BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–3

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: Small nucleolar RNAs form a major class of non-coding RNAs that has received quite a bit of attention both experimentally and computationally. Nevertheless, no comprehensive annotation efforts have been undertaken, presumably due to the large number of snoRNA families and their relatively rapid pace of sequence evolution that makes orthology assignment for these short sequences a difficult task.

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, and secondary structure and predicts putative target sites.

Availability: The `snoStrip` web server is freely 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 evolutionary 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 ncRNAs classes and nucloelytic 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). The 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 have been devised to identify snoRNAs in *de novo* searches of genomic DNA, see e.g., **?**. Although it has turned out that homologous snoRNAs are often hard to find due to the small size, poor sequence conservation, and – in the case of box C/D snoRNAs – lack of a strong secondary structure, so far no specific tool for homology-based snoRNA search has been devised. On the other hand, the `Rfam` database covers only a subset of the known snoRNAs and many of the seed alignments contain only very few independent sequences check this – can we be more specific here? Available snoRNA databases, on the other hand, focus on human **?** and yeast sequences **?** and provide a rather incomplete coverage of homologs in other species.

*to whom correspondence should be addressed

2 RESULTS AND DISCUSSION

The `snoStrip`-pipeline has been designed to fill this gap. It embraces four 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 as a means to increase specificity, (3) a module for extracting additional features including the prediction of secondary structure and putative targets, and (4) the computation of family-wide alignments. Each novel snoRNA and its corresponding snoRNA-derived information are subsequently stored in an internal database called `snoBoard`.

Step 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 openening, and gap extension parameters $q = -1$, $G = 2$, and $E = 1$) is employed. If no candidate(s) are returned, a covariance model (CM) is generated from all snoRNA sequences currently assigned to the family under consideration using `infernal` 1.0.2 (Nawrocki *et al.*, 2009). To increase the 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 $E < 0.01$.

Step 2 – Box filtering and target site extraction. Short conserved box motifs are characteristic for *bona fide* snoRNAs.

Several nucleotides have to be present to ensure functionality of a C/D box snoRNAs: both sheared G•A dinucleotides in box C (RU**GA**UGA) and box D (CU**GA**) are essential (Watkins *et al.*, 2000); the 5' uridine of box D is also indispensable (Cahill *et al.*, 2002). Several surveys indicate the importance of an impeccable structure of the Kink-turn motif in which box C and D play a crucial role (Watkins *et al.*, 2002; Xia *et al.*, 1997). In particular, stem II of the Kink-turn is required to contain at least one uridine of the U-U pair (RUGA**U**GA and C**U**GA) and two Watson-Crick base-pairs (formed by RUGAU**GA** and **NC**UGA). In general, we accept one mutation in the remaining positions of each of the two terminal boxes. At most two mutations are tolerated in the internal C' and D' box motifs since these are often more divergent (Cahill *et al.*, 2002). For H/ACA box snoRNAs, both box motifs are inherently more variable. Merely the adenine residues (**A**N**A**NN**A** and **A**C**A**) seem to underlie a certain evolutionary pressure. Nevertheless, not every single adenine has to be present to preserve functionality (Normand *et al.*, 2006). To account for this, we allow one mutated adenine in each box motif. The ACA motif is also known to be slightly

variable in its middle position. Hence, instead of a mutated adenine we further accept box motifs such as AUA and AAA.

In order to estimate the conservation of a particular box motif in the current snoRNA-family S, `snoStrip` employes `MUSCLE` (Edgar, 2004) to compute a temporary alignment of $s_1, \ldots, s_n$ with the new snoRNA candidate s_{new}. If the location of certain box motifs in the alignment agrees for all a sequences s_i and the corresponding aligned box b_{new} fits our restrictions, 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 $s_1, \ldots, s_n$.

A similar approach is employed to search for putative target sites in potential C/D box snoRNAs. Here, the start positions of box D and D' (if present in the query family) is determined and a window of size 18nts preceeding these boxes will be used as target region. Within this window, distinct PWMs of length 9nt are generated for each of the 10 possible start positions of the target region. Each of these PWMs is used to score all 9mers of the candidate sequence. A region in the new sequence is interpreted as target sequence provided ... subsequence providing an appropriate score ??? what is this ??? C/D box snoRNA candidates encoding both box motifs and putative target sites will be treated as true homologs and will be included in our database. Remaining sequences are stored for manual inspection.

Step 3 – Property extraction. An important feature of snoRNAs is their type-specific secondary structure. Within our pipeline we therefore use the predictions of `RNAsubopt` (Wuchty *et al.*, 1999). Furthermore, we apply type-specific folding constraints, such that C/D box snoRNAs are forced to contain an internal loop embraced by box C and D, while H/ACA box snoRNAs are prohibited to form base pairs in their hinge and tail region, respectively. Appropriately folded C/D box sequences are either cut after their closing base pairs or after at most 10bp. Otherwise, the snoRNAs are cut after eight nucleotides upstream of box C and downstream of box C. H/ACA box snoRNAs are assumed to terminate 3nt after the ACA box.

Target predictions for novel snoRNA sequences are either performed by `RNAsnoop` (HACA) (Tafer *et al.*, 2010) or by `PLEXY` (CD) (Kehr *et al.*, 2011). In order to consider the internal structure of the target RNAs, we also provide precomputed accessibility profiles from `RNAup` (Mückstein *et al.*, 2006). In contrast to pseudouridines, accessibility around methylated residues does not significantly differ from overall nucleotide accessibility (data not shown). Thus, no internal structure is considered for C/D box snoRNA target prediction. 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.

