the MSA.

To meet the need for high quality MSAs in applications such as RNA structure modeling, especially covariance-based secondary structure prediction, we developed rMSA to automate RNA MSA construction by sequence-sequence and profile-sequence alignment (Figure 1). Compared to existing programs for generating RNA MSAs, rMSA consistently and significantly improves prediction of rSS and contacts. While rMSA uses a set of component methods (blastn, Infernal and RNAfold) similar to a previous method (RNAcmap), our algorithm uses a novel five-stage hierarchical sequence search strategy that avoids the excessive incorporation of unrelated sequences. It also features a unique covariance-based MSA selection strategy that consistently improves the final alignment quality. Both strategies are not found in any previous RNA MSA approach, such as RNAcmap. The rMSA webserver and source code are available at https://zhanggroup.org/rMSA/ and https://github.com/pylelab/rMSA, respectively.

Results

Dataset

The rMSA pipeline was tested on a benchmark dataset of 361 non-redundant RNA chains that are collected from the PDB database, where each chain has 30 to 750 nucleotides and at least 10 intra-chain canonical base pairs assigned by DSSR.[28] Only structures with resolution better than 4 Å are included. Any two chains in the dataset must share < 80% sequence identity according to CD-HIT-EST.[29] This 80% identity cutoff has been used in previous studies[4–5] and is the minimal cutoff of CD-HIT-EST. The dataset includes diverse types of RNAs including rRNAs, tRNAs, introns and many others (Table S1).

Overview of the rMSA pipeline

The rMSA pipeline consists of five stages of nucleotide sequence searches and alignments through the standard RNAcentral and nt databases. Each stage corresponds to one column in Figure 1. While RNAcentral collects and annotates many kinds of non-coding RNAs,[30] ~90% of annotated RNAs in this database are tRNAs and rRNAs with molecular function "GO:0003735 structural constituent of ribosome" and "GO:0030533 triplet codon-amino acid adaptor activity", respectively. In this sense, RNAcentral can be considered a subset of the nt database, which collects both genomic and transcriptomic nucleotide sequences, including non-coding RNAs, protein-coding RNAs and non-transcribed regions. As of Nov 9, 2020, RNAcentral and nt contain approxi-

