Codon usage diversity in city microbiomes

Haruo Suzuki[1,2]

1. *Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan*
2. *Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, Japan*

E-mail: haruo@sfc.keio.ac.jp

Introduction

It has been reported that almost half of metagenomic read data from the New York City subway systems did not match any known organism[1]. To shed some light on these unknown sequences, we need alternative approaches that do not rely on taxonomic or functional assignment. Synonymous codon usage varies between organisms and among genes within a genome, and reflects various factors, including mutational biases shaping G+C content, compositional skew between leading and lagging strands of replication, rRNA and tRNA gene numbers, translational efficiency and accuracy, growth rate, and life style[2]. Previous studies compared codon usage of highly expressed genes (i.e. those annotated as 'ribosomal proteins') and all genes in metagenomes to predict gene expression levels[3] and maximal growth rates[4]. Here, we apply annotation-independent approaches for synonymous codon usage to the microbiomes of three cities: New York City[1], Boston[5], and Sacramento.

Materials and Methods

Metagenomic data for the MetaSUB Inter-City Challenge were downloaded from the CAMDA 2017 website (http://contest.camda.info). Filtering and assembly of raw read data was performed using MOCAT2 (http://mocat.embl.de) with default parameters[6]. Samples in which the number of processed reads was too low were subsequently excluded from the analysis. Prodigal[7] was used to predict protein-coding sequences in the assembled metagenome. The final data set included 51 metagenomes: 19 from Boston, 19 from NY, and 13 from Sacramento (Table 1).

For each metagenome, principal component analysis (PCA) was performed to identify major trends of variation in synonymous codon usage among genes[8]. To interpret principal components (PCs), we analyzed correlations between the PC scores and three gene features: the proportion of G and C and that of G and T at the third codon positions (GC3 and GT3), and the P2 index which represents the proportion of codons conforming to the intermediate strength of codon-anticodon interaction[9,10]. Figure 1 shows scatter plots of the first and second principal components (PC1 and PC2) scores obtained by PCA, plotted against nucleotide contents (GC3 and GT3) for genes from the metagenomic sample SRR1749476 in NY (Table 1). At the threshold correlation coefficient (r) value of 0.70, GC3 values were significantly correlated with PC1 scores ($|r| = 0.966$) and GT3 values were significantly correlated with PC2 scores ($|r| = 0.704$). GC3 and GT3 were thus identified as the main trends of variation among genes on PC1 and PC2, respectively.

For each metagenomic sample, the mean distance (*D*mean) between all pairs of genes was calculated to measure diversity in synonymous codon usage[11]. The distance between two genes was measured as 1 – r, where r is the Pearson's product moment correlation coefficient between the two vectors of normalized codon usage data called relative adaptiveness (W). We used the W value to avoid effects of gene length, amino acid composition, and codon degeneracy.

We assessed the robustness of our results by varying sequence data sets (e.g. excluding genes of <100 or <200 codons in length). The codon usage analyses were conducted using the G-language Genome Analysis Environment version 1.9.1 (http://www.g-language.org)[12]. Statistical analyses were implemented using the R version 3.3.3 (https://www.R-project.org).

Results and discussion

The PCA method identified three gene features (GC3, GT3, and P2) as major trends of variation in synonymous codon usage among genes for the city metagenomes (Table 1):

1. GC3 was detected in all the 51 samples. This is consistent with the previous report that synonymous codon usage is affected primarily by the overall G+C content of the genome[13]. GC3 shows a wide variation among bacteria and has thus been used to detect genes acquired by horizontal transfer[14].
2. GT3 was detected in most (15 out of 19) Boston samples, 1 out of 19 NY samples, and none of the 13 Sacramento samples. GT3 is higher in the leading strand than in the lagging strands of DNA replication and reflects strand-specific mutation biases in single bacterial genomes.
3. P2 was detected in 6 out of 19 NY samples but it was not detected in any metagenomic samples from Boston and Sacramento. P2 indicates the efficiency of the codon-anticodon interaction and highly expressed genes tend to have high P2 values in *Escherichia coli* and yeast[15]. This suggests that synonymous codon usage in these NY metagenomic samples could be subject to translational selection although there is no obvious common feature (e.g. geographical locations, surface types and materials) in these samples.

Thus, one can detect trends of synonymous codon usage variation among genes at the level of metagenomes as well as single bacterial genomes.

The *D*mean values (Figure 2) indicated that synonymous codon usage diversity was high in Sacramento, intermediate in Boston, and low in the New York City. The differences were significant (Kruskal-Wallis rank sum test; p-value = 9.435e-08). This suggests that Sacramento metagenomes contained diverse bacteria with different codon preferences. We checked that this is not due to a systematic compositional bias in the Sacramento metagenomic samples.

Our results suggest that codon usage can provide additional information on genetic diversity in microbiomes, and be used to predict genes under mutational biases and translational selection (e.g. highly expressed genes) from sequence data alone.

Table 1: Gene features (GC3, GT3, and P2) detected by PCA in city metagenomes.

			Principal components		
Sample	City	*D*mean	PC1	PC2	PC3
Sample_1A	Sacramento	0.466	GC3	nd	nd
Sample_1C	Sacramento	0.778	GC3	nd	nd
Sample_2A	Sacramento	0.745	GC3	nd	nd
Sample_2B	Sacramento	0.760	GC3	nd	nd
Sample_2C	Sacramento	0.800	GC3	nd	nd
Sample_3A	Sacramento	0.781	GC3	nd	nd
Sample_3B	Sacramento	0.729	GC3	GT3	nd
Sample_3C	Sacramento	0.684	GC3	nd	nd
Sample_5A	Sacramento	0.857	GC3	nd	nd
Sample_5B	Sacramento	0.819	GC3	nd	nd
Sample_5C	Sacramento	0.838	GC3	nd	nd
Sample_6A	Sacramento	0.761	GC3	nd	nd
Sample_6B	Sacramento	0.725	GC3	nd	GT3
SRR1749406	NY	0.366	GC3	P2	nd
SRR1749410	NY	0.248	GC3	nd	nd
SRR1749412	NY	0.594	GC3	nd	nd
SRR1749419	NY	0.259	GC3	nd	nd
SRR1749421	NY	0.186	GC3	nd	nd
SRR1749422	NY	0.230	GC3	nd	nd
SRR1749423	NY	0.366	GC3	P2	nd
SRR1749437	NY	0.257	GC3	nd	nd
SRR1749454	NY	0.267	GC3	nd	nd
SRR1749457	NY	0.398	GC3	nd	P2
SRR1749476	NY	0.370	GC3	GT3	nd
SRR1749495	NY	0.671	GC3	nd	nd
SRR1749512	NY	0.506	GC3	nd	P2
SRR1749516	NY	0.173	GC3	nd	nd
SRR1749519	NY	0.212	GC3	nd	nd
SRR1749529	NY	0.402	GC3	nd	nd
SRR1749544	NY	0.254	GC3	nd	nd
SRR1749671	NY	0.369	P2	GC3	nd
SRR1750012	NY	0.608	GC3	nd	P2
SRR3545898	Boston	0.555	GC3	GT3	nd
SRR3545919	Boston	0.722	GC3	nd	nd
SRR3545934	Boston	0.509	GC3	GT3	nd
SRR3545941	Boston	0.484	GC3	GT3	nd
SRR3545948	Boston	0.692	GC3	nd	nd
SRR3545955	Boston	0.500	GC3	GT3	nd
SRR3545963	Boston	0.547	GC3	GT3	nd
SRR3546354	Boston	0.484	GC3	GT3	nd
SRR3546356	Boston	0.673	GC3	nd	nd
SRR3546358	Boston	0.486	GC3	GT3	nd
SRR3546361	Boston	0.646	GC3	nd	nd
SRR3546363	Boston	0.494	GC3	GT3	nd
SRR3546365	Boston	0.519	GC3	GT3	nd
SRR3546367	Boston	0.544	GC3	GT3	nd
SRR3546371	Boston	0.497	GC3	GT3	nd
SRR3546373	Boston	0.694	GC3	GT3	nd
SRR3546375	Boston	0.647	GC3	nd	GT3
SRR3546380	Boston	0.557	GC3	GT3	nd
SRR3555059	Boston	0.486	GC3	GT3	nd

nd, any gene features considered were not detected.

