**RESEARCH NOTE**

# Comparative studies on sequence characteristics around translation initiation codon in four eukaryotes

QINGPO LIU and QINGZHONG XUE*

*Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, 310029 Hangzhou, People's Republic of China*

## Introduction

The initiation of protein biosynthesis is a major determinant of the efficiency of gene expression at the translational level. As indicated in many studies, the nucleotide sequences around the AUG translation initiation codon (AUG context) act as an important signal to trigger the initiation of translation event (Kozak 1987; Miyasaka *et al*. 2002). It is also known that the occurrence of a purine at the –3 and a G at the +4 position is the most important feature of the AUG context in eukaryotes, and the strong preference for G at the +4 position promotes the process of recognition of the actual initiation codon by the ribosome subunits (Kozak 1997, 1999). However, there is presently little knowledge about the sequence characteristics around – especially before – the initiator codon, and their effects on translation initiation efficiency. Recently, the whole genome sequences of *Arabidopsis thaliana*, *Homo sapiens*, *Drosophila melanogaster* (Kaul *et al*. 2000; Adams *et al*. 2000; Venter *et al*. 2001), and the draft genome sequence of *Oryza sativa* have been determined (Yu *et al*. 2002; Goff *et al*. 2002). Therefore, it is of interest to determine whether hitherto unknown sequence elements might affect translation initiation in the genes of these eukaryotes. We have done this through a systematic analysis of codon and nucleotides biases around the AUG codon in these four species (a monocot, dicot, invertebrate and vertebrate). Moreover, we also examined the relationship between the conservation of nucleotides around the initiator codon and gene expression level.

*For correspondence. E-mail: liuqp@genomics.org.cn; qzhxue@hotmail.com.

## Materials and methods

### *Sequence data*

Publicly available full-length cDNA sequences of *A. thaliana*, *O. sativa*, *D. melanogaster* and *H. sapiens* were downloaded from the GenBank database (release 140.0). To minimise sampling errors, the redundant sequences were excluded, as were sequences: (1) that had incorrect initiation and termination codons, and (2) in which the length of 5′-UTR was less than 60 bases. Finally, we used 27238 sequences for *A. thaliana*, 23175 for *O. sativa*, 14933 for *D. melanogaster*, and 26749 for *H. sapiens* for further analysis.

### *Evaluation of the biases*

The *G*-test (Sokal and Rohlf 1993) was employed for evaluating the observed deviation of codons distribution from the expected distribution at each codon position. The AUG codon was taken as 0, and the codon immediately before or following the initiator codon was referred to as codon –1 or codon +1, respectively. The bias in codon appearance at position *i* was evaluated by the *G*-value, defined as:

$$G_i = N\sum\nolimits_{(x)} 2O_i^{(x)} \ln\frac{O_i^{(x)}}{E^{(x)}}.$$

Where $N$ was the number of cDNA sequences, $O_i^{(x)}$ was the frequency of codon $x$ at a given position, and $E^{(x)}$ was the frequency of codon $x$ in the coding sequence region (+1 ~ +10) or noncoding region (–1 ~ –20) of all cDNA sequences.

### *Characteristic analysis of nucleotide distribution around the initiator codon*

Each base frequency from the –30 to +33 positions, and

**Keywords.** *Arabidopsis; Drosophila*; rice; translation initiator codon; sequence characteristic; gene expression level.

the average frequencies of the four bases in CDS and non-CDS regions were calculated by writing PERL script. Then the frequency of each base was taken to subtract the average frequency in CDS (+4 ~ +33) and non-CDS (–30 ~ –1). The distribution of base bias at each position is illustrated in figure 1.

***Calculation of information content***

To identify highly conserved positions, information content (Schneider *et al*. 1986) was calculated for positions through –30 to +33 with respect to the initiator codon. The calculation formula of information content was defined as follows:

$$I(L) = \sum_{b=A,U,C,G} f(b,L)\log_2 \frac{f(b,L)}{p(b)}.$$

Where *f (b,L)* and *p (b)* were the frequency and probability of nucleotide *b* (A, U, C or G) at the *L* position respectively.

***Expression profile***

As was indicated, mRNA abundance of a given gene assessed by EST (expressed sequence tag) data could be used to estimate the expression level of that gene. Therefore, expression profile was determined by counting the number of occurrence of each gene among EST sequences. Targeted CDSs were aligned with all of the EST sequences according to species by using BLASTN. If the BLASTN alignment showed at least 95% identity over 100 nucleotides, it was counted as a sequence match (Duret and Mouchiroud 1999). In each species, the sequences for analysing were equally ranked and clustered into ten groups according to their mRNA abundance values.

