J. Appl. Genet. 44(1), 2003, pp. 1-19

Invited Editorial

Noncoding RNA transcripts

Maciej SZYMAŃSKI, Mirosława Z. BARCISZEWSKA, Marek ŻYWICKI,
Jan BARCISZEWSKI

Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

Abstract. Recent analyses of the human genome and available data about the other higher eukaryotic genomes have revealed that, in contrast to Eubacteria and Archaea, only a small fraction of the genetic material (ca 1.5%) codes for proteins. Most of genomic DNA and its RNA transcripts are involved in regulation of gene expression, which can be exerted at either the transcriptional level, controlling whether a gene is transcribed and to what extent, or at the post-translational level, regulating the fate of the transcribed RNA molecules, including their stability, efficiency of their translation and subcellular localization. Noncoding RNA genes produce functional RNA molecules (ncRNAs) rather than encoding proteins. These stable RNAs act by multiple mechanisms such as RNA-RNA base pairing, RNA-protein interactions and intrinsic RNA activity, as well as regulate diverse cellular functions, including RNA processing, mRNA stability, translation, protein stability and secretion. Non-protein-coding RNAs are known to play significant roles. Along with transfer RNAs, ribosomal RNAs and mRNAs, ncRNAs contribute to gene splicing, nucleotide modification, protein transport and regulation of gene expression.

Key words: gene expression, imprinting, noncoding RNA, silencing.

Introduction

The regulation of gene expression is a fundamental aspect of biological phenomena, such as the responses to environmental conditions, development of multicellular organisms, morphology and disease. Gene regulatory patterns are extraordinarily diverse and complex, yet the regulation of each gene is precise

Received: June 6, 2002. Accepted: September 12, 2002.

Correspondence: J. BARCISZEWSKI, Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Noskowskiego 12/14, 61-704 Poznań, Poland, e-mail: jan.barciszewski@ibch.poznan.pl

with respect to when and how much expression occurs. Gene regulation is remarkably flexible, both to rapidly alter the network of genes expressed in response to new conditions and to accommodate to evolutionary demands.

Currently, we are far away from an initial view of molecular mechanisms underlying cellular functions that was established over forty years ago. The central dogma of molecular biology defined a general pathway for the expression of genetic information stored in DNA, transcribed into transient messenger RNAs and decoded on ribosomes with the help of adapter RNAs (transfer RNAs) to produce proteins that were supposed to perform all enzymatic and structural functions in the cell. According to that model, ribonucleic acids (RNAs) plays a rather accessory role and the complexity of organisms is defined solely by the number of proteins encoded in a genome according to the "one gene-one protein" rule. That rather simple picture got complicated with finding of primary transcripts of eukaryotic protein genes in which coding sequences were interrupted by noncoding fragments (introns) that are excised and discarded during pre-mRNA maturation (pre-mRNA splicing). Subsequently, it was realized that in some cases they also provide means for synthesis of more than one protein product from a single gene by alternative splicing.

During the past twenty years it has been shown that in the cell there is a variety of RNA molecules that display a remarkable range of functions far beyond those already known for messenger (mRNA), ribosomal (rRNA) and transfer RNA (tRNA). This huge versatility is mainly due to chemical properties of RNA, which allow it to form complex tertiary structures capable of performing many roles that for many years were thought to be an exclusive domain of proteins.

General properties of ribonucleic acids

RNA is a ubiquitous cellular biopolymer (20% of *E. coli* dry weight), (MATTICK 2001). It is involved in all aspects of the maintenance, transfer and processing of genetic information. RNA shows unique properties as a biomolecule, since it can serve a role in the coding and decoding (by specific Watson-Crick base pairing) as well as processing of genetic information by forming intricately structured, often catalytically active components of the processing machinery. It acquires complex folded conformations that can participate in sophisticated recognition processes. RNAs provide recognition elements for protein binding, form large macromolecular complexes, and directly (RNA catalysis) or indirectly (RNP catalysis) catalyse numerous chemical reactions in the cell. RNA tertiary structures can form a virtually unlimited number of highly specific ligand-binding sites. RNA can interact with chemically and structurally diverse sets of small compounds, which exert profound effects on the biological function of the target. Detailed structural studies of antibiotics bound to the ribosome have revealed that the small-molecular-compound recognition mechanism involving rRNA is based on combination of shape recognition and both electrostatic and hydrogen bonding interactions (MOORE, STEITZ 2002). However, the dynamic nature of RNA structures, together with the presence of associated proteins that could displace the most strongly bound ligand, makes RNA an especially difficult species to target with high-affinity molecules (aptamers).

The higher-order structures of many RNAs remain unknown, and the catalytic mechanisms for RNAs are poorly understood in general. The genetic information encoded as DNA in most living organisms is copied into mature RNAs which are folded into arrays of tertiary structures. Although there are many steps at which mRNA expression can be regulated, the only ones where stable higher-order complexes are known to reproducibly and predictably inhibit mRNA function are at the level of splicing and translation (FILIPOWICZ, POGACIC 2002). The high complexity of regulation of gene expression through RNA metabolism increases with organism and tissue organization, e.g. brain cells provide unusually abundant examples of regulation by alternative RNA processing and small noncoding RNAs (EDDY 2001). When RNA and protein bind each other, recognition occurs by induced fit mechanism rather than by rigid "lock and key" docking (MOORE, STEITZ 2002).

In addition to protein synthesis, several RNA-based processes are known and regulatory mechanisms have been documented (ERDMANN et al. 2001). Many fascinating discoveries of the last two decades, together with a fast-growing number of new functional RNAs, led to a hypothesis of a primordial RNA world, where both information and enzymatic functions are carried out by RNA molecules (EDDY 2001). However, in the course of evolution, most of the catalytic functions were taken over by proteins and the role of a major carrier of genetic information was acquired by chemically more stable DNA. It seems that those catalytic RNAs are not only molecular fossils left from all-RNA organisms, but they play important roles in extant organisms. This is particularly clear seeing the results of genome sequencing, which show that the protein-coding genes alone are not enough to account for the observed complexity of higher organisms.

