

Article

## Clustering Rfam 10.1: Clans, Families, and Classes

Felipe A. Lessa<sup>1</sup>, Tainá Raiol<sup>2</sup>, Marcelo M. Brigido<sup>2</sup>, Daniele S. B. Martins Neto<sup>3</sup>,  
Maria Emília M. T. Walter<sup>1,\*</sup> and Peter F. Stadler<sup>4,5,6,7,8,9</sup>

<sup>1</sup> Department of Computer Science, Institute of Exact Sciences, University of Brasília, Brasília 70910-900, Brazil; E-Mail: felipe.lessa@gmail.com

<sup>2</sup> Department of Cellular Biology, Institute of Biology, University of Brasília, Brasília 70910-900, Brazil; E-Mails: tainaraiol@gmail.com (T.R.); brigido@unb.br (M.M.B.)

<sup>3</sup> Department of Mathematics, University of Brasília, Brasília 70910-900, Brazil; E-Mail: daniele@mat.unb.br

<sup>4</sup> Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; E-Mail: studla@bioinf.uni-leipzig.de

<sup>5</sup> Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, Leipzig D-04103, Germany

<sup>6</sup> Fraunhofer Institut für Zelltherapie und Immunologie–IZI Perlickstraße 1, D-04103 Leipzig, Germany

<sup>7</sup> Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, Wien A-1090, Austria

<sup>8</sup> Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

<sup>9</sup> Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

\* Author to whom correspondence should be addressed; E-Mail: mariaemilia@unb.br; Tel.: +55-61-3107-7358; Fax: +55-61-3107-7359.

Received: 5 May 2012, in revised form: 4 June 2012 / Accepted: 15 June 2012 /

Published: 5 July 2012

---

**Abstract:** The Rfam database contains information about non-coding RNAs emphasizing their secondary structures and organizing them into families of homologous RNA genes or functional RNA elements. Recently, a higher order organization of Rfam in terms of the so-called clans was proposed along with its “decimal release”. In this proposition, some of the families have been assigned to clans based on experimental and computational data in order to find related families. In the present work we investigate an alternative classification for the RNA families based on tree edit distance. The resulting clustering recovers some of the Rfam clans. The majority of clans, however, are not recovered by the structural clustering. Instead, they get dispersed into larger clusters, which correspond roughly to

well-described RNA classes such as snoRNAs, miRNAs, and CRISPRs. In conclusion, a structure-based clustering can contribute to the elucidation of the relationships among the Rfam families beyond the realm of clans and classes.

**Keywords:** Rfam; non-coding RNA; secondary structure; clans; clusters

---

## 1. Introduction

The Rfam database systematically collects sequences, alignments, consensus secondary structures, covariance models (CMs) and the corresponding annotation for RNAs with evolutionarily conserved secondary structures [1,2]. The database is constructed using a small manually curated “seed alignment” for each RNA family which is then expanded by a large-scale search for homologs in nucleotide databases. Rfam-families consist of homologous sequences that can be reasonably well aligned and that share some common function.

Rfam release 10.0 [3] introduced the concept of a *clan* as a means for describing explicit relationships between Rfam families for which homology is recognizable but sequence similarities are too faint for good alignments (such as archaeal RNase P, nuclear RNase P, and the two bacterial RNase P types a and b) or which are classified into different Rfam families because of diverse functions (for example RNase MRP RNA and the four RNase P RNA families). Rfam clans thus correspond to the definition of RNA families as used e.g., in [4].

At an even higher level of aggregation, *RNA classes* comprise families that share characteristic sequence and/or structure features, without necessarily being evolutionarily related. The best known examples of RNA classes are animal microRNAs, both classes of small nucleolar RNAs (the box C/D snoRNAs and the box H/ACA snoRNAs) and transfer RNAs. At the level of RNA classes, it is not required for class members to be related by common descent.

There is strong support for the hypothesis that all tRNAs are homologs deriving from a single clover-leaf-structured ancestor [5,6]. Additional RNA families such as mascRNA [7], menRNAs [8], or BC1 [9] are also descendants of tRNAs and hence belong to the tRNA clan. MicroRNA families, on the other hand, frequently arise *de novo* [10–12]. In fact, presence/absence patterns of microRNA families have turned out to be a valuable and nearly homoplasy-free phylogenetic marker [13]. It has been argued that novel microRNA families can easily arise in transcribed regions, considering that stem-loop structures resembling microRNA precursors frequently occur in random RNA sequences. This mechanism is most easily seen in the expansion of microRNA clusters by hairpins that are unrelated to more ancient cluster’s components [14]. Topics of innovation and expansion of microRNA families are reviewed e.g., in [15,16]. In regard to snoRNAs it is not clear to what extent families are ancestrally related. While distant homologies can be established in some cases, e.g., U87 and U88 [17], there is also some evidence for the lineage-specific innovation of snoRNA families, e.g., in birds [18, Figure S8].

Clustering of RNAs based on their sequences and/or structural characteristics is probably the simplest approach for identifying families, clans, or classes, see e.g., [19–21]. Here we are only concerned with

higher levels of aggregation beyond the level of Rfam families; thus we focus on structural similarity. A wide variety of different algorithmic schemes have been proposed to quantify (dis)similarities among known secondary structures. Since secondary structures have canonical representations in the form of ordered trees, tree alignments (RNAforester [22]) and tree editing [23,24] are the most natural and elegant means of comparison. For the specific purpose of clustering, however, it is a fundamental shortcoming of alignments that the cost functions violate the triangle inequality and hence do not form a metric on the set of labeled ordered trees [22]. Hence we employ here a tree-edit distance.

