A Novel Phylogenetic Classification Method of Genomic Sequence Fragments from Uncultured Microbe Mixtures in Environmental and Clinical Samples

Project Representative
Toshimichi Ikemura Sokendai, HCAR

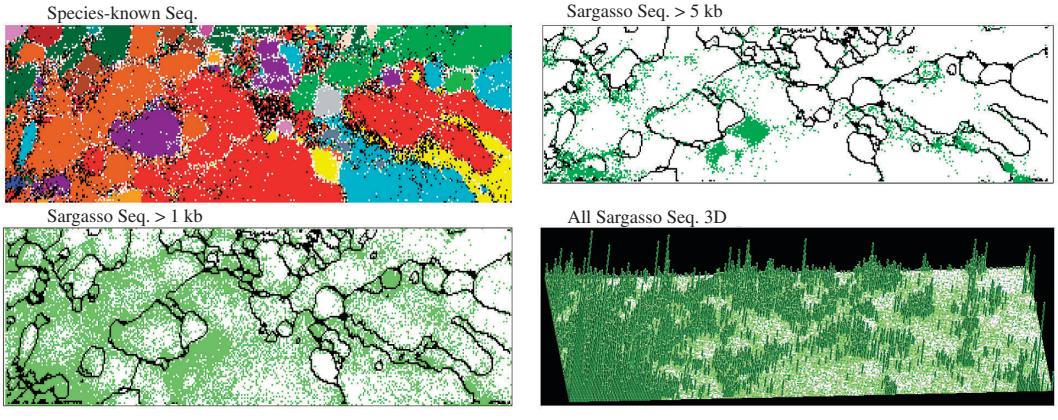
Authors
Takashi Abe National Institute of Genetics and Sokendai
Toshimichi Ikemura Sokendai, HCAR

Self-Organizing Map (SOM) is an effective tool for clustering and visualizing high-dimensional complex data on a two-dimensional map. We modified the conventional SOM to genome informatics, making the learning process and resulting map independent of the order of data input, and developed a novel bioinformatics tool for phylogenetic classification of sequence fragments obtained from pooled genome samples of microorganisms in environmental and clinical samples. This phylogenetic classification was possible without either orthologous sequence sets or sequence alignments. We first constructed SOMs for tetranucleotide frequencies in 210,000 5-kb sequence fragments obtained from 1502 prokaryotes for which at least 10 kb of genomic sequence has been deposited in public DNA databases (Species-known Seq. in Fig.). The sequences could be classified primarily according to phylogenetic groups without information regarding the species. We used the SOM to classify sequence fragments derived from the Sargasso Sea near Bermuda. The Sargasso sequences were effectively visualized on a single map.

Keywords: Self-organizing Map, SOM, oligonucleotide frequency, bioinformatics

Venter *et al.* applied shotgun sequencing to mixed genomes collected from the Sargasso Sea and deposited approximately 811,000 sequence fragments in the DNA database. The environmental sequences were analyzed with the newly developed SOM. We first mapped sequences longer than 5 kb on the SOM constructed for the 1,502 known prokaryotes (Sargasso Seq. > 5 kb in Fig.). All dominant species were detected. We then mapped 1-kb fragments derived from 134,600 entries longer than 1 kb and all 811,000 sequences (Sargasso Seq. > 1 kb and All Sargasso Seq. 3D, respectively). Although the sequences were not clustered tightly, skewed distributions were observed. This is more easily visualized with a 3-dimensional (3D) representation in which the number of Sargasso sequences classified into each lattice point is indicated by the height of a bar.

G+C% has been used for a long period as a fundamental parameter for phylogenetic classification of microbes, but the G+C% is apparently too simple a parameter to differentiate a wide variety of known species. SOM used oligonucleotide frequency effectively to distinguish the species because oligonucleotide frequencies vary significantly among their genomes. In the case of the homology-based



For details, see Abe *et al.* (*DNA Research*, 12, 281–290, 2005)

phylogenetic classification, a set of orthologous sequences from a wide range of species (e.g. rDNA) is an absolute requirement, and therefore, this conventional method is difficult to apply to the classification of sequences for novel genes from poorly characterized species. Because the SOM does not require the orthologs, the present strategy can enhance novel metagenomic studies of uncultured microbes.

Bibliographies

1) 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.

2) 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.

環境中ならびに臨床検査サンプル中の微生物混合ゲノムに由来するゲノム断片配列についての新規な系統分類法

プロジェクト責任者
池村　淑道　　総合研究大学院大学・葉山高等研究センター
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　　総合研究大学院大学・葉山高等研究センター

SOMは生物種に関する予備知識なしに、断片配列の大半を生物系統に分類（自己組織化）が可能である。この優れた特徴を、環境由来のゲノム配列の新規な情報的解析に適用した。環境に生育する微生物の大半が実験室での培養が困難であり、未開拓なゲノム資源として残されてきた。環境微生物類については、培養やクローン化を行わずに、混合状態のままDNAを抽出し、DNA断片としてクローン化し配列決定をする研究が開始されている。米国のVenterらは、バーミューダ沖の海水中の微生物集団から混合状態でゲノムDNAを回収し、80万本の断片塩基配列を決定し、約120万の遺伝子の候補を推定している。これらの環境由来の大量ゲノム配列の全体を対象に、生物系統の推定や多様性の実体を知ることを目的とした大規模SOM解析を行った。従来の系統推定法と異なり、オルソログ配列のセットや配列間のアラインメントが必要でなく、連続塩基の出現頻度のみで系統推定が可能であり、新規性の高い未知の配列類には最適の解析手法である。

キーワード：自己組織化マップ, SOM, 環境微生物, オリゴヌクレオチド頻度