Noncoding RNA

The draft of the human genome, though still incomplete, clearly reveals that coding sequences account only for less than 2% of its total (Table 1). A similar phenomenon is observed in other eukaryotic genomes (SZYMAŃSKI, BARCISZEWSKI 2002). Repeated sequences make up at least 50% of the total human genome. Among the different types of repeats, transposon-derived ones predominate (~45% of the genome), particularly retroelements including short interspersed nuclear elements (SINEs or Alu repeats, ~13%), long interspersed nuclear elements (LINEs, ~20%) and long terminal repeats (LTR) containing retroelements (~8%). Of other types of repeat, the most frequent are short tandem repeats (STRs), such

Table 1. Results of an analysis of features of genomes. based on their complete DNA sequences. Protein-coding part size means size of all open reading frames; number of genes means number of protein-coding genes; gene size means length of open reading frame. ΔG/C means difference between G/C rate in coding and noncoding part.

A. Eubacterial genomes. *Mezorhizobium loti* genome has not been analysed. because its file is deposited without a CDS (coding sequence annotation)

Organism	Accession number	Genome size (bp)	Noncoding part size (%)	Noncoding part size (bp)	Protein-coding part size (%)	Protein-coding part size (bp)	Gene number	Max. gene size (bp)	Min. gene size (bp)	Average gene size (bp)	G+C in genome	G+C in coding part (%)	Max. G+C in coding part (%)	Min. G+C in coding part (%)	G+C in coding part / G+C in genome	G+C in noncoding part (%)	G+C in noncoding part / G+C in genome	Δ G+C (%)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nostoc sp. PCC 7120	BA000019	6413771	19	1229654	81	5184117	5230	9789	54	991	41	42	56	24	1.024	37	0.902	5
P. aeruginosa	AE004091	6264403	15	923230	85	5341173	5236	16884	72	1020	67	67	76	30	1.000	64	0.955	3
S. typhimurium LT2	AE006468	4857432	17	833058	83	4024374	4144	16680	45	971	52	53	70	23	1.019	47	0.904	6
S. enterica subsp. *enterica serovar Typhi*	AL513382	4809037	16	787416	84	4021621	4314	10875	30	932	52	53	73	21	1.019	47	0.904	6
Y. pestis	AL590842	4653728	19	901595	81	3752133	3840	11118	45	977	48	49	63	24	1.021	43	0.896	6
E. coli	U00096	4639221	16	743997	84	3895224	3987	7152	45	977	51	52	67	27	1.017	45	0.882	7
M. tuberculosis CDC1551	AE000516	4403836	16	694741	84	3709095	3760	12456	93	986	66	66	81	48	1.000	65	0.985	1
B. halodurans	BA000004	4202353	18	775154	82	3427199	3845	5451	36	891	44	44	55	18	1.000	41	0.932	3
V. cholerae chl	AE003852	2961149	17	490847	83	2470302	2531	13677	81	976	48	48	56	26	1.000	45	0.937	3
V. cholerae ch2	AE003853	1072315	18	194872	82	877443	1020	9792	93	860	47	48	56	26	1.021	44	0.936	4
C. crescentus	AE005673	4016947	15	600730	85	3416217	3417	7440	93	1000	67	68	76	36	1.015	65	0.970	3
C. acetobutylicum	AE001437	3940880	17	651569	83	3289311	3526	8613	87	933	31	32	46	17	1.032	28	0.903	4
R. solanacearum	AL646052	3716413	15	569014	85	3147399	3231	12807	63	974	67	68	81	39	1.015	64	0.955	4
S. meliloti	AL591688	3654135	17	635364	83	3018771	3161	8499	123	955	63	63	71	48	1.000	59	0.936	4
Synechocystis PCC6803	AB001339	3573470	16	564746	84	3008724	3023	12600	85	995	48	49	59	25	1.021	43	0.896	6
M. leprae	AL450380	3268203	27	866227	73	2401976	2584	9231	82	930	58	59	68	42	1.017	55	0.948	4

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
L. innocua	AL592022	3011208	15	437928	85	2573280	2804	6504	114	918	37	38	52	23	1.027	35	0.946	3
L. monocytogenes EGD-e	NC_003210	2944528	14	419965	86	2524563	2698	6135	87	936	38	38	52	24	1.000	35	0.921	3
S. aureus subsp. *aureus Mu50*	BA000017	2878040	19	559987	81	2318053	2558	20142	51	906	33	34	53	19	1.030	30	0.909	4
A. tumefaciens	AE007869	2841581	16	457745	84	2383836	2481	8409	93	961	59	60	69	37	1.017	55	0.932	5
D. radiodurans ch1	AE000513	2648638	17	452620	83	2196018	2366	5823	114	928	67	68	77	34	1.015	64	0.955	4
D. radiodurans ch2	AE001825	412348	19	80215	81	332133	308	4881	168	1078	67	67	76	37	1.000	65	0.970	2
X. fastidiosa	AE003849	2679306	21	564504	79	2114802	2538	10368	93	833	53	54	72	20	1.019	49	0.924	5
L. lactis subsp. *lactis*	AE005176	2365589	18	429929	82	1935660	2148	5952	96	901	35	36	47	19	1.028	31	0.886	4
N. meningitidis	AE002098	2272351	24	547687	76	1724664	1938	8112	69	890	52	53	65	18	1.019	46	0.885	7
P. multocida	AE004439	2257487	13	295205	87	1962282	1941	11760	114	1011	40	41	48	23	1.025	36	0.900	5
S. pneumoniae	AE007317	2038615	18	374374	82	1664241	1872	7656	63	889	40	41	49	21	1.025	36	0.900	5
T. maritima	AE000512	1860725	17	317516	83	1543209	1555	5073	93	992	46	46	68	30	1.000	45	0.978	1
H. influenzae Rd	L42023	1830138	17	310596	83	1519542	1628	5085	66	933	38	39	50	21	1.026	35	0.921	4
H. pylori 26695	AE000511	1667867	15	252830	85	1415037	1449	8682	39	977	39	40	61	22	1.026	35	0.897	5
C. jejuni	AL111168	1641481	15	238121	85	1403360	1429	4554	45	982	31	31	44	10	1.000	28	0.903	3
A. aeolicus	AE000657	1551335	18	284576	82	1266759	1267	4725	144	1000	43	44	51	26	1.023	42	0.975	2
R. conorii	AE006914	1268755	23	289972	77	978783	1269	6066	69	771	32	33	46	17	1.031	31	0.969	2
C. pneumoniae AR39	AE002161	1229858	16	191771	84	1038087	1024	5481	93	1014	41	41	53	20	1.000	37	0.902	4
T. pallidum	AE000520	1138011	16	177288	84	960723	909	4602	93	1057	53	53	69	38	1.000	54	1.019	−1
R. prowazekii	AJ235269	1111523	27	302393	73	809130	788	7023	69	1027	29	30	42	19	1.034	25	0.862	5
M. pulmonis	AL445566	963879	15	140511	85	823368	714	9651	114	1153	27	27	53	17	1.000	22	0.815	5
M. pneumoniae	U00089	816394	19	152476	81	663918	620	5649	114	1071	40	41	56	28	1.025	36	0.900	5
U. urealyticum	AF222894	751719	12	90609	88	661110	577	15018	105	1146	25	26	38	16	1.040	23	0.920	3
M. genitalium	L43967	580074	19	109254	81	470820	414	5418	114	1137	32	32	44	21	1.000	32	1.000	0

