Characterization of Genetic Signal Sequences with Batch-Learning SOM

Takashi Abe[a,b]*, Shun Ikeda[a], Shigehiko Kanaya [c], Kennosuke Wada[d], and Toshimichi Ikemura[a]

[a]Nagahama Institute of Bio-Science and Technology, Tamura-cho 1266, Nagahama-shi, Shiga-ken 526-0829, Japan. [b]National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan. [c]Nara Institute of Science and Technology, Ikoma, Japan. [d]Untrod Inc., Keihanna Plaza, Seika-cho Soura-gun, Nara 619-0237, Japan.
E-mail: takaabe@nagahama-i-bio.ac.jp

Keywords: batch learning SOM, BL-SOM, oligonucleotide frequency, the Earth Simulator, genome informatics

Abstract**—** An unsupervised clustering algorithm Kohonen's SOM is an effective tool for clustering and visualizing high-dimensional complex data on a single map. We previously modified the conventional SOM for genome informatics, making the learning process and resulting map independent of the order of data input on the basis of Batch Learning SOM (BL-SOM). We generated BL-SOMs for tetra- and pentanucleotide frequencies in 300,000 10-kb sequences from 13 eukaryotes for which almost complete genomic sequences are available. BL-SOM recognized species-specific characteristics of oligonucleotide frequencies in most 10-kb sequences, permitting species-specific classification of sequences without any information regarding the species. We next constructed BL-SOMs with tetra- and pentanucleotide frequencies in 37,086 full-length mouse cDNA sequences. With BL-SOM we also analyzed occurrence patterns of the oligonucleotides that are thought to be involved in transcriptional regulation on the human genome.

1 Introduction

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. 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 parameter to characterize individual genomes and genomic portions. The G+C%, however, is too simple to differentiate wide varieties of genomes. Since oligonucleotide frequency, which is an example of high-dimensional data, varies significantly among genomes, this frequency can be used for comparative genome study (1; see also references cited by this paper). 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 (2-4). On the basis of Batch-Learning SOM (BL-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 (1, 5-10). 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 (i.e., self-organized) according to species without any information regarding the species, and increasing the length of the oligonucleotides from di- to tetranucleotides increased the clustering power (1). For investigating the power to detect differences among a wide range of eukaryotes, tri- and tetranucleotide frequencies in 10- and 100-kb sequence fragments derive from 40 eukaryotic genomes, which have been sequenced extensively, were next analyzed (10). To analyze a massive amount of eukaryotic genome sequences and visualize on a single map, the Earth Simulator (11), which is one of the highest performance supercomputers in the world, was used in that study. In the present study, to furtherer improve the power for detecting differences among closely related eukaryotes and resolving intraspecies differences, we examined the frequencies of tetra- and pentanucleotides in a wide rage of eukaryotes, using the Earth Simulator and focused on the oligonucleotide frequencies in connection with characterization of genetic signal sequences including those for transcriptional regulation.

2 Methods

Multivariate analyses such as factor corresponding analysis and principal component analysis (PCA) have been used successfully to investigate variations in gene sequences. 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. SOM implements

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

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 (2-4). 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 (5,6). 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 (6). Weight vectors (wij) were arranged in the two-dimensional lattice denoted by i (=0, 1, .., I-1) and j (=0, 1, .., J-1). I was set as 250, and J was defined by the nearest integer greater than (σ2/σ1) x I. σ1 and σ2 were the standard deviations of the first and second principal components, respectively. Weight vectors (wij) were set and updated as described previously (1). The BL-SOM program could be obtained from G-inforBIO (http://wdcm.nig.ac.jp/inforbio/) and from UNTROD, Inc. (http://untrod.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.

