Volume 289, number 1, 123–125 FEBS 10115 September 1991
© 1991 Federation of European Biochemical Societies 00145793/91/$3.50
ADONIS 001457939100831O

Two types of linkage between codon usage and gene-expression levels

Takeshi Nakamura[1], Akira Suyama[2] and Akiyoshi Wada[1,*]

[1]*Department of Physics, Faculty of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113, Japan* and [2]*Technological University of Nagaoka, Nagaoka, Niigata 940-21, Japan*

Received 5 June 1991

The relation between codon usage and gene-expression levels is an intensively investigated and discussed topic in the field of molecular evolution. We statistically analyzed 25 *Escherichia coli* gene sequences by a new classification of synonymous codons and found that (i) there are two distinct types of linkage between codon usage and gene-expression levels in *E. coli*, and (ii) one of the two kinds of codon preferences (the codon preference concerned with interaction of GC/AT choice at three codon positions) is observed significantly in weakly expressed genes.

Codon usage; Gene expression; Third letter usage; (G+C)-Content; tRNA content; Neutral theory

1. INTRODUCTION

Well-established constraints on codon choice in unicellular organisms are a tRNA-content constraint [1] and an optimization of codon–anticodon interaction energy [2]. In this text, codon preference attributed to the former constraint is termed 'type I preference' and that attributed to the latter constraint is called 'type II preference' for simplicity. We investigated directly, by a new distinction of synonymous codons, how these two kinds of preferences vary in 25 *E. coli* genes with different levels of expression, the genes* analyzed are listed in Table I (*GenBank Release 35.0).

Type I preference correlates well with gene expression levels in *E. coli*. This correlation has been clearly demonstrated using the concept of 'optimal codons' [1]. The dichotomy between optimal codons and non-optimal codons is based mainly on the constraint imposed by organism-specific populations of iso-accepting tRNAs. We, on the other hand, refer to a dichotomy leading to the type II preference as a 'counterbalanced/ uncounterbalanced dichotomy' for the following reason. This dichotomy discriminates codons in terms of the criterion of whether or not GC/AT choice at the codon third position is affected by GC/AT choice at the codon first and second positions. In a wide range of organisms, the local (G+C)-content at the codon third position in a gene has a negative correlation with the local (G+C)-content at the codon first and second positions (*the third letter counterbalance*) [3,4]. The third letter counterbalance in *E. coli* was found to be composed of intracodon adjustment and intercodon adjustment (unpublished); in highly expressed genes, our recent finding concerned with intracodon counterbalance adjustment agreed with the proposal of the optimization of codon–anticodon interaction energy [2].

2. RESULTS AND DISCUSSION

All codons concerned (47 codons specifying 14 amino acids) were classified into the following four groups: (1) optimal and counterbalanced codons, (2) optimal and uncounterbalanced codons, (3) non-optimal and counterbalanced codons, and (4) non-optimal and uncounterbalanced codons. Discrimination between optimal

Table I

The *E. coli*-gene sample (25 genes)

Group	Genes	Number of molecules*
A	lpp, rplL, tufA, ompA, rplA, rplK, rpsL, rplJ	1.5×10^5~ 1.5×10^4
B	lpd, uncA, uncD, rpoB	$6 \times 10^3 \sim 2 \times 10^3$
C	glyS, glnS, thrS	$1.5 \times 10^3 \sim 1 \times 10^3$
D	trpB, fol, trpC, trpA, thrA, lacY	
E	galR, lacI, trpR, araC	

Groups A–E are arranged in order by level of gene expression: (A) highly expressed genes; (B), (C) and (D) genes with intermediate levels of expression; (E) weakly expressed genes. Gene expression levels for groups D and E are estimated.

*Number of corresponding protein molecules per *E. coli* genome. Data follows the table developed by Ikemura [1].

**Present address:* Sagami Chemical Research Center, Sagamihara, Kanagawa 229, Japan.

Correspondence address: T. Nakamura, Biotechnological Laboratories, Central Research Division, Takeda Chemical Industries, Yodogawa-ku, Osaka 532, Japan.

codons and non-optimal codons followed the table made by Ikemura [1]. This classification of codons into four groups was made in parallel with an X·Z pair (consisting of a first letter and third letter in an identical codon) and with a YZ pair (consisting of a second letter and third letter in an identical codon). An example of this new codon-classification is listed in Table II for only X·Z pairs. We have tentatively called this a '2 × 2 distinction', because it is a combination of the preceding two types of codon dichotomies. Codons without degeneracy and these which could not be identified as either optimal or non-optimal were omitted in this analysis. There are two reasons that the analysis can be quite favorable in *E. coli*: (i) (G+C)-content constraint affects the choice of synonymous codons in each organism [5,6]. In *E. coli* (G+C)-content is almost 50% and thus (G+C)-content constraint can be neglected; (ii) a context effect is another factor which can affect codon usage [7], however, we found the context effect in the 25 *E. coli* genes, analyzed here, weaker than intracodon adjustment as an overall tendency (unpublished).

The strength of the two kinds of codon preferences, type I and II, was evaluated by measuring the deviation from the expected number of occurrences of the above four codon-groups. This deviation from expectation was quantified for each sequence by standard measure, D, i.e. $D = (O - E)/S$, where O is the observed number of occurrences; E and S are the expected number of occurrences and the standard deviation of the number of occurrences in a reference set of theoretical sequences, respectively. The reference set is made up by all permutations of synonymous codons under the following limitations: (a) an amino acid sequence encoded by the theoretical DNA sequence is the same as a natural one; (b) choice within each set of synonymous codons is made according to the mean base composition at the codon third position of a gene; this mean value is approximated by the mean base composition at the codon third position of the 158 *E. coli* genes used in our previous study. The expected number and standard deviation of occurrences of codons which belong to each codon group are analytically calculated using a statistical formula for the sum of the sample mean.

Fig. 1 presents D values of the four codon groups in the 25 *E. coli* genes. We found that the mode of linkage between the type I preference and gene-expression levels is different from the mode of linkage between the type II preference and gene-expression levels. The actual existence of two types of linkage was clearly demonstrated for the first time.

The fact elucidated here is as follows. Optimal codon groups (○ and ●) are preferred to non-optimal codon groups (□ and ■) in groups A, B and C, which have higher levels of gene expression. However, this type I preference diminishes and then almost disappears in genes with lower expression levels. For counterbalanced/uncounterbalanced dichotomy, Fig. 1 demonstrates two types of separate codon preferences in genes with higher and lower levels of gene expression, respectively. In genes with higher expression levels, counterbalanced codons, e.g. UAC, are preferred to uncounterbalanced codons. e.g. UAU; this preference is indicated by an upward arrow (*the upward arrow preference*). Conversely, in genes with lower levels of expression, uncounterbalanced codons are preferred; this is indicated by a downward arrow (*the downward arrow preference*). Strictly speaking, the existence of *the upward arrow preference* in an optimal codon group (○ and ●) in highly expressed genes is not significant. However, as a whole, Fig. 1 demonstrates the existence of two types of linkage between codon usage and gene-expression levels.

Kurland et al. comprehensively discussed the codon preference reported to date, and stated that the type II preference exists only in highly expressed genes and it may be a reflection of other phenomena [8,9]. But Fig. 1 actually demonstrated the existence of *the downward arrow preference* in weakly expressed genes.

The antagonism between selective pressure and randomization effect caused by mutation is another important subject [10–12]. In connection with this antagonism, two conclusions can be deduced from Fig. 1. (i) Ikemura proposed that codon choice becomes neutral or nearly neutral in genes with lower levels of expression, due to the dominance of the randomization effect caused by mutation [1]. This is essentially plausible for the type I preference. (ii) The same proposal does not, however, apply to the type II preference. In genes with lower levels of expression, the choice between counterbalanced codons and uncounterbalanced codons is not random. Therefore, the discovery of two types of linkage between codon usage and gene expression levels requires a novel explanation of how the antagonism

Table II

2 × 2 distinction with X·Z pair** of synonymous codons in *E. coli*

(1) optimal and counterbalanced codons (○)
CGU GCU GCA GUU GUA GGU
ACC AUC AAC UUC UAC GAA

(2) optimal and uncounterbalanced codons (●)
CUG CGC CCG CAG AAA GCG
GUG GGC ACU

(3) non-optimal and counterbalanced codons (□)
UUG CUU CUA CGA AGG CCU
CCA CAA AAG GGA ACG UAU

