BIOINFORMATICS

Vol. 22 no. 14 2006, pages e197–e202
doi:10.1093/bioinformatics/btl257

Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data

Jana Hertel[1,*] and Peter F. Stadler[1,2,3]

[1]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany,
[2]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria and
[3]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

ABSTRACT

Summary: Recently, genome-wide surveys for non-coding RNAs have provided evidence for tens of thousands of previously undescribed evolutionary conserved RNAs with distinctive secondary structures. The annotation of these putative ncRNAs, however, remains a difficult problem. Here we describe an SVM-based approach that, in conjunction with a non-stringent filter for consensus secondary structures, is capable of efficiently recognizing microRNA precursors in multiple sequence alignments. The software was applied to recent genome-wide RNAz surveys of mammals, urochordates, and nematodes.

Availability: The program `RNAmicro` is available as source code and can be downloaded from http://www.bioinf.uni-leipzig/Software/RNAmicro

Contact: Jana Hertel, Tel: ++49 341 97 16704, Fax: ++49 341 97 16709, {jana,studla}@bioinf.uni-leipzig.de

1 INTRODUCTION

MicroRNAs (miRNAs) form an abundant class of non-coding RNA genes that have an important function in post-transcriptional gene regulation and in particular modulate the expression of developmentally important genes in both multi-cellular animals and plants. In both kingdoms they act as negative regulators of translation. They are transcribed as longer primary transcripts from which approximately 70 nt precursors (pre-miRNAs) with a characteristic stem-loop structure are extracted; after export to the cytoplasm, the mature miRNAs, approximately 22 nt in length, are cut out from one side of the precursor stem structure. For reviews on the discovery and function of miRNAs we refer to the literature, see e.g. (Ambros, 2004; Kidner & Martienssen, 2005). At present, several hundred distinct miRNA families are known in metazoan animals (Griffiths-Jones *et al*., 2005; Hertel *et al*., 2006), and a few dozens have been described in plants (Griffiths-Jones *et al*., 2005; Zhang *et al*., 2005; Axtell & Bartel, 2005). In contrast to other major RNA classes, in particular tRNAs, there is no recognizable homology between different families, so that it is unclear whether they arose independently in evolution or whether they derive from a single ancestral microRNA gene.

There are two basic strategies to detecting novel miRNAs. The simpler one uses sequence homology to experimentally known miRNAs as well as the characteristic hairpin structure of the pre-miRNA (Weber, 2005; Legendre *et al*., 2005; Hertel *et al*., 2006; Dezulian *et al*., 2006). A specialized machine learning approach that is specifically designed to search for distant homologs of human miRNA families is described in (Nam *et al*., 2005). Clearly, this approach is not capable of finding miRNAs for which no family member is already known.

Several approaches have focused on detecting novel miRNAs based on the secondary structure of their precursor, sequence conservation in related organisms, and the sequence conservation patterns of the 3′ and 5′ arms precursor hairpin. The programs `miRscan`[1] (Lim *et al*., 2003b), `miRseeker` (Lai *et al*., 2003), and `miralign`[2] (Wang *et al*., 2005) have lead to the discovery of a large number of novel microRNAs in nematodes (Lim *et al*., 2003b), insects (Lai *et al*., 2003;Wang *et al*., 2005) and vertebrates (Lim *et al*., 2003a). Grad *et al*., (2003) developed a computational method for predicting miRNAs in the *C. elegans* genome using both sequence and structure homology with known miRNAs. A similar procedure was employed in the plant-specific `harvester` approach (Dezulian *et al*., 2006). Berezikov *et al*. (2005) use phylogenetic shadowing to find regions that are under stabilizing selection and exhibit the characteristic variations in sequence conservation between stems, loop, and mature miRNA. In this case, secondary structure is used in a later filtering step. Genomic context also can give additional information: `Mirscan-II`, for example, takes conservation of surrounding genes into account (Ohler *et al*., 2004). Altuvia *et al*., (2005) utilize the propensity of miRNAs to appear in genomic clusters (often in the form of polycistronic transcripts) as an additional selection criterion.

MicroRNA detection without the aid of comparative sequence analysis is a very hard task but unavoidable when species-specific miRNAs are of prime interest. The `miR-abela`[3] approach first searches for hairpins that are robust against changes in the folding windows (and also thermodynamically stabilized) and then uses a support vector machine (SVM) to identify microRNAs among these candidates (Sewer *et al*., 2005). A related technique is described by Xue *et al*. (2005). The program `PalGrade` scores hairpins in a somewhat similar way (Bentwich *et al*., 2005). A quite different

*To whom correspondence should be addressed.

