A NOVEL BIOINFORMATICS STRATEGY FOR PHYLOGENETIC STUDY OF GENOMIC SEQUENCE FRAGMENTS: SELF-ORGANIZING MAP (SOM) OF OLIGONUCLEOTIDE FREQUENCIES

Takashi Abe[1], Toshimichi Ikemura[2], Shigehiko Kanaya[3], Makoto Kinouchi[4], and Hideaki Sugawara[1].

[1] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, and The Graduate University for Advanced Studies (Sokendai), Mishima, Shizuoka 411-8540, Japan. [2] The Graduate University for Advanced Studies (Sokendai), Hayama Center for Advanced Research, Hayama-cho, Kanagawa 240-0193, Japan. [3]Department of Bioinformatics and Genomes, Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0101, Japan. [4]Department of Bio-System Engineering, Faculty of Engineering, Yamagata University, Yonezawa, Yamagata 992-8510, Japan.

Abstract – *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 two-dimensional array of weight vectors [1-3]. We adapted SOM to genome informatics, making the learning process and resulting map independent of the order of data input. In the present study, a novel bioinformatics strategy for phylogenetic classification of sequence fragments obtained from pooled genome samples of uncultured microbes in environmental samples was developed. Using the SOM method we could visualize microbial diversity and relative abundance of microorganisms within an environmental sample on a single map. First we constructed SOMs of tri- and tetranucleotides in a total of 3.3 Gb sequence derived from 113 prokaryotic and 13 eukaryotic genomes, for which complete genomic sequences are available. SOMs classified 330,000 10-kb sequences from the 126 genomes mainly according to species without information of species. Importantly, the classification was possible without orthologous sequence sets and thus was especially useful for analyses of novel sequences from poorly characterized species, which have attracted wide industrial attentions. Using the SOM method, sequences that were derived from a single genome but cloned independently in a metagenome library could be reconstructed in silico. Because the classification power is very high, SOM is an efficient and fundamental bioinformatic strategy for extracting a wide range of genome information from a vast amount of sequence data. Biological significance of the clustering on SOM and usefulness in metagenome studies were discussed.*

Key words – self-organizing map, oligonucleotide frequencies, genome signatures, phylogenetic classification, environmental microorganism, metagenome analysis.

1 Introduction

Most environmental microorganisms can not be cultured easily under laboratory conditions. Genomes of uncultured organisms have remained mostly uncharacterized and are thought to contain a wide range of novel genes of scientific and industrial interest. Metagenomic

approaches, which are analyses of mixed populations of uncultured microbes, have been developed to identify novel and industrially useful genes and to study microbial diversity in a wide variety of environments [4,5,6]. With the metagenomic approach, genomic DNAs are extracted directly from an environmental sample that contains multiple organisms, and the DNA 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 sequence fragments, it is difficult to predict from what phylotypes individual sequences are derived. This is because the conventional phylogenetic classification of sequences is based on sequence homology searches, which require orthologous sequence sets; and therefore, this strategy can not be applied to poorly characterized or novel sequences. A new phylogenetic classification method could be developed on the basis of an unsupervised neural network algorithm, Kohonen's self-organizing map (SOM). We previously modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input [7, 8, 9, 10, 11]. In the present study, the SOM method was optimized for phylogenetic classification of genomic sequences from environmental and clinical samples.

