
©1996-2009 All Rights Reserved. *Online Journal of Bioinformatics*. You may not store these pages in any form except for your own personal use. All other usage or distribution is illegal under international copyright treaties. Permission to use any of these pages in any other way besides the before mentioned must be gained in writing from the publisher. This article is exclusively copyrighted in its entirety to OJB publications. This article may be copied once but may not be, reproduced or re-transmitted without the express permission of the editors. [This journal satisfies the refereeing requirements \(DEST\) for the Higher Education Research Data Collection \(Australia\).](#) Linking: To link to this page or any pages linking to this page you must link directly to this page only here rather than put up your own page.

OJB™

Online Journal of Bioinformatics®

Volume 10 (1):165-179, 2009

Comparative genomic study on context-dependence of CpG mutations: Acceleration effect of 5' T nucleotides and new evidence of strand asymmetry in genes

Wang Y, Leung FC

School of Biological Sciences and Genome Research Centre, University of Hong Kong, Pokfulam, Hong Kong

ABSTRACT

Wang Y, Leung FC., Comparative genomic study on context-dependence of CpG mutations: Acceleration effect of 5' T nucleotides and new evidence of strand asymmetry in genes, Online J Bioinformatics 10 (2):165-179, 2009. Previous studies have reported context-dependence of CpG mutations. In this study, we demonstrate the effect of four CpG flanking nucleotides through comparative genomic analyses. We obtained orthologous genes of *C. elegans*, fruitfly, sea squirt, zebrafish and human. Analyses on two 5' flanking positions reveal that T at -2 position can affect T/A to G/C changes at -1 position more significantly than the other nucleotides. As a co-effect of the T/A to G/C changes and CpG mutations, TTCG motif is significantly lower than AACG motif in the zebrafish and human genes. We then studied observed/expected values of dicodons that have a central CpG. The value of TTC•GAA is lowest except in *C. elegans*, supporting again the context-dependent effect in genes. In addition, we calculated substitutional rates of CpG and four flanking sites. The rate of G is much lower than that of C, and even those of two most adjacent flanking positions for all the species. Mutational rate of CpG sites is facilitated by 5' flanking T nucleotides, and substitutions on CpG sites in genes are more frequently observed on sense strand.

Key words: Comparative genomics; context-dependent mutation; CpG deficiency; strand asymmetry.

INTRODUCTION

In the human genome, CpG sites are notable mutation hotspots and, correspondingly, CpG frequency is merely 20% of the expected frequency. Evidence shows that point mutations on CpG dinucleotides (CG→TG or CG→CA) are responsible for about 25% of the pathological disorders and 33% of the genetic diseases in humans (Cooper and Youssoufian 1988). This is largely ascribed to DNA methylation (Bestor 1990; Bird 1980), which is involved in gene expression regulation and proviral DNA suppression in vertebrates (Harbers et al. 1981; Hu et al. 1984; Robertson and Jones 2000). The DNA methylation is used for epigenetic modification of genomic structure, playing an important role in embryonic development and tissue-specific expression of genes (Baylin 2000; Okano et al. 1999). This is strongly supported by recent reports showing that DNA methylation pattern is stable in the same tissue from different human sex and age groups, whereas the same genomic regions are differentially methylated for different tissues (Eckhardt et al. 2006; Oakes et al. 2007).

The side effect of DNA methylation is the loss of CpG sites in vertebrate genomes due to the high mutational rate of methylated CpG sites. The mutations of CpG sites were shown to be coupled with an enigmatic context-dependent effect (Fryxell and Moon 2005; Zhang and Zhao 2004). Generally, TA-rich flanking sequences will impose higher mutational pressure on CpG sites. A report on CpG mutations in the human p53 genes shows that methylated TCGA and TCGG sites exhibit higher mutational rates than GCGG (Ollila et al. 1996). In the mouse lacZ gene, a 5' pyrimidine of methylated CpG sites can obviously improve the deamination rate (Ikehata et al. 2000). What instigates the context-dependent mutations? All the CpG sites are the targets for DNA methylation (Ikehata et al. 2000; You et al. 1998) and the preference for flanking nucleotides of target sites has not been reported yet. Since the context-dependence is not likely a result of site specification in the process of methylation, alternative mechanisms associated with the effect of context dependence are therefore inferred to be involved in facilitating mutations on methylated CpG sites.

Insects were the lowest invertebrates in which DNA methylation was discovered (Lyko et al. 2000; Wang et al. 2006), whereas their methylation level is very low and the contribution to CpG loss has not been assessed yet (Field et al. 2004). In sea urchins and sea squirts, fractional DNA methylation (less than 50%) was noticed by studying the methylation level at the boundary of invertebrates and vertebrates (Tweedie et al. 1997), and a later study found that the methylation level on genes was higher than in other genomic regions (Simmen et al. 1999). Maintenance of global CpG methylation in vertebrates accelerates the loss of CpG sites from a genome. The mutational rate of methylated CpG sites even varies among vertebrates in function of body temperature. Warm-blooded vertebrates are losing CpG sites 20.6-fold faster than cold-blooded ones because of positive correlation between body temperature and deamination rate of methylated CpGs (Fryxell and Zuckerkandl 2000).

