The distribution of RNA motifs in natural sequences

Véronique Bourdeau, Gerardo Ferbeyre, Marie Pageau, Bruno Paquin* and Robert Cedergren

Département de Biochimie, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, QC H3C 3J7, Canada

Received July 16, 1999; Revised and Accepted September 28, 1999

ABSTRACT

Functional analysis of genome sequences has largely ignored RNA genes and their structures. We introduce here the notion of 'ribonomics' to describe the search for the distribution of and eventually the determination of the physiological roles of these RNA structures found in the sequence databases. The utility of this approach is illustrated here by the identification in the GenBank database of RNA motifs having known binding or chemical activity. The frequency of these motifs indicates that most have originated from evolutionary drift and are selectively neutral. On the other hand, their distribution among species and their location within genes suggest that the destiny of these motifs may be more elaborate. For example, the hammerhead motif has a skewed organismal presence, is phylogenetically stable and recent work on a schistosome version confirms its *in vivo* biological activity. The under-representation of the valine-binding motif and the Rev-binding element in GenBank hints at a detrimental effect on cell growth or viability. Data on the presence and the location of these motifs may provide critical guidance in the design of experiments directed towards the understanding and the manipulation of RNA complexes and activities *in vivo*.

INTRODUCTION

The realization that conserved amino acid motifs in proteins can often be related to function has greatly aided the evaluation of unidentified open reading frames in sequence databases. Detailed comparison of protein sequence, structure and function has now provided motif databases (1–4), which greatly facilitates the task of inferring function for otherwise uncharacterized coding sequences. As sequences have accumulated, so has the number of recognizable motifs, thereby guaranteeing that an ever-increasing role will be played by functional inference or *in silico* analysis of sequence motifs.

RNA remains an enigma in gene sequence research since few systematic attempts have been made to identify RNA coding genes in sequence databases other than those belonging to a few well-known families, such as transfer RNAs (tRNAs) or small nucleolar RNAs (snoRNAs; in particular see ref. 5), as well as other RNAs like Group I and II introns. The degeneracy built into RNA molecules due to their composition being based on only four major nucleotides renders the primary sequence of RNA insufficient, by itself, in defining motifs. Secondary and tertiary structural aspects must therefore be made part of RNA motif definitions. In spite of these complications, evidence is accumulating that RNA motifs will provide the ultimate basis for an understanding of RNA structure and function (6,7).

To introduce our concept of RNA genomics, 'ribonomics', we define and present here the distribution of a number of RNA motifs in the GenBank sequence database (8). Given the lack of consensus on what an RNA motif might be (6,9–11), the use of motif in this work will refer to RNA molecules or parts thereof which have a chemical or ligand-binding activity in a defined context. Our motifs could be called 'functional motifs' because of their demonstrated biological activity under known conditions, but prudence dictates caution: it is unlikely that these motifs would behave similarly in all environments due to factors of conformational variability, accessibility, etc. In light of these complications, we will arbitrarily refer to the motifs under investigation here as 'biological motifs' in order to acknowledge their known activities under certain conditions and to underline their potential to play a biological role in other contexts.

Our search of GenBank employed RNAMOT, a computer search engine developed in our laboratory, which defines primary and secondary structural information within computer readable 'descriptors' (9,10). Previously, features of this strategy have been illustrated in searches to identify tRNAs (9), alternative folding patterns of cytoplasmic tRNAs (12), putative Tat-binding elements (TBEs) in viruses linked to human immunodeficiency infections (13) and a catalytic RNA domain in the repetitive DNA of schistosome (14) and of cricket (Rojas *et al.*, manuscript in preparation). Here, we present extensive searches of the GenBank database for RNA biological motifs implicated in chemical or ligand-binding activities.

MATERIAL AND METHODS

The search

The program RNAMOT (9,10), written in C, requires a sequence file in the IUPAC/IUB format and a 'descriptor' defining the motif under investigation. In the course of this study, we have used the release of October 15, 1998 of the

*To whom correspondence should be addressed. Tel: +1 514 343 6111 ext. 1938; Fax: +1 514 343 2210; Email: paquinb@magellan.umontreal.ca
Present address:
Gerardo Ferbeyre, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY 11724, USA

Dedicated to the late Robert Cedergren

sequence data bank: GenBank (NCBI-GenBank flat file release 109.0). Searches were carried out on both strands and all occurrences of motifs involving unidentified bases denoted by N in the database were disregarded. A Power Challenge XL with 32 CPUs IP 19, R4400, 150 MHz processor (3072 Mb) running UNIX IRIX 6.2 was used.

In order to help establish the significance of their presence, frequencies of each motif in the database were compared with frequencies in a random sequence database generated by a uniform pseudo-random number generator (15) with a period length near 2^{121}. The random sequence databases contained 10 000 sequences of 100 000 nucleotides each; the four nucleotides A, C, G and T were used with equal probabilities. An 'expected' frequency in GenBank (**N**) was calculated from the number of occurrences of each motif in the random databases (**M**) by the following: $\mathbf{N} = (\mathbf{a} \times \mathbf{M})/(10^4 \times 10^5)$, where **a** is the number of nucleotides in GenBank (2.009×10^9 in release 109.0).

The analysis

Subsequent to the compilation of motif frequency in sequence fragments, the location within the fragment was identified and extracted from the associated documentations by SITE, a suite of programs written in C with Perl, AWK and Bourne Shell scripts. SITE determined the strand sense, and whether the motif was contained partially or wholly within features defined in the documentation of the sequence fragment. From the overall list of some 67 features in GenBank, we have combined and compiled the following subset of features for our classification: 1) mRNAs: sequenced mRNAs or cDNAs, CDS (coding sequences), mat_peptides (maturation peptides) and UTR (untranslated regions); 2) introns; 3) control regions: the CAAT_signals, TATA_signal, enhancers, promoters, –10 and –35_signals; 4) LTR (long terminal repeat); 5) rRNAs (ribosomal RNAs); 6) tRNAs; 7) other RNAs: including pre_RNAs (precursor RNAs), prim_transcript (primary transcript RNAs), guide RNAs, scRNA (small cytoplasmic RNAs) and snRNAs (small nuclear RNAs); 8) satellite & repeat: sequences with satellite DNA features or repeated sequence entries; 9) artificial: patent, synthetic, artificial and oligonucleotide entries not described with features; and finally 10) a miscellaneous category including all other minor features or no description at the location of the result. The occurrence and location files were then converted into the HTML format and linked to GenBank files. The compilation of the results in some cases involved correcting the location assignment of individual sequence entries especially when an mRNA feature is identified, but the motif falls outside identified CDSs, indicating that it is an intron or a UTR.

A possible phylogenetic distribution was also evaluated through the compilation of repeated gene names. For this analysis, the GenBank files EST (expressed sequence tags), GSS (genome survey sequences), HTG (high throughput genome sequencing), PAT (patent sequences), STS (sequence tagged sites), SYN (synthetic and chimeric sequences) and UNA (unannotated sequences) were not considered.

