FEBS Letters 464 (1999) 60–62 FEBS 23104

Isochore specificity of AUG initiator context of human genes

Graziano Pesole[a,]*, Giorgio Bernardi[b], Cecilia Saccone[c]

[a] *Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26, 20133 Milan, Italy*
[b] *Stazione Zoologica 'Anton Dohrn', Villa Comunale 1, 80121 Naples, Italy*
[c] *Dipartimento di Biochimica e Biologia Molecolare, Università di Bari e Centro di Studio sui Mitocondri e Metabolismo Energetico, C.N.R., via Orabona 4, 70126 Bari, Italy*

Received 4 October 1999; received in revised form 26 November 1999

Edited by Takashi Gojobori

Abstract The efficiency of AUG start codon recognition in translation initiation is modulated by its sequence context. Here we investigated a non-redundant set of 5914 human genes and show that this context is different in genes located in different isochores. In particular, of the two main consensus start sequences, RCCaugR is five-fold more represented than AARaugR in genes from the GC-rich H3 isochores compared to genes from the GC-poor L isochores. Furthermore, genes located in GC-rich isochores have shorter 5′ UTRs and stronger avoidance of upstream AUG than genes located in GC-poor isochores. This suggests that genes requiring highly efficient translation are located in GC-rich isochores and genes requiring fine modulation of expression are located in GC-poor isochores. This is in agreement with independent data from the literature concerning the location of housekeeping and tissue-specific genes, respectively.

© 1999 Federation of European Biochemical Societies.

Key words: Translation initiation; mRNA 5′ untranslated region; GC content; Upstream open reading frame; Statistical analysis

1. Introduction

The efficiency of AUG start codon recognition in translation initiation is modulated by its sequence context [1–4] and by other features of the 5′ untranslated (UTR) leader such as length, presence of stable secondary structures and of upstream open reading frames (ORFs) [5]. Previous analyses have shown that the most frequent, and therefore regarded as optimal, start codon context in vertebrate genes is (G/A)-CC<u>AUG</u>G [6,7]. Start codons deviating from the optimal context at one or more of the crucial positions, mainly A/G^{-3} and G^{+4} (numbering AUG +1 to +3), may be recognized less efficiently and allow for a translation regulated by the cellular conditions and requirements. Furthermore, the features of the 5′ leader sequence may modulate translation factor requirements. Indeed, in regulated genes, maximal levels of expression are not required and downstream translation has been observed in mammalian, plant and yeast cells to be strongly influenced by physiological conditions or tissue-specific effects [5]. We report here an extensive survey of codon context and 5′ UTR leader features by analyzing a non-redundant set of 5914 human genes from UTRdb (release 10.0 [8]).

Following the finding that the human genome (as well as that of warm-blooded vertebrates in general) is a mosaic of isochores, i.e. long ($>$ 300 kb) DNA segments homogeneous in base composition (see [9] for a review), we investigated the features of 5′ UTRs and of the AUG start codon context considering separately the genes belonging to the light GC-poor isochores, i.e. L1 and L2 accounting for 63% of the genome, and to the heavy GC-rich isochores, H1, H2 and H3, accounting for 24%, 7.5% and 4.7% of the genome respectively. The linear correlation existing between the GC level of gene third codon positions (GC3) and that of the corresponding isochore [10] allowed the isochore assignment of all genes considered.

2. Materials and methods

The UTRdb specialized database [8] was used as the source of sequence data for the present study. It contains non-redundant collections of 5′ and 3′ untranslated sequences of eukaryotic mRNAs (5′ UTR, 3′ UTR). The computer program UTRstat (kindly provided by Giorgio Grillo) was used to selectively extract from UTRdb the 5′ UTR sequences of genes located in the four different isochore classes considered here, namely L (i.e. L1+L2), H1, H2 and H3. An option of UTRstat has been used to calculate the average sequence length of 5′ UTRs annotated as complete regions in UTRdb being derived from genomic sequence entries clearly indicating mRNA start sites in the Feature Table. The isochore assignment of human genes, based on the GC content observed at the third codon position (GC3) of the coding region, was made as follows according to Zoubak et al. [10]: isochore L, GC3 $<$ 47%; isochore H1, 47% $\leq$ GC3 $<$ 61%; isochore H2, 61% $\leq$ GC3 $<$ 74%; isochore H3, GC3 $\geq$ 74%.

