Development of a novel genome informatics strategy on the basis of Self-Organizing Map (SOM)

Takashi Abe

Doctor of Philosophy

Acknowledgments

I am deeply grateful to Prof. Toshimichi Ikemura, Department of Genetics, School of Life Science, The Graduated University for Advanced Studies, and Department of Evolutionary Genetics, National Institute of Genetics, for his helpful advises and kind support throughout this work. He kindly accepted me as his laboratory member, and he has spent a lot of time to guide me to what I am now.

I am also greatly indebted to Prof. Shigehiko Kanaya, Nara Institute of Science and Technology, and Prof. Makoto Kinouchi, Yamagata University, for their valuable suggestions, discussions, and constant encouragement. I am also greatly indebted to Mr. Tokio Kozuki, Dr. Satoshi Nakagawa, and Mr. Tamotsu Eguchi, Xanagen Inc., for their valuable suggestions and constant encouragement. I am also greatly indebted to Dr. Atsushi Fukushima, and Mr. Yuta Ichiba, Department of Evolutionary Genetics, National Institute of Genetics, for discussion and constant encouragement. Acknowledgments are also made to Mr. Yutaka Ado and Ms. Chizuko Ohuchi, Koyowa Hakko Kogyo Co., Ltd., and Mr. Akira Ohyama, Mitsui Knowledge Industry Co., Ltd. The helpful assistance and collaborations of Ms. Yoko Kosaka, Ms. Nanayo Ishihara, Ms. Kimie Amano, Ms. Song Jian-Ping, and Ms. Kiyomi Kita, Department of Evolutionary Genetics, National Institute of Genetics.

Acknowledgments are also made to Prof. Hideaki Sugawara, Prof. Takashi Gogobori, and Prof. Toshihiko Shiroishi, Department of Genetics, School of Life Science, The Graduated University for Advanced Studies, and Prof. Masahira Hattori, Kitasato Institute for Life Sciences, Kitasato University, for serving on the supervisory committee for my thesis.

I wish to express my gratitude to all the members of the Department of Evolutionary Genetics, National Institute of Genetics, Xanagen Inc., Laboratory of Comparative Genomics, Division of Structural and Functional Genomics, Department of Bioinformatics and Genomics, Nara Institute of Science and Technology, and Department of Bio-System Engineering, Faculty of Engineering, Yamagata University.

During my study, I used the supercomputer of DDBJ, which sometimes interrupted DDBJ activities. I thank all the DDBJ staff for their patience in allowing me to use the super computer.

Finally, I would like to express special thanks to my family for their never changing support.

Contents

ABSTRACT

With the increasing amount of available genomic sequences, novel tools are needed for comprehensive analysis of species-specific sequence characteristics for a wide variety of genomes. Self-Organizing Map (SOM), which was developed by Kohonen to study memory and recall/association mechanisms, can identify and associate similar types of information and localize such information in close vicinity on a two-dimensional map. SOM has been proven to be a powerful unsupervised algorithm and applied in various fields of science and technology (e.g., complex industrial processes, document and image databases, and financial applications) but rarely been applied to analysis of genome sequences. In this thesis study, on the basis of batch-learning SOM (BL-SOM), I modified the conventional SOM for genome informatics to make the learning process and resulting map independent on the order of data input. The initial weight vectors of the Kohonen's conventional SOM were usually set by random values, but the vectors in my method were initialized by principal component analysis (PCA) to obtain the same result between different caluculations. I furthere modified BL-SOM to execute parallel processing with supercomputers and PC-clusters and thus could analyze a vast amount of available genomic sequences.In this thesis study, I used the modified SOM to analyze short oligonucleotide frequencies (di- to pentanucleotide frequency) in a wide variety of prokaryotic and eukaryotic genomes.

When only fragments of genomic sequences (e.g., 10-kb sequences) from mixed genomes of multiple organisms are available, it would appear to be impossible to identify how many and what types of genomes are present in the collected sequences. However, I found that the modified SOM could classify the sequence fragments according to species without any information other than oligonucleotide frequencies. I constructed SOMs of di-, tri-, and tetranucleotide frequencies in 1- and 10-kb sequences from prokaryotic and eukaryotic genomes for which complete sequences are available. SOM recognized, in most 10-kb sequences, species-specific characteristics of oligonucleotide frequencies (key combinations of oligonucleotide frequencies), permitting species-specific classification of sequences without any information regarding species.

A majority of environmental microorganisms, especially those living in

extreme environments, are difficult to culture in the laboratory. Because conventional experimental approaches have been unsuccessful, these genomes have remained uncharacterized, and there is the possibility that such genomes contain a wide range of novel genes that would be of scientific and/or industrial interest. Metagenomics, which is genomic analysis of uncultured microorganisms, has been proposed to study microorganism diversity in a wide variety of environments and to identify novel and industrially useful genes. In the metagenome analysis of uncultured microorganisms, genome DNAs are extracted directly from an environmental specimen that contains multiple organisms, and the genomic fragments are then cloned and sequenced. With a simple collection of fragmental sequences, it appears to be impossible to predict what kinds and the ratios of species present in an environmental sample, to which lineages the species belong, and how the genomes are novel. To establish SOM as a methodology suitable to this purpose, I constructed SOMs of tetranucleotide frequencies in 1- and 5-kb sequences from approximately 80 bacterial genomes for which complete sequences are available. Sequences were clustered primarily according to species and to 11 major bacterial groups without any information regarding the species. With this SOM method, all sequences in DNA databases that were from unidentified or uncultured bacteria and longer than 1 kb were classified into 11 major bacterial groups. The result indicated that the method is useful also for survey of pathogenic microorganisms causing novel, unclear infectious diseases.

Next, I analyzed tetra- and pentanucleotide frequencies in the human genome, and found that frequencies and distributions of oligonucleotide sequences involved in transcriptional regulation were often biased significantly from random occurrence. I could categorize occurrence patterns and frequencies of known signal sequences in the human genome. When known signal sequences from various species with sufficient experimental data are characterized and categorized systematically with SOMs, it should be possible to develop an *in silico* method to predict signal sequences, which is thought to be most useful for identification of signal sequences in genomes for which only sequence data are available. Because the number of such poorly characterized genomes becomes high, development of such an *in silico* method has become increasingly important. I have developed SOM as a methodology just suitable to this purpose.

In addition to protein-coding sequences (CDSs), the flanking regions upstream of transcription start sites and the 5’ and 3’ untranslated regions (UTRs) have attracted attention because of their crucial roles in transcriptional and post-transcriptional regulation of gene expression. By combining analyses on cDNA and genomic sequences of human and mouse, I developed SOM to characterize the six functional regions, 5’ and 3’ UTRs, CDSs, introns, 5’ flanking regions, and ncRNAs, in these genomes and to identify hidden sequence characteristics in the functional regions. Because clustering power of SOM is very high, I propose that SOM can provide fundamental guidelines for understanding molecular processes and mechanisms that have established sequence characteristics of individual genomes and genomic regions during evolution.

Chapter I: Introduction

It is one of the most important tasks of life science to obtain unknown basic knowledge from information of a large amount of genome sequences. With the increasing amount of available sequences, novel tools are needed for comprehensive analysis of species-specific sequence characteristics for a wide variety of genomes. Multivariate analyses such as factor corresponding analysis and principal component analysis (PCA) have been used successfully to investigate variation in gene sequences (Grantham et al. 1980; Sharp et al. 1994; Kanaya et al. 1996). However, the clustering powers of the multivariate analyses are inadequate when massive amounts of sequence data from a wide variety of genomes are analyzed collectively (Kanaya et al. 2001). Self-Organizing Map (SOM), which was developed by Kohonen (Kohonen 1982, 1990, 1998; Kohonen et al. 1996) to study memory and recall/association mechanisms, can identify and associate similar types of information and localize such information in close vicinity on a two-dimensional map. SOM implements nonlinear projection of multi-dimensional data onto a two-dimensional array of weight vectors, and this preserves effectively the topology of the high-dimensional data space. SOM has been proven to be a powerful unsupervised algorithm and applied in various fields of science and technology (e.g., complex industrial processes, document and image databases, and financial applications) but rarely to analysis of genomic sequences. This is partly because the conventional SOM is dependent on the order of data input in the learning process. In the case of memory, the order of data input is important, but is not suitable for genome sequence analysis. In this and the previous studies, I modified the conventional SOM, on the basis of batch-learning SOM, for genome informatics to make the learning process and resulting map independent on the order of data input (Abe et al. 1999; Kanaya et al. 2001). In this thesis study, I used the modified SOM to analyze frequencies of short oligonucleotide (di- to pentanucleotide) in a wide variety of prokaryotic and eukaryotic genomes.

When only fragments of genomic sequences (e.g., 10-kb sequences) from mixed genomes of multiple organisms are available, it would appear to be impossible to identify how many and what types of genomes are present in the collected sequences. However, I found that the modified

SOM could classify the sequence fragments according to species without any information other than oligonucleotide frequencies. In Chapter II, I constructed SOMs of di-, tri-, and tetranucleotide frequencies in 1- and 10-kb sequences from prokaryotic and eukaryotic genomes for which complete sequences are available. SOM recognized in most 10-kb sequences species-specific characteristics (key combinations of oligonucleotide frequencies, i.e., genome signature), permitting species-specific classification of sequences without any information regarding species. In Chapter III, di-, tri-, and tetranucleotide frequencies in 90 prokaryotic and eukaryotic genomes were analyzed on a single SOM. The separation of eukaryotic and prokaryotic sequences on a single map could prove the resolving power of SOM, which should provide a powerful tool for phylogenetic classification of genomic sequences obtained from environmental uncultured microorganisms, including symbiotic/parasitic microorganisms associated with macroorganisms.

A majority of microorganisms, especially those living in extreme environments, are difficult to culture in the laboratory. Because conventional experimental approaches have been unsuccessful, these genomes have remained to uncharacterize, and there is the possibility that such genomes contain a wide range of novel genes that would be of scientific and/or industrial interest. Metagenomics, which is genomic analysis of uncultured microorganisms, has been proposed to study microorganism diversity in a wide variety of environments and to identify novel and industrially useful genes. In the metagenome analysis of uncultured microorganisms, genome DNAs are extracted directly from an environmental specimen that contains multiple organisms and the genomic fragments are then cloned and sequenced. This technology analyzing mixed genomes is expected to be an effective, powerful methodology to understand the diversity of microorganisms in an environment. However, with a simple collection of fragmental sequences, it is difficult to answer the following questions: what are the types and proportions of species in a specimen, to what phylogenetic groups do the species belong, and how novel are the genomes that are present in the specimen. In Chapter IV, I developed SOM as a methodology just suitable to this purpose and analyzed a vast amount of sequences derived from the pooled DNA samples obtained from the Sargasso Sea water (Venter et al, 2004)

In Chapter V, I analyzed tetra- and pentanucleotide frequencies in the human genome, and found that frequencies and distributions of oligonucleotide sequences involved in transcriptional regulation were often biased significantly from random occurrence. I could categorize occurrence patterns and frequencies of known signal sequences in the human genome. When known signal sequences from various species with sufficient experimental data are characterized and categorized systematically with SOMs, it should be possible to develop an *in silico* method to predict signal sequences, which is thought to be most useful for identification of signal sequences in genomes for which only sequence data are available. Because the number of such poorly characterized genomes becomes high, development of such an *in silico* method has become increasingly important. In Chapter V, I developed SOM as a methodology just suitable to this purpose.

In addition to protein-coding sequences (CDSs), the regions upstream of transcription start sites and the 5’ and 3’ untranslated regions (UTRs) of eukaryotic genes have attracted attention because of their crucial roles in transcriptional and post-transcriptional regulation of gene expression. Non-protein-coding transcripts (ncRNA) are also involved in the gene regulation or in guiding RNA modifications. In Chapter VI, by combining analyses on cDNA and genomic sequences of human and mouse, I developed bioinformatics based on SOM to characterize these functional regions and identify hidden sequence characteristics in these functional regions.

Of importance in genome analysis is prediction of the function of proteins that are identified through genome sequencing but lack significant sequence homology with function-known proteins, which are left as function-unknown proteins. In Chapter VII, I describe protein analysis as future prospect.

Chapter II: Development of Novel Bioinformatics for Unveiling Hidden Genome Signatures

II-1: INTRODUCTION

In addition to protein-coding information, genomic sequences contain a wealth of information of interest in many fields of biology from molecular evolution to genome engineering. G+C% has been used as a fundamental characteristic of individual genomes, but the G+C% is apparently too simple a parameter to differentiate a wide variety of genomes of known sequences. Oligonucleotide frequency can be used to distinguish genomes because oligonucleotide frequencies vary significantly among genomes; dinucleotide frequencies, for example, are shown to be genome signatures for both prokaryotes and eukaryotes (Nussinov 1984; Karlin 1998; Karlin et al. 1998; Gentles et al. 2001). Comprehensive analyses of oligonucleotide frequencies in a wide variety of genomes are thought to provide fundamental knowledge of individual genomes, namely, key combinations of oligonucleotides responsible for the biological properties of the different genomes and genome portions. A few researchers including me have used SOMs to characterize codon usage patterns of a wide variety of prokaryotes (Abe et al. 1999; Kanaya et al. 1998; Kanaya et al. 2001; Wang et al. 2001). I introduced a new feature to the SOM for studies of genomic sequences that makes the learning process independent of the order of data input, first by characterizing codon usage in 60,000 genes from 29 prokaryotic species. SOM was particularly useful not only in searching for horizontally transferred genes but also in predicting the donor genomes of the transferred genes (Kanaya et al. 2001).

In this chapter, I constructed SOMs with di-, tri-, and tetranucleotide frequencies for a total of 17,000 10-kb and 170,000 1-kb genomic sequences of 65 prokaryote genomes and for a total of 46,000 10-kb and 460,000 1-kb segments of 6 eukaryote genomes. The resulting SOMs for the 16-, 64-, and 256-dimensional spaces (for di-, tri-, and tetranucleotide frequencies, respectively) revealed clear separations between inter- and intraspecies sequences that generally corresponded to biological categories. Comparative analysis of interspecies differences in oligonucleotide frequencies could provide insight into hidden signatures in genome sequences established during evolution.

II-2: METHODS

Genome sequences

The DNA sequences of 65 prokaryotic genomes, itemized in the Fig. II.1 legend, were obtained from the DDBJ GIB web site (http://www.ddbj.nig.ac.jp/), and those of the six eukaryotes itemized in the Fig. II.4 legend were obtained from the DDBJ/EMBL/GenBank web site. For the calculation of oligonucleotide frequency for a window, when the number of undetermined nucleotides (Ns) exceeded 10% of the window size, the respective sequences were omitted from the analysis. When the number of Ns was less than 10% of the window size, oligonucleotide frequencies were normalized to the length of the sequence without Ns and included in the analysis.

SOM adapted for genome informatics

The neural network algorithms can be supervised or unsupervised. The supervised training is accomplished by presenting a sequence of training vectors, each with an associated target output vector. An essential requirement of the supervised learning is the availability of an external teacher such as class. We may think of the teacher as having knowledge of the environment that is represented by a set of input-output examples. Knowledge of the environment available to the teacher is transferred to the neural network learning algorithms. In unsupervised there is no external teacher to oversee the learning process. The learning normally is driven by a similarity measure without specifying target vectors. The self-organizing net modifies the weights to that the most similar vectors are assigned to the same output unit, which is represented by an example vector.

SOM is an unsupervised neural network algorithm that implements a characteristic nonlinear projection from the high-dimensional space of input data onto a two-dimensional array of weight vectors (Kohonen 1982, 1990, 1998; Kohonen et al. 1996). In the conventional SOM developed by Kohonen, the map is a two-layered network that can organize a topological map of cluster units from a random starting point. The network combines an input layer with a competitive layer of processing units. During the self-organization process, the cluster unit, whose weight vector matches the input pattern most closely (typically based on minimum Euclidean distance), is chosen as the winner. The winning unit and its neighboring units update their weights. After training is complete, pattern relationships and group are observed from the competitive layer. This yields the graphical organization of pattern relationships. These maps result from an information compression that retains only the most relevant common features of the set of input signals. This preserves effectively the topology of the high-dimensional data space. It is thought of as a flexible net that is spread into the multi-dimensional "data cloud". Because the net is a two-dimensional array, it can be visualized easily. The weight vectors ($\mathbf{w}_{ij}$) are arranged in the two-dimensional lattice denoted by i (=0, 1,..., I-1) and j (= 0, 1, ..., J-1).

The learning process of the SOM developed by me was designed to be independent of the order of input of vectors on the basis of batch-learning SOM (BL-SOM) as I previously reported (Abe et al. 1999; Kanaya et al. 2001). In the conventional SOM (Kohonen 1982, 1990, 1998; Kohonen et al. 1996), the initial weight vectors $\mathbf{w}_{ij}$ are set by random values as noted above, but in my method the vectors are initialized by PCA (Step 1). For mapping multi-dimensional space data onto a plane, PCA rotates the vector space with the eigenvectors (the principal components) of the covariance matrix as a new basis. The principal components are orthogonal, and the plane spanned by the two first components, PC1 and PC2, was usually used for a linear data projection. Weights in the first dimension (the number of lattice points in the first dimension is denoted by I) were arranged into 150 nodes for 10-kb sequences or 350 nodes for 1-kb sequences corresponding to a width of five times the standard deviation ($5\sigma_1$) of the first principal component; and the second dimension (J) was defined by the nearest integer greater than (σ_2/σ_1) x I.

The weight vector on the *ij*th lattice ($\mathbf{w}_{ij}$) was represented as follows:

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

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 $\mathbf{w}_{i'j'}$) satisfied in minimal distance. After associating all input vectors with weight vectors, updating was done according to Step 3.

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

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

where components of set S_{ij} are input vectors associated with $\mathbf{w}_{i'j'}$ satisfying $i - \beta(r) \le i' \le i + \beta(r)$ and $j - \beta(r) \le j' \le j + \beta(r)$. Here, $\mathbf{w}_{ij}^{(new)}$ is an updated vector. The two parameters $\alpha(r)$ and $\beta(r)$ are learning coefficients for the *r*th cycle, and N_{ij} is the number of components of S_{ij}. $\alpha(r)$ and $\beta(r)$ are set by

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

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

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

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

where N is the total number of sequences analyzed.

II-3: RESULTS

Species-specific oligonucleotide frequencies in prokaryotic genomes

SOMs were constructed with di-, tri-, and tetranucleotide frequencies for approximately 17,000 genomic 10-kb sequences derived from the 65 prokaryotic genomes whose complete sequences are known. As the first step to obtain the initial weight vectors, frequencies for the 17,000 non-overlapping segments were analyzed by principal component analysis (PCA). This is based on the knowledge that multivariate analyses including PCA successfully classified gene sequences into groups corresponding to known biological categories when the numbers of sequences and species were much smaller than that analyzed here (Grantham et al. 1980; Medigue et al. 1991; Sharp and Matassi 1994; Andersson and Sharp 1996). After 40 learning cycles, oligonucleotide frequencies of genomic sequences were effectively reflected as the weight vectors in SOMs (Fig. II.1A-C). Comparison of the sequence classification into lattice nodes of the final weight vectors with that of the initial vectors set by the first and second principal components of PCA (Fig. II.1G) showed that sequences of a single species were much more tightly clustered in the final vectors. Nodes that contain sequences from a single species are indicated in color, and those including sequences of more than one species are indicated in black. In the SOMs, sequences of most species were separated into species-specific non-overlapping zones (Fig. II.1A-C). In contrast, the resolving power of the conventional PCA method that could be estimated with the initial vectors (Fig. II.1G) was poor. The contiguous non-intermingling zones that contained sequences of a single species were very limited when compared with the contiguous non-overlapping zones obtained with SOMs.

Analysis of the weight vectors for individual nodes showed that strongly biased vectors were localized to the edge of the map, whereas those with weakly biased vectors were in the center. The G+C% for each weight vector in di-, tri-, and tetranucleotide SOMs (abbreviated as Di-, Tri-, and Tetra-SOMs) was reflected mainly in the horizontal axis and increased from left to right; sequences of AT- and GC-rich prokaryotes were distributed on the left- and right-hand sides of the SOMs, respectively (Fig. II.1D-F). Importantly, sequences with the same G+C% are separated by a complex combination of oligonucleotide frequencies resulting in species-specific separations. In other words, most of the 10-kb segments in each genome have a combination of oligonucleotides that reflect the respective genome like a signature, and SOMs can reveal the signature as representative weight vectors. The 170,000 non-overlapping 1-kb sequences were also analyzed (Fig. II.2). Species-specific separations were again observed, though the resolution was somewhat reduced. This shows that species-specific signatures are detectable even in a major population of the 1-kb sequences.

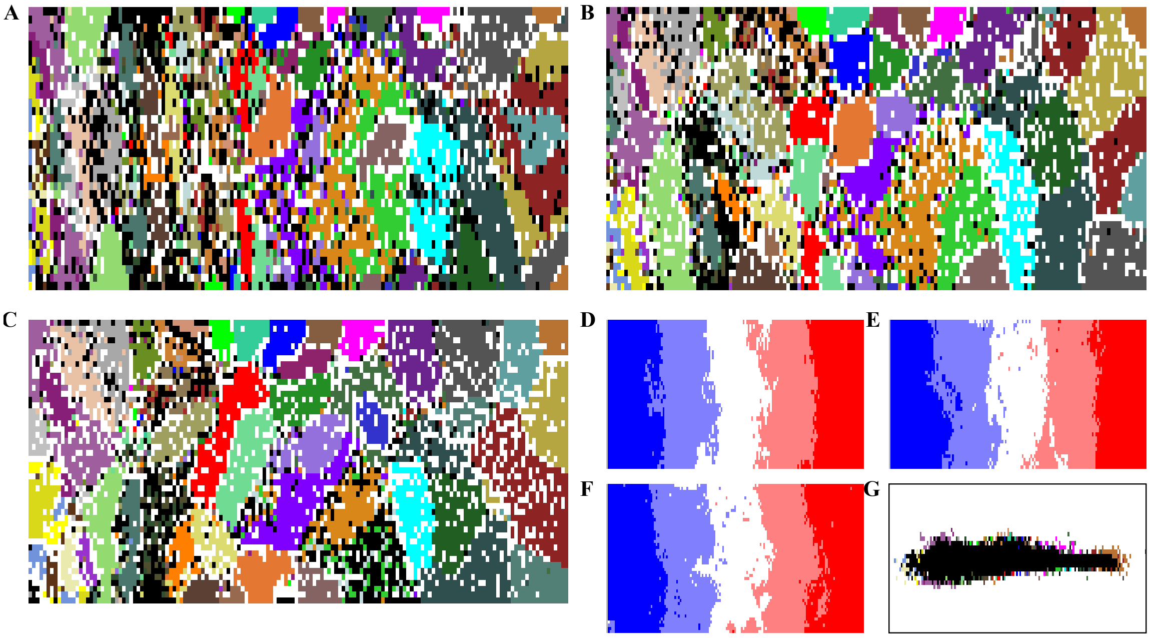


Figure II.1 SOMs for 10-kb sequences of 65 prokaryotic genomes. (A, B, and C) Di-, Tri-, and Tetra-SOMs, respectively. Nodes that contain sequences from more than one species are indicated in black, and those that contain sequences from a single species are indicated in color as follows: *Aquifex aeolicus* (🟩), *Archaeoglobus fulgidus* (🟦), *Aeropyrum pernix* (🟪), *Agrobacterium tumefaciens* (🟦), *Borrelia burgdorferi* (🟨), *Bacillus halodurans* (🟥), *Bacillus subtilis* (🟩), *Buchnera sp.* (🟫), *Clostridium acetobutylicum* (🟪), *Caulobacter crescentus* (🟨), *Campylobacter jejuni* (🟨), *Chlamydia muridarum* (🟥), *Chlamydophila pneumoniae* (🟫), *Chlamydia trachomatis* (🟫), *Deinococcus radiodurans* (🟦), *Escherichia coli* (🟧), *Halobacterium* sp. (🟧), *Haemophilus influenzae* (🟧), *Helicobacter pylori* (🟫), *Listeria innocua* (🟩), *Lactococcus lactis* (🟦), *Listeria monocytogenes* (🟩), *Mycoplasma genitalium* (🟪), *Methanococcus jannaschii* (🟪), *Mycobacterium leprae* (🟪), *Mesorhizobium loti* (⬛), *Mycoplasma pneumoniae* (🟫), *Mycoplasma pulmonis* (🟦), *Methanothermobacter thermautotrophicus* (🟫), *Mycobacterium tuberculosis* (⬛), *Neisseria meningitidis* (🟫), *Pyrococcus abyssi* (🟧), *Pseudomonas aeruginosa* (🟥), *Porphyromonas gingivalis* (🟩), *Pyrococcus horikoshii* (🟧), *Pasteurella multocida* (🟨), *Rickettsia conorii* (⬜), *Rickettsia prowazekii* (🟦), *Staphylococcus aureus* (🟩), *Sinorhizobium meliloti* (🟩), *Streptococcus pneumoniae* (🟨), *Streptococcus pyogenes* (⬜), *Sulfolobus solfataricus* (⬜),

Sulfolobus tokodaii (🟫), *Salmonella typhimurium* (🟩), *Synechocystis* sp. (🟧), *Thermoplasma acidophilum* (🟪), *Thermotoga maritima* (🟩), *Treponema pallidum* (🟦), *Thermoplasma volcanium* (🟩), *Ureaplasma urealyticum* (🟨), *Vibrio cholerae* (🟪), *Xylella fastidiosa* (🟩), and *Yersinia pestis* (🟩). (D, E, and F) G+C% for each weight vector in Di-, Tri-, and Tetra-SOMs, respectively. G+C% for each node vector was divided into five categories containing an equal number of nodes. The highest, second-highest, middle, second-lowest, and lowest G+C% categories are shown in dark red, light red, white, light blue, and dark blue, respectively. (G) Classification by the initial weight vectors set by PCA for the Di-SOM. Nodes are colored as described in A-C.

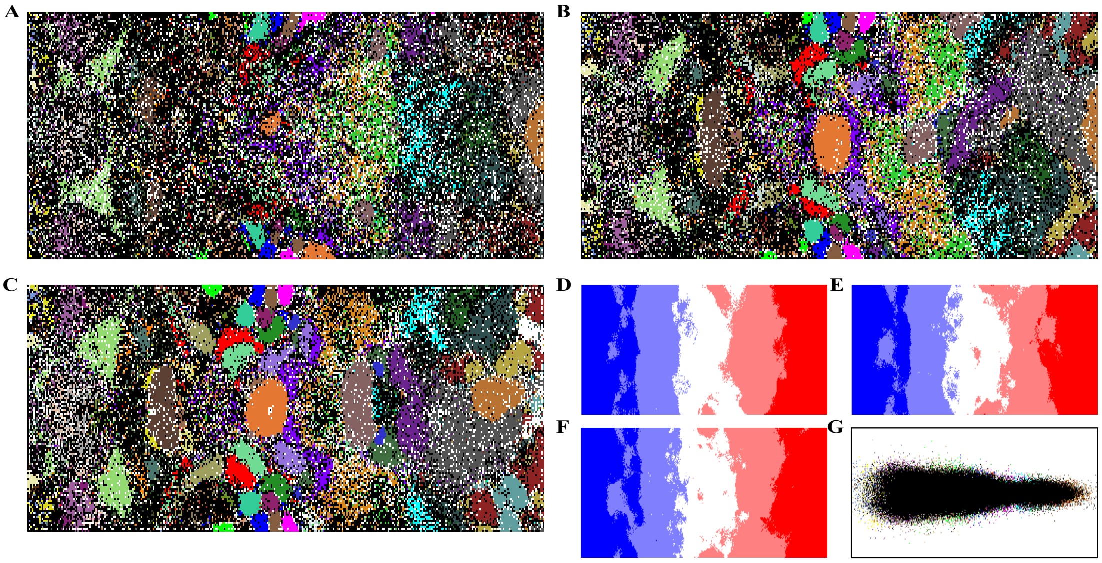


Figure II.2 SOMs for 1-kb sequences of 65 prokaryotic genomes. (A, B, and C) Di-, Tri-, and Tetra-SOMs, respectively. Nodes are colored as described in Fig. II.1A-C. (D, E, and F) G+C% for each weight vector is shown as described in Fig. II.1D-F. (G) Classification by the initial weight vectors for the Di-SOM.

