De-novo identification of homologous RNA secondary structure domains using base-pairing probabilities

Stefan E Seemann[*,2,3], Martin A Smith[*,3,4], XiuCheng Quek[3,4], John S Mattick[3,4]

[*] *Contributed equally*

[2] *Garvan Institute of Medical Research, 384 Victoria Street, Sydney 2010, Australia*

[3] *University of Copenhagen, Groennegaardsvej 3, Frederiksberg, Denmark*

[4] *St Vincents Clinical School, UNSW, Sydney 2010, Australia*

seemann@rth.dk

Keywords: RNA secondary structure, Basepair probability, Structural alignment, Clustering, RNA -protein interactions, RNA immunoprecipitation, high-throughput sequencing

Abstract: Non-protein coding RNAs (ncRNAs) are the prevalent transcriptional product of higher eukaryote genomes. Their varied biological functions are governed by both their sequence composition and their higher-order structural conformation. The uncertainty of secondary structure prediction algorithms for single RNA sequences in conjunction with the limited diversity of well-characterised RNA structures have restricted the identification and annotation of novel functional ncRNA domains. Here, we present a unified computational methodology for the identification of common RNA secondary structures from a set of sequences, requiring little to no user intervention while being fully customisable. We compare the performance of several state of the art tools for pairwise secondary structure alignment with `DotAligner`, a novel algorithm we developed that considers the ensemble of sub-optimal RNA base pairings between two RNA sequences simultaneously. Through hierarchical clustering and bootstrapping analysis, our method identifies statistically significant clusters of homologous, structured RNA domains with no limitations on the sequence composition of the input. We successfully identify known RNA secondary structures mixed in with randomised controls, as well as novel structured domains from various previously published transcriptomic datasets.

1 INTRODUCTION

The structure of RNA molecules is an essential functional criteria of many non-coding RNAs (ncRNAs), such as the stem-loop of microRNAs and the double stem-loop RNA motifs of the HOTAIR long ncRNA (Gupta et al., 2010). NcRNAs can be divided in RNA families of similar inherent functionality, structures, or composition. The largest collection of RNA families is the Rfam database with 2,208 families in its version 11.0 (Burge et al., 2013). However, high-throughput sequencing continuously uncovers novel non-coding RNA transcripts and genome-wide RNA structure predictions have revealed hundreds of thousands putative conserved RNA secondary structures. We hypothesize that the RNA secondary structure is the scaffold for intermolecular interactions of many ncRNA-driven regulatory pathways. Protein binding domains of RNA molecules may evolve totally independent from sequence and, instead, may be solely determined by structure. It has been shown that if the sequence similarity falls below 60% sequence comparison will not find anymore domain similarities that are based on structure (Gardner et al., 2005). In addition, competing structures and suboptimal structures may support or even drive the functionality of an RNA domain. Hence, methods are needed that find structural similarity independent from sequence conservation and freed from one single optimal RNA secondary structure.

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

each string. This problem variant is MAX-SNP-hard. However, the problem can be attacked by intelligent heuristics that avoid the examination of all possible aligning states.

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

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

2 IMPLEMENTATION

As described in (Palù et al., 2010) 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. Here, we present the program `DotAligner` which solves the related problem of aligning two basepair probability matrices (dot plots). `DotAligner` employs the heuristics alignment-envelope, which imposes constraints to sub-optimal string alignments, and fold-envelope, which imposes constraints to pre-calculated base pairing probabilities, to built pairwise sequence-structure alignments. We make 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 (Muckstein et al., 2002). A major criteria for the implementation was a fast running time to make `DotAligner` applicable for RNA structure clustering of large data sets. The alignment procedure consists of two steps:

1. pairwise probabilistic string alignments,
2. stochastic backtracking of string alignments and combined weight of corresponding dot plot alignments.

In the following we describe the alignment procedure and its weight functions implemented in `DotAligner`.

