The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures

Jean-Pierre Séhi Glouzon, Jean-Pierre Perreault and Shengrui Wang

Bioinformatics, 33(8), 2017, 1169–1178
doi: 10.1093/bioinformatics/btw773
Advance Access Publication Date: 14 January 2017
Original paper

Structural bioinformatics

The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures

Jean-Pierre Séhi Glouzon[1,2], Jean-Pierre Perreault[2] and Shengrui Wang[1,*]

[1]Department of Computer Science, Faculty of Science, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada and [2]RNA Group, Department of Biochemistry, Faculty of Medicine and Health Sciences, Applied Cancer Research Pavilion, Université de Sherbrooke, Sherbrooke, QC J1E 4K8, Canada

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on December 11, 2015; revised on September 16, 2016; editorial decision on December 1, 2016; accepted on Month 0, 0000

Abstract

Motivation: Comparing ribonucleic acid (RNA) secondary structures of arbitrary size uncovers structural patterns that can provide a better understanding of RNA functions. However, performing fast and accurate secondary structure comparisons is challenging when we take into account the RNA configuration (i.e. linear or circular), the presence of pseudoknot and G-quadruplex (G4) motifs and the increasing number of secondary structures generated by high-throughput probing techniques. To address this challenge, we propose the super-n-motifs model based on a latent analysis of enhanced motifs comprising not only basic motifs but also adjacency relations. The super-n-motifs model computes a vector representation of secondary structures as linear combinations of these motifs.

Results: We demonstrate the accuracy of our model for comparison of secondary structures from linear and circular RNA while also considering pseudoknot and G4 motifs. We show that the super-n-motifs representation effectively captures the most important structural features of secondary structures, as compared to other representations such as ordered tree, arc-annotated and string representations. Finally, we demonstrate the time efficiency of our model, which is alignment free and capable of performing large-scale comparisons of 10 000 secondary structures with an efficiency up to 4 orders of magnitude faster than existing approaches.

Availability and Implementation: The super-n-motifs model was implemented in C++. Source code and Linux binary are freely available at http://jpsglouzon.github.io/supernmotifs/.

Contact: Shengrui.Wang@Usherbrooke.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Exploring the relationships between ribonucleic acids (RNAs) by comparing their secondary structures provides critical insight into their functions. In fact, complex molecules such as RNAs can fold into secondary and tertiary structures to perform various functions involved in the regulation of translation, transcription, splicing, and so on (Wan *et al.*, 2011). However, because RNA tertiary structure is largely determined by its secondary structure (Brion and Westhof, 1997; Tinoco and Bustamante, 1999), RNAs with similar secondary structures will likely have the same or related functions. Thus, comparing RNA secondary structures can significantly contribute to understanding RNA functions.

In this paper, we consider three important aspects of secondary structure data in designing our model. We consider the nature of the

RNA (linear or circular), the presence of functional motifs such as pseudoknots and RNA G-quadruplexes (G4s) and finally, the growing number of secondary structures. First, while most RNAs are linear, recent studies suggest that circular RNA transcripts are abundant and have a potential role in gene regulation (Jeck *et al.*, 2013; Kosik, 2013). Many well-known pathogens such as viroids and the hepatitis delta virus have a circular RNA genome (Flores *et al.*, 2012). Second, both pseudoknot and G4 motifs are known to be involved in translation and splicing regulation (Millevoi *et al.*, 2012; Staple and Butcher, 2005). Pseudoknots are secondary structure topologies comprising additional base pairs between loops and are pervasive in many RNA families such as transfer messenger RNA (tmRNA), ribosomal RNA (rRNA), ribonuclease P RNA (RNase P), and so on. G4s, on the other hand, are formed by the stacking of non-canonical interactions of guanines; many such motifs have been found in the untranslated regions of mRNA (Huppert *et al.*, 2008). Finally, high-throughput methods for probing RNAs, such as FragSeq (Underwood *et al.*, 2010) and SHAPE-Seq (Loughrey et *al.*, 2014), yield a large number of secondary structures (Bellaousov *et al.*, 2013; Lorenz *et al.*, 2011).

Comparing secondary structures of arbitrary size from linear and circular RNAs while also considering pseudoknot and G4 motifs is a challenging task. Most of the algorithms for comparing secondary structures are not capable of handling circular RNAs or pseudoknots and G-quadruplexes because of their underlying representations of secondary structure. Existing algorithms for comparing secondary structures can be grouped into four categories according to their representations. The first group, based on an ordered tree representation of secondary structures, includes RNAdistance (Lorenz *et al.*, 2011), RNAforester (Schirmer and Giegerich, 2013), MiGal (Allali and Sagot, 2008) and RNAstrat (Guignon *et al.*, 2005). The second group, based on the string-encoded representation, includes RNAdistance and BEAR (Mattei *et al.*, 2014). The third group, based on the arc-annotated sequence representation, includes Gardenia (Blin *et al.*, 2010) and Efficient alignment of RNA secondary structures (ERA) (Zhong and Zhang, 2013). Finally, the fourth group, based on a representation combining sequence and structure ensemble information on RNA, called ensemble-based representation, includes LocARNA (Will *et al.*, 2007) and SPARSE (Will et al., 2015).

The majority of algorithms for comparing secondary structures do not handle secondary structures from circular RNA. This is because they are based on representations such as the ordered tree, string-encoded, arc-annotated sequence and ensemble-based approaches, which consider a secondary structure as an ordered set of nested base pairs beginning at the 5′ extremity and ending at the 3′ extremity of the RNA. While this is an appropriate way of looking at linear RNAs, it is not meaningful for circular RNAs because they do not have any 5′ and 3′ extremities. In fact, the 5′ and 3′ ends of circular RNA are joined, resulting in no directionality and, consequently, no intrinsic base pair ordering. Thus, these representations are not suitable to handle secondary structures from circular RNA. In this context, a meaningful alignment of secondary structures from circular RNA, whether global or local, can be hard to compute since alignment-based approaches inherit limitations from their respective structural representations. In fact, while it is possible to linearize circular RNA at a given position in order to compute the structural representation and perform global or local alignments, choosing the optimal starting position to produce the best alignment is not a trivial task. To address this challenge, specialized tools have been developed in the context of cyclic sequence alignment (Fernandes et al., 2009; Mosig *et al.*, 2006; Will and Stadler, 2014). However, there is no such approach specifically designed for aligning or comparing cyclic secondary structures.

Being based on secondary structure representations that consider only nested base pairs, the algorithms for comparing secondary structures cannot handle pseudoknots and G4 motifs, which in fact involve non-nested base interactions. It is important to note that an arc-annotated sequence representation can support comparison of secondary structures with non-nested base pairs but (so far) at a high computational cost (NP-hard; Schirmer *et al.*, 2014). Thus, algorithms based on ordered tree, string-encoded and arc-annotated sequence representations are not appropriate for processing secondary structures of circular RNA or for handling pseudoknots and G4 motifs. An extensive survey of secondary structure comparison algorithms and representations is given in the study by Schirmer et al. (2014).

Alignment-based algorithms for comparing secondary structures are effective, but they are computationally expensive, rendering them inappropriate for large-scale secondary structure comparisons. In general, alignment-based approaches, usually implemented using dynamic programming (Eddy, 2004), are known to be time-consuming, especially in the context of sequence analysis (Bonham-Carter *et al.*, 2013; Vinga, 2014). Advances have therefore been made toward alignment-free models in order to meet the need to process the thousands of sequences generated by high-throughput sequencing techniques (Haubold, 2014; Pinello *et al.*, 2014; Song *et al.*, 2014). Secondary structures are much more complex than sequences because nucleotides at the sequence level are involved in pairing. This renders secondary structure alignment computationally more intensive than sequence alignment. It is therefore necessary to develop alignment-free approaches to efficiently compute similarities between RNA secondary structures.

To address the need for an efficient way of comparing secondary structures from linear and circular RNA comprising pseudoknots and G4s, we propose a new model named *super-n-motifs*, based on the idea that similar secondary structures share similar combinations of motifs. Since secondary structures can be decomposed into building blocks, i.e. basic motifs such as stems or hairpin loops (Hendrix *et al.*, 2005), the secondary structures can be seen as being formed by multiple combinations of motifs. It is thus likely that secondary structures comprising shared or similar combinations of motifs are similar and belong to the same RNA family. As an example, a transfer RNA (tRNA) has a cloverleaf-shaped secondary structure formed by the combination of three hairpins and a stem (a hairpin being a combination of a stem and a loop). A secondary structure that possesses combinations of motifs similar to those in a tRNA secondary structure is likely a tRNA.

The super-n-motifs model takes as input given secondary structures and relies on three consecutive steps to build an effective and efficient vector-based representation of secondary structures. The proposed model first computes a bag-of-n-motifs model of secondary structures, where "i-motifs," for $0 \leq i \leq n$, are built from $(i-1)$-motifs and their neighbor relations from one level of abstraction to another, with 0-motifs being basic motifs. The value of n is the highest level of abstraction. Among the basic motifs considered are pseudoknots, G4s, single-stranded regions or external loops at the 5′ end and single-stranded regions or external loops at the 3′ end, and so on. The bag-of-n-motifs model explicitly handles pseudoknots and G4 motifs and the nature of the RNA (linear or circular). For circular RNAs, it simply ignores the external loop motifs at the 5′ and 3′ ends in the description of secondary structures from these RNAs. The n-motifs (Unless otherwise stated, we use the term "n-motifs" to represent an ensemble of i-motifs for all $0 \leq i \leq n$, and use i-motif or 2-motif to represent a motif at a particular level of abstraction, i or 2 here.) are thus designed to capture local and increasingly global structural features of secondary structures.

Second, the model computes the relative importance of each generated *i*-motif in order to select a reduced set of *i*-motifs, or features, for representing secondary structures. We call this the *n*-motifs representation. In fact, since each *i*-motif is computed from a single secondary structure, there can be a great number of *i*-motifs even for a small value of *n*. Not all the *i*-motifs are discriminative features for representing secondary structures. This step allows the user to choose the most discriminative ones based on an analysis of their frequencies of occurrence.

Third, the *n*-motifs representation obtained at the previous step is further transformed to obtain the super-n-motifs representation. This step makes use of singular value decomposition (SVD) to create feature variables as linear combinations of the *i*-motifs retained at Step 2. This latent representation makes it possible to capture some of the intrinsic similarity between secondary structures even if they do not share many *i*-motifs. Moreover, it also reduces the number of features in the final representation of secondary structures.

Finally, the vector representation of our model greatly facilitates comparisons between secondary structures. The super-n-motifs model is efficient because it is vector based and alignment free and effective because the super-n-motifs representation contains rich information about each secondary structure, including not only motifs and relations between motifs but also combinations of them, characterizing the secondary structure as a whole. The contributions of the super-n-motifs model are summarized in three major points are as follows:

- It explicitly captures structural information on RNA (linear or circular) since it relies on a description of secondary structures by n-motifs, which are hierarchically built, taking neighbor relations into account (see Section 2.1). The model is general and considers various basic motifs such as pseudoknots, G4s and single-stranded regions at both the 5′ and 3′ ends, in addition to many other common motifs.
- It allows an effective comparison of secondary structures because it computes the similarity of secondary structures based on their most informative structural features, which are the best *n*-motif combinations, i.e. the super-n-motifs (see Sections 2.3 and 3.1).
- It yields fast comparisons of secondary structures because it relies on an alignment-free approach that computes secondary structure similarities based on vectors in a low-dimensional space (see Section 3.2).

2 Materials and Methods

In this section, we describe in details the three main steps of the super-n-motifs model as they are outlined in the previous section, and we present the comparison metric and a complexity analysis of the model.

2.1 Bag-of-n-motifs model

The bag-of-n-motifs model yields a description of an RNA secondary structure in terms of multiple motifs built at different levels of abstraction. The description is designed to capture local and increasingly global structural features. From a secondary structure, it extracts basic motifs and their properties: for instance, an internal loop motif and its property, which is its symmetry or asymmetry; or a stem motif and its property, corresponding to its number of base pairs (Supplementary Figure S1 illustrates the motifs and properties for an arbitrary secondary structure). Motif properties yield specific structural information about the nature or size of the motifs and do not consider specific bases. After extracting basic motifs (also called 0-motifs) and their properties, the model computes 1-motifs by considering the neighbor relations of the 0-motifs. Similarly, it builds 2-motifs by considering the neighbor relations of the 1-motifs. The bag-of-n-motifs yields a set of structural features of a secondary structure that is the union of 0-motifs, 1-motifs, 2-motifs, ... and *n*-motifs. Figure 1 presents a simple example of the bag-of-n-motifs model computed from a secondary structure composed of a stem of five base pairs and a hairpin loop with two single-stranded nucleotides. A more complex example of the bag-of-n-motifs model, derived from a secondary structure of a circular RNA comprising motifs such as pseudoknots, G-quadruplexes, multiloops, stems, internal loops and hairpin loops, is illustrated in Supplementary Figure S2.

To generate the structural description of a secondary structure, the bag-of-n-motifs model builds a series of $n+1$ undirected graphs denoted by $\Theta = (G_0, G_1, \ldots, G_i, \ \ldots, G_n)$, where G_0 is a graph built from basic motifs and their neighbor relations and every other G_i is created from G_{i-1} by agglomerating nodes in G_{i-1} and by extending neighbor relations. A node of G_i is called an *i*-motif. The union of all the sets of *i*-motifs constitutes the bag-of-n-motifs, i.e. bag-of-n-motifs $= \cup_{i=0}^{n}$ node_set_of_G_i. As an example, in Figure 1, *S* and *H* are 0-motifs in G_0, while $S[H]$ is a 1-motif in G_1. $S[H]$ is created as an agglomerated motif around *S* with a surrounding *H*, the closest motif to *S* in G_0. Each motif is also associated with a description of its properties in a dotted square: for instance, *S* with *S_5* in G_0 and $S[H]$ with $S\_5[H\_2]$ in G_1. We can formally define the graph G_0 and each G_i, as in the following.