Due to context-dependent mutagenicity, CpG sites are supposed to show a strong compositional bias for flanking nucleotides. The bias may be displayed by comparative genomics analysis, although previous statistical methods have been employed to show the effect using single nucleotide polymorphisms (SNPs) and mutation database of human genes (Fryxell and Moon 2005; Krawczak et al. 1998; Zhang and Zhao 2004). Strikingly, nucleotide compositional bias at 500 flanking positions of SNPs was evaluated by Zhang and Zhao (2004). Present knowledge of the influence of flanking nucleotides on mutagenicity of methylated CpGs is limited to the two

most adjacent flanking positions (± 1) of CpGs in genic regions. Our previous study indicates positive selection of G/C at ± 1 positions of CpGs in fish genes, compared to sea squirt orthologs (Wang and Leung 2008). This study extended the study on the context-dependence to ± 2 flanking positions of CpG sites. We obtained orthologous genes from *Caenorhabditis elegans*, fruitfly (*Drosophila melanogaster*), sea squirt (*Ciona intestinalis*), zebrafish (*Danio rerio*) and human (*Homo sapiens*). The species represent those lacking DNA methylation and showing slight, fractional and global DNA methylation respectively. Zebrafish and human were selected to represent cold-blooded and warm-blooded vertebrates that show different deamination rates of methylated CpG sites. Using homologous fragments between orthologous genes, we showed the important role of T at -2 position in affecting T/A to G/C changes at -1 position and significantly lower frequency of TTCCG relative to AACG motif. Evidence also came from dicodon usage bias and high substitution rates of CpG and its flanking sites. Finally, we showed strand bias of the substitutions on CpG sites. The importance is that this was also found in *C. elegans* and fruitfly orthologs.

MATERIALS AND METHODS

Orthologous gene collection:

We obtained tables for pairwise orthologous genes from the BIOMART database (<http://www.biomart.org>) in which Ensembl 45 Homology database was used. The tables showed IDs of the orthologous genes between zebrafish (*D. rerio*; ZFISH 6) and the following species: human (*H. sapiens*; NCBI36), sea squirt (*C. intestinalis*; JGI2), fruitfly (*D. melanogaster*; BDGP4.3) and *C. elegans* (WB170). To remove redundancy, we only kept the first orthologue in case of multiple orthologues in another species. We compact the tables into one by matching the zebrafish IDs between the tables. There were 5798 groups of orthologous genes for the five species. Using the IDs, we downloaded coding sequences (CDS) from the EMBL database (<http://www.ensembl.org/>).

Collection of homologous fragments in orthologous genes:

We first extracted homologous fragments from CDSs of the orthologous genes. Pairwise alignments were performed on orthologous genes for species pairs of the five species. The alignment starts from finding an identical sequence seed of 5 bp on both sequences. Homologous fragments were obtained from extension of the seed at both ends. The extension terminated while continual two mismatches were found. Homologous fragments longer than 50 bp were taken into a dataset. The alignment might be resumed at a new site until the searching reached the end of CDSs. Therefore, we sometimes could collect more than one homologous fragment in one CDS.

Investigating CpG sites and the flanking nucleotides:

We located CpG sites within the homologous fragments. CpG might present in both of the homologous fragments or just one of the fragments. We named the first case as conserved CpG site and the second as substituted CpG site. We recorded the dinucleotides at the flanking positions of both conserved and substituted CpG sites (two at 5' end, and another two at 3' end) and computed the nucleotide frequencies to assess their influence on CpG substitutional rate. On the other hand, we also documented nucleotide changes at the flanking positions of conserved CpG sites to uncover mutational bias of G/C for protection of CpG sites.

Calculation of obs/exp value of a dicodon:

We used the CDSs of the above orthologous genes to compute dicodon obs/exp value that has also been described as relative abundance of a dicodon (Karlin and Mrazek 1996). It was calculated as per the formula: $(\sum L * \sum N_{\alpha \bullet \beta}) / (\sum N_{\alpha} * \sum N_{\beta})$, where $\sum L$ is the number of the codons (start and stop codons were not accounted) in all the orthologous genes, $\sum N_{\alpha \bullet \beta}$ is the count of dicodon $\alpha \bullet \beta$, and $\sum N_{\alpha}$ is the count of codon α . Deviation of obs/exp value from 1 denotes under or over-representation of a dicodon. We calculated obs/exp values for dicodons TTC•GAA and AAC•GTT. To compare the usage of synonymous codon before GAA and GTT codons, we also obtained obs/exp values for dicodons TTT•GAA and AAT•GTT. The second codon was then converted to GGG and GCC. The obs/exp values of these dicodons were used to assess the influence of 3' GC content of CpG sites on the usage of the synonymous codons. Therefore, we calculated obs/exp values for 16 dicodons. Because CCC, CCT, GGC and GGT are four-fold degenerate, we could not use to them assess the influence of 5' GC content.

