

# Comparison of Cell Envelope Gene Structure in *Chlamydia* Genomes

Girish Rao and David K.Y. Chiu

Department of Computing and Information Science  
University of Guelph, Canada

## Abstract

This paper analyzes the gene location-pattern structures across related microbial genomes in an aim to extract their organizational and functional differences. We propose a representation of sequence ordered sets to define the whole set of selected genes of a genome. To reduce the combinatoric comparison, plots depicting sequence similarity of the specified gene sets were constructed. Pair-wise comparisons were then performed based on global rearrangement operations of: (1) shifts, (2) reversals, or (3) split-reversals. Experiments were done using five closely related *Chlamydia* genomes that have substantial pathogenic differences. To analyze the relationship of pathogenic variation, the cell envelope set of genes was chosen. Results show consistent hierarchical gene rearrangements among genomes displaying similar pathogenicity. We hypothesize that an organizational structure may be intrinsic to the overall pathogenic functionality of these genomes. Results generally supported the hypothesis that genes of the cell envelope provide the variable loci required to form an effective source of variation among species that are otherwise very similar.

**Keywords:** *Chlamydiae* pathogenicity; cell envelope genes; organizational hierarchy; gene structure

## 1. Introduction

It is generally intuitive that comparing genomes as a whole can reveal important properties across genomes. Comparisons among a carefully selected set of genomes can reflect common characteristics as well as unique properties of each of the genomes under study. Different whole genome comparative schemes can be used. When a genome is compared exhaustively along its sequence using a method such as the dot-plot [1], all the sequences are considered for comparison. The approach proposed in this paper considers a specified set of coding sequences, and

compares the properties of these subsequences across several genomes.

Five *Chlamydia* genomes of scientific import that have been sequenced were selected. By comparing these varying species, an understanding of Chlamydial infection and its pathogenicity may be attained. *C. pneumoniae* is responsible for 10% of human pneumoniae cases in the United States, and 5% of bronchitis cases [2]. This paper examines three different strains of the *C. pneumoniae* species, namely, AR39, CWL029, and J138. The fourth genome, *Chlamydia trachomatis* (serovar D strain), has been implicated in human ocular infection and sexually transmitted disease (STDs) [2]. The fifth genome, *Chlamydia muridarum* (strain Nigg), has been implicated as the mouse pneumonitis strain of *C. trachomatis* [3]. The first three genomes belong to one species, having closely related disease expression, while the last two belong to a different species of *C. trachomatis* [4]. All share a common biology as obligate eubacterial parasites [4].

Though these genomes share common biological functionality, extreme diversity exists in their pathogenic activity. The genetic basis for the variation of disease expression remains a major unanswered question in Chlamydial biology [4]. As in recent studies, we have focused on the cell envelope genes [4,5,6]. Thus, the proposed cell envelope hypothesis states that the cell envelope genes are the source of these pathogenic variations.

In general, the comparison of the selected gene sets between genomes is analogous to a graph morphism problem that can be exponentially explosive. In order to study the similarity of cell envelope gene sequences among the five genomes, a comparison scheme based on sequence ordered set is proposed. An ordered set is distinguished both by the identity of each element and by the order of the elements. A sequence ordered set is such that the elements are themselves sequences. We also propose a sequence-ordered-set maximum-similarity plot, as a special type of sequence-score plot that provides a flexible, visual, comparative representation. These plots allow for visual analysis, and aid in revealing

organizational gene patterns unique to a given genome, or shared among many genomes. The proposed analyses based on sequence-ordered-set comparisons then provide a basic description such that more complex comparisons can be established.

