A Large-Scale Self-Organizing Map for Metagenome Studies for Surveillance of Microbial Community Structures

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

Authors
Takashi Abe[*1], Yuki Iwasaki[*2], Kennosuke Wada[*2], Yoshiko Wada[*2] and Toshimichi Ikemura[*2]
*1 Department of Information Engineering, Faculty of Engineering, Niigata University
*2 Department of Bioscience, Nagahama Institute of Bio-Science and Technology

Metagenome analyses, which directly sequence mixed genomes of uncultured environmental microorganisms, have become widely used in earth sciences for clarifying a microbial community structure in an environmental ecosystem. We have previously developed an unsupervised clustering method "Batch-learning SOM: BLSOM" for oligonucleotide compositions in genomic fragments, which recognizes species-specific characteristics of oligonucleotide composition in individual genomes and can cluster tens millions of genomic sequences according to phylogenetic groups, solely depending on their oligonucleotide composition. For phylogenetic classification of metagenomic sequences obtained from environmental samples, we have annually updated a large-scale tetranucleotide BLSOM for all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles. By mapping metagenomic sequences from an environmental sample on this large-scale BLSOM constructed by ES, we can clarify the microbial community structure in the ecosystem. This strategy, however, cannot be properly applied to short sequences (e.g. those shorter than 100 bp) and thus sequences obtained with new generation sequencers widely used currently. To solve this issue, we have developed an oligonucleotide BLSOM for tRNA genes (tDNAs) since their lengths are mostly shorter than 90 bp and their sequences have been stably conserved during evolution. When constructing BLSOM with species-unknown tDNAs obtained from metagenomic sequences plus species-known microbial tDNAs (ca. 0.6 million tDNAs in total), a large portion of metagenomic tDNAs has self-organized with species-known tDNAs, giving information on microbial communities in environmental samples. BLSOM also allows us to identify tDNAs suitable as phylogenetic markers for rare phylotypes.

Keywords: batch learning SOM, metagenome, microbial community, phylogenetic marker, tRNA, oligonucleotide composition, new-generation sequencer

1. Introduction

Metagenome analyses, which directly sequence mixed genomes of uncultured environmental microorganisms, have been widely used in earth and environment sciences. A massive amount of environmental metagenomic sequences has been registered in the International DNA Sequence Databanks (DDBJ/EMBL/GenBank) but poorly characterized, especially in cases of short sequences obtained from new generation sequencers, and therefore, a large portion of short metagenomic sequences has been stored in a less useful manner. In more detail, while homology search for nucleotide and amino-acid sequences such as BLAST has widely been used for a basic bioinformatics tool for phylogenetic characterization of gene/protein sequences, phylotypes of most metagenomic sequences, especially of short sequences obtained by new generation sequencers, cannot be properly assigned. This is because reliable phylogenetic trees required for their phylogenetic assignment cannot be constructed due to their sequence novelty and short length. To solve this issue, we have developed an unsupervised clustering method "Batch-learning SOM: BLSOM" for oligonucleotide composition [1-4]. This unsupervised clustering method recognizes species-specific characteristics of oligonucleotide composition in genomic fragments of individual genomes, permitting clustering (self-organization) of fragment sequences according to species without need for species information during BLSOM calculation. Furthermore, this BLSOM is suitable for actualizing high-performance parallel-computing with the high-performance supercomputer, ES [3-5].

Because BLSOM with oligonucleotide (e.g. tetranucleotide) composition can cluster genomic fragments relatively short in length (e.g. 500-bp sequences) according to phylotype, this method has been successfully applied to the phylogenetic

classification of a massive amount of metagenomic sequences [4-7]. When considering phylogenetic classification of species-unknown sequences obtained from environmental samples, it is important to construct and update, at least once a year, a large-scale BLSOM for all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles. By mapping metagenomic sequences obtained from environmental samples on the newest BLSOM constructed with ES, we have clarified the microbial community structure in individual ecosystems analysed in Japan [7].

