Bioinformatics, 37(20), 2021, 3494–3500
doi: 10.1093/bioinformatics/btab391
Advance Access Publication Date: 22 May 2021
Original Paper

Structural bioinformatics

RNAcmap: a fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis

Tongchuan Zhang [1], Jaswinder Singh [2], Thomas Litfin[1], Jian Zhan[1], Kuldip Paliwal[2] and Yaoqi Zhou [1,3,]*

[1]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr. Southport, Queensland 4222, Australia, [2]Signal Processing Laboratory, School of Engineering and Built Environment, Griffith University, Brisbane, Queensland 4111, Australia and [3]Institute for Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on February 17, 2020; revised on March 27, 2021; editorial decision on May 14, 2021; accepted on May 18, 2021

Abstract

Motivation: The accuracy of RNA secondary and tertiary structure prediction can be significantly improved by using structural restraints derived from evolutionary coupling or direct coupling analysis. Currently, these coupling analyses relied on manually curated multiple sequence alignments collected in the Rfam database, which contains 3016 families. By comparison, millions of non-coding RNA sequences are known. Here, we established RNAcmap, a fully automatic pipeline that enables evolutionary coupling analysis for any RNA sequences. The homology search was based on the covariance model built by INFERNAL according to two secondary structure predictors: a folding-based algorithm RNAfold and the latest deep-learning method SPOT-RNA.

Results: We showed that the performance of RNAcmap is less dependent on the specific evolutionary coupling tool but is more dependent on the accuracy of secondary structure predictor with the best performance given by RNAcmap (SPOT-RNA). The performance of RNAcmap (SPOT-RNA) is comparable to that based on Rfam-supplied alignment and consistent for those sequences that are not in Rfam collections. Further improvement can be made with a simple meta predictor RNAcmap (SPOT-RNA/RNAfold) depending on which secondary structure predictor can find more homologous sequences. Reliable base-pairing information generated from RNAcmap, for RNAs with high effective homologous sequences, in particular, will be useful for aiding RNA structure prediction.

Availability and implementation: RNAcmap is available as a web server at https://sparks-lab.org/server/rnacmap/ and as a standalone application along with the datasets at https://github.com/sparks-lab-org/RNAcmap_standalone. A platform independent and fully configured docker image of RNAcmap is also provided at https://hub.docker.com/r/jaswindersingh2/rnacmap.

Contact: zhouyq@szbl.ac.cn

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

1 Introduction

RNA structures are the foundations for their diverse functional roles ranging from catalysis, cell-signalling, to transcriptional regulation (Geisler and Coller, 2013). Determining RNA structures by traditional experimental techniques, such as X-ray crystallography, nuclear magnetic resonance and cryogenic electron microscopy, are costly and time-consuming. In fact, only 3% or 99 of 3016 RNA families from Rfam (Kalvari *et al.*, 2018) have experimentally solved structures and the number of solved RNA-only structures per year stay the same for the past two decades (50–80/year). By comparison, the number of non-coding RNAs collected in RNAcentral has doubled from 8 million in 2015 to 16 million in 2019 (Petrov *et al.*, 2015; The RNAcentral Consortium, 2018). The fast-increasing gap between the number of non-coding RNA sequences and the number of experimentally solved structures makes computational approaches highly desirable.

Computational RNA structure predictions were evaluated by RNA-Puzzles (Cruz *et al.*, 2012; Miao *et al.*, 2015, 2017), which were blind experiments in RNA 3-D structure prediction, similar to Critical Assessment of Structure Prediction (CASP) for blind protein structure prediction (Cheng *et al.*, 2019; Kinch *et al.*, 2016; Schaarschmidt *et al.*, 2018). Results of recent three rounds of RNA-Puzzles showed that predicting near-native models (RMSD < 10 Å) remained challenging for most current methods (Cruz *et al.*, 2012; Miao *et al.*, 2015, 2017). However, there is a significant improvement in ab initio protein structure prediction since CASP12 (Schaarschmidt *et al.*, 2018). Such improvement is largely due to employing significantly improved prediction in protein contact maps as the restraints for 3D structure prediction. Prediction of protein contact map was improved by increasingly accurate mutational coupling analysis due to the fast-expanding protein sequence database and more powerful, deep contextual learning for enhancing coupling signals (Cheng *et al.*, 2019; Hanson *et al.*, 2019; Jones *et al.*, 2012, 2015; Wang *et al.*, 2017b).

Contact maps inferred from mutational coupling have been demonstrated to improve RNA secondary (Bernhart *et al.*, 2008; Singh *et al.*, 2021a; Zhang *et al.*, 2020) and tertiary structure prediction (De Leonardis *et al.*, 2015; Wang *et al.*, 2017a; Weinreb *et al.*, 2016). However, tools like RNAalifold (Bernhart *et al.*, 2008) rely on mutual information that is unable to separate direct and indirect coupling (Morcos *et al.*, 2011), while more accurate evolutionary coupling (Wang *et al.*, 2017a; Weinreb *et al.*, 2016) relied on family homologs from the Rfam database, which has 3016 families only as of January 2019. This limitation prevents a wide application of evolutionary coupling for RNA secondary and tertiary structure prediction.

An accurate mutational coupling analysis requires a large number of sequence homologs. RNA homology search is a challenging problem because most structural RNAs are known to preserve the secondary structure rather than the primary sequence (Menzel *et al.*, 2009). The covariance-model-based search enabled by INFERNAL was shown to outperform both sequence-based and profile HMM-based methods with very high sensitivity and specificity, as it incorporates information from both sequence and secondary structure (Freyhult *et al.*, 2007). Recent studies on genome-wide search for pseudoknotted non-coding RNA (Huang *et al.*, 2008; Vasavada *et al.*, 2015) and comparison of RNA multiple sequence alignment tools (Pucci *et al.*, 2019) confirmed the state-of-the-art performance of INFERNAL.

