Sequence analysis

SnoReport: computational identification of snoRNAs with unknown targets

Jana Hertel1,2,*  Ivo L. Hofacker1  and Peter F. Stadler1–4

1Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria, 2Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelestraße 16–18, D-04107 Leipzig, 3Fraunhofer Institut für Zelltherapie und Immunologie Deutscher Platz 5e, 04103 Leipzig, Germany and 4Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received on June 13, 2007; revised on August 20, 2007; accepted on September 8, 2007

Advance Access publication September 25, 2007

Associate Editor: Limsoon Wong

ABSTRACT
Summary: Unlike tRNAs and microRNAs, both classes of snoRNAs, which direct two distinct types of chemical modifications of uracil residues, have proved to be surprisingly difficult to find in genomic sequences. Most computational approaches so far have explicitly used the fact that snoRNAs predominantly target ribosomal RNAs and spliceosomal RNAs. The target is specified by a short stretch of sequence complementarity between the snoRNA and its target. This sequence complementarity to known targets crucially contributes to sensitivity and specificity of snoRNA gene finding algorithms.

The discovery of ‘orphan’ snoRNAs, which either have no known target, or which target ordinary protein-coding mRNAs, however, begs the question whether this class of ‘housekeeping’ non-coding RNAs is much more widespread and might have a diverse set of regulatory functions. In order to approach this question, we present here a combination of RNA secondary structure prediction and machine learning that is designed to recognize the two major classes of snoRNAs, box C/D and box H/ACA snoRNAs, among ncRNA candidate sequences. The SnoReport software is implemented in ANSIC. The source code is available under the GNU Public License at http://www.bioinf.uni-leipzig.de/Software/snoReport. The supplementary material is available from the http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-015

Contact: jana@bioinf.uni-leipzig.de

*To whom correspondence should be addressed.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
High-throughput studies employing fundamentally different experimental approaches have demonstrated that a substantial fraction of the transcriptional output of eukaryotic cells does not code for proteins (Bertone et al., 2004; Carninci et al., 2005; Cheng et al., 2005). These studies not only generated a plethora of data on poorly understood novel genes but also led to a rapid increase in the number of documented members of well-established structured RNA families. Tools of comparative genomics, in particular RNAz (Washietl et al., 2005b) and Evofold (Pedersen et al., 2006), furthermore demonstrated that RNA secondary structure is under stabilizing selection at thousands of loci in the genomes of higher eukaryotes. A significant fraction of these loci are most likely independent non-coding RNAs (Pedersen et al., 2006; Washietl et al., 2005a), albeit the overwhelming majority of these computational predictions so far has remained without functional annotation.

Non-protein coding RNAs (ncRNAs) form a very heterogeneous group of transcripts with diverse evolutionary origins and histories. Indeed, several classes of enigmatic small ncRNAs have been discovered during the last year: piwi-related piRNAs (Aravin et al., 2006; Lau et al., 2006), 21U-RNAs (Ruby et al., 2006) and three classes of small ncRNAs associated with promoters and termini of protein-coding genes (Kapranov et al., 2007). The classification and annotation of both experimentally discovered and computationally predicted RNAs thus have become an important topic in computational biology, reviewed, e.g. by The Athanasius F. Bompfünewerer RNA Consortium et al. (2007).

Increased interest in ncRNAs in general has also motivated a series of experimental studies in diverse organisms aiming at increasing the set of documented members of the two best-understood RNA classes: microRNAs and snoRNAs. Most of the recent experimental approaches in this direction were...
assisted by bioinformatics employed to predict candidates for experimental verification. So far, however, of all classes of small RNAs, only tRNAs can be detected reliably solely based on the genomic sequence with a sensitivity and specificity that allows nearly perfect genome-wide annotation without the need for comparative genomics approaches [using tRNAscan-SE (Lowe and Eddy, 1997)]. A large array of different approaches, reviewed, e.g. in Yoon and De Micheli (2006), has recently been developed to detect microRNAs. The continuing increase in the number of human microRNAs from one release of the miRBase (Griffiths-Jones, 2004) to the next shows that none of the existing tools can in practice find microRNA precursors with specificity and sensitivity comparable to that of tRNAscan-SE. Despite significant improvements in recent years, the tools for discovering snoRNAs (Edvardsson et al., 2002; Lowe and Eddy, 1999; Schattner et al., 2004) are even less reliable.

