D162–D164 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 14 December 2006*
doi:10.1093/nar/gkl994

Noncoding RNAs database (ncRNAdb)

Maciej Szymanski*, Volker A. Erdmann[1] and Jan Barciszewski

Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Noskowskiego 12, 61-704 Poznan, Poland and [1]Institute fur Chemie/Biochemie, Freie Universitat Berlin, Thielallee 63, 14195 Berlin, Germany

Received September 15, 2006; Revised October 27, 2006; Accepted October 30, 2006

ABSTRACT

The noncoding RNA database (ncRNAdb) was created as a source of information on RNA molecules, which do not possess protein-coding capacity. It is now widely accepted that, in addition to constitutively expressed, housekeeping or infrastructural RNAs, there is a wide variety of RNAs participating in mechanisms involved in regulation of gene expression at all levels of transmission of genetic information from DNA to proteins. Noncoding RNAs' activities include chromatin structure remodeling, transcriptional and translational regulation of gene expression, modulation of protein function and regulation of subcellular distribution of RNAs as well as proteins. Noncoding transcripts have been identified in organisms belonging to all domains of life. Currently, the ncRNAdb contains >30 000 ncRNA sequences from Eukaryotes, Eubacteria and Archaea, but does not include housekeeping transcripts or microRNAs and snoRNAs for which more specialized databases are available. The contents of the database can be accessed via the WWW at http://biobases.ibch.poznan.pl/ncRNA/.

INTRODUCTION

In recent years, we have witnessed a growing interest in the involvement of RNA molecules in controlling gene expression. Numerous studies demonstrated that regulatory noncoding or non-protein-coding RNAs (ncRNAs and npcRNAs) play equally important role as protein transcription factors in determining the repertoire of expressed genes virtually in all living organisms. The significance of ncRNAs is evident from the results of analyses of the protein-coding capacity of sequenced genomes. The contribution of protein-coding regions decreases with the increase of complexity of organisms. In bacteria, unicellular eukaryotes and invertebrates, the coding sequences constitute ~95, 30 and 20% of the genomic DNA, respectively. In mammals, the open reading frames account only for ~1.5–2% of the genomes (1,2). Similar proportions of the coding and noncoding regions are observed within the sequences which are actually transcribed (3). There are also very little differences between the mammalian proteomes which suggests that the diversity observed on the phenotypic level is not determined by different sets of proteins, but rather by programs which govern their expression (4). A support for this view came recently from the discovery of the brain-specific HAR1F RNA expressed during cerebral cortex development. Accelerated evolution of HAR1F-encoding region may have contributed to the evolution of human brain (5).

Regulatory ncRNAs have been identified in all domains of life and they have been shown to be involved in numerous mechanisms controlling expression of genes at all levels of transmission of genetic information from DNA to proteins (6). They include epigenetic modification of chromatin structure (methylation and modification of histones), regulation of transcription by modulation of activity of RNA polymerase and transcription factors, RNA modification, mRNA stability and translation. Unlike the infrastructural ncRNAs (e.g. tRNAs, rRNAs, snRNA and snoRNAs), regulatory RNAs are not transcribed constitutively in all cells. In many cases, their expression depends on the cell or tissue type, developmental stage or is controlled by epigenetic factors or environmental conditions (e.g. hormones and stress). Specific changes in the expression of particular ncRNAs in humans have been also linked to human neurobehavioral and developmental disorders and cancer (7).

Initially, virtually all ncRNAs were discovered by chance and there were no systematic approaches to identify the whole contents of the transcriptome. At the time of the publication of the first ncRNA database in 1999 (8), there was only a handful of known ncRNAs which were regarded as curiosities. Since then, there has been a steady growth of reports on identification of new noncoding transcripts. In recent years, a large number of ncRNAs sequences have been determined in large scale sequencing projects of cDNAs (9,10), and small cytoplasmic RNAs (11). There has also been a growing number of reports describing novel methods for computational identification of ncRNA-encoding genes (12,13).

Despite the advances in identification of new ncRNAs both in prokaryotes and eukaryotes, our knowledge of the spectrum of their activities is rather limited. There are relatively few noncoding transcripts which have been at least

*To whom correspondence should be addressed. Tel: +48 61 8528503; Fax: +48 61 8520532; Email: mszyman@ibch.poznan.pl

partially characterized in terms of function or expression. This growth of interest in RNA biology was a motivation for several databases published in recent years (e.g. 14,15).

THE DATABASE

The purpose of the database is to serve and organize information concerning regulatory noncoding transcripts from all groups of organisms. In addition to the RNAs for which regulatory activities have been documented, the database also contains sequences of ncRNAs which are known to be expressed, but their role in a cell is still unknown. In comparison with other ncRNA databases, which appeared in recent years, ncRNAdb does not focus on any particular class of transcripts or taxonomic groups. The RNAs included in the database have been demonstrated or are suspected to function without being translated into proteins and they are not constitutively expressed housekeeping transcripts (e.g. tRNAs, rRNAs and snRNAs).

