PLOS | ONE

RESEARCH ARTICLE

Accurate Classification of RNA Structures Using Topological Fingerprints

Jiajie Huang[1,2], **Kejie Li**[1,3], **Michael Gribskov**[1,4]*

1 Department of Biological Sciences, Purdue University, West Lafayette, Indiana, United States of America, **2** Life Sciences Solutions Group, Thermo Fisher Scientific, South San Francisco, California, United States of America, **3** Computational Biology Department, Biogen Idec, Cambridge, Massachusetts, United States of America, **4** Department of Computer Science, Purdue University, West Lafayette, Indiana, United States of America

* gribskov@purdue.edu

OPEN ACCESS

Citation: Huang J, Li K, Gribskov M (2016) Accurate Classification of RNA Structures Using Topological Fingerprints. PLoS ONE 11(10): e0164726. doi:10.1371/journal.pone.0164726

Editor: Danny Barash, Ben-Gurion University, ISRAEL

Received: April 1, 2016

Accepted: September 29, 2016

Published: October 18, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by The U.S. National Science Foundation [DBI-0850148] (http://www.nsf.gov/). This funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The first author, Jiajie Huang, is currently employed by Thermo Fisher Scientific. The second author, Kejie Li, is currently employed by Biogen Idec. These commercial affiliations provided support in the form of salaries for these two authors, but did not have any additional role in the study design, data

Abstract

While RNAs are well known to possess complex structures, functionally similar RNAs often have little sequence similarity. While the exact size and spacing of base-paired regions vary, functionally similar RNAs have pronounced similarity in the arrangement, or topology, of base-paired stems. Furthermore, predicted RNA structures often lack pseudoknots (a crucial aspect of biological activity), and are only partially correct, or incomplete. A topological approach addresses all of these difficulties. In this work we describe each RNA structure as a graph that can be converted to a topological spectrum (RNA fingerprint). The set of subgraphs in an RNA structure, its RNA fingerprint, can be compared with the fingerprints of other RNA structures to identify and correctly classify functionally related RNAs. Topologically similar RNAs can be identified even when a large fraction, up to 30%, of the stems are omitted, indicating that highly accurate structures are not necessary. We investigate the performance of the RNA fingerprint approach on a set of eight highly curated RNA families, with diverse sizes and functions, containing pseudoknots, and with little sequence similarity–an especially difficult test set. In spite of the difficult test set, the RNA fingerprint approach is very successful (ROC AUC > 0.95). Due to the inclusion of pseudoknots, the RNA fingerprint approach both covers a wider range of possible structures than methods based only on secondary structure, and its tolerance for incomplete structures suggests that it can be applied even to predicted structures. Source code is freely available at https://github.rcac.purdue.edu/mgribsko/XIOS_RNA_fingerprint.

Introduction

Once seen as a simple scaffold, RNA is now known to play important regulatory and catalytic roles. RNA is involved in processes including transcriptional regulation [1], RNA maturation and modification [2], and RNA splicing [3]. The structural motifs in RNA that are responsible for its functions are evolutionarily conserved; however, unlike DNA and protein, for which conserved functional motifs can be identified based on sequence similarity, the functional motifs in RNA may have little or no sequence similarity [4], and instead conserve patterns of

collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the "author contributions" section.

Competing Interests: The first author, Jiajie Huang, is currently employed by Thermo Fisher Scientific. The second author, Kejie Li, is currently employed by Biogen Idec. These commercial affiliations do not alter the authors' adherence to all PLOS ONE policies on sharing data and materials.

base-pairing (stems) and topological relationships between base-paired regions, for instance nesting of stems, multi-loops, and pseudoknots [5, 6]. This topological view of RNA structure has been discussed by Giegerich *et al.* who point out that, in a family of RNAs with the same function, the global arrangements of structural elements (topology) are conserved, but there is considerable variation in the length of stems, presence of bulge loops and unpaired bases, and type of base-pairs. Therefore, in the study of RNA functions, it may be more relevant to look at global topological patterns than individual base-pairs [7, 8]. RNAs with similar functions, for example those in ribonuclease P (RNase P), the ribosome, or self-splicing introns, typically have strongly conserved topologies [5, 9–11]. One of the notable topological aspects of RNA structure is the importance of pseudoknots in many classes of molecules. For example, in Hepatitis Delta Virus (HDV), a double-pseudoknotted structure contained in a self-cleaving ribozyme is a key factor in HDV infection [12]; in Group I self-splicing introns, the catalytic core is formed by pseudoknots [13]; in ribosomal RNA, pseudoknots at the catalytic site are the key structures that mediate microbial resistance to antibiotics [14] and stimulate viral frame-shifting [15].

As only a small number of functional RNA classes have been identified, we believe that the majority of regulatory and functional RNA motifs are yet to be identified. Eukaryotic genomes are pervasively transcribed [16]; almost every base can be found in an RNA transcript. This is surprising since, in most genomes, protein-coding sequences comprise only a small fraction of the genome. Much of this RNA is therefore likely to be regulatory in nature, and will almost certainly contain functionally important structures, including pseudoknots.

Just as conserved structural topologies are important for RNA function, the identification of novel conserved topologies provides an approach to discovering the functions of currently unknown classes of biologically important RNAs. An analogy can be made to the importance of sequence alignment and database searching programs in identifying novel proteins and DNA regulatory elements. While typical functional RNA structures are pseudoknotted, the current computational approaches to RNA structure comparison only consider structures without pseudoknots. Because of their importance to RNA function, we believe that incorporating pseudoknots in structural comparisons is critical to identifying biologically important classes of molecules. In this paper we propose a straightforward approach to comparing RNA structural topologies, including pseudoknots, and identifying known and unknown conserved topologies. Waterman [17] introduced the first graphical representation of RNA structure, the full-graph, where the nodes represents nucleotides. The tree-graph representation was later introduced by Shapiro *et al.*, which is an abstract tree where the nodes represent structural elements [18–20], and this coarse-grained representation was implemented in the ViennaRNA package [21]. Fontana *et al.* implemented the homeomorphically irreducible tree (HIT) that represents an RNA secondary structure as a contracted topology in which each node represents a structural element weighted by size [22]. Shu *et al.* have developed the element-contact graphs (ECGs) with size-weighted nodes as well [23], which uses topological indices, such as the Randić index [23, 24], the Wiener index, and Balaban index, to measure graph connectivity [23, 24]. Although the ECGs framework and an extended Wiener index [25] were developed to be able to identify small ncRNAs such as miRNAs, no evidence was shown for its ability to classify larger RNAs (for example, 23S rRNA are usually over 1000nt long) with low sequence similarity. The RNAshapes package [7, 8] of Giegerich *et al.*, which represents RNA structures as abstract shapes and aims for efficient RNA structure comparisons, has been shown useful in topologically clustering RNA families; however, RNAshapes does not perform well on families with pseudoknots [26]. Building on this work, Heyne *et al.* developed a graph-based pipeline called GraphClust [27] for fast clustering of RNA molecules. In this approach, RNA secondary structures are generated by the RNAshapes package from input sequences, encoded by graphs preserving nucleotide connectivity, and clustered by a graph kernel, the Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [28]. However, given data sets of small RNA sequences (sequence length $<$ 400nt, similarity up to 80%) the precision and recall of GraphClust only reaches around 85%. In addition, these approaches do not include pseudoknots in either the representation or the analysis.

The Schlick group has developed the RNA-As-Graphs procedure, which combines elements from several sources to develop a database of mathematically possible RNA graphs in which RNA structures are represented as Shapiro tree graphs, without pseudoknots, or as dual graphs, with inclusion of pseudoknots [29–32]. Numerical descriptors have been applied to comparison of these RNA topological patterns. The eigenvalue spectrum of the Laplacian matrix measures graph compactness and connectivity; λ_2, the second eigenvalue of the Laplacian matrix [33–35], measures RNA graph similarity. The Schlick group used several structural invariants, including λ_2 and linear combinations of α and β (the intercept and slope of the eigenvalues of the Laplacian matrix), for categorizing the structural similarity of RNA graphs, and for predicting whether randomly generated RNA topologies are similar to biological examples (RNA-like). These numerical descriptors, however, have never been shown to be able to group RNAs into structural/functional classes. Moreover, these approaches, which rely on a small number of numeric descriptors, cannot identify similarity between specific substructures nested within fairly large graphs (for instance graphs of the size of RNase P RNA, which may have up to 20 vertices).