3 Results

3.1 BL-SOMs for 13 eukaryotic genomes

To investigate clustering power of BL-SOM for eukaryotic sequences, we first analyzed tetra- and pentanucleotide frequencies in 300,000 non-overlapping 10-kb sequences and overlapping 100-kb sequences with a 10-kb sliding step from 13 eukaryotic genomes (a total of 3 Gb). These genomes included human *Homo sapiens*, puffer fish *Fugu rubripes*, zebrafish *Danio rerio*, rice *Oryza sativa, Arabidopsis thaliana*, *Medicago truncatula*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Dictyostelium discoideum, Plasmodium falciparum*, *Entamoeba histolytica, Schizosaccharomyces pomb*, and *Saccharomyces cerevisiae*. The BL-SOM, which was adapted for genome informatics, was constructed as described previously (1). First, oligonucleotide frequencies in the 300,000 10-kb sequences were analyzed by PCA, and the first and second principal components were used to set the initial weight vectors, which were arranged as a two-dimensional array. After 80 learning cycles, oligonucleotide frequencies in the 10-kb sequences were represented by the final weight vectors in the two-dimensional array, and the resulting BL-SOM revealed clear species-specific separations (Fig. 1). The sequences were clustered (self-organized) primarily into species-specific territories; nodes 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 each species were clustered tightly on these tetra- and pentanucleotide BL-SOMs (abbreviated as Tetra- and Penta-SOMs, respectively); e.g., 97% and 98% of analyzed human sequences were classified into the human territories (🟥 in Fig. 1) in the 10-kb Tetra- and Penta-SOMs, respectively.

The G+C% obtained from the weight vector for each node in these BL-SOMs was found to reflect in the horizontal axis and increased from left to right. Sequences with high G+C% were located on the right side of the Tetra- and Penta-SOMs (data not shown). Sequences with the same G+C% were separated by a complex combination of oligonucleotide frequencies resulting in species-specific separations (1). In the 10-kb BL-SOMs, intraspecies separations were observed; e.g., human was divided into two major territories in the 10-kb Tetra-SOM (Fig. 1A). In the Penta-SOM, however, human sequences were classified into a single continuous territory. This indicated that despite wide variations among 10-kb segments of human sequences, the BL-SOM recognized common features of pentanucleotide frequencies in human sequences. In the 100-kb BL-SOMs, interspecies (but not intraspecies) separations were more prominent than in the 10-kb BL-SOMs; in the 100-kb Tetra- and Penta-SOMs, all species had one major territory. Furthermore, the species territories were surrounded by contiguous white lattice points, which contained no genomic sequences. The species borders could be drawn automatically on the basis of the contiguous white nodes, because the vectors of the species-specific nodes that were located even near a territory border were distinct between territories.

3.2 Diagnostic oligonucleotides for species separations

BL-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. The frequency of each oligonucleotide in each weight vector in the 100-kb BL-SOMs was calculated and normalized with the level expected from the mononucleotide composition at each node, and the observed/expected ratios are illustrated in red (overrepresented), blue (underrepresented), or white (moderately represented) in Fig. 2. This normalization allowed us to study oligonucleotide frequencies in each node independently of mononucleotide composition (1). Transitions between red (overrepresentation) and blue (underrepresentation) for various tetra- and pentanucleotides often coincided exactly with species borders. Several diagnostic examples of tetranucleotides for the species separations are presented in Fig. 2A. AATT was overrepresented in rice, *Drosophila*, and *C. elegans*, underrepresented in *Fugu* and zebrafish, and moderately represented in human and *Arabidopsis*. CAGT was overrepresented in all three vertebrates but underrepresented in rice, *Arabidopsis, Plasmodium,* and *Dictyostelium*. Results for a pair of complementary tetranucleotides was nearly identical, and only data for one tetranucleotide are presented. Fifteen examples of pentanucleotides diagnostic for species separation are presented in Fig. 2B. BL-SOMs utilized a complex combination of many oligonucleotides for sequence separations, which resulted in classification (self-organization) of sequences according to species.

3.3 Classification of gene and genomic sequences of mouse and human according to functional categories

In the era of extensive genome sequencing, it is important to predict numerous functions of genomic sequences utilizing increasingly available DNA sequences. Sequencing of cDNAs derived from RNA transcripts is one of most promising source of information useful for functional prediction of gene sequences. Efforts to determine full-length cDNA sequences and catalogue the transcripts provide essential tools to facilitate functional analysis of the transcripts, and recent studies have been extended to non-protein-coding transcripts (ncRNA) (12,13). For example, an international collaborative study to analyze the full-length mouse cDNA sequences reported that ncRNAs may be one major component of the transcriptome (14). In addition to the roles in protein synthesis (ribosomal and transfer RNAs), ncRNAs have been implicated in roles that require highly specific nucleic acid recognition, such as in directing post-transcriptional regulation of gene expression or in guiding RNA modifications. Even in the case of protein-coding cDNAs, it has become increasingly important to predict functions of untranslated regions (UTRs). The 5'- and 3'-UTRs of eukaryotic mRNAs play a crucial role in post-transcriptional regulation of gene expression modulating nucleo-cytoplasmic mRNA transport, translation efficiency, subcellular localization, and stability. New systematic approaches are needed for comprehensive analyses of massive amounts of available cDNA sequences, which can be aimed not only at protein-coding sequences (CDSs) but also at UTRs and ncRNAs.

We next constructed BL-SOMs for tetra- and pentanucleotide frequencies (Tetra- and Penta-SOMs in Fig. 3) in the 37,086 full-length mouse cDNA sequences (14). Nodes that contain only the sequences for protein-coding cDNAs are marked in violet, those containing only the sequences for ncRNAs are marked in red, and those containing the sequences of both categories are marked in black. A major portion of the sequences of the two categories was separated from each other on these BL-SOMs. In the right-hand side, there was one broad satellite zone of ncRNAs, where ncRNAs with CpG islands were enriched. Detailed investigation of the number of cDNAs classified into each node showed that evident clustering of protein-coding cDNAs in many nodes. To show this graphically, the number of sequences classified into each node was represented by the height of the vertical rod using a color specifying the category (3D in Fig. 3). One factor responsible for the separation between protein-coding and non-coding cDNAs might be the characteristics derived from codon usage pattern which can be defined only in CDSs in the protein-coding cDNAs. However, it is also conceivable that characteristics even of UTR sequences differ from those of ncRNAs. To examine these possibilities, the 5' and 3' UTR and CDS sequences of protein-coding cDNAs were separately analyzed, together with ncRNAs. To avoid potential artefacts caused by redundant sequences of UTRs, we used mouse UTR sequences compiled in UTRdb (http://www.ba.itb.cnr.it/UTR/), which is a specialized database of 5' and 3' UTRs of eukaryotic mRNAs cleaned from redundancy. Furthermore, in UTRdb, polyA-tail sequences in 3' UTRs are removed systematically, and this is crucial for omitting trivial, evident effects of polyA-sequences on oligonucleotide frequency in 3' UTRs. To get statistically meaningful results, UTR sequences shorter than 100 nucleotides were omitted from this analysis. Clear separation among the four functional categories (5' and 3' UTRs, CDSs, and ncRNAs) was observed on Tetra- and Penta-SOMs, and results of Penta-SOM are presented in Fig. 4. A major portion of 3' UTRs was located in the left-hand and bottom part and a major portion of 5' UTRs was located in the right side. A major portion of ncRNAs was located in the upper part, but there was one satellite zone in the right, lower side closely associated with the 5' UTR territory, where ncRNAs with CpG islands were enriched. The finding that a major portion of ncRNAs was located separately from 3' and 5' UTRs showed that the separation between protein-coding and non-coding sequences found in Fig. 3 was not due to a simple reflection of codon usage patterns in CDSs.

3.4 Characterization of upstream region of transcription stat site

Regulatory signals of transcription are typically located in the close proximity of the transcription stat site (TSS). Next we focused on 5-kb sequences upstream of human TSSs available from UCSC Genome Bioinformatics Site (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/). We divided the 20,647 5-kb sequences into 1-kb fragments and constructed Penta-SOM with the 103,235 1-kb sequences (Fig. 5A). Distribution pattern of the 0-1 kb region from TSS was most characteristic and different from patterns of other 1-kb segments. This should be because a major portion of signal sequences for transcriptional regulation are located in the close proximity of TSS and thus the 0-1 kb was richest of the characteristic oligonucleotides compared with other segments. Actually the diagnostic pentanucleotides for the 0-1 kb corresponded often to transcription regulatory signals or their constitute elements (Fig. 5B). BL-SOM can properly point out candidates of functional oligonucleotides in the functionally important genome regions, by showing their occurrence level is clearly distinct from other genomic regions.

