Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering

Sebastian Will[1], Kristin Reiche[2], Ivo L. Hofacker[3], Peter F. Stadler[2], and Rolf Backofen[1*]

[1] Chair for Bioinformatics · Institute of Computer Science · Albert-Ludwigs-Universität
Georges-Koehler-Allee, Geb. 106 · D-79110 Freiburg, Germany · {*will,backofen*}*@informatik.uni-freiburg.de*
[2] Bioinformatics Group, Department of Computer Science · University of Leipzig
Härtelstraße 16-18 · D-04107 Leipzig, Germany · {*kristin,studla*}*@bioinf.uni-leipzig.de*
[3] Department of Theoretical Chemistry · University of Vienna
Währingerstraße 17 · A-1090 Wien, Austria · {*ivo,studla*}*@tbi.univie.ac.at*

Abstract

The RFAM database defines families of ncRNAs by means of sequence similarities that are sufficient to establish homology. In some cases, such as microRNAs, box H/ACA snoRNAs, functional commonalities define classes of RNAs that are characterized by structural similarities, and typically consist of multiple RNA families. Recent advances in high-throughput transcriptomics and comparative genomics have produced very large sets of putative non-coding RNAs and regulatory RNA signals. For many of them, evidence for stabilizing selection acting on their secondary structures has been derived, and at least approximate models of their structures have been computed. The overwhelming majority of these hypothetical RNAs cannot be assigned to established families or classes.

We present here a structure-based clustering approach that is capable of extracting putative RNA classes from genome-wide surveys for structured RNAs. The LocARNA tool implements a novel variant of the Sankoff algorithm that is sufficiently fast to deal with several thousand candidate sequences. The method is also robust against false positive predictions, i.e., a contamination of the input data with unstructured or non-conserved sequences.

We have successfully tested the LocARNA-based clustering approach on the sequences of the RFAM-seed alignments. Furthermore, we have applied it to a previously published set of 3332 predicted structured elements in the *Ciona intestinalis* genomes (Missal *et al.*, Bioinformatics **21(S2), i77-i78)**. In addition to recovering e.g. tRNAs as a structure-based class, the method identifies several RNA families, including microRNA and snoRNA candidates, and suggests several novel classes of ncRNAs for which to-date no representative has been experimentally characterized.

1 Introduction

Starting with the discovery of microRNAs [1, 2, 3] and the advent of genome-wide transcriptomics [4, 5, 6], it has become obvious that RNA plays a large variety of important, often regulatory, roles in living organism that extend far beyond being a mere intermediate in protein biosynthesis. The elucidation of the functional roles of the plethora of newly discovered non-coding RNAs (ncRNAs) has thus become a central research interest in molecular biology.

Recent advances in computational RNomics have resulted in numerous software packages that can be employed to detect ncRNAs with evolutionarily conserved secondary structures [7, 8, 9, 10, 11, 12]. Two of these, EvoFold [10] and RNAz [9, 13] are efficient enough to be applied to genome-wide surveys in mammals [10, 13] and other metazoan clades [14, 15]. Both approaches start from multiple sequence alignments. While EvoFold uses the SCFG approach pioneered by qrna [7], RNAz is based on evaluating the folding thermodynamics. Both approaches classify input alignments either as unstructured or as possessing a common RNA secondary structure; in the latter case they provide a prediction for the consensus structure of the aligned sequences.

Just as in the case of proteins, ncRNA-sequences can be grouped into *families* that are characterized by clear homologies. Usually the members in a family share functional characteristics as well as conserved

*To whom correspondence should be addressed, email: *backofen@informatik.uni-freiburg.de*

sequence and structure motifs. Indeed, the RFAM database [16] compiles several hundred families of ncRNAs based on this observation. Examples include the individual snRNAs U1, U2, U4, U5, and U6, 5S rRNA, RNAse P RNA, the RNA component of telomerase, more than a hundred families of snoRNAs and several hundred microRNA families collected in mirbase [17]. In many cases, RNA families can be grouped together, forming a *ncRNA class* whose members have no discernible homology at sequence level, but still share common structural and functional properties. The best-known classes are tRNAs (although is well established that all tRNAs derive from a common ancestor [18]), the two distinct classes of snoRNAs (box H/ACA and box C/D), RNAse P and MRP RNAs, and microRNAs. It is thus natural to ask whether the many ncRNA candidates that have been predicted computationally can be grouped into families or even classes, and in particular, whether there is evidence for novel families and classes for which we have not yet seen experimentally verified representatives.

As sequence similarity is often remote even within well-established RNA families, we cannot rely on pure sequence alignment techniques for this task. Indeed, it has been shown that sequence alignments of structured RNAs fail at pairwise sequence identities below about 60% [19]. Several different algorithmic approaches have been introduced in the past to determine structural similarities and to derive consensus structure patterns for RNAs that are too diverse to be alignable at sequence level. The corresponding software tools, such as MARNA [20], PMmulti [21], RNAforrester [22], cannot be applied without modifications to the problem of clustering predicted structures from RNAz or EvoFold surveys, however. The main reason is that these ncRNA detectors are not guaranteed to find the complete ncRNA genes; rather they usually detect particularly conserved substructures and sometimes the predictions are contaminated with spurious predictions in the flanking sequences. Thus a local structure-based alignment algorithm is necessary. This is already implemented in RNAforrester [22], which is based on tree-alignment, and in the local sequence-structure alignment approach described [23], which in addition can detect also *structurally* local motifs. A related approach detects exact local sequence structure patterns in $O(n^2)$ [24]. However, all these approaches require a *single* known or predicted input structure. Tree-alignment and tree-editing in addition have only limited capabilities to repair incorrect base-pairs. Tree-alignment is particularly restrictive in this respect since even broken arcs must be nested. As a consequence, RNAforrester tends to produce many alignment columns that contain mostly gap characters in the multiple alignment mode.

In contrast, derivates of the Sankoff algorithm [25] solve the problem of simultaneous folding and alignment, which turned out to be more appropriate. However, the large number of predicted ncRNAs, several thousands in the case of nematode and urochordate genomes and close to 100000 in the case of mammals, calls for more efficient variants of these algorithms.

In this contribution we introduce LocARNA, a local pairwise structural alignment algorithm for pseudoknot-free RNA secondary structures, and its multiple version mLocARNA. (m)LocARNA is a Sankoff-style algorithm, similar to PMmulti, that is efficient enough to be used for large clustering of predicted ncRNAs. We have successfully tested the LocARNA-based clustering approach on the sequences of the RFAM-seed alignments to demonstrate the feasibility of the approach, and to evaluate the results. Furthermore, we use the data from a survey of the ascidians *Ciona intestinalis* and *Ciona savignyi* [14] to achieve the following goals: (1) We search for novel, clade-specific RNA families in Ciona, which is of interest in itself. (2) In doing so, we can increase the credibility of some of the predicted ncRNAs, since being part of larger family of related RNAs with similar structure reduces the likelihood of being a false positive prediction. (3) We improve the genome annotation be assigning additional ncRNAs to known families. (4) The inferred consensus structures of novel families form a starting point for subsequent searches in related organisms.

