

A Novel Bioinformatic Strategy for Unveiling Hidden Genome Signatures of Eukaryotes: Self-Organizing Map of Oligonucleotide Frequency

Takashi Abe^{1,2,3} **Shigehiko Kanaya**^{3,4,5} **Makoto Kinouchi**^{3,5,6}
tajaabe@lab.nig.ac.jp skanaya@gtc.aist-nara.ac.jp kinouchi@yz.yamagata-u.ac.jp
Yuta Ichiba^{1,3} **Tokio Kozuki**^{2,3} **Toshimichi Ikemura**^{1,3}
yichiba@lab.nig.ac.jp kozuki@xanagen.com tikemura@lab.nig.ac.jp

- ¹ Department of Population Genetics, National Institute of Genetics, Mishima, Shizuoka-ken 411-8540, Japan
- ² Xanagen Inc., Sakado, Takatsu-ku, Kawasaki, Kanayagawa-ken 213-0012, Japan
- ³ ACT-JST (Japan Science and Technology Corp.)
- ⁴ Department of Bioinformatics and Genomes, Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara-ken 630-0101, Japan
- ⁵ CREST JST (Japan Science and Technology Corp.)
- ⁶ Department of Bio-System Engineering, Faculty of Engineering, Yamagata University, Yonezawa, Yamagata-ken 992-8510, Japan

Abstract

With the increasing amount of available genome sequences, novel tools are needed for comprehensive analysis of species-specific sequence characteristics for a wide variety of genomes. We used an unsupervised neural network algorithm, Kohonen's self-organizing map (SOM), to analyze di- and trinucleotide frequencies in 9 eukaryotic genomes of known sequences (a total of 1.2 Gb); *S. cerevisiae*, *S. pombe*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *Fugu*, and rice, as well as *P. falciparum* chromosomes 2 and 3, and human chromosomes 14, 20, 21, and 22, that have been almost completely sequenced. Each genomic sequence with different window sizes was encoded as a 16- and 64-dimensional vector giving relative frequencies of di- and trinucleotides, respectively. From analysis of a total of 120,000 nonoverlapping 10-kb sequences and overlapping 100-kb sequences with a moving step size of 10 kb, derived from a total of the 1.2 Gb genomic sequences, clear species-specific separations of most sequences were obtained with the SOMs. The unsupervised algorithm could recognize, in most of the 120,000 10-kb sequences, the species-specific characteristics (key combinations of oligonucleotide frequencies) that are signature representations of each genome. Because the classification power is very high, the SOMs can provide fundamental bioinformatic strategies for extracting a wide range of genomic information that could not otherwise be obtained.

Keywords: self-organizing map, oligonucleotide frequency, genome signatures

1 Introduction

Multivariate analysis methods such as factor corresponding analysis and principal component analysis (PCA) have been successfully used for analyzing heterogeneities of genome characteristics such as species-specific codon usage [5, 11, 12, 17, 19]. However, clustering powers of the conventional multivariate analysis methods become rather poor when a large amount of sequences in a wide range of species are analyzed. Here, we used a novel neural-network algorithm with high clustering power, a self-organizing map (SOM). The unsupervised neural network algorithm is an effective tool for

clustering and visualizing high-dimensional data; it converts complex nonlinear relations among high-dimensional data into simple geometric relations that can be viewed in two dimensions [14, 15, 16]. This method can be used to identify categories from raw data with a high clustering power and trace factors reflected in individual categories. We and others have used SOMs to characterize codon usage patterns of a wide variety of bacteria [1, 9, 10, 20]. We introduced a new feature to the SOM for genome sequence study, that makes the learning process and the resultant map structures independent of the input order of data [1], and we characterized codon usage of 60,000 genes from 29 bacterial species [9].

