*Plant Biotechnology* 26, 469–477 (2009)

***Original Paper***



# Novel bioinformatics for inter- and intraspecies comparison of genome signatures in plant genomes

Takashi Abe*, Kennosuke Wada, Yuki Iwasaki, Toshimichi Ikemura

Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan
* E-mail: takaabe@nagahama-i-bio.ac.jp Tel & Fax: +81-749-64-8126

Received September 16, 2009; accepted October 16, 2009 (Edited by S. Kanaya)

**Abstract** Novel tools are needed for comprehensive comparisons of the inter- and intraspecies characteristics of a large amounts of available genome sequences. 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 (BLSOM), making the resulting map independent of the order of data input. We generated BLSOMs for oligonucleotide frequencies in fragment sequences (e.g. 10-kb) from 13 plant genomes for which almost complete genome sequences are available. BLSOM recognized species-specific characteristics (key combinations of oligonucleotide frequencies) in most of the fragment sequences, permitting classification (self-organization) of sequences according to species without any information regarding the species during computation. To disclose sequence characteristics of a single genome independently of other genomes, we constructed BLSOMs for sequence fragments from one genome plus computer-generated random sequences. Genomic sequences were clearly separated from random sequences, revealing the oligonucleotides with characteristic occurrence levels in the genomic sequences. We discussed these oligonucleotides diagnostic for genomic sequences, in connection with genetic signal sequences. Because the classification and visualization power is very high, BLSOM is thought to be an efficient and powerful tool for extracting a wide range of genomic information.

**Key words:** Batch-learning SOM, oligonucleotide frequency, random sequences, genome signature.

Genome sequences contain a wealth of genetic information, and massive quantities of genome sequences have accumulated in the International Nucleotide Sequence Databases. It is an important task of life science to extract unknown basic knowledge from a large amount of available genome sequences. The G+C content (G+C%) has been used for a long period as a basic parameter for characterizing individual genomes and genomic portions but is too simple a parameter to differentiate and characterize 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 (Nussinov 1984; Karlin et al. 1997; Karlin 1998a, b; Rocha et al. 1998; Gentles and Karlin 2001; Pride et al. 2003). 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 (Kohonen 1982, 1990; Kohonen et al. 1996). On the basis of Batch Learning SOM (BLSOM), 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 (Kanaya et al. 1998, 2001; Abe et al. 2002, 2003). We used the BLSOM for phylogenetic classification of genomic fragment sequences obtained from mixed genomes of environmental microorganisms by analyzing tetranucleotide frequencies (Abe et al. 2005, 2006b; Hayashi et al. 2005; Uchiyama et al. 2005).

In the present study, to test the power of BLSOM to detect the differences among plant genomes and the intraspecies differences within one plant genome, we examined the frequencies of tri- and tetranucleotide in most (if not all) of the plant genomes for which complete (or nearly complete) sequence data are available and compared genome signatures among interspecies. BLSOMs allowed us to extract species-specific characteristics of oligonucleotide frequencies in each plant genome (genome signature). It should be noted, however, that the species-specific characteristics thus extracted were inevitably dependent on a set of the species included in the BLSOM analysis because the analysis is focusing on the comparison among the species (a comparative genomics). As a strategy for disclosing characteristics of a single genome independently of other genomes, we examined the BLSOM for sequence fragments (e.g. 10 kb) from one

This article can be found at http://www.jspcmb.jp/

genome plus computer-generated random sequences. In the resulting BLSOM including random sequences, the genomic sequences were clearly separated from random sequences, and furthermore, the genomic sequences were clustered according to functional and structural categories in a single genome without any information other than the oligonucleotide frequencies. We then focused on oligonucleotides with characteristic frequencies in connection with genetic signal sequences.

## Materials and methods

### *Batch-Learning Self-Organizing Map for genome sequences*

Multivariate analyses such as factor corresponding analysis and principal component analysis (PCA) have been used successfully to investigate variations in gene sequences (Grabtham et al. 1980; Kanaya et al. 1996, 1999). However, the clustering powers of conventional multivariate analyses are inadequate when massive amounts of sequence data from a wide variety of genomes are analyzed collectively (Kanaya et al. 1998, 2001; Abe et al. 2002, 2003). 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 (Kohonen 1982, 1990; Kohonen et al. 1996). We modified the conventional SOM for genome informatics on the basis of batch-learning SOM (BLSOM) to make the learning process and resulting map independent of the order of data input (Kanaya et al. 1998, 2001; Abe et al. 2002, 2003). The initial weight vectors were defined by PCA instead of random values on the basis of the finding that PCA can classify gene sequences into groups of known biological categories when relatively small amounts of sequence data were analyzed in advance (Kanaya et al. 1996, 1999). 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 350 nodes for 10-kb sequences (Figures 1), and J was defined by the nearest integer greater than $(\sigma_2/\sigma_1)\times 350$. $\sigma_1$ and $\sigma_2$ were the standard deviations of the first and second principal components, respectively. Weight vectors ($w_{ij}$) were set and updated as described previously (Kanaya et al. 1998, 2001). A BLSOM program suitable for PC cluster systems and a PC program for mapping of new sequences on a large-scale BLSOM constructed with high-performance supercomputers can be obtained from UNTROD, Inc. (k_wada@nagahama-i-bio.ac.jp).