The use of direct structure comparison becomes quite limited in practical applications, because secondary structures of individual sequences are unknown in most cases. Computational prediction of secondary structures for individual sequences, on the other hand, is not sufficiently accurate. This limitation may be overcome, or at least alleviated, however, by using comparative information, see e.g., [25]. Successful applications of RNA clustering [19,21] typically use combined sequence and structure alignments based on the Sankoff algorithm [26] to combine thermodynamic rules with conservation information. For each Rfam family, a high-quality manual sequence alignment together with a matching consensus secondary structure model is available. These consensus structures can be readily used for structure-based clustering.

In this contribution we explore to what extent the manually annotated clans of related Rfam families are detectable by unsupervised clustering, whether RNA classes such as microRNA and the two snoRNA classes are recognizable, and whether there are good candidates for clans or classes of Rfam families that have remained so far undetected.

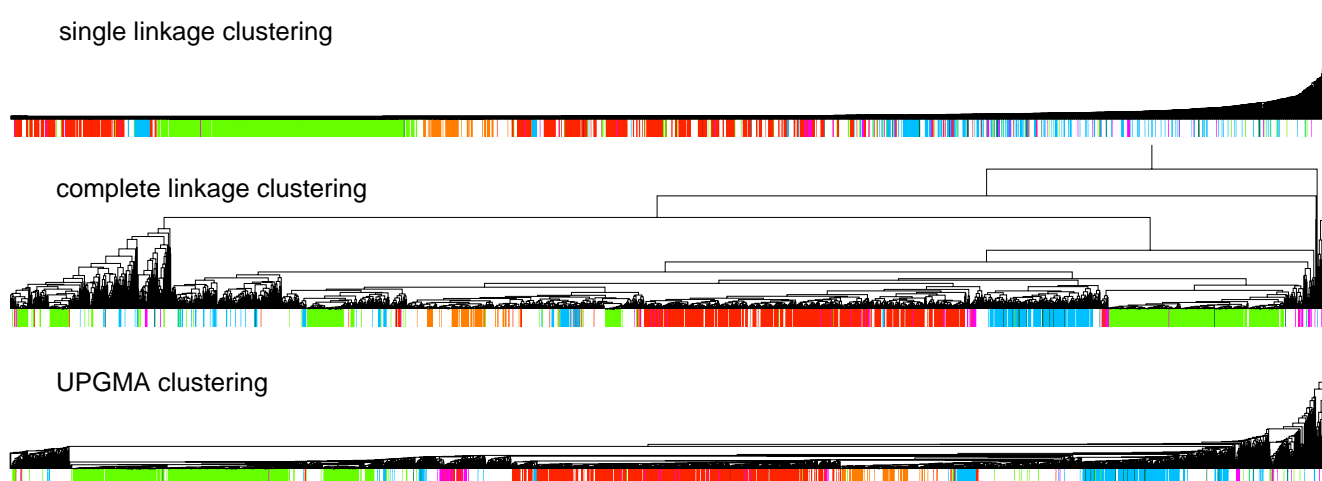
## 2. Results and Discussion

### 2.1. Clusters and RNA Classes

Starting from the matrix of tree edit distances between the 1973 families collected in Rfam 10.1 we computed hierarchical clusterings using UPGMA [27], single linkage [28], and complete linkage [29] methods. Results for each computation are represented as ultrametric trees (Figure 1). High-resolution versions of all diagrams are available at <http://www.biomol.unb.br/rfam/>.

Firstly the three hierarchies were compared to each other. By doing this, it was observed that the single linkage hierarchy is clearly different from the other two, considering that it has a caterpillar-like shape with only a few discernible clusters. Figure 1 shows a strong tendency of the microRNA, snoRNA and, to a lesser extent, the CRISPR families to cluster together. All three clustering methods also show a pronounced, although not total, separation among animal, plant, and viral microRNAs. In contrast to viral microRNAs, differences between plant and animal microRNAs have been well recorded in the literature, e.g., reviewed in [30]. Most viral microRNA families in Rfam are from Herpesviridae (16 of 20 viral microRNA families), which clustered with animal microRNA families, as expected. Although viral microRNAs are wide-spread [31,32], they are often poorly conserved and hence have not been included as an Rfam family.

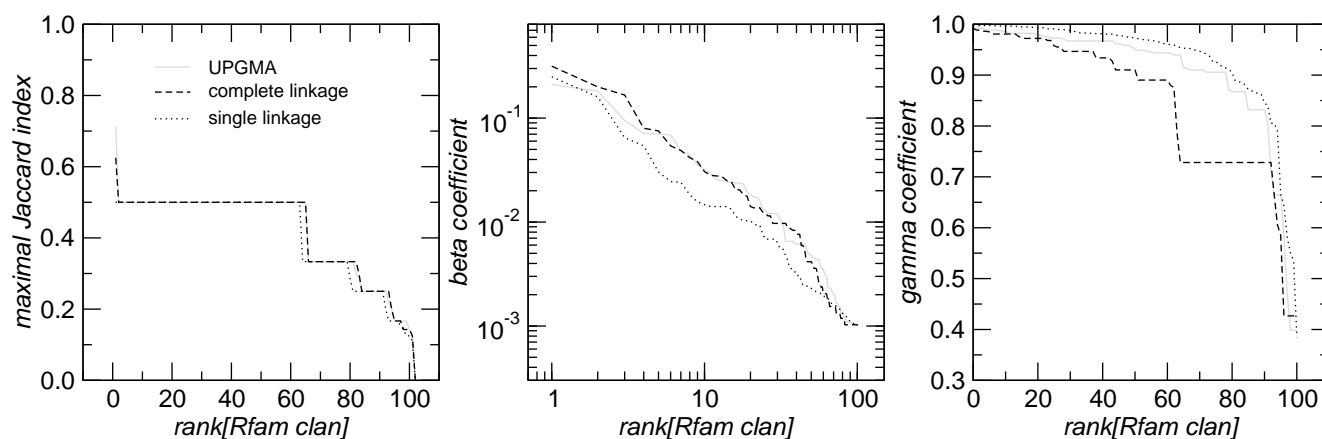
**Figure 1.** Dendrograms of the consensus structures of all Rfam 10.1 families computed with three different hierarchical clustering methods. Large important classes of ncRNAs are highlighted. Reddish colors denote three classes of microRNAs animal (scarlet), plant (fuchsia), and viral (brown). Box C/D snoRNAs are represented by bright green, while light blue indicates box H/ACA snoRNAs. Prokaryotic CRISPR families are shown in orange.