In addition to protein-coding information, genome sequences contain a wealth of information of interest in many fields of biology from molecular evolution to genome engineering. From biology we know that during evolution genome sequences of individual species have been structured in both distinct and common ways. G+C% has been used as a fundamental characteristic of individual genomes, but the G+C% is apparently too simple a parameter to differentiate a wide variety of genomes. Oligonucleotide frequency can be used to distinguish genomes because oligonucleotide frequencies vary significantly among genomes; dinucleotide frequencies, for example, were shown to be genome signatures for both prokaryotes and eukaryotes [13]. Comprehensive analyses of oligonucleotide frequencies in eukaryote genomes may provide fundamental knowledge of individual genomes, namely, key combinations of oligonucleotides responsible for the biological properties of the different genomes. In the present study we applied Kohonen's self-organizing map (SOM) to create graphical representations of oligonucleotide frequencies from which we could extract a wide range of genomic information. We constructed SOMs for di- and trinucleotide frequencies in 10- and 100-kb genome segments of 9 eukaryotic genome sequences currently available (a total of 1.2 Gb). The resulting two-dimensional projections of the 16- or 64-dimensional sequence space revealed clear species-specific separations. Comparative analysis of interspecies oligonucleotide frequencies could provide insight into hidden signatures in the genome sequences that have established during evolution.

2 Method and Results

The DNA sequences of *S. cerevisiae*, *S. pombe*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *P. falciparum*, and human were obtained from the GenBank web site (<http://www.ncbi.nlm.nih.gov/Genbank/>), and those of *Fugu* and rice were from the site (<http://fugu.hgmp.mrc.ac.uk/>) and the site (<http://rgp.dna.affrc.go.jp/>), respectively. In calculation of oligonucleotide frequency for a window, when the number of undermined nucleotides (Ns) exceeded 10% of the window size, the respective genomic fragments were omitted from the SOM analysis. In the case where the number of Ns was less than 10% of the window size, oligonucleotide frequencies were normalized with the length without Ns and included in the SOM analysis. The self-organizing map (SOM) is a neural-network algorithm that implements a characteristic nonlinear projection from the high-dimensional space of input signals onto a two-dimensional array of weights [15, 16, 17]. The weights (\mathbf{w}_{ij}) are arranged in a two-dimensional lattice denoted by i ($=0, 1, \dots, I-1$) and j ($=0, 1, \dots, J-1$). The standard learning algorithm for a two-dimensional SOM is as follows: Step 1, setting initial weights \mathbf{w}_{ij} ; Step 2, finding weight $\mathbf{w}_{i'j'}$ with minimum distance to the k th input vector \mathbf{x}_k ($k = 1, 2, \dots, N$); and Step 3, updating the $i'j'$ th weight vectors including set $S_{ij}=\{i'-\beta \leq i \leq i'+\beta, j'-\beta \leq j \leq j'+\beta\}$ by the following:

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij} + \alpha(\mathbf{x}_k - \mathbf{w}_{ij}) \quad (1)$$

Here, α and β are called learning coefficients. In Step 4, input vector \mathbf{x}_k is classified into weight \mathbf{w}_{ij} with the nearest distance after learning processes of Steps 2 and 3.

Learning process of the conventional SOM was designed to be dependent of the order of input data because it was developed for the study of memory. On the basis of the batch-learning SOM, we modified the conventional SOM for genome informatics to make the learning process and resulting map

independent of the order of data input. Results of codon-usage analyses for bacterial genomes with the conventional and modified methods were reported previously by [1, 9, 10]. In the modified method used also in the present study, weight vectors were initialized under PCA, which is a statistical method that performs a linear mapping to extract optimal features from an input distribution in the mean squared error sense and can be used by self-organizing neural networks to form unsupervised neural preprocessing modules for classification problems [14, 15, 16]. Hence, the initial weight vectors are set based on the widest scale of the sequence distribution in the oligonucleotide frequency space with PCA. Weights in the first dimension were arranged into two hundred lattices ($I = 200$) corresponding to the width of five times the standard deviation ($5\sigma_1$) of the first principal component. The second dimension (J) was defined by the nearest integer greater than $100\sigma_2/\sigma_1$. The weight vector on the ij th lattice was represented as follows:

$$\mathbf{w}_{ij} = \mathbf{x}_{av} + 5\sigma_1[\mathbf{b}_1(i - I/2)/I + \mathbf{b}_2(j - J/2)J] \quad (2)$$

Here, \mathbf{x}_{av} is the average vector for codon usage patterns; \mathbf{b}_1 and \mathbf{b}_2 are eigen vectors for the first and second principal components. In Step 2, the Euclidean distances between the input vector \mathbf{x}_k and the weight vector \mathbf{w}_{ij} were calculated, and \mathbf{x}_k was classified into the weight vector (called $\mathbf{w}_{i'j'}$) with the smallest distance among them. After classifying all input vectors into the weight vectors, updating process was done according to Step 3.

