Multiple Gene Sequence Analysis Using Genes of The Bacterial DNA Repair Pathway

Miguel Rotelok Neto¹, Carolina Weigert Galvão¹, Leonardo Magalhães Cruz², Dieval Guizelini², Leilane Caline Silva², Jarem Raul Garcia¹ and Rafael Mazer Etto¹*
¹Departamento de Biologia Estrutural, Molecular e Genética; Universidade Estadual de Ponta Grossa; Ponta Grossa - PR - Brasil. ²Departamento de Bioquímica e Biologia Molecular; Universidade Federal do Paraná; Curitiba - PR - Brasil

ABSTRACT

The ability to recognize and repair abnormal DNA structures is common to all forms of life. Physiological studies and genomic sequencing of a variety of bacterial species have identified an incredible diversity of DNA repair pathways. Despite the amount of available genes in public database, the usual method to place genomes in a taxonomic context is based mainly on the 16S rRNA or housekeeping genes. Thus, the relationships among genomes remain poorly understood. In this work, an approach of multiple gene sequence analysis based on genes of DNA repair pathway was used to compare bacterial genomes. Housekeeping and DNA repair genes were searched in 872 completely sequenced bacterial genomes. Seven DNA repair and housekeeping genes from distinct metabolic pathways were selected, aligned, edited and concatenated head-to-tail to form a super-gene. Results showed that the multiple gene sequence analysis using DNA repair genes had better resolution at class level than the housekeeping genes. As housekeeping genes, the DNA repair genes were advantageous to separate bacterial groups at low taxonomic levels and also sensitive to genes derived from horizontal transfer.

Key words: Bacterial genome, multiple gene sequence analysis, taxonomic assignment, DNA repair genes, HGT

INTRODUCTION

DNA repair processes are indispensable for maintaining the integrity of genetic information in all organisms. Environmental agents such as chemicals, UV light, and ionizing radiation, as well as errors in DNA metabolism challenge the chemical structure and stability of the genome. These threats lead to a variety of alternations in the normal DNA structure such as single- and double-strand breaks, chemically modified bases, abasic sites, inter- and intra-strand cross-links, and base-pairing mismatches. Given this diversity of threats and their effects, it is not surprising that there is a corresponding diversity of DNA repair pathways (Eisen et al. 1999). The diversity of specificity, functions, and complexity of repair pathways is best understood by comparing the mechanisms of action among pathways. Such comparisons are simplified by the division of repair processes into three major classes based on general mechanism of action: (1) direct repair, in which abnormalities are chemically reversed; (2) recombinational repair, in which homologous recombination is used to repair abnormalities; and (3) excision repair, in which a section of the DNA strand containing an abnormality is removed and a repair patch is synthesized using the intact strand as a template. Within each of these classes, there are multiple types and sometimes, even subtypes of repair (Eisen et al. 1999). Despite the high diversity of DNA repair genes and its importance for
maintaining the integrity of genetic information, most studies use 16S rRNA or others housekeeping genes to establish relationships among the bacterial genomes.

The multiple gene sequence analysis has to be used to understand the relationships among genomes. There are many ways of multiple gene sequence analysis (de Queiroz et al. 1995; Huelsenbeck et al. 1996; Yang 1996; Nei et al. 2001; Suchard et al. 2003), but two fundamentally different ways are more often considered. In one, the analyses are done after the gene sequences are concatenated head-to-tail to form a super-gene alignment. In the other, the analyses are done separately for each gene and the resulting dendrograms are used to generate a consensus dendrogram. The dendrogram based on the concatenation approach has been more used because of its presumed statistical advantages and accurate generation of dendrograms, even when the concatenated sequences have evolved with very different substitution patterns (Gadagkar et al. 2005; Lang et al. 2013).

