Journal of Theoretical Biology 404 (2016) 303–311

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Comparative analysis of contextual bias around the translation initiation sites in plant genomes

Paras Gupta, Latha Rangan*, T. Venkata Ramesh, Mudit Gupta

Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam 781039, India

HIGHLIGHTS

- The current study builds up on earlier findings that TIS is an important determinant of translation efficiency.
- The study was extended to plants by using a much more representative dataset of seven monocots and seven dicots species.
- Nucleotide bias around TIS directly correlates with protein abundance.
- Extensive three base periodicity around TIS observed.
- Most representative TIS sequences for each taxon also identified.

ARTICLE INFO

Article history:
Received 1 February 2016
Received in revised form
17 May 2016
Accepted 10 June 2016
Available online 15 June 2016

Keywords:
Consensus sequence
Dicots
Monocots
Kozak sequence
Translation initiation site

ABSTRACT

Nucleotide distribution around translation initiation site (TIS) is thought to play an important role in determining translation efficiency. Kozak in vertebrates and later Joshi et al. in plants identified context sequence having a key role in translation efficiency, but a great variation regarding this context sequence has been observed among different taxa. The present study aims to refine the context sequence around initiation codon in plants and addresses the sampling error problem by using complete genomes of 7 monocots and 7 dicots separately. Besides positions −3 and +4, significant conservation at −2 and +5 positions was also found and nucleotide bias at the latter two positions was shown to directly influence translation efficiency in the taxon studied. About 1.8% (monocots) and 2.4% (dicots) of the total sequences fit the context sequence from positions −3 to +5, which might be indicative of lower number of housekeeping genes in the transcriptome. A three base periodicity was observed in 5′ UTR and CDS of monocots and only in CDS of dicots as confirmed against random occurrence and annotation errors. Deterministic enrichment of GCN**AUG**GC in monocots, AAN**AUG**GC in dicots and GCN**AUG**GC in plants around TIS was also established (where **AUG** denotes the start codon), which can serve as an arbiter of putative TIS with efficient translation in plants.

1. Introduction

Proteins are synthesized from mRNAs in a process called translation. Translation, as a process, follows expression of messenger RNAs (mRNA) in the form of proteins and the amount of energy expended in this process is a testament to its fundamental importance (Schwanhäusser et al., 2011). The process can be divided into three distinct stages: initiation, elongation of the polypeptide chain, and termination. The region at which translation initiates (that is the initiating aminoacyl-tRNA pairs with the start codon AUG) signals the beginning of polypeptide synthesis and is called the translation initiation site (TIS). Control of gene expression at the principal level is mediated by translation initiation regulation, for which sequence context around the start codon plays a major role in identification of TIS and subsequent enhancement of translation efficiency. It is known that recognition of the translation start codon by eukaryotic ribosomes depends upon its sequence context. Examination of context sequences can inform us about key features specifying TIS. The most widely accepted model for translation initiation is the scanning model (Kozak, 1978). According to the scanning model, the 40S ribosomal subunit, along with a number of initiation factors, attaches to the 7-methyl guanosine cap at the 5′ end of mRNA and begins scanning along the mRNA in a linear fashion till it encounters an AUG codon. While translation initiation from the first AUG holds true in many cases, there are also a considerable number of exceptions (Kozak, 2002; Pelletier and Sonenberg, 1988). In these exceptions, the main determining factor in AUG choice is the context of the

* Corresponding author.
E-mail address: latha_rangan@yahoo.com (L. Rangan).

http://dx.doi.org/10.1016/j.jtbi.2016.06.015
0022-5193/

respective codon i.e. the sequence of nucleotides surrounding the putative TIS. To this effect, from 699 vertebrate mRNAs, a context sequence was derived (gccgcC(A/G)cc**AUG**G) where **AUG** represents the start codon) known as Kozak's consensus (Kozak, 1987). By mutational studies, the presence of a purine at −3 position and guanine at +4 has been deemed as crucial for optimum translation efficiency (Kozak, 1986, 1997), although, in yeast, nucleotide change at −3 position showed no marked effect on translation efficiency (Mark Cigan and Donahue, 1987; Yun et al., 1996).

A decade later, this context sequence was extended to plants by Joshi et al. (1997) who derived the context sequence for monocots (1127 genes) as GCGGC(A/C)(A/G)(A/C)C**AUG**GCG and for dicots (3643 genes) as AAAAAAA(A/C)A**AUG**GCU respectively (Joshi et al., 1997). Various other researchers further extended this context sequence determination to other taxonomic groups as and when more genomic sequences became available (Cavener and Ray, 1991). Thus, a huge inter-taxon variation has been observed, contrary to the initial belief that the Kozak's consensus is universal and the same for all eukaryotes (Kozak, 1984).

Consensus sequence for the context of the AUG codon in higher plants was proposed on the basis of very limited number of sequences (Joshi et al., 1997) and extended to two model crops with the availability of complete genome sequences (Rangan et al., 2008). Moreover, a three base periodicity around TIS has also been observed with potential role in translation (Nakagawa et al., 2008). With the current boom in the availability of genomic data, it is of practical use to further refine the context sequence around initiation codon in plants and verify the context rules derived by Joshi et al. and Kozak, which in both the cases have been marred by a small number of datasets and lower statistical confidence. We mitigate the latter two issues by using complete genomes of 14 plant species and G statistic (Sokal and Rohlf, 1995), which helps to account for the compositional base bias of a species while determining nucleotide conservation.

