A novel bioinformatics strategy for searching industrially useful genome resources from metagenomic sequence libraries

Hiroshi Uehara, Yuki Iwasaki, Chieko Wada, Toshimichi Ikemura and Takashi Abe*

Nagahama Institute of Bio-Science and Technology, Tamura-cho 1266, Nagahama-shi, Shiga 526-0829, Japan

(Received 19 November 2010, accepted 27 December 2010)

Although remarkable progress in metagenomic sequencing of various environmental samples has been made, large numbers of fragment sequences have been registered in the international DNA databanks, primarily without information on gene function and phylotype, and thus with limited usefulness. Industrial useful biological activity is often carried out by a set of genes, such as those constituting an operon. In this connection, metagenomic approaches have a weakness because sets of the genes are usually split up, since the sequences obtained by metagenome analyses are fragmented into 1-kb or much shorter segments. Therefore, even when a set of genes responsible for an industrially useful function is found in one metagenome library, it is usually difficult to know whether a single genome harbors the entire gene set or whether different genomes have individual genes. By modifying Self-Organizing Map (SOM), we previously developed BLSOM for oligonucleotide composition, which allowed classification (self-organization) of sequence fragments according to genomes. Because BLSOM could reassociate genomic fragments according to genomes, BLSOM may ameliorate the abovementioned weakness of metagenome analyses. Here, we have developed a strategy for clustering of metagenomic sequences according to phylotypes and genomes, by testing a gene set contributing to environment preservation.

Key words: bioinformatics, environment preservation, metagenome, Self-Organizing Map, uncultured microorganism

INTRODUCTION

More than 99% of microorganisms inhabiting natural environments are difficult to culture under laboratory conditions. While genomes of the unculturable organisms have remained primarily uncharacterized, these should contain a wide range of novel genes of scientific and industrial interest. To explore such an enormous quantity of novel genome resources, metagenomic analyses have been developed to find industrially and scientifically valuable genes and to study microbial communities in a wide variety of environments (Amann et al., 1995; Hugenholtz and Pace, 1996; Rondon et al., 2000; DeLong, 2002; Lorenz et al., 2002; Schloss and Handelsman, 2003). The metagenomic analysis is a culture-independent approach that performs shotgun sequencing on mixed genome DNAs extracted from environmental samples.

The accelerated increase of sequences obtained from environmental samples because of revolutionary development of DNA sequencing technologies has attracted attention, not only from scientific fields, but also from industrial and medical fields. Vast numbers of fragment sequences have been deposited in the International Nucleotide Sequence Databases (INSD), but, notably, the majority of the sequences are stored in the INSD with no phylogenetic or functional annotation, and thus, with the limited usefulness.

The metagenomic sequencing is undoubtedly a powerful strategy for comprehensive study of a microbial community in an ecosystem, but for most of the sequences, it is difficult to predict from what phylotypes each sequence is derived. This situation has arisen because orthologous sequence sets, which cover a broad phylogenetic range for constructing reliable phylogenetic trees through sequence homology searches, are unavailable for novel gene sequences. G plus C percentage (%GC) has long been used as a fundamental parameter for phylogenetic
classification of microorganisms, but the %GC is apparently too small a parameter to differentiate a wide variety of species. Oligonucleotide composition, however, can be used even to distinguish species with the same %GC, because oligonucleotide composition varies significantly among microbial genomes and has been called the “genome signature” by Karlin and his colleagues (Karlin et al., 1998).

Phylogenetic clustering and classification in the present study is designed as an extension of the single parameter “%GC” to the multiple parameters “oligonucleotide frequencies”. In this connection, we previously modified the SOM (Self-Organizing Map) developed by Kohonen’s group (Kohonen, 1990; Kohonen et al., 1996) for genome informatics on the basis of batch-learning SOM “BLSOM” (Kanaya et al., 2001; Abe et al., 2003), which makes the learning process and resulting map independent of the order of data input. The BLSOM thus developed could recognize species-specific characteristics of oligonucleotide composition in a wide range of genomes and permitted clustering (self-organization) of genomic fragments according to genomes with neither the orthologous sequence set nor the troublesome and possible error-prone processes of sequence alignments. Therefore, BLSOM is thought to be the most suitable strategy for phylogenetic classification and characterization of a huge quantity of sequence fragments obtained from pooled genomic samples of uncultured microbes in an environmental or clinical sample. In our previous BLSOM analysis of metagenomic sequences (Abe et al., 2005), we initially constructed a large-scale BLSOM with all known genome sequences to characterize oligonucleotide composition of all species-known microorganisms currently available. Then, on this large-scale BLSOM, each metagenomic sequence derived from an environmental sample was mapped by finding the lattice point with the closest similarity of oligonucleotide composition, as described in MATERIALS AND METHODS. In order to construct this large-scale BLSOM, we had to use high performance supercomputers such as the Earth simulator (Abe et al., 2006a, 2006b). While this strategy is undoubtedly powerful and has already been applied to various metagenome studies (Uchiyama et al., 2005; Hayashi et al., 2005; Kosaka et al., 2008), use of high performance supercomputers running user-implemented programs have not been popular and therefore its applicable fields and users appeared to be limited. In the present study, we developed a BLSOM method for metagenome studies that can be conducted with PCs or PC clusters.

