Software Development Based on BLSOM for Unveiling Microbial Diversities Hidden in a Massive Number of Metagenomic Sequences

Project Representative

Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

Authors

Takashi Abe Graduate school of Science and Technology, Niigata University

Kennosuke Wada Nagahama Institute of Bio-Science and Technology

Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

We have previously modified the conventional Self-Organizing Map (SOM) for genome and protein informatics, on the basis of batch-learning SOM, which makes the learning process and resulting map independent of the order of data input. BLSOM thus developed has become suitable for actualizing high-performance parallel-computing and revealed species-specific characteristics of oligonucleotide (e.g., tetranucleotide) composition in individual genomes. This permits clustering (self-organization) of genomic fragments (e.g., 1 kb or less) according to species and phylotype without phylogenetic information during the calculation. When using ES, this alignment-free clustering method BLSOM could analyze far more than 10,000,000 sequences simultaneously. Therefore, sequence fragments from almost all prokaryotic, eukaryotic and viral genomes currently available can be classified (self-organized) according to phylotype on a single two-dimensional map. We have annually updated this large-scale BLSOM by analyzing all sequence data available at the concerned year. By mapping metagenomic sequences obtained from environmental or clinical samples on this large-scale BLSOM, we can predict phylotype compositions of the samples.

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

1. Introduction

One of the most important tasks in the environmental life science is to unveil unknown basic knowledge from big data of genomic sequences currently accumulated at an accelerating pace in the International DNA Databanks. We have developed a novel bioinformatics tool for large-scale comprehensive, phylogenetic studies on big sequence data, which is applicable to any environmental sample: a tool that can overview all available sequences from prokaryotic, eukaryotic, organelle and viral genomes at once. 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 [1-3]. We have modified the SOM for the genome analysis by developing a Batch-Learning SOM (BLSOM) [4], and applied the BLSOM to the analysis of short oligonucleotide composition (di- to pentanucleotide composition) in a wide range of prokaryotic, eukaryotic, organelle and viral genomes [5-7].

Suppose only a massive amount of fragmental sequences (e.g., 1 kb sequences) derived from mixed genomes of multiple organisms in an environmental sample are available, it appears impossible to identify how many and what types of genomes are present in the sample. However, we found that BLSOM can classify the genomic fragments according to phylotype without any information other than oligonucleotide composition; BLSOM can properly recognize species-specific characteristics of oligonucleotide composition in most genomic fragments, permitting phylotype-specific clustering (self-organization) of sequences and unveiling diagnostic oligonucleotides responsible for the phylotype-specific clustering [5-9].

Metagenomics studies of uncultivable microorganisms in environmental and clinical samples should allow extensive surveys of genes useful in medical and industrial applications, and furthermore, have become increasingly important in the environmental geoscience, as an indispensable tool for revealing microbial diversities in ecosystems. Traditional methods of phylogenetic assignment have been based on sequence homology searches and, therefore, inevitably focused on well-characterized genes, for which orthologous sequences required for constructing a reliable phylogenetic tree are available. However, most of the well-characterized genes are not industrially attractive. The present alignment-free clustering method, BLSOM, is a suitable method for characterizing novel genes, for which orthologous sequences covering a wide range of phylogenetic groups are not available.

When we consider phylogenetic classification of species-unknown sequences obtained from a poorly-studied environmental sample, BLSOMs have to be constructed in advance with all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles, because novel genomes may occur in the ecosystem. When ES is used, even ten millions of sequences can be clustered (self-organized) on BLSOM according to phylotypes with high accuracy [5-7]. By mapping a large number of environmental genomic sequences on this large-scale BLSOM, we can predict phylotype of each of these environmental sequences, and therefore, reveal an ecological structure in each ecosystem.

2. Methods

BLSOM for oligonucleotide composition was conducted as described previously [5, 10].

