Sequences from Almost All Prokaryotic, Eukaryotic, and Viral Genomes Available Could be Classified According to Genomes on a Large-Scale Self-Organizing Map Constructed with the Earth Simulator

Takashi Abe[1]*, Hideaki Sugawara[1], Shigehiko Kanaya[2] and Toshimichi Ikemura[3]

[1] *Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, and The Graduate University for Advanced Studies (Sokendai), Mishima, Japan*

[2] *Department of Bioinformatics and Genomes, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan*

[3] *The Graduate University for Advanced Studies (Sokendai), Hayama Center for Advanced Research, Hayama, and Nagahama Institute of Bio-Science and Technology, Nagahama, Japan*

Abstract With the increasing and massive amount of available genome sequences, novel tools are needed for comprehensive analysis of genome-specific sequence characteristics for a wide variety of organisms. An unsupervised neural network algorithm, Self-Organizing Map (SOM), is an effective tool for clustering and visualizing high-dimensional complex data on a map. We modified the conventional SOM for genome informatics, making the resulting map independent of data input order. Using the Earth Simulator, we generated SOMs for oligonucleotide frequencies in a massive amount of genomic sequence fragments derived from a wide range of prokaryotes, eukaryotes, organelles, and viruses. SOM recognized genome-specific characteristics of oligonucleotide frequencies in individual genomes, permitting classification (self organization) of sequence fragments according to genomes.

Keywords: Self-Organizing Map (SOM), Environmental Microorganisms, Genome Informatics, Phylogenetic Classification, Metagenome

1. Introduction

Genome sequences, even protein-noncoding sequences, contain a wealth of information. Many groups have reported that oligonucleotide frequency, which is an example of high-dimensional data, varies significantly among genomes and can be used to analyze 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]. 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. We modified the conventional SOM for genome informatics on the basis of batch learning SOM to make the learning process and resulting map independent of the order of data input [8–10]. The newly developed SOM is suitable for actualizing high-performance paral-lel-computing and thus for a large scale computation using the Earth Simulator. In our previous studies [11–13], the SOM method was optimized for phylogenetic classification of genomic sequences obtained from mixed genomes of environmental microorganisms, analysing tetranucleotide frequencies (256-dimensional vectorial data) in 5-kb sequence fragments. In the present study, we further improved the classification method to be applicable for sequences not only from prokaryotes but also from eukaryotes, organelles, and viruses.

2. Methods

The initial weight vectors were set based on the widest scale of the sequence distribution in the oligonucleotide frequency space with PCA, as described in our previous papers [8–11]. 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 $\sigma_2/\sigma_1 \times I$; and I was set in the present study as the average number of sequence data per neuron becomes approximately four. σ_1 and σ_2 were the standard deviations of the first and second principal components, respectively. The weight vector on the ijth lattice ($\mathbf{w}_{ij}$) was represented as follows:

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

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 ijth weight vector was updated by

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

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 rth cycle, and N_{ij} is the number of components of S_{ij}. $\alpha(r)$ and $\beta(r)$ are set by

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

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

where, $\alpha(1)$ and $\beta(1)$ are the initial values for the T-cycle of the learning process. In the present study, we selected $60 \sim 100$ for T, 0.6 for $\alpha(1)$, and $40 \sim 80$ 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

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

where N is the total number of sequences analyzed. This batch learning 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 oligonucleotide frequencies in 90 genomes

To investigate clustering power of SOM for a wide range of prokaryotic and eukaryotic sequences, we first analyzed oligonucleotide frequencies in 90 genomes, which represent a wide rage of prokaryotic and eukaryotic phylogenetic groups. SOMs were constructed with di-, tri-, and tetranucleotide frequencies (16-, 64-, and 256-demensional vectorial data, respectively) for 140,000 nonoverlapping 10-kb sequences and overlapping 100-kb sequences with a moving step size of 10 kb. To set the initial weight vectors, frequencies for the 140,000 sequences were analyzed by PCA.

After 80 learning cycles for SOM, the sequences of many species were separated (self-organized) into species-specific territories (Fig. 1a–c). SOM separations obtained without any species information closely fit the sequence classification according to species. Lattices that included sequences from a single species are indicated in color, those including sequences from more than one species are indicated in black, and those with no sequences are indicated in white. Comparison of sequence classification with the initial vectors set by PCA (Fig. 1e) with those for the final vectors (Fig. 1a) revealed that sequences from a single species were far more tightly clustered on the final map. In all SOMs, most of the eukaryotic sequences were effectively classified into the species-specific territories. In the 10-kb SOMs, the clustering was most evident in the tetranucleotide SOM, and almost all eukaryotic sequences were classified according the species. For example, 95, 98, and 99% of human sequences were classified into human territories (■ in Fig. 1a-c) of the tri- and tetranucleotide SOMs (tri- and tetra-SOMs), respectively. In the 100-kb SOMs, the species-specific separations became more evident, and yeasts *S. cerevisiae* (■) and *S. pombe* (■) and many bacteria also occupied species-specific territories. The species territories were surrounded with contiguous white lattices into which no genomic sequences were classified, showing that vectors of species-specific lattices located even near the species border were clearly distinct from each other. The unsupervised algorithm recognized the species-specific characteristic (a key combination of oligonucleotide frequencies) that is the representative signature of each genome.

The underlying representation in SOMs enabled us to identify characteristic oligonucleotide usage patterns for individual genomes [9]. The frequencies of each tri- and tetranucleotide in each weight vector in the 100-kb SOMs were calculated and represented as different levels of red and blue (Fig. 2). Transitions between the red and blue lev-
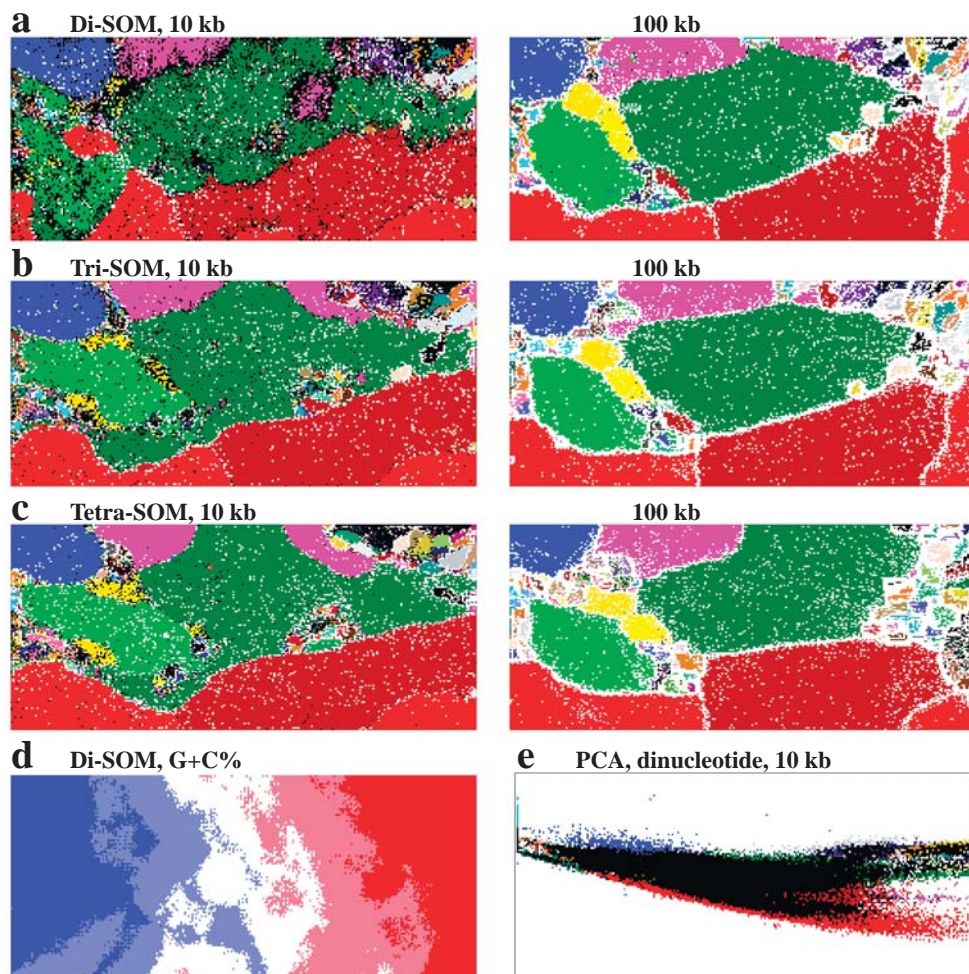
Fig. 1 SOMs for nonoverlapping 10-kb and overlapping 100-kb sequences of 90 genomes. 10-kb and 100-kb di- (a), tri- (b) and tetranucleotide (c) SOMs. (d) G+C% for the weight vector of each lattice was calculated and divided into five categories with an equal number of lattices. The lattices belonging 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. G+C% for each weight vector in the 10-kb di-SOM. (e) Sequence classification by the initial weight vectors set by PCA (Principal Component Analysis) for the 10-kb di-SOM.