The purpose of this work is to develop a fully automatic pipeline (RNAcmap) for RNA evolutionary coupling analysis that does not rely on well-curated Rfam families. RNAcmap first employs BLAST-N (Altschul *et al.*, 1997) to perform an initial homolog search from the NCBI nucleotide database. The resulting homologous sequences and the predicted secondary structure are then employed for building the covariance model for the second-round search by INFERNAL (Nawrocki and Eddy, 2013), the same tool employed in Rfam to facilitate the comparison. Unlike Rfam that utilizes experimentally validated secondary structures or consensus prediction, RNAcmap employs a folding-based algorithm RNAfold (Lorenz *et al.*, 2011) or a recent deep-learning-based method SPOT-RNA (Singh *et al.*, 2019) for secondary structure prediction to ensure that the method is fully automatic. The resulting multiple sequence alignment from the second-round search is then employed for evolutionary coupling analysis to yield base-pairing and distance-based contact maps. Three methods for evolutionary coupling analysis were examined (GREMLIN, plmc and mfDCA) (De Leonardis *et al.*, 2015; Kamisetty *et al.*, 2013; Weinreb *et al.*, 2016). The pipeline was further tested on two large scale datasets [the PseudoBase++ (Taufer *et al.*, 2008) set and the RNA structure Atlas (Petrov *et al.*, 2013) representative set] and compared with existing tools [RNAalifold (Bernhart *et al.*, 2008) and R-scape (Rivas *et al.*, 2017)].

We showed that the resulting contact maps from RNAcmap (RNAfold/SPOT-RNA) are comparably accurate to those based on Rfam-aligned homologous sequences. Similar accurate results can be achieved for those sequences that are not curated in Rfam and two independent large datasets. The streamlined pipeline should be useful for RNA secondary and tertiary structure prediction tasks.

2 Materials and methods

2.1 The RNAcmap pipeline

In the homology search step, homologs with high sequence similarity (MSA-1 in Fig. 1)is first obtained by running BLAST-N (Altschul *et al.*, 1997) to search the NCBI nucleotide database (parameters: E-value = 0.001, line-length = 1000, num-alignments = 50 000). The

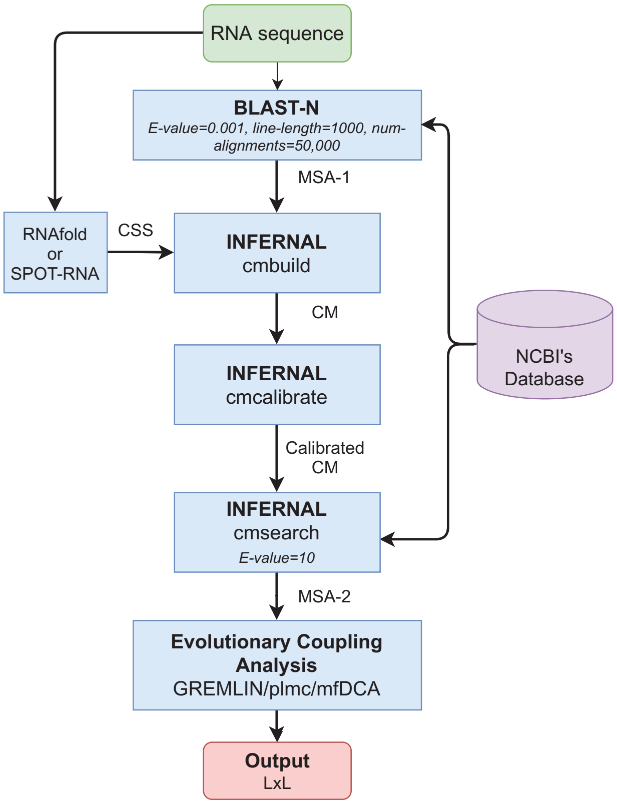

Fig. 1. The architecture of the RNAcmap pipeline. CSS, consensus secondary structure; CM, covariance model; L, length of the input RNA sequence

consensus secondary structure (CSS) for the MSA-1 is obtained from single-sequence-based predictor either RNAfold or SPOT-RNA. Using the homolog sequences (MSA-1) and the predicted CSS as an input, a covariance model (CM) is built by cmbuild from INFERNAL tool (Nawrocki and Eddy, 2013) and then calibrated with cmcalibrate program from INFERNAL as shown in Figure 1. Afterwards, the calibrated covariance model (CM) is employed to perform the second round of search against the NCBI database by using cmsearch program from INFERNAL with E-value of 10.0. This cutoff is chosen in order to include more homologs with low sequence identity (Hanumanthappa *et al.*, 2021; Sun *et al.*, 2019). Finally, the aligned homologous sequences (MSA-2 in Fig. 1) are used for evolutionary coupling analysis with a chosen tool.

2.2 RNA secondary structure prediction tool

We employed either a folding-based algorithm RNAfold (Lorenz *et al.*, 2011) or our recently developed deep-learning method SPOT-RNA (Singh *et al.*, 2019) for secondary structure prediction. SPOT-RNA improves prediction of secondary structure over existing folding-based algorithms, not only in canonical but also in non-canonical and non-nested (pseudoknot) base pairs. The improvement is the largest for non-nested and non-canonical base pairs (Singh *et al.*, 2019).