[1]http://genes.mit.edu/mirscan/
[2]http://bioinfo.au.tsinghua.edu.cn/miralign
[3]http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi

approach starts with the analysis of overrepresented patterns in phylogenetic footprints located in the 3′UTRs of mRNAs. These motifs constitute putative microRNA target sites and are used to guide the search for corresponding pre-miRNA candidates (Xie *et al*., 2005).

Advances in computational RNomics have most recently made it feasible to perform genome-wide surveys for non-coding RNAs that are not *a priori* restricted to particular RNA classes. Programs such as `qrna` (Rivas & Eddy, 2001), `EvoFold` (Pedersen *et al*., 2006), and `RNAz` (Washietl *et al*., 2005b) attempt to discover evolutionarily conserved RNA secondary structures in given multiple sequence alignments. Two distinct approaches have been realized: `EvoFold` and `qrna` are based on SCFGs (stochastic context free grammars) to evaluate the probability that the aligned sequences have evolved under the constraint of conserving secondary structure. `RNAz`, in contrast, is based on energy-directed RNA folding and assesses both thermodynamic stabilization of the secondary structure relative to a randomized control and structural conservation as measured by the relative folding energy of an alignment consensus consensus (Hofacker *et al*., 2002). A support vector machine (SVM) is then employed to classify the multiple sequence alignment as ''structured RNA''. Both `RNAz` and `Evofold` have been applied to surveying the human genome providing evidence for tens of thousands of genomic loci with signatures of evolutionarily conserved secondary structure (Washietl *et al*., 2005b; Pedersen *et al*., 2006) and detected tens of thousands of putative structured RNAs. Further `RNAz` surveys have been conducted for urochordates (Missal *et al*., 2005), nematodes (Missal *et al*., 2006), and yeasts (Steigele *et al*., 2006).

These surveys produced extensive lists of candidates for functional RNAs without using (or providing) information on membership in a particular class of RNAs. The large number of putative ncRNAs (from a few thousands in invertebrates to about 100000 in mammals) prompts the development of efficient automatic tools for their further classification and annotation.

With the exception of a small number of evolutionarily very well conserved RNAs (in particular rRNAs, tRNAs (Lowe & Eddy, 1997), the U5 snRNA (Collins *et al*., 2004), RNAse P and MRP (Piccinelli *et al*., 2005)), most ncRNAs are not only hard to discover *de novo* in large genomes, but they are also surprisingly hard to recognize if presented without annotation. Indeed, given an alignment not more than a few hundred nucleotides in length that is known to contain an conserved secondary structure, it should be very easy to decide whether these sequences belong to a known class of ncRNAs or not. Conceptually, this is a simple classification task that should be solvable efficiently by most machine learning techniques.

In the case of non-coding RNAs, however, machine learning approaches severely suffer from the very limited amount of available positive training data and the fact that negative training data are almost never known at all. Even for the most benign case, microRNA precursors, there is only a few hundred independent known examples, namely the miRNA families listed in the `mir-base` (Griffiths-Jones, 2004; Griffiths-Jones *et al*., 2005; Hertel *et al*., 2006). Over-training is thus a serious problem. As a consequence, it is necessary to restrict oneself to a small set of descriptors. This constraint, however, makes the choice of the descriptors a crucial task. Since most ncRNAs have well-conserved secondary structures, it seems natural to include

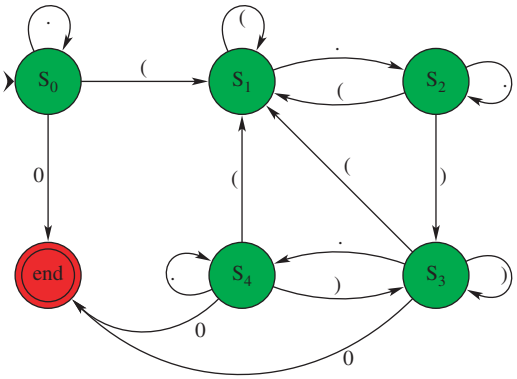

Fig. 1. Secondary structure automaton. The automaton reads an RNA secondary structure string in dot parantheses notation, recognizes all substructures, and stores their start positions and lengths.

structural descriptors in the classification procedure. RNA structure prediction, however, is less than perfect even when covariation information from the alignments can be used (Hofacker *et al*., 2002). This is true in particular when the exact ends of structured sequence within the multiple sequence alignment are not known.

