Gene Classification by Self-Organization Mapping of Codon Usage in Bacteria with Completely Sequenced Genome

Shigehiko Kanaya [145] kanaya@eie.yz.yamagata-u.ac.jp
Yoshihiro Kudo [1] ykudo@eie.yz.yamagata-u.ac.jp
Takashi Abe [1] a95510@eie.yz.yamagata-u.ac.jp
Takanori Okazaki [1] a95550@eie.yz.yamagata-u.ac.jp
Carlos Del Carpio [2] carlos@translell.eco.tut.ac.jp
Toshimichi Ikemura [3] tikemura@ddbj.nig.ac.jp

[1] Department of Electrical Information Engineering, Faculty of Engineering, Yamagata University, Yonezawa, Yamagata 992-8510, Japan
[2] Department of Ecological Engineering, Faculty of Engineering, Toyohashi University of Technology, Toyohashi, Aichi 441-8580, Japan
[3] Department of Population Genetics, National Institute of Genetics, and the Graduate University for Advanced Studies, Mishima, Shizuoka, 441-8540, Japan.
[4] Department of Developmental Genetics, National Institute of Genetics
[5] CREST, JST (Japan Science and Technology)

1 Introduction

The goal of our study is to understand the heterogeneous codon usage among intra- and inter-species specific codon usage in view of biological functions. In the previous study, we heve developed the procedure for estimating species-specific heterogeneous codon usage on the basis of principal component analysis [1, 2, 3] and applied this procedure for estimating ORFs in the complete nucleotide sequences of ϕCTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa* [4]. In the present study, we propose a procedure to categorize genes in interspecies. Codon usage in a gene can be described by a vector consisting of codon-usage frequencies. The categorization procedure is based on an artificial self-organization algorithm of the vectors representing codon usage in genes. The self-organization process, suggested originally by Kohonen [5, 6, 7], has the special property of effectively creating spatially organized "internal representations" of various features of input signals and their abstractions. We have applied the methodology to classification of genes in sixteen bacteria whose genomes have been sequenced completely.

2 Methodology

Reference vectors are placed in two-dimensional array (called Self-organization map (SOM)). Here $\mathbf{w}^{uv}$ represents the vector at the uvth position of the array. The initial reference vectors are set by the cumulative codon usage patterns for bacteria, and placed using the first and second axes obtained by a principal component analysis. The vectors are adjusted by the competitive learning as follows. Firstly, we find $\mathbf{w}^{CuCv}{}_i$ closest to $\mathbf{x}_i$, where $\mathbf{x}_i$ represents the codon usage pattern for the ith gene. Then, the vectors in the region between Cu-β and Cu+β, and between Cv-β and Cv+β are adjusted by *Eq*.(1).

$$\mathbf{w}^{uv(New)} = \mathbf{w}^{uv} + \alpha(\mathbf{x}_i - \mathbf{w}^{uv}) \tag{1}$$

where $0<\alpha<1$. The two parameters α and β are decreased as the training process proceeds. The following error function in *Eq* .(2) is used for assessing the efficiency of organized "internal representations" of input vectors.

$$E = \sum_i D(\mathbf{x}_i, \mathbf{w}^{CuCv}{}_i)^2 \tag{2}$$

where D($\mathbf{x}_i$, $\mathbf{w}^{CuCv}{}_i$) represents Euclidean distance between the two vectors. Then, the ith gene is classified into $CuCv$-position of the array.

3 Results and Discussion

The data set consists of 29596 ORFs with longer than 299 nts in length (1489 for *Aquifex aeolicus*, 2088 for *Archaeoglobus fulgidus*, 3788 for *Bacillus subtilis*, 772 for *Borrelia burgdorferi*, 833 for *Chlamydia trachomatis*, 3913 for *Escherichia coli*, 1572 for *Haemophilus influenzae*, 1392 for *Helicobacter pylori*, 1522 for *Methanobacterium thermoautotrophicum*, 1646 for *Methanococcus jannaschii*, 450 for *Mycoplasma genitalium*, 657 for *Mycoplasma pneumoniae*, 3675 for *Mycobacterium tuberculosis*, 1973 for *Pyrococcus horikoshii*, 2909 for *Synechocystis sp.*, and 917 for *Treponema pallidum*). The training was done at five cycles for the data set, that is, 29596 x 5 iterations. The two parameters are set as follows, α=0.40, 0.35, 0.30, 0.25, 0.20, and β=10, 8, 6, 4, 2 for five cycles.

Fig. 1 shows the number of genes classified by SOM. Vertical and horizontal axes are initially determined by the first and second principal components of cumulative codon usage for the sixteen bacteria. Dots on SOM represent 1 to 9 genes. This figure indicates very biased distribution of genes on SOM. More than hundred genes are clusterized at thirteen positions (which are underlined in the SOM.) Gene for *E. coli* are dominant at the position 37 in vertical axis and 15 in horizontal axis (denoted by v37-h15) and genes for *H. pylori* are dominant at the position of v61-h24. Genes for *B. subtilis* are dominant at four positions, v100-h1, v100-h4, v100-h5, and v100-h41. Genes for *A. fulgidus* are dominant at two positions, v78-h43 and v80-h44. Genes for *A. aeolicus* are dominant at four positions, v81-h45 and v82-h45, v70-h59, and v71-h59. Multi-species genes are clusterized at one position, v71-h35. In this way, SOM makes it possible to understand the biased distribution of genes in codon usage patterns. The biological meanings of the reference vectors will be discussed in poster at GIW98.