3. Results

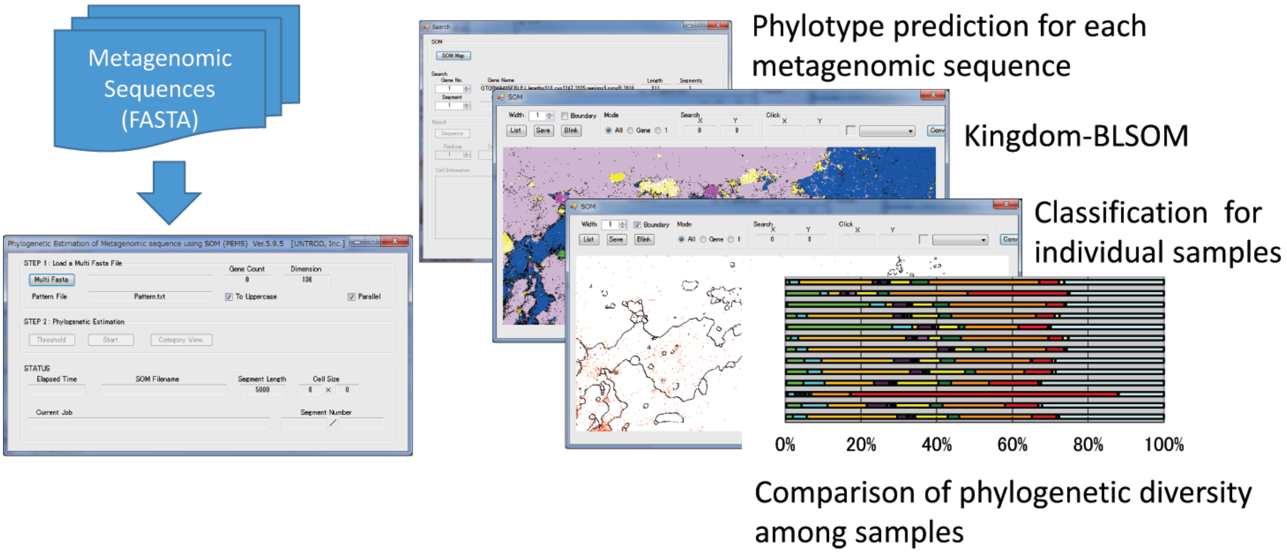
3.1 Software development for unveiling of microbial diversities present in metagenomic sequences

Because BLSOM does not require orthologous sequence sets for phylogenetic assignment, this alignment-free method can provide a systematic strategy for revealing both microbial diversity and relative abundance of different phylotype members of uncultured microorganisms, including viruses, occurring even in a novel environment. Here, we introduce a software, with which researchers can easily predict the phylotype for each of a large number of metagenomics sequences, as described in Fig. 1. The software, which is named PEMS (Phylogenetic Estimation of Metagenomic sequences on the basis of batch-learning Self-organizing map), can be downloaded freely at http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft_e.

To estimate phylotypes of the metagenomic sequences, we have equipped three types of large-scale BLSOMs, namely Kingdom-, Prokaryote- and Genus group-BLSOM, in advance, using all genome sequences deposited in DDBJ/ENA/GenBank. Kingdom-BLSOM have been constructed with tetranucleotide composition in all 5-kb sequences derived from the whole-genome sequences of 111 eukaryotes, 2,813 prokaryotes, 1,728 mitochondria, 110 chloroplasts and 31,486 viruses (Fig. 2A). To obtain more detailed phylotype information for prokaryotic sequences, Prokaryote- and Genus group-BLSOM have been constructed with a total of 3,500,000 5-kb sequences from 3,157 species, for which at least 10 kb of sequence was available from DDBJ/ENA/GenBank (Fig. 2B).

Mapping of metagenomic sequences longer than 300 bp on Kingdom-BLSOMs, after normalization of the sequence length, can be conducted by finding the lattice point with the minimum Euclidean distance in the multidimensional space. Importantly, oligonucleotide occurrences in the both terminal sequences derived from one metagenomic fragment can be added, even the two terminal sequences are segmentalized by an undermined sequence. To identify further detailed phylogenies of the metagenomic sequences that have been mapped to the prokaryotic territories on Kingdom-BLSOM, these are successively mapped on Prokaryote-BLSOM. Similar stepwise mappings of metagenomic sequences on BLSOMs constructed with sequences from more detailed phylogenetic categories (e.g., phylum and genus) are then conducted, to obtain further detailed phylogenetic information.

The PEMS has successfully unveiled microbial diversities in metagenomics sequences [11-15]. We previously developed also a BLSOM that can predict protein functions on the basis of similarity in short oligopeptide (e.g., di- and tripeptide) composition [10].


Metagenomic Sequences (FASTA)
Phylotype prediction for each metagenomic sequence
Kingdom-BLSOM
Classification for individual samples
0%
20%
40%
60%
80%
100%
Comparison of phylogenetic diversity among samples
Search
SOM
SOM Map
Width
Boundary
Mode
List
Save
Blink
Multi Fasta
Threshold
Start
Category View

Phylogenetic Estimation of Metagenomic sequence using SOM (PEMS) Ver.5.9.5 [UNTROD, Inc.]

STEP 1 : Load a Multi Fasta File

Gene Count: 0
Dimension: 136
Pattern File: Pattern.txt
☑ To Uppercase
☑ Parallel

STEP 2 : Phylogenetic Estimation

STATUS

Elapsed Time
SOM Filename
Segment Length: 5000
Cell Size: 0 × 0
Current Job
Segment Number: /

Fig. 1 Overview of PEMS. The workflow for phylogenetic prediction of metagenomic sequences. For details, see our URL (http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft_e).

