生物多様性を俯瞰するための大規模ゲノム情報基盤の整備

課題責任者
阿部　貴志　　新潟大学　工学部

著者
阿部　貴志　　新潟大学　工学部
池村　淑道　　長浜バイオ大学

次世代シークエンサーの登場以降、ゲノム解読プロジェクトに拍車がかかり、ゲノム配列データの増加は想像を遥かに超える。大量のゲノム配列データを包括的、かつ俯瞰的に把握する強力な手法が必要である。我々はこれまで、連続塩基組成のみに着目してゲノム配列断片を高精度にクラスタリングし可視化できる、一括学習型自己組織化マップ (BLSOM) を開発し、この AI 技術を様々なゲノム科学分野に応用してきた。BLSOM は並列計算に適しており、大規模解析が可能であるが、データ量の増加に伴い膨大な計算資源と計算時間が必要となっている。

我々は、ES を活用した大規模解析として、現在取得可能な全生物のゲノム配列データ 600Gbase を対象とした超大規模 BLSOM 解析を行い、更新を続けている。また、大規模 BLSOM 解析結果を用いて、水平伝播候補遺伝子探索システムの開発を行った。超大規模ゲノム配列データからの効率的な知識発見に向けた新規性の高い情報基盤を提供している。

キーワード：一括学習型自己組織化マップ（BLSOM）, 連続塩基組成, 環境メタゲノム, 水平伝播候補遺伝子, AI

1. はじめに

次世代シークエンサーに代表されるゲノム解読技術のハイスループット化に伴い、多種多様な生物種に対するゲノム解読が行われている。大量に蓄積している広範囲の生物種のゲノム塩基配列から、生物種間の類似関係に基づき、生物種ごとの特徴を明らかにすることは、遺伝学や進化生物学を含むゲノム科学の重要課題である。

ゲノム塩基配列の解読が困難であった時期には、実験で測定可能な GC% が各生物種ゲノムや、さらにはゲノムの内部構造を特徴付ける基本的な量として用いられてきた。多くのゲノムが解読されている現在では、同じ GC% を持つゲノムが多数存在し、GC% のみでは特徴を理解するのは困難である。一方、塩基配列を文章のように扱い、単語の出現頻度解析（Word Count）を行うことで、ゲノム配列に潜む多様な情報を効率的に抽出できる。ここで単語とは、2 連・3 連・4 連塩基のような連続塩基（オリゴヌクレオチド）を意味する。同一の GC% を持つ生物種でも、2 連塩基について同一な出現頻度特性を持つ生物種は少なく、連続塩基が 3 連や 4 連と長くなるにつれ、同一の頻度特性を持つ生物種の可能性は極端に小さくなる。この連続塩基頻度解析を、DDBJ/ENA/GenBank に収録されているゲノム配列の全体を対象にして、大規模な解析を行うことで、新規視点での知識発見を可能とした。

我々は、広範な生物種に由来する超大量ゲノム配列を対象に、ゲノム配列の連続塩基塩基の頻度に着目することで、生物種固有の特徴を俯瞰的に把握可能とする一括学習型自己組織化マップ（Batch-Learning Self-Organizing Map, BLSOM）を開発した [1-3]。BLSOM は生物種の情報を計算の途中で一切与えずに、連続塩基の出現頻度の類似性のみで、生物種ごとに高精度に分離（自己組織化）する強力なクラスタリング能を持ち、その結果を容易に可視化できる。さらに、並列計算に適したアルゴリズムになっており、地球シミュレータなどの高性能計算機を用いた超大規模解析をも、世界に先駆けて可能とした [4]。ゲノム配列上には、断片配列であっても生物種を特徴づけるサイン (genome signature) が内在しており、BLSOM がそのゲノム配列の個性を見分けている。これまで、比較ゲノムによるウイルス種固有な特徴の解明 [5]、連続塩基組成に基づく生物種固有なシグナル配列の探索法 [6]、環境中の微生物叢より取得されたメタゲノム配列に対する系統推定法 [7]、たんぱく質アミノ酸配列の連続アミノ酸組成に着目したたんぱく質機能推定法などへの様々なゲノム配列解析 [8] への応用を行ってきた。

環境メタゲノム配列に対する系統推定では、地球上に生息する全既知生物が持つゲノムの特徴を、網羅的かつ俯瞰的に把握する大規模 BLSOM を作成し、かつ更新していく必要があり、現時点で取得可能な全生物のゲノム配列データ 635 Gbase を対象とした超大規模 BLSOM 解析を行った。また、大規模 BLSOM 解析結果を用いて、水平伝播候補遺伝子探索システムの開発を行った。

2. 方法

2.1 一括学習型自己組織化マップ（Batch-Learning Self-Organizing Map, BLSOM）

コホネン博士が開発した自己組織化マップ（Self-Organizing Map, SOM）は大量で複雑な情報について、似た情報を自ずと集める（自己組織化する）ことを計算機上で実現している [10-12]。工学・経済学・言語学のような大量で複雑な情報を解析する分野で普及してきたが、ゲノム塩基配列の解析においては我々のグループが先導的に技術開発を進めてきた。従来型のコホネン SOM の場合、大量データの解析には長い計算時間を必要とし、出

来上がった地図がデータの入力順に依存する問題があった。我々は、従来型 SOM の長所を生かしながら、再現性のある分類結果を得る形式にアルゴリズムを変更した「一括学習型自己組織化マップ (BLSOM)」を、ES1 の時期より開発してきた [1, 2]。大量データに対する大規模な並列処理が可能となり、大量データ解析に適したアルゴリズムとなった [4-5]。

3. 結果と考察

3.1 現在入手可能な全生物ゲノム配列データを対象にした大規模 BLSOM の更新

現時点で 100 kb 以上の配列断片が INSD（国際塩基配列データベース）に収録されている原核生物の 5645 種、既知真核生物の 1793 種、既知ウイルス 622 種、ミトコンドリア 642 種、葉緑体 848 種について、塩基配列を 5000 塩基 (5kb) ごとに断片化したゲノム配列断片 127 百万件 (650 ギガ塩基) を対象に、縮退 4 連塩基頻度に基づく BLSOM 解析した結果を図 1 に示す。DNA データベースには 2 本鎖 DNA の片方の配列が登録されており、2 本鎖配列の選択に関する自由度に起因する影響を除くため、相補的な連続塩基（例えば AAAA と TTTT）を同一のものとみなすことを、「縮退」と定義している。BLSOM の計算過程では、各配列断片がどの生物種に由来するのかを、計算機には与えない（教師なし学習アルゴリズム）。BLSOM の学習後に、格子点が同一の生物系統の配列のみからなる場合にはその系統を示す色で、複数の系統の配列が混在する場合には黒色で表示した。真核と原核生物については 95% レベルの高精度で分離されており、オルガネラとウイルス相互や、これらと核ゲノムとの分離は 80% レベルであった。80% レベルである理由の一つはウイルスやオルガネラと核ゲノム間でのゲノム断片の水平伝播と考えられる。

連続塩基組成を対象にした BLSOM 解析を、比較ゲノム解析による共生微生物間の水平伝播遺伝子の予測とゲノム進化の解明、インフルエンザウイルス等のゲノム配列の宿主特異的な特徴解明や、着目生物が持つシグナルやモチーフを含む特徴配列の探索などに適用しており、全生物の特徴を俯瞰的に理解する手法を提供してきた意義は大きい。

3.2 水平伝播候補遺伝子探索への BLSOM の応用

生物種間での遺伝子の水平伝播が生物進化や環境適応に大きな役割を果たしてきたことは定説となっているが、受け手側のゲノムに埋め込まれた外来遺伝子は長い進化の過程で痕跡化しつつある。従来の相同性検索では、高精度かつ網羅的に外来遺伝子とその送り出し側の由来種を特定することが困難であり、相同性検索とは異なった原理に基づく超大規模ゲノムデータ解析法の確立が求め