There are several aspects of RNA structure that make it particularly hard to identify topologically similar structures. Structures from the same functional family may have little or no sequence similarity; they typically have a similar arrangement of stems (topology), but different local base-pairing; our knowledge of the structures may be incomplete due to lack of a high-quality three-dimensional structure or structural prediction; structures may lack biologically important pseudoknots since tractable computational approaches based on dynamic programming often do not include these important features; or in the case of graph comparison, the computation itself may require infeasible amounts of time. The RNA XIOS graph [36] explicitly represents serial, nested, pseudoknotted, and mutually exclusive stems, but finding topologically similar RNA structures requires identifying isomorphous subgraphs common to one or more structures. The coarse-grained approach we describe here builds on the XIOS approach, addresses the problems described above, and provides a feasible approach to clustering and identifying biological RNAs with topologically similar structures. A coarse-grained RNA secondary structure representation has also been used to predict RNA deleterious mutations by RNAMute [37] and RNAmutants [38]. In addition, a coarse-grained approach is also applied for RNA design such as RNAexinv [39] and Nanofolder [40]. We demonstrate the utility of the XIOS approach by classifying a representative set of pseudoknot-including RNA structural families that have very low levels of sequence similarity–the high accuracy of the classification indicates that this approach can be broadly applied to identifying RNAs with conserved topologies, whether their function is known or unknown.

Materials and Methods

Curated RNA families

A set of curated RNA structures have been collected from the literature and a variety of biological databases [41], and is extended in this work (S1 Table). This set of known structures has been carefully selected to contain pseudoknots, to cover a broad range of lengths, and to have been the subject of extensive expert curation by the biological community. This curated set includes 206 structures of transfer RNA, Ribonuclease P RNA, transfer-messenger RNA, group I and group II self-splicing introns, and 5S, 16S and 23S ribosomal RNA. The structures in this curated set have been reviewed to ensure they reflect expert opinion on the correct structure, and to ensure that the reported structures are as accurate as possible given existing experimental data such as X-ray crystallography [42, 43] and covariance analysis [44]. The curated structures have been screened to ensure that all structures are full-length, and no pair of structures has greater than 50% sequence identity. Multiple families of the curated structures contain pseudoknots. While several large databases of RNA structures exist, for instance Rfam [45] and RNAStrand [46], these resources could not be directly used for testing in this work because many families lack pseudoknots, lack a consensus of expert opinion on the correct structure(s), have only a family consensus structure (rather than individual structures for each RNA), have high levels of sequence identity, or comprise incomplete structures or structures in which single stranded regions (or other regions judged to be unimportant) have been removed.

XIOS graphs

In a XIOS graph, RNA stems are shown as vertices and the relationships between stems are shown as edges [36]. Edges may be one of four types: X–mutually exclusive (stems with base conflicts, such as those in two alternative structures that use the same RNA sequence); I–included (nested); O–overlapping (pseudoknotted); S–serial (adjacent) (Fig 1). Because there are exactly four classes, and each pair of stems can have one and only one type of relationship, we can omit S relationships without loss of generality (any pair of vertices without an edge have an implicit S edge). In this work, none of the structures have X edges; the graphs therefore have only two edge types, I and O. S1 Fig shows the XIOS graph representation of the Hepatitis D Virus (HDV) ribozyme RNA.

Curated XIOS graphs

S1 Table shows the vertex number, edge number, and average degree of the XIOS graphs of the curated RNA structures. Graph matching is highly dependent on the size of the graph

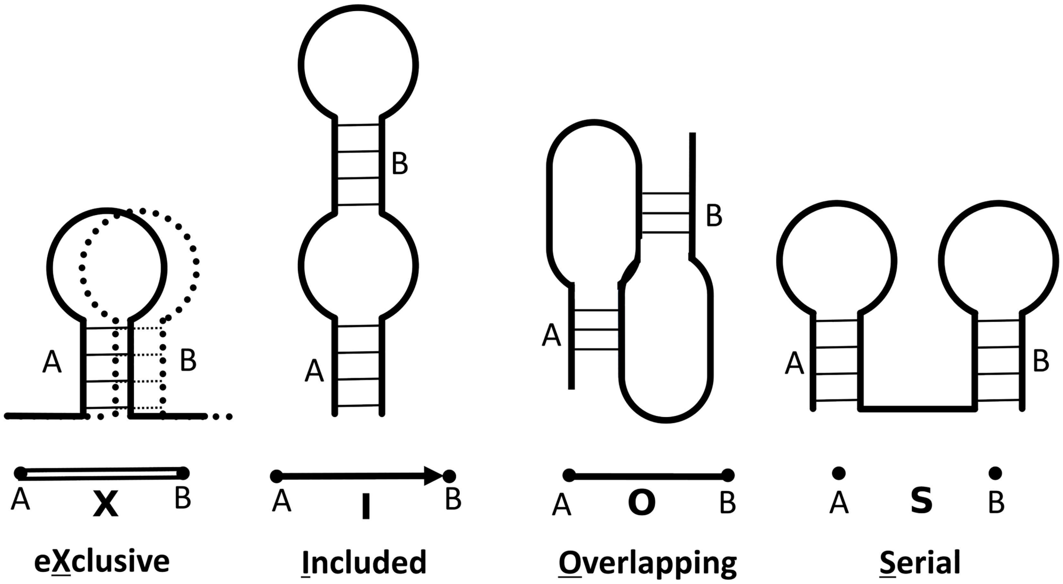

Fig 1. XIOS graph stem-stem relationships. Edges show the relationship between two stems, and may be one of four types: X (mutually exclusive), I (included or nested), O (overlapping or pseudoknotted), or S (serial or adjacent).

doi:10.1371/journal.pone.0164726.g001

(described by the number of vertices and edges) and the average degree of the vertices in the graph; the characteristics of the curated RNA structures differ significantly between families making this a representative set for RNAs in general.

Results

This work focuses on the topological similarity between RNA structures, that is, similarity in the relative location and nesting of stems, and the location of pseudoknots. In principle, this should provide the broadest range of matching since individual structures often differ in the length of stems and the length of single-stranded regions between stems. As mentioned before, the sequences themselves can be even more variable with little or no sequence conservation detectable, even between RNAs with similar structures. Topologically similar substructures in a pair of RNAs correspond to isomorphous subgraphs in their respective XIOS graphs. The maximal common subgraph (MCS) represents the greatest possible topological match between RNAs, similar to the maximal alignment between two sequences. But the MCS is difficult to identify because of the large size of biologically important structures; *e.g.*, the 23S rRNA can have more than 50 stems [47]. Finding the MCS of a set of graphs, corresponding to the largest conserved topological motif in a group of RNA structures, is an NP-hard problem [48], making the computational identification of the MCS time consuming. In order to decrease the inefficient scaling inherent in graph matching, we characterize each graph as a set of smaller subgraphs. We call this set of subgraphs the RNA topological fingerprint, or more simply, the RNA fingerprint. There are two key elements needed to determine an RNA fingerprint: a comprehensive dictionary of RNA topological motifs, and an approach to identifying the motifs that are present in a XIOS graph.

Enumerating a comprehensive set of RNA topologies

We have exhaustively enumerated a non-redundant set of all physically possible RNA topological motifs containing from one to seven stems (Table 1). The graphs in this set are all IO-connected, that is, all vertices (stems) can be reached by traversing I and O edges. Briefly, a complete set of topologies for an N-stem RNA structure can be created by generating all the permutations of an ordered set of 2N numbers; the numbers represent N objects (stems), numbered 1 to N, each with two instances (corresponding to the two base-paired halves of the stem). For three stems (N = 3), the ordered unpermuted set would be (1, 1, 2, 2, 3, 3), with each pair of matching numbers representing the two base-paired halves of a stem. The unpermuted set, above, would thus correspond to three serial stems, and a permuted set such as (1, 2, 3, 2, 3, 1) would indicate a pair of pseudoknotted stems, 2 and 3, found within the loop of stem 1.

Obviously, this procedure generates multiple copies (isomorphs) of some topologies, for instance (1, 2, 2, 3, 3, 1) and (3, 1, 1, 2, 2, 3), as well as some graphs that are not connected (for

Table 1. Topological Motif Library.