In Step 3, the ij th weight vector was updated by

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij} + \alpha(r) \left(\sum_{\mathbf{x}_k \in S_{ij}} \mathbf{x}_k / N_{ij} - \mathbf{w}_{ij} \right) \quad (3)$$

Here, the components of Set S_{ij} are input vectors classified into $\mathbf{w}_{i'j'}$ satisfying $i - \beta(r) \leq i' \leq i + \beta(r)$ and $j - \beta(r) \leq j' \leq j + \beta(r)$. The two parameters $\alpha(r)$ and $\beta(r)$ are learning coefficients for the r th cycle, and N_{ij} is the number of components of S_{ij} . The learning process is monitored by the total distance between \mathbf{x}_k and the nearest weight vector $\mathbf{w}_{i'j'}$, represented as

$$Q(r) = \sum_{k=1}^N \{ \|\mathbf{x}_k - \mathbf{w}_{i'j'}\|^2 \} \quad (4)$$

where N is the total number of sequences analyzed.

The SOM program used for the sequence analyses was obtained from Xanagen Inc. (Sakado, Takatsu-ku, Kawasaki-shi, Kanagawa-ken, Japan).

3 Results

3.1 Species-Specific Oligonucleotide Frequencies in 9 Eukaryote Genomes

SOMs were constructed with di- and trinucleotide frequencies (Figs. 1 and 2, respectively) in non-overlapping 10-kb genomic segments and in overlapping 100-kb sequences with a moving step size of 10 kb, that were derived from 9 eukaryote genomes (a total of 1.2 Gb); *S. cerevisiae*, *S. pombe*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *Fugu*, and rice, as well as *P. falciparum* chromosomes 2 and 3, and human chromosomes 14, 20, 21, and 22, that have been almost completely sequenced. As the first step to obtain the initial weight vectors, frequencies of a total of 120,000 nonoverlapping 10-kb sequences and overlapping 100-kb sequences with a moving step size of 10 kb, derived from a total of 1.2 Gb eukaryotic sequences were analyzed by principal component analysis (PCA) as was done in the codon analysis [9]. This is based on the knowledge that multivariate analyses including PCA successfully classified gene sequences into groups corresponding to known biological categories [5, 11, 12, 19], when numbers of sequences and species were much smaller than those analyzed here. This also saved considerably the calculation time of learning processes. After 60 learning cycles, oligonucleotide frequencies of genome

sequences were reflected as the final weight vectors in the respective SOMs. Lattices that include sequences from a single species are indicated by a letter corresponding to the species name, those including sequences of more than one species are indicated by black dots, and those with no sequences are indicated in white. The sequences were separated into nonoverlapping species-specific zones (Figs. 1 and 2). This shows that the SOM separations obtained without any species information closely fit the sequence clustering among species. The species-specific clustering was more evident in tri- and 100-kb SOMs than in the respective di- and 10-kb SOMs. For example, more than 99% of the human 100-kb sequences were located in the human territories that are marked by “H”. This shows that the SOM separations, obtained without any species information, closely fit separations among species, and thus the unsupervised algorithm can recognize, in most of the sequences, the species-specific characteristic (a key combination of oligonucleotide frequencies) that is the representative signature of each genome. This observation can be used as a basis for searching for hidden signatures in genome sequences.