2.3 Evolutionary coupling analysis tool

Evolutionary coupling analysis, or covariance analysis, is the process to infer evolutionary coupling signals from the sequence alignment. There are two sources of noises in the process: one is the phylogenetic bias and the indirect-coupling effect (Lapedes *et al.*, 1999). Various covariance scores were developed, which can be separated into two categories:

- Local covariance score: Local means that only two sites of alignment columns are considered when calculating the score. Mutual information (MI) is a common method. Recently, R-scape is developed to identify conserved RNA pairs from the alignment using local covariance scores. The phylogenetic bias is overcome by using a null hypothesis to generate synthetic alignments, accounting for phylogenetic correlation and base composition bias (Rivas *et al.*, 2017).

- Global covariance scores: Global means that all covariation scores are calculated under a global probabilistic graph model. This type of scores is designed to overcome indirect-coupling bias. Lapedes and colleagues related the problem of decoupling sequence covariation in alignment with the using model in statistical physics (Lapedes *et al.*, 1999). Later, a variety of methods were developed to learn the model parameters, including message-passing (Weigt *et al.*, 2009), Bayesian network approach (Burger and van Nimwegen, 2010), mean-field [PSICOV (Jones *et al.*, 2012), mfDCA (Morcos *et al.*, 2011)] and pseudo-likelihood [plmDCA (Ekeberg *et al.*, 2013), GREMLIN (Kamisetty *et al.*, 2013), plmc (Weinreb *et al.*, 2016)]. All these methods are collectively given the name of direct coupling analysis (DCA).

In this work, we considered three DCA method, GREMLIN, plmc and mfDCA because they were applied to RNA alignments before or had the option to deal with RNA alignment. GREMLIN (Kamisetty *et al.*, 2013) is obtained from https://github.com/sokrypton/GREMLIN_CPP. The method is based on the pseudo-likelihood inference of direct coupling analysis with L2 regularization. GREMLIN is run with the recommended parameter for RNA (-alphabet rna-gap_cutoff 1.0-lambda 0.01-eff_cutoff 0.8-max_iter 100). The plmc method was obtained from https://github.com/debbiemarkslab/plmc. It employed similar pseudo-likelihood to infer the parameters in the DCA model (Weinreb *et al.*, 2016). We utilized the recommended parameters for RNA (-a -.ACGU -le 20 -lh 0.01 -m 50). The mfDCA method was obtained from http://dca.rice.edu/portal/dca/download. It employed the inverse covariance matrix to infer the coupling parameters (De Leonardis *et al.*, 2015; Morcos *et al.*, 2011). An additional Average Production Correction (APC) (Dunn *et al.*, 2008) was applied to the original mfDCA to be consistent with other DCA methods. In addition to above three DCA predictors, we also consdered R-scape (Rivas *et al.*, 2017) and alignment based folding method RNAalifold (Bernhart *et al.*, 2008) for comparison. They were downloaded from http://www.eddylab.org/R-scape/ and https://www.tbi.univie.ac.at/RNA/, respectively.

3 Datasets

3.1 PDB dataset

We downloaded a total of 4528 structures containing 6294 RNA chains from the Protein Data Bank (PDB). Among them, 4281 RNA chains were selected with sequence length between 32 and 500. Using cmfind from INFERNAL and Rfam database (Version 14.1), these chains were further split into two sets: 3182 RNA chains were mapped to existing Rfam families and 1099 RNA chains were not mapped to any Rfam families. The majority of structure-mapped Rfam families (77%, 2461 of 3182) are tRNA, 5S rRNA and 5.8S rRNA. These two sets were further reduced by limiting to X-ray-determined structures with resolution < 3.5 Å and clustered by CD-HIT-EST (Fu *et al.*, 2012; Li and Godzik, 2006) with sequence identity cut off of 0.8.

For those RNAs mapped to known Rfam families, we required the minimal aligned length to be 80% of the RNA length and the aligned length of the Rfam covariance model to be greater than 50%. If one RNA was aligned to two or more Rfam families, the family with the highest score (or the lowest E-value) was taken as the correctly aligned family. Finally, for each family, we selected one RNA. Dataset-1 contained 43 structured, non-redundant RNAs as shown in Supplementary Table S1.

For those RNAs that were not mapped to Rfam families, we required the minimal number of base pairs to be 10, which is the lowest number of base pairs in the Rfam set. Dataset 2 set contained 117 non-redundant RNAs as shown in Supplementary Table S1. Further, Supplementary Table S2 shows different types of base-pairs, median and maximum sequence length for dataset-1,2.

3.2 PseudoBase++ dataset

The PseudoBase++ database (Taufer *et al.*, 2008) collected over 300 records of pseudoknot RNA secondary structures. We downloaded 304 RNA sequences from PseudoBase++. After excluding the sequences with large gaps or ambiguous bases a total of 274 RNA sequences were obtained. Further to avoid any potential bias, we removed sequences with more than 80% identity with SPOT-RNA training data. The final PseudoBase++ dataset (dataset-3), consists of 31 RNAs with the number of effective homologous sequences $N_{eff}/L > 0.2$ as shown in Supplementary Table S1.

3.3 RNA structure atlas dataset

The RNA Structure Atlas organized all RNA-containing 3D structures from PDB into non-redundant classes and selects high-quality representative structure from each class. We extracted 366 single-chain RNAs from RNA Structure Atlas (Version 3.126) at 4.0 Å resolution with sequence lengths ranging from 32 to 500. By mapping the sequences to existing Rfam families, it is found that 77 sequences are tRNA (RFAM ID: RF00005), eight times more than the second largest group (9 sequences, Purin riboswitch, RFAM ID: RF00167). We excluded tRNA to avoid test bias, resulting in 266 RNAs in this set. Furthermore, these 266 RNAs were filtered against the SPOT-RNA training data at 80% sequence identity cut-off and with N_{eff}/L value cut-off of 0.2. The final Atlas dataset (dataset-4), consists of 133 RNAs as shown in Supplementary Table S1. Supplementary Table S2 also shows different types of base-pairs, median and maximum sequence length for dataset-3,4.