Estimate of substitutional rate:

We calculated ratio of transitions to transversions (Ts/Tv) on CpG sites using the homologous fragments. In this test, we intended to evaluate the influence of CpG flanking nucleotides on the Ts/Tv ratio. In case that the nucleotide at the closest flanking position of CpG sites is identical between the fragments, we recorded substitutions on C or G, as well as the flanking nucleotide. We then estimated substitutional rates of CpG and its four flanking sites using Kimura's two-parameter method (Nei 1987). Pairs of hexanucleotides that have a central CpG sites in at least one of the homologous fragments were collected. They were used to count the numbers of identical, transitional and transversional nucleotide pairs at the six positions separately. Then, transition and transversion frequency denoted by P and Q respectively were calculated for individual positions. Thus, we had $R+P+Q=1$, in which R is the frequency of identical nucleotides. The estimated number of substitutions at a given site is computed as:

$$-\frac{1}{2} \log_e [(1 - 2P - Q) * \sqrt{(1 - 2Q)}]$$

RESULTS**Context-dependent mutations detected by pairwise alignments:**

Orthologous genes from five representative species were used to demonstrate the impact of DNA methylation on CpG sites. We first attempted to display the effect of ± 2 flanking nucleotides on CpG mutations. The sequences used were homologous fragments between orthologous genes. We counted substituted CpG sites (substitutions occurred on C or G in one of the homologous fragments) and conserved CpG sites (Figure 1). The ratio of substituted to conserved CpG sites (s/c ratio) was used to estimate the influence of 5' flanking nucleotides. However, the result did not obviously support the expected effect of some nucleotide at the ± 2 positions. Therefore, the most adjacent flanking nucleotides are more powerful in the context-dependent effect.

Due to the context-dependence, abundant T/A to G/C replacements at -1 flanking position of CpG sites have been reported as a result of positive selection for reduction of mutational pressure on CpG sites (Wang and Leung 2008). We therefore exhibited the influence of ± 2 flanking positions on the frequency of the T/A to G/C changes at the ± 1 flanking positions. Using the homologous fragments between a subjected species and *C. elegans*, conserved CpG sites

from a species pair were located. We specified the nucleotide at -2 position and recorded T/A to G/C and G/C to T/A changes at -1 flanking position of conserved CpG sites. We found that T/A to G/C substitutions were significantly more frequent than those in opposite direction, except for the presence of C at -2 flanking position (Table 1). Note that the significant changes were in direction of G/C to T/A for CNCG pattern in human and the sea squirt. The result again confirms the positive selection of G/C at the -1 flanking position. Moreover, we found that the nucleotides at -2 flanking position may affect the frequency of T/A to G/C changes. T at the -2 position appears to be able to induce more frequent T/A to G/C changes than other nucleotides simply on the basis of the number difference between T/A to G/C changes and those in opposite direction. We created 2x2 contingency tables consisting of the amounts of the changes under different -2 flanking nucleotides. When the amounts under -2 flanking T were compared with others, the difference is always significant in human ($P < 0.05$). This is also true for the fruitfly ($P < 0.05$). The significance level $P < 0.001$ was observed in one table for sea squirt and zebrafish. These results indicate that -2 flanking nucleotides can indirectly affect CpG mutations. Although we showed the effect using 5' flanking nucleotides, similar result was expected for the complementary ones at 3' flanking positions.

Table 1 GC enrichment at -1 flanking position of conserved CpG sites

	TNCG				ANCG				CNCG				GNCG			
	hu	ze	sq	fly	hu	ze	sq	fly	hu	ze	sq	fly	hu	ze	sq	fly
T/A→C/G	144	178	88	109	120	162	70	77	26	80	12	28	148	185	150	51
C/G→T/A	40	46	21	19	58	56	38	26	44	51	28	21	66	61	23	41
P value	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.002	$<10^{-4}$	0.03	0.01	0.01	0.3	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.3

The conserved CpG sites were present in both of the homologous fragments from orthologous genes between *C. elegans* and four other species including human (hu), zebrafish (ze), sea squirt (sq) and fruitfly (fly). The substitutions occurred at -1 flanking position (N) of the conserved CpG sites were surveyed. In particular, substitutions of T/A (T or A) to G/C (G or C) or in opposite direction were counted and chi-square tests were used to determine the significance of the difference in amount between the two substitution directions.

To support the strong inducing effect of 5' flanking Ts on CpG mutations, we focused on the substitutional rate of CpG sites with 5' TT and AA dinucleotides. The similar inducing effect of the -1 flanking T has been revealed by a survey on a human gene (Ollila et al. 1996). In the homologous fragments, we located the CpG sites following TT or AA dinucleotides and counted substituted CpG sites (substitutions occurred on C or G in one of the homologous fragments) and conserved CpG sites (Figure 1). The s/c ratio was used to estimate the effect. We found that the s/c ratios increased notably when the homologous fragments of zebrafish and human were examined against those from other organisms (Table 2), suggesting an abundance of depletions of TTCG and AACG motifs in vertebrates. The difference is that TTCG is associated with a higher s/c ratio than AACG in most of the species pairs, whereas the three ratios for sea squirt are against the rule. We used Fisher's exact test to evaluate significance of the difference between TT and AA in affecting substitutions on CpG sites. The amounts of substituted and conserved CpG sites with 5' TT and AA were used to construct a 2x2 contingency table for the statistical test. Significance was shown in three pairwise comparisons ($P < 0.05$), all of which include *C. elegans* (Table 2). Therefore, 5' TT dinucleotide is more influential in context-dependent mutagenicity. This suggests a similar effect of 3' AA dinucleotide.

