Estimation of prokaryote genomic DNA G+C content by sequencing universally conserved genes

Pierre-Edouard Fournier,1 Karsten Suhre,1 Ghislain Fournous2 and Didier Raoult2

1Information Génomique et Structurale, CNRS UPR2589, Case 934, 163 Avenue de Luminy, 13288 Marseille cedex 09, France
2Unité des rickettsies, IFR 48, CNRS UMR 6020, Faculté de Médecine, Université de la Méditerranée, 27 Boulevard Jean Moulin, 13385 Marseille cedex 05, France

Determination of the DNA G+C content of prokaryotic genomes using traditional methods is time-consuming and results may vary from laboratory to laboratory, depending on the technique used. We explored the possibility of extrapolating the genomic DNA G+C content of prokaryotes from gene sequences. For this, 127 universally conserved genes were studied from 50 prokaryotic genomes in the Clusters of Orthologous Groups database. Of these, 57 genes were present as a single copy in the genomes of 157 different prokaryotic species available in GenBank. There was a strong correlation [coefficient of determination (r²) > 95 %] between the DNA G+C contents of 20 genes and their corresponding genomes. For each of the 157 prokaryotic genomes studied, the DNA G+C content of the 20 genes was used to determine a ‘calculated’ genome DNA G+C content (CGC) and this value was compared with the ‘real’ genome DNA G+C content (RGC). In order to select the most suitable gene for the determination of CGC values, we compared the r² and median mol% difference between CGC and RGC as well as the sensitivity of each gene to provide CGC values for prokaryotic genomes that differ by less than 5 mol% from their RGC. The highly conserved ftsY gene (median size 1144 nucleotides), a vertically inherited member of the GTPase superfamily, showed the highest r² value of 0·98, the smallest median mol% difference between CGC and RGC of 1·06 and a sensitivity of 100 %. Using ftsY DNA G+C content values, the CGC values of 100 genomes not included in the calculation of r² differed by less than 5 mol% from their RGC values. These data suggest that the genomic DNA G+C content of prokaryotes may be estimated easily and reliably from the ftsY gene sequence.

INTRODUCTION

The current taxonomic classification of prokaryotes is based on polyphasic taxonomy (Vandamme et al., 1996). This approach combines the genomic and phenotypic characteristics of a strain. The minimum amount of genomic information required for the description of a novel bacterial species must include its phylogenetic classification, DNA–DNA relatedness and the mol% G+C content of DNA. The current taxonomic classification of prokaryotes is based on polyphasic taxonomy (Vandamme et al., 1996). This approach combines the genomic and phenotypic characteristics of a strain. The minimum amount of genomic information required for the description of a novel bacterial species must include its phylogenetic classification, DNA–DNA relatedness and the mol% G+C content of DNA.

Abbreviations: CGC, calculated genome DNA G+C content; GGC, DNA G+C content from each essential single-copy gene; RGC, real genome DNA G+C content.

Tables detailing the prokaryotic species for which complete genome sequences are available in GenBank, the differences between RGC and CGC values obtained in the study for 100 prokaryotic genomes and a list of genes conserved in prokaryotic genomes in the COG database are available as supplementary data in IJSEM Online.

Previously, it has been suggested that micro-organisms showing more than 10 mol% difference in DNA G+C contents might not belong to the same genus and that 5 mol% is the common range found within a species (Goodfellow et al., 1997). Of the various methods available for the determination of DNA G+C content (De Ley, 1970; Ko et al., 1977; Marmur & Doty, 1962; Mesbah & Whitman, 1989; Owen et al., 1969; Schildkraut et al., 1962; Xu et al., 2000), the thermal denaturation temperature (Tm) method is most commonly used (Marmur & Doty, 1962). However, thermal denaturation is a time-consuming method that requires a large amount of DNA and lacks intra- and inter-laboratory reproducibility, and the Tm is calculated using a formula proposed by Mandel et al. (1970) which is not suitable for prokaryotes with very low or elevated DNA G+C contents (Ezaki et al., 1990).