The evolution of species has been largely assumed to be strictly bifurcating, a tree-like process (Felsenstein 2004). However, many argue that such a tree-like depiction of the history of prokaryote species is not valid because of the high incidences of horizontal gene transfer (HGT), which obscures a vertical line of descent (Hilario and Gogarten 1993; Baptiste et al. 2005; Dagan et al. 2008; Koonin and Wolf 2008). In general, HGT tends to occur between the closely related organisms, more extensively between the genera and species than at the level of phylum or family (Staley 2009). In this work, bacterial genomes were compared using multiple gene sequence analysis of genes involved on DNA repair pathway. To validate this approach, the same methodology of analysis was applied also on the classical housekeeping genes. Both analyses were compared with the dendrogram based on 16S rRNA gene. The influence of possible HGT on the super-gene alignments of ubiquitous housekeeping or DNA repair genes from many completely sequenced genomes was evaluated.

MATERIAL AND METHODS

Selection of bacterial genes
The 52 housekeeping genes commonly used in the multilocus sequence typing approach and the 60 genes that encode proteins involved in DNA repair were searched in 872 bacterial genomes completely sequenced from National Center for Biotechnology Information (NCBI) database in June 2010 using the GenBank Database Explorer program (GDE) (Guizelini 2010). The ubiquity and presence in distinct pathways were the criteria used to choose the genes. Seven genes of the DNA repair and seven housekeeping genes were selected (Table 1). The genes that had more than one copy per genome were compared to the NCBI database using the BLAST tool and the copy that had the highest identity score was chosen.

Alignment and edition
Individual gene alignments were carried out with Clustal W (Larkin 2007), using the translated sequences and applying a gap opening penalty of 3.0 and a gap extension penalty of 1.8 as suggested by Hall (2011). After alignment, the amino acid sequences were converted again to nucleotides and edited with GBlocks (Catresana 2000) to remove highly divergent regions and to reduce the influence of the length of gene in the dendrogram. Fragments of approximately 500 bp were generated. The parameters used in the program GBlocks are shown in Table 1.

Table 1 - Parameter of the GBlocks program used for each gene.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Minimum block length (bp)</th>
<th>Permission of Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>adk</td>
<td>2</td>
<td>none</td>
</tr>
<tr>
<td>fusA</td>
<td>100</td>
<td>none</td>
</tr>
<tr>
<td>guaA</td>
<td>38</td>
<td>none</td>
</tr>
<tr>
<td>gyrB</td>
<td>70</td>
<td>none</td>
</tr>
<tr>
<td>leuS</td>
<td>140</td>
<td>none</td>
</tr>
<tr>
<td>rplB</td>
<td>10</td>
<td>none</td>
</tr>
<tr>
<td>rpoB</td>
<td>40</td>
<td>none</td>
</tr>
<tr>
<td>dnaE</td>
<td>220</td>
<td>50%</td>
</tr>
<tr>
<td>mutS</td>
<td>2</td>
<td>none</td>
</tr>
<tr>
<td>nth</td>
<td>5</td>
<td>none</td>
</tr>
<tr>
<td>recA</td>
<td>144</td>
<td>none</td>
</tr>
<tr>
<td>ruvB</td>
<td>85</td>
<td>none</td>
</tr>
<tr>
<td>uvrC</td>
<td>50</td>
<td>none</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>2</td>
<td>50%</td>
</tr>
</tbody>
</table>

*The gene showed low homology, so in this case was permitted the presence of 50% gaps in the sequences. The chosen parameters generated sequences of approximately 500 bp, except for 16S rDNA which was full used. Thus, it was possible to minimize the effect of the length of genes in analysis.
Concatenation of genes to form super-gene alignments
The gene sequences edited were concatenated head-to-tail to form a super-gene alignment. The order of super-DNA-repair-gene alignment was: dnaE, dut, mutS, nth, recA, ruvB and uvrC. The order of super-housekeeping-gene alignment was: adk, fusA, guaA, gyrB, leuS, rplB and rpoB.

Dendrogram of super-gene alignments
The aligned concatenated sequence of DNA repair and housekeeping genes were analyzed by MEGA 5.05 program (Tamura et al. 2011) to choose the best nucleotide substitution model. The dendrograms were constructed by Maximum Likelihood (ML). The statistical test used 500 bootstrap replications and the cutoff value was set at 65 Bootstrap values.

Dendrogram of 16S rRNA gene
A dendrogram based only on the 16S rRNA was constructed and compared to the dendrogram of the super-gene alignment. The 16S rRNA gene sequences were searched in the NCBI database also with the help of the GDE program (Guizelini 2010). The 16S rRNA sequences were aligned with Clustal W using default parameters (Larkin 2007). Alignments were edited with GBLOCKS.

