METHOD

Identification and classification of common RNA structure motifs

Martin A Smith[1,2*†], Stefan E Seemann[3†], XiuCheng Queck[1,2] and John S Mattick[2]

*Correspondence:
m.smith[at]garvan.org.au
[1]Garvan Institute of Medical Research, 384 Victoria Street, NSW 2010 Sydney, Australia
Full list of author information is available at the end of the article
[†]Contributed equally

Abstract

The abundance and diversity of processed transcripts in eukaryotic genomes possesses 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.

Keywords: RNA structure clustering; Functions of RNA structures; RNA–protein binding; Regulation by non-coding RNAs

Background

The human genome is pervasively transcribed into RNA with less than 2% encoding protein sequences. As genomic technologies progress, an ever increasing amount of non-protein coding RNAs (ncRNA) are being discovered. Many ncRNAs lack sequence conservation or sequence motifs, in contrast to the open reading frame of protein-coding RNA. Instead, the higher-order structure of RNA molecules is an essential feature of ncRNAs that can be used for their classification and the inference of their biological function.

Long noncoding RNAs (lncRNAs) are of particular interest for functional genome annotation given the large expanses they encompass. 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. 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 [1, 2]. 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 [3, 4, 5, 6, 7, 8, 9]. Protein-interacting regions of lncRNAs are likely to contain a combination of sequence and structure motifs that confer binding specificity.

Most small ncRNAs have well characterised secondary and tertiary structures, as evidenced in RFAM, the largest collection of curated RNA families (2,588 families as of version 12.2 [10]). In contrast, determining the structural features of lncRNAs is a 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 [11, 12], despite evolutionary and biochemical support of conserved base-pairing interactions [7, 13, 14].

Identifying RNAs with similar functions involves comparing both their primary sequence and higher-order structures simultaneously. If the sequence similarity falls below 60%, sequence comparison will not find anymore domain similarities that are based on structure [15]. Hence, methods are needed that find structural similarity independent from sequence conservation and freed from one single optimal RNA secondary structure. The Sankoff algorithm resolves the optimal sequence-structure alignment of two RNAs [16], but its computational complexity limits its practicality. Alternative strategies often employ precalculated secondary structure ensembles for each sequence, e.g. basepair probabilities in thermodynamically equilibrated RNA structure ensembles [17]. The latter can substantially speed up the calculation of structure-based alignments [18], of which there are many variants. The programs PMcomp [18], LocaRNA [19], and ProbAlign [20] use the precomputed base-pair probability matrices of both sequences and score the alignment based on the notion of a common secondary structure. The sequence-structure alignment problem is reduced to a two-dimensional problem by RNApaln [21] and StrAL [22] which reduce base-pair probabilities to base specific probabilities (such as unpaired probability). All these methods do not explicitely consider suboptimal structures in the alignment. The pairwise alignment of basepair probability matrices (dot plots) has been first introduced by CARNA [23, 24]. CARNA finds iterativelly better alignments with an effective constraint programming technique using a branch and bound scheme. Another heuristic is pruning of the dynamical programming matrix without pre-folding constraints, which is implemented by FoldAlign [25, 26].

Will *et al.* [19] first showed that a (dis)similarity matrix can be constructed from all-vs-all pairwise RNA structure alignments with the pairwise alignment tool `LocaRNA`, identifying known and novel groups of homologous RNAs using hierarchical clustering [19]. However, this strategy involves applying a subjective threshold to the resulting dendrogram to extract structurally related sequences. Alternative methods to identify clusters of homologous RNAs include `NoFold`, which clusters query sequences based on their relative similarity to reference structures [27], and `GraphClust`, an alignment-free approach that decomposes RNA structures into graph-encoded features [28]. `RNAscClust`, an extension of `GraphClust`, utilizes the evolutionary signatures of RNA structures as additional classification feature [29].

Here, we describe a computational pipeline for the identification and classification of homologous RNA structures from a large set of query sequences. At its core lies `DotAligner`, a heuristic pairwise sequence alignment algorithm that considers suboptimal base-pairing probabilities. We compare `DotAligner` with other pairwise RNA structure alignment algorithms to highlight its speed and accuracy at classifying know RNA families. We combine

DotAligner with density based clustering for the impartial identification of RNA structural motifs, which can identify known RFAM families and novel RNA structural motifs from ENCODE eCLIP data. The resulting clusters of homolgous RNA structures can then be used to search for homologous structures across reference genomes and transcriptomes.

Results and Discussion

Ensemble-guided pairwise RNA structure alignment

DotAligner 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. The algorithm overcomes the limitations of comparing unique RNA secondary structures (such as minimum free energy predictions) to yield an optimal sequence alignment that considers mutual base pair probabilities. **Figure 1** illustrates a structural alignment performed with DotAligner in contrast to an alignment that considers only sequence composition.

A major criteria for the implementation was a fast running time to make DotAlignerapplicable for RNA structure clustering of large data sets. Consistently, the algorithm performs pairwise sequence-structure alignments from pre-calculated RNA dot plots an using alignment-envelope heuristic, which impose constraints on sub-optimal string alignments, and fold-envelope heuristics, which impose constraints to pre-calculated base pairing probabilities. The alignment procedure thus consists of two steps, each considering base pairing probabilities: (i) Computation of a partition function over all canonical pairwise string alignments, and (ii) structure-weighted stochastic backtracking of all string alignments. The detailed implementation and mathematical description of DotAligner can be found in the **Supplementary Information**.

After initial parameter optimisation (see **Methods**), we applied DotAlignerto BRAliBase 2.1 pairwise RNA structure alignments, a reference dataset specifically designed for algorithm benchmarking [15, 30]. In this application, DotAligner seldom produces alignments of better quality than those generated by three other state of the art algorithms, namely CARNA [24], FOLDALIGN [31, 32] and LocaRNA [19] (**Figure 2**).

Interestingly, many of the pairwise structure alignments produced Structural Conservation Index (SCI) scores above those from the BRAliBase 2.1 reference alignments (**Figure 2B**). The SCI represents the alignment consensus energy normalized by the average energy of the single sequences folded independently [33]; It has been shown to be one of the most reliable metrics for conserved RNA structure detection [34]. With the exception if DotAligner, all other surveyed algorithm produce a substantial amount of alignments with SCI values above that of the reference alignment **Figure 2B** , suggesting that many optimization algorithms tend to overestimate the amount of paired bases in consensus RNA structure predictions. However, another possibility is that BRAliBase 2.1 alignments do not correctly depict several RNA familes, which may have been automatically generated in RFAM based on similarity to a covariance model.

Fast and accurate classification of RNA structures

The intended application of `DotAligner` 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 between distinct structured RNA species from a heterogeneous sample of known RNA structure families. We performed all versus all pairwise structure alignments of stochastically sampled RFAM sequences, which were selected with constraints on their sequence composition (PID) to control for and ascertain any sequence-dependent biases (see **Methods**). The alignment scores of `DotAligner` and other algorithms were then converted into a similarity matrix and compared to a binary classification matrix of sampled RFAM entries.