In our laboratory, we have been studying the rpoB gene, encoding the β-subunit of the DNA-dependent RNA polymerase (Stackebrandt et al., 2002).
polymerase (RNAP) (Cramer, 2002; Murakami & Darst, 2003), for several years (Drancourt et al., 2004; Khamis et al., 2003, 2004; Mollet et al., 1997, 1998; Renesto et al., 2000, 2001a, b; Taillardat-Bisch et al., 2003). We have observed that, in prokaryotes, the DNA G+C content of the \textit{rpoB} gene correlates well with that of the genome. This led to the suggestion that the \textit{rpoB} gene might serve as a measure of genomic DNA G+C content. We investigated whether the genomic DNA G+C content of prokaryotes would correlate well with that of their genes. More specifically, our objective was to identify the most suitable gene which had a DNA G+C content that could be used as an acceptable (not differing by more than 5 mol% from that of the genome) measure of prokaryotic genomic DNA G+C content.

**METHODS**

**Study design.** The study design is summarized in Fig. 1.

**Identification of essential genes.** We focused our attention on universally conserved genes (i.e. genes present in all prokaryotes). For this, we searched the Clusters of Orthologous Groups (COG) database (as of 1 April 2005) containing reliably annotated genes of 50 complete prokaryotic genomes (Tatusov et al., 2001). Nucleotide sequences of the genes identified as universally conserved genes were then retrieved from the 206 prokaryotic genomes available in GenBank (Benson et al., 2005) as of 1 April 2005 (see Supplementary Tables S1 and S2 in IJSEM Online) using BLAST software (Altschul et al., 1997). Genes that were present as a single copy within all 206 genomes were selected from the 127 universally conserved genes.

**Correlation between gene and genomic DNA G+C contents.** For a given prokaryotic genome, the DNA G+C content from each essential single-copy gene (GGC) and each genome (‘real’ DNA G+C content, RGC) was determined from its nucleotide sequence using the EMBOSS software package (Olson, 2002). For correlation analysis, we analysed only 157 genomes, one genome per species. This was to eliminate the possibility of misinterpreting data that might have resulted from the overrepresentation of some species for which several strains have already been completely sequenced. Correlations between GGC and RGC were studied by means of scatter plots using EXCEL 2003 software (Microsoft). Tendency curves and coefficients of determination ($r^2$) between GGC and RGC were examined for each gene. For the final analyses, we retained only those universally conserved single-copy genes that exhibited an $r^2$ value of >95%.

We also determined a ‘calculated’ genomic DNA G+C content (CGC) for each gene and each species. This was inferred from the GGC using the tendency curve equation (Table 1). For each gene, the variable $y$ in the tendency curve equation represents GGC and $x$ represents GGC. The median mol% difference between CGC and RGC was calculated for each gene. In addition, the sensitivity (i.e. the probability that the RGC and CGC values of two species differ by more than 5%) of each gene was calculated. The best candidate gene was defined as the gene that showed the highest $r^2$ and sensitivity and the smallest median mol% difference.

**Statistical analysis.** The statistical significance of the correlations between GGC and RGC was determined using the $t$-test [$t = r(n-2)/(1-r^2)$, where $r$ is the correlation coefficient, $r^2$ is the coefficient of determination and $n$ represents the number of genomes studied] (Bailey, 1995). Contrasts in gene sensitivities were tested using Fisher’s exact test. All statistical tests were two-tailed. Observed differences were considered significant when $P<0.05$.

**Estimation of the suitability of the selected gene.** In order to examine the suitability of the selected gene for the estimation of prokaryotic genome DNA G+C content, we inferred the CGC from the GGC for each of the 49 prokaryotic strains that had not been used previously in gene selection. This was because the species of each of these prokaryotic strains had already been considered in the calculations of DNA G+C content correlation. We also tested 51 prokaryotic genomes that became available after our study had begun (Supplementary Table S2 in IJSEM Online).

