

# 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 工学部情報工学科

著者

阿部 貴志 新潟大学 工学部情報工学科

池村 淑道 長浜バイオ大学

次世代シーケンサの登場以降、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新た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。

我々は大規模なゲノム解析を高速に行うために自己圧縮 BLSOM (SC-BLSOM) を開発した。自己圧縮 BLSOM では、入力データの分類に基づいた BLSOM の階層構造を構築することで高速化を実現した。原核生物完全長ゲノムを用いて SC-BLSOM と BLSOM の比較を行ったところ、SC-BLSOM は大幅に計算時間を短縮しており、クラスタリング能も BLSOM より上昇していた。SC-BLSOM はゲノム配列データを始めとする膨大なデータから効率よく知識を得るための強力な AI 手法となりえる。

キーワード：一括学習型自己組織化マップ (BLSOM), 自己圧縮型 BLSOM (SC-BLSOM), 連続塩基組成, 環境メタゲノム, AI

## 1. はじめに

次世代シーケンサ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。ゲノム解読データは指数関数的に増加し、大量のゲノム配列データが国際塩基配列データベース (DDBJ/EMBL/GenBank) に蓄積され、今やペタスケールに届く勢いである。特に、全地球レベルでの環境メタゲノム解析が産出するゲノム配列は、その大量性だけでなく、原核・真核生物およびウイルスを含む、広範な生物種のゲノム配列が混在するのが特徴と言える。このようなゲノムピックデータからの効率的な知識発見には、公開されている全生物ゲノムの全体像を俯瞰的に把握しておくことが重要である。

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 3 連や 4 連続塩基の頻度に注目することで、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、いち早く可能とした [4]。現在公開されている全既知生物のゲノム情報を対象にした大規模 BLSOM 解析により、原核生物や真核生物ばかりではなく、ウイルス類についても、連続塩基頻度の類似度のみで、配列断片が高精度に分離することを明らかにした。この知見を基に、メタゲ

ノム配列に対する系統推定が可能なことをも、世界に先駆けて報告した [3]。国内外の実験研究者との共同研究を通じて、メタゲノム配列データを対象に BLSOM 解析を行い、約 30 報の論文発表を行ってきた [代表例は 5-8]。上述のように、環境由来のメタゲノム配列には、原核・真核生物およびウイルス類の、多種多様なゲノムに由来する配列が混在している。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 を予め作成しておく必要がある。地球シミュレータ上で高度な並列化を行い、大規模計算を実現させてきたが、全既知生物のゲノム配列を対象にした大規模 BLSOM を更新する場合、現在ではデータの増加量が想定を遥かに超えており、1 ヶ月以上もの計算時間が必要となる。ゲノムピックデータのための、超高速化した解析手法の開発が必須となってきた。

今年度の成果としては、従来型の大規模 BLSOM の更新に加えて、これまでの BLSOM の可視化や分離能などの特長は損なわず、爆発的なゲノム配列データの増加に対応する新規解析手法として、自己圧縮型 BLSOM (Self-Compressing BLSOM, SC-BLSOM) [9] を開発し、検証を行った。

## 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

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は殆ど用いられずにきた経緯がある。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、従来型のコホネン 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するために、「一括学習型自己組織化マップ (BLSOM)」を ES1 と ES2 を用いて開発してきた [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]。

## 2.2 自己圧縮 BLSOM (Self-Compressing BLSOM, SC-BLSOM)

BLSOM の計算時間増加率は、おおよそ (データの増加量  $\times 2$ )<sup>2</sup> と近似でき、データ件数が 2 倍増加すれば、計算時間は約 16 倍増加する。ゲノムビッグデータ時代において、最新データへの大規模 BLSOM の更新は容易ではない。超大量データに対応できる新規解析手法として、自己圧縮 BLSOM の開発を行った。アルゴリズムの概要を図 1 に示す。

Step 1 として、入力データをデータの分類基準に基づいて分割する。今回は、入力データとして既知原核生物のゲノム配列データを使用したため、既知原核生物の生物系統情報を分類基準とした。Step 2 として、分割した入力データごとに BLSOM 解析を行う。ここで行う BLSOM 解析のことを、1 階層目での BLSOM とする。ここで、作成する BLSOM のリファレンスベクトル数は、分割した入

