氏　　名	阿 部 貴 志
学位（専攻分野）	博士(理学)
学 位 記 番 号	総研大乙第127号
学位授与の日付	平成１６年３月２４日
学位授与の要件	学位規則第４条第２項該当
学 位 論 文 題 目	Development of novel genome informatics strategy on the basis of Self-Organizing Map(SOM)
論 文 審 査 委 員	主　査　教授　菅原　秀明
	教授　五條堀　孝
	教授　城石　俊彦
	教授　服部　正平（北里大学）
	助教授　金谷　重彦（奈良先端科学技術大学院大学）

With the increasing amount of available genomic sequences, novel tools are needed for comprehensive analysis of species-specific sequence characteristics for a wide variety of genomes. Self-Organizing Map (SOM), which was developed by Kohonen to study memory and recall/association mechanisms, can identify and associate similar types of information and localize such information in close vicinity on a two-dimensional map. SOM has been proven to be a powerful unsupervised algorithm and applied in various fields of science and technology (e.g., complex industrial processes, document and image databases, and financial applications) but rarely been applied to analysis of genome sequences. In this thesis study, on the basis of batch-learning SOM (BL-SOM), I modified the conventional SOM for genome informatics to make the learning process and resulting map independent on the order of data input. The initial weight vectors of the Kohonen's conventional SOM were usually set by random values, but the vectors in my method were initialized by principal component analysis (PCA) to obtain the same result between different caluculations. I furthere modified BL-SOM to execute parallel processing with supercomputers and PC-clusters and thus could analyze a vast amount of available genomic sequences.In this thesis study, I used the modified SOM to analyze short oligonucleotide frequencies (di- to pentanucleotide frequency) in a wide variety of prokaryotic and eukaryotic genomes.

When only fragments of genomic sequences (e.g., 10-kb sequences) from mixed genomes of multiple organisms are available, it would appear to be impossible to identify how many and what types of genomes are present in the collected sequences. However, I found that the modified SOM could classify the sequence fragments according to species without any information other than oligonucleotide frequencies. I constructed SOMs of di-, tri-, and tetranucleotide frequencies in 1- and 10-kb sequences from prokaryotic and eukaryotic genomes for which complete sequences are available. SOM recognized, in most 10-kb sequences, species-specific characteristics of oligonucleotide frequencies (key combinations of oligonucleotide frequencies), permitting species-specific classification of sequences without any information regarding species.

A majority of environmental microorganisms, especially those living in extreme environments, are difficult to culture in the laboratory. Because conventional experimental approaches have been unsuccessful, these genomes have remained uncharacterized, and there is the possibility that such genomes contain a wide range of novel genes that would be of scientific and/or industrial interest. Metagenomics, which is genomic analysis of uncultured microorganisms, has been proposed to study microorganism diversity in a wide variety of environments and

to identify novel and industrially useful genes. In the metagenome analysis of uncultured microorganisms, genome DNAs are extracted directly from an environmental specimen that contains multiple organisms, and the genomic fragments are then cloned and sequenced. With a simple collection of fragmental sequences, it appears to be impossible to predict what kinds and the ratios of species present in an environmental sample, to which lineages the species belong, and how the genomes are novel. To establish SOM as a methodology suitable to this purpose, I constructed SOMs of tetranucleotide frequencies in 1- and 5-kb sequences from approximately 80 bacterial genomes for which complete sequences are available. Sequences were clustered primarily according to species and to 11 major bacterial groups without any information regarding the species. With this SOM method, all sequences in DNA databases that were from unidentified or uncultured bacteria and longer than 1 kb were classified into 11 major bacterial groups. The result indicated that the method is useful also for survey of pathogenic microorganisms causing novel, unclear infectious diseases.

Next, I analyzed tetra- and pentanucleotide frequencies in the human genome, and found that frequencies and distributions of oligonucleotide sequences involved in transcriptional regulation were often biased significantly from random occurrence. I could categorize occurrence patterns and frequencies of known signal sequences in the human genome. When known signal sequences from various species with sufficient experimental data are characterized and categorized systematically with SOMs, it should be possible to develop an *in silico* method to predict signal sequences, which is thought to be most useful for identification of signal sequences in genomes for which only sequence data are available. Because the number of such poorly characterized genomes becomes high, development of such an *in silico* method has become increasingly important. I have developed SOM as a methodology just suitable to this purpose.

