bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Stochastic Sampling of Structural Contexts Improves the Scalability and Accuracy of RNA 3D Modules Identification

Roman Sarrazin-Gendron[1], Hua-Ting Yao[1,2], Vladimir Reinharz[3], Carlos G. Oliver[1], Yann Ponty[2], and Jérôme Waldispühl[1]

[1] School of Computer Science, McGill University, Montreal, Canada
[2] LIX, CNRS UMR 7161, Ecole Polytechnique, Palaiseau, France
[3] Center for Soft and Living Matter, Institute for Basic Science, Ulsan, South Korea

Abstract. RNA structures possess multiple levels of structural organization. Secondary structures are made of canonical (i.e. Watson-Crick and Wobble) helices, connected by loops whose local conformations are critical determinants of global 3D architectures. Such local 3D structures consist of conserved sets of non-canonical base pairs, called RNA modules. Their prediction from sequence data is thus a milestone toward 3D structure modelling. Unfortunately, the computational efficiency and scope of the current 3D module identification methods are too limited yet to benefit from all the knowledge accumulated in modules databases.

Here, we introduce `BayesPairing2`, a new sequence search algorithm leveraging secondary structure tree decomposition which allows to reduce the computational complexity and improve predictions on new sequences. We benchmarked our methods on 75 modules and 6360 RNA sequences, and report accuracies that are comparable to the state of the art, with considerable running time improvements. When identifying 200 modules on a single sequence, `BayesPairing2` is over 100 times faster than its previous version, opening new doors for genome-wide applications.

Keywords: RNA structure prediction · RNA 3D modules · RNA modules identification in sequence.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

RNAs use complex and well organized folding processes to support their many non-coding functions. The broad conservation of structures across species highlights the importance of this mechanism [35,14]. RNAs can operate using folding dynamics [25] or hybridization motifs [2]. Yet, many highly specific interactions need sophisticated three dimensional patterns to occur [15,11,13].

RNAs fold hierarchically [36]. First, Watson-Crick and Wobble base pairs are rapidly assembled into a secondary structure that determine the topology the RNA. Then, unpaired nucleotides form non-canonical base pairs interactions [16], stabilizing the loops while shaping the tertiary structure of the molecule. These non-canonical base pairing networks have thus been identified as critical components of the RNA architecture [4] and several catalogs of recurrent networks along with their characteristic 3D geometries are now available [10,7,27,28,30,12]. They act has structural organizers and ligand-binding centers [8] and we call them *RNA 3D modules*.

In contrast to well-established secondary structure prediction tools [20,22], we are still lacking efficient computational methods to leverage the information accumulated in the module databases. Software such as `RMDetect` [8], `JAR3D` [34] and our previous contribution `BayesPairing 1` [32] have been released, but their precision and scalability remains a major bottleneck.

The significance of a module occurrence is typically assessed from recurrence: substructures that are found in distinct RNA structures are assumed to be functionally significant [30]. Based on this hypothesis, three approaches have been developed so far for the retrieval and scoring of 3D modules from sequence. The first one, `RMDetect`, takes advantage of Bayesian Networks to represent base pairing tendencies learned from sequence alignments. Candidate modules found in an input sequence are then scored with Bayesian probabilities. However, while showing excellent accuracy, `RMDetect` suffers from high computational costs, and minimal structure diversity among modules predicted [32]. Another option is `JAR3D` [34], which refined the graphical model-based scoring approach introduced by `RMDetect` and represents the state of the art for module scoring. However, it was not designed to maximize input sequence scanning efficiency and is limited in module diversity, only being applied to hairpin and internal loops. Finally, `BayesPairing 1` [32], a recently introduced tool combining the Bayesian scoring of `RMDetect` to a regular expression based sequence parsing, is able to identify junction modules in input sequences and showed improved computational costs compared to `RMDetect`, which it was inspired from. Unfortunately, none of the aforementioned software can be used for the discovery of many RNA 3D modules in new sequences at the genome scale.

In this paper, we present `BayesPairing 2`, a efficient tool for high-throughput search of RNA modules in sequences. `BayesPairing 2` analyzes the structural landscape of an input RNA sequence through secondary structure stochastic sampling and uses this information to identify candidate module insertion sites and select modules occurring in a favorable structural context. This pre-scoring stage enables us to dramatically reduce the number of putative matches and thus to (i) simultaneously search for multiple modules at once and (ii) eliminate false positives. `BayesPairing 2` shows comparable performance to the state of the art while scaling gracefully with the number of modules searched. It also supports alignment search, a feature of `RMDetect` which could not be integrated in the `BayesPairing 1` framework. All these improvements support potential applications at the genome scale.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