The occurrence frequency of AUG located upstream of the initiator codon in the 5′ UTRs (upstream AUG or uAUG) was calculated using the computer program Findpatterns [11] on the four previously generated isochore-specific sequence datasets. The uAUG Obs/Exp value was calculated assuming a zero-order Markov chain and calculating the four nucleotide frequencies over 40 nt long sequences spanning from position −20 to +20 with respect to the start codon.

The computer program AUGscan (kindly provided by Giorgio Grillo) was used to calculate the occurrence of the various oligonucleotide initiator contexts in the four isochore-specific gene collections above defined.

3. Results and discussion

Table 1 shows the 10 most represented heptanucleotides, including three nucleotides upstream and one downstream of the initiator AUG, for the genes belonging to the different isochore compartments. Two main consensus start codon contexts were observed: AARaugR and RCCaugR (R = purine), mainly differing at positions −1 and −2, with other optimal initiator codons more similar to either of the two. Furthermore, the relative abundance of these two initiator contexts

*Corresponding author. Fax: (39)-2-70632811.
E-mail: graziano.pesole@unimi.it

0014-5793/99/$20.00 © 1999 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(99)01675-0

Table 1
The 10 most represented heptamers spanning from position −3 to +4 with respect to the initiator AUG observed in genes belonging to different isochore compartments as inferred from their GC3 content

Initiator AUG context	Count	%
Isochore L (1626 genes)		
AAGAUGG	84	5.17
GCCAUGG	64	3.94
ACCAUGG	53	3.26
AAAAUGG	47	2.89
AAAAUGA	47	2.89
AAGAUGA	41	2.52
ACCAUGA	31	1.91
AACAUGG	31	1.91
ATCAUGG	30	1.85
AACAUGA	29	1.78
Isochore H1 (1443 genes)		
GCCAUGG	82	5.68
ACCAUGG	76	5.27
AAGAUGG	62	4.30
GCCAUGA	37	2.56
AAAAUGG	35	2.43
GAGAUGG	32	2.22
ACCAUGT	30	2.08
AGCAUGG	27	1.87
AAGAUGA	26	1.80
ATCAUGG	25	1.73
Isochore H2 (1342 genes)		
GCCAUGG	102	7.60
ACCAUGG	63	4.69
ACCAUGA	42	3.13
AAGAUGG	38	2.83
AGGAUGG	38	2.83
GCCAUGA	37	2.76
GCCAUGC	36	2.68
GCCAUGT	36	2.68
AAGAUGA	27	2.01
GACAUGG	26	1.94
Isochore H3 (1503 genes)		
GCCAUGG	177	11.78
ACCAUGG	76	5.06
GCCAUGA	47	3.13
ACCAUGA	44	2.93
GCCAUGC	42	2.79
AGGAUGG	41	2.73
AGCAUGG	39	2.59
CCCAUGG	36	2.40
AAGAUGG	33	2.20
GGCAUGG	31	2.06

The isochore assignment of human genes was made as described in Section 2.

was very different in the different isochores with RCCaugR less represented than AARaugR in the L isochore (i.e. 171 vs. 234 occurrences respectively) but five-fold more abundant in the H3 isochore (i.e. 347 vs. 69 occurrences respectively). We also observed a much stronger context preference for genes belonging to GC-rich isochores (data not shown).

According to the scanning mechanism proposed by Kozak [7] the translation of eukaryotic mRNAs in most cases starts at the AUG nearest to its 5′ end. A survey of 1083 complete human 5′ UTRs showed that adherence to the first AUG rule was observed in only 55% of the genes but again striking differences were observed for genes belonging to different isochores, with upstream AUG (uAUG) more abundant in genes located in GC-poor than in GC-rich isochores whose 5′ UTR showed a stronger avoidance, as inferred from the Obs/Exp value of the trinucleotide AUG (see Table 2).

