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

Vladimir Reinharz, Yann Ponty, Jérôme Waldispühl

Published online XXXXX *Nucleic Acids Research, XXXXX, Vol. XXXXX, No. XXXXX* ***1–10***
doi:10.1093/nar/gkn000

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

Vladimir Reinharz[1], **Yann Ponty**[2], **Jérôme Waldispühl**[1]

[1] School of Computer Science, McGill University, Montreal, Canada
[2] Laboratoire d'informatique, École Polytechnique, Palaiseau, France.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Systematic structure probing experiments (e.g. SHAPE) of RNA mutants such as the mutate-and-map 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 mutate-and-map 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) (35) and parallel analysis of RNA structure (PARS) (10), which provide transcriptome-wide measurements of RNA structure at single-nucleotide resolution *in vitro*. Combined with classical RNA nearest neighbor energy minimization models (33), this data allowed for a significant improvement of the accuracy of RNA secondary structure prediction methods (3, 34).

Recently, R. Das and colleagues introduced the mutate-and-map protocol, which consists in obtaining SHAPE data simultaneously for an RNA and for (a large number of) its 1-point mutants (11). 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 mutate-and-map data for RNA secondary structure prediction uses empirical rules and pseudo-energy bonuses within classical dynamic programming prediction algorithms (3).

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 (6, 31).

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 mutate-and-map 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 mutate-and-map experiments – and the evolutionary information available for this molecule – usually contained in multiple sequence alignments.

This paper attempts to look beyond RNA structure determination, and introduces a novel concept to leverage the information embedded in experimental structure probing data sets of mutant RNAs. We apply neutral theory principles (26) to detect functional dependencies between distant nucleotides in a single stranded RNA. More precisely, we first use mutate-and-map 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 multiple sequence alignments (Rfam database (18)) homologous sequences containing those destabilizing mutations, and compare their nucleotide distribution to the background distribution observed in the Rfam multiple sequence alignment. 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 mutate-and-map data sets available on the RNA Mapping Database (2). 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 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.

METHODS

Definitions

We abstract an RNA sequence w of length n as a string in $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U}\}^n$. 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)**:

- **Stems** consist in one or more base pairs $\{(i_1,j_1),\cdots,(i_l,j_l)\}$ such that $i_m+1=i_{m+1}$ and $j_m-1=j_{m+1}$;
- **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;
- **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;
- **Multi-loops** are a set of base pairs $\{(i_1,j_1),\cdots,(i_l,j_l)\}$ where $l>2, i_1<i_2<j_2<\cdots<i_l<j_l<j_1$ and all k such that $i_1<k<i_2 \bigcup j_l<k<j_1 \bigcup_{2\leq m<l} j_m<k<i_{m+1}$ and k does not belong to any base pair.
- **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 secondary structure elements is given in Fig. 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.

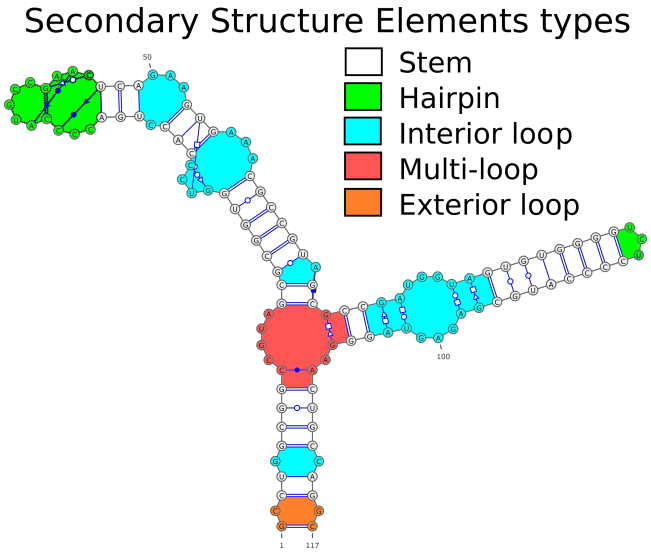


Figure 1. RNA secondary structure elements. Blue edges represent canonical base pairs, while black indicate the phosphodiester backbone. The hairpins are colored in green, the interior loops in blue, the multi-loops in red and the exterior loop in orange. The base pairs adjacent to those elements are part of them (e.g. the base pair 9-20 both belongs to a stem and to the interior loop).

