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 Graduate school of Science and Technology, Niigata University

Shigehiko Kanaya Graudate School of Information Science, Nara Institute of Science and Technology

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), on the basis of batch-learning SOM, to 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 have established the alignment-free clustering method BLSOM that could analyze far more than 10,000,000 sequences simultaneously. Therefore, 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 this large-scale BLSOM and updated annually by analyzing all available genomic sequence data at that time. By mapping the metagenomic sequences obtained from a mixed genome sample on this large-scale BLSOM, we can predict phylotype compositions of environmental and clinical samples. BLSOM for oligopeptide compositions can also be used for prediction of protein functions on the basis of similarity in oligopeptide compositions of proteins.

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

1. Introduction

One of the most important tasks in the life and environmental science is to unveil unknown basic knowledge from big data of genomic sequences accumulated in the International DNA Databanks. Therefore, it is important to develop a novel bioinformatics tool for large-scale comprehensive studies of phylotype-specific characteristics of genomes, which can overview almost all available sequences from prokaryotic, eukaryotic 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], and we have modified the SOM for the genome analyses by developing a Batch-Learning SOM (BLSOM) [4]. We have used the BLSOM to analyze short oligonucleotide frequencies (di- to pentanucleotide frequency) in a wide range of prokaryotic and eukaryotic genomes [5-7].

Suppose only fragmental sequences (e.g., 10 kb sequences) derived from mixed genomes of multiple organisms are available, it appears impossible to identify how many and what types of genomes are present in these collected sequences. However, we found that BLSOM could classify the sequence fragments according to species without any information other than oligonucleotide frequencies. BLSOM recognized, in most sequence fragments, species-specific characteristics of oligonucleotide frequencies, 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 also important in the environmental science. 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 the most suitable method for this purpose.

When we consider phylogenetic classification of species-unknown sequences obtained from environmental and clinical samples, BLSOMs have to be constructed in advance with all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles. When ES is used, sequences can be clustered (self-organized) on BLSOM according to phylotypes with high accuracy. By mapping a large number of environmental genomic sequences on this large-scale BLSOM, we can predict phylotypes of these environmental sequences. Because BLSOM does not require orthologous sequence sets, the present alignment-free method can provide a systematic strategy for revealing microbial diversity and relative abundance of different phylotype members of uncultured microorganisms including viruses in an environmental sample.

2. Methods

BLSOM for oligonucleotide compositions and that for peptide composition were conducted as described previously [5, 10].

3. Results

3.1 Unveiling of microbial diversities within tick guts by analyzing metagenomic sequences

In this report, we introduce the study, which has been conducted under the collaboration with Prof. Sugimoto's group (Division of Collaboration and Education, Research Center for Zoonosis Control, Hokkaido University) and has been recently published by the ISME Journal [11].

Ticks in various regions of the world are vectors for bacterial, viral and protozoal pathogens. Ticks may act not only as vectors but also as reservoirs of tick-transmitted microbes (e.g., certain Rickettsia spp. and Borrelia spp.). Because of this medical and social importance, it has become an urgent necessity to understand and survey the bacterial pathogens of the microbes in ticks. To decipher the content, diversity of the microbial gut community in various ticks, we analyzed metagenomic sequences from bacteria-enriched fraction prepared form 6 tick species and predicted microbial diversity and relative abundance of microorganisms, by using BLSOM. In order to analyze phylotypes of the metagenomic sequences in each tick sample, three types of large-scale BLSOMs, namely Kingdom-, Prokaryote-, and Genus group-BLSOM, were constructed in advance, using all genome sequences available from DDBJ/EMBL/GenBank.

Kingdom-BLSOM was constructed with tetranucleotide frequencies for 5-kb sequences derived from the genome sequences of 111 eukaryotes, 2,813 prokaryotes, 1,728 mitochondria, 110 chloroplasts, and 31,486 viruses. To get more detailed phylotype information for prokaryotic sequences, Prokaryote- and Genus group-BLSOM were constructed with a total of 3,500,000 of 5-kb sequences from 3,157 species, for which at least 10 kb of sequence was available from DDBJ/EMBL/GenBank (Fig. 1).

