A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide range of eukaryotic genomes

Takashi Abe[1,2], Hideaki Sugawara[1,2], Shigehiko Kanaya[3], Makoto Kinouchi[4], Yasaburo Matsuura[5], Heizo Tokutaka[5] and Toshimichi Ikemura[2]

[1] National Institute of Genetics, Mishima and [2]the Graduate University for Advanced Studies (SOKENDAI), Hayama, Japan. [3] Nara Institute of Science and Technology, Ikoma, Japan. [4] Yamagata University, Yonezawa, Japan. [5]Tottori University, Tottori, Japan.

Abstract – ***Novel tools are needed for comprehensive comparisons of interspecies characteristics of massive amounts of genomic sequences currently available. An unsupervised neural network algorithm, Kohonen's Self-Organizing Map (SOM), is an effective tool for clustering and visualizing high-dimensional complex data on a single map. We modified the conventional SOM for genome informatics on the basis of batch-learning SOM, making the learning process and resulting map independent of the order of data input. We generated the SOMs for tetranucleotide frequencies in 10- and 100-kb sequence fragments from 38 eukaryotes for which almost complete genomic sequences are available. A massive amount of data points (e.g. approximately 600,000 data points for 10-kb sequences) in the 256-dimensional space for the tetranucleotide frequencies was analyzed using the Earth Simulator, which is one of the highest performance supercomputers in the world. SOM recognized species-specific characteristics (key combinations of oligonucleotide frequencies) in the genomic sequences, permitting species-specific classification of the sequences without any information regarding the species. Because the classification power is very high, SOM is an efficient and powerful tool for extracting a wide range of genome information.***

Key words – Self-organizing Map (SOM), batch learning SOM, oligonucleotide frequency, the Earth Simulator, genome informatics, tetranucleotide frequency

1 Introduction

Genome sequences, even protein-noncoding sequences, contain a wealth of information. The G+C content (G+C%) is a fundamental characteristic of individual genomes and used for a long period as a basic phylogenetic parameter to characterize individual genomes. The G+C%, however, is too simple a parameter to differentiate wide varieties of genomes. Many groups have reported that oligonucleotide frequency, which is an example of high-dimensional data, varies significantly among genomes and can be used to study genome diversity [1-4]. An unsupervised neural network algorithm, Kohonen's Self-organizing Map (SOM), is a powerful tool for clustering and visualizing high-dimensional complex data on a two-dimensional map [5-7]. On the basis of batch learning SOM, we have developed a modification of the conventional SOM for genome sequence analyses, which makes the learning process and resulting map independent of the order of data input [8-13]. We previously constructed the SOMs for di-, tri-, and tetranucleotide frequencies in 10-kb genomic sequences from 65 bacteria and 6 eukaryotes. In the resulting SOMs, the sequences were clustered according to species without any information regarding the species, and increasing the length of the oligonucleotides from di- to tetranucleotides increased the clustering power [11]. In the present study, for investigating the power to detect differences among closely related eukaryotes, tetranucleotide frequencies in 10- and 100-kb sequence fragments derive from 38 eukaryotic genomes, which have been sequenced extensively, were analyzed. To characterize and visualize a massive amount of eukaryotic genome sequences on a single map, the Earth Simulator, which is one of the highest performance supercomputers in the world, was used.

2 Methods

SOM implements nonlinear projection of multi-dimensional data onto a two-dimensional array of weight vectors, and this effectively preserves the topology of the high-dimensional data space (5-7). We modified the conventional SOM for genome informatics on the basis of batch learning SOM (BL-SOM) to make the learning process and resulting map independent of the order of data input [8-12]. The initial weight vectors were set based on the widest scale of the sequence distribution in the oligonucleotide frequency space with PCA. Weights in the first dimension (I) were arranged into lattices corresponding to a 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 σ_2/σ_1 x I; and I was set in the present study as the average number of sequence data per neuron becomes four. σ_1 and σ_2 were the standard deviations of the first and second principal components, respectively. The weight vector on the *ij*th lattice ($\mathbf{w}_{ij}$) was represented as follows:

$$\mathbf{w}_{ij} = \mathbf{x}_{av} + \frac{5\sigma_1}{I}\left[\mathbf{b}_1\left(i - \frac{I}{2}\right) + \mathbf{b}_2\left(j - \frac{J}{2}\right)\right] \qquad (1)$$

where $\mathbf{x}_{av}$ is the average vector for oligonucleotide frequencies of all input vectors, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are eigenvectors for the first and second principal components. In Step 2, the Euclidean distances between the input vector $\mathbf{x}_k$ and all weight vectors $\mathbf{w}_{ij}$ were calculated; then $\mathbf{x}_k$ was associated with the weight vector (called $\mathbf{w}_{i'j'}$) with minimal distance. After associating all input vectors with weight vectors, updating was done according to Step 3.