## 2. Methods

### 2.1. Sequence-ordered-sets and sequence-score plots

A sequence-ordered-set can be defined as an ordered set  $S = \langle s_1, s_2, \dots, s_n \rangle$  where the elements  $s_1, s_2, \dots, s_n$  are themselves sequences. The integer  $n$  specifies the size of the set and can vary for different sets. The elements of  $S$  are ordered by a relation between consecutive members, that is,  $s_i$  “precedes”  $s_{i+1}$ . Note that two ordered sets  $S1 = \langle s_{11}, s_{12}, \dots, s_{1n} \rangle$  and  $S2 = \langle s_{21}, s_{22}, \dots, s_{2m} \rangle$  are equal if and only if:

(i)  $n = m$ , (ii) Corresponding elements  $s_{1k} = s_{2k}$  for all  $k=1, 2, \dots, n$ . Note that each element needs not be consecutive; in fact, most of the cell envelope genes are non-adjacently distributed throughout the genome. A sequence-ordered-set representation for gene set comparisons is important here because the specified set can then be considered and analyzed as a whole.

To evaluate a comparison between two sequence ordered sets, we propose the following Sequence-Ordered-Set Maximum Similarity Score ( $\Psi$ ) that computes: 1) the maximum similarity score of each element in a set to all the elements in the other set, 2) the summation of all the maximum similarity scores normalized. Let the sequence ordered sets be denoted as  $S1 = \langle s_{11}, s_{12}, \dots, s_{1n} \rangle$  and  $S2 = \langle s_{21}, s_{22}, \dots, s_{2m} \rangle$ , then  $\Psi$  can be defined formally as:

$$\Psi(S1, S2) =$$

$$(\sum_{k=1, n} d(s_{1k}, S2) + \sum_{k'=1, m} d(s_{2k'}, S1)) / (n + m)$$

A distance function  $d$  can be applied to each element in one set to the other set. The distance function  $d$  can be defined to map an element from one set and the other set to a real number between zero and one. For our purpose in this paper, we choose the minimum distance function of edit distance, using operations of substitutions, insertions, and deletions based on the BLAST matching program, to find the element of maximum similarity. The metric property of  $\Psi(S1, S2)$  then depends on that of  $d$ . A minimum distance search (MDS) algorithm was implemented. Given  $S1 = \langle s_{11}, s_{12}, \dots, s_{1n} \rangle$  and  $S2 = \langle s_{21}, s_{22}, \dots, s_{2m} \rangle$ , the minimum distance  $d(s_{1k}, S2)$  is defined as:

$$d(s_{1k}, S2) = \min_{k'=1, m} \delta(s_{1k}, s_{2k'}) ;$$

$$s_{1k} \in S1; s_{2k'} \in S2,$$

where  $\delta(s_{1k}, s_{2k'})$  is a predefined distance function between an element of  $S1$  and an element of  $S2$ .

To facilitate further comparisons between two sequence ordered sets, a graphic plot based on the distance values of  $d(s_{1k}, S2)$  and  $d(s_{2k'}, S1)$  with respect to the sequence order (or location indices) is proposed. We call this plot a *sequence-ordered-set maximum-similarity plot* (SOS-MS plot), or just *sequence-score plot* (S-S plot) in this paper. That is, instead of summing all the distances, the distance values are plotted along the location indices of the sequence elements  $s_{1k}$  and  $s_{2k'}$ ,  $k=1, 2, \dots, n$  and  $k'=1, 2, \dots, m$ . In this way, the ordering position or the actual location index of the sequence elements of the two ordered sets can then be incorporated, similar to the problem of finding graph morphism.

Using the sequence-score plots, the presence of hypothesized gene rearrangement operations can also be evaluated. This process of globally rearranging the ordered set is described next.

### 2.2. Global operations

By analyzing the comparative sequence-score plots, three global operations are proposed based on: (1) shifts of the locations of a subset of consecutive elements, (2) splits of a whole set of consecutive elements into segments, and (3) reversals of the locations of consecutive elements in a segment (similar to inversions). These operations act in a uniform manner upon specific subsets of elements to alter the overall relationships between genomes.

