A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*

Dmitry A. Samarsky* and Maurille J. Fournier

Department of Biochemistry and Molecular Biology, Program in Molecular and Cellular Biology, University of Massachusetts, Amherst, MA 01003, USA

ABSTRACT

Small nucleolar RNAs (snoRNAs) are involved in cleavage of rRNA, modification of rRNA nucleotides and, perhaps, other aspects of ribosome biogenesis in eukaryotic cells. Scores of snoRNAs have been discovered in recent years from various eukaryotes, and the total number is predicted to be up to 200 different snoRNA species per individual organism. We have created a comprehensive database for snoRNAs from the yeast *Saccharomyces cerevisiae* which allows easy access to detailed information about each species known (almost 70 snoRNAs are featured). The database consists of three major parts: (i) a utilities section; (ii) a master table; and (iii) a collection of tables for the individual snoRNAs. The utilities section provides an introduction to the database. The master table lists all known *S.cerevisiae* snoRNAs and their major properties. Information in the individual tables includes: alternate names, size, family classification, genomic organization, sequences (with major features identified), GenBank accession numbers, occurrence of homologues, gene disruption phenotypes, functional properties and associated RNAs and proteins. All information is accompanied with appropriate literature references. The database is available on the World Wide Web (http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html), and should be useful for a wide range of snoRNA studies.

BACKGROUND

Small nucleolar RNAs (snoRNAs) are stable RNA species localized in the eukaryotic nucleolus. Extensive studies in recent years have revealed the presence of snoRNAs in a broad variety of eukaryotes including fungi, protists, plants and animals (for reviews on snoRNAs see refs 1–7). Individual organisms are predicted to contain from about 75–100 (yeast) to perhaps nearly 200 (mammals) different snoRNAs. SnoRNAs have been demonstrated to define sites of nucleotide modifications in rRNA, specifically 2′-O-ribose methylation and formation of pseudouri-

dines (Ψ). In addition, a few snoRNAs are required for cleavage of precursor rRNA. Roles of snoRNAs in other aspects of ribosome biogenesis are also possible, such as rRNA folding and assembly of ribosomal subunits.

Structurally, all known snoRNAs (with one exception) fall into two groups: the box C/D family and the box H/ACA family (8,9). The exception is the MRP RNA which is involved in a specific pre-rRNA cleavage reaction, perhaps as a ribozyme (2,3). Members of the box C/D family contain one or two sets of universally conserved short sequence elements known as boxes C and D, and boxes C′ and D′ (the second set; 10). Most box C/D snoRNAs also contain long (>10 nt) sequences complementary to rRNA. Boxes C and D, as well as boxes C′ and D′, are usually located in close proximity, and form a structure known as the box C/D motif. This motif is important for snoRNA stability, processing, nucleolar targeting and function (11, and references therein). A small number of box C/D snoRNAs are involved in rRNA processing, most, however, are known or predicted to serve as guide RNAs in ribose methylation of rRNA. Targeting involves direct base pairing of the snoRNA at the rRNA site to be modified and selection of a rRNA nucleotide a fixed distance from box D or D′ (reviewed in refs 3,4,7; see also 12,13–15).

Members of the box H/ACA family contain an ACA triplet, exactly 3 nt upstream from the 3′ end and an H-box in a hinge region that links two structurally similar functional domains of the molecule. Both boxes are important for snoRNA biosynthesis and function. A few box H/ACA snoRNAs are involved in rRNA processing; most others are known or predicted to participate in selection of uridine nucleosides in rRNA to be converted to pseudouridines. Site selection is mediated by direct base pairing of the snoRNA with rRNA through one or both targeting domains (reviewed in refs 3,4; see also 16,17).

The DNA coding units for the snoRNAs occur in both traditional and novel genetic arrangements. Some are transcribed from independent promoters which serve mono- or polycistronic snoRNA coding units. Others are encoded within introns of protein (or protein-like) genes. Regardless of the diverse genomic organization, snoRNA synthesis appears to involve a number of pathways with common steps: (i) folding of the precursor to form a box C/D or box H/ACA protein binding motif; (ii) binding of protein(s) to this motif; (iii) processing of the precursor to the mature RNA; (iv) partial or complete assembly of the snoRNP

*To whom correspondence should be addressed at present address: HHMI/University of Massachusetts Medical Center, Program in Molecular Medicine, Worcester, MA 01606, USA. Tel: +1 508 856 5330; Fax: +1 508 856 5473; Email: dmitry.samarsky@ummed.edu

particle; and (v) transport to the nucleolus (11, and references therein).