2 Methods

On the basis of batch learning SOM, we modified the conventional Kohonen's SOM for genome informatics to make the learning process and resulting map independent of the order of data input [7, 8, 9, 10]. The initial weight vectors were defined by PCA instead of random values. Weight vectors (w_{ij}) were arranged in the two-dimensional lattice denoted by i (= 0, 1, .., I-1) and j (= 0, 1, .., J-1). I was set as 150 and 350 in Fig. 1A and B, respectively, and as 350 and 400 in Figs. 3 and 4, respectively. J was defined by the nearest integer greater than (σ_2/σ_1) x I. σ_1 and σ_2 were the standard deviations of the first and second principal components, respectively. Weight vectors (w_{ij}) were set and updated as described previously [11]. Detailed procedures of the present SOM method are explained by another presentation from our group in this WSOM 2005 meeting [12]. Nucleotide sequences were obtained from http://www.ncbi.nlm.nih.gov/GenBank/ [13]. When the number of undetermined nucleotides (Ns) in a sequence exceeded 10% of the window size, the sequence was omitted from the analysis. When the number of Ns was less than 10%, the oligonucleotide frequencies were normalized to the length without Ns and included in the analysis.

3 Results and Discussion

3.1 SOMs for oligonucleotide frequencies in 126 genomes.

To test the classification power of SOM for a wide range of genome sequences, we analyzed short oligonucleotide frequencies in the 126 genomes for which complete sequences are available: 0.3 Gb for 113 prokaryotes and 3.0 Gb for 13 eukaryotes (see Fig. 1 legend). SOMs were constructed with tri- and tetranucleotide frequencies for 330,000 nonoverlapping 10-kb sequences and overlapping 100-kb sequences with a sliding step size of 10 kb derived from the total of 3.3 Gb sequence. To set the initial weight vectors, frequencies for the 330,000 sequences were analyzed by PCA [7, 8, 9, 10, 11]. After 100 learning cycles, the sequences of most species were separated into species-specific territories (Fig. 1A, B). SOM separation, which was obtained without any species information, closely fit the sequence classification according to species. Lattice points that contained 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. Most of the sequences were effectively classified into the species-specific territories. Even in the

10-kb SOMs, most eukaryotic sequences were classified according the species. For example, 98 and 99% of human sequences were classified into human territories (■ in Fig. 1A, B) of the tri- and tetranucleotide SOMs (tri- and tetra-SOMs), respectively. In the 100-kb SOMs, the species-specific separations became more evident, and many prokaryotes also occupied clear 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, and therefore, the species borders could be drawn automatically on the basis of the contiguous white lattices. Collectively, the unsupervised algorithm recognized the species-specific characteristic (a key combination of oligonucleotide frequencies) that is the representative signature of each genome.

