Estimation of Protein-production levels in *Escherichia coli* Genes on the basis of Multivariate Diversity in Codon Usage

Shigehiko Kanaya[1]
kanaya@eie.yz.yamagata-u.ac.jp
Yoshihiro Kudo[1]
ykudo@eie.yz.yamagata-u.ac.jp
Yasukazu Nakamura[2]
yanakamu@ddbj.nig.ac.jp
Toshimichi Ikemura[2]
tikemura@ddbj.nig.ac.jp

[1] Dep. of Electric and Inf. Eng., Fac. of Eng., Yamagata Univ.
Yonezawa, Yamagata-ken 992,Japan

[2] Dep. of Evol. Genet., Natl. Inst. of Genet., and Grad. Univ. for Advanced Studies
Mishima, Shizuoka-ken 411, Japan

1 Introduction

Estimation of protein-production levels, along with peptide-motif search, gives valuable information for prediction of gene function. Choice among synonymous codons in both prokaryotic and eukaryotic genes is clearly non-random and the codon-usage pattern is undoubtedly important characteristic for determining protein-production levels of genes.

In the present work, we have constructed measures which reflect diversity of *E.coli* genes in codon usage by means of a combined method of relative representation for codon usages of genes and principal component analysis(PCA). Then, the factors of the widest scales constructed could be connected with protein-production levels. Protein production levels were estimated for 1500 CDSs proposed by the *E.coli* genome projects of Japan and USA.

2 Method

Taking the number of synonymous codons for each amino acid into consideration, a codon-usage pattern of the ith gene was represented by a 61-dimensional vector consisting of $x_{ij(m)}$ ($Eq.(1)$).

[1]金谷重彦, 工藤喜弘 : 山形大学工学部・電子情報工学科・生体システム講座, 992 米沢市城南 4-3-16
[2]中村保一, 池村淑道 : 国立遺伝学研究所・集団遺伝学研究所・進化遺伝研究部門, 441 三島市谷田 111

$$x_{ij(m)} = f_{ij(m)} / [\sum_{j=1}^{M(m)} f_{ij(m)} / M(m)] \quad (1)$$

where $f_{ij(m)}$ denotes frequencies of jth codon in the mth amino acid, and M(m) denotes the number of synonymous codons in the mth amino acid.

In order to assess diversity of genes in the 61-dimensional space representing codon-frequencies, PCA was applied to a reference data set consisting of 610 *E.coli K12* genes extracted from DDBJ (Release 18). To standerdize the scale of the principal components, Z'_k, each of them for the reference data set was normalized by *Eq.*(2).

$$Z_k = (Z'_k - Av[Z'_k]) / SD[Z'_k] \quad (2)$$

Here, Av$[Z'_k]$ and SD$[Z'_k]$ are average and standard deviation of Z'_k for the reference data, respectively. Z-parameters, Z_k, were correlated to the protein-production level by a regression analysis.

3 Results and Discussion

The first three components (Z'_1, Z'_2, and Z'_3) are significant axes according to the Kaiser's rule[1]. Of the 61 variables, twenty-five contribute positively to Z'_1. It should be stressed that most of them correspond to the optimal codons assigned by Ikmura[2] [3] based on experimental data of *E.coli* tRNA contents. Correlations between Z_k and protein-production levels were examined using *E.coli* protein contents in cells determined under four different growth conditions by Neidhardt and his colleagues[4]. A representative correlation is expressed by *Eq.*(3).

$$log(Rich) = 2.44 + 0.55 Z_1 (r = 0.74, n = 31) \quad (3)$$

where log(Rich) represents common logarithm of the amount of proteins (molecules) per genome grown in cells in rich medium. The protein-production levels for the 1500 CDSs were estimated by *Eq.*(3). Among them, CDSs with the highest protein-production levels are as follows; *tufA*, *tufB*, *mopA*, *rpsI*, *rplW*, *rplL*, *rpmB*, *rplD*, *fusA* *rpoC* and *atpD* whose production levels estimated are larger than 6.0x10^3 molecules per genome. We could also detect CDSs in CDS-undiscovered regions for *E.coli* DNA sequences by application of the present methodology.

References

[1] H.F.Kaiser, *Edu.Psychol.Meas.*, Vol.20,pp.141-151,1960.

[2] T.Ikemura, *J.Mol.Biol.*, Vol.2, pp.13-34,1985.

[3] T.Ikemura, In D.L.Hatfield, B.J.Lee, R.M.Pirlte (eds), *Transfer RNA in protein synthesis*,CRC Press, London, pp.87-111, 1992.

[4] R.A.VanBogelen et al, *Electrophoresis*, Vol.13, pp.1014-1054,1992.