A Large-scale Genomics and Proteomics Analyses Conducted by the Earth Simulator

Project Representative
Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

Authors
Takashi Abe*1, Hiroshi Uehara*1, Yuki Iwasaki*1, Kennosuke Wada*1 and Toshimichi Ikemura*1
*1 Nagahama Institute of Bio-Science and Technology

Although remarkable progress in metagenomic sequencing of various environmental samples icluding those from oceans has been made, large numbers of fragment sequences have been registered in DNA databanks without information on gene function and phylotype, and thus with limited usefulness. Scientific and industrial useful 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 short segments. Therefore, even when a set of genes responsible for a scientifically and/or industrially useful function is found in one metagenome library, it is 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 self-organization of sequences according to genomes. Because BLSOM could reassociate genomic fragments according to genome, BLSOM should ameliorate the abovementioned weakness of metagenome analyses. Here, we developed a strategy for clustering of metagenomic sequences according to phylotypes and genomes, by testing a gene set for a metabolic pathway contributing to environment preservation.

Keywords: batch learning SOM, oligonucleotide frequency, protein function, metagenomics

1. 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, which are culture-independent approaches performing shotgun sequencing on mixed genome DNA samples, have been developed, and vast numbers of fragment sequences have been deposited in the International Nucleotide Sequence Databases (INSD). 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 is because orthologous sequence sets, which cover a broad phylogenetic range needed 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 simple 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 thus are called "genome signature". Phylogenetic clustering and classification in the present study is designed as an extension of the single parameter "%GC" to the multiple parameters "oligonucleotide frequencies".

We previously modified the SOM developed by Kohonen's group [1-3] for genome informatics on the basis of batch-learning SOM (BLSOM), which makes the learning process and resulting map independent of the order of data input [4-6]. The BLSOM thus developed could recognize phylotype-specific characteristics of oligonucleotide frequencies in a wide range of genomes and permitted clustering (self-organization) of genomic fragments according to phylotypes with neither the orthologous sequence set nor the troublesome and potentially mistakable processes of sequence alignment. Furthermore, the BLSOM was suitable for actualizing high-performance parallel-computing with the high-performance supercomputer "the Earth Simulator", and permitted clustering (self-organization) of almost all genomic sequences available in the International DNA Databanks on a single map [7-9]. By focusing on the frequencies of oligonucleotides (e.g., tetranucleotides), the BLSOM allowed highly accurate classification (self-organization) of most genomic sequence fragments on a species basis without providing species-related

information during BLSOM computation. Therefore, the present unsupervised and alignment-free clustering method should be most suitable for phylogenetic clustering of sequences from novel unknown organisms [10-12]. We have employed BLSOM for metagenomic studies on a large amount of environmental sequences, in joint research with experimental research groups analyzing various environmental and clinical samples [10,11].

Biological activity with industrial usefulness, such as processes responsible for environmental clean-up and preservation, is often carried out by a set of genes rather than a single gene; e.g., an operon responsible for one metabolic activity. However, in the case of metagenomic approaches, contigs with a significant length, such as those covering an operon, were obtained only for very dominant species after assembling of sequences. Even 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 has 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 having a full set or, at least, a major portion of the gene set. In the present report, we explained 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.

2. Methods

Genomic fragment sequences derived from metagenome analyses 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 were segmented into the window size, and the residual sequences, which were shorter than the window size, were omitted from the BLSOM analysis.

BLSOM learning was conducted as described previously [4-6], and the BLSOM program was obtained from UNTROD Inc. (y_wada@nagahama-i-bio.ac.jp).

3. Results

3.1 BLSOMs for sequences obtained by metagenome analyses

To test the clustering power of BLSOM for oligonucleotide composition in metagenomic sequences, we analyzed a large quantity of fragment sequences obtained from eight typical metagenomic libraries currently available; for details about metagenomic libraries, refer to [12]. To develop an informatics strategy useful to search gene candidates contributing to environmental cleanup and preservation, we focused on