## 2.2. Clusters and Rfam Clans

Rfam 10.1 defines 102 clans, the majority of which comprises only two Rfam families. Figure 2 shows that they have not been well recovered by the clustering of the consensus structure. In fact, the coefficient  $\beta$  which measures a clan's tightness within the dendrogram shows a power-law like behavior, indicating that only a few clans are tightly clustered while most clans spread out over large areas of the dendrogram.

**Figure 2.** Distribution of  $\alpha$  (maximal Jaccard index),  $\beta$ , and  $\gamma$  for all of the 102 Rfam clans. These data show that most clans do not appear tightly clustered w.r.t. any of the three methods. The clans shown in the x-axis, together with  $\alpha$ ,  $\beta$ , and  $\gamma$  are listed in the supplementary material.



This is not entirely unexpected for several distinct reasons. Firstly, many clans consist of structurally related microRNA families. As all animal microRNA precursors share a very similar stem-loop structure,

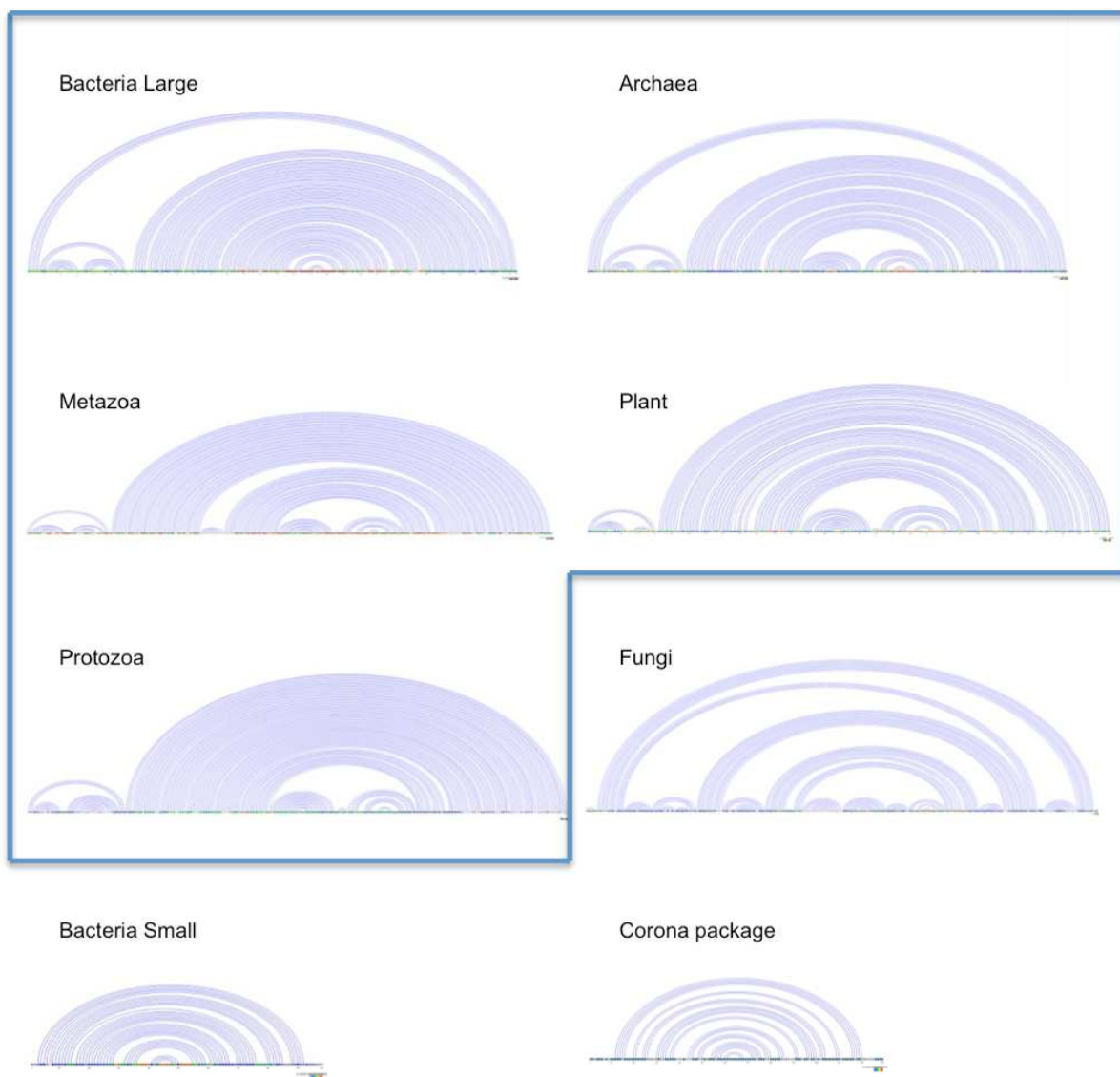
similarities inside a clan are not stronger than those shared with other microRNA families. These clans are thus dispersed throughout the microRNA clusters in Figure 1, leading to small values of  $\alpha$  and  $\beta$ . Secondly, this intra-clan similarity clusters most of the animal microRNAs in a single tree branch. In contrast, some other clans combine functionally related families that are clearly structurally distant in their supposed homologous structure: the tRNA clan (CL0001) for instance contains four families of bacterial tmRNAs along with the tRNA and the tRNA-Sec families.