Despite the seemingly poor quality of pairwise alignments generated by `DotAligner`, it reproduces the known classification of RFAM structures more accurately than almost all other pairwise RNA structure alignment tools (**Figure 3A-C**). Only `CARNA`, another ensemble-based structural alignment algorithm, presents classification accuracies comparable to `DotAligner`. However, `CARNA` requires substantially more time to perform the comparisons (**Figure 3D**) as it will indefinitely continue to compute the alignment until it converges on an optimal result, or a hard time constraint is enforced. In this regard, `DotAligner` performs better than all other RNA structure alignment tools, highlighting the efficacy of the implemented heuristics it employs. In contrast, the heuristics implemented in `RNApaln` increased the speed of `pmcomp` by 2 orders of magnitude, but at a slight reduction in accuracy (Figure 3C,D). Only a C++ implementation of the Needleman-Wunsch Algorithm (NWA) [35]–a classical sequence alignment algorithm that ignores secondary structure information–performs faster than `DotAligner` on average, most likely due to the presence of stretches of homologous sequences within biologically related RNAs.

Density-based clustering of homologous RNA structures

Given `DotAligner`'s accurate classification of known structured RNA, we subjected its output to cluster analysis to identify and extract input sequences which display common sequence-structure motifs. The previous work by Will et al. applied hierarchical clustering to the dissimilarity matrices produced by `LocaRNA` to organise sequences based on their structural homology [19], yet this does not apply a cut-off that can be used to extract meaningful clusters of structurally homologous sequences in an unsupervised manner. We attempted to achieve this by applying a statistical threshold derived from bootstrapping the underlying data using `pvclust` [36], but this generated clusters of variable size that often spanned across many disjoint families (data not shown).

We therefore opted for a density-based clustering strategy that, in theory, can decipher clusters of varying density (i.e. subsets of the data with greater sequence-structure homology). The OPTICS (Ordering Points To Identify the Clustering Structure) algorithm [37] was chosen for this purpose, as it has very few parameters to optimise. OPTICS is a derivative of the Density-BaSed Clustering for Application with Noise (`DBSCAN`) [38] algorithm that, as its name states, is suitable for noisy data, such as RNA immunoprecipitation followed by high-throughput sequencing (RIPseq).

We then benchmarked the two main OPTICS clustering parameters–Xi steepness threshold and the minimum number of points in a cluster (**Supplementary Figure 1**)–on a pooled set of 580 stochastically sampled RFAM sequences encompassing various ranges of sequence similarity, as well as a corresponding set of 580 dinucleotide shuffled controls (see **Methods**). After performing all vs all pairwise alignments with `DotAligner`, the alignment scores were normalised from 0 to 1 for the minimal and maximal values, respectively. The resulting similarity matrix was then used to evaluate the effect of OPTICS parameters on clustering performance, revealing that a minimum of 4 points (or sequences) and a steepness threshold of 0.006 gave the best results **Supplementary Figure 1A**. Extracting only the minimum cluster when

When compared to other state of the art methods, namely `NoFold` and `GraphClust`, the combination of `DotAligner` and OPTICS performs comparably well (**Figure 4, Table 1**). The default version of `NoFold` nonetheless outshines `DotAligner` at clustering known RFAM families. However, it intrinsically employs RFAM covariance models when comparing input sequences, therefore is likely to be subject to over-fitting in this example. We thus removed the CMs associated to the RFAM sequences in our benchmarking dataset from the `NoFold` algorithm, which yielded lower sensitivity and less accurate qualitative cluster metrics than the `DotAligner` and OPTICS combination. Interestingly, the specificity of `NoFold` increased slightly despite removing 72 CMs from its classification set.

Identifying protein-binding RNA motifs from eCLIP data

The optimised parameters for OPTICS clustering of `DotAligner` output were incorporated into a high-performance computing pipeline that extracts clusters of homologous RNA structural motifs from a set of input sequences (see **Methods**). This pipeline was applied to enhanced cross-linked RNA immunoprecipitation (eCLIP) sequencing data from 44 RNA binding proteins from the ENCODE consortium [39].

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.

Conclusion

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 reasonable computational power.

In addition to its speed, `DotAligner`'s strength lies in it's capacity to accurately score structurally homoglous RNA sequences. The algorithm appears to generate pairwise alignments

that differ somewhat to the reference structural alignments. Despite this, `DotAligner` can harness the information content of base pair probability ensembles to output a reliable structural similarity score of two RNA sequences. Our results show that this can nonetheless be sufficient to extract structurally and functionally related sequences from a large amount of noisy input; an ideal application for screening high-throughput sequencing data, such as RNA immunoprecipitaion data, for common structural motifs.

As the increasing accessibility of next generation sequencing coalesces with precision medicine, in-depth transcriptome profiling will help elucidate the function and clinical impact of disease-associated, non-coding single nucleotide variants. Indeed, 80% of disease-associated single nucleotide polymorphisms occur in non-coding regions [40, 41]. Thus, elucidating the structural features of RNAs associated to RNA-binding proteins and ribonucleoprotein (RNP) complexes, combined to the systematic classification of their genome-wide occurrence, can identify novel riboSNitches (functional RNA structures that are disrupted in disease) and help pinpoint the molecular function of non-coding mutations. Something about riboswitches.

weaknesses: boundaries, paramter-choice of algorithms.

Methods

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 [30]. 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 [34], and the Matthews Correlation Coefficient (MCC) were also calculated (see below) for all resulting alignments.
Baseline parameters were then selected via a product rank these 2 metrics (**supplementary data?**).

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\text{Structural Conservation Index (SCI)} = MFE_{consensus}/\overline{MFE}_{single}$$

$$\delta\text{SCI} = SCI_{observed}/SCI_{reference}$$

Classification of RNA secondary structure families

Further refinement of the optimal parameters was performed using a binary classifier two sets of 200 stochastically sampled RFAM entries with published structures: (i) a low pairwise sequence identity set, and (ii) a high pairwise sequence identity set, where any two sequences from the same family share between 0-55% and 56-95% pairwise sequence identity, respectively. A binary classification matrix was then constructed, where sequences `x` and `y` present a score of 1 if they belong to the same RFAM family, versus a score of 0 if they do not. The similarity matrix resulting from all-vs-all pairwise comparisons with

`DotAligner` was tested for accuracy using the Area Under the Curve AUC of the ROC, as calculated R package `pROC` [42]. A more restricted range of parameter values were then tested on both datasets, namely t k e o, which had the highest impact on alignment accuracy in preliminary testing (SUPP FIG X??). Finally, a ranked sum for both datasets of the AUC (first) and lowest average runtime (second) was performed to determine the default runtime parameters for `DotAligner` Supplementary table X.

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 (PID) and a low PID set, where any two sequences from the same set present greater than or less than 55% PSI. The

Clustering RNA structures with randomised controls

OPTICS benchmarking was performed by stochastically sampling the collection of RFAM 12.0 seed alignments using the accompanying `JAVA` program `GenerateRFAMsubsets.java` (see **Supplementary Information**) with parameters GET PARAMETERS FROM SERVER, how many sequences minimum/maximum, with 3 ranges of pairwise sequence identity: 1-55%, 56-75%, and 75-95%. The resulting 580 unique sequences were then randomised while controlling their dinucleotide content with the `easel` program included in the `Infernal` (v1.1.2) software package [43] with option "-k 2". The 1160 sequences were submitted to all-vs-all pairwise comparisons with `DotAligner` and the scores were inverted and a normalised (min=1, max=0) into a dissimilarity matrix, which was then imported into the `R` statistical programming language, converted into a 'dist' object without transformation, and subjected to `OPTICS` clustering as implemented in the 'dbscan' CRAN repository with a range of parameters (see **Figure 4A,B**).