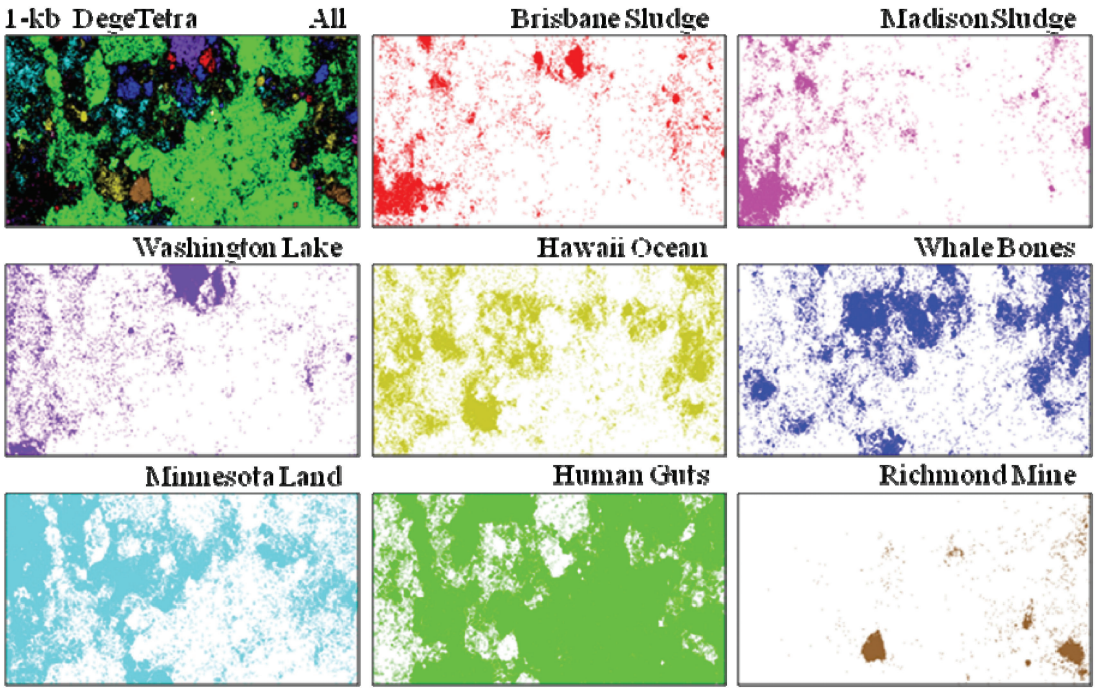

Fig. 1 DegeTetra-BLSOM for 1-kb metagenomic sequences of 8 environmental samples. Lattice points that include sequences from more than one environmental sample are indicated in black, and those containing sequences only from one sample are indicated in color as follows: Brisbane Active Sludge (■), Madison 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.

metagenomic sequences longer than 1 kb, which likely harbored an intact protein-coding sequence. 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. 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, which primarily reflect a transcriptional polarity for individual genes [6]. When we constructed BLSOM in which the frequencies of a pair of complementary oligonucleotides (e.g., AACC and GGTT) in each fragment were summed, the tendency of the splitting into a few territories was diminished for most of species [7]. For phylogenetic clustering of metagenomic sequences, it is unnecessary to know the transcriptional polarity for 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 result of DegeTetra-BLSOM for 1- or 2-kb sequence fragments (i.e., a window size of 1 or 2 kb) was listed in Figs. 1 or 2, respectively; DegeTri-BLSOM gave a similar result [12]. Sequences longer than 2 kb (approximately 15% of the sequences longer than 1 kb) should represent contig sequences obtained by the assembling process of shot-gun sequencing, and therefore, 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" panel 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. 1 and 2, 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 thus 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. On the two BLSOMs listed in Figs. 1 and 2, 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" panel, 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 was much simpler on both 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

