Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
BioSystems 80 (2005) 123–131
BioSystems
www.elsevier.com/locate/biosystems

Synonymous codon usage and gene function are strongly related in *Oryza sativa*

Qingpo Liu[a], Shijuan Dou[b], Zhijuan Ji[c], Qingzhong Xue[a,∗]

[a] *Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*
[b] *Department of Biochemistry and Cell Biology, College of Life Science, Agricultural University of Hebei, Baoding 071001, China*
[c] *State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China*

Received 27 August 2004; received in revised form 28 October 2004; accepted 29 October 2004

Abstract

The relationship between codon usage and gene function was investigated while considering a dataset of 2106 nuclear genes of *Oryza sativa*. The results of standard χ^2 test and *F*-statistic showed that for every 59 synonymous codons, a strongly significant association with gene functional categories existed in rice, indicating that codon usage was generally coordinated with gene function whether it was at the level of individual amino acids or at the level of nucleotides. However, it could not be directly said that the use of every codons differed significantly between any two functional categories. Notably, there existed large difference both in selection for biased codons or selection intensity among functional categories. Therefore, we identified at least two classes of genes: one group of genes, mainly belonging to the "METABOLISM" category, was tended to use G- and/or C-ending codons while the other was more biased to choose codons ending with A and/or U. The latter group contained genes of various functions, especially those genes classified into the "Nuclear Structure" category. These observations will be more important for molecular genetic engineering and genome functional annotation.

Keywords: Rice; Codon usage; Gene function

---

1. Introduction

Non-random use of synonymous codons universally exists both within and between organisms. The results of numerous studies demonstrated that there was a species-specific pattern of codon usage. Especially, closely related organisms always shared similarities in codon frequency (Sharp et al., 1988). However, it was observed that there were big differences in codon usage among genes within one species. For example, correspondence analysis identified at least two classes of genes in *Arabidopsis thaliana* according to codon usage, in which one group of genes was highly biased to G/C; the other had a weaker preference for A/T-biased codons (Chiapello et al., 1998). In some unicellular prokaryotes, the codon usage was attributable to

---

∗ Corresponding author. Tel.: +86 571 86971611;
fax: +86 571 86971117.
E-mail address: qzhxue@genomics.org.cn (Q. Xue).

0303-2647/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2004.10.008

the equilibrium between natural selection and compositional mutation bias (Bulmer, 1988; Sharp et al., 1993). In contrast, diverse patterns of codon usage might arise from compositional constraints of the genomes in some prokaryotes with extremely high AT or GC contents (Sharp et al., 1993) and human (Karlin and Mrázek, 1996). It was interesting to note that with the development of genome projects, rather complex patterns of codon usage were reported both in prokaryotes and eukaryotes (Romero et al., 2000; Gupta et al., 2004; Liu et al., 2004).

Analysis of codon usage data has both theoretical and practical importance in understanding the basics of molecular biology. In *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Oryza sativa*, there was a strongly significant correlation between gene expression level and codon usage bias (Ikemura, 1981; Sharp et al., 1986; Duret and Mouchiroud, 1999). Highly expressed genes displayed much more significant variation in codon usage than genes expressed at lower levels, suggesting that codon usage patterns had a functional significance. Using an experimental approach, Iannacone et al. (1997) and Rouwendal et al. (1997) demonstrated that there was a positive relationship between codon usage bias and gene expression level by transforming plants with vectors expressing genes with modified codon usage, which suggested stronger natural selection constrains on highly expressed genes to optimize translation efficiency and accuracy by the use of optimal codons than the lowly expressed genes did (Bulmer, 1988).

In all life forms, it appeared that codon usage bias was determined by diverse factors, such as expression levels, gene length, and protein secondary structure, etc. as indicated by many studies (Gouy and Gautier, 1982; Sharp and Li, 1986; Moriyama and Powell, 1998; Xie and Ding, 1998; Duret and Mouchiroud, 1999; Gupta et al., 2000). Recently, an important issue whether synonymous codon usage and gene function is related or not, which is significant and valuable for genetic engineering, has been reported, although there are only a few of mentions of this event (Chiapello et al., 1998; Ma et al., 2002). The correlation between codon usage and gene function was found in four mammal species, and the latter was the dominant factor in shaping codon usage (Ma et al., 2002). This observation was also found in *A. thaliana*, in which amino acids of different functional roles might be distinguished by the patterns of synonymous codon usage (Chiapello et al., 1998). Rice is the most important crop for human consumption, and about one-third of the population depends on rice for more than 50% of caloric intake. Currently, the draft genome sequences of rice have been determined (Yu et al., 2002; Goff et al., 2002). So it is of interest to investigate how about this relationship in *Oryza sativa*.

2. Materials and methods

2.1. Dataset

Oryza sativa complete Coding DNA Sequences (CDS) were downloaded from the GenBank database (release 140.0, http://www.ncbi.nlm.nih.gov). A PERL script developed by us was used to extract the CDS and protein sequences. To minimize sampling errors, only those CDS that have correct initial and termination codons and no internal stop codon were kept in the dataset.

2.2. Functional classification

For the well-characterized *A. thaliana* genome available from GenBank, genes have been assigned to 25 functional categories according to the role of the gene product plays for the living cell. In this study, a total of 13,745 *A. thaliana* protein sequences were retrieved (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/xindex.html). Using BLASTP, 17097 selected *Oryza sativa* protein sequences were aligned with *A. thaliana*. If the BLASTP alignment shows at least 95% identity and the length difference is not over 50 bases, it is counted as a sequence match. Accordingly, 2106 rice genes that belong to 23 functional categories were extracted for further analysis (Table 1).

2.3. Measures of codon usage bias

To normalize codon usage within datasets without the influence of amino acid compositions, relative synonymous codon usage (RSCU) values were calculated by dividing the observed codon usage by that expected when all codons for the same amino acid are used equally (Comeron and Aguade, 1998). RSCU values