3.2 Characteristic Sequence Patterns for Individual Eukaryotic Genomes

Underlying representation in SOMs enables us to retrieve characteristic sequence patterns for individual genomes and genome portions. The frequency of each dinucleotide in the weight vector representative for each lattice in the 100-kb di-SOM (Fig. 1) is illustrated in black and gray (Fig. 3). Lines in all panels in Fig. 3 represent the species borders observed in the di-SOM. Species borders coincide with regions of sharp transition between the black and gray levels for several dinucleotides, that correspond to the diagnostic dinucleotides for the species border formation. For example, levels of CG, GC, CT, TC, AG, and GA contributed appreciably to the species separation. It should be stressed that the SOM utilizes complex combinations of many more dinucleotides for the sequence separations, importantly, in area-dependent manners. This is due to the principle that SOM implements the nonlinear projection from the multidimensional space of input data onto a two dimensional array of weight vectors [14, 15, 16].



Figure 1: SOMs with dinucleotide frequencies in nonoverlapping 10-kb and overlapping 100-kb sequences of 9 eukaryote genomes. Lattices for each species are noted by a capital letter as follows; *S. cerevisiae* (S), *S. pombe* (P), *P. falciparum* (B), *C. elegans* (C), *Arabidopsis* (A), rice (R), *Drosophila* (D), *Fugu* (F), and human (H).

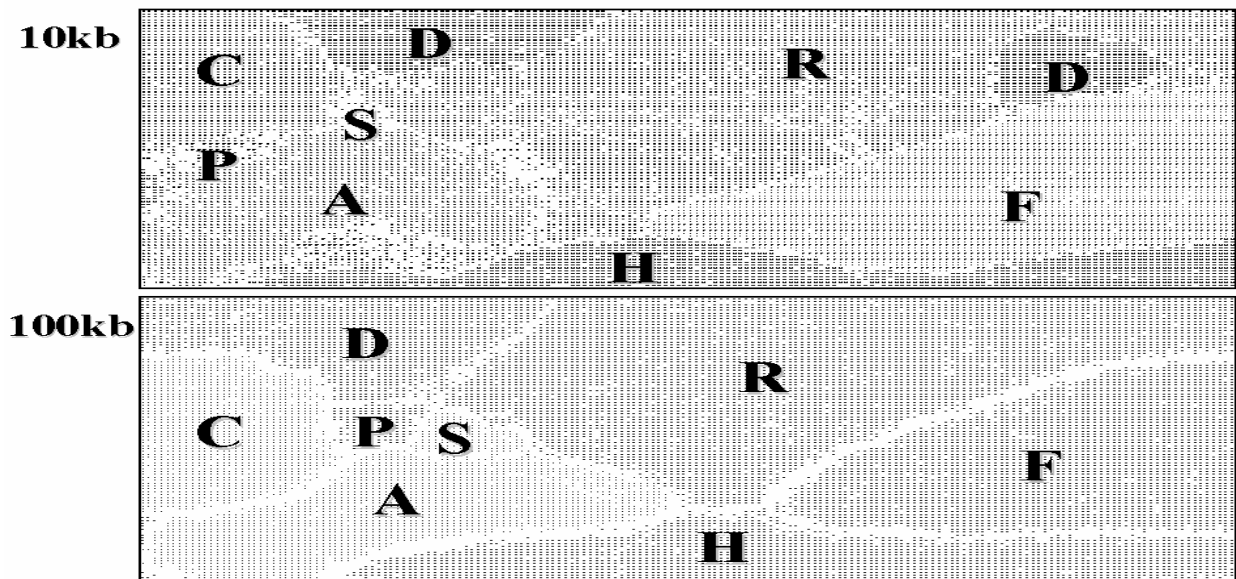


Figure 2: SOMs with trinucleotide frequencies in nonoverlapping 10-kb and overlapping 100-kb sequences of 9 eukaryote genomes. Lattices for each species are noted as described in Fig. 1.

In similar fashion, trinucleotide levels for each representative vector in the tri-SOM were analyzed. Again, species borders often coincided with regions of sharp transition between levels for various diagnostic trinucleotides and the diagnostic examples for species separations were listed (Fig. 4). The SOM utilizes complex combinations of many more trinucleotides for species separation in area-dependent manners. Collectively, the SOM, which can cluster and project complex data efficiently, was shown to be an excellent tool for analyzing global characteristics of genome sequences and for revealing key combinations of oligonucleotides representing individual genomes.