## Results

***Analysis of base biases with respect to the translation initiation codon***

The *G* value of each codon position from –20 to +10 with respect to the AUG codon was calculated (data not shown). An interesting finding was that the codons immediately before (–1 and –2 codons) and following (+1 codon) the AUG codon clearly displayed larger biases than other codon positions. In these four eukaryotes, the –1 codon was the most biased among all the examined positions, and next was the +1 codon, with the exception of *D. melanogaster* for which the bias at codon –2 was slightly higher than at the +1 codon. Notably, the differ-

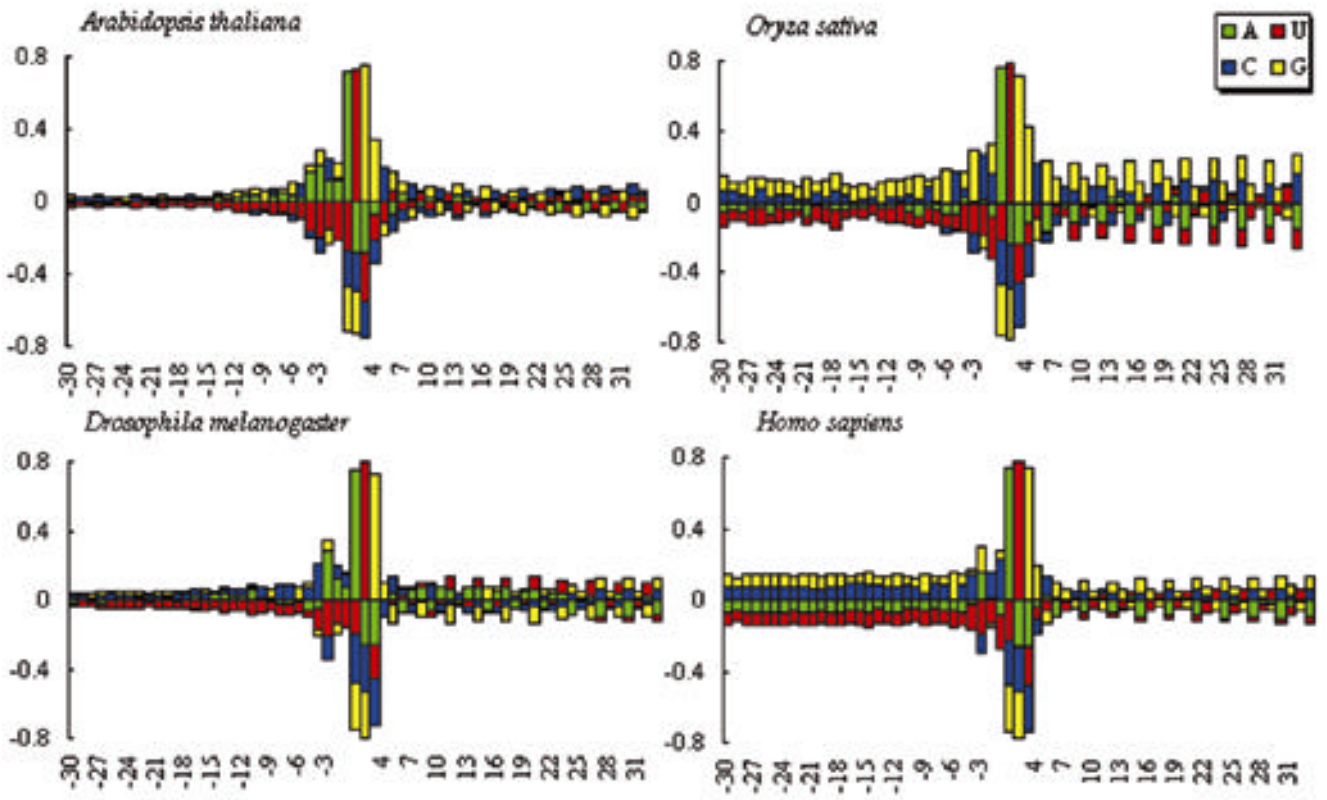


**Figure 1.** Distribution of biases of four kinds of nucleotides at each base position around the initiator codon in the four eukaryotes studied.

ence between the –1 codon and +1 codon was significantly larger for *D. melanogaster* and *H. sapiens*, compared to the two plant species.

Then a detailed analysis concerning of the –1 and +1 codons in the appearance of base and codon was performed, revealing that G and C were preferentially used at the first and second letters respectively in the +1 codon for every examined eukaryote (figure 1). Otherwise, high frequency of G at the third letter in the +1 codon for *H. sapiens*, *A. thaliana* and *O. sativa* was found. In agreement with these observations, the most frequently biased +1 codon position might be GCG encoding an alanine for the above three species. However, alanine encoding codons except for GCA were much favoured at the +1 codon in *D. melanogaster*, although their biases were not outstanding compared with *H. sapiens* and plants. As for the –1 codon, it was found that the first letter was highly biased toward purines. The biggest discrepancies were at the second and third bases of the –1 codon among the four eukaryotes. For *H. sapiens* and *O. sativa*, C was the greatly favoured nucleotide at the second letter and C or G at the third letter. In contrast, A or C at the second letter, and A, G or C at the third letter were biased to use in *A. thaliana* and *D. melanogaster*. Overall, there was no clear tendency for preferentially used codon in the –1 position among them that could be regarded as species-specific characteristic. However, according to the results illustrated in figure 1, it could be deduced that the consensus sequence around the AUG codon for each species would be A(A/C)(A/G)AUGGC (G/U) for *A. thaliana*, GC(G/C) AUGGCG for *O. sativa*, (A/G)C(C/G)AUGGCG for *H. sapiens*, and A(A/C)(A/C) AUGGCB for *D. melanogaster* where B means U, C or G, a finding which has relevance for functional genome annotation and further improvement of gene prediction tools.