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

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij} + \alpha(r)\left(\frac{\sum_{\mathbf{x}_k \in S_{ij}} \mathbf{x}_k}{N_{ij}} - \mathbf{w}_{ij}\right) \qquad (2)$$

where components of set S_{ij} are input vectors associated with $\mathbf{w}_{i'j'}$ satisfying $i - \beta(r) \le i' \le i + \beta(r)$ and $j - \beta(r) \le j' \le j + \beta(r)$. 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}. $\alpha(r)$ and $\beta(r)$ are set by

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

$$\beta(r) = \max\{1, \beta(1) - r\} \qquad (4)$$

where, $\alpha(1)$ and $\beta(1)$ are the initial values for the T-cycle of the learning process. In the present study, we selected 60~100 for T, 0.6 for $\alpha(1)$, and 40 ~ 80 for $\beta(1)$ depending on the map size (approximately a fourth of I). The learning process is monitored by the total distance between $\mathbf{x}_k$ and the nearest weight vector $\mathbf{w}_{i'j'}$, represented as

$$Q(r) = \sum_{k=1}^{N} \left\{ \left\| \mathbf{x}_k - \mathbf{w}_{i'j'} \right\|^2 \right\} \qquad (5)$$

where N is the total number of sequences analyzed. This batch learning SOM (BL-SOM) is suitable for actualizing high-performance parallel-computing and thus for a large scale computation using the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

3 Results

3.1 SOMs for 38 eukaryote genomes

To investigate clustering power of SOM for a wide range of eukaryote sequences, we analyzed tetranucleotide frequencies in 590,000 and 59,000 10- and 100-kb sequence fragments derived from 38 eukaryote genomes listed in the Fig. 1 legend (a total of 5.9 Gb), which represented a wide rage of eukaryotic phylotypes. To prevent excess contribution of a large size of the human genome, sequences from a half of human chromosomes were used. The SOM adapted for genome informatics was constructed as described previously [11]. First, oligonucleotide frequencies in the sequence fragments were analyzed by PCA, and the first and second principal components were used to set the initial weight vectors that were arranged as a two-dimensional array. SOMs obtained after 100 and 60 learning cycles for the 10- and 100-kb sequences, respectively, revealed clear species-specific separations of the sequence fragments (Fig. 1). Lattice points that contain sequences from a single species are indicated in color, and those that include sequences from more than one species are indicated in black. Most of the 100-kb sequences were classified primarily into species-specific territories. Comparison of classification on the 100-kb tetranucleotide SOM (100-kb Tetra-SOM) with that by the initial vectors set by the first and second principal components (100-kb Tetra-PCA), revealed that sequences from each species were clustered far more tightly on the SOM.

In DNA databases, only one strand of a pair of complementary double-helix sequences is registered, and choice between the two complementary sequences of genomic fragments is often arbitrary in the database registration. When global characteristics of oligonucleotide frequencies in the genome are considered, distinction of frequencies between the complementary oligonucleotides (e.g. AAAC versus GTTT) is not important in most cases. To reduce computation time, SOM was constructed in separate with frequencies for degenerate sets in which the frequencies of a pair of complimentary tetranucleotides were added (DegeTetra-SOM in Fig. 1). This roughly halved the computation time and the level of the species-specific classification is almost equivalent. On the 100-kb Tetra- and DegeTetra-SOMs, 99.8 and 99.9% of the sequences were classified into the proper species territories, respectively. The species separation pattern was simpler on the DegeTetra-SOM than on the Tetra-SOM.

The classification power according to species was shown to be significantly lower for 10-kb sequences (10-kb DegeTetra-SOM in Fig. 1) than for 100-kb SOM; and a large portion of the sequences derived from the two closely related fishes, *Fugu* and *Tetraodon*, each of which composed its own distinct territory on the 100-kb SOM, could not be separated from each other.