2 Structure-Based Clustering

In this work, we set up a pipeline for automated clustering of ncRNAs (or ncRNA candidates) and semi-automated selection of novel, complex clusters of RNAs. The input is a set of RNAs $R_1, \ldots, R_m$, which are given by their sequences, and the output is a hierarchical clustering of these RNAs. In addition, we will generate a fast, pre-sorted, and annotated overview of the clusters for further inspection by an expert. Our pipeline is built from the following steps:

1. For each of the RNAs, we compute structural information using McCaskill's algorithm, implemented in RNAfold. This algorithm computes a matrix of pair probabilities based on a complete energy model of RNAs.
2. The next step is to compute all pairwise alignments of the structurally annotated sequences using LocARNA. Note that this requires to compute $\mathcal{O}(m^2)$ pairwise sequence/structure alignments for

determining the distance matrix. Note further that performing all pairwise comparisons cannot be reasonably circumvented or replaced in a full-featured clustering procedure. For genomic-scale data sets, $\mathcal{O}(m^2)$ comparisons are way too costly for most existing sequence-structure approaches. The computational efficiency remains crucial, even if this computationally most intensive procedure is distributed for parallel computation, which we do in a straightforward manner. As result we assign a LocARNA-alignment score $\mathrm{score}(i,j)$ to each pair of RNAs (R_i, R_j).

3. A cluster-tree is generated by applying the weighted pair group method algorithm (WPGMA), which is also known as average-linkage clustering, to a matrix of pairwise distances of the RNAs. There, the distances $d(i,j)$ correspond directly to our LocARNA-scores. Instead of computing distances as $max_{ij} - \mathrm{score}(i,j)$, we define distances by

$$d(i,j) = max(0, q - \mathrm{score}(i,j)),$$

where q is the x-quantile (e.g. $x = 99\%$) of all pairwise scores. This decision avoids that exceptionally large scores influence the distance-transformation. In the resulting tree, internal nodes correspond to clusters of RNAs. Their heights correspond to the mean pairwise LocARNA-scores of their constituents and thus give a single-value measure of cluster quality.

4. A good overview and a true quality assessment of the clusters can be provided best through multiple alignments of each cluster. We simultaneously construct all multiple sequence/structure alignments, i.e. one for each cluster, by only $O(n)$ runs of the pairwise alignment algorithm. This can be done by constructing the multiple alignments progressively, using the, already constructed, cluster-tree as guide tree.

   From each of the multiple alignments, we collect information that can guide a quality assessment of the cluster. We compute the mean pairwise sequence identity (MPI) and, using RNAalifold, the structure conservation index (SCI), the consensus minimum free energy (MFE), the consensus MFE structure, and the consensus base pair probabilities. Sorting the list of generated clusters by the quantities size of cluster, SCI, MPI, and MFE provides the expert with a automatically proposed order for his manual inspection of the clusters. The multiple alignment itself and the consensus structure information facilitate the selection of "interesting" clusters.

This pipeline crucially depends on pairwise sequence/structure alignments. Therefore, we require the following algorithmic components which we describe in some detail in the following subsections:

2.1 We need an efficient algorithm for high quality pairwise alignments of RNAs that considers both sequence and structure information. For this purpose, the best results are achieved with Sankoff-style algorithms. We provide the new method called LocARNA, which is much more efficient than current approaches and uses base pair probabilities as structural input.

2.2 For the selection of clusters, one important sub-problem is to extract consensus structure information from the clustered RNAs, which is done by using RNAalifold. For producing the input for RNAalifold, we introduce the *local* multiple alignment method mLocARNA.

2.1 LocARNA: Efficient pairwise local sequence/structure alignment

For the pairwise alignments of RNA we use our novel tool LocARNA, which computes local alignments of RNA. It is a Sankoff-style algorithm in the spirit of PMcomp, but goes beyond its ancestor by introducing local alignment and significantly improving the efficiency.

The Sankoff algorithm [25] provides a general solution to the problem of simultaneously computing an alignment and the common secondary structure of the two aligned sequences. In its full form, the problem requires $\mathcal{O}(n^6)$ CPU time and $\mathcal{O}(n^4)$ memory, where n is length of the RNA sequences to be aligned. In general, one can distinguish two variants of the Sankoff algorithms: Programs such as foldalign [26, 27] and dynalign [28] implement a more or less complete energy model for the RNA folding part. In contrast, PMcomp [21] assumes that a structure model for the two input sequences is already known and given in the form of weights for the individual base pairs. However, note that such a structure model is reasonably obtained using McCaskill's algorithm [29] again on the basis of a full featured energy model.

Consider two sequences A and B with associated base pairing weight matrices Ψ^A and Ψ^B, respectively. The goal is to compute a sequence alignment $\mathbb{A}$ of A and B together with secondary structure $\mathbb{S}$ on $\mathbb{A}$. $\mathbb{A}$ consists of a set of (mis)matches written as pairs (i,k), where i is a position in A, and k a position in B. The consensus secondary structure $\mathbb{S}$ for an alignment $\mathbb{A}$ consists of a set of quadruples $(ij;kl)$, where $(i,k) \in \mathbb{A}$ and $(j,l) \in \mathbb{A}$ are two matches in $\mathbb{A}$, (i,j) is a base pair on sequence A and (k,l) is a base pair on sequence B. Furthermore, denote by $\mathbb{A}_s$ the single-stranded part of the alignment, i.e., if $(i,k) \in \mathbb{A}_s$

then there is no pair (j, l) such that $(ij; kl) \in \mathbb{S}$ or $(ji; kl) \in \mathbb{S}$. The goal is to determine the pair $(\mathbb{A}, \mathbb{S})$ that maximized the score function

$$\sum_{(ij;kl)\in\mathbb{S}} (\Psi^A_{ij} + \Psi^B_{kl}) + \sum_{(i,k)\in\mathbb{A}_s} \sigma(A_i, B_k) - N_{\text{gap}}\gamma, \tag{1}$$

where $\sigma : \{A, C, G, U\}^2 \to \mathbb{R}$ is the similarity score for (mis)matches, γ is the gap score parameter and N_{gap} is the number of insertions and deletions in the alignment $\mathbb{A}$.

Both PMcomp and our novel tool, LocARNA("**Loc**al **A**lignment of **RNA**"), use base pair scores that are derived from the base pairing probability matrices of the two individual sequences. More precisely, we use here

$$\Psi_{ij} = \begin{cases} \log \frac{P_{ij}}{p_0} \big/ \log \frac{1}{p_0} & \text{if } P_{ij} \geq p^* \\ -\infty & \text{otherwise,} \end{cases} \tag{2}$$

where P_{ij} is equilibrium pairing probability as computed by McCaskill's algorithm [29], p_0 is the expected probability for a pairing to occur at random and p^* is the cut-off probability, below which the arcs are ignored. Formally, this is expressed by assigning $-\infty$ as weight in this case. The term $\log P_{ij}/p_0$ is the log-odds score for having a specific base pairing against the null model of a random pairing, and $\log 1/p_0$ is a normalization factor that transforms the weights to a maximum of 1. The reason for this normalization is just that it is more easier to balance the sequence score against the structure score.

LocARNA improves the PMcomp approach in several ways. First of all, it uses a modified dynamic programming approach that allows us to utilize the fact that typically the number of significant base pairs does not grow with $\mathcal{O}(n^2)$, i.e., that the Ψ matrices are usually sparse. In particular, if p^* is constant for different n, then each base can take part in at most $1/p^*$, and thus $\mathcal{O}(1)$, many base pairs. Hence, there are only $m = \mathcal{O}(n)$ significant entries in Ψ.

