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

Self-Organizing Map (SOM) developed by Kohonen is an effective tool for clustering and visualizing high-dimensional complex data on a two-dimensional map. We previously modified the conventional SOM to genome informatics, making the learning process and resulting map independent of the order of data input [1, 2]. The BLSOM thus developed on the basis of batch-learning SOM became suitable for actualizing high-performance parallel-computing with a high-performance vectorial supercomputer [3]. The BLSOM revealed species-specific characteristics of oligonucleotides (e.g., tetranucleotides) frequencies in individual genomes, permitting clustering (self-organization) of genome fragments (e.g., 10 kb or less) according to species without species information during the calculation. Using ES, sequence fragments from almost all prokaryotic, eukaryotic, and viral sequences currently available could be classified (self-organized) according to phylotypes. Utilizing results of this large-scale BLSOM, phylotypes of a massive amount of genomic fragments obtained by metagenome analyses can be predicted.

We developed also the BLSOM method to predict protein function on the basis of similarity in oligopeptide composition (di-, tri- and tetrapeptide compositions in this study) of proteins. Oligopeptides are component parts of a protein and are involved in formation of functional motifs and structural parts of proteins. Concerning the oligopeptide frequencies in the 110,000 proteins, which had been classified into 2853 function-known COGs (clusters of orthologous groups of proteins), BLSOMs could faithfully reproduce the COG classifications. Proteins whose functions have been unknown because of lack of significant global sequence homology to function-known proteins detectable with conventional sequence homology searches (e.g. BLAST), could be related to function-known proteins using the large-scale BLSOM.

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

1. Introduction

Unculturable environmental microorganisms should contain a wide range of novel genes of scientific and industrial usefulness. Recently, sequencing analyses of mixed genome samples that directly extracts the mixed genome DNA of uncultured environmental microorganisms, i.e., metagenome analyses, have been established. A large portion of the environmental sequences has been registered in the International DNA Sequence Databanks with almost no functional and phylogenetic annotation, and therefore, in a less useful manner.

The homology search for nucleotide and amino-acid sequences such as BLAST has become widely accepted as a basic bioinformatics tool not only for phylogenetic characterization of gene/protein sequences but also for prediction of their biological functions when genomes and genomic segments are decoded. Whereas usefulness of this sequence homology search is apparent, it has became clear that homology search can predict the protein function of only 50% of genes, or fewer, when a novel genome is decoded. To complement the sequence homology search, it is urgently required to establish methods for predicting protein functions based on different principles.

Previously, we developed a batch-learning SOM (BLSOM) that depends on neither the order of data input nor the initial conditions, for oligonucleotide frequencies in genome sequences [1–4]. The BLSOM recognized species-specific characteristics of oligonucleotide frequencies in individual genomes, permitting clustering of genome fragments according to species without the need for species information during the calculation. This BLSOM was suitable for actualizing high-performance parallel-computing with a high-performance supercomputer [3–5]. In the present report, we describe use of the BLSOM method not only for the phylogenetic classification of genomic fragments but also for prediction of protein function on the basis of similarity in composition of oligopeptides of proteins.

2. Methods

Nucleotide sequences were obtained from URL (http://www.ddbj.nig.ac.jp/anoftp-e.html) and amino acid sequences were from URL (http://www.ncbi.nlm.nih.gov/COG). We modified the conventional SOM 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 on the basis of the finding that PCA can classify gene sequences into groups of known biological categories. The genomic sequences were analyzed as described previously [1–4].