2. Materials and Methods

We have previously modified the conventional SOM for genome informatics on the basis of batch-learning SOM to make the learning process and resulting map independent of the order of data input [1,2]. The initial weight vectors were defined by PCA instead of random values, and genomic sequences were analysed as described previously [1-4].

We have constructed and updated a large-scale database for tRNA genes (tDNAs) obtained not only from the completely sequenced genomes but also from draft sequences of prokaryotic genomes in WGS (Whole Genome Shotgun) division in DDBJ/EMBL/GenBank. In accord with the remarkable progress of DNA sequencing technology, a vast quantity of metagenomic sequences obtained from a wide variety of environmental and clinical samples have been compiled in DDBJ/EMBL/GenBank, and short metagenomic sequences obtained even with new-generation sequencers contain a large number of full-length tRNAs because tRNA lengths are short. Therefore, tDNAs found from metagenomic sequences have also been included in the tRNA gene database [8-10]. To enhance the completeness and accuracy of tDNAs compiled in "tRNADB-CE" (http://trna.ie.niigata-u.ac.jp), three computer programs, tRNAscan-SE [11], ARAGORN [12], and tRNAfinder [13] were used in combination, since their algorithms partially differed and rendered somewhat different results. The tDNAs found concordantly by three programs were stored in tRNADB-CE after briefly checking anticodon sequences. Discordant cases among programs were manually checked by experts in tRNA experimental fields before inclusion into the database (8-10). In the newest version, approximately 0.4 million bacterial tDNAs obtained from more than 7000 complete or near complete genomes and approximately 0.2 million tDNAs obtained from more than 200 environmental samples have been compiled [10], and these tDNAs have been used in the present BLSOM analysis.

3. Results

3.1 Oligonucleotide BLSOMs for bacterial tDNAs

Because the annual updating of the large-scale tetranucleotide BLSOM with ES and its application to metagenome studies have been described in detail a few times in our previous annual reports, we will focus in this report on a newly developed BLSOM strategy suitable for phylogenetic analyses using short metagenomic sequences obtained from new generation sequencers, because of the following reason. Our previous oligonucleotide BLSOMs have shown that genomic sequences of 1 kb or longer can be clustered according to species with high accuracy and those of 300 bp with sufficient accuracy. However, this species-dependent clustering reduces evidently for sequences shorter than 100 bp, which are obtained mainly from new generation sequencers widely used. To solve this issue, we have focused on tRNA genes because their lengths are mostly less than 90 bp and their sequences have been stably conserved during evolution at least at a phylum level [9].

It should also be pointed out that oligonucleotides such as penta- and hexanucleotides often represent motif sequences responsible for sequence-specific protein binding (e.g. transcription factor binding). Occurrences of such motif oligonucleotides should differ from occurrences expected from the mononucleotide composition in the respective genome and may differ among genomic portions within a single genome. Actually, we have recently found that a pentanucleotide-BLSOM for the human genome can detect characteristic enrichment of many transcription-factor-binding motifs in pericentric heterochromatin regions [14]; i.e. BLSOM can effectively detect the characteristic and combinatorial occurrences of functional motif oligonucleotides in genomic sequences.

Each tRNA has characteristic and combinatorial occurrences of motif oligonucleotides, which are required for fulfilment of its function (e.g. binding to proper enzymes and other RNAs) and its structural formation (e.g. clover leaf), and these functionary important motif oligonucleotide are thought to be stably conserved during evolution. Taking these into account, we have constructed BLSOM for pentanucleotide compositions in prokaryotic tRNAs in tRNADB-CE (Fig. 1) as described previously [1,5]. Interestingly, the tRNAs are primarily separated according to amino acid, without giving information other than the oligonucleotide composition (manuscript in preparation); lattice points containing only tRNAs belonging to one amino acid are marked by the color representing the amino acid (Figs. 1A and 1B). This shows that the BLSOM can detect characteristic combinations of motif oligonucleotides required for proper recognition by various enzymes including aminoacyl-tRNA synthetases. The tDNAs for one amino acid form one or a few major territories and many tiny satellite-type spots (Fig. 1B). The number of tDNAs in each lattice point is important when considering biological significance of minor territories and tiny satellites, and thus the vertical bar in Fig. 1C presents the number of tDNAs for three examples of amino acids. Lattice points in major and minor territories of each amino acid apparently contain many tDNAs, and even some tiny satellites have multiple tDNAs (Fig. 1C). Satellites with multiple tDNAs may not represent improper cases, such