2.1 Pairwise probabilistic string alignments

In step 1 the computation of the partition function over all canonical pairwise string alignments is adapted from probA (Muckstein et al., 2002). 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 (Dalli et al., 2006). 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 (Klein and Eddy, 2003) 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.

2.2 Stochastic backtracking and combined weight of dot plot alignments

In step 2 a properly weighted sample of stochastic pairwise string alignments in the alignment ensemble is examined for their sequence-structure similarity. The stochastic backtracking is adapted from probA (Muckstein et al., 2002) 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 (Klein and Eddy, 2003) 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 (Will et al., 2007):

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

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

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

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

3 MATERIALS AND METHODS

All data of modest size, scripts, and pipelines described herein are available in the associated GitHub repository (**REFERENCE TO GITHUB REPO**).

3.1 *Parameter optimization using pairwise alignments*

Initial parameter optimization was performed on 8,976 pairwise RNA structure alignments (7,859 unique sequences from 36 RNA structure families) curated in the `BRAliBase 2.1` reference dataset (Wilm et al., 2006). Postscript files representing the base-pairing probabilities were generated using the implementation of McCaskill's partition function algorithm in `RNAfold` from the Vienna RNA package (version 2.1.3) on the raw, unaligned `BRAliBase 2.1` sequences. The postscript files were then converted to pairwise probability matrices using an *ad hoc* java script.

The following combination of parameters were tested on the reference sequences (*equation reference*){*min value; max value; increment*}:

- κ weight of sequence similarity versus structural similarity (11){0; 1; 0.1};
- θ relative weight of sequence/basepair similarity *versus* the similarity of unpaired/pairing probability (4)(10){0.2; 1; 0.2};
- g_o and g_{ext} are the gap-open and gap-extension penalties, respectively (8){0; 1; 0.1}.

The output was then contrasted to the reference alignments using two key metrics previously shown to be the most accurate at detecting structural conservation (Gruber et al., 2008):

1. The difference in Structural Conservation Index (SCI) between `DotAligner` alignments and the `BRAliBase 2.1` reference alignment. The SCI is a robust thermodynamic measure of structural compatibility, where the Minimum Free Energy (MFE) of the alignment consensus—calculated with RNAalifold from the Vienna RNA package (Lorenz et al., 2011)—is normalized by the average MFE of individual sequences. For pairwise alignments, the ΔSCI is calculated as:
$$\Delta SCI = SCI_{DotAligner} - SCI_{BRAliBase} \quad (15)$$
where
$$SCI = MFE_{Cons} / \left(\frac{MFE_1 + MFE_2}{2}\right) \quad (16)$$
2. The topological edit distance between the experimental alignment consensus secondary structure and that from the reference using `RNAdistance` from the Vienna RNA package. Quek, which version of these software did you employ? Also, was RNAalifold used in (1.) or just RNAz ?

3.2 *Stochastic sampling of structured RNA families from RFAM*

The seed alignments from RFAM 12.0 (Nawrocki et al., 2014) were downloaded, split into families, and converted from Stockholm to Fasta file formats. Only RFAM entries with published tertiary or crystal structures, without nested base-pairs (pseudoknots—natively excluded in RFAM v12.0), and with at least 10 representatives were retained for subsequent sampling. A modified version of the RFAM sampling software described in (Smith et al., 2013) was used to extract representative sequences from the seed alignments with restrictions on their sequence composition and length. The program (`GenerateRFAMsubsets.java`) incrementally samples each family (starting from RF00001) like so:

1. Random selection of an 'anchor' sequence within the RFAM entry that satisfies the size constraints. If no sequence is found after 250 random draws, the next RFAM entry is sampled;
2. Each sequence of compatible size in the current RFAM entry is compared to the 'anchor'. If the pairwise sequence identity of both sequences is within the minimum and maximum constraints, the sequence is retained. Pairwise sequence identity is calculated as the number of non-indel matches divided by the length of the smaller sequence;
3. Step 2 is repeated. However, each additional sequence must satisfy the min and max sequence identity constraints against all previously retained sequences, ensuring that the overall sequence identity of the sampled RFAM entry lies within the constrained range;
4. Once a maximal amount of sequences are selected (default 20), or if no match to the 'anchor' is found, the next RFAM entry is sampled.
5. The sampling continues until all RFAM entries have been surveyed or a user-defined ceiling is reached.

3.3 *Performance benchmarking using binary classification matrices*

The RFAM accessions from sequences sampled using the above-mentioned strategy were used to populate a binary matrix serving as a reference classifier. In other words, if any 2 sequences being compared belong to the same structure family, the corresponding position in the matrix is instantiated with "1", or "0" otherwise. Another matrix is then populated with the alignment/comparison score from a given algorithm (using the same sequences) that have been normalized between 0 and 1 (min and max score, respectively). The empirical matrix is evaluated against the binary classification matrix via Receiver Operating Characteristic (ROC) and the Area Under the Curve (AUC) of the ROC using the `pROC` R package (Robin et al., 2011).

This approach was used for both refined parameter optimization of `DotAligner`and for comparative performance benchmarking of pairwise RNA structure/sequence alignment algorithms (on separate datasets generated via the stochastic sampling).

Additional parameter range for refined parameter optimization:
First round (coarse)
k in 0.25 0.5 0.75
t in 0.25 0.5 0.75
s in 5000 1000 200
T in 10 5 1
o in 0.75 1
e in 0.05 0.2

Second round (fine)
k in 0.3 0.4 0.5 0.6
t in 0.5 0.6 0.7 0.8
s in 1 5 20 50
T in 10 5 1
o in 1
e in 0.05

In addition to `DotAligner`, the algorithms used for performance benchmarking include:

- `LocaRNA`
- `CARNA`
- `FOLDALIGN`
- `pmcomp`
- `Needleman-Wunsch`

3.4 *Cluster analysis and extraction*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996) was used via the DBSCAN R package (https://github.com/mhahsler/dbscan).

3.5 *Processing of eCLIP data*

The 53 narrowPeak bed files (available in April 2016) asociated with the human eCLIP datasets described in (Van Nostrand et al., 2016) were downloaded from the ENCODE data hub (https://www.encodeproject.org/).

4 RESULTS

Parameter optimization

The `DotAligner` algorithm implements several theoretical parameters that first need to be tuned before applying this tool to biological sequence analysis. All combinations of core parameters were tested on the 8,976 pairwise RNA structure alignments curated in the `BRAliBase 2.1` reference dataset (Wilm et al., 2006). The resulting 26,398,295 pairwise alignments were filtered to retain only those **with an RNAdistance score equal to 0** (8,078,665), indicating that the alignment generated the exact structure reported

in the reference. Of these, 7,624,073 (94.4%) produced a SCI score as good as the reference alignment, including 79,511(1.0%) where the SCI from DotAligner alignments scored better. Interestingly, the latter can be assigned to most of the RNA families represented in BRAliBase 2.1(25 out of 36). Furthermore, many of the results encountered while optimizing DotAligner parameters were associated with greater SCI scores than the reference alignments (2,774, 733; 34.3%), but were ignored given a non-null edit distance with the reference. This suggests that there is some room for improvement in the representation of the structures in the reference alignments, *which may have been automatically generated based on similarity to a covariance model*. Although some of the alignments produced may be closer to the biological reality than the BRAliBase 2.1 representatives, which encompass many sequences automatically added into RFAM given their similarity to covariance models, we dismissed this possibility to ensure both metrics were compatible with and directly comparable to the reference structures.

We performed a product rank selection of the globally optimal parameters for DotAligner. As several combinations of parameters gave optimal results, we selected those that were consistently present in the top scoring alignments across as many families as possible by ranking the alignments in function of their SCI score, ensuring they had a null edit distance to the reference.

