Gene Classification Method Based on Batch-Learning SOM

Takashi Abe [1] m99804@eie.yz.yamagata-u.ac.jp
Shigehiko Kanaya [14] kanaya@eie.yz.yamagata-u.ac.jp
Makoto Kinouchi [1] kinouchi@eie.yz.yamagata-u.ac.jp
Yoshihiro Kudo [1] ykudo@eie.yz.yamagata-u.ac.jp
Hirotada Mori [45] hmori@gtc.aist-nara.ac.jp
Hideo Matsuda [46] matsuda@ics.es.osaka-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] CREST, JST (Japan Science and Technology)
[5] Research and Education Center for Genetic Information, Nara Institute of Science and Technology,8916-5 Takayama, Ikoma, Nara 630-0101, Japan
[6] Department of Informatics and Mathematical Science, Gradute School of Engineering Science Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

1 Introduction

We previously examined species-specific heterogeneous codon usage by principal component analysis [1], and similarity of genes for several species in codon usage by an artificial self-organization algorithm originally proposed by T. Kohonen [2]. The self-organizing map (SOM) implements a characteristic nonlinear projection from the high-dimensional space of input signals onto a low-dimensional array of neurons [2]. The projection obtained by the standard learning algorithm of SOM is highly dependent on learning order of inputs. Consequently, fine classification is obtained for the latter inputs. In the present study, we report a batch-learning algorithm for constructing self-organizing map and examined codon-usage similarity of genes among bacteria.

2 Methodology

2.1 Initial neuron weights

Neurons were arranged in two dimensional lattice denoted by $s(= 0, 1, ..., S-1)$ and $t(= 0, 1, ..., T-1)$. Initial neuron vectors were determined using principal component analysis. The number of neurons in the first dimension was set to 100. The neuron number of the second dimension (T) was set to the nearest integer greater than $S\sigma_2/\sigma_1$, and the neuron vectors on the stth lattice was determined by

$$\mathbf{w}_{st} = \mathbf{x}_{av} + 5\sigma_1\{\mathbf{b}_1\frac{s - S/2}{S} + \mathbf{b}_2\frac{t - T/2}{T}\}. \quad (1)$$

Here, $\mathbf{x}_{av}$ is the average vector for codon usage patterns of genes; $\mathbf{b}_1$ and $\mathbf{b}_2$ represent eigen vectors for the first and second principal component; σ_1 and σ_2 represent variances for the two components.

2.2 Classification of genes into st-lattice point and updating neuron vectors

We classified the ith gene into st-lattice point satisfied with $s' - \beta(r) \leq s \leq s' + \beta(r)$, and $t' - \beta(r) \leq t \leq t' + \beta(r)$. Here, $s't'$ is the neuron vector with the smallest Euclidean distance of $\mathbf{x}_i$ to $\mathbf{w}_{st}$ among $S \times T$ neuron vectors, and $\beta(r) = \max\{0, 100/4 - r\}$ is the coefficient for region affected by $\mathbf{x}_i$ in the rth cycle of learning process. After all genes were classified into lattice points, neuron vectors were updated by

$$\mathbf{w}_{st}^{(new)} = \mathbf{w}_{st} + \alpha(r)\{\sum_{i=1}^{N_{st}} \frac{\mathbf{x}_i}{N_{st}} - \mathbf{w}_{st}\}. \qquad (2)$$

Here $\mathbf{x}_i$ represents codon usage pattern of the ith gene classified into the stth lattice point, and N_{st} is the total number of genes classified into this lattice point. $\alpha(r)$ was determined by

$$\alpha(r) = \max\{0.01, 0.06(1 - r/100)\}. \qquad (3)$$

2.3 Classification of genes

After learning process, genes were plotted in the lattice points with the nearest Euclidean distance of its codon usage pattern to neuron vector.

3 Results and Discussion

The neuron vectors were developed using codon usage patterns for 29596 ORFs with longer than 299 nts. The following three remarks were obtained.