as those raised by DNA sequencing errors, but represent real tDNAs with certain nonstandard characteristics. The tDNAs belonging to a sharp peak in a satellite spot located away from the correspondent major territories have been found to primarily represent isoaccepting tDNAs (isoacceptors) of various species belonging to one phylogenetic family, which often differ in sequence for few bases. This finding indicates that the nonstandard-type tDNAs are candidates for molecular phylogenetic markers representing a specific phylotype; i.e. this BLSOM can provide a strategy to find phylogenetic marker tDNAs.

3.2 BLSOM for species-known plus species-unknown tDNAs

The tRNADB-CE has included tDNAs obtained from a massive amount of metagenomic sequences obtained from approximately 2000 environmental samples; tDNAs predicted concordantly by all three programs have been included and are abbreviated as metagenomic tDNAs. Since metagenomic sequences should be derived not only from bacteria, but also archaea and fungi, we have constructed BLSOM with species-unknown metagenomic tDNAs plus species-known bacterial, archaeal, and fungal tDNAs; 0.6 million tDNAs in total (Both in Fig. 2A). Species-unknown metagenomic and species-known microbial tDNAs are visualized separately in Metagenome and Known in Fig. 2A. Amino acid-dependent clustering is apparent, but their separation patterns are more complex than those for species-known bacterial tDNAs listed in Fig. 1A, and there are more black lattice points in Fig. 2A than in Fig. 1A. A major portion of black lattice points are observed for metagenomic tDNAs (Metagenome in Fig. 2A) while a minor portion is observed also for species-known tDNAs (Known in Fig. 2A). Detailed inspection of the species-known tDNAs belonging to black lattice points has revealed these to be primarily archaeal and fungal tDNAs, showing that BLSOM has separated archaeal and fungal tDNAs from bacterial tDNAs and that a significant level of metagenomic tDNAs should be derived from archaea and fungi.

The observation that a large portion of archaeal and fungal tDNAs are located in black lattice points in Fig. 2A indicates their self-organization to depend largely on their sequence characteristics distinct from bacterial tDNAs, rather than on distinctions between amino acids. To study amino acid-dependent clustering of archaeal and fungal tDNAs, BLSOMs

