J Mol Evol (2004) 58:692–700
DOI: 10.1007/s00239-004-2591-1

JOURNAL OF MOLECULAR EVOLUTION

Cytosine Methylation Is Not the Major Factor Inducing CpG Dinucleotide Deficiency in Bacterial Genomes

Yong Wang,[1,3] Eduardo P.C. Rocha,[2] Frederick C.C. Leung,[1] Antoine Danchin[2,3]

[1] Department of Zoology, University of Hong Kong, Pokfulam, Hong Kong SAR, China
[2] Unite GGB, URA 2171, Institut Pasteur, 28 rue Dr. Roux, 75015, Paris, France
[3] HKU-Pasteur Research Centre, 8 Sassoon Road, PokFulam Hong Kong SAR, China

Received: 14 June 2003 / Accepted: 10 December 2003

Abstract. CpG dinucleotide deficiency has been found in viruses, mitochondria, prokaryotes, and eukaryotes. The consensual explanation is that it is due to deamination of methylated cytosines, as established for vertebrate and plants. However, we still do not know whether C5 cytosine methylation is also the major cause of CpG deficiency in bacteria. By combining annotation and experimental data identifying the presence of C5 cytosine methyltransferases with analysis of CpG relative abundance in 67 bacterial species, we found that CpG relative abundance in most bacterial genomes that have cytosine C5 methyltransferases tends to be in the normal range (observed/expected values between 0.82 and 1.21). In contrast, many bacterial species likely to be lacking C5 cytosine methylation showed CpG deficiency. Furthermore, when comparing genomes with one another, TpG and CpA relative abundances were found to be independent from CpG relative abundance. This contrasted with intragenome analyses, where C_3pG_1 relative abundance (the subscripts refer to position of a nucleotide in a codon) was found to be generally positively correlated with T_3pG_1 relative abundances when plotted against GC content in protein coding sequences (CDSs). This suggests the existence of alternative mechanisms contributing to CpG deficiency in bacteria.

Key words: CpG deficiency — C5-specific methylation — C5 methyltransferase — Recognition sites — Bacterial genomes — GC content

Introduction

CpG deficiency was first observed in vertebrates (Josse et al. 1961; Swartz et al. 1962), then in some species of archaea, bacteria, and fungi, as well as in mitochondria belonging to many organisms (Cardon et al. 1994; Karlin et al. 1998). CpG dinucleotides play an important role in cell differentiation and in the regulation of gene expression in vertebrates (Bestor 1990). CpG deficiency can also influence codon usage bias (De Amicis and Marchetti 2000) and the relative abundance of oligonucleotides, thereby indirectly affecting a variety of cell functions. This triggered many studies aiming at understanding genome base composition biases (Karlin et al. 1998). Several hypotheses have been put forward to explain CpG deficiency, including counter-selection at the translation level (Subak-Sharpe et al. 1966), DNA methylation (Bird 1980), DNA structural constraints (Antri et al. 1993), DNA–protein interaction, and stressful environments (Karlin et al. 1994b). Among them, DNA methylation is the most popular hypothesis.

Cytosine deamination is a major cause of mutation in living organisms, especially in open DNA

Correspondence to: Antoine Danchin; *email:* adanchin@pasteur.fr

structures (for recent references and discussion see Lobry and Sueoka 2002). It is, however, readily repaired, since deamination leads to uracil, subject to proofreading in DNA. It is widely documented that methylated cytosine is even more prone to spontaneous deamination and this induces transition mutations to the natural base thymine (Coulonder et al. 1978). Such mutations are hard to repair (Coulonder et al. 1978). Since methylated cytosines were predominantly found within CpG dinucleotides in vertebrates, CpG deficiency was naturally linked to CpG methylation (Bird 1980). The presence of highly methylated CpG dinucleotides in both male and female germ cells provided strong evidence for the relationship between DNA methylation and CpG deficiency in the human genome (El-Maarri et al. 1998). However, cytosine methylation may not be the ultimate or only explanation for CpG deficiency. For example, CpG deficiency in most mitochondrial genomes is unlikely to be related to DNA methylation, because DNA methylase has not yet been discovered in these organelles. One of the few reports on methylation in mitochondria identified an RNA methylation by a nucleus-encoded RNA adenine methyltranferase (McCulloch et al. 2002). CpG deficiency was also found in many bacterial species and their phages (Karlin et al. 1994a, 1997), where cytosine methylation is not widespread (see below).

This prompted us to revisit the association between DNA methylation and CpG deficiency in bacterial genomes. In bacteria, DNA methylation is generally associated with restriction-modification systems (RM systems) (Wilson 1988). These elements may prevent the invasion of the cell by bacteriophages. So far, more than 2000 different RM systems have been identified and over 700 methyltransferases are known to recognize at least 300 different DNA sites (http://www.neb.com/rebase) (Roberts and Macelis 2001). Three kinds of DNA methylation systems were found in bacteria: A6-adenine methylation, N4-cytosine methylation, and C5 cytosine methylation (Bestor 1990). In this report, we focus our attention on C5 cytosine-specific methylation, the same DNA methylation process that is assumed to induce CpG deficiency in eukaryotes. Due to versatile functions and recognition sites of DNA methylation in bacteria compared to vertebrates, DNA methylation is unlikely to share a common role in all bacterial genomes. This was previously suggested in a study on the *Mycoplasma genitalium* genome in which CpG deficiency was suspected to be unrelated to DNA methylation (Goto et al. 2000). The suspicion was based on the finding that the high substitution rate from C to T was not specific to CpG and TpG dinucleotides and the fact that there was no reported methylation activity in mycoplasmas (Goto et al. 2000). In the present study, we further document that deamination of methylated cytosine is probably not the reason for the CpG deficiency in bacteria.

