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 Graduate school of Science and Technology, Niigata University
Nagahama Institute of Bio-Science and Technology
Kennosuke Wada Nagahama Institute of Bio-Science and Technology
Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

Self-Organizing Map (SOM) developed by Kohonen is an effective tool for clustering and visualizing high-dimensional complex data on a two-dimensional map. We previously modified the conventional Self-Organizing Map (SOM) to genome and protein informatics, making the learning process and resulting map independent of the order of data input. BLSOM thus developed on the basis of batch-learning SOM became suitable for actualizing high-performance parallel-computing, and revealed species-specific characteristics of oligonucleotides (e.g., tetranucleotides) frequencies in individual genomes, permitting clustering (self-organization) of genomic fragments (e.g., 5 kb or less) according to species without species information during the calculation. Using ES, we established the alignment-free clustering method BLSOM that could analyze far more than 10,000,000 sequences simultaneously. Sequence fragments from almost all prokaryotic, eukaryotic, and viral genomes currently available could be classified (self-organized) according to phylotypes on a single two-dimensional map. Here, we have further developed a strategy for predicting phylotypes of a massive amount of genomic fragments obtained by metagenome analyses, by mapping of each metagenomic sequence on a large-scale BLSOM that was constructed for almost all genomic sequences currently available from the International Nucleotide Sequence Databases.

Keywords: batch learning SOM, oligonucleotide frequency, phylogenetic classification, metagenomics

1. Introduction

Although remarkable progress in metagenomic sequencing of various environmental samples including the most extensively studied samples from oceans has been made, huge numbers of fragment sequences are registered without information on gene function and phylotype in International Nucleotide Sequence Databases, and thus with 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 required 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 has been called the “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”. For this purpose, 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 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 Nucleotide Sequence Databases on a single map [7-9]. By focusing on the frequencies of oligonucleotides (e.g., tetranucleotides), the BLSOM has allowed highly accurate classification (self-

organization) of most genomic sequence fragments on a species basis without providing species-related information during BLSOM computation. The present unsupervised and alignment-free clustering method is thought to be the most suitable one for phylogenetic estimation for a massive amount of genomic fragments obtained by metagenome analyses. This was done by mapping of each metagenomic sequence on the large-scale BLSOM constructed with ES, which could classify almost all known prokaryotic, eukaryotic, and viral sequences according to phylotypes [10-12]. We have already employed the BLSOM method for studies of environmental genomic fragments in joint research with experimental research groups analyzing various environmental and clinical samples [10, 11].

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 (1 kb) were segmented into the window size, and the residual sequences, which were shorter than the window size, were omitted from the BLSOM analysis.

Using a high-performance supercomputer “the Earth Simulator”, we could analyze almost all 5-kb genomic sequences derived from 5600 prokaryotes, 411 eukaryotes, 31486 viruses, 4479 mitochondria, and 225 chloroplasts, which were obtained from the recent version released from the International Nucleotide Sequence Databanks. The 5600 prokaryotes were selected because at least 10-kb genomic sequences were registered in the Databases. One important target of the phylogenetic classification of metagenome sequences is the sequences derived from species-unknown novel microorganisms. To keep good resolution for microorganism sequences on the BLSOM, it is necessary to avoid excess representation of sequences derived from higher eukaryotes with large genomes. Therefore, in the cases of higher eukaryotes, 5-kb sequences were selected randomly from each large genome up to 200 Mb in Fig. 1A. In this way, the total quantities of prokaryotic and eukaryotic sequences were made almost equal. The separation between eukaryotes and prokaryotes was achieved with a high accuracy of 95%; the separation between organelles and viruses and between nuclear and viral genomes was also achieved with an accuracy of approximately 80%. Clear separation of the species-known prokaryote sequences into 38 major families was also observed (refer to Fig. 2A). The separation of eukaryotic sequences according to families was also observed on this 5-kb BLSOM (data not shown). During BLSOM computation, no information was given to the computer regarding which species each sequence fragment belonged to (unsupervised learning algorithm).

