METHOD

Identification and clustering of RNA structure motifs

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

*Correspondence:
m.smith[at]garvan.org.au
[1]RNA Biology and Plasticity Group, 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 diversity of processed transcripts in eukaryotic genomes poses a challenge for the classification of their biological functions. Sparse sequence conservation in non-coding sequences and the unreliable nature of RNA structure predictions further exacerbate this conundrum. Here, we describe a computational method for the unsupervised discovery and classification of homologous RNA structure motifs from a set of sequences of interest. Our approach outperforms comparable algorithms at clustering known RNA structure families, both in time and accuracy. It identifies clusters of known and novel structure motifs from ENCODE immunoprecipitation data of 44 RNA-binding proteins.

Keywords: Functions of RNA structures; RNA structure clustering; Machine learning; RNA–protein interactions; Functional genome annotation; Regulation by non-coding RNAs

Background

As genomic technologies progress, an ever increasing amount of non-protein coding RNAs (ncRNA) are being discovered. Long non-coding RNAs (lncRNAs) are of particular interest for functional genome annotation given their abundance throughout the genome. So far, few lncRNAs have been functionally characterised, and those that have seem to be involved in regulation of gene expression and epigenetic states [1, 2]. 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.

Many ncRNAs lack sequence conservation, in contrast to protein-coding genes. 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 [3]). 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 [4, 5], despite evolutionary and biochemical support of conserved base pairing interactions [6, 7, 8]. Nonetheless, 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.

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 [9, 10, 11, 12, 6, 13, 14]. Protein-interacting regions of lncRNAs are likely to contain a combination of sequence

and structure motifs that confer binding specificity, which may be present in multiple target transcripts. For example, there is evidence that sequence and structure components of transposable elements, which are frequent in lncRNAs [15, 16], have been co-opted into mammalian gene regulatory networks [17, 18]. Identifying and annotating the genomic occurrence of homologous RNA structure motifs from sets of biologically related sequences will improve our understanding of the structure-function relationship of lncRNAs and the molecular mechanisms underlying their regulatory features. Resolving this challenge can be beneficial for the analysis of high-throughput RNA sequencing experiments that measure how RNAs interact with other molecules, such as crosslinked RNA immunoprecipitation or RNAse footprinting methodologies.

The identification of RNAs with similar functions involves comparing both their primary sequence and higher-order structures simultaneously. However, sequence-based methods to identify common structural features perform poorly when sequence identity falls below 60% [19]. Hence, methods are needed that find structural similarity independent from sequence conservation and freed from single RNA secondary structure predictions. The Sankoff algorithm resolves the optimal sequence-structure alignment of two RNAs [20], but its computational complexity limits its practicality. Its most comparable implementation, `FoldAlign`, employs a minimum free energy-based strategy with pruning of the associated dynamical programming matrix [21, 22]. Alternative strategies often employ pre-calculated secondary structure probability distributions (thermodynamically equilibrated canonical ensembles) for each sequence [23]. These can substantially speed up the calculation of structure-based alignments [24], of which there are many variants. The programs `PMcomp` [24], `LocaRNA` [25], and `ProbAlign` [26] use the pre-computed 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 in `RNApaln` [27] and `StrAL` [28], which derive probabilities for individual bases (such as the probability of being unpaired) from all base pairing probabilities. These methods all fail to explicitly consider suboptimal structures in the alignment. The pairwise alignment of entire base pairing probability matrices (RNA *dot plots*) was first introduced by `CARNA` [29, 30], which iteratively improves alignments using a constraint programming technique implementing a branch and bound scheme.

These pairwise RNA structure alignment algorithms can be used to identify clusters of homolgous RNA structure motifs from a set of sequences of interest. Will *et al.* 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 [25]. However, this strategy involves applying a subjective threshold to the resulting dendrogram to extract structurally related sequences. Alternative approaches to all-vs-all pairwise comparisons for RNA structure clustering include `NoFold`, which clusters query sequences based on their relative similarity to a panel of reference structure motif profiles [31], and `GraphClust`, an alignment-free approach that decomposes RNA structures into graph-encoded features [32]. `RNAscClust`, an extension of `GraphClust`, utilises the evolutionary signatures of RNA structures (when available) as an additional classification feature [33].

Here, we describe a computational pipeline for the identification and clustering of homologous RNA structures from a large set of query sequences. At its core lies `DotAligner`, a heuristic pairwise sequence alignment algorithm that considers the ensemble of base pair probabilities for each queried sequence. We benchmark the performance of `DotAligner` with other pairwise RNA structure alignment algorithms through several iterations of a stochastic sampling strategy across all RFAM seed alignments, highlighting the speed and accuracy of our method. We combine `DotAligner` with density based clustering for the unsupervised identification of RNA structural motifs, which can identify both known RFAM families and novel RNA structural motifs from ENCODE enhanced cross-linked immunoprecipitation (eCLIP) data. Finally, we exemplify how clusters of homologous RNA structures identified by our method can be used to search for homologous structures across reference genomes and transcriptomes to generate a map of functionally related RNA structure motifs.

Results

Ensemble-guided pairwise RNA structure alignment

We developed an algorithm that leverages the diversity of suboptimal solutions from a partition function of RNA alignments to identify an optimal sequence-structure alignment of two RNAs. The algorithm, termed `DotAligner`, overcomes the limitations of comparing unique RNA secondary structures (such as minimum free energy predictions) to yield a pairwise alignment that considers mutual base pair probabilities. A schematic of how `DotAligner` functions is illustrated in Figure 1.

`DotAligner` was developed with emphasis on runtime performance to facilitate all-vs-all pairwise comparisons of RNA structures on large data sets. Consequently, it uses pre-calculated RNA dot plots to perform alignments. It also makes use of the observation that a significant subset of stochastic sequence alignments between two RNAs will overlap the correct structure-based alignment, even though the optimal sequence alignment deviates significantly from the structural alignment [34]. The algorithm combines an alignment-envelope heuristic with a fold-envelope heuristic, which impose constraints on suboptimal sequence alignments and pre-calculated base pair probabilities, respectively. The alignment procedure consists of two steps, each considering base pair probabilities: (1) Generating a partition function of pairwise probabilistic string alignments; (2) Stochastic sampling of string alignments and scoring of aligned dot plots. Existing building blocks are integrated to `DotAligner` in a novel way. A `StrAL`-like score is applied during the dynamic programming in step 1, then a `CARNA`-like score is used to score the aligned dot plots in step 2, and, lastly, the partition function in step 1 and sampling in step 2 are adapted from `ProbA` [34]. The detailed implementation and mathematical description of `DotAligner` can be found in Additional file 1.

Evaluation of pairwise alignment quality

We first tested `DotAligner` on `BRAliBase 2.1` pairwise RNA structure alignments, a reference dataset specifically designed for algorithm benchmarking [19, 35] (see Methods). In this application, `DotAligner` seemingly performs worse than three other state of the art algorithms, namely `CARNA` [30], `FOLDALIGN` [36, 37] and `LocaRNA` [25], as well as the

Needleman-Wunsch pairwise sequence alignment algorithm, which ignores RNA structure content (Fig. 2). When comparing how well the algorithms perform in function of the pairwise sequence identity of BRAliBase 2.1 reference alignments, DotAligner produces alignments of lesser quality than comparable RNA structure alignment tools, particularly below 60% sequence identity, albeit with better accuracy than sequence-only alignments. Upon closer inspection, DotAligner outperforms the other tools around the 65-80% sequence identity range. As mentioned in the next section, this roughly corresponds to the average pairwise intra-family sequence identity of RFAM clans.