Table 2. Substitutions at CpG sites following TT and AA dinucleotides

	fruitfly		sea squirt		zebrafish		human	
	TTCG	AACG	TTCG	AACG	TTCG	AACG	TCG	AACG
<i>C. elegans</i>	0.4	0.42	1.04	1.69	1.99	1.48	3.28	1.96
fruitfly			1.56	1.45	1.82	1.56	2.57	2.43
sea squirt					1.42	1.56	2.38	2.45
zebrafish							2.25	2.18

The ratios in the table are the number of substituted TTCG (or AACG) sites to that of conserved TTCG (or AACG) sites. The datasets used were homologous fragments from pairwise alignment between orthologous genes. The substituted TTCG site means that the TTCG site was found to have substitutions on CpG in one of the homologous fragments; The conserved AACG means that both of the homologous fragments have an AACG at the same place. We used Fisher's exact test to detect the difference between substitutions on TTCG and those on AACG in the homologous fragments. We constituted a 2x2 contingency table with the numbers of conserved TTCG and substituted TTCG, and those of conserved AACG and substituted AACG. Homologous fragments belonging to species pairs *C. elegans*-sea squirt, *C. elegans*-zebrafish, and *C. elegans*-human were found to show significant difference ($P < 0.05$). The ratios for the three species pairs are shown in block format.

```

TGCCGCGCGCTGGTCATCGCGCCCTTTTCGGCATCGCA
TGTCCGTGCACTGGTCATCGCGCCGCTGTTTGGCATCGTA
      ↑   ↑                               ↑
  
```

Figure 1: Alignment example of homologous fragments: The homologous fragments are obtained from a pair of orthologous genes: human ENST00000320230 and zebrafish ENSDART00000081198. The substituted CpG sites are marked with an arrow; the conserved CpG sites are underlined

Evidence from dicodon usage bias:

Evidence was also obtained from synonymous codon usage inferred by observed/expected (obs/exp) values of 16 dicodons containing a central CpG (Table 3). We used the orthologous genes to make the analysis. Overall, obs/exp value of dicodon TTC•GAA was lowest in human and zebrafish, which is in strong contrast to AAC•GTT which had an opposite arrangement of TT and AA. On the other hand, complementary dicodons of AAC•GAA and TTC•GTT show nearly equal obs/exp values. The order of obs/exp values in increasing order is: TTC•GAA < AAC•GAA ≈ TTC•GTT < AAC•GTT. This rule was not found in the invertebrates.

Dicodons with TTT and AAT show higher obs/exp values (Table 3), exhibiting codon usage biases among synonymous codons (TTC and TTT) for Phe and those (AAC and AAT) for Asn before GNN codons in sea squirt, zebrafish and human. This was also shown in the results of *C. elegans* and the fruitfly in spite of the small difference between the obs/exp values. Basically, obs/exp values of dicodons TTC•GGG, AAC•GGG, TTC•GCC and AAC•GCC were higher than those of TTC•GAA, AAC•GAA, TTC•GTT and AAC•GTT, implying the importance of high GC content in preserving CpG sites.

Table 3. Obs/exp values of dicodons showing influence of central embedded CpG on codon usage.

Species	1 st codon	2 nd codon			
		GTT	GAA	GGG	GCC
human	TTC	0.23	0.14	0.56	0.46
	TTT	1.70	1.66	1.66	1.30
	AAC	0.34	0.21	0.67	0.48
	AAT	1.75	1.77	1.46	1.15
zebrafish	TTC	0.33	0.18	0.85	0.68
	TTT	1.57	1.30	1.78	1.38
	AAC	0.48	0.34	0.69	0.67
	AAT	1.68	1.83	1.54	1.34
sea squirt	TTC	0.71	0.48	0.93	0.87
	TTT	1.42	1.20	1.22	1.40
	AAC	0.52	0.63	0.85	0.56
	AAT	1.40	1.43	1.08	0.93
fruitfly	TTC	0.88	0.65	1.10	0.94
	TTT	1.34	1.07	2.16	1.47
	AAC	0.84	1.10	1.24	0.73
	AAT	1.33	0.93	2.22	1.46
<i>C. elegans</i>	TTC	0.55	0.81	0.96	0.73
	TTT	1.26	1.24	1.73	1.42
	AAC	0.65	0.82	1.82	1.23
	AAT	1.29	0.90	2.38	0.75

The CDSs used were from 5798 orthologous genes between the five species. Obs/exp value was calculated as per the formula: $(\sum L * \sum N_{\alpha\beta}) / (\sum N_{\alpha} * \sum N_{\beta})$, where $\sum L$ is the count of codons (start and stop codons were not accounted) of all the CDSs, $\sum N_{\alpha\beta}$ is the total count of dicodon $\alpha\beta$, and $\sum N_{\alpha}$ is the total count of codon α . The obs/exp values <0.82 indicate significant under-representation of a dicodon; the values >1.20 indicate significant over-representation (Karlin and Mrazek 1996).

Strand asymmetric CpG loss due to the context-dependent mutations:

Due to massive positive selection of G/C at CpG flanking position of CpG sites, zebrafish genome preserves more CpG sites than sea squirt genome (Wang and Leung 2008). This was verified using the orthologous genes from five species in large phylogenetic distance. In order to estimate the loss of CpG sites, we used the homologous fragments to count the numbers of substituted and conserved CpG sites. Substitutions on CpG sites were restricted to changes of

CpG to TpG/CpA in this test. The ratio of substituted to conserved CpG sites (*s'/c* ratio) was used as a measure of DNA methylation-induced CpG loss. Because we could not detect the substitutions that resulted in the same pattern in both species, the ratio was an estimate.