Nucleotide sequences were obtained from http://www.ncbi.nlm.nih.gov/Genbank/. 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.

### *Generation of random nucleotide sequences*

For each genomic fragment sequence, a random sequence that had nearly the same oligonucleotide (e.g., trinucleotide) composition as the respective genomic sequence was generated by the Markov chain model. For example, in the case of the random sequence with the same trinucleotide composition (designated Tri-Random), we initially calculated trinucleotide composition of the respective genomic fragment and chose the first trinucleotide randomly but with a weight level to reflect the level of the trinucleotide composition of the genomic fragment. If the first trinucleotide thus chosen in the randomization process was "AGC", the next nucleotide "X" which follows the "AGC" was chosen randomly but with a weight level reflecting the "GCX" composition in the respective genomic sequence. This random choice process was continued until the length of the Tri-Random sequence reached to that of the respective genomic sequence. In the present study, two random sequences were generated for each genomic sequence.

## Results

### *BLSOMs for 13 plant genomes*

To investigate the clustering power of BLSOM for a wide range of plant genomes, we initially analyzed tri- and tetranucleotide frequencies in 5- or 10-kb fragment sequences derived from a total of 2 Gb sequence of 13 plant genomes. These genomes included *Arabidopsis thaliana*, papaya *Carica papaya*, *Chlamydomonas reinhardtii*, rice *Oryza sativa* Indica, rice *Oryza sativa* Japonica, rice *Oryza sativa* Japonica Nihon, *Ostreococcus lucimarinus*, moss *Physcomitrella patens*, poplar *Populus trichocarpa*, and grape *Vitis vinifera*, as well as all of the chloroplast, mitochondrion and plastid genomes compiled in the International Nucleotide Sequence Databases. The BLSOM adapted for genome informatics was constructed as described previously (Abe et al. 2002, 2003). First, oligonucleotide frequencies in the 5- or 10-kb sequences 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. BLSOMs obtained after 140 and 90 learning cycles for the 5- and 10-kb sequences, respectively, revealed clear species-specific separations (self-organization) of the sequence fragments (Figures 1).The sequences were classified primarily into species-specific territories; lattice points that include sequences from a single species are indicated in color, and those that include sequences from more than one species are indicated in black. Sequences from a single species were clustered more tightly on the tetranucleotide BLSOM (Tetra-BLSOM) than on the trinucleotide BLSOM (Tri-BLSOM). For example, 85.4 and 74.3 % of *Arabidopsis* sequences were classified into the *Arabidopsis* territories (■ in Figure 1) in the 10-kb Tetra- and Tri -BLSOMs, respectively.

In DNA databases, only one strand of a pair of complementary double-stranded sequences is registered. When global characteristics of oligonucleotide frequencies in the genome are considered, distinction of frequencies between the complementary