To obtain the initial I × J map, the oligonucleotide frequency data was projected linearly as lattices onto the two-dimensional space obtained with the first and second principal components of PCA. Weights in the first dimension (I = 350) were arranged into the lattice 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 $\sigma_2/\sigma_1 \times$ I. Then, SOM learning was conducted as described in Methods section. Lattices that include sequences from more than one species are indicated in black, those including no sequences are indicated in white, and those including sequences from a single species are indicated in color as follows: Human (■), *Fugu rubripes* (■), Rice (■), *A. thaliana* (■), *C. elegans* (■), *D. melanogaster* (■), *P. falciparum* (■), *S. cerevisiae* (■), *S. pombe* (■), *A. aeolicus* (■), *A. fulgidus* (■), *A. pernix* (■), *A. tumefaciens* (■), *B. burgdorferi* (■), *B. halodurans* (■), *B. melitensis* (■), *B. subtilis* (■), *Buchnera* sp. (■), *C. acetobutylicum* (■), *C. crescentus* (■), *C. jejuni* (■), *C. muridarum* (■), *C. perfringens* (■), *C. pneumoniae* (■), *C. trachomatis* (■), *D. radiodurans* (■), *E. coli* (■), *F. nucleatum* (■), *Halobacterium* sp. (■), *H. influenzae* (■), *H. pylori* (■), *L. innocua* (■), *L. lactis* (■), *L. monocytogenes* (■), *M. acetivorans* (■), *M. genitalium* (■), *M. jannaschii* (■), *M. kandleri* (■), *M. leprae* (■), *M. loti* (■), *M. pneumoniae* (■), *M. pulmonis* (■), *M. thermoautotrophicum* (■), *M. tuberculosis* (■), *N. meningitides* (■), *P. aerophilum* (■), *P. abyssi* (■), *P. aerophilum* (■), *P. aeruginosa* (■), *P. furiosus* (■), *P. horikoshii* (■), *P. multocida* (■), *R. conorii* (■), *R. prowazekii* (■), *R. solanacearum* (■), *S. aureus* (■), *S. coelicolor* (■), *S. meliloti* (■), *S. pneumoniae* (■), *S. pyogenes* (■), *S. solfataricus* (■), *S. tokodaii* (■), *S. typhimurium* (■), *Synechocystis* sp. (■), *T. acidophilum* (■), *T. maritime* (■), *T. pallidum* (■), *T. tengcongensis* (■), *T. volcanium* (■), *U. urealyticum* (■), *V. cholerae* (■), *X. axonopodis* (■), *X. campestris* (■), *X. fastidiosa* (■), and *Y. pestis* (■). In the case of the initial map constructed on the basis of PCA, most of the lattice points were associated with sequences derived from more than one species. In contrast, after SOM learning, most of lattice points were associated with sequences derived from one species and thus colored.