Isochore-specific features were also observed calculating the average length of 5′ UTRs which were inversely correlated to the GC richness of the relevant isochore (Table 2). Even if the length of the 5′ UTR does not in itself downregulate translation, long 5′ UTRs are more likely to contain *cis*-acting oligonucleotides or secondary structures reducing translation efficiency [12].

Considering that mRNA with long 5′ UTRs, uAUG and initiator codons with suboptimal context are translated less efficiently, our results suggest that genes requiring highly efficient translation should be mostly located in GC-rich isochores, whereas genes requiring fine modulation of expression should be predominantly located in GC-poor isochores. These indications are in agreement with independent data indicating a preferential location of housekeeping and tissue-specific genes in GC-rich and GC-poor isochores, respectively [9,13]. Moreover, an analysis of Tables 1 and 2 stresses the stronger similarities of the features observed in isochores L/H1, on the one hand, and H2/H3, on the other. This is in agreement with other independent observations pointing in the same direction (see [14] for a review).

Our data generally suggest different functional features for genes belonging to different isochore compartments which thus should be considered separately for any kind of analysis. Interestingly, whereas the most represented initiator context in mammalian genes, as well as in warm-blooded vertebrate genes in general, is GCCaugG, in cold-blooded vertebrates it is AAAaugG, very similar to that observed in GC-poor human isochores, as expected for genomes in which GC-rich isochores are almost absent.

Acknowledgements: This work was supported by Training and Mobility Research European Project ERB-FMRX-CT98-0221, EU Grant ERB-BIO4-CT96-0030 and Programma Biotecnologie legge 95/95 (MURST 5%). We are grateful to Giorgio Grillo for providing computer programs to carry out some of the analyses.

References

[1] Kozak, M. (1991) J. Biol. Chem. 266, 19867–19870.
[2] Kozak, M. (1992) Annu. Rev. Cell Biol. 8, 197–225.
[3] Kozak, M. (1995) Proc. Natl. Acad. Sci. USA 92, 2662–2666.

Table 2
Features of genes belonging to different isochores

Isochore	Genes adhering to the first AUG rule (%)	Average 5′ UTR length (nt)	AUG Obs/Exp in 5′ UTRs
L (133 genes)	41.2	287.3	0.85
H1 (388 genes)	46.7	212.9	0.67
H2 (309 genes)	61.4	180.0	0.54
H3 (153 genes)	72.2	151.7	0.52

The AUG Obs/Exp value was calculated assuming a zero-order Markov chain and calculating the four nucleotide frequencies on 40 nt long sequences spanning from position −20 to +20 with respect to the start codon. The average length was calculated on complete 5′ UTRs from genomic DNA entries where the coordinates of the mRNA start were clearly indicated in the Feature Table of the relevant EMBL database entries.

[4] Kozak, M. (1996) Mamm. Genome 7, 563–574.
[5] Gray, N.K. and Wickens, M. (1998) Annu. Rev. Cell Dev. Biol. 14, 399–458.
[6] Kozak, M. (1987) Nucleic Acids Res. 15, 8125–8148.
[7] Kozak, M. (1999) Gene 234, 187–208.
[8] Pesole, G., Liuni, S., Grillo, G., Ippedico, M., Larizza, A., Makalowski, W. and Saccone, C. (1999) Nucleic Acids Res. 27, 188–191.
[9] Bernardi, G. (1995) Annu. Rev. Genet. 29, 445–476.
[10] Zoubak, S., Clay, O. and Bernardi, G. (1996) Gene 174, 95–102.
[11] GCG (1994) Genetic Computer Group (GCG), 575 Science Drive, Madison, WI 53711, USA.
[12] Gallie, D.R. (1996) Plant Mol. Biol. 32, 145–158.
[13] Bickmore, W. and Craig, J. (1997) Chromosome Bands: Patterns in the Genome, R.G. Landes Company, Austin, TX.
[14] Bernardi, G. (2000) Gene (in press).