Θ is initialized by G_0, which corresponds to the graph of basic motifs with their properties. It is an undirected labeled graph defined as $G_0 = (V_0, E_0, P_0)$, where V_0 is the set of vertices corresponding to basic motifs. Let's define $V_0 = \left\{v_0^1, \ldots, v_0^j, \ldots, v_0^J\right\}$, where J is the total number of basic motifs. E_0 is the set of edges, indicating adjacency of motifs. Two motifs represented by $v_0^j \in V_0$ and $v_0^k \in V_0$ are considered adjacent, i.e. $\left(v_0^j,\ v_0^k\right) \in E_0$, if they share at least one nucleotide. P_0 is a set of J phantom nodes, each of which is attached to a vertex in V_0 to describe the property of the associated motif. Other elements in the construction of G_0 include:

- A set of node labels Ω_M initialized by $\{H, S, I, M, B, E5, E3, P, G4\}$ corresponding to basic motifs. Ω_M is enriched subsequently in the following steps. *H* stands for hairpin loop, *S* for stem, *I* for

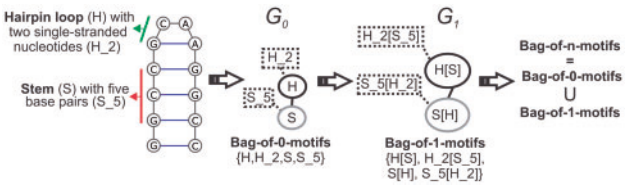

Fig. 1. Bag-of-n-motifs model of a secondary structure composed of a hairpin loop of two single-stranded nucleotides and a stem of five base pairs. It builds the list of motifs, i.e. the bag-of-n-motifs, from the graph of motifs (G_0) corresponding to a stem (*S*), a stem with 5 base pairs (*S_5*), a hairpin loop (*H*), and a hairpin loop with 2 single-stranded nucleotides (*H_2*). Then it computes the list (or bag) of 1-motifs associated with the graph of 1-motifs (G_1), by considering motifs with their neighbors. This yields a hairpin loop and stem motif ($H[S]$) where the neighbor of the hairpin loop is the stem, a two single-stranded nucleotides hairpin and stem of five base pairs ($H\_2[S\_5]$), a stem and hairpin loop motif ($S[H]$), and a five-base-pair stem and hairpin loop of two single-stranded nucleotides ($S\_5[H\_2]$). The bag-of-n-motifs model yields structural features of the secondary structure and comprises all the *n*-motifs: $\{S,\ H,\ S\_5,\ H\_2, S[H], S\_5[H\_2], H\_2[S\_5]\}$.

internal loop, M for multiloop or multi-branched loop, B for bulge loop, $E5$ for single-stranded region or external loop at $5'$ end, $E3$ for single-stranded region or external loop at $3'$ end, P for pseudoknot and $G4$ for G-quadruplex.

- A set of property labels Ω_P initialized by $\{(H|M|B|E5|E3)\_nb\text{-}of\text{-}single\text{-}stranded\text{-}nucleotides, I_{symmetry}, (S|P)\_nb\text{-}of\text{-}base\text{-}pairs, G4\_nb\text{-}of\text{-}quartets\}$. The property associated with motifs $H, M, B, E5$ and $E3$ is the number of single-stranded nucleotides. For I, it is the symmetry. The property for S and P is the number of base pairs. For $G4$, it is the number of G-quartets, i.e. the number of stacked tetrads composed of interacting guanines.
- Two functions f_0 and g_0 assigning to each vertex v_0^j a label from Ω_M and a label from Ω_P. In particular, $p_0^j = g_0\left(v_0^j\right)$ allows one to generate P_0, i.e. $P_0 \equiv g_0(V_0)$. For the model description, the use of the function f_0 is redundant. But it is a useful element in the software development of the model. For the example shown in Figure 1, the bag of 0-motifs from G_0 is $V_0 = \{S, H\}$ with the corresponding property set $P_0 = \{S\_5, H\_2\}$.

For any i, $0 < i \leq n$, G_i is built from G_{i-1}. Similar to G_0, G_i is represented as $G_i = (V_i, E_i, P_i)$, where $V_i = \left\{v_i^1, \ldots, v_i^j, \ldots, v_i^J\right\}$ represents the set of vertices that are the i-motifs. We need the following set function $N_{G_{i-1}}()$ to generate an i-motif: for an $(i-1)$-motif $v_{i-1}^j \in V_{i-1}$. $v_i^j = N_{G_{i-1}}\left(v_{i-1}^j\right) \equiv \left\{v_{i-1}^j[\cdots v_{i-1}^k \cdots], \left(v_{i-1}^j,\ v_{i-1}^k\right) \in E_{i-1}\right\}$. Here, the bracket notation, "[and]", represent the neighbor relations. From this definition of the set function $N_{G_{i-1}}()$, we can see that the jth i-motif is obtained as an agglomeration of the $(i-1)$-motifs around the jth $(i-1)$-motif. The edge set E_i of G_i is defined as follows: for two vertices $v_i^j \in V_i$ and $v_i^k \in V_i$, $\left(v_i^j,\ v_i^k\right) \in E_i$ if and only if $N_{G_{i-1}}(v_{i-1}^j) \cap N_{G_{i-1}}(v_{i-1}^k) \neq \varnothing$. In other words, there is an edge between v_i^j and v_i^k if they have at least one $(i-1)$-motifs in common. The property set P_i is defined similarly as $P_i = g_i(V_i)$, where the property function $g_i()$ is defined as follows: for any $v_i^j \in V_i$, $p_i^j = g_i\left(v_i^j\right) \equiv g_{i-1}\left(N_{G_{i-1}}\left(v_{i-1}^j\right)\right) = \{g_{i-1}(v_{i-1}^l), v_{i-1}^l \in N_{G_{i-1}}\left(v_{i-1}^j\right)\} = \{p_{i-1}^l | v_{i-1}^l \in N_{G_{i-1}}\left(v_{i-1}^j\right)\}$. The set of property labels Ω_P is then augmented by the new property labels generated at this level, i.e. $\Omega_P \equiv \Omega_P \cup P_i$. Similar operations are applied to create the node labeling function f_i and to update the node labels set Ω_M.

Therefore, for each secondary structure in our dataset, a set of graphs Θ is created and the union of all the computed bags of i-motifs results in a bag-of-n-motifs. All the bags-of-n-motifs from the secondary structures in a dataset provide a possibly very large ensemble of motifs. From this ensemble of motifs, we will select a set of most relevant motifs that will be used as features for representing the secondary structures from our dataset. To alleviate the text, from now on, the terms "motif," "i-motif" and "n-motif" will be used interchangeably, unless specified otherwise.

2.2 n-motifs representation of secondary structure

The n-motifs obtained in the bag-of-n-motifs model can be thought of as playing a similar role to that of n-grams in n-gram models. If we arrange all the n-motifs in some order, we can create a matrix representation of all the secondary structures. We introduce the matrix $S \in \mathbb{R}^{w*r}$, where w is the total number of secondary structures in the dataset, r is the number of unique n-motifs, and each element of the matrix, s_{ij}, is the frequency of occurrence of a unique n-motif j in the secondary structure i. We denote by s_i the ith row vector of S, corresponding to the ith secondary structure, and by z_j the jth column vector of S, corresponding to the jth n-motif. Not all the unique n-motifs convey the same amount of information. Therefore, it is desirable to select relevant n-motifs to get a refined vector representation of secondary structures, called the n-motifs representation.

Extracting relevant n-motifs from S leads to a better description of secondary structures. The method proposed here for selecting relevant n-motifs is based on an analysis of their frequencies of occurrence. We hypothesize that the relevance of an n-motif is proportional to its occurrence frequency. For instance, among important motifs of a tRNA, the single-stranded region at the $3'$ end is necessary for the binding of amino acids during the translation process. Therefore, we would expect a high occurrence of this motif in a population of secondary structures comprising tRNA. To automatically identify and remove irrelevant n-motifs, i.e. more rarely occurring n-motifs, we utilize the head/tail division rule (Jiang and Liu, 2011). This rule specifies that in the context of a heavy-tailed distribution, the arithmetic mean of the n-motif occurrences yields a natural division between the head (high-occurrence n-motifs) and the tail (low-occurrence n-motifs) of the heavy-tailed distribution.

The distribution of n-motifs ranked by decreasing order of total occurrence follows a heavy-tailed distribution (Supplementary Figure S2). We denote this distribution by f. With x the rank of an n-motif z_j and $f(x) = \sum_{i=1}^{w} s_{ij}$ its total occurrence, f satisfies the condition $\forall x, f(x) \geq f(x+1)$. $f(x) \geq 1$ since an n-motif is present in at least one secondary structure. f follows a heavy-tailed distribution in two regards. On the one hand, it satisfies the long-tailed distribution property: $\lim_{x \to +\infty} f(x+y)/f(x) = 1$, with $y > 0$ (Foss et al., 2011). In fact, low-ranked n-motifs tend to have an occurrence of one because they are present in at least one secondary structure. On the other hand, the long-tailed distribution is a subclass of the heavy-tailed distribution. Thus, we can apply the head/tail division rule to extract relevant n-motifs since f follows a heavy-tailed distribution. An n-motif z_j is considered relevant if the condition $f(x) \geq f_{\text{mean}}$ is satisfied, where $f_{\text{mean}} = \sum_{j=1}^{r} f(x_j) * 1/r$ with r the total number of n-motifs.

After removing irrelevant n-motifs, it is useful to reduce the effect of extreme n-motif occurrences arising from large secondary structures and increase the relative importance of medium n-motif occurrences specific to groups of structures. In fact, large secondary structures naturally tend to have many more n-motifs than small secondary structures. Moreover, while n-motifs specific to groups of structures bear valuable structural information about those groups, they are underestimated because the corresponding n-motifs tend to have medium to low occurrences. To alleviate these effects, we use the logarithm transformation such that $s'_{ij} = \log(s_{ij} + 1)$. s'_{ij} represents the relative relevance or importance of an n-motif z_j to a secondary structure s_i. The n-motifs representation of s_i is defined by the vector s'_i.

The n-motifs representation emphasizes relevant characteristics of secondary structures. It removes irrelevant (i.e. rare) n-motifs, reduces the impact of extremely high occurrences of n-motifs due to large secondary structures and increases the relative importance of n-motifs specific to groups of structures. Given that the n-motifs representation has, in fact, dependent features, in the next subsection, we propose to explore the relationships between n-motifs in order to find the best n-motif combinations characterizing secondary structures.

2.3 Super-n-motifs representation of secondary structure

From the previous two steps, we obtain individual n-motifs that capture geometric properties of RNA secondary structures. However,

some secondary structures sharing few *n*-motifs can belong to a same family. In this subsection, we develop a final vector representation of an RNA secondary structure by computing *n*-motif combinations to effectively represent the structure information. Such combinations should make it possible to capture latent statistical relationships between the *n*-motifs so as to create non-correlated features on which comparisons of secondary structures will be made. We search for the best linear combinations of *n*-motifs using SVD (Golub and Reinsch, 1970).

Let $S' \in \mathbb{R}^{w*m}$ be the matrix of *n*-motifs representations, where w is the total number of secondary structures and m is the total number of relevant *n*-motifs. S' is decomposed by SVD to a product of matrices, as follows:

$$S' = U\Sigma V^T,$$

where $U^T U = 1$ with $U \in \mathbb{R}^{w*w}$ and $V^T V = 1$ with $V \in \mathbb{R}^{m*m}$. Columns of U are the eigenvectors of $S'*S'^T$ and are linear combinations of *n*-motifs. Columns of V are the eigenvectors of S'^T*S' and are linear combinations of S' rows representing secondary structures. Σ is a $w * m$ diagonal matrix containing the square roots of the non-zero eigenvalues of $S'*S'^T$ or singular values of S' in decreasing order, i.e. $\sigma_{1,1} > \sigma_{2,2} > \ldots > \sigma_{r,r} > 0$. By selecting the first k largest singular values in Σ, we obtain a truncated matrix S'_k, where S'_k is defined by:

$$S' \approx S'_k = U_k \Sigma_k V_k^T.$$

S'_k yields the best low-rank approximation of S', such that $S'_k - S'$ (the Frobenius norm) is minimized.

The sum of the first k singular values in Σ_k represents the largest amount of information, in terms of variability, possibly retained using only k variables (features). We consider the matrices $U'_k = U_k \Sigma_k$, where U_k is weighted relative to Σ_k. The space defined by the k vectors of U'_k is called the super-n-motifs space. The super-n-motifs space is a low-dimensional space relative to the original space, since k is typically chosen such that $k \ll m$. The super-n-motifs representation of a structure s'_i is denoted by u'_i.

The super-n-motifs gather particular information concerning the relevant *n*-motifs and their relationships in order to provide a global picture of the structural features of the RNA. The super-n-motifs representation yields a vector representation of a secondary structure in a low-dimensional space. By comparing these representations, secondary structure relationships can be explored. The next subsection presents the comparison of secondary structures in the super-n-motifs representation.

2.4 Comparison of secondary structures based on the super-n-motifs representation

The exploration of secondary structure relationships is facilitated using the super-n-motifs representation. Secondary structures can be effectively compared by computing the cosine dissimilarity (Bonham-Carter *et al.*, 2013; Vinga and Almeida, 2003) between their super-n-motifs representations. Given two secondary structures, s_i and s_j, and their respective super-n-motifs representations, u'_i and u'_j, the cosine dissimilarity between u'_i and u'_j is defined as:

$$d_{\cos}\left(u'_i, u'_j\right) = 1 - \left(u'_i \cdot u'_j\right) / \left(u'_i u'_j\right)$$

s_i and s_j, are considered similar or close to one another when $d_{\cos}\left(u'_i, u'_j\right) \approx 0$ and dissimilar or far from one another when $d_{\cos}\left(u'_i, u'_j\right) \approx 2$.

2.5 Complexity of the super-n-motifs model

The time complexity of the super-n-motifs model is $O(w * m * \min(w, m))$, where w is the number of secondary structures and m is the number of relevant *n*-motifs. It corresponds to the SVD time complexity, since our approach is dominated by the SVD computation. The overall time complexity of the super-n-motifs model is $O(w) + O(r) + O(w*m*\min(w, m))$ where:

- $O(w)$ is associated with the computation of the bag-of-n-motifs model on the w structures;
- $O(r)$ represents the computation required for the selection of the m relevant *n*-motifs out of the total number of *n*-motifs denoted by r;
- $O(w*m*\min(w, m))$ represents the time complexity of SVD (Golub and Van Loan, 1996).

