Differences in the substitution frequency of CpG/CpG to TpG/CpA based on the presence/absence of DNA cytosine methyltransferase

Original Article

Journal of the Tsuruma HEALTH SCI. MED. KANAZAWA UNIV. Vol. 29(2)
1～9 2005

Differences in the substitution frequency of CpG/CpG to TpG/CpA based on the presence/absence of DNA cytosine methyltransferase

Naoki Takahashi Hiroshi Nakashima

ABSTRACT

DNA sequence comparison was used to analyze nucleotide substitutions from CpG/CpG to TpG/CpA among three vertebrates and two invertebrates. Comparison of the G+C content at the third codon position showed that fruit fly has the most conservative sequences among the five species studied. Assuming that the conservative sequences are close to ancestral ones, the sequences of fruit fly were used as references to count the number of substitutions. Differences in the substitution frequencies from CpG to TpG across the species suggested that the DNA methylation enzyme has accelerated mutations in the DNA. Approximately 22% of CpGs in fruit fly were estimated to have mutated to TpGs in human. Most substitutions were observed at the third position in synonymous codons. Substitution frequencies from CpG to TpG were in the order zebrafish < mouse < human. This order is in accordance with the branching order shown in the dendrogram and indicates cumulative substitutions during the process of evolution. The CpG dinucleotides within the CpG islands are normally unmethylated and indicate a lesser number of CpG → TpG mutations. The substitution frequency of CpG to CpA in the coding sequences was considerably low, i.e., less than half that of the CpG to TpG substitutions. Plausible reasons for this difference are discussed.

KEY WORDS

CpG/CpG→TpG/CpA mutation, Sequence comparison, Cytosine methylation

Introduction

The mean guanine plus cytosine (G+C) content of genomic DNA among bacteria varies from approximately 25% to 75%[1]. A large variation has been reported in the occurrence of dinucleotide frequencies in the genes of different species[2-7], however, the reason for this variation is unclear. Only the CpG dinucleotide deficiency in vertebrates is clearly understood. This is explained by the fact that methylated cytosines are mutational hotspots that lead to CpG depletion during evolution. Vertebrates possess DNA cytosine methyltransferase that methylates cytosine only in the CpG dinucleotide, thereby yielding 5-methylcytosine. On the other hand, it is considered that invertebrates do not possess such methylation activity[8]. However, recent reports have indicated that cytosine residues in fruit fly are subjected to low levels of methylation[9,10]. Most methylation was observed at the CpT dinucleotide rather than at the CpG dinucleotide[10]. It has been reported that the inactivation of methyltransferases results in lethality, indicating that DNA methylation is an essential process in mammalian development[11]. Both methylated and unmethylated cytosines undergo spontaneous deamination to form thymine and uracil, respectively. The rate of deamination of methylated

Department of Clinical Laboratory Science, Graduate Course of Medical Science and Technology, Division of Health Science Kanazawa University

cytosine is higher than that of unmethylated cytosine. Repair of the T:G mismatch by the very short patch repair system is inefficient when compared with the repair of the U:G mismatch by uracil-DNA glycosylase. Therefore, cytosine methylation seems to increase the potential for mutation from C to T. The mutations triggered by cytosine methylation could be inferred by simple sequence comparison by counting the occurrence of TpG/CpA dinucleotides in the corresponding positions of CpGs in ancestral sequences. However, it is very difficult to infer ancestral sequences from extant species. To overcome this problem, we compared the G+C content at the third codon position in order to identify conservative species. We inferred the G+C content of the common ancestor by pairwise species comparison and estimated the species that had undergone major G+C content change. The number of nucleotide substitutions was counted assuming that the sequences of conservative species are close to ancestral ones. In this study, the protein-coding DNA sequences of nematode (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), zebrafish (*Danio rerio*), mouse (*Mus musculus*), and human (*Homo sapiens*) were analyzed based on the presence/absence of DNA cytosine methyltransferase.