Table 1B. Archaeal genomes. Two already known archaeal genomes have not been taken into account: *Pyrococcus furiosus* genome file is deposited without CDS annotation, but annotation of *Aeropyrum pernix* genome file is erroneous (NATALE et al. 2000)

Organism	Accession number	Genome size (bp)	Noncoding part size (%)	Noncoding part size (bp)	Protein-coding part size (%)	Protein-coding part size (bp)	Gene number	Max. gene size (bp)	Min. gene size (bp)	Average gene size (bp)	G+C in genome	G+C in coding part (%)	Max. G+C in coding part (%)	Min. G+C in coding part (%)	G+C in coding part / G+C in genome	G+C in noncoding part (%)	G+C in noncoding part / G+C in genome	Δ G+C (%)
S. solfataricus	AE006641	2992245	23	702612	77	2289633	2610	4281	123	877	36	37	58	23	1.028	33	0.917	4
S. tokodaii	BA000023	2694765	23	617652	77	2077113	2436	4329	153	853	33	34	67	19	1.030	30	0.909	6
P. aerophilum	AE009441	2222430	20	446760	80	1775670	2255	8358	57	787	51	52	75	17	1.020	49	0.961	3
A. fulgidus	AE000782	2178400	19	414967	81	1763433	2035	7278	78	867	49	49	61	28	1.000	45	0.918	6
Halobacterium sp. *NRC-1*	AE004437	2014239	17	340995	83	1673244	1908	4113	93	877	68	68	78	40	1.000	65	0.956	3
P. abyssi	AL096836	1765118	18	315813	82	1449305	1523	6369	57	952	45	45	56	32	1.000	42	0.933	3
M. thermo-autotrophicum	AE000666	1751377	14	239149	86	1512228	1742	5364	108	868	50	51	60	26	1.020	43	0.860	8
P. horikoshii	BA000001	1738505	13	223958	87	1514547	1636	13311	71	926	42	42	70	29	1.000	39	0.929	3
M. jannaschii	L77117	1664970	17	279717	83	1385253	1599	8685	69	866	31	32	47	21	1.032	29	0.935	3
T. volcanium	BA000011	1584804	20	316583	80	1268221	1384	6231	153	916	40	41	50	20	1.025	36	0.900	5
T. acidophilum	AL139299	1564906	19	290368	81	1274538	1339	6246	138	952	46	47	60	28	1.022	40	0.870	7

Table 1C. Eukaryotic genomes. Analysis is based on published information on human protein set from SWISSPROT (The Arabidopsis Genome Initiative, 2000; Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408, 796-815; The *C. elegans* Sequencing Consortium, 1998; Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012-2018; see also ADAMS et al. 2000; KATINKA et al. 2001, VENTER et al. 2001, WOOD et al. 2002). Protein-coding part size means total size of exons; gene size means total exons size of gene. *Guilardia theta* nucleomorph genome has not been taken into account, because of high selectivity of nucleomorph genome

Organism	Genome size (bp)	Noncoding part size (%)	Noncoding part size (bp)	Protein-coding part size (%)	Protein-coding part size (bp)	Gene number	Average gene size (bp)	G+C in genome (%)	G+C in coding part (%)	G+C in coding part / G+C in genome	G+C in noncoding part (%)	G+C in noncoding part / G+C in genome	Δ G+C (%)
H. sapiens	2910000000	98	2849607984	2	60392016	39114	1544*	41	50	1.219	42	1.024	8
D. melanogaster	180000000	86.6	155900000	13.4	24100000	13601	1773	57.7	–	–	–	–	–
A. thaliana	115409949	71.2	82160699	28.8	33249250	25498	1303	34.9	44	1.294	33	0.971	11
C. elegans	97000000	73	70810000	27	26190000	18424	1421	36	–	–	–	–	–
S. pombe	12462637	43	5358934	57	7103703	4824	1426	36	40	1.111	31	0.861	9
S. cerevisiae	12000000	29.5	3540000	70.5	8460000	5651	1497	38.3	40	1.053	33	0.868	7
E. cuniculi	2900000	10	290000	90	2610000	1997	1307	46	47	1.022	45	0.978	2