DATABASE DESCRIPTION

Purpose of the database

The database serves the following purposes.

(i) A list of all known snoRNAs from the yeast *Saccharomyces cerevisiae. Saccharomyces cerevisiae* is predicted to possess ~75–100 different snoRNA species. Nearly 70 have already been identified and at least partially characterized. This master list will be useful for analyzing the structure, function and evolution of snoRNAs and their genes.

(ii) Ready access to key information available for individual snoRNAs and their genes. Since the discovery of the first snoRNA from *S.cerevisiae* in 1983 (18), a substantial number of papers devoted to different snoRNAs has been published. Though several excellent reviews summarizing information about sno-RNAs are available, electronic access to an updated compilation of data is not possible. The primary aim of the present database is to fill this void.

(iii) Easy comparison of the major features of snoRNAs. SnoRNAs and their cognate genes, as well as those from different organisms, possess a number of common features. Comparison of these features is helpful in defining the structure, mechanism of production and function of individual snoRNAs. The HTML format of the database allows easy and efficient manipulations of the RNA and DNA sequences.

(iv) Key literature references. Publications on snoRNA studies are growing at a fast pace. Orderly, comprehensive access to this literature through the database featured will be a useful aid to both specialists and newcomers to the field.

Database access and manipulations

The database is available on the World Wide Web at http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html , and can be viewed using major www-browsers (e.g., Netscape and Microsoft Explorer). The home page (not shown) provides links with the utilities pages and a master table. Utilities pages (not shown) include general information about snoRNAs, an outline of the structure of the database, an introduction to the authors, and a list of related RNA links.

The first functional page is the master table (Fig. 1). It provides a list of all known snoRNAs from *S.cerevisiae*. General information about each individual snoRNA is also presented, including: the size, family assignment, organization of the DNA coding sequence, existence of homologues in other organisms and function. The title of each snoRNA in the master table is linked to the individual snoRNA table.

The individual tables (an example is shown in Fig. 2) provide more detailed information about each snoRNA, including: alternate name(s), genomic neighbors, associated proteins, principal investigators of major defining studies and key literature references. The individual tables also provide links to the snoRNA and corresponding gene sequences, and to GenBank sites containing these sequences. The functional link in an individual table, in a box titled 'Process', provides information about the role of the snoRNA in rRNA maturation. It also connects to the individual page which specifies the pre-rRNA

Figure 1. Master table of snoRNAs. The master table provides a list of all known snoRNAs from *S.cerevisiae* and summarizes their most common features. The snoRNA titles (highlighted with red) are clickable and provide links to individual snoRNA tables (see Fig. 2). The master table can be accessed from the database home page, as well as from all pages related to the individual snoRNAs.

nucleotides or regions affected by a particular snoRNA (not shown).

Figure 2. Individual snoRNA page. A page that provides detailed information about an individual snoRNA (snR39 in this case) is shown. Such tables are accessed by clicking the corresponding titles in the master table (see Fig. 1). By using the various links highlighted (with blue) in a table, additional detailed information can be obtained about the snoRNA featured, including: gene structure in the form of descriptive and plain texts, GenBank data on the snoRNA coding and gene sequences, its immediate genomic neighbors, the snoRNA structure in the form of descriptive and plain texts and the nature of the biological processes in which the snoRNA is involved.

Behind the scenes

Maintenance is critical to the future value of any database. The present database encompasses almost 500 files, and the total number could well expand to a thousand. To simplify future manipulations of this material, the database has been organized as follows.