Interestingly, many of the pairwise structure alignments produced Structural Conservation Index (SCI) scores above those from the BRAliBase 2.1 reference alignments (Fig. 2). The SCI represents the alignment consensus energy normalised by the average energy of the single sequences folded independently [38]; it has been shown to be one of the most reliable metrics for conserved RNA structure detection [39]. With the exception of DotAligner, the other RNA structure alignment tools display, on average, SCI values above 0 in the 45-60% identity range, suggesting the underlying optimization algorithms tend to overestimate the amount of paired bases in consensus RNA structure predictions.

DotAligner's capacity to produce competitive pairwise alignments is demonstrated via a 5S-Adenosyl Methionine (SAM) riboswitch (RFAM clan RF00634, Additional file 2: Figure S1). In the RFAM alignment, the two representative sequences (AM420293_1 and CP000580_2_6) have a sequence identity of 55%. Pure sequence alignment increases this to 69%, but fails to align most structural features. DotAligner's pairwise probabilistic string alignment (step 1) creates an alignment of PID=56%, which is increased to PID=63% through DotAligner's sampling. The number of correctly aligned suboptimal base pairs increases via DotAligner's sampling. In this example, the alignment scores do not differ very much between DotAligner's optimal string alignment (step 1) and the best sample (step 2) (0.58 and 0.60, respectively), despite of a $\sim$25x increase of runtime through sampling (s=1000 in this example). As justified below, the benefits of sampling are outweighed by other properties of the algorithm.

Fast and accurate classification of RNA structures

The intended application of DotAligner is the identification and clustering 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-vs-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). DotAligner alignment scores were then compared to a binary classification matrix representing the distinct RFAM families (Additional file 2: Figure S3).

Despite the seemingly poor quality of pairwise alignments generated by DotAligner, it reproduces the known classification of RFAM structures more accurately, in general, than the other surveyed pairwise RNA structure alignment tools (Fig. 3 and Additional file 2:

Table S1). In fact, only when the average pairwise sequence identity drops below 55% for a given set of homologous RNA structures do the other algorithms perform comparably to DotAligner (Fig. 3C). Interestingly, the sequence alignments produced by Needleman-Wunsch are able to cluster RFAM sequences into their respective clans comparably well to more specialised RNA structure alignment tools, suggesting that most RFAM clans present sufficient stretches of local sequence identity to cluster them appropriately. Indeed, re-aligning the sequences from RFAM seed alignments based on their sequence alone, while permitting free end-gaps to evaluate local sequence similarity, shifts the median pairwise sequence identity from 59% to 72% (Additional file 2: Figure S4).

The efficacy of the heuristics implemented in DotAligner are further accentuated by its runtime, which consistently lies between simple sequence alignment and more sophisticated RNA structure alignment algorithms (Fig. 3C and Additional file 2: Fig. S5). The impact of sequence length does not correlate with AUC scores, but it increases runtime exponentially (Additional file 2: Fig. S6).

Density-based clustering of homologous RNA structures

Given DotAligner's accurate clustering of known structured RNA using binary classification, 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 [25]. However, this does not apply a cut-off that can be used to accurately extract novel 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 [40], 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 significantly greater sequence-structure homology). The OPTICS (Ordering Points To Identify the Clustering Structure) algorithm [41] 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) [42] algorithm that, as its name states, is suitable for noisy data, such as RNA immunoprecipitation followed by high-throughput sequencing (RIPseq). We benchmarked the two main OPTICS clustering parameters–Xi steepness threshold and the minimum number of points in a cluster (Additional file 2: Figure S7)–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 versus all pairwise alignments with DotAligner, we evaluated 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 (Additional file 2: Figure S7A).

In comparison to GraphClust, the combination of DotAligner and OPTICS performs comparably well (Fig. 4, Table 1, Additional file 2: Table S2). The default version of NoFold nonetheless outshines DotAligner at clustering known RFAM families. However, it intrinsically employs RFAM covariance models (CMs) that are also present in the test

data, therefore this specific application is likely to be subject to over-fitting. We thus removed 72 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, while its specificity increased slightly despite removing 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 [43], with 100 positive control sequences from RFAM (Additional file 2: Table S3). From 2,650 high-confidence (>8-fold fold-enrichment versus background, P-value $< 10^{-4}$) eCLIP peaks that overlap evolutionarily conserved secondary structure predictions, 25 significant clusters of homologous RNA were detected, including all 11 positive controls (Fig. 5).

Indeed, the *spike in* RFAM sequences facilitate the identification of similar RNA structures, such as the homologs to SNORNA72 depicted in (Fig. 5C-D). The 4 additional sequences that co-cluster with SNORNA72 controls are all associated to the DKC1 protein, which binds to H/ACA snoRNAs. Furthermore, 3 of the DKC1-bound peaks are annotated as snoRNAs in the Gencode 24 reference, while the 4th is not annotated as a snoRNA despite strong sequence and structure similarity, highlighting how this method can successfully identify and cluster new RNA structure motifs. Another example is the Y RNA cluster, which contains 3 sequences homologous to this RFAM family that are also associated to the TROVE2 protein, which binds to misfolded non-coding RNAs, pre-5S rRNA, and Y RNAs.

Our method also identifies RNA structure families impartially, as exemplified by several clusters of DKC1-associated sequences which present consensus secondary structures indicative of snoRNAs (Fig. 5E). Closer inspection of the corresponding eCLIP peaks revealed that these sequences are indeed annotated as snoRNAs in Gencode. There are also examples of de Novo structural motifs that are associated to RNA-binding proteins with no previously known binding sites, such as an UPF1-dominated cluster (Fig. 5F), composed of a structure motif belonging to ALU repeats (Additional file 2: Figure S8). When searching the human genome for homology to the RNA structure motif derived from this cluster, most ALU elements are detected, as well as a few other repeat-containing sequences. Interestingly, 998 homologs to the motif did not localise to ALU elements (Additional file 2: Figure S8C-D), 58% of which overlap miTranscriptome reference transcripts [44].

Discussion

The increasing accessibility of next generation sequencing and immunoprecipitation protocols provides large resources for in-depth transcriptome and interactome profiling. Elucidating the structural features of RNAs associated to RNA-binding proteins and ribonucleoprotein (RNP) complexes, combined with the systematic classification of their genome-

or transcriptome-wide occurrence, can identify recurrent functional motifs that may form components of regulatory networks. Pragmatically, the method we describe facilitates this process by enabling rapid and unsupervised clustering of RNA structure motifs with reasonable accuracy. We also show that clustering eCLIP sequences can identify new RNA structures and their homologs throughout the genome (Additional file 2: Figure S8A-C), which can be used to assign putative functions to non-coding loci and categorize them accordingly.

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, the strength of `DotAligner` lies in its capacity to accurately score structurally homologous RNA sequences and the suboptimal structural landscape of RNAs, reducing several dimensions of information into a single discriminative numeric value. Our results show that this can 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 immunoprecipitation data–for common structural motifs.

The algorithm generates pairwise alignments that differ qualitatively to reference structural alignments at lower ranges of sequence identity, but performs better than more complex algorithms within ranges of sequence similarity that substantially overlap those of functionally related RNAs, as presented in RFAM. This could be a consequence of refining the runtime parameters through testing on independently and stochastically sampled RFAM sequences; it is not impossible that other algorithms could undergo comparable parameter optimisation. However, the significantly higher computational complexity of other related tools compared to our method make it fairly difficult (and resource-intensive) to perform such brute-force parameter optimisation.

High-throughput CLIPseq data poses a challenge for consensus motifs detection since several molecules that are in close physical proximity to the target molecule get co-precipitated together. Consequently, other RNA sequences may be present that do not directly bind to the target protein. We have shown that our method is nonetheless suitable for such noisy biological data. For example, the UPF1 cluster we describe may be an example of an indirect binding event, as UPF1 directly interacts with STAU1, a double stranded RNA-binding protein which has been reported to target ALU sequences [45]. Other clusters identified in our eCLIP analysis cluster together sequences from more than one target protein, which raises the possibility that a common RNA structure motif may be bound by different proteins, either as part of a quaternary complex or as a common, competing binding target. We privilege this hypothesis over one of spurious false-positive clustering given our benchmark results and the observation that very few clusters were observed when analysing less stringently filtered eCLIP peaks (data not shown).