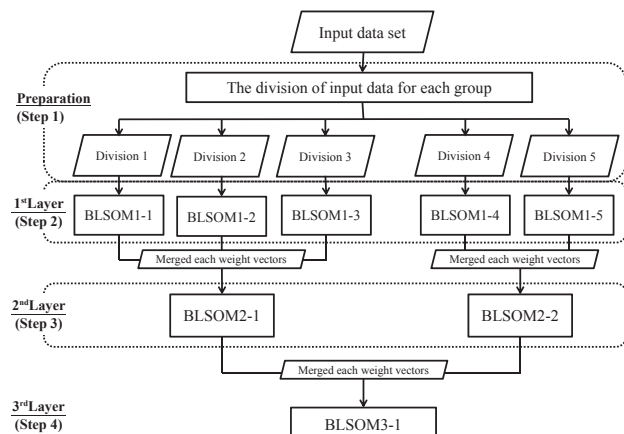


図 1 SC-BLSOM の概要

力データ数の半分とした。Step 3 として、1 階層目の分割した入力データごとの BLSOM で取得したリファレンスベクトルを次階層でのデータ分類基準に従い、集約する。Step 4 として、分割基準ごとに集約したリファレンスベクトルを用いて、BLSOM 解析を行う。ここで行う BLSOM 解析のことを、2 階層目での BLSOM とする。次の階層がある場合は、Step3 と Step4 を繰り返すことで階層を拡張していくことができる。

SC-BLSOM では BLSOM を行う回数が増えているが、1 階層目での入力データの分割により、データ数が大幅に減少するため、全入力データを対象にした BLSOM よりも計算時間が大幅に少なくなる。また、2 階層目の BLSOM で使用するリファレンスベクトルは、元データ数よりも少ないため計算時間が減少する。よって、SC-BLSOM は BLSOM より計算時間が短くなることが期待できる。

## 3. 結果と考察

### 3.1 自己圧縮 BLSOM (Self-Compressing BLSOM, SC-BLSOM) と BLSOM との比較

SC-BLSOM の基本的な性能をテストするために、計算時間とクラスタリング能を、従来の BLSOM と比較した。使用した入力データは、原核生物完全長ゲノム 817 種よりランダムに 10kb ごと抜き出し、元の配列長の 1/10 になるまで併合した塩基配列データとし、断片化サイズ 5kb、縮退 4 連続塩基頻度の条件で、従来の BLSOM と SC-BLSOM を作成した。従来の BLSOM では、リファレンスベクトル数 (ノード数) を入力データ数の 50% となるように設定した。今回の SC-BLSOM は 2 階層とし、各階層での BLSOM のリファレンスベクトル数は入力データ数の 50% となるように設定した。1 階層目での入力データの分割数は、生物系統の Phylum を利用し、対象ゲノムが属する Phylum 数である 20 とした。入力データ数は 90,998 配列断片である。従来の BLSOM と SC-BLSOM の 2 階層目の分類結果を図 2 に示す。また 1 階層目で、各 Phylum で作成した BLSOM マップを図 3 に示す。

SC-BLSOM は従来の BLSOM と比べ、計算時間が約 1/20 にまで減少させることができ、SC-BLSOM が非常に高速であることが判明した。計算時間の減少率は、各階層での入力データ数に対するリファレンスベクトル数の