***Relationship between information content and gene expression level***

Most of the three base positions either in –1 or +1 codons were highly conserved compared with other positions, especially the –3 and +4 positions for *A. thaliana*, *O. sativa* and *H. sapiens*. As for *D. melanogaster*, the most biased base position was located at –3; however, no apparent differences between the +4, +5, +6 positions and others were found (figure 2). Interestingly, the +4 was the highest conserved base position in plants, while the –3 position was highly conserved in animals.

From figure 3, it could be observed that the base positions around the AUG codon displayed significant conservation in highly expressed genes, especially the –3, +4, +5, and +6 positions in *A. thaliana*, *D. melanogaster* and *H. sapiens*, and the –2 and +5 positions in *O. sativa*. Significant correlations were also found at several other

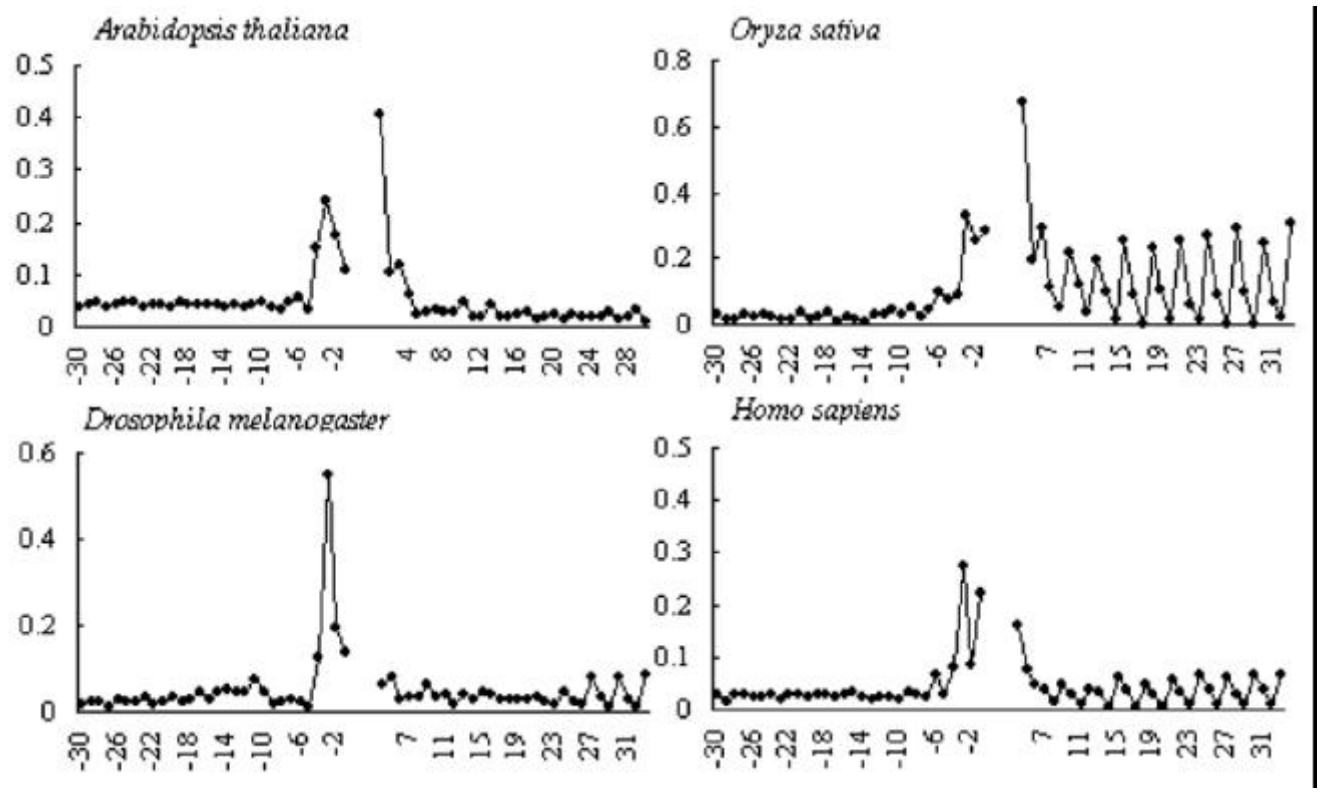


**Figure 2.** Information content of each base position from –30 to +33 in the four eukaryotes studied.

positions (see table 1 of the electronic supplement on the journal website www.ias.ac.in/jgenet). It is worth noting that most of the positions before the AUG codon in *H. sapiens* were greatly conserved in highly expressed genes, as compared to genes expressed at lower levels, suggesting that the genes of the former group possessed strong consensus to regulate the efficiency of translation initiation. However, the opposite was true in the case of *A. thaliana*, leading us to speculate that those positions being significantly correlated with gene expression level might involve in translation initiation, regardless of the possible positive or negative regulation they performed.

