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 Nagahama Institute of Bio-Science and Technology

Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

Since the development of next-generation DNA sequencers, complete genome sequences of more than 2,000 species have been determined and metagenomic analyses covering a large number of novel species in various environments have become common (Genomes Online Database, http://www.genomesonline.org/). Microorganisms in diverse environments should contain an abundance of novel genes, and therefore, intense research activities are underway using samples obtained from a wide variety of environments, such as seawater and soil, and human intestines. The present BLSOM is an unsupervised algorithm that can separate most genomic sequence fragments based only on the similarity of oligonucleotide frequencies. It can separate sequences on a species basis with no additional information other than the oligonucleotide frequencies. Unlike the conventional phylogenetic estimation methods based on sequence homology searches, the BLSOM requires neither orthologous sequence set nor sequence alignment, and therefore, this method is most suitable for phylogenetic estimation for novel gene sequences. It can be used to visualize an environmental microbial community on a plane and to accurately compare it between different environments.

Keywords: batch learning SOM, oligopeptide frequency, protein function, bioinformatics

1. Introduction

Large-scale metagenomic analyses using recently released next-generation sequencers have actively been underway. The number of fragmental sequences obtained and registered in the International Nucleotide Sequence Databases (INSD) has soared above 17 million. For most of these genomic sequence fragments, however, it is difficult to estimate the phylogeny of organisms from which individual fragmental sequences are derived or to determine the novelty of such sequences. Most metagenomic sequences registered in the databases have limited utility because of lack of the phylogenetic information and the functional annotation; this situation has arisen because orthologous sequence sets, which cover a broad phylogenetic range and are required for the creation of reliable phylogenetic trees through sequence homology searches, are unavailable for novel gene sequences. A method for estimating the phylogeny and gene function that is based on principles totally different from sequence homology searches is urgently needed. We previously modified the SOM developed by Kohonen's group 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 [1-2]. 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 process of sequence alignment. Furthermore, the BLSOM was suitable for actualizing high-performance parallel-computing with the high-performance supercomputer "the Earth Simulator", and permitted clustering (self-organization) of almost all genomic sequences available in the International DNA Databanks on a single map [3-5]. 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 sequences from novel unknown organisms [3, 6-7] and for comparative genome analyses [8-9].

We employed BLSOM for analyses of environmental genomic fragments in joint research with experimental research groups analyzing various environmental and clinical samples [6-7]. This report introduces a strategy how to efficiently explore the genomic sequences from novel unknown microorganisms, including viral genomes, by utilizing numerous metagenomic sequences and how to determine the diversity and novelty of genomes in environmental microbial communities.

2. Methods

Nucleotide sequences were obtained from DDBJ (DNA Databank of Japan, http://www.ddbj.nig.ac.jp/anoftp-e.html). We modified the conventional SOM for genome informatics on the basis of batch-learning SOM (BLSOM) 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 on the basis of the finding that PCA can efficiently classify gene sequences into groups of known biological categories.

3. Results

3.1 Phylogenetic estimation for environmental DNA sequences and microbial community comparison

More than 17 million genomic sequence fragments obtained from various environments through metagenomic analysis have been registered in the International Nucleotide Sequence Databases. A major portion of them is novel but has a limited utility because of lack of phylogenetic and functional annotation. The phylogeny estimation of genomic sequence fragments of novel microorganisms, based on the BLSOM, requires in advance the characterization of oligonucleotide frequency of all species-known microorganisms currently available. Therefore, a large-scale BLSOM covering all known prokaryotic sequences, including those of viruses, mitochondria, chloroplasts, and plasmids was first constructed (Prokaryotes or Eukaryotes in Fig. 1). On the 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. The 210 thousand sequences with a fragment size of 1 kb or more, which were collected from the Sargasso Sea near Bermuda [10], were thus mapped. This mapping of the sequence fragments obtained from an environmental sample can estimate proportions of species present in the sample. Approximately 70% of sequences from the Sargasso Sea were mapped to the prokaryotic territories, while the rest was mapped to the eukaryotic, viral or organelle territories.