Compared to a tRNA, a tmRNA has a big loop insertion that reflects its role as an mRNA-like repair template for ribosomal protein synthesis. Thus tmRNAs and tRNAs are separated by very large tree-edit distances. The four tmRNA families (RF00023, RF01849, RF01850, RF01851) also feature major differences. The most common one-piece type and distinct families of the two-piece tmRNAs can be respectively found in  $\alpha$ -proteobacteria,  $\beta$ -proteobacteria, and cyanobacteria. In addition, some tmRNAs with a permuted organization have also been described [33,34].

Despite the usually poor representation of Rfam clans in the structure-based clustering, there are a few clans that at least partially agree with the clustering data. The SRP clan (CL00003) is the one that was most coherently recovered as measured by the maximal Jaccard index  $\alpha \approx 0.71$ . It consists of 5 phylogenetically defined subgroups of signal recognition particle RNAs, whose homology is well established [35]: Metazoa SRP (RF00017), Bacteria large SRP (RF01854), Plant SRP (RF01855), Archaea SRP (RF01857) and Protozoa SRP (RF01856) (Figure 3). In addition to these families, which cluster together in our analysis, the SRP clan also contains Bacteria small SRP (RF00169) and Fungi SRP (RF01502), which were located far apart from the clustered SRP families in the UPGMA tree. The bacterial small SRP RNA (4.5S RNA) family (RF00169) is fully functional in mycoplasma and gram-negative bacteria, and harbors the conserved helices 5 and 8, which can be found in all kingdoms [35]. The reduced size of such RNAs prevented them from clustering with other SRP families. In our analysis, they clustered with RF00182, an unrelated viral element without any obvious functional or phylogenetic association. The fungal SRP RNA components (RF01502), on the other hand, possessed a highly conserved structure when compared to the other clan member. It contains several extra helices, however, that explain the large values of the tree distance in comparison to the other clan members. Thus, it appeared separated in a distinct tree branch. This case exemplifies that clustering with global distance functions alone cannot cope with those cases where there are dramatic structural differences between homologous RNAs.

Therefore, tree edit distance does not necessarily respect phylogenetic relationships. Nevertheless, homologous structures are frequently located in close proximity in the tree. Ribonuclease P, an ubiquitous ribozyme required for tRNA processing [36,37], for instance, is represented by four Rfam families in the clan (CL0002). Although they are located within the same subtree, they cluster with functionally and evolutionarily unrelated Rfam families such as fungi\_U3 (RF01846), U1\_yeast (RF00488), and the internal ribosomal entry sites of hepatitis A virus, IRES\_HepA (RF00228). Interestingly, the alpha\_tmRNA is also found in the RNase P sub-tree, very close to RNaseP\_nuc (RF00009) and RUF21 (RF01825), an yeast ncRNA with unknown functions. On the other hand, the prototypic tmRNA (RF00023) clusters together with the RNase\_MRP (RF00030), a distant homolog of RNase P [38,39]. Therefore, one can speculate that RNase P and tmRNA might share a common evolutionary history.

**Figure 3.** Linear representations of the secondary structures of the SRP clan members. The seven depicted SRP clan members display a conserved stem loop structure. The five of them that appear as a cluster in the UPGMA tree are delimited by a blue boundary. Fungal SRP family contains extra loops, while the Small Bacterial SRP families contain only a conserved stem loop domain, thus they have been both excluded from the cluster. An UPGMA neighbor of the Small Bacterial SRP family, an unrelated virus derived RFAM family (Corona package), is shown for comparison.



It is worth to mention that functional similarity, on the other hand, does not necessarily imply structural similarity. The IRES structures, frequently found in RNA virus and several cellular genes [40], are fine examples, although there are neither common designs, nor common signatures, or even common origins. Therefore, it is not surprising that they are found dispersed all over the cluster tree, gathering diverse functionally unrelated families.

The large plateau at  $\alpha = 0.5$  in Figure 2 mostly consists of size two clans, which are not recovered as cherries in the clustering tree. The CRISPR-2 clan is among the few larger clans that cluster together with most of its families. Using the dispersion coefficient  $\beta$  as a measure, clans CRISPR-2 and CRISPR-1

are well clustered. In addition a few microRNA and snoRNA clans with only two members received relatively large  $\beta$  values. A visual inspection of the UPGMA tree (Figure 1) reveals that differences in structural complexity can be directly inferred from the tree: larger and more complex RNAs, in particular rRNAs, appear isolated in this tree's leftmost. In the other extreme, simple structural elements such as microRNA's precursor hairpins and snoRNAs' simple structures were clearly clustered together.