Clustering performance metrics were calculated like so:

- True Positives (TP) = Number of representatives from the dominant RFAM family in a cluster;
- False Positives (FP) = Number of non-dominant RFAM family representatives in cluster, or clusters where there is no dominant RFAM family (i.e. equally represented families), or clusters where dominant sequence is a negative control;
- False Negatives (FN) = RFAM sequences that fail to cluster;
- True Negatives (TN) = Negative control sequences that fail to cluster.
- Sensitivity (recall) = TP / (TP + FN) ;
- Specificity = TN / TN + FP ;
- False positive rate = 1 - Specificity ;
- Precision = TP / (TP + FP) ;
- Accuracy = (TP + TN) / (TP + TN + FP + FN) ;

Clustering of protein-bound evolutionarily-conserved RNAseq reads

The genomic coordinates of ENCODE eCLIP peaks were downloaded in .bed format from the April 2016 release via the ENCODE portal (`https://www.encodeproject.org/search`). The resulting 5,040,096 peaks were filtered to keep only those with $\geq$8x fold enrichment over the total input background and an associated P-value $\leq 10^{-4}$. Furthermore, peaks were merged if they overlapped by more than 50nt to avoid over-representing

the same sequence (see **Supplementary Information** section 2). The remaining peaks were subsequently filtered by retaining only those that present same-strand overlap with any evolutionarily conserved structure (ECS) predictions from [7]. Finally, the associated genomic sequences were extracted into a .fasta file, which was supplemented with 100 reference RNA structures from 11 RFAM families (see Supplementary Table 1). Merging, overlap, and sequence extraction operations were performed with `bedtools` version v2.26.0.

The normalised similarity matrix resulting from all vs all pairwise comparisons with `DotAligner` was then subjected to clustering with the `dbscan` 1.1-1 R package from Michael Hahsler https://github.com/mhahsler/dbscan using the command '`opticsXi(optics(D, eps=1, minPts=4, search="dist"), xi = 0.006, minimum=T)`'. The sequences for each cluster were then extracted and submitted to multiple structure alignment with `mLocaRNA` version 1.9.1 using parameters '`--probabilistic --iterations=10 --consistency-transformation --noLP`'.

The complete analytical pipeline is available at https://github.com/noncodo/BigRedButton, which will identify and extract homologous RNA structural motifs from a set of input sequences in fasta format.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
MAS and JSM are partially supported by a Cancer Council NSW project grant (RG 14-18). SES was supported by a Carlsberg Foundation grant (2011_01_0884) and the Innovation Fund Denmark (0603-00320B). Michael Hahsler for developing and disseminating a fast DBSCAN R package.

Author details
[1]Garvan Institute of Medical Research, 384 Victoria Street, NSW 2010 Sydney, Australia. [2]St Vincent's Clinical School, UNSW Australia, Victoria Street, NSW 2010 Sydney, Australia. [3]Center for non-coding RNA in Technology and Health (RTH), University of Copenhagen, Groennegaardsvej 3, 1870 Frederiksberg, Denmark.

References

1. Morris, K.V., Mattick, J.S.: The rise of regulatory rna. Nature Reviews Genetics **15**(6), 423–437 (2014)
2. Engreitz, J.M., Ollikainen, N., Guttman, M.: Long non-coding rnas: spatial amplifiers that control nuclear structure and gene expression. Nature Reviews Molecular Cell Biology (2016)
3. Zappulla, D., Cech, T.: Rna as a flexible scaffold for proteins: yeast telomerase and beyond. In: Cold Spring Harbor Symposia on Quantitative Biology, vol. 71, pp. 217–224 (2006). Cold Spring Harbor Laboratory Press
4. Hogg, J.R., Collins, K.: Structured non-coding rnas and the rnp renaissance. Current opinion in chemical biology **12**(6), 684–689 (2008)
5. Rinn, J.L., Chang, H.Y.: Genome regulation by long noncoding rnas. Annual review of biochemistry **81**, 145–166 (2012)
6. Mercer, T.R., Mattick, J.S.: Structure and function of long noncoding rnas in epigenetic regulation. Nature structural & molecular biology **20**(3), 300–307 (2013)
7. Smith, M.A., Gesell, T., Stadler, P.F., Mattick, J.S.: Widespread purifying selection on rna structure in mammals. Nucleic acids research, 596 (2013)
8. Chujo, T., Yamazaki, T., Hirose, T.: 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 (2016)
9. Blythe, A.J., Fox, A.H., Bond, C.S.: 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 (2016)
10. 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.*: Rfam 12.0: updates to the rna families database. Nucleic acids research **43**(D1), 130–137 (2015)
11. Eddy, S.R.: Computational analysis of conserved rna secondary structure in transcriptomes and genomes. Annual review of biophysics **43**, 433–456 (2014)
12. Rivas, E., Clements, J., Eddy, S.R.: A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. Nature Methods (2016)