We define $D_{i\,j;k\,l}$ as the maximal similarity score of an alignment for the subsequences $A[i..j]$ and $B[k..l]$ with the additional condition that $(i\,j; k\,l)$ is part of consensus secondary structure. To profit from the reduced number of significant basepairs in time and space complexity, we calculate and store only $D_{i\,j;k\,l}$ that correspond to significant base pairs. Due to this modification, we need to take special care to avoid redundant computation. Therefore, we compute the entries $D_{i\,j;k\,l}$ by fixing i and k and varying only j and l. We introduce the notation $D_{i\,\cdot;k\,\cdot}$ to denote the matrix slice where i and k are fix. The efficient calculation of $D_{i\,\cdot;k\,\cdot}$ in $\mathcal{O}(n^2)$ time requires auxiliary matrices M, where the entries $M_{i\,j;k\,l}$ are the optimal similarity score of subsequences $A[i + 1..j]$ and $B[k + 1..l]$, and leads to computation order that differs from PMcomp. Finally, the dynamic programming recursion for M and D takes the usual form of a Sankoff-style algorithm:

$$\begin{aligned} M_{i\,j;k\,l} &= \max \begin{cases} M_{i\,j-1;k\,l-1} + \sigma(A_j, B_l) \\ M_{i\,j-1;k\,l} + \gamma \\ M_{i\,j;k\,l-1} + \gamma \\ \max\limits_{j'l'} M_{i\,j'-1;k\,l'-1} + D_{j'\,j;l'\,l} \end{cases} \\ D_{i\,j;k\,l} &= M_{i\,j-1;k\,l-1} + \Psi^A_{ij} + \Psi^B_{kl} \end{aligned} \tag{3}$$

The important observation is that the last, computationally most expensive, alternative in the M recursion needs to be evaluated only for $P^A_{j'l'} \geq p^*$ and $P^B_{jl} \geq p^*$, and, analogously, D needs to be stored only for matching base pairs. We observe that $D_{i\,\cdot;k\,\cdot}$ depends only on $M_{i\,\cdot;k\,\cdot}$, which in turn can be computed from other $M_{i\,\cdot;k\,\cdot}$ entries. Thus we only need to store the entries of M for the current values of i and k, i.e. $\mathcal{O}(n^2)$ entries. The recursion can therefore be evaluated in $\mathcal{O}(m^2 + n^2)$ memory and $\mathcal{O}(n^2(n^2 + m^2))$ time.

From the matrices M and D, we can now compute the score of the best global alignment as well as the score of the best local alignment. In our study, we are only interested in the latter. Global alignment is only explained for better understanding and for comparison to the global alignment algorithm PMcomp. The score of the global alignment can be computed by evaluating the recursion for $M_{0\,j;0\,l}$, i.e. the optimal global alignment score is $M_{0\,|A|;0\,|B|}$.

Concerning local alignment, in a Sankoff-style approach usually we compute a four-dimensional matrix of alignment scores for *each* pair of subsequences $A_i \dots A_j$ and $B_k \dots B_l$. In this case, we could trivially obtain the best local alignment score by searching for the maximal score.

In our case however, we cannot apply this simple method, since we do not compute entries for all possible pairs of subsequences. Rather, we compute only scores for subsequences that are closed by

(significant) base pairs or prefixes of them. Those scores are either stored in $D_{i\,j;k\,l}$ (in the case of closing a base-pair match) or in $M_{i\,j;k\,l}$.

Instead we will borrow, slightly tailored for our purpose, the trick of standard sequence alignment, which is to add an additional 0 entry in the recursion for cutting off dissimilar prefix-alignments. The best local alignment is then obtained as the maximal entry of the matrix.

However, note that we must not change the recursion equations for all $M_{i\,j;k\,l}$ which serve for computing some entry of D. Only for alignments of subsequences $A_i \dots A_j$ and $B_k \dots B_l$, where at least one of the subsequences is not enclosed by a (significant) base pair, it is correct to cut off dissimilar prefix-alignments. All these cases are accounted for when considering the alignments of all pairs of prefixes of A and B, which are stored in the $M_{0\,\bullet;0\,\bullet}$ slice. Therefore, we introduce a variant of Eq. (3) by

$$M^{\mathrm{T}}_{0\,j;0\,l} = \max(0, M_{0\,j;0\,l}).$$

Instead of computing the entries $M_{0\,j;0\,l}$, we will then compute the entries $M^{\mathrm{T}}_{0\,j;0\,l}$. Note that the entries $M_{0\,j;0\,l}$ will not needed to compute any entry $D_{i'\,j';k'\,l'}$.

By computing the maximum of score 0 and $M_{0\,j;0\,l}$, we ensure that entries in $M^{\mathrm{T}}_{0\,\bullet;0\,\bullet}$ are nonnegative. Since negative scores are considered dissimilar, we thereby remove prefix-alignments that do not belong to the local alignment. The optimal local alignment score is then $\max_{jl}(M^{\mathrm{T}}_{0\,j;0\,l})$.

The corresponding optimal alignment and consensus secondary structure can now be obtained by backtracing, i.e. for local alignment we start from the maximal entry in $M^{\mathrm{T}}_{0\,\bullet;0\,\bullet}$ and stop when similarity drops to its minimal value of 0. In addition, for every pair $(i\,j;k\,l)$ in the consensus structure we have to re-compute the $M_{i\,\bullet;k\,\bullet}$ at a cost of $\mathcal{O}(n^2+m^2)$. Since there are at most $\mathcal{O}(n)$ pairs in the consensus structure, the cost of backtracing stays negligible.

`LocARNA` is implemented in `C++`, which results in a further performance gain relative to the `Perl` implementation of `PMcomp`. While it fully exploits speed and memory reductions that can be obtained by limiting possible consensus structures, additional performance gains are possible by restricting the possible sequence alignments. This is done e.g. in `stemloc` [30] by using "alignment envelops". A similar but more easily implemented technique is used by `CONSAN` [31], where high confidence matches ("pins") are derived from local sequence alignments. The algorithm then considers only alignments that contain all pins.

2.2 Local Multiple Sequence Structure Alignments

Based on the pairwise `LocARNA` algorithm, we construct a progressive multiple alignment method, `mLocARNA`, which is similar in spirit to `PMmulti`, the `PMcomp`-derived multiple alignment tool [21]. `mLocARNA` differs from `PMmulti` in the algorithm for computing base paring weights $\Psi^{A\circ B}$ for the combined alignment of A and B from the base paring weights of the sub-alignments (or sequences) A and B. For a pair of columns pq in the alignment of A and B, `PMcomp` defines the combined base pair weight by

$$\Psi^{A\circ B}_{pq} = \begin{cases} \sqrt{\Psi^{A}_{i_p i_q} \times \Psi^{B}_{k_p k_q}} & \text{if } p \text{ and } q \text{ are gap-less} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where i_p and i_q are the positions corresponding to p and q in the sub-alignment A, respectively. k_p and k_q are defined analogously for sub-alignment B. This has the effect that whenever one sub-alignment contains a gap at p or q or has a very low base pair probability then the structural information between p and q from the other sub-alignment is effectively lost. In consequence, `PMmulti` tends to remove most base pairs when aligning many sequences.

In order to avoid this undesired effect, we introduce the new definition

$$\Psi^{A\circ B}_{pq} = \sqrt{\bar{\Psi}^{A}_{pq} \times \bar{\Psi}^{B}_{pq}}, \tag{5}$$

where

$$\bar{\Psi}^{A}_{pq} = \begin{cases} \max(p_0, \Psi^{A}_{i_p i_q}) & \text{if } p \text{ and } q \text{ are gap-less} \\ p_0 & \text{otherwise} \end{cases}$$

and $\bar{\Psi}^{B}_{pq}$ is defined analogously.

As usual, the order of pairwise alignments is directed by a guide tree. We use for that purpose the sub-trees produced by the hierarchical clustering.