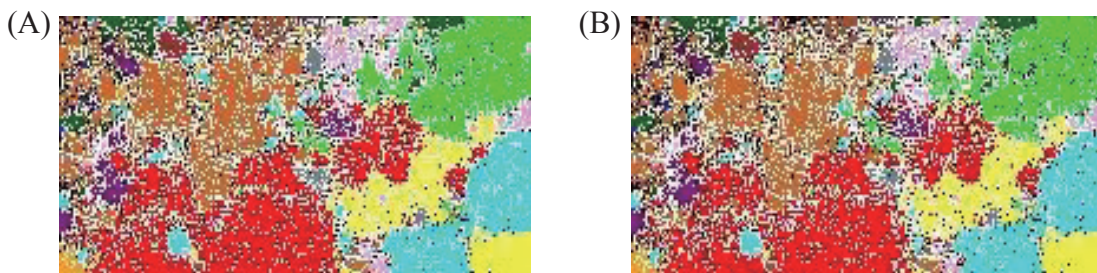


図 2 原核生物 817 種を対象に、縮退 4 連続塩基、断片化サイズ 5kb での SC-BLSOM (A) と BLSOM (B) の解析結果。ある生物系統のみが分類されている格子点をその生物系統固有の色で表示し、複数の生物種が混在している点は黒としている。生物系統と色との対応は以下のとおりである。Actinobacteria (■), Alphaproteobacteria (■), Aquificae (■), Bacteroidetes (■), Betaproteobacteria (■), Chlamydiae (■), Chlorobi (■), Chloroflexi (■), Crenarchaeota (■), Cyanobacteria (■), Deinococcus-Thermus (■), Deltaproteobacteria (■), Epsilonproteobacteria (■), Euryarchaeota (■), Firmicutes (■), Fusobacteria (■), Gammaproteobacteria (■), Spirochaetes (■), Tenericutes (■), Thermotogae (■)。



割合の2乗以下と近似できる。この場合、入力データ数の50%で、かつ、2階層であるため、 $1/16$  ( $(0.5 \times 0.5)^2$ ) 以下であり、理論値通りの結果となった。また、クラスタリング能をみると、従来のBLSOMに対し約3%向上していた。さらに、SC-BLSOMがオリジナルの入力ベクトルデータの特徴を十分に反映していたかを調べるために、オリジナルの入力データを2階層目のSC-BLSOMにプロットし、クラスタリング能を計測した。SC-BLSOMは入力データの減少に伴いマップサイズも小さくするため、元の入力データをプロットすると1リファレンスベクトルあたりの入力ベクトル数が約4件となり、従来のBLSOMの倍の密度となる。その分、クラスタリング能が悪くなると予想されたが、従来のBLSOMよりも1%以上高いという結果であった。

BLSOMは教師無しの機械学習であり、生物種の情報を実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）さ

せる能力を持つ。しかしながら、生物種が既知の全ゲノム配列で大規模BLSOMを作成しておき、その上へ生物種未知のメタゲノム配列をマップし、系統を推定する目的においては、大規模BLSOMの作成時に、生物種の知識を活用することは有用な戦略となる。SC-BLSOMでは生物系統別にマップを作成しているが、この分類は生物種についての知識を学習時に取り入れていることになる。クラスタリング能が高まる理由と考えられる。

環境メタゲノム中には新規性の高い生物種のゲノムが存在する可能性が高い。従来型の大規模BLSOMでは、既知ゲノムのなかでも、大量に存在する（言い換えれば研究が十分に進んでいる）系統に属するゲノム配列が、マップ形成において、寄与する割合が高い。新規性の高いマイナーな生物系統の特徴が、十分に反映しているとは言い難い。SC-BLSOMでは、マイナーな生物系統の寄与を高めることも可能である。