Our previous studies (Abe et al., 2005, 2006a, 2006b) using high performance supercomputers showed that genomic fragments of several kb (e.g., 2 kb) or longer gave a high accuracy of clustering of fragment sequences according to genomes. However, when the fragment size was shortened to 1 kb, the level of accuracy was reduced and sequences derived from one species made a few distinct territories rather than one major territory (Abe et al., 2003). In the case of metagenomic sequence analyses, contigs of several kb or longer in length were only obtained from genomes of dominant species in an environment, and a major portion of metagenome sequences were far less than several kb. In the present study, using PCs and PC clusters, we searched for BLSOM conditions suitable for the metagenomic sequences by analyzing 1- and 2-kb sequences.

Biological activity with industrial usefulness, such as processes responsible for environmental preservation, is often carried out by a set of genes rather than a single gene; e.g., an operon responsible for one metabolic activity. In this connection, metagenomic approaches have a weakness because a set of the genes was usually split up, because the sequences obtained by metagenome analyses were primarily fragmented into kb-level or shorter segments. Contigs with a significant length, such as those covering an operon, were obtained only in the cases of very dominant species. Therefore, when a set of genes responsible for an industrially useful function is found in one metagenome library, it is difficult to know whether a single genome harbors the gene set of interest or whether different genomes coexisting in the sample have individual genes. From the industrial and scientific view, it is valuable to find a metagenomic library that may have a single genome harboring a full set or, at least, a major portion of the gene set. Because BLSOM has the ability to reassociate fragmented sequences according to a genome, BLSOM may ameliorate the weakness of metagenomic sequencing. In the present study, we developed a strategy for the in-silico association of fragmented sequences according to phylotype (hopefully even to species), by focusing on a set of genes contributing to environmental cleanup and preservation.

**MATERIALS AND METHODS**

**DNA sequences** Genomic fragment sequences derived from metagenome analyses and from bacterial species were obtained from http://www.ncbi.nlm.nih.gov/GenBank/. Metagenome sequences shorter than 1 kb in length were not included in the present study. When the number of undetermined nucleotides (Ns) in a sequence exceeded 10% of the window size, the sequence was omitted from the BLSOM analysis. When the number of Ns was less than 10%, the oligonucleotide frequencies were normalized to the length without Ns and included in the BLSOM analysis. Sequences that were longer than a window size (1 or 2 kb) were segmented into the window size, and the residual sequences, which were shorter than the window size, were omitted from the BLSOM analysis.

**BLSOM construction** SOM is an unsupervised neural
network algorithm that implements a characteristic non-linear projection from the high-dimensional space of input data onto a two-dimensional array of weight vectors (Kohonen, 1990; Kohonen et al., 1996). We previously modified the conventional SOM for genomeinformatics to make the learning process and resulting map independent of the order of data input, on the basis of batch learning SOM “BLSOM” (Kanaya et al., 2001; Abe et al., 2003). The initial weight vectors were defined by Principal Component Analysis (PCA) instead of random values. BLSOM learning was conducted as described previously (Abe et al., 2003), and the BLSOM program was obtained from UNTROD Inc. (y_wada@nagahama-i-bio.ac.jp). Mapping of a metagenomic sequence on BLSOMs was conducted by finding the lattice point that had the minimum Euclidean distance in the multidimensional space. In the case of mapping of metagenomic sequences harboring candidate genes for PCB degradation on 2-kb BLSOMs, sequences even less than 2 kb (but longer than 1 kb) were included after normalization to the length. The program for this mapping on BLSOMs was also available from UNTROD Inc. (y_wada@nagahama-i-bio.ac.jp).