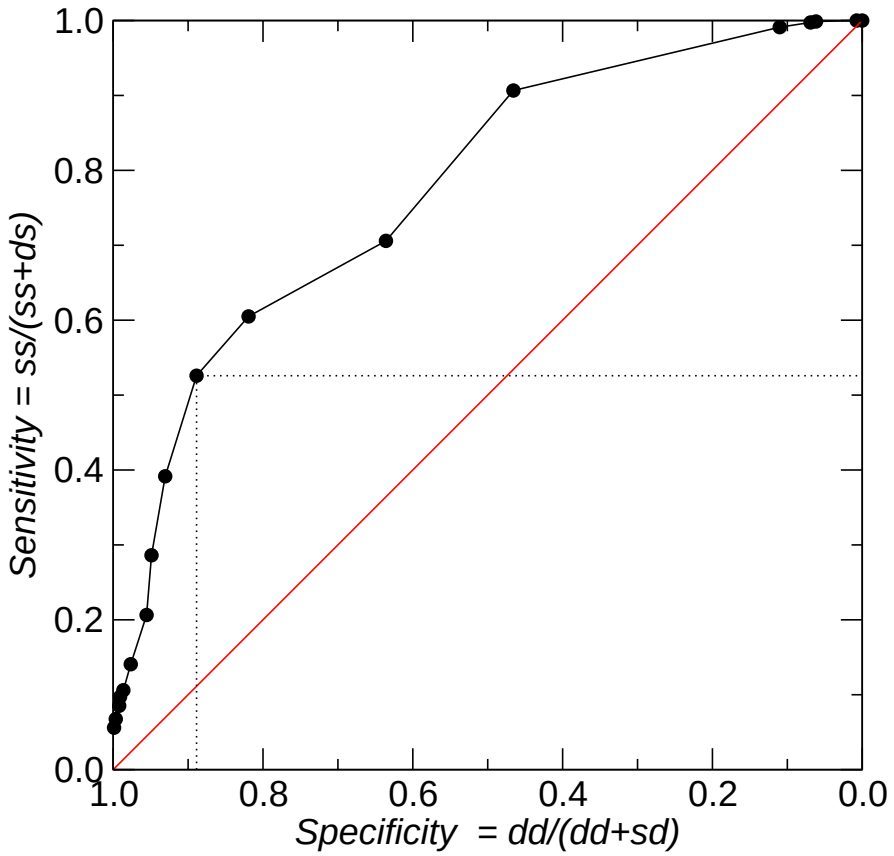

Figure 1: ROC curve the global comparison of clustering and RFAM families. At a false positive rate of 12% we achieve a sensitivity of 52% (correctly grouping together sequences of the same family), which is more than sufficient to detect families.

3 Results

3.1 Evaluation of the Clustering Procedure

To evaluate the quality of our clustering approach we have applied our procedure to the sequences in the RFAM seed alignments. Our test set consists of all seed sequences which have no more than 80% sequence identity and do not exceed 400nt in length, resulting in 3901 sequences from 504 families. Normally, quality measures such as sensitivity and specificity are defined for binary classification problems, while here we face the problem of comparing our hierarchical clustering with the family assignment in RFAM. In principle, there are two ways of looking at the problem, namely globally (considering the complete set of clusters), and locally (considering the quality for each family separately).

Concerning the global view, the complete RFAM defines a partition of the set of all sequences into families (or clusters), and we can compare the degree of agreement between the partition defined by our clustering with the partition defined by RFAM. Since we have a hierarchical clustering, different sets of clusters can be defined by cutting the tree at different thresholds ϑ, and we have to compare all these thresholds to find the set of clusters with the best agreement. The problem of comparing the partition defined by a given set of clusters (generated by cutting the tree at some specific level), with the partition defined by RFAM is now transformed into a classification problem as follows. We consider all possible pairs of sequences, and define the number of true positives (ss) as the number of sequence pairs from the same family that lie in the same cluster. Analogously, the number of false positives, false negatives, and true negatives are given by the number of pairs from different families but same cluster (ds), same family but different clusters (sd), and different families and different clusters (dd), respectively. Sensitivity and specificity are then defined as usual, namely $spec = dd/(dd + ds)$ and $sens = ss/(ss + sd)$. The receiver operating characteristic (ROC), obtained by plotting the sensitivity against the false positive rate (1-specificity) for different values of the cut-off ϑ, is shown in figure 3.1.

A problem in the comparison to RFAM families is that different families exhibit very different diversity:

Table 1: Average precision and F-measure for different minimum recall levels.

Minimum recall level	Average recall	Average precision	Average F-measure
0.50	0.5818	0.8280	0.6079
0.55	0.6996	0.7819	0.6475
0.60	0.7277	0.7530	0.6391
0.65	0.7596	0.7117	0.6191
0.70	0.8092	0.6831	0.6158
0.75	0.8519	0.5949	0.5650
0.80	0.8763	0.5701	0.5526
0.85	0.9381	0.4794	0.4964
0.90	0.9599	0.4419	0.4647
0.95	0.9766	0.3907	0.4173

Averages are weighted by family size. Families which are only represented by one sequence do not contribute to average as their precision is always 1.

some families consist only of closely related sequences while others accommodate significant variation in sequence and structure. Therefore one should not expect that the RFAM family division can be modeled by using one fixed threshold ϑ for all families. We therefore consider a local, family-wise, criterium for the clustering quality. For a given RFAM family R and a cluster C we define the recall $r(R, C)$ as the fraction of members from R contained in C, i.e. $r(R, C) = |R \cap C|/|R|$. For each family and a given minimum recall $0.5 < r <= 1$ we can always determine the minimal threshold ϑ such that there is a unique cluster C with $r(R, C) \geq r$. A measure how well the clustering reconstructs the family R is then the associated precision $p(R, C) = |R \cap C|/|C|$. An equal assessment of precision and recall is given with the F-measure:

$$f_{0.5}(R, C) = \frac{2 * r(R, C) * p(R, C)}{r(R, C) + p(R, C)}$$

Table 1 shows the average precision and F-measure weighted by family size for different minimum recall levels between 0.5 and 0.95. If we require that least 70% of a family (= minimal recall level) are grouped within the same cluster level, we get in fact on average a recall of 80%. In this case, we observe on average 32% false positive sequences within this cluster. Of course, we have much better values for some families like 5S rRNA, where we have a precision of 100% at a recall level of 95%. The complete RFAM tree constructed with our method is given as supplementary material.

Concerning the formation of classes comprising several families, this makes mainly sense for classes like tRNAs and miRNAs which have a similar structure, but e.g. not for ribosomal RNAs where there are four structurally different families. The best classification is observed for the class of all tRNAs. They still have a precision of 96% at a recall level of 95%. Concerning the class of all miRNAs, they are (not surprisingly) grouped in several separate cluster. However we have a large cluster comprising 85% of all 213 miRNAs and only 18% false positive sequences.

3.2 Clustering of ncRNA-candidates in *Ciona intestinalis*

The data set resulting from the RNAz-based survey for conserved non-coding RNAs in the genomes of the ascidians *Ciona intestinalis* and *Ciona savignyi* [14] consists of 3332 predicted structured RNAs, of which only about 500 could be annotated as members of well-known RNA families. The overwhelming majority of the known RNAs are the 301 tRNAs recognized by RNAz. Fig. 2 summarizes the results of the clustering procedure.