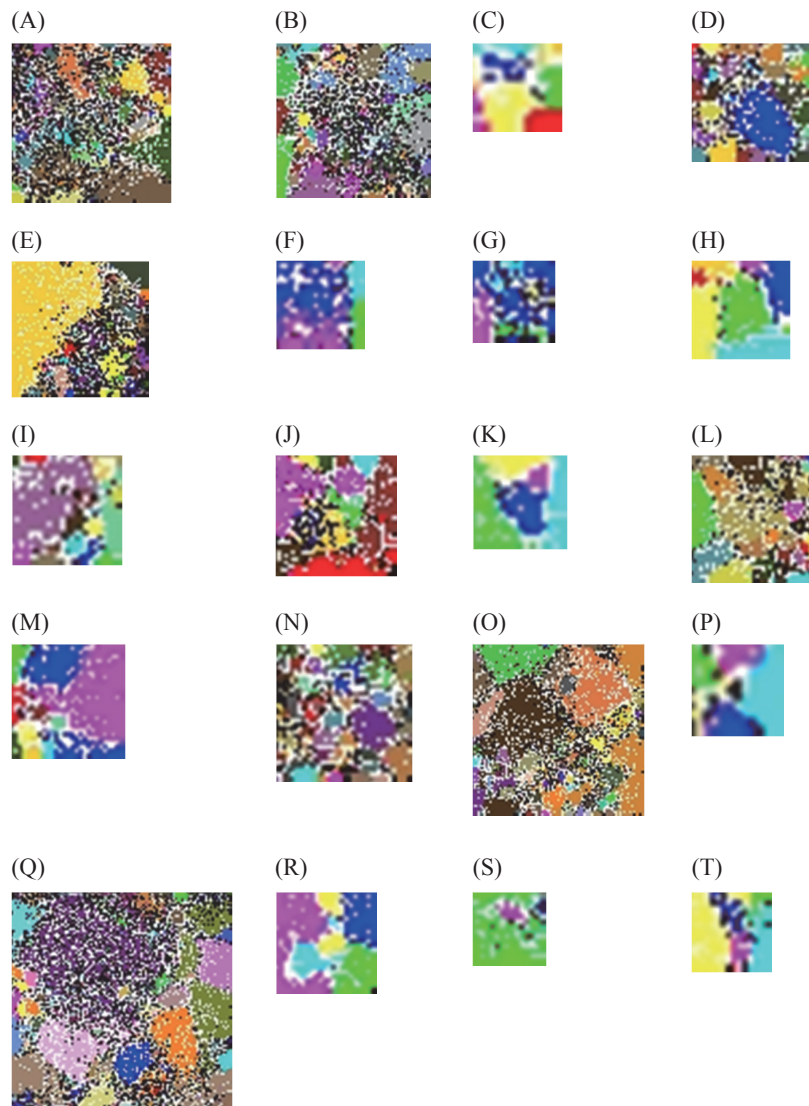


図3 SC-BLSOM解析時の1階層目でのBLSOM解析結果。生物系統別に示しており、図と生物系統との関係は以下のとおりである。(A) Actinobacteria, (B) Alphaproteobacteria, (C) Aquificae, (D) Bacteroidetes, (E) Betaproteobacteria, (F) Chlamydiae, (G) Chlorobi, (H) Chloroflexi, (I) Crenarchaeota, (J) Cyanobacteria, (K) Deinococcus-Thermus, (L) Deltaproteobacteria, (M) Epsilonproteobacteria, (N) Euryarchaeota, (O) Firmicutes, (P) Fusobacteria, (Q) Gammaproteobacteria, (R) Spirochaetes, (S) Tenerricutes, (T) Thermotogae.

#### 4. まとめ

高速かつ BLSOM と同等のクラスタリング能力のある手法として自己圧縮 BLSOM (Self Compressing BLSOM, SC-BLSOM) を開発した。SC-BLSOM は入力データを分割し BLSOM を階層的に構築することで高速化を果たし、高精度なクラスタリング能も可能であった。さらに、比較ゲノム解析への応用を行うことで手法の有用性を示した。

階層的に BLSOM を実行する SC-BLSOM は、データの分割数が多いほど高速になる。データの分割が多いほど 1 階層目の各 BLSOM に使うデータ数が少なくなり、さらにその分マップサイズも小さくなるため計算時間が大幅に短くなる。また、1 階層目の BLSOM はそれぞれ完全に独立しているため、SC-BLSOM は特別な仕組みを使わなくても簡便に並列化でき、より大量のデータを扱うことが可能である。

SC-BLSOM は、従来の BLSOM を用いたメタゲノム配列に対する生物系統推定法に加え、比較ゲノム解析 [13, 14] やタンパク質機能推定法 [15, 16] など、これまで、我々が BLSOM で発表してきた解析への応用が可能である。より超大量化してきているゲノム配列データのみならず、BLSOM 解析が可能な多次元データで、階層的に属性付けられるデータであれば解析可能であり、より高速で効果的な知識発見が可能な、汎用性が高い AI 手法となる。

#### 謝辞

本研究は、JSPS 科研費 26330327, 26330334 の助成を受けたものです。本研究成果の一部は、地球シミュレータにより得られたものです。

#### 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, pp. 7-23, 2006.
- [5] 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," *Canadian Journal of Microbiology*, vol. 51, pp. 251-259, 2005.
- [6] 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 & Genetic Systems*, vol. 86, pp. 53-66, 2011.
- [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] A. Yamamuro, A. Kouzuma, T. Abe, and K. Watanabe. "Metagenomic analyses reveal the involvement of syntrophic consortia in methanol/electricity conversion in microbial fuel cells," *PLoS ONE*, vol. 9 pp. e98425, 2014.
- [9] A. Kikuchi, T. Ikemura, and T. Abe. "Development of Self-Compressing BLSOM for Comprehensive Analysis of Big Sequence Data," *BioMed Research International*, vol. 2015, article ID 506052, 2015.
- [10] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59-69, 1982.
- [11] T. Kohonen, "The self-organizing map," *Proceeding of IEEE*, vol. 78, pp. 1464-1480, 1990.
- [12] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceeding of IEEE*, vol. 84, pp. 1358-1384, 1996.
- [13] T. Kosaka, S. Kato, T. Shimoyama, S. Ishii, T. Abe, and K. Watanabe, "The genome of *Pelotomaculum thermopropionicum* reveals niche-associated evolution in anaerobic microbiota," *Genome Res.*, vol. 18, pp. 442-448, 2008.
- [14] 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.
- [15] 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.
- [16] S. Ikeda, T. Abe, Y. Nakamura, K. Nelson, A. H. Morita, A. Nakatani, N. Ono, T. Ikemura, K. Nakamura, Md. Altaf-Ul-Amin, and S. Kanaya, "Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSACk Motorcycle database," *Plant Cell Physiol.*, vol. 54, pp. 711 - 727, 2013.