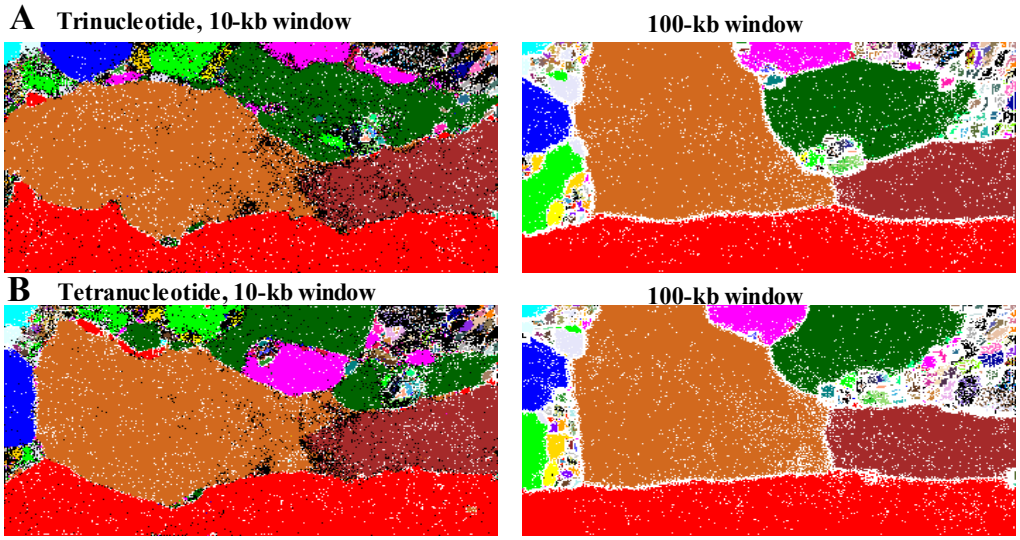

Fig. 1 SOMs for nonoverlapping 10-kb and overlapping 100-kb sequences of 126 genomes. 10-kb and 100-kb tri- (A) and tetranucleotide (B) SOMs. Lattice points that contain sequences from more than one species are indicated in black, those including no sequences are indicated in white, and those including sequences from a single species are indicated in color as follows *Saccharomyces cerevisiae* (■), *Schizosaccharomyces pombe* (■), *Dictyostelium discoideum* (■), *Entamoeba histolytica* (■), *Plasmodium falciparum* (■), *Arabidopsis thaliana* (■), *Medicago truncatula* (■), rice *Oryza sativa* (■), *Caenorhabditis elegans* (■), *Drosophila melanogaster* (■), puffer fish *Fugu rubripes* (■), zebrafish *Danio rerio* (■), *Homo sapiens* (■), *Aeropyrum pernix* (■), *Agrobacterium tumefaciens* (■), *Anabaena* sp. (■), *Aquifex aeolicus* (■), *Archaeoglobus fulgidus* (■), *Bacillus anthracis* (■), *Bacillus halodurans* (■), *Bacillus subtilis* (■), *Bifidobacterium longum* (■), *Borrelia burgdorferi* (■), *Bradyrhizobium japonicum* (■), *Brucella melitensis* (■), *Brucella suis* (■), *Buchnera aphidicola* (■), *Buchnera* sp. (■), *Campylobacter jejuni* (■), *Caulobacter crescentus* (■), *Chlamydia muridarum* (■), *Chlamydia trachomatis* (■), *Chlamydophila pneumoniae* (■), *Chlorobium tepidum* (■), *Clostridium acetobutylicum* (■), *Clostridium perfringens* (■), *Corynebacterium efficiens* (■), *Corynebacterium glutamicum* (■), *Deinococcus radiodurans* (■), *Escherichia coli* (■), *Fusobacterium nucleatum* (■), *Haemophilus influenzae* (■), *Halobacterium* sp. (■), *Helicobacter pylori* (■), *Lactobacillus plantarum* (■), *Lactococcus lactis* (■), *Listeria innocua* (■), *Listeria monocytogenes* (■),*Mesorhizobium loti* (■), *Methanobacterium thermoautotrophicum* (■), *Methanococcus jannaschii* (■), *Methanopyrus kandleri* (■), *Methanosarcina acetivorans* (■), *Methanosarcina mazei* (■), *Mycobacterium leprae* (■), *Mycobacterium tuberculosis* (■), *Mycoplasma genitalium* (■), *Mycoplasma pneumoniae* (■), *Mycoplasma pulmonis* (■), *Neisseria meningitidis* (■), *Oceanobacillus iheyensis* (■), *Pasteurella multocida* (■), *Pseudomonas aeruginosa* (■), *Pseudomonas putida* (■), *Pseudomonas syringae* (■), *Pyrobaculum aerophilum* (■), *Pyrococcus abyssi* (■), *Pyrococcus furiosus* (■), *Pyrococcus horikoshii* (■), *Ralstonia solanacearum* (■), *Rickettsia conorii* (■), *Rickettsia prowazekii* (■), *Salmonella enterica* (■), *Salmonella typhimurium* (■), *Shewanella oneidensis* (■), *Shigella flexneri* (■), *Sinorhizobium meliloti* (■), *Staphylococcus aureus* (■), *Streptomyces coelicolor* (■), *Staphylococcus epidermidis* (■), *Streptococcus agalactiae* (■), *Streptococcus pneumoniae* (■), *Streptococcus pyogenes* (■), *Sulfolobus solfataricus* (■), *Sulfolobus tokodaii* (■), *Synechocystis* sp. (■), *Thermoplasma acidophilum* (■), *Thermosynechococcus elongatus* (■), *Thermotoga maritime* (■), *Treponema pallidum* (■), *Tropheryma whipplei* (■), *Thermoanaerobacter tengcongensis* (■), *Thermoplasma volcanium* (■), *Ureaplasma urealyticum* (■), *Vibrio cholerae* (■), *Vibrio parahaemolyticus* (■), *Vibrio vulnificus* (■), *Xanthomonas axonopodis* (■), *Xanthomonas campestris* (■), *Xylella fastidiosa* (■), and *Yersinia pestis* (■). Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/env/Fig1.html).

