Identification and classification of common RNA structure motifs

Martin A. Smith[1,2,*#], Stefan E. Seemann[2,3,*], XiuCheng Quek[1,2] & John S. Mattick [2]

May 4, 2017

[1] *Garvan Institute of Medical Research, 384 Victoria Street, Sydney 2010, Australia*
[2] *St Vincents Clinical School, UNSW Australia*
[3] *University of Copenhagen, Groennegaardsvej 3, Frederiksberg, Denmark*
[*] *Contributed equally*
[#] *Corresponding author*
m.smith[at]garvan.org.au

Abstract

The abundance and diversity of processed transcripts in eukaryotic genomes poses a challenge for the systemic classification of their biological functions. Sparse sequence conservation in non-coding sequences and the unreliable nature of RNA structure prediction algorithms further exacerbate this conundrum. Here, we present a unified computational approach for the unsupervised discovery and classification of homologous RNA structure motifs from a set of sequences of interest. At its core lies DotAligner, a heuristic pairwise RNA structure alignment algorithm that considers both sequence similarity and the ensemble of sub-optimal RNA base pairings. Our approach outperforms other state of the art tools at classifying known RNA structure families, both in time and accuracy. When combined to density-based clustering using an empirically measured threshold, this method identifies both known and novel RNA structure motifs from ENCODE immuno-precipitation data for 44 proteins, further expanding the lexicon of functional transcriptomic motifs.

1 Introduction

The human genome is pervasively transcribed into RNA with less than 2% encoding protein sequences. As genome technologies progress, an ever increasing amount of non-protein coding RNAs (ncRNA) are being discovered. Once class of ncRNAs–long noncoding RNAs (lncRNAs)–are of particular interest for functional genome annotation given the large expanses they encompass. So far, only a relatively small quantity of lncRNAs have been functionally characterised, with regulation of gene expression and epigenetic states recurring as common biological functions (Morris and Mattick, 2014; Engreitz et al., 2016). Understanding the molecular mechanisms underlying the biological functions of lncRNAs–and how they are disrupted in disease–is required to improve the functional annotation of the human genome. This is of particular importance in the era of personalised genomics, as over 80% of trait-associated single nucleotide polymorphisms occur in non-coding regions (Hindorff et al., 2009; Ritchie et al., 2014).

The higher-order structure of RNA molecules is an essential feature of non-coding RNAs that can be used for their classification and the inference of their biological function. Most small ncRNAs have well characterised secondary and tertiary structures, as evidenced by RFAM, the largest collection of curated RNA families (2,588 families as of version 12.2 (Nawrocki et al., 2015)). However, determining the structural features of long ncRNAs (lncRNAs) is a more complex problem given their size and, in general, faster evolutionary turnover. These challenges have raised doubts concerning the prevalence of functional structural motifs in lncRNAs (Eddy, 2014; Rivas et al., 2016), despite evolutionary and biochemical support of conserved base-pairing interactions (Smith et al., 2013; Spitale et al., 2015; Lu

et al., 2016).

We, and others, hypothesise that lncRNAs act as scaffolds for the recruitment of proteins and assembly of ribonucleoproteins (RNPs), mediated by the presence of modular RNA structures, akin to the domain organisation of proteins (Zappulla and Cech, 2006; Hogg and Collins, 2008; Rinn and Chang, 2012; Mercer and Mattick, 2013; Smith et al., 2013; Chujo et al., 2016; Blythe et al., 2016). Protein-interacting regions of lncRNAs are likely to contain a combination of sequence and structure motifs that confer binding specificity. The higher-order structural features of protein-binding RNAs are subjected to different evolutionary dynamics than sequence constrained regions, where selective pressures to preserve higher-order structures facilitate compensatory and covariating mutations (Pang et al., 2006; Smith et al., 2013; Johnsson et al., 2014).

