192–195 Nucleic Acids Research, 1999, Vol. 27, No. 1

Collection of mRNA-like non-coding RNAs

Volker A. Erdmann, Maciej Szymanski[1], Abraham Hochberg[2], Nathan de Groot[2] and Jan Barciszewski[1,*]

Institut fur Biochemie, Freie Universitat Berlin, Thielallee 63, 14195 Berlin, Germany, [1]Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Noskowskiego 12, 61704 Poznan, Poland and [2]Department of Biological Chemistry, The Hebrew University, IL-91904 Jerusalem, Israel

Received September 3, 1998; Revised and Accepted November 3, 1998

ABSTRACT

In last few years much data has accumulated which shows that in different cells various RNA transcripts are synthesized. They lack protein coding capacity and do not produce mature protein. It seems that they work mainly or exclusively on the RNA level. Their function and mechanism of action is poorly understood. In this paper we have collected all known RNA transcript and prepared a database for further structural and functional studies. This is the first collection of the nucleotide sequences of RNAs of this kind. The data can be accessed via WWW at: http://www.man.poznan.pl/5SData/ncRNA/index.html

INTRODUCTION

Eukaryotic genes contain clearly identifiable open reading frames (ORFs) that direct the translation of a biologically functional protein products. However, not all RNA molecules (other than tRNA or rRNA) are translated into polypeptides. Certain RNA molecules possess catalytic activity by themselves and act as ribozymes, show peptidyltransferase activity or as aptamers binding small molecular compounds. In addition, in the cell there are also stable mRNA-like transcripts (spliced and polyadenylated) without defined ORFs which are suggested to function on the RNA level in the absence of protein products and can be recognized as RNA riboregulators. They have been found in animals as well as plants. Although many efforts have been taken, their mechanisms of action still remain a mystery. This will be possible upon better understanding of their structure and the nature of their interactions with other components of the cell.

The aim of this work is to collect all non-translable RNAs known to date and provide a database for further studies in different laboratories.

CHARACTERISTICS OF DIFFERENT RNA TRANSCRIPTS

In this paper we collected the following six types of RNA transcripts: *Xist*, *H19*, *gadd7/adapt15*, *ENOD40*, *His*-1, *hsr*-omega and CR20. They have been shown to be expressed in different cells. However, all share common features: they are spliced, polyadenylated and lack long ORFs. It seems that they function as stable RNA molecules. Detailed analysis of their origin and properties enabled us to suggest some classes of RNA transcripts: (i) gene regulators, silencers (*Xist*, *roX1*, *roX2*, *H19*); (ii) abiotic stress signals (*gadd7*/*adapt15*, *adapt33*, *hsr*ω); (iii) biotic stress signals (*His-1*; *ENOD40*, CR20).

This classification is very preliminary and is based on a limited amount of functional data rather than on structural information. To verify this, many more examples of RNA transcripts with similar function are necessary.

Below we briefly describe all known RNA transcripts and show cross-reference data (Table 1).

Dosage-compensation regulatory RNAs (Xist, roX)

One of the fundamental differences between male and female cells is the number of X chromosomes. The difference in chromosome content and the requirement of equal expression of the chromosome X genes in both sexes led to the evolution of several types of dosage-compensation. In mammals and in *Drosophila*, this process involves expression of specific genes, whose products lack long ORFs, and seem to act as stable RNA transcripts that, together with specific proteins are responsible for chromatin remodelling.

In mammals, both X chromosomes are transcriptionally active at the early stages of XX embryo development. However, upon differentiation, a single X chromosome is inactivated. This process is usually random and there is an equal probability of inactivation of either of the X chromosomes.

This process is dependent on the X inactive specific transcript gene (*Xist*) located in a 450 kb long X inactivation center locus (Xic). The *Xist* gene is expressed exclusively from the inactive X chromosome and produces a large processed RNA, for which no protein product has been identified. This RNA is not exported from the nucleus and was found to be associated with the inactive X chromosome. In fact the inactivation of X chromosomes in mammals follows the '*n*–1' rule, that results in inactivation of all, except for one, X chromosome in the cell.