`DotAligner` has several variables that will influence the clustering results and speed depending on the type of input data. The most influential variables are the weight between sequence and structure similarity, and the exploration depth of suboptimal alignments in the stochastic backtracking. We have shown that stochastic sampling of suboptimal string

alignments improves the alignment of RNA dot plots. However, the performance increase does not outweigh the increase in runtime associated with sampling suboptimal sequence alignments. Our RFAM clustering benchmark using a binary classification strategy has shown that the best trade-off between alignment accuracy and speed comes at the abandonment of sampling, as supported by the de novo structures identified from the ENCODE eCLIP data. Future optimization of `DotAligner` parameters will likely increase its usability. For example, dynamic parameters could be implemented that adjust the degree of sampling diversity and number of samples based on the sequence identity obtained from step 1 of `DotAligner`. This could tune the algorithm's performance based on the nature of the input, potentially improving `DotAligner`'s performance across all ranges of sequence identity. Another potential enhancement could be achieved in the stochastic sampling by only considering elements of the ensemble with probabilities larger than a threshold. By doing so we could (1) reduce the number of useless samples, (2) guarantee that cells of high probability are passed (suboptimal structures), and (3) leave time/samples to explore the ensemble space (slightly modified alignments by limiting sample diversity) around these suboptimals.

Another great challenge lies in the accurate depiction of RNA structure motif boundaries. Whereas global structures may stabilize the RNA molecule, local structural domains are often sufficient for recognition by RNA binding proteins. A strategy to find optimal local alignments would be desirable for this purpose. `DotAligner` can search for semi-local alignments by introducing penalty-free gaps at the sequence extremities (N.B., `LocaRNA` also supports this functionality). In this study we did not investigate in the optimization of these local pairwise similarity scores, because they may miss parts of the functional units (RNA structure) and, hence, hinder the search for optimal clusters. Instead, we circumvented this issue by overlapping eCLIP peaks to evolutionarily conserved RNA secondary structure predictions with well-characterised flanking helices supported by base pair covariation [6]. While preparing this manuscript, a complementary and comprehensive dataset of evolutionarily conserved RNA secondary structures was published [46]. Its application can further increase the amount of eCLIP peaks with accurate structural motif boundaries. Alternatively, RNA structure boundaries can be refined by, for example, using alternative strategies such as computational boundary refinement with `LocaRNA-P` [47], or improving the biological data with enzymatic probing with the double-stranded RNase T1 endoribonuclease.

Conclusion

An efficient pairwise RNA sequence alignment heuristic, which intrinsically considers suboptimal base pairings, accurately discriminates between distinct structured RNA families. When combined with a noise tolerant density based clustering algorithm, this approach identifies known and novel RNA structure motifs from a set of input sequences without a priori knowledge. The resulting RNA structure motifs are subsequently used to identify homologs in the human genome, improving the annotation of long non-coding RNAs and increasing the repertoire of functional genetic elements.

Methods

Benchmarking and parameter optimisation

The `DotAligner` algorithm implements several 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 [35]. We first tested all combinations of the following parameters: ***k*** and *t* from 0 to 1 in increments of 0.1; ***o*** and ***e*** from 0.2 to 1 in increments of 0.2. 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 [39] based on Minimum Free Energy (MFE), and the Matthews Correlation Coefficient (MCC) of *predicted* and *reference* RNA secondary structure were also calculated for all resulting alignments:

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

$$\Delta\text{SCI} = SCI_{predicted}/SCI_{reference}, \text{ where SCI} = MFE_{consensus}/\overline{MFE}_{single}$$

Baseline parameters were then selected via a product rank of these 2 metrics and subjected to refinement using a binary classification approach, described in the next section.

Binary classification of RNA secondary structure families

Further refinement of the optimal parameters was performed using a binary classifier for two sets of 200 stochastically sampled RFAM entries with published structures: (i) a low Pairwise Sequence Identity (PSI) set, and (ii) a high PSI set, where any two sequences from the same family share between 0-55% and 56-95% PSI, respectively (Fig. 3A-B). The `JAVA` implementation of this algorithm, derived from [6], can be found in the Additional file 1. Further investigation of the impact of local sequence similarity on algorithmic performance was done by sampling all seed alignments of RFAM version 12.3 via 3 replicates of our stochastic sampling procedure. The sequences were then striped of gaps and pseudo-knots (not present in the preliminary RFAM version 12.0 alignments), and realigned with a variant of Needleman-Wunsch enabling free end gaps. The samplings were limited to 5 ranges of sequence identity, as presented in Fig. 3C.

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 by the R package `pROC` [48]. A more restricted range of parameter values were then tested on both datasets, where a ranked sum for both datasets of the AUC was performed to determine the default runtime parameters for `DotAligner`, namely $\theta = 0.5$, $\kappa = 0.3$, $g_o = 1$, and $g_{ext} = 0.05$ (Additional file 2: Table S4). Parameter θ (or -t in the command line) is the weight of sequence similarity compared to similarity of unpaired probabilities, κ (or -k) is the weight between sequence-based similarity and dot plot similarity, g_o (or -go) is gap opening penalty, and g_{ext} (-go) is gap extension penalty. Sampling specific parameters s (number of samples) and T (measure of sampling diversity) had minimal impact in

the refined parameter optimization from sampled RFAM clans, although the parameters can increase alignment scores in low and medium pairwise sequence identity ranges (Additional file 2: Figures S1 and S2A). We also show that in average the alignment score saturates after 1000 samples of the stochastic backtracking for $T = 0.25$ (Additional file 2: Figure S2B). `CARNA` version 1.2.5 was run with parameters "–write-structure –noLP –time-limit=120000" ; `LocaRNA` version 1.7.13 was run with parameter "–noLP"; `FOLDALIGN` version 2.1.1 was run with and without parameter "-global". Default parameters were used for `pmcomp`, downloaded from https://www.tbi.univie.ac.at/RNA/PMcomp/ and `RNApaln` version 2.3.5. The custom implementation of Needleman-Wunsch can be found in the GitHub repository associated to this work as well as the benchmark implementation scripts.

Clustering RNA structures with randomised controls

OPTICS benchmarking was performed by stochastically sampling the collection of RFAM 12.0 seed alignments using the `JAVA` program `GenerateRFAMsubsets.java` (see Additional file 1) with three ranges of pairwise sequence identity: 1-55%, 56-75%, and 75-95%, a minimum of 5 representative sequences per family, and sizes ranging between 70 and 170 nt. The resulting 580 unique sequences were then shuffled while controlling their dinucleotide content with the `easel` program included in the `Infernal` (v1.1.2) software package [49] with option "-k 2". The 1160 sequences were submitted to all-vs-all pairwise comparisons with `DotAligner` and the scores were inverted and 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 Fig. 4A,B).

Other tested RNA clustering approaches were `GraphClust` and `NoFold`. We ran `GraphClust` version 0.7.6 inside the docker image provided with `RNAscClust` with default parameters. `NoFold` version 1.0.1 uses 1,973 RFAM covariance models by default as empirical feature space. In the `NoFold` (all CMs) variant we ran the program with default parameters, whereas in the `NoFold` (filtered) variant we reduced the feature space to 1,902 covariance models by removing RFAM families from our benchmark set.

The following clustering performance metrics was used: 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 $\geq 8\times$ 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 50 nt to avoid over-representing the same sequence (Additional file 1). The remaining peaks were subsequently filtered by retaining only those that present same-strand overlap with any evolutionarily conserved structure (ECS) predictions from [6]. Finally, the associated genomic sequences were extracted into a .fasta file, which was supplemented with 100 reference RNA structures from 11 RFAM families (see Additional file 2: Table S3). 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`'.

