CHANGE IN CpG CONTEXT IS A LEADING CAUSE OF CORRELATION BETWEEN RATES OF NON-SYNONYMOUS AND SYNONYMOUS SUBSTITUTIONS IN RODENTS

Bazykin G.A.*, Ogurtsov A.Y., Kondrashov A.S.

1 Dept. of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08540, USA; 2 National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894, USA

* Corresponding author: e-mail: gbazykin@princeton.edu

Keywords: evolution, point substitution rate, mutation bias, CpG deamination, dinucleotides

Summary

Motivation: Correlation between the rates of synonymous (silent) and non-synonymous (amino acid-changing) nucleotide substitutions in genes is a wide-spread and yet unexplained genome-level phenomenon which is in disagreement with the neutral theory of molecular evolution (Kimura, 1983). In mammals, this correlation can be caused by mutational dependence of the point mutation events.

Results: Comparison of 7,732 alignments of mouse and rat genes confirms the previously observed correlation between rates of substitutions in non-degenerate (Kₐ) and four-fold degenerate (K₄) nucleotide sites. In rodents, this correlation is primarily caused by tandem substitutions and, in particular, by CpG mutation bias leading to doublet nucleotide substitutions. The nature of correlation between Kₐ and K₄ in seven pairs of prokaryotic genomes is unclear.

Introduction

Synonymous (silent) nucleotide sites are often assumed to evolve “neutrally” and therefore are frequently used as a measure of non-synonymous substitutions. This assumption, however, conflicts with the well-described phenomenon of variation of rates of synonymous substitutions across the genome and, in particular, of correlation between rates of non-synonymous and synonymous substitutions. Selection for translation efficiency (Chamary, Hurst, 2004) or RNA structure (Smith, Hurst, 1999) acting on silent sites were suggested as possible explanations, as well as methodological biases in distance estimation (Bielański et al., 2000).

It has been claimed that the correlation of rates of synonymous and non-synonymous substitutions is dependent upon the particular method used for estimation of substitution rates (Bielański et al., 2000). Therefore, to reveal the leading cause of this correlation, it is preferable to use closely related species. At low evolutionary distances, substitutional saturation is negligible, and different methods of estimation of divergence converge.

Methods

Mouse and rat coding sequences were obtained from version 30 of the mouse genome (Mouse Genome Sequencing Consortium, 2002) and version 2 of the rat genome (Rat Genome Project Sequencing Consortium, 2004) from NCBI. Orthologs were identified according to the two-directional best-hit approach using protein BLAST (Altschul et al., 1997). Alignments of the amino acid sequences for each pair of the orthologs was made using ClustalW (Thompson et al., 1994) and reverse transcribed to get the nucleotide alignments. Rates of nucleotide substitutions in different groups of sites were obtained using a PERL script available from the authors. All suitable triplets of bacterial genomes were obtained from the NCBI Entrez database and processed analogously. Genes with doublets removed are those in which adjacent nucleotide sites were excluded from
analysis if both carried substitutions. A substitution at site 1 of the doublet was assumed to change the CpG context of the following site 2 when one of the species carried “C” at site 1 and the other species carried some other nucleotide. A substitution at site 2 of the doublet was assumed to change the CpG context of the preceding site 1 when one of the species carried “G” at site 2 and the other species carried some other nucleotide.

Outliers can have a profound effect on the value of correlation coefficient. In order to ensure that only high-quality (unambiguous) alignments are included in the analysis, we excluded all genes with divergences in non-degenerate sites exceeding 1.5 average amino-acid divergences between corresponding species, and divergences in 4-fold-degenerate sites exceeding 10 average amino-acid divergences (therefore the abrupt left and top boundaries of region with data points at Fig. 1a). This approach is conservative in regard to determination of correlation.

**Results and Discussion**

Our data confirms the previously observed significant correlation between per gene substitutions rates in non-synonymous and synonymous nucleotide sites (Fig. 1a). This correlation, however, is primarily caused by doublet substitutions occurring in adjacent nucleotides. When sites with double substitutions were excluded from analysis, the magnitude of correlation was greatly reduced (Fig. 1b).