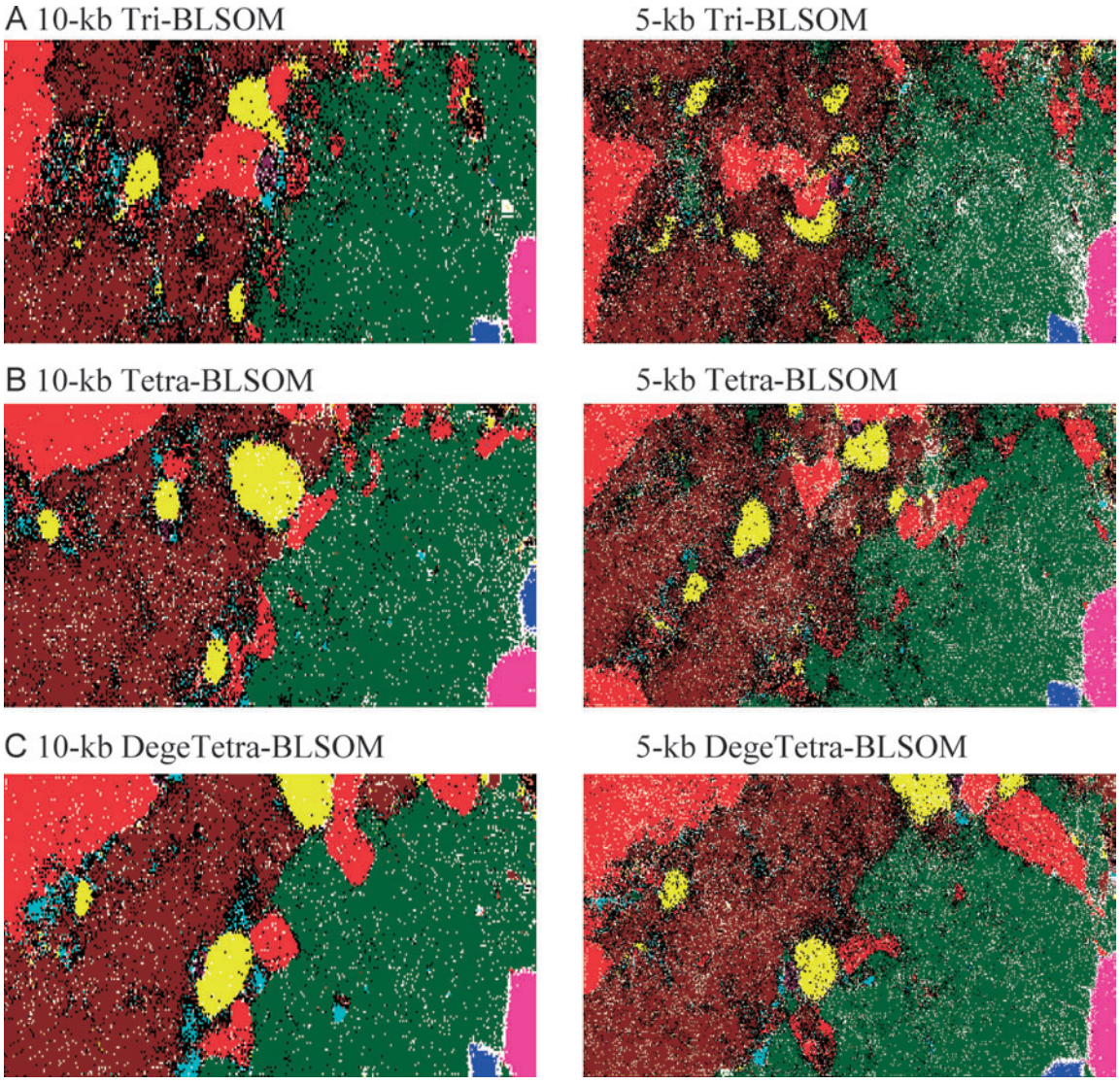


Figure 1. BLSOMs for 5- and 10-kb sequences of 13 plant genomes. (A) Tri-BLSOM. (B) Tetra-BLSOM. (C) DegeTetra-BLSOM. Lattice points that include sequences from more than one species are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single species are indicated in color as follows: *Arabidopsis thaliana* (■), papaya *Carica papaya* (■), *Chlamydomonas reinhardtii* (■), rice *Oryza sativa* Indica (■), rice *Oryza sativa* Japonica (■), rice *Oryza sativa* Japonica Nihon (■), *Ostreococcus lucimarinus* (■), moss *Physcomitrella patens* (■), poplar *Populus trichocarpa* (■), grape *Vitis vinifera* (■), chloroplast (■), mitochondrion (■), and plastid (■) genomes.

oligonucleotides (e.g., AAAC versus GTTT) is not important in most cases; e.g. the choice between the two complementary sequences of genomic fragments is often arbitrary in the database registration. To reduce computation time, BLSOM was also constructed with frequencies for degenerate sets in which the frequencies of a pair of complementary tetranucleotides were added (DegeTetra-BLSOM in Figure 1). This roughly halved the computation time and the pattern of the species-specific classification became simpler. Furthermore, a slight increase in the classification power according to species was observed. For example, on the 10-kb Tetra- and DegeTetra-BLSOMs, 85.4% and 92.4% of *Arabidopsis* sequences were classified into the *Arabidopsis* territories, respectively. The species-specific separation pattern on the 5-kb BLSOM was more complex than on the 10-kb BLSOM, and the classification power according to species was lower on the former BLSOM. Collectively, the 10-kb DegeTetra-BLSOM gave the highest power of species-specific separation in the present study.

When we inspected the 10-kb BLSOMs in detail, the dicot genomes (*Arabidopisis*, papaya, poplar) and monocot genomes (rice) were classified to the right and left side of the map, respectively, showing that the BLSOMs recognized common features of these two types of plant genomes. There were several minor territories composed of small numbers of sequences with specific characteristics. For example, a very small territory for *Arabidopsis* (■) located between the rice

(■) territories was composed primarily of sequences from its centromeric and subcentromeric regions. Analysis of such intraspecies separations may provide profound information regarding the structural details of individual genomes. Because the lineage relationship is relatively close, a large portion of sequences derived from the two related dicots, Poplar and Grape, could not be separated well from each other.

