BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–4

snoStrip: A snoRNA annotation pipeline

Sebastian Bartschat[*], Stephanie Kehr, Hakim Tafer, Peter F. Stadler, Jana Hertel

Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Although small nucleolar RNAs form an important class of non-coding RNAs no comprehensive annotation efforts have been undertaken, presumably because the task is complicated by both the large number of distinct snoRNA families and their relatively rapid pace of sequence evolution.

Results: With snoStrip we present an automatic annotation pipeline developed specifically for comparative genomics of snoRNAs. It makes use of sequence conservation, canonical box motifs, as well as secondary structure and predicts putative targets.

Availability: The snoStrip web service and the download version is available at http://snostrip.bioinf.uni-leipzig.de/

Contact: sebastian@bioinf.uni-leipzig.de

1 INTRODUCTION

Small nucleolar RNAs (snoRNAs) are one of the most abundant and evolutionarily ancient groups of functional non-coding RNAs dating back at least 2-3 billion years to the last common ancestor of Archaea and Eukarya. They fulfill an impressive variety of cellular functions ranging from specifying the locations of chemical modifications in several ncRNA classes and nucleolytic processing of rRNAs to the synthesis of telomeric DNA and an involvement in genomic imprinting and alternative splicing, reviewed e.g. by (Bachellerie *et al.*, 2002; Matera *et al.*, 2007). They broadly fall into two classes distinguished by secondary structure and characteristic sequence boxes, after which they are named box C/D and box H/ACA snoRNAs. A variety of computational tools has been devised to identify snoRNAs *de novo* in searches of genomic DNA, see e.g., (Hertel *et al.*, 2008; Yang *et al.*, 2006). Homologous snoRNAs are often hard to find due to their small size, poor sequence conservation, and – in the case of box C/D snoRNAs – lack of a conserved secondary structure. So far no specific tool for homology-based snoRNA search has been devised. At the same time, the Rfam database covers only a subset of the known snoRNAs and many of the seed alignments contain only very few independent sequences (70% of the snoRNA alignments contain less than 16 sequences). Available snoRNA databases, on the other hand, mainly focus on single organisms, e.g. snoRNA-LBME-db on human and the Umass-database on yeast. Lacking overall sequence conservation and structural elements combined with characteristic sequence motifs makes it hard to detect snoRNAs by means of sequence homology, i.e., NCBI-blast, only.

2 RESULTS AND DISCUSSION

The snoStrip-pipeline has been designed to fill this gap. It embraces five parts: (1) a homology-based search procedure to accumulate potential snoRNA candidates, (2) a post-filter that uses the conservation of box motifs and putative target sites to increase specificity, (3) a module for extracting additional features including secondary structure and putative target predictions, (4) the computation of family-wide alignments, and (5) an optional validation check. Each novel snoRNA candidate and its corresponding snoRNA-derived information are subsequently stored in an internal database called snoBoard. The snoStrip-pipeline can either be run with single or multiple query families, each of which may contain one or more query sequences.

(1) – Homology search. The snoStrip-pipeline utilizes a set of known snoRNA sequences $\{s_1, s_2, \ldots, s_n\}$ of a given family S as queries to identify their homologs in a given target genome. First blastn with relaxed parameters (word size $W = 8$, E-value 10^{-3}, mismatch, gap opening, and gap extension parameters $q = -1$, $G = 2$, and $E = 1$) is employed. If no candidate is returned, a covariance model (CM) is generated from S using infernal 1.0.2 (Nawrocki *et al.*, 2009). To increase sensitivity, the model is calibrated with --exp-cmL-loc set to 3.0 Mb. An Infernal-derived candidate for a genome of length N is accepted if its bitscore exceeds $\log_2(2N)$ and $E < 0.01$.

(2) – Box filtering and target site extraction. Short conserved box motifs are characteristic for *bona fide* snoRNAs. However, several specific nucleotides and structural components have to be present to ensure their functionality. For detailed information and references, please have a look at our manual on the web server.

Given a snoRNA-family S and a snoRNA candidate s_{new}, snoStrip uses MUSCLE (Edgar, 2004) to obtain a temporary alignment of $s_1, \ldots, s_n$ with s_{new}. If the location of a box motif in the alignment agrees for all sequences s_i and the box of s_{new} fits certain restrictions (see manual), this position is selected as box location in the candidate. Otherwise, a gap-free search window roughly delimited by the minimal and maximal start positions of the boxes in the known sequences is used to determine the location that best fits a PWM created from the corresponding box motifs of $s_1, \ldots, s_n$.