For protein sequence analyses, we provided a window of 200 amino acids that is moved with a 50-amino acid step for proteins longer than 200 amino acids. BLSOM with tripeptide frequency (20^3 = 8000 dimensional data) required very long computation times, which exceeded the limit available for our group. To reduce the computation time, BLSOM was constructed with the tripeptide frequencies of the degenerate eleven groups of residues categorized according to the physico-chemical properties; {V, L, I}, {T, S}, {N, Q}, {E, D}, {K, R, H}, {Y, F, W}, {M}, {P}, {C}, {A}, and {G}; 11^3 = 1331 dimensional data. We analyzed also the tetrapeptide frequencies of degenerate six groups of residues; {V, L, I, M}, {T, S, P, G, A}, {E, D, N, Q}, {K, R, H}, {Y, F, W}, and {C}; 6^4 = 1296 dimensional data.

3. Results

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

A large-scale metagenomics study of uncultivable microorganisms in environmental and clinical samples should allow extensive surveys of genes useful in medical and industrial applications and assist in developing accurate views of the ecology of uncultivable microorganisms. Traditional methods of phylogenetic classification have been based on sequence homology searches and therefore inevitably focused on well-characterized genes (e.g., rDNA), for which orthologous sequences from a wide range of phylotypes are available. However, most of the well-characterized genes are not industrially attractive. It would be best if microbial diversity could be assessed during the process of screening for novel genes with industrial and scientific significance. An unsupervised clustering method BLSOM is thought to be the most suitable method for this purpose [4–7]. When we consider phylogenetic classification of species-unknown sequences obtained from environmental and clinical samples, it is 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 known 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. On the basis of our previous study on phylogenetic classification of prokaryotic sequences, BLSOM was constructed with frequencies of degenerate sets of tetranucleotides (DegeTetra-SOM) in 5-kb sequence fragments where the frequencies of a pair of complimentary tetranucleotides (e.g. AAAC versus GTTT) were added [4]. In the present study, using the ES, we could analyze almost all genomic sequences available from 2813 prokaryotes, 111 eukaryotes, 31486 viruses, 1728 mitochondria, and 110 chloroplasts. The 2813 prokaryotes were selected because at least 10-kb genomic sequences were registered in DDBJ/EMBL/NCBI. 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. Therefore, in the cases of higher eukaryotes, 5-kb eukaryotic sequences were selected randomly from each genome up to 25 Mb. This enabled us to analyze an equivalent number of prokaryotic and eukaryotic 5-kb sequences, and DegeTetra-SOM was constructed with the 5-kb sequences (Fig. 1A). The power of BLSOM to separate prokaryotic and eukaryotic sequences from each other was very high (ca. 97% accuracy). We also observed the clear separation of prokaryotic sequences into 28 major prokaryote families (Fig. 1B), confirming our previous study [3]. The separation of eukaryotic sequences according to species was also observed (data not shown).

3.2 A large-scale BLSOM constructed with almost all available protein sequences derived from the function-known COG categories

Next we introduce use of BLSOM for prediction of protein function on the basis of similarity in composition of oligopeptides of proteins. For the test dataset to examine whether proteins are clustered (i.e., self-organized) according to function by BLSOM, we chose proteins that had been classified into function known 2853 COGs by NCBI [8, 9]. Dipeptide composition (20^2 = 400 dimensional vectorial data) in 110,000 proteins belonging to the 2853 COG categories was investigated by BLSOM. In addition to the BLSOM for the dipeptide composition of 20 amino acids (abbreviated as Di20-BLSOM), we tested the BLSOM for the dipeptide or tripeptide composition after classification into 11 groups according to the physico-chemical properties of amino acids (see **Methods**); 121(= 11^3) or 1331(= 11^3) dimensional data (abbreviated as Di11- or Tri11-BLSOM,