In *Drosophila* another mechanism of dosage-compensation has evolved. In contrast to the mammalian system where one of the X chromosomes in females is transcriptionally inactivated, this is achieved by hypertranscription of the single X chromosome in

*To whom correspondence should be addressed. Tel: +48 61 852 8503; Fax:+48 61 852 0532; Email:jbarcisz@ibch.poznan.pl

Table 1. List of the non-coding RNAs included in the database

RNA	Size	EMBL/GenBank Acc. No.	Notes
Xist, roX, PAT1			
Homo sapiens	16.5 kb	M97168	
Mus musculus	14.7 kb	L04961	
Equus caballus	nd	U50911	partial sequence
Oryctolagus cuniculus	nd	U50910	partial sequence
Drosophila melanogaster roX1	3749 nt	U85980	
Drosophila melanogaster roX2	1293 nt	U85981	
HZ-1 virus PAT-1	2937 nt	U03488	
H19			
Homo sapiens	2313 nt	M32053	
Mus musculus	1899 nt	X58196	
Rattus rattus	2297 nt	X59864	
Oryctolagus cuniculus	1842 nt	M97348	partial sequence
gadd7/adapt15, adapt33, vseap1			
Cricetulus griseus gadd7	754 nt	L40430	
Cricetulus griseus adapt15	746 nt	U26833	adapt15-P9
	753 nt	U26834	adapt15-P8
Cricetulus griseus adapt33	1290 nt	U29660	adapt33A
	1186 nt	U29661	adapt33B
Cricetulus griseus vseap1	0.9 kb	AJ003192	
	3.1 kb		
His-1			
Homo sapiens	nd	U56440	gene sequence, exon structure unknown
Mus musculus	3053 nt	U09772	alternatively spliced forms of the same pre-mRNA
	3003 nt	U10269	
hsr-ω			
Drosophila melanogaster	1174 nt	U18307	alternative poly(A) sites
	1190 nt		
Drosophila hydei	1129 nt	M14558; J02629	
Drosophila pseudoobscura	1213 nt	X16337; X16157;	
ENOD40			
Glycine max	679 nt	X69154	ENOD40-1
	617 nt	X69155	ENOD40-2
Pisum sativum	702 nt	X81064	
Phaseolus vulgaris	600 nt	X86441	
Vicia sativa	718 nt	X83683	
Tifolium repens	631 nt	AJ000268	
Lotus japonicus	770 nt	AF013594	
Medicago sativa	626 nt	X80263	
	733 nt	L32806	
Medicago truncatula.	920 nt	X80262	
Nicotiana tabacum	470 nt	X98716	
Vigna radiata	331 nt	AF061818	partial sequence
Sesbania rostrata	638 nt	Y12714	
CR20			
Cucumis sativus	1108 nt	D79216	
Arabidopsis thaliana	758 nt	D79218	

nd: not determined.

male cells. This process involves, among other factors, the products of the two genes *roX1* and *roX2*. The products of these genes, like *Xist* RNA, do not encode proteins, and their expression is male-specific. The mechanism of action of these RNAs resembles that of *Xist* RNA. *roX1* RNA becomes associated with the X chromosome at sites determined by binding of the *msl* (male specific lethal) gene products. This probably leads to the remodelling of the chromatin and allows increased transcription (1–6).

RNA species involved in dosage-compensation known to date include: human Xist RNA, mouse Xist RNA, *Drosophila roX1* and *roX2* RNA. Partial Xist RNA sequences have been determined for *Oryctolagus cuniclulus* and *Equus caballus*. Hz-1 virus persistence associated gene 1 (pag1) encoding a 2.9 kb non-coding viral PAT1 RNA shows similarity to Xist.

H19 RNA

H19 RNA is an oncofoetal RNA expressed at high levels in placenta and several foetal tissues. After birth, H19 expression in both man and mouse is strongly decreased in different tissues and drops to undetectable levels except for skeletal muscles, heart and mammary cells. It is re-expressed in tumors derived from embryonic tissues, which express the H19 gene. The gene is one of known imprinted genes. It is expressed exclusively from a maternal chromosome and is linked and co-regulated with the insulin-like growth factor 2 gene, which is also imprinted but expressed primarily from the paternal chromosome. The imprinting results from the methylation of a 7–9 kb domain surrounding the paternal allele of H19. On the other hand, the maternal allele is relatively unmethylated and displays an open chromatin conformation. The loss or inactivation of the

maternal copy of H19 was found to be associated with several tumors, including Wilms' tumor.