4 Data processing

4.1 Secondary structure annotation

The terms for characterizing RNA structures differ greatly in the literature. Here we followed the definition of the terms as in bpRNA (Danaee *et al.*, 2018). Briefly, a 'canonical' base pair is a base pair with the type of AU or GC and 'Wobble' base pair with type of GU. A 'stem' is defined as a region of more than two uninterrupted base pairs with no intervening loops or bulges. An 'isolated canonical base pair' is defined as a canonical base pair without stacking interaction or not belonging to a stem. A 'Pseudoknot' exists when two non-nested base pairs (i, j) and (a, b) satisfy $(i < a < j < b)$. The 'Pseudoknot base pairs' in a pseudoknotted RNA are those base pairs that require the least to remove in order to yield a pseudoknot-free secondary structure.

We used X3DNA-DSSR (Lu and Olson, 2003) to annotate the secondary structure from the PDB structure in dataset 1, 2 and 4. The secondary structure for dataset-3 was directly downloaded from PseudoBase++ webserver. For secondary structures in dataset $1-4$, we used the bpRNA script to annotate the RNA structural elements (Danaee *et al.*, 2018).

4.2 Multiple sequence alignment annotation

The performance of the RNAcmap contact prediction is determined by the quality of the homologous sequence profile. We calculated the Number of EFFective (N_{eff}) homologous sequences for each multiple sequence alignment using GREMLIN. N_{eff} is defined as the sum of weights after down-weighting each sequence by the number of neighbors above a pairwise sequence similarity cutoff of 0.8. The MSA depth is defined as N_{eff}/L, where L is the sequence length of the RNA.

5 Performance measures

5.1 Base pairs

The performance of a single RNA was evaluated by the sensitivity $(\mathrm{SN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}))$, precision $(\mathrm{PR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}))$ and Matthews Correlation Coefficient (MCC) using top $L/6$, $L/4$, $L/2$

and L predicted pairs, where L is the length of RNA sequence and MCC is given by

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Here, TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively. Only non-local pairs were evaluated ($|i-j| > 4$, i and j are the sequence positional indices of the nucleotides).

Here, we defined non-hydrogen-bonded tertiary contacts if (i) the nearest-heavy atom distance between two nucleotides is less than 8 or 12 Å and (ii) these two nucleotides are not adjacent to the existing base pairs. This definition follows the work by De Leonardis *et al.* (2015). Tertiary contacts were evaluated after removing base pairs and two nucleotides neighboring to the base pairs. The overall performance was evaluated by MCC, sensitivity and precision.

5.2 RNA topology

It is frequently observed that only a few base pairs within a stem have detectable evolutionary coupling signals. Therefore, it is necessary to evaluate the prediction of stems beside the prediction of base-pair contacts. This is because that for the RNA modeling problem in a real-world scenario, capturing all stems of an RNA is more valuable than capturing all base pairs of a stem. To evaluate the performance on the stem level, we define that a stem is correctly predicted if one or more base pairs within the stem are correctly predicted. The overall performance was evaluated by MCC, sensitivity and precision.

6 Results

6.1 The RNAcmap pipeline

6.1.1 Comparison of covariance methods

We first examined how different covariance scores would impact the outcome of the contact prediction step. For 43 RNAs in dataset-1, we generated MSA profiles for the non-redundant PDB set and applied GREMLIN, mfDCA_apc, plmc and R-scape to calculate coupling scores for all possible base pairs. In addition to DCA predictors, we also included alignment based folding method RNAalifold for comparison. Figure 2 compares base-pair prediction evaluated by MCC, precision and sensitivity respectively. Results for top L/6, L/4, top L/2 and top L predictions are presented. As expected, increasing the number of predictions from top L/6 to L leads to an increase in sensitivity but a decrease in precision. Except R-scape, top L/4 predictions reached the highest MCC for all evolutionary coupling methods, suggesting that using top L/4 predictions have the optimal balance of sensitivity and precision.

As shown in Figure 2A, GREMLIN has a comparable performance with RNAalifold based on the average MCC at top L/4 with not much statistical significant performance difference (P-value= 0.03, the paired t-test). This is followed by the comparable performance of plmc and R-scape, then by mfDCA. However, the performance improvement of GREMLIN is statistical significant over mfDCA (P-value= 1×10^{-7}), plmc (P-value= 1×10^{-5}) and R-scape (P-value= 1×10^{-3}) when evaluated based on MCC of top L/4 using paired t-test. Because of the statistical significant difference among GREMLIN and other predictors (except RNAalifold), GREMLIN will be used as the default for all subsequent analysis. We preferred, GREMLIN over RNAalifold because GREMLIN can predicts non-canonical and pseudoknots base-pairs as we shall see later. Moreover, RNAalifold is a folding-based algorithm, not a method that extracts co-evolutional information.

6.2 Comparison of secondary structure predictors

We compared the effect of using RNAfold and SPOT-RNA in the homology search step. To make a fair comparison, we excluded those RNAs in dataset 1 and 2 with sequence similarity greater than 80% to any of RNAs in the SPOT-RNA training set. This sequence-identity cut off was the lowest cutoff allowed by CD-HIT-EST (Fu *et al.*, 2012; Li and Godzik, 2006) and employed previously for separating training and independent test sets (Guruge *et al.*, 2018; Singh *et al.*, 2019; 2021b; Yang *et al.*, 2017). This leads to a total of 77 RNAs as a combined test set (see Supplementary Table S1, dataset 1 + 2).