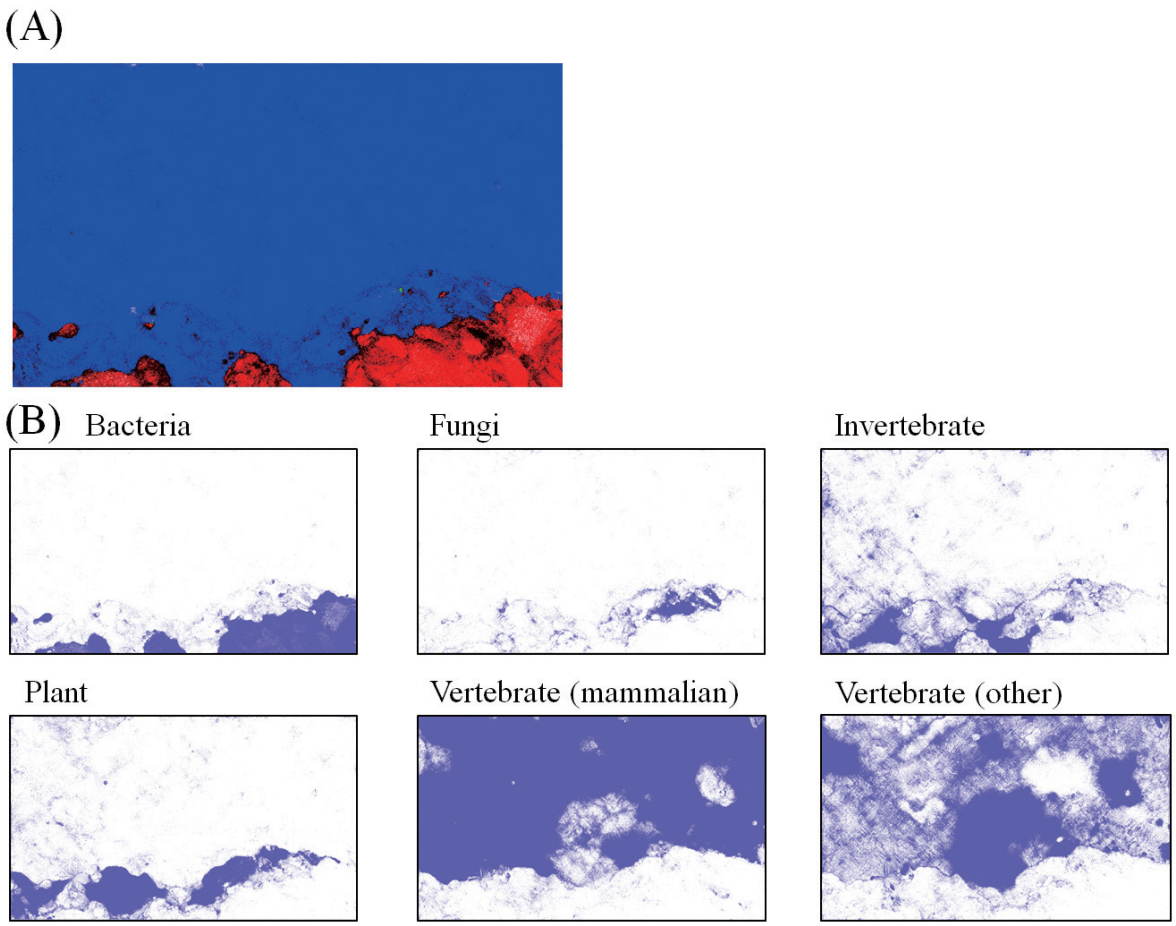

図 1　国際塩基配列データベースに収録されている原核生物 5645 種（■）、既知真核生物の 1793 種（■）、ウイルス 622 種（■）、ミトコンドリア 642 種（■）、葉緑体 848 種（■）の塩基配列情報を対象とした BLSOM マップ (A) と生物系統別の分布図 (B)。

られている。BLSOM は配列相同性に依らず、連続塩基組成のみで生物系統ごとに高精度に分離できることから、着目生物種に固有の GC%やコドン組成とは異なる水平伝播遺伝子の検出にも有用性が期待できる。

我々は、北海道大学人獣共通感染症リサーチセンターの杉本千尋先生らのグループと共同で、ツェツェバエとその共生微生物間での水平伝播候補遺伝子の検出を試みるため、連続塩基組成に基づく BLSOM を用いた水平伝播候補遺伝子の検出法の開発を行った [9]。まず、ツェツェバエゲノムと全既知微生物ゲノムとを混合させた BLSOM 解析を行った（図 2）。ツェツェバエゲノムと全既知微生物とでは明瞭な分離が見られ、ツェツェバエゲノムの 3.8% が微生物由来の領域に分離されており、それらを水平伝播候補とした。その由来をみると、既知のツェツェバエ共生微生物に加え、これまで報告されていない多様な微生物種からの水平伝播候補領域を検出することができた。宿主と共生生物間での水平伝播様式を解明する上で、相同性検索とは異なる観点での検出が可能となった。水平伝播遺伝子は着目生物種に固有の GC% やコドン組成などと異なることが多く、むしろその由来種（送り出し側）に近い特徴を示している。網羅的に、かつ高精度に水平伝播遺伝子の由来種を予測することができれば、様々な環境下での生育環境適応戦略や共生関係が及ぼしてきた共進化過程の解明に向けた基盤情報を提供できる。