**RESULTS**

**Selection of essential genes**

In the COG database, we identified 127 universally conserved genes present in each of the 50 prokaryotic genomes (Tatusov et al., 2001) (see Supplementary Table S3 in IJSEM Online). The nucleotide sequences of these 127 genes from each of the 206 available prokaryotic genomes were then retrieved from the GenBank database (Supplementary Tables S1 and S2 in IJSEM Online). Subsequently, we eliminated 70 of these genes, including \textit{rpoB} (Ishikawa et al., 2004), that were present in more than one copy within at least one genome. A total of 57 genes that were present as a single copy in all 206 genomes were chosen for further analyses.

**Correlation between GGC and RGC**

Among the 57 single-copy universally conserved genes, $r^2$ ranged from 0.73 for the \textit{rpmC} gene to 0.98 for the \textit{ftsY} gene (Supplementary Table S3 in IJSEM Online). For 20 of these genes, $r^2$ exceeded a value of 0.95 (Supplementary Table S3

Fig. 1. Study design.
Table 1. Suitability comparison of 20 essential genes for the estimation of genomic DNA G+C content

Median DNA G+C content difference is the difference between the real and calculated genome DNA G+C contents (mol%). Gene sensitivity was determined as the number of prokaryote species in which the difference between the real and calculated genomic G+C contents was <5% out of the 157 species tested.

<table>
<thead>
<tr>
<th>COG number</th>
<th>Gene</th>
<th>Tendency curve equation</th>
<th>Median DNA G+C content difference (mol%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG0012</td>
<td>–</td>
<td>y=1.0904x−5.5241</td>
<td>1.31</td>
<td>97.4</td>
</tr>
<tr>
<td>COG0016</td>
<td>pheS</td>
<td>y=1.0404x−3.2462</td>
<td>1.28</td>
<td>98.0</td>
</tr>
<tr>
<td>COG0052</td>
<td>rpsB</td>
<td>y=1.1309x−6.7165</td>
<td>1.58</td>
<td>96.8</td>
</tr>
<tr>
<td>COG0072</td>
<td>pheT</td>
<td>y=0.9378x+1.1912</td>
<td>1.09</td>
<td>98.0</td>
</tr>
<tr>
<td>COG0195</td>
<td>nusA</td>
<td>y=1.0862x−5.0726</td>
<td>1.25</td>
<td>96.1</td>
</tr>
<tr>
<td>COG0216</td>
<td>prfA</td>
<td>y=1.0782x−6.0966</td>
<td>1.24</td>
<td>97.4</td>
</tr>
<tr>
<td>COG0233</td>
<td>frr</td>
<td>y=1.1102x−4.118</td>
<td>1.77</td>
<td>94.2</td>
</tr>
<tr>
<td>COG0264</td>
<td>tf</td>
<td>y=1.0993x−5.498</td>
<td>1.49</td>
<td>93.6</td>
</tr>
<tr>
<td>COG0275</td>
<td>–</td>
<td>y=0.969x−0.5475</td>
<td>1.46</td>
<td>97.4</td>
</tr>
<tr>
<td>COG0319</td>
<td>–</td>
<td>y=0.9295x+3.6058</td>
<td>1.59</td>
<td>98.7</td>
</tr>
<tr>
<td>COG0336</td>
<td>trnD</td>
<td>y=1.0053x−2.5542</td>
<td>1.66</td>
<td>94.2</td>
</tr>
<tr>
<td>COG0359</td>
<td>rplI</td>
<td>y=1.0228x−0.524</td>
<td>1.69</td>
<td>96.1</td>
</tr>
<tr>
<td>COG0481</td>
<td>lepA</td>
<td>y=1.1714x−9.152</td>
<td>1.22</td>
<td>99.3</td>
</tr>
<tr>
<td>COG0525</td>
<td>valS</td>
<td>y=1.0339x−3.5434</td>
<td>1.75</td>
<td>97.4</td>
</tr>
<tr>
<td>COG0528</td>
<td>pyrH</td>
<td>y=1.1333x−8.132</td>
<td>1.44</td>
<td>93.6</td>
</tr>
<tr>
<td>COG0536</td>
<td>obg</td>
<td>y=1.0663x−6.1973</td>
<td>1.4</td>
<td>97.4</td>
</tr>
<tr>
<td>COG0552</td>
<td>ftsY</td>
<td>y=0.9509x+0.4351</td>
<td>1.06</td>
<td>100</td>
</tr>
<tr>
<td>COG0691</td>
<td>snpB</td>
<td>y=1.0558x−1.4047</td>
<td>1.6</td>
<td>93.6</td>
</tr>
<tr>
<td>COG0781</td>
<td>nusB</td>
<td>y=0.9082x+4.4571</td>
<td>1.29</td>
<td>98.7</td>
</tr>
<tr>
<td>COG0858</td>
<td>rbfA</td>
<td>y=0.9626x+2.4083</td>
<td>1.87</td>
<td>94.2</td>
</tr>
</tbody>
</table>