# 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 previously developed a Batch-Learning Self-Organizing Map (BLSOM), which can cluster genomic fragment sequences according to phylotype solely dependent on oligonucleotide composition. BLSOM is suitable for high-performance parallel-computing and can analyze big data simultaneously. Even though, with the remarkable current increase in genomic sequences from various organisms, novel tools are needed for comprehensive analyses of all sequence data currently available. We thus developed Self-Compressing BLSOM (SC-BLSOM) for reduction of computation time. The strategy of SC-BLSOM is to hierarchically construct BLSOMs according to data class such as phylotype. The first BLSOM was constructed with each of the divided input data that represent distribution of the data subclass, resulting in remarkable compression of the number of data for a single BLSOM. The second BLSOM was constructed with a set of the first BLSOM to summarize the class distribution.

In this year, we have developed the above SC-BLSOM for preparing the unavoidable data expansion, in addition to the annual updating of the standard large-scale BLSOM for all genomic sequences publicized by the International DNA Sequence Databank. We compared SC-BLSOM with the conventional BLSOM by focusing on bacterial genome sequences. SC-BLSOM could be constructed faster than the standard BLSOM and cluster the sequences according to phylotype with a higher accuracy, showing the method's suitability for efficient knowledge discovery from big sequence data, such as environmental metagenome sequences.

**Keywords:** Batch-Learning Self-Organizing Map (BLSOM), Self-Compressing BLSOM (SC-BLSOM), Oligonucleotide usage, Metagenome, AI

## 1. Introduction

By focusing on oligonucleotide composition in genomic sequences, we previously developed a “batch-learning self-organizing map (BLSOM)”, which allows us to panoramically grasp the characteristics of sequences unique to organism species from an ultra-large amount of genomic sequences, and have so far applied it to gene, genome, and metagenome analyses [1-3]. The method provides a strong clustering ability, with its result easily visible, under which sequences of genome fragments for each species are separated (“self-organized”) with high accuracy, based only on similarities in oligonucleotide composition with absolutely no information given on the species during learning. Furthermore, the algorithm can be optimized for parallel computations, enabling ultra-large-scale analyses performed by supercomputers, such as the “Earth Simulator” [4]. As the BLSOM takes computation time proportional to approximately the cube of the quantity of input data, a large-scale BLSOM requires huge amounts of computational time and resources. With the appearance of next-generation sequencers, which has prompted genomic sequence data to grow at an exponential rate, enhancement of computer performance alone will not suffice, and a higher-speed, larger-scale analysis strategy is now required. It should be stressed here that, for

the application to metagenome studies of environmental samples, which should contain genomes from a wide range of prokaryotes, eukaryotes and viruses, a large scale-BLSOM for all species-known genomes must be constructed in advance for mapping of metagenome sequences on the large-scale BLSOM.

We developed the “Self-Compressing BLSOM (SC-BLSOM)”, which provides higher-speed computation and clustering performance equal to or more than the conventional BLSOM [5]. The SC-BLSOM achieves higher speed by dividing input data into subclasses and structuring BLSOMs in a hierarchical manner. The present report demonstrated the effectiveness of the SC-BLSOM by means of comparative studies of its computation time and clustering performance for almost all prokaryotic genomes currently available.

## 2. Material & Method

### 2.1 Batch-Learning Self-Organizing Map (BLSOM)

We previously modified the conventional SOM [6-8] 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 analyzed as described previously [1-4]. A BLSOM program

suitable for PC cluster systems and a PC program for mapping of new sequences on a large-scale BLSOM constructed with the Earth Simulator can be obtained from our web site (<http://bioinfo.ie.niigata-u.ac.jp/?BLSOM>). We have published applications of BLSOM to a wide variety of topics [9-14].