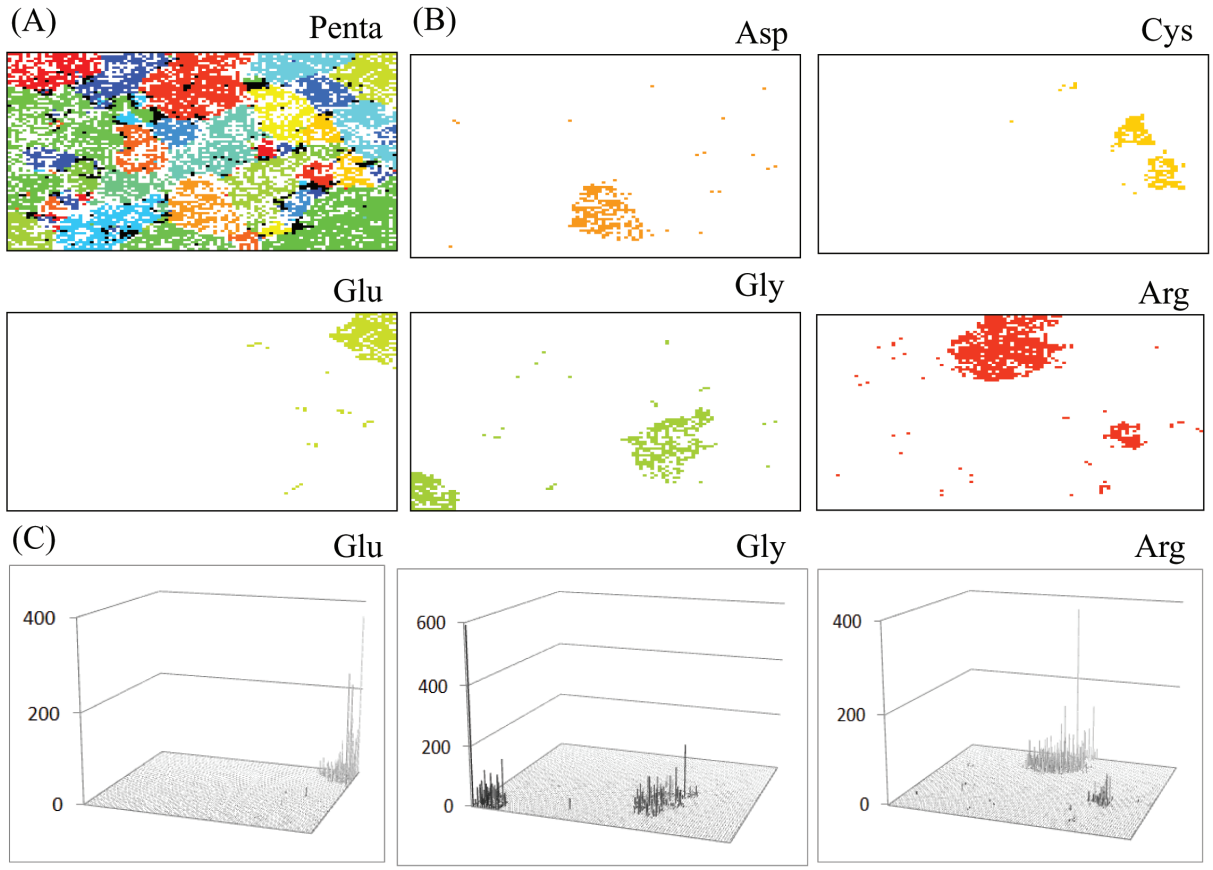

Fig. 1 Oligonucleotide-BLSOM for bacterial tDNAs. (A) BLSOM for tri-, tetra-, and pentanucleotide compositions (Tri-, Tetra-, and Penta). Lattice points containing tDNAs of multiple amino acids are indicated in black, and those containing tDNAs of a single amino acid are coloured as follows: Ala (■), Arg (■), Asn (■), Arp (■), Cys (■), Gln (■), Glu (■), Gly (■), His (■), Ile (■), Leu (■), Lys (■), Met (■), Phe (■), Pro (■), Ser (■), Thr (■), Trp (■), Tyr (■), and Val (■). (B) Lattice points containing tDNAs of four examples of amino acids on Penta in Fig. 1A are visualized separately with the colour used there. (C) Number of tDNAs in each lattice point on Penta is presented for individual amino acids by the height of the vertical bar. Lattice points containing multiple tDNAs, but not one or a few tDNAs, turn out to be detectable.

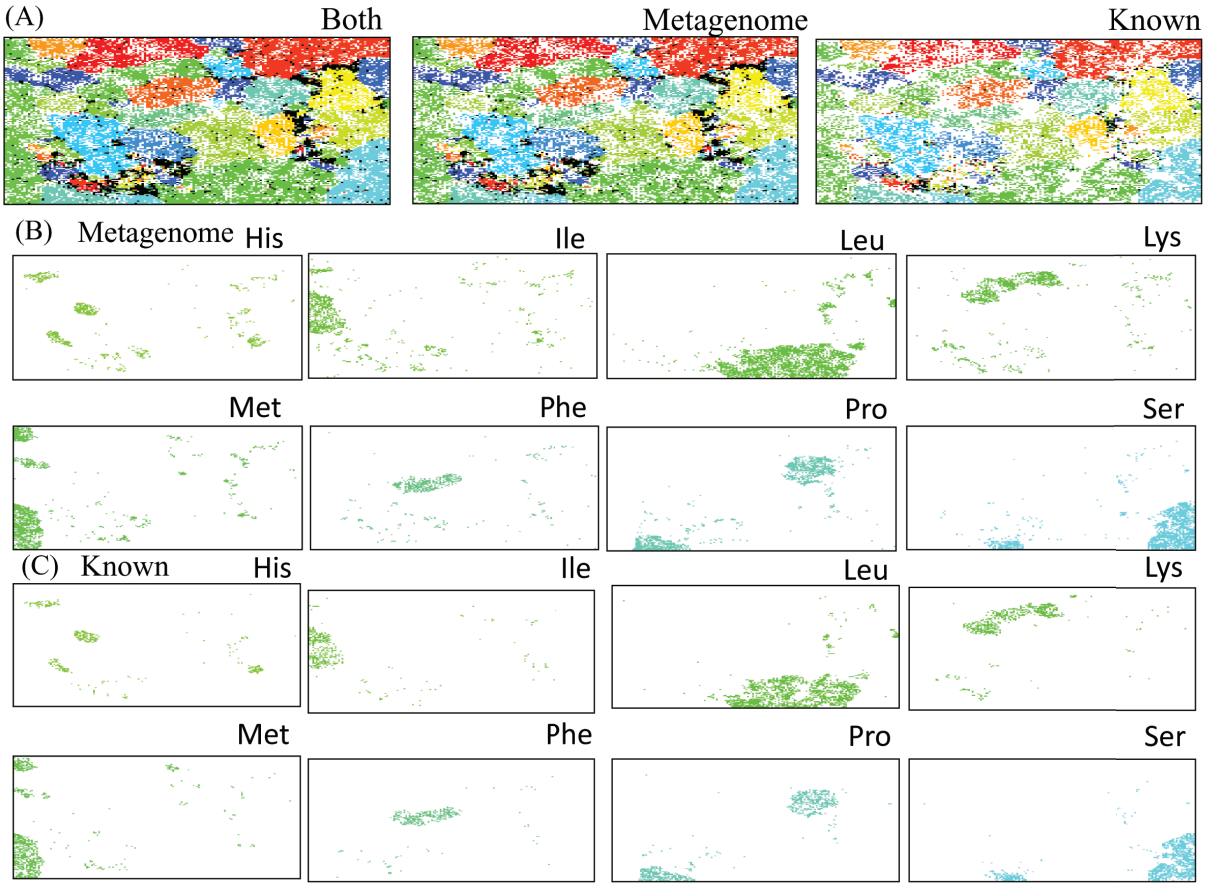

Fig. 2 BLSOM for metagenomic plus species-known microbial tDNAs. (A) Both; Lattice points are marked for both types of tDNAs as described in Fig. 1A. Metagenome or Known; lattice points containing only metagenomic or species-known microbial tDNAs are marked as described in Fig. 1A. (B,C) Lattice points containing metagenomic or species-known microbial tDNAs of individual amino acids are marked as described in Fig. 1B.