The H19 gene product is 2.3 kb long RNA, and is spliced and polyadenylated. The genes isolated from human and mouse consist of five exons separated by four unusually short introns. No conserved ORF was found by sequence comparison. In human sequence, there is a putative ORF that would encode a 26 kDa protein, but no translation product has been identified so far. Apparently, the gene product plays its role as an RNA. Until today, nothing is known about the function of H19 RNA. The re-expression in certain types of tumours led to the conclusion that it can act as a tumour suppressor, but this has not been confirmed (7–9; C.Lipmann, T.Specht, A.Hochberg and V.A.Erdmann, submitted).

H19 RNAs known to date are from human, mouse, rat and rabbit.

Oxidative stress response RNAs (*gadd7*/*adapt15*, *adapt33*)

The oxidative stress caused by exposure to UV radiation or reactive oxygen species is responsible for significant damage in biological cells. The effects of this stress involve inactivation of enzymes and transport proteins, peroxidation of lipids, DNA damage and mutations, as well as cleavage of cellular macromolecules. Oxidative stress has been shown to be responsible for several human disorders, including, among others, cataracts, artherosclerosis, cancer as well as ageing. Although much is known about the chemical and biochemical consequences of oxidant exposure, little is known about its effects on gene expression. In bacteria, a number of genes are known to be modulated by oxidative stress. These include catalase, superoxide dismutase and glutathione reductase, that are involved in detoxication of the cells. In mammals several groups of genes are expressed in response to UV irradiation or exposure to hydrogen peroxide and other reactive oxygen species. Transcripts of some of these genes apparently lack protein products and are likely to act as RNA.

One of these genes, *gadd7*, is expressed in response to treatment with UV radiation. It belongs to the family of *gadd* genes induced by various types of growth arrest signals and by DNA damage. The *gadd7* transcript is a 0.9 kb long polyadenylated RNA species, lacking a long ORF. Sequence analysis showed that there are three short ORFs for 38, 37 and 43 amino acid long peptides. However, *in vitro* translation of the *gadd7* transcript did not reveal any protein product. The *gadd7* RNA may play its function in the regulation of other genes following DNA damage.

Other RNA species produced in response to oxidative stress, induced by hydrogen peroxide are *adapt15* and *adapt33*. Their induction was correlated with the adaptive response to H_2O_2. Transcription products of both of these genes lack long ORFs and similarly to *gadd7* RNA are polyadenylated. The *adapt15* RNA is 0.95 kb long and is almost identical to *gadd7* RNA, whereas for *adapt33*, two homologous RNA species of 1.46 and 0.99 kb have been isolated (10–13).

gadd/adapt RNAs known to date have been isolated from Chinese hamster cells and include *gadd7*, *adapt15 and adapt33* (two species). Also v-src end associated peptide 1 RNA (vseap1) shows high similarity to gadd7/adapt 15 RNA.

Heat shock response RNA (hsrω)

Protection against various environmental stresses is mainly conferred by the induction of heat shock genes. This mechanism is common to both prokaryotic and eukaryotic organisms. The amino acid sequences of heat shock-induced proteins isolated from a variety of organisms are highly conserved during evolution. In *Drosophila* a major site of transcription in heat shock is the *hsr*ω gene. It is located in the polytene region 93D. An interesting feature of this region is that it is induced independently of the rest of the heat shock genes, encoding all major groups of heat shock proteins. The product of this gene is polyadenylated and spliced RNA and has only very small ORFs, that show very little conservation among different species. The expression of *hsr*ω is constitutive and it is elevated by heat shock. Most of the *hsr*ω transcripts are located in the nucleus. In both normal and stressed cells the transcription from the *hsr*ω locus gives rise to three transcripts: ω*1*, ω*2* and ω*3*. All of them have the same transcription start site. The hsrω*1* transcript is ~10 kb long and results from transcription of the whole locus. This RNA is accumulated at very high levels in the nucleus. At the 3′-end of hsrω*1* transcript there is a 7–8 kb long region consisting of a short tandem repeat unit. The hsrω*2* (~1.9 kb) transcript also accumulates in the nucleus and results from the use of alternate polyadenylation site located just upstream of the tandem repeats region. The ω*3* transcript (1.2–1.3 kb) is produced from ω*2* by removal of a single intron and is a cytoplasmic species. This RNA does not contain any long ORFs. One of the short ORFs, that is conserved in three *Drosophila* species would encode a peptide 23–27 amino acids long. It is likely that this RNA does not act as a messenger for protein synthesis but performs other function as RNA. Some results suggest that the short ORF is translated, but no accumulation of protein product was observed. It seems that the sole act of translation and not generation of functional protein product may be important (14–17).