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 A large-scale BLSOM constructed with all sequences available from species-known genomes

In the present study, we examined the BLSOM ability for phylogenetic separation of prokaryotic sequences and applyed the BLSOM to the phylogenetic prediction of sequences obtained by metagenome analyses. Large-scale metagenomic studies of uncultivable microorganisms in environmental and clinical samples have recently been conducted to survey genes useful in industrial and medical applications and to assist in developing accurate views of the ecology of uncultivable microorganisms in each environment. Conventional methods of phylogenetic classification of gene/genomic sequences have been based on sequence homology searches and therefore, the phylogenetic studies focused inevitably on well-characterized gene sequences, for which orthologous sequences from a wide range of phylotypes are available for constructing a reliable phylogenetic tree. The well-characterized genes, however, often are not industrially attractive. It would be best if microbial diversity and ecology could be assessed during the process of screening for novel genes with industrial and scientific significance. The present unsupervised and alignment-free clustering method, BLSOM, is thought to be the most suitable one for this purpose because there was no need of orthologous sequence sets [4-6].

Metagenomic analyses can be applied to not only environmental but also medical samples such as clinical samples, and therefore, are applicable to exploring unknown pathogenic microorganisms that cause novel infectious diseases. It should be noted that the mixed genome samples in the medical and pharmaceutical fields may contain DNA from a wide range of eukaryotes, as well as from humans. Therefore, when we consider phylogenetic classification of genomic sequences derived from species-unknown environmental microorganisms obtained by metagenome studies, it is necessary to construct BLSOM in advance with all available sequences from species-known prokaryotes, eukaryotes, viruses and organelles compiled in the International DNA Databanks. According to our previous studies of metagenome sequences [7], BLSOM was constructed with oligonucleotide frequencies in 5-kb sequence fragments. In DNA databases, only one strand of a pair of complementary sequences is registered. Our previous analyses revealed that sequence fragments from a single prokaryotic genome are often split into two territories that reflect the transcriptional polarities of the genes present in the fragment [7]. For phylotype

classification of sequences from uncultured microbes, it is not required to know the transcriptional polarity of the sequence, and the split into two territories complicates assignment to species. Therefore, we previously introduced a BLSOM in which frequencies of a pair of complementary oligonucleotides (e.g., AACC and GGTT) were summed, and the BLSOM for the degenerate sets of tetranucleotides were designated as DegeTetra-BLSOM.

3.2 Phylogenetic estimation for environmental DNA sequences and microbial community comparison

More than 17 million genomic sequence fragments obtained from various environments through metagenomic analyses have been registered in the International Nucleotide Sequence Databases. A major portion of them is novel but at present moment have a limited utility because of lacks of phylogenetic and functional annotations. The phylogeny estimation of genomic sequence fragments of novel microorganisms requires the feature extraction of oligonucleotide composition of all species-known microorganism genomes, in advance on BLSOM. Therefore, a large-scale BLSOM (Fig. 1A) covering all known sequences, including those of viruses, mitochondria, chloroplasts, and plasmids, was constructed as described in **Methods**. Next on this BLSOM, numerous sequence fragments derived from an environmental sample were mapped; i.e., the similarity of the oligonucleotide frequency in fragmental sequences from environmental samples with that of sequences from species-known genomes was examined. In Fig. 1B, 210 thousand sequences with a fragment size of 1 kb or more, which were collected from the Sargasso Sea near Bermuda [12], were mapped (Fig. 1B). This analysis of all sequence fragments obtained from one environmental sample can estimate numbers and proportions of species present in the sample. Approximately 70% of sequences from the Sargasso Sea were mapped to the prokaryotic territories, and the rest was mapped to the eukaryotic, viral or organelle territories.