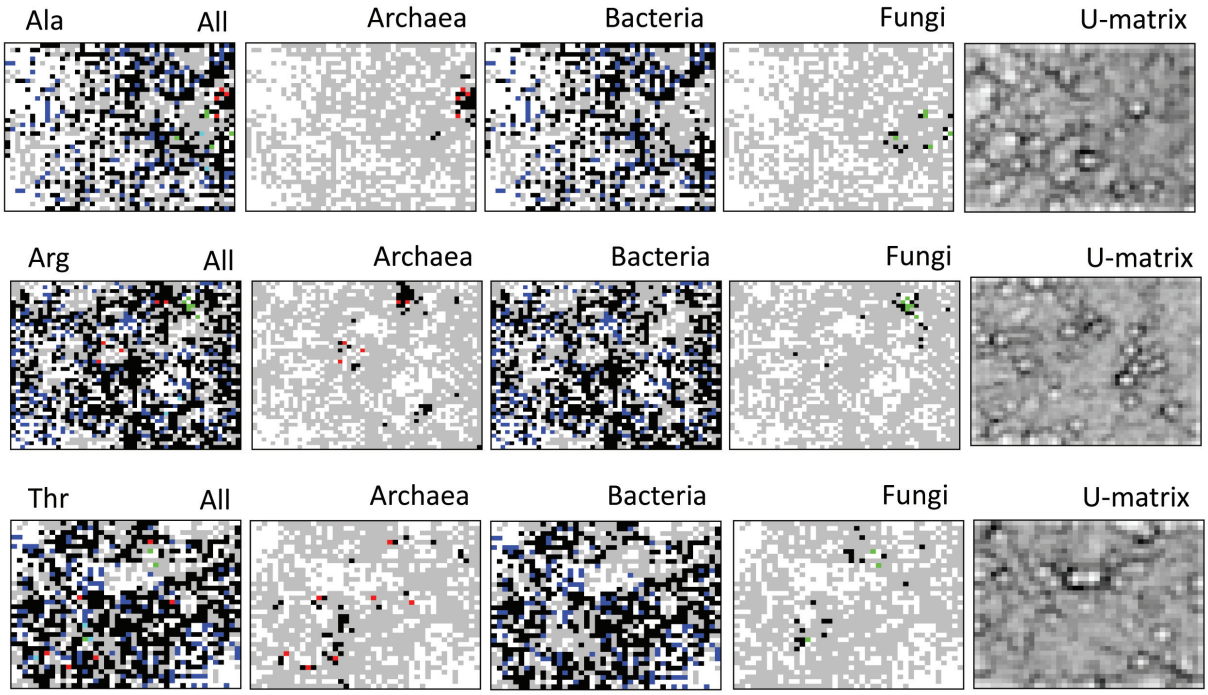

Fig. 3 BLSOM for metagenomic plus species-known microbial tDNAs of one amino acid. All; lattice points containing tDNAs derived from only bacterial, archaeal, fungal, or metagenomic sequences are coloured in blue, red, green, or gray, respectively, and those containing tDNAs from more than one category are marked in back. Bacteria, Archaea, or Fungi panel; lattice points containing metagenomic tDNAs plus bacterial, archaeal, or fungal tDNAs are separately coloured as described for the All panel.

have to be constructed only for archaeal and fungal tDNAs. The association of a large number of metagenomic tDNAs with archaeal and fungal tDNAs can be supported by the following analysis.

3.3 BLSOM for tDNAs of one amino acid

To investigate the phylotype-dependent separation of tDNAs more in detail, we have constructed a BLSOM for each amino acid for species-known plus species-unknown tDNAs, as shown by three examples of amino acids (Fig. 3). On the All panel, lattice points containing tDNAs from only bacterial, archaeal, fungal, and metagenomic sequences are coloured in blue, red, green, and gray, respectively; those containing tDNAs from more than one category are marked in black. On the Bacteria, Archaea, or Fungi panel, lattice points containing metagenomic tDNAs (gray) plus bacterial, archaeal, or fungal tDNAs are separately coloured as described for the All panel. A large portion of lattice points on the Bacteria panel are marked in black, showing that many metagenomic tDNAs are clustered (self-organized) along with known bacterial tDNAs, predicting their phylogenetic attribution. On the Bacteria panel, some clear gray contiguous areas contain metagenomic, but not bacterial, tDNAs, and a portion of the gray areas contain archaeal and fungal tDNAs (black on the Archaea or Fungi panel), providing phylogenetic attribution of these metagenomic tDNAs.

The U-matrix [15] listed in Fig. 3 visualizes the dissimilarity level of oligonucleotide composition between neighbouring lattice points as a blackness level. It should be mentioned that Dick et al. (2009) [16] has successfully applied the U-matrix method of an oligonucleotide-SOM to the phylogenetic clustering of environmental metagenomic sequences. On the U-matrix panel in Fig. 3, many white or pale black areas are surrounded by dark black circles. Because white and pale black on U-matrix means similar oligonucleotide compositions between neighbouring lattice points, and thus, between tDNAs located in neighbouring lattice points, phylogenetic predictions for metagenomic tDNAs within a white and pale black zone surrounded by a dark black circle can be obtained by referring to species-known tDNAs colocalizing in this zone, as described by Dick et al. (2009) [14].