### *Biological meanings of BLSOM clustering*

The G+C% is known to be a fundamental characteristic not only of individual genomes but also of genomic portions within one genome. For example, *Arabidopsis* is known to be composed of long-range segmental G+C% distributions "AT- and GC-rich isochores" (Zhang and Zhang 2004), which have originally been found for warm-blood vertebrates (Bernardi et al. 1985; Bernardi 2004; Ikemura 1985; Ikemura and Aota 1988). The G+C% obtained from the weight vector for each lattice point in the Tri-BLSOM (G+C% in Figure 2) was reflected in the horizontal axis and increased from left to right. 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 Tetra-BLSOMs (data not shown). Sequences with the same G+C% were separated by a complex combination of oligonucleotide frequencies resulting in species-specific separations. Territories of *Arabidopsis* (a dicot plant) extended significantly in the horizontal direction, showing sequences of distinct G+C% levels (i.e., AT-rich and GC-rich sequences) present within one genome. Furthermore, its territory was split into a few sub-territories with significant sizes (larger than the centromeric and subcentromeric regions mentioned above), which may reflect the AT- and GC-rich isochores. BLSOM could differentiate genomic portions with distinct characteristics within one genome, and therefore, the detailed analyses on intraspecies separations may provide profound information regarding the structural details of individual genomes.

BLSOM recognized the species-specific combination of oligonucleotide frequencies that is the representative

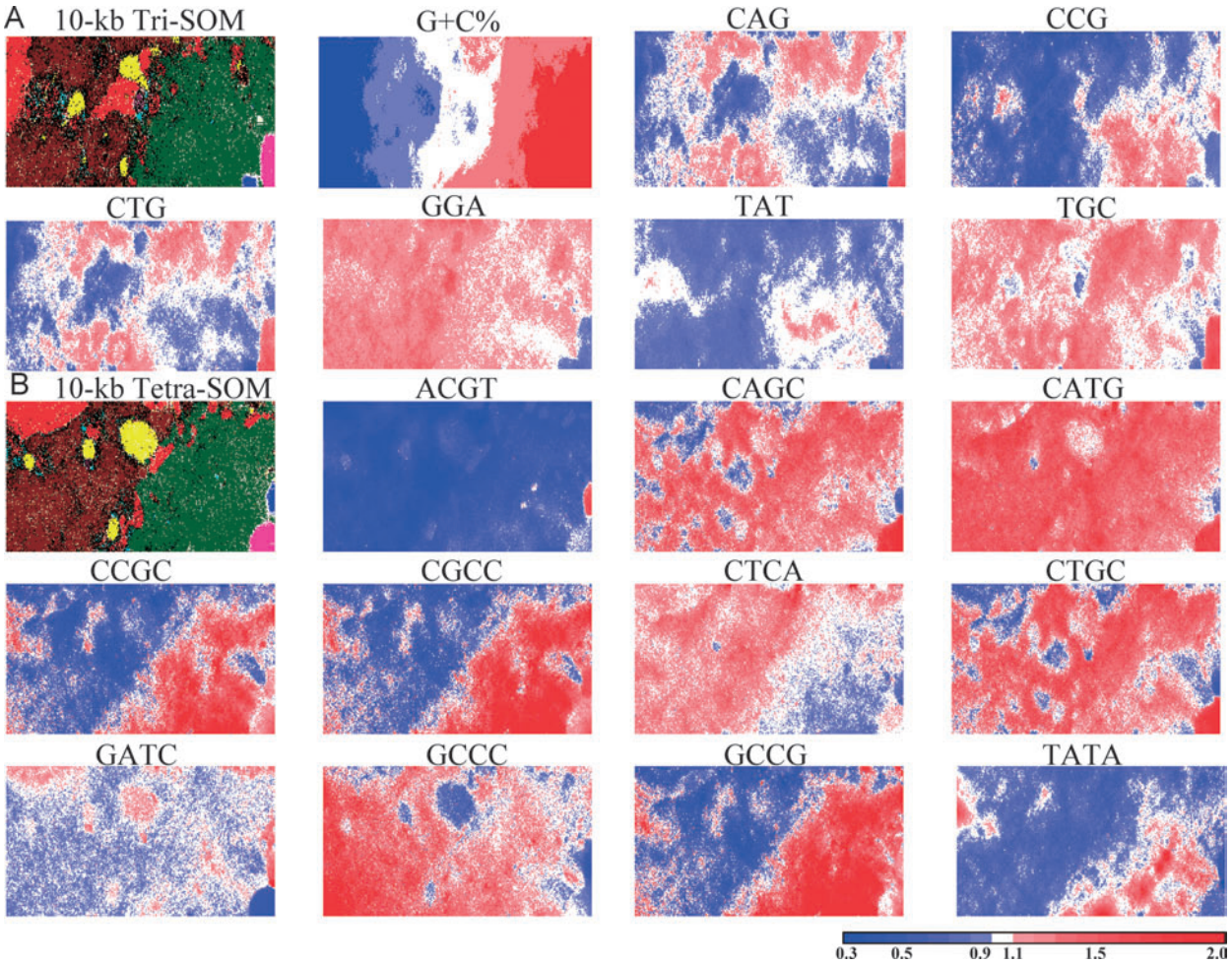


Figure 2. Level of each tri or tetranucleotide in a 10-kb BLSOM. Diagnostic examples of species-specific separations are presented. Level of each tri- or tetranucleotide in each lattice point in the 10-kb Tri- or Tetra-BLSOM in Figures 1 was calculated and normalized with the level expected from the mononucleotide composition of the lattice point. The observed/expected ratio is indicated in colors shown at the bottom of the figure.