To further identify more detailed phylogenies of the environmental sequences thus mapped to the prokaryotic

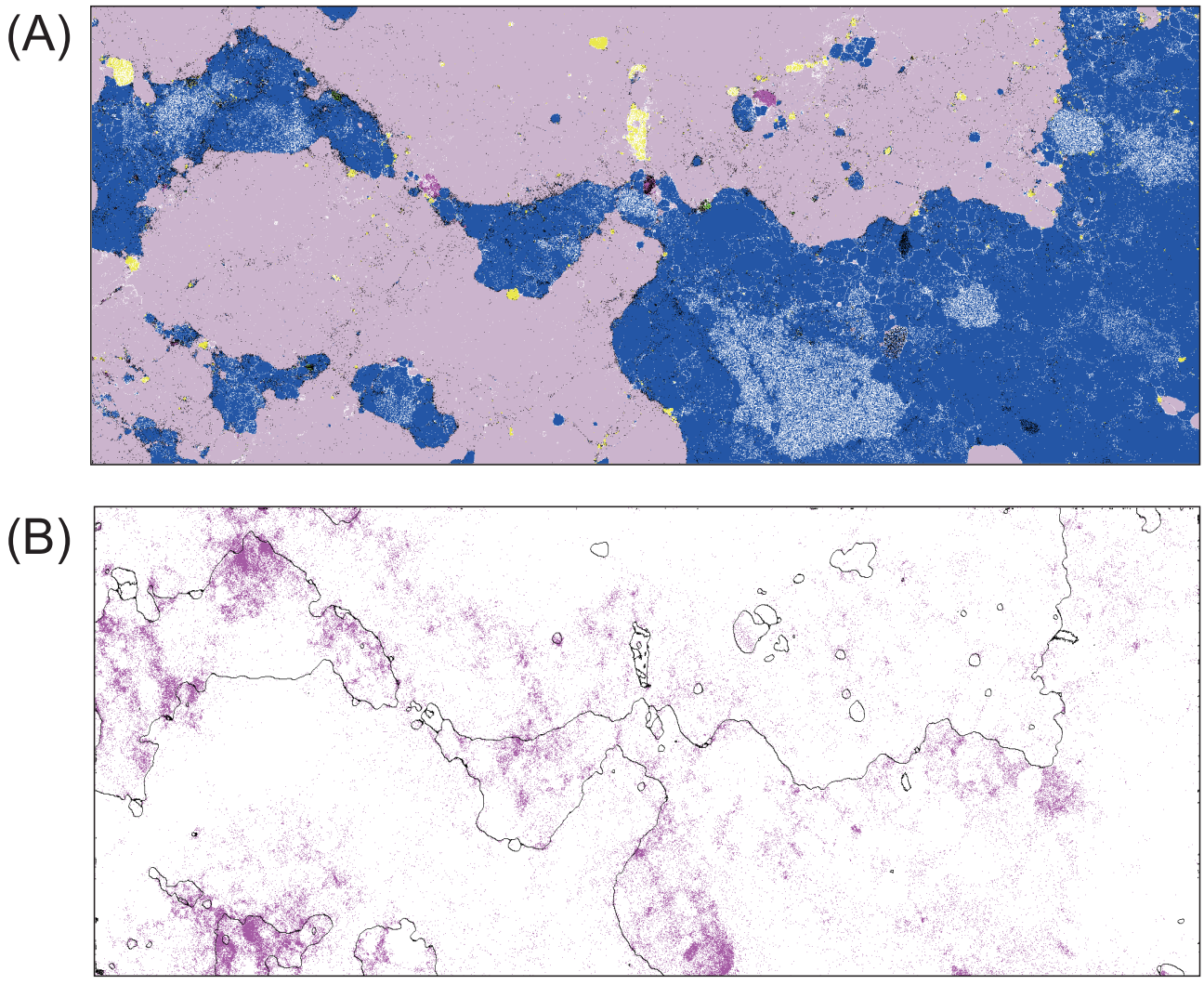

Fig. 1 BLSOM for phylogenetic classification of environmental sequence. (A) DegeTetra-BLSOM of 5-kb sequences derived from species-known 5,600 prokaryotes, 411 eukaryotes, 1,728 mitochondria, 225 chloroplasts, and 31,486 viruses. (B) Sargasso sequences longer than 1 kb were mapped on the 5-kb DegeTetra-BLSOM, after normalization of the sequence length.