The space complexity of the super-n-motifs model is due largely to the space required to store the matrix of raw *n*-motifs, which is on the order of $O(w*r)$. In detail, the space complexity of the super-n-motifs model is $O(w*r) + O(w*m) + O(w*w) + O(w * m) + O(m * m)$, where:

- $O(w*r)$ is associated with the space required to store the matrix of *n*-motifs, S.
- $O(w*m)$ represents the storage space associated with the matrix of relevant *n*-motifs, S''.
- $O(w*w)$, $O(w * m)$ and $O(m * m)$ are, respectively, the space taken by the matrices U, Σ and V associated with SVD.

Typically, $r > w$, so the space complexity is $O(w*r)$, corresponding to the space necessary to store the matrix associated with space S.

3 Results

We evaluated the model's capacity to perform accurate, efficient comparisons between secondary structures of various sizes from linear and circular RNA comprising pseudoknots and G4s. The accuracy of RNA secondary structure comparison is expressed in terms of discriminative power, i.e. the ability to bring secondary structures from the same family closer together and push secondary structures from different families farther apart. The efficiency is evaluated in terms of the time required to compare sets comprising increasing numbers of secondary structures.

We used the normalized mutual information (NMI) and the f-measure (FMEAS; Manning *et al.*, 2008) to compute the discriminative power of secondary structure comparison algorithms, grouped by their underlying representations, on a medium-sized dataset of 2368 secondary structures from 12 RNA families (Fig. 2 and Supplementary Figure S4) and a large-sized dataset of 15 287 structures from 76 RNA families (Fig. 3). These datasets comprise, respectively, 9 and 66 linear RNA families and 3 and 10 circular RNA families (Andronescu *et al.*, 2008; Garant *et al.*, 2015; Giguère *et al.*, 2014; Nawrocki *et al.*, 2015). Supplementary Section 1 provides a detailed description of these datasets. When the discriminative power in terms of NMI or FMEAS is close to one, it means that all the members of each secondary structure family are close to one another and far from other secondary structures. Details on the computation of NMI and FMEAS are provided in Supplementary Section 2.

For the evaluation on the medium-sized dataset, the approaches and representations carefully chosen from the survey (Schirmer *et al.*, 2014) are as follows: the super-n-motifs model based on 5-gram

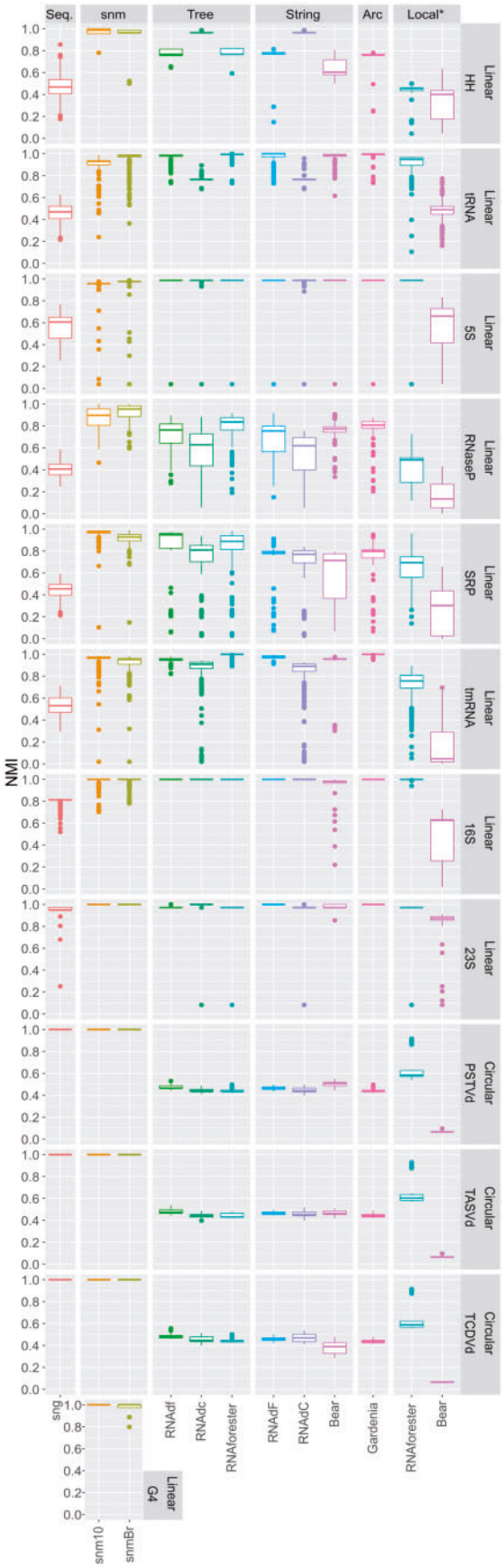

Fig. 2. Algorithms' discriminative power in terms of NMI on RNA secondary structure families. NMI distribution close to one means high discriminative power. Algorithms were grouped according to their representations: sequence (Seq.) super-n-motifs (snm), ordered tree (Tree), string-encoded (String) and arc-annotated sequence (Arc). Local* refers to local alignment version of approaches such as RNAforester and Bear. Families of secondary structures from linear RNA are HH ribozymes, tRNA, 5S rRNA, RNaseP, signal recognition particle (SRP), tmRNA, 16S rRNA, 23S rRNA) and structures with *G*-quadruplexes motifs (G4). Families of secondary structures from circular RNA are potato spindle tuber viroid (PSTVd), tomato apical stunt viroid (TASVd) and tomato chlorotic dwarf viroid (TCDVd).

(sng), the super-n-motifs model based on the super-n-motifs representation with $k = 10$ super-n-motifs (snm10) and the number of super-n-motifs automatically determined using the broken stick model (snmBr; see Supplementary Section 3); RNAdistance with the full tree (RNAdf) and the coarse-grained tree (RNAdc), and RNAforester based on the rooted ordered tree (Tree); RNAdistance with full string (RNAdF) and coarse-grained string (RNAdC), and BEAR based on the string-encoded representation (String); and Gardenia based on the arc-annotated sequence (Arc). In addition to the previously described approaches versions of RNAforester and BEAR. Algorithm parameters used are described in the Supplementary Section 3. For the evaluation on the large-sized dataset, the models compared are sng, snmBr and RNAdF. This latter was chosen because it is one of the approaches achieving the best trade-off between efficiency and discriminative power.

We evaluated the efficiency of the secondary-structure comparison algorithms by computing the time required to perform all-against-all comparisons on sets of 56, 104, 506, 1007 and 15 287 secondary structures (Tables 1 and 2), corresponding, respectively, to 1540, 5356, 127 765, 506 521 and ≈117 million pairwise comparisons. Each of these sets comprises from small (≈36 nt) to large secondary structures (≈2900 nt) of linear RNA (see the sets of secondary structures in Supplementary File S1), corresponding to the size variability generally found in RNA secondary structures. To avoid redundancy, it is only necessary to compute $(w^2 - w)/2$ comparisons, where w is the total number of secondary structures in a dataset.

3.1 Analysis of discriminative power

3.1.1 Performance on linear RNA families with high structural variability

From the results shown in Figure 2 and Supplementary Figure S4, we can remark that the super-n-motifs model demonstrates comparable or superior discriminative power compared to other approaches on families with high structural variability such as RNaseP, SRP and HH. In fact, the super-n-motifs model performs better on the RNaseP family compared to other methods, showing its capability to identify substructures or local structures shared among the set of highly variable secondary structures. Indeed, RNaseP structures from our dataset come from three domains, Archea, Bacteria and Eukaryotes, and the structures corresponding to these domains are known to be highly variable while showing a central conserved core (Evans *et al.*, 2006). Specifically, RNaseP structures exhibit very diverse conformations, due to more structural rearrangement than extended stems or loops, as these structures do not differ much in size, i.e. ≈337.66 nt ± 26.25 SD (see Supplementary Table S1). The diversity of RNaseP structures can be visualized in Supplementary Figure S5, where the structures of ASE_00264, ASE_00099 and ASE_00346 belong, respectively, to the Eukarya, Bacteria and Archea domains.

The super-n-motifs model also shows comparable discriminative power on the SRP family compared to the ordered tree-based

approach, RNAdf, and superior discriminative power compared to all the other approaches. Given that SRP structures in our dataset are highly variable in terms of size (Rosenblad et al., 2009), this suggests that the super-n-motifs model is much less affected by the size variability of structures. In fact, the SD relative to the mean size of structures in the SRP family is very high, ≈266.30 nt ± 65.46 SD (Supplementary Table S1).

As for circular permuted structures, the results shown in Figure 2 and Supplementary Figure S4 suggest that the super-n-motifs model performs very well on the HH family, similar to a coarse-grained approach such as RNAdc and RNAdC. This shows its capacity to handle circular permuted structures. Structural variability in the HH family is due to the fact that our dataset comprises hammerhead (HH) structures of types I and III with a large variability in size, ≈61.71 nt ± 24.16 SD (Supplementary Table S1). It is known that the HH family is comprised of three groups, types I, II and III, which are circularly permuted forms of HH distinguished by the open-ended helix that connects the motifs with the flanking sequences (Hammann et al., 2012). The structural variability of HH in our dataset can be seen by examining structures RFA_00660 and RFA_00407, representing HH of types I and III, in Supplementary Figure S5.

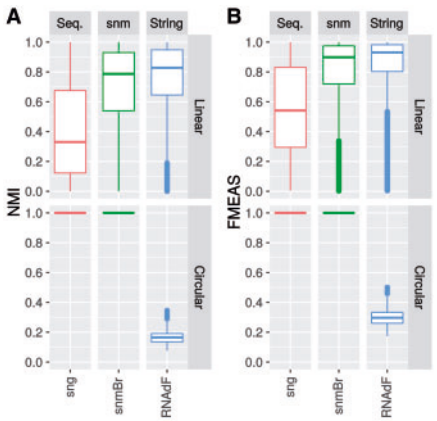

Fig. 3. Discriminative power of sng, snmBr, RNAdF and RNAdC in terms of NMI and FMEAS on RNA secondary structure families from a large dataset of 15 287 secondary structures belonging to 76 families. The discriminative power of each model is reported separately according to its performance on linear and circular RNA families.

3.1.2 Performance on linear RNA families with conserved structures

Performance comparison on families of conserved secondary structures shows that the super-n-motifs model displays high discriminative power similar to that of the majority of the tested approaches. Indeed, the super-n-motifs model, as well as RNAdf, RNAdF, RNAforester, BEAR and Gardenia, generate excellent NMI and FMEAS results on families with conserved structures such as the tRNA, 5S, tmRNA, 16S and 23S families (Fig. 2 and Supplementary Figure S4). The structural conservation of these families can be explained by the fact that not much size variability for any of these families is observed in our dataset. In fact, a small SD compared to the average size of structures is observed for the tRNA, 5S, tmRNA, 16S and 23S families. For each of these families, respectively, we have average sizes and SDs of ≈77 nt ± 5.42 SD, ≈120 nt ± 7.55 SD, ≈362.46 nt ± 6.72 SD, ≈1577.35nt ± 110.83 SD and ≈2908.09 nt ± 28.38 SD (Supplementary Table S1).

3.1.3 Performance on circular RNA families, *G*-quadruplexes and pseudoknots

The super-n-motifs model is designed to handle secondary structures from circular RNA families seamlessly. In fact, our model shows by

Table 2. Running times (in d: days and s: seconds) of sng, snmBr and RNAdF on the large-sized dataset of 15 287 secondary structures corresponding to ≈117 million pairwise comparisons.

Representation	Algorithm	Time complexity	Running times
Sequence	sng	$O(wm\ min(w,m))$	1697s
Super-n-motifs	snmBr	$O(wm\ min(w,m))$	**179s**
String	RNAdF	$O(l^2)$	41d[a]

Note: sng and snmBr were run on an 8 CPU (1.8 GHz) desktop pc with 5.8 GB of RAM and RNAdF was run in parallel on a 24 core node of the super computer Mammouth-mp2. w, m and l refer to the number of structures, the number of computed relevant n-motifs and the length of the strings representing the secondary structures. The best result is shown in bold.

[a]Estimated CPU time, that is, the sum of CPU time consumed by all of the CPUs.

Table 1. Running times (in d: days, m: minutes and s: seconds) on 56, 104, 506 and 1007 secondary structures, corresponding to 1540, 5356, 127 765 and 506 521 pairwise comparisons.

Representation	Algorithm	Time complexity for one pair. comp.	56 structures 1540 pair. comp.	104 structures 5356 pair. comp.	506 structures 127765 pair. comp.	1007 structures 506521 pair. comp.
Sequence	sng	$O(wm\ min(w,m))$	**0.12s**	**0.22s**	**3s**	35s
Super-n-motifs	snmBr	$O(wm\ min(w,m))$	**0.12 s**	0.66 s	**3 s**	**4 s**
Tree	RNAdf	$O(l^2d^2)$	4 m 29 s (10^3 f.)	16 m 24 s (10^3 f.)	5 h 44 m (10^3 f.)	14 h 7 m (10^4 f.)
	RNAdc	$O(l^2d^2)$	2 s (10^1 f.)	7 s (10^1 f.)	2 m17 s (10^1 f.)	5 m 4 s (10^1 f.)
	RNAforester	$O((l^2/r)q^2)$	1 h 31 m (10^4 f.)	10 h 5 m (10^5 f.)	>1d	>1d
String	RNAdF	$O(l^2)$	31 s (10^2 f.)	1 m 47 s (10^2 f.)	35 m 58 s (10^3 f.)	1 h 38 m (10^3 f.)
	RNAdC	$O(l^2)$	1 s	4 s	1 m 14 s (10^1 f.)	3 m 22 s (10^1 f.)
	BEAR	$O(l^2)$	2 m 31 s (10^3 f.)	8 m 34 s (10^2 f.)	3 h 17 m (10^3 f.)	7 h 54 m (10^3 f.)
Arc	Gardenia	$O(l^4)$	18 m 52 s (10^3 f.)	1 h 32 m (10^4 f.)	>1d	>1d
	ERA	$O(l^3)$	8 h 51 m (10^5 f.)	>1d	>1d	>1d
Ensemble	Sparse	$O(l^2)$	>1d	>1d	>1d	>1d

Note: The algorithms were run on an 8 CPU (1.8 GHz) desktop pc with 5.8 GB of RAM. f. and pair comp. stand for fold and pairwise comparisons. w and m refer to the number of structures and the number of computed relevant n-motifs. l represents the size of trees or forests, strings or arc-annotated sequences or number of base pairs representing secondary structures. d is the depth of trees. r and q are, respectively, the number of anchors and the maximum degree of forests. The best results in each column are shown in bold.

far the best discriminative power compared to all other methods, regardless of whether they are based on global or local alignment, on secondary structures from the 3 circular RNA families of the medium-sized dataset and the 10 circular families of the large-sized dataset (Fig. 2, Supplementary Figure S4 and Fig. 3). The clear advantage of our approach can be explained by the fact that, contrary to the other methods that assume contiguity of structural features, our model does not. Indeed, our model defines a secondary structure as an unordered collection of structural features, the *n*-motifs. Consequently, it captures structural features irrespective of their orders, which is an important property that allows it to effectively handle not only circular RNA but also RNA with the circularly permuted structures found in the HH family.

