188–191 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

UTRdb: a specialized database of 5′ and 3′ untranslated regions of eukaryotic mRNAs

Graziano Pesole[1,2,*], Sabino Liuni[2,3], Giorgio Grillo[4], Matilde Ippedico[4], Alessandra Larizza[4], Wojciech Makalowski[5] and Cecilia Saccone[2,3,4]

[1]Dipartimento di Biologia, D.B.A.F., Università della Basilicata, via Anzio 10, 85100 Potenza, Italy, [2]Area di Ricerca di Bari, C.N.R., via Amendola 166/5, 70126 Bari, Italy, [3]Centro di Studio sui Mitocondri e Metabolismo Energetico C.N.R. and [4]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona 4, 70126 Bari, Italy and [5]National Center for Biotechnology Information, NLM-NIH, Bethesda, MD, USA

Received September 28, 1998; Accepted September 30, 1998

ABSTRACT

The 5′ and 3′ untranslated regions of eukaryotic mRNAs may play a crucial role in the regulation of gene expression controlling mRNA localization, stability and translational efficiency. For this reason we developed UTRdb (http://bigarea.area.ba.cnr.it:8000/BioWWW/#UTRdb), a specialized database of 5′ and 3′ untranslated sequences of eukaryotic mRNAs cleaned from redundancy. UTRdb entries are enriched with specialized information not present in the primary databases including the presence of nucleotide sequence patterns already demonstrated by experimental analysis to have some functional role. All these patterns have been collected in the UTRsite database so that it is possible to search any input sequence for the presence of annotated functional motifs. Furthermore, UTRdb entries have been annotated for the presence of repetitive elements.

INTRODUCTION

Understanding the basic mechanisms of cell growth, differentiation and response to environmental stimuli, i.e. the program controlling the temporal and spatial order of molecular events, is becoming a real challenge in molecular biology. Indeed, although most of the regulatory elements are thought to be embedded in the non-coding part of the genomes, nucleotide databases are biased by the presence of expressed sequences mostly corresponding to the protein coding portion of the genes. Among non-coding regions, the 5′ and 3′ untranslated regions (5′-UTR and 3′-UTR) of eukaryotic mRNAs have often been experimentally demonstrated to contain sequence elements crucial for many aspects of gene regulation and expression (1–6).

The main functional roles so far demonstrated for 5′- and 3′-UTR sequences are: (i) control of mRNA cellular and subcellular localization (4,7); (ii) control of mRNA stability (1,8); (iii) control of mRNA translation efficiency (9,10).

Several regulatory signals have been already identified in 5′- or 3′-UTR sequences, usually corresponding to short oligonucleotide tracts, also able to fold in specific secondary structures, which are protein binding sites for various regulatory proteins.

The analysis of large collections of functionally equivalent sequences (11,12), such as 5′-and 3′-UTR sequences, could be indeed very useful for defining their structural and compositional features as well as for searching the alleged function-associated sequence patterns (13–15). For this reason we constructed UTRdb, a specialized sequence collection, deprived from redundancy, of 5′- and 3′-UTR sequences from eukaryotic mRNAs.

UTRdb entries have been enriched with specialized information, not present in the primary databases, including the presence of sequence patterns demonstrated by experimental evidence to play some functional role. Additionally, because ~10% of mammalian mRNAs contain repetitive elements in their UTRs (16) which are not usually annotated in the original records, we decided to include this information in our database as well.

We also created UTRsite, a collection of functional sequence patterns located in the 5′- or 3′-UTR sequences which could prove very useful for automatic annotation of anonymous sequences generated by sequencing projects as well as for finding previously undetected signals in known gene sequences.

ASSEMBLING UTRdb COLLECTIONS

The specialized database of UTR sequences was generated by UTRdb_gen, a computer program we devised for this task. Seven sequence collections were generated for both 5′- and 3′-UTR sequences, one for each of the eukaryotic divisions of the EMBL/GenBank nucleotide database, namely: (i) human; (ii) rodent; (iii) other mammal; (iv) other vertebrate; (v) invertebrate; (vi) plant; (vii) fungi.