The *s'/c* ratios were 0.29 and 0.57 for the fruitfly and *C. elegans* respectively in the fruitfly-*C. elegans* pair (Table 4). This implies that the fruitfly genome has more CpG sites than *C. elegans* because substituted CpG sites of fruitfly in the homologous fragments was counted to be 29% of conserved CpG sites, whilst those of *C. elegans* occupied 57% of the conserved CpG sites. The ratios obtained in the other species pairs are 1.28, 0.97 and 1.17 for the sea squirt, zebrafish and human respectively (Table 4). The ratio of the sea squirt is higher than that of the zebrafish and human, suggesting that the sea squirt genome has undergone more CpG depletions than the zebrafish and human. Considering the nearly equal ratios (0.57-0.58) of *C. elegans* in all the species pairs except human-*C. elegans*, at least the ratios of sea squirt and zebrafish reflect a real difference in CpG loss. The result is consistent with our expectation.

Table 4. Estimate of CpG loss

subjected species	Substitutions in subjected species			Substitutions in <i>C. elegans</i>			Conserved CpG sites
	TpG	CpA	total/conserved CpGs	TpG	CpA	total/conserved CpGs	
human	2193	379	1.17	609	277	0.40	2202
zebrafish	2520	435	0.96	1411	363	0.58	3071
sea squirt	1562	257	1.28	634	205	0.58	1435
fruitfly	1170	310	0.29	2162	729	0.57	5032

The numbers of substituted and conserved CpG sites were counted on homologous fragments from orthologous genes between *C. elegans* and the subjected species. The ratio of total to conserved CpGs, where total is referred to the sum of TpG and CpA substitutions, was applied to estimate CpG loss.

Our result unexpectedly shows strand asymmetry of CpG mutations in genes. CpG→TpG substitutions basically occur over five-fold more frequently than CpG→CpA substitutions for the sea squirt, the zebrafish and human (Table 4), indicating that CpG mutations have a strong strand bias because a CpG→CpA substitution is equivalent to CpG→TpG substitution on the anti-sense strand. The amounts of the two types of substitutions are significantly different in all the species (Chi-square test; $P < 0.0001$).

Substitutional rates of CpG and its flanking sites:

A hypothetical, combined effect of the context-dependence of CpG mutations and positive selection is increased substitutional rates of CpG flanking sites. We then compared the rates of CpG sites and those of the flanking sites. Before the calculation, we studied the transition/transversion (Ts/Tv) ratio at C and G of CpG sites under scenarios of different 5' and 3' flanking nucleotides. A ratio around 0.5 is the equivalent to equal rates of transition and transversion. Homologous fragments between *C. elegans* and the remaining species were used in the test. At 5' side, high Ts/Tv ratios on C were exhibited regardless of the nucleotide at -1 position (Table 5). Particularly, the 5' flanking A was accompanied with extremely high Ts/Tv ratios. Only for the fruitfly, the ratio on C with 5' A is lower than that with 5' T. Statistical analysis showed that the ratios on C with different 5' flanking nucleotides are significantly different (ANOVA, $P = 0.0023$). The ratios on G are generally lower than 0.5 except for those with 5' C, indicating a high proportion of nonsynonymous substitutions of all. Significant difference

was also found among the ratios in different categories (ANOVA, $P=0.005$). We also computed Ts/Tv on C and G of CpG sites when 3' flanking nucleotides were specified. The results suggest that 3' flanking nucleotides do not affect Ts/Tv as strongly as 5' ones. Similarly, Ts/Tv ratios on G with different 3' nucleotides are mostly less than 0.5. The difference here is that there is no significant variance among the ratios on C, indicating independence to 3' flanking nucleotides. Significant difference was not observed for these ratios in different categories of flanking nucleotides (ANOVA, $P>0.1$).

Table 5 Ts/Tv at CpG sites with different 5' neighboring nucleotides

	Ts/Tv of C				Ts/Tv of G			
	5' T	5' A	5' C	5' G	5' T	5' A	5' C	5' G
human	1.14	9.43	0.84	0.88	0.14	0.26	0.54	0.05
zebrafish	1.14	8.78	0.84	1.08	0.19	0.27	0.46	0.13
sea squirt	2.00	4.52	1.34	1.09	0.39	0.18	0.56	0.22
fruitfly	3.93	3.39	2.25	1.93	0.39	0.32	1.12	0.14

The dataset is described in Table 4. The table shows Ts/Tv ratios of C and G in CpG sites. The influence of the 5' nucleotides on the ratios was demonstrated by specification of nucleotide at the -1 position of CpG sites.