Intraspecies Separation for Prokaryotic Genomes

Detailed inspection of SOMs showed that each species was often split into two major zones that were composed of roughly equal numbers of data points. To illustrate this split more clearly, data points for each of seven representative species in the 10-kb Tri-SOM (Fig. II.1B) and in the 1-kb Tri- and Tetra-SOMs (Fig. II.2B and C) were plotted with a single color (Fig. II.3A-C). In the 1-kb SOMs, intraspecies separations were observed for all seven species, but in the 10-kb SOM, the separation was not observed for three species. In order to investigate the biological significance of the two zones, correlation with transcription polarities of protein coding sequences (CDSs) in the respective genomic segments was studied focusing on CDSs data of *E. coli* and *B. subtilis* compiled in the DDBJ Genome Information Broker (http://gib.genes.nig.ac.jp/). Sequences belonging to each major zone in the 1-kb Tetra-SOMs (Fig. II.3C) are illustrated separately as red and blue bands below the diagrams that show transcription polarities of CDSs in the 200-kb segment (Fig. II.3D and E). A red or blue band in each species showed clustering of contiguous 1-kb sequences belonging to one of the two zones in the 1-kb SOM. Each band coincided with the clustering of CDSs with one polarity, and borders between red and blue bands were usually located at positions corresponding to the switch positions for transcription polarity. In the cases of the three species for which the intraspecies separation was lost in the 10-kb SOM, switching of transcriptional polarities occurs within a 10-kb segment at higher probabilities than observed for the other four species (data not shown). These findings indicate that codon usage patterns contribute to the intraspecies separations and probably also to the interspecies separations.

Genome segments introduced through horizontal transfer from distantly related organisms are known to retain the sequence characteristics of the donor genome and can be distinguished from those of the acceptor genome (Lawrence and Ochman 1997, 1998). For example, genes transferred from other genomes often have codon usage patterns distinct from those of their intrinsic genomes (Grantham et al. 1980; Ikemura 1985; Medigue et al. 1991). I showed previously that SOMs are useful for identifying horizontally transferred genes and, importantly, for predicting the donor genomes of the transferred genes (Kanaya et al. 2001). There are characteristic data points in Fig. II.3A that are located away from the major zones of individual species. The sequences that have oligonucleotide frequencies clearly distinct from those of major zones should correspond, at least in part, to genome portions that have been transferred horizontally from other genomes. To test this possibility, I examined 10-kb sequences from *E. coli* that were located outside the territories of both *E. coli* and a closely related bacterium *S. typhimurium.* When the sequences in the *S. typhimurium* territory were excluded, the next highest number of *E. coli* sequences was found in the *Y. pestis* territory. I then focused the five sequences in the *Y. pestis* territory commonly found in the Di-, Tri-, and Tetra-SOMs. Within these sequences, there were 37 known genes, 23 of which had significant homology with *Y. pestis* genes. For example, at the amino acid level, 6 out of 23 proteins had identity levels greater than 60%, and the highest was 80%, which was significantly higher than the 40% calculated for the average identity for the orthogonal pairs of *E. coli* and *Y. pestis* proteins (Deng et al. 2002). Furthermore, three genes were homologous with the phage-encoded genes, and one was homologous with a transposon gene. These findings support the prediction that these genes may have been transferred horizontally into the *E. coli* genome from other organisms.

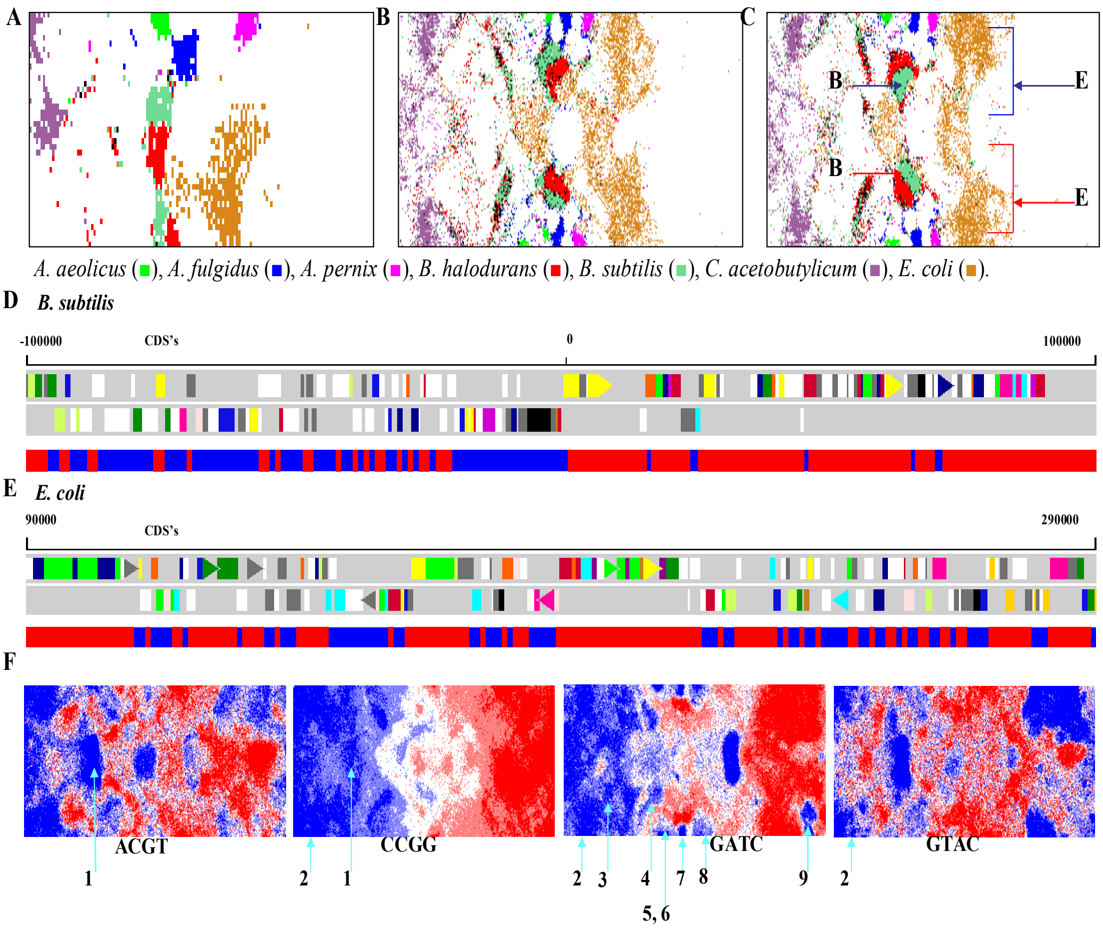


Figure II.3 Intraspecies separations and tetranucleotide distributions in SOMs for prokaryotic genomes. (A, B, and C) The 10-kb Tri-, 1-kb Tri-, and 1-kb Tetra-SOMs. Seven representative species with two major zones are indicated in color as follows: *A. aeolicus* (🟩), *A. fulgidus* (🟦), *A. pernix* (🟪), *B. halodurans* (🟥), *B. subtilis* (🟩), *C. acetobutylicum* (🟪), and *E. coli* (🟧). In C, the two major zones of *B. subtilis* or *E. coli* are noted with red or blue arrow with the letter B or E, respectively. (D) Transcriptional polarity and SOM separation for *B. subtilis* sequences. Two transcriptional polarities of CDSs in the 200-kb *B. subtilis* segment with a replication origin are presented separately in the top two panels; this was obtained from the DDBJ web site (http://gib.genes.nig.ac.jp/). Below the two panels, contiguous 1-kb sequences within the 200-kb segment and belonging to the two major zones marked with red and blue arrows in C are shown separately with the red and blue bands, respectively. (E) Transcriptional polarity and SOM separation for *E. coli* sequences. A 200-kb *E. coli*

segment lacking a replication origin was analyzed as described in D. (F) Tetranucleotide distribution in the 1-kb Tetra-SOM for prokaryotes. Levels of each tetranucleotide for all node vectors in the Tetra-SOM of Fig. II.2C were divided into five categories containing an equal number of nodes, and the highest, second-highest, middle, second-lowest, and lowest categories are shown with different levels of red and blue as described in Fig. II.1D-F. Zones for prokaryotes that have genes encoding a restriction enzyme that recognizes the respective tetranucleotide are noted by light blue lines with the following numbers to show the species name: 1, *H. pylori*; 2, *M. jannaschii*; 3, *S. aureus*; 4, *S. pneumoniae*; 5, *P. abyssi*; 6, *P. horikoshii*; 7, *A. fulgidus*; 8, *A. pernix*; and 9, *D. radiodurans*. For other palindromic tetranucleotides, see Supplementary Data II.2. Of 17 restriction enzymes from 11 prokaryotes, the respective tetranucleotides were underrepresented in 15 instances.

SOMs for eukaryotic genomes

The protein-coding portion of each eukaryotic genome, especially in higher eukaryotes, is reduced appreciably in comparison to that of prokaryotic genomes. Therefore, genome signatures derived from species-specific codon usage should be less prevalent than those observed for prokaryotes. I examined di-, tri-, and tetranucleotide frequencies for 6 eukaryote genomes (a total of 460 Mb) including 4 genomes (*S. cerevisiae, C. elegans, A. thaliana, D. melanogaster*) that have been sequenced completely, as well as *P. falciparum* chromosomes 2 and 3, and human chromosomes 20, 21, and 22. The 46,000 nonoverlapping 10-kb segments from these 6 eukaryote genomes were analyzed (Fig. II.4A-C). Most of the 10-kb segments were separated according to species. For example, more than 95% of the human sequences were located in the human territories, which are marked in red in Fig. II.4A-C. This shows that SOM separations, that were obtained without any species information, closely fit separations among species, and thus the unsupervised algorithm can recognize, in most 10-kb sequences, the species-specific characteristic (a key combination of oligonucleotide frequencies) that is the representative signature of each genome.

The G+C% calculated for each weight vector in Di-, Tri-, and Tetra-SOMs are shown in Fig. II.4D-F. It is apparent that sequences with the same G+C% are separated by a complex combination of oligonucleotide frequencies resulting in species-specific separations. Underlying representation in SOMs enables us to retrieve characteristic sequence patterns for individual genomes and genome regions. The frequency of each dinucleotide in the weight vector for each node in the Di-SOM is illustrated in red and blue (Fig. II.5). Complementary pairs of dinucleotides (e.g. AA versus TT) had similar distribution patterns. This indicates that when the sequence complimentary to the sequence registered by the International DNA Databank is used for a certain genome, general patterns may not change appreciably. Lines in all panels in Fig. II.5 represent the species borders observed in the Di-SOM. Species borders coincide with regions of transition between the red and blue levels for several dinucleotides, which correspond to the diagnostic dinucleotides for the species border formation. For example, the CG dinucleotide deficiency (dark blue zones in the CG panel) is a factor responsible for separation of human sequences (red in Fig. II.4A) from *Drosophila* (pink) and

Arabidopsis (green) sequences. Levels of CG, GC, AG, and GA contributed appreciably to separation of *Drosophila* sequences from others. It should be stressed that the SOM utilizes complex combinations of many more dinucleotides for the sequence separations in an area-dependent manner. This is because SOMs implement nonlinear projection from the multi-dimensional space of input data onto a two-dimensional array of weight vectors (Kohonen 1982, 1990, 1996).

In similar fashion, trinucleotide levels for each representative vector in the Tri-SOM were analyzed. Again, species borders often coincided with regions of sharp transition between the red and blue levels for various diagnostic trinucleotides (Fig. II.6). For humans, high levels of AGG, CAG, CCC, CCT, CTG, and GGG as well as low levels of ACG, CGA, CGT, and TCG were observed, and for *Drosophila*, high levels of GCA and TGC and low levels of AGA, TCA, TCT, and TGA were observed. The underrepresentation of CNG in a major portion of the *Arabidopsis* territory is thought to be related to cytosine methylation in CNG trinucleotides (Lindroth et al. 2001). The SOM utilizes complex combinations of many trinucleotides for species separation in an area-dependent manner; the 64 panels for all trinucleotides are presented as Supplementary Data II.1.

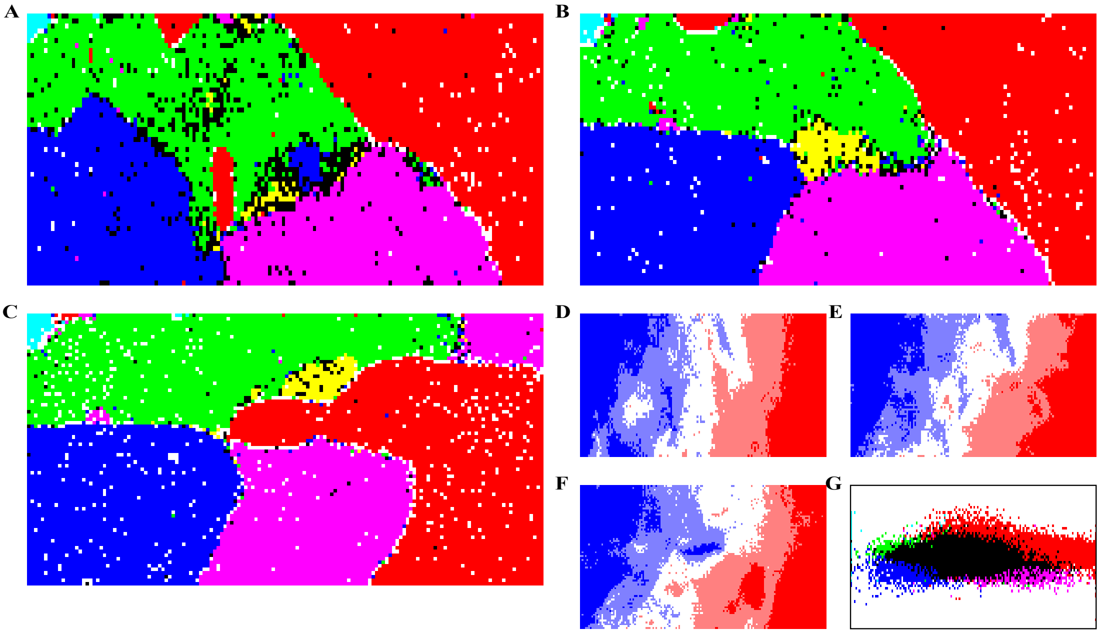


Figure II.4 SOM for 10-kb sequences of six eukaryotes. (A, B, and C) Di-, Tri-, and Tetra-SOMs, respectively. Nodes that contain sequences from more than one species are indicated in black, and those that contain sequences from a single species are indicated in color as follows: *A. thaliana* (🟩), *C. elegans* (🟦), *D. melanogaster* (🟪), *P. falciparum* (■), *S. cerevisiae* (🟨), and human chromosomes 20, 21, and 22 (🟥). (D, E, and F) G+C% for each weight vector in Di-, Tri-, and Tetra-SOMs, respectively. G+C% for each node vector is shown as described in Fig. II.1D-F. (G) Classification by the initial weight vectors for the Di-SOM.

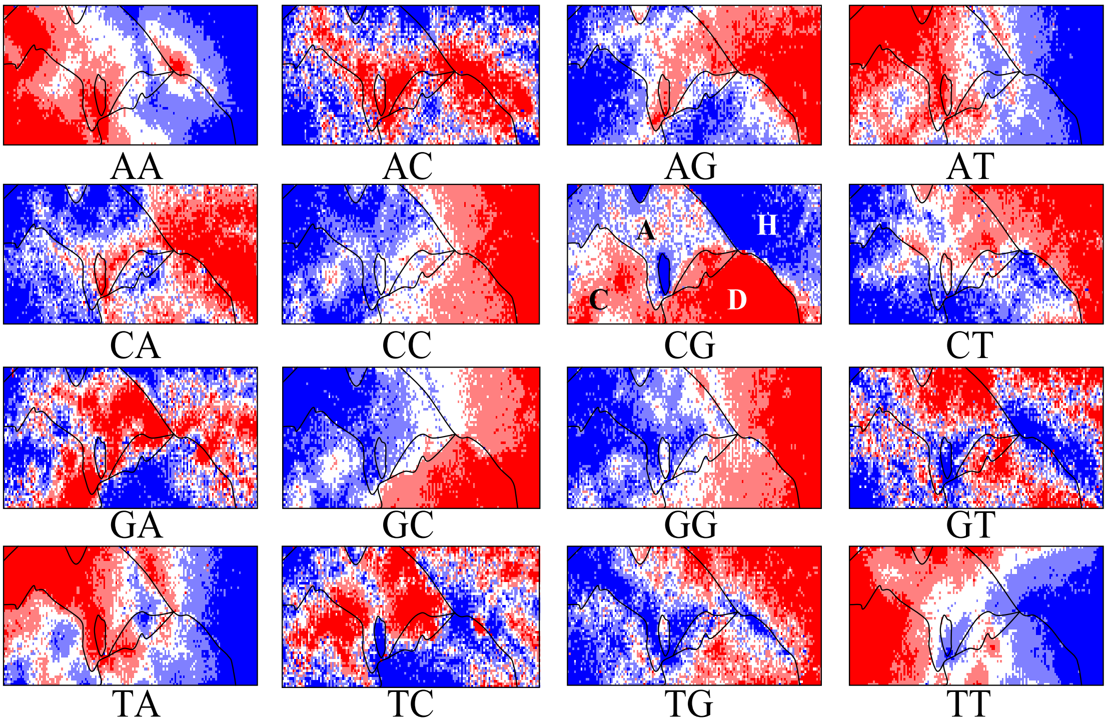


Figure II.5 Dinucleotide distribution in 10-kb Di-SOM for six eukaryotes. Levels of each dinucleotide for all node vectors in the Di-SOM of Fig. II.4A were divided into five categories containing an equal number of nodes and the categories are shown as described in Fig. II.3F. Species borders in the Di-SOM (Fig. II.4A) are marked by lines. Major zones for four species were noted in the CG panel as follows: *A. thaliana* (A), *C. elegnas* (C), *D. melanogaster* (D), and human (H).

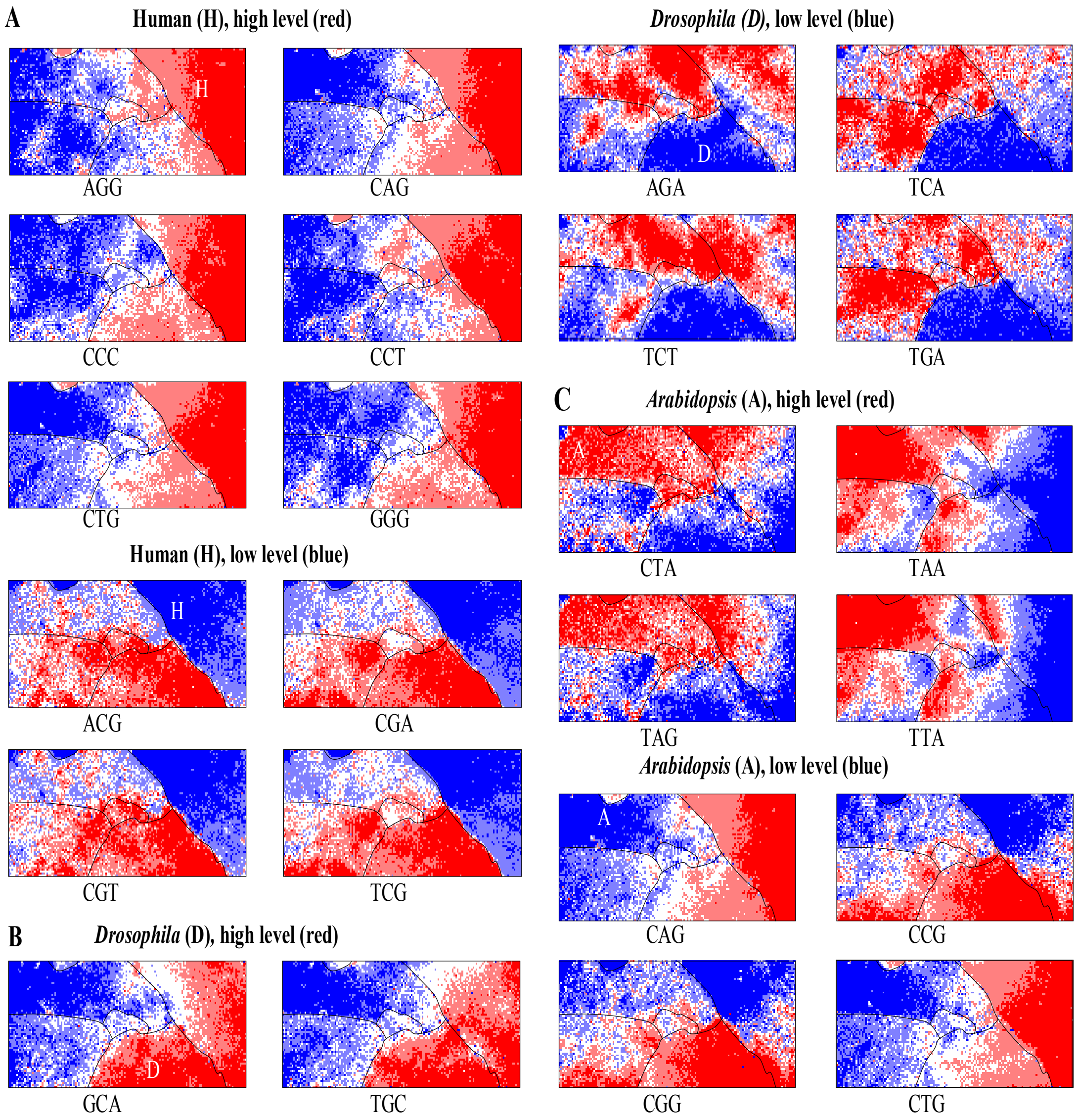


Figure II.6 Trinucleotide distribution in 10-kb Tri-SOM for six eukaryotes. Levels of each trinucleotide for all node vectors in the Tri-SOM of Fig. II.4B were divided into five categories and shown as described in Fig. II.3F. Species borders are shown as described in Fig. II.5. (A) Human. Six diagnostic trinucleotides with high frequencies and four with low frequencies are shown. (B) *D. melanogaster.* Two diagnostic trinucleotides with high frequencies and four with low frequencies are shown. (C) *A. thaliana*. Four diagnostic trinucleotides with high frequencies and four with low frequencies (CNG) are shown.

Intraspecies separation observed for eukaryote genomes

Human sequences had two and one satellite zone in the Di- and Tri-SOMs, respectively (red minor zones in Fig. II.4A and B). Genomes of warm-blooded vertebrates such as humans are known to be composed of long-range segmental G+C% distributions designated as “isochores” by Bernardi (1985) (Ikemura 1985; Ikemura and Aota 1988; Bernardi 1989; Gautier 2000; Eyre-Walker and Hurst 2001). Correlation of the segmental G+C% distributions with SOM separations was observed. For example, approximately 500 10-kb sequences belonging to the satellite red zone located at the top at the left side in the Di- and Tri-SOMs were practically common between the two SOMs, and the G+C% was between 30 to 33%, which corresponds to very AT-rich sequences in the human genome. Four-fifths of the sequences were derived from very AT-rich “gene-desert regions” on chromosome 21 (Hattori et al. 2000) that correspond to L1 isochores (Saccone et al. 1999) and replicate very late during S phase (Watanabe et al. 2002). The SOM can unveil specific genome portions with distinct characteristics as intraspecies separations. *Drosophila* sequences were split into two major zones in the Tetra-SOM (pink in Fig. II.4C), and this split was associated with G+C%.

SOMs with 1-kb eukaryote sequences

To determine the usefulness of SOMs for analysis of genomes with respect to functional aspects, I investigated the effects of shortening the sequence length on SOM separations because 10-kb segments appear to be too long for such studies. SOMs were constructed with the dinucleotide frequencies for a total of 460,000 nonoverlapping 1-kb sequences from 6 eukaryotes. Clear separations of species were observed, but territories of individual species were split into several zones (Fig. II.7A). Mirror symmetric distributions were apparent for sequences of each genome. The G+C% in the SOM was reflected mainly on the horizontal axis, and the complementarity of oligonucleotides was reflected on the vertical axis. I examined the possible factors responsible for the separations in the 1-kb SOM by analyzing dinucleotide levels for each node. The best example was the CG dinucleotide level shown in Fig. II.7B. All *Drosophila* zones (pink in Fig. II.7A) corresponded primarily to the CG-rich zones (red in Fig. II.7B), and all human zones (red in Fig. II.7A) corresponded primarily to the CG-poor zones (blue in Fig. II.7B) except for one clear characteristic zone that is marked by an arrow. This CG-rich human zone is thought to have CpG-island sequences that are often present in the regulatory regions for transcription. In order to examine this possibility, I mapped human UTR sequences compiled in UTRdb (Graziano et al. 1998, 2002), which is a specialized database of 5' and 3' UTRs of eukaryotic mRNAs cleaned from redundancy. To study the characteristics irrespective of length differences in the UTR sequences, oligonucleotide frequencies in UTR sequence were calculated and were mapped to the node of the SOM with the shortest distance in the multidimensional frequency space (Fig. II.7C). The 5' and 3' UTR sequences were mapped in different zones of human territories, and the 5' UTR sequences were enriched characteristically in the CG-rich human zone marked by an arrow. This demonstrates the usefulness of SOM for discovery of local, functional sequence characteristics, and the usefulness is explained in detail in Chapter VI.

The average number of sequences per node in the 1-kb SOM was 11. The actual number of sequences classified into each node that is composed of sequences from a single species is shown by the height of the colored rod (Fig. II.7D). There were many apparently high rods. Systematic analyses of these characteristic rods might provide unique, biologically significant information. It should be noted that many of the 1-kb segments are free of species-specific ubiquitous repetitive elements, such as *Alu* or L1 elements in the human genome. The sequences with or without repetitive elements were found to be colocalized in the major zones of individual species. Detailed inspection showed that 10-kb human sequences with or without *Alu* or L1 elements were also colocalized in the human major zones on the 10-kb SOMs. Therefore, the major factors responsible for the species-specific separations of eukaryote sequences do not appear to be ubiquitous repetitive elements. Factors responsible for the separations could be characteristics that are more extensively embedded than repetitive elements.

