COMPUTATIONAL TRANSCRIPTOMICS

Identifying structured protein-binding domains in long noncoding RNAs

Stefan E. Seemann[1,2], Martin A. Smith[1,3], John S. Mattick[1,3]

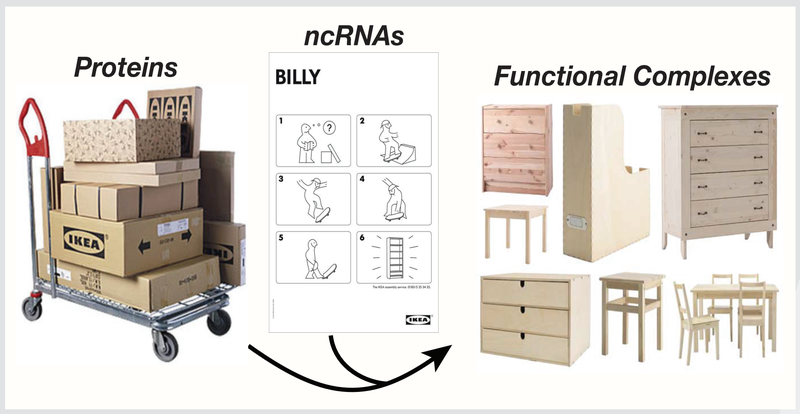
[1] *RNA Biology and Plasticity, Garvan Institute of Medical Research, Sydney, Australia*

[2] *Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Medical Sciences, University of Copenhagen*

[3] *St-Vincent's Clinical School, UNSW Australia, Sydney*

INTRODUCTION

Long non-coding RNAs (lncRNAs) encompass a substantial fraction of mammalian genomes, yet their detailed functional characterization lies in stark contrast to their abundance in the literature. Their extensive regulatory roles depend on their inter- and intra-molecular interactions, namely through the binding of effector proteins to discrete RNA structure conformations.



In particular, lncRNAs direct the action of epigenetic modification complexes, dynamically regulating gene expression during development and differentiation. Their function can be compared to a set of instructions for the assembly or use of specialised equipment. For example, the lncRNA HOTAIR is known to interact with and guide the action of the Polycomb Repressor Complex 2 in *trans* through structured RNA domains [1]. Given the abundance of evolutionarily conversed RNA structure motifs [2] and the pervasive nature of mammalian transcription [3], **we propose that** families of RNA structure motifs form a network of functional domains for the recruitment of specific RNA-binding proteins.

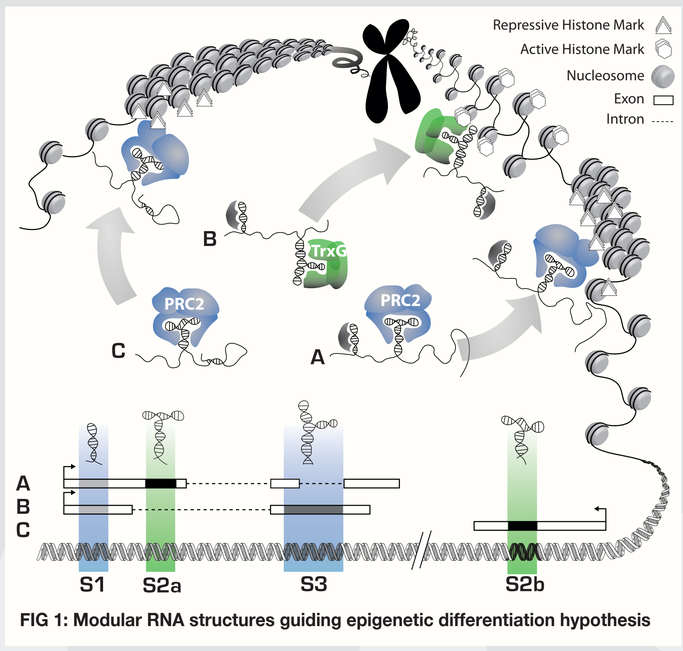


FIG 1: Modular RNA structures guiding epigenetic differentiation hypothesis

Here, we describe a new program and a systematic bioinformatics approach for the identification of common RNA structures within a subset of sequences, considering an ensemble of sub-optimal base-pairings.