Material and Methods

1. DNA sequence alignments

The cDNA sequences of nematode, fruit fly, zebrafish, mouse, and human were obtained from the web site of Reference Sequence release 3 (ftp://ftp.ncbi.nih.gov/refseq/release/) of the National Center for Biotechnology Information. Protein-encoding nucleotide sequences were selected according to the feature table of the data, and the amino acid sequences were obtained by translation. The dataset included 21,124 cDNA sequences of nematode, 18,746 cDNAs of fruit fly, 3,267 cDNAs of zebrafish, 26,211 cDNAs of mouse, and 27,405 cDNAs of human. Since the number of zebrafish sequences was considerably small than those of the others, the amino acid sequences of zebrafish were individually compared against the datasets of nematode, fruit fly, mouse, and human by using the BLASTP program[12]. The best hit pairs were selected if they showed greater than 30% amino acid identity over three-fourths of the total length of a given sequence. Proteins from the same species showing greater than 30% sequence identity were excluded to avoid any bias. The ClustalW program[13] was used for multiple sequence alignment of homologous sequences from the five species. Gap-free alignment regions longer than 100 amino acid residues were selected, and the corresponding DNA sequences were employed for the analysis. As a result, 38 pairwise alignments between the five species were employed. The alignment included 25,674 nucleotides and 8,558 codons per species. The DNA sequence similarity between human and nematode was 41% – 78%, and the amino acid sequence similarity was 32% –99%. The most conserved protein in the dataset was histone H3, while the least conserved was the adaptor-related protein. The phylogenetic tree was constructed by the ClustalW program using aligned amino acid sequences.

2. Analysis of dinucleotide composition

For a given DNA sequence, the occurrence of 16 types of dinucleotides was counted, and the expected dinucleotide composition was estimated as a product of the mononucleotide composition obtained in the same sequence. To study the dinucleotide frequencies, the log odds ratio was used, which is defined as the logarithm of the observed versus the expected dinucleotide composition. The log odds ratio was multiplied by 100 and expressed as an integer.

Log odds ratio = 100 × log (observed dinucleotide composition/expected dinucleotide composition)

The value of the log odds ratio obtained by this calculation was used in this study.

3. Identification of conservative species based on the G+C content at the third codon position

The codons were classified into three groups according to the definition of Bellgard and Gojobori [14]. Group 1 codons, referred to as IA, are different but code for identical amino acids. Codons with identical nucleotides at the first position were considered. Group 2 codons, DA, are different and code for different amino acids. Group 3 codons, IC, are identical codons. The average G+C content at the third position was calculated individually for the three groups of codons.

Pairwise comparisons of the G+C content of

three groups of codons were made across the five species. Assuming that the G+C content of IC codons represents their common ancestor, the species with less deviation between the IC and IA codons was considered to be conservative. Among the five species, the sequences from the most conservative species were regarded to be closest to the ancestral ones. These sequences were used as references when the nucleotide substitutions were counted.

4. Estimation of CpG/CpG → TpG/CpA substitutions

The substitution from CpG to TpG was estimated based on the sequence comparison between query and reference sequences. If a CpG dinucleotide in a reference sequence matched with a TpG dinucleotide in a query sequence, it was counted as one CpG → TpG substitution. It is well known that different proteins mutate at characteristically different rates. In reality, some proteins are highly conserved and show little or no change across species. In this study, the total number of CpG → TpG substitutions between two species was counted in 38 alignments. The total number of substitutions indicated the accumulated mutations of the two species since they diverged. The substitution frequency was expressed as the total number of substitutions versus the total number of CpG dinucleotides in the reference sequences. The CpG → CpA substitution was treated in a manner similar to that of the CpG → TpG substitution.

Results

1. Dinucleotide compositions

The mononucleotide and dinucleotide compositions of the aligned sequences were calculated. In this study, we focused on the substitution from CpG/CpG to TpG/CpA. The compositions of only the CpG, TpG, and CpA dinucleotides as well as their log odds ratios are listed in Table 1.

A dinucleotide with a positive or negative value of the log odds ratio indicates whether its occurrence relative to the random expectation is favorable or unfavorable, respectively. For instance, the log odds ratio of CpG in human has a large negative value of -27, indicating that its occurrence is unfavorable. On the other hand, both TpG and CpA have positive values indicating that their occurrence is favorable. This is consistent with previously reported results [2,4,6]. The CpG deficiency and TpG and CpA richness in vertebrates is explained by the fact that cytosine methylation tends to convert CpG/CpG to TpG/CpA, resulting in an increase in TpG and CpA. The log odds ratios of CpG in invertebrates were nearly zero, i.e., close to the random level. However, those of TpG and CpA were positive, indicating favorable occurrence. This result suggests that the increase in the amount of TpG and CpA in invertebrates might be attributed to some other mutational mechanisms.