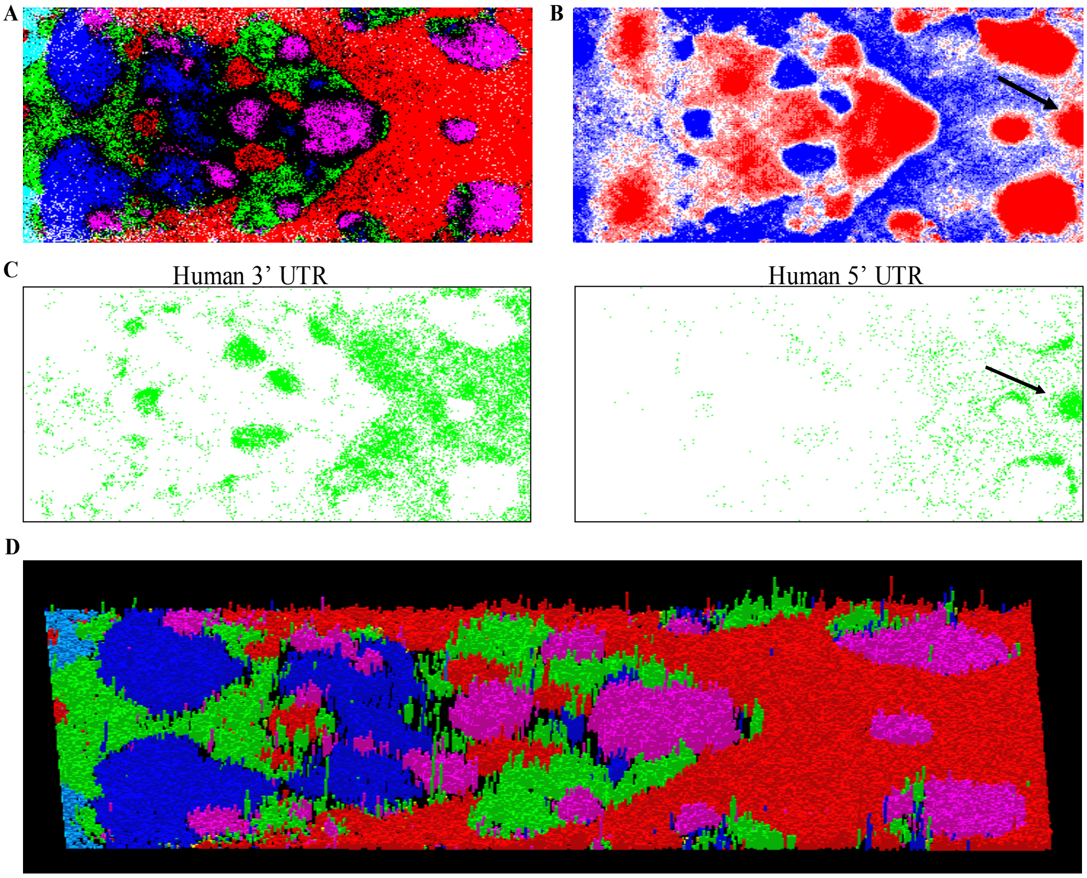


Figure II.7 Di-SOM for 1-kb sequences of six eukaryotes. (A) Di-SOM. Nodes are colored as described in Fig. II.4A. (B) CG dinucleotide levels for all weight vectors were calculated and shown as described in Fig. II.5. The CG-rich zone in the human territories is noted with an arrow. (C) Mapping of sequences of human 5' and 3' UTRs. I used human UTRs compiled in UTRdb (http://bighost.area.ba.cnr.it/BIG/UTRHome/), which is a specialized database of eukaryotic 5' and 3' UTRs that has been cleaned of redundant sequences. To get statistically meaningful results, sequences shorter than 100 nucleotides were excluded from this analysis.

(D) Three-dimensional presentation of the Di-SOM. Number of sequences classified into each node that has sequences from a single species is presented with the height of the colored rod.

II-4: DISCUSSION

Biological implications of SOM separations and genome signatures

To investigate the biological significance of diagnostic oligonucleotides for SOM separations, I examined the correlation of levels of palindromic tetranucleotides with respective restriction enzyme systems by referring to the restriction enzyme database (REBASE; http://vent.neb.com/~vincze/genomes/), as explained in detail in Chapter III. Restriction site tetranucleotides were underrepresented in 10 of the 11 prokaryotes that have genes encoding 4-base cutter enzymes (blue in Fig. II.3F). This finding is consistent with that of Karlin et al. (1997) on compositional biases of prokaryotic genomes, again indicating that SOMs can effectively classify sequences according to biological categories. The 256 panels for all tetranucleotides in the prokaryote and eukaryote SOMs are presented as Supplementary Data II.2 and II.3, respectively. I then considered the biological significance of diagnostic tetranucleotides in eukaryotes. One possible explanation is a contributory effect of levels of the di- and trinucleotide components of tetranucleotides. For example, tetranucleotides containing the CG dinucleotide were clearly underrepresented in the human territory (Supplementary Data II.3). Transition zones of various other tetranucleotides (e.g., CTCA and CTGA) were also sharp and exactly coincided with species-borders. Such sharp transitions and exact coincidences were not typical for the dinucleotide components (e.g., CT, TC, and CA in Fig. II.5). As found in the restriction enzyme analysis, some tetranucleotides may have biological significance. Species-specific characteristics for DNA synthesis and repair enzymes, as well as sequence preferences of ubiquitous DNA-binding proteins, may be responsible for differences in oligonucleotide distribution between species. In the cases of signal and motif sequences, such as transcription factor binding sites, they may be biased from the random occurrence statistically calculated from the genome base composition, as explained in Chapter V. This prediction is consistent with the finding that GAGA/TCTC, which is a transcription signal in *Drosophila* (Soeller et al. 1993), was underrepresented in the *Drosophila* genome (Supplementary Data II.3).

II-5: SUMMARY

From analysis of 1-kb and 10-kb genomic sequences derived from 65 prokaryotes (a total of 170 Mb) and from 6 eukaryotes (460 Mb), clear species-specific separations of major portions of the sequences were obtained with the Di-, Tri-, and Tetra-SOMs. The unsupervised algorithm could recognize, in most 10-kb sequences, the species-specific characteristics (key combinations of oligonucleotide frequencies) that are signature features of each genome. I could classify DNA sequences within one and between many species into subgroups that corresponded generally to biological categories.

Chapter III: Self-Organizing Map Reveals Sequence Characteristics of 90 Prokaryotic and Eukaryotic Genomes on a Single Map

III-1: INTRODUCTION

Systematic analyses of oligonucleotide frequencies in a wide variety of genomic sequences can reveal fundamental genome characteristics, namely, key combinations of oligonucleotides responsible for biological properties of different genomes and genomic regions, as described in Chapter II. I 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. Furthermore, the initial weight vectors were defined using principal component analysis (PCA) instead of random values. Therefore, the modified SOM was independent of not only the data input order but also the initial condition. In the resulting SOMs, the sequences were clustered according to species without any information regarding the species, and increasing the oligonucleotides in length from di- to tetranucleotides increased the clustering power.

In this chapter, di-, tri- and tetranucleotide frequencies in 90 prokaryotic and eukaryotic genomes were analyzed on a single SOM. The separation of eukaryotic and prokaryotic sequences on a single map can clarify the resolving power of SOM, which may provide a powerful tool for phylogenetic classification of genomic sequences obtained from environmental uncultured microorganisms, including symbiotic/parasitic microorganisms associated with macroorganisms (refer to Chapter IV, for details).

III-2: METHODS

Genome sequences

Sequences were obtained from DDBJ/EMBL/GenBank. For the calculation of oligonucleotide frequency for a window, when the number of undetermined nucleotides (Ns) exceeded 10% of the window size, the respective sequences were omitted from the analysis. When the number of Ns was less than 10% of the window size, oligonucleotide frequencies were normalized to the length without Ns and included in the analysis, as described in Chapter II.

SOMs adapted for genome informatics

SOMs were constructed as described in Chapter II. The initial weight vectors were set based on the widest scale of the sequence distribution in the oligonucleotide frequency space with PCA. Weights in the first dimension (I) were arranged into lattice nodes corresponding to a width of five times the standard deviation ($5\sigma_1$) of the first principal component; I was set as 250 in this study. The second dimension (J) was defined by the nearest integer greater than (σ_2/σ_1) x I.

III-3: RESULTS

SOMs for oligonucleotide frequencies in 90 genomes

I analyzed short oligonucleotide frequencies in the 90 genomes for which complete sequences are available: 0.2 Gb for 81 prokaryotes and 1.2 Gb for nine eukaryotes (see Fig. III.1 legend). For the human genome, sequences from four chromosomes that were almost completely sequenced, were analyzed. SOMs were constructed with di-, tri-, and tetranucleotide frequencies for 140,000 nonoverlapping 10-kb sequences and overlapping 100-kb sequences with a sliding step size of 10 kb derived from a total of 1.4 Gb from the 90 genomes. To set the initial weight vectors, frequencies for the 140,000 sequences were analyzed by PCA. After 80 learning cycles, the sequences of many species were separated into species-specific territories (Fig. III.1A-C). SOM separations obtained without any species information closely fit the sequence classification according to species. Nodes 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. Comparison of sequence classification with the initial vectors set by PCA (Fig. III.1E) with those for the final vectors (Fig. III.1A) revealed that sequences from a single species were far more tightly clustered with the final vectors. 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. III.1A-C) of the di, tri-, and tetranucleotide SOMs (Di-, 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 nodes into which no genomic sequences were classified, showing that vectors of species-specific nodes located even near the species border were clearly distinct from each other and the species borders primarily could be drawn automatically on the basis of the contiguous white nodes. The unsupervised algorithm recognized the species-specific characteristic (a key combination of oligonucleotide frequencies) that is the representative signature of each genome.

The G+C% for the weight vector of each node was reflected in the horizontal axis and increased from left to right in the Di-SOM (Fig. III.1D), and similar results were obtained in Tri- and Tetra-SOMs. Sequences with the same G+C% were separated by a complex combination of oligonucleotide frequencies resulting in species-specific separations. The underlying representation in SOMs allowed us to identify characteristic sequence patterns for individual genomes. The frequencies of each di-, tri-, and tetranucleotide in each weight vector in the 100-kb SOMs were calculated and represented as different levels of red and blue (Fig. III.2). Complementary oligonucleotides had similar distribution patterns as shown in AA and TT panels in Fig. III.2A, and therefore, only one example is presented except the AA/TT pair. Transitions between the red and blue levels coincided often with the species borders. For example, TA was diagnostic for separation of *Fugu* (F) from rice (R). Underrepresentation of CG was apparent in most regions of human, *Arabidopsis* (A), and *Fugu*. In the Tri- and Tetra-SOMs, only diagnostic examples for species separations are listed (Fig. III.2B and C). One clearest example was CATG, which was overrepresented in human and rice, underrepresented in *Drosophila* (D), and moderately represented in *Arabidopsis* and *Fugu*. So far as judged from one oligonucleotide, even in the 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 biological categories (according to species in this case).

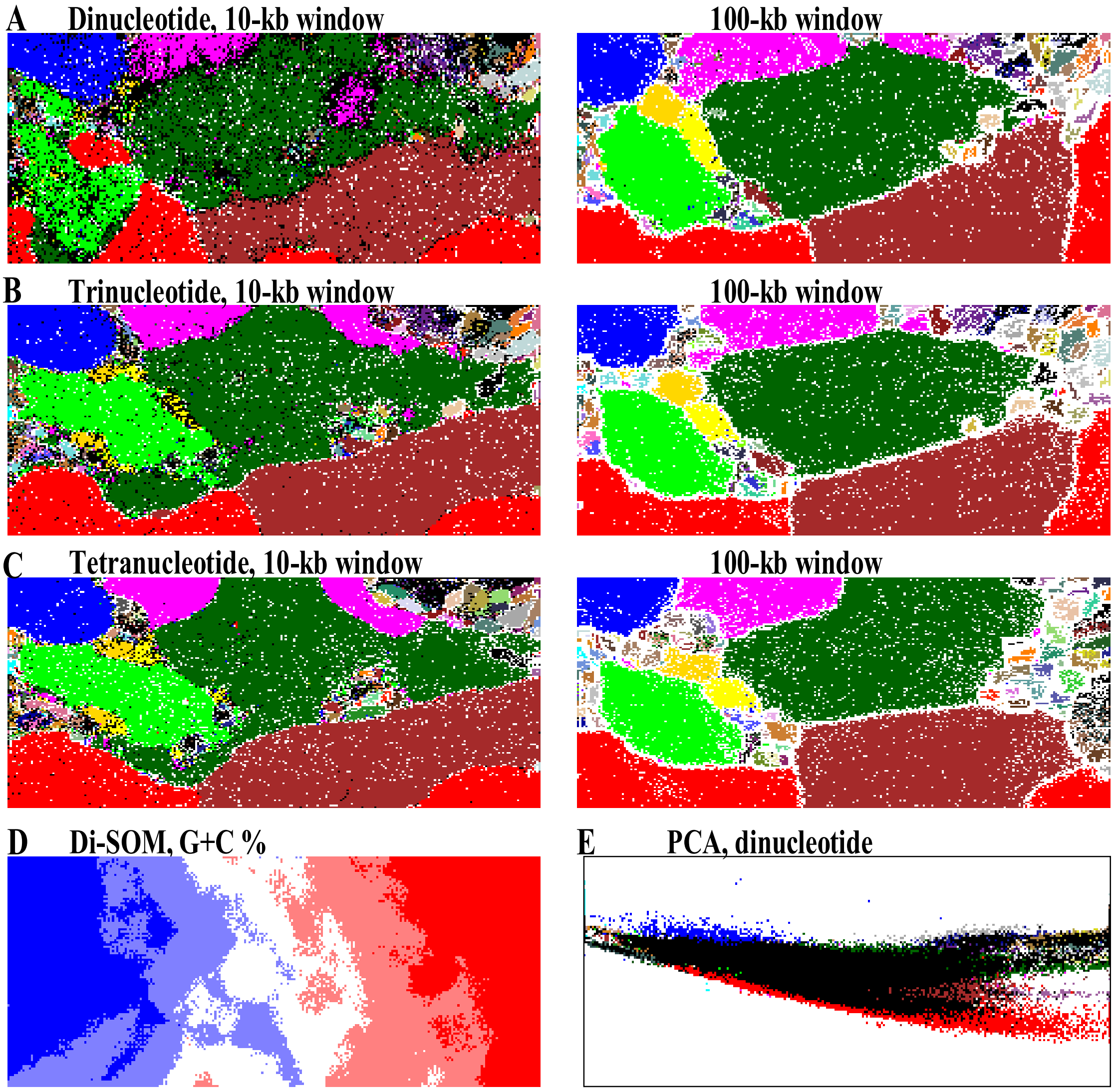


Figure. III.1 SOMs for nonoverlapping 10-kb and overlapping 100-kb sequences of 90 genomes. 10-kb and 100-kb di- (A), tri- (B), and tetranucleotide (C) SOMs. Nodes that contain sequences from more than one species are indicated in black, those including no sequences are indicated in white, and those including sequences from a single species are indicated in color as follows: 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* (🟫). (D) G+C% for the weight vector of each node was calculated and divided into five categories with an equal number of nodes. The nodes 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. (E) Sequence classification by the initial weight vectors set by PCA for the 10-kb Di-SOM. G+C% for each weight vector in the 10-kb Di-SOM.

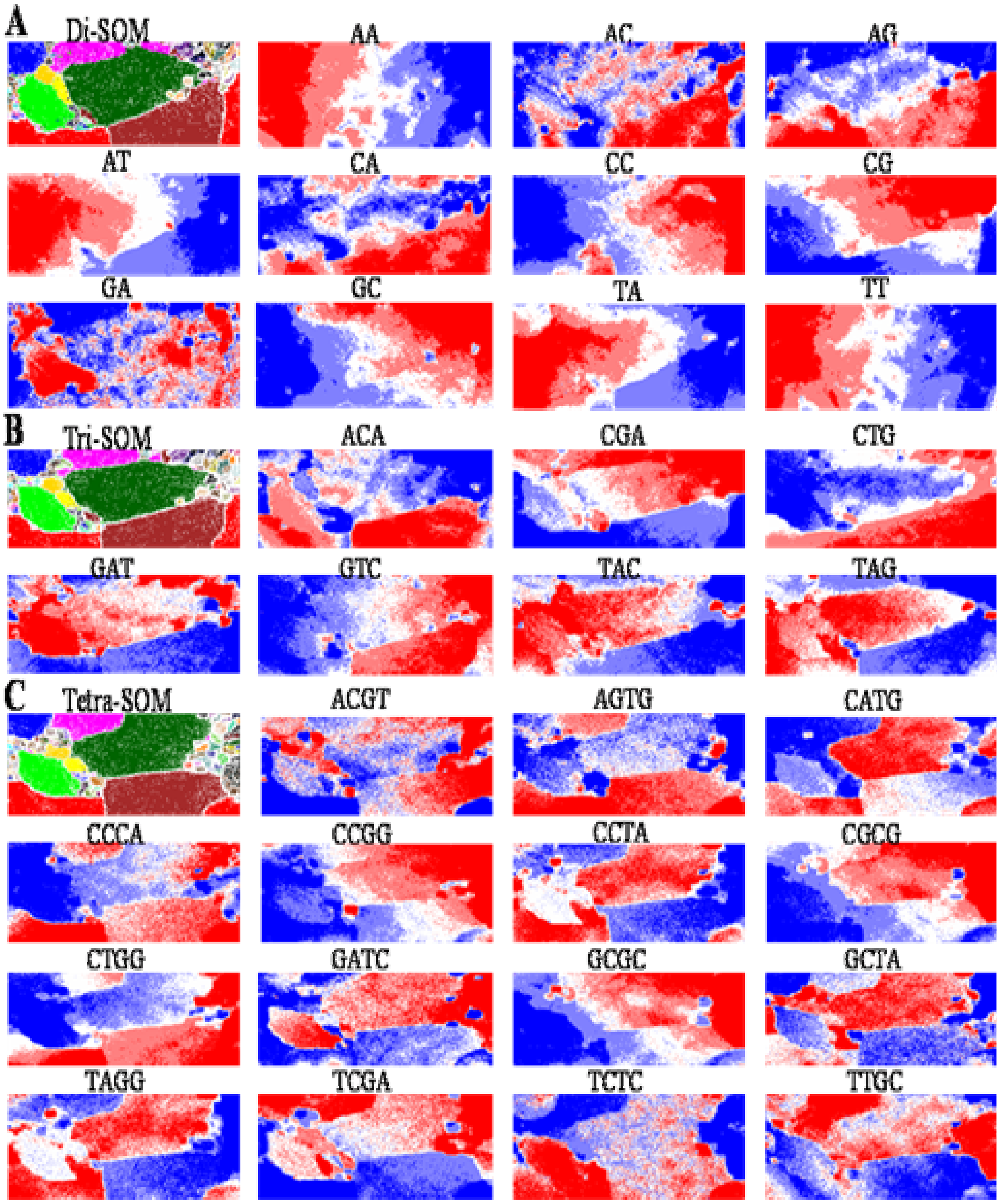


Figure III.2 Level of each di-(A), tri-(B), and tetranucleotide (C) in 100-kb SOMs. Diagnostic examples of species separations are presented. Levels of the di-, tri-, and tetranucleotide for the weight vector of each node in the respective SOMs of Fig. III.1 were divided into five categories with an equal number of nodes and are shown as described in Fig. III.1D. The 100-kb SOMs in Fig. III.1 are presented in the first panel; *C. elegans* (C), *Arabidopsis* (A), rice (R), *Drosophila* (D), *Fugu* (F), and human (H).

Biological implications of SOM separation and of intraspecies separation

In the DNA database, only one strand of a pair of complimentary sequences is registered. Our previous analysis of prokaryotic genomes showed that sequence fragments from a single genome are often split into two SOM territories that reflect the transcriptional polarities of the genes present in the fragment (Abe et al. 2003), as described in Chapter II; such a split was also observed in the SOM in Fig. III.1. For a few eukaryotes in Fig. III.1, intraspecies separations in SOMs were apparent, most evidently for *Drosophila* (🟪) and human (🟥), resulting in satellite territories for one eukaryote. This shows that SOMs can effectively depict intraspecies sequence differences. Distinct territories of one eukaryote were separated mainly in the horizontal direction, showing that the intraspecies separations were related to G+C% difference. The human genome contains long-range segmental G+C% distributions “isochores”, which are related with chromosomal bands (Bernardi et al. 1985; Ikemura 1985; Ikemura and Aota, 1988; Bernardi 1989), and correlation of these regional G+C% distributions with SOM separations was observed (Abe et al. 2003). The separation of *Drosophila* sequences may relate with euchromatic and heterochromatic chromosomal bands (Lindsley et al. 1992). In the 100-kb Tetra-SOM , the intraspecies separations were least evident and even *Drosophila* had one territory, indicating that the Tetra-SOM recognized similarities of tetranucleotide frequencies among sequences in a single genome while their G+C% varied significantly. When I inspected the 100-kb SOMs in detail, there were several minor territories composed of small numbers of sequences with specific characteristics. For example, a minor territory for *Arabidopsis* located near the rice territory was composed primarily of sequences from centromeric and subcentromeric regions. Analysis of intraspecies separations may provide fundamental information regarding structures of individual genomes.

Transitions between the red and blue levels for the diagnostic tetranucleotides often coincided exactly species borders including prokaryotic borders (Fig. III.2C). As an attempt to investigate the biological significance of the diagnostic tetranucleotides, I examined the correlation of levels of palindromic tetranucleotides with respective restriction enzyme systems for prokaryotic species by referring to the restriction enzyme database (REBASE; http://vent.neb.com/~vincze/genomes/) (Fig. III.3), as discussed in Chapter II. Of 16 restriction enzymes from 11 prokaryotes that have genes encoding 4-base cutter enzymes, the restriction site tetranucleotides were underrepresented in 14 instances (blue in Fig. III.3). This showed that SOM recognized the biological properties of genomic sequences properly and classified sequences according to biological categories.

I then considered the biological significance of diagnostic oligonucleotides in eukaryotes. The species-specific characteristics of DNA-synthesizing and -repairing enzymes, such as the sequence-recognition specificity of DNA primase and the context-dependent repair mechanism, may be the major factors responsible for the species-specific separations. Sequence preferences for ubiquitous DNA-binding proteins of individual species, such as histones, may be factors. Transition zones of various tetranucleotides (e.g., CATG and TAGG) were very sharp and coincided exactly the eukaryotic species-borders. As found for the restriction enzyme case of prokaryotic genomes, some tetranucleotides may have biological significance by themselves. Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). In the cases of signal and motif sequences, such as transcription signals with high, specific binding activities to transcription factors, they are thought to be significantly biased from the random occurrence statistically calculated from the genome base composition, as explained in detail in Chapter V.

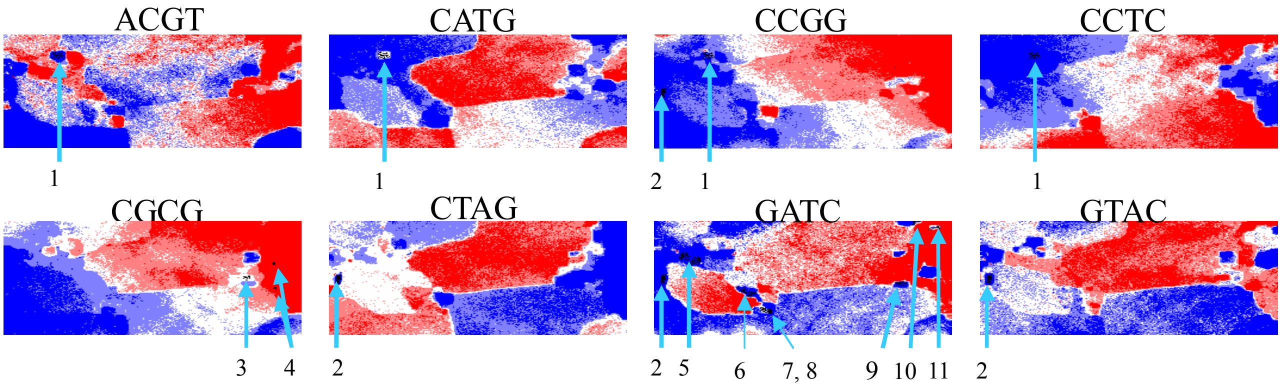


Figure III.3: Tetranucleotide distribution for prokaryotic genomes containing restriction enzyme genes on the 100-kb Tetra-SOM. Zones for prokaryotes that have genes encoding a restriction enzyme that recognizes the respective tetranucleotide are noted by light blue lines with the following numbers to show the species name: 1, *H. pylori*; 2, *M. jannaschii*; 3, *T. maritime*; 4, *T. pallidum*; 5, *S. aureus*; 6, *S. pneumoniae*; 7, *P. abyssi*; 8, *P. horikoshii*; 9, *A. fulgidus*; 10, *A. pernix*; and 11, *D. radiodurans*.

SOMs with human, mouse, and rat draft sequences

The present analysis of genome signatures is an example of comparative genomics, and the results are affected by the genomes analyzed. To reduce the effect, I analyzed most (if not all) eukaryotes for which almost complete genomic sequences are available, and the organisms were phylogenetically distant eukaryotes. When high-quality draft sequences are considered, those from additional genomes with biological and biomedical importance are available. To examine the clustering power of SOMs for closely related species, SOMs were generated from 4,600,000 non-overlapping 10-kb and overlapping 100-kb sequences from draft sequences of human, mouse and rat (2.4, 1.6, and 0.6 Gb, respectively). Nodes that contained sequences from one or two species are indicated in color and those that contained sequences from all three species are indicated in black (Fig. III.4A). Significant separation between human (red) and rodent (mouse and rat; light blue) sequences was observed even in the 10-kb SOMs. In the 10-kb Tetra-SOM, 41% of the human sequences was classified into human-specific territories, and 5% and 4% of mouse and rat sequences, respectively, were classified into mouse and rat territories. In the 100-kb SOMs, separation between human and rodents was very apparent, and 99% of human sequences were classified into the human territory. Furthermore, partial separations between mouse and rat were observed; 50% and 21% of mouse and rat sequences, respectively, were classified into mouse (dark blue) and rat (green) territories. Thus, SOMs can recognize unique sequence characteristics even in closely related species. Diagnostic tetranucleotides for species separations in the 100-kb Tetra-SOM are listed in Fig. III.4B. In the mouse and rat territories, ACAC and AATT were over- and underrepresented, respectively, in comparison to levels in human. ACAA and GACA were overrepresented in mouse, and GTGA was overrepresented in human.