DOTALIGNER: A base-pair probability aligner

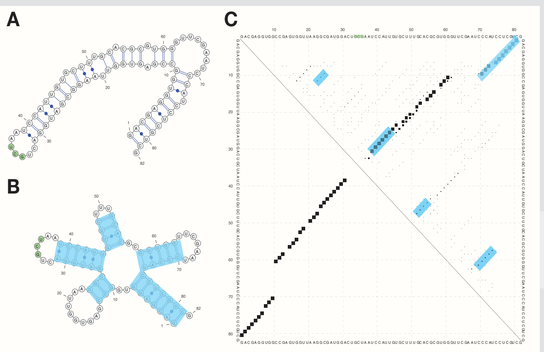


FIG 2: Dot-plots for (sub-optimal) RNA structure prediction
(**A**) The miminum free energy prediction for a SER-tRNA; (**B**) The biological structure of the same tRNA; (**C**) RNA dot-plot highlighting the biological structure within the sub-optimal pairs.

There are several RNA structure alignment algorithms that employ dot-plots, however only one intrinsically aligns dot-plots (CARNA [4]). We developed a fast 2 step heuristic approach using dynamic programming to directly align dot-plots called DotAligner.

RNA structure dot-plots contain all base-pairing probabilities, indicated by the size of the dots, of optimal and sub-optimal structures. Considering suboptimal base-pairing can improve the physical realism of RNA stucture predictions significantly, particularly in comparative approaches.

DotAligner Get pairwise alignment A of dotplots D_A and D_B

Require: sequence S_A and its dotplot $D_A = Pr(x_{11} \dots x_{NN})$;
sequence S_B and its dotplot $D_B = Pr(y_{11} \dots y_{MM})$
{STEP 1: Global alignment of the pairing probabilities of each base in S_a and S_b}
for i = 1 to N **do**
  for j = 1 to M **do**
    Z_{ij} = Needleman-Wunsch $[Pr(x_{i1} \dots x_{iN}), Pr(y_{j1} \dots y_{jM})]$
  end for
end for
{STEP 2: find "best" local path through similarity matrix Z}
for i = 1 to N **do**
  for j = 1 to M **do**
    S_{ij} = max $\{S_{i-1j}$ + gap $, S_{i-1j-1} + Z_{i-1j-1}, S_{ij-1}$ + gap $\}$
  end for
end for
$A(D_A, D_B)$ = BACKTRACKING(S)

BENCHMARKING

We compared the performance of DotAligner to 4 other RNA structure alignment algorithms [4-7] and a pure sequence based tool on 2 datasets, consisting of randomly chosen RFAM [8] sequences presenting limited pairwise sequence identity (PSI) values: (i) [10-55%]; and (ii) [56-95%].Sensitivity and specificity were measured from all vs. all pairwise comparisons followed by hierarchical clustering of the scores. Significance was measured using an alpha statistic of 0.98 after 10,000 bootstraps (via the R package pvclust).

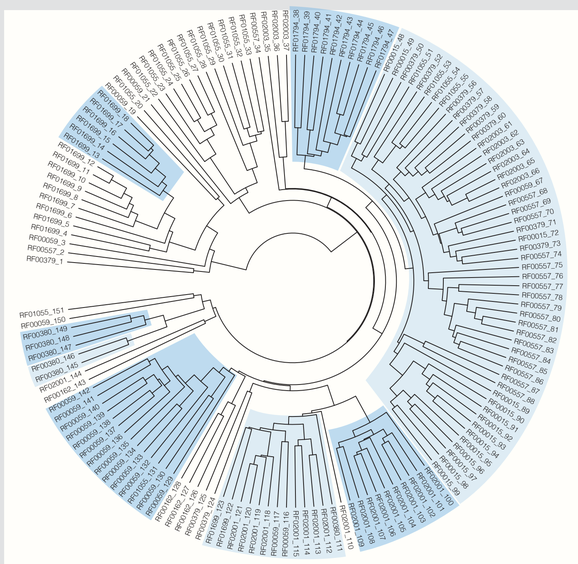


FIG 3: Hierarchical clustering of benchmark data from DotAligner
Non-random clusters are highlighted in blue (AU statistic ≥ 0.98).