The goal of performing an operation is two-fold. First, by rearranging a segment of a genome, further similarities may appear. Finding such relationships may provide insight into the evolutionary and organizational structure shared between the elements of the ordered sets. Second, if strong relationships exist between sets of similar functionality or subclass, the plausibility of the operations and their significance can be inferred.

The three global operations proposed here represent global events have also been shown to occur biologically [7,8,9]. While the sequence-score plots provide valuable visual information, the quantitative position patterns are difficult to derive from the visualization alone. Thus, currently a computational program LocusDistance is implemented to compare two ordered sets at a time; an operation is performed on a single ordered set ( $S1$ ), and the new locations of these genes are compared to the locations of the unchanged ordered set ( $S2$ ).

## 3. Cell envelope gene comparisons

Gene sequence data of the five complete genomes, including the cell envelope genes, were downloaded from the Comprehensive Microbial Resource (CMR) [10], an extensive database housed at The Institute for Genomic Research (TIGR). All five genomes were made available in the Version 1.0 January-December 2000 build of the database.

From the five genomes, we obtain a total of ten non-redundant pair-wise comparisons: six comparisons that are *inter*-species, three that are *intra*-species, and one that is between *Chlamydia trachomatis* (serovar D strain) and *Chlamydia muridarum* (strain Nigg). Sequence-score plots were generated for these pair-wise comparisons. The plots were then visually analyzed, and suitable global operations were selected and applied. New sequence-score plots were generated, effectively displaying the relationships between genomes after applying a hypothesized operation.

In Figure 1, the original comparison between *C. pneumoniae* CWL029 (denoted CP029) and *C. pneumoniae* AR39 is displayed. An initial observation of the plot reveals extremely high BLAST Bit scores (and low e-values, data not shown), which signifies high pattern similarity from gene element to gene element; however, it is evident that gene order is not as highly conserved.

However, upon applying a Split-Reversal operation, significant similar relational organizations were discovered. The CP\_AR39 cell envelope gene set was split at its 5' gene position 932429 into proximal and distal segments. Each segment was then independently reversed. Figure 2 displays the result of reversing the proximal segment.

It is evident that the rearrangement due to applying the global operation produces a comparison that shows the overall increase of similar location order between the two proximal segments of the genomes. Likewise, the reversal of the distal segment of *C. pneumoniae* AR39 also resulted in an increase in such similarity (Fig. 3).

These two sequence-score plots reveal two organizational segments, based on the selected gene clusters that "structurally" rearrange together, according to a specifically identified global operation.

Calculating the distance between each gene element further strengthened the relationship between CP029 and CP\_AR39. That is, for  $S1 = \langle s_{11}, s_{12}, \dots, s_{1n} \rangle$  and  $S2 = \langle s_{21}, s_{22}, \dots, s_{2m} \rangle$ , the difference in base pair position was calculated between  $s_{11}$  and  $s_{21}$ ,  $s_{12}$  and  $s_{22}$ , for each  $s_{1n}$  and  $s_{2m}$ . The tabular data generated by the program for this comparison is provided in abbreviated version as a graphical figure (Figure 4).

Figure 4 plots the physical distance between pairs of genes (as determined by the MDS algorithm) prior to the split-reversal operation, and after the operation. It is evident that the operation causes the distances to become highly similar. The post-operational indices also reflect the nature of the proximal-distal split:

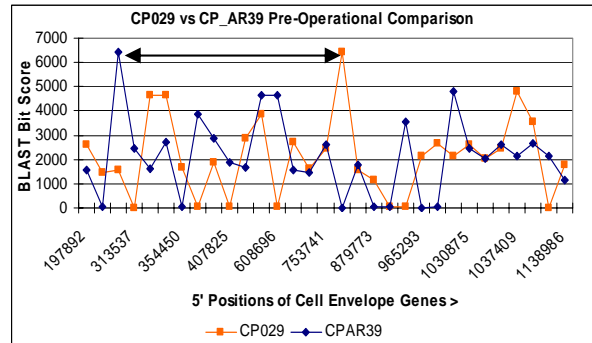


Fig.1. Sequence-score plot displaying intra-species comparison between *C. pneumoniae* CWL029 and *C. pneumoniae* AR39. Note the very high BLAST scores, indicating very high similarity. However, the overall relational organization of the gene elements between the two genomes is not as high. The black, double-sided arrow demarcates the location difference between the 6400-scoring gene on CP029 and the 6400-scoring gene on CP\_AR39.

