BIOINFORMATICS
Vol. 00 no. 00 2008
Pages 1–4

Structural profiles of human miRNA families from pairwise clustering

Bogumił Kaczkowski[1,†], Elfar Torarinsson[1,†], Kristin Reiche[2], Jakob Hull Havgaard[1], Peter F. Stadler[3,2,4,5], and Jan Gorodkin[1*]

[1]Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Frederiksberg C, Denmark
[2]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstr. 1, 04103 Leipzig, Germany
[3]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstrasse 16-18, Leipzig, Germany
[4]Santa Fe Institute, Santa Fe NM87501, USA
[5]Institute of Theoretical Chemistry, University of Vienna, Währingerstr. 17, A-1090 Vienna, Austria

[†]these authors contributed equally
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

MicroRNAs are a group of small, ∼21nt long, riboregulators inhibiting gene expression at a posttranscriptional level. Their most distinctive structural feature is the foldback hairpin of their precursor pre-miRNAs. Even though each pre-miRNA deposited in miRBase has its secondary structure already predicted, little is known about the patterns of structural conservation among pre-miRNAs. We address this issue by clustering the human pre-miRNA sequences based on pairwise, sequence and secondary structure alignment using FOLDALIGN, followed by global multiple alignment of obtained clusters by WAR. As a result, the common secondary structure was successfully determined for four FOLDALIGN clusters: the RF00027 structural family of the Rfam database and three clusters with previously undescribed consensus structures.

Contact: gorodkin@genome.ku.dk

1 INTRODUCTION

MicroRNAs (miRNAs) are a group of small, 21-23nt long, non-protein coding RNAs, which negatively influence gene expression at the post-transcriptional level (reviewed in Maroney et al., 2006). Recent research has provided growing evidence of the profound role of miRNAs in cancer, stem cell, and various diseases (reviewed in Zhang et al., 2006; Kato and Slack, 2008). Target prediction studies demonstrate that miRNAs might be involved in regulating the expression of as much as one third of the human gene set (Lewis et al., 2005).

The biogenesis of miRNAs begins with transcription by RNA polymerase II. The product, the primary miRNA (pri-miRNA), is processed by the RNase III *Drosha* to a precursor miRNA (pre-miRNA). Pre-miRNAs have distinctive fold-back hairpin structures with an estimated average length of about 70nt. Pre-miRNAs are transported to the cytoplasm where they are cleaved by RNase III Dicer. Mature, 21-23nt long miRs enter the miRNA-induced silence complex (miRISC), which then inhibit expression of target mRNAs by either cleaving the mRNA or preventing translation (reviewed in Maroney et al., 2006; Kato and Slack, 2008).

MicroRNAs have been grouped into families based on sequence conservation of the hairpin sequences deposited in miRBase (Griffiths-Jones et al., 2006) using a manually curated BLAST clustering, and some pre-miRNAs have been grouped together in 46 different families by Rfam (Griffiths-Jones et al., 2003) based on sequence and structure. A systematic investigation of the structural variability of pre-miRNAs, however, is still lacking despite the enormous growth in miRNA-related literature (Lindow and Gorodkin, 2007).

We present here a systematic study of combined sequence and structure similarities among human pre-miRNAs using FOLDALIGN (Havgaard et al., 2007) with the aim of identifying clusters that correspond to additional miRNA families with well-defined sequence and structure conservation beyond the mature miRNA. There are at least two mechanisms that may have lead to a significant clustering of miRNA structures. Since structural evolution is usually slower than sequence evolution (Schuster et al., 1994), structural clusters can reveal common ancestry of families whose sequences have already diverged beyond recognition. On the other hand, there is evidence that the processing of pre-miRNAs is specifically regulated; reviewed by Schmittgen (2008). These differential processes potentially may have caused selection of distinctive structural features that are involved in discriminating regulatory interactions. The analysis of the structural differences between plant and animal pre-microRNA stem-loops has already provided some insights into the mechanisms of microRNA biogenesis (Rabani et al., 2008). A detailed structural classification of miRNA precursors thus not only helps fill the gap in the information provided by Rfam and miRBase, respectively, but also may provide insights into the intricacies of miRNA processing. In *Drosophila*, Argonaute protein association is mediated by the secondary structure of the miRNA precursor (Förstemann et al.,