There are additional DotAligner parameters that may contribute to alignment accuracy (see below). Based on the initial optimization results, their contribution to alignment accuracy on the BRAliBase 2.1 reference set is less significant than the aforementioned variables. These parameters include:

- T measure of the relative importance of the optimal string alignment.
- The usage of statistical substitution matrices R_{seq} and R_{paired} instead of similarity matrices M_{seq} and M_{paired}, and
- The relative weight ζ of the unpaired probability compared to the ratio of upstream pairing probability if substitution matrix R_{seq} is used;
- The number s of suboptimal alignments to consider ;
- The minimal unpaired/pairing probability p_0 (default 0.0005).

0.2 1.0 k0.7 t0.5

4.1 Benchmarking performance using heterogenous samples of structured RNAs

As our intended application of pairwise RNA structure alignments is for the identification of homologous motifs from a pool of biological sequences, we measured the accuracy of DotAligner using a more suitable benchmark than the cummulative score distribution of individual comparisons. We subsampled the structured RNA reference alignments from RFAM (Nawrocki et al., 2014) using a stochastic sampling approach, which ensures that the sampled sequences present user-constrained sequence characteristics (see **Materials and Methods**). Two test samples of 200 sequences (no more than 20 per family) were generated for initial benchmarking: a low pairwise sequence identity (low-pi) and a high pairwise sequence identity (high-pi) set, where all sequences from the same family share between 0-55% and 56-95% sequence identity, respectively. These pooled sets of sequences with heteregeneous RNA structures provide a more practical benchmarking test set, with potential background interference between structurally related sequences.

The optimal DotAligner parameters identified through the BRAliBase 2.1 analysis were refined by performing all *vs* all pairwise alignments on the low-pi and high-pi datasets, using a range of different parameter combinations. The output of each parameter combination was then compared to a binary classification matrix and subjected to Receiver Operating Characteristic (ROC) analysis (see **Materials and Methods**). Parameters with the

To assess how well DotAligner reproduces known classifications of RNA structure, we compared the normalized scores from all vs all pairwise comparisons on the stochastically sampled RFAM input (similarity matrix) to a binary matrix representing the RFAM classifications (**FIGURE XXX**). Something about the conclusions here.

DotAligner to other RNA structure alignment and clustering tools using the following framework:

1. Generate dissimilarity matrix dM_A from $\frac{n(n-1)}{2}$ pairwise structure comparisons with each algorithm
2. Hierarchical clustering of RNA secondary structures and significance testing with pvclust (Suzuki R and Shimodaira H. Bioinformatics 2006).
3. Generate dissimilarity matrix dM_R from scoring metric of (1.) from curated RFAM alignments

(constrained alignment).

4. Calculate the correlation coefficient between dM_A and dM_R using the Mantel correlation statistic (the cross-product between the standardised distances).

The accuracy of the proposed algorithm is assessed using the specificity (SP) and the sensitivity (SN), which are defined as follows:

$$SP = \frac{TN}{TN+FP}, \quad SN = \frac{TP}{TP+FN} \tag{17}$$

where TP is the number of correctly predicted positives, FP is the number of incorrectly predicted positives, TN is the number of correctly predicted negatives, and FN is the number of incorrectly predicted negatives. Furthermore, the area under the receiver operating characteristic (ROC) curve was used to optimize the different combinations of parameters. The ROC curve plots the true positive rates (SN) as a function of the false positive rates (1 - SP) for varying parameters.

As benchmark data set we selected 300 sequences of 10 H/ACA-box snoRNA families from Rfam version 11.0 seed alignments with average pairwise sequence identity (APSI) $< 90\%$ and sequence lengths of $>$ 130bp and $<$ 140bp: *SNORA1*, *SNORA13*, *SNORA14*, *SNORA15*, *SNORA16*, *SNORA17*, *SNORA18*, *SNORA19*, *SNORA2*, *SNORA22*. We chose only sequences of similar length because step 1 of `DotAligner` performs global alignments.

