D112–D115 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki041

NONCODE: an integrated knowledge database of non-coding RNAs

Changning Liu[1,2,3], Baoyan Bai[1,3], Geir Skogerbø[2], Lun Cai[2,3], Wei Deng[1], Yong Zhang[1,3], Dongbo Bu[2], Yi Zhao[2] and Runsheng Chen[1,2,*]

[1]Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, [2]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology and [3]Graduate School of the Chinese Academy of Science, Beijing 100080, China

Received August 11, 2004; Revised and Accepted September 27, 2004

ABSTRACT

NONCODE is an integrated knowledge database dedicated to non-coding RNAs (ncRNAs), that is to say, RNAs that function without being translated into proteins. All ncRNAs in NONCODE were filtered automatically from literature and GenBank, and were later manually curated. The distinctive features of NONCODE are as follows: (i) the ncRNAs in NONCODE include almost all the types of ncRNAs, except transfer RNAs and ribosomal RNAs. (ii) All ncRNA sequences and their related information (e.g. function, cellular role, cellular location, chromosomal information, etc.) in NONCODE have been confirmed manually by consulting relevant literature: more than 80% of the entries are based on experimental data. (iii) Based on the cellular process and function, which a given ncRNA is involved in, we introduced a novel classification system, labeled *process function class*, to integrate existing classification systems. (iv) In addition, some 1100 ncRNAs have been grouped into nine other classes according to whether they are specific to gender or tissue or associated with tumors and diseases, etc. (v) NONCODE provides a user-friendly interface, a visualization platform and a convenient search option, allowing efficient recovery of sequence, regulatory elements in the flanking sequences, secondary structure, related publications and other information. The first release of NONCODE (v1.0) contains 5339 non-redundant sequences from 861 organisms, including eukaryotes, eubacteria, archaebacteria, virus and viroids. Access is free for all users through a web interface at http://noncode.bioinfo.org.cn.

INTRODUCTION

Traditionally, most RNA molecules were regarded as carriers conveying information from the gene to the translation machinery. The most prominent exceptions to this are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are directly involved in the process of translation. However, since the late 1990s, it has been widely acknowledged that other types of non-protein-coding RNA molecules are present in organisms ranging from bacteria to mammals, which affect a large variety of processes including plasmid replication, phage development, bacterial virulence, chromosome structure, DNA transcription, RNA processing and modification, development control and others (1–16). These observations suggest that the traditional view of the structure of the genetic regulatory systems in organisms is far from complete. Therefore, further research on non-protein-coding RNA will give us a new framework for considering and understanding the genomic programming of biological complexity. However, the unsystematic naming of non-protein-coding RNAs may be an impediment to effective research. The term small RNAs (sRNAs) has been predominantly used for such RNAs in bacteria, whereas the term non-coding RNAs (ncRNAs) has been the most common term for eukaryotic RNAs of this kind (17,18). To have a common term for all such RNAs, we have opted to apply the term ncRNA to all these functional RNAs, irrespective of the realm of life in which they might appear.

The understanding of the importance of ncRNAs in basic cellular processes is ever increasing, and new members and

[*]To whom correspondence should be addressed. Tel: +86 10 6488 8543; Fax: + 86 10 6487 7837; Email: crs@sun5.ibp.ac.cn
Correspondence may also be addressed to Yi Zhao. Tel: +86 10 6256 5533 ext. 5717; Fax: + 86 10 6256 7724; Email: biozy@ict.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors classes of ncRNAs are continuously being reported. Thus, over the years, several databases have been established to collect, organize and classify ncRNA sequences and information. Some databases are intended to collect only certain category of ncRNAs, such as SRP RNAs, tmRNAs or RNase P RNAs, whereas others, such as the Small RNA Database, the Noncoding RNA Database and the Rfam Database, have collected ncRNAs of several categories (19–24). However, even in the latter kind of databases certain ncRNA members or classes are missing. Another problem with all the current databases is that the classification systems for ncRNAs used nowadays are not uniform and only a few attempts have been made to integrate the various classification systems. In these classification systems, some ncRNA groups are named according to cellular localizations, such as snRNAs, snoRNAs or scRNAs, some are named according to functions, like pRNAs (package RNAs), gRNAs (guide RNAs) or tmRNAs (transfer-messenger RNAs), and others again are simply labeled according to their sedimentation coefficients (6S RNA, 5.3S RNA, etc.). Furthermore, because of this lack of integration, one type of ncRNA often appears under several names or in more than one category (7,12,25–30).