4 Discussion

Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). Because BL-SOM analysis is dependent only on oligonucleotide frequencies, this method is applicable even for the genomes sequenced but with little additional experimental data. In order to know TSS, experimental data of transcription were required, but to know start sites of protein-coding sequences, such experimental data were not required in most cases because a large portion of protein-coding ORFs can be predicted computationally. Because TSSs are mainly located in the close proximity of the start site of protein coding ORFs, analyses on regulatory signals for gene expression will be conducted even using the start site of protein coding ORFs. When known signal sequences of various species with enough experimental data are characterized systematically, we can develop an *in silico* method of signal sequence prediction for a wide range of species. Genetic signals, such as transcription regulatory signals, are often longer than pentanucleotides, and therefore, analyses of longer oligonucleotides will be needed to test this possibility.

Recognition mechanisms of genetic signal sequences with the cognate proteins and occurrence levels of the respective oligonucleotide sequences are thought to be related with each other. When an oligonucleotide sequence has a distinct activity, such as high-affinity binding with a specific functional protein, the occurrence level of the respective oligonucleotide may be biased from that predicted by random assortment and may vary significantly across the genome. For example, an 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. In other words, such sequences would be underrepresented across the entire zone of the BL-SOM with a wide window (e.g., 100 kb) but would occur at higher frequencies in restricted portions of the BL-SOM with a much narrower window (e.g., 1 ~ 10 kb). In contrast, when some signal sequences occur across the genome at frequencies similar to or higher than those predicted by random occurrence, combination with other signal sequences closely situated should be a prerequisite for the sequence to function as a regulatory signal. The TRANSCompel database (http://compel.bionet.nsc.ru/new/compel/compel.html) contains data regarding combinatorial regulatory units composed of two *cis* binding elements closely situated. BL-SOM data concerning levels of oligonucleotides may provide insight into the mutual role of oligonucleotides that comprise combinatorial units for transcriptional regulation. By referring to the behaviors of signal sequences determined for well-studied organisms, we can possibly develop an *in silico* method to predict signal sequences in genomes that have been sequenced but for which there is little additional experimental data. Recently, we have applied the BL-SOM method to phylogenetic classification of novel genomic sequences derived from environmental, mixed genomes (15-18).

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) and for Grant-in-Aid for Scientific Research on Priority Areas "Applied Genomics" 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. 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.

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

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

[4] 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.

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

[6] 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.

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

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

[9] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, Y. Matsuura, H. Tokutaka, and T. Ikemura, "A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide range of eukaryotic genomes" *WSOM 2005,* pp. 187-194, 2005.

[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, S. Kanaya, and T. Ikemura, "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", *Journal of the Earth Simulator*, **vol. 6**, pp.17-23, 2006.

[12] A. Huttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J. Bachellerie, and J. Brosius, "RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse", *EMBO* **vol. 20**, pp. 2943-2953, 2001.

[13] C. Marker, A. Zemmann, T. Terhorst, M. Kiefman, J. Kastenmayer, P. Green, J. Bachellerie, J. Brosius, and A. Huttenhofer, "Experimental RNomics: Identification of 140 Candidates for Small Non-Messenger RNAs in the Plant *Arabidopsis thaliana*", *Current Biology* **vol. 12**, pp. 2002-2013, 2002.

[14] Y. Okazaki, M. Furuno, T. Kasukawa, et al., "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs", *Nature* **vol. 420**, pp. 563-573, 2002.

[15] 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 Res.,* **vol. 12**, pp. 281-290, 2005.

[16] 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.