3.3 Intraspecies Separations in SOMs

Human sequences had three and two zone in the 10- and 100-kb di-SOMs, respectively (H zones in Fig. 1). Genomes of warm-blooded vertebrates such as humans are known to be composed of long-range segmental G+C% distributions “isochores” [2, 3, 7, 8, 18]. Correlation of the segmental G+C% distributions with SOM separations was observed. For example, in the case of approximately 500 10-kb sequences belonged to the satellite zone at the left bottom side in the 10-kb di-SOMs, the G+C% was between 30 to 33%, corresponding to very AT-rich sequences in the human genome. Four-fifths of the sequences were derived from very AT-rich “gene-desert region” in chromosome 21 [6], corresponding to L1 isochore [18] and replicating very late during S phase [21]. The SOM can unveil specific genome portions with distinct characteristics as intraspecies separations. Detailed analyses of sequences of equivalent G+C% but separated distantly in SOMs may uncover their genomic uniqueness, which might relate to their functional differentiation.

4 Discussion

It should be noted that the sequences with or without ubiquitous repetitive elements were found to be colocalized in the major zones of individual species in the 10-kb SOM. For example, detailed inspection showed that 10-kb human sequences with or without *Alu* or L1 were colocalized in the human major zones on the 10-kb SOMs. Therefore, the major factors responsible for the species-specific separations