As of August 2006, the noncoding regulatory RNA database includes over 30 000 sequences from 99 organisms. A significant growth of the amount of data, compared with previous edition published in 2003 which contained 300 sequences (16), is primarily due to systematic identification of noncoding transcripts in mammals by means of large scale cDNA sequencing (8,9). These sequences account for over 90% of the data. On the other hand, certain groups of RNAs (microRNAs and snoRNAs), which were present in previous editions, were removed from our service to avoid the redundancy with other specialized databases.

DATA SOURCES

The primary source of sequences included in the database were the GenBank (17) records. Human and mouse ncRNA sequences are in part derived from H-Invitational (18) and FANTOM3 (8) full length cDNA databases, respectively. Computationally identified small cytoplasmic, bacterial RNAs which are not annotated in the genomic sequences were derived from the Rfam—the database of RNA families (15).

Since many of the primary transcripts of the eukaryotic ncRNAs are subject to alternative splicing, various splicing variants derived from the same gene are presented as separate entries. For ncRNAs, for which there are no individual GenBank records, predicted transcripts were extracted from genomic sequences using information provided in the feature tables or based on the multiple sequence alignments.

Apart from the sequences, the entries contain supplementary information (when available) on experimentally verified activities, expression patterns and chromosomal localization. The annotations and genome mapping information for the sequences rely on data provided in original GenBank records, H-Invitational Database of Annotated Human Genes (release 3.4, July 2006), FANTOM3 Database (March 2006 release) and UCSC Genome Browser Database (19). Most of the entries for eukaryotic transcripts are linked to the UCSC Genome Browser (20).

DATABASE ACCESS

Individual nucleotide sequences can be retrieved in FASTA format as separate entries or downloaded as batch files. The data can be searched using transcript names, accession numbers or organism names. In addition to the access to the database records, the search results also linked to full GenBank entries. In the current version of the database we also included the BLAST server which allows to perform sequence similarity searches using the full ncRNA database (~64 Mb). The search results are linked to the full database records.

The browser section of the database is intended as a source of basic information on noncoding transcripts which have been at least partially characterized in terms of function or expression patterns. For such RNAs we provide short descriptions of known activities described in the literature and relevant citations linked to the Medline database. The browser entries also give access to the nucleotide sequences from the database (FASTA) or the entire GenBank records.

The database can be accessed through the WWW at the following URL: http://biobases.ibch.poznan.pl/ncRNA.

ACKNOWLEDGEMENTS

This work was supported by grants from the Polish State Committee for Scientific Research (M.S. and J.B.), the Fonds der Chemischen Industrie e.V. (V.A.E.) and the National Foundation for Cancer Research (M.S., V.A.E. and J.B.). Funding to pay the Open Access publication charges for this article was provided by the Fonds der Chemischen Industrie e.V.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewor,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,M., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Frith,M.C., Pheasant,M. and Mattick,J.S. (2005) Genomics: the amazing complexity of the human transcriptome. *Eur. J. Hum. Genet*., **13**, 894–897.
4. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*., **2**, 986–991.
5. Pollard,K.S., Salama,S.R., Lambert,N., Lambot,M.A., Coppens,S., Pedersen,J.S., Katzman,S., King,B., Onodera,C., Siepel,A. *et al*. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443**, 167–172.
6. Szymanski,M. and Barciszewski,J. (2003) Regulation by RNA. *Int. Rev. Cytol*., **231**, 197–258.
7. Szymanski,M., Barciszewska,M.Z., Erdmann,V.A. and Barciszewski,J. (2005) A new frontier for molecular medicine: noncoding RNAs. *Biochim. Biophys. Acta*, **1756**, 65–75.
8. Erdmann,V.A., Szymanski,M., Hochberg,A., de Groot,N. and Barciszewski,J. (1999) Collection of mRNA-like noncoding RNAs. *Nucleic Acids Res*., **27**, 192–195.
9. Maeda,N., Kasukawa,T., Oyama,R., Gough,J., Frith,M., Engstrom,K.M., Lenhard,B., Aturaliya,R.N., Batalov,S., Beisel,B.W. *et al*. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet*., **2**, e62.
10. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al*. (2004) Complete sequencing and characterization of 21 243 full-length human cDNAs. *Nature Genet*., **36**, 40–45.

11. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*., **20**, 2943–2953.
12. Lipovich,L. and King,M.C. (2006) Abundant novel transcriptional units and unconventional gene pairs on human chromosome 22. *Genome Res*., **16**, 45–54.
13. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2006) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol*., **23**, 1383–1390.
14. Pang,K.C., Stephen,S., Engstrom,P.G., Tajul-Arifin,K., Chen,W., Wahlestedt,C., Lenhard,B., Hayashizaki,Y. and Mattick,J.S. (2005) RNAdb—a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res*., **33**, D125–D130.
15. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,SR. and Bateman,A. (2005) Rfam: annotating noncoding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121–D124.
16. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res*., **31**, 429–431.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res*., **33**, D34–D38.
18. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*., **2**, e256.
19. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
20. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The Human Genome Browser at UCSC. *Genome Res*., **12**, 996–1006.