On Low Frequency of CpG Dinucleotides in Bacterial Genomes

Mami Goto \(^{13}\)  
mgoto@sfc.keio.ac.jp  
Takanori Washio \(^{13}\)  
washy@sfc.keio.ac.jp  
Masaru Tomita \(^{23}\)  
mt@sfc.keio.ac.jp

1 Laboratory for Bioinformatics  
2 Graduate School of Media and Governance  
3 Department of Environmental Informaion, Keio University, Endo 5322, Fujisawa, Kanagawa 252-0816, Japan.

CpG depletion is a phenomenon known to be caused by CpG DNA methylation. Although CpG DNA methylation is believed to be a feature of vertebrates and plants [1], we have found that some procaryote genomes such as \(M.\) genitalium show significantly low CpG frequencies. On the other hand, the genome of \(M.\) pneumoniae, which is a closely related species to \(M.\) genitalium, does not show such clear sign of CpG depletion.

In order to discuss possible causes of the bacterial CpG depletion, we conducted computer analyses of frequencies of CpG dinucleotides in twelve complete procaryote genomes, including \(M.\) genitalium and \(M.\) pneumoniae.

For all of the complete genome sequences, we first analyzed CpG frequencies in coding and non-coding regions separately. The results of the analysis are shown in Table 1. CpG observed/expected (O/E) values are significantly lower in coding sequences than non-coding sequences in the genomes of \(M.\) genitalium, \(B.\) burgdorferi and \(M.\) jannaschii. In the genome of \(M.\) thermoautotrophicum, on the other hand, CpG O/E values are about the same in both coding and non-coding sequences.

Comprehensive analysis of mutation patterns of single nucleotide and dinucleotide substitutions has been conducted using homologous sequence pairs of the genomes of of \(M.\) genitalium and \(M.\) pneumoniae. These homologous sequences were aligned, and we counted substituted nucleotides and dinucleotides only if six bases directly flanking them (three bases upstream and three bases downstream) are perfectly conserved. The results are shown in Table 2.

While the point mutations from C (cytosine) in \(M.\) pneumoniae to T (thymine) in \(M.\) genitalium are frequent in general, C to T mutations occurred most frequently on the cytocine in the CpG dinucleotide (Table 3).

Table 1: CpG Observed/Expected (O/E) ratio. Expected CpG frequencies are products of C and G single nucleotide frequencies.

<table>
<thead>
<tr>
<th>species</th>
<th>Whole sequence</th>
<th>Coding sequence</th>
<th>Non coding sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M.) genitalium</td>
<td>0.39</td>
<td>0.37</td>
<td>0.54</td>
</tr>
<tr>
<td>(M.) pneumoniae</td>
<td>0.82</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>(E.) coli</td>
<td>1.16</td>
<td>1.16</td>
<td>1.09</td>
</tr>
<tr>
<td>(H.) influenzae</td>
<td>1.09</td>
<td>1.09</td>
<td>1.07</td>
</tr>
<tr>
<td>(B.) subtilis</td>
<td>1.04</td>
<td>1.05</td>
<td>0.87</td>
</tr>
<tr>
<td>(H.) pylori</td>
<td>0.93</td>
<td>0.94</td>
<td>0.84</td>
</tr>
<tr>
<td>(B.) burgdorferi</td>
<td>0.48</td>
<td>0.47</td>
<td>0.73</td>
</tr>
<tr>
<td>(S.) PCC6803</td>
<td>0.75</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>(A.) aciculus</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>(A.) fulgidus</td>
<td>0.78</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>(M.) jannaschii</td>
<td>0.32</td>
<td>0.27</td>
<td>0.67</td>
</tr>
<tr>
<td>(M.) thermoautotrophicum</td>
<td>0.51</td>
<td>0.51</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Table 2: Nucleotide variations in homologous genes of *M. genitalium* and *M. pneumoniae*.

<table>
<thead>
<tr>
<th>Change</th>
<th>M. genitalium (%)</th>
<th>M. pneumoniae (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → T</td>
<td>734</td>
<td>5.67%</td>
</tr>
<tr>
<td>A → G</td>
<td>1003</td>
<td>7.75%</td>
</tr>
<tr>
<td>A → C</td>
<td>401</td>
<td>3.10%</td>
</tr>
<tr>
<td>T → A</td>
<td>922</td>
<td>7.12%</td>
</tr>
<tr>
<td>T → G</td>
<td>284</td>
<td>2.19%</td>
</tr>
<tr>
<td>T → C</td>
<td>1036</td>
<td>8.01%</td>
</tr>
<tr>
<td>G → A</td>
<td>2205</td>
<td>17.04%</td>
</tr>
<tr>
<td>G → T</td>
<td>941</td>
<td>7.27%</td>
</tr>
<tr>
<td>G → C</td>
<td>399</td>
<td>3.08%</td>
</tr>
<tr>
<td>C → A</td>
<td>1189</td>
<td>9.19%</td>
</tr>
<tr>
<td>C → T</td>
<td><strong>3469</strong></td>
<td><strong>26.81%</strong></td>
</tr>
<tr>
<td>C → G</td>
<td>358</td>
<td>2.77%</td>
</tr>
</tbody>
</table>

12941 100.00%  

Table 3: Dimucleotide variations in homologous genes of *M. genitalium* and *M. pneumoniae*.

<table>
<thead>
<tr>
<th>Occurrences</th>
<th>Frequency(MP=MG(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>3054</td>
</tr>
<tr>
<td>AT</td>
<td>1179</td>
</tr>
<tr>
<td>AG</td>
<td>1554</td>
</tr>
<tr>
<td>TA</td>
<td>1555</td>
</tr>
<tr>
<td>TT</td>
<td>838</td>
</tr>
<tr>
<td>TG</td>
<td>3398</td>
</tr>
<tr>
<td>TC</td>
<td>2428</td>
</tr>
<tr>
<td>CT</td>
<td>1391</td>
</tr>
<tr>
<td>CC</td>
<td>358</td>
</tr>
</tbody>
</table>

We therefore conclude that this type of mutations, Cpg to TpG/CpA, is the primary force of the Cpg depletion in *M. genitalium*. However, since *M. genitalium* does not have a gene homologue to Cpg DNA methylase, the cause of the frequent Cpg mutation is yet to be known.

Acknowledgements

This work is supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, “Genome Informatics”, from The Ministry of Education, Science, Sports and Culture in Japan.

References