To identify the detailed phylogenies of the environmental sequences thus mapped to the prokaryotic territories, a BLSOM analyzing 5-kb genomic sequence fragments derived from 2,389 known prokaryotes, which have been compiled in the International Nucleotide Sequence Databases, was created with tetranucleotide frequencies (Phylum-BLSOM in Fig. 2). For the 5-kb genomic sequences from the 2,813 species-known prokaryotes used to create this BLSOM, their separation into 28 phylogenetic groups was examined, revealing that 85% of the sequences separated according to their phylotypes. The reason why 100% separation was not achieved is thought to be mainly because of horizontal gene transfer between the genomes of different microbial species [2-3].

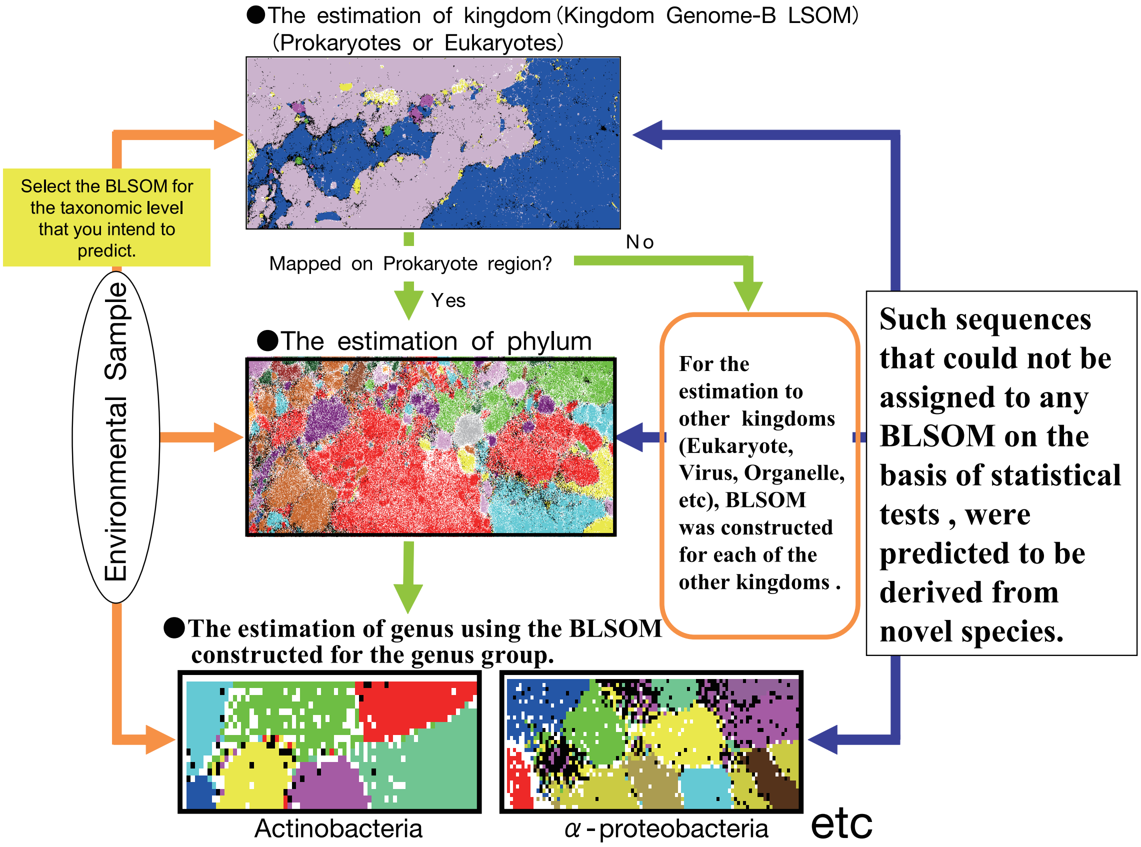
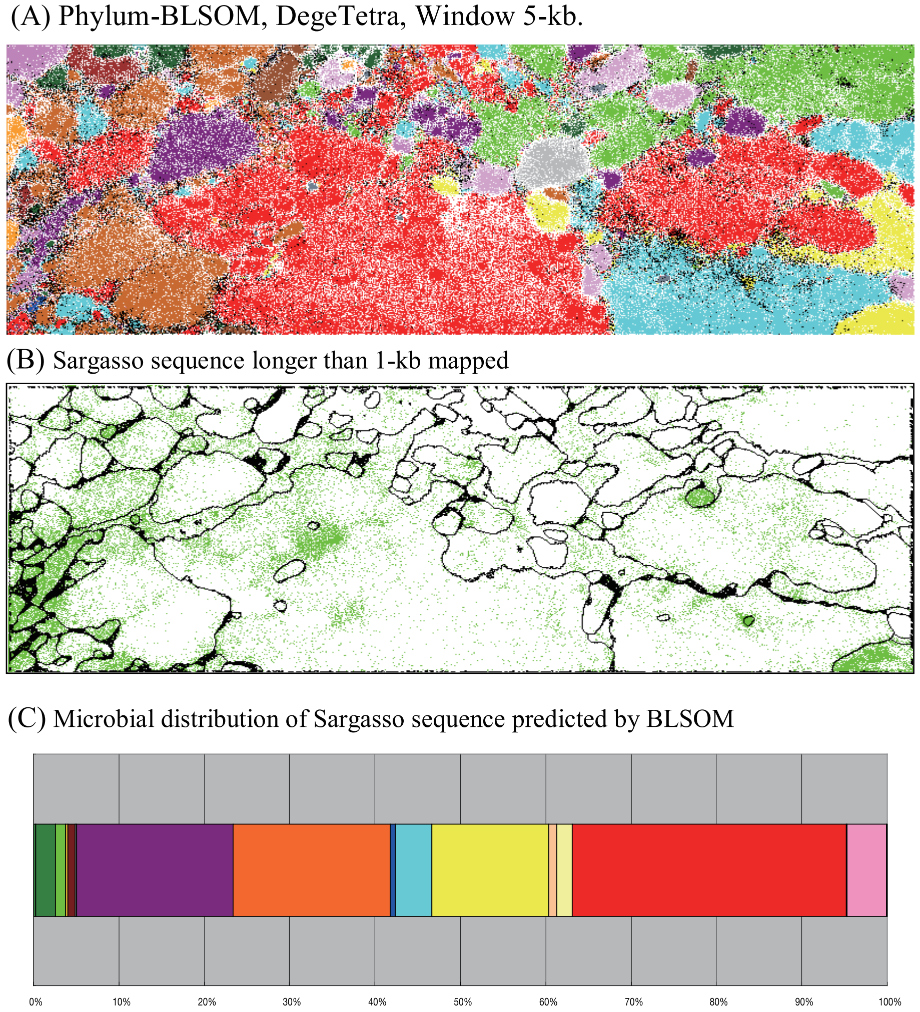