Figure 3 compares the performance in MCC for the top L/4 predictions by RNAcmap (SPOT-RNA) and by RNAcmap (RNAfold), respectively. The former has 29 RNAs with higher MCC, compared to 20 RNAs with higher MCC by the latter. The mean MCC is 0.30 for RNAcmap (SPOT-RNA) and 0.28 for RNAcmap (RNAfold). The performance difference is not statistically significant with P-value 0.137 obtained through paired t-test for this dataset.

6.3 Comparison between RFAM curated and RNAcmap (RNAfold) generated multiple sequence alignment

Using RNAs in dataset 1, it is possible to compare MSA generated by RNAcmap (RNAfold) and MSA supplied by RFAM. Figure 4 compares the base-pair prediction performance using MSA from RNAcmap (RNAfold) and RFAM, respectively. For a reference, the first-round MSA-1 based on the BLAST-N search is also shown. BLAST-N-based alignment provides poor prediction with MCC close to zero. Manually curated MSA from Rfam improves over RNAcmap (RNAfold) with 0.01 to 0.1 higher MCC at all prediction cutoffs. The improvement is observed for both sensitivity and precision. Rfam-alignment improves over RNAcmap (RNAfold) statistically significant for top L/6 predictions (P-value= 0.0007, paired t-test on MCC of 43 RNAs), top L/4 predictions (P-value= 0.0008)

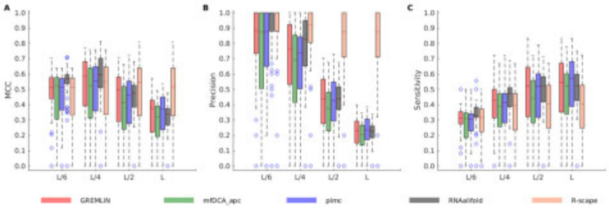

Fig. 2. Boxplot of MCC (**A**), Precision (**B**) and Sensitivity (**C**) of predicted base pairs by RNAcmap (RNAfold) based on three evolutionary coupling methods GREMLIN, mfDCA_apc, plmc, R-scape and RNAalifold, respectively, for 43 RNAs in the Rfam set. The distribution is shown in terms of median, 25th and 75th percentile with outlier shown by dots

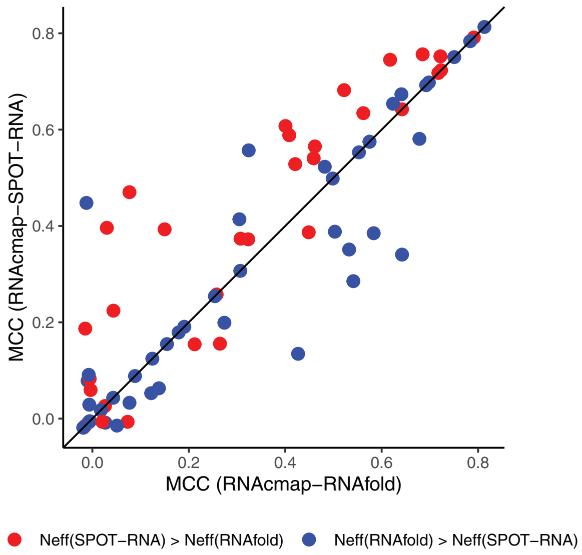

Fig. 3. MCC by RNAcmap with SPOT-RNA versus RNAcmap with RNAfold for 77 RNAs in a combined test set. RNAs with higher N_{eff} by RNAcmap (SPOT-RNA) than RNAcmap (RNAfold) are shown in red (in blue, otherwise)

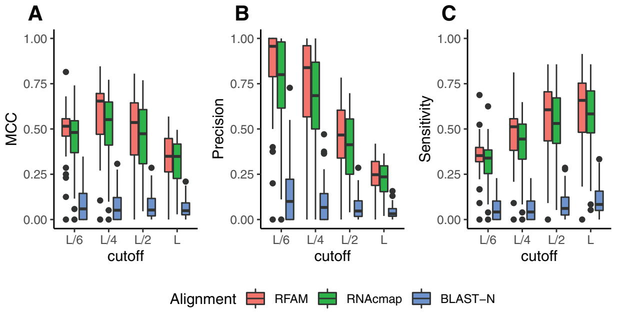

Fig. 4. Boxplot of MCC (**A**), Precision (**B**) and Sensitivity (**C**) by Rfam-supplied alignment in comparison to RNAcmap (RNAfold) and BLAST-N for 43 RNAs in the Rfam set. The distribution is shown in terms of median, 25th and 75th percentile with outlier shown by dots. All employed GREMLIN for base-pair prediction