1. Projection obtained by this method is not dependent on learning order of inputs, and its learning effectiveness estimated by Eq. (4) for batch learning is nearly the same as that for iterative learning.

$$\mathbf{Q} = \sum_{i=1}^{N} \{\mathbf{x}_i - \mathbf{w}_{s't'}^{(i)}\} \qquad (4)$$

Here, $\mathbf{w}_{s't'}^{(i)}$ represents neuron vector closest to $\mathbf{x}_i$ among them. **2**. Fig. 1 shows projection of genes into neurons. Neurons including genes for only one species are denoted by capitals. ('A', *A. fulgidus*; 'B', *A. aeolicus*; 'C', *B. burgdorferi*; 'D', *B. subtilis*; 'E', *C. trachomatis*; 'F', *E. coli*; 'G', *H. pylori*; 'H', *H. influenzae*; 'I', *M. jannaschii*; 'J', *M. thermoautotrophicum*; 'K', *M. tuberculosis*; 'L'. *M. genitalium*, 'M'; *M. pneumoniae*; 'N', *P. horikoshii*; 'O', *Synechocystis sp.*; 'P', *T. pallidum*). Most of neurons are characterized by only one species. This indicates that codon usage patterns are intrinsic for these 16 bacteria. **3**. The first array is roughly explained by G+C% at the codon third position. (data not shown). Upper and lower sides correspond to low and high G+C%, respectively.

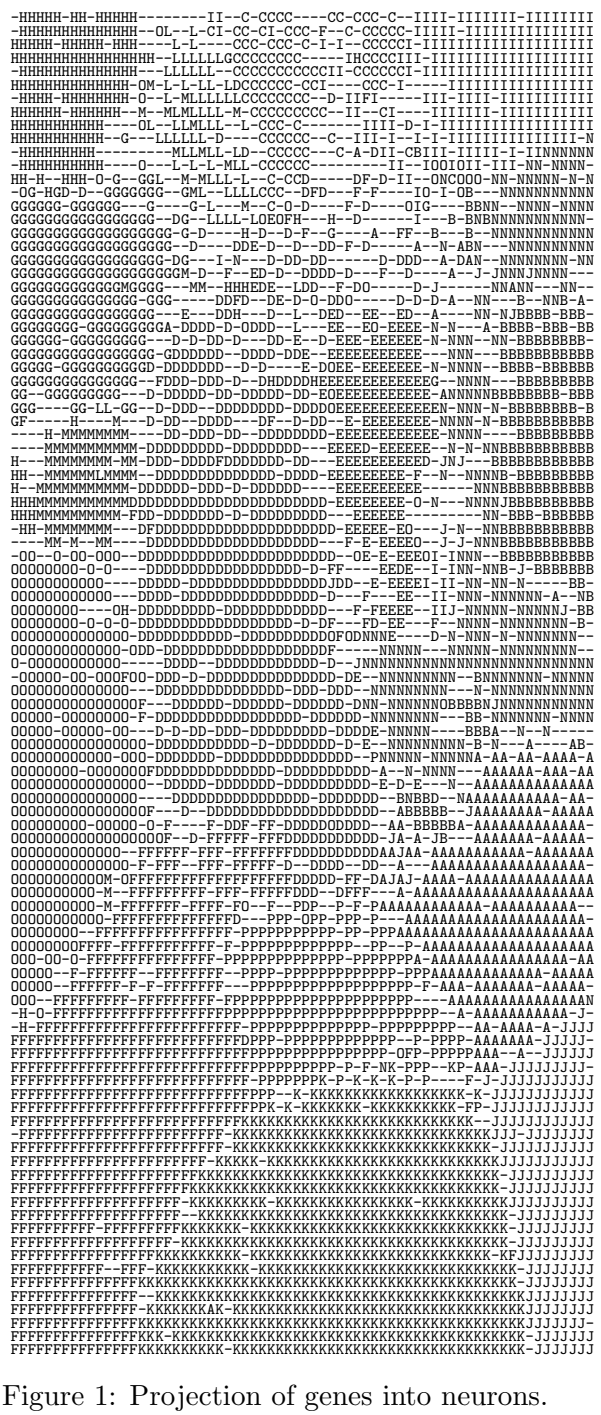

Figure 1: Projection of genes into neurons. Vertical and horizontal axes correspond to the first and second dimensions in neuron array, respectively.

References

[1] Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T., Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs, *GENE*, in press, 1999.

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