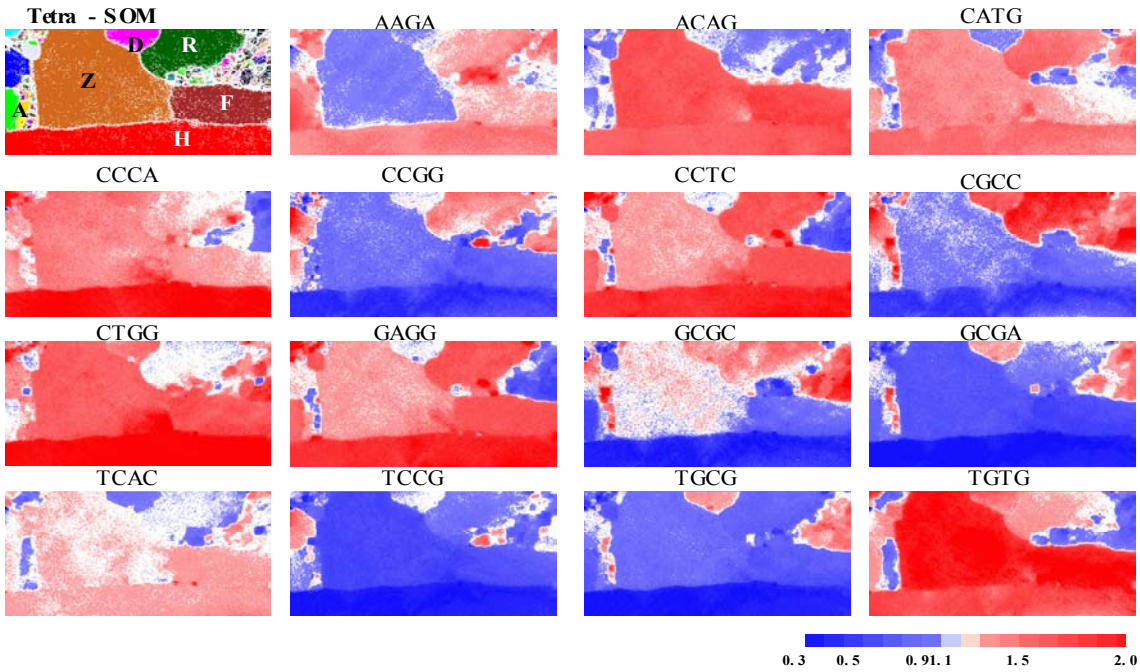

Fig. 2 Level of each tetranucleotide in the 100-kb tetranucleotide SOM. Diagnostic examples of species separations are presented. Level of each tetranucleotide for each lattice point in the 100-kb Tetra-SOM was calculated and normalized with the level expected from the mononucleotide composition of the lattice point. The observed/expected ratio is indicated in colors at the bottom of the figure. The 100-kb tetranucleotide SOM in Fig. 1 is presented in the first panel; *Arabidopsis* (A), rice (R), *Drosophila* (D), Fugu (F), Zebra fish (Z), and human (H). Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/env/Fig2.html).