RESULTS

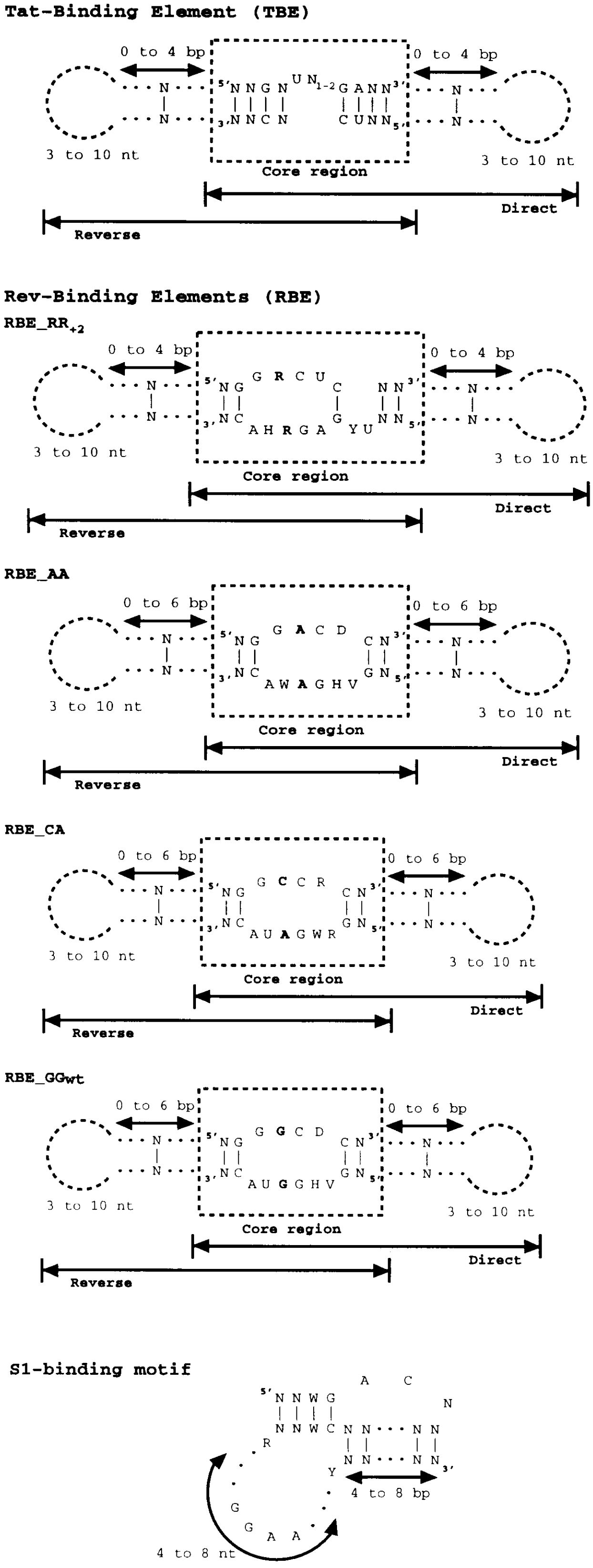
The lack of a bona fide list of accepted RNA motifs prompted us to make a rather arbitrary selection and definition of motifs. We chose examples of defined individual structures rather than structures derived from the known families of cellular RNAs. Thus with the exception of the known natural occurrences of the hammerhead motif (16) or of the UV-loop motif (17) little or no prior information was available indicating whether these structures would be found in the database. Whenever it was possible, we composed two descriptors which maintained the central core region of the motifs but gave two possible positions for the 5′ and 3′ ends (see legend of Fig. 1).

Protein-binding motifs

The Tat-binding element (TBE). The interaction of the Tat protein with a specific RNA-binding site, the trans-activation response region (TAR), is involved in the replication of primate lentiviruses like HIV (human immunodeficiency virus) and SIV (simian immunodeficiency virus) (18,19). Basal transcription from the HIV LTR allows the synthesis of short RNA transcripts, but in the presence of the Tat protein, transcription is enhanced and RNA transcripts are longer (20). The search descriptor for the TBE was defined by the minimal consensus sequence of the TAR element found in different HIV isolates and chemical interference analysis of the binding site (21,22; Fig. 1). In the TBE consensus structure, the two base pairs in the upper and lower stem and the one uridine (U) at the 5′ terminus of the internal loop are invariant, since they provide key interactions in the conformation of the TBE when bound to Tat (23).

The RNAMOT search of the GenBank database using the TBE descriptor produced a list of 52 698 occurrences of the motif in its two possible orientations: 26 102 occurrences for the direct orientation and 26 596 for the reverse one (Table 1). After removal of the motifs from the patent and the synthetic or chimeric sequences, 25 518 occurrences of the direct motif and 24 976 of the reverse one are present in the 'natural' GenBank (Table 1). An identical search with the random sequence bank produced 32 320 presences for the direct orientation and 32 887 for the reverse. The frequency of the motif in GenBank is thus on the order of that expected in random sequences or slightly lower. Next, we established the organismal distribution of the motif. From Table 2, it can be seen that the motif is distributed among the organismal classes in roughly the proportion that each class is represented in GenBank (compare 'TBE' versus 'GenBank distribution' columns) and not only in viral sequences where the natural motif has been identified (see also Fig. 2). The distribution of the TBE motif among genetic features was determined and is shown in Table 3. Particularly interesting is the high number of occurrences of the direct motif in LTR features compared with the reverse motif (on the plus strand) and inversely its low representation in the mRNAs of the minus strand whereas the reverse motif has a quite high incidence.

The Rev-binding element (RBE). The Rev protein and its binding site, the Rev responsive element (RRE), promote the transport of unspliced transcripts of the HIV genome to the cytoplasm (19,24,25). The primary interaction between Rev and the RRE has been shown to be largely determined by a small, 30 nt region of the RRE, called the Rev-binding element (RBE; 26). Descriptors for the RBE motif were derived from the sequences of RNAs possessing high binding affinity to the Rev protein as isolated by selection from partially randomized



sequence pools *in vitro* (see SELEX below; 27). In Figure 1, consensus structures for the four different classes used in our searches are shown. The classes are named according to the non-Watson–Crick interaction bridging two nucleotides in the internal loop (Fig. 1, in bold). Note that the RR_{+2} class contains a bulge of 2 nt not present in the other classes and that the GGwt class includes the wild-type RBE motif.

The number of occurrences for all the RBE classes in GenBank is: 87 for CA class, which represents 0.31 times what we expected; 122 for RR_{+2}, thus 0.22 times the amount expected; 1059 for AA class with 0.49 times the expected frequency; and 1068 for GGwt class, which is around the expected value (1.01 times; Table 1). The expected number and the observed number of occurrences in GenBank are both quite low except for the GGwt class. The AA class seems to produce twice as many occurrences in the reverse orientation of the motif versus the direct one. From the distribution of the hits in the GenBank files (Table 2; Fig. 2), it is obvious that the frequency of the occurrences of the GGwt class is biased by a huge representation of the motif in the viral sequences because of numerous HIV sequence entries. In fact, if we remove the number of occurrences due to HIV/SIV sequences, the frequency obtained dropped slightly lower than the expected level (Table 1). The distribution among features found in Table 3 shows that the RBE has a relatively high frequency in mRNA and coding sequences.

The S1-binding motif. This RNA motif contains a pseudoknot with highly conserved sequence elements in its loops (Fig. 1). The motif binds both the S1 ribosomal protein and the 30S ribosomal subunit from *Escherichia coli* (28). Such an RNA motif on the 5′ UTR of an mRNA might have a regulatory role in translation initiation.

A descriptor of this S1-binding motif was used to search GenBank. Of 135 identified occurrences only two were in the patent or synthetic sequences (Table 1). The remaining 133 represent a frequency slightly higher than what we were expecting (1.38 times the expectation). Surprisingly, the distribution of this motif in GenBank shows a presence higher than expected for a random distribution in the mammalian sequences (especially in primate, Pri, and other mammalian, Mam, files) and the EST, whereas in the bacterial sequences we obtained only half of the expected number (Table 2, Fig. 2). This could mean that the motif has been restricted in bacteria.