### 2.3. Significant Structure-Based Clusters

It is natural to ask whether the hierarchical clustering trees can be used to identify significant structure-based clusters, and whether the families with strong similarities share biological functions. The Rfam clans have been constructed manually and consist of groups of families sharing common ancestors too divergent to be aligned, or also presenting good alignments but distinct functions so that they could not be included in the same Rfam family [3]. Some of these clans clearly share a common evolutionary ancestry, considering that similarities in their biological functions were experimentally verified. Hierarchical clustering, on the other hand, suggests how Rfam families that are not contained in clans could be related to each other and how clans may be organized at even higher levels of aggregation. It thus provides a more inclusive annotation.

Generally speaking, clan families were not tightly clustered together by our structure-based distance trees. In particular, larger, more complex secondary structures appeared widely separated from smaller, simpler ones (Figure 4(a)). Figure 4(b) shows in more detail the right hand side of Figure 4(a). As expected, C/D box snoRNAs (SNORD), H/ACA box snoRNAs (SNORA), miRNAs, and CRISPRs were clustered tightly together. Our data suggest then six clusters: SNORD1, SNORD2, SNORA, miRNA1, miRNA2 and CRISPR. The families included in these clusters are listed in Table 1. Such clusters possess very simple and clear secondary structure relationships and aggregate a large fraction of the families in the Rfam database.

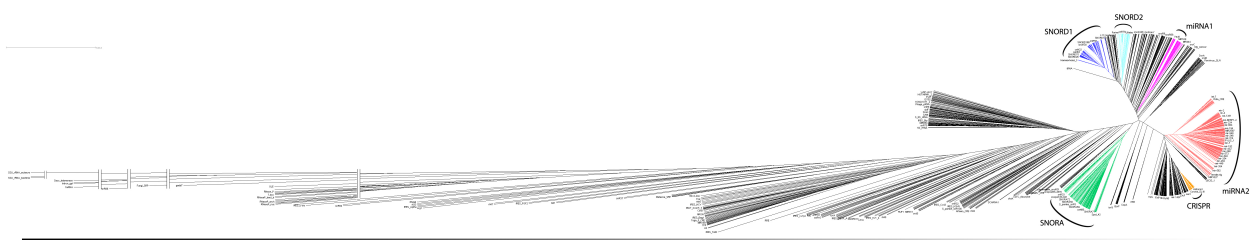
## 3. Experimental Section

Stockholm formatted alignments were retrieved as well as some metadata (such as accession numbers and short descriptions) for each of the 1,973 families of the Rfam database version 10.1 [3]. Then, the consensus structure was extracted from each alignment. Both Rfam and Vienna RNA Package [41] use string representations for secondary structures in which each base pair is denoted by a matching pair of parentheses and each unpaired bases by dots. This string has a natural interpretation as an ordered forest: dots denote leaf nodes and pairs of parentheses, interior nodes.

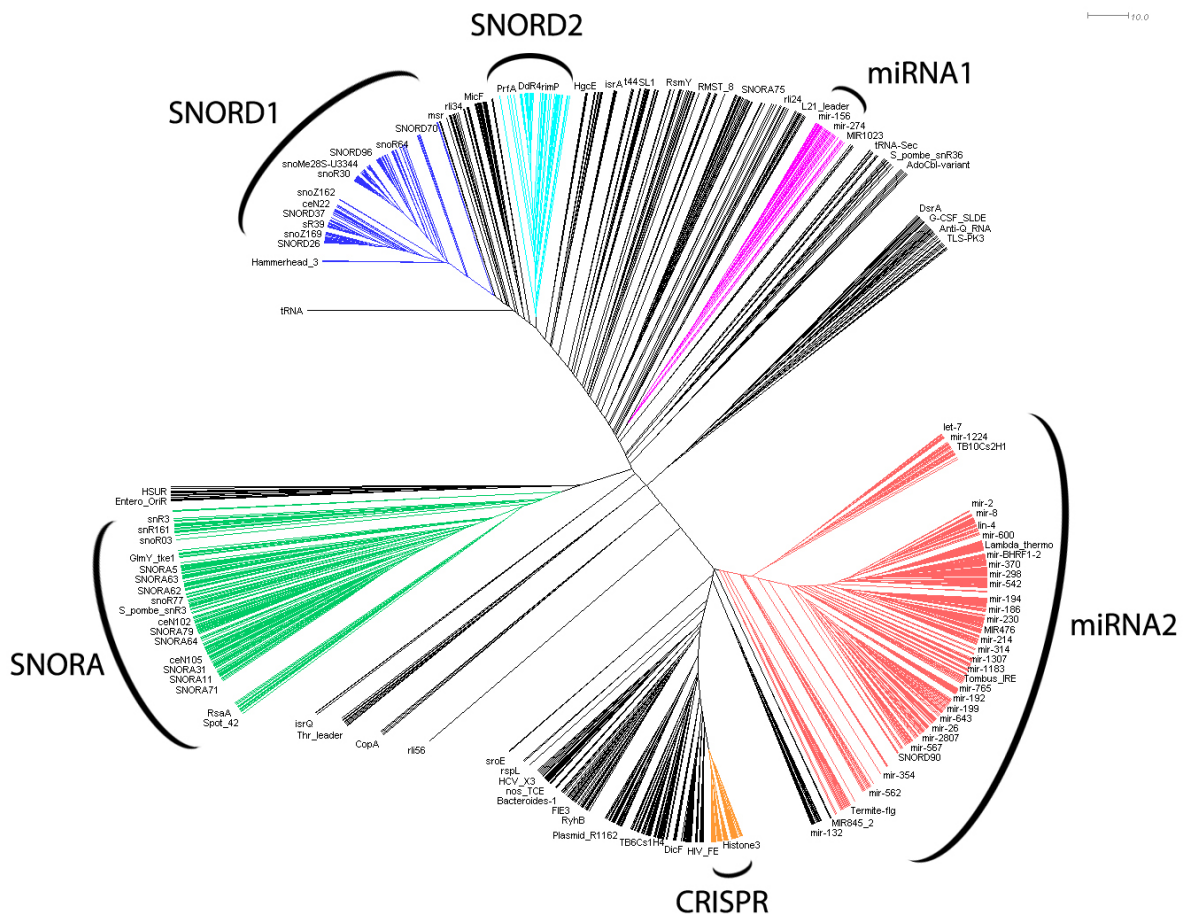
In this study, a full tree editing distance was used, in which every unpaired base and base pair were represented as nodes in the tree representation of the secondary structure. Canonical edit operations for forests were done by insertions and deletions of nodes. Upon node's deletion, its children became parent-node's children. Interior nodes represented base pairs and hence two nucleotides, while leaf nodes referred to single nucleotides, therefore in/del costs were set to 2 for interior nodes and 1 for leaves. Edit distance  $d(x, y)$  of two ordered forests  $x$  and  $y$  can in general be computed efficiently by means of a dynamic programming algorithm [42]. Here we used an implementation available in the RNAdistance program, a component of the Vienna RNA Package (version 1.8.4). One disadvantage of using this

