Genetic Relationships among 527 Gram-Negative Bacterial Plasmids

Yunyun Zhou¹, Douglas R. Call¹,², and Shira L. Broschat¹,²*

¹School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA
²Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA

Running title: Genetic relationships among bacterial plasmids

Correspondence to:
Shira L. Broschat
School of Electrical Engineering and Computer Science
Washington State University
PO Box 642752
Pullman, WA 99164-2752
Phone: 509-335-5693
Fax: 509-335-3818
Email: shira@eecs.wsu.edu
Abstract

Plasmids are mosaic in composition with a maintenance “backbone” as well as “accessory” genes obtained via horizontal gene transfer. This horizontal gene transfer complicates the study of their genetic relationships. We describe a method for relating a large number of Gram-negative (GN) bacterial plasmids based on their genetic sequences. Complete coding gene sequences of 527 GN bacterial plasmids were obtained from NCBI. Initial classification of their genetic relationships was accomplished using a computational approach analogous to hybridization of “mixed-genome microarrays.” Because of this similarity, the phrase “virtual hybridization” is used to describe this approach. Protein sequences generated from the gene sequences were randomly chosen to serve as “probes” for the virtual arrays, and virtual hybridization for each GN plasmid was achieved using BLASTp. Each resulting intensity matrix was used to generate a distance matrix from which an initial tree was constructed. Relationships were refined for several clusters by identifying conserved proteins within a cluster. Multiple-sequence alignment was applied to the concatenated conserved proteins, and maximum likelihood was used to generate relationships from the results of the alignment. While it is not possible to prove that the genetic relationships among the 527 GN bacterial plasmids obtained in this study are correct, replication of identical results produced in a separate study for a small group of IncA/C plasmids provides evidence that the approach used can correctly predict genetic relationships. In addition, results obtained for clusters of Borrelia plasmids are consistent with the expected exclusivity for plasmids from this genus.

Keywords:

Gram-negative plasmid, horizontal gene transfer, broad host range, IncA/C plasmid, Borrelia
plasmids, antibiotic resistance genes, classification, cluster refinement

Abbreviations:
BHR, broad host range; HGT, horizontal gene transfer; GN, Gram-negative; MGM, mixed-genome microarray; CDS, coding gene sequence; ADD, average absolute difference

1. Introduction
Plasmids are extrachromosomal DNA molecules that are found in many species of bacteria and within taxa from archaea, eukaryota, and bacteria (Bapteste et al., 2004). Sequenced plasmids vary in size from less than 1 kbp to more than 2500 kbp, and plasmids vary in their compatibility with different hosts and with other plasmids within the same host cell (Couturier et al., 1988). Plasmids are considered “mosaic” in composition containing both backbone genes for maintenance and mobile and transmissible genes that encode “accessory” traits (Christopher, 2000). Plasmid genes can be obtained from multiple sources (Boyd et al., 1996) and disseminated by horizontal gene transfer (HGT). HGT is responsible for the dissemination of many of the undesirable traits associated with bacteria, including antibiotic resistance and virulence. In addition, broad-host-range plasmids play an important role in bacterial adaptation to new environments. This provides much of the motivation for understanding the relationships among plasmids. Knowledge of these relationships will help us to better understand how genes are shared horizontally across species boundaries as well as to understand microbial evolution.

There are several ways to identify genes that have arisen from divergent sources, including comparison of GC frequency, codon usage, and genomic signatures (Campbell et al., 1999; Karlin 2001; Karlin et al., 1995; Mark et al., 2000; Suzuki et al., 2008). However, there is some debate...
over whether plasmid mosaicism can be understood from such features (Campbell et al., 1999; Mark et al., 2000). In addition, while molecular methods are frequently used to characterize plasmids (Smalla et al., 2000), there is no sequence analogous to the 16S rRNA sequence in bacteria with which to examine their phylogenetic relationships. Several network-based representations have been used to explore genetic relationships among plasmids (Halary and Leigh, 2009; Popa et al., 2011; Brilli et al., 2008). In particular, Brilli et al. (2008) studied the evolutionary relationships of several Gram-negative bacterial plasmids, including those hosted by Escherichia, Salmonella, and Shigella, using the Blast2Network method. Our work is the first to study the genetic relationships of a broad and diverse group of Gram-negative bacterial plasmids.