At the first glance the result might look disappointing as we find a large number of predictions that do not belong to any tight cluster. This is not surprising, however, given that we expect a very high noise level in this data set: (1) The RNAz screen has an estimated false discovery rate of about 18%. (2) No attempts have been made to correct the fairly unreliable strand-prediction of RNAz, which has an error rate up to 30% [32]. (3) We can expect that a significant fraction of structured elements have been predicted only partially. (4) Thermodynamic consensus structure predictions based on only pair-wise alignments are far

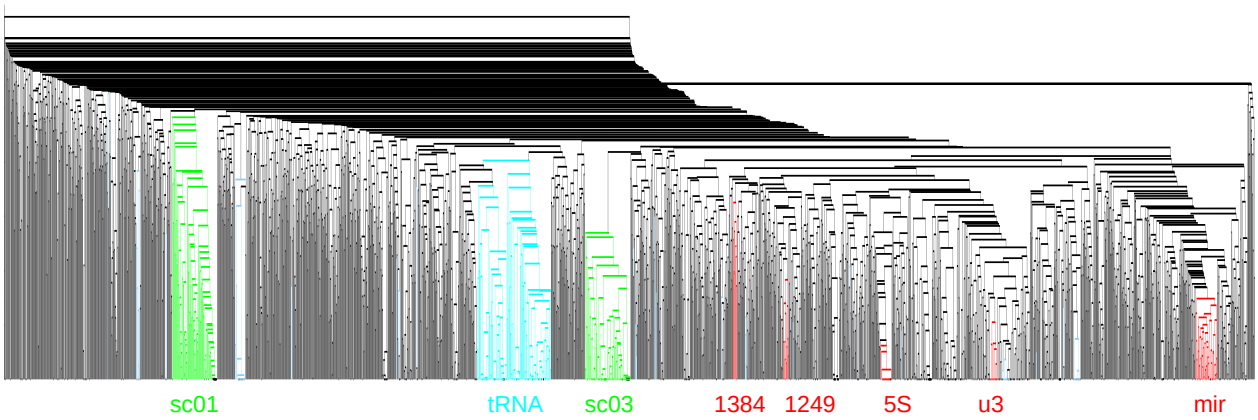

Figure 2: Summary of the clustering procedure. The WPGMA tree contains 3332 putative ncRNAs. A few large, prominent clusters are indicated. Among them are tRNAs and U3 snRNA, and a miRNA cluster, Fig 3, which contains the known miRNAs *mir-124-a/b* and *let-7* as well as candidates for *mir-126* and *mir-7*. Clusters 1384, Fig 4, and 1249, Fig 5, are good candidates for novel ncRNA classes. sc01 and sc03 are both example clusters based on high sequence similarity.

from perfect [33, 19]. It is thus not surprising that only a fraction of the input data can be assigned to meaningful clusters.

As expected, the largest and most prominent cluster is comprised of tRNAs. As discussed in some detail in [34], this tRNA cluster is composed of subclusters corresponding to homologous tRNAs with common anti-codons. Several other well-known multi-gene families are easily identifiable as structural clusters, including the U5 snRNAs, U3 snRNAs and 5S rRNAs. Several families of multi-copy genes with common secondary structure are present in the Ciona genomes [34]. Most of them are also readily identifiable in the structural cluster tree. Since these subclusters are easily detectable already on sequence level, they are of little interest for the structured based approach pursued here.

A more interesting example is a cluster, Fig. 3, that contains two paralogs of *mir-124* and one copy of *let-7* microRNAs that were previously described in computational screens of *Ciona intestinalis* [36, 37], as well as good candidates for *mir-126* and *mir-7*. The other members of the cluster have no sequence similarity with known microRNA families compiled in `miRBase release 9.0` (`blast` $E \leq 0.001$). Both *mir-124* candidates occur within introns of known mRNAs of `Ciona intestinalis` (JGI2.0), while *mir-126* and *mir-7* do not seem to be located in an intron. That a large cluster of known and putative miRNAs was detected demonstrates that annotation of ncRNA candidates is highly improved by structure based clustering. The majority of cluster members could not be identified as miRNA candidates by sequence comparison alone [14]. Further a comprehensive comparative screen for miRNAs across the metazoan species identified only few homologs with high sequence similarity within the urochordates [37] raising the question if there may exist a group of yet unknown miRNA families within the urochordates.

The following two panels 4 and 5 highlight two novel clusters of structurally similar predictions for which no functional or class assignment is available. The neighbor-net graphs in the insets show the sequence distance within the example cluster. Since the sequence distance is on average larger than 0.5, this confirms that the clusters are defined essentially based on structural similarities. While our examples usually contain some subsets of related sequences, overall there is little or no detectable sequence conservation so that the clusters could not have been detected by sequence similarity alone. Since many ncRNAs, in particular snRNAs, tend to form multi-gene families (often evolving under some form of concerted evolution that keeps the family members nearly identical), a moderate copy number in the genome can be interpreted as supporting the hypothesis that the candidate is indeed a true ncRNA.

In cluster 1384, Fig 4, for example, sequences with a well conserved secondary structure but low sequence similarity are grouped. 9 of 11 sequences of cluster 1384 could be exactly mapped to the new *Ciona intestinalis* assembly JGI2.0. The structural cluster contains three sub-clusters, 1378, 1381 and 1383, that have overall structural features in common. All sub-clusters have three stem loops originating from one single multiloop as consensus structure. But their length and number of internal loops differ. Their grouping into the super-clusters 1382 and 1384 are justified by compensatory mutations. Two