●The estimation of kingdom (Kingdom Genome-B LSOM)
(Prokaryotes or Eukaryotes)
Select the BLSOM for the taxonomic level that you intend to predict.
Environmental Sample
Mapped on Prokaryote region?
No
Yes
●The estimation of phylum
For the estimation to other kingdoms (Eukaryote, Virus, Organelle, etc), BLSOM was constructed for each of the other kingdoms .
Such sequences that could not be assigned to any BLSOM on the basis of statistical tests , were predicted to be derived from novel species.
●The estimation of genus using the BLSOM constructed for the genus group.
Actinobacteria
α-proteobacteria
etc

Fig. 1 The workflow of phylogenetic estimation using Genome-BLSOMs.

The 140 thousand metagenomic sequences from the Sargasso Sea that were mapped in advance to the prokaryotic territories (Fig. 1) were remapped on the BLSOM for the detailed prokaryotic phylotype assignment. They broadly spread across the BLSOM, demonstrating that the sequences belonged to a wide range of phylogenies (Fig. 2B). Interestingly, there were areas on the map where metagenomic sequences were densely mapped, which should 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 (Fig. 2C). The procedure above can be used to establish the phylogenetic distribution of microbial communities living in any subject environments, e.g., floras.

Further detailed phylogenetic estimation at the genus or species level becomes possible, through successive remapping the subject sequences on a BLSOM created with the sequences from known genomes of each phylogenetic group such as one family. Such systematic and detailed phylogenetic estimation in the stepwise manner from the domains of organisms (e.g., eukaryotes and prokaryotes), through the phylogenetic groups to the genus or species level, was explained in Fig. 1. This procedure can also determine the novelty of environmental sequences at various phylogenetic levels, allowing the efficient detection of sequences with high novelty.



(A) Phylum-BLSOM, DegeTetra, Window 5-kb.

(B) Sargasso sequence longer than 1-kb mapped

(C) Microbial distribution of Sargasso sequence predicted by BLSOM

Nodes that include sequences from plural 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: Acidobacteria (🟩), Actinobacteria (🟩), Alphaproteobacteria (🟦), Aquificae (🟦), Bacteroidetes (🟫), Betaproteobacteria (🟨), Chlamydiae (🟥), Chlorobi (⬜), Cenibacterium (🟥), Chloroflexi (⬜), Crenarchaeota (🟩), Cyanobacteria (🟪), Deinococcus-Thermus (🟩), Deltaproteobacteria (🟧), Dictyoglomi (🟩), Epsilonproteobacteria (🟧), Euryarchaeota (🟩), Fibrobacteres (🟫), Firmicutes (🟧), Fusobacteria (🟦), Gammaproteobacteria (🟥), Nanoarchaeota (🟩), Nitrospirae (🟫), Planctomycetes (⬜), Spirochaetales (🟪), Thermodesulfobacteriales (🟫), Thermotogales (🟪), Verrucomicrobiae (🟥)

Fig. 2 Phylogenetic classification of sequences from an environmental sample. (A) DegeTetra-BLSOM of 5-kb sequences derived from species-known 2,813 prokaryotes. (B) Sargasso sequences that were classified into prokaryotic territories in Fig. 1 were mapped on the 5-kb DegeTetra-BLSOM constructed with the sequences only from the species-known 2,813 prokaryotes. (C) Microbial distribution of Sargasso sequences predicted by BLSOM.

3.2 Visualization of all virus genome sequences on one plane

Currently, metagenomic analyses focusing on an abundance of viruses in seawater have been reported [11]. Since virus genomes contain no rDNA, conventional methods of phylogenetic estimation based on rRNA sequence cannot be used, and therefore, a new method is urgently required. To test the clustering power of BLSOM for wide varieties of virus sequences, we first analyzed tri- and tetranucleotide frequencies in sequences of 43,828 virus genomes, which have been compiled by GIB-V (Genome Information Broker for Viruses, http://gib-v.genes.nig.ac.jp/) in DDBJ (DNA Databank of Japan). BLSOM was constructed with tri- and tetranucleotide frequencies (Tri- and Tetra-BLSOM) in all 0.5- and 1-kb fragment sequences derived from virus genomes (Fig. 3A and 3B, respectively). Then, lattice points that contained sequences from a single phylogenetic family are indicated in color, and those that included sequences from more than one family are indicated in black. A major portion of the BLSOMs was colored, showing a major portion of the fragmented sequences to be separated (self-organized) according to phylotype. The level of the phylotype-specific clustering was slightly higher for Tetra-BLSOM than for Tri-BLSOM, and the level for the 1-kb sequences was higher than that of the 0.5-kb fragment sequences. It should be pointed out that no information of virus phylotype was given during the BLSOM calculation; i.e., unsupervised self-organization.