2. G+C content at the third codon position

Figure 1 shows the linear correlation of the average G+C content between total sequences and the third codon positions for the five species (in filled circle). This is consistent with previous reports[1,14]. Among the five species, fruit fly has the highest G+C content (55.1%), while nematode has the lowest (48.0%). The three vertebrates have a similar G+C content. The G+C contents at the third position for the three groups of codons in the case of both fruit fly and nematode are marked in Figure 1 (squares for IA codons, crosses for DA codons, and triangles for IC codons). The difference in the case of IA codons was very large (40.1%). Assuming that the G+C content of IC codons represents their common ancestor, the G+C content of the nematode had changed significantly toward reduction in the G+C content, whereas the G+C content of fruit fly has remained

Table 1. Average dinucleotide compositions (%) and their log odds ratio, 100log(obs/exp).

species	dinucleotide			log odds ratio		
	CpG	TpG	CpA	CpG	TpG	CpA
nematode	5.78	6.50	7.77	0	7	6
fruit fly	7.22	7.02	7.62	-2	10	5
zebrafish	4.07	7.83	8.12	-19	12	10
mouse	3.51	8.17	7.84	-27	14	9
human	3.42	8.33	7.83	-27	15	9

unchanged. This result indicated that fruit fly is more conservative than nematode. Similar pairwise G+C content comparison indicated that fruit fly is the most conservative of the five species used in this study. Therefore, the sequences of fruit fly were used as the reference sequences. The phylogenetic tree indicates that the branching order is nematode < fruit fly < zebrafish < mouse < human; this result was obtained from the majority of amino acid sequences. However, the branching order between nematode and fruit fly was obscure when a phylogenetic tree was constructed from the amino acid sequences of certain proteins. Although, nematode branched earlier than fruit fly in the phylogenetic tree, the fruit fly sequence was used as the reference because fruit fly was found to be the most conservative species. Recently, Denver et al. have reported that the mutation rate of nematode is approximately 10 times higher than previous estimates[15].

3. CpG → TpG substitution frequencies

The DNA sequence alignment of the *ras* oncogene is shown in Figure 2. In this alignment, five CpG dinucleotides in fruit fly were substituted with TpGs in nematode. When we compared the sequences of fruit fly and human, it was found that 14 CpGs in fruit fly have changed to TpGs in human. A similar sequence comparison indicated that 11 and 15 CpGs in fruit fly have changed to TpGs in zebrafish and mouse, respectively. These results indicated that DNA cytosine methyltransferase in vertebrates accelerated the mutations. In the case where CpG was substituted with CpA, seven CpGs in fruit fly have changed to CpAs in human. The substitution frequency from CpG to CpA was very low when compared with that from CpG to TpG. The reason for this low frequency is discussed later.

Sequence comparison between fruit fly and nematode indicated that 273 CpGs in fruit fly have changed to TpGs in nematode in the 38 aligned sequences that were examined. Since there were 1829 CpGs in fruit fly, 15% of CpGs (273/1829) were found to have mutated to TpGs during the course of evolution. Since nematode does not possess DNA cytosine methyltransferase and fruit fly has very low CpG dinucleotide methylation activity, the above mutation ratio would correspond to the rate of spontaneous deamination that results in the formation of thymine.

The substitution frequencies of CpG to TpG are plotted in Figure 3. The substitution frequencies between two species, fruit fly (as reference) versus