**Generation of random nucleotide sequences** For each genomic or metagenomic sequence, a random sequence that had nearly the same oligonucleotide (e.g., dinucleotide) composition as the respective sequence was generated by the Markov chain model (Abe et al., 2009). In the case of the random sequence with nearly the same dinucleotide composition (designated Di-Random), we initially calculated dinucleotide composition of one genomic or metagenomic sequence and chose the first dinucleotide randomly but with a weight level to reflect the level of the dinucleotide composition of this sequence. When the first dinucleotide thus chosen in this process was “GC”, the next nucleotide “X” which followed the “GC” was chosen randomly, but with a weight level reflecting the “CX” dinucleotide composition in the respective genomic or metagenomic sequence. This process was repeated until the length of the Di-Random sequence reached to that of the respective sequence. In the present study, ten random sequences were generated for each genomic or metagenomic sequence. The program for the generation of random nucleotide sequences was also available from UNTROD Inc. (y_wada@nagahama-i-bio.ac.jp or http://octonion.jp/genome/).

**RESULTS AND DISCUSSION**

**BLSOMs for sequences obtained by metagenome analyses** To test the clustering power of BLSOM for oligonucleotide composition in metagenomic sequences obtained from various environmental samples, we analyzed a large quantity of fragment sequences obtained from eight typical metagenomic libraries currently available (552,303 sequence fragments) with BLSOMs. In the present study, however, a huge number of sequences obtained from Sargasso Sea (Venter et al., 2004) and from human gut samples of 124 Europeans (Qin et al., 2010) were not included, but the sequences from Hawaii Ocean (DeLong et al., 2006; Frias-Lopez et al., 2008) and from human guts of 13 Japanese (Kurokawa et al., 2007) were included. One main purpose of this study was to develop a bioinformatics strategy for searching industrially useful genes contributing to environmental preservation from a wide variety of environments. An excessively large contribution of sequences from the first two samples was thought to interfere with knowledge discovery from other environments. Furthermore, a major portion of the characteristics of the two environments was thought to be appreciably represented in the latter two samples. Another reason for avoidance of the two gigantic samples was that the purpose of the present study is to develop BLSOM methods that can be conducted with PCs or PC clusters, rather than with high-performance supercomputers.

To develop an informatics strategy useful to search gene candidates contributing to environmental cleanup and preservation, we focused on metagenomic sequences longer than 1 kb, which likely harbored an intact protein-coding sequence. Previous BLSOM analyses of a wide range of completely-sequenced prokaryotes showed that genomic fragments of 5 kb or longer derived from one species were clustered (self-organized) according to the species, resulting in one species-specific territory (Abe et al., 2003; Abe et al., 2005). Much shorter fragments (e.g., 1-kb fragments), however, formed a few split, but species-specific, territories (Abe et al., 2003). This split into a few sub-territories reflected in part the transcriptional polarity of the gene present in the fragment. In DNA databases, only one strand of a pair of complementary sequences is registered. Some sequences represent the coding sequences of the protein-coding genes but others represent the template sequences, and these two types of sequences have somewhat different characteristics of oligonucleotide composition, resulting in the split of the species-specific territory into at least two separate territories. When we constructed BLSOM in which the frequencies of a pair of complementary oligonucleotides (e.g., AAC and GTT) in each fragment were summed, the tendency of the splitting into a few territories was diminished for most of species (Abe et al., 2005). For phylogenetic clustering of metagenomic sequences, it is unnecessary to know the transcriptional polarity of the sequence, and the split into a few territories complicates the clustering according to genome. Therefore, in the present study, we constructed BLSOMs for the degenerate sets of tri- and tetranucleotides (DeGeTri- and DeGeTetra-BLSOMs, respectively).

The results of DeGeTri- and DeGeTetra-BLSOMs for 1-
kb sequence fragments (i.e., a window size of 1 kb) were listed in Fig. 1, A and B, respectively. As a separate analysis, we constructed DegeTri- and DegeTetra-BLSOMs with a 2-kb sequence window (Fig. 2, A and B, respectively). Sequences longer than 2 kb (approximately 15% of the sequences longer than 1 kb) should rep-