4. Conclusion, discussion, and future prospects

Most environmental microorganisms cannot be cultured easily under laboratory conditions, and genomes of unculturable microorganisms have remained mostly uncharacterized but are thought to contain a wide range of novel genes of scientific and industrial usefulness. The most important contribution of the alignment-free clustering method BLSOM is to clarify microbial community structures in environmental ecosystems. When analysing a dataset composed mainly of metagenomic sequences shorter than 100 bp, an oligonucleotide BLSOM for tDNAs is useful. However, if the dataset is composed mainly of sequences longer than 500 bp, BLSOMs with tri- and tetranucleotide compositions in all genomic fragments should be more suitable than tRNA-BLSOM, because all genomic sequences are informative.

When searching for a certain genome of particular interest (e.g. from the view of earth sciences) by surveying a massive number of short metagenomic sequences, phylogenetic marker tDNAs should become very useful because a conventional sequence homology search can be used. Our group has started to search for tDNAs useful as phylogenetic markers, especially for rare genomes, and has planned to publish such markers in tRNADB-CE.

Unsupervised data mining (e.g. BLSOM) not requiring advanced knowledge, hypotheses, or particular models may provide the least expected knowledge and will become increasingly important in studies not only of metagenomic sequences, but also of genomic sequences from a wide variety of phylogenetic groups. Oligonucleotide BLSOM, which can analyse more than ten million sequences at once, is suitable for unveiling novel knowledge hidden within big sequence data, providing a timely tool for researches following remarkable progresses of high-throughput sequencing technology. Our recent oligonucleotide-BLSOM study of a wide range of fishes, including coelacanth, has revealed a characteristic oligonucleotide composition in the coelacanth genome evidently distinct from other fish genomes, and the characteristic composition found for coelacanth has been connected with the lowest dinucleotide CG occurrence (i.e. the highest CG suppression) among fishes, which is rather equivalent to that of tetrapods [17]. This evident CG suppression in coelacanth is thought to reflect molecular evolutionary processes of epigenetic systems including DNA methylation during vertebrate evolution. Actually, sequence of a *de novo* DNA methylase (Dntm3a) of coelacanth has been found more closely related to that of tetrapods than that of other fishes.

Oligonucleotide BLSOM, which can analyse more than ten million sequences at once by using high performance supercomputers as ES, is most suitable for unveiling novel knowledge hidden within big sequence data, providing a timely tool for researches following the remarkable progress of high-throughput sequencing technology.

Acknowledgements

This work was supported by a Grant-in-Aid for Publication of Scientific Research Results (no. 228056) for Scientific Research, from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and by a NIG Collaborative Research Program (A). The computation for constructing a large-scale BLSOM was done with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

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

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

[3] Abe, T., Sugawara, H., Kanaya, S., and Ikemura, T. (2006) 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*, 6, 17-23.

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

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

[6] Uchiyama T., Abe T., Ikemura T., and Watanabe T. (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nature Biotech*, 23, 88-93.

[7] Nakao R., Abe T., Nijhof A. M., Yamamoto S., Jongejan F., Ikemura T., and Sugimoto C. (2013) A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks. *ISME J, 7*, 1003-1015.

[8] Abe, T., Ikemura, T., Ohara, Y., Uehara, H., Kinouchi, M., Kanaya, S., Yamada, Y., Muto, A., and Inokuchi, H. (2009) tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res*., 37, D163-D168.

[9] Abe, T., Ikemura, T., Sugahara, J., Kanai, A., Ohara, Y., Uehara, H., Kinouchi, M., Kanaya, S., Yamada, Y., Muto, A., and Inokuchi, H. (2011) tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Res*., 39, D210-D213.

[10] Abe, T., Inokuchi, H., Yamada, Y., Muto, A., Iwasaki, Y., and Ikemura, T. (2014) tRNADB-CE: tRNA gene database well-timed in the era of big sequence data and its use for metagenome studies. *Front. Genet*., 01 May 2014 | doi: 10.3389/fgene.2014.00114.

