

Combining structure probing data on RNA mutants with evolutionary information reveals RNA-binding interfaces

Vladimir Reinharz¹, Yann Ponty^{2,*} and Jérôme Waldispühl^{1,*}

¹School of Computer Science, McGill University, Montreal, Québec H3A 0E9, Canada and ²Laboratoire d'informatique, École Polytechnique, 91128 Palaiseau, France

Received June 23, 2015; Revised March 6, 2016; Accepted March 21, 2016

ABSTRACT

Systematic structure probing experiments (e.g. SHAPE) of RNA mutants such as the mutate-and-map (MaM) protocol give us a direct access into the genetic robustness of ncRNA structures. Comparative studies of homologous sequences provide a distinct, yet complementary, approach to analyze structural and functional properties of non-coding RNAs. In this paper, we introduce a formal framework to combine the biochemical signal collected from MaM experiments, with the evolutionary information available in multiple sequence alignments. We apply neutral theory principles to detect complex long-range dependencies between nucleotides of a single stranded RNA, and implement these ideas into a software called aRNhAck. We illustrate the biological significance of this signal and show that the nucleotides networks calculated with aRNhAck are correlated with nucleotides located in RNA–RNA, RNA–protein, RNA–DNA and RNA–ligand interfaces. aRNhAck is freely available at <http://csb.cs.mcgill.ca/arnhack>.

INTRODUCTION

A recent surge of experimental technologies allows us to rapidly access the structural profile of RNA molecules. Such approaches include *in vitro* methods such as selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (1) and parallel analysis of RNA structure (PARS) (2), which provide transcriptome-wide measurements of RNA structure at single-nucleotide resolution *in vitro*. Combined with classical RNA nearest neighbor energy minimization models (3), this data allowed for a significant improvement of the accuracy of RNA secondary structure prediction methods (4,5).

Recently, Das *et al.* introduced the mutate-and-map (MaM) protocol, which consists in obtaining SHAPE data

simultaneously for an RNA and for (a large number of) its 1-point mutants (6). By revealing the perturbation of base-pairing properties, this data provides an information to estimate the contribution of a specific position to the stability of the native structure, which can in turn be used to determine the structure of the molecule. The current approach to exploit MaM data for RNA secondary structure prediction uses empirical rules and pseudo-energy bonuses within classical dynamic programming prediction algorithms (5).

A distinct, yet complementary, approach to analyze structural and functional properties of non-coding RNAs makes use of the evolutionary information encapsulated within multiple sequence alignments (MSAs). The latter provides an alternate signal which is often key to understand and characterize the origin and structure of functional motifs (7,8).

To date, both approaches have not been combined and even less reconciled. Nonetheless, an important observation is that the systematic mutations such as those conducted in the MaM protocol enable us to probe the evolutionary landscape of a molecule, which in turn can be used to reveal nucleotide patterns in the fitness landscape. To capture this signal, it is essential to design a formal framework that calculates correlations between the genetic robustness of the structural profile—obtained from MaM experiments—and the evolutionary information available for this molecule—usually contained in MSAs.

This paper attempts to look beyond RNA structure determination, and introduces a novel concept to leverage the information embedded in experimental structure probing datasets of mutant RNAs. We apply neutral theory principles (9) to detect functional dependencies between distant nucleotides in a single stranded RNA. More precisely, we first use MaM data to identify mutations that significantly destabilize the native structure of the molecule, i.e. the mutations associated with the most divergent SHAPE profiles. Then, we retrieve from RNA MSAs (Rfam database (10)) homologous sequences containing those destabilizing mutations, and compare their nucleotide distribution to the

*To whom correspondence should be addressed. Tel: +1 514 398 5018; Fax: +1 514 398 3883 Email: jeromew@cs.mcgill.ca
Correspondence may also be addressed to Yann Ponty. Tel: +33 1 69 33 41 12; Fax: +33 1 74 85 42 42; Email: yann.ponty@lix.polytechnique.fr

background distribution observed in the Rfam MSA. Finally, the ensemble of positions with highest mutual information is used to reveal nucleotide networks of functional dependencies. This protocol aims to capture non-trivial covariations or geometric conservations that are key to guarantee the stability and specificity of the binding site structure.

We implement our model in a software named aRNhAck. We illustrate potential applications of the signal captured with our theoretical framework, and apply aRNhAck to analyze MaM datasets available on the RNA Mapping Database (11). Our experiments reveal non-trivial long-range dependencies within ncRNA primary structures of 5S ribosomal RNA, the yeast phenylalanine tRNA and the cobalamin, adenine and glycine riboswitches. We investigate the biological significance of these patterns by looking at the distribution of these nucleotides on the RNA three-dimensional (3D) structures. Interestingly, we find significant correlations between the sets of nucleotides produced by our method and those identified as participating in RNA–RNA, RNA–protein, RNA–DNA and RNA–ligand interfaces.

MATERIALS AND METHODS

Definitions

We abstract an RNA sequence w of length n as a string in $\{A, C, G, U\}^n$. A **secondary structure** S for w is a set of base pairs (i, j) , $0 < i < j \leq n$, which are pairwise non-crossing, i.e. if (i, j) and (k, l) are in S and $i < k$, then $i < j < k < l$ or $i < k < l < j$.

Any secondary structure S can be decomposed in five types of **secondary structure elements (SSEs)**:

- (i) **Stems** consist in one or more base pairs $\{(i_1, j_1), \dots, (i_l, j_l)\}$ such that $i_m + 1 = i_{m+1}$ and $j_m - 1 = j_{m+1}$;
- (ii) **Hairpins** are composed of a base pair (i, j) and, for any position $k \in [i + 1, j - 1]$, k is not involved in any base pair;
- (iii) **Interior loops** are a set of two base pairs $\{(i_1, j_1), (i_2, j_2)\}$, where $i_1 < i_2 < j_2 < j_1$ and all k such that $i_1 < k < i_2$ or $j_2 < k < j_1$ and k does not belong to any base pair;
- (iv) **Multi-loops** are a set of base pairs $\{(i_1, j_1), \dots, (i_l, j_l)\}$ where $l > 2$, $i_1 < i_2 < j_2 < \dots < i_l < j_l < j_1$ and all k such that $i_1 < k < i_2 \cup j_l < k < j_1 \cup \bigcup_{2 \leq m < l} j_m < k < i_{m+1}$ and k does not belong to any base pair.
- (v) **Exterior loops** are the remaining unpaired positions, all k such that there is no base pair (i, j) with $i < k < j$, and their adjacent base pairs.

An illustration of the different types of SSEs is given in Figure 1. We use the term of **loop** to denote indiscriminately either an hairpin, an interior loop, a multi-loop or the exterior loop.

By definition, the SSEs are not disjoint. Every base pair at the interface between two structural elements belongs to both of these SSEs. Any unpaired position, however, only belongs to a single SSE while a base pair always belongs to a stem and at most two other SSEs, for lonely base pairs.

SHAPE and structural information

SHAPE experiments (12) on single stranded RNA provide a measure of the flexibility of individual nucleotides. The reactivity score that it produces has been shown to discriminate between base-paired versus unconstrained or flexible residues (1). Many methods (4,5) have been developed to leverage SHAPE data to increase the accuracy of secondary structure prediction.

The result of a SHAPE experiment for an RNA of length n is a vector of length n where every position is associated with a positive value, indicating the reactivity of a given nucleotide. The resulting vector is called a **SHAPE profile**. Beyond providing partial structural information, SHAPE profiles can be analyzed in a differential setting, e.g. to monitor structural differences between related RNAs. In such a setting, a normalization step is required, and we used a procedure introduced by Deigan *et al.* (5) to preprocess all data. It consists in dividing each value by the mean of the top 10% of the data after excluding outliers. The outliers are defined as the value $> 1.5 \times$ the interquartile range. For simplicity, any future reference to SHAPE values will indicate normalized values.