4.2 Performance benchmarking

4.3 Complete RFAM sequences

We compare `DotAligner` with sequence alignments (in-house implementation of Needleman-Wunsch algorithm with the blastn parameters match = 2, mismatch = -3 and gap penalty = 5 which are optimized for sequence identity of 90%) and the structure alignment tools `pmcomp` (using default parameters or larger values for parameter *-D* if the length difference of two sequences is $>$ 5 bp), `CARNA`, and `LocaRNA`. Figure **??** shows that the sequence aligner (SP = 0.80, SN =0.97) performs very well on our benchmark set with a very high sensitivity which is most likely due to the fact that the input sequences have some degree of sequence information. `pmcomp` (SP = 0.72, SN = 0.74) performed with a medium sensitivity and specificity. With `DotAligner` we are able to find very well defined clusters (SP = 0.99), however, at the cost of sensitivity (SN = 0.61), see Figure **??**.

4.4 Fragmented RFAM sequences

Local alignment, simulating genomic screens. More emphasis on qualitative clustering

4.5 A unified RNA structure clustering pipeline

We implemented a user-friendly pipeline that automates all steps required for the detection of homologous RNA secondary structure motifs from a set of user- provided sequences. The pipeline is implemented in BASH programming language and is designed for execution on a high performance computing server (currently, only SGE is supported). This enables non-specialists to complete such an analysis with minimal bioinformatics knowledge, while facilitating parameter modification and customization for advanced users.

In summary, the pipeline performs the following tasks on a fasta file input:

1. Generates base-pairing probability matrices for each sequence with RNAfold's parition function algorithm
2. Performs all-vs-all pairwise alignment in parallel with DotAligner (and/or CARNA, locarna, ...)
3. Generate (dis)similarity matrix from pairwise alignment scores
4. Perform hierarchical clustering and bootstrap significance testing with `pvclust` (Suzuki R and Shimodaira H. Bioinformatics 2006).
5. Extract the sequences and associated guide trees for significant clusters
6. Render a consensus secondary structure motif using the multiple structure alignment tool mlocarna

4.5.1 On consensus hierarchical clustering

We are investigating the practicality and efficiency of a consensus hierarchical clustering approach, where the (dis)similarity matrices of different pairwise structure alignment algorithms are concurrently employed for cluster analysis. **This is cutting-edge stuff and Luis will report back soon.**

4.5.2 On multiple structure alignment and 2D motif rendering

Generating a multiple structural alignment at the end of the pipeline is an important but tricky step. Right now, we are using mlocarna for this, which to my knowledge is the only tool that can produce

such output without too much fuss. However, there is a substantial concern that arises from its use: DotAlignerand mlocarna use fundamentally different alignment algorithms. This caveat is somewhat resolved by enforcing mlocarna to use the guide tree produced from the all-vs-all (dis)similarity matrix from DotAligner. Mlocarna will still align the sequences based on their consensus structure, therefore some additional benchmarking may be required. N.B., we can dictate which pairwise aligner (or probabilistic aligner) to be used by mlocarna its execution parameters, although mlocarna may ignore this when a guide-tree is provided—thus employing locarna to produce intermediate alignments even when only 2 sequences are involved. **CHECK THIS WITH SEBASTIAN WILL**

Some more specific points to consider:

- –threads=X seemingly doesn't affect mlocarna performance. Is this only implemented for pairwise comparisons? Generating the intermediary alignments uses one CPU. Perhaps –cpu=X will work?
- RNAPLfold is used for longer input sequences, right? Is this because it overcomes sequence length discrepancies? Should we enfore a size limit on the input sequences (either trim or extend the input to XX nucleotides divergence)?
- Try option `--pw-aligner path/to/DotAligner` and see if it will give more reliable consensus structure
- Test whether the speed limitation of iterative refinement (`--iterations=XX`) will be compensated by better quality alignments
- Will this cause (m)locarna to use the entire dot plots for the alignment? Test the effect of the following parameters `--probabilistic --consistency-transformation --it-reliable-structure=XX`.