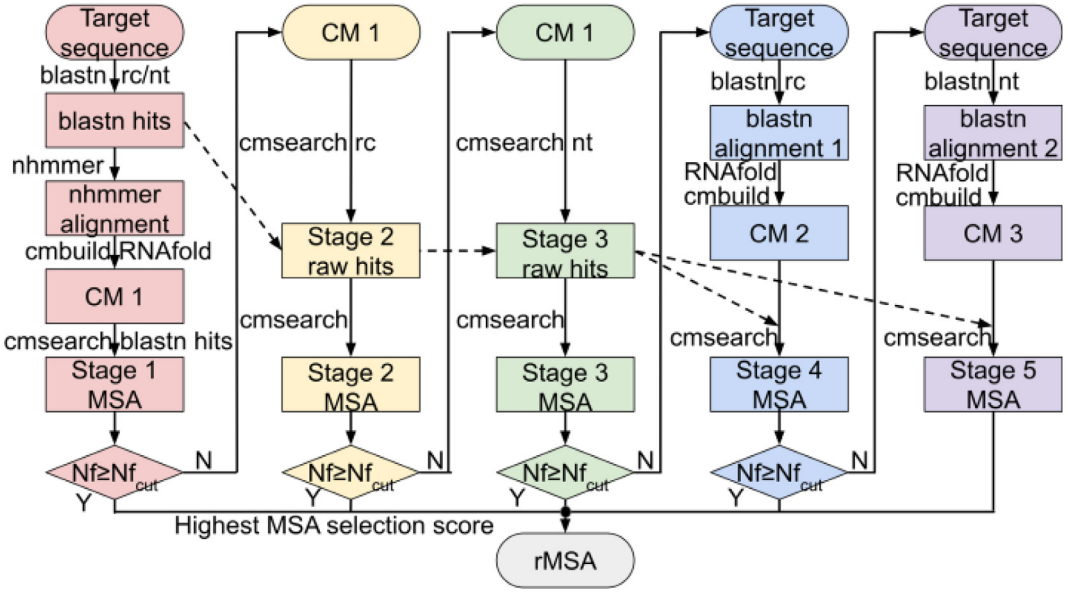

Figure 1. The rMSA pipeline generates five different MSAs. "CM" and "rc" are short for Covariance Model and the RNAcentral database, respectively. Nf_{cut} = 128. The blastn searches are performed with "-max_target_seqs 50,000 -strand plus" and "-max_target_seqs -strand both" options to search only the plus strand in RNAcentral and both strands in nt, respectively, as RNAcentral is a transcriptomic database while nt is a genomic database. For similar reasons, cmsearch was performed using the "–toponly –incE 10.0" option for the plus strand in RNAcentral and "–incE 10.0" for both strands in nt. The e-value cutoff –incE 10.0 is the same as that used in a previous study,[6] where it was found to improve MSA quality. The nhmmer search was performed using "–watson" to only consider alignments with directions that are consistent with blastn alignments.

mately 18 billion and 332 billion nucleotides, respectively.

Depth of MSAs from rMSA

Table 1 compares the depth of alignments from rMSA to those from a state-of-the-art MSA generator (RNAcmap) and three commonly used sequence search tools (Infernal, nhmmer and blastn). Infernal, RNAcmap and rMSA all require an initial rSS for CM construction; this rSS is predicted by RNAfold[20] in this benchmark. In Table 1, the difference between "RNAcmap" and "Infernal" is on how the query CM was constructed: "RNAcmap" constructs the CM from blastn MSA; "Infernal" constructs the CM from the target sequence alone. Similarly, "nhmmer" and "blastn" in Table 1 also start from the target sequence alone instead of a pre-aligned sequence profile. All four third-party MSA programs were used with the default search parameters and they each employ the same sequence databases (RNAcentral and nt) as rMSA. The alignment depth is quantified by the number of unique sequences (*N*) and the number of effective sequences (*Nf*), which is the number of sequences that are non-redundant at 80% sequence identity cutoff divided by the square root of sequence length, in the MSA (Text S1). *Nf* is one of the most commonly used metrics in the structure prediction community for quantifying the depth of an MSA.[31] Here, the 80% sequence identity cut-

Table 1 Average alignment depth of different MSA construction schemes, measured by the number of sequences (*N*) and the number of effective sequences (*Nf*).

MSA	*N*	*Nf*	P-value
rMSA stage 1	3722.6	62.8	**9.83E-21**
rMSA stage 2	4933.5	85.4	**5.05E-15**
rMSA stage 3	5154.6	93.1	**2.51E-5**
rMSA stage 4	5273.0	96.7	**1.02E-4**
rMSA stage 5	5239.2	96.2	**8.01E-5**
rMSA †	5292.0	98.1	*
rMSA (*Nf*) †	5308.8	98.7	9.97E-1
rMSA (*SCI*) †	4285.8	73.9	**2.08E-12**
RNAcmap	23226.8	70.8	**1.83E-3**
Infernal	9177.0	21.9	**2.57E-29**
nhmmer	3111.3	2.9	**4.15E-42**
blastn	1014.1	0.3	**5.07E-44**
RNAlien	2.0	0.1	**1.33E-34**

* All p-values are calculated by one-tail t-test to check if rMSA has higher *Nf* than the respective MSA schema. P-values < 0.05 are in bold.

† "rMSA" is the final MSA output by the standard rMSA. Since the final MSA is selected by a covariance-based MSA score rather than the alignment depth, it is possible for the final MSA to be shallower than the MSAs with the highest *Nf*. The highest *Nf* MSA and highest *SCI* MSA among the MSAs from the five stages are denoted as "rMSA (*Nf*)" and "rMSA (*SCI*)", respectively, in this table.

off in *Nf* is established by a previous study[13] on RNA covariance analysis.