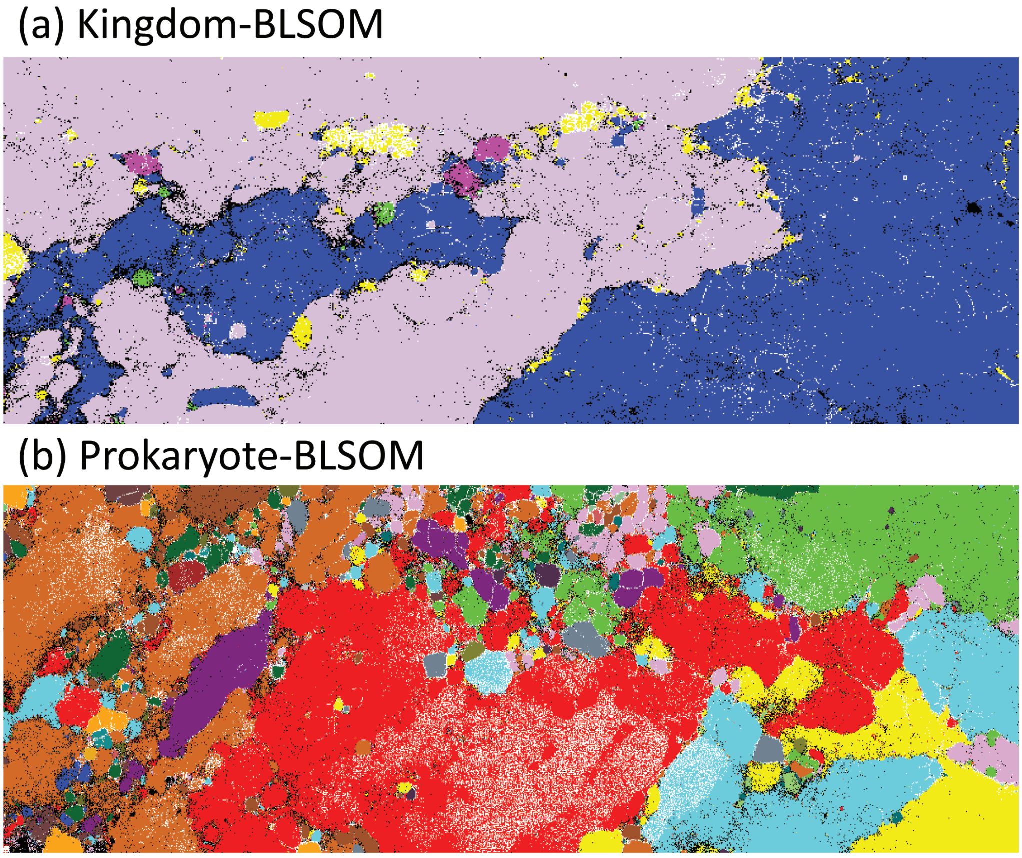

Fig. 1 BLSOMs for phylogenetic classification of environmental sequences. (a) Kingdom-BLSOM: DegeTetra-BLSOM of 5-kb sequences derived from prokaryotic, eukaryotic and viral genomes currently available. (b) Prokaryote-BLSOM: DegeTetra-BLSOM of 5-kb sequences derived from species-known prokaryotes currently available.

By the mapping on these BLSOMs, we could predict phylotypes of metagenomic sequences from guts of 6 individual ticks; the mapping of the metagenomic sequences longer than 300 bp on Kingdom-BLSOM, after normalization of the sequence length, was conducted by finding the lattice point with the minimum Euclidean distance in the multidimensional space (Fig. 2). To identify further detailed phylogenies of the metagenomic sequences that had been mapped to the prokaryotic territories on the Kingdom-BLSOM, these sequences were successively mapped on Prokaryote-BLSOM (Fig. 3). Similar stepwise mappings of metagenomic sequences on BLSOMs that were constructed with sequences from more detailed phylogenetic categories (e.g., phylum and genus) were conducted, in order to get further detailed phylogenetic