In addition to protein-coding sequences (CDSs), the flanking regions upstream of transcription start sites and the 5' and 3' untranslated regions (UTRs) have attracted attention because of their crucial roles in transcriptional and post-transcriptional regulation of gene expression. By combining analyses on cDNA and genomic sequences of human and mouse, I developed SOM to characterize the six functional regions, 5' and 3' UTRs, CDSs, introns, 5' flanking regions, and ncRNAs, in these genomes and to identify hidden sequence characteristics in the functional regions. Because clustering power of SOM is very high, I propose that SOM can provide fundamental guidelines for understanding molecular processes and mechanisms that have established sequence characteristics of individual genomes and genomic regions during evolution.

論文の審査結果の要旨

阿部貴志君が提出した審査論文「Development of a novel genome informatics strategy on the basis of Self-Organizing Map (SOM) は、ゲノム配列データにおける生物種や機能に特徴的な偏りを抽出することができる新規なデータマイニングの手法を論じている。この手法は、教師なしニューラルネットワークアルゴリズムを適用した結果を 2 次元空間に展開する自己組織化地図法 (Self-Organizing Map; SOM) を、学位申請者がゲノム解析用に適した手法へ改良したものである。記憶の組織化をモデルとして提案された従来の SOM はゲノム配列データのマイニングに適していなかったが、第1に、SOMの本質的な問題であった入力データの入力順に依存しない結果を得られるように改良し、第2に主成分分析の前処理をして、入力データの特徴を初期値に反映するように改良することによって、ゲノム配列データのマイニングに適用可能とした。また、大量データへの対応を意識して多数の CPU を利用した並列計算が可能な計算手法を構築した。

配列が既知の微生物と真核ゲノム配列の全体を対象に、10 kb と 100 kb 断片配列の 2 ～ 5 連続塩基頻度を入力データとする計算機実験を行い、新手法が、配列のアノテーションを一切使用することなく大半の断片配列が生物種ごとにクラスターを形成することを示した。すなわち、生物種のゲノム配列に潜む生物種固有のサイン (genome signature) を機械的に検出することを可能とした。また、配列が既知な全バクテリアを対象にした解析では、各クラスターに含まれる断片配列の視覚的比較から、ゲノム断片の水平移動の様式を把握できることを示した。さらに、バクテリアゲノムの解析結果から得たSOM地図をもとに、培養が困難な環境微生物類の混合試料に由来する塩基配列の系統推定の有効性を示した。したがって、環境微生物の多様性とその進化の研究、ならびに新規ゲノムを探索に有用な新たなデータマイニングの可能性を示した。

バクテリアに加えて、ヒトとマウス等の近縁な生物種間でも明瞭な差違を検出しており、ヒトゲノム配列の解析では、転写制御シグナル等の既知シグナル類は SOM 上で特徴的な出現パターンを示す傾向を見出して、新手法が、機能的意味をもつシグナル類の網羅的な探索法として有望なことを示した。

当該研究について申請者はすでに 5 編の原著論文を発表し、そのうち 3 編が申請者を筆頭著者とする国際誌における発表である。なお、生物分野のデータマイニングに SOM を応用した論文は、2000 年までは年に 1～5編であったが 2001 年に 1515 編、2002 年、2003 年とそれぞれ 25 編を超えている。

以上のことから、申請者の研究は、ゲノムデータおよび配列データのマイニングに新しい手法をもたらすとともに、SOM による生物データマイニングの先駆的研究と位置づけられる。

このように本申請者の研究は、特に優れた業績に基づいていることから、審査委員会は、申請論文を、論文博士の学位を授与するに足るものと評価した。

なお、申請者は、協和発酵工業株式会社（株式会社ザナジェンに出向）に所属し、2001 年 5 月から現在までの約 2 年半にわたり、国立遺伝学研究所進化遺伝部門の受託研究員として、池村淑道教授のもとで、「大量なゲノム配列からの効率的な知識発見を行うための手法の開発ならびに、ゲノム配列に潜む生物種の個性の解明」についての研究に従事してきた。