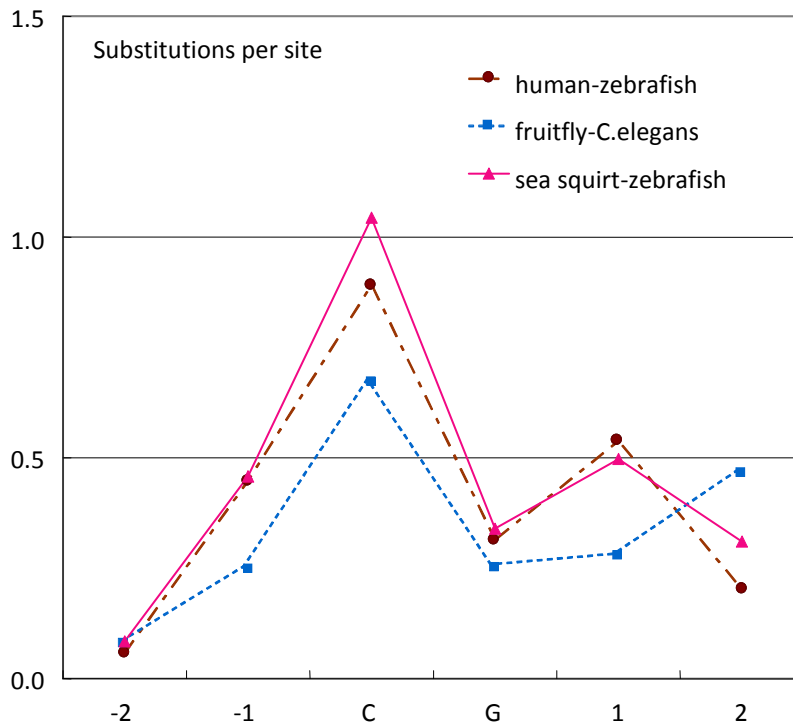


Figure 2. Substitutional rates of CpG site and four flanking positions: The substitutional rate was calculated with Kimura's two-parameter method. Homologous fragments from three species pairs: human-zebrafish, zebrafish and sea squirt and fruitfly-*C. elegans*, were used in the calculation.

Using Kimura's two-parameter method, we estimated substitution rates of the CpG and its four flanking positions. The calculation was carried out for species pairs of *C. elegans*-fruitfly, sea squirt-zebrafish and zebrafish-human. We found that the rate of C was 2.8-3.0 times higher than that of G in the three species pairs (Figure 2), showing again that substitutions at CpG sites in CDSs are strand-asymmetric in both invertebrates and vertebrates. At least two flanking positions (± 1) show higher substitutional rates than corresponding rates of G except -1 for fruitfly-*C.elegans*, indicating elevated substitutional rates at CpG flanking sites. In contrast, the rates at -2 position are much lower than those of the other sites, accounting for less than one-third. Between the species pairs, fruitfly-*C.elegans* shows obviously lower substitutional rates at CpG and ± 1 positions than the other two species pairs as we had expected. Moreover, the substitutional rate of C for sea squirt-zebrafish is 17% higher than that for human-zebrafish, supporting rapid CpG loss at the invertebrate-vertebrate boundary. In both of the species pairs, the rates of ± 1 positions are 44%-60% of those of C sites, but at least 35% higher than those of G sites. The surprising finding in this test is the high substitutional rate of C for *C. elegans* and the fruitfly, which implies that CpG sites are being deficient in the invertebrates as well, at least in some genic regions.

DISCUSSION

Context-dependent and strand-asymmetric CpG mutations:

In this study context-dependence of CpG mutations has been supported by comparative genomic analyses conducted to show subtle differences among orthologous genes after divergence from a common ancestor. Although the genes selected are conservative in a long evolutionary history from worm to human, the context-dependent effect was exhibited by a series of tests using the homologous fragments obtained from pairwise alignment. The result is consistent with those from previous studies on the basis of survey on SNPs (Fryxell and Moon 2005; Zhang and Zhao 2004). Furthermore, we observed GC-skewed substitutions at CpG flanking positions as a result of positive selection, in accord with the result in our previous study (Wang and Leung 2008). A new finding in this study is that T nucleotide at -2 position can strongly drive A/T to G/C substitutions at -1 position of CpG sites. Since most of our results came from analyses on genes, selection is also an essential factor affecting CpG sites in the genes. However, the contribution of negative selection to CpG depletion is probably trivial, compared with mutations induced by DNA methylation in vertebrates.

Our result strongly indicates a strand-asymmetric effect on CpG mutations. This was found not only in vertebrates but also in the fruitfly and worm that do not have CpG mutational hotspots. The strand asymmetry effect was described previously in a few studies (Krawczak et al. 1998; Leader et al. 1995). Inference of the effect was first derived from different relative abundances of CpG dinucleotides in two codon positions {2,3} and {3,1} (the CpG at silent position and the first position of the following codon). The relative abundances were supposed to be equal under the assumption that CpG sites at sense and anti-sense strand mutate at a nearly equal rate. But the result was that the value of CpG at {3,1} is significantly higher than that at {2,3}, suggesting deamination rate of methylated CpG sites is higher on sense strand (Leader et al. 1995). The second proof was obtained from statistical analyses on human gene mutations, showing that transitional rate of CG→CA was estimated to be 1.4-fold higher than that for CG→TG (Krawczak et al. 1998). The two studies reach opposite conclusions and our study solves the conflict by supporting the former. Our study provides a direct demonstration of the strand asymmetry of

CpG mutations by calculation of substitutional rates using homologous fragments from orthologous genes. The substitutional rates of C and G differ by a factor of ~ 3 , suggesting that the mutational rate of methylated CpG sites on the anti-sense strand is much lower.

The strand asymmetry has not been understood up to now. One hypothesis is that there is a probable difference in repair efficiency on the two strands by methylated DNA-binding domain 4 (MBD4), a protein responsible for converting mutations back to normal at methylated CpG sites (Hendrich et al. 1999). However, over-abundant C to T transitions on the sense strand were shown in genes of MBD4-knocked-out mice (Wong et al. 2002). Therefore, involvement of MBD4 is not convincing at present. An alternative hypothesis is that transitions of CpG \rightarrow TpG or CpA perhaps differ in interrupting protein function of the gene involved (Krawczak et al. 1998).