The frequencies of each tetranucleotide in each weight vector in the tetranucleotide 100-kb SOM were calculated and represented as different levels of red and blue (Fig. 2). Transitions between the red and blue levels coincided often with the species borders (Fig. 2). One clearest example was CATG, which was overrepresented in human (H), Arabidopsis (A), Zebra fish (Z) and rice (R), underrepresented in Drosophila (D), and moderately represented in Fugu (F). So far as judged from one oligonucleotide, even in this clear example, resolving power between species was dependent on map positions. It was shown that SOMs utilized complex combinations of multiple oligonucleotides for sequence separations in map position-dependent manners resulting in effective classification according to species.

3.2 Application to metagenomic studies.

In the metagenomic study of environmental microorganisms, DNA is extracted directly from mixed genomes in an environmental sample without cultivation, and the DNA fragments are cloned into vectors, then sequenced. An unsupervised algorithm SOM should be most useful for predicting how many and what types of genomes are present in an environmental sample, because genomic sequences with specific characteristics were self-organized without information of spices. Recently, Tyson *et al.* [14] applied the metagenome shotgun-sequencing to mixed genomes collected from an acidophilic biofilm growing in acid mine drainage (AMD), which is a worldwide environmental problem. To reconstruct dominant genomes with shotgun sequencing, they focused on this biofilm because of low-complexity of the mixed genomes in the sample and deposited approximately 2,455 sequence fragments in DDBJ/EMBL/GenBank. We phylogenetically classified these biofilm sequences in the following way.

In the DNA database, only one strand of a pair of complementary sequences is registered. Our previous analyses revealed that sequence fragments from a single prokaryotic genome are often split into two territories that reflect the transcriptional polarities of the genes present in the fragment [11]. For phylogenetic classification of sequences from uncultured microbes, it is not necessary to know the transcriptional polarity of the sequence, and the split into two territories complicates assignment to species. Therefore, we tested a new type of SOM in which frequencies of a pair of complementary oligonucleotides (e.g., AACC and GGTT) were summed, and the SOM for the degenerate sets of tetranucleotides was designated DegeTetra-SOM. We constructed a DegeTera-SOM with all available sequences of known species rather than only those from completely sequenced genomes. This is because environmental samples of interest presumably contain poorly characterized and novel species. In Fig. 3, DegeTetra-SOM constructed with 210,000 non-overlapping 5-kb sequences (a total of 1.05 Gb) from 1502 species-known prokaryotes for which at least 10 kb of sequence has been deposited in DDBJ/EMBL/GenBank is shown. These 1502 prokaryotes were then classified into 25 phylotypes with reference to the NCBI Taxonomy Database. The classification according to phylotype was apparent (Species-known Seq. in Fig. 3A); as a separate study, the classification on the 1-kb SOM was found to be significantly lower than that on the 5-kb SOM (should be published elsewhere). Then, the biofilm sequences were mapped on the 5-kb DegeTetra-SOM (Biofilm sequences in Fig. 3B). Most of the biofilm sequences, which were derived from the low-complexity metagenome library, were located mostly in distinct and restricted territories, confirming that most sequence fragments derived from a single genome but cloned independently could be reassociated *in silico*.

When phylogenetic classification of uncultured environmental microorganisms, especially from clinical samples, was considered, it is necessary to construct SOMs in advance with both prokaryotic and eukaryotic sequences because varieties of eukaryotic genomes may be included in the samples. Furthermore, when microorganisms symbiotic/parasitic with a higher eukaryote are analyzed, sequences from this eukaryote are included inevitably in the sequence collection. To examine the SOM separation of prokaryotic sequences from a wider range of eukaryotic sequences, 5-kb sequences from 13 eukaryotes sequenced extensively were analyzed simultaneously with 5-kb sequences from the 147 prokaryotes. To avoid excess representation of eukaryotic genomes with large sizes and to analyze an equivalent number of prokaryotic and eukaryotic sequences, 5-kb eukaryotic sequences were selected randomly from each eukaryote genome up to 25 Mb and DegeTetra-SOM was constructed with the 5-kb prokaryotic and eukaryotic sequences (Fig. 4). The power of SOM to separate prokaryotic from eukaryotic sequences was very high. 99.5% of prokaryotic sequences were classified into prokaryotic territories, and 0.2% and 0.1% were classified into yeast *S. pombe* and *S. cerevisiae* territories, respectively. Separation among eukaryotic genomes again was apparent also on the 5-kb SOM (Fig. 4B). Next we mapped the biofilm sequences on this SOM (Biofilm Seq. in Fig. 4C) after normalization of the sequence length. It is apparent that a major portion of the biofilm sequences were classified into specific prokaryote territories.

