Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*

Kishore R. Sakharkar,1 Meena K. Sakharkar,2 Chandra Verma1 and Vincent T. K. Chow3

Correspondence
Vincent T. K. Chow
micctk@nus.edu.sg

1BioInformatics Institute, Matrix, 30 Biopolis Street, Singapore
2Nanyang Centre for Supercomputing and Visualization, School of Mechanical and Production Engineering, Nanyang Technological University, Singapore
3Human Genome Laboratory, Department of Microbiology, Faculty of Medicine, National University of Singapore, Kent Ridge, Singapore 117597

Overlapping genes have been proposed as a means of achieving genome reduction by compressing the maximum amount of information in limited sequence space. In this report, comparative analyses of the overlapping genes of genomes of nine bacteria with different lifestyles were performed. The results clearly suggest that overlapping genes may be a result of evolutionary pressure to minimize genome size. The genomes of two closely related obligatory intracellular parasites – *Rickettsia prowazekii* and *Rickettsia conorii* – were investigated further. Detailed analyses of these two genomes revealed that mutations at the ends of coding regions and elimination of intergenic DNA are the main forces that determine overlapping of genes.

INTRODUCTION

Overlapping genes are a common occurrence in prokaryotic genomes (Normark et al., 1983). In overlapping genes, the same DNA sequence encodes two proteins using different reading frames. The evolution of overlapping genes has been approached theoretically and empirically (Sander & Schulz, 1979; Smith & Waterman, 1981; Pavesi et al., 1997). Clark et al. (2001) suggested that overlapping genes occur as a result of mutational bias towards deletion. Overlapping genes are more conserved between species than non-overlapping genes (Lipman, 1997; Yelin et al., 2003), mostly because a mutation in the overlapping region causes changes in both genes, and selection against such mutations should therefore be stronger. Miyata & Yasunaga (1978) reported that rates of evolution are slower in overlapping genes. Keese & Gibbs (1992) suspected that overlapping genes arise as a result of overprinting – a process of generating new genes from pre-existing nucleotide sequences (when more than one out-of-phase reading frame may lie in a single nucleotide sequence). The role of the genetic code in generating an open reading frame (ORF) inside a coding region has been elaborated (Cebrat et al., 1997). Overlap is thought to be important as (1) a means of compressing the maximum amount of information into short sequences of structural genes and may be a result of evolutionary pressure to minimize genome size and increase the density of genetic information; and (2) as a mechanism for regulating gene expression through translational coupling of functionally related polypeptides (Normark et al., 1983; Chen et al., 1990; Inokuchi et al., 2000). Some ORFs belong to clusters of more than two overlapping ORFs (Normark et al., 1983; Krakauer, 2000). Overlapping genes may evolve as a result of the extension of an ORF caused by a switch to an upstream initiation codon, substitutions in initiation or termination codons, and deletions and frameshifts that eliminate initiation or termination codons (Rogozin et al., 2002). Comparative analysis of the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* showed that most overlapping genes are generated by mutations at the end of coding regions (Fukuda et al., 1999). The analysis revealed that the loss of a stop codon causes the gene to elongate to the next stop codon. Recently, it was suggested that evolution of overlapping genes occurs at a universal mutation rate across bacterial genomes (Fukuda et al., 2003). However, little is known about the origin, evolution and cross-species conservation of overlapping genes, and about the frequency and genome-wide distribution of overlapping genes in different genomes. We therefore sought to more closely examine how the relative usage of the genome...
changes with genome size in micro-organisms with similar lifestyles. Here, we report the number of overlapping genes in genomes of bacteria with different lifestyles, and compare overlapping genes in two completely sequenced genomes of obligatory intracellular parasites for humans – Rickettsia prowazekii and Rickettsia conorii. This analysis should help explain the role of overlapping genes in obligatory intracellular parasites of humans, and shed light on their evolution and dynamics in bacterial genomes in relation to genome reduction. Analysis of overlapping genes is hampered by sequencing and annotation errors present in genomes, and by the limitations of gene-finding algorithms to handle multiple reading frames (Burge & Karlin, 1998). The authors thus advise caution in drawing inferences from these data.