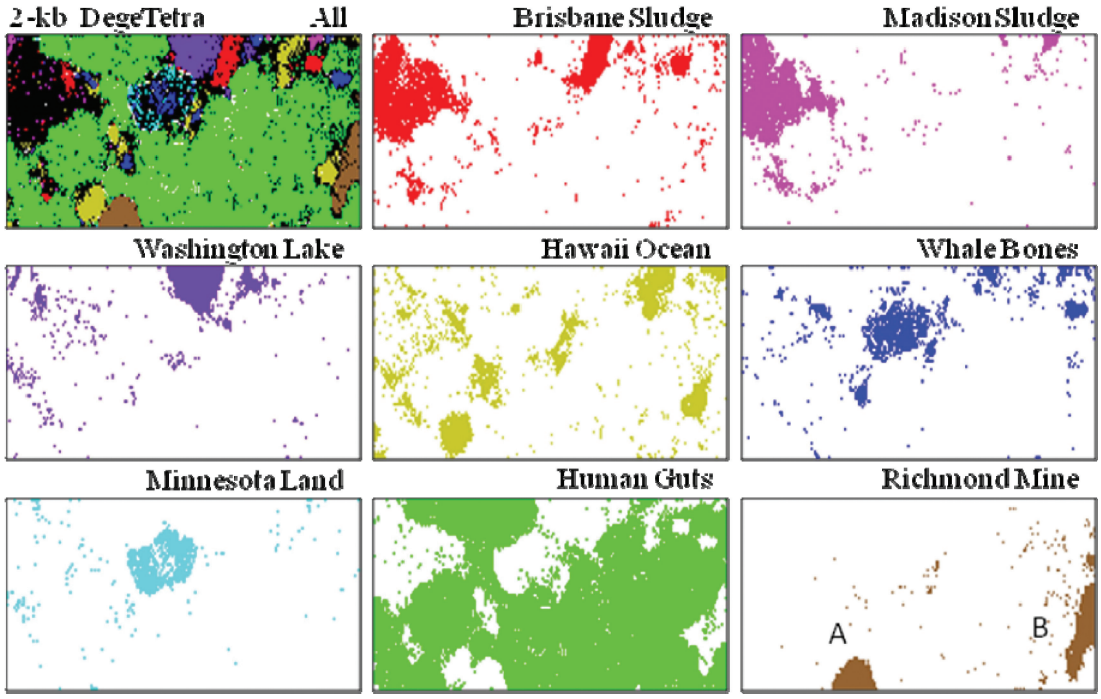

Fig. 2 DegeTetra-BLSOM for 2-kb metagenomic sequences of 8 environmental samples. Lattice points are indicated in color as described in Fig. 1.

Bones or the Minnesota Farm Land were distributed widely on the 1-kb BLSOM, but were much localized on the 2-kb BLSOM. This indicated that these samples contained a wide variety of genomes but phylotypes of dominant species were rather limited. In the case 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 ALL panels in Figs. 1 and 2), indicating that the microbial community in the human body environment differed significantly from that of natural environments.

3.2 Reassociation of genomic fragments according to phylotypes and genomes

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. [13] 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 [13]. In the "Richmond Mine" panel of 2-kb DegeTetra BLSOMs (Fig. 2), 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 territories, 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, but a major portion of the sequences in the extended B territory showed the highest similarity to sequences from Ferroplasma. This finding obtained by BLSOM supported the view that BLSOM has a potentiality for reassociating genomic fragments in a metagenome library according to genome, even in the presence of a massive quantity of sequences derived from a wide variety of genomes.

For the case of environmental sequences with a low genome complexity such as those from biofilms in the acid mine, reassociation of the metagenomic sequences according to genome can be obtained for a dominant species, by constructing one complete genome after sequence-assembling with conventional sequence homology searches. This reassociation, however, becomes increasingly difficult, when subdominant or minor populations are concerned. If 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 template 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 that inevitably dependents on a template genome is apparently inappropriate. In the case of BLSOM, reassociation (self-organization) of genomic fragments according to genome can be attained without the template genome, showing its wide applicability.

3.3 Genes contributing to environmental preservation

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 those constituting an operon. However, contigs with a significant length, e.g., 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 has 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. To test a feasibility of the abovementioned informatics strategy, we searched gene candidates useful for environmental preservation such as degradation of hazardous compounds. As a model case, we chose a set of ten genes, which are responsible for the metabolic activity of PCB degradation: *bphA1, A2, A3, A4, B, C, D, E, F, and G*. Using amino-acid sequences of the ten enzymes as queries, we searched the candidate genes of interest from the present metagenomic sequences derived from the eight environmental samples, with the DDBJ-tBLASTn search under a strict criterion (e values less than 1e-20); for details, refer to [12]. Candidate genes representing nine and eight out of the ten enzyme genes were found in the Washington Lake and Hawaii Ocean samples, respectively. 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. In usual metagenomic data, however, a coverage density by fragmental sequences may not reach to a level that completely covers a certain genome. In a practical experimental 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 of the pathway genes is found in a certain environmental sample, a larger scale of metagenomic sequencing, which

may completely cover the possible candidate genome, will be conducted on the environmental sample. 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 two environmental samples (Washington Lake and Hawaii Ocean samples) having a major part of the degradation pathway.