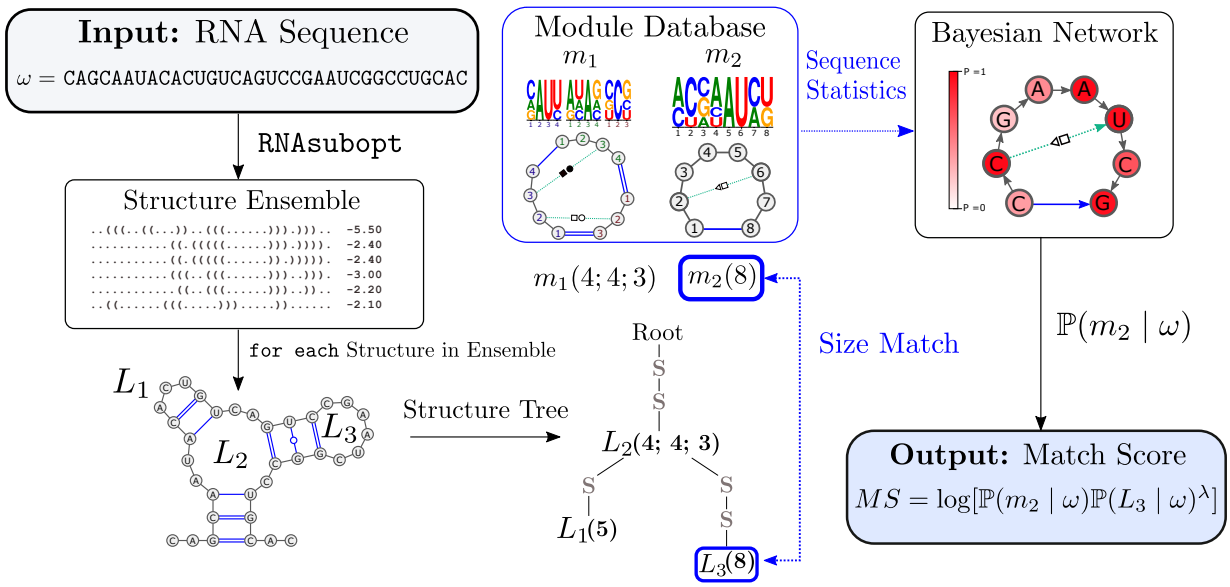


Fig. 1: The `BayesPairing2` workflow addresses the identification of non-canonical 3D modules, *i.e.* arrangements of canonical and non canonical base pairs that are essential to the 3D architecture of RNAs. It takes as input either an RNA transcript or a multiple sequence alignment, possibly supplemented with a (shared) secondary structure, and returns an ordered list of occurrences for candidate modules. Its key idea is to match predicted secondary structure loops, highly likely to occur in thermodynamically-stable models, against a database of local modules learned from sequence data filtered for isostericity [19]. In this figure, we show the identification pipeline for one module on one structure of the ensemble. This is then repeated for all modules, for all structures.

2 Methods

Concepts and model. A **non-canonical 3D module** consists in a set of non-canonical base pairs [17]. Modules occur within a **secondary structure loop**, consisting of one or several stretches of unpaired positions within an RNA transcript, also called **regions**, delimited by classic Watson-Crick/Wobble base pairs.

At the **thermodynamic equilibrium**, an RNA sequence w is expected to behave stochastically and adopt any of its secondary structure S, **compatible** with w with respect to canonical Watson-Crick/Wobble base pairing rules, with probability proportional to its **Boltzmann factor** [23]. The **Boltzmann probability** of a secondary structure S for an RNA sequence w is then given by

$$\mathbb{P}(S \mid w) = \frac{e^{-E_{S,w}/RT}}{\mathcal{Z}_w}$$

where $E_{S,w}$ represents the free-energy assigned to the (S, w) pair by the experimentally-established Turner energy model [37], $\mathcal{Z}_w = \sum_{S'} e^{-E_{S',w}/RT}$ is the partition function [23], R is the Boltzmann constant and T the absolute temperature. By extension, the Boltzmann probability of a given loop to occur within a sequence w is simply defined as

$$\mathbb{P}(\text{loop} \mid w) = \sum_{\substack{S \text{ compatible with } w \\ \text{loop} \in S}} \mathbb{P}(S \mid w).$$

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the current absence of thermodynamic data for non-canonical base-pairs and modules, we adopt a probabilistic approach, and model the sequence preferences associated with a module statistically as a **Bayesian network**, following Cruz *et al* [8]. The structures of Bayesian networks are systematically derived from the base pairs occurring within **recurrent 3D motifs** [30]. Such motifs are typically mined within available 3D RNA structures in the PDB [5], and clustered geometrically. The network is then transformed into **conditional emission probabilities** for individual positions in the module, using a **tree decomposition** [6] of the network to minimize the maximum number of prior observations for each position, akin to the junction tree method [3]. Maximum likelihood emission probabilities are then learned for each module using pseudo-counts.

The **emission probability** for the positions of a module m to be assigned to a nucleotide content A is then given by

$$\mathbb{P}(\text{assignment } A \mid \text{module } m) = \prod_{i \in m} \mathbb{P}(p_i = A_i \mid p_j = A_j \wedge p_{j'} = A_{j'} \wedge \ldots). \tag{1}$$

where p_j, $p_{j'}, \ldots$ represent the content of positions $j, j', \ldots$, the positions conditioning the content p_i of position i, as derived using the tree decomposition, and A_i represents the content of the i-th position in A. Using Bayes Theorem while assuming uniform priors for both assignments and modules, we obtain

$$\begin{aligned}\mathbb{P}(\text{module } m \mid \text{assignment } A) &= \frac{\mathbb{P}(A \mid m) \times \mathbb{P}(m)}{\mathbb{P}(A)} \\ &= \frac{4^{|m|} \prod_{i \in m} \mathbb{P}(p_i = A_i \mid p_j = A_j \wedge \ldots)}{|\mathcal{M}|}.\end{aligned}$$

where $\mathcal{M}$ represents the set of admissible modules.