territories, a BLSOM analyzing 5-kb genomic sequence fragments only from 5,600 known prokaryotes was constructed for degenerate tetranucleotide composition (Fig. 2A). On this species-known prokaryote BLSOM, the separation into 38 phylogenetic groups was examined, revealing that 85% of the sequences were properly clustered according to the phylogenetic group resulting in formation of the phylogenetic group-specific territory. The reason why 100 % separation was not achieved was mainly because of horizontal gene transfer between the genomes of different microbial species [6, 7]. The 140 thousand metagenomic sequences from the Sargasso Sea that were mapped previously to the prokaryotic territories (Fig. 1B) were remapped on the BLSOM that was constructed for sequences derived only from species known-prokaryotes, in order to get the detailed prokaryotic phylotype assignment. The mapped sequences broadly spread across the BLSOM, demonstrating that the sequences belonged to a wide range of phylotypes (Fig. 2B). Interestingly, there were areas where metagenomic sequences were densely mapped, which may indicate dominant species/genera. In sum, the estimation of prokaryotic phylogenetic groups could provide phylogenetic information for almost half of sequence fragments from the Sargasso Sea. The present procedure can be used to estimate the phylogenetic structures of microbial communities living in a subject environment and thus to understand the diversity of microbial floras.

Through successive mapping of the subject sequences on a BLSOM created with the sequences from species-known genomes of a much restricted phylogenetic group, further detailed phylogenetic estimation, such as at the genus or species level, becomes possible. In other words, this stepwise tracking allowed us to estimate detailed structures of microbial communities present in an environmental sample, to determine the novelty of obtained environmental sequences at various phylogenetic levels and to find the novel sequences efficiently.

Recent metagenomic analyses showed an abundance of viruses in seawater [13]. Since virus genomes contain no rDNA, conventional methods of phylogenetic estimation based on