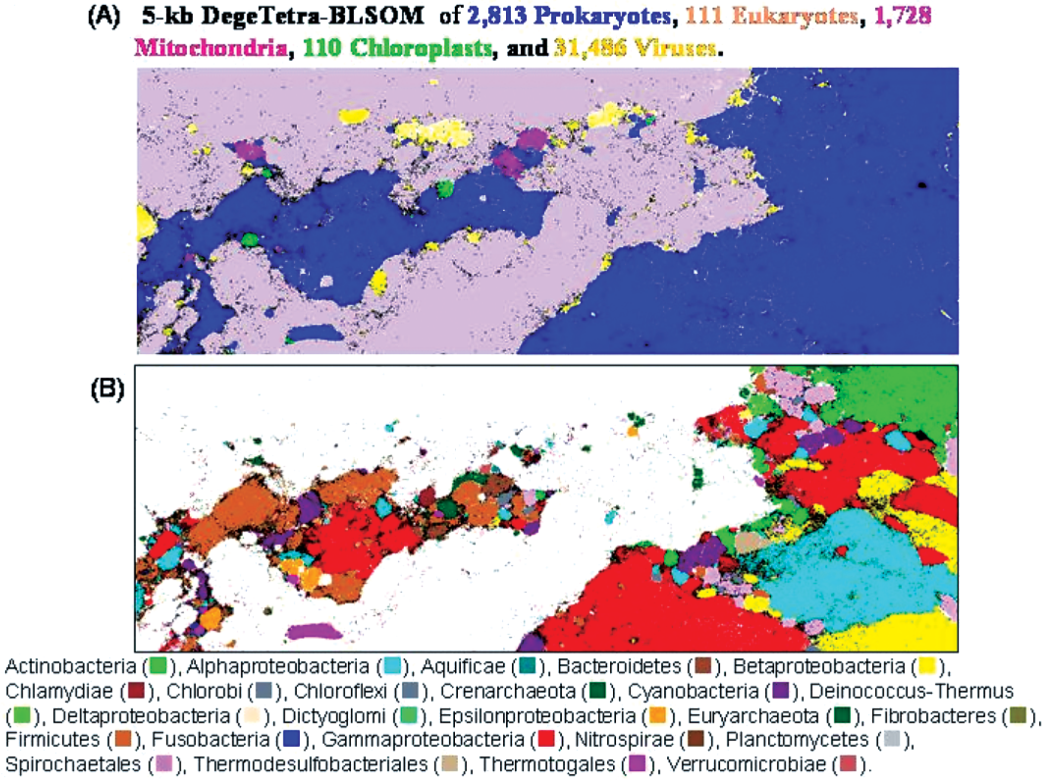

Fig. 1 5-kb DegeTetra-BLSOM for almost all genomic sequences available in DDBJ/EMBL/GenBank. (A) Nodes that contain only the sequences from prokaryotes (🟦), eukaryotes (🟪), viruses (🟨), mitochondria (🟪), or chloroplasts (🟩) were separately colored, and those containing the sequences of more than one category are marked in black. (B) Nodes that contain only the sequences from one prokaryotic family were in colors shown at the bottom of the figure, and those containing the sequences of more than one family are marked in black.

respectively). We also tested the tetrapeptide composition after classification of amino acids into 6 groups; 1296 (= 6^4) dimensional data (Tetra6-BLSOM). These four different BLSOM conditions were examined to what degree the similar results were obtained among the four conditions and which gave the best accuracy. It should be noted that BLSOMs for much higher dimensional data such as those for the tripeptide composition of 20 amino acids (8000-dimensional data) and for the tetrapeptide composition after grouping 11 categories (14641-dimensional data) was difficult in the present study because of limitation of ES resources available for our group.

To establish a method that is less dependent on the amino acid sequence length, we provided a window of 200 amino acids that is moved with a 50-amino acid step for proteins longer than 200 amino acids, and the BLSOM was constructed for the overlapped 200-amino acid sequences. Introduction of a window with a shifting step enable us to analyze multi-functional and -domain proteins, which originated often from the fusion of distinct proteins during evolution, collectively with smaller proteins.