The final **match log-odds score** MS associated with a motif m being embedded within a given loop (*i.e.* at a given position) for an RNA sequence w is given by

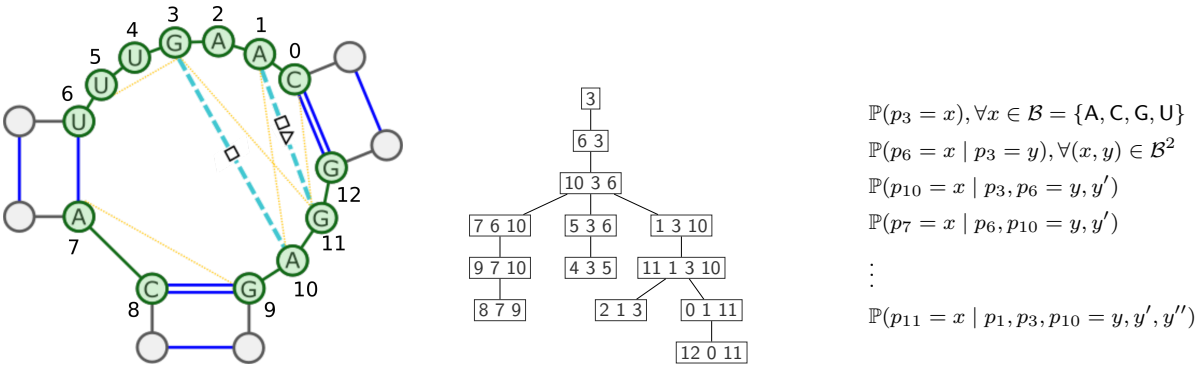
$$\text{MS} = \lambda \times \log\big(\mathbb{P}(\text{loop} \mid w)\big) + \sum_{i \in m} \log\big(\mathbb{P}(p_i = A_i \mid p_j = A_j \wedge \ldots)\big) + |m| \log 4 - \log(|\mathcal{M}|) \tag{2}$$

where λ is a term that allows to control the relative weight of the structure and local sequence composition.

Algorithmic considerations and complexity. On an algorithmic level, for given sequence w and module m, we remark that it suffices to optimize for the first two terms of the above equations, the others being constant for a given module. A list of loops having highest Boltzmann probability $\mathbb{P}(\text{loop} \mid w)$ is first estimated from a statistical sample, generated using (non-redundant) stochastic backtrack [9,24,31]. The second term, *i.e.* the probability of the module content, is only evaluated for the loops that are compatible with the size constraints of the module, with tolerance for a size mismatch of up to one base per strand ($-\infty$ otherwise). Its evaluation uses conditional probabilities, learned from a tree-decomposition of the module, as described in Figure 2. Matches featuring scores higher than a **cut-off** α are then reported as candidates.

The **overall complexity** of the method, when invoked with a module m and a transcript w of length n is in $\mathcal{O}(n^3 + kn \log n + \min(k, n^{2h(m)}) \times n \times |m|))$, where k denotes the number of sampled secondary structures and $h(m)$ is the total number of helices in m. It follows a sequence-agnostic precomputation in $\mathcal{O}(4^{w(m)} + |m| \times D)$, where $w(m)$ represents the tree-width of m, and D represents the overall size of the dataset used for training the model.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) RNA module (b) Tree decomposition (c) Conditional probabilities

Fig. 2: An RNA 3D module (2a), here the three-way junction of the TPP riboswitch, represented in green, drawn in its structural context. Dashed and dotted lines respectively represent non-canonical base pairs and stacking interactions. A tree decomposition (2b) of the module represents the dependencies between the module positions, leading to conditional probabilities (2c), estimated from available sequence data

Remark that, while our reliance on sampling formally makes our method a heuristic in the context of optimizing the objective in Equation (2), it must be noted that sampling provides a **statistically consistent** estimator for the probabilities of loops. Moreover, the probabilities associated with all possible loops could be computed exactly using constrained dynamic programming in time $\mathcal{O}(n^{3+2h(m)})$ [20].

Implementation. Secondary structures are non-redundantly sampled from the whole ensemble if the structure is not provided in the input, using `RNAsubopt` for a single sequence, or `RNAalifold` for a set of pre-aligned sequences [20,24,31]. The tree decomposition of a module is computed by `htd` library [21], then turned into conditional probabilities that are learned within `pgmpy` [1]. `BayesPairing 2` is freely available as a downloadable software at (http://csb.cs.mcgill.ca/BP2).

Positioning against previous version. Using stochastic sampling allows `BayesPairing 2` to efficiently score all modules of a dataset in a single sequence search, unlike the previous version, which requires multiple regex searches on the sequence for each module. While searching structure-first improves the sensitivity, especially on modules without a strong sequence signal, it can add potential false positives, especially for small modules which appear a lot in secondary structures. This translates into more candidates scored, but scoring a candidate is much faster than scanning a sequence. Thus, `BayesPairing 2` is much more more efficient when searching for many modules. In addition, the ability to sample with `RNAalifold` allows `BayesPairing 2` to take full advantage of aligned sequences.