els coincided often with the species borders, and in Fig. 2, diagnostic examples for the species separations are listed; complementary oligonucleotides had similar distribution patterns, and therefore, only one example is presented. One clearest example was CATG, which was overrepresented in human and rice, underrepresented in *Drosophila* (D), and moderately represented in *Arabidopsis* and *Fugu*. Underrepresentation of CG-containing tri- and tetranucleotides was apparent in human, *Arabidopsis* (A), and *Fugu*. So far judged from one oligonucleotide, even in this clear example, resolving power between species was clearly dependent on map positions along the species border. It should be stressed that SOMs utilized complex combinations of multiple oligonucleotides for sequence separations in map position-dependent manners resulting in effective classification according to genome categories. This is because SOMs incorporate the nonlinear projection from the multi-dimensional space of input data onto a two-dimensional array of weight vectors [5–7].

3.2 Application for phylogenetic classification of sequences derived from environmental samples

Most environmental microorganisms can not be cultured easily under laboratory conditions. Genomes of uncultivable microorganisms have remained mostly uncharacterized and are thought to contain a wide variety of novel genes of scientific and industrial interest. Metagenomic approaches, which are analyses of mixed populations of uncultured microorganisms, have been developed to identify novel and industrially useful genes and to study microbial diversity in a wide range of environments [14–16]. With the metagenomic approach, genome DNAs are extracted directly from an environmental sample that contains multiple organisms, and genomic fragments are cloned and sequenced. This is a powerful strategy for comprehensive analysis of biodiversity in an ecosystem. However, with a simple collection of many genomic sequence fragments, it is difficult to predict from what phylotypes individual sequences are derived. This is because the conventional phylogenetic classification of genomic sequences is based on sequence homology searches, which require orthologous sequence sets; and therefore, the strate-
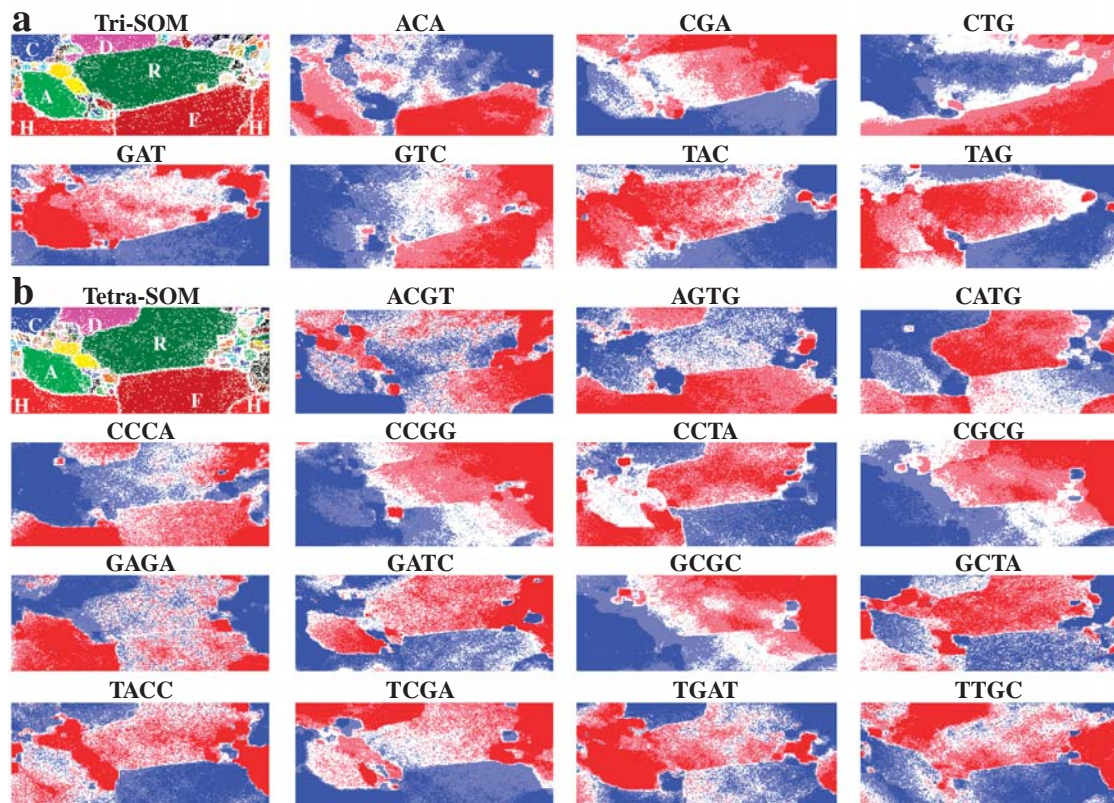