Number of Stems	Unique Topologies
1	1
2	2
3	8
4	46
5	368
6	3914
7	51390

doi:10.1371/journal.pone.0164726.t001

instance the unpermuted set, above). Some of the isomorphs can be eliminated by imposing two restrictions. First, the graph must be connected, and second, the first instances (left half stem) of each object (stem) must occur in numerical order. Even these restrictions do not entirely eliminate permutations that correspond to isomorphic XIOS graphs. For instance, the sets (1, 2, 1, 3, 3, 2) and (1, 2, 2, 3, 1, 3) are mirror images of each other, and correspond to the same XIOS graph. These symmetry-related topologies are detected and removed using the gSpan [36, 49] approach. In gSpan, a graph is described using a canonical labeling called the minimum DFS code; Isomorphic graphs are guaranteed to have identical minimum DFS codes.

Using this approach, we have enumerated a library of all unique physically possible RNA topologies with 2 to 7 stem structures (Table 1). Because the minimum DFS code provides a unique description for each topology, we index the motif library with a compressed version of the minimum DFS code. The index of any structure within the library can be easily determined by simply determining its minimum DFS code.

The topologies in the library are not independent; two unique 5-stem XIOS graphs, for instance, may share a common 4-stem subgraph as shown in Fig 2. In this situation, we say that the 4-stem subgraph is the parent of both 5-stem graphs because they each have had one stem added to the parent subgraph (Fig 2). When comparing topological motifs, subgraphs that share a parent are clearly more similar than subgraphs that only share a grandparent or great-grandparent. The topological motif library includes all the parent and child relationships between the enumerated graphs in order to allow for partial matching.

Determining RNA fingerprints using random sampling

A XIOS graph corresponding to a single structure can be characterized by the set of fixed-size subgraphs it contains. This set of constituent subgraphs is the RNA fingerprint (Fig 3), which can be thought of as a subgraph spectrum that is characteristic of a specific topology. Currently we use a library comprising all 7-stem and smaller subgraphs; this number has been chosen to cover both large and small biological structures, without requiring excessive computation. For even a relatively small graph, for instance a graph with 25 to 30 vertices, exhaustively enumerating the complete set of 7-vertex subgraphs within it can be time consuming. The subgraph sampling approach we describe here allows the determination of the fingerprint in reasonable time on parallel hardware. Briefly, given a XIOS graph, we randomly sample a fixed number, currently seven, of connected vertices from the graph (Table 2). Sampling continues until a suitable termination condition is met, typically when all observed subgraphs have been

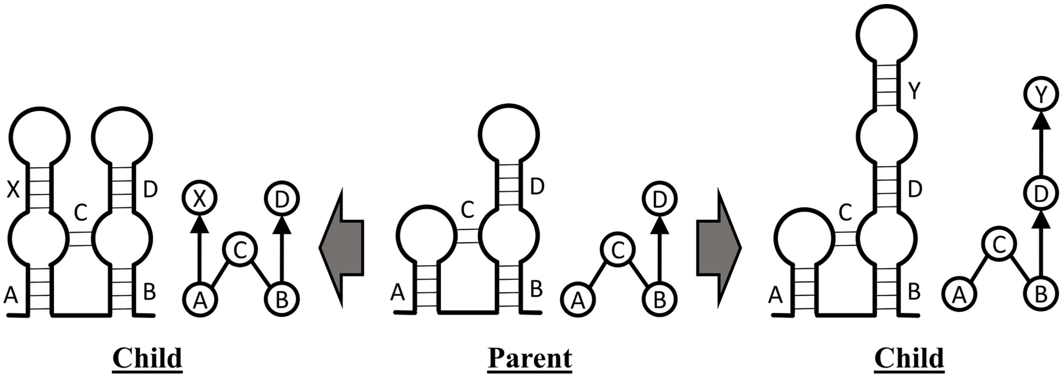

Fig 2. Parent-Child relationships. The parent graph is a 4-stem motif; two different child graphs are created by adding one stem to the parent graph.

doi:10.1371/journal.pone.0164726.g002

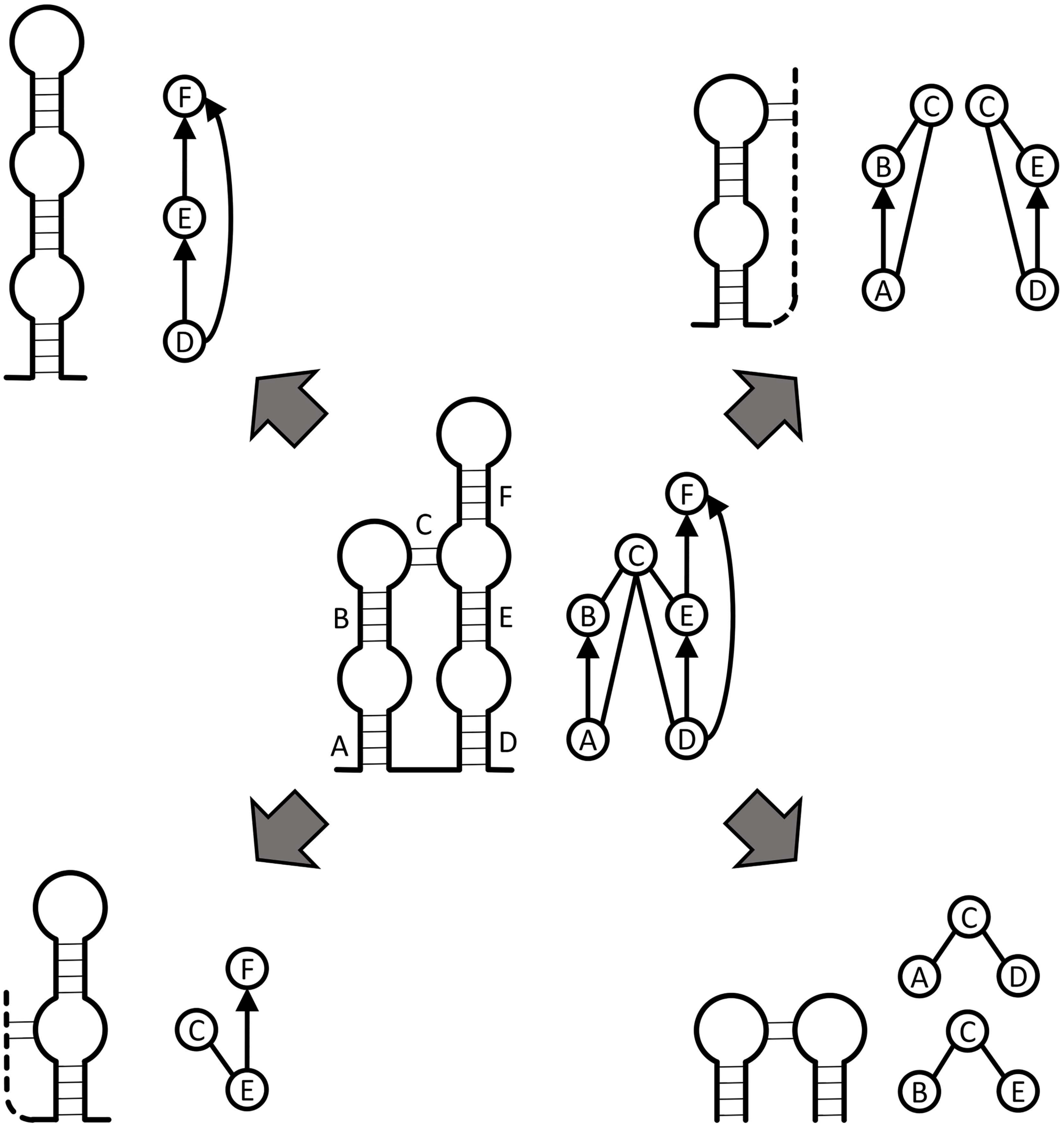

Fig 3. Example of an RNA fingerprint. All 3-vertex motifs (corners) in a 6-vertex RNA graph (center) are shown. The thick solid lines represent RNA chain, the thin solid lines represent base pairs, and the dotted lines represent RNA sequences whose connectivity is not completely specified.

doi:10.1371/journal.pone.0164726.g003

independently sampled 10 times. In each iteration, one subgraph is sampled and uniquely identified by its minimum DFS code, which is used as a reference to identify the subgraph in the RNA structural motif library. The complete fingerprints of 151 RNA structures computed by an exhaustive method (not shown) have been used to validate the correctness of the RNA fingerprints computed by random sampling (Fig 4).

RNA fingerprints identify topologically similar RNA structures

The set of subgraph motifs sampled in a query graph is its ***simple fingerprint***. We define the ***extended fingerprint*** as the simple fingerprint plus all of the ancestral subgraphs (i.e., parent, grandparent, etc., see Fig 2) of the simple fingerprint motifs. In this section we use both the simple fingerprint and the extended fingerprint to identify RNAs with similar topologies. The average numbers of motifs in simple and extended fingerprints are shown in S2 Fig.