We next identified lattice points, on which the metagenomic sequences having the gene candidates for the PCB-degradation pathway were mapped, by finding the lattice point that had the minimum Euclidean distance in the multidimensional space for each candidate sequence. Lattices points containing the candidate gene sequences were widely scattered in the Washington Lake sample (Fig. 3A), indicating that the candidate sequences were presumably derived from various genomes. In contrast, 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 out of the initial eight genes, showing a powerful function of BLSOM for identifying potentially useful genome resources. In the "All" panel 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 present in this territory were derived from the species specifically present in the Hawaii Ocean, rather than those present in ubiquitous environments.

4. Discussion and Perspective

We established a method for identifying potentially useful genome resources from environments. It should be mentioned here that 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 might be composed of closely related, multiple genomes. We have recently developed a wide applicable strategy, which could separate metagenomic sequences present in one compact territory (e.g., that found in the Hawaii Ocean) according to phylotypes and hopefully to species, without effects of other metagenomic sequences coexisting [12]. In the strategy, random sequences with nearly the same mono-, di- or trinucleotide composition to each metagenomic sequence in the compact territory were generated [14]. Then, we constructed DegeTetra-BLSOM for the metagenomic sequences plus the random sequences. In the presence of the random sequences, two major compact territories, a few small territories and many scattered points were separately observed, which were surrounded by random sequences [12]. In one of the major territory, a major portion of PCB-degradation pathway genes was found, showing the usefulness of the strategy to specify a genome resource that should have a major portion (if not all) of the metabolic pathway genes.

Concerning a set of genes responsible for a certain metabolic pathway derived from one genome, homology search methods can provide information only if a very large amount of metagenomic sequences is available for constructing a nearly complete genome by their assembling or if there is a template genome sequence that is derived from a closely related species with the respective environmental one and thus is usable for mapping of metagenomic sequences. In contrast, the present *in-silico* association with BLSOM can be achieved without a template genome for mapping and thus is applicable to the really novel, environmental genomes.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C, 20510194) and for Young Scientists (B, 20700273)

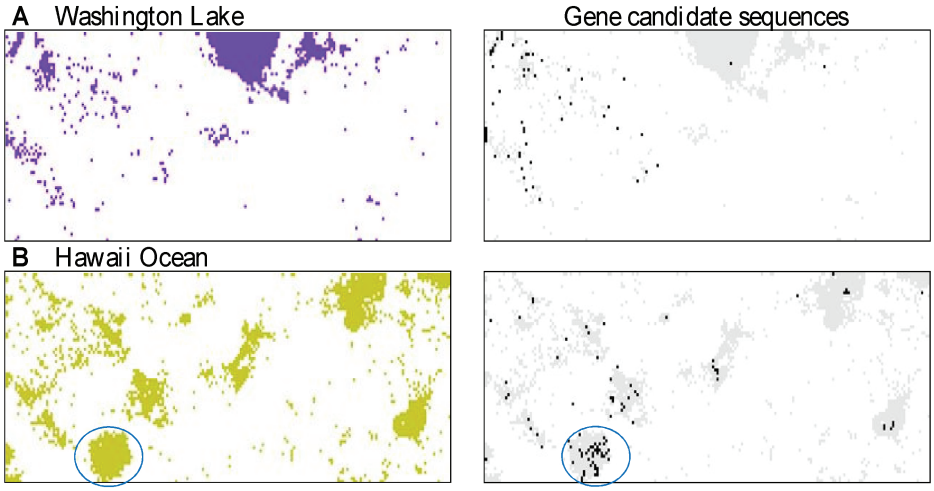

Fig. 3 Lattice points containing sequences harboring gene candidates for PCB-degradation pathway. (A), (B): 2-kb DegeTetra-BLSOM for Washington Lake or Hawaii Ocean listed in Fig. 2, respectively. In the right-side panel, lattice points containing sequences harboring the gene candidate sequences were indicated by dots. A compact territory in the Hawaii Ocean containing a cluster of gene candidate sequences was marked with a circle.

from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The present computation was done with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] T. Kohonen, "The self-organizing map", Proceedings of the IEEE, vol. 78, pp. 1464-1480, 1990.