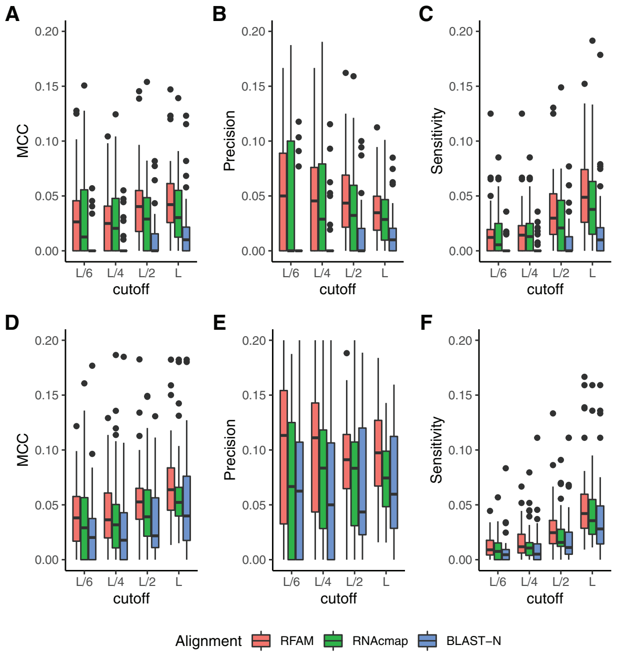

Fig. 5. Boxplot of MCC (**A** and **D**), Precision (**B** and **E**) and Sensitivity (**C** and **F**) for tertiary contact prediction using alignment from Rfam, RNAcmap (RNAfold) and BLAST-N for the Rfam dataset (43 RNAs). Tertiary contact is defined based on the distance of nearest-heavy atoms between two nucleotides, <8 Å for A, B and C, < 12 Å for D, E and F, respectively. The distribution is shown in terms of median, 25th and 75th percentile with outlier shown by dots

and top L/2 predictions (P-value= 0.004) but not top L predictions (P-value= 0.013). This result confirmed that using top L/4 predictions achieved a balanced performance over precision and sensitivity for RNA on both RNAcmap and RFAM MSA.

Figure 5 further examines the ability to predict tertiary contacts (based on a cutoff < 8 Å and < 12 Å, respectively) by comparing the results of different alignments. Performance for all methods is poor with < 0.05 in median precision and < 0.02 in median sensitivity for all top $L/6$, $L/4$, $L/2$ and L predictions. Rfam-based alignment and RNAcmap (RNAfold) are significantly better than Blast-N (with P-value < 0.0002 for all cases).

6.4 Beyond Rfam families

Although Rfam-based alignment has a slight edge in performance than RNAcmap, one advantage of RNAcmap (RNAfold) is that it can predict contacts for sequences that are not in the Rfam collection. Using 160 RNAs in the PDB dataset, Figure 6 shows MCC of base pair prediction as a function of the MSA depth (N_{eff}/L). While the Rfam-based alignment improves over RNAcmap (RNAfold) for the Rfam set, the performance of RNAcmap (RNAfold) for 117 RNAs in the non-Rfam set is nearly the same as that for 43 RNAs in the Rfam set. All showed a trend of improved prediction with increased MSA depth (N_{eff}/L). Particularly, reasonably accurate predictions (MCC $>$ 0.5) are made for MSA depth $>$ 1 for 21 of 27 RNAs (78%) using Rfam alignment and 31 of 39 RNAs using RNAcmap alignment. Two outliers with high N_{eff}/L and low MCC in Figure 6 are both resulted from poorly predicted secondary structures (Red : 6ASO-I, Blue: 4QJD-B), which led to incorrect homologous sequences.

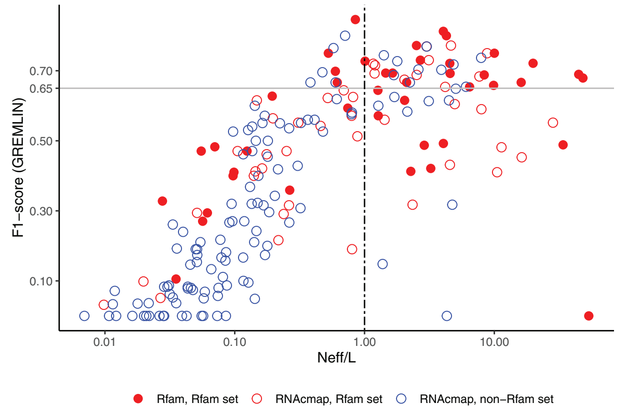

Fig. 6. F1 of base-pair prediction as a function of alignment effective size normalized by length (N_{eff}/L). GREMLIN was used for evolutionary coupling analysis using top L/4 pair predictions by Rfam-supplied alignment (filled red circle) and RNAcmap (RNAfold, red circle) for the Rfam set, and RNAcmap (RNAfold, blue) for the non-Rfam set

Figure 7 shows the results for 18 RNAs in the combined test set (dataset-1 and dataset-2) with $N_{eff}/L > 1$ as we considered that the evolutionary information is not reliable for $N_{eff}/L < 1$ as shown in Figure 6. We evaluated the performance on the base-pair level and on the stem-level. On the base-pair level, we examined different types of base pairs including canonical and Wobble base pairs in helical regions, in non-helical regions (unstacked, isolated single base pairs), non-canonical base pairs and nested base pairs. SPOT-RNA and RNAcmap (SPOT-RNA) can correctly predict more canonical and Wobble base pairs in a helical region as compare to other predictors. On the other hand, RNAcmap (RNAfold) significantly improves over RNAfold in predicting non-canonical base pairs and base pairs in pseudoknots. SPOT-RNA is slightly better in predicting non-canonical base pairs and much better in predicting base pairs in pseudoknots than RNAcmap (SPOT-RNA). This is because RNAfold was not built for predicting pseudoknots or non-canonical base pairs whereas SPOT-RNA, a deep learning technique, was trained for predicting any base pairs including pseudoknots and non-canonical base pairs. Moreover, pseudoknots and non-canonical base-pairs are not employed by INFERNAL for building covariance models. What is more revealing is the evaluation at the stem level (Fig. 7 last columns). RNAcmap (SPOT-RNA) and RNAcmap (RNAfold) achieved higher F1-score than SPOT-RNA (or RNAfold).