In this comparative study, nucleotide bias around TIS is investigated across seven monocots and seven dicots respectively. A strong positive inter-relationship was found between highly conserved positions and protein abundance and a prevalence of repeats of three nucleotides around TIS was observed. Sequences GCN**AUG**GC, AAN**AUG**GC and GCN**AUG**GC were identified as the archetypal TIS signal in monocots, dicots and plants respectively.

2. Materials and methods

2.1. Data collection and filtering

Coding sequences (CDS) and mRNAs were downloaded from different databases as listed in Table 1. Throughout the study, the words *monocots* and *dicots* imply the seven species for each taxon respectively while *plants* refer to the combined fourteen species. Only those species were considered whose sequence data was freely available and without any embargo over first use in a publication. Initially, only genes bearing a corresponding CDS were included in the analysis. The final CDSs and mRNAs used for analysis were obtained by filtering at four levels as follows: redundant sequences were first removed followed by selection of sequences having the canonical start codon (AUG) and a stop codon (UAA/UAG/UGA). If alternative splicing information was available, for such variants only one representative sequence bearing the longest CDS was selected at stage three and lastly, only those sequences were considered which had a 23 bp window constitutive of 10 nucleotides upstream and 10 nucleotides downstream of AUG.

Table 1
List of species used in the analysis.

S no.	Crop	Generic name	Database
A.	**Monocots**		
1.	Sorghum	*Sorghum bicolor*	Phytozome
2.	Maize	*Zea mays*	Phytozome
3.	Rice	*Oryza sativa*	Phytozome
4.	Purple false brome	*Brachypodium distachyon*	Phytozome
5	Wheat	*Triticum aestivum*	Genbank
6.	Barley	*Hordeum vulgare*	Genbank
7.	Banana	*Musa acuminate*	Ensembl Biomarts
B.	**Dicots**		
1.	Arabidopsis	*Arabidopsis thaliana*	TAIR 10
2.	Soyabean	*Glycine max*	Phytozome
3.	Potato	*Solanum tuberosum*	Phytozome
4.	Grapevine	*Vitis vinifera*	Phytozome
5.	Barrel clover	*Medicago truncatula*	Phytozome
6.	Black cottonwood	*Populus trichocarpa*	Phytozome
7.	Tomato	*Solanum lycopersicum*	Ensembl Biomarts

The websites for the following databases are as follows. *Phytozome*, (http://www.phytozome.net/); *Genbank*, (http://www.ncbi.nlm.nih.gov/), the terms '*Triticum aestivum*' and '*Hordeum vulgare* sub sp. *vulgare*' in the 'organism' field along with 'complete CDS' in 'title' field were used in the search field for downloading sequences; *Ensembl Biomarts*, (http://plants.ensembl.org/biomart/martview/); *TAIR*, (http://arabidopsis.org/).

2.2. Data analysis

For a particular species, 10 nucleotides both upstream and downstream of the AUG were extracted and aligned together without any gaps and subsequently nucleotide frequency at each position across this 23 bp window was calculated. All the analysis was done using programs written in Python, BioPython (Cock et al., 2009), ggplot2 (Wickham, 2009), and R (RC Team, 2013).

Context sequence was determined by two methods: *Cavener's criteria and G statistic*. For Cavener's criteria, a single base was given *consensus* status and indicated by capital letter if the relative frequency of a single nucleotide at a certain position is greater than 50% and greater than twice the relative frequency of the second most frequent base. When no single base fulfilled the above-mentioned conditions, a pair of bases was suggested *co-consensus* if the sum of relative frequencies of those two nucleotides exceeded 75%. If neither of these two criteria was fulfilled, the position was denoted by the most frequent or *dominant* nucleotide in lower case and if two bases have the same higher frequency, they were recognized as *co-dominant bases* (Cavener, 1987).

In G statistic, the observed frequency of nucleotides is compared with the expected frequency and any significant difference between the two is reported. Hence G statistic was preferred over Cavener's criteria in estimating the nucleotide bias.

The *G*-value at position *i* was calculated by the formula:

$$G^{(i)} = 2\sum_{n} O_n^{(i)} \ln\left(\frac{O_n^{(i)}}{E_n^{(i)}}\right) \quad (1)$$

where $O_n^{(i)}$ is the observed number of nucleotide *n* (A, U, G and C) at position *i*, and $E_n^{(i)}$ is the expected number of nucleotide *n* at position *i*.

For each species, expected frequency was computed separately at the first, second and third base positions of a codon and then averaged for all the codons in CDS and a separate expected frequency for 5′ untranslated region (UTR) was also used. When the sample size is large, the distribution of G statistic can be represented by Chi-square distribution with $f-1$ degrees of freedom, where f is the number of different classes. Each term in formula (1) represents the contribution of each nucleotide to the bias. To normalize the G value against the different number of sequences found in each genome, a value has been defined (Nakagawa et al., 2008) such that: