

Strategies for Genome Reduction in Microbial Genomes

Kishore R. Sakharkar Vincent T.K. Chow
ksakharkar@gmail.com micctk@nus.edu.sg

Human Genome Laboratory, Department of Microbiology, Yong Loo Lin School of
Medicine, National University of Singapore, Kent Ridge, Singapore

Abstract

Niche dependent differential gene loss and overlapping genes have been proposed as means of achieving genome reduction by retaining indispensable genes and compressing maximum amount of information in available sequence space. Herein, we analyzed the differential gene loss and overlapping genes in bacterial genomes with different lifestyles. Our results clearly suggest that gene loss and overlapping genes could be a result of evolutionary pressure to minimize genome size. Comparative analysis of the genomes shows that the genomes display marked similarities in patterns of protein length and frequency. It is clear from our analysis that habitat is a major factor contributing to genome reduction. These comparisons increase our knowledge of the forces that drive the extreme specialization of the bacteria and its association to the host.

Keywords: overlapping genes, genome reduction, minimal genomes, obligatory intracellular parasites, overprinting

1 Introduction

Obligatory intracellular parasites have small genomes and a tendency towards further genome reduction [1]. There have been many reports on the evolution of these bacteria from larger genomes by genome deterioration [4]. Zomorodipour and Andersson [15] have provided examples of reductive convergent evolution in the genomes of *Rickettsia prowazekii* and *Chlamydia trachomatis* and have associated it with metabolic parasitism in response to intracellular habitat [15]. Gene degradation is a common feature of obligatory intracellular parasites targeting overlapping subsets of potentially dispensable genes while adapting to the selective pressures of different niches [2, 12]. Therefore, genes found as multiple copies may outline their specific adaptations [3].

Also, overlapping genes are a common occurrence in prokaryotic genomes [9]. Overlap is thought to be important as: (1) a means of compressing a maximum amount of information into short sequences of structural genes and could 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 [5, 7, 9]. Recently, we reported on the overlapping genes in two closely related obligatory intracellular parasites – *Rickettsia prowazekii* and *Rickettsia conorii* and reported that mutations at the end of coding regions and elimination of intergenic DNA are the main forces that determine the overlap [13]. Still, 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.

The flux, streamlining and elimination of genes in genomes of obligate intracellular parasitic species is an ongoing process. Thus, obligatory intracellular parasitism is an excellent model to study how bacteria exploit their host cell's functions and understand the strategies the bacteria 'embark on' to reduce genome size while maintaining minimal sets of genes for its existence in an intracellular environment or a niche.

Complete genome sequence is available for 16 eubacterial obligate intracellular parasites that are pathogenic for humans. This study is an attempt to understand the role of differential gene loss and overlapping genes, as recourse to genome reduction and understand the dynamics of microbial genomes addressing the critical questions concerning the size, the processes and the content of genes that are involved in such processes. This may shed light on the differential loss of genes in response to life-style. Knowledge in this area may contribute to elucidating the fundamental mechanism of host-pathogen interaction specifically with reference to identification of determinants responsible for host specificity and virulence, disease pathogenesis and identification of new targets for vaccine and drug design. This analysis will also help us examine how the relative usage of the genome changes with genome size in organisms with same or different life-styles.

2 Method and Results

The genome sequences for the 16 obligatory intracellular parasites – *C. trachomatis* (NC_000117); *C. pneumoniae* CWL029 (NC_000922); *R. prowazekii* (NC_000963); *M. leprae* (NC_002677); *R. conorii* (NC_003103); *R. typhi* (NC_006142); *C. muridarum* (NC_002620); *C. abortus* (NC_004552); *C. caviae* (NC_003361); *C. burnetii* RSA 493 (NC_002971); *T. whipplei* (NC_004551); *T. whipplei* Twist (NC_004572); *E. ruminantium* str. Gardel (NC_006831); *E. welgevonden* (NC_006832); *E. ruminantium* str. Welgevonden (NC_005295); *C. pneumoniae* AR39 (NC_002179); *C. pneumoniae* J138 (NC_002491); *C. pneumoniae* TW_183 (NC_005043), one endosymbiotic bacterium – *B. aphidicola* (NC_004545) and one free living bacterium – *Clostridium perfringens* (NC_003366), were downloaded from the National Centre for Biotechnology Information (NCBI) (<ftp://ftp.ncbi.nlm.nih.gov/genomes/bacteria>) along with COG files (Database of Cluster of Orthologous Groups of proteins) [8] describing the gene categories. Programs were written for clustering and tabulation of protein sequences according to their length distributions and percentage change in genomes. The CDS feature annotation was used to extract the genes showing overlap. We define overlapping genes as pairs of adjacent genes whose coding regions (CDS) partly/completely overlap. The percentage of genes that is present as overlapping gene pairs in each genome and the direction of their overlap are calculated. There is a high possibility that mis-annotated open reading frames (ORFs) are included in the genome data, thus we also used the authentic (genuine) ORFs in some analyses to improve the accuracy. We defined the authentic/genuine ORFs as those that are not annotated as hypothetical and putative ORFs, or are conserved in the NCBI COGs database of orthologous genes (<http://www.ncbi.nlm.nih.gov/COG/>).