signature of each genome and enabled us to identify the frequency patterns that are characteristic of individual genomes; genome signatures. The frequency of each oligonucleotide in each lattice vector in the 10-kb Tri- and Tetra-BLSOMs presented in Figure 1 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 Figure 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 tri- and tetranucleotides often coincided exactly with species borders, and several diagnostic examples of species separations are presented in Figure 2. For example, difference in the frequencies of CG- and GC-containing oligonucleotides was detected and this was found in many other cases not presented in Figures 2; the CG-containing oligonucleotides (but not the GC-containing oligonucleotides) were primarily underrepresented in many species. This may be related to the methylation at CG and CNG site in plant genomes (Cha et al. 2005). CCGC was underrepresented in almost all the sequences from dicot genomes (*Arabidopisis*, papaya, and poplar) and overrepresented in sequences from rice genomes. CAGC was underrepresented in moss and overrepresented in other genomes. Collectively, the frequencies of CG- and CNG-containing oligonucleotides in the dicot genomes were shown to be significantly lower than in the rice (monocot) genome. BLSOMs utilized a complex combination of many oligonucleotides for the sequence separation, which results in classification on the basis of phylogenetic groups such as species.

### *Intraspecies differences revealed in a single genome by addition of random nucleotide sequences*

BLSOM classified genomic sequences into known biological categories (into species in Figure 1) without any information other than oligonucleotide frequencies. Because the classification and visualization power is very high, BLSOM should be a powerful bioinformatics tool for extracting a wide range of genome information. For example, in our previous study that studied the biological implications of diagnostic tetranucleotides in Tetra-BLSOM for bacterial genomes, we found correlation of the frequencies of palindromic tetranucleotides with restriction enzyme systems in the bacteria that produce 4-base cutter restriction enzymes (Abe et al. 2003).