SHAPE and Structural Information

SHAPE experiments (16) 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 (35). Many methods (3, 34) 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* (3) 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 greater than $1.5\times$ 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 **mutate-and-map** (11) 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 a sa. Thus, the sequence of the mutant associated with position i is entirely characterized

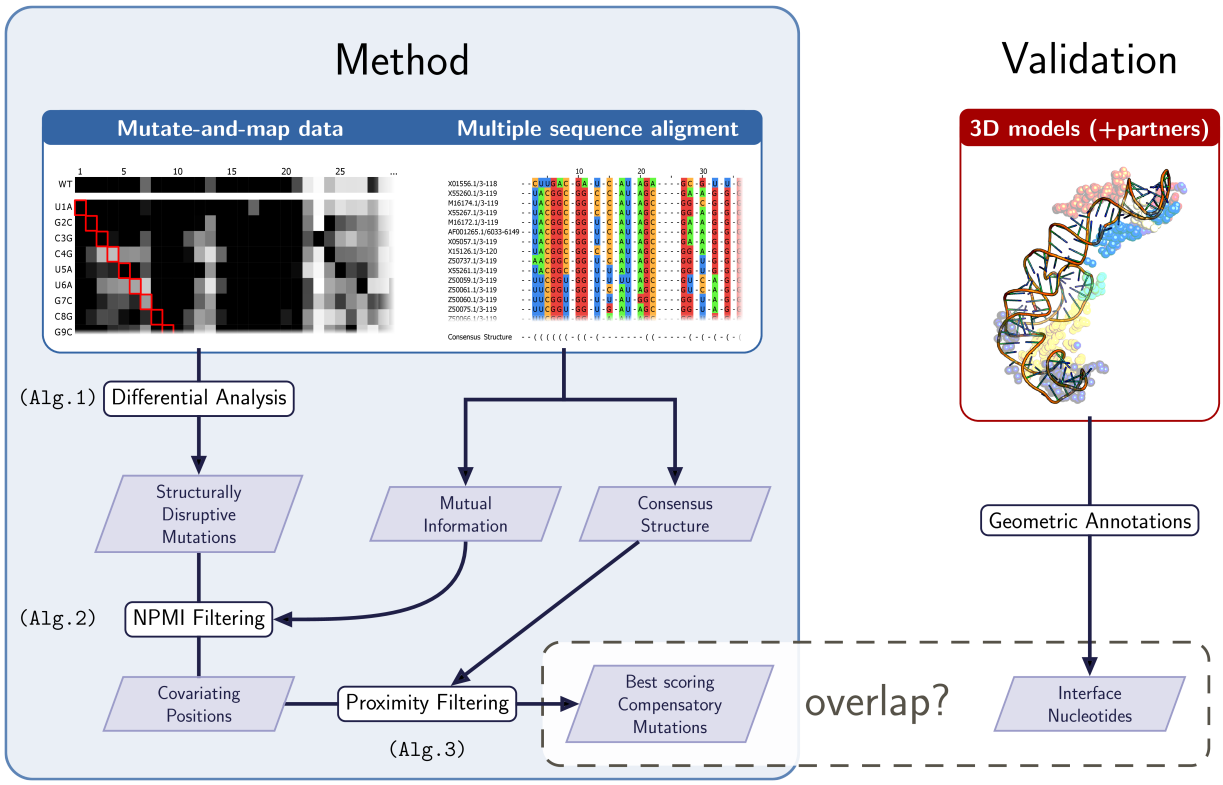


Figure 2. General overview of our method and validation procedure. Starting from mutate-and-map data, supplemented by an Rfam multiple sequence alignment, our method identifies positions that are simultaneously distant, yet co-evolve jointly, with positions that are structurally disruptive.

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 mutate-and-map 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 mutate-and-map 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 (Fig. S1); The second is a variant of Δ which considers the maximally contributing window over the whole sequence (instead of only considering the one centered on i for Δ), and aims at identifying non-local structural rearrangements (Fig. 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 three-dimensional disruption of the SHAPE profile (Fig. 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 identity mutations associated with significant changes of the structure. More precisely, given a percentile δ and a mutate-and-map (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 Algo. 1).