It has been shown that if sequence similarity falls below 60%–the 'twilight zone' of multiple sequence alignment–sequence-centric approaches for the identification of RNA structures perform poorly (Gardner et al., 2005).

In addition, competing structures and suboptimal structures may support or even drive the functionality of an RNA domain. Hence, methods are needed that find structural similarity independent from sequence conservation and freed from one single optimal RNA secondary structure.

For clustering of RNA domains a dissimilarity measurement of all pairs of query structures is needed. The dissimilarity is described through a pairwise weighted string alignment with arbitrary pairwise dependencies (for base pairings). The Needleman-Wunsch algorithm solves the maximum weight string alignment problem by dynamic programming in $O(N^2)$ by preserving the sequence order and maximizing the similarity. The consideration of pairs of nucleotides in each sequence that form intra-molecular interactions extends the problem to pairwise dependencies among positions in each string. This problem variant is MAX-SNP-hard. However, the problem can be attacked by intelligent heuristics that avoid the examination of all possible alignment states.

Simultaneous alignment and folding (Sankoff, 1985) is the acknowledged gold standard to predict the consensus structure and alignment of a set of related RNA sequences. Because the Sankoff algorithm is practically not applicable, the pre-calculation of the structure ensemble of each sequence, e.g. basepair probabilities in thermodynamically equilibrated RNA structure ensembles (McCaskill, 1990), is used by different methods to speed up the calculation of structure-based alignments. include a review of structural alignments . In particualr, tThe programs `pmcomp` for pairwise and `pmmulti` for multiple alignments (Hofacker et al., 2004), as well as `LocaRNA` (Will et al., 2007) score the alignment based on the notion of a common secondary structure. Despite of the usage of the basepair probability matrices these methods extract the maximum-weight common secondary structure but do not explicitely consider suboptimal structures in the alignment. The pairwise alignment of basepair probability matrices (dot plots) has been first introduced by `CARNA` (Palù et al., 2010; Sorescu et al., 2012). `CARNA` finds iteratively better alignments with an effective constraint programming technique using a branch and bound scheme (propagator).

Beside of `LocaRNA` and a method based on directed acyclic graph kernels (Sato et al., 2008), the alignment-free approach `ClustGraph` (Heyne et al., 2012) has been used to cluster RNA structure in common domains. Here, we propose an alternative heuristic for the pairwise weighted string alignment with arbitrary pairwise dependencies that can deliver dissimilarity scores of dot plots in time close to an Needleman-Wunsch alignment which makes the approach applicable for clustering of large numbers of putative RNA domains. We combine this algorithm with cluster analysis for applications with noise to impartially and reliably extract homologous structural RNA motifs.

2 Results

Ensemble-guided pairwise RNA structure alignment

A structural alignment of two RNAs with similar higher-order structures involves aligning both sequence similarities and homologous paired, helical regions of the molecules. This often produces results that can differ significantly from the optimal sequence alignment since the conservation of sequence may be under less selective pressure than conservation of structural features. However, the relative importance of sequence versus structural integrity varies across all ncRNAs; micro RNAs have strong sequence and structure constraints, whereas tRNAs are mostly constrained with respect to their tertiary structures, with mostly the anti-codon being constrained for sequence compostion. It is therefore essential to accommodate a dynamic range of sequence-structure weight functions when comparing two RNA sequences impartially.

Consequently, we developed a new heuristic method of aligning two RNA basepair probability matrices (dot plots) called `DotAligner`. It leverages the diversity of suboptimal solutions from a partition function of RNA secondary structure predictions to identify an optimal sequence-structure alignment of two RNAs. `DotAligner` thus overcomes the limitations of comparing unique RNA secondary structures (such as minimum free energy predictions) to yield an optimal sequence alignment that considers mutual basepair probabilities.

`DotAligner` employs the heuristics alignment-envelope, which imposes constraints to sub-optimal string alignments, and fold-envelope, which imposes constraints to pre-calculated base pairing probabilities, to build pairwise sequence-structure alignments from pre-calculated RNA dot plots.