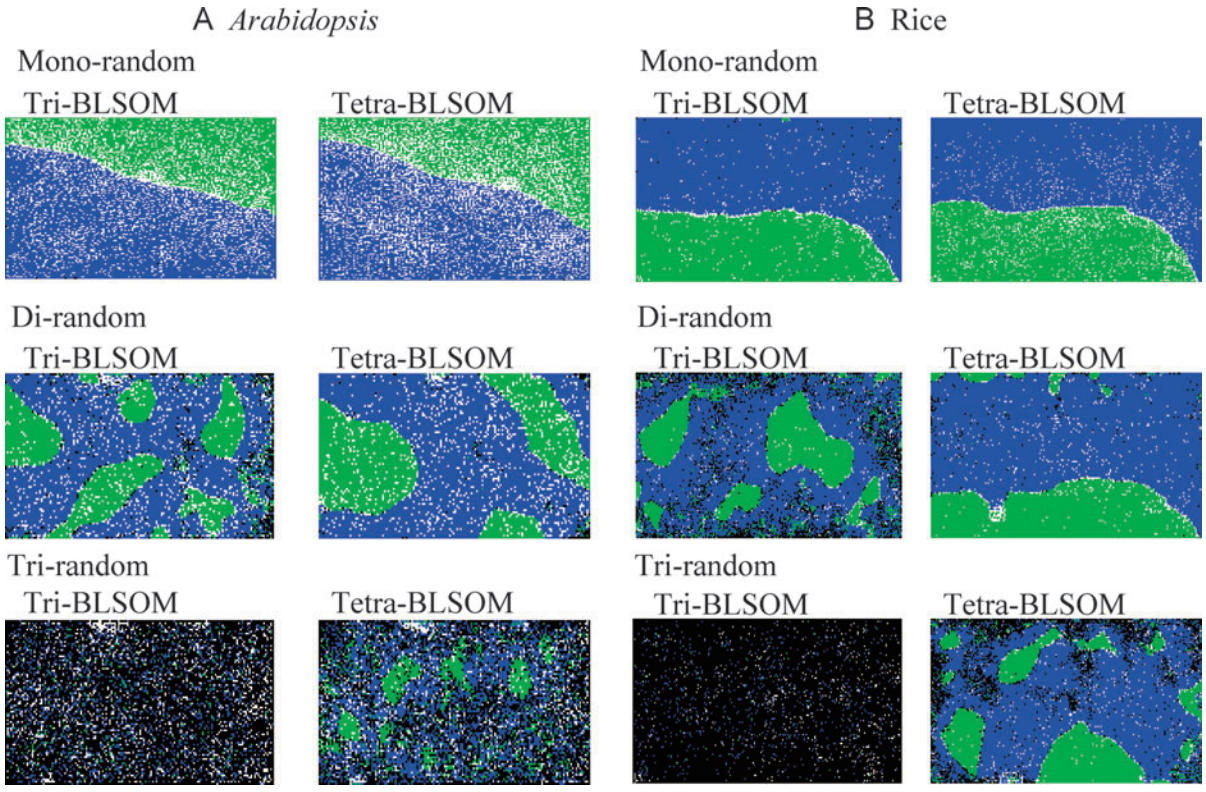


Figure 3. Tri- or Tetra-BLSOM for 10-kb fragment sequences from the *Arabidopsis*. (A) or Rice (B) genome plus computer-generated 10-kb random sequences. Mono-, Di- or Tri-random sequences were added. Lattice points containing only genomic sequences or random sequences are shown in green and blue, respectively, and lattice points including both types of sequences are shown in black.

Restriction site tetranucleotides were characteristically underrepresented in bacteria that produce 4-base cutter enzymes, showing that the Tetra-BLSOM properly recognized this biological property of the genome sequences.

The species-specific characteristics found in Figures 1 and 2 were inevitably dependent on a set of the species included in the BLSOM analysis because the analysis is focusing on the interspecies comparison (a comparative genomics). In order to study the characteristics of oligonucleotide composition in a single plant genome independently of other genomes, we next analyzed the Tri- and Tetra-BLSOMs for 10-kb fragment sequences from the *Arabidopsis* genome plus computer-generated random sequences. For each of the 10-kb sequences from the *Arabidopsis* genome (120 Mb), two 10-kb random sequences with almost the same mono-, di- or trinucleotide composition with the corresponding genomic sequence were generated; for generation of random sequences, refer to Materials and methods. Then, Tri- and Tetra-BLSOMs for the 12,000 10-kb genomic sequences of *Arabidopsis* plus the 24,000 random sequences thus generated were constructed (Mono-, Di- and Tri-random in Figure 3A). In the same way, 10-kb random sequences were generated for each of the rice genomic fragments, and Tri- and Tetra-BLSOM were constructed with 10-kb rice genomic sequences plus the random sequences (Figure 3B).

In the case of addition of the Mono-random sequences, genomic sequences were clearly separated from the random sequences on all BLSOMs (Figure 3), showing BLSOMs able to effectively recognize characteristic bias of oligonucleotide frequencies in the genomic sequences. When Di-random sequences were included in the BLSOM construction, contribution of the characteristics of dinucleotide composition in the genomic sequences was removed, and thus the characteristics of trinucleotide or tetranucleotide composition independent of the characteristics of dinucleotide composition is thought to be sensitively detected by Tri- or Tetra-BLSOM, respectively. In the case of Tri-random addition, Tetra-BLSOM (but not Tri-BLSOM) can detect the characteristics of the genomic sequences.

In the Di- or Tri-random addition shown in Figure 3, not only the separation from the random sequences but also the intraspecies separation was clear with a trivial exceptional case of Tri-BLSOM for Tri-random addition. Because the *Arabidopsis* genome can be methylated at CG and CNG sites, tetranucleotides including such di- and trinucleotides may be the diagnostic oligonucleotides that differentiated genomic sequences from the random sequences. This prediction was confirmed in most of the oligonucleotides containing CG or CNG (Figure 4).