Correlation between substitutions in adjacent sites can arise if one mutational event simultaneously affects two successive nucleotides. However, such double substitutions are extremely rare (Kondrashov, 2003), and the observed effect has to be caused by separate point mutation events. Such correlation can also be due to selection on silent substitutions that restore codon bias following an amino acid change (Lipman, Wilbur, 1984).
The nature of correlation is revealed by consideration of the sites of adjacent substitutions in which one of the substitutions can affect the CpG context of the neighboring nucleotide site. Removal of the subset of such sites is sufficient to achieve the strong reduction in correlation (Fig. 1c). Conversely, only a minor reduction in the correlation coefficient is achieved by removal of sites of neighboring substitutions in which both substitutions leave the CpG context of the other one invariant (Fig. 1d).

The simplest explanation for correlation between K₁ and K₄ that is consistent with these findings is interdependence of mutational events in adjacent nucleotides due to CpG deamination. CpG dinucleotide is hypermutable in vertebrates. If the first substitution (regardless of whether it occurs in a non-synonymous or synonymous site) creates the CpG dinucleotide, the second substitution at the adjacent nucleotide site is facilitated. This is expected to result in the observed pattern of substitutions coupling.

This explanation is further supported by analysis of seven pairs of closely related bacterial genomes. All the pairs of bacterial species indicated significant correlation between K₁ and K₄ of various magnitude. However, removal of doublets and, in particular, of doublets involving change in CpG context did not lead to a profound decrease in correlation comparable with that observed in rodents. Therefore, some other factor has to be responsible for correlation between K₁ and K₄ in prokaryotes.

An obvious next step would be to reveal the order of substitutions – whether the change of context in non-synonymous site facilitates the synonymous substitution or vice versa. This can be achieved if a third orthologous gene from an outgroup species (e.g., human) is employed.

**Acknowledgements**

G.B. was supported by BWF graduate fellowship.

**References**


**Table.** Correlation coefficients between divergences in non-degenerate and four-fold degenerate nucleotide sites in 8 pairs of genomes

<table>
<thead>
<tr>
<th></th>
<th>No. of genes</th>
<th>Fraction of amino acid differences¹</th>
<th>All sites</th>
<th>Doublets removed²</th>
<th>Doublets with change in CpG context removed²</th>
<th>Doublets without change in CpG context removed²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muridae</td>
<td>7.732</td>
<td>4.3 %</td>
<td>0.3</td>
<td>0.09</td>
<td>0.11</td>
<td>0.26</td>
</tr>
<tr>
<td>Bacillus</td>
<td>1.915</td>
<td>3.5 %</td>
<td>0.44</td>
<td>0.34</td>
<td>0.36</td>
<td>0.42</td>
</tr>
<tr>
<td>Bordetella</td>
<td>2.696</td>
<td>0.4 %</td>
<td>0.12</td>
<td>0.10</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Ecoli</td>
<td>3.122</td>
<td>1.2 %</td>
<td>0.27</td>
<td>0.21</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>Salmonella</td>
<td>2.531</td>
<td>0.8 %</td>
<td>0.21</td>
<td>0.17</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>Staphylococcus</td>
<td>1.591</td>
<td>0.5 %</td>
<td>0.23</td>
<td>0.20</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Streptococcus</td>
<td>1.065</td>
<td>0.7 %</td>
<td>0.28</td>
<td>0.20</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>Vibrio</td>
<td>579</td>
<td>0.7 %</td>
<td>0.29</td>
<td>0.25</td>
<td>0.26</td>
<td>0.28</td>
</tr>
</tbody>
</table>


¹ Fraction of mismatches in alignments of orthologous proteins between genomes; ² see Methods for details. All correlations were significant at P < 0.05.

An obvious next step would be to reveal the order of substitutions – whether the change of context in non-synonymous site facilitates the synonymous substitution or vice versa. This can be achieved if a third orthologous gene from an outgroup species (e.g., human) is employed.