13. 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.*: Structural imprints in vivo decode rna regulatory mechanisms. Nature **519**(7544), 486–490 (2015)
14. 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.*: Rna duplex map in living cells reveals higher-order transcriptome structure. Cell **165**(5), 1267–1279 (2016)
15. Gardner, P.P., Wilm, A., Washietl, S.: A benchmark of multiple sequence alignment programs upon structural {RNAs}. Nucleic Acids Res **33**(8), 2433–2439 (2005). doi:10.1093/nar/gki541
16. Sankoff, D.: Simultaneous solution of the {RNA} folding, alignment and protosequence problems. SIAM J. Appl. Math. **45**, 810–825 (1985)
17. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for {RNA} secondary structure. Biopolymers **29**(6-7), 1105–1119 (1990). doi:10.1002/bip.360290621
18. Hofacker, I.L., Bernhart, S.H., Stadler, P.F.: Alignment of RNA base pairing probability matrices. Bioinformatics **20**(14), 2222–2227 (2004). doi:10.1093/bioinformatics/bth229
19. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., Backofen, R.: Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol **3**(4), 65 (2007). doi:10.1371/journal.pcbi.0030065
20. Roshan, U., Livesay, D.R.: Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics **22**(22), 2715–2721 (2006). doi:10.1093/bioinformatics/btl472
21. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA package 2.0. Algorithms Mol Biol **6**, 26 (2011). doi:10.1186/1748-7188-6-26
22. Dalli, D., Wilm, A., Mainz, I., Steger, G.: STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. Bioinformatics **22**(13), 1593–1599 (2006). doi:10.1093/bioinformatics/btl142
23. Palù, A., Möhl, M., Will, S.: A Propagator for Maximum Weight String Alignment with Arbitrary Pairwise Dependencies. In: Cohen, D. (ed.) Principles and Practice of Constraint Programming – CP 2010, Lecture no edn., pp. 167–175 (2010). doi:10.1007/978-3-642-15396-9_16. http://dx.doi.org/10.1007/978-3-642-15396-9_16
24. Sorescu, D.A., Möhl, M., Mann, M., Backofen, R., Will, S.: CARNA–alignment of RNA structure ensembles. Nucleic acids research **40**(Web Server issue), 49–53 (2012). doi:10.1093/nar/gks491
25. Havgaard, J.H., Torarinsson, E., Gorodkin, J.: Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol **3**(10), 1896–1908 (2007). doi:10.1371/journal.pcbi.0030193
26. Sundfeld, D., Havgaard, J.H., de Melo, A.C., Gorodkin, J.: Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. Bioinformatics **32**(8), 1238–1240 (2016). doi:10.1093/bioinformatics/btv748
27. Middleton, S.A., Kim, J.: Nofold: RNA structure clustering without folding or alignment. RNA **20**(11), 1671–1683 (2014). doi:10.1261/rna.041913.113
28. Heyne, S., Costa, F., Rose, D., Backofen, R.: GraphClust: alignment-free structural clustering of local {RNA} secondary structures. Bioinformatics **28**(12), 224–32 (2012). doi:10.1093/bioinformatics/bts224
29. Miladi, M., Junge, A., Costa, F., Seemann, S.E., Hull Havgaard, J., Gorodkin, J., Backofen, R.: RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics (2017). doi:10.1093/bioinformatics/btx114
30. Wilm, A., Mainz, I., Steger, G.: An enhanced rna alignment benchmark for sequence alignment programs. Algorithms for molecular biology **1**(1), 1 (2006)
31. Havgaard, J.H., Torarinsson, E., Gorodkin, J.: Fast pairwise structural rna alignments by pruning of the dynamical programming matrix. PLoS Comput Biol **3**(10), 193 (2007)
32. Sundfeld, D., Havgaard, J.H., de Melo, A.C., Gorodkin, J.: Foldalign 2.5: multithreaded implementation for pairwise structural rna alignment. Bioinformatics, 748 (2015)
33. Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding rnas. Proceedings of the National Academy of Sciences of the United States of America **102**(7), 2454–2459 (2005)
34. Gruber, A.R., Bernhart, S.H., Hofacker, I.L., Washietl, S.: Strategies for measuring evolutionary conservation of rna secondary structures. BMC bioinformatics **9**(1), 122 (2008)
35. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48**(3), 443–453 (1970)
36. Suzuki, R., Shimodaira, H.: Pvclust: an r package for assessing the uncertainty in hierarchical clustering. Bioinformatics **22**(12), 1540–1542 (2006)
37. Ankerst, M., Breunig, M., Kriegel, H., *et al.*: Ordering points to identify the clustering structure. In: Proc. ACM SIGMOD, vol. 99 (1999)
38. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.*: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231 (1996)
39. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., *et al.*: Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). Nature methods **13**(6), 508–514 (2016)
40. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A.: 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 (2009)
41. Ritchie, G.R., Dunham, I., Zeggini, E., Flicek, P.: Functional annotation of noncoding sequence variants. Nature methods **11**(3), 294–296 (2014)
42. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M.: proc: an open-source package for r and s+ to analyze and compare roc curves. BMC bioinformatics **12**(1), 1 (2011)
43. Nawrocki, E.P., Eddy, S.R.: Infernal 1.1: 100-fold faster rna homology searches. Bioinformatics **29**(22), 2933–2935 (2013)
44. Muckstein, U., Hofacker, I.L., Stadler, P.F.: Stochastic pairwise alignments. Bioinformatics **18 Suppl 2**, 153–60 (2002)
45. Klein, R.J., Eddy, S.R.: RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics **4**, 44 (2003). doi:10.1186/1471-2105-4-44

Figures

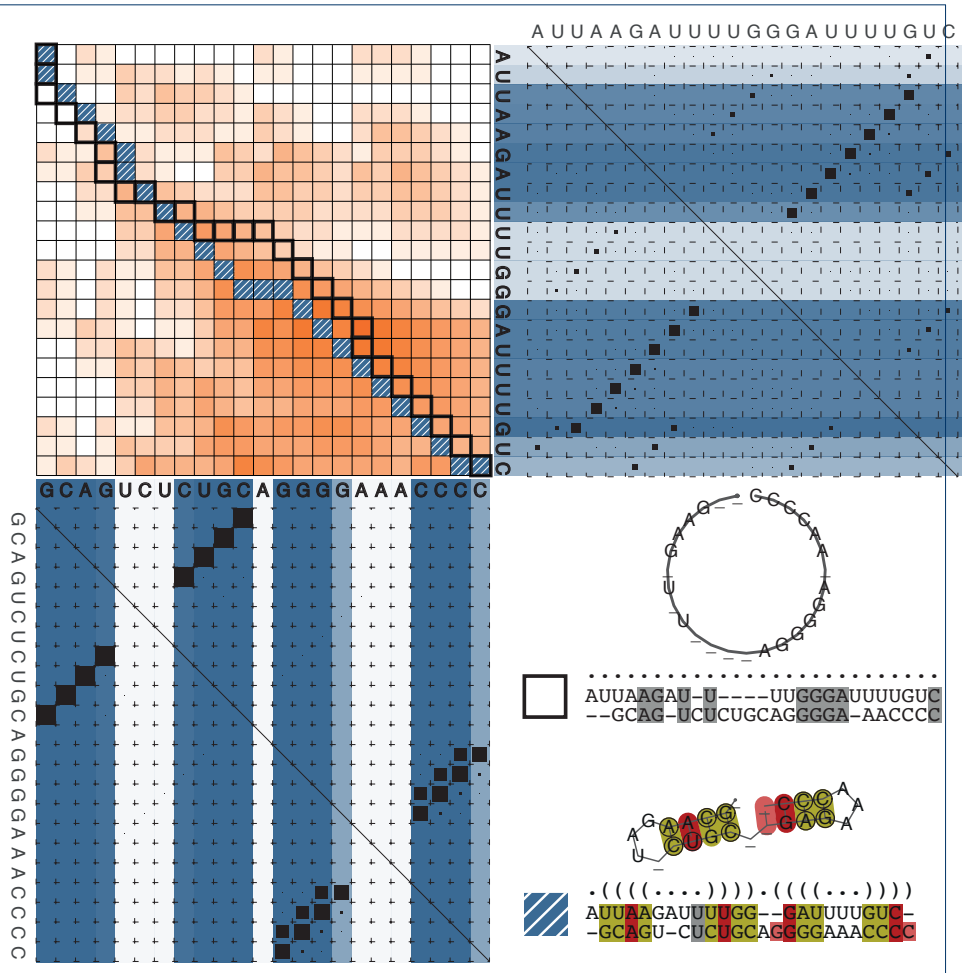

Figure 1: **Schematic of a pairwise alignment with DotAligner.** A dynamic programming matrix is first filled in based on sequence similarity (top left–color intensity indicates cumulative sequence similarity score). A partition function over all pairwise sequence alignments is then calculated and interrogated for structural compatibility by stochastic backtracking. Two ensemble of suboptimal secondary structures are considered for this purpose (bottom left and top right dot plots; blue lines indicate cumulative basewise pairing probabilities) This 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.

Table 1: Comparative clustering performance.