Mutate-and-Map. Given an RNA sequence w of length n , the MaM (6) strategy consists in completing an initial SHAPE experiment on w with n additional SHAPE experiments on single-point mutants. For each sequence, a distinct position i is selected and the associated nucleotide is mutated into his Watson-Crick complementary base (i.e. C to G, G to C, A to U, or U to A). Thus, the sequence of the mutant associated with position i is entirely characterized by i , and we will denote by w_i the sequence of the i th mutant in the following. Each SHAPE experiment produced n reactivities, and the MaM scheme results in a $(n + 1) \times n$ matrix, where each row corresponds to the SHAPE profile. We will denote by R the SHAPE profile of the wild-type sequence w , and by R_i the profile of the i th mutants w_i .

Structural disruption. Given the high correlation between an RNA SHAPE profile and its structure, it is generally accepted that a significant SHAPE profile disruption reflects a change of structure. Using MaM data, we estimate the SHAPE profile disruption of the point-wise mutation at position i , by comparing the SHAPE profiles of w and w_i using different metrics.

In this work, we quantify the profile disruption induced by the i th mutation, by taking the l^2 norm between R and R_i , restricted to a window of width $2\lambda + 1$, i.e. by considering positions in the interval $[i - \lambda, i + \lambda]$, for λ a parameter. In this work, λ is fixed to 10. We denote $\Delta(w, w_i)$ this distance measure defined as:

$$\Delta(w, w_i) = \sqrt{\sum_{k=i-\lambda}^{i+\lambda} (R(k) - R_i(k))^2}.$$

We also tested three alternative measures: The first one is l^2 norm between the whole profiles of w and w_i , to evaluate the global SHAPE disruption (Supplementary Figure S1); The second is a variant of Δ which considers the maximally contributing window over the whole sequence (instead of

Secondary Structure Elements types

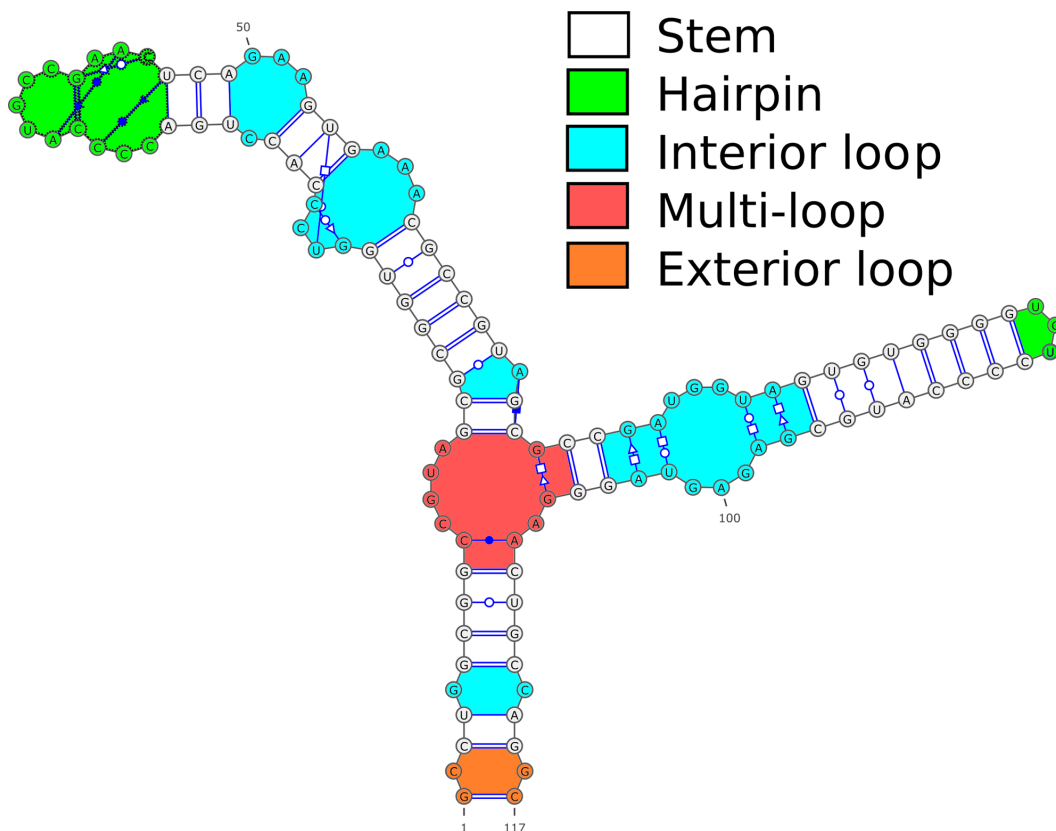


Figure 1. RNA secondary structure elements (SSEs). Blue edges represent canonical base pairs, while black indicate the phosphodiester backbone. The hairpins are colored in green, the interior loops in blue, the multi-loops in red and the exterior loop in orange. The base pairs adjacent to those elements are part of them (e.g. the base pair 9–109 belongs to a stem and to the interior loop).

only considering the one centered on i for Δ), and aims at identifying non-local structural rearrangements (Supplementary Figure S2); The third restrains the positions for the ℓ^2 norm between w and w_i to the SSE where the mutation lies, in order to assess the local 3D disruption of the SHAPE profile (Supplementary Figure S3). Although some of these measures showed potential for applications, we only report our analysis on the Δ measure, whose signal was clearest.

In the following, we use a parameter δ to identify mutations associated with significant changes of the structure. More precisely, given a percentile δ and a MaM experiment, we select the mutations at position i with a SHAPE profile disruption $\Delta(w, w_i)$ in the δ percentile of all profile disruptions (See Algorithm 1).

Evolutionary information

Evolutionary information can be used to witness how nature *repairs* non-lethally disruptive mutations to preserve or re-establish a given phenotype/function. It has been used for a wealth of applications, ranging from the detection of RNA 3D modules (13) to the correction of pyrosequencing errors (14).

On the structural level, a simple, yet powerful, example lies in the paradigm of compensatory mutations. When the function of an RNA secondary structure hinges on its

capacity to adopt a stable structure, which typically requires the presence of canonical base-pairs A-U and G-C, any mutation that occurs within one of the paired bases disrupts the structure stability, and therefore the fitness of the RNA. Evolution will then favor mutations which restore the canonical status of the base-pair, either by reverting the mutation or by compensating it. Compensatory mutations may also be witnessed in positions which are not immediately structurally (in a broad definition, including non-canonical motifs and pseudoknots) related to the disrupting mutation. In this case, we hypothesize that these mutations are most likely involved in the quaternary structure, and may reside at the interface between the chain of interest and other molecules in the formation complexes. The detection of such interactions is therefore the primary application of our method.

To that purpose, we rely on covariation models, an information theoretical tool that has been successfully used for RNA for sequence alignment and structure prediction (15). The Rfam database (10) is a repository of RNA families, composed of aligned homologous sequences which are gathered from an hypothesis of structural homogeneity within a given functional family. In this work, we will use the hand-curated MSA of Rfam to supplement structural disruption data, and use this additional knowledge, in con-

junction with information theory, to identify specific pairs of positions and nucleotides having high affinity with each other.

Normalized point-wise mutual information. To identify those pairs of nucleotides at specific positions which vary together, we use the **normalized point-wise mutual information** (NPMI) measure (15). Given x and y two mutations, each indicated by a column of an alignment and a nucleotide present at the position, the NPMI is defined as:

$$\text{NPMI}(x, y) = \frac{\log \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}}{-\log \mathbb{P}(x, y)} \in [-1, 1]$$

where, probabilities $\mathbb{P}(\cdot)$ are estimated from their frequencies in the MSA.

