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

Graziano Pesole[1,2,*], Sabino Liuni[2,3], Giorgio Grillo[4] 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., via Orabona 4, 70126 Bari, Italy and [4]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona 4, 70126 Bari, Italy

Received August 11, 1997; Revised and Accepted September 22, 1997

ABSTRACT

The important role the untranslated regions of eukaryotic mRNAs may play in gene regulation and expression is now widely acknowledged. 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 functional patterns already demonstrated by experimental analysis to have some functional role. A collection of such patterns is being collected in UTRsite database (http://bio-www.ba.cnr.it:8000/srs5/) which can also be used with appropriate computational tools to detect known functional patterns contained in mRNA untranslated regions.

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); (ii) control of mRNA stability (1,7); (iii) control of mRNA translation efficiency (8,9).

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 (10,11), 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 (12–14). 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.

We also created UTRsite, a collection of functional sequence patterns located in the 5′- or 3′-UTR sequences which could reveal 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

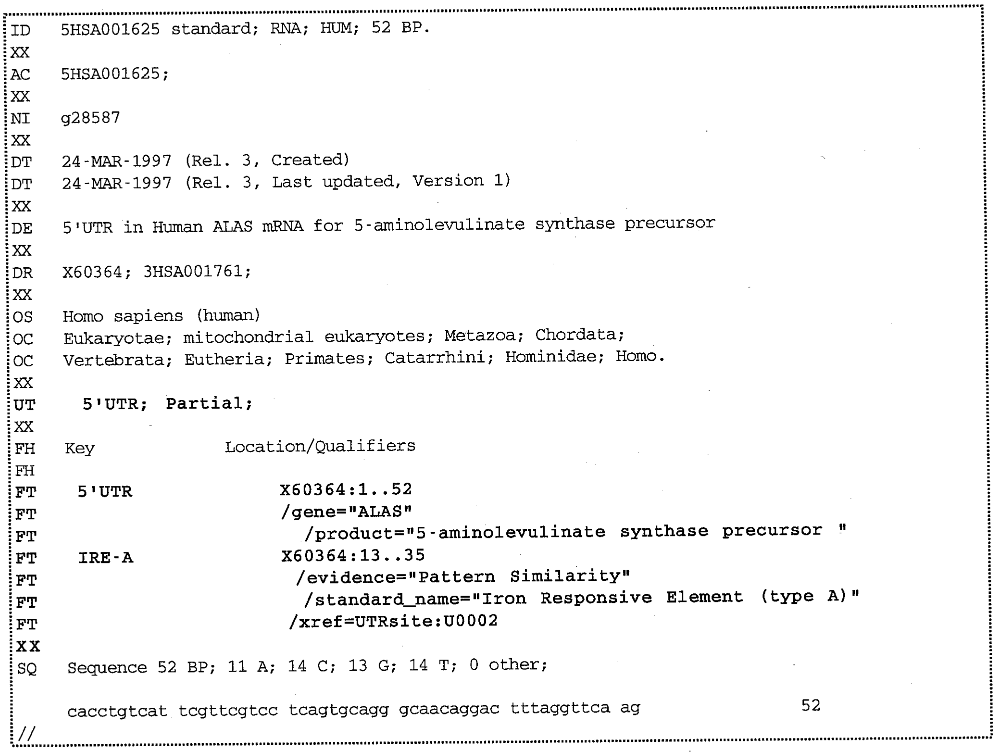
The specialized database of UTR sequences was generated by UTRdb_gen, a computer program we devised for this task. Seven sequence collections were generated for both 5′- and 3′-UTR sequences, one for each of the eukaryotic division of the EMBL/GenBank nucleotide database, namely: (i) Human; (ii) Rodent; (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’ should be present in the EMBL/GenBank entries, only a small percentage of the entries are adequately annotated. Indeed, of the 80 649 primary entries where UTRdb_gen was able to extract 5′- or 3′-UTR sequences only 13.8% contained the 5′UTR or 3′UTR feature key. UTRdb_gen is able to define UTR region boundaries even when these keys are not reported in the primary entry by using a predetermined 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 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 (15) which is able to generate automatically, in a very fast way, cleaned collections by removing entries having a similarity and overlapping degree