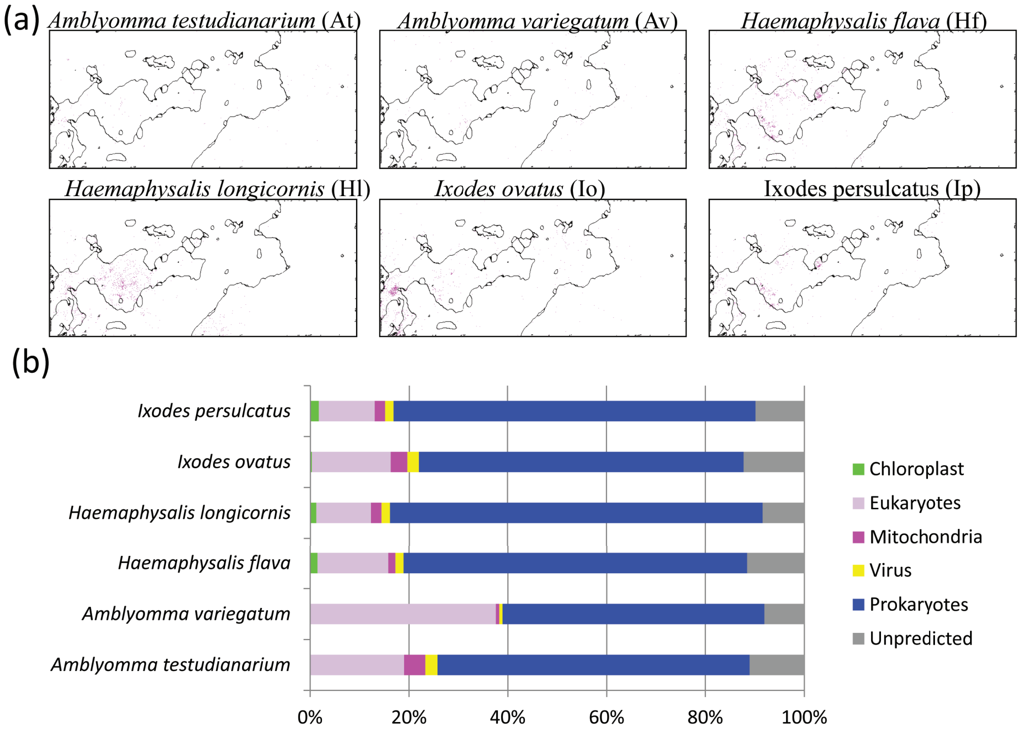

Fig. 2 Comparison of kingdom diversity among tick samples. (a) Mapping of tick metagenome sequences to kingdom-BLSOM. (b) Proportion of predicted kingdom category on basis of Kingdom-BLSOM.