Table 1D. Viral RNA genomes

Particle name	Accession number	Particle size	Noncoding part size (%)	Noncoding part size (bp)	Coding part size (%)	Coding part size (bp)	Number of genes	G+C in particle (%)	G+C in coding part (%)	G+C in coding part / G+C in particle	G+C in noncoding part (%)	G+C in noncoding part / G+C in particle	G+C in structural part (%)	G+C in structural part / G+C in particle	ΔG+C
Tobacco mosaic virus	NC_001367	6395	4	272	96	6123	6	43	43	1.000	43	1.000	55	1.279	0
Kennedya yellow mosaic virus	NC_001746	6362	3	163	97	6199	3	54	55	1.018	51	0.944	65	1.204	4
Eggplant mosaic virus	NC_001480	6331	4	233	96	6098	3	54	54	1.000	54	1.000	63	1.167	0
Turnip yellow mosaic virus	NC_001509	6319	3	191	97	6128	3	55	55	1.000	48	0.873	60	1.091	7
Satellite tobacco mosaic virus	NC_003796	1058	44	468	56	590	2	46	46	1.000	46	1.000	52	1.130	0
Escherichia coli tmRNA	ECU68074	363	91	330	9	33	1	53	42	0.792	54	1.019	56	1.057	–12

as $(A)_n$, $(CA)_n$ or $(CGG)_n$. Although these occupy only about 1% of the genome, the total number of such repeats in the genome is about 10^5 (MATTICK 2001).

In proteome-oriented analyses of genomic sequences, only mRNA-coding genes are taken into account, and those that produce non-protein-coding transcripts are often ignored. From genomic analyses it is however evident that with an increase in organism complexity, the protein-coding contribution of a genome decreases (Table 1). It is estimated, that up to 98% of the transcriptional output of eukaryotic genomes consist of RNA that does not encode protein (ADAMS et al. 2000). This includes introns and transcripts from other non-protein-coding genes, which can account for 50-75% of all transcription in higher eukaryotes. Over the past 10 years, RNA molecules not encoding proteins, called noncoding RNAs (ncRNAs) turned out to be remarkably versatile and to play various roles in prokaryotic and eukaryotic cells (EDDY 2001, ERDMANN et al. 2000, 2001).

Numbers of protein-coding RNA genes in complete eukaryotic genomes are much lower than initially expected (Table 1) (WOOD et al. 2002). *Caenorhabditis elegans* and *Drosophila melanogaster* genomes contain only twice as much genes as yeasts or some bacteria. In the human genome the number is doubled relative to invertebrates (ADAMS et al. 2000, VENTER et al. 2001). About 99% of the estimated 40 000 human protein-coding genes have orthologs within the mouse genome. The analyses of the human genomic sequences showed that about 99.9% of differences between genomes of individual humans are located outside sequences encoding proteins. They essentially do not work at all for one class of genes – the noncoding RNA genes, which produce transcripts that function directly as structural, catalytic or regulator RNAs (Table 2). The knowledge of ncRNAs is still limited to biochemically abundant species and occasional discoveries (Table 3). Due to the lack of rigorous methods of detection it is not known how many ncRNA genes exist, how important they are and what functions they play, what the relative amounts of them are, when and how they are expressed, how stable they are, whether they contain modified bases, what their secondary and tertiary structures are like and finally why they substitute proteins (EDDY 2001, ERDMANN et al. 2001, SCHATTNER 2002, STORZ 2002).

Noncoding RNAs range in size from about 20 nucleotides for the large family of microRNA that modulate development in *C. elegans*, *D. melanogaster* and mammals, to 100-200 nucleotides for small RNAs commonly found as translational regulators in bacteria, and finally to over 10 000 nucleotides for RNA involved in gene silencing in higher eukaryotes (HUTVAGNER, ZAMORE 2002, STORZ 2002).

Functions of ncRNAs

It is now clear that all organisms contain a wealth of small untranslated RNAs that function in a variety of cellular processes. The widespread use of RNA molecules as riboregulators may in part be due to their quick and easy production, as no pro-

Table 2. Functional classification of non-protein-coding RNA transcripts (Szymanski, Barciszewski 2002)

Protein-coding transcripts	Noncoding transcripts	
	housekeeping RNAs	regulatory RNAs
mRNA	**tRNA** translation of genetic information **rRNA** ribosome components, catalysis of peptide bond formation **snRNA** pre-mRNA splicing, spliceosome components **snoRNA** RNA modification – 2'-O-methylation and pseudourydilation **RNase P RNA** maturation of 5'-ends of pre-tRNA **telomerase RNA** telomeric DNA synthesis, component of telomerase **4.5S RNA** protein export in bacteria **7SL RNA** protein export in eucaryotes **tmRNA** *trans*-translation **hY RNA** Ro RNP components, function unknown **RNase MRP** RNA processing	**transcriptional regulators** chromatin remodelling structure associated with X-chromosome inactivation and dosage compensation in eukaryotes (roX RNAs, Xist/Tsix transcripts), regulation of expression of imprinted genes (H19, antisense transcripts from imprinted chromosomal regions) **post-transcriptional regulators** antisense RNA:RNA interactions repress or stimulate translation of regulated mRNAs in eukaryotic and prokaryotic cells (DsrA, MicF, lin-4, let-7, microRNAs) **protein function modulators** RNA-protein interactions modulate activity of protein (6S RNA, OxyS, SRA RNA) **RNA distribution regulators** specific subcellular location of RNA influences localization of mRNA or pre-mRNA (hsr-, Xlsirt, BC1, BC200)

tein synthesis is required. RNA molecules may be destroyed, making them well suited for transient modulation of gene expression. In addition to transfer and ribosomal RNAs, many new non-protein-coding transcripts, with diverse functions, have been identified. There is a growing number of untranslated RNAs involved in regulation of gene expression in eukaryotes (Table 2).