<table>
<thead>
<tr>
<th>A 1-kb DegeTri</th>
<th>All</th>
<th>Brisbane Sludge</th>
<th>Madison Sludge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington Lake</td>
<td>Hawaii Ocean</td>
<td>Whale Bones</td>
<td></td>
</tr>
<tr>
<td>Minnesota Land</td>
<td>Human Guts</td>
<td>Richmond Mine</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B 1-kb DegeTetra</th>
<th>All</th>
<th>Brisbane Sludge</th>
<th>Madison Sludge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington Lake</td>
<td>Hawaii Ocean</td>
<td>Whale Bones</td>
<td></td>
</tr>
<tr>
<td>Minnesota Land</td>
<td>Human Guts</td>
<td>Richmond Mine</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. BLSOMs for 1-kb metagenomic sequences of 8 environmental samples. (A) and (B) DegeTri- and DegeTetra-BLSOM, respectively. Lattice points that include sequences from more than one environmental sample are indicated in black, and those containing sequences only from one sample are indicated in color as follows: Brisbane Active Sludge (■), Madison Active Sludge (▲), Washington Lake (●), Hawaii Ocean (○), Whale Bones (◆), Minnesota Farm Land (□), Human Guts (●), and Richmond Mine (■). In eight panels with the sample names above the panels, all lattice points containing sequences from one sample are indicated in a color representing the sample, regardless of coexistence of sequences from other samples.
resent contig sequences obtained by the assembling process of shot-gun sequencing, and thus, primarily represent sequences derived from dominant or subdominant species in each environment. In other words, BLSOMs constructed with the 2-kb sequences is suitable for determining the characteristics of dominant and subdominant species in the environment.

In the “All” panels in Figs. 1 and 2, lattice points that contained sequences from one environment are indicated by the color representing that environment, and those that included sequences from more than one environment are indicated in black. In each of other eight panels in Figs. 1A, 1B, 2A, and 2B, all lattice points containing sequences derived from one environment were indicated.
by the color representing that environment. Difference in characteristics of individual environmental samples could be visualized on a single plane, supporting efficient knowledge discovery from a large number of metagenomic sequences and showing a powerful function of BLSOM. The observation that DegeTri- and DegeTetra-BLSOMs gave similar results showed that the separation patterns should represent basal characteristics of the environmental samples. We will explain individual separation patterns in more detail, and first compare patterns of two active sludge samples: Brisbane and Madison Sludge (Martin et al., 2006). On all four BLSOMs, global patterns of the two sludge samples (Brisbane and Madison Sludge panels) resembled each other, but there were clear compact zones that were specifically found only in one sludge sample. Since the compact zones were colored in red or pink even in the “All” panels, the sequences were derived presumably from characteristic species in the environment, rather than the species ubiquitously present in various environments. Visualization power of BLSOM could support this kind of efficient and luminous knowledge discovery.

The pattern of the Washington Lake (Kalyuzhnaya et al., 2008) was much simpler on all four BLSOMs than that of the Hawaii Ocean. A few large but isolated territories were observed in the Washington Lake, indicating that microorganisms with close phylogenetic relations may dominate in the sample. Sequences derived from the Whale Bones (Tringe et al., 2005) or the Minnesota Farm Land (Edwards et al., 2006) were distributed widely on 1-kb DegeTri- and DegeTetra-BLSOMs, but were much localized on 2-kb DegeTri- and DegeTetra-BLSOMs. This indicated that these samples contained a wide variety of genomes but phylotypes of dominant species were rather limited. In the cases of the Human Guts, the patterns were very complex both on 1- and 2-kb BLSOMs. This showed a high complexity of genomes present in this sample, which was a mixture of gut samples from 13 different Japanese individuals. There were wide green zones that were primarily composed of sequences derived from Human Guts (green zones in the “All” panels in Figs. 1 and 2), indicating that the microbial community in the human body environment differed significantly from that of natural environments.

Reassociation of genomic fragments according to phylotype and genome The pattern of sequences derived from an acid mine drainage at the Richmond Mine was very simple (“Richmond Mine” in Figs. 1 and 2). Tyson et al. (2004) selected acidophilic biofilms in this acid mine drainage for metagenome shotgun-sequencing because of the low-complexity of constituent genomes. They attempted to reconstruct dominant genomes by assembling a large number of sequences obtained with the shotgun sequencing, and actually, reconstructed one nearly complete genome of Leptospirillum group II, and many scaffold sequences for Ferroplasma type II (Tyson et al., 2004).

In the “Richmond Mine” panel of 2-kb DegeTri- and DegeTetra BLSOMs (Fig. 2, A and B), there were two major compact territories: one was quite compact (A territory) but the other was rather extended (B territory). To examine the sequences present in these zones, a BLAST search of each sequence in A or B territory against NCBI RefSeq (non-redundant database of sequences) was conducted. More than 99% of sequences from the compact A territory were assigned to the sequences from Leptospirillum (Bacteria; Nitrospirae; Nitrospira (class); Nitrospirales), showing the A territory to represent sequences primarily derived from Leptospirillum. In contrast, 74% of the sequences in the extended B territory showed the highest similarity to sequences from Ferroplasma (Archaea; Euryarchaeota; Thermoplasmata; Thermoplasmatales), and 10% and 5% of the B territory sequences showed the highest similarity to sequences from Thermoplasmata and Picrophilus, respectively, which belong to the Archaea order, as does Ferroplasma. This showed that there was a more complex composition of sequences than the composition in the A territory and that even a continuous territory (e.g., the B territory) did not necessarily represent one genome.