Availability of data and materials
Source code, pipelines and data can be obtained at https://github.com/noncodo/BigRedButton.

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

Funding
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).

Author's contributions
MAS, SES and JSM conceived the study. SES and MAS wrote the manuscript and performed data analysis. MAS performed benchmarking analyses and developed analytic pipelines. SES created and implemented `DotAligner` source code. XQ assisted in `DotAligner` parameter optimisation and benchmarking.

Acknowledgements
We would like to thank DR Eva Maria Novoa Pardo for her scientific council, R programming tricks, and critical manuscript reading. Thanks to Prof Oliver Mülhemann for discussions relating to data interpretation. Thanks to Luis Renato Arriola-Martinez for his advice concerning covariance models. Thanks to Michael Hahsler for developing and disseminating a fast DBSCAN R package.

Author details
[1]RNA Biology and Plasticity Group, Garvan Institute of Medical Research, 384 Victoria Street, NSW 2010 Sydney, Australia. [2]St Vincent's Clinical School, Faculty of Medicine, UNSW Australia, , 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. 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)
4. Eddy, S.R.: Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annual review of biophysics **43**, 433–456 (2014)
5. 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)
6. Smith, M.A., Gesell, T., Stadler, P.F., Mattick, J.S.: Widespread purifying selection on RNA structure in mammals. Nucleic acids research, 596 (2013)
7. 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)
8. 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)
9. 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

10. Hogg, J.R., Collins, K.: Structured non-coding RNAs and the RNP Renaissance. Current opinion in chemical biology **12**(6), 684–689 (2008)
11. Rinn, J.L., Chang, H.Y.: Genome regulation by long noncoding RNAs. Annual review of biochemistry **81**, 145–166 (2012)
12. 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)
13. 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)
14. 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)
15. Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., Feschotte, C.: Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS genetics **9**(4), 1003470 (2013)
16. Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., Ulitsky, I.: Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell reports **11**(7), 1110–1122 (2015)
17. Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., Bourque, G.: Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature genetics **42**(7), 631–634 (2010)
18. Kelley, D., Rinn, J.: Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome biology **13**(11), 107 (2012)
19. 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
20. Sankoff, D.: Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. **45**, 810–825 (1985)
21. 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
22. 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
23. 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
24. 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
25. 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
26. 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
27. 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
28. 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
29. 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
30. 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
31. 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
32. 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
33. 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
34. Muckstein, U., Hofacker, I.L., Stadler, P.F.: Stochastic pairwise alignments. Bioinformatics **18 Suppl 2**, 153–60 (2002)
35. Wilm, A., Mainz, I., Steger, G.: An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms for molecular biology **1**(1), 1 (2006)
36. 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)
37. Sundfeld, D., Havgaard, J.H., de Melo, A.C., Gorodkin, J.: Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. Bioinformatics, 748 (2015)
38. 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)
39. 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)
40. Suzuki, R., Shimodaira, H.: Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics **22**(12), 1540–1542 (2006)
41. Ankerst, M., Breunig, M., Kriegel, H., *et al.*: Ordering points to identify the clustering structure. In: Proc. ACM SIGMOD, vol. 99 (1999)
42. 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)
43. 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)
44. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., *et al.*: The landscape of long noncoding RNAs in the human transcriptome. Nature genetics **47**(3), 199–208 (2015)
45. Gong, C., Maquat, L.E.: LncRNAs transactivate Staufen1-mediated mRNA decay by duplexing with 3'UTRs via Alu elements.

Nature **470**(7333), 284 (2011)
46. Seemann, S.E., Mirza, A.H., Hansen, C., Bang-Berthelsen, C.H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C.T., Pociot, F., et al.: The identification and functional annotation of RNA structures conserved in vertebrates. Genome Research, 208652 (2017)
47. Will, S., Joshi, T., Hofacker, I.L., Stadler, P.F., Backofen, R.: LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. Rna **18**(5), 900–914 (2012)
48. 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)
49. Nawrocki, E.P., Eddy, S.R.: Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**(22), 2933–2935 (2013)
50. Gotoh, O.: An improved algorithm for matching biological sequences. J Mol Biol **162**(3), 705–708 (1982)
51. 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 the similarity in sequence and cumulative base-wise pairing probabilities (top left–colour intensity indicates cumulative similarity score). A partition function over all pairwise alignments is calculated and interrogated for structural compatibility by stochastic backtracking. Two ensembles over all secondary structures are considered for this purpose (bottom left and top right dot plots–blue lines indicate cumulative base-wise pairing probabilities). The final scoring makes usage of the base pair probabilities in the dot plots. 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.

Figure 2: **Comparison of RNA structure alignment quality in function of sequence identity.** `BRAliBase 2.1` reference RNA structure alignments were submitted to 5 different pairwise alignment algorithms, including the Needleman-Wunsch sequence-only alignment algorithm. **(Top)** The total number of surveyed alignments in function of pairwise sequence identity. The Matthews Correlation Coefficient (MCC), difference in Structural Conservation Index (Δ-SCI), and `RNAdistance` calculated topological edit distance between the `RNAalifold` consensus of the computed alignment and the reference `BRAliBase 2.1` alignment consensus are compared in the lower 3 plots.

Tables

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

Figure 3: **Classification of known RNA structures.** (**A**) Receiving Operator Characteristic (ROC) curves measuring the classification accuracy of surveyed algorithms by contrasting their computed similarity matrices to a binary classification matrix of RFAM sequences (1 if sequences are in same family; 0 if different). High PID = 56-95% pairwise sequence identity from the provided RFAM alignment; Low PID = 1-55%. (**B**) Precision-recall curve; (**C**) Area Under the Curve (AUC) of ROC values with 95% confidence intervals for the top 4 performing algorithms across 5 ranges of pairwise sequence identity, as calculated from a variant of the Needleman-Wunsch algorithm with free end gaps. The 3 replicates correspond to stochastically sampled sequences from RFAM 12.3 (see Additional file 2: Table S1); (**D**) Runtime distribution of single thread computation on a 2.6 GHz AMD Opteron processor (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,B**) Reachability plots of OPTICS clustering display the OPTICS-derived ordering of points (x axis) and their distance to the nearest neighbour (y axis). Colours represent significant clusters: (**A**) Clustering of RFAM benchmarking data indicating the distance to the nearest neighbour, (**B**) Clustering of 2,650 ENCODE eCLIP peaks + 100 RFAM controls that overlap evolutionarily conserved secondary structure predictions. The dominant RNA binding protein in each cluster is displayed next to significant clusters; those with an asterisk are 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. (**D-F**) `RNAalifold` predicted consensus RNA secondary structures: (**D**) Structure of the alignment displayed in (**C**), (**E**) Structure of a cluster of impartially detected DKC1-bound snoRNAs, and (**F**) Structure of a novel UPF1-bound motif.

Additional files

Additional file 1 — Supplementary methods
Link to GitHub repository, detailed description of the `DotAligner` implementation, RNA structure clustering, and eCLIP data processing.

Additional file 2 — Supplementary tables and figures

Additional file 1 — Supplementary methods

The pipelines, scripts and some programs used in this work can be found in the GitHub repository associated to this work at https://github.com/noncodo/BigRedButton.

1. DotAligner implementation

The weight W of alignment A of two arc-annotated sequences (S_a, P_a) and (S_b, P_b) has been defined by [29] as

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

where S is a sequence and P is a base pair 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 practise heuristics are used to find near-optimal solutions.

DotAligner solves the related problem of aligning two base pair 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 [25]. 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 pair probabilities, to build pairwise sequence-structure alignments.

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

1 Partition function of pairwise probabilistic string alignments;
2 Stochastic sampling of string alignments and scoring of aligned dot plots.