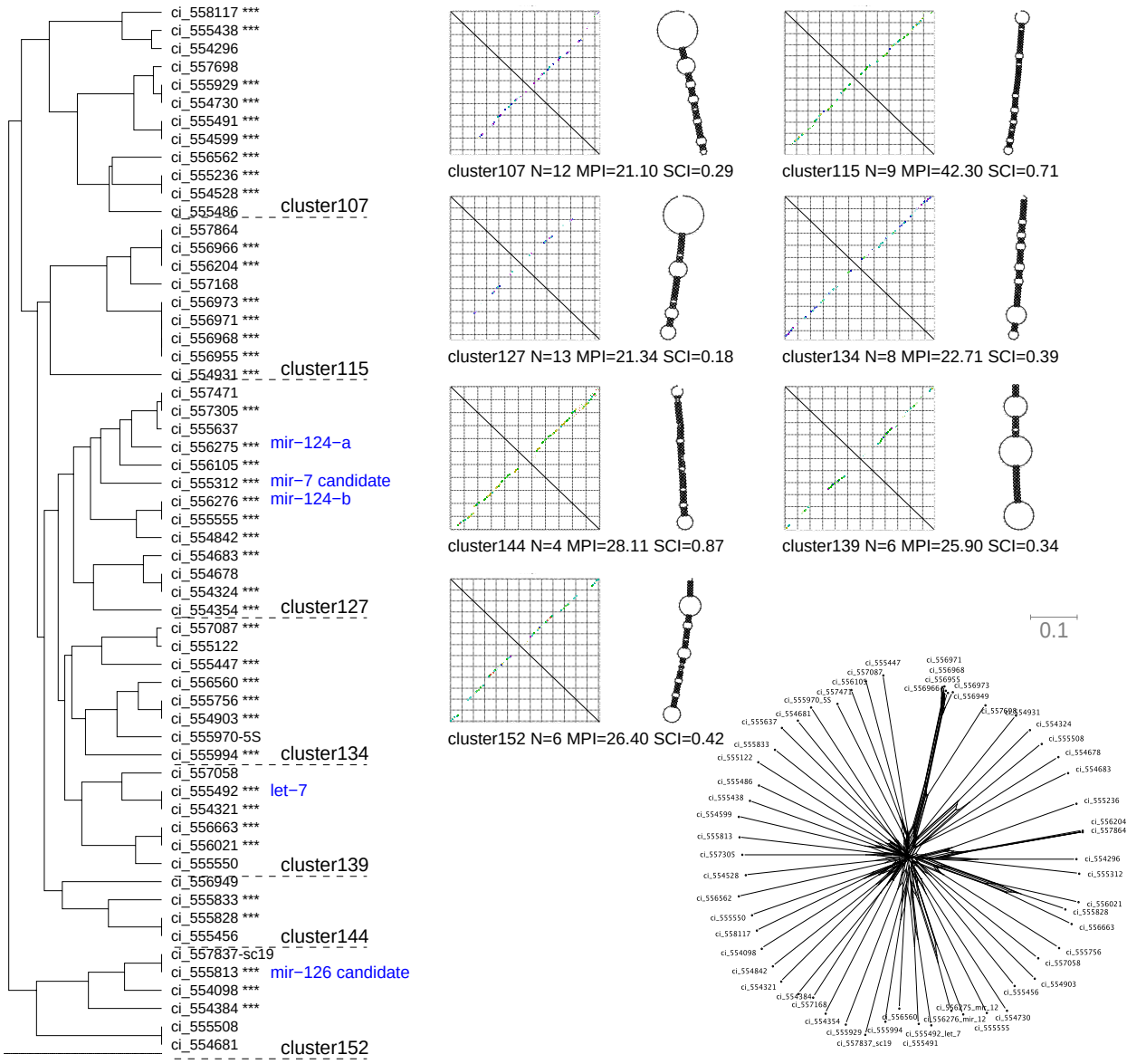

Figure 3: Cluster containing known and predicted *Ciona intestinalis* microRNAs. The two known *mir-124* paralogs are members of sub-cluster 127. Whereas the known *let-7* is found in sub-cluster 139. Sequence ci_555813 in sub-cluster 152 contains a *mir-126* candidate (UCGUACCGUGAGUAAUAAAGC) and ci_555312 in sub-cluster 127 a *mir-7* candidate (UGGAAGACUAGUGAUUUUGUUGU). 40 of the 58 cluster members (marked with ***) are classified as putative microRNAs by `RNAmicro` [35]. The fourth known microRNA in urochordates, *mir-92* does not fall into this structural cluster. Members of the cluster are no sequence related (`NeighborNet` in the bottom right corner). N...number of sequences in cluster. MPI...mean pairwise identity of multiple alignment. SCI...structure conservation index.

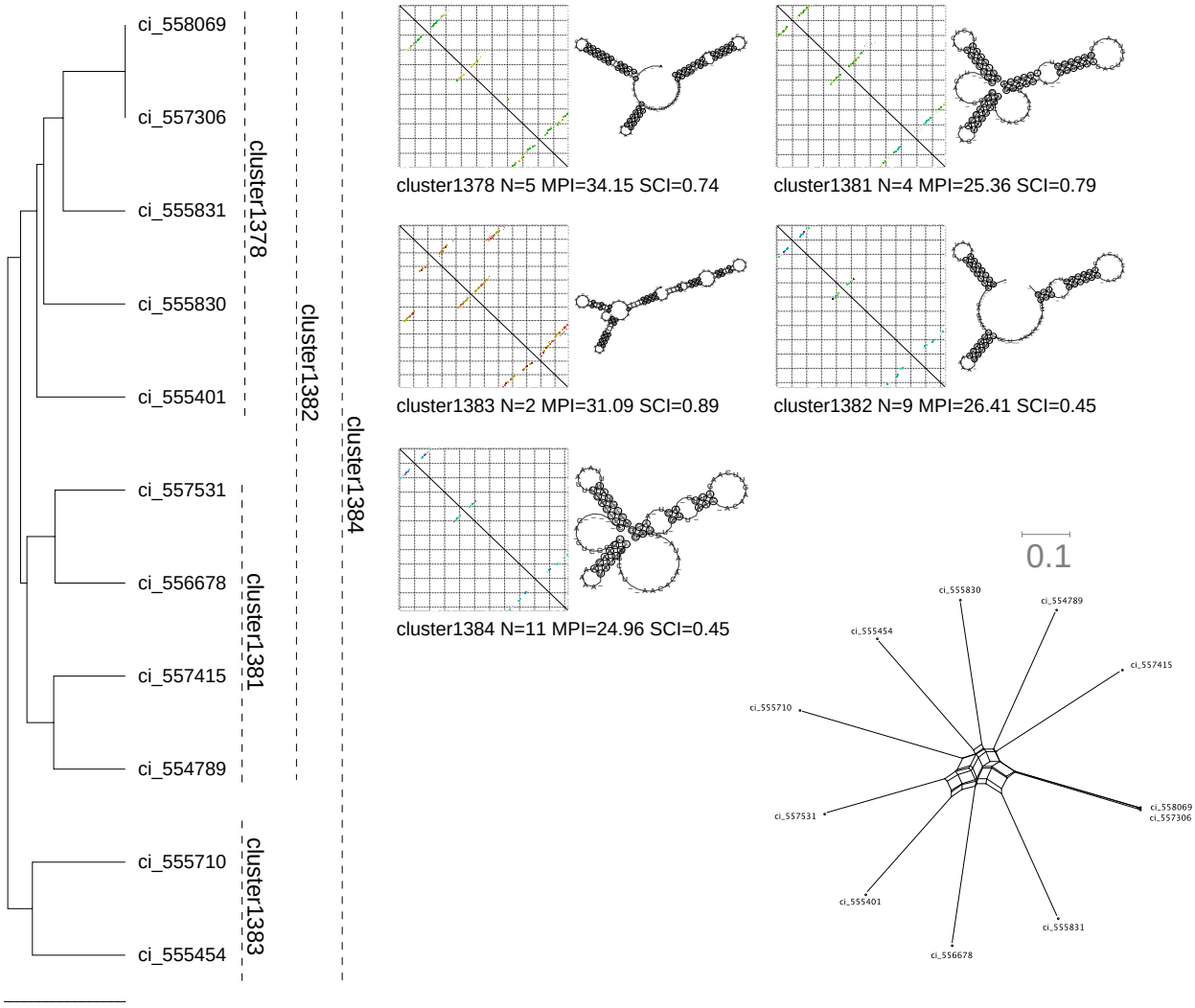

Figure 4: Cluster 1384 groups sequences with a well conserved secondary structure consisting of three stem loops. Whereas the sequence identity is low we observe a high structural conservation. N...number of sequences in cluster. MPI...mean pairwise identity of multiple alignment. SCI...structure conservation index.