4. まとめ

我々が開発してきたゲノム配列解析手法である一括学習型自己組織化マップ（BLSOM）を要素技術として、ES を中心とした HPC を用いて、様々なメタゲノム解析により取得された大量メタゲノム配列データを対象に、微生物生態系理解のための生物系統推定、ゲノム毎の再構築を行い、加えて有用遺伝子探索のためのタンパク質機能推定手法等を開発してきた。開発した情報解析システムを活用して全地球レベルでの多様な環境における微生物生態系と生物浄化能の全体像が把握することで、各々の環境に応じた環境の保全や修復を目的とした生物浄化能の最適化が可能となり、グリーン・イノベーションにおける生物浄化能の利用促進にもつながる。地球環境科学への更なる貢献が期待できる。

次世代シークエンサーの活用に伴い、ゲノムならびにメタゲノム解析は情報爆発の時代を迎えた。ゲノムビックデータに対応するためには、より高速化した解析手法の開発が求められている。BLSOM の可視化や分離能などの特長を生かして、爆発的なゲノム配列データの増加に対応できる新規解析手法を開発しており、ゲノムビックデータからの効率的なデータマイニング手法として、自然科学分野のみならず、産業界や医学分野など広い分野での更なる応用を目指した研究開発を行っていきたい。

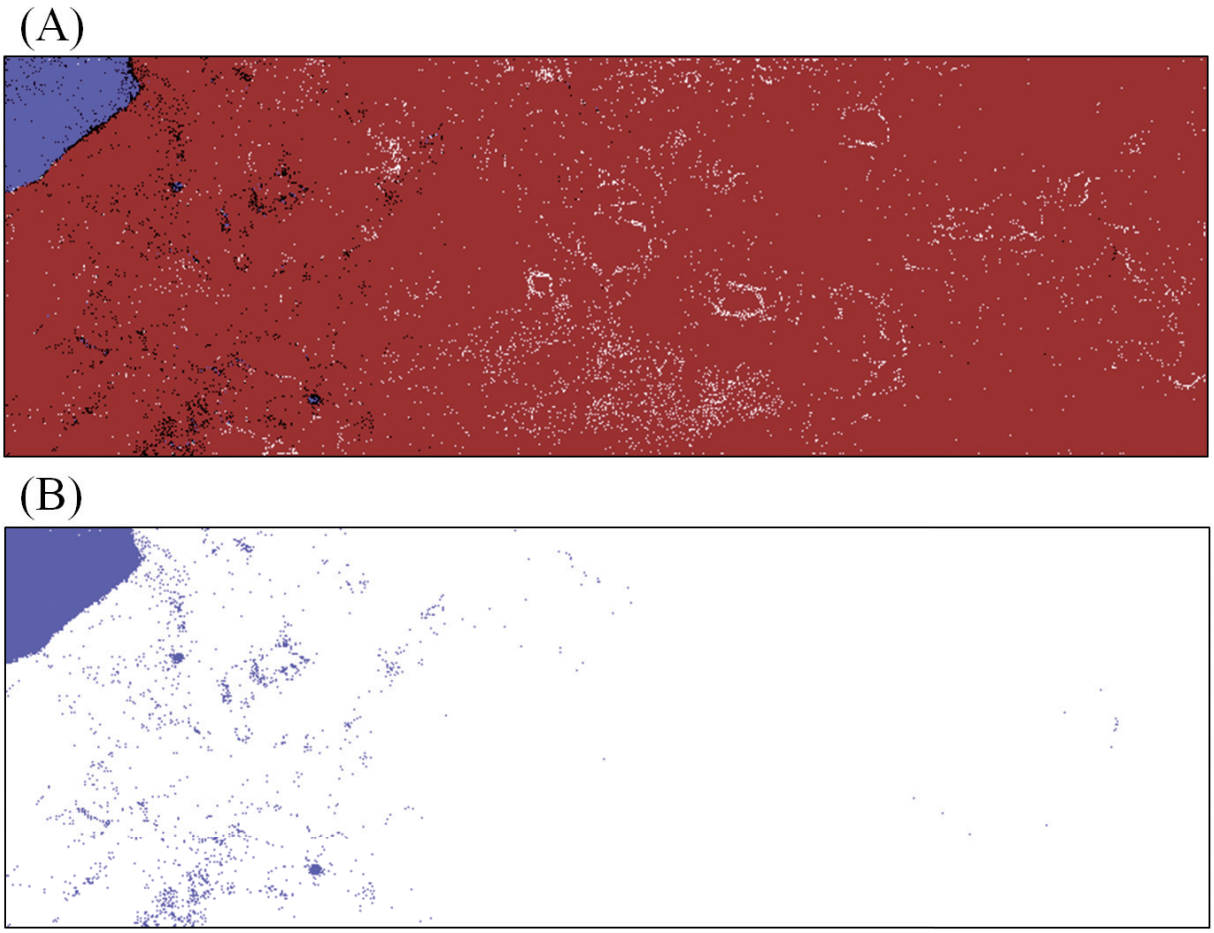

図 2　ツェツェバエ（■）と全既知微生物（■）との BLSOM マップ (A) とツェツェバエの分布図 (B)。

謝辞

本研究は、JSPS 科研費 17K00401 の助成を受けたものです。本研究成果は，地球シミュレータを主に用いて得られたものです。

文献

[1] 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.

[2] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, “Informatics for unveiling hidden genome signatures,” *Genome Research*, vol. 13, no. 4, pp. 693-702, 2003.

[3] 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 Research*, vol. 12, no. 5, pp. 281-290, 2005.

[4] 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, 1pp. 7-23, 2006.

[5] Y. Iwasaki, T. Abe, K. Wada, M. Itoh, and T. Ikemura, “Prediction of Directional Changes of Influenza A Virus Genome Sequences with Emphasis on Pandemic H1N1/09 as a Model Case,” *DNA Research*, vol. 18, no. 2, pp. 125-136, 2011.

[6] Y. Iwasaki, K. Wada, Y. Wada, T. Abe and T. Ikemura. “Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance,” *Chromosome Research*, Vol. 21, pp. 461-474, 2013.

[7] R. Nakao, T. Abe, A. M. Nijhof, S. Yamamoto, F. Jongejan, T. Ikemura, and C. Sugimoto, “A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks,” *ISME Journal*, vol. 7, no. 5, pp. 1003-1015, 2013.

[8] T. Abe, S. Kanaya, H. Uehara, and T. Ikemura, “A Novel Bioinformatics Strategy for Function Prediction of Poorly-Characterized Protein Genes Obtained from Metagenome Analyses,” *DNA Research*, vol. 16, no. 5, pp. 287-297, 2009.

[9] R. Nakao*, T. Abe*, S. Funayama, and C. Sugimoto (*equal contribution). “Horizontally Transferred Genetic Elements in the Tsetse Fly Genome: An Alignment-Free Clustering Approach Using Batch Learning Self-Organising Map (BLSOM),” *BioMed Research International*, Vol. 2016, Article ID 3164624, 2016.

A Large-Scale Batch-Learning Self-Organizing Map for Surveillance of Microbial Community Structures