The sequences of *hsr*ω genes have been determined for *Drosophila melanogaster*, *Drosophila hydei* and *Drosophila pseudoobscura*.

His-1 RNA

Upon viral insertion in murine myeloid leukemias, the *His-1* gene is activated. It is not expressed in uninfected cells. The *His-1* gene is a single-copy sequence that has been found in a variety of vertebrate species. *His-1* RNA is expressed as a 3 kb long spliced and polyadenylated RNA. An analysis of the RNA sequence did not reveal any ORF that would exceed 219 nt. The lack of long ORF for this RNA suggests that it can function in the absence of an encoded protein. *His-1* gene was shown to be expressed at low levels in the epithelial cells of the adult mouse stomach, prostate, seminal vesicle and the developing choroid plexus. The *His-1* RNA transcription is correlated with viral insertion and carcinogenesis, since no transcripts were detected in normal tissues. They can be readily identified in mouse leukemias and carcinomas. This finding suggests that the expression of the *His-1* gene is highly restricted and that its inappropriate activation may contribute to carcinogenesis (18–20).

His-1 RNAs known to date are those from human and mouse.

Early nodulin 40 (*ENOD40*)

The genes that are activated in plants by Nod (nodulin) factors are called nodulins. In the symbiosis between rhizobia and legumes, nitrogen-fixing nodules are formed as the outcome of a complex process that includes new organ development, microbial invasion of plant tissues, internalization of bacteria in plant cells and

functional differentiation of the two partners. During nodule development, several plant genes are expressed in subsequent stages. Genes that are transcribed early in the interaction (*ENOD* genes) are hypothesized to play a role in organogenesis and bacterial invasion. Several *ENOD* clones have been isolated, but their individual contribution to nodule formation is often unclear. Some of the nodulins, including *ENOD40*, are employed as early markers to study the initiation of nodulation. One of them, *ENOD40*, an early nodulin gene, is expressed following inoculation with *Rhizobium meliloti* or by adding *R.meliloti*-produced nodulation (Nod) factors or the plant hormone cytokinin to uninoculated roots. It is detectable in the root pericycle opposite the nodule primordium even before the appearance of infection threads, and is also found later on, associated with vascular strands in mature nodules. Comparison of the *enod40* sequence isolated from several legumes and one non-legume did not reveal any conserved large ORF. Instead, a conserved region was found which may allow the production of a particularly stable cytoplasmic RNA. Therefore it has been proposed that *enod40* encodes an RNA with regulatory function. In addition *ENOD40* also shows a very short ORF of only 10–13 amino acids, but the 3′-untranslated region appears to have important functions as a riboregulator (21–24).

ENOD40 RNA known to date are those from: *Glycine max*, *Pisum sativum*, *Phaseolus vulgaris*, *Vicia sativa*, *Tifolium repens*, *Lotus japonicus*, *Medicago sativa*, *Medicago truncatula*, *Nicotiana tabacum*, *Vigna radiata* and *Sesbania rostrata*.

Cytokinin response RNA (CR20)

The CR20 gene has been found to be one of several genes repressed by cytokinins in excised cotyledons of cucumber. Detailed analysis showed that there are several CR20 transcripts generated by alternative splicing of the precursor RNA. However, none of the isolated and sequenced CR20 transcripts contained a long ORF. A comparison of the two known CR20 sequences from cucumber and *Arabidopsis thaliana* revealed that, although they show little overall homology, there is a highly conserved 180 nt region that seems to form a stable secondary structure (25). This suggests that this RNA is not translated into a protein and may function as RNA.

DESCRIPTION AND AVAILABILITY OF THE DATABASE

The database provides an access to the described sequences of the various non-coding mRNA-like RNAs. The files contain the nucleotide sequences as well as necessary accessory information in EMBL format. The sequences are available from hypertext lists for each particular RNA type, which makes access straightforward.

To make the database access as fast as possible, all individual sequences are stored, and can be retrieved as separate files. The database is accessible via the World Wide Web at the following URL: http://www.man.poznan.pl/5SData/ncRNA/index.html . Any comments, suggestions or corrections are welcome at the Email address: jbarcisz@ibch.poznan.pl.