The ncRNA database NONCODE was created against this background. NONCODE comprises almost all ncRNAs now publicly available (except tRNAs and rRNAs) that are either confirmed experimentally or predicted computationally. The first release of NONCODE (v1.0) contains 5339 non-redundant sequences from 861 organisms, including eukaryotes, eubacteria, archaebacteria, virus and viroids. Furthermore, to integrate existing classification systems, a new classification system labeled the *process function class* (PfClass) has been introduced, based on the cellular process and function in which a given ncRNA is involved. PfClass provides a unified classification system and a concise functional annotation of ncRNAs. According to the cellular process involved, 5339 ncRNAs were assigned one or more of 26 PfClasses. The PfClass classification system is the first attempt of a unified classification system for ncRNAs. It is our hope that this integrated system will help in clearing up the classification problem. In conclusion, the aim of the NONCODE database is to be a unified gateway to search, retrieve and update information about ncRNAs in order to facilitate research on ncRNAs, gene networks and functional genomics. Through a user-friendly web interface at http://noncode.bioinfo.org.cn, access is free for all users.

METHODS AND IMPLEMENTATION

NONCODE pipeline

GenBank entries were the major source of data, and the PubMed database was used as the starting point for the data collection (31). PubMed was first filtered using queries from a table of keywords, which includes ‘ncRNA’, ‘snoRNA’, ‘snRNA’, ‘tmRNA’, ‘SRP RNA’, ‘gRNA’, etc. The publications that matched with these queries were then examined and the ncRNA sequences were extracted from the obtained literature. By reading the filtered literature, a new set of ncRNA keywords were gained and added into the keywords table. This new keywords table was used to filter the GenBank BCT, INV, MAM, PHG, PLN, PRI, ROD, VRL and VRT divisions automatically, and the filtered result was then manually confirmed. The original sequence and annotation information were imported into the database powered by MySQL. All the data are integrated and organized in such a manner that users can efficiently query and browse information.

NONCODE annotation

One significant characteristic of NONCODE is its content of additional information on the ncRNAs obtained from the related literature. Briefly, seven steps were carried out after the GenBank screening. (i) For each sequence filtered from GenBank, we manually checked whether or not it represented an actual ncRNA and assigned the confirmed sequence an accession number (NcID, i.e. ncRNA id). (ii) Basic information—name, alias, length, organisms, references, etc.—of confirmed sequences was collected from GenBank. (iii) Additional information concerning function, cellular role, cellular location, etc. was included, by consulting relevant literature. Each ncRNA has also been annotated with one of the five specific mechanisms (sequence base pairing, structural complementarity, spatial blocking, catalysis or epimodification), through which it exerts its function. (iv) According to our PfClass classification system, one or more of the 26 PfClasses were assigned to all ncRNAs. Moreover, a subset of 1114 ncRNAs have been divided into nine additional categories according to whether they are specific to gender or tissue or associated with tumors and diseases, etc. (v) To visualize the location of an ncRNA in the genome or in a specific DNA fragment, along with regulatory elements in the flanking sequences, GenBank annotations were used to create figures for all ncRNAs. (vi) Each ncRNA sequence was checked for redundancies using Perl scripts, and each cluster of redundant sequences was given a non-redundant accession number (UniqID, i.e. unique ncRNA id). (vii) The secondary structures of non-redundant ncRNA sequences were predicted using the Vienna RNA Package (32). The predicted result in the PDF format is available through the website.

NONCODE process function classification

Ever since the beginning of ncRNA research there has not been in place any integrated system for classification, and therefore, exists a considerable measure of confusion with respect to naming of ncRNAs. This frequently brings about difficulties when ncRNAs from different sources are collected for analysis. Therefore, when the NONCODE database was established it was carefully considered as how to establish classification criteria that might increase the usefulness of the database resource.

The cellular process and function of an ncRNA was chosen as the basic criterion for a unified classification system called PfClass in NONCODE. When labeled according to this system, each kind of ncRNA is named after its cellular process and corresponding function. The actual category is given according to two or three levels of keywords connected by an underscore. The first keyword will be DNA, RNA or Protein, representing a cellular process in which either of the three molecular types is a crucial component. The second keyword describes the actual process, and if the ncRNA is involved in a complex process with several