The results reported in Figure 2 and Supplementary Figure S4 demonstrate that the proposed approach has the capacity to handle pseudoknots and G4s. In fact, the super-n-motifs model demonstrates high discriminative power on secondary structures from linear RNA families comprising pseudoknots, such as RNaseP, tmRNA, 16S and 23S rRNA (see Supplementary Table S1). Moreover, our model effectively processes G4s, since it is capable of discriminating secondary structures that contain a single G4 from those without any G4s. It is important to note that while all the other approaches simply ignore these motifs, our model processes pseudoknots and G4 effectively, thanks to the flexibility of the bag-of-n-motifs that allows the explicit integration of any kind of motif (here, pseudoknots and G4 motifs) in the structural description of secondary structures. Our model makes no distinction between various types of pseudoknots, such as the H-type or the three-stemmed RNA pseudoknot (Staple and Butcher, 2005).

3.1.4 Large-scale performance analysis

A very important advantage of our approach is that it can be used to handle very large datasets in a very time-efficient way while maintaining high discriminative power as the number of structures grows. As shown in Figure 2, it achieves discriminative power comparable or superior to that of the tested approaches on linear RNA and consistently outperforms them on circular RNA in the medium-sized dataset of 2368 structures with 12 families. On the large-sized dataset of 15 287 structures with 76 families, we compared our approach with sng and RNAdF (Fig. 3 and Table 2). RNAdF was chosen because it is one of the most efficient approaches, combining good time complexity, $O(l^2)$, and high discriminative power rivaling that of approaches based on ordered-tree and arc-annotated sequences, which have a time complexity of at least $O(l^2d^2)$ (see Table 1 and Fig. 2). From Figure 3 and Table 2, it can be seen that our approach yields discriminative power comparable to that of RNAdF and superior to that of sng, yet significantly outperforms them in term of running times. This evaluation shows at least that our approach is indeed effective and efficient on a large dataset, i.e. 3 min as compared to the equivalent of 41 days for RNAdF. Currently, a thorough comparative study of different approaches on large datasets like the one used here is not possible since most competing approaches would need more than 1 month to produce their results on a single workstation.

3.1.5 Structural information at the coarse-grained and fine-grained levels

The results in Figure 2 and Supplementary Figure S4 show that the super-n-motifs model yields consistently high discriminative power compared to approaches using coarse-grained representation of structures (in which subsets of nucleotides or base pairs forming motifs such as stems or hairpin loops are considered as elements of the representation), for instance, RNAdc and RNAdC, on the HH family. It performs equally well as, and sometimes better than, approaches using fine-grained representation (in which each nucleotide or each base pair is considered as an element of the representation). These approaches include RNAdf, RNAdF, RNAforester, BEAR and Gardenia (Fig. 2 and Supplementary Figure S4). We observe that approaches based on fine-grained representation of structures perform better than approaches based on coarse-grained representation because the latter often results in a loss of structural information. Our approach combines the advantages of both fine-grained and coarse-grained representation and performs well on all the families. This can be explained by the fact that the structural features captured by our model represent coarse-grained features such as motifs and neighboring motifs, on the one hand, and fine-grained features such as the number of base pairs in a stem or the number of string-stranded nucleotides in loops on the other.

3.1.6 Sufficiency of structural information for separating families

We performed a comparison between structure-based and sequence-based methods to see whether structural information could be sufficient to separate families. We observed that structure-based methods, including snmBr, yield high discriminative power, while sequence-based methods like sng consistently have low discriminative power on most linear RNA families (Fig. 2, Supplementary Figure S4 and Fig. 3). In fact, for the medium-sized dataset families such as HH, tRNA, 5S, RNaseP, SRP and tmRNA and the large-sized dataset families, sng yields low performance, indicating that most RNAs, based on sequence information, have been assigned to the wrong family. These results show that, in our context, structural information is more important than sequence information for distinguishing RNA families. It is important to note that sng performs well on circular RNA, since sng, like snmBr, is insensitive to the RNA direction by the fact that it captures the unordered statistical patterns of sequences, the n-grams. Consequently, sng can be a good candidate to compare circular RNA sequences.

3.2 Efficiency assessment

The efficiency of the super-n-motifs model is shown by Tables 1 and 2. We observed that it is faster, by up to 4 or 5 orders of magnitude, than all the tested approaches, on data varying from very small-sized sets of 56 structures, corresponding to 1540 pairwise comparisons, to the large-sized set of 15 287 structures corresponding to ≈117 million pairwise comparisons. It exhibits a linear running time: it took 0.15 s, 0.27 s, 4 s, 7s, 12s and 182s, respectively, to compute all-against-all pairwise comparisons of 56, 104, 506, 1007, 2330 and 15 287 secondary structures of various sizes (from ≈36 to ≈2900 nt), corresponding to 1540, 5356, 127 765, 506 521, ≈2 million, ≈117 million pairwise comparisons (see Tables 1 and 2 and Supplementary Table S2). For the large-sized dataset of 15 287 structures, our approach is 4 orders of magnitude faster than RNAdF, since it took 182 s. (3 min) to compute ≈117 million comparisons from the set of 15 827 secondary structures, a task that would take ≈41 days of estimated CPU time for RNAdF (Table 2).

Tables 1 and 2 suggest that the alignment-based methods are convenient for low or medium-scale analysis but not suitable for large-scale analysis. RNAforester, Gardenia, ERA and SPARSE are more suitable for low-scale analysis, since even to compare 506 structures they already require days of computation. It is worth mentioning that ERA and SPARSE exhibit the highest running times: 8 h 51 m and more than a day to compare 56 structures

corresponding to 1540 comparisons. For this reason, we were unable to evaluate their respective discriminative power on the medium-sized datasets requiring computation of ≈2 million pairwise comparisons.

The results presented in the previous paragraphs are to be expected, because the majority of the tested approaches typically use DP as a foundation to compute alignment. Indeed, it is known that DP-based approaches are computationally demanding (Bonham-Carter *et al.*, 2013). Alignment-free approaches scale well and are of interest in sequence analysis to handle the vast numbers of transcripts generated by high-throughput sequencing methods (Bonham-Carter *et al.*, 2013; Vinga, 2014; Vinga and Almeida, 2003;). As the number of validated secondary structures grows, mainly due to high-throughput probing techniques, an alignment-free approach such as the super-n-motifs model offers a good alternative to the existing methods.

4 Discussion

Understanding RNA functions by comparing RNA secondary structures is challenging for several reasons. On the one hand, secondary structures are complex due to the nature of RNA, which can take a linear or a circular configuration and may contain pseudoknots and G4 motifs. On the other hand, thousands of secondary structures are generated by high-throughput probing techniques. In this paper, we proposed the super-n-motifs model and demonstrated its effectiveness and efficiency at comparing RNA secondary structures.

Our model computes accurate comparisons of secondary structures and naturally tends to cluster structures in a way that reflects known secondary structure families. This is an important property, because it is expected that members of a family of RNA secondary structures performing related functions form a natural cluster. This can be particularly helpful for RNA annotation, structure-based phylogeny, homology search in databases and identification of new families in populations of RNA. Since our model efficiently handles pseudoknots and G4 motifs, it can also help in understanding their functional roles.

In future work, our approach can be extended to compare RNA on the basis of sequences and secondary and tertiary structures. In fact, in addition to pseudoknots and G4 motifs, our model can be extended to handle other motifs such as sarcin-ricin, kink turn or c-loop motifs, thus making it possible to combine motifs from the secondary structures with the tertiary structures. Sequence information can be incorporated with n-grams as in sng method. Our model can thus yield a rich and global view of RNA by combining sequence and structural features.

Funding

This work was supported by a joint grant from the Fonds de Recherche du Québec—Nature et Technologies (FRQ-NT) and the Université de Sherbrooke Research Chair on RNA Structure and Genomics to JPP and the Natural Sciences and Engineering Research Council of Canada to SW. The Mammouth parallèle II supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the FRQ-NT.

Conflict of Interest: none declared.

References

Allali, J. and Sagot, M.F. (2008) A multiple layer model to compare RNA secondary structures. *Softw. Pract. Exp.*, **38**, 775–792.

Andronescu, M. *et al.* (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Bellaousov, S. *et al.* (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.*, **41**, W471–W474.

Blin, G. *et al.* (2010) Alignments of RNA structures. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **7**, 309–322.

Bonham-Carter, O. *et al.* (2013) Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinformatics*, **15**, 890–905.

Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 113–137.

Eddy, S.R. (2004) What is dynamic programming? *Nat. Biotechnol.*, **22**, 909–910.

Evans, D. *et al.* (2006) RNase P: interface of the RNA and protein worlds. *Trends Biochem. Sci.*, **31**, 333–341.

Fernandes, F. *et al.* (2009) CSA: an efficient algorithm to improve circular DNA multiple alignment. *BMC Bioinformatics*, **10**, 230.

Flores, R. *et al.* (2012) Viroids and hepatitis delta virus. *Semin. Liver Dis.*, **32**, 201–210.

Foss, S. *et al.* (2011) Heavy-tailed and long-tailed distributions. In: *An Introduction to Heavy-Tailed and Subexponential Distributions SE - 2*, Springer Series in Operations Research and Financial Engineering. Springer, New York, pp. 7–38.

Garant, J.M. *et al.* (2015) G4RNA: an RNA G-quadruplex database. *Database*, doi: 10.1093/database/bav059.

Giguère, T. *et al.* (2014) Comprehensive secondary structure elucidation of four genera of the family Pospiviroidae. *PLoS One*, **9**, e98655.

Golub, G.H. and Van Loan,C.F. (1996) Matrix computations. *Phys. Today*, **10**, 48.

Golub, G.H. and Reinsch, C. (1970) Singular value decomposition and least squares solutions. *Numer. Math.*, **14**, 403–420.

Guignon, V. *et al.* (2005) An edit distance between RNA stem-loops. In: Consens, M. and Navarro, G. (eds), *String Processing and Information Retrieval SE 38*, Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 335–347.

Hammann, C. *et al.* (2012) The ubiquitous hammerhead ribozyme. *RNA*, **18**, 871–885.

Haubold, B. (2014) Alignment-free phylogenetics and population genetics. *Brief. Bioinformatics*, **15**, 407–418.

Hendrix, D.K. *et al.* (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, **38**, 221–243.

Huppert, J.L. *et al.* (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.

Jeck, W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.

Jiang, B. and Liu, X. (2011) Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *Int. J. Geogr. Inf. Sci.*, **26**, 215–229.

Kosik, K.S. (2013) Circles reshape the RNA world. *Nature*, **495**, 4–6.

Lorenz, R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Loughrey, D. *et al.* (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res.*, **42**, e165–e165.

Manning, C.D. *et al.* (2008) Introduction to Information Retrieval. *J. Am. Soc. Inf. Sci. Technol.*, **1**, 496.

Mattei, E. *et al.* (2014) A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res.*, **42**, 6146–6157.

Millevoi, S. *et al.* (2012) G-quadruplexes in RNA biology. *Wiley Interdiscip. Rev. RNA*, **3**, 495–507.

Mosig, A. *et al.* (2006) Comparative Analysis of Cyclic Sequences: viroids and other Small Circular RNAs. In: *Lecture Notes in Informatics, German Conference on Bioinformatics,* **83**, pp. 93–102.

Nawrocki, E.P. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.

Pinello, L. *et al.* (2014) Applications of alignment-free methods in epigenomics. *Brief. Bioinformatics*, **15**, 419–430.

Rosenblad, M.A. *et al.* (2009) Kinship in the SRP RNA family. *RNA Biol.*, **6**, 508–516.

Schirmer, S. *et al*. (2014) Introduction to RNA secondary structure comparison. In: Gorodkin, J. and Ruzzo,W.L. (eds), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods SE 12*, Methods in Molecular Biology. Humana Press, New York, pp. 247–273.

Schirmer, S. and Giegerich, R. (2013) Forest alignment with affine gaps and anchors, applied in RNA structure comparison. In: *Theoretical Computer Science*, pp. 51–67.

Song, K. *et al*. (2014) New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinformatics*, **15**, 343–353.

Staple, D.W. and Butcher, S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*., **3**, 0956–0959.

Tinoco, I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol*., **293**, 271–281.

Underwood, J.G. *et al*. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.

Vinga, S. (2014) Editorial: alignment-free methods in computational biology. *Briefings Bioinformatics*, **15**, 341–342.

Vinga, S. and Almeida, J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

Wan, Y. *et al*. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet*., **12**, 641–655.

Will, S. *et al*. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol*., **3**, 680–691.

Will, S. *et al*. (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.

Will, S. and Stadler, P.F. (2014) Algorithms in Bioinformatics. In: Brown, D. and Morgenstern, B. (eds) *Proceedings of 14th International Workshop, WABI 2014*, Wroclaw, Poland, September 8–10, 2014, pp. 135–147 Springer, Berlin/Heidelberg.

Zhong, C. and Zhang, S. (2013) Efficient alignment of RNA secondary structures using sparse dynamic programming. *BMC Bioinformatics*, **14**, 269.

