

Heterogeneity in Synonymous Codon Usage among Genes of Diverse Bacterial Genomes

Haruo Suzuki^{1,2}

haruo@sfc.keio.ac.jp

Rintaro Saito^{1,3}

rsaito@sfc.keio.ac.jp

Masaru Tomita^{1,3}

mt@sfc.keio.ac.jp

¹ Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan

² Graduate School of Media and Governance, Bioinformatics Program, Keio University

³ Department of Environmental Information, Keio University

Keywords: codon usage heterogeneity, distance metric, genomic base composition, mutation-selection balance

1 Introduction

It has become clear that in most bacterial species there is considerable heterogeneity in codon usage among genes, and the codon usage of any gene is thought to reflect a balance between the forces of mutation, selection, and random genetic drift [4]. The strength and direction of these forces vary for different species, depending on their lifestyles [1]. In this study, we investigated the level of synonymous codon usage heterogeneity among genes of diverse bacterial species with different genome compositions.

2 Sequences and Methods

The protein-coding sequences (CDS) of complete genomes were retrieved from the NCBI FTP server [5]. The mean distance (D_{mean}) between all pairs of CDS, based on the differences in relative synonymous codon usage (RSCU) values, is calculated to quantify the level of heterogeneity in synonymous codon usage. We used two distance metrics: absolute distance given in the following equation [2] and $1 - r$ (where r is the correlation coefficient between RSCU values of two CDS).

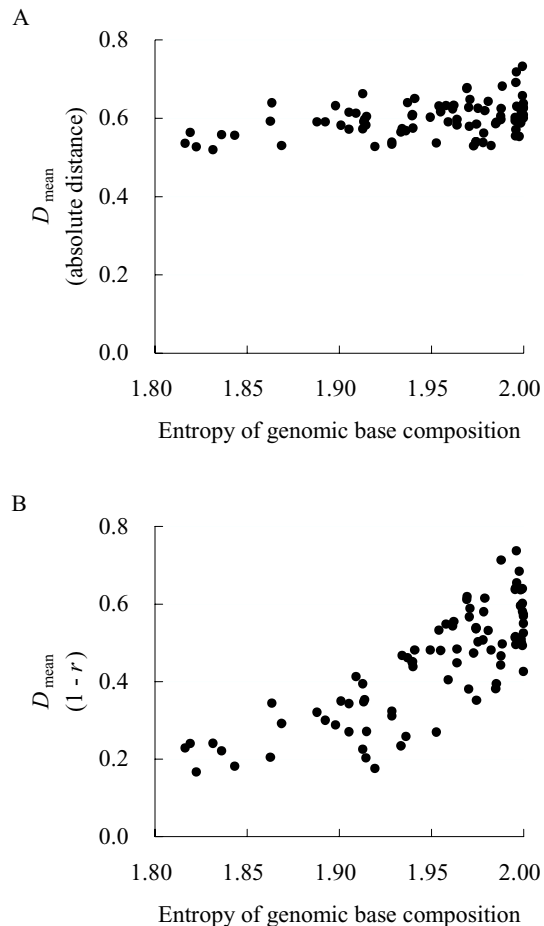
$$(\text{absolute distance})_{ij} = \frac{\sum_{c=1}^k |RSCU_{ic} - RSCU_{jc}|}{k}$$

where k is the number of synonymously degenerate codons for that amino acid, ic represents codon c on CDS i , and jc represents codon c on CDS j .

3 Results and Discussion

The heterogeneity analysis based on correlation coefficients shows that there is a strong correlation between the level of synonymous codon usage heterogeneity among all CDS and the degree of genomic base composition bias ($r = 0.807$), while in the case of absolute distances the correlation is weaker ($r = 0.373$). Because CDS pairs whose codon usage patterns have the same shape but different magnitudes will yield large absolute distances, the use of absolute distances may mask the effects that are directly linked to heterogeneity in direction of codon usage bias. Our result suggests that the heterogeneity analysis based on correlation coefficients instead of absolute distances is effective to explore shifts in the mutation-selection balance across bacterial species with different lifestyles.

Figure 1: Mean distance (D_{mean}) between all pairs of CDS as a heterogeneity index. Distances were measured by using (A) absolute distance and (B) $1-r$, respectively. The GenBank accession numbers of the 87 bacterial genomes examined are as follows: NC_000117, NC_000853, NC_000854, NC_000907, NC_000908, NC_000909, NC_000911, NC_000912, NC_000913, NC_000915, NC_000916, NC_000917, NC_000918, NC_000919, NC_000922, NC_000961, NC_000962, NC_000963, NC_000964, NC_001263, NC_001318, NC_002162, NC_002163, NC_002179, NC_002488, NC_002491, NC_002505, NC_002516, NC_002528, NC_002578, NC_002607, NC_002620, NC_002655, NC_002662, NC_002677, NC_002689, NC_002695, NC_002696, NC_002737, NC_002745, NC_002754, NC_002755, NC_002758, NC_002771, NC_002932, NC_003028, NC_003030, NC_003047, NC_003098, NC_003103, NC_003106, NC_003112, NC_003143, NC_003197, NC_003198, NC_003210, NC_003212, NC_003272, NC_003295, NC_003304, NC_003317, NC_003364, NC_003366, NC_003413, NC_003450, NC_003454, NC_003551, NC_003869, NC_003901, NC_003902, NC_003919, NC_004061, NC_004070, NC_004113, NC_004116, NC_004193, NC_004307, NC_004310, NC_004337, NC_004342, NC_004347, NC_004350, NC_004368, NC_004431, NC_004432, NC_004459, NC_004461.



References

- [1] Ermolaeva, M.D., Synonymous codon usage in bacteria, *Curr. Issues Mol. Biol.*, 3(4):91–97, 2001.
- [2] McInerney, J.O., GCUA: general codon usage analysis, *Bioinformatics*, 14(4):372–373, 1998.
- [3] Ovreas, L. and Torsvik, V., Microbial diversity and community structure in two different agricultural soil communities, *Microbial Ecology*, 36(3):303–315, 1998.
- [4] Sharp, P.M., Stenico, M., Peden, J.F., and Lloyd, A.T., Codon usage: mutational bias, translational selection, or both?, *Biochem. Soc. Trans.*, 21:835–841, 1993.
- [5] <ftp://ftp.ncbi.nih.gov/>