3 Results

3.1 `Rna3Dmotif` dataset

In order to assess the performance of `BayesPairing 2` on its own and in context with that of `BayesPairing 1`, we assembled a representative sequence-based dataset of local RNA 3D modules. We ran `Rna3Dmotif` on the non-redundant RNA PDB structure database [18]. Identified modules

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were then matched to `Rfam` family alignments via 3D structure positions. Sequences from these alignments were filtered to remove poorly aligned sequences, using isostericity substitution cutoffs ensuring that the extracted sequences could adopt their hypothesized structure. Modules matched to at least 35 sequences were added to the dataset. 75 modules, totaling 20125 training sequences, were collected. To assess the presence and potential impact of false positives and true negatives, a negative dataset was assembled. To build this dataset, each sequence in the true positive dataset was shuffled while preserving its dinucleotide distribution. We assume motif occurrences to be homogeneous in length.

3.2 Validation on the `Rna3Dmotif` dataset

Validating searches on sequences with known structure. A first aspect to validate is the ability of our method to retrieve the module when the native secondary structure is provided, ensuring the availability of a suitable loop for the module. For this test, the sequences were obtained from the positive dataset, and the structures accommodating their respective modules were generated with `RNAfold` hard constraint folding. As expected, structure-informed BP2 recovers every existing module.

Joint prediction of secondary structure loops and module occurrences. To assess the performance of `BayesPairing2` on sequences of unknown structure, we performed two-fold cross-validation on 100 randomly sampled unique sequences (or on all sequences when fewer were available), for each module, amounting to a total of 6360 sequences. For each sequence-module pair, the candidate with highest score S through 20000 sampled structures was considered a true positive if its match score MS was above the score cutoff $T = -2.16$, and if its predicted position matched its real three-dimensional structure location. For the *5 candidates* sensitivity, any correct prediction within the top 5 candidates could be considered a true positive.

	F1 score	MCC	FDR	Sensitivity (1 candidate)	Sensitivity (5 candidates)
Avg accuracy per module	0.932	0.863	0.061	0.745	0.842

Table 1: `BayesPairing` module identification accuracy on `Rna3Dmotif` dataset

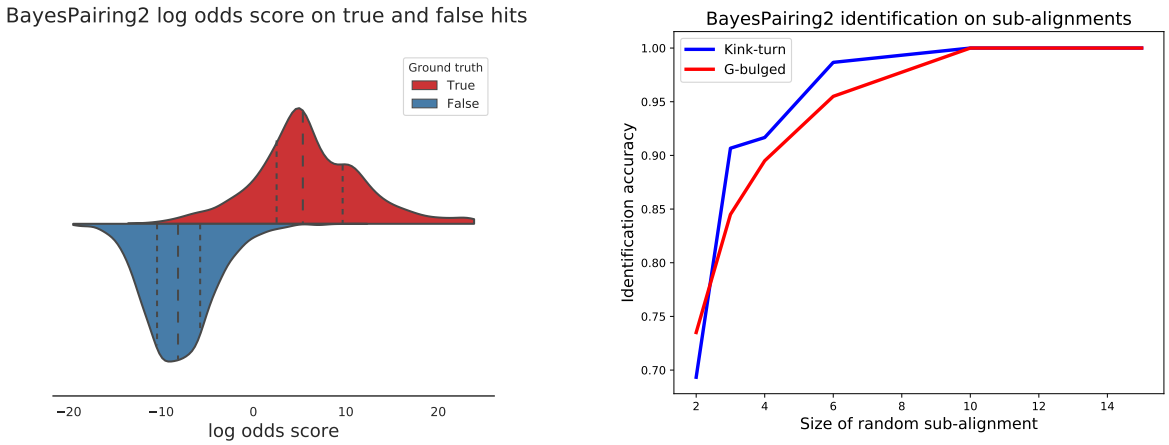
Prediction score distribution and false discovery rate. We executed the same two-fold cross-validation experiment on the shuffled sequences described in section 3.1. `BayesPairing2` found no on 92% of the 6360 sequences. It should be noted that it is not impossible for a shuffled sequence to contain a good hit for a module.

We obtained distributions of true and false hit scores from the cross-validation dataset. The score distributions, presented in Figure 3a, are clearly distinct, and a score cutoff of -2.16 produced a false discovery rate of 0.061, as reported along with other common metrics in Table 1. This cutoff was computed on the full dataset, and the optimal cutoff for each module was observed to rarely vary by more than 0.2.

3.3 Validation on known module alignments from `Rfam`

Sequence search. To complement the cross-validation experiments, `BayesPairing2` was also tested on `Rfam` alignments of the kink-turn and G-bulged internal loop modules. In these experiments, the modules were associated with their respective families through the `Rfam` motif

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) Identification probabilistic scores output by `BayesPairing 2` for 4500 true hits and 4500 false hits

(b) Prediction accuracy on sub-alignments for the kink-turn and g-bulged modules, by size of sub-alignment

Fig. 3: Evaluating `BayesPairing 2` scores and accuracy.

Trained on	**Identified on/with**					
Family	RF00162		RF02540		RF02541	
Software	BP1	BP2	BP1	BP2	BP1	BP2
RF00162	0.963	0.963	0.466	0.716	0.663	0.634
RF02540	0.300	0.820	0.989	0.807	0.673	0.723

(a) Kink-Turn

Trained on	**Identified on/with**			
Family	RF02540		RF02541	
Software	BP1	BP2	BP1	BP2
RF02540	0.977	0.977	0.911	0.990
RF02541	0.818	0.966	0.931	0.990

(b) G-bulged

Table 2: `Rfam` cross-family results for kink-turn (left) and G-bulged (right)

database, then trained on one family and tested on the other. The results, for `BayesPairing 1` and `BayesPairing 2`, are displayed in Tables 2a and 2b. We used standard parameters and selected the cutoffs associated to the same false discovery rate of 0.1 for both methods.