An NPMI of -1 indicates that x and y never appear together. On the opposite side of the spectrum, a value of 1 signifies a perfect correlation. If x and y can be considered as two independent random variables, then the NPMI will be 0 . Starting from an Rfam alignment of total length m , the NPMI of every $25 \binom{m}{2}$ pairs of possible mutations is computed, where m is the length of the alignment. For every $\binom{m}{2}$ pair of positions, the nucleotides can be either A, C, G, U or a gap -. The set of all NPMI s greater than -1 is called ζ .

The procedure to compute the positions over a cutoff percentile ζ_c given a mutation m , a list of positions p and a MSA is described in Algorithm 2.

Structures as graphs. Most disruptive mutations, when observed in multiple alignments, are found in combination with compensatory mutations which re-establish the structure. Since a common secondary structure is posited in this work, such local covariations are scarcely informative and should be ignored. However, RNAs are 3D, and thus one cannot use the sequence distance between the mutation and positions of interests. In order to assess a realistic notion of distance, we transform the secondary structure into a graph G where the positions are the vertices. The edges are composed of the phosphodiester bonds and the canonical base pairs. For G to adequately reflect the pairwise proximity of nucleotides involved in a loop, an edge is added between every pair of position belonging to the same loop. Effectively every loop becomes a clique. Figure 2 shows the full graph of a secondary structure. The distance from a loop to a position x is defined as the maximal shortest path in G from x to any position in that loop.

Proximity filtering. In RNA, a large proportion of observed covariations are adequately explained by the necessity to preserve the secondary structure. Since the secondary structure is, to a large extent, already revealed by comparative analysis (and already present in the Rfam profile taken as input to the method), it does not constitute the primary object of interest of our study. In order to minimize the probability of detecting local structural compensations, we require a minimal distance γ between the index of the mutation and the position of the loops selected for their good NPMI values. This criterion is formally implemented in Algorithm 3.

Binding interfaces positions. Since both negative and positive correlations can indicate positions of interest, we use two different, ζ^- and ζ^+ , thresholds for the NPMI s. ζ^+ will be a bound on the positive values of the NPMI and ζ^- on the negative ones. Due to the high number of possible combinations, NPMI s having values -1 are frequent and uninformative. They are discarded.

For those loops deemed as regions of interest, we predict that the set of positions with an NPMI above ζ^+ or below ζ^- are nucleotides in binding interfaces while the others are not.

Algorithm 1: disruptiveMutations(MaM, δ)

```

 $l = D = \emptyset$ 
for  $m \in MaM$  do
   $l \leftarrow l \cup \text{shapeDisruption}(m, MaM)$ 
for  $m \in MaM$  do
  if  $\text{shapeDisruption}(m, MaM) \geq$ 
     $\text{percentile}(l, \delta)$  then
     $D \leftarrow D \cup \{m\}$ 
return  $D$ 

```

Algorithm 2: filterNPMI(m, p, MSA, ζ_c)

```

 $q = \emptyset$ 
 $a \leftarrow \text{getAllNPMIs}(MSA)$ 
 $\zeta^+ \leftarrow \text{percentile}(a[a > 0], \zeta_c)$ 
 $\zeta^- \leftarrow -1 \times \text{percentile}(-1 \times a[-1 < a < 0], \zeta_c)$ 
for  $x \in p$  do
  for  $l \in \{A, C, G, U, -\}$  do
    if  $\text{getNPMI}(m, x, l) > \zeta^+$  then
       $q \leftarrow q \cup \{x\}$ 
    else if  $-1 < \text{getNPMI}(m, x, l) < \zeta^-$  then
       $q \leftarrow q \cup \{x\}$ 
return  $q$ 

```

Algorithm 3: filterNearbyPositions(m, S, γ)

```

 $p = \emptyset$ 
 $g \leftarrow \text{SGraph}(S)$ 
for  $u \in \text{getSSEs}(g)$  do
  if  $\text{distance}(\{m\}, u) \geq \gamma$  then
     $p \leftarrow p \cup \{u\}$ 
return  $p$ 

```

Algorithm 4: aRNhAck($S, MaM, MSA, \delta, \gamma, \zeta_c$)

```

 $M_{\text{disruption}} \leftarrow \text{disruptiveMutations}(MaM, \delta)$ 
for  $m \in M_{\text{disruption}}$  do
   $p \leftarrow \text{filterNearbyPositions}(m, S, \gamma)$ 
   $p \leftarrow \text{filterNPMI}(m, p, MSA, \zeta_c)$ 
return  $p$ 

```

Implementation

Our software, aRNhAck, is implemented in python 2.7. To identify mutations of interest, the threshold of SHAPE profile disruption was tested between the 95 and 99 percentiles. The parameter λ was set to 10 creating windows of size 21 to measure the local SHAPE profile disruptions. The parameter γ was evaluated for values between 0 and 30. The whole implementation is freely available at: <http://csb.cs.mcgill.ca/arnhack>

It requires BioPython (16) for reading the MSAs, networks (17) for modeling the graphs and matplotlib (18) for visualizing the results. The 3D complex analysis used in our validation is performed using the Python API provided by the PyMol software (19).

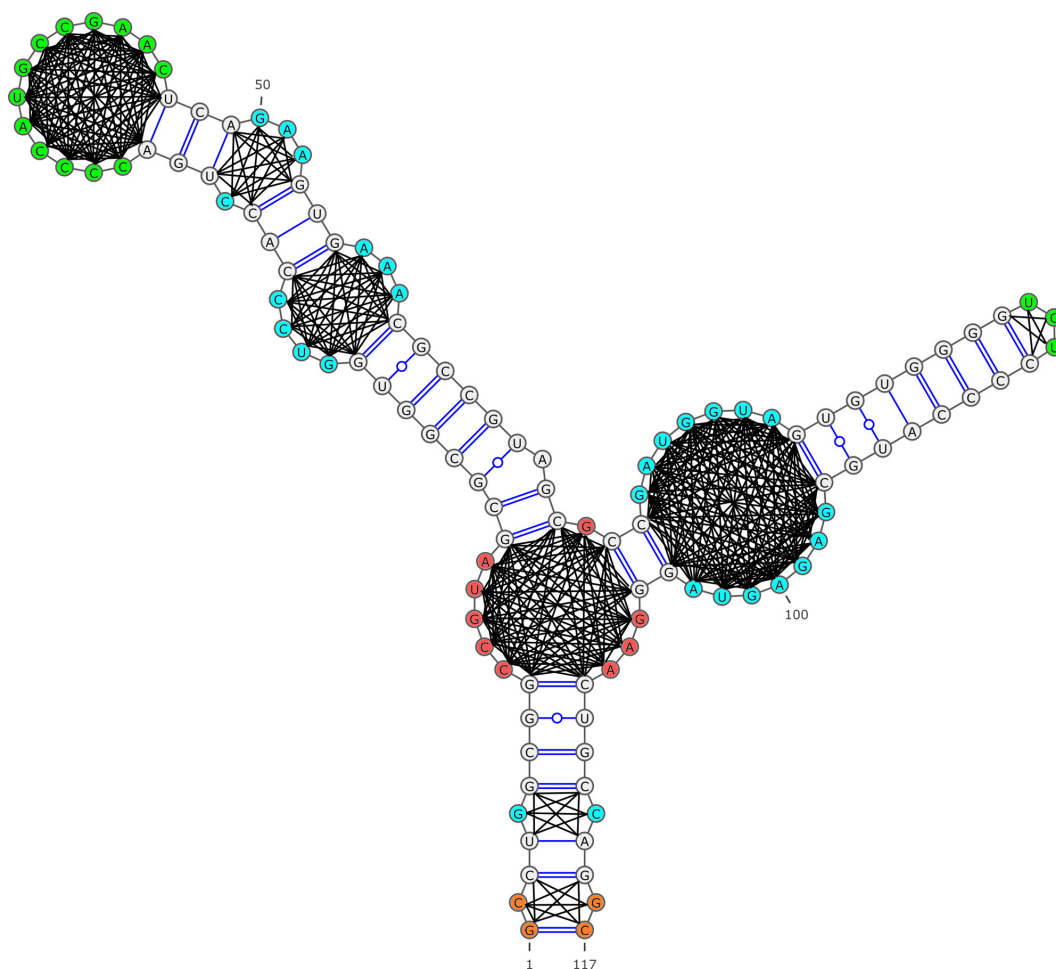


Figure 2. RNA secondary structure graph used for proximity filtering.