Algorithm	# Clusters	Sensitivity	Specificity	Accuracy
`DotAligner+OPTICS`	53	0.716	0.886	0.802
`GraphClust`	201	**0.990**	0.110	0.635
`NoFold` (all CMs)	**62**	0.866	0.965	**0.916**
`NoFold` (filtered)	45	0.674	**0.976**	0.826

Tables

Additional Files

Additional file 1 — Supplemental Material

More detailed description of the `DotAligner` implementation, RNA structure clustering and eCLIP data processing, and Supplemetal Tables and Figures summarizing study results.

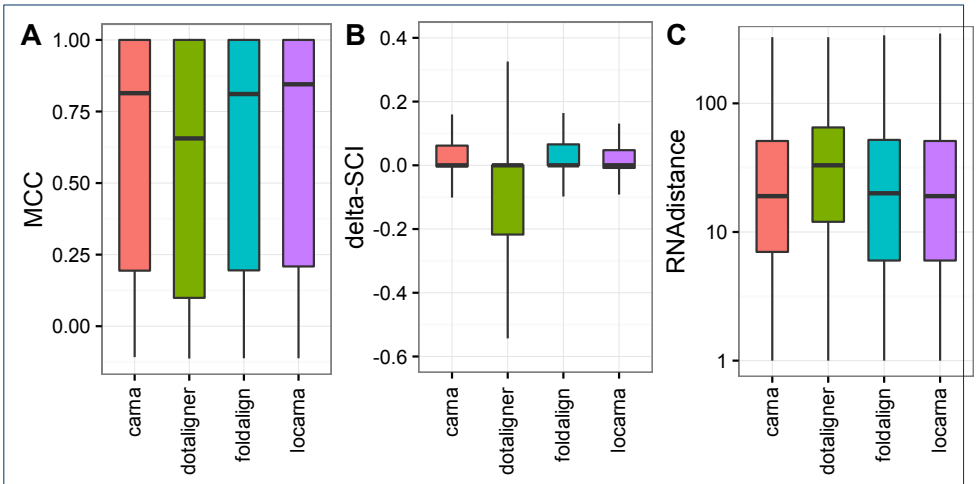

Figure 2: **Comparison of RNA structure alignment qualities.** The Matthews Correlation Coefficient (**A**), Structural Conservation Index (**B**), and `RNAdistance` calculated topological edit distance (**C**) are compared between the `RNAalifold` consensus of the computed alignment and the and reference `BRAliBase 2.1` alignment consensus.

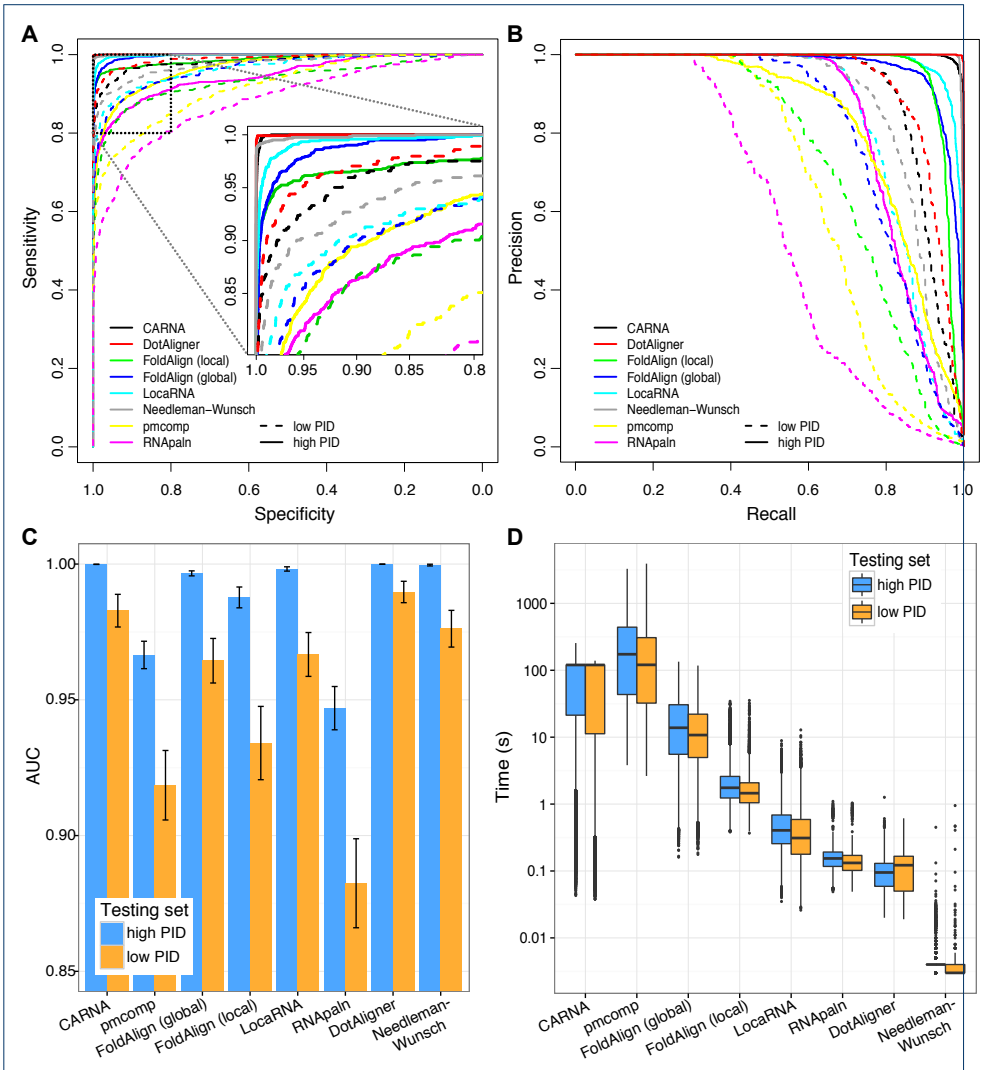

Figure 3: **Classification of known RNA structures.** (**A**) Receiving Operator Characteristic (ROC) curves measuring the classification accuracy by contrasting the computed similarity matrices of each algorithm to a binary classification matrix of RFAM sequences (1 if same family; 0 if different).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

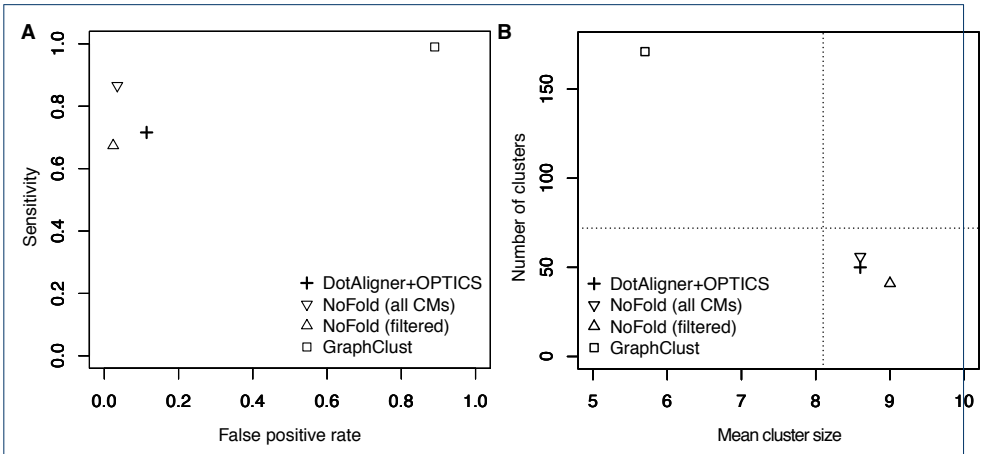

Figure 4: **Comparative clustering benchmark of RFAM sequences and their shuffled controls**. Clustering performance metrics of 3 algorithms on 580 reference RFAM structures and their dinucleotide-shuffled controls. (**A**) Sensitivity vs false positive rate; (**B**) Qualitative cluster statistics (horizontal dashed line indicate real amount of clusters from unique RFAM families).