As observed in section 3.2, `BayesPairing 2` is slightly weaker at identifying modules with a strong sequence signal than `BayesPairing 1`, but considerably stronger when there is significant sequence variation as its signal appears to be more robust. This is particularly well illustrated by the capacity of `BayesPairing 2` to identify the ribosomal kink-turn module on SAM riboswitch sequences. While the considerable sequence difference between the ribosome and riboswitch causes a sharp drop of 47% in `BayesPairing 1` accuracy when predicting off-family, `BayesPairing 2` only loses 25%.

Alignment search improvement. Despite positive results in module identification on sequences taken from `Rfam`, sequence-based methods cannot fully take advantage of the common structure of an alignment. We show the relevance of including module identification on alignments in `BayesPairing 2` by improving the results presented in section 3.3. If, instead of parsing individual sequences for modules, we parse randomly sampled sub-alignments, the predictions rise with the size of the sub-alignment until they reach 100%, up from 50 to 95% with sequence predictions by both software tools. Despite very low sample size (500 secondary structure sampled with `RNAalifold`), the alignment quickly outperforms the sequence predictions for all modules, on all tested families, as shown in Figure 3b.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

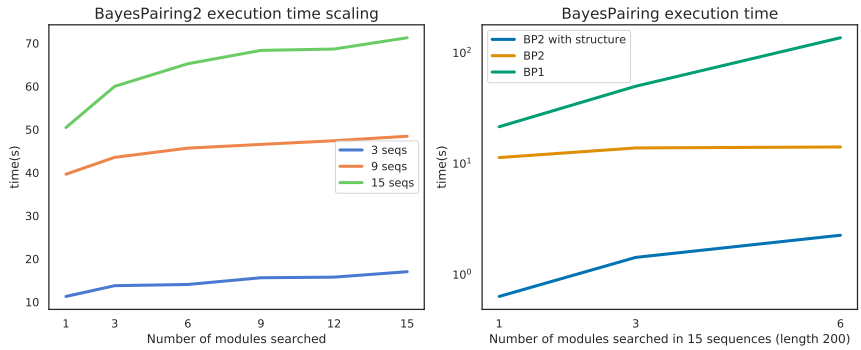


Fig. 4: Execution time of `BayesPairing 2`, as a function of numbers of modules and sequences (left), and compared to `BayesPairing 1` (right)

3.4 Time benchmark

The execution time of `BayesPairing 2` was measured on 15 sequences (average size of $\sim$ 200 nucleotides) containing a module each, with 5 hairpins, 5 internal loops and 5 multi-branched loops. We searched for 1, 3, 9 and 15 modules, and the execution time as a function of the sequence length and number of modules is displayed in Figure 4. While the software takes in the vicinity of two to three seconds to identify a module in a sequence of length 200, increasing the number of modules searched by a factor of fifteen only doubles its execution time.

Every test mentioned in this paper was run on Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz, Ubuntu 16.0.4 with 23 cores, with a total physical memory of 792 gigabytes.

4 Discussion

Comparison to the state of the art. The first software to tackle the specific task of identifying 3D motifs in full RNA sequences was `RMDetect` (2011) [8], which showed good accuracy but was severely limited in the variety of motifs it could identify. `BayesPairing 1` improved on this method by adding more flexibility and improving its search efficiency [32]. Another method, `JAR3D`, does not undertake full sequence searches but scores hairpin and internal loops against a database of models from the `RNA 3D Motif Atlas`. `BayesPairing 2` can be adapted to fulfill the same task, and their purposes are close enough to be comparable. Because `BayesPairing 1` has been shown to be a clear improvement on `RMDetect`, we focus our comparison on the former and `JAR3D`.

The good performances of `BayesPairing 1` [32] relies on the assumption that the structural motif searched has a strong sequence signal. Indeed, the tool identifies motif location candidates through regular expressions. Thus `BayesPairing 1` struggles with motifs trained on a large number of distinct sequences with no dominant sequence pattern.

While it performed well on structure-based datasets with high sequence conservation, our `Rfam`-based dataset, with an average of 268 sequences from multiple `Rfam` families for each module, appears challenging for the method and is clearly outperformed by `BayesPairing 2` on the dataset described in section 3.1, as shown in Table 3. We also show in Figure 4 that `BayesPairing 2` scales much better in the number of modules searched.

`JAR3D` was also shown to outperform `RMDetect` in the identification of new variants of RNA 3D modules [40]. However, it does not scan full sequences, and it is limited to the motifs of the `RNA 3D Motif Atlas`. In order to compare the software, we cross-referenced our motif database with the `RNA 3D Motif Atlas` and identified eight common motifs. This relatively low number is due to the differences in 3D structure annotations between `Rna3Dmotif` and the `RNA 3D Motif Atlas`.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	F1 score	MCC	FDR	Sensitivity (1 candidate)	Sensitivity (5 candidates)
BP1	0.715	0.510	0.178	0.219	0.348
BP2	**0.932**	**0.863**	**0.061**	**0.745**	**0.842**

Table 3: Performance of the two versions of BayesPairing on Rna3Dmotif dataset

Method		**Ratio positive hits**	
Software	# sequences	BP2	JAR3D
BayesPairing 2 dataset	160	**0.981**	0.525
RNA 3D Motif Atlas	148	0.458	**0.963**
All sequences - shuffled	308	**0.133**	0.394

Table 4: Performance of BayesPairing 2 and JAR3D on 308 sequences over 8 motifs

We then tested both software on all unique sequences in both datasets for each module, as well as on dinucleotide-shuffled versions of those sequences. The results are presented in table 4.