Step 4 – Family-wide alignments. After extracting homologous candidate sequences for query families from one or more given target genomes, `MUSCLE` is utilized to produce multiple alignments of all snoRNA sequences assigned to a specific family.

The `snoStrip`-pipeline can be run with multiple query families providing a convenient means for comprehensive snoRNA annotation. What if we only have single sequence to start with???

The `snoStrip` web server provides easy access to this snoRNA annotation pipeline. To-date, we have compiled snoRNA family data for deuterostomes and fungi; further major groups will follow. The `snoStrip` service can be deployed in two operating modes:

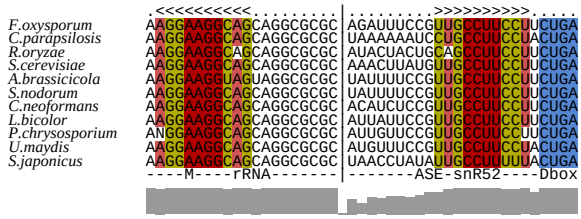

Fig. 1: Automatic `snoStrip`-based snoRNA annotation of the genome of the fungus *Fusarium oxysporum* took about 5 hours and resulted in a total of 51 C/D box and 20 H/ACA box snoRNAs. As an example of the web server output, the target conservation of box C/D snoRNA family snR52 is shown.

(1) genome-wide snoRNA annotation and (2) single sequence conservation. Due to resource constraints, the web version accepts only moderate size genomes (60MB) as input. For the much larger animal genomes, the pipeline must be downloaded. In the genome-wide mode, the taxonomic range that is to be used as query can be specified. In single sequence mode, the putative boxes in the query sequence must be annotated by the user. The service provides a variety of results that can be downloaded, e.g., mfasta- and gff-files, family-wide alignments, and alignments displaying the conserved snoRNA-targetRNA interactions, see Fig 1.

ich glaub, wir sollten schon die deterostomia hier mit erwaehnen.

The initial fungi dataset was composed of 233 experimentally verified snoRNAs in 5 fungi species. We were able to enlarge it to a total amount of over 3500 snoRNAs in 63 fungi genomes using our `snoStrip`-pipeline, forming the most comprehensive collection of fungi snoRNAs today. This resource sets the stage for a detailed investigation into their evolution (Manuscript in preparation).

	CD-snoRNAs		HACA-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

During our analysis it turned out that the 1^{st} WC base-pair as well as the U-U base-pair of stem II of the K-turn motifs are present in the majority of box C/D snoRNAs. The 2^{nd} WC base-pair is formed by 68% only, hence we declared it as arbitrary.

To further examine our pipeline, we analyzed 30 *Giardia lamblia* snoRNAs retrieved from Isolate A. Sequence-wise, 29 families can be found in both Isolates B and E, whereby our pipeline discarded three families completely due to 2 mutated positions within box C.

Here, we present a **comprehensive pipeline for automatic snoRNA annotation**, reducing manual curation to a minimum. Lacking overall sequence conservation and structural elements combined with characteristic sequence motifs makes it hard to detect snoRNAs by means of `BLAST` only. Our approach, on the other hand, provides a fully automatic pipeline for the genome-wide annotation of snoRNAs, including box motif extraction, secondary structure and target prediction. Our web service `snoStrip` provides a convenient starting point for homology-based snoRNA annotation in novel fungi genomes. Furthermore, `snoStrip` can be used to check experimentally derived, putative snoRNA sequences for conservation across all fungi lineages. This helps to

verify the candidates as true snoRNA with conserved box motifs and target regions.

Our huge dataset can be used to address several highly interesting topics. On the one side, we are able to investigate the evolution of snoRNAs quite accurately. Large scale analyses of target conservation and co-evolution of target and snoRNA can be realized. Additionally, the reliable set of `snoStrip` derived snoRNAs is suitable for training different machine learning algorithms that aim to detect novel snoRNAs computationally.

vielleicht doch noch ein paar takte ueber andere viecher/pflanzen...

Funding: PFS holds secondary affiliations with the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, the Fraunhofer Institut 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. This work was funded in part by the European Union under the auspices of the FP-7 project QUANTOMICS.

REFERENCES

Bachellerie, J. P., Cavaill, J. & Httenhofer, A. (2002) The expanding snorna world. *Biochimie*, **84** (8), 775–90.

Cahill, N. M., Friend, K., Speckmann, W., Li, Z. H., Terns, R. M., Terns, M. P. & Steitz, J. A. (2002) Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box c/d snornp. *EMBO J*, **21** (14), 3816–28.

Edgar, R. C. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32** (5), 1792–7.

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.

Normand, C., Capeyrou, R., Quevillon-Cheruel, S., Mougin, A., Henry, Y. & Caizergues-Ferrer, M. (2006) Analysis of the binding of the n-terminal conserved domain of yeast cbf5p to a box h/aca snorna. *RNA*, **12** (10), 1868–82.

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.

Watkins, N. J., Dickmanns, A. & Lührmann, R. (2002) Conserved stem ii of the box c/d motif is essential for nucleolar localization and is required, along with the 15.5k protein, for the hierarchical assembly of the box c/d snornp. *Mol Cell Biol*, **22** (23), 8342–52.

Watkins, N. J., Ségault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C. & Lührmann, R. (2000) A common core rnp structure shared between the small nucleoar box c/d rnps and the spliceosomal u4 snrnp. *Cell*, **103** (3), 457–66.

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.

Xia, L., Watkins, N. J. & Maxwell, E. S. (1997) Identification of specific nucleotide sequences and structural elements required for intronic u14 snorna processing. *RNA*, **3** (1), 17–26.