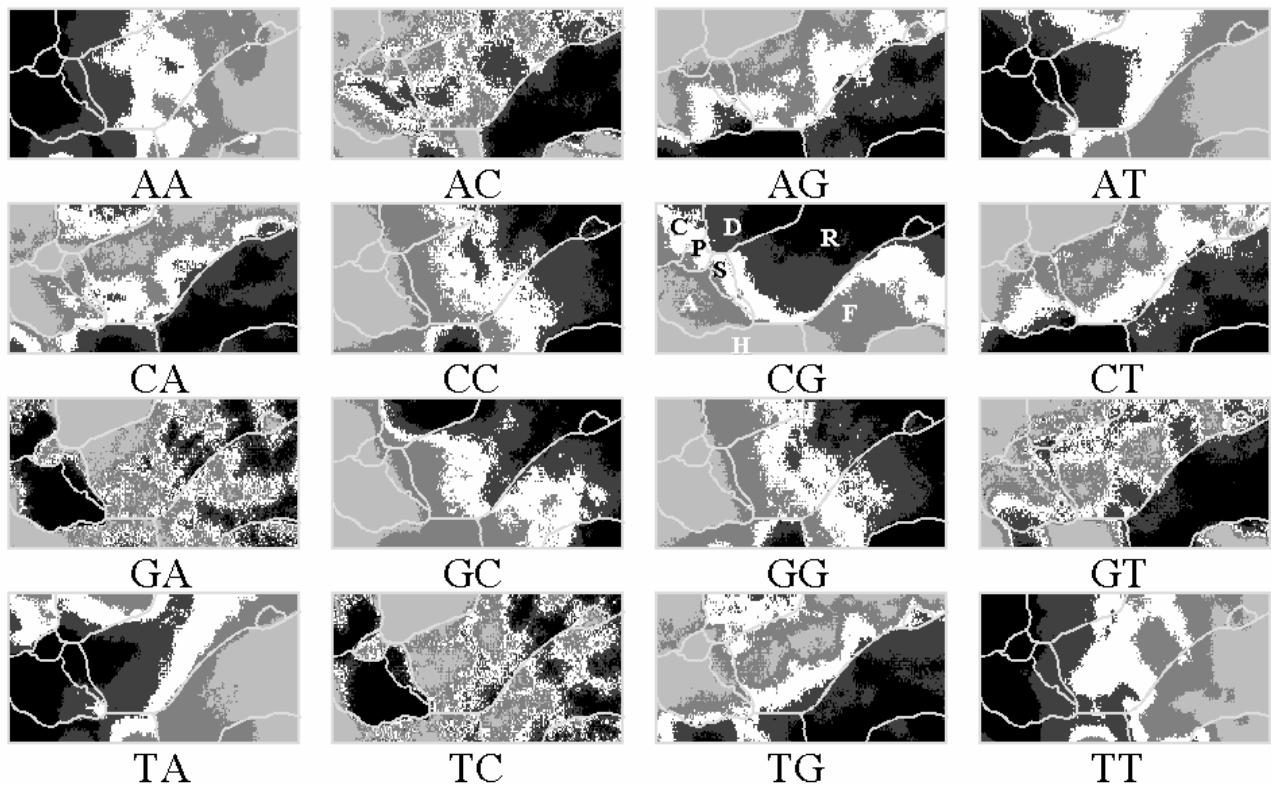


Figure 3: Dinucleotide distribution in 100-kb SOMs for 9 eukaryote genomes. Levels of each dinucleotide for all weight vectors in the 100-kb di-SOM (Fig. 1) were divided into 5 categories with an equal number of sequences. The highest, second-highest, middle, second-lowest, and lowest levels are shown in black, dark gray, white, gray, and light gray, respectively. Species borders for 9 eukaryotes in the SOM (Fig. 1 100kb) are marked by lines. Major zones of the species were noted in the CG panel by the letters noted in Fig. 1.

of eukaryote sequences should not be ubiquitous repetitive elements. Factors responsible for the separations could be characteristics that are more extensively embedded than repetitive elements. One clear example of such a characteristic is CG deficiency in the human and *Fugu* genomes, that is clearly observed in Fig. 3. The species-specific characteristics of DNA-synthesizing and -repairing enzymes, such as the sequence-recognition specificity of DNA primase and the context-dependent repair mechanisms, are candidate factors. Sequence preferences for ubiquitous DNA-binding proteins of individual species, such as histones, are also candidate factors. Furthermore, signal sequences are thought to be biased from the random occurrence and presumably underrepresented except for those very densely present in the genome. This prediction is consistent with our unpublished finding that GAGA/TCTC, which is a transcription signal in *Drosophila*, was characteristically underrepresented in the *Drosophila* territory in the tetranucleotide SOM for eukaryote genomes. SOMs with longer oligonucleotides such as penta- and hexanucleotides may systematically reveal a wide range of signal sequences in genomes.

The present analysis is an example of comparative genomics, and the obtained results were undoubtedly affected by choice of the genomes used. However, when the same set of genomes for which a substantial amount of sequences became available was analyzed, the clustering pattern was stable. For example, among the 9 eukaryotes analyzed here, human sequences increased significantly during course of this study and inclusion of the additional human chromosomes which have been newly sequenced completely, did not change the global clustering pattern.