A major criteria for the implementation was a fast running time to make `DotAligner` applicable for RNA structure clustering of large data sets. The alignment procedure consists of two steps:

Fast and accurate classification of RNA structures

We compared the quality of the alignments produced by `DotAligner` with the results of three state of the art RNA structure alignment algorithms: `CARNA`, `FOLDALIGN` and `LocaRNA` **(ADD REFS)** (**Figure 1**). In all three qualitative metrics we assessed, `DotAligner` performed the worst on average. Interestingly, many of the pairwise structure alignments produced SCI scores above those from the `BRAliBase 2.1` reference alignments. Unless the reference alignments are poorly annotated (a possibility given that RFAM entries may have been automatically generated based on similarity to a covariance model), this comparison exposes the possible tendency of global optimization algorithms to overestimate the amount of paired bases in consensus RNA structure predictions. In this regard, `DotAligner` stands apart as it reports this phenomenon less than the other surveyed tools (**Figure 1B**).

Despite its underwhelming performance in reproducing benchmark alignments, `DotAligner` excels elsewhere. The intended application of this algorithm is the identification and classification of RNA structural motifs from a large and diverse set of sequences of interest. Therefore, we evaluated the ability of `DotAligner` to distinguish distinct RNA structure topologies from a heterogeneous sample of known RNA structure families. We performed all versus all pairwise structure alignment of stochastically sampled RFAM sequences, which were selected with constraints on their sequence composition, highlighting any sequence-dependent bias (see **Methods** and **Supplementary Figure 2**). Despite the seemingly poor quality of its pairwise alignments, `DotAligner` reproduces the known classification of RFAM structures as well or better than other pairwise RNA structure alignment tools (**Figure 2A,B**).

The classification accuracy of `DotAligner` is most comparable to `CARNA`, another ensemble-based structural alignment algorithm, but the latter requires substantially more time to perform the comparisons