Bioinformatics, 33(8), 2017, 1169–1178
doi: 10.1093/bioinformatics/btw773
Advance Access Publication Date: 14 January 2017
Original paper

Structural bioinformatics

The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures

Jean-Pierre Séhi Glouzon[1,2], Jean-Pierre Perreault[2] and Shengrui Wang[1,*]

[1]Department of Computer Science, Faculty of Science, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada and [2]RNA Group, Department of Biochemistry, Faculty of Medicine and Health Sciences, Applied Cancer Research Pavilion, Université de Sherbrooke, Sherbrooke, QC J1E 4K8, Canada

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on December 11, 2015; revised on September 16, 2016; editorial decision on December 1, 2016; accepted on Month 0, 0000

Abstract

Motivation: Comparing ribonucleic acid (RNA) secondary structures of arbitrary size uncovers structural patterns that can provide a better understanding of RNA functions. However, performing fast and accurate secondary structure comparisons is challenging when we take into account the RNA configuration (i.e. linear or circular), the presence of pseudoknot and G-quadruplex (G4) motifs and the increasing number of secondary structures generated by high-throughput probing techniques. To address this challenge, we propose the super-n-motifs model based on a latent analysis of enhanced motifs comprising not only basic motifs but also adjacency relations. The super-n-motifs model computes a vector representation of secondary structures as linear combinations of these motifs.

Results: We demonstrate the accuracy of our model for comparison of secondary structures from linear and circular RNA while also considering pseudoknot and G4 motifs. We show that the super-n-motifs representation effectively captures the most important structural features of secondary structures, as compared to other representations such as ordered tree, arc-annotated and string representations. Finally, we demonstrate the time efficiency of our model, which is alignment free and capable of performing large-scale comparisons of 10 000 secondary structures with an efficiency up to 4 orders of magnitude faster than existing approaches.

Availability and Implementation: The super-n-motifs model was implemented in C++. Source code and Linux binary are freely available at http://jpsglouzon.github.io/supernmotifs/.

Contact: Shengrui.Wang@Usherbrooke.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Exploring the relationships between ribonucleic acids (RNAs) by comparing their secondary structures provides critical insight into their functions. In fact, complex molecules such as RNAs can fold into secondary and tertiary structures to perform various functions involved in the regulation of translation, transcription, splicing, and so on (Wan *et al.*, 2011). However, because RNA tertiary structure is largely determined by its secondary structure (Brion and Westhof, 1997; Tinoco and Bustamante, 1999), RNAs with similar secondary structures will likely have the same or related functions. Thus, comparing RNA secondary structures can significantly contribute to understanding RNA functions.

In this paper, we consider three important aspects of secondary structure data in designing our model. We consider the nature of the

RNA (linear or circular), the presence of functional motifs such as pseudoknots and RNA G-quadruplexes (G4s) and finally, the growing number of secondary structures. First, while most RNAs are linear, recent studies suggest that circular RNA transcripts are abundant and have a potential role in gene regulation (Jeck *et al.*, 2013; Kosik, 2013). Many well-known pathogens such as viroids and the hepatitis delta virus have a circular RNA genome (Flores *et al.*, 2012). Second, both pseudoknot and G4 motifs are known to be involved in translation and splicing regulation (Millevoi *et al.*, 2012; Staple and Butcher, 2005). Pseudoknots are secondary structure topologies comprising additional base pairs between loops and are pervasive in many RNA families such as transfer messenger RNA (tmRNA), ribosomal RNA (rRNA), ribonuclease P RNA (RNase P), and so on. G4s, on the other hand, are formed by the stacking of non-canonical interactions of guanines; many such motifs have been found in the untranslated regions of mRNA (Huppert *et al.*, 2008). Finally, high-throughput methods for probing RNAs, such as FragSeq (Underwood *et al.*, 2010) and SHAPE-Seq (Loughrey et *al.*, 2014), yield a large number of secondary structures (Bellaousov *et al.*, 2013; Lorenz *et al.*, 2011).

Comparing secondary structures of arbitrary size from linear and circular RNAs while also considering pseudoknot and G4 motifs is a challenging task. Most of the algorithms for comparing secondary structures are not capable of handling circular RNAs or pseudoknots and G-quadruplexes because of their underlying representations of secondary structure. Existing algorithms for comparing secondary structures can be grouped into four categories according to their representations. The first group, based on an ordered tree representation of secondary structures, includes RNAdistance (Lorenz *et al.*, 2011), RNAforester (Schirmer and Giegerich, 2013), MiGal (Allali and Sagot, 2008) and RNAstrat (Guignon *et al.*, 2005). The second group, based on the string-encoded representation, includes RNAdistance and BEAR (Mattei *et al.*, 2014). The third group, based on the arc-annotated sequence representation, includes Gardenia (Blin *et al.*, 2010) and Efficient alignment of RNA secondary structures (ERA) (Zhong and Zhang, 2013). Finally, the fourth group, based on a representation combining sequence and structure ensemble information on RNA, called ensemble-based representation, includes LocARNA (Will *et al.*, 2007) and SPARSE (Will et al., 2015).

The majority of algorithms for comparing secondary structures do not handle secondary structures from circular RNA. This is because they are based on representations such as the ordered tree, string-encoded, arc-annotated sequence and ensemble-based approaches, which consider a secondary structure as an ordered set of nested base pairs beginning at the 5′ extremity and ending at the 3′ extremity of the RNA. While this is an appropriate way of looking at linear RNAs, it is not meaningful for circular RNAs because they do not have any 5′ and 3′ extremities. In fact, the 5′ and 3′ ends of circular RNA are joined, resulting in no directionality and, consequently, no intrinsic base pair ordering. Thus, these representations are not suitable to handle secondary structures from circular RNA. In this context, a meaningful alignment of secondary structures from circular RNA, whether global or local, can be hard to compute since alignment-based approaches inherit limitations from their respective structural representations. In fact, while it is possible to linearize circular RNA at a given position in order to compute the structural representation and perform global or local alignments, choosing the optimal starting position to produce the best alignment is not a trivial task. To address this challenge, specialized tools have been developed in the context of cyclic sequence alignment (Fernandes et al., 2009; Mosig *et al.*, 2006; Will and Stadler, 2014). However, there is no such approach specifically designed for aligning or comparing cyclic secondary structures.

Being based on secondary structure representations that consider only nested base pairs, the algorithms for comparing secondary structures cannot handle pseudoknots and G4 motifs, which in fact involve non-nested base interactions. It is important to note that an arc-annotated sequence representation can support comparison of secondary structures with non-nested base pairs but (so far) at a high computational cost (NP-hard; Schirmer *et al.*, 2014). Thus, algorithms based on ordered tree, string-encoded and arc-annotated sequence representations are not appropriate for processing secondary structures of circular RNA or for handling pseudoknots and G4 motifs. An extensive survey of secondary structure comparison algorithms and representations is given in the study by Schirmer et al. (2014).

Alignment-based algorithms for comparing secondary structures are effective, but they are computationally expensive, rendering them inappropriate for large-scale secondary structure comparisons. In general, alignment-based approaches, usually implemented using dynamic programming (Eddy, 2004), are known to be time-consuming, especially in the context of sequence analysis (Bonham-Carter *et al.*, 2013; Vinga, 2014). Advances have therefore been made toward alignment-free models in order to meet the need to process the thousands of sequences generated by high-throughput sequencing techniques (Haubold, 2014; Pinello *et al.*, 2014; Song *et al.*, 2014). Secondary structures are much more complex than sequences because nucleotides at the sequence level are involved in pairing. This renders secondary structure alignment computationally more intensive than sequence alignment. It is therefore necessary to develop alignment-free approaches to efficiently compute similarities between RNA secondary structures.

To address the need for an efficient way of comparing secondary structures from linear and circular RNA comprising pseudoknots and G4s, we propose a new model named *super-n-motifs*, based on the idea that similar secondary structures share similar combinations of motifs. Since secondary structures can be decomposed into building blocks, i.e. basic motifs such as stems or hairpin loops (Hendrix *et al.*, 2005), the secondary structures can be seen as being formed by multiple combinations of motifs. It is thus likely that secondary structures comprising shared or similar combinations of motifs are similar and belong to the same RNA family. As an example, a transfer RNA (tRNA) has a cloverleaf-shaped secondary structure formed by the combination of three hairpins and a stem (a hairpin being a combination of a stem and a loop). A secondary structure that possesses combinations of motifs similar to those in a tRNA secondary structure is likely a tRNA.

The super-n-motifs model takes as input given secondary structures and relies on three consecutive steps to build an effective and efficient vector-based representation of secondary structures. The proposed model first computes a bag-of-n-motifs model of secondary structures, where "i-motifs," for $0 \leq i \leq n$, are built from $(i-1)$-motifs and their neighbor relations from one level of abstraction to another, with 0-motifs being basic motifs. The value of n is the highest level of abstraction. Among the basic motifs considered are pseudoknots, G4s, single-stranded regions or external loops at the 5′ end and single-stranded regions or external loops at the 3′ end, and so on. The bag-of-n-motifs model explicitly handles pseudoknots and G4 motifs and the nature of the RNA (linear or circular). For circular RNAs, it simply ignores the external loop motifs at the 5′ and 3′ ends in the description of secondary structures from these RNAs. The n-motifs (Unless otherwise stated, we use the term "n-motifs" to represent an ensemble of i-motifs for all $0 \leq i \leq n$, and use i-motif or 2-motif to represent a motif at a particular level of abstraction, i or 2 here.) are thus designed to capture local and increasingly global structural features of secondary structures.

Second, the model computes the relative importance of each generated *i*-motif in order to select a reduced set of *i*-motifs, or features, for representing secondary structures. We call this the *n*-motifs representation. In fact, since each *i*-motif is computed from a single secondary structure, there can be a great number of *i*-motifs even for a small value of *n*. Not all the *i*-motifs are discriminative features for representing secondary structures. This step allows the user to choose the most discriminative ones based on an analysis of their frequencies of occurrence.

Third, the *n*-motifs representation obtained at the previous step is further transformed to obtain the super-n-motifs representation. This step makes use of singular value decomposition (SVD) to create feature variables as linear combinations of the *i*-motifs retained at Step 2. This latent representation makes it possible to capture some of the intrinsic similarity between secondary structures even if they do not share many *i*-motifs. Moreover, it also reduces the number of features in the final representation of secondary structures.

Finally, the vector representation of our model greatly facilitates comparisons between secondary structures. The super-n-motifs model is efficient because it is vector based and alignment free and effective because the super-n-motifs representation contains rich information about each secondary structure, including not only motifs and relations between motifs but also combinations of them, characterizing the secondary structure as a whole. The contributions of the super-n-motifs model are summarized in three major points are as follows:

- It explicitly captures structural information on RNA (linear or circular) since it relies on a description of secondary structures by n-motifs, which are hierarchically built, taking neighbor relations into account (see Section 2.1). The model is general and considers various basic motifs such as pseudoknots, G4s and single-stranded regions at both the 5′ and 3′ ends, in addition to many other common motifs.
- It allows an effective comparison of secondary structures because it computes the similarity of secondary structures based on their most informative structural features, which are the best *n*-motif combinations, i.e. the super-n-motifs (see Sections 2.3 and 3.1).
- It yields fast comparisons of secondary structures because it relies on an alignment-free approach that computes secondary structure similarities based on vectors in a low-dimensional space (see Section 3.2).

2 Materials and Methods

In this section, we describe in details the three main steps of the super-n-motifs model as they are outlined in the previous section, and we present the comparison metric and a complexity analysis of the model.

2.1 Bag-of-n-motifs model

The bag-of-n-motifs model yields a description of an RNA secondary structure in terms of multiple motifs built at different levels of abstraction. The description is designed to capture local and increasingly global structural features. From a secondary structure, it extracts basic motifs and their properties: for instance, an internal loop motif and its property, which is its symmetry or asymmetry; or a stem motif and its property, corresponding to its number of base pairs (Supplementary Figure S1 illustrates the motifs and properties for an arbitrary secondary structure). Motif properties yield specific structural information about the nature or size of the motifs and do not consider specific bases. After extracting basic motifs (also called 0-motifs) and their properties, the model computes 1-motifs by considering the neighbor relations of the 0-motifs. Similarly, it builds 2-motifs by considering the neighbor relations of the 1-motifs. The bag-of-n-motifs yields a set of structural features of a secondary structure that is the union of 0-motifs, 1-motifs, 2-motifs, ... and *n*-motifs. Figure 1 presents a simple example of the bag-of-n-motifs model computed from a secondary structure composed of a stem of five base pairs and a hairpin loop with two single-stranded nucleotides. A more complex example of the bag-of-n-motifs model, derived from a secondary structure of a circular RNA comprising motifs such as pseudoknots, G-quadruplexes, multiloops, stems, internal loops and hairpin loops, is illustrated in Supplementary Figure S2.

To generate the structural description of a secondary structure, the bag-of-n-motifs model builds a series of $n+1$ undirected graphs denoted by $\Theta = (G_0, G_1, \ldots, G_i, \ \ldots, G_n)$, where G_0 is a graph built from basic motifs and their neighbor relations and every other G_i is created from G_{i-1} by agglomerating nodes in G_{i-1} and by extending neighbor relations. A node of G_i is called an *i*-motif. The union of all the sets of *i*-motifs constitutes the bag-of-n-motifs, i.e. bag-of-n-motifs $= \cup_{i=0}^{n}$ node_set_of_G_i. As an example, in Figure 1, *S* and *H* are 0-motifs in G_0, while $S[H]$ is a 1-motif in G_1. $S[H]$ is created as an agglomerated motif around *S* with a surrounding *H*, the closest motif to *S* in G_0. Each motif is also associated with a description of its properties in a dotted square: for instance, *S* with *S_5* in G_0 and $S[H]$ with $S\_5[H\_2]$ in G_1. We can formally define the graph G_0 and each G_i, as in the following.