The present method may be useful for survey of pathogenic microorganisms in clinical laboratory samples (e.g., feces, sputum and snivel), especially those that can not be cultured easily. Because no species information is required in advance, the method is most useful for identification of novel pathogenic microorganisms that cause unclear infectious diseases.

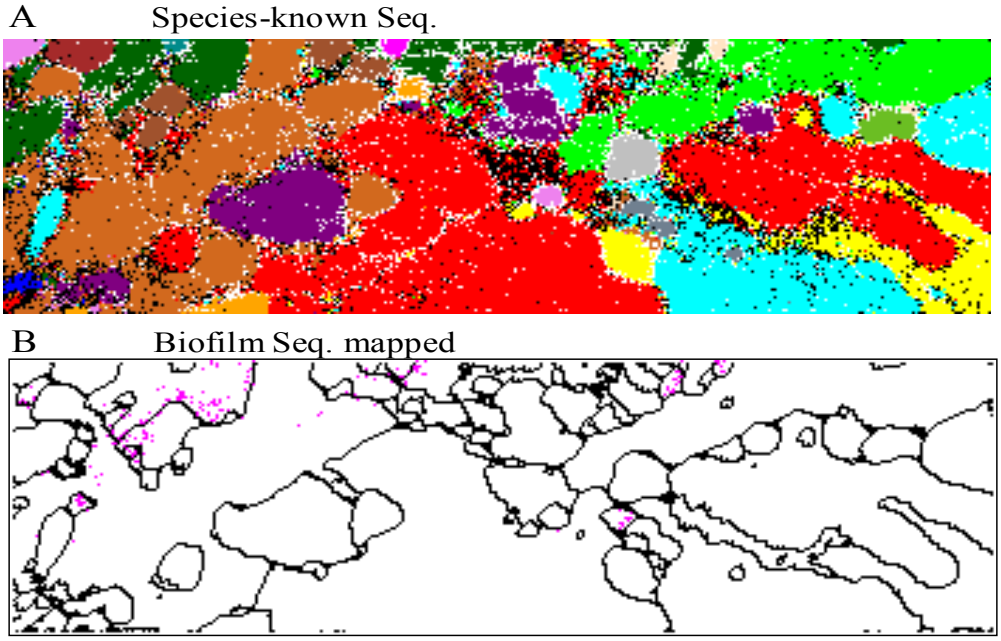

Fig. 3 SOM for phylogenetic classification of sequences from an environmental sample. (A) DegeTetra-SOM of 5-kb sequences of species-known prokaryotes. The genomic sequences from 1502 prokaryotes were used. Species-known Seq.; prokaryotic sequences were classified into 25 phylotypes. Lattice points that include sequences from more than one phylotype are indicated in black, and those that contain sequences from a single phylotype are indicated in color as follows: α-proteobacteria (■), β-proteobacteria (■), γ-proteobacteria (■), δ-proteobacteria (■), ε-proteobacteria (■), Actinobacteria (■), Aquificae (■), Bacteroidetes (■), Chlamydiae (■), Chlorobi (■), Chloroflexi (■), Crenarchaeota (■), Cyanobacteria (■), Deinococcus-Thermus (■), Dictyoglomi (■), Euryarchaeota (■), Fibrobacteres (■), Firmicutes (■), Fusobacteria (■), Nitrospirae (■), Planctomycetes (■), Spirochaetales (■), Thermodesulfobacteriales (■), Thermotogales (■), and Verrucomicrobiae (■).(B) The biofilm sequences were mapped on the DegeTetra-SOM. The lattice points, on which biofilm sequences were mapped, are indicated in pink. Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/env/Fig3.html).