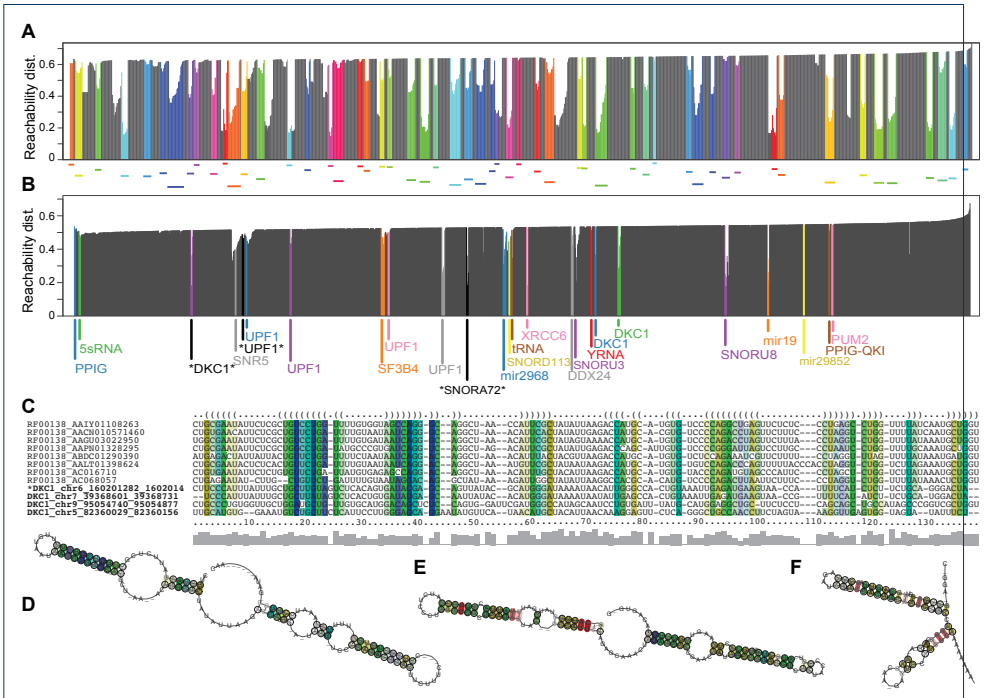

Figure 5: **De novo homologous RNA motif identification**. (**A**) Reachability plot of OPTICS clustering on RFAM benchmarking data indicating the distance to the nearest neighbour. Colors represent significant clusters. (**B**) Reachability plot of 2,650 ENCODE eCLIP peaks + 100 RFAM controls overlapping evolutionarily conserved secondary structure predictions. The dominant RNA binding protein in each cluster is displayed next to significant clusters; those with an asterix and portrayed below. (**C**) Multiple structure alignment generated by `mLocaRNA` on the sequences from a cluster containing both RFAM snoRNA72 sequences, and DKC1 (a snoRNA-binding protein) eCLIP peaks. An unannotated DKC1-bound sequence is marked with an asterix. (**E**) `RNAalifold` consensus of the alignment displayed in (**C**), a cluster of impartially detected DKC1-bound snoRNAs (**E**), and a novel UPF1-bound motif (**F**).

Additional file 1 — Supplemental Material

1. `DotAligner` implementation

As described in [23] the weight W of alignment A of two arc-annotated sequences (S_a, P_a) and (S_b, P_b) is defined by

$$\begin{aligned} W(A) &= \sigma(A) + \tau(A) + \gamma(A) \\ &= \sum_{(i,i')\in A} \sigma(i,i') + \sum_{\substack{(i,j)\in P_a,\\ (i',j')\in P_b,\\ (i,i')\in A,\\ (j,j')\in A}} \tau(i,j,i',j') + \gamma \times N \end{aligned} \tag{1}$$

where S is a sequence and P is a base pairing probability matrix, $\sigma(i,i')$ is the similarity of sequence positions $S_a[i]$ and $S_b[i']$, $\tau(i,j,i',j')$ is the similarity of arcs $(i,j) \in P_a$ and $(i',j') \in P_b$, and γ is the gap cost associated with each sequence position that is not matched ($N = |S_a| + |S_b| - 2|A|$). The alignment problem finds the maximal $W(A)$. As its solution is MAX-SNP-hard, in praxis heuristics are used to find near-optimal solutions.

`DotAligner` solves the related problem of aligning two basepair probability matrices (dot plots). A major criteria for the implementation was a fast running time enable all-vs-all pairwise structural alignments and the associated distance (dissimilarity) matrix, which can be used for cluster analysis of large data sets [19]. Consequently, it employs a heuristic alignment-envelope, which imposes constraints to sub-optimal string alignments, and a fold-envelope, which imposes constraints to pre-calculated base pairing probabilities, to build pairwise sequence-structure alignments. `DotAligner` makes use of the observation that large samples from the ensemble of stochastic sequence alignments contain the correct structure-based alignment with significant probability, even though the optimal sequence alignment deviates significantly from the structural alignment [44].

Below, we describe the alignment procedure and weight functions. The alignment procedure consists of two steps:

1. Pairwise probabilistic string alignments;
2. Stochastic backtracking of string alignments and combining weights of base pairing probability matrices.

1.1 Pairwise probabilistic string alignments

In step 1 the computation of the partition function over all canonical pairwise string alignments is adapted from probA [44]. The probability of an alignment A in the ensemble of all alignments $Z(T)$ is

$$Pr(A;T) = \frac{1}{Z(T)} exp(\beta W(A)), \tag{2}$$

where $\beta = 1/T$. The parameter T is analogous to the temperature in the thermodynamic interpretation of the alignment problem and determines the relative importance of the optimal string alignment. If $T = 1$ then we recover the 'true' probability, if $T \to 0$ then $Pr(A;0) = 0$ for all alignments with a score $W(A)$ less then the score of the optimal string alignment,

and if $T \to \infty$ then all alignments have the same $Pr(A,\infty) = 1/Z(\infty)$. Hence, T controls the search space of suboptimal alignments for step 2. The algorithm runs in $O(N^3)$ for calculating the partition function. The weight function $W(A)$ of the probA implementation is changed to explore the ensemble of dot plot alignments. We reduce the sequence-structure alignment problem to a two-dimensional problem similar to the metric introduced in StrAL [22]. Hence, step 1 considers only the similarity σ and the gap cost γ described in equation 1:

$$W_{\text{Step1}}(A) = \sigma(A) + \gamma(A) \tag{3}$$

The similarity $\sigma(i,i')$ for matched sequence positions $S_a[i]$ and $S_b[i']$ takes into account sequence similarity M_{Seq} and the similarity in their unpaired probabilities $\Delta\omega(i,i')$ weighted by the parameter θ:

$$\sigma(i,i') = \theta \times M_{Seq}^{(i,i')} + (1-\theta) \times \Delta\omega(i,i') \tag{4}$$