Candidate anti sense elements (ASE, 9-20nts in length) are located immediately upstream of box D and/or D'. We extracted corresponding PWMs of these lengths to score snoRNA candidates. Testing on randomized and true data returned 13nts as the most

[*] to whom correspondence should be addressed

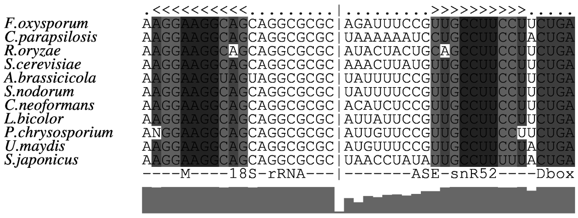
sensible window size and 0.7 as score threshold for acceptance (in accordance with Chen *et al.* (2007), see our website).

(3) – Property extraction. An important feature of snoRNAs is their type-specific secondary structure. We use RNAsubopt (Wuchty *et al.*, 1999) with type-specific folding constraints: box C/D snoRNAs are required to contain an internal loop delimited by the boxes C and D, while box H/ACA snoRNAs are prohibited from forming base pairs in their hinge and tail regions, resp. Correctly folded box C/D sequences are pruned at the base of the closing stem (or after at most 10bp), other candidates are truncated 8 nucleotides upstream of box C and downstream of box D. Box H/ACA snoRNAs are assumed to terminate 3nt after the ACA box.

Target predictions for putative snoRNA sequences are either performed by RNAsnoop (H/ACA) (Tafer *et al.*, 2010) or by PLEXY (C/D) (Kehr *et al.*, 2011). For box H/ACA snoRNAs we utilize target RNA accessibility profiles precomputed by RNAup (Mückstein *et al.*, 2006). We omit this step for box C/D snoRNAs since the accessibility around methylation sites doesn't significantly differ from the overall accessibility (data not shown). Finally, all single sequence predictions are mapped to the positions in target RNA alignments to facilitate the analysis of the conservation of the predicted modification sites.

(4) – Family-wide alignments. All potential snoRNA sequences assigned to a specific family are aligned with MUSCLE.

(5) – Validation check. In an optional postprocessing validation check we analyze all detected candidates with respect to their alignment score and target binding affinity (details in the manual).



```
                 .<<<<<<<<<<..........|..........>>>>>>>>>>.....
F.oxysporum      AAGGAAGGCAGCAGGCGCGC|AGAUUUCCGUUGCCUUCCUUCUGA
C.parapsilosis   AAGGAAGGCAGCAGGCGCGC|UAAAAAAUCCUGCCUUCCUACUGA
R.oryzae         AAGGAAGGCAGCAGGCGCGC|AUACUACUGCAGCCUUCCUUCUGA
S.cerevisiae     AAGGAAGGCAGCAGGCGCGC|AAACUUAUGUUGCCUUCCUUCUGA
A.brassicicola   AAGGAAGGUAGUAGGCGCGC|UAUUUUCCGUUGCCUUCCUUCUGA
S.nodorum        AAGGAAGGCAGCAGGCGCGC|UAUUUUCCGUUGCCUUCCUUCUGA
C.neoformans     AAGGAAGGCAGCAGGCGCGC|ACAUCUCCGUUGCCUUCCUUCUGA
L.bicolor        AAGGAAGGCAGCAGGCGCGC|AUUAUUCCGUUGCCUUCCUUCUGA
P.chrysosporium  ANGGAAGGCAGCAGGCGCGC|AUUGUUCCGUUGCCUUCCUUCUGA
U.maydis         AAGGAAGGCAGCAGGCGCGC|AUGUUUCCGUUGCCUUCCUACUGA
S.japonicus      AAGGAAGGCAGCAGGCGCGC|UAACCUAUAUUGCCUUUUUACUGA
                 ----M----18S-rRNA---|-------ASE-snR52----Dbox
```

Fig. 1: As an example of the web server output, the conservation of the target interaction between the 18S rRNA (left) and the ASE of box C/D snoRNA family snR52 (right) is shown.