These genes encode RNAs that lack open reading frames and function as their final products. Small nontranslated RNAs are engaged in a wide variety of molecular tasks and perform a multitude of functions in the cell, e.g. tRNAs function as adapters in translation, small nuclear RNAs are involved in RNA splicing, and small nucleolar RNAs direct modification of ribosomal RNAs (Mignone et al. 2002, Filipowicz, Pogacic 2002). Alu elements and many other types of repeats, can produce noncoding RNAs, which are potentially capable of taking part in the regulation of gene activity through mechanisms of post-transcriptional gene

Table 3. Examples of non-coding RNAs and their characteristics (kb = kilo bases; n/d = not determined; n/a = not available; nt = nucleotides) (ERDMANN et al. 2001)

Noncoding RNA	Size	EMBL/GenBank Acc. No. or Ref.	Remarks
1	2	3	4
A. DNA markers			
1. Dosage compensation RNAs			
Homo sapiens	16.5 kb	M97168	
Mus musculus	14.7 kb	L04961	
Bos taurus	n/d	AF104906	partial sequence
Equus caballus	n/d	U50911	partial sequence
Oryctolagus cuniculus	n/d	U50910	partial sequence
Drosophila melanogaster roX1	3749 nt	U85980	
Drosophila melanogaster roX2	1293 nt	U85981	
HZ-1 virus PAT-1	937 nt	U03488	
Homo sapiens Tsix	40 kb	Ref. 12	
Mus musculus XistAS	n/d	Ref. 13	
2. H19			
Homo sapiens	2313 nt	M32053	
Mus musculus	1899 nt	X58196	
Rattus rattus	2297 nt	X59864	
Oryctolagus cuniculus	1842 nt	M97348	partial sequence
Pongo pygmaeus	1644 nt	AF190058	partial sequence
Felis catus	1747 nt	AF190057	partial sequence
Lynx lynx	879 nt	AF190056	partial sequence
Ovis aries	397 nt	AF105429	partial sequence
Thomomys monticola	875 nt	AF190055	partial sequence
Elephantidae gen. sp.	856 nt	AF190054	partial sequence
Peromyscus maniculatus	2094 nt	AF214115	
3. IPW			
Homo sapiens	2075 nt	U12897	
Mus musculus	734 nt	U69888	partial sequence
B. Gene regulators			
1. NTT			
Homo sapiens	17 kb	U54776	
2. DGCR5			
Homo sapiens	1284 nt	X91348	
3. KvLQT$_1$–AS			
Homo sapiens	n/d	n/a	
Mus musculus	n/d	AF119385	partial intron sequence
4. Nesp/GNAS			
Homo sapiens	828 nt	AJ251760	partial sequence
Mus musculus	1083 nt	AF173359	
5. SCA8			
Homo sapiens	32.3 kb	AF252279	partial sequence

1	2	3	4
6. CMPD associated RNA			
Homo sapiens	3414	D43770	
7. Developmental timing			
Caenorhabditis brigsae let-7	21 nt	AF210771	
Caenorhabditis elegans lin-4	22 nt / 61 nt	U01830	
8. Other noncoding RNA transcripts			
Homo sapiens UBE3A antisense	n/d	n/a	
Homo sapiens DISC2	15 kb	AF222981	
Homo sapiens ZNF 127 AS	1827 nt	U19107	
Styela clava SCYc RNA	1.1 kb	L42757	
9. Antisense plasmid			
sok RNA	66 nt	AP000342	R100 plasmid
finP RNA	2778 nt	AP000342	R100 plasmid
CopA	186 nt	V00326	R1 plasmid
RNA I-	107 nt	J01566	ColE1 plasmid
Selenomonas ruminantium CtRNA	88 nt	Z49917	pJJM1 plasmid
Escherichia coli Incl	363 nt	M34837	Col Ib-P9 plasmid
Streptococcus pneumoniae RNA II	111 nt	S81045	pLS1 plasmid
Streptococcus agalactiae RNA II		L03355	pIP501 plasmid
Escherichia coli RNA I	73 nt	M28718	pMU 720 plasmid
C. Abiotic stress signals			
1. gadd7/adapt15, adapt33, vseap1			
Cricetulus griseus gadd7	754 nt	L40430	
Cricetulus griseus adapt15	746 nt 753 nt	U26833 U26834	adapt15-P9 adapt15-P8
Cricetulus griseus adapt33	1290 nt 1186 nt	U29660 U29661	adapt33A adapt33B
Cric etulus griseus vseap1	0.9 kb 3.1 kb	AJ003192	
2. hsr-ω			
Drosophila melanogaster	1174 nt 1190		
Drosophila hydei	1129 nt	M14558; J02629	
Drosophila pseudoobscura	1213 nt	X16337; X16157;	
3. G90			
Mus musculuss	1357 nt	AJ132433	
4. OxyS			
Escherichia coli	110 nt	U87390	
5. DsrA			
Escherichia coli	86 nt	U17136	putative
Salmonella typhimurium	82 nt	AF090431	
Klebsiella pneumoniae	82 nt	AF090431	
6. DD3/PCGEM1			
Homo sapiens	3800 nt 1600 nt	AF103907 AF22389	

Table 3 (cont.)

1	2	3	4
D. Biotic stress signals			
1. His-1			
Homo sapiens	nd	U56440	gene sequence, exon structure unknown
Mus musculus	3053 nt 3003	U09772 U10269	alternatively spliced forms of same pre-mRNA
2. ENOD40			
Glycine max	679 nt 617 nt	X69154 X69155	ENOD40-1 ENOD40-2
Pisum sativum	702 nt	X81064	
Phaseolus vulgaris	600 nt	X86441	
Vicia sativa	718 nt	X83683	
Trifolium repens	631 nt	AJ000268	
Lotus japonicus	770 nt	AF013594	
Medicago sativa	626 nt 733 nt	X80263 L32806	
Medicago truncatula	920 nt	X80262	
Nicotiana tabacum	470 nt	X98716	
Vigna radiata	331 nt	AF061818	partial sequence
Sesbania rostrata	638 nt	Y12714	
3. lbiRNA			
Bacteriophage Acm1	97 nt	Z30964	
4. CR20			
Cucumis sativus	1108 nt	D79216	
Arabidopsis thaliana	758 nt	D79218	
5. GUT15			
Arabidopsis thaliana	1377 nt	U84973	
Nicotiana tabacum	1670 nt	U84972	
E. Other functions			
1. Bsr RNA			
Rattus norvegicus	4723 nt 920 nt 2032 nt 1198 nt 1773 nt 2244 nt 1755 nt	AB014883 AB014882 AB014881 AB014880 AB014879 AB014878 AB014877	isolated clones contain various number of ~0.9 kb repeat units
BC1 RNA			
Rattus rattus	152 nt	M16113	
Peromyscus maniculatus	391 nt	U33851	
Peromyscus californicus	359 nt	U33850	
Meriones unguiculatus	350 nt	U33852	
Mus musculuss	152 nt	U01310	
Mesocricetus auratus	142 nt	U01309	
Cavia porcellus	165 nt	U01304	