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

[18] T. Abe, H. Sugawara, S. Kanaya, and T. Ikemura, "A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of environmental uncultured microbes", *Polar Bioscience,* **vol. 20**, pp. 103-112, 2006.

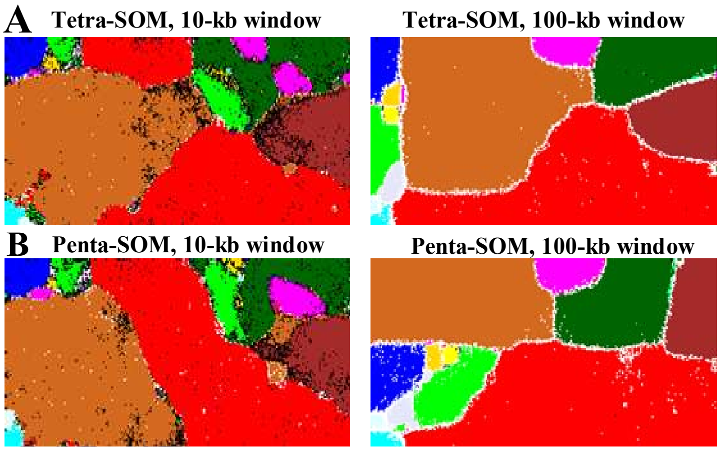

Fig. 1 BL-SOMs for non-overlapping 10-kb and overlapping 100-kb sequences of 13 eukaryotic genomes. (A) Tetra-SOMs. (B) Penta-SOMs. Nodes that include sequences from plural species are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single species are indicated in color as follows: *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* (🟧), and *Homo sapiens* (🟥).