1.1 Partition function of pairwise probabilistic string alignments

In step 1 the computation of the partition function $Z(T)$ over all canonical pairwise string alignments A is adapted from probA [34]:

$$Z(T) = \sum_A exp(\beta W(A)), \tag{2}$$

where $\beta = 1/T$. Then the probability of a specific alignment A is defined as:

$$Pr(A;T) = \frac{1}{Z(T)} exp(\frac{W(A)}{T}), \tag{3}$$

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 uses the dynamic programming algorithm of Gotoh [50] which has running time of $O(N^2)$. 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 [28]. Hence, step 1 considers only the similarity σ and the gap cost γ described in equation 1:

$$W_{\text{Step1}}(A) = \sigma(A) + \gamma(A) \tag{4}$$

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:

$$\sigma(i,i') = \theta \times M_{Seq}^{(i,i')} + (1-\theta) \times \Delta\omega(i,i') \tag{5}$$

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

$$\Delta\omega(i,i') \in \begin{cases} 0 & \text{if } \omega(i) == 0 \\ & \text{and } \omega(i') == 0 \\ 1 - |\omega(i) - \omega(i')| & \text{else} \end{cases} \tag{6}$$

so that $\Delta\omega = (0,1)$.

The gap term in equation 1 is replaced with affine gap costs:

$$\gamma(A) = l \times g_o + (N-l) \times g_{ext} \tag{7}$$

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 can be considered as free (set parameter `--free-endgaps`).

1.2 Stochastic sampling of string alignments and scoring of aligned dot plots

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 [34] for selecting s suboptimal string alignments A_s. The combined score (weight) W_{Step2} is a variant of equation 1 to explore the similarity of the corresponding dot plot alignments:

$$W_{\text{Step2}}(A_s) = \kappa \times \frac{W_{\text{Step1}}(A_s)}{|A_s|} + (1-\kappa) \times \frac{\tau(A_s)}{|\text{Match}_{A_s}|^2} \tag{8}$$

where the parameter κ weights for each alignment A_s between the sequence-based similarity $W_{\text{Step1}}(A_s)$ normalised by alignment length $|A_s|$ and dot plot similarity $\tau(A_s)$ normalised by the number of aligned bases $|\text{Match}_{A_s}|$ in alignment A_s. Similar to equation 5 the dot plot similarity τ sums the parameter θ weighted similarity of aligned base pairs M_{paired} and the similarity in their pairing probabilities $\Delta\psi$:

$$\tau(i,j,i',j') = \theta \times M_{paired}^{(i,j,i',j')} + (1-\theta) \times \Delta\psi(i,j,i',j') \tag{9}$$

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 base pairs (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 calculated by

$$\Delta\psi(i,j,i',j') \in \begin{cases} 0 \text{ if } \psi(i,j) == 0 \text{ and } \psi(i',j') == 0 \\ 1 - |\psi(i,j) - \psi(i',j')| \text{ else} \end{cases} \tag{10}$$

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 [25]:

$$\psi(i,j) = max\left(0, log\frac{P(i,j)}{p_0} / log\frac{1}{p_0}\right) \tag{11}$$

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 base pair probabilities are compared, but stronger similarities for high base pair probabilities. Unpaired probabilities are handled in a similar way by

$$\omega(i) = max\left(0, log\frac{1-\sum_k P(i,k)}{p_0} / log\frac{1}{p_0}\right) \tag{12}$$

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

1.3 Alternative model using substitution rates

Alternatively, the sequence and base pair similarities M_{Seq} and M_{paired} in equations 5 and 9 can be replaced by the statistical substitution models R_{Seq} and R_{paired}, respectively. In this (non-default) model of `DotAligner` R_{Seq} is multiplied with the ζ weighted sum of the similarity of unpaired probabilities $\Delta\omega$ and the similarity of upstream pairing probabilities $\Delta\omega^{up}$ (set parameter `--mutation-rates`):

$$\begin{aligned}\sigma(i,i') = R_{Seq}^{(i,i')} \times \zeta \times \Delta\omega(i,i') + \\ R_{Seq}^{(i,i')} \times (1-\zeta) \times \Delta\omega^{up}(i,i')\end{aligned} \tag{13}$$

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 [51] 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

$$\omega^{up}(i) = \sum_{k=1}^{i-1} \psi(k,i) / \sum_{k=1}^{|S|} \psi(k,i) \tag{14}$$

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 base pair similarity matrix M_{paired} can be replaced by a statistical substitution model R_{paired} which describes the probability for observing a given base pair substitution relative to background nucleotide frequencies:

$$\tau(i,j,i',j') = R_{paired}^{(i,j,i',j')} \times \Delta\psi(i,j,i',j') \tag{15}$$

The log-odd scores L from the RIBOSUM85-60 matrix [51] are transformed to probabilities R_{paired} by $2^{L(i,j,i',j')}/(1+2^{L(i,j,i',j')})$.

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.

```
## R code -- R version 3.4.1
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