Θ is initialized by G_0, which corresponds to the graph of basic motifs with their properties. It is an undirected labeled graph defined as $G_0 = (V_0, E_0, P_0)$, where V_0 is the set of vertices corresponding to basic motifs. Let's define $V_0 = \left\{v_0^1, \ldots, v_0^j, \ldots, v_0^J\right\}$, where J is the total number of basic motifs. E_0 is the set of edges, indicating adjacency of motifs. Two motifs represented by $v_0^j \in V_0$ and $v_0^k \in V_0$ are considered adjacent, i.e. $\left(v_0^j,\ v_0^k\right) \in E_0$, if they share at least one nucleotide. P_0 is a set of J phantom nodes, each of which is attached to a vertex in V_0 to describe the property of the associated motif. Other elements in the construction of G_0 include:

- A set of node labels Ω_M initialized by $\{H, S, I, M, B, E5, E3, P, G4\}$ corresponding to basic motifs. Ω_M is enriched subsequently in the following steps. *H* stands for hairpin loop, *S* for stem, *I* for

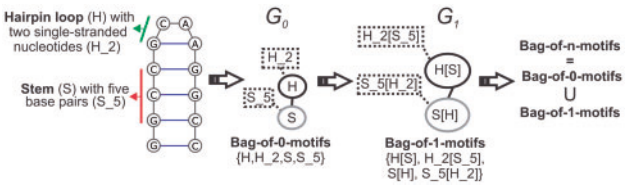

Fig. 1. Bag-of-n-motifs model of a secondary structure composed of a hairpin loop of two single-stranded nucleotides and a stem of five base pairs. It builds the list of motifs, i.e. the bag-of-n-motifs, from the graph of motifs (G_0) corresponding to a stem (*S*), a stem with 5 base pairs (*S_5*), a hairpin loop (*H*), and a hairpin loop with 2 single-stranded nucleotides (*H_2*). Then it computes the list (or bag) of 1-motifs associated with the graph of 1-motifs (G_1), by considering motifs with their neighbors. This yields a hairpin loop and stem motif ($H[S]$) where the neighbor of the hairpin loop is the stem, a two single-stranded nucleotides hairpin and stem of five base pairs ($H\_2[S\_5]$), a stem and hairpin loop motif ($S[H]$), and a five-base-pair stem and hairpin loop of two single-stranded nucleotides ($S\_5[H\_2]$). The bag-of-n-motifs model yields structural features of the secondary structure and comprises all the *n*-motifs: $\{S,\ H,\ S\_5,\ H\_2, S[H], S\_5[H\_2], H\_2[S\_5]\}$.

internal loop, M for multiloop or multi-branched loop, B for bulge loop, $E5$ for single-stranded region or external loop at $5'$ end, $E3$ for single-stranded region or external loop at $3'$ end, P for pseudoknot and $G4$ for G-quadruplex.

- A set of property labels Ω_P initialized by $\{(H|M|B|E5|E3)\_nb\text{-}of\text{-}single\text{-}stranded\text{-}nucleotides, I_{symmetry}, (S|P)\_nb\text{-}of\text{-}base\text{-}pairs, G4\_nb\text{-}of\text{-}quartets\}$. The property associated with motifs $H, M, B, E5$ and $E3$ is the number of single-stranded nucleotides. For I, it is the symmetry. The property for S and P is the number of base pairs. For $G4$, it is the number of G-quartets, i.e. the number of stacked tetrads composed of interacting guanines.
- Two functions f_0 and g_0 assigning to each vertex v_0^j a label from Ω_M and a label from Ω_P. In particular, $p_0^j = g_0\left(v_0^j\right)$ allows one to generate P_0, i.e. $P_0 \equiv g_0(V_0)$. For the model description, the use of the function f_0 is redundant. But it is a useful element in the software development of the model. For the example shown in Figure 1, the bag of 0-motifs from G_0 is $V_0 = \{S, H\}$ with the corresponding property set $P_0 = \{S\_5, H\_2\}$.

For any i, $0 < i \leq n$, G_i is built from G_{i-1}. Similar to G_0, G_i is represented as $G_i = (V_i, E_i, P_i)$, where $V_i = \left\{v_i^1, \ldots, v_i^j, \ldots, v_i^J\right\}$ represents the set of vertices that are the i-motifs. We need the following set function $N_{G_{i-1}}()$ to generate an i-motif: for an $(i-1)$-motif $v_{i-1}^j \in V_{i-1}$. $v_i^j = N_{G_{i-1}}\left(v_{i-1}^j\right) \equiv \left\{v_{i-1}^j[\cdots v_{i-1}^k \cdots], \left(v_{i-1}^j,\ v_{i-1}^k\right) \in E_{i-1}\right\}$. Here, the bracket notation, "[and]", represent the neighbor relations. From this definition of the set function $N_{G_{i-1}}()$, we can see that the jth i-motif is obtained as an agglomeration of the $(i-1)$-motifs around the jth $(i-1)$-motif. The edge set E_i of G_i is defined as follows: for two vertices $v_i^j \in V_i$ and $v_i^k \in V_i$, $\left(v_i^j,\ v_i^k\right) \in E_i$ if and only if $N_{G_{i-1}}(v_{i-1}^j) \cap N_{G_{i-1}}(v_{i-1}^k) \neq \varnothing$. In other words, there is an edge between v_i^j and v_i^k if they have at least one $(i-1)$-motifs in common. The property set P_i is defined similarly as $P_i = g_i(V_i)$, where the property function $g_i()$ is defined as follows: for any $v_i^j \in V_i$, $p_i^j = g_i\left(v_i^j\right) \equiv g_{i-1}\left(N_{G_{i-1}}\left(v_{i-1}^j\right)\right) = \{g_{i-1}(v_{i-1}^l), v_{i-1}^l \in N_{G_{i-1}}\left(v_{i-1}^j\right)\} = \{p_{i-1}^l | v_{i-1}^l \in N_{G_{i-1}}\left(v_{i-1}^j\right)\}$. The set of property labels Ω_P is then augmented by the new property labels generated at this level, i.e. $\Omega_P \equiv \Omega_P \cup P_i$. Similar operations are applied to create the node labeling function f_i and to update the node labels set Ω_M.

Therefore, for each secondary structure in our dataset, a set of graphs Θ is created and the union of all the computed bags of i-motifs results in a bag-of-n-motifs. All the bags-of-n-motifs from the secondary structures in a dataset provide a possibly very large ensemble of motifs. From this ensemble of motifs, we will select a set of most relevant motifs that will be used as features for representing the secondary structures from our dataset. To alleviate the text, from now on, the terms "motif," "i-motif" and "n-motif" will be used interchangeably, unless specified otherwise.

2.2 n-motifs representation of secondary structure

The n-motifs obtained in the bag-of-n-motifs model can be thought of as playing a similar role to that of n-grams in n-gram models. If we arrange all the n-motifs in some order, we can create a matrix representation of all the secondary structures. We introduce the matrix $S \in \mathbb{R}^{w*r}$, where w is the total number of secondary structures in the dataset, r is the number of unique n-motifs, and each element of the matrix, s_{ij}, is the frequency of occurrence of a unique n-motif j in the secondary structure i. We denote by s_i the ith row vector of S, corresponding to the ith secondary structure, and by z_j the jth column vector of S, corresponding to the jth n-motif. Not all the unique n-motifs convey the same amount of information. Therefore, it is desirable to select relevant n-motifs to get a refined vector representation of secondary structures, called the n-motifs representation.

Extracting relevant n-motifs from S leads to a better description of secondary structures. The method proposed here for selecting relevant n-motifs is based on an analysis of their frequencies of occurrence. We hypothesize that the relevance of an n-motif is proportional to its occurrence frequency. For instance, among important motifs of a tRNA, the single-stranded region at the $3'$ end is necessary for the binding of amino acids during the translation process. Therefore, we would expect a high occurrence of this motif in a population of secondary structures comprising tRNA. To automatically identify and remove irrelevant n-motifs, i.e. more rarely occurring n-motifs, we utilize the head/tail division rule (Jiang and Liu, 2011). This rule specifies that in the context of a heavy-tailed distribution, the arithmetic mean of the n-motif occurrences yields a natural division between the head (high-occurrence n-motifs) and the tail (low-occurrence n-motifs) of the heavy-tailed distribution.

The distribution of n-motifs ranked by decreasing order of total occurrence follows a heavy-tailed distribution (Supplementary Figure S2). We denote this distribution by f. With x the rank of an n-motif z_j and $f(x) = \sum_{i=1}^{w} s_{ij}$ its total occurrence, f satisfies the condition $\forall x, f(x) \geq f(x+1)$. $f(x) \geq 1$ since an n-motif is present in at least one secondary structure. f follows a heavy-tailed distribution in two regards. On the one hand, it satisfies the long-tailed distribution property: $\lim_{x \to +\infty} f(x+y)/f(x) = 1$, with $y > 0$ (Foss et al., 2011). In fact, low-ranked n-motifs tend to have an occurrence of one because they are present in at least one secondary structure. On the other hand, the long-tailed distribution is a subclass of the heavy-tailed distribution. Thus, we can apply the head/tail division rule to extract relevant n-motifs since f follows a heavy-tailed distribution. An n-motif z_j is considered relevant if the condition $f(x) \geq f_{\text{mean}}$ is satisfied, where $f_{\text{mean}} = \sum_{j=1}^{r} f(x_j) * 1/r$ with r the total number of n-motifs.

After removing irrelevant n-motifs, it is useful to reduce the effect of extreme n-motif occurrences arising from large secondary structures and increase the relative importance of medium n-motif occurrences specific to groups of structures. In fact, large secondary structures naturally tend to have many more n-motifs than small secondary structures. Moreover, while n-motifs specific to groups of structures bear valuable structural information about those groups, they are underestimated because the corresponding n-motifs tend to have medium to low occurrences. To alleviate these effects, we use the logarithm transformation such that $s'_{ij} = \log(s_{ij} + 1)$. s'_{ij} represents the relative relevance or importance of an n-motif z_j to a secondary structure s_i. The n-motifs representation of s_i is defined by the vector s'_i.

The n-motifs representation emphasizes relevant characteristics of secondary structures. It removes irrelevant (i.e. rare) n-motifs, reduces the impact of extremely high occurrences of n-motifs due to large secondary structures and increases the relative importance of n-motifs specific to groups of structures. Given that the n-motifs representation has, in fact, dependent features, in the next subsection, we propose to explore the relationships between n-motifs in order to find the best n-motif combinations characterizing secondary structures.

2.3 Super-n-motifs representation of secondary structure

From the previous two steps, we obtain individual n-motifs that capture geometric properties of RNA secondary structures. However,

some secondary structures sharing few *n*-motifs can belong to a same family. In this subsection, we develop a final vector representation of an RNA secondary structure by computing *n*-motif combinations to effectively represent the structure information. Such combinations should make it possible to capture latent statistical relationships between the *n*-motifs so as to create non-correlated features on which comparisons of secondary structures will be made. We search for the best linear combinations of *n*-motifs using SVD (Golub and Reinsch, 1970).

Let $S' \in \mathbb{R}^{w*m}$ be the matrix of *n*-motifs representations, where w is the total number of secondary structures and m is the total number of relevant *n*-motifs. S' is decomposed by SVD to a product of matrices, as follows:

$$S' = U\Sigma V^T,$$

where $U^T U = 1$ with $U \in \mathbb{R}^{w*w}$ and $V^T V = 1$ with $V \in \mathbb{R}^{m*m}$. Columns of U are the eigenvectors of $S'*S'^T$ and are linear combinations of *n*-motifs. Columns of V are the eigenvectors of S'^T*S' and are linear combinations of S' rows representing secondary structures. Σ is a $w * m$ diagonal matrix containing the square roots of the non-zero eigenvalues of $S'*S'^T$ or singular values of S' in decreasing order, i.e. $\sigma_{1,1} > \sigma_{2,2} > \ldots > \sigma_{r,r} > 0$. By selecting the first k largest singular values in Σ, we obtain a truncated matrix S'_k, where S'_k is defined by:

$$S' \approx S'_k = U_k \Sigma_k V_k^T.$$

S'_k yields the best low-rank approximation of S', such that $S'_k - S'$ (the Frobenius norm) is minimized.

The sum of the first k singular values in Σ_k represents the largest amount of information, in terms of variability, possibly retained using only k variables (features). We consider the matrices $U'_k = U_k \Sigma_k$, where U_k is weighted relative to Σ_k. The space defined by the k vectors of U'_k is called the super-n-motifs space. The super-n-motifs space is a low-dimensional space relative to the original space, since k is typically chosen such that $k \ll m$. The super-n-motifs representation of a structure s'_i is denoted by u'_i.

The super-n-motifs gather particular information concerning the relevant *n*-motifs and their relationships in order to provide a global picture of the structural features of the RNA. The super-n-motifs representation yields a vector representation of a secondary structure in a low-dimensional space. By comparing these representations, secondary structure relationships can be explored. The next subsection presents the comparison of secondary structures in the super-n-motifs representation.

2.4 Comparison of secondary structures based on the super-n-motifs representation

The exploration of secondary structure relationships is facilitated using the super-n-motifs representation. Secondary structures can be effectively compared by computing the cosine dissimilarity (Bonham-Carter *et al.*, 2013; Vinga and Almeida, 2003) between their super-n-motifs representations. Given two secondary structures, s_i and s_j, and their respective super-n-motifs representations, u'_i and u'_j, the cosine dissimilarity between u'_i and u'_j is defined as:

$$d_{\cos}\left(u'_i, u'_j\right) = 1 - \left(u'_i \cdot u'_j\right) / \left(u'_i u'_j\right)$$

s_i and s_j, are considered similar or close to one another when $d_{\cos}\left(u'_i, u'_j\right) \approx 0$ and dissimilar or far from one another when $d_{\cos}\left(u'_i, u'_j\right) \approx 2$.

2.5 Complexity of the super-n-motifs model

The time complexity of the super-n-motifs model is $O(w * m * \min(w, m))$, where w is the number of secondary structures and m is the number of relevant *n*-motifs. It corresponds to the SVD time complexity, since our approach is dominated by the SVD computation. The overall time complexity of the super-n-motifs model is $O(w) + O(r) + O(w*m*\min(w, m))$ where:

- $O(w)$ is associated with the computation of the bag-of-n-motifs model on the w structures;
- $O(r)$ represents the computation required for the selection of the m relevant *n*-motifs out of the total number of *n*-motifs denoted by r;
- $O(w*m*\min(w, m))$ represents the time complexity of SVD (Golub and Van Loan, 1996).