[*] To whom correspondence should be addressed at: Area di Ricerca del CNR, via Amendola 166/5, 70126 Bari, Italy. Tel: +39 80 5482176; Fax: +39 80 5484467; Email: graziano@area.ba.cnr.it



```
ID   5HSA001625 standard; RNA; HUM; 52 BP.
XX
AC   5HSA001625;
XX
NI   g28587
XX
DT   24-MAR-1997 (Rel. 3, Created)
DT   24-MAR-1997 (Rel. 3, Last updated, Version 1)
XX
DE   5'UTR in Human ALAS mRNA for 5-aminolevulinate synthase precursor
XX
DR   X60364; 3HSA001761;
XX
OS   Homo sapiens (human)
OC   Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
OC   Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
UT    5'UTR; Partial;
XX
FH   Key            Location/Qualifiers
FH
FT    5'UTR                X60364:1..52
FT                         /gene="ALAS"
FT                           /product="5-aminolevulinate synthase precursor "
FT    IRE-A                X60364:13..35
FT                          /evidence="Pattern Similarity"
FT                           /standard_name="Iron Responsive Element (type A)"
FT                         /xref=UTRsite:U0002
XX
SQ   Sequence 52 BP; 11 A; 14 C; 13 G; 14 T; 0 other;

     cacctgtcat tcgttcgtcc tcagtgcagg gcaacaggac tttaggttca ag              52
//
```

Figure 1. Sample entry of UTRdb. Specialized information not present in the primary EMBL/GenBank database is shown in boldcase. 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 of an 'Iron Responsive Element' (17) (UTRsite entry 'IRE-A', AC number: U0002) in this sequence entry has been also annotated.

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 and S.Brunetta) including information about the location of experimentally defined patterns collected in UTRsite. The UTRsite entries describe the various regulatory elements present in UTR regions and whose functional role has been established on 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 4.0) which in total contain about 60 000 entries and 18 500 000 nucleotides. On average, over 25% of entries are redundant and were then 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 addition site.

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

Table 2 reports functional patterns included in UTRsite (release 1.0). Many 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 experts in the field, were described by using the pattern description syntax used in PATSCAN program (kindly provided by Ross Overbeek and modified by Sandra Brunetta). The PATSCAN program is available at:
http://bio-www.ba.cnr.it:8000/BioWWW/#Patscan

AVAILABILITY OF UTRdb

UTRdb is publicly available by anonymous FTP (IP address: ftp.ba.cnr.it; dir: /pub/embnet/database/utr) and UTRdb entries can be retrieved on the Web by using SRS (16) at the EBI WWW server (http://srs.ebi.ac.uk:5000/) or at the BioWWW server (http://bio-www.ba.cnr.it:8000/srs5). SRS retrieval allows linking between UTRdb, EMBL/GenBank and UTRsite entries. UTR-scan, a computer program searching patterns included in UTRsite will be soon made available on the Web.

CONCLUSIONS AND PERSPECTIVES

The important role that untranslated regions 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 untranslated regions are involved in crucial biological functions.

Table 1. Number of entries (N) and nucleotide length (L) of UTRdb collections (release 4.0) after redundancy cleaning

	N	L	Redundancy %N	%L
5′-UTR				
Human	6220	1 084 947	33.8	21.3
Other_mammal	2195	247 933	25.1	14.5
Rodent	6982	1 182 728	28.1	16.4
Other_vertebrate	2693	359 692	19.5	14.6
Invertebrate	3916	697 168	19.2	12.3
Plant	5409	525 766	17.9	9.9
Fungi	995	161 197	11.8	6.6
TOTAL	28 410	4 259 431		
3′-UTR				
Human	6600	4 350 835	35.8	24.5
Other_mammal	2353	1 024 534	32.8	23.2
Rodent	7027	3 990 573	31.9	21.7
Other_vertebrate	3099	1 435 895	19.9	13.1
Invertebrate	4513	1 662 453	18.4	15.3
Plant	6757	1 578 937	14.9	12.8
Fungi	1112	257 566	10.3	7.7
TOTAL	31 461	14 300 793		