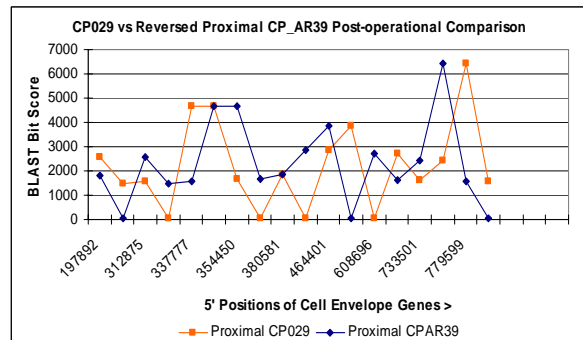


Fig.2. Plot displaying intra-species, post-operational comparison between *C. pneumoniae* CWL029 and proximally reversed *C. pneumoniae* AR39. The degree of order similarity compared to Figure 1 is greatly increased after the split-reversal operation.

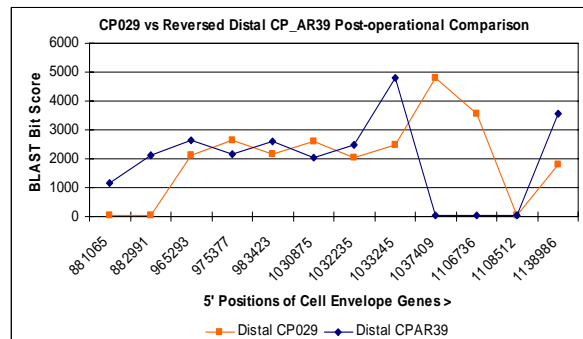


Fig.3. Plot displaying intra-species, post-operational comparison between *C. pneumoniae* CWL029 and distally reversed *C. pneumoniae* AR39. The degree of similarity compared to Figure 1 is greatly increased after applying the split-reversal operation.

proximally reversed genes are offset by about 124751 and 125053 bases; the distally reversed genes are offset by about 84906 bases.

The distinct, cross-genome patterns displayed in Figures 2, 3 and 4 were typical of comparisons involving genomes of similar pathogenicity, while comparisons involving species such as *C. pneumoniae* and *C. trachomatis* (differing pathogenic phenotypes) failed to display any such similarity.

## 4. Analysis and Discussion

The intra-species comparisons between *C. pneumoniae* AR39 and each of *C. pneumoniae* CWL029 and *C. pneumoniae* J138 provided very consistent results.

The increase in sequence-location similarity after applying the same split-reversal operation confirmed the close evolutionary relationship between each of these genomes. That such consistency can be achieved by transforming the unsystematic location indices using a single formula is indicative of the close structural-functional relationship these gene sets share.

In addition, the split-reversal operation reveals that the *C. pneumoniae* cell envelope gene set is composed of two gene clusters, structures that rearrange as a unit. The fact that these non-adjacent genes rearrange together provides insight into the nature of possible functional structure. It is possible that each of these two clusters is composed of genes that perform a particular role within the cell envelope. Studies have shown that functionally related genes tend to organize themselves co-linearly in order to optimize transcription and translation [11]. Thus, by maintaining the grouping of specific structures, biological processes may take place more efficiently.