Fig. 2 Level of each tri-(a) and tetranucleotide (b) in 100-kb SOMs. Seven and fifteen examples of component plane to show species separations are presented for Tri- and Tetra-SOMs, respectively. Levels of the tri- and tetranucleotide for the weight vector of each lattice in the respective SOMs of Fig. 1 were divided into five categories with an equal number of lattices and are shown as described in Fig. 1d. The 100-kb SOMs in Fig. 1 are presented in the first panel; *C. elegans* (C), *Arabidopsis* (A), Rice (R), *Drosophila* (D), Fugu (F), and Human (H).

Table 1 Percentage of sequences classified into each genome category.

	Eukaryote	Mitochondrion	Chloroplast	Virus	Prokaryote
Eukaryote	95.6%	0.2%	0.2%	1.6%	2.4%
Mitochondrion	4.9%	87.0%	3.2%	0.8%	4.1%
Chloroplast	8.6%	5.2%	72.2%	0.8%	13.2%
Virus	9.7%	0.2%	0.2%	78.5%	11.4%
Prokaryote	1.7%	0.1%	0.4%	1.6%	96.3%
Sargasso Seq.	8.3%	0.1%	1.4%	3.8%	85.7%

gy can not be applied to poorly characterized or novel gene sequences. We developed a SOM method for phylogenetic classification of novel sequences obtained from an environmental or clinical sample, which contains a wide variety of microorganisms including viruses.

In DNA databases, only one strand sequence of a pair of double strand 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. SOMs for phylogenetic classification were constructed previously with frequencies for degenerate sets of tetranucleotides, where the frequencies of a pair of complimentary tetranucleotides (e.g. AAAC versus GTTT) were added (DegeTetra-SOM). This roughly halved the computation time and the level of the species-specific classification was almost equivalent [11–13].

3.2.1 A large-scale SOM constructed with almost all available sequences derived from species-known genomes

When we consider phylogenetic classification of species-unknown sequences obtained from environmental and clinical samples, it is important to construct SOMs in advance with all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles. This is because various eukaryotic and viral DNAs are thought to be present in environmental and clinical samples. Furthermore, when microorganisms symbiotic/parasitic with a higher eukaryote are analyzed with a metagenomic strategy, sequences from the eukaryote are included inevitably in the sequence collection. On the basis of our previous study on phylogenetic classification of prokaryotic sequences, SOM was constructed with frequencies of degenerate sets of tetranucleotides (DegeTetra-SOM) in 5-kb sequence fragments [11–13]. In the present study, using the Earth Simulator, we could analyze almost all prokaryotic genomic sequences (from 1,502 species)

plus sequences from 40 eukaryotes extensively sequenced and those from 1,065 viruses, 642 mitochondria, and 42 chloroplasts, which have been completely sequenced. The 1,502 prokaryotes were selected because at least 10-kb genomic sequences were registered in DDBJ/EMBL/NCBI. Our main target of the phylogenetic classification, however, is the sequences derived from species-unknown microorganisms present in environmental and clinical samples. To keep good resolution for microorganism sequences, it is necessary to avoid excess representation of sequences derived from higher eukaryotes with large genomes. Therefore, 5-kb eukaryotic sequences were selected randomly from each eukaryote genome up to 25 Mb. This enabled us to analyze an equivalent number of prokaryotic and eukaryotic 5-kb sequences, and DegeTetra-SOM was constructed with the 5-kb sequences (Fig. 4). The power of SOM to separate prokaryotic, eukaryotic, viral, and organelle sequences from each other was very high (Table 1). We also observed the clear separation of prokaryotic sequences into 25 major prokaryote families, confirming our previous study [11]. As found in Fig. 1c, the separation of eukaryotic sequences according to species was also observed (data not shown).