UTRdb_gen, performing an accurate parsing of the Feature Table of the relevant EMBL entries is able to automatically generate the various UTRdb collections. Although the Feature keys ‘5′UTR’ and ‘3′UTR’ are valid features for the EMBL/Gen-

*To whom correspondence should be addressed at: Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona 4, 70126 Bari, Italy. Tel. +39 80 544 3305; Fax: +39 80 544 3317; Email: graziano@area.ba.cnr.it

Bank entries, only a small percentage of the entries are adequately annotated. Indeed, of the 65 084 primary entries where UTRdb_gen was able to extract 5′- or 3′-UTR sequences, only 10.9% contained the 5′UTR or 3′UTR feature key. UTRdb_gen is able to define UTR regions even when these keys are not reported in the primary entry by using a predefinite syntactic parsing of other relevant feature keys, such as mRNA, CDS, exon, intron, etc.

UTRdb_gen automatically annotates generated UTR entries by adding some specialized information such as completeness or not of the UTR region, number of spanned exons and cross-referencing to the primary database entry. A cross reference between 5′- and 3′-UTR sequences from the same mRNA has also been established.

The generation of UTR entries cleaned from redundancy has been obtained by using CLEANUP program (17) which is able to generate automatically, and fast, cleaned collections by removing entries having a similarity and overlapping degree with longer entries present in the database above a user-fixed threshold. In this case the cut-off parameters we used for the CLEANUP application were 95% for similarity and 90% for overlapping.

The UTR entries have been further enriched, by using the program UTRnote (kindly provided by G. Grillo, S. Brunetta and D. Colangelo) including information about the location of experimentally defined patterns collected in UTRsite and of repetitive elements present in the Repbase database (18). The UTRsite entries describe the various regulatory elements present in UTR regions whose functional role has been established on an experimental basis. Each UTRsite entry is constructed on the basis of information reported in the literature and revised by scientists experimentally working on the functional characterization of the relevant UTR regulatory element.

CONTENT OF UTRdb

Table 1 reports a summary description of UTRdb (release 7.0) which in total contains 65 084 entries and 22 489 119 nucleotides. On average, >30% of entries proved to be redundant and were removed from the database.

5′-UTR sequences were defined as the mRNA region spanning from the cap site to the starting codon (excluded), whereas 3′-UTR sequences were defined as the mRNA region spanning from the stop codon (excluded) to the poly-A starting site.

A sample entry of UTRdb is shown in Figure 1. The UTRdb entries have been formatted according to the EMBL database format.

Table 2 reports functional patterns and repetitive elements included in UTRsite (release 2.0). More entries will be included in further releases. A sample UTRsite entry is reported in Figure 2. Functional patterns, defined on the basis of the information reported in the literature and/or advice by the scientists expert in the field, were described by using the pattern description syntax used in the PATSCAN program (19).

We have made available a novel version of PATSCAN (Sandra Brunetta and Donata Colangelo) at http://bio-www.ba.cnr.it:8000/BioWWW/patscanGCG.html

Table 1. Number of entries (N) and nucleotide length (L) of UTRdb collections (release 7.0) after redundancy cleaning

Collection			Redundancy	
	N	L	%N	%L
5′-UTR				
Fungi	931	161 789	25.1	13.2
Human	6669	1 380 128	39.0	21.8
Invertebrate	4054	829 754	28.1	25.4
Other_mammal	2106	291 282	35.5	17.0
Other_vertebrate	2846	431 565	27.2	20.7
Plant	5993	633 846	24.0	11.2
Rodent	7056	1 357 490	35.6	19.7
Total	**29 655**	**5 085 854**		
3′-UTR				
Fungi	1154	286 567	13.2	8.1
Human	7503	5 541 031	37.9	24.5
Invertebrate	5067	2 006 870	20.3	19.8
Other_mammal	2457	1 172 955	37.9	24.9
Other_vertebrate	3499	1 697 054	22.6	15.0
Plant	8116	1 965 316	15.2	13.1
Rodent	7633	4 733 472	35.1	21.7
Total	**35 429**	**17 403 265**		

UTRdb 7.0 was generated from EMBL release 54. Relevant redundancy percentages calculated with respect to the number of entries (%N) and to the nucleotide length (%L) are also indicated.