```
#!/bin/bash

# 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
    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
```

Additional file 2 — Supplementary Tables and Figures

Table S1. Uniqueness and diversity of stochastically sampled RFAM subsets

			# RFAM families		
Pairwise identity range	# sequences	% unique	Rep. 1	Rep. 2	Rep. 3
0-55	178	94.4	33	19	32
56-65	900	92.2	113	108	110
66-75	899	92.6	83	76	74
76-85	900	91.7	80	82	79
86-99	900	93.6	58	47	59

Table S2. 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

Table S3. 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

Table S4. Rank-product of best `DotAligner` parameters

Parameters	low_PI rank	high_PI rank	rank product	low_PI AUC	high_PI AUC	AUC sum	Combined Time
k=0.3 t=0.5 o=1 e=0.05	1	112	112	0.983297903	0.996178994	1.97948	0.140273
k=0.3 t=0.8 o=1 e=0.05	181	1	181	0.959342489	0.997188985	1.95653	0.133496
k=0.3 t=0.5 o=1 e=0.05	2	110	220	0.983297903	0.996178994	1.97948	0.135262
k=0.3 t=0.5 o=1 e=0.05	3	109	327	0.983297903	0.996178994	1.97948	0.134188
k=0.3 t=0.8 o=1 e=0.05	184	2	368	0.959342489	0.997188985	1.95653	0.144565
k=0.3 t=0.5 o=1 e=0.05	4	113	452	0.983297903	0.996178994	1.97948	0.150288
k=0.3 t=0.8 o=1 e=0.05	182	3	546	0.959342489	0.997188985	1.95653	0.142137
k=0.3 t=0.5 o=1 e=0.05	5	114	570	0.983297903	0.996178994	1.97948	0.156738
k=0.3 t=0.5 o=1 e=0.05	6	111	666	0.983297903	0.996178994	1.97948	0.155101
k=0.3 t=0.8 o=1 e=0.05	185	4	740	0.959342489	0.997188985	1.95653	0.146729
k=0.3 t=0.5 o=1 e=0.05	7	115	805	0.983297903	0.996178994	1.97948	0.186257
k=0.3 t=0.5 o=1 e=0.05	8	116	928	0.983297903	0.996178994	1.97948	0.192388
k=0.3 t=0.8 o=1 e=0.05	186	5	930	0.959342489	0.997188985	1.95653	0.154183
k=0.3 t=0.5 o=1 e=0.05	9	117	1053	0.983297903	0.996178994	1.97948	0.210514
k=0.3 t=0.8 o=1 e=0.05	183	6	1098	0.959342489	0.997188985	1.95653	0.154234
k=0.3 t=0.5 o=1 e=0.05	10	119	1190	0.983297903	0.996178994	1.97948	0.285647
k=0.4 t=0.6 o=1 e=0.05	13	97	1261	0.983273039	0.996343919	1.97962	0.133738
k=0.3 t=0.5 o=1 e=0.05	11	118	1298	0.983297903	0.996178994	1.97948	0.269801
k=0.3 t=0.8 o=1 e=0.05	187	7	1309	0.959342489	0.997188985	1.95653	0.187293
k=0.4 t=0.6 o=1 e=0.05	14	101	1414	0.983273039	0.996343919	1.97962	0.144514

