On Low Frequency of CpG Dinucleotides in Bacterial Genomes

Mami Goto [13] **Takanori Washio** [13] **Masaru Tomita** [23]

mgoto@sfc.keio.ac.jp washy@sfc.keio.ac.jp mt@sfc.keio.ac.jp

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

CpG depletion is a phenomenon known to be caused by CpG DNA methylation. Although CpG DNA methylation is believed to be a feature of vertebrates and plants [1], we have found that some procaryote genomes such as *M. genitalium* show significantly low CpG frequencies. On the other hand, the genome of *M. pneumoniae*, which is a closely related species to *M. genitalium*, does not show such clear sign of CpG depletion.

In order to discuss possible causes of the bacterial CpG depletion, we conducted computer analyses of frequencies of CpG dinucleotides in twelve complete procaryote genomes, including *M. genitalium* and *M. pneumoniae*.

For all of the complete genome sequences, we first analyzed CpG frequencies in coding and non-coding regions separately. The results of the analysis are shown in Table 1. CpG observed/expected (O/E) values are significantly lower in coding sequences than non-coding sequences in the genomes of *M. genitalium*, *B. burgdorferi* and *M. jannaschii*. In the genome of *M. thermoautotrophicum*, on the other hand, CpG O/E values are about the same in both coding and non-coding sequences.

Comprehensive analysis of mutation patterns of single nucleotide and dinucleotide substitutions has been conducted using homologous sequence pairs of the genomes of of *M. genitalium* and *M. pneumoniae*. These homologous sequences were aligned, and we counted substituted nucleotides and dinucleotides only if six bases directly flanking them (three bases upstream and three bases dowstream) are perfectly conserved. The results are shown in Table 2.

While the point mutations from C (cytosine) in *M. pneumoniae* to T (thymine) in *M. genitalium* are frequent in general, C to T mutations occurred most frequently on the cytocine in the CpG dinucleotide (Table 3).

Table 1: CpG Observed/Expected (O/E) ratio. Expected CpG frequencies are products of C and G single nucleotide frequencies.

species	Whole sequence	Coding sequence	Non coding sequence
Mycoplasma genitalium	**0.39**	**0.37**	**0.54**
Mycoplasma pneumoniae	0.82	0.82	0.83
Escherichia coli	1.16	1.16	1.09
Haemophilus influenzae	1.09	1.09	1.07
Bacillus subtilis	1.04	1.05	0.87
Helicobacter pylori	0.93	0.94	0.84
Borrelia burgdorferi	**0.48**	**0.47**	**0.73**
Synechocystis PCC6803	0.75	0.75	0.68
Aquifex aeolicus	0.87	0.87	0.87
Archaeoglobus fulgidus	0.78	0.77	0.83
Methanococcus jannaschii	**0.32**	**0.27**	**0.67**
Methanobacterium thermoautotrophicum	**0.51**	**0.51**	**0.54**

Table 2: Nucleotide variations in homologous genes of *M. genitalium* and *M. pneumoniae*.

A → T	734	5.67 %
A → G	1003	7.75 %
A → C	401	3.10 %
T → A	922	7.12 %
T → G	284	2.19 %
T → C	1036	8.01 %
G → A	2205	17.04 %
G → T	941	7.27 %
G → C	399	3.08 %
C → A	1189	9.19 %
C → T	**3469**	**26.81 %**
C → G	358	2.77 %
	12941	100.00 %

Table 3: Dinucleotide variations in homologous genes of *M. genitalium* and *M. pneumoniae*. $\boxed{\frac{occurrences}{frequency(MP \to MG)(\%)}}$

