Strong negative selection on CpG dinucleotides in the human genome: a comparison of de novo and inherited mutation rates

Sergey A. Spirin

Department of Mathematical Methods in Biology, Belozersky Institute, Moscow State University, sas@belozersky.msu.ru

Sofya A. Medvedeva

Department of Bioengineering and Bioinformatics, Moscow State University, sof.medv@gmail.com

Alexander Y. Panchin

Institute for Information Transmission Problems, Russian Academy of Science, alexpanchin@yahoo.com

Andrey V. Alexeevski

Department of Mathematical Methods in Biology, Belozersky Institute, Moscow State University,, aba@belozersky.msu.ru

Yuri V. Panchin

Institute for Information Transmission Problems, Russian Academy of Sciences, ypanchin@yahoo.com

In a recent article Kong et al.[1] measured *de novo* mutations rates in 78 human parent-offspring trios. The reported rate of *de novo* C to T (and also G to A) mutations is 18 times higher in CpG sites than in non-CpG sites and this difference is greater than previous estimates. *De novo* mutations in parent-offspring trios seem to be the best available representation of mutations that have not passed through long-term filters of natural selection. Other sources used to obtain mutation data such as SNPs or cross-species variations are influenced by natural selection to a much greater extent because they provide information on mutations that were probably passed down through many generations.

We compared the *de novo* mutation data obtained by Kong et. al and mutation data derived from a set of inherited (transmitted) mutations that resulted in human single nucleotide polymorphisms (SNPs). The direction of these inherited mutations was established by reconstructing the ancestral states of SNPs using the alignments of human, chimp and orangutan genomes. The comparison of these two mutation datasets adds new prospective to the results obtained by Kong et al. Both *de novo* and inherited mutation datasets indicate a trend towards the decrease of G+C content in the human genome, but CpG dinucleotides are accumulated according to inherited mutations data, yet lost according to *de novo* mutations data. By comparing mutation datasets we were able to estimate the fraction of CpG dinucleotides experiencing the pressure of stabilizing selection. Our analysis supports the hypothesis that natural selection strongly favors CpG preservation on the genome scale in humans. Computations based on dataset comparisons suggest that at least 50% of CpG dinucleotides in the human genome are under the pressure of stabilizing selection. This hypothesis is also supported by additional analysis of the allele frequencies of novel variants resulting from gained and lost CpG dinucleotides.

1. Kong, A. et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk, *Nature*, **488**:471-475