*to whom correspondence should be addressed

2007; Tomari et al., 2007). Whether secondary structure also affects Argonaute loading in mammals remains to be discovered (Peters et al., 2007; Farazi et al., 2008).

In order to cluster human pre-miRNAs based on sequence and structure, we extracted all experimentally verified mature miRNAs in human from miRBase, version 10.0 (Griffiths-Jones et al., 2006), added the flanking regions to include the hairpin structure, and ran FOLDALIGN (Havgaard et al., 2007) on all pairwise combinations. This resulted in a score for all pre-miRNA pairs based on sequence and structure. These FOLDALIGN scores were used to cluster the pre-miRNAs using the R environment package (Ihaka and Gentleman, 1996) Pvclust (Suzuki and Shimodaira, 2006). In order to check the stability of the clustering, we also employed a different approach to extract relevant clusters from the hierarchical cluster-tree adapted from the *Duda rule* (Duda et al., 2001).

Since the FOLDALIGN clusters are based on sequence and structure conservation, all clusters containing at least 4 pre-miRNAs were extracted and the WAR webserver (Torarinsson and Lindgreen, 2008) was used to perform multiple alignments to see if any of the clusters had clear and well defined secondary structure.

2 METHODS

The pre-miRNAs deposited in miRBase 10.0 differ in length. The reason is that there is little experimental evidence on pre-miRNA's exact ends so that precursor sequences are taken from genomic context with flanking sequences that give pre-miRNA a length from 60 to about 120 nt. To eliminate the impact of varying flanks on the alignments, we re-extracted the pre-miRNA sequences from their genomic context with uniform flank lengths. More specifically, we extracted experimentally verified mature miRNAs from pre-miRNAs containing only a single mature sequence. If the mature miRNA was in the $5'$ stemloop precursor arm, we added 20 nts upstream and 80 nts downstream from the mature miRNA $5'$ end, otherwise we added 20 nts downstream and 80 nts upstream from the mature miRNA $3'$ end.

Local alignments were computed using FOLDALIGN (Havgaard et al., 2007), a variant of the Sankoff algorithm (Sankoff, 1985), which simultaneously uses sequence and structure information. We compared all-against-all pairs of the 427 extracted human miRNAs. The resulting scores where then clustered using the R statistical environment package Pvclust (Suzuki and Shimodaira, 2006) (see Supp. Mat.[1] for details). In Pvclust, for each cluster in hierarchical clustering, p-values are calculated via multiscale bootstrap re-sampling. The agglomerative method, average linkage, was used and 10,000 bootstrap replications were run, with relative sample size were set from 0.5 to 1.4, incrementing in steps of 0.1.

For a cluster with probability $p \geq 0.95$, the hypothesis that "the cluster does not exist" is rejected with significance level 0.05; roughly speaking, we can think that these clusters not only "seem to exist" due to sampling error, but may stably be observed if we increase the number of observation (Suzuki and Hayashizaki, 2004). Using the *pvpick* function of Pvclust, we extracted all clusters with $p \geq 0.95$.

In addition to extracting highly significant clusters by Pvclust we retrieved a complete partition of the hierarchical cluster-tree into distinct subtrees by applying an adaption of the "*Duda rule*" (Duda et al., 2001). A complete partition of the cluster-tree yields a wider range of evidence for biologically relevant structural clusters of human miRNAs. The hierarchical cluster-tree is traversed starting from the leaves towards the root testing each internal node whether the two subtrees defined by this node define two distinct clusters or are believed to be members of the same cluster. This decision is based on evaluating the sum of squared errors of the minimum free energies of all miRNAs in the subtree relative to the minimum free energy of the consensus secondary structure of this subtree. (For a detailed description of this method see the Supp. Mat.).