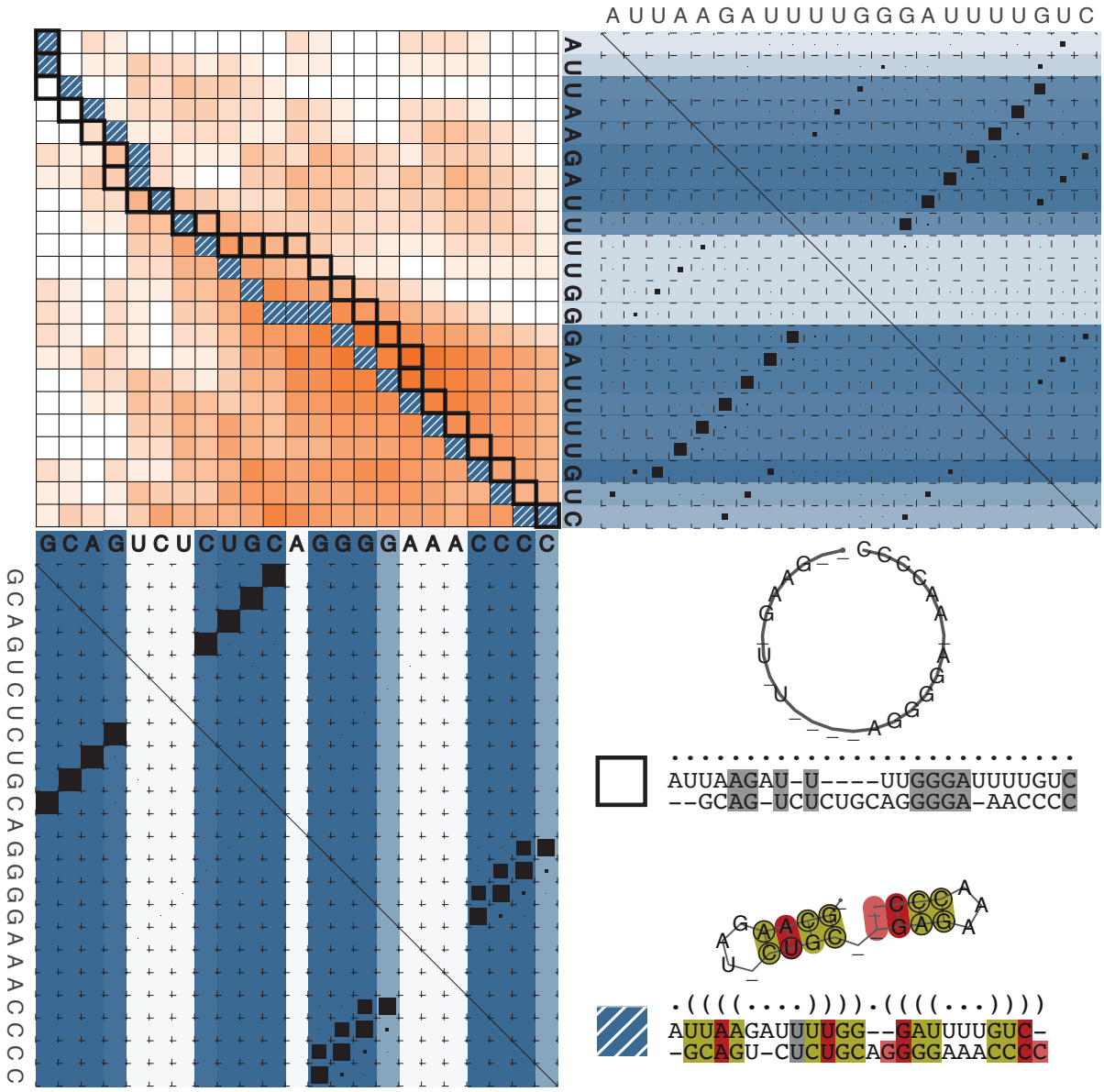

Figure 1:
Schematic of a pairwise alignment with DotAligner. A dynamic programming matrix is first filled first filled in based on sequence similarity (top left–color intensity indicates cumulative sequence similarity). A partition function over all pairwise sequence alignments is then calculated and interrogated for structural similarities by stochastic backtracking. Considering the ensemble of suboptimal secondary structures–represented by dot plots in the upper right and lower left quadrants–effectively warps the optimal sequence alignment path (top left, black outline) towards one that includes structural features (striped blue cells). In the bottom right, the optimal sequence alignment and associated consensus secondary structure is contrasted to that produced by DotAligner, exposing the common structural features hidden in the suboptimal base pairing ensemble of both sequences.