In general, we feel that positions with high information content are more likely to exhibit base conservation in the course of evolution. For example, C was significantly positively correlated with information content at position +5 in the four eukaryotes, and was the preferred base in

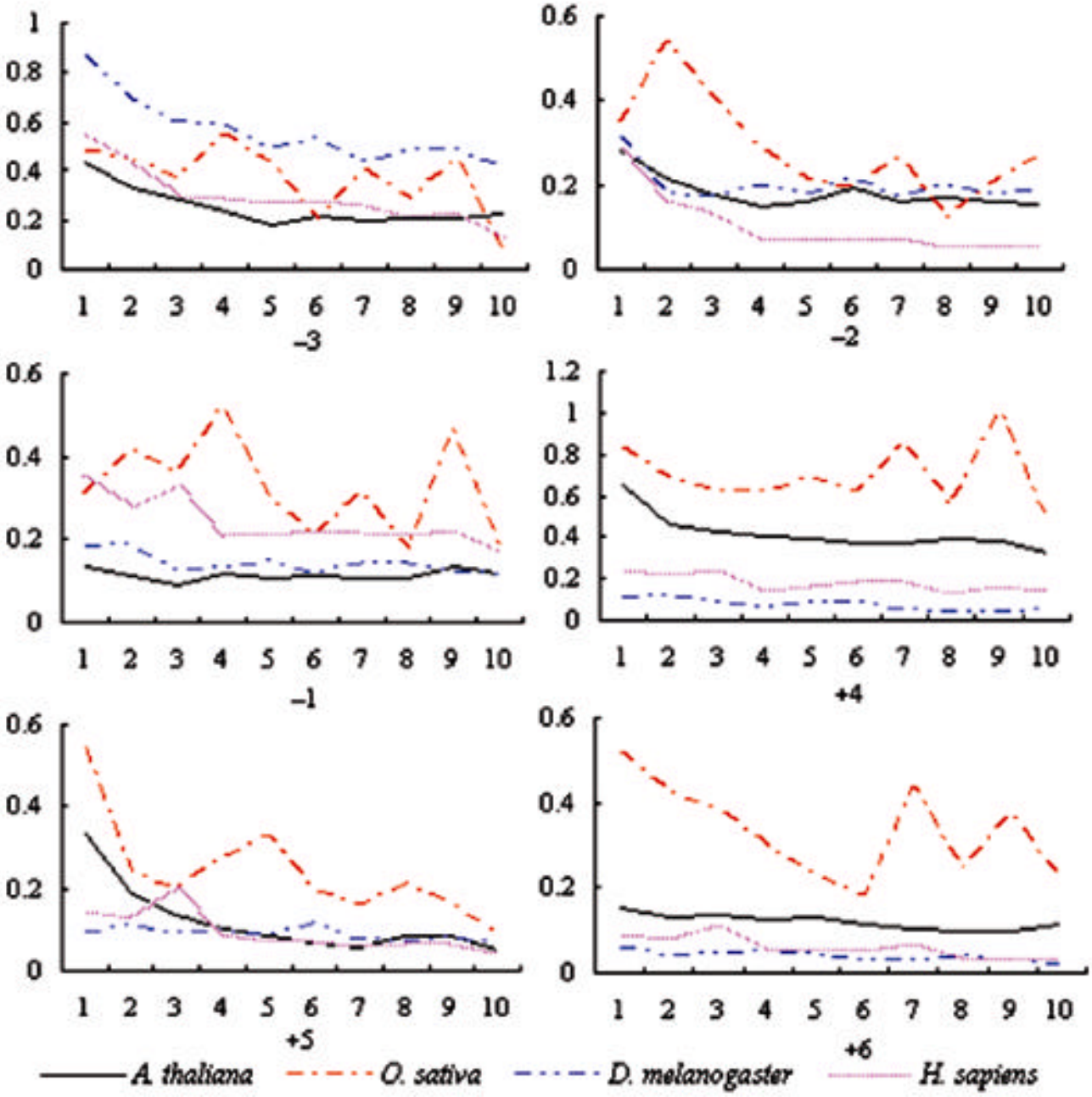


**Figure 3.** Relationship between information content (from –3 to +6 positions) and gene expression level assessed by mRNA abundance. The horizontal axis represents gene groups ranked by their mRNA abundance (the genes in group 1 have the highest mRNA abundance values). The vertical axis represents the information content.

highly expressed genes. Moreover, it was observed that highly conserved positions, especially those positions around the AUG codon showed very strong preference for C or G, irrespective of the genomic environment (see table 2 of the electronic supplement on the journal website www.ias.ernet.in/jgenet).

## Discussion

Using 699 vertebrate mRNA sequences, Kozak (1987) proposed a GC-rich sequence, GCCGCC(A/G)CCAUGG as the consensus sequence for the context of functional AUG codon. Nevertheless, higher plants were seen to possess an AC-rich consensus sequence, caA(A/C) aAUGGCg as a context of translation initiator codon (Joshi *et al*. 1997). However, it has also been suggested that the AUG context might play a less significant role on translation efficiency in plants than in animals (Joshi *et al*. 1997). In this study, we observed that the –1 and +1 codons, especially the former one, were highly significantly conserved, with GCG being the most biased +1 codon while the –1 codon showed a species-specific pattern. As reported by Stenström *et al*. (2001) and Gonzalez de Valdivia and Isaksson (2004), high A content of the codon immediately downstream of the AUG codon was associated with high gene expression and gave high translation initiation in *Escherichia coli*. Therefore, such biases in codon and base appearances immediately upstream and downstream of the AUG codon could be an extended signal for efficiency of translation initiation.