Table 1. PfClasses in NONCODE v1.0 and their corresponding traditional classes

PfClass	Corresponding traditional classes
DNA_imprinting	XIST, roX, H19, MHM, KvLQT1-AS, Tsix, Air
DNA_packaging	pRNA
DNA_repair	RNA a, b, c, d
DNA_replication_initiation	RNAII
DNA_replication_regulation	ctRNA, RNA I
DNA_replication_repression	incA, RNA I
DNA_stability	telomerase RNA
DNA_transcription_initiation	RNA II
DNA_transcription_regulation	inc RNA, copA RNA, SRA
DNA_transcription_regulation of RNA polymerase	6S RNA, 7SK
DNA_transcription_repression	RNAI, GcvB RNA
RNA_editing	gRNA
RNA_modification_methylation	snoRNA
RNA_modification_methylation&pseudouridylation	scaRNA
RNA_modification_pseudouridylation	snoRNA
RNA_processing_cleavage	RNase P RNA, RNase MRP RNA, snoRNA
RNA_processing_splicing	snRNA, self-splicing ribozyme RNA, PAN
RNA_reverse_transcription	msr RNA
RNA_translation_enhancement	csrB RNA, DsrA RNA
RNA_translation_regulation	ANTI-RAF1, RprA, sok RNA, VA RNA, RyhB, sar RNA, NaPi-2b1, 5.3S RNA, aHIF
RNA_translation_suppression	miRNA, DicF, Spot 42, Finp, MicF, OxyS, flmB, PrrB_RsmZ, NTT, GcvB RNA, etc.
RNA_translation_surveillance	tmRNA
RNA_translocation	ScYC RNA, hsr-omega RNA, Xlsirt
Protein_transport	SRP_7SL RNA, SRP_4.5S RNA
Miscfunction_mRNAlike	BORG, IGF2AS, CR20, meuRNA, Rian, Ks-1, GNAS1-as RNA, IPW, etc.
Miscfunction_snm	Bsr RNA, Y RNA, dsrB, vault RNA, 4.5S RNA, 6Sa RNA, G8, etc.

The first column represents the PfClass classification system. Each PfClass is given according to two or three levels of keywords connected by an underscore ('_'). The first keyword will be DNA, RNA or Protein, representing a cellular process in which either of the three molecular types has a crucial function. The second keyword describes the actual process, and if the ncRNA is involved in a complex process with several aspects, a third keyword may further indicate a more specific function of the ncRNA. The second column lists corresponding traditional classes.

aspects, a third keyword may further indicate a more specific function of the ncRNA. For example, the snRNA U1 will be assigned to the PfClass *RNA_processing_splicing*, and RNase P RNAs to the PfClass *RNA_processing_cleavage* (for details see Table 1).

The PfClass classification system represents the first attempt of a unified classification system for ncRNAs. In the future, as our understanding of ncRNAs deepen, and the content of NONCODE further expands, steps will be taken to further extend and perfect the PfClass system in order to increase its usefulness. To further harmonize the exchange of data between different systems, application of Gene Ontology (GO) (33) annotation on our PfClass system will be considered.

CURRENT STATUS AND FUTURE DEVELOPMENTS

Till date, more than 10 000 sequences filtered from GenBank by our in-house program have been manually examined. The current release (v.1.0) of NONCODE contains a total of 6232 entries assigned to 26 PfClasses, and covers 109 traditional classes such as snRNA, snoRNA, microRNA and RNase P RNA. More than 80% of the entries are based on experimental data. Basic information on each entry is provided, including accession number in GenBank, traditional class, name, PfClass, organism, reference, UniqID (accession number without redundancy in NONCODE) and NcID (accession number with redundancy in NONCODE), all of which can be used as keywords for data search. NONCODE also provides additional information on function and cellular role, cellular location, chromosomal information, alternative names, secondary structure and whether or not the ncRNA has undergone splicing. Each ncRNA has also been annotated with one of the five specific mechanisms (sequence base pairing, structural complementarity, spatial blocking, catalysis or epimodification), through which it exerts its function. Figures showing genomic locations for all ncRNAs and their regulatory elements have been included, and a subdivision into nine additional classes (outside the PfClass system) has also been applied to a number of ncRNAs. NONCODE also offers an efficient search option, allowing recovery of sequence, related publications and other information.

In the near future, several aspects of NONCODE will be improved. (i) For a number of ncRNAs, information on function, location, etc. is still lacking, and this information will be completed as soon as it becomes available. (ii) As the information on ncRNAs increases and the content of NONCODE further expands, the PfClass system will be further extended and perfected in order to increase its usefulness. GO annotation on the PfClass system will also be considered seriously, with the aim of harmonized exchange of data between the different systems. (iii) Additional services such as BLAST alignment, ncRNAs prediction and possibilities for submission and registration of users' sequences will be provided. In addition, two large-scale screens for novel ncRNAs in *Caenorhabditis elegans* and human tissues are being carried out in our laboratory (Y. Wang, Z.Y. Sun, Y. Zhao, C.N. Liu, G. Skogerbø, W. Deng, Z. Fu, Y.D. Wang, L. Cai and H.S. He, unpublished data), and the results will be added in the next version of NONCODE. NONCODE is thus designed to adapt and to reflect the most current information on ncRNAs available. It will continue to grow in both content and functionality, and will be updated every six months to include any new data from literature and GenBank.