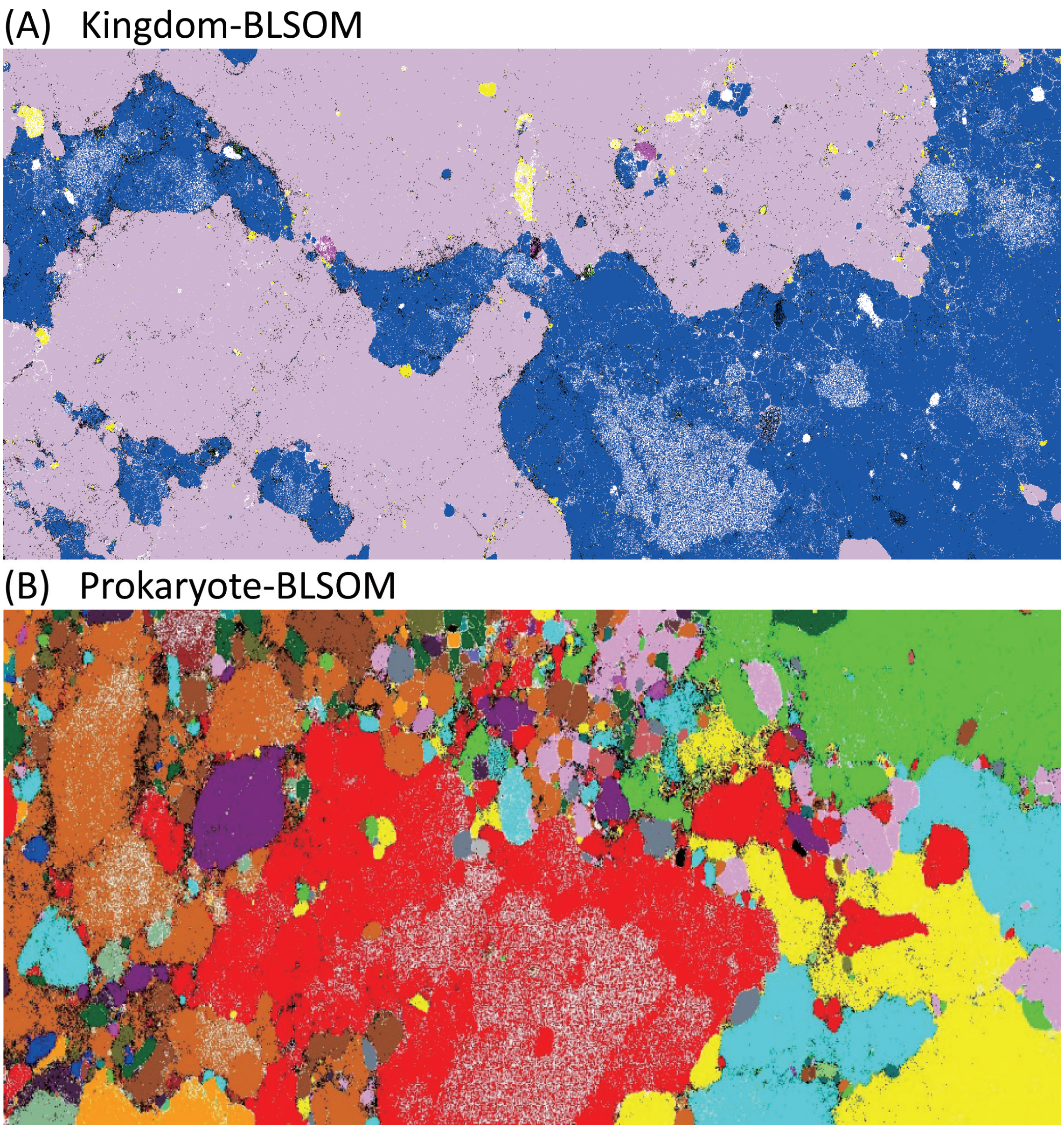


Fig. 2 BLSOMs for phylogenetic classification of environmental sequences. (A) Kingdom-BLSOM: DegTetra-BLSOM of 5-kb sequences derived from prokaryotic, eukaryotic and viral genomes currently available. In the Databanks, only one strand of complementary sequences is registered, and the strand is chosen rather arbitrarily in the registration of fragment sequences. Because the obtained BLSOM should not be affected by the choice of strands registered in the Databanks, we constructed a BLSOM, for which the frequencies of a pair of complementary oligonucleotides in each fragment were summed up [7]. The BLSOM for the degenerate set of a pair of complementary tetranucleotides is abbreviated as DegTetra. (B) Prokaryote-BLSOM: DegTetra-BLSOM of 5-kb sequences derived from species-known prokaryotes currently available. For colors, see our original paper [10].