In this contribution we present an SVM-based classificator for microRNA precursors that is designed to evaluate the information contained in multiple sequence alignments. The program `RNAmicro` is designed specifically to work as a 'sub-screen' for large-scale ncRNA surveys with `RNAz` or `Evofold`. The goal of `RNAmicro` is thus a bit different from that of specific surveys for miRNAs in genomic sequences: in the latter case one is interested in very high specificity so that the candidates selected for experimental verification contain as few false positives as possible. `RNAmicro`, in contrast, tries to provide an annotation of the `RNAz` survey data, so that we are interested in a more balanced trade-off between sensitivity and specificity similar to that of annotating protein motifs in known predicted protein coding genes.

2 METHODS

`RNAmicro` consists of (1) a preprocessor that identifies conserved 'almost-hairpins' in a multiple sequence alignment, (2) a module that computes a vector of numerical descriptors from each 'almost-hairpin', and (3) a support vector machine used to classify the candidate based on its vector of descriptors.

2.1 Detecting 'Almost Hairpins'

The outer loop of `RNAmicro` extracts windows of length L in 1-nucleotide steps from the input alignment. For each window, consensus sequence and consensus structure are computed using the `RNAalifold` algorithm (Hofacker *et al*., 2002) implemented in the `Vienna RNA Package` (Hofacker *et al*., 1994; Hofacker, 2003). The automaton in Fig. 1 is then used to analyze the consensus secondary structure, which is obtained in 'dot-parenthesis' notation[4].

Alignment windows whose consensus structure does not contain a stem with at least 10 base pairs or which contains two or more hairpins with at least 5 base pairs each are classified as 'not a miRNA precursor' without further analysis. Otherwise, the starting position and the length ℓ of the

[4]In this string notation for secondary structures, each unpaired nucleotide is represented by a dot, while base pairs correspond to matching pairs of parentheses.

Table 1. Descriptors used for SVM classification

Property	#	Descriptors
Structure	2	l_s, l_h
Sequence composition	1	G+C
Sequence conservation	4	$S_{5'}$, $S_{3'}$, S_0, $S_{\min}$
Thermodynamic stability	4	$\bar{E}$, $\bar{\epsilon}$, $\bar{\eta}$, $\bar{z}$
Structure conservation	1	E_{cons}
Total	12	

See text for definitions.

'almost-hairpin' which constituted the pre-miRNA candidate, are recorded and the corresponding alignment window is used to compute the descriptors. This filter, which on purpose is not very stringent, thus also accepts stem-loop structures with short 'branches' as candidates. Some important animal microRNAs are known to have structures of this type, for example *let-7*.

2.2 Descriptors

The lengths l_s and l_h of stem and hairpin loop regions recognized by the automaton form the first two descriptors provided the alignment window passes the structure filter. In addition we use the G+C content.

The second class of descriptors summarizes the thermodynamic properties of local sequence interval. MicroRNA precursors are known to be more stable than other RNAs with the same sequence composition (Bonnet *et al*., 2004; Clote *et al*., 2005). We thus use the average $\bar{z}$ of the energy z-scores

$$z = (E - \langle E \rangle_{\text{random}})/\sigma \tag{1}$$

where E is the folding energy of the given sequence. The mean $\langle E \rangle_{\text{random}}$ and σ of the distribution of randomized sequences is computed from a regression model as described by Washietl *et al*. (2005b) instead of using a shuffling procedure. Zhang *et al*. (2006) reported two folding energy scores that efficiently distinguish pre-miRNAs from other ncRNAs. The '*adjusted mfe*' is defined as $\epsilon = 100 \times E/\ell$; the '*mfe index*' η is the ratio of ϵ and the G+C content. We use their average values $\bar{\epsilon}$ and $\bar{\eta}$ as descriptors.

Structural conservation can be assessed by the *structure conservation index* (Washietl *et al*., 2005b), i.e. the ratio of the average folding energy of the aligned sequences and the energy of the consensus secondary structure. We use here $\bar{E}$ and E_{cons} separately.