For fungi, the **snoStrip web server** provides easy access to this snoRNA annotation pipeline. This service can be deployed in two operating modes: (1) genome-wide snoRNA annotation and (2) single sequence conservation. Due to resource constraints, the web version accepts moderate size genomes (60MB) as input. While in (1), the taxonomic range that is to be used as query can be specified, for mode (2), it is necessary to provide sequence specific box motifs. The service returns a variety of results that can be downloaded, e.g., mfasta- and gff-files, family-wide alignments, and alignments displaying conserved snoRNA-targetRNA interactions, see Fig 1.

For (large) genomes of multicellular plants and animals the snoStrip pipeline is easily applicable in a locally installed version. We have, for instance, conducted an extensive survey of metazoan snoRNAs that will be reported elsewhere. In the following we briefly outline snoStrip results on fungi and *G.lamblia*.

The initial query set consisted of 231 experimentally verified snoRNAs from five **fungal species** (see detailed manual on our web server). Running the snoStrip-pipeline resulted in more than 3500 putative snoRNAs in 63 fungal genomes. A more detailed overview is given in the table below. This provides by far the most comprehensive collection of fungal snoRNAs today and sets the stage for a detailed investigation into their evolution. Overall, we compared our candidates against several snoRNA prediction tools and the results can be found on the snoStrip website. A whole genome snoRNA annotation in *F.oxysporum* took about 5 hours and resulted in a total of 51 box C/D and 20 box H/ACA candidates.

	CD-snoRNAs		H/ACA-snoRNAs		# genomes
	families	seq.	families	seq.	
overall	67	2565	56	999	63
basal	29	76	6	14	4
Basidiomycota	28	161	8	34	7
Taphrinomycotina	31	89	23	57	3
Saccharomycotina	46	696	33	312	18
Pezizomycotina	58	1543	25	582	31

To test whether snoStrip can accomodate divergent sequence patterns we analyzed 30 validated snoRNAs from ***Giardia lamblia*** Isolate A (Hudson *et al.*, 2012). By sequence, 29 families were recovered in both Isolates B and E. With default settings, the pipeline rejected three of these families due to their aberrant box C sequences, which harbor two 2 substitutions.

In summary, with our snoStrip pipeline we provide a convenient and efficient way to annotate homologous snoRNAs in newly sequenced genomes. Complementarily, single snoRNA genes can be evolutionary traced across a widespread of species. Our snoStrip generated collections of snoRNA data constitute a valuable resource for large-scale studies, e.g., on snoRNA evolution and target interaction. It further enables a more generalized characterization of snoRNA species, e.g., for improving the accuracy of machine learning approaches for *de novo* snoRNA prediction.

Funding: This work was funded in part by the European Union under the auspices of the FP-7 project QUANTOMICS. PFS holds secondary affiliations with the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, the Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany, the Institute for Theoretical Chemistry, University of Vienna, Austria, and the The Santa Fe Institute, Santa Fe, USA.

REFERENCES

Bachellerie, J. P., Cavaillé, J. & Hüttenhofer, A. (2002) The expanding snoRNA world. *Biochimie,* **84** (8), 775–90.

Chen, C. L., Perasso, R., Qu, L. H. & Amar, L. (2007) Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J Mol Biol,* **369** (3), 771–83.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *NAR,* **32** (5), 1792–7.

Hertel, J., Hofacker, I. L. & Stadler, P. F. (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics,* **24** (2), 158–64.

Hudson, A. J., Moore, A. N., Elniski, D., Joseph, J., Yee, J. & Russell, A. G. (2012) Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in Giardia lamblia. *NAR,* **40** (21), 10995–1008.

Kehr, S., Bartschat, S., Stadler, P. F. & Tafer, H. (2011) PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics,* **27** (2), 279–80.

Matera, A. G., Terns, R. M. & Terns, M. P. (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol,* **8** (3), 209–20.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F. & Hofacker, I. L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics,* **22** (10), 1177–82.

Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics,* **25** (10), 1335–7.

Tafer, H., Kehr, S., Hertel, J., Hofacker, I. L. & Stadler, P. F. (2010) RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics,* **26** (5), 610–6.

Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers,* **49** (2), 145–65.

Yang, J. H., Zhang, X. C., Huang, Z. P., Zhou, H., Huang, M. B., Zhang, S., Chen, Y. Q. & Qu, L. H. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *NAR,* **34** (18), 5112–23.