Table 2. Subgraph random sampling pseudocode.

```
Algorithm: Subgraph Random Sampling
Input: Query graph $G = (V, E)$, subgraph size $n$
Output: Sampled subgraph $S = (V_s, E_s)$
Select a random vertex $v_i \in V$
Initialize the set of vertices $V_s = \{v_i\}$
Initialize the set of edges $E_s = \{\}$
WHILE $/ V_s / < n$ DO
  Identify $N_{Vs}$, the vertices adjacent to $V_s$
  IF $N_{Vs} = = \{\}$ DO
    BREAK
  ELSE DO
    Select a random vertex $v_j$ from $N_{Vs}$
    Update $E_s = E_s \cup \{ (v_i, v_j) \} \forall v_i, v_j \in V_s$
    Update $V_s = V_s \cup \{v_j\}$
  END IF
END WHILE
RETURN subgraph $S = (V_s, E_s)$
```

doi:10.1371/journal.pone.0164726.t002

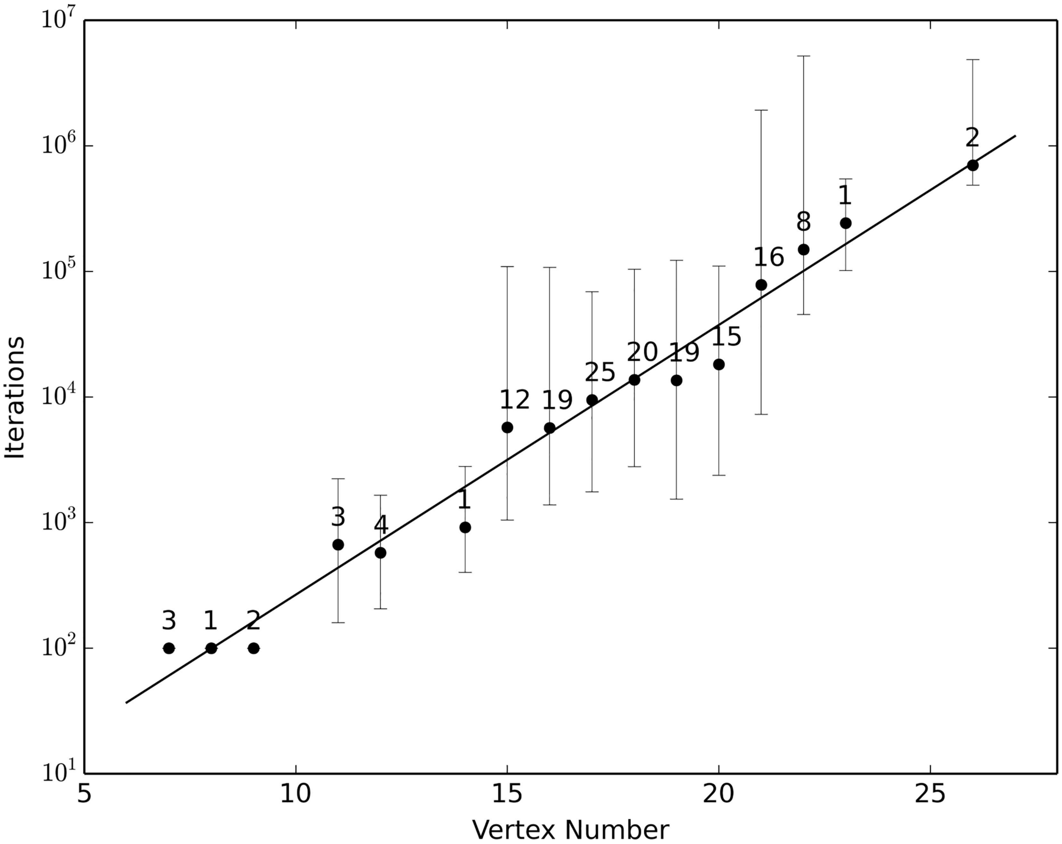

Fig 4. Scaling of sampling with graph size. Fingerprints for 151 RNA graphs in the curated set were determined multiple times (10 times per RNA graph) by random sampling. Numbers above the dots indicate the number of different graphs with the same size (vertex number); each dot represents the average number of iterations needed to determine the complete fingerprint for this specific size group, with bars showing the maximum and minimum iterations as well.

doi:10.1371/journal.pone.0164726.g004

Consider the simple or extended fingerprints, X and Y, of RNA R_X and RNA R_Y; X = {x_1, x_2, x_3,. . ., x_m} and Y = {y_1,y_2, y_3,. . ., y_n} where x_1, x_2, x_3, . . ., x_m and y_1,y_2, y_3, . . ., y_n are the subgraph motifs found in RNAs R_X and R_Y. We have evaluated five similarity functions (Table 3) for their ability to identify topologically similar structures.

Fig 5 shows the classification performance of the different similarity functions as measured by Receiver Operating Characteristic (ROC) curves [55]. Jaccard Similarity works best in the classification of RNA structures, with an area under the ROC curve (AUC) greater than 0.95 for the extended fingerprint. The increase in AUC from 0.870 for the simple fingerprint to 0.952 for the extended fingerprint using Jaccard Similarity indicates that the inclusion of parent subgraphs substantially improves the detection of topologically similar structures. The classification performance of Jaccard Similarity using the extended fingerprint on different RNA classes is around 0.95 for all groups except for 16S rRNA and group II introns (S2 Table). Fig 6 shows the ability of the extended-Jaccard similarity to effectively classify the test structures into functional groups. As can be seen in the upper triangle of Fig 6, the level of sequence similarity is very low between these structures and would be insufficient for correct clustering (not shown). The 23S rRNAs form a single group, and also share some similarity with 16S rRNAs, which may be explained by the topological similarity of the two subunits of rRNA [47]. The 5S rRNAs form two separate groups, one with archaeal and eukaryotic nuclear structures, and the other with bacterial structures. Self-splicing introns, especially the Group II Introns, share a high topological similarity with the 23S and 16S rRNAs. The accuracy of the classification confirms that our topological approach can identify topologically similar RNAs, and potentially functionally similar RNAs, as well. In addition, a neighbor-joining tree [56] (Fig 6, S3 and S4 Figs), using the extended-Jaccard similarity, correctly groups almost all the curated RNA families into the correct categories, with only one Group I Intron falling onto a branch outside of its curated group (Fig 6, tree on the right side).

Similarity of incomplete graphs can be detected using RNA fingerprints

In most cases, topological comparisons must be based on predicted structures, because three-dimensional structures or high-quality comparative structures are usually unavailable. Although structures with pseudoknots can be predicted [41, 57–60], such predicted structures will typically be inaccurate or incomplete. It is highly desirable that a similarity function be able to correctly identify similar RNAs, even when their structures are incomplete. To test the effects of graph incompleteness on the extended-fingerprint Jaccard Similarity function, incomplete RNA graphs were generated by randomly removing a percentage (10%, 30%, 50%, 60%, and 70%, respectively) of the vertices (stems) in the curated structures (Fig 5F). The extended-fingerprint Jaccard Similarity can identify similar structures when only 70% of the original stems are present (AUC = 0.810), and performs better than random even when only 30% of the stems remain. In addition, since pseudoknots are important structural motifs in RNAs, for the 160 RNA structures that have pseudoknots, we generated incomplete RNA

Table 3. RNA fingerprint similarity functions. X and Y are fingerprints of the two structures being compared.

