Assessment of Species-specific Diversity of Genes in Codon Usage

Shigehiko Kanaya$^1$  
kanaya@eie.yz.yamagata-u.ac.jp
Yasukazu Nakamura$^2$  
yanakamu@ddb.j.nig.ac.jp
Yoshihiro Kudo$^1$  
ykudo@eie.yz.yamagata-u.ac.jp
Toshimichi Ikemura$^2$  
tikemura@ddb.j.nig.ac.jp

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

Mishima, Shizuoka-ken 411, Japan

1 Introduction

Species-specific diversity of genes in codon-usage is fundamentally important characteristic for determining suitability of genes in genomes and estimating protein-production levels of genes.

We have developed measures which reflect diversity of genes in codon usage by means of a multivariate statistical method and assess species-specific diversity of genes in codon usage for four organisms, Bacillus subtilis(Bs), Escherichia coli(Ec), Pseudomonas aeruginosa(Pa), and Salmonella typhimurium (St).

2 Method

To exclude effects of amino acid compositions for the respective genes and take the number of synonymous codons for each amino acid into consideration,a codon-usage pattern of the $i$th gene was represented by a 61-dimensional vector consisting of $x_{ij(m)}$ (Eq.(1)).

$$x_{ij(m)} = \frac{f_{ij(m)}}{\sum_{j=1}^{M(m)} f_{ij(m)}}$$

(1)
where $f_{j(m)}$ denotes frequencies of $j$th codon in the $m$th amino acid, and $M(m)$ denotes the number of synonymous codons in the $m$th amino acid.

In order to assess species-specific diversity of genes in the 61-dimensional space representing codon-frequencies, principal component analysis was applied to a data set for each species. The specific-diversity is assessed by contribution (called factor loadings expressed by Eq. (2)) of the $j$th codon frequency to the $k$th principal component ($PC$), $Z_k$.

$$r(Z_k, X_j) = \frac{\text{Cov}[Z_k, X_j]}{(\text{Var}[Z_k], \text{Var}[X_j])^{1/2}}$$

where Cov[A, B] and Var[A] denote covariance between two variables, A and B, and variance on variable A, respectively.

### 3 Results and Discussion

To exclude confusion derived from differences of strains, only sequences annotated as *B.subtilis 168*, *E.coli K12*, *P.aeruginosa PAO*, and *S.typhimurium LT2*, were extracted from bacterial sequences in DDBJ (Release 18 and 21). We selected from them respective data sets consisting of 150 genes for Bs, 610 genes for Ec, 130 genes for St, and 98 genes for Ps, respectively, which are longer than 500 nucleotides.

The significant components according to the Kaiser's rule[1] are the first four PCs for Bs, three PCs for Ec, two PCs for St, and three PCs for Ps. In Ec, most of the original variables ($X_j$) which contribute positively to PC1 ($r(Z_k, X_j) > 0$) correspond to the optimal codons assigned by Ikemura[2][3] and PC1 ($Z_1$) for genes is highly correlated to optimal codon index (Fop[3], $r=0.95$). PC2 ($Z_2$) is correlated to GC% at the codon third position. These indicate that the diversity by optimal codon is much larger than that by the codon third position.

The highest five correlations between factor loadings among species are 0.98 between PC1s for Ec and St, 0.94 between PC2s for Ec and St, 0.70 between PC2s for Ec and Bs, 0.70 between PC2s for St and Bs, and 0.63 between PC1s between St and Ps. The structure of the diversity is highly conserved between Ec and St regardless of the numbers of samples.

PC2s for Bs and St are highly correlated to GC% at the codon third position (0.782 and 0.805). Because PC1s for four species are not highly correlated to GC% at the codon third position, the most important factors on the diversity of genes in codon usages are not correlated to selection of bases at the third codon position. In Bs, most of the genes involved in translation machinery and synthesis of purine-nucleotides (these are representative highly expressed genes in Ec) have positive large $Z_1$, suggesting that PC1 reflects preferential codon structure in Bs. The present methodology is fundamentally applicable to other organisms and may give clues to estimate species-specific diversity of genes in codon usage.

### References

