The small nucleolar ribonucleoprotein (snoRNP) database

J. CHRISTOPHER ELLIS,[1] DANIEL D. BROWN,[2] and JAMES W. BROWN[3]

[1]Department of Biology, Duke University, Durham, North Carolina 27708, USA
[2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA
[3]Department of Microbiology, North Carolina State University, Raleigh, North Carolina 27695, USA

ABSTRACT

Small nucleolar ribonucleoproteins (snoRNPs) are widely studied and characterized as guide RNAs for sequence-specific 2′-O-ribose methylation and psuedouridylation of ribosomal RNAs. In addition, snoRNAs have also been shown to interact with some tRNAs and direct alternative splicing in mRNA biogenesis. Recent advances in bioinformatics have resulted in new algorithms able to rapidly identify noncoding RNAs generally and snoRNAs specifically in genomic and metagenomic sequences, resulting in a rapid increase in the number and diversity of identified snoRNA sequences. The snoRNP database is a web-based collection of snoRNA and snoRNA-associated protein sequences from a wide range of species. The database currently contains 8994 snoRNA sequences from Bacteria, Archaea, and Eukaryotes and 589 snoRNA-associated protein sequences. The snoRNP database can be found at: http://evolveathome.com/snoRNA/snoRNA.php.

Keywords: sRNA; sRNP; small nucleolar RNA; small nucleolar RNP

INTRODUCTION

The most wellknown function of small nucleolar ribonucleoprotein particles (designated snoRNPs in Eukaryotes and sRNPs in Archaea) is to catalyze the 2′-O-ribose methylation (C/D snoRNPs) and pseudouridinylation (H/ACA snoRNPs) of ribosomal RNAs. The RNA components of snoRNP complexes are responsible for sequence specificity to the RNA target and possess no inherent catalytic or modification activity. Direct base pairing between the snoRNA and the target RNA provides site specificity for the catalytic protein subunits associated with the guide snoRNA, resulting in post-transcriptional modification of the target RNA.

Ribosomal RNAs are not the only target for snoRNP-directed nucleoside modification; for example, in Archaea, some C/D sRNPs (as snoRNA homologs are known in these organisms) direct post-transcriptional 2′-O-methylation of tRNA (Clouet d'Orval et al. 2001; Tang et al. 2005). In addition to their most well-known roles in 2′-O-methylation and psuedouridylation of ribosomal RNAs, some snoRNPs perform essential roles in pre-rRNA cleavages, and others function as rRNA chaperones during the assembly of the ribosome (Sáez-Vasquez et al. 2004). snoRNAs have also been shown to be associated with Prader-Willi syndrome, and other snoRNAs have putative binding sites flanking alternative splice junctions (Cavaillé et al. 2000; Skryabin et al. 2007; Bazeley et al. 2008; Ding et al. 2008).

snoRNAs fall into two major families based on their function and secondary structure: C/D and H/ACA. The primary function of the H/ACA snoRNP complex is to convert specific uridine ribonucleotides in the target RNA to pseudouridines. The H/ACA small nucleolar motif is composed of two conserved sequence elements: the Hinge box (H) between the two RNA hairpins, and an ACA sequence at the 3′ end of the RNA (Balakin et al. 1996). C/D snoRNAs act as guide RNAs for site-specific 2′-O-methylation. The C/D box motif in eukaryotic and archaeal sRNPs is characterized by two terminal conserved sequences, box C (AUGAUGA) and box D (CUGA), and two similar internal C′/D′ motifs.

The proteins in the snoRNP complex are remarkably diverse but share some common features. These proteins can be divided into four subtypes: archaeal H/ACA sRNA proteins (L7Ae, CBF5, NOP10, and GAR1), eukaryotic H/ACA snoRNA (Gar1p, Cbf5p, Nhp2p, and Nop10p), archaeal C/D sRNA (L7Ae, Nop56/58, fibrillarin), and eukaryotic C/D snoRNA (Snu13p, Nop56p, Nop58p, and fibrillarin). This diversity underscores the need for a snoRNP database that includes both the RNA and protein sequences.