Evolutionary Information

Evolutionary information can be used to witness how nature *repairs* non-lethally disruptive mutations to preserve or reestablish a given phenotype/function. It has been used for a wealth of applications, ranging from the detection of RNA 3D modules (32) to the correction of pyrosequencing errors (21).

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 (5). The Rfam database (18) 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 multiple sequence alignment of Rfam to supplement structural disruption data, and use this additional knowledge, in conjunction 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 (5). 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:

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

where probabilities $\mathbb{P}(\cdot)$ are estimated from their frequencies in the multiple sequence alignment.

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 NPMIs 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 multiple sequence alignment MSA is described in Algo. 2.

Structures as Graphs Most disruptive mutations, when observed in multiple alignments, are found in combination with compensatory mutations which reestablish 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 three dimensional, 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. Fig 3 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.

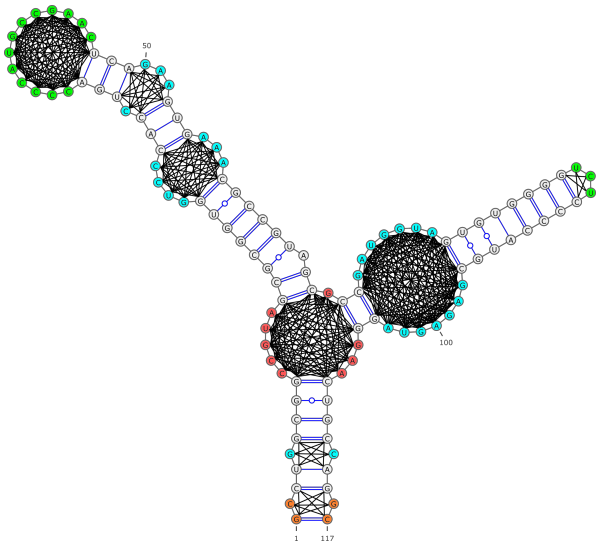


Figure 3. RNA secondary structure graph used for proximity filtering.

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 Algo. 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 NPMIs. 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, NPMIs 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)

$l = D = \varnothing$
for $m \in MaM$ **do**
  $l \leftarrow l \cup$ `shapeDisruption`(m, MaM)
for $m \in MaM$ **do**
  if `shapeDisruption`$(m, MaM) \geq$ `percentile`(l, δ) **then**
    $D \leftarrow D \cup \{m\}$
return D

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

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

Algorithm 3: filterNearbyPositions(m, S, γ)

$p = \varnothing$
$g \leftarrow$ `SGraph`(S)
for $u \in$ `getSSEs`(g) **do**
  if `distance`$([m], u) \geq \gamma$ **then**
    $p \leftarrow p \cup \{u\}$
return p

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

$M_{\text{disruption}} \leftarrow$ `disruptiveMutations`(MaM, δ)
for $m \in M_{disruption}$ **do**
  $p \leftarrow$ `filterNearbyPositions`(m, S, γ)
  $p \leftarrow$ `filterNPMI`(m, p, MSA, ζ_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` (1) for reading the multiple sequence alignments, `networkx` (7) for modeling the graphs, and `matplotlib` (9) for visualizing the results. The 3D complex analysis used in our validation is performed using the Python API provided by the `PyMol` software (25).

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

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

Dataset

We evaluated our methods on molecule for which we could obtain simultaneously (i) a mutate-and-map experiment data set, (ii) a determined three-dimensional structures interacting with other chain(s), and (iii) a Rfam alignment. 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 faciliate 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 (Fig. 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 mutate-and-map protocol was applied to the consensus sequence of 4 structures which have as PDB identifiers `2WWQ` (28), `3OAS` and `3OFC` (4), and `3ORB` (14). Those four determined structures have almost 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 identifier `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 identifier `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 4 structures, with PDB identifiers `3IWN` (12) and `3MXH`, `3MUV`, `3MUT` (30). Importantly, c-di-GMP is known to bind a pocket inside the 3-way junction at positions 11-13, 40-41 and 85 of the sequence on which the mutate-and-map experiments were run (13, 29), 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 identifier `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 identifier `1Y26`. Three different MaM experiments were conducted on this molecule. Experiments `Adenine_2` and `Adenine_3` where 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` (19) software was used with default parameter values to: 1) create a covariance model for each alignment, and; 2) align the sequence from the mutate-and-map 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 data set was composed of the regions of interest given γ, i.e. the set of positions returned by the Algo 3.

