# CpG Dinucleotide Distribution and DNA Methylation

Tom Shimizu[1,2] tom@sfc.keio.ac.jp
Kouichi Takahashi[1,3] t94249kt@sfc.keio.ac.jp
Masaru Tomita[1,3] mt@sfc.keio.ac.jp

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

It is known that the dinucleotide CpG is significantly underrepresented in genomic sequences of organisms which extensively methylate their DNA[1]. In these species, most cytosine bases of CpG dinucleotides are found to be methylated and this extensive CpG methylation is thought to have caused the depletion of the dinucleotide over the course of evolution[2]. Thus, the extent of CpG depletion in the genomic sequence can serve as an index of the extent of CpG methylation in an organism.

CpG islands are small regions of these CpG-depleted genomes which have remained relatively CpG-rich, and are usually unmethylated[3]. They are associated with most housekeeping genes and many tissue-specific genes and are most often found in the $5'$ flanking region[4]. It is also known that the methylation state of CpG islands is sometimes associated with gene suppression.

In the present work, we have attempted to characterize the different types of DNA methylation in various species by focusing on (a)the extent of CpG depletion and (b)the size of CpG islands around the $5'$ end of genes where most CpG islands are located.

In order to assess the degree of CpG depletion and visualize the average CpG island profile(the distribution of CpG dinucleotides around the $5'$ end of genes) in a given set of sequences, CpG O/E(Observed/Expected ratio) values were calculated in 500bp windows moving across each sequence in 10bp steps. CpG O/E is defined as the ratio of the actual CpG density to that expected from the local base composition, and was calculated by the formula

$$CpG\ O/E = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \cdot \frac{W^2}{W-1}$$

where $W$ is the number of nucleotides in a window(adapted from [5]). The sequences were aligned at the start codon and the average CpG O/E at each window position along the sequence was plotted on a two dimensional graph as the mean CpG island profile. The start codon was used as the aligning point because most CpG islands are known to exist at the $5'$ end of genes[4]. Sequences from GenBank Release 94.0 were used for analysis.

Our results are consistent with previous observations in that many vertebrate genes are associated with CpG islands at their $5'$ end. Although CpG DNA methylation is generally known to be a feature of vertebrate and plant species, our results revealed significant CpG depletion and CpG island-like features in other organisms. We will discuss the different types of CpG depletion and the their implications in the context of DNA methylation.

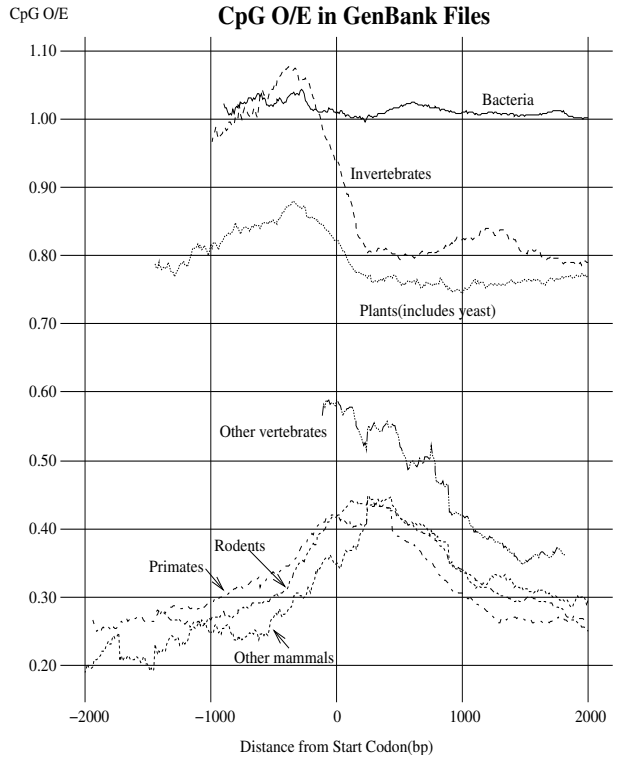


**Figure 1**

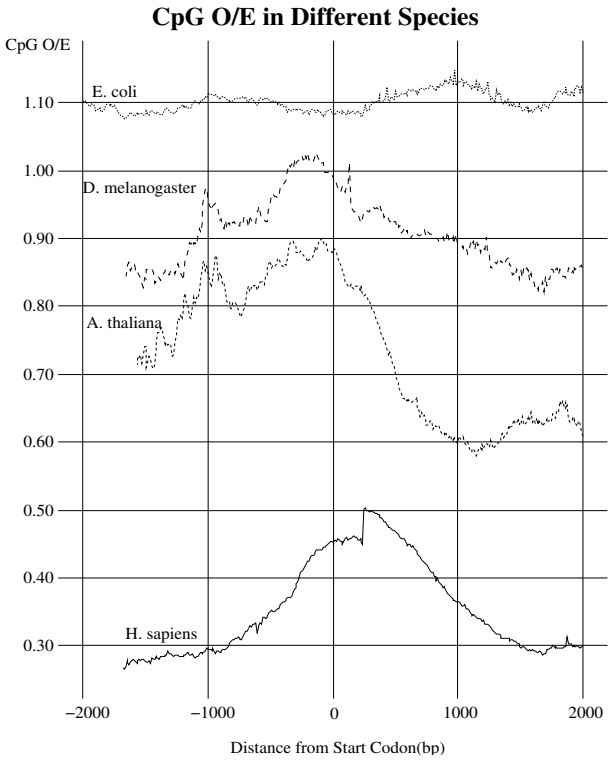


**Figure 2**

We first analyzed all sequences in the classification provided by the GenBank flat file release: primates, rodents, other mammals, other vertebrates, plants(includes yeast), invertebrates, and bacteria(**Figure 1**). We then classified the sequences by species; all genes for each species with more than 1000 entries in GenBank were analyzed. Human, *Drosophila*, *Arabidopsis* and *E. coli* profiles are shown in **Figure 2** to represent vertebrates, invertebrates, plants and bacteria.

## Acknowledgement

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

## References

[1] Bird, A. P., and Taggart, M. H. "Variable patterns of total DNA and rDNA methylation in animals," *Nucleic Acids Res.*, Vol. 8, pp. 1485-97, 1980.

[2] Bird, A. P. "DNA methylation and the frequency of CpG in animal DNA," *Nucleic Acids Res.*, Vol. 8, pp. 1499-104, 1980.

[3] Bird, A. P., Taggart, M. H., Frommer, M., Miller, J. M. and Macleod, M. "A Fraction of the Mouse Genome That Is Derived from Islands of Nonmethylated, CpG-Rich DNA," *Cell*, Vol. 40, pp. 91-99, 1985.

[4] Gardiner-Garden, M. and Frommer, M. "CpG Islands in Vertebrate Genomes," *J. Mol. Biol.*, Vol. 196, pp. 261-282, 1987.

[5] Matsuo K., Clay, O., Takahashi, T., Silke, J. and Schaffner, W. "Evidence for Erosion of Mouse CpG Islands during Mammalian Evolution ," *Som. Cell and Mol. Gen.*, Vol. 19, pp. 543-555, 1993.