ACKNOWLEDGEMENTS

We are indebted to all of our colleagues who focus much of their research efforts on the analyses of ncRNAs. Sequence data were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genbank). This work was supported by the Chinese Academy of Sciences Grant No. KSCX2-2-27, National Sciences Foundation of China Grant Nos 39890070 and 60496320, the National High Technology Development Program of China under Grant No. 2002AA231031, National Key Basic Research & Development Program 973 under Grant Nos 2002CB713805 and 2003CB715900, and Beijing Science and Technology Commission Grant No. H010210010113.

REFERENCES

1. Hildebrandt,M. and Nellen,W. (1992) Differential antisense transcription from the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. *Cell*, **69**, 197–204.
2. Wagner,E.G. and Simons,R.W. (1994) Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.*, **48**, 713–742.
3. Lankenau,S., Corces,V.G. and Lankenau,D.H. (1994) The *Drosophila* micropia retrotransposon encodes a testis-specific antisense RNA complementary to reverse transcriptase. *Mol. Cell. Biol.*, **14**, 1764–1775.
4. Morfeldt,E., Taylor,D., von Gabain,A. and Arvidson,S. (1995) Activation of alpha-toxin translation in *Staphylococcus aureus* by the trans-encoded antisense RNA, RNAIII. *EMBO J.*, **14**, 4569–4577.
5. Sharp,T.V., Schwemmle,M., Jeffrey,I., Laing,K., Mellor,H., Proud,C.G., Hilse,K. and Clemens,M.J. (1993) Comparative analysis of the regulation of the interferon-inducible protein kinase PKR by Epstein-Barr virus RNAs EBER-1 and EBER-2 and adenovirus VAI RNA. *Nucleic Acids Res.*, **21**, 4483–4490.
6. Avner,P. and Heard,E. (2001) X-chromosome inactivation: counting, choice and initiation. *Nature Rev. Genet.*, **2**, 59–67.
7. Wassarman,K.M. and Storz,G. (2000) 6S RNA regulates *E.coli* RNA polymerase activity. *Cell*, **101**, 613–623.
8. Yang,Z., Zhu,Q., Luo,K. and Zhou,Q. (2001) The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature*, **414**, 317–322.
9. Nguyen,V.T., Kiss,T., Michels,A.A. and Bensaude,O. (2001) 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*, **414**, 322–325.
10. Frank,D.N. and Pace,N.R. (1998) Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.*, **67**, 153–180.
11. Will,C.L. and Luhrmann,R. (2001) Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biol.*, **13**, 290–301.
12. Guthrie,C. and Patterson,B. (1988) Spliceosomal snRNAs. *Annu. Rev. Genet.*, **22**, 387–419.
13. Kiss,T. (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, **20**, 3617–3622.
14. Kiss-Laszlo,Z., Henry,Y., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–1088.
15. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The *C.elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, **75**, 843–854.
16. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.
17. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.*, **2**, 919–929.
18. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
19. Rosenblad,M.A., Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T. (2003) SRPDB: Signal Recognition Particle Database. *Nucleic Acids Res.*, **31**, 363–364.
20. Zwieb,C., Gorodkin,J., Knudsen,B., Burks,J. and Wower,J. (2003) tmRDB (tmRNA database). *Nucleic Acids Res.*, **31**, 446–447.
21. Brown,J.W. (1999) The Ribonuclease P Database. *Nucleic Acids Res.*, **27**, 314.
22. Gu,J., Chen,Y. and Reddy,R. (1998) Small RNA database. *Nucleic Acids Res.*, **26**, 160–162.
23. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res.*, **31**, 429–431.
24. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
25. Balakin,A.G., Smith,L. and Fournier,M.J. (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.
26. Nakamura,K., Minemura,M., Nishiguchi,M., Honda,K., Nakamura,A. and Yamane,K. (1992) Conserved residues and secondary structure found in small cytoplasmic RNAs from thirteen *Bacillus* species. *Nucleic Acids Res.*, **20**, 5227–5228.
27. Hendrix,R.W. (1998) Bacteriophage DNA packaging: RNA gears in a DNA transport machine. *Cell*, **94**, 147–150.
28. Sugisaki,H. and Takanami,M. (1993) The 5′-terminal region of the apocytochrome b transcript in *Crithidia fasciculata* is successively edited by two guide RNAs in the 3′ to 5′ direction. *J. Biol. Chem.*, **268**, 887–891.
29. Keiler,K.C., Waller,P.R. and Sauer,R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, **271**, 990–993.
30. Zhanybekova,S.S.h., Polimbetova,N.S., Nakisbekov,N.O. and Iskakov,B.K. (1996) Detection of a new small RNA, induced by heat shock, in wheat seed ribosomes. *Biokhimiia*, **61**, 862–870 (in Russian).
31. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
32. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
33. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.