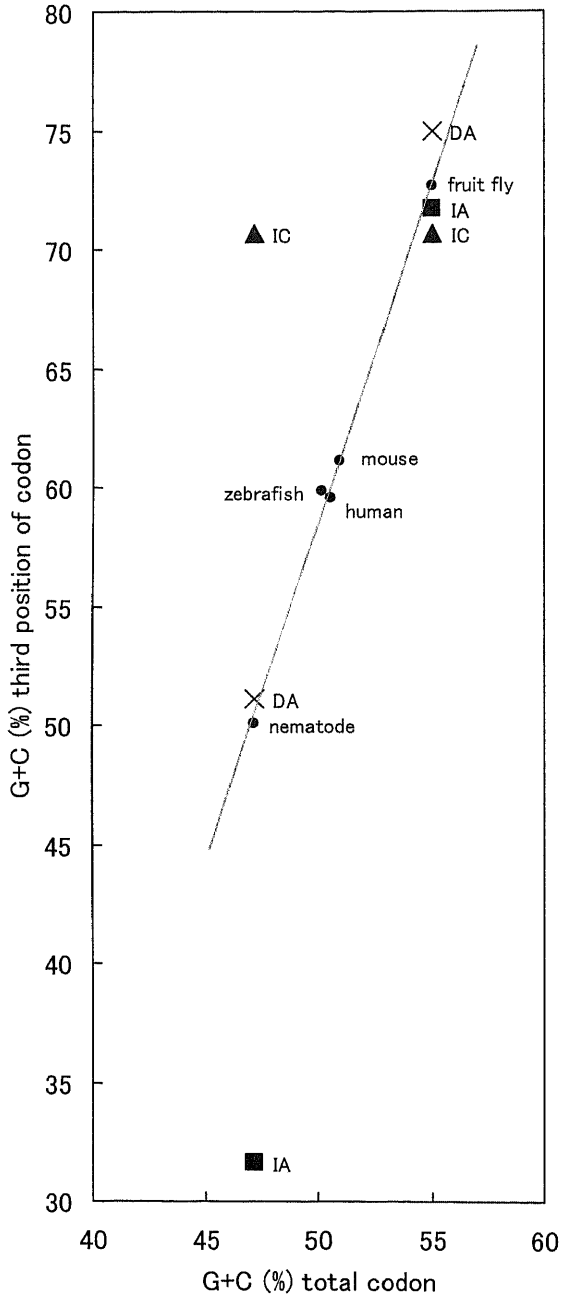

Figure 1:G+C content at the third codon position versus the total G+C content for the aligned sequences of nematode, zebrafish, human, mouse, and fruit fly (in filled circles). The G+C content at the third position for the three groups of codons of fruit fly and nematode is plotted: square for IA codons, cross for DA codons, and triangle for IC codons.