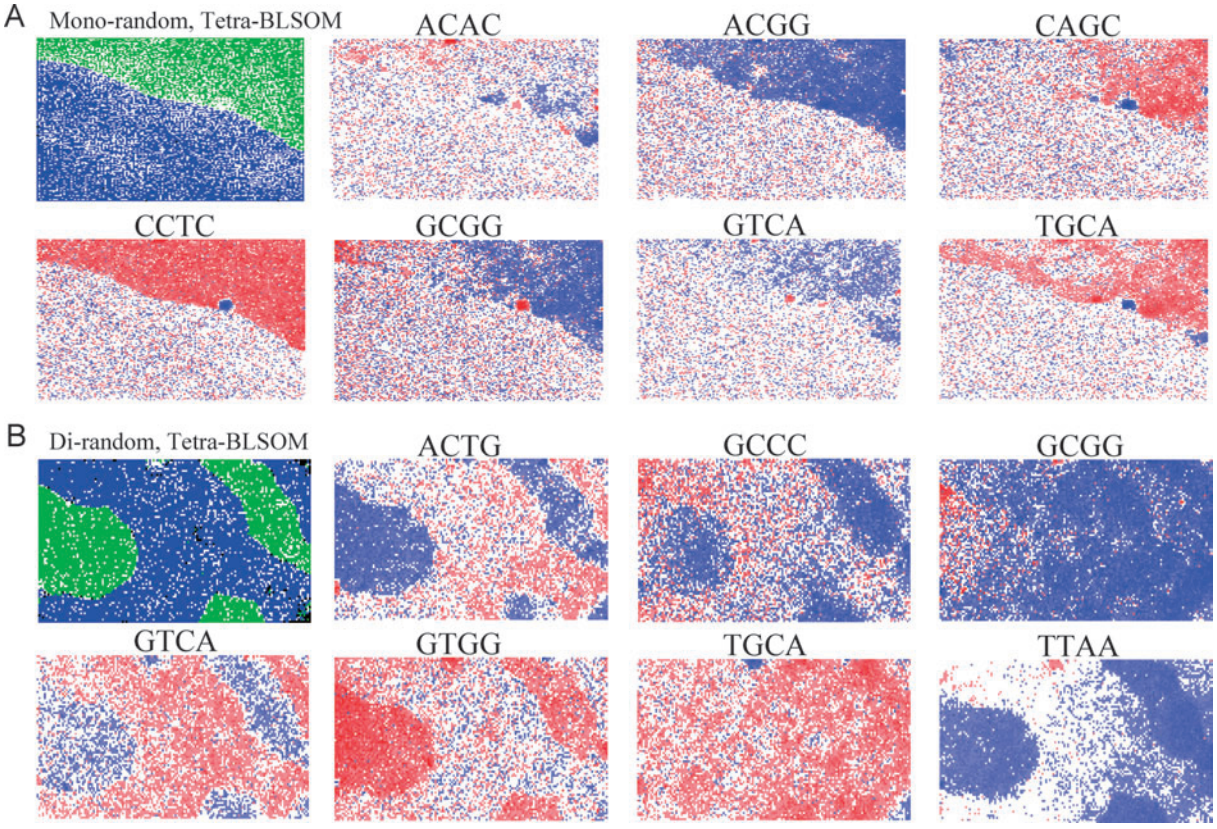


Figure 4. Level of each tetranucleotide in a Tetra-BLSOM for 10-kb *Arabidopsis* sequences plus Mono- or Di-random sequences (A or B, respectively). Diagnostic examples of species-specific separations are presented.

There were also other oligonucleotides diagnostic for genomic sequences. Detailed studies of such characteristic oligonucleotides from various viewpoints may provide fundamental guidelines for understanding biological properties of these oligonucleotides diagnostic for genomic sequences; e.g., genetic signal sequences (Abe et al. 2006a, c).

Transcription regulatory signals are typically located in the close vicinity of the transcription stat site. When we preliminarily investigated the diagnostic tetranucleotides that were underrepresented in the *Arabidopsis* genomic sequences, they often corresponded to the component of regulatory elements at transcription start site (Yamamoto et al. 2007a, b) (Figure 4). Important genetic signals are thought to be primarily underrepresented across the genome. There were, however, the tetranucleotides that were components of regulatory elements but overrepresented in the genomic sequences. These might correspond to the localization-insensitive classes of regulatory elements (Yamamoto et al. 2007a, b).

Genetic signals, such as transcription signals, are typically longer than tetranucleotides, and therefore, analyses of longer oligonucleotides are needed to accurately study the signal sequnces. To conduct BLSOM with longer oligonucleotides such as hexa- and heptanucleotides (4,096- and 16,384-dimensional data), a large-scale computation using a high-performance supercomputer such as "the Earth Simulator" becomes essential because of their high-dimensional vectorial data. BLSOM (but not the conventional SOM) is suitable for actualizing high-performance parallel-computing with high-performance supercomputers such as "the Earth Simulator" (Abe et al. 2006b).

## Discussion

### *Characterization of genetic signal sequences by addition of random sequences*

Because inter- and intraspecies separations on BLSOMs are very clear, extensive studies using the present method should provide fundamental information for understanding the detailed molecular mechanisms that have established species-specific sequence characteristics during evolution. When characteristic oligonucleotides, both underrepresented and overrepresented in each genome, are considered, various factors, including DNA conformational tendencies and context-dependent mutation, repair, and modification, are thought to be responsible (Nussinov 1984; Karlin et al. 1997; Karlin 1998a, b; Rocha et al. 1998; Gentles and Karlin 2001; Pride et al. 2003). With respect to overrepresented sequences, preferences for sequences recognized by ubiquitous DNA-binding proteins and abundant repetitive elements must be considered.

Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). Occurrence levels of the respective signal sequences across the genome are thought to be related to the mechanisms, in which the signal sequences are detected by the regulatory proteins in charge. When an oligonucleotide sequence has a distinct activity such as high-affinity binding with a specific regulatory protein, its occurrence level may be biased from the level predicted as a random event and may vary significantly across the genome. We previously found that occurrence levels of oligonucleotides corresponding to genetic signals such as transcriptional signals were often biased significantly from the levels of random occurrence expected from the nucleotide composition of the genome (Abe et al. 2006a). This indicated the possibility that BLSOM may be useful for a novel tool for characterization and *in silico* prediction of genetic signal sequences. The oligonucleotide sequence with a high affinity for a transcription factor would be underrepresented across most regions of the genome but would be more prevalent in regions that regulate gene expression; such sequences would be underrepresented across the entire zone of the BLSOM with a wide window (e.g., 100-kb) but would occur at higher frequencies in restricted portions of the BLSOM with a much narrower window (e.g., 1-kb). In contrast, when a certain signal sequence occurs across the genome at frequencies similar to or higher than that predicted by random occurrence, coexistence of this sequence with other closely-situated signal sequences being detected with other regulatory proteins should be a prerequisite for the sequence to function as a regulatory signal. Such information regarding frequency of the oligonucleotide with a binding activity to a regulatory protein may provide insight into the mutual role of oligonucleotides that comprise combinatorial units for transcriptional regulation.