As rMSA progresses from Stage 1 to 4, the alignment depth gradually increases, while the alignment depth of Stage 4 and 5 is comparable (Table 1). On average, 34.2% of the sequences in each stage differ from those of the previous stage. While the highest *Nf* stage is not always selected by rMSA as the final alignment, the *Nf* of the final alignment and that of the deepest alignment is comparable (rMSA versus rMSA (*Nf*) in Table 1). Although rMSA does not produce the deepest alignment in terms of the number of sequences (*N*), it is the method with the highest number of effective sequences (*Nf*). For example, although the average *N* of rMSA is only 22.8% of that of RNAcmap, the average *Nf* of rMSA is 38.6% larger. Since *Nf* considers sequence redundancy up to 80% sequence identity, these data suggest that rMSA alignment is more diverse and includes fewer redundant sequences.

A more detailed analysis is shown in Figure S1, which shows the per-target alignment depths of rMSA alignments versus those of RNAcmap, as well as the overall distribution of alignment depth. A large majority of targets have greater alignment depth by rMSA than by RNAcmap (red in Figure S1(**A**) **and** (**C**)). Nonetheless, RNAcmap has greater alignment depth (blue in Figure S1(**A**) **and** (**C**)) for a small fraction of targets, most of which are tRNAs and large subunit rRNAs. Moreover, the alignment depth of RNAcmap is more unstable, with 22.1% and 1253.4% higher standard deviations for *Nf* and *N*, respectively, than those of rMSA. In other words, the alignment depth of rMSA is more consistent, where ultra-shallow or extra-deep MSAs rarely occur. This is thanks to the multi-stage setup of rMSA, which terminates the pipeline when an intermediate stage already achieves a sufficient alignment depth. More uniform alignment depths are advantageous for downstream covariance analysis, which cannot effectively extract co-mutation statistics from ultra-shallow MSAs and will run out of memory when fed with unnecessarily large MSAs. Since not all input RNAs will go through all five stages of the rMSA, the running time only has a moderate dependency to input sequence length (Pearson Correlation Coefficient = 0.71, Figure S2).

In addition to the standard rMSA pipeline where the final rMSA alignment is selected by the covariance-based MSA score from Equation (1), we also tested the selection of rMSA alignments by *Nf* or Structure Conservation Index (*SCI*; see details in Text S2). Among the alignments from these three selection strategies, the *SCI* selected alignment is the shallowest in terms of both *Nf* and *N*. This is because *SCI* is a measurement of consistency among different sequences in the same MSA,[32] causing high *SCI* alignments to have less diverse sequences.

Deeper MSAs are not always of higher qualities: very large MSAs can be incorrectly constructed, for example, by including too many unrelated sequences or by aligning related sequences to incorrect positions. Therefore, to objectively assess the qualities of the MSAs, the following sections designed two MSA-based RNA structure prediction tasks (rSS and contact predictions).

rMSA improves rSS prediction

The MSA is the sole input for covariance-based rSS prediction. Therefore, the accuracy of covariance-based rSS prediction should be a sensitive indicator of the MSA quality. Two covariance programs are included: PLMC and R-scape. It was recently reported[33] that incorporating thermodynamic parameters into covariance analysis of R-scape via the --RAFSp option (i.e., RNAalifold with stacking) can improve its rSS prediction accuracy for some challenging cases. Therefore, Table 2 includes both the default R-scape and the R-scape running with RAFSp statistics. Table S2 also included another two rSS predictors, PETfold[34] and RNAalifold,[35] both of which combine evolutionary conservation and thermodynamics parameters.

In Table 2, the accuracy of rSS prediction is quantified by two metrics: F1-score and Mathews Correlation Coefficient (MCC) of top *Ln* predicted canonical base pairing, where *Ln* is the number of canonical base pairs in the native structure assigned by DSSR. Details of F1-score and MCC calculation are explained in Text S3. This section only treats the PLMC and R-scape programs as predictors for rSS, regardless of whether the rSS is evolutionarily conserved. Therefore, Table 2 reports all top *Ln* predicted canonical base pairs by R-scape (with and without the RAFSp statistics) regardless of whether the pairs are considered significant (with E-value < 0.05) by R-scape.