METHODS

The genome sequences of five obligatory intracellular parasites of humans [Chlamydia pneumoniae (Kalman et al., 1999), Chlamydia trachomatis (Stephens et al., 1998), Mycobacterium leprae (Cole et al., 2001), R. prowazekii (Andersson et al., 1998) and R. conorii (Ogata et al., 2001)], two minimal genomes [Mycoplasma genitalium (Fraser et al., 1995) and Mycoplasma pneumoniae (Himmelreich et al., 1996)], one endosymbiotic bacterium (Buchnera aphidicola; Shigenobu et al., 2000) and one free-living bacterium (Clostridium perfringens; Shimizu et al., 2002) were downloaded from the National Center for Biotechnology Information website (ftp://ftp.ncbi.nlm.nih.gov/genomes/bacteria). The coding sequence (CDS) annotation feature was used for extracting the genes showing overlap. We define overlapping genes as pairs of adjacent genes whose coding regions partly or completely overlap (Table 1). We calculated the percentage of genes that are present as overlapping gene pairs in each genome, and ascertained the direction of their overlap. We selected the two rickettsial species for further detailed analysis on the possible mechanisms for the genes to overlap. We classified the overlapping gene pairs of the rickettsial species into four categories: (1) gene pairs that occur as overlapping genes in both genomes; (2) gene pairs that overlap in both genomes with different numbers of overlapping nucleotides are tabulated for each category (Supplementary Tables A, B, C, D and E, available in IJSEM Online). The gene pairs that are homologous in one rickettsial species but do not exactly overlap or occur as split genes in another species (either next to each other or at some distance) were manually examined in detail to identify the possible reasons of their emergence during the course of evolution. As there is a high possibility that incorrectly annotated ORFs may be included in the genomic data, we removed ‘non-genuine ORFs’ from our detailed analysis. We define ‘non-genuine ORFs’ as those that are annotated as ‘hypothetical’ or ‘unknown’ in the genome databases.

RESULTS AND DISCUSSION

Genomes and overlapping genes

Reductive evolution is documented in obligatory intracellular parasites as genes become inactivated once their functions are no longer required in highly specialized niches (Andersson & Andersson, 1999). This process may have naturally defined the minimal gene set for a pathogenic organism based on lifestyle. Very small genomes have adopted obligate associations with the host.

In order to further investigate the retention of overlapping genes in organisms with different and similar lifestyles, we determined the proportion of genomes represented by overlapping genes in five obligatory intracellular parasites, two reduced genomes, one endosymbiont and one free-living bacterium (Table 1). It was noteworthy that a substantial portion of the genomes is represented by overlapping genes in all the organisms, clearly suggesting an important role for overlapping gene pairs in bacterial genomes. Interestingly, Mycoplasma genitalium has the smallest genome but the largest percentage representation of overlapping gene pairs (Fig. 1). Mycoplasma pneumoniae, the closest relative of Mycoplasma genitalium, also shows a greater proportion of overlapping gene pairs.

<table>
<thead>
<tr>
<th>Bacterial species</th>
<th>Genome size (Mb)</th>
<th>Total no. of ORFs</th>
<th>No. of olp gene pairs</th>
<th>No. of 'genuine' olp gene pairs</th>
<th>Orientation of overlaps (all genes)</th>
<th>Orientation of overlaps (‘genuine’ genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rickettsia prowazekii</td>
<td>1.11</td>
<td>901</td>
<td>111</td>
<td>40</td>
<td>+/−/−</td>
<td>−−</td>
</tr>
<tr>
<td>Rickettsia conorii</td>
<td>1.27</td>
<td>1407</td>
<td>225</td>
<td>66</td>
<td>87/110</td>
<td>17</td>
</tr>
<tr>
<td>Chlamydia trachomatis</td>
<td>1.018</td>
<td>975</td>
<td>173</td>
<td>101</td>
<td>68/68</td>
<td>34</td>
</tr>
<tr>
<td>Chlamydia pneumoniae</td>
<td>1.201</td>
<td>1092</td>
<td>199</td>
<td>107</td>
<td>72/78</td>
<td>44</td>
</tr>
<tr>
<td>Mycobacterium leprae</td>
<td>3.191</td>
<td>2765</td>
<td>341</td>
<td>179</td>
<td>169/122</td>
<td>39</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>0.566</td>
<td>523</td>
<td>179</td>
<td>81</td>
<td>95/53</td>
<td>27</td>
</tr>
<tr>
<td>Mycoplasma pneumoniae</td>
<td>0.797</td>
<td>729</td>
<td>209</td>
<td>121</td>
<td>117/67</td>
<td>22</td>
</tr>
<tr>
<td>Buchnera aphidicola</td>
<td>0.625</td>
<td>596</td>
<td>49</td>
<td>40</td>
<td>24/20</td>
<td>5</td>
</tr>
<tr>
<td>Clostridium perfringens</td>
<td>2.960</td>
<td>2855</td>
<td>174</td>
<td>89</td>
<td>78/90</td>
<td>5</td>
</tr>
</tbody>
</table>
Obligatory intracellular parasites follow this trend, thereby endorsing the fact that overlapping genes are a means of compressing the maximum amount of information into a short sequence. Mycobacterium leprae is an exceptional facultative intracellular parasite that has a higher proportion of overlapping genes compared to the free-living anaerobe Clostridium perfringens. This may be explained by the notion that Mycobacterium leprae is still undergoing downsizing and genome reduction, as it is often considered a genome ‘in decay’. This explanation is supported by the fact that Mycobacterium leprae has the maximum number of pseudogenes (>1000) compared to only 12 pseudogenes in R. prowazekii, the obligatory intracellular parasite with the most extensive genome degradation (Fig. 1). These observations clearly implicate the role of overlapping genes and their contribution to genome reduction.