4 Concluding Remarks

Novel tools are needed for comprehensive studies of massive amounts of genomic sequences currently available. An unsupervised neural network algorithm, self-organizing map (SOM), is an effective tool for clustering and visualizing high-dimensional complex data on a single map. SOM recognized species-specific characteristics (key combinations of oligonucleotide frequencies) in sequence fragments permitting species-specific classification of the sequences without any information regarding the species. Because species-specific clustering on SOMs is very clear, SOM is a powerful tool for phylotype classification of genomic sequences, especially for sequence fragments obtained from mixed genomes of uncultured environmental microorganisms [15, 16, 17]. When considering phylogenetic classification of environmental microorgansms, it is worthwhile to construct SOMs of all sequences from all known species for extraction of sequence characteristics in a wider rage of genomes. In the case of sequences from totally novel organisms, sequences even from related species might not be represented on the SOM. Importantly, such novel sequences can be identified accurately by calculating distance between the vector of the respective sequence data and that of the sequence-mapped lattice point (i.e., the node with the minimum distance from the sequence data in the multidimensional space).

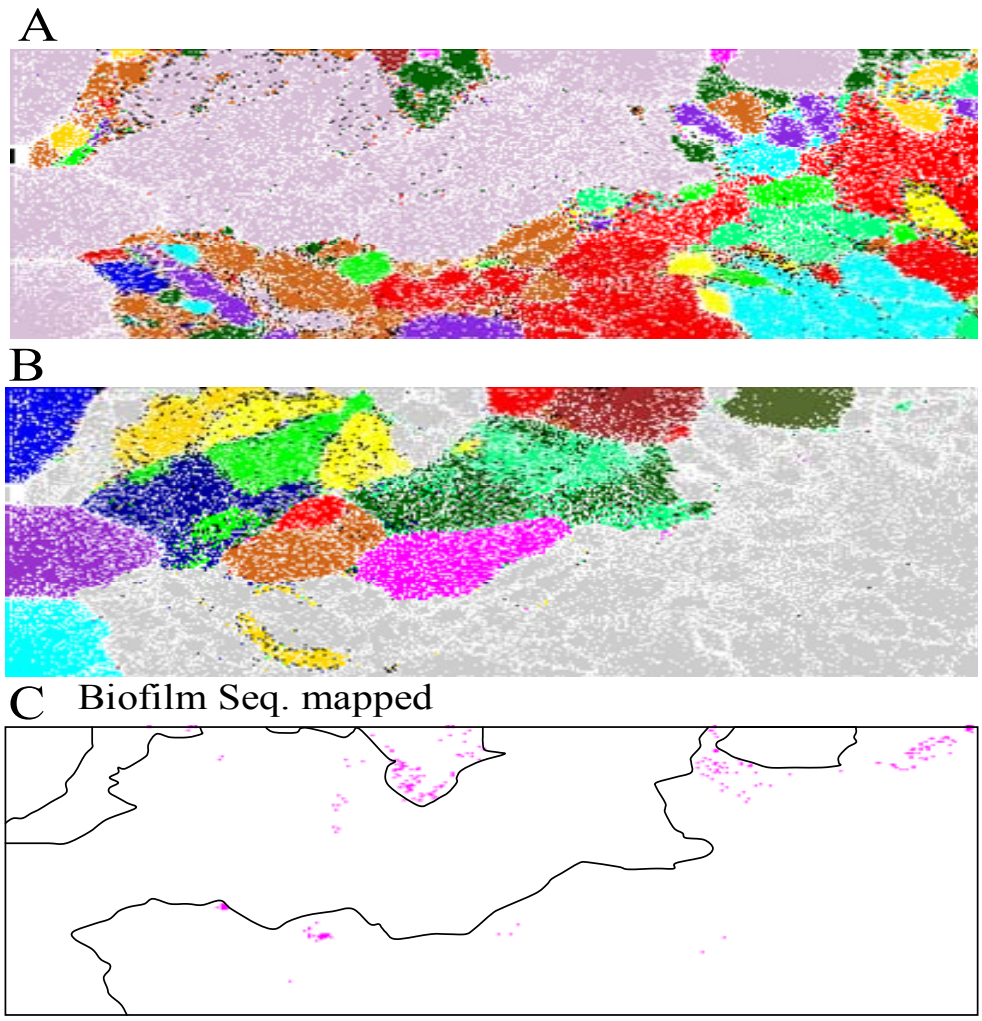

Fig. 4 DegeTetra-SOM of 5-kb sequences from 147 prokaryotes and 13 eukaryotes. (A) Lattice points that contain prokaryotic sequences from more than one phylogenetic group are indicated in black, and those that contain sequences from a single group are indicated in color as follows: α-proteobacteria (■), β-proteobacteria (■), γ-proteobacteria (■), δ-proteobacteria (■), Archaea (■), Chlamydia (■), Firmicutes (■), Actinobacteria (■), Fusobacteria (■), Thermotogae (■), Cyanobacteria (■), and others (■). Nodes that contain sequences from both prokaryotic and eukaryotic sequences are also indicated in black and those contain sequences only from eukaryotic genomes are indicated in color (■). (B) Lattice points that contain sequences only from prokaryotic genomes are indicated in color (■). Lattice points