A Large-scale Batch-learning Self-organizing Map for Function Prediction of Poorly-characterized Proteins Progressively Accumulating in Sequence Databases

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

As a result of decoding of extensive genome sequences, a large number of proteins whose function cannot be predicted by the homology search of amino acid sequences is progressively accumulated and thus remains of no use in science and industry. A method to predict the protein function that does not depend on the sequence homology search is in urgent need. We previously developed a Batch-Learning SOM (BLSOM) for genome informatics, and in the present report, we describe use of the BLSOM method for prediction of protein function on the basis of similarity in composition of oligopeptides (di-. tri- and tetrapeptide in this study) of proteins. Oligopeptides are elementary components of a protein and involved in formation of functional motifs and structural organization of proteins. BLSOM for oligopeptides could extract characteristics of oligopeptide composition actualizing protein structure and function and thus predict functions.

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

1. Introduction

The development of DNA sequencing technology has drastically accelerated the process of genome sequencing, and the genomes of more than 750 species, ranging from eukaryotes including human, to prokaryotes including *Escherichia coli*, lactic acid bacteria, and various archaea, have been completely sequenced and published. Furthermore, nearly 3,000 genome projects are currently in progress. Homology search of base and amino acid sequences has become a basic analysis method in the bioinformatics field that is essential not only for analyzing the evolution and phylogeny of genes and proteins but also for predicting the function of each protein gene in the sequenced genomes. While sequence homology search is obviously useful, it has also been revealed that this method cannot predict protein functions for almost half of genes in newly sequenced genomes, especially of novel species. For protein functions, three-dimensional configurations of functional components, which are composed of oligopeptides, have great significance, and therefore, there are many cases in which no significant homology is found over the entire one-dimensional sequence of amino acids between proteins with identical or similar functions. Considering this fact, large-scale projects have been promoted to determine the three-dimensional structure of function-unknown proteins by X-ray crystallographic analysis and NMR methods and to predict their functions based on the similarity of higher-level structure with function-known proteins. However, with limitations in terms of cost, workforce, and technology, such projects may be inadequate to predict the functions of function-unknown proteins, especially those of environmental microorganisms that will be increasingly identified in the near future.

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 (1–5). 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 such as the Earth Simulator (3–5). In the present report, we describe use of the BLSOM method for prediction of protein function on the basis of similarity in composition of oligopeptides of proteins.

2. Methods

Amino acid sequences were obtained from URL (http://www.ncbi.nlm.nih.gov/COG). Proteins shorter than 200 amino acids in length were not included in the present study. 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 tripeptide frequencies of the degenerate eleven groups of residues; {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): and with 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).

In the case of the conventional SOM, the initial vectorial data are set at a random value but this results in a final map which is changed by each initial data set and thus inconvenient for interpretation of map results. In our previous BLSOM analyses for genome sequences (1–5), we obtained a reproducible map by using the first and secondary primary components in the PCA analysis of oligonucleotide composition for the initial vectorial data. We again used this strategy in the present study. When the PCA method was applied to dipeptide composition in proteins in a preliminary study, the first component tended to reflect the length of proteins. While the length of a protein undoubtedly relates to its function, the length sometimes differs significantly even between proteins with the same function. Furthermore, lengths of proteins with similar functions are known to differ significantly from each other and eukaryotes have many multifunctional and -domain proteins. Even prokaryotes have large proteins which originated often from the fusion of different proteins during evolution. Because the main purpose of the present BLSOM method is to predict functions of proteins obtained from rapidly accumulated sequences derived from a wide range of novel phylotypes, we introduced a method that is less dependent on the length of proteins, providing a window of 200 amino acids that is moved with a 50-amino acid step for proteins longer than 200 amino acids. BLSOMs were constructed with di-, tri- and tetrapeptide frequencies in the overlapped 200-amino acid sequences.

3. Results

For the dataset to examine whether proteins are clustered (i.e., self-organized) according to function by BLSOM, we chose proteins that have been classified into function known 2,853 COGs (clusters of orthologous groups of proteins) by NCBI (6, 7), which have been identified on the basis of an all-against-all sequence comparison of the proteins encoded in complete genomes using the sequence homology search (6, 7). Proteins belonging to a single COG have significant homology of amino acid sequences over the whole range of the proteins and most likely have the same function. We prepared BLSOMs under various conditions to search for conditions that would most faithfully reproduce the COG classification. One important criterion of the separation according to functional category is at what level individual nodes on a BLSOM contain 200-amino acid fragments derived from proteins belonging to a single COG category. Dipeptide composition (20^2 = 400 dimensional vectorial data) in ca. 120,000 proteins belonging to the 2853 function known COGs was first investigated. In addition to the BLSOM for the dipeptide composition (abbreviated as Di20-SOM), the BLSOM for the tripeptide composition after classification into 11 groups, 1331(= 11^3) dimensional data (abbreviated as Tri11-SOM), or tetrapeptide composition after classification into 6 groups, 1296 (= 6^4) dimensional data (abbreviated as Tetra6-SOM). These three different BLSOM conditions were examined how similar prediction results were obtained and which gave the best accuracy.