$M_{Seq}^{(i,i')}$ is 1 if sequence positions $S_a[i]$ and $S_b[i']$ match and else 0. The similarity of unpaired probabilities is defined as

$$\Delta\omega(i,i') = \begin{cases} 0 & \text{if } \omega(i) == 0 \\ & \text{and } \omega(i') == 0 \\ 1 - |\omega(i) - \omega(i')| & \text{else} \end{cases} \tag{5}$$

so that $\Delta\omega = (0,1)$. Alternatively a statistical substitution model R_{Seq} replaces the sequence similarity and is multiplied with the ζ weighted sum of $\Delta\omega$ and the similarity in ratios of upstream pairing probability $\Delta\omega^{up}$:

$$\begin{aligned} \sigma(i,i') = {} & R_{Seq}^{(i,i')} \times \zeta \times \Delta\omega(i,i') + \\ & R_{Seq}^{(i,i')} \times (1-\zeta) \times \Delta\omega^{up}(i,i') \end{aligned} \tag{6}$$

R_{Seq} is a 4×4 matrix of probabilities for observing a given substitution relative to background nucleotide frequencies. We use the log-odd scores L from the RIBOSUM85-60 matrix introduced in [45] which are transformed to probabilities R_{Seq} by $2^{L(i,i')}/(1+2^{L(i,i')})$. The ratio of upstream pairing probability ω^{up} is defined as

$$\omega^{up}(i) = \sum_{k=1}^{i-1} \psi(k,i) / \sum_{k=1}^{|S|} \psi(k,i) \tag{7}$$

where $i \in S$, $|S|$ is the length of sequence S, and $\psi(k,i)$ is the pairing probability of sequence positions $S[k]$ and $S[i]$. The downstream pairing probability is implicitly considered in the

weight function through the usage of unpaired probability and upstream pairing probability. The gap term in equation 1 is replaced with affine gap costs:

$$\gamma(A) = l \times g_o + (N - l) \times g_{ext} \tag{8}$$

where l is the number of initiation gaps, N is the number of all gaps, g_o is the penalty for opening a gap and g_{ext} is the penalty for gap extensions. Start and end gaps are considered as free.

1.2 Stochastic backtracking and combined weight of dot plot alignments

Here, a properly weighted sample of stochastic pairwise string alignments in the alignment ensemble is examined across both sequences for sequence-structure similarity. The stochastic backtracking is adapted from probA [44] for selecting s suboptimal string alignments A_s. The combined weight W_{Step2} is a variant of equation 1 to explore the similarity of the corresponding dot plot alignments:

$$W_{\text{Step2}}(A_s) = \kappa \times \frac{W_{\text{Step1}}(A_s)}{|A_s|} + (1 - \kappa) \times \frac{\tau(A_s)}{|\text{Match}_{A_s}|^2} \tag{9}$$

where the parameter κ weights for each alignment A_s between the sequence-based similarity $W_{\text{Step1}}(A_s)$ normalized by alignment length $|A_s|$ and dot plot similarity $\tau(A_s)$ normalized by the number of aligned bases $|\text{Match}_{A_s}|$ in alignment A_s. Similar to equation 4 the dot plot similarity τ sums the parameter θ weighted similarity of aligned basepairs M_{paired} and the similarity in their pairing probabilities $\Delta\psi$:

$$\tau(i, j, i', j') = \theta \times M_{paired}^{(i,j,i',j')} + (1 - \theta) \times \Delta\psi(i, j, i', j') \tag{10}$$

where $M_{paired}^{(i,j,i',j')}$ is 1 if $S_a[i]$ and $S_a[j]$ as well as $S_b[i']$ and $S_b[j']$ form canonical basepairs (G-C, C-G, A-U, U-A, G-U or U-G) and else 0. The similarity in pairing probabilities $\Delta\psi$ is then calculate by

$$\Delta\psi(i, j, i', j') = \begin{cases} 0 \text{ if } \psi(i, j) == 0 \text{ and } \psi(i', j') == 0 \\ 1 - |\psi(i, j) - \psi(i', j')| \text{ else} \end{cases} \tag{11}$$

Similar to M_{Seq} in equation 4, the basepair similarity matrix M_{paired} can be replaced by a statistical substitution model R_{paired} which describes the probability for observing a given basepair substitution relative to background nucleotide frequencies:

$$\tau(i, j, i', j') = R_{paired}^{(i,j,i',j')} \times \Delta\psi(i, j, i', j') \tag{12}$$

Again, the log-odd scores L from the RIBOSUM85-60 matrix [45] are transformed to probabilities R_{paired}.

For both sequences S_a and S_b, the pairing probability matrices P_a and P_b are computed in advance using McCaskill's algorithm, implemented in RNAfold or RNAplfold. The robustness of the alignment is improved by applying log-odds scores ψ of having a specific base pairing against the null model of a random pairing [19]:

$$\psi(i,j) = max\left(0, log\frac{P(i,j)}{p_0} / log\frac{1}{p_0}\right) \tag{13}$$

where p_0 is the expected probability for a pairing to occur at random. The term $log\frac{1}{p_0}$ is a normalization factor that transforms the scores to a maximum of 1. $P == 1$ results in $\psi = 1$, $P > p_0$ results in $\psi > 0$, and $P \leq p_0$ results in $\psi = 0$. This transformation gives weaker similarities if low basepair probabilities are compared, but stronger similarities for high basepair probabilities. Unpaired probabilities are handled in a similar way by

$$\omega(i) = max\left(0, log\frac{1-\sum_k P(i,k)}{p_0} / log\frac{1}{p_0}\right) \tag{14}$$

where p_0 is the expected probability for an unpaired base to occur at random.

2. Clustering RNA structures with randomised controls

Below is the code used to calculate the accuracy and other performance metrics of the clustering benchmark of stochastically sampled RFAM entries. All files can be found on the associated `GitHub` repository https://github.com/noncodo/BigRedButton.

```
cat("File name","TP","TN","FP","FN","SENS","SPEC","ACC","\n",sep="\t",
    file="accuracies.tsv")
file.names <- dir(pattern="*_clust.tsv$")
for(x in 1:length(file.names)){
gc <- read.delim(file.names[x], header=F)
# for 1 - max V2
TP=0
FP=0
NumClust <- max(gc$V2)
for ( cl in 0:NumClust) {
  if ( cl %in% gc$V2 ) {
  v <- as.vector( gc$V1[ gc$V2 == cl ] )
  t <- sort( table( v ), decreasing=T )
  best <- as.integer( t[1] )
  cID <- names( t[ 1 ] )
  if ( cl == 0 ) {
    if ( cID == "shuffled" ) {
      FN <- length(v)-best
      TN <- best
    }
    else
      cat("Houston, we have a TN problem")
  }
  else {
    if ( cID == "shuffled" ) {
```

```
      FP = FP + length(v)
    }
    if ( is.na( as.integer( t[2] )) || as.integer( t[2] ) < best ) {
      TP = TP + best
      FP = FP + length(v)-best
    }
    else if ( as.integer( t[2] ) == best ) {
      # treat both as false positives
      FP = FP + length(v)
    }
  }
}}
TP
TN
FP
FN
SENS=TP / (TP + FN )
SENS
SPEC=TN / ( TN + FP )
SPEC
ACC=(TP + TN) / ( TP + TN + FP + FN )
ACC
cat(file.names[x],TP,TN,FP,FN,SENS,SPEC,ACC,"\n",sep="\t",
     file="accuracies.tsv", append=T)
}
```