An important characteristic of pre-miRNAs is the difference in the sequence conservation between the mature miRNA, which may be contained at either the 3′ or the 5′ side of the stem-loop structure, other parts of the stem, and the hairpin loop region, respectively, see e.g. (Lim *et al*., 2003b; Lai *et al*., 2003). We compute the average columnwise entropies $S_{5'}$, $S_{3'}$, and S_0, separately for 5′ and 3′ sides of the stem region and the hairpin loop. For a region (i.e., a subset of alignment positions) ξ we define

$$S_\xi = -\frac{1}{\text{len}(\xi)} \sum_{i \in \xi} \sum_{\alpha = A, C, G, U} p_{i,\alpha} \ln p_{i,\alpha} \tag{2}$$

where $p_{i,\alpha}$ is the fraction of α nucleotides at sequence position i. Since the mature miRNA is typically extremely well conserved, we determine the sequence window of length 23 with the lowest entropy $S_{\min}$ and use this value as an additional descriptor, Table 1.

2.3 SVM implementation

For classification we used a support vector machine as implemented in the `libsvm` package, version 2.8, (Chang & Lin, 2001). Descriptor vectors were scaled linearly to the interval $[-1, +1]$ before training using the binary version of `svm-scale` which is included in the `libsvm` package. The SVM was then trained using a radial basis function (RBF) kernel with $\gamma = 2$ and probability estimates. Default settings as listed in the README file of the `libsvm` package were used for all other parameters. The RBF kernel was used based on the recommendation of the `libsvm` documentation and positive experience with this kernel in the `RNAz` program. As we shall see below, these settings give satisfactory results in our context.

For alignments of length at most L, a single classification is performed. For longer alignments, we used a sliding window of length L with step-size 1. In this case, only the best (w.r.t. to SVM classification confidence value p) non-overlapping windows of length L were retained for each input alignment.

Table 2. Initial training and performance of RNAmicro SVM

Classification	Test sets *Positive*	*Negative*
miRNA	134	2
not miRNA	13	381
Total	147	383

Half of the positive and negative sets were used for training and testing, respectively.

2.4 SVM Training

Due to the relative sparseness of the available training data we used a stepwise training scheme. The positive training set is constructed from the union of animal microRNAs contained in the `miRNA registry` 6.0 and orthologous and paralogous sequences that have been obtained by a homology search in all metazoan genomes (Hertel *et al*., 2006). This set consisted of 295 alignments of distinct microRNA families composed by 2 up to 20 sequences from nematodes, insects, and vertebrates. Care was taken to avoid any sequence similarity between different alignments by using the family definition of (Hertel *et al*., 2006), which identifies several groups of microRNAs with different `mirbase` numbers as homologs. The antagonistic data was obtained by randomly shuffling the columns of each *true* miRNA alignment until the consensus sequence of the shuffled alignment folded again into a hairpin structure. This was successful for all but one true miRNA alignment. We have to rely at least in part on artificial examples since it seems hard to obtain a large collection of mutually independent evolutionarily conserved hairpin structures that are *known* not to be pre-miRNAs. The artificial set of negatives was complemented by a collection of 483 tRNA alignments which also passed the hairpin check. Note, however, that tRNAs are fairly similar to each other and hence cover only a relatively small part of the descriptor space.

In order to assess the quality of the descriptors, we divided both the positive and the negative set randomly into two halves, one used for training the SVM and the other used as test set. Consequently, there was no significant phylogenetic bias in the training set versus the test set.

We used `RNAmicro` with three different window sizes, $L = 70, 100, 130$, to scan the input alignments. An alignment is classified as putative microRNA if at least one window of at least one of the three values of L is classified with $p > 0.5$ by the SVM. We achieve a sensitivity of about 90% (134/147) and a specificity of about 99% (381/383) on the test dataset, Table 2. As an alternative training and testing we divided the available data into 90% for training and tested if the remaining 10% were classified correctly. This yields in a sensitivity of about 84% (26/31) and a specificity of about 99% (153/155).

Since the different training schemes yield consistent results and the training and test alignments are unrelated at sequence level, over-training thus does not seem to be a serious issue. We therefore trained the SVM using the entire positive and negative sets. We then tested the program on the results of `RNAz` screens of nematodes (Missal *et al*., 2006) and seasquirts (Missal *et al*., 2005). Although we could classify almost all known miRNAs that were contained in these data as miRNA, we found that in addition a