The size of BLSOMs was chosen so as to provide a mean ca. 8 sequences per node. The average probability that all 8 sequences were derived from a single COG category 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 designated here the node that contained fragments derived only from a single COG as "pure node". Considering the probability of occurrence of a pure node as an accidental event to be extremely low (e.g., 2.3×10^{-28}), we compared the occurrence level of pure nodes on different BLSOMs. Even no COG information was given during calculation, high percentages of correct clustering (self organization) of proteins according to the COG category was observed on all three BLSOMs, and the highest occurrence of pure nodes was observed on the Tripeptide BLSOM (Tri11-SOM). Approximately 45, 34 and 17% of nodes of Tri11-, Di20- and Tetra6-SOMs, respectively, contained sequences derived only from a single COG. Concerning these pure nodes, 20 examples of clustering according to COG on these BLSOMs are shown in Fig. 1, where the number of sequences associated with each pure node (and thus derived from a single COG) was shown with the height of the vertical bar with a color representing each COG. Sequences belonging to a single COG were localized often in the neighbouring nodes, resulting in a high peak compose of adjacent high bars.

The figures show that classification (self-organization) with high accuracy has been achieved for each functional category. There also observed a few high peaks located far apart from each other. Detailed inspection showed that these detached high peaks are mostly due to the different 200-