The space complexity of the super-n-motifs model is due largely to the space required to store the matrix of raw *n*-motifs, which is on the order of $O(w*r)$. In detail, the space complexity of the super-n-motifs model is $O(w*r) + O(w*m) + O(w*w) + O(w * m) + O(m * m)$, where:

- $O(w*r)$ is associated with the space required to store the matrix of *n*-motifs, S.
- $O(w*m)$ represents the storage space associated with the matrix of relevant *n*-motifs, S''.
- $O(w*w)$, $O(w * m)$ and $O(m * m)$ are, respectively, the space taken by the matrices U, Σ and V associated with SVD.

Typically, $r > w$, so the space complexity is $O(w*r)$, corresponding to the space necessary to store the matrix associated with space S.

3 Results

We evaluated the model's capacity to perform accurate, efficient comparisons between secondary structures of various sizes from linear and circular RNA comprising pseudoknots and G4s. The accuracy of RNA secondary structure comparison is expressed in terms of discriminative power, i.e. the ability to bring secondary structures from the same family closer together and push secondary structures from different families farther apart. The efficiency is evaluated in terms of the time required to compare sets comprising increasing numbers of secondary structures.

We used the normalized mutual information (NMI) and the f-measure (FMEAS; Manning *et al.*, 2008) to compute the discriminative power of secondary structure comparison algorithms, grouped by their underlying representations, on a medium-sized dataset of 2368 secondary structures from 12 RNA families (Fig. 2 and Supplementary Figure S4) and a large-sized dataset of 15 287 structures from 76 RNA families (Fig. 3). These datasets comprise, respectively, 9 and 66 linear RNA families and 3 and 10 circular RNA families (Andronescu *et al.*, 2008; Garant *et al.*, 2015; Giguère *et al.*, 2014; Nawrocki *et al.*, 2015). Supplementary Section 1 provides a detailed description of these datasets. When the discriminative power in terms of NMI or FMEAS is close to one, it means that all the members of each secondary structure family are close to one another and far from other secondary structures. Details on the computation of NMI and FMEAS are provided in Supplementary Section 2.

For the evaluation on the medium-sized dataset, the approaches and representations carefully chosen from the survey (Schirmer *et al.*, 2014) are as follows: the super-n-motifs model based on 5-gram

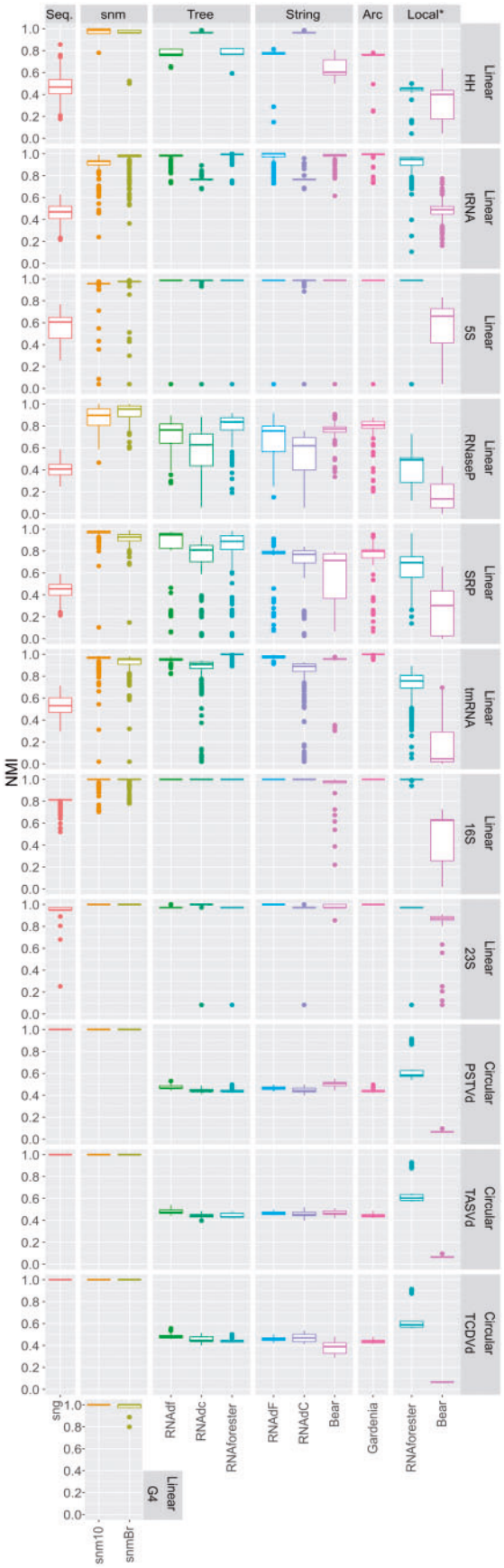

Fig. 2. Algorithms' discriminative power in terms of NMI on RNA secondary structure families. NMI distribution close to one means high discriminative power. Algorithms were grouped according to their representations: sequence (Seq.) super-n-motifs (snm), ordered tree (Tree), string-encoded (String) and arc-annotated sequence (Arc). Local* refers to local alignment version of approaches such as RNAforester and Bear. Families of secondary structures from linear RNA are HH ribozymes, tRNA, 5S rRNA, RNaseP, signal recognition particle (SRP), tmRNA, 16S rRNA, 23S rRNA) and structures with *G*-quadruplexes motifs (G4). Families of secondary structures from circular RNA are potato spindle tuber viroid (PSTVd), tomato apical stunt viroid (TASVd) and tomato chlorotic dwarf viroid (TCDVd).

(sng), the super-n-motifs model based on the super-n-motifs representation with $k = 10$ super-n-motifs (snm10) and the number of super-n-motifs automatically determined using the broken stick model (snmBr; see Supplementary Section 3); RNAdistance with the full tree (RNAdf) and the coarse-grained tree (RNAdc), and RNAforester based on the rooted ordered tree (Tree); RNAdistance with full string (RNAdF) and coarse-grained string (RNAdC), and BEAR based on the string-encoded representation (String); and Gardenia based on the arc-annotated sequence (Arc). In addition to the previously described approaches versions of RNAforester and BEAR. Algorithm parameters used are described in the Supplementary Section 3. For the evaluation on the large-sized dataset, the models compared are sng, snmBr and RNAdF. This latter was chosen because it is one of the approaches achieving the best trade-off between efficiency and discriminative power.

We evaluated the efficiency of the secondary-structure comparison algorithms by computing the time required to perform all-against-all comparisons on sets of 56, 104, 506, 1007 and 15 287 secondary structures (Tables 1 and 2), corresponding, respectively, to 1540, 5356, 127 765, 506 521 and ≈117 million pairwise comparisons. Each of these sets comprises from small (≈36 nt) to large secondary structures (≈2900 nt) of linear RNA (see the sets of secondary structures in Supplementary File S1), corresponding to the size variability generally found in RNA secondary structures. To avoid redundancy, it is only necessary to compute $(w^2 - w)/2$ comparisons, where w is the total number of secondary structures in a dataset.

3.1 Analysis of discriminative power

3.1.1 Performance on linear RNA families with high structural variability

From the results shown in Figure 2 and Supplementary Figure S4, we can remark that the super-n-motifs model demonstrates comparable or superior discriminative power compared to other approaches on families with high structural variability such as RNaseP, SRP and HH. In fact, the super-n-motifs model performs better on the RNaseP family compared to other methods, showing its capability to identify substructures or local structures shared among the set of highly variable secondary structures. Indeed, RNaseP structures from our dataset come from three domains, Archea, Bacteria and Eukaryotes, and the structures corresponding to these domains are known to be highly variable while showing a central conserved core (Evans *et al.*, 2006). Specifically, RNaseP structures exhibit very diverse conformations, due to more structural rearrangement than extended stems or loops, as these structures do not differ much in size, i.e. ≈337.66 nt ± 26.25 SD (see Supplementary Table S1). The diversity of RNaseP structures can be visualized in Supplementary Figure S5, where the structures of ASE_00264, ASE_00099 and ASE_00346 belong, respectively, to the Eukarya, Bacteria and Archea domains.

The super-n-motifs model also shows comparable discriminative power on the SRP family compared to the ordered tree-based

approach, RNAdf, and superior discriminative power compared to all the other approaches. Given that SRP structures in our dataset are highly variable in terms of size (Rosenblad et al., 2009), this suggests that the super-n-motifs model is much less affected by the size variability of structures. In fact, the SD relative to the mean size of structures in the SRP family is very high, ≈266.30 nt ± 65.46 SD (Supplementary Table S1).

As for circular permuted structures, the results shown in Figure 2 and Supplementary Figure S4 suggest that the super-n-motifs model performs very well on the HH family, similar to a coarse-grained approach such as RNAdc and RNAdC. This shows its capacity to handle circular permuted structures. Structural variability in the HH family is due to the fact that our dataset comprises hammerhead (HH) structures of types I and III with a large variability in size, ≈61.71 nt ± 24.16 SD (Supplementary Table S1). It is known that the HH family is comprised of three groups, types I, II and III, which are circularly permuted forms of HH distinguished by the open-ended helix that connects the motifs with the flanking sequences (Hammann et al., 2012). The structural variability of HH in our dataset can be seen by examining structures RFA_00660 and RFA_00407, representing HH of types I and III, in Supplementary Figure S5.

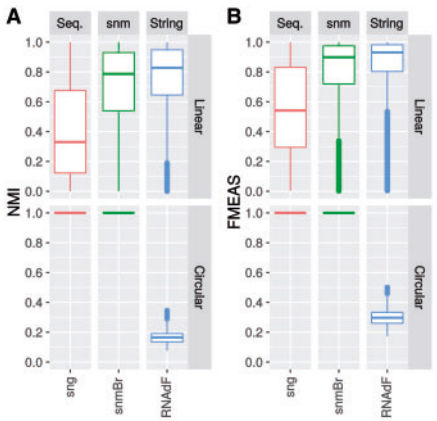

Fig. 3. Discriminative power of sng, snmBr, RNAdF and RNAdC in terms of NMI and FMEAS on RNA secondary structure families from a large dataset of 15 287 secondary structures belonging to 76 families. The discriminative power of each model is reported separately according to its performance on linear and circular RNA families.

3.1.2 Performance on linear RNA families with conserved structures

Performance comparison on families of conserved secondary structures shows that the super-n-motifs model displays high discriminative power similar to that of the majority of the tested approaches. Indeed, the super-n-motifs model, as well as RNAdf, RNAdF, RNAforester, BEAR and Gardenia, generate excellent NMI and FMEAS results on families with conserved structures such as the tRNA, 5S, tmRNA, 16S and 23S families (Fig. 2 and Supplementary Figure S4). The structural conservation of these families can be explained by the fact that not much size variability for any of these families is observed in our dataset. In fact, a small SD compared to the average size of structures is observed for the tRNA, 5S, tmRNA, 16S and 23S families. For each of these families, respectively, we have average sizes and SDs of ≈77 nt ± 5.42 SD, ≈120 nt ± 7.55 SD, ≈362.46 nt ± 6.72 SD, ≈1577.35nt ± 110.83 SD and ≈2908.09 nt ± 28.38 SD (Supplementary Table S1).

3.1.3 Performance on circular RNA families, *G*-quadruplexes and pseudoknots

The super-n-motifs model is designed to handle secondary structures from circular RNA families seamlessly. In fact, our model shows by

Table 2. Running times (in d: days and s: seconds) of sng, snmBr and RNAdF on the large-sized dataset of 15 287 secondary structures corresponding to ≈117 million pairwise comparisons.

Representation	Algorithm	Time complexity	Running times
Sequence	sng	$O(wm\ min(w,m))$	1697s
Super-n-motifs	snmBr	$O(wm\ min(w,m))$	**179s**
String	RNAdF	$O(l^2)$	41d[a]

Note: sng and snmBr were run on an 8 CPU (1.8 GHz) desktop pc with 5.8 GB of RAM and RNAdF was run in parallel on a 24 core node of the super computer Mammouth-mp2. w, m and l refer to the number of structures, the number of computed relevant n-motifs and the length of the strings representing the secondary structures. The best result is shown in bold.

[a]Estimated CPU time, that is, the sum of CPU time consumed by all of the CPUs.

Table 1. Running times (in d: days, m: minutes and s: seconds) on 56, 104, 506 and 1007 secondary structures, corresponding to 1540, 5356, 127 765 and 506 521 pairwise comparisons.

Representation	Algorithm	Time complexity for one pair. comp.	56 structures 1540 pair. comp.	104 structures 5356 pair. comp.	506 structures 127765 pair. comp.	1007 structures 506521 pair. comp.
Sequence	sng	$O(wm\ min(w,m))$	**0.12s**	**0.22s**	**3s**	35s
Super-n-motifs	snmBr	$O(wm\ min(w,m))$	**0.12 s**	0.66 s	**3 s**	**4 s**
Tree	RNAdf	$O(l^2d^2)$	4 m 29 s (10^3 f.)	16 m 24 s (10^3 f.)	5 h 44 m (10^3 f.)	14 h 7 m (10^4 f.)
	RNAdc	$O(l^2d^2)$	2 s (10^1 f.)	7 s (10^1 f.)	2 m17 s (10^1 f.)	5 m 4 s (10^1 f.)
	RNAforester	$O((l^2/r)q^2)$	1 h 31 m (10^4 f.)	10 h 5 m (10^5 f.)	>1d	>1d
String	RNAdF	$O(l^2)$	31 s (10^2 f.)	1 m 47 s (10^2 f.)	35 m 58 s (10^3 f.)	1 h 38 m (10^3 f.)
	RNAdC	$O(l^2)$	1 s	4 s	1 m 14 s (10^1 f.)	3 m 22 s (10^1 f.)
	BEAR	$O(l^2)$	2 m 31 s (10^3 f.)	8 m 34 s (10^2 f.)	3 h 17 m (10^3 f.)	7 h 54 m (10^3 f.)
Arc	Gardenia	$O(l^4)$	18 m 52 s (10^3 f.)	1 h 32 m (10^4 f.)	>1d	>1d
	ERA	$O(l^3)$	8 h 51 m (10^5 f.)	>1d	>1d	>1d
Ensemble	Sparse	$O(l^2)$	>1d	>1d	>1d	>1d

Note: The algorithms were run on an 8 CPU (1.8 GHz) desktop pc with 5.8 GB of RAM. f. and pair comp. stand for fold and pairwise comparisons. w and m refer to the number of structures and the number of computed relevant n-motifs. l represents the size of trees or forests, strings or arc-annotated sequences or number of base pairs representing secondary structures. d is the depth of trees. r and q are, respectively, the number of anchors and the maximum degree of forests. The best results in each column are shown in bold.

far the best discriminative power compared to all other methods, regardless of whether they are based on global or local alignment, on secondary structures from the 3 circular RNA families of the medium-sized dataset and the 10 circular families of the large-sized dataset (Fig. 2, Supplementary Figure S4 and Fig. 3). The clear advantage of our approach can be explained by the fact that, contrary to the other methods that assume contiguity of structural features, our model does not. Indeed, our model defines a secondary structure as an unordered collection of structural features, the *n*-motifs. Consequently, it captures structural features irrespective of their orders, which is an important property that allows it to effectively handle not only circular RNA but also RNA with the circularly permuted structures found in the HH family.