In fact, with the exception of a small number of evolutionarily very well conserved RNAs—such as the rRNAs, tRNAs (Lowe and Eddy, 1997), the U5 snoRNA (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. While it is often impossible to use relatively faint sequence homologies to find homologs of known RNAs in a complete genome, it can be much easier to recognize the very same sequences in the output of genomic ncRNA screens, due to the enrichment of functional RNAs by several orders of magnitude. Moreover, we found in previous ncRNA screens based on RNAz (Missal et al., 2005, 2006; Washietl et al., 2005), that a comparison with Rfam alignments using structure-based methods such as infernal (Nawrocki and Eddy, 2007) identifies very few RNAz hits that are not already recognizable by means of blast. Thus, while homologs of known sequences usually can be recognized reliably, determining class membership of novel examples is a much harder problem.

In this contribution, we introduce a combination of RNA secondary structure prediction and machine learning that is capable of recognizing snoRNAs in single sequences. In contrast to previous methods for snoRNA recognition [with the exception of snoSeeker (Yang et al., 2006)], we deliberately do not include information on putative target sites within ribosomal or spliceosomal RNAs although these additional constraints can dramatically enhance the specificity (Lowe and Eddy, 1999; Schattner et al., 2004, 2006). Notably, an increasing number of orphan snoRNAs, which lack antisense to rRNAs or snRNAs, has been experimentally identified from different eukaryotes (e.g. Huang et al., 2005; Hüttenhofer et al., 2001, 2004). In particular, a subgroup of snoRNAs is expressed in the mammalian brain and appears not to be involved in modification of rRNAs and snRNAs. Instead, some of these snoRNAs target specific mRNAs and interfere with A-to-I editing, see, e.g. the work of Vitali et al. (2005) and the recent review by Rogejl (2006).

For microRNAs, we recently presented the machine-learning approach, RnaMicro, that has been designed specifically to annotate genome-wide comparative genomics data (Hertel and Stadler, 2006): based on about a dozen variables describing features of sequence variation and consensus structure in a multiple sequence alignment, a support vector machine (SVM) is trained to distinguish between microRNA precursors and other types of hairpin-like structures. The observation that box H/ACA snoRNAs acted as decoys in early versions of RnaMicro's SVM model suggested to use a similar approach, albeit with adapted sets of descriptors for snoRNA recognition and a specific pre-filter based on predicted secondary structures. Since the two sub-classes of snoRNAs, box C/D and box H/ACA sequences are mutually unrelated in terms of sequence and structure (Fig. 3) snoReport employs distinct models derived from a similar collection of descriptors for them.

## 2 MATERIALS AND METHODS

### 2.1 Data sources

Since snoReport in essence implements a machine-learning method, positive and negative training sets play an essential role for the performance of our approach. As is often the case in bioinformatic applications, the available positive datasets are relatively small. We use here the collection of the snoRNA database Lestrade and Weber, 2006, while other non-coding RNAs from the Rfam (Griffiths-Jones et al., 2005) and miRBase (Griffiths-Jones, 2004) and a set of random sequences serve as negative examples. Clusters of sequences with substantial sequence homology are determined using blastclust (Altschul et al., 1990) with an E-value threshold of 10\(^{-3}\). From each cluster, only a single representative is retained. In this way, we avoid correlations between test and training sets in validation experiments. The total numbers of available independent test and training sequences are summarized in Table 1.