In this paper we introduce a method for investigating the genetic relationships of 527 Gram-negative (GN) bacterial plasmids using their complete gene sequences. Prior to the availability of these sequences, methods such as the one we describe in this paper were not possible, and it was necessary to rely on other approaches--e.g., supertree algorithms [a supertree is a single phylogenetic tree assembled from a combination of smaller phylogenetic trees based on different datasets (Gordon, 1986)]--to estimate the genetic relationships of a large number of plasmids. The significant advantage of our approach is that it exploits all the genetic information available in a systematic and comprehensive manner. We start with a modified virtual mixed-genome microarray (MGM) method to create an initial tree that describes overall genetic similarity for these plasmids (Wan et al., 2007) using proteins rather than DNA for both “probes” and “targets.” Because virtual hybridization of MGMs is an entirely computational method, protein sequences can be used as readily as DNA sequences. We choose to use protein “probes” and “targets” because doing so is more efficient computationally (amino acid sequences are one-third as long as their
nucleotide counterparts) and because differences in silent nucleotide mutations are absent in amino acid sequences. To overcome representational bias due to gene repetition, we use BLASTp on the concatenated amino acid sequences of a plasmid with itself and remove duplicate proteins for each plasmid. After removal of the duplicate proteins, protein sequences are randomly chosen to serve as “probes” for the virtual arrays, and virtual hybridization for each GN plasmid is achieved using BLASTp. Each resulting intensity matrix is used to generate a distance matrix from which the initial tree is constructed. After completion of the initial tree, conserved proteins within a cluster are identified and used to refine the relationships within the cluster by means of multiple sequence alignment of the conserved proteins.

2. Materials and Methods

2.1 Data Preparation.

In July 2010 the complete gene sequences for 2,171 bacterial plasmids were available in the NCBI genome database (http://www.ncbi.nlm.nih.gov/). Of these, 527 sequences were for Gram-negative (GN) bacterial plasmids with more than 50 putative coding genes (CDS) (supplemental file 1). These were downloaded in FASTA format and translated into amino acid sequences based on putative open reading frames. BLASTp with default parameters was used to remove duplicate proteins within plasmid sequences by blasting the sequence with itself. Duplicate proteins were not removed across plasmids because of the need to reflect a representative distribution within the entire protein population. A protein was considered to be a duplicate for the similarity value as (length of matching sequence) \*(BLAST similarity score)/(length of reference protein) \(\geq 0.45\) (Call et al., 2010). The resulting set of proteins for all 527 GN plasmids after removal of duplications–more than 97,000 in total–was used to obtain “probes” that were randomly se-
lected to create the virtual arrays. Each array consisted of 20,000 proteins, roughly 20% of the total protein population. The probe selection procedure utilized independent sampling without replacement.

2.2 Selection of number of arrays.

Using 20% of the protein pool to construct an array corresponds, on average, to 20% representation of each plasmid on an array; this degree of representation is sufficient for discrimination (Wan et al., 2007). Nevertheless, because probe selection is random, there is no guarantee that plasmids will have equal representation, and therefore sampling bias might be a concern. To overcome potential bias, we can construct a number of virtual arrays and generate the initial tree using a consensus method based on all the array results. The problem then is to determine the number of arrays needed for the analysis. In terms of accuracy of the relationship results, we assume the more arrays that are used, the better. However, the computational expense involved in using BLASTp or “virtual hybridization” for each array makes it necessary to determine an optimum number of arrays—i.e., a number that minimizes the computational cost while minimizing variance. In theory sampling bias can be removed simply by using the entire set of proteins to construct one single array—i.e., an array with more than 97,000 probes rather than multiple arrays of 20,000 probes. However, the computational expense both in terms of CPU time (estimated at more than 16 days of continuous CPU time using one powerful processor, which was unavailable for this work) and memory is impractical.

To determine the optimum number of arrays, we used the average absolute difference (ADD) (Fig. 1) as a function of the number of arrays. After virtual hybridization of an array for N different plasmids, an N x N distance matrix is obtained. Pair-wise comparison of two distance matric-
es results in an $M = N(N-1)/2$ distance vector. For the ADD metric, we sum the absolute difference of the mean distances for $n+1$ arrays and $n$ arrays over all $M$ values. The formula is as follows:

$$ADD = \sum |<D(n+1,i)> - <D(n,i)>|$$

where

- $i$ takes on values between 1 and $M$,
- $n$ is the number of virtual arrays, and
- $<D(n,i)>$ is the mean distance value for the $n$ virtual arrays.