Similarity Function	Definition
Intersection	$\mathbf{S_B(X,Y) = \|X \cap Y\|}$
Cosine [50]	$S_C(X,Y) = \frac{\|X\cap Y\|}{\sqrt{\|X\|\|Y\|}}$
Dice [51, 52]	$S_D(X,Y) = \frac{2\|X\cap Y\|}{\|X\|+\|Y\|}$
Hamming [53]	$\mathbf{S_H(X,Y) = \|(X \cap Y) \cup (X \cup Y)^C\|}$
Jaccard [54]	$S_J(X,Y) = \frac{\|X\cap Y\|}{\|X\cup Y\|}$

doi:10.1371/journal.pone.0164726.t003

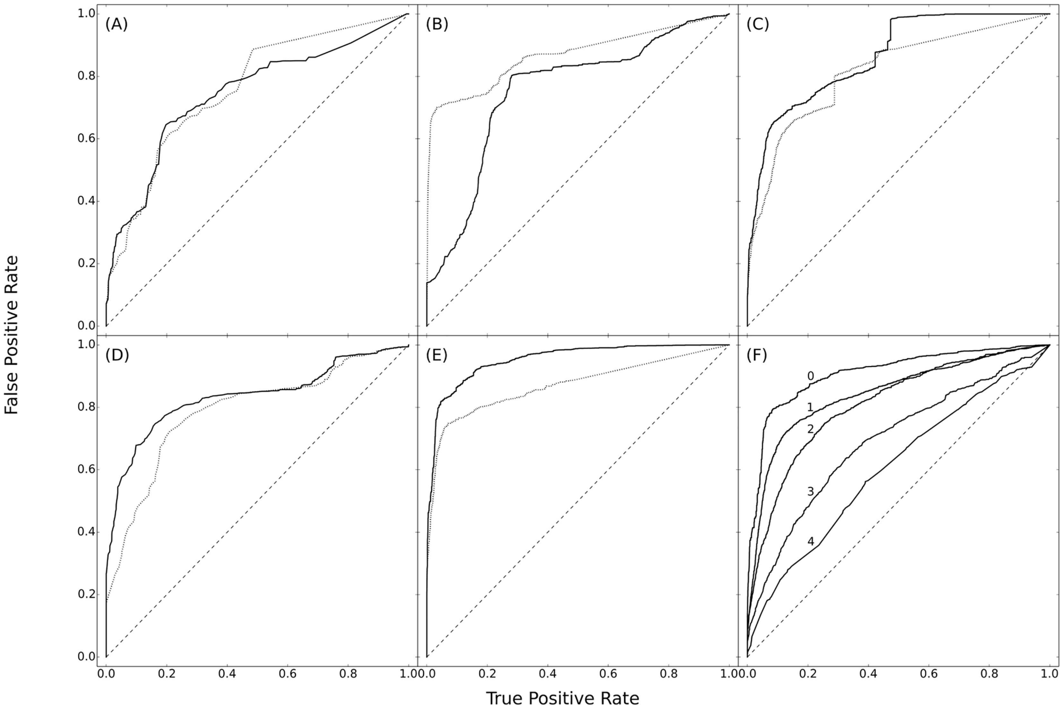

Fig 5. Classification performance of similarity functions. Pairwise similarities were calculated, using the indicated similarity functions, for all RNAs in the curated dataset and ranked from high to low. A pair of RNAs from the same curated family is considered a positive match; otherwise they are considered to be a negative match. In all panels, the dashed line indicates the simple fingerprint, and the solid line the extended fingerprint. The AUC for the simple and extended fingerprints, respectively, are indicated in parentheses, below. (A) Intersection Similarity (AUC simple, 0.759; extended, 0.746), (B) Cosine Similarity (0.867; 0.753), (C) Dice Similarity (0.821; 0.864), (D) Hamming Similarity (0.789; 0.834), and (E) Jaccard Similarity (0.870; 0.952). (F) Classification after random removal of vertices from RNA graphs. All RNAs (except for tRNA and 5S rRNA which are too small for 70% stem removal) are included. The five lines show ROC curves with differing fractions of stems removed (AUC in parentheses): (0) no stem removal (AUC = 0.909), (1) 10% stem removal (0.844), (2) 30% stem removal (0.810), (3) 50% stem removal (0.691), and (4) 70% stem removal (0.605).

doi:10.1371/journal.pone.0164726.g005

graphs by first removing all the pseudoknot-forming vertices (stems), and continuing removing random vertices until 30% of vertices were removed. The extended-fingerprint Jaccard Similarity correctly identifies similar structures with pseudoknots removed (AUC = 0.915, data not shown).

Fingerprint similarity is not an artifact of graph size

The structures within each curated family generally have very similar numbers of stems. Indeed, one can classify the structures into the correct groups using graph size alone (not shown). It is essential, therefore to consider whether the results in Figs 5 and 6 are merely due to the similarity in sizes. In order to test the effect of size, we have created a test data set in which the graphs have been expanded to the same size (number of vertices) by randomly adding additional vertices and edges to the graphs. In order to ensure that these expanded graphs are typical of real biological structures we use a procedure in which we sample substructures from the set of curated structures, and add them to the curated graphs. In order to do this, we created a database (decoy database) of the 2 to 5 stem motifs found in the curated structures, and randomly added these subgraphs to the curated structures according their frequency in the entire curated set (which should reflect the biological background distribution).

We selected a set of 177 RNA graphs containing up to 25 vertices from the curated data set (S1 Table), and created an expanded set by embedding subgraphs, randomly selected according

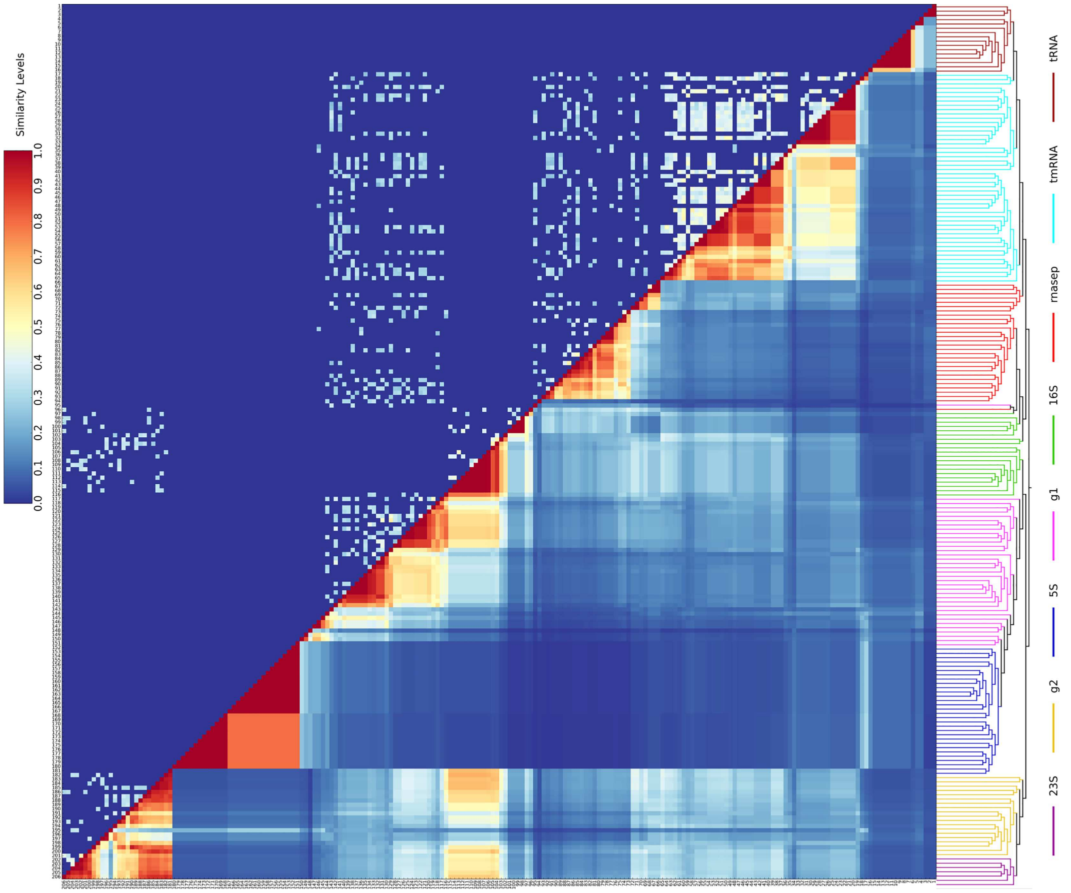

Fig 6. Extended-fingerprint Jaccard similarity between biological RNAs. Upper triangle. Sequence identity. Lower triangle. Extended-fingerprint Jaccard Similarity of all the curated RNA structures (see S3 Fig and S5 Table for IDs). Sequence identity is shown in color, ranging from 0 (blue) to 1 (red) at steps of 0.1. A neighbor-joining dendrogram calculated according to the extended-fingerprint Jaccard similarity is shown on the right side of the heat map.

doi:10.1371/journal.pone.0164726.g006

to probability of occurrence, from the decoy database into these RNA graphs until each RNA graph contained 30 vertices. As a control, we also created a decoy set of graphs with 30 vertices, by random embedding of subgraphs from the decoy database only, *i.e.*, graphs with no information from real biological structures except the frequency of occurrence of subgraphs in the known structures. Both the expanded and the decoy graph sets should be completely free of size effects since they all have exactly the same number of stems. The two sets were mixed and graphs compared using the Extended Fingerprint Jaccard Similarity. There is only a minor decrease in performance (S3 Table, Extended Fingerprint Jaccard Similarity: AUC = 0.840) when compared to the results obtained from the classification of the original dataset (Fig 5, Extended Fingerprint Jaccard Similarity: AUC = 0.952). As expected, the decoy set of graphs have AUC values close to 0.5, indicating that the decoy structures are random with respect to each other.