When we correlated the genome sizes of the nine bacterial species with the numbers of overlapping genes, a weak correlation of 0.56 was found. Inclusion of Escherichia coli in our data raised the correlation to 0.86. Fukuda et al. (2003) reported a complete analysis of overlapping genes and genome sizes of 50 prokaryotes, and found a good correlation between the number of overlapping genes and genome size. Thus, the niche and lifestyle of the organism also play an important role in determining the proportion of genes showing overlapping configuration, and may be considered as constraints that favour smaller genomes. In particular, these adaptive pressures explain how a given quantity of information can be represented by a relatively small message.

Overlapping gene pairs can assume one of three structures, namely, ‘convergent’ (→←), ‘unidirectional’ (→→) or ‘divergent’ (←→) (Rogozin et al., 2002). A significant proportion (>90%) of overlapping gene pairs in all the genomes were identified to be unidirectional. These results support the earlier hypothesis by Eyre-Walker (1995) that most overlapping gene pairs have unidirectional structure (Table 1). The frequent occurrence of the unidirectional overlapping structure probably reflects the commonest orientation of adjacent genes in the chromosomes, as prokaryotic genes are often organized into operons or clusters of genes that are transcribed together. Since all genes in an operon must be transcribed in the same direction, this organization will be reflected by a tendency for adjacent genes to have the same orientation. Fewer gene pairs have the two inverted orientations (←→ and ↔→). The lower proportion of the divergent structure may be attributed to the evolutionary constraints at the 5′ end of the gene and the upstream region, which incorporate essential structures such as promoters. In addition, a frameshift mutation at the 5′ end may destroy the entire gene. The unidirectional and convergent structures are more easily formed due to the loss of stop codons or a frameshift. These results concur with those of Rogozin et al. (2002) and Fukuda et al. (2003), and highlight the fact that gene orientation, genome reduction and evolutionary constraints work together during the organism’s adaptation in its niche.

**Comparative study of overlapping genes in R. prowazekii and R. conorii**

Gene pairs that occur as overlapping genes in both genomes. R. prowazekii is thought to have essentially appeared as a subset of R. conorii 40 to 80 million years ago. One hundred and thirty seven genes of R. conorii do not have any sequence similarity with the R. prowazekii genome (Ogata et al., 2001). Supplementary Table A (in IJSEM Online) summarizes the overlapping gene pairs in two genomes – R. prowazekii and R. conorii. The overlapping gene pairs are unidirectional, and are found on the same strand in both the genomes except for RP884/RP885 (↔←) and RC1373/RC1374 (→→). The number of unidirectional overlapping gene pairs is more on the leading strand than on the lagging strand for both the
To overlap or not to overlap. Although there are six gene pairs that are overlapping in *R. prowazekii*, these are split in *R. conorii*. However, there are only three gene pairs that are overlapping in *R. conorii*, but are split in *R. prowazekii*, which has a smaller genome than *R. conorii* (Table 1). These differences support the notion that overlapping genes may be a means of compressing the maximum amount of information into the available short sequence space, and may be a result of evolutionary pressure to minimize genome size. It was observed that overlapping genes are generated due to loss of a stop codon or start codon of either gene that results in extension of the 3′ end or reassignment of the start codon. This can happen as a result of deletion of the stop/start codons, point mutations at the stop/start codons or frame shift anywhere in the coding region. The results are elaborated below.

**Gene pairs that overlap in *R. prowazekii* but are split in *R. conorii*.** Six gene pairs overlap in *R. prowazekii*, but are non-overlapping or split in *R. conorii*. Out of these, two are present with zero intergenic distance between them in *R. conorii*, and have 4 bp overlap in *R. prowazekii* (Supplementary Table C, in IJSEM Online). However, four of them are overlapping in *R. prowazekii*, and have intergenic distances ranging from 3 to 188 bp in *R. conorii* (Supplementary Table D, in IJSEM Online).

**Gene pairs that overlap in *R. conorii* but are split in *R. prowazekii*.** Three gene pairs were identified as overlapping in *R. conorii*, but as split genes in *R. prowazekii*. The intergenic distance for these genes ranges from 1 to 40 bp. The possible causes for their emergence are elaborated in Supplementary Table E (in IJSEM Online).

**Conclusion**

Whole genome sequencing of micro-organisms is providing an opportunity for computer-based genetic analysis that allows us to highlight important features such as overlapping genes in the genomes. From our analysis, mutations at the ends of the coding region are the main force that determines gene overlaps. It also appears that gene overlaps arise from the reduction or elimination of intergenic regions caused by mutational bias towards deletion that helps in genome compression while retaining information content. Furthermore, most of the overlapping genes are not mutually exclusive in function. These studies thus emphasize that there is substantial plasticity among obligatory intracellular parasites, and that overlapping genes facilitate genome reduction and functional coupling.

**REFERENCES**