We used different strategies to determine the interaction sites (i.e. positive data set), 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 5 Å from any atom of another chain the the complex. An implementation of this script is included in the distribution of `aRNhAck`

For the tRNA, we extracted positions that are at most 5 Å away from another chain in the two tRNAs found inside the structure of the yeast 80S ribosome-tRNA complexes (PDB identifier `3J78`). However, because those were not phenylalanine tRNAs, we aligned them to the MaM sequence with `LocaRNA`(36), 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 (27), 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` (24).

For the riboswitches, we used `Ligand Explorer` (17) to identify nucleotide at most 5 Å from the ligand in their respective crystal structures.

The set of all positions is found in the Supplementary Material Table 1.

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 pairs 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 0^{th} to the 100^{th} percentile of the positive values of ζ (resp. negative values of ζ).

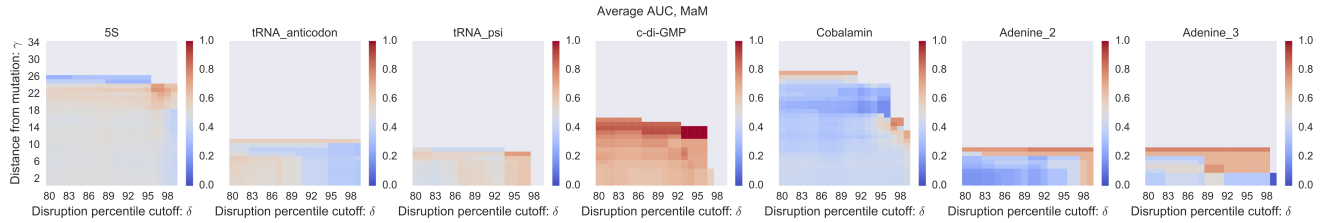
Thus, for each SHAPE profile distance measure, each SHAPE distance threshold δ, each significant mutation given that measure, each γ, each PDB model and for every threshold pair (ζ^-,ζ^+), we obtained standard sensitivity and specificity scores. Those with a given SHAPE profile distance measure, SHAPE distance threshold δ, γ, PDB model and pair (ζ^-,ζ^+) where averaged together. The workflow of the method is illustrated in Fig. 2.

RESULTS

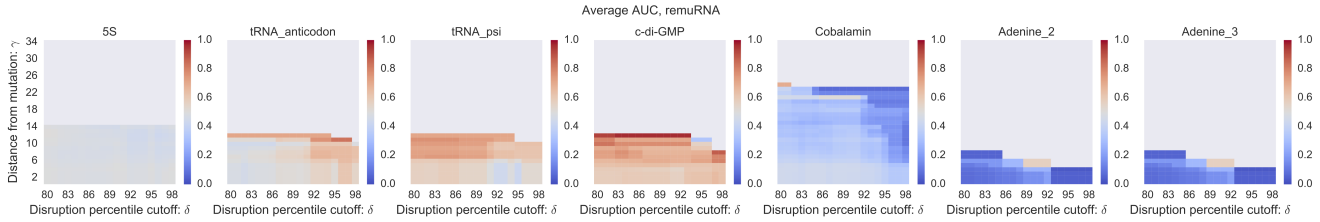
We evaluated `aRNhAck` on a comprehensive set of values for δ the SHAPE profile distance measure and γ the proximity threshold. For each (δ,γ) pair, we computed a Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC). The data are shown in Fig. 4a. Noticeably, we averaged the 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 Algo. 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 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 Fig. 4a. In all cases, there exists a pair of parameters (δ,γ) for which `aRNhAck` achieve good predictive performance. Importantly, the maximal AUCs are



(a) MaM



(b) `remuRNA`

Figure 4. Overall performances of `aRNhAck` using experimental and computationally-predicted structural disruption data. 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 litterature. The c-di-GMP, cobalamin and adenine riboswitches positive sets are composed of the positions at most 5 Å from their ligands in their PDB structures. Four different models exist for c-di-GMP and the AUC values are averaged.