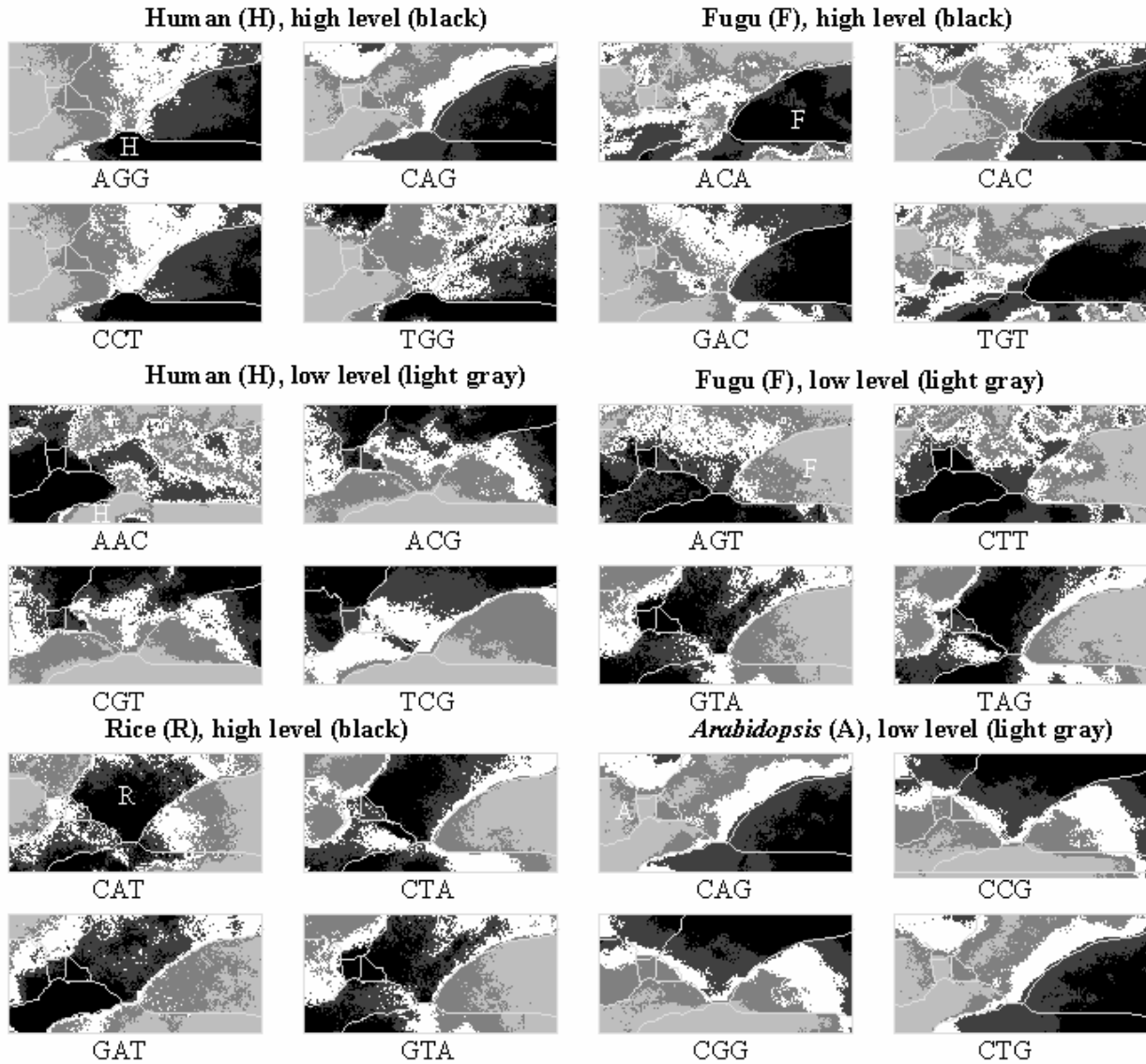


Figure 4: Trinucleotide distribution in 100-kb SOMs for 9 eukaryote genomes. Diagnostic examples for species separations are presented. Levels of each trinucleotide for all weight vectors in the 100-kb tri-SOM (Fig. 2) were divided into 5 categories and shown as described in Fig. 3.

5 Concluding Remarks

In the SOMs, 10- and 100-kb sequences of 9 eukaryotes were separated into species-specific non-overlapping zones (Figs. 1 and 2). In contrast, the clustering power of the conventional PCA method was poor and rather useless for comprehensive comparison of many genomic sequences of a wide variety of species (data not shown: refer to Kanaya *et al.* 2001 [9] for codon analysis). PCA and SOM perform a mapping of the multi-dimensional sequence space onto the plane. PCA rotates the vector space using the eigenvectors (the principal components, PC) of the covariance matrix as a new basis. The principal components are orthogonal, and the plane spanned by the two first components, PC1 and PC2, was used for a linear data projection and display. In contrast to the PCA, the SOM technique performs a nonlinear projection. This can avoid erroneous interpretations of linear projection from a high-dimensional space onto the plane. The SOM algorithm belongs to the field of artificial neural networks. A SOM consists of a grid of formal neurons. The principle of SOM is to associate each data vector with a neuron. Data vectors that are close to each other in high-dimensional space are mapped

onto adjacent neurons. Each neuron can be interpreted as a cluster grouping together data vectors that are most similar to each other, thereby introducing a classification scheme that preserves the topology of the high-dimensional data space. It should be noted that, as a characteristics of nonlinear projection, the distance of two lattices in the grid map is not proportional to the similarity levels of the representative vectors for the lattices. In other words, the similarity levels for the two representative vectors could not be reflected proportionally by their distance in the map. However, the finding that the species territories were surrounded with contiguous white lattices into which no data sequences were classified (e.g., see 100-kb SOM in Fig. 2) showed that vectors of species-specific lattices located even near the species border were clearly distinct from each other.