4. Conclusion

Large-scale metagenomic analyses on environmental samples using recently released next-generation sequencers are actively underway on a global mass scale, and the obtained numerous sequences have been registered in the public databases. Large-scale computations using various, novel bioinformatics tools, which are suitable for big data analyses, are undoubtedly needed for efficient knowledge-findings from the massive number of genomic sequences accumulated in studies in the environmental geoscience. The present BLSOM can phylogenetically classify most genomic sequence fragments, based only on similarity in oligonucleotide composition, and visualize a microbial community structure in an environment on a two-dimensional plane, thus supporting an accurate comparison among a wide variety of environments [11-15].

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

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C, 26330327, 26330334) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The present computation was done with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] T. Kohonen, “The self-organizing map”, Proceedings of the IEEE, vol. 78, pp. 1464-1480, 1990.

[2] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, “Engineering applications of the self-organizing map”, Proceedings of the IEEE, vol. 84, pp. 1358–1384, 1996.

[3] T. Kohonen, *Self-Organizing Maps*. Berlin, Springer, 1997.

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

[5] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, “A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency”, Genome Inform., vol. 13, pp. 12–20, 2002.

[6] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, “Informatics for unveiling hidden genome signatures”, Genome Res., vol. 13, pp. 693–702, 2003.

[7] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, “Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples”, DNA Res., vol. 12, pp. 281–290, 2005.

[8] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, “Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes”, Gene, vol. 365, pp. 27–34, 2006.

[9] T. Abe, H. Sugawara, S. Kanaya, and T. Ikemura, “Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator”, Journal of the Earth Simulator, vol. 6, pp.17–23, 2006.

[10] Y. Iwasaki, T. Abe, K. Wada, Y. Wada, and T. Ikemura, “A Novel Bioinformatics Strategy to Analyze Microbial Big Sequence Data for Efficient Knowledge Discovery: Batch-Learning Self-Organizing Map (BLSOM)”, Microorganisms, vol. 1, pp. 137–157, 2013.

[11] 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 J., vol. 7, pp.1003–1015, 2013.

[12] M. Mitsumori, S. Nakagawa, H. Matsui, T. Shinkai, and A. Takenaka, “Phylogenetic diversity of gene sequences isolated from the rumen as analysed using a self-organizing map (SOM)”, Journal of Applied Microbiology, vol. 109, no. 3, pp.763–770, 2010.

[13] A. Kouzuma, T. Kasai, G. Nakagawa, A. Yamamoto, T. Abe, and K. Watanabe, “Comparative metagenomics of anode- associated microbiomes developed in rice paddy-field microbial fuel cells”, PLoS One, vol. 8, doi:10.1371/journal.pone.0077443, 2012.

[14] 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, doi:10.1371/journal.pone.0098425, 2013.

[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, 287–298, 2009.

全ゲノム・全タンパク質配列の自己組織化マップを用いた大規模ポストゲノム解析

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　　新潟大学　大学院自然科学研究科

和田健之介　　長浜バイオ大学　バイオサイエンス学部

池村　淑道　　長浜バイオ大学　バイオサイエンス学部

次世代シーケンサーの登場による DNA シークエンサーの最近の飛躍的な高速化に伴い、ゲノム配列のデータベースへの蓄積が加速されている。さらには、海洋や土壌等の様々な環境やヒト腸内などから取得される混合ゲノム試料を対象として、メタゲノム解析プロジェクトが世界的に進行しており、特に海洋はその主対象の一つである。全地球レベルでの解析が進行しており、ビッグデータ解析が重要となっている。我々が開発した一括学習型自己組織化マップ（BLSOM）は、断片ゲノム配列を生物種ごとに高精度に分離（自己組織化）する能力を持つ。解読済の全ゲノム配列を対象にして大規模 BLSOM を逐次に更新して行けば、メタゲノム解析で新たに得られる大量塩基配列をマップすることで、各環境中で生息する生物集団の全体像を正確に把握することが可能になり、併せて新規性の高い有用遺伝子類を発掘できる。大規模な BLSOM マップの更新を行いつつ、作成した BLSOM マップを用いてメタゲノム配列に対する系統推定を行うためのソフトウェア PEMS の公開を行い、ES での研究成果の普及を図った。さらに、国内外の共同研究者との共同研究にて本ソフトウェアを利用した研究成果の発表を行った。

キーワード : 自己組織化マップ, BLSOM, メタゲノム解析, オリゴヌクレオチド頻度, 生物系統推定, 微生物生態, 環境生態