BIOINFORMATICS TOOLS FOR EVALUATING MICROBIAL RELATIONSHIPS

By

DA MENG

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

May 2009

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of DA MENG find it satisfactory and recommend that it be accepted.

Shira L. Broschat, Ph.D., Chair

John H. Miller, Ph.D.

Douglas R. Call, Ph.D.

ACKNOWLEDGMENTS

My first and most earnest acknowledgment must go to my advisor and chair of my Thesis Committee Dr. Shira L. Broschat. Shira has been instrumental in ensuring my academic, professional, financial, and moral well being ever since I joined her group. In every sense, none of this work would have been possible without her. Many thanks also go to my committee members Drs. Douglas R. Call and John H. Miller for their valuable time and suggestions, which significantly improved the quality of this dissertation.

Far too many people to mention individually have assisted me in so many ways during my work at WSU. In particular, I would like to thank Drs. Margaret A. Davis and Thomas E. Besser who provided valuable biological data and knowledge I needed for my work. I also want to thank all the members of the Call Lab and Shira's group. As co-workers, they provided me with a great working environment. I want to thank Mr. David Seamans and all his family members for their kind help.

I would like to thank the School of EECS and the Department of Veterinary Microbiology and Pathology for providing help with my research. Support from Drs. Ali Saberi and Zhe Dang made my graduate studies possible.

My final, and most heartfelt, acknowledgment is to my wife and daughter. I want to thank them for their patience and endless support. Without them I just could not have finished all this work.

BIOINFORMATICS TOOLS FOR EVALUATING MICROBIAL RELATIONSHIPS

Abstract

Da Meng, Ph.D.
Washington State University
May 2009

Chair: Shira L. Broschat

Recent years have seen the rapid development of microbial genomics. The vast amounts of microbial genomic information make it possible to study microorganisms systematically. However, how to manipulate the huge amount of available data, how to retrieve genomic information effectively, and how to effectively process the large scale data are big challenges. To cope with these difficulties, more sophisticated informatics methods have been widely used and have become an essential part of genomics.

My research work is focused on developing new strategies for several of the challenges mentioned above. First, we describe a new software tool, PLasmid Analysis System and Marker IDentification (PLASMID), for selecting an optimal set of probes for the design of a classification microarray. The tool provides the user with several clustering methods, a probe ranking method, probe redundancy reduction, and probe selection using stepwise discriminant analysis. The software package has been applied to data from a mixed-plasmid microarray, a virtual mixed-genome microarray, and an expression microarray.

Second, to increase discrimination accuracy, we have developed a fusion algorithm that combines the information obtained from both Pulsed-Field Gel Electrophoresis (PFGE) and Multiple-Locus Variable-Number Tandem Repeat Analysis (MLVA) assays to obtain phylogenetic relationships. Results are assessed by comparison with phage-typing assays and with known epidemiological relationships. Our analysis shows that the fusion algorithm provides an improved ability to discriminate between bacterial isolates and to infer phylogenetic relationships compared with using either PFGE or MLVA analysis alone.

Many studies have shown that horizontal gene transfer (HGT) is common among microbes. HGT can lead to mosaic-like gene sequences in plasmids which makes it a challenge to build robust phylogenetic trees. Simply applying existing phylogenetic methods to study the evolution of plasmids may lead to questionable results. When multiple sequence alignment is used, most phylogenetic methods assume the sequences are homologous. We consider several features of plasmids that affect phylogeny analysis and introduce a method for establishing reliable phylogenetic relationships.

TABLE OF CONTENTS

CHAPTER

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Figure **Page**

DEDICATION

To my wife Si Wu and daughter Catherine Meng

To my parents Zhaozhu Meng and Xuezhi Zhang

CHAPTER 1

Introduction

1.1 Microbiology in the genomic era

The last 20 years has seen the rapid development of prokaryotic genomics. Since *Haemophilus influenzae* was sequenced in 1985 [1], about 880 complete bacterial genomes have been sequenced, and currently there are about 1000 ongoing genome projects (as of April 2009 in the NCBI database http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). Prokaryotic genomics has had a revolutionary impact on our view of the microbial world and also on the methodologies for microbiological studies.

1.1.1 Complexity of microbial genomes

Comparative genomics has revealed that microbial genomes are very diverse. This is due to the complicated nature of microbial evolution. Differing from eukaryotic genomes where mutations play a key role in evolution, the contents of prokaryotic genomes are also changed by gene losses, gene rearrangements, horizontal gene transfer, and so on [2, 3]. This means that even strains from the same species can differ significantly. For example, two *Escherichia coli* strains (O157:H7 and K-12) have more than 1000 different genes [4].

The dynamic nature of microbial genomes complicates several tasks in microbiology studies. One of these is the development of strategies to prevent and treat microbe-related diseases. Since microbe-related diseases are common threats to the public health, microbes (especially bacteria) have been studied for many years. One point of progress was the introduction of antibiotics to treat bacterial infections. However, the use of antibiotics has been challenged by the emergence of antibiotic resistance among bacteria.

Antibiotic resistance has been found to exist widely in bacterial species [5]. It is believed that antibiotic resistance evolves via natural selection. However, antibiotic resistance also can be introduced to bacteria via horizontal gene transfer [5]. Plasmids play an important role in this process. Plasmids are extrachromosomal genetic elements that constitute up to 10% of the total DNA found in many species of bacteria [6]. Because plasmids are capable of cell-to-cell transfer between bacterial species, genes harbored by plasmids are widely shared, playing a critical role in the evolution of bacteria [7]. Establishing accurate relationships between plasmids will help us to understand an important factor in the dissemination of antibiotic resistance genes, and establishing accurate relationships between bacteria will help us to identify the factors that cause diseases, the risks of outbreaks, and methods for preventing disease transmission. Unfortunately, the complexity of microbial genomes is apparent when we try to compare the genetic contents of strains and to build a phylogeny tree from them [3].

1.1.2 Molecular methods to obtain data

One characteristic of microbiology studies in the genomics era currently is that we can generate a huge amount of data efficiently. Numerous different genomics-based experimental methods are available. These methods are usually called molecular methods since they are often based on genetic characteristics. Compared to traditional phenotype-based methods, molecular methods are cost effective, easy to implement, and generate highly-discriminatory data [8]. Of these methods, pulse-field gel electrophoresis (PFGE) is considered the gold standard. Multiple-locus variable-number tandem repeat analysis (MLVA) assays are also a potentially powerful alternative or complementary typing tool.

Pulsed-field gel electrophoresis

Pulsed-field gel electrophoresis (PFGE) is one of the most reproducible and highly discriminatory typing techniques, and it has been widely and successfully used for subtyping a variety of *Salmonella enterica* serovars [9–11]. One of the primary advantages of PFGE is that protocols are relatively simple to standardize, results are robust, and for many situations the technique is able to discriminate between closely related strains. In addition, extension of the assay to new serovars does not require a great deal of modification as might be required with polymerase chain reaction (PCR) dependent procedures. PFGE involves size separating chromosomal DNA macro-restriction fragments in agarose gels, and strains are typed depending on the resulting band pattern observed after electrophoresis has been completed. Difficulties arise when strains are very closely related (i.e., poor discrimination) or when bands either co-migrate in the gel or when identically sized bands represent completely different fragments of chromosomal DNA [12]. These complications are more pronounced when a large number of bands are generated by the restriction digest. In addition, while band patterns convey a crude degree of genetic relatedness, a large number of independent restriction digests would be needed to infer an accurate phylogeny between isolates [12]. A final and somewhat unavoidable criticism of PFGE is that the procedure is time-consuming and not conducive to high-throughput screening [13].

Multiple-locus variable-number tandem repeat analysis

Multiple-locus variable-number tandem repeat analysis (MLVA) is a PCR-based technique that relies on amplification of chromosomal or plasmid DNA that encompasses short tandem repeats of a DNA sequence. The tandem repeats are prone to higher than background mutation rates due to DNA-strand slippage during DNA replication [14], and thus the amplified fragments will vary in length depending on the number of repeats harbored at a given locus. Different fragment lengths are tallied either as the total length (base pairs) or the estimated number of repeat units, and each discretely sized fragment is considered a unique allele for the locus under investigation. Because the technique can be multiplexed and automated, it is conducive to rapid and relatively high throughput typing needs. MLVA assays are relatively robust [15], and while not perfect, these assays can provide phylogenetically informative information even with a limited number of loci [16]. While access to a sequenced genome dramatically speeds the ability to establish new assays [17], this is not a requisite to assay development. The primary limitations of the technique include the potential need for a new set of loci for every species or serovar under investigation and the fact that some loci are very unstable and can disappear from some strains or lineages (this produces the equivalent of an uninformative null allele). Mutation rates can also vary between loci [18, 19], which, if ignored, can introduce bias into phylogenetic analyses.

Other molecular methods

Several other molecular methods are also widely used. Multi-locus sequence typing (MLST) characterizes isolates of bacterial species using the sequences of internal fragments (usually 450-500 bp) of housekeeping genes [20, 21]. The great advantage of MLST is that it is an unambiguous typing method. However, it becomes very costly for a large number of strains. Randomly amplified polymorphic DNA (RAPD) is based on the amplification of a random DNA segment under the assumption that the patterns of bands may be different for individuals in a population or closely related species [22, 23].

Much work has been done to compare different molecular methods, and it has been shown that the discriminatory power depends not only on the techniques themselves, but also on the types of strains to be studied [8, 24].

1.1.3 High-throughput technology: DNA microarrays

DNA microarrays provide a powerful, high-throughput genomic method that has been widely used in biological studies. To construct a DNA microarray, single-strand fragments of DNA (also called probes) representing the genes of an organism are attached to a surface of glass or plastic. Each fragment can bind to a complementary DNA or RNA strand. Typically, more than 30,000 spots can be put on one slide, and it is possible to create a microarray representing every gene in a genome. Thus, microarrays can provide genome-wide information which allows a comprehensive genetic analysis of an organism or a sample.

DNA microarrays have been used for genotyping, expression analysis, and studies of protein-DNA interactions [25]. When used for assessing the genetic relationships of bacterial strains, microarrays may be whole-genome microarrays composed of open reading frames (ORFs) of one complete genome sequence [26, 27]. However, this type of microarray is limited by the requirement of representing one complete reference sequence which may not contain genetic content specific to nonsequenced strains. One possible improvement is to include specific genes from multiple whole-genome sequences or to use mixed-genome microarrays (MGMs) which use randomly-selected gene fragments from many strains of bacteria as probes [28–32].

1.2 Bioinformatics: From data to knowledge

The vast amount of microbial genomic information makes it possible to study microorganisms systematically. However, how to manipulate the huge amount of available data, how to retrieve genomic information effectively, and how to process the large scale data efficiently are all challenging problems. Because of these problems, the field of bioinformatics has emerged and has become an integral part of microbial studies [33].

1.2.1 Databases

Various databases have been established for storing genomic data, and the internet makes it possible for these data to be accessed and shared by the public. Since there are different types of genomic data, it is impossible to build one database containing all data. Currently there are two types of genomic databases. Primary databases contain sequences and structures (for example, NCBI GenBank) and related annotations, bibliographies, and cross-references to other databases and provide the basis for biological studies; secondary databases contain biological knowledge obtained by analyzing genomic sequences and structure data. The database of Clusters of Orthologous Groups of proteins (COGs http://www.ncbi.nlm.nih.gov/COG), for example, contains information for phylogenetic analysis [34]. The Ribosomal Database Project (RDP) provides ribosome related data and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences [35]. Several secondary databases are listed in Table 1, and the number of such databases continues to increase. Knowledge from these databases can help to process biological data efficiently. For example, the Gene Ontology database has been used to process microarray datasets [36].

1.2.2 Data retrieval methods

In order to use the information available in databases, an efficient information retrieval method should be used to obtain all related information quickly. Such methods differ, depending on the type of data to be retrieved. FASTA and BLAST are the two most widely used methods for retrieving sequence data.

FASTA was the first fast sequence searching algorithm used for comparing a query sequence against a database [37]. The FASTA algorithm performs a rapid and approximate search for matched sequence segments followed by application of the Smith-Waterman alignment algorithm [38] to these segments.

BLAST (Basic Local Alignment Search Tool) is a rapid sequence database search tool which is more efficient than FASTA. BLAST generates a list of all possible words of length k (protein sequence: $k = 3$ and nucleotide sequences: $k = 12$) in the query sequence which is then expanded by finding high-scoring words based on a scoring matrix (such as Blosum62 or PAM120) [37]. After forming the query words, the database is searched for these words. When an exact match is found, the algorithm looks in both directions for the rest of the sequence. The output of BLAST is a list of high-scoring segment pairs (HSPs) and an "E value" which is an estimate of the probability of finding an HSP with score S. The E value is often used as a standardized measure for estimating the statistical significance of sequence similarity.

1.2.3 Computational methods

A number of computational methods have been developed and used in genomic studies. Of these methods, genetic sequence alignment is the foundation for many other methods and widely used in comparative genomics. A good alignment method should give biologically meaningful results and at the same time be computationally efficient.

There are two types of alignment methods, local alignments and global alignments. The former methods try to identify similar segments between two sequences while the latter try to align the entire length of two sequences. Methods for aligning two sequences are called pairwise alignment methods. BLAST and FASTA are two widely used pairwise alignment methods. These methods can be extended to multiple sequences; however, multiple sequence alignment (MSA) is more complicated.

ClustalW [39] is a widely used MSA method which is efficient for aligning protein sequences and short nucleotide sequences. However, it may fail for distantly related sequences [40]. PSI-BLAST [41] is a very successful method for detecting weak similarities. Two recently developed algorithms, MLAGAN and MAVID [42, 43], are designed for global alignment of both evolutionarily close and distant megabase-length genomic sequences. However, a phylogenetic tree is assumed to be known for use with MLAGAN. MAVID is a progressive global alignment program that works by recursively aligning the 'alignments' at ancestral nodes of the guide phylogenetic tree. MAUVE is used for comparing long genome sequences efficiently and takes into account possible large-scale evolutionary events among sequences [44].

1.2.4 From data driven to principle driven: Current status and future directions

Currently, the focus of bioinformatics is to create new computational methods for collecting and analyzing data for specific biological problems as well as to create databases for saving more and more biological knowledge. Methodologies are data driven. However, this is only the first step toward understanding the complexity of biological systems [33]. The future will see more mathematics, physics, and chemistry in the modeling of underlying biological processes in such a way that the biological phenomena can be explained precisely. Systems biology is the first attempt in this direction; its goals are to understand the structure of a biological system, for example, gene networks at the cellular level, and to understand the dynamics of the system [45].

1.3 Phylogenetic analysis

The goal of phylogenetic analysis is to reconstruct the evolutionary history of a set of organisms. In molecular epidemiology, it helps to elucidate mechanisms that lead to microbial outbreaks and epidemics.

Phylogenetic analysis usually begins with multiple sequence alignment of the sequences of a set of organisms. After obtaining an MSA, a number of different phylogenetic methods can be used to compute phylogenetic trees. These methods can be broadly classified into maximum parsimony, distance, and maximum likelihood methods [46–49]. The difference between these methods is how they define which tree is best among all possible trees. Maximum parsimony tries to find an evolutionary tree or trees which require a minimum number of changes from the common ancestral sequences. For maximum likelihood methods, given the MSA, the probability of a specific tree occurring is computed, and the one or ones with the highest values are considered to be the evolutionary tree or trees. Distance-based methods construct a tree by hierarchical clustering methods using a distance matrix for all organisms that is computed using MSA. To use MSA for phylogenetic analysis, it is necessary to assume an underlying mutation model. Of the ones that have been proposed, the Jukes-Cantor (JC) model [50] is the simplest one. In the JC model, each base in a DNA sequence has an equal mutation rate and all complementary pairs of the four nucleotides A, T, C, and G have equal substitution rates. These assumptions are not realistic in practice, so many complex models have been proposed and tried. Successful phylogenetic analysis requires a suitable model.

Phylogenetic analysis of microbial strains is problematic due to its dynamic nature [51]. Different genes among strains may contain contradictory information about their evolution. Consensus trees have been suggested as a solution. An alternative is the introduction of networks that represent the evolutionary relationships between microbial strains [52]. According to a recent summary, currently there are about 386 available packages and 52 servers for phylogenetic analysis. These tools are different with respect to the methods implemented and the mutation models used.

1.4 Bioinformatics challenges

Many bioinformatics tools have been borrowed from the fields of artificial intelligence, data mining, and statistical methods. However, the characteristics of biological data may differ significantly from those of the original data for which the methods were developed. Though many computational methods have been introduced for genomic data analysis based on these methods, several challenges still exist.

One challenge is the high dimensionality of data that results from high-throughout methods, particularly microarray data. A typical DNA microarray might have thousands of features (probes) for, at most, one hundred samples. For traditional methods, including clustering methods and classification methods, the rule of thumb is to use at least ten samples for each feature [53]. Feature reduction is typically required before these sorts of analyses can be performed [54].

Another challenge is integrating data from different sources. These datasets might show a high degree of heterogenity and might also vary in quality. They might be generated using different experimental platforms or based on different molecular methods. Using these data together efficiently requires developing suitable bioinformatics methods. Of these methods, the simplest one is to put several datasets together to build a larger dataset and then analyze this larger dataset. However, this method will not work if the formats of the original datasets differ. Furthermore, the best processing methods for different datasets are not the same. For example, Dice coefficents worked well for some PFGE data we used but did not work well for some VNTR data. Thus, it might be an impossible task to choose an optimal method for a combined dataset. An alternate method is to process different datasets separately and then combine the results to obtain the final result [55–58]. The difficulties with this kind of method, however, are determining the extent to which the different sources of data should contribute and explaining the combined results.

1.5 Consideration of programming languages

Selecting a programming language to use for a study might seem to be a trivial matter at first. However, given the need for maintenance and the possibility of use on different platforms, it is not actually trivial. Several factors play important roles in the selection. The first is issue of speed. If speed is a major problem, C or C++ is usually preferred. If it is important to interact with databases to retrieve data or to write the output into a database, choosing a language such as Java, Perl, or python will simplify the work. In our case, we chose to use C and Java together or C and python together.

A key challenge in designing new computational methods to analyze genomic data is verifying the results. One possibility is to build a simulator to generate simulated data. In phylogenetic analysis, there are many simulators available for generating simulated data. However, the simulated data may be biased in the evaluation. For such cases, accumulating more analytical data and building a gold standard would be very helpful. Finally, choosing a standard format for representing data is important because it decreases the time needed to change data formats when different tools are used.

1.6 Objectives of this research

The discovery of genes and DNA made it possible to study organisms directly at the genetic level [59]. Compared to traditional biotechnology methods, modern techniques give more precise results in less time. These methods, especially high-throughput genomic technologies, provide efficient ways to study many important biological questions, from antibiotic resistance to cancer. More and more data have been generated, and data acquisition is increasing exponentially.

Considering the complexity of bacterial genetics, however, acquiring more biological data and even genome sequences is just the beginning. It is still very challenging work to obtain biologically interesting information from these data. To do so requires the development of suitable bioinformatics methods and associated software tools. This is the focus of this thesis. In Chapter 2 we introduce a fusion method for improving the accuracy of phylogenetic relationships by combining heterogeneous information. In Chapter 3 a method for improving the design of classification microarrays is presented. In Chapter 4 a method for constructing reliable evolutionary relationships among plasmids is discussed. Possible directions for future work are given in Chapter 5.

Table 1. List of databases

	Database	URL	
Primary	GenBank	http://www.ncbi.nlm.nih.gov/Genbank/	DNA sequences
	EMBL	http://www.ebi.ac.uk/embl/	DNA sequences
	SwissProt	http://www.ebi.ac.uk/uniprot/	protein sequences
	EC-ENZYME	http://ca.expasy.org/enzyme/	enzyme nomenclature
	RCSB PDB	http://www.rcsb.org/pdb/home/home.do	biological structures
Secondary	InterPro	http://www.ebi.ac.uk/interpro/	protein function
	PROSITE	http://ca.expasy.org/prosite/	protein function
	Pfam	http://pfam.sanger.ac.uk/	protein function
	SMART	http://smart.embl-heidelberg.de/	protein function
	COG	http://www.ncbi.nlm.nih.gov/COG/	ortholog groups
	GO	http://www.geneontology.org/	gene ontology

List of References

[1] R. Fleischmann, *et al.*, “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd,” *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[2] C. R. Woese, “Bacterial evolution.” *Microbiol. Mol. Biol. Rev.*, vol. 51, no. 2, pp. 221–271, 1987.

[3] W. F. Doolittle, “Phylogenetic classification and the universal tree,” *Science*, vol. 284, no. 5423, pp. 2124–2128, 1999.

[4] N. T. Perna, *et al.*, “Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7,” *Nature*, vol. 409, pp. 529–533, 2001.

[5] P. Boerlin and R. J. Reid-Smith, “Antimicrobial resistance: its emergence and transmission,” *Anim Health Res Rev*, vol. 9, no. Special Issue 02, pp. 115–126, 2008.

[6] C. M. Thomas, *The Horizontal gene pool : bacterial plasmids and gene spread.* Amsterdam, The Netherlands: Harwood Academic, 2000.

[7] D. K. Summers, *The Biology of Plasmids.* Oxford: Blackwell Science, 1996.

[8] F. C. Tenover, R. D. Arbeit, and R. V. Goering, “How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists,” *Infect. Control Hosp. Epidemiol.*, vol. 18, p. 426439, 1997.

[9] K. Kubota, T. J. Barrett, M. L. Ackers, P. S. Brachman, and E. D. Mintz, “Analysis of *Salmonella enterica* serotype typhi pulsed-field gel electrophoresis patterns associated with international travel,” *J Clin Microbiol*, vol. 43, no. 3, pp. 1205–1209, 2005.

[10] I. L. Ross and M. W. Heuzenroeder, “Use of AFLP and PFGE to discriminate between *Salmonella enterica* serovar Typhimurium DT126 isolates from separate food-related outbreaks in Australia,” *Epidemiol Infect*, vol. 133, no. 04, pp. 635–644, 2005.

[11] J. M. Ling, N. W. S. Lo, Y. M. Ho, K. M. Kam, N. T. T. Hoa, L. T. Phi, and A. F. Cheng, “Molecular methods for the epidemiological typing of *Salmonella enterica* serotype typhi from Hong Kong and Vietnam,” *J Clin Microbiol*, vol. 38, no. 1, pp. 292–300, 2000.

[12] M. A. Davis, D. D. Hancock, T. E. Besser, and D. R. Call, “Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of *Escherichia coli* O157:H7,” *J Clin Microbiol*, vol. 41, no. 5, pp. 1843–9, 2003.

[13] L. J. Hathaway, S. Brugger, A. Martynova, S. Aebi, and K. Muhlemann, "Use of the Agilent 2100 bioanalyzer for rapid and reproducible molecular typing of *streptococcus pneumoniae*," *J Clin Microbiol*, vol. 45, no. 3, pp. 803–809, 2007.

[14] A. Van Belkum, S. Scherer, L. van Alphen, and H. Verbrugh, "Short-sequence DNA repeats in prokaryotic genomes," *Microbiol Mol Biol Rev*, vol. 62, no. 2, pp. 275–93, 1998.

[15] K. L. Hopkins, C. Maguire, E. Best, E. Liebana, and E. J. Threlfall, "Stability of multiple-locus variable-number tandem repeats in *Salmonella enterica* serovar typhimurium," *J Clin Microbiol*, vol. 45, no. 9, pp. 3058–61, 2007.

[16] B. A. Lindstedt, T. Vardund, L. Aas, and G. Kapperud, "Multiple-locus variable-number tandem-repeats analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium using PCR multiplexing and multicolor capillary electrophoresis," *J Microbiol Methods*, vol. 59, no. 2, pp. 163–72, 2004.

[17] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic Acids Res*, vol. 27, no. 2, pp. 573–580., 1999.

[18] A. J. Vogler, C. Keys, Y. Nemoto, R. E. Colman, Z. Jay, and P. Keim, "Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7," *J Bacteriol*, vol. 188, no. 12, pp. 4253–63, 2006.

[19] A. J. Vogler, C. E. Keys, C. Allender, I. Bailey, J. Girard, T. Pearson, K. L. Smith, D. M. Wagner, and P. Keim, "Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*," *Mutat. Res.*, vol. 616, no. 1-2, pp. 145–58, 2007.

[20] M. C. J. Maiden, *et al.*, "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms," *PNAS*, vol. 95, no. 6, pp. 3140–3145, 1998.

[21] R. Urwin and M. C. J. Maiden, "Multi-locus sequence typing: a tool for global epidemiology," *TIM*, vol. 11, no. 10, pp. 479 – 487, 2003.

[22] J. Welsh and M. McClelland, "Fingerprinting genomes using PCR with arbitrary primers," *Nucleic Acids Res*, vol. 18, no. 24, pp. 7213–7218, 1990.

[23] J. G. Williams, A. R. Kubelik, K. J. Livak, J. Rafalski, and S. V. Tingey, "DNA polymorphisms amplified by arbitrary primers are useful as genetic markers," *Nucleic Acids Res*, vol. 18, no. 22, pp. 6531–6535, 1990.

[24] K.-W. Chen, H.-J. Lo, Y.-H. Lin, and S.-Y. Li, "Comparison of four molecular typing methods to assess genetic relatedness of *Candida albicans* clinical isolates in Taiwan," *J Med Microbiol*, vol. 54, no. 3, pp. 249–258, 2005.

[25] V. Trevino, F. Falciani, and H. A. Barrera-Saldaa, “DNA microarrays: a powerful genomic tool for biomedical and clinical research,” *Mol Med.*, vol. 13, no. 9-10, pp. 527 – 541, 2007.

[26] S. Naser, F. L. Thompson, B. Hoste, D. Gevers, K. Vandemeulebroecke, I. Cleenwerck, C. C. Thompson, M. Vancanneyt, and J. Swings, “Phylogeny and identification of Enterococci by *atpa* gene sequence analysis,” *J Clin Microbiol*, vol. 43, no. 5, pp. 2224–2230, 2005.

[27] S. Porwollik, R. M.-Y. Wong, and M. McClelland, “Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis,” *PNAS*, vol. 99, no. 13, pp. 8956–8961, 2002.

[28] D. Call, M.-S. Kang, J. Daniels, and T. Besser, “Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray,” *J Appl Microbiol*, vol. 100, no. 1, pp. 15–28, 2006.

[29] M. K. Borucki, M. J. Krug, W. T. Muraoka, and D. R. Call, “Discrimination among *Listeria monocytogenes* isolates using a mixed genome DNA microarray,” *Vet Microbiol*, vol. 92, no. 4, pp. 351–362, 2003.

[30] M. K. Borucki, S. H. Kim, D. R. Call, S. C. Smole, and F. Pagotto, “Selective discrimination of *Listeria monocytogenes* epidemic strains by a mixed-genome DNA microarray compared to discrimination by pulsed-field gel electrophoresis, ribotyping, and multilocus sequence typing,” *J Clin Microbiol*, vol. 42, no. 11, pp. 5270–5276, 2004.

[31] D. R. Call, M. K. Borucki, and T. E. Besser, “Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of *Listeria monocytogenes*,” *J Clin Microbiol*, vol. 41, no. 2, pp. 632–639, 2003.

[32] M. Soule, E. Kuhn, F. Loge, J. Gay, and D. Call, “Using DNA microarrays to identify library-independent markers for bacterial source tracking,” *Appl. Environ. Microbiol.*, vol. 72, pp. 1843–1851, 2006.

[33] M. Kanehisa and P. Bork, “Bioinformatics in the post-sequence era,” *Nat Genet*, vol. 33, pp. 305 – 310, 2003.

[34] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin, “The COG database: new developments in phylogenetic classification of proteins from complete genomes,” *Nucleic Acids Res*, vol. 29, no. 1, pp. 22–28, 2001.

[35] J. R. Cole, *et al.*, “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis,” *Nucleic Acids Res*, vol. 37, no. suppl_1, pp. D141–145, 2009.

[36] G. Gamberoni, S. Storari, and S. Volinia, “Finding biological process modifications in cancer tissues by mining gene expression correlations,” *BMC Bioinformatics*, vol. 7, no. 1, p. 6, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/6

[37] D. Lipman and W. Pearson, “Rapid and sensitive protein similarity searches,” *Science*, vol. 227, no. 4693, pp. 1435–1441, 1985.

[38] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *J Mol Bio*, vol. 147, no. 1, pp. 195 – 197, 1981.

[39] J. D. Thompson, D. G. Higgins, and T. J. Gibson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” *Nucleic Acids Res*, vol. 22, pp. 4673–4680, 1994.

[40] R. Van Hellemont, P. Monsieurs, G. Thijs, B. De Moor, Y. Van de Peer, and K. Marchal, “A novel approach to identifying regulatory motifs in distantly related genomes,” *Genome Biol.*, vol. 6, no. 13, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2005-6-13-r113

[41] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucl. Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[42] N. Bray and L. Pachter, “MAVID: Constrained ancestral alignment of multiple sequences,” *Genome Research*, vol. 14, no. 4, pp. 693–699, 2004.

[43] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. C. S. Program, E. D. Green, A. Sidow, and S. Batzoglou, “LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA,” *Genome Research*, vol. 13, no. 4, pp. 721–731, 2003.

[44] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna, “Mauve: multiple alignment of conserved genomic sequence with rearrangements,” *Genome Res*, vol. 14, no. 7, pp. 1394–1403, 2004. [Online]. Available: http://genome.cshlp.org/content/14/7/1394.abstract

[45] N. A. Van Riel, “Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments,” *Brief Bioinform*, vol. 7, no. 4, pp. 364–374, 2006. [Online]. Available: http://bib.oxfordjournals.org/cgi/content/abstract/7/4/364

[46] H. Li and W. Wang, “Dissecting the transcription networks of a cell using computational genomics,” *Curr Opin Genetics Dev*, vol. 13, no. 6, pp. 611 – 616, 2003.

[47] M. J. Aardema and J. T. MacGregor, “Toxicology and genetic toxicology in the new era of “toxicogenomics”: impact of “-omics” technologies,” *Mutat. Res. Fundam. Mol. Mech. Mugag.*, vol. 499, no. 1, pp. 13–25, 2002.

[48] J. Raes, K. U. Foerstner, and P. Bork, “Get the most out of your metagenome: computational analysis of environmental sequence data,” *Curr Opin Microbiol*, vol. 10, no. 5, pp. 490 – 498, 2007.

[49] G. G. Loots, “Chapter 10 genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis,” in *Long-Range Control of Gene Expression*, ser. Advances in Genetics, V. van Heyningen and R. E. Hill, Eds. Academic Press, 2008, vol. 61, pp. 269 – 293.

[50] J. T. H. and C. R. Cantor, “Evolution of protein molecules,” in *Mammalian protein metabolism*, H. N. Munro, Ed. New York: Academic Press, 1969, pp. 21–132.

[51] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, “Genome-scale approaches to resolving incongruence in molecular phylogenies,” *Nature*, vol. 425, pp. 798 – 804, 2003.

[52] D. H. Huson and D. Bryant, “Application of phylogenetic networks in evolutionary studies,” *Mol Biol Evol*, vol. 23, no. 2, pp. 254–267, February 2006.

[53] A. K. Jain, R. P. Duin, and J. Mao, “Statistical pattern recognition: a review,” *IEEE Trans Pattern Anal Mach Intell*, vol. 22, pp. 4–37, 2000.

[54] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,” *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[55] E. V. Koonin and M. Y. Galperin, “Prokaryotic genomes: the emerging paradigm of genome-based microbiology,” *Curr Opin Genetics Dev*, vol. 7, no. 6, pp. 757 – 763, 1997.

[56] I. Uchiyama, “MBGD: microbial genome database for comparative analysis,” *Nucleic Acids Res*, vol. 31, no. 1, pp. 58–62, 2003.

[57] R. Zhang and C.-T. Zhang, “The impact of comparative genomics on infectious disease research,” *Microbes Infect*, vol. 8, no. 6, pp. 1613 – 1622, 2006.

[58] D. M. Raskin, R. Seshadri, S. U. Pukatzki, and J. J. Mekalanos, “Bacterial genomics and pathogen evolution,” *Cell*, vol. 124, no. 4, pp. 703 – 714, 2006.

[59] J. R. Stinchcombe and H. E. Hoekstra, “Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits,” *Heredity*, vol. 100, no. 2, pp. 158–170, 2007.

CHAPTER 2

A Java-based Tool for the Design of Classification Microarrays

2.1 Abstract

Mixed-plasmid and mixed-genome microarrays can be used to compare the genetic content of plasmid and bacterial genomes for classification purposes. Selection of probes is a key factor in designing successful mixed microarrays because redundant sequences are inefficient and limited representation of diversity can restrict their application. We have developed a Java-based software tool, called PLASMID, for use in selecting the minimum set of probe sequences needed to distinguish between different groups of plasmids or bacteria.

The software program was successfully applied to several different sets of data. The utility of PLASMID was illustrated using existing plasmid microarray data as well as data from a virtual mixed-genome microarray constructed from different strains of *Streptococcus*. Moreover, use of public data from expression microarray experiments demonstrated the generality of PLASMID.

In this chapter we describe a new software tool for selecting a set of probes for a classification microarray. While the tool was developed for the design of mixed microarrays—and mixed-plasmid microarrays in particular—it can also be used to design expression arrays. The user can choose from several clustering methods (including hierarchical, non-hierarchical, and a model-based genetic algorithm), several probe ranking methods, and several different display methods. A novel approach is used for probe redundancy reduction, and probe selection is accomplished via stepwise discriminant analysis. Data can be entered in different formats (including Excel and comma-delimited text), and dendrogram, heat map, and scatter plot images can be saved in several different formats (including jpeg and tiff). Weights generated using stepwise discriminant analysis can be stored for analysis of subsequent experimental data. Additionally, PLASMID can be used to construct virtual microarrays with genomes from public databases.

2.2 Background

How to use genetic information for classification is part of the larger question of how to "capture" and quantify genetic diversity for any group of heterogeneous entities. The majority of DNA microarrays in use today are created from single genomes that do not reflect the heterogeneity of most populations. Mixed-DNA microarrays offer an alternative for "capturing" genetic diversity for classification purposes. Mixed-genome or mixed-plasmid microarrays include DNA from one or more reference strains or plasmids that is shotgun-cloned, and a microarray is generated from randomly selected, PCR-amplified clone inserts [1, 2]. Unlike most fingerprinting tools, the mixed-array format permits identification of informative probes that can be retrieved from the clone library for sequencing [3]. However, redundant sequences and limited representation of diversity can limit the application of these tools [2, 4]. Fortunately, a growing public database of genomes offers a new opportunity to incorporate non-redundant and diverse sequences into a mixed-microarray format. These arrays can be used to quickly assess the distribution of genetic diversity across multiple species and niches.

This work focuses on the optimal design of classification arrays. By optimal we mean minimizing the complexity and cost of an array by using as few probes as possible while still rendering sufficient information to discriminate between strains and to avoid bias. Selection of an optimal set of probes is a key factor in designing a successful mixed microarray to suit a particular need. The effects of probe length and the number of probes per gene have been discussed in [5]. A method for finding unique and valid oligonucleotides or probes was proposed in [6], which tries to identify probes for a gene such that there is no similar occurrence in other locations of a genome. A tool for choosing optimal DNA oligos is reported in [7], which identifies oligo sequences that occur in members of the target group but not in the non-target group. However, these methods are used for genome-wide probe selection and are not intended to identify minimum probe sets for classification problems.

A number of methods have been introduced for designing optimal probe sets. Pre-filtering methods [8] use clustering of all probes to find similar probe groups. Similar probes are discarded; the remaining probes are ranked, and top-ranked probes are kept for further analysis. A similar method [9] uses K-means to cluster all genes, and the means of different gene clusters are used as prototype genes. The limitation of these methods is that the number of clusters must be specified. A hybrid approach [10] ranks the probes first and selects a set of top-ranked probes. Hierarchical clustering is then used on these probes to generate a dendrogram. The optimal probes are selected by collapsing dense clusters. In this manner a small set of probes is identified that has a similar prediction accuracy to one that uses more probes.

The methods described above identify optimal probes using training data when the structure of the data is given. Such information, however, is usually unavailable for microarray data sets. A tool is still needed to help design mixed microarrays when prior knowledge of a microarray data set is unavailable. The focus of this chapter is a software program, PLASMID, used for selecting a minimum set of probe sequences needed to distinguish between groups of plasmids or bacteria. Data used to identify probe candidates can be either existing microarray data (or similar hybridization data) or sequence data from a public database such as GenBank. The latter are converted to “probe” sequences, and virtual hybridization is used to generate data for probe selection [11]. To demonstrate the generality of PLASMID, we include an example whereby the program can also be applied to develop a minimum probe set to distinguish between two classes of leukemia using data from an expression array.

2.3 Methods

2.3.1 Finding meaningful clusters in hybridization data

Finding meaningful clusters in a given set of hybridization or sequence data is a key task in optimal microarray design; our tool provides several clustering options. Clustering methods can be classified into two general groups: distance-based methods and model-based methods. Distance-based methods are either non-hierarchical or hierarchical, and each method has its particular strengths and weaknesses. Currently our tool includes the K-means non-hierarchical clustering algorithm and hierarchical clustering by means of Unweighted Pair Group Method with Arithmetic mean (UPGMA), neighbor joining, or Ward’s minimum variance method, all of which are widely used in microarray data analysis. A model-based method is also implemented.

Distance metrics. Distance-based methods require use of a distance metric. There are different types of distance metrics that can be used to compute the distance between two objects to be clustered. Selection of a suitable metric is very important for the obtainment of reasonable clustering results. Unfortunately, there are no selection guidelines except to choose the metric that gives the “best” results based on an error function or the ability to classify particular data points. Euclidean distance is the most commonly used metric; a large distance implies low similarity. Pearson’s correlation coefficient is another commonly used metric that measures the extent to which two objects are linearly related. The value of the correlation coefficient ranges from -1 to +1, and values of zero indicate a random relationship between objects. If we assume a microarray data set with n samples and p probes and x_{gs} is the intensity value for sample s at probe g, then distances are calculated using:

1. Euclidean distance
$$d_{ij} = \sqrt{\sum_{g=1}^{p} (x_{gi} - x_{gj})^2} \tag{1}$$

2. Pearson's correlation coefficient
$$d_{ij} = \frac{\sum_{g=1}^{p} (x_{gi} - x_{.i})(x_{gj} - x_{.j})}{\sqrt{\sum_{g=1}^{p} (x_{gi} - x_{.i})^2}\sqrt{\sum_{g=1}^{p} (x_{gj} - x_{.j})^2}} \tag{2}$$

where $x_{.i}$ is the mean intensity of sample i across all probes.

Hierarchical clustering algorithms. In hierarchical clustering, a distance metric is used to calculate the distance (or similarity) matrix for the N samples to be clustered. The clustering algorithm functions as follows [12]:

1. Start by assigning each sample to a cluster. Thus, for N samples, there are N clusters, each containing just one item. Let the distances between the clusters be the same as the distances between the samples they contain.

2. Find the closest pair of clusters and merge them into a single cluster; there are now $N - 1$ clusters.

3. Compute distances between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all samples are clustered into a single group of size N.

After completing these steps, a dendrogram is constructed. It is up to the user to decide how to subdivide the dendrogram into meaningful clusters according to the problem under consideration. In step 3, there are different ways of calculating distances between two clusters: the single-linkage method (also called neighbor joining or minimum method), complete-linkage method (or maximum method), and average-linkage method (also known as UPGMA). In the single-linkage method, the distance between one cluster and another cluster is defined as the shortest distance from any member of one cluster to any member of the other cluster; in complete-linkage, distance is defined as the greatest distance from any member of one cluster to any member of the other cluster; and in average-linkage clustering, distance is the average distance from any member of one cluster to any member of another cluster. Ward's minimum variance is another hierarchical clustering method that is widely used. At each step in the analysis, the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion.

Non-hierarchical clustering algorithm. The K-means clustering method is the most widely used non-hierarchical clustering method [12]. In K-means clustering, all samples in question are initially assigned to k clusters; the number of clusters is assigned *a priori*. The k centroids, one for each cluster, are calculated, and each sample is associated with its nearest centroid. After all samples have been reassigned to different clusters, the centroids are recalculated. This process continues until the centroids no longer change locations. The objective of the K-means clustering method is for the solution to have the minimum intra-cluster variance for all possible cluster partitions.

Model-based genetic clustering. Distance-based methods are simple to use, and the clustering results are easy to explain. However, it is hard to obtain information about the number of clusters, the confidence level of the clustering results, and so on, from these methods. To avoid some of these issues, model-based clustering methods can be used as an alternative. Model-based clustering methods assume that the data can be clustered according to a set of underlying distributions. These underlying distributions can be modeled, and finding a suitable model can be construed as an optimization problem. We assume that M is the underlying model for a data set X, and the best clustering result is represented by partition P. A measure is used to determine which P is most likely for X. In our tool the measure is the likelihood of all possible partitions P. A number of different optimization methods can be used to find the solution for P. In our tool, we have chosen to use a genetic algorithm because of its simplicity and efficiency in addition to its ability to find the optimal solution. Usually model-based clustering methods are based on the Expectation-Maximization (EM) method. However, EM algorithms tend to break down for microarray data because an inversion of the covariance matrix must be performed. In genetic algorithms, a search method is used to circumvent the need for this computation, thereby making genetic model-based methods more stable.

The simplest case is to assume that all samples (e.g., plasmids) x_{ij} in the same cluster i for a given probe j share the same set of normal distributions, $x_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$ where μ is the mean and σ^2 is the variance. To find the best partition

P we want to find the maximum value of the likelihood function $\mathcal{L}(x, \mu, \sigma^2, P)$

$$\begin{aligned} \mathcal{L}(x, \mu, \sigma^2, P) &= \mathcal{L}(x|\mu, \sigma^2, P)\mathcal{L}(\mu, \sigma^2, P) \\ &= \mathcal{L}(x|\mu, \sigma^2, P)\mathcal{L}(\mu, \sigma^2|P)\mathcal{L}(P) \end{aligned}$$

where Bayes' theorem has been used to obtain the equalities. Ideally we can assign the likelihood of different partitions $\mathcal{L}(P)$ according to prior knowledge; including this information will improve the clustering performance. However, we may not have this information *a priori*. In this case, we assume all partitions are equally likely and set $\mathcal{L}(P)$ to one.

We can obtain the likelihood of the product $\mathcal{L}(x|\mu, \sigma^2, P)\mathcal{L}(\mu, \sigma^2|P)$ by choosing μ and σ^2 for a particular distribution. For example, we could assume a normal distribution centered about zero and obtain μ and σ^2; this would give us the maximum likelihood. Instead we assume all possible values of μ and σ^2 for a chosen distribution for each and integrate over these values to obtain the average likelihood for all μ and σ^2. This gives us the marginal likelihood $\mathcal{L}(x|P) = \mathcal{L}(x|\mu, \sigma^2, P)\mathcal{L}(\mu, \sigma^2|P)$. For this work, we assume a normal distribution for μ and an inverse-Γ distribution for σ^2. This gives:

$$\mathcal{L}(x|P) = \prod_k \prod_j \frac{2\sigma_0^2}{\Gamma(1)} \frac{(2\pi)^{-(n_k/2)}}{\sqrt{n_k + 1}} \frac{\Gamma(n_k/2 + 1)}{(2\sigma_0^2 + 0.5(\sum_i x_{k_ij}^2 + \mu_0^2 - \frac{\sum_i (x_{k_ij} + \mu_0)^2}{n_k + 1}))^{(n_k/2+1)}} \tag{3}$$

where k is the index of clusters, j is the index of probes, n_k is the number of samples in the kth cluster, k_i is the index of samples in the kth cluster, and μ_0

and σ_0^2 are the overall mean and variance of all the data [13].

Using this as a measure, the genetic algorithm is used to find the partition that maximizes the likelihood. The steps of the genetic algorithm are summarized as follows:

1. Generate N random partitions. Each partition is represented by a vector $[1\ 2\ 1 \cdots]$ where each term is the index of a cluster.

2. Prior knowledge of pairs of samples highly unlikely to be in the same cluster can be incorporated into the partition likelihood by creating a text file with each pair of samples, together with a small weighting factor, on one line. The weighting factor must be smaller than 1, but how much smaller has to be determined empirically based on the end result. A weighting factor of zero indicates that the pair cannot be in the same cluster.

3. Compute the likelihood $\mathcal{L}$ for all partitions.

4. Repeat the following steps until the maximum iterations (Max) has been reached or the difference between the likelihood of two successive iterations is less than ϵ, where Max and ϵ are given.

   (a) Select the two partitions with the highest scores.

   (b) Do crossover and mutation on these two partitions to generate new partitions. Crossover is accomplished by randomly selecting sections of equal length from each partition and exchanging them. Mutation is performed following crossover and is accomplished by randomly selecting one term in each of the partitions and changing it to a different value.

   (c) Compute the likelihood $\mathcal{L}$ for these two new partitions (offspring).

   (d) Replace the two lowest-ranked partitions with the offspring.

Other measures can be used including Bayesian Information Criteria and minimum description length. These measures will be included in future versions of PLASMID.

2.3.2 Probe ranking for classification

In a DNA microarray data set there are usually many more probes (features) than the number of objects to be classified, and often many of these probes are redundant. Thus, in the design of an optimal probe set for object classification, our goal is to identify and remove irrelevant and redundant probes. Irrelevant probes can be removed using probe ranking. There are two basic approaches to probe ranking: filter techniques and wrapper techniques. Because of their simplicity filter procedures are used most commonly for DNA microarrays. The filter procedure ranks each probe using a metric based on its classification relevance. Top-ranked probes are then selected to perform classification. Numerous filter metrics are described in the literature [14]: probabilistic and distance metrics, dependence measures, scores based on information theory, etc. In our tool, filter metrics are determined using two different statistical tests, the ANOVA-F and Brown-Forsythe tests. Other tests considered were the Welch, adjusted Welch, Cochran, and Kruskal-Wallis test statistics [15].

The test statistic is used as a metric to evaluate the discriminating power of a probe. Higher values represent more discriminating probes. If we assume that for a microarray dataset with n samples and p probes, x_{gs} is the intensity value for sample s at probe g, then the microarray data set can be written in the following matrix form:

$$G = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

Assuming that these n samples belong to k classes, we use y_{ij}^{k} to represent the intensity for the jth sample of the ith class on the kth probe. The ANOVA-F and Brown-Forsythe formulas are given by:

1. ANOVA-F test statistic

$$F = \frac{(n-k)\sum n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{(k-1)\sum(n_i-1)s_i^2} \tag{4}$$

where

$$\bar{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i, \tag{5}$$

$$\bar{y}_{\cdot\cdot} = \sum_{i=1}^{k} n_i\bar{y}_{i\cdot}/n, \tag{6}$$

and

$$s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/(n_i - 1) \tag{7}$$

Under the null hypothesis and assuming no differences in variance, this test statistic follows the F distribution $F_{k-1,n-k}$.

2. Brown-Forsythe test statistic

$$B = \frac{\sum n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum(1 - n_i/n)s_i^2} \tag{8}$$

Under the null hypothesis and assuming no differences in variance, this test statistic follows the F distribution of $F_{k-1,v}$, where

$$v = \frac{[\sum (1 - n_i/n)s_i^2]^2}{\sum (1 - n_i/n)^2 s_i^4/(n_i - 1)} \tag{9}$$

For some applications, clusters may include an insufficient number of samples for meaningful statistical analysis. Such cases can be handled by generating random samples that differ only slightly from the original samples. These samples can be included in the statistical analysis and then discarded without compromising the probe ranking procedure. The purpose of adding these samples is for computational convenience only; they do not add more information.

The end result of the probe ranking function is a list of all probes ranked by their classification relevance. At this point, the user can either stop and use some chosen number of the top-ranked probes for the array probe set or continue with probe reduction and stepwise discriminant analysis to remove redundant probes and assign weights to the probes.

2.3.3 Stepwise discriminant analysis

While the top-ranked probes are informative, at least some of them are likely to convey redundant information. In our tool, redundancy is removed using K-means clustering of probes followed by stepwise discriminant analysis (SDA) [16]. A set of top-ranked probes is clustered into κ groups; probes in the same group are highly correlated with each other but uncorrelated or loosely correlated with probes in other groups. The probe closest to the center of a group is chosen to be representative of that group, and the κ representative probes are used in the SDA that identifies the optimal probe set $\mathcal{G}$. At each step of the SDA, an $\mathcal{F}$ statistic is computed for each probe; this value is used to determine whether including the probe or excluding the probe from $\mathcal{G}$ will significantly improve sample differentiation. The SDA process starts with an empty probe set $\mathcal{G}$, and an iterative process of adding a probe to $\mathcal{G}$ or removing a probe from $\mathcal{G}$ continues until no probes can be added or removed. $\mathcal{F}_{remove}$ is used for the probes in $\mathcal{G}$, and $\mathcal{F}_{enter}$ is used for the probes not in $\mathcal{G}$. The probe in $\mathcal{G}$ with the smallest value of $\mathcal{F}_{remove}$ less than a chosen threshold value, usually 1.0, is removed; the probe not in $\mathcal{G}$ with the largest value of $\mathcal{F}_{enter}$ greater than the threshold value is added to $\mathcal{G}$. The formulas used to compute $\mathcal{F}$ are

$\mathcal{F}$ values:

$$\mathcal{F}_{remove} = \frac{n-r-q+1}{q-1}\frac{\Lambda(\mathcal{G}\setminus p)-\Lambda(\mathcal{G})}{\Lambda(\mathcal{G})} \tag{10}$$

$$\mathcal{F}_{enter} = \frac{n-r-q}{q-1}\frac{\Lambda(\mathcal{G})-\Lambda(\mathcal{G}|p)}{\Lambda(\mathcal{G}|p)} \tag{11}$$

Wilks' Λ:

$$\Lambda(\mathcal{G}) = \frac{\det(W)}{\det(T)} \tag{12}$$

Within-group covariance matrix:

$$W(\mathcal{G}) = \sum_{m=1}^{q}\sum_{m=1}^{n_m}(x_{mki}-x_{mi.})(x_{mkj}-x_{mj.}) \tag{13}$$

Among-group covariance matrix:

$$T(\mathcal{G}) = \sum_{m=1}^{q}\sum_{m=1}^{n_m}(x_{mki}-x_{i..})(x_{mkj}-x_{j..}) \tag{14}$$

where q is the number of clusters, n_m is the number of samples in the cluster m, x_{mki} is the value of the ith probe for the kth sample in the mth cluster, n is the total number of samples, r is the number of probes currently included in $\mathcal{G}$, $\mathcal{G}|p$ denotes a new group of probes which is obtained by adding the probe p to $\mathcal{G}$, $\mathcal{G}\setminus p$

denotes a new group of probes which is obtained by removing the probe p from $\mathcal{G}$.

At the conclusion of SDA, the optimal probe set is determined based on the prediction accuracy of the selected probes. Because there are typically a small number of samples associated with microarray data, prediction accuracy is computed using the leave-one-out (LOO) cross validation method [10, 15]. The set of probes associated with the highest LOO predication accuracy are written to a file together with their associated weights. It is important to note that when SDA is used to obtain the final probe set, the weights associated with the probes must be used for classification of new empirical data obtained using the probes. The probes should not be treated with equal weight.

2.3.4 Probe selection for a classification microarray

In summary, the steps in our design of an optimal probe set are:

1. Cluster the microarray or sequence data set and select clusters of interest using a hierarchical, non-hierarchical, and/or model-based method. *A priori* clustering is also permitted.

2. Use the probe ranking procedure to rank the probes for relevance.

3. Repeat K-means clustering for probe reduction until satisfied:

    (a) Select j top-ranked probes.

    (b) Repeat for κ in a chosen range:

        i. Cluster the j top-ranked probes into κ clusters.

        ii. Choose κ representative probes, one from each cluster.

        iii. Use SDA to find a set of probes from the κ representative probes and compute the LOO prediction accuracy.

4. Save the set of probes associated with the highest LOO prediction accuracy together with their weights. After constructing the optimized microarray, a set of independent control samples should be hybridized to empirically assess the accuracy of the microarray results.

A flowchart of the process is shown in Fig. 3. It should be pointed out that the optimal number of probes computed by this process does not take into account the effects of noise and other random experimental effects. The sample-to-feature (SFR) ratio gives the minimum number of probes that should be used to create a microarray. The rule of thumb is given by [17]:

$$SFR = \frac{number\ of\ samples}{number\ of\ features} \leq \frac{1}{5}.$$

In this chapter we refer to features as probes. The SFR should be used in conjunction with the results to choose the optimal probe set.

2.4 System Overview

Our software tool PLASMID, which stands for PLasmid Analysis System and Marker IDentification, is implemented as a Java application. The NetBeans platform was chosen for development because addition of new features is easily implemented. Also, many of the tasks common to desktop applications are provided by NetBeans. These include user interface management (e.g., menus and toolbars), user settings management, storage management (saving and loading any kind of data), window management, and wizard framework (supporting step-by-step dialogs). Each function is implemented as a NetBeans module and can be installed or removed easily. Java is a platform-independent programming language, so although PLASMID has been developed using the Windows operating system, it will be relatively easy to adapt it to other operating systems. We intend to extend PLASMID for use on both Apple and Linux computers. In addition to using Java, PLASMID uses code written using the C programming language. C is needed for computationally intensive tasks that require greater speed and efficiency. However, the use of two different programming languages is transparent to the user.

PLASMID provides an integrated environment for designing an optimal classification microarray. As such, PLASMID v0.9 includes the following services:

1. Loading and management of different kinds of input data, including plasmid sequence data, hybridization data, virtual hybridization data, and probe sequences. Data may be in tab-delimited or comma-delimited text format or in Microsoft Excel spreadsheet format.

2. Different methods for processing hybridization data. The tool provides several data preprocessing methods, including normalization and noise filtering. It also provides hierarchical, non-hierarchical, and model-based methods for clustering samples; two different statistical tests for ranking probes; use of K-means clustering for reduction of probe redundancy; and stepwise discriminant analysis with assignment of weights to probes.

3. Design of mixed arrays using existing hybridization data or virtual hybridization data. An optimal set of probes is identified, and weights associated with each probe are stored for analysis of experimental results.

4. Construction of virtual microarrays to obtain virtual hybridization data using genomes from the National Center for Biotechnology Information (NCBI) database. Genomes for probes can be chosen by accession number or by gene sequence.

5. Visualization of microarray data and data processing results, including dendrograms, heat maps, and scatter plots. Plots can be saved in different image formats.

6. Automatic probe design after the user has specified the parameters. A step-by-step wizard guides the user through the various steps.

Experimental data obtained from microarrays designed using PLASMID can be used as input data and analyzed using the weighted classification function obtained in step 3.

2.5 Implementation

Implementation of the tool is based on the NetBeans system, using the Java and C programming languages, where each function is implemented as a module. A new function can easily be added without affecting existing functions. The program was written for the Windows operating system, but in the future it will be implemented for both the Linux and Mac OS X operating systems.

2.6 Results and Discussion

In this section we present results obtained using PLASMID to analyze a mixed-plasmid microarray data set [4] and a simulated mixed-genome microarray data set [11]. We also present results for publically-available leukemia expression array data [18]. For this latter data set, clusters (i.e., types of leukemia) are pre-assigned so only probe ranking, reduction of probe redundancy, and stepwise discriminant analysis (SDA) are used to determine the optimal probe set. PLASMID's performance in probe selection is evaluated using the leave-one-out (LOO) approach for which one sample is excluded and the remaining samples are used to obtain the discriminant functions. Each sample is, in turn, excluded and a corresponding set of discriminant functions is used to classify it. The prediction accuracy, the percentage of times a set of discriminant functions correctly classifies the withheld sample, is used as the performance metric.

Mixed-plasmid microarray data. A mixed-plasmid microarray has been used to compare the genetic composition of plasmids [4]. The microarray consists of 576 probes composed of randomly selected fragments of plasmid DNA, and the samples consist of data from hybridization experiments with 43 plasmids. The sample data are composed of hybridization signal intensities for each microarray probe.

First we used the Ward's minimum variance hierarchical clustering algorithm to create a dendrogram. To test the two-class problem, we divided the dendrogram into two clusters. One cluster consisted of 15 plasmids which, with one exception (the *peSSuTet* plasmid), have the $bla_{\mathrm{CMY-2}}$ antibiotic resistance gene; the other cluster consisted of 28 plasmids. We then used the probe ranking function, choosing the ANOVA-F test statistic, and generated a scatter plot (Fig.1). The scatter plot shows that the majority of the probes have statistical values close to zero and, thus, that ANOVA-F test statistics can be used to distinguish between informative ($F > 0$) and noninformative ($F \approx 0$) probes. This result also serves to highlight the need for optimization algorithms, as the majority of probes provide limited discrimination.

For the two-cluster case, we chose 1, 20, and 200 top-ranked probes for comparison. Using reduction of probe redundancy and SDA, we found that one probe (5-E3, a transposase gene associated with the $bla_{\mathrm{CMY-2}}$ element [19]) correctly classified all but two of the plasmids into the two original dendrogram clusters [4]. Interestingly, in the original study one of these two plasmids (*pe1171sT*) was grouped with plasmids that harbor the $bla_{\mathrm{CMY-2}}$ gene even though it does not carry this gene. The present analysis separated *pe1171sT* from the $bla_{\mathrm{CMY-2}}$ plasmids. In addition, a different plasmid (*pe7594T*) that harbors the $bla_{\mathrm{CMY-2}}$ gene was rejoined with other $bla_{\mathrm{CMY-2}}$ positive plasmids in the current analysis. Thus, analysis using PLASMID more accurately reflects the phenotypic properties of the plasmids included in the study. The one exception was the *peSSuT* plasmid that was consistently grouped with $bla_{\mathrm{CMY-2}}$ plasmids while not harboring this gene [7].

Next we divided the original dendrogram into five clusters and ranked probes as before. As expected, the number of probe clusters κ specified for the reduction of probe redundancy affects the prediction accuracy (Table 2). Small values of κ certainly reduce redundancy, but they also reduce specificity. The optimal set of probes is identified using SDA with the LOO method to determine the highest prediction accuracy. In this case, the smallest number of probes from the topmost ranked probes with the highest prediction accuracy is 10. Thus, PLASMID analysis reduced the original data set of 576 probes to 10 probes that are needed to accurately assign plasmids to one of five clusters. Note that both fewer and greater numbers of probes can be used to achieve the same prediction accuracy, and the choice of the number of probes to use must be made by the user. Non-hierarchical clustering followed by probe ranking, probe reduction, and SDA gave similar results (data not shown).

In addition to hierarchical and non-hierarchical clustering methods, we can obtain classification results using our model-based method, which is based on a genetic algorithm. The genetic algorithm predicted that the most likely number of sample clusters is five (Table 3). Comparison of Tables 2 and 3 shows that prediction accuracies depend on the initial clustering method used. For this case, the prediction accuracies for the model-based clustering method are larger for a given number of probe clusters than those obtained via the hierarchical method. Furthermore, the variance in prediction accuracies is lower as a function of the number of top-ranked probes when clusters are initially assigned using the model-based method. However, for other data sets another clustering model might give the best results.

Based on the sample to feature ratio (SFR), at least nine probes (features) are required for classifying 43 plasmids (samples). Table 3 shows several choices for ten probes with equivalent performance. When additional information is available, it should be used to assist with the choice of a set.

Virtual Streptococcus mixed-genome microarray data. A virtual *Streptococcus* mixed-genome microarray was constructed by Wan *et al.* [11]. To create the equally-represented, 4000-probe virtual array, 800 gene segments each 600-bp long were randomly selected from genomes of fifteen strains of five bacterial species—that is, each species was represented by 800 different probes. Virtual hybridization was accomplished using BLAST scores as proxies for array probe intensities, and PLASMID was used to analyze the data. In the initial analysis one species was excluded from the study because it was represented by only a single strain (*S. mutans* UA159). Because we knew *a priori* that the samples belonged to four different species, the goal was to find an optimal set of probes to differentiate these four. ANOVA-F tests were used to rank the 4000 probes, and LOO analysis was performed on different numbers of the highest ranked probes. In fact, we found the LOO prediction accuracy to be 100% for differentiating the four different species using only the single top-ranked probe. Use of additional probes had no effect on the results. While it appears that successful classification can be achieved with a single probe when classification relies on differences in hybridization signal, given inherent sources of variation in microarray hybridization data, it would be prudent to include additional probes to increase classification confidence using empirical data. For example, the minimum recommended probe set in this case would be 3 (SFR).

In the second analysis, our model-based clustering method identified two clusters, one with the two *S. pneumoniae* strains and the other with the remaining 13 strains. After probe ranking, reduction of probe redundancy, and SDA, it was found that one probe could be used to differentiate these two groups. We also used non-hierarchical clustering of the samples followed by probe ranking, probe reduction, and SDA. When the number of clusters was chosen to be $k = 2$, the result was identical to the result obtained using our model-based cluster method. When the number of clusters was chosen to be $k = 3$ or $k = 4$, the two *S. pneumoniae* genomes were placed in different clusters. A dendrogram constructed using the neighbor joining method shows a clear distinction between the two *S. pneumoniae* samples and the remaining bacteria (Fig.2). When these two samples are excluded, PLASMID groups the remaining thirteen samples correctly into four species clusters. The results shown in Table 4 are obtained using non-hierarchical clustering, probe ranking, probe reduction, and SDA. As this table illustrates, only two probes are needed to obtain 100% prediction accuracy by species. These two probes are from the genomes of *S. pneumoniae* TIGR4 and either *S. pyogenes* M1 GAS or *S. pyogenes* MGAS5005. Based on the SFR rule of thumb, at least three probes are needed. Several choices exist that suffice for this condition.

Public ALL/AML leukemia data. The ALL/AML leukemia data set, obtained from expression arrays, has been widely used in the literature. It consists of two classes of leukemia, acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML), and there are 72 samples (47 ALL and 25 AML) and 7129 probes. Table 5 shows prediction accuracy results after probe ranking, probe redundancy reduction, and SDA have been performed. When the top 50 probes were selected, the highest accuracy was achieved when probes were clustered into 10 groups. A set of 10 probes was identified with a prediction accuracy of 97.22%. Using additional probes does not lead to improvement. According to the SFR rule of thumb, at least 20 probes should be used in the actual microarray design; several choices of 20 probes exist and all produce robust prediction results (Table 5).

2.7 Conclusions

In this chapter we describe a new software tool, PLASMID, for selecting an optimal set of probes for the design of a classification microarray. The tool provides the user with several clustering methods, a probe ranking method, probe redundancy reduction, and probe selection using stepwise discriminant analysis. Images can be saved in several different formats, and weights generated using SDA can be stored for use in analysis of experimental data. In addition, PLASMID can be used to construct virtual microarrays with genomes from public databases. The software package has been applied to data from a mixed-plasmid microarray, a virtual mixed-genome microarray, and an expression microarray. Robust results have been obtained for all three sets of data.

Although many methods are available for determining a set of features for a given microarray data set, these methods require the classification information to be known in advance. PLASMID was designed to be used prior to implementation of a microarray when no such information is available, although the program can also be used when clusters are known *a priori.*

PLASMID can be obtained by following the link from http://www.vetmed.wsu.edu/research_vmp/microArrayLab/.

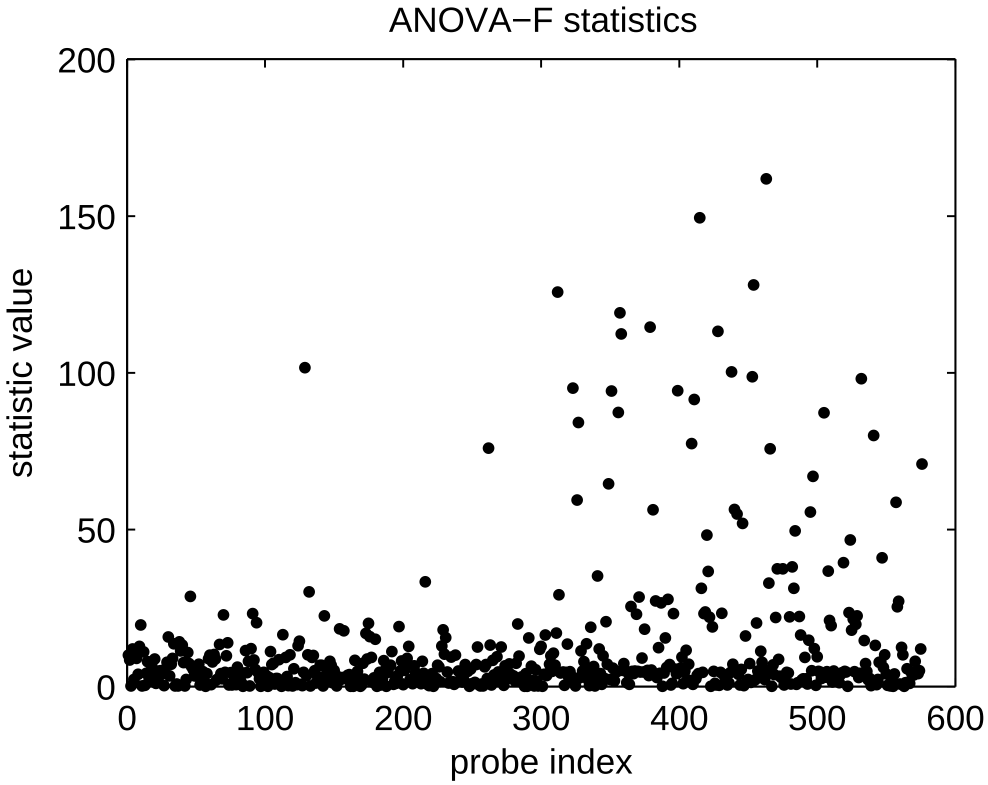


Figure 1. Scatter plot of ANOVA-F test statistics for the mixed-plasmid microarray probes: The scatter plot shows that the majority of the probes have statistical values close to zero and, thus, that ANOVA-F test statistics can be used to distinguish between informative ($F > 0$) and noninformative ($F \approx 0$) probes.

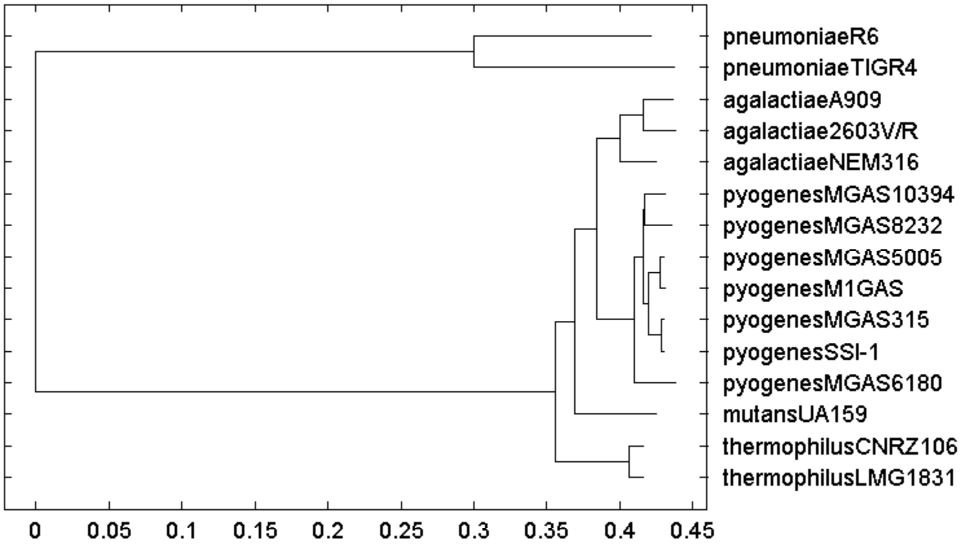


Figure 2. Dendrogram for *Streptococcus* MGM data: The dendrogram constructed using the neighbor joining method shows a clear distinction between the two *S. pneumoniae* samples and the remaining bacteria.

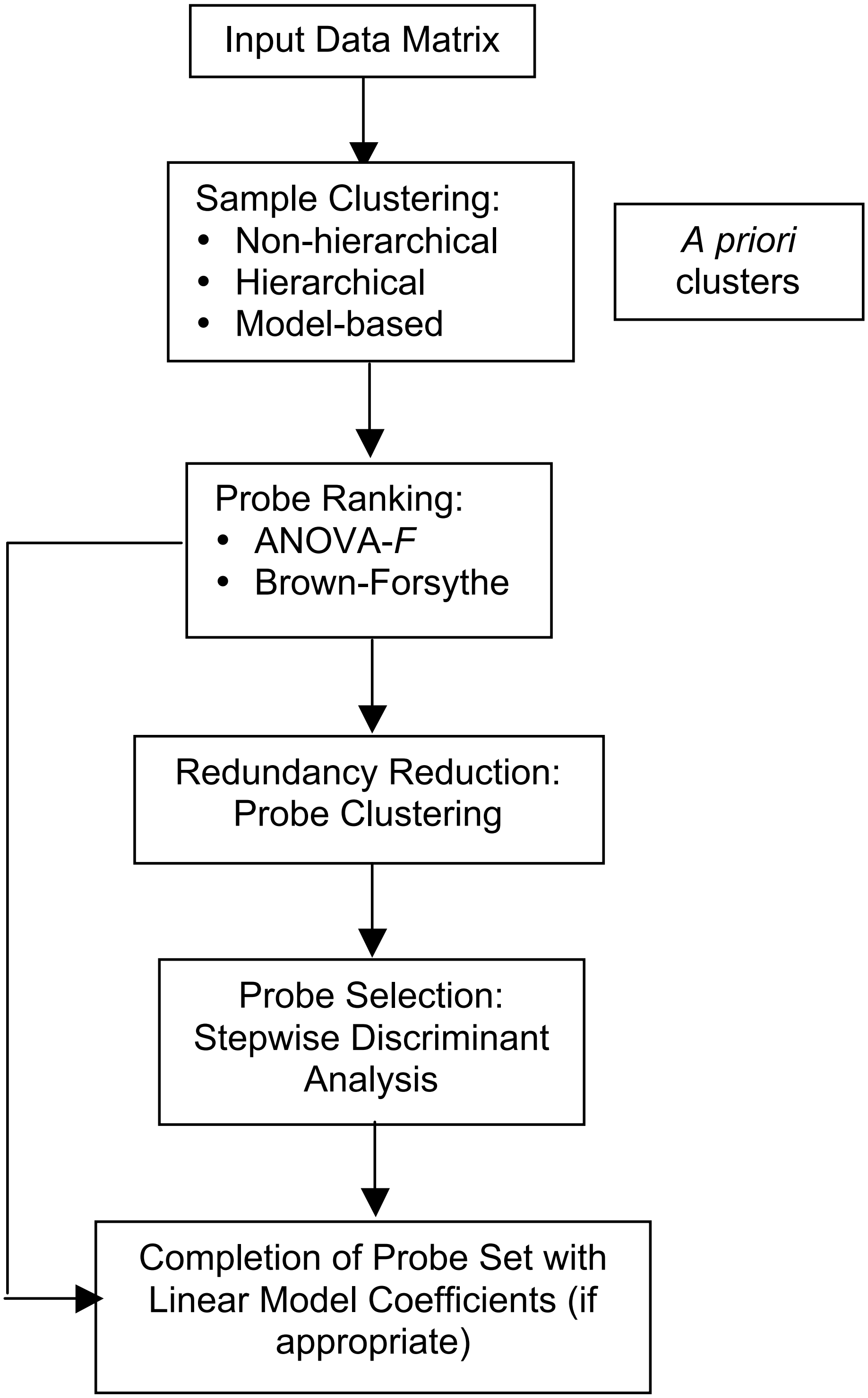

Input Data Matrix
Sample Clustering:
• Non-hierarchical
• Hierarchical
• Model-based
A priori clusters
Probe Ranking:
• ANOVA-*F*
• Brown-Forsythe
Redundancy Reduction: Probe Clustering
Probe Selection: Stepwise Discriminant Analysis
Completion of Probe Set with Linear Model Coefficients (if appropriate)

Figure 3. Flowchart of PLASMID

Table 2. Classification accuracy of mixed-plasmid data using hierarchical clustering with five sample clusters. PA is the prediction accuracy.

	Number of clusters of probes, κ											
	2		5		10		20		30		40	
Number of top-ranked probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes
100	72.09	2	72.09	5	72.09	10	69.77	19	69.77	29	69.77	32
150	86.05	2	93.02	5	95.35	10	95.35	20	95.35	29	95.35	36
200	74.42	2	90.70	5	93.02	10	93.02	20	93.02	30	95.35	35
250	76.74	2	95.35	5	95.35	10	95.35	20	95.35	30	90.70	34
300	46.51	2	88.37	5	93.02	10	93.02	20	93.02	30	90.70	35
350	76.74	2	93.02	5	93.02	10	95.35	20	95.35	30	95.35	33
400	69.77	2	93.02	5	90.70	10	93.02	20	90.70	30	93.02	35

Table 3. Classification accuracy of mixed-plasmid data with model-based clustering. PA is the prediction accuracy.

	Number of clusters of probes, κ											
	2		5		10		20		30		40	
Number of top-ranked probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes
100	83.72	2	95.35	5	95.35	10	95.35	19	95.35	36	95.35	33
150	53.49	2	90.70	5	93.02	10	93.02	20	93.02	36	93.02	28
200	79.07	2	93.02	5	93.02	10	93.02	20	93.02	36	93.02	35
250	76.74	2	95.35	5	95.35	10	95.35	20	95.35	35	93.02	32
300	69.77	2	93.02	5	93.02	10	95.35	20	95.35	34	95.35	35
350	67.44	2	93.02	5	93.02	10	93.02	20	93.02	35	93.02	35
400	69.77	2	93.02	5	93.02	10	93.02	20	93.02	37	95.35	35

Table 4. Classification accuracy using mixed-genome array data with non-hierarchical clustering for four sample (species) clusters. PA is the prediction accuracy.

	Number of clusters of probes, κ											
	2		5		10		20		30		40	
Number of top-ranked probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes
50	100	2	100	5	100	7	100	7	100	1	75	1
100	100	2	100	5	100	7	100	7	100	7	100	1

Table 5. Classification accuracy using ALL/AML leukemia data. PA is the prediction accuracy.

	Number of clusters of probes, κ											
	2		5		10		20		30		40	
Number of top-ranked probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	PA (%)	No. of probes	Pa (%)	No. of probes	PA (%)	No. of probes
50	94.44	2	94.44	5	97.22	10	97.22	20	97.22	30	97.22	39
100	88.89	2	95.83	5	94.44	10	97.22	20	97.22	30	97.22	40
150	83.33	2	95.83	5	97.22	10	97.22	20	97.22	30	97.22	39
200	79.17	2	80.56	5	97.22	10	97.22	20	97.22	30	97.22	39
250	79.17	2	79.17	5	97.22	10	97.22	20	97.22	30	97.22	39

List of References

[1] M. K. Borucki, M. J. Krug, W. T. Muraoka, and D. R. Call, "Discrimination among *Listeria monocytogenes* isolates using a mixed genome DNA microarray," *Vet Microbiol*, vol. 92, no. 4, pp. 351–362, 2003.

[2] D. R. Call, M. K. Borucki, and T. E. Besser, "Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of *Listeria monocytogenes*," *J Clin Microbiol*, vol. 41, no. 2, pp. 632–639, 2003.

[3] M. Soule, E. Kuhn, F. Loge, J. Gay, and D. Call, "Using DNA microarrays to identify library-independent markers for bacterial source tracking," *Appl. Environ. Microbiol.*, vol. 72, pp. 1843–1851, 2006.

[4] D. Call, M.-S. Kang, J. Daniels, and T. Besser, "Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray," *J Appl Microbiol*, vol. 100, no. 1, pp. 15–28, 2006.

[5] C.-C. Chou, C.-H. Chen, T.-T. Lee, and K. Peck, "Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression," *Nucleic Acids Res*, vol. 32, no. 12, pp. e99–, 2004.

[6] H. Hyyro, M. Juhola, and M. Vihinen, "Genome-wide selection of unique and valid oligonucleotides," *Nucleic Acids Res*, vol. 33, no. 13, pp. e115–, 2005.

[7] S. J. Emrich, M. Lowe, and A. L. Delcher, "PROBEmer: a web-based software tool for selecting optimal DNA oligos," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3746–3750, 2003.

[8] J. Jaeger, R. Sengupta, and W. Ruzzo, "Improved gene selection for classification of microarrays," *Pac. Symp. Biocomput.*, pp. 53–64, 2003.

[9] B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clément, and J.-D. Zucker, "Improving classification of microarray data using prototype-based feature selection," *SIGKDD Explor. Newsl.*, vol. 5, no. 2, pp. 23–30, 2003.

[10] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, no. 8, pp. 1530–1537, 2005.

[11] Y. Wan, S. L. Broschat, and D. R. Call, "Validation of mixed-genome microarrays as a method for genetic discrimination," *Appl. Environ. Microbiol.*, vol. 73, no. 5, pp. 1425–1432, 2007.

[12] A. Jagota, *Microarray Data Analysis and Visualization.* Bioinformatics By The Bay Press, 2001.

[13] Z. S. Qin, "Clustering microarray gene expression data using weighted chinese restaurant process," *Bioinformatics*, vol. 22, no. 16, pp. 1988–1997, 2006.

[14] Y. Su, T. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, "RankGene: identification of diagnostic genes based on expression data," *Bioinformatics*, vol. 19, no. 12, pp. 1578–1579, 2003.

[15] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by test statistics," *J Biomed Biotechnol.*, vol. 2, pp. 132–138, 2005.

[16] R. I. Jennrich, "Stepwise discriminant analysis," in *Statistical methods for digital computers*, K. Enslein, Ed. John Wiley & Sons Inc, 1977, vol. III, pp. 76–95.

[17] R. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

[18] T. R. Golub, *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[19] M.-S. Kang, T. E. Besser, and D. R. Call, "Variability in the region downstream of the $bla_{\mathrm{CMY-2}}$ β–lactamase gene in *Escherichia coli* and *Salmonella enterica* plasmids," *Antimicrob. Agents Chemother.*, vol. 50, no. 4, pp. 1590–1593, 2006.

CHAPTER 3

A Fusion Algorithm for Assessing Intra-specific Genetic Relationships between Bacterial Pathogens

3.1 Abstract

Determining phylogenetic relationships between bacterial strains is important in molecular epidemiology studies. Two molecular typing methods, pulsed-field gel electrophoresis (PFGE) and multiple-locus variable-number tandem repeat analysis (MLVA), are widely used in such studies. In this work, we propose a fusion algorithm that combines the information obtained from both PFGE and MLVA assays to obtain phylogenetic relationships. Two sets of *Salmonella enterica* are examined; one set includes serovar Typhimurium isolates from a wide range of sampling dates, locations, and host species while the other set includes a group of serovar Newport isolates collected over a limited geographic and temporal scale. Results are assessed by comparison with phage-typing assays and with known epidemiological relationships. The analysis shows that the fusion algorithm provides an improved ability to discriminate between isolates and to infer phylogenetic relationships compared with using either the PFGE or MLVA method alone.

3.2 Introduction

Salmonellosis is one of the most common food-borne diseases in the United States [1]. Consequently, it is important to understand how *Salmonella* strains disseminate within and between reservoirs and environments. For this purpose many molecular typing tools have been used to elucidate the genetic relationships between strains [2]. Of these methods pulse-field gel electrophoresis (PFGE) is considered the gold standard for strain typing, and multiple-locus variable-number tandem repeat analysis (MLVA) assays are powerful alternative or complementary typing tools. Both methods offer a high degree of resolution for strain typing depending on several factors.

PFGE is one of the most reproducible and highly discriminatory typing techniques, and it has been widely and successfully used for typing a variety of *Salmonella enterica* serovars [3, 4]. One of the primary advantages of PFGE is that protocols are relatively simple to standardize, results are robust, and for many situations the technique is capable of discriminating between closely related strains. In addition, extension of the assay to new serovars does not require a great deal of modification as might be required with procedures that are dependent on polymerase-chain reaction (PCR). PFGE involves size separating chromosomal DNA macro-restriction fragments in agarose gels, and strains are typed depending on the resulting band pattern observed after electrophoresis has been completed. Difficulties arise when strains are very closely related (i.e., poor discrimination; [5, 6]) or when bands either co-migrate in the gel or when identically-sized bands represent completely different fragments of chromosomal DNA [7]. These complications are more pronounced when a large number of bands are generated by the restriction digest. In addition, while band patterns convey a crude degree of genetic relatedness, a large number of independent restriction digests would be needed to infer an accurate phylogeny between isolates [7]. A final and somewhat unavoidable criticism of PFGE is that the procedure is time-consuming (requiring days), and thus it is not conducive to high-throughput screening.

Multiple-locus variable-number tandem-repeat analysis (MLVA) is a PCR-based technique that relies on amplification of chromosomal or plasmid DNA that encompasses short tandem repeats of a DNA sequence. The tandem repeats are prone to higher than background mutation rates due to DNA-strand slippage during DNA replication [8], and thus the amplified fragments will vary in length depending on the number of repeats harbored at a given locus. Different fragment lengths are tallied either as the total length (base pairs) or the estimated number of repeat units, and each discretely-sized fragment is considered a unique allele for the locus under investigation. Because the technique can be multiplexed and automated, it is conducive to rapid and relatively high-throughput strain typing. MLVA assays are relatively robust [5, 9–11], and while not perfect, these assays can provide phylogenetic information even with a limited number of loci [12, 13]. While access to a sequenced genome dramatically speeds the ability to establish new assays [14], this is not a requisite to assay development. The primary limitations of the technique include the potential need for a new set of loci for every species or serovar under investigation and the fact that some loci are very unstable and can disappear from some strains or lineages (this produces the equivalent of an uninformative null allele). Mutation rates can also vary between loci [15, 16], which, if ignored, can introduce bias into phylogenetic analyses.

The objective of this study was to determine if we could combine the information that is obtained from both PFGE and MLVA assays to produce more rigorous and discriminatory analyses of *Salmonella* isolates. Two sets of *Salmonella* isolates were used in this study; one set included serovar Typhimurium isolates from a wide range of sampling dates, locations, and host species while the other set included a group of serovar Newport isolates collected over a limited geographic and temporal scale. The results of different typing methods were assessed by comparison with phage-typing assays and with known epidemiological relationships. To interpret MLVA data we employed a metric that incorporates a stepwise-mutation model, and to interpret the PFGE data we employed Dice coefficients to construct a distance matrix. Our analysis shows that the fusion of the two typing methods provides an improved ability to discriminate between isolates and to infer phylogenetic relationships compared with using either method alone. We provide a PC-based, stand-alone software package to assist practitioners with this type of analysis (available at http://www.vetmed.wsu.edu/research_vmp/MicroArrayLab/).

3.3 Materials and Methods

***Salmonella* strains**. Two sets of isolates were used for this study. Set A included 44 *S. enterica* serovar Typhimurium strains that were previously collected from eight types of animal hosts and from different locations and different time periods (see figures for strain designations and descriptors). Because these isolates were epidemiologically unrelated, we assumed that they encompassed a high degree of genetic variability. Set B included 69 *S. enterica* serovar Newport isolates, mostly collected from Washington State cattle in 2006, and this set was assumed to represent less genetic diversity. *Salmonella* serovar Typhimurium strains were phage typed at the National Microbiology Laboratory, Canadian Science Center for Human and Animal Health, Winnipeg, Manitoba. Serovar Newport isolates were tested for antibiotic resistance using a disc diffusion method [17] according to Clinical and Laboratory Standards Institute guidelines [18, 19]. Northwestern bovine-origin isolates were tested for susceptibility to a panel of antimicrobials that included ampicillin (10 μg), chloramphenicol (30 μg), gentamicin (10 μg), kanamycin (30 μg), streptomycin (10 μg), tetracycline (30 μg), triple-sulfa (a combination of sulfadiazine, sulfamethazine, and sulfamerazine) (250 μg), trimethoprim-sulfamethoxizole (1.25 μg-23.75 μg), ceftazidime (30 μg), amoxicillin-clavulanic acid (20/10 μg), and nalidixic acid (30 μg) (BD Diagnostics, Sparks, Maryland, USA). The northeastern isolates were tested with the same panel, except that a sulfisoxazole disc (250 μg) was substituted for the triple-sulfa disc.

Pulse-field gel electrophoresis (PFGE). We followed a standard PFGE protocol for *Salmonella enterica* using an XbaI restriction digest [20]. Briefly, genomic DNA was digested in agarose plugs with the restriction enzyme, XbaI, and resulting DNA fragments were gel-separated using a CHEF DR II (BioRad, Hercules, CA) apparatus. Electrophoresis conditions included an initial pulse time of 2.2 s, final pulse time of 63.8 s, running temperature of 14°C, and a run time of 18-20 hours at 6 V/cm. PFGE gels were stained with ethidium bromide and visualized on a UV transilluminator. Gel images were analyzed using Bionumerics version 4.6 (Applied Maths, Sint-Martens-Latem, Belgium). Estimated band sizes were exported from Bionumerics for the current study.

Multiple-locus variable-number tandem repeat analysis (MLVA). For *S. enterica* serovar Typhimurium isolates, four of five previously described variable-number tandem repeat (VNTR) loci were employed (STTR5, STTR6, STTR9, and STTR10pl) following a published protocol [12]. For the *S. enterica* serovar Newport strains, two Typhimurium loci (STTR5, and STTR6) [12] and four published Newport-specific loci were employed (NewportA, NewportB, NewportM, and NewportL) [21]. The PCR reactions for MLVA were completed in two separate reactions, PCR1 and PCR2, to avoid overlap in combinations of fragment size and dye color in the same reaction. Primers, including forward fluorophore-conjugated primers, were purchased from Applied Biosystems (Foster City, CA). For each isolate, a single colony was suspended in dH2O (100 μL) and boiled for 20 min followed by centrifugation at 14,000 rpm for 5 min. The boiled lysate suspension (100 μL) was used for template in the PCR reactions. Two separate triplex reactions (Table 6) were run using an iCycler thermal cycler (BioRad, Hercules, CA) in 25 L volumes. Cycling conditions for both reactions included an initial denaturation at 94°C for 15 min followed by 25 cycles of 94°C for 30 s, 55°C for 1 min, and 72°C for 1.5 min with a final extension step at 72°C for 10 min. Size standard

(LIZ600, Applied Biosystems) (0.125 μL) and Hi Di formamide (19.375 μL) were added to PCR products (0.5 μL) (total volume 20 μL for capillary electrophoresis). Capillary electrophoresis was carried out at the Washington State University Genomics Core using a 3730 DNA Analyzer with Pop-7 polymer (Applied Biosystems). The resulting electropherograms were analyzed using GeneMarker software (Softgenetics LLC, State College, PA, USA).

Data Analysis. Dice similarity coefficients were calculated using Bionumerics (Applied Maths, Sint-Martens-Latem, Belgium) from PFGE data to generate the distance matrix and the unweighted pair group method with arithmetic mean (UPGMA) algorithm was used to construct a dendrogram. For the MLVA data we divided the total length of the tandem repeats by the estimated size of each repeat to obtain the number of tandem repeats for each locus and each strain. There were five loci with tandem repeats for *S. enterica* serovar Typhimurium, but data from one locus were not used because they were from a plasmid locus and only a subset of the 44 bacteria isolates were positive for this locus. Initially, data from the remaining four loci were used, but there was very little variability among isolates for two of the loci, and excluding these data did not alter the results. For *S. enterica* serovar Newport there were six loci, all of which were used. Because passage experiments indicate that VNTR mutations are usually composed of a single step [9, 15], we modeled our data using a single-step stepwise-mutation model (SMM). Based on this statistical model, we estimated the distance S (the total number of single steps) between two lineages (two different isolates) and their most recent common ancestor (MRCA) using the number of tandem repeats, X_L, of the two lineages. The distances S for all lineage pairs were then used to construct the distance matrix and UPGMA was used to obtain a dendrogram.

If μ is the rate of stepwise mutations per generation and if we assume the gain or loss of a repeat is equally probable, then the following conditional probabilities P characterize the single-step SMM:

$$\begin{aligned} P(X_{t+1} = i+1 | X_t = i) = P(X_{t+1} = i-1 | X_t = i) &= \frac{\mu}{2} \\ P(X_{t+1} = i | X_t = i) &= 1 - \mu \\ P(\| X_{t+1} - X_t \| \leq 2 | X_t = i) &= 0 \end{aligned} \tag{15}$$

where i denotes the number of tandem repeats at distance t, where t is an integer number between zero and infinity. Based on these conditional probabilities, the probability of the distance t is given by [22]:

$$P(t | n_0, \ldots, n_k) = N(t)/D(t) \tag{16}$$

where

$$\begin{aligned} N(t) &= e^{(\lambda + 2\mu n)t} \prod_{j=0}^{k} [I_j(2\mu t)]^{n_j} \\ D(t) &= \int_0^{\inf} e^{(\lambda + 2\mu n)t} \prod_{j=0}^{k} [I_j(2\mu t)]^{n_j} dt \end{aligned} \tag{17}$$

The equation for P assumes that the mutation rate μ is constant for all loci which is approximately true for our *S. enterica* serovar Typhimurium data; n_m denotes the locus number where the subscript $m = 0, 1, 2, \ldots, k$ is the difference between the number of tandem repeats for two lineages and $m = k$ is the maximum number of differences; n is the number of loci used; λ is a parameter associated with the distance to a MRCA (in this work $\lambda = 0.0002$ was found to give satisfactory results [22]); I_j is the jth order modified Bessel function of type 1. The distance S is the value of t with the maximum probability.

Fusion algorithm. The dendrograms constructed using the PFGE-only or

MVLA-only data differ substantially for both *S. enterica* serovar Typhimurium and serovar Newport. Consequently, if we assume that both sets of data contain useful information as well as error, it is possible that better results can be obtained by combining the data. In fact, it is known that if two different algorithms used with the same data give different results, if the error for each is less than the error associated with randomly generated results, and if the error for both is uncorrelated, a combination of these algorithms will give better results than either of the two algorithms alone [23]. For our problem we have two different sets of data but they are for the same sets of samples. We can safely assume that the error for both the MLVA and PFGE algorithms is less than the error for random clustering because both PFGE and MLVA can recapitulate epidemiological relationships [21, 24]. Furthermore, we can assume that the error is uncorrelated because the PFGE and MVLA assays measure different types of genetic differences. Consequently, because the two methods provide different results, it is likely that combining the PFGE and MLVA data will provide a more comprehensive picture of the underlying population genetics of these strains.

Several strategies can be used to combine different types of data. One strategy is to treat each data type independently and produce two independent dendrograms that are then combined to form a single dendrogram. While conceptually simple, this approach weights all sources of information equally, which may not be an optimal approach. Another strategy is to combine the data sets together before generating a dendrogram, but it may be difficult to combine the data if they are different types (e.g., discrete and continuous), and even if this is accomplished, a suitable approach for evaluating the combined data may not exist.

An alternative approach is to process each type of data using an algorithm that is appropriate for that data type and then combine the results at some midpoint in the process; this is the approach we employed. We begin with two distance matrices, one for the PFGE data set and one for the MLVA data set as described previously. One distance matrix is used to construct a dendrogram, and a threshold is selected (see below) to define distinct clusters from this dendrogram. If two strains in the second distance matrix are grouped together in one of the clusters formed from the first distance matrix, the distance between them is reduced (see below); if these two strains are not grouped together, then the distance in the second matrix is left unchanged. After all pairwise distance values are adjusted based on the clusters from the first matrix, a final dendrogram is generated. Both sets of data can be used in alternating roles: PFGE data are used to create the clusters while MLVA data are used to create the distance matrix to be modified, and MLVA data are used to create the clusters while PFGE data are used to create the distance matrix to be modified.

For the fusion algorithm described above, values for two parameters must be chosen. The first is the threshold value, thr, which divides one dendrogram into distinct clusters. The second is the degree that each distance value should be reduced in the second distance matrix. If D is the distance matrix to be modified with elements $d(i, j)$ and D^* is the modified distance matrix, the elements of D^* are given in terms of $d(i, j)$ by:

$d^*(i, j) = d(i, j)$ if bacteria samples i and j are not in the same cluster

$d^*(i, j) = r\ d(i, j)$ if bacteria samples i and j are in the same cluster

with $0 < r \leq 1$. Thus, the weight parameter r dictates how much the distance value will be reduced. The choice of these two parameters, thr and r, changes the results, and there is no obvious way of knowing what values are the best to use. Our approach was to use the entire range of values for the threshold thr, i.e., 0.05-1 and several ranges of values for the weight parameter r. For the former, this range corresponds to having each strain form its own cluster ($thr = 0.05$) and having all strains clustered into a single group ($thr = 1$). For each set of ranges we created a dendrogram using UPGMA and the modified distance matrix; from the resulting set of dendrograms we constructed a generalized tree using Consense from the software package Phylip [25] with default parameters. In addition to the question of the optimum parameter values is the question of whether results for the generalized tree would be more accurate with the set of dendrograms obtained using PFGE data to form the initial clusters (referred to as PFGE clusters), using the set of dendrograms obtained using the MLVA data to form the initial clusters (referred to as MLVA clusters), or using the combined sets of dendrograms (referred to as All-clusters, i.e., data from both PFGE clusters and MLVA clusters). When both PFGE and MLVA data are combined to obtain the All-clusters data, the PFGE and MLVA data will create a conflict if they disagree completely on the relationship between two strains. For example, the PFGE data may indicate that strains A and B always occur as a pair while the MLVA data may indicate that strains A and C always occur as a pair. If this happens, Consense [25] will construct a tree that depends on the order of the input. To prevent such an occurrence, we break the tie by multiplying the values of one set of data by 0.501 and the other set by 0.499.

3.4 Results and Discussion

Comparison of genetically diverse strains of *S. enterica* serovar Typhimurium. Generalized trees were constructed using the fusion, MLVA, and PFGE algorithms described above. For the fusion algorithm, sets of dendrograms were generated for the weight parameter r ranging between 0 and 1 at increments of 0.1 and for the threshold thr ranging between 0.05 and 1 at increments of 0.05. This produced a set of 220 dendrograms using PFGE clusters to implement the process and a set of 220 dendrograms using MLVA clusters to implement the process. The Consense program was then used with these sets of dendrograms to construct three generalized trees: one for PFGE cluster data, one for MLVA cluster data, and one for All-clusters data. In the latter case we weighted discrepant assignments by multiplying PFGE cluster data by a factor of 0.499 and multiplying the MLVA cluster data by a factor of 0.501. This weights the analysis in favor of MLVA under the assumption that there is more phylogenetically relevant information available from MLVA data compared with PFGE [7].

Three additional generalized trees were constructed using a subset of 80 of each of the sets of 220 dendrograms. These subsets correspond to the weight parameter r ranging between 0.3 and 0.6 at increments of 0.1 and the same threshold range as previously stated. Thus, a total of eight algorithms were examined including the six variations of the fusion algorithm described above together with analysis of the PFGE-only and MLVA-only data.

Assessing the validity of our analysis is complicated by the lack of a gold standard with which to compare our results. Indeed, barring a complete genome sequence for each strain and suitable algorithms for assessing genetic relationships, the only potential gold standards available are multi-locus sequence typing (MLST) and phenotypic characteristics. Given the potential lack of genetic variation for intra-serovar MLST comparisons [26], we chose to compare the *S. enterica* serovar Typhimurium strains using susceptibility to a panel of lytic bacteriophage as a measure of relatedness. Phage typing involves identification of the susceptibility of each strain to a panel of lytic bacteriophage. Our analysis assumes that strains with similar phage susceptibilities are more closely related than strains with dissimilar phage susceptibilities. All of the strains were subjected to susceptibility testing using a panel of 31 bacteriophage. Strains that were judged untypable with this panel were subject to testing with an additional 16 bacteriophage. Only one strain (8745) was considered untypable using the combined panel of 47 lytic bacteriophage (Tables 8 and 9).

The generalized tree constructed from the dendrograms that were obtained when All-clusters data were used (r range 0-1) highlights the potential for the discriminating power of the fusion algorithm (Fig. 4). For this tree, there are 14 instances of paired strains, i.e., strains that form a single node. Of these pairs, 6 are perfect matches, meaning that they share an identical response to the lytic phage panel (Tables 7, 8, and 9). Three pairs are considered close matches because they differ by 3 or fewer bacteriophage susceptibility tests. Three pairs are distant matches with phage susceptibilities differing by 4 to 7 reactions. Two pairs do not match according to the lytic phage panel (differ by at least 8 phage reactions and usually many more). The presence of paired strains with incongruous bacteriophage susceptibilities is not entirely surprising given that horizontal transmission of plasmids may be sufficient to alter the phage susceptibility profiles of *S. enterica* strains [27, 28]. Nevertheless, this analysis indicates that according to the fusion algorithm the majority of pairs are perfect or close matches based on their bacteriophage susceptibilities. An analysis across a subset of the range of r values (r range 0.3-0.6) produced a similar pattern of matches (Table 7).

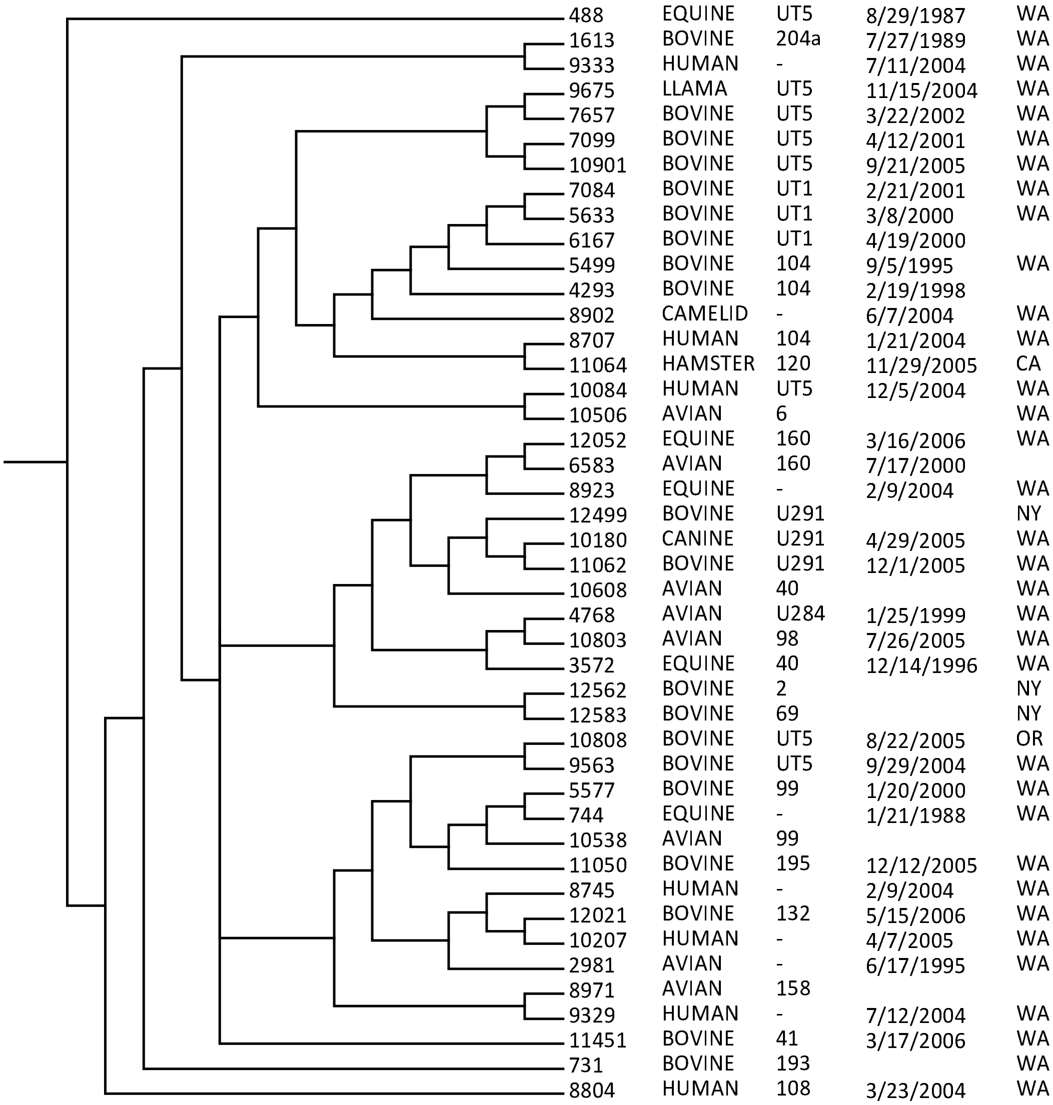
When the fusion algorithm was implemented using MLVA clusters or PFGE clusters, there were more defined pairs and fewer perfect matches according to bacteriophage susceptibility (Tables 8 and 9). Nevertheless, the results for all six fusion analyses demonstrate that strain matching results were relatively robust for both the full range of r values and a subset of these values (Table 7). The MLVA-only results (Fig. 5) were comparable to the results obtained using MLVA clusters, but are not as good as those obtained using All-clusters. While the number of perfect matches for both All-clusters and MLVA-only is comparable, MLVA-only mismatches five pairs whereas All-clusters mismatches two pairs. Moreover, if we consider strains coupled with pairs, All-clusters gives better results. For example, in the All-clusters dendrogram shown in Fig. 4, all three U291 phage types form a cluster, and the two 99 phage types form a cluster with the closely related strain 744 (Table 8). Interestingly, PFGE-only analysis (Fig. 6) produced more strain pairs [4], but relative to phage typing, PFGE was incapable of arriving at any perfect matches (Table 7). Consequently, while PFGE-only data may be useful for discriminating between strains, our results clearly demonstrate a failure of PFGE to link closely related strains. We surmise that this reduces the likelihood that the broader genetic relationships are captured accurately when this approach is used as has been suggested elsewhere [7]. Because the fusion algorithm produces much more robust matching results, we conclude that this strategy is more likely to reflect accurate intra-specific genetic relationships when examining epidemiologically unrelated strains of *S. enterica*.

Comparison of genetically similar *S. enterica* serovar Newport strains. Eight generalized dendrograms were constructed for *S. enterica* serovar Newport using the PFGE-only data, MLVA-only data, and the fusion algorithm with the same parameters discussed in the previous section except that the PFGE cluster data were weighted by a factor of 0.501 and the MLVA cluster data were weighted by a factor of 0.499 in the event of a completely discrepant assignment. This weights the analysis is favor of PFGE under the assumption that there is more phylogenetically relevant information available from PFGE data compared with MLVA data. The results for All-clusters (r between 0 and 1), MLVA-only, and PFGE-only are shown in Figs. 7, 8, and 9, respectively.

As with the serovar Typhimurium analysis, we lacked a gold standard for assessing the performance of our results relative to the true genetic relationships. In addition, for serovar Newport, we did not have a bacteriophage panel to use as a phenotypic surrogate for relatedness. Instead, we considered grouping within farms as well as susceptibility to antibiotics. On this basis, the MLVA-only results were poorest, and the fusion and PFGE-only algorithms performed comparably. The All-clusters and PFGE-only trees grouped isolates within farm with relatively high consistency (e.g., Farms A, J, and L; Figs. 7 and 9), whereas the MLVA-only dendrogram did poorly at grouping within farm isolates (Fig. 8). Comparison of the All-clusters tree (Fig. 7), with r ranging between 0 and 1, and the PFGE-only tree (Fig. 9), shows that All-clusters gives somewhat better results than PFGE-only. For example, the All-clusters tree groups seven Farm L strains and ten Farm J strains while the PFGE-only tree groups six Farm L strains and nine Farm J strains. Also, in the All-clusters tree all ten antibiotic-susceptible strains are grouped within the same cluster. We conclude from this analysis that the All-clusters fusion algorithm is best at identifying genetically related isolates and, as a consequence, the broader topological relationships using All-clusters are more likely to reflect the underlying genetic relationships.

3.5 Conclusions

In this chapter we presented a fusion algorithm that combines information from two widely used methods for molecular typing, pulsed-field gel electrophoresis (PFGE) and multiple-locus variable-number tandem repeat analysis (MLVA), to obtain phylogenetic relationships for two different sets of *Salmonella* strains. The strains from one set were *S. enterica* serovar Typhimurium obtained from diverse geographic locations and diverse animal hosts over a long period of time; we assumed these strains to represent a high degree of genetic diversity. The strains from the other set were *S. enterica* serovar Newport obtained mostly from the state of Washington within a short period of time from bovine hosts only; we assumed these strains to represent less genetic diversity. Results for the fusion algorithm were compared with those obtained using PFGE-only and MLVA-only results, and it was found that for both sets of data the fusion algorithm gave better results than either the PFGE-only or MLVA-only method. Thus, the fusion algorithm worked well to identify intra-specific genetic relationships for both strains with a high degree of genetic diversity and strains more closely related. In the future, we hope to have the opportunity to test the fusion algorithm with additional data.



488	EQUINE	UT5	8/29/1987	WA
1613	BOVINE	204a	7/27/1989	WA
9333	HUMAN	-	7/11/2004	WA
9675	LLAMA	UT5	11/15/2004	WA
7657	BOVINE	UT5	3/22/2002	WA
7099	BOVINE	UT5	4/12/2001	WA
10901	BOVINE	UT5	9/21/2005	WA
7084	BOVINE	UT1	2/21/2001	WA
5633	BOVINE	UT1	3/8/2000	WA
6167	BOVINE	UT1	4/19/2000	
5499	BOVINE	104	9/5/1995	WA
4293	BOVINE	104	2/19/1998	
8902	CAMELID	-	6/7/2004	WA
8707	HUMAN	104	1/21/2004	WA
11064	HAMSTER	120	11/29/2005	CA
10084	HUMAN	UT5	12/5/2004	WA
10506	AVIAN	6		WA
12052	EQUINE	160	3/16/2006	WA
6583	AVIAN	160	7/17/2000	
8923	EQUINE	-	2/9/2004	WA
12499	BOVINE	U291		NY
10180	CANINE	U291	4/29/2005	WA
11062	BOVINE	U291	12/1/2005	WA
10608	AVIAN	40		WA
4768	AVIAN	U284	1/25/1999	WA
10803	AVIAN	98	7/26/2005	WA
3572	EQUINE	40	12/14/1996	WA
12562	BOVINE	2		NY
12583	BOVINE	69		NY
10808	BOVINE	UT5	8/22/2005	OR
9563	BOVINE	UT5	9/29/2004	WA
5577	BOVINE	99	1/20/2000	WA
744	EQUINE	-	1/21/1988	WA
10538	AVIAN	99		
11050	BOVINE	195	12/12/2005	WA
8745	HUMAN	-	2/9/2004	WA
12021	BOVINE	132	5/15/2006	WA
10207	HUMAN	-	4/7/2005	WA
2981	AVIAN	-	6/17/1995	WA
8971	AVIAN	158		
9329	HUMAN	-	7/12/2004	WA
11451	BOVINE	41	3/17/2006	WA
731	BOVINE	193		WA
8804	HUMAN	108	3/23/2004	WA

Figure 4. Generalized tree for 44 *S. enterica* serovar Typhimurium isolates generated from 160 dendrograms. The dendrograms were obtained using the fusion algorithm with All-clusters (see text), the weight parameter *r* between 0 and 1, and the threshold parameter *thr* between 0.05 and 1. Information includes isolate designation, source, phage type, collection date, and state where the isolate was collected.

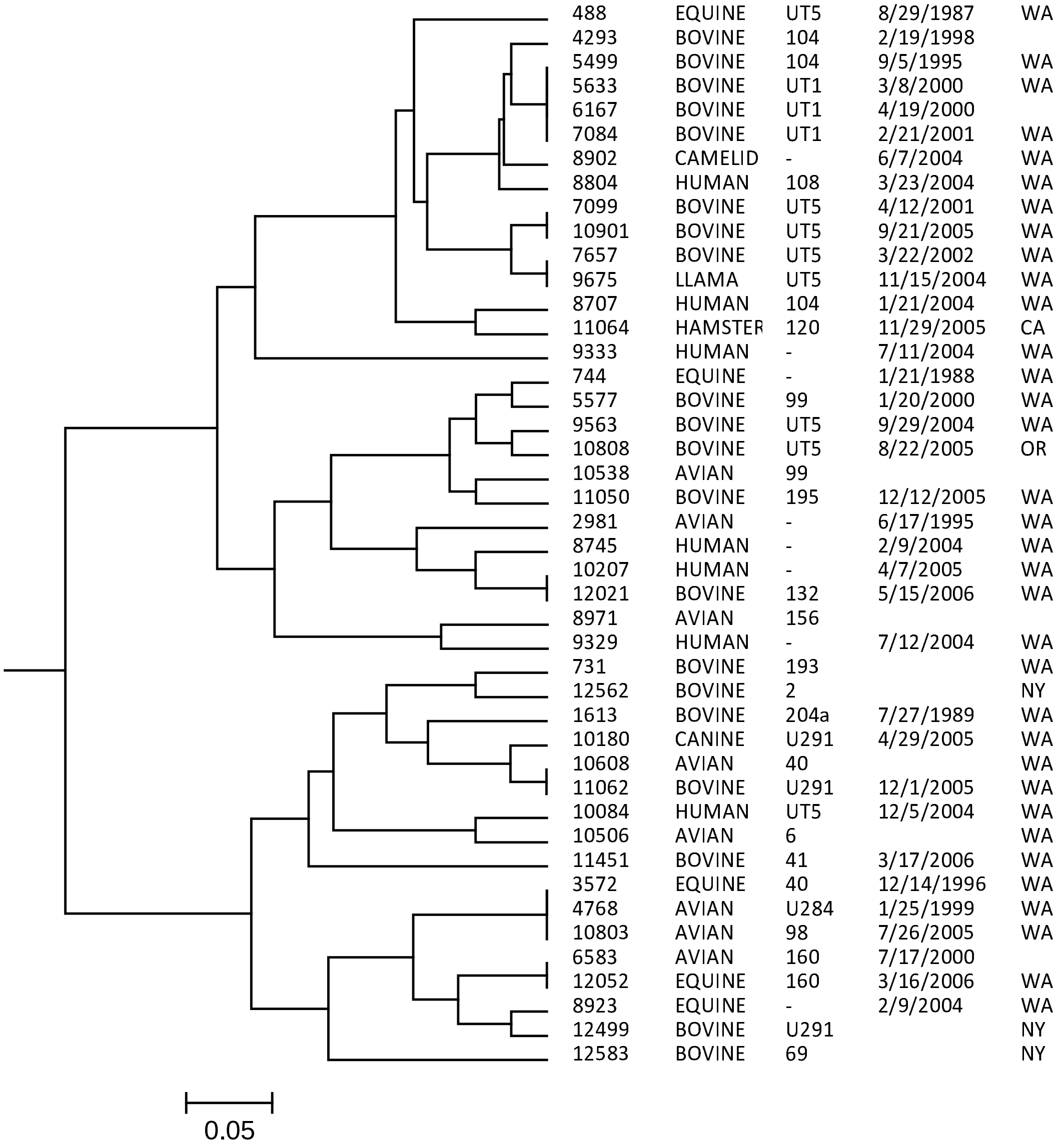

488 EQUINE UT5 8/29/1987 WA
4293 BOVINE 104 2/19/1998
5499 BOVINE 104 9/5/1995 WA
5633 BOVINE UT1 3/8/2000 WA
6167 BOVINE UT1 4/19/2000
7084 BOVINE UT1 2/21/2001 WA
8902 CAMELID - 6/7/2004 WA
8804 HUMAN 108 3/23/2004 WA
7099 BOVINE UT5 4/12/2001 WA
10901 BOVINE UT5 9/21/2005 WA
7657 BOVINE UT5 3/22/2002 WA
9675 LLAMA UT5 11/15/2004 WA
8707 HUMAN 104 1/21/2004 WA
11064 HAMSTER 120 11/29/2005 CA
9333 HUMAN - 7/11/2004 WA
744 EQUINE - 1/21/1988 WA
5577 BOVINE 99 1/20/2000 WA
9563 BOVINE UT5 9/29/2004 WA
10808 BOVINE UT5 8/22/2005 OR
10538 AVIAN 99
11050 BOVINE 195 12/12/2005 WA
2981 AVIAN - 6/17/1995 WA
8745 HUMAN - 2/9/2004 WA
10207 HUMAN - 4/7/2005 WA
12021 BOVINE 132 5/15/2006 WA
8971 AVIAN 156
9329 HUMAN - 7/12/2004 WA
731 BOVINE 193 WA
12562 BOVINE 2 NY
1613 BOVINE 204a 7/27/1989 WA
10180 CANINE U291 4/29/2005 WA
10608 AVIAN 40 WA
11062 BOVINE U291 12/1/2005 WA
10084 HUMAN UT5 12/5/2004 WA
10506 AVIAN 6 WA
11451 BOVINE 41 3/17/2006 WA
3572 EQUINE 40 12/14/1996 WA
4768 AVIAN U284 1/25/1999 WA
10803 AVIAN 98 7/26/2005 WA
6583 AVIAN 160 7/17/2000
12052 EQUINE 160 3/16/2006 WA
8923 EQUINE - 2/9/2004 WA
12499 BOVINE U291 NY
12583 BOVINE 69 NY
0.05

Figure 5. Dendrogram for 44 *S. enterica* serovar Typhimurium isolates constructed using UPGMA and a distance matrix obtained using a single-step stepwise mutation model for VNTR data. See Fig. 4 for isolate information.

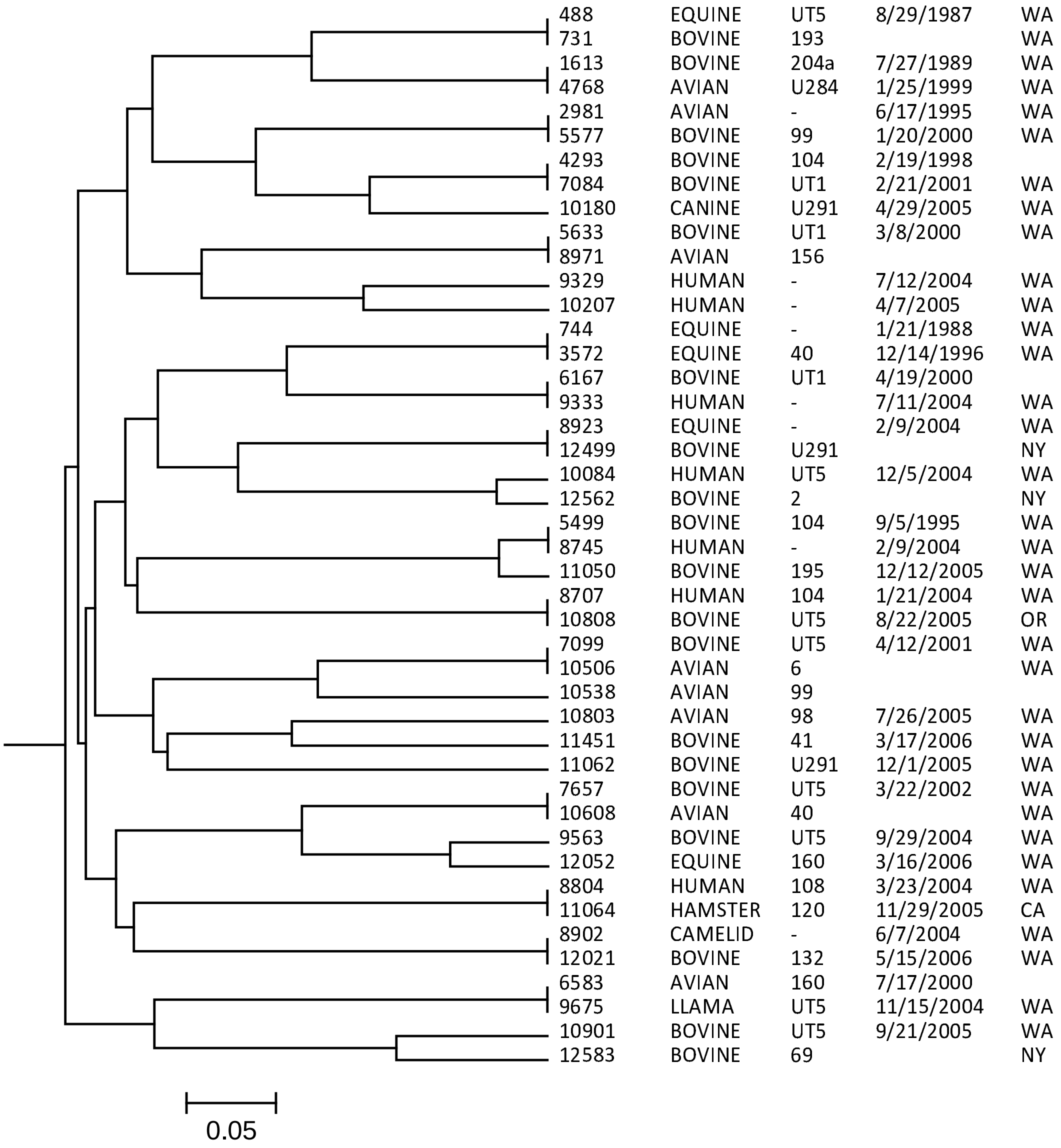
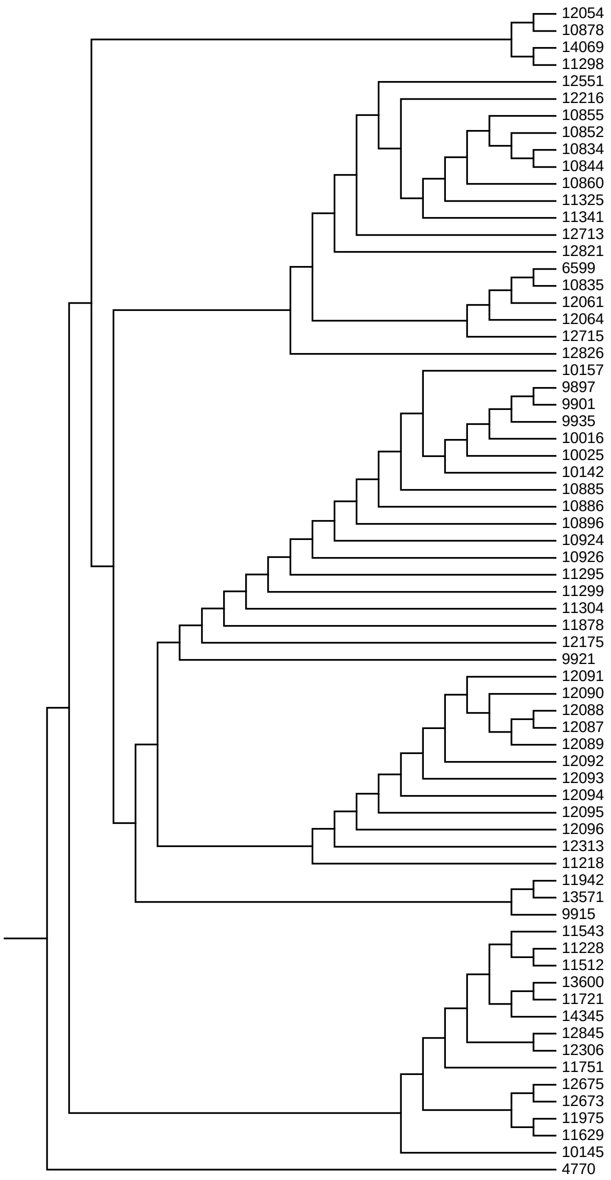
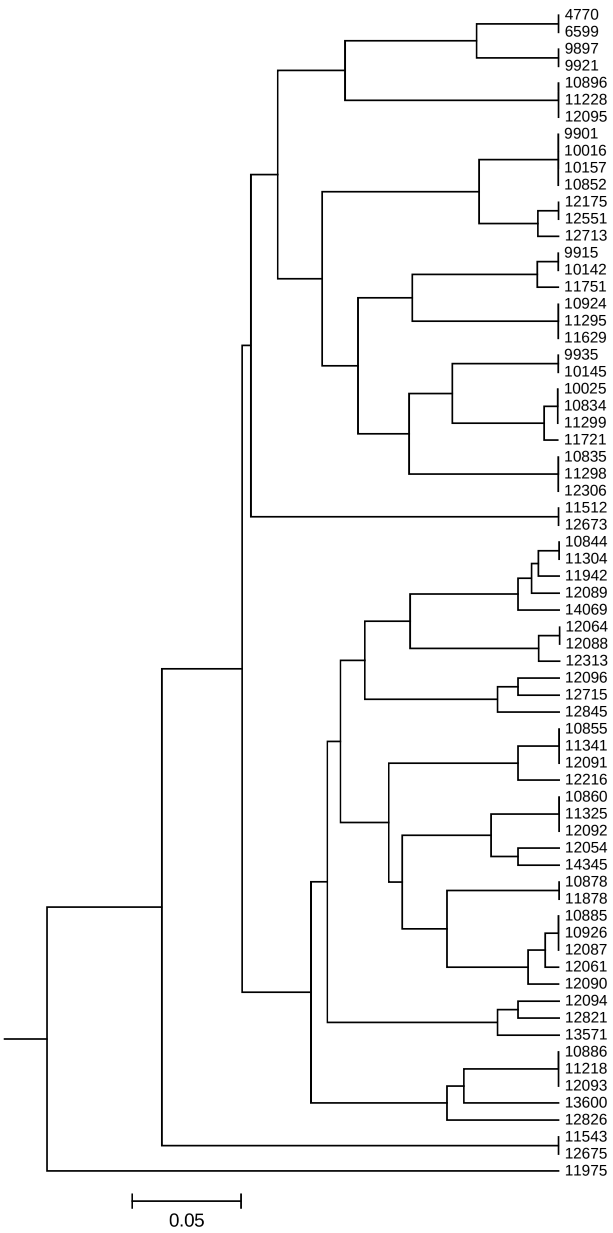

488 EQUINE UT5 8/29/1987 WA
731 BOVINE 193 WA
1613 BOVINE 204a 7/27/1989 WA
4768 AVIAN U284 1/25/1999 WA
2981 AVIAN - 6/17/1995 WA
5577 BOVINE 99 1/20/2000 WA
4293 BOVINE 104 2/19/1998
7084 BOVINE UT1 2/21/2001 WA
10180 CANINE U291 4/29/2005 WA
5633 BOVINE UT1 3/8/2000 WA
8971 AVIAN 156
9329 HUMAN - 7/12/2004 WA
10207 HUMAN - 4/7/2005 WA
744 EQUINE - 1/21/1988 WA
3572 EQUINE 40 12/14/1996 WA
6167 BOVINE UT1 4/19/2000
9333 HUMAN - 7/11/2004 WA
8923 EQUINE - 2/9/2004 WA
12499 BOVINE U291 NY
10084 HUMAN UT5 12/5/2004 WA
12562 BOVINE 2 NY
5499 BOVINE 104 9/5/1995 WA
8745 HUMAN - 2/9/2004 WA
11050 BOVINE 195 12/12/2005 WA
8707 HUMAN 104 1/21/2004 WA
10808 BOVINE UT5 8/22/2005 OR
7099 BOVINE UT5 4/12/2001 WA
10506 AVIAN 6 WA
10538 AVIAN 99
10803 AVIAN 98 7/26/2005 WA
11451 BOVINE 41 3/17/2006 WA
11062 BOVINE U291 12/1/2005 WA
7657 BOVINE UT5 3/22/2002 WA
10608 AVIAN 40 WA
9563 BOVINE UT5 9/29/2004 WA
12052 EQUINE 160 3/16/2006 WA
8804 HUMAN 108 3/23/2004 WA
11064 HAMSTER 120 11/29/2005 CA
8902 CAMELID - 6/7/2004 WA
12021 BOVINE 132 5/15/2006 WA
6583 AVIAN 160 7/17/2000
9675 LLAMA UT5 11/15/2004 WA
10901 BOVINE UT5 9/21/2005 WA
12583 BOVINE 69 NY
0.05

Figure 6. Dendrogram for 44 *S. enterica* serovar Typhimurium isolates constructed using UPGMA with Dice coefficients for PFGE data. See Fig. 4 for isolate information.



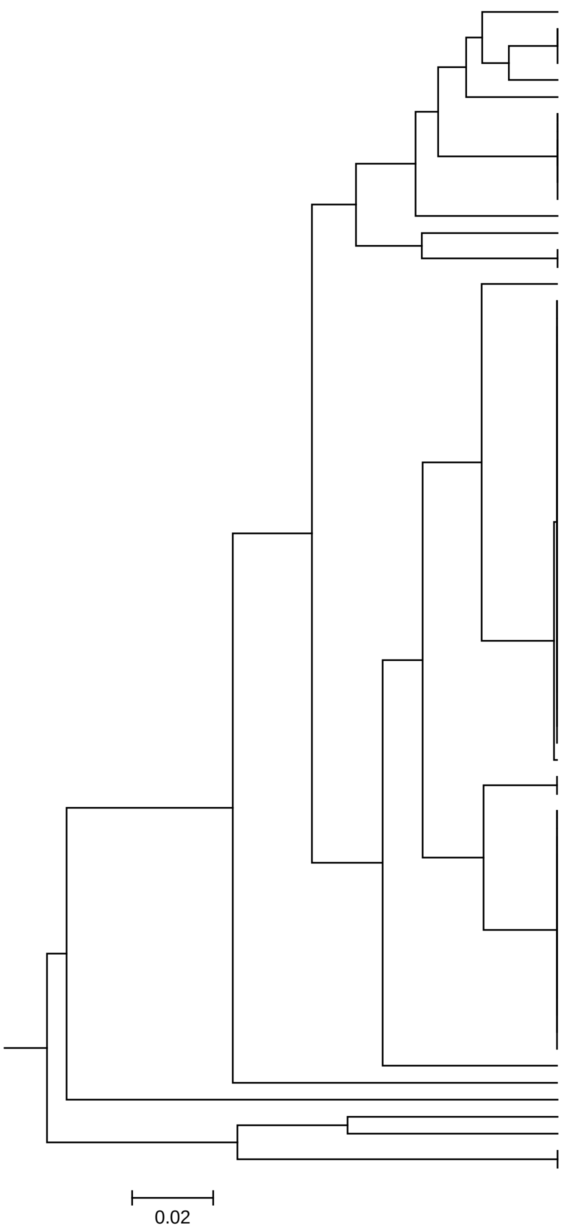
12054	U	4/24/2006	KING	ACSTAmcSuCaz
10878	A	10/17/2005	SNOHOMISH	ACSTAmcSuCaz
14069	DD	1/29/2007	TWIN FALLS	ACSTAmcSuCaz
11298	BB	2/6/2006	WHATCOM	ACKSxtSTAmcSuCaz
12551	Y		CLINTON	ACSTAmcSu
12216	G	6/5/2006	WHATCOM	ASTAmcSuCaz
10855	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10852	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10834	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10844	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10860	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11325	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
11341	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
12713	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
12821	Z	8/15/2006	YAKIMA	ACSTAmcSuCaz
6599	K	8/24/2000	JEROME	ACSTSu
10835	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
12061	P	5/25/2006	GOODING	ACSTAmcSuCaz
12064	P	5/3/2006	GOODING	ACSTAmcSuCaz
12715	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
12826	AA	8/15/2006	YAKIMA	ACKSTAmcSuCaz
10157	X	3/14/2005	WHATCOM	ASTAmcSuCaz
9897	H	12/22/2004	WHATCOM	ASTSuCaz
9901	N	1/3/2005	SNOHOMISH	ASSuCaz
9935	EE	2/1/2005	WHATCOM	ASTSuCaz
10016	D		WHATCOM	ASTSuCaz
10025	D	1/31/2005	WHATCOM	ASTAmcSuCaz
10142	F		WHATCOM	ASTAmcSuCaz
10885	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10886	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10896	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10924	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10926	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
11295	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11299	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11304	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11878	D		WHATCOM	ASTAmcSuCaz
12175	D		WHATCOM	ASTAmcSuCaz
9921	EE	1/27/2005	WHATCOM	ASTSuCaz
12091	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12090	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12088	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12087	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12089	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12092	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12093	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12094	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12095	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12096	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12313	R	6/20/2006	ADAMS	ASTAmcSuCaz
11218	C	2/1/2006	CLAY	ACSTAmcSuCaz
11942	BB	5/2/2006	WHATCOM	ASTAmcSuCaz
13571	D		WHATCOM	ASTAmcSuCaz
9915	M	1/3/2005	JEROME	ACKSTCaz
11543	C	2/1/2006	CLAY	ACSTSu
11228	C	2/1/2006	CLAY	ACSTSu
11512	C	2/1/2006	CLAY	ACSTSu
13600	V	10/23/2006	WHATCOM	SUSCEPT
11721	A	4/17/2006	SNOHOMISH	SUSCEPT
14345	E	3/30/2007	WHATCOM	SUSCEPT
12845	B	7/10/2006	GRANT	SUSCEPT
12306	Q	6/19/2006	GRANT	SUSCEPT
11751	D			SUSCEPT
12675	CC	7/5/2006	YAKIMA	SUSCEPT
12673	CC	7/5/2006	YAKIMA	SUSCEPT
11975	GG	5/8/2006	ADAMS	SUSCEPT
11629	S	3/15/2006	UTAH	SUSCEPT
10145	W	4/12/2005	JEROME	ACSTAmcSuCaz
4770	I	1/26/1999	SNOHOMISH	Su

Figure 7. Generalized tree for 69 *S. enterica* serovar Newport isolates generated from 160 dendrograms. The dendrograms were obtained using the fusion algorithm with All-clusters (see text), the weight parameter *r* between 0 and 1, and the threshold parameter *thr* between 0.05 and 1. Information includes isolate designation, collection date, county where the isolate was collected (Washington State), and antibiotic resistance phenotype (see methods). Resistance profile abbreviations: A, ampicillin; C, chloramphenicol; K, kanamycin; Sxt, trimethoprim-sulfa; S, streptomycin; T, tetracycline; Amc, amoxicillin-clavulanic acid; Su, triple-sulfa; Caz, ceftazidime.



4770	I	1/26/1999	SNOHOMISH	Su
6599	K	8/24/2000	JEROME	ACSTSu
9897	H	12/22/2004	WHATCOM	ASTSuCaz
9921	EE	1/27/2005	WHATCOM	ASTSuCaz
10896	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
11228	C	2/1/2006	CLAY	ACSTSu
12095	J	6/6/2006	WHATCOM	ASTAmcSuCaz
9901	N	1/3/2005	SNOHOMISH	ASSuCaz
10016	D		WHATCOM	ASTSuCaz
10157	X	3/14/2005	WHATCOM	ASTAmcSuCaz
10852	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
12175	D		WHATCOM	ASTAmcSuCaz
12551	Y			ACSTAmcSu
12713	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
9915	M	1/3/2005	JEROME	ACKSTCaz
10142	F		WHATCOM	ASTAmcSuCaz
11751	D			SUSCEPT
10924	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
11295	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11629	S	3/15/2006	UTAH	SUSCEPT
9935	EE	2/1/2005	WHATCOM	ASTSuCaz
10145	W	4/12/2005	JEROME	ACSTAmcSuCaz
10025	D	1/31/2005	WHATCOM	ASTAmcSuCaz
10834	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11299	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11721	A	4/17/2006	SNOHOMISH	SUSCEPT
10835	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11298	BB	2/6/2006	WHATCOM	ACKSxtSTAmcSuCaz
12306	Q	6/19/2006	GRANT	SUSCEPT
11512	C	2/1/2006	CLAY	ACSTSu
12673	CC	7/5/2006	YAKIMA	SUSCEPT
10844	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11304	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11942	BB	5/2/2006	WHATCOM	ASTAmcSuCaz
12089	J	6/6/2006	WHATCOM	ASTAmcSuCaz
14069	DD	1/29/2007	TWIN FALLS	ACSTAmcSuCaz
12064	P	5/3/2006	GOODING	ACSTAmcSuCaz
12088	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12313	R	6/20/2006	ADAMS	ASTAmcSuCaz
12096	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12715	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
12845	B	7/10/2006	GRANT	SUSCEPT
10855	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11341	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
12091	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12216	G	6/5/2006	WHATCOM	ASTAmcSuCaz
10860	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
11325	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
12092	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12054	U	4/24/2006	KING	ACSTAmcSuCaz
14345	E	3/30/2007	WHATCOM	SUSCEPT
10878	A	10/17/2005	SNOHOMISH	ACSTAmcSuCaz
11878	D		WHATCOM	ASTAmcSuCaz
10885	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10926	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
12087	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12061	P	5/25/2006	GOODING	ACSTAmcSuCaz
12090	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12094	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12821	Z	8/15/2006	YAKIMA	ACSTAmcSuCaz
13571	D		WHATCOM	ASTAmcSuCaz
10886	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
11218	C	2/1/2006	CLAY	ACSTAmcSuCaz
12093	J	6/6/2006	WHATCOM	ASTAmcSuCaz
13600	V	10/23/2006	WHATCOM	SUSCEPT
12826	AA	8/15/2006	YAKIMA	ACKSTAmcSuCaz
11543	C	2/1/2006	CLAY	ACSTSu
12675	CC	7/5/2006	YAKIMA	SUSCEPT
11975	GG	5/8/2006	ADAMS	SUSCEPT

Figure 8. Dendrogram for 69 *S. enterica* serovar Newport isolates constructed using UPGMA and a distance matrix obtained using a single-step stepwise mutation model for VNTR data. See Fig. 7 for isolate information



4770	I	1/26/1999	SNOHOMISH	Su
6599	K	8/24/2000	JEROME	ACSTSu
12061	P	5/25/2006	GOODING	ACSTAmcSuCaz
12064	P	5/3/2006	GOODING	ACSTAmcSuCaz
9897	H	12/22/2004	WHATCOM	ASTSuCaz
12826	AA	8/15/2006	YAKIMA	ACKSTAmcSuCaz
11228	C	2/1/2006	CLAY	ACSTSu
11512	C	2/1/2006	CLAY	ACSTSu
11543	C	2/1/2006	CLAY	ACSTSu
11721	A	4/17/2006	SNOHOMISH	SUSCEPT
13600	V	10/23/2006	WHATCOM	SUSCEPT
14345	E	3/30/2007	WHATCOM	SUSCEPT
14069	DD	1/29/2007	TWIN FALLS	ACSTAmcSuCaz
11218	C	2/1/2006	CLAY	ACSTAmcSuCaz
12306	Q	6/19/2006	GRANT	SUSCEPT
12845	B	7/10/2006	GRANT	SUSCEPT
9901	N	1/3/2005	SNOHOMISH	ASSuCaz
9921	EE	1/27/2005	WHATCOM	ASTSuCaz
9935	EE	2/1/2005	WHATCOM	ASTSuCaz
10142	F		WHATCOM	ASTAmcSuCaz
10157	X	3/14/2005	WHATCOM	ASTAmcSuCaz
10885	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10886	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10896	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10924	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
10926	A	10/24/2005	SNOHOMISH	ASTAmcSuCaz
11295	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11299	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11304	BB	2/6/2006	WHATCOM	ASTAmcSuCaz
11878	D		WHATCOM	ASTAmcSuCaz
11942	BB	5/2/2006	WHATCOM	ASTAmcSuCaz
12087	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12088	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12089	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12090	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12091	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12092	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12094	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12095	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12096	J	6/6/2006	WHATCOM	ASTAmcSuCaz
12175	D		WHATCOM	ASTAmcSuCaz
12216	G	6/5/2006	WHATCOM	ASTAmcSuCaz
12313	R	6/20/2006	ADAMS	ASTAmcSuCaz
13571	D		WHATCOM	ASTAmcSuCaz
12093	J	6/6/2006	WHATCOM	ASTAmcSuCaz
10016	D		WHATCOM	ASTSuCaz
10025	D	1/31/2005	WHATCOM	ASTAmcSuCaz
10145	W	4/12/2005	JEROME	ACSTAmcSuCaz
10834	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10835	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10844	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10852	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10855	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10860	L	9/26/2005	SNOHOMISH	ACSTAmcSuCaz
10878	A	10/17/2005	SNOHOMISH	ACSTAmcSuCaz
11325	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
11341	L	2/21/2006	SNOHOMISH	ACSTAmcSuCaz
12054	U	4/24/2006	KING	ACSTAmcSuCaz
12551	Y			ACSTAmcSu
12713	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
12715	MM	7/17/2006	YAKIMA	ACSTAmcSuCaz
12821	Z	8/15/2006	YAKIMA	ACSTAmcSuCaz
9915	M	1/3/2005	JEROME	ACKSTCaz
11298	BB	2/6/2006	WHATCOM	ACKSxtSTAmcSuCaz
11751	D			SUSCEPT
11629	S	3/15/2006	UTAH	SUSCEPT
11975	GG	5/8/2006	ADAMS	SUSCEPT
12673	CC	7/5/2006	YAKIMA	SUSCEPT
12675	CC	7/5/2006	YAKIMA	SUSCEPT

Figure 9. Dendrogram for 69 *S. enterica* serovar Newport isolates constructed using UPGMA with Dice coefficients for PFGE data. See Fig. 7 for isolate information.

Table 6. Primer sequences used for six VNTR loci from *S. enterica* serovar Newport

	Locus set	Forward primer[a]	Reverse primer	Repeat size[b]	Amplicon size[c]
PCR1					
	STTR6	5′-6FAM-TCGGGCATGCGTTGAAA	5′-CTGGTGGGGAGAATGACTGG	6	397
	NEWPORT-A	5′-PET-ACTGAAAGGAAGGGGAGAGC	5′-GTCAGGGTGGAATAGAATGC	9	429
	NEWPORT-L	5′-VIC-GAAGTACCGAAGTGGGTGAT	5′-CGTCCGTTAGAGGAACGTAT	51	529
PCR2					
	STTR5	5′-PET-ATGGCGAGGCGAGCAGCAGT	5′-GGTCAGGCCGAATAGCAGGAT	6	564
	NEWPORT-B	5′-VIC-GGCCGATATAGCTCAGTTGG	5′-GAACCTCGCTTAGGGTTGTG	12	350
	NEWPORT-M	5′-6FAM-GGTCATAGAGGGTCTGCAT	5′-ATGGAGCACAGACCACTAAC	36	378

[a] Forward primers (arbitrarily designated) were conjugated to one of three fluorescent dyes (6FAM, PET, or VIC).
[b] Approximate size (bp) of single repeat units
[c] Size (bp) of resulting amplicon based on reference sequence.

Table 7. Correspondence between classification algorithm and phage typing for 44 *S. enterica*

Algorithm[a]			No. of		
	Total no. of pairs[b]	Perfect Matches[c]	Close Matches	Distant Matches	No Matches
All-clusters, r: 0-1	14	6	3	3	2
All-clusters, r: 0.3-0.6	14	6	2	3	3
MLVA clusters, r: 0-1	16	4	6	2	4
MLVA clusters, r: 0.3-0.6	15	4	4	2	5
PFGE clusters, r: 0-1	15	4	4	2	5
PFGE clusters, r: 0.3-0.6	15	4	4	2	5
MLVA-only	15	4	4	2	5
PFGE-only	20	0	3	7	10

[a] All-clusters = generalized dendrogram for MLVA clusters and PFGE clusters; MLVA clusters = fusion algorithm initiated with MLVA dendrogram; PFGE clusters = fusion algorithm initiated with PFGE dendrogram; ranges of the weight parameter r as shown. MLVA-only = dendrogram from MLVA data only (Fig. 5); PFGE-only = dendrogram from PFGE data only (Fig. 6)

[b] The number of paired isolates in the respective dendrogram

[c] A match refers to the similarity of each pair to phage susceptibility (Tables 8 and 9); a perfect match means that the pair has identical susceptibility; a close match means that a pair differs by 3 or fewer susceptibilities; a distant match indicates a difference of 4 to 7; and no match means a difference of at least 8 and typically 15 or more.

Table 8. Phage panel results for 44 *S. enterica* serovar Typhimurium isolates and 31 bacteriophage. A plus sign indicates a lytic (positive) reaction. Common phage types are listed in the first column in regular font. Atypical isolates are listed by their identification numbers in italic font.

PT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
8745																															
UT5																															
UT1																															
195																															
193																															
99											+																				
204a														+																	
120																		+													
744											+		+																		
98											+										+										
104												+	+					+													
108												+	+																		+
9333													+	+	+																+
8923											+		+					+	+			+									
160											+		+				+	+	+	+		+								+	
156											+	+	+	+		+		+							+	+	+		+		+
6						+										+											+		+		+
69										+						+				+		+	+						+	+	
U291				+	+	+					+			+	+	+		+	+						+	+	+				+
9329				+							+		+						+		+	+									+
U284		+									+								+	+	+									+	
10207			+	+	+	+					+		+	+	+	+			+		+			+					+		+
8902				+	+	+				+	+	+	+	+	+	+		+	+		+			+		+			+		+
2981		+		+		+					+	+	+						+	+	+	+								+	+
132		+	+	+	+	+					+	+	+	+	+	+	+		+		+	+		+		+	+		+		+
2		+	+	+	+	+				+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+		+	+	+
41	+	+	+	+	+	+	+			+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+
40	+	+	+	+	+	+	+			+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+

Table 9. Additional phage panel results for 44 *S. enterica* serovar Typhimurium isolates and 16 bacteriophage. Isolates without a positive response to the first panel of 31 bacteriophage were subjected to further testing. A plus sign indicates a lytic (positive) reaction. Common phage types are listed in the first column in regular font. Atypical isolates are listed by their identification number in italic font. nd means not determined. One isolate, 8745, was untypable using the 47 bacteriophage.

PT	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
8745																
UT5																+
UT1														+		+
195			+	+	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
193	+	+	+	+	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd

Table 10: *S. enterica* serovar Typhimurium VNTR data. The strain number is given in the first column, and the remaining five columns contain the total number of tandem repeats for five different loci. Only STTR5 and STTR6 data were used in the MLVA analysis. STTR11 and STTR9 provided very little variation, and too many values were missing for STTR10pl.

SNUM	STTR11	STTR9	STTR5	STTR6	STTR10pl
488	589	169	268	320	0
731	589	169	256	309	0
744	589	169	280	303	0
1613	598	160	233	309	348
2981	598	160	298	332	348
3572	589	160	210	279	0
4293	598	169	262	344	478
4768	589	160	210	279	0
5499	598	169	262	338	323
5577	589	178	286	303	341
5633	598	169	262	338	341
6167	0	169	262	338	341
6583	598	160	216	291	416
7084	598	169	262	338	341
7099	598	160	280	332	379
7657	598	160	274	332	379
8707	598	169	274	350	403
8745	589	169	304	309	0
8804	589	169	268	338	0
8902	0	169	262	332	471
8923	598	160	228	285	385
8971	589	169	280	0	341
9329	589	0	298	0	341
9333	598	160	304	350	354
9563	598	160	274	309	379
9675	598	160	274	332	385
10084	598	160	245	320	385
10180	598	160	239	291	372
10207	598	160	298	314	354
10506	598	160	245	332	379

10538	598	160	280	291	397
10608	598	0	239	297	366
10803	589	160	210	279	0
10808	598	160	280	309	379
10901	598	160	280	332	385
11050	598	169	274	297	452
11062	598	160	239	297	354
11064	598	169	268	356	428
11451	598	160	245	0	360
12021	598	160	298	314	0
12052	0	160	216	291	379
12499	598	160	228	291	366
12562	589	169	256	297	0
12583	589	178	0	309	0

Table 11: *S. enterica* serovar Newport VNTR data. The strain number is given in the first column, and the remaining six columns contain the total number of tandem repeats for six different loci.

SNUM	STTR5	STTR6	NWPTA	NWPTB	NWPTM	NWPTL
4770	227	325	406	349	380	531
6599	216	325	406	349	380	531
9897	216	313	406	349	380	0
9901	216	313	406	349	380	531
9915	221	319	406	349	380	531
9921	210	313	406	349	380	531
9935	216	313	406	349	380	531
10016	216	313	406	349	380	531
10025	216	313	406	349	380	531
10142	216	313	406	349	380	531
10145	216	319	388	337	380	531
10157	216	313	406	349	380	531
10834	216	319	406	349	380	531
10835	216	325	406	349	380	531
10844	216	319	406	349	380	531
10852	216	319	406	349	380	531
10855	216	319	406	349	380	531
10860	216	319	406	349	380	531
10878	216	343	406	349	380	531
10885	216	313	406	349	380	531
10886	216	313	406	349	380	531
10896	216	313	406	349	380	531
10924	216	313	406	349	380	531
10926	216	313	406	349	380	531
11218	216	307	415	349	380	531
11228	233	313	352	349	380	408
11295	216	313	406	349	380	531
11298	216	331	406	349	380	531
11299	216	313	406	349	380	531
11304	216	313	406	349	380	531
11325	216	319	406	349	380	531
11341	216	319	406	349	380	531
11512	233	313	352	349	380	408
11543	233	313	352	349	380	408
11629	221	0	388	337	335	430

11721	233	307	352	349	380	408
11751	251	0	352	349	380	408
11878	216	313	406	349	380	531
11942	221	313	406	349	380	531
11975	227	0	379	0	344	430
12054	216	337	406	349	380	531
12061	216	325	406	349	380	531
12064	216	325	406	349	380	531
12087	216	307	406	349	380	0
12088	216	307	406	349	380	531
12089	216	307	406	349	380	531
12090	216	307	406	349	380	531
12091	216	307	406	349	380	531
12092	216	307	406	349	380	531
12093	216	307	406	349	380	531
12094	216	0	406	349	380	531
12095	216	307	406	349	380	531
12096	216	307	406	349	380	531
12175	216	313	406	349	380	531
12216	216	319	406	349	380	0
12306	233	307	370	349	380	0
12313	216	307	406	349	380	0
12551	216	319	406	349	380	0
12673	221	0	460	0	344	430
12675	221	0	460	0	344	430
12713	216	319	406	349	380	0
12715	216	325	406	349	380	0
12821	216	319	406	349	380	0
12826	210	325	406	349	380	0
12845	233	307	370	349	380	0
13571	221	313	406	349	380	0
13600	233	307	0	349	380	408
14069	216	331	406	349	380	0
14345	233	307	352	349	380	408

List of References

[1] CDC, "National antimicrobial resistance monitoring system for enteric bacteria (NARMS). Human isolates final report, 2005." *U.S. Department of Health and Human Services*, 2008.

[2] S. L. Foley, S. Zhao, and R. D. Walker, "Comparison of molecular typing methods for the differentiation of *Salmonella* foodborne pathogens," *Foodborne Pathog Dis*, vol. 4, no. 3, pp. 253–76, 2007.

[3] P. Gerner-Smidt, K. Hise, J. Kincaid, S. Hunter, S. Rolando, E. Hyytia-Trees, E. M. Ribot, and B. Swaminathan, "PulseNet USA: a five-year update," *Foodborne Pathog Dis*, vol. 3, no. 1, pp. 9–19, 2006.

[4] S. Lukinmaa, U. M. Nakari, M. Eklund, and A. Siitonen, "Application of molecular genetic methods in diagnostics and epidemiology of food-borne bacterial pathogens," *APMIS*, vol. 112, no. 11-12, pp. 908–29, 2004.

[5] B. A. Lindstedt, E. Heir, E. Gjernes, and G. Kapperud, "DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci," *J Clin Microbiol*, vol. 41, no. 4, pp. 1469–79, 2003.

[6] J. Zheng, C. E. Keys, S. Zhao, J. Meng, and E. W. Brown, "Enhanced subtyping scheme for *Salmonella enteritidis*," *Emerg Infect Dis*, vol. 13, no. 12, pp. 1932–5, 2007.

[7] M. A. Davis, D. D. Hancock, T. E. Besser, and D. R. Call, "Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of *Escherichia coli* O157:H7," *J Clin Microbiol*, vol. 41, no. 5, pp. 1843–9, 2003.

[8] A. Van Belkum, S. Scherer, L. van Alphen, and H. Verbrugh, "Short-sequence DNA repeats in prokaryotic genomes," *Microbiol Mol Biol Rev*, vol. 62, no. 2, pp. 275–93, 1998.

[9] D. R. Call, L. Orfe, M. A. Davis, S. Lafrentz, and M. S. Kang, "Impact of compounding error on strategies for subtyping pathogenic bacteria," *Foodborne Pathog Dis*, vol. 5, no. 4, pp. 505–16, 2008.

[10] K. L. Hopkins, C. Maguire, E. Best, E. Liebana, and E. J. Threlfall, "Stability of multiple-locus variable-number tandem repeats in *Salmonella enterica* serovar typhimurium," *J Clin Microbiol*, vol. 45, no. 9, pp. 3058–61, 2007.

[11] B. A. Lindstedt, M. Torpdahl, E. M. Nielsen, T. Vardund, L. Aas, and G. Kapperud, "Harmonization of the multiple-locus variable-number tandem repeat

analysis method between Denmark and Norway for typing *Salmonella* typhimurium isolates and closer examination of the VNTR loci," *J Appl Microbiol*, vol. 102, no. 3, pp. 728–35, 2007.

[12] B. A. Lindstedt, T. Vardund, L. Aas, and G. Kapperud, "Multiple-locus variable-number tandem-repeats analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium using PCR multiplexing and multicolor capillary electrophoresis," *J Microbiol Methods*, vol. 59, no. 2, pp. 163–72, 2004.

[13] P. Gulati, R. K. Varshney, and J. S. Virdi, "Multilocus variable number tandem repeat analysis as a tool to discern genetic relationships among strains of *Yersinia enterocolitica* biovar 1A," *J Appl Microbiol*, 2009.

[14] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic Acids Res*, vol. 27, no. 2, pp. 573–580., 1999.

[15] A. J. Vogler, C. Keys, Y. Nemoto, R. E. Colman, Z. Jay, and P. Keim, "Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7," *J Bacteriol*, vol. 188, no. 12, pp. 4253–63, 2006.

[16] A. J. Vogler, C. E. Keys, C. Allender, I. Bailey, J. Girard, T. Pearson, K. L. Smith, D. M. Wagner, and P. Keim, "Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*," *Mutat. Res.*, vol. 616, no. 1-2, pp. 145–58, 2007.

[17] A. Bauer, M. Kirby, J. Sherris, and M. Turck, "Antibiotic susceptibility testing by a standard single disk method," *Am J Clin Path*, vol. 45, pp. 493–496, 1966.

[18] National Committee for Clinical Laboratory Standards, "Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: Approved standard M7-A6," *NCCLS Villanova, PA, USA*, 2003.

[19] National Committee for Clinical Laboratory Standards, "Performance standards for antimicrobial susceptibility testing, 14th informational supplement, 13th ed. approved standard M100-S13." *NCCLS, Wayne, Pa.*, 2003.

[20] E. M. Ribot, M. A. Fair, R. Gautom, D. N. Cameron, S. B. Hunter, B. Swaminathan, and T. J. Barrett, "Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet," *Foodborne Pathog Dis*, vol. 3, no. 1, pp. 59–67, 2006.

[21] M. Davis, K. Baker, D. Call, L. Warnick, Y. Soyer, M. Wiedmann, Y. Gröhn, P. McDonough, D. Hancock, and T. Besser, "Multiple locus variable number of tandem repeats typing method for *Salmonella enterica* serovar Newport," *J Clin Microbiol*, vol. in press., 2009.

[22] B. Walsh, “Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals,” *Genetics*, vol. 158, no. 2, pp. 897–912, 2001. [Online]. Available: http://www.genetics.org/cgi/content/abstract/158/2/897

[23] T. G. Dietterich, “Ensemble methods in machine learning,” *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[24] H. Harbottle, D. G. White, P. F. McDermott, R. D. Walker, and S. Zhao, “Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates,” *J Clin Microbiol*, vol. 44, no. 7, pp. 2449–57, 2006.

[25] J. Felsenstein, “Phylip (phylogeny inference package) version 3.5c. distributed by the author. department of genetics, university of washington, seattle,” 1993.

[26] M. K. Fakhr, L. K. Nolan, and C. M. Logue, “Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella enterica* serovar Typhimurium,” *J Clin Microbiol*, vol. 43, no. 5, pp. 2215–9, 2005.

[27] B. Içgen, G. C. Gürakan, and G. Özcengiz, “Effects of plasmid curing on antibiotic susceptibility, phage type, lipopoly saccharide and outer membrane protein profiles in local *Salmonella* isolates,” *Food Microbiol*, vol. 18, no. 6, pp. 631 – 635, 2001.

[28] G. Bezanson, R. Khakhria, and R. Lacroix., “Involvement of plasmids in determining bacteriophage sensitivity in *Salmonella typhimurium*: genetic and physical analysis of phagovar 204,” *Can J Microbiol*, vol. 28, pp. 993 – 1001, 1982.

CHAPTER 4

Studying the Evolution of Bacterial Plasmids

4.1 Introduction

Understanding the evolution of microbial organisms is one of the most important objectives in microbiology. With the development of many bacterial genome projects, molecular sequences have become widely used in phylogenetic analyses of microbial organisms [1]. Often the sequences of housekeeping genes or conserved structural sequences are used for these analyses. In particular, 16S rRNA sequences are considered to be the gold standard in building a taxonomy for bacteria [2]. However, considerable incompatibility with the results of phylogenetic analysis has been reported. Inconsistencies can occur for closely related species and even for distantly separated taxa [3].

Studies have shown that a number of factors may affect the accuracy of phylogenetic analysis, including the biological processes underlying the molecular sequences and the computational methods used in the analysis [3]. From the perspective of biological sequences, it has been shown that evolution of molecular sequences can only provide indirect and incomplete information for species evolution. Different genes may contain inconsistent and even conflicting information about species evolution. The heterogeneity of genes can be due to gene duplication, horizontal gene transfer, and the merging of genes (coalescence) [4]. Different computational methods may also lead to incongruent results because of differing assumptions about the underlying mutation models and also because of different schemes for selecting the optimum phylogenetic results.

To overcome the effect of biological and computational factors, the analysis of multiple genes rather than only one gene has been suggested [3, 5]. The results show that many conflicts can be resolved by using more genes in the phylogenetic analysis; in addition, there is strong support for monophyly in phylogenetic analysis with multiple gene usage [3, 5]. Nevertheless, the use of multiple genes may cause issues because of their heterogeneity. Methods that model different evolution rates for the genes have been proposed, and the results have shown that the effect of heterogeneity can be reduced [6–8]. It has also been found that genes evolving quickly may cause the phenomenon of long-branch attraction (LBA) in which divergent but unrelated species appear to be closely related in phylogeny [6]. Thus, it is suggested that slowly evolving genes are preferred for use in phylogenetic analysis [5]. After gene selection, the "correct" evolution rates for different sites among the genes can be modeled for phylogenetic analysis [9].

With the emergence and spread of antimicrobial resistance, interest in plasmids has grown. Because plasmids are capable of cell-to-cell transfer between bacterial species, horizontal gene transfer (HGT) is common and can happen between genetically distant species. Over time plasmids can gain or lose DNA segments from different host chromosomes and other plasmids resulting in a mosaic structure that often includes multiple genes for resistance to antibiotics. Because of their heterogeneous structure, constructing a phylogenetic tree for plasmids is challenging.

The focus of this work is constructing a reliable phylogenetic tree for microbes with dynamic genetic contents, especially for plasmids. We begin by studying the effects of different factors, from selecting different set of genes to using different computational methods for phylogenetic analysis. In particular, we identify a set of conserved genes based on their similarity and use only these conserved genes in our phylogenetic analyses. The heterogeneity of genes is compared using estimates of their evolution rates. We use our method on two different data sets, a set of eight plasmids [10] that harbor a bla$_{CMY_2}$ antibiotic resistance gene and a set of 106 Gram-negative, enteric plasmids from the NCBI database (http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2157). Statistical tests and expert knowledge are used to evaluate the results.

4.2 Methods

4.2.1 Detection of conserved regions

Because the genetic makeup of plasmids is often due to HGT, the composition of different plasmids will reflect totally different evolutionary scenarios. Thus, a phylogenetic analysis should begin by determining which DNA sequences should be used and which should be excluded from the analysis. For a group of closely related plasmids, most genetic content should come from a recent common ancestor; however, even these plasmids can contain diverse genetic content. Genes from a common ancestor should contain more information about the evolutionary relationship between plasmids so we refer to these as "conserved genes." After conserved genes have been identified, multiple sequence alignment can be performed with these genes and the alignment results used for further analysis. Biologically, the conserved genes are often important for plasmid transfer and maintenance. Deleterious mutations in these sequences are likely to lead to plasmid loss [11].

A conserved region among several nucleotide sequences can be identified by sequence comparison. For a set S of n plasmids $S = (p^1, p^2, \ldots, p^n)$, a conserved gene can be represented by the tuple $G_i = (g_{i1}, g_{i2}, \ldots, g_{in})$ where i denotes the ith conserved gene and g_{ik} denotes the specific gene from the kth plasmid. Similarity scores for each pair of genes in a tuple can be obtained via BLAST [12] using the bl2seq program with the score defined as (length of matching sequence)*(BLAST identity score)/(length of reference gene + length of matching sequence gene). A tuple is considered conserved when all scores are above a chosen threshold. Doing a

BLAST search for all pairs of genes is very time consuming and is only practical for a small set of plasmids. An alternative method is to build a BLAST database using all the genes from all the plasmids under consideration. Each plasmid sequence can then be compared against the database and the results used to identify conserved tuples.

As mentioned above, the selection of conserved genes is based on a threshold value for the score obtained via BLAST. The maximum possible value, 0.5, occurs when there is a perfect match between a gene pair. A threshold value of 0.3 is used in this study. Typically it was found that scores are either between 0.3 and 0.5 or else less than about 0.1. Usually each gene has just one high score, but when there is more than one high score for a particular gene, it is assumed to be the result of gene duplication. When this occurs, the gene is removed from the pool of conserved genes.

Virtual mixed-plasmid microarray

For a very large set of plasmids, it may be impossible to find conserved genes shared by all the plasmids. In this case, a virtual mixed-plasmid microarray (MPM) can be used efficiently to construct a preliminary tree for all the plasmids based on gross similarity [13]. This tree can then be used to identify several smaller subgroups of plasmids. If the similarity of all plasmids in a subgroup is higher than a given threshold, shared genes can be identified and used to more precisely determine the relationship between these plasmids.

For the study of 106 plasmids described below, 100 gene segments from each plasmid sequence were randomly selected to create a virtual MPM. Virtual hybridization of these probes was accomplished using BLAST scores as proxies for array probe intensities. To construct a reliable dendrogram using a virtual MPM, a number of virtual MPMs should be generated. The optimal dendrogram is obtained from the best virtual MPM which in turn is obtained when probes are of optimal length. In practice this means that each virtual MPM is generated using different probe lengths that range between 100 and 1000 bp. Distance matrices are calculated from the "hybridizations," and correlation coefficients between a given distance matrix and all other distance matrices are calculated. Because each distance matrix represents one probe length, the optimal probe length occurs when its correlation coefficient is high and those of its two neighboring probe lengths are close in value, i.e., when a plot of the correlation coefficients as a function of probe length is flattest. The optimal probe length for "hybridizations" representing the 106 plasmids was found to be 500 bp.

4.2.2 Multiple alignment of conserved genes

Having obtained the conserved genes for a group of plasmids, the next step is to complete a multiple sequence alignment for these conserved genes. There are two ways of doing this. If we assume that K conserved genes have been found $G_1 = (g_{11}, g_{12}, \ldots, g_{1n})$, $G_2 = (g_{21}, g_{22}, \ldots, g_{2n})$, …, and $G_K = (g_{K1}, g_{K2}, \ldots, g_{Kn})$, then by concatenating all conserved genes for each plasmid we get $P_1 = [g_{11}g_{21} \ldots g_{K1}]$, $P_2 = [g_{12}g_{22} \ldots g_{K2}]$, …, and $P_n = [g_{1n}g_{2n} \ldots g_{Kn}]$. The alignment can be done for P_1 to P_n. This method works when all conserved genes can be assumed to be orthologous. This method does not scale very well when P_1 to P_n is very long.

Another approach is to align each tuple separately, i.e., sequences of g_{11} to g_{1n} are aligned, and we denote the result as a_1. The process is repeated for G_2 to G_K. Assuming that the g_{mn}'s are orthologous, a_1 to a_K can be concatenated to obtain the entire alignment. When ClustalW is used to perform the alignment using this approach, the time complexity is reduced considerably. For example, to compare a set of eight plasmids of average length 10 kbp, the time is reduced from approximately 20 hours for concatenated sequences to approximately 30 minutes for separate tuple alignment. This approach also allows the use of different mutation models for different a_j to compensate for the heterogeneity of different genes in the plasmids.

Distance based phylogenetic methods were used to generate the evolutionary tree from the results of the alignment. Various mutation models can be used to generate the distance matrix. Of these methods the Jukes-Cantor (JC) model is the simplest [14]. In the JC model, each base in a DNA sequence has an equal mutation rate, and all complementary pairs of the four nucleotides A, T, C, and G have an equal substitution rate, i.e., $p(A \leftrightarrow G) = p(A \leftrightarrow C) = \ldots = \alpha$, where α is the substitution rate. The JC distance between two DNA sequences S_0 and S_1 is defined as:

$$d_{JC} = -\frac{3}{4} ln(1 - \frac{4}{3} p) \tag{18}$$

where p is the fraction of the sites that disagree in S_0 and S_1.

After generating the distance matrix using the JC model, we chose to use the Neighbor-Joining (NJ) method [15, 16] to generate a phylogenetic tree.

4.2.3 Heterogeneity of mutation rates among conserved genes

Several strategies can be used to deal with the possible heterogeneity of mutation rates among conserved genes. The Gamma distribution is widely used to model these mutation rate differences. Models for estimating the heterogeneity of different genes have also been proposed in [8] where model parameters were estimated using a maximum-likelihood method.

Among all conserved genes, those that undergo active selection and evolve rapidly are usually the cause of a questionable tree structure. Removing these genes should improve the resulting phylogenetic tree [5]. For this purpose, the evolution rate for each gene is estimated. This requires an initial dendrogram which can either be one that has already been constructed or one that is constructed using all the genes to be used in the analysis. From this initial dendrogram the species that differs the most is chosen as the outlier species S_{out} and all other species are denoted as inner species S_{in}. The idea is to calculate a distance between a gene in the outermost species and the respective gene in each of the inner species [5]. We assume that this distance is related to the evolution rate of the gene under consideration. For each shared gene, a distance matrix is computed based on this shared gene assuming a JC model for mutations. An average branch length is computed using this distance matrix by finding the distances between S_{out} and each S_{in}, adding these values, and dividing by the total number of terms. The average branch length for each gene is used as its evolution rate. Genes with high values can be removed a few at a time and the effect on the resulting phylogenetic tree can be observed [1].

4.2.4 Bootstrap test

After obtaining a final phylogenetic tree for the plasmids, we must verify it. While it is possible that results can be verified based on expert knowledge, it is important to develop a method for verifying results based on the data given when such knowledge is unavailable. Bootstrapping is the most commonly used method for evaluating the reliability of a phylogenetic tree [3]. To find the bootstrap values in the phylogenetic tree, we assume that K conserved genes are used to construct a branch of the tree. We randomly select L conserved genes with replacement from these K conserved genes and use them to construct a branch. The process is repeated N times, and the branches with different structures are counted. The number of times a particular branch appears as a percentage of all N branches is used as a bootstrap confidence value. This is repeated for all successive branches.

4.3 Results and discussion

4.3.1 Analysis of bla$_{CMY_2}$ sequences

Bla$_{CMY_2}$ A/C plasmids have been reported in [10, 17–19]. In these papers, eight plasmid sequences (including pSN254, pAR060302, peH4H, and pAM04528 which were presented therein and four other A/C plasmids, pYR1, pIP1202, pP91278, and pP99018 from the NCBI database) were used for phylogenetic analysis. A second and independent phylogenetic analysis was conducted by comparing the gain and loss of large sequence segments among these plasmids. Through analysis of these segments, a parsimonious tree was constructed. The dendrogram presented in Fig. 11 shows that pYR1 is the outlier and is separate from all other groups. To evaluate the evolution rate for different genes, a distance matrix D_a for each gene a was computed based on the JC model. From this distance matrix, we found the average distance between pYR1 and the other seven strains which was considered the evolution rate for gene a. A histogram of all 91 conserved genes for all eight sequences is shown in Fig. 10. From this figure we see that the estimated evolution rates of three genes differ from those of the 88 remaining genes. We obtained dendrograms and performed bootstrapping tests first using all 91 conserved genes and then removing the three genes with different evolution rates and using the remaining 88 conserved genes. Comparing the results in Figs. 11 (a) and (b), we see that the results are almost identical. This suggests that divergent mutation rates for a minority of sequences (3/91) had no effect on the predicted phylogeny.

4.3.2 Analysis of Gram-negative plasmids

To test our method with a large dataset, 315 Gram-negative plasmid sequences were downloaded from the NCBI website, and a primitive tree was obtained using a virtual MPM. From this tree, 106 plasmids from Gammaproteobacteria were selected for further analysis. Because these 106 plasmids are not closely related, we were unable to find a set of shared genes for all of them. Instead we used a second virtual MPM to obtain a dendrogram of the 106 plasmid sequences Fig. 12, and from this dendrogram we identified several subgroups of plasmids. A subgroup of seven plasmids was chosen for which 17 conserved genes were identified and two dendrograms were generated as shown in Fig. 13. For one dendrogram, it was assumed that the rates of evolution were the same for all sites in all 17 genes; for the other, a Γ distribution of evolution rates was assumed for these sites. The differences between the two dendrograms indicate the importance of including evolution rates in the phylogenetic analysis.

4.4 Conclusions

In this chapter, we studied the problem of constructing a dendrogram for a set of plasmids that establishes their evolutionary relationships. Since the genetic contents of plasmids are mosaic and dynamic, we proposed a method for constructing a phylogenetic tree that uses only the conserved genes common to all plasmids. Using a set of bla$_{CMY_2}$ plasmids, we were able to show that the conserved genes that had been identified were homogeneous despite a difference in evolution rates. The high bootstrapping values obtained indicate that the dendrogram is very stable. However, the analysis of the method on a set of 106 Gram-negative plasmids suggests that different evolution rates in different sites of a gene may need to be considered. Other factors that might be important in building a phylogenetic tree for plasmids will be studied later.

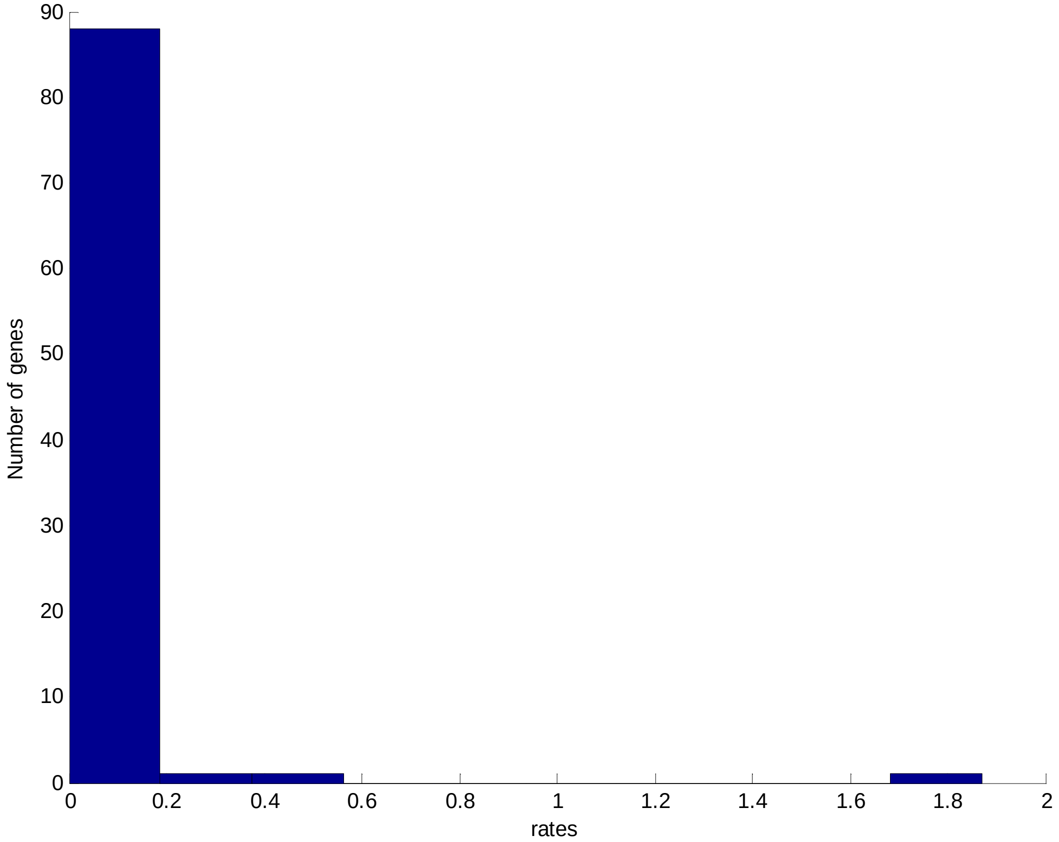


Figure 10. Histogram of the distribution of evolution rates for all 91 conserved genes.

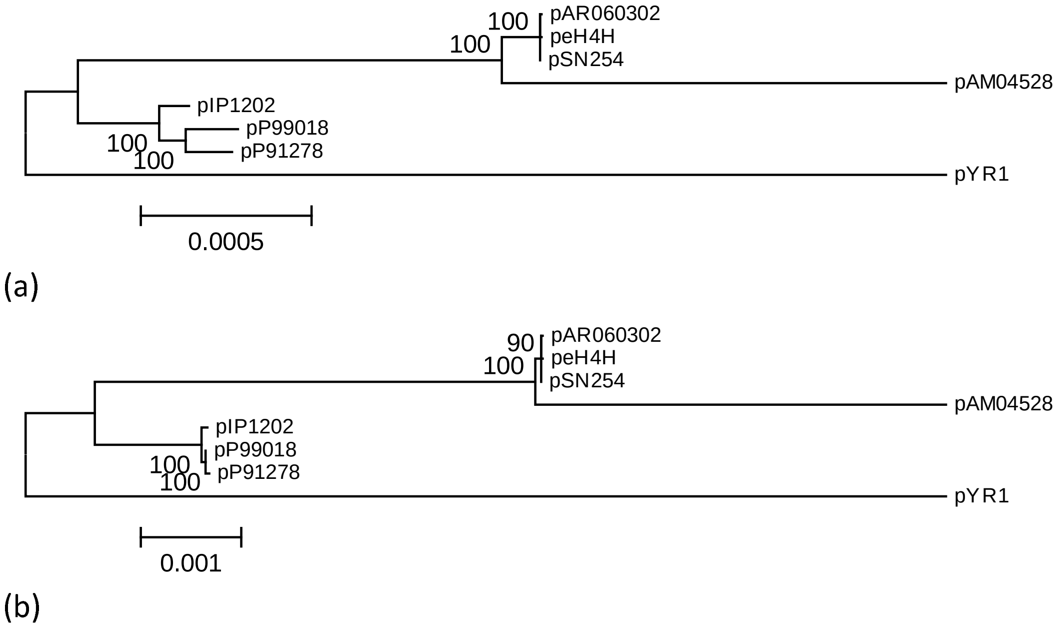


Figure 11. Dendrogram obtained using the Neighbor-Joining method with the Jukes-Cantor model and (a) all 91 conserved genes and (b) 88 conserved genes (3 genes with high evolution rates were removed).

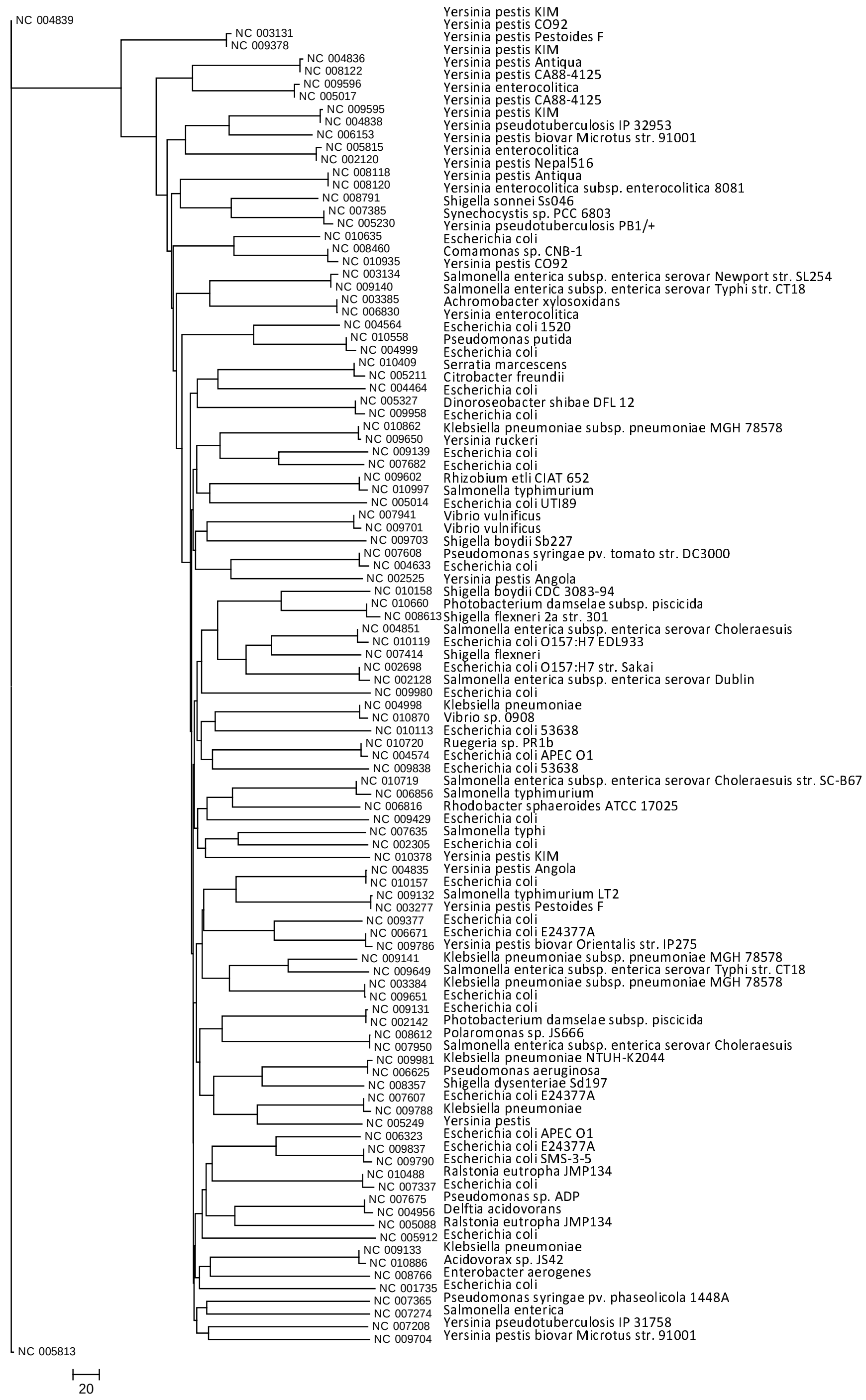


Figure 12. Dendrogram for 106 Gram-negative plasmids constructed using a virtual mixed-plasmid microarray.

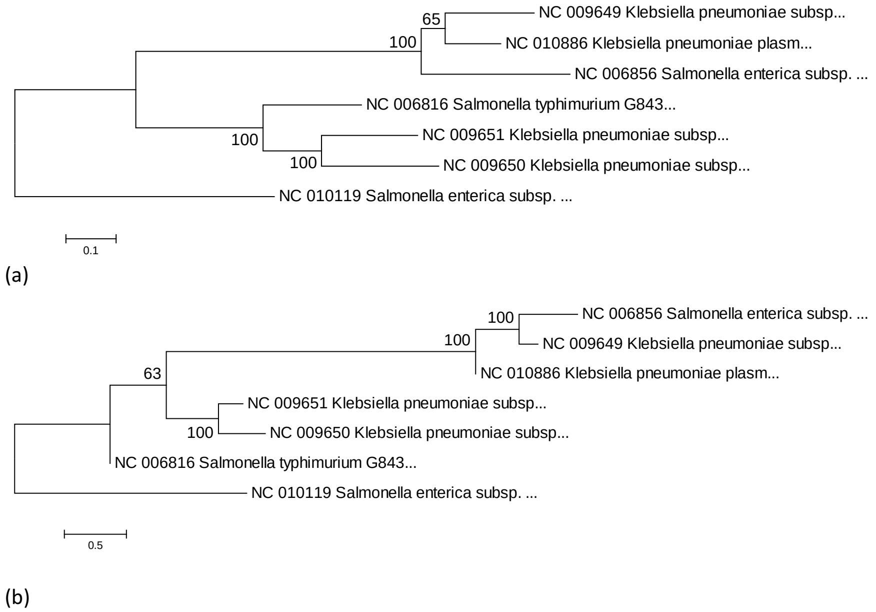

NC 009649 Klebsiella pneumoniae subsp...
NC 010886 Klebsiella pneumoniae plasm...
NC 006856 Salmonella enterica subsp. ...
NC 006816 Salmonella typhimurium G843...
NC 009651 Klebsiella pneumoniae subsp...
NC 009650 Klebsiella pneumoniae subsp...
NC 010119 Salmonella enterica subsp. ...
65
100
100
100
0.1
(a)
NC 006856 Salmonella enterica subsp. ...
NC 009649 Klebsiella pneumoniae subsp...
NC 010886 Klebsiella pneumoniae plasm...
NC 009651 Klebsiella pneumoniae subsp...
NC 009650 Klebsiella pneumoniae subsp...
NC 006816 Salmonella typhimurium G843...
NC 010119 Salmonella enterica subsp. ...
100
100
63
100
0.5
(b)

Figure 13. Dendrogram for a subset of 106 plasmids constructed from 17 conserved genes (a) assuming a Γ distribution of the evolution rates across sites and (b) assuming all sites have the same evolution rate.

List of References

[1] P. Lopez, P. Forterre, and H. Philippe, "The root of the tree of life in the light of the covarion model," *J Mol Evol*, vol. 49, no. 4, pp. 496–508, 1999.

[2] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, 2007.

[3] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies," *Nature*, vol. 425, pp. 798 – 804, 2003.

[4] S. V. Edwards, "Is a new and general theory of molecular systematics emerging?" *Evolution*, vol. 63, pp. 1–19, 2008.

[5] H. Philippe, N. Lartillot, and H. Brinkmann, "Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia," *Mol Biol Evol*, vol. 22, no. 5, pp. 1246–1253, 2005. [Online]. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/22/5/1246

[6] B. Kolaczkowski and J. W. Thornton, "A mixed branch length model of heterotachy improves phylogenetic accuracy," *Mol Biol Evol*, vol. 25, no. 6, pp. 1054–1066, 2008. [Online]. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/25/6/1054

[7] N. Lartillot and H. Philippe, "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Mol Biol Evol*, vol. 21, no. 6, pp. 1095–1109, 2004.

[8] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *J Mol Evol*, vol. 42, no. 5, pp. 587–596, May 1996. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/8662011

[9] Z. Yang, "Among-site rate variation and its impact on phylogenetic analyses," *Trends Ecol Evol*, vol. 11, no. 9, pp. 367 – 372, 1996.

[10] D. R. Call, R. Singer, D. Meng, S. L. Broschat, L. H. Orfe, J. Anderson, D. R. Herndon, L. S. Kappmeyer, J. B. Daniels, and T. E. Besser, "$bla_{\mathrm{CMY-2}}$ positive Inc A/C plasmids from *Escherichia coli* and *Salmonella enterica* are a distinct component of a larger lineage of plasmids," *Antimicrob. Agents Chemother. (In review).*

[11] L. Gross, "Are "ultraconserved" genetic elements really indispensable?" *PLoS Biology*, vol. 5, no. 9, p. e253, 2007.

[12] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool." *J Mol Biol*, vol. 215, no. 3, pp. 403–410, October 1990.

[13] Y. Wan, S. L. Broschat, and D. R. Call, "Validation of mixed-genome microarrays as a method for genetic discrimination," *Appl. Environ. Microbiol.*, vol. 73, no. 5, pp. 1425–1432, 2007.

[14] J. T. H. and C. R. Cantor, "Evolution of protein molecules," in *Mammalian protein metabolism*, H. N. Munro, Ed. New York: Academic Press, 1969, pp. 21–132.

[15] T. Mailund, G. Brodal, R. Fagerberg, C. Pedersen, and D. Phillips, "Recrafting the neighbor-joining method," *BMC Bioinformatics*, vol. 7, no. 1, p. 29, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/29

[16] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol Biol Evol*, vol. 4, no. 4, pp. 406–425, 1987.

[17] A. Carattoli, F. Tosini, W. P. Giles, M. E. Rupp, S. H. Hinrichs, F. J. Angulo, T. J. Barrett, and P. D. Fey, "Characterization of plasmids carrying CMY-2 from expanded-spectrum cephalosporin-resistant *salmonella* strains isolated in the United States between 1996 and 1998," *Antimicrob. Agents Chemother.*, vol. 46, no. 5, pp. 1269–1272, 2002.

[18] T. J. Welch, *et al.*, "Multiple antimicrobial resistance in plague: An emerging public health risk," *PLoS ONE*, vol. 2, no. 3, p. e309, 2007.

[19] M.-J. Kim, I. Hirono, K. Kurokawa, T. Maki, J. Hawke, H. Kondo, M. D. Santos, and T. Aoki, "Complete DNA sequence and analysis of the transferable multiple-drug resistance plasmids (R plasmids) from *photobacterium damselae* subsp. *piscicida* isolates collected in Japan and the United States ," *Antimicrob. Agents Chemother.*, vol. 52, no. 2, pp. 606–611, 2008.

CHAPTER 5

Conclusions

5.1 Conclusions

In Chapter 2 we described a new software tool for selecting a set of probes for a classification microarray. While the tool was developed for the design of mixed microarrays—and mixed-plasmid microarrays in particular—it can also be used to design expression arrays. The user can choose from several clustering methods (including hierarchical, non-hierarchical, and a model-based genetic algorithm), several probe ranking methods, and several different display methods. A novel approach is used for probe redundancy reduction, and probe selection is accomplished via stepwise discriminant analysis. Data can be entered in different formats (including Excel and comma-delimited text), and dendrogram, heat map, and scatter plot images can be saved in several different formats (including jpeg and tiff). Weights generated using stepwise discriminant analysis can be stored for analysis of subsequent experimental data. Additionally, PLASMID can be used to construct virtual microarrays with genomes from public databases, which can then be used to identify an optimal set of probes.

Determining phylogenetic relationships between bacterial strains is important in molecular epidemiology studies. Two molecular typing methods, pulse-field gel electrophoresis (PFGE) and multiple-loci variable-number tandem repeat analysis (MLVA), are widely used in such studies. In Chapter 3, we proposed a fusion algorithm that combines the information obtained from both PFGE and MLVA assays to obtain phylogenetic relationships. Two sets of *Salmonella enterica* were examined; one set included serovar Typhimurium isolates from a wide range of sampling dates, locations, and host species while the other set included a group of serovar Newport isolates collected over a limited geographic and temporal scale. Results were assessed by comparison with phage-typing assays and with known epidemiological relationships. The analysis showed that the fusion algorithm provides an improved ability to discriminate between isolates and to infer phylogenetic relationships compared with using either the PFGE or MLVA method alone.

In Chapter 4, we discussed methods for improving the accuracy of phylogenetic analysis. Phylogenetic analysis of plasmids is complicated by the dynamic nature of plasmid gene sequences. Considerable genetic content is obtained via horizontal gene transfer. We proposed a method for including only conserved genes for phylogenetic analysis. Experimental results on a set of eight plasmids showed that the effect of gene heterogeneity had been reduced. Further analysis of a group of Gram-negative, enteric plasmids from the NCBI database showed that using a model that incorporates the different evolution rates among different sites of gene sequences can improve the phylogenetic analysis of plasmids.

5.2 Future Work

More work can be done to improve PLASMID. Additional functions can be included in the tool, for example, functions to design comprehensive microarrays and to design PCR mapping assays. With more annotated information available in public databases, it would be useful to extend PLASMID to utilize this information to identify biomarkers. Finally, it is hoped that PLASMID will be used to develop actual mixed microarrays and that its utility will be fully realized and tested.

We limited our discussion of the fusion algorithm to using two different datasets to improve our understanding of epidemiological relationships among bacteria. In real applications, there is increasingly more information that can be combined and exploited for greater precision and understanding. For example, text mining methods have been proposed to search in public publication databases for available knowledge related to a biological problem. To handle mixed and complicated data or knowledge together, more complex computation schemes need to be developed. For example, a Bayesian network has been used to detect the causal associations among genes and diseases by combining microarray data and genotypic data together [1].

For phylogenetic analyses of plasmids, a better model is needed for the dynamics of plasmid gene sequences. Different characteristics may contribute more information beyond that of the sequence itself. For example, conservation of synteny has been used to study plasmid evolution [2]. Another possibility is to combine the insertions and deletions of genes in the phylogenetic analysis. Novel methods should also be considered such as considering phylogenetic networks rather than phylogenetic trees; network methods have been used to handle the complexity of microbes with dynamic gene content [3]. In addition to finding a robust phylogenetic relationship among plasmids, we need to develop a classifier for inserting new plasmids into an existing tree.

Determining the relationships between microbial strains is only the first step. Our ultimate goal is to understand the underlying biological mechanisms, e.g., how the strains evolved over time. Ideally, we would like to be able to create realistic models of processes such as evolution which would allow us to predict future events. In order to accomplish this, we will need to develop more advanced computational methods to use with the vast amount of data available both now and in the future.

List of References

[1] J. Zhu, M. C. Wiener, C. Zhang, A. Fridman, E. Minch, P. Y. Lum, J. R. Sachs, and E. E. Schadt, "Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations," *PLoS Comput Biol*, no. 4, p. e69, 04 2007.

[2] E. E. Eichler and D. Sankoff, "Structural dynamics of eukaryotic chromosome evolution," *Science*, vol. 301, no. 5634, pp. 793–797, August 2003.

[3] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Mol Biol Evol*, vol. 23, no. 2, pp. 254–267, February 2006.

BIBLIOGRAPHY

Aardema, M. J. and MacGregor, J. T., "Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies," *Mutat. Res. Fundam. Mol. Mech. Mugag.*, vol. 499, pp. 13–25, 2002.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., "Basic local alignment search tool." *J Mol Biol*, vol. 215, pp. 403–410, October 1990.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucl. Acids Res.*, vol. 25, pp. 3389–3402, 1997.

Bauer, A., Kirby, M., Sherris, J., and Turck, M., "Antibiotic susceptibility testing by a standard single disk method," *Am J Clin Path*, vol. 45, pp. 493–496, 1966.

Benson, G., "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic Acids Res*, vol. 27, pp. 573–580., 1999.

Bezanson, G., Khakhria, R., and Lacroix., R., "Involvement of plasmids in determining bacteriophage sensitivity in *Salmonella typhimurium*: genetic and physical analysis of phagovar 204," *Can J Microbiol*, vol. 28, pp. 993 – 1001, 1982.

Boerlin, P. and Reid-Smith, R. J., "Antimicrobial resistance: its emergence and transmission," *Anim Health Res Rev*, vol. 9, pp. 115–126, 2008.

Borucki, M. K., Kim, S. H., Call, D. R., Smole, S. C., and Pagotto, F., "Selective discrimination of *Listeria monocytogenes* epidemic strains by a mixed-genome DNA microarray compared to discrimination by pulsed-field gel electrophoresis, ribotyping, and multilocus sequence typing," *J Clin Microbiol*, vol. 42, pp. 5270–5276, 2004.

Borucki, M. K., Krug, M. J., Muraoka, W. T., and Call, D. R., "Discrimination among *Listeria monocytogenes* isolates using a mixed genome DNA microarray," *Vet Microbiol*, vol. 92, pp. 351–362, 2003.

Bray, N. and Pachter, L., "MAVID: Constrained ancestral alignment of multiple sequences," *Genome Research*, vol. 14, pp. 693–699, 2004.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Program, N. C. S., Green, E. D., Sidow, A., and Batzoglou, S., "LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA," *Genome Research*, vol. 13, pp. 721–731, 2003.

Call, D. R., Orfe, L., Davis, M. A., Lafrentz, S., and Kang, M. S., "Impact of compounding error on strategies for subtyping pathogenic bacteria," *Foodborne Pathog Dis*, vol. 5, pp. 505–16, 2008.

Call, D. R., Borucki, M. K., and Besser, T. E., "Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of *Listeria monocytogenes*," *J Clin Microbiol*, vol. 41, pp. 632–639, 2003.

Call, D. R., Singer, R., Meng, D., Broschat, S. L., Orfe, L. H., Anderson, J., Herndon, D. R., Kappmeyer, L. S., Daniels, J. B., and Besser, T. E., "$bla_{\mathrm{CMY-2}}$ positive Inc A/C plasmids from *Escherichia coli* and *Salmonella enterica* are a distinct component of a larger lineage of plasmids," *Antimicrob. Agents Chemother. (In review).*

Call, D., Kang, M.-S., Daniels, J., and Besser, T., "Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray," *J Appl Microbiol*, vol. 100, pp. 15–28, 2006.

Carattoli, A., Tosini, F., Giles, W. P., Rupp, M. E., Hinrichs, S. H., Angulo, F. J., Barrett, T. J., and Fey, P. D., "Characterization of plasmids carrying CMY-2 from expanded-spectrum cephalosporin-resistant *salmonella* strains isolated in the United States between 1996 and 1998," *Antimicrob. Agents Chemother.*, vol. 46, pp. 1269–1272, 2002.

CDC, "National antimicrobial resistance monitoring system for enteric bacteria (NARMS). Human isolates final report, 2005." *U.S. Department of Health and Human Services*, 2008.

Chen, D., Liu, Z., Ma, X., and Hua, D., "Selecting genes by test statistics," *J Biomed Biotechnol.*, vol. 2, pp. 132–138, 2005.

Chen, K.-W., Lo, H.-J., Lin, Y.-H., and Li, S.-Y., "Comparison of four molecular typing methods to assess genetic relatedness of *Candida albicans* clinical isolates in Taiwan," *J Med Microbiol*, vol. 54, pp. 249–258, 2005.

Chou, C.-C., Chen, C.-H., Lee, T.-T., and Peck, K., "Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression," *Nucleic Acids Res*, vol. 32, pp. e99–, 2004.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje, J. M., "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis," *Nucleic Acids Res*, vol. 37, pp. D141–145, 2009.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T., "Mauve: multiple alignment of conserved genomic sequence with rearrangements," *Genome Res*, vol. 14, pp. 1394–1403, 2004. [Online]. Available: http://genome.cshlp.org/content/14/7/1394.abstract

Davis, M. A., Hancock, D. D., Besser, T. E., and Call, D. R., "Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of *Escherichia coli* O157:H7," *J Clin Microbiol*, vol. 41, pp. 1843–9, 2003.

Davis, M., Baker, K., Call, D., Warnick, L., Soyer, Y., Wiedmann, M., Gröhn, Y., McDonough, P., Hancock, D., and Besser, T., "Multiple locus variable number of tandem repeats typing method for *Salmonella enterica* serovar Newport," *J Clin Microbiol*, vol. in press., 2009.

Dietterich, T. G., "Ensemble methods in machine learning," *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

Doolittle, W. F., "Phylogenetic classification and the universal tree," *Science*, vol. 284, pp. 2124–2128, 1999.

Edwards, S. V., "Is a new and general theory of molecular systematics emerging?" *Evolution*, vol. 63, pp. 1–19, 2008.

Eichler, E. E. and Sankoff, D., "Structural dynamics of eukaryotic chromosome evolution," *Science*, vol. 301, pp. 793–797, August 2003.

Emrich, S. J., Lowe, M., and Delcher, A. L., "PROBEmer: a web-based software tool for selecting optimal DNA oligos," *Nucleic Acids Res*, vol. 31, pp. 3746–3750, 2003.

Fakhr, M. K., Nolan, L. K., and Logue, C. M., "Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella enterica* serovar Typhimurium," *J Clin Microbiol*, vol. 43, pp. 2215–9, 2005.

Felsenstein, J., "Phylip (phylogeny inference package) version 3.5c. distributed by the author. department of genetics, university of washington, seattle," 1993.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., and al., e., "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd," *Science*, vol. 269, pp. 496–512, 1995.

Foley, S. L., Zhao, S., and Walker, R. D., "Comparison of molecular typing methods for the differentiation of *Salmonella* foodborne pathogens," *Foodborne Pathog Dis*, vol. 4, pp. 253–76, 2007.

Gamberoni, G., Storari, S., and Volinia, S., "Finding biological process modifications in cancer tissues by mining gene expression correlations," *BMC Bioinformatics*, vol. 7, p. 6, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/6

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytia-Trees, E., Ribot, E. M., and Swaminathan, B., “PulseNet USA: a five-year update,” *Foodborne Pathog Dis*, vol. 3, pp. 9–19, 2006.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S., “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531–537, 1999.

Gross, L., “Are “ultraconserved” genetic elements really indispensable?” *PLoS Biology*, vol. 5, p. e253, 2007.

Gulati, P., Varshney, R. K., and Virdi, J. S., “Multilocus variable number tandem repeat analysis as a tool to discern genetic relationships among strains of *Yersinia enterocolitica* biovar 1A,” *J Appl Microbiol*, 2009.

H., J. T. and Cantor, C. R., “Evolution of protein molecules,” in *Mammalian protein metabolism*, Munro, H. N., Ed. New York: Academic Press, 1969, pp. 21–132.

Hanczar, B., Courtine, M., Benis, A., Hennegar, C., Clément, K., and Zucker, J.-D., “Improving classification of microarray data using prototype-based feature selection,” *SIGKDD Explor. Newsl.*, vol. 5, pp. 23–30, 2003.

Harbottle, H., White, D. G., McDermott, P. F., Walker, R. D., and Zhao, S., “Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates,” *J Clin Microbiol*, vol. 44, pp. 2449–57, 2006.

Hathaway, L. J., Brugger, S., Martynova, A., Aebi, S., and Muhlemann, K., “Use of the Agilent 2100 bioanalyzer for rapid and reproducible molecular typing of *streptococcus pneumoniae*,” *J Clin Microbiol*, vol. 45, pp. 803–809, 2007.

Hopkins, K. L., Maguire, C., Best, E., Liebana, E., and Threlfall, E. J., “Stability of multiple-locus variable-number tandem repeats in *Salmonella enterica* serovar typhimurium,” *J Clin Microbiol*, vol. 45, pp. 3058–61, 2007.

Huson, D. H. and Bryant, D., “Application of phylogenetic networks in evolutionary studies,” *Mol Biol Evol*, vol. 23, pp. 254–267, February 2006.

Hyyro, H., Juhola, M., and Vihinen, M., “Genome-wide selection of unique and valid oligonucleotides,” *Nucleic Acids Res*, vol. 33, pp. e115–, 2005.

Içgen, B., Gürakan, G. C., and Özcengiz, G., “Effects of plasmid curing on antibiotic susceptibility, phage type, lipopoly saccharide and outer membrane protein profiles in local *Salmonella* isolates,” *Food Microbiol*, vol. 18, pp. 631 – 635, 2001.

Jaeger, J., Sengupta, R., and Ruzzo, W., “Improved gene selection for classification of microarrays,” *Pac. Symp. Biocomput.*, pp. 53–64, 2003.

Jagota, A., *Microarray Data Analysis and Visualization.* Bioinformatics By The Bay Press, 2001.

Jain, A. K., Duin, R. P., and Mao, J., “Statistical pattern recognition: a review,” *IEEE Trans Pattern Anal Mach Intell*, vol. 22, pp. 4–37, 2000.

Jennrich, R. I., “Stepwise discriminant analysis,” in *Statistical methods for digital computers*, Enslein, K., Ed. John Wiley & Sons Inc, 1977, vol. III, pp. 76–95.

Kanehisa, M. and Bork, P., “Bioinformatics in the post-sequence era,” *Nat Genet*, vol. 33, pp. 305 – 310, 2003.

Kang, M.-S., Besser, T. E., and Call, D. R., “Variability in the region downstream of the $bla_{\mathrm{CMY-2}}$ β–lactamase gene in *Escherichia coli* and *Salmonella enterica* plasmids,” *Antimicrob. Agents Chemother.*, vol. 50, pp. 1590–1593, 2006.

Kim, M.-J., Hirono, I., Kurokawa, K., Maki, T., Hawke, J., Kondo, H., Santos, M. D., and Aoki, T., “Complete DNA sequence and analysis of the transferable multiple-drug resistance plasmids (R plasmids) from *photobacterium damselae* subsp. *piscicida* isolates collected in Japan and the United States ,” *Antimicrob. Agents Chemother.*, vol. 52, pp. 606–611, 2008.

Kolaczkowski, B. and Thornton, J. W., “A mixed branch length model of heterotachy improves phylogenetic accuracy,” *Mol Biol Evol*, vol. 25, pp. 1054–1066, 2008. [Online]. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/25/6/1054

Koonin, E. V. and Galperin, M. Y., “Prokaryotic genomes: the emerging paradigm of genome-based microbiology,” *Curr Opin Genetics Dev*, vol. 7, pp. 757 – 763, 1997.

Kubota, K., Barrett, T. J., Ackers, M. L., Brachman, P. S., and Mintz, E. D., “Analysis of *Salmonella enterica* serotype typhi pulsed-field gel electrophoresis patterns associated with international travel,” *J Clin Microbiol*, vol. 43, pp. 1205–1209, 2005.

Lartillot, N. and Philippe, H., “A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process,” *Mol Biol Evol*, vol. 21, pp. 1095–1109, 2004.

Li, H. and Wang, W., “Dissecting the transcription networks of a cell using computational genomics,” *Curr Opin Genetics Dev*, vol. 13, pp. 611 – 616, 2003.

Lindstedt, B. A., Heir, E., Gjernes, E., and Kapperud, G., "DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci," *J Clin Microbiol*, vol. 41, pp. 1469–79, 2003.

Lindstedt, B. A., Torpdahl, M., Nielsen, E. M., Vardund, T., Aas, L., and Kapperud, G., "Harmonization of the multiple-locus variable-number tandem repeat analysis method between Denmark and Norway for typing *Salmonella* typhimurium isolates and closer examination of the VNTR loci," *J Appl Microbiol*, vol. 102, pp. 728–35, 2007.

Lindstedt, B. A., Vardund, T., Aas, L., and Kapperud, G., "Multiple-locus variable-number tandem-repeats analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium using PCR multiplexing and multicolor capillary electrophoresis," *J Microbiol Methods*, vol. 59, pp. 163–72, 2004.

Ling, J. M., Lo, N. W. S., Ho, Y. M., Kam, K. M., Hoa, N. T. T., Phi, L. T., and Cheng, A. F., "Molecular methods for the epidemiological typing of *Salmonella enterica* serotype typhi from Hong Kong and Vietnam," *J Clin Microbiol*, vol. 38, pp. 292–300, 2000.

Lipman, D. and Pearson, W., "Rapid and sensitive protein similarity searches," *Science*, vol. 227, pp. 1435–1441, 1985.

Loots, G. G., "Chapter 10 genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis," in *Long-Range Control of Gene Expression*, ser. Advances in Genetics, van Heyningen, V. and Hill, R. E., Eds.   Academic Press, 2008, vol. 61, pp. 269 – 293.

Lopez, P., Forterre, P., and Philippe, H., "The root of the tree of life in the light of the covarion model," *J Mol Evol*, vol. 49, pp. 496–508, 1999.

Lukinmaa, S., Nakari, U. M., Eklund, M., and Siitonen, A., "Application of molecular genetic methods in diagnostics and epidemiology of food-borne bacterial pathogens," *APMIS*, vol. 112, pp. 908–29, 2004.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G., "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms," *PNAS*, vol. 95, pp. 3140–3145, 1998.

Mailund, T., Brodal, G., Fagerberg, R., Pedersen, C., and Phillips, D., "Recrafting the neighbor-joining method," *BMC Bioinformatics*, vol. 7, p. 29, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/29

Naser, S., Thompson, F. L., Hoste, B., Gevers, D., Vandemeulebroecke, K., Cleenwerck, I., Thompson, C. C., Vancanneyt, M., and Swings, J., “Phylogeny and identification of Enterococci by *atpa* gene sequence analysis,” *J Clin Microbiol*, vol. 43, pp. 2224–2230, 2005.

National Committee for Clinical Laboratory Standards, “Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: Approved standard M7-A6,” *NCCLS Villanova, PA, USA*, 2003.

National Committee for Clinical Laboratory Standards, “Performance standards for antimicrobial susceptibility testing, 14th informational supplement, 13th ed. approved standard M100-S13.” *NCCLS, Wayne, Pa.*, 2003.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A., and Blattner, F. R., “Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7,” *Nature*, vol. 409, pp. 529–533, 2001.

Philippe, H., Lartillot, N., and Brinkmann, H., “Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia,” *Mol Biol Evol*, vol. 22, pp. 1246–1253, 2005. [Online]. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/22/5/1246

Porwollik, S., Wong, R. M.-Y., and McClelland, M., “Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis,” *PNAS*, vol. 99, pp. 8956–8961, 2002.

Qin, Z. S., “Clustering microarray gene expression data using weighted chinese restaurant process,” *Bioinformatics*, vol. 22, pp. 1988–1997, 2006.

Raes, J., Foerstner, K. U., and Bork, P., “Get the most out of your metagenome: computational analysis of environmental sequence data,” *Curr Opin Microbiol*, vol. 10, pp. 490 – 498, 2007.

Raskin, D. M., Seshadri, R., Pukatzki, S. U., and Mekalanos, J. J., “Bacterial genomics and pathogen evolution,” *Cell*, vol. 124, pp. 703 – 714, 2006.

Ribot, E. M., Fair, M. A., Gautom, R., Cameron, D. N., Hunter, S. B., Swaminathan, B., and Barrett, T. J., “Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet,” *Foodborne Pathog Dis*, vol. 3, pp. 59–67, 2006.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B., “Genome-scale approaches to resolving incongruence in molecular phylogenies,” *Nature*, vol. 425, pp. 798 – 804, 2003.

Ross, I. L. and Heuzenroeder, M. W., “Use of AFLP and PFGE to discriminate between *Salmonella enterica* serovar Typhimurium DT126 isolates from separate food-related outbreaks in Australia,” *Epidemiol Infect*, vol. 133, pp. 635–644, 2005.

Saitou, N. and Nei, M., “The neighbor-joining method: A new method for reconstructing phylogenetic trees,” *Mol Biol Evol*, vol. 4, pp. 406–425, 1987.

Smith, T. F. and Waterman, M. S., “Identification of common molecular subsequences,” *J Mol Bio*, vol. 147, pp. 195 – 197, 1981.

Somorjai, R., Dolenko, B., and Baumgartner, R., “Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions,” *Bioinformatics*, vol. 19, pp. 1484–1491, 2003.

Soule, M., Kuhn, E., Loge, F., Gay, J., and Call, D., “Using DNA microarrays to identify library-independent markers for bacterial source tracking,” *Appl. Environ. Microbiol.*, vol. 72, pp. 1843–1851, 2006.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S., “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,” *Bioinformatics*, vol. 21, pp. 631–643, 2005.

Stinchcombe, J. R. and Hoekstra, H. E., “Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits,” *Heredity*, vol. 100, pp. 158–170, 2007.

Su, Y., Murali, T., Pavlovic, V., Schaffer, M., and Kasif, S., “RankGene: identification of diagnostic genes based on expression data,” *Bioinformatics*, vol. 19, pp. 1578–1579, 2003.

Summers, D. K., *The Biology of Plasmids.* Oxford: Blackwell Science, 1996.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V., “The COG database: new developments in phylogenetic classification of proteins from complete genomes,” *Nucleic Acids Res*, vol. 29, pp. 22–28, 2001.

Tenover, F. C., Arbeit, R. D., and Goering, R. V., “How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists,” *Infect. Control Hosp. Epidemiol.*, vol. 18, p. 426439, 1997.

Thomas, C. M., *The Horizontal gene pool : bacterial plasmids and gene spread.* Amsterdam, The Netherlands: Harwood Academic, 2000.

Thompson, J. D., Higgins, D. G., and Gibson, T. J., "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, pp. 4673–4680, 1994.

Toolan, T. M. and Tufts, D. W., "Detection and estimation in non-stationary environments," in *Proceedings IEEE Asilomar Conference on Signals, Systems & Computers*, Nov. 2003, pp. 797–801.

Trevino, V., Falciani, F., and Barrera-Saldaa, H. A., "DNA microarrays: a powerful genomic tool for biomedical and clinical research," *Mol Med.*, vol. 13, pp. 527 – 541, 2007.

Uchiyama, I., "MBGD: microbial genome database for comparative analysis," *Nucleic Acids Res*, vol. 31, pp. 58–62, 2003.

Urwin, R. and Maiden, M. C. J., "Multi-locus sequence typing: a tool for global epidemiology," *TIM*, vol. 11, pp. 479 – 487, 2003.

Van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H., "Short-sequence DNA repeats in prokaryotic genomes," *Microbiol Mol Biol Rev*, vol. 62, pp. 275–93, 1998.

Van Hellemont, R., Monsieurs, P., Thijs, G., De Moor, B., Van de Peer, Y., and Marchal, K., "A novel approach to identifying regulatory motifs in distantly related genomes," *Genome Biol.*, vol. 6, 2005. [Online]. Available: http://dx.doi.org/10.1186/gb-2005-6-13-r113

Van Riel, N. A., "Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments," *Brief Bioinform*, vol. 7, pp. 364–374, 2006. [Online]. Available: http://bib.oxfordjournals.org/cgi/content/abstract/7/4/364

Vogler, A. J., Keys, C., Nemoto, Y., Colman, R. E., Jay, Z., and Keim, P., "Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7," *J Bacteriol*, vol. 188, pp. 4253–63, 2006.

Vogler, A. J., Keys, C. E., Allender, C., Bailey, I., Girard, J., Pearson, T., Smith, K. L., Wagner, D. M., and Keim, P., "Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*," *Mutat. Res.*, vol. 616, pp. 145–58, 2007.

Walsh, B., "Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals," *Genetics*, vol. 158, pp. 897–912, 2001. [Online]. Available: http://www.genetics.org/cgi/content/abstract/158/2/897

Wan, Y., Broschat, S. L., and Call, D. R., "Validation of mixed-genome microarrays as a method for genetic discrimination," *Appl. Environ. Microbiol.*, vol. 73, pp. 1425–1432, 2007.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R., "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, vol. 73, pp. 5261–5267, 2007.

Wang, Y., Makedon, F. S., Ford, J. C., and Pearlman, J., "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, pp. 1530–1537, 2005.

Welch, T. J., Fricke, W. F., McDermott, P. F., White, D. G., Rosso, M.-L., Rasko, D. A., Mammel, M. K., Eppinger, M., Rosovitz, M., Wagner, D., Rahalison, L., LeClerc, J. E., Hinshaw, J. M., Lindler, L. E., Cebula, T. A., Carniel, E., and Ravel, J., "Multiple antimicrobial resistance in plague: An emerging public health risk," *PLoS ONE*, vol. 2, p. e309, 2007.

Welsh, J. and McClelland, M., "Fingerprinting genomes using PCR with arbitrary primers," *Nucleic Acids Res*, vol. 18, pp. 7213–7218, 1990.

Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J., and Tingey, S. V., "DNA polymorphisms amplified by arbitrary primers are useful as genetic markers," *Nucleic Acids Res*, vol. 18, pp. 6531–6535, 1990.

Woese, C. R., "Bacterial evolution." *Microbiol. Mol. Biol. Rev.*, vol. 51, pp. 221–271, 1987.

Yang, Z., "Maximum-likelihood models for combined analyses of multiple sequence data," *J Mol Evol*, vol. 42, pp. 587–596, May 1996. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/8662011

Yang, Z., "Among-site rate variation and its impact on phylogenetic analyses," *Trends Ecol Evol*, vol. 11, pp. 367 – 372, 1996.

Zhang, R. and Zhang, C.-T., "The impact of comparative genomics on infectious disease research," *Microbes Infect*, vol. 8, pp. 1613 – 1622, 2006.

Zheng, J., Keys, C. E., Zhao, S., Meng, J., and Brown, E. W., "Enhanced subtyping scheme for *Salmonella enteritidis*," *Emerg Infect Dis*, vol. 13, pp. 1932–5, 2007.

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., Sachs, J. R., and Schadt, E. E., “Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations,” *PLoS Comput Biol*, p. e69, 04 2007.