Recently, Banfield and her colleagues (Dick et al., 2009), developed a new type of SOM “ESOM (emergent SOM)” and obtained clear phylotype-specific classification of metagenomic sequences by analyzing several acidophilic biofilms in the Richmond mine with the tetranucleotide ESOM. In their ESOM, U-Matrix method (Ultsch, 1993) could successfully visualize and assign territory borders between different phylogenetic groups. In the BLSOMs used in the present study, territory borders of different taxonomic groups were assigned and visualized by mixing metagenomic sequences from a wide verity of environments, while the U-Matrix method could be used also in our BLSOM, as described previously (Abe et al., 2006c).

The findings obtained by both BLSOM and ESOM supported the view that SOM methods have a potentiality for reassociating genomic fragments in a metagenome library according to genome. Even using the conventional sequence homology searches, this reassociation could be obtained for a dominant species, especially in a sample with a low genome complexity, by constructing one complete genome after assembling a large number of fragment sequences. This reassociation of metagenome sequences according to genome, however, becomes increasingly difficult, when subdominant or minor populations are concerned. When a completely-sequenced genome with a very close phylogenetic relationship is available, contigs of metagenomic sequences even derived from a subdominant species may be mapped on the tem-
plate genome and thus classified according to genome. However, a good template genome would not be available for novel, poorly-characterized phylogenetic groups. Because one main purpose of metagenome analyses was to find novel species in environments, the method inevitably dependent on a template genome is apparently insufficient. In the case of BLSOM and ESOM methods, reassociation (self-organization) of genomic fragments according to genome can be attained without the template genome, showing wide applicability.

**Genes contributing to environmental preservation and restoration** Metagenome approaches should allow extensive surveys of sequences useful in scientific and industrial applications. Biological activity with industrial usefulness is often carried out by a set of genes rather than a single gene, such as that constituting an operon. However, contigs of a significant length, such as those covering an operon, could not be obtained, except in the case of very dominant species. Therefore, even when a set of genes of interest is found in an environmental sample with sequence homology searches, it is difficult to know whether a single genome harbors a set of the genes or different genomes in the sample happen to have these genes as a whole. From the industrial and scientific view, the former case, especially representing novel genome, should be valuable, and a bioinformatics strategy to distinguish the two cases becomes important for effectively utilizing metagenomic sequences. Because BLSOM have a potentiality to reassociate fragmental sequences according to genome, it may distinguish the two cases and thus ameliorate a weakness of the metagenome approaches. Because the distribution patterns of sequences from individual environments on DegetTri- and DegeTetra-BLSOMs resembled each other and the pattern of 2-kb sequences was much simpler than that of 1-kb sequences (Figs. 1 and 2), we used 2-kb DegeTetra- BLSOM in the following analyses.

To test a feasibility of the abovementioned informatics strategy, we searched gene candidates useful for environmental preservation such as degradation of hazardous compounds and present in specific environments rather than those in ubiquitous environments. At the first step, by referring to the literatures concerning degradation of PCB (Shimizu et al., 2001), dioxin (Armengaud et al., 1999), polyethylene glycol (Tani et al., 2007), and fluorene (Schuler et al., 2008), we chose a set of genes responsible for one metabolic activity, whose functions have been proven by experiments, or a set of genes known to be clustered in a restrictive genomic portion primarily with the same transcriptional polarity; the detail procedures were described later for a gene set for PCB degradation. Using amino-acid sequences of the gene products as queries, genes from other species that most likely have the same functions were searched against Non-Redundant Protein Sequence Database (nr) in NCBI on the basis of strict criteria for the DDBJ-BLASTp search (more than 80% level of identity and coverage). Then, using amino-acid sequences derived from these newly-expanded genes as queries, we searched the candidate genes of interest from a large quantity of metagenomic sequences with the DDBJ-tBLASTn search under a strict criterion (e values less than 1e–20). For each environmental source analyzed in Figs. 1 and 2, the numbers of candidate genes found for each metabolic activity were listed (Table 1).