Methods

Sources of Data

First, the fully sequenced bacterial genomes were surveyed, after being retrieved from the NCBI (http://www.ncbi.nlm.nih.gov). We searched for potential C5 methyltransferase genes using the annotation files. When such a cytosine methyltransferase was identified, the bacterial identification was used to search for the corresponding enzyme in the REBASE database (http://rebase.neb.com) (Roberts and Macelis 2001). Almost all the cytosine methyltransferases were C5 methyltransferases, and the one case of N4 cytosine methylation was discarded. When more than one C5 methyltransferase was found in a genome, only the one including a CpG dinucleotide at the restriction site was included.

Cytosine-specific methyltransferase genes are labeled as "putative" in some bacteria. This makes in-depth analysis difficult because the biochemical properties of their products are not substantiated in REBASE. Therefore, the latter approach is only feasible for well-studied bacteria in which the presence of cytosine methylation has been studied. As a complement of explicit identification, we used the BLASTP tool provided by REBASE to ascertain that a CDS putatively coding for a C5 methyltransferase is highly similar to a known C5 methyltransferase CDS.

Second, utilizing REBASE, we also identified C5 methyltransferases in the unfinished genomes of several bacterial species. When such a gene was found, we collected the available DNA sequences from NCBI, extending our study to the corresponding organisms. By exploring REBASE in addition to two other protein databases, Pfam at the Sanger Centre (http://www.sanger.ac.uk/Software/Pfam/) and TIGRFAMs at TIGR (http://www.tigr.org/TIGRFAMs/), we collected DNA sequences from all the bacteria that are likely to express C5 methyltranferases. Finally, only bacterial species for which more than 20 nonredundant sequences (excluding ribosomal DNA) could be retrieved from GenBank were included in the analysis.

Relative Abundance of Dinucleotides

To measure the frequency of dinucleotides in a long genomic sequence, the value of relative abundance was calculated by computing the relevant odds ratio (Burge et al. 1992). In the case of CpG dinucleotide, the formula is $\rho_{CpG} = F_{CpG}/F_C{*}F_G$, where ρ_{CpG} denotes relative abundance of CpG and F_{CpG} denotes the frequency of CpG dinucleotide. If ρ_{CpG} falls between 0.81 and 1.20, the CpG dinucleotide is considered to be at a normal level. If it is lower than 0.81, the CpG relative abundance is classified as being deficient. However, the relative abundance of this dinucleotide can be further classified as follows: 0.78–0.81 is marginally low, 0.70–0.78 is significantly low, 0.50–0.70 is very low, and ≤ 0.50 is extremely low (Burge et al. 1992). In this study, the bacteria with CpG relative abundances lower than 0.78 were considered to be CpG deficient.

GC Content and CpG Deficiency at Neutral Positions of CDS

Generally bacterial CDSs are short in size, so the variance of CpG relative abundances of the CDSs with the same GC content is very large. Especially in low-GC content CDSs, the values will highly deviate from the trend line when they are plotted against GC content. The deviation could strongly mask the changing tendency

Table 1. The recognition sites of the C5 methyltransferases lacking a methylated CpG dinucleotide

Bacteria	GC content (%)	ρ_{CpG}	C5 methyltransferase	Recognition site
*Acetobacter pasteurianus**	55.4	0.93	M. *Apa*LI	GTGCAC
*Anabaena variabilis**	42.4	0.83	M. *Ava*IX	R^{m}CCGGY
*Bacillus brevis**	44.1	1.06	M. *Bb*VI	G^{m}CAGC
*Bacillus cereus**	36.8	0.95	M. *Hae*III	GG^{m}CC
*Bacillus firmus**	41.1	0.86	M. *Bfi*IB	A^{m}CTGGG
Bacillus halodurans (NC_002570)	51.7	1.31	M. *Bha*II	GGCC
*Bacillus pumilus**	40.2	0.86	M. *Bpu*10IA	CCTNAGC
*Bacillus sphaericus**	35.9	0.89	M. *Bsp*RI	GG^{m}CC
Bacillus subtilis (AL009126)	43.5	1.04	M. *Bsu*FI	^{m}CCGG
*Citrobacter freundii**	59.3	1.14	M. *Cfr*10I	R^{m}CCGGY
Clostridium acetobutylicum (NC_003030)	30.9	0.45(- - -)	M. *Cac*824I	GCNGC
Corynebacterium glutamicum (NC_003450)	53.7	0.97	M. *Cgl*I	GCSGC
*Enterobacter aerogenes**	53.6	1.19	M. *Eae*I	YGG^{m}CCR
*Enterobacter cloaceae**	54.6	1.1	M. *Ecl*18kI	C^{m}CNGG
Escherichia coli K12 (U00096)	50.8	1.16	M. *Eco*KDcm	C^{m}CWGG
Escherichia coli O157:H7 Sakai (NC_002695)	50.4	1.12	M. *Eco*KO157DcmP	CCWGG
Lactococcus lactis subsp. Cremoris	35.5	0.83	M. *Scr*FIA	CCNGG
*Neisseria gonorrhoeae**	52.6	1.33	M. *Ngo*PII	GG^{m}CC
*Neisseria lactamica**	49.7	1.33	M. *Nla*IV	GGNNCC
Neisseria meningitidis (NC_003116)	51.7	1.31	M. *Nme*AORF191P	CCWGG
Neisseria meningitidis MC58 (NC_003112)	51.4	1.31	M. *Nme*BIA	GGNNCC
Nostoc sp. PCC 7120 (NC_003272)	41.3	0.79	M. *Ava*IX	R^{m}CCGGY
*Salmonella enteritidis**	48.8	1.07	M. *Sen*PI	C^{m}CNGG
Salmonella typhi CT18 (NC_003198)	52.0	1.24	M. *Sty*CDcmP	CCWGG
Salmonella typhimurium (NC_003197)	52.2	1.24	M. *Sty*LT2DcmP	CCWGG
*Shigella sonnei**	43.1	1.02	M. *Sso*II	C^{m}CNGG
Yersinia pestis (NC_004088)	47.6	0.99	M. *Ype*ORF391P	CCWGG