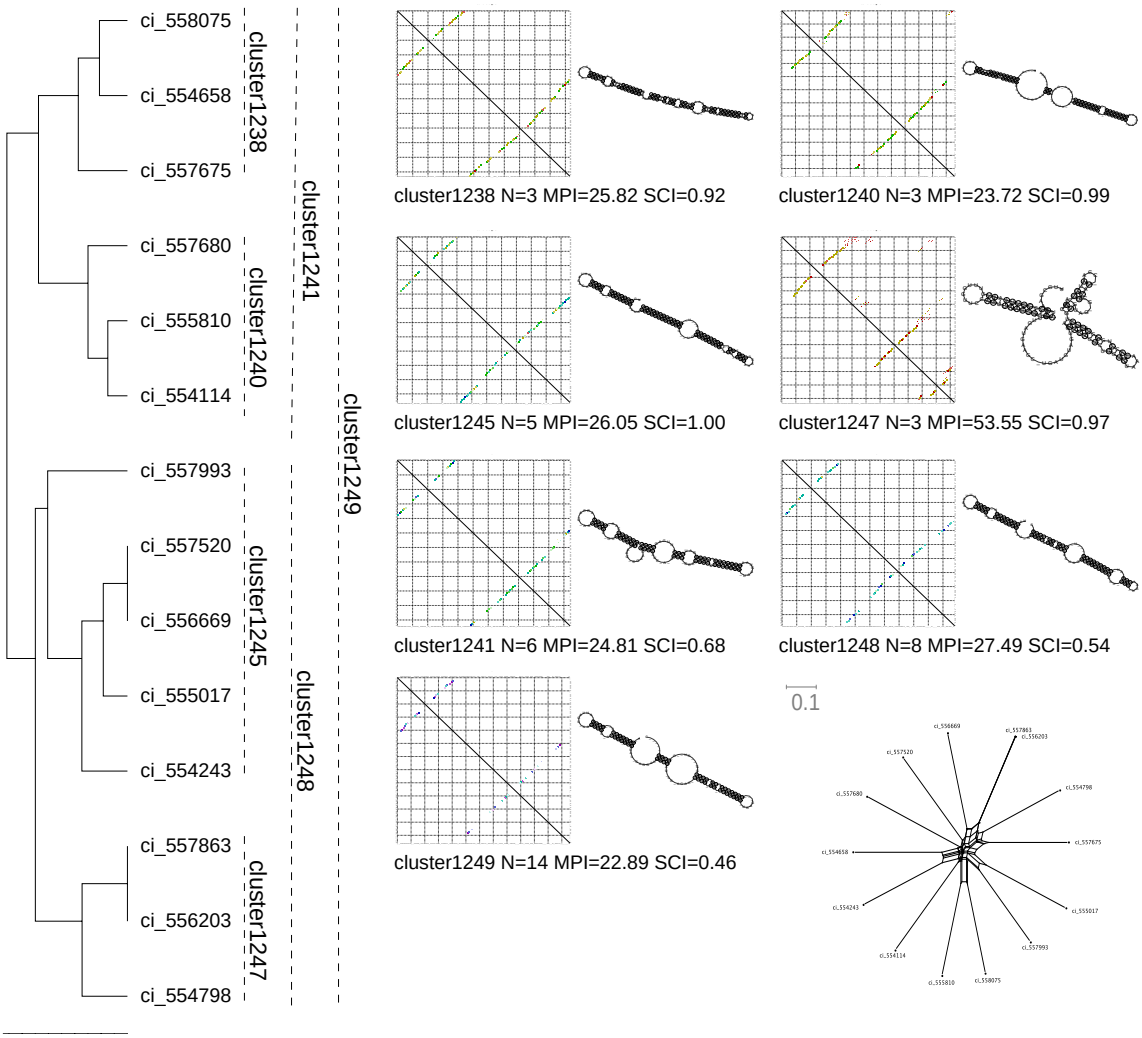

Figure 5: Example of structure based clustering of very diverse sequences which might form a novel ncRNA class. The consensus structure models thus show a large number of compensatory mutations. N...number of sequences in cluster. MPI...mean pairwise identity of multiple alignment. SCI...structure conservation index.

sequences of sub-cluster 1378 and one of sub-cluster 1381 appear within an intron of Ciona-mRNA AK113484. Whereas the two sequences of sub-cluster 1378 appear within the same copy of mRNA AK113484 on chr01p, the sequence in sub-cluster 1381 occurs in a copy on chr04q. Six different genomic copies of AK113484 exist in JGI2.0 but none of the intronic regions where the ncRNA candidates are found are associated with repeats. This allows the conclusion that those ncRNA candidates are indeed functional ncRNAs as their sequences are highly diverged whereas they share common structural features and appear within the same Ciona-mRNA. One sequence of sub-cluster 1383 occurs in an exon of the known protein coding Ciona-mRNA AK114007. All other elements are intergenic or at least the corresponding mRNAs are not yet known.

Cluster 1249 is also composed of highly divergent sequences but similar secondary structures. Two sequences of sub-cluster 1247 appear within an intron of the Ciona-mRNA AK174830. Sub-clusters 1238 and 1245 contain one sequence occuring in an intron of Ciona-mRNA AK222260 and AK116291, respectively.

Clusters 1384 and 1249 are good candidates for novel classes of urochordate-specific ncRNAs, since none of the sequences has detectable ncRNA homologs in vertebrates.

3.3 Clustering of ncRNA-candidates in *Gammaproteobacteria*

A RNAz screen of six related gammaproteobacteria resulted in a ncRNA candidate set of 123 unique loci of the reference organism *Escherichia coli.* The screen follows the same pipeline as in [14, 15] but includes a new approach to built multiple alignments. Only alignments with homolog sequences of at least three genomes, with maximal pairwise blast e-values of $1e-10$ and a minimal length of 40nt were retained for input to the RNAz-pipeline.

That the majority of ncRNA candidates could be annotated with known *Escherichia coli* ncRNAs (labeled with *EC[...]* in Fig. 6.) is not surprising as the screen was set up with a restrictive e-value for the initial blast search. Further only candidates with homologs in at least three gammaproteobacteria genomes are reported. This provides us with a second ncRNA candidate set to validate the clustering approach, which in contrast to the RFAM seed sequences in section 3.1 was detected by RNAz. A candidate was annotated to be a known *Escherichia coli* ncRNA if their genomic regions overlap to at least 70%. If such an annotation was not available a blast search against the RFAM database ($E < 1e-6$) identified further homolog ncRNAs.

In Fig. 6 the complete WPGMA tree is depicted. It is nicely seen that again tRNAs get grouped in one separate cluster. Even tRNAs coding for the same nucleotide are mostly found within the same subclusters. Different families of rRNAs appear also in several separate clusters, although there exist none single cluster for each family.

4 Discussion

Genome-wide studies, both experimental and computational, have uncovered tens of thousands of transcripts in higher Eukaryotes, that have little or no protein-coding capacity. For a large subset of these, there is evidence for selection acting to preserve secondary structure motifs. Many classes of functional RNAs, on the other hand, can be recognized based on structural similarities. It is thus natural to ask if the available data contain evidence for novel families and classes of structured RNAs, for which so far no representative has been well characterized experimentally. To answer this question, it is necessary to cluster the candidate RNAs based on their structural features, a task that is computationally much harder than clustering based on sequence similarity.

We present here a new tool, LocARNA, which implements a novel, more efficient variant of the Sankoff algorithm. We have demonstrated that LocARNA is fast enough to make structure-based clustering of thousands of putative structured RNAs feasible. The main reason for its superior efficiency is due to the pre-filtering of the basepairs by their probability, and an efficient computation scheme that is able to profit from the reduced number of basepairs considered. The method is also robust enough to find significant clusters in fairly noisy, realistic data that contain a substantial fraction of false positive predictions. We have successfully tested the tool on the sequences of the RFAM seed alignments.