Chemically and catalytically active motifs

The UV-loop motif. The photoreactive UV-loop motif was adapted from a consensus of similar RNA loop structures

Figure 1. The protein-binding motifs. (Top) Secondary structure of the TBE. (Middle four) Secondary structure of the four Rev-binding elements: RBE_RR_{+2}, RBE_AA, RBE_CA and RBE_GGwt, which were named according to the postulated interaction of the bold nucleotides (59). Note that RBE_GGwt includes the wild-type motif found in HIV and SIV. (Bottom) Structure of the S1-binding motif. Whenever possible, the RNA motifs were given two orientations (direct and reverse) by maintaining the core region and by varying the position of the stem–loop that completes the motif. The letter code represents the following nucleotides: B = C, G or U; D = A, G or U; H = A, C or U; K = G or U; M = A or C; N = A, C, G or U; R = A or G; S = C or G; V = A, C or G; W = A or U; Y = C or U (IUPAC-IUB code). Note that N-N constraints imply Watson–Crick or G-U pairing. T and U are considered the same.

Table 1. Frequency of RNA motifs in GenBank

	Frequency in GenBank		Frequency in ‘natural’	Expected frequency	Ratio frequency over
	Total	Direct or reverse	GenBank		expected
FAD	10	–	10	0	–
Theophylline-d	12	3	0	4	0
Theophylline-r		9	3	2	1.50
Valine-d	25	9	9	55	0.16
Valine-r		16	15	43	0.35
DNAzyme_8-17	54	–	54	110	0.49
RBE_CA-d	87	35	35	155	0.23
RBE_CA-r		52	51	136	0.38
RBE_RR$_{+2}$-d	122	16	10	47	0.21
RBE_RR$_{+2}$-r		106	10	43	0.23
S1	135	–	133	96	1.38
Neomycin	402	–	391	915	0.43
FMN-d	469	209	203	177	1.15
FMN-r		260	255	193	1.32
RBE_AA-d	1059	369	357	1019	0.35
RBE_AA-r		690	641	1025	0.63
RBE_GGwt-d	1068	860 *(361)	792 *(356)	486	1.63 *(0.73)
RBE_GGwt-r		208	202	533	0.38
Hammerhead	2788	–	414	515	0.80
Leadzyme-d	2808	1517	1487	1806	0.82
Leadzyme-r		1291	1231	1804	0.68
UV-loop-d	2956	1565	1543	718	2.15
UV-loop-r		1391	1371	734	1.87
ATP-d	7693	3438	3319	1918	1.73
ATP-r		4255	4207	2085	2.02
tRNA	5841	–	5664	0	–
TBE-d	52 698	26 102	25 518	32 320	0.79
TBE-r		26 596	24 976	32 887	0.76
Paromomycin-d	831 965	418 554	407 954	466 255	0.87
Paromomycin-r		413 411	404 188	462 539	0.87

The number of occurrences found in the GenBank for both orientations of the RNA motifs is shown together (total column) or separately (direct or reverse column). The ‘natural’ GenBank refers to the database after removal of entries that are patented sequences (PAT) and synthetic or chimeric sequences (SYN). These values are compared to the results obtained in a random sequence database as indicated in ‘expected frequencies’ (see Material and Methods). The ratio is the number of occurrences found in the ‘natural’ GenBank over what was expected. *, values excluding HIV and SIV sequences for RBE_GGwt-d.

found in viroids, 5S rRNA, the sarcin-ricin loop of 28S rRNA and the hairpin ribozyme (17). This internal loop includes a G and a U (Fig. 3, bold) which are covalently cross-linked upon UV radiation (29).

The two descriptors used represent the two orientations (5′ or 3′) of the motif (Fig. 3). There are 2914 occurrences found in the ‘natural’ GenBank (1452 occurrences expected, Table 1). The distribution of this motif shows an over-representation in the invertebrates (Table 2, Fig. 2) as well as in rRNA genes (reverse motif; Table 3).

The hammerhead motif. The hammerhead ribozyme motif was initially defined as a self-cleaving domain found in plant virusoids and satellite RNAs (reviewed in 16). This motif is composed of

Table 2. The organismal distribution of RNA motifs in GenBank

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Rna	Est	Misc.
S1	21	2	5	2	5	7	3	1	0	0	41	14
Neomycin	11	5	1	8	10	5	15	6	0	0	22	18
FMN-d	17	3	<1	2	7	10	4	1	0	0	31	24
FMN-r	9	9	1	9	8	7	7	3	0	0	28	17
DNAzyme_8-17	11	0	<1	2	4	11	24	6	0	0	33	9
RBE_AA-d	18	6	6	1	8	7	8	3	0	0	22	20
RBE_AA-r	16	2	1	<1	5	4	6	1	<1	0	47	17
RBE_GG-d	5	2	<1	1	1	1	4	63	0	0	17	5
RBE_GG-r	17	6	3	1	4	2	15	6	0	0	29	15
Hammerhead	3	1	0	2	20	9	9	26	<1	0	21	7
Leadzyme-d	17	5	2	2	3	3	10	2	<1	0	43	13
Leadzyme-r	14	5	2	1	6	4	11	3	<1	0	42	12
UV-loop-d	14	2	1	1	15	7	5	3	0	<1	28	24
UV-loop-r	14	2	1	1	17	11	5	2	<1	1	21	25
ATP-d	25	5	2	1	2	3	4	2	0	<1	37	20
ATP-r	26	4	1	1	3	2	4	2	<1	0	35	20
tRNA	2	1	<1	1	15	30	38	0	1	5	1	5
TBE-d	15	3	1	1	7	8	7	6	<1	<1	34	19
TBE-r	13	4	1	1	7	8	7	3	<1	<1	37	19
Paromomycin-d	15	3	1	1	8	8	6	2	<1	<1	35	20
Paromomycin-r	15	3	1	1	8	7	6	3	<1	<1	34	21
GenBank distribution	15	3	1	1	7	7	6	3	<1	<1	36	21

Percentage of occurrences in the different groups of organisms or sub-parts of the ‘natural’ GenBank for the motifs having a total number of occurrences over 100 in Table 1. The GenBank distribution is the relative size of each GenBank group relative to the total size of the ‘natural’ GenBank. Pri, primate sequences; Rod, rodent sequences; Mam, other mammalian sequences; Vrt, other vertebrate sequences; Inv, invertebrate sequences; Pln, plant sequences (including fungi and algae); Bct, bacterial sequences; Vrl, viral sequences; Phg, phage sequences; Rna, structural RNA sequences; Est, expressed sequence tag sequences; Misc., genome survey, high throughput genomic sequencing, sequence tagged site and unannotated sequences.

three helices surrounding a single-stranded, catalytic core region. Extensive mutagenic analysis has defined the sequence requirements for efficient self-cleavage: changes in the unpaired core region are not tolerated, whereas few sequence restrictions constrain the base paired regions (Fig. 3).

The descriptor used in the GenBank search for the hammerhead motif did not include the base-pairing requirements derived from helices I and III of the consensus hammerhead RNA motif in order that only the catalytic portion of the motif would be found (bold region in Fig. 3). This definition makes it possible to find the substrate portion of the motif at a distant site consistent with a *trans*-cleavage mode *in vivo*, where the cleavage site and the catalytic core could be in different molecules (14). The hammerhead catalytic motif occurs 2788 times in all the GenBank but 85% of these occurrences correspond to artificial sequences, leaving 414 occurrences in the ‘natural’ GenBank (Table 1), compared to 515 occurrences expected, a ratio of 0.80. The organismal distribution shown in Table 2 and Figure 2 suggests concentrations of hammerhead motifs in both invertebrate and viral sequences, whereas primates seem to be inhospitable to the motif. The motif is also under-represented in the ESTs. The data in Table 3 eloquently support and extend the apparent preference of this motif for repetitive DNA of eukaryotes (14).

