Gene 507 (2012) 106–111

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Characterization and evolution of 5′ and 3′ untranslated regions in eukaryotes

Honglei Liu [a,1], Jiaming Yin [a,1], Meili Xiao [a], Caihua Gao [a], Annaliese S. Mason [b], Zunkang Zhao [c], Yingchun Liu [a], Jiana Li [a], Donghui Fu [d,*]

[a] Engineering Research Center of South Upland Agriculture of Ministry of Education, P.R. China, College of Agronomy and Biotechnology, Southwest University, Chongqing, China
[b] Centre for Integrative Legume Research and School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4072, Australia
[c] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
[d] Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Jiangxi Agricultural University, Nanchang 330045, China

ARTICLE INFO

Article history:
Accepted 18 July 2012
Available online 27 July 2012

Keywords:
Repetitive sequences
Genome size
Species gene number

ABSTRACT

Untranslated regions (UTRs) in eukaryotes play a significant role in the regulation of translation and mRNA half-life, as well as interacting with specific RNA-binding proteins. However, UTRs receive less attention than more crucial elements such as genes, and the basic structural and evolutionary characteristics of UTRs of different species, and the relationship between these UTRs and the genome size and species gene number is not well understood. To address these questions, we performed a comparative analysis of 5′ and 3′ untranslated regions of different species by analyzing the basic characteristics of 244,976 UTRs from three eukaryote kingdoms (Plantae, Fungi, and Protista). The results showed that the UTR lengths and SSR frequencies in UTRs increased significantly with increasing species gene number while the length and G + C content in 5′ UTRs and different types of repetitive sequences in 3′ UTRs increased with the increase of genome size. We also found that the sequence length of 5′ UTRs was significantly positively correlated with the presence of transposons and SSRs while the sequence length of 3′ UTRs was significantly positively correlated with the presence of tandem repeat sequences. These results suggested that evolution of species complexity from lower organisms to higher organisms is accompanied by an increase in the regulatory complexity of UTRs, mediated by increasing UTR length, increasing G + C content of 5′ UTRs, and insertion and expansion of repetitive sequences.

1. Introduction

Intact messenger RNA (mRNA) in eukaryotes consists of a number of elements: a 7-methylguanosine cap which exists in the 5′ terminal, a 5′ untranslated region (5′ UTR), a coding region, a 3′ untranslated region (3′ UTR), and a 3′ poly(A) tail. The 5′ UTR, as the leading mRNA element, begins at the transcription start site and ends just before the coding region initiator. The 3′ UTR starts at the end of the coding region, and ends just before the poly(A) tail. Untranslated regions (UTRs) in eukaryotes aid in regulating translation and mRNA half-life, interacting with species RNA-binding proteins. The 5′ UTR may have 100 or more nucleotides, and the 3′ UTR may be even longer, up to several kilobases in size. 3′ UTRs are generally twice as long as 5′ UTRs (Pesole et al., 2001). The lengths of both 5′ UTRs (Kochetov et al., 1998) and 3′ UTRs (Sandberg et al., 2008) have known correlations to gene expression level. However, little is known about the relationship between evolution of organism complexity and the structural characteristics of UTRs, including UTR length, UTR G + C content and the presence of repetitive sequences in UTRs.

UTR conformations range from simple primary structures to highly ordered secondary structures which are involved in the regulation of gene expression. 5′ UTRs can contain binding sites for signaling proteins which affect transcription stability or translation efficiency of mRNA (Avramovich Tirosh et al., 2008; Zimmer et al., 2008). In addition, some proteins can bind with 3′ UTRs to affect mRNA expression (Zearfoss et al., 2011; Zeng et al., 2009). 3′ UTRs are also used as miRNA target sites to downregulate target mRNA (Ott et al., 2011; Wynendaele et al., 2010), and for the specific subcellular targeting of transcripts (Andreassi and Riccio, 2009).