2.1 Figures

The results of our analysis are presented below.

3 Discussion

We analyzed genome size, gene distributions in various COG categories [8], protein lengths and overlapping genes from 16 obligatory intracellular prokaryotes and compared them with those of endosymbiotic bacteria and a free-living bacterium. In general prokaryotic obligate intracellular parasites of humans have reduced genomes as compared to free-living bacterium *C. perfringens*, indicating a continual selective pressure for a minimal genome. One reason proposed is the intracellular habitat preventing the uptake of foreign DNA by these organisms as suggested earlier [14]. The other reason is the differential loss of genes in these organisms with loss of gene function or genome decay with increase in adaptation to the host [4]. Hence, reductive convergent evolution can be a result of prolonged intracellular life. The results of our investigation are described below.

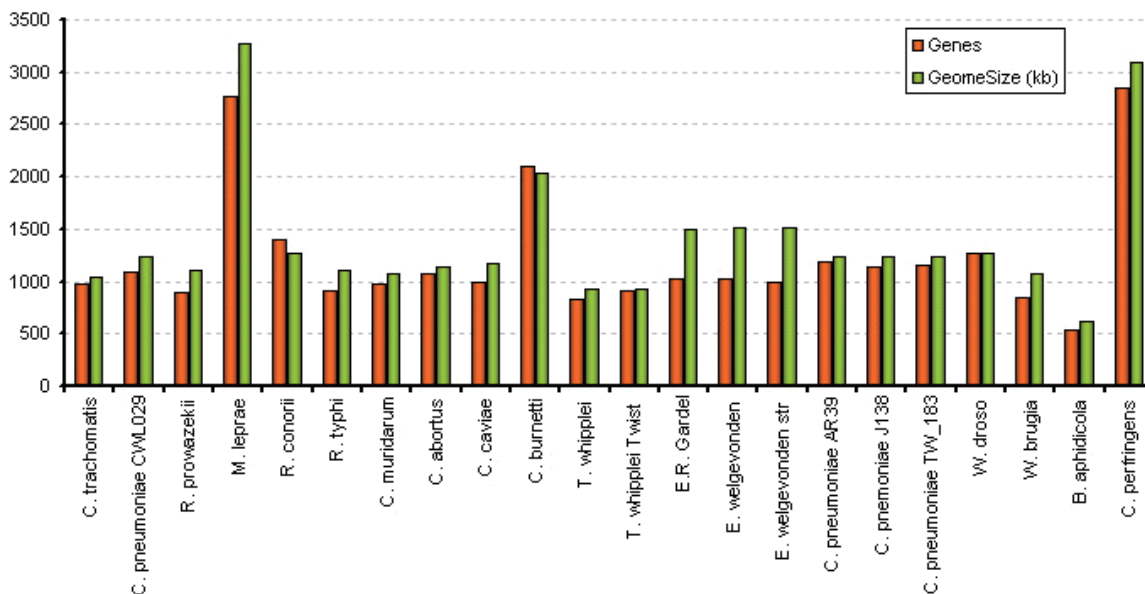


Figure 1: Distribution for genome size and gene numbers in genomes.

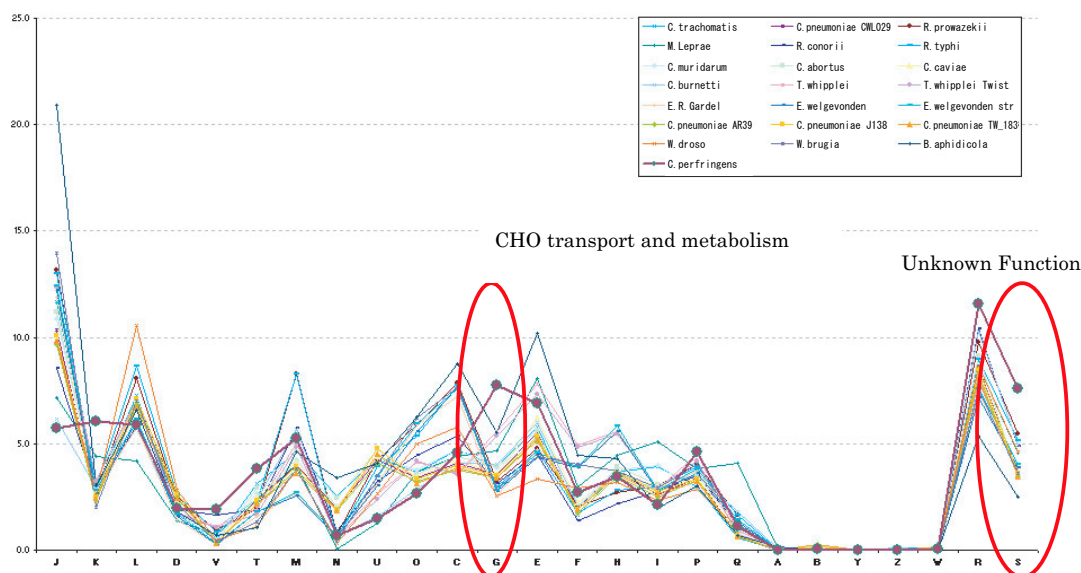


Figure 2: A plot of COG category distributions. COG categories plotted against their percentage representation in the genomes. Encircled region at G shows a clear decrease in proteins for carbohydrate transport and metabolism for obligatory lifestyle and region at S shows proportion of proteins of unknown biological function.

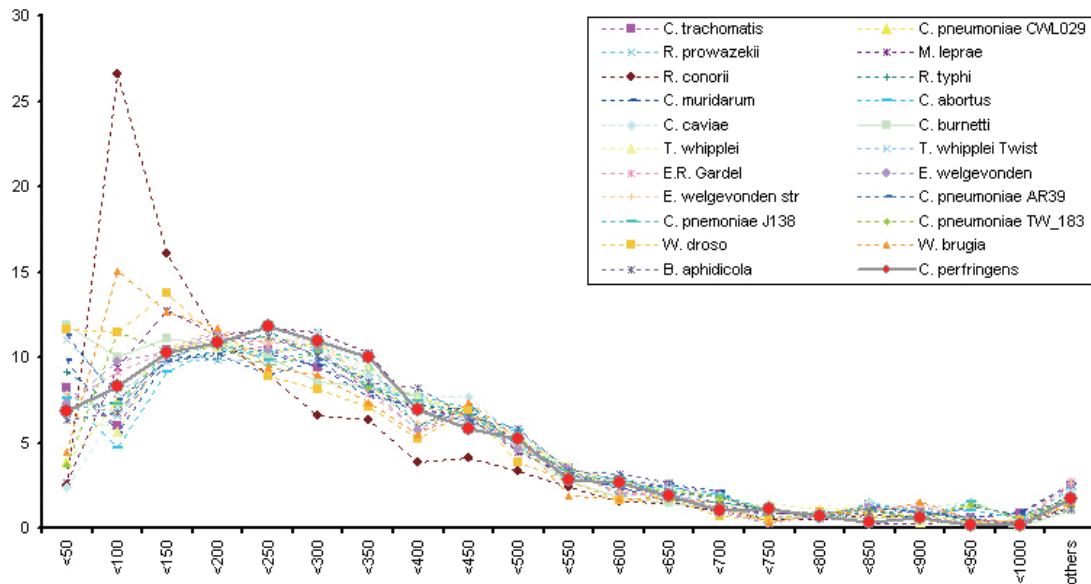


Figure 3: Percentage change in protein length distributions in the sixteen obligate intracellular parasites when compared to *C. perfringens*, a free living bacterium (red dots).