We chose the ftsY gene as the best candidate gene. It had the highest $r^2$ value (0.98; Fig. 2) and sensitivity (100%). Furthermore, it showed the smallest median mol% difference between RGC and CGC (1.06). The median size of the ftsY gene in prokaryotes was found to be 1144 nucleotides (range 669–2325 nucleotides).

Estimation of the suitability of the ftsY gene for the estimation of genomic DNA G+C content

The CGC [CGC = (0.9509 × GGC) + 0.4351] was inferred from the sequences of ftsY genes (Table 1). The CGC values of all 49 prokaryotic strains not used previously in the calculations of correlations differed from RGC values by less than 5 mol% (Supplementary Table S2 in IJSEM Online). Likewise, for all 51 species that were not available for the initial calculations, the difference between CGC and RGC was <5 mol% (see Supplementary Table S2 in IJSEM Online).

DISCUSSION

In this study, we demonstrate that the genomic DNA G+C content of prokaryotes can be estimated reliably from gene nucleotide sequences. The currently used methods for the determination of genomic DNA G+C content lack in IJSEM Online). These genes include pheS (COG0016), rpsB (COG0052), pheT (COG0072), nusA (COG0195), prfA (COG0216), frr (COG0233), tsf (COG0264), trnD (COG0336), rplI (COG0359), lepA (COG0481), valS (COG0525), pyrH (COG0528), obg (COG0536), ftsY (COG0552), snpB (COG0691), nusB (COG0781), rbfA (COG0858) and genes encoding a predicted GTPase (COG0012), a predicted S-adenosylmethionine-dependent methyltransferase involved in cell envelope biogenesis (COG0275) and a metal-dependent hydrolase (COG0319). For all these 20 genes, the correlations between GGC and RGC were statistically significant ($P < 10^{-6}$). The median mol% difference between RGC and CGC ranged from 1.06 for the ftsY gene to 1.87 for the rbfA gene (Table 1). The sensitivities of the 20 genes ranged from 93.6% (147/157) for the tsf, pyrH and snpB genes to 100% (157/157) for the ftsY gene (Table 1). The ftsY gene was significantly more sensitive (157/157) than each of the tsf, pyrH and snpB genes (147/157, $P < 10^{-5}$) and the frr, trnD and rbfA genes (148/157, $P = 0.01$). The lepA gene was significantly more sensitive (156/157) than the tsf, pyrH and snpB genes (147/157, $P = 0.01$) and the frr, trnD and rbfA genes (148/157, $P = 0.02$). Individually, both the predicted metal hydrolase (COG0319) and nusB (155/157) genes were significantly more sensitive than the tsf, pyrH and snpB genes (147/157, $P = 0.04$). All other comparisons provided $P$ values >0.05 and thus were considered non-significant.
reproducibility and inter-laboratory comparability. Also, the accuracy of the results may not always be optimal. For example, the genomic DNA G+C content of *Tropheryma whippelii* strain Twist-Marseille determined using conventional methods (59-4 mol%; La Scola et al., 2001) differed by more than 13 mol% from that obtained after the complete genome sequence of the strain had been sequenced (46-3 mol%; Raoult et al., 2003). This difference exceeds the 5 mol% regarded as an acceptable range for the DNA G+C contents of strains within a species (Goodfellow et al., 1997).