Runtime analysis

Determination of whether a query RNA graph contains a subgraph isomorphic to a specific graph in the structural motif library, is an NP-complete problem [48]. The brute-force comparison requires comparing the query RNA graph with every graph in the library, and its computational complexity is $O(nm^m)$, where n is the number of graphs in the library (55,728), and m is the number of edges in the query graph. The subgraph random sampling algorithm can be parallelized by simultaneously running independent instances on multiple processors. The algorithm identifies the fingerprint of all 206 curated RNA graphs in a reasonable time, especially when it is run on multiple cores (S5 Fig). The average runtime for calculating the fingerprint of RNAs in each functional family is shown in S4 Table.

Discussion

A great deal of work has focused on identifying similar RNAs based on the comparison of RNA secondary structures. This is readily accomplished using approaches such as tree edit distance [22, 61] or string related measures such as those used in RNAshapes [7]. Other approaches include the information of sequence alignment and folding of RNA sequences, for example, Saito *et al.* developed an algorithm that clusters RNAs by all possible sequence alignments, and all possible secondary structures computed from dynamic programming and partition function calculations [62–64]. This approach correctly discriminated short RNA sequences (around 100 bases) from different families. Unfortunately, secondary structures, and in particular minimum free-energy predicted structures based on dynamic programming approaches, do not predict pseudoknots [65–67], which are important in biological structures. Even if predicted pseudoknots are available [68–70], it is not simple to add them to tree or string based methods because of their non-nested nature. In addition, structure matching methods based on dynamic programming have the additional problem of determining gap penalties; it is not at all clear how to weight insertions and deletions in RNA structures.

Statistical algorithms, such as kernel methods, have been developed to classify RNA sequences and structures. Kin *et al.* developed a marginalized kernel to measure RNA sequence similarity [71], and this kernel was later implemented by Karklin *et al.* to measure the similarity of RNA secondary structures represented by dual graphs [29, 72]; Liu *et al.* developed a fuzzy kernel to cluster the secondary structure ensemble generated from a single sequence [73]. The GraphClust pipeline developed by Heyne *et al.* encodes RNA sequence-structure information as graphs and measures RNA graph similarities using a decomposition kernel and computing the summed similarity of pairs of neighborhood subgraphs [27]. However, no pseudoknotted structures were included in these approaches. Sakakibara *et al.* developed a stem kernel that could discriminate between functional RNA sequences and randomly shuffled sequences using structural features including pseudoknots [74]; however, no result was shown in which the stem kernel could discriminate between sequences from different functional RNA groups, in addition, the randomly shuffled sequences they generated only retain nucleotide composition, while preserving dinucleotide composition is known to be important in generating randomized negative controls for predicted RNA structures [75, 76]. In summary, none of these approaches have demonstrated that they can succeed on the difficult test case presented here: classifying a diverse set of functional families, with diverse sizes, containing pseudoknots, and with little sequence similarity. Topological methods have the dual advantage of easily representing pseudoknots and not requiring an insertion/deletion penalty. In the RNA-As-Graphs procedure [29, 31, 77, 78], RNA topologies are represented either as “tree graphs” (without pseudoknots) or “dual graphs” (with pseudoknots), and the topological properties of an RNA graph are summarized using the eigenvalue of its Laplacian matrix (constructed from the adjacency and degree matrices of the graph). They have developed a database, with all mathematically possible RNA graphs enumerated, including “existing graphs” (RNA structures experimentally solved or obtained from comparative analysis) and “missing graphs” (mathematically possible RNA structures that have not yet been experimentally observed). Using “existing graphs” as training data, "missing graphs" in the database were classified as either "RNA-like" or "non-RNA-like" by applying regression analysis on Laplacian eigenvalue spectra [34]. These approaches, which target the identification of novel RNA topologies, however, are not sufficient for matching specific RNA functional families.

Graph matching is a computationally intensive process that scales exponentially with the size of graph (in general, graph matching is an NP-hard process) [79]. Functional RNA molecules can include dozens of stems/loops, especially with the current advance in high-throughput technologies, and long non-coding RNAs including hundreds of stems/loops are not uncommon [80]. The RAG database, however, only includes dual graphs up to 9 vertices and tree graphs up to 10 vertices [77], which can cover RNA topologies only up to about 200nt, while the XIOS graph approach can handle RNA topologies with 70 vertices and 2000nt (S1 and S5 Tables). Moreover, in a follow-up study, the discrimination between structures predicted to be RNA-like (naturally existing) and non-RNA-like was not impressive; out of 42 newly discovered RNA topologies, only 24 of them had been predicted as RNA-like, while 18 of them had been predicted to be non-RNA like [77]. In addition, the numerical descriptors used in the RAG procedures have never been shown to be able to group RNAs into structural/functional classes.

The XIOS graph is a topological graph approach [36] that specifically distinguishes pseudoknots as a distinct type of edge. In addition to incorporating pseudoknots (O edge, Overlapping), one of the most important characteristics in RNA structure, the XIOS approach also includes embedding (I edge, Included) and juxtaposition (S edge, Serial), which are the two of the RNA structural principles in the RNAshapes framework. The increased number of edge-types in XIOS improves one's ability to match graphs, for example using gSpan; however, the time required to find the maximal common subgraph in two moderately large RNA graphs, for instance with twenty to thirty stems in each graph, is prohibitive using exhaustive approaches such as gSpan. Using the XIOS approach, we can easily enumerate a complete set of biologically possible RNA graphs, permitting the construction of a complete dictionary of all graphs that may occur in a RNA molecule up to a specified size. This allows us to characterize any RNA topology in terms of the spectrum of subgraphs it contains, its RNA fingerprint, and to identify topologically similar RNA structures based on their fingerprints. This approach is successful with known RNA families, and is relatively insensitive to both the completeness of the RNA graph, and the presence of extraneous added vertices in the graph. Similarities between RNA structures in the same family are still detectable when the graphs are expanded to the same size, indicating that the ability to identify topologically similar structures is not simply an artifact of the similar sizes of RNAs within known families. These characteristics of RNA fingerprint matching are highly important in real-world settings where comparisons are made between predicted structures in which only 60–80% of the true stems may be correctly predicted [58, 81], and a substantial number of mispredicted stems may be present. As mentioned before, no previously reported RNA structure comparison method has shown that it can accurately identify/classify RNAs according to topological similarity using the particularly difficult set of pseudoknot containing graphs used here.

Exhaustively enumerating the set of subgraphs present in a XIOS graph is time consuming because each subgraph in the entire motif database must be separately tested against the query to determine whether there is a match. Because the dictionary of subgraphs is large (55,728 graphs with seven or fewer stems), a brute force approach is slow. In this work we suggest a sampling approach to enumerating the subgraph spectrum. The computational complexity of motif sampling depends on both the size and structure of the query graph, and on the number of vertices sampled in each iteration. As most of RNA XIOS graphs are highly connected, an increase in graph size can result in a large increase in the time required to completely sample the fingerprint. Fortunately, the motif sampling is completely parallelizable; any number of processors can independently sample subgraphs from the query, and the time required per query graph is modest. Furthermore, our results suggest that a complete fingerprint may not be necessary; that even incomplete fingerprints (such as fingerprints derived from structures where part of the structure has been removed) are sufficient to identify topologically similar structures. The question of whether absolutely every subgraph has been detected, which a sampling strategy cannot guarantee, is therefore somewhat moot.