Datasets

We evaluated our methods on molecule for which we could obtain simultaneously (i) an MaM experiment dataset, (ii) a determined 3D structures interacting with other chain(s) and (iii) a Rfam alignment Table 1 shows the quality of the alignment of the mutate-and-map sequences with their corresponding Rfam family. This search resulted in six RNAs: the 5S ribosomal RNA, the c-di-GMP riboswitch, the cobalamin riboswitch (Puzzle 6), the phenylalanine tRNA, the adenine riboswitch and the glycine riboswitch.

However, the experimental structure of the glycine riboswitch found in the PDB contains an artificial stem loop binding a protein used to stabilize the RNA structure and facilitate the crystallization. Since this protein is missing in the MaM experiments, we decided to exclude this riboswitch from our test set. Nonetheless, we show our results in the supplementary material (Supplementary Figure S7).

The 5S ribosomal RNA is the family RF00001 on Rfam. Its seed alignment consist of 713 sequences. The family also provides the consensus structure. The MaM protocol was applied to the consensus sequence of four structures which have as PDB IDs: 2WWQ (20), 3OAS and 3OFC (21), and 3ORB (22). Those four determined structures have almost

Table 1. Bit scores of mutate-and-map sequences on covariance model built from related Rfam alignments using *infernal*

RNA	Bit score
5S	48.65
c-di-GMP ribo.	44.23
cobalamin ribo.	118.96
adenine ribo.	62.68
tRNA	43.91
glycine ribo.	52.85

the same sequence with slight differences in the length on their 5' and 3' extremities.

The yeast phenylalanine tRNA is included in the Rfam family RF00005 which has 960 seed sequences from various tRNAs. Its structure has been crystallized in presence of magnesium and manganese (PDB ID: 1EHZ). Although, for a complete characterization of its structural context and interactions with other molecules, we also considered structures of the two tRNAs in the structure of the yeast 80S ribosome-tRNA complexes (PDB ID: 3J78).

The c-di-GMP riboswitch is present in family RF01051 in Rfam, which contains 156 sequences in its seed alignment, and a consensus structure. The consensus sequence

was also built from four structures, with PDB IDs: 3IWN (23) and 3MXH, 3MUV, 3MUT (24). Importantly, c-di-GMP is known to bind a pocket inside the three-way junction at positions 11 – 13, 40 – 41 and 85 of the sequence on which the MaM experiments were run (25,26), and the MaM experiment was done in presence of its ligand. It is also worth noting that, in order to facilitate the crystallization, the hairpin loop L2 of this molecule has been artificially designed to bind the U1A protein. Here, we included only the positions binding the ligand. Nonetheless, for completeness, we also show in the supplementary material the results including the stem loop L2 binding interface.

The MaM cobalamin riboswitch sequence can be found in the Rfam family RF00174 which has 430 seed sequences. The PDB contains the structure bounded to its ligand (PDB ID: 4GXY). Noticeably, the MaM experiments were done in the presence of cobalamin ligands.

The adenine riboswitch belongs to family RF00167 which has 133 seed sequences. The structure with the adenine ligand has PDB ID: 1Y26. Three different MaM experiments were conducted on this molecule. Experiments Adenine_2 and Adenine_3 were done in presence of the ligand, and are used in this paper. The third experiment Adenine_4 has been performed in absence of the ligand, and thus was omitted from this benchmark since disruptive mutations cannot be used to detect key structural elements of the ligand-bound structure. Nonetheless, the results are indicated in the supplementary material.

To complete our benchmark, we also built a secondary test set of Rfam families with experimentally determined 3D structures, but for which MaM experiments were not available. We selected all Rfam families with sequences having a size ranging from 35 to 150 nucleotides, and with PDB files containing at least one other molecule in the vicinity of the RNA. In total, we found 14 families matching 729 different structures.

We omitted the shortest sequences (i.e. Rfam families RF00032, RF00037 and RF00175) because our distance metric δ would be too coarse to extract a signal on such small structures. Similarly, we also removed large molecules (i.e. more than 150 nucleotides) because the accuracy of the nearest-neighbor model decreases significantly beyond this size. So it is the case for computational tools with which we simulated MaM data (i.e. remuRNA).

Experimental design

The *Infernal* 1.1 (27) software was used with default parameter values to: (i) create a covariance model for each alignment, and; (ii) align the sequence from the MaM experiment with the generated covariance model. The consensus secondary structure was then restricted to gapless positions within the aligned sequence w .

For each mutation over the SHAPE profile percentile cutoff δ , the dataset was composed of the regions of interest given γ , i.e. the set of positions returned by the Algorithm 3.

We used different strategies to determine the interaction sites (i.e. positive dataset), depending of the nature and context of these interactions. All interactions were manually verified.

For the 5S RNA, we implemented a PyMOL script to extract nucleotides of each PDB model, whose position any of their atom is at most at 5Å from any atom of another chain in the complex. An implementation of this script is included in the distribution of aRNhAck

For the tRNA, we extracted positions that are at most 5Å away from another chain in the two tRNAs found inside the structure of the yeast 80S ribosome-tRNA complexes (PDB ID: 3J78). However, because those were not phenylalanine tRNAs, we aligned them to the MaM sequence with *Lo-caRNA*(28), and used this alignment to map the interaction sites on the latter. We identified the positions 1, 19, 34 – 36, 56 – 57, 73 – 76 (containing the anticodon) in this positive set. Among them, only the anticodon and T-ψ-C-G, known to bind the 5S RNA in the 50S ribosomal subunit (29), motif appeared to us to be strongly conserved. Thus, we considered only these two interactions sites in our experiments and presented the results separately. For completeness, the results obtained on other positions have been included in the supplementary material. Finally, we also confirmed the location of the anticodon using *tRNAscan-SE* (30).

For the riboswitches, we used *Ligand Explorer* (31) to identify nucleotide at most 5Å from the ligand in their respective crystal structures.

The set of all positions is found in the Supplementary Table S1.

All other remaining positions compose the negative dataset. The positions not present in the model were ignored. This highlights one of the challenges of this benchmark. For the 5S rRNA, out of 121 positions, two models had 3 nucleotides missing, one had 4 missing and the other 6. For c-di-GMP, out of 103 positions, one model had 8 nucleotides missing, two others 21 and the last 22. Which explains some discrepancies between the models.

The set ζ is composed of the NPMI between every pair of positions and every possible nucleotide (i.e. A, C, G, U and –) in the resulting alignment. The thresholds on the NPMIs, ζ^+ (resp. ζ^-) was sliced from the 0th to the 100th percentile of the positive values of ζ (resp. negative values of ζ).