The leadzyme motif. The leadzyme is a catalytic RNA having the unusual property of being able to cleave a target RNA in the presence of lead, whereas the classical catalytic RNAs require magnesium, manganese or calcium divalent cations (30,31). The original leadzyme was isolated from *in vitro* experiments where partially randomized RNA molecules derived from a tRNA structure were selected for their ability to self-cleave in the presence of lead ion; there are thus no known naturally occurring leadzymes which cleave *in vivo*. [Note that many RNA molecules do show site-specific cleavages in the presence of divalent lead ion *in vitro* (32).] The consensus structure for a catalytically active leadzyme has been determined by extensive chemical and enzymatic characterization (31) and is shown in Figure 3.

Table 1 shows that the frequency of the leadzyme motif is only slightly lower than that expected based on the constraints used in the descriptors (1487/1806 for the direct orientation and 1231/1804 for the reverse one). The motif shows a slight overabundance among the mammals and bacteria (Table 2, Fig. 2), but its presence among the features may be more significant, since an over-representation among mRNA sequences is evidence that the motif is being expressed in RNA and may therefore be involved in some cleavage activity (Table 3).

In contrast to the search for the hammerhead motif above, the entire leadzyme motif, its catalytic and substrate portions, were combined in the descriptor used in this case. If active, leadzyme occurrences in coding sequences would likely be involved in self-cleavage (*cis*-cleavage) of the cited sequence *in vivo*, although transcription of the leadzyme could lead to

Table 3. Distribution of RNA motifs among GenBank features

	Plus Strand										Minus Strand					
	mRNA	introns	Control region	LTR	rRNA	tRNA	others RNAs	Satellite & repeat	artificial	Misc.	mRNA	introns	Control region	others RNAs	artificial	Misc.
FAD	3									4						3
Theophylline-d									3							
Theophylline-r	1								6							2
Valine-d	2									2	1					4
Valine-r	3								1	6						6
DNAzyme_8-17	16			1						10	1					26
RBE_CA-d	4	1								8	2					20
RBE_CA-r	16									9						27
RBE_RR$_{+2}$-d	5								6	2	1					2
RBE_RR$_{+2}$-r	3								96	2						5
S1	38								2	29	4	2				60
Neomycin B	89	8	2		5				7	76	14				5	196
FMN-d	62	2							4	45	4	4			2	86
FMN-r	74	4							27	49	7	1			1	97
RBE_AA-d	126	3						1	3	76	12	2			6	140
RBE_AA-r	186	5	2		2				43	107	16					329
RBE_GG-d	512	7			2				52	132	5				3	147
RBE_GG-r	61	2			2				2	54	3	2			1	81
Hammerhead	115	4			3			30	2344	88	16	1			19	168
Leadzyme-d	621	9	1		4		1	4	10	212	33	6			14	602
Leadzyme-r	514	17	1		4	1	2	1	17	166	24	2			33	509
UV-loop-d	292	32	1	2	4		1	4	3	441	17	15			2	751
UV-loop-r	252	27	2		35		4	3	6	359	31	11		1	7	653
ATP-d	1155	57	5		8		3	5	18	874	48	21		4	16	2041
ATP-r	942	40	2		6			17	39	665	28	13		1	51	1634
tRNA	31	5	13		17	3431	49		128	548	4	4	1	1082	9	519
TBE-d	6484	303	82	503	118	24	50	33	226	5406	382	124		25	160	12 182
TBE-r	6198	294	44	10	133	6	99	41	202	5352	1072	111		12	285	12 737

Occurrences of the RNA motifs in different gene regions as annotated in the feature section of the GenBank report of each entry (see Material and Methods). Plus strand, sequence submitted to GenBank; Minus strand, complementary sequence from the sequence submission; Misc., not described. Note that on the minus strand the ‘other RNAs’ column includes rRNA and tRNA.

trans-cleavage as well. The fact that the normal intracellular concentration of lead ion would be below the 10–100 µM required for leadzyme activity *in vitro* does not augur well for *in vivo* activity. These identified leadzyme motifs, however, might play a role in lead poisoning.

The RNA-cleaving DNA enzyme motif. Santoro and Joyce (33) isolated, by *in vitro* selection, DNA molecules capable of recognizing RNAs by Watson–Crick base pairing and cleaving them. One of them, DNAzyme_8-17, when paired to its target has the consensus structure shown in Figure 3. We decided to use this motif in our search although it is not an RNA motif because single-stranded DNA shares many characteristics with single-stranded RNA (34–36), even though the availability of such a single-stranded DNA motif is unwarranted. Fifty-four occurrences of the DNAzyme_8-17 can be found in the ‘natural’ GenBank, which represents only 49% of the expected rate (Table 1). Compared to other ‘catalytic’ RNA motifs, this shows the lowest frequency of occurrences.

Small molecule-binding RNA motifs

Aptamer motifs. Aptamers represent a class of RNA molecules that have been isolated and characterized *in vitro* by a technique called SELEX (37). In the first step, a partially or fully randomized pool of RNA molecules is challenged by a potential ligand (37–39). Those RNA molecules bound to the ligand are amplified by PCR after being reverse transcribed into DNA. Cycles of binding and amplification follow until the selected mixture is judged appropriate. Individual molecules are then isolated, sequenced and consensus structures proposed. In the following database searches, we have constructed descriptors based on the consensus structures of the RNA molecules reported in the original publications.

The neomycin-binding motif B. Neomycin and other aminoglycosides bind to 16S rRNA in the A site causing misincorporation of amino acids during protein synthesis (40). An oligoribonucleotide, motif A, mimicking the decoding region of 16S rRNA was partially randomized and used by Famulok and Hüttenhoffer (41) for *in vitro* selection. They identified a new group of neomycin-binding RNAs whose consensus structure was named motif B (Fig. 4). This motif was used to screen the GenBank for neomycin-binding sites and a total of 391 occurrences were found among natural sequences, less than half of the 915 expected (Table 1). A particularly high representation of this motif can be noted among the bacterial and vertebrate sequences (Table 2, Fig. 2); however, the expectation that these surplus occurrences might be in rRNA genes finds little support in the present data (Table 3).