Several snoRNA databases exist (Samarsky and Fournier 1999; Brown et al. 2003; Xie et al. 2007). All of these are focused on specific types of snoRNAs or snoRNAs from

Reprint requests to: James W. Brown, Department of Microbiology, North Carolina State University, Campus Box 7615, Raleigh, NC 27695-7615, USA; e-mail: james_brown@ncsu.edu; fax: (919) 515-7867.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1871310.

specific sources, and do not include snoRNA-associated proteins. The snoRNP database is the most comprehensive assembly of experimentally confirmed and computationally predicted snoRNAs and their proteins. The database, with its search engine, collection of ~9000 snoRNA sequences, and nearly 600 snoRNP protein sequences, serves the snoRNP community and can be found at http://evolveathome.com/snoRNA/snoRNA.php.

DATABASE OVERVIEW

The identification of noncoding (e.g., microRNA, RNase P RNA, 6S RNA, and snoRNA) RNAs in genomic and metagenomic sequences has proven to be a difficult computational task. Numerous algorithms have been created to identify putative noncoding RNAs, including snoRNAs. The snoRNA community has benefited from the creation of several important algorithms capable of identifying putative snoRNAs, such as snoTARGET, snoReport, Snoscan, snoGPS, and snoSeeker (Lowe and Eddy 1999; McCutcheon and Eddy 2003; Accardo et al. 2004; Hüttenhofer et al. 2004; Ghazal et al. 2005; Schattner et al. 2005; Yang et al. 2006; Bazeley et al. 2008; Hertel et al. 2008). This rapid expansion of robust bioinformatics tools has resulted in the identification of a large number of putative snoRNAs embedded in chromosomal, contig, and metagenomic sequences. However, because many of the predicted snoRNAs are embedded, but not annotated, in large sequences and are not available individually, it is difficult for researchers to access individual snoRNA sequences. Furthermore, none of the current snoRNA databases contain the proteins associated with snoRNP complexes. To address the lack of a central specialized repository, facilitate rapid searches of snoRNA/protein sequences, and support novel snoRNA-related algorithms in the future, the snoRNP database was created.

About the snoRNP database

The snoRNP database currently contains nearly 9000 snoRNA sequences and 600 snoRNP protein sequences, making it the largest specialized repository of not only snoRNA sequences but also snoRNP protein sequences. To facilitate queries of the database sequences, the snoRNP database has a search engine that supports a number of different search criteria. The most common search themes are: genus (e.g., *Drosophila*), species (e.g., *melanogaster*), accession number, snoRNA class (C/D or H/ACA), sequence, and journal. Because some snoRNAs contain conserved sequence elements, snoRNA researchers can also search the database for conserved sequences, such as the box C (AUGAUGA) and box D (CUGA) motifs, for more refined searches. All matches to the researcher's search terms are given in hyperlink list format, allowing users to choose the sequences they wish to view. Two sequence formats exist in the database: one for snoRNAs and another for snoRNA-associated proteins. Each format contains important information such as sequence type (e.g., RNA or protein), genus name, species name, accession number, publication (title and journal), description (usually an informal name), and the sequence itself. The protein sequence includes a brief description of the protein if available. Unique format features for RNA sequences include the start and stop locations of the RNA itself. This is important when the sequence is embedded in a larger genomic or metagenomic sequence. Furthermore, the snoRNA family (C/D or H/ACA) and class (U14, U19, etc.) are given, if known. Sequences for either proteins or snoRNAs can be downloaded in FASTA format individually or as an entire list of search results.

User support

The snoRNP database is a user-supported database that allows researchers to make contributions to the research community by submitting relevant data to the database to be shared. Although the database is actively seeking help in secondary structure determination of snoRNAs in the database, users are also encouraged to send sequences of snoRNAs or snoRNA proteins not currently found in the database. With the help of the snoRNA community, we hope to become a repository for snoRNA-related algorithms to make the snoRNP database a central location for snoRNA-related computational resources. We encourage all users to contact us via our contact page with all comments, suggestions, and contributions.

Database creation

One of the primary difficulties for snoRNA researchers today is that many snoRNAs are embedded in chromosomal sequences, contigs, or metagenomic sequences, making it time consuming and difficult to analyze snoRNA sequences. Researchers needed a central repository of snoRNA and associated protein sequences that did not require them to wade through large segments of genomic data. To address this deficiency in the research field, it was important to develop novel algorithms to extract snoRNAs and their associated protein sequences from large genomic sequences. To accomplish this goal, Perl and some PHP scripts were used to extract snoRNA and protein sequences from the National Center of Biological Institute (NCBI) database. Annotations that the users may find helpful, such as the GI and accession numbers, associated publication, and start and stop sites for each snoRNA sequence, were retained, and unrelated sequences were removed.

ACKNOWLEDGMENT

This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Received August 7, 2009; accepted January 5, 2010.

REFERENCES

Accardo MC, Giordano E, Riccardo S, Digilio FA, Iazzetti G, Calogero RA, Furia M. 2004. A computational search for box C/D snoRNA genes in the *Drosophila melanogaster* genome. *Bioinformatics* **20:** 3293–3301.

Balakin AG, Smith L, Fournier MJ. 1996. The RNA world of the nucleolus: Two major families of small RNAs defined by different box elements with related functions. *Cell* **86:** 823–834.

Bazeley PS, Shepelev V, Talebizadeh Z, Butler MG, Fedorova L, Filatov V, Fedorov A. 2008. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene* **408:** 172–179.

Brown JW, Echeverria M, Qu LH, Lowe TM, Bachellerie JP, Hüttenhofer A, Kastenmayer JP, Green PJ, Shaw P, Marshall DF. 2003. Plant snoRNA database. *Nucleic Acids Res* **31:** 432–435.

Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J, Hüttenhofer A. 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci* **97:** 14311–14316.

Clouet d'Orval B, Bortolin ML, Gaspin C, Bachellerie JP. 2001. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNA^{Trp} intron guides the formation of two ribose-methylated nucleosides in the mature tRNA^{Trp}. *Nucleic Acids Res* **29:** 4518–4529.

Ding F, Li HH, Zhang S, Solomon NM, Camper SA, Cohen P, Francke U. 2008. SnoRNA Snord116 (Pwcr1/MBII-85) deletion causes growth deficiency and hyperphagia in mice. *PLoS ONE* **3:** e1709. doi: 10.1371/journal.pone.001709.

Ghazal G, Ge D, Gervais-Bird J, Gagnon J, Abou Elela S. 2005. Genome-wide prediction and analysis of yeast RNase III-dependent snoRNA processing signals. *Mol Cell Biol* **25:** 2981–2994.

Hertel J, Hofacker IL, Stadler PF. 2008. snoReport: Computational identification of snoRNAs with unknown targets. *Bioinformatics* **24:** 158–164.

Hüttenhofer A, Cavaillé J, Bachellerie JP. 2004. Experimental RNomics: A global approach to identifying small nuclear RNAs and their targets in different model organisms. *Methods Mol Biol* **265:** 409–428.

Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.

McCutcheon JP, Eddy SR. 2003. Computational identification of noncoding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res* **31:** 4119–4128.

Sáez-Vasquez J, Caparros-Ruiz D, Barneche F, Echeverría M. 2004. A plant snoRNP complex containing snoRNAs, fibrillarin, and nucleolin-like proteins is competent for both rRNA gene binding and pre-rRNA processing in vitro. *Mol Cell Biol* **24:** 7284–7297.

Samarsky DA, Fournier MJ. 1999. A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res* **27:** 161–164.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan, and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33:** W686–W689.

Skryabin BV, Gubar LV, Seeger B, Pfeiffer J, Handel S, Robeck T, Karpova E, Rozhdestvensky TS, Brosius J. 2007. Deletion of the MBII-85 snoRNA gene cluster in mice results in post-natal growth retardation. *PLoS Genet* **3:** e235. doi: 10.1371/journal.pgen.0030235.

Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Hüttenhofer A. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **55:** 469–481.

Xie J, Zhang M, Zhou T, Hua X, Tang L, Wu W. 2007. sno/scaRNAbase: A curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res* **35:** D183–D187.

Yang JH, Zhang XC, Huang ZP, Zhou H, Huang MB, Zhang S, Chen YQ, Qu LH. 2006. snoSeeker: An advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res* **34:** 5112–5123.