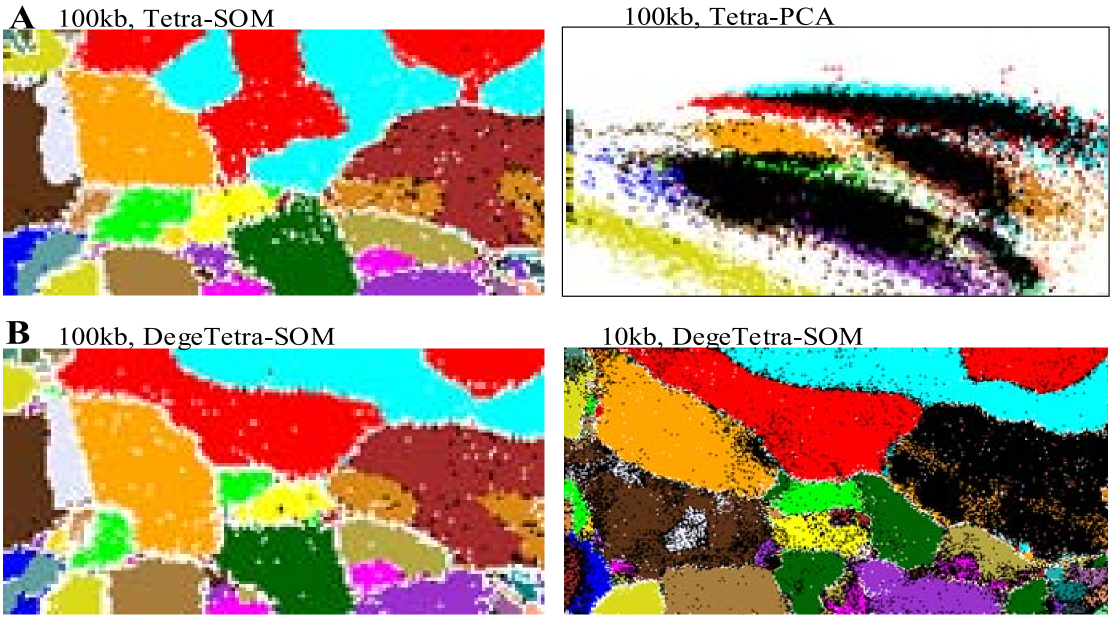

Fig. 1. SOMs for 10- and 100-kb sequences of 38 eukaryotic genomes. (A) 100-kb Tetra-SOM and Tetra-PCA. Tetra-PCA represents the sequence classification by the initial weight vectors set by PCA. (B) 100- and 10-kb DegeTetra-SOMs. Lattice points that include sequences from more than one species are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single species are indicated in color as follows: *Homo sapiens* (■), *Gallus gallus* (■), *Fugu rubripes* (■), *Tetraodon nigroviridis* (■), zebrafish *Danio rerio* (■), *Arabidopsis thaliana* (■), rice *Oryza sativa* (■), *Ciona savignyi* (■), *Ciona intestinalis* (■), *Apis mellifera* (■), *Bombyx mori* (■), *Drosophila pseudoobscura* (■), *Drosophila melanogaster* (■), *Caenorhabditis elegans* (■), Caenorhabditis briggsae (■), *Plasmodium falciparum* (■), *Plasmodium yoelii* (■), *Giardia lamblia* (■), *Kluyveromyces waltii* (■), *Magnaporthe grisea* (■), *Neurospora crassa* (■), *Anopheles gambiae* (■), *Aspergillus terreus* (■), *Aspergillus nidulans* (■), *Candida albicans* (■), *Coprinopsis cinerea* (■), *Cryptococcus neoformans* (■), *Cryptosporidium parvum* (■), *Gibberella zeae* (■), *Saccharomyces* sp. (■), and *Schizosaccharomyces pombe* (■). For human, sequences from chromosomes 2, 6, 7, 13, 14, 20, 21, 22, X, and Y were analyzed. Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/euka/Fig1.html).