One important point of this test data analysis is at what level each lattice-point on a BLSOM grid contains fragments derived from a single COG. The number of COG groups analyzed is 2835, and the size of the BLSOM was established so as to provide 8 data points per lattice-point. If sequences were randomly chosen, the probability that all fragments associated to one lattice-point were derived from a single COG by chance should be extremely low, e.g. $(1/2853)^8 = 2.3 \times 10^{-28}$, while this value depends on the number of fragments derived from proteins belonging to the respective COG. We designate here the lattice-point that contained fragments derived only from a single COG as "pure lattice-point"; in this definition, the lattice-points that contained only one sequence fragment were not included. Considering that an occurrence probability of pure lattice-point as an accidental event is extremely low, a high percentage of pure lattice-points was observed. The highest occurrence level of pure lattice-points was observed on the Tri11-BLSOM; approximately 45% of lattice-points contained sequences derived only from a single COG (Fig. 2). To graphically show the difference among these BLSOMs, pure lattice-points were colored as red, those contained sequence fragments derived from two different COGs were colored as pale red, and those from more than three COGs were colored as blue (Fig. 3A-D). This again showed clear clustering (self organization) of proteins according to a COG category.

3.3 Mapping of protein sequences from environmental samples on large-scale BLSOMs constructed for all NCBI-COG sequences

The most important contribution of the present BLSOM method is thought to predict functions of increasingly vast amount of function-unknown proteins derived from the less characterized microbe genomes as those found in the metagenomic approaches. To test the feasibility of BLSOMs for function prediction of such environmental proteins, we next focused on protein candidates that were found from about 1 million genomic fragments derived from metagenome libraries originating in the Sargasso Sea [10] and compared the results obtained by the conventional sequence homology search and the BLSOM method, in the following way. In the first step, using the conventional sequence homology search BLAST, we searched for Sargasso proteins (> 200 amino acids) which showed significant global homology with NCBI-COG proteins on a criterion that 70% or more identity of the amino acid sequence was observed over 70% region of the Sargasso protein; this criteria is analogous to that used for the NCBI-COG identification. A total of 4240 Sargasso protein sequences was thus found, and were tentatively called as Sargasso-COG proteins. The 4240 Sargasso-COG proteins were divided into 200-amino acid segments with a sliding step of 50 amino acids and these 200-amino acid segments were mapped on each of the BLSOM constructed in advance with NCBI-COG proteins such as those in Fig. 2. Then, as to each lattice-point on which Sargasso fragments were mapped, the most abundant NCBI-COG protein were identified, and the mapped Sargasso segments were tentatively assumed to belong to this most abundant COG category. By summing up this tentative assignment data of 200-amino acids segments for each of the 4240 Sargasso-COG protein, each Sargasso protein was finally assigned to the most abundant COG category among COGs tentatively assigned.

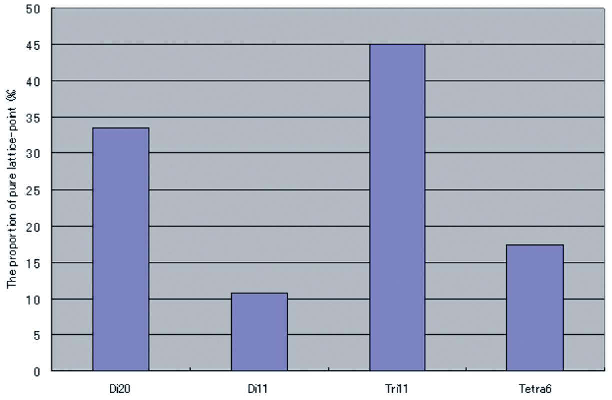

Fig. 2 The proportion of pure lattice-point for each analysis condition.

While BLSOM is an alignment-free clustering method, which is clearly different from sequence homology searches, 91, 87 or 79% of the 4240 Sargasso COG proteins were assigned to the original COG categories on Tri11-, Di20- or Tetra6-BLSOM, respectively. As expected from the results of Fig. 2, the highest identity level was obtained on Tri11-BLSOM. Detailed inspection of the miss-assigned cases showed that different COGs with similar functions were confounded. This may help us to construct similarity map of COGs on the basis of oligopeptide composition.