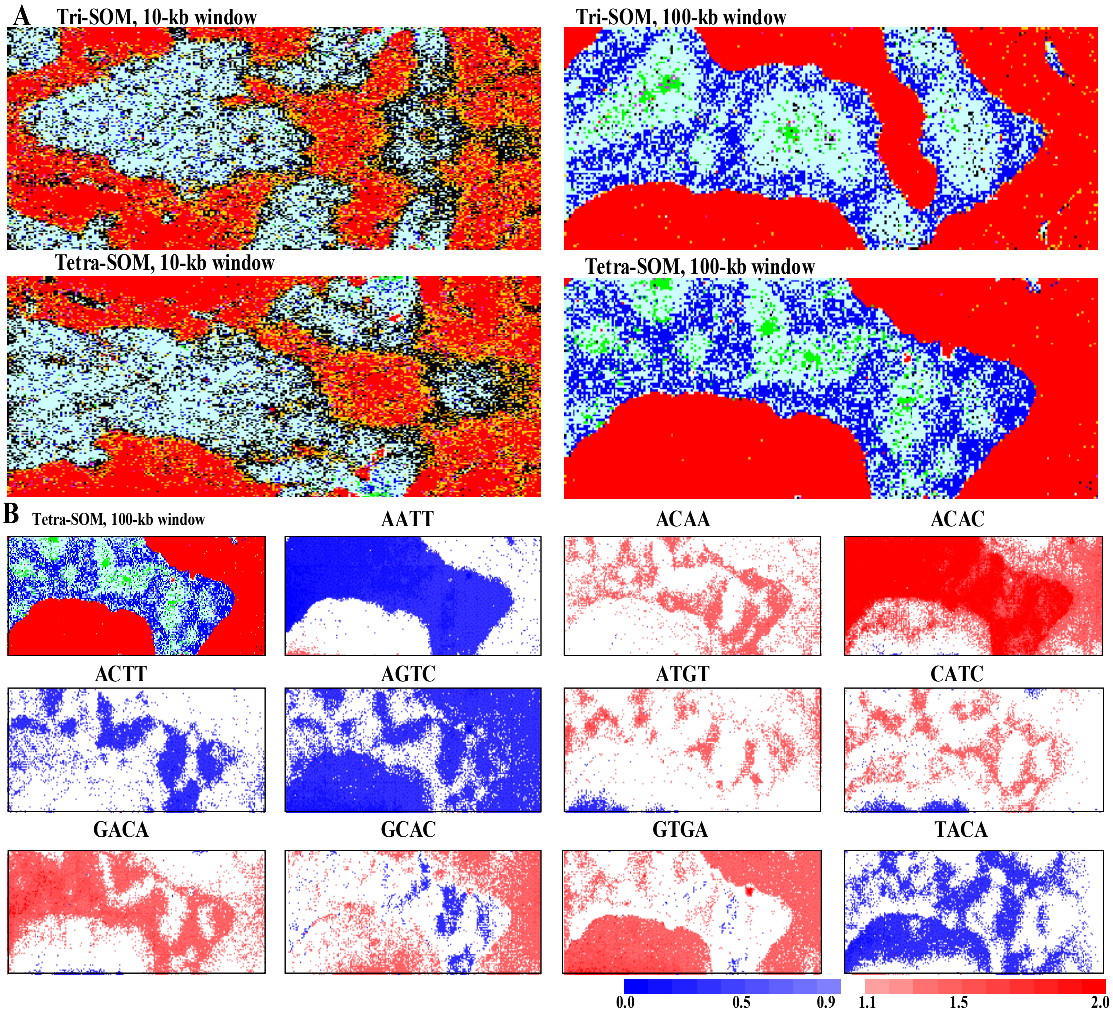


Figure III.4 SOMs for human, mouse and rat draft sequences. (A) Tri-and Tetra-SOMs were constructed with nonoverlapping 10-kb and overlapping 100-kb sequences with a 10-kb sliding step. Nodes that contain sequences from one or two species are indicated in color (human, red; mouse, dark blue; rat, green; mouse + rat, light blue; human + mouse, yellow; human + rat, violet), those from three species are indicated in black, and those that contained no sequences are indicated in white. (B) Levels of individual tetranucleotides in the weight vector of each node in the 100-kb Tetra-SOM was calculated after normalization of the mononucleotide composition of the node weight vector and presented as described in Fig. III.2. Diagnostic examples for species separations are shown.

III-4: DISCUSSION

SOMs can systematically extract profound genomic information from the oligonucleotide frequency in each genome. Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). In the cases of signal and motif sequences, such as transcription signals, they are thought to be significantly biased from the random occurrence statistically calculated from the genome base composition, as described in detail in Chapter V. Genetic signals are typically longer than tetranucleotides. Therefore, SOMs for longer oligonucleotides such as penta- and hexanucleotides may systematically reveal a wide range of genetic signal sequences in genomes. Because clustering power of SOM is very high, SOM should provide fundamental guidelines for understanding molecular mechanisms that have established sequence characteristics of individual genomes during evolution.

III-5: SUMMARY

I used the SOM to characterize oligonucleotide frequencies in a total of 1.4 Gb sequences derived from 90 prokaryotic and eukaryotic genomes for which complete genomic sequences are available. SOMs classified 140,000 10-kb sequences from the 90 genomes mainly according to species and could unveil hidden sequence characteristics of each genome. Because the classification power is very high, SOM is an efficient and fundamental bioinformatic strategy for extracting a wide range of genomic information from a vast amount of eukaryotic and prokaryotic sequence data.

Chapter IV: Novel Bioinformatics for Phylogenetic Classifications of Genomic Sequences from Uncultured Microorganisms in Environmental, Clinical Samples

IV-1: INTRODUCTION

The vast majority of environmental microorganisms have not been cultured under laboratory conditions, and conventional experimental studies have been limited. The genomes that lack experimental characterization may have a wide range of novel, valuable genes of both scientific and industrial interest (Amann et al. 1995; Hugenholtz et al. 1996). Metagenomic approach, which is genomic analysis of uncultured microorganisms, has been developed to identify novel and industrially useful genes and to study microbial diversity in a wide variety of environments (Amann et al. 1995; Hugenholtz et al. 1996; Rondon et al. 2000; MacNeil et al. 2001; Lorenz et al. 2002; Courtois et al. 2003; Schloss et al. 2003). In the approach, DNA is extracted directly from mixed genomes in an environmental sample without cultivation of microorganisms, and the DNA fragments are cloned into vectors, then sequenced. Therefore, bioinformatic strategies to predict how many and what types of genomes are present in a sample are needed. It is also important to clarify how novel the genomic sequences are. To address such issues, I attempted to use SOM. In Chapters II and III, I constructed SOMs for di-, tri-, and tetranucleotide frequencies in sequence fragments from a wide rage of prokaryotic and eukaryotic genomes. In the resulting SOMs, the sequences were clustered according to species without any information regarding the species, and increasing the oligonucleotides in length from di- to tetranucleotides increased the clustering power. In this Chapter, SOM of tetranucleotide frequencies was improved for phylogenetic classification of genomic sequences from uncultured prokaryotes in environmental and clinical samples.

IV-2: METHODS

The initial weight vectors were defined by PCA instead of random values, and weight vectors (w_{ij}) were arranged in the two-dimensional lattice denoted by i (=0, 1, .., I-1) and j (=0, 1, .., J-1), as described in Chapter II. I was set as 350, 300 and 100 in Figs. IV.1, IV.2A and IV.2B, respectively, and 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. When pentanucleotide frequencies were analyzed, it took four times of the computation time of Tetra-SOM but the improvement of the resolving power for 1-kb sequences was not significant.

IV-3: RESULTS

SOMs of prokaryotic genomes for which complete sequences are available

To investigate the clustering power of SOM for a wide range of prokaryotic genome sequences, I first analyzed 81 prokaryotic genomes for which complete sequences are available (a total of 226 Mb); in this analysis, only one genome representing different strains of one species or closely related species with a relatively large genome was used to avoid overrepresentation of certain large genomes. SOMs were constructed for tetranucleotide frequencies of 226,000 nonoverlapping 1-kb (Fig. IV.1A) and overlapping 5-kb sequences with a 1-kb sliding step (Fig. IV.1B). These short sequences were tested because such fragments can be efficiently cloned and sequenced even for samples obtained from extreme environments. To obtain the initial weight vectors, tetranucleotide frequencies in the sequences were analyzed by PCA. After 100 learning cycles, the frequencies of tetranucleotides in these sequences were reflected as the weight vectors in the Tetra-SOMs in Fig. IV.1. Nodes that contained sequences from a single species are indicated in color, and those that contain sequences from more than one species are indicated in black (Species in Fig. IV.1). The clustering power of the 5-kb SOM was much higher than that of the 1-kb SOM.

In the DNA database, only one strand of a pair of complimentary sequences is registered. The previous analysis of prokaryotic genomes in Chapter II showed that sequence fragments from a single genome are often split into two SOM territories that reflect the transcriptional polarities of the genes present in the fragment (Abe et al. 2003); such a split was also observed in the Tetra-SOM in Fig. IV.1. For phylogenetic classification of sequences from uncultured prokaryotes, it is not necessary to know the transcriptional polarity in the sequence fragment and the split into two territories complicates assignment to species. Therefore, I tested a new type of SOM in which a pair of complementary tetranucleotides (e.g., AATC and GATT) was added, and the SOM for the degenerate sets of tetranucleotides was designated DegeTetra-SOM. This approximately halved the computation time, and the clustering power was higher than that of the Tetra-SOM (Table IV.1). In the 5-kb DegeTetra-SOM (Fig. IV.1B), sequences were separated primarily into nonoverlapping species-specific territories. For example, approximately 75% of 5-kb sequences were classified into the correct species territory (Table IV.1). SOM recognized in the 5-kb sequences the key combinations of tetranucleotide frequencies that are the signature features of each genome. The reason that increasing the sequence length from 1 kb to 5 kb substantially improved the resolving power may be due to reduced statistical scattering of tetranucleotide frequencies in 5-kb sequences. Because increasing the length from 5 kb to 10 kb did not improve the resolving power appreciably (data not shown), I used the 5-kb SOM in later analyses of sequences from species-unidentified prokaryotes.

In the phylogenetic classification of uncultured prokaryotes, especially those found in novel or extreme environments, classification into major phylogenetic groups rather than into individual species is important. Therefore, I tested in advance the classification of the 226,000 sequences of known prokaryotes into 11 major phylogenetic groups (Group in Fig. IV.1); nodes that contain sequences from one phylogenetic group are indicated in color, and those that contain sequences from more than one group are shown in black. Number of black nodes, which contained sequences of more than one group, decreased significantly in comparison to that of the species classification. Approximately 88% of 5-kb sequences were classified into the appropriate group territory in the DegeTetra-SOM (Table IV.1). Phylogenetically related species often had neighboring territories and, therefore, produced united territories in the group classification. Interestingly, even in the united territories, species borders were often visible as continuous white nodes that contained no genomic sequences in the 5-kb SOMs (see Group in Fig. IV.1B). The vectors of species-specific nodes, even near a species-territory border, were distinct between species, and species borders could be drawn automatically on the basis of the contiguous white nodes.

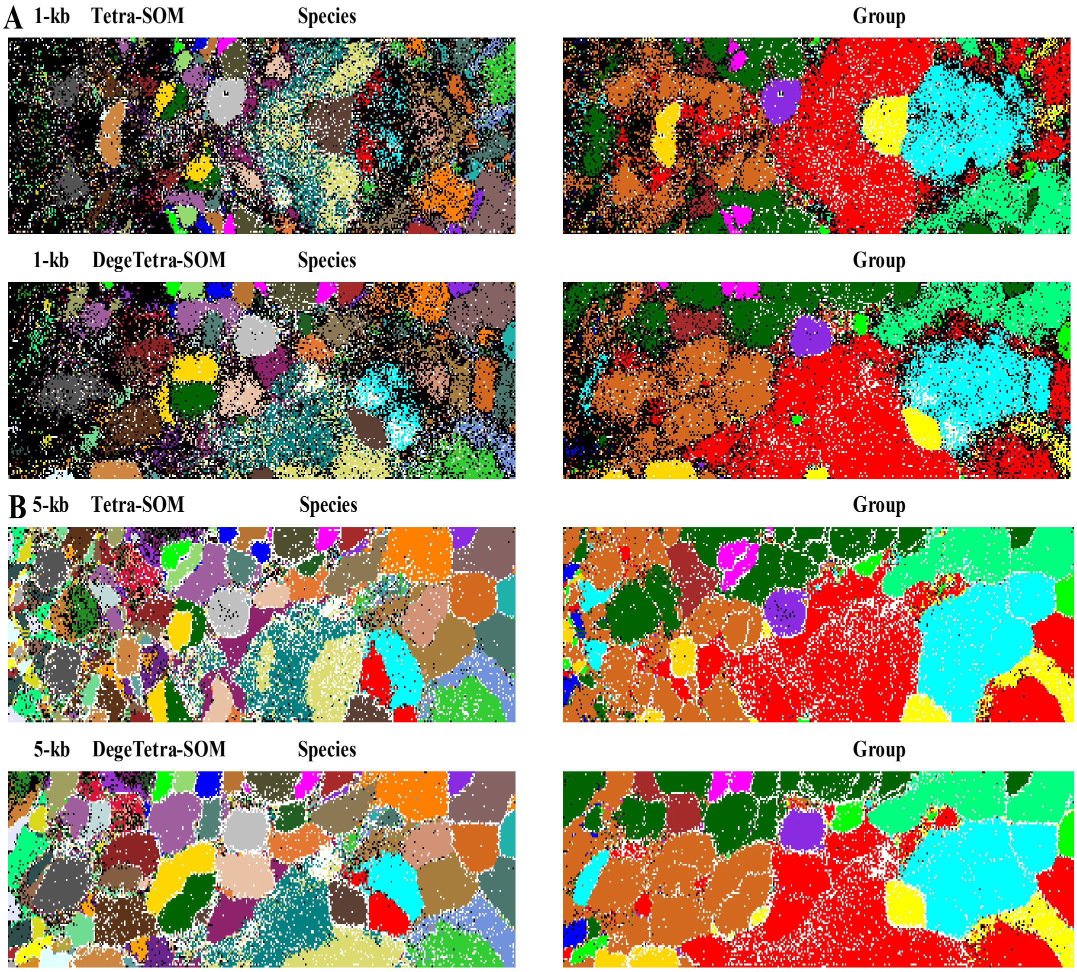


Figure IV.1 Tetra- and DegeTetra-SOMs for nonoverlapping 1-kb (A) and overlapping 5-kb sequences (B) of 81 genomes of 65 prokaryotic species. Species: sequence classifications into each species. Nodes that contain sequences from more than one species are indicated in black, and those containing sequences from a single species are indicated in color as follows: *Aquifex aeolicus* (🟩), *Archaeoglobus fulgidus* (🟦), *Aeropyrum pernix* (🟪), *Agrobacterium tumefaciens* (🟦), *Borrelia burgdorferi* (🟨), *Bacillus halodurans* (🟨), *Brucella melitensis* (🟥), *Bacillus subtilis* (🟩), *Buchnera* sp. (🟫), *Clostridium acetobutylicum* (🟩), *Caulobacter crescentus* (🟧), *Campylobacter jejuni* (□), *Chlamydia muridarum* (🟩), *Clostridium perfringens* (□), *Chlamydophila pneumoniae* (🟥), *Chlamydia trachomatis* (🟧), *Deinococcus radiodurans* (🟦), *Escherichia coli* (🟦), *Fusobacterium nucleatum* (⬛), *Halobacterium* sp. (🟪), *Haemophilus influenzae* (🟧),

Helicobacter pylori (🟧), *Lactococcus lactis* (🟩), *Listeria monocytogenes* and *innocua* (🟫), *Methanosarcina acetivorans* (🟪), *Mycoplasma genitalium* (🟨), *Methanococcus jannaschii* (🟨), *Methanopyrus kandleri* (🟥), *Mycobacterium leprae* (🟫), *Mesorhizobium loti* (🟫), *Mycoplasma pneumoniae* (🟦), *Mycoplasma pulmonis* (🟧), *Methanobacterium thermoautotrophicum* (🟧), *Mycobacterium tuberculosis* (🟧), *Neisseria meningitidis* (🟫), *Pyrococcus abyssi* (🟩), *Pseudomonas aeruginosa* (🟦), *Pyrobaculum aerophilum* (🟩), *Pyrococcus furiosus* (🟪), *Pyrococcus horikoshii* (🟪), *Pasteurella multocida* (🟪), *Rickettsia conorii* (⬛), *Rickettsia prowazekii* (🟫), *Ralstonia solanacearum* (🟦), *Streptococcus agalactiae* (🟫), *Staphylococcus aureus* (⬛), *Streptomyces coelicolor* (🟫), *Sinorhizobium meliloti* (🟧), *Streptococcus pneumoniae* (🟥), *Sulfolobus solfataricus* (🟩), *Sulfolobus tokodaii* (🟧), *Salmonella typhimurium* (🟨), *Synechocystis* sp. (⬜), *Thermoplasma acidophilum* (🟦), *Thermotoga maritima* (🟩), *Teponema pallidum* (🟩), *Thermoanaerobacter tengcongensis* (🟩), *Thermoplasma volcanium* (⬜), *Ureaplasma urealyticum* (⬜), *Vibrio cholerae* (🟧), *Xanthomonas campestris* and *axonopodis* (🟩), *Xylella fastidiosa* (🟧), and *Yersinia pestis* (🟪). Sequences of closely related prokaryotes such as different strains of one species are indicated by the same color. Group: sequence classifications into 11 phylogenetic groups. Nodes that contain sequences from more than one group are indicated in black, and those containing 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 (🟪), and others (🟩).

Table IV.1 The proportion (%) of sequences classified into correct

Sequence on SOM	Tetra-SOM		DegeTetra-SOM	
	Species (%)	Phylogenetic group (%)	Species (%)	Phylogenetic group (%)
1 kb on 1kb-SOM	37.0	61.2	40.6	69.2
5 kb on 5kb-SOM	67.8	84.9	74.6	88.0
1 kb on 5kb-SOM	62.1	77.9	66.0	78.9

Colored lattice on SOM	Tetra-SOM		DegeTetra-SOM	
	Species (%)	Phylogenetic group (%)	Species (%)	Phylogenetic group (%)
1 kb on 1kb-SOM	50.6	70.8	55.4	73.8
5 kb on 5kb-SOM	72.4	94.1	77.2	96.6

Sequences introduced through horizontal transfer

Genome segments introduced through horizontal transfer from phylogenetically distant species tend to retain the sequence characteristics of the donor genome and can be distinguished from those of the host genome (Jeltsch et al. 1996). Even in the 5-kb DegeTetra-SOM for the group classification (Group in Fig. IV.1B), there are nodes marked in black, which should contain sequences with tetranucleotide frequencies distinct from those of the major portion of the respective genome. Such sequences should correspond, at least in part, to segments transferred horizontally from a phylogenetically distant genome. When *B. subtilis* sequences located outside the Firmicutes territory were investigated, many were found to be derived from A+T-rich islands where prophage and prophage-like sequences are clustered (Kunst et al. 1997). In shotgun sequencing studies of uncultured prokaryotes, SOMs (as well as conventional sequence homology searches) classify such horizontally transferred sequences into the donor genome group. Although this information is interesting, it creates problems in understanding microbial diversity within an environmental sample. Therefore, it is important to study known genomes in advance to estimate the proportion of sequences presumably transferred from other genomes and to find possible strategies to solve the problems caused by horizontal transfer (refer to Discussion). When oligonucleotide frequencies were calculated and normalized for the sequence length, 1-kb sequences could be mapped onto the 5-kb SOM. The proportions of 1-kb sequences classified into the correct group territory were 69% and 79% on the 1- and 5-kb DegeTetra-SOMs, respectively (Table IV.1). The increased hit level in mapping on the 5-kb SOM is because SOM can extract genome-specific characteristics of oligonucleotide frequencies more accurately as the analyzed sequences become longer. This means that for phylogenetic classification of sequences from uncultured mixed prokaryotes, even short sequences must be mapped on the SOM constructed with longer sequences (e.g., 5 kb) of species-known genomes. It is also conceivable that sizes of many horizontally transfer segments found in the current genomes are much shorter than 5 kb. The observation that the hit level in the group classification was higher than that in the species classification may reflect, at least in part, that chance of horizontal transfer across distinct phylogenetic groups is lower than that across the species within one group.

Sequences from species-unidentified prokaryotes

One goal of metagenomic studies is to reconstruct individual genomes of uncultured microorganisms by sequencing a large amount of the genomic DNAs contained in metagenomic libraries derived from an environmental sample (Rondon et al. 2000; MacNeil et al. 2001; Courtois et al. 2003; Schloss et al. 2003). Such technology should assist in developing accurate views of the ecology of environmental microorganisms and allow extensive surveys of sequences useful in industrial and scientific applications. However, with a simple compilation of a large number of sequence fragments, it appears to be impossible to predict what kinds and the ratios of species present in an environmental sample, to which lineages the species belong, and how the genomes are novel. A conventional approach is to catalog 16S rRNA gene (16S rDNA) sequences (Pace et al. 1985) mainly by polymerase chain reaction (PCR) amplification followed by sequencing (Amann et al. 1995; Hugenholz et al. 1996). However, for totally novel genomes, it might be difficult to design 16S rDNA-specific primers. Furthermore, 16S rDNAs are not thought to be industrially useful. It is most desirable, in addition to the 16S rDNA method, if we can develop a method by which microbial biodiversity is assessed automatically during the process of searching for and identifying genes with industrial and scientific significance. I propose here the development of such a method based on SOMs.

As the first practical application of the proposed method, I characterized non-rDNA sequences from unidentified or uncultured prokaryotes that were submitted to DDBJ/EMBL/GenBank by many groups (Thomsen et al. 2001; McMahon et al. 2002; Stokes et al. 2001; Knietsch et al. 2003); sequences were submitted directly to DDBJ/EMBL/GenBank often without publication by literatures. I found 660 non-rDNA genomic sequences longer than 1 kb in the Release of DDBJ and classified the 660 sequences into 11 prokaryotic groups in the following way. DegeTetra-SOM was constructed in advance with nonoverlapping 5-kb sequences from the 147 prokaryotic genomes for which almost complete sequences are currently available (Known prokaryotes in Fig. IV.2A). For this SOM, instead of the 81 genomes used in Fig. IV.1, 147 genomes were used to analyze a wider range of sequences because sequence redundancy due to inclusion of closely related genomes, such as those of different strains of one species, is not problematic for the group classification. Nodes that contained sequences from one group are marked in colors, and those that contained sequences of more than one group are shown in black. The 660 non-rDNA sequences from species-unidentified prokaryotes were then mapped on this 5-kb DegeTetra-SOM (Unidentified prokaryotes in Fig. IV.2A). Detailed inspection of the sources of these non-rDNA sequences with reference to DDBJ annotations revealed that approximately 50% corresponded to sequences obtained from rumen contents. These sequences are specified in pink. There was a characteristic zone composed primarily of the respective sequences (11% of sequences from the rumen contents; specified with a green arrow in Unidentified prokaryotes in Fig. IV.2A) in an Archaea territory that contained *Methanobacterium, Methanosarcina* and *Thermoplasma* sequences. This indicated that the rumen sequences were derived from Methanogens. Another characteristic clustering for the rumen sample was in the δ-proteobacteria territory containing *Desulfovibrio desulfuricans* (8% of the rumen sequences; a dark yellow arrow), indicating the sequences were derived from sulphate-reducing bacteria. The anaerobic species are expected to be present in rumen. The map locations of all 660 non-rDNA sequences on the 5-kb DegeTetra-SOM are presented on Supplementary data IV.1.

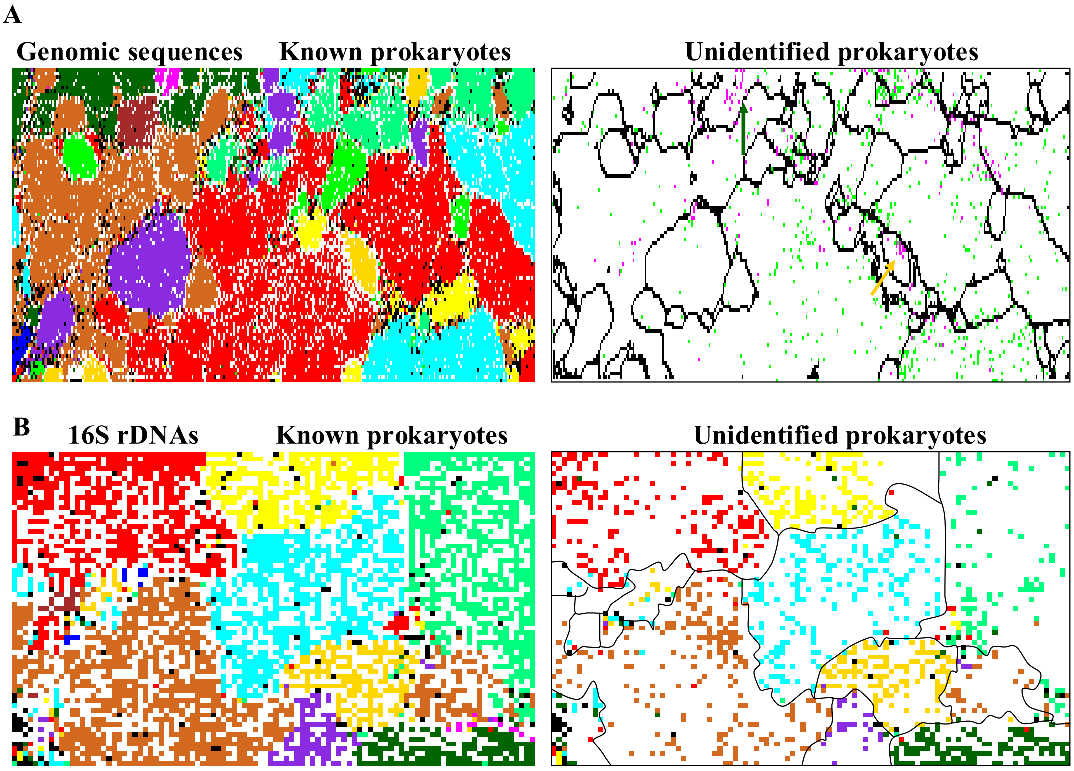


Figure IV.2 SOMs for phylogenetic classification of sequences from species-unidentified prokaryotes. (A) 5-kb DegeTetra-SOM of 147 prokaryotic genomes. Known prokaryotes: sequence classification into 11 phylogenetic groups for species-known prokaryotes. Nodes are marked as described in the group classification in Fig. IV.1. Unidentified prokaryotes; 660 non-rRNA sequences from species-unidentified prokaryotes were mapped on the 5-kb DegeTetra-SOM. The 343 sequences from rumen contents submitted directly to DDBJ/EMBL/GenBank by Matsui et al. (AB085236-AB085579) are specified in pink. (B) Tetra-SOM of 16S rDNAs from known prokaryotes. Known prokaryotes: classification into 11 phylogenetic groups. Nodes are marked as described in the group classification in Fig. IV.1; others were not included. Unidentified prokaryotes; rDNA sequences from species-unidentified prokaryotes, for which the group classification is annotated in DDBJ/EMBL/GenBank, were mapped on the 16S rDNA Tetra-SOM. Nodes where sequences of a single group were mapped are colored showing the group and those containing sequences of more than one group are marked in black.

SOMs for rDNA sequences

Because 16S rDNA sequences were highly conserved during evolution, their sequences have been used for detailed phylogenetic classification of prokaryotic species, including uncultured prokaryotes (Pace et al. 1985; Ludwig et al. 1994). Approximately 22,800 16S rDNA sequences longer than 1 kb from 6,080 known prokaryotes are compiled in DDBJ/EMBL/GenBank. Combination of SOMs for genomic and 16S rDNA sequences will provide a tool for detailed phylogenetic studies of these sequences from environmental prokaryotes. A full length of 16S rDNA is approximately 1.5 kb and represents approximately 0.1% of each prokaryotic genome and, therefore, the data from 16S rDNA sequences did not contribute significantly to the SOMs shown in Figs. IV.1 and IV.2A. Because the G+C% of rDNAs is clearly higher than those of other genome regions, rDNAs were located near borders of or slightly apart from the territory of the respective phylogenetic group.

During shotgun sequencing-based screens for industrially useful genes from environmental microorganisms, many rDNA sequences are also determined. In the shotgun approach, PCR primer design is unnecessary, and unbalanced sequence amplification does not occur. I constructed Tetra- and DegeTetra-SOMs with the 22,800 16S rDNA sequences from known prokaryotes after normalization for the sequence length. Clear clustering according to phylogenetic group was observed; 97% and 95% of sequences were classified into the appropriate group territory on the Tetra- and DegeTetra-SOMs, respectively, and the result of the Tetra-SOM is presented (Known prokaryotes in Fig. IV.2B). The reason that the Tetra-SOM gave the better classification than the DegeTetra-SOM may be due to that directions of all rDNA sequences compiled in the DNA databases represented one polarity corresponding to rRNA sequences and that the Tetra-SOM could detect the sequence characteristics of this polarity, which could not be detected by the DegeTetra-SOM. The finding that the hit level of 16S rDNA classification into the correct group territory was higher than that of genomic sequences may indicate that the occurrence of horizontal transfer of rDNAs, if present, is lower than that of other genome portions. I then searched for 16S rDNA sequences from species-unidentified prokaryotes in DDBJ/EMBL/GenBank and found 9,540 sequences longer than 1 kb. Because the purpose of 16S rDNA sequencing of the unidentified species is phylogenetic assignment with reference to 16S rDNA sequences of species-known prokaryotes, the phylogenetic groups assigned with conventional phylogenetic methods such as sequence-homology searches are annotated in DDBJ/EMBL/GenBank for a large portion of the species-unidentified sequences. I selected 3,320 sequences for which classification into the major prokaryotic groups is indicated in DDBJ/EMBL/GenBank and mapped them on the 16S rDNA Tetra-SOM for known prokaryotes. Nodes that contain the sequences of a single group are indicated in color of the annotated group, and those having sequences from more than one group are marked in black (Unidentified prokaryotes in Fig. IV.2B). The major zones were marked in color, and the color pattern resembled that of known prokaryotes (Known prokaryotes in Fig. IV.2B), showing that assignments based on SOM were almost the same to those obtained with conventional phylogenetic methods. It should be stressed that with this metagenomic method, PCR primer design is not needed, and unbalanced amplification among distinct rDNAs is not a concern.

I next mapped the residual 6,220 rDNAs for which the phylogenetic assignment is not noted in DDBJ/EMBL/GenBank. Although global pattern for these uncharacterized sequences resembled that of the group-assigned sequences, the proportion of sequences located near the borders of group territories appeared to increase; map locations of all 6,220 rDNAs on the Tetra-SOM are presented on Supplementary data IV.2. Sequences mapped near territory borders may correspond to sequences derived from prokaryotes for which 16S rDNA sequences have not been accumulated and underrepresented in the present SOM.

Separation between prokaryotic and eukaryotic sequences

When considering phylogenetic classification of uncultured environmental microorganisms, it is necessary to construct SOMs with both prokaryotic and eukaryotic sequences because varieties of eukaryotic microorganisms are present in the environmental sample. Furthermore, when microorganisms symbiotic/parasitic with a higher eukaryote are analyzed, sequences from this eukaryote may be included in the sequence collection. To examine the SOM separation between prokaryotic and eukaryotic sequences for a wider range of species than that analyzed in Chapter III, 5-kb sequences from 13 eukaryotic genomes, which have been sequenced extensively, were analyzed simultaneously with 5-kb sequences from the 147 prokaryotes used in Fig. IV.2. To avoid excess representation of eukaryotic sequences and 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. IV.3). 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 was also apparent (Fig. IV.3B). Examples of SOM for simultaneous analysis of eukaryotic and prokaryotic sequences, which was constructed without the random selection for eukaryotic sequences, are presented on Fig. III.1

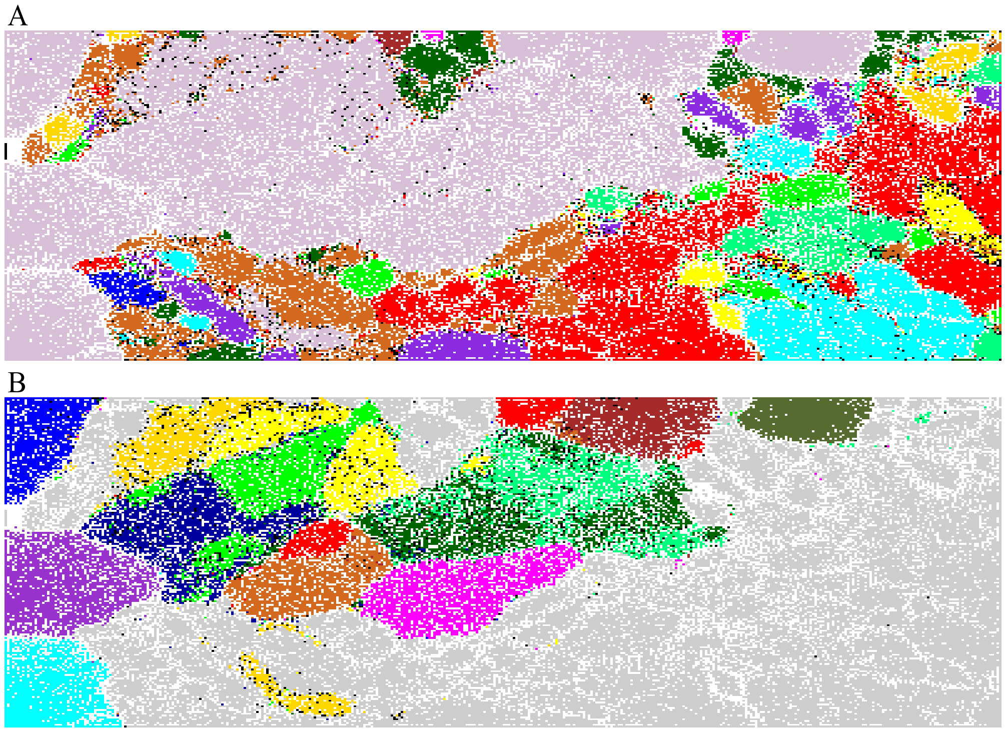


Figure IV.4. DegeTetra-SOMs of nonoverlapping 5-kb sequences from 147 prokaryotes and 13 eukaryotes. (**A**) Nodes that contain prokaryotic sequences from more than one phylogenetic group are indicated in black, and those that contain sequences from a single group are indicated in color as follows:α-proteobacteria (🟦), 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**) Nodes that contain sequences only from prokaryotic genomes are indicated in color (⬜). Nodes 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* (🟥). Nodes that contain sequences from more than one eukaryotic species or from both eukaryotic and prokaryotic species are indicated in black.

Sequences from Sargasso Sea samples.

Venter *et al.* (2004) recently applied “whole-genome shotgun sequencing” to uncultured microbial populations collected from seawater samples from the Sargasso Sea and determined a total of 1.05-Gb sequence: approximately 811,000 sequence fragments have been deposited in GenBank. To classify the species-unknown sequences into phylotypes, we constructed in advance a DegeTetra-SOM with 210,000 non-overlapping 5-kb sequences from 1502 species-known prokaryotes for which at least 10-kb of sequence has been deposited in the DNA databases (Known Species in Fig. IV.5A).The genomic sequences from 1502 prokaryotes were used to obtain an SOM that is better suited for phylogenetic classification of species-unknown sequences. With reference to the NCBI Taxonomy Database, these 1502 prokaryotes were classified into 25 phylotypes, and lattice points that contained sequences from one type are marked in color, and those that contained sequences from more than one type are shown in black. I next mapped Sargasso Sea sequences on the 5-kb DegeTetra-SOM constructed with the sequences from the 1502 species-known prokaryotes (Sargasso Sequences Fig. IV.5A). Evidently skewed distribution was observed; there were zones where a large portion of the Sargasso sequences were localized and also zones where the sequences were rarely classified. This was clearly visualized by the 3D representation in which the number of Sargasso sequences classified into each lattice point is indicated by the height of the bar (3D in Fig. IV.5A). For example, a typical abundant zone was found at the bottom left, which was composed primarily of very AT-rich sequences. In this zone, sizes of individual territories of known phylotypes were rather small and associated with each other in a rather complex manner, indicating that sequences from phylotypes underrepresented in the DNA databanks were located. A large number of A+T-rich Sargasso sequences, for which there were no related sequences in the DNA databanks, might be classified into the zone. To examine this possibility, I next constructed a DegeTetra-SOM with the sequences from the 1502 species plus Sargasso sequences (Fig. IV.5B), and colored the lattice points that contained sequences from a single phylotype concerning the species-known sequences, as described in Fig. IV.5A. Lattice points that contained sequences from more than one group are shown in black and those that contained only Sargasso sequences by a color (🟩). The territories of species-known sequences shrank appreciably when compared with the territories in Fig. IV.5A, and a large area, especially at the left side representing very AT-rich sequences, was occupied primarily by Sargasso sequences. This is because four times more Sargasso sequences than the species-known sequences were included in this SOM and a large portion of the Sargasso sequences had oligonucleotide frequencies distinct from those of known species.

In the SOM method, distances between reference vectors of the neighboring lattice points, which can be visualized with gray levels by a Umatrix method (Kraaijveld et al., 1992; Ultsch, 1993; Livarinen et al., 1994), provides valuable information concerning levels of differences in characteristics of sequences classified into the neighboring lattice points in the respective zone (Umatrix in Fig. IV.5B). Clustering of lattice points with very low gray levels (and thus white) represents zones composed primarily of sequences with similar oligonucleotide frequencies and thus most likely of sequences derived from the same or closely related genomes. In contrast, clustering of lattice points with high gray levels in Umatrix may represent the regions containing sequences derived from heterogeneous genomes of low abundance. In the case of sequences from novel genomes of lower abundance even in an environmental sample, sequences from the same or related species could not be represented significantly in either the DNA databanks or the environmental sample. The novelty of such sequences can be determined by calculating the distance between the vector of the sequence and the reference vector of the sequence-mapped lattice point (i.e., the lattice point with the minimum distance from the sequence vector in the multidimensional space).

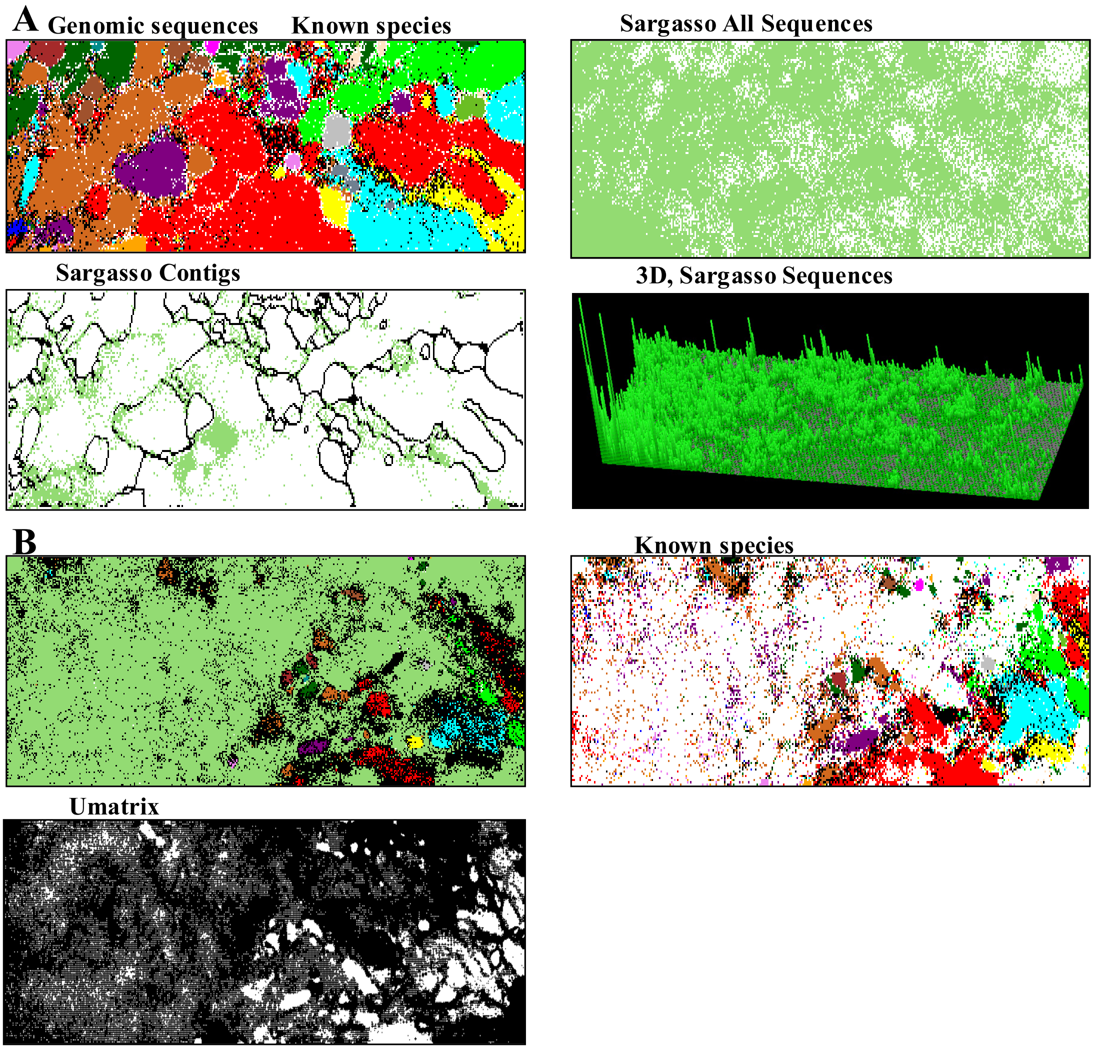


Figure IV.5 SOMs for phylogenetic classification of sequences from Sargasso sequences. (A) DegeTetra-SOM of non-overlapping 5-kb sequences of 1502 prokaryotes. Known species, sequence classification into 25 prokaryotic groups. Lattice points that include sequences from more than one phylogenetic group are indicated in black, and those that contain sequences from a single group are indicated in color as follows: α-proteobacteria (🟦), β-proteobacteria (🟨), δ-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 (🟥). Sargasso All Sequences, 811,000 sequences from the Sargasso Sea sample were mapped on the 5-kb DegeTetra-SOM after normalization of the sequence length. Sargasso Contigs, 1-kb sequences derived from contigs longer than 1 kb. 3D, 3D representation of Sargasso sequences. Logarithm of the number of Sargasso sequences mapped on each lattice points is indicated by the height of the bar. (B) SOM of species-known sequences plus Sargasso sequences. DegeTetra-SOM was constructed with 211,000 5-kb sequences from 1502 prokaryotes and 811,000 Sargasso sequences. Lattice points that contain sequences from a single prokaryotic group regarding the species-known sequences are indicated in color as described in Fig. 2B, those that contain sequences more than one group are indicated in black, and those that contain only Sargasso sequences are indicated in color (🟩). known species, representation of Known species, sequence classification into 25 prokaryotic groups. Umatrix, distances of reference vectors between the neighboring lattice points visualized with gray levels. Umatrix was calculated according to umat in the SOM_PAK obtained from the URL (http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml).

IV-4: DISCUSSION

The number of 23S rDNA sequences compiled in DNA databases was less than 5% of that of 16S rDNAs. When 23S rDNA sequences longer than 1 kb were analyzed with the Tetra-SOM, they were again classified primarily according to the phylogenetic groups (data not shown), indicating that 23S rDNA sequences can also be used for the phylogenetic classification. During course of extensive shotgun sequencing of an environmental sample, substantial amounts of non-rDNA and rDNA sequences will become available. Combined SOM analyses of non-rDNA and rDNA sequences may provide detailed information for collective genomes in an environmental sample on the basis of profound knowledge of the phylogeny of rDNAs. This may also solve, at least in part, the complications caused by horizontal transfer of sequences. For example, when no rDNA sequences are found in a certain phylogenetic group territory in the rDNA SOM but a statistically relevant amount of non-rDNA sequences is found in this group territory in the SOM of genomic sequence, the non-rDNA sequences may be candidates for horizontally transferred sequences. In the case where a statistically relevant amount of rDNA sequences maps to a certain group territory in the rDNA SOM and assigned to be derived from one (or a few closely related) species with conventional phylogenetic methods, a major portion of the non-rDNA sequences mapped to this group territory in the genomic sequence SOM can be predicted to be primarily derived from the respective genomes. When 16S rDNA data are obtained by conventional PCR-based methods, the combination of these results with the SOM data may also provide detailed phylogenetic characterization of non-rDNA sequences. Collectively, the combination of SOM for rDNA and non-rDNA sequences will reconstruct *in silico* the rDNA and non-rDNA sequences that belong to a single genome but were cloned independently. Integration of SOM into metagenomic strategies can further enhance genomic studies of biodiversity and screens for useful genomic sequences in collective genomes obtained from specific environments.

The finding of rumen samples (Fig. IV.2A) indicates that the present method is useful also 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 sequence information is required in advance, the method may be useful for identification of pathogenic microorganisms that cause novel, unclear infectious diseases. It is worthy noting that viral sequences were accurately separated from prokaryotic sequences on the 5-kb DegeTetra-SOM (data not shown).

In the present study, I constructed SOMs of the genomes, for which almost complete sequences are available, to avoid complications caused by redundant genomic sequences in DDBJ/EMBL/GenBank. When considering phylogenetic classification of environmental microorgansms, it would be worthwhile to construct SOMs of all sequences from known species for extracting sequence characteristics in a wider rage of genomes. In the case of sequences from totally novel organisms, sequences even from related species are underrepresented in 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 node (i.e., the node with the minimum distance from the sequence data) in the multidimensional space.

IV-5: SUMMARY

I constructed SOMs of tetranucleotide frequencies in 1- and 5-kb sequences from prokaryotic genomes for which complete sequences are available. Sequences were clustered primarily according to species and to 11 major prokaryotic groups without any information regarding the species. SOM is a powerful tool for phylogenetic classification of genomic sequences and is thought to be most useful for classification of sequence fragments obtained from mixed genomes of uncultured environmental microorganisms. With this method, all non-rRNA sequences in DDBJ/EMBL/GenBank that were from unidentified or uncultured prokaryotes and longer than 1 kb were classified into 11 major prokaryotic groups. The result indicated that the method is also useful for survey of pathogenic microorganisms causing novel, unclear infectious diseases. SOM also classified 16S rRNA sequences accurately into phylogenetic groups.

Chapter V: Inter- and Intraspecies Characterizations of Eukaryotic Genome Sequences and *In Silico* Prediction of Genetic Signal Sequences

V-1: INTRODUCTION

Novel tools are needed for comprehensive comparisons of not only inter- but also intraspecies characteristics of massive amounts of increasingly available genomic sequences. In this chapter, I first constructed SOM with the frequencies of tri-, tetra-, and pentanucleotides in most (if not all) eukaryotic genomes for which almost complete sequence data are available. Then, I constructed SOM with oligonucleotide frequencies in 10-kb sequences from 2.8 Gb of human sequences and focused on oligonucleotides with frequencies characteristically biased from random occurrence in connection with possible *in silico* prediction of genetic signal sequences. The identification and analysis of a wide range of genetic functional signals (e.g., protein binding sites or gene regulatory sites such as promoters, ribosome-binding sites, and transcriptional initiating and terminating sites) are important. Common approaches to finding functional signals include the consensus sequence method, the weight matrix method, and the neural network method. In this chapter, I propose that SOM can provide a novel, systematic method to search for candidates for signal sequences.

V-2: RESULTS

SOMs for 13 eukaryotic genomes

To investigate the clustering power of SOM for a wide range of eukaryotic sequences, I first analyzed tri-, tetra-, and pentanucleotide frequencies in 300,000 non-overlapping 10-kb sequences and overlapping 100-kb sequences with a 10-kb sliding step derived from the 13 eukaryotic genomes (a total of 3 Gb) listed in Fig. V.1. The SOM adapted for genome informatics was constructed as described in Chapter II. First, frequencies for 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. After 80 learning cycles, oligonucleotide frequencies of the 10-kb sequences were represented by the final weight vectors arranged as a two-dimensional array, and the resulting SOM revealed clear species-specific separations. The sequences were clustered primarily into species-specific territories (Fig. V.1); nodes that contain sequences from a single species are indicated in color and those that contain sequences from more than one species are indicated in black. Comparison of the sequence classification in the 10-kb trinucleotide SOM (Tri-SOM, Fig. V.1A) with classification by the initial vectors (PCA in Fig. V.1A), which were set by the first and second principal components, revealed that sequences from one species were far more tightly clustered in the Tri-SOM. Species clustering appeared to increase in the tetra- and pentanucleotide SOMs (Tetra- and Penta-SOMs; Fig. V.1B and C). For example, 94%, 97%, and 98% of human sequences were classified into the human territories (🟥 in Fig. V.1) in the 10-kb Tri-, Tetra-, and Penta-SOMs, respectively.

When global characteristics of oligonucleotide frequencies in the genome are considered, distinction in frequencies between the complementary oligonucleotides (e.g. AAAC versus GTTT) may not be important. It was noted in Chapter IV that Tetra- and Penta-SOMs require long computation times. Therefore, SOMs were also constructed with frequencies of Degenerate sets in which frequencies of a pair of complimentary oligonucleotides were added (DegeTetra- and DegePenta-SOMs). This roughly halved the computation time without appreciable loss of clustering power (Fig. V.1B and C).

The G+C% obtained from the weight vector representative of each node in the Tri-SOM was reflected in the horizontal axis and increased from left to right (G+C% in Fig. V.1A); high G+C% sequences (red in the G+C% panel) were located on the right side of the map, and similar results were obtained for the Tetra- and Penta-SOMs (data not shown). In the 10-kb SOMs, intraspecies separations were evident. For example, human was divided into two major territories in the 10-kb Tri- and Tetra-SOMs. In the Penta-SOM (Fig. V.1C), however, human sequences were classified into a single territory, indicating that despite wide variations among human 10-kb sequences, the SOM recognized the common features in the pentanucleotide usages. In the 100-kb SOMs, interspecies separations were very prominent, and the species territories were surrounded by contiguous white nodes, which contained no genomic sequences. The vectors of the species-specific nodes even near a territory border were distinct between territories, and the species borders primarily could be drawn automatically on the basis of the contiguous white nodes as noted in Chapter III. The intraspecies separations were less evident in the 100-kb SOMs and all species had one major territory in the 100-kb Tetra- and Penta-SOMs (Fig. V.1B and C). When I inspected the 100-kb SOMs in detail, there were several minor territories composed of small numbers of sequences with specific characteristics. For example, a minor territory for *Arabidopsis* (🟩) located between the rice (🟩) and *Fugu* (🟫) territories was composed primarily of sequences from centromeric and subcentromeric regions. Analysis of intraspecies separations may provide fundamental information regarding structures of individual genomes.

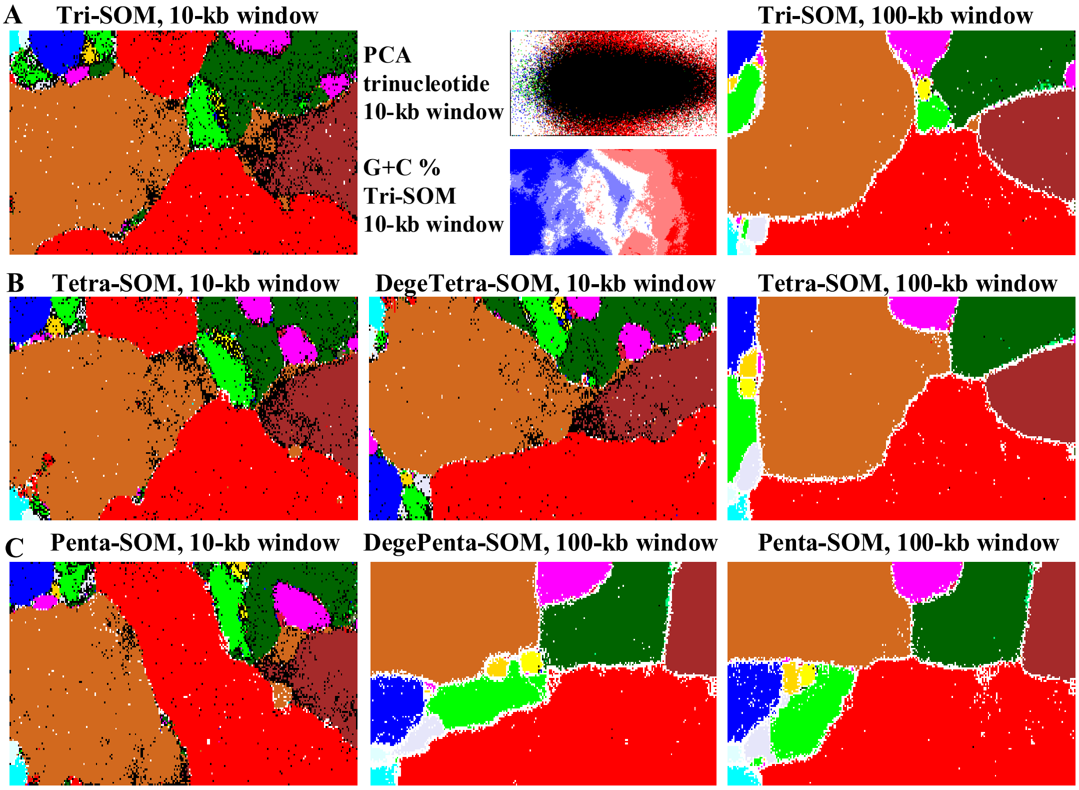


Figure V.1. SOMs for non-overlapping 10-kb and overlapping 100-kb sequences of 13 eukaryotic genomes. (A) Tri-SOMs. PCA, sequence classification by the initial weight vectors set by PCA for the 10-kb Tri-SOM. G+C% for each node in the 10-kb Tri-SOM was calculated and divided into five categories with an equal number of nodes. The nodes 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. (B) Tetra- and DegeTetra-SOMs. (C) Penta- and DegePenta-SOMs. Nodes that contain 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: *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* (🟥).

Diagnostic oligonucleotides for species separations

SOM recognized the species-specific combinations of oligonucleotide frequencies that is the representative signature of each genome, and therefore, allowed to identify the sequence patterns that are characteristic of individual genomes. The frequency of each oligonucleotide in each node in the 100-kb SOMs was calculated and normalized with the level expected from the mononucleotide composition in each node, and the observed/expected ratios thus normalized are illustrated in red (overrepresented), blue (underrepresented), or white (moderately represented) in Fig. V.2. This normalization allowed oligonucleotide frequencies in each node to be studied independently of mononucleotide compositions. For example, differences in the frequencies of CG- and GC-containing oligonucleotides can be detected sensitively irrespective of G+C% differences, just as the CG-deficiency in mammalian genomes was detected (Karlin et al. 1998, 2002; Gentles et al. 2001; Bird et al. 1985).