3.2 Biological significance of SOM separation

The G+C% is known to be a fundamental characteristic not of individual genomes but also of genomic portions in a single genome. For example, each genome of warm-blooded vertebrates such as human and chicken is composed of long-range segmental G+C% distributions; AT- and GC-rich isochors [14-17]. G+C% obtained from the weight vector for each lattice point in the 100-kb Tetra-SOM (G+C% in Fig. 2A) was reflected in the horizontal axis and increased from left to right. In other words, sequences with high G+C% (red in the G+C% panel) were located on the right side of the map, and similar results were obtained for the DegeTetra-SOMs (data not shown). Territories of the two warm-blooded vertebrates, human and chicken, extended significantly in the horizontal direction, showing sequences of distinct G+C% levels (i.e., AT-rich and GC-rich sequences) present within each genome. Furthermore, the territory of each species was split into a few sub-territories, which most likely reflect the AT- and GC-rich isochores. SOM could differentiate genomic portions with distinct characteristics within one genome. Therefore, the analysis on intraspecies separations may provide profound information regarding the structural details of individual genomes.

3.3 Diagnostic oligonucleotides for species separations

Underlying representation in SOMs enables us to retrieve characteristic oligonucleotide frequencies for individual genomes and genomic portions. The frequency of each tetranucleotide obtained from the weight vector for each lattice point in the 100-kb SOMs was calculated and normalized with the level expected from the mononucleotide composition at each lattice point, and the observed/expected ratios are illustrated in red (overrepresented), blue (underrepresented), or white (moderately represented) in Fig. 2. This normalization allowed oligonucleotide frequencies in each lattice point to be studied independently of mononucleotide compositions. Transitions between red (overrepresentation) and blue (underrepresentation) for various tetranucleotides often coincided exactly with species borders, showing that SOM recognized the species-specific combination of oligonucleotide frequencies that is the representative signature of each genome and enabled us to identify the frequency patterns that are characteristic of individual genomes. Twenty diagnostic cases for the species-specific territory formation are listed in Fig. 2. SOM utilizes complex combinations of many tetranucleotides for the species 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.