The paromomycin-binding motif. Paromomycin is another aminoglycoside antibiotic binding to rRNA at the ribosome A site. Recht *et al.* (42) developed a consensus structure of this motif based on the analysis of the critical nucleotides essential for paromomycin (Fig. 4). A total of 831 965 occurrences was found. This number should serve as a warning: the utility of finding a given RNA motif in the database is strictly dependent on the careful definition of the motif descriptor. The paromomycin

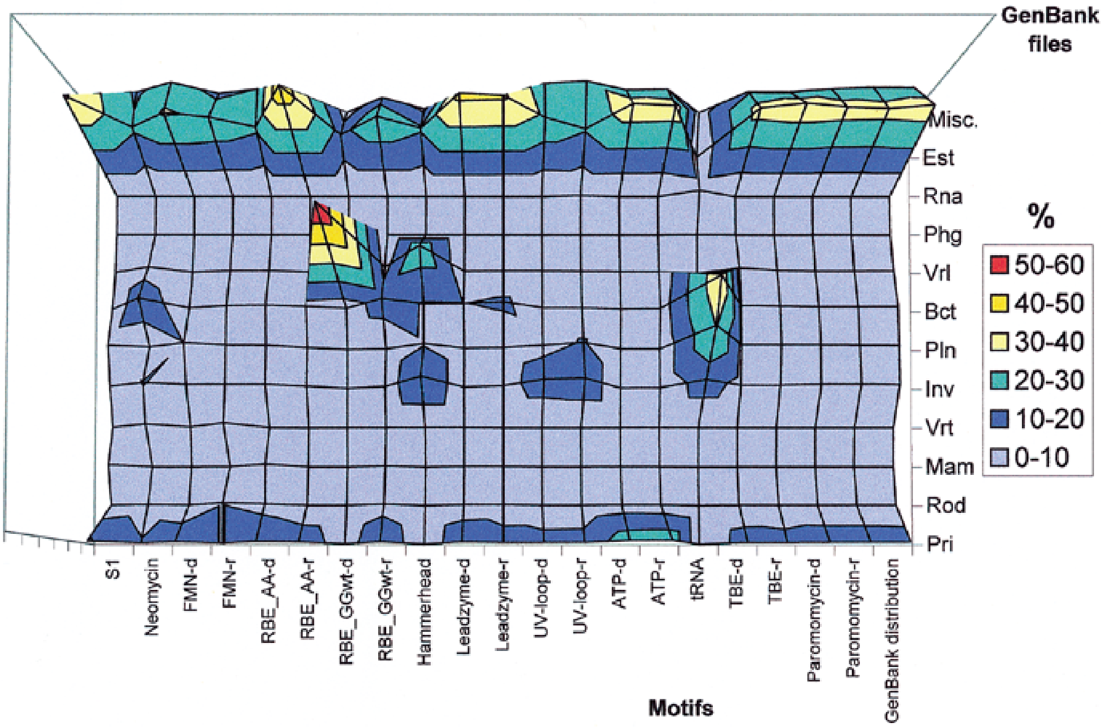


Figure 2. The organismal distribution of RNA motifs in GenBank. Graphic representation of the percentages of occurrences of the RNA motifs and the GenBank distribution as shown in Table 2.

motif in this case is insufficiently constrained, and the search produces an unmanageable result list. Estimation of the frequency (i.e. in a small database) could avoid a useless search.

The valine-binding motif. The valine-binding motif has been exclusively defined by the SELEX technique, and thus there is no evidence that this RNA motif plays any biological role *in vivo* (43). On the other hand, the motif is highly constrained as shown in Figure 4. This motif proves to be very under-represented in GenBank; we found only 24 'natural' occurrences when 98 are expected from the random database (Table 1). These few occurrences are concentrated in the mRNAs (Table 3). Perhaps the clustering of sites in mRNA and the avoidance of this motif signal a functional role for valine.

The theophylline-binding motif. RNA ligands have been isolated by Jenison *et al.* (44) following a SELEX for theophylline binding and a counter-SELEX against binding to caffeine. The consensus secondary structure is shown in Figure 4. We use two descriptors that correspond to the two orientations of the theophylline-binding motif found by selection. Putative theophylline-binding motifs were found only 12 times in GenBank. Most of these are in patent and synthetic sequences, leaving only three occurrences of the reverse motif (Table 1). Since the motif contained a lot of constraints, this result is close to what was expected. Because of the low level of incidence, it is not significant to look at the distribution.

The FMN- and the FAD-binding motifs. Burgstaller and Famulok (45) performed a SELEX to isolate RNA ligands to flavin adenine mononucleotide (FMN) and flavin adenine dinucleotide (FAD). The consensus motifs are presented in Figure 4. The occurrence of FMN-binding motifs of the direct orientation was comparable to the random distribution (203/177) whereas that of the reverse orientation was slightly higher (255/193) (Table 1). It was found at a frequency nine times higher than expected in vertebrate sequences (Table 2, Fig. 2) and the FAD-binding motif was found more frequently than expected (10/0) (Table 1).

The ATP-binding motif. ATP-binding RNAs have been isolated by *in vitro* selection (46). The RNA aptamer consensus obtained recognizes the adenine part of ATP (Fig. 4). The same consensus structure was obtained independently in another experiment that selected RNAs for binding to NAD (45). With two descriptors of the ATP-binding motif, one for each orientation of the motif, we found a total of 7526 'natural' occurrences (Table 1). This is nearly twice that expected in a random situation. Table 2 clearly highlights a high distribution of occurrences in mammalian sequences (primates, rodents and other mammals; see also Fig. 2). A low frequency is observable in invertebrate, plant and bacterial sequences. The ATP-binding motif seems to have no location limitation through the genomes (Table 3).

The tRNA motif. We used a general motif for tRNA (Fig. 5) as a positive control for the search. Our scan found 5664 occurrences, which is far in excess over the expected number (none were obtained in the random sequence database we used). Table 2 and Figure 2 show a significantly increased presence of the motif in structural RNA sequences (RNA), invertebrate, plant and bacterial sequences which could come from a bias of known tRNA sequences for these organisms in GenBank. ESTs contain an extremely low frequency of tRNA motifs, as expected. Table 3 ascertains that most tRNAs are in the tRNA feature.

Phylogenetic analysis

We also looked for a possible phylogenetic distribution of these motifs. We were thus searching for multiple occurrences of a particular motif within homologous genes of different species. We found many such examples, but in most cases the distribution of the motifs between closely related species did not match their relationships as inferred from molecular phylogeny. In fact, the evolutionary pattern followed that of the encoded proteins. Two exceptions, however, are the occurrence of the ATP-binding motif in an intron of the adenine phosphoribosyl transferase (APRT) gene of several species of rodents (see below) and the presence of hammerhead motifs in satellite DNAs of several eukaryotes (14,47; Rojas *et al.*, manuscript in preparation).

DISCUSSION

We have reported the results obtained in the building of an RNA motif database. RNA structures have been defined, converted to computer-usable descriptors and used to search the GenBank database. The principal issues rising from these data are the origin of the motifs, the significance of the wide distribution of the RNA motifs in the database and the usefulness of this information.

Origin of the motifs

Among the origins for the RNA motifs that can be envisaged are random evolutionary drift of sequences, descent from an ancestral organism (phylogenetic origin) and horizontal transfer between organisms. In the first case, one would expect frequencies of occurrences to reflect the probability of a random formation of motifs and a uniform distribution of motifs throughout natural sequences. Frequencies of motifs in a given organism would be proportional to genome size and the restrictiveness of constraints used in their definition. Motifs would be found indiscriminately among different organismal sources and in transcribed or non-transcribed regions. The distribution of the motifs studied here corresponds very well with these criteria. Moreover, our results confirm the conclusion of Schuster *et al.* (48) that sequences able to fold in the same secondary structure can be found randomly in a space of artificial sequences. They are also consistent with reports showing that a subset of evolved RNAs have a similar distribution of shape elements, as do natural RNAs (49,50). Based on these observations, we favor the random drift origin of the vast majority of RNA motifs in the present version of GenBank. Even if the frequencies of some motifs (RBEs, DNAzyme_8-17, valine-binding motif, UV-loop and the ATP motif) vary significantly from the expected (Table 1), this could reflect an origin by random drift, with strong negative or positive selectivity.