1	2	3	4
BC200 RNA			
Homo sapiens	200 nt	AF020057, U01306	
Saguinus oedipus	195 nt	AF067788	
Saguinus imperator	194 nt	AF067787	
Aotus trivirgatus	196 nt	AF067786	
Macaca fascicularis	200 nt	AF067785	
Macaca mulatta	200 nt	AF067784	
Chlorocebus aethiops	205 nt	AF067783	
Papio hamadryas	197 nt	AF067782	
Hylobates lar	203 nt	AF067781	
Pongo pygmaeus	198 nt	AF067780	
Gorilla gorilla	204 nt	AF067779	
Pan paniscus	205 nt	AF067778	
2. SRA			
Homo sapiens	875 nt	AF092038	
Mus musculus	829 nt	AF092039	
3. meiRNA			
Schizosaccharomyces pombe	508 nt	D31852	
4. UHG			
Homo sapiens U22HG	1114 nt	U40580	
Mus musculus U22HG	590 nt	U40654	
Homo sapiens U17HG	885 nt	AJ006834	variant A
	2139 nt	AJ006835	variant AB
Mus musculus U17HG	1682 nt	AJ006836	
	383 nt	AJ006837	
Homo sapiens U19HG	681 nt	AJ224167	
	785 nt	AJ224166	
	310 nt	AJ224170	
	375 nt	AJ224169	
	666 nt	AJ224168	
Homo sapiens Gas5	4055 nt	AF141346	
5. Xlsirt RNA			
Xenopus laevis	76 nt	S67412	single repeat sequences
	79 nt	S67413	
	78 nt	S67414	
	80 nt	S67415	

silencing. It probably acts to suppress transposon activity and repress the expression of other repeated genes. It is known that a strong increase in the number of Alu transcripts occurs when cells are placed under stress, as a result of viral infection or inhibition of translation (EDDY 2001). Genes encoding housekeeping RNAs (transfer, ribosomal, small nuclear, nucleolar, vault, telomerase, etc.) have been annotated (Table 2). Rapidly accumulating evidence indicates that ncRNAs can play critical roles in a wide range of cellular processes, from protein secretion to gene regulation. In many cases important developmental decisions such as gene dosage, silencing or genome imprinting, are related to expression of small ncRNAs (ERDMANN et al. 2001). RNA is detected as component of chromatin, but this has traditionally been attributed to the presence of either nascent transcripts or small RNAs involved in splicing or transcript processing. Many studies show that untranslated RNA play a role in maintaining and regulation of chromosome structure, and several recent findings suggest that RNA molecules may have more central roles in silencing of genes and chromatin regulation than previously believed. The chromodomain, a motif common in proteins that have a role in regulation of gene expression, has been shown to interact with ncRNAs *in vitro*. Human *Xist* RNA is required for X-chromosome inactivation and mouse *AIR* RNA is required for autosomal gene imprinting (SLEUTELS el al. 2002). The *Xist* RNA is produced by the inactive X-chromosome and spreads in *cis* along the chromosome. The chromosome-associated RNA has been proposed to recruit proteins that affect chromatin structure and establish or maintain gene silencing (ERDMANN et al. 2000, SZYMANSKI, BARCISZEWSKI 2002).

The variety of ncRNA genes known today is fairly small relative to protein-coding genes, although the number of members within a single RNA gene family is substantial.

Some are naturally occurring antisense RNAs, whereas others have more complex structures (BRANDL 2002). In addition to that, there are small interfering RNAs (siRNA) and small temporary RNA (stRNA), which mediate down regulation of gene expression. siRNA involves RNA interference, for which target mRNAs are degraded but stRNAs inhibit expression of target RNAs after translation initiation without affecting mRNA stability (HUTVAGNER, ZAMORE 2002). Most mammalian imprinted genes occur in clusters that contain noncoding RNAs. Their expression from one parental allele correlates with repression of linked protein-coding genes, which suggests that ncRNAs are involved in the silencing mechanism (Table 2).

There are many pieces of evidence that suggest a functional role for sense-antisense pairings in mammalian gene regulation at a surprising variety of levels (BRANDL 2002). It includes genomics imprinting, RNA interference, translational regulation, alternative splicing, X-chromosome inactivation and RNA editing. Although in a number of cases, where the mode of regulation has been explored in detail, they have proved uniquely intriguing, such that it is difficult to make generalizations.

Endogenous antisense RNAs can be broadly divided into two categories:

(i) antisense RNAs (*trans*-antisense) transcribed from loci distinct from their putative targets are generally short and have the potential to form imperfect duplexes with complementary regions of their sense counterparts,

(ii) antisense transcripts (*cis*-antisense RNAs) that originate from the same genomic region but with opposing orientation have, by virtue of their common but complementary origin, the potential to form long perfect duplexes. Various *cis*-antisense RNAs have been observed in prokaryotes, plants and animals, and their roles are unlikely to be limited to those in imprinting and chromatin structure. Mutations in one *cis*-antisense RNA in humans (SCA 8) are found in patients with spinocerebellar ataxia (SZYMAŃSKI, BARCISZEWSKI 2002).