## 2.2 Self-Compressing BLSOM (SC-BLSOM)

A conventional BLSOM performs as reflecting characteristics of input data onto the weight vectors, which are arranged on a two-dimensional map in the same format as the input data. In other words, characteristics of the input data are summarized or compressed into weight vectors. The SC-BLSOM is an analytical method that fully takes advantage of the BLSOM, and Fig. 1 shows the algorithm of the SC-BLSOM.

In Step 1, input data are divided according to data classification criteria. Here we use the genome sequence data of known prokaryotes for test input data; phylogenetic affiliation of the known prokaryote is used as the classification criterion: Division 1-5 in Fig. 1. In Step 2, BLSOM analysis is conducted on each group of the divided input data, and such BLSOM is called the first layer BLSOM. The number of BLSOM nodes (lattice points) created in this step is determined to be half the number of the divided data. In Step 3, weight vectors obtained in the first-layer BLSOMs of the divided input data are merged, according to the next layer classification criterion representing again phylogenetic affiliation: BLSOM1-1, -2, and -3 and BLSOM2-1 and -2 in Fig. 1. In Step 4, the BLSOM analyses are performed using weight vectors summarized according to the second division criterion, and the BLSOM analyses in this step are defined as the BLSOMs in the second layer; BLSOM2-1 and -2 in Fig. 1. Steps 3 and 4 can be repeated for additional layers, where applicable.

In the SC-BLSOM, more BLSOMs are constructed than in the conventional BLSOM, but they need much shorter computation time than the BLSOM for all data input at once, because the division of data in the first layer BLSOM has significantly reduced the number of input data for each divided BLSOM; as above mentioned, BLSOM takes computation

time proportional to approximately the cube of the quantity of input data. Additionally, smaller number of vectorial data in the second layer BLSOM than the original amount of data also cut computation time. Hence, shorter computation time is expected for the SC-BLSOM compared with the conventional BLSOM.

## 3. Result & Discussion

### 3.1 Performance Comparison between SC-BLSOM and BLSOM

To test the basic performance of SC-BLSOM, its computation time and clustering performance were measured and compared with those for the conventional BLSOM. This test used genomic sequence data obtained by randomly picking out genomic sequences of 10 Kb from 817 different complete genomes of prokaryotes and merging them until the sequence length was one tenth of the original data; BLSOM and SC-BLSOM were constructed with a degenerated tetranucleotide composition in a window size of 5 Kb; the frequencies of pairs of complementary tetranucleotide (e.g., AAAC and GTTT) in each fragment were summed up and abbreviated as DegeTetra [3]. The number of lattice points for the conventional BLSOM and for the SC-BLSOM in two layers were each set to be 50% of the quantity of input data. Input data for the first layer of SC-BLSOM were divided into 20 divisions using the number of phyla for the analyzed organisms as the classification criterion. The number of input data amounted to 90,998 sequence fragments.

Figure 2 shows the results for the conventional BLSOM and the second layer of the SC-BLSOM; Fig. 3 shows the BLSOM maps created for each phylum in the first layers. Here, the clustering performance is presented as the percentage of the lattice points on which only a single phylum was classified on the obtained map. The SC-BLSOM accomplished a reduced computation time approximately one sixth of the time required for the conventional BLSOM, and as for the clustering performance the SC-BLSOM showed an improvement of about 3% over the conventional BLSOM.

As the SC-BLSOM reduces the map size in proportion to the decrease of input data, plotting the original input data produces about four data per one node, resulting in a density twice that of the conventional BLSOM. It might be expected to degrade the clustering performance, but the actual performance was found to be better than that of the conventional BLSOM. We will next discuss this enhancement. BLSOM is an unsupervised machine learning, and with no phylogenetic information the genomic fragments for each species are self-organized with high accuracy according to phylotype, showing its powerful performance. However, when considering the large-scale BLSOM for all species-known genomes for mapping species-unknown metagenomic sequences, the knowledge of species can be included during learning in order to enhance the phylotype-dependent clustering. Actually, the subdivision of sequences into phylotype categories was a kind of this teaching and thus