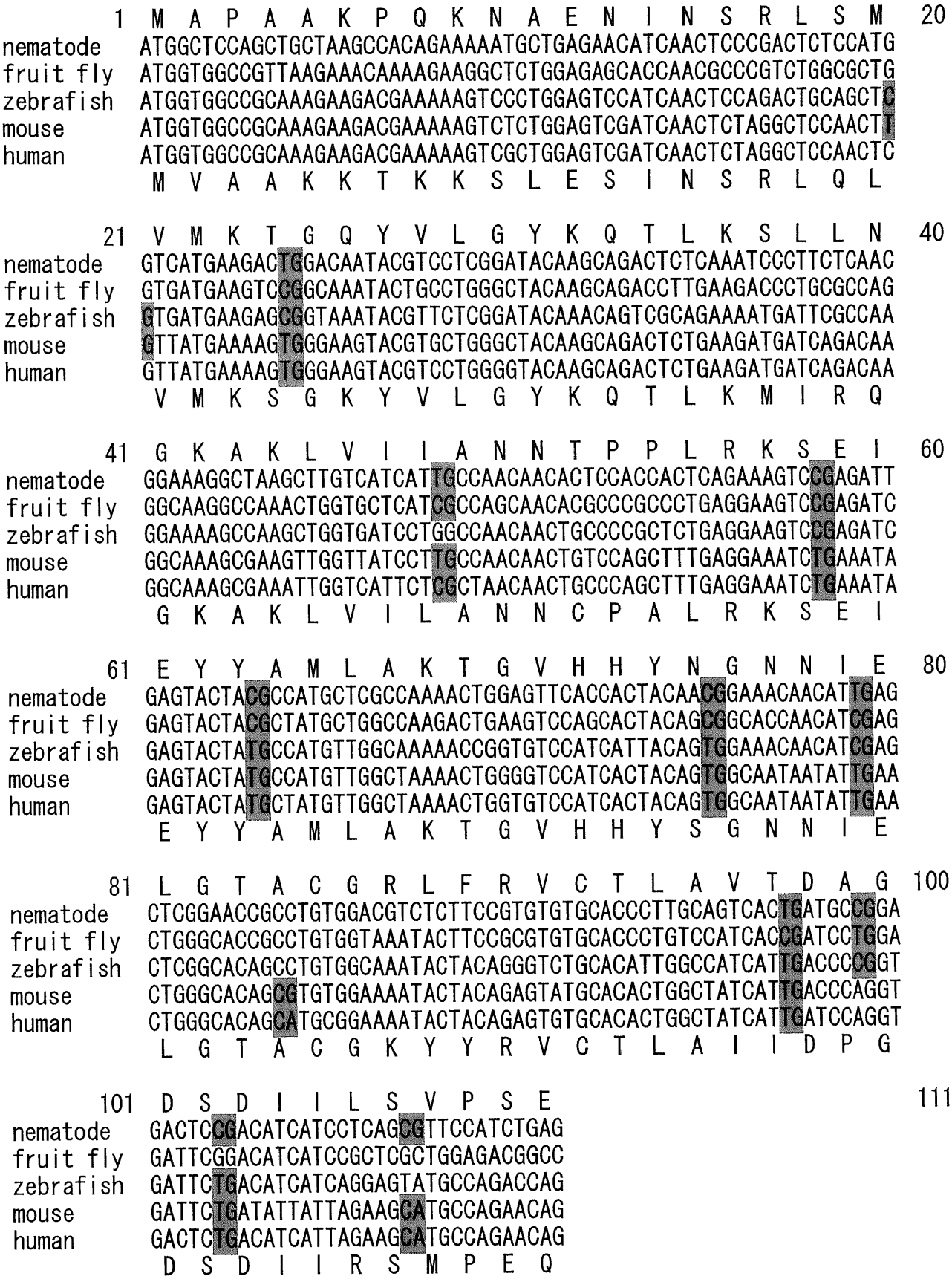

Figure 2:DNA sequence alignment of the *ras* oncogene with the residue numbers; the amino acid sequences of fruit fly and human are attached. Inferred substitution sites from CpG/CpG to TpG/CpA are highlighted in gray.

the query species were of the order zebrafish < mouse < human. This order is in accordance with the branching order in the dendrogram, and indicates that cumulative substitutions have occurred. This is reasonable because the amount of mutational changes is considered to be almost proportional to the evolutionary time. The substitution frequencies of CpG in invertebrates to TpG in vertebrates were estimated to be approximately 150% more than those of invertebrates. Sequence comparison between fruit fly and human indicated the largest substitution frequency of 22.4% (409/1829).

It is known that CpG dinucleotides within CpG islands are normally unmethylated, whereas most CpGs outside CpG islands are methylated. CpG islands are regions of more than 500 base pairs in size with a G+C content greater than 55% and an observed CpG / expected CpG ratio greater than 0.65 [16] (greater than -19 for the log odds ratio). They have been conserved during evolution because they are normally devoid of methylation. To examine the relationship between the substitution frequency and the CpG content in a simple manner, the relationship between the log odds ratios of CpG dinucleotides and the amount of CpGs was studied. Figure 4 indicates the plots (correlation coefficient, 0.84) of the log odds ratios of CpG versus the CpG content in human. Genes with a high CpG content exhibited less CpG → TpG mutations, indicating that the cytosines present in sequences with high CpG content might be unmethylated. Genes that had more than 5% of the CpG dinucleotide had log odds ratios of value close to -10. Ribosomal proteins L8 and L35, ADP-ribosylation factor 6, H3 histone, and guanine nucleotide binding protein, which are significantly conserved proteins, were included in this group. Genes that had less than 1% CpGs had log odds ratios of values less than -60. Calmodulin 2, signal recognition particle, and proteasome (macropain) were included in this group. Similar plots for genes from vertebrates showed strong correlations; however, those from invertebrates showed weak correlations (the correlation coefficient was 0.48 in fruit fly). Monroe et al. [17] have reported that in rodents, the