DNA G+C content is relatively constant in prokaryotic genomes, in particular in coding regions (Forsdyke & Mortimer, 2000; Sandberg et al., 2003), and correlates well with synonymous codon choice (Knight et al., 2001), amino acid usage (Lobry, 2005) and genomic signatures (Deschavanne et al., 1999). This prompted us to investigate whether gene sequences could be used to extrapolate values of genomic DNA G+C content. In our study, we chose genes that were conserved in all prokaryotes as a single copy (Supplementary Table S3 in IJSEM Online; Koonin, 2003), as the objective was to identify a gene that could be used for all prokaryotes and that had a CGC value that did not differ from the RGC value by more than 5 mol%. The *ftsY* gene emerged as the best candidate gene, exhibiting the highest coefficient of determination between GGC and RGC, the smallest median mol% difference between CGC and RGC and a sensitivity of 100 %. In prokaryotes, the *ftsY* gene is present as a single copy (Cao & Saier, 2003) and there is no evidence to suggest horizontal transfer of this gene (Caldon et al., 2001; Gribaldo & Cammarano, 1998). The CGC values inferred, as above, from the GGC value for this gene were within a range of 5 mol% from the RGC for 100 prokaryotic strains that had not been previously included in the determination of correlation. Furthermore, we observed that the CGC values obtained for such prokaryotic species as *Campylobacter jejuni* (31-0 and 31-2 mol%), *Ureaplasma urealyticum* (25-9 mol%) or *T. whippelii* (49-3 mol%) were closer to their RGC [30-3 and 30-5 (Owen, 1983), 25-5 (Razin, 1985) and 46-3 mol% (Raoult et al., 2003), respectively] than the DNA G+C content values obtained using traditional methods [31-5 (Owen, 1983), 26-9–28 (Razin, 1985) and 59-4 mol% (La Scola et al., 2001), respectively]. Finally, the *ftsY* gene has a median size of 1144 nucleotides and the genome sequences available cover all phylogenetic prokaryote clades. This makes it easy to select primers from phylogenetically close genomes and makes this gene easy to sequence. Thus, based on our results and the characteristics of the gene, we believe that the *ftsY* gene offers an accurate way of estimating genomic DNA G+C content.

Compared with conventional methods, our method is rapid, less labour-intensive and reproducible. Our method also requires a smaller quantity of DNA than required for the conventional methods and the results are easily comparable between laboratories. It may even be suitable for uncultured bacteria. In addition, unlike recently described methods such as those using a LightCycler thermal cycler (Xu et al., 2000), our method does not require the use of any specific equipment as sequencing facilities are now available in many academic and non-academic laboratories. Such laboratories could even be sent PCR products from remote areas. Furthermore, our method is not affected by differences in affinity to the SYBR Green nucleic acid stain between prokaryotic chromosomes. In summary, the use of the *ftsY* gene GGC is a rapid and reliable means of estimating genomic DNA G+C content that may easily be used by any laboratory.

**ACKNOWLEDGEMENTS**

The authors thank Professor Michel Drancourt for helpful discussions and Dr Mohamad Khan for reviewing the English.

**REFERENCES**