**Genes for PCB degradation** As a model system to develop a bioinformatics strategy, we focused on a complex metabolic system composed of many (rather than a few) enzymatic genes and chose the metabolic activity for PCB degradation, which was composed of ten enzyme genes, bphA1, bphA2, bphA3, bphA4, bphB, bphC, bphD, bphE, bphF, and bphG, harbored by *Rhodococcus* sp. strain RHA1 (Shimizu et al., 2001). Then, we expanded this set of genes by using Non-Redundant Protein Sequence Database in NCBI, as described above, and using the expanded amino-acid sequences as queries in the tBLASTn search, candidate genes displaying the activity of the PCB degradation were searched from a large quantity of metagenomic sequences and listed the number of found candidates for each of the ten genes in various environments (Table 1). Candidate genes representing nine out of the ten enzyme genes were found in the Washington Lake samples, and eight out of the ten genes were found in the Hawaii Ocean, Human Guts, Minnesota Land, and Whale Bone samples. It should be mentioned here that a full set of the gene candidates of interest was not found in any environmental samples analyzed here. This might indicate the absence of genomes having a full set of the genes in these environments. However, in usual metagenomic data, a coverage density by metagenomic sequences may not reach to a level that completely covers a certain genome. In a practical approach, the first trial may be a search for genome resources to cover a significant portion of the metabolic pathway of interest, by analyzing various environmental samples. If a significant portion is found in a certain sample, a larger scale of metagenomic sequencing will be conducted on the basis of findings in the first trial. A purpose of the present study was to develop an informatics strategy, rather than an actual search for the PCB-degradation pathway, and therefore, we focused on the above five environmental samples harboring a major part of the degradation pathway.

We next identified lattice points, on which the metagenomic sequences harboring the respective gene candidates were mapped, as described in MATERIALS AND METHODS. Lattices points containing the candidate gene sequences were widely scattered in the Washington Lake sample (Fig. 3A). While nine out the ten gene can-
candidates were found in Table 1, there was only one gene candidate sequence in the major territory of the Washington Lake. These findings indicated that the candidate sequences were presumably derived from various genomes. Similar results were found in Minnesota Land and Whale Bone samples (data not shown). In contrast,
Bioinformatics for metagenome studies

in the Hawaii Ocean, sequences of gene candidates were located in a restricted zone (marked by a circle in Fig. 3B) and covered seven genes (bphA1, bphA2, bphB, bphC, bphD, bphE, and bphF) out of the initial eight genes listed in Table 1, showing a powerful function of BLSOM for identifying potentially useful genome resources. In the Human Guts, clusters of the candidate gene sequences were also found in restricted zones on BLSOMs (Fig. 3C). Because candidate gene sequences from the Hawaii Ocean were located in a more compact territory than those from the Human Guts, we selected the former compact territory, as a test example for developing a bioinformatics method.

In the “All” panels in Figs. 1 and 2, the compact territory in the Hawaii Ocean (marked by a circle in Fig. 3B) was colored in brownish yellow, showing that the sequences locating in this territory were derived from the characteristic species in the Hawaii Ocean, rather than those present in ubiquitous environments. However, there was a possibility that this compact territory might be composed of multiple species, which formed a joint continuous territory, as found in the B territory in the Richmond Mine (Fig. 2). In the case of novel and closely related species present only in this environment, these could not be separated from each other because metagenomic sequences derived from other environments could not sectionalize their joint territory. In order to investigate whether this compact territory of the Hawaii Ocean was composed of sequences from a single genome or multiple genomes, we compared the amino acid sequences derived from metagenomic sequences mapped in this territory, for each candidate protein. If the amino acid sequences derived from metagenomic candidate genes for one enzyme were different from each other, the territory should be composed of multiple genomes because one sequence for one enzyme was expected for most prokaryotic genomes. In this compact territory, three or more different amino-acid sequences were observed for each of the enzyme genes of interest, showing the presence of multiple genomes.

**Addition of computer-generated random sequences**

The resolving power for individual species on BLSOM in Figs. 1 and 2 was inevitably dependent on the metagenomic sequences included in the analysis, and the compact territory specific to the Hawaii Ocean was indicated.