* The whole genomic data are not available. Accession numbers of whole bacterial genomes are indicated in parentheses. A relative abundance (ρ_{CpG}) followed by (-), (- -), (or) (- - -) is significantly low, very low, or extremely low, respectively. Methylated cytosines are preceded by a superscript m. W denotes A or T; S denotes G or C; N denotes any nucleotide.

of CpG relative abundance. Since the calculated ρ_{CpG} for longer sequences do not deviate from actual values as much as those for shorter sequences (i.e., decreasing magnitude of deviation from actual value as CDS length increases), we first listed all the CDSs according to their GC contents. We then concatenated every 40 CDSs (every 20 CDSs for some small bacterial genomes, like the *C. trachomatis* genome) to generate long coding sequences for this study. The third position of a codon is under less selective pressure due to the redundancy in the genetic code, therefore we chose C_3pG_1 (C in the third position of a codon; G in the first position of the following codon) to study the mutation pattern of CpG dinucleotides. The relative abundances of C_3pG_1 and T_3pG_1 in each sequence were calculated and then plotted against the GC content of the CDS.

Results

Classification of Genomes According to C5 Methyltranferase

A total of 47 bacterial species whose genomes contain C5 methyltransferases were analyzed in terms of GC content, CpG relative abundance, C5 methyltransferase, and C5 methylation site. These species were categorized into three groups according to their C5 methyltransferase recognition sites (the length of the recognition sites was in the range of four to seven nucleotides). Some of these sites contain a methylated CpG dinucleotide, while others do not. In our first group, non-CpG dinucleotides are methylated in the recognition sites of the C5 methyltransferases (Table 1). In our second group, the presence of methylated CpG dinucleotides in recognition sites is uncertain (Table 2). In our third group, a methylated CpG dinucleotide can be found in the recognition sites (Table 3). Although we still do not know which cytosine is methylated in the recognition site of CGATCG (for *Escherichia coli* O157:H7 EDL933) in Table 3, the recognition site must have a methylated CpG dinucleotide because both cytosines in the recognition site are within the CpG dinucleotides. In addition, 20 bacterial species were found to be lacking C5 methyltransferases (their CpG relative abundances are listed in Table 4).

Is CpG Deficiency a Result of Horizontal Transfer of RM Systems?

RM systems in free-living bacteria are often horizontally transferred by means of linkage with mobility-related elements such as phages and plasmids (Kobayashi 2001 and references therein). RM systems act like an infectious agent, by rendering the bacteria dependent on the functioning of the methy-

Table 2. The recognition sites of the C5 methyltransferases possibly containing a methylated CpG dinucleotide

Bacteria	GC content (%)	ρ_{CpG}	C5 methyltransferase	Recognition site
*Bacillus stearothermophilus**	47.4	1.33	M. *Bsr*FI	RCCGGY
Caulobacter crescentus CB15 (NC_002696)	67.1	1.16	M. *Ccr*MORF1033P	Unknown
Haemophilus influenzae (NC_000907)	38.1	1.09	M. *Hind*V	GRCGYC
*Herpetosiphon giganteus**	41.8	1.01	M. *Hgi*GI	GRCGYC
Listeria monocytogenes EGD (NC_003210)	38.0	1.11	M. *Lmo*EORF2316P	Unknown
*Nostoc punctiforme**	41.6	0.84	M. *Npu*ORFC230P	RCCGGY
*Ralstonia solanacearum**	66.6	1.2	M. *Rso*ORF3438P	Unknown
Sinorhizobium meliloti (AE006469)	62.6	1.29	M. *Sme*ORF3763P	Unknown
Streptococcus pneumoniae (AE000514)	39.6	0.69(- -)	M. *Spn*ORF1336P	Unknown
Streptococcus pyogenes M1 (AE009949)	38.5	0.71(-)	M. *Spy*ORF1077P	Unknown
Ureaplasma urealyticum (AF222894)	25.4	0.88	M. *Uur*ORF528P	Unknown
Vibrio cholerae (AE003852)	47.5	1.04	M. *Vch*AORF198P	Unknown
Xylella fastidiosa (NC_002488)	52.6	1.01	M. *Xfa*ORF1774P	Unknown

Note. See Table 1, *Note*.