The mechanisms of action for ncRNAs can be grouped into several types:
– ncRNAs where base-pairing with another RNA (ca 10 base pairs) is central to function, e.g. snoRNAs that direct RNA modification, the bacterial RNAs that modulate translation by forming base pairs with specific target mRNAs, microRNAs involved in silencing;
– ncRNAs resembling structures of other nucleic acids, e.g. 6S RNA is reminiscent of an open bacterial promoter or tmRNA that has features of both tRNAs and mRNAs;
– ncRNAs with catalytic function, e.g. ribonuclease P.

Most ncRNAs are associated with proteins that augment their functions, but some ncRNAs (snRNAs, SRP, telomere RNA, 7SK RNA) serve key structural roles in protein-RNA complexes (SZYMANSKI, BARCISZEWSKI 2002).

Taking into account the versatility of RNA and the fact that the properties of RNA provide advantages over peptides for some mechanisms, it is likely that a number of ncRNAs have evolved more recently.

A search for ncRNA genes in genomes

There is a lack of generalized computational methods for identifying new classes of RNA genes. Most of ncRNAs recognized to date were identified genetically or by accident, although recent data indicate that systematic approach should reveal many more cases. One of such approaches involves expressed RNA sequence tags (similarity to ESTs). However, the diversity of ncRNAs discovered is so great that it is difficult to categorize them except broadly on the basis of their occurrence and function but not on primary or secondary structure (ERDMANN et al. 2001). The first ncRNAs were identified almost 50 years ago on the basis of their high expression, direct labelling and isolation on polyacrylamide gel. Others were identified by fractionation of nuclear extracts or by association with specific proteins. In the "pregenomic era", one of the best approaches to understand the physiological role of an unknown gene product was to examine the phenotype of mutant strains and take clever guesses for its function, which would lead to biochemical experiments. Such process could sometimes take years to accomplish (SCHATTNER 2002).

In the genomic era and beyond, the best starting approach is to use an *in silico* method to compare the deduced sequence of the gene product with those of known function, and with a little bit of luck, one might get to the same place in a few hours. The search for ncRNAs has included large gaps between protein-coding genes, extended stretches of conservation between species with the same gene order, orphan promoter or terminator sequences, presence of G+C-rich regions in organisms with conserved RNA secondary structures with high A+T content (Table 1). As one can see, the amount of noncoding sequences increases from Eubacteria to Eukaryota. In Eubacteria and Archaea the amount of noncoding sequences is similar but in eukaryotes it is much higher. Generally, the level of noncoding sequences in Eubacteria is between 12-27% (average 17.6%) and 13-23% (average 18.5%), while in Archaea seems to be random. There is no correlation with genome size. On the other hand, the amount of noncoding sequences in Eukaryota varies dramatically. One can suggest that its role is to protect DNA against random damage by lowering the possibility of damage to occur in regions important for cell functions (MATTICK 2001).

To better understand the role of G+C content, we compared some viral RNA genomes and transfer-messenger RNA (tmRNA). Interestingly, the G+C content of the tRNA-like part is much higher than those of viral genomes. Also the G+C content of the coding and noncoding part is different. However, there are two interesting observations. For TMV, EMV and STMV, the G+C content is the same within the coding and noncoding part. In addition, the G+C content of the noncoding part of tmRNA is close to that of the tRNA-like part (Table 1D). Bacterial genomes are gene-rich and noncoding DNA represents usually regulatory sequences (promoters) and non-transcribed mRNA portions. In contrast, less than 2% of the human genome encode proteins (Table 1). A question is what information, if any, is contained within the remaining 98% of a genome? How to find ncRNAs in the genome? One can look for the amount of G+C bases (Table 1). Generally there is no direct correlation between the contribution of G+C bases to coding and noncoding RNAs, although the differences between genomes are clearly visible. It varies from 30 to 70% among prokaryotes and is around 40-50% in eukaryotes although, G+C content of noncoding regions is smaller (Table 1).

Coding and noncoding genomic parts can be characterized by the ratio of G+C content of coding and noncoding regions to the G+C content of a whole genome. For Eubacteria and Archaea the ratio in protein-coding regions is above 1, but for the noncoding part ca 0.9 (Table 1). Greater differences are seen for eukaryotic genomes. The value for the coding part is 1-1.3 but for noncoding part it is 0.8-1. The same tendency can be observed for viral RNA genomes (Table 1). The ratio of G+C content of the coding part to the total G+C content of tobacco mosaic virus (TMV), is ca 1.3, almost identical to that the of coding part of *Arabidopsis thaliana*. Generally the G+C patterns of eukaryotic and viral genomes are very similar.

There are some limitations of current methods. Most of the computation approaches have focused on intergenic regions. It has been recently shown that some of the ncRNAs are processed from longer protein or RNA-encoding transcripts. It is also possible that ncRNAs are expressed from the opposite strand of protein-coding genes. If so, expression-based methods might miss ncRNAs that are synthesized under highly specific conditions (developmental signals, environmental signals, cell type). However, when results of these two approaches do not appear to make any sense with each other, new experiments must be done.

Perspectives

The ultimate goal of genome projects from bacterial to human, is not only to sequence their entire genomes in order to identify their complete set of genes, but also to obtain information as to when and where these genes are being expressed and whether their expression is possibly altered during unfavourable circumstances, such as disease, aging or stress. A great challenge is to understand how the genetic information results in the concerted action of gene products in time and space to generate function. The ever-growing realization of the variety of biochemical roles of RNA in all living organisms is leading to an increasing appreciation that cellular RNAs provide inviting targets to treat a variety of diseases.