In the next analysis, to attempt to predict functions of

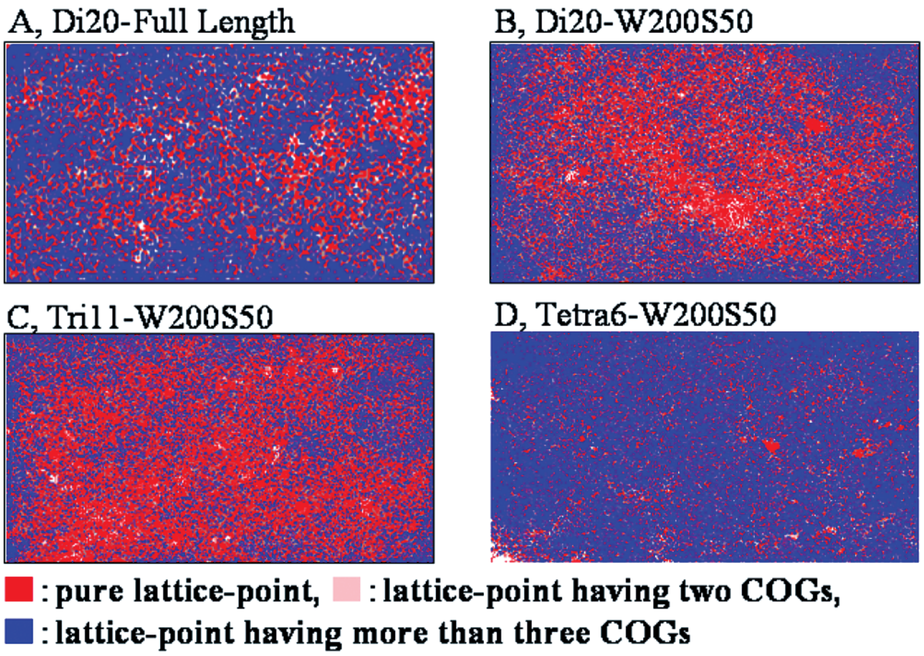

Fig. 3 Level of pure lattice-point for different BLSOM conditions.

Sargasso proteins that were not detected by the conventional sequence homology search, we mapped 200-amino acid segments derived from all Sargasso proteins longer than 200 amino acids on BLSOMs, and more than 3000 Sargasso proteins could be newly assigned to COG categories. We plan to publicize the results of the assignments obtained concordantly with three BLSOM conditions (Tri11-, Di20-, Tetra6-BLSOMs).

4. Conclusion and Perspective

We established a method of phylogenetic prediction for individual genomic fragments obtained by metagenomic analysis, by using BLSOMs of oligonucleotide frequencies. We introduced also the BLSOM method for predicting functions of proteins found by metagenomic analyses. For function-unknown proteins for which the consistency of the predicted function is observed by BLSOMs for the frequencies of dipeptides, tripeptides, and tetrapeptides, their predicted functions are thought to be reliable. For these large-scale BLSOM analyses, use of the high-performance supercomputer ES is essential. The data obtained only by ES are unique datasets in genomics and proteomics fields and provide a guideline for research groups including those in industry to study functions of novel genes with scientific and industrial usefulness through experiments.

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] 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.

[2] 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.

[3] 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.

[4] 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.

[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] 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.

[8] R.L. Tatusov, E.V. Koonin, and D.J. Lipman, "A genomic perspective on protein families", *Science*, vol.278, pp.631–637, 1997.

[9] R.L. Tatusov, et al., "The COG database: an updated version includes eukaryotes", *BMC Bioinformatics*, vol.4, pp.41.

[10] J.C. Venter, et al. 2004, Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, 304, 66–74.