References

[1] Kanaya, S., Kudo, Y., Nakamura, Y. and Ikemura, T., Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage, *CABIOS*, 12:213–225, 1996.

[2] Kanaya, S., Kudo, Y., Suzuki, S., and Ikemura, T., Systematization of species-specific diversity of genes in codon usage: Comparison of the diversity among bacteria and prediction of protein production levels in cells, *Genome Informatics 1996*, 61–71, 1996.

[3] Kanaya, S., Okumlura, T., Miyauchi, M., Fukasawa, H., and Kudo, Y., Assessment of protein coding sequences in *Bacillus subtilis* genome using species-specific diversity of genes in codon usage based on multivariate analysis, *Res. Communications in Biochem. and Cell & Mol. Biol.*, 1:82–92, 1997.

[4] Nakayama, K., Kanaya, S., Ohnishi, M., Terawaki, Y., and Hayashi, T., The complete nucleotide sequence of ϕCTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa.*, *Mol. Biol.*, (in press), 1998.

[5] Kohonen, T., The self-organizing map, *Proc. IEEE*, 78:1464–1480, 1990.

[6] Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J., Engineering applications of the self-organizing map, *Proc. IEEE*, 84:1358–1384, 1996.

[7] Kohonen, T., *Self-Organizing Maps*, Springer-Verlag Berlin Heidelberg, 1995.

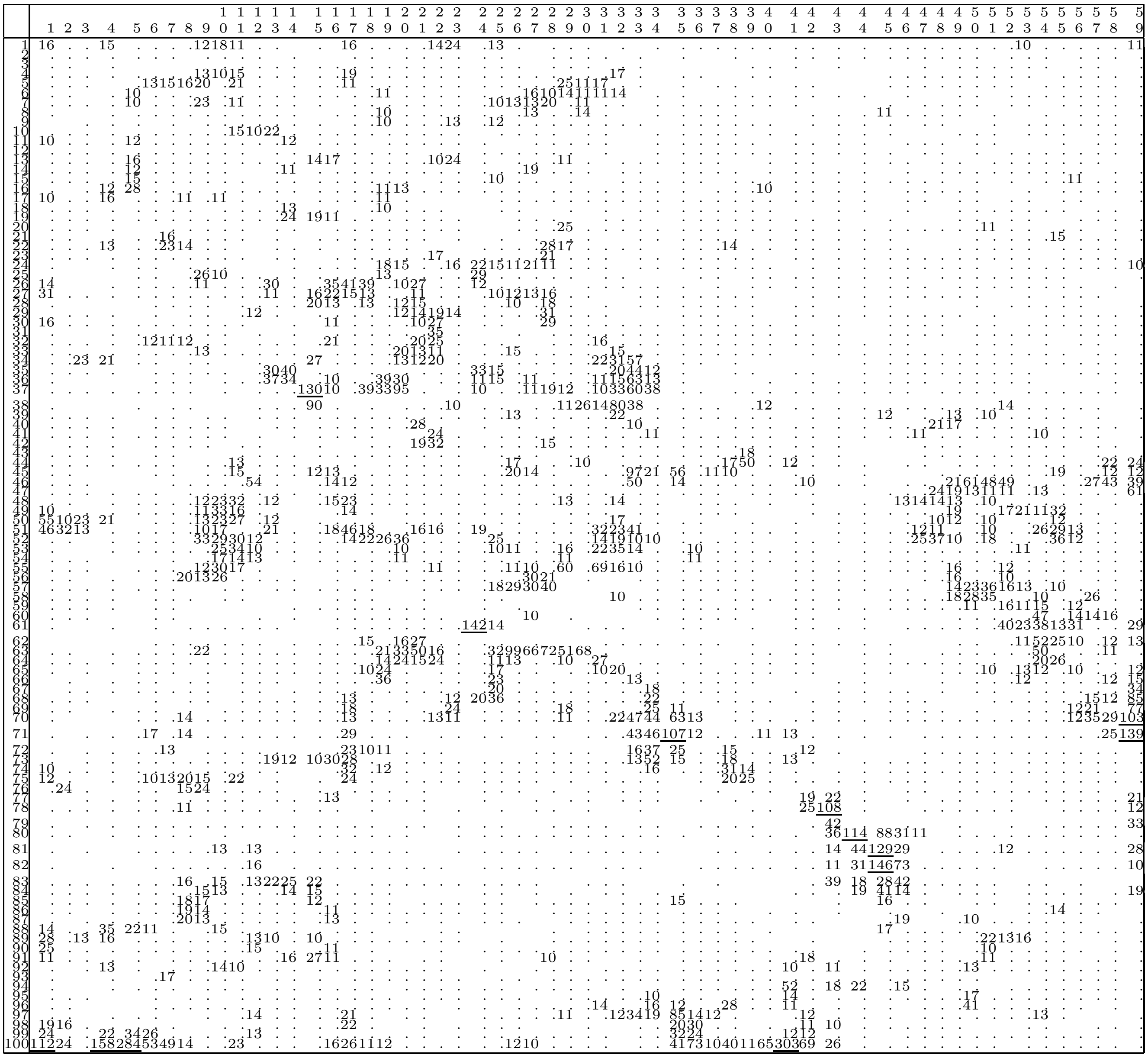

Figure 1: Classification by SOM (Details are described in the text).