While those results are partial, they appear to show a tendency of JAR3D to correctly identify loop variants, but at the cost of admitting a fair number of false positives. Further work is necessary to fully compare the two software, but given they are not in direct competition as JAR3D could be used downstream of BayesPairing 2 as a refined scoring tool, we believe these results indicate there is room for our tool to contribute in the field of module scoring, as well as module identification.

Applications. The most obvious application for an efficient and parallelizable motif identification framework is to parse sequences for local 3D structure signal. Modular approaches for RNA 3D structure construction like RNA-MoIP [29] have been shown to successfully take advantage of local tertiary structure information. In particular, RNA-MoIP leverages 3D module matches to select the most stable secondary structures to use as a scaffold for the full structure. Indeed, secondary structures that can accommodate known 3D modules are often more predictive of the real structure than those who cannot [8]. To this day, RMDetect, BayesPairing 1 and BayesPairing 2 are the only known full sequence probabilistic module identification tools to be able to identify hairpins, internal loops and junctions, which are key components of many well-known structures, namely several riboswitches. Of the three, BayesPairing 2 is the most scalable. This scalability is essential as many datasets include hundreds of modules [27,30], and this number will keep increasing as more structures are crystallized and mining methods improve.

While the tertiary structure signal encodes information that can be leveraged to build a full 3D structure, its implied functional significance can be taken advantage to refine tasks like sequence classification. Traditional methods for sequence classification include k-mer based techniques [38], as well as sequence and structure motifs [39], but those only use the sequence and secondary structure signals. 3D modules are highly complementary to those methods.

Identifying multi-branched loops in sequences; applications to riboswitch discovery. One of the distinctive characteristics of BayesPairing 2 is its ability to identify multi-branched loops. These motifs happen to be very common in riboswitches, in which they are often closely related to function, namely in the tyrosine pyrophosphate (TPP) riboswitch, the Cobalamin riboswitch, and the S-adenosyl methionine I (SAM-I) riboswitch [33]. We can use sequences from

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

`Rfam` riboswitch families to train 3D module models, and then use those models to label new sequences as putative riboswitches.

The software also provides insight on the role of those of 3D modules in the folding dynamics of the riboswitch. Because `BayesPairing 2` searches secondary structure ensembles for loops matching known structural modules, it can be used to observe, within the assumptions of the `RNAfold` library, how easily riboswitch sequences appear to fold into their junction. For instance, the TPP riboswitch's junction is very present in its Boltzmann ensemble, as its small (13 bases) three-way junction was correcetly identified by our software on 81% of the sequences from the TPP `Rfam` family.

Because we could hypothesize the frequency of identification of a specific loop to be correlated with its size, it could be expected that the SAM-I riboswitch four-way junction, which counts 28 bases, would be identified less frequently. This is indeed the case as it was identified on 35% of the sequences of its family with a similar pipeline.

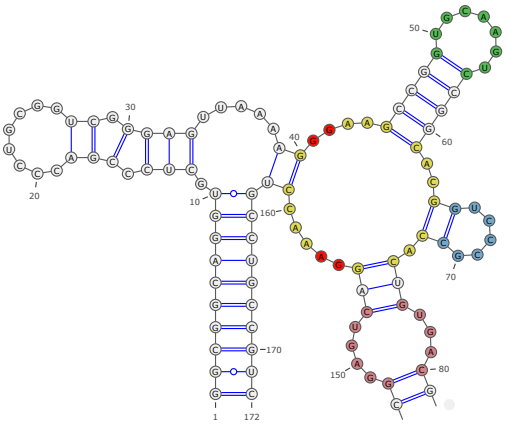


Fig. 5: The cobalamin riboswitch four-way junction (in yellow) in PDB structure 4GXY [26]. The adjacent structural motifs used to refine the structural search are highlighted. Bases within 3 Angstrom of the cobalamin molecule in the bound structure are indicated in bright red.

The much smaller (17 bases) cobalamin riboswitch junction would then be expected to be found with a frequency somewhere in between 35% and 81%, based on this size assumption. Surprisingly, it was only successfully identified on 3.5% of the `Rfam` cobalamin family sequences.

However, interestingly, identifying small structural modules (two hairpins and one internal loop) around the junction with a first run of `BayesPairing 2` and then using the position of those modules as constraints for a second run raises the frequency of identification of the multi-loop to 32%. The more adjacent motifs are found, the higher the identification confidence was observed to be. In contrast, applying the same method to the SAM riboswitch, or on shuffled cobalamin riboswitch sequences, does not leave to a significant improvement.