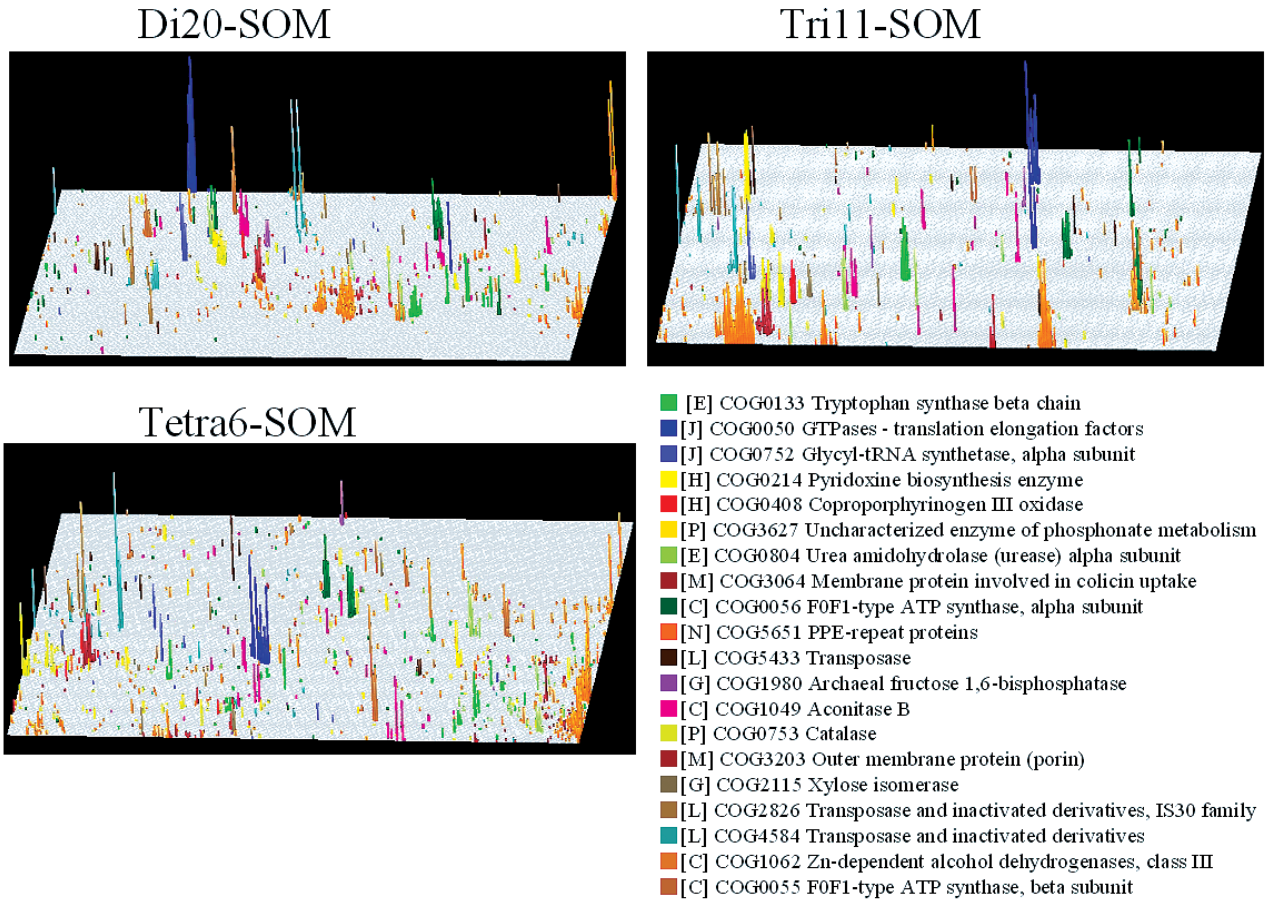

Fig. 1 Clustering of protein sequences according to COG (20 examples).

amino acid segments (e.g, anterior and posterior portions) derived from one protein, which have distinct oligonucleotide compositions and possibly represented distinct structural and/or functional domains. This type of distinct, major peaks may be informative for prediction of functions of multifunctional multidomain proteins. Based on the classification with high accuracy for each functional category, BLSOM is shown to be a powerful tool for function prediction of function-unknown proteins.

4. Conclusion and Perspective

Recently, a sequencing method for mixed genome samples that directly extracts the mixed genome DNA of uncultured microorganisms from environmental samples without cultivation (metagenomic analysis) is increasingly used. Sequences determined by metagenomic analysis are very novel, and their data are registered to international DNA sequence databases with almost no annotation regarding gene functions, and therefore, in a useless manner. Previously, we reported a method of phylogenetic prediction using BLSOMs of oligonucleotide frequencies for genome fragments determined by metagenomic analysis. In the present study, we introduced the BLSOM method developed for conducting function prediction for protein sequences determined by metagenomic analysis. For the verification of its usefulness, we have used the data of protein sequences among metagenome sequences from the Sargasso Sea (manuscript in preparation).

The prediction of functions of function-unknown proteins requires the BLSOM analysis that covers all function-known proteins in databases in advance. In addition, the analysis on all proteins, including function-known and unknown proteins, on a single BLSOM is valuable. To understand which function-unknown proteins self-organize with which function-known proteins provides an important guideline for function estimation. For the large-scale BLSOM analysis, use of high-performance supercomputers is essential. In the cases of function-unknown proteins for which the consistency of the predicted function is determined by analyzing dipeptide, tripeptide and tetrapeptide frequencies, their predicted functions will be published. Such data are unique datasets for function estimation and provide a guideline for research groups of academia and industries to prove the functions of proteins through experiments.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) and for Grant-in-Aid for

Scientific Research on Priority Areas "Applied Genomics" 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] 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.

[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, "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.

[5] 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.

[6] RL. Tatusov, EV. Koonin, DJ. Lipman, "A genomic perspective on protein families", *Science*, DJ. vol.278, pp.631–637, 1997.

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

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

塩基配列解読技術の発展に伴い、ゲノム配列の解読は飛躍的に加速しており、ヒトを始めとする真核生物から微生物まで、750種を超える生物種のゲノム配列解読が公開されており、3,000近くのゲノムプロジェクトが全世界で進行中である。塩基やアミノ酸配列の相同性検索法は、これらの解読されたゲノムの各タンパク質遺伝子の機能推定に不可欠の技術として利用され、バイオインフォマティクスの基本手法となった。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、配列相同性検索でタンパク質の機能が推定できない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ェ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な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連や3連アミノ酸頻度を対象にしたBLSOM解析を開発した。本解析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おり、機能が既知な機能カテゴリー (COGID数は2,853：配列数は113,738) に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。windowを設けることで、通常の大きさのタンパク質と大型タンパク質とを同時に解析することが可能となる。

地球シミュレータを用いて、データベースに収録された機能既知の大量なタンパク質を対象に、連続アミノ酸頻度に着目したBLSOMを試みたところ、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（トリペプチド）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。

近年、難培養性微生物類のゲノムDNA混合物を環境試料から培養過程を経ずに直接抽出し、混合ゲノム試料を対象とした配列決定法（メタゲノム解析）が普及しつつある。このメタゲノム解析由来の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国際配列データベースに登録されている。データベースに収録された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、同じ機能が推定できた機能未知タンパク質から、その推定機能を公開する予定である。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、産業界への大きな貢献が期待できる。

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