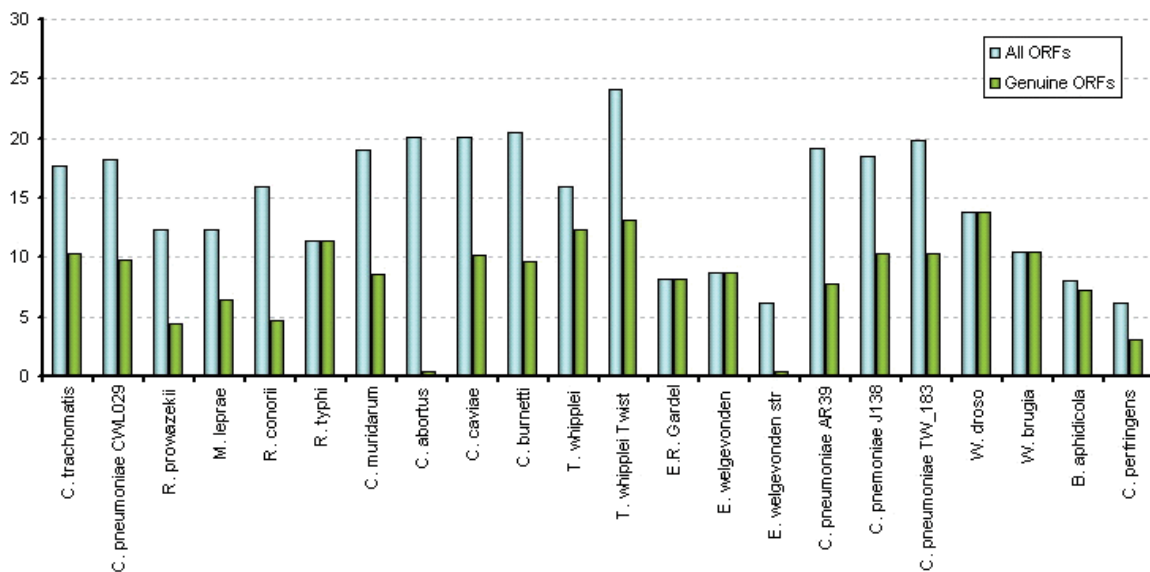


Figure 4: The percentage of genes involved in overlap. Grey: All the genes. Green: Genuine genes.

3.1 Genome Size and Number of Genes

Comparison of genome sequences from 16 completely sequenced obligatory intracellular parasites of humans with the symbiont *B. aphidicola* and a free-living bacterium *C. perfringens* reveals wide variation in size as well number of genes/proteins. Generally, it can be inferred that the number of genes decrease with decrease in genome size ($r = 1.0$). Obligatory intracellular parasites have small genomes. Since most bacterial genomes contain primarily coding DNA, genome reduction in prokaryotes must involve the loss of metabolic functions and physiological capacities with important phenotypic implications [3, 10]. It is clear that there is massive genome reduction and convergent evolution in response to life-style in all the 16 obligatory intra-cellular parasites. It is interesting to see that the smallest genome does not contain the least number of genes. This suggests for the phenomenon of differential gene loss in response to niche as most of these intracellular parasites display extreme diversity in tissue trophism and disease expression. Genome reduction pattern is also seen in the symbiont – *B. aphidicola*. It is interesting to see that *C. burnetti* and *M. leprae* have genome sizes and number of genes comparable to free-living bacteria. This can be explained based on the fact that both *C. burnetti* and *M. leprae* are still undergoing downsizing and genome reduction and are often considered as a genome ‘in decay’.

3.2 Genes Lost and Genes Retained

The 16 obligatory intracellular genomes display marked similarities in COG category distributions (Figure 2). A substantial proportion of the proteins in all the Obligatory intracellular parasites as well as *C. perfringens*, the free living bacterium chosen (at random) for comparison, are of unknown biological function. Many genes are present in one organism but absent from the other. Identification of such genes is of particular importance for the mutually exclusive biological, virulence and pathogenetic capabilities of each species. It is clear from COG category distribution that habitat is a major factor contributing to genome reduction. Supply of energy, nutrients and metabolites from the host supplements the bacterium’s potential to synthesise them. So genes involved in many metabolic pathways (enzymes) are lost partly or completely from these parasites and the loss of genes from these pathways underlie the loss in number of genes and corresponding genome size. However, this must be complemented by increase in transporter systems for their uptake from the milieu. Thus, parasitic life-style gives rise to problems that must be solved by homologous or analogous systems from the host or environmental niche. It is interesting to see that this differential gene loss occurs maximum in the range 250-600 amino acids (Figure 3). Most of the genes (> 80%) in this length range are metabolic enzymes. Comparative analyses also reveals that these genomes display marked similarities in patterns of protein length and frequency distribution, with substantial sharing of a ‘backbone genome’. These results support the fact that gene loss is niche dependent and is independent of protein length. Longer proteins, if essential are retained (e.g. gyrases) [12].