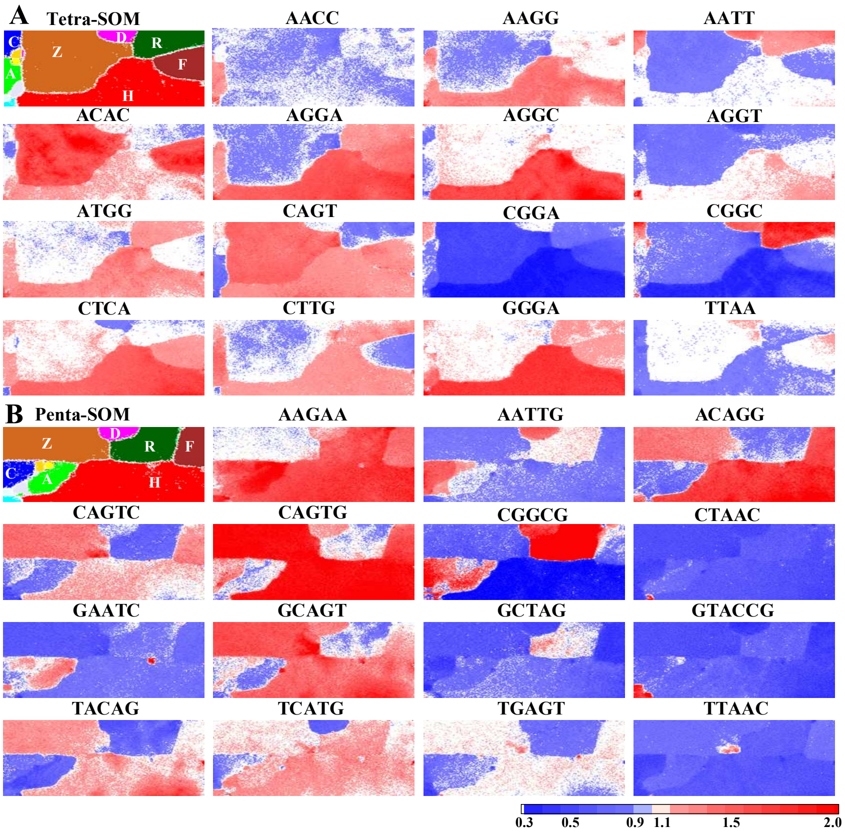

Fig. 2 Level of each tetranucleotide (A) and pentanucleotide (B) in 100-kb BL-SOMs. Diagnostic examples of species separations are presented. Level of each tetra- and pentanucleotide in each node in the 100-kb Tetra- and Penta-SOMs (Fig. 1) was calculated and normalized with the level expected from the mononucleotide composition of the node. The observed/expected ratio is indicated in colors shown at the bottom of the figure. Results for other oligonucleotides are presented by our Web site (http://lavender.genes.nig.ac.jp/takaabe/WSOM2007/WSOM2007.html). The 100-kb BL-SOMs in Fig. 1A and B are presented in the first panel with letters indicating species name: *C. elegans* (C), *Arabidopsis* (A), rice (R), *Drosophila* (D), *Fugu* (F), zebrafish (Z), and human (H). For other species, refer to the colors in the Fig. 1 legend.

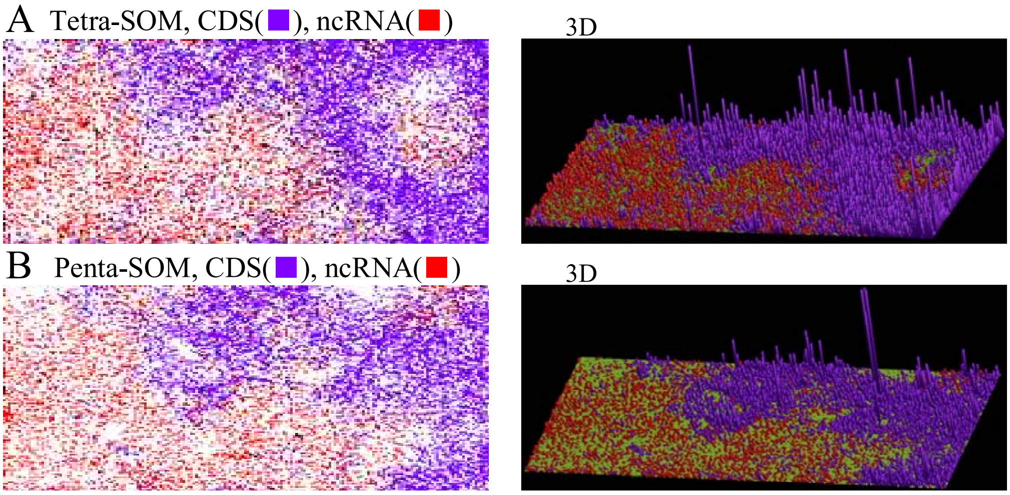

Fig. 3 BL-SOM for mouse cDNA sequences. (A and B) Tetra- and Penta-SOMs, respectively. Nodes that contain only the sequences for protein-coding cDNAs are marked in violet, those containing only the sequences for ncRNAs are marked in red, and those containing the sequences of both categories are marked in black. 3D: three-dimensional presentation of the BL-SOMs.