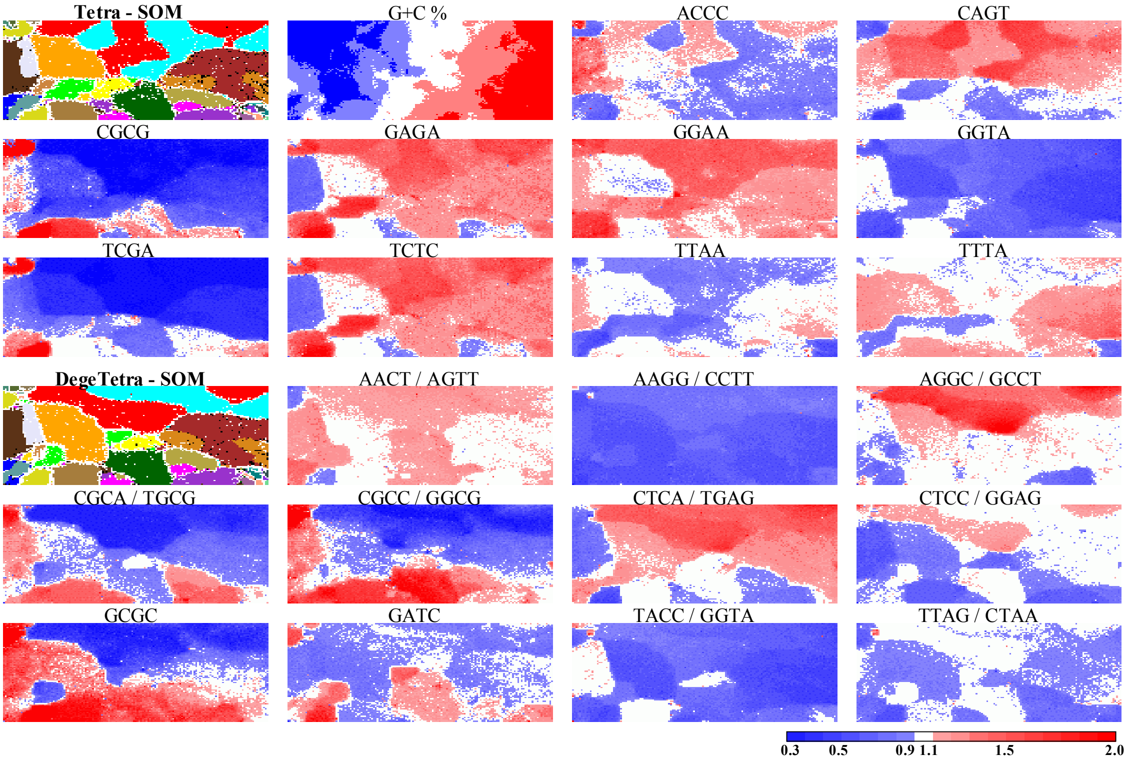

Fig. 2. Levels of each tetranucleotide and of each pair of complimentary tetranucleotides in 100-kb SOMs. G+C% for each lattice point in the 100-kb SOMs was calculated and divided into five categories with an equal number of lattices [11]. The lattice points belonging to the categories of the highest, second-highest, middle, second-lowest, and lowest G+C% are shown in dark red, light red, white, light blue, and dark blue, respectively. Examples of tetranucleotide component planes diagnostic for species separations are presented. Levels of individual tetranucleotides and of complimentary tetranucleotide pairs for each lattice point in the 100-kb Tetra- and DegeTetra-SOMs, respectively, were calculated and normalized with the level expected from the mononucleotide composition of the lattice point. The observed/expected ratio is indicated in colors at the bottom of the figure. Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/euka/Fig2.html).

3.4 SOMs with human, mouse, and rat sequences

The present study is an example of comparative genomics, and the resulting SOM is clearly affected by the genomes included in the analysis. In Fig.1, we analyzed 38 eukaryotic genomes,