The ADD is a robust estimator of the absolute difference between two populations of distance matrices. When two populations are exact, the ADD will be zero. As we increase the number of arrays used, we expect the differences between the mean values to decrease (i.e., the mean values will not change as much, for example, between 50 and 51 arrays compared to between 2 and 3 arrays), and thus we expect the value of the ADD to decrease as the number of arrays used increases. For $n=1$ the ADD is approximately 60,000 (Fig. 1). It then declines quickly, and for $n=20$, the value is $< 2,000$. For 50 and 99 arrays, the values are 679 and 379, respectively. Note that to calculate the value of the ADD for 99 arrays, 100 arrays are used. If we assume the distance between each location in the two matrices is the maximum distance we obtained minus the minimum possible distance of zero, the maximum value for the ADD is 301,707. The ADDs for 50 and 99 arrays are then approximately 0.002 and 0.001 of this maximum value, respectively. While the small fractional difference of 0.002 indicates that the use of 50 arrays is probably sufficient, we conservatively chose to use 100 arrays.

3. Results
3.1 Initial tree construction.

As indicated previously, each array was populated by 20,000 randomly chosen protein “probes” (entire proteins were used) and a total of 100 arrays were “constructed.” For each array virtual hybridization was performed for the 527 GN plasmids. Hybridization was simulated using stand-alone BLASTp with default parameters and normalized intensities between 0 and 1 were obtained using the formula \( 2 \times \text{(length of matching sequence)} \times \text{(BLASTp similarity score)} / (\text{length of reference protein} + \text{length of matching sequence protein}) \) resulting in 100 \((527 \times 20,000)\) normalized intensity matrices. From each intensity matrix, a distance matrix was calculated using pairwise Euclidean distances. Two methods were used to obtain consensus trees from the 100 distance matrices: majority voting (Margush and Mcmorris, 1981) and averaging. Because majority voting is order dependent for an unrooted tree, human and cattle proteins were used as an outgroup to force clustering of the GN plasmids (these proteins were added to the arrays and during virtual hybridization all intensity values for these proteins were zero for the plasmids). Consense from Phylip was then used to obtain the initial tree (Felsenstein, 1989). For the second method the 100 distance matrices were averaged to obtain a single distance matrix and neighbor joining was used to construct the initial tree using Matlab.

The sheer size of the consensus trees for the 527 GN bacterial plasmids prevents inclusion of them as figures. Instead, files for these trees in Newick standard format are included as supplemental files 2 and 3 that may be viewed in circular format using freely available, tree-generating software packages such as MEGA5 (Tamura et al, 2011). To compare results for the two types of consensus trees, we considered the intensity results for a subset of 50 of the 527 GN plasmids. Each virtual array was populated by randomly selecting 1,400 probes from a protein pool of
As with the 527 plasmids, 100 arrays were used to obtain 100 distance matrices. Figures 2 and 3 show the majority voting and averaged distance results, respectively. There are some limitations to comparing the two trees because; for example, one is unrooted (majority voting) and the other is rooted. Nevertheless, recall that both trees are generated using all proteins and reflect only the initial relationships. Hence, distances are unimportant, and we focus on comparison of the clusters. While the trees have different topologies, the clusters are virtually identical.

### 3.2 Cluster refinement.

The initial tree is constructed using the entire pool of proteins including those obtained via HGT. Relationships between clusters can be further refined using conserved genes (Call et al., 2010). Conserved genes are the genes shared among clusters and are considered to be orthologous genes and, thus, are less likely to represent cases of HGT—i.e., genetic content from a recent common ancestor can be used to refine relationships within clusters, while genes from a common ancestor should contain more information about the evolutionary relationship among plasmids. It would take considerable computational time to refine all the clusters comprising the 527-plasmid tree, and there would be no meaningful way to assemble the resulting clusters because distances would be inconsistent across clusters. In addition, there is no benefit to refining a cluster unless it is of interest to a particular investigator. To illustrate the results of refining a cluster, we consider the two clusters shown in Figs. 4(a) and 5(a), one from the *Borrelia* group and the other a group of plasmids from several species. Both clusters form part of the full 527-plasmid tree obtained using majority voting. Conserved proteins were identified via BLASTp searches using the same approach described in (Call et al., 2010) except using proteins rather than genes; the similarity cut-off value was 0.3. In addition, when either the length of an aligned sequence was short relative to the reference sequence or the BLASTp identity score was low, the protein was excluded.
from the analysis. The number of conserved proteins identified at each cluster node is shown.

Figures 4(b) and 5(b) shows the results after multiple-sequence alignment has been performed using MEGA5 with default settings. Bootstrap values (500 iterations) are given as a measure of confidence for each node.

4. Discussion

The method described in this paper provides a new in silico approach to study genetic relationships among plasmids. One useful outcome of this type of analysis is that we can consider questions about the dissemination of antibiotic resistance genes among bacteria as discussed below.

While there is no “gold standard” to assess the “accuracy” of the proposed relationships (Figs. 2-5), there are several independent observations that support the validity of our results. One consistent finding was the clear demarcation of Borrelia plasmids as separate from plasmids from Enterobacteriaceae (Baker et al., 2007) both in the 50-plasmid analysis (Fig. 2) and in the full plasmid analysis (supplemental files 2 and 3). Borrelia is a distinct genus of organisms from the spirochete phylum and is mostly known for pathogenic species that are responsible for vector-borne zoonotic infections (e.g., Lyme disease). Plasmids available for Borrelia in Genbank include both circular and linear plasmids. The latter are rare for members of Enterobacteriaceae so we would not expect them to be genetically related. Furthermore, the only plasmid sequences from relapsing fever cluster closely together (supplemental files 4 and 5) as should be expected given that B. recurrentis (NC_011246) is considered to be very closely related to B. duttonii (NC_011247) (Lescot, 2008). The Borrelia plasmids that are classified as “cp32” are also grouped together (additional files 4 and 5; NC_011720, NC_011842, NC_012106, NC_012253) as expected given that these plasmids are considered to arise from prophage (Eggers and Samuels, 1999). Borrelia plasmids that are required for infectivity, including lp25 (NC_012166,
NC_011856), lp36 (NC_011857, NC_011867, NC_001855, NC_012184, NC_012202), and lp38 (NC_012167, NC_012182) are also clustered together (Purser and Norris, 2000) (supplemental files 4 and 5). Another notable and consistent grouping in our analysis includes incompatibility A/C plasmids (IncA/C; Fig. 5) that have been well described in the literature (Call et al., 2010; Fricke et al., 2009; Welch et al., 2007) and should be expected to cluster closely together.

4.1 Distribution of Four Antibiotic-Resistance Genes

Importantly, the 527-plasmid tree can be used to study the distribution of genes of particular interest. To illustrate this we considered four common antibiotic resistance genes, tetA, tetB, cat, and floR. We first performed a sensitivity analysis to identify orthologous resistance genes using the coefficient of variation to identify a similarity threshold (45% similarity in this case; data not shown). Next a BLASTp search of all 527 plasmid sequences resulted in identification of 40 plasmids with one or two of the four AR genes (Fig. 6). In addition to the relatively few plasmids with these antibiotic resistance genes, it is striking that 37 of the 40 plasmids group within a single 115-plasmid cluster (Fig. 6). We examined the 115-plasmid cluster further (supplemental file 6) and found several features of interest. Plasmids from Escherichia coli compose 51.3% of the 115 plasmids. This contrasts with their representation in the remaining 412 plasmids, which is only 1.5%. Of the 37 plasmids with antibiotic resistance genes within the cluster, 15 (41.7%) are from E. coli. While most of the 115 plasmids are from bacteria belonging to the Enterobacteriaceae family, a number of them are not, although all belong to the class Gammoproteobacteria. Importantly, not all plasmids associated with the Enterobacteriaceae family or the Gammoproteobacteria class are in the cluster. Removal of the antibiotic resistance genes had an insignificant impact on the clustering of the 115 plasmids confirming that this grouping was not a consequence of the antibiotic resistance genes themselves. Finally, the size distribution (number of...
proteins) of the 115 plasmids is significantly smaller than the size distribution of the entire set of 527 plasmids. Thus, the antibiotic resistance genes we examined are most closely associated with smaller plasmids of the Enterobacteriaceae family, which indicates that a subpopulation of plasmids is probably responsible for dissemination of these resistance traits in nature.

5. Conclusions

We present an *in silico* approach for simultaneously establishing the genetic relationships among 527 Gram-negative bacterial plasmids. The method uses complete gene sequences for plasmids with at least 50 coding genes to create 100 virtual arrays. These arrays are used to construct an initial tree by consensus that can be refined cluster by cluster using multiple-sequence alignment of conserved proteins within each cluster. While it is not possible to confirm the accuracy of the consensus trees, known relationships from both *Borrelia* and IncA/C plasmids were reflected accurately in our analysis. Based on our results, one can construct additional hypotheses about both inter- and intra-genus transmission of plasmids.

Authors’ contributions

YZ and SLB performed the research for this paper, and all three authors shared in the preparation of the manuscript.

Acknowledgments

The authors are grateful to Carl M. Hansen Foundation for partial support of YZ. In addition, the authors thank Prof. Troy Bankhead for sharing his expertise on *Borrelia*. This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract NOI-
A1-30055 and by the Agricultural Animal Health Program and Agricultural Research Center at Washington State University.

References


Call, D.R., Singer, R.S., Meng, D., Broschat, S.L., Orfe, L.H., et al., 2010. blaCMY-2-positive Inc A/C plasmids from Escherichia coli and Salmonella enterica are a distinct component of a larger lineage of plasmids. Antimicrobial Agents and Chemotherapy 54, 590-596.


Figure Captions

**Fig. 1.** Absolute distance difference \((ADD)\) as a function of the number of arrays \((n)\) used for averaging. ADD value decreases as the increasing number of arrays. i.e. ADD curve is more steep if \(n=20\) than \(n=50\) or \(99\).

**Fig. 2.** Majority voting consensus tree. The majority value of clusters has been listed on each node. A cluster is confirmed if it is in a majority (> 50%) of a tree.

**Fig. 3:** Averaged distance consensus tree.

**Fig. 4:** (a) *Borrelia* cluster from initial tree; NC_012175 is the reference plasmid for determining the conserved proteins. (b) *Borrelia* cluster after refinement.

**Fig. 5:** (a) Mixed bacterial plasmids cluster from initial tree including 8 IncA/C plasmids; NC_012693 is the reference plasmid. (b) Mixed bacterial plasmids cluster after refinement.

**Fig. 6:** Distribution of antibiotic resistance genes *tetA* (red), *cat* (yellow), *tetB* (green), *tetA* and *floR* (blue), and *tetB* and *cat* (black). Thirty-seven of the 40 plasmids carrying antibiotic resistance genes are within a 115-plasmid cluster. Labeling of taxa in the figure is not possible due to size limitations.

Supplemental Material

**File 1** - Excel spreadsheet of 527 Gram-negative plasmids used in the study

**File 2** - Newick standard file for 527 Gram-negative plasmids using majority voting. Because an unrooted consensus tree is order dependent, we used human X and cattle X reproductive proteins as an out group to force clustering of the plasmids.

**File 3** - Newick standard file for 527 Gram-negative plasmids using averaged distances

**File 4** - *Borrelia* sub-group, majority consensus tree

**File 5** - *Borrelia* sub-group, averaged distance consensus tree
File 6 - 115-plasmid cluster with 36 AR genes
Figure 4a

65 conserved proteins

58 conserved proteins

<table>
<thead>
<tr>
<th>Plasmid Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>94a_lp54</td>
</tr>
<tr>
<td>CA-11.2A_lp54</td>
</tr>
<tr>
<td>72a_lp54</td>
</tr>
<tr>
<td>64b_lp54</td>
</tr>
<tr>
<td>WI91-23_lp54</td>
</tr>
<tr>
<td>29805_lp54</td>
</tr>
<tr>
<td>lp54</td>
</tr>
<tr>
<td>Bol26_lp54</td>
</tr>
<tr>
<td>N40_lp54</td>
</tr>
<tr>
<td>JD1_lp54</td>
</tr>
</tbody>
</table>
Figure 4b

Click here to download high resolution image

<table>
<thead>
<tr>
<th>Plasmid Designation</th>
<th>Borrelia burgdorferi</th>
</tr>
</thead>
<tbody>
<tr>
<td>72a_lp54</td>
<td></td>
</tr>
<tr>
<td>94a_lp54</td>
<td></td>
</tr>
<tr>
<td>CA-11.2A_lp54</td>
<td></td>
</tr>
<tr>
<td>N40_lp54</td>
<td></td>
</tr>
<tr>
<td>29805_lp54</td>
<td></td>
</tr>
<tr>
<td>WI91-23_lp54</td>
<td></td>
</tr>
<tr>
<td>JD1_lp54</td>
<td></td>
</tr>
<tr>
<td>Bol26_lp54</td>
<td></td>
</tr>
<tr>
<td>64b_lp54</td>
<td></td>
</tr>
<tr>
<td>B31_lp54</td>
<td></td>
</tr>
</tbody>
</table>