The files corresponding to the main page, master table and utilities pages (e.g., snoRNA overview, pre-rRNA sequence, etc.) have been placed in the root directory. Files relating to the individual snoRNAs are all grouped in corresponding sub-directories. The principles in naming the files inside these sub-directories are consistent: (i) all file names start with the name of the respective snoRNA (e.g., snR39_ta.html), and (ii) the second half of the file name refers to the context of the corresponding page ('…_ta' stands for the snoRNA table; '…_fu' for the snoRNA function; '…_rd' for the RNA descriptive text; etc.). This internal organization was designed to facilitate addition of new information by investigators in the field directly or through an author of the present database (D.A.S.).

Data sources

The conceptual information in the database originates from the published literature. Sequence data were retrieved from the GenBank database. References and links to relevant publications and internet sites have been provided. In some cases, unpublished data from our laboratory and information shared by other laboratories has been used.

FUTURE OF THE DATABASE

Database applications

The most valuable application of the database will be comparative analysis of sequences of different snoRNAs and corresponding genes from *S.cerevisiae* and other organisms. In addition, since the database summarizes features of all known *S.cerevisiae* snoRNAs, it should become a valuable resource for preparing reviews and original papers about snoRNAs. Users of the database are respectfully asked to cite the present article in publications.

Database updates and expansions

The yeast *S.cerevisiae* is predicted to posses up to 100 different snoRNAs. The current database describes about two thirds of these, leaving room for significant updating and expansion in the future. The database will be updated on a regular basis and on arrival of novel information about new or previously characterized snoRNAs.

Future modifications to the database may include: (i) direct links from the publications cited to the corresponding sites in the MEDLINE database; (ii) lists of functional and structural homologues from other organisms; (iii) schematic figures of predicted and demonstrated snoRNA secondary structures; and (iv) with permission, links to principal investigators responsible for discovery or characterization of individual snoRNAs.

We are open to all suggestions concerning modification, correction and improvement of the database. We also welcome new information about known, or previously uncharacterized snoRNAs from the yeast *S.cerevisiae* and homologues from other organisms.

ACKNOWLEDGEMENTS

REFERENCES

1 Gerbi,S.A. (1995) *Biochem. Cell. Biol.*, **73**, 845–858.
2 Maxwell,E.S. and Fournier,M.J. (1995) *Annu. Rev. Biochem.*, **35**, 897–934.
3 Tollervey,D. and Kiss,T. (1997) *Curr. Opin. Cell Biol.*, **9**, 337–342.
4 Smith,C.M. and Steitz J.A. (1997) *Cell*, **89**, 669–672.
5 Peculis,B.A. and Mount,S.M. (1996) *Curr. Opin. Cell Biol.*, **6**, 1413–1415.
6 Maden,T. (1996) *Nature*, **383**, 675–676.
7 Bachellerie,J.-P. and Cavaille,J. (1997) *Trends Biochem. Sci.*, **22**, 257–261.
8 Balakin,A.G., Smith,L. and Fournier,M.J. (1996) *Cell*, **86**, 823–834.
9 Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) *Genes Dev.*, **11**, 941–956.
10 Kiss-Laszlo,Z., Henry,Y. and Kiss,T. (1998) *EMBO J.*, **17**, 797–807.
11 Samarsky,D.A., Fournier,M.J., Singer,R.H. and Bertrand,E. (1998) *EMBO J.*, **17**, 3747–3757.
12 Kiss-Laszlo,Z., Henry,Y., Bachellerie,J.-P., Caizergues-Ferrer,M. and Kiss,T. (1996) *Cell*, **85**, 1077–1088.
13 Cavaille,J., Nicoloso,M. and Bachellerie,J.-P. (1996) *Nature*, **383**, 732–735.
14 Nicoloso,M., Qu,L.H., Michot,B. and Bachellerie,J.-P. (1996) *J. Mol. Biol.*, **260**, 178–195.
15 Tycowski,K.T., Smith,C.M., Shu,M.D. and Steitz,J.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 14480–14485.
16 Ni,J., Tien,A.L. and Fournier,M.J. (1997) *Cell*, **89**, 565–573.
17 Ganot,P., Bortolin,M.L. and Kiss,T. (1997) *Cell*, **89**, 799–809.
18 Wise,J.A., Tollervey,D., Maloney,D., Swerdlow,H., Dunn,E.J. and Guthrie,C. (1983) *Cell*, **35**, 743–751.