[2] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas. "Engineering applications of the self-organizing map", Proceedings of the IEEE, vol. 84, pp. 1358-1384, 1996.

[3] T. Kohonen, Self-Organizing Maps. Berlin, Springer, 1997.

[4] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, and T. Ikemura, "Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome", Gene, vol. 276, pp.89-99, 2001.

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

[6] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures", Genome Res., vol. 13, pp. 693-702, 2003.

[7] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples", DNA Res., vol. 12, pp. 281-290, 2005.

[8] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, "Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes", Gene, vol. 365, pp. 27-34, 2006.

[9] T. Abe, H. Sugawara, S. Kanaya, and T. Ikemura, "Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator", Journal of the Earth Simulator, vol. 6, pp.17-23, 2006.

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

[11] H. Hayashi, T. Abe, M. Sakamoto, H. Ohara, T. Ikemura, K. Sakka, and Y. Benno, "Direct cloning of genes encoding novel xylanases from human gut ", Can. J. Microbiol., vol. 51, pp. 251-259, 2005.

[12] H. Uehara, Y. Iwasaki, C. Wada, T. Ikemura, and T. Abe, "A novel bioinformatics strategy for searching industrially useful genome resources from metagenomic sequence libraries", Genes Genet. Sys., vol. 86, pp. 53-66, 2011.

[13] G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen et al, "Community structure and metabolism through reconstruction of microbial genomes from the environment", Nature, vol. 428, pp. 37-43, 2004.

[14] T. Abe, K. Wada, Y. Iwasaki, and T. Ikemura, "Novel bioinformatics for inter- and intraspecies comparison of genome signatures in plant genomes", Plant Biotechnology, vol. 26, pp. 469-477, 2009.

ES を用いた大規模ゲノム・プロテオミクス解析：多様な環境由来の大量メタゲノム配列から、重要な代謝経路を保持する有用ゲノム資源を探索する新規戦略

プロジェクト責任者
池村　淑道　　長浜バイオ大学　バイオサイエンス学部
著者
阿部　貴志 *1，上原　啓史 *1，岩崎　裕基 *1，和田健之介 *1，池村　淑道 *1
*1 長浜バイオ大学　バイオサイエンス学部

海洋を代表例とする、多様な地球環境で生育する微生物類は培養が困難なため膨大なゲノム資源が未開拓・未利用に残されてきた。環境中の生物群集から培養せずにゲノム混合物を回収し、断片ゲノム配列を解読し有用遺伝子を探索する「メタゲノム解析法」が開発され、科学的・産業的に注目を集めている。我々が開発した一括学習型自己組織化マップ法（BLSOM）は、断片ゲノム配列を生物種ごとに高精度に分離（自己組織化）する能力を持つ。既知の全生物種由来のゲノム断片配列を一枚の BLSOM 上で俯瞰し、特徴抽出をする事も可能である。

国際 DNA データベースに収録されている、代表的な 8 環境由来の混合ゲノム試料に関しての、大規模メタゲノム解析で得られた大量塩基断片配列を対象に、3 連並びに 4 連塩基頻度の BLSOM 解析を行った所、各環境に特異的に存在するゲノム類に由来する配列を特定出来た。メタゲノム解析では、一つの代謝系を構成する酵素群の遺伝子が、オペロンを構成して近傍に位置していても、断片配列の解読の過程で、泣き別れを起こしてしまう。BLSOM はこれらの泣き別れを起こした配列を、*in silico* で再集合させる能力を持つ。環境汚染物質の分解に関与する代謝系を構成する酵素群（例えば、PCB 分解に関与する 10 酵素）をモデル系として、その大半を持つゲノムを探索する目的の BLSOM 法を確立した。

広範囲のゲノムが解読された結果、アミノ酸配列の相同性検索では機能が推定できない、機能未知なタンパク質が大量に蓄積し、産業的にも未利用なまま残されてきた。オリゴペプチド頻度の BLSOM を用いれば、これらの大量なタンパク質類の機能推定が可能である。

キーワード: 自己組織化マップ, BLSOM, 環境微生物, オリゴヌクレオチド, 環境浄化, メタゲノム解析