3.2.2 Phylogenetic classification of environmental sequences

A large-scale metagenome study of uncultivable microorganisms in environmental and clinical samples should allow extensive surveys of genes useful in medical and industrial applications and assist in developing accurate views of the ecology of uncultivable microorganisms. Traditional methods of phylogenetic classification have been based on sequence homology searches and therefore inevitably focused on well-characterized genes such as rDNA, for which orthologous sequences from a wide range of phylotypes are available. However, most of the well-characterized genes, including rDNA, are not industrially attractive. It would be best if microbial diversity could be assessed during the process of screening for novel genes with industrial and scientific significance. SOM is the most suitable method for this purpose.
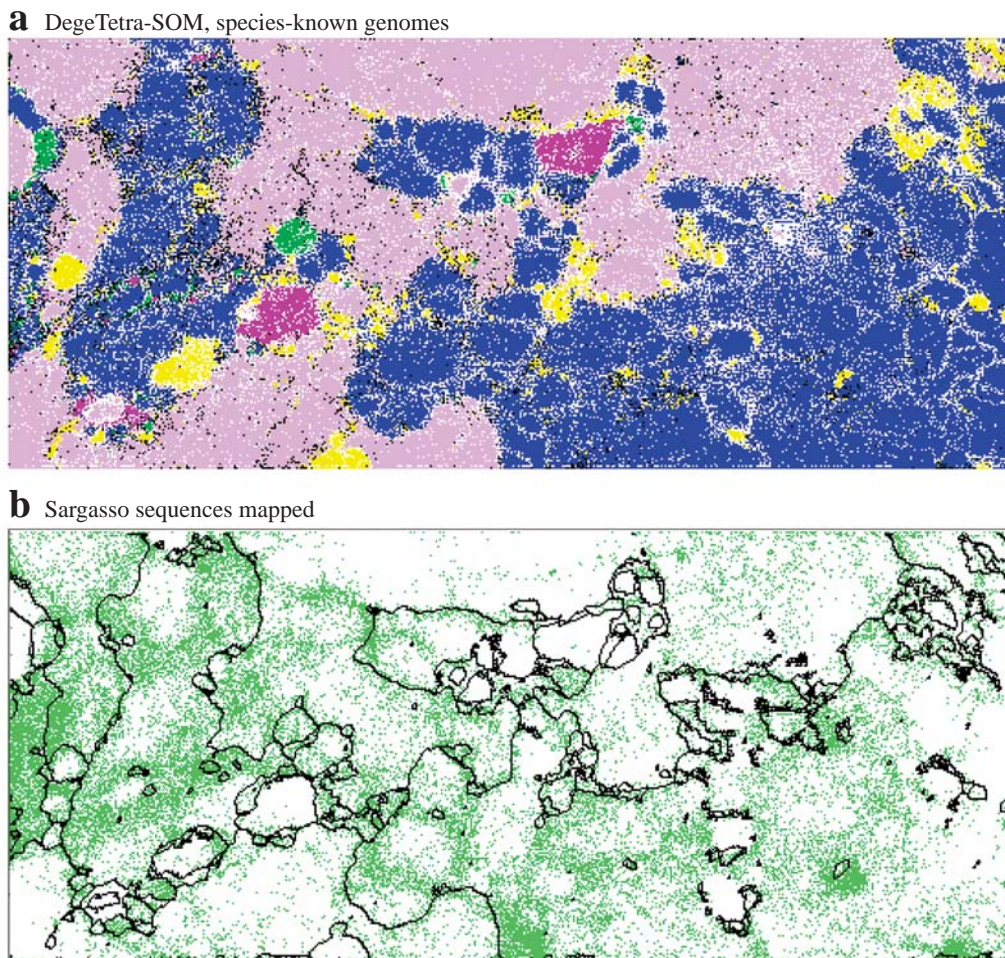
a DegeTetra-SOM, species-known genomes