To evaluate the agreement of these clusters with the miRBase (Griffiths-Jones et al., 2006) and Rfam (Griffiths-Jones et al., 2003) classification is quantified by the Matthews correlation coefficient (MCC). We define true positives (TP) as those pairs of miRNAs that are clustered together both by our clustering and the reference while true negatives cluster differently in both approaches. A pair is a false positive (FP) if clustered together by us but appears separately in the reference. Conversely, a false negative pair (FN) is clustered together in the reference but separated in our approach. The MCC is a value between $+1$ (perfect correlation) and -1 (perfect anticorrelation), with 0 indicating uncorrelated (random) data (Matthews, 1975).

3 RESULTS AND DISCUSSION

We obtained 42 Pvclust clusters with $p \geq 0.95$, containing 220 miRNAs. The correlation with Rfam and miRBase was quite good with MCCs of 0.76 and 0.74 for Rfam and miRBase, respectively. Of these 220 miRNAs, 148 miRNAs were present in only ten clusters of more than three members.

The adapted Duda rule resulted in a partition of all human 427 miRNAs into 60 clusters. Each of the Pvclust clusters was completely contained in a single Duda cluster. The clusters defined by the adapted Duda rule have lower values for MCCs (0.46 for both Rfam and miRBase), because they comprise the entire cluster-tree, not just the most highly significant clusters. It should of course be kept in mind that Rfam and miRBase contains miRNAs from many organisms whereas we only cluster human pre-miRNAs, hence we only compare to human sequences in Rfam and miRBase.

The ten FOLDALIGN clusters with $p \geq 0.95$ and containing more than three miRNAs were selected and subjected to multiple alignment and structure prediction to study if there was a common well defined secondary structure underlying the clustering. Global multiple alignment was performed using the webserver WAR (Torarinsson and Lindgreen, 2008), which uses seven different programs to perform multiple alignment and secondary structure predictions of the given sequences.

We are quite strict and define good clusters as those clusters where at least 3 of the seven programs in WAR agreed well on both the multiple alignment and the corresponding predicted secondary structures, as shown in the consensus heatmap generated by WAR (see Supp. Mat.). It may therefore be the case that we erroneously reject good clusters for which one or more of the programs predicted stable structures because of a lack of agreement between the programs. Out of the ten clusters, four clusters, containing 70 miRNAs, satisfied these stringent criteria, Figure 1. These four clusters were also the clusters with the highest average pairwise sequence identity, which in general was very low for the ten clusters. Details, such as size of the clusters and overlap with miRBase and Rfam for all ten clusters are compiled in table 1.

Three of the clusters, **A**, **C**, and **I**, corresponded quite well to known miRBase families. Cluster **C** is listed in Rfam, annotated with the same secondary structure as predicted by the WAR consensus. Clusters **A** and **I**, on the other hand, are not contained in Rfam, hence it was previously unknown whether they share a common secondary structure as well as having a conserved sequence. Among the 25 members of cluster **I**, 13 belong to the miRBase family mir-154. On these data, we also performed a pure sequence based clustering using SSearch (Pearson, 1991), which clustered 22 of the 25 members together. Indeed, cluster **I**

[1] `http://genome.ku.dk/resources/mirclust`

Table 1. Overview of the ten Foldalign/Pvclust clusters subjected to global multiple alignment on web server WAR. In parenthesis the ID and size of the corresponding Duda cluster is given. We list membership in Rfam and miRBase families as well as the average pairwise identity (API), the quality of the cluster, and (in parenthesis) the number of programs (out of seven) from which the consensus structure was determined.