(4) non-optimal and uncounterbalanced codons (■)
UUA CUC CGG AGA CCC GCC
GUC GGG ACA AUU AUA AAU
UUU GAG

**X indicates the first letter in a codon, and Z indicates the third letter in the identical codon.

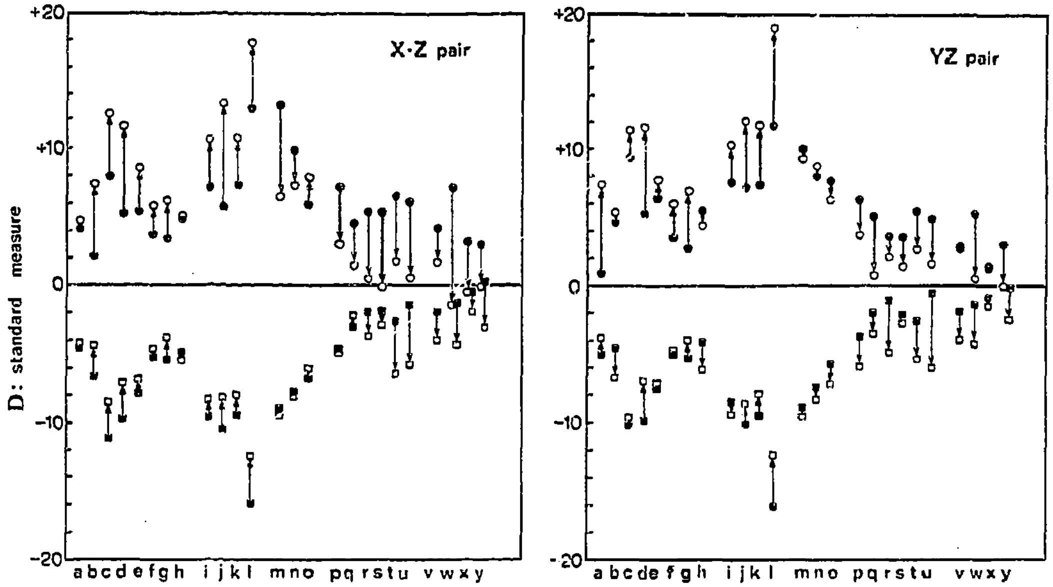

Fig. 1. Plots of *D*, standard measure of the number of occurrences of the following four codon groups for the 25 *E. coli* genes in Table I: (○) optimal and counterbalanced codons; (●) optimal and uncounterbalanced codons; (□) non-optimal and counterbalanced codons; (■) non-optimal and uncounterbalanced codons. Each of the letters (a)–(y) positioned under the panels denotes respectively a member of the 25 genes in the order named in Table I: (a) lpp, (b) rplL, (c) tufA, and so on. Gene groups with a higher expression level are positioned at the left half in each panel, and the ones with lower expression level positioned at the right half in each panel. For the direction of arrows, see text.

between selective pressure and randomization effect caused by mutation determines the two separate types of codon-choice.

Acknowledgement: This work was supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan.

REFERENCES

[1] Ikemura, T. (1985) Mol. Biol. Evol. 2, 13–34.
[2] Grosjean, H. and Fiers, W. (1982) Gene 18, 199–209.
[3] Wada, A. and Suyama, A. (1985) FEBS Lett. 188, 291–294.
[4] Wada, A. and Suyama, A. (1986) Prog. Biophys. Mol. Biol. 47, 113–157.
[5] Wada, A., Suyama, A. and Hanai, R.J., Mol. Evol., in press.
[6] Bernardi, G. and Bernardi, G. (1985) J. Mol. Evol. 22, 363–365.
[7] Yarus, M. and Folley, L.S. (1985) J. Mol. Biol. 182, 529–540.
[8] Kurland, C.G. (1987) Trends Biochem. 12, 126–128.
[9] Andersson, S.G.E. and Kurland, C.G. (1990) Microbiol. Rev. 54, 198–210.
[10] Kimura, M. (1981) Proc. Natl. Acad. Sci. USA 78, 5773–5777.
[11] Sharp, P.M. and Li, W.-H. (1986) J. Mol. Evol. 24, 28–38.
[12] Bulmer, M. (1987) Nature 325, 728–730.