氏名	中村保一		
学位（専攻分野）	博士(理学)		
学位記番号	総研大乙第87号		
学位授与の日付	平成１３年3月２３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	Date analysis and presentation of large-scale nucleotide sequence information		
論文審査委員	主査	教授	五條堀　孝
		教授	城石　俊彦
		助教授	藤山　秋佐夫
		教授	久原　哲（九州大学）
		助教授	角谷　徹仁（国立遺伝学研究所）

論文内容の要旨

Summary

Rapid, automated sequencing technologies with related advances in computational analysis and informatics have transformed the nature of biological research. The huge amounts of sequence data challenge the scientific community to understand and use this new information effectively. In this thesis, I describe the construction of data analysis and presentation systems for sequence information; these systems will assist in the identification of gene and implementation of a high-throughput genome sequencing era.

Chapter 1

CUTG (codon usage tabulated from GenBank) is a comprehensive database of codon usage. To generate an electronic data set for codon usage for each gene and for codon choice trends in each genome, Ikemura *et al.* have compiled codon usage in genes encoding proteins contained within the international DNA sequence database. The data files are available on ftp sites at Kazusa DNA Research Institute, National Institute of Genetics and European Bioinformatics Institute.

The compilation is synchronized with major releases of GenBank. The latest data source available during the preparation of this thesis was NCBI-GenBank Flat File Release 120.0. The frequencies of each of the 382,241 complete protein-coding sequences (CDSs) was compiled from the taxonomic divisions of the DNA sequence database. The sum of the codons used by 11,388 organisms has also been calculated. A list of the codon usage of genes and the sum of the codons used by each organism can be viewed at http://www.kazusa.or.jp/codon/. A new WWW interface has been developed to provide data in a format compatible with that of the CodonFrequency output in the GCG Wisconsin Package^{TM}. Also, for each species, there is a query box to search for information in the comments for each gene. The user can choose CDSs by keyword and then generate codon usage tables from the selected genes. This tool provides researchers with the ability to examine intra-species variations in codon usage.

As an application of codon usage-based analysis for microbial genome sequencing efforts, I used only the sequences of the ribosomal protein genes as standards for calculation when I performed a modified codon adaptation index (CAI) analysis. This is in contrast to the traditional method of analysis, which relies on prior knowledge of the sequences of the most highly expressed genes.

To begin, I tabulated the patterns of codon-anticodon recognition in the following microorganisms whose genomes have been sequenced completely: *Haemophilus influenzae* Rd, *Methanococcus jannaschii,* and *Synechocystis* sp. strain

PCC6803. For *Escherichia coli*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, and *Saccharomyces cerevisiae*, the previously adopted codon-anticodon combination was used.

I then used a modified CAI (Sharp and Li, 1987) as a measure of synonymous codon bias. The original CAI value for each gene was measured with the codon preferences of the genes for highly expressed proteins such as ribosomal proteins and elongation factors, as a basis. To generalize this method to organisms for which only sequence information exists, I modified the procedure of extraction by simply taking into account the sequences of the ribosomal protein-coding genes, and the codon usage biases of the ribosomal protein genes of each of the seven microbial genomes was recalculated. With these values, $\mathrm{CAI}_{\mathrm{rp}}$, a CAI that depended on the codon biases of the ribosomal protein genes, was calculated for all of the protein-coding genes of the genome.

Of the seven genomes examined, a clear correlation between the $\mathrm{CAI}_{\mathrm{rp}}$ score and the level of protein-coding gene expression was observed for all but the genes of *M. genitalium*. For the six genomes, elongation factors, and chaperonins, and ribosomal proteins had high CAI scores In contrast, genes for transposases and genes of prophage origin, which are expressed at lower levels, had the lowest CAI scores. This result indicates that codon usage analysis based on ribosomal protein gene sequences may be useful for predicting the expression levels of unknown genes. This method would be particularly useful for microbes where the entire genomes is being sequenced, since the DNA sequences of most, if not all, genes would be available.