Position-specific weight matrices (PWMs) are used to represent the characteristic sequence motifs. These are extracted from snoRNAdb (Lestrade and Weber, 2006) using the web-version of mene (Bailey and Elkan, 1994), see Figure 1 for the resulting motifs. We apply

<table>
<thead>
<tr>
<th>CD positive samples</th>
<th>135</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD negative samples</td>
<td>1770</td>
</tr>
<tr>
<td>HACA positive samples</td>
<td>81</td>
</tr>
<tr>
<td>HACA negative samples</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 1. Training and test data

Positive examples correspond to the known box C/D and box H/ACA snoRNAs from snoRNAdb that pass the initial test for the presence of the sequence motifs and structural features. The negative set consists of selected Rfam families that are not snoRNAs (e.g. snRNAs, RNase P, 3S rRNAs, tRNAs), as well as miRNAs taken from miRBase that pass the initial motif and structure filter. Some C/D snoRNAs also show sequence motifs similar to H box and H/ACA also may show a low-scoring C and/or D box, due to the fact that the box motifs are PWM's and the matching algorithm also finds similar boxes—naturally with a lower score. These sequences are included as negative examples of box H/ACA snoRNAs, and vice versa.

![Fig. 1. Frequency plots of characteristic sequence motifs. Box ACA in contrast is an exact sequence pattern. See Supplementary Material for corresponding PWMs.](weblogo.berkeley.edu)
snoReport to sets of predicted and partially experimentally verified C/D and H/ACA snoRNAs from the studies of (Lowe and Eddy, 1999; Schattner et al., 2004; Huang et al., 2005; Zemann et al., 2006 and Yang et al., 2006) to assess the performance of our approach.

2.2 Software components

RNA structure prediction is performed using the current release 1.6 of the Vienna RNA Package (Hofacker, 2003; Hofacker et al., 1994). PWMs are matched to the genomic sequence using psmatch, a reimplementation (Mosig et al., 2006) of the algorithm described in Kel et al. (2003). We use libsvm 2.83 (Chang and Lin, 2001) to implement the SVM. The program snoReport is written in ANSI C and links these components.

2.3 The snoReport workflow

The two major classes of snoRNAs, the C/D box and the H/ACA box snoRNAs, can clearly be distinguished by characteristic sequence motifs and structural features. A third class of snoRNAs, the so-called scaRNAs, often shows features of the two other classes combined within a single molecule. Our goal here is a tool that is able to distinguish snoRNAs from the rest of the modern ‘RNA world’, and that in particular reliably recognizes members of the two major snoRNA classes. Both C/D box and H/ACA box snoRNAs are distinguished by their characteristic secondary structures; naturally, hence, we start with structure prediction.

In the prototypical structure of both classes, the characteristic sequence boxes are located in unpaired regions. Since predicted structures that are computed from a single sequence are rather unreliable (see, e.g. Hofacker et al., 2002), we first map the conserved sequence boxes to the input sequence. Pairs of boxes C and D, or H and ACA boxes, respectively, that are located within a maximum distance of 200 or 120 nt, respectively, from each other serve as seeds for a snoRNA candidate. Since the length of the sequence influences the outcome of structure predictions, we estimate the boundaries of the snoRNA as indicated in Figure 2.

Constrained secondary structures (Hofacker et al., 1994) are then computed for all pairs of C and D, and H and ACA boxes, respectively, that are within a maximal distance of 200 and 120 nt (indicated by $D_{CD}$ and $D_{HACA}$, respectively, in Fig. 2). In the case of box C/D snoRNAs, the entire region between boxes C and D must be unpaired, for H/ACA only the boxes themselves are prevented from base pairing. The constraint folding algorithm as implemented in the Vienna RNA Package is used for secondary structure computation.

If, using the constraints described above, the sequence folds into a prototypical snoRNA structure, the feature vector is computed from this sequence/structure pair and passed to the corresponding SVM. Both folding constraint and feature vector are of course different for box C/D and H/ACA snoRNAs. With an additional option, the user can also scan the reversed complement of the sequence. In order to keep memory requirements low, only positive results are immediately written to an output file (there is also an option to print all results). A candidate is classified as positive, i.e. as putative snoRNA candidate, if the SVM classification returns a classification $p_{SVM} > 0.5$. Candidates with $p_{SVM} > 0.9$ are regarded as high-scoring candidates.

