 Nucleic Acids Research, 2000, Vol. 28, No. 1 193–196

UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs

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

[1]Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26, 20133 Milano, Italy, [2]Area di Ricerca di Bari, Consiglio Nazionale delle Ricerche (CNR), via Amendola 166/5, 70126 Bari, Italy, [3]Centro di Studio sui Mitocondri e Metabolismo Energetico del Consiglio Nazionale delle Ricerche (CNR), via Orabona 4, 70126 Bari, Italy, [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 30, 1999; Accepted October 4, 1999

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, 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. All internet resources implemented for retrieval and functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs are accessible at http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/

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 (1–7).

The main functional roles so far demonstrated for 5′- and 3′-UTR sequences are: (i) control of mRNA cellular and subcellular localization (4,7,8); (ii) control of mRNA stability (1,9); and (iii) control of mRNA translation efficiency (10,11).

Several regulatory signals have already been identified in 5′- and 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 (12,13), such as 5′- and 3′-UTR sequences, could indeed be very useful for defining their structural and compositional features as well as for searching the alleged function-associated sequence patterns (14–16). 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 (17) which are not usually annotated in the original records, we decided to include this information in our database.

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. Eight sequence collections were generated for both 5′- and 3′-UTR

*To whom correspondence should be addressed at: Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26, 20133 Milano, Italy. Tel: +39 02 7064 4803; Fax: +39 02 7063 2811; Email: graziano.pesole@unimi.it

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; and (viii) Patent.

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/Genbank entries, only a small percentage of the entries are adequately annotated. Indeed, of the 120 767 primary entries where UTRdb_gen was able to extract 5′- or 3′-UTR sequences, only 15.8% contained the 5′UTR or 3′UTR feature key in the corresponding EMBL entry. 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 (18) which is able to generate automatically, very quickly, 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, Area de Ricerca di Bari del Consiglio Nazionale delle Ricerche) including information about the location of experimentally defined patterns collected in UTRsite and of repetitive elements present in the Repbase database (19). 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 distinguished 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 12.0) which in total contains 120 767 entries and 37 353 172 nucleotides. On average, >29.3% 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 UTRdb entry 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 3.0). More entries will be included in further releases. A sample UTRsite entry is reported in Figure 2. Functional patterns, defined on the basis

Table 1. Number of entries (N) and nucleotide length (L) of UTRdb collections (release 12.0) after redundancy cleaning

			Redundancy	
	N	L	%N	%L
5′-UTR				
Fungi	1136	195 215	23.91	13.04
Human	8785	1 887 755	38.61	28.15
Invertebrate	5376	1 033 413	27.63	15.52
Other_mammal	2429	339 321	36.06	27.62
Other_vertebrate	3564	519 656	25.63	18.19
Plant	8499	924 695	24.91	13.98
Rodent	8496	1 629 025	34.98	24.92
Patent	213	55 918	29.00	41.86
TOTAL	38 498	6 584 998		
3′-UTR				
Fungi	1415	338 564	13.61	9.47
Human	10 207	8 367 057	36.91	30.95
Invertebrate	6677	2 607 959	19.89	17.06
Other_mammal	3202	1 457 422	29.14	24.27
Other_vertebrate	4419	2 195 694	21.22	14.36
Plant	11 548	2 777 812	15.16	14.15
Rodent	9181	5 737 426	34.66	27.41
Patent	232	91 287	27.04	43.03
TOTAL	46 881	23 573 221		