Chapter 2

A WWW database system that provides information for deduced protein-coding genes was constructed for the cyanobacteria sequencing project.

Cyanobacteria are prokaryotic microorganisms that carry a complete set of genes for oxygenic photosynthesis. In 1996, Kaneko *et al.* reported the complete 3.57 megabase (Mb) sequence of the genome of *Synechocystis* sp. strain PCC6803, which contains 3,168 potential protein-coding genes.

CyanoBase (http://www.kazusa.or.jp/cyano/) is an online resource for accessing genomic data for the cyanobacterium. The core portion of CyanoBase contains annotations for each of the 3,168 protein genes deduced from the entire nucleotide sequence of the *Synechocystis* sp. strain PCC6803 genome. The annotation for each protein-coding gene is accessible through three menus on the main page of this database: map image, gene classification lists, and keyword and similarity search engines. The aim of this database is to provide detailed information on potential protein-coding genes through a user-friendly interface that includes clickable genome maps and a hypertext classification list.

The database also contains repository facilities that store and offer experimental information and proposed function of each gene. Of the 3,168 deduced genes on the *Synechocystis* genome, 1,722 are annotated as functionally unassigned, which included 1,270 putative genes, 418 genes similar to hypothetical ones, and 34 genes similar to expressed sequence tags (ESTs) of other genomes. To analyze the functions of these genes, systematic disruption of each gene and characterization of the resulting mutants is thought to be a promising strategy.

CyanoMutants (http://www.kazusa.or.jp/cyano/mutants/) is a cumulative database that allows users to stores and access mutant information through the WWW. Each entry in CyanoMutants contains three sections: identification of the mutated gene, information about the phenotype, and person to whom correspondence should be addressed. Each entry is linked to the corresponding annotation in CyanoBase. The corresponding page in CyanoBase contains a link to the page in CyanoMutants that provides mutant information. These linked information will prevent unnecessary overlaps in experiments and promote communication among scientists to elucidate the functions of putative genes in cyanobacteria.

As of December 2000, CyanoMutants contained 431 mutant entries, 134 of which have phenotype description. The number of genes registered is expected to increase continuously since a large number of gene disruption experiments have been carried out since the release of the genomic sequence of *Synechocystis* sp. strain PCC6803.

Chapter 3

A protocol to automate the execution of similarity searches and gene prediction programs was developed for the *Arabidopsis thaliana* genome sequencing project. High-throughput annotation of 27 Mb genomic sequences of *A. thaliana* has been carried out with the assistance of the system.

The 125 Mb genome of *A. thaliana* is organized into five chromosomes and contains an estimated 25,500 genes. To understand the entire genetic system in this plant, an international sequencing project of the *A. thaliana* genome has been initiated 1996, and currently it is in completion phase. Our research group is participating in sequencing the entire bottom arm and portions of the top arm of chromosome 5 and also the top arm of chromosome 3. During the process of annotating the genomic sequences of clones on the chromosomes, I have constructed a computer-aided system for high-throughput gene identification.

In this system, nucleotide sequences are translated in six frames with use of the universal codon table, and each frame is subjected to a similarity search against the non-redundant protein database, nr, with use of the BLAST program. Each local alignment, that shows an E-value < 0.001 to known protein sequences, is extracted and stored. Potential exons for protein-coding genes are predicted with the computer programs Grail and GENSCAN. For localization of exon-intron boundaries, donor/acceptor sites for splicing are predicted by NetGene2 and SplicePredictor. To identify transcribed regions and structural RNA genes, the BLAST program is used to compare nucleotide sequences with the EST and RNA gene data sets. For assignment of tRNA genes and tRNA structures, tRNA-scanSE is used.