3. eCLIP data processing

Data in .bigBed format was acquired from the ENCODE data hub from the following link: https://www.encodeproject.org/search/?type=Experiment&assay_term_name=eCLIP&files.file_type=bigBed+narrowPeak&month_released=April%2C+2016

```
# Convert accessions to protein IDs
cut -f 1,16,29 metadata.tsv | sed 's/-human /_/g' | while read line
do
  F1=$(echo $line | awk '{print $1".bed"}' )
  F2=$( echo $line | awk '{ print $2".bed"}')
  cp $F1 $F2
done

# Rename files accordingly
for file in *bed
do
  mv $file $(head -n 1 $file | cut -f 4).bed
done

# Filter for greater than or equal to 8x fold enrichment
# And -log10( P-value ) greater than or equal to 4
for file in *rep0?.bed
do
  awk '{if ($7 >= 4 && $8 >= 4) print }' $file > ../filtering/${file}_filt3
done

#Intersect both replicates (>1 overlap)
for file in *rep01.bed_filt3 ; do
    >&2 echo "Processing "$file
    bedtools intersect -s -u -f 0.5 -a <( cut -f 1-6 $file ) -b <( cut -f 1-6
        ${file//rep01/rep02} ) > ${file}_1
```

Supplementary Table 1. List of RFAM families from benchmark that did not cluster

Sequence count	RFAM ID	RFAM family
2	RF00005	tRNA
5	RF00015	U4 spliceosomal RNA
8	RF00020	U5 spliceosomal RNA
5	RF00021	Spot 42 RNA
1	RF00026	U6 spliceosomal RNA
10	RF00059	TPP riboswitch (THI element)
5	RF00167	Purine riboswitch
11	RF00169	Bacterial small signal recognition particle RNA
13	RF00199	SL2 RNA
4	RF00374	Gammaretrovirus core encapsidation signal
11	RF00378	Qrr RNA
6	RF00386	Enterovirus 5' cloverleaf cis-acting replication element
6	RF00389	Bamboo mosaic virus satellite RNA cis-regulatory element
4	RF00444	PrrF RNA
17	RF00494	Small nucleolar RNA U2-19
2	RF00515	PyrR binding site
4	RF00550	Hepatitis E virus cis-reactive element
7	RF01685	6S-Flavo RNA
7	RF01697	Chlorobi-RRM RNA
6	RF01705	Flavo-1 RNA
4	RF01725	SAM-I/IV variant riboswitch
2	RF01728	STAXI RNA
7	RF01734	crcB RNA
1	RF01750	pfl RNA
6	RF01754	radC RNA
4	RF01764	yjdF RNA
5	RF02033	HNH endonuclease-associated RNA and ORF (HEARO) RNA

```
    bedtools intersect -s -u -f 0.5 -b <( cut -f 1-6 $file ) -a <( cut -f 1-6
        ${file//rep01/rep02} ) > ${file}_2

    # merge peaks if they are close together
    bedtools merge -d 50 -s -delim "|" -c 4,5,6 -o first,count,first -i <( cat
        ${file}_1 ${file}_2 | sort -k 1,1 -k 2,2n ) >
        ${file%*.bed_filt3}_filt_0.5_merged_50_s.bed

   # intersect with ECS (in file ECS_congruous_sorted.bed6)
    bedtools intersect -wo -s -b ${file%*.bed_filt3}_filt_0.5_merged_50_s.bed
        -a ECS_congruous_sorted.bed6 >
        ${file%*.bed_filt3}_filt_0.5_merged_50_s_anyECS.bed
done

#merge all files into one
cat *_50_s_anyECS_merged.bed > All_ECS_merged_50nt_peaks.bed
# wc -l All_ECS_merged_50nt_peaks.bed
## 2650

#edit sequence names and get sequence from reference genome (hg19)
awk 'OFS="\t"{print $1,$2,$3,$4"_"$1"_"$2"_"$3"_"$6,$5,$6}'
    ./All_ECS_merged_50nt_peaks.bed > ./All_ECS_merged_50nt_peaks_renamed.bed
bedtools getfasta -s -name -fi ~/data/fasta/hg19.fa -bed
    ./All_ECS_merged_50nt_peaks_renamed.bed -fo
    ./All_ECS_merged_50nt_peaks_renamed.fasta

#combine with known control RNA structure
cat All_ECS_merged_50nt_peaks_renamed.fasta spike-ins.fasta >
    All_ECS_merged_50nt_peaks_renamed_spikeIns.fasta
```

Supplementary Table 2. List of control RNA structures

Sequences	RNA family	RFAM ID
5	5SRNA	RF00002
8	SNORA72	RF00138
10	SNORD113	RF00181
10	SNORU3	RF00012
10	SNORU8	RF00096
8	SNR5	RF01252
9	YRNA	RF00019
10	mir19	RF00245
7	mir2968	RF02093
6	mir29852	RF02095
17	tRNA	RF00005

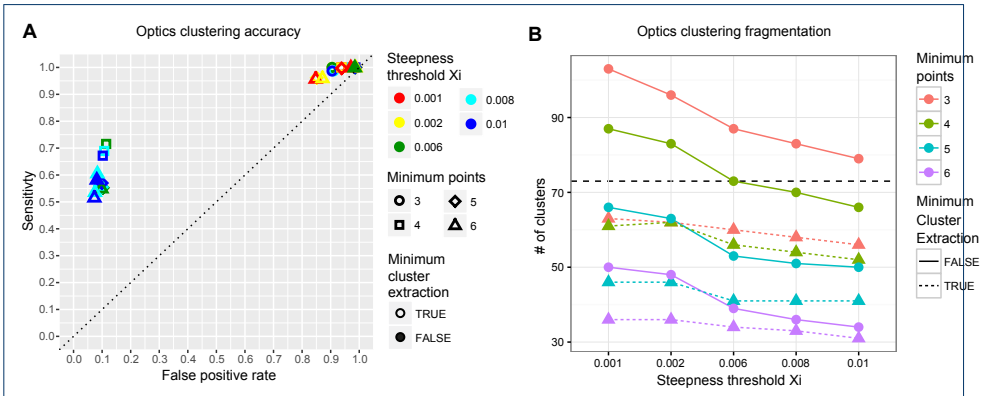

Figure S1. OPTICS clustering optimisation

Effect of OPTICS parameters on clustering accuracy (**A**) and amount of clusters (**B**) from a `DotAligner` dissimilarity matrix of 580 reference RFAM structures and their dinucleotide-shuffled controls (horizontal dashed line indicates expected amount of clusters, or unique RFAM families).