Table 3. The recognition sites of the C5 methyltransferases containing a methylated CpG dinucleotide

Bacteria	GC content (%)	ρ_{CpG}	C5 methyltransferase	Recognition site
Escherichia coli O157:H7 EDL933 (AE005174)	50.2	1.12	M. *Eco*O157ORF2389P	CGATCG
*Haemophilus parainfluenzae**	39.2	1.06	M. *Hpa*II	C^mCGG
Helicobacter pylori 26695 (AE0005II)	38.8	0.93	M. *Hpy*AVIII	G^mCGC
Helicobacter pylori J99 (AE001439)	39	0.94	M. *Hpy*99XI	A^mCGT
Mycoplasma pulmonis (NC_002771)	26.6	0.28 (- - -)	M. *Mpu*CORF430P	^mCG
Synechocystis sp. PCC 6803 (AB001339)	47.6	0.75(-)	M. *Ssp*6803I	CGATCG
*Xanthomonas oryzae**	62.4	1.13	M. *Xor*II	CGATCG

Note. See Table 1, *Note*.

Table 4. Relative abundance of CpG dinucleotide in bacteria devoid of C5 methyltransferases

Bacteria	GC content (%)	ρ_{CpG}
Brucella melitensis (AE008917)	57	1.20
Buchnera aphidicola AP (NC_002528)	26.2	0.87
Campylobacter jejuni (NC_002163)	30.5	0.62 (- -)
Chlamydia muridarum (AE002160)	40.3	0.75 (-)
Chlamydia trachomatis (AE001273)	41.2	0.79
Chlamydophila pneumoniae (AE002161)	40.5	0.73 (-)
Clostridium perfringens (BA000016)	29.4	0.21 (- - -)
Fusobacterium nucleatum (AE009951)	27	0.16 (- - -)
Lactococcus lactis IL1403 (AE005176)	35.3	0.77 (-)
Listeria innocua (NC_003212)	37.3	1.11
Mycoplasma genitalium (NC_000908)	31.6	0.39 (- - -)
Mycobacterium leprae (NC_002677)	57.7	1.12
Mycoplasma pneumoniae (NC_000912)	39.9	0.82
Mycobacterium tuberculosis (NC_002755)	65.5	1.18
Pasteurella multocida (AE004439)	40.3	1.07
Rickettsia conorii (NC_003103)	32.4	1.03
Rickettsia prowazekii (NC_000963)	28.9	0.77 (-)
Ralstonia solanacearum (AL646052)	66.8	1.19
Staphylococcus aureus Mu50 (BA000017)	32.7	0.94
Treponema pallidum (AE000520)	52.7	1.08

Note. The relative abundance of CpG (ρ_{CpG}) at a level of significantly low, very low, or extremely low is labeled with (-), (- -), or (- - -), respectively.

lase to avoid chromosome degradation by the nuclease. These bacteria thus suffer a selective pressure for the avoidance of restriction sites (Rocha et al. 2001). Since most of the underrepresented sites are not recognition sites for the known RM systems of a given bacterium, the avoidance on these sites indi-

cates the impact of RM systems in bacteria's evolutionary history (Rocha et al. 2001). Therefore the current status of DNA methylation does not allow investigating the avoidance of the sites that may have been methylated in the past due to RM systems that were lost. Because free-living bacteria can often contact with other bacteria living in the surrounding environment, they can easily obtain a new RM system through horizontal transfer. Obligatory intracellular parasites and symbionts cannot do so due to their occlusive living environment. Such bacteria are currently devoid of such systems, and are generally thought to lack horizontal transfer. Thus, one may suppose that they have not been in contact with such systems for a large period of their recent evolution. We therefore made a comparative analysis of obligatory intracellular bacteria with the free-living bacteria holding at least one RM system. We observed that only two free-living bacterial species, *Streptococcus pneumoniae* and *Streptococcus pyogenes*, are CpG deficient. In contrast, 6 of 12 intracellular pathogens or symbionts show CpG deficiency. Thus, CpG dinucleotides are more significantly depleted in intracellular pathogens or symbionts than in proteobacteria (χ^2 test, $p < 0.01$). This is the opposite of what was expected under the cytosine deamination theory via the spread of RM systems.

Lack of Association Between Cytosine Methylation and CpG Deficiency

Among the 34 recognition sites identified in bacterial genomes (Tables 1 and 3), only seven methylated CpG dinucleotides were found within the recognition sites. Therefore, cytosine methylation in bacteria is not generally associated with CpG dinucleotide methylation.

Surprisingly, we find CpG deficiency in eight bacterial species (*Campylobacter jejuni*, *Chlamydia muridarum*, *Chlamydophila pneumoniae*, *Clostridium perfringens*, *Fusobacterium nucleatum*, *Lactococcus lactis* IL1403, *Mycoplasma genitalium*, and *Rickettsia prowazekii*) that are devoid of C5 methyltransferase (Table 4), and this is in contrast to five species (*Clostridium acetobutylicum*, *Mycoplasma pulmonis*, *S. pneumoniae*, *S. pyogenes*, and *Synechocystis* sp. 6803) that contain C5 methyltransferase but are significantly CpG deficient (Tables 1, 2, and 3). This suggests that CpG dinucleotide deficiency is more frequent in bacteria lacking cytosine methylation (χ^2 test, $p < 0.01$). We cannot exclude, however, that this is due to a genome sampling effect since genome programs did not select the bacteria of interest in a random way.

