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 National Institute of Genetics and SOKENDAI
Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology

Homology searches for nucleotide and amino-acid sequences have been used widely to predict functions of genes and proteins when genomes are decoded and thus become a basic bioinformatics tool. Whereas usefulness of the sequence homology search is apparent, it has become increasingly clear that homology search can predict the protein function of only 50% of genes, or fewer, when a novel genome is decoded. As a result of decoding of extensive genome sequences from a wide variety of phylotypes, 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 estimate the protein function that does not depend on the sequence homology search is in urgent need. We previously developed a Batch-Learning SOM (BL-SOM) for genome informatics, which does not depend on the order of data input. This report focuses on BL-SOM analyses on di to tri continuous amino acid frequencies. Concerning the di- and tripeptide frequencies in the 110,000 proteins which have been classified into 2,853 function-known COGs (clusters of orthologous groups of proteins to represent individual functional categories), BL-SOMs that faithfully reproduced the COG classifications were obtained. This indicated that proteins, whose functions are presently unknown because of lack of significant homology with function-known proteins, can be related to function-known proteins with the BL-SOM.

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

1. Introduction

Self-Organizing Map (SOM) is an unsupervised neural network algorithm developed by Kohonen and his colleagues, which is an efficient and easy interpretation of clustering of high-dimensional complex data with visualization on a two-dimensional plane. Previously, we developed a modified type SOM (batch-learning SOM: BL-SOM) that depends on neither the order of data input nor the initial conditions, for oligonucleotide frequencies in genome sequences (1–5). The BL-SOM recognizes species-specific characteristics of oligonucleotide frequencies in individual genomes, permitting clustering (self-organization) of genome fragments according to species without the need for species information during the calculation. The BL-SOM 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 BL-SOM method for prediction of protein function on the basis of similarity in composition of oligopeptides (dipeptides and tripeptides in this study) of proteins. Oligopeptides are elementary components of a protein and involved in formation of functional motifs and structural organization of proteins. BL-SOM for oligopeptides may extract characteristics of oligopeptide composition which actualizes protein structure and function.

2. Methods

SOM implements nonlinear projection of multi-dimensional data onto a two-dimensional array of weight vectors, and this effectively preserves the topology of the high-dimensional data space. We modified previously the conventional SOM for genome informatics on the basis of batch-learning SOM (BL-SOM) to make the learning process and resulting map independent of the order of data input (1). The initial weight vectors were defined by PCA instead of random values as described previously (1–5). The method is fully automatic, does not require prealigned sequences, and generates a mapping from a high-dimensional input vectorial space to a two-dimensional output space.

Amino acid sequences were obtained from http://www.ncbi.nlm.nih.gov/Genbank/. 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. When the remainder of the final segment was 50 amino acids or less, the data in question was discarded, but a quantity of 200 amino acids from the remaining amino acids was added to the dataset if 50 amino acids or more remained. BL-SOM 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, SOM was constructed in this report with frequencies for the following 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.

3. Results

Sequences of approximately 2 million proteins are registered in the public databases, and about 200,000 proteins have been classified into approximately 5000 COGs (clusters of orthologous groups of proteins), which are the functional categories identified with bidirectional best-hit relationships between the completely sequenced genomes using the sequence homology search. 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. While there are many COGs whose function is unknown at the present moment, COG is undoubtedly a useful categorization of proteins according to functions. In the present study, we focused on oligopeptide compositions in the 110,000 proteins, which have been classified into 2,853 function-known COGs, and prepared BL-SOMs under various conditions to search for conditions that would most faithfully reproduce the COG classification.

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 BL-SOM 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.

Lengths of proteins with similar functions are known to differ significantly from each other and eukaryotes have many multi-functional 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 BL-SOM 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. BL-SOMs were constructed with di- and tri-peptide frequencies in these overlapped 200-amino acid sequences.

For testing the feasibility of the BL-SOM method for function prediction of proteins, one important criterion of the separation according to the functional category is at what level individual nodes on a BL-SOM contain 200-amino acid fragments derived from the proteins belonging to a single COG category. The number of function-known COGs analyzed here was 2,835, which gave 472,574 200-amino acid fragments derived from the 110,000 proteins. The size of BL-SOMs was chosen so as to provide a mean ca. 8 sequence fragments per node. The average probability that all 8 sequences were derived from a single COG category can be estimated roughly to be $(1/2853)^8$ ($= 2.3 \times 10^{-28}$) if the sequence fragments are randomly distributed on a map; the probability depends on various factors such as the proportion of the fragments derived from the respective COG in the total number of fragments derived from all 2,853 COGs. 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 BL-SOMs.

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 these BL-SOMs, and the highest occurrence of pure nodes was observed on the Tripeptide-SOM (Tri-SOM); approximately 35 and 45% of nodes of Dipeptide- and Tripeptide-SOMs contained sequences derived only from a single COG, respectively. Concerning these pure nodes, 12 examples of clustering according to COG on Tri-SOM 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. 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-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.