The results reported in Figure 2 and Supplementary Figure S4 demonstrate that the proposed approach has the capacity to handle pseudoknots and G4s. In fact, the super-n-motifs model demonstrates high discriminative power on secondary structures from linear RNA families comprising pseudoknots, such as RNaseP, tmRNA, 16S and 23S rRNA (see Supplementary Table S1). Moreover, our model effectively processes G4s, since it is capable of discriminating secondary structures that contain a single G4 from those without any G4s. It is important to note that while all the other approaches simply ignore these motifs, our model processes pseudoknots and G4 effectively, thanks to the flexibility of the bag-of-n-motifs that allows the explicit integration of any kind of motif (here, pseudoknots and G4 motifs) in the structural description of secondary structures. Our model makes no distinction between various types of pseudoknots, such as the H-type or the three-stemmed RNA pseudoknot (Staple and Butcher, 2005).

3.1.4 Large-scale performance analysis

A very important advantage of our approach is that it can be used to handle very large datasets in a very time-efficient way while maintaining high discriminative power as the number of structures grows. As shown in Figure 2, it achieves discriminative power comparable or superior to that of the tested approaches on linear RNA and consistently outperforms them on circular RNA in the medium-sized dataset of 2368 structures with 12 families. On the large-sized dataset of 15 287 structures with 76 families, we compared our approach with sng and RNAdF (Fig. 3 and Table 2). RNAdF was chosen because it is one of the most efficient approaches, combining good time complexity, $O(l^2)$, and high discriminative power rivaling that of approaches based on ordered-tree and arc-annotated sequences, which have a time complexity of at least $O(l^2d^2)$ (see Table 1 and Fig. 2). From Figure 3 and Table 2, it can be seen that our approach yields discriminative power comparable to that of RNAdF and superior to that of sng, yet significantly outperforms them in term of running times. This evaluation shows at least that our approach is indeed effective and efficient on a large dataset, i.e. 3 min as compared to the equivalent of 41 days for RNAdF. Currently, a thorough comparative study of different approaches on large datasets like the one used here is not possible since most competing approaches would need more than 1 month to produce their results on a single workstation.

3.1.5 Structural information at the coarse-grained and fine-grained levels

The results in Figure 2 and Supplementary Figure S4 show that the super-n-motifs model yields consistently high discriminative power compared to approaches using coarse-grained representation of structures (in which subsets of nucleotides or base pairs forming motifs such as stems or hairpin loops are considered as elements of the representation), for instance, RNAdc and RNAdC, on the HH family. It performs equally well as, and sometimes better than, approaches using fine-grained representation (in which each nucleotide or each base pair is considered as an element of the representation). These approaches include RNAdf, RNAdF, RNAforester, BEAR and Gardenia (Fig. 2 and Supplementary Figure S4). We observe that approaches based on fine-grained representation of structures perform better than approaches based on coarse-grained representation because the latter often results in a loss of structural information. Our approach combines the advantages of both fine-grained and coarse-grained representation and performs well on all the families. This can be explained by the fact that the structural features captured by our model represent coarse-grained features such as motifs and neighboring motifs, on the one hand, and fine-grained features such as the number of base pairs in a stem or the number of string-stranded nucleotides in loops on the other.

3.1.6 Sufficiency of structural information for separating families

We performed a comparison between structure-based and sequence-based methods to see whether structural information could be sufficient to separate families. We observed that structure-based methods, including snmBr, yield high discriminative power, while sequence-based methods like sng consistently have low discriminative power on most linear RNA families (Fig. 2, Supplementary Figure S4 and Fig. 3). In fact, for the medium-sized dataset families such as HH, tRNA, 5S, RNaseP, SRP and tmRNA and the large-sized dataset families, sng yields low performance, indicating that most RNAs, based on sequence information, have been assigned to the wrong family. These results show that, in our context, structural information is more important than sequence information for distinguishing RNA families. It is important to note that sng performs well on circular RNA, since sng, like snmBr, is insensitive to the RNA direction by the fact that it captures the unordered statistical patterns of sequences, the n-grams. Consequently, sng can be a good candidate to compare circular RNA sequences.

3.2 Efficiency assessment

The efficiency of the super-n-motifs model is shown by Tables 1 and 2. We observed that it is faster, by up to 4 or 5 orders of magnitude, than all the tested approaches, on data varying from very small-sized sets of 56 structures, corresponding to 1540 pairwise comparisons, to the large-sized set of 15 287 structures corresponding to ≈117 million pairwise comparisons. It exhibits a linear running time: it took 0.15 s, 0.27 s, 4 s, 7s, 12s and 182s, respectively, to compute all-against-all pairwise comparisons of 56, 104, 506, 1007, 2330 and 15 287 secondary structures of various sizes (from ≈36 to ≈2900 nt), corresponding to 1540, 5356, 127 765, 506 521, ≈2 million, ≈117 million pairwise comparisons (see Tables 1 and 2 and Supplementary Table S2). For the large-sized dataset of 15 287 structures, our approach is 4 orders of magnitude faster than RNAdF, since it took 182 s. (3 min) to compute ≈117 million comparisons from the set of 15 827 secondary structures, a task that would take ≈41 days of estimated CPU time for RNAdF (Table 2).

Tables 1 and 2 suggest that the alignment-based methods are convenient for low or medium-scale analysis but not suitable for large-scale analysis. RNAforester, Gardenia, ERA and SPARSE are more suitable for low-scale analysis, since even to compare 506 structures they already require days of computation. It is worth mentioning that ERA and SPARSE exhibit the highest running times: 8 h 51 m and more than a day to compare 56 structures

corresponding to 1540 comparisons. For this reason, we were unable to evaluate their respective discriminative power on the medium-sized datasets requiring computation of ≈2 million pairwise comparisons.

The results presented in the previous paragraphs are to be expected, because the majority of the tested approaches typically use DP as a foundation to compute alignment. Indeed, it is known that DP-based approaches are computationally demanding (Bonham-Carter *et al.*, 2013). Alignment-free approaches scale well and are of interest in sequence analysis to handle the vast numbers of transcripts generated by high-throughput sequencing methods (Bonham-Carter *et al.*, 2013; Vinga, 2014; Vinga and Almeida, 2003;). As the number of validated secondary structures grows, mainly due to high-throughput probing techniques, an alignment-free approach such as the super-n-motifs model offers a good alternative to the existing methods.

4 Discussion

Understanding RNA functions by comparing RNA secondary structures is challenging for several reasons. On the one hand, secondary structures are complex due to the nature of RNA, which can take a linear or a circular configuration and may contain pseudoknots and G4 motifs. On the other hand, thousands of secondary structures are generated by high-throughput probing techniques. In this paper, we proposed the super-n-motifs model and demonstrated its effectiveness and efficiency at comparing RNA secondary structures.

Our model computes accurate comparisons of secondary structures and naturally tends to cluster structures in a way that reflects known secondary structure families. This is an important property, because it is expected that members of a family of RNA secondary structures performing related functions form a natural cluster. This can be particularly helpful for RNA annotation, structure-based phylogeny, homology search in databases and identification of new families in populations of RNA. Since our model efficiently handles pseudoknots and G4 motifs, it can also help in understanding their functional roles.

In future work, our approach can be extended to compare RNA on the basis of sequences and secondary and tertiary structures. In fact, in addition to pseudoknots and G4 motifs, our model can be extended to handle other motifs such as sarcin-ricin, kink turn or c-loop motifs, thus making it possible to combine motifs from the secondary structures with the tertiary structures. Sequence information can be incorporated with n-grams as in sng method. Our model can thus yield a rich and global view of RNA by combining sequence and structural features.

Funding

This work was supported by a joint grant from the Fonds de Recherche du Québec—Nature et Technologies (FRQ-NT) and the Université de Sherbrooke Research Chair on RNA Structure and Genomics to JPP and the Natural Sciences and Engineering Research Council of Canada to SW. The Mammouth parallèle II supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the FRQ-NT.

Conflict of Interest: none declared.

References

Allali, J. and Sagot, M.F. (2008) A multiple layer model to compare RNA secondary structures. *Softw. Pract. Exp.*, **38**, 775–792.

Andronescu, M. *et al.* (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Bellaousov, S. *et al.* (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.*, **41**, W471–W474.

Blin, G. *et al.* (2010) Alignments of RNA structures. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **7**, 309–322.

Bonham-Carter, O. *et al.* (2013) Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinformatics*, **15**, 890–905.

Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 113–137.

Eddy, S.R. (2004) What is dynamic programming? *Nat. Biotechnol.*, **22**, 909–910.

Evans, D. *et al.* (2006) RNase P: interface of the RNA and protein worlds. *Trends Biochem. Sci.*, **31**, 333–341.

Fernandes, F. *et al.* (2009) CSA: an efficient algorithm to improve circular DNA multiple alignment. *BMC Bioinformatics*, **10**, 230.

Flores, R. *et al.* (2012) Viroids and hepatitis delta virus. *Semin. Liver Dis.*, **32**, 201–210.

Foss, S. *et al.* (2011) Heavy-tailed and long-tailed distributions. In: *An Introduction to Heavy-Tailed and Subexponential Distributions SE - 2*, Springer Series in Operations Research and Financial Engineering. Springer, New York, pp. 7–38.

Garant, J.M. *et al.* (2015) G4RNA: an RNA G-quadruplex database. *Database*, doi: 10.1093/database/bav059.

Giguère, T. *et al.* (2014) Comprehensive secondary structure elucidation of four genera of the family Pospiviroidae. *PLoS One*, **9**, e98655.

Golub, G.H. and Van Loan,C.F. (1996) Matrix computations. *Phys. Today*, **10**, 48.

Golub, G.H. and Reinsch, C. (1970) Singular value decomposition and least squares solutions. *Numer. Math.*, **14**, 403–420.

Guignon, V. *et al.* (2005) An edit distance between RNA stem-loops. In: Consens, M. and Navarro, G. (eds), *String Processing and Information Retrieval SE 38*, Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 335–347.

Hammann, C. *et al.* (2012) The ubiquitous hammerhead ribozyme. *RNA*, **18**, 871–885.

Haubold, B. (2014) Alignment-free phylogenetics and population genetics. *Brief. Bioinformatics*, **15**, 407–418.

Hendrix, D.K. *et al.* (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, **38**, 221–243.

Huppert, J.L. *et al.* (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.

Jeck, W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.

Jiang, B. and Liu, X. (2011) Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *Int. J. Geogr. Inf. Sci.*, **26**, 215–229.

Kosik, K.S. (2013) Circles reshape the RNA world. *Nature*, **495**, 4–6.

Lorenz, R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Loughrey, D. *et al.* (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res.*, **42**, e165–e165.

Manning, C.D. *et al.* (2008) Introduction to Information Retrieval. *J. Am. Soc. Inf. Sci. Technol.*, **1**, 496.

Mattei, E. *et al.* (2014) A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res.*, **42**, 6146–6157.

Millevoi, S. *et al.* (2012) G-quadruplexes in RNA biology. *Wiley Interdiscip. Rev. RNA*, **3**, 495–507.

Mosig, A. *et al.* (2006) Comparative Analysis of Cyclic Sequences: viroids and other Small Circular RNAs. In: *Lecture Notes in Informatics, German Conference on Bioinformatics,* **83**, pp. 93–102.

Nawrocki, E.P. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.

Pinello, L. *et al.* (2014) Applications of alignment-free methods in epigenomics. *Brief. Bioinformatics*, **15**, 419–430.

Rosenblad, M.A. *et al.* (2009) Kinship in the SRP RNA family. *RNA Biol.*, **6**, 508–516.

Schirmer, S. *et al*. (2014) Introduction to RNA secondary structure comparison. In: Gorodkin, J. and Ruzzo,W.L. (eds), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods SE 12*, Methods in Molecular Biology. Humana Press, New York, pp. 247–273.

Schirmer, S. and Giegerich, R. (2013) Forest alignment with affine gaps and anchors, applied in RNA structure comparison. In: *Theoretical Computer Science*, pp. 51–67.

Song, K. *et al*. (2014) New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinformatics*, **15**, 343–353.

Staple, D.W. and Butcher, S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*., **3**, 0956–0959.

Tinoco, I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol*., **293**, 271–281.

Underwood, J.G. *et al*. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.

Vinga, S. (2014) Editorial: alignment-free methods in computational biology. *Briefings Bioinformatics*, **15**, 341–342.

Vinga, S. and Almeida, J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

Wan, Y. *et al*. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet*., **12**, 641–655.

Will, S. *et al*. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol*., **3**, 680–691.

Will, S. *et al*. (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.

Will, S. and Stadler, P.F. (2014) Algorithms in Bioinformatics. In: Brown, D. and Morgenstern, B. (eds) *Proceedings of 14th International Workshop, WABI 2014*, Wroclaw, Poland, September 8–10, 2014, pp. 135–147 Springer, Berlin/Heidelberg.

Zhong, C. and Zhang, S. (2013) Efficient alignment of RNA secondary structures using sparse dynamic programming. *BMC Bioinformatics*, **14**, 269.