Experimental determination of RNA structures by X-ray crystallography or NMR is difficult, and a relatively small number of complete structures are available. Instead, structures are often predicted using a combination of biochemical information (chemical modification, nuclease sensitivity, and mutational sensitivity), secondary structure prediction, and phylogenetic conservation (covariance). This results in "known" structures that are incomplete (missing important stems) or inaccurate (containing stems that do not exist, or are unimportant in the function of the RNA). It is therefore important that the structural/topological comparison be robust with respect to incompleteness or error in the structures, a salient characteristic of the RNA fingerprint comparison we describe here. The extended-fingerprint Jaccard Similarity correctly identifies topologically similar RNAs across a broad range of sizes, and biological functions, but its potential application is far more general. RNA structure prediction is commonly judged to be 60 to 80 percent accurate [59, 68, 82]. The ability of the RNA fingerprint to correctly identify/classify structural topologies even when 30% or more of the true stems are removed (Fig 5F), suggests that this approach can be applied to broadly search for topologically similar structures based on structures predicted from sequence (work in progress). Currently, GraphClust [27] is probably the most widely used program for comparing and clustering RNAs according to sequence and structure without pseudoknots. We believe that our approach, as a coarse-grained method including pseudoknots, complements that of GraphClust.

Supporting Information

S1 Fig. XIOS RNA graph representation of a Hepatitis D Virus (HDV) ribozyme RNA.
(PDF)

S2 Fig. Numbers of motifs in Simple and Extended Fingerprints.
(PDF)

S3 Fig. Heatmap dendrogram.
(PDF)

S4 Fig. Neighbor-joining tree showing the classification using Extended Jaccard Similarity.
(PDF)

S5 Fig. Runtime analysis of the subgraph random sampling algorithm.
(PDF)

S1 Table. Curated RNA structures.
(PDF)

S2 Table. Classification performance of Extended Fingerprint Jaccard Similarity for 8 curated families.
(PDF)

S3 Table. Classification performance for expanded graphs using different similarity functions.
(PDF)

S4 Table. Run time analysis.
(PDF)

S5 Table. Complete list of curated RNA structures used in this study.
(XLSX)

Author Contributions

Conceptualization: JH KL MG.

Data curation: JH KL MG.

Formal analysis: JH KL.

Funding acquisition: MG.

Investigation: JH KL MG.

Methodology: JH KL MG.

Project administration: MG.

Resources: JH MG.

Software: JH KL MG.

Supervision: MG.

Validation: JH.

Visualization: JH.

Writing – original draft: JH MG.

Writing – review & editing: JH MG.

References

1. Barrandon C, Spiluttini B, Bensaude O. Non-coding RNAs regulating the transcriptional machinery. Biol Cell. 2008; 100(2):83–95. doi: 10.1042/BC20070090 PMID: 18199047.
2. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. 1983; 35(3 Pt 2):849–57. PMID: 6197186.
3. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell. 1982; 31 (1):147–57. PMID: 6297745.
4. Pang K, Frith M, Mattick J. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 2006; 22(1):1–5. doi: 10.1016/j.tig.2005.10.003 PMID: 16290135
5. Pace NR, Brown JW. Evolutionary perspective on the structure and function of ribonuclease P, a ribozyme. J Bacteriol. 1995; 177(8):1919–28. PMID: 7536728; PubMed Central PMCID: PMCPMC176831.
6. Johnsson P, Lipovich L, Grandér D, Morris K. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta. 2014; 1840(3):1063–71. doi: 10.1016/j.bbagen.2013.10.035 PMID: 24184936
7. Giegerich R, Voss B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Res. 2004; 32 (16):4843–51. doi: 10.1093/nar/gkh779 PMID: 15371549; PubMed Central PMCID: PMCPMC519098.
8. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 2006; 22(4):500–3. doi: 10.1093/bioinformatics/btk010 PMID: 16357029.
9. Ellis JC, Brown JW. The RNase P family. RNA Biol. 2009; 6(4):362–9. PMID: 19738420.

10. Hedberg A, Johansen SD. Nuclear group I introns in self-splicing and beyond. Mob DNA. 2013; 4 (1):17. doi: 10.1186/1759-8753-4-17 PMID: 23738941; PubMed Central PMCID: PMCPMC3679873.

11. Lehmann K, Schmidt U. Group II introns: structure and catalytic versatility of large natural ribozymes. Crit Rev Biochem Mol Biol. 2003; 38(3):249–303. doi: 10.1080/713609236 PMID: 12870716.

12. Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions. PLoS Biol. 2005; 3 (6):956–9. e21310.1371/journal.pbio.0030213. PMID: WOS:000229992900004.

13. Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. Crystal structure of a self-splicing group I intron with both exons. Nature. 2004; 430(6995):45–50. doi: 10.1038/nature02642 PMID: 15175762.

14. Powers T, Noller HF. A functional pseudoknot in 16S ribosomal RNA. EMBO J. 1991; 10(8):2203–14. PMID: 1712293; PubMed Central PMCID: PMCPMC452909.

15. Egli M, Sarkhel S, Minasov G, Rich A. Structure and Function of the Ribosomal Frameshifting Pseudoknot RNA from Beet Western Yellow Virus. Helvetica Chimica Acta. 2003; 86(5):1709–27.

16. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005; 308(5725):1149–54. doi: 10.1126/science.1108625 PMID: 15790807.

17. Waterman M. Secondary structure of single-stranded nucleic acids. Adv Math. 1978; 1:167–212 (suppl.).

18. Shapiro BA. An algorithm for comparing multiple RNA secondary structures. Comput Appl Biosci. 1988; 4(3):387–93. PMID: 2458170.

19. Shapiro BA, Zhang KZ. Comparing multiple RNA secondary structures using tree comparisons. Comput Appl Biosci. 1990; 6(4):309–18. PMID: 1701685.

20. Margalit H, Shapiro BA, Oppenheim AB, Maizel JV. Detection of common motifs in RNA secondary structures. Nucleic Acids Res. 1989; 17(12):4829–45. PMID: 2473442; PubMed Central PMCID: PMCPMC318035.

21. Hofacker IL FW, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem. 1994; 125:167–88.

22. Fontana W, Konings DA, Stadler PF, Schuster P. Statistics of RNA secondary structures. Biopolymers. 1993; 33(9):1389–404. doi: 10.1002/bip.360330909 PMID: 7691201.

23. Shu W, Bo X, Zheng Z, Wang S. A novel representation of RNA secondary structure based on element-contact graphs. BMC Bioinformatics. 2008; 9:188. doi: 10.1186/1471-2105-9-188 PMID: 18402706; PubMed Central PMCID: PMCPMC2373570.

24. Benedetti G, Morosetti S. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. Biophys Chem. 1996; 59(1–2):179–84. PMID: 8867337.

25. Churkin A, Gabdank I, Barash D. On topological indices for small RNA graphs. Comput Biol Chem. 2012; 41:35–40. doi: 10.1016/j.compbiolchem.2012.10.004 PMID: 23147564.

26. Janssen S, Reeder J, Giegerich R. Shape based indexing for faster search of RNA family databases. BMC Bioinformatics. 2008; 9:131. doi: 10.1186/1471-2105-9-131 PMID: 18312625; PubMed Central PMCID: PMCPMC2277397.

27. Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics. 2012; 28(12):i224–i32. doi: 10.1093/bioinformatics/bts224 PMID: 22689765; PubMed Central PMCID: PMCPMC3371856.

28. Costa F, Grave KD, editors. Fast neighborhood subgraph pairwise distance kernel. Proceedings of the 27th International Conference on Machine Learning (ICML-10) 2010; Haifa, Israel: Omnipress.

29. Gan HH, Pasquali S, Schlick T. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucleic Acids Res. 2003; 31(11):2926–43. PMID: 12771219; PubMed Central PMCID: PMCPMC156709.

30. Laing C, Schlick T. Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol. 2011; 21(3):306–18. doi: 10.1016/j.sbi.2011.03.015 PMID: 21514143; PubMed Central PMCID: PMCPMC3112238.

31. Gan HH, Fera D, Zorn J, Shiffeldrim N, Tang M, Laserson U, et al. RAG: RNA-As-Graphs database—concepts, analysis, and features. Bioinformatics. 2004; 20(8):1285–91. doi: 10.1093/bioinformatics/bth084 PMID: 14962931.

32. Fera D, Kim N, Shiffeldrim N, Zorn J, Laserson U, Gan HH, et al. RAG: RNA-As-Graphs web resource. BMC Bioinformatics. 2004; 5:88. doi: 10.1186/1471-2105-5-88 PMID: 15238163; PubMed Central PMCID: PMCPMC471545.

33. Barash D, Comaniciu D. A Common Viewpoint on Broad Kernel Filtering and Nonlinear Diffusion. Lect Notes Comput Sc. 2003; 2695 683–98.