Evolutionary dynamics of the motifs

An auxiliary issue to origin is evolutionary flux: can a motif be ‘fixed’ in a population or is it transitory? Evaluation of this property requires detailed phylogenetic analysis, which is not possible in the current organismally sparse database. The

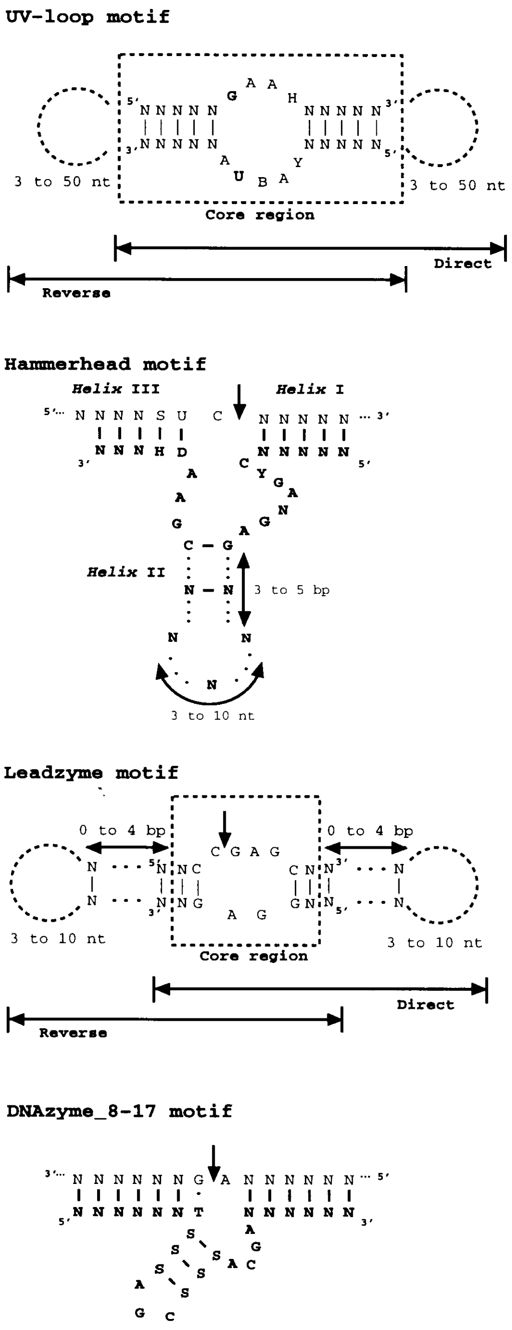


Figure 3. Chemically and catalytically active motifs. (Top) The secondary structure of the UV-loop motif is presented. The nucleotides in bold are involved in cross-linking upon UV irradiation. (Second from top) The hammerhead motif is shown in bold with its target RNA. This motif is composed of three helical regions Helix I, II, and III surrounding a catalytic, non-helical region. In searches with this motif, only the lower bold part of the structure has been encoded in the descriptor to allow distant potential substrates to be found. (Third from top) The leadzyme motif, note that in contrast to the hammerhead motif, both catalytic and substrate portions of this motif were encoded into the descriptor for the searches. (Bottom) The DNAzyme_8-17 motif is presented in bold with its target RNA. As with the hammerhead, only the portion in bold was used in searches. As explained in Figure 1, two orientations were given to the motifs when possible (direct and reverse). The arrows indicate the position of the catalytic cleavage site. The letter code is defined in the legend of Figure 1.

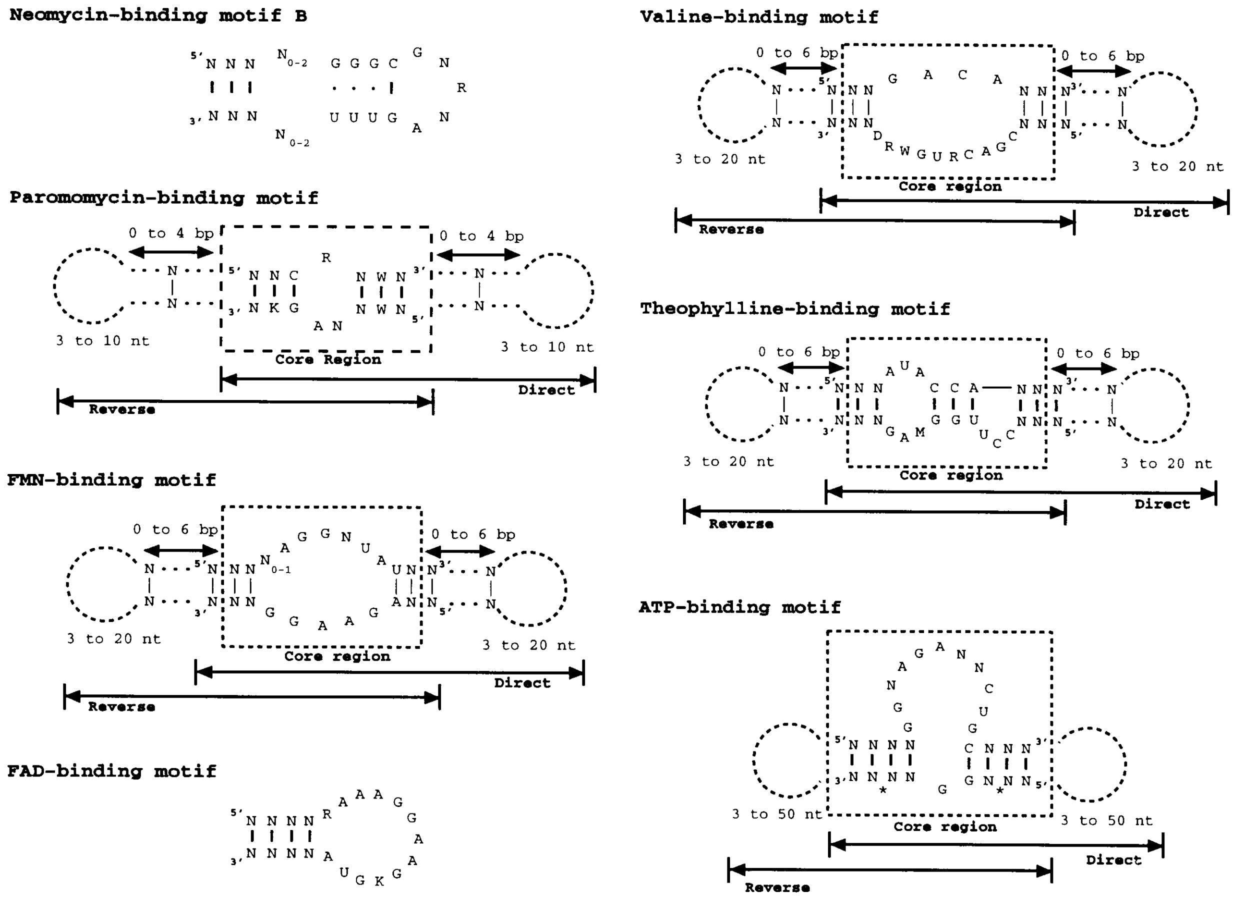


Figure 4. Small molecule-binding RNA motifs. The structure of the neomycin-binding motif B (top left), the paromomycin motif (second left), the FMN- (flavine adenine mononucleotide) binding motif (third left), the flavine adenine (FAD) motif (bottom left), the valine-binding motif (top right), the theophylline-binding motif (middle right) and the ATP-binding motif (bottom right) are presented. In the ATP-binding motif, the asterisks indicate that one mismatch is permitted in the stem. The letter code is defined in the legend of Figure 1 and two orientations were given to the motifs when possible (direct and reverse).

limited analysis we performed confirms quite a wide distribution with few exceptions of conservation such as the ATP motif (see 'ATP in APRT' below) and the hammerheads in satellite DNA. Indeed, in the study of the schistosome catalytic RNA domain, the distribution of the hammerhead motif was determined in related species (14). These data clearly indicate that the hammerhead motif is evolutionarily stable among closely related species of schistosomes. The distribution in conjunction with the biochemical data generated for the motif also shows that the schistosome hammerhead RNA is catalytically active *in vivo*. Thus, even when a motif is generated by a random drift of sequences, this 'evolutionary accident' can be put to profit by the host.