This difference in behavior between riboswitches could be rooted in different factors like co-transcriptional folding, RNA-RNA and RNA-protein interactions and/or the intrinsic difficulty of predicting riboswitch structural element with models learned from bound structures. However, the contrast between the constrained and unconstrained results in the cobalamin riboswitch tends to indicate that some, but not all multi-branched structure are strongly correlated with surrounding loops conformations.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Limitations and Future Work Our approaches presents two main limitations. First, the assumption that motif occurrences have a consistent size is not a trivial one to make. For small modules, it is a reasonable assumption that the vast majority of occurrences will have the same size since adding or removing a base would have a large impact on the local 3D structure. However, for larger motifs, and especially junctions, the size constraint can prevent us from identifying some variants. This is something we alleviate in `BayesPairing 2` by allowing imperfect matches, with a tolerated difference of up to one base per strand, but further work remains to be done to fully identify motifs bigger than 20 bases, for which this fuzzy matching might not be sufficient.

Second, a consequence of searching secondary structures before sequence is that in the rare cases when the sequence is better conserved than its secondary structure, the accuracy of the tool will suffer. It could however be argued that not overfitting to currently known sequences could be worth losing a bit of accuracy, although this can only be evaluated quantitatively as new structures and module occurrences become available, since the current structure datasets do not show sufficient sequence variability.

Interestingly, a large majority of the modules that cannot be predicted from sequence only by `BayesPairing 2` occur in secondary structures that are never generated by `RNAsubopt`. In many of those cases, a base pair stacking was removed to allow the insertion of the module, at a considerable energy cost. We hypothesize that those small modifications, although not energetically favorable at the secondary structure level, are stabilized by 3D interactions which cannot be inferred from sequence. Going further with this hypothesis, differences in performances are then indicative of the stabilizing effect of non-canonical modules. This assumption could be tested in the future using coarse-grain molecular dynamics to correlate those two metrics.

The other notable limitation of the method is that the loop-based module definition used in our study does not allow the prediction of pseudoknots, nor canonical helices.

5 Conclusion

We presented `BayesPairing 2`, a software for efficient identification of RNA modules in sequences and alignments. `BayesPairing 2` strictly outperforms its previous version in execution time, search on provided secondary structures, and sequence search accuracy. It also appears to have complementary strengths to `JAR3D`, the state of the art for scoring. Finally, its structure-based approach brings a perspective on the place of the motif in the sequence's Boltzmann ensemble. This added context helps improve identification accuracy, but also the interpretation of the results, and can provide additional information about the role of a module in the folding process. Moreover, the time complexity improvement opens new doors for genome-wide sequence mining for local 3D structure patterns. As new RNA structures and sequences become available, more modules will be discovered, and `BayesPairing 2` is fast enough to take advantage of its customizability to contribute to filling the gap between secondary and tertiary structure prediction tool by associating a wide selection of RNA modules of interest to those new sequences.

Acknowledgements

The authors are greatly indebted to Anton Petrov for providing us with alignments between RNA PDB structures and `Rfam` families, which helped us match 3D modules to sequence alignments.

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Ankan, A., Panda, A.: pgmpy: Probabilistic graphical models using python. In: Proceedings of the 14th Python in Science Conference (SCIPY 2015). Citeseer (2015)
2. Argaman, L., Altuvia, S.: fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. J. Mol. Biol. **300**(5), 1101–1112 (Jul 2000)
3. Bach, F.R., Jordan, M.I.: Thin junction trees. In: Advances in Neural Information Processing Systems. pp. 569–576 (2002)
4. Beelen, R.H., Fluitsma, D.M., van der Meer, J.W., Hoefsmit, E.C.: Development of different peroxidatic activity patterns in pertoneal macrophages in vivo and in vitro. J Reticuloendothel Soc **25**(5), 513–523 (May 1979)
5. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**, 235–242 (Jan 2000). https://doi.org/10.1093/nar/28.1.235
6. Bodlaender, H.L.: Dynamic programming on graphs with bounded treewidth. In: International Colloquium on Automata, Languages, and Programming. pp. 105–118. Springer (1988)
7. Chojnowski, G., Walen, T., Bujnicki, J.M.: Rna bricks–a database of rna 3d motifs and their interactions. Nucleic Acids Res **42**(Database issue), D123–31 (Jan 2014). https://doi.org/10.1093/nar/gkt1084
8. Cruz, J.A., Westhof, E.: Sequence-based identification of 3D structural modules in RNA with RMDetect. Nat Methods **8**(6), 513–21 (Jun 2011). https://doi.org/10.1038/nmeth.1603
9. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for rna secondary structure prediction. Nucleic acids research **31**, 7280–7301 (Dec 2003). https://doi.org/10.1093/nar/gkg938
10. Djelloul, M., Denise, A.: Automated motif extraction and classification in rna tertiary structures. RNA **14**(12), 2489–97 (Dec 2008). https://doi.org/10.1261/rna.1061108
11. Du, Z., Lind, K.E., James, T.L.: Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. Chem. Biol. **9**(6), 707–712 (Jun 2002)
12. Ge, P., Islam, S., Zhong, C., Zhang, S.: De novo discovery of structural motifs in RNA 3D structures through clustering. Nucleic Acids Research **46**(9), 4783–4793 (03 2018). https://doi.org/10.1093/nar/gky139
13. Huck, L., Scherrer, A., Terzi, L., Johnson, A.E., Bernstein, H.D., Cusack, S., Weichenrieder, O., Strub, K.: Conserved tertiary base pairing ensures proper RNA folding and efficient assembly of the signal recognition particle Alu domain. Nucleic Acids Res. **32**(16), 4915–4924 (2004)
14. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I.: Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research **46**(D1), D335–D342 (11 2017). https://doi.org/10.1093/nar/gkx1038
15. Lancaster, L., Lambert, N.J., Maklan, E.J., Horan, L.H., Noller, H.F.: The sarcin-ricin loop of 23S rRNA is essential for assembly of the functional core of the 50S ribosomal subunit. RNA **14**(10), 1999–2012 (Oct 2008)
16. Leontis, N.B., Westhof, E.: Geometric nomenclature and classification of RNA base pairs. RNA **7**(4), 499–512 (Apr 2001)
17. Leontis, N.B., Westhof, E.: Geometric nomenclature and classification of rna base pairs. RNA (New York, N.Y.) **7**, 499–512 (Apr 2001). https://doi.org/10.1017/s1355838201002515
18. Leontis, N.B., Zirbel, C.L.: Nonredundant 3d structure datasets for rna knowledge extraction and benchmarking. In: RNA 3D structure analysis and prediction, pp. 281–298. Springer (2012)
19. Lescoute, A., Leontis, N.B., Massire, C., Westhof, E.: Recurrent structural rna motifs, isostericity matrices and sequence alignments. Nucleic acids research **33**, 2395–2409 (2005). https://doi.org/10.1093/nar/gki535
20. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA package 2.0. Algorithms Mol Biol **6**, 26 (Nov 2011). https://doi.org/10.1186/1748-7188-6-26
21. mabseher: A small but efficient c++ library for computing (customized) tree and hypertree decompositions., https://github.com/mabseher/htd
22. Mathews, D.H.: RNA secondary structure analysis using RNAstructure. Curr Protoc Bioinformatics **Chapter 12**, Unit 12.6 (Mar 2006). https://doi.org/10.1002/0471250953.bi1206s13
23. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for rna secondary structure. Biopolymers **29**, 1105–1119 (1990). https://doi.org/10.1002/bip.360290621
24. Michálik, J., Touzet, H., Ponty, Y.: Efficient approximations of RNA kinetics landscape using non-redundant sampling. Bioinformatics (Oxford, England) **33**, i283–i292 (Jul 2017). https://doi.org/10.1093/bioinformatics/btx269
25. Mustoe, A.M., Brooks, C.L., Al-Hashimi, H.M.: Hierarchy of RNA functional dynamics. Annu. Rev. Biochem. **83**, 441–466 (2014)
26. Peselis, A., Serganov, A.: Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. Nature structural & molecular biology **19**(11), 1182 (2012)
27. Petrov, A.I., Zirbel, C.L., Leontis, N.B.: Automated classification of rna 3d motifs and the rna 3d motif atlas. RNA **19**(10), 1327–40 (Oct 2013). https://doi.org/10.1261/rna.039438.113