---

**Fig. 3.** Lattice points containing sequences harboring gene candidates for PCB-degradation pathway. (A), (B), (C): 2-kb DegeTetra-BLSOM for Washington Lake, Hawaii Ocean and Human Guts listed in Fig. 2B, respectively. In the right-side panel, lattice points containing sequences harboring the gene candidate sequences were indicated by dots. A compact territory in the Hawaii Ocean containing a cluster of gene candidate sequences was marked with a circle.
to be composed of multiple genomes. We next developed a wide applicable method, which could separate metagenomic sequences present in one compact territory according to phylotypes and hopefully to species, without effects of other metagenomic sequences coexisting. As a specialized strategy to disclose sequence characteristics of a single genome without coexistence of other genome sequences, we previously developed one method, in which BLSOM was constructed for fragment sequences (e.g., a window size of 10-kb) derived from one genome plus computer-generated random sequences (Abe et al., 2009). In the present study, we extended this method for use in metagenomic analyses. To test the feasibility of the method and to search suitable analysis conditions for the kb-level sequences, we first analyzed 2-kb sequences from three bacterial species (Escherichia coli, Bacillus subtilis and Haemophilus influenzae), as a model case. When only the genomic sequences were analyzed, most of the genomic sequences were clustered (self-organized) according to species (the right panel in Fig. 4A), as expected; a minor portion of sequences was located far away from the respective species-specific territory and these may correspond primarily to horizontally-transferred genomic segments, which had the base composition different from the recipient genome (Abe et al., 2003). If we do not know the species from which each sequence is derived, we cannot color the lattice points and thus cannot detect borders between species (the left panel in Fig. 4A). This is the case of metagenomic analyses because we don’t know usually the origin of the metagenomic sequences.

To obtain random sequences that had sequence characteristics appreciably similar to each 2-kb genomic sequence of these bacteria, we generated three types of random sequences as described in MATERIALS AND METHODS: random sequences having nearly the same mono-, di- or trinucleotide composition (Mono-random, Di-random or Tri-random, respectively). After generating ten random sequences of one type for each genomic sequence, we constructed DegeTetra-BLSOM with the genomic plus computer-generated random sequences (Fig. 4, B, C and D). In the presence of Mono-random sequences, all genomic sequences were clearly separated
from random sequences, but genomic sequences from the three bacteria formed an extended continuous territory without barrier between the bacterial species (Fig. 4B). This indicated that the tetranucleotide compositions in Mono-random sequences were too distinct from tetranucleotide compositions in real genomic sequences to differentiate these genomic sequences according to species.

In the presence of Di-random sequences (Fig. 4C), genomic sequences derived from different species were clearly separated from each other and were surrounded by the random sequences derived from the respective genome, showing that the tetranucleotide compositions in genomic sequences of one species were more similar to the tetranucleotide compositions in the Di-random sequences generated from the respective genomic sequences than the tetranucleotide compositions in the genomic sequences derived from other species. A minor portion was again located away from the species-specific territory, and these may primarily represent horizontally-transferred genomic segments (Abe et al., 2003). In the presence of Tri-random sequences (Fig. 4D), a major portion of *E. coli* sequences again formed a compact major territory, but sequences of other two species were widely scattered. This indicated, so far concerning oligonucleotide composition, the trinucleotide composition was already a good representative of genome signature (Karlin et al., 1998) for *B. subtilis* or *H. influenza*, but was not enough to represent the genome signature of *E. coli*. While biological significance and causative factors to produce this difference among bacterial species appeared to be very interesting, we did not consider this point in the present study but considered an actual strategy to apply this method to metagenomic analyses.

The results obtained in Fig. 4 suggests a possibility that, when DegeTetra BLSOM is constructed in the presence of Di-random sequences for each metagenomic sequence, metagenomic sequences derived from one environmental species may be surrounded by random sequences representing this species and may be separated from metagenomic sequences derived from other species, without information concerning the constituent species. In the case of metagenomic sequences, however, phylogenetic relationship between the constituent genomes was unknown in most cases and there should be various cases, and difficult in the genome-specific clustering will increases as the increase of phylogenetic closeness between the constituent genomes. While the mixing with Di-random sequences gave a good separation among the abovementioned three species, analyses on various conditions may be required in actual metagenomic samples.

**Separation of metagenomic sequences in one compact territory by mixing with random sequences**

We applied the abovementioned method to the metagenomic sequences, which were clustered in the compact territory in the Hawaii Ocean (marked by a circle in Fig. 3B). In order to obtain random sequences that had sequence characteristics appreciably similar to metagenomic sequences, ten random sequences with nearly the same di- or trinucleotide composition to each metagenomic sequence in the compact territory were generated. Then, we constructed DegeTetra-BLSOM for the metagenomic sequences plus the Di- or Tri-random sequences (Fig. 5). In the presence of Di-random sequences (Fig. 5A), one major compact territory, a few small territories and many scattered points were observed. Then, we mapped the Hawaii Ocean sequences harboring the candidate genes present in the compact territory in Fig. 3B. A major portion of the sequences, which contained six gene candidates (*bphA1, bphA2, bphB, bphC, bphE*, and *bphF*), was localized in the major territory (Gene candidates in Fig. 5A). Because more than one amino-acid sequence were found for one gene candidate, sequences derived from more than one genome should be present.