b Sargasso sequences mapped

Fig. 3 Phylogenetic classification of environmental sequences. (a) DegeTetra-SOM of 5-kb sequences derived from species-known 1,502 prokaryotes, 40 eukaryotes, 642 mitochondria, 42 chloroplasts, and 1065 viruses. Names of these known species are available from our Web page (http://lavender.genes.nig.ac.jp/takaabe/SPT1.htm). Lattice points that contain sequences only prokaryotic or eukaryotic sequences are indicated in colors (■ or ■, respectively); those that contain only mitochondria, chloroplast, or virus sequences are indicated in colors (■, ■, or ■, respectively); those that include more than one category are indicated in black. (b) Sargasso sequences longer 1 kb were mapped on the 5-kb DegeTetra-SOM, after normalization of the sequence length.

Venter *et al*. [17] applied a large-scale shotgun sequencing to mixed genomes collected from the Sargasso Sea near Bermuda and deposited approximately 811,000 sequence fragments (a total of 1 Gb) in DDBJ/EMBL/GenBank. The Sargasso sequences longer than 1 kb were mapped onto the DegeTetra-SOM presented in Fig. 3a; for each Sargasso sequence, the lattice point with the shortest distance in the multidimensional vectorial space was assigned (Fig. 3b). More than 85% of the Sargasso sequences were classified into the prokaryotic territories, and skewed distribution even in the prokaryotic territory was clear. In the bottom low in Table 1, the proportion of Sargasso sequences classified into individual categories is listed. In the phylogenetic classification of uncultivable and poorly characterized species, classification into phylotypes rather than into individual species is important. Detailed phylotype predictions for individual Sargasso sequences are available on our Web page (http://lavender.genes.nig.ac.jp/takaabe/SPT1.htm and http://lavender.genes.nig.ac.jp/takaabe/SPT2.htm). In the Web page, 92 genera, into which Sargasso sequences were assigned, are listed together with the ID number of each of the assigned Sargasso sequences. Most of these Sargasso sequences have not been characterized phylogenetically because there were no orthologs for these sequences. In the case of sequences from totally novel organisms, sequences even from related species are not represented on the SOM in Fig. 3a. Importantly, such novel sequences can be identified by calculating distance between the vectorial data of the respective sequence and that of the sequence-mapped lattice point.

Acknowledgement

(This article is reviewed by Dr. Tetsuya Sato.)

References

[1] S. Karlin, Global dinucleotide signatures and analysis of genomic heterogeneity, *Curr. Opin. Microbiol.*, vol.**1**, pp.98–610, 1998.

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

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

[4] E. P. Rocha, A. Viari, and A. Danchin, Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons, *Nucleic Acids Res.*, vol.**26**, pp.2971–2980, 1998.

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

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

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

[8] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, and T. Ikemura, 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**, pp.89–99, 2001.

[9] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, Informatics for unveiling hidden genome signatures, *Genome Res.*, vol.**13**, pp.693–702, 2003.

[10] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, A large-scale Self-Organizing Map (SOM) unveils sequence characteristics of a wide range of eukaryote genomes, *Gene*, vol.**365**, pp.27–34, 2006.

[11] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples, *DNA Research*, vol.**12**, pp.281–290, 2005.

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

[13] H. Hayashi, T. Abe, M. Sakamoto, et al., Direct cloning of genes encoding novel xylanases from human gut, *Can. J. Microbiol.*, vol.**51**, 251–259, 2005.

[14] R. I. Amann, W. Ludwig, and K. H. Schleifer, Phylogenetic identification and in situ detection of individual microbial cells without cultivation, *Microbiol. Rev.*, vol.**59**, pp.143–169, 1995.

[15] P. Hugenholtz, and N. R. Pace, Identifying microbial diversity in the natural environment: a molecular phylogenetic approach, *Trends Biotechnol.*, vol.**14**, pp.190–197, 1996.

[16] M. R. Rondon, P. R. August, A. D. Bettermann, et al., Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms, *Appl. Environ. Microbiol.*, vol.**66**, pp.2541–2547, 2000.

[17] J. Venter, K. Remington, J. F. Heidelberg, et al., Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, vol.**304**, 66–74, 2004.