As shown in Table 2, rMSA produces significantly more accurate MSAs with 19.3%, 15.0% and 8.9% higher rSS prediction F1-score by PLMC, R-scape and R-scape --RAFSp, respectively, than RNAcmap, which in turn has higher F1-scores than other existing MSA approaches. rMSA internally uses nhmmer, blastn and Infernal. RNAcmap uses blastn and Infernal. It is therefore not surprising that rMSA and RNAcmap outperform nhmmer, blastn and Infernal when these three programs are used individually. On the other hand, although RNAlien also internally uses blastn and Infernal, its performance is much worse than any other pipeline listed in Table 2. This is caused by the shallowness of RNAlien alignments, most of which have no more than 2 sequences (Table 1). Although the comparison among rMSA, RNAcmap and RNAlien is adequate to prove the value of rMSA, we nonetheless keep

Table 2 Average rSS prediction accuracies by different MSA construction and covariance-based rSS prediction schemes. Accuracies are measured by F1-score and MCC, where a perfect prediction would achieve 1 for both metrics.

rSS predictor	MSA[§]	F1	P-value	MCC	P-value
PLMC[†]	rMSA	0.648	*	0.646	*
	rMSA (*Nf*)	0.641	**5.29E-3**	0.639	**5.22E-3**
	rMSA (*SCI*)	0.589	**4.97E-13**	0.586	**5.23E-13**
	RNAcmap	0.543	**3.01E-22**	0.539	**3.34E-22**
	Infernal	0.402	**1.97E-47**	0.396	**2.54E-47**
	nhmmer	0.214	**1.15E-98**	0.208	**1.71E-98**
	blastn	0.040	**1.78E-145**	0.033	**1.90E-145**
	RNAlien	0.024	**4.34E-147**	0.016	**4.81E-147**
R-scape	rMSA	0.575	*	0.572	*
	rMSA (*Nf*)	0.573	2.88E-1	0.570	2.88E-1
	rMSA (*SCI*)	0.546	**2.90E-5**	0.542	**2.90E-5**
	RNAcmap	0.500	**1.50E-13**	0.496	**1.73E-13**
	Infernal	0.390	**1.43E-35**	0.386	**1.97E-35**
	nhmmer	0.230	**2.42E-83**	0.226	**7.66E-83**
	blastn	0.086	**2.92E-122**	0.080	**5.65E-122**
	RNAlien	0.030	**9.54E-130**	0.028	**2.37E-128**
R-scape –RAFSp	rMSA	0.561	*	0.558	*
	rMSA (*Nf*)	0.559	2.78E-1	0.556	2.76E-1
	rMSA (*SCI*)	0.543	**3.06E-3**	0.540	**3.08E-3**
	RNAcmap	0.515	**8.06E-7**	0.511	**8.23E-7**
	Infernal	0.462	**5.39E-17**	0.459	**5.95E-17**
	nhmmer	0.275	**8.73E-74**	0.272	**9.15E-74**
	blastn	0.215	**8.34E-87**	0.209	**1.19E-86**
	RNAlien	0.063	**7.47E-128**	0.061	**5.37E-127**

[*] All p-values are calculated by one-tail t-test to check if rMSA is better (higher F1 and higher MCC) than the respective MSA schema. P-values < 0.05 are in bold.

[†] Apart from canonical base pairs, a covariance analysis can also report other pairwise interactions, such as the coupling between nucleotide pairs adjacent to each other in the sequence. To exclude these non-canonical interactions, the output of covariance analysis is filtered by the following steps before calculating the accuracy: firstly, only Watson-Crick (A:U and G:C) and Wobble (G:U) base pairs are included; secondly, the two nucleotides must be separated by at least 4 positions in the sequence; thirdly, if a base is predicted to simultaneously paired to another two or more bases, only a single base pair with the best covariance score is reported.

[§] Version number of all MSA and rSS prediction programs are listed in Table S3.

Infernal, nhmmer and blastn in our benchmark to understand the extent to which rMSA and RNAcmap provide improvements over their component methods.

The rMSA alignment from the current MSA selection strategy where MSAs from different stages are scored by Equation (1) results in a small but consistent improvement in MSA quality, with 1.1%, 0.3% and 0.4% higher rSS F1-score by PLMC, R-scape and R-scape --RAFSp, respectively, compared to an alternatively rMSA implement where the highest *Nf* MSA is always selected. On the other hand, the rMSA alignment selected by *SCI* is significantly worse than the current MSA selection strategy, with 9.1%. 5.0% and 3.2% reductions in rSS F1-score by PLMC, R-scape and R-scape --RAFSp, respectively, where the p-values are all < 3.06E-3. These data highlight the importance of MSA selection.