that contain sequences from a single eukaryotic species are indicated in color as follows: *Saccharomyces cerevisiae* (■), *Schizosaccharomyces pombe* (■), *Dictyostelium discoideum* (■), *Entamoeba histolytica* (■), *Plasmodium falciparum* (■), *Arabidopsis thaliana* (■), *Medicago truncatula* (■), rice *Oryza sativa* (■), maize *Zea mays* (■), *Caenorhabditis elegans* (■), *Drosophila melanogaster* (■), puffer fish *Fugu rubripes* (■), zebrafish *Danio rerio* (■), and *Homo sapiens* (■). Lattice points that contain sequences from more than one eukaryotic species or from both eukaryotic and prokaryotic species are indicated in black. (C) The biofilm sequences were mapped on the DegeTetra-SOM. The lattice points, on which biofilm sequences were mapped, are indicated in pink. Color version of this figure can be obtained from our URL (http://lavender.genes.nig.ac.jp/takaabe/wsom2005/env/Fig4.html).

Acknowledgements

This work was supported by grants from ACT-Japan Science and Technology Corporation and 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 Japan Agency for Marine-Earth Science and Technology.

References

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

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

[3] 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.

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

[5] E.F. DeLong (2002) Microbial population genomics and ecology. *Curr. Opin. Microbiol.,* **vol. 5,** p. 520-524.

[6] P. Hugenholtz and N.R. Pace (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.*, **vol. 14**, p. 190-197.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M., Matsuura, Y., Tokutaka, H.. and Ikemura, T. (2005) A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide rage of eukaryotic genomes. WSOM 2005 in press.

[13] http://www.ncbi.nlm.nih.gov/Genbank/

[14] G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar and J.F. Banfield (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **vol. 428**, p. 37-43.

[15] 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.

[16] H. Hayashi, T. Abe, M. Sakamoto, H. Ohara, T. Ikemura, K. Sakka and Y. Benno, Direct cloning of genes encoding novel xylanases from human gut, *Canadian Journal of Microbiology*, in press.