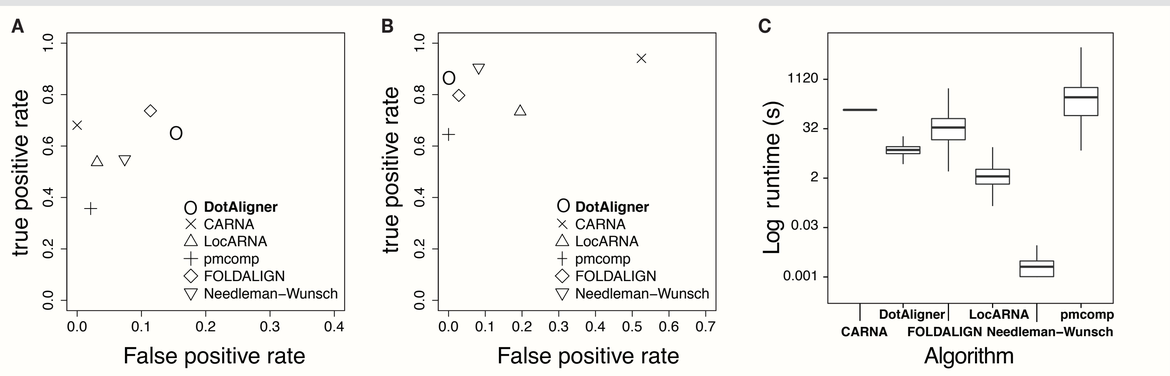


FIG 4: Benchmarking performance of pairwise RNA structure alignment algorithms
(**A**) Receiver Operating Characteristic (ROC) plot for the low (≤55%) pairwise sequence identity RFAM dataset; (**B**) ROC plot for the higher (>55%) mean pairwise identity RFAM dataset. SP=TN/(TN+FP), SN=TP/(TP+FN). TP= most common RFAM ID in a cluster, FN=RFAM IDs outside main cluster (**C**) Runtime distributions on 2.4GHz AMD CPU (N.B., *CARNA* was limited to 120 seconds).

APPLICATION TO RIPSEQ DATA

DNMT1 is an RNA-binding repressor of transcription that has DNA methylation activity. DNMT1-bound transcripts have been shown to target its activity to specific genomic loci [9]. It has also been shown to preferentially bind double-stranded RNA, although no common sequence or structure motifs have been reported in the DNMT1-bound transcriptome.

We applied an analytic pipeline similar to Fig 5. to the associated RIPseq data [9], selecting 152 sequences preferentially bound to DNMT1 (≥3 fold more abundant than IgG control and >15x coverage).

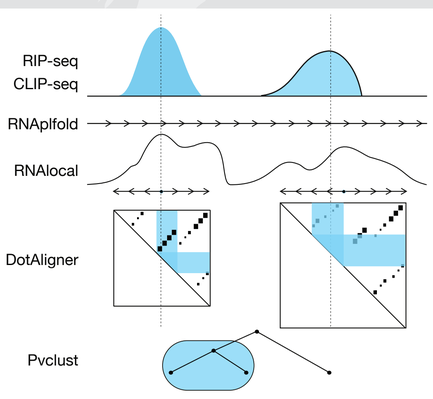


FIG 5: Experimental workflow for RNA structure clustering using DotAligner

This approach identifies 7 new statistically significant RNA structure motifs that share little sequence identity, yet present consensual 2D conformations (Fig 6).