### *In silico prediction of genetic signal sequences*

As mentioned above, 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 often diagnostic for the species-specific separations (Abe T et al. 2006a). When known signal sequences of various species with enough experimental data are characterized and categorized systematically with BLSOMs, we can possibly develop an *in silico* method of signal prediction most useful for genomes that were 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. In this respect, analyses

of density distribution of the candidate oligonucleotide for the genetic signals along the genome sequence will become important. We previously reported examples of such analyses of mammalian genomes (Abe et al. 2006c).

### *BLSOM application for predicting gene functions*

For almost half of genes from novel genomes sequenced, it has become clear that protein functions cannot be estimated through sequence homology search. We have recently applied BLSOM to protein sequence studies to analyze the frequency of oligopeptides and found the separation (self-organization) of proteins according to their functions (Abe et al. 2009). This indicates that the BLSOM can be used as a protein function estimation that does not rely on the sequence homology search and the troublesome and mistakable sequence-alignment. Large-scale BLSOM analyses of a large amount and a wide variety of genome and protein sequences facilitates efficient extraction of fundamental information, which supports researches and developments in a broad range of life sciences and industrial fields.

## Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C) and for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

## References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2002) A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Inform* 13: 12–20

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. *Genome Res* 13: 693–702

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel Phylogenetic Studies of Genomic Sequence Fragments Derived from Uncultured Microbe Mixtures in Environmental and Clinical Samples. *DNA Res* 12: 281–290

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

Abe T, Sugawara H, Kanaya S, Ikemura T (2006b) 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. *J Earth Simulator* 6: 17–23

Abe T, Sugawa H, Kanaya S, Kosaka Y, Ikemura T (2006c) Characterization of transcriptional regulatory signals with novel bioinformatics tools. In: Kiyama R, Shimizu M (eds) *DNA structure, Chromatin and Gene Expression*, Transworld Research Network, India, pp 1–16

Abe T, Kanaya S, Ikemura T (2009) A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses. *DNA Res* 16: 287–298

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

Bernardi, G (2004) *Structural and evolutionary genomics: natural selection in genome evolution.* Elsevier, Amsterdam; New York

Chan SW, Henderson IR, Jacobsen SE (2005) Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat Rev Genet* 6: 351–360

Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* 11: 540–546

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. *Nucl Acids Res* 8: r49–r62

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2: 13–34

Ikemura T, Aota S (1988) Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures. *J Mol Biol* 203: 1–13

Hayashi H, Abe T, Sakamoto M, Ohara H, Ikemura T, Sakka K, Benno Y (2005) Direct cloning of genes encoding novel xylanases from human gut. *Can J Microbiol* 51: 251–259

Kanaya S, Kudo Y, Nakamura Y, Ikemura T (1996) 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

Kanaya S, Kudo Y, Abe T, Okazaki T, Carlos DC, Ikemura T (1998) Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome. *Genome Inform* 9: 369–371

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) 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

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nshi T, Mori H, Ikemura T (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map: characterization of horizontally transferred genes with emphasis on E. coli O157 genome. *Gene* 276: 89–99

Karlin S, Mrazek J, Campbell A (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 179: 3899–3913

Karlin S (1998a) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* 1: 598–610

Karlin S, Campbell A, Mrazek, J (1998b) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* 32: 185–225

Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43: 59–69

Kohonen T (1990) The self-organizing map. *Proc IEEE* 78: 1464–1480

Kohonen T, Oja E, Simula O, Visa A, Kangas J (1996) Engineering applications of the self-organizing map. *Proc IEEE* 84: 1358–1384

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. *Nucl Acids Res* 12: 1749–1763

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 13: 145–158

Rocha EP, Viari A, Danchin A (1998) Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nucl Acids Res* 26: 2971–2980

Sharp PM, Matassi G (1994) Codon usage and genome evolution. *Curr Opin Gen Dev* 4: 851–860

Uchiyama T, Abe T, Ikemura T, Watanabe K (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol* 23: 88–93

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J (2007a) Differentitation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucl Acid Res* 35: 6219–6226

Yamamoto YY, Ichida H, Matui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T (2007b) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* 8: 67

Zhang R, Zhang CT (2004) Isohore structures in the genome of the Plant Arabidopsis thaliana. *J Mol Evol* 59: 227–238