Prediction of putative Horizontal Gene Transfer
The Alien Hunter program was used to check the possibility of horizontal gene transfer (HGT). The program analyzed the genome of the organism, predicted putative HGT events with the implementation of Interpolated Variable Order Motifs (Vernikos 2006) and marked them with a color code compatible with the Artemis genomic analysis program (Rutherford 2000).

RESULTS AND DISCUSSION
The results of DNA repair genes in the 872 completely sequenced bacterial genomes from NCBI, confirmed that the recA was the main DNA repair gene, because it was found in every bacterial genomes. The universality of RecA indicated both that it was an ancient gene and that its activity was irreplaceable. Despite the existence of more than 60 genes involved in DNA repair pathway, only seven genes were ubiquitous in 182 bacterial genomes. Although the DNA repair genes were indispensable for maintaining the genetic information, it was surprising that only seven genes were omnipresent in 20% bacterial genomes analyzed. Since the analyzed bacteria occupy different habitat, this result suggested that the environmental threats were responsible for maintaining the DNA repair genes diversity in bacterial genomes.

Based on an approach of gene sequences concatenation, a comparison was made between the bacterial genomes present in the NCBI database using super-gene alignments. Seven housekeeping and seven DNA repair genes present in 182 bacterial genomes were utilized, corresponding to seven phyla, fourteen classes and thirty one orders. Since Gammaproteobacteria is the class, which has more completely sequenced genomes deposited in databases (Lagensen 2010), it was the most representative in this analysis.

The dendrogram based on the concatenated sequences of housekeeping or repair genes was more efficient to distinct the bacterial groups at low taxonomic levels. On the other hand, the 16S rRNA gene dendrogram showed better resolution at high taxonomic levels. The Escherichia and Shigella genera were grouped into the same branch in the 16S rRNA dendrogram (Fig. 1) but they were separated in the housekeeping (Fig. 2) and in the DNA repair (Fig. 3) genes super-alignment dendrogram. In contrast, the housekeeping genes dendrogram showed incongruities at the class level. Some alpha proteobacteria clustered with epsilonproteobacteria and two species of Bacteroidetes grouped with Chlamydiae (Fig. 2). These clustering incongruities were not showed on the DNA-repair genes dendrogram (Fig. 3). The housekeeping or the DNA repair genes dendrograms grouped taxonomically distant species. The Brachyspira hyodysenteriae from Epsilonchaetes phylum was grouped with Clostridium botulinum from Firmicutes phylum. B. hyodysenteriae genome analysis using the Alien Hunter program suggested that rpoB, fusA, dut and dnaE genes could have been acquired via horizontal transfer (Fig. 4). Both B. hyodysenteriae and C. botulinum inhabited the large intestine of pigs (Bellgard et al. 2009; Hafströmm et al. 2011). It justified the dendrogram results of housekeeping or repair genes concatenated sequences (Figs. 2 and 3). The relationship between the environmental and genetic similarity was observed also in the Bacteroidetes group. The Salinibacter ruber bacterium was distant from other bacteria of...
the same phylum Bacteroidetes (Cytophaga hutchinsonii and Gramella forsetii). However, none of the genes used in the construction of the super alignment was identified as coming from the horizontal transfer by Alien Hunter program (data not shown). The result observed by super-gene alignment could be attributed to evolutionary convergence with species of Halobacteria class of the Euryarchaeota phylum that were common in the habitat of S. ruber (Mongodin 2005). In the dendrogram based on Bacteria and Archaea 16S rRNA gene, it was possible to see that S. ruber grouped with the archaea Haloarcula marismortui with a strong bootstrap support and did not group with other Bacteroidetes species (Fig. 5).