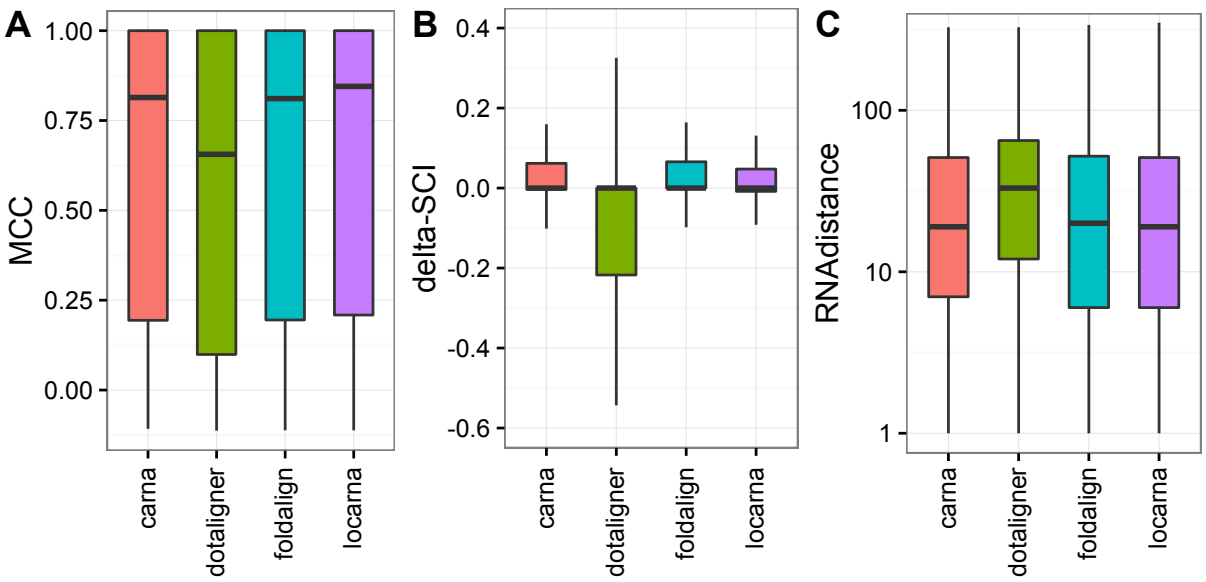

Figure 2: RNA structure alignment quality. (A) Matthews Correlation Coefficient; (B) Structural Conservation Index difference between computed and reference alignments; (C) Topological edit distance between the `RNAalifold` consensus of the computed alignment and the and reference consensus.

(**Figure 2C**). Indeed, `DotAligner` is faster that all surveyed algorithms, with the notable exception of the Needleman-Wunsch Algorithm (NWA), a sequence alignment algorithm that ignores secondary structure information. NWA performs surprisingly well at classifying RNA sequences, most likely due to the presence of stretches of homologous sequence within biologically related RNAs. On larger datasets derived from experimental data, NWA would likely suffer from higher false positive rates given the repetitive nature of higher-eukaryote genomes.

De novo identification of homologous RNA structures in noisy samples

Identifying protein-binding RNA motifs from eCLIP data

We applied this clustering methodology to recently published Enhanced Cross-Linking and ImmunoPrecipitation (eCLIP) sequencing data from the ENCODE consortium [**REF**], which contains eCLIP peaks associated to 44 RNA binding proteins (RBPs). We submitted the sequences from eLCIP peaks with strong, highly significant enrichment over background to the above-mentioned clustering strategy (see **Methods**).

[initial clustering data results].

If a structural motif is targeted by a RBP, it is highly probably that the associated CLIPseq peaks do not encompass the entire sequence that forms the structural motif, since reverse transcription will terminate at the covalent bond formed between the protein and the RNA. Consequently, large RNA structures bound by RBPs are difficult to model without additional structure probing data. This experimental approach also introduces a 3' bias in the sequence-specificity of RBP binding sites given the nature of 3' to 5' reverse transcription required for cDNA synthesis.

To address these caveats, we surveyed the potential to form locally-stable RNA secondary structures in the regions directly flanking 36,161 filtered eCLIP peaks (see **Methods**). XXX% overlap a predicted RNA secondary structure, compared to XXX random peaks.

Discussion

Given its relative speed and accuracy, `DotAligner` can be used to generate larger (dis)similarity matrices for cluster analysis than other pairwise structure alignment algorithms, or at least produce them with