Finally, a *t*-test shows that the CpG relative abundances in bacteria containing RM systems methylating CpG dinucleotides (Table 3) are not significantly lower than those of other bacteria (Tables 1 and 4; $p > 0.1$), indicating that the presence of methylated CpG dinucleotides in recognition sites does not give rise to CpG deficiency.

Table 5. Relative abundances of NpG and CpN in 13 bacterial genomes that show CpG deficiency

Bacteria	CpG	TpG	ApG	GpG	CpA	CpT	CpC
C. acetobutylicum	0.45	1.02	1.13	1.22	1.02	1.13	1.21
C. jejuni	0.62	1.03	1.09	1.11	1.03	1.09	1.1
C. muridarum	0.75	0.96	1.14	1.09	0.97	1.15	1.07
C. perfringens	0.21	0.94	1.24	1.34	0.93	1.23	1.29
C. pneumoniae	0.73	0.96	1.19	1.05	0.96	1.18	1.06
F. nucleatum	0.16	1.05	1.18	1.27	1.03	1.17	1.27
L. lactis IL1403	0.77	1.13	0.96	1.05	1.13	0.97	1.05
M. genitalium	0.39	1.17	1.06	1.12	1.15	1.06	1.15
M. pulmonis	0.28	1.12	1.12	1.04	1.11	1.13	1.07
R. prowazekii	0.77	1.02	1.06	1.03	1.03	1.06	1.03
S. pneumoniae	0.69	1.11	1.07	1.03	1.09	1.09	1.03
S. pyogenes	0.71	1.12	1.04	1.03	1.11	1.05	1.04
Synechocystis sp. 6803	0.75	1.05	0.85	1.36	1.05	0.85	1.36

The above analyses do not support the idea that cytosine methylation is responsible for CpG deficiency. Therefore, we have performed a set of analyses to further explore potential reasons behind CpG deficiency in bacteria.

Associations Between CpG Deficiency and Other Dinucleotide Biases

According to the cytosine methylation hypothesis, CpG dinucleotide is depleted through deamination of methylated cytosines, leading to the concurrent increase in relative abundances of TpG and CpA. In our present study we found that the relative abundances of both TpG and CpA are not significantly higher than that of ApG, GpG, CpT, and CpC ($p > 0.1$, *t*-test) among the bacterial species that show CpG deficiency (Table 5). In *Chlamydiae* and *Clostridia*, the relative abundances of TpG and CpA are lower than that of ApG, GpG, CpT, and CpC. The reasons for this are presently unknown. CpG relative abundance of the bacterial species showing CpG deficiency was plotted against TpG and CpA relative abundances (Fig. 1). The regression of TpG on CpG (Fig. 1A) results in a nearly horizontal line ($R^2 = 0.0002$, slope $= -0.005$, $p < 0.001$), indicating that the change in CpG relative abundance is not correlated with that of TpG relative abundance. In sharp contrast, a negative correlation of the two values was found in the human genome (addressed below). The regression of CpA on CpG (Fig. 1B) also results in a nearly horizontal line ($R^2 = 0.006$, slope $= 0.023$, $p < 0.001$). These findings indicate that CpG variation is not significantly negatively correlated with TpG or CpA abundances. As such, it

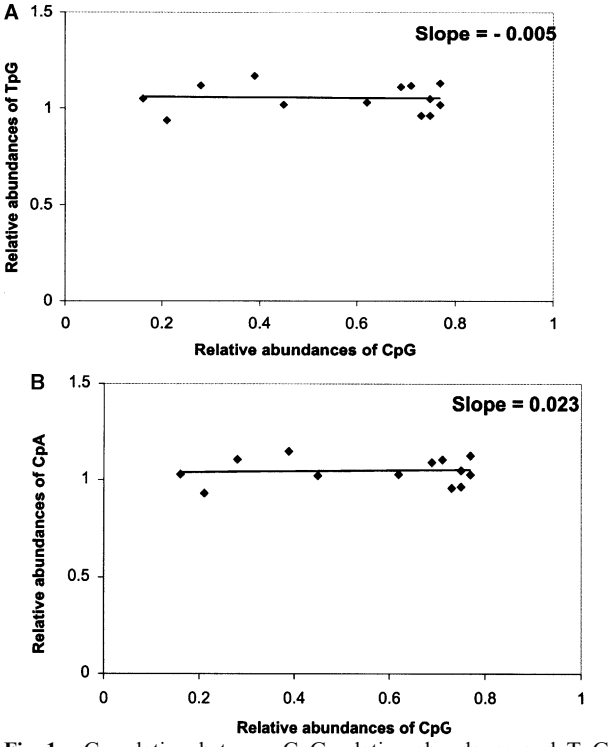

Fig. 1. Correlation between CpG relative abundance and TpG (**A**) and CpA (**B**) relative abundances in the bacterial genomes that show CpG deficiency.

seems unlikely that CpG variation in bacteria can be attributed to different rates of methylated cytosine deamination.

Analysis of Covariation Between CpG Relative Abundance and GC Content

It has been pointed out that the negative correlation between CpG and TpG in different GC contents is an artifact ascribed to deamination of methylated cytosine in the human genome (Duret and Galtier 2000). In order to further test the hypothetical relationship between cytosine methylation and CpG deficiency in bacteria, we analyzed the covariation among dinucleotides CpG, TpG, and CpA under different contents.