distance measure is that two secondary structures of very distinct lengths always become very distant, even if they possess similar motifs. In addition, one of the most time consuming steps was running RNAdistance, since almost 2 million structure comparisons had to be completed. Nevertheless, this procedure took only a few hours.

**Figure 4.** Circular view of the UPGMA structural distance-based tree. **(a)** Circular view of the complete UPGMA dendrogram. One can see below the tree a gradient indicating more complex structures on the left and simpler ones on the right. Vertical bars represent shortening of branch length; **(b)** Closer view of clustered snoRNAs (SNORD1 (dark blue), SNORD 2 (light blue) and SNORA (green)), miRNAs (miRNA1 (pink) and miRNA2 (red)) and CRISPR (orange).



(a)



(b)



**Table 1.** Clusters of snoRNAs, miRNAs, and CRISPRs.

Cluster	Number of Rfam families included	Percentage of Rfam families of the expected ncRNA	Clans (name and identification) with all families included in Cluster
SNORD1	334	94.9%	SNORD52 (CL00063), U54 (CL00008), SNORD26 (CL00050), SNORD44 (CL00060), SNORD58 (CL00064), SNORD101 (CL00074), SNORD105 (CL00075), SNORND104 (CL00077), SNORD61 (CL00067), SNORD39 (CL00057), SNORD18 (CL00047), SNORD34 (CL00055), SNORD96 (CL00072), SNORD110 (CL00076), SNORD30 (CL00052), SNORD19 (CL00048), SNORD100 (CL00073)
SNORD2	86	81.4%	SNORD15 (CL00045)
SNORA	158	81.0%	SNORA7 (CL00025), SNORA28 (CL00033), SNORA44 (CL00036), SNORA17 (CL00029), SNORA35 (CL00034), SNORA5 (CL00024), SCARNA4 (CL00019)
miRNA1	45	86.6%	MIR171 (CL00099)
miRNA2	472	85.6%	mir-34 (CL00087), mir-216 (CL00094), mir-279 (CL00095), mir-36 (CL00088), mir-81 (CL00091), mir-182 (CL00093), mir-3 (CL00084), mir-50 (CL00089), mir-BART (CL00097), mir-137 (CL00092), mir-73 (CL00090)
CRISPR	100	59.0%	CRISPR-1 (CL00014), CRISPR-2 (CL00015)

Various agglomerative clustering methods differ only in their definition of the distance measure  $D$  between clusters. In each step, the two closest clusters,  $p$  and  $q$ , are united to a single cluster  $p \cup q$ . The distance of  $p \cup q$  to all clusters  $c$  is then obtained recursively starting from  $D(\{x\}, \{y\}) = d(x, y)$  for clusters consisting of individual points. The form of the recursion determines the particular clustering method [43]. For UPGMA  $D(c, p \cup q) = |p|/(|p| + |q|)D(c, p) + |q|/(|p| + |q|)D(c, q)$ , for single linkage  $D(c, p \cup q) = \min [D(c, p), D(c, q)]$ , and for complete linkage  $D(c, p \cup q) = \max [D(c, p), D(c, q)]$ . The resulting hierarchy of clusters is conveniently represented as a dendrogram  $T$ , in which leaves are the individual points. Each cluster  $c$  is uniquely identified by a node  $c'$  as the set of leaves of subtree rooted at  $c'$ . Dendrograms are drawn with a custom-made tool that allows to highlight sets of leaves using regular expressions that match against the extracted Rfam metadata.

Distance-based clusters are compared to Rfam clans and some other groupings, such as microRNAs or snoRNAs, using three quantitative measures. Denote by  $q$  an externally defined group, and let  $c$  be a cluster of  $T$ . Then we define

$$\alpha(q) = \max_c \frac{|q \cap c|}{|q \cup c|} \quad \beta(q) = \frac{|q|}{\min_{c:q \subset c} |c|} \quad \gamma(q) = 1 - \frac{\text{height}(q)}{\min_{c:q \subset c} \text{height}(c)}. \quad (1)$$

where  $\text{height}(q)$  is the height of a given subtree  $q$  as calculated by the cluster distance  $D(\cdot, \cdot)$ , *i.e.*, zero for a leaf  $q$ , or  $D(l, r)$  for a subtree  $q$  having children  $l$  and  $r$ . The maximal Jaccard index  $\alpha(q)$  compromises between coverage and contamination.  $\beta(q)$  measures how dispersed  $q$  is in  $T$  by computing the fraction

of members of  $q$  which compose the smallest cluster  $c \in T$  that entirely contains  $q$ . This measure is quite sensitive to individual outliers. On the other hand,  $\gamma(q)$ , which is also a measure of dispersion, takes the dendrogram cluster heights avoiding to assign bad scores to groups of very similar families. If the pre-defined group  $q$  appears as a cluster in  $T$ ,  $\alpha(q) = \beta(q) = 1$ .