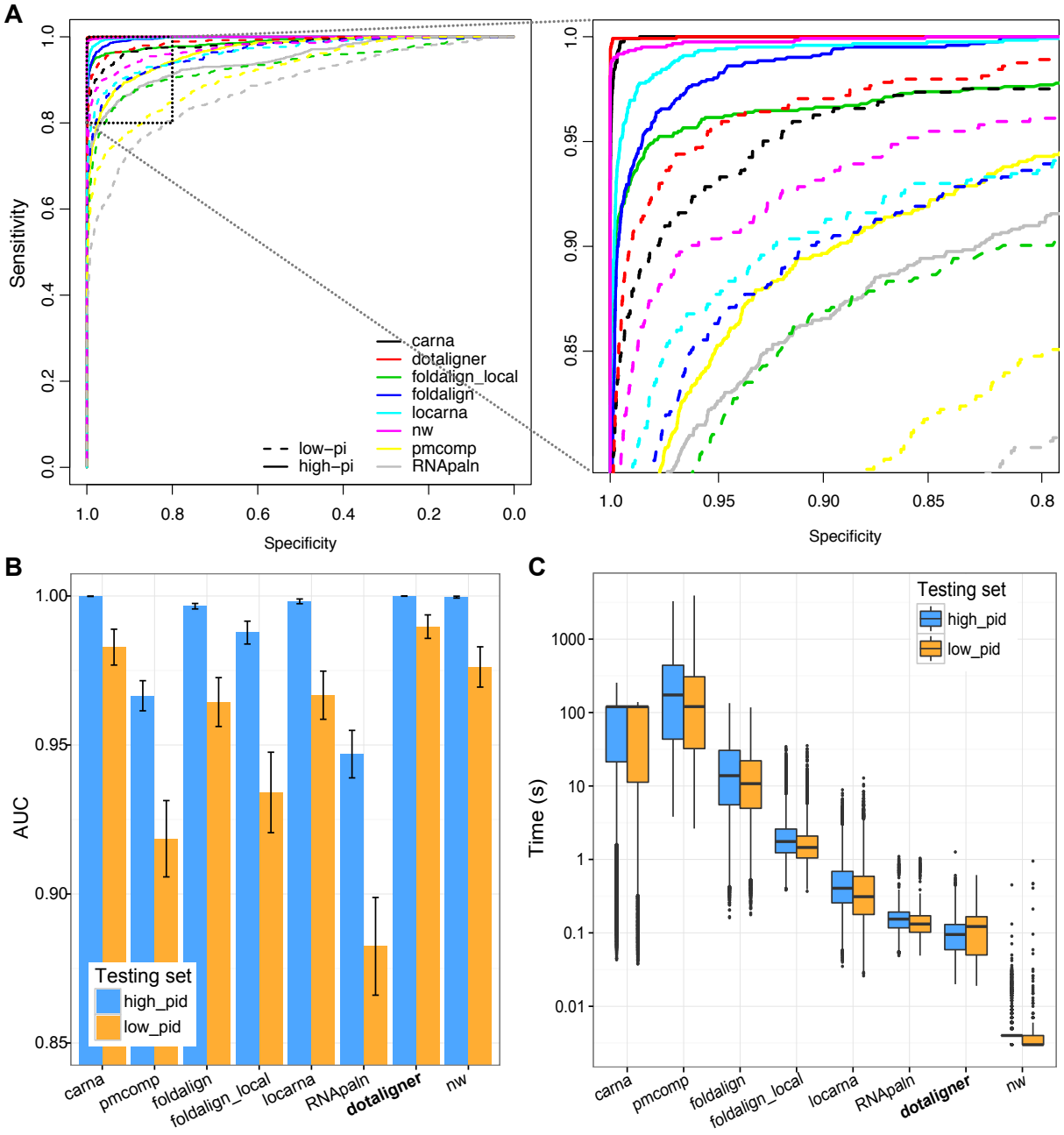

Figure 3: Classification of known RNA structures. (A) Receiving Operator Characteristic (ROC) curves measuring the classification accuracy by contrasting the computed similarity matrices of each algorithm to a binary classification matrix of RFAM sequences (1 if same family; 0 if different). High PID = 56-95% pairwise sequence identity; Low PID = 1-55%. (B) Area Under the Curve (AUC) of the ROC values with 95% confidence intervals; (C) Distribution of computation time (N.B. a fixed upper limit of 120 s was imposed for `CARNA`).