Transitions between red (overrepresentation) and blue (underrepresentation) for various tetra- and pentanucleotides often coincided exactly with species borders. Several diagnostic examples for the species separations are presented in Fig. V.2. 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 both plants. Results for a pair of complementary tetranucleotides were nearly identical, and therefore, only data for one tetranucleotide are presented. For all panels of individual tetranucleotides, see Supplementary Data V.1. In the case of pentanucleotides, examples for DegePenta-SOM are presented (Fig. V.2B). (ACAGG+CCTGT) and (CGACG+CGTCG) were over- and underrepresented, respectively, in all three vertebrates; and (CGAAA+TTTCG) was overrepresented in *Drosophila* and *C. elegans* and moderately represented in *S. pombe.* SOMs utilized a complex combination of many oligonucleotides for sequence separations, which results in classification according to species.

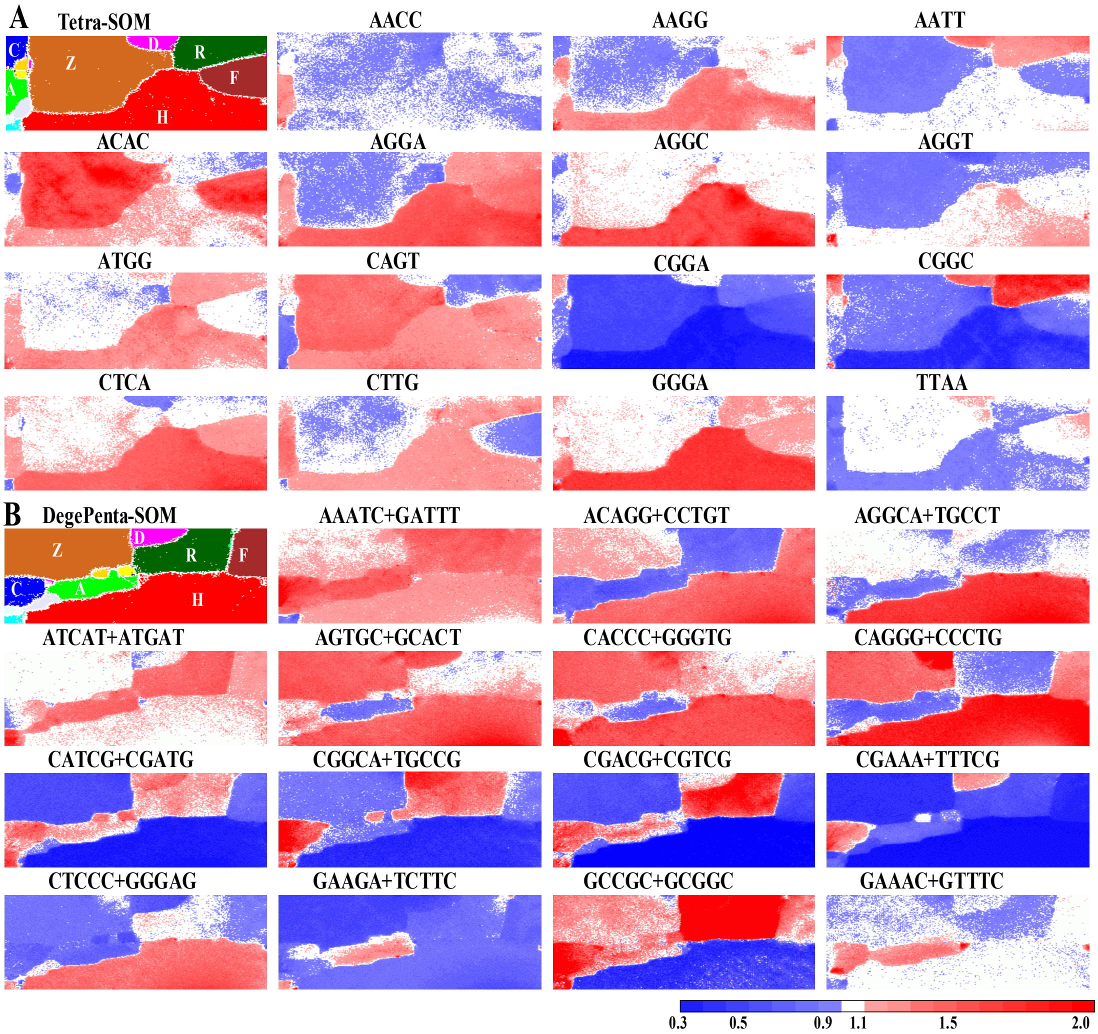


Figure V.2. Level of each tetranucleotide (A) and that of each pair of complimentary pentanucleotides (B) in 100-kb SOMs. Diagnostic examples of species separations are presented. Levels of individual tetranucleotides and of pentanucleotide pairs for each node in the 100-kb Tetra- and DegePenta-SOMs, respectively, of Fig. V.1 were calculated and normalized with the level expected from the mononucleotide composition of the node. The observed/expected ratio is indicated in colors at the bottom of the figure. The 100-kb SOMs in Fig. V.1B and C 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 Fig. V.1.

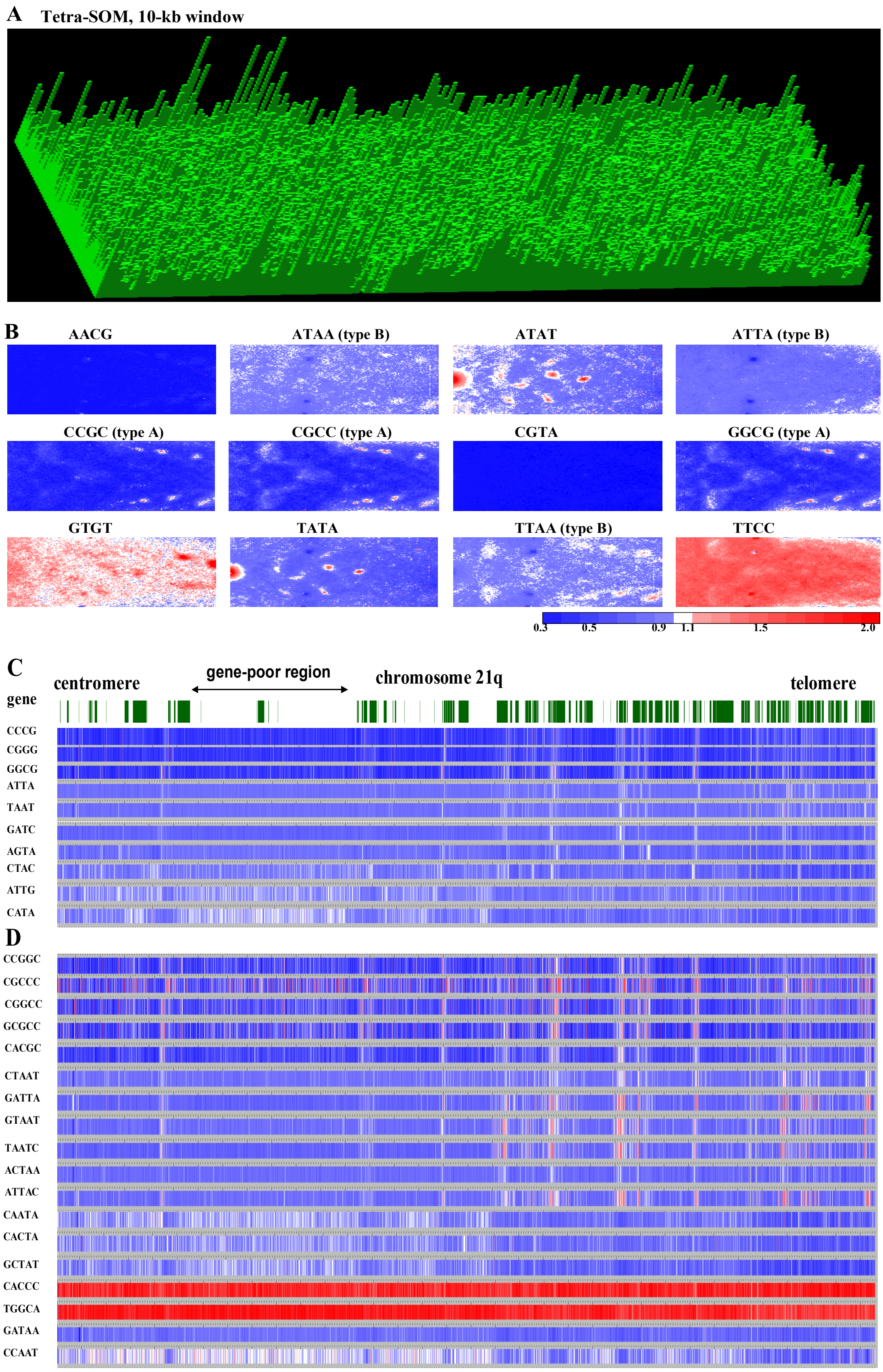
SOM with sequences from one genome and *in silico* prediction of genetic signals

To investigate the clustering power of SOM for intraspecies separations of sequences from one genome, human genomic sequences were analyzed. I constructed Tetra- and Penta-SOMs with nonoverlapping 10-kb and overlapping 100-kb sequences derived from 2.8 Gb of human sequence and present the 10-kb Tetra-SOM (Fig. V.3A). I calculated the levels of individual oligonucleotides in each node after normalization for the mononucleotide composition of each node. Several representative tetranucleotide patterns in the 10-kb SOM, including those underrepresented (blue) and overrepresented (red) across the entire zone (see AACG and TTCC, respectively), are presented in Fig. V.3B. For all patterns, refer to Supplementary Data V.2. In the present study, I focused on tetranucleotides that were represented significantly in restricted portions in the 10-kb SOM (red and white spots in the blue zone) but were underrepresented across the entire zone in the 100-kb SOM. One type of the examples corresponded to tetranucleotides containing one CG plus two C/G, for which similar patterns of local overrepresentation were observed (type A). These tetranucleotides corresponded to the constituent elements of GC-box, which is a well-characterized transcription signal (Philipsen et al. 1999). TTAA, ATAA, and ATTA, which are the constituent elements of TATA-box (Roeder et al. 1996), had patterns similar to those of type A tetranucleotides while the sequences were totally different (type B). Other characteristic patterns were observed, and some were similar. To investigate the biological significance of tetranucleotides with local overrepresentation patterns, the level of each tetranucleotide in the node, which represented the level of 10-kb sequences classified into this node, was plotted along the chromosome 21q sequence (Fig. V.3C). Distribution patterns of types A and B tetranucleotides and several others are presented; for all tetranucleotide panels, see Supplementary Data V.3. Types A and B had similar patterns, and significant representations (red and white) were observed in gene-rich regions. In various portions of the chromosome, distribution patterns of GATC and AGTA also resembled those of types A and B. All these tetranucleotides were represented at higher levels in gene-rich regions also in chromosomes 20 and 22 (data not shown), suggesting that these tetranucleotides are related to gene structure, function, and/or regulation. Karlin and colleagues (Karlin et al. 1998, 2002; Gentles et al. 2001) found that di-, tri-, and tetranucleotide frequencies in one genome were highly correlated, and therefore, dinucleotide frequencies capture major characteristics of the species-specific oligonucleotide frequencies. Because of this fundamental genome feature, I could sensitively select specific tetranucleotides that were underrepresented in most regions in the Tetra-SOM but whose constituent trinucleotides were rather overrepresented in the Tri-SOM. Because underrepresentation of the tetranucleotides was not due to that of constituent trinucleotides, the underrepresentation may reflect the biological significance of the tetranucleotide, as observed for the restriction site sequences in prokaryotic genomes (Karlin et al. 1997). Underrepresentation of ATTG could not be explained by levels of ATT and TTG and thus belonged to this type. The distribution of ATTG was concentrated in gene-poor regions (Fig. V.3C), suggesting that this sequence might not be related directly to gene function or structure.

Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). When genetic signal sequences are considered with respect to their occurrence patterns in the genome, they may be classified into different categories. In one category, the oligonucleotide sequence occurs across the genome at such a level as that predicted by the random occurrence based on the mononucleotide composition, and a combination with other sequences is a prerequisite for the sequence to function as a signal. In contrast, when an oligonucleotide sequence has a distinct activity, such as binding a target protein, its occurrence may be biased from the random level and vary significantly across the genome. For example, binding sequences for gene-regulatory proteins that belong to the latter category, may be underrepresented in comparison with random occurrence across most regions of the genome but would be more prevalent in gene-regulatory regions. In other words, such signals would be underrepresented across the entire zone of the SOM with a wide window (e.g., 100-kb SOM) but could occur at higher frequencies in restricted portions in the SOM with a much narrower window (e.g., 10-kb SOM). Furthermore, other transcription signals may also be overrepresented in these restricted portions because multiple transcription signals are usually clustered in gene-regulatory regions. The finding that various tetranucleotides including types A and B had characteristics predicted for the signals indicates that SOM can provide a tool for *in silico* prediction of genetic signal sequences. When signal sequences predicted with SOMs are compared with actual signals determined from experimental data, behaviors of signal sequences on SOMs may be cataloged. On the basis of the knowledge gleaned from the organisms well studied with various molecular techniques, it may be possible to develop an *in silico* method that is most useful for signal prediction in genomes that have been sequenced but for which there is little additional data. Because the number of such genomes has increased rapidly, development of this *in silico* method has become increasingly important.

Genetic signals, such as transcription signals, are typically longer than tetranucleotides. Therefore, analyses of longer oligonucleotides are needed to test the feasibility of the proposed method. It is also conceivable that detailed comparisons between the SOMs for different length oligonucleotides could elucidate the sequence length of the actual, functional signals. I thus examined underrepresented pentanucleotides in the human genome in the same manner as tetranucleotides. Examples where distribution patterns were distinct between the gene-rich and -poor regions are presented in Fig. V.3D. The local, specific representation patterns were often clearer than those of tetranucleotides, suggesting that many tetranucleotides of interest are constituent elements of signals with longer lengths. If the window sizes of SOMs are narrowed to less than 10 kb, different categories of signal sequences may be detected. Comparisons among the SOMs with different length oligonucleotides and different size windows may reveal various features of genetic signal sequences in the genome.

As a preliminary application of this method, I analyzed the zebrafish genome, which has been sequenced recently (http://www.sanger.ac.uk/Projects/D_rerio/). The 10- and 100-kb DegeTetra- and DegePenta-SOMs were constructed, and degenerate pentanucleotides that were underrepresented across most portions in the 100-kb SOM but were represented significantly in restricted portions in the 10-kb SOM were focused on. Representative patterns of local, specific representations in the 10-kb DegePenta SOM are presented in Fig. V.3E. These includes pentanucleotides whose constituent tetranucleotides were rather overrepresented across most regions in the 10-kb DegeTetra-SOM (type A) or did not show local, specific representations in the DegeTetra-SOM (type B).


A
Tetra-SOM, 10-kb window
B
AACG
ATAA (type B)
ATAT
ATTA (type B)
CCGC (type A)
CGCC (type A)
CGTA
GGCG (type A)
GTGT
TATA
TTAA (type B)
TTCC
0.3
0.5
0.9
1.1
1.5
2.0
C
centromere
gene-poor region
chromosome 21q
telomere
gene
CCCG
CGGG
GGCG
ATTA
TAAT
GATC
AGTA
CTAC
ATTG
CATA
D
CCGGC
CGCCC
CGGCC
GCGCC
CACGC
CTAAT
GATTA
GTAAT
TAATC
ACTAA
ATTAC
CAATA
CACTA
GCTAT
CACCC
TGGCA
GATAA
CCAAT

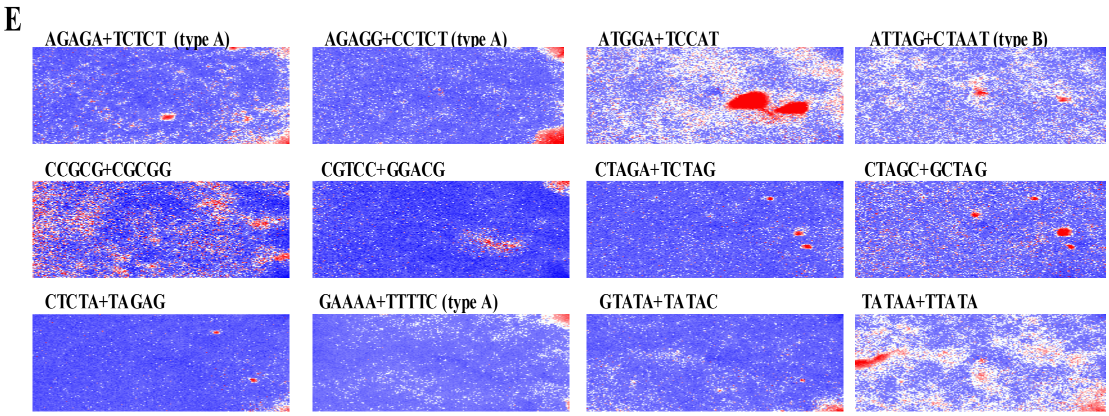


Figure V.3. SOMs with human genomic sequences. (A) Three-dimensional presentation of the 10-kb Tetra-SOM for human sequences. The number of sequences classified into each node is indicated by the height of the bar. (B) Characteristic examples of tetranucleotide levels in the 10-kb SOM are presented. The level of each tetranucleotide was calculated and presented as described in Fig. V.2. (C and D) Tetra- and pentanucleotide levels, respectively, are plotted from the centromere to the telomere along human chromosome 21q in blue, white, and red. Locations of genes are indicated in green at the top of the panel, and the 7-Mb gene-poor region is noted with a bidirectional arrow. (E) Representative examples of pentanucleotide levels in the 10-kb DegePenta-SOM for the zebrafish genome are presented.

V-3: DISCUSSION

Wide varieties of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). My finding that various tetranucleotides (e.g., types A and B) have characteristics consistent with those of transcription signals indicates the possibility that SOM may be a novel tool for characterization and *in silico* prediction of genetic signal sequences. Genetic signals, such as transcription signals, are typically longer than tetranucleotides, and therefore, analyses of longer oligonucleotides are needed to test this possibility. I examined underrepresented pentanucleotides in the human genome in the same manner as described for tetranucleotides. Examples where distribution patterns were distinct between the gene-rich and -poor regions are presented in Fig. V.3D along with the four signal pentanucleotides described later. The local, specific representation patterns of pentanucleotides were often clearer than those of tetranucleotides, suggesting that many tetranucleotides of interest are constituent elements of signal sequences with longer lengths (e.g., GC-box).

Recognition mechanisms of genetic signal sequences and occurrence levels of the respective sequences across the genome are thought to be related. When an oligonucleotide sequence has a distinct activity, such as high-affinity binding to a specific target protein, the occurrence level 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; such sequences would be underrepresented across the entire zone of the SOM with a wide window (e.g., 100-kb) but would occur at higher frequencies in restricted portions of the SOM with a much narrower window (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. Information regarding the frequencies of the oligonucleotides with transcription factor binding activities may provide insight into the mutual role of oligonucleotides that comprise combinatorial units for transcriptional regulation (e.g., differential contribution in specificity determination). Collectively, SOM data concerning levels of oligonucleotides with factor binding activities will enable me to categorize and visualize distinct behaviors of a variety of genetic signal sequences on SOMs. By referring to the behaviors of signal sequences determined for well-studied organisms, I 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.

As a preliminary step toward developing such an *in silico* approach, I characterized pentanucleotide sequences known to have transcription factor binding activities. I screened the TRANSFAC database (http://transfac.gbf.de/TRANSFAC) for pentanucleotides that are considered binding sequences for mammalian transcription factors. In the database, there are 22 pentanucleotides that are reported to be factor binding sequences by literatures. However, when I checked these sequences in detail by referring to the MATRIX table in the database, most were constituents of much longer signal sequences. Four pentanucleotides were selected as main determinant sequences for transcription factor binding: NF-Y binding site CCAAT (Mantovani et al. 1998), GATA-1 factor binding site GATAA (Evans et al. 1988), KLF binding site CACCC (Philipsen et al. 1999), and NF-1 binding site TGGCA (Borgmeyer et al. 1984; Nowock et al. 1985). In Fig. V.4A and B, I show the distribution patterns of these four pentanucleotides, together with a typical pattern of the GC-box core element (CGCCC), in the 10- and 100-kb Penta-SOMs, respectively, for the 2.8 Gb of human sequences. GATAA was underrepresented in most regions of both 10- and 100-kb SOMs. CCAAT was underrepresented in most regions of the 100-kb SOM but was represented at higher levels in restricted regions of the 10-kb SOM. Detailed comparison of the distribution of CCAAT with that of the core element of the GC-box showed that zones with a high frequency of CCAAT were distinct from those of the GC-box sequence in the 10-kb SOM (Fig. V.4A). The distribution on chromosome 21q (the lowest panel in Fig. V.3D) showed that the CCAAT pentanucleotide was more abundant in gene-poor regions than gene-rich regions; in Fig. V.3D, distribution patterns of four signal pentanucleotides are presented. Interestingly, CCAAT appeared to be most dense in the 7-Mb of the gene-poorest region found by Hattori et al. (Hattori et al. 2000). This pentanucleotide is recognized with high affinity and specificity by the heterotrimeric transcription factor NF-Y with two histonic subunits NF-YB and NF-YC, which resemble histones H2A and H2B (Ronchi et al. 1995). NF-Y can bind DNA at different steps during nucleosome formation and bend DNA strands. Considering these functions of NF-Y and its high affinity for CCAAT, the clustering in gene-poor regions might be related to specific chromatin structures in these regions. CACCC and TGGCA were overrepresented across most regions of the 10- and 100-kb SOMs (Fig. V.4A and B) and on chromosome 21q (Fig. V.3D). Such abundant sequences may require additional, specific sequences for proper regulation of gene expression and/or may correspond to binding sites for ubiquitous DNA-binding factors. In the case of TGGCA, the functional signal sequence for transcriptional regulation is a pair of two complimentary pentanucleotides closely situated (Borgmeyer et al. 1984; Nowock et al. 1985). This could explain, at least in part, the overrepresentation of this pentanucleotide. Systematic categorization of the frequencies of known signal sequences across a genome is fundamental information that is valuable for understanding the molecular mechanisms that underlie proper signal recognition. When characteristic oligonucleotides, both underrepresented and overrepresented in the genome, are considered, various factors, including DNA conformational tendencies and context-dependent mutation and modification of DNA, are thought to be responsible (Karlin et al. 1998;Rocha et al. 1998; Gentles et al. 2001; Pride et al. 2003).

SOM can illustrate oligonucleotide frequencies for a wide variety of genomes on a single map. In Fig. V.4C and D, I present the occurrence pattern in the 10- and 100-kb Penta-SOMs for 13 eukaryotes (Fig. V.1C) of each of four mammalian signal sequences aforementioned. GATAA was underrepresented in all these eukaryotes except *Plasmodium* and *Dictyostelium*; CCAAT was underrepresented in most genomic regions of the three vertebrates but overrepresented in the two plants and the two invertebrates. These findings may provide fundamental information regarding the mechanisms of signal recognition in individual species and the evolutionary processes that established the signal recognition system.

SOM can identify oligonucleotides with frequencies that are biased from randomness on a two-dimensional map. Furthermore, because the genomic sequences with the specific characteristics were self-organized on the map, the genomic locations of such sequences could be easily plotted along the chromosomes (Fig. V.3C and D). When known signal sequences like transcription factor binding activities of various species with enough experimental data are characterized and categorized systematically with SOMs, I 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.

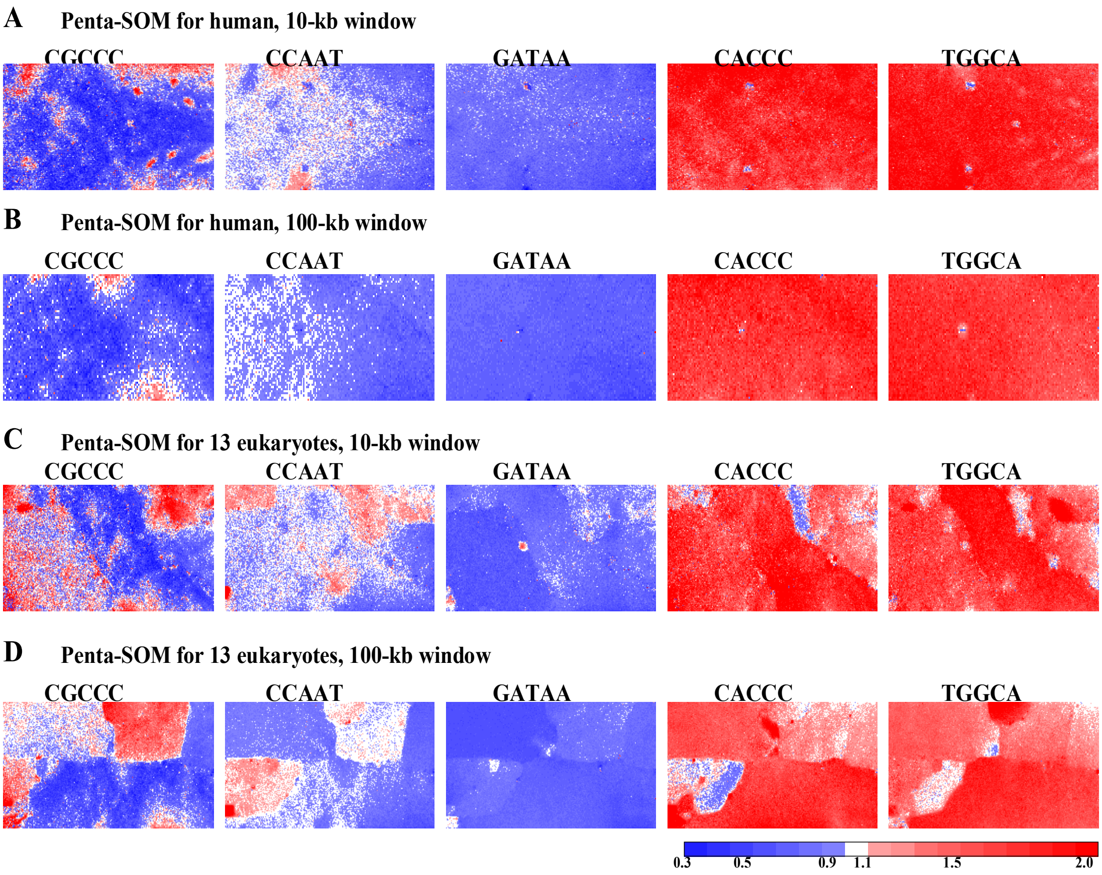


Figure V.4. Characterization of genetic signal sequences. (A and B) Pentanucleotide levels in the 10- and 100-kb SOMs for 2.8 Gb of human sequences, respectively. (C and D) Pentanucleotide levels in the 10- and 100-kb SOMs for the 13 eukaryotes presented in Fig. V.1C, respectively. The level of each pentanucleotide was calculated and presented as described in Fig. V.2.

V-4: SUMMARY

I generated SOMs for tri-, tetra-, and pentanucleotide frequencies in 300,000 10-kb sequences from 13 eukaryotes for which almost complete genomic sequences are available (a total of 3 Gb). SOM recognized in most 10-kb sequences species-specific characteristics (key combinations of oligonucleotide frequencies), permitting species-specific classification of sequences without any information regarding the species. Because the classification power is very high, SOM is an efficient and powerful tool for extracting a wide range of genomic information. SOM constructed with oligonucleotide frequencies in 10-kb sequences from 2.8 Gb of human sequences identified oligonucleotides with frequencies characteristically biased from random occurrence predicted from the mononucleotide composition, and 10-kb sequences rich in these oligonucleotides were self-organized on a map. Because these oligonucleotides often corresponded to genetic signals or the constituent elements, I propose an *in silico* method that should be useful for identification of genetic signal sequences in genomes for which large amounts of sequence data like transcription factor binding activities are available but additional experimental data are lacking.

Chapter VI: *In Silico* Classification of Gene and Genomic Sequences of Human and Mouse according to Functional Categories

VI-1: INTRODUCTION

In the era of extensive genome sequencing, it is important to predict numerous functions of gene and genomic sequences utilizing increasingly available 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) (Huttenhofer et al. 2001; Marker et al. 2002). 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 (Okazaki et al. 2002). 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 (Eddy 2001; Mattick 2002). 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 (Curtis et al. 1995; Decker and Parker 1995; Chen and Shyu 1995; Mazumder et al. 2003).

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. In this chapter, I constructed SOMs with oligonucleotide frequencies in 37,086 full-length mouse cDNA sequences whose expression was confirmed in separate experimental approaches (Okazaki et al. 2002), and found separation between protein-coding and noncoding cDNAs on the SOMs. I also analyzed 5’ and 3’ UTR sequences compiled in UTRdb (Graziano et al. 1998, 2002) and found that these sequences tended to cluster according to the functional categories. Additionally, I constructed SOMs of oligonucleotide frequencies in 1-kb genomic sequences from mouse and human genomes and found that SOM could detect differential sequence characteristics of 5’ and 3’ UTRs, CDSs, introns, and ncRNAs.

VI-2: RESULTS

SOMs for full-length mouse cDNA sequences

I constructed SOMs with tri-, tetra-, and pentanucleotide frequencies in the 37,086 full-length mouse cDNA sequences (Tetra- and Penta-SOMs are shown in Fig. VI.1). 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 in all SOMs constructed. In the right-hand side, there was one broad satellite zone of ncRNAs (Fig. VI.1), where ncRNAs with CpG island and/or with antisense activity were enriched. Detailed investigation of the number of cDNAs assigned to each node showed that evident clustering of protein-coding cDNAs in multiple nodes. To show this graphically, the number of sequences classified into each node was represented by the height of the vertical rod; in the nodes containing sequences of two categories, the number of the major and minor category in one node is represented by the height of the upper and lower part of the vertical rod, respectively, using a color specifying the category (3D in Fig. VI.1). 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. 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 artifacts caused by redundant sequences of UTRs, we used mouse UTR sequences compiled in UTRdb, 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 Tri-, Tetra-, and Penta-SOM (Penta-SOM is presented in Fig. VI.2). 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 island were enriched. When UTRs and ncRNAs with antisense activity were focused on, these tended to be located away from the respective major territory and often to be closely associated with CDS territories. 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. VI.1 was not due to a simple reflection of codon usage patterns in CDSs. The finding that territories of one functional category were split into several zones will be discussed below.

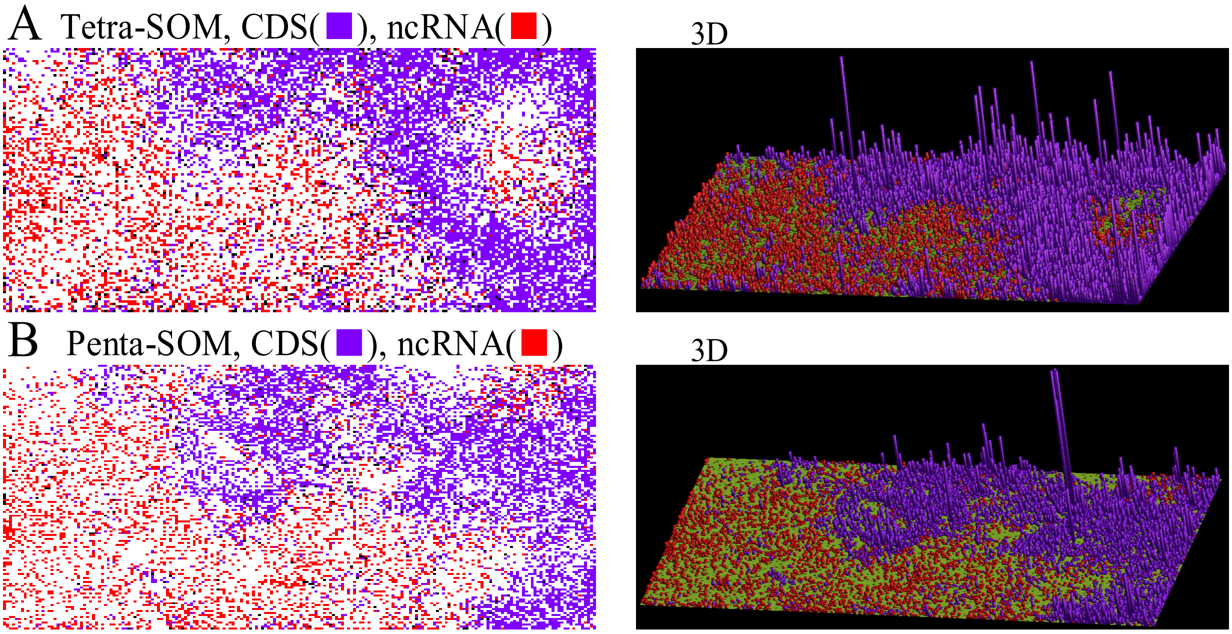


Fig. VI.1 SOM with mouse cDNA sequences. (A and B) Tetra- and Penta-SOMs, respectively. Nodes that contain only the sequences for protein-coding cDNAs are marked in violet, those containing only the sequences for ncRNAs are marked in red, and those containing the sequences of both categories are marked in black. 3D: three-dimensional presentation of the SOMs.

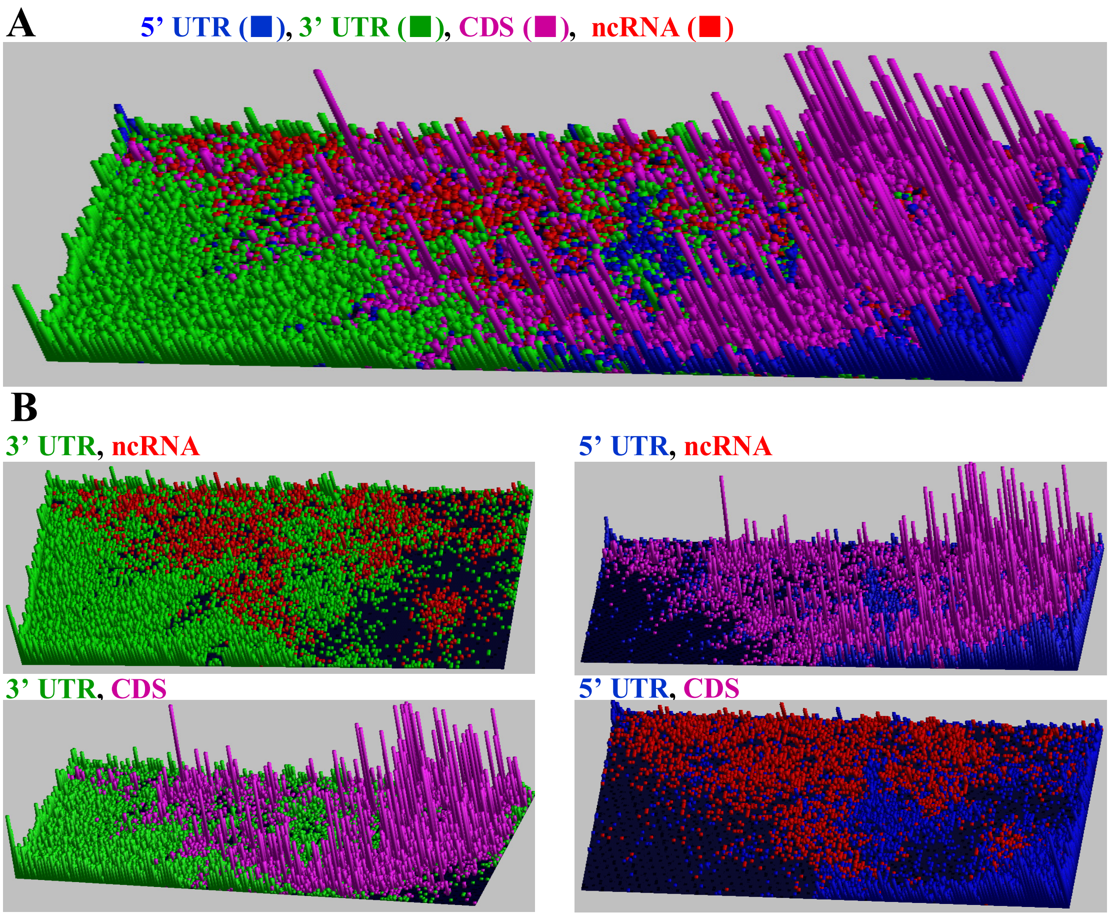


Fig. VI.2 SOM with mouse cDNA and UTRs sequences. (A) Three-dimensional presentation of four functional categories (5’ and 3’ UTRs, CDSs, and ncRNAs) on Penta-SOM. Nodes that contain only the sequences for 3’ and 5’ UTR are marked in green and blue, respectively, and those containing only the sequences 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.

SOMs for mouse and human genomic sequences and mapping of UTR, CDS and intron sequences on the SOMs

In the SOM analyses of cDNA sequences, it is impossible to compare sequence characteristics of UTRs, CDSs, and ncRNAs with those of introns and franking sequences. Therefore, I constructed Tri-, DegeTri-, and DegeTetra-SOMs with 1-kb sequences derived from 2.3 Gb mouse genomic sequences (DegeTetra-SOM is shown in Fig. VI.3A). Then, five functional categories of sequences (5’ and 3’ UTRs, CDSs, ncRNAs, and introns) were mapped on the 1-kb SOMs (Fig. VI.3B) in order to compare and clarify characteristics of oligonucleotide frequencies in these functional sequences. Sizes of many exons and introns were far shorter than 1 kb. In order to get statistically meaningful results, contiguous exons or introns of each functional sequence (e.g., each CDS) were concatenated and followed by segmentation into 1-kb sequences, and oligonucleotide frequencies in these sequences were mapped on the SOMs constructed with 1-kb mouse genomic sequences. While introns distributed rather dispersedly on the SOM, punctuated and differential distributions of 5’ and 3’ UTRs, CDSs, and ncRNAs were observed. This revealed distinct characteristics in oligonucleotide frequency between the functional categories, showing that SOM could detect the sequence characteristics specific to the functional regions. Importantly, the territory of each functional category was divided into multiple zones. It is also worthy of noting that there was a characteristic zone in the right and lower part of the DegeTetra-SOM, where sequences of 5’ and 3’ UTRs and CDSs were closely associated. This may reflect, at least in part, their closely related functions such as antisense activity; the SOM was constructed with the frequencies of the degenerate set of complimentary tetranucleotides, and therefore, two complimentary sequences tend to locate close to each other.

To examine generality of the finding obtained with mouse sequences to other mammalian genomes, Tri-, DegeTri-, and DegeTetra-SOM were constructed with 1-kb sequences from 2.8 Gb human genomic sequences and sequences of four functional categories (5’ and 3’ UTRs, CDSs, and introns in protein-coding genes) were mapped on the SOM (Fig. VI.4B,C). Confirming the results of mouse sequences, inter- and intra-category separations were observed. Introns again distributed more dispersedly than UTRs and CDSs. In the analysis of human sequences, ncRNAs were not included because systematically curated compilation of human ncRNAs was reported in no public databases. However, in the case of the human genome, a curated dataset of the 5’ franking sequences of a wide rage of genes can be obtained from the TRANSFAC database. I selected 2-kb sequence upstream of the transcriptional start site of each gene and segmented into 1-kb sequences, followed by mapping of the two sets of the segmented 1-kb sequences on the 1-kb DegeTetra-SOM (Fig. VI.4C). Because the two sets gave similar results, the collective 1-kb sequences were treated as the 5' franking sequences. The location of the 5' franking sequences was most restrictive and many sequences were overlapped partly with a portion of the 5' UTR territory (Fig. VI.4C). Close association of the 5' franking sequences with a portion of the 5' UTRs may relate, at least in part, with multiple start sites of transcription, as well as the presence of CpG islands.

The 5' and 3' UTRs of eukaryotic mRNAs play a crucial role in post-transcriptional regulation of gene expression by modulating mRNA transport, translation efficiency, subcellular localization, and stability. The separations in either 5' or 3' UTRs on SOMs may correspond, at least in part, to these differential functions. As a preliminary attempt to know correlation of the intra-category separation with functional differences, functions of human genes corresponding to the highest and the second highest peaks for 3' UTRs, which were located in the left-hand and right-hand sides in the top part of the human DegeTetra-SOM (Fig. VI.4B), respectively, were investigated referring to Ensembl (http://www.ensembl.org/). A major portion of the genes whose 3' UTR was mapped in the left-side peak was found to be genes for nucleic-acid binding proteins, and that of the genes whose 3' UTR were mapped in the right-side peak corresponded to genes for catalytic enzymes. It is also worthy of noting that there were UTR zones closely associated with CDS zones and the UTRs in these zones were found often to be sequences with antisense activity involved in post-transcriptional regulation.

To construct the SOM with mouse and human genomic sequences (Figs. VI.3 and 4), almost all available sequences were used, and therefore, regions poorly-characterized with molecular techniques other than sequencing were included in the analysis. Because no information other than oligonucleotide frequencies is required for the SOM formation, I propose that SOM is a novel in silico method that is useful for prediction of functional sequences in the genome regions with little characterization with experiments. Sequence localized in a zone where sequences of known functional categories are clustered should have sequence characteristics similar to those of the functional sequences, suggesting that the sequences also have the respective function. Furthermore, a diagnostic combination of oligonucleotide frequencies responsible for individual functions should be clarified.

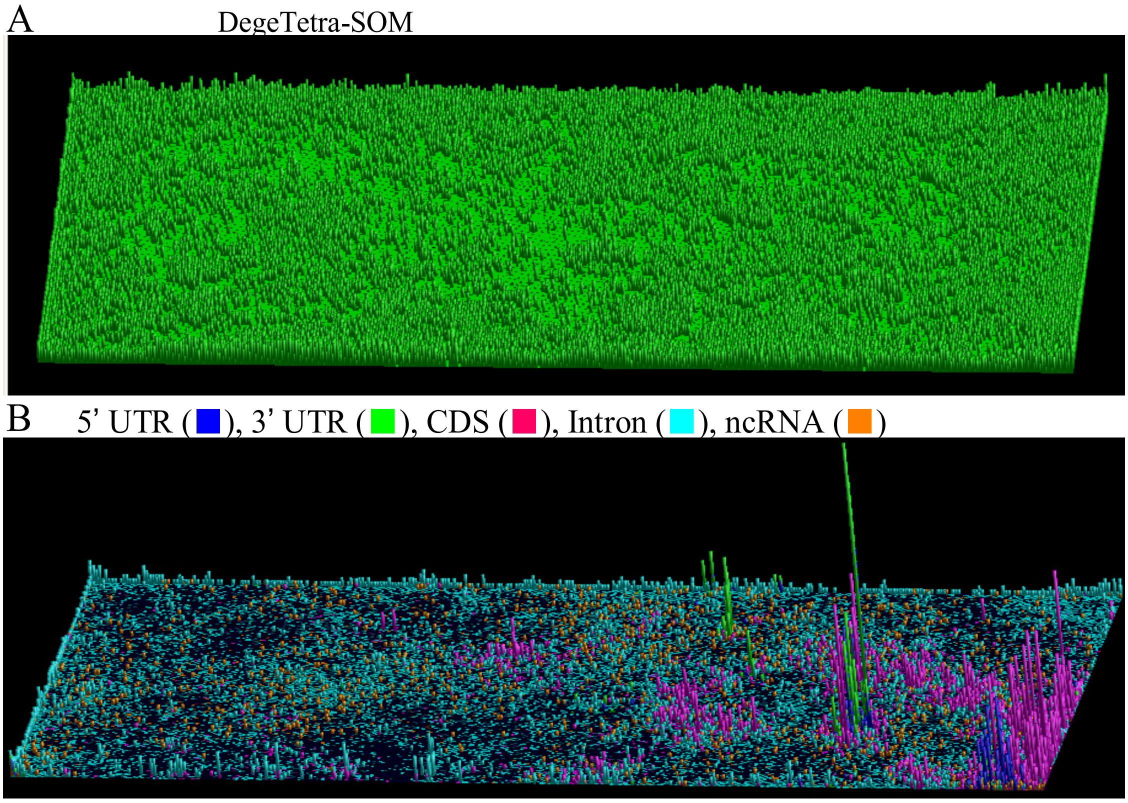


Figure VI.3 SOMs with mouse genomic sequences. (A) Three-dimensional presentation of the 1-kb DegeTetra-SOM for mouse sequences. The square root of a number of sequences classified into each node is indicated by the height of the vertical rod. (B) Mapping of sequences of 5’ and 3’ UTRs, CDSs, and introns on the 1-kb DegeTetra-SOMs. To avoid potential artifacts caused by redundant sequences, I used mouse UTRs compiled in UTRdb (http://bighost.area.ba.cnr.it/BIG/UTRHome/), which is a specialized database of eukaryotic 5’ and 3’ UTRs that has been cleaned of redundant sequences. To get statistically meaningful results, sequences shorter than 100 nucleotides were excluded from this analysis. Detailed inspection of 3’ UTR data in UTRdb revealed that a small number of sequences have poly-A tails of varying lengths while the sequences in UTRdb should not contain poly-A tail sequences. Such sequences were excluded from the analysis. Sizes of many exons and introns are far smaller than 1 kb. The purpose of the present analysis is to clarify and compare sequence characteristics of 5’ and 3’ UTRs, CDSs, and introns. Taking this into account, the contiguous exons, as well as of introns, in each gene were concatenated and then segmented into 1-kb sequences; for the last segment

less than 1 kb, the 1-kb sequence from the 3’ end is includes in the analysis instead of the last segment less than 1 kb. Tetranucleotide frequency in each sequence was calculated and followed by mapping to the node of the 1-kb DegeTetra-SOM with the shortest distance in the multidimensional frequency space.

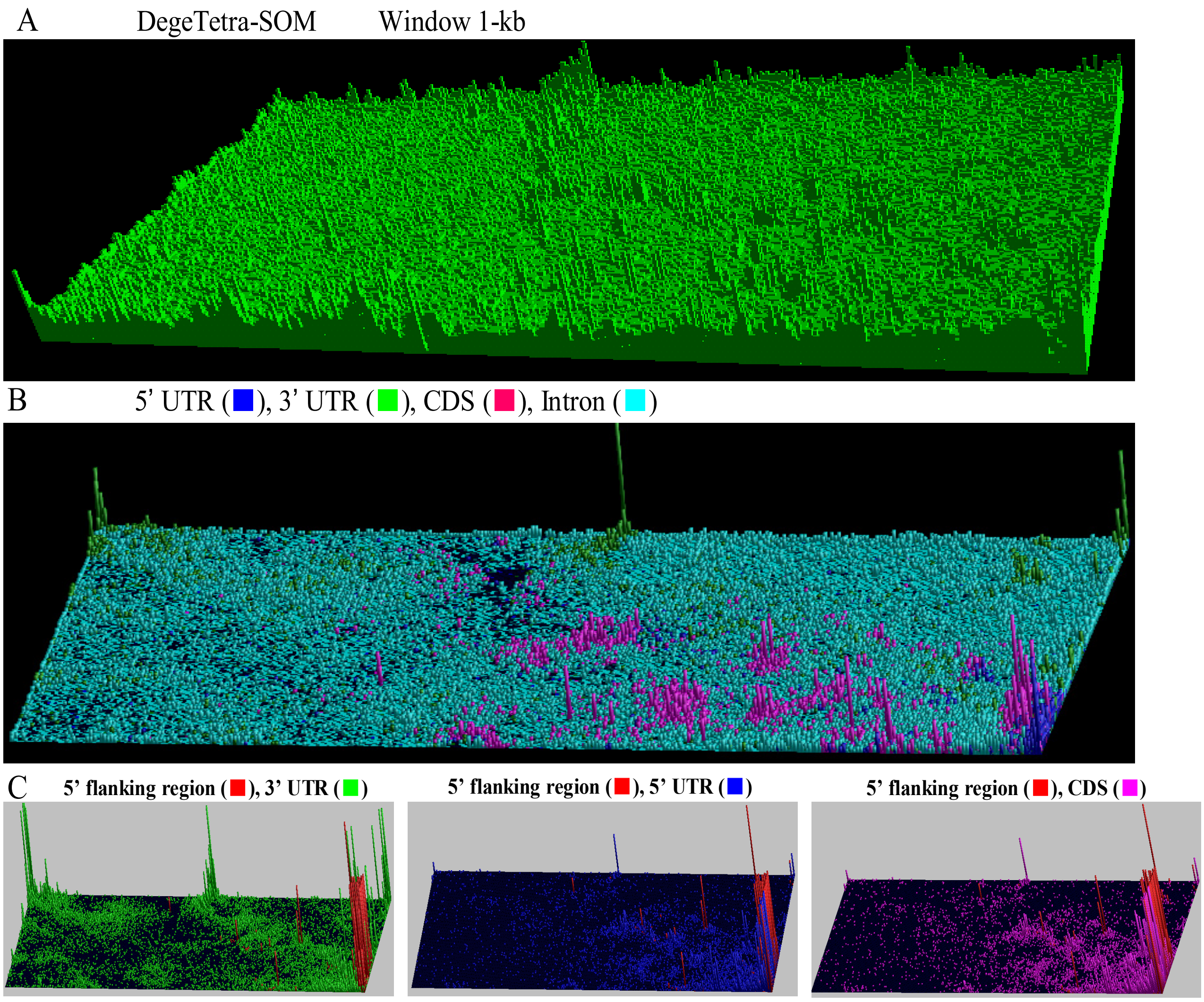


Figure VI.4 SOMs with human genomic sequences. (A) Three-dimensional presentation of the 1-kb DegeTetra-SOM for human sequences. The square root of a number of sequences classified into each node is indicated by the height of the bar. (B) Mapping of sequences of 5’ and 3’ UTRs, CDSs, and introns on the 1-kb DegeTetra-SOMs. Four functional categories used as described as Fig. VI.3. (C) Comparison of sequence location between two functional categories (5’ flanking region and 3’ UTR; 5’ flanking region and 5’ UTR; 5’ flanking region and CDS; 5’ UTR). Nodes that contain only the sequences of one category are marked in the color representing the category.

VI-3: DISCUSSION

One important area of the applications of neural networks for nucleotide sequence analysis is for gene identification. The gene identification with neural networks is studied by two complementary approaches: gene search by content and gene search by signal (Staden 1990; Fickett 1996). The searches by content use various coding measures to determine the protein-coding potential. The searches by signal methods identify signal sequences such as splice sites. Various neural networks provide useful models in which sequence features for both signals and content can be combined and weighted to improve accuracy (Uberbacher et al. 1996; Snyder and Stormo 1995). Other applications of neural networks are sequence classification and feature detection. The detection of significant sequence features and understanding of biological rules governing gene structure and regulation are important problems that can be addressed. In this chapter, I showed that characteristic features in individual functional regions such as 5' and 3' UTRs in mammalian genomes could be effectively extracted by SOM. Either the 5' or 3' UTR sequences were divided into multiple zones in SOMs and there were nodes enriched with the 3' UTR sequences of the genes for a particular functional category of proteins. When information regarding differential functions of UTRs has accumulated for many sequences, the present method may be useful as an *in silico* method to predict the functions of individual UTRs. Inclusion of my finding into the known gene identification tool may improve the identification power and provide additional information regarding sequence characteristics of UTRs, which are involved in the post-transcriptional regulation.

VI-4: SUMMARY

In this chapter, I showed that SOM is an effective tool for comparing and identifying sequence characteristics of differentiated functional regions in the genome. I first constructed SOMs with oligonucleotide frequencies in mouse full-length cDNA and found separation between protein-coding and non-coding cDNAs. Next, I constructed SOMs with oligonucleotide frequencies in sequences of four functional categories (5' and 3' UTRs, CDSs, and ncRNAs), and found that these sequences tended to cluster according to the functional categories. This showed that SOM detected distinct characteristics in these functional sequences and that the separation between protein-coding and non-coding cDNAs was not a simple reflection of characteristics of codon usage pattern in CDSs. I then constructed SOMs with oligonucleotide frequencies in

1-kb genomic sequences from mouse and human genomes followed by mapping of sequences of 5’- and 3’-UTRs, CDSs, and introns on the SOMs. Sequences of these distinct categories tended to cluster according to the functional categories, confirming that SOM detected sequence characteristics in the distinct functional regions in the genome. Therefore, a combination of diagnostic oligonucleotides responsible for individual functions should be clarified.

Chapter VII: Future Prospect: Application of SOM to Protein Sequence Analyses

VII-1: Prediction of protein function

Of importance in genome analysis is prediction of the function of proteins that are identified through genome sequencing but lack significant sequence homology with function-known proteins, which are left as function-unknown proteins. For example, in metagenome analyses (Chapter IV) of environmental microorganisms, prediction of the functions of novel proteins is essential for identifying industrially useful genes. A uniquely important problem for protein sequence analysis is prediction of conformations, which are directly related with functions. Approximately 10 years ago, clustering of proteins according to tertiary structure families was studied with both unsupervised Kohonen's self-organizing classifiers (Ferran et al. 1994) and supervised BP classifiers (Wu et al. 1995). Ferran et al. (1994) analyzed dipeptide frequencies in proteins with the conventional SOM and reported that clustering by tertiary structure and function is possible. However, this methodology was seldom applied to prediction of protein functions because the studies were performed prior to large-scale genome sequencing and, therefore, there were not many proteins with unknown functions. Furthermore, the computation times were long and biologists were not familiar with neural networks.

I have attempted to develop a system to cluster proteins by structure and function on the basis of the modified SOM, focusing on di- and tripeptide frequencies in protein sequences compiled in databases. SOM should cluster proteins with similar oligopeptide distributions, permitting to predict functions even of proteins that lack significant homology to function-known proteins detectable with conventional sequence homology searches. I analyzed the 400- and 8000-dimensional vectors representing di- and tripeptide frequencies, respectively, in 50,000 protein sequences from a wide rage of prokaryotes. I found that SOM clustering was dependent not only on functions but also on species and that a major source of species-dependent separation was related to the amino acid frequencies that are characteristic of individual species (for example, between the microbes in extreme environment etc.). As a preliminary task, I have been developing a method to eliminate the influence of species-specific amino acid frequencies. As an extension of the present study, I plan to establish SOM as a method to characterize proteins whose functions are currently unknown. With respect to the function-unknown proteins, I may make predictions for their functions on the basis of the functions of known proteins that are located similarly on SOMs. This technology may become a useful bioinformatics strategy for predicting functions of proteins without significant sequence homology to function-known proteins detectable with conventional sequence homology searches.

Because the classification power of SOM is very high for a large amount of complex data, this unsupervised algorithm can be established as the wide applicable, fundamental bioinformatics for effectively extracting a wide range of knowledge from a massive amount of genome and protein sequences, importantly without prior knowledge of the sequences. SOM will provide fundamental knowledge for understanding molecular processes and mechanisms that have established sequence characteristics of individual genomes during evolution.

References

Abdurashidova, G., Deganuto, M., Klima, R., Riva, S., Biamonti, G., Giacca, M., and Falaschi, A. 2000. Start sites of bidirectional DNA synthesis at the human lamin B2 origin. *Science* **287,** 2023-2026.