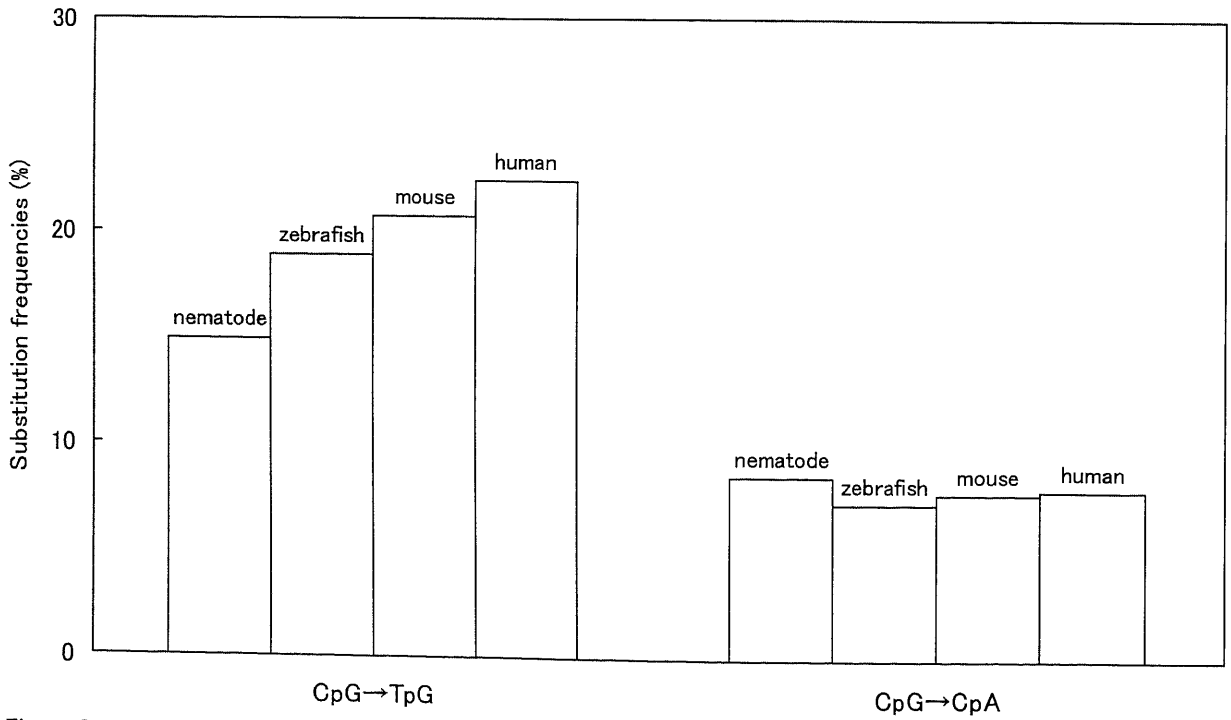

Figure 3: Inferred substitution frequencies (%) from CpG to TpG and from CpG to CpA are plotted. The number of substitutions was estimated as the number of replacements between the two species from the CpG dinucleotide in the reference sequence, i.e., the fruit fly sequence to the TpG dinucleotide in the query sequence. The substitution frequency was calculated as the total number of substitutions divided by the total number of CpG dinucleotides in the reference sequences.

substitution frequency is not related to the methylated cytosine content in CpGs. This is inconsistent with our results. Recently, Shiota has reported that during mammalian development the DNA methylation status is changeable, and cell differentiation is associated with both methylation and demethylation[18]. The changeable cytosine methylation might be related to this inconsistency.

Discussion

The information contained in DNA sequences has been obscured by numerous nucleotide substitutions that occurred during millions of years of diverging evolution. The branching order of the tree only indicates the emergence of lineages. Each extant species has accumulated numerous substitutions from its ancestors. Therefore, inferring ancestral sequences is very difficult. In this work, most conservative species were assessed by comparison of the G+C content at the third codon position. The fruit fly was found to be the most conservative of the five species studied. The number of nucleotide substitutions was counted assuming that the sequences of fruit fly are close to the ancestral sequences. Fossil flies trapped in amber and preserved intact for millions of years are excellent sources of ancient DNA. However, these sequences are not available at present. Similar looking between fossil and extant flies suggests the similarities in their genes.

Since DNA is a double-stranded molecule, the substitution of CpG/CpG to TpG/CpA produces the same amount of TpG and CpA. However, in the coding sequence, the substitution frequency of CpG to CpA was very low when compared with that of CpG to TpG. To examine this difference, the substitution sites in the codons and the effect of the nucleotide substitutions on the amino acids were surveyed. For example, sequence comparison between fruit fly and human showed the highest substitution frequency, i.e. 409 CpGs in fruit fly were substituted with TpGs in human. In this case, 396 substitutions i.e. 97% (396/409) occurred at the third codon position, three substitutions at the first position, and ten substitutions at the second position. Most of the substitutions