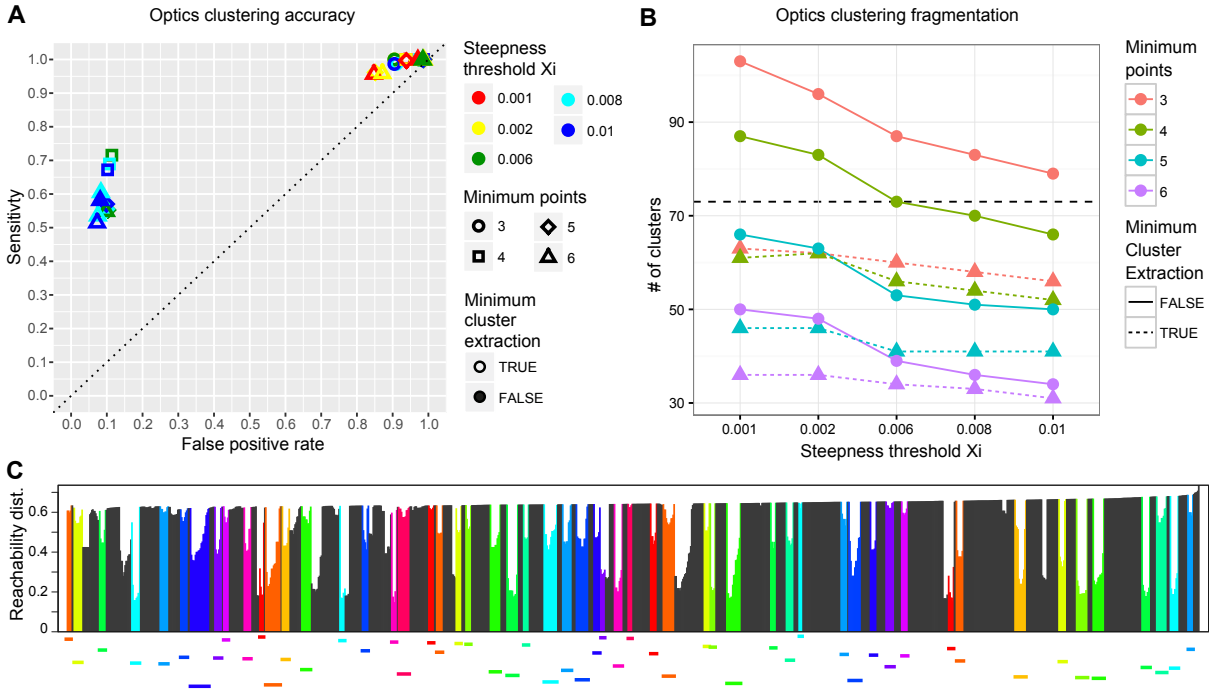

Figure 4: Density based clustering of known RNA structures.

reasonable computational power.

3 Materials and Methods

3.1 Benchmarking and parameter optimisation.

The `DotAligner` algorithm implements several theoretical parameters that first need to be tuned before being applied to biological sequence analysis. **All combinations** of core parameters were tested on the 8,976 pairwise RNA structure alignments curated in the `BRAliBase 2.1` reference dataset (Wilm et al., 2006). For each set of parameter combinations, the amount of alignments producing identical structural topologies to the reference alignment was determined using *RNAdistance*. The Structural Conservation Index (SCI), a robust measure of RNA structural alignment integrity, were also calculated for all resulting alignments. Baseline parameters were then selected via a product rank of the 2 aforementioned metrics (**supplementary data?**).

3.2 Classification of RNA secondary structure families.

This was achieved by sampling the entire collection of RFAM entries with published structures in a stochastic manner, while ensuring that all sampled sequences respected constraints on their sequence composition. Specifically, we extracted a high Pairwise Sequence Identity (PSI) and a low PSI set, where any two sequences from the same set present greater than or less than 55% PSI. The

References

Blythe, A. J., Fox, A. H., and Bond, C. S. (2016). The ins and outs of lncrna structure: How, why and what comes next? *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(1):46–58.