Project Representative
Takashi Abe Graduate School of Science and Technology, Niigata University

Authors
Takashi Abe Graduate School of Science and Technology, Niigata University
Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

We have previously modified the conventional Self-Organizing Map (SOM), on the basis of batch-learning SOM, for genome and protein informatics, which makes the learning process and resulting map independent of the order of data input. BLSOM thus developed 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 100,000,000 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. We have constructed the large-scale BLSOM and updated it annually by analyzing all available genomic data in that time. We here apply this large-scale BLSOM to detect horizontal gene transfer candidates for unveiling symbiotic evolutionary history of the tsetse fly genome.

Keywords: Batch-Learning Self-Organizing Map, Oligonucleotide usage, Metagenome, Horizontal Gene Transfer, AI

1. Introduction

One of the most important current task of life science is to unveil unknown basic knowledge from big data of genomic sequences accumulated in the International DNA Databanks. An unsupervised neural network algorithm, self-organizing map (SOM), is an effective tool for clustering and visualizing high-dimensional complex data on a single map, and we have modified the SOM for the genome analyses by developing a Batch-Learning SOM (BLSOM) [1,2]. We have used the BLSOM to analyze short oligonucleotide frequencies (di- to pentanucleotide frequency) in a wide range of prokaryotic and eukaryotic genomes [1-4].

When only fragmental sequences (e.g., 10 kb sequences) from mixed genomes derived from multiple organisms are given, it appears impossible to identify how many and what types of genomes are present in the collected sequences. However, we found that BLSOM could classify the sequence fragments according to phylotype without any information other than oligonucleotide frequencies. BLSOM recognized, in most sequence fragments, phylotype-specific characteristics of oligonucleotide frequencies, permitting phylotype-specific clustering (self-organization) of sequences and unveiling diagnostic oligonucleotides responsible for the phylotype-specific clustering [3,4].