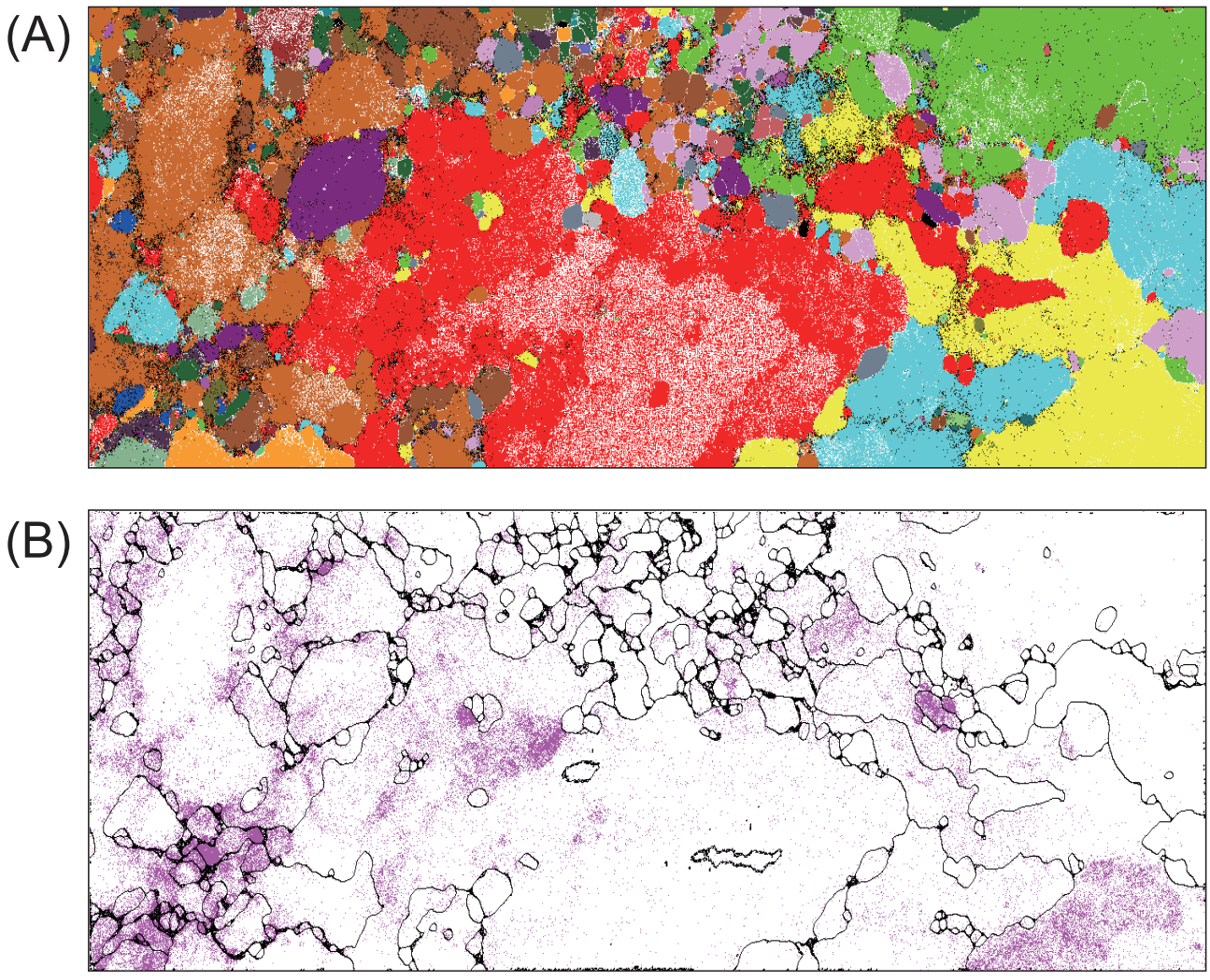

Fig. 2 Phylogenetic classification of sequences from an environmental sample. (A) DegeTetra-BLSOM of 5-kb sequences derived from species-known 5,600 prokaryotes. (B) Sargasso sequences that were classified into prokaryotic territories in Fig. 1B were mapped on the 5-kb DegeTetra-BLSOM constructed with the sequences only from the species-known 5,600 prokaryotes.

rRNA sequences cannot be used. BLSOM analysis for fully sequenced virus genomes showed a separation according to their phylogenies, allowing us to conduct phylogenetic estimation in the viral kingdom without relying on orthologous sequence sets or sequence alignments. The publication of a large-scale BLSOM, which separated all available genomic sequences including viral and organelle sequences, will provide fundamental data required in large-scale metagenomic studies. The BLSOM is applicable to a broad range of life sciences, such as medical and pharmaceutical sciences, and related industrial fields.

The mapping of novel sequences on the large-scale BLSOM that was constructed with ES can be performed using a PC-level computer; our group has made a PC software program for the BLSOM mapping. This software can be freely available at http://bioinfo.ie.niigata-u.ac.jp.

4. Conclusion and Perspective

Large-scale metagenomic analyses using recently released next-generation sequencers are actively underway on a global basis, and the obtained numerous environmental sequences have been registered in the public databases. Large-scale computations using various, novel bioinformatics tools are undoubtedly needed for knowledge-findings from the massive amount of sequence data. The present BLSOM is an unsupervised algorithm that can separate most sequence fragments based only on the similarity of oligonucleotide frequencies. Unlike the conventional phylogenetic estimation methods, the BLSOM requires no orthologous sequence set or sequence alignment, and therefore, is suitable for phylogenetic estimation for novel gene sequences. It can also be used to visualize an environmental microbial community on a plane and to accurately compare it between different environments.