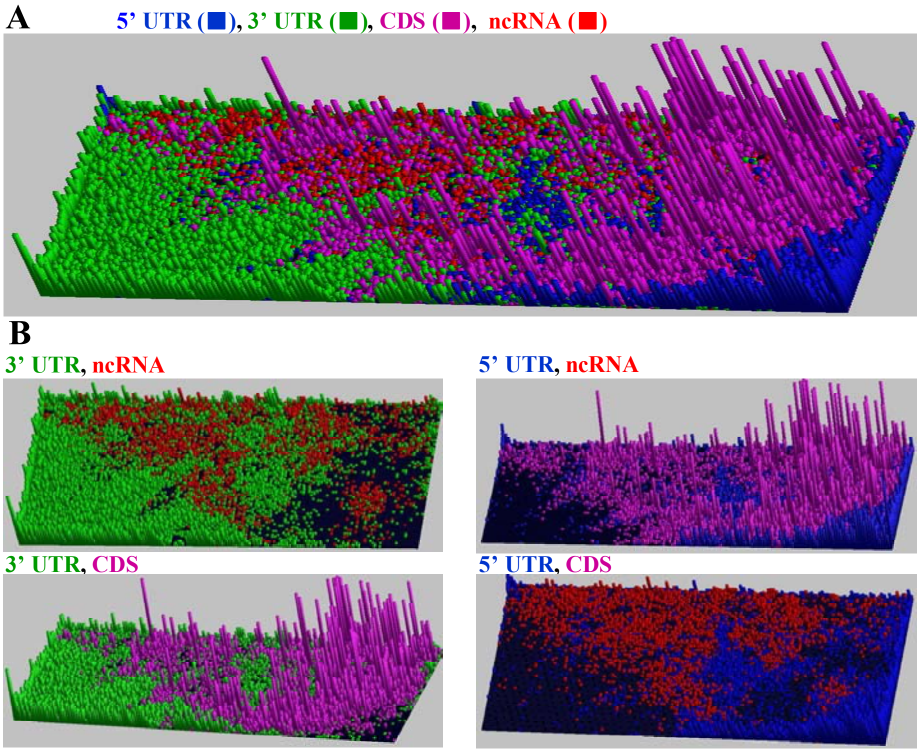

Fig. 4 BL-SOM for mouse cDNA and UTR sequences. (A) Three-dimensional presentation of four functional categories (5' and 3' UTRs, CDSs, and ncRNAs) on Penta-SOM. Nodes that contain only the sequences for 3' and 5' UTRs are marked in green and blue, respectively, and those containing only the sequences for protein-coding cDNAs and ncRNAs are marked in violet and red, respectively. (B) Comparison of sequence location between two functional categories (3' UTR and ncRNA; 3' UTR and CDS; 5' UTR and ncRNA; 5'UTR and CDS). Nodes that contain only the sequences of one category are marked in the color representing the category.

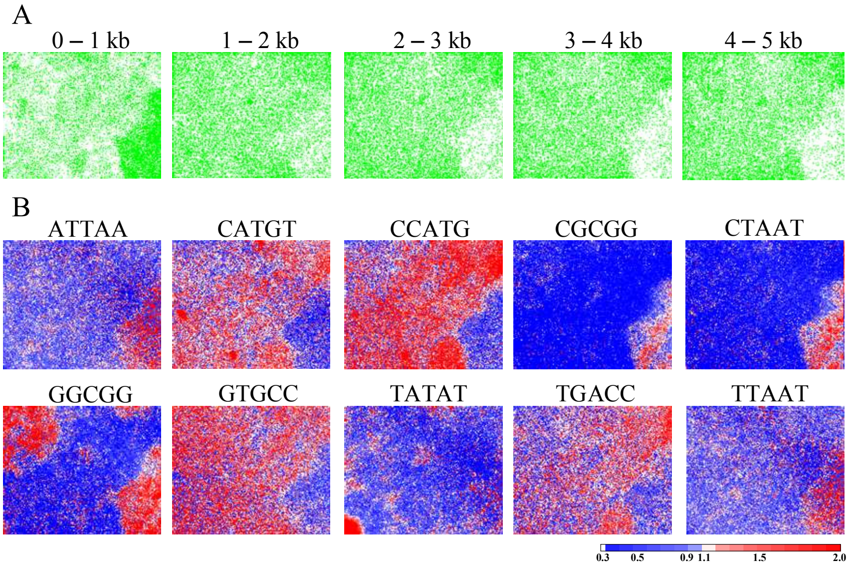

Fig. 5 Penta-SOM for human genomic sequences upstream of transcriptional state sites. (A) Each of the 20,647 5-kb sequences upstream of transcriptional state sites (TSSs) was divided into 1-kb fragments as follows: 0-1 kb, 1-2 kb, 2-3 kb, 3-4 kb, and 4-5 kb from TSS. Penta-SOM was constructed with the 103,235 non-overlapping 1-kb sequences. (B) Level of a pentanucleotide at each lattice point in the Penta-SOM was calculated and is indicated in colors as described in Fig. 2. Patterns of ten pentanucleotides, which is known to be a transcription factor-binding sequence, are presented.