The snoReport package additionally includes a perl script that summarizes the output by extracting the best hit (based on the resulting probability) from a group of predictions at overlapping locations of the input string.

2.3.1 Box H/ACA snoRNAs

Box ACA is located immediately after the last paired base of the second stem loop. Using the fact that both boxes must be unpaired as a constraint, the secondary structure of the sequence (truncated after ACA) is computed. The initial structure test is passed if it forms a single stem-loop structure between the H and the ACA box and if there is a stem-loop structure preceding box H. Now the sequence is truncated again at the 5’ end before the first base pair of this first stem loop. As a cross-check, the constrained structure is recomputed and the first stem loop is adjusted if necessary. Based on this predicted structure and the truncated sequence, the following attributes are extracted for SVM classification:

- $E$ Minimum free energy when folding without constraint
- $E_c$ Minimum free energy when folding with constraint
- $E_{av}$ Mean of the MFE
- $E_{std}$ SD of the MFE
- $L_{s5}$ Number of paired bases in 5’ stem
- $L_{s3}$ Number of paired bases in 3’ stem
- $L_{u5}$ Number of unpaired bases in 5’ hairpin loop
- $L_{u3}$ Number of unpaired bases in 3’ hairpin loop
- $D_{HACA}$ Distance from H box to first unpaired base of first interior loop of 5’ stem
- $D_{A313}$ Distance from ACA box to first unpaired base of first interior loop of 3’ stem
- $D_{HACA}$ Distance from H to ACA box
- $GC$ GC content of the subsequence
- $\sigma_H$ psmatch score of the H box.

**H/ACA snoRNA**

![Fig. 3. Schematic representation of the two major snoRNA classes: box H/ACA snoRNA (left) and box C/D snoRNA (right). The characteristic sequence motifs, i.e. the boxes C = ‘AUGAUGA’, D = ‘CUGA’, H = ‘ANANNA’ (consensus sequence), and ACA, as well as the secondary structure features are crucial for the function of snoRNAs (Bachellerie et al., 2002; Lafontaine and Tollervey, 2002). See Figure 2 for sequence logos of the three boxes.](image-url)
To provide both $E$ and $E_c$ helps to get an idea how much 'effort' is necessary to force the sequence into the requested structure or if the sequence likes to fold into another more or less stable structure. The $E_{avg}$ and $E_{adv}$ values represent the mean and SD of folding energies for random sequences with identical nucleotide frequency. Rather than compute these values via shuffling, we use the same regression SVM as used in the RNAz program for zscore calculations (Washietl et al., 2005b). Lengths of both stems and hairpin-loop regions as well as the values for the distances ensure to classify only biologically sensible structures. GC content often is an indicator of thermodynamical stability of the structure for GC base pairs are connected with three hydrogen bonds instead of two in the AT base pairs.

### 2.3.2 Box C/D snoRNAs

The complete region from the start of $C$ box to the end of $D$ box must remain unpaired. This long loop is enclosed by a short stem of usually 4–10 stacked base pairs. Depending on the predicted start position $i_C$ of the $C$ box and end position $j_D$ of the $D$ box motif in the sequence, we truncate the sequence 15 nt upstream at the 5' on the predicted start position enclosed by a short stem of usually 4–10 stacked base pairs. Depending $C$ box to the end of $D$ box must remain unpaired. This long loop is hydrogen bonds instead of two in the AT base pairs. Depending

In addition, $i_C$ and $j_D$ must be at least 3 nt away from start and end positions of the input sequence.

The remaining sequence is then folded forcing the region between both boxes to stay unpaired. For box C/D snoRNAs we use the following descriptors:

- $E$ Minimum free energy when folding without constraint
- $E_c$ Minimum free energy when folding with constraint
- $E_{avg}$ Mean of the MFE
- $E_{adv}$ SD of the MFE
- $L_s$ Number of paired bases in terminal stem
- $L_i$ Number of unpaired bases in hairpin loop
- $D_{CD}$ Distance from $C$ to $D$ box
- $\sigma_C$ Matching score of box $C$
- $\sigma_D$ Matching score of box $D$
- GC GC content of the subsequence.