The probability of the involvement of DNA structural constraints :

Up to present, the context-dependent effect has not been understood yet. Since the theory of deamination of methylated CpG sites cannot solely explain all the characteristics of CpG mutation, it is probably a result of co-effect of all related factors. As we know, DNA methylation is not a universal explanation for CpG deficiency because bacteria and viruses showing CpG deficiency do not have a methylation system as that in vertebrates. Alternative hypotheses proposed to explain CpG deficiency at least include DNA structural constraints for bacterial genomes and host immunological stimulation for small viral genomes (Antri et al. 1993; Laura et al. 2006; Wang and Leung 2004).

Basically, two hypotheses for CpG deficiency on the basis of DNA structural constraints have been proposed. For the first one, CpG deficiency was connected to high stacking energy of CpG in the DNA double helix (Breslauer et al. 1986; Hunter 1993); for the second one, CpG deficiency was considered as a result of thermal instability of CpG sites in DNA double helix, particularly when 5' TT and/or 3' AA dinucleotides were present (Antri et al. 1993). The latter hypothesis is more promising as an interpretation to the puzzles due to its relatedness to the context dependence. In reference to previous reports, the DNA structural constraints have been suggested to be the mechanism responsible for the context-dependent CpG deficiency in bacterial genomes (Wang and Leung 2004). The same context-dependent effect discovered in eukaryotic organisms suggests the involvement of the DNA structural constraints in the mutation process.

The employ of the hypothesis in eukaryotes is bold because DNA molecules in eukaryotes differ from those in prokaryotes by being entangled with histone complexes, and consequently most DNA structural constraints are possibly suppressed by chromatin. Indeed, CpG deficiency was not shown in *C. elegans* and insects. After establishment of CpG methylation, the DNA structural constraints seem to be released by some means and methylated TTCGAA and relevant motifs (CpG with flanking 5' T and 3'A nucleotides) are exposed again to be substitution hotspots, conferring a favorable DNA structural modification by intervention of methyl groups. Although the influence of cytosine methylation on DNA structure was found to be context-dependent (Hodges-Garcia and Hagerman 1992), methylated CpG sites could not pronouncedly alter DNA conformation, in spite of small local modifications like DNA flexibility (Derreumaux et al. 2001; Mayer-Jung et al. 1997; Norberg and Vihinen 2001). Whether such a modification is indeed the cause of mutations at CpG sites is still an open question.

Given the fact that DNA structural constraints are suppressed by histones, the influence of the constraint could not be removed completely. During DNA duplication and gene transcription, naked DNA is vulnerable to its essential structural constraints again. This is the possible interpretation for our observation of somewhat context-dependent CpG suppression in *C. elegans* and the fruitfly. For example, dicodons of TTC•GAA and TTC•GTT are under-represented, and substitutional rates of CpG and its flanking sites are surprisingly high in *C. elegans* and the fruitfly. We could not exclude the probability that all this was driven by the DNA structural constraints.

Learning more about DNA structural constraints will help us to understand intrinsic nature of DNA methylation in vertebrates. We suggest that the DNA structural constraints are the fundamental reason for CpG deficiency, and CpG methylation as a type of DNA structural modification in vertebrates simply accelerates the mutation process. To uncover biological processes leading to attenuate the intrinsic DNA structural fault is the task in priority for future studies. Molecular biology experiments and physiochemical calculations are expected to cast lights on the mutation process at methylated CpG sites in varied conditions. All this is of great importance for genetic, pathological, and molecular biological researches.

GC3 level and codon usage bias:

Codon usage bias is an unequal usage of synonymous codons. The current most accepted hypothesis for codon usage bias is the selection-mutation-drift (SMD) theory that states that codon usage pattern arises as a result of a balance between selection of optimal synonymous codons for translation efficiency, mutation and drift for the persistence of non-optimal codon (Bulmer 1991; Ikemura 1981; Robinson et al. 1984; Rocha 2004). It is a general rule that GC3 (GC content at the third position of a codon) is even higher than global GC content of corresponding coding region (Aissani et al. 1991; Bernardi 2004). Therefore, codon usage bias was to some extent accounted for the selection of G/C at the silent position. The high GC3 is believed to be an adaptive mechanism for high gene expression efficiency (Kim et al. 1997; Kudla et al. 2006) and high translational efficiency (Konu and Li 2002; Rocha 2004). Here, we proposed another possibility on the basis of the context-dependent CpG mutations. That is the high GC3 is at least partially caused by increased GC content at CpG flanking sites. In vertebrates, codons with high GC3 are favored in order to prevent CpG mutations in the following codon positions, and to increase GC content at silent position is the main approach to providing a stable environment for CpG-containing codons.

Acknowledgments

We thank Dr. Alessandra Riva for comments and critical reading of the draft.