Thus, for each SHAPE profile distance measure, each SHAPE distance threshold δ , each significant mutation given that measure, each γ , each PDB model and for every threshold pair (ζ^- , ζ^+), we obtained standard sensitivity and specificity scores. Those with a given SHAPE profile distance measure, SHAPE distance threshold δ , γ , PDB model and pair (ζ^- , ζ^+) were averaged together. The workflow of the method is illustrated in Figure 3.

RESULTS

We evaluated aRNhAck on a comprehensive set of values for δ the SHAPE profile distance measure and γ the proximity threshold. For each (δ , γ) pair, we computed a receiver operating characteristic curve and its area under the curve (AUC). The data are shown in Figure 4A. Noticeably, we averaged the values for the 5S rRNA and c-di-GMP riboswitch who have four PDB models. Importantly, we remind that the set of positives and negatives is influenced by the value of γ , as calculated by Algorithm 3. We also note that we only show the pair of parameters δ , γ such that for all structures for each RNA the positive and negative sets

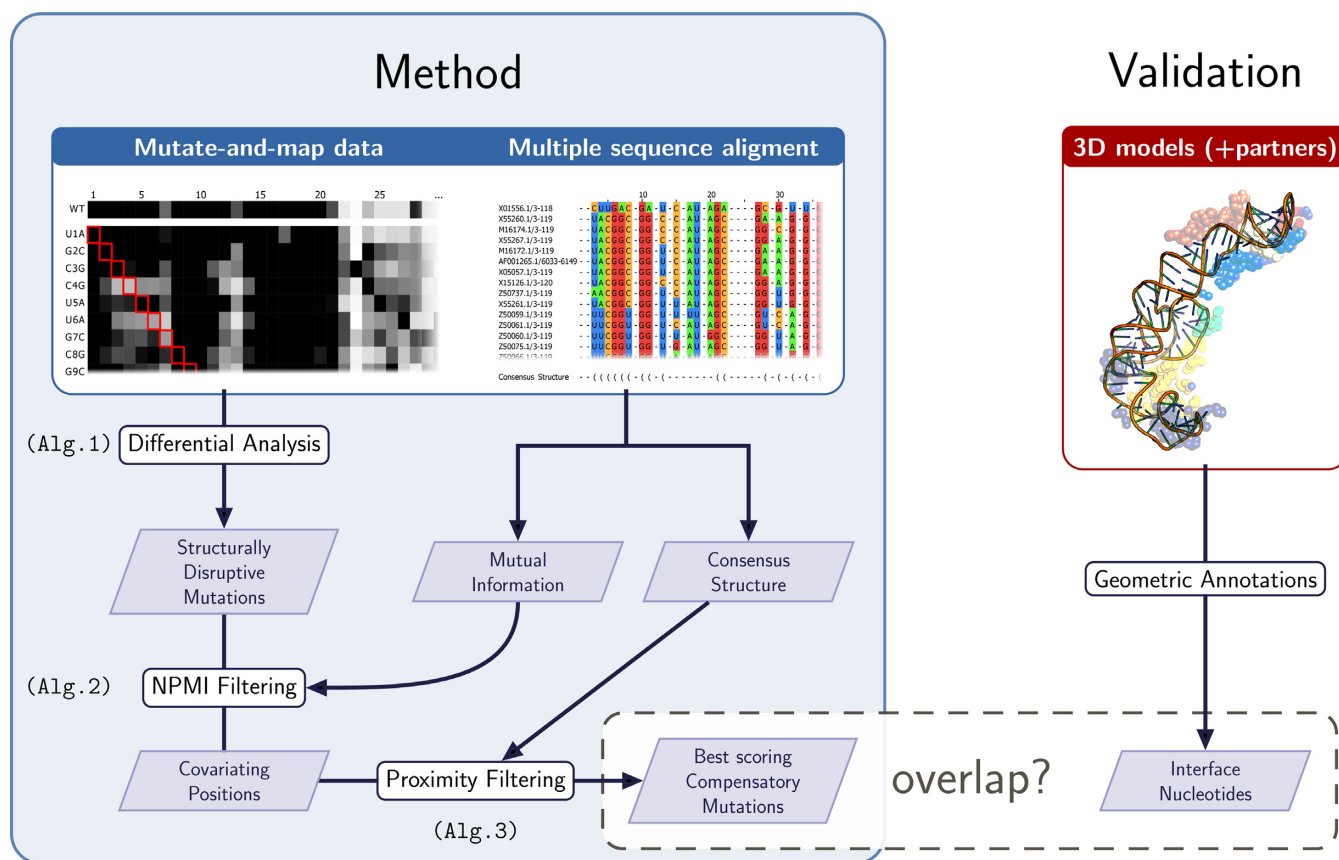


Figure 3. General overview of our method and validation procedure. Starting from mutate-and-map data, supplemented by an Rfam multiple sequence alignment, our method identifies positions that are simultaneously distant, yet co-evolve jointly, with positions that are structurally disruptive.

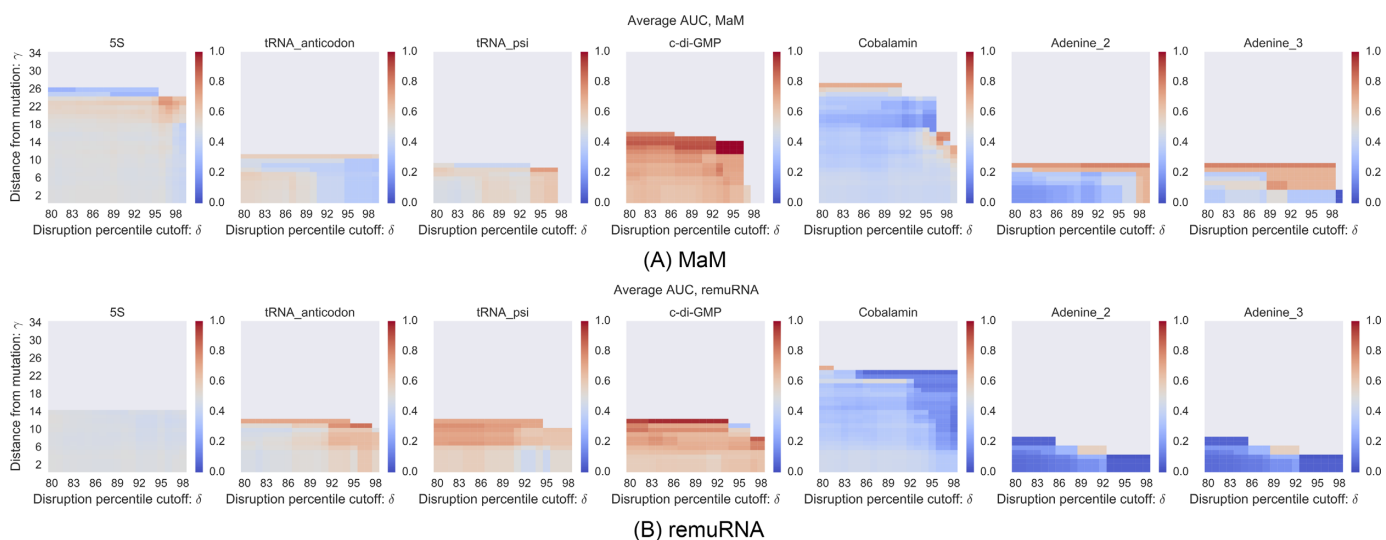


Figure 4. Overall performances of aRNhAck using experimental (A) and computationally-predicted structural disruption data (B). For a set of extreme percentile cutoff of the SHAPE profile disruption in the first row (computational remuRNA disruption in the second row) δ and a minimal distance γ from the mutation we show the average AUC. 5S positive set composed of the binding interfaces with other chains present in its four PDB models. The tRNA positive set is divided between the anticodon positions and the A- ψ -C-G motif positions, obtained from the literature. The c-di-GMP, cobalamin and adenine riboswitches positive sets are composed of the positions at most 5Å from their ligands in their PDB structures. Four different models exist for c-di-GMP and the AUC values are averaged.

