Analysis of CpG Dinucleotide Frequency in Bacterial Genomes with Respect to Genomic Regions and Codon Positions

Mami Goto[1,2] mgoto@sfc.keio.ac.jp
Takanori Washio[1,2] washy@sfc.keio.ac.jp
Masaru Tomita[1,3] mt@sfc.keio.ac.jp

[1] Laboratory for Bioinformatics
[2] Graduate School of Media and Governance
[3] Department of Environmental Information, Keio University,
5322 Endo, Fujisawa 252-0816, Japan

Keywords: CpG dinucleotide, comparative genomics, nucleotide substitution

1 Introduction

CpG suppression is generally known as a characteristic of higher organisms such as vertebrates and plants. CpG dinucleotides are believed to have been depleted through the point mutations from CpG to TpG/CpA caused by the methylation of the 5' end of the cytosine of CpG [1].

There are some bacterial genomes showing low CpG frequencies. Since these organisms do not have CpG methyltransferase, the cause of CpG depletion in bacteria is an open question.

In this study, we systematically analyzed the CpG frequencies using 19 bacterial complete genomes. We divided the genome sequences into three regions: protein coding region, non-coding region, and RNA coding region, and obtained CpG frequencies for each region for each organism. We also conducted a detailed investigation of the individual base substitutions, by comparing the orthologous gene pairs in the two closely-related species *M. genitalium* and *M. pneumoniae*.

2 Results and Discussion

CpG frequencies in each region Table 1 shows the CpG frequencies (observed/expected ratio) in protein coding region, RNA coding region, and non-coding region of low CpG genomes. CpG frequencies are low in both protein coding region and non-coding region, while those in RNA coding region are significantly higher.

Comparison between *M. genitalium* and *M. pneumoniae* In protein coding region, the substitution rates of CpG-TpG/CpA in the mycoplasmas are very high (29.8% and 14.2%, respectively). The C-T and G-A substitution rates are also high in many other dinucleotides (ApC-ApT, 34.1%; TpC-TpT. 31.5%; CpT-TpT, 32.0%; ApG-ApA, 30.8%). In RNA coding regions of *M. genitalium* and *M. pneumoniae*, on the other hand, the substitution rate of the CpG is quite low; in fact, 95.5% of CpG are conserved.

CpG depletion and codon usage At codon positions {1,2}, 52.5% of CpG in *M. pneumoniae* are conserved in *M.genitalium*, which is markedly high compared with 8.3% at codon positions {2,3} and 6.9% at codon positions {3,4}(4=1 of the next codon). The following codons are particularly avoided in *M. genitalium* compared with those in *M. pneumoniae*: CGN (arginine), NCG (proline, serine, threonine, and alanine). This result indicates that the CpG depletion in the protein coding region is largely related to the biased usage of synonymous codons.

Table 1: CpG frequenciesO/E ratioof low CpG genomes with respect to genomic regions

species	whole genome	protein coding region	RNA coding region	non-coding region
M. genitalium	0.39(− − −)	0.37(− − −)	0.81	0.44(− − −)
B. burgdorferi	0.48(− − −)	0.47(− − −)	0.89	0.54(− −)
Synechocystis sp.	0.75(−)	0.75(−)	0.89	0.67(− −)
C. pneumoniae	0.73(−)	0.72(−)	0.92	0.73(−)
R. prowazekii	0.77(−)	0.77(−)	0.93	0.69(− −)
A. fulgidus	0.78(−)	0.78(−)	0.84	0.77(−)
M. jannaschii	0.32(− − −)	0.27(− − −)	0.83	0.57(− −)
M. thermoautotrophicum	0.51(− −)	0.51(− −)	0.74(−)	0.49(− − −)
P. horikoshii	0.61(− −)	0.60(− −)	0.86	0.54(− −)
P. abyssi	0.71(−)	0.70(−)	0.88	0.69(− −)

We refer to values in the range $Fxy <= 0.78$ as low.
Underrepresentation is indicated by $-(0.70 < Fxy <= 0.78)$, $- -$ $(0.50 < Fxy <= 0.70)$, $- - -(Fxy <= 0.50)$ [2].

Table 2: Substitutions in orthologous genes of *M. genitalium* and *M. pneumoniae*(%)

MG↘ MP↓	AA	AT	AG	AC	TT	TG	TC	GG	CA	CT	CG	CC
AA	78.14	1.19	7.09	1.32	0.40	0.25	0.36	0.29	2.57	0.20	0.06	0.12
AT	2.08	80.26	0.71	5.62	3.12	0.02	0.12	0.05	0.17	1.50	0.02	0.02
AG	**30.80**	2.72	49.73	1.52	0.38	5.06	0.76	3.67	0.79	0.28	0.57	0. 19
AC	5.63	**34.13**	2.12	44.84	1.47	0.18	1.86	0.26	0.50	0.58	0.03	1.41
TT	0.35	3.34	0.05	0.39	77.86	1.20	4.75	0.02	0.32	6.00	0.03	0.16
TG	0.92	0.35	5.68	0.27	9.16	65.98	2.09	1.74	0.79	0.41	2.04	0.38
TC	0.90	0.90	0.68	2.79	**31.53**	4.37	40.32	0.18	0.54	1.04	0.00	5.32
GG	3.17	0.73	15.77	1.31	0.43	8.70	0.43	54.12	0.31	0.04	0.50	0.08
CA	11.30	0.42	0.65	0.49	0.81	1.38	0.68	0.08	48.94	5.58	1.22	2.68
CT	1.28	**6.80**	0.36	0.86	**32.00**	0.33	0.95	0.03	8.86	41.10	0.98	3.16
CG	3.03	0.55	**9.35**	0.55	0.95	**29.75**	0.65	1.39	**14.18**	**15.72**	**16.67**	4.88
CC	1.67	1.38	0.43	**3.33**	2.79	1.23	10.87	0.00	13.98	23.25	1.48	33.39

3 Concluding Remarks

At present, the explanation for the CpG decreases in some bacterial genomes is not clearly known. Our results show that the CpG depletion is not due to codon bias, because the same level of depletion is observed in non-coding region. We consider it likely that a certain global pressure exists, which may have the effect of decreasing CpG over the entire genome. That pressure, then, results in the codon bias, and CpG is avoided in coding region by the use of synonymous codons without CpG

Acknowledgments

This work was supported in part by Japan Science and Technology Corporation and a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Bestor, T. H. and Coxon, A., The pros and cons of DNA methylation, *Curr. Biol.*, 6:384–386, 1993.

[2] Karlin, S. and Burge, C., Dinucleotide relative abundance extremes: a genomic signature, *Trends Genet.*, 11:283–90, 1995.