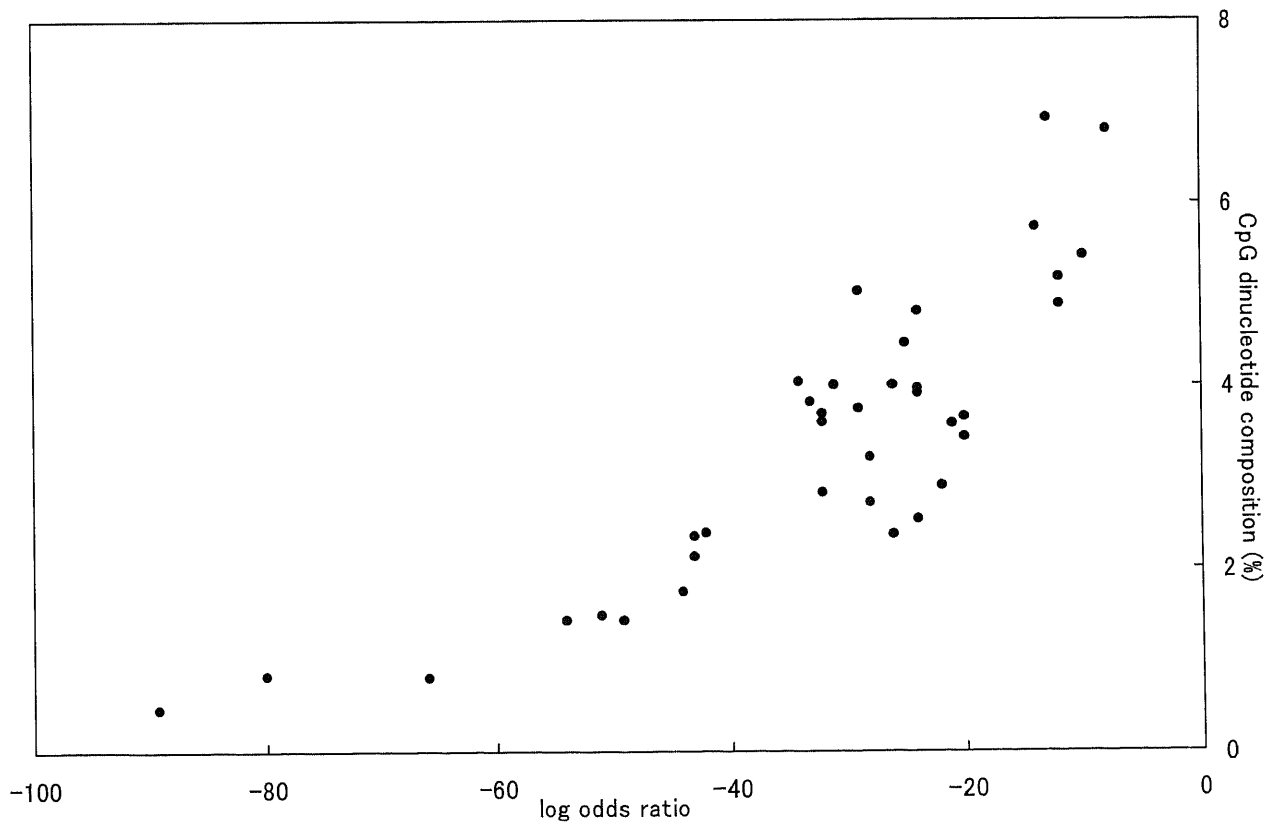

Figure 4: The log odds ratio, 100 log (obs/exp), versus the CpG dinucleotide composition (%) for 38 human sequences.

occurred at the third codon position, and this trend was commonly observed in other cases. Among the amino acid residues, 327 remained identical, and 82 were replaced by other amino acids. Regarding the substitution from CpG to CpA, 145 CpGs in fruit fly were substituted with CpAs in human. Therefore, the substitution frequency is approximately one-third that of CpG to TpG. In this case, 88 substitutions, i.e., 68% (88/145) were observed at the third codon position, 48 substitutions at the first position, and nine at the second position. Among the amino acid residues, 72 remained identical, and 73 were replaced by other amino acids. Nucleotide substitutions of CpG to CpA indicated higher frequencies of amino acid substitutions than those of CpG to TpG. Amino acid substitutions might be detrimental to protein function. This might be the reason for the lower substitution frequency of CpG to CpA when compared with that of CpG to TpG. Nucleotide substitution between pyrimidines at the third codon position would not yield an amino acid replacement. In addition to the CpG → TpG substitutions, a majority of the amino acid substitutions were accompanied by other nucleotide substitutions. In case of the Met and Ile residues, nucleotide substitutions of purines at the third codon position might yield amino acid substitutions. With regard to synonymous codons, substitution between pyrimidines at the third codon positions could occur without changing eight amino acid residues—Cys, Asp, Phe, His, Ile, Asn, Ser, and Tyr. The same substitution between purines could occur in five residues — Glu, Lys, Leu, Gln, and Arg. Therefore, the chances of substitution between pyrimidines at the third codon position without amino acid replacement are more when compared with that between purines. This is consistent with the report of Grantham[19] in which it was shown that pyrimidines are generally preferred to purines at the third position of synonymous codons.

The substitution frequencies from CpG to TpG showed an approximately 150% increase if the species have DNA cytosine methyltransferase. Previously, Lutsenko and Bhagwat [20] have reported that cytosine methylation results in a four-fold increase in mutation frequency. The differences in mutation frequency might arise from the dataset of the sequences. In this study, we employed rather conserved sequences, i.e., sequences with amino acid identity greater than 30% in order to analyze gap-free regions. Conservative sequences have lower mutation frequencies, therefore, this might be the reason for the inconsistency of the mutation frequency by cytosine methylation.