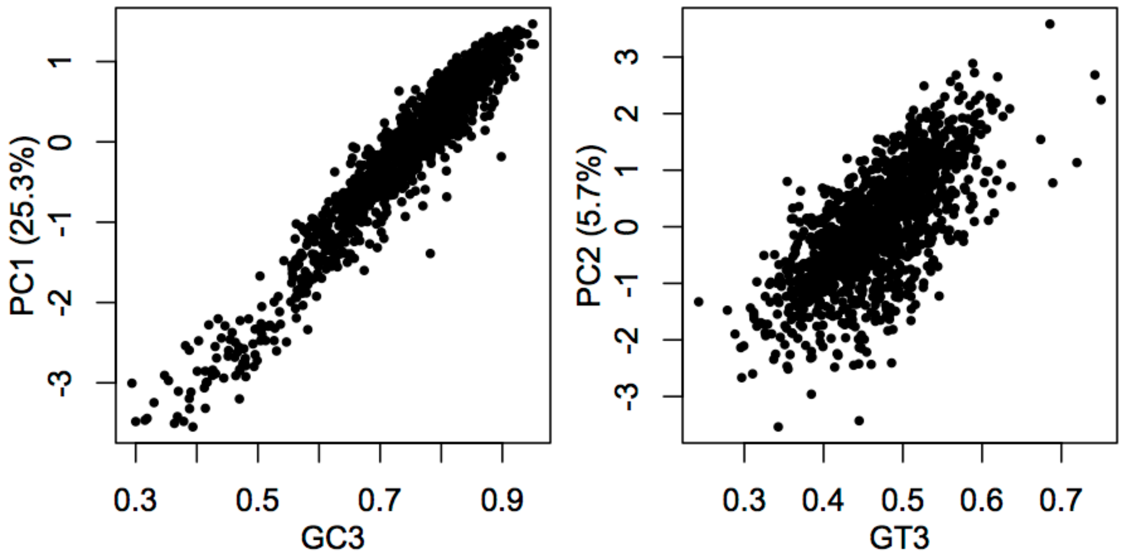

Figure 1: Scatter plot showing PC1 and PC2 scores obtained by principal component analysis (PCA) of codon usage in metagenomic sample SRR1749476, plotted against GC3 and GT3, respectively. Each dot represents a gene. Proportions of variance explained by each PC are shown in parentheses.

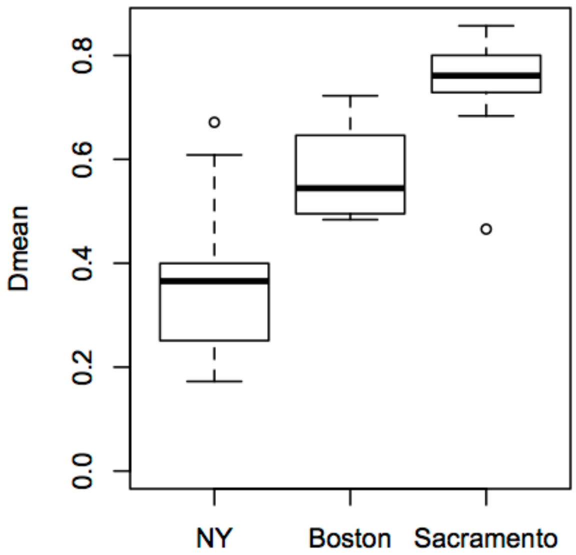

Figure 2: Diversity in synonymous codon usage among genes for the metagenomes of three cities (NY, Boston, and Sacramento), measured by a mean distance (*D*mean) between all pairs of genes.

References

1. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1,** 72–87 (2015).
2. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33,** 1141–53 (2005).
3. Roller, M., Lucić, V., Nagy, I., Perica, T. & Vlahoviček, K. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res.* **41,** 8842–8852 (2013).
4. Vieira-Silva, S. & Rocha, E. P. C. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet.* **6,** e1000808 (2010).
5. Hsu, T. *et al.* Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. *mSystems* **1,** e00018-16 (2016).
6. Kultima, J. R. *et al.* MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS One* **7,** e47656 (2012).
7. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
8. Suzuki, H., Saito, R. & Tomita, M. A problem in multivariate analysis of codon usage data and a possible solution. *FEBS Lett.* **579,** 6499–504 (2005).
9. Grosjean, H. & Fiers, W. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18,** 199–209 (1982).
10. Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10,** 7055–74 (1982).
11. Suzuki, H., Saito, R. & Tomita, M. Measure of synonymous codon usage diversity among genes in bacteria. *BMC Bioinformatics* **10,** 167 (2009).
12. Arakawa, K. *et al.* G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* **19,** 305–6 (2003).
13. Lynn, D. J., Singer, G. A. C. & Hickey, D. A. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30,** 4272–7 (2002).
14. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring horizontal gene transfer. *PLoS Comput. Biol.* **11,** e1004095 (2015).
15. Shields, D. C. & Sharp, P. M. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res.* **15,** 8023–40 (1987).