Table 2. Functional patterns included so far in UTRsite (v2.0)

Functional Patterns	Reference	Hits found in UTRdb 7.0
Iron-responsive element (IRE)	20	65
Histone 3′-UTR stem–loop structure	21	27
AU-rich class II destabilizing element	22	175
TGE translational regulation element	23	45
Selenocysteine insertion sequence (SECIS)	24,25	189
APP 3′-UTR stability control element	26	7
Cytoplasmatic polyadenylation element (CPE)	27	4614
Nanos	28	397
15-LOX-DICE	29	83
Repetitive elements		5476

For each pattern the number of hits with UTRdb entries is also reported.

AVAILABILITY OF UTRdb

UTRdb is publicly available by anonymous FTP (ftp://area.ba.cnr.it/pub/embnet/database/utr/) and UTRdb entries can be

```
ID   3MMU006284 standard; DNA; ROD; 3044 BP.
XX
AC   CC051318;
XX
DT   03-APR-1998 (Rel. 7, Created)
DT   03-APR-1998 (Rel. 7, Last updated, Version 1)
XX
DE   3'UTR in Mus musculus insulin-like growth factor II (Igf2) gene, complete
DE   cds.
XX
DR     EMBL; U71085;
DR     UTR; BB045319;
DR     UTR; BB045320;
DR     UTR; BB045321;
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Rodentia;
OC   Sciurognathi; Muridae; Murinae; Mus.
XX
UT     3'UTR; Complete; 1 exon(s).
XX
FH   Key             Location/Qualifiers
FH
FT   3'UTR               EMBL::U71085:24594..27637
FT                      /gene="Igf2"
FT                       /product="insulin-like growth factor II"
FT   IRE                 EMBL::U71085:25658..25680
FT                       /evidence="Pattern Similarity"
FT                        /standard_name="Iron Responsive Element"
FT                       /db_xref="UTRsite:U0002"
FT   CPE                 EMBL::U71085:26356..26438
FT                       /evidence="Pattern Similarity"
FT                        /standard_name="Cytoplasmic Polyadenylation Element"
FT                       /db_xref="UTRsite:U0005"
FT   NANOS               EMBL::U71085:27558..27572
FT                       /evidence="Pattern Similarity"
FT                        /standard_name="nanos Translation Control Element"
FT                       /db_xref="UTRsite:U0007"
FT   RSINE2              EMBL::U71085:25414..25462
FT                       /evidence="Pattern Similarity"
FT                       /repeat_name="RSINE2"
FT                       /repeat_class="SINE/B4"
XX
SQ   Sequence 3044 BP; 775 A; 825 C; 710 G; 734 T; 0 other;
     ATCAAATTAT GTGGTAATTC TGCAATGTAG TACCATCAAT CTGTGACCTC CTCTTGAGCA        60
     GGGACAGTTC CATCACGTCC CACACTAAGA TCTCTCTGCT CCACTCCCTC CCCAGGTTTC       120
     .......... .......... .......... .......... .......... ..........
//
```

Figure 1. Sample entry of UTRdb. Specialized information not present in the primary EMBL/GenBank database is shown in boldcase with active crosslinks with other databases underlined. The 'UT' line reports information about completeness or not of the relevant UTR entry (e.g. complete or partial) as well as the number of spanned exons in the case of genomic DNA sequences. The presence in this sequence entry of an 'iron responsive element' (20) (UTRsite entry: U0002), of a 'cytoplasmatic polyadenylation element' (27) (UTRsite entry: U0005), of a Nanos element (28) (UTRsite entry: U0007) and of a repetitive element SINE/B4 has been also annotated.

```
<Entry>
IRON-RESPONSIVE ELEMENT; U0002
<Description>
The "iron-responsive element" (IRE) is a particular hairpin structure located in the 5'-
untranslated region (5'-UTR) or in the 3'-untranslated region (3'-UTR) of various mRNAs coding
for proteins involved in cellular iron metabolism. The IREs are recognized by trans-acting
proteins known as Iron Regulatory Proteins (IRPs) that control mRNA translation rate and
stability. Two closely related IRPs, denoted as IRP-1 and IRP-2, have been identified so far
which bind IREs and become inactivated (IRP-1) or degradated (IRP-2) when the iron level in the
cell increases. IRPs show a significant degree of similarity to mitochondrial aconitase (EC
4.2.1.3). It has been shown that under high iron conditions IRP-1, which contains a 4Fe-4S
cluster that possibly acts as a cellular iron biosensor, has enzymatic activity and may act as a
cytosolic aconitase.
Cellular iron homeostasis in mammalian cells is maintained by the coordinate regulation of the
expression of "Transferrin receptor", which determines the amount of iron acquired by the cell,
and of "Ferritin", an iron storage protein, which determines the degree of intracellular iron
sequestration. Thus if the cell requires more iron, the level of transferrin receptor has to
increase and conversely the level of ferritin has to decrease.
Ferritin, in vertebrates, consists of 24 protein subunits of two types, type H with Mr of 21 kDa
and type L with Mr of 19-20 kDa. The apoprotein (Mr 450 kDa) is able to store up to 4500 Fe
(III) atoms.
The 5'-UTR of H- and L ferritin mRNAs contain one IRE whereas multiple IREs are located in the
3'-UTR of transferrin receptor mRNA.
In the case of low iron concentration, IRPs are able to bind the IREs in the 5'-UTR of H- and L-
Ferritin mRNAs repressing their translation and the IREs in the 3'-UTR of transferrin mRNA
increasing its stability. Conversely, if iron concentration is high, IRP binding is diminished,
which increases translation of ferritins and downregulate expression of the transferrin
receptor.
IREs have also been found in the mRNAs of other proteins involved in iron metabolism like
"erythroid 5-aminolevulinic-acid synthase (eALAS) " involved in heme biosynthesis, the mRNA
encoding the mitochondrial aconitase (a citric acid cycle enzyme) and the mRNA encoding the
iron-sulfur subunit of succinate dehydrogenase (another citric acid cycle enzyme) in Drosophila
melanogaster.
Two alternative IRE consensus (type A or type B) have been found. In certain IREs the bulge is
best drawn with a single unpaired cytosine, whereas in others the cytosine nucleotide and two
additional bases seem to oppose one free 3' nucleotide. Some evidences also suggest a structured
loop with an interaction between nucleotide one and nucleotide five (in boldcase).

   G  W          G  W
 A    G        A    G
  C  H          C  H
   NN            NN
   NN            NN
   NN            NN
   NN            NN
   NN            NN
  C              C
   NN          N   N
   NN            N
   NN           NN
   NN           NN
   NN           NN

The lower stem can be of variable length and is AU-rich in transferrin mRNA. W=A,U and D=not G.
<Pattern>
r1={au,ua,gc,cg,gu,ug} ! r1 represents pairing rules
(p1=2...8 c p2=5...5 CAGWGH r1~p2 r1~p1 | p1=2...8 nnc p2=5...5 CAGWGH r1~p2 n r1~p1)
!(type A|type B)
<Bibliography>
Hentze MW and Kuhn LC (1996) Molecular control of vertebrate iron metabolism: mRNA based
regulatory circuits operated by iron, nitric oxide, and oxidative stress. Proc. Natl. Acad. Sci.
USA 93: 8175-8182.
```

Figure 2. Sample entry of UTRsite describing the 'iron responsive element (IRE)' (20). The IRE functional pattern which consists of both primary and secondary structure information is described in the 'Pattern' section according to the format adopted by PATSCAN program (http://bio-www.ba.cnr.it:8000/BioWWW/patscanGCG.html).

retrieved on the Web by using SRS at the EBI WWW server (http://srs.ebi.ac.uk:80/) or at the BioWWW server (http://bio-www.ba.cnr.it:8000/srs/). SRS retrieval allows crosslinking between UTRdb, EMBL/GenBank and UTRsite entries. The computer program UTRscan (http://bio-www.ba.cnr.it:8000/cgi-bin/BioWWW/UTRscanHTML.pl) has been made available to search UTRsite patterns in users submitted sequences.

CONCLUSIONS AND PERSPECTIVES

The important role that UTRs of eukaryotic mRNAs may play in gene regulation and expression is now widely recognized. Indeed, experimental studies have demonstrated that sequence motifs located in the UTRs are involved in crucial biological functions.

The huge amount of functionally equivalent sequences stored in UTRdb now makes possible the study of their structural and compositional features and the application of statistical methods for the identification of significant signals. Previous cleaning-up of databases was, however, necessary to avoid artefacts caused by redundant sequences. Even if statistical significance does not necessarily mean biological significance, it may provide useful indications for further experimental work, such as site-directed mutagenesis.

UTRdb will be updated with the new EMBL database releases and UTRsite will be continuously updated by adding new entries describing functional patterns whose biological role has been experimentally demonstrated.

ACKNOWLEDGEMENTS

For revision of UTRsite entries we would like to thank Jim Malter (APP 3′-UTR stability control element), Alain Krol (SECIS), Matthias Hentze (IRE and 15-LOX DICE), Bill Marzluff (histone stem–loop structure), Ann-Bin Shyu (ARE), Arturo Verrotti (CPE), Robin Wharton (nanos) and Elizabeth Goodwin (TGE). This work was supported by EU grants BIO4-CT95-0130 and ERB-BIO4-CT96-0030.

REFERENCES

1 Decker,C.J. and Parker,R. (1994) *Trends Biochem. Sci.*, **19**, 336–340.
2 Kaufman,R.J. (1994) *Curr. Opin. Biotechnol.*, **5**, 550–557.
3 Klausner,R.D., Rouault,T.A. and Harford,J.B. (1993) *Cell*, **72**, 19–28.
4 Singer,R.H. (1992) *Curr. Opin. Cell Biol.*, **4**, 15–19.
5 Wilhelm,J.E. and Vale,R.D. (1993) *J. Cell. Biol.*, **123**, 269–274.

6 McCarthy,J.E.G. and Kollmus,H. (1995) *Trends Biochem. Sci*., **20**, 191–197.
7 Johnston,D. (1995) *Cell*, **81**, 161–170.
8 Beelman,C.A. and Parker,R. (1995) *Cell*, **81**, 179–183.
9 Curtis,D., Lehman,R. and Zamore,P.D. (1995) *Cell*, **81**, 171–178.
10 Sonenberg,N. (1994) *Curr. Opin. Gen. Dev*., **4**, 310–315.
11 Mengeritsky,G. and Smith,T.F. (1987) *Comput. Applic. Biosci*., **3**, 223–227.
12 Konopka,A.K. (1994) In Smith,D.W. (ed.), *Informatics and Genome Projects*. Academic Press, San Diego, CA.
13 Pesole,G., Liuni,S., Grillo,G. and Saccone,C. (1997) *Gene*, **205**, 95–102.
14 Pesole,G., Grillo,G. and Liuni,S. (1996) *Comp. Chem*., **20**, 141–144.
15 Pesole,G., Fiormarino,G. and Saccone,C. (1994) *Gene*, **140**, 219–225.
16 Makalowski,W., Zhang,J. and Boguski,M. (1996) *Genome Res*., **6**, 846–857.
17 Grillo,G., Attimonelli,M., Liuni,S. and Pesole,G. (1996) *Comput. Applic. Biosci*., **12**, 1–8.
18 Jurka,J. (1998) *Curr. Opin. Struct. Biol*., **8**, 333–337.
19 Dsouza,M., Larsen,N. and Overbeek,R. (1997) *Trends Genet*., **13**, 497–498.
20 Hentze,M.W. and Kuhn,L.C. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.
21 Williams,A.S. and Marzluff,W.F. (1995) *Nucleic Acids Res*., **23**, 654–662.
22 Chen,C. and Shyu,A. (1995) *Trends Biochem. Sci*., **20**, 465–470.
23 Goodwin,E.B., Okkema,P.G., Evans,T.C. and Kimble,J. (1993) *Cell*, **75**, 329–339.
24 Hubert,N., Walczak,R., Sturchler,C., Schuster,C., Westhof,E., Carbon,P. and Krol,A. (1996) *Biochimie*, **78**, 590–596.
25 Walczak,R., Westhof,E., Carbon,P. and Krol,A. (1996) *RNA*, **2**, 367–379.
26 Zaidi,S.H.E. and Malter,J.S. (1994) *J. Biol. Chem*., **269**, 24007–24013.
27 Verrotti,A., Thompson,S., Wreden,C., Strickland,S. and Wickens,M. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 9027–9032.
28 Dahanukar,A. and Wharton,R. (1996) *Genes Dev*., **10**, 2610–2620.
29 Ostareck-Lederer,A., Ostareck,D., Standart,N. and Thiele,B. (1994) *EMBO J*., **13**, 1476–1481.