ACKNOWLEDGEMENTS

This work has been supported by the Deutsche Forschungsgemeinschaft (Gottfried Wilhelm Leibnitz Prize to V.A.E.), the Sonderforschungsbereich 344-C8, the DFG Trilateral Research Project ‘Genomic Imprinting in Human Bladder Cancer’, the Deutsche Agentur fur Raumfahrtangelegenheiten GmbH, the Fonds der Chemischen Industrie e.V. and the Polish State Committee for Scientific Research.

REFERENCES

1 Willard,H.F. and Salz,H.K. (1997) *Nature*, **386**, 228–229.
2 Kuroda,M.I. and Meller,V.H. (1997) *Cell*, **91**, 9–11.
3 Panning,B. and Jaenisch,R. (1998) *Cell*, **93**, 305–308.
4 Clerc,P. and Avner,P. (1998) *Nature Genet.*, **19**, 249–253.
5 Brockdorf,N. (1988) *Curr. Opin. Genet. Dev.*, **8**, 328–333.
6 Moulton,C.M., Chow,J.C., Brown,C.J. and Lawrence,J.B. (1998) *J. Cell. Biol.*, **142**, 13–23.
7 Hao,Y., Crenshaw,T., Moulton,T., Newcomb,E. and Tycko,B. (1993) *Nature*, **365**, 764–767.
8 Joubel,A., Curgy,J.-J., Pelczar,H., Begue,A., Lagrou,C., Stehelin,D. and Coll,J. (1996) *Cell. Mol. Biol.*, **42**, 1159–1172.
9 Pfeifer,K., Leighton,P.A. and Tilghman,S.M. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 13876–13883.
10 Hollander,M.C., Alamo,I. and Fornace,A.J.,Jr (1996) *Nucleic Acids Res.*, **24**, 1589–1593.
11 Wang,Y., Crawford,D.R. and Davies,K.J.A. (1996) *Arch. Biochem. Biophys.*, **332**, 255–260.
12 Crawford,D.R., Schools,G.P., Salmon,S.L. and Davies,K.J.A. (1996) *Arch. Biochem. Biophys.*, **325**, 256–264.
13 Mizenina,O., Yanushevich,Y., Musatkina,E., Rodina,A., Camonis,J., Tavitian,A. and Tatosyjan,A. (1998) *FEBS Lett.*, **422**, 79–84.
14 Fini,M.E., Bendena,W.G. and Pardue,M.L. (1989) *J. Cell. Biol.*, **108**, 2045–2057.
15 Lakhotia,S.C. and Sharma,A. (1995) *Chromosome Res.*, **3**, 151–161.
16 Lakhotia,S.C. and Sharma,A. (1996) *Genetica*, **97**, 339–348.
17 McKechnie,S.W., Halford,M.M., McColl,G. and Hoffmann,A.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 2423–2428.
18 Askew,D.S., Li,J. and Ihle,J.N. (1994) *Mol. Cell Biol.*, **14**, 1743–1751.
19 Li,J., Rhodes,J.C. and Askew,D.S. (1997) *Gene*, **184**, 169–176.
20 Li,J., Witte,D.P., Van Dyke,T. and Askew,D.S. (1997) *Am. J. Pathol.*, **150**, 1297–1305.
21 Crespti,M.D., Jurkevitch,E., Poiret,M., d’Aubenton-Carafa,Y., Petrovics,G., Kondorosi,E. and Kondorosi,A. (1994) *EMBO J.*, **13**, 5099–5112.
22 van de Sande,K., Pawlowski,K., Czaja,I., Wieneke,U., Schell,J., Schmidt,J., Walden,R., Matvienko,M., Wellink,J., van Kammen,A., Franssen,H. and Bisseling,T. (1996) *Science*, **273**, 370–373.
23 van de Sande,K. and Bisseling,T. (1997) *Essays Biochem.*, **32**, 127–142.
24 Corich,V., Goormachting,S., Lievens,S., Van Montagu,M. and Holsters,M. (1998) *Plant Mol. Biol.*, **37**, 67–76.
25 Teramoto,H., Toyama,T., Takeba,G. and Tsuji,H. (1996) *Plant Mol. Biol.*, **32**, 797–808.