REFERENCES

- Aissani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C, Bernardi G (1991) The compositional properties of human genes. *J Mol Evol* 32:493-503
- Antri SE, Mauffret O, Monnot M, Lescot E, Convert O (1993) Structural deviations at CpG provide a plausible explanation for the high frequency of mutation at this site. *J. Mol. Biol.* 230:373-378

- Baylin SB (2000) DNA methylation: tying it all together: epigenetics, genetics, cell cycle and cancer. *Science* 277:1948-1949
- Bernardi G (2004) Structural and evolutionary genomics: Natural selection in genome evolution. *Genome Biol.* 5:361-364
- 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* 8:1499-504
- Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. *Proc. Nat. Acad. Sci. USA* 83:3746-3750
- Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897-907
- Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. *Human Genet* 78:151-155
- Derreumaux S, Chaoui M, Tevanian G, Femandjian S (2001) Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucl. Acids Res.* 29:2314-2326
- Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38:1378-1385
- Fryxell KJ, Moon W (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22:650-658
- Fryxell KJ, Zuckerkandl E (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* 17:1371-1383
- Harbers K, Schnieke A, Stuhlmann H, Jahner D, Jaenisch R (1981) DNA Methylation and gene expression: Endogenous retroviral genome becomes infectious after molecular cloning. *Proc. Nat. Acad. Sci. USA* 78:7609-7613
- Hendrich B, Hardeland U, Ng H-H, Jiricny J, Bird A (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* 401:301-304
- Hodges-Garcia Y, Hagerman PJ (1992) Cytosine methylation can induce local distortions in the structure of duplex DNA. *Biochemistry* 31:7595-7599
- Hu WS, Fanning TG, Cardiff RD (1984) Mouse mammary tumor virus: specific methylation patterns of proviral DNA in normal mouse tissues. *J Virol.* 49:66-71
- Hunter CA (1993) Sequence-dependent DNA structure the role of base stacking interactions. *J Mol Biol* 230:1025-1054
- Ikehata H, Takatsu M, Saito Y, Tetsuya O (2000) Distribution of spontaneous CpG-associated G:C-->A:T mutations in the lacZ gene of mutaTM mice: effects of CpG methylation, the sequence context of CpG sites, and severity of mutations on the activity of the lacZ gene product. *Environ. Mol. Mutagen.* 36 301-311
- Ikemura T (1981) Correlation between the abundance of yeast transfer RNAs and the occurrence of the response codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E.coli translational system. *J. Mol. Evol.* 15:389-409
- Karlin S, Mrazek J (1996) What drives codon choices in human genes? *Journal of Molecular Biology* 262:459-472

- Kim CH, Oh Y, Lee TH (1997) Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* 199:293-301
- Konu O, Li MD (2002) Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *J Mol Evol* 54:35-41
- Krawczak M, Ball E, Cooper D (1998) Neighboring-nucleotide effects on the rates of germline single-base-pair substitution in human genes. *Am. J. Hum. Genet.* 63:474-488
- Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* 4:933-942
- Laura AS, Colin RP, Edward CH (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.* 62:551-563
- Leader DP, Peter B, Ehmer B (1995) Analysis of CpG dinucleotide frequency in relationship to translational reading frame suggests a class of genes in which mutation of this dinucleotide is asymmetric with respect to DNA strand. *FEBS Letters* 376:125-129
- Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in *Drosophila melanogaster*. *Nature* 408:538-540
- Mayer-Jung C, Moras D, Timsit Y (1997) Effect of cytosine methylation on DNA-DNA recognition at CpG steps. *J. Mol. Biol.* 270:328
- Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, p 73-74
- Norberg J, Vihinen M (2001) Molecular dynamics simulation of the effects of cytosine methylation on structure of oligonucleotides. *J. Mol. Struct.* 546:51-62
- Oakes CC, Salle SL, Smiraglia DJ, Robaire B, Trasler JM (2007) A unique configuration of genome-wide DNA methylation patterns in the testis. *Proc. Nat. Acad. Sci. USA* 104:228-233
- Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for De Novo methylation and mammalian development. *Cell* 99:247-257
- Ollila J, Lappalainen I, Vihinen M (1996) Sequence specificity in CpG mutation hotspots. *FEBS letters* 396:119-122
- Robertson KD, Jones PA (2000) DNA methylation: past, present and future directions. *Carcinogenesis* 21:461-467
- Robinson S, Lilley R, Little S, Emtage JS, Yamamoto G (1984) Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res.* 12:6663-6671
- Rocha EPC (2004) Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* 14:2279-2286
- Simmen MW, Leitgeb S, Charlton J, Jones SJM, Harris BR, Clark VH, Bird A (1999) Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* 283:1164-1167
- Tweedie S, Charlton J, Clark V, Bird A (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol. Cell. Biol.* 17:1469-1475
- Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE (2006) Functional CpG methylation system in a social insect. *Science* 314:645-647
- Wang Y, Leung FCC (2004) DNA structure constraint is probably a fundamental factor inducing CpG deficiency in bacteria. *Bioinformatics* 22:3336-3345
- Wang Y, Leung FCC (2008) GC Content increased at CpG flanking positions of fish genes compared with sea squirt orthologs as a mechanism for reducing impact of DNA methylation. *PLoS ONE* 3:e3612

- Wong E, Yang K, Kuraguchi M, Werling U, Avdievich E, Fan K, Fazzari M, Jin B, Brown AMC, Lipkin M, Edelmann W (2002) Mbd4 inactivation increases C->T transition mutations and promotes gastrointestinal tumor formation. *Proc. Nat. Acad. Sci. USA* 99:14937-14942
- You YH, Halangoda A, Buettner V, Hill K, Sommer S, Pfeifer G (1998) Methylation of CpG dinucleotides in the *lacI* gene of the Big Blue transgenic mouse. *Mutat Res.* 420:55-65
- Zhang F, Zhao Z (2004) The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics* 84:785-795