Runtime is reduced by avoiding unnecessary examinations of unlikely candidates, i.e. those with box matching scores below 0.5, distances between boxes that are too short or too long, and those for which the characteristic secondary structures cannot be attained.

### 3 RESULTS

#### 3.1 Test statistics

To test each of the two models, we used cross-validation with randomly distributed datasets for training and testing. To this end, we partitioned the available data for positive and negative samples into sets of 50% training and test data, 80% train and 20% test data and 90% train and 10% test data, respectively. The training vectors are scaled such that the values are in a range from $[-1,1]$ before creating the model. The SVM type of our models is C-SVC, kernel-type is radial basis function (rbf) kernel and values for gamma and C are 2 and 1, respectively. When applying snoReport to the test sequences, we obtained a sensitivity of 0.96 and specificity of 0.91 for the classification of C/D snoRNAs. For H/ACA snoRNA classification, we find a sensitivity of 0.78 and a specificity of 0.89. In both cases those values correspond to the 80% test and 20% train model. The accuracy values of course increase for the 90%/10% model. Due to the small datasets, this increase is not significant, however.

The human genome contains a region that is particularly rich in snoRNAs chromosome 11, positions 93103000–93108500 (hg18); it contains 2 C/D and 6 H/ACA snoRNAs. Figure 4 gives a short overview of this region. We applied snoReport and two target-based snoRNA prediction tools to this region in order to evaluate the sensitivity.

In order to assess the false discovery rate of snoReport, 100 sequences are created using a genomic snoRNA rich region on human chromosome 11 (see, Fig. 4) and applying the dinucleotide shuffling procedure included in the SQUID toolkit by (Durbin et al., 1998) to this sequence. These randomized sequences are of course devoid of true snoRNAs.

These 100 ‘random sequences’ were scanned with snoReport. Additionally, we scanned those sequences with the box H/ACA snoRNA detector snoGPS (Schattner et al., 2004) and box C/D snoRNA detector snoScan (Lowe and Eddy, 1999) to compare true and false prediction rates. Both programs require target information (that is either included in the program packages for human and yeast or can be downloaded from the corresponding web pages). A more extensive target set for snoGPS was constructed here by declaring all Us in human rRNAs and snRNAs as potential targets. Note that this is still more restrictive than the approach taken by snoReport.

SnoGPS found 17 H/ACA snoRNA candidates on the human snoRNA rich region. Two overlap ACA1, three with ACA2, two with ACA25 and one with ACA40. However, the two C/D snoRNAs were also classified as H/ACA and there were eight additional candidates predicted that could not be mapped to one of the known snoRNAs in this region. Using the more extensive target set that contains every possible U as a putative target site, SnoGPS found 226 H/ACA snoRNA candidates on the human genomic snoRNA rich region.
including all C/D snoRNAs. SnoGPS reported 926 candidates in the 100 shuffled sequences. In summary, nine candidates predicted by SnoGPS could be assigned to known H/ACA snoRNAs while six candidates were misclassified on the human snoRNA rich region using known targets only. Using the extended target set, the false discovery rate increases significantly to ~80%. Additionally, SnoGPS reported 927 candidates on the randomized sequences, which were also regarded as false positive predictions.

SnoScan reported 19 C/D snoRNA candidates on the human snoRNA rich region. Three contain the complete C/D snoRNA mgh28S2409 and 2 mgh28S2411 while the remaining 14 candidates overlap with less than a half of the length of the known gene with the two known C/D snoRNA candidates and, thus are slightly dislocated. On the 100 shuffled sequences SnoScan predicted 257 snoRNAs.