3.3 Genome Reduction and Overlapping Genes

In order to investigate the occurrence of overlapping genes in organisms with different and similar lifestyle, we determined the proportion of genomes represented by overlapping genes in 16 obligatory intracellular parasites, one endosymbiont and one free living bacteria (Figure 4). It is interesting to see that a substantial portion of the genomes in represented by overlapping genes in all the organisms. These observations clearly suggest an important role of overlapping gene pairs in genomes. All the 16 Obligatory intracellular parasites and the endosymbiont – *B. aphidicola* show a substantial amount of overlapping genes, endorsing the fact that overlapping genes are a means of compressing maximum amount of information into short sequence space and could be a result of evolutionary pressure to minimize genome size. These observations for the first time clearly suggest the role of overlapping genes and their contribution in genome reduction. These results also suggest that the niche and life

style of the organism also play an important role in determining the proportion of genes showing overlapping configuration and could be considered as pressures/constraints favoring smaller genomes. In particular these adaptive pressures explain how a given quantity of information can come to be represented by a small message.

It is interesting to see the frequent occurrence of unidirectional overlapping structure ($\rightarrow\rightarrow$ or $\leftarrow\leftarrow$) probably reflecting the most common orientation of adjacent genes in the chromosomes, as prokaryotic genes are often organized into operons (clusters of genes that are transcribed together). Because all genes in an operon must be transcribed in the same direction, this organization will be reflected in a tendency for nearby genes to have the same orientation. The two inverted orientations ($\leftarrow\rightarrow$ and $\rightarrow\leftarrow$) have lesser number of gene pairs. The lower number of the divergent structure could probably be due to the evolutionary constraints on the 5'end of the gene and the upstream region, which have structures that have essential structures like promoters. In addition, a frame shift mutation at the 5'end could destroy the entire gene. The unidirectional and convergent structures are more easily formed due to the loss of stop codons or a frameshift [6, 11]. These results are in accord with previous reports by [6, 11] and highlight the fact that gene orientation, genome reduction and evolutionary constraints work together during adaptation of the organism in its niche. These studies thus highlight that there is substantial plasticity among obligatory intracellular parasites and overlapping genes and differential gene loss convey genome reduction.

Acknowledgments

This study was supported by a grant from the Biomedical Research Council, Singapore.

References

- [1] Andersson, J. O. and Andersson, S. G., Insights into the evolutionary process of genome degradation, *Curr. Opin. Genet. Dev.*, 9:664–671, 1999.
- [2] Andersson, J. O. and Andersson, S. G., Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes, *Mol. Biol. Evol.*, 18:829–839, 2001.
- [3] Andersson, S. G. and Kurland, C. G., Reductive evolution of resident genomes, *Trends Microbiol.*, 6:263–268, 1998.
- [4] Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H., and Kurland, C. G., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, *Nature*, 396:133–140, 1998.
- [5] Chen, S. M., Takiff, H. E., Barber, A. M., Dubois, G. C., Bardwell, J. C., and Court, D. L., Expression and characterization of RNase III and Era proteins. Products of the *rnc* operon of *Escherichia coli*, *J. Biol. Chem.*, 265:2888–2895, 1990.
- [6] Fukuda, Y., Nakayama, Y. and Tomita, M., On dynamics of overlapping genes in bacterial genomes, *Gene*, 323:181–187, 2003.
- [7] Inokuchi, Y., Hirashima, A., Sekine, Y., Janosi, L., and Kaji, A., Role of ribosome recycling factor (RRF) in translational coupling, *EMBO J.*, 19:3788–3798, 2000.
- [8] Natale, D. A., Galperin, M. Y., Tatusov, R. L., and Koonin, E. V., Using the COG database to improve gene recognition in complete genomes, *Genetica*, 108:9–17, 2000.

- [9] Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F. P., and Olsson, O., Overlapping genes, *Annu. Rev. Genet.*, 17:499–525, 1983.
- [10] Ochman, H. and Moran, N., Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis, *Science*, 292:1096–1098, 2001.
- [11] Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L. and Koonin, E. V., Purifying and directional selection in overlapping prokaryotic genes, *Trends Genet.*, 18:228–232, 2002.
- [12] Sakharkar, K. R., Dhar, P. K., and Chow, V. T., Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis, *Int. J. Syst. Evol. Microbiol.*, 54:1937–1941, 2004.
- [13] Sakharkar, K. R., Sakharkar, M. K., Verma, C., and Chow, V. T., Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*, *Int. J. Syst. Evol. Microbiol.*, 55:1205–1209, 2005.
- [14] Tamas, I., Klasson, L. M., Canback, B., Naslund, A. K., Eriksson, A., Wernegreen, J. J., Sandstorm, J. P., Moran, N. A., and Andersson, S. G., 50 Million years of genomic stasis in endosymbiotic bacteria, *Science*, 296:2376–2379, 2002.
- [15] Zomorodipour, A. and Andersson, S. G., Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis*, *FEBS Lett.*, 452:11–15, 1999.