

# Gene-Transpositions on Functional Categories in Prokaryotic Genomes

Nobuyoshi Sugaya<sup>1</sup>  
sugaya@ims.u-tokyo.ac.jp

Hiroo Murakami<sup>1</sup>  
hiroo@ims.u-tokyo.ac.jp

Sachiyo Aburatani<sup>1</sup>  
sachiyo@ims.u-tokyo.ac.jp

Kunio Shimizu<sup>2</sup>  
shimizu@math.keio.ac.jp

Katsuhisa Horimoto<sup>1</sup>  
khorimot@ims.u-tokyo.ac.jp

<sup>1</sup> Laboratory of Biostatistics, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan.

<sup>2</sup> Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan.

**Keywords:** number of genes, gene content, gene location, gene function, gene transposition

## 1 Introduction

Recent accumulation of completely sequenced prokaryotic genomes has prompted the comparisons of functional gene content between the genomes with a wide range of size variation. Some studies [4, 5] provide a new insight with the relation between gene content and genome sizes: the number of genes in protein biosynthesis is almost equally distributed among various sizes of genomes, but that the gene numbers in metabolisms, gene regulation and metabolite transport change in accordance with the genome sizes.

In consideration with the above relation, we estimate the transposition degrees of genes categorized by their functions, by taking into account not only the numbers of genes but also the gene locations. In this study, we focus on the variation of gene number and gene-transposition degrees estimated by the two measures in the functional categories, in comparison with the prokaryotic genomes in different genome sizes.

## 2 Method and Results

### 2.1 Genomic Data

We analyzed 10 genomes from the organisms in distinct genome sizes: *Escherichia coli* K-12 MG1655, *Haemophilus influenzae* Rd, *Neisseria meningitidis* MC58, *Caulobacter crescentus*, *Bacillus subtilis* 168, *Streptococcus pneumoniae* R6, *Mycobacterium tuberculosis* H37Rv, *Synechocystis* sp. PCC6803, *Methanococcus jannaschii* DSM2661 and *Sulfolobus solfataricus* P2. The sizes of the above genomes are ranged from 1.7Mbp to 4.6Mbp.

All genes in the genomes were classified into 14 categories, according to their gene function annotated in the TIGR-CMR database [7]. As a preprocessing of gene location comparison, ortholog pairs between the compared genomes are defined as the best-hit pairs with sequence similarity ( $P < 0.01$ ) by the BLASTP program, and the pairs were also classified into 14 categories according to TIGR-CMR database [7].

### 2.2 Two Measures for Gene Location

We estimate the gene transposition degree by two statistical measures: one is a measure of similarity ( $S^*$ ) based on the Von Mises distributions [6], and another is the gene-location distance (GLD) [2, 3] based on the correlation coefficient for circular data [1]. The former measures the similarity of the distribution form of all gene locations on two circular genomes, irrespective of orthologous relationship between genes, and the latter measures the similarity of the configuration of orthologous genes.

### 2.3 Gene Number and Gene-Transposition Degree on Functional Categories

The gene number and the two gene-transposition degrees between the 10 genomes were compared in 14 functional categories (Table 1). Since we focus on the variation of the above three values in this study, we listed the coefficient of variation (CVs) that was defined as the standard deviation divided by the average in the compared genomes. Thus, a large value of CV in a functional category indicates that the corresponding measure shows variety in 10 compared genomes, and a small CV indicates that the measure shows uniformity in the genomes.

The gene numbers were partly consistent with the trend in previous studies [4, 5]. In some metabolism-related categories, CVs are large (Category No. 5 and 8), being consistent with the previous trend, but CVs are small (1, 2 and 11).

The similarity of distribution form shows uniformity in the categories, and is not correlated with the CVs of gene content. Note that the large CVs were found in 1, 10, 11, and 13, in which many operon structures are known in a few genomes.

The similarity of gene configuration shows a similar trend of that of gene content. Although the correspondence between the gene content and GLD in each genome is further required, the change of gene configuration is closely related with that of gene content, as a first approximation.

Table 1: Comparison of gene number and the two gene-transposition degrees between the 10 genomes on 14 functional categories

No.	Function	No. of genes	S*	GLD
		CV	CV	CV
1.	Amino acid biosynthesis	0.25	0.04	0.14
2.	Biosynthesis of cofactors, prosthetic groups, and carriers	0.36	0.01	0.19
3.	Cell envelope	0.58	0.01	0.32
4.	Cellular processes	0.75	0.01	0.41
5.	Central intermediary metabolism	0.72	0.01	0.57
6.	DNA metabolism	0.43	0.01	0.19
7.	Energy metabolism	0.48	0.01	0.16
8.	Fatty acid and phospholipid metabolism	0.60	0.01	0.59
9.	Protein fate	0.54	0.01	0.24
10.	Protein synthesis	0.16	0.04	0.18
11.	Purines, pyrimidines, nucleosides, and nucleotides	0.30	0.04	0.21
12.	Regulatory functions	0.73	0.01	0.33
13.	Transcription	0.62	0.06	0.56
14.	Transport and binding proteins	0.55	0.00	0.18

### 3 Discussion

The present study focused on the variation of the three measures in 10 genomes with distinct sizes. The two measures for gene locations reveal distinct features of gene-transposition degrees. We will further discuss the comparison of the three measures from the evolutionary relationship between the compared genomes.

### References

- [1] Fisher, N. I. and Lee, A. J. A correlation coefficient for circular data. *Biometrika*, 70:327-332, 1983.
- [2] Horimoto, K., Suyama, M., Toh, H., Mori, K., and Otsuka, J., A method for comparing circular genomes from gene locations: application to mitochondrial genomes, *Bioinformatics*, 14:789-802, 1998.
- [3] Horimoto, K., Fukuchi S., and Mori, K., Comprehensive comparison between locations of orthologous genes on archaeal and bacterial genomes, *Bioinformatics*, 17:791-802, 2001.
- [4] Nimwegen, E.V., Scaling laws in the functional content of genomes, *Trends Genet.*, 19:479-484, 2003.
- [5] Ranea, J.A.G., Buchan, D.W.A., Thornton, J.M., and Orengo, C.A., Evolution of protein superfamilies and bacterial genome size, *J. Mol. Biol.*, 336:871-887, 2004.
- [6] Shimizu, K. and Iida, K., Pearson type VII distributions on spheres, *Commun. Statist.-Theory Meth.*, 31:513-526, 2002.
- [7] <http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl>