# Relationship between Gene Function and Codon Usage in *Escherichia coli* on the basis of Principal Component Analysis

Shigehiko Kanaya[1]
kanaya@eie.yz.yamagata-u.ac.jp

Toshimichi Ikemura[2]
tikemura@ddbj.nig.ac.jp

Yoshihiro Kudo[3]
ykudo@eie.yz.yamagata-u.ac.jp

[1] Department of Electrical and Information Engineering, Faculty of Engineering,
Yamagata University
Yonezawa, Yamagata 992 Japan

[2,3] National Institute of Genetics; The Graduate Univ. for Advanced Studies
Mishima, Shizuoka 311 Japan

## Abstract

Rapid advance in experimental and theoretical techniques in genetics has afforded both abundant and useful information, and also difficulty in data processing and interpretation. Indeed, by increase of nucleotide sequence data accumulated by pioneers in the field, some relations between gene function and codon usage have been clarified. As the data increase, it becomes more difficult to characterized these relation systematically, because we must recognize the 64 kinds of codon frequencies on a great number of genes. It is expected that multivariate analysis make it possible to overcome this difficulty and to characterize genes in terms of codon usage.

In order to investigate the factors involved in the diversity of *Escherichia coli* genes in terms of codon usage, and clarify some relations between codon usage and gene function, we have constructed a data set consisting of about two thousand genes with the following information : gene name, gene function and codon usage. In the present paper, we recognize some relations between codon usage and gene function by means of a principal component analysis of this data set.

To exclude the effect of amino acid compositions on codon usage, firstly, frequencies of codons in each synonymous group were normalized to unity, and all of data were

[1,3] 金谷重彦，工藤喜弘：山形大学工学部電子情報工学科, 992 米沢市城南 4-3-16

[2] 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, 411 三島市谷田 111

represented in form of a matrix, $\mathbf{X}_{ij}$, where $i$=1,2,...,N and $j$=1,2,...,M (N and M denote the number of genes and codons used, respectively). For an $i$th gene ($i$=1,2,..,N), the vector consisting of the normalized codon frequencies, $(x_{i1}, .., x_{ij}, .., x_{iM})$, is transformed to the vector consisting of principal components, $(z_{i1},..,z_{ij},..,z_{iM})$, according to the following conditions. (1) A correlation of principal components between $Z_k$ and $Z_{k'}$ is zero, and (2) the first principal component, $Z_1$, is the linear combination of the variables, $X_j$, with the largest variance, and the second principal component, $Z_2$, is the linear combination with the second largest variance, and so on. $Z_k = b_{k1}X_1 + ...... + b_{km}X_M$ ($k$=1,2,...,M) where $\sum_{j=1}^{M} b_{ij}^2 = 1$.

By scattering genes on a map consisting of the first $k$ principal components, we can comprehend proximities among them from a viewpoint of the structure of codon usage.

We have assembled a data set with the following information: gene name, category name [M.Riley, Microbiol Rev.,**54**,862-952,1992] of cellular function of the gene product, genomic map position, and complete coding sequence extracted from DDBJ (Release 18,1994). The data set consists of 1528 genomic coding genes, 26 transposon-related genes, 106 plasmid genes, and 574 function-unknown genomic open reading frames (simply called ORFs). The first two components (PC1 and PC2) account for more than 10% of the original variance (24.8% and 14.5%, respectively). It is observed that the loadings of the PC1 is negatively correlated to the preference codons reported[T.Ikemura, In Hatfield,D.L., Lee,B.J. and Pirtle,R.M.(Ed.), Transfer RNA in protein synthesis.,pp.87-111, CRC Press, London.]. This suggests that the largest diversity of genomic genes is mainly explained in terms of the preference codon usage.

In order to investigate some relations between PC1 value ($z_{i1}$) of gene and gene functions, we examined distribution of the genes in PC1 for each of the categories of gene function. An approximate median of the PC1 in the 1528 genomic genes are used as a critical to characterize the gene distributions for the categories. Most of genes contained in the following four categories are of PC1 values larger than the critical: IE, ATP proton motive force interconversion; IIB1, Biosynthesis of purine ribonucleotides; IIIA2, Ribosomal proteins and their modification; and IIIA4, Aminoacyl-tRNA synthetases and their modifications. The categories involve genes which are constantly utilized in the cell. On the other hand, most of genes contained in the following seven categories are of PC1 values smaller than the critical: IIA4, Biosynthesis of aromatic amino acid family; IID, Biosynthesis of Cofactors; IVB, Murein sacculus, IVC,Surface polysaccharides and antigens; VIB, Phage-related functions and prophages; VIC, Colicin-related functions; VID, Plasmid-related functions. The categories involve genes which are utilized in some specific cases or are constantly utilized in a small amount. These may be reflected in the diversity of codon usage in E.coli genes.

The first two components of the transposon-related genes, plasmid genes, and ORFs are calculated using the eigen vectors of the PC1 ($k = 1$) and PC2 ($k = 2$), $(b_{k1}, .., b_{km})$ ,of genomic genes. The configuration by the first two components indicates that all of the genes of the transposon-related and plasmid genes are located in a different region from a region that the genes involved in the four categories are concentrated. Since six of the 574 ORFs are located in this specific region, it is plausible that these are the constantly utilized genes in the cell. In conclusion, this method may play an important role in the determination whether a gene is constantly utilized or not in the cell.