In the bacteria studied here, CpG relative abundance is found to be higher in the DNA sequences with a high GC content. No bacterial species showing overall CpG deficiency has more than a 50% GC content (Tables 1–4). We then analyzed the correlation between CpG relative abundance and GC content at the intragenome level. The GC content within a genome is not uniform, so we might expect CpG relative abundances in different genomic regions to correlate with the GC content. Because a bacterial genome is largely composed of CDSs, the effect of codon usage bias on CpG dinucleotide must not be ignored. For example, a study in plants showed that

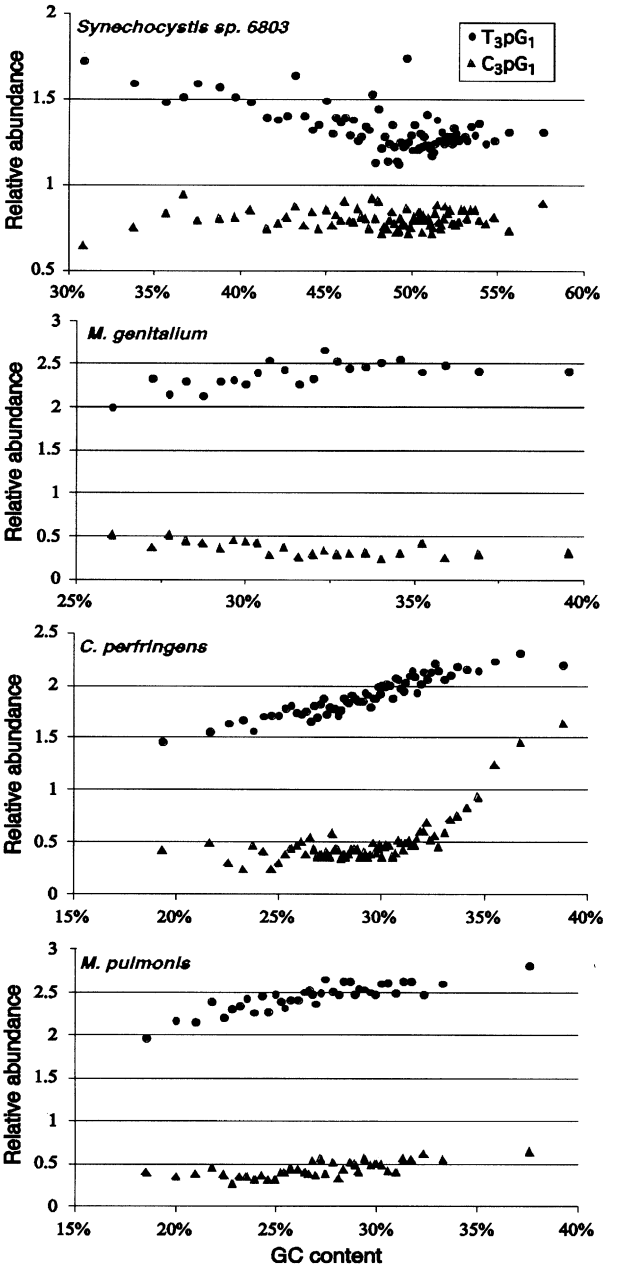

Fig. 2. C_3pG_1 and T_3pG_1 relative abundances plotted against GC contents of long coding sequences. C_3pG_1 (C in the third position of a codon, G in the first position of the following codon); T_3pG_1 (T in the third position of a codon, G in the first position of the following codon). The dots in these patterns represent the data from the long coding sequences. The long coding sequences were generated by listing all CDSs according to their GC contents and then integrating every 40 CDSs.

the negative correlation between C_3pG_1 and T_3pG_1 relative abundances was significant (De Amicis and Marchetti 2000). This was considered to be a consequence of heavy DNA methylation in plants. Therefore, we compared the relative abundance of the neutral dinucleotide sites, C_3pG_1 and T_3pG_1, in a CDS.

We then analyzed the CDSs of the 13 bacterial species showing CpG deficiency for the covariation of

Table 6. Slope values of the changing pattern of C_3pG_1 and T_3pG_1 relative abundances plotted against GC contents of long coding sequences

Organism	C_3pG_1	T_3pG_1
C. acetobutylicum	0.97	3.57
C. jejuni	2.79	2.82
C. muridarum	1.79	3.37
C. perfringens	4.70	4.87
C. pneumoniae	1.02	1.46
F. nucleatum	0.69	5.76
L. lactis IL1403	3.12	5.66
M. genitalium	−0.96	0.71
M. pulmonis	3.27	1.62
R. prowazekii	0.53	3.82
S. pneumoniae	0.60	2.11
S. pyogenes	0.45	3.25
Synechocystis sp. 6803	0.06	−1.67

Note. Description of C_3pG_1 and T_3pG_1 is as in Fig. 2.

dinucleotide relative abundance with GC content. The relative abundances of C_3pG_1 and T_3pG_1 were plotted against the GC content of all the CDSs. The results for *C. perfringens* and *M. pulmonis* are shown in Fig. 2, indicating that C_3pG_1 relative abundance increases somewhat in parallel with T_3pG_1 relative abundances in different GC contents. In comparison, the relative abundance of C_3pG_1 is negatively correlated with that of T_3pG_1 in *Homo sapiens* (Duret and Galtier 2000). This distinctive correlation pattern in humans probably results from methylated cytosine deamination.