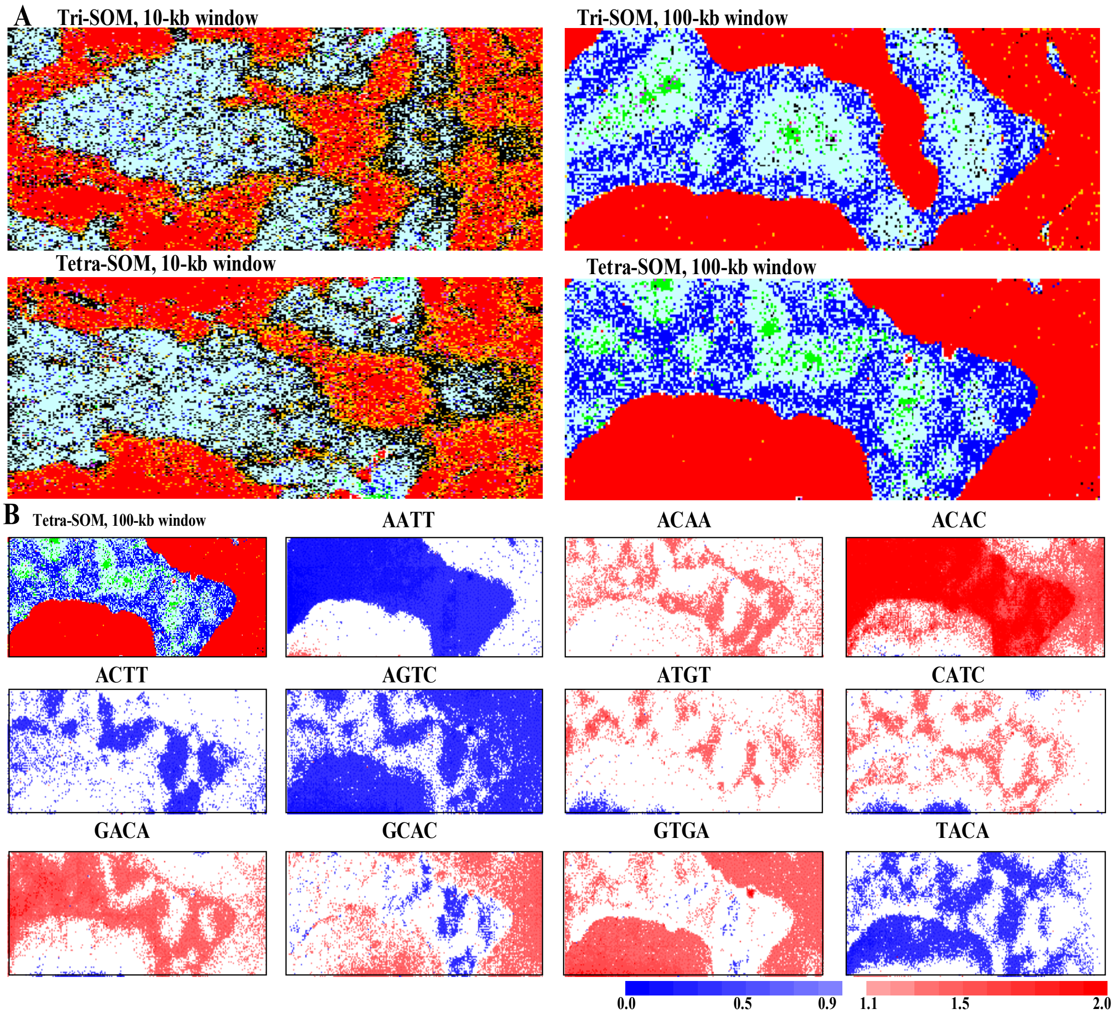

Fig. 3 SOMs for human, mouse and rat draft sequences. (A) Tri- and Tetra-SOMs were constructed with nonoverlapping 10-kb and overlapping 100-kb sequences with a 10-kb sliding step. Lattice points that contain sequences from one or two species are indicated in color (human, red; mouse, dark blue; rat, green; mouse + rat, light blue; human + mouse, yellow; human + rat, violet), those from three species are indicated in black, and those that contained no sequences are indicated in white. (B) Levels of individual tetranucleotides in the weight vector of each lattice point in the 100-kb Tetra-SOM was calculated after normalization of the mononucleotide composition of the lattice weight vector and presented as described in Fig. 2. Component planes diagnostic for species separations are shown. Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/euka/Fig3.html).

which covered a wide range of phylogenetically distinct eukaryotes. To furtherer test the clustering power of SOM for closely related eukaryotes, SOMs were constructed with nonoverlapping 10-kb and overlapping 100-kb sequences with a step size of 10 kb derived from human, mouse and rat genomes. Lattice points that included sequences from one or two species are indicated in color and those that included sequences from all three species are indicated in black (Fig. 3A). Significant separation between human (red) and rodent (mouse and rat; light blue) sequences was observed even in the 10-kb SOMs. In the 10-kb Tetra-SOM, 41% of the human sequences were classified into human-specific territories, and 5% and 4% of mouse and rat sequences, respectively, were classified into mouse and rat territories. In the 100-kb SOMs, separation between human and rodents was very clear, and 99% of human sequences were classified into the human territory. Furthermore, partial separations between mouse and rat were observed; 50% and 21% of mouse and rat sequences, respectively, were classified into mouse (dark blue) and rat (green) territories. Thus, SOMs can recognize unique sequence characteristics even in closely related species.

Diagnostic tetranucleotides for species separations in the 100-kb Tetra-SOM are listed in Fig. 3B. In the mouse and rat territories, ACAC and AATT were over- and underrepresented, respectively, in comparison to levels in human. ACAA and GACA were overrepresented in mouse, and GTGA was overrepresented in human.

4 Discussion

The present SOM is an unsupervised algorithm and thus differs from the supervised algorithm, Learning Vector Quantization (LVQ), proposed by Kohonen *et al.* [18]. The main conclusion of the present study is that, even using the unsupervised algorithm (*i.e.*, without information regarding species), genomic sequences were self-organized according to species on the basis of oligonucleotide frequencies. Because the classification power is very high, the SOM can provide a powerful bioinformatic tool for classifying and extracting a wide range of genomic information even for the categories undiscovered in advance, especially when sequence fragments obtained from a single genome were analyzed; for example, classification according to functional categories unassigned in advance (our unpublished results). However, when we consider the classification of sequences only according to known categories (*e.g.*, according to species), LVQ might have better resolution. If so, LVQ should provide a powerful bioinformatic tool such as the phylogenetic assignment of species-unknown sequences obtained from mixed genomes in a novel environmental sample [13], which was also analyzed in another paper of this meeting [19]. This is because a map is constructed in advance with all available sequences from known species being compiled in DNA databases and on this map the species-unknown sequences from an environmental sample are mapped [19]. High accuracy of species classification for the species-known sequences is prerequisite to the accurate phylogenetic prediction for the species-unknown sequences. Recently we have also analyzed difference of results obtained with plane and spherical SOMs.