It appears that rMSA does not outperform RNAcmap, Infernal or nhmmer for thermodynamics-based rSS prediction by PETfold and RNAalifold (Table S2). This is because, unlike PLMC and R-scape, which can parse the deep MSAs originally produced by different MSA programs, PETfold and RNAalifold require an aggressive filtering of input MSA (Text S4). After this filtering procedure, the rMSA alignment has lost 88.7%, on average, of its original effective sequences, which is consistently higher than RNAcmap, Infernal, nhmmer and blastn, which lose 74.6%, 71.2%, 75.9%, and 66.7% of their original effective sequences, respectively (*Nf* column in Table S4). How to modify rMSA, RNAalifold and PETfold to better leverage a deeper alignment is a topic worthy of future investigation, particularly given that the use of more accurate PETfold rSS prediction can in turn improves the quality of rMSA alignments (Table S5).

Comparison of the rSS prediction performance by different predictors using the same MSA generator in Table 2 reveals two trends that are not in complete agreement with previous studies. Firstly, in our test of rMSA, R-scape is actually slightly more accurate than R-scape --RAFSp, although the latter was suggested by our previous study[33] to produce more accurate rSS predictions under certain constraints. Nonetheless, for all other MSA generators tested, R-scape --RAFSp indeed results

in better accuracy than R-scape. These discrepancies are likely due to the higher *Nf* in rMSA alignments. When *Nf* is small (*Nf* < 22), there are twice as many targets where R-scape --RAFSp outperforms R-scape than targets where R-scape is better; but when *Nf* is big (Nf > 22), most targets have R-scape outperforming R-scape --RAFSp (Figure S3).

Additionally, for rMSA and RNAcmap, the global covariance algorithm, PLMC, is respectively shown to give 12.7% and 8.6% better rSS F1-scores than the local covariance algorithm, R-scape. This is consistent with the conclusions from almost all previous studies[31,36] on protein covariance analysis, where global covariance almost always outperforms local covariance algorithms. Yet, it is in contradiction to the original R-scape study,[14] which concludes no advantage in using a global covariance algorithm over the R-scape local covariance algorithm. These differences in relative performance comparisons among different MSA-based rSS predictors may be caused by differences in MSA construction approaches compared to the original R-scape study, as R-scape indeed outperforms PLMC on shallower MSAs generated by nhmmer and blastn.

Deeper MSAs are not automatically better for rSS prediction. In fact, the rSS prediction by PLMC only has a modest Pearson Correlation Coefficient of 0.55 to logarithm of *Nf* of rMSA alignment for our benchmark targets (Figure S4). This is partially explained by Figure S5: although the rSS prediction accuracy initially improves with increasing *Nf* when *Nf* is small, the accuracy starts to plateau when *Nf* > 64. This is similar to what we observed in our previous study for protein MSAs.[1]

rMSA for RNA long-range contact prediction

While it is evident that MSA is central to covariance-based rSS prediction where all statistics are derived only from the MSA, it is less clear whether MSA is equally important for more complicated machine learning based RNA structure prediction tasks, where not all features are from the MSA. Therefore, this section includes another benchmark using RNAcontact,[37] a deep learning predictor of RNA tertiary contacts. In RNA, rSS is related to but more narrowly defined than contact, as the former usually only includes canonical base pairs while the latter includes all pairs of nucleotides with minimal atomic distance < 8 Å.[13,37] Similar to the previous study,[37] contacts are considered only if they are long-range, i.e. separated by $\geq$ 24 nucleotides in the sequence. The specific sequence separation of 24 for long-range contact is used by many prior studies[1,37–39] in covariance- and machine learning-based contact prediction to define the set of contacts that are most influential for tertiary structure folding. Similar to evaluation of rSS, the accuracy of RNA contact prediction is also evaluated by F1-score and MCC of the top Ln^{long} contacts, where Ln^{long} is the number of long-range contacts in the experimental structure. Four short targets (PDB IDs: 1et4 chain A, 2xdb chain G, 4ato chain G, 5kk5 chain B) are excluded from this benchmark due to lack of long-range contact in their experimental structures.