Table 4. Genome size and number of chromosomes for various organisms

Organism	Genome size	Number of chromosomes
Amoeba dubia	670,000,000,000	Several hundred
Trumpet lily *(Lilium longiflorum*)	90,000,000,000	12
Mouse (*Mus musculus*)	3,454,200,000	20
Human (*Homo sapiens*)	3,200,000,000	23
Carp (*Cyprinus carpio*)	1,700,000,000	49
Chicken (*Gallus gallus*)	1,200,000,000	39
Housefly (*Musca domestica*)	900,000,000	6
Tomato (*Lycopersicon esculentum*)	655,000,000	12
Yeast (*Saccharomyces cerevisiae*)	12,000,000	16
Escherichia coli	4,639,221	1
Agrobacterium tumefaciens	2,841,581	1
Mycoplasma genitalium	580,074	1

In molecular medicine this is reflected in numerous disorders based on polygenic traits and the notion that the number of human diseases exceeds the number of genes in the genome. The availability of the complete human genome sequence has highlighted the need for tools to analyse its contents. It is known already that the total number of human genes which show only a tiny part of the whole genome does not differ substantially from the number of genes of *Arabidopsis thaliana*, although both genomes of 3.4×10^9 bp and 120×10^6 bp in size, respectively, varied strongly (Table 1). At the biochemical level, proteins or ribonucleic acids rarely act alone, but rather they interact with other proteins or RNA to perform a specific cellular job. It suggests that the organism complexity may partly rely on the contextual combination of the protein gene products and noncoding transcripts. These assembles represent more than the sum of their parts by showing new functions. Most biologists and genome researchers concentrate mainly on protein-coding genes, and thus are not aware of the special issues involved in detecting RNA genes.

Various data discussed above indicate that a potentially important class of genes has largely escaped our detection. It seems that there is a large group of functional RNA molecules, which remains hidden between and sometimes within protein-coding regions (introns) and are unaccounted for. Recent discoveries in molecular and cellular biology encouraged structural biologists to analyse new ways in which RNAs can fold, interact with proteins and be catalytically active. Bioinformatics has made a strong entry into RNA research and it seems to be a safe prediction that this discipline will engage into a very close symbiosis with RNA biologists. Today's version of "pure RNA world" is the ribonucleoprotein world (RNP world), whose fertilizing winds blow across the entire RNA landscape from transcription, processing, editing, translation and RNP remodeling (HENTZE et al. 2000).

After years devoted to the isolation of individual genes involved in physiological or developmental processes, biology has entered the world of whole genome analysis (Table 1). One of the great achievements of molecular biology was the sequencing of the human genome, which is huge, but not the largest one. Still there are larger genomes (Table 4). Current genomic approaches rely primarily on technological innovations, such as large-scale DNA sequencing and DNA microarrays, which allow researchers to study all the genes involved in a given process. More important is the conceptual revolution induced by these innovations. Implications of genomics for the understanding and treatment of human diseases is obvious. Genomics will also increase our knowledge of genome organization and evolution, e.g. of gene content, genome organization at both the sequence and cytogenetic levels, promoter usage or alternative splicing. Faced with the avalanche of genomic sequences and data on their expression, scientists are confronting a frightening prospect: piles of information but only flakes of knowledge? How can genomic sequences being determined and deposited, and the thousands of expression profiles being generated by the new arrays methods, be synthesized into useful knowledge. The recent discovery of hundreds of new ncRNAs illustrates that the "RNome" (similar to genome) will need to be characterized before a complete tally of the number of genes encoded by a genome can be achieved. What form will this knowledge take? Can we throw some new light on RNA? These are questions to be addressed in the future. It is clear that the more we learn about RNA, the more is to explore. Still there is much to investigate before genome is over and there is much beyond genome.

Now it is clear that the human DNA sequence is not enough for complete interpretation of the entire human genome.

Acknowledgements. This work was supported within KBN research grants 6PO4A 063 21 and 6PO4C 105 20.

REFERENCES

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G. et al. (2000). The genome sequence of *Drosophila melanogaster*. Science 287: 2185-2195.

Brandl S. (2002). Antisense-RNA regulation and RNA interference. Biochim. Biophys. Acta 1575: 15-25.

Eddy S.R. (2001). Non-coding RNA genes and the modern RNA world. Nature Rev. 2: 919-929.

Erdmann V.A., Szymański M., Hochberg A., De Groot N., Barciszewski J. (2000). Non-coding, mRNA-like RNAs database Y2K. Nucleic Acids Research 28: 197-200.

Erdmann V.A., Barciszewska M.Z., Hochberg A., De Groot N., Barciszewski J. (2001). Regulatory RNAs. Cell. Mol. Life Sci. 58: 1-18.

Filipowicz W., Pogacic V. (2002). Biogenesis of small nucleolar ribonucleoproteins. Curr. Opin. Cell Biol. 14: 319-327.

Hentze M.W., Izaurralde E., Seraphin B. (2000). A new era for the RNA world. EMBO Rep. 1: 394-398.

Hutvagner G., Zamore P.D. (2002). RNAi: nature abhors a double-strand. Curr. Opin. Gen. Develop. 12: 225-232.

Katinka M.D., Duprat S., Cornillot E., Metenier G., Thomarat F., Prensier G. et al. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414: 450-453.

Mattick J.S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2: 986-991.

Mignone F., Gissi C., Liuni S., Pesole G. (2002). Untranslated regions of mRNAs. Gen. Biol. 3: 0004.1-0004.10.

Moore P.B., Steitz T.A. (2002). The involvement of RNA in ribosome function. Nature 418: 229-235.

Natale D.A., Shankavaram U.T., Galperin M.Y., Wolf Y.I., Aravind L., Koonin E.V. (2000). Towards understanding the first genome sequence of a crenarcheon by genome annotation using clusters of orthologous groups of proteins (COGs). Gen. Biol. 1: 0009.1-0009.19.

Schattner P. (2002). Searching for RNA genes using base-composition statistics. Nucleic Acids Res. 30: 2076-2082.

Sleutels F., Zwart R., Barlow D.P. (2002). The non-coding *Air* RNA is required for silencing autosomal imprinted genes. Nature 415: 810-813.

Storz G. (2002). An expanding universe of noncoding RNAs. Science 296: 1260-1263.

Szymański M., Barciszewski J. (2002). Beyond the proteome: non-coding regulatory RNAs. Gen. Biol. 3(5): 0005.1-0005.8.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G. et al. (2001). The sequence of the human genome. Science 291: 1304-1351.

Wood V., Gwillian R., Rajandream M.A., Lyne R., Stewart A. Sgouros J. et al. (2002). The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871-880.