In contrast, snoReport recovered all of the H/ACA and C/D snoRNAs with the exception of ACA25 and did not report any additional candidates on chromosome 11 region. ACA25 was not found by snoReport because the H box is located in the paired region of the S stem. If this pairing is prevented, the whole stem is destroyed. The sequence hence does not pass the structure filter. On the shuffled sequences, we found 90 box C/D candidates and only 6 H/ACA candidates. In these examples, the specificity of snoReport is an order of magnitude better than that of both SnoGPS and SnoScan while the sensitivity is at least comparable.

We remark that our data contain a nice example of an experimentally known orphan snoRNA that was detected by snoReport but was not found by the target-dependent approaches: ACA18 H/ACA snoRNA has no known targets. It was correctly classified as H/ACA snoRNA by snoReport, while it was not detected by SnoGPS even with the extended target set.

### 3.2 Validation on real data

For further validation of the prediction quality of snoReport, we applied our program to a set of elsewhere predicted and partially verified snoRNA sequences in human, nematodes, drosophilids, and the very distant Leishmania species.

Yang et al. (2006) reported 54 novel snoRNAs in the human genome (21 C/D and 32 H/ACA box snoRNAs) using their computational approach snoSeeker that is based on probabilistic models, pairwise whole genomic alignments of eukaryotes and regions of sequence complementary to rRNAs or snoRNAs. Screening their candidate sequences with snoReport resulted in 11 of 21 putative C/D candidates (including 7 of the 10 candidates whose expression was confirmed by northern blot analysis) as well as 23 of the 33 H/ACA candidates (8 of the 10 experimentally confirmed ones).

We therefore applied our program to predicted and partially experimentally confirmed snoRNAs reported in several recent publications in order to get an impression how reliably snoReport predicts canonical snoRNAs beyond the phylogenetic range of its training set, i.e. outside the vertebrates. A short summary of the results on already reported snoRNA candidates is given in Table 2.

Table 2. Results of snoReport applied to reported snoRNAs in human, nematodes, drosophilids and leishmania

<table>
<thead>
<tr>
<th>Human</th>
<th>C/D: 11/21</th>
<th>HACA: 23/32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmed</td>
<td>CD: 7/10</td>
<td>HACA: 8/10</td>
</tr>
<tr>
<td>Nematodes</td>
<td>Deng et al. (2006)</td>
<td>CD: 16/40</td>
</tr>
<tr>
<td>Zemann et al. (2004)</td>
<td>CD confirmed: 11/27</td>
<td>CD not confirmed: 19/70</td>
</tr>
<tr>
<td>Zemann et al. (2004)</td>
<td>CD: 8/17</td>
<td>HACA: 10/16</td>
</tr>
<tr>
<td>Leishmania</td>
<td>Liang et al. (2007)</td>
<td>CD: 7/62</td>
</tr>
</tbody>
</table>

We display the number of candidates that were positively classified by snoReport and the number of candidates reported in the cited references.

Applying a novel experimental protocol, Deng et al. (2006) cloned numerous small non-coding RNAs in Caenorhabditis elegans. Based on their genomic environment and transcriptional characteristics they found two-thirds independently transcribed, among them many intronic snoRNAs. We applied snoReport to their whole set of sequences that also contains RNase P, snoRNAs, snoRNA-like and ‘stem-bulge RNAs’ as well as 42 putative novel C/D and 47 H/ACA snoRNAs and novel not further classified candidates.

We found 22 of the C/D snoRNAs, while 3 sequences were misclassified as C/D snoRNA. Additionally, 31 H/ACA snoRNAs could be verified with our program and none of the sequences was falsely classified as H/ACA snoRNA. All previously known snoRNAs were found and 5 of the previously unassigned novel sequences could be classified as C/D snoRNA, and 1 as H/ACA snoRNA.

Zemann et al. (2006) detected 121 snoRNAs in C. elegans by a combination of high-throughput cDNA library screening and computational search strategies. Applied to those sequences, snoReport recognized 33 of 77 C/D snoRNAs and 44 of 57 H/ACA snoRNAs, including all known genes.