As shown in Table S6, although rMSA is not specifically optimized for contact prediction, it nonetheless is able to significantly outperform the best third-party MSAs (Infernal) by 4.6% higher F1-score. Overall, the differences between different MSA schemes are smaller than those shown in Table 2. This is mainly because the deep neural network in RNAcontact can extract rich information from very shallow alignments or even the target sequence alone ("Single" in Table 2), thereby making its accuracies less dependent on MSAs.

Conclusions

We have developed rMSA, a free and open-source package for generating high quality RNA MSAs, which significantly improves MSA-based rSS and contact prediction. Detailed analysis showed that the advantage of rMSA lies in its incremental MSA generation scheme that ensures sufficient depth and coverage while preventing redundant or unrelated sequencing populating large MSAs.

We originally developed rMSA for the prediction of RNA structural features such as rSS and contacts, regardless of whether the structural features are evolutionarily conserved. This is reflected in several technical details underlying the implementation of rMSA, such as the use of thermodynamics-based single-sequence rSS prediction (rather than covariance-based prediction), as well as the lack of a built-in subroutine for removal of potential pseudogenes through species-based filtering. An MSA method dedicated to evolutionarily conserved rSS detection still requires further development.

One limitation of rMSA is that it only searches the standard RNAcentral and nt databases, without utilizing metagenome sequence databases. Given the successful application of metagenome database search in protein MSA construction,[1–3] it should be expected that a future version of rMSA incorporating metagenome sequences should result in a further improvement in MSA quality.

Materials and Methods

Details for each of the five stages in rMSA are explained below. In Stage 1, sequence-sequence alignment is performed by blastn to search the target RNA against RNAcentral and nt databases to obtain initial blast hits. These hits are re-aligned by nhmmer[40] to construct an initial alignment. Both blastn and nhmmer are used because the former is

faster while the latter generates alignments with more aligned positions, which is helpful for making the initial alignment. This initial alignment is converted into a CM using the cmbuild/cmcalibrate tools of Infernal. Since a CM cannot be constructed without the pseudoknot-free rSS of a target sequence, a single-sequence-based rSS prediction is performed by RNAfold and fed into cmbuild. The CM is used by the cmsearch program of Infernal to perform a profile-sequence search through the blastn hits to derive the Stage 1 MSA.

In Stage 2 and 3, CM from Stage 1 is searched by cmsearch against the RNAcentral and nt databases, respectively. The raw cmsearch hits are combined with hits from previous stages and re-aligned into the Stage 2 and 3 MSAs by cmsearch.

In Stage 4, the target sequence is searched by blastn against the RNAcentral database. In Stage 5, the target is searched by blastn against the nt database. The two blastn MSAs from the two searches are separately converted into another two CMs (denoted CM 2 and CM 3, respectively), without nhmmer re-alignment. This avoids missing blastn hits that are not included in nhmmer re-alignment, thereby complementing the nhmmer-based CM used in Stage 1 to 3. When building the CMs, the same rSS predicted by RNAfold is used in Stages 1, 4 and 5. CM 2 and CM 3 are searched by cmsearch through sequences collected from the first three stages to obtain Stage 4 MSA and Stage 5 MSA, respectively. To avoid constructing unnecessarily large MSAs, Stage 2 to 5 are only performed if the previous stage has a length-normalized number of effective sequences (*Nf*) < 128. Moreover, Stage 2 to Stage 5 always first check if the MSA derived from the smaller RNAcentral database is sufficiently large before attempting to search the much larger nt database, thereby further avoiding unnecessarily large MSAs. While the *Nf* cutoff of 128 is from our previous study[1] for protein MSAs, it also works well for RNA MSAs in this study.