Another result that must be examined is the overlapping regions displayed by the sequence-score plots (data not shown) generated by the intra-species comparisons. As stated, such patterns have implications for gene organization. The extent of overlapping is typical only of intra-species comparisons. Thus, not only do these gene clusters rearrange similarly on both genomes, the lengths between each gene within the cluster are also consistent on both genomes.

The inter-species, different-pathogenicity comparisons (such as the CP\_AR39-CT experiment) provided more insight. The lack of a distinct rearrangement relationship between the two cell envelope gene sets indicates that these genes, organizationally speaking, are not as closely related as other pathogenically related genomes. Thus, the variation in cell envelope gene organization is emulated as variation in pathogenicity.

The final comparison type involved *C. trachomatis* and *C. muridarum*. The fact that the genomes are related by a simple shift operation, while the intra-species genomes are related by a more costly split-reversal relationship, supports the strong link between the species [12], and also highlights intrinsic properties of these genomes. First, the existence of

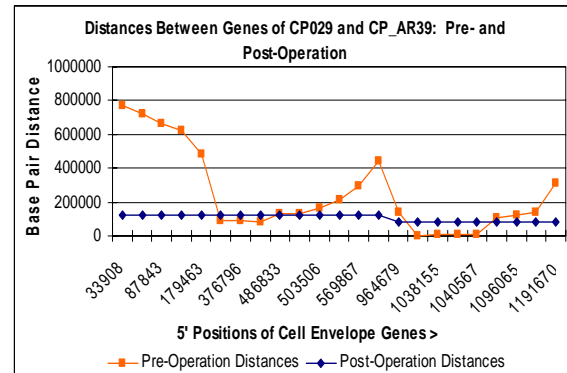


Fig.4. Intra-species comparison between *C. pneumoniae* CWL029 and split-reversed *C. pneumoniae* AR39 is displayed. The comparison displays the difference between base pair position of genes in CP029 and CP\_AR39, pre-operationally and post-operationally. It is evident that the operation causes the comparison to become more similar.

gene clustering on the *C. pneumoniae* genomes is easier to substantiate than on the *C. trachomatis* and *C. muridarum* genomes: certain genes on the *C. pneumoniae* genomes tend to group and cluster together according to rearrangement patterns. Such observations suggest that gene rearrangements occur more frequently in *C. pneumoniae* genomes. Earlier reports have indicated that simple sequence repeats (SSRs) increase the occurrence of gene rearrangements; moreover, it has been found that such SSR elements occur more frequently in *C. pneumoniae* genomes than in other *Chlamydia* genomes [5,13]. Thus, our results substantiate evidence of specific organizational gene structures.

Given these results, it becomes evident that the distribution of the cell envelope genes of the *Chlamydia* genomes and their organization can be evaluated using our proposed sequence-score plots and global operations. By applying such operations, organizational clusters become more evident. Though the degree of clustering can be controlled by other sequences within the genome (SSRs), gene clusters may associate based on an overall functionality (cell envelope proteins) as well. The quantification of such functional genetic networks has become the aim of several research studies [14]. Further study will broaden our understanding of such intrinsic phenomena. These interpretations can be gained from analyses initiated by the proposed methods, which clearly show detailed comparisons incorporating both

sequence similarity and distribution relationships across whole gene sets.

## 5. Conclusions

The comparative study provides evidence regarding the source of pathogenic variations, as well as comparative organizational structure among the *Chlamydia* genomes. By analyzing the non-adjacent cell envelope genes, questions regarding gene rearrangements and structural comparisons were further evaluated. In addition, the rearrangement experiments using our proposed global operations offered insight into genome dynamics and a possible organizational structure that may affect, or be intrinsically linked to, the overall pathogenicity of the species. Such organizational and sequential consistencies among the cell envelope regions of pathogenically related genomes provide further evidence that the genes of the cell envelope may generally be involved in imparting unique pathogenicity to different species of the *Chlamydia* genus.

## Acknowledgements

The research is supported by the National Sciences and Engineering Research Council of Canada, Discovery Grant and the Korea Research Foundation Grant (KRF-2004-042-C00020).

