Estimation of protein–production levels for ORFs found by *E. coli* and yeast genome projects, basing on levels of "optimal codon" usage, in connection with feasibility of their protein coding ability and with assignment to foreign–type genes.

Y. Nakamura
yanakamu@ddbj.nig.ac.jp

T. Fukagawa
tfukagaw@ddbj.nig.ac.jp

K. Sugaya
ksugaya@ddbj.nig.ac.jp

T. Ikemura
tikemura@ddbj.nig.ac.jp

Nat. Inst. Genet. and Grad. Univ. Advanced Studies, Mishima, Sizuoka 411 Japan

Choice among synonymous codons in both prokaryotic and eukaryotic genes is clearly non-random, although it does not affect the nature of proteins synthesized. Among genes of each unicellular organism, there is a clear similarity of codon choice patterns, regardless of gene function; "codon dialect" found for individual unicellular organisms [1]. Taxonomically related organisms have similar dialects but those distantly related have distinct ones. For example, characteristics of *E. coli* codon–choice (*E. coli* dialect) differ considerably from those of yeast *S. cerevisiae*, but are similar to those of *Salmonella*. By measuring cellular tRNA contents of these three species, we have shown that their codon dialect to be related to the isoaccepting tRNA population of individual organisms [2][3]. We also found the extent of codon bias (accent of codon dialect) is related to protein production level of each gene [1][2][3][4]. Codon usage in genes encoding abundant proteins is almost always much more dependent on tRNA content (strong accent) than that in moderately or poorly expressed genes (moderate accent); i.e., highly expressed genes almost always have a strong accent but those with moderate or low expression have a moderate accent. It has also been found that foreign–type genes such as those of transposons, plasmids, and viruses often have quite different codon patterns from the respective host dialect. To examine these features quantitatively, the frequency of "optimal" codon use was previously defined as

$$\mathrm{Fop} = \frac{\text{the numbers of "optimal" codons}}{\text{sum of the numbers of "optimal" and "nonoptimal" codons}}$$

中村 保一, 深川 竜朗, 菅谷 公彦, 池村 淑道：遺伝研, 総研大, 〒 411 静岡県三島市谷田 1111

The "optimal codon" of each amino acid is the codon translated by the most abundant isoacceptor and thus optimal for translational efficiency of the respective organism. Spectrum of the optimal codon has been shown to correspond strikingly well to the preferred codon of each organism, i.e. codon dialect [1]. The cellular content of a wide variety of *E. coli* proteins has been measured by Neidhardt. It is thus possible to analyze the correlation between the frequency of optimal codon use (Fop) and cellular protein contents. The definite correlation between Fop and protein content was revealed, showing Fop in each gene to be related to its production level [4]; Fop for highly expressed genes was almost always high (e.g., > 0.8), but low for weakly expressed genes. Fop of foreign–type genes was found significantly lower than those of intrinsic genes, supporting that the former genes do not necessarily use the *E. coli* dialect.

In this paper, we calculated Fop of 123 ORFs that were sequenced and registered by *E. coli* genome project of Japan (GenBank LOCUS name; ECO110K and ECO82K). Out of the 123 ORFs, 65 ORFs was assigned to known *E. coli* genes by the project. 43 ORFs was noted to have certain sequence homology with proteins of other species, and the residual 15 not to exhibit such homology. Basing on Fop levels, we could classify these less characterized 58 ORFs into highly–, moderately– or weakly–expressed genes, as well as into foreign–type genes. We calculated also Fop of 480 ORFs registered by *E. coli* genome projects of other countries and classified them as above. It should be mentioned that, in their registration, they often include relatively shorter ORFs than those of Japan. We analyzed feasibility of these short ORFs (and also of long ORFs) as the protein–coding region.

Genome G+C% is also an important factor to determine the codon usage pattern, especially the G+C% at the codon third position. In the foreign–type genes, the G+C% distribution at this position is much wider than that of the intrinsic *E. coli* genes and there is a significant number of AT–rich ones. This characteristic was also useful to assign foreign–type genes. We are now analyzing the "optimal codon usage" and the G+C% at the codon third position of *S. cerevisiae* ORFs registered by the yeast genome projects. It is worthwhile to mention that *S. cerevisiae* genes with low Fop often correspond nuclear genes encoding mitochondrial proteins or mating–related factors [4].

References

[1] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms" *Mol. Biol. Evol.*, Vol. 2, 13, 1985.

[2] T. Ikemura, "Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes" *J. Mol. Biol.*, Vol. 146, 1, 1981.

[3] T. Ikemura, "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes." *J. Mol. Biol.*, Vol. 158, 573, 1982.

[4] T. Ikemura, "Correlation between codon usage and tRNA content in microorganisms." In Transfer RNA in Protein Synthesis (CRC Press), 87, 1992.