The results of the regressions of C_3pG_1 and T_3pG_1 relative abundances in function of GC content are listed in Table 6. A positive slope value means a positive correlation between GC content and dinucleotide relative abundance. Except for two cases, the slopes are positive in all the bacterial species. If the two slope values of a given species in Table 6 are positive, the relative abundances of C_3pG_1 and T_3pG_1 increase with the GC content. This seems to be a general trend with only two exceptional species, *M. genitalium* and *Synechocystis* sp. PCC 6803 (Fig. 2). The negative slope of C_3pG_1 relative abundance in *M. genitalium* is small and we do not know at present how to explain it. With a positive C_3pG_1 slope value and a negative T_3pG_1 slope value, *Synechocystis* sp. PCC 6803 has a trend that is similar to *H. sapiens* except that the relative abundance of C_3pG_1 remains quite constant as the GC content increases (Fig. 2). The explanation to this exception probably lies in the relatively higher GC content (47.6%) and larger genome size (3.6 Mb) of *Synechocystis* sp. PCC 6803. The above results are in agreement with the rule that CpG deficiency is related to lower GC content but do not support the prediction of the cytosine methylation hypothesis.

Discussion

Evaluation of the Potential Effects of RM Systems on CpG Deficiency in Bacteria

In vertebrates, it is widely accepted that CpG deficiency is a consequence of CpG methylation (Bird 1980; Jeltsch 2002). The DNA methylation pattern on CpG dinucleotides is largely maintained by DNA methyltransferase1 (Dnmt1) (Lyko et al. 1999). Some essential differences in the properties of DNA methyltransferases in vertebrates and bacteria may explain the observed differences in CpG deficiency. First, bacteria vary widely in both the content and the size of their C5 methyltransferase recognition sites. Most of the recognition sites do not contain a methylated CpG dinucleotide, suggesting that cytosine methylation is not a determinant of CpG deficiency in bacteria. Although some RM systems have a methylated CpG dinucleotide, the large size of these recognition sites determines that most CpG dinucleotides are not methylated because of the low occurrence of these sites in the genome (i.e., CpG methylation mediated by a single methyltransferase in a rare site such as CGATCG is too weak to induce CpG deficiency).

Second, the DNA methylation in bacteria is a kind of *de novo* methylation (Bestor 1990). This is different from that in vertebrates because Dnmt1 can only function on hemimethylated DNA (Lyko et al. 1999). *De novo* methylation mediated by Dnmt3a and Dnmt3b indeed occurs in vertebrates, but it is restricted in very early embryonic stage (Ramsahoye et al. 2000; Gowher and Jeltsch 2001). These differences between bacterial C5 methyltransferases and those of vertebrates reinforce the idea that C5 methylation is not the major source of CpG deficiency in bacteria. It is possible that a more fundamental mechanism is affecting dinucleotide relative abundance and distribution in bacterial genomes, rather than cytosine methylation.

Third, RM systems are frequently gained and lost by horizontal transfer (Kobayashi 2001). As such, the presence of C5 methyltransferase is intermittent, and possibly rare, which necessarily implicates a much lower bias than methylated cytosine deamination that in genomes containing C5 methyltransferase in permanence, such as in humans. Most free-living bacteria are not CpG deficient compared to pathogen/symbionts. Therefore, the contribution of RM systems to CpG deficiency in bacteria appears suspicious in analysis involving either current or historic parameters. Interestingly, it was reported that free-living pathogens had a significantly higher GC content than intracellular pathogens and symbionts (Rocha and Danchin 2002). Here we show that CpG deficiency correlates with GC content and lifestyle.

Association of GC Content and CpG Deficiency

In this study we find that C_3pG_1 relative abundance and GC content are generally positively correlated in those bacterial species that show CpG deficiency. We obtained qualitatively similar correlations using C_1pG_2 and C_2pG_3 in this analysis (results not shown). This strengthens the link between CpG dinucleotide relative abundance and GC content in bacteria. Identical correlations have been found in humans (Aissani and Bernardi 1991; Pesole et al. 1997) and RNA viruses (Rima and McFerran 1997). It was subsequently pointed out that this could be a mathematical artifact caused by the high mutation rate on methylated CpG dinucleotide (Duret and Galtier 2000). As methylated CpG deaminates to TpG or CpA dinucleotides, the number of C and G decreases in this process. This would lead to a lower expected number of CpG dinucleotides in the new sequence compared to the original sequence. This effect is found to be more evident when the GC content increases (Duret and Galtier 2000). However, the mutation process from methylated CpG to TpG dinucleotide is not present in most of the bacteria that show CpG deficiency. This is implied by parallel changing patterns of CpG and TpG in different GC contents in bacteria. As a result, Duret and Galtier's artifact hypothesis does not explain satisfactorily the association of GC content and CpG deficiency in the bacterial context.

CpG Deficiency in Vertebrates May Be the Cost of a Newly Developed Function of DNA Methylation

Two functions have been suggested for DNA methylation. A primary function is to defend a genome against the invasion of bacteriophages or transposon elements, and a secondary function, a new-developed function in evolution history, is connected with the regulation of gene expression (Yoder et al. 1997). We classify the organisms having DNA methylation into two groups according to the different functions: the first group includes bacteria, fungi, and invertebrates; and the second group includes vertebrates and plants. Only in the second group, CpG dinucleotides are massively methylated or demethylated in order to regulate gene expression activity. In conclusion, only the DNA methylation playing the secondary function in vertebrates and plants can be persuasively linked to CpG deficiency.