6.5 Performance on pseudoknotted RNAs

Pseudoknot structures in RNA are known difficult to model. Supplementary Figure S1 shows the basepair prediction F1-score as a function of different types of base-pairs for dataset-3. Only 9 of 266 RNAs has MSA depth greater than 1, however, we noticed that 31 RNAs in this dataset with MSA depth $>$ 0.2 have reasonably accurate predictions (MCC $>$ 0.5).

Supplementary Figure S1 shows the F1-score of prediction on different base-pairs and at stem levels of these 31 RNAs. In the canonical and Wobble pair category, RNAcmap (SPOT-RNA) predicts fewer pairs than SPOT-RNA and RNAalifold but better than the R-scape. In the pseudoknot category, RNAcmap (SPOT-RNA)

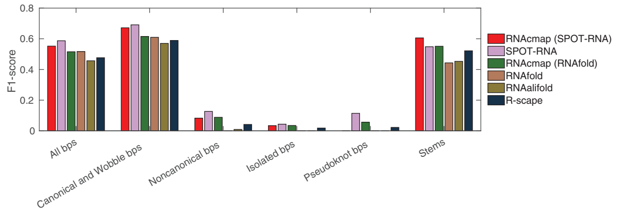

Fig. 7. F1-score of predicted base pairs reported according to all base-pairs (bps), canonical and Wobble bps, non-canonical bps, isolated bps, pseudoknot bps and stems. The metrics are evaluated on 18 RNAs with 'deep' RNAcmap alignment ($N_{eff}/L > 1$)

performs better than the RNAalifold and R-scape but underperforms as compare to SPOT-RNA as pseudoknot information from SPOT-RNA was not utilzed by the INFERNAL tool. We do not make comparison on non-canonical base-pairs because there were only 4 non-canonical base-pairs in 31 RNAs which was not statistically meaningful comparison.

6.6 Performance on RNA atlas datasets

The RNA structure atlas dataset (dataset-4) provides a representative RNA structure set for testing. We predicted contacts using RNAcmap (SPOT-RNA) pipeline and compared results with R-scape and RNAalifold using the same RNAcmap (SPOT-RNA) generated MSA as input. 133 out of 288 RNAs have MSA depth > 0.2 and used for the comparison.

Supplementary Figure S2 shows F1-score for these 133 RNAs for different base-pairs and stem levels. The same trend is observed as RNAcmap predicts less canonical and Wobble pairs than SPOT-RNA and RNAalifold while capturing more pairs than the R-scape. In non-canonical and pseudoknot category, RNAcmap (SPOT-RNA) performs better than RNAalifold but comparable to R-scape.

6.7 Computational efficiency

RNAcmap performs computation demanding database search (cmsearch and BLAST-N) as well as covariance analysis using DCA tools. In our test on the RNA atlas datasets, a median of 30 CPU hours for each sequence is required. Most jobs finished with 17 to 51 CPU hours (25–75 percentile). Therefore, we recommend using RNAcmap with 20 GB RAM and multicore support.

7 Discussion

In this article, we have established a fully automatic pipeline that can predict contact maps directly from any given RNA sequences by homology search and evolutionary coupling analysis. The performance of RNAcmap is comparable to that from manually curated Rfam alignments. More importantly, the performance is robust for those sequences not belonging to Rfam families, pseudoknot RNAs and non-redundant RNA sets. Thus, RNAcmap is expected to be useful to generate structural restraints for RNA secondary and tertiary structure prediction, as demonstrated previously (De Leonardis *et al.*, 2015; Wang *et al.*, 2017a; Weinreb *et al.*, 2016; Zhang *et al.*, 2020).

It is found that the performance of RNAcmap is less dependent on the tools for evolutionary coupling analysis. The difference between GREMLIN, mfDCA_apc and plmc is small (Fig. 2). This result is consistent with an independent study (Pucci *et al.*, 2019). However, the performance of RNAcmap is more strongly dependent on the secondary structure predictor (Fig. 3). SPOT-RNA, that has more accurate secondary structure prediction, improves over RNAfold in generating alignments that yielded improved contact prediction. In particular, more stem regions were captured by using RNAcmap (SPOT-RNA) (Fig. 7), indicating more accurate topological connections in base pairing patterns. A simple meta predictor RNAcmap (SPOT-RNA/RNAfold) was established by using the secondary structure predictor that will yield a higher number of effective homologous sequences (N_{eff}). This meta predictor further improves over RNAcmap (RNAfold). It is not entirely surprising as RNAfold (a folding-based algorithm) and SPOT-RNA (a deep-learning-based method) are likely complementary to each other. It should be noted that because the covariance model by INFERNAL cannot use the pseudoknot and non-canonical base pair information from input secondary structure, therefore, improvement for non-canonical base-pairs and pseudoknots are independent of the input secondary structure predictor employed.

Contact map results for RNAs are different from those of proteins. For homologous sequence alignment, our results showed that sequence-only similarity search (BLAST-N) missed many homologous sequences with low sequence identity, resulting in poor prediction in the downstream covariance analysis. Using a predicted secondary structure in the RNAcmap greatly expanded the coverage of homologous sequences, resulting in a much more accurate prediction. This is different in the case of protein homologous search, where sequence-only similarity is sufficient to capture most homologous sequences (Remmert *et al.*, 2011). Moreover, the contact maps for RNAs are dominated by the hydrogen-bonded base pairs. The accuracy for predicting distance-based tertiary contacts is only marginally better than random (Fig. 5). This result is consistent with previous studies (De Leonardis *et al.*, 2015; Pucci *et al.*, 2019).

We also experimented with INFERNAL E-value cut-off for MSA-2 generation to see if different E-value cut-offs yields better results. As shown in Supplementary Figure S3, with increase in E-value performance of RNAcmap consistently improved on 27 RNAs from dataset-2 with $N_{eff} > 1$ for lowest E-value (1×10^{-4}). However, the improvement by increasing the E-value cutoff is at the cost of significant increase in the computing time.

Not all base pairs are of equal importance when inferring the RNA structure. As showing in the comparison between RNAcmap and secondary structure predictors (SPOT-RNA, RNAfold), RNAalifold and R-scape, RNAcmap-predicted base pairs are more enriched with isolated, pseudoknotted and non-canonical base pairs, which bring richer information for the overall topology of the RNA. Even for helical stem regions, RNAcmap predicts more stems than other methods, although the average number of predicted canonical and Wobble base pairs within a stem is less than that of secondary-structure predictors. This is because evolutionary coupling can only capture the strongest signals that show a marked difference in structural and functional stabilities between deleterious single and rescuing double mutations. In other words, the results from evolutionary coupling analysis offer a topology frame that can be further improved by a post-processing method. Indeed, Zhe *et al.* showed that a simple Monte-Carlo simulated annealing can recover nearly all base pairs of two ribozymes using pairing probabilities from mutational coupling analysis as a part of the energy function for folding secondary structure (Zhang *et al.*, 2020).

Acknowledgements

The authors gratefully acknowledge the use of the High Performance Computing Cluster Gowonda to complete this research, and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). They also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Funding

This work was supported by Australia Research Council [DP210101875 and DP180102060 to Y.Z. and K.P.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bernhart,S.H. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Burger,L. and van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633.

Cheng,J. *et al.* (2019) Estimation of model accuracy in CASP13. *Proteins Struct. Funct. Bioinf.*, **87**, 1361–1377.

Cruz,J.A. *et al.* (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.

Danaee,P. *et al.* (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res.*, **46**, 5381–5394.

De Leonardis,E. *et al.* (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, **43**, 10444–10455.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.

Freyhult,E.K. *et al.* (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, **14**, 699–712.

Guruge,I. *et al.* (2018) B-factor profile prediction for RNA flexibility using support vector machines. *J. Comput. Chem.*, **39**, 407–411.

Hanson,J. *et al.* (2019) Getting to know your neighbor: protein structure prediction comes of age with contextual machine learning. *J. Comput. Biol.*, **27**, 796–814.

Hanumanthappa,A.K. *et al.* (2021) Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network. *Bioinformatics*, **36**, 5169–5176.

Huang,Z. *et al.* (2008) Fast and accurate search for non-coding RNA pseudoknot structures in genomes. *Bioinformatics*, **24**, 2281–2287.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Jones,D.T. *et al.* (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999–1006.

Kalvari,I. *et al.* (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.

0.02w?>Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA*, **110**, 15674–15679.

Kinch,L.N. *et al.* (2016) Evaluation of free modeling targets in CASP11 and ROLL. *Proteins Struct. Funct. Bioinf.*, **84**, 51–66.

Lapedes,A.S. *et al.* (1999) Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lect. Notes Monogr. Ser.*, **33**, 236–256.

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

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

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.

Menzel,P. *et al.* (2009) The tedious task of finding homologous noncoding RNA genes. *RNA*, **15**, 2075–2082.

Miao,Z. *et al.* (2015) RNA-puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, **21**, 1066–1084.

Miao,Z. *et al.* (2017) RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, **23**, 655–672.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA*, **108**, E1293–E1301.

Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Petrov,A.I. *et al.* (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, **19**, 1327–1340.

Petrov,A.I. *et al.*; RNAcentral Consortium. (2015) RNAcentral: An international database of ncRNA sequences. *Nucleic Acids Res.*, **43**, D123–D129.

Pucci, F.et al. (2020) Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. RNA, **26**, 794–802.

Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rivas,E. *et al.* (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.

Schaarschmidt,J. *et al.* (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins Struct. Funct. Bioinf.*, **86**, 51–66.

Singh,J. *et al.* (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.*, **10**, 1–13.

Singh,J. *et al.* (2021a) Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics*, **37**, 2589–2600.

Singh,J. *et al.* (2021b) RNA backbone torsion and pseudotorsion angle prediction using dilated convolutional neural networks. J. Chem. Inf. Model., DOI: 10.1021/acs.jcim.1c00153.

Sun,S. *et al.* (2019) Enhanced prediction of RNA solvent accessibility with long short-term memory neural networks and improved sequence profiles. *Bioinformatics*, **35**, 1686–1691.

Taufer,M. *et al.* (2008) PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Res.*, **37**, D127–D135.

The RNAcentral Consortium. (2018) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.*, **47**, D221–D229.

Vasavada,M. *et al.* (2015) Genome-wide search for pseudoknotted noncoding RNA: a comparative study. In: Elloumi, M. et al. (eds.) *Pattern Recognition in Computational Molecular Biology*. John Wiley & Sons, Ltd., pp. 155–164.

Wang,J. *et al.* (2017a) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.*, **45**, 6299–6309.

Wang,S. *et al.* (2017b) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Comput. Biol.*, **13**, e1005324.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci.*, **106**, 67–72.

Weinreb,C. *et al.* (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.

Yang,Y. *et al.* (2017) Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. *RNA*, **23**, 14–22.

Zhang,Z. *et al.* (2020) Accurate inference of the full base-pairing structure of RNA by deep mutational scanning and covariation-induced deviation of activity. *Nucleic Acids Res.*, **48**, 1451–1465.