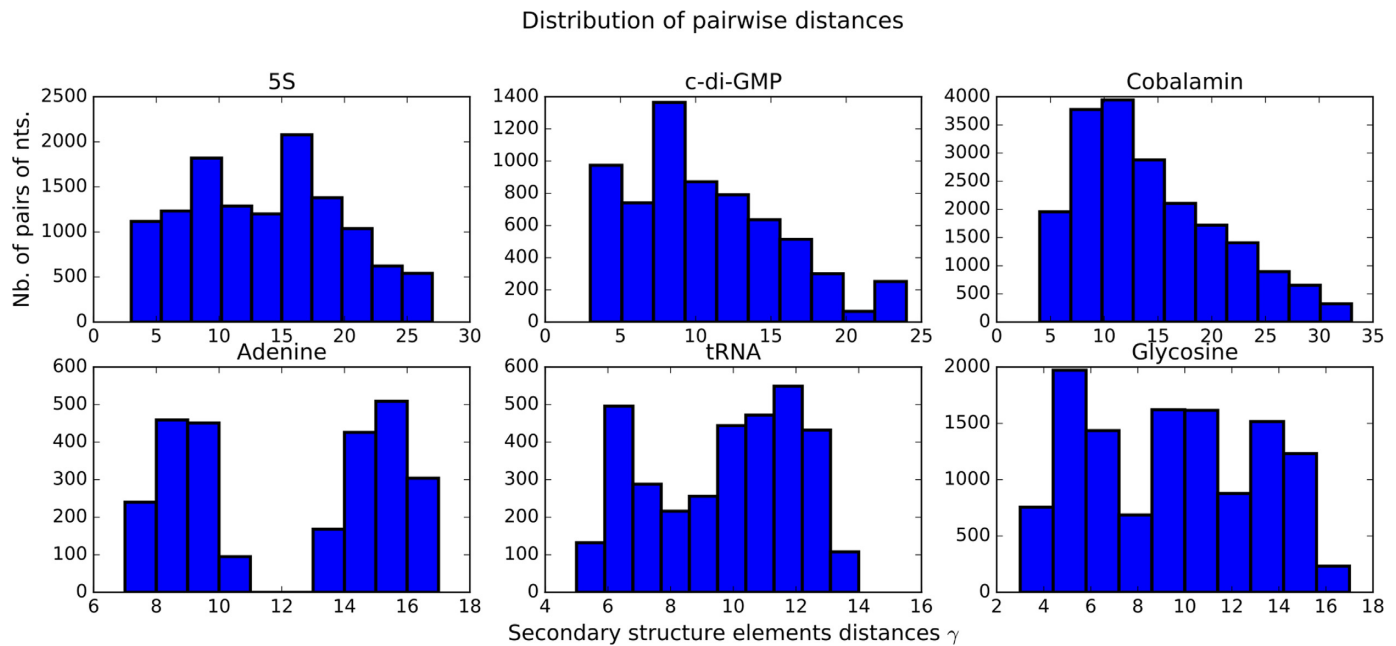


Figure 5. Distance distribution for pairs of SSEs, weighted by the numbers of non-shared nucleotides. The different distribution affect the quantity of positions selected by the parameter γ , distance from the mutation.

are non-empty. As discussed before, many nucleotides are missing in the PDB models of 5S and c-di-GMP.

Evolutionary stabilization of *in vitro* disruptive mutations reveal binding sites

We show our results in Figure 4A. In all cases, it exists a pair of parameters (δ , γ) for which aRNhACK achieve good predictive performance. Importantly, the maximal AUCs are found when the SHAPE disruption percentile cutoff δ is around 97%, with the largest possible distance γ from those mutations.

Results on 5S RNA and tRNA appear to have a slightly more diffuse pattern than other experiments. We hypothesize that it is due to the complexity of the nucleotide interaction network used to stabilize their 3D structures. Such a network would be easily disrupted by any mutation, hence the amplified noise in MaM experiments.

By contrast, the c-di-GMP riboswitch exhibits one of the strongest signals, most likely because of a strong evolutionary conservation and the central location of the binding site.

Interestingly, the cobalamin riboswitch exhibits a negative correlation with smaller disruption cutoff δ than the optimal value for which a positive correlation is found. In fact, these values of δ are strongly associated with positions in the leftmost hairpin of this structure. This suggests a conserved, yet currently unannotated, structural motif or binding interface that would warrant further investigations.

Finally, the two MaM experiments on the adenine riboswitch show that, although similar results are observed, the variation between experiments remains a concern and that the quality of the SHAPE experiments must be taken into account. The necessity of the correct structure when applying the MaM protocol is necessary as negative results

are obtained when using the unbound form (See Supplementary Figure S9).

We conjecture that the differences in the influence of the γ parameter, minimal distance from the mutation, are due to structural differences. We present in Figure 5 the distribution of path lengths (distance) between every pair of SSEs, weighted by the number of combinations of positions that are not in the intersection of the SSE. We observe that the distribution of distances on 5S rRNA tends toward a normal distribution, while on the c-di-GMP and cobalamin riboswitches it seems instead to follow a Poisson pattern. By contrast, the tRNA and adenine riboswitches have distributions tending toward bimodal modes. Those distributions determine how smoothly the number of positions considered could decrease as the parameter γ , minimal distance from the mutation, increases.

In Figure 6, we visualize one prediction from our method, for the 5s chain of 30FC. The disrupting mutation (red) is found in the top right corner behind the black spheres, and the positions with high mutual information are showed in green. The black spheres represent the subset of the residues for a chain of the complex, that are positioned at less than 5Å from the RNA. The other spheres belong to other molecules, each being color-coded to indicate its chain. Interestingly, we notice that, although many different chains are close to the RNA, and the position with high mutual information are far from the mutation, there is chain A (in black) close to the mutation and interacting with chain Z (purple) which binds with the mutation. Chain A builds a bridge up to chains O (beige) and F (yellow), themselves interacting with the compensatory mutations. We believe this example suggests the existence of mechanisms similar to compensatory mutations, but at the level of the quaternary structure.

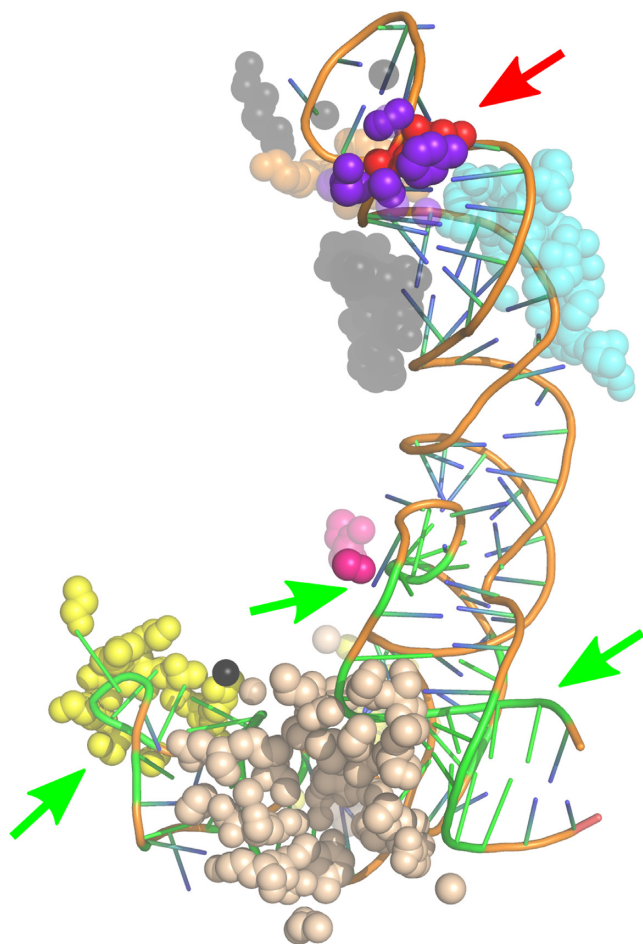


Figure 6. Predicted positions and interacting chains of the 5S rRNA $30FC$ structure. In red on the top right behind purple spheres is the disrupting mutation, in green the predicted position with high mutual information. The spheres around the RNA represent the subset of nucleotides at most at 5 Å from the rRNA, from other chains in the complex. The other spheres belong to other molecules. Each sphere is color-coded to indicate its chain as follows. Chain A is black, Z purple, W pink, V light blue, O beige, F yellow and M orange.