Chujo, T., Yamazaki, T., and Hirose, T. (2016). Architectural rnas (arcrnas): A class of long noncoding rnas that function as the scaffold of nuclear bodies. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(1):139–146.

Eddy, S. R. (2014). Computational analysis of conserved rna secondary structure in transcriptomes and genomes. *Annual review of biophysics*, 43:433–456.

Engreitz, J. M., Ollikainen, N., and Guttman, M. (2016). Long non-coding rnas: spatial amplifiers that control nuclear structure and gene expression. *Nature Reviews Molecular Cell Biology*.

Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural {RNAs}. *Nucleic Acids Res*, 33(8):2433–2439.

Heyne, S., Costa, F., Rose, D., and Backofen, R. (2012). GraphClust: alignment-free structural clustering of local {RNA} secondary structures. *Bioinformatics*, 28(12):i224–32.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227.

Hogg, J. R. and Collins, K. (2008). Structured non-coding rnas and the rnp renaissance. *Current opinion in chemical biology*, 12(6):684–689.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long non-coding rnas; sequence, structure, function. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1840(3):1063–1071.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., et al. (2016). Rna duplex map in living cells reveals higher-order transcriptome structure. *Cell*, 165(5):1267–1279.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for {RNA} secondary structure. *Biopolymers*, 29(6-7):1105–1119.

Mercer, T. R. and Mattick, J. S. (2013). Structure and function of long noncoding rnas in epigenetic regulation. *Nature structural & molecular biology*, 20(3):300–307.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory rna. *Nature Reviews Genetics*, 15(6):423–437.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2015). Rfam 12.0: updates to the rna families database. *Nucleic acids research*, 43(D1):D130–D137.

Palù, A., Möhl, M., and Will, S. (2010). A Propagator for Maximum Weight String Alignment with Arbitrary Pairwise Dependencies. In Cohen, D., editor, *Principles and Practice of Constraint Programming – CP 2010*, pages 167–175. Lecture no edition.

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding rnas: lack of conservation does not mean lack of function. *Trends in Genetics*, 22(1):1–5.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding rnas. *Annual review of biochemistry*, 81:145–166.

Table 1: This is a table with scientific results.

1	2	3	4	5
aaa	bbb	ccc	ddd	eee
aaaa	bbbb	cccc	dddd	eeee
aaaaa	bbbbb	ccccc	ddddd	eeeee
aaaaaa	bbbbbb	cccccc	dddddd	eeeeee
1.000	2.000	3.000	4.000	5.000

Ritchie, G. R., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–296.

Rivas, E., Clements, J., and Eddy, S. R. (2016). A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. *Nature Methods*.

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

Sato, K., Mituyama, T., Asai, K., and Sakakibara, Y. (2008). Directed acyclic graph kernels for structural {RNA} analysis. *BMC Bioinformatics*, 9:318.

Smith, M. A., Gesell, T., Stadler, P. F., and Mattick, J. S. (2013). Widespread purifying selection on rna structure in mammals. *Nucleic acids research*, page gkt596.

Sorescu, D. A., Möhl, M., Mann, M., Backofen, R., and Will, S. (2012). CARNA–alignment of RNA structure ensembles. *Nucleic acids research*, 40(Web Server issue):W49–53.

Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., et al. (2015). Structural imprints in vivo decode rna regulatory mechanisms. *Nature*, 519(7544):486–490.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring noncoding {RNA} families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65.

Wilm, A., Mainz, I., and Steger, G. (2006). An enhanced rna alignment benchmark for sequence alignment programs. *Algorithms for molecular biology*, 1(1):1.

Zappulla, D. and Cech, T. (2006). Rna as a flexible scaffold for proteins: yeast telomerase and beyond. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 71, pages 217–224. Cold Spring Harbor Laboratory Press.

4 Acknowledgments