Because BLSOM could reassociate genomic fragments according to genomes, we have recently developed a widely applicable method, which could separate metagenomic sequences according to phylotypes and hopefully to species, without coexistence of sequences from species-known prokaryotes or metagenomic sequences from other environmental samples [14]. In that study, random sequences with nearly the same mono-, di- or trinucleotide composition to each metagenomic sequence in one environmental sample of interest were generated. Then, we constructed DegeTetra-BLSOM for the metagenomic sequences plus the computer-generated random sequences. Under the presence of the random sequences, metagenomic sequences formed many well-separated territories surrounded with the random sequences. In that study, a major portion of PCB degradation enzyme genes was found within one clear territory, indicating that this territory contained the genome harboring the metabolic pathway genes for PCB degradation. This is another application of BLSOM for metagenome analyses. In the cases of conventional methods, which were based on sequence homology searches, information concerning a gene set of one species responsible for a certain biological activity can be obtained only when a very large amount of sequences is available for constructing a nearly complete genome by sequence assembling or when a complete genome sequence is available as a reliable template for mapping of metagenomic sequences. In contrast, the BLSOM method can be achieved without a template genome for mapping and thus is applicable to the really novel, environmental genomes.

BLSOM can also be used to predict functions of proteins, for which the sequence similarity search at an amino-acid level cannot predict functions. This is because proteins with the same or similar functions can be clustered (self-organized) primarily according to functions on BLSOM for oligopeptide composition [15].

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C, 23500371) and for Young Scientists (B, 23710242) 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] J. C. Venter et al., “Environmental genome shotgun sequencing of the Sargasso Sea”, Science, vol. 304, pp. 66-74, 2004.

[13] R. A. Edward and F. Rohwer, “Viral metagenomics”. Nat Rev Microbiol, vol. 3, pp. 504-10, 2005.

[14] 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.

[15] Takashi Abe, Shigehiko Kanaya, Hiroshi Uehara, and Toshimichi Ikemura, “A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses”, DNA Research, vol. 16, pp. 287-298, 2009.

全ゲノム・全タンパク質配列の自己組織化マップを用いた大規模ポストゲノム解析

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　　新潟大学　大学院自然科学研究科
　　長浜バイオ大学　バイオサイエンス学部
和田健之介　　長浜バイオ大学　バイオサイエンス学部
池村　淑道　　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、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）する能力を持つ知見を元に、メタゲノム配列に対する系統推定法の開発を行なってきた。

メタゲノム解析により取得され断片ゲノム配列には、原核生物のみならず、真核生物やウイルス由来の断片ゲノム配列も豊富に含まれている。これら断片ゲノム配列からの真核生物やウイルスを探索するために、断片化サイズ 5kb での 4 連続塩基頻度にて、既知真核生物 412 種（ゲノム配列断片数：618 万件）、既知ウイルス 30,000 種（12 万件）、原核生物 1294 属（488 万件）、ミトコンドリア 4479 種（2 万件）、葉緑体 225 種（6 千件）のゲノム配列断片 1120 万件（56 ギガ塩基）の全生物種を対象にした大規模 BLSOM 解析を行なった。また、上記解析で使用した真核生物、原核生物、ウイルスについて、各生物カテゴリに推定された配列群に対し、より詳細な系統推定を行うことを目的に、断片サイズ 5kb での 4 連続頻度での BLSOM 解析を行なった。

タンパク質の機能推定にも BLSOM が有用な事を見出している。

既に 8 大学 11 グループ、6 公的研究機関 7 グループ、4 民間企業との共同研究を行っており、我が国で行われている大規模メタゲノム解析の大半で活用されている。世界の研究グループにとって必須の技術とするべく、本研究成果より得られた大規模 BLSOM 解析結果を用いたメタゲノム配列に対する系統推定を行うためのソフトウェアの公開を行っている（http://bioinfo.ie.niigata-u.ac.jp/）。

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頻度, 生物系統推定,
　　バイオインフォマティクス