It is reported that genes have species-specific nucleotide compositions [2-7], and this feature is more significant in bacteria. Therefore, nonrandom mutations might occur in bacteria. However, there is insufficient information on nonrandom mutations. Availability of more information on sequence-dependent mutations, such as CpG→TpG mutations, would increase our understanding of not only species-specific compositions but also the process of evolution.

References

1) Muto, A., Osawa, S. : The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA, 84 : 166-169, 1987.
2) Nussinov, R. : Nearest neighbor nucleotide patterns. J. Biol. Chem., 256 : 8458-8462, 1981.
3) Nussinov, R. : Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res., 12 : 1749-1763, 1984.
4) Karlin, S., Burge, C. : Dinucleotide relative abundance extremes: a genomic signature. Trend Genetics, 11 : 283-290, 1995.
5) Karlin, S. et al. : Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol., 179 : 3899-3913, 1997.
6) Nakashima, H. et al. : Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* genes. DNA Res., 4 : 185-192, 1997.
7) Nakashima, H. et al. : Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res., 5 : 251- 259, 1998.
8) Bird, A. : DNA methylation patterns and epigenetic memory. Genes & Dev., 16 : 6-21, 2002.
9) Gowher, H. et al. : DNA of *Drosophila melanogaster* contains 5-methycytosine. EMBO J., 19 : 6918-6923, 2000.
10) Lyko, F. et al. : DNA methylation in *Drosophila melanogaster*. Nature, 408 : 538-540, 2000.
11) Okano, M. et al. : DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell, 99 : 247-257, 1999.
12) Altschul, S.F. et al. : Basic local alignment search tool. J. Mol. Biol., 215 : 403-410, 1990.
13) Thompson, J.D. et al. : CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22 : 4673–4680, 1994.

14) Bellgard, M.I., Gojobori, T. : Significant differences between the G+C content of synonymous codons in orthologous genes and the genomic G+C content. Gene, 238 : 33-37, 1999.
15) Denver, D.R. et al. : High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature, 430 : 679-682, 2004.
16) Takai, D., Jones, P.A. : Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA, 99 : 3740-3745, 2002.
17) Monroe, J.J. et al. : Extent of CpG methylation is not proportional to the in vivo spontaneous mutation frequency at transgenic loci in Big Blue™ rodents. Mutat. Res., 476 : 1-11, 2001.
18) Shiota, K. : DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet. Genome Res., 105 : 325-334, 2004.
19) Grantham, R. : Workings of the genetic code. Trend Biochem. Sci., 5 : 327-331, 1980.
20) Lutsenko, E., Bhagwat, A.S. : Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells A model, its experimental support and implications. Mutat. Res., 437 : 11-20, 1999.

DNAシトシンメチル化酵素の有無による２塩基CpG/CpGからTpG/CpAへの置換頻度の違い

高橋　直生，中島　広志

要　　旨

脊椎動物は、2塩基CpGのシトシンのみをメチル化して5-メチルシトシンに変えるDNAシトシンメチル化酵素を持っている。メチル化されたシトシンは脱メチル化しチミンに変わりやすい。一方、無脊椎動物は一般にシトシンメチル化酵素を持たない。最近、ハエが非常に弱いメチル化活性を持っていると報告された。シトシンは脱メチル化されるとウラシルとなる。脱メチル化で生じたチミンとウラシルはＤＮＡ修復酵素により復元されるが、ウラシルの修復力が高いと知られている。ここではゼブラフィッシュ、マウス、ヒトの脊椎動物3種と、ハエ、線虫の無脊椎動物2種の合計5種の相同な塩基配列38本を用いCpG/CpGからTpG/CpAへの置換頻度の解析をおこなった。２つの生物種の相同配列において同一コドンを先祖型のコドンに近いと仮定し、同一コドン3位と同義語コドン3位のＧ＋Ｃ量の差の小さい生物種を先祖型に近いと判定した。ここで用いた5種の生物種においてハエが最も先祖型に近いと得られた。よって、ハエの配列の2塩基CpGに対応する２塩基がTpG/CpAとなっている数を線虫、ゼブラフィッシュ、マウス、ヒトの配列で数えそれぞれの塩基置換数を推定した。脊椎動物のCpGからTpGの置換頻度は無脊椎動物の約１．５倍であった。大部分の置換は同義語コドン3位で起こっていた。ＤＮＡは2本鎖だからCpGからCpAの置換頻度はCpGからTpGの頻度と同じと予想されるが実際の置換頻度は半分以下と低かった。この理由について議論する。