When characteristic oligonucleotides, both underrepresented and overrepresented in each genome, are considered, various factors, including DNA conformational tendencies and context-dependent mutation and modification of DNA, are thought to be responsible [1-4, 11]. All CpG-containing tetranucleotides (CGCG, TCGA, CGCA, TGCG, CGCC, GGCG, and GCGC in Fig. 2) showed clear underrepresentation characteristically in all vertebrate territories. This properly reflected their genome characteristics, which is related with DNA methylation. With respect to overrepresented sequences, preferences for sequences recognized by ubiquitous DNA-binding proteins and abundant repetitive elements must be considered. Because inter- and intraspecies separations are very clear, SOMs should provide fundamental guidelines for understanding the detailed molecular mechanisms that have established genome-specific sequence characteristics during evolution. A wide variety of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). We found that occurrence levels of oligonucleotide sequences corresponding to important functional signals were often biased significantly from the random occurrence level, which is expected from the base composition of the genome, and were diagnostic for the species separations (our unpublished results). When known signal sequences of various species with enough experimental data are characterized systematically with SOMs in advance, we can develop an *in silico* method of signal prediction, which is most useful for genomes that are sequenced but for which there is little additional experimental data. Because the number of such genomes has increased rapidly, development of such an *in silico* method has become increasingly important. Functional signals, such as transcription-regulatory signals, are typically longer than tetranucleotides, and therefore, analyses of longer oligonucleotides become important. To conduct SOM with longer oligonucleotides such hexa- and heptanucleotides (4,096- and 16,384-dimensional data) for a massive amount of genome sequences currently available, a large-scale computation using a high-performance supercomputer such as the Earth Simulator becomes essential.

Acknowledgements

This work was supported by grants from the Advanced and Innovational Research Program in Life Sciences and a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan. A part of the present computation was done with the Earth Simulator of JAMSTECT.

References

[1] S. Karlin (1998) Global dinucleotide signatures and analysis of genomic heterogeneity, *Curr. Opin. Microbiol.*, **vol. 1,** p. 98-610,.

[2] R. Nussinov (1984) Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res* **vol. 12**, p. 1749-1763.

[3] A.J. Gentles and S. Karlin (2001) Genome-scale compositional comparisons in eukaryotes, *Genome Res.*, **vol. 11**, p. 540-546.

[4] E.P. Rocha, A. Viari and A. Danchin (1998) Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res.* **Vol. 26**, p. 2971-2980.

[5] T. Kohonen (1982) Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, **vol. 43**, p. 59-69.

[6] T. Kohonen (1990) The self-organizing map, *Proc. IEEE*, **vol. 78**, p. 1464-1480.

[7] T. Kohonen, E. Oja, O. Simula, A. Visa and J. Kangas (1996) Engineering applications of the self-organizing map, *Proc. IEEE*, **vol. 84**, p. 1358-1384.

[8] S. Kanaya, Y. Kudo, T. Abe, T. Okazaki, D.C. Carlos, and T. Ikemura (1998) Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome, *Genome Informatics Series,* **vol. 9**, p. 369-371.

[9] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori and T. Ikemura (2001) 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*, **vol. 276**, p. 89-99.

[10] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura (2002) A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: Self-organizing map of oligonucleotide frequency, *Genome Informatics Series*, **vol. 13**, p. 12-20.

[11] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki and T. Ikemura (2003) Informatics for unveiling hidden genome signatures, *Genome Res.*, **vol. 13**, p. 693-702.

[12] T. Abe, T. Kozuki, Y. Kosaka, A. Fukushima, S. Nakagawa, and T. Ikemura (2003) Self-organizing map reveals sequence characteristics of 90 prokaryotic and eukaryotic genomes on a single map. *WSOM 2003*, p. 95-100.

[13] T. Uchiyama, T. Abe, T. Ikemura and K. Watanabe (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes, *Nature Biotechnology,* **vol. 23**, p. 88-93.

[14] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier (1985) The mosaic genome of warm-blooded vertebrates, *Science*, **vol. 228**, p. 953-958.

[15] T. Ikemura (1985) Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, **vol. 2**, p. 13-34,.

[16] T. Ikemura and S. Aota (1988) Global variation in G+C content along vertebrate genome DNA: possible correlation with chromosome band structures, *J. Mol. Biol.*, **vol. 203**, p. 1-13.

[17] G. Bernardi, *Structural and Evolutionary Genomics*, Amsterdam, Netherlands, Elsevier, 2004.

[18] T. Kohonen, G. Barna, and R.Chrisley (1988) Statistical pattern recognition with neural networks: benchmarking studies, *Proc. ICNN,* **vol. I**, p.61-68.

[19] T. Abe, T. Ikemura, S. Kanaya, M. Kinouchi, and Sugawara, H. (2005) A novel bioinformatics strategy for phylogenetic study of genomic sequences: Self-Organizing Map (SOM) of oligonucleotide frequencies. *WSOM* 2005 in press.