MG↘ MP↓	AA	AT	AG	AC	TA	TT	**TG**	TC	GA	GT	GG	GC	**CA**	CT	**CG**	CC	
AA	5050	77	458	85	178	26	16	23	308	14	19	18	166	13	4	8	6463
	78.14	1.19	7.09	1.32	2.75	0.40	0.25	0.36	4.77	0.22	0.29	0.28	2.57	0.20	0.06	0.12	
AT	85	3272	29	229	7	127	1	5	8	227	2	15	7	61	1	1	4077
	2.08	80.26	0.71	5.62	0.17	3.12	0.02	0.12	0.20	5.57	0.05	0.37	0.17	1.50	0.02	0.02	
AG	973	86	1571	48	19	12	160	24	46	26	116	20	25	9	18	6	3159
	30.80	2.72	49.73	1.52	0.60	0.38	5.06	0.76	1.46	0.82	3.67	0.63	0.79	0.28	0.57	0.19	
AC	215	1303	81	1712	27	56	7	71	36	77	10	127	19	22	1	54	3818
	5.63	34.13	2.12	44.84	0.71	1.47	0.18	1.86	0.94	2.02	0.26	3.33	0.50	0.58	0.03	1.41	
TA	225	9	20	6	2413	277	183	49	51	3	5	2	265	10	3	10	3531
	6.37	0.25	0.57	0.17	68.34	7.84	5.18	1.39	1.44	0.08	0.14	0.06	7.50	0.28	0.08	0.28	
TT	22	212	3	25	264	4935	76	301	10	63	1	14	20	380	2	10	6338
	0.35	3.34	0.05	0.39	4.17	77.86	1.20	4.75	0.16	0.99	0.02	0.22	0.32	6.00	0.03	0.16	
TG	34	13	209	10	358	337	2428	77	5	6	64	6	29	15	75	14	3680
	0.92	0.35	5.68	0.27	9.73	9.16	65.98	2.09	0.14	0.16	1.74	0.16	0.79	0.41	2.04	0.38	
TC	20	20	15	62	197	700	97	895	12	7	4	38	12	23	0	118	2220
	0.90	0.90	0.68	2.79	8.87	31.53	4.37	40.32	0.54	0.32	0.18	1.71	0.54	1.04	0.00	5.32	
GA	561	19	37	30	185	15	21	6	2783	49	62	41	147	4	3	5	3968
	14.14	0.48	0.93	0.76	4.66	0.38	0.53	0.15	70.14	1.23	1.56	1.03	3.70	0.10	0.08	0.13	
GT	29	533	19	84	8	238	10	11	155	1903	91	209	22	92	2	9	3415
	0.85	15.61	0.56	2.46	0.23	6.97	0.29	0.32	4.54	55.72	2.66	6.12	0.64	2.69	0.06	0.26	
GG	82	19	408	34	26	11	225	11	192	92	1400	63	8	1	13	2	2587
	3.17	0.73	15.77	1.31	1.01	0.43	8.70	0.43	7.42	3.56	54.12	2.44	0.31	0.04	0.50	0.08	
GC	104	47	62	245	23	26	7	113	122	430	66	1051	24	45	1	46	2412
	4.31	1.95	2.57	10.16	0.95	1.08	0.29	4.68	5.06	17.83	2.74	43.57	1.00	1.87	0.04	1.91	
CA	435	16	25	19	840	31	53	26	138	8	3	7	1884	215	47	103	3850
	11.30	0.42	0.65	0.49	21.82	0.81	1.38	0.68	3.58	0.21	0.08	0.18	48.94	5.58	1.22	2.68	
CT	43	228	12	29	9	1073	11	32	9	77	1	15	297	1378	33	106	3353
	1.28	6.80	0.36	0.86	0.27	32.00	0.33	0.95	0.27	2.30	0.03	0.45	8.86	41.10	0.98	3.16	
CG	61	11	188	11	34	19	**598**	13	2	8	28	3	**285**	316	**335**	98	2010
	3.03	0.55	9.35	0.55	1.69	0.95	**29.75**	0.65	0.10	0.40	1.39	0.15	**14.18**	15.72	**16.67**	4.88	
CC	46	38	12	92	71	77	34	300	14	47	0	39	386	642	41	922	2761
	1.67	1.38	0.43	3.33	2.57	2.79	1.23	10.87	0.51	1.70	0.00	1.41	13.98	23.25	1.48	33.39	
	7985	5903	3149	2721	4659	7960	3927	1957	3891	3037	1872	1668	3596	3226	579	1512	57642

We therefore conclude that this type of mutations, CpG to TpG/CpA, is the primary force of the CpG depletion in *M. genitalium*. However, since *M. genitalium* does not have a gene homologue to CpG DNA methylase, the cause of the frequent CpG mutation is yet to be known.

Acknowledgements

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

References

[1] Shimizu, T.S., Takahashi, K. and Tomita, M., CpG distribution patterns in methylated and non-methylated species, *Gene*, 205:103–107, 1997.