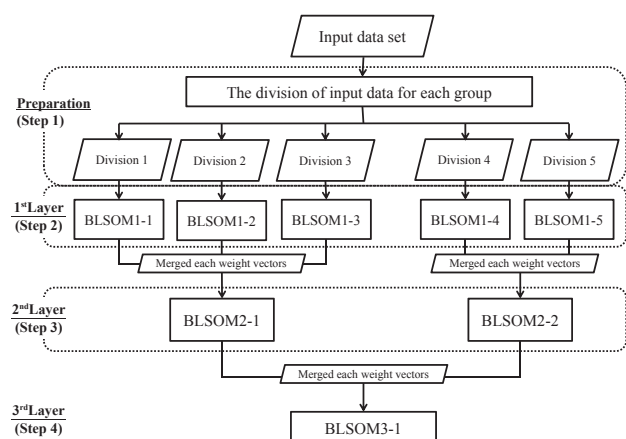


Fig. 1 Overview of SC-BLSOM algorithm



enhanced the clustering performance. Accordingly, the SC-BLSOM can be described as a method, which is fast, capable of high clustering performance and effective for large-scale genome analyses.

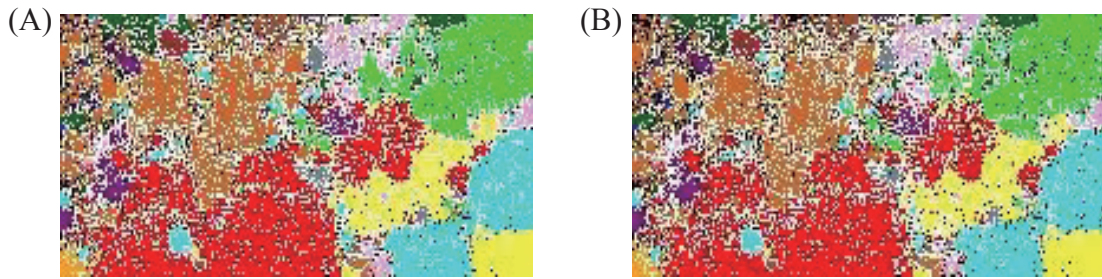


Fig. 2 SC-BLSOM and BLSOM with DegeTetra in 5-Kb sequences from 817 microbial genomes. (A) SC-BLSOM. (B) BLSOM. Lattice points that include sequences from more than one species are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single species are indicated in color as follows: Actinobacteria (■), Alphaproteobacteria (■), Aquificae (■), Bacteroidetes (■), Betaproteobacteria (■), Chlamydiae (■), Chlorobi (■), Chloroflexi (■), Crenarchaeota (■), Cyanobacteria (■), Deinococcus-Thermus (■), Deltaproteobacteria (■), Epsilonproteobacteria (■), Euryarchaeota (■), Firmicutes (■), Fusobacteria (■), Gammaproteobacteria (■), Spirochaetes (■), Tenericutes (■), Thermotogae (■).