3.3 BLSOM for prediction of protein function

For almost half of protein-gene candidates predicted from novel genomes newly sequenced, protein functions cannot be estimated through sequence homology searches. To complement the homology searches, the establishment of a protein function estimation method based on totally different principles is important. We have applied BLSOM to protein studies by analyzing oligopeptide frequencies and found the separation (self-organization) of proteins according to their functions [12]. This shows that the BLSOM can be used for a protein function estimation that does not rely on sequence homology searches, providing a novel, valuable method to find scientifically or industrially important protein genes that have not been found by sequence homology searches. Large-scale BLSOMs, which analyse vast quantities of genomic sequences and protein sequences, can facilitate the efficient extraction of useful information that supports development in a broad range of life sciences and industrial fields.

4. Conclusion and Perspective

We established a method of phylogenetic prediction for individual genomic fragments obtained by metagenomic analysis, by using BLSOM of oligonucleotide frequencies. The publication of large-scale BLSOM constructed with ES, which can separate all genomic sequences currently available on a plane, will provide a foundation of novel and large-scale genomic information useful for a broad range of life sciences, such as medical and pharmaceutical sciences, and related industrial fields. The mapping of newly obtained sequences

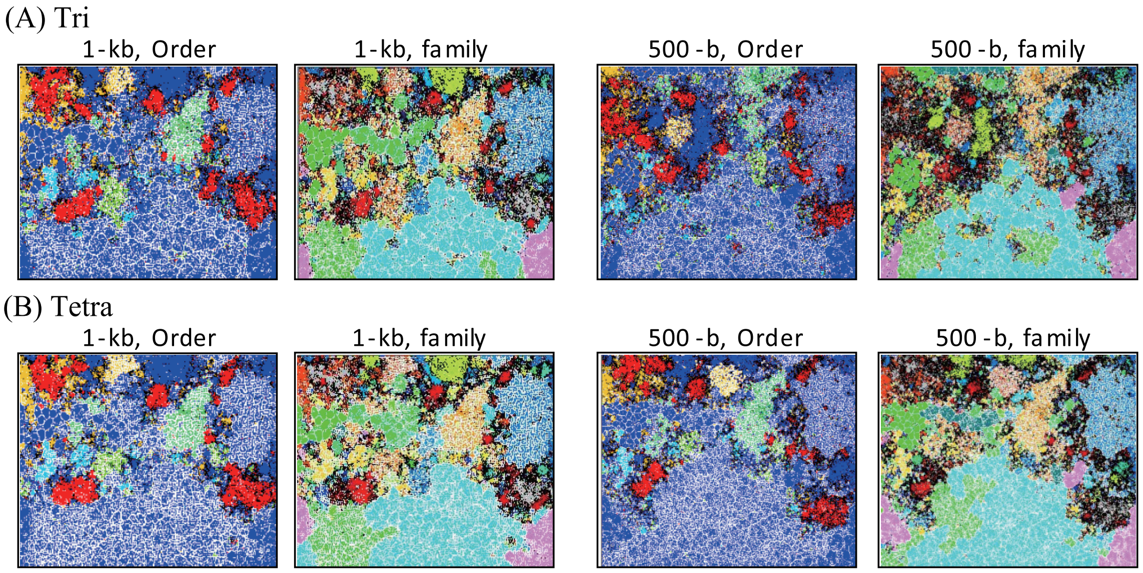


Fig. 3 BLSOMs for non-overlapping 1-kb and 500 bp sequences of all virus genomes. (A) Tri-BLSOMs. (B) Tetra-BLSOMs. Lattice points that include sequences from plural species are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single species are indicated with difference in levels of blackness and with letters as follows: Order: Caudovirales (🟥), Herpesvirales (🟨), Mononegavirales (🟩), Nidovirales (🟩), Nidovirales (🟦), Orderunclassified (🟦). Family: Coronaviridae (🟧), Siphoviridae (⬜), Hepadnaviridae (🟩), Flaviviridae (🟨), Poxviridae (🟦), Retroviridae (🟪), Orthomyxoviridae (🟦).

on the large-scale BLSOM can be performed using a PC-level computer; our group has created a PC software program for the BLSOM mapping.

We introduced also the BLSOM method for predicting functions of proteins obtained by genome analyses. For function-unknown proteins for which the consistency of the predicted function is observed by BLSOMs with the frequencies of di-, tri-, and tetrapeptides, their predicted functions are thought to be reliable. Use of the high-performance supercomputer ES is essential for these large-scale BLSOM analyses. The data obtained by ES are unique datasets in genomics and proteomics fields and provide a valuable guideline for research groups including those in industry to study functions of novel genes with scientific and industrial usefulness.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C, 20510194) and for Young Scientists (B, 20700273) 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] 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 Res*., vol. 13, 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 Res*., vol. 12, 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.17-23, 2006.

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

[6] 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.

[7] 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.

[8] 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*., 18, pp. 442-448, 2008.

[9] K. Yasui, M. Tabata, S. Yamada, T. Abe, T. Ikemura, R. Osawa, and T. Suzuki, "Intra-Species Diversity between Seven Bifidobacterium adolescentis Strains Identified by Genome-Wide Tiling Array Analysis", *Biosci. Biotechnol. Biochem*., vol 73, pp. 1422-1424, 2009.

[10] J. C. Venter, K. Remington, J. F. Heidelberg et al., "Environmental genome shotgun sequencing of the Sargasso Sea", *Science*, vol. 304, pp. 66-74, 2004.

[11] R. A. Edwards, and F. Rohwer, "Viral metagenomics", *Nature Rev. Microbiol*., vol 3, pp. 504-510, 2005.

[12] T. Abe, H. Uehara, S. Kanaya, and T. Ikemura, "A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses", *DNA Res*., vol. 16, pp. 469-477, 2009.

地球シミュレータで可能になる大規模なゲノミクスとプロテオミクス研究

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　　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

生命の設計図・シナリオであるゲノムはATGCの4種類の塩基からなり、特徴はその配列長が極めて長いことにある。最近では「次世代シーケンサー」と呼ばれるDNA配列解読のための革命的とも言われる新実験技術・装置類が登場し、既に2000を超える生物種のゲノム配列が解読されており、配列情報のデータベースへの蓄積は爆発的な増加を見せている。ゲノム解読技術の発展は、「メタゲノム解析」と呼ばれる新実験技術を生み、全地球レベルでの生物生態系の把握を目標にした大規模解析も可能になってきた。「メタゲノム解析」とは、環境中に生息する生物群集に由来する、多種類のゲノムの混合物を対象にしたゲノム配列解読である。自然環境で生息する99%以上の微生物類は実験室での培養が困難であり、通常の実験的な研究がなされておらず、膨大なゲノム資源が未開拓・未利用に残されてきた。この難培養性微生物類のゲノムは新規な遺伝子を豊富に保有すると考えられ、産業的・医学的にも注目を集めている。環境問題における重要性も明らかとなり、多様な環境由来の混合ゲノム試料を対象にした「大規模メタゲノム解析」が普及しつつある。既に、大量なゲノム断片配列が公的データベースに収録されているが、新規性の高い配列類であることから参照配列がなく「どの系統の生物種に由来するのか」や「どのような機能を有するのか」を、配列相同性検索のような従来法で適確に推定することは不可能である。生物系統やタンパク質機能についての情報なしに、利用価値が低いままに、配列だけがデータベースに収録されている。

我々のグループの場合は、超大量のゲノム配列が解読される時代の到来を予測し、その状況に適した情報解析法を開発してきた。既に特許化したBLSOM（一括学習型自己組織化マップ）は、大量ゲノム情報からの知識発見において、当初予想を遥かに超える優れた能力を持つことが判明した。高度な並列化にも適しており、地球シミュレータを使用して、大規模ゲノム解析を行っている。オリゴヌクレオチド頻度に着目したBLSOMは、メタゲノム解析で得られる大量な断片ゲノム配列について、その由来する生物系統の推定を可能にし、オリゴペプチド頻度に着目したBLSOMは各断片配列にコードされるタンパク質遺伝子の機能推定を可能にした。

キーワード: 自己組織化マップ, 環境微生物, オリゴヌクレオチド頻度, オリゴペプチド頻度, 生物系統推定,
タンパク質機能推定