UTRdb 12.0 was generated from EMBL release 59. Relevant redundancy percentages calculated with respect to the number of entries (%N) and to the nucleotide length (%L) are also indicated.

```
ID   5HSA012029 standard; DNA; HUM; 253 BP.
XX
AC   BB046362;
XX
DT   14-OCT-1998 (Rel. 9, Created)
DT   14-OCT-1998 (Rel. 9, Last updated, Version 1)
XX
DE   5'UTR in Homo sapiens chromosome 7q22 sequence, complete sequence.
XX
DR   EMBL; AF053356;
DR   UTR; CC052750;
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheria;
OC   Primates; Catarrhini; Hominidae; Homo.
XX
UT   5'UTR; Complete; 2 exon(s)
XX
FH   Key             Location/Qualifiers
FH
FT   5'UTR           join(complement(EMBL::AF053356:101757..101920),
FT                   complement(EMBL::AF053356:99415..99503))
FT                   /product="GNB2"
FT   5'TOP           complement(EMBL::AF053356:101916..101920)
FT                   /evidence="Pattern Similarity"
FT                   /standard_name="Ribosomal mRNA 5'terminal oligopyrimidine Element"
FT                   /db_xref="UTRsite:U0010"
FT   repeat_region   complement(EMBL::AF053356:101820..101861)
FT                   /evidence="Pattern Similarity"
FT                   /repeat_type="(CCG)n"
FT                   /repeat_family="Simple_repeat"
XX
SQ   Sequence 253 BP; 29 A; 127 C; 82 G; 15 T; 0 other;
     ctccgctctg gggaggcagc gctggcggcg gggctggggc cactgaggaa atccatccgc        6
     gccgccgccg ccgccgccgc cgccgccgcc gcctccgccg cggaggaaga cagcgccgcc       12
     cgcgcaccgc cagcgacccc cgccgcagag tcccaccgcc acaggcctcg ggccagcggc       18
     caggagctgc ctcccccagc cccgtcccg cggcccccag ccgcccccaa ccctgcccca       24
     cgggcccggc gcc                                                         25
//
```

Figure 1. Sample entry of UTRdb. Specialized information not present in the primary EMBL/GenBank database is shown in bold case 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 a '5′ribosomal mRNA TOP' (32–34) (UTRsite entry: U0010) and of a microsatellite element has also been annotated.

Table 2. Functional patterns included so far in UTRsite (v3.0)

Functional patterns	Reference	Hits found in UTRdb 12.0
Iron-responsive element (IRE)	23	65
Histone 3′UTR stem–loop structure	24	27
AU-rich class II destabilising element	25	175
TGE translational regulation element	26	45
Selenocysteine insertion sequence (SECIS)	27,28	189
APP 3′-UTR stability control element	29	7
Cytoplasmatic polyadenylation element (CPE)	30	4614
Nanos	31	397
ribosomal protein mRNA 5′ TOP	32–34	298
TNF mRNA translation repression element	35	14
Vimentin 3′UTR mRNA element	36	12
GLUT1 mRNA stabilising element	37	48
15-LOX-DICE	38	83
Repetitive elements		44 806

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

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 (20).

AVAILABILITY OF UTRdb

UTRdb and UTRsite are publicly available by anonymous FTP (ftp://area.ba.cnr.it/pub/embnet/database/utr/). All internet resources we implemented for retrieval and functional analysis of 5′- and 3′-UTR sequences are accessible at http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/ (21). These include SRS retrieval (22) of UTRdb and UTRsite, also available at the EBI WWW server (http://srs.ebi.ac.uk:80/), UTRscan and UTRfasta. The UTRscan utility allows the enquirer to search user-submitted sequences for any of the patterns collected in UTRsite. The UTRfasta utility allows database searches against fully annotated UTRdb entries.

CONCLUSIONS AND PERSPECTIVES

The important role that untranslated regions of eukaryotic mRNAs may play in gene regulation and expression is now

<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)' (23). 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 the PATSCAN program (http://bio-www.ba.cnr.it:8000/BioWWW/patscanGCG.html).

widely recognized. Indeed, experimental studies have demonstrated that sequence motifs located in the untranslated regions 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 is necessary however to avoid artefacts caused by redundant sequences. Even if statistical significance does not necessarily mean biological significance, it may provide a useful indication 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), Elizabeth Goodwin (TGE), Roger Kaspar (ribosomal protein mRNA TOP), Danuta Radzioch (TNF mRNA translation repression element), Ruben Boado (GLUT1 mRNA stabilising element) and Zendra E. Zehner (Vimentin 3′UTR mRNA element). This work was supported by EU grant ERB-BIO4-CT96-0030 and by Programma Biotecnologie legge 95/95 (MURST 5%).