5 DISCUSSION

The application of DotAligner is a fastly calculated similarity score between two probability matrices to enable their subsequent clustering. Similar to CARNA, which also does not garanty the optimal solution, DotAligner is not deterministic.

We plan to integrate the proposed method in a pipeline that screens regions of interest for structured RNA domains in a collection of RNA molecules. The so far presented approach finds only global alignments. This strategy is applicable for input sequences of similar lengths or if one sequence is considerably shorter (due to the usage of free end gaps). However, local alignment is favorable if both input sequences are long. Despite of the partition function version of the local alignment problem is available, its application dramatically increases the search space and, thus, the running time. As alternative, a possible screening pipeline may comprise window based thermodynamic folding, *e.g.* by RNAplfold (Bernhart et al., 2006), and filter regions of high intra-molecular binding probabilities in a pre-processing step, *e.g.* by using RNAlocal (Dotu et al., 2010), followed by the presented alignment tool DotAligner. SHOULD WE INCLUDE THIS ANALYSIS IN THIS PAPER AND IF YES WHERE? The pre-selection of local structural potential may improve the boundaries of common structured RNA domain.

DotAligner can also be extended for multiple alignments, similar to the strategy implemented in pmmulti (Hofacker et al., 2004), and the generation of phylogenetic trees. This may replace or support the hierarchical clustering approach used here. In addition, both may serve as input for RNA secondary structure predictors, such as PETfold (Seemann et al., 2008) unifying thermodynamic and evolutionary information.

ACKNOWLEDGEMENTS

I thank the Carlsberg foundation for my travel grant. Skål!
MAS is funded in part by a Cancer Council NSW project grant and

REFERENCES

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22:614–615.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2013). Rfam 11.0: 10 years of {RNA} families. *Nucleic Acids Res*, 41(Database issue):D226–32.

Dalli, D., Wilm, A., Mainz, I., and Steger, G. (2006). STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, 22(13):1593–1599.

Dotu, I., Lorenz, W. A., Van Hentenryck, P., and Clote, P. (2010). {RNA} structural segmentation. *Pac Symp Biocomput*, pages 57–68.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

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

Gruber, A. R., Bernhart, S. H., Hofacker, I. L., and Washietl, S. (2008). Strategies for measuring evolutionary conservation of rna secondary structures. *BMC bioinformatics*, 9(1):122.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076.

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

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

Klein, R. and Eddy, S. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4:44.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):1.

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

Muckstein, U., Hofacker, I., and Stadler, P. (2002). Stochastic pairwise alignments. *Bioinformatics*, 18 Suppl 2:S153–60.

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. (2014). Rfam 12.0: updates to the rna families database. *Nucleic acids research*, page gku1063.

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1.

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

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

Seemann, S. E., Gorodkin, J., and Backofen, R. (2008). Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic acids research*, 36(20):6355–62.

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

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

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. (2016). Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature methods*, 13(6):508–514.

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

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

Table 1: Table Title

$\Delta SCI < 0$	$\Delta SCI \geq 0$
5S rRNA	Cobalamin
5.8S rRNA	Entero CRE
Entero 5 CRE	HCV SLIV
Entero OriR	Hammerhead 1
HCV SLVII	HepC CRE
HIV FE	Histone3
HIV GSL3	Lysine
HIV PBS	SRP bact
Hammerhead 3	U1
IRES HCV	UnaL2
IRES Picorna	yybP-ykoY
Intron gpII	
K chan RES	
Retroviral psi	
SECIS	
SRP euk arch	
S box	
T-box	
TAR	
1U2	
U6	
gcvT	
sno 14q I II	
tRNA	

The caption without a number

SUPPLEMENTARY INFORMATION