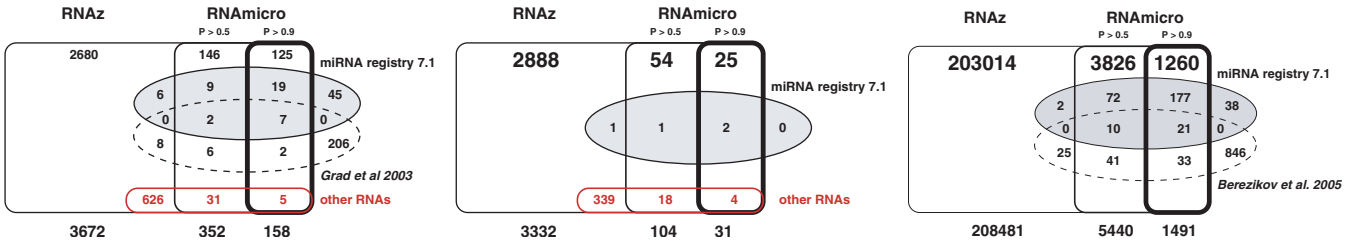

Fig. 2. Venn diagrams of RNAmicro-classifications of RNAz survey data with a RNAz cutoff of 0.5. The subsets of structured RNAs that are classified as miRNA candidates by RNAmicro are shown with bold outlines for $p = 0.5$ and $p = 0.9$ confidence levels. The subset of known microRNAs are shown with a grey background. Red numbers are other known ncRNAs or UTR elements that constitute known false positives in the $0.5 < p \leq 0.9$ and the $p > 0.9$ confidence classes, respectively. Numbers below the Venn diagram are the total number of RNAz alignments that were screened by RNAmicro, and the total numbers of signals classified as positive at confidence values $p = 0.5$ and $p = 0.9$, respectively. (a) Data from a pairwise screen of the nematoda *C. elegans* and *C. briggsae* (Missal *et al.,* 2006). In this case many known ncRNAs are contained in the data set allowing at least a rough estimate of false positive rates. (b) In the case of the two urochordates *Ciona intestinalis* and *Ciona savignyi* only 4 miRNAs are known. (c) For the screen of mammalian genomes comprising sequences that are conserved at least in human, dog, mouse, and rat (Washietl *et al.,* 2005a) almost all known non-coding RNAs were not available in the input alignments because they are marked as repetitive (tRNAs, snRNA, some microRNAs), so that a meaningful estimate for the false positive rate cannot be derived.

significant number of other known ncRNAs was mis-classified as pre-miRNAs. This indicates that our initial negative set does not sufficiently cover the descriptor space. The reason is that hairpins are common motifs in many other ncRNAs and that several other ncRNA families are also known to be thermodynamically very stable (Clote *et al*., 2005).

We therefore extracted alignments of noncoding RNAs from the Rfam database, focusing on a subset of snoRNAs, rRNAs, additional tRNAs, and RNAseP sequences and scored those with RNAmicro. False positives were added to the negative set and RNAmicro was retrained and tested with the 50% method as described above. The sensitivity was still around 90% while the specificity dropped to 78%. Thus, the mis-classified alignment slices of the negative input alignments were added to the training set. This procedure was iterated until no significant improvement was achieved on the Rfam dataset. This procedure is not statistically sound, of course. The alignments from the RNAz surveys contain in part different combinations of species and have been produced with different methods than those used for training, so that we can at least check the sensitivity of the model on the RNAz-alignments of the known microRNA precursors. Furthermore, other known ncRNAs in these data serve as a negative control.

3 APPLICATIONS

Three extensive surveys of metazoan genomes using RNAz (Washietl *et al*., 2005b) have been published recently. The screen of vertebrate genomes (Washietl *et al*., 2005a) was based on the top 5% conserved multiz alignments (Blanchette *et al*., 2004) as determined by phastcons (Siepel *et al*., 2005). For nematodes and urochordates, alignments were constructed using clustalw based on initial blast hits, see (Missal *et al*., 2005, 2006) for details. In all three cases, only non-repetitive non-protein-coding sequences were investigated.

In order to identify putative miRNAs among them we screened all individual alignment slices that were classified as potentially structured RNA with SVM classification confidence of $p_{\text{RNAz}} > 0.5$. Note that in all three studies individual alignment slices are combined to single 'RNAz hits' when they overlapped on the genome of the species. Hence the number of alignment slices is much larger than the number of 'RNAz hits' reported in these studies. Redundancies arising from miRNAs that appear in more than one alignment slice have been removed. The Venn diagrams in Fig. 2 summarize our classification.