データベースに蓄積の著しい機能未知のタンパク質類の機能推定のための一括学習型の自己組織化マップ法

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

ゲノム配列の解読は飛躍的に加速しており、約900の生物種のゲノム配列が公開され、4,000近くのゲノムプロジェクトが進行中である。さらには、140の環境由来の混合ゲノムを対象にしたメタゲノム解析も進行しており、2000万件を超えるゲノム断片配列がデータバンクに登録されているが、大半の配列について由来する生物種も遺伝子機能も不明で、利用価値が低いままに残されている。本研究では、部分的なゲノム配列しか解読されていない生物種も含めて、現時点でデータバンクに登録されている大半のゲノム配列を対象にした大規模BLSOM解析を行った。具体的には、2813原核生物、111真核生物、31486ウイルス、1728ミトコンドリア、110葉緑体に由来する5kb断片配列に関する4連塩基頻度のBLSOMを行い、生物系統を正確に反映した分離（自己組織化）を得ている。メタゲノム解析で得られたゲノム断片配列の生物系統の推定法が確立できた。

アミノ酸配列の相同性検索法は、ゲノムが解読された際に、各タンパク質遺伝子の機能を推定する基本技術として利用されている。この有用性が明らかになる一方で、新規性の高いゲノムが解読された際には、配列相同性検索で機能が推定できないタンパク遺伝子は半数近くに及ぶことも明らかになった。タンパク質の機能については、機能部品類の3次元上での立体配置が重要であり、同一ないしは類似の機能を持つタンパク質間でも、アミノ酸の1次元配列上での全域に渡っての有意な相同性を見付けられない例が多い。この視点からX線結晶構造解析やNMR法でタンパク質の3次元構造を決定し、機能既知タンパク質との高次構造上の類似性で機能を推定する大規模なプロジュクトが推進されてきた。しかしながら、費用や労力ならびに技術上の限界から、今後ますます急増する膨大な数の機能未知タンパク質類の機能推定には不十分と考えられる。配列相同性検索を補完する、異なった原理に基づくタンパク質の機能推定法の確立が急務と言える。

我々はタンパク質の2～4連アミノ酸頻度を対象にしたBLSOM解析を開発した。本研究開発では微生物を中心とした機能カテゴリー別のデータベースであるCOG (Cluster of Orthologous Group) に収録されたタンパク質を対象にして、機能が特定されているタンパク質類を解析に用いた。2連アミノ酸頻度ならびに、20のアミノ酸を物理化学的な類似性で11のカテゴリーに集約した上での3連アミノ酸頻度、ならびに6カテゴリーに集約した上での4連アミノ酸頻度に着目して、BLSOM解析を行った。タンパク質が機能や構造により自己組織化する傾向を示し、特に20のアミノ酸を物理化学的な性質の類似度で、11カテゴリーへグループ化した3連アミノ酸頻度のBLSOMは機能に基づく分離の度合いが最も高かった。相同性検索に依存しないタンパク質の機能推定法として有用性の高い新規手法を確立できたので以下の解析を行った。メタゲノム法で得られる塩基配列は新規性が非常に高く、遺伝子機能に関するアノテーションもほとんどついておらず、利用価値が低いままに残されている。機能既知の大量なタンパク質とメタゲノム解析で得られた多数の機能未知のタンパク質を混合した集合データを対象に、大規模BLSOM解析を行い、メタゲノム解析で得られた多数のタンパク質の機能推定を行った。2連アミノ酸頻度、11に集約した3連アミノ酸頻度、6に集約した4連アミノ酸頻度の解析で共通して同じCOG機能が推定できるが、相同性検索では機能が未知に分類されるタンパク質が既に約3000件得られている。世界的に類例の無い、タンパク質機能推定のデータセットであり、実験グループが機能を証明する実験を行う上での指針を提供できる。有用遺伝子を探索している産業界からの期待も大きい。

キーワード：自己組織化マップ, BLSOM, 環境微生物, オリゴペプチド頻度, タンパク質機能推定, バイオインフォマティクス