Abe, T., Kanaya, S., Kinouchi, M., Kudo, Y., Mori, H., Matsuda, H., Carlos, D. C., and Ikemura T. 1999. Gene classification method based on batch-learning SOM. *Genome Inform. Ser.* **10**, 314-315.

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

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

Abe, T., Kozuki, T., Kosaka, Y., Fukushima, A., Nakagawa, S., Ikemura, T., 2003. Self-organizing map reveals sequence characteristics of 90 prokaryotic and eukaryotic genomes on a single map. *Workshop 2003 on Self-Organizing Maps* 95-100.

Abremski, K., Sirotkin, K., and Lapedes, A. 1993. Application of neural networks and information theory to the identification of' E. *coli* transcriptional promoters. *Math. Model. Sci. Computing* **2**, 636-641.

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

Andersson, S.G., and Sharp, P.M. 1996. Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiol.* **142,** 915-925.

Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A., and Damiani, G. 1991. Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *CABIOS.* **7**, 353-357.

Bahary, N., and Zon, L.I. 1998. Use of the zebrafish (Danio rerio) to define hematopoiesis. *Stem Cells* **16**, 89-98.

Bartlett, E.B. 1994. Dynamic node architecture learning: an information theoretic approach. *Neural Netw.* **7**, 129-140.

Baum, E.B., and Haussler, D. 1989. What size net gives valid generalizations. *Neural Comput.* **1**, 151-160.

Bernardi, G. 1989. The isochore organization of the human genome. *Annu. Rev. Genet.* **23,** 637-661.

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

Bernstein, E., Caudy, A.A., Hammond, SM., and Hannon, GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363-366.

Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. 1995. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91-99.

Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature*. **321**, 209-213.

Bisant, D., and Maizel, J. 1995. Identification of' ribosome binding sites in *Escherichia coli* using neural network models. *Nucleic Acids Res.* **23,** 1632-1639.

Blattner, F.R., Plunkett, G.3rd., Bloch C.A., and Perna, N.T. et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277,** 1453-1474.

Blom, N., Hansen, J., Blaas, D., and Brunak, S. 1996. Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci.* **5**, 2203-2216.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M., and Lautrup, B.,. Norskov, L., Olsen, O.H., and Petersen, S.B. 1988. Protein secondary structure and homology by neural networks. *FEBS Letters* **241**, 223-228.

Bohr, H., Bohr, J., Brunak, S., Cottenll, R.M., and Fredholm, H., Norskov, L., Olsen, O.H., and Petersen, S.B. 1990. A novel approach to prediction of' the 3-dimensional structures of protein backbones by neural networks. *FEBS Letters* **261**, 43-46.

Bolshakov, V.N., Topalis, P., Blass, C., Kokoza, E. della., Torre, A., Kafatos, F.C., and Louis, C. 2002. A comparative genomic analysis of two distant diptera, the fruit fly, Drosophila melanogaster, and the malaria mosquito, Anopheles gambiae. *Genome Res.* **12,** 57-66.

Borgmeyer, U., Nowock, J., and Sippel, A.E. 1984. The TGGCA-binding protein: a eukaryotic nuclear protein recognizing a symmetrical sequence on double-stranded linear DNA. *Nucleic Acids Res*. **12**, 4295-4311.

Brunak, S., Engelbrecht, J., and Knudsen, S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49-65

Buchko, G.W., Ni, S., Thrall, B.D., and Kennedy, M.A. 1998. Structural features of the minimal DNA binding domain (M98-F219) of human nucleotide excision repair protein XPA. *Nucleic Acids Res.* **26**, 2779-2788.

Cachin, C. 1994. Pedagogical pattern selection strategies. *Neural Networks* **7,** 175-181.

Callard, G.V., Tchoudakova, A.V., Kishida, M., and Wood, E. 2001. Differential tissue distribution, developmental programming, estrogen regulation and promoter characteristics of cyp19 genes in teleost fish. *J Steroid Biochem. Mol. Biol.* **79**, 305-314.

Carpenter, G.A., and Grossberg, S. 1988. The ART of adaptive pattern recognition by a self-organizing neural network. *Computer* **21**, 77-88.

Chandonia. J.M., and Karplus. M. 1995. Neural networks for secondary structure and structural class predictions *Protein Sci.* **4**, 275-285.

Chen, C.Y., Shyu, A.B. 1995. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20**: 465-470.

Chen, C.Y., Xu, N., Shyu, A.B. 2002. Highly selective actions of HuR in antagonizing AU-rich element-mediated mRNA destabilization. *Mol. Cell Biol.* **22**: 7268-7278.

Cheng, B., and Titterington. D.M. 1994. Neural networks: a review from a statistical perspective. *Statistical Science* **9**, 2-54.

Cherkauer, K.J ., and Shavlik, J.W. 1993. Protein structure prediction: selecting salient features from large candidate pools. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1,** 74-82.

Comtat, C., and Morel, C. 1995. Approximate reconstruction of PET data with a self-organizing neural network. *IEEE. Trans. Neur.* **6, 3**, 783-789.

Courtois, S., Cappellano, C.M., Ball, M., Francou, F.X., Normand, P., Helynck, G., Martinez, A., Kolvek, S.J., Hopke, J., Osburne, M.S., *et al.* 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol.* **69**, 49-55.

Cross, S.H., and Bird, A.P. 1995. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5,** 309-314.

Curtis, D., Lehmann, R., Zamore, P.D. 1995. Translational regulation in development. *Cell* **81**: 171-178.

DeLuka, D, Roseto, G., Bucciarelli, G., and Bernardi, G. 2002. An analysis of the genome of Ciona intestinalis. *Gene* **295,** 311-316.

Decker, C.J., and Parker, R. 1995. Diversity of cytoplasmic functions for the 3' untranslated region of eukaryotic transcripts. *Curr. Opin. Cell Biol.* **7**: 386-392.

Deng, W., Burland, V., Plunkett III, G., Boutin, A., Mayhew, G.F., Liss, P., Perna, N.T., Rose, D.J., Mau, B., Zhous, S., et al. 2002. Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184,** 4601-4611.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., DeTomaso, A., Davidson, B., DiGregorio, A., Gelpke, M., Goodstein, D.M, et al. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298,** 2157-2167.

Demeler, B., and Zhou, G.W. 1991. Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Res.* **19**, 1593-1599.

Detrich, H.W.3rd., Kieran, M.W., Chan, F.Y., Barone, L.M., Yee, K., Rundstadler, J.A., Pratt, S., Ransom, D., and Zon, L.I. 1995. Intraembryonic hematopoietic cell migration during vertebrate development. *Proc. Natl. Acad. Sci. U S A.* **92**, 10713-10717.

Diffley, J.F., and Labib, K. 2002. The chromosome replication cycle. *J. Cell Sci. 115* **5,** 869-872.

Dubchak, I., Holbrook, S.R., and Kim, S.H. 1993a. Prediction of protein folding class from amino acid composition. *Proteins* **16**, 79-91

Dubchak, I., Holbrook, S. R., and Kim. S.H. 1993b. Comparison of two variations of neural network approach to the prediction of protein folding pattern. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1,** 118-126.

Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N.M., Bertone, P., Zhang, Z., and Gerstein, M. 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res.* **30,** 2515-2523.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919-929.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188-200.

Erkki, O. 1989. Neural networks, principal components, and subspaces. *Inter. J. Neur. Sys.* **1**, 61-69.

Evans, T., Reitman, M., and Felsenfeld, G. 1988. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc. Natl. Acad. Sci. U S A.* **85**, 5976-5980.

Eyre-Walker and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev.* **2,** 549-555.

Farber, R., Lapedes, A., and Sirotkin, K. 1992. Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.* **226**, 471-479.

Fariselli, P., Compiani, M., and Casadio, R. 1993. Predicting secondary structure of membrane proteins with neural networks. *Eur. Biophys. J.* **22**, 41-51.

Fariselli, P., and Casadio, R. 1996. HTP: a neural network-based method for predicting the topology of helical transmembrane domains in protein. *Comput. Appl. Biosci.* **12**, 41-48

Ferran, E.A., and Ferrara, P. 1992. Clustering proteins into families using artificial neural networks. *Comput. Appl. Biosci.* **8**, 39-44.

Ferran, E.A., and Pflugfelder, B. 1993. A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *Comput. Appl. Biosci.* **9**, 671-680

Ferran. E.A., Pflugfelder, B., and Ferrara, P. 1994. Self-organized neural maps of human protein sequences. *Protein Sci.* **3**, 507-521.

Fickett, J.W. 1996. The gene identification problem: an overview for developers. *Comput. Chem.* **20**, 103-l 18.

Francino, M.P., and Ochman, H. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* **13,** 240-245.

Gaffney, P.M., Pierce, J.C., Mackinley, A.G., Titchen, D.A., and Glenn, W.K. 2003. Pearl, a novel family of putative transposable elements in bivalve mollusks. *J. Mol. Evol.* **56,** 308-316.

Gautier C. 2000. Compositional bias in DNA. *Curr. Opin. Genet. Dev.* **10,** 656-661.

Gentles, A.J., and Karlin, S. 2001. Genome-Scale Compositional Comparisons in Eukaryotes. *Genome Res.* **11,** 540-546.

Ghirlando, R., and Trainor, C.D. 2000. GATA-1 bends DNA in a site-independent fashion. *J. Biol. Chem.* **275**, 28152-28156.

Giuliano, F., Amgo, P., Scalia, F., Cardo, P. P., and Daminani, G. 1993. Potentially functional regions of nucleic acids recognized by a Kohonen's self'-organizing map. *Comput. Appl. Biosci.* **9**, 687-693.

Gogarten J.P, Olendzenski L. 1999. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**, 630-636.

Granjeon, E., and Tarroux. P. 1995. Detection of compositional constraints in nucleic acid sequences using neural networks. *Comput. Appl. Biosci.* **11**, 29-37.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, 49-62.

Graziano, P., Sabino, L., Giorgio, G., and Cecila S. 1998. UTRdb: a specialized database of 5'- and 3'-untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **26**, 192-195.

Graziano, P., Sabino, L., Giorgio, G., Flavio, L., Flavio, M., Carmela G., and Cecila S. 2002. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.* **30**, 335-340.

Gribskov, M., Devereux, J., and Burgess, R.R., 1984. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* **12**, 539-549.

Hamilton, P., and Reeve, J., 1985. Structure of genes and an insertion element in the methane producing archaebacterium *Msthanobrevibacter* smithii. *Mol. Gen. Genet.* **200**, 47-59.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. 2001. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293,** 1146-1150.

Hannon, G.J. 2002. RNA interference. *Nature* **418,** 244-251.

Hansen, B.M., and Hendriksen, N.B. 2001. Detection of enterotoxic Bacillus cereus and Bacillus thuringiensis strains by PCR analysis. *Appl. Environ. Microbiol.* **67**, 185-189.

Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., and Nielsen, J.O., Hansen J.E. 1995. Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-Gal-NAc:polypeptide N-acetylgalactosaminyl transferase. *Biochem. J.* **308**, 801-813.

Haring, D., and Kypr, J. 1999. Variations of the mononucleotide and short oligonucleotide distributions in the genomes of various organisms. *J. Theor. Biol.* **201**,141-156.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K. *et al.* 2000. The DNA sequence of human chromosome 21. *Nature* **405,** 311-319.

Hecht-Nielsen, R. 1987. Counterpropagation networks. *Appl. Optics.* **26**, 4979-4984.

Henne, A., Daniel, R., Schmitz, R.A., and Gottschalk, G. 1999. Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl. Environ. Microbiol.* **65**, 3901-3907.

Hirst, J.D., and Sternberg, M.J.E. 1992. Prediction of structural and functional features of' protein and nucleic acid sequences by artificial neural networks. *Biochemistry* **31**, 7211-7218.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of representative protein data sets *Protein Sci.* **1** 409-417

Holbrook, S.R., Dubchak, I., and Kim, S.H. 1993. PROBE: a computer program employing an integrated neural network approach to protein structure prediction. *Biotechniques* **14**, 984-989.

Holley, H. L., and Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U S A.* **86**, 152-156.

Holmes, A.J., Gillings, M.R., Nield, B.S., Mabbutt, B.C., Nevalainen, K.M., and Stokes, H.W. 2003. The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ. Microbiol.* **5**, 383-394.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M. et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298,** 129-149.

Horton, P.B., and Kanehisa, M. 1992. An assessment of' neural network and statistical approaches for prediction of *E. coli* promoters sites. *Nucleic Acids Res.* **20**, 4331-4338.

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

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

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-838.

Hutvagner, G., and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297,** 2056-2060.

Iivarinen, J., Kohonen, T., Kangas, J., and Kaski, S. 1994. Visualizing the clusters on the self-organizing map. *Multiple Paradigms for Artificial Intelligence* 122-126.

Ikemura, T., 1981a. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1-21.

Ikemura, T. 1981b. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* **151,** 389-409.

Ikemura, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *J. Mol. Biol.* **158**, 573-597.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13-34.

Ikemura, T., and 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.

Izsvak, Z., Ivics, Z., Garcia, Estefania, D., Fahrenkrug, S.C., and Hackett, P.B. 1996. DANA elements: a family of composite, tRNA-derived short interspersed DNA elements associated with mutational activities in zebrafish (Danio rerio). *Proc. Natl. Acad. Sci. U S A.* **93,** 1077-1081.

Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H., and Hackett, P.B. 1999. Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.* **48**, 13-21.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. 1991. Adaptive mixtures of local experts. *Neural Comput.* **3**, 79-87.

Jeltsch, A., and Pingoud, A. 1996. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.* **42**, 91-92.

Jordan, M.I., and Jacobs, R.A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6**, 181-214.

Kadonaga, J.T. 2002. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.* **34,** 259-264.

Kanaya, S., Ikemura, T., and Kudo, Y. 1994. Relationship between gene function and codon usage in *Escherichia coli* on the basis of principal component analysis. *Genome Inform. Ser.* **5**, 186-187.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., and Ikemura, T. 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.* **276**, 89-99.

Kanaya, S., Kudo, Y., Suzuki, S., and Ikemura T. 1996. Systematization of species-specific diversity of genes in codon usage: comparison of the diversity among bacteria and prediction of the protein production levels in cells. *Genome Inform. Ser.* **7**, 61-71.

Kanaya, S., Kudo, Y., Abe, T., Okazaki, T., Carlos, D.C., and Ikemura, T. 1998. Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome. *Genome Inform. Ser.* **9,** 369-371.

Kanaya, S., Kudo, Y., Nakamura, Y., and Ikemura, T. 1995. Estimation of protein-production levels in *Escherichia coli* genes on the basis of multivariate diversity in codon usage. *Genome Inform. Ser.* **6**, 86-87.

Kanaya, S., Kudo, Y., Nakamura, Y., and 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., Okumura, T., Miyauchi, M., Fukagawa, H., and Kudo, Y. 1997. Assessment of protein coding sequences in *Bacillus subtilis* genome using species-specific diversity of genes in codon usage based on multivariate analysis: comparison of the diversity between *B. subtilis* and *Escherichia coli*. *Res.Comm. in Biochem. and Cell & Mol. Biol.* **1**, 82-92.

Kanaya, S., Yamada, Y., Kudo, Y. and 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., and Kudo, Y., 1993. Assessment of species-specific codon usage by principal component analysis. *Genome Inform. Ser.* **4**, 231-238.

Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* **1,** 598-610.

Karlin, S., Brocchieri, L., Trent, J., Blaisdell, BE., and Mrazek, J. 2002. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. *Theor. Popul. Biol.* **61,** 367-390.

Karlin, S., Campbell, A., and Mrazek, J. 1998. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32,**185-225.

Karlin, S., Doerfler, W., and Cardon, L.R. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J. Virol.* **68,** 2889-2897.

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

Karlin, S., and Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11,** 283-290.

Karlin, S., and Ladunga, I. 1994. Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. U S A.* **91,** 12832-12836.

Karlin, S., and Mrazek, J. 1996. What drives codon choices in human genes ? *J Mol. Biol.* **262,** 459-472.

Karlin, S., and Mrazek, J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U S A.* **94**, 10227-10232.

Katti, M.V., Ranjekar, P.K., and Gupta, V.S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18,** 1161-1167.

Kneller, D.G., Cohen, F.E., and Langridge, R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171-182.

Knietsch, A., Waschkowitz, T., Bowien, S., Henne, A., and Daniel, R. 2003. Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on Escherichia coli. *J. Mol. Microbiol. Biotechnol.* **5**, 46-56.

Knoepfler, P.S., Lu, Q., and Kamps, M.P. 1996. Pbx-1 Hox heterodimers bind DNA on inseparable half-sites that permit intrinsic DNA binding specificity of the Hox partner at nucleotides 3' to a TAAT motif. *Nucleic Acids Res.* **24,** 2288-2294.

Kobayashi, M., Nishikawa, K., and Yamamoto, M. 2001. Hematopoietic regulatory domain of gata1 gene is positively regulated by GATA1 protein in zebrafish embryos. *Development* **128**, 2341-2350.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43,** 59-69.

Kohonen, T. 1988. Learning vector quantization. *Neural Netw.* **11**, 303.

Kohonen, T. 1990. The self-organizing map. *Proc. IEEE* **78,** 1464-1480.

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

Kohonen, T. 1998. The Self-Organizing Map. *Neurocomputing* **21**, 1-6.

Kohonen,T., and Somervuo, P. 2002. How to make large self-organizing maps for nonvectorial data. *Neural Netw.* **15,** 945-52.

Kormng, P.G., Hebsgaard, S.M., Rouze, P., and Brunak, S. 1996. Cleaning the GenBank *Arabidopsis thaliana* data set. *Nucletic Acids Res.* **24**, 316-320.

Kraaijveld, M.A., Mao, j., and Jain, A.K. 1992. A non-linear projection method based on Kohonen's topology preserving maps. *Proceedings of the 11th International Conference on Pattern Recognition*, 41-45.

Kunisawa, T., Kanaya, S., and Kutter, E. 1998. Comparison of Synonymous Distribution Patterns of Bacteriophage and Host Genomes .*DNA Reserch* **5**, 319-326.

Kunkel, T.A. 1992. Biological asymmetries and the fidelity of eukaryotic DNA replication. *Bioessays* **14,** 303-308.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S. et al.. 1997.

The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**, 249-256.

Ladunga, I., Czako, F., Csabai, I., and Geszti, T. 1991. Improving signal peptide prediction accuracy by simulated neural network. *Comput. Appl. Biosci.* **7**, 485-487.

Lawrence, J.G., and Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44,** 383-397.

Lawrence, J.G., and Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U S A.* **95,** 9413-9417.

Le, Q.H., Turcotte, K., and Bureau, T. 2001. Tc8, a Tourist-like transposon in Caenorhabditis elegans. *Genetics* **158**, 1081-1088.

Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292,** 2077-2080.

Lorenz, P., Liebeton, K., Niehaus, F., and Eck, J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr. Opin. Biotechnol.* **13**, 572-577.

Lu, Q., and Kamps, M.P. 1996. Structural determinants within Pbx1 that mediate cooperative DNA binding with pentapeptide-containing Hox proteins: proposal for a model of a Pbx1-Hox-DNA complex. *Mol. Cell Biol.* **16,** 1632-1640.

Ludwig, W., and Schleifer, K.H. 1994. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15**, 155-173.

Lyons, S.E., Lawson, N.D., Lei, L., Bennett, P.E., Weinstein, B.M., and Liu, P.P. 2002. A nonsense mutation in zebrafish gata1 causes the bloodless phenotype in vlad tepes. *Proc. Natl. Acad. Sci. U S A.* **99**, 5454-5459.

MacNeil, I.A., Tiong, C.L., Minor, C., August, P.R., Grossman, T.H., Loiacono, K.A., Lynch, B.A., Phillips, T., Narula, S., Sundaramoorthi, R. et al. 2001. Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J. Mol. Microbiol. Biotechnol.* **3,** 301-308.

Mantovani, R. 1998. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res.* **26**, 1135-1143.

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

Matticsk, J.S. 2003 Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930-939.

Matuda, H. 1999. Detection of conserved domains in protein sequence using a maximum-density subgraph algorithm. *IEICE. TRANS. FUND.* **E83-A**, 4.

Mazumder, B., Seshadri, V., Fox, P.L. 2003. Translational control by the 3'-UTR: the ends specify the means. *Trends Biochem. Sci.* **28**: 91-98.

McMahon, K.D., Dojka, M.A., Pace, N.R., Jenkins, D., and Keasling, J.D. 2002. Polyphosphate kinase from activated sludge performing enhanced biological phosphorus removal. *Appl. Environ. Microbiol.* **68**, 4971-4978.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222,** 851-856.

Meng, A., Tang, H., Yuan, B., Ong, B.A., Long, Q., and Lin, S. 1999. Positive and negative cis-acting elements are required for hematopoietic expression of zebrafish GATA-1. *Blood* **93**, 500-508.

Myllykallio, H., Lopez, P., Lopez, Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H., and Forterre, P. 2000. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288,** 2212-2215.

Nowock, J., Borgmeyer, U., Puschel, A.W., Rupp, R.A., and Sippel, A.E. 1985. The TGGCA protein binds to the MMTV-LTR, the adenovirus origin of replication, and the BK virus enhancer. *Nucleic Acids Res.* **25**, 2045-2061.

Nussinov, R. 1984. Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res.* **12**, 1749-1763.

Nussinov, R. 1986. Some guidelines for identification of recognition sequences: regulatory sequences frequently contain (T)GTG/CAC(A), TGA/TCA and (T)CTC/GAG(A). *Biochim. Biophys. Acta.* **866,** 93-108.

Nussinov, R. 1991. Distinct patterns in the dinucleotide nearest neighbors to G/C and A/T oligomers in eukaryotic sequences. *J. Mol. Evol.* **33**, 259-266.

Ochman, H., Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299-304.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420,** 563-573.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* **51**, 4-12.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. 2002. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16,** 948-958.

Pavlov, Y.I., Newlon, C.S., and Kunkel, TA. 2002. Yeast origins establish a strand bias for replicational mutagenesis. *Mol. Cell* **10,** 207-213.

Pavlov, Y.I., Rogozin, I.B., Galkin, A.P., Aksenova, A.Y., Hanaoka, F., Rada, C., and Kunkel, T.A. 2002. Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase eta during copying of a mouse immunoglobulin kappa light chain transgene. *Proc. Natl. Acad. Sci. U S A.* **99,** 9954-9959.

Philipsen, S., and Suske, G. 1999. A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Res.* **27**, 2991-3000.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res*. **13,** 145-158.

Hecht-Nielsen, R. 1988. Application of Counterpropagation networks. *Neural Netw.* **1**, 131-140.

Radman, M. 1998. DNA replication: one strand may be more equal. *Proc. Natl. Acad. Sci. U S A.* **95,** 9718-9719.

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

Rodriguez, Valera, F. 2002. Approaches to prokaryotic biodiversity: a population genetics perspective. *Environ. Microbiol.* **4**, 628-633.

Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21**, 327-335.

Ronchi, A., Bellorini, M., Mongelli, N., and Mantovani, R. 1995. CCAAT-box binding protein NF-Y (CBF, CP1) recognizes the minor groove and distorts DNA. *Nucleic Acids Res*. **23**, 4565-4572.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C. at al. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541-2547.

Saccone, S., Federico, C., Solovei, I., Croquette, M.F., Valle, G.D., and Bernardi, G. 1999. Identification of the gene-richest bands in human prometaphase chromosomes. *Chromosome Res.* **7,** 379-386.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blocker, H., Perez-Alonso, M., Obermaier, B., et al. 2000. Sequence and analysis of chromosome 3 of the plat *Arabidopsis thaliana*. European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute. *Nature* **408**, 820-2.

Satoh, N. 2003. The ascidian tadpole larva: comparative molecular development and genomics. *Nat. Rev. Genet.* **4,** 285-295.

Schloss, P.D., Handelsman, J. 2003. Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* **14**, 303-310.

Sharp, P.M., and Li, W., The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281-1295.

Sharp, P.M., and Matassi, G. 1994. Codon usage and genome evolution. *Curr. Opin. Gen. Dev.* **4**, 851-860.

Shioiri, C., and Takahata, N. 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.* **53,** 364-376.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657-663.

Smit, AF., and Riggs, AD. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. U S A*. **93,** 1443-1448.

Snyder, E.E., and Stormo, G.D. 1995. Identification of protein coding regions in genomic DNA. *J Mol Biol.* **248**, 1-18.

Soeller, W.C., Oh, C.E., and Kornberg, T.B. 1993. Isolation of cDNAs encoding the *Drosophila* GAGA transcription factor. *Mol. Cell Biol.* **13,** 7961-7970.

Staden, R. 1990. Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol.* 183, 193-211.

Stokes, H.W., Holmes, A.J., Nield, B.S., Holley, M.P., Nevalainen, K.M., Mabbutt, B.C., and Gillings, M.R. 2001. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl. Environ. Microbiol.* **67**, 5240-5246.

Tchoudakova, A., Kishida, M., Wood, E., and Callard, G.V. 2001. Promoter characteristics of two cyp19 genes differentially expressed in the brain and ovary of teleost fish. *J. Steroid. Biochem. Mol. Biol.* **78**, 427-439.

Thomsen, T.R., Finster, K., and Ramsing, N.B. 2001. Biogeochemical and molecular signatures of anaerobic methane oxidation in a marine sediment. *Appl. Environ. Microbiol.* **67**, 1646-1656.

Torsvik, V., and Ovreasm L. 2002. Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microbiol.* **5**, 240-245.

Toth, G., Gaspari, Z., and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*. **10,** 967-981.

Trainor, C.D., Ghirlando, R., and Simpson, M.A. 2000. GATA zinc finger interactions modulate DNA binding and transactivation. *J. Biol. Chem.* **275**, 28157-28166.

Trainor, C.D., Omichinski, J.G., Vandergon, T.L., Gronenborn, A.M., Clore, G.M., and Felsenfeld, G. 1996. A palindromic regulatory site within vertebrate GATA-1 promoters requires both zinc fingers of the GATA-1 DNA-binding domain for high-affinity interaction. *Mol. Cell. Biol.* **16**, 2238-2247.

Turcotte, K., and Bureau, T. 2002. Phylogenetic analysis reveals stowaway-like elements may represent a fourth family of the IS630-Tc1-mariner superfamily. *Genome* **45**, 82-90.

Uberbacher, E.C., Xu, Y., Mural, R.J. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol*. **266**, 259-81.

Ultsch, A. 1993. Self organized feature maps for monitoring and knowledge aquisition of a chemical process. *Proceedings of the International Conference on Artificial Neural Networks* 864-867.

Venter, J.C., Karin, R., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea**,** *Science* 304, 66-74.

Wang, H.C., Badger, J., Kearney, P., and Li, M. 2001. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol. Biol. Evol.* **18,** 792-800.

Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T. 2002. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.* **11,** 13-21.

Wu, C.H., Berry, M., Shivakumar, S., and Mclarty, J. 1995. Neural networks for full-protein sequence classification: sequence encoding with singular value decomposition. *Machine Learning* **21**, 177-193.

Yang, G., and Hall, T.C. 2003. MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res.* **31**, 3659-3665.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25-33.

Zdobnov, E.M., VonMering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, GM., et al. 2002. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* **298,** 149-159.

Zhang, H., Kolb, F.A., Brondani, V., Billy, E., and Filipowicz, W. 2002. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO* **21,** 5875-5885.