All outputs are then parsed and stored in the General Feature Format (GFF). When required, the results are parsed and loaded into a WWW-based information display system called *Arabidopsis* Genome Displayer. This display system shows the positional relation of genome features along a genomic sequence. Simultaneously, an annotation composing interface allows manual editing of the gene model showing tentative nucleotide and protein sequences and exon-intron organization. The annotator performs similarity searches as needed on the working model during the gene-modeling process. After careful editing, the most reasonable model of a genomic region is saved in the in-house database as a deduced gene.

In conclusion, 6,124 potential protein-coding genes were assigned to the 27 Mb regions of *Arabidopsis* chromosomes 3 and 5 covered by 461 clones and gap-closing units. The average density of genes was estimated to be 1 gene per 4.4 kb. One hundred twenty-seven RNA genes were deduced by similarity searches and computer predictions. Of 6,124 deduced protein-coding genes, 2,808 carried EST sequences, indicating that 46% of the total genes in *A. thaliana* may be represented in the current EST databases.

論文の審査結果の要旨

各種生物のゲノムプロジェクトが確実に進行する中、大量のDNA配列データが産出され、それに伴ってバイオインフォマティクスと呼ばれる研究領域が極めて重要な位置づけで発展してきている。特にゲノムデータのもつ大量性と複雑性という大きな課題を克服しながら、ゲノムデータから生物学的に有用な情報を抽出していく研究や、そのような研究に役立つデータベースの構築および情報システムの確立に関する研究は、バイオインフォマティクスの更なる発展において、とりわけ重要と考えられる。

中村氏は、博士論文において、大きく分けて次の３つの研究を行った。

１つは、「CUTG (Codon Usage Tabulated from GenBank)」と名付けられたコドン使用頻度の体系的データベースの構築と、それを用いた同義コドン使用頻度のバイアスに関する研究である。まず、国際DNAデータベースから約1,400の生物種にわたる約400,000個に及ぶ完全タンパク質翻訳領域を抽出し、コドン使用頻度を計算して、データベースを構築した。
また、種内や種間においてコドン使用頻度の差異が直ちに検索できるシステムを構築した。このデータベースとその検索システムを自ら用いて、コドン使用頻度データから遺伝子発現の相対的な量を定性的に予測する研究を行った。特に、従来、同義コドン使用頻度のバイアスの指標として使われていたCAIを独自のアイデアで改変し、リボゾームタンパク質遺伝子のDNA配列データを基準とする新たな指標を考案した。この新たな指標を用いれば、ゲノムのDNA配列データしかわかっていなくとも、各種の遺伝子の相対的な発現量を定性的に予測できることを、多数のバクテリアゲノムデータに応用して示すことに成功した。

２つ目は、らん藻のゲノムプロジェクトに参加し、「CyanoBase」というゲノムデータベースを構築した。らん藻は光合成に関与するすべての遺伝子を保有する原核生物として知られ、3.57Mbに及ぶ完全ゲノムのDNA配列が解読された。特に、3,168個のタンパク翻訳領域に注目し、これらのそれぞれについて様々な情報を付加するアノテーションを行って、付加価値の高いゲノムデータベースの構築に成功した。また、「CyanoMutants」というデータベースも構築し、突然変異やその表現型に対する影響等に関する情報を集積させた。これらのデータベースは、らん藻遺伝子の機能解析に重要な役割を果たすと考えられる。

３つ目は、アラビドプシスという植物の125Mbからなるゲノムプロジェクトに、バイオインフォマティクスの立場で参加し、ゲノム中に存在する遺伝子の予測とそれに対するアノテーションを体系的に行った。遺伝子の予測は、現在でも難しい研究といわれているが、中村氏はデータ処理の流れや既存ソフトウェアの組み合わせや改良を考えるなどして独自の遺伝子予測システムを作り上げた。特に、３番と５番の染色体の約27Mbにわたる領域において、6,124個の遺伝子候補領域の同定に成功した。

以上の研究を含めて、28編の原著論文をNatureなどの国際誌に発表しており、この分野への貢献が多大と評価し、審査委員一同は、博士論文が本学の基準を充分に満たすと判断した。