In the presence of Tri-random sequences (Fig. 5B), two compact territories and many scattered points were observed. In a major territory we found clustering of sequences for *bphA1, bphB, bphC, and bphD*, and in a minor territory, *bphA1, bphA2, bphC, and bphF* (Gene candidates in Fig. 5B). Importantly, the sets of the gene candidates differed in amino-acid sequences between the two compact territories, showing two territories to represent different genomes. Candidate genes in the smaller territory had one amino acid sequence for each gene, but more than one amino acid sequence were found in the larger territory, indicating that at least the larger territory had sequences derived from more than one genome. On many scattered lattice points, many various sequences harboring the gene candidates with different amino-acid sequences were mapped. This showed that the sequences derived from various minor genomes coexisting in the initial compact territory could be separated from sequences derived from the major genomes of interest.

When we separately generate the new random sequences using metagenomic sequences present in each of the newly-formed compact territories and mix these with the metagenomic sequences, a new separation of the metagenomic sequences may be obtained if the territory is composed of more than one genome, and this process can be repeated. This is an advantage of the present method and therefore, we generated Tri-random sequences for the metagenomic sequences belonging to two territories in Fig. 5B, and constructed DegeTetra-BLSOM in the presence of the newly-generated random sequences (Fig. 5, C and D). All sequences from the minor territory formed one territory (Fig. 5C), supporting that this territory (and thus the minor territory in Fig. 5B) may be composed primarily of one genome; there were four candidate genes (*bphA1, bphA2, bphC, and bphF*)...
In contrast, sequences from the major territory again formed two territories and scattered points (Fig. 5D). In this new large territory, three candidate genes (bphA1, bphB, and bphC) were found and more than one amino acid sequence was found for one enzyme candidate (Gene candidates in Fig. 5D). The results in Fig. 5 predicted that the genome present in the territory in Fig. 5C may be the most suitable genome resource for PCB-degradation in the Hawaii Ocean sample. The average %GC of the lattice points harboring the PCB-degradation genes in Fig. 5C and that of the sequence fragments harboring the genes were 63.0% and 60.2%, respectively. Judged from nucleotide compositions both from BLSOM and from gene sequences, the genome of attention in the Hawaii Ocean sample was GC-rich. It should be noted here that a total length of the metagenomic sequences even in this territory became less than 500 kb, which apparently could cover only a part of one genome. Therefore, even one amino acid sequence was found for each gene candidate, this was not a complete proof that this territory represented a single genome at the present moment.

A main purpose of the present study was to develop a bioinformatics strategy, rather than to find a practical enzyme system for PCB degradation. Therefore, we did not conduct further analyses but considered possible reasons why only a half of the genes responsible for one biological activity were found in the final territory. 1) Random sampling of sequences inevitably happened in the metagenomic sequencing and the density of the coverage of metagenomic sequences did not reach to a level to fully cover the respective genome. 2) Some candidate genes might be the horizontally transferred genes, and these might be separated from the compact territory harboring the main body of the genome of interest. 3) The sequences even derived from a single genome may be split up in the presence of Tri-random sequences, as found for B. subtilis or H. influenza. To solve this third problem, we have to study in detail the causative factors to form more than one territory in the presence of random sequences (e.g., B. subtilis or H. influenza in Fig. 4D) even when sequences from one genome were analyzed, by analyzing a wide range of species-known genomes. 4) The environmental species of interest may have an enzymatic pathway slightly different from a pathway described in the literature and used in the start point.

This type of information concerning a gene set responsible for an interested pathway in actual environmental samples should be valuable for designing practical experiments of search for industrially useful genomes. Importantly, conventional methods could provide information concerning a gene set responsible for a biological activity,
only when a very large amount of sequences was available for constructing a nearly complete genome by assembling or when a complete genome sequence was available for a reliable template for mapping of metagenomic sequences. It should be stressed again that the present in-silico association could be achieved without a template genome for mapping and thus was applicable to the really novel, environmental genomes.

This work was supported by the Integrated Database Project and Grant-in-Aid for Scientific Research (C, No. 20510194) and for Young Scientists (B, No. 20700273) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

REFERENCES
Abe, T., Sugawara, H., Kanaya, S., and Ikemura, T. (2006b) Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator. J. Earth Simulator 6, 17–23.