A further analysis on nematode snoRNAs (Huang et al., 2005) revealed 17 C/D and 16 H/ACA genes of which 8 and 11, respectively, could also be classified by snoReport. Accardo et al. (2004) computationally searched for C/D snoRNAs in the Drosophila melanogaster genome. Out of 27 confirmed sequences snoReport correctly assigned 11 candidates and additionally classified 19 of their ‘not confirmed’ candidates.

A very recent genome-wide analysis of snoRNAs in Leishmania major (Liang et al., 2007) found 62 C/D box snoRNAs and 37 H/ACA-like snoRNAs. SnoReport detected only seven of the C/D but none of the H/ACA-like candidates. A closer inspection showed that the H/ACA-like snoRNAs in Leishmanias are quite different from the canonical box H/ACA snoRNAs of yeast and vertebrates. For example, they lack a recognizable H box and they have an AGA instead of an ACA box.
4 DISCUSSION

We have presented here a combination of secondary structure prediction and SVM—classification that is capable of recognizing and classifying both major classes of snoRNAs. In contrast to most other tools [with the notable exception of snoSeeker (Yang et al., 2006), see below] that have been constructed for this purpose, snoReport does not rely on putative modification targets in rRNAs or snRNAs. It uses individual sequences as input. Trained almost exclusively on mammalian sequences it performs satisfactorily, with a sensitivity on the order of 50% and a false discovery rate that is an order of magnitude lower than that of other approaches, on nematodes and insects, and to a certain extent even on distantly related eukaryotes such as Leishmania. Recent work by Yang et al. (2006) as well as the further analysis of the RINAz survey of the human genome (Washietl et al., 2005a) suggests that there is still a larger number of orphan snoRNAs hidden in mammalian (and probably also in other eukaryotic) genomes.

Our approach differs from snoSeeker in two respects. This program uses a library of pairwise (whole genomic) alignments to which the input sequence is aligned. The resulting alignment is then used to identify the boxes and an HMM is employed to assess the structure. Thus, snoSeeker is inherently limited to snoRNAs that have homologs which can be aligned by blast/ multiz. In contrast, snoReport is completely independent of homology information and does not depend on any further a priori information. SnoReport also differs from snoSeeker by its purpose. While the snoSeeker program is designed for screening whole genomic alignments for putative snoRNA candidates and searching for putative target sites, snoReport has been designed to either annotate sequences that resulted from other non-coding RNA prediction tools or to screen complete chromosomes/genomes for novel snoRNA candidates.

We have also produced a separate variant of snoReport (not part of the current public distribution) that directly scores multiple sequence alignments. We observed that in practice there are too many alignment errors in automatically generated genome-wide alignments that lead to mis-alignments in particular of the sequence boxes. As a consequence, the direct evaluation of alignments does not lead to an improved classification on real data, e.g. from RINAz. In contrast, we observed a substantial improvement on manually curated alignments, in which the alignment of the boxes are repaired or mis-aligned sequences were removed. Thus, we propose to further explore a two-step classification procedure: in the first step, use snoReport separately on all individual sequences of an input alignment. If sufficient evidence is accumulated for a candidate, then a high-quality alignment can be constructed [using a structure-based-alignment approach such as locarna (Will et al., 2007) or possibly also a specialized alignment tool that knows about snoRNA specific features]. In the second step, this alignment could then be re-investigated by the alignment version of snoReport.

Our computational experiments show that snoRNAs show substantial variation between distantly related eukaryotes. The snoReport software therefore has been designed to be easily retrained as additional snoRNA sequences become available. In particular, the sensitivity for non-mammalian species could be improved by using clade-specific models. While this option is built into the software already, its practical applicability is severely limited by small sets of snoRNAs that have been characterized for any given species beyond metazoan animals and yeast.

ACKNOWLEDGEMENTS

Financial support by the German DFG in the framework of the Bioinformatics Initiative (BIZ-6/1-2) and the SPP ‘Metazoan Deep Phylogeny’, as well as the Austrian GEN-AU project ‘non-coding RNA’ is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES


