Plasmid 68 (2012) 133–141

Contents lists available at SciVerse ScienceDirect

Plasmid

journal homepage: www.elsevier.com/locate/yplas

Genetic relationships among 527 Gram-negative bacterial plasmids

Yunyun Zhou [a], Douglas R. Call [a,b], Shira L. Broschat [a,b,*]

[a] School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA
[b] Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA, USA

ARTICLE INFO

Article history:
Received 23 February 2012
Accepted 7 May 2012
Available online 12 May 2012
Communicated by Dr. W.F. Fricke

Keywords:
Gram-negative plasmid
Horizontal gene transfer
Antibiotic resistance genes
Classification
Cluster refinement

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. Finally, the 527-plasmid tree was used to study the distribution of four common antibiotic resistance genes.

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

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

* Corresponding author. Address: School of Electrical Engineering and Computer Science, Washington State University, PO Box 642752, Pullman, WA 99164-2752, USA. Fax: +1 509 335 3818.
E-mail address: shira@eecs.wsu.edu (S.L. Broschat).

0147-619X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.plasmid.2012.05.002

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 and Burge, 1995; Suzuki et al., 2008; van Passel et al., 2006). However, there is some debate over whether plasmid mosaicism can be understood from such features (Campbell et al., 1999; van Passel et al., 2006). 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 2171 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) (Supplementary 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) $\geqq$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 selected 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) and memory is impractical.

To determine the optimum number of arrays, we used the average distance difference (*ADD*) (Fig. 1) as a function of the number of arrays. After virtual hybridization of an array for N different plasmids, an $N \times N$ distance matrix is obtained. Pair-wise comparison of two distance matrices results in an $M = N(N-1)/2$ distance vector. For the *ADD* metric, we sum the absolute difference of the mean