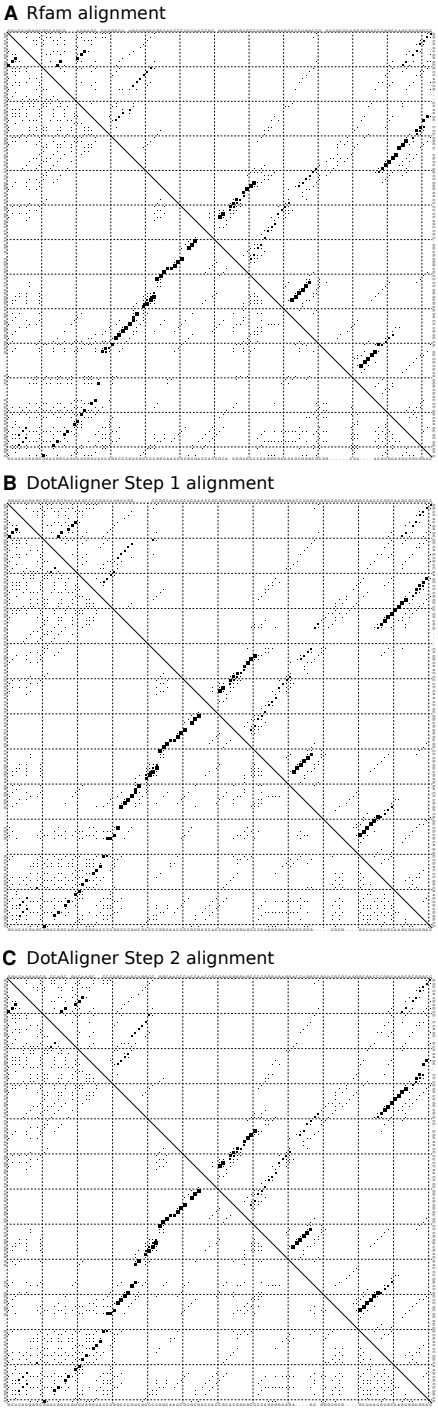

Figure S1. Pairwise alignments of dot plots of SAM riboswitch.
The two sequences AM420293_1 (upper triangles of the dot plots) and CP000580_2_6 (lower triangles) of the 5S-adenosyl methionine (SAM) riboswitch (Rfam family RF00634) are aligned (**A**) as in the Rfam reference alignment, (**B**) through `DotAligner`'s pairwise probabilistic string alignment (step 1), and (**C**) through `DotAligner`'s sampling of stochastic alignments (step 2). `DotAligner`'s sampling increases the combined score of the alignment from 0.58 to 0.60 (and the sequence identity from 56% to 63%), and improves the quality of the alignment compared to the Rfam reference.

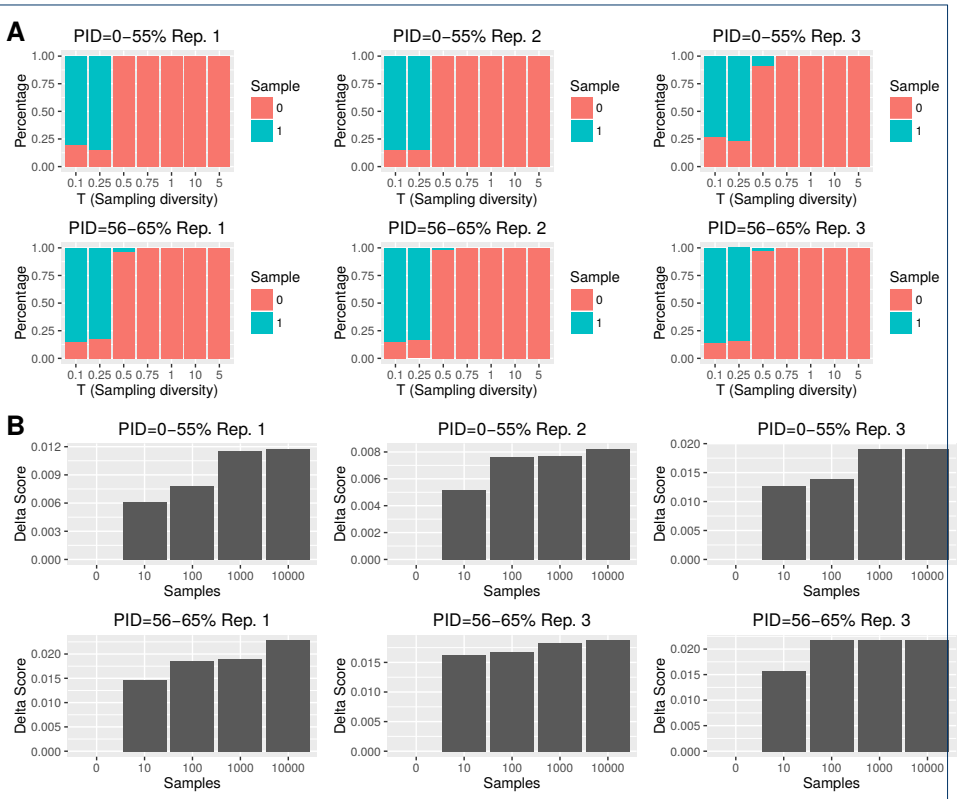

Figure S2. Impact of sampling of stochastic alignments on the alignment score.
We executed `DotAligner` (default runtime parameters) while (**A**) varying parameter T (sampling diversity) with 1000 samples (parameter s) and (**B**) varying the number of samples with $T = 0.25$ on the RFAM binary classification benchmark datasets corresponding to 0-55% and 56-65% sequence identity (PID) (3 replicates each). For parameter T equal 0.1 and 0.25 the majority of pairwise alignments are optimized through the sampling procedure (sample increased the alignment score) (**A**). In few cases, T=0.5 also produced an optimized alignment through sampling. In average the alignment score saturates after 1000 samples of the stochastic backtracking for $T = 0.25$ (**B**).

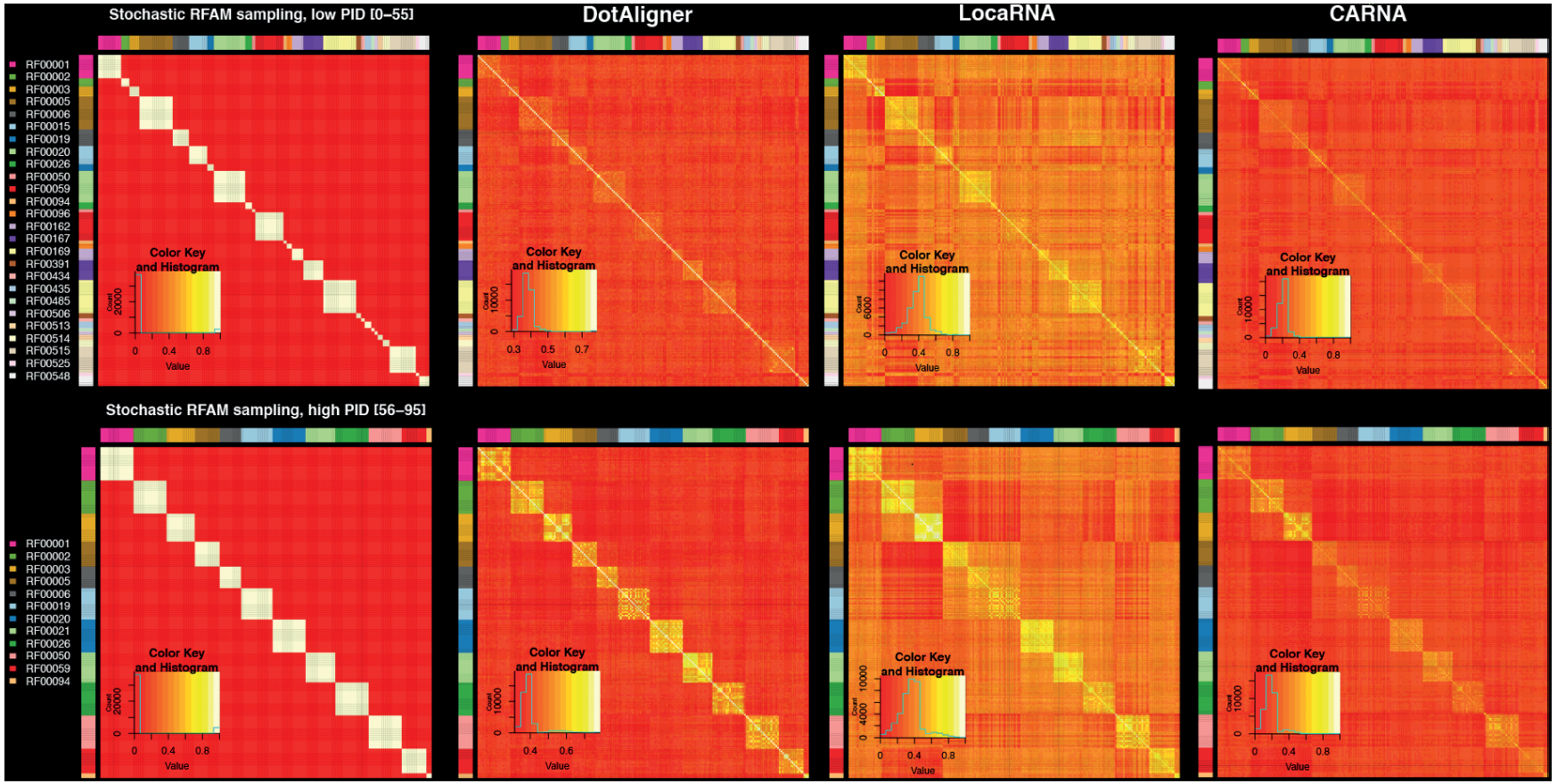

Figure S3. Binary classification of similarity matrices

Stochastically sampled RFAM version 12.0 sequences are labelled as belonging to the same family in white, and in red when not (Left). Heat maps of the similarity matrices produced by `DotAligner`, `LocaRNA` and `CARNA` are listed in columns 2, 3 and 4, respectively. (top) Low mean pairwise identity samples, where each sequence within a family shares between 0 and 55% sequence identity; (bottom) Higher (56-95%) mean pairwise identity samples.

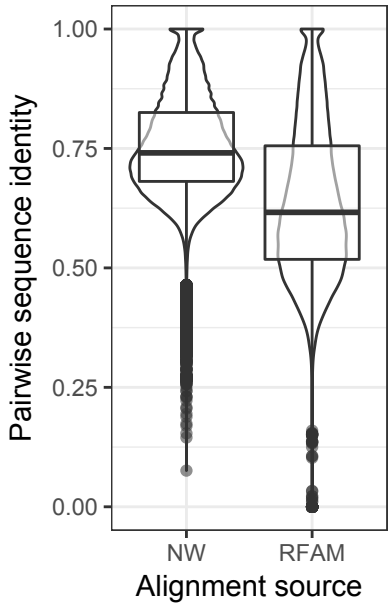

Figure S4. Difference in sequence identity between structural and sequence alignments