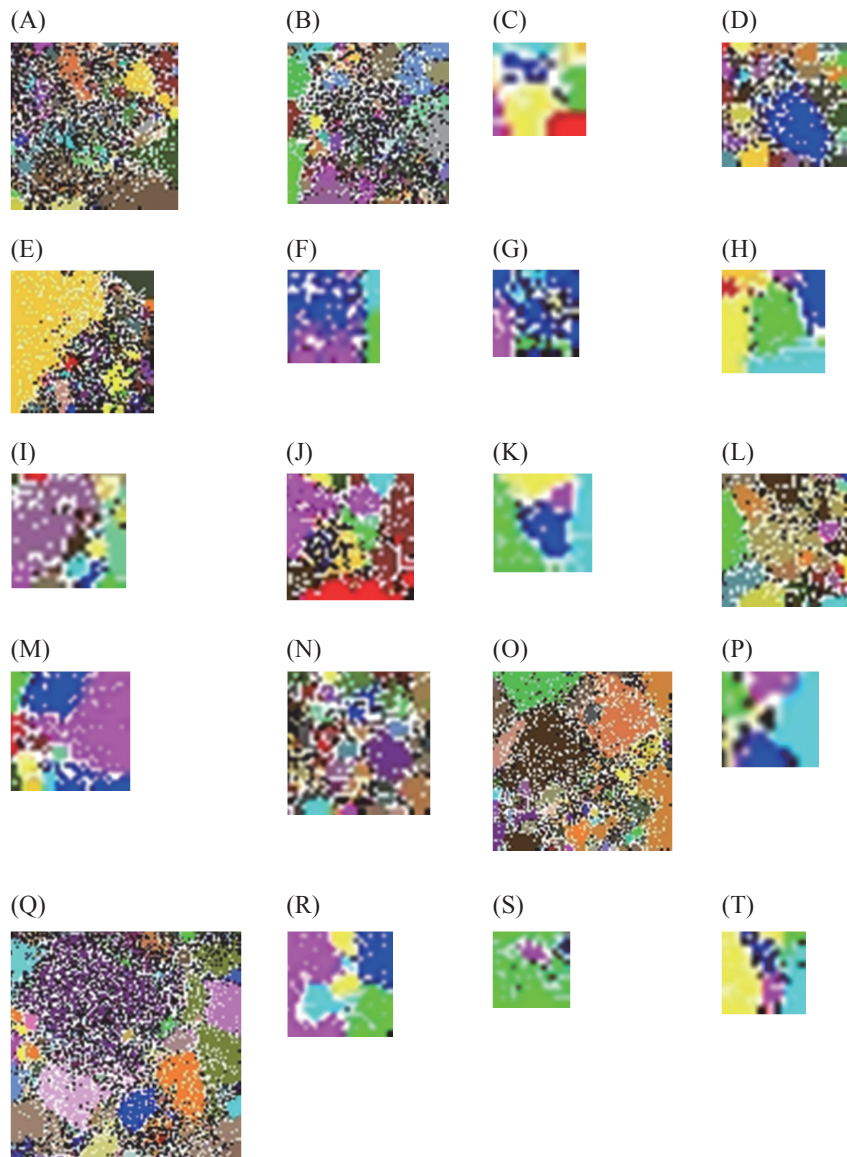


Fig. 3 BLSOM with DegeTetra in 5-Kb sequences from each phylum in the first layers. (A) Actinobacteria, (B) Alphaproteobacteria, (C) Aquificae, (D) Bacteroidetes, (E) Betaproteobacteria, (F) Chlamydiae, (G) Chlorobi, (H) Chloroflexi, (I) Crenarchaeota, (J) Cyanobacteria, (K) Deinococcus-Thermus, (L) Deltaproteobacteria, (M) Epsilonproteobacteria, (N) Euryarchaeota, (O) Firmicutes, (P) Fusobacteria, (Q) Gammaproteobacteria, (R) Spirochaetes, (S) Tenericutes, (T) Thermotogae.

#### 4. Conclusion

We have developed a Self-Compressing BLSOM (SC-BLSOM), which provides high-speed and the higher level of clustering performance than the conventional BLSOM. This high-speed is achieved by dividing input data according to phylogenetic group and structuring the layered BLSOMs. Actual application to the comparative genome analyses demonstrated the method's effectiveness [5]. As the SC-BLSOM performs the analyses in a layered manner, it does faster as the data are divided further; i.e., the fewer number of data in the first layer BLSOM reduces the map size, resulting in shorter computation time.

#### Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C) Grant Number 26330327 and 26330334. The computation was done in part 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] A. Kikuchi, T. Ikemura, and T. Abe. "Development of Self-Compressing BLSOM for Comprehensive Analysis of Big Sequence Data," *BioMed Research International*, vol. 2015, article ID 506052, 2015.
- [6] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59-69, 1982.
- [7] T. Kohonen, "The self-organizing map," *Proceeding of IEEE*, vol. 78, pp. 1464-1480, 1990.
- [8] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceeding of IEEE*, vol. 84, pp. 1358-1384, 1996.
- [9] 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.
- [10] 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.
- [11] S. Ikeda, T. Abe, Y. Nakamura, K. Nelson, A. H. Morita, A. Nakatani, N. Ono, T. Ikemura, K. Nakamura, Md. Altaf-Ul-Amin, and S. Kanaya, "Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSACk Motorcycle database" *Plant Cell Physiol.*, vol 54, pp.711 – 727, 2013.
- [12] 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.
- [13] Y. Iwasaki, T. Abe, Y. Wada, K. Wada, and T. Ikemura, "Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains" *BMC Infectious Diseases*, 201313:386, DOI: 10.1186/1471-2334-13-386.
- [14] Y. Iwasaki, T. Abe, K. Wada, Y. Wada, and T. Ikemura, "An artificial intelligence approach fit for tRNA gene studies in the era of big sequence data", *Genes & Genetic Systems*, Article ID: 16-00068, doi.org/10.1266/ggs.16-00068.