Cluster	**A** (8)	**B** (41)	**C** (18)	**D** (38)	**E** (36)
Cluster size	26 (31)	23 (63)	9 (19)	4 (6)	8 (9)
API	0.79	0.23	0.59	0.33	0.29
Rfam families	not present	not present	7 in RF00027	not present	RF00245,-074,-258
miRBase families	26 in mir-515	mir-548, -32, -122	9 in let-7	4 various	mir-19, -22, 2 various
Consensus structure	Well defined (4)	Not found	Well defined (5)	Not found	Not found
Cluster	**F** (12)	**G** (28)	**H** (18)	**I** (6)	**J** (60)
Cluster size	10 (10)	20 (20)	5 (19)	25 (25)	18 (26)
API	0.62	0.22	0.25	0.56	0.26
Rfam families	not present	not present	not present	not present	not present
miRBase families	mir-506, -509, -892	mir-392, -374, 8 various	4 various	mir-154,-368,-329,-379	mir-941, -744, -484
Consensus structure	Well defined (7)	Borderline	Not found	Well defined (3)	Not found

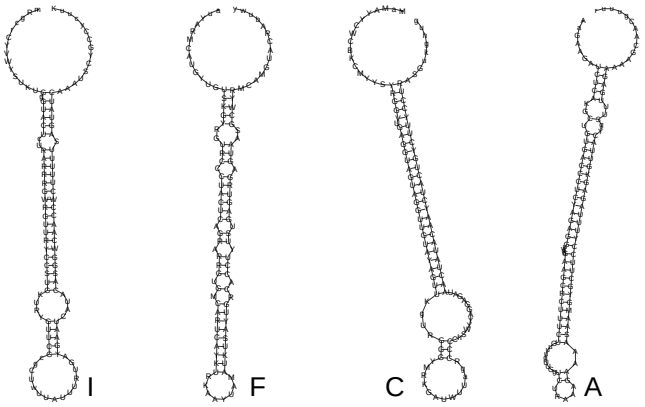

Fig. 1. The consensus secondary structure, as predicted by the WAR consensus, for four selected clusters

comprises a subclass of the **mir-134** "supercluster" on Chr.14, for which common ancestry has been proposed based on faint sequence similarities (Hertel et al., 2006).

Cluster **F** was neither in miRBase nor Rfam. It is predicted to have highly similar secondary structure by all seven programs, in good agreement with each other. Five of the programs in WAR predict very stable consensus secondary structures for cluster **B** but they do not completely agree on their predictions, probably affected by the very low average pairwise identity, so we do not consider cluster **B** as a good cluster.

The Duda-rule clustering agreed very well with our good clusters **A**, **F** and **I**. Furthermore, it agreed perfectly with **G**, which we do not define as a good cluster since the programs in WAR do not agree. Still, both FoldalignM and RNASampler predict a stable consensus structure for **G**, so this cluster is likely valid. The Duda-rule clustering merged **C** and **H**. Upon inspection, we do not believe this to be correct because cluster **C**, the let-7 family, agrees well with an already defined Rfam cluster, and there do not seem to be any data that link the members of cluster **H** to the let-7 family either evolutionary or in terms of function. The Duda-rule clustering also predicts several clusters not predicted by Pvclust. In general, these were small and of extremely low sequence similarity. We believe that most if not all of them are spurious results. The best Duda-rule clusters missing from Pvclust, 15 and 34, have a well-defined WAR consensus. Cluster 15 contains the mir-103 and mir-107 families, which are obvious paralogs. Further details on all clusters, including their WAR results, are available in the Supp. Mat.

For each of the three novel clusters with well defined consensus alignment and structure (**A**, **F**, and **I**; **C** was previously known) we constructed a covariance model using *cmbuild* from the Infernal package (Nawrocki and Eddy, 2007). Flanking ends which were not part of the common predicted structure were removed before running *cmbuild*. Together with the search engine Infernal, these covariance models provide a very sensitive and discriminative tool for homology search that is much more specific than other computational approaches for miRNA discovery. Using RaveNnA (Weinberg and Ruzzo, 2006), which is basically a fast HMM based filter to speed up Infernal, we searched the whole human genome for new instances of our three clusters at an E-value cutoff $< 10^{-5}$. In comparison to miRBase, this resulted in 49, 27 and 16 additional members of **A**, **F**, and **I**, respectively, (see Supp. Mat.). One of the additional miRNAs for cluster **A** is also contained in the corresponding Duda cluster 8. Within each of these three clusters, both the known miRNAs and the Infernal-predicted candidates are highly spatially clustered, with the vast majority of members located in the same region of the same chromosomes. This lends further credibility to these predictions.

Interestingly, more than half of the new members of cluster **A** are close to being located perfectly antisense to the known members. This does not necessarily imply that these are new functional miRNAs, although not impossible, but rather this could indicate that the pre-miRNA in cluster **A** is of palindromic origin. In plants, a connection between short miniature inverted-repeat transposable elements (MITEs) and miRNAs has been reported (Mette et al., 2002). More recently Piriyapongsa and Jordan (2007) reported that the human miRNA genes, hsa-mir-548, are derived from the MITE elements, Made1, which consist of two 37 base pair (bp) terminal inverted repeats that flank 6 bp of internal sequence. Thus, Made1 elements are nearly perfect palindromes, and when expressed as RNA they form highly stable hairpin loops. Furthermore, Piriyapongsa and Jordan (2007) discuss a principle

whereby full length DNA-type transposable elements that encode multiple siRNAs become degraded into short non-autonomous MITEs, which if transcribed can form short hairpins that are processed to yield single mature miRNA sequences. Near perfect, antisense, high scoring matches were also seen in clusters **F** and **I**, although not nearly as many as for cluster **A**, probably due to the significantly higher sequence conservation of cluster **A**.

The known cluster **C** has quite well conserved mature miRNA sequences, all located in the 5' arm of the hairpin. In contrast, for clusters **A**, **F** and **I** a clear consensus mature sequence is absent. However, these clusters contain the mature sequence in the 5' arm and the 3' arm. For clusters **A** and **F** the mature sequences are conserved within the the 5' arm and 3' arm, respectively. Therefore, it seems natural to speculate that the dominant mature product has switched from one arm to the other in some of the cluster members. This hypothesis is further supported by the observation that, in these clusters, the location of the mature miRs is quite well-conserved on both arms (Supp. Mat.). A miR-switching scenario also strongly supports the conservation of the hairpin structure and is particularly plausible for appromixately palindromic precursors, which we observe for exactly these three clusters. For cluster **I**, the majority of the matures are located in 3' arm, but less well conserved. However, it has been observed that matures of the 3' arm are over-all less conserved (Gorodkin et al., 2006).

CONCLUDING REMARKS

Clustering of the human miRNAs based on both sequence and structure leads to the identification of four significant clusters, of which only one was known previously to have a conserved structure. Using a covariance model for each of these new clusters allowed us to predict 92 new miRNAs, many of which are located antisense, indicating a possible palindromic origin of the clusters. In one case, cluster **I**, the structural clustering corroborates earlier suggestions of common ancestry based on both faint sequence similarities and the location in a single genomic location. At least in some cases the structural conservation could be explained by migration or switching of the dominant mature miR to the other arm of the precursor, a mechanism by which highly divergent mature microRNAs can originate from a common ancestral precursor.

Our analysis suggests that large structural clustering can help to uncover ancient homologies and provides some additional information on the early evolution of microRNA families. Limited to human miRNAs only, we have not encountered evidence for structural subclasses that could be associated with processing differences of the precursors. It will be interesting to see, however, if an extension of this work to covering the complete collection of metazoan microRNAs will reveal phylum-specific differences and/or clusters that cannot be explained by common ancestry.