The LocARNA implements a local sequence structure alignment method, which is required when applied to candidate ncRNA sequences where the exact region of interest is not exactly known (of course, the tool can also be applied to global alignment problems). Clearly, there is a length dependency in the scores, which has several sources, one being the calculation of pair probabilities. This influences both pairwise alignment and the clustering, which implies that the ncRNAs to be clustered should not diverge

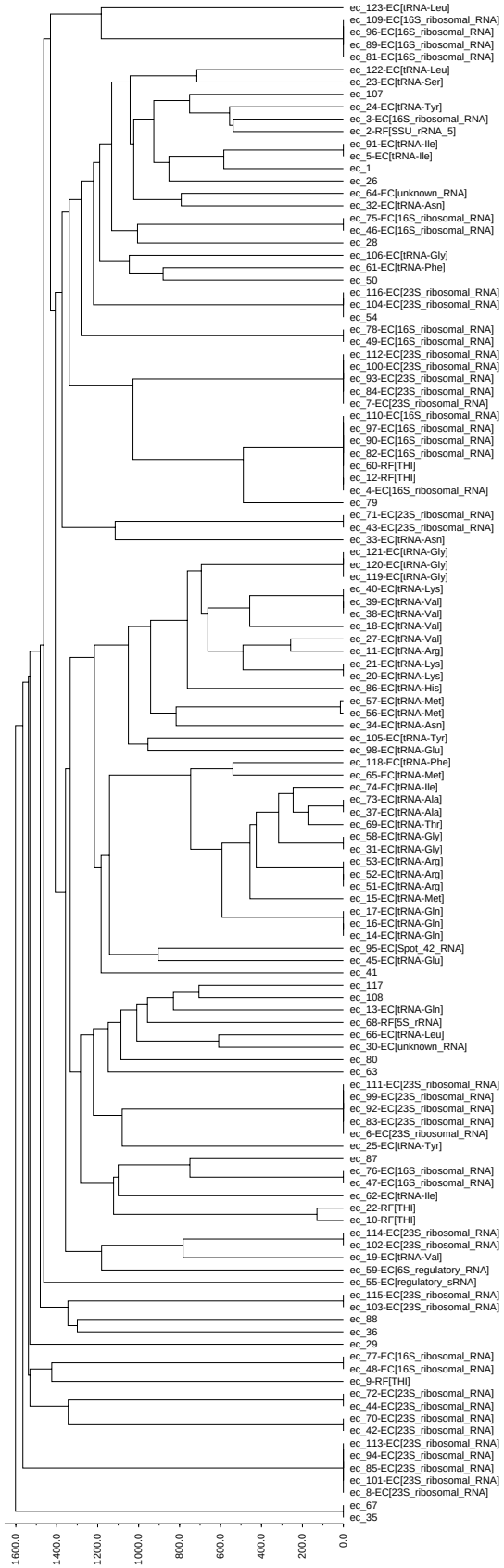

Figure 6:

The complete WPGMA clustering tree for ncRNA candidates in *Gammaproteobacteria Escherichia coli.* Candidates are annotated with known *Escherichia coli* ncRNAs (*EC[...]*) or if such not exist with ncRNAs from the RFAM database (*RF[...]*).

to much in length. This is the case in many applications like the clustering of predicted ncRNAs. A more precise treatment of the different kinds of dependencies (like GC-content) is planned for a future version.

The application of the tool to a dataset of more than 3000 predicted structured RNAs in urochordates showed that the clustering approach not only recovers known RNA families and classes such as tRNAs, but also predicts several candidates for novel ncRNA classes. In some cases we find that additional sequences are identified as structural relatives of known RNA families. In this way we have, for example, identified a *mir-126* and a *mir-7* homolog which were not detected in previous computational studies. More importantly, however, we also find structure-based clusters that are candidates for novel, presumably urochordate-specific, RNA classes. We find that these clusters often contain sub-clusters consisting of multi-copy sequences. Comparing this with the characteristics of several well-studied ncRNA families, in particular tRNAs, the snRNAs associated with the major spliceosome, and SL RNAs lends further credibility to the hypothesis that these sequences indeed form a *bona fide* RNA class.

Supporting Information

See `http://www.bioinf.uni-freiburg.de/Supplements/locarna-06-12`.

Acknowledgments

We thank Dominic Rose and Jana Hertel for the data of the `RNAz` screen of gammaproteobacteria. We also would like to thank the anonymous referees for many helpful comments. This work has been funded, in part, by the Austrian GEN-AU projects “bioinformatics integration network II” and “non-coding RNA”, the German DFG Bioinformatics Initiative BIZ-6/1-2, and the Federal Ministry of Education and Research in the context of the Jena Center for Bioinformatics.

References

[1] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. Science 294:853–857.

[2] Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. Science 294:858–862.

[3] Lee R, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294:862–864.

[4] Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

[5] Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154.

[6] Bertone P, Stoc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242–2246.

[7] Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8 [epub].

[8] Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 342:19–39.

[9] Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102:2454–2459.

[10] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. (2006) Classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2:e33.

[11] Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res 16:885–889.

[12] Uzilov AV, Keegan JM, Mathews DH (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 7:173 [epub].

[13] Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. Nature Biotech 23:1383–1390.

[14] Missal K, Rose D, Stadler PF (2005) Non-coding RNAs in *Ciona intestinalis*. Bioinformatics 21 S2:i77–i78. Proceedings ECCB/JBI'05, Madrid.

[15] Missal K, Zhu X, Rose D, Deng W, Skogerbø G, et al. (2006) Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. J Exp Zool B: Mol Dev Evol 306:379–392.

[16] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) `Rfam`: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121–D124.

[17] Griffiths-Jones S (2004) The microRNA Registry. Nucl Acid Res 32:D109–D111.

[18] Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress AWM, et al. (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. Science 244:673–679.

[19] Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33:2433–2439.

[20] Siebert S, Backofen R (2005) `MARNA`: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics 21:3352–3359.

[21] Hofacker IL, Bernhart SHF, Stadler PF (2004) Alignment of RNA base pairing probability matrices. Bioinformatics 20:2222–2227.

[22] Höchsmann M, Töller T, Giegerich R, Kurtz S (2003) Local similarity in RNA secondary structures. In: Proc of the Computational Systems Bioinformatics Conference, Stanford, CA, August 2003 (CSB 2003). pp. 159–168.

[23] Backofen R, Will S (2004) Local sequence-structure motifs in RNA. Journal of Bioinformatics and Computational Biology (JBCB) 2:681–698.

[24] Backofen R, Siebert S (2005) Fast detection of common sequence structure patterns in RNAs. Journal of Discrete Algorithms Accepted.

[25] Sankoff D (1985) Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. SIAM J Appl Math 45:810–825.

[26] Gorodkin J, Heyer L, Stormo G (1997) Finding common sequences and structure motifs in a set of RNA molecules. In: Gaasterland T, Karp P, Karplus K, Ouzounis C, Sander C, et al., editors, Proceedings of the ISMB-97. Menlo Park, CA: AAAI Press, pp. 120–123.

[27] Hull Havgaard JH, Lyngsø R, Stormo GD, Gorodkin J (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 21:1815–1824.

[28] Mathews D, Turner D (2002) Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317:191–203.

[29] McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105–1119.

[30] Holmes I (2005) Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics 6:73 [epub].

[31] Dowell R, Eddy SR Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. Preprint, *http://selab.wustl.edu/index.html*.

[32] Missal K, Stadler PF `RNAstrand`: Reading direction of structured RNAs in multiple sequence alignments. Preprint, *http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/06-006.pdf*.

[33] Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059–1066.

[34] Bompfünewerer AF, Backofen R, Bernhart SH, Flamm C, Fried C, et al. (2006) RNAs everywhere: Genome-wide annotation of structured RNAs. J Exp Zool B: Mol Dev Evol In press.

[35] Hertel J, Stadler PF (2006) Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data. Bioinformatics 22:e197–e202. ISMB 2006 contribution.

[36] Legendre M, Lambert A, Gautheret D (2005) Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 21:841–845.

[37] Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, et al. (2006) The expansion of the metazoan microRNA repertoire. BMC Genomics 7:15 [epub].