Repetitive sequences such as simple sequence repeats (SSRs) are often located in UTRs, and hence participate in UTR variation and gene regulation. The rapid variation of SSRs through elongation and contraction of repeats due to slipped strand mispairing (slippage) during DNA replication or through unequal crossing over during recombination (Chambers and MacAvoy, 2000; Ellegren, 2004; Vergnaud and Denoeud, 2000) is an important force for molecular evolution, and may also lead to an increase of biological complexity (Borstnik and Pumpernik, 2002). More SSRs are located in UTRs

Abbreviations: UTR, untranslated region; mRNA, messenger RNA; 5′ UTR, 5′ untranslated region; 3′ UTR, 3′ untranslated region; miRNA, MicroRNA; SSR, simple sequence repeat; SINE, short interspersed element; LINE, long interspersed element; LTR, long terminal repeat; ORF, open reading frame; IRES, internal ribosome entry site.

* Corresponding author. Tel.: +86 791 83813142; fax: +86 791 83813185.
E-mail address: fudhui@163.com (D. Fu).

[1] These authors contributed equally.

0378-1119/$ – see front matter.
http://dx.doi.org/10.1016/j.gene.2012.07.034

than in other transcribed regions (Morgante et al., 2002; Wren et al., 2000), where they often serve as protein binding sites to regulate gene translation (Calkhoven et al., 1994).

Increasing organismal complexity has been proposed to be reflected by larger genome size, increases in species gene/protein number and in numbers of noncoding regulatory elements, and by the growth of biological networks (Barrett et al., 2012; Chen et al., 2011b; Lynch and Conery, 2003). Whether UTR frequency and characteristics are related to genome size and species gene number has not been well investigated. Hence, we downloaded the known eukaryotic UTR sequences (244,976 in total from kingdoms Plantae, Fungi, and Protista) to determine the following characteristics: a) typical lengths; b) G + C content; and c) the composition of their different repetitive sequences, specifically simple sequence repeats (SSRs), short interspersed elements (SINES), long interspersed elements (LINES), and long terminal repeats (LTRs). We then investigated the relationship between the basic characteristics of the UTRs and organismal complexity across a wide range of phyla in order to shed light on the evolution of UTRs.

2. Materials and methods

2.1. Materials

5′ UTR and 3′ UTR sequences were downloaded from the UTRscan site (http://utrdb.ba.itb.cnr.it/home/download). These sequences were classified into seven categories: Protista (1 species), basidiomycetes (2 species), ascomycetes (12 species), green algae (2 kinds), mosses, monocotyledonous plants (2 kinds), and dicotyledonous plants (3 kinds).

2.2. Methods

2.2.1. Length and G + C content of UTRs

UTR length and G + C content for 5′ UTRs and 3′ UTRs in each category were collected and recorded using Microsoft Office Excel 2003. BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) (Tippmann, 2004) was used to analyze the nucleotide composition of the UTRs.

2.2.2. SSRs in the UTRs

SSR Locator software (da Maia et al., 2008) was used to mine SSRs with mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octa-, nova-, or decanucleotide motifs, which contained a minimum of 10, 8, 6, 5, 4, 4, 3, 3, 3, and 3 repeats, respectively. Hence, only SSRs with a total length ≥10 bp were identified.

2.2.3. UTR transposons and tandem repeat sequences

The LTR-finder program (http://tlife.fudan.edu.cn) was used to predict intact LTR retrotransposons. The parameters for minimum LTR length and minimum distance between LTRs were 50 bp and 100 bp respectively. The Eukarya tRNA database was used as the reference database, and the other parameters were set as defaults. The different transposon types were predicted using RepeatMasker (http://www.repeatmasker.org/) (Smith et al., 2007). Results from the LTR-Finder were used to substitute results from RepeatMasker as some of the LTR retrotransposons predicted by RepeatMasker were not intact. Tandem repeats (motif length > 10 bp) were predicted using Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) with default parameters.

2.2.4. Correlation analysis among the frequency and characteristics of UTR, genome size and species gene number

We performed correlation analysis among length of UTRs, G + C content of UTRs, the frequency of SSRs, UTR transposons and tandem repeat sequences, genome size, and species gene number using non-parametric paired data for multiple tests (120 pairwise comparisons) by SAS 9.0 software (Institute, 1996). The data of genome size, and species gene number were collected from the public database or published references.