ACKNOWLEDGMENT

This work was supported by Danish Research Council for production and technology and the Danish Center for Scientific Computation and by 6th Framework Programme of the European Union (SYNLET). The work on the Duda rule has been done by RK while she was a member of the Bioinformatics Group, University of Leipzig.

REFERENCES

R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. John Wiley & Sons, Inc., 2001.

T. A. Farazi, S. A. Juranek, and T. Tuschl. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, 135: 1201-1214, 2008.

K. Förstemann, M. D. Horwich, L. M. Wee, Y. Tomari, and P. D. Zamore. Drosophila microRNAs are sorted into functionally distinct Argonaute complexes after production by Dicer-1. *Cell*, 130:287-297, 2007.

J. Gorodkin, J. H. Havgaard, M. Ensterö, M. Sawera, P. Jensen, M. Ohman, M. Fredholm. MicroRNA sequence motifs reveal asymmetry between the stem arms. *Comput Biol Chem.*, 30:249-254, 2006.

S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31:439–41, 2003.

S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34:, 1362-4962, 2006.

J. H. Havgaard, E. Torarinsson, and J. Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput Biol*, 3, 2007.

J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I.L. Hofacker, P.F. Stadler, and The Students of Bioinformatics Computer Labs 2004 and 2005. The Expansion of the Metazoan MicroRNA Repertoire. *BMC Genomics*, 7:15, 2006.

R. Ihaka and R. Gentleman. R: A Language for Data Analysis and Graphics *J Comput Graphical Statistics*, 5:299–314, 1996.

M. Kato and F. J. Slack. MicroRNAs: small molecules with big roles – *C. elegans* to human cancer. *Biology of the cell*, 100:71–81, 2008.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120:15–20, 2005.

M. Lindow and J. Gorodkin. Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol.*, 26:339–51, 2007.

P. A. Maroney, Y. Yu, and T. W. Nilsen. MicroRNAs, mRNAs, and translation. *Cold Spring Harb Symp Quant Biol*, 71:531–535, 2006.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405:442–451, 1975.

M. F. Mette, J. van der Winden, M. Matzke, and A. J. Matzke. Short RNAs can identify new candidate transposable element families in arabidopsis. *Plant Physiol*, 130:6–9.

E. P. Nawrocki and S. R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol.*, 3:e56, 2007.

W. R. Pearson. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11:635–650, 1991.

L. Peters, and G. Meister. Argonaute proteins: mediators of RNA silencing. *Molecular Cell*, 26:611-623, 2007.

J. Piriyapongsa and K. I. Jordan. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*, 2, e203, 2007.

M. Rabani, M. Kertesz, E. Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA*, 2008 [Epub PMID: 18815376]

D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45:810–825, 1985.

T.D. Schmittgen. Regulation of microRNA processing in development, differentiation and cancer. *J Cell Mol Med*, 2008 [Epub; PMID: 18752632]

P. Schuster, W. Fontana, P.F. Stadler, I.L. Hofacker. From Sequences to Shapes and Back: A case study in RNA secondary structures, *Proc. Roy. Soc. Lond. B* 255 279–284, 1994.

M. Suzuki and Y. Hayashizaki. Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *Bioessays.*, 26:833–43, 2004.

R. Suzuki and H. Shimodaira. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22:1540–1542, 2006.

Y. Tomari, T. Du, and P. D. Zamore. Sorting of Drosophila small silencing RNAs. *Cell*, 130:299-308, 2007.

E. Torarinsson and S. Lindgreen. WAR: Webserver for aligning structural RNAs. *Nucleic Acids Res.*, doi:10.1093/nar/gkn275, 2008

Z. Weinberg and W. L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics.*, 22:35–9, 2006.

B. Zhang, X. Pan, and T. A. Anderson. MicroRNA: a new player in stem cells. *J. Cell. Physiol.*, 209: 266–269, 2006.