In human genes, the preference for the optimal nucleotide of the mammalian translation initiation AUG context (GCCGCC(A/G)CCAUGG) was generally more pronounced in the highly expressed genes at the –9 through –1 positions (Miyasaka *et al*. 2002), indicating that gene expression level was significantly affected by the conservation of sequence flanking AUG initiation codon. A similar observation was also made in our study. It was inferred that highly expressed genes were likely to have strong consensus sequences around the initiator codon, and those preferred bases would be conserved in the course of evolution in order to make their translation initiation efficiency (Ozawa *et al*. 2002); while weakly expressed genes were less selective in the regions around the initiator codon and its effect on translation initiation. The above explanation might be suitable for *H. sapiens* and *D. melanogaster*. However, no clear relationship between gene expression level and information content was found in *O. sativa* except for the –16, –5, –4, –2, and +5 positions. In the case of *A. thaliana*, all of the significant correlations were negative and located at those positions before –3 (see table 1 of the electronic supplement: http://www.ias.ac.in/jgenet/index.htmlCurrentNumber.htm), a result that is opposite to the pattern seen in *H. sapiens*. In addition, A and U were preferentially used in highly expressed genes in the AT-rich organism *A. thaliana*. In contrast, in the three GC-rich species, highly expressed genes were biased to choose C and G (see table 2 of the electronic supplement on the journal website www.ias.ernet.in/jgenet). Overall, it could be deduced that the strong consensus sequences in highly expressed genes are likely to play crucial roles in translation initiation, and the sequence patterns upstream of the –3 position might be involved in the regulation of translation inititation.

**Acknowledgements**

This work was supported by the grant from the Key Research Project of Zhejiang Province (2003C22007). We are also thankful for the valuable suggestions and comments of Dr Amitabh Joshi and the two anonymous reviewers.

## References

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F., *et al*. 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Duret L. and Mouchiroud D. 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487.

Goff S. A., Ricke D., Lan T. H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., *et al*. 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100.

Gonzalea de Valdivia E. I. and Isaksson L. A. 2004 A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res*. **32**, 5198–5205.

Joshi C. P., Zhou H., Huang X. Q. and Chiang V. L. 1997 Context sequences of translation initiation codon in plants. *Plant Mol. Biol*. **35**, 993–1001.

Kaul S., Koo H. L., Jenkins J., Rizzo M., Rooney T., Tallon L. J., Feldblyum T., Nierman W., Benito M., Lin X., *et al*. 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Kozak M. 1987 An analysis of 5′-noncoding sequences from 699 vertebrates messenger RNAs. *Nucleic Acids Res*. **15**, 8125– 8148.

Kozak M. 1997 Recognition of AUG and alternative-initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J*. **16,** 2482–2492.

Kozak M. 1999 Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**, 187–208.

Miyasaka H., Kanai S., Tanaka S., Akiyama H. and Hirano M. 2002 Statistical analysis of the relationship between translation initiation AUG context and gene expression level in humans. *Biosci. Biotech. Biochem*. **66**, 667–669.

Ozawa Y., Hanaoka S., Saito R., Washio T., Nakano S., Shinagawa A., Itoh M., Shibata K., Carninci P., Konno H., Kawai, J., Hayashizaki Y. and Tomita M. 2002. Comprehensive sequence analysis of translation termination sites in various eukaryotes. *Gene* **300**, 79–87.

Schneider T. D., Stormo G. D., Gold L. and Ehrenfeucht A. 1986

Information content of binding sites on nucleotide sequence. *J. Mol. Biol.* **188**, 415–431.

Sokal R. R. and Rohlf F. J. 1993 *Biometry*, 3rd Edn. W. H. Freeman and Co., New York.

Stenström C. M., Jin H., Major L. L., Tate W. P. and Isaksson L. A. 2001 Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* **263**, 273–284.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., *et al*. 2001 The sequence of the human genome. *Science* **291**, 1304–1351.

Yu J., Hu S., Wang J., Wong G. K., Li S., Liu B., Deng Y., Dai L., Zhou Y., Zhang X., *et al*. 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92

.

Received 5 April 2005

# Supplementary data: Comparative studies on sequence characteristics around translation initiation codon in four eukaryotes

Qingpo Liu and Qingzhong Xue

*J. Genet.* **84**, 317–322

**Table 1.** Spearman's rank correlation coefficients of 39 base positions around the AUG codon in the four eukaryotic species studied.