In this vision, a set of mutations contributes to re-establish the opportunity for participating in complexes, by compensating the effect of a disruptive mutation.

Computationally predicted structural disruptions yield weaker signal

To justify the use of SHAPE experimental data, we evaluate the performance of a fully automated pipeline in which mutations altering the RNA conformational landscape are predicted with a computer software instead of MaM data. Here, we predict these destabilizing mutations with *remuRNA* (32). Alternatively, for longer sequences, *RNA_{snp}* (33) can also be used.

The Boltzmann conformational ensemble \mathbb{B}_w of a sequence w is the probability distribution of valid RNA secondary structures on the sequence w . Given a wild type sequence wt and a mutant m , *remuRNA* (32) computes the relative entropy (or Kullback–Leibler divergence) between the two probability distributions \mathbb{B}_{wt} and \mathbb{B}_m . The latter

provides an estimate of the destabilization of the conformational landscape induced by the mutation. Given the set of all secondary structure S , the *relative entropy* is defined as:

$$\sum_{S \in \mathcal{S}} \mathbb{P}(S | \mathbb{B}_{wt}) \log \left(\frac{\mathbb{P}(S | \mathbb{B}_{wt})}{\mathbb{P}(S | \mathbb{B}_m)} \right).$$

We report our results in Figure 4B. Here again, our data unveil a signal that shows a correlation between the mutation identified with *aRNhAcK* and the RNA-binding interfaces. Nonetheless, the strength of the signal extracted with *remuRNA* is of lower magnitude than the one achieved with the SHAPE experiments and the MaM protocol. An exception is the tRNA for which *aRNhAcK* achieves better performance with *remuRNA* than MaM data. We conjecture that this could be due to a difficulty of the MaM protocol to capture a clear signal on these structures.

To further validate our model, we applied this protocol based on *remuRNA* prediction on a dataset made of Rfam families with experimentally determined 3D structures (See ‘Materials and Methods’ section). It took us 1 CPU year to complete this experiment on each of the 727 structures. For each family, we extracted the sequences annotated by Rfam as having the best matching score (i.e. the Bit score measuring the fitness of the PDB sequence to the Rfam covariance model). This restrained the set to 52 sequences since some families had many sequences with the same score. We present in Figure 7 our analysis on those top scoring sequences, showing the same trend. Complete results including omitted (short) families RF00032, RF00037 and RF00175 are available in the supplementary material (See Supplementary Figure S6).

To a lesser extent, the same trend is observed through all the matches annotated by Rfam (See Supplementary Figure S6). The full list of matching families is shown in the Sup. Mat. It is important to notice that the set of positive and negative positions was automatically retrieved from the PDB. Although *c-di-GMP* is to the best of our knowledge the only family with a designed sequence incorporated in the structure, interactions not provided in the PDB structures are considered as negative.

The poorest results are achieved in family RF01118. Interestingly, one of the conserved feature of this family structure is the presence of a pseudoknot, which is not modeled in the thermodynamic model underlying *remuRNA*. For those particular cases, only chemical experiments such as MaM can provide us trustworthy information about the destabilization produced by single point mutations. This reinforces the importance of producing further experimental data to reach the best performances.

These observations validate our methodology and at the same time justify the use of SHAPE data.

To complete this analysis, we also investigate the ratio and size of the overlap between the structurally-disruptive positions predicted using *remuRNA* and SHAPE experiments, for different percentiles. As indicated in Supplementary Figure S4, we notice a clear linear decrease in the size of the overlap. At the 50th percentile, the size of the intersection is cut by half. When combining the results of *remuRNA* and SHAPE experiments together, we quickly reach results that are almost as good as those obtained with SHAPE data

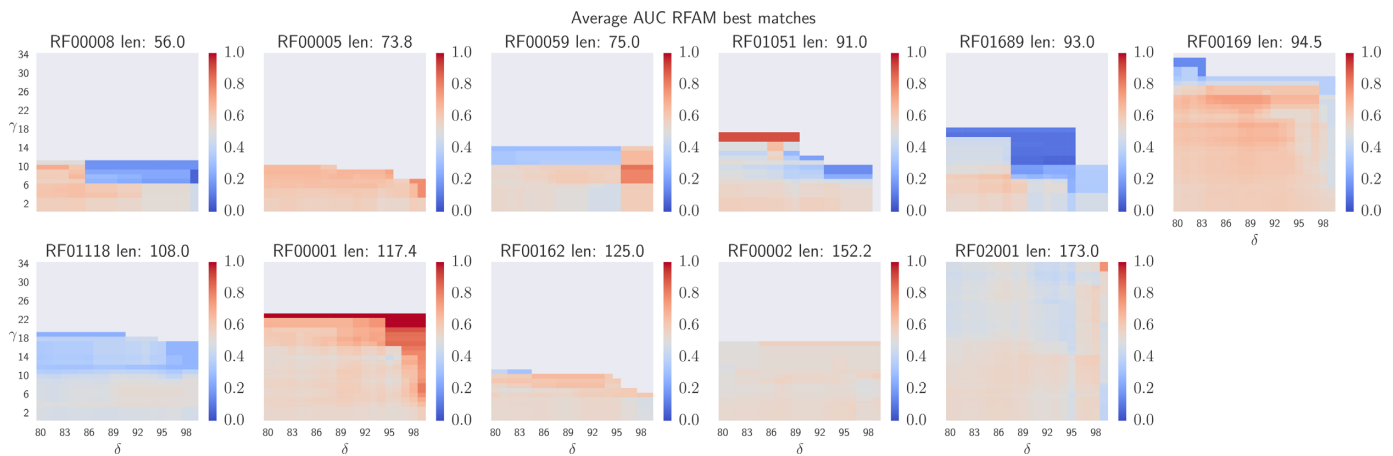


Figure 7. Performance of aRNhAck for remuRNA-predicted disruptions. For each Rfam family, we consider all PDBs having <150 nucleotides, and having maximal matching score to family. For a set of extreme percentile cutoff of the SHAPE profile disruption in the first column (computational remuRNA disruption in the second column) δ and a minimal distance γ from the mutation. Note that the PDB models considered for the 5S family (RF0001) do not match those investigated by MaM, which explains the discrepancies observed between the results above and those of Figure 4.

alone, but then we also lose all specificity since the intersection sets are too small and appear to mainly identify mutations not found in the MSA. Those results are shown in Supplementary Figure S5.

These observations implies that although a theoretical model do capture part of the complexity of the structural conformation ensemble, it is currently too noisy to identify fine grain differences captured by the SHAPE experiments.

CONCLUSION

We have presented a novel paradigm for analyzing non-coding RNA sequences combining the biochemical signal collected from structure probing experiments on RNA mutants, with the evolutionary information available in MSAs. We applied this model using MaM and Rfam data, and show that the signal extracted with our technology yields promising performance for identifying nucleotides involved in molecular interfaces.

A broad range of methods have been produced to predict RNA–Protein interactions (34) or RNA–RNA interactions (35–37). Yet, the vast majority of these programs aim to identify potential molecular targets from a library, and predict the best fits. By contrast, aRNhAck focuses on the sole biochemical and evolutionary properties of the RNA being analyzed. It enables, for the first time without prior knowledge of potential partners, the identification of *hot-spots* in RNA, involved in RNA–RNA, RNA–Protein, RNA–DNA and RNA–ligand interfaces, i.e. sets of critical nucleotides possibly implicated in the molecular functions. This information could then in turn be used to identify molecular targets or more realistically restrict the degree of freedom of molecular docking software (38).

This result illustrates the usefulness of the signal extracted by aRNhAck, but the scope of application of these concepts should not remain limited to quaternary structures. For instance, we envision to use the nucleotide networks detected with aRNhAck to predict non-canonical interactions and 3D motifs within an RNA molecules.

The main contribution of this work is to show that neutral theory principles can be combined with structure probing experiments to calculate complex evolutionary signals embedded in ncRNA sequences. aRNhAck aims to be a model for a new family of RNA sequence/structure analysis methods.

The volume of applications of aRNhAck is currently limited by the number of available datasets. Nonetheless, we showed that, to some extent, experimental data could be replaced by computationally-predicted data. Moreover, the democratization of molecular probing experiments suggests that this framework will be a valuable resource to exploit new datasets and discover elaborated networks essential for the functional properties of RNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Olivier Tremblay-Savard for his useful comments and suggestions about tRNA binding interfaces.

FUNDING

NSERC CGS Fellowship (to V.R.); French Fondation pour la Recherche Médicale [FRM DBI20141231337 to Y.P.]; FQRNT new researchers grant [239215]; NSERC Discovery grants [219671, 241015 to J.W.]. Funding for open access charge: NSERC.

Conflict of interest statement. None declared.

REFERENCES

- Wilkinson, K.A., Merino, E.J. and Weeks, K.M. (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.

2. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y. and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
3. Turner, D.H. and Mathews, D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.
4. Washietl, S., Hofacker, I.L., Stadler, P.F. and Kellis, M. (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.*, **40**, 4261–4272.
5. Deigan, K.E., Li, T.W., Mathews, D.H. and Weeks, K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.
6. Kladwang, W., VanLang, C.C., Cordero, P. and Das, R. (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.*, **3**, 954–962.
7. Theis, C., Höner Zu Siederdisen, C., Hofacker, I.L. and Gorodkin, J. (2013) Automated identification of RNA 3D modules with discriminative power in RNA structural alignments. *Nucleic Acids Res.*, **41**, 9999–10009.
8. Gardner, P.P. and Eldai, H. (2015) Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res.*, **43**, 691–698.
9. Schuster, P., Fontana, W., Stadler, P.F. and Hofacker, I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.*, **255**, 279–284.
10. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
11. Cordero, P., Lucks, J.B. and Das, R. (2012) An RNA mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics*, **28**, 3006–3008.
12. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. and Weeks, K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127**, 4223–4231.
13. Theis, C., zu Siederdisen, C.H., Hofacker, I.L. and Gorodkin, J. (2013) Automated identification of RNA 3D modules with discriminative power in RNA structural alignments. *Nucleic Acids Res.*, **41**, 9999–10009.
14. Reinharz, V., Ponty, Y. and Waldspühl, J. (2013) Using structural and evolutionary information to detect and correct pyrosequencing errors in noncoding RNAs. *J. Comput. Biol.*, **20**, 905–919.
15. Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
16. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
17. Hagberg, A.A., Schult, D.A. and Swart, P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, pp. 11–15.
18. Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
19. Schrödinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1. PyMOL The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.
20. Seidelt, B., Innis, C.A., Wilson, D.N., Gartmann, M., Armache, J.-P., Villa, E., Trabuco, L.G., Becker, T., Mielke, T., Schulten, K. *et al.* (2009) Structural insight into nascent polypeptide chain-mediated translational stalling. *Science*, **326**, 1412–1415.
21. Dunkle, J.A., Xiong, L., Mankin, A.S. and Cate, J.H. (2010) Structures of the Escherichia coli ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 17152–17157.
22. Llano-Sotelo, B., Dunkle, J., Klepacki, D., Zhang, W., Fernandes, P., Cate, J.H. and Mankin, A.S. (2010) Binding and action of CEM-101, a new fluoroketolide antibiotic that inhibits protein synthesis. *Antimicrob. Agents Chemother.*, **54**, 4961–4970.
23. Kulshina, N., Baird, N.J. and Ferré-D'Amaré, A.R. (2009) Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nat. Struct. Mol. Biol.*, **16**, 1212–1217.
24. Smith, K.D., Lipchock, S.V., Livingston, A.L., Shanahan, C.A. and Strobel, S.A. (2010) Structural and biochemical determinants of ligand binding by the c-di-GMP riboswitch. *Biochemistry*, **49**, 7351–7359.
25. Smith, K.D., Lipchock, S.V., Ames, T.D., Wang, J., Breaker, R.R. and Strobel, S.A. (2009) Structural basis of ligand binding by a c-di-GMP riboswitch. *Nat. Struct. Mol. Biol.*, **16**, 1218–1223.
26. Kulshina, N., Baird, N.J. and Ferré-D'Amaré, A.R. (2009) Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nat. Struct. Mol. Biol.*, **16**, 1212–1217.
27. Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
28. 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**, e65.
29. Schwarz, U., Menzel, H.M. and Gassen, H.G. (1976) Codon-dependent rearrangement of the three-dimensional structure of phenylalanine tRNA, exposing the T-ψ-CG sequence for binding to the 50S ribosomal subunit. *Biochemistry*, **15**, 2484–2490.
30. Schattner, P., Brooks, A.N. and Lowe, T.M. (2005) The tRNAscan-SE, snSCAN and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**(Suppl. 2), W686–W689.
31. Moreland, J.L., Gramada, A., Buzko, O.V., Zhang, Q. and Bourne, P.E. (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*, **6**, 21.
32. Salari, R., Kimchi-Sarfaty, C., Gottesman, M.M. and Przytycka, T.M. (2012) Detecting SNP-induced structural changes in RNA: application to disease studies. In: *Research in Computational Molecular Biology*. Springer, Berlin Heidelberg, pp. 241–243.
33. Sabarinathan, R., Tafer, H., Seemann, S.E., Hofacker, I.L., Stadler, P.F. and Gorodkin, J. (2013) RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum. Mutat.*, **34**, 546–556.
34. Puton, T., Kozłowski, L., Tuszyńska, I., Rother, K. and Bujnicki, J.M. (2012) Computational methods for prediction of protein–RNA interactions. *J. Struct. Biol.*, **179**, 261–268.
35. Hofacker, I.L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi1202s04.
36. Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M. and Pierce, N.A. (2011) NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.*, **32**, 170–173.
37. Wright, P.R., Georg, J., Mann, M., Sorescu, D.A., Richter, A.S., Lott, S., Kleinkauf, R., Hess, W.R. and Backofen, R. (2014) CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res.*, **42**, W119–W123.
38. Lopes, A., Sacquin-Mora, S., Dimitrova, V., Laine, E., Ponty, Y. and Carbone, A. (2013) Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. *PLoS Comput. Biol.*, **9**, e1003369.