REFERENCES

1. Decker,C.J. and Parker,R. (1994) *Trends Biochem. Sci.*, **19**, 336–340.
2. Kaufman,R.J. (1994) *Curr. Opin. Biotech.*, **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. Bashirullah,A., Cooperstock,R.L. and Lipshitz,H.D. (1998) *Annu. Rev. Biochem.*, **67**, 335–394.
8. Johnston,D. (1995) *Cell*, **81**, 161–170.
9. Beelman,C.A. and Parker,R. (1995) *Cell*, **81**, 179–183.
10. Curtis,D., Lehman,R. and Zamore,P.D. (1995) *Cell*, **81**, 171–178.
11. Sonenberg,N. (1994) *Curr. Opin. Genet. Dev.*, **4**, 310–315.
12. Mengeritsky,G. and Smith,T.F. (1987) *Comput. Appl. Biosci.*, **3**, 223–227.
13. Konopka,A.K. (1994) In Smith,D.W. (ed.), *Informatics and Genome Projects*. Academic Press, San Diego, CA.
14. Pesole,G., Liuni,S., Grillo,G. and Saccone,C. (1997) *Gene*, **205**, 95–102.
15. Pesole,G., Grillo,G. and Liuni,S. (1996) *Comp. Chem.*, **20**, 141–144.
16. Pesole,G., Fiormarino,G. and Saccone,C. (1994) *Gene*, **140**, 219–225.
17. Makalowski,W., Zhang,J. and Boguski,M. (1996) *Genome Res.*, **6**, 846–857.
18. Grillo,G., Attimonelli,M., Liuni,S. and Pesole,G. (1996) *Comput. Appl. Biosci.*, **12**, 1–8.
19. Jurka,J. (1998) *Curr. Opin. Struct. Biol.*, **8**, 333–337.
20. Dsouza,M., Larsen,N. and Overbeek,R. (1997) *Trends Genet.*, **13**, 497–498.
21. Pesole,G. and Liuni,S. (1999) *Trends Genet.*, **15**, 379–380.
22. Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
23. Hentze,M.W. and Kuhn,L.C. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.
24. Williams,A.S. and Marzluff,W.F. (1995) *Nucleic Acids Res.*, **23**, 654–662.
25. Chen,C. and Shyu,A. (1995) *Trends Biochem. Sci.*, **20**, 465–470.
26. Goodwin,E.B., Okkema,P.G., Evans,T.C. and Kimble,J. (1993) *Cell*, **75**, 329–339.
27. Hubert,N., Walczak,R., Sturchler,C., Schuster,C., Westhof,E., Carbon,P. and Krol,A. (1996) *Biochimie*, **78**, 590–596.
28. Walczak,R., Westhof,E., Carbon,P. and Krol,A. (1996) *RNA*, **2**, 367–379.
29. Zaidi,S.H.E. and Malter,J.S. (1994) *J. Biol. Chem.*, **269**, 24007–24013.
30. Verrotti,A., Thompson,S., Wreden,C., Strickland,S. and Wickens,M. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 9027–9032.
31. Dahanukar,A. and Wharton,R. (1996) *Genes Dev.*, **10**, 2610–2620.
32. Amaldi,F. and Pierandrei-Amaldi,P. (1997) *Prog. Mol. Subcell. Biol.*, **18**, 1–17.
33. Kaspar,R.L., Kakegawa,T., Cranston,H., Morris,D.R. and White,M.W. (1992) *J. Biol. Chem.*, **267**, 508–514.
34. Morris,D.R., Kakegawa,T., Kaspar,R.L. and White,M.W. (1993) *Biochemistry*, **32**, 2931–2937.
35. Hel,Z., Di Marco,S. and Radzioch,D. (1998) *Nucleic Acids Res.*, **26**, 2803–2812.
36. Zehner,Z.E., Shepherd,R.K., Gabryszuk,J., Fu,T.F., Al-Ali,M. and Holmes,W.M. (1997) *Nucleic Acids Res.*, **25**, 3362–3370.
37. Boado,R.J. and Pardridge,W.M. (1998) *Brain Res. Mol. Brain Res.*, **59**, 109–113.
38. Ostareck-Lederer,A., Ostareck,D., Standart,N. and Thiele,B. (1994) *EMBO J.*, **13**, 1476–1481.