The proposed clusters were constructed under careful examination of the most representative Rfam families of a particular ncRNA in the UPGMA dendrogram, *i.e.*, a cluster was formed containing the largest number of relatively close Rfam families in the tree. The parsing of the Rfam, the hierarchical clustering, and the drawing of the resulting dendrograms were implemented in Haskell, the source code of which is available at <http://hackage.haskell.org/package/> under the open source 3-clause BSD license. As mentioned before, the most time-consuming step was the computation of the distance matrix with RNAdistance, while all the other computations took only a few minutes on a notebook.

#### 4. Conclusions

We have developed a tool that calculates and draws automatically dendrograms showing the relationship among all the Rfam's ncRNA families. Such dendrograms have demonstrated to be able to confirm already expected relationships, such as snoRNAs and microRNAs, and also to be able to expose unknown ones, such as those discussed in Section 2.2.

The computational clustering reported here includes all of the 1,973 Rfam families, compared to only 306 families that were manually annotated as members of Rfam clans. Our analysis suggests that the automatic and manual methods should be combined to comprehensively reveal the structural and evolutionary relationships of the entire content of the Rfam database.

As future work, the dendrograms could be further analyzed, considering that it is suspected that there may be more information to be extracted than what we have already covered with this work. We are interested in studying putative ncRNAs' classes that can be defined based on the information we obtained from the dendrograms. For instance, to measure the clusters' consistency derived from those trees, it would be useful to compute the probability of finding the correct cluster for a given sequence. This could be used to know if a particular clustering could be used to predict a snoRNA or a miRNA class based on the secondary structure. It would also be interesting to see if there is a biological basis for the appearance of two clusters of box C/D snoRNAs and microRNAs, respectively, in Figure 4(b). Finally, there is also room for testing different distance measures, including ones that are less affected by length differences or measures that explicitly take into account conserved sequence elements.

#### Acknowledgments

This work was supported in part by CAPES (to F.A.L. and T.R.), by the *Deutsche Forschungsgemeinschaft* (grant STA 850/7-1 within SPP-1258 to P.F.S.), and by CNPq and FINEP 01.08.0166.00 (to M.E.M.T.W).

**References**

1. Griffiths-Jones, S.; Bateman, A.; Marshall, M.; Khanna, A.; Eddy, S.R. Rfam: An RNA family database. *Nucleic Acids Res.* **2003**, *31*, 439–441.
2. Griffiths-Jones, S.; Moxon, S.; Marshall, M.; Khanna, A.; Eddy, S.R.; Bateman, A. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **2005**, *33*, 121–124.
3. Gardner, P.P.; Daub, J.; Tate, J.; Moore, B.L.; Osuch, I.H.; Griffiths-Jones, S.; Finn, R.D.; Nawrocki, E.P.; Kolbe, D.L.; Eddy, S.R.; Bateman, A. Rfam: Wikipedia, clans and the “decimal” release. *Nucleic Acids Res.* **2011**, *39*, D141–D145.
4. The Athanasius F. Bompfünowerer RNA Consortium.; Backofen, R.; Flamm, C.; Fried, C.; Fritsch, G.; Hackermüller, J.; Hertel, J.; Hofacker, I.L.; Missal, K.; Mosig, Axel Prohaska, S.J.; Rose, D.; *et al.* RNAs everywhere: Genome-wide annotation of structured RNAs. *J. Exp. Zool. B: Mol. Dev. Evol.* **2007**, *308B*, 1–25.
5. Eigen, M.; Lindemann, B.F.; Tietze, M.; Winkler-Oswatitsch, R.; Dress, A.W.M.; von Haeseler, A. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **1989**, *244*, 673–679.
6. Rodin, A.S.; Szathmáry, E.; Rodin, S.N. One ancestor for two codes viewed from the perspective of two complementary modes of tRNA aminoacylation. *Biol. Direct* **2009**, *4*, doi:10.1186/1745-6150-4-4.
7. Wilusz, J.E.; Freier, S.M.; L., S.D. 3' End Processing of a Long Nuclear-Retained Noncoding RNA Yields a tRNA-like Cytoplasmic RNA. *Cell* **2008**, *135*, 919–932.
8. Sunwoo, H.; Dinger, M.E.; Wilusz, J.E.; Amaral, P.P.; Mattick, J.S.; Spector, D.L. MEN  $\epsilon/\beta$  nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* **2009**, *19*, 347–359.
9. Rozhdestvensky, T.S.; Kopylov, A.M.; Brosius, J.; Hüttenhofer, A. Neuronal BC1 RNA structure: Evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure. *RNA* **2001**, *7*, 722–730.
10. Hertel, J.; Lindemeyer, M.; Missal, K.; Fried, C.; Tanzer, A.; Flamm, C.; Hofacker, I.L.; Stadler, P.F. The students of bioinformatics computer labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **2006**, *7*, 1–15.
11. Sempere, L.F.; Cole, C.N.; McPeck, M.A.; Peterson, K.J. The phylogenetic distribution of metazoan microRNAs: Insights into evolutionary complexity and constraint. *J. Exp. Zool. B. Mol. Dev. Evol.* **2006**, *306*, 575–588.
12. Niwa, R.; Slack, F.J. The evolution of animal microRNA function. *Curr. Opin. Genet. Dev.* **2007**, *17*, 145–150.
13. Heimberg, A.M.; Cowper-Sal-lari, R.; Sémon, M.; Donoghue, P.C.; Peterson, K.J. MicroRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 19379–19383.
14. Tanzer, A.; Stadler, P.F. Molecular evolution of a microRNA cluster. *J. Mol. Biol.* **2004**, *339*, 327–335.