Because species-specific clustering in SOMs is very clear, SOMs may provide fundamental guidelines for identifying actual molecular mechanisms responsible for establishing the genome characteristics of individual species (genome signatures) during evolution. Application of the SOM in genome informatics should provide a new systematic strategy for revealing hidden genome characteristics that could not otherwise be obtained.

Acknowledgments

This work was supported by ACT- Japan Science and Technology Corporation and by Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan. We thank to Ms Nanayo Ishihara and Yoko Kosaka for technical assistance. The DNA sequences of *Fugu* has been provided freely by the Fugu Genome Consortium for use in this publication/correspondence only.

References

- [1] Abe, T., Kanaya, S., Kinouchi, M., Kudo, Y., Mori, H., Matsuda, H., Carlos, D.C., and Ikemura, T., Gene classification method based on batch-learning SOM, *Genome Informatics*, 10:314–315, 1999.
- [2] Bernardi, G., The isochore organization of the human genome, *Annu. Rev. Genet.*, 23:637–661, 1989.
- [3] Bernardi, G., Olofsson, B., Filipinski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F., The mosaic genome of warm-blooded vertebrates, *Science*, 228:953–958, 1985.
- [4] Gentles, A.J. and Karlin, S., Genome-scale compositional comparisons in eukaryotes, *Genome Res.*, 11:540–546, 2001.
- [5] Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A., Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.*, 8:r49–r62, 1980.
- [6] Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K. *et al.*, The DNA sequence of human chromosome 21, *Nature*, 405:311–319, 2000.
- [7] Ikemura, T., Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, 2:13–34, 1985.
- [8] Ikemura, T. and Aota, S., Global variation in G+C content along vertebrate genome DNA: Possible correlation with chromosome band structures, *J. Mol. Biol.*, 203:1–13, 1988.

- [9] Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., and Ikemura, T., Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): Characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome, *Gene*, 276:89–99, 2001.
- [10] Kanaya, S., Kudo, Y., Abe, T., Okazaki, T., Carlos, D.C., and Ikemura, T., Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome, *Genome Informatics*, 9:369–371,1998.
- [11] 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.
- [12] 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 expression level and species-specific diversity of codon usage based on multivariate analysis, *Gene*, 238:143–155, 1999.
- [13] Karlin, S., Global dinucleotide signatures and analysis of genomic heterogeneity, *Curr. Opin. Microbiol.*, 1:598–610, 1998.
- [14] Kohonen, T., Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, 43:59–69, 1982.
- [15] Kohonen, T., The self-organizing map, *Proc. IEEE*, 78:1464–1480, 1990.
- [16] Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J., Engineering applications of the self-organizing map, *Proc. IEEE*, 84:1358–1384, 1996.
- [17] Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A., Evidence for horizontal gene transfer in *Escherichia coli* speciation, *J. Mol. Biol.*, 222:851–856, 1991.
- [18] Saccone, S., Federico, C., Solovei, I., Croquette, M.F., Valle, G.D., and Bernardi, G., Identification of the gene-richest bands in human prometaphase chromosomes, *Chromosome Res.*, 7:379–386, 1999.
- [19] Sharp, P.M. and Matassi, G., Codon usage and genome evolution, *Cur. Op. in Genetics and Dev.*, 4:851–860, 1994.
- [20] Wang, H.C., Badger, J., Kearney, P., and Li, M., Analysis of codon usage patterns of bacterial genomes using the self-organizing map, *Mol. Biol. Evol.*, 18:792–800, 2001.
- [21] Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T., Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions, *Human Molecular Genetics*, 11:13–21, 2002.
- [22] <http://fugu.hgmp.mrc.ac.uk/>
- [23] <http://rgp.dna.affrc.go.jp/>
- [24] <http://www.ncbi.nlm.nih.gov/Genbank/>