3. Results

3.1. Sequence downloads and the classification of UTRs

A total of 244,976 UTR sequences were downloaded from the UTRscan database (http://utrdb.ba.itb.cnr.it/home/download). These sequences contained 113,901 5′ UTR and 131,075 3′ UTR sequences from 23 species in the kingdoms Plantae, Fungi and Protista. The collected UTR sequences were split up across three categories: the Plantae (5′ UTR vs 3′ UTR = 95,697 vs 105,312), Fungi (16,498 vs 22,315), and Protista kingdoms (1706 vs 3448), and in seven subcategories: diatoms (1706 vs 3448), basidiomycetes (6114 vs 7705), ascomycetes (10,384 vs 14,610), green algae (6605 vs 6242), moss (5612 vs 7152), monocots (33,545 vs 36,655), and dicots (49,935 vs 55,263).

3.2. UTR length and its relation with genome size and species gene number

UTR length determines gene translation efficiency and transcription stability. 5′ UTR sequence length (22 species) ranged from 1 bp to 5550 bp, with an average of 164 bp, whereas 3′ UTR sequence length (21 species) ranged from 1 bp to 12,060 bp with an average of 287 bp. 5′ UTRs and 3′ UTRs were classified into 11 groups based on length, in 100 bp intervals. The majority (92.7%) of 5′ UTR sequences ranged from 1 bp to 400 bp, and 5′ UTR sequences with lengths less than 100 bp accounted for almost half of the 5′ UTR sequences (46.6%). The majority of the 3′ UTR sequences (93.1%) were between 1 bp and 600 bp, and more than half (53.5%) of the 3′ UTR sequences ranged from 100 bp to 300 bp. On average, the length of the 3′ UTR sequences was nearly twice that of 5′ UTR sequences, with a greater mode, maximum range and proportion of sequences in larger size categories for the 3′ UTRs.

A significant positive correlation ($r = 0.59$, $P < 0.0001$) (Table 1) was detected between the sequence lengths of 5′ UTRs and 3′ UTRs. In addition, a significant correlation was detected between genome size and average 5′ UTR length ($r = 0.31$, $P = 0.0005$), but not between genome size and average 3′ UTR length ($r = 0.15$, $P = 0.1038$). The correlations between species gene number and UTR length were also positive and significant ($r = 0.33$, $P = 0.0002$ for 5′ UTRs; $r = 0.27$, $P = 0.0025$ for 3′ UTRs).

3.3. Distribution of UTR G + C content and its relationship with UTR length, genome size and species gene number

The G + C content of the 5′ UTRs ranged from 0% to 100%, with an average of 48.3%. The G + C content of the 3′ UTRs also ranged from 0% to 100%, but with an average of 38.7%. Most 5′ UTR sequences (59.3%) had a G + C content lower than 50%, but almost all 3′ UTRs (91.9%) had a G + C content lower than 50%. Nine-tenths of the 5′ UTRs had a G + C content between 30% and 70%, whereas nine-tenths of the 3′ UTR had a G + C content between 20% and 50%, confirming that the 3′ UTRs are AU-rich elements (Fig. 1).

A correlation analysis (Table 1) was performed to determine the relationship between genome size, species gene number and G + C content of the UTRs. There was a significant positive correlation ($r = 0.65$, $P < 0.0001$) between the average G + C contents of 5′ UTRs and 3′ UTRs. A significant positive correlation was detected between G + C content and sequence length of 3′ UTRs ($r = 0.53$, $P < 0.0001$), but not between the G + C content and sequence length of 5′ UTRs ($r = 0.13$, $P = 0.1498$, not significant). Genome size and G + C content of 5′ UTRs were also correlated ($r = 0.26$, $P = 0.0035$), but genome size and G + C content of 3′ UTRs and species gene number and G + C content of UTRs were not significantly associated.