34. Barash D. Spectral Decomposition for the Search and Analysis of RNA Secondary Structure. J Comput Biol. 2004; 11(6):1169–74. doi: 10.1089/cmb.2004.11.1169 PMID: 15662204

35. Barash D. Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation. Bioinformatics. 2004; 20(12):1861–9. doi: 10.1093/bioinformatics/bth157 PMID: 14988109.

36. Li K, Rahman R, Gupta A, Siddavatam P, Gribskov M. Pattern matching in RNA structures. Bioinformatics Research and Applications. 2008; 4983:317–30. PMID: WOS:000255941000029.

37. Churkin A, Barash D. An efficient method for the prediction of deleterious multiple-point mutations in the secondary structure of RNAs using suboptimal folding solutions. BMC Bioinformatics. 2008; 9:222. doi: 10.1186/1471-2105-9-222 PMID: 18445289; PubMed Central PMCID: PMCPMC2386494.

38. Waldispühl J, Devadas S, Berger B, Clote P. Efficient algorithms for probing the RNA mutation landscape. PLoS Comput Biol. 2008; 4(8):e1000124. doi: 10.1371/journal.pcbi.1000124 PMID: 18688270; PubMed Central PMCID: PMCPMC2475669.

39. Avihoo A, Churkin A, Barash D. RNAexinv: An extended inverse RNA folding from shape and physical attributes to sequences. BMC Bioinformatics. 2011; 12:319. doi: 10.1186/1471-2105-12-319 PMID: 21813013; PubMed Central PMCID: PMCPMC3176266.

40. Bindewald E, Afonin K, Jaeger L, Shapiro BA. Multistrand RNA secondary structure prediction and nanostructure design including pseudoknots. ACS Nano. 2011; 5(12):9542–51. doi: 10.1021/nn202666w PMID: 22067111; PubMed Central PMCID: PMCPMC3263976.

41. Gupta A, Rahman R, Li K, Gribskov M. Identifying complete RNA structural ensembles including pseudoknots. RNA Biol. 2012; 9(2):187–99. doi: 10.4161/rna.18386 PMID: 22418849.

42. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 2011; 39(Database issue):D392–401. doi: 10.1093/nar/gkq1021 PMID: 21036868; PubMed Central PMCID: PMCPMC3013649.

43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–42. PMID: 10592235; PubMed Central PMCID: PMCPMC102472.

44. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res. 1994; 22 (11):2079–88. PMID: 8029015; PubMed Central PMCID: PMCPMC308124.

45. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015; 43(Database issue):D130–7. doi: 10.1093/nar/gku1063 PMID: 25392425; PubMed Central PMCID: PMCPMC4383904.

46. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics. 2008; 9:340. doi: 10.1186/1471-2105-9-340 PMID: 18700982; PubMed Central PMCID: PMCPMC2536673.

47. Petrov AS, Bernier CR, Hershkovits E, Xue Y, Waterbury CC, Hsiao C, et al. Secondary structure and domain architecture of the 23S and 5S rRNAs. Nucleic Acids Res. 2013; 41(15):7522–35. doi: 10.1093/nar/gkt513 PMID: 23771137; PubMed Central PMCID: PMCPMC3753638.

48. Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-completeness. New York: W. H. Freeman; 1979.

49. Yan XF, Han JW. gSpan: Graph-based substructure pattern mining. Proceedings of the 2002 IEEE International Conference on Data Mining. 2002:721–4. PMID: WOS:000180274000107.

50. Singhal A. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001; 24(4):35–42.

51. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945; 26 (3):297–302.

52. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab: København, I kommission hos E. Munksgaard; 1948. 1–34 p.

53. Hamming R. Error detecting and error correcting codes. Bell Syst Tech J. 1950; 29(2):147–60.

54. Jaccard P. The distribution of the flora in the alpine zone. New Phytologist. 1912; 11(2):37–50.

55. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem. 1996; 20(1):25–33. PMID: 16718863.

56. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406–25. PMID: 3447015.

57. Reeder J, Steffen P, Giegerich R. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic Acids Res. 2007; 35(Web Server issue):W320–4. doi: 10.1093/nar/gkm258 PMID: 17478505; PubMed Central PMCID: PMCPMC1933184.

58. Ren J, Rastegari B, Condon A, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA. 2005; 11(10):1494–504. doi: 10.1261/rna.7284905 PMID: 16199760; PubMed Central PMCID: PMCPMC1370833.

59. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011; 27(13):i85–93. doi: 10.1093/bioinformatics/btr215 PMID: 21685106; PubMed Central PMCID: PMCPMC3117384.

60. Janssen S, Giegerich R. The RNA shapes studio. Bioinformatics. 2015; 31(3):423–5. doi: 10.1093/bioinformatics/btu649 PMID: 25273103.

61. Chen S, Zhang K. An improved algorithm for tree edit distance with applications for RNA secondary structure comparison. J Comb Optim. 2014; 27:778–97.

62. Saito Y, Sato K, Sakakibara Y. Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures. BMC Bioinformatics. 2011; 12 Suppl 1:S48. doi: 10.1186/1471-2105-12-S1-S48 PMID: 21342580; PubMed Central PMCID: PMCPMC3044305.

63. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981; 147 (1):195–7. PMID: 7265238.

64. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29(6–7):1105–19. doi: 10.1002/bip.360290621 PMID: 1695107.

65. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008; 453:3–31. doi: 10.1007/978-1-60327-429-6_1 PMID: 18712296.

66. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011; 6:26. doi: 10.1186/1748-7188-6-26 PMID: 22115189; PubMed Central PMCID: PMCPMC3319429.

67. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. 2004; 101(19):7287–92. doi: 10.1073/pnas.0401799101 PMID: 15123812; PubMed Central PMCID: PMCPMC409911.

68. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. 2010; 16(10):1870–80. doi: 10.1261/rna.2125310 PMID: 20699301; PubMed Central PMCID: PMCPMC2941096.

69. Beyer W. RNA Secondary Structure Prediction including Pseudoknots: University of Vienna; 2010.

70. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol. 1999; 285(5):2053–68. doi: 10.1006/jmbi.1998.2436 PMID: 9925784.

71. Kin T, Tsuda K, Asai K. Marginalized kernels for RNA sequence data analysis. Genome Inform. 2002; 13:112–22. PMID: 14571380.

72. Karklin Y, Meraz RF, Holbrook SR. Classification of non-coding RNA using graph representations of secondary structure. Pac Symp Biocomput. 2005:4–15. PMID: 15759609.

73. Liu Q, Zhang Y, Xu Y, Ye X. Fuzzy kernel clustering of RNA secondary structure ensemble using a novel similarity metric. J Biomol Struct Dyn. 2008; 25(6):685–96. doi: 10.1080/07391102.2008.10507214 PMID: 18399702.

74. Sakakibara Y, Popendorf K, Ogawa N, Asai K, Sato K. Stem kernels for RNA sequence analyses. J Bioinform Comput Biol. 2007; 5(5):1103–22. PMID: 17933013.

75. Schattner P. Searching for RNA genes using base-composition statistics. Nucleic Acids Res. 2002; 30 (9):2076–82. PMID: 11972348; PubMed Central PMCID: PMCPMC113829.

76. Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 2008; 9:192. doi: 10.1186/1471-2105-9-192 PMID: 18405375; PubMed Central PMCID: PMCPMC2375906.

77. Izzo JA, Kim N, Elmetwaly S, Schlick T. RAG: an update to the RNA-As-Graphs resource. BMC Bioinformatics. 2011; 12:219. doi: 10.1186/1471-2105-12-219 PMID: 21627789; PubMed Central PMCID: PMCPMC3123240.

78. Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. Methods Mol Biol. 2012; 905(99–122).

79. Fan W, Li J, Ma S, Tang N, Wu Y, Wu Y. Graph Pattern Matching: From Intractable to Polynomial Time. Proceedings of the VLDB Endowment. 2010; 3(1–2):264–75.

80. Novikova IV, Hennelly SP, Sanbonmatsu KY. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res. 2012; 40(11):5034–51. doi: 10.1093/nar/gks071 PMID: 22362738; PubMed Central PMCID: PMCPMC3367176.

81. Zhang B, Yehdego DT, Johnson KL, Leung MY, Taufer M. Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MapReduce. BMC Struct Biol.

2013; 13 Suppl 1:S3. doi: 10.1186/1472-6807-13-S1-S3 PMID: 24564983; PubMed Central PMCID: PMCPMC3952952.

82. Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA. 2009; 15(10):1805–13. doi: 10.1261/rna.1643609 PMID: 19703939; PubMed Central PMCID: PMCPMC2743040.