[11] Lowe, T. M. and Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., 25, 955-964.

[12] Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*., 32, 11-16.

[13] Kinouchi, M. and Kurokawa. K. (2006) tRNAfinder: A software system to find all tRNA genes in the DNA sequence based on the cloverleaf secondary structure. *J. Comp. Aided Chem*., 7, 116-126.

[14] Iwasaki, Y., Wada, K., Wada, Y., Abe, T., and Ikemura, T. (2013) Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance. *Chromosome Res*., 5, 461-474.

[15] Ultsch, A. (1993) Self organized feature maps for monitoring and knowledge acquisition of a chemical process. In *Proc. ICANN'93*, Int. Conf. on Artificial Neural Networks, edited by S Gielen, B Kappen. London: Springer, pp. 864–867.

[16] Dick G. J., Andersson A. F., Baker B. J., Simmons S. L., Thomas B. C., Yelton A. P., Banfield J. F. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol*, 10, R85.

[17] Iwasaki, Y., Abe, T., Okada, N., Wada, K., Wada, Y., and Ikemura, T. (2014) Evolutionary Changes in Vertebrate Genome Signatures with Special Focus on Coelacanth. *DNA Res*., doi: 10.1093/dnares/dsu012. First published online: May 6.

地球環境変動・保全に係わる全地球レベルでの微生物群集構造把握のためのゲノム情報基盤整備

課題責任者
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，岩崎　裕貴 *2，和田健之介 *2，和田　佳子 *2，池村　淑道 *2
*1 新潟大学　工学部
*2 長浜バイオ大学　バイオサイエンス学部

地球環境の変動が生物生態系に多大な影響を与えると共に、生物生態系側も地球環境へ多大な影響を与えて来たが、そこに関与している微生物類の地球レベルでの生態については未知に残されている重要な課題が多い。地球環境への影響を知るには、個々の微生物種ではなく、微生物群集構造の包括的な把握が必須であるが、自然環境の微生物類の 99% 以上が実験室では培養が困難であり、関与する微生物類の特定を困難にして来た。最近の DNA 配列の解読技術の飛躍的な高速化は、「メタゲノム解析法」と呼ばれる革新的な手法を確立させ、全地球レベルでの大規模なメタゲノム解析が進行し、超大量なゲノム配列が集積している。我々が開発して来た、一括学習型自己組織化マップ法（BLSOM）は、オリゴヌクレオチド組成（連続塩基組成）だけで、断片ゲノム配列を生物系統ごとに高精度に分離（自己組織化）する能力を持つ。国際 DNA データバンクに収録された全ゲノム配列を対象に、ES で大規模 BLSOM を作成・更新し、メタゲノム解析で得られる大量配列をマップすることで、各環境中で生息する生物群集の全体像の把握が可能になり、加えて、環境保全・浄化に役立つ新規微生物類やそれらの保持する有用遺伝子類を発掘できる。本年度も、新たに国際 DNA 配列データベースに登録されたゲノム配列を追加して、4 連塩基組成の大規模 BLSOM を作成し、我が国の実験グループが解読をした大量メタゲノム配列の系統推定を行い、微生物集団構造を明らかにした。現時点で普及している新世代シーケンサーが解読する配列は 100bp 以下が大半であり、従来から我々が開発して来た連塩基組成の BLSOM 法に適しているとは言えない。この問題を解決する手段として、本年度は tRNA 遺伝子（tDNA）を対象にした BLSOM を開発して、新世代シーケンサーが解読するメタゲノム配列を用いて、環境中の微生物群集を推定する方法の開発も行った。tDNA の完全長が 100bp 以下であり、遺伝子配列が、特に機能モチーフ配列部位が進化的に保存性が高いことから、分子系統マーカーとしての有用性が高く、新世代シーケンサーを用いたメタゲノム研究における有用性が示された。

キーワード: 自己組織化マップ, BLSOM, 環境微生物 , メタゲノム解析 , 次世代シーケンサー , ビッグデータ