| | **– 30** | **– 29** | **– 28** | **– 27** | **– 26** | **– 25** | **– 24** | **– 23** | **– 22** | **– 21** | **– 20** | **– 19** | **– 18** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *A. thaliana* | – 0.248 | – 0.806** | – 0.612 | – 0.709* | – 0.030 | – 0.321 | – 0.576 | – 0.539 | – 0.188 | – 0.552 | – 0.636* | – 0.442 | 0.006 |
| *O. sativa* | 0.236 | 0.442 | – 0.139 | 0.152 | 0.406 | – 0.049 | 0.467 | 0.176 | – 0.273 | 0.273 | – 0.224 | 0.430 | – 0.127 |
| *D. melanogaster* | – 0.709* | – 0.030 | 0.188 | – 0.418 | – 0.273 | – 0.030 | 0.224 | 0.139 | – 0.273 | – 0.067 | – 0.358 | 0.697* | 0.661* |
| *H. sapiens* | 0.879** | 0.915** | 0.806** | 0.903** | 0.867** | 0.758* | 0.358 | 0.697* | 0.879** | 0.467 | 0.915** | 0.442 | 0.867** |
| | **– 17** | **– 16** | **– 15** | **– 14** | **– 13** | **– 12** | **– 11** | **– 10** | **– 9** | **– 8** | **– 7** | **– 6** | **– 5** |
| *A. thaliana* | – 0.915** | – 0.515 | – 0.042 | – 0.224 | – 0.394 | – 0.479 | – 0.758* | – 0.564 | – 0.648* | – 0.830** | – 0.224 | 0.030 | – 0.055 |
| *O. sativa* | 0.248 | – 0.697* | 0.018 | 0.176 | – 0.127 | – 0.067 | 0.297 | – 0.042 | – 0.285 | 0.552 | 0.345 | 0.455 | 0.661* |
| *D. melanogaster* | 0.067 | 0.358 | 0.782** | 0.079 | 0.612 | 0.200 | 0.164 | 0.152 | 0.067 | 0.879** | 0.503 | 0.648* | 0.455 |
| *H. sapiens* | 0.891** | 0.806** | 0.903** | 0.733* | 0.442 | 0.721* | 0.782** | 0.939** | 0.745* | 0.939** | 0.636* | 0.855** | 0.818** |
| | **– 4** | **– 3** | **– 2** | **– 1** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| *A. thaliana* | 0.564 | 0.648* | 0.624 | 0.030 | 0.891** | 0.903** | 0.855** | – 0.382 | 0.576 | – 0.624 | 0.152 | – 0.152 | – 0.515 |
| *O. sativa* | 0.705* | 0.624 | 0.697* | 0.382 | 0.243 | 0.770** | 0.552 | 0.333 | – 0.467 | 0.164 | 0.285 | – 0.552 | 0.309 |
| *D. melanogaster* | 0.721* | 0.939** | 0.200 | 0.661* | 0.903** | 0.721* | 0.830** | 0.770** | 0.067 | 0.770** | 0.564 | 0.491 | 0.018 |
| *H. sapiens* | 0.818** | 0.976** | 0.939** | 0.709* | 0.733* | 0.903** | 0.879** | 0.770** | 0.309 | 0.915** | – 0.042 | – 0.394 | 0.842** |

*Note*: * and ** indicate statistical significance at the 0.05 and 0.01 level respectively.

**Table 2.** Spearman's rank correlation coefficients of four kinds of bases with information content in the four eukaryotic species studied.