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 amplifed noise in MaM experiments.

By constrast, the c-di-GMP riboswitch exhibits one of the strongest signal, 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 Fig. S9 in supplementary material).

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 Fig. 5 the distribution of path lengths (distance) between every pair of secondary structure elements, weighted by the number of combinations of positions that are not in the intersection of the secondary structure element. 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 follows 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.

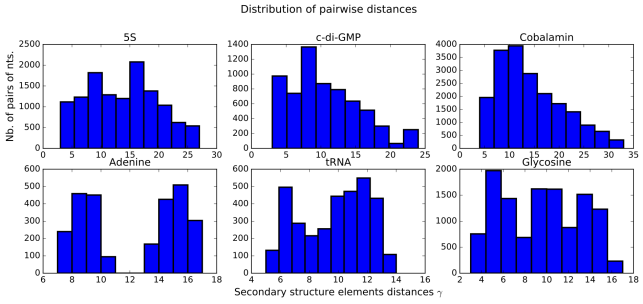


Figure 5. Distance distribution for pairs of secondary structure elements, weighted by the numbers of non-shared nucleotides. The different distribution affect the quantity of positions selected by the parameter γ, distance from the mutation.

In Fig. 6 we visualize one prediction from our method, for the 5s chain of `3OFC`. 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 5 Å 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.

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

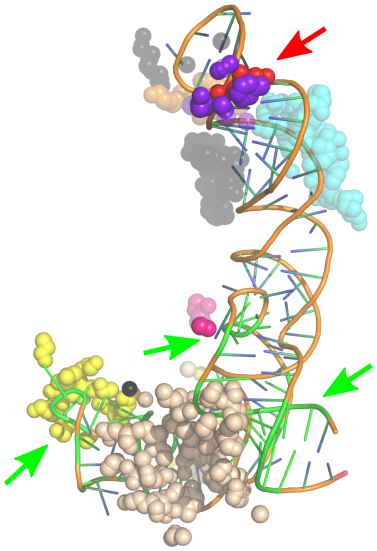


Figure 6. Predicted positions and interacting chains of the 5S rRNA 3OFC structure. In red on the top right behind purple spheres is the disrupting mutation, in green the predicted position with high mutual information. The spheres around the RNA represent the subset of nucleotides at most at 5 Å from 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.

Computationally predicted structural disruptions yield weaker signal

To justify the use of SHAPE experimental data, we evaluate the performance of a fully automated pipe-line in which mutations altering the RNA conformational landscape are predicted with a computer software instead of mutate-and-map data. Here, we predict these destabilizing mutations with `remuRNA` (23). Alternatively, for longer sequences, `RNAsnp` (22) 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` (23) 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 $\mathcal{S}$, 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 Fig. 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 mutate-and-map 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 data set made of Rfam families with experimentally determined 3D structures (See Methods). 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 Fig. 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 Fig. S6).

To a lesser extend, the same trend is observed through all the matches annotated by Rfam (See Fig. S6). The full list of matching families in 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 Fig. S4, we notice a clear linear decrease in the size of the overlap. At the 50^{th} 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 alone, but then we also lose all specificity since the intersection sets are too small and

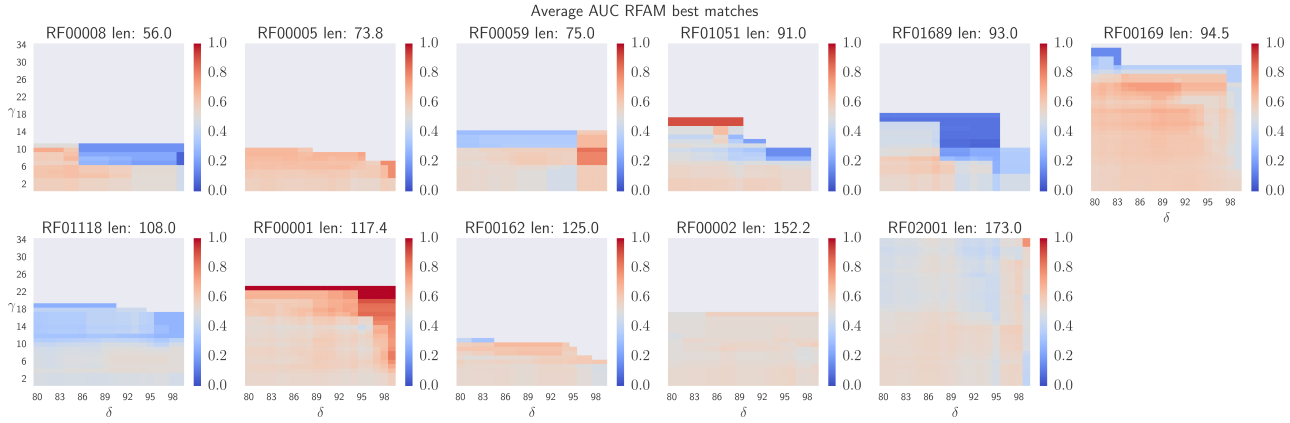


Figure 7. Performance of `aRNhAck` for `remuRNA`-predicted disruptions. For each Rfam family, we consider all PDBs having less than 150 nucleotides, and having maximal matching score to family. For a set of extreme percentile cutoff of the SHAPE profile disruption in the first column (computational `remuRNA` 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 Fig. 4.

appear to mainly identify mutations not found in the multiple sequence alignment. Those results are shown in Fig. 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 multiple sequence alignments. We applied this model using mutate-and-map 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 (20) or RNA-RNA interactions (8, 37, 38). 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 (15).

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 data sets. 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 data sets and discover elaborated networks essential for the functional properties of RNAs.

ACKNOWLEDGMENTS

The authors would like to thank Olivier Tremblay-Savard for his useful comments and suggestions about tRNA binding interfaces. This work was funded by a NSERC CGS fellowship (to VR), a French *Fondation pour la Recherche Médicale* grant (to YP), and a FQRNT team grant 239215 and NSERC Discovery grants 219671 & 241015 (to JW).

Conflict of interest statement. None declared.

REFERENCES

1. Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.
2. Pablo Cordero, Julius B Lucks, and Rhiju Das. An RNA mapping database for curating RNA structure mapping experiments. *Bioinformatics*, 28(22):3006–8, Nov 2012.
3. Katherine E Deigan, Tian W Li, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A*, 106(1):97–102, Jan 2009.
4. Jack A Dunkle, Liqun Xiong, Alexander S Mankin, and Jamie HD Cate. Structures of the escherichia coli ribosome with antibiotics bound near the

peptidyl transferase center explain spectra of drug action. *Proceedings of the National Academy of Sciences*, 107(40):17152–17157, 2010.
5. Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, 1994.
6. Paul P Gardner and Hisham Eldai. Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res*, Dec 2014.
7. Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.
8. Ivo L Hofacker. RNA secondary structure analysis using the vienna RNA package. *Current Protocols in Bioinformatics*, pages 12–2, 2009.
9. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.
10. Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7, Sep 2010.
11. Wipapat Kladwang, Christopher C VanLang, Pablo Cordero, and Rhiju Das. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem*, 3(12):954–62, Dec 2011.
12. Nadia Kulshina, Nathan J Baird, and Adrian R Ferré-D'Amaré. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nature structural & molecular biology*, 16(12):1212–1217, 2009.
13. Nadia Kulshina, Nathan J Baird, and Adrian R Ferré-D'Amaré. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nature structural & molecular biology*, 16(12):1212–1217, 2009.
14. Beatriz Llano-Sotelo, Jack Dunkle, Dorota Klepacki, Wen Zhang, Prabhavathi Fernandes, Jamie HD Cate, and Alexander S Mankin. Binding and action of cem-101, a new fluoroketolide antibiotic that inhibits protein synthesis. *Antimicrobial agents and chemotherapy*, 54(12):4961–4970, 2010.
15. Anne Lopes, Sophie Sacquin-Mora, Viktoriya Dimitrova, Elodie Laine, Yann Ponty, and Alessandra Carbone. Protein-protein interactions in a crowded environment: An analysis via cross-docking simulations and evolutionary information. *PLoS Comput Biol*, 9(12):e1003369, 12 2013.
16. Edward J Merino, Kevin A Wilkinson, Jennifer L Coughlan, and Kevin M Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *Journal of the American Chemical Society*, 127(12):4223–4231, 2005.
17. John L Moreland, Apostol Gramada, Oleksandr V Buzko, Qing Zhang, and Philip E Bourne. The molecular biology toolkit (mbt): a modular platform for developing molecular visualization applications. *BMC bioinformatics*, 6(1):21, 2005.
18. Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, et al. Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, page gku1063, 2014.
19. Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.
20. Tomasz Puton, Lukasz Kozlowski, Irina Tuszynska, Kristian Rother, and Janusz M Bujnicki. Computational methods for prediction of protein–rna interactions. *Journal of structural biology*, 179(3):261–268, 2012.
21. Vladimir Reinharz, Yann Ponty, and Jérôme Waldispühl. Using structural and evolutionary information to detect and correct pyrosequencing errors in noncoding rnas. *Journal of Computational Biology*, 20(11):905–919, 2013.
22. Radhakrishnan Sabarinathan, Hakim Tafer, Stefan E Seemann, Ivo L Hofacker, Peter F Stadler, and Jan Gorodkin. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Human mutation*, 34(4):546–556, 2013.
23. Raheleh Salari, Chava Kimchi-Sarfaty, Michael M Gottesman, and Teresa M Przytycka. Detecting SNP-induced structural changes in RNA: application to disease studies. In *Research in Computational Molecular Biology*, pages 241–243. Springer, 2012.
24. Peter Schattner, Angela N Brooks, and Todd M Lowe. The trnascan-se, snoscan and snogps web servers for the detection of trnas and snornas. *Nucleic acids research*, 33(suppl 2):W686–W689, 2005.
25. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. PyMOL The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC., August 2010.
26. P Schuster, W Fontana, P F Stadler, and I L Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci*, 255(1344):279–84, Mar 1994.
27. Ulrich Schwarz, Heinrich M Menzel, and Hans G Gassen. 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(11):2484–2490, 1976.
28. Birgit Seidelt, C Axel Innis, Daniel N Wilson, Marco Gartmann, Jean-Paul Armache, Elizabeth Villa, Leonardo G Trabuco, Thomas Becker, Thorsten Mielke, Klaus Schulten, et al. Structural insight into nascent polypeptide chain–mediated translational stalling. *Science*, 326(5958):1412–1415, 2009.
29. Kathryn D Smith, Sarah V Lipchock, Tyler D Ames, Jimin Wang, Ronald R Breaker, and Scott A Strobel. Structural basis of ligand binding by a c-di-GMP riboswitch. *Nature structural & molecular biology*, 16(12):1218–1223, 2009.
30. Kathryn D Smith, Sarah V Lipchock, Alison L Livingston, Carly A Shanahan, and Scott A Strobel. Structural and biochemical determinants of ligand binding by the c-di-gmp riboswitch. *Biochemistry*, 49(34):7351–7359, 2010.
31. Corinna Theis, Christian Höner Zu Siederdissen, Ivo L Hofacker, and Jan Gorodkin. Automated identification of RNA 3d modules with discriminative power in RNA structural alignments. *Nucleic Acids Res*, 41(22):9999–10009, Dec 2013.
32. Corinna Theis, Christian Höner zu Siederdissen, Ivo L Hofacker, and Jan Gorodkin. Automated identification of RNA 3d modules with discriminative power in RNA structural alignments. *Nucleic acids research*, 41(22):9999–10009, 2013.
33. Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(Database issue):D280–2, Jan 2010.
34. Stefan Washietl, Ivo L Hofacker, Peter F Stadler, and Manolis Kellis. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*, 40(10):4261–72, May 2012.
35. Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (shape): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc*, 1(3):1610–6, 2006.
36. Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding rna families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, 2007.
37. Patrick R Wright, Jens Georg, Martin Mann, Dragos A Sorescu, Andreas S Richter, Steffen Lott, Robert Kleinkauf, Wolfgang R Hess, and Rolf Backofen. Coprarna and intarna: predicting small RNA targets, networks and interaction domains. *Nucleic acids research*, page gku359, 2014.
38. Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.