The present study updated the large-scale comprehensive analyses of phylotype-specific characteristics, by focusing on almost all currently available sequences from prokaryotic, eukaryotic and viral genomes. We also applied the large-scale BLSOM to detect horizontal gene transfer candidates for unveiling symbiotic evolutionary history of Tsetse (tsetse fly), which are obligate parasites that live by feeding on the blood of vertebrate animals and have been extensively studied because of their role in transmitting disease.

2. Method

BLSOM for oligonucleotide compositions and that for peptide composition were conducted as described previously (Abe et al., 2003) [2] and (Abe et al., 2009) [5].

3. Result & Discussion

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

Most environmental microorganisms cannot be cultured easily under laboratory conditions. Genomes of uncultivable microorganisms have remained largely uncharacterized and are thought to contain a wide variety of novel genes of scientific and industrial interest. Metagenomic approaches, which are analyses of mixed populations of uncultured microorganisms, have thus been developed to identify novel and industrially useful genes and to study microbial diversity in a wide range of environments. In the metagenomic approach, genome DNAs are extracted directly from an environmental sample that contains multiple organisms, and genomic fragments are sequenced. This

is a powerful strategy for comprehensive analysis of biodiversity in an ecosystem. However, with a simple collection of many genomic sequence fragments, it is difficult to predict from what phylotypes individual sequences are derived.

When we consider phylogenetic classification of species-unknown sequences obtained from environmental and clinical samples, it becomes important to construct BLSOMs in advance with all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles. This is because various eukaryotic and viral DNAs are thought to be present in environmental and clinical samples. Furthermore, when microorganisms symbiotic/parasitic with a higher eukaryote are analyzed with a metagenomic strategy, sequences from the eukaryote are included inevitably in the sequence collection. Basing on our previous results of phylogenetic classification for prokaryotic sequences, BLSOM was constructed with frequencies of degenerate sets of tetranucleotides (DegeTetra-BLSOM) in 5-kb sequence fragments [2-4].

Using the Earth Simulator, we could analyze almost all prokaryotic genomic sequences (from 5,645 species) plus genomic sequences from 1,793 eukaryotes extensively sequenced and those from 622 viruses, 642 mitochondria and 848 chloroplasts, for which at least 100 kb of sequence was available from Refseq in NCBI (Fig. 1). It should be noted that our main target of the phylogenetic classification is the sequences derived from species-unknown microorganisms present in environmental and clinical samples. To keep good resolution for microorganism sequences, it is necessary to avoid excess representation of sequences derived from higher eukaryotes with large genomes. As previously found, the power of BLSOM to separate prokaryotic, eukaryotic, viral, and organelle sequences from each other was very high; the clear separation of each taxonomic group is observed in Fig. 1.

3.2 Application to detect horizontal gene transfer candidates.

In the report of this year, we focus on the study that has been conducted under the collaboration with Prof. Sugimoto's group (Division of Collaboration and Education, Research Center for Zoonosis Control, Hokkaido University) and recently published by BioMed Research International [6].

Tsetse flies (*Glossina* spp.) are the primary vectors of trypanosomes, which can cause human and animal African trypanosomiasis in sub-Saharan African countries. The objective of this study was to explore the genome of *Glossina morsitans*

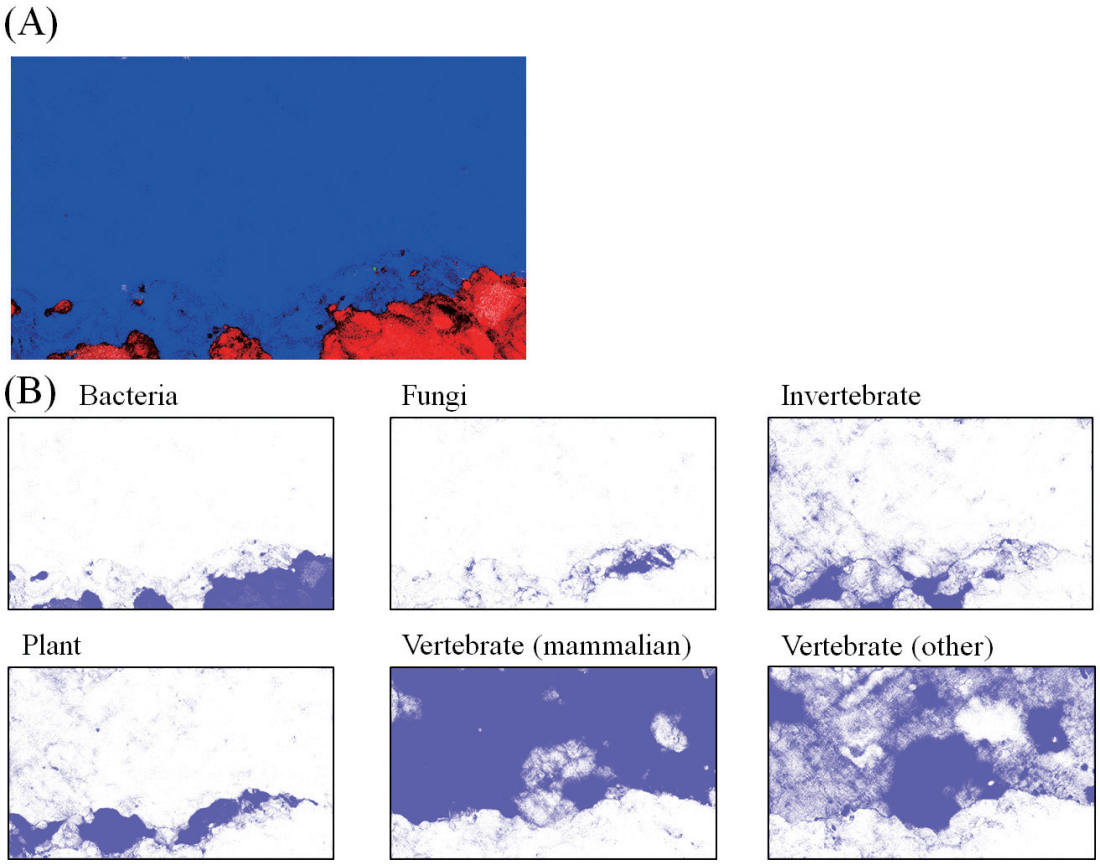

Fig. 1 Phylogenetic classification of almost all known-species. (A) DegeTetra-BLSOM of 5 kb sequences derived from almost all prokaryotic genomic sequences (from 5,645 species) plus sequences from 1,793 eukaryotes extensively sequenced and those from 622 viruses, and 642 mitochondria. Lattice points that contain sequences only prokaryotic or eukaryotic sequences are indicated in colors (■ or ■, respectively); those that contain only mitochondria, virus and chloroplasts sequences are indicated in colors (■, ■ or ■, respectively); those that include more than one category are indicated in black. (B) Distributions of each taxonomic group.