UTRdb 4.0 was generated from EMBL release 51. Relevant redundancy percentages calculated with respect to the number of entries (%N) and to the nucleotide length (%L) were also indicated.

Table 2. Functional patterns included so far in UTRsite (v1.0). For each pattern the number of hits with UTRdb entries is also reported

Functional patterns	Reference	Hits found in UTRdb 4.0
Iron-responsive element	(17)	73
Histone 3′-UTR stem–loop structure	(18)	37
AUUUA destabilising element	(19)	129
tra-2 sex determining repeats	(20)	1
Selenocysteine insertion sequence	(21,22)	80
APP 3′-UTR stability control element	(23)	7

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 however necessary to avoid artefacts caused by redundant sequences. Even if statistical significance does not necessarily mean biological significance, it may provide useful indication for further experimental work, such as site-directed mutagenesis.

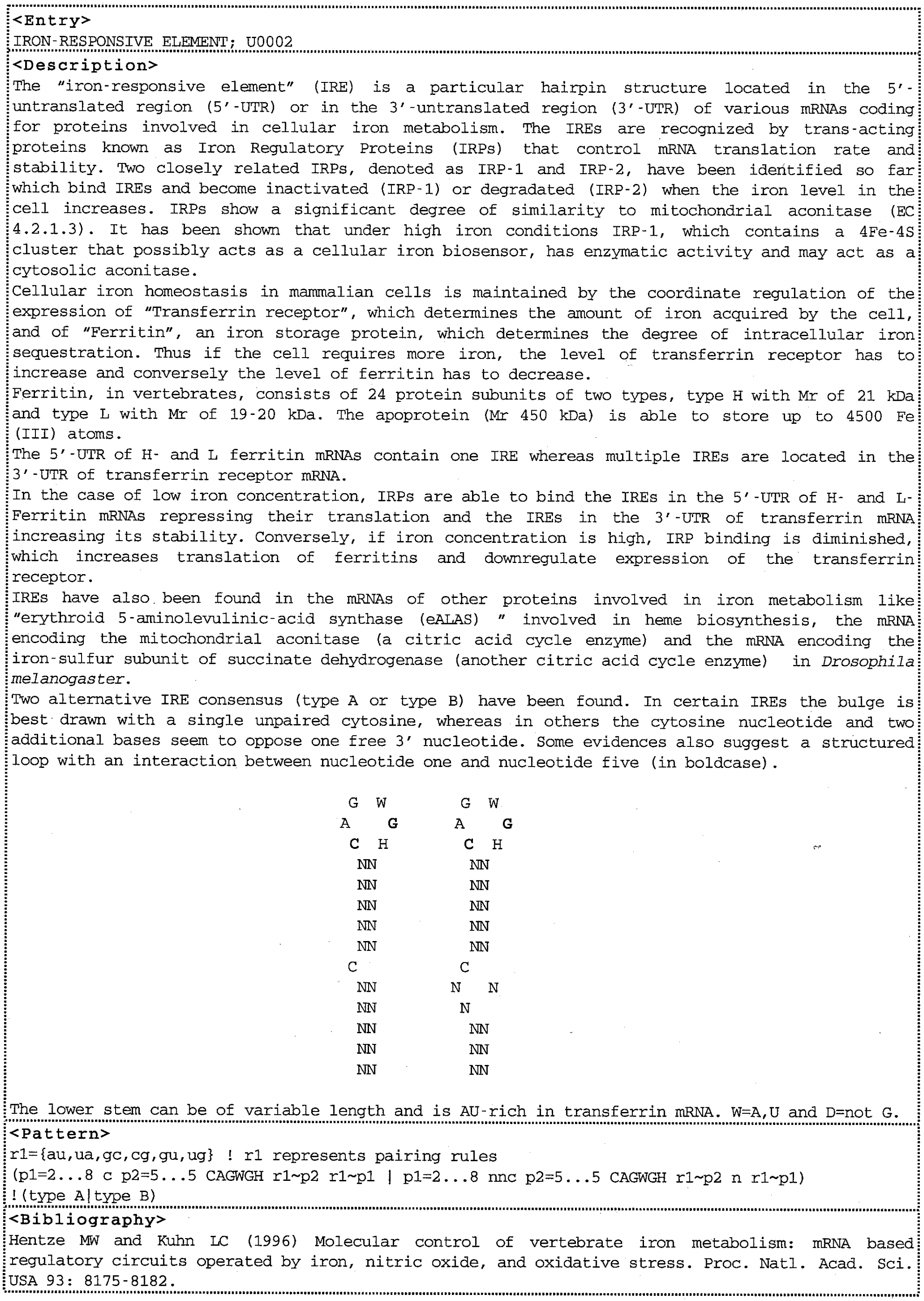
UTRdb will be updated with the new EMBL database releases and UTRsite will be continuously updated by adding new entries describing functional patterns whose biological role has been experimentally demonstrated.