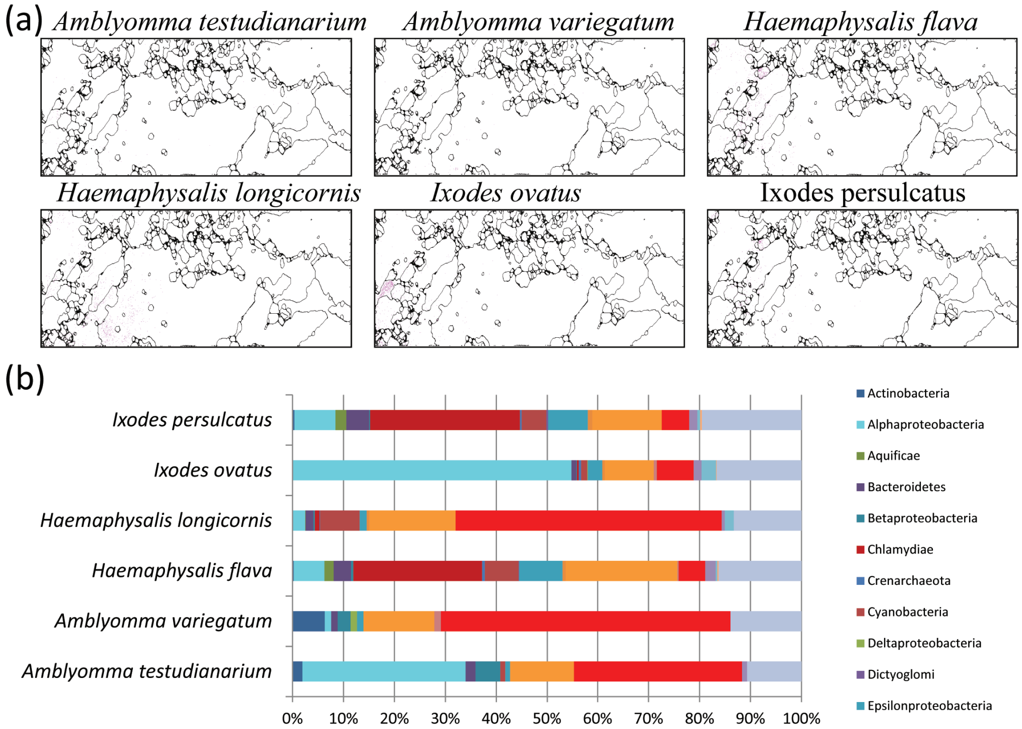

Fig. 3 Comparison of prokaryotic phylotypes diversity among tick samples. (a) Mapping of tick metagenome sequences to Prokaryote-BLSOM. (b) Proportion of predicted kingdom category on basis of Prokaryote-BLSOM.

information.

Each sample was dominated by different species, indicating that each tick was reservoirs of specific pathogens. In addition to bacteria previously associated with human/animal diseases, such as *Anaplasma, Bartonella, Borrelia, Ehrlichia, Francisella,* and *Rickettsia*, the present BLSOM analysis detected microorganisms belonging to the phylum Chlamydiae in some tick species.

3.2 Characterization of secondary metabolic enzyme groups based on peptide sequence

When we constructed BLSOM for peptide compositions of proteins, we found that proteins tended to cluster according to functions of proteins [10], indicating that the peptide-BLSOM should be usable for predicting functions of function-unknown proteins. In this report, we introduce the study, which has been conducted under collaboration with Prof. Kanaya's group (Nara Institute of Science and Technology) and recently been published by Ikeda et al. [12].

To compare diversities of enzymes of secondary metabolism based on the peptide BLSOM, we initially made a map reflecting the diversity of peptide sequences based on dipeptide frequencies. Amino acid sequences were obtained from Non-Redundant Protein Sequences of PlantGDB (http://www.plantgdb.org/download/download.php/) and the clusters of orthologous groups (COGs) were from http://www.ncbi.nlm.nih.gov/COG/. Initially, we selected proteins longer than 200 amino acids and divided each protein into segments of 200 amino acids starting from N-termini with 50 amino-acid step size. Consequently we obtained 4,892,003 peptide fragments from 721,266 protein sequences (596,974 proteins of 59,165 plants and 124,292 protein sequences of 66 bacterial species). A 1,752,300×400 data matrix was then created by counting the frequency of 400 dipeptides in the fragments and then this matrix was utilized for constructing BLSOM. Then, we assessed diversity of enzymes concerning secondary metabolic pathways including terpenoids, alkaloids, flavonoids and cytochrome P450 enzymes.

Terpenes are the largest group of plant natural products with variety of core chemical structures comprising at least 30,000 compounds. Terpene diversity is caused by the large number of different terpene synthases corresponding to the first step to synthesize terpenes and some terpene synthases produce multiple products. Figure 4a shows distribution of fragments