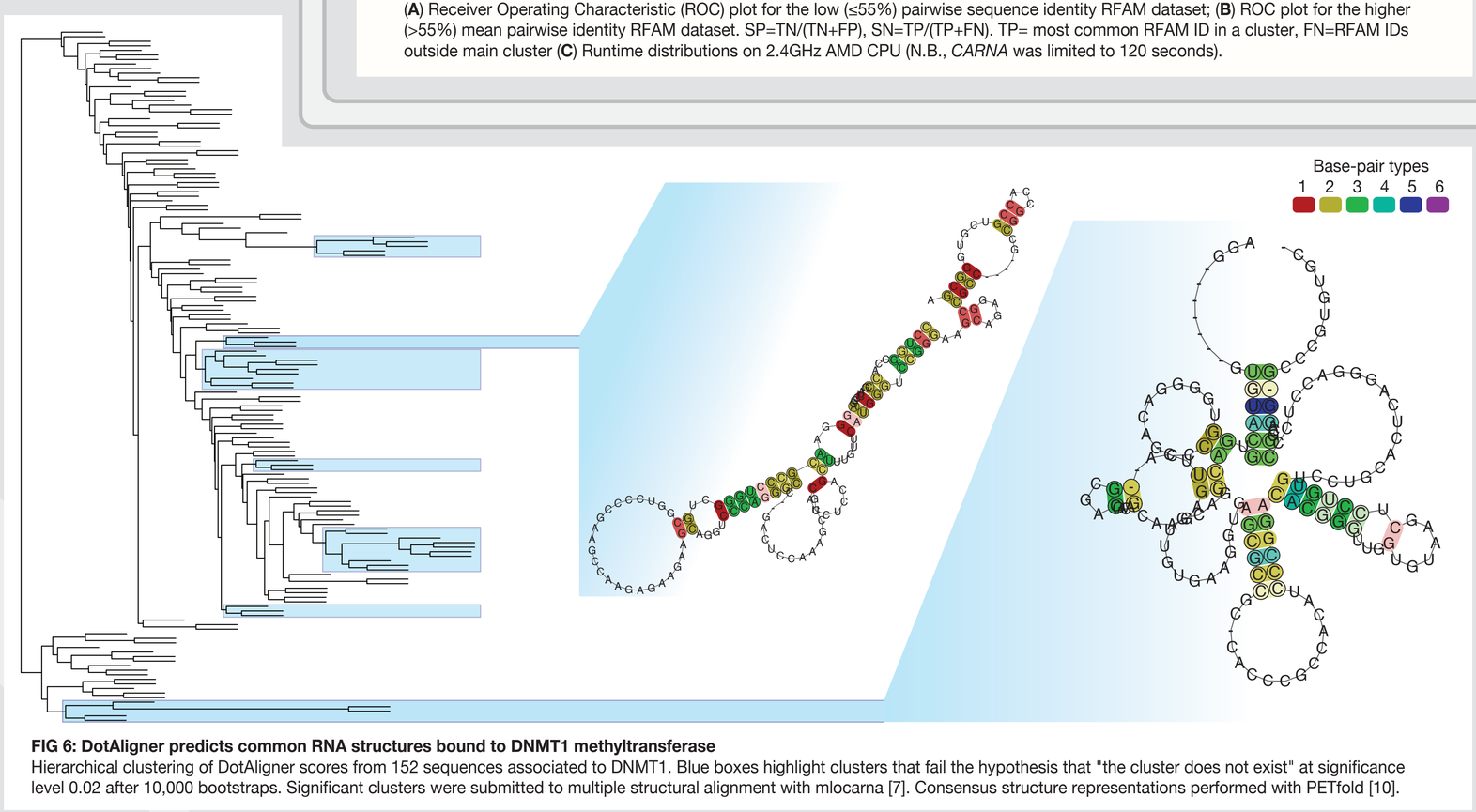


FIG 6: DotAligner predicts common RNA structures bound to DNMT1 methyltransferase
Hierarchical clustering of DotAligner scores from 152 sequences associated to DNMT1. Blue boxes highlight clusters that fail the hypothesis that "the cluster does not exist" at significance level 0.02 after 10,000 bootstraps. Significant clusters were submitted to multiple structural alignment with mlocarna [7]. Consensus structure representations performed with PETfold [10].

Acknowledgments

SES wants to acknowledge the Carlsbergfondet which supports his visit to the Garvan Institute with a travel grant, and the Center for non-coding RNA in Technology and Health, University of Copenhagen. MAS is funded in part by a Cancer Council NSW project grant. The authors would like to thank Warren Kaplan and Derrick Lin for their assistance with and access to the Wolfpack HPC cluster.

References

[1] Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010; [2] Smith, M A et al. Widespread purifying selection on {RNA} structure in mammals. NAR, 2013; [3] Djebali, S et al. Landscape of transcription in human cells. Nature, 2012; [4] Sorescu, D A et al. CARNA–alignment of RNA structure ensembles. NAR, 2012; [5] Havgaard, J H et al. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol, 2007; [6] Hofacker, I L et al. Alignment of RNA base pairing probability matrices. Bioinformatics, 2004; [7] Will, S et al. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol, 2007; [8] Burge, S W. et al. Rfam 11.0: 10 years of {RNA} families. NAR, 2013; [9] Di Ruscio, A et al. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature, 2013; [10] Seemann, S E et al. The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. NAR, 2011