bioRxiv preprint doi: https://doi.org/10.1101/834762; this version posted November 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E.K., Blazewicz, J., Adamiak, R.W.: RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics **11**, 231 (May 2010). https://doi.org/10.1186/1471-2105-11-231
29. Reinharz, V., Major, F., Waldispühl, J.: Towards 3D structure prediction of large rna molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. Bioinformatics **28**(12), i207–14 (Jun 2012). https://doi.org/10.1093/bioinformatics/bts226
30. Reinharz, V., Soulé, A., Westhof, E., Waldispühl, J., Denise, A.: Mining for recurrent long-range interactions in rna structures reveals embedded hierarchies in network families. Nucleic Acids Research **46**(8), 3841–3851 (2018)
31. Rovetta, C., Michálik, J., Lorenz, R., Tanzer, A., Ponty, Y.: Non-redundant sampling and statistical estimators for RNA structural properties at the thermodynamic equilibrium (2019), under review. Preprint available at https://hal.inria.fr/hal-02288811
32. Sarrazin-Gendron, R., Reinharz, V., Oliver, C.G., Moitessier, N., Waldispühl, J.: Automated, customizable and efficient identification of 3d base pair modules with bayespairing. Nucleic acids research (2019)
33. Serganov, A., Nudler, E.: A decade of riboswitches. Cell **152**(1-2), 17–24 (2013)
34. Theis, C., Zirbel, C.L., Zu Siederdissen, C.H., Anthon, C., Hofacker, I.L., Nielsen, H., Gorodkin, J.: RNA 3D modules in genome-wide predictions of RNA 2D structure. PLoS One **10**(10), e0139900 (2015). https://doi.org/10.1371/journal.pone.0139900
35. Thiel, B.C., Ochsenreiter, R., Gadekar, V.P., Tanzer, A., Hofacker, I.L.: RNA Structure Elements Conserved between Mouse and 59 Other Vertebrates. Genes (Basel) **9**(8) (Aug 2018)
36. Tinoco, I., Bustamante, C.: How RNA folds. J Mol Biol **293**(2), 271–81 (Oct 1999). https://doi.org/10.1006/jmbi.1999.3001
37. Turner, D.H., Mathews, D.H.: Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic acids research **38**, D280–D282 (Jan 2010). https://doi.org/10.1093/nar/gkp892
38. Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology **15**(3), R46 (2014)
39. Xue, C., Li, F., He, T., Liu, G.P., Li, Y., Zhang, X.: Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine. BMC bioinformatics **6**(1), 310 (2005)
40. Zirbel, C.L., Roll, J., Sweeney, B.A., Petrov, A.I., Pirrung, M., Leontis, N.B.: Identifying novel sequence variants of RNA 3D motifs. Nucleic Acids Res **43**(15), 7504–20 (Sep 2015). https://doi.org/10.1093/nar/gkv651