The final MSA is selected from among the generated MSAs by a covariance-based MSA selection score derived from the Pseudo-Likelihood Maximization algorithm implemented by the PLMC program.[13] To this end, all pairs of nucleotides are ranked in descending order of PLMC covariance scores. Pairs of nucleotides are excluded from consideration if they cannot form canonical base pairs, i.e. not Watson-Crick (A:U and G:C) or Wobble (G:U) base pairs, or separated by less than 4 positions in the sequence. For the remaining base pairs, the top |*BP*| base pairs are used to calculate the MSA selection score, where *BP* is the set of all base pairs predicted by the single-sequence-based rSS predictor (RNAfold) in Stage 1. The MSA score is defined as:

$$MSAscore = \sum_{p=1}^{|BP|}(2 \cdot I[p \epsilon BP] - 1) \cdot plmc_p \quad (1)$$

Here, p is the ranking of base pairs in descending order of PLMC score; $plmc_p$ is the PLMC score of the p-th base pair and it is positively correlated to base pair probability (Figure S6); the Iverson bracket operator $I[\ ]$ equals to 1 if p-th base pairs in PLMC prediction is also within the set of single-sequence rSS prediction, or 0 otherwise. The MSA score measures the agreement between single-sequence-based and MSA-based rSS predictions. The motivation behind this score is that MSAs with higher quality (i.e., more homologous sequences, greater diversities, and less alignment errors) should result in stronger covariance signals and more sensitive prediction of base pairings and contacts, as was found by our recent studies on protein MSA generations.[41–42] The MSA with the highest MSA score among the five MSAs of the five stages is selected as the final rMSA alignment.

All five stages of rMSA uses CM constructed with rSS predicted by RNAfold, which may be less accurate than rSS predicted by deep learning[4,34] or determined by chemical probing such as SHAPE-MaP.[26] Although more accurate rSS may result in a higher quality final MSA, we choose RNAfold to be consistent with previous studies.[6,43] Meanwhile, rMSA provides an option for users to specify their own secondary structure for the input RNA.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests that could have appeared to influence the work reported in this paper. AMP is an associate editor of Journal of Molecular Biology. This does not alter the editorial policy or peer review process for this manuscript.

CRediT authorship contribution statement

Chengxin Zhang: Methodology, Software, Writing – original draft, Writing – review & editing. **Yang Zhang:** Conceptualization, Writing – review & editing. **Anna Marie Pyle:** Supervision, Writing – review & editing.

Acknowledgement

We thank Rafael de Cesaris Araujo Tavares in technical assistances on lncRNA rSS, as well as Drs Xiaoqiong Wei and Li-Tao Guo for insightful discussions. This work used the Extreme Science and Engineering

Discovery Environment (XSEDE), which is supported by the National Science Foundation (ACI1548562). This work was supported by the National Human Genome Research Institute [HG011868 to A.M.P.], National Institute of General Medical Sciences [GM136422 and S10OD026825 to Y.Z.], and the National Institute of Allergy and Infectious Diseases [AI134678 to Y.Z.]. A. M.P. is supported as an Investigator of the Howard Hughes Medical Institute. Funding for open access charge: Howard Hughes Medical Institute.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167904.

Received 2 September 2022;
Accepted 26 November 2022;
Available online 1 December 2022

Keywords:
RNA secondary structure;
multiple sequence alignment;
structure prediction