It is reassuring that most of the RNAmicro predictions have high confidence values in the original RNAz screens: For example, 3850 (70%) of the 5440 $p_{\text{RNAmicro}} > 0.5$ candidates in the mammalian screen have $p_{\text{RNAz}} > 0.9$. Conversely, Only 204 (14%) of the 1491 $p_{\text{RNAmicro}} > 0.9$ have $p_{\text{RNAz}} < 0.9$. At least a rough estimate for the false discovery rate can be obtained from the distribution of the classification confidence values. For the three RNAz surveys we expect that about 1/5 to 1/4 of the putative ncRNAs are false positives at $p > 0.5$ classification confidence (not shown).

Berezikov *et al*. (2005) predicted 976 miRNAs by scanning whole-genome human/mouse and human/rat alignments. Their method, however, highlights evolutionary recent microRNAs so that it is not too surprising that there is relatively little overlap between these candidates and the RNAz screen (Washietl *et al*., 2005a), which focuses on evolutionary well-conserved RNA structures.

In order to compare our prediction with related classification methods, we re-evaluated the positive RNAmicro predictions using the SVM approach by Xue *et al*. (2005), which is designed for finding miRNAs *ab initio* in genomic sequences. Their procedure employs a very restrictive check for hairpin structures which in particular rejects the majority (180) of the 249 known microRNA precursors. Only 3077 of our 5440 $p > 0.5$ candidates and only 953 of our 1481 $p > 0.9$ candidates pass the hairpin filter. Of these, 1590 and 657, resp., are scored as microRNAs. Screening the $p_{\text{RNAz}} \geq 0.9$ subset with mir-abela returned 981 candidates, of which RNAmicro classifies 515 as microRNA precursors.

Several computational searches for miRNAs have been performed for nematodes. Grad *et al*. (2003) predicted 222 microRNA candidates (beyond those known at the time of publication) for *C. elegans*. Since most of the candidates are not conserved in *C. briggsae*, these sequence were not in the input set of RNAz survey. Thus, this set shows little overlap with our classification. Nevertheless it is interesting to note that the estimated total number of miRNAs is comparable. In contrast, based on the

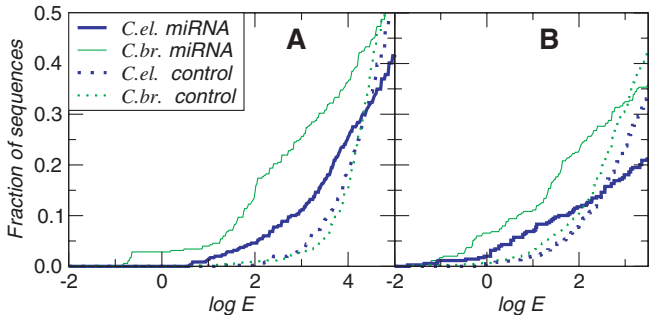

Fig. 3. Distribution of two closely related upstream motifs (A) and (B) reported for both *C. elegans* and *C. briggsae* (Ohler *et al.*, 2004, Fig.2). We plot the fraction of RNAmicro candidates for which mast (Bailey & Gribskov, 1998) recovers at least one copy A or B within 2000 nt upstream of the miRNA candidate as a function of the mast *E*-value cutoff. For small cutoffs, the miRNA specific sequence elements are overrepresented in true data versus a control set of RNAz hits that were not classified as microRNAs.

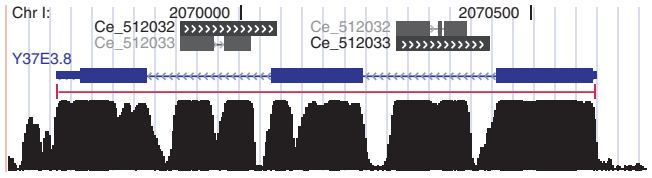

Fig. 4. Typical example of a pair of related putative intronic microRNAs in ***C. elegans*** extracted from the USCS genome browser. The gene ***Y37E3.8*** is a hypothetical protein of unknown function. The 'mountain range' on the bottom displays the sequence conservation between *C. elegans* and *C. briggsae*.

results of experimental verification of mirscan predictions, Lim *et al*. (2003b) and Ohler *et al*. (2004) conclude that the overwhelming majority of *C. elegans* miRNAs should have been found already.

Ohler *et al*. (2004) reported upstream sequence motifs specific to independently transcribed miRNAs in *C. elegans* and *C. briggsae*. We have therefore searched 2000 nt upstream for approximate occurrences of these patterns using mast. We find that both approximate patterns are substantially overrepresented in sequences classified as miRNAs relative to the remainder of the data, (Fig. 3). This provides additional *statistical* evidence that a substantial fraction of the RNAmicro-predictions indeed are microRNAs. As noted by Ohler *et al*. (2004), these sequence patterns, which are presumably transcription factor binding sites, do not occur associated with intronic miRNAs. We find that 176 (50%) of the 351 *C. elegans* candidates are located in introns (Fig. 4).

In the human data, 4245 candidates are not associated with known protein-coding genes, while 1107 candidates (20%) are located in introns (of which 36 are known microRNAs). This is in agreement with a recent study reporting that intronic microRNAs are much more frequent than previously thought (Ying & Lin, 2006). The remaining 88 sequences map to exons of known genes and are probably false positives.

MicroRNAs have a tendency to appear in clusters, probably because they are frequently processed from a polycistronic transcript. This fact has been utilized by (Altuvia *et al*., 2005; Sewer *et al*., 2005) to identify additional miRNAs in the vicinity of known ones. Using a rather conservative distance cutoff of <1000 nt between adjacent miRNAs, we found 143 clusters of miRNA candidates in the human genome, which contain 316 individual candidate sequences. Among them are 58 known miRNAs (according to mirbase 7.1) in 33 clusters. Most prominently, we recover the extensive imprinted cluster at human locus 14q32 discovered by (Lagos-Quintanta *et al*., 2002) (in total, we found 54 candidates in multiple tight clusters between positions 100M and 101M of the hg17 assembly) and the paralogs of the *mir-17* cluster (Tanzer & Stadler, 2004). In *C. elegans* we find 30 clusters with 131 members, in *C. intestinalis* there are 5 clusters with 10 members. Note that these are conservative estimates since in some cases, such as the *C. elegans mir-42* cluster, it is known that the distance between clustered miRNAs can be larger.

4 DISCUSSION

In contrast to other related approaches to miRNA detection, RNAmicro does not directly search a genome or genomes. Instead it is designed to classify the raw results of large-scale comparative genomics surveys for putative RNAs that are *conserved in both sequence and secondary structure*. Consequently, RNAmicro uses a different tradeoff between sensitivity and specificity. In the spirit of protein annotation methods, we aim for very high sensitivity rather than minimizing the expected number of false positives. As classificators become available for other classes of ncRNAs and common UTR motifs, conflicting class assignments from different classificators will eventually help to improve the specificity of miRNA detection.

Clearly, the performance of RNAmicro depends on the sensitivity and specificity of the initial screen for structured RNA candidates. However, RNAz exhibits a sensitivity of more than 80% at 99% specificity already on pairwise alignments (Washietl *et al*., 2005b, Table 2). In practice, it recovered 157 of the 163 human microRNAs in the input alignments that were known when the RNAz survey was performed (Washietl *et al*., 2005a). We therefore argue that this first step does not dramatically influence the overall sensitivity for microRNAs. Instead, the main limitations rather lie in (a) the coverage and quality of the input alignments and (b) the phylogenetic conservation of microRNAs, which of course limits all comparative approaches.

We have applied RNAmicro to three recent RNAz-bases studies of mammalian, nematode, and urochordate ncRNAs. In each case a large number of novel miRNA candidates have been detected. We have therefore investigated whether there is confounding evidence that a significant fraction of these predictions should be true positives: In *C. elegans*, for example, we find a strong association of RNAmicro predictions with a miRNA specific upstream motif previously reported by Ohler *et al*. (2004).. Furthermore, we found several hundred miRNA candidates that occur in tight genomic clusters. In particular in the human data, a large number of predictions are located within 1000 nt of a known microRNA. In line with recent reports (Ying & Lin, 2006),, we furthermore observed a substantial fraction (20% in human, 50% in *C. elegans*) of candidates are located in introns. Thus we argue that a large part of the RNAmicro candidates corresponds to real microRNAs. It is well conceivable that we have seen only a small fraction of the true miRNA repertoire to due to small expression levels and expression

patterns restricted to a few cell-lines (Ambros, 2004; Bartel & Chen, 2004; Mattick, 2004).

ACKNOWLEDGEMENTS

Financial support by the German *DFG* in the framework of the Bioinformatics Initiative (BIZ-6/1-2) and the SPP 'Metazoan Deep Phylogeny' is gratefully acknowledged.

REFERENCES

Altuvia,Y., Landgraf,P., Lithwick,G., Elefant,N., Pfeffer,S., Aravin,A., Brownstein,M.J., Tuschl,T. and Margalith,H. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res.*, **33**, 2697–2706.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Axtell,M.J. and Bartel,D.P. (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell*, **17**, 1658–1673.

Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Bartel,D.P. and Chen,C.-Z. (2004) Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs. *Nat. Genet.*, **5**, 396–400.

Bentwich,I., Avniel,A.A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E., Sharon,E., Spector,Y. and Bentwich,Z. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.

Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E. and Ronald Plasterk,H.A. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.

Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F.A., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D., Haussler,D. and Miller,W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Bonnet,E., Wuyts,J., Rouzé,P. and van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Clote,P., Ferré,F., Kranakis,E. and Krizanc,D. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.

Collins,L.J., Macke,T.J. and Penny,D. (2004) Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif. *J. Integ. Bioinf.*, **6**, 15p.

Dezulian,T., Remmert,M., Palatnik,J.F., Weigel,D. and Huson,D.H. (2006) Identification of plant microRNA homologs. *Bioinformatics*, **22**, 359–360.

Grad,Y., Aach,J., Hayes,G.D., Reinhart,B.J., Church,G.M., Ruvkun,G. and Kim,J. (2003) Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell*, **11**, 1253–1263.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.

Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.

Hertel,J., Lindemeyer,M., Missal,K., Fried,C., Tanzer,A., Flamm,C., Hofacker,I.L. and Stadler,P.F. & The Students of Bioinformatics Computer Labs 2004 and 2005 (2006). The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 25.

Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Kidner,C.A. and Martienssen,R.A. (2005) The developmental role of microRNA in plants. *Curr. Opin. Plant Biol.*, **8**, 38–44.

Lagos-Quintanta,M., Rauhut,R., Yalcin,A., Meyer,J., Lendeckel,W. and Tuschl,T. (2002) Identification of tissue specific microRNAs from mouse. *Curr. Biol.*, **12**, 735–739.

Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of drosophila microRNA genes. *Genome Biol.*, **4**, R42, [Epub].

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

Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003a) Vertebrate microRNA genes. *Science*, **299**, 1540–1540.

Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,P.B. (2003b) The microRNAs of *Caenorhabditis elegans*. *Genes & Development*, **17**, 991–1008.

Lowe,T.M. and Eddy,S. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Mattick,J.S. (2004) RNA regulation: a new genetics? *Nat. Genet.*, **5**, 316–323.

Missal,K., Rose,D. and Stadler,P.F. (2005) Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, **21**, i77–i78.

Missal,K., Zhu,X., Rose,D., Deng,W., Skogerbø,G., Chen,R. and Stadler,P.F. (2006) Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. *J. Exp. Zool.: Mol. Dev. Evol.*, DOI: 10.1002/jez.b.21086.

Nam,J.-W., Shin,K.-R., Han,J., Lee,Y., Kim,V.N. and Zhang,B.-T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, **33**, 3570–3581.

Ohler,U., Yekta,S., Lim,L.P., Bartel,D.P. and Burge,C.B. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309–1322.

Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

Piccinelli,P., Rosenblad,M.A. and Samuelsson,T. (2005) Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.*, **33**, 4485–4495.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Sewer,A., Paul,N., Landgraf,P., Aravin,A., Pfeffer,S., Brownstein,M.J., Tuschl,T., van Nimwegen,E. and Zavolan,M. (2005) Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics*, **6**, 267, [epub].

Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S., Weinstock,G.M., Wilson,R.K., Gibbs,R.A., Kent,W.J., Miller,W. and Haussler,D. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034–1050.

Steigele,S., Stadler,P.F. and Nieselt,K. (2006) Computational prediction and annotation of structured RNAs in yeasts. RECOMB poster.

Tanzer,A. and Stadler,P.F. (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, **339**, 327–335.

Wang,X., Zhang,J., Li,F., Gu,J., He,T., Zhang,X. and Li,Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.

Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005a) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.

Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005b) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.

Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.

Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Xue,C., Li,F., He,T., Liu,G., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310, [epub].

Ying,S.-Y. and Lin,S.L. (2006) Current perspectives in intronic microRNAs (miRNAs). *J. Biomed. Sci.*, **13**, 5–15.

Zhang,B.H., Pan,X.P., Cox,S.B., Cobb,G.P. and Anderson,T.A. (2006) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246–254.

Zhang,B.H., Pan,X.P., Wang,Q.L., Cobb,G.P. and Anderson,T.A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.*, **15**, 336–360.