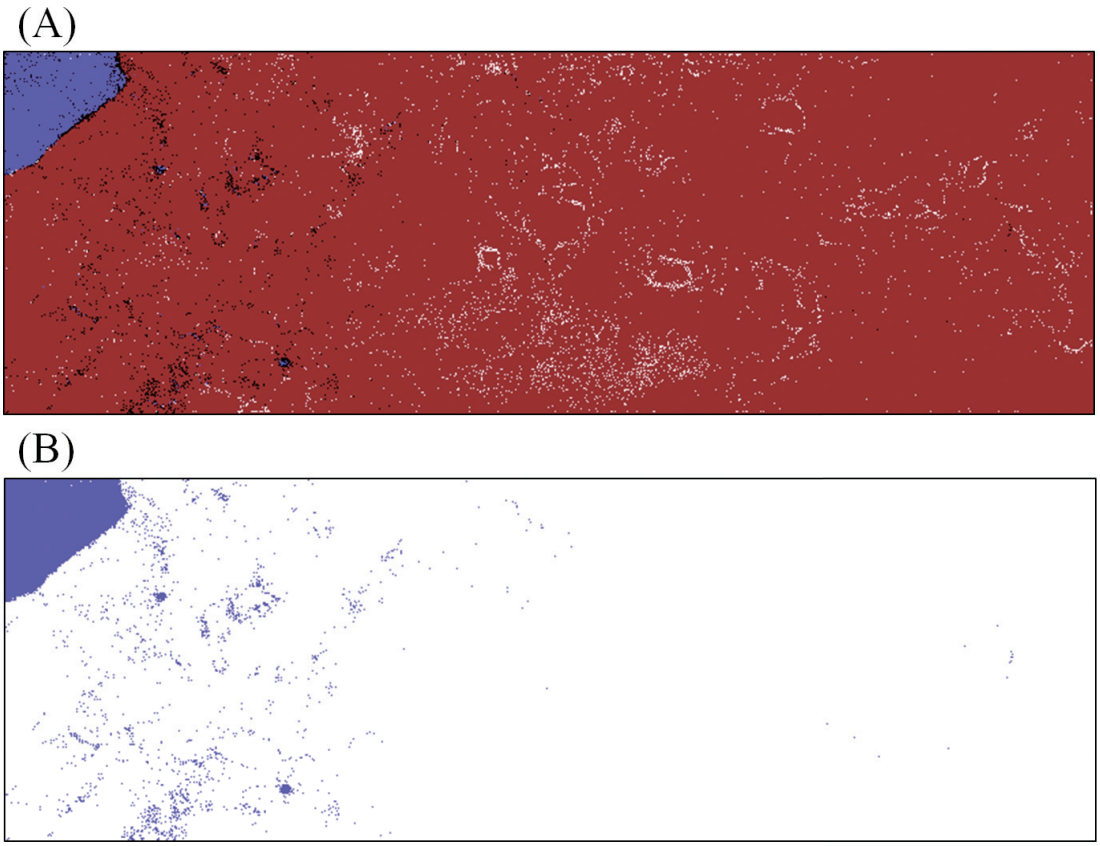

Fig. 2 Tsetse+Prokaryotes-BLSOM. (A) DegeTetra-BLSOM of the tsetse fly plus 5,600 identified prokaryotes. Lattice points that include the sequences from the tsetse fly are indicated in purple, those that contain no genomic sequences are indicated in white and those containing sequences from a prokaryote are indicated in brown. Lattice points that include both tsetse fly- and prokaryote-sequences are shown in black. (B) Distribution of tsetse fly genome sequences. Only purple lattice points are shown.

morsitans for evidencing the horizontal gene transfer (HGT) from microorganisms.

To identify HGT candidates in the tsetse fly genome derived from prokaryotes, a Tsetse+Prokaryotes-BLSOM was constructed using all genome sequences deposited in DDBJ/ENA/GenBank (Fig. 2A). In more detail, the BLSOM was constructed with a degenerate tetranucleotide composition for all 5-kb sequences derived from tsetse fly genome sequences of longer than 5 kb plus 5,600 identified prokaryote genomes, for which at least 10 kb of sequence was available from DDBJ/ENA/GenBank.

For tsetse fly contigs of longer than 5 kb (9,710 contigs), a 5-kb window with a 1-kb step was set to obtain 303,250 segments, which were mapped to the Tsetse+Prokaryotes-BLSOM by identifying the lattice point with the minimum Euclidian distances in the multidimensional space (Fig. 2B). For every lattice point, at which tsetse fly genomic segments were mapped to prokaryotic territories, the most abundant phylum was identified, and the mapped fly genomic segments were tentatively assumed to be transferred from the phylum. Finally, when the most abundant phylum in more than 40% of the segments derived from a single fly contig was the same, the contig was assigned to be derived from this phylum.

After an initial scan of HGT events using BLSOM, we identified 3.8% of the tsetse fly genome as HGT candidates. The predicted donors of these HGT candidates included known symbionts, such as *Wolbachia*, as well as bacteria that have not previously been associated with the tsetse fly. We detected HGT candidates from diverse bacteria such as *Bacillus* and *Flavobacteria*, suggesting a past association between these taxa.

4. Conclusion

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 efficient 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.

BLSOM can also be used to detect horizontal gene transfer candidates, for which the sequence similarity search could

not predict their origins. These findings provide a basis for understanding the co-evolutionary history of a host and its microbes and prove the effectiveness of BLSOM for the detection of HGT events.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C) Grant Number 17K00401. The computation was done mainly with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] 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.

[2] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, no. 4, pp. 693-702, 2003.

[3] 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 Research*, vol. 12, no. 5, pp. 281-290, 2005.

[4] 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, 1pp. 7-23, 2006.

[5] T. Abe, S. Kanaya, H. Uehara, and T. Ikemura, "A Novel Bioinformatics Strategy for Function Prediction of Poorly-Characterized Protein Genes Obtained from Metagenome Analyses," *DNA Research*, vol. 16, no. 5, pp. 287-297, 2009.

[6] R. Nakao*, T. Abe*, S. Funayama, and C. Sugimoto (*equal contribution). "Horizontally Transferred Genetic Elements in the Tsetse Fly Genome: An Alignment-Free Clustering Approach Using Batch Learning Self-Organising Map (BLSOM)," *BioMed Research International*, Vol. 2016, Article ID 3164624, 2016.