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

Jana Hertel1,* and Peter F. Stadler1,2,3

1Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtestraβe 16-18, D-04107 Leipzig, Germany, 2Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria and 3Santa 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 precursor (Lim et al, 2003b), mirSeeker (Lai et al, 2003), and miralign2 (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-abela3 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

1http://genes.mit.edu/mirscan/
2http://bioinfo.au.tsinghua.edu.cn/miralign
3http://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 miRNAs. 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 tRNAs, rRNAs (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 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 classifier 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.

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

\[ e198 \]

> 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.

In this string notation for secondary structures, each unpaired nucleotide is represented by a dot, while base pairs correspond to matching pairs of parentheses.
where \( \pi \) mature miRNA is typically extremely well conserved, we determine the structure. We use here the 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., 2003). We thus use the average \( \varepsilon \) of the energy \( z\)-scores

\[
z = (E - \langle E \rangle_{\text{random}}) \sigma
\]

where \( E \) is the folding energy of the given sequence. The mean \( \langle E \rangle_{\text{random}} \) and \( \sigma \) 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 \( \epsilon = 100 \times E/F \); the ‘mfe index’ \( \eta \) is the ratio of \( \epsilon \) and the G+C content. We use their average values \( \epsilon \) and \( \eta \) 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 \( \hat{E} \) and \( E_{\text{cons}} \) 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 \( S_{\epsilon}, S_{\eta} \), and \( S_\pi \), separately for 5 and 3 sides of the stem region and the hairpin loop. For a region (i.e., a subset of alignment positions) \( \xi \) we define

\[
S_\xi = \frac{1}{\text{len}(\xi)} \sum_{i \in \xi} \sum_{a=A,C,G,U} p_{i,a} \ln p_{i,a}
\]

where \( p_{i,a} \) is the fraction of \( a \) 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 \( S_{\text{min}} \) 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.

<table>
<thead>
<tr>
<th>Property</th>
<th>#</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>2</td>
<td>( l_s, l_h )</td>
</tr>
<tr>
<td>Sequence composition</td>
<td>1</td>
<td>G+C</td>
</tr>
<tr>
<td>Sequence conservation</td>
<td>4</td>
<td>( S_\epsilon, S_\eta, S_{\text{min}} )</td>
</tr>
<tr>
<td>Thermodynamic stability</td>
<td>4</td>
<td>( E, \epsilon, \eta, \zeta )</td>
</tr>
<tr>
<td>Structure conservation</td>
<td>1</td>
<td>( E_{\text{cons}} )</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

See text for definitions.

<table>
<thead>
<tr>
<th>Table 1. Descriptors used for SVM classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Property</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Structure</td>
</tr>
<tr>
<td>Sequence composition</td>
</tr>
<tr>
<td>Sequence conservation</td>
</tr>
<tr>
<td>Thermodynamic stability</td>
</tr>
<tr>
<td>Structure conservation</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Table 2. Initial training and performance of RNAmicro SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>miRNA</td>
</tr>
<tr>
<td>not miRNA</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification</th>
<th>Test sets</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA</td>
<td>134</td>
<td>2</td>
<td>381</td>
</tr>
<tr>
<td>not miRNA</td>
<td>13</td>
<td>381</td>
<td>383</td>
</tr>
<tr>
<td>Total</td>
<td>147</td>
<td>383</td>
<td>383</td>
</tr>
</tbody>
</table>

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 sea squirts (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...
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 snRNAs, tRNAs, 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 $p_{RNAz} > 0.5$. Note that in all three studies individual alignment slices are combined to single ‘RNAz hits’ when they overlap 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, $8350 \ (70\%)$ of the 5440 RNAmicro > 0.5 candidates in the mammalian screen have $p_{RNAz} > 0.9$. Conversely, Only 204 (14%) of the 1491 RNAmicro > 0.9 have $p_{RNAz} < 0.9$. 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 $p_{RNAz} \geq 0.9$ 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
miRNA specific sequence elements are overrepresented in true data versus a candidate as a function of the 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. 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.

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) 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-Quintana 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 (Tananer & 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 classifiers become available for other classes of ncRNAs and common UTR motifs, conflicting class assignments from different classifiers 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., 2005a, 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-based 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