| *A. thaliana* | **– 30** | **– 29** | **– 28** | **– 27** | **– 26** | **– 25** | **– 24** | **– 23** | **– 22** | **– 21** | **– 20** | **– 19** | **– 18** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.636* | – 0.333 | – 0.394 | – 0.733* | – 0.418 | – 0.745* | – 0.903** | – 0.903** | – 0.578 | – 0.818** | – 0.685* | – 0.867** | – 0.770** |
| U | 0.370 | – 0.201 | – 0.127 | – 0.115 | 0.418 | 0.661* | 0.401 | 0.729* | 0.248 | 0.225 | – 0.091 | 0.309 | 0.632* |
| C | 0.632* | 0.915** | 0.778** | 0.794** | 0.600 | – 0.673* | 0.806** | 0.794** | 0.770** | 0.721* | 0.596 | 0.855** | 0.382 |
| G | – 0.164 | – 0.479 | – 0.067 | 0.200 | – 0.042 | – 0.055 | – 0.673* | – 0.067 | – 0.152 | 0.103 | 0.042 | – 0.091 | – 0.345 |
| | **– 17** | **– 16** | **– 15** | **– 14** | **– 13** | **– 12** | **– 11** | **– 10** | **– 9** | **– 8** | **– 7** | **– 6** | **– 5** |
| A | – 0.806** | – 0.321 | – 0.552 | – 0.503 | – 0.394 | 0.103 | – 0.231 | – 0.067 | – 0.152 | – 0.693* | – 0.018 | – 0.358 | – 0.455 |
| U | – 0.224 | – 0.079 | 0.224 | 0.248 | – 0.067 | – 0.383 | – 0.394 | – 0.648* | – 0.503 | – 0.778** | – 0.539 | 0.091 | – 0.745* |
| C | 0.927** | 0.413 | 0.444 | 0.685* | 0.579 | 0.576 | 0.418 | 0.612 | 0.479 | 0.818** | 0.602 | – 0.636* | 0.790** |
| G | 0.292 | – 0.097 | – 0.018 | – 0.212 | – 0.207 | – 0.115 | 0.297 | 0.515 | – 0.139 | 0.042 | – 0.073 | 0.673* | – 0.103 |
| | **– 4** | **– 3** | **– 2** | **– 1** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| A | 0.564 | 0.588 | – 0.122 | – 0.333 | – 0.891** | – 0.709* | – 0.855** | – 0.867** | – 0.830** | – 0.673* | – 0.345 | – 0.709* | – 0.733* |
| U | – 0.648* | – 0.578 | – 0.576 | – 0.733* | – 0.535 | – 0.867** | 0.564 | 0.188 | – 0.103 | – 0.503 | 0.333 | 0.127 | 0.455 |
| C | – 0.176 | – 0.891** | 0.867** | 0.842** | – 0.539 | 0.927** | – 0.067 | 0.139 | 0.903** | 0.709* | – 0.273 | 0.733* | 0.770** |
| G | 0.103 | – 0.164 | – 0.806** | 0.224 | 0.915** | – 0.830** | 0.818** | 0.661* | – 0.602 | 0.394 | 0.394 | – 0.006 | – 0.067 |
| *O. sativa* | **– 30** | **– 29** | **– 28** | **– 27** | **– 26** | **– 25** | **– 24** | **– 23** | **– 22** | **– 21** | **– 20** | **– 19** | **– 18** |
| A | 0.134 | 0.237 | – 0.261 | 0.200 | – 0.723* | 0.012 | 0.030 | 0.049 | 0.537 | – 0.049 | 0.587 | – 0.024 | – 0.195 |
| U | – 0.430 | – 0.301 | – 0.055 | – 0.413 | 0.103 | 0.345 | – 0.503 | – 0.134 | – 0.152 | – 0.378 | – 0.309 | – 0.457 | 0.073 |
| C | 0.460 | 0.372 | 0.503 | 0.067 | 0.555 | – 0.311 | 0.575 | 0.231 | 0.195 | – 0.092 | – 0.079 | 0.055 | 0.564 |
| G | – 0.334 | – 0.310 | – 0.311 | – 0.079 | – 0.267 | 0.024 | 0.500 | 0.030 | – 0.482 | 0.055 | – 0.251 | 0.677* | – 0.406 |
| | **– 17** | **– 16** | **– 15** | **– 14** | **– 13** | **– 12** | **– 11** | **– 10** | **– 9** | **– 8** | **– 7** | **– 6** | **– 5** |
| A | – 0.097 | 0.456 | 0.091 | – 0.012 | – 0.049 | 0.244 | 0.517 | – 0.390 | – 0.129 | 0.000 | 0.426 | – 0.543 | – 0.418 |
| U | – 0.178 | – 0.030 | – 0.025 | – 0.280 | – 0.261 | 0.098 | – 0.092 | 0.406 | 0.390 | 0.098 | – 0.231 | – 0.018 | – 0.360 |
| C | 0.139 | 0.767** | – 0.055 | 0.444 | 0.091 | – 0.401 | – 0.109 | – 0.030 | 0.268 | 0.438 | – 0.418 | – 0.037 | 0.890** |
| G | – 0.239 | – 0.584 | 0.036 | 0.249 | – 0.267 | 0.049 | – 0.176 | 0.018 | – 0.415 | – 0.624 | 0.043 | 0.632* | – 0.407 |
| | **– 4** | **– 3** | **– 2** | **– 1** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| A | 0.117 | 0.456 | – 0.177 | 0.384 | – 0.086 | – 0.538 | – 0.826** | 0.109 | – 0.109 | – 0.784** | 0.262 | 0.650* | 0.000 |
| U | – 0.267 | – 0.772** | – 0.128 | – 0.585 | 0.433 | – 0.243 | 0.267 | – 0.320 | 0.433 | 0.451 | – 0.465 | 0.067 | – 0.213 |
| C | 0.143 | – 0.153 | 0.661* | 0.596 | – 0.681* | 0.890** | 0.122 | – 0.462 | 0.178 | 0.697* | 0.323 | 0.176 | – 0.128 |
| G | – 0.049 | 0.195 | – 0.584 | – 0.394 | 0.166 | – 0.369 | 0.018 | 0.285 | – 0.721* | – 0.732* | – 0.224 | – 0.600 | 0.227 |

**Table 2.** (*Contd.*)

| *D. melanogaster* | **– 30** | **– 29** | **– 28** | **– 27** | **– 26** | **– 25** | **– 24** | **– 23** | **– 22** | **– 21** | **– 20** | **– 19** | **– 18** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.782** | – 0.030 | 0.297 | – 0.503 | – 0.334 | – 0.139 | 0.042 | 0.248 | – 0.462 | – 0.042 | – 0.248 | 0.721* | 0.648* |
| U | 0.406 | 0.188 | – 0.430 | – 0.219 | – 0.024 | – 0.321 | – 0.564 | – 0.103 | 0.782** | 0.127 | – 0.438 | – 0.794** | – 0.778** |
| C | 0.091 | 0.067 | – 0.766** | 0.794** | 0.176 | 0.200 | 0.200 | – 0.273 | – 0.188 | 0.212 | 0.539 | – 0.401 | 0.565 |
| G | 0.103 | – 0.042 | 0.782** | – 0.152 | 0.309 | – 0.188 | – 0.170 | – 0.164 | 0.188 | – 0.055 | 0.527 | – 0.152 | – 0.699* |