The difference in pairwise sequence identity for 1,189,675 randomly sampled RFAM version 12.3 seed alignments is shown for sequence-only alignments using a variant of the Needleman-Wunsch algorithm permitting free end gaps (NW) and the native RFAM seed alignments. Only sequences within the same family are compared, exposing the presence of local sequence similarity within the sequences. Pairwise sequence identity is defined by the number of matching nucleotides divided by the length of the shortest sequence.

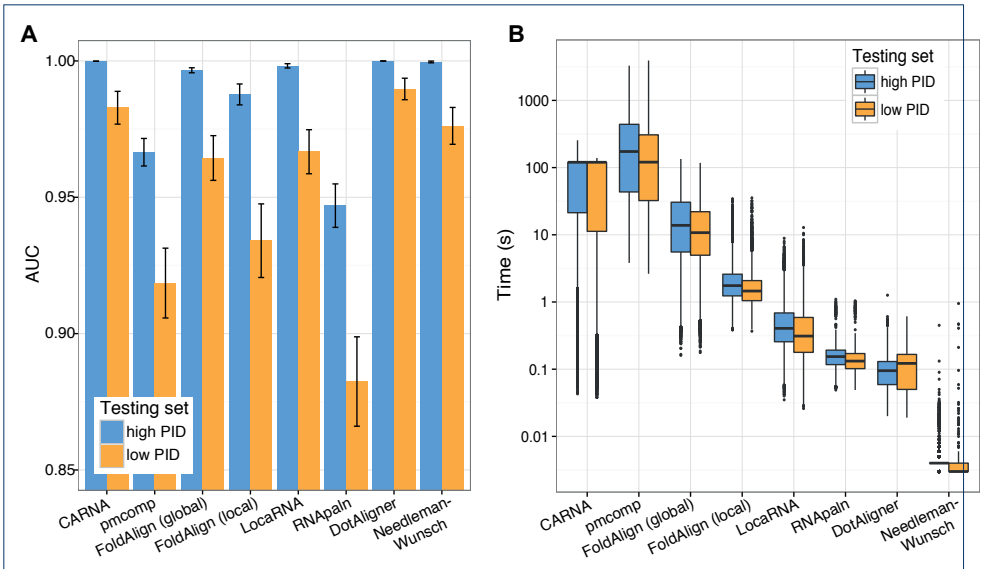

Figure S5. Classification of known RNA structures. (**A**) Area Under the Curve (AUC) of ROC values with 95% confidence intervals associated to Figure 3A. (**B**) Associated runtime distribution of single thread computation on a 2.6 GHz AMD Opteron processor (N.B. a fixed upper limit of 120 s was imposed for `CARNA`).

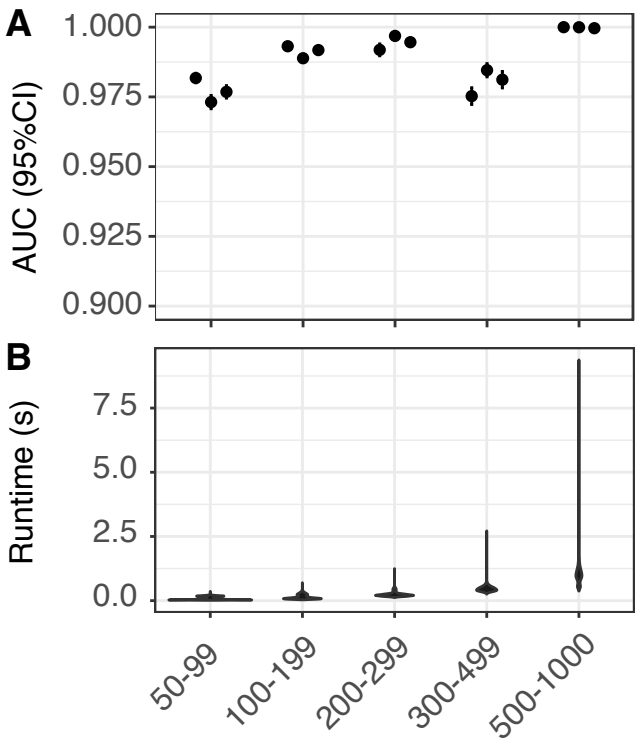

Figure S6. `DotAligner` clustering performance in function of sequence length. (**A**) Area Under the Curve (AUC) of ROC values with 95% confidence intervals for 3 replicates of stochastically sampled RFAM version 12.3 clans, controlling for sequence length (x-axis). N.B. the 500-1000 set only includes between 17-20 sequences given their rarity in the RFAM datasets, compared to 299-300 for the other samples. (**B**) Associated runtime distribution of single thread computation on a 2.6 GHz AMD Opteron processor.

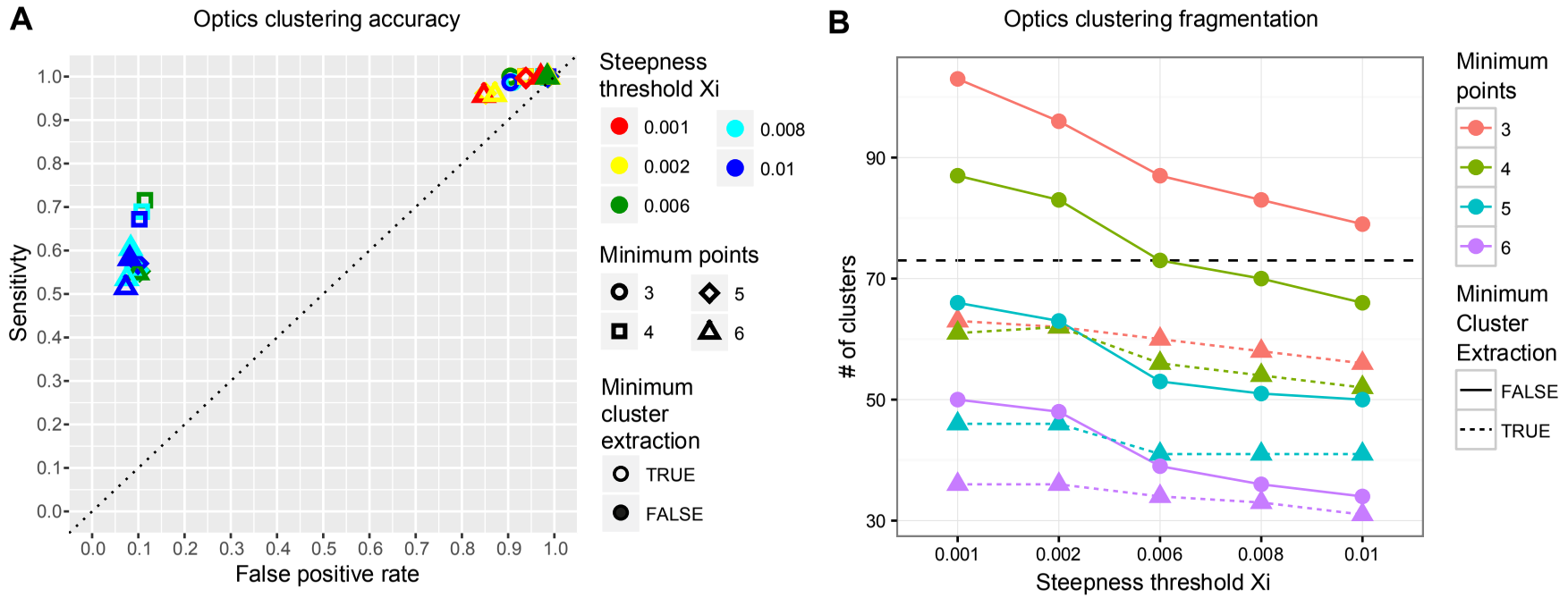

Figure S7. OPTICS clustering optimisation

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).

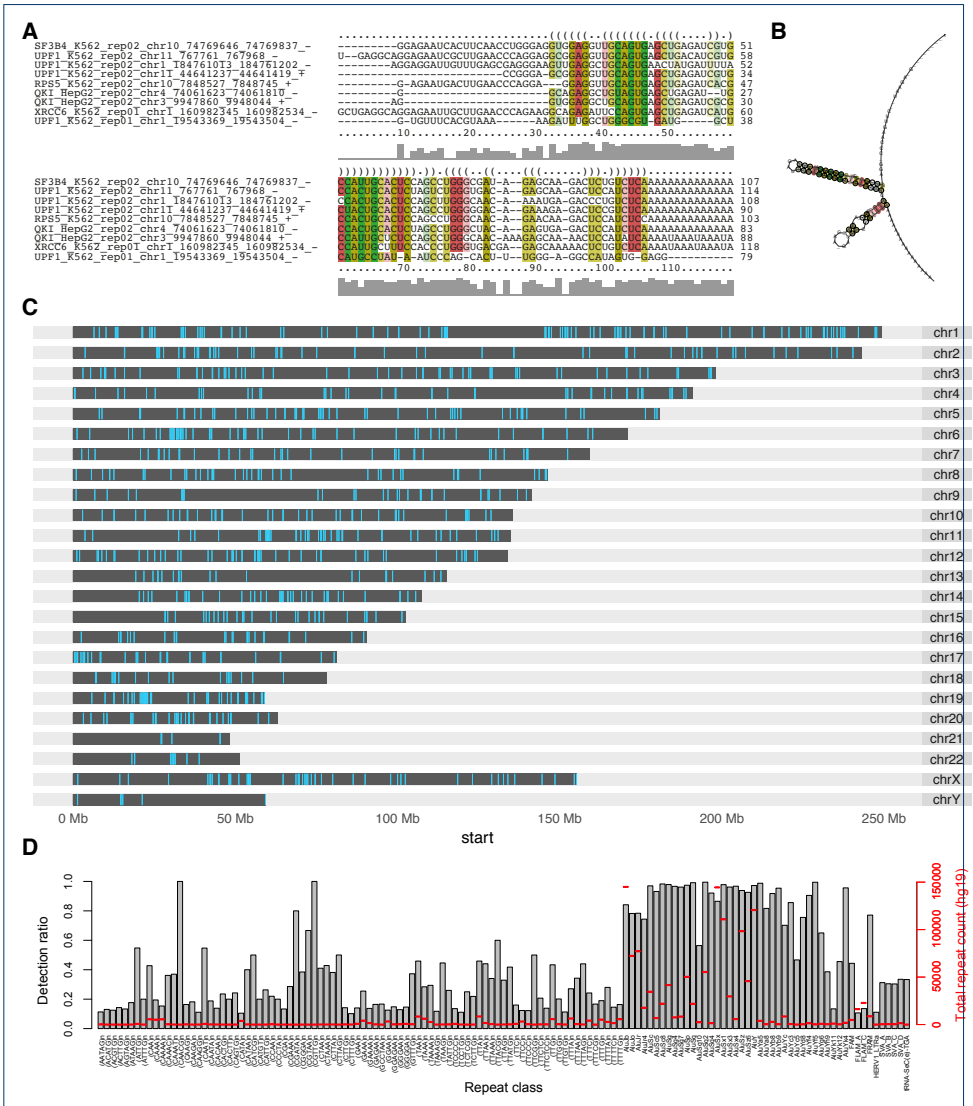

Figure S8. Genomic distribution of a UPF1-associated RNA structure motif

(**A**) Multiple sequence alignment of a significant cluster from **Figure 5F** as produced by `mLocaRNA` and `RNAalifold`, and its associated consensus secondary structure prediction (**B**). (**C**) Karyogram illustrating the human genomic coordinates (Grch37) of structural motif homologs, to this motif that do not overlap RepeatMasker annotations [**?**], as identified with `cmsearch` from the Infernal software package [49]. (**D**) Distribution of homologs within repeat elements (only repeats classes where $> 10\%$ of the repeats overlap homologs are displayed).