Figure 1 - Maximum Likelihood tree of bacterial 16S rRNA gene. The dendrogram was estimated by neighbor joining using the k2+G model, with MEGA 5.05 program. Support values are calculated from 500 bootstrap replicates, with a bootstrap cutoff of 65. The colored symbols represent different taxonomic groups. This representation is a radial cladogram, in which branch length is not proportional to time, and some branches may be elongated so that the names of the taxa appear on the circumference of the circle.
Figure 2 - Maximum Likelihood tree of concatenated housekeeping gene. Dendrogram inferred from a concatenated, partitioned alignment of 7 housekeeping genes. The dendrogram was estimated by neighbor joining using the k2+G model, with MEGA 5.05 program. Support values were calculated from 500 rapid bootstrap replicates, with a bootstrap cutoff of 65. The colored symbols represent different taxonomic groups. This representation is a radial cladogram, in which branch length is not proportional to time, and some branches may be elongated so that the names of the taxa appear on the circumference of the circle.
Figure 3 - Maximum Likelihood tree of concatenated DNA repair gene. Dendrogram inferred from a concatenated, partitioned alignment of 7 DNA repair gene. The dendrograms was estimated by neighbor joining using the k2+G model, with MEGA 5.05 program. Support values were calculated from 500 rapid bootstrap replicates, with a bootstrap cutoff of 65. The colored symbols represent different taxonomic groups. This representation is a radial cladogram, in which branch length is not proportional to time, and some branches may be elongated so that the names of the taxa appear on the circumference of the circle.

In all dendrograms, the thermophilic bacterium *Pelotomaculum thermopropionicum*, present in anaerobic reactors (Kosaka et al. 2008), and the photosynthetic bacteria *Heliobacterium modesticaldum* (Sattley et al. 2008) were separated from the *Firmicutes* phylum cluster. Alien Hunter genomic analyses of these two species suggested that *P. thermopropionicum* rpoB, fusA, rplB and mutS genes (Fig. 6) and *H. modesticaldum* dut and dnaE genes (Fig. 7), used in the analyzes, could have been acquired via horizontal transfer.
Figure 4 - Chromosomal islands in *Brachyspira hyodysenteriae*. Putative horizontal transferred genes are mapped in the genome in different color tones according to the Alien Hunter program scores. Dark Red = highest HGT scores and Light red = lowest HGT scores. The target genes used in DNA repair and housekeeping super-gene construction are indicated.

Figure 5 - Maximum Likelihood tree of bacterial and archaeal 16S rRNA gene. The dendrogram was estimated by neighbor joining using the k2+G model, with MEGA 5.05 program. Support values are calculated from 500 bootstrap replicates, with a bootstrap cutoff of 65. The colored symbols represent different taxonomic groups. This representation is a radial cladogram, in which branch length is not proportional to time, and some branches may be elongated so that the names of the taxa appear on the circumference of the circle.
Figure 6 - Chromosomal islands in *Pelotomaculum thermopropionicus*. Putative horizontal transferred genes are mapped in the genome in different color tones according to the Alien Hunter program scores. Dark Red = highest scores and Light red = lowest scores. The target genes used in DNA repair and housekeeping super-gene construction are indicated.

Figure 7 - Chromosomal islands in *Helio bacterium modesticaldum*. Putative horizontal transferred genes are mapped in the genome in different color tones according to the Alien Hunter program scores. Dark Red = highest HGT scores and Light red = lowest HGT scores. The target genes used in DNA repair and housekeeping super-gene construction are indicated.

In conclusion, the results of DNA repair genes in 872 bacterial genomes showed that few genes were present in most species analyzed. The diversity of DNA repair genes seemed to be a result of different environmental threats to which bacterial genomes were exposed. Nevertheless, the
multiple gene sequence analysis using DNA repair genes showed better resolution at class level than the multiple gene sequence analysis of housekeeping genes. As housekeeping genes, the dendrogram based on DNA repair genes showed to be an accurate approach to distinct bacterial groups at low taxonomic levels and also sensitive to genes derived from horizontal transfer. These results also suggested that the genes of DNA repair pathway could substitute the classical housekeeping genes on the multiple gene sequence analysis.

ACKNOWLEDGMENTS

We would like to thank the Brazilian Program of National Institutes of Science and Technology-INCT/Brazilian Research Council-CNPq/MCT, CAPES and Fundação Araucária of Paraná State for the financial support.

REFERENCES


Received: December 12, 2014; Accepted: March 01, 2015.