References

1. Zhang, C., Zheng, W., Mortuza, S.M., Li, Y., Zhang, Y., (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105–2112.
2. Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G.A., Kim, D.E., (2017). Protein structure determination using metagenome sequence data. *Science* **355**, 294–298.
3. Wang, Y., Shi, Q., Yang, P., Zhang, C., Mortuza, S., Xue, Z., (2019). Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biol.* **20**, 1–14.
4. Singh, J., Hanson, J., Paliwal, K., Zhou, Y.Q., (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.*, 10.
5. Hanumanthappa, A.K., Singh, J., Paliwal, K., Singh, J., Zhou, Y., (2020). Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network. *Bioinformatics* **36**, 5169–5176.
6. Zhang, T., Singh, J., Litfin, T., Zhan, J., Paliwal, K., Zhou, Y., (2021). RNAcmap: A Fully Automatic Pipeline for Predicting Contact Maps of RNAs by Evolutionary Coupling Analysis. *Bioinformatics*, btab391.
7. Eggenhofer, F., Hofacker, I.L., Siederdissen Honer Zu, C., (2016). RNAlien - Unsupervised RNA family model construction. *Nucleic Acids Res.* **44**, 8433–8441.
8. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
9. Nawrocki, E.P., Eddy, S.R., (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935.
10. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., Backofen, R., (2007). Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol.* **3**, e65.
11. Tan, Z., Fu, Y., Sharma, G., Mathews, D.H., (2017). TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.* **45**, 11570–11581.
12. Do, C.B., Foo, C.S., Batzoglou, S., (2008). A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics* **24**, i68–i76.
13. Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., Marks, D.S., (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975.
14. Rivas, E., Clements, J., Eddy, S.R., (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48.
15. Bindewald, E., Shapiro, B.A., (2006). RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* **12**, 342–352.
16. Rivas, E., (2021). Evolutionary conservation of RNA sequence and structure. WIREs. *RNA* **n/a**, e1649.
17. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200.
18. Pang, P.S., Jankowsky, E., Wadley, L.M., Pyle, A.M., (2005). Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *J. Exp. Zool. B Mol. Dev. Evol.* **304B**, 50–63.
19. Zhang, H., Zhang, L., Mathews, D.H., Huang, L., (2020). LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics* **36**, i258–i267.
20. Lorenz, R., Bernhart, S.H., Zu Siederdissen, C.H., Tafer, H., Flamm, C., Stadler, P.F., (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 1–14.
21. Fox, G.E., Woese, C.R., (1975). 5S RNA secondary structure. *Nature* **256**, 505–507.
22. Michel, F., Jacquier, A., Dujon, B., (1982). Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* **64**, 867–881.
23. Liu, F., Somarowthu, S., Pyle, A.M., (2017). Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat. Chem. Biol.* **13**, 282.
24. Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., Pyle, A.M., (2015). HOTAIR forms an intricate and modular secondary structure. *Mol. Cell.* **58**, 353–361.
25. Novikova, I.V., Hennelly, S.P., Sanbonmatsu, K.Y., (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.* **40**, 5034–5051.
26. Huston, N.C., Wan, H., Strine, M.S., de Cesaris Araujo Tavares, R., Wilen, C.B., Pyle, A.M., (2021). Comprehensive in vivo secondary structure of the SARS-

CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol Cell*. **81** 584-98 e5.
27. Rivas, E., Clements, J., Eddy, S.R., (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* **36**, 3072–3076.
28. Lu, X.-J., Bussemaker, H.J., Olson, W.K., (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142 -e.
29. Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682.
30. RNAcentral Consortium, (2021). RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* (49), D212–D220.
31. Shrestha, R., Fajardo, E., Gil, N., Fidelis, K., Kryshtafovych, A., Monastyrskyy, B., (2019). Assessing the accuracy of contact predictions in CASP13. *Proteins* **87**, 1058–1068.
32. Washietl, S., Hofacker, I.L., Stadler, P.F., (2005). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**, 2454–2459.
33. Tavares, R.C., Pyle, A.M., Somarowthu, S., (2019). Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. *J. Mol. Biol.* **431**, 1592–1603.
34. Seemann, S.E., Gorodkin, J., Backofen, R., (2008). Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.* **36**, 6355–6362.
35. Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F., (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinf.* **9**, 474.
36. Li, Y., Zhang, C., Bell, E.W., Zheng, W., Zhou, X., Yu, D.-J., (2021). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *Plos Comput Biol.* (In press).
37. Sun, S., Wang, W., Peng, Z., Yang, J., (2020). RNA inter-nucleotide 3D closeness prediction by deep residual neural networks. *Bioinformatics*.
38. Ruiz-Serra, V., Pontes, C., Milanetti, E., Kryshtafovych, A., Lepore, R., Valencia, A., (2021). Assessing the accuracy of contact and distance predictions in CASP14. *Proteins* **89**, 1888–1900.
39. Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M., (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190.
40. Wheeler, T.J., Eddy, S.R., (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489.
41. Li, Y., Zhang, C., Zheng, W., Zhou, X., Bell, E.W., Yu, D.J., (2021). Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins*.
42. Zheng, W., Li, Y., Zhang, C.X., Zhou, X.G., Pearce, R., Bell, E.W., (2021). Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins*.
43. Sun, S., Wu, Q., Peng, Z., Yang, J., (2019). Enhanced prediction of RNA solvent accessibility with long short-term memory neural networks and improved sequence profiles. *Bioinformatics* **35**, 1686–1691.