Actually the above boundary, within the animal kingdom, should be moved forward to the sea urchin, the only invertebrate species in which Dnmt1-like methyltransferase was identified (Aniello et al. 1996, 2003). As such, it should be distinguished from the other invertebrates. Dnmt1 is critical in playing the secondary function (Ramsahoye et al. 2000), so the presence of Dnmt1-like protein in sea urchin is probably a strong requirement of developmental regulation. Therefore, the evolution of methyltransferase genes from bacteria to human reflects the requirement of functions specialized in more complex organisms, making DNA methylation evolve from a protection mechanism to an epigenetics mechanism. This enables an organism to have an increased life span and to survive under more complex environmental conditions. This benefit comes at a cost. For one, vertebrate genomes confront a huge mutation pressure on the recognition sites for DNA methylation. Until now, no study has shown that vertebrates have found a strategy to compensate for the depleted CpG dinucleotides. Theoretically, continued CpG depletion will lead to a vertebrate genome crisis.

Conclusion

We studied the link between C5 methylation and CpG content in bacteria and found no significant correlation. Thus, C5 methylation is probably not the major factor inducing CpG deficiency in bacteria and more effort should be invested in looking for alternative explanations for this phenomenon. Finally, this study indicates that CpG dinucleotide deficiency is related to GC content. This can be taken as a clue in the search for factors that induce CpG deficiency in bacteria.

Acknowledgments. We would like to thank X.H. Xia and K.Y. Yuen for their interest in the early phases of this work. Special thanks are given to two anonymous reviewers for their critical reading of the manuscript. This work was supported by the BIOSUPPORT programme and a RGC grant from the Hong Kong government.

References

Aissani B, Bernardi G (1991) CpG islands, genes and isochores in the genomes of vertebrates. Gene 106:185–195

Aniello F, Locascio A, Fucci L, Geraci G (1996) Isolation of cDNA clones encoding DNA methyltransferase of sea urchin *P. lividus:* Expression during embryonic development. Gene 178:57–91

Aniello F, Villano G, Corrado M, Locascio A, Russo MT, D'Aniello S, Franscone M, Fucci L, Branno M (2003) Structural organization of the sea urchin DNA (cytosine-5)-methyltransferase gene and characterization of five alternative spliced transcripts. J Gene 302:1–9

Arber W, Linn S (1969) DNA modification and restriction. Annu Rev Biochem 38:467–500

Bestor TH (1990) DNA methylation: Evolution of a bacterial immune function into a regulator to gene expression and genome structure in higher eukaryotes. Phil Trans R Soc Lond B 326:179–187

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1358–1362

Cardon LR, Burge C, Claytion DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA 91:3799–3803

Coulonder C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775–780

DeAmicis F, Marchetti S (2000) Intercodon dinucleotides affect codon choice in plant genes. Nucleic Acids Res 28:3339–3345

Duret L, Galtier N (2000) The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. Mol Biol Evol 17:1620–1625

El-Maarri O, Olek A, Balaban B (1998) Methylation levels at selected CpG sites in the factor VIII and FGFR3 genes, in mature female and male germ cells: Implications for male-driven evolution. Am J Hum Gene 63:1001–1008

Goto M, Washio T, Tomita M (2000) Causal analysis of CpG suppression in the Mycoplasma genome. Microbial Comp Genomics 5:51–58

Gowher H, Jeltsch A (2001) Enzymatic properties of recombinanat Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG sites. J Mol Biol 309:1201–1208

Jeltsch A (2002) Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. Chembiochem 3:274–293

Josse J, Kaiser AD, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid. J Biol Chem 236:864–871

Karlin S, Doerfler W, Cardon LR (1994a) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68:2889–2897

Karlin S, Ladunga I, Blaisdell BE (1994b) Heterogeneity of genomes: measures and values. Proc Natl Acad Sci USA 91:12837–12841

Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185–225

Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29:3742–3756

Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. Genome Biol 3:1–14

Lyko F, Ramsahoye BH, Kashevsky H, Tudor M, Mastrangelo MA, Orr-Weaver TL, Jaenisch R (1999) Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in Drosophila. Nat Genet 23:363–366

McCulloch V, Seidel-Rogol LB, Shadel GS (2002) A human mitochondrial transcription factor is related to RNA adenine methyltransferases and binds S-adenosylmethionine. Mol Cell Biol 22:1116–1125

Pesole G, Luini S, Grille G, Saccone C (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene 205:95–102

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R (2000) Non-CpG methylation is prevalent in embryonic stem cell and may be mediated by DNA methyltransferse 3a. Proc Natl Acad Sci USA 97:5237–5242

Rima BK, McFerran NV (1997) Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Virol 78:2859–2870

Roberts RJ, Macelis D (2001) REBASE—Restriction enzymes and methylases. Nucleic Acids Res 29:268–269

Rocha EPC, Danchin A (2002) Base composition bias in genomes might result from competition for scare metabolic resources. TIG 18:291–294

Rocha EPC, Danchin A, Viari A (2001) Evolution role of restriction/modification systems as revealed by comparative genome analysis. Genome Res 11:946–958

Subak-Sharpe H, Burk RR, Crawford LV (1966) An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. Cold Spring Harbor Symp Quant Biol 31:737–748

Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic synthesis of deoxyribonucleic acid. J Biol Chem 237:1961–1967

Wilson GG (1988) Type II restriction-modification systems. TIG 4:314–318

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. TIG 13:335–340