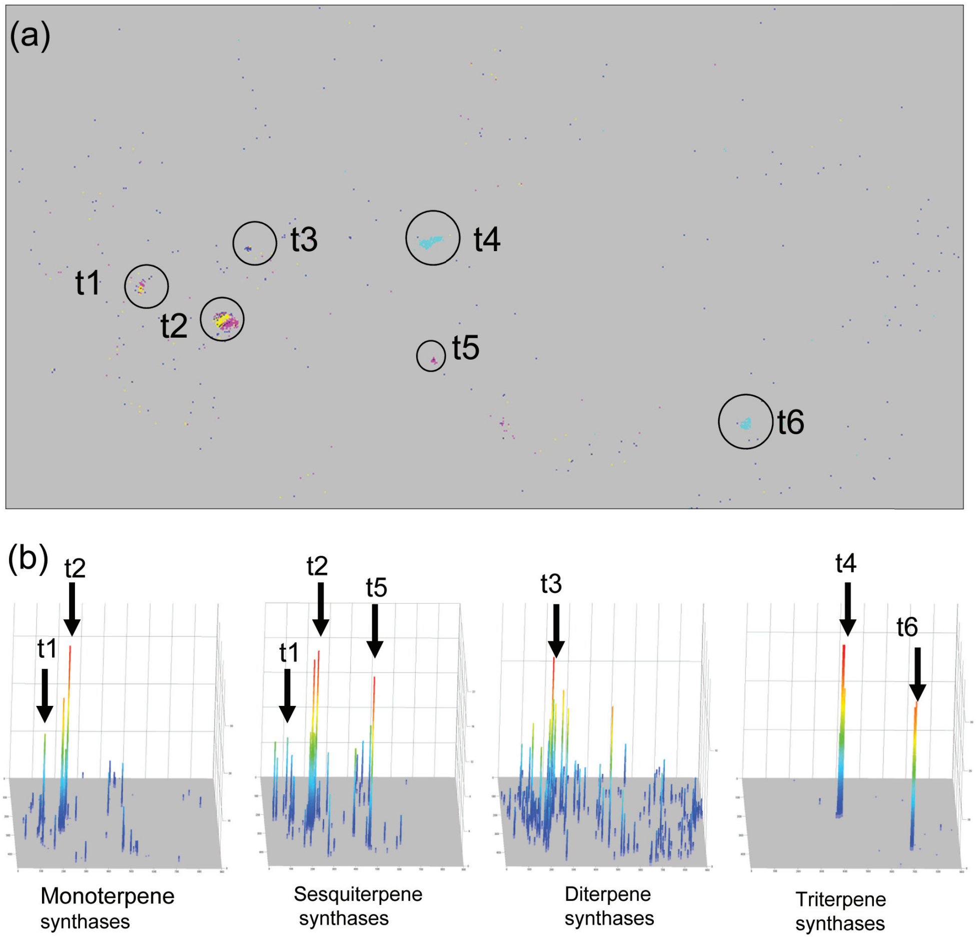

Fig. 4 Four groups of terpene synthases are plotted in the map. (a) Self-organizing map for four terpene synthases (lattice points occupied by monoterpene, sesquiterpene, diterpene, and triterpene synthases are represented by yellow, pink, blue and sky blue, respectively and six clusters with abundant fragments are denoted by c1 to c6.). (b) 3D bar graph indicating counts of monoterpene, sesquiterpne, diterpene, and triterpene synthases, respectively.

of terpene synthases which mainly clusterized in six regions t1 to t6. Figure 4b shows distribution of those in four terpene synthases. It is noted that three types of terpene synthases except diterpene synthases are less diverged in peptide sequence level, that is, small changes in peptide sequences in terpene synthases; this makes it possible to synthesize a much diverged terpenoid compounds. Those properties can be explained in the projection of terpene synthases to BLSOM.

4. Conclusion

Large-scale metagenomic analyses on environmental samples using recently released next-generation sequencers are actively underway on a global basis, and the obtained numerous 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 (i.e. big sequence data).

The present BLSOM is an unsupervised algorithm that can separate most sequence fragments based only on similarity in oligonucleotide composition. 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 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 predict functions of proteins, even for which the sequence similarity search at an amino-acid level cannot predict functions. This is because proteins with the same or similar functions can be clustered (self-organized) primarily according to functions on BLSOM for oligopeptide composition.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C, 23500371) and for Young Scientists (B, 23710242) 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] 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.

[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] 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.

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

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　　新潟大学大学院　自然科学研究科
金谷　重彦　　奈良先端科学技術大学院大学　情報科学研究科
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。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。本年の Annual Report では、北海道大学の杉本研究室と行った、ダニの体内のメタゲノム配列の BLSOM 解析の結果を紹介した。ダニを媒介する新規感染症が我が国でも報告されており、時宜を得た解析となっている。

データベースへの蓄積の著しい、機能未知なタンパク質の機能推定を可能にする新技術として、昨年度までにオリゴペプチドの BLSOM を確立したが、この技術開発では、原核生物由来のタンパク質を対象にしてきた。タンパク質機能推定法の更なる産業的利用を目指し、植物の２次代謝物推定に着目し、現時点で知られている植物の全タンパク質を対象にした大規模 BLSOM の作成を行った。本年の Annual Report では、奈良先端大の金谷研究室と行った、植物の二次代謝物の生産に係わる酵素類に関する BLSOM 解析の結果を紹介した。植物の二次代謝物は医薬やサプリメントとして利用価値が高く、産業的にも注目されている。

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, オリゴペプチド頻度, タンパク質機能推定