4. Conclusion

Protein sequences contain all the following information:

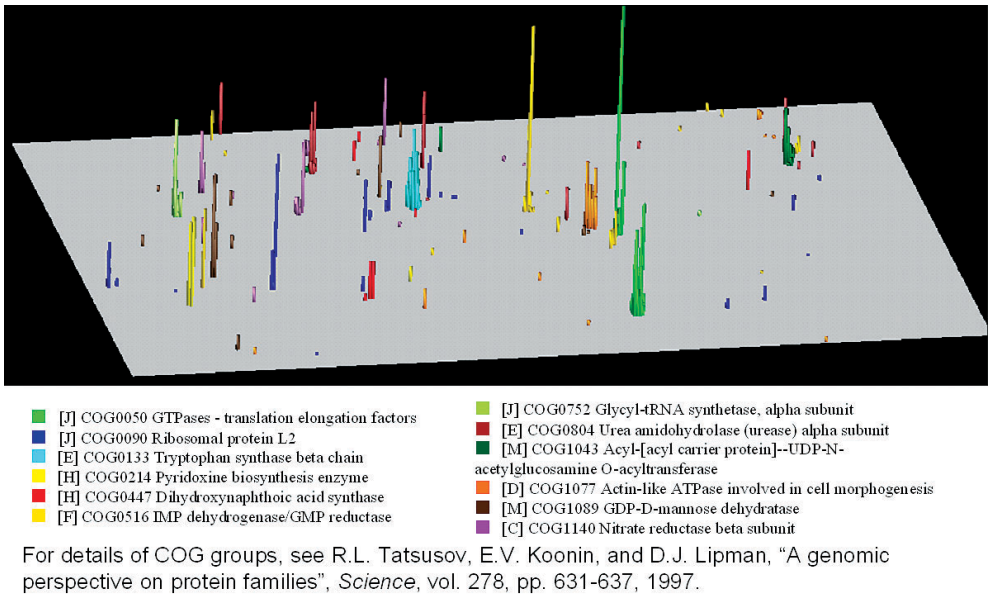

Fig. 1 Tri-SOM: Clustering of protein sequences according to COG (12 examples).

(1) evolutionary origin of the sequence; (2) functional requirements (e.g., the need to form an active reaction center); and (3) structural demands (e.g., the need to form a certain globular structure). When we consider prediction of functions especially of the proteins derived from novel genomes, such genes that originated by ancient diversifications of one gene during evolution become important to predict functions because these may hold similar functions even now. These may share common functional domains in parts even in the absence of significant global homologies detectable by the conventional homology search methods. To predict function of novel proteins, comparison of sequences with similar but not the identical function derived from a wide rage of phylotypes becomes important, and therefore, a large-scale BL-SOM has to be constructed in advance with all available proteins with known functions.

Most environmental microorganisms cannot be cultured easily under laboratory conditions, and genomes of the unculturable microorganisms have remained mostly uncharacterized but are thought to contain a wide range of novel protein genes of scientific and industrial usefulness. Metagenomic approaches, which are sequence analyses of mixed genomes of uncultured environmental microbes, have been developed recently to identify wide varieties of novel and industrially useful genes. The most important contribution of the present alignment-free method is thought to predict functions of increasingly vast amount of function-unknown proteins derived from the less characterized genomes as those studied in the metagenomic approaches. The present unsupervised, self-classifying strategy to find association of function-unknown proteins with function-known proteins on a large scale BL-SOM constructed with a high performance supercomputer "the Earth Simulator" should serve as a new and powerful tool in function prediction of vast amount of novel proteins collectively, systematically, and thus efficiently.

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.

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　　国立遺伝学研究所・生命情報DDBJ研究センター
池村　淑道　　長浜バイオ大学　バイオサイエンス学部

環境中に存在する微生物類は培養が困難であり、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のままに残されてきた。培養を行わずにゲノムDNAの混合物を抽出し、大量なゲノム断片の配列決定を行うメタゲノム解析は、環境中の生物多様性を研究する有力な方法である。しかしながら、新規性の高いゲノム由来の断片配列が解読された際に、配列相同性検索で各断片配列の由来する生物系統や配列中に特定されたタンパク質の機能が推定できる例は限定されている。異なった原理に基づく遺伝子機能や系統の推定法の確立が急務である。我々のグループは、塩基配列のオリゴヌクレオチド頻度に着目したBL-SOM法を用いて、地球シミュレータにより、データベースに収録された生物種既知のゲノム由来の大量断片配列と環境由来の大量断片配列を混合したデータセットをBL-SOM解析して、メタゲノム解析で得られた多数の配列の系統推定を可能にした。

環境微生物類のゲノム解析の主目的は、生物学的また産業的に重要なタンパク質遺伝子を発掘することにある。新規性の高い遺伝子の場合には、アミノ酸配列の相同性検索でタンパク質の機能推定は困難な例が多い。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BL-SOM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た。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したトリペプチドのBL-SOMは機能に基づく分離の傾向が高かった。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確立できる可能性が考えられる。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データを対象に、それらのダイとトリペプチド頻度の大規模BL-SOM解析を行って、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試みている。

キーワード：自己組織化マップ, BL-SOM, 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