## References

- [1] Y. Huang, and L. Zhang, "Rapid and sensitive dot-matrix methods for genome analysis," *Bioinformatics*, pp.460-466, 2004.
- [2] S. Kalman, W. Mitchell, R. Marathe, C. Lammel, J. Fan, R. Hyman, L. Olinger, J. Grimwood, R. Davis, and R. Stephe, "Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*," *Nature Genetics*, pp. 385-9, 1999.
- [3] J. Perfettini, T. Darville, A. Dautry-Varsat, R. Rank, and D. Ojcius, "Inhibition of apoptosis by gamma interferon in cells and mice infected with *Chlamydia muridarum* (the mouse pneumonitis strain of *Chlamydia trachomatis*)," *Infect Immun*, pp.2559-65, 2002.
- [4] T. Read, R. Brunham, C. Shen, S. Gill, J. Heidelberg, O. White, E. Hickey, J. Peterson, T. Utterback, K. Berry, S. Bass, K. Linher, J. Weidman, H. Khouri, B. Craven, C. Bowman, R. Dodson, M. Gwinn, W. Nelson, R. DeBoy, J. Kolonay, G. McClarty, S. Salzberg, J. Eisen, and C. Fraser, "Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39," *Nucleic Acids Research*, pp.1397-1406, 2000.
- [5] E. Rocha, O. Pradillon, H. Bui, C. Sayada, and E. Denamur, "A new family of highly variable proteins in the *Chlamydia pneumoniae* genome," *Nucleic Acids Res*, pp. 4351-60, 2002.
- [6] B. Vandahl, A. Pedersen, K. Gevaert, A. Holm, J. Vandekerckhove, G. Christiansen, and S. Birkelund, "The expression, processing and localization of polymorphic membrane proteins in *Chlamydia pneumoniae* strain CWL029," *BMC Microbiol*, p. 36, 2002.
- [7] D. Snustad, and M. Simmons, *Principles of Genetics* 2<sup>nd</sup> Edition John Wiley & Sons, Inc. pp.477-493, 2000.
- [8] C. Chen, K. Wu, Y. Chang, C. Chang, H. Tsai, T. Liao, Y. Liu, H. Chen, A. Shen, J. Li, T. Su, C. Shao, C. Lee, L. Hor, and S. Tsai, "Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen," *Genome Res*, pp. 2577-87, 2003.
- [9] M. McLeod, X. Qin, S. Karpathy, J. Gioia, S. Highlander, G. Fox, T. McNeill, H. Jiang, D. Muzny, L. Jacob, A. Hawes, E. Sodergren, R. Gill, J. Hume, M. Morgan, G. Fan, A. Amin, R. Gibbs, C. Hong, X. Yu, D. Walker, and G. Weinstock, "Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae," *J Bacteriol*, pp. 5842-55, 2004.
- [10] J. Peterson, L. Umayam, T. Dickinson, E. Hickey, and O. White, "The Comprehensive Microbial Resource," *Nucleic Acids Research*, pp. 123-125, 2001.
- [11] R. Svetic, C. MacCluer, C. Buckley, K. Smythe, and J. Jackson, "A metabolic force for gene clustering," *Bull Math Biol*, pp. 559-81, 2004.
- [12] C. Ortutay, Z. Gaspari, G. Toth, E. Jager, G. Vida, L. Orosz, and T. Vellai, "Speciation in *Chlamydia*: genome wide phylogenetic analyses identified a reliable set of acquired genes," *J Mol Evol*, pp.672-80, 2003.
- [13] E. Rocha, "DNA repeats lead to the accelerated loss of gene order in bacteria," *Trends in Genetics*, pp. 600-603, 2003.
- [14] M. Xiong, J. Zhao, and H. Xiong, "Network-based regulatory pathways analysis," *Bioinformatics*, pp. 2056-66, 2004.