| | **– 17** | **– 16** | **– 15** | **– 14** | **– 13** | **– 12** | **– 11** | **– 10** | **– 9** | **– 8** | **– 7** | **– 6** | **– 5** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.073 | – 0.395 | 0.733* | 0.115 | 0.564 | 0.164 | 0.103 | 0.309 | – 0.213 | 0.467 | 0.455 | 0.479 | – 0.467 |
| U | – 0.673* | – 0.903** | – 0.285 | – 0.736* | – 0.673* | – 0.778** | – 0.571 | – 0.467 | 0.079 | – 0.875** | – 0.842** | – 0.515 | – 0.539 |
| C | 0.321 | 0.624 | – 0.152 | 0.479 | 0.754* | 0.555 | – 0.261 | – 0.539 | – 0.479 | 0.794** | 0.042 | – 0.055 | 0.552 |
| G | 0.292 | 0.309 | – 0.612 | – 0.018 | – 0.588 | 0.091 | 0.661* | 0.200 | 0.067 | – 0.255 | – 0.382 | 0.067 | 0.103 |

| | **– 4** | **– 3** | **– 2** | **– 1** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.347 | 0.830** | 0.127 | 0.267 | – 0.648* | – 0.830** | – 0.903** | 0.200 | 0.292 | 0.139 | – 0.248 | 0.455 | – 0.122 |
| U | – 0.939** | – 0.988** | – 0.321 | – 0.782** | – 0.067 | – 0.624 | 0.588 | – 0.467 | – 0.855** | – 0.663* | – 0.760* | – 0.236 | 0.127 |
| C | 0.673* | – 0.717* | – 0.067 | – 0.103 | – 0.685* | 0.855** | 0.280 | – 0.782** | 0.176 | 0.505 | 0.127 | 0.255 | 0.370 |
| G | – 0.406 | – 0.517 | 0.006 | 0.608 | 0.867** | 0.188 | – 0.347 | 0.879** | 0.624 | – 0.127 | 0.685* | – 0.430 | – 0.442 |

| *H. sapiens* | **– 30** | **– 29** | **– 28** | **– 27** | **– 26** | **– 25** | **– 24** | **– 23** | **– 22** | **– 21** | **– 20** | **– 19** | **– 18** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.879** | – 0.915** | – 0.450 | – 0.782** | – 0.879** | – 0.505 | – 0.407 | – 0.648* | – 0.903** | – 0.588 | – 0.624 | – 0.333 | – 0.687* |
| U | – 0.661* | – 0.139 | – 0.758* | – 0.505 | – 0.309 | – 0.442 | – 0.212 | – 0.345 | – 0.685* | – 0.091 | – 0.939** | – 0.309 | – 0.830** |
| C | 0.418 | 0.915** | 0.588 | 0.225 | 0.274 | 0.612 | 0.018 | 0.419 | 0.515 | 0.382 | 0.661* | 0.869** | 0.709* |
| G | 0.855** | 0.503 | 0.486 | 0.721* | 0.867** | – 0.146 | 0.661* | 0.818** | 0.576 | 0.006 | 0.811** | 0.030 | 0.733* |

| | **– 17** | **– 16** | **– 15** | **– 14** | **– 13** | **– 12** | **– 11** | **– 10** | **– 9** | **– 8** | **– 7** | **– 6** | **– 5** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.939** | – 0.139 | – 0.648* | – 0.486 | – 0.503 | – 0.867** | – 0.600 | – 0.006 | – 0.754* | – 0.588 | – 0.024 | – 0.632* | – 0.915** |
| U | – 0.624 | – 0.806** | – 0.697* | – 0.770** | 0.225 | – 0.079 | – 0.721* | – 0.661* | – 0.382 | – 0.863** | – 0.620 | – 0.903** | 0.103 |
| C | 0.709* | 0.758* | 0.912** | 0.418 | 0.273 | 0.772** | 0.842** | 0.842** | – 0.164 | 0.903** | 0.612 | – 0.419 | 0.794** |
| G | 0.188 | 0.067 | 0.815** | 0.733* | 0.055 | 0.345 | 0.164 | – 0.012 | 0.770** | – 0.055 | – 0.370 | 0.806** | 0.049 |

| | **– 4** | **– 3** | **– 2** | **– 1** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | – 0.552 | 0.915** | – 0.830** | – 0.527 | – 0.815** | – 0.721* | – 0.809** | – 0.468 | 0.432 | – 0.863** | – 0.442 | – 0.588 | – 0.879** |
| U | – 0.806** | – 0.939** | – 0.915** | – 0.766** | 0.280 | – 0.709* | – 0.503 | – 0.389 | – 0.285 | – 0.685* | 0.164 | 0.152 | – 0.442 |
| C | 0.855** | – 0.709* | 0.964** | 0.745* | – 0.723* | 0.818** | 0.600 | – 0.261 | 0.139 | 0.830** | 0.224 | 0.353 | 0.152 |
| G | – 0.758* | 0.903** | – 0.815** | – 0.626 | 0.709* | – 0.358 | 0.790** | 0.576 | – 0.115 | 0.470 | 0.139 | 0.212 | 0.612 |

*Note*: * and ** indicate statistical significance at the 0.05 and 0.01 level respectively.