ACKNOWLEDGEMENTS

This work was supported by EU grant BIO4-CT95-0130. We would like to thank Sandra Brunetta for providing some of the computer programs, Matteo diTommaso and Rodrigo Lopez for assistance in the construction of the Web server at the EBI. 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 Bill Marzluff (Histone stem–loop structure).

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 Beelman,C.A. and Parker,R. (1995) *Cell*, **81**, 179–183.
8 Curtis,D., Lehman,R. and Zamore,P.D. (1995) *Cell*, **81**, 171–178.
9 Sonenberg,N. (1994) *Curr. Opin. Gen. Dev.*, **4**, 310–315.
10 Mengeritsky,G. and Smith,T.F. (1987) *Comput Appl. Biosci.*, **3**, 223–227.
11 Konopka,A.K. (1994) In Smith,D.W. (ed.), *Informatics and Genome Projects*. Academic Press, San Diego.
12 Pesole,G., Liuni,S., Grillo,G. and Saccone,C. (1997) *Gene*, in press.
13 Pesole,G., Grillo,G. and Liuni,S. (1996) *Comput. Chem.*, **20**, 141–144.
14 Pesole,G., Fiormarino,G. and Saccone,C. (1994) *Gene*, **140**, 219–225.
15 Grillo,G., Attimonelli,M., Liuni,S. and Pesole,G. (1996) *Comput. Appl. Biosci.*, **12**, 1–8.
16 Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci.*, **9**, 49–57.
17 Hentze,M.W. and Kuhn,L.C. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 8175–8182.
18 Williams,A.S. and Marzluff,W.F. (1995) *Nucleic Acids Res.*, **23**, 654–662.
19 Bohjanen,P.R., Petryniak,B., June,C.H., Thompson,C.B. and Lindsten,T. (1992) *J. Biol. Chem.*, **267**, 6302–6309.
20 Goodwin,E.B., Okkema,P.G., Evans,T.C. and Kimble,J. (1993) *Cell*, **75**, 329–339.
21 Hubert,N., Walczak,R., Sturchler,C., Schuster,C., Westhof,E., Carbon,P. and Krol,A. (1996) *Biochimie*, **78**, 590–596.
22 Walczak,R., Westhof,E., Carbon,P. and Krol,A. (1996) *RNA*, **2**, 367–379.
23 Zaidi,S.H.E. and Malter,J.S. (1994) *J. Biol. Chem.*, **269**, 24007–24010.



<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)' (17). 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/#Patscan).