A Statistical Analysis of Gene Arrangement Patterns in Bacterial and Yeast Genomes

Kentaro Tomii
tomii@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611, Japan

Abstract

The complete sequences of prokaryotic genomes and a eukaryotic genome have been reported. The time has come to analyze entire genome features among different organisms. We show here the results of a statistical analysis of the correlation between gene arrangements in the genome and biological functitons of gene products.

1 Introduction

We can often see a set of proteins, for instance, proteins in an operon, that are encoded closely in the genome sequence and have related biological functions, for example, in the metabolic pathway. Is there any rules for the arrangement of related genes in the genome sequence? If so, are the rules conserved over different organisms? Recently the sequences of complete genomes of several bacteria and yeast have been determined[1][2][3]. By using these data, we analyze whether there is any characteristic arrangement of ORFs in the genome sequence.

2 Data and Method

We use the six organisms, *H.influenzae*, *M.genitalium*, *Synechocystis*, *E.coli*, *B.subtilis*, and *S.cerevisiae*, and have collected the genome sequence data. The data of the nucleotide positions and functional assignments are taken from KEGG (Kyoto Encyclopedia of Genes and Genomes) which also contains the data sets of known metabolic pathways and enzyme gene catalogs of several organisms. Our analysis is based on observing the frequencies of nearest neighbor pairs of ORFs categorized by functional properties, such as by Riley's classification[4]. For example, *Synechocystis* sp. PCC6803 gene products are classified into 14 functional categories[3]. We consider all possible (14×14) pairs of combinations of categorized ORFs and estimate, not rigorously, the expected frequency of observing each pair in random arrangement in the genome sequence. We also check the statistical significance of real frequencies observed for these pairs.

3 Results

There is a tendency that proteins implicated in photosynthesis and respiration are encoded nearby in the genome of *Synechocystis* sp. PCC6803. Most of these proteins are encoded as a set of subunits, for example, ATP synthetase, cytochrome *c* oxidase, and NADH dehydrogenase. Ribosomal proteins are also encoded as a set. We compare ORF arrangements in the genomes of different organisms in view of the function and the relative position in the genome sequence. We also discuss the relationships with operons and transcription mechanisms.

Acknowledgment

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] R.D. Fleischmann, *et al.* “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd,” *Science*, Vol. 269, pp. 496-512, 1995.

[2] C.M. Fraser, *et al.* “The minimal gene complement of *Mycoplasma genitalium*,” *Science*, Vol. 270, pp. 397-403, 1995.

[3] T. Kaneko, *et al. DNA Res.*, Vol. 3, pp. 109-136, 1996.

[4] M. Riley, “Functions of the gene products of *Escherichia coli*,” *Microbiol. Rev.*, Vol. 57, pp. 862-952, 1993.