The RNA motifs presented here can be useful or detrimental to the organism and their relatively high (but expected) frequency hints that most of them have little or no effect. Furthermore, it must be kept in mind that distal recognition elements well known *in vivo* have been ignored in our search descriptors. The RBEs, the DNAzyme_8-17 and, in particular, the valine-binding motifs are clearly exceptions. The simplest, obvious explanation for these distributions is that the motifs are somehow detrimental to the host organism.

The activity and utility of motifs

The presence of RNA motifs in the database raises important and cogent issues dealing with not only the activity of an RNA motif in a novel context, but also the ability of the organism or an outside entity to take advantage of the motif in that context. It is unlikely that all of the occurrences identified in this work are active since many unfavorable aspects are to be taken into account such as transcription, compartmentalization, alternative folding, co-localization with the interacting protein or molecule and with co-ions, etc. However, because of its intrinsic properties, RNA can accumulate mutations without changing its secondary structure, providing access to new shapes and motifs (48,51). Since RNA structures are dynamic, the presence of a new, putative motif can confer a potential, novel activity to the RNA. Environmental changes can induce alternative folding within the RNA and encourage the formation of the motifs. Thus knowledge of the presence of a motif by itself is useful information because some of these may be bona fide motifs. For example, our recent study of the hammerhead motif found in schistosome repetitive DNA shows that it is expressed and active in schistosomes and may be involved in the regulation of the synaptobrevin-like protein gene via a *trans*-cleavage of the

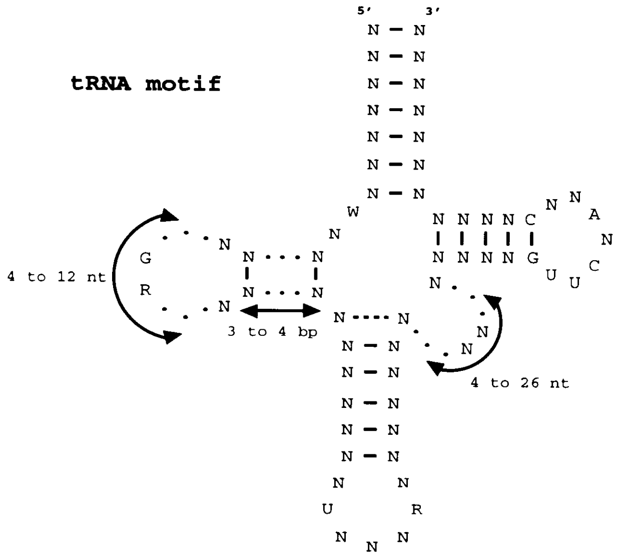


Figure 5. The tRNA motif. Secondary structure of the canonical tRNA motif containing no terminal 'CCA' or introns. The letter code is defined in the legend of Figure 1.

mRNA (14). Other occurrences of the RNA motifs are equally intriguing even if not yet proven. The presence of a neomycin motif B in *Giardia* could explain its high sensitivity to this antibiotic compared with other eukaryotes (52). In the case of the TBE motif, we have suggested that the presence of putative TBE structures in Kaposi sarcoma associated herpes virus (or HHV8) and hepatitis C virus might be related to the fact that these viral infections are exacerbated by the HIV virus (13). Putative TBEs were also found in viruses like shope fibroma that stimulate HIV replication (53). Here are two particularly interesting occurrences.

TBE motif in vaccinia virus. Park *et al.* (54) have reported that HeLa cells express a 'TAR-binding protein' that is a potent inhibitor of the interferon-induced, ribosome-associated protein kinase, PKR, which mediates the antiviral and antiproliferative effects of interferon (55). Vaccinia virus also possesses a similar protein, called E3L whose absence in the replication defective mutant virus can be complemented by the human TAR-binding protein. Park *et al.* (54) have suggested that E3L could provide a means by which the virus could escape the interferon induced antiviral pathway. Finding a TBE motif in the coding region of a subunit of the vaccinia virus RNA polymerase (accession no. VACRNAPSA; position 6000–6034) provides a mechanism: the cellular TAR-binding protein or the viral-encoded counterpart could inhibit PKR by binding to the TBE.

ATP-binding motif in APRT genes. Occurrences of the ATP-binding motif have been found in the second intron of the APRT gene in four closely related rodents: *Mus pahari* (accession no. MPU28721; position 1128–1163), *Stochomys longicaudatus* (accession no. SLU28723; position 1060–1113), *Rattus norvegicus* (accession no. RATAPRT; position 1104–1138) and *Gerbillus campestris* (accession no. GCU28961; position 1193–1227). On phylogenetic trees either based on rodent APRT genes or on other morphological and biochemical analysis (56), the ones with an ATP-binding motif in the intron II cluster in the center of the tree and are phylogenetically linked. In two related rodents, the motif seems to have been lost by a deletion in the intron II of *Mus musculus* and *Mus spicilegus* (accession nos M11310 and U28720, respectively). Since the APRT gene is involved in the salvage pathway of adenine synthesis in mammals, the presence of the motif might play a regulatory role.

Taking advantage of fortuitous targets

RNA motifs do not have to be used by or be useful to the host cell to provide an important entry point to metabolic manipulation of a cell. As the number of defined small molecule- and protein-binding motifs grows, RNA-based intervention could become the method of choice in inhibiting, stimulating or modulating biological processes. Equally exciting is the possible use of the RNA motif database in the identification of secondary targets, when evaluating drugs for which RNA aptamers have been selected. Since it has been proved that the presence of an RNA aptamer in the mRNA of a given gene inhibit its expression upon binding to the ligand (57), it is quite probable that the fortuitous occurrence of a motif in RNA influences its expression.

Conclusion

As the number of unknown sequences accumulates in GenBank databases, recourse to motif search programs will be increasingly useful for *in silico* functional analysis (58). A research strategy based on database searches and experimental approaches was developed. This strategy when coupled to the ability to define new motifs using *in vitro* selection could lead to a virtually limitless source of new concepts in the understanding and use of RNA structures. Using such a strategy, we have already been successful in identifying functional catalytic motifs (14). Likewise, a similar strategy was used by Lowe and Eddy (5) to identify new snoRNA genes in yeast. The occurrences and location files of our searches are available on the web at the URL: http://www.centrcn.umontreal.ca/~bourdeav/Ribonomics

ACKNOWLEDGEMENTS

We thank Alice Rae, Nicholas Delihas and Franz Lang for reviewing the manuscript. The authors wish to express their gratitude to NSERC of Canada who financed this project. V.B. holds a doctoral fellowship from NSERC of Canada. The late R.C. was Richard Ivey Scholar of the Canadian Institute for Advanced Research (CIAR) program in Evolutionary Biology. In addition, we thank Bernard Lorazo, Daniel Raymond and André Fourrier of DiTER (Direction des infrastructures Technologiques d'Enseignement et de Recherche) at the Université de Montréal for their assistance.

REFERENCES

1. Murvai,J., Vlahovicek,K., Barta,E., Szepesvári,C., Acatrinei,C. and Pongor,S. (1999) *Nucleic Acids Res.*, **27**, 257–259.
2. Attwood,T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) *Nucleic Acids Res.*, **27**, 220–225.
3. Henikoff,J.G., Henikoff,S. and Pietrokovski,S. (1999) *Nucleic Acids Res.*, **27**, 226–228.

4. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) *Nucleic Acids Res.*, **27**, 215–219.
5. Lowe,T.M. and Eddy,S.R. (1999) *Science*, **283**, 1168–1171.
6. Michel,F. and Westhof,E. (1996) *Science*, **273**, 1676–1677.
7. Westhof,E., Masquida,B. and Jaeger,L. (1996) *Fold. Des.*, **1**, R78–R88.
8. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
9. Gautheret,D., Major,F. and Cedergren,R. (1990) *Comput. Appl. Biosci.*, **6**, 325–331.
10. Laferrière,A., Gautheret,D. and Cedergren,R. (1994) *Comput. Appl. Biosci.*, **10**, 211–212.
11. Dandekar,T. and Hentze,M.W. (1995) *Trends Genet.*, **11**, 45–50.
12. Steinberg,S. and Cedergren,R. (1995) *RNA*, **1**, 886–891.
13. Ferbeyre,G., Bourdeau,V. and Cedergren,R. (1997) *Trends Biochem. Sci.*, **22**, 115–116.
14. Ferbeyre,G., Smith,J.M. and Cedergren,R. (1998) *Mol. Cell. Biol.*, **18**, 3880–3888.
15. L'Écuyer,P. and Andres,T.H. (1997) *Math. Comput. Simulation*, **44**, 99–107.
16. Bratty,J., Chartrand,P., Ferbeyre,G. and Cedergren,R. (1993) *Biochim. Biophys. Acta*, **1216**, 345–359.
17. Burke,J.M. (1996) *Biochem. Soc. Trans.*, **24**, 608–615.
18. Sodroski,J., Rosen,C., Wong-Staal,F., Salahuddin,S.Z., Popovic,M., Arya,S., Gallo,R.C. and Haseltine,W.A. (1985) *Science*, **227**, 171–173.
19. Cullen,B.R. and Greene,W.C. (1989) *Cell*, **58**, 423–426.
20. Karn,J. and Graeble,M.A. (1992) *Trends Genet.*, **8**, 365–368.
21. Weeks,K.M., Ampe,C., Schultz,S.C., Steitz,T.A. and Crothers,D.M. (1990) *Science*, **249**, 1281–1285.
22. Weeks,K.M. and Crothers,D.M. (1991) *Cell*, **66**, 577–588.
23. Puglisi,J.D., Tan,R., Calnan,B.J., Frankel,A.D. and Williamson,J.R. (1992) *Science*, **257**, 76–80.
24. Zapp,M.L. and Green,M.R. (1989) *Nature*, **342**, 714–716.
25. Cullen,B.R. and Malim,M.H. (1991) *Trends Biochem. Sci.*, **16**, 346–350.
26. Tan,R., Chen,L., Buettner,J.A., Hudson,D. and Frankel,A.D. (1993) *Cell*, **73**, 1031–1040.
27. Giver,L., Bartel,D., Zapp,M., Pawul,A., Green,M. and Ellington,A.D. (1993) *Nucleic Acids Res.*, **21**, 5509–5516.
28. Ringquist,S., Jones,T., Snyder,E.E., Gibson,T., Boni,I. and Gold,L. (1995) *Biochemistry*, **34**, 3640–3648.
29. Branch,A.D, Benenfeld,B.J., Baroudy,B.M., Wells,F.V., Gerin,J.L. and Robertson,H.D. (1989) *Science*, **243**, 649–652.
30. Pan,T. and Uhlenbeck,O.C. (1992) *Nature*, **358**, 560–563.
31. Chartrand,P., Usman,N. and Cedergren,R. (1997) *Biochemistry*, **36**, 3145–3150.
32. Ciesiolka,J., Michalowski,D., Wrzesinski,J., Krajewski,J. and Krzyzosiak,W.J. (1998) *J. Mol. Biol.*, **275**, 211–220.
33. Santoro,S.W. and Joyce,G.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 4262–4266.
34. Paquette,J., Nicoghosian,K., Qi,G.R., Beauchemin,N. and Cedergren,R. (1990) *Eur. J. Biochem.*, **189**, 259–265.
35. Breaker,R.R. and Joyce,G.F. (1994) *Chem. Biol.*, **1**, 223–229.
36. Cuenoud,B. and Szostak,J.W. (1995) *Nature*, **375**, 611–614.
37. Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–510.
38. Joyce,G.F. (1989) *Gene*, **82**, 83–87.
39. Ellington,A.D. and Szostak,J.W. (1990) *Nature*, **346**, 818–822.
40. Davies,J. and Davis,B.D. (1968) *J. Biol. Chem.*, **243**, 3312–3316.
41. Famulok,M. and Huttenhofer,A. (1996) *Biochemistry*, **35**, 4265–4270.
42. Recht,M.I., Fourmy,D., Blanchard,S.C., Dahlquist,K.D. and Puglisi,J.D. (1996) *J. Mol. Biol.*, **262**, 421–436.
43. Majerfeld,I. and Yarus,M. (1994) *Nature Struct. Biol.*, **1**, 287–292.
44. Jenison,R.D., Gill,S.C., Pardi,A. and Polisky,B. (1994) *Science*, **263**, 1425–1429.
45. Burgstaller,P. and Famulok,M. (1994) *Angew Chem. Int. Ed. Engl.*, **33**, 1084–1087.
46. Sassanfar,M. and Szostak,J.W. (1993) *Nature*, **364**, 550–553.
47. Green,B., Pabon-Peña,L., Graham,T.A., Peach,S.E., Coats,S.R. and Epstein,L.M. (1993) *Mol. Biol. Evol.*, **10**, 732–750.
48. Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) *Proc. R Soc. Lond. B Biol. Sci.*, **255**, 279–284.
49. Fontana,W., Konings,D.A., Stadler,P.F. and Schuster,P. (1993) *Biopolymers*, **33**, 1389–1404.
50. Reidys,C., Stadler,P.F. and Schuster,P. (1997) *Bull. Math. Biol.*, **59**, 339–397.
51. Huynen,M.A. (1996) *J. Mol. Evol.*, **43**, 165–169.
52. Andrews,B.J., Panitescu,D., Jipa,G.H., Vasile-Bugarin,A.C., Vasiliu,R.P. and Ronnevig,J.R. (1995) *Am. J. Trop. Med. Hyg.*, **52**, 318–321.
53. Tseng,C.K., Hughes,M.A., Hsu,P.L., Mahoney,S., Duvic,M. and Sell,S. (1991) *Am. J. Pathol.*, **138**, 1149–1164.
54. Park,H., Davies,M.V., Langland,J.O., Chang,H.-W., Nam,Y.S., Tartaglia,J., Paoletti,E., Jacobs,B.L., Kaufman,R.J. and Venkatesan,S. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 4713–4717.
55. McMillan,N.A., Chun,R.F., Siderovski,D.P., Galabru,J., Toone,W.M., Samuel,C.E., Mak,T.W., Hovanessian,A.G., Jeang,K.T. and Williams,B.R. (1995) *Virology*, **213**, 413–424.
56. Fieldhouse,D., Yazdani,F. and Golding,G.B. (1997) *Heredity*, **78**, 21–31.
57. Werstuck,G. and Green,M.R. (1998) *Science*, **282**, 296–298.
58. Segovia,L. (1998) *Nature Biotechnol.*, **16**, 25.
59. Leclerc,F., Cedergren,R. and Ellington,A.D. (1994) *Nature Struct. Biol.*, **1**, 293–300.