15. Chen, K.; Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **2007**, *8*, 93–103.
16. Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genetics* **2011**, *12*, 846–860.
17. Marz, M.; Gruber, A.R.; Höner zu Siederdisen, C.; Amman, F.; Badelt, S.; Bartschat, S.; Bernhart, S.H.; Beyer, W.; Kehr, S.; Lorenz, R.; *et al.* Animal snoRNAs and scaRNAs with Exceptional Structures. *RNA Biol.* **2011**, *8*, 938–946.
18. Dalloul, R.A.; Long, J.A.; Zimin, A.V.; Aslam, L.; Beal, K.; Blomberg, L.A.; Bouffard, P.; Burt, D.W.; Crasta, O.; Crooijmans, R.P.M.A.C.; *et al.* Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biol.* **2010**, *8*, doi:10.1371/journal.pbio.1000475.
19. Will, S.; Missal, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.* **2007**, *3*, doi:10.1371/journal.pcbi.0030065.
20. Liu, Q.; Olman, V.; Liu, H.; Ye, X.; Qiu, S.; Xu, Y. RNACluster: An integrated tool for RNA secondary structure comparison and clustering. *J. Comput. Chem.* **2008**, *29*, 1517–1526.
21. Torarinsson, E.; Havgaard, J.H.; Gorodkin, J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* **2007**, *23*, 926–932.
22. Höchsmann, M.; Töller, T.; Giegerich, R.; Kurtz, S. Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.* **2003**, *2*, 159–168.
23. Shapiro, B.A.; Zhang, K.Z. Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.* **1990**, *6*, 309–318.
24. Jiang, T.; Lin, G.; Ma, B.; Zhang, K. A general edit distance between RNA structures. *J. Comput. Biol.* **2002**, *9*, 371–388.
25. Hofacker, I.L.; Fekete, M.; Stadler, P.F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **2002**, *319*, 1059–1066.
26. Sankoff, D. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.* **1985**, *45*, 810–825.
27. Sokal, R.R.; Michener, C.D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **1958**, *28*, 1409–1438.
28. Sibson, R. SLINK: An optimally efficient algorithm for the single-link cluster method. *Comput. J. (BCS)* **1973**, *16*, 30–34.
29. Sorensen, 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. *Biologiske Skrifter* **1948**, *5*, 1–34.
30. Axtell, M.J.; Westholm, J.O.; Lai, E.C. Vive la différence: Biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.* **2011**, *12*, doi:10.1186/gb-2011-12-4-221.
31. Li, S.C.; Shiau, C.K.; Lin, W.C. Vir-Mir db: Prediction of viral microRNA candidate hairpins. *Nucleic Acids Res.* **2008**, *36*, D184–D189.
32. Cullen, B.R. Viruses and microRNAs: RISCy interactions with serious consequences. *Genes Dev.* **2011**, *25*, 1881–1894.

33. Sharkady, S.M.; Williams, K.P. A third lineage with two-piece tmRNA. *Nucleic Acids Res.* **2004**, *32*, 4531–4538.
34. Mao, C.; Bhardwaj, K.; Sharkady, S.M.; Fish, R.I.; Driscoll, T.; Wower, J.; Zwieb, C.; Sobral, B.W.; Williams, K.P. Variations on the tmRNA gene. *RNA Biol.* **2009**, *6*, 355–361.
35. Rosenblad, M.A.; Larsen, N.; Samuelsson, T.; Zwieb, C. Kinship in the SRP RNA family. *RNA Biol.* **2009**, *6*, 508–516.
36. Piccinelli, P.; Rosenblad, M.A.; Samuelsson, T. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.* **2005**, *33*, 4485–4495.
37. Walker, S.C.; Engelke, D.R. Ribonuclease P: The evolution of an ancient RNA enzyme. *Crit. Rev. Biochem. Mol. Biol.* **2006**, *41*, 77–102.
38. Schmitt, M.E.; Bennett, J.L.; Dairaghi, D.J.; Clayton, D.A. Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. *FASEB J.* **1993**, *7*, 208–213.
39. Woodhams, M.D.; Stadler, P.F.; Penny, D.; Collins, L.J. RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.* **2007**, *7*, doi:10.1186/1471-2148-7-S1-S13.
40. Pisarev, A.V.; Shirokikh, N.E.; Hellen, C.U. Translation initiation by factor-independent binding of eukaryotic ribosomes to internal ribosomal entry sites. *C R Biologie* **2005**, *328*, 589–605.
41. Hofacker, I.L.; Fontana, W.; Stadler, P.F.; Bonhoeffer, L.S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **1994**, *125*, 167–188.
42. Zhang, K.; Shasha, D. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.* **1989**, *18*, 1245–1262.
